

**Multifidelity Methods for Design of Transition Metal  
Complexes**

by

Jon Paul Janet

B.Sc. Eng., The University of Cape Town (2012)

M.Sc., KTH Royal Institute of Technology (2015)

M.Sc., Technical University of Berlin (2015)

Submitted to the Department of Chemical Engineering & Center for  
Computational Engineering

in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy in Chemical Engineering and Computation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author.....

Department of Chemical Engineering & Center for Computational Engineering

18<sup>th</sup> December, 2019

Certified by .....

Heather J. Kulik

Department of Chemical Engineering

Thesis Supervisor

Certified by .....

Youssef Marzouk

Department of Aeronautics and Astronautics

Thesis Supervisor

Accepted by .....

Patrick S. Doyle

Robert T. Haslam (1911) Professor of Chemical Engineering

Graduate Officer

Accepted by .....

Nicolas Hadjiconstantinou

Professor of Mechanical Engineering

CCE Co-Director

**ARCHIVAL COPY 3: Center for Computational Engineering**



# Multifidelity Methods for Design of Transition Metal Complexes

by

Jon Paul Janet

Submitted to the Department of Chemical Engineering & Center for  
Computational Engineering  
on 18<sup>th</sup> December, 2019, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy in Chemical Engineering and Computation

## Abstract

The rational design of materials with tightly controlled properties is crucial to addressing future challenges in energy, electronics and catalysis. While improvements in computing power have made simulation with density functional theory (DFT) an essential tool in screening new materials, it remains too costly to address truly high-dimensional design spaces. This problem is especially acute for open-shell transition metal (TM) complexes, which are of central importance in homogeneous catalysis and have applications in solar energy and electronics. The space of TM complexes is enormous and poorly characterized, while DFT calculations for these systems are expensive and sensitive to method choice, making it impractical to simulate large numbers of candidates indiscriminately. This makes the search for TM complexes with desired properties a formidable challenge. This thesis addresses this challenge by formulating strategies for materials design that exploit insights from data-driven surrogate models together with first-principles simulations. A framework for data-driven inference of the quantum properties of TM complexes is developed, using artificial neural networks (ANNs) and graph-based molecular representations that facilitate rapid screening while retaining physical meaning such that chemical insights can be extracted. Multiple sources of uncertainty that would limit the application of these methods to TM complexes are addressed. Surrogate models are trained to estimate system-specific DFT uncertainty by including data from DFT calculations with different fractions of exact exchange, and a novel uncertainty metric for data-driven discovery is proposed that quantifies the ability of ANNs to generalize to unseen

data based on similarity in the learned latent space. This metric is shown to offer superior performance over existing methods. The application of these methods to virtual design problems is demonstrated with two case studies: 1) identifying spin crossover complexes from a design space of thousands using an evolutionary strategy and 2) probabilistic, multiobjective optimization of redox couples over a 3 million-complex space. The utility of this surrogate-assisted approach is evident and orders-of-magnitude accelerations are obtained over screening purely with DFT. Such strategies open the door for *in silico* design of some of the most challenging molecular systems at a far greater scale than ever before.

Thesis Supervisor: Heather J. Kulik  
Title: Department of Chemical Engineering

Thesis Supervisor: Youssef Marzouk  
Title: Department of Aeronautics and Astronautics

## Acknowledgments

I would like to thank my primary advisor, Professor Heather Kulik, for the enormous amount of guidance, support and advice on all aspects of my graduate experience, and for being deeply involved in every aspect of this work. Heather is not only an amazing scientist, but a principled leader who gives generously of her time to develop her students both scientifically and professionally. I am extremely grateful that Heather took me on in my first year, with no relevant experience whatsoever, and enabled me to conduct the research I wanted to do. I would also like to thank Professor Youssef Marzouk for his excellent advice and contagious curiosity. Our discussions, and the many readings groups, have helped me to broaden my scientific horizons and frame the challenges in this project in a consistent and logical manner. Thanks are also owed to my committee members, Professors Jeffrey Grossman and Richard Braatz, for their valuable input on the progress of this thesis.

I have been highly fortunate to benefit from the advice and camaraderie of group members past and present, and in particular the inorganic subgroup, whose efforts have supported the later chapters of this thesis. My gratitude extends to my friends, both preexisting and those I have acquired along the way, for the company and levity. To my family and especially my parents: thank you for setting me on this course and encouraging me at every turn.

Finally; to my partner, Astrid, without whose support I would never had the courage, tenacity or self-discipline to make it through these years: thank you sincerely for all of your efforts over the last decade – not least of which your critical reading of this document – and I look forward to our next steps.



# Contents

<b>List of Figures</b>	<b>11</b>
<b>List of Tables</b>	<b>19</b>
<b>1 Introduction</b>	<b>25</b>
1.1 Rational design of transition metal complexes . . . . .	25
1.2 Thesis overview . . . . .	29
<b>2 Background</b>	<b>33</b>
2.1 First-principles virtual screening . . . . .	33
2.1.1 Overview . . . . .	33
2.1.2 Density functional theory . . . . .	37
2.1.3 Challenges for transition metal chemistry . . . . .	40
2.2 Machine learning in chemistry . . . . .	46
2.2.1 Overview . . . . .	46
2.2.2 Representations . . . . .	58
2.2.3 Models . . . . .	70
2.3 Using surrogate models for chemical design . . . . .	102

<b>3</b>	<b>Surrogate models for transition metal complexes</b>	<b>113</b>
3.1	Introduction . . . . .	114
3.2	Methods . . . . .	117
3.2.1	Test set construction and simulation details . . . . .	117
3.2.2	Descriptor selection . . . . .	121
3.2.3	Training and uncertainty quantification of ML models . . . . .	127
3.3	Results and discussion . . . . .	130
3.3.1	Overview of data set spin-state energetics . . . . .	130
3.3.2	Spin-state splittings from an ANN . . . . .	133
3.3.3	Predicting exchange sensitivity with an ANN . . . . .	138
3.3.4	Predicting metal-ligand bond length with an ANN . . . . .	140
3.3.5	Expanding the test set with experimental complexes . . . . .	143
3.3.6	Extrapolating GGA functionals to hybrids with an ANN . . . . .	148
3.4	Conclusions . . . . .	149
<b>4</b>	<b>Mapping transition metal complex space for machine learning</b>	<b>151</b>
4.1	Introduction . . . . .	152
4.2	Approach to feature construction and selection . . . . .	154
4.2.1	Autocorrelation functions as descriptors . . . . .	154
4.2.2	Revised autocorrelations for transition metal complexes . . . . .	159
4.2.3	Feature selection methods . . . . .	163
4.3	Computational details . . . . .	167
4.3.1	Organization of data sets . . . . .	167
4.3.2	First-principles simulation methodology . . . . .	170
4.4	Results and discussion . . . . .	171
4.4.1	Spin splitting energy . . . . .	171



4.4.2	Descriptor transferability to bond length prediction . . . . .	180
4.4.3	Descriptor transferability to redox data . . . . .	183
4.4.4	Overall comparison of best descriptor subsets . . . . .	187
4.5	Conclusions . . . . .	192
<b>5</b>	<b>Design of spin crossover materials with ANNs and DFT</b>	<b>195</b>
5.1	Introduction . . . . .	196
5.2	Design methodology . . . . .	199
5.3	Results and discussion . . . . .	204
5.4	Conclusions . . . . .	208
5.5	Computational details . . . . .	209
<b>6</b>	<b>Uncertainty and extrapolation of ANNs for chemical discovery</b>	<b>211</b>
6.1	Introduction . . . . .	212
6.2	Results and discussion . . . . .	216
6.3	Conclusions . . . . .	228
6.4	Computational details . . . . .	228
6.5	Addendum: improving generalization with smart data acquisition . .	229
<b>7</b>	<b>Multiobjective, multifidelity redox couple design</b>	<b>235</b>
7.1	Introduction . . . . .	236
7.2	Approach and methods . . . . .	238
7.2.1	First-principles calculations . . . . .	238
7.2.2	Design space . . . . .	240
7.2.3	Machine learning . . . . .	242
7.2.4	Multiobjective design framework . . . . .	244
7.3	Results and discussion . . . . .	248

7.3.1	Model selection and extrapolation from existing data . . . . .	248
7.3.2	Active learning process . . . . .	251
7.3.3	Analysis of leads . . . . .	255
7.4	Conclusions . . . . .	262
7.5	Computational details . . . . .	263
7.5.1	First-principles methods . . . . .	263
7.5.2	Machine learning methods . . . . .	265
<b>8</b>	<b>Concluding remarks</b>	<b>267</b>
8.1	Conclusions . . . . .	267
8.2	Future directions . . . . .	271
	<b>Bibliography</b>	<b>273</b>
<b>A</b>	<b>Surrogate models for transition metal complexes</b>	<b>331</b>
<b>B</b>	<b>Mapping transition metal complex space for machine learning</b>	<b>389</b>
<b>C</b>	<b>Design of spin crossover materials with ANNs and DFT</b>	<b>423</b>
<b>D</b>	<b>Uncertainty and extrapolation of ANNs for chemical discovery</b>	<b>441</b>
<b>E</b>	<b>Multiobjective, multifidelity redox couple design</b>	<b>475</b>

# List of Figures

1-1	Examples of transition metal complexes . . . . .	27
2-1	Nominal cost and accuracy for quantum chemical methods . . . . .	36
2-2	Illustration of spin splitting for an octahedral complex . . . . .	41
2-3	Challenges for inorganic chemistry . . . . .	43
2-4	Effect of exact exchange variation on spin splitting energy . . . . .	45
2-5	Overview of supervised learning . . . . .	48
2-6	Illustration of over-fitting . . . . .	54
2-7	Diagram of CV process . . . . .	56
2-8	Basic overview of the featurization process . . . . .	59
2-9	Comparison of some common featurization strategies . . . . .	62
2-10	Example of graph representation of the oxalate ion . . . . .	64
2-11	Comparison of linear and Gaussian kernels . . . . .	78
2-12	Hyperparameter estimation for KRR by CV . . . . .	80
2-13	Gaussian process regression example . . . . .	85
2-14	Diagram of single neural network node . . . . .	88
2-15	Activation functions for neural networks . . . . .	89
2-16	Overview of a multilayer perceptron model . . . . .	91

2-17	Illustration of a single 2D convolutional filter . . . . .	95
2-18	Comparison between predictive and generative models . . . . .	98
2-19	Schematic of surrogate-assisted virtual screening workflow . . . . .	105
3-1	Ligands used to spin splitting data set . . . . .	118
3-2	Schematic of ANN for transition metal chemistry . . . . .	119
3-3	Summary of LASSO-selected variables for spin state prediction . . . . .	124
3-4	Decision tree for ground spin state of transition metal complexes . . . . .	132
3-5	DFT and ANN predictions of $\Delta E_{H-L}$ for $Fe^{3+}$ complexes . . . . .	133
3-6	Test and training errors for spin splitting prediction by ANN . . . . .	134
3-7	Test and training errors for $\Delta E_{H-L}$ prediction by ANN . . . . .	137
3-8	Test and training errors for $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$ prediction by ANN . . . . .	140
3-9	Test and training errors for spin splitting prediction by ANN . . . . .	141
3-10	Test and training errors for bond length by ANN . . . . .	143
3-11	Representative CSD test set molecules . . . . .	144
3-12	Spin splitting energy predictions on CSD structures . . . . .	145
3-13	Comparison of CSD crystal bond distances and spin states . . . . .	147
4-1	Autocorrelation depth and prediction accuracy . . . . .	157
4-2	Learning rate for different descriptor sets . . . . .	159
4-3	Illustration of RACs . . . . .	161
4-4	Schematic of feature selection methods . . . . .	165
4-5	Data sets used for feature selection . . . . .	168
4-6	PCA for different representations of inorganic complexes . . . . .	174
4-7	Data sets used for feature selection . . . . .	177
4-8	Descriptor locality analysis for spin splitting energy . . . . .	179
4-9	Descriptor locality analysis for bond lengths . . . . .	182

4-10	Descriptor locality analysis for redox potential . . . . .	187
4-11	Relative locality and character of different feature sets . . . . .	189
4-12	Positioning of homoleptic and heteroleptic complexes in PCA space .	191
5-1	Feature space distances and representative complexes . . . . .	198
5-2	Comparison of evolutionary algorithm control strategies . . . . .	203
5-3	Assessment of ANN-assisted SCO design strategy . . . . .	204
5-4	Validation of ANN-designed SCO leads . . . . .	206
6-1	Schematic of ANN uncertainty quantification methods . . . . .	213
6-2	PCA comparison of $\Delta E_{H-L}$ predictions to CSD test set . . . . .	218
6-3	Relationship between $\Delta E_{H-L}$ errors and uncertainty metrics . . . . .	222
6-4	Atomization energy errors and three UQ metrics . . . . .	225
6-5	Latent distance error control on organic and inorganic data . . . . .	226
6-6	PCA of OHLDB complexes and ANN prediction errors . . . . .	230
6-7	ANN error reduction with OHLDB complexes . . . . .	233
7-1	Design space of potential RFB candidates . . . . .	241
7-2	Illustration of 2D active learning workflow . . . . .	247
7-3	Relative extrapolation errors for initial $\Delta G_{solv}$ ML models . . . . .	250
7-4	Acquisition function values for 2.8M complex space from ANN model	252
7-5	Active learning lookahead errors for $\Delta G_{solv}$ prediction . . . . .	254
7-6	Evolution of DFT Pareto frontier . . . . .	256
7-7	Composition of final Pareto set . . . . .	259
7-8	Boxplot of multiobjective design results . . . . .	261
A-1	Variable selection plots . . . . .	353
A-2	Binary ground state classification tree for heteroleptic compounds . .	354

A-3	Splitting energy prediction: Mn . . . . .	355
A-4	Splitting energy prediction: Co . . . . .	356
A-5	Splitting energy prediction: Cr . . . . .	357
A-6	Splitting energy prediction: Fe(II) and Ni(II) . . . . .	358
A-7	Splitting energy prediction: comparison of uncertainty . . . . .	359
A-8	Test set error histogram for different machine learning models . . . . .	360
A-9	Scatter plot of HFX sensitivity . . . . .	362
A-10	HF slope prediction: error boxplot . . . . .	362
A-11	HF slope prediction: Co . . . . .	363
A-12	HF slope prediction: Cr . . . . .	364
A-13	HF slope prediction: Fe(II) and Ni(II) . . . . .	365
A-14	HF slope prediction: Mn . . . . .	366
A-15	HF slope prediction: comparison of uncertainty . . . . .	367
A-16	Bond distance prediction: LS error boxplot . . . . .	367
A-17	Bond distance prediction: comparison of LS uncertainty . . . . .	368
A-18	Bond distance prediction: LS Co . . . . .	369
A-19	Bond distance prediction: LS Cr . . . . .	370
A-20	Bond distance prediction: LS Fe(II) and Ni(II) . . . . .	371
A-21	Bond distance prediction: LS Mn . . . . .	372
A-22	Bond distance prediction: HS error boxplot . . . . .	373
A-23	Bond distance prediction: comparison of HF uncertainty . . . . .	373
A-24	Bond distance prediction: HS Co . . . . .	374
A-25	Bond distance prediction: HS Cr . . . . .	375
A-26	Bond distance prediction: HS Fe(II), Fe(III) and Ni(II) . . . . .	376
A-27	Bond distance prediction: HS Mn . . . . .	377
A-28	Metal-ligand gradient bond projection . . . . .	379

A-29 CSD errors and Tanimoto/FP2 dissimilarity . . . . .	381
A-30 CSD errors and Euclidean and Pearson distances . . . . .	382
A-31 CSD LS geometry errors and dissimilarity comparison . . . . .	385
A-32 CSD HS geometry errors and dissimilarity comparison . . . . .	386
A-33 Measured CSD bond distances and DFT predicted spin states . . . . .	387
A-34 Comparison of HFX interpolation and direct prediction . . . . .	387
B-1 Random forest tree number convergence for spin-splitting . . . . .	392
B-2 Random forest tree number convergence for bond lengths . . . . .	393
B-3 Random forest tree number convergence for ionization prediction . . . . .	393
B-4 Random forest tree number convergence for redox prediction . . . . .	394
B-5 Learning rate for splitting energy prediction using RAC-155 . . . . .	399
B-6 PCA variance die-off with RAC-155 principal components . . . . .	400
B-7 PCA variance die-off with CM-ES principal components . . . . .	400
B-8 Elastic net parameter response for spin splitting energy . . . . .	401
B-9 Univariate filter scores for spin splitting energy . . . . .	403
B-10 RFE mean-CV error with number of features retained . . . . .	404
B-11 Random forest variable importance scores . . . . .	405
B-12 KRR error distributions for spin splitting energy predictions . . . . .	407
B-13 Elastic net parameter response for bond lengths . . . . .	410
B-14 Random forest variable importance scores for bond lengths . . . . .	411
B-15 Elastic net parameter response for ionization potential . . . . .	415
B-16 Random forest variable importance scores for ionization potential . . . . .	415
B-17 Locality and performance of feature sets on ionization potential . . . . .	416
B-18 Error comparison for RAC-155 on ionization/redox potential . . . . .	417
B-19 Outlier complex for redox prediction . . . . .	417

B-20	Elastic net parameter response for redox potential . . . . .	418
B-21	Random forest variable importance scores for redox prediction . . . . .	419
B-22	Triazolyl-pyridine ligands from the redox data set . . . . .	420
B-23	Schematic of feature set locality and character including LASSO-28 . . . . .	421
C-1	Structures of design space ligands . . . . .	424
C-2	Response of fitness function to changing splitting energy . . . . .	426
C-3	Number of unique complexes sampled per repeat using ANN . . . . .	430
C-4	GA trajectories on a 2D histogram of compound space . . . . .	431
C-5	Number of unique complexes sampled across all repeats . . . . .	432
C-6	Histogram of complexes explored with varied control mode . . . . .	432
C-7	Error vs. distance to closest training point . . . . .	435
C-8	Wall time for DFT single point energies . . . . .	435
C-9	Wall time for DFT geometry optimizations . . . . .	436
C-10	Wall time for DFT SP GA runs . . . . .	436
C-11	Ligand incidence in ANN GA leads with 15% exchange . . . . .	438
D-1	Ligands for training inorganic complex ANN . . . . .	445
D-2	CSD structures: ACEYOW-CERZII . . . . .	446
D-3	CSD structures: COMTED02-DEDKOO . . . . .	447
D-4	CSD structures: DOQRAC-EKOTUV . . . . .	448
D-5	CSD structures: ETEKIX-FARHOV . . . . .	449
D-6	Variance decay of PCA for inorganic data . . . . .	450
D-7	Comparison of CSD predictions from single and ensemble models . . . . .	451
D-8	Distribution of CSD prediction errors . . . . .	452
D-9	Comparison of test/train distributions for two prediction tasks . . . . .	453
D-10	Distribution of CSD distances to training data . . . . .	454



D-11	Distance thresholds, numbers of neighbors and retained errors . . . . .	455
D-12	Latent distance threshold performance with number of neighbors . . . . .	456
D-13	Correlation between UQ metrics and model errors on CSD data . . . . .	457
D-14	Maximum retained errors for CSD using different UQ metrics . . . . .	458
D-15	Average retained errors for CSD using different UQ metrics . . . . .	459
D-16	Error rate for CSD prediction using different UQ metrics . . . . .	460
D-17	Neural network architecture for QM9 predictions . . . . .	461
D-18	Distribution of QM9 atomization prediction errors . . . . .	464
D-19	Correlation between UQ metrics and model errors on QM9 data . . . . .	465
D-20	Average retained errors for QM9 using different UQ metrics . . . . .	466
D-21	Quantitative UQ from ensemble or latent distance and model error . . . . .	468
D-22	Retained errors for combined uncertainty on CSD data . . . . .	470
D-23	Calibrated feature space distance vs. model error on CSD set . . . . .	471
D-24	Analysis of PCA and UMAP of the latent space . . . . .	472
D-25	Latent distance comparison for MNIST and Fashion–MNIST . . . . .	474
E-1	Solvent and thermodynamic corrections for redox potentials . . . . .	476
E-2	Structures of core heterocycles and C3 modifications . . . . .	479
E-3	Ring functionalization approaches . . . . .	480
E-4	Distribution of metals and connecting atoms in ‘hot start’ data . . . . .	481
E-5	Size and denticity comparison of ‘hot start’ data with design space . . . . .	481
E-6	Initial cluster size and degree of isolation . . . . .	483
E-7	Silhouette analysis of clusters . . . . .	484
E-8	Schematic of active learning workflow . . . . .	485
E-9	Uncertainty metric comparison for uniform test data . . . . .	487
E-10	Relative extrapolation errors for initial logP ML models . . . . .	488

E-11 Comparison of hot start and design space property distributions . . .	488
E-12 Comparison of UQ metrics for initial medoids from hot start data . .	490
E-13 Evolution of E[I] and P[I] distributions during active learning . . . .	491
E-14 Lookahead and random-test errors for logP . . . . .	493
E-15 Distribution of metals in simulated complexes . . . . .	495
E-16 Full bidentate ligand structure selected in two complexes . . . . .	495
E-17 Most commonly-sampled unfunctionalized bidentate ligands . . . . .	496
E-18 Histogram of base ligand occurrence in simulated . . . . .	497
E-19 Redox and logP values of simulated complexes by base heterocycle . .	498
E-20 Distributions of redox/logP for randomly-sampled complexes . . . . .	501
E-21 Q-Q plots of redox/logP values for randomly-sampled complexes . . .	502
E-22 Neural network architectures . . . . .	504
E-23 UQ parameters for multitask ANN . . . . .	505
E-24 GP hyperparameter CV response surfaces . . . . .	506

# List of Tables

2.1	Some commonly used kernels their hyperparameters . . . . .	79
3.1	Discrete/continuous feature combinations . . . . .	123
3.2	Optimal heuristic feature set MCDL-25 . . . . .	127
3.3	Train, test and CSD errors for predicting $\Delta E_{\text{H-L}}$ . . . . .	138
3.4	Test set prediction errors for predicting $\frac{\partial \Delta E_{\text{H-L}}}{\partial \alpha_{\text{HF}}}$ by metal . . . . .	139
4.1	Comparison of feature sets for inorganic chemistry . . . . .	172
4.2	Peformance of RAC-subsets for spin splitting energy . . . . .	175
4.3	Peformance of RAC-subsets for bond lengths . . . . .	181
4.4	Peformance of RAC-subsets for redox potential . . . . .	185
A.1	Ligand properties . . . . .	334
A.2	Core homoleptic ligands . . . . .	335
A.3	CSD structure references . . . . .	336
A.4	Relaxed tolerances for some CSD cases . . . . .	337
A.5	Excluded structures: spin contamination . . . . .	337
A.6	Excluded structures: geometric brekaup . . . . .	337
A.7	Variable scaling constants . . . . .	338

A.8	Variable Selection for splitting energy: set a . . . . .	339
A.9	Variable Selection for splitting energy: set b . . . . .	340
A.10	Variable Selection for splitting energy: set c . . . . .	341
A.11	Variable Selection for splitting energy: set d . . . . .	342
A.12	Variable Selection for splitting energy: set e . . . . .	343
A.13	Variable selection for splitting energy: set f . . . . .	344
A.14	Variable selection for splitting energy: set g . . . . .	345
A.15	Variable selection for HF slope: set a . . . . .	346
A.16	Variable selection for HF slope: set b . . . . .	347
A.17	Variable selection for HF slope: set c . . . . .	348
A.18	Variable selection for HF slope: set d . . . . .	349
A.19	Variable selection for HF slope: set e . . . . .	350
A.20	Variable selection for HF slope: set f . . . . .	351
A.21	Variable selection for HF slope: set g . . . . .	352
A.22	Hyperparameters for different KRR and SVR models . . . . .	354
A.23	RMS prediction errors for splitting energy . . . . .	361
A.24	Average HFX sensitivity by metal and ligand connection atom . . . . .	361
A.25	Error breakdown for bond distance predictions . . . . .	378
A.26	molSimplify gradients for ANN assisted structure design . . . . .	378
A.27	CSD split energy prediction . . . . .	380
A.28	LS CSD geometry prediction . . . . .	383
A.29	HS CSD geometry prediction . . . . .	384
B.1	Learning rates by descriptor for QM9 atomization energies . . . . .	391
B.2	Learning rates for <i>3d</i> -AC on QM9 dipole moments . . . . .	391
B.3	Hyperparameters for spin splitting energy prediction . . . . .	394

B.4	Hyperparameters for bond length prediction . . . . .	395
B.5	Hyperparameters for ionization potential prediction . . . . .	395
B.6	Hyperparameters for redox prediction . . . . .	395
B.7	Spin splitting data set ligands from previous work . . . . .	396
B.8	Redox data set ligands from previous work . . . . .	396
B.9	Redox/ionization spin multiplicities . . . . .	397
B.10	Features included in the RAC-155 descriptor set. . . . .	398
B.11	Features included in the LASSO-28 descriptor set. . . . .	401
B.12	Features included in the UV-86 descriptor set. . . . .	402
B.13	Features included in the RFE-43 descriptor set. . . . .	403
B.14	Features included in the randF-41 descriptor set. . . . .	405
B.15	Features included in the randF-26 descriptor set. . . . .	406
B.16	KRR error metrics with random forest cutoffs on spin splitting . . . . .	406
B.17	Features in the common RAC-12 set . . . . .	408
B.18	Features in the proximal-only PROX-23 set. . . . .	408
B.19	Features in the bond-length-selected LASSO-83B descriptor set. . . . .	409
B.20	Features in randF-48B/randF-49B . . . . .	410
B.21	Neighborhoods for Fe(III)(pisc) <sub>6</sub> in different feature sets . . . . .	412
B.22	Features in LASSO-19I (ionization) . . . . .	413
B.23	Features in randF-28I (ionization) . . . . .	414
B.24	KRR test and train errors for ionization potentials . . . . .	414
B.25	Features in LASSO-19G (redox) . . . . .	418
B.26	Features in randF-38G (redox) . . . . .	419
B.27	Comparison of spin-state choice in RAC-155 redox predictions . . . . .	420
C.1	Design space of ligands. . . . .	425

C.2	Allowed ligand combinations. . . . .	426
C.3	Spin multiplicities used in this work . . . . .	426
C.4	Summary of sampled complexes with varied control mode . . . . .	431
C.5	Comparison of ANN predictions to DFT results . . . . .	434
C.6	Evaluation of DFT-SP-GA hits with full optimization . . . . .	437
C.7	Basis set dependence . . . . .	439
D.1	Ligands used in inorganic training data . . . . .	443
D.2	Overall inorganic spin splitting ANN performance . . . . .	450
D.3	Points used to calibrate latent-distance uncertainty model . . . . .	459
D.4	Parameters for latent-distance dependent CSD uncertainty model . . . . .	460
D.5	Hyperparameters and topology for organic atomization energy ANN . . . . .	462
D.6	Improvement in predictions with residual layers . . . . .	462
D.7	Repetition test for QM9 benchmark . . . . .	463
D.8	Overall QM9 atomization energy ANN performance . . . . .	463
D.9	Parameters for latent-distance dependent QM9 uncertainty model . . . . .	467
D.10	Active learning CSD experiment . . . . .	468
D.11	Hyperparameters and topology for inorganic spin splitting ANN . . . . .	469
D.12	Error/distances relationship and dimensionality reduction . . . . .	471
D.13	Hyperparameters and topology for image classification CNNs . . . . .	473
E.1	Spin multiplicities selected for ionization processes . . . . .	476
E.2	Original sources of ‘hot start’ data . . . . .	480
E.3	Design space and cluster comparison . . . . .	484
E.4	Error metrics of initial models . . . . .	486
E.5	Correlation of absolute test errors for initial models . . . . .	486
E.6	Correlation of absolute out-of-distribution errors for initial models . . . . .	489

E.7	Summary of ‘hot start’ and active learning data set sizes . . . . .	491
E.8	Redox potential lookahead error metrics . . . . .	492
E.9	logP lookahead error metrics . . . . .	494
E.10	Most commonly-selected functional groups on ligands . . . . .	499
E.11	Properties of complexes at the Pareto front . . . . .	500
E.12	DFT results and structures for final Pareto complexes . . . . .	501
E.13	Revised geometry check tolerances . . . . .	502
E.14	ANN hyperparameters . . . . .	503
E.15	GP hyperparameters . . . . .	505





# Chapter 1

## Introduction

### 1.1 Rational design of transition metal complexes

The rational design of functional materials at a molecular level is a central objective in chemistry and materials science<sup>1-3</sup>. Serious challenges in sustainable energy<sup>4,5</sup>, chemical process design<sup>6</sup>, molecular electronics<sup>7</sup> and drug discovery<sup>3</sup> require the design of molecular systems with exotic and tightly controlled properties that are fundamentally quantum-mechanical in nature<sup>8</sup>, for example excitation energies for photochemical systems<sup>9</sup> or reaction barriers in catalysis<sup>10,11</sup>.

The vast nature of chemical space makes computational tools essential for screening for novel molecules and materials with targeted properties<sup>12</sup>, from the design of solvents<sup>13</sup>, light emitting diodes (LEDs)<sup>14</sup>, dye sensitizers<sup>15-17</sup>, polymers<sup>18</sup>, redox couples<sup>19</sup>, catalysis<sup>20</sup>, and nonlinear optical materials<sup>21</sup>. Many commercial drugs, dating back to the 1980s have at least benefited from computational design<sup>3</sup>. The potential for designing novel molecules with targeted properties with computational chemistry is enormous.

Despite increases in computational power to simulate new compounds, the diversity of chemical space and cost of these calculations means that the fraction of possible designs that can be probed directly by these methods is vanishingly small<sup>22</sup>. To address these difficulties, chemists have turned to data-driven *surrogate models* to help interpolate between, and extrapolate from, computational and experimental observations. While this idea is not new - concern about hype over neural networks in chemistry dates back thirty years<sup>23</sup> - there has been a contemporary boon in the development of data-driven models of unprecedented complexity and accuracy<sup>24-29</sup>, exploiting recent developments in the field of machine learning<sup>30</sup> and the availability of ever-larger datasets<sup>31-33</sup>. These models are increasingly being utilized to supplement virtual screening and identify exceptional novel materials *in silico* which can be realized synthetically, including light emitting materials<sup>34</sup>, metallic glasses<sup>35</sup>, singlet fission complexes<sup>36</sup> and magnetic materials<sup>37</sup>. The future of molecular design and discovery is expected increasingly to leverage both physics-based and data-driven methods to guide experiments<sup>38</sup>.

This thesis concerns the combination of first-principles high throughput virtual screening (HTVS) and data-driven modeling to a challenging design problem: open-shell *transition metal complexes*. These transition metal complexes, consisting of a set of ligands coordinating to a one or more of the elements in the periodic table that have partially filled *d* shells (Figure 1-1), are of great practical importance. Transition metal complexes are ubiquitous in homogeneous/biomimetic<sup>39-43</sup>, where they show promise for achieving some of the most challenging and industrially-relevant reactions including selective partial oxidation of hydrocarbons<sup>44-46</sup> and reduction of CO<sub>2</sub><sup>47,48</sup>.

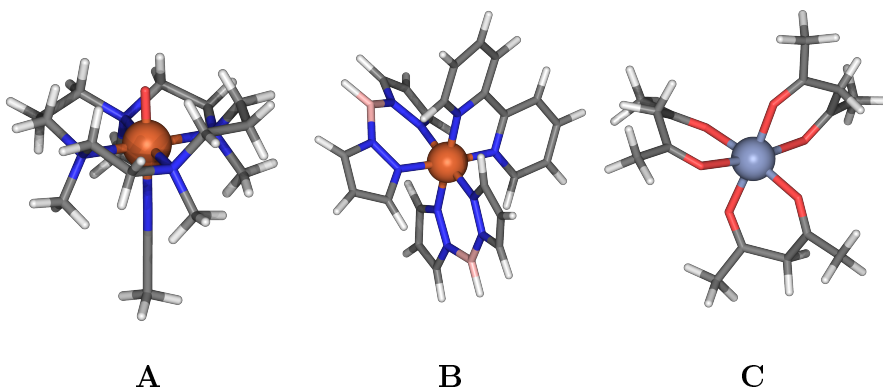


Figure 1-1: Some noteworthy examples of octahedral transition metal complexes: **A**:  $\text{Fe}[(\text{O})(\text{TMC})(\text{NCCH}_3)]$  (TMC = 1,4,8,11-tetramethyl-1,4,8,11-tetraazacyclotetradecane), a reactive intermediate step in catalytic cycle possessing a iron(IV)=O moiety for olefin epoxidation<sup>46</sup>; **B**:  $\text{Fe}[(\text{H}_2\text{B}(\text{pz})_2)_2(\text{bipy})]$  (bipy = 2,2'-bipyridine,  $\text{H}_2\text{B}(\text{pz})_2$  = dihydrobis(1-pyrazolyl)borate), a spin crossover complex<sup>49</sup>; **C**:  $\text{Cr}[(\text{acac})_3]$  (acac = acetylacetonate), a redox couple for non-aqueous redox flow batteries<sup>50</sup>. The metal centers are shown as large spheres (orange = Fe, blue = Cr), while other atoms are drawn as sticks (gray = C, red = O, blue = N, pink = B, white = H).

The open-shell 3d electron structure of mid-row transition metals allows them to access different spin states<sup>51</sup>, and complexes with finely-tuned spin state orderings have potential applications as sensors<sup>52,53</sup>, electronics<sup>7</sup>, thermochromatic materials<sup>54</sup> and dye sensitizers<sup>16,55,56</sup>. In these applications, the energy difference between the high- and low-spin states plays a key role in determining material properties, for example determining the frequency of light that can be absorbed or emitted. Spin state ordering is essential for understanding reactions catalyzed by open-shell complexes<sup>57,58</sup> as the spin state can dramatically alter the free energy landscape of the reaction. Spin crossover complexes (SCOs), transition metal complexes that exhibit spin bistability<sup>52</sup>, are of particular interest for rational design<sup>49,53,59-61</sup>. These mate-

rials possess high- and low-spin states are close enough in energy that entropic effects can alter the spin state ordering in response to changing temperature, mechanical strain<sup>62</sup> or upon absorbing light, creating functional, single-molecule devices<sup>63</sup>.

Another important application of transition metal complexes is in redox flow batteries (RFBs), which are a promising grid-level energy storage technology<sup>64,65</sup> that decouples power delivery and cell capacity by holding the redox active species in a liquid solution which is stored separately and pumped to a cell when required. RFBs are appealing for high capacity and low cost, making them ideal for fixed storage energy applications in support of renewable energy<sup>66,67</sup>. Transition metal complexes have been used as redox couples in RFBs for decades<sup>68</sup> due their redox stability and low membrane permeation propensity<sup>69</sup>. By tuning their ligand field<sup>70,71</sup>, important characteristics, such as cell potential and solubility, can be manipulated to create more efficient non-aqueous flow batteries<sup>50</sup>.

In all cases, the promise of molecular transition metal chemistry derives from the ability to manipulate these electronic properties precisely through control of the ligand field. However, since the space of conceivable ligands is practically infinite<sup>72</sup>, and matching ligand fields and metal centers is a fundamentally combinatorial problem, the challenge of designing novel transition metal complexes is considerable. This is compounded by the general lack of understanding and intuition compared to organic chemistry<sup>51,58</sup>, the complicated electronic structure of these materials<sup>73,74</sup>, and the lack of large databases<sup>12,31,75</sup>, open source tools<sup>76,77</sup> and well-established machine learning (ML) approaches<sup>24,27,78-81</sup> that support the design of organic and bulk materials. Evidently, there is a need to develop new methods that address these challenges and facilitate the *de novo* design of these systems. This thesis will address these challenges.

## 1.2 Thesis overview

This thesis develops a framework for first-principles design of transition metal complexes with targeted properties that integrates data-driven surrogate modeling to accelerate, guide and understand density functional theory (DFT) calculations. The thesis is structured as follows:

- Chapter 2 reviews the theory behind first-principles screening, how it can be bypassed by machine learning and how these methods have been combined to solve to chemical design problems in other application domains such as small organic molecules and bulk crystalline systems.
- Chapter 3 presents the first data-driven surrogate models for the spin-dependent quantum chemical properties of open-shell transition metal complexes, uniquely addressing the idiosyncratic challenges of this region of chemical space. Heuristic descriptors based on chemical knowledge are used to represent complexes and demonstrated to have good correlation with quantum mechanical (QM) outcomes. The capacity of these models to extrapolate to new data is probed with a test of very dissimilar complexes and this provides important insights that lay the groundwork for the handling of model uncertainty in later chapters.
- Chapter 4 develops novel representations for transition metal complexes and proposes a new family of descriptors based on molecular graphs that can encode flexibility metal-local and global atomic information. These representations lead to predictive models with increased accuracy, but simultaneously retain interpretability. Feature selection is performed to identify the most important features for different outcomes, providing to physical insight into the first-principles data used to parameterize the models. In particular, by analysis

of the features correlated with different QM outcomes, specific length-scales of atomic influence on the metal center can be detected, suggesting new orthogonal design strategies.

- Chapter 5 applies the developed surrogate models to discover new SCOs posed on a space of a few thousand transition metal complexes. An artificial neural network (ANN) surrogate model is equipped with a geometric measure of extrapolative uncertainty and combined with a newly-developed, diversity-seeking, evolutionary design strategy to identify lead candidates to simulate with DFT. A set of these leads are simulated using DFT and the majority are verified as having the target DFT-level spin state properties, yielding DFT-level SCOs in a fraction of the time required for conventional HTVS. The impact of the proposed uncertainty-constrained optimization on the optimization process is explored and shown to provide increased diversity relative to the naïve approach.
- Chapter 6 examines uncertainty quantification for ANN models for chemical discovery and develops a novel method for assigning confidence based on geometric extrapolation using the learned nonlinear transformations in the ANN latent space. This metric provides as-good-or-better performance compared to existing ensembles methods without additional overheads inherent in committee-based uncertainty. A probabilistic error model based on this metric is introduced and shown to provide superior quantitative error estimation on both transition metal complexes and traditional organic benchmarks, as well as allowing for efficient active learning.
- Chapter 7 exploits the developed machine learning framework to tackle a mul-

tiobjective design problem in the space of RFBs, identifying transition metal complexes with suitable redox and solubility properties from a space of nearly 3 million candidates. A combinatorial strategy to construct potential transition metal complexes is devised and used to generate a diverse and densely-sampled chemical space. A two-dimensional Bayesian optimization strategy based on expected improvement (EI) is implemented to balance exploration and exploitation in this design space using a multitask ANN surrogate model, coupled with an uncertainty model on based exploration in the learned latent space. The algorithm is able to identify and iteratively refine a Pareto frontier of candidate complexes using a few hundred DFT simulations that are substantially enriched relative to random sampling.

- Chapter 8 provides conclusions on the research presented in the previous chapters and discusses how the developed methods can be applied to other important areas such as catalysis.





# Chapter 2

## Background

### 2.1 First-principles virtual screening

#### 2.1.1 Overview

Experimental characterization of materials is generally expensive, time consuming and requires expert knowledge. While advances in automated, high throughput experimental setups<sup>82-84</sup>, possibly assisted by data-driven methods<sup>85,86</sup>, show great promise for testing large numbers of candidates<sup>5</sup>, for example drugs<sup>3,87</sup>, catalysts<sup>88</sup> or solid materials<sup>89</sup>; these approaches are inherently restricted to certain chemical regimes due to the range of reagents and synthesis conditions that can be automated. The practically infinite variety of possible candidates,  $\mathcal{O}(10^{60})$  organic molecules alone<sup>90</sup>, guarantee that we will only ever access a tiny fraction of chemical space synthetically. In light of these challenges, high throughput virtual screening (HTVS) of materials with computers has emerged as a critical paradigm<sup>5,8,12,91</sup>, complementing experimental investigations while being far less limited with respect to the types and numbers of systems that can be studied.

Techniques of computational chemistry have been instrumental in developing understanding of complex processes occurring on length- and time-scales that are difficult to access, for example catalysis<sup>92</sup>, electrochemistry<sup>93</sup>, proteins and enzymes<sup>94,95</sup>, cellular processes<sup>96</sup>, batteries<sup>97</sup> and solar devices<sup>15,98</sup>. However, the ability to discover new materials ‘from scratch’ using computational techniques is a fairly recent development<sup>99–101</sup>. Increases in computational resources along with more efficient algorithms and methods, including GPU-accelerated computing<sup>102–106</sup>, have facilitated the use accurate quantum methods for screening of large numbers of materials on a scale that was previously unobtainable<sup>10,107</sup>, and improved understanding of the accuracy of these methods relative to experimental measurements has allowed even complex phenomena such as catalysis to be screened with useful fidelity<sup>6</sup>.

Large-scale HTVS efforts have so far been mainly applied to bulk crystalline inorganic materials, with prominent examples being the Materials Project<sup>108</sup>, which includes computational properties of over 120k inorganic crystals and 500k nanoporous materials, the Open Quantum Materials Database (OQMD)<sup>12</sup> which includes 300k quantum simulations of periodic systems, the AFLOW library<sup>75</sup> consisting of 3M crystal structures with 500k calculated properties, or organic chemistry as in the Harvard Clean energy project<sup>98,109</sup>, consisting of density functional theory (DFT) calculations for over 2 million molecules for solar cell applications. The exhaustive screening of thousands of candidates at a quantum mechanical level has led to the discovery of real materials with exotic electronic properties, for example singlet fission materials<sup>36</sup> (with applications in photovoltaic materials) or metal carbides with exceptional hardness<sup>110</sup>. Importantly, we draw a distinction between these exhaustive simulations of a large number of candidates (‘pure’ HTVS) and the ideas of rational design discussed in Section 2.3, where complexes are selectively simulated. One important question in any simulation of a chemical system is the choice of

method, which relates a chosen chemical structure (a molecule for example) to a property of interest (energy, redox potential, affinity for given protein target etc). Different methods can be classified by the degree to which they approximate fundamental physics with empirical parameters. On the one end of the spectrum, purely-empirical models, such as quantitative structure–activity relationships (QSARs) or quantitative structure–property relationships (QSPRs)<sup>111,112</sup>, group-additivity<sup>113</sup>, fragmentation methods<sup>114,115</sup> or the myriad of machine learned models that have been developed (*vide infra*), make no specific recourse to physics and rely on being calibrated from or fit to existing data, which makes them cheap to compute for a large number of inputs but fundamentally limits their transferability to systems unlike their training data.

While even the simplest quantum method remain more expensive than classical Newtonian methods<sup>116,117</sup> (such as force fields<sup>118–120</sup>), they offer an unparalleled ability to generalize to new systems, making these *ab initio* methods incredibly powerful for HTVS. In addition, many important applications, such as catalyst discovery<sup>6,101</sup>, discovering new reactions<sup>100</sup>, spintronics<sup>7</sup> and photovoltaic systems<sup>16</sup> involve the electrons explicitly and thus require quantum treatment.

According to (non-relativistic) quantum mechanics, all of the observable properties of a system can be obtained from its wavefunction<sup>121</sup>,  $\Psi$ , which, for the ground state of an atomistic system, in turn obeys the (time-independent) Schrödinger equation<sup>122</sup>:

$$H\Psi = E\Psi$$

Here,  $H$  is the Hamiltonian and  $E$  is the energy of the system in state  $\Psi$ . Methods that approximate solutions to the Schrödinger equation are termed *ab initio* methods, since they make relatively few assumptions and have correspondingly few parame-

ters. Therefore, they are applicable to a wide variety of systems, without the need for case-by-case calibration. These methods typically employ the Born–Oppenheimer approximation<sup>123</sup>, meaning that electrons are treated quantum-mechanically and nuclei are treated classically<sup>121</sup>. However, the solution of the Schrödinger equation for a multi-electron system remains challenging and it must be solved numerically, requiring approximations. The number and nature of approximations determine the accuracy-cost trade off of the method (Figure 2-1).

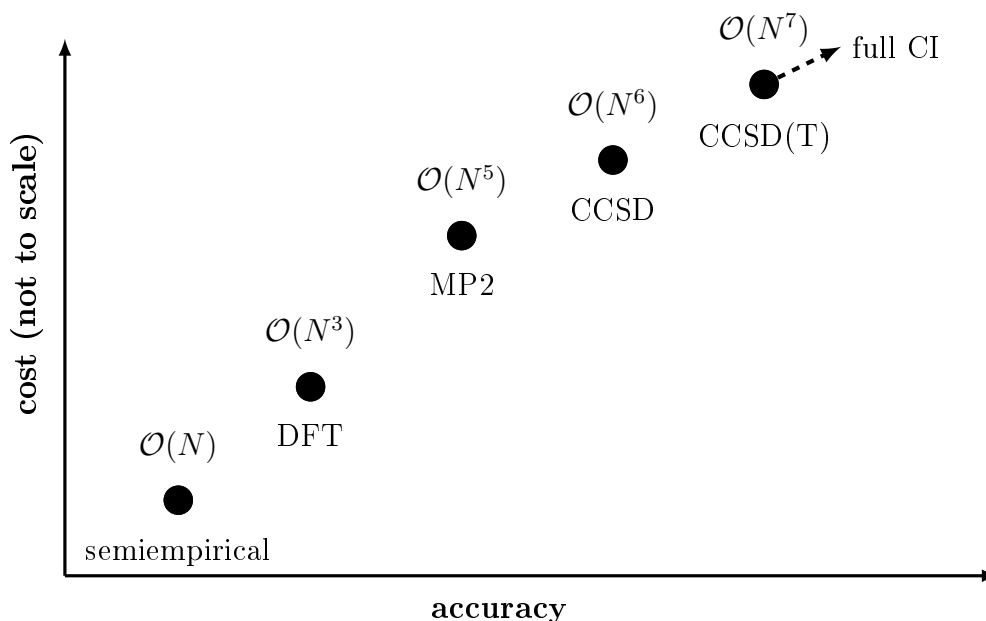


Figure 2-1: Different quantum chemistry methods by formal scaling (in terms of the number of basis functions,  $N$ ) and nominal accuracy. DFT = density functional theory, CCSD = coupled cluster with singles and doubles, CCSD(T) = coupled cluster with singles and doubles with perturbative triples, MP2 = Møller-Plesset perturbation theory, order 2, CI = configuration interaction. Based on Ref.<sup>124</sup>.

These techniques range from semiempirical methods<sup>125–127</sup>, which attempt to use parametric simplifications to the full Hamiltonian<sup>128</sup> (and hence are not truly *ab initio*), through mainstay methods such as DFT, which are affordable enough to

make quantum simulations of hundreds of atoms routine, to methods which address electron correlation explicitly<sup>122</sup>, including coupled-cluster models<sup>129–131</sup>, configuration interaction<sup>132,133</sup> and Möller-Plesset perturbation theory<sup>134,135</sup>. While these methods can match experimental accuracy<sup>136</sup>, the costs are severe – for example, coupled-cluster models\* formally scale as  $\mathcal{O}(N^7)$  (*c.f.*  $\sim \mathcal{O}(N^3)$  for DFT), where  $N$  is the number of basis functions<sup>74</sup>, and to screen databases of thousands of species using these methods would take decades<sup>137</sup>.

### 2.1.2 Density functional theory

DFT is the most popular<sup>91</sup> electronic structure method with a good compromise between computational cost and useful accuracy<sup>91,138</sup>, resulting in the 1998 Nobel prize in chemistry, and generating in excess of 10k publications per year<sup>138</sup>. The fundamental idea behind DFT is to avoid working with the high-dimensional wavefunction and instead work with the electronic density (the amplitude of the wavefunction,  $\rho \propto \int \Psi^* \Psi$ ). Hohenberg and Kohn<sup>139</sup> proved the one-to-one correspondence of the ground state electron density, external potential and wavefunction by the existence<sup>†</sup> of a universal density functional ( $E[\rho]$ ), which means it is possible to evaluate the energy of a system in terms of either the wavefunction or the electron density:

$$E[\Psi] \iff E[\rho]$$

This direct formulation is still in use as ‘orbital-free’ DFT, but is complicated by the inability to express the kinetic energy contribution analytically in terms of electron density, requiring empirical approximations<sup>141–143</sup>. Instead, Kohn and Sham<sup>144</sup>

---

\*with single and double excitations and perturbative treatment of triple excitations

†subject to some weak representability restrictions<sup>140</sup>

reformulated the electron density in terms of the orbitals of a system of fictitious, non-interacting particles which nonetheless share the same total density<sup>145</sup>,  $\Psi_{KS}$ . Since this fictitious system shares the same density as the original system, it is ‘as good’ as the original system:

$$\begin{array}{ccccc}
 \Psi & \implies & \rho & \longleftarrow & \Psi_{KS} \\
 \Downarrow & & \Downarrow & & \Downarrow \\
 E[\Psi] & = & E[\rho] & = & E[\Psi_{KS}]
 \end{array}$$

The existence of such an equivalent system is theoretically guaranteed<sup>139,140</sup> for some external potential,  $E_{xc}$  that accounts for the differences between the interacting real system and this non-interacting, Kohn-Sham system. The difference can be broken down into contributions from correlation and exchange terms, giving rise to the *exchange-correlation* functional. It is this term that is not known analytically, and therefore this is the only approximation necessary to describe a real system in the Kohn-Sham framework<sup>138</sup>. The re-expression of the problem as a non-interacting system of electrons acting under a modified potential is the key to practical Kohn-Sham DFT, resulting in a method that formally scales as  $\mathcal{O}(N^3)$ , and allowing it to benefit from the well-developed computational methods for finding the ground state wavefunctions of such system of non-interacting electrons<sup>140,145</sup>: the same machinery used to solve the *Hartree-Fock* problem via the self-consistent field method<sup>122</sup> for molecular and periodic systems, with the addition of the parameterized exchange-correlation potential added in.

Many different methods for parameterizing this functional have been proposed, and are sometimes grouped by complexity (and nominally accuracy) into a Jacob’s lad-

der<sup>146</sup>. The simplest are termed local density approximations (LDAs), which depend only on the local density,  $\rho$ , and are parameterized to reproduce the correct behavior for the asymptotic case of a uniform electron gas<sup>147-149</sup>. In practice, this is a strong approximation that makes LDAs most applicable to bulk metallic systems where the density changes slowly<sup>146</sup>, but not sufficiently accurate in the presence of strong local variations, such as those characteristic of molecular orbitals<sup>150</sup>. In order to address these issues, generalized gradient approximations (GGAs) such as BLYP<sup>146,151,152</sup> or PBE<sup>146</sup> include an explicit dependence on the gradient of the density<sup>153</sup>,  $\nabla^2\rho$ . This improves upon the description of molecular systems<sup>146</sup>, with up to an order of magnitude reduction of error compared to LDAs for atomization benchmarks<sup>91</sup>. Meta-GGAs are a further development in this line that incorporate information from the second derivative<sup>154</sup> of the density (i.e. the Laplacian,  $\Delta\rho$ ).

The most important class of functionals from the perspective of this thesis are (global) hybrid-GGA functionals, which use a linear combination of the exchange functional from a GGA and the so-called *exact exchange* contribution from the Hartree-Fock method. This is a judicious choice because the Hartree-Fock method tends to over-localize electron density (relative to more accurate methods), while the GGAs (or LDAs) will tend to over-delocalize the density<sup>155</sup>. A hybrid which lies between these two can therefore be substantially more accurate than either method alone. A prototypical hybrid GGA exchange functional might be expressed as<sup>156</sup>:

$$E_x^{\text{hybrid}} = E_x^{\text{GGA}} + \alpha (E_x^{\text{exact}} - E_x^{\text{GGA}})$$

where  $E_x^{\text{exact}}$  is the exchange term based on Hartree-Fock and  $\alpha \in (0, 1)$  is the mixing fraction. The hybrid functional B3LYP<sup>152,157,158</sup> (based on BLYP) uses  $\alpha = 0.20$ , determined empirically, while PBE0<sup>159,160</sup> (based on PBE) uses 25%, derived theoret-

ically. Hybrid-GGAs have shown improvement in accuracy on diverse systems<sup>91,161</sup> including transition metal complexes<sup>162</sup>, leading to their popularity<sup>138</sup> and motivating their use in this thesis. Another type of hybrid functional, called *range-separated* hybrids, applies the exact exchange correction over specific length-scales only, either to improve long-range interactions<sup>163–165</sup> or to reduce the computational overhead of evaluating the exact exchange component in periodic systems<sup>166,167</sup>.

In spite of its popularity, DFT suffers from issues surrounding functional choice<sup>138,168</sup> and simple functionals are known to have systematic deficiencies. Some major difficulties faced by hybrid-GGA level DFT include the underprediction of reaction barriers, band gaps and charge transfer energies<sup>136,155</sup>, as well as poor descriptions of weak, non-covalent interactions<sup>138,168</sup> (e.g. dispersion, which can be addressed with *post-facto* empirical corrections<sup>169–171</sup>).

### 2.1.3 Challenges for transition metal chemistry

Unfortunately, open-shell transition metal complexes are very challenging systems to study with electronic structure methods<sup>74</sup> for several reasons. At the most basic level, transition metals are heavier than main group organic elements and hence have more electrons and require either effective core potentials<sup>172</sup> or simply more basis functions. Transition metal complexes are also typically larger molecules – the largest non-trivial octahedral transition metal complex (for example Fe(III)[(CO)<sub>6</sub>]) has thirteen heavy (non-H) atoms, compared to large databases of DFT calculations on organic molecules<sup>31,32</sup>, which typically contain molecules with < 10 heavy atoms. The partially occupied d orbitals of mid-row transition metal complexes can be populated in multiple, non-equivalent ways, distinguished by spin state. For octahedral complexes such as the Fe(III)[(CO)<sub>6</sub>] example, the fivefold degeneracy of the iron



3d orbitals is broken by differing overlap with the orbitals of the ligands, resulting in splitting the orbitals into  $e_g$  and  $t_{2g}$  sets with an energy gap that depends on the nature of the ligand field (Figure 2-2). A metal with five 3d electrons such as Fe(III) may exist in a high spin configuration with unpaired electrons populating the  $e_g$  set, or in low spin state with the  $e_g$  set empty (or in intermediate states between these two extremes). The energetic difference between these states is termed the spin splitting energy,  $\Delta E_{H-L}$  (Figure 2-2), and will be a key prediction target in this thesis (particularly Chapters 3 and 5). Accurate knowledge of the relative energetics between these different states is of great importance as the spin state can dramatically the geometric<sup>51</sup>, magnetic<sup>173</sup>, electronic<sup>52-54,63</sup> and catalytic properties<sup>57,58</sup> of the transition metal complex.

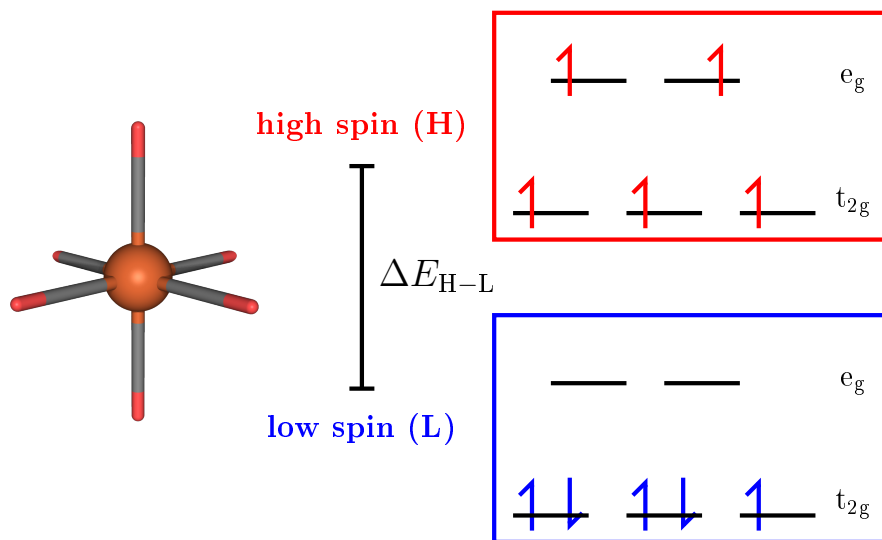


Figure 2-2: An octahedral Fe(III)[(CO)<sub>6</sub>] complex (left) and two possible spin states (right), showing the d-orbital splitting into  $t_{2g}$  and  $e_g$  sets with either high- (red) or low-spin (blue) occupations of the five 3d electrons indicated with barbed arrows. The spin splitting energy,  $\Delta E_{H-L}$ , is the difference in energy between these states.

This ambiguity in ground spin state typically motivates studying multiple spin states simultaneously, driving up the computational cost of simulation, particularly for HTVS. Spin is a fundamental quantum mechanical property and hence computational prediction of the spin preferences of a given compound requires first principles simulation. In addition, transition metal complexes may exist in multiple different oxidation states (i.e. number of d electrons), which adds and additional complexity to screening efforts.

Beyond questions of size, oxidation state and spin, inorganic chemistry lags behind the open-source infrastructure that facilitates HTVS for organic and periodic systems (Figure 2-3). Computational screening in organic chemistry benefits from mature, feature-rich, open-source toolboxes such as RDKit<sup>76</sup> and OpenBabel<sup>77</sup>, compact and efficient representations such as SMILES strings<sup>174</sup>, accurate and broadly-applicable classical methods<sup>175</sup> (i.e. force fields), freely-accessible databases of millions potential candidates such as ChEMBL<sup>176</sup> or ZINC<sup>177</sup>, and well-defined notions of chemical similarity, for example the Tanimoto similarity<sup>178</sup>. For periodic systems, the Atom Simulation Environment<sup>179</sup>, pymatgen and AFLOW<sup>75</sup>/OQMD<sup>12</sup> provide some similar capacity. Machine learning on organic chemistry has benefited from large quantum chemical data sets that have been assembled specifically to facilitate data-driven modeling, such as QM9<sup>31</sup>, including multiple DFT-derived properties for 130k small organic molecules, or the ANI-1 databases<sup>32</sup>, consisting of 20 million off-equilibrium DFT energy calculations on small organic molecules.

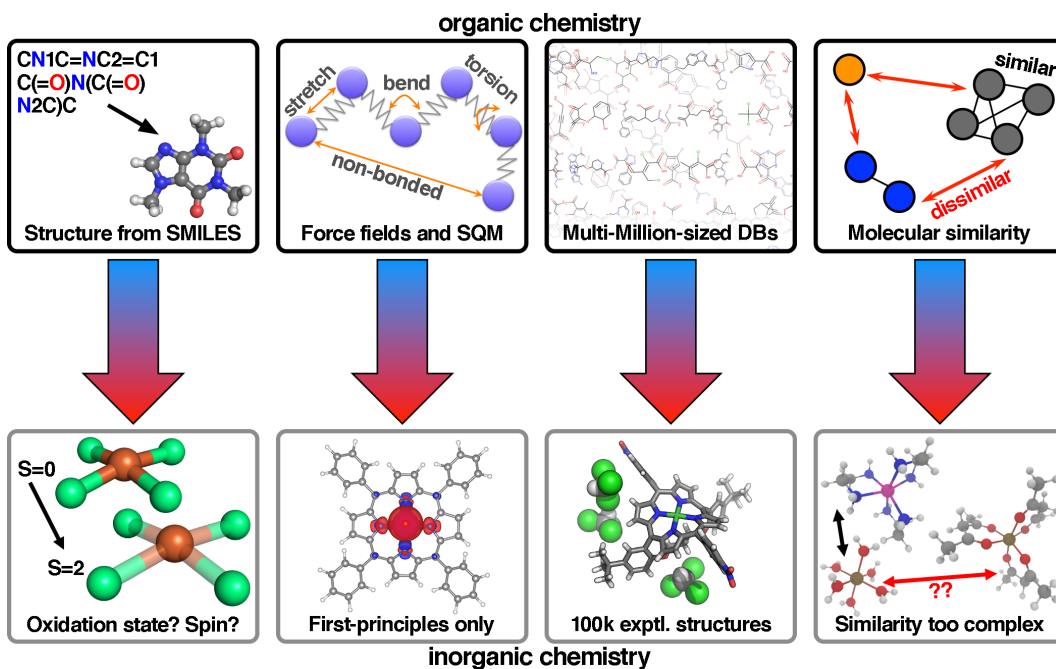


Figure 2-3: Differences between HTVS in organic (top) and inorganic chemistry (bottom). From left to right: 3D structure generation, applicable simulation methods, databases and concepts of molecular similarity. Reproduced from Ref. <sup>180</sup>.

By contrast, many of these advances are not applicable to transition metal complexes<sup>181</sup>: what few force fields are applicable<sup>182</sup> are insufficiently flexible to handle spin and oxidation dependence of metal-ligand bonding<sup>180</sup>, common<sup>76,77</sup> implementations are unable to decode SMILES strings to good starting geometries, notions of similarity are not clearly defined and what few databases are available<sup>183,184</sup> are an order of magnitude smaller ( $\sim 10^5$  complexes) and consist of experimentally characterized structures, limiting their potential for new chemical discovery.

Fortunately, molSimplify<sup>181</sup>([github.com/hjkgrp/molSimplify](https://github.com/hjkgrp/molSimplify)), an open source inorganic complex construction toolbox recently developed at MIT, has begun to address these deficiencies, providing a scriptable framework for creating 3D geometries for transition metal complexes and automating HTVS for these systems. This thesis

makes extensive use of, and in turn contributes to the development of, molSimplify. Even if sufficiently many complexes can be simulated, the accuracy of DFT for transition metal complexes is highly variable<sup>185</sup>. The variety of low-lying electronic d states that facilitates the tunability of transition metal complexes, and hence their useful properties as functional materials, simultaneously creates difficulties for first-principles simulations<sup>73,186</sup> due to many near-degenerate orbitals. Rigorous handling of these effects requires specialized multi-configurational methods<sup>51,185,187</sup> such as multi-configurational self-consistent field (MCSCF)<sup>188</sup> or complete active-space second-order perturbation theory (CASPT2)<sup>189</sup>, which are not ‘black box’ and require expert input<sup>187</sup> (though efforts to automate these decisions have shown promise<sup>190,191</sup>). Nonetheless, DFT has been applied extensively to study properties of transition metal complexes such as spin state ordering<sup>74,162,192</sup> and redox<sup>69</sup>. Studies have indicated surprisingly good qualitative performance from hybrid functionals<sup>186</sup> for calculations of redox potentials and spin splitting energies, though empirical corrections have been suggested to improve agreement with experimental measurements<sup>193</sup>. In particular, different amounts of exact exchange are advocated for transition metal complexes in the literature, ranging from 0%<sup>194</sup> or 15%<sup>74</sup> (so-called B3LYP\*) to 30%–50% or more<sup>195,196</sup>. The impact of varying the exact exchange fraction on spin state ordering has been thoroughly investigated<sup>59,192,194,197–199</sup>, with three main conclusions (Figure 2-4): 1) the spin state ordering, as characterized by the energy difference between high- and low-spin configurations, is a nearly linear function of the exchange fraction over the range 0%–30%; 2) the sensitivity to the exchange fraction is a strong function of ligand field, i.e. the gradient of this linear relationship is typically larger for strong field ligands (those that bias toward low spin states); 3) the amount of exchange needed to capture the correct spin state ordering is different for different systems – for example, exchange of 15% was found to suffice

for a series of Fe(II/III) spin crossover complexes (SCOs)<sup>59,200</sup>, while a study across different metals and ligand fields found 40% to give the best average agreement with high-accuracy wavefunction methods<sup>201</sup>. In summary, choice of exchange fraction serves as a lever to correct errors in DFT methods but the appropriate fraction for a given calculation is not known *a priori*, acting as source of uncertainty for simulation results.

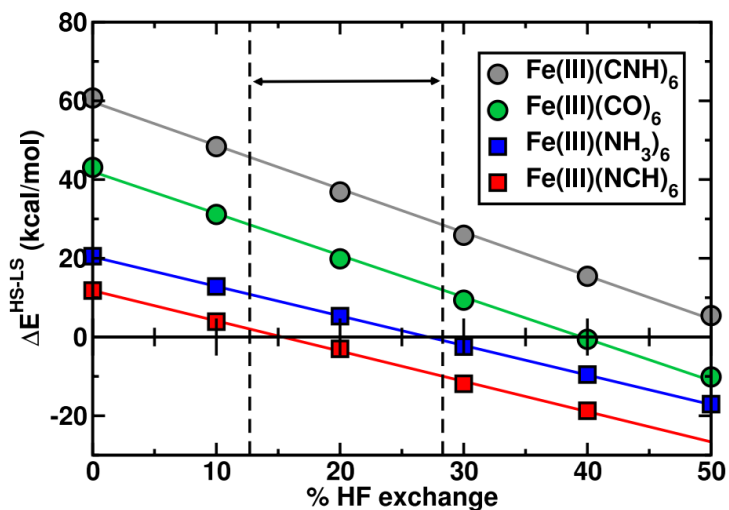


Figure 2-4: The effect of changing the fraction of exact exchange used on the DFT-predicted energetic difference between high-spin and low-spin states ( $\Delta E_{H-L}$ ) for four octahedral, homoleptic Fe(III) complexes with ligands: NCH, NH<sub>3</sub>, CNH and CO with a  $3\sigma$  confidence interval from normal distribution poll data on hybrid exchange functionals, as indicated with black dashed lines and black arrow. Results are calculated with B3LYP-like DFT. Reproduced from Ref.<sup>197</sup>.

Practical comparisons with experimental systems also require corrections for solvation<sup>202,203</sup> and finite-temperature thermodynamic effects<sup>192,204</sup>, further increasing the computational burden and requiring additional approximations – for example, the conductor like polarizable continuum model<sup>205</sup>. In spite of all of these factors, accuracy for calculation of spin splitting energies with hybrid functions is  $\sim 10$ kcal/mol

out-of-the-box and qualitatively-correct behavior can be obtained, with results improved after adjusting exchange fractions<sup>59,196</sup> or adding post-facto empirical corrections<sup>206</sup>. The accuracy of hybrid DFT for transition metal complexes has also been investigated, finding excellent correlation between computed and experimental redox potentials ( $R^2 > 0.95$  according to Ref.<sup>203</sup>) with typical errors of around 0.15eV<sup>19</sup> or 0.1 eV<sup>207</sup>. For quantitatively comparing redox potentials to experimentally measured values, it is necessary to use a reference potential calculated with the same method<sup>203,207</sup> to benefit from cancellations of systematic errors.

## 2.2 Machine learning in chemistry

Note: Parts of this section have been submitted for publication in the ACS Elements series, intended to provide an pedagogical introduction to machine learning in chemical sciences. They have been abridged and reformatted for consistency.

### 2.2.1 Overview

Machine learning (ML) methods are applied to a great variety of problems in chemistry, and will doubtless continue to find novel applications in the future. While all of these application areas necessitate specific adaptations, (almost) all fit into the basic framework of supervised learning. This section provides a review of the basic ideas underlying supervised learning that will be used in this thesis, and is largely based on the books by Vapnik<sup>208</sup> and Hastie *et al.*<sup>209</sup>.

#### Supervised learning

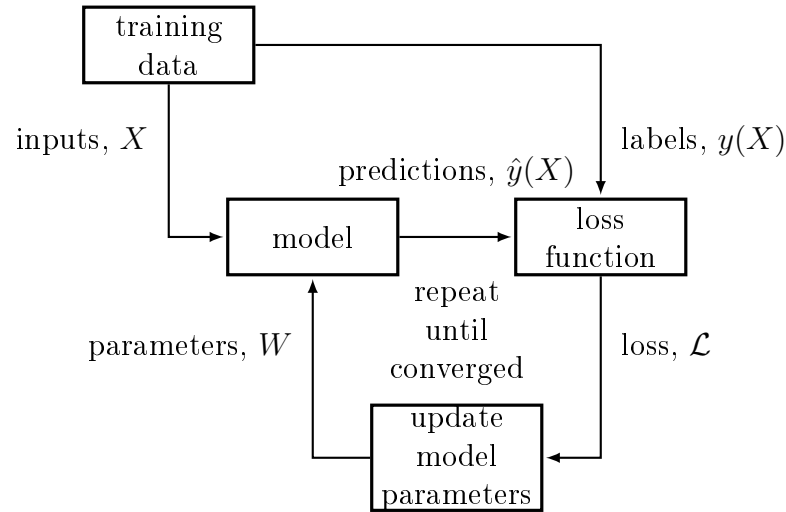
The fundamental objective of supervised learning is to develop a regression model or function,  $f$ , that is capable of making predictions in response to some supplied

inputs,  $x$ , denoted generally as  $\hat{y}(x) = f(x)$ . There are few restrictions on what  $x$  and  $\hat{y}$  may be: scalars, vectors, images, graphs and more exotic types of data are possible. In chemistry the inputs are often different arrangements of atoms in 3D space and the outputs could be the corresponding energy of the system – this is the same task that is the main objective of computational chemistry (e.g. molecular mechanics or first-principles simulations). The choice of model  $f$  is more standardized and typically comes from a few major families: kernel methods, random forests, or neural networks models. In either case, the model will depend on some parameters, denoted  $W$ :

$$\hat{y}(x) := f(x, W) \tag{2.1}$$

The choice of these parameters will uniquely determine the behavior of the model and the central task in training the model, is the selection of these parameters. This is one way in which the soft distinction between ‘machine learning’ and general regression might be drawn: in machine learning, there is typically little effort made to adapt the structure of the model to the task at hand, instead preferring a very flexible model family with many parameters<sup>210</sup>, and making an intelligent choice of  $W$ . For example, neural network potentials (NNPs), a family of neural network models that directly relate atomic coordinates to energies<sup>24</sup>, are similar in purpose to molecular mechanics force fields and both are parameterized to agree with experimental observations or first-principles simulations. However, while force fields assume structured nonlinear equations based on polynomials (e.g. electrostatic repulsion and harmonic oscillators), NNPs make no explicit assumptions about the type of functions relating the atomic positions and energies and instead start with a very general form and learn to reproduce the structure-energy relation them from the data directly.

## 1: Training phase



## 2: Testing phase

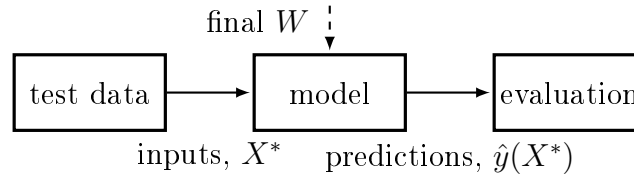


Figure 2-5: Overview of a typical iterative supervised learning training and testing process.

Supervised learning here refers to the use of  $n$  labeled *training data* pairs of previous examples,  $(x_i, y_i)$ , to infer the relationship between inputs and outputs, as controlled by the model parameters,  $W$ . We will use  $X$  and  $y$  without subscripts to refer to the collection of all the data and with subscripts to refer to individual data points, and adopt the convention of forming the training data into a *data matrix*,  $X \in \mathbb{R}^{n \times d}$ , which has the  $n$  individual observations as rows and the descriptors for each the observations as columns. This immediately requires a way of determining how good a given value of  $W$  is for capturing the relationship between  $X$  and  $y$  with a *loss function*. A loss function is a way to measure how well the model predictions match



with the data, and the obvious choice is compute the ( $l^2$ ) norm of the error between  $y$  and  $\hat{y}$  for some input  $X$ :

$$\mathcal{L}(y, \hat{y}(X)) := \|y - \hat{y}(X)\|_2^2 = \|y - f(X, W)\|_2^2 = \sum_{i=1}^n (y_i - f(x_i, W))^2 \quad (2.2)$$

This is the well known "least-squares" error function and has number of useful properties that are worth discussing briefly. First, it is a strictly non-negative, even function of the errors  $y_i - \hat{y}(x_i)$ , and is differentiable and moreover a convex function (of  $\hat{y}$ ), which is extremely important for optimization. This provides optimality guarantees for any minima if  $f$  itself is convex, as in the case of linear least-squares regression. There are alternative loss functions as well, particularly for classification tasks where the output  $\hat{y}$  is a probability associated with a certain label<sup>209</sup>, but this simple metric is used in many applications of ML in chemical sciences. The learning task can then be understood as a problem of minimizing the chosen loss function. This is often done iteratively by starting with some initial choice, perhaps normally distributed about zero, and then optimizing  $W$  using a traditional optimization routine to update  $W$  and reduce the loss. In some cases this optimization can be done directly, as is the case with linear and kernel methods, but generally it must be carried out numerically, for example by gradient descent and derivatives thereof (Figure 2-5). The amount of data needed is highly application dependent and the availability of data is related to the source and cost of acquiring the data. For the construction of high-quality NNPs, thousands to hundreds of thousand of observations of molecular configurations evaluated using density functional theory (DFT) are routinely used<sup>24,211,212</sup>, as sufficient data is required to infer relationships between the bond angles and distances and the relative energy of the atomic configuration. Modern general purpose NNPs for organic chemistry have been trained on

more than 20 million DFT geometries<sup>213</sup>, but NNPs that are specialized for limited element compositions and conditions, for example those developed by Behler and coworkers for zinc oxides<sup>214</sup> or water clusters<sup>215</sup>, can achieve high accuracy with tens of thousands of DFT geometries or less. Data points for ML can be drawn from large databases<sup>216</sup> or measured from experiments directly<sup>217</sup>, but the most common source is from the output of computational chemistry calculations.

Once the training procedure is complete, and some finalized model parameters are selected, the model is typically tested on some *out of sample (OOS)* or *test* data, which are observations that are not drawn from the training data. This a critical step in the assessment of any fitted ML model, especially those with large numbers of parameters, because of the risk of over-fitting for sufficiently complicated models.

## Statistical learning theory

Over-fitting is defined as a mismatch between the *true risk* – the loss function over all inputs, and the *empirical risk*, the loss on the finite observations used to calibrate the model, for a certain choice of model  $f$  (uniquely determined by the parameters  $W$ ):

$$\mathcal{E}_{\text{emp}}(f) = \mathcal{E}_{\text{emp}}(W) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(y_i, f(x_i, W)) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i, W))^2 \quad (2.3)$$

where the second equality only applies in the case of the least-squares loss function. The basic process of supervised learning (Figure 2-5) is to update the model

parameters,  $W$ , to make  $\mathcal{E}_{\text{emp}}$  as small as possible. We denote this function

$$\hat{f} := \arg \min_{f \in \mathcal{T}} \mathcal{E}_{\text{emp}}(f) \quad (2.4)$$

This is the best function over a family of models,  $\mathcal{T}$  considered. However, the objective of constructing a surrogate model is to build a useful proxy for a physical process that is applicable to new values of  $x$  not in the training data. For example, in drug design one might train a model to predict activity of some known reference molecules and want to apply the model to screen unknown targets<sup>218</sup>, or one might want to use a NNP to assess the energy of unknown conformations<sup>24</sup>. This is called *generalization* and can be mathematically formulated as the expected risk for any fixed function,  $f$ :

$$\mathcal{E}(f) := \int_{\mathcal{X} \times \mathbb{R}} \mathcal{L}(y, f(x)) p(x, y) dx dy = \mathbb{E}[\mathcal{L}(Y, f(X))] \quad (2.5)$$

Here, the integral is taken over all possible inputs and outputs, and  $p(x, y)$  represents the true, usually unknowable joint probability distribution between the random variables  $X$  and  $Y$ . Application of the laws of probability gives that the function that minimizes  $\mathcal{E}(f)$  is  $f^*(x) := \mathbb{E}[\mathcal{L}(Y, f(X)) | X = x]$ , i.e., the conditional mean value of  $Y$  given  $x$ . This formulation allows for non-deterministic relationships between  $x$  and  $y$ , as in the case of processes with measurement noise. The field of *statistical learning theory* is concerned with the analysis of equations 2.3–2.5 to determine when models found by minimizing empirical risk can be expected to generalize, i.e. give true low risk. A central concept<sup>208</sup> is that *generalization error* can be decomposed

into two contributions:

$$f^\dagger := \arg \min_{f \in \mathcal{T}} \mathcal{E}(f) \tag{2.6}$$

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) = \left( \mathcal{E}(\hat{f}) - \mathcal{E}(f^\dagger) \right) + \left( \mathcal{E}(f^\dagger) - \mathcal{E}(f^*) \right) \tag{2.7}$$

the first term is the *estimation error*, which arises when from making a sub-optimal choice of function based on training data, commonly called *over-fitting*. The second term is the *approximation error*, the error made by restricting the choice of model family. For example, using a linear model, for an application that might not be captured by a linear relationship. There are two critical ideas that are worth noting<sup>208</sup>: under mild assumptions one can show that:

1. Assume that the space of possible models is complex enough to have near-zero approximation error, for example using a very nonlinear model that can fit any data exactly. Then, if we draw training points from a fixed distribution, our empirical risk will converge to the risk for that distribution as the number of training points increases, i.e. our model will generalize.
2. The rate of convergence as we add more data is inversely related to size of  $\mathcal{T}$  i.e., more complex model spaces require more data to generalize.

The rigorous definition of ‘size’ is the Vapnik–Chervonenkis dimension<sup>219</sup>, but the above can be intuitively interpreted as the set of model functions  $\mathcal{T}$  to be large/complex enough to have low approximation error, but no more complex. With limited data, it may be better searching for a model in a simpler, smaller family of models that ‘learn’ more robustly and quickly as opposed to very complex models with many parameters. Conversely, a simple model will stop improving with more data past a certain point – where the approximation error dominates.

## Regularization

Regularization techniques allow control model complexity in a fine-grain manner, allowing us to search for the correct level of model complexity using a smooth parameterization by adding a penalty,  $R(f(X))$ , to the loss function, which depends only on the complexity of the model and not on how well it fits the data.

$$\mathcal{L}'(y, f(X)) = \mathcal{L}(y, f(X)) + \lambda R(f(X)) \quad (2.8)$$

Here,  $\lambda \geq 0$  balances between making the model fit the data and forcing the model to be as simple as possible. As a concrete example, the most common type of regularization<sup>208,209</sup> is *Tikhonov* or  $\ell_2$  regularization, which is given by the  $\ell_2$  norm of the model parameters,  $R(f(X, W)) := \|W\|_2^2$ . Combining this with the square loss function and writing out the explicit dependence on the model parameters gives:

$$\mathcal{L}'(y, f(X, W)) = \frac{1}{n} \|y - f(X, W)\|_2^2 + \lambda \|W\|_2^2 \quad (2.9)$$

By optimizing  $\mathcal{L}'$  in place of  $\mathcal{L}$ , some of the fit to training data will be sacrificed to reduce the magnitude of the weights. However, the model will be simpler and will often generalize better (i.e. have a lower true risk). It is useful to note that, via Lagrange multipliers, minimizing eq. 2.9 with respect to some  $p$ -dimensional parameters  $W$  is equivalent to minimizing:

$$\min_{W \in \mathbb{R}^p} \frac{1}{n} \|y - f(X, W)\|_2^2 \quad (2.10)$$

$$\text{s.t. } \|W\|_2^2 \leq r \quad (2.11)$$

This establishes that setting a value of  $\lambda$  defines a budget for parameters (how large we allow  $t$  to become), such that large values of  $\lambda$  result in small coefficients  $W$  and vice-versa.

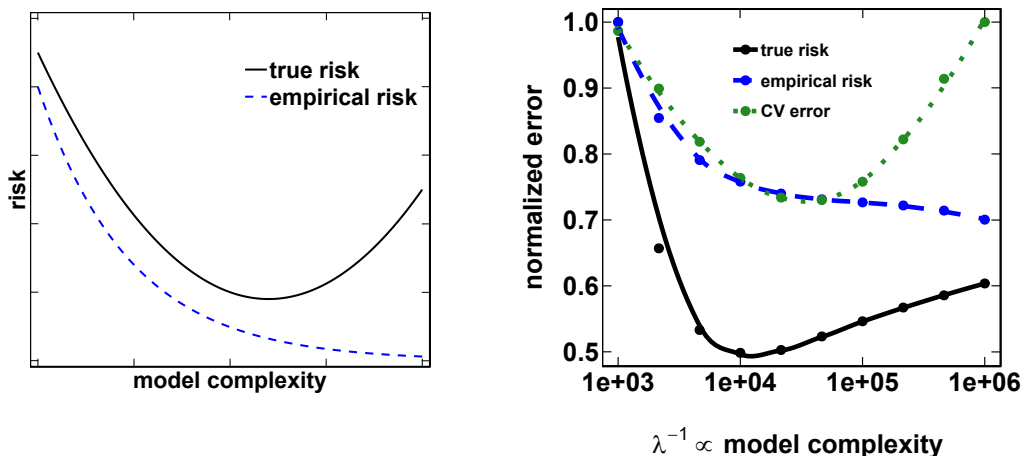


Figure 2-6: The ‘classical’ picture of over-fitting: (left) illustration of the relationship between model complexity and true and empirical risk, showing that sufficiently flexible models will always fit training data arbitrarily well but will tend to generalize poorly (right) CV error estimate as an indicator of true risk, showing fitting of  $y = \sin(\pi x)$  with a 15<sup>th</sup> order polynomial 30 points uniformly sampled in  $[0, 1]$  with measurement noise  $\mathcal{N}(0, 0.15)$  and different levels of Tikhonov regularization  $\lambda$ , which is inversely correlated with model complexity. An LOOCV estimate of the error is computed for each value of  $\lambda$  (green), and compared to the empirical (blue) and true (black) risks.

$\lambda$  is called a *hyperparameter*, as opposed to a parameter, because while both affect model performance, it is not selected based on the training data – it is set for the model before we begin the training process. Note that trying to set the derivative of eq. 2.9 with respect to  $\lambda$  to zero in order to minimize our training loss would trivially result in  $\lambda = 0$ , aligning with the notion that more complicated models can always fit training data better. However, while regularization provides us with a fine-grained control on model complexity, it does not help to answer the important question: *how*

*complicated should my model be?*

Unfortunately, errors on training data cannot help determine how complicated a model the data set can support. Figure 2-6 shows a schematic illustration of this issue – although the empirical risks, that is the error on the training data, continues to decrease as complexity is added, the real ability of the model to predict new results (true risk) begins to degrade.

### **(Cross)-validation and model selection**

Instead, one can use *validation* data to answer this question. That is, the ‘training data’ should be further divided into a set of data that will be used for training and a validation set to estimate how well each model generalizes. The distinction between validation and test data is that the ‘validation data’ is used to help construct the model (e.g., by selecting how much regularization to use), whereas ‘test data’ is not used at any stage of model training.

In order to reduce dependence on the particular way training and validation data are partitioned, it is standard practice to use a technique known as *cross-validation* (CV), which involves sub-dividing the training data into  $k$  equal folds. Then, remove the first fold is removed from the training data and the remaining  $k - 1$  folds are used to train the model, with a fixed choice of the hyperparameters. This model is tested on the left-out fold to generate an error estimate for OOS data. This error is recorded, and then the training process is repeated repeated  $k$  times, such that each fold has been left out exactly once. These errors are averaged to produce the *cross-validation error* (inner loop in Figure 2-7).

This metric for error is an estimator for the generalization error made by the model since it is only based on data not used to train the model at each step. Of course,

this increases the cost of training the model by a factor of  $k$  for each model that is to be evaluated.

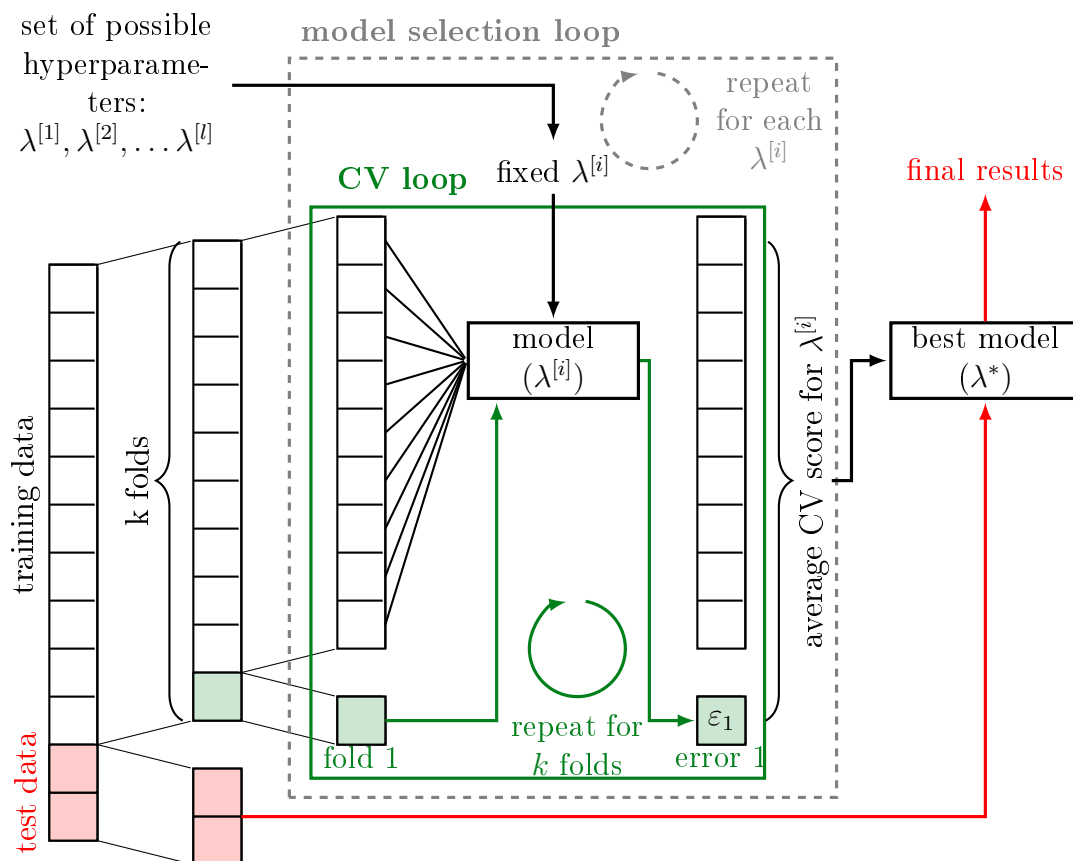


Figure 2-7: Illustration of the machinery for model selection via double loop grid search  $k$ -fold cross-validation (CV) for a single hyperparameter  $\lambda$ . The full dataset is split into test and train fractions. Then, for each choice of the hyperparameter  $\lambda$  (the gray outer loop), the training data is further split into  $k$  equal folds and then a series of  $k$  models are trained with one fold left out of the training process each time (green inner loop). Each model is evaluated on the left-out fold to generate  $k$  errors  $\epsilon_1, \epsilon_2, \dots, \epsilon_k$ . The average error for each  $\lambda$  is compared and the best  $\lambda^*$  is chosen to train a model on the full training set. The test data is only used to evaluate the final model.



Because the average cross-validation (CV) error is an estimate of the generalizability of a given model and hyperparameter combination, it can be used to select among different models and hyperparameters (Figure 2-6).

The choice of  $k$  is typically between 5 and the amount of data available,  $n$ . The special case where  $k = n$  is called *leave one out cross-validation* (LOOCV, used in Figure 2-6) because only one point is left out in each fold.

In practice 5 or 10 are common choices, with the smaller values of  $k$  being cheaper to evaluate since the training process is repeated fewer times. The larger the value of  $k$ , the less biased the CV error estimator is relative to generalization potential of the model trained on all the training data\*. When using small data sets, small values of  $k$  may be undesirable since the model trained on only  $\frac{k-1}{k}$  of the data may be highly unstable. When a large amount of data is available, the overhead of repeating the training process is impractical, and a large  $k$  may not be feasible.

Since in practice choosing good values of the hyperparameters is difficult, and hyperparameters are typically interrelated, a standard approach is to perform *grid search cross-validation* (outer loop in Figure 2-7). Here, a range of values is selected for each hyperparameter, and this range is discretized into sample values to try for each hyperparameter. Every combination of hyperparameters can then be scored using CV, and the best performing model is then re-trained on all of the training data (and evaluated against the held-out test data). This provides a robust method for model selection and can be used to decide between different types of models (e.g., linear vs. non-linear or maximum degree of non-linearity) as well. This pipeline is essentially a double loop, requiring training the model  $k$  times for each hyperparameter combination. For robustness to how the test/train data is partitioned, it is also possible to repeat this entire process for different test/train splits, forming a triple loop scheme

---

\*This bias is introduced precisely because a smaller fraction of data is used.

sometimes called *nested cross-validation*. This can be important for small or highly heterogeneous data sets, and testing robustness to test/train splits is always a good practice.

Alternatives to CV do exist, for example bootstrapping<sup>220</sup>, where random training sets of consistent size are assembled by drawing from the original data set with repetition. These approaches face similar costs of model training time. For more than a handful of hyperparameter values, an exhaustive grid search becomes infeasible, and so it is natural to seek more optimal strategies, particularly for neural networks or other models where the training cost is substantial. One approach is to use Bayesian methods to optimally sample the hyperparameter space<sup>221,222</sup>. For complex models, practitioners may be forced to do with a single validation split.

## 2.2.2 Representations

All machine learning models, from linear regression to neural networks, take numbers, vectors, and matrices as inputs and transform them to outputs. Therefore, in order to apply machine learning to molecules or materials consisting of atoms, they must first be converted into numerical *features* or *descriptors* that describe each observation. Molecules or materials generally correspond to discrete objects in chemical space,  $c^{(i)}$ , which we may then featurize in a manner that produces discrete vectors,  $x^{(i)}$ , in some  $d$ -dimensional vector space  $\mathcal{X} \subset \mathbb{R}^d$  (Figure 2-8). For efficient inference it is desirable to build a feature space in which molecules or materials with similar properties are proximal to each other<sup>223</sup> (i.e., a small  $d(x^{(i)}, x^{(j)})$  in Figure 2-8), but it can be challenging to know *a priori* how to build such a feature space.

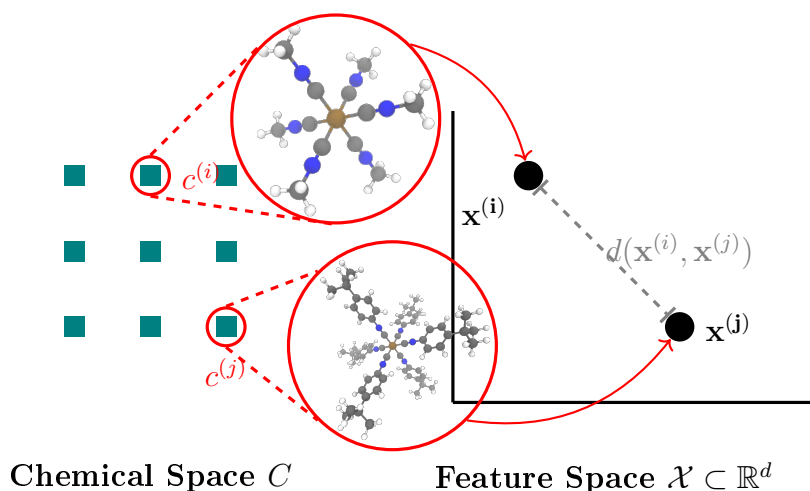


Figure 2-8: Basic overview of the featurization process, mapping elements in chemical space to vectors that represent them in descriptor space. The choice of the featurization governs the geometric similarity and overall geometry of the descriptor space.

For these reasons, a significant body of work has focused on the development of representations for atomistic systems<sup>81,223–225</sup>. The earliest developments date back to the 1960s-1980s in the cheminformatics community and featurization continues to be an area of active research<sup>223,226</sup>. Unfortunately, one cannot expect there to be a single ideal featurization for the broad range of challenges of interest to researchers in the chemical sciences. Before getting into specific featurization strategies, we will consider what makes a feature set broadly suitable for a given problem. There are some ideal characteristics of representations that are all typically only imperfectly realized by real feature sets:

1. The representation should be invariant to equivalent inputs that correspond to the same physical system and conversely should provide a unique encoding for each distinct input.
2. The representation should encode property (i.e., chemical) similarity, i.e. the

distribution of observations in the feature space should correspond with the distribution in property. The worse the input-output correspondence is, the more nonlinearity is required of the model to make predictions.

3. The representation should be as affordable and straightforward to compute as possible<sup>227</sup>. This requirement is particularly applicable to training of ML models based on first principles calculation where attributes of the wavefunction may correlate very well with output properties but may be too expensive to compute over large numbers of inputs..
4. The representation should be easy to interpret to enable extraction or encapsulation into derived heuristics or physical principles. While not critical in terms of model performance, it is much easier for the practitioner to understand why models make a certain prediction or how to improve predictions when features are interpretable.
5. Finally, a highly desirable aspect of a representation would be bijection, that is any point in the feature space should be able to be inverted back to a real molecule or material<sup>227</sup>. Representations that have such qualities enable so-called inverse design strategies, where optimization can be conducted in the continuous descriptor space<sup>228</sup>.

Another important distinction is to be drawn between what system or portion of a system the descriptors represent. In chemistry, it is natural to focus on the properties of individual molecules, as has historically been done in quantitative structure–property relationship (QSPR) modelling. In this case, features are computed on a per-molecule basis, even when the ultimate property being predicted may depend

on many other molecules present, such as solvent or the human body, as is the case in quantitative-structure activity relationship building for therapeutic drug design. There are many applications where representations must be computed for multiple molecules at a time or for systems for which the chemical bonding is not strictly covalent and the definition of a molecule becomes challenging. For example, predicting reaction outcomes from multiple reactant species<sup>229,230</sup> to one or more products or in predicting the binding energy of reactant species on catalysts on surfaces<sup>231,232</sup>. For periodic systems, such as metal-organic frameworks<sup>233</sup> or crystalline materials<sup>80</sup>, challenges in choosing how to represent the unit cell can also be apparent. In the case of multiple molecules being involved, one option is to consider all molecules present as parts of the single, partially disconnected system. Both of these issues can be handled naturally by representing the systems through their connectivity in a molecular graph representation introduced below.

Alternatively, in analogy with classical molecular mechanics force fields, local atomic environments can be featurized. Thus, each molecule This essentially converts each observation of a molecule into a large number of atomic features. The output property used in training machine learning models is usually obtained on a per-molecule basis, but some properties, especially the atomization energy<sup>24</sup> or local atomic partial charges<sup>234</sup>, can be naturally broken down into a sum of contributions of individual atoms. Thus, a mapping to local atomic environments may be more suitable for some property predictions than others. Increasing ease of training larger neural networks with modern computational resources has motivated the use of *learned representations*, which sidestep complex, pre-computed features and instead learn an internal representation at the same time as the model is trained starting only from simple atomic identity and connectivity information<sup>81,235</sup>

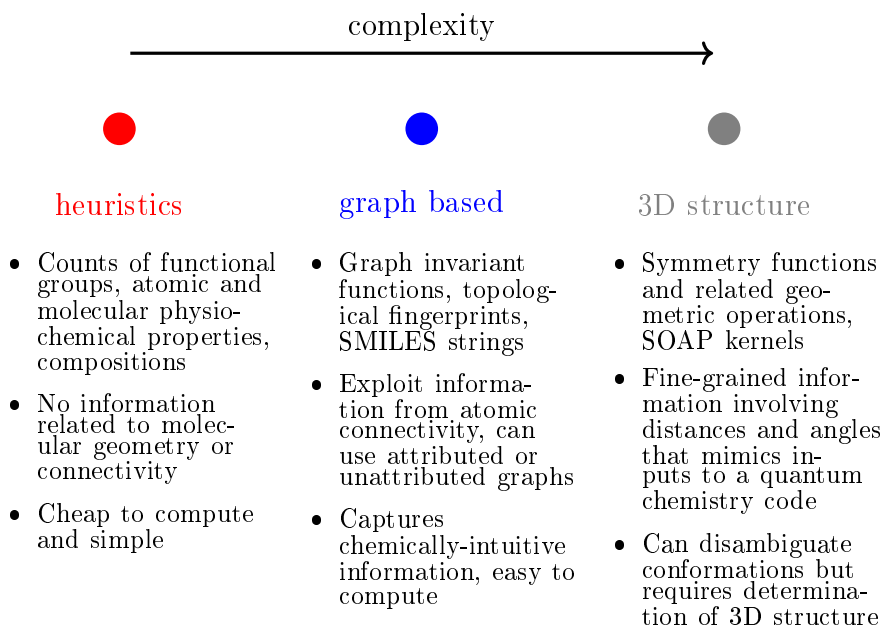


Figure 2-9: Comparison of some common featurization strategies, organized by the level of detail included/complexity.

The simplest possible feature sets are collections of *ad hoc* system properties that are counted or inferred from the input molecule directly. Examples of these features range from the number of each element type present to tabulated, semi-empirical or quantum mechanically-calculated, physicochemical properties of molecules such as molecular weight or lipophilicity. These types of properties have been used in the development of quantitative structure property or quantitative structure activity relationship (i.e., QSPR and quantitative structure–activity relationship (QSAR)) models for decades<sup>224</sup>. When empirical *ad hoc* features are employed, they are typically cheap to compute. However, because constructing these feature sets assumes some knowledge of the most important predictors of a chemical property *a priori*, models built on these features will not be unique or can be insensitive to small

changes in chemistry. These features have shown good baseline results in many areas, including in challenging problems, especially those that are not well-described by covalent bonding. Some examples are periodic ‘materials’ systems<sup>236,237</sup> and heterogeneous catalysis<sup>231,238</sup>.

*Fingerprints* are an important subset of classical cheminformatics descriptors that are vectorial representations of molecules computed by breaking the molecule into fragments. There is a great degree of flexibility in how this is done and, as a result, a large number of fingerprints. The most basic molecular fingerprints, such as the FP2 fingerprint<sup>77</sup>, are binary vectors that count if a specific functional group (e.g., a phenyl ring) is present. Each position in the vector corresponds to a functional group or collection of functional groups. To include additional information about functional group proximity, *topological* features are used. These feature sets use information about the connectivity of atoms in the molecule, sometimes including information about the bond order between atoms, without depending on any specific 3D arrangement of atoms. We will also refer to these as features based on the *molecular graph*, which is a mathematical graph with atoms as vertices and bonds as the edges (Figure 2-10). Graph representations may include vertex and edge attributes. The representation can be modified by adding more attributes to the graph, e.g., atomic nuclear charges, bond orders, or even bond lengths from 3D geometry. Graphs are an attractive mathematical representation for machine learning in chemistry because they inherently reflect the locality of chemistry and the proximity or separation between different components of a molecule or material.

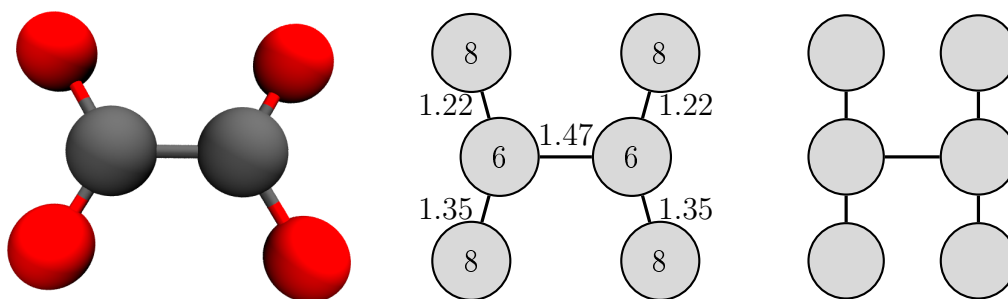


Figure 2-10: Graph representations of oxalate ion,  $\text{C}_2\text{O}_4^{2-}$  (left); a weighted representation with atomic numbers and bond distances (center); and an unweighted representation (right).

For the simplest representations, graphs may be unattributed with indistinguishable vertices that only map connectivity and no information about elemental identities. Even this simplified graph model generates useful descriptors. The Randić<sup>239</sup> index, which was proposed over 40 years ago, is a measure of molecular branching that is known to correlate well with boiling points of alkanes and is given by the equation:

$$r_\alpha := \sum_i \sum_j (\text{deg}(i) \text{deg}(j))^\alpha \quad (2.12)$$

Here,  $\text{deg}(i)$  refers to the degree, or number of bonds that node/atom  $i$  has in the graph, the sum is taken over all bonds  $i, j$  in the graph, and  $\alpha$  is a constant, originally  $-\frac{1}{2}$  or  $-1$ . Other well-known features based on unattributed graphs are the Wiener<sup>240</sup> index and the many descriptors proposed by Kier and Hall<sup>241-243</sup>, which are variants of bond counting methods that differentiate bonding patterns in molecules.

Attributed graphs readily incorporate more chemical information by associating nodes and edges with specific chemical properties. For example, autocorrelation descriptors<sup>22,244</sup> are calculated as functions of physical properties of the atoms in the



graph that are separated by a fixed  $d$  bonds:

$$AC_d = \sum_i \sum_j P_i P_j \delta(d_{ij}, d) \quad (2.13)$$

where  $P$  refers to an atomic property, for example nuclear charge or electronegativity, and  $\delta(d_{ij}, d)$  is one if atoms  $i$  and  $j$  are separated by  $d$  bonds. By varying the number of bonds considered and using different atomic properties, a large number of possible features can be generated. Graph convolutions<sup>28,235</sup> are a recent generalization of these graph-kernel methods that learns the function to be applied on neighbouring atoms at training time and are presented in more detail in the Section on artificial neural networks (ANNs). The text based SMILES representation<sup>174</sup> contains the same information as the atom-type attributed molecular graph and has been used as a representation, e.g., in reaction outcome prediction<sup>245,246</sup>.

Many fingerprint methods that use topological information are also available, for example Morgan or extended connectivity fingerprints<sup>247,248</sup>. These fingerprints work by collecting a description of the atomic types a certain number of bonds from each atom. However, rather than producing a numerical value, extended connectivity fingerprints (ECFPs) concatenate the identifiers of adjacent atoms, resulting in a dynamically-derived list of fragments for each atom that can be hashed into a binary vector of fixed length. This is similar to the FP2 fingerprints introduced above but the key difference is that there is no pre-defined library of substructures. The ECFPs is instead dynamically generated based on the molecular graph.

None of the featurizations introduced so far incorporate 3D structural information, and thus they would not be suitable for property predictions that require 3D information (e.g., NNPs). Most 3D structural descriptors make use of internal coordinates to maintain rotation and translation invariance, although some models include symme-

try operations as part of the model and are able to use raw Cartesian coordinates as inputs<sup>249</sup>. In addition, features should not be sensitive to the atomic ordering in the molecule. Thus, pairwise Euclidean distances  $\mathbf{r}_{ij}$  between atoms  $i$  and  $j$  are a natural starting point. The Coulomb matrix<sup>250</sup>, named for its relation to the Coulomb operator, is a simple example of a 3D structural descriptor. For a molecule with  $m$  atoms, the Coulomb matrix is defined as an  $m \times m$  matrix with elements:

$$C_{i,j} := \begin{cases} 0.5 (Z_i Z_j)^{0.5} & i = j \\ \frac{Z_i Z_j}{\mathbf{r}_{ij}} & \text{otherwise} \end{cases} \quad (2.14)$$

where  $Z_i$  is the nuclear charge of atom  $i$ . The exponent of the diagonal elements in the Coulomb matrix was obtained from an empirical fitting procedure on organic molecules. By definition, matrix elements decay with  $1/r$ , leading to stronger matrix elements between chemically-bonded atoms. Although simple and intuitive to build, this representation has some limitations. The labeling order of atoms alters the structure of the matrix, and the Coulomb matrix also changes when the number of atoms changes. The first limitation can be solved in a number of ways, e.g., by using the eigenvalues of the matrix instead of the elements, by sorting the rows and columns, or by randomly-sampling different orderings<sup>251</sup>. The issue of dependence on molecule size must be solved by padding smaller matrices with zeros in order to match the size of the largest molecule either in training data or on which one would like to carry out model prediction.

Many extensions to the basic idea of eq. 2.14 have been proposed, most commonly to include angular information in a similar manner. The ‘Bag of Bonds’<sup>25</sup> descriptor explicitly factorizes and sorts the elements in the matrix by the atom types involved in each pairwise distance, e.g., grouping all of the C–C terms. The ‘Bonds, an-

gles’ machine learning (BAML) descriptor<sup>223</sup> replaces the distances with Morse and Lennard-Jones potentials with parameters from force fields. It also includes explicit three- and four-body angular terms. A slightly different family of features can be obtained by inspection of the histograms of all pairwise distances and three- and four-body angles for each group of atom types, i.e. one histogram of C–C distances, one histogram of C–H distances per molecule. Discretizing these histograms into a fixed length vector results in the histogram of distances, angles and dihedral angles (HDAD<sup>212</sup>) descriptor. A different set of features can be obtained by consideration of atomic environments, instead of molecules. As mentioned, these methods are typically associated with NNPs. The atomic environment is defined by a set of *symmetry functions* that encode pairwise distances to neighbors using a square-exponential kernel<sup>24,213</sup>, giving a scalar  $G_i^r$  for each atom  $i$ . A cutoff function,  $f_c(\mathbf{r}_{ij})$ , is used to enforce locality, decaying smoothly to zero for distances larger than a given threshold. These thresholds are typically around 6 Å or less to reduce the computational cost of the featurization, which is notably shorter range than electrostatic and van der Waals force lengthscales typically considered in molecular mechanics force fields. To address this concern, explicit physics-derived terms<sup>252,253</sup> can be alongside these necessarily short-range features. An additional angular term obtained by replacing the square-exponential with trigonometric functions is also used to give angular dependence. These terms then account for local interactions at each atom by measuring how crowded the local environment is<sup>254</sup>:

$$k_{ij}^r := e^{-\eta(\mathbf{r}_{ij}-r_s)^2} \quad (2.15)$$

$$G_i^r := \sum_{i \neq j} k_{ij}^r f_c(\mathbf{r}_{ij}) \quad (2.16)$$

Typically a number of different values of  $\eta$  and  $r_s$  are used to give a vector of sym-

metry functions for each atomic environment which decay more or less rapidly and modulate the degree of locality in 3D space. One difficulty is that this approach does not naturally distinguish atom types. To get around this limitation, a separate environment fingerprint can be constructed for each possible atom type, e.g., one for carbon and one for hydrogen. Angular symmetry functions use two inter-atomic distances for a three-body angle and so one symmetry function is needed for each pair of atom types. A typical NNP implementation<sup>213</sup> in organic chemistry (i.e., C, N, O, H, F elements) uses 32 radial symmetry functions for each atom type and 8 angular symmetry functions for each atom type pair, leading to a final dimension of about 750 for each atomic environment.

A distinct approach called the *smooth overlap of atomic densities*<sup>255</sup> generalizes the square-exponential basis used in symmetry functions to a position-dependent Gaussian function, as shown in eq. 2.17, over values of  $r_s$ . The density of neighbors around atom  $i$  is defined as:

$$\rho_i(r_s) := \sum_j e^{-\eta(\mathbf{r}_{ij}-r_s)^2} \quad (2.17)$$

This is a continuous function of  $r_s$ . Rather than constructing a fixed, finite-dimensional fingerprint from this function, the overlap of two environments can be explicitly calculated by integration, giving a similarity score between two atomic environments  $\tilde{K}_{ij}$ . While the dependence on pairwise distances ensures translational invariance, this approach is not rotationally invariant by default and therefore requires an additional integration over all possible rotations in 3D space,  $\hat{R}$ . Practical computation of the integral in eq. 2.18 is achieved with a set of radial basis functions and spherical

harmonic functions, which allows the rotational integral to be computed easily<sup>255</sup>.

$$\tilde{K}_{ij} = \int d\hat{R} \left| \int dr_s \rho_i(r_s) \rho_j(\hat{R}r_s) \right|^2 \quad (2.18)$$

$$K_{ij} = \frac{\tilde{K}_{ij}}{\sqrt{\tilde{K}_{ii} \tilde{K}_{jj}}} \quad (2.19)$$

A final normalization in eq. 2.19 provides a scaled similarity measure that is suitable for direct use as the kernel term between any two environments. Extension to systems with multiple chemical species can be accomplished by treating each pairing of atom types separately. Other proposed approaches exploit properties of the molecular wavefunction calculated with quantum chemical methods. Such features can be highly informative because the wavefunction describes the electronic configuration of the system of interest, and historical QSPR relationships have exploited this link<sup>256,257</sup>. In recent examples, atomic partial charges and vibrational modes<sup>258</sup>, bond orders<sup>259</sup>, or matrix elements from the electronic Hamiltonian<sup>217,260,261</sup> have been employed in machine learning models as features. Samples of the raw electron density grid<sup>262</sup> have also been used as input to convolutional ANNs. These approaches incur the additional cost of solving for the wavefunction at Hartree-Fock (HF) or DFT level but provide a rich source of information for statistical learning since they transfer some of the difficulty of the inference task to the electronic structure calculation. Such approaches are beneficial when the target property is still more costly than the quantum chemical evaluation. For example, geometry optimizations consist of many sequential calculations, and so wavefunction properties have been used to evaluate and predict outcomes of such calculations as they were being performed<sup>259</sup>. Similarly, orbitals from low level methods have been used as inputs to estimate higher-scaling CCSD(T) energies<sup>261</sup>. Another advantage of this

type of featurization is that it may lead to more transferable models than those based on more basic geometric or graph-based features. Beyond the examples above, promising results have also been obtained for predicting DFT exchange-correlation functional energies<sup>217,260</sup>, direct relationships between electron density and energy for orbital-free DFT<sup>143</sup>, and experimental reaction yields<sup>258</sup>.

### 2.2.3 Models

In this section, some of the main nonlinear model forms that are used to construct surrogates throughout this thesis are presented: kernel methods and artificial neural networks.

#### Linear and kernel Models

The simplest model that can be interpreted as with the kernel framework is multiple linear regression (MLR). An MLR model produces estimates,  $\hat{y}$ , for a quantity of interest,  $y$ , as linear function of the data matrix of  $n$ ,  $d$ -dimensional observations  $X \in \mathbb{R}^{n \times d}$ , by choosing  $d$  weights,  $w$ :

$$\hat{y}_{\text{MLR}}(X) := Xw, \quad w \in \mathbb{R}^d, \quad X \in \mathbb{R}^{n \times d} \quad (2.20)$$

The model can be fit using the least-squares loss function with  $l_2$  regularization (eq. 2.9) by taking the derivative with respect to our parameters and setting it equal to zero:

$$\begin{aligned} \mathcal{L}(w) &= \|\hat{y}_{\text{MLR}}(X) - y\|_2^2 + \lambda \|w\|_2^2 \\ \frac{\partial \mathcal{L}(w)}{\partial w} = 0 &\iff (X^T X + I_d \lambda)^{-1} X^T y \end{aligned} \quad (2.21)$$

Setting  $\lambda = 0$  returns to a fully standard linear least-squares regression problem, in which case the *normal equation* is recovered, representing a projection<sup>263</sup> of the data  $y$  into the column space of  $X^*$ .

The computational cost of this operation classically scales with  $d^3$  and the result is  $d$  best-fit weights. For numerical reasons the equation  $(X^T X + I_d \lambda)w = X^T y$  should be solved through a factorization technique, for example the lower-upper ( $LU$ )-decomposition (see, for example Trefethen and Bau<sup>263</sup>). However, there is another way to formulate this process that facilitates extension to non-linear (i.e., general kernel methods). For  $n > d$ , the rows of  $X$  or columns of  $X^T$  are full rank and eq. 2.21 can be rewritten to express  $w = X^T a$  for  $a \in \mathbb{R}^n$ . This corresponds to a shift of basis for  $w$  in the row space of  $X$ , instead of the column space. The use of regularization with  $\lambda > 0$  favors the smallest solution for  $a$ , as measured by the  $\ell_2$  norm (for Tikhonov regularization). Now, the MLR equations can be expressed in terms of this transform at new (test) point  $x^* \in \mathbb{R}^{1 \times d}$ . Here, each observation is a row vector:

$$\hat{y}_{\text{MLR}}(x^*) = x^* w = \sum_{j=1}^d x_j^* w_j = x^* X^T a = \sum_{i=1}^n x^* (x^{(i)})^T a_i = \sum_{i=1}^n k(x^*, x^{(i)}) a_i \quad (2.22)$$

where  $k(x', x) = \langle x', x \rangle : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  is the *linear kernel function*. To avoid confusion, parenthesis are included in the superscripts to refer to different observation indices, not powers. Eq. 2.22 provides two equivalent formulations of the linear model. The linear kernel defines the similarity of inputs by the euclidean inner product  $\langle x', x \rangle = (x')^T x$ . This means orthogonal points such as  $[0, 1]$  and  $[1, 0]$  are maximally dissimilar, resulting in a zero kernel term, while points that are co-linear

---

\*It is easily shown that  $(y - \hat{y}) \perp Xw$ , i.e. any remaining error is orthogonal to the model family

are the ‘most similar’. The corresponding vector form of the kernel equation can be expressed in terms of the kernel matrix  $K$ , with  $K_{i,j} = k(x^{(i)}, x^{(j)})$ :

$$\hat{y}(X) = Ka \tag{2.23}$$

$$a = (K + I_n\lambda)^{-1}y \tag{2.24}$$

Owing to its simplicity and interpretability, MLR has a long history of application to problems in chemical science domains. Drug design in particular has seen widespread application of linear models<sup>264,265</sup>. MLR is still routinely used to model complex systems, such as catalysis<sup>266</sup>, solubility<sup>267</sup> and reaction selectivity prediction<sup>268</sup>. Low training cost also makes regularized MLR models a good toolbox<sup>227</sup> for testing different feature sets and extracting the most correlated descriptors. However, since these model does not include any non-linearity by definition, it is essential that the target property is well correlated with the descriptors used, which may necessitate using difficult-to-compute descriptors (such as those from quantum mechanical calculations<sup>257</sup> or experiments, for example linear free energy relationships) or the range of applicability is limited to a small set of systems. This can be addressed somewhat by dividing the input space into regions and training only local models<sup>269</sup>, but MLR models typically cannot match the quantitative accuracy of more flexible models. Nonetheless, a simple model that is readily understood by non-specialists is extremely valuable

Not all functions are well expressed as linear combinations of their properties, for example we do expect the energy of a chemical bond to be a linear function of the distance between them. In general, it is required to be able to model arbitrary functional forms as well as interactions the features. To demonstrate how to extend this framework for nonlinear regression, we beginning with a simple example: fitting a



quadratic polynomial to a case with two features per input (two properties for each observation  $x^{(j)}$ :  $x_1^{(j)}$  and  $x_2^{(j)}$ ):

$$y_{\text{QUAD}}(x) := w_1 + w_2\sqrt{2}x_1 + w_3\sqrt{2}x_2 + w_4\sqrt{2}x_1x_2 + w_5x_1^2 + w_6x_2^2 \quad (2.25)$$

that is, a general quadratic function with an cross term, involving both features, controlled by  $w_4$ . The presence of the  $\sqrt{2}$  terms will simplify the expressions later, but they could be combined with the coefficients  $w$ . Notice that, although the features are transformed nonlinearly, the model is linear in its parameters. The nonlinear *feature transform*  $\varphi$  that maps the original observations in  $\mathbb{R}^{1 \times 2}$  to expanded features in  $\mathbb{R}^{1 \times 6}$  is:

$$\varphi\left(\begin{bmatrix} x_1 & x_2 \end{bmatrix}\right) = \begin{bmatrix} 1 & \sqrt{2}x_1 & \sqrt{2}x_2 & \sqrt{2}x_1x_2 & x_1^2 & x_2^2 \end{bmatrix} \quad (2.26)$$

This enlarged feature space is sometimes called the ‘lifted feature space’<sup>270</sup>. The problem can be expressed in terms of the the matrix that is obtained by applying  $\varphi$  to each row of  $X$  as  $\varphi(X) \in \mathbb{R}^{n \times 6}$ :

$$y_{\text{QUAD}}(x) := \varphi(X)w = \begin{bmatrix} 1 & \sqrt{2}x_1^{(1)} & \sqrt{2}x_2^{(1)} & \sqrt{2}x_1^{(1)}x_2^{(1)} & (x_1^{(1)})^2 & (x_2^{(1)})^2 \\ \vdots & & & \vdots & & \\ 1 & \sqrt{2}x_1^{(n)} & \sqrt{2}x_2^{(n)} & \sqrt{2}x_1^{(n)}x_2^{(n)} & (x_1^{(n)})^2 & (x_2^{(n)})^2 \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_6 \end{bmatrix} \quad (2.27)$$

Eq. 2.27 shows that the proposed model is linear in the lifted feature space and there the best fit coefficients can be obtained in direct analogy to the linear case:

$$w = (\varphi(X)^T \varphi(X) + I_{d'} \lambda)^{-1} \varphi(X)^T y \quad (2.28)$$

Note that the dimension of the inverse in eq. 2.28 is now  $d' = 6$ , larger than the  $d = 2$  linear case and hence the computational cost of solving the equation is in principle  $\sim 30\times$  larger. In general, the dimension of the lifted feature space will grow rapidly with the dimension of the underlying features. For example, extending general second-order polynomials to a four variable case (including all cross-terms) leads to 16 coefficients. Fortunately, the kernel formulation proposed in eqs. 2.23–2.24 will also apply. One can equivalently operate on the  $n \times n$  kernel matrix, which is defined for this case as:

$$K = \varphi(X)\varphi(X)^T \in \mathbb{R}^{n \times n} \quad (2.29)$$

$$K_{i,j} = \langle \varphi(x^{(i)}), \varphi(x^{(j)}) \rangle = \begin{bmatrix} 1 & \sqrt{2}x_1^{(i)} & \dots & (x_1^{(i)})^2 & (x_2^{(i)})^2 \end{bmatrix} \begin{bmatrix} 1 \\ \sqrt{2}x_1^{(j)} \\ \vdots \\ (x_1^{(j)})^2 \\ (x_2^{(j)})^2 \end{bmatrix} \quad (2.30)$$

Again, the kernel matrix elements are inner products, but in the transformed space instead of the underlying features. As eqs. 2.23–2.24 apply, the number of parameters in the kernel version of the regression problem and the dimension of the matrix that needs to be inverted is only  $n$ , the number of training points, regardless of the dimension of the feature transform. This is useful if the computation of these inner products is straightforward and of low computational cost. Fortunately, this is often the case – for a quadratic model, the have the following expression is available for the kernel<sup>271</sup>:

$$K_{i,j} = ((x^{(1)})^T x^{(2)} + 1)^2 \quad (2.31)$$

By using this expression to form the kernel, one can calculate all the terms needed for the regression problem using only vector products in the dimension of the original features (in this case 2) and solve it using eqs. 2.23–2.24 . This means that calculating and inverting the kernel matrix scales with the number training points as  $n^3$ , but does not scale with the dimension of our transformed feature space, as long as the kernel can be efficiently. These concepts lead to one of the best-developed and understood branches of machine learning: the study of reproducing kernel Hilbert spaces (RKHSs)<sup>271</sup>. Hilbert spaces are spaces of functions, and the theory underlying kernel methods is thus very general<sup>272,273</sup>, applying to problems posed in the space of

both functions and finite collections of transforms such as the quadratic polynomials and even infinite expansions of such basis functions.

Notice that the variable transform introduced above is characterized completely by its kernel function, eq. 2.30. This function defines data similarity as the inner product in the transformed space. Reversing this observation, if a metric to define similarity is available, the explicit transformation can be bypassed by selecting a kernel  $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} = \langle \varphi, \varphi \rangle$  without ever explicitly considering the transformation function  $x \rightarrow \varphi(x)$ . One natural question is whether any function can be the kernel, and the theoretical requirements for kernels are given in Mercer's theorem<sup>273</sup>. In general seek to formulate kernel functions that satisfy the following, slightly stricter criteria<sup>271</sup>:

1. The function must be non-negative and symmetric:  $k(x', x) = k(x, x') \geq 0$  for all inputs.
2. The kernel must be positive definite, that is  $\sum_{i=1}^n \sum_{j=1}^n k(x^{(i)}, x^{(j)}) c_i c_j \geq 0$  for all possible sets of  $n$  training data  $x^{(i)}, x^{(j)}$  and real coefficients  $c_i, c_j$

These requirements are a direct consequence of the kernel needing to correspond to an inner product in the transformed space. One can select a suitable function to be the definition of similarity, even if the corresponding feature transform,  $\varphi$ , is not known. This simplifies the machine learning task because determining the most suitable transform in high dimensions is challenging, whereas quantifying similarity between points can be more straightforward. To be concrete, general kernel ridge

regression (KRR) models can be formulated as:

$$\min_w \sum_{i=1}^n (y^{(i)} - y_{\text{KRR}}(x^{(i)}))^2 + \lambda \|w\|^2 \quad (2.32)$$

$$\text{where } y_{\text{KRR}}(x^{(i)}) = \sum_{j=1}^{d'} w_j \varphi_j(x^{(i)}) = \sum_{k=1}^n a_k k(x^{(i)}, x^{(k)}) \quad (2.33)$$

One can always express the optimal solution to the regression problem as a sum over  $n$  coefficients  $a$ , i.e., one per training point, instead of  $d'$  coefficients  $w$ , i.e., one per transformed dimension, even in the case where there are infinitely many basis functions. This result is known as the Representer Theorem<sup>272</sup> and is a direct consequence of the properties of Hilbert spaces\*. The key idea is that the minimizer of the training error can be expressed as a function of the training points only and an orthogonal component, and the presence of the regularization term in eq. 2.32 ensures that this orthogonal term increase the regularization penalty while not decreasing the loss function value.

Having outlined the theory of KRR, we now turn to the practical question of selecting kernel functions. The most obvious idea is to base kernels on the geometric distance between points in their feature space,  $\|x^{(i)} - x^{(j)}\|_2^2$ . This type of translation invariant kernel ensures that that only the relative similarity between inputs matters. However, using the raw distance would not make a reasonable kernel because it would assign nearly equal points no similarity ( $\|x^{(i)} - x^{(j)}\|_2^2 \approx 0$ ) while far away points would have large kernel terms. Therefore, a family of kernels can be obtained by taking the negative exponent of the pairwise distance, giving the Gaussian or

---

\*for a recent, general proof see Ref.<sup>272</sup>

radial basis function (RBF) kernel:

$$K_{i,j} = \exp\left(-\gamma\|x^{(i)} - x^{(j)}\|_2^2\right) = \exp\left(-\gamma\sum_{k=1}^d\left(x_k^{(i)} - x_k^{(j)}\right)^2\right) \quad (2.34)$$

$$y(x^*) = \sum_{i=1}^n a_i k(x^{(i)}, x^*)$$

$$k(x^{(i)}, x^*) = \langle x^{(i)}, x^* \rangle \quad k(x^{(i)}, x^*) = \exp\left(-\gamma\|x^{(i)} - x^*\|_2^2\right)$$

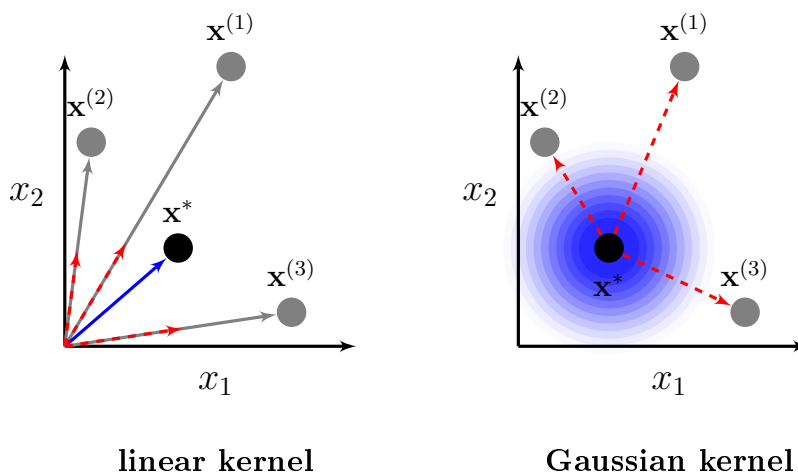


Figure 2-11: Comparison of linear and Gaussian kernel similarity in a function of two input dimensions  $(x_1, x_2)$  evaluated at a new point  $x^*$ : the linear kernel (left) considered the inner product between  $x^*$  and the training points  $x^{(1-3)}$  while the Gaussian kernel (right) considers a decaying exponential of the Euclidean distance between  $x^*$  and the training points  $x^{(1-3)}$ , illustrated as a fading blue region of influence. Kernel expressions and the overall regression equation are given above.

In comparison with the linear kernel function, pairwise similarity in this case is simply an exponentially decaying function of the distance between each point. The prediction of new points is taken as a linear function of these pairwise distance-based similarities (Figure 2-11). Note that this kernel depends on one hyperparameter,  $\gamma > 0$ , referred to as the inverse correlation length, which controls how quickly the

similarity function decays as the distance between points increases. Large values of  $\gamma$  correspond to rapid decay and very ‘local’ influence in feature space, while small values of  $\gamma$  correspond to long-range interactions.

It is useful to examine what nonlinear feature map,  $\varphi(X)$ , underlies the kernel in eq. 2.34. For this case, the underlying feature map is given by the Taylor series expansion of the exponential function, shown for  $x \in \mathbb{R}$  for simplicity:

$$\varphi(x) = e^{-\gamma x^2} \left[ 1 \quad \sqrt{\frac{2\gamma}{1!}}x \quad \sqrt{\frac{2^2\gamma^2}{2!}}x^2 \quad \sqrt{\frac{2^3\gamma^3}{3!}}x^3 \quad \dots \right]$$

Therefore, performing KRR with this kernel is equivalent to linear regression in this functional space. Many other choices of one-hyperparameter kernel function are possible (Table 2.1). Generally, the different kernels can be distinguished by how quickly they decay with respect to the distance between points, and it is difficult to infer which kernel to use *a-priori*. In practice, it is best to test a number of kernels and select the best through cross-validation. In practical chemical applications<sup>79,274</sup>, eq. 2.34 is often preferred.

Table 2.1: Some commonly used kernels their hyperparameters

Name	Expression
Gaussian/RBF	$k(x^{(i)}, x^{(j)}) = \exp\left(-\gamma \ x^{(i)} - x^{(j)}\ _2^2\right)$
Exponential	$k(x^{(i)}, x^{(j)}) = \exp\left(-\gamma \ x^{(i)} - x^{(j)}\ _2\right)$
Laplacian	$k(x^{(i)}, x^{(j)}) = \exp\left(-\gamma \ x^{(i)} - x^{(j)}\ _1\right)$
Matérn 3/2	$k(x^{(i)}, x^{(j)}) = \left(1 + \gamma \ x^{(i)} - x^{(j)}\ _2\right) \exp\left(-\gamma \ x^{(i)} - x^{(j)}\ _2\right)$

The primary difficulty and computational expense incurred when training KRR models is therefore rigorous hyperparameter selection. Selecting the wrong hyperparam-

eters can have a large impact on model performance. A typical supervised learning task using a one-parameter kernel from Table 2.1 involves selecting two hyperparameters, the kernel length scale,  $\gamma$ , and the regularization strength,  $\lambda$ . Both of these hyperparameters can be thought of in terms of tuning the smoothness of the function. As  $\gamma$  becomes large, the support of the kernel function around each data point becomes narrower and the fitting function sharper or higher frequency. Taking large values of  $\lambda$  biases towards smooth functions.

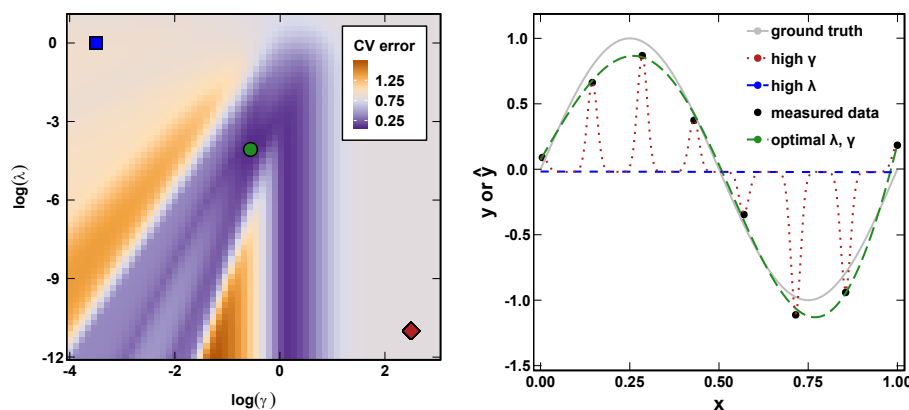


Figure 2-12: Comparison of fitting the function  $y = \sin(\pi x)$  based on 8 points uniformly sampled in  $[0, 1]$  with measurement noise  $\mathcal{N}(0, 0.15^2)$  using KRR with a Gaussian kernel. Hyperparameter estimation by 2D grid search (left) shows complicated dependence of the leave-one-out cross validation (CV) error as a function of  $\gamma$  and  $\lambda$ . The optimal values,  $\gamma = 0.28$  and  $\lambda = 8.6 \times 10^{-5}$  are shown as a green circle, while an example of a too-high value of  $\gamma = 10^{2.5}$  is shown as a red diamond while a too-high value of  $\lambda = 10$  is shown as blue square. The result on the predicted function for these hyperparameters is shown with lines in corresponding colors (right). Sampled points are shown in black, with the ground truth shown with a gray line.

These effects may be explored by considering what happens to eq. 2.34 as  $\gamma$  becomes very large. Then, all off-diagonal entries of the kernel matrix will decay to zero,



giving  $K \approx I$ . Applying this to eqs. 2.23–2.24 in limit of large  $\gamma$  gives:

$$y(X) = Ka = I_n(I_n + I_n\lambda)^{-1}y = \begin{bmatrix} \frac{y_1}{1+\lambda} \\ \vdots \\ \frac{y_n}{1+\lambda} \end{bmatrix}$$

Therefore, this will return the training data  $y$  exactly for small  $\lambda$ . It is also clear that including the regularization parameter will worsen our fit to training data. What about the predictions on a new point? Consider that the vector  $K(X, x^*) \approx 0$  for large enough  $\gamma$  if  $x^*$  is not in the training set. Then, we see that the prediction for out of sample points will be zero:

$$y(x^*) = K(X, x^*)a = \begin{bmatrix} K(x^{(1)}, x^*) \\ \vdots \\ K(x^{(n)}, x^*) \end{bmatrix} a = 0$$

If the out-of-sample data is scaled by the training data average, the net result will be that we will make the average training prediction,  $\bar{y}$ , for all out-of-sample points. For the large  $\lambda$  limit, one can neglect the kernel term in eqs. 2.23–2.24, leading to:

$$y(X) = Ka = K(I_n\lambda)^{-1}y = K \begin{bmatrix} \frac{y_1}{\lambda} \\ \vdots \\ \frac{y_n}{\lambda} \end{bmatrix} = 0$$

In this limit, the training predictions are zero or the average if the data has been

scaled. Since in this case, coefficients  $a$  are also zero, the predictions on any out-of-sample point will also be zero. Therefore, both hyperparameter limits give the same out-of-sample predictions but for distinct reasons (Figure 2-12). The high  $\gamma$  model is a corrugated surface, rising up to meet each training point and falling away just as quickly to zero in between them. The high  $\lambda$  model is flat everywhere and very smooth. Since the two hyperparameters are coupled, they must be chosen simultaneously. A standard approach would be to conduct a two-dimensional grid search cross-validation scheme, evaluating all combinations for a range of  $\lambda$  and  $\gamma$  values. KRR models have also seen widespread application to predictions of energies and various electronic and orbital properties<sup>79,212,251,275,276</sup>, electron densities<sup>277</sup> and catalytic properties for heterogeneous<sup>278</sup> and homogeneous<sup>279</sup> systems based on structural molecular descriptors. Many of these examples use relatively small numbers of training examples ( $\leq 10^4$  observations), especially compared to ANN models trained for similar purposes, which suggests KRR methods are a good choice for applications with more limited data. Another promising application of KRR methods is in ‘ $\Delta$ ’-learning<sup>137</sup>, where predictions at a cheap level of theory (for example, Hartree-Fock) are ‘corrected’ to match results from more expensive simulations (such as correlated wavefunction theory methods). Owing to their transparent dependence on pairwise distances in the feature space, KRR models retain a degree of interpretability that is lacking from, for example, neural networks. This property, as well as the relative speed of training and good accuracy, helps explain the popularity of kernel-based methods. Accuracy of kernel-based models may meet or exceed that of neural networks in some cases<sup>79,212,280</sup>, meaning they should not be overlooked based on the popularity of deep learning. While most applications use the Gaussian kernel, Laplacian<sup>79</sup> and Matern<sup>261</sup> kernels (as in Table 2.1) have been reported to offer superior performance in some applications. One drawback to KRR methods is that calculat-

ing the best-fit coefficients for eq. 2.24 involves inverting a matrix with dimension equal to the square of the number of training points, and even making predictions once trained requires calculating the kernel function between the new point and all training data. Thus, these methods do not scale well as the number training examples becomes large (though they have been routinely applied to  $\sim 10^4$  examples). Recent work<sup>281</sup> has attempted to address this issue by fragmenting organic molecules into local environments that generalize with fewer (100s–1000s) training examples.

## Gaussian process regression

Gaussian process regression (GPR) provides an alternative theoretical basis for preceding discussion of kernel methods. GPR produces similar predictive model forms but via distinct reasoning. Similar to KRR, the literature and interpretation of GPR is rich, and a detailed theoretical review is given in Ref.<sup>282</sup>. For completeness, GPR is briefly reviewed here in relation to the previous KRR framework.

Within the Gaussian process (GPs) framework, one models the data-generating process as a probabilistic relationship between inputs  $X$  and outputs  $y$ . The probability distribution of the output variable(s) can be determined based on the set of accumulated observations and conditioned on observing specific a new input,  $x^*$ . A Gaussian process is defined as a collection of random variables where any finite subset has a joint Gaussian distribution, which we assume to have a zero mean. Modeling the regression function as a stationary Gaussian process yields:

$$y_{\text{GP}}(x) \sim \mathcal{N}(0, k(x, x')) \quad (2.35)$$

The function value at a given position is related to the evaluation at all other inputs by a covariance function,  $k(x, x')$ . Covariance functions can be selected using the same criteria as for the kernels. While the Gaussian kernel is a popular choice, this is not required or implied for it to be a *Gaussian process* and specifying this covariance function completely defines the GP. When it comes to the regression task, one aims to determine the distribution around a given new point,  $x^*$ , based on the training data  $X$  and  $y$ . Their joint distribution may be written as follows<sup>282</sup>:

$$\begin{bmatrix} y_{\text{GP}}(X) \\ y_{\text{GP}}(x^*) \end{bmatrix} \sim \begin{bmatrix} K(X, X) & K(X, x^*) \\ K(x^*, X) & K(x^*, x^*) \end{bmatrix}$$

It is also standard to interpret the observations  $X$  as having some inherent noise  $\sigma^2$ . For such a case, one replaces the self-interaction kernel, that is the diagonal elements of the kernel matrix, with  $K(X, X) + I\sigma^2$ . The GP need not exactly interpolate the previous observations, thus improving numerical stability in the same manner as ridge regression in KRR. Through manipulation of the above expression for the conditional distribution, where  $K = K(X, X)$ , one can show:

$$p(y_{\text{GP}}(x^*)|X, y) = \mathcal{N}\left(K(K + I\sigma^2)^{-1}y, K(x^*, x^*) - K(X, x^*)(K + I\sigma^2)^{-1}K(X, x^*)\right) \tag{2.36}$$

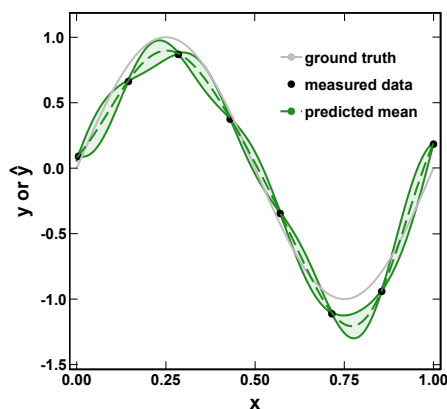


Figure 2-13: Fitting the function  $y = \sin(\pi x)$  based on 8 points uniformly sampled in  $[0, 1]$  with measurement noise  $\mathcal{N}(0, 0.15^2)$  using a GP with a Gaussian kernel and hyperparameters estimated as per Figure 2-12. The mean predicted value is shown as a dashed green line with the shaded region representing one standard deviation above and below the mean. Sampled points are shown in black with the ground truth shown with a gray line.

This expression enables prediction of the average and most likely value for the function at  $x^*$ , given by  $K(K + I\sigma^2)^{-1}y$ . Setting  $a = (K + I\sigma^2)^{-1}y$  and identifying  $\sigma^2 = \lambda$ , we also recover the original KRR expressions, eqs. (2.23–2.24). Thus, one obtains the same prediction for the same choice of kernel from the GP mean value, although with a slightly different interpretation.

The key advantage of the GPR approach is that eq. 2.36 gives not only a point value prediction but a distribution of possible values for the function at each new point, meaning that the GP comes with built-in estimates of system-specific uncertainty, with larger variance predicted for points with which the model is least certain. This predicted uncertainty, given by the variance in eq. 2.36, is a direct function of the pairwise distance between a new point and the training data (Figure 2-13). Note that this variance goes to zero if  $x^* = X$  and  $\sigma^2$  is small, meaning that the uncertainty at the training points goes zero (or at least to  $\sigma^2$ ).

This built-in uncertainty estimate makes GPs a popular tool for optimization or ex-

ploration problems where data-driven decisions need to be made about which new points to simulate.

Another advantage of the GP probabilistic framework is that the likelihood of observing the training data for a fixed set of hyperparameters can be explicitly computed. This gives another metric for how well the proposed model represents that data. Since this probability can be obtained as an analytic, differentiable function of the hyperparameter, one may choose hyperparameters by maximizing this probability directly rather than relying on the validation-set strategies discussed earlier. In practice, such an optimization typically suffers from multiple local minima, and so, if attempted, should be carried out multiple times with distinct initial guesses<sup>282</sup>.

Kernel models in the guise of GPs have also been widely applied to chemical applications. One notable example are so-called Gaussian approximation potentials (GAPs)<sup>78,255</sup>. These are a class of GPs that utilize a custom kernel similarity definition based on 3D geometry to recreate the potential energy surfaces of DFT simulations. GAPs have been used to create low-cost, accurate potentials for various bulk and atomic systems<sup>226,283–285</sup>, exploiting the built-in uncertainty estimates to decide when new simulations are needed for a particular atomic configuration. GPs have also been used to accelerate geometry optimizations directly<sup>286</sup>, to predict dispersion interactions<sup>287</sup>, correlation energies<sup>261</sup> and reaction outcomes<sup>288</sup>, and to interpolate between levels of theory for band-gap predictions<sup>236</sup>. Generally the GPR framework is invoked when it is desirable to use the ML model to recommend to data points to acquire, or to track model uncertainty.

## Neural Networks

Artificial neural networks are a type of regression and classification models with significant contemporary popularity but a long history, first developed in the 1960s<sup>289–291</sup>. Their simpler predecessor called the *perceptron*, was introduced<sup>292</sup> in the 1950s, and the development of the original, underlying theories<sup>293–295</sup> can be traced back to the 1940s. ANNs are so-named because they were initially conceptualized as a model of how neurons in the brain communicate by collecting input signals from other neurons and propagating this signal to other neurons. A detailed history of the development of ANNs is provided in Ref.<sup>289</sup>. d ANNs are best thought of as a flexible class of non-linear models that can have billions of parameters<sup>296</sup> and still remain efficient to train. There has been a renewed interest in both chemistry and beyond in ANNs models, driven in part by the excellent progress obtained in some benchmarks<sup>30</sup>. Alexnet<sup>297</sup>, an ANN developed in 2012, achieved state-of-the-art performance on the 1000-class ImageNet Large-Scale Visual Recognition Challenge<sup>298</sup>, beating the previous best known model by 60% and revolutionizing the field of computer vision. ANN models also exceeded the performance of other models for natural language processing with long short-term memory (LSTM) layers<sup>299,300</sup>. These advances were due to both algorithmic improvements, i.e. convolutional and long-term short-term memory (LTSM) layers, as well as increased computational power through the use of graphics processing units (GPUs) that made it possible for larger networks to be trained<sup>289,301</sup>. These developments in ANN architectures led to increasing interest in the chemistry community<sup>302</sup>. The simplest model of an ANN is the multi-layer perceptron. A multi-layer perceptron (MLP) is a series of transformations from one vector space to another. Each one of these transforms is called a neuron or node, and the basic model of a node is as follows<sup>303</sup>: the node calculates a weighted sum over

the input vector, combining elements of the feature space,  $x_i$  with unique weights,  $w_i$  (Figure 2-14). The output of an idealized neuron<sup>209</sup> is zero unless the total computed sum exceeds a certain threshold, at which point the node ‘turns on’ and propagates a signal downstream.

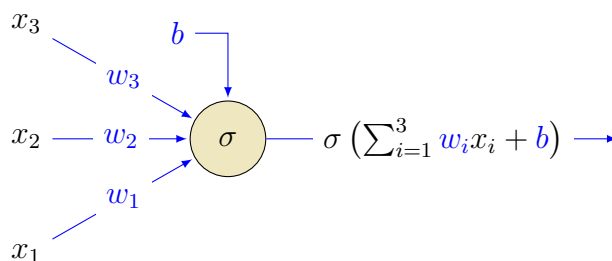


Figure 2-14: Diagram of a single neuron, showing a combination of inputs  $x_i$ , weighted by weights  $w_i$ , and a possible bias term  $b$ . The output of the neuron is given by passing the summation to the activation function,  $\sigma$ . Parameters of the model that can be learned to change the output are indicated in blue.

This non-linear response is governed by an activation function  $\sigma$ , which receives the weighted sum of inputs  $w_i x_i$  and a possible bias term  $b$  as inputs. Originally, non-differentiable step function were used to mimic biology<sup>209</sup>. These non-smooth functions were replaced with smooth approximations such as the sigmoid or hyperbolic tangent functions (Figure 2-15). Recently, these functions have been supplanted<sup>30</sup> by so-called rectified linear units (ReLUs)<sup>304</sup>, which have the following form:

$$\sigma\left(\sum w_i x_i + b\right) = \begin{cases} 0 & \sum w_i x_i + b \leq 0 \\ \sum w_i x_i + b & \sum w_i x_i + b > 0 \end{cases} \quad (2.37)$$



This shift to ReLU-like functions has been motivated mainly by the lower computational cost of evaluating the ReLU function and its derivatives, especially in modern ‘deep learning’ where a great many activation functions are used<sup>305</sup> simultaneously. Only the activation function can make the neural network become non-linear, and if a linear activation function with  $\sigma(x) \propto x$  were to be used, the entire MLP model would collapse to MLR. However, the incorporation of non-linearities in the activation function makes MLPs extremely flexible, powerful regression functions. Output of nodes is control by changing the input weights  $w_i$  and bias term  $b$  (blue labels in Figure 2-14). The derivative of the node output with respect to any of the weights can be obtained by:

$$\frac{\partial \sigma(\sum w_i x_i + b)}{\partial w_i} = x_i \sigma'(\sum w_i x_i + b) \quad (2.38)$$

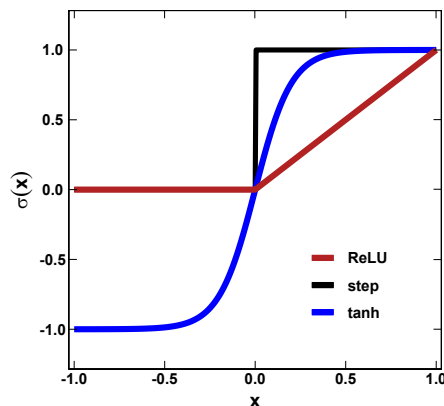


Figure 2-15: Activation functions for neural networks, showing initially-proposed step function (black) and approximations hyperbolic tangent (tanh, blue) and rectified linear (ReLU, red) functions.

However, the capacity of a single neuron are fairly limited, so many such nodes are

grouped together into a layer, and many such layers are grouped into a network (Figure 2-16). The input representation,  $I$ , is passed into multiple layers of nodes, called hidden layers because their output is not directly observed, denoted with an  $H$ . Networks with more than one hidden layer<sup>209</sup> are called deep neural networks (DNNs), although in practice this usually implies a large<sup>289</sup> number of layers. DNNs such as ResNet<sup>306</sup> have hundreds of hidden convolutional layers. Each hidden layer in simple MLP consists of multiple nodes that all receive the same input signals, although each neuron has its own vector of input weights, denoted  $w_{i,j}^{[l]}$  for the weight of input  $i$  to hidden node  $H_j^{[l+1]}$  at layer  $l + 1$ . To avoid confusion with indices of the vector, given by subscripts, and different training examples, indicated with superscript parenthesis,  $(i)$ , we use superscripts with brackets,  $[l]$ , to refer to the sequence of layers in the network. The output of each hidden node in layer  $(l - 1)$  is used as the input for all of the nodes in layer  $l$ . The MLP model is also called a *fully connected network*.

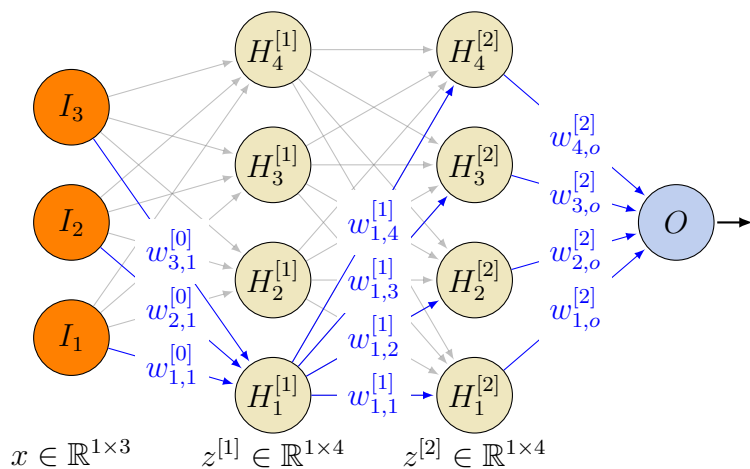


Figure 2-16: Overview of a multilayer perceptron model, showing how three input nodes,  $I$ , connect to two layers of four hidden nodes,  $H$ , and a scalar output node  $O$ . One possible pathway through the network passing through node  $H_2^{[1]}$ , and the associated learnable weights,  $w$ , is highlighted in blue. Other pathways through the other hidden nodes in layer one are shown in gray. The dimensions of the representation at each stage are shown below the nodes:  $x$  for the feature space and  $z^{[1]}$  and  $z^{[2]}$  for the two latent representations.

The collection of outputs from all the nodes in layer  $l$  the *latent state*,  $z^{[l]}$ . The latent state is a vector space of the dimension of the number of nodes in layer  $l$ , and the neural network as a whole can be understood as a mapping from the input,  $x$ , through a series of learned latent representations  $x \mapsto z^{[1]} \mapsto z^{[2]} \dots \mapsto y$ . The final output of the network is constructed from a linear combination of the outputs of the final layer, meaning that there is an additional set of weights that convert the the final latent space to the output ( $w_{i,o}^{[2]}$  in Figure 2-16). Because the relationship between the final latent space and the output is a linear one, inspecting how the latent representation of an input (i.e., molecule or data) varies for different inputs provides insight into how the ANN operates. Latent representations have been exploited for interpolation in the model space between molecules<sup>228</sup>, and to enrich data sets optimally<sup>307</sup>.

By combining a large number of nodes, highly non-linear models can be constructed. The number of parameters for a fully-connected network of  $L$  layers, each with  $N$  nodes, a single scalar output, and a  $d$ -dimensional input is  $\mathcal{O}(N^2L + dN)$ . The number of layers should in principal be selected based on fit to validation data, though in practice small changes to the number of nodes has limited effect. Typical implementations<sup>24,79,143,213,308,309</sup> of MLP in chemistry use 50–500 nodes per layer and  $\leq 3$  layers, corresponding to around  $10^4$ – $10^5$  parameters.

Multiple output nodes can be used to predict multiple properties from the same latent representation using a multi-task ANN. By predicting multiple properties simultaneously, for example dipole moments and energies<sup>253,310</sup>, multi-task ANNs increase utility and can sometimes show improved performance over multiple, single-task models as in the DeepTox<sup>311</sup> multi-task ANN which predicts binding affinity of drug-like molecules to 12 types of receptors.

Fortunately, the required derivatives for all internal weights can be obtained in a straightforward, analytical manner using the chain rule. The resulting expressions are simplest for the final layer and increase in complexity as the derivatives propagate back through the network. Thus, the gradient calculation should start at the final layer and move progressively back through the network, reusing already calculated terms. However, derivatives for the weights at layer  $l$  depend on the hidden state variables at early layers  $z^{[l-1]} \dots z^{[1]}$ . Therefore, during training, the states and outputs of the network are calculated in a forward pass by inputting the features of the training data to enable the loss function to be calculated. Then, the gradients of the loss function with respect to the weights are calculated in a backwards pass from the final layer to the input and used to update the weights. This process is known as back propagation and is critical to the efficiency of modern neural network training<sup>303</sup>.

In principle any numerical optimization procedure could be applied to update the weights of the ANN to minimize the loss function, but by far the most common<sup>30,303</sup> in practice is *stochastic minibatch gradient descent*. The basic idea is to update the weights of the network iteratively by taking a step of length  $\alpha$ , commonly known as the learning rate, in the negative direction of the loss function:

$$w^{[t+1]} = w^{[t]} - \alpha \sum_{i \in \text{batch}} \nabla \mathcal{L}_w(x_i, w^{[t]}) \quad (2.39)$$

The dependence on only  $\nabla \mathcal{L}$ , instead of second or higher derivatives, is important since higher derivatives are not as readily computed. This algorithm follows the gradient of the loss function in a descent direction,  $\nabla \mathcal{L}_w(X, w_i)$ , calculated at the current weights and summed over a randomly-sampled subset of the data in each step<sup>312,313</sup>, called a *batch*. The size of the batch can range from a single observation, called stochastic gradient descent (SGD), to the full data set, which is ordinary gradient descent. A full pass through all the batches is called one *epoch*.

The use of batches serves two purposes: firstly, loading all examples into memory can become a bottleneck in model training when using very large data sets or large-footprint inputs, for example high-resolution images<sup>297</sup>, which can be mitigated by using a series of smaller batches. Secondly, since the optimization problem is not convex<sup>303,312</sup>, gradient descent can only find local minima. Since the number of parameters in the model is often extremely large, and the model is highly nonlinear, the optimization landscape is generally very challenging and local minima and saddle points occur frequently. The number of local minima can be expected to grow exponentially with the dimension<sup>314</sup> and the number of saddle points may be even larger<sup>315</sup>. The use random batches introduces noise into the optimization procedure and can allow the model to escape locally minima. Generally, the larger the batch

size, the smoother the optimization and the faster it will converge but the more susceptible the procedure is to getting stuck in unproductive minima.

The selection of the learning rate,  $\alpha$ , can have a significant effect on the stability of the training process and the quality of the final model<sup>305</sup>. The learning rate is often decreased during training time<sup>313</sup> using *learning rate decay*, which helps the model optimization process travel quickly to productive regions in the initial steps and then fine-tune the estimates with smaller stepsizes later. While eq. 2.39 is the underlying method, modern optimizers for ANN training such as Adam<sup>316</sup> (a portmanteau of ‘adaptive moments’) and AdaDelta<sup>317</sup> make use of these and other enhancements to offer better performance in comparison to raw SGD.

Fully-connected layers are only one of the possible types of neural network models. Alternative layers leading to distinct architectures<sup>303</sup> have been proposed. Convolutional neural network (CNN) layers are a type of ANN layer that encodes spatial invariance and reduces the number of free parameters in the model. Originally proposed in analogy to the visual cortex in the 1980s in a model known as the neocognitron<sup>318</sup>, CNNs gained prominence in image recognition tasks<sup>319,320</sup>. CNNs have spurred some of the increased interest in deep learning from the early 2010s by powering record-setting image classification models such as AlexNet<sup>297</sup> and ResNet<sup>306</sup>. The CNN uses *convolutional filters* to extract predictive information from a raw input. A convolutional filter is comprised of a small receptive field that is repeatedly applied to different parts of the input, a process called convolution. In an image context, a block of the image at a time is fed into a small MLP, until the process has covered the entire image. The parameters of the filter are fixed throughout the translation across the image (i.e., the convolution process), and therefore each filter in a CNN applies the same operation to each sub-region. In the context of image classification, this filter, once trained, can be interpreted as a means of detecting

a particular input signature (e.g. a cat's ear) wherever it is located in the image. Multiple filters can be defined, which detect the presence different features. Figure 2-17 shows the application of a single convolutional filter to a 2D input. The receptive field is translated across the input image, producing a scalar output each time. Therefore, the output of each single feature is an array of reduced dimension compared to the input.

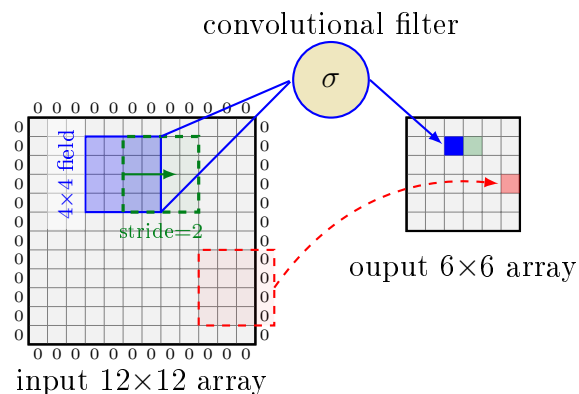


Figure 2-17: Illustration of a single 2D convolutional filter applied to an input array of dimension  $12 \times 12$  units with a zero-padding of one. The filter has a receptive field of  $4 \times 4$  units and stride of 2 units, illustrated being applied to the blue sub-region of the input array. The unit applies its activation function,  $\sigma$ , to this  $4 \times 4$  sub-region and stores the scalar result in the corresponding unit of the output array, illustrated with a blue square. Next, the receptive field will be translated to the right by the stride, and this is shown in with a dashed green region, whose output will occupy the green square in the output array. An example of what happens when the receptive field overhangs the edge of the array and utilizes the zero padding is shown by the red dashed region, with the corresponding output region shown with a red square.

Since CNNs explicitly encode translational invariance, they are most appropriate for processing 2D images or 3D coordinates where an object or feature's absolute location is not important to the model prediction. In three dimensions, voxels or

point-cloud data<sup>321</sup> are used in the place of pixels, and the filters are cubes instead of rectangles. Convolutions in 1D are also sometimes useful, for example if the dimension is time, they can be employed to detect a feature in the same way at all times, and have been applied in analyzing medical electrocardiography data<sup>322</sup>. In chemistry, convolutional networks have been applied to input images or 3D geometries, for example, protein docking<sup>323</sup>, energies of atoms in planar materials<sup>324</sup>, and the relationship between electron density and energies in orbital-free DFT<sup>143</sup> or exchange correlation energies<sup>262</sup>. Other examples include making predictions of molecular properties from images of 2D chemical skeleton structures<sup>325</sup> or, alternatively, converting images of 2D skeleton structures into machine-readable SMILES strings<sup>326</sup>.

One subfamily of CNNs that is particularly relevant to chemistry is graph convolutional neural networks (GCNNs)<sup>235</sup>. These are similar to standard CNNs, except that they use the connections in the molecular graph (i.e., chemical bonds) to define proximity rather than 2D or 3D space. While in an image convolution, nearby pixels are assessed by the same filter, in a GCNN, atoms that are connected or near each other in the molecular graph are treated together. GCNNs assign a ‘fingerprint’ vector to each atom<sup>235,327</sup> and this fingerprint is then updated by comparing it with all the atoms bonded to it, and the process is repeated multiple times. It is a convolution because the same operation is applied to all atoms in the molecule at each stage. This framework of iterative updates based on graph connectivity can be more broadly classified as a message passing neural networks (MPNNs)<sup>327</sup>, and a large number of varieties are possible. The ‘message’ is the information that is passed to each atom from its neighbors, a function of the neighbors’ own fingerprints. Information only propagates one bond away per iteration. For a molecule with a longest bond path of  $N$  atoms,  $N/2$  iterations of the convolutional operation are required for the most



distant atoms to see each other and share information. Some alternative formulations known as weave<sup>328</sup> or wave<sup>329</sup> networks attempt to address this shortcoming by passing information more non-locally. Many extensions, for example updating bond properties and atom properties at each step<sup>328,330</sup> or treating molecules as directed graphs in which each bond has a start and end atom<sup>81,331</sup> have been attempted. Crystalline systems can be handled by considering periodic versions of the molecular graph<sup>80</sup>. Geometric information can be included through edge features or incorporating additional bond types for non-covalent interactions<sup>332</sup>. Another related class of models uses 3D distances in place of graph bonding information, and so these models pass information between regions of 3D space centered on atoms instead but have similar update functions based on the state of their neighbors. Some examples of this approach are *continuous filter convolutions* in SchNet<sup>27,211</sup>, the Hierarchically Interacting Particle Neural Network (HIP-NN)<sup>26</sup> or the atoms-in-molecule network (AIMNet)<sup>310</sup>.

Applying these specialized convolutional filters on the molecular graph produces a set of atomic features for either direct use in property prediction, such as atomic energies<sup>26,27</sup>, or to compute a representative ‘molecule fingerprint’ from all of the atoms in the graph in conjunction with another neural network<sup>327</sup>. The molecular fingerprint can be used as an input to a final MLP to give some of the most accurate predictions of molecular properties<sup>28,81,310</sup> to date, or to predict reaction sites and outcomes<sup>229</sup>.

Neural networks also support some unique types of modeling that have garnered substantial interest<sup>333</sup>, such as is that of *generative modeling*. The principle goal of generative modeling is not to assign values to previously unknown input samples but rather to generate new inputs that are similar and yet still distinct from the original inputs<sup>303</sup> (Figure 2-18). The model should ‘imagine’ new samples that are

diverse and representative of the original data. In chemistry, this would correspond to the generation of new molecules, based on but distinct from existing sets that largely follow the chemical rules implicitly encoded in the source data. In theory, such models could enable significant augmentation of such databases.

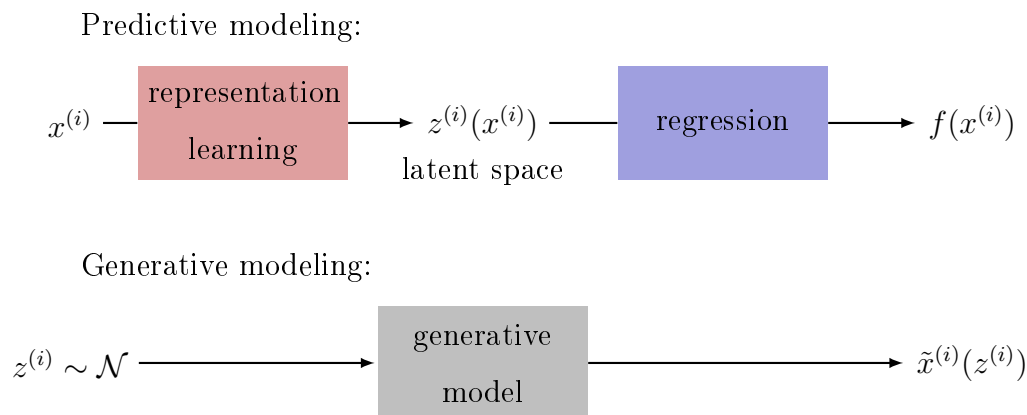


Figure 2-18: Comparison between predictive (top) and generative (bottom) modeling approaches. In predictive modeling, the learned function maps from the input  $x^{(i)}$  to the output  $f(x^{(i)})$  through the latent representation  $z^{(i)}$ . By contrast, generative modeling attempts to map a (usually Gaussian) distribution of vectors in latent space,  $z^{(i)} \sim \mathcal{N}$ , to a distribution of synthetic inputs  $\tilde{x}^{(i)}(z^{(i)})$ , such that these synthetic inputs should be indistinguishable from samples from the original data distribution.

There are two primary types of generative model: variational autoencoders (VAEs) and generative adversarial networks (GANs). These two types differ primarily in how the generative model is trained, but, once trained, both models function similarly. The trained generative model takes in samples from a random latent variable, which is almost always drawn from a Gaussian distribution<sup>303,334</sup>. The mean and covariance are learned during training in VAEs, but are typically set at the outset when using GANs. This use of random variables is essential to ensure that each time the model

is called it will generate distinct output. The generative model can be interpreted as a mapping between a distribution over the latent space to a distribution over the space of real input data. The sampled latent vectors are passed through a neural network model that outputs vectors that imitate the distribution of the ‘real’ data as closely as possible<sup>303</sup>, extracting syntactic meaning from variations in the latent vector and reconstructing real data (Figure 2-18).

VAEs are related to a simpler class of models called autoencoders. A standard autoencoder consist of an encoder, which is a neural network that maps the inputs to a latent vector, just like a standard ANN. However, this encoder is followed by another neural network that acts as a decoder and converts points from the latent space back into their input format<sup>303</sup>. During training, the loss function is set as the difference between the original data points and the reconstructed data  $\tilde{x}$ . This approach creates low-dimensional latent representation that can be used to interpret high-dimensional data more readily. Autoencoders have been used in chemistry to cluster molecular dynamics trajectories<sup>335,336</sup> and extract collective variables.

The VAE also consists of an encoder, but this encoder maps the inputs to parameters of the latent space distribution instead of the latent vectors directly. In practice, the mapping is to the mean and variance or covariance of a normal distribution, which sampled to generate latent points that are passed to the decoder. In order to train the model, real inputs are fed to the encoder network, and the resulting latent distribution is sampled and used to generate a distribution of synthetic data. This distribution is scored by how similar it is to the original distribution, and the gradient of this loss function can be used to update the network weights for both the encoder and generator. Care must be taken in selecting a suitable metric for measuring the similarity between the synthetic and original data distributions in order encourage a diverse synthetic distribution, as opposed to perfectly reconstructing the

input as in the standard autoencoder. The approach typically taken is to optimize the *variational lower bound* on the probability of observing the original data under the proposed distribution from the generator<sup>334,337</sup>. One advantage of autoencoders over GANs is that any known molecule can be injected into the latent space with the encoder, making it possible to identify paths between known molecules in the latent space<sup>338</sup> or to optimize molecular properties in latent space<sup>228</sup> and to then decode them to real molecules.

GANs are the other main type of generative model. GANs also use a decoder network, here, called a *generator*, which maps samples of the latent space to elements in the input or feature space. Unlike VAEs, there is no encoder stage, and the generator is not trained to reproduce the existing data. Instead GANs are trained simultaneously with a second network called the *discriminator*. The objective of the discriminator is to distinguish between the synthetic samples from the generator and samples from the original data<sup>339</sup>. The generator is trained to make the discriminator confuse its output with the real data while the discriminator is trained to correctly distinguish real and fake data. These two networks are trained in a stagewise, *adversarial* manner. This simple approach can generate very convincing synthetic results, for example realistic human faces using convolutional layers<sup>340</sup>. One issue with GANs is that the entire synthetic data distribution can concentrate on a limited set of outputs regardless of the latent space input. This is called ‘mode collapse’, and can be partly addressed by attempting to match the distribution of the synthetic data to the real data using the Wasserstein loss function<sup>341</sup>.

Small organic molecules have been generated either SMILES strings or as graphs with both VAEs<sup>342–345</sup> and GANs<sup>346–348</sup>. Both representations of molecular structures are challenging to generate in comparison to fixed size vectors or images because the size of the molecules can vary substantially. SMILES strings have been generated

sequentially, with the starting character generated first and then subsequent letters added to the string until a termination key is selected<sup>346</sup>, making use of a technique known as policy gradient optimization<sup>349</sup>. Due to the probabilistic nature of the process, taking the derivative of the so-called policy is not as simple as the deterministic derivatives, but an expectation or average gradient can be calculated using multiple samples. This approach was first demonstrated for text (in SeqGAN<sup>350</sup>) and then adapted to yield SMILES-generating models<sup>346,347,351</sup> in chemistry.

Generation of molecular graphs, as opposed to SMILES strings, can be realized ‘all-at-once’ by outputting a full graph<sup>343,352</sup> up to a maximum fixed size or by learning step-wise construction rules<sup>344,353,354</sup>, as with the sequential assembly of SMILES. One potential problem with generative models is the ease with which one can generate invalid structures. This issue can be resolved though enforcing grammar in SMILES strings<sup>355,356</sup>, introducing text-based SMILES alternatives that have simpler grammar<sup>357</sup>, or limiting the types of graphs that can be generated by only adding chemically-reasonable bonds<sup>353</sup>. Generative models are usually assessed based on the fraction of valid and unique generated molecules as well as the diversity of the generated molecules. In the case that targeted molecules have been generated, they could also be judged on the distribution of the molecules’ targeted physical properties such as molecular weight and solubility. Since generative modeling for chemistry is a relatively new field, it is not immediately apparent which, if any, of the presented strategies is best. In a comparison of SMILES-based GANs and SMILES- and graph-based VAE, similar performance was obtained for all models<sup>358</sup>. In a comparison between an all-at-once graph GAN and a sequence-based SMILES GAN, it was reported<sup>352</sup> that the graph approach resulted in higher validity and better ability to steer generation towards specific properties but at the cost of a much lower uniqueness score compared to SMILES.

## 2.3 Using surrogate models for chemical design

### Overview

Having reviewed both ‘physics-based’ virtual screening with first-principles methods in Section 2.1 and data-driven surrogates in Section 2.2, we now turn to the question of rational design or discovery. The problem of chemical design can be formulated as an optimization problem: from the space of all possible arrangements of atoms  $C_f$ , find candidates  $x_{\text{opt}}$  that have desirable properties, as scored by a given utility function  $f$  which judges the suitability of a proposed candidate:

$$x_{\text{opt}} = \arg \min_{x \in C_f} [-f(x)] \quad (2.40)$$

Some severe simplifications are necessary to address this problem:

1. The search of possible materials must be heavily constrained, by limiting the range of possible candidates to a finite and manageable subset,  $C_{f'} \subset C_f$ , which should still be sufficiently large to contain useful complexes. This is typically a specific set of molecules, selected based on chemical intuition – for example, fragments that are common in photoactive materials or pharmacologically relevant. For the purpose of this thesis, the search space will be restricted to transition metal complexes comprised of first-row transition metals and discrete collections of ligands based on those commonly used in inorganic chemistry.
2. The true utility function for a new molecule is incredibly complicated, involving a great number of dimensions – having the intended properties, stability, synthetic accessibility, cost and toxicity (especially for drug candidates<sup>3</sup>). Because of these challenges,  $f$  must also be restrictively approximated, usually to a

one-dimensional descriptor for high throughput virtual screening (HTVS) – for example, the excitation energy for photoemitters<sup>34</sup> or the energy of the d-band center in catalysis<sup>92</sup>. Additionally, rather than experimentally determining this property, the evaluation is carried out using either a first-principles method, or a surrogate thereof, in the interests of being able to evaluate a larger number of designs. This approximation introduces uncertainty into the real utility of any proposed candidates and therefore motivates identifying as many ‘good’ candidates as possible, in the hopes that some of them will have the expected property when evaluated at a higher level of theory or experimentally synthesized. This also immediately invokes the need for uncertainty quantification, requiring that the fidelity, both of the simulation to the physical system and the of the data-driven surrogate to the simulation, be estimated.

3. Often, solutions to eq. 2.40 will be found iteratively, that is candidates are proposed and evaluated, and then the results of this evaluation are used to select future candidates. In order to be a ‘rational strategy’, the iterative sampling proposed by the method must outperform random search<sup>359</sup>.

Computational molecular design strategies to address eq. 2.40 have been extensively reported in the literature<sup>1</sup>; thorough reviews are given in Refs<sup>359–361</sup>. In principle, any discrete optimization strategy could be employed to select candidates<sup>362</sup>. Where the model for  $f$  is affordable or the search space is small enough, direct HTVS has been applied to a large diversity of applications, from perovskites<sup>363</sup> and chromophores<sup>17</sup> to drugs<sup>87</sup>. Metropolis Monte Carlo strategies have also been employed for this discrete optimization, particularly in pharmacological applications, for example the virtual design of enzyme–inhibitor<sup>364</sup> and protein–ligand<sup>365</sup> interactions. However, all of these methods are restricted by the need to perform many property

evaluations, which become prohibitively expensive, particularly for the screening of quantum mechanical properties and when addressing large design spaces.

Use of data-driven surrogate models can greatly improve the ratio between the number of materials screened and the number of expensive simulations or experiments needed<sup>359,366</sup>, for example allowing for identification of photovoltaic materials from a design space of thousands using only hundred of first-principles simulations<sup>237</sup> or training artificial neural network (ANN) surrogate models to predict the CO adsorption energy of one hundred bimetallic surfaces and using it to explore a much larger space of candidates<sup>231</sup>. In drug discovery, the application of quantitative structure–activity relationship (QSAR) methods to screening is well established<sup>112</sup>, and applications<sup>367</sup> of deep neural network (DNN) and modern machine learning (ML) methods are rapidly proliferating<sup>368,369</sup>. In one study, antimicrobial peptides<sup>370</sup> were successfully identified from a design space of around 100k candidates with ANNs trained on only 1400 experimental observations.

These approaches are subsets of *multifidelity optimization*<sup>371</sup>, because they exploit a low-cost, less-reliable surrogate model in together with the more expensive, more accurate physics-based simulations or experiments<sup>372,373</sup>. The basic idea of these surrogate-based optimization routines<sup>373</sup> is to investigate initial candidates, use these candidates to parameterize a surrogate model, and then use this surrogate model to explore the design space and select new candidates to investigate (Figure 2-19).

This method is classified as model adaption<sup>371</sup> since information from the high fidelity model is used to correct the lower order model adaptively. A typical iteration of such a method involves solving the optimization problem, evaluating the high-fidelity model at the optimum, using this new evaluation to adapt the surrogate, and then repeating the process. The cheaper model is used to explore the space more completely and reduce the number of ‘bad’ calls to the expensive model. Since



the surrogate model is only an approximation to the truth, this creates a competing incentive to select new points not only on the basis of their estimated suitability in terms of eq. 2.40 (blue circles Figure 2-19), but also their ability to enrich the surrogate model (regions of low confidence in Figure 2-19) – active learning<sup>374</sup>, which has been demonstrated to drastically lower the amount of data needed to fit ML models to chemical data<sup>375</sup>. These two objectives define a trade-off between exploitation (i.e. improving properties) and exploration (i.e. improving surrogate confidence)<sup>372</sup>.

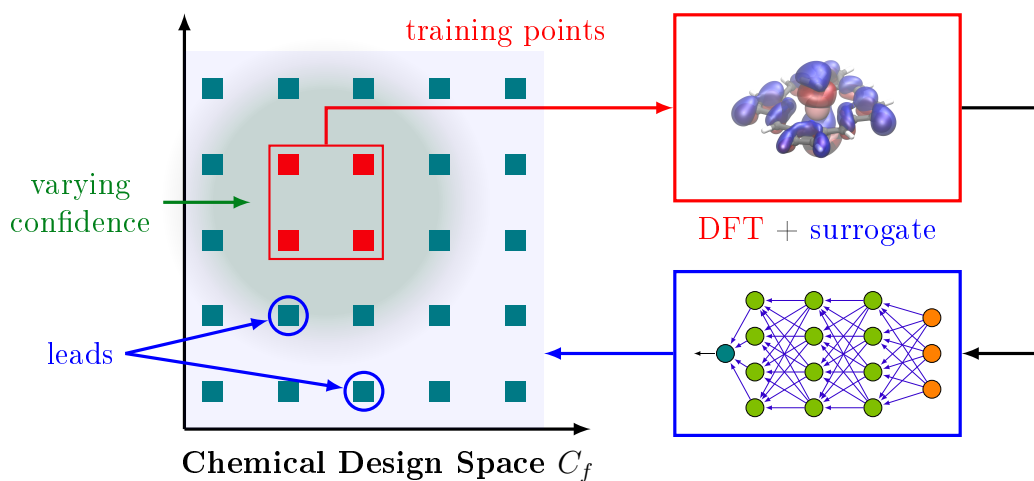


Figure 2-19: Schematic of surrogate-assisted virtual screening workflow representing a discrete chemical design space (squares). Some limited available training data (red) is screened using a first-principles method (DFT) and used to train a surrogate model (an ANN), which is used to search the space for new leads to simulate next (blue circles). Decaying model confidence far from the training data is illustrated by continuous green shading.

Many authors have highlighted the obvious synergy<sup>5,33</sup> between modern data-driven methods, computational chemistry and high-throughput synthesis, which has recently led to experimental discovery of several new materials. Some examples include

the most efficient, blue, organic light emitting compound<sup>34</sup>, discovered by screening a 1M complex space using an ANN and density functional theory (DFT), demonstrating an order-of-magnitude increase in the number of materials that can be assessed relative to the examples of real physical materials discovered from ‘pure’ HTVS presented in Section 2.1.1. Other real materials discovered using surrogate-assisted design processes and subsequently experimentally validated include novel magnetic materials<sup>37</sup> from nearly 300k candidates using regression based on quantum mechanics (QM) descriptors and metallic glasses<sup>35</sup> screened with random forest model initially based on  $\sim 7000$  experimental compositions.

### **Genetic algorithms**

Perhaps the most popular strategies for searching through chemical space are genetic algorithms (GAs). Genetic or evolutionary algorithms for molecular design all follow the same general principle. To begin with, a set of genes is assembled, where each gene codes one choice in the discrete design space. The quality of the candidates is assessed, and then the most fit genes are retained. In order to explore new combinations<sup>14</sup>, each generation the pool is enriched by randomly changing some of the genes (‘mutation’), meaning changing one of the connected functional groups, and by exchanging groups between candidates (‘reproduction’). These changes can be naturally mapped onto chemically-reasonable transformations, for example connection of molecular fragments<sup>14</sup> or functional groups<sup>376</sup>, or as text modifications directly to SMILES strings<sup>377</sup>.

First-principles-powered GAs have been applied to challenging virtual design problems, such as finding nearly 4000 fluorophores from a design space of over 1M candidates at the cost of 7500 DFT calculations<sup>14</sup> and reactive core-shell nanoparticle

catalysts of different sizes<sup>378</sup> were identified from spaces of thousands of candidates at much greater efficiency than by random search. While GAs can be based on first-principles methods only, they can naturally be accelerated by replacing expensive function evaluations with data-driven surrogate models<sup>379</sup>.

Some early examples of molecular design with GAs and simple data-driven models are to optimize the solubility of drug candidates<sup>380</sup> based on QSAR models and the polymer properties using group-additivity methods<sup>18</sup>, which both outperformed random search strategies. A GA with a Gaussian process regression (GPR) surrogate has recently been demonstrated<sup>381</sup> to provide 50-fold reduction in the number of required energy calculations (relative to a pure-DFT GA) to extract a convex hull over a space of  $10^{44}$  homotops of  $\text{Pt}_x\text{Au}_{147-x}$  alloys. This thesis will apply these ideas to aid design of spin crossover complexes (SCOs) with ANN surrogates in Chapter 5.

Another advantage of GAs is their extensibility to multiobjective optimization<sup>382,383</sup> and inherent parallelism. GAs have been suggested to be substantially simpler than, and provide competitive results with generative models for optimizing solubility<sup>377</sup>. However, genetic algorithms are limited in that they are localized and cannot explore chemical space more than one mutation from existing structures<sup>360,384</sup> – thus focusing on *exploitation* at the expense of *exploration*.

## Probabilistic and Bayesian approaches

An alternative approach to solve eq. 2.40 with surrogate models is to account directly for uncertainty in the design process by formulating a probabilistic model of utility of potential candidates,  $f(x)$ , conditioned on the previously observed values,

$f(X) = \{f(x^{(1)}), f(x^{(2)}) \dots\}$ . In this way, a distribution  $p(f(x)|F(X), X)$  is predicted instead of a point estimate. This approach can account for system-specific uncertainty naturally by having the distribution be more or less concentrated at different sample points  $x$ . The most common way to construct an estimate of such a distribution is by formulating a posterior distribution over functional values using Bayes' rule<sup>282</sup>. Recalling Section 2.2.3, the Gaussian process regression produces exactly such an estimate,

$$p(f(x)|X, f(X)) = \mathcal{N}(\mu(x), \sigma^2(x)) \quad (2.41)$$

where the predictive mean and variance have been abbreviated to  $\mu(x^*)$  and  $\sigma^2(x^*)$  respectively from full expressions in eq. 2.36. This construction of the surrogate model enables the use of global acquisition functions that choose new sample points in way that seeks out global (as opposed to local) minima. Efficient Global Optimization<sup>372</sup> (EGO) is staple of Bayesian optimization<sup>385</sup> that makes use of Gaussian processes (GPs) to balance optimally exploitation and exploration of the solution space during optimization. To illustrate the method, consider an iterative minimization problem, and denote the current minimum value of the utility function at the best input,  $x^*$ , of out currently observed inputs  $X$ , as  $f^* = f(x^*)$ . Then *improvement* can be defined as:

$$I(f) := I(f(x)) = \max(0, f^* - f(x)) \quad (2.42)$$

This can be combined with eq. 2.41 to calculate the probability that a new point will be better (i.e. lower) than  $f^*$ . Since the distribution is Gaussian, these integrals

can be solved analytically, yielding:

$$P[I] := P[I(f) > 0] = \int_{-\infty}^{\infty} \mathbb{1}_{I(f)} p(f|f(X), X) df \quad (2.43)$$

$$= \int_{-\infty}^{f^*} \frac{1}{\sqrt{2\pi\sigma^2(x)}} \exp\left(-\frac{(f - \mu(x))^2}{2\sigma^2(x)}\right) df \quad (2.44)$$

$$= \Phi\left[\frac{f^* - \mu(x)}{\sigma(x)}\right] \quad (2.45)$$

where  $\Phi$  is the cumulative standard normal distribution function and  $\mathbb{1}$  is an indicator function, i.e.  $\mathbb{1}_x = 1 \iff x > 0$ , else  $\mathbb{1}_x = 0$ . Eq. 2.45 is defined entirely as a function the posterior GP distribution for a given input  $x$ , and could be used to choose which points to simulate in the next iteration. However, this score is insensitive to the magnitude of the improvement, whereas in optimization contexts such as chemical design, a large improvement is much more desirable than a small one. The expected value of this improvement is therefore proposed instead<sup>372,386</sup>, and is defined as:

$$E[I] := \mathbb{E}(I(f)) = \int_{-\infty}^{\infty} I(f) p(f|f(X), X) df \quad (2.46)$$

$$= \int_{-\infty}^{f^*} [f^* - f(x)] \frac{1}{\sqrt{2\pi\sigma^2(x)}} \exp\left(-\frac{(f - \mu(x))^2}{2\sigma^2(x)}\right) df \quad (2.47)$$

$$= (f^* - \mu(x)) \Phi\left(\frac{f^* - \mu(x)}{\sigma(x)}\right) + \sigma(x) \phi\left(\frac{y^* - \hat{y}(x)}{\sigma(x)}\right) \quad (2.48)$$

where  $\phi$  is the standard normal distribution. The first term in eq. 2.48 corresponds to exploitation of the model – i.e. seeking to sample points with the greatest potential gains – while the second term favors exploration, i.e. seeking out regions of the space with high variance, which corresponds in practice to regions that are far from the previously sampled values. Using this expression as the objective for surrogate optimization encourages the model to visit new regions of the design space. In

molecular discovery applications, expected improvement strategies have been used to search for low-energy conformers<sup>387</sup>, stable perovskites<sup>388</sup> and crystals with targeted melting temperatures<sup>389</sup> in an uncertainty-aware manner.

A drawback of GP regression is that the complexity is cubic in the number of data evaluations, as opposed to ANN models which have linear complexity<sup>390</sup>; however, relative to the cost of DFT, which might take hours, inverting a matrix corresponding to hundreds or thousands of observations may not be a substantial computational cost. efficient global optimization (EGO) is also a sequential approach as described, though modifications to generate parallel samples have been proposed<sup>391</sup>. The *knowledge gradient*<sup>392,393</sup> is a more recent acquisition function in the spirit of eq. 2.48 that seeks to maximize information content of the surrogate while accounting for cost explicitly. Expected improvement techniques can also be extended to treat multi-objective optimization<sup>394</sup>, and these will be applied to the problem of redox flow batteries (RFB) design in Chapter 7.

General alternatives to expected improvement have also been proposed, most notably Thompson sampling<sup>395</sup>. Thompson sampling can be used to conduct optimization in the presence of model uncertainty by sampling one realization of the model parameters and optimizing under this belief, then repeating this process many times. It has a stronger focus on exploitation than expected improvement, is naturally parallel and does not depend on the structure of the posterior (i.e. non-Gaussian posteriors). Thompson sampling has been coupled to Bayesian neural networks and applied to chemical discovery problems, showing greater sample efficiency than EGO on some benchmark datasets<sup>396</sup>. The Phoenix optimizer<sup>397</sup> is an acquisition function developed especially for planning chemical discovery experiments that also claims superior sample efficiency and stability compared to vanilla EGO.

## Continuous and generative approaches

Not all approaches to solve eq. 2.40 operate in discrete chemical space. Some problems can be cast naturally in terms of continuous variables, for example the design of lithography targets<sup>398,399</sup>, or optimization of reaction conditions<sup>400</sup>. Converting the discrete chemical optimization problem into optimization over a continuous vector space has obvious advantages in terms of optimization, being amenable to gradient-based optimization machinery, and has therefore been attempted in various guises over many decades. An early inverse design idea<sup>401</sup> was based on finding a wavefunction with intended properties, back-calculating a corresponding Hamiltonian and then inverting back to molecule structure, but this was restricted to very simple systems. Another approach<sup>402</sup> demonstrated a shell-wise method for constructing molecular catalysts by optimizing the number and position of nuclear charges using gradient-based optimization, but reported difficulties in generating in inverting these positions back to chemical space.

A similar proposal is the linear combination of atomic potentials (LCAP) method<sup>403</sup>, which proposes casting the search in chemical space as continuous optimization over nuclear potential functions,  $V(\mathbf{r})$ . Every molecular system generates a  $V(\mathbf{r})$  based on atomic types and positions, which uniquely defines all properties of the system. However, the relationship is not surjective; it is not possible to find a unique set of nuclei and positions that generate a given  $V(\mathbf{r})$ . To get around this, the method fixes possible positions for nuclei, and then considers continuous variations in the potential field from different atom types at these locations. Setting all coefficients except one to zero at a given nuclear position allows recovery of real systems, while other values of the coefficients correspond to non-physical hybrid systems that interpolate between the potential maps of physical systems. This approach allows for

fully continuous, differentiable optimization in the space of chemical configurations, and was successfully demonstrated for a two-site system<sup>403</sup>. However, when expanding to more complicated systems, the method failed due to extremely non-smooth property response behaviors being observed for the hybrid potential fields – to the extent that moving in the interpolated space is infeasible<sup>404</sup>. Instead, methods have been proposed that utilize the gradients to inform search directions in a Monte Carlo search of the discrete space<sup>405</sup>, and these methods have been demonstrated for designing non-linear optical materials from fixed libraries with hundreds of thousands of candidates<sup>21,404</sup>.

In all of these cases, the primary difficulty is mapping the continuous variable back to a real system after the optimization is complete. Deep generative models, as described in Section 2.2.3 (Page 98), have emerged as a promising way of addressing this challenge<sup>333</sup>. Additionally, generative models have the unique potential to dispense with the need to create a human-engineered design space and are not limited to the extents of existing databases or motifs. variational autoencoders (VAEs) are by their nature able to invert from an arbitrary coordinates in latent space to representation of chemistry, and optimization in their learned latent space has been shown<sup>228</sup> to be able to design novel molecules with targeted properties such as solubility and synthetic accessibility (as judged by quantitative structure–property relationship (QSPR) methods). In the case of generative adversarial networks (GANs), it is possible to ‘steer’ the generation process toward specific regions of chemical space by designing a ‘reward function’ that provides a higher score to generated samples that have a desired property, as has been demonstrated by generating drug-like molecular graphs<sup>352</sup> and SMILES<sup>346</sup>. However, generative models to date have only been applied to generate organic molecules, where large databases already exist, and make extensive use of tools such as RDKit<sup>76</sup> for property prediction and validation.



# Chapter 3

## Surrogate models for transition metal complexes

Note: This chapter was originally published as “Janet, J. P., Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* 2017, 8, 5137–5152” and has been formatted for consistency. Supporting information provided with the manuscript, available online at <http://www.doi.org/10.1039/C7SC01247K>, has been placed in Appendix A.

### Chapter summary

Direct density functional theory (DFT) simulation of inorganic materials and molecular transition metal complexes is often used to describe subtle trends in inorganic bonding and spin-state ordering, but these calculations are computationally costly and properties are sensitive to the exchange–correlation functional employed. This chapter begins to overcome these challenges by using artificial neural networks (ANNs) to predict quantum-mechanically-derived properties, including spin-state ordering, sensitivity to Hartree–Fock exchange, and spin-state specific bond length variation. The ANNs are trained on a small set of heuristic inorganic-chemistry-appropriate empirical inputs that are both maximally transferable and do not require three-dimensional structural information for prediction. Using these descriptors, our ANN predicts spin-state splittings of single-site transition metal complexes (i.e.,

Cr–Ni) at arbitrary amounts of Hartree–Fock exchange to within 3 kcal/mol accuracy of DFT calculations. Our exchange-sensitivity ANN enables improved predictions on a diverse test set of experimentally-characterized transition metal complexes by extrapolation from semi-local DFT to hybrid DFT. The ANN also outperforms other machine learning models (i.e., support vector regression and kernel ridge regression), demonstrating particularly improved performance in transferability, as measured by prediction errors on the diverse test set. The ability to generalize to diverse complexes from experimental databases is examined and heuristics are developed to identify when a compound of interest is likely to be poorly predicted by the ANN. The ANNs developed in this chapter provide foundation for data-driven models for screening transition metal complexes.

### 3.1 Introduction

High-throughput computational screening has become a leading component of the workflow for identifying new molecules<sup>34,406</sup>, catalysts<sup>6</sup>, and materials<sup>108</sup>. First-principles simulation remains critical to many screening and discovery studies, but relatively high computational cost of direct simulation limits exploration of chemical space to a small fraction of feasible compounds<sup>22,90</sup>. In order to accelerate discovery, lower levels of theory, including machine-learning models, have emerged as alternate approaches for efficient evaluation of new candidate materials<sup>407</sup>. Artificial neural networks (ANNs) have recently found wide application in the computational chemistry community<sup>254,408,409</sup>. Machine learning approaches were initially appreciated for their flexibility to fit potential energy surfaces and thus force field models<sup>24,213,214,410–413</sup>. Broader applications have recently been explored, including in exchange-correlation functional development<sup>260,408</sup>, general solutions to the Schrödinger equation<sup>414</sup>, orbital free density functional theory<sup>143,415</sup>, many body expansions<sup>416</sup>, acceleration of dynamics<sup>417–419</sup>, band-gap prediction<sup>236,420</sup>, and molecular<sup>34,406</sup> or heterogeneous catalyst<sup>231</sup> and materials<sup>421–424</sup> discovery, to name a few.

Essential challenges for ANNs to replace direct calculation by first-principles methods include the appropriate determination of broadly applicable descriptors that enable the use of the ANN flexibly beyond molecules in the training set, e.g. for larger molecules or for those with diverse chemistry. Indeed, the most successful applications of ANNs at this time beyond proof-of-concept demonstration in replacement of direct first-principles simulation have been in the development of force fields for well-defined compositions, e.g. of water<sup>215,425</sup>. Within organic chemistry, structural descriptors such as a Coulomb matrix<sup>426</sup> or local descriptions of the chemical environment and bonding<sup>223,427</sup> have been useful to enable predictions of energetics as long as a relatively narrow range of compositions is considered (e.g., C, H, N, O compounds). These observations are consistent with previous successes in cheminformatics for evaluating molecular similarity<sup>428</sup>, force field development<sup>429</sup>, quantitative structure-activity relationships<sup>430</sup>, and group additivity<sup>431</sup> theories. For transition metal complexes, few force fields have been established that can capture a full range of inorganic chemical bonding<sup>432</sup>, and the spin-state- and coordination-environment-dependence of bonding<sup>173</sup> suggests that more careful development of descriptors is required to broadly predict properties of open-shell transition metal complexes. Similarly, descriptors that worked well for organic molecules have been demonstrated to not be suitable in inorganic crystalline materials<sup>433</sup>. It is well-known<sup>197,434,435</sup> that there is a strong relationship between sensitivity of electronic properties (e.g., spin-state splitting) and the direct ligand-atom and ligand field strength<sup>436,437</sup> in transition-metal complexes. Since ligands with the same direct metal-bonding atom can have substantially different ligand-field strengths (e.g., C for both weaker field CH<sub>3</sub>CN versus strong-field CO), whereas distant substitutions (e.g., tetraphenylporphyrin vs. base porphine) will have a limited effect, a transition-metal complex descriptor set that carefully balances metal-proximal and metal-distant descriptors

is needed.

Within transition metal chemistry and correlated, inorganic materials, a second concern arises for the development of ANN predictions of first-principles properties. Although efficient correlated wavefunction theory methods (e.g., MP2) may be straightforwardly applied to small organic molecules, such methods are not appropriate for transition metal complexes where best practices remain an open question<sup>438</sup>. Although promising avenues for ANNs include the mapping of lower-level theory results, e.g. from semi-empirical theory<sup>439</sup>, to a higher-level one, as has been demonstrated on atomization energies<sup>137</sup> and more recently reaction barriers<sup>440</sup>, suitable levels of theory for extrapolation are less clear in transition metal chemistry.

Additionally, uncertainty remains about the amount of exact (Hartree-Fock, HF) exchange to include in study of transition metal complexes, with recommendations ranging from no exchange, despite disproportionate delocalization errors in approximate DFT on transition metal complexes<sup>155,436,441</sup>, to alternately low<sup>59,198,199</sup> or high<sup>196</sup> amounts of exact exchange in a system-dependent manner. Indeed, there has been much interest recently in quantifying uncertainty with respect to functional choice in energetic predictions<sup>442-444</sup>, including through evaluation of sensitivity of predictions with respect to inclusion of exact exchange<sup>196,197</sup>. Spin-state splitting is particularly sensitive to exchange fraction<sup>197,434,435</sup>, making it a representative quantity for which it is useful to both obtain a direct value and its sensitivity to varying the exchange fraction. Thus, a machine-learning model that predicts spin-state ordering across exchange values will be useful for translating literature predictions or providing sensitivity measures on computed data.

Overall, a demonstration of ANNs in inorganic chemistry, e.g. for efficient discovery of new spin-crossover complexes<sup>60,445</sup> 63-64, for dye-sensitizers in solar cells<sup>1665</sup>, or for identification of reactivity of open-shell catalysts<sup>58</sup> via rapid evaluation of spin-

state ordering should satisfy two criteria: i) contain flexible descriptors that balance metal-proximal and metal-distant features and ii) be able to predict spin-state ordering across exchange-correlation mixing. In this work, we make progress toward both of these aims, harnessing cheminformatics-inspired transition metal-complex structure generation tools<sup>181</sup> and established structure-functional sensitivity relationships in transition metal complexes<sup>196,437</sup> to train ANNs for transition metal complex property prediction.

The outline of the rest of this work is as follows. In Section 3.2, we review the computational details of data set generation, we discuss our variable selection procedure, and we review details of the artificial neural network trained. In Section 3.3, we provide the results and discussion on the trained neural networks for spin-state ordering, spin-state exchange sensitivity, and bond-length prediction on both training-set-representative complexes and diverse experimental complexes. Finally, in Section 3.4, we provide our conclusions.

## 3.2 Methods

### 3.2.1 Test set construction and simulation details

#### Data set construction

Our training set consists of octahedral complexes of first-row transition metals in common oxidation states:  $\text{Cr}^{2+/3+}$ ,  $\text{Mn}^{2+/3+}$ ,  $\text{Fe}^{2+/3+}$ ,  $\text{Co}^{2+/3+}$ , and  $\text{Ni}^{2+}$ . High-spin (H) and low-spin (L) multiplicities were selected for each metal from the ground, high-spin state of the isolated atom and the higher-energy, lowest-spin state within 5 eV that had a consistent d-orbital occupation for both states, as obtained from the National Institute of Standards and Technology atomic spectra database<sup>446</sup>. The

selected H-L states were: triplet-singlet for  $\text{Ni}^{2+}$ , quartet-doublet for  $\text{Co}^{2+}$  and  $\text{Cr}^{3+}$ , quintet-singlet for  $\text{Fe}^{2+}$  and  $\text{Co}^{3+}$ , quintet-triplet for  $\text{Cr}^{2+}$  and  $\text{Mn}^{3+}$ , and sextet-doublet for  $\text{Mn}^{2+}$  and  $\text{Fe}^{3+}$ .

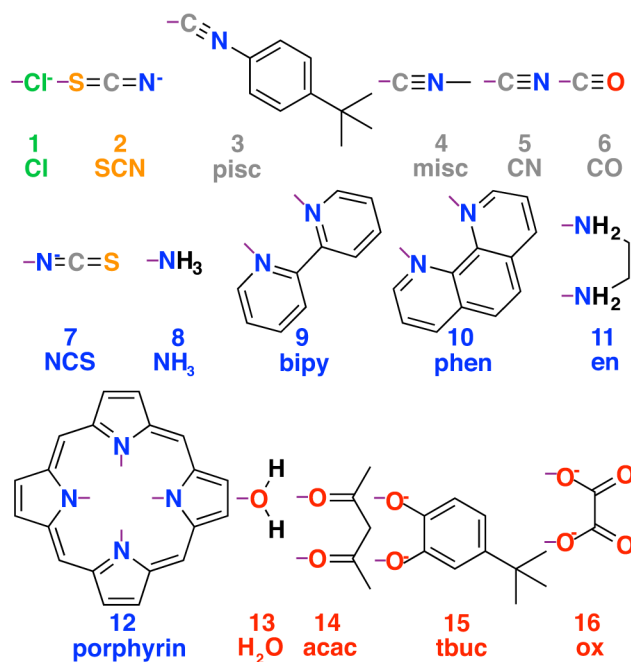


Figure 3-1: Set of ligands used to generate the transition metal complex data set. Ligands are numbered 1-16 and colored according to the atom type that coordinates with the metal, with chlorine in green, carbon in gray, sulfur in orange, nitrogen in blue, and oxygen in red. Purple lines indicate the bonds formed to metal-coordinating atoms in the ligand complexes. Abbreviations for each ligand used in the text are also shown. Full chemical names are provided in Appendix A, Table A.3.

A set of common ligands in inorganic chemistry was chosen for variability in denticity, rigidity, and size (nine monodentate, six bidentate, and one tetradentate in Figure 3-1 and Appendix A, Table A.1). These ligands span the spectrochemical series from weak-field chloride (1,  $\text{Cl}^-$ ) to strong-field carbonyl (6,  $\text{CO}$ ) along with representative intermediate-field ligands and connecting atoms, including S (2,  $\text{SCN}^-$ ), N (e.g.,

9,  $\text{NH}_3$ ), and O (e.g., 14, acac). All possible homoleptic structures with all metals/oxidation states were generated from ten of these ligands (90 molecules) using the molSimplify toolkit<sup>181</sup> (Appendix A, Table A.2). Additional heteroleptic complexes (114 molecules) were generated with molSimplify with one mono- or bidentate axial ligand type ( $L_{\text{ax}}$ ) and an equatorial ligand type ( $L_{\text{eq}}$ ) of compatible denticity (ligands shown in Figure 3-1, schematic shown in Figure 3-1). We also selected 35 molecules from the Cambridge Structural Database<sup>184</sup> (Appendix A, Table A.3).

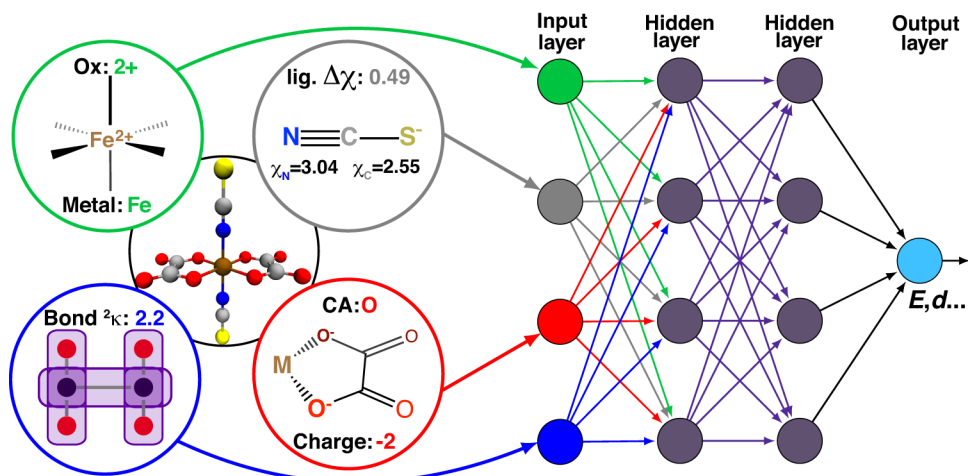


Figure 3-2: Schematic diagram of descriptors (left) as inputs to the ANN (right), along with hidden layers, and output (e.g., spin-state splittings) layers with additive bias term in each node omitted.

### First-principles geometry optimizations

DFT gas-phase geometry optimizations were carried out using TeraChem<sup>105,447</sup>. DFT calculations employ the B3LYP hybrid functional<sup>152,157,158</sup> with 20% Hartree-Fock (HF) exchange ( $a_{\text{HF}} = 0.20$ ) and a variant45 ( $a_{\text{HF}} = 0.00$  to 0.30 in 0.05 increments) that holds the semi-local DFT portion of exchange in a constant ratio. We

calculate and predict sensitivities with respect to HF exchange,  $\frac{\partial \Delta_{\text{H-L}}}{\partial a_{\text{HF}}}$ , as approximated from linear fits, in units of kcal/mol.HFX<sup>-1</sup>, where 1 HFX corresponds to varying from 0% to 100% HF exchange. B3LYP<sup>152,157,158</sup> is chosen here due to its widespread use and our prior experience<sup>45</sup> with tuning it to study HF exchange sensitivity, where we also observed<sup>45</sup> similar behavior with other GGA hybrids, e.g. PBE0, as long as the same HF exchange fraction was compared.

The composite basis set used consists of the LANL2DZ effective core potential<sup>172</sup> for transition metals and the 6-31G\* basis for the remaining atoms. All calculations are spin-unrestricted with virtual and open-shell orbitals level-shifted<sup>448</sup> by 1.0 and 0.1 eV, respectively, to aid self-consistent field (SCF) convergence to an unrestricted solution.

For all training and test case geometry optimizations, default tolerances of 10<sup>-6</sup> hartree for SCF energy changes between steps and a maximum gradient of 4.5 × 10<sup>-4</sup> hartree/bohr were employed, as implemented in the DL-FIND interface<sup>449</sup> with TeraChem (Appendix A, Table A.4). Entropic and solvent effects that enable comparison to experimental spin-state splittings have been omitted, and we instead evaluate the DFT adiabatic electronic spin state splitting, as in previous work because our goal is to predict DFT properties and sensitivity to functional choice<sup>194,197</sup>. In high-throughput screening efforts ongoing in our lab, entropic and solvent effects that influence catalytic and redox properties will be considered in the future.

For each molecular structure (90 homoleptic, 114 heteroleptic) 14 geometry optimizations were carried out at 7 exchange fractions (from 0.00 to 0.30) and in high- or low- spin, for a theoretical maximum of 2856 geometry optimizations. In practice, 166 structures were excluded due to i) large spin contamination, as defined by an expectation value of  $\langle S^2 \rangle$  that deviated more than 1  $\mu\text{B}$  from the exact value (< 1%, 26 of 2856, see Appendix A, Table A.5), ii) dissociation in one or both spin states,



especially of negatively charged ligands, leading to loss of octahedral coordination (4%, 126 of 2856, see Appendix A, Table A.6), or iii) challenges associated with obtaining a stable minimized geometry ( $< 1\%$ , 14 of 2856, see Appendix A, Table A.2). Eliminating these cases produced a final data set of 2,690 geometry optimizations (Appendix A, Text A.1). Although these excluded cases are a fraction of our original data set, they highlight considerations for application of the ANN in high-throughput screening: highly negatively charged complexes should be avoided, and single point DFT calculations should be used to confirm that a high-fitness complex does not suffer from large  $\langle S^2 \rangle$  deviations.

### 3.2.2 Descriptor selection

High-throughput screening of transition-metal complex properties with direct prediction from an ANN requires mapping of an empirical feature space that represents the complex,  $\mathcal{X}$ , to quantum-mechanical predictions. This feature space should be balanced to avoid i) too few descriptors with insufficient predictive capability or ii) too many descriptors that lead to over-fitting of the ANN. Molecular descriptors<sup>384</sup> that have been used for parameterizing chemical space include: atomic composition, electronegativity<sup>427</sup>, formal charges, and representations of the geometric structure. This last class of descriptors may be divided into those that depend either on 3D structural information<sup>25,143,214,450,451</sup> or on graph-theoretic connectivity maps<sup>452</sup> (e.g., the Randić<sup>239</sup>, Wiener shape<sup>240</sup>, or Kier<sup>453</sup> indices). Graph-theoretic methods are preferable to 3D structural information to avoid sensitivity to translation/rotation or molecule size<sup>250</sup>, though we note that subsystem descriptors<sup>214,451,454</sup> and element-specific pairwise potentials<sup>25,250</sup> have been employed successfully to overcome some challenges. A secondary reason to avoid use of 3D structural information is the im-

explicit requirement of equilibrium geometries obtained from a geometry optimization, which are readily achieved with semi-empirical methods on small organic molecules<sup>452</sup> but would be prohibitive and error-prone for transition metal complexes.

We use  $L_1$ -regularized, least absolute shrinkage and selection operator (LASSO) linear least-squares regression<sup>209</sup>, as implemented in the `glmnet`<sup>455</sup> package in R3.2.5<sup>456</sup>, to evaluate candidate descriptor sets. LASSO is used to reduce over-fitting, force the coefficients of the least-powerful indicators to zero, and avoid monotonic decrease of model error as feature space dimensionality increases. Given observed input-output pairs  $(x_i, y_i)$  for  $i = 1, 2, \dots, n$  with  $x \in \mathcal{X} \subset \mathbb{R}^m$  assembled as rows of data matrix  $X$  and  $\lambda \in \mathbb{R}^+$ , the output is modeled as:

$$y_{pred} = \beta^T X + \beta_0 \mathbb{1} \quad (3.1)$$

for  $\{\beta, \beta_0\} \in \mathbb{R}^m \times \mathbb{R}$ , where

$$\{\beta, \beta_0\} = \arg \min_{\beta, \beta_0} \left( \|y - \beta^T X - \beta_0 \mathbb{1}\|_2^2 + \lambda \sum_{i=1}^m |\beta_i| \right) \quad (3.2)$$

The parameter  $\lambda$  is selected by ten-fold cross-validation with values typically between  $10^{-1}$  and  $10^{-6}$ . Our descriptors include both continuous variables that are normalized and discrete variables that are described by zero-one binary coding (Appendix A, Table A.7). Metal identity represents a descriptor best described by a set of discrete variables: 4 binary variables are chosen to represent Cr, Mn, Fe, and Ni, and Co corresponds to the case where all 4 variables are zero. This leads to a higher number of overall variables than for continuous descriptors (see Table 3.1).

Based on previous observations<sup>197,437</sup> we hypothesize that spin-state ordering is predominantly determined by the immediate chemical environment around the metal

center, potentially enabling predictive descriptors that are widely transferable across a range of molecule sizes. We compare 7 descriptor sets on the data and select the subset of descriptors that give the best simultaneous predictive performance for spin-state splitting,  $\Delta E_{\text{H-L}}$ , and its sensitivity with respect to HF exchange variation,  $\frac{\partial \Delta E_{\text{H-L}}}{\partial \alpha_{\text{HF}}}$ , as indicated by the prediction root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_{i,\text{pred}} - y_i)^2} \quad (3.3)$$

When two variable sets perform comparably, we select the variable set that will enable broader application of the ANN. All sets include the metal identity as a discrete variable and metal oxidation state, ligand formal charge, and ligand denticity as continuous variables (Figure 3-3, some descriptors shown in Figure 3-2).

Table 3.1: Comparison of variable sets by root-mean-squared errors (RMSE) after regularization in  $\Delta E_{\text{H-L}}$  and  $\frac{\partial \Delta E_{\text{H-L}}}{\partial \alpha_{\text{HF}}}$  prediction along with number of discrete variables (with all binary levels of the discrete variables counted in parentheses) and the number of continuous variables.

set	RMSE( $\Delta E_{\text{H-L}}$ ) (kcal/mol)	RMSE( $\frac{\partial \Delta E_{\text{H-L}}}{\partial \alpha_{\text{HF}}}$ ) (kcal/mol.HFX )	Discrete variables	Continuous variables
a	14.6	20.6	3 (37)	6
b	15.1	21.7	3 (15)	8
c	15.2	21.2	3 (15)	11
d	15.1	21.3	3 (15)	10
e	14.9	21.1	3 (15)	12
f	15.1	23.5	3 (15)	10
g	14.9	21.3	3 (15)	12

Set **a** represents our most specific model, where we explicitly code the full axial or equatorial ligand identity as a discrete variable, limiting the application of the model but producing one of the lowest RMSEs for  $\Delta E_{\text{H-L}}$  and  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$  (Table 3.1). Elimination of ligand identity in favor of ligand connecting atom elemental identity and total number of atoms in set **b** increases  $\Delta E_{\text{H-L}}$  MSE slightly and decreases  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$  MSE (Table 3.1).

Property	Variable set						
	a	b	c	d	e	f	g
<b>Complex-based</b>							
Metal identity	■	■	■	■	■	■	■
Oxidation state	■	■	■	■	■	■	■
$a_{\text{HF}}$	■	■	■	■	■	■	■
sum( $\Delta\chi$ )	■	■	■	■	■	■	■
min( $\Delta\chi$ )	■	■	■	■	■	■	■
max( $\Delta\chi$ )	■	■	■	■	■	■	■
<b>Ligand-based</b>							
Identity	■	■	■	■	■	■	■
Connection atom	■	■	■	■	■	■	■
Charge	■	■	■	■	■	■	■
Denticity	■	■	■	■	■	■	■
Number of atoms	■	■	■	■	■	■	■
Bond order	■	■	■	■	■	■	■
Truncated Kier index	■	■	■	■	■	■	■

Figure 3-3: Summary of variables chosen for each set **a** through **g**. Employed variables are indicated in shaded gray and grouped by whether they are assessed on the whole complex (complex-based) or on each individual axial or equatorial ligand (ligand-based).  $\Delta\chi$  is the difference in Pauling electronegativity between the ligand connecting atom and all atoms bonded to it, and the sum, maximum or minimum values are obtained over all ligands.

The shift from set **a** to **b** increases the model applicability but at the cost of omitting subtler ligand effects. For instance, ethylenediamine (11, en) and phenanthroline (10, phen) have the same ligand charge/denticity and direct ligand atom (N), making them equivalent in set **b** except for the larger size of phen. System size alone is not expected to be a good predictor of field strength (e.g., the small CO is one of the strongest field ligands). In set **c**, we introduce properties that depend on the empirical pairwise Pauling electronegativity difference ( $\Delta\chi$ ) between the ligand connecting atom (LC) and any  $i$ th atom connected (CA) to it:

$$\Delta\chi_{LC,i} = \chi_{LC} - \chi_i \quad (3.4)$$

These whole-complex differences include the maximum,  $\max(\Delta\chi)$  and minimum,  $\min(\Delta\chi)$ , as well as sum:

$$\text{sum}(\Delta\chi) = \sum_{\text{lig}} \sum_{LC \in \text{lig}} \sum_{i \in CA} \Delta\chi_{LC,i} \quad (3.5)$$

which is taken over the direct ligand atom and all atoms bonded to it for all ligands (lig.) in the complex. These additional set **c** descriptors reduce  $\Delta E_{\text{H-L}}$  MSE slightly and decrease the  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$  MSE to its lowest value (see Table 3.1). In set **d**, we eliminate  $\min(\Delta\chi)$  expecting it to be redundant with the max and sum, at the cost of a small increase in both MSEs.

Finally, in sets **e-g**, we replace ligand size (i.e., number of atoms) with general descriptors to enable prediction on molecules larger than those in any training set. For example, tetraphenylporphyrin will have comparable electronic properties to unfunctionalized porphyrin (12), despite a substantial size increase. In set **e**, we introduce the maximum bond order of the ligand connecting atom to any of its nearest neigh-

bors, a measure of the rigidity of the ligand environment, which is zero if the ligand is atomic (see Appendix A, Table A.1). In set **f**, we eliminate the number of atoms and bond order metric, replacing them with a broader measure of the ligand geometry adjacent to the metal. After trial and error, we have selected the truncated Kier shape index<sup>453</sup>,  ${}^2\kappa$  which is defined by the inverse ratio of the square of number unique paths of length two ( ${}^2P$ ) in the molecular graph of heavy atoms to the theoretical maximum and minimum for a linear alkane with the same number of atoms:

$${}^2\kappa = \frac{2^2P_{max} \ 2P_{min}}{({}^2P)^2} \quad (3.6)$$

and set to zero for any molecules that do not have paths of length two. The truncation means that only the ligand atoms within three bonds of the connecting atom are included in the graph. The set **f** MSEs are comparable to or a slight increase from sets with molecule size, but they beneficially eliminate system size dependence. In set **g**, we reintroduce the bond order metric as well, providing the lowest MSEs except for set **a** or **c**, both of which are much less transferable than set **g**. Thus, the comparable performance of set **g** to a full ligand descriptor (set **a**) supports our hypothesis that a combination of metal-centric and ligand-centric in a heuristic descriptor set can be predictive and transferable.

This final feature space is 15-dimensional with five per-complex descriptors and five per-ligand descriptors for each equatorial or axial ligand (see Table 3.2 for ranges of values and descriptions). A comparison of all errors and weights of variables across the seven data sets is provided in Appendix A, Tables A.7–A.21 and Figure A-1.

Table 3.2: Optimal (set **g**) input space descriptors and their range in the training set.  $\Delta\chi$  is the difference in Pauling electronegativity between the ligand connecting atom and all atoms bonded to it. Here, a continuous descriptor corresponds to a single input, whereas discrete descriptors correspond to one input per level.

Symbol	Type	Descriptor	Values or Range
whole-complex descriptors			
M	Discrete	metal identity	Cr, Mn, Fe, Co, Ni
O	Continuous	oxidation state	2 to 3
me	Continuous	max $\Delta\chi$ over all ligands	-0.89 to 1.20
se	Continuous	sum of $\Delta\chi$ over all ligands	-5.30 to 7.20
aHF	Continuous	HF exchange fraction	0.00 to 0.30
ligand-specific descriptors			
L	Discrete	ligand connection atom	Cl, S, C, N, or O
C	Continuous	ligand charge	0 to -2
k	Continuous	truncated Kier index	0.00 to 6.95
b	Continuous	ligand bond order	0 to 3
D	Continuous	ligand denticity	1 to 4

### 3.2.3 Training and uncertainty quantification of ML models

ANNs enable complex mapping of inputs to outputs<sup>457</sup> beyond multiple linear regression and support the use of both discrete (i.e., binary choices such as metal identity) and continuous (e.g., the % of HF exchange) variables. Here, we apply an ANN with an input layer, two intermediate hidden layers, and an output layer (Figure 3-2). The network topology was determined by trial and error, with additional hidden layers yielding no improved performance. All analysis is conducted in R 3.2.5<sup>456</sup>, using the H2O package<sup>458</sup> with tanh non-linearity and linear output.

As with many ML models, ANNs are sensitive to over-fitting due to the number of weights to be trained<sup>459</sup>. We address overfitting using dropout<sup>460,461</sup>, wherein robustness of the fit is improved by zeroing out nodes in the network with an equal probability,  $p_{drop}$ , at each stage of training (5% for spin-state splitting, 15% for HF exchange sensitivity, and 30% for bond lengths, selected by trial and error). Dropout has been shown to address overfitting when training feedforward ANNs on small datasets<sup>96</sup>, with larger values of  $p_{drop}$  giving more aggressive regularization that worsens training errors but improves test errors. We use  $L_2$  weight regularization with a fixed penalty weight  $\lambda$ , as is applied in standard ridge regression, with an effective loss function for training:

$$\{W\} = \arg \min_W \left( \sum_{i=1}^N (y_{pred}(x_i) - y_i)^2 + \lambda \sum_{l=1}^L (\|W_l\|_2^2 + \|b_l\|_2^2) \right) \quad (3.7)$$

Here,  $W_l$  refers to the weights from layer  $l$  to  $l + 1$ ,  $b_l$  are the corresponding biases, and  $y_{pred}(x_i)$  is the ANN prediction for the input-output pair  $(x_i, y_i)$ , and the sums run over  $N$  training pairs and  $L$  layers.

During network training, we randomize the order of data points and partition the first 60% as training data and the last 40% for testing. Dropout networks, consisting of two hidden layers of 50 nodes each, are trained on the data set for varying values of  $\lambda$  ranging from  $10^{-1}$  to  $10^{-1}$  using 10-fold cross validation. For each  $\lambda$ , the training data is partitioned into ten groups, a network is trained on nine of the groups and scored based on eq. 3.7 on the left-out group to select the best regularization parameter:  $5 \times 10^{-4}$  for spin-state splitting,  $10^{-2}$  for HF exchange sensitivity, and  $\times 10^{-3}$  for bond lengths. We varied and optimized<sup>462</sup> the learning rate between 0.05 and 1.5, and optimal rates were selected as 1.0 (bond lengths) and 1.5 (spin-state splitting or HF exchange sensitivity). We use batch optimization for training (batch



size = 20) for 2000 epochs. The training algorithm minimizes eq. 3.7 over the training data using stochastic gradient descent<sup>320,462–464</sup>.

It has been challenging to estimate ANN model uncertainty<sup>460,465</sup> with the possible exception of bootstrapping<sup>466</sup> by training the ANN on numerous subsamples of available training data. Model uncertainty will be due to either high-sensitivity to descriptor changes or test molecule distance in chemical space to training data (see also Section 3.3). Recent work<sup>459</sup> showed that minimization of the loss function in eq. 3.7 is equivalent to approximate variational optimization of a Gaussian process (GP), making previously suggested ANN sampling for different dropout realizations<sup>460</sup> a rigorously justified<sup>459</sup> model uncertainty estimate. We sample  $J = 100$  distinct networks each with output  $y_{pred}^j$  with different nodes dropped at the optimized weights and average over the predictions:

$$\bar{y}_{pred}(x_i) = \frac{1}{J} \sum_{j=1}^J y_{pred}^j(x_i) \quad (3.8)$$

The ANN predictive variance is estimated as<sup>459</sup>:

$$\text{var}(\bar{y}_{pred}(x_i)) \approx \tau^{-1}I + \frac{1}{J} \sum_{j=1}^J (y_{pred}^j(x_i)^T y_{pred}^j(x_i) - \bar{y}_{pred}(x_i)^T \bar{y}_{pred}(x_i)) \quad (3.9)$$

Here,  $\tau$  is

$$\tau = \frac{(1 - p_{drop}) l^2}{2N\lambda} \quad (3.10)$$

where  $N$  is the number of training data points, and  $l$  is a model hyperparameter for the GP that affects the estimation of predictive variance but does not enter into the

ANN training. The contribution of  $\tau$  in eq. 3.9 is a baseline variance inherent in the data, whereas the second term represents the variability of the GP itself. We obtain  $\tau$  values of 0.6 for spin-state splitting, 0.07 for HF exchange sensitivity, and  $10^4$  for bond lengths (see Section 3.3). We choose  $l$  by maximizing the log predictive likelihood of the corresponding GP based on the training data (details are provided in the Appendix A, Text A.2).

We selected an ANN model based on the successful demonstrations<sup>24,79,410</sup> of ANN-based models for predicting quantum chemical properties but also provide a comparison to two other common machine learning models<sup>209</sup>: kernel ridge regression (KRR) and a support vector regression model (SVR), both using a square-exponential kernel. We used the R package kernlab<sup>467</sup> and selected hyperparameters (the width of the kernel, and the magnitude of the regularization parameters which are given in the Appendix A, Table A.22) using a grid search and ten-fold cross-validation using the R package CVST<sup>468</sup>. We also compared training on our descriptor set to a KRR model with a kernel based on the  $L_1$  distance between sorted Coulomb matrix representations<sup>250</sup>, as demonstrated previously<sup>31,79</sup>.

## 3.3 Results and discussion

### 3.3.1 Overview of data set spin-state energetics

Analysis of the qualitative and quantitative features of the spin-state splitting data set motivates the training of an ANN to move beyond ligand field arguments. We visualize qualitative ground states (i.e., high-spin or low-spin) for the homoleptic subset of the data using a recursive binary tree (Figure 3-4, descriptor definitions provided in Table 3.2), as previously outlined<sup>106</sup> and implemented in the open source

rpart package<sup>469</sup> for R 3.2.5<sup>456</sup>. A recursive binary tree is a list of “branches” of the data ordered by statistical significance that gives the most homogeneous final “leaves” (here, with at least 10 data points) after a given number of permitted divisions (here, 6). Using descriptor set **g**, the data are partitioned into branches by testing which descriptors provide the “best” division to produce majority high- or low-spin states in leaves based on the concept of information impurity<sup>469</sup> and pruning to remove statistically insignificant branches. The resulting electronic structure spectrochemical “tree” simultaneously addresses metal-specific strengths of ligands and exchange-correlation sensitivity. As expected, strong field direct carbon ligands (no Cl, N, O or S in Figure 3-4) provide the root division of the tree, producing low-spin ground states for 92% of all Mn, Fe, and Co complexes (far right box on the third tier in Figure 3-4). Next level divisions include the M(II) oxidation state for  $a_{HF} > 0.05$  that are predominantly (96%) high-spin. Spin-state ordering is well-known<sup>196,197</sup> to be sensitive to HF exchange, and the tree reveals Mn<sup>3+</sup> with nitrogen ligands to have the strongest  $a_{HF}$  dependence, since they are 69% high-spin for  $a_{HF} > 0.1$  but 90% low-spin for  $a_{HF} \leq 0.1$ . Extension of the recursive binary tree to heteroleptic compounds produces a second-level division based on sum ( $\Delta\chi$ ), validating the relevance of the identified electronegativity descriptors for predicting heteroleptic spin-state ordering (Appendix A, Figure A-2).

Quantitatively, the maximum  $\Delta E_{H-L}$  in the data set is 90.7 kcal/mol for the strong-field Co(III)(misc)<sub>6</sub> complex at  $a_{HF} = 0.00$ , and the minimum value is -54.2 kcal/mol for the weak-field Mn(II)(NCS<sup>-</sup>)<sub>6</sub> at  $a_{HF} = 0.30$ . These extrema are consistent with i) the ordering of metals in the spectrochemical series<sup>173</sup> and ii) the uniform effect of stabilizing high-spin states with increasing HF exchange. By comparing compound trends in the data set, we are able to identify whether additivity in ligand field effects, which has been leveraged previously in heuristic DFT correction models<sup>193,206,470</sup>, is

a universally good assumption. For the  $\text{Fe(III)}(\text{Cl}^-)_{6-n}(\text{pisc})_n$  complexes (denoted 1-1 through 3-3 in Figure 3-5), increasing  $n$  from 0 to 2 through the addition of two axial pisc ligands increases the spin-state splitting by 15.1 kcal/mol per replaced chloride. Transitioning to a complex with all equatorial pisc ligands ( $n = 4$ ) increases the spin-state splitting by only 10.4 kcal/mol per additional ligand, and the homoleptic structure pisc ( $n = 6$ ) only adds 7.5 kcal/mol per additional ligand beyond the  $n = 4$  case. An additive model cannot precisely reproduce diminishing ligand effects.

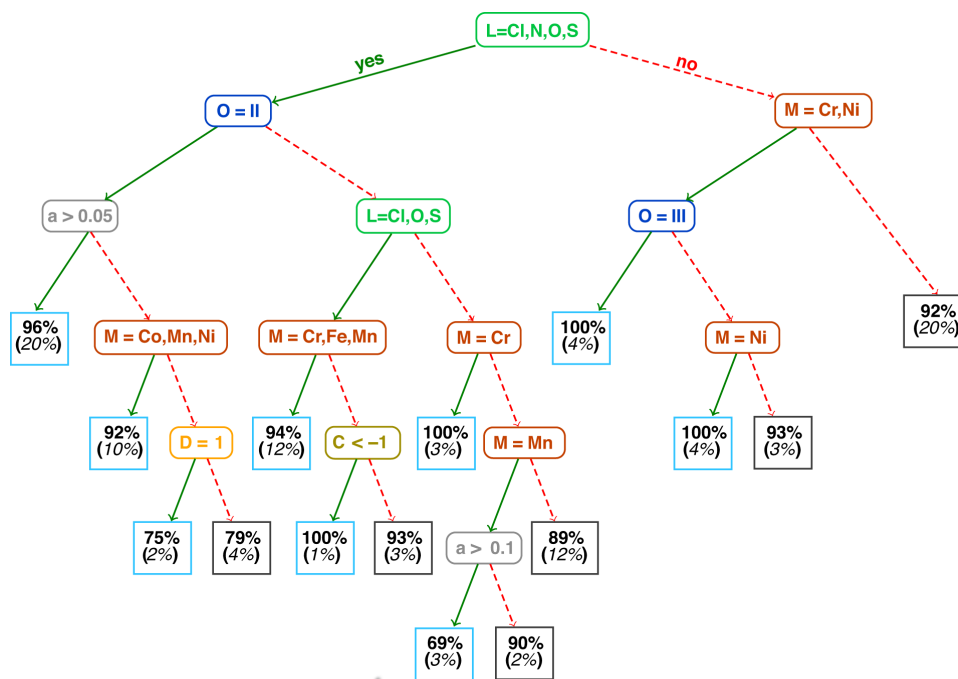


Figure 3-4: Binary ground state classification tree for homoleptic compounds. M indicates metal identity, L ligand connection atom, O oxidation state,  $a$  the fraction of HF exchange, C the charge, and D the ligand denticity. Each leaf node indicates the percent of elements in that leaf (light blue boxes for high-spin and dark gray boxes for low-spin) in bold font and percentage of total homoleptic population in the node (italic font, in parentheses).

As a stronger example for the need for nonlinear models such as an ANN, replacing two axial ligands from the strong-field  $\text{Mn(II)(CO)}_6$  complex with the weaker-field  $\text{NCS}^-$  (6-6 and 6-7 in Appendix A, Figure A-3) alters  $\Delta E_{\text{H-L}}$  by  $< 1$  kcal/mol, as strong-field ligands (e.g., CO,  $\text{CN}^-$ ) have an overriding effect on spin-state splitting.

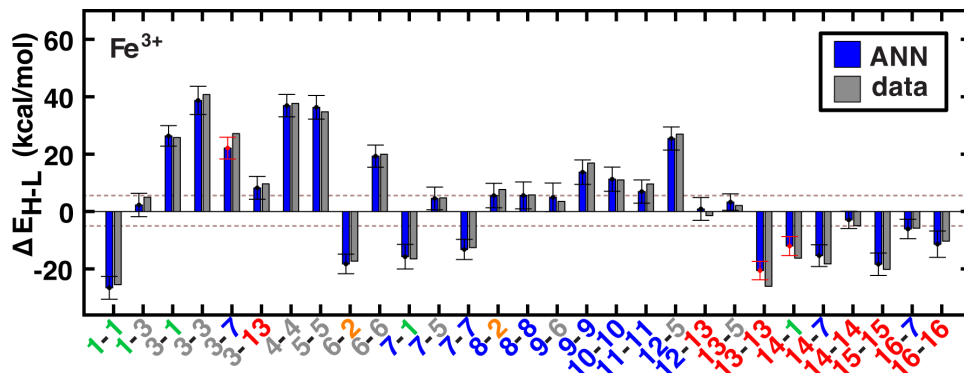


Figure 3-5: ANN model predictions (ANN, blue bars) and computed (data, gray bars) spin-state splittings,  $\Delta E_{\text{H-L}}$ , for the B3LYP functional ( $a_{\text{HF}} = 0.20$ ) in kcal/mol. Complexes are labeled by equatorial and then axial ligands according to the numbering indicated in Figure 3-1 and color-coded by direct ligand atom (green for chlorine, gray for carbon, blue for nitrogen, red for oxygen, and orange for sulfur). The error bars represent an estimated  $\pm 1$  standard deviation credible interval from the mean prediction, and error bars that do not encompass the computed value are highlighted in red. Brown dashed lines correspond to a  $\pm 5$  kcal/mol range around zero  $\Delta E_{\text{H-L}}$ , corresponding to near-degenerate spin states.

### 3.3.2 Spin-state splittings from an ANN

Motivated by non-linear effects in ligand additivity, we trained an ANN using a heuristic descriptor set (see Section 3.2.2) to predict qualitative spin-state and quantitative spin-state splitting. The ANN predicts the correct ground state in 98% of the test cases (528 of 538) and 96% of training cases (777 of 807). All of the misclassifications are for cases in which DFT  $\Delta E_{\text{H-L}}$  is  $< \pm 5$  kcal/mol (Appendix A).

The ANN spin-state prediction errors are not sensitive to HF exchange mixing, and thus our trained ANN is able to predict ground states of transition metal complexes from the pure GGA limit to hybrids with moderate exchange.

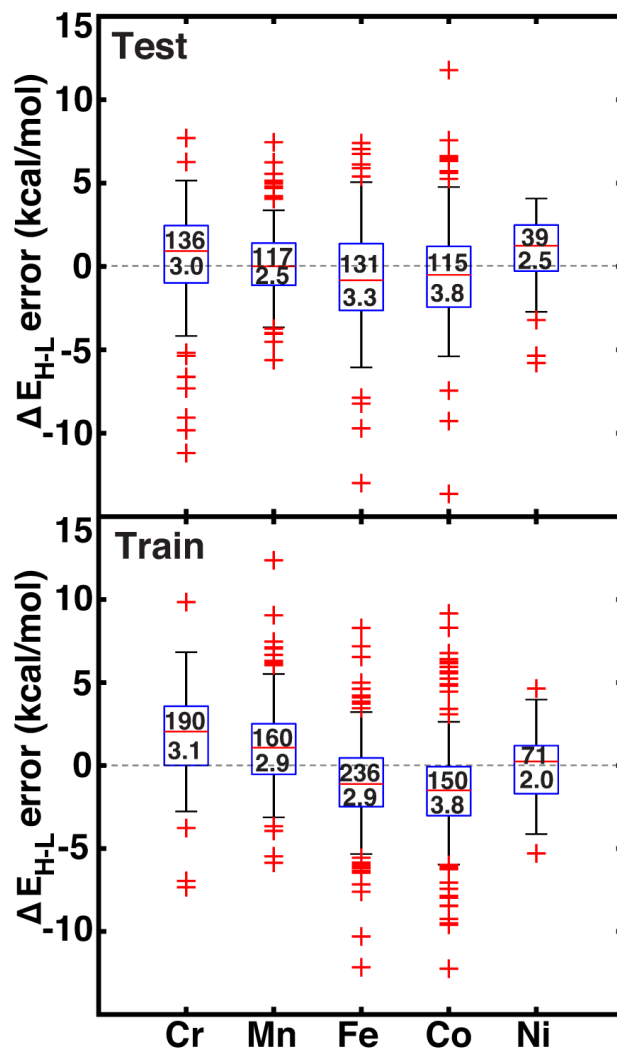


Figure 3-6: Error boxplots for  $\Delta E_{H-L}$  in kcal/mol using the ANN for test (top) and training (bottom) data partitioned by metal identity. The top number inside the box indicates the number of cases in each set, and the bottom number indicates the RMSE in kcal/mol. The range for both graphs is from 15 kcal/mol to  $-15$  kcal/mol.

We assess quantitative performance with root mean squared errors (RMSE) of the ANN (eq. 3.3), overall and by metal (Figure 3-6, Appendix A, Table A.23, and Appendix A, Figures A-3–A-6). The comparable RMSE of 3.0 and 3.1 kcal/mol for the test and training data, respectively, indicate an appropriate degree of regularization. The ANN predicts DFT spin-state splittings within 1 kcal/mol (i.e., “chemical accuracy”) for 31% (168 of 538) of the test data and within 3 kcal/mol (i.e., “transition metal chemical accuracy”<sup>187</sup> for 72% (389 of 538) of the test data. Only a small subset of 49 (4) test molecules have errors above 5 (10) kcal/mol, and correspond to strong-field Co and Cr complexes, e.g., Cr(II)(NCS<sup>-</sup>)<sub>2</sub>(pisc)<sub>4</sub> (Appendix A, Figure A-5). The model is equivalently predictive for homoleptic and heteroleptic compounds at 2.2 and 2.3 kcal/mol average unsigned error respectively.

The training and test RMSEs broken down by metal reveal comparable performance across the periodic table (Figure 3-6). Slightly higher test RMSEs (maximum unsigned errors) for Co and Fe complexes at 3.8 (15.7) and 3.3 (13.0) kcal/mol, respectively, are due to the train/test partition and more variable ligand dependence of spin-state ordering in these complexes (Figure 3-6 and Appendix A, Table A.23). When the ANN performs poorly, the errors are due to both under- and over-estimation of  $\Delta E_{H-L}$  for both strong- and weak-field ligands, regardless of HF exchange fraction: e.g.,  $\Delta E_{H-L}$  for Co(III)(CN<sup>-</sup>)<sub>6</sub> at  $a_{HF} = 0.00$  and Co(III)(en)<sub>3</sub> at  $a_{HF} = 0.20$  are overestimated by 14 and 9 kcal/mol, respectively, but  $\Delta E_{H-L}$  for Fe(III)(Cl<sup>-</sup>)<sub>6</sub> at  $a_{HF} = 0.10$  and Co(II)(H<sub>2</sub>O)<sub>2</sub>(CN<sup>-</sup>)<sub>4</sub> at  $a_{HF} = 0.30$  and are underestimated by 9 and 7 kcal/mol, respectively.

Quantified uncertainty estimates correspond to a baseline standard deviation in the model of approximately 1.5 kcal/mol ( $\sqrt{\tau^{-1}}$ ) and a mean total estimated standard deviation across the training and test cases of 3.8 and 3.9 kcal/mol, respectively (see sec 2.3 and error bars on Figure 3-5). These credible intervals are not rigor-

ously confidence intervals but can highlight when prediction uncertainty is high: a  $\pm 1$  ( $\pm 2$ ) standard deviation (std. dev.) interval on ANN predictions captures 83% (98%) of computed values for test set (see Appendix A, Figure A-7). Highest std. dev. values of around 5 kcal/mol are observed for Fe(II) and Mn(II) complexes and the lowest are around 3 for Cr and Co complexes. A single std. dev. around the ANN prediction contains the calculated  $\Delta E_{H-L}$  for 26 of 29 Fe(III) complexes at  $a_{HF} = 0.20$  but misses heteroleptic oxygen coordinating complexes, 13-13 and 14-1, and underestimates the effect of C/N ligands in 3-7 (Figure 3-5). The model performs consistently across different ligand sizes, from porphyrin Fe(III) complexes (12-13, 12-5) to Fe(II)(NH<sub>3</sub>)<sub>6</sub> and Fe(II)(CO)<sub>6</sub> (6-6 and 8-8). For ligand-specific effects, the ANN performs well, reversing splitting magnitude as equatorial and axial ligands are swapped (e.g., 1-3 versus 3-1).

Review of other metals/oxidation states reveals comparable performance for cases where the high-spin state is always favored (e.g., Mn(II), Cr(III), or Ni(II)), low-spin state is always favored (e.g., Cr(III)), and those where ligands have strong influence over the favored spin state (e.g., Fe(II) and Cr(II)) (see Appendix A, Figure A-3-A-6). For instance, metal-specific effects examined through comparison of M(II)(CO)<sub>6</sub> complexes (Figure 3-7) reveal good ANN performance both for where the strong-field ligand strongly favors the low-spin state (i.e., Fe and Ni) and where the spin-states are nearly degenerate (i.e., Cr, Mn, Co). The trends outlined here for 20% HF exchange hold at other exchange mixing values. Thus, our ANN trained on a modest data set with heuristic descriptors predicts spin-state splitting within a few kcal/mol of the DFT result.



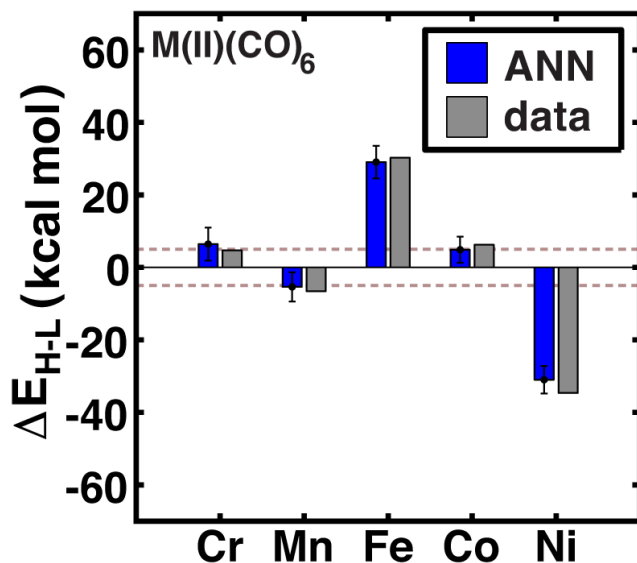


Figure 3-7: ANN model predictions (ANN, blue bars) and computed (data, gray bars) spin-state splittings,  $\Delta E_{H-L}$ , with the B3LYP functional ( $a_{HF} = 0.20$ ) in kcal/mol on  $M(II)(CO)_6$  complexes, where  $M = Cr, Mn, Fe, Co,$  or  $Ni$ . The error bars represent an estimated  $\pm 1$  standard deviation credible interval from the mean prediction, and brown dashed lines correspond to a  $\pm 5$  kcal/mol range around zero  $\Delta E_{H-L}$ , corresponding to near-degenerate spin states.

Comparing our results to KRR, SVR, and LASSO regression reinforces the choice of an ANN (Table 3.3 and Appendix A, Figure A-8). The ANN outperforms KRR with either our descriptor set or the sorted Coulomb matrix descriptor both on the full data set or at fixed HF exchange (Appendix A, Text A.3). The ANN also performs slightly better than SVR on test data with our descriptors. Linear LASSO regression was employed for feature selection (Section 3.2.2) but is outperformed by all other methods (Table 3.3). We will revisit the performance of these models on a more diverse molecule test set in Section 3.3.5 to assess the question of transferability.

Table 3.3: Train/test data and CSD test set RMSEs and max UEs in kcal/mol.HFX-1 for different machine learning methods and descriptor sets compared: KRR, kernel ridge regression, using square-exponential kernel for descriptor set g and the L1 matrix distance<sup>52</sup> for the sorted Coulomb matrix descriptor; SVR, support vector regression using square-exponential kernel; ANN, artificial neural network. Results are also given for the KRR/Coulomb case, restricted to B3LYP only since the Coulomb matrix does not naturally account for varying HF exchange.

Model	Descriptor	Training		Test		CSD	
		RMSE	max UE	RMSE	max UE	RMSE	max UE
LASSO	set g	16.1	89.7	15.7	93.5	19.2	72.5
KRR	set g	1.6	8.5	3.9	17.0	38.3	88.4
SVR	set g	2.1	20.9	3.6	20.4	20.3	64.8
ANN	set g	3.0	12.3	3.1	15.6	13.1	30.4
KRR	sorted Coulomb	4.3	41.5	30.8	103.7	54.5	123.9
KRR, B3LYP only	sorted Coulomb	17.2	58.0	28.1	69.5	46.7	118.7

### 3.3.3 Predicting exchange sensitivity with an ANN

Spin-state splittings exhibit high sensitivity to exchange<sup>196,197</sup> with linear behavior that we previously identified<sup>197</sup> to be strongly dependent on direct ligand identity and field strength when we compared a set of Fe complexes. Over this data set, computed exchange sensitivities are indeed linear, ranging from  $-174$  kcal/mol.HFX-1 for strong-field Fe(II)(CO)<sub>6</sub> to  $-13$  kcal/mol.HFX<sup>-1</sup> for weak-field Cr(III)(en)<sub>2</sub>(NH<sub>3</sub>)<sub>2</sub>. Cr(III) is the least exchange-sensitive metal in our test set, whereas Fe(II) and Mn(II) are the most sensitive (Appendix A, Table A.24 and Figure A-9).

We therefore generalize previous observations<sup>45</sup> in an ANN that predicts HF exchange sensitivity of spin-state ordering,  $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$ , using the same descriptors as for direct spin-state splitting, excluding only  $a_{HF}$ . The smaller size of this data set ( $\frac{1}{7}$  the size of the  $\Delta E_{H-L}$  data set) leads to overfitting, with lower RMSE values

of 13 kcal/mol.HFX<sup>-1</sup> for the training data versus 22 kcal/mol.HFX<sup>-1</sup> for the test set (Table 3.4, Appendix A, Figure A-10). Although results are reported in units of HFX (from 0 to 100% exchange), for typical 20% variation in exchange, a 20 kcal/mol.HFX<sup>-1</sup> sensitivity error only corresponds to a 4 kcal/mol energy difference. Both maximum unsigned errors (UE) and RMSEs are largest for Mn(II/III) and Cr(II) complexes, with the largest case producing an 92 kcal/mol.HFX<sup>-1</sup> underprediction for Mn(III)(H<sub>2</sub>O)<sub>4</sub>(pisc)<sub>2</sub>. Overall, the ANN prediction errors are less than less than 20 (40) kcal/mol.HFX<sup>-1</sup> for 65% (95%) of the test data. The ANN provides a valuable strategy for predicting exchange sensitivity, reproducing nonmonotonic and nonconvex ligand sensitivity in heteroleptic compounds: a Fe(III) complex with ox, 16, and NCS<sup>-</sup>, 7, ligands is more sensitive to HFX than the respective homoleptic complexes (Figure 3-8, other metals in Appendix A, Figures A-11–A-14).

Table 3.4: Test set RMSEs in kcal/mol.HFX<sup>-1</sup> separated by metal and oxidation state along with minimum and maximum unsigned test errors (UE). The number of test cases is indicated in parentheses.

Species	RMSE	min. UE	max. UE
Cr(II)	21(14)	4	45
Cr(III)	17(8)	2	37
Mn(II)	24 (6)	3	40
Mn(III)	38(8)	4	92
Fe(II)	18 (9)	2	41
Fe(III)	15(12)	<1	32
Co(II)	17(8)	<1	26
Co(III)	20(8)	<1	46
Ni(II)	9(4)	1	15

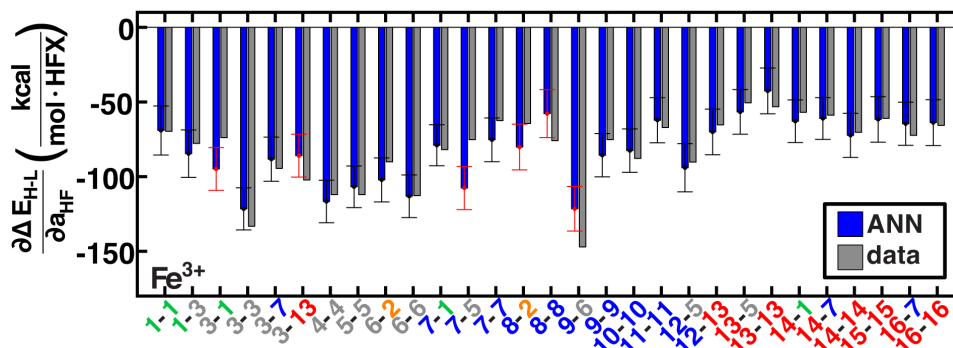


Figure 3-8: ANN model predictions (ANN, blue bars) and computed (data, gray bars) spin-state splitting sensitivities to HF exchange,  $\frac{\partial\Delta E_{H-L}}{\partial a_{HF}}$ , in kcal/mol.HFX<sup>-1</sup>, for Fe<sup>3+</sup> complexes. Complexes are labeled as equatorial and then axial ligands according to the numbering indicated in Figure 3-1 and color-coded by direct ligand atom (green for chlorine, gray for carbon, blue for nitrogen, red for oxygen, and orange for sulfur). The error bars represent an estimated  $\pm 1$  standard deviation credible interval from the mean prediction, and error bars that do not encompass the computed value are highlighted in red.

Uncertainty intervals of ANN predictions for HFX sensitivity yield a narrow range from 14 kcal/mol.HFX<sup>-1</sup> to 17 kcal/mol.HFX<sup>-1</sup>. For the 29 Fe(III) complexes studied, 23 (80%) of the ANN credible intervals span the computed exchange sensitivity (Figure 3-8). Across the full metal and oxidation state data set, 70% (83%) of the computed data is contained by  $\pm 1$  ( $\pm 2$ ) std. dev. intervals (Figure 3-8 and Appendix A, Figure A-15). This performance can be further improved by extending the training data. Exchange-sensitivity provides value both for extrapolation of computed (see Section 3.3.6) or literature values obtained at an arbitrary exchange mixing and in identification of cases of high-sensitivity to DFT functional choice.

### 3.3.4 Predicting metal-ligand bond length with an ANN

Using our descriptor set, we trained an ANN on the minimum metal-ligand bond distances for both low-spin and high-spin geometries ( $\min(R_{LS/HS})$ ), which only dif-

fer from the exact metal-ligand bond length for distorted or heteroleptic compounds. This ANN for bond length prediction extends capabilities we have recently introduced for generating high-quality transition metal complex geometries<sup>181</sup> in order to enable spin-state dependent predictions without requiring extended geometry-optimization.

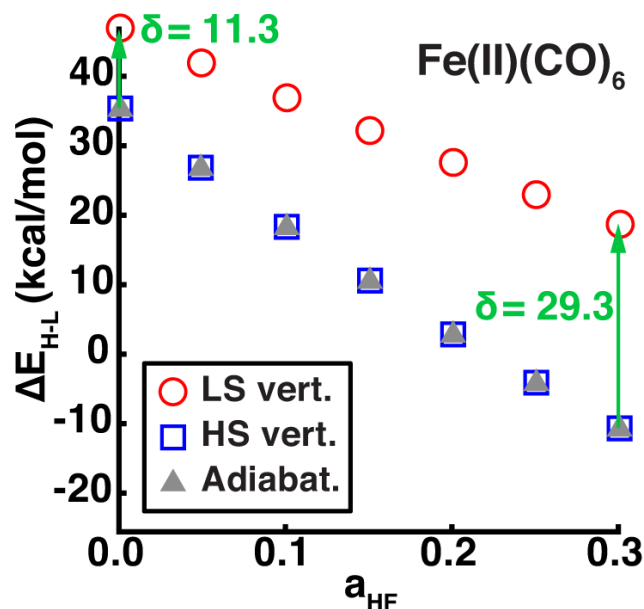


Figure 3-9: The vertical or adiabatic spin-state splittings,  $\Delta E_{H-L}$ , in kcal/mol as a function of HF exchange,  $a_{HF}$ , for  $\text{Fe(II)(CO)}_6$ . Spin-state splittings evaluated at the HS or LS geometries are indicated by open blue squares and open red circles, respectively. The adiabatic spin-state splitting is shown as filled gray triangles. The HS vertical and adiabatic splittings overlap, whereas the LS vertical splitting overestimates  $\Delta E_{H-L}$ , as indicated by the green arrow and annotated  $\delta$  in kcal/mol for  $a_{HF} = 0.00$  and  $a_{HF} = 0.30$ .

Furthermore, comparison of adiabatic and vertical spin-state splittings computed either at the low- or high-spin optimized geometries reveals that the vertical splitting at the HS geometry is indistinguishable from the adiabatic splitting, but the LS geometry vertical splitting favors the LS state by 10 – 30 kcal/mol, increasing with  $a_{HF}$  (Figure 3-9). Thus, if the ANN bond length predictions are accurate, adiabatic

spin-state splittings can be obtained from DFT single points at ANN-predicted HS-only or both LS/HS geometries.

Metal-ligand bond distances in the  $a_{HF} = 0.20$  data set vary from  $\min(R_{LS}) = 1.81\text{\AA}$  (in  $\text{Fe(II)(pisc)}_2(\text{Cl}^-)_4$ ) to  $\min(R_{HS}) = 2.55\text{\AA}$  (in  $\text{Fe(III)(Cl}^-)_6$ ). The metal-ligand bond length ANN produces comparable RMSE across training (0.02 Å for HS and LS) and test (0.02 Å for LS and 0.03 Å for HS) data with comparable errors regardless of metal identity and oxidation- or spin- state (Appendix A, Table A.25 and Figures A-16–A-27). ANN bond length std. devs. Range from 0.026 to 0.045 Å with a  $\sim 0.01$  Å baseline contribution. For low-spin (high-spin) complexes, 79% (81%) and 96% (96%) of the calculated values fall within one and two std. dev. of ANN-predicted bond lengths, respectively (Appendix A, Figures A-17 and A-23).

The ANN overestimates bond lengths of low-spin Fe(III) complexes by more than a full standard deviation for seven cases, e.g., underestimating Fe-C distances in CN (7-5, 13-5) and pisc (3-7, 3-13) complexes (Figure 3-10). However, it also reproduces subtle trends, e.g. replacing axial ligands in homoleptic LS  $\text{Fe(III)(pisc)}_6$  (3-3 in Figure 3-10,  $\min(R_{LS}) = 1.92\text{\AA}$ ) with  $\text{Cl}^-$  increases the minimum bond distance to 1.94 Å (3-1 in Figure 3-10), but replacing equatorial pisc ligands instead with  $\text{Cl}^-$  (1-3 in Figure 3-10) decreases the minimum bond distance to 1.90 Å, a feature reproduced by the ANN. Non-additive bond length effects motivate the use of the ANN in initial geometry construction<sup>437</sup>. Indeed, when we use ANN-predicted metal-ligand bond lengths in structure generation instead of our previous strategy based on a discrete database of DFT bond lengths<sup>437</sup>, we reduce the metal-ligand component of the gradient by 54–90% (Appendix A, Text A.4, Figure A-28 and Table A.26). The ANN-predicted bond lengths and spin states are now available in molSimplify<sup>437</sup> as an improved tool for structure generation.

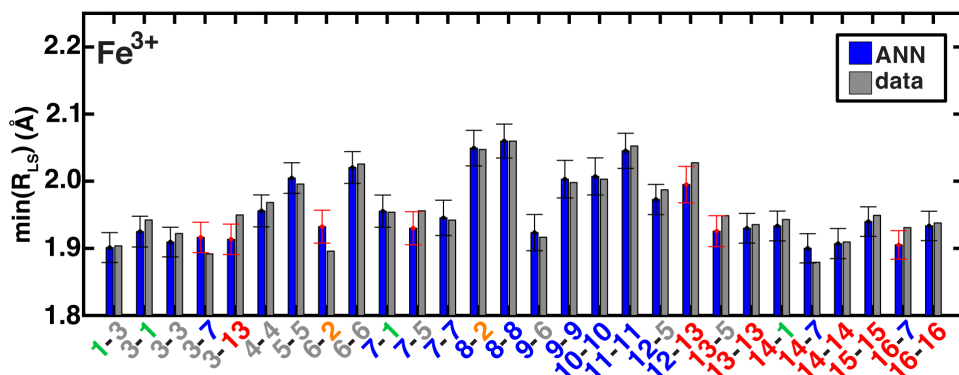


Figure 3-10: ANN model predictions (ANN, blue bars) and computed (data, gray bars) minimum LS  $\text{Fe}^{3+}$  bond lengths,  $\min(R_{LS})$ , in Å. Complexes are labeled as equatorial and then axial ligands according to the numbering indicated in Figure 3-1 and color-coded by direct ligand atom (green for chlorine, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation credible interval around the mean prediction, and error bars that do not encompass the computed value are highlighted in red.  $\text{Fe(III)(Cl)}_6$  (1-1) is excluded due to being off scale: it has a predicted/calculated bond length of 2.44/2.45 Å, and an error standard deviation of  $\pm 0.02$ .

### 3.3.5 Expanding the test set with experimental complexes

In order to test the broad applicability of the trained ANNs, we selected 35 homoleptic and heteroleptic octahedral complexes from the Cambridge Structural Database<sup>184</sup> (CSD) with a range of metals (Cr to Ni) and direct ligand atom types (N, C, O, S, Cl) (Appendix A, Table A.26). The CSD test cases span a broader range of compounds than the training set, containing i) larger macrocycles, e.g. substituted porphyrins (tests 9, 25), clathrochelates (test 16), phthalocyanines (tests 4, 7), and cyclams (tests 5, 12, 14, 17, 24, 29, and 33, 12 and 33 shown in Figure 3-11) and ii) coordination combinations or functional groups, e.g., OCN in test 30, absent from the training set. Indeed, large CSD test molecule sizes, e.g. up to 103 atoms in a single equatorial ligand, further motivates our relatively size-independent descriptor set over forms that do not scale well with molecule size.

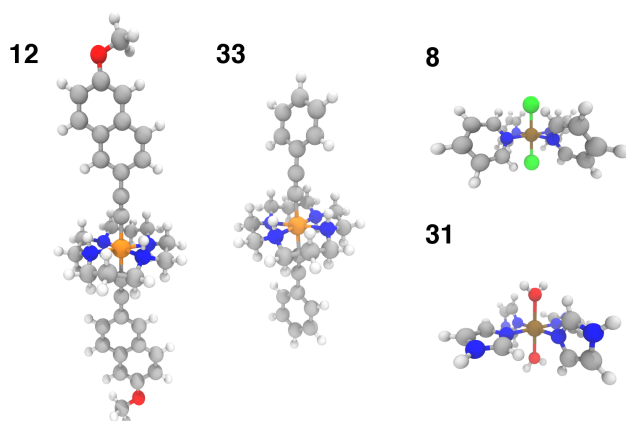


Figure 3-11: Representative CSD test set molecules shown in ball and stick representation with carbon atoms in gray, nitrogen atoms in blue, oxygen in red, hydrogen in white, chlorine in green, chromium in orange, and iron in brown. Test molecules 12 (CSD ID: SUMLET) and 33 (CSD ID: YUJCIQ) are Cr(III) cyclams for which the ANN performs least well, and test molecules 8 (CSD ID: TPYFEC04) and 31 (CSD ID: BIPGEN) are cases for which the ANN predicts  $\Delta E_{\text{H-L}}$  within 3 kcal/mol.

The ANN predicts CSD test case spin-state splittings within 5 kcal/mol for 15 of the 35 complexes, an overall mean unsigned error of 10 kcal/mol, and RMSE of 13 kcal/mol (see Appendix A, Table A.27). The large RMSE is due in part to poor performance on early-transition-metal cyclams (red symbols in left panel of Figure 3-12) for which the ANN overestimates spin-state splitting by at about 30 kcal/mol (Cr-cyclams, tests 12 and 33 in Figure 3-11). The ANN predicts spin-state splittings within around 3 kcal/mol for several non-macrocyclic complexes that are better represented in the training data (e.g., test cases 8 and 31 in Figure 3-11). The correct ground state is assigned in 90% of CSD test cases (96% after excluding cyclams); the only incorrect, non-cyclam spin state assignment is a spin-crossover complex, test 25 (calculated  $\Delta E_{\text{H-L}} = -0.2$  kcal/mol). Compared to other machine learning models (KRR and SVR), the ANN is more transferable to dissimilar CSD structures (Table 3.3), outperforming the next-best model, SVR, by 30%. The relative success of the



ANN on the CSD data is partially attributable to the use of dropout regularization, which has been shown<sup>461</sup> to improve robustness.

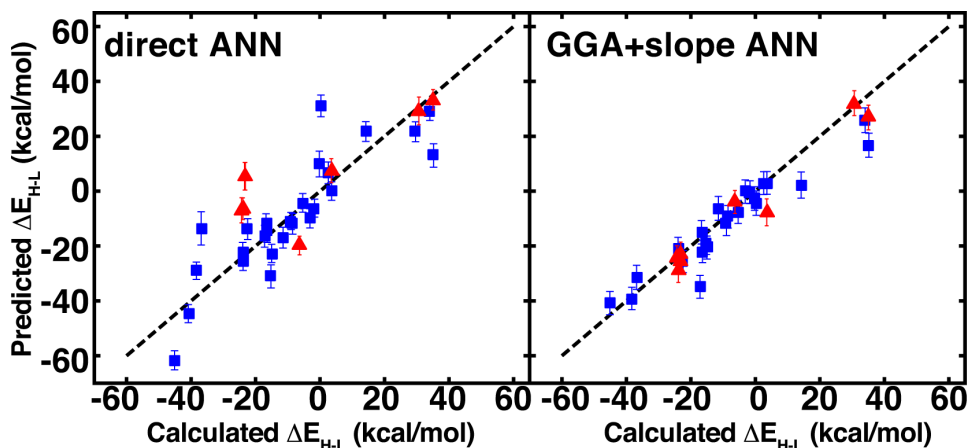


Figure 3-12: ANN spin-state splitting energy,  $\Delta E_{\text{H-L}}$ , prediction on CSD test structures vs. DFT-calculated values, both at  $a_{\text{HF}} = 0.20$  and in kcal/mol. Direct prediction (left) is compared to GGA calculations and extrapolation using the predicted slope from the ANN (right). Error bars represent a credible interval of one standard deviation from the model uncertainty analysis (either in direct ANN at left or slope ANN at right), and a parity line (black, dashed) is indicated. Cyclams are indicated in red triangles, as described in main text, and the remaining test cases are indicated by blue squares.

The observation of good performance with reasonable similarity between CSD structures and the training data but poor performance when the CSD structure is not well-represented motivates a quantitative estimate of test compound similarity to training data. We first computed overall molecular similarity metrics (e.g., FP2 fingerprint via Tanimoto<sup>428,471</sup> as implemented in OpenBabel<sup>77</sup>) but found limited correlation ( $R^2 = 0.1$ ) to prediction error (see Appendix A, Figure A-29 and Text A.5). Comparing the Euclidean and uncentered Pearson distances in descriptor space between the CSD test cases and the closest training data descriptors provides improved correlation to prediction error of  $R^2 = 0.3$  and  $R^2 = 0.2$ , respectively (Appendix A,

Figure A-30). Large errors (i.e.,  $> 15$  kcal/mol) are only observed at a Euclidean norm difference exceeding 1.0 (half of the CSD data), providing an indication of lack of reliability in ANN prediction. This high distance to training data does not guarantee inaccurate prediction, e.g., CSD test case 8, a Fe(II) tetrapyridine complex, is predicted with fortuitously good  $\sim 2$  kcal/mol error but has a Euclidean norm difference  $> 1.4$ . We have implemented the Euclidean norm metric alongside the ANN in our automated screening code<sup>181</sup> to detect complexes that are poorly represented in training data and advise retraining or direct calculation.

ANN-predicted equilibrium metal-ligand bond lengths for both HS and LS CSD geometries produced RMSEs of 0.10 and 0.07 Å, respectively (Appendix A, Tables A.28–A.29). Trends in bond length prediction error differ from those obtained for spin-state splitting. For instance, bond length errors are average in the cyclams even though spin-state splitting predictions were poor. The large Euclidean distance to training data heuristic ( $> 1.0$ ) is observed for five of the seven large (i.e.,  $> 0.1$ Å) HS bond distance errors (see Appendix A, Texts A.4–A.5 and Figures A-31–A-32). The highest HS prediction errors ( $> 0.2$ Å) occur for tests 8 and 35, underestimating the Fe-N bond length by 0.2Å (2.1Å ANN vs. 2.3 Å calculated) in the former case. Despite poor geometric predictions, the ANN predicts test 8  $\Delta E_{\text{H-L}}$  to within 3 kcal/mol, and this differing performance is due to the fact that predictions of these two outputs are independent. Interligand effects that are ignored by our descriptor set can restrict bond length extension, e.g. in test 16, where an O...O- interligand hydrogen-bond produces an unusually short 1.9Å high-spin Fe-N bond distance (vs. ANN prediction of 2.1 Å). Future work will focus on incorporating extended metrics of rigidity to account for these effects.

We investigated the relationship between the experimental CSD bond distances and the ANN-predicted bond distances. If the experimentally measured bond distance

lies close to one spin state’s predicted bond length, then the complex may be expected to be in that spin state, assuming i) the ANN provides a good prediction of the spin-state specific bond lengths and ii) that the gas-phase optimized DFT and CSD bond distances are comparable. The majority of experimental bond lengths are near the extrema of the ANN predictions (subset where ANN predicts LS-HS bond distance of at least  $0.05 \text{ \AA}$  shown in Figure 3-13, full set in Appendix A, Figure A-33).

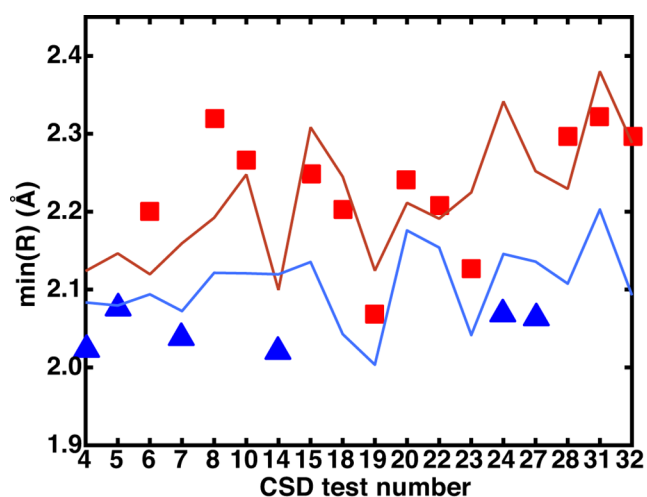


Figure 3-13: Comparison of measured CSD bond distances in the crystal phase, represented by symbols (red squares for high-spin or blue triangles for low-spin based on DFT assignment at  $a_{HF} = 0.20$  with the ANN predicted HS (red line) and LS (blue line) bond distances. Only the CSD test cases where the difference between ANN-predicted LS and HS bond distances is  $\geq 0.05 \text{ \AA}$  are shown for clarity. For all of these cases, the ANN correctly predicts the DFT spin state.

Nine of the twelve (9 of 9 in Figure 3-13) experimental bond lengths that are on or above the predicted HS bond distance boundary have an HS ground state, eleven of the fifteen (6 of 6 in Figure 3-13) experimental bond lengths that are on or below the predicted LS bond distance have an LS ground state, and remaining structures (3 in Figure 3-13) reside at intermediate distances. Some discrepancies are due to

differences between the gas phase geometries and those in the crystal environment (e.g., test 27 in Figure 3-13 and see Appendix A, Tables A.27–A.29). This bond-length-based spin-assignment thus provides a strategy for corroboration of direct spin-state prediction.

### 3.3.6 Extrapolating GGA functionals to hybrids with an ANN

Linear spin-state HF exchange sensitivity may be exploited to predict properties at one  $a_{HF}$  value from computed properties obtained at another, e.g., to translate literature values or to accelerate periodic, plane-wave calculations where incorporation of HF exchange increases computational cost. We carry out comparison of the utility of this  $\Delta$ -ML-inspired<sup>137</sup> strategy on the 35 CSD test set to identify if prediction errors are improved, especially for molecules poorly-represented in the training set. On the CSD molecules, extrapolating  $a_{HF} = 0.00$  spin-state ordering to  $a_{HF} = 0.20$  with the exchange-sensitivity ANN reduces the maximum error to 23 kcal/mol and decrease the mean unsigned error and RMSE to 5 kcal/mol and 7 kcal/mol (the right pane of Figure 3-11). For the GGA + slope ANN approach, excluding the nine cyclams does not change the RMSE/MUE values, confirming good ANN exchange-sensitivity prediction even when spin-state splitting prediction is poor.

These reduced average errors are quite close to the uncertainty introduced by the slope prediction performance at around 4 kcal/mol over a 20% exchange interval. Although this approach does eliminate the largest outliers and improve prediction across the CSD test set, it necessitates semi-local DFT geometry optimizations or a judicious bond length choice for vertically-approximated spin-state ordering. This approach also has limited benefit for cases well-represented in the training data set due to the sparser data set in the exchange sensitivity ANN. Indeed, over the origi-

nal test set molecules, extrapolated ANN exchange sensitivities on top of calculated  $a_{HF} = 0.00$  splittings produce an RMSE of around 4 kcal/mol comparable to or slightly worse than direct prediction (Appendix A, Figure A-34).

## 3.4 Conclusions

We have presented a series of ANN models trained using 2,690 DFT geometry optimizations of octahedral transition metal complexes generated from a set of 16 candidate axial and equatorial ligands and transition metals (Cr-Ni) at varying fractions of HF exchange. From the unseen test cases of a 60-40% train-test partition, we demonstrated good accuracy on spin-state splitting predictions of around 3 kcal/mol and metal-ligand bond distances around 0.02-0.03 Å. Our simple descriptor set, including: i) the ligand connection atom, ii) electronegativity and bonding of the coordinating ligand atom environment, iii) ligand formal charge, iv) ligand denticity, and v) metal identity and oxidation state ensures transferability of the ANN. Importantly, the employed connectivity models are not 3D-structure-based, instead relying on a truncated graph-theoretic representation of the ligand, making the approach suitable for screening large numbers of complexes without precise structural information. Although we have trained ANNs to predict bond lengths and spin-state splitting, the data set and descriptors could be used to predict other quantities such as ionization potential, redox potential, or molecular orbital energies. Such efforts are currently underway in our lab.

A test of our ANN on diverse molecules obtained from an experimental database indicated good performance, with MUEs of 5 kcal/mol for spin states for compounds within our proposed Euclidean distance reliability criteria and 10 kcal/mol for the full set. In both diverse and representative cases, the ANN outperforms other ma-

chine learning models. Our ANN predictions of HF exchange sensitivity provide a tool for interpolating between exchange-correlation functionals or extrapolating from semi-local GGAs to a hybrid result, which we demonstrated on CSD cases, improving MUE to 5 kcal/mol across the full 35 molecule set.

Natural extensions to this work include the development of the current ANN for extrapolation of GGA to hybrid functional properties in condensed matter systems and generalizing the coordination definition to enable prediction of properties of unsaturated metals in catalytic cycles. Overall, we have demonstrated a relatively sparse feature space to be capable of predicting electronic structure properties of transition metal complexes, and we anticipate that this strategy may be used for both high-throughput screening with knowledge of functional choice sensitivity and in guiding assessment of sources of errors in approximate DFT.

# Chapter 4

## Mapping transition metal complex space for machine learning

Note: This chapter was originally published as “Janet, J. P., Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships. *J. Phys. Chem. A* 2017, 121, 8939–8954” and has been formatted for consistency. Supporting information provided with the manuscript, available online at <http://www.doi.org/10.1021/acs.jpca.7b08750>, has been placed in Appendix B.

### Chapter summary

While the choice of representation is always important in ML, the small size of available data sets and lack of chemical intuition transition metal chemistry, where an appropriate molecular representation becomes a critical ingredient in ML model predictive accuracy. This chapter introduces a series of revised autocorrelation descriptors (RACs) that encode relationships between the heuristic atomic properties (e.g., size, connectivity, and electronegativity) on a molecular graph. By manipulating the starting point, scope, and nature of the quantities evaluated in standard ACs, we make these RACs efficient graph representations for inorganic chemistry. On an organic molecule set, we first demonstrate superior standard AC performance to other presently-available topologica-only descriptors for predictive inference. For inorganic chemistry, our RACs yield  $\sim 1$  kcal/mol ML MUEs on set-aside test molecules in

spin-state splitting in comparison to an order-of-magnitude higher errors from feature sets that encode whole-molecule structural information. Systematic feature selection methods including univariate filtering, recursive feature elimination, and direct optimization (e.g., random forest and LASSO) are used to down sample the high dimensional RAC space and identify the most important dimensions. Random-forest- or LASSO-selected subsets 4–5× smaller than the full RAC set produce even better predictive performance and show good transferability to metal-ligand bond length and redox potential prediction. Evaluation of feature selection results across property sets reveals the relative importance of local, electronic descriptors (e.g., electronegativity, atomic number) in spin-splitting and distal, steric effects in redox potential and bond lengths, leading to insights into the nature of ligand functionalization and the electronic properties of the metal center.

## 4.1 Introduction

Computational high-throughput screening is key in chemical and materials discovery<sup>14,75,99,109,472–478</sup>, but high computational cost has limited chemical space exploration to a small fraction of feasible compounds<sup>22,90</sup>. Machine-learning (ML) models have emerged as alternative approaches especially for efficient evaluation of new candidate materials<sup>407</sup> or potential energy surface fitting and exploration through sophisticated force field models<sup>24,213,214,254,410–413</sup>. Examples of recent ML applications in computational chemistry include exchange-correlation functional development<sup>260,479</sup>, general solutions to the Schrödinger equation<sup>414</sup>, orbital free density functional theory<sup>143,415</sup>, many body expansions<sup>416</sup>, acceleration of dynamics<sup>417,419,480</sup>, band-gap prediction<sup>236,420</sup>, and molecular<sup>34,406</sup> or heterogeneous catalyst<sup>231,481</sup> and materials<sup>421–424</sup> discovery, to name a few.

Essential challenges for ML models to augment or replace first-principles screening are model selection and transferable feature set identification. For modest sized data sets, descriptor set selection is especially critical<sup>223,227,255,426</sup> for successful ML



modeling. Good feature sets should<sup>227</sup> be cheap to compute, as low dimensional as possible, and preserve target similarity (i.e. materials with similar should properties have similar feature representations). Within organic chemistry, structural descriptors such as a Coulomb matrix<sup>426</sup> or local descriptions of the chemical environment and bonding<sup>223,427</sup> have been useful to enable predictions of energetics as long as a relatively narrow range of elements (e.g., C, H, N, O, F) is considered. These observations are consistent with previous successes in evaluating molecular similarity<sup>428</sup>, force field development<sup>429</sup>, quantitative structure-activity relationships<sup>430</sup>, and group additivity<sup>431</sup> theories on organic molecules.

Descriptors that work well for organic molecules have proven unsuitable for inorganic materials<sup>433</sup> or molecules<sup>308</sup>. This lack of transferability can be readily rationalized: it is well-known<sup>197,308,434,435</sup> that some electronic properties of transition metal complexes (e.g., spin state splitting) are much more sensitive to direct ligand atom identity that dominates ligand field strength<sup>436,437</sup>. Unlike organic molecules, few force fields have been established that can capture the full range of inorganic chemical bonding<sup>432</sup>. The spin-state- and coordination-environment-dependence of bonding<sup>173</sup> produces a higher-dimensional space that must be captured by sophisticated descriptors or functions. In spite of these challenges, suitable data-driven models for inorganic chemistry will be crucial in the efficient discovery of new functional materials<sup>60,445</sup>, for solar energy<sup>16</sup>, and for catalyst discovery<sup>58,402</sup>.

With the unique challenges of inorganic chemistry<sup>74</sup> in mind, we recently trained a neural network to predict transition metal complex quantum mechanical properties<sup>308</sup>. From several candidate descriptor sets, we demonstrated good performance, i.e., 3 kcal/mol root mean squared error for spin-splitting and 0.02–0.03 Å for metal-ligand bond lengths, of heuristic, topological-only near-sighted descriptors. These descriptors required no precise three-dimensional information and outperformed es-

tablished organic chemistry ML descriptors that encode more whole-complex information.

In this work, we introduce systematic, adaptable-resolution heuristic and topological descriptors that can be tuned to encode molecular characteristics ranging from local to global. As these descriptors require no structural information, rapid ML model prediction without prior first-principles calculation is possible, and such ML models can improve structure generation<sup>181</sup> through bond length prediction<sup>308,482</sup>. We apply this adaptable descriptor set to both organic and inorganic test sets, demonstrating excellent transferability. We use rigorous feature selection tools to quantitatively identify optimal locality and composition in machine learning feature sets for predicting electronic (i.e., spin-state and redox potential) and geometric (i.e., bond length) properties. The outline of the rest of this work is as follows. In Section 4.2, we review our new descriptors, methods for subset selection, and the ML models trained in this work. In Section 4.3, we provide the computational details of first-principles data sets and associated simulation methodology. In Section 4.4, we present Results and discussion on the trained ML models for spin-state splitting, bond-lengths, and ionization/redox potentials. Finally, in Section 4.5, we provide our conclusions.

## 4.2 Approach to feature construction and selection

### 4.2.1 Autocorrelation functions as descriptors

Autocorrelation functions<sup>244</sup> (ACs) are a class of molecular descriptors that have been used in quantitative structure-activity relationships for organic chemistry and drug design<sup>483-485</sup>. ACs are defined in terms of the molecular graph, with vertices for atoms and unweighted (i.e., no bond length or order information) edges for bonds.

Standard ACs<sup>244</sup> are defined as:

$$P_d = \sum_i \sum_j P_i P_j \delta(d_{ij}, d) \quad (4.1)$$

where  $P_d$  is the AC for property  $P$  at depth  $d$ ,  $\delta$  is the Dirac delta function, and  $d_{ij}$  is the bond-wise path distance between atoms  $i$  and  $j$ . Alternatives to the eq. 4.1 AC sums are motivated and discussed in Section 4.2.2. The AC depth  $d$  thus encodes relationships between properties of atoms separated by  $d$  bonds; it is zero if  $d$  is larger than the longest molecular path, and 0-depth ACs are just sums over squared properties. The five atomic, heuristic properties used in our ACs are: i) nuclear charge,  $Z$ , as is used in Coulomb matrices<sup>250</sup>; ii) Pauling electronegativity,  $\chi$ , motivated by our previous work<sup>482</sup>; iii) topology,  $T$ , which is the atom’s coordination number; iv) identity,  $I$ , that is 1 for any atom, as suggested in Ref.<sup>22</sup>; and v) covalent atomic radius,  $S$ . Although i, ii, and v are expected to be interrelated, the  $S$  quantity uniquely imparts knowledge of spatial extent, and covalent radii follow different trends than  $Z$  or  $\chi$  (e.g. the covalent radius of Co is larger than Fe and Ni). ACs are compact descriptors, with  $d + 1$  dimensions per physical property encoded at maximum depth  $d$ , that depend only on connectivity and do not require Cartesian or internal coordinate information. Although inclusion of geometric information improves predictive capabilities of machine learning models in organic chemistry<sup>486</sup>, reliance on structural information requires explicit calculation or knowledge of it prior to ML prediction, which is not practical for transition metal complexes. AC sets also are vectorial descriptors that are invariant with respect to system size and composition, unlike frequently-used symmetry functions<sup>24</sup>, bag-of-bonds<sup>25</sup>, and Coulomb matrices<sup>79,250</sup>.

Despite their promise in therapeutic drug design<sup>244,483,485</sup> or in revealing inorganic

complex structure-property relationships<sup>482</sup>, ACs have not yet been tested as features in machine learning models that predict quantum mechanical properties. We first apply ACs to the QM9 database<sup>31</sup> of 134k organic (C, H, O, N, and F elements) molecules consisting of up to nine heavy atoms. This database contains B3LYP/6-31G-calculated properties, including atomization energies and HOMO-LUMO gaps, making it a frequent test set for machine learning models and descriptor sets<sup>212,223,486,487</sup>. The QM9 test set allows us to both identify if there is an optimal maximum depth for ACs and to determine the baseline predictive capability of ACs in comparison to established descriptors<sup>25,79,250</sup>. Throughout this work, we score feature sets by training Kernel ridge regression (KRR) models<sup>209</sup> with a Gaussian kernel. KRR is a widely-employed<sup>25,79,250</sup> ML model, that has produced sub-kcal/mol out-of-sample property prediction error on large organic databases and crystals<sup>137,212,488</sup>. We have selected KRR for the i) ease of retraining, ii) transparency of differences in KRR models<sup>209</sup>, as predictions are related to arrangement of data points in feature space, and iii) wide use of KRR in computational chemistry<sup>25,79,137,250,488</sup> (Appendix B, Text B.1).

First, we test the effect of increasing the maximum AC depth to incorporate increasingly nonlocal ACs on AE prediction test set errors using a 1000 molecule training set repeated five times (Figure 4-1). We evaluate prediction test set mean unsigned error (MUE) on the remaining 133k molecules in the QM9 set. Test set MUEs first decrease with increasing depth from 18 kcal/mol MUE at zero-depth (i.e., only sums of constituent atom properties) and reach a minimum of 8.8 kcal/mol MUE at maximum three-depth ACs. Without any further feature reduction, maximum three-depth ACs (3d-ACs) correspond to a 20-dimensional feature set (i.e., 4 length scales  $\times$  5 properties). Increasing the maximum depth beyond three increases test errors slightly up to 9.2 kcal/mol for maximum six-depth ACs (Figure 4-1). Minimum train/test MUEs

with  $3d$ -ACs emphasizes the length scale of chemically relevant effects, in line with previous observations<sup>197,308,486</sup> and increasing train/test MUEs due to the addition of more poorly correlating non-local descriptors emphasizes the importance of careful feature selection (Section 4.2.3). Regardless of maximum depth chosen, AC-derived prediction accuracy is impressive since the KRR model is trained with  $< 1\%$  of the QM9 data set, which has a large overall AE mean absolute deviation of 188 kcal/mol.

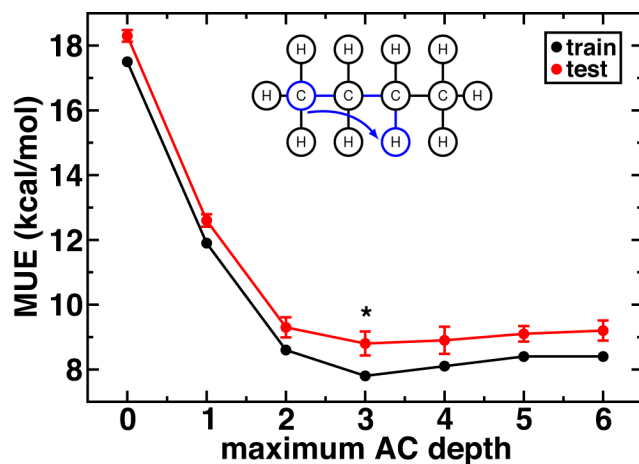


Figure 4-1: Train (black line) and test (red line) MUEs (in kcal/mol) for QM9<sup>31</sup> AEs predicted by KRR models trained on AC feature sets with increasing maximum depth. Each model is trained on 1,000 molecules and tested on the 133k remaining molecules. Error bars on test set error correspond to standard deviations from training the KRR model on five different samples, and the red circles correspond to the mean test error. The lowest MUE maximum-depth, 3, is indicated with an asterisk. An example of a term in a 3-depth AC is shown on butane in inset.

We now compare  $3d$ -AC performance and learning rates (i.e., over increasingly large training sets) to i) the Coulomb matrix eigenspectrum<sup>250</sup> (CM-ES) representation, which is an easy to implement 3D-derived descriptor<sup>79</sup>; ii) the recently-developed<sup>486</sup> 2B descriptor that, like ACs, does not require explicit 3D information and encodes connectivity and bond order information for atom pairs; and iii) and more com-

plex<sup>486</sup> 12NP3B4B descriptors. The 12NP3B4B descriptors, which encode a continuous, normal distribution of bond distances for each bond-type in a system-size invariant manner, require 3D information but have demonstrated performance similar to the best results reported<sup>275,488</sup> QM9 AEs<sup>486</sup>. We trained the CM-ES KRR model using the recommended<sup>79</sup> Laplacian kernel, but we selected a Gaussian kernel for 3d-ACs after confirming it produced lower MUEs (Appendix B, Text B.2). For our ultimate goal of inorganic complex property prediction (Section 4.3), 3D information, even from semi-empirical geometries, is not readily achievable from currently available semi-empirical theories. However, we compare our two trained KRRs to reported performance of 2B and 12NP3B4B descriptors from the literature, which we select as the best-reported 3D-structure-free descriptor and as a high-accuracy, 3D-structure-dependent descriptor, respectively<sup>486</sup>.

For the largest 16000 molecule training set, the 3d-AC test set MUEs are 68% and 43% lower than CM-ES and 2B descriptors, respectively. The 3d-AC descriptors are only outperformed by 12NP3B4B by 74% or 4.5 kcal/mol, owing to the bond distance information encoded in this set (Figure 4-2 and Appendix B, Table B.1). This improved performance of 12NP3B4B and other comparably-performing<sup>486</sup> descriptors (e.g., superposition of atomic densities<sup>427</sup> or the many-body tensor representation<sup>488</sup>) comes at a severe cost of requiring accurate geometries before predictions can be made, whereas 3d-AC significantly outperforms the previous best-in-class topology-only descriptors set 2B. Learning rates (i.e., training-set size test set MUE dependence) are comparable among 3d-AC, 2B, and 12NP3B4B descriptors but slightly steeper for the poorer performing CM-ES representation (Figure 4-2). For dipole moment prediction, 3d-AC performs nearly as well as 12NP3B4B: the 3d-AC test MUE at 1,000 training points is only 2% higher than 12NP3B4B and 19% higher at 16,000 training points (Appendix B, Table B.2). Thus, ACs are promising size-

invariant, connectivity-only descriptors for machine learning of molecular properties. However, we have previously observed limited transferability of organic representations for inorganic complexes<sup>308</sup>, and we next identify the transferability of our present descriptors as well as beneficial inorganic chemistry adaptations.

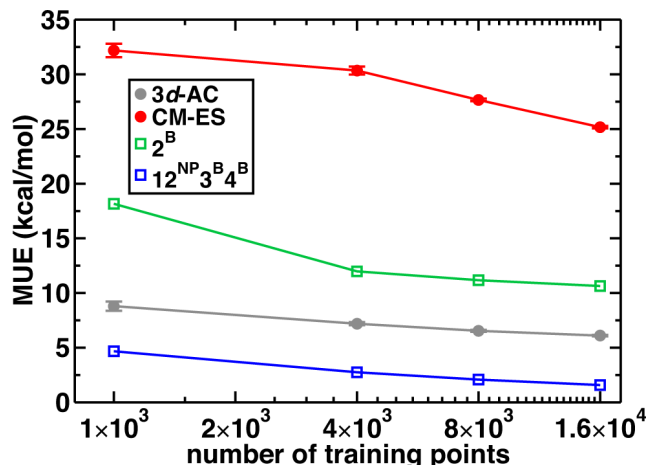


Figure 4-2: Training set size dependence of test set MUEs (in kcal/mol) for KRR model prediction of QM9<sup>31</sup> AEs for four feature sets. In all cases, the test set consists of the remainder of the 134k molecule set not in the training set. For the maximum 3-depth autocorrelation (3d-AC, gray circles) and Coulomb matrix eigenspectrum<sup>250</sup> (CM-ES, red circles) trained in this work, standard deviations (error bars) and mean test errors are reported from training results on five samples selected for each training set size. The 2B (green open square) and 12NP3B4B (blue open square) KRR test set MUEs from literature<sup>486</sup> are provided for comparison.

## 4.2.2 Revised autocorrelations for transition metal complexes

We previously proposed<sup>308</sup> a mixed continuous (e.g., electronegativity differences) and discrete (e.g., metal and connecting atom identity) set of empirical, topological descriptors (referred to in this work as MCDL-25) that emphasized metal-proximal properties for predictive modeling of transition metal complexes with an artificial

neural network. The MCDL-25 set is metal-focused in nature with the longest range effects only up to two bonds through a truncated Kier shape index<sup>453</sup>. This imparted good accuracy (i.e., root mean squared error, RMSE, of 3 kcal/mol) for spin-state splitting predictions and superior transferability to test set molecules with respect to commonly-employed descriptors<sup>250</sup> used in machine learning for organic chemistry that encode complete, 3D information. In addition to standard ACs (eq. 4.1 in Section 4.2.1), we now introduce revised ACs (RACs) inspired by descriptors in the metal-focused MCDL-25 set. In these RACs, we both restrict where the sums in eq. 4.1 start (i.e., to account for potentially greater importance of the metal and inner coordination sphere) and which other atoms are in the scope (Figure 4-3). In the extended notation of the broader AC set, the standard ACs starts on the full molecule (*f*) and has all atoms in the scope (*all*), i.e.,  ${}^f P_d$ . As in ref.<sup>482</sup>, we compute restricted-scope ACs that separately evaluate axial or equatorial ligand properties:

$${}_{ax/eq}{}^f P_d = \frac{1}{|\text{ax/eq ligands}|} \sum_i^{n_{ax/eq}} \sum_j^{n_{ax/eq}} P_i P_j \delta(d_{ij}, d) \quad (4.2)$$

where  $n_{ax/eq}$  is the number of atoms in the corresponding axial or equatorial ligand and properties are averaged within the ligand subtype. We introduce restricted-scope, metal-centered (*mc*) descriptors, in which one of the atoms, *i*, in the *i,j* pair is a metal center:

$${}_{all}{}^{mc} P_d = \sum_i^{mc} \sum_j^{all} P_i P_j \delta(d_{ij}, d) \quad (4.3)$$

For the complexes in this work there is only one metal-center, which simplifies the sum, but there is no inherent restriction to a single metal center (see green arrows in Figure 4-3).



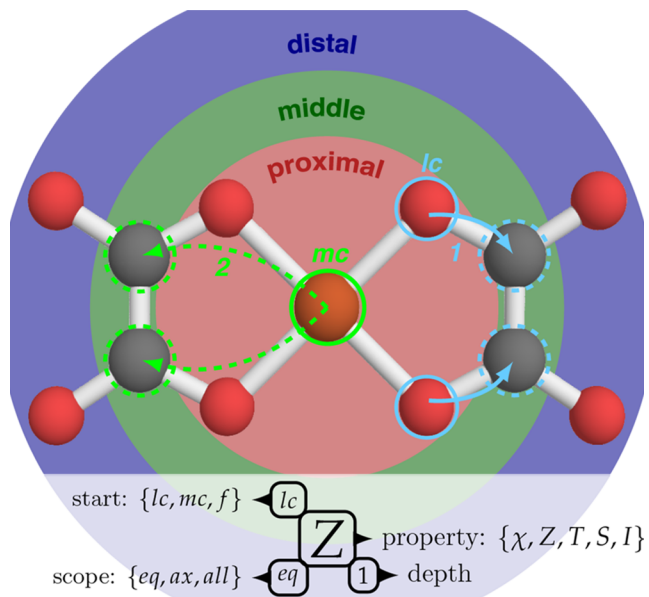


Figure 4-3: Schematic of ACs in the equatorial plane of an iron octahedral complex with two eq. oxalate ligands shown in ball and stick representation (iron is brown, oxygen is red, and carbon is gray). Regions of the molecule used to classify descriptors are designated as proximal (metal and first coordination shell, in red), middle (second coordination shell, in green) and distal (third shell and beyond, in blue) throughout the text. Light green circles and arrows depict terms in a 2-depth  $mc$  RAC (e.g.,  $^{mc}_{all}Z_2$ ), and the light blue circles and arrows depict terms in a 1-depth  $lc$  RAC (e.g.,  $^{lc}_{ax}Z_1$ ).

A second restricted-scope, metal-proximal AC definition is the ligand-centered ( $lc$ ) sum in which one of the atoms,  $i$ , in the  $i,j$  pair is the metal-coordinating atom of the ligand:

$$^{lc}_{ax/eq}P_d = \frac{1}{|ax/eq \text{ ligands}|} \frac{1}{|lc|} \sum_i^{lc} \sum_j^{n_{ax/eq}} P_i P_j \delta(d_{ij}, d) \quad (4.4)$$

We average the ACs over all  $lc$  atoms and over all ligands in order to treat ligands of differing denticity on equal footing (see light blue arrows in Figure 4-3).

Inspired by our previous success<sup>308,489</sup> in employing electronegativity differences be-

tween atoms to predict electronic properties, we also modify the AC definition,  $P'$ , to property differences rather than products for a minimum depth,  $d$ , of 1:

$${}_{ax/eq/all}^{lc/mc}P'_d = \frac{1}{|ax/eq \text{ ligands}|} \frac{1}{|lc|} \sum_i^{lc/mc n_{ax/eq/all}} \sum_j (P_i - P_j) \delta(d_{ij}, d) \quad (4.5)$$

where the scope can be axial, equatorial, or all ligands, the start must be  $lc$  or  $mc$  because a sum of differences over all will be zero, and these ACs are not symmetric so the ordering of indices  $i, j$  is enforced for consistency.

We combine all six types of AC or RAC start/scope definitions ( $f/all$ ;  $mc/all$ ;  $lc/ax$ ;  $lc/eq$ ,  $f/ax$ ; and  $f/eq$ , eqs. 4.1–4.5) with both products and differences of the five atomic properties for depths from zero, where applicable, to maximum depth  $d$ . There are  $6d + 6$  descriptors for six product AC/RACs (eqs. 4.1–4.4) with each of the five atomic properties (i.e., a total of  $30d + 30$  product AC/RACs). For difference RACs (eq. 4.5), there are no zero-depth descriptors, and three non-trivial start/scope definitions ( $mc/all$ ;  $lc/ax$ ; and  $lc/eq$ ), producing  $3d$  descriptors for all of the atomic properties excluding  $I$ , giving  $12d$  difference descriptors for a total of  $42d + 30$  product or difference RACs. These ACs represent a continuous vector space that is increasingly nonlocal with increased maximum  $d$  and dimension invariant with respect to system size. This descriptor set also does not depend on any 3D information, which is valuable for structure prediction<sup>308,482</sup>.

We classify relative locality of ACs into three categories (see Figure 4-3): 1) proximal: depends only on atom types and connectivity in first coordination shell; 2) middle: depends on information from two coordination shells; and 3) distal: all remaining descriptors based on the molecular graph. This broad AC set naturally recovers well-known quantities: i)  ${}_{all}^{mc}I_1$  is the metal coordination number and ii)  ${}_{all}^fI_0$  is the total

number of atoms. We also recover continuous descriptor analogues to the variables in MCDL-2552: i)  ${}_{all}^{mc}Z_0$  is the metal identity, ii)  ${}_{ax/eq}^{lc}Z_0$  is the coordinating atom identity, and iii)  ${}_{all}^{mc}\chi'_1$  is  $\Delta\chi$ . Some ACs are redundant (e.g.,  ${}_{all}^{mc}I_1$  and  ${}_{all}^{mc}T_0$  are the same). Before model training, all ACs are normalized to have zero-mean and unit variance based on training data, and any constant features in training data are filtered out.

### 4.2.3 Feature selection methods

Feature reduction from a large descriptor space improves the ratio of training points to the dimension of the feature space, decreasing training time and complexity<sup>490</sup> for non-linear models (e.g., neural networks) or improving predictions in kernel-based methods with isotopic kernels by eliminating uninformative features. In linear models, feature reduction increases stability, transferability, and out-of-sample performance<sup>490</sup>. Reducing feature space, without impact on model performance<sup>490</sup>, is also useful<sup>491</sup> for providing insight into which characteristics are most important for determining materials properties. Starting from  $n$  observations (e.g., spin-state splitting, bond length, or redox potential) of  $y_{data}(x_i)$  and molecular descriptors  $x_i \in \mathcal{X}_m$  in an  $m$ -dimensional feature space,  $\mathcal{X}_m \subset \mathbb{R}^m$ , we use established<sup>491</sup> feature selection techniques to obtain a lower-dimensional representation of the data,  $\mathcal{X}_d \subset \mathcal{X}_m$  that maximizes out-of-sample model performance while having the smallest possible dimension.

Feature selection techniques may be broadly classified<sup>491</sup> as (Figure 4-4): 1) simple filters, 2) wrapper methods, and 3) direct optimization or shrinkage methods<sup>209</sup>. Type 1 univariate filtering (UVF) acts on each descriptor individually, discarding

those that fail a statistical test (here, the p-value for a linear model being above a cutoff of 0.05). UVF is amenable to very high-dimensional problems<sup>491</sup> but neglects interactions between descriptors that may occur<sup>490</sup>, and the significance test in a linear model may not relate well to the final machine learning model.

Type 2 wrapper methods require multiple steps<sup>490,491</sup>: iterative feature subset choice along with model training and scoring (Figure 4-4). Combinatorial testing of every possible subset is only feasible for small feature sets (e.g., < 40 variables with simple predictive models<sup>209</sup>). The model used in training and scoring is flexible, but the repeated model training time may become prohibitive. Stepwise search<sup>490</sup>, with greedy recursive feature addition or elimination (i.e., RFA or RFE) on most improvement or least penalty, respectively, or randomized searches less prone to local minima<sup>491</sup>, are employed for larger feature sets. Cross-validation (CV) scoring, which is unaffected by feature space size changes, will usually produce a minimum for an optimal number of features<sup>490</sup>. We recently used<sup>482</sup> RFE with an embedded linear model to select variables to use in multiple linear regression (MLR) to identify four key RACs from a larger 28-dimensional space for redox potential prediction. In this work, we primarily employ RFE-MLR to select features to be used for KRR training, despite potentially eroded model transferability between MLR and KRR. The fine hyperparameter grid search needed to produce a robust KRR model at each RFE iteration would take around 30 days in parallel on a 4-core Intel 3.70 GHz Core i7-4820K when starting from a large (ca. 150) descriptor set, making some initial reduction in feature space necessary for practical RFE-KRR (Appendix B, Text B.3).

Type 3 shrinkage or direct optimization methods use regularization (e.g., elastic net or L1-regularized linear regression, LASSO<sup>492</sup>) or a model (e.g. random forests) that determines variable importance in one shot during training, making Type 3 methods much more computationally efficient than Type 2.

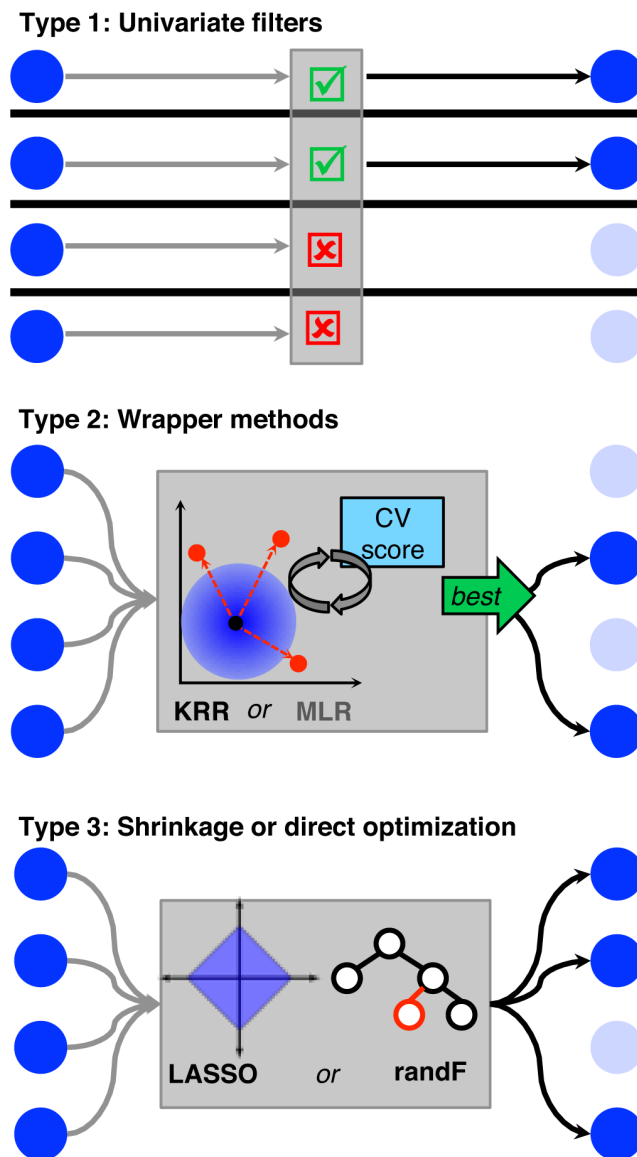


Figure 4-4: Schematic of three main types of feature selection approaches with retained and input features represented by dark blue circles. Type 1 (top) univariate filters evaluate features one at a time; type 2 (middle) wrapper methods train a model (e.g., KRR or MLR) and use a cross validation score to recursively eliminate features; and type 3 (bottom) shrinkage or direct optimization models such as LASSO and random forests (randF) carry out one-shot feature selection and regularization or model training, respectively.

However, it remains uncertain if the typically lower complexity of the combined feature-selection and fitting model (e.g, L1 regularized regression in LASSO) produces results that are transferable to the subsequent ML model to be trained (e.g., KRR). In this work, we use an elastic net, a generalization of LASSO that we previously used to select descriptors for machine learning models<sup>308</sup>, in which a blend of L2 and L1 regularization is applied<sup>493</sup>, giving the loss function as:

$$\mathcal{L}(W) = \|xW - y_{data}(x)\|_2^2 - \lambda (\alpha \|W\|_1 + (1 - \alpha) \|W\|_2^2) \quad (4.6)$$

Here,  $W$  are the regression coefficients,  $\lambda$  is the regularization strength, and  $\alpha$  interpolates between ridge ( $\alpha = 0$ ) and LASSO ( $\alpha = 1$ ) regression. Higher  $\alpha$  aggressively reduces the feature space, and the best  $\alpha$  is selected by cross-validation with  $\lambda$ , with intermediate  $\alpha$  often favored for balancing prediction with feature reduction<sup>209</sup>.

Random forests<sup>494</sup>, which are based on an ensemble of sequential binary decision trees, are another Type 3 method (Figure 4-4). Each tree is trained on a bootstrapped data sample and uses a random input variable set. Integrated feature selection is achieved by comparing tree performance when descriptors are randomly permuted<sup>495</sup> to yield an importance score for each descriptor and discard those below a threshold value. Here, we use 1000 trees and discard descriptors with an increase of  $< 1\%$  (or higher, where specified) in normalized MSE on out-of-model samples upon removal (see convergence details in Appendix B, Figures B-1–B-4).

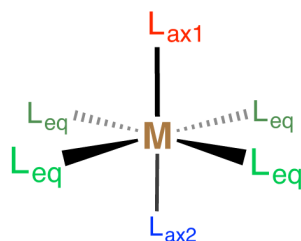
We now compare feature selection methods on our transition metal complex data sets, as judged by performance on 60%–40% and 80%–20% train-test partitions for the larger spin-splitting and smaller redox data set (see Section 4.3.1), respectively. Feature selection is only carried out on the training data, and KRR models are used for judging performance of a feature set using identical cross-validation for hy-

hyperparameter estimation. All analysis is conducted in R version 3.2.3<sup>456</sup>. We use the kernlab<sup>467</sup> package for KR regression, CVST<sup>468</sup> for cross-validation, glmnet<sup>455</sup> for elastic net regression, caret<sup>496</sup> for feature selection wrapper functions and randomForest<sup>497</sup> for random forests. All kernel hyperparameter values are provided in Appendix B, Tables B.3–B.6.

## 4.3 Computational details

### 4.3.1 Organization of data sets

Feature selection and model training is carried out on two data sets of single-site octahedral transition metal complexes, which were generated from extension of data collected in previous work<sup>308,482</sup> (Figure 4-5). These data sets are derived from around 3,300 DFT geometry optimizations of molecules up to over 150 atoms in size, which is a smaller number of training points than has been feasible in machine learning on small (i.e., up to 9 heavy atoms) organic molecules<sup>31</sup> but slightly larger than has successfully been used in bulk catalysis<sup>231</sup>. For both sets, the complexes contain Cr<sup>2+/3+</sup>, Mn<sup>2+/3+</sup>, Fe<sup>2+/3+</sup>, Co<sup>2+/3+</sup>, or Ni<sup>2+</sup> first row transition metals. High-spin (H) and low-spin (L) multiplicities were selected for each metal from those of the isolated ion in National Institute of Standards and Technology atomic spectra database<sup>446</sup>: triplet-singlet for Ni<sup>2+</sup>, quartet-doublet for Co<sup>2+</sup> and Cr<sup>3+</sup>, quintet-singlet for Fe<sup>2+</sup> and Co<sup>3+</sup>, quintet-triplet for Cr<sup>2+</sup> and Mn<sup>3+</sup> (due to the fact that there is no data available for Mn<sup>3+</sup> singlets<sup>446</sup>), and sextet-doublet for Mn<sup>2+</sup> and Fe<sup>2+</sup>.



	spin-split (1345)	redox (226)	
		new (185)	Fe-N (41)
M	Cr,Mn,Fe,Co,Ni	Cr,Mn,Fe,Co	Fe
L types	16	5	41
L CA	C,N,O,S,Cl	C,N,O	N
L denticity	1 to 4	1	1 to 2
symmetry	ax≠eq	ax1≠ax2≠eq	ax=eq
properties	$\Delta E_{H-L}$ , $\min(R_L)$	$\Delta E_{III-II}$ , $E^0$	

Figure 4-5: (top) Schematic of octahedral transition metal complex illustrating possible unique ligands (one equatorial ligand type,  $L_{eq}$ , and up to two axial ligand types,  $L_{ax1}$  and  $L_{ax2}$ ) in the spin-splitting and redox data sets. (bottom) Characteristics of each data set: metal identity, number of ligand types (L types), connecting atom identity of the ligand to the metal (L CA), range of denticities (L denticity), ligand symmetry corresponding to the schematic complex representation, and associated quantum mechanical properties. Spin-splitting and redox Fe-N sets were previously published<sup>308,482</sup>, but the “new” subset of the redox data set was generated in this work.

For all data sets, the molSimplify<sup>181</sup> code was used to generate starting geometries from the above metals and a ligand list (ligands provided in Appendix B, Table B.7). Incompatible ligand combinations are disabled (e.g., equatorial porphyrin ligands can occur once and only with monodentate axial ligands).

The spin-state splitting data set<sup>308</sup> consists of 1345 unique homoleptic or heteroleptic complexes with up to one unique axial and equatorial ligand type with ligands selected from 16 common ligands of variable ligand field strength, connecting atom



identity, and denticity (Figure 4-5). For this data set, the structures were evaluated using hybrid density functional theory (DFT) at 7 percentages of Hartee-Fock (HF) exchange from 0 to 30% in 5% increments. This set was previously used to train models that predict i) the adiabatic, electronic spin-state splitting energy,  $\Delta E_{\text{H-L}}$ , ii) the exchange sensitivity of the spin-state splitting, and iii) the spin-state dependent minimum metal-ligand bond lengths (e.g.,  $\min(R_L)$  or  $\min(R_H)$ ) that differ from the average metal-ligand bond length only for distorted homoleptics or heteroleptic complexes. In this work, we only train and test models on  $\Delta E_{\text{H-L}}$  and  $\min(R_L)$ .

The redox data set (226 unique structures) is comprised of 41 previously studied<sup>482</sup> Fe-nitrogen monodentate and bidentate homoleptic complexes and 185 newly generated structures (Figure 4-5 and Appendix B, Table B.8). The new complexes were obtained by generating combinations of metals (Cr, Mn, Fe, Co) and five small, neutral monodentate ligands (CO, pyridine, water, furan, and methyl isocyanate) with up to two axial ligand types and one equatorial ligand type. Axial ligand disengagement occurred during optimization in several of the 300 theoretically possible cases, reducing the final data set.

In all cases, we calculate the M(II/III) redox couple starting from the adiabatic ionization energy of the reduced complex’s ground state spin:

$$\Delta E_{\text{III-II}} = \Delta E_{\text{III}} - \Delta E_{\text{II}} \quad (4.7)$$

At minimum, this ionization energy requires M(II) low-spin and high-spin geometry optimizations as well as the selected lowest energy M(III) state that differs by a single electron detachment (Appendix B, Table B.9).

To compute the redox potential, we also include solvent and thermodynamic (i.e. vibrational enthalpy and zero point vibrational energy) corrections in a widely adopted

thermodynamic cycle approach<sup>19,203,207</sup>98-100. We estimate the M(II/III) redox potential in aqueous solution at 300K,  $\Delta G_{\text{solv}}$ :

$$\Delta G_{\text{solv}} = G_{\text{gas}}(\text{M(III)}) - G_{\text{gas}}(\text{M(II)}) + \Delta G_s(\text{M(III)}) - \Delta G_s(\text{M(II)}) \quad (4.8)$$

where  $G_{\text{gas}}$  is the gas phase energy with thermodynamic corrections and  $\Delta G_s$  is the solvation free energy of the gas phase structure. We then compute the redox potential:

$$E^0 = -\frac{G_{\text{solv}}}{nF} \quad (4.9)$$

where the number of electrons transferred is  $n = 1$  and  $F$  is Faraday’s constant.

### 4.3.2 First-principles simulation methodology

Our simulation methodology was the same for all generated data sets. All DFT calculations employ the B3LYP hybrid functional<sup>152,157,158</sup> with 20% HF exchange ( $a_{\text{HF}} = 0.20$ ), except for cases where HF exchange is varied<sup>197</sup> while holding the semi-local DFT exchange ratio constant. In inorganic complexes, the optimal amount of HF exchange is highly system dependent<sup>194,196,197,437</sup>, motivating our earlier training of an ANN to predict spin-state ordering and bond length in an HF exchange dependent manner as well as the sensitivity of properties to HF exchange fraction<sup>308</sup>. Exchange-sensitivity is not the focus of the present work, as our prior work demonstrated<sup>308</sup> that ANN accuracy was not sensitive to functional choice. We use the LANL2DZ effective core potential<sup>172</sup> for all transition metals, bromine, and iodine and the 6-31G\* basis for the remaining atoms. The use of a modest basis set is motivated by our previous observations<sup>482</sup> that extended basis sets did not substantially

alter trends in redox or spin-state properties. Gas phase geometry optimizations were conducted using the L-BFGS algorithm implemented in the DL-FIND<sup>449</sup> (for the spin-splitting data set) or in translation rotation internal coordinates<sup>498</sup> (for the redox data set) interfaces to TeraChem<sup>105,447</sup> to the default tolerances of  $4.5 \times 10^{-4}$  hartree/bohr for the maximum gradient and  $1 \times 10^{-6}$  hartree for the change in self-consistent field (SCF) energy between steps. All calculations were spin-unrestricted with virtual and open-shell orbitals level-shifted<sup>448</sup> by 1.0 eV and 0.1 eV, respectively, to aid SCF convergence to an unrestricted solution. Deviations of  $\langle S^2 \rangle$  from the expected value by more than 1  $\mu\text{B}$  led to exclusion of that data point from our data set. The aqueous solvent environment, where applicable, was modeled using an implicit polarizable continuum model (PCM) with the conductor-like solvation model (COSMO<sup>205,499</sup>) and  $\epsilon = 78.39$ . The solute cavity was built using Bondi’s van der Waals radii<sup>500</sup> for available elements and 2.05 Å for iron, both scaled by a factor of 1.2. Vibrational entropy and zero-point corrections were calculated from a numerical Hessian obtained with built-in routines in TeraChem<sup>105,447</sup>.

## 4.4 Results and discussion

### 4.4.1 Spin splitting energy

We evaluate our RACs (i.e., both standard ACs and the modified start, scope, and difference ACs defined in Section 4.2.2) for KRR training on the spin-splitting data set and compare to both previous MCDL-25 descriptors<sup>52</sup> and widely-employed<sup>79,250</sup> Coulomb-matrix-derived descriptors. Based on our results for organic molecules (Section 4.2.1), we use a maximum depth of 3 in the  $42d + 30$  RACs, producing 156 potential descriptors, which reduce to 151 after discarding 5 descriptors that are

constant (e.g.,  ${}^{lc}I_0$  and  ${}^{mc}T_0$ ) due to unchanged octahedral coordination in the data sets in this work (Appendix B, Table B.10). We add four variables (i.e., oxidation state, HF exchange and axial/equatorial ligand denticity) from our MCDL-25 set<sup>308</sup> to produce a final 155-variable set (RAC-155). The RAC-155 set is transferable to inorganic chemistry, with already good MCDL-25/KRR (Gaussian kernel) test set RMSE and MUE of 3.88 and 2.52 kcal/mol reduced to 1.80 and 1.00 kcal/mol with RAC-155 (Table 4.1). This performance is also superior to Coulomb matrix (CM)-based descriptors computed on high-spin geometries. Using either i) an L1 matrix difference kernel on sorted Coulomb matrices<sup>137,250</sup> (CM-L1) or ii) eigenvalues<sup>250</sup> and a Laplacian kernel, as recommended in Ref.<sup>79</sup> (CM-ES), we obtain 10-30× higher RMSE and MUEs than for RAC-155 or MCDL-25 (Table 4.1, learning rates for RAC-155 in Appendix B, Figure B-5).

Table 4.1: Coulomb matrix eigenspectrum representation with a Laplacian kernel (CM-ES)<sup>25</sup>, our prior hybrid discrete-continuous descriptors (MCDL-25)<sup>308</sup> with a Gaussian kernel, and the full RAC-155 set introduced in this work with a Gaussian kernel.

Feature set	RMSE (kcal/mol)	MUE (kcal/mol)
CM-L1	30.80	20.84
CM-ES	19.19	14.96
MCDL-25	3.88	2.52
RAC-155	1.80	1.00

Visualization with principal component analysis (PCA) of the key descriptor space dimensions with spin-splitting or molecular size variation overlaid reveals why CM-ES performs poorly in comparison to RACs (Figure 4-6). The first two principal

components encode the majority of the feature space variation for both sets: 85% of CM-ES and 55% of RAC-155 (Appendix B, Figures B-6–B-7). As expected<sup>250,308</sup> CM-ES shows excessive molecule-size-dependent clustering that is not predictive of how metal electronic properties vary. As an example, homoleptic Fe(III) complexes with strong-field t-butylphenylisocyanide (pisc) and methylisocyanide (misc) ligands have comparable  $\Delta E_{\text{H-L}}$  of 41 and 38 kcal/mol but differ in size substantially at 151 and 37 atoms, respectively (structures in Figure 4-6 inset). Despite comparable spin splitting, these molecules are on opposite ends of PC1 in the CM-ES PCA with no intermediate data (Figure 4-6). More broadly, no clustering is apparent in spin-splitting energies with CM-ES in comparison to the strong system size clustering (Figure 4-6).

In contrast, RAC-155 distributes data more evenly in the PCA with smaller size-dependence due to using both metal-centered and ligand-centered ACs in addition to truncating the depth of descriptors to three prior to feature selection (Figure 4-6). Improved RAC performance is also due to better representation of molecular similarity with apparent weak-field and strong-field groupings, assisting KRR learning<sup>25</sup> that relies on nearest neighbor influence for property prediction (Figure 4-6).

Spin splitting energies are well predicted by KRR with RAC-155, outperforming our previous MCDL-25 representation but at the initial cost of an order-of-magnitude increase (from 25 to 155) in feature space dimension. We thus apply feature selection techniques (Section 4.2.3) to identify if AC subsets maintain predictive capability with smaller feature space size.

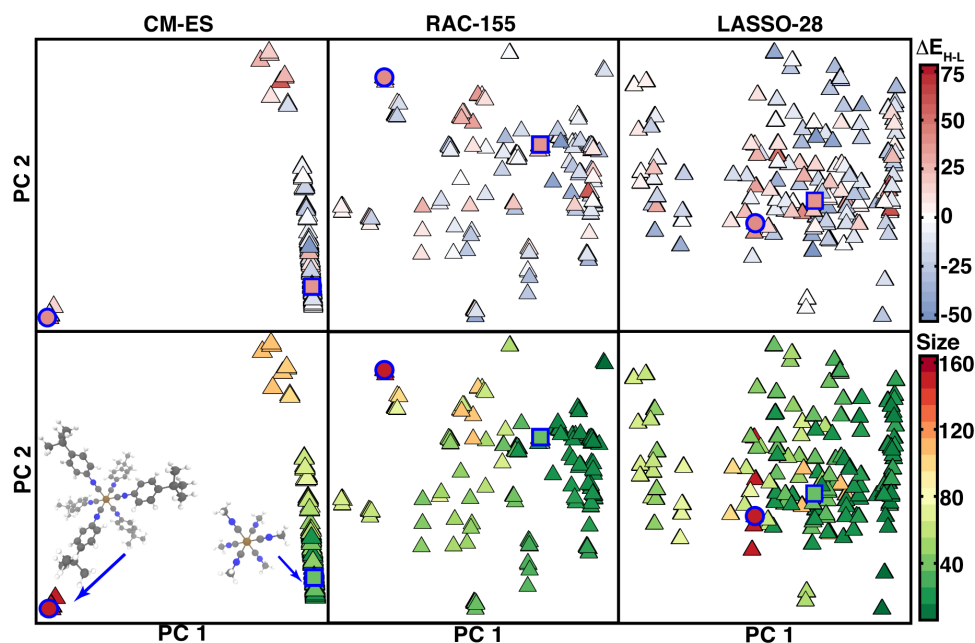


Figure 4-6: Projection of spin-splitting data set onto the first two principal components (arbitrary units) for the Coulomb matrix eigenspectrum (CM-ES, left), full revised AC set (RAC-155, center), and the LASSO-selected subset (LASSO-28, right). The PCA plots are colored by DFT-calculated spin splitting energy (top, scale bar in kcal/mol at right) and size (bottom, scale bar in number of atoms at right). Ball and stick structures of representative complexes Fe(III)(pisc)<sub>6</sub> and Fe(III)(misc)<sub>6</sub> (iron brown, nitrogen blue, carbon gray, and hydrogen white) are inset in the bottom left, and the associated data points are highlighted with a blue circle and square, respectively, in each plot.

Starting with Type 3 shrinkage methods we have previously employed<sup>308</sup>, we carried out feature selection with an elastic net. Comparable CV scores were obtained for all  $\alpha$ , and so we chose  $\alpha = 1$  (i.e., LASSO) (Appendix B, Figure B-8). LASSO retained 28 features, eliminating over 80% of the features in RAC-155 with a 0.2 kcal/mol decrease in test RMSE and the best overall, sub-kcal/mol MUE (Table 4.2 and Appendix B, Table B.11). PCA on LASSO-28 reveals even weaker size dependence than RAC-155 and closer pisc and misc species in PC space (Figure 4-6).

Table 4.2: Train and test set KRR model prediction errors (RMSE for train/test and MUE for test) for spin-splitting energy (in kcal/mol) for RAC-155 and down-selected subsets based on spin-splitting data using LASSO, univariate filters (UV), recursive feature elimination (RFE) based on MLR, and random forest (randF). The last results presented for comparison are the common feature subset (RAC-12), a proximal-only subset (PROX-23) of RAC-155, and the full RAC-155.

Feature set	train		test
	RMSE	RMSE	MUE
	(kcal/mol)	(kcal/mol)	(kcal/mol)
LASSO-28	0.60	1.65	0.96
UV-86	0.43	1.78	0.99
RFE-43	0.41	2.50	1.20
randF-41	0.40	1.87	1.01
randF-26	1.18	2.12	1.28
RAC-12	1.31	2.90	1.86
PROX-23	5.43	6.03	3.70
RAC-155	0.55	1.80	1.00

Type 1 feature selection with UV filters ( $p \leq 0.05$ ) retains 86 features (UV-86, Appendix B, Table B.12) and comparable performance to RAC-155, suggesting elimination of descriptors that have weak univariate correlation does not reduce KRR accuracy (Table 4.2 and Appendix B, Figure B-9). Type 3 RFE with an embedded MLR model produces a flat CV error, with an absolute CV minimum at 43 retained features (i.e., RFE-43, Appendix B, Table B.13 and Figure B-10). The RFE-43 KRR model shows 0.5 kcal/mol and 0.2 kcal/mol worsened test RMSE and MUE, respectively, compared to RAC-155. Improved performance could possibly be obtained with a higher fidelity embedded model but at the cost of prohibitive computational

time for feature selection (see Section 4.2.3).

In addition to LASSO, we also employed the Type 3 random forest (randF) model, which has a suggested 1% MSE cutoff for feature selection, and by varying this cutoff we can vary feature set size. The standard 1% cutoff with random forest selects 41 features (randF-41), yielding KRR test RMSE/MUE within 0.1 kcal/mol of RAC-155 (Table 4.2 and Appendix B, Figure B-11 and Table B.14). We also truncate at 2% randF MSE to retain only 26 variables (randF-26), favorably reducing the feature space but slightly worsening test MUE relative to randF-41 or LASSO-28 by 0.2–0.3 kcal/mol, with other cutoffs yielding no KRR test error improvement (Table 4.2 and Appendix B, Tables B.15–B.16). In addition to average errors, error distributions are symmetric, and maximum errors track with RMSE/MUE: LASSO-28 yields the smallest ( $< 9$  kcal/mol) maximum error (Appendix B, Figure B-12).

The best-performing LASSO-28 set contains some features equivalent to those in MCDL-25: i) LASSO-28  $^{mc}_{all}\chi'_2$  and  $^{mc}_{all}\chi'_3$  are similar to MCDL-25  $\Delta\chi$ , ii) LASSO-28  $^{mc}_{all}S'_1$ ,  $^{mc}_{all}Z_1$ , and  $^{lc}_{ax}\chi_1$  encode the size and identity of the ligand connecting atoms also present in MCDL-25, and iii) metal-identity, which was a discrete variable in MCDL-25, is represented by  $^{mc}_{all}Z_0$  and  $^{mc}_{all}\chi_0$  in LASSO-28. Our new difference-type RACs are well-represented (10 of 28 in LASSO-28), and only 5 of 28 are whole-ligand(4, e.g.,  $^f_{ax}I_3$ ) or whole-complex (1,  $^f_{ax}\chi_2$ ). Thus,  $mc$ ,  $lc$ , and difference-derived RACs, all motivated by our prior observations of inorganic chemistry, are key to high accuracy predictions.

It is useful to understand the effect of feature selection method choice by identifying the number of common features among the three best-performing selected feature sets, LASSO-28, UV-86, and randF-41 (Figure 4-7). Only 12 features are common to the three subsets, which we designate RAC-12 (Appendix B, Table B.17). In RAC-12, 7 of the retained descriptors are proximal, and 5 of 12 descriptors incorporate  $\chi$



or  $\Delta\chi$ . All four of the retained distal properties in RAC-12 (e.g.,  $^{mc}_{all}\chi'_d$ ,  $d = 1, 2, 3$  and  $^{mc}_{all}S'_1$ ) are of the newly introduced difference-derived AC type. A KRR model trained on RAC-12 produces test set RMSE and MUE 1.1 and 0.9 kcal/mol above the  $13\times$  larger RAC-155 but still significantly lower than the twice as large MCDL-25 (see Table 4.1 and 4.2). Broadly, two thirds of all features are selected by at least one of the three best feature selection methods (Figure 4-7).

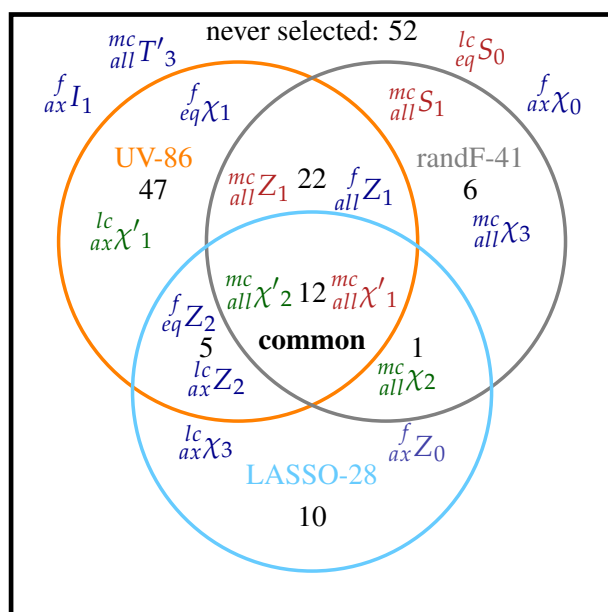


Figure 4-7: Venn diagram showing common descriptors among the three best performing subsets of RAC-155 returned by feature selection algorithms: UV-86, LASSO-28, and randF-41. A total of 12 common variables are found among all three sets, and other numbers refer to unique or common variables between sets. Example features are indicated, colored by classification (proximal in red, middle in green, and distal in blue).

Over 80% of the descriptors in randF-41 are also found in the larger UV-86, but fewer (31% of randF-41) are present in the smaller LASSO-28. Unique descriptors in randF-41 are  $mc$ -type, whereas unique LASSO-28 descriptors are non-local 2-depth

or 3-depth standard ACs on ligands.

We further classify the degree of locality in each feature set, as designated by the bond-wise path-length scales of information in the descriptors (i.e., proximal, middle, and distal, defined in Section 4.2.2). We quantify the fraction of descriptors corresponding to each category in a feature set, e.g. the proximal fraction:

$$\text{frac (proximal)} = \frac{\text{num. of proximal RACS} + 2}{\text{num. RACS} + 2} \quad (4.10)$$

where the denominator only contains the RACs that can be assigned to proximal (the two ligand denticity variables are also included here), middle, or distal portions of the molecule, not oxidation state or HF exchange. Relative to RAC-155, all feature selection methods increase the proximal fraction, and we observe lowest MUEs in subsets with higher proximal fractions, i.e., over 0.3 in the best-performing LASSO-28 or in randF-41 and increased to nearly 0.5 when a higher MSE cutoff is used in random forest (i.e., randF-26, Figure 4-8). The higher-dimensional Type 1 UV-86 subset and Type 2 RFE-43 subset possess the most similar distributions to RAC-155 with still good performance likely due to relatively large feature set size (Figure 4-8). Modest feature space dimension ( $< 30$ ) always gives higher proximal fraction than larger subsets.

Given the high fraction of retained proximal descriptors in randF-26 and RAC-12, we also tested the suitability of a full proximal-only set of RACs and denticity variables along with oxidation state and HF exchange (PROX-23) for KRR model training (Appendix B, Table B.18). This PROX-23 KRR model is the worst performing of all KRR models, including MCDL-25, with test RMSE and MUE of 6.0 and 3.7 kcal/mol, emphasizing the importance of beyond-proximal information present in both MCDL-25 and the feature sets selected in this work (Table 4.2).

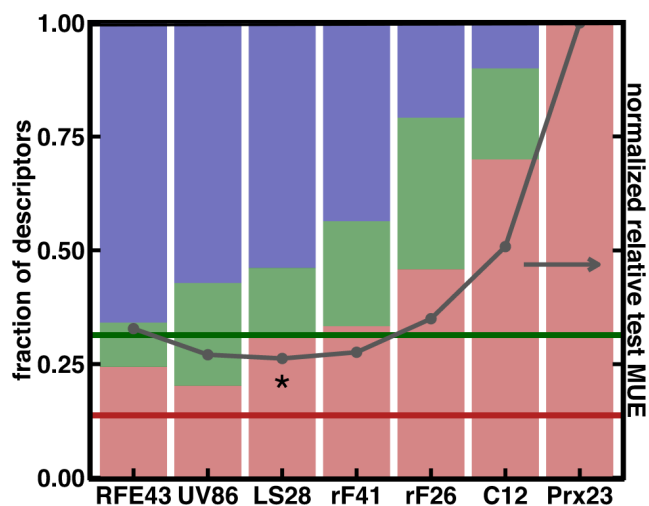


Figure 4-8: Fraction of selected descriptors that are proximal (red), middle (green) or distal (blue), as defined in the main text and depicted in Figure 4-3 compared against RAC-155 reference fractions (dark red proximal fraction and green middle fraction as horizontal lines) along with their performance for spin-splitting prediction with KRR. The normalized relative test set spin-splitting MUE from a KRR model is shown in dark grey for each set, and the lowest test MUE is indicated with an asterisk. Sets are sorted left to right in decreasing distal fraction: RFE with MLR (RFE43); UV filter (UV86); LASSO (LS28), random forest with 1% (rF41) or 2% cutoff (rF26), common set (C12), and proximal-only (Prx23). HF exchange and oxidation state are not shown but are used in all models

The superior performance of the LASSO-28 subset over the similarly-sized randF-26 also highlights the importance of second-shell and global descriptors, as 78% of the 18 features present in LASSO-28 that are absent from randF-26 are distal (e.g.,  ${}^{mc}_{all}\chi'_3$ ,  ${}^{lc}_{eq}\chi'_2$ , and  ${}^{lc}_{ax}T_3$ ). Comparing randF-26 to the larger randF-41 set, which has a 0.3 kcal/mol lower test MUE, we observe that 12 of the 15 features present in randF-41 but omitted in randF-26 are distal.

## 4.4.2 Descriptor transferability to bond length prediction

A key advantage of our geometry-free RACs is that they enable bond length prediction<sup>308</sup> to facilitate accurate structure generation<sup>180,181</sup>. We first evaluate the predictive performance of our full AC set (RAC-155), the proximal subset (PROX-23), and spin-state splitting-selected feature sets (LASSO-28, randF-41, and randF-26) as well as the common subset (RAC-12) for training KRR models on minimum low-spin metal-ligand bond lengths (i.e.,  $\min(\text{RL})$ ) in the low-spin, DFT geometry-optimized structures of complexes in the spin-splitting data set. If the complex is homoleptic and symmetric, there is only a single metal-ligand bond length in the low-spin complex that corresponds exactly to  $\min(R_L)$ , otherwise we take the minimum of the equatorial or axial metal-ligand bond length in order to predict a single property.

Except for PROX-23, all feature subsets yield RMSEs and MUEs around 1.4 and 0.5 pm (i.e., 0.014 Å and 0.005 Å), respectively, with RAC-12 performing nearly as well (test RMSE: 1.6 pm, MUE: 0.6 pm) (Table 4.3). The overall best RMSE performance is observed for LASSO-28, better than for RAC-155, and all subsets have very slightly degraded (i.e., 0.05 pm worse) MUE performance compared to RAC-155 (Table 4.3). The PROX-23 set yields 2–3× larger errors (test RMSE: 2.7 pm and MUE: 1.8 pm), which is significantly worse than the smaller common set (RAC-12), indicating the critical importance of middle and distal features (Figure 4-9). Nevertheless, nearly all feature sets yield better prediction with a KRR model than our prior, proximally-weighted MCDL-25 set (neural network test RMSE: 2 pm)<sup>308</sup>.

Table 4.3: Train and test set KRR model prediction errors (RMSE for train/test and MUE for test) for minimum low-spin bond length (in pm) for down-selected subsets of RAC-155 using LASSO and random forest (randF) on bond length data (denoted with suffix “B”) shown first, as well as original spin-splitting feature sets (LASSO-28, randF-41, and randF-26), shown next. The randF-49B contains manually added HF exchange, which is excluded from automatically selected randF-48B. The last results presented for comparison are the common feature subset (RAC-12), a proximal-only subset (PROX-23) of RAC-155, and the full RAC-155.

Feature set	train	test	
	RMSE (pm)	RMSE (pm)	MUE (pm)
LASSO-83B	0.15	1.33	0.42
randF-48B	1.25	2.06	1.21
randF-49B	0.18	1.34	0.45
LASSO-28	0.12	1.28	0.47
randF-41	0.16	1.38	0.47
randF-26	0.20	1.37	0.48
RAC-12	0.16	1.62	0.59
PROX-23	2.37	2.67	1.76
RAC-155	0.16	1.33	0.42

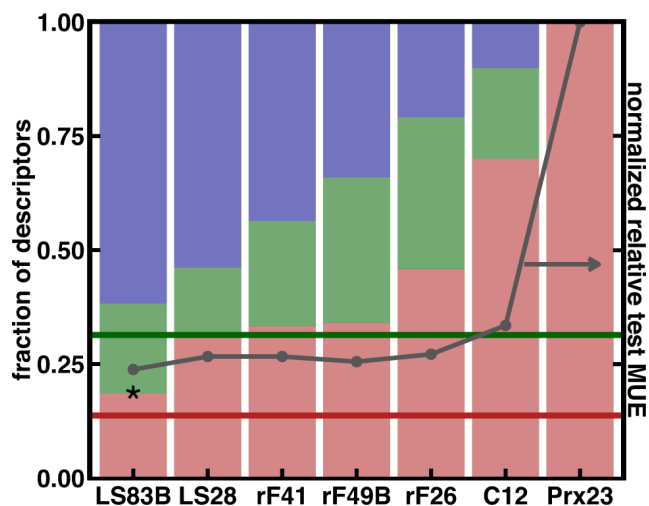


Figure 4-9: Fraction of selected descriptors that are proximal (red), middle (green) or distal (blue), as defined in the main text and depicted in Figure 4-3 compared against RAC-155 reference fractions (dark red proximal fraction and green middle fraction as horizontal lines) along with their performance for low-spin bond length prediction with KRR. The normalized relative test set bond length MUE from a KRR model is shown in dark grey for each set, and the lowest test MUE is indicated with an asterisk. Sets are sorted left to right in decreasing distal fraction: LASSO on bond length (LS83B) or on spin-splitting data (LS28); random forest on spin-splitting (1%, rF41), on bond length data (1%, rF49B), higher cutoff on spin-splitting (or 2%, rF26); the spin-splitting-derived common set (C12); and proximal-only (Prx23). HF exchange and oxidation state are not shown but are used in all models.

We also carried out feature selection on the bond length data with LASSO and random forest to obtain new feature sets (denoted with a “B” suffix). With bond length data, LASSO and random forest retain larger feature sets of 83 and 48 RACs, respectively (LASSO-83B and randF-48B in Table 4.3, and Appendix B, Tables B.19–B.20 and Figures B-13–B-14). In KRR model training, LASSO-83B performs exactly the same as RAC-155 with half the features, whereas randF-48B has 2-3x larger errors (test RMSE: 2.1 pm, MUE: 1.2 pm). This degraded randF-48B performance occurs because HF exchange has been dropped at the 1% MSE random forest cutoff, producing a discontinuous jump in kernel hyperparameters (Table 4.3 and Appendix

B, Table B.4 and Figure B-14). The indirect effect of HF exchange on bond length within a single complex is apparent<sup>197</sup>, but across a wide data set of complexes, the role of HF exchange in bond length data is more easily missed by random forest than in the case of spin splitting. Manually adding HF exchange to the feature set (randF-49B) makes this set perform comparably in KRR model training to the other feature subsets (Table 4.3).

Comparison of random forest feature sets selected on bond lengths (randF-49B) and on spin splitting (randF-41) reveals differences in the underlying structure-property relationships. Both sets have 34 features in common, with an increased proximal fraction relative to RAC-155, but there is a slight bias toward middle features for the bond-length selected set (15 middle in randF-49B instead of 9 in randF-41) (Figure 4-9). The 15 unique features present in randF-49B but absent from randF-41 are weighted toward topological, size-derived effects with 5 T-type (e.g.,  $^{mc}_{all}T_1$ ), 2 I-type (e.g.,  $^{lc}_{ax}I_1$ ), and 4 S-type (e.g.,  $^{lc}_{eq}S_0$ ) RACs. Conversely, four of the seven features in randF-41 but absent from randF-49B are middle/distal and  $\chi/Z$ -type (e.g.,  $^{mc}_{all}Z'_3$  and  $^{lc}_{eq}\chi'_1$ ). Comparable KRR bond length prediction accuracy with both feature sets is due to similar data clustering: the ten nearest complexes to Fe(III)(pisc)<sub>6</sub> are largely unchanged between randF-49B and randF-41, but would differ substantially for RAC-12 and PROX-23 (Appendix B, Table B.21). Thus evaluation of random forest feature set selection reveals structure-property-error relationships that may not be apparent from evaluating KRR model errors alone.

### 4.4.3 Descriptor transferability to redox data

We now test the transferability of RAC descriptor sets to our redox data set for the prediction of M(II/III) gas phase ionization potentials (IPs) and aqueous redox

potentials (see Figure 4-5). Here, all calculations are with B3LYP (20% exchange), and the oxidation state is no longer a fixed variable. Therefore, all feature sets have two fewer variables, but we retain the sets' original names. It might be expected that direct gas phase IPs are easier to learn than redox potentials, which incorporate composite and potentially opposing solvent and thermodynamic effects. However, we observe qualitatively similar KRR model performance and feature selection trends, and we thus summarize gas phase IP results briefly (Appendix B, Text B.4, Tables B.22–B.24, and Figures B-15–B-16). After removal of a single outlier molecule, RAC-155 yields test set RMSE and MUE values of 0.46 and 0.35 eV, respectively, or a 3% or 2% error relative to the 14.4 eV data set mean, and spin-splitting-selected subsets randF-41 or LASSO-28 produce the next lowest but slightly larger errors (Appendix B, Table B.24 and Figures B-17–B-19).

Redox potentials (i.e., including thermodynamic and aqueous implicit solvent corrections) in the full redox data set range from 3.3 to 10.4 eV with a mean of 6.7 eV, and the gas phase IP outlier is not a redox potential outlier (Appendix B, Figure B-18). The full RAC-155 set produces lower absolute errors with respect to gas phase IP (test: RMSE 0.40 eV, 6% error and MUE: 0.32 eV, 5% error) but higher relative errors due to the lower data set mean (Table 4.4). Feature selection on redox potentials from the redox data set retains 19 variables with LASSO (LASSO-19G), comparable to the size selected on gas phase IP but smaller than feature sets selected by LASSO on spin-splitting or bond length (Appendix B, Figure B-20 and Table B.25). LASSO-19G improves very slightly over RAC-155 (test RMSE: 0.38 eV and MUE: 0.31 eV), despite being 12% of the size of the full set (Table 4.4). Random forest on redox potential retains 38 features (randF-38G), improving over both LASSO-19G and RAC-155 (test RMSE: 0.31 eV, 5% error and MUE: 0.26 eV, 4% error) (Table 4.4 and Appendix B, Figure B-21 and Table B.26). Thus, comparable



or reduced absolute errors and only slightly increased relative errors indicates that the combination of ionization potential, solvent, and thermodynamic corrections is only slightly more challenging to capture than IP alone.

Table 4.4: Train and test set KRR model prediction errors (RMSE for train/test and MUE for test) for redox potential (in eV) for down-selected subsets of RAC-155 using LASSO and random forest (randF) on redox data (denoted with suffix “G”) shown first, as well as original spin-splitting feature sets (LASSO-28, randF-41, and randF-26), shown next. The last results presented for comparison are the common feature subset (RAC-12) from all methods, a proximal-only subset (PROX-23) of RAC-155, and the full RAC-155.

Feature set	train		test
	RMSE	RMSE	MUE
	(pm)	(pm)	(pm)
LASSO-83B	0.15	1.33	0.42
randF-48B	1.25	2.06	1.21
randF-49B	0.18	1.34	0.45
LASSO-28	0.12	1.28	0.47
randF-41	0.16	1.38	0.47
randF-26	0.20	1.37	0.48
RAC-12	0.16	1.62	0.59
PROX-23	2.37	2.67	1.76
RAC-155	0.16	1.33	0.42

Evaluating the spin-splitting-selected feature subsets (LASSO-28, randF-41, and randF-26) and the common set (RAC-12) on the redox data set for redox potential prediction produces some of the lowest test errors of all sets (Table 4.4). The spin-splitting-selected randF-26 performs best (test RMSE: 0.29 eV, 4% error and

MUE: 0.23 eV, 3% error), with the larger randF-41 performing nearly as well, whereas LASSO-28 has larger errors (e.g., test MUE of 0.35 eV) more comparable to RAC-155. The RAC-12 set exhibits its best relative performance for any property prediction so far (test RMSE: 0.37 eV and MUE: 0.32 eV), equivalent to the 13× larger full RAC-155 and substantially better than the proximal-only PROX-23 (test MUE: 0.78 eV, Table 4.4). The better performance of spin-splitting-selected sets on redox data could be due to i) the larger, more diverse data in the spin-splitting training set or ii) that our redox calculation implicitly requires knowledge of spin, as the redox potential is always evaluated from the ground state of the reduced species. However, separate prediction of high- or low-spin redox potentials yields similar accuracy, suggesting combined ground state and redox potential prediction does not increase the difficulty of the learning task (Appendix B, Table B.27).

Within the redox potential prediction subsets, a relationship between the prediction accuracy and fraction of descriptor type (i.e., proximal vs. distal) is less clear than for spin splitting or bond length (Figure 4-10). Simultaneously comparing locality and test set MUE across feature sets shows comparable performance for i) randF-38G with a proximal fraction below that of RAC-155, ii) the relatively high proximal and middle fractions in randF-26, and iii) and even relatively good performance in the RAC-12 minimal, proximal-heavy subset (Figure 4-10). Comparing the poorer performing spin-splitting-selected LASSO-28 to the redox-selected LASSO-19G reveals missing middle/distal S- or I-type RACs (e.g.,  ${}_{ax/eq}^{lc}I_3$ ,  ${}_{ax/eq}^{lc}S'_1$ ) in the former.

Examining descriptors in the better-performing, redox-selected randF-38G that are absent from similarly-sized spin-splitting-selected randF-41 reveals 10 T-type and 3 I-type RACs, seven lc 3-depth RACs, and two whole-ligand  ${}_{eq}^f\chi_1$  and  ${}_{eq}^f\chi_0$  RACs, indicating a preference for whole-complex-derived, and, in particular, connectivity information, consistent with observations of the importance of whole-ligand RACs in

redox potentials<sup>482</sup>. Comparing instead the 17 common features in randF-38G and randF-41 reveals mostly *mc* RACs (e.g.,  ${}^{mc}_{all}Z_0$  and  ${}^{mc}_{all}\chi'_1$ ) similar to the metal and connecting atom information in MCDL-25<sup>308</sup>.

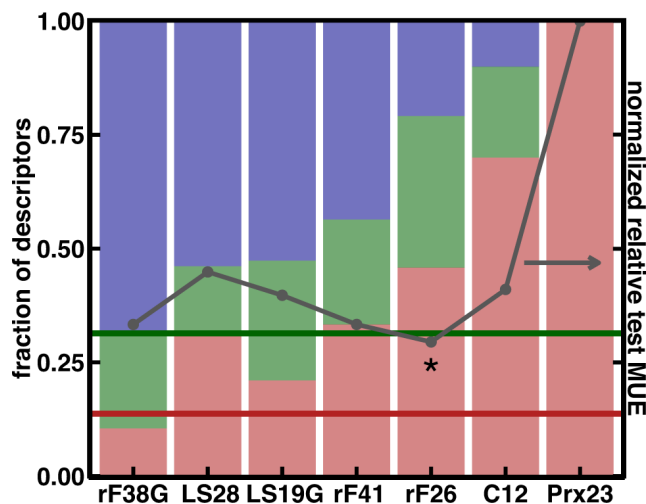


Figure 4-10: Fraction of selected descriptors that are proximal (red), middle (green) or distal (blue), as defined in the main text and depicted in Figure 4-3 compared against RAC-155 reference fractions (dark red proximal fraction and green middle fraction as horizontal lines) along with their performance for redox potential prediction with KRR. The normalized relative test set redox potential MUE from a KRR model is shown in dark grey for each set, and the lowest test MUE is indicated with an asterisk. Sets are sorted left to right in decreasing distal fraction: random forest on redox potential (rF38G); LASSO on redox potential (LS19G) or spin-splitting (LS28); random forest on spin-splitting (1%, rF41 or 2%, rF26); spin-splitting common set (C12); and proximal-only (Prx23). HF exchange and oxidation state are not used in any models

#### 4.4.4 Overall comparison of best descriptor subsets

Overall, Type 3 LASSO or random forest methods have provided the best price-performance trade-off for feature selection in KRR model training of transition metal complex properties on the data sets studied in this work. Although LASSO-28 pro-

duced the lowest KRR model test MUE of 0.96 kcal/mol, randF-41 (1% cutoff) and randF-26 (2% cutoff) produce similarly good 1.01–1.28 kcal/mol test MUEs on the spin-splitting data set and demonstrate somewhat better transferability to redox potential prediction on the redox data set. All three of these subsets are accurate for low-spin bond length prediction, with 1.3-1.4 pm test RMSE and 0.5 pm MUE that is only slightly worse relative to larger, bond-length-selected feature sets, randF-49B or LASSO-83B. The best redox potential prediction performance is achieved not with redox-selected randF-38G (test MUE: 0.26 eV), LASSO-19G (test MUE: 0.31 eV), or even the full RAC-155 (test MUE: 0.33 eV), but with the smaller spin-splitting selected randF-26 (test MUE: 0.23 eV). As an overall recommendation, we thus would select randF-26 for broad spin-splitting, bond length, and gas phase IP/redox potential prediction or LASSO-28 for only spin-splitting and bond length prediction.

To explore how feature space topology differs when using spin-splitting-selected features (randF-26 or randF-41) versus redox-selected features (randF-38G), we consider the example of Fe(II/III) complexes with triazolyl-pyridine ligands from the redox data set. In two cases, these homoleptic, bidentate complexes have a methyl group on the carbon adjacent to pyridinyl nitrogen (ligand 9,  $E^0 = 6.1$  eV and ligand 23,  $E^0 = 6.0$  eV), but in one case the methyl group is in the meta position with respect to the metal-coordinating pyridinyl nitrogen (ligand 8 with  $E^0 = 5.5$  eV) (Appendix B, Figure B-22). Ligands 8 and 9 contain a 1,2,3-triazole, whereas ligand 23 contains 1,2,4-triazole. Within randF-26 and randF-41, the high fraction of proximal or middle mc descriptors emphasizes differences between 1,2,3-triazole and 1,2,4-triazole rather than capturing the importance of the ligand-connecting atom adjacent methyl group. The additional distal *T*-, *I*- and *S*-type descriptors in randF-38G increase the relative importance of the metal-adjacent methyl groups over the order of ring substituents, correctly identifying the nearest neighbor of the ligand 9 complex as

the ligand 23 complex (Appendix B, Text B.5).

Although we have identified a feature set that is transferable across multiple properties when paired with a KRR model, there are still noteworthy differences in optimal feature sets obtained from random forest (i.e., spin-splitting randF-26/41, bond-length randF-49B, and redox randF-38G) that can inform our understanding of the degree of locality and nature of features needed for differing property prediction. To simplify this analysis, we classify  $\chi$ - and  $Z$ -derived RACs as electronic and  $S$ -,  $I$ -, and  $T$ - as topological (Figure 4-11).

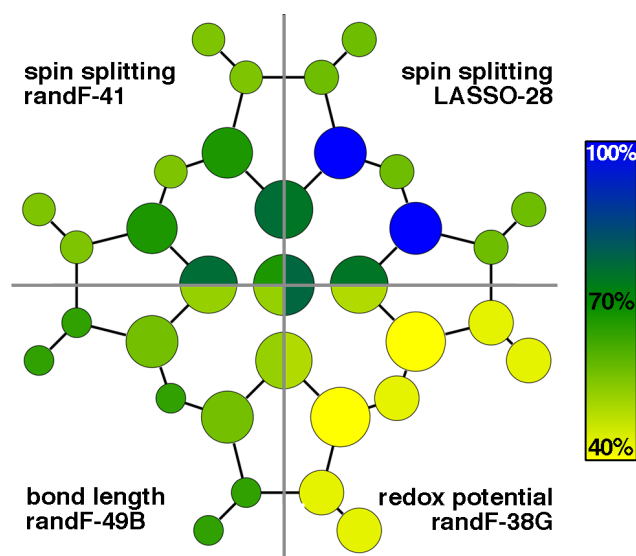


Figure 4-11: Schematic of relative proximity and electronic (blue) or topological (yellow) of feature sets on an iron-porphyrin complex. Feature sets are designated by their training data: spin splitting (randF-41 and randF-26, top), bond length (randF-49B, bottom left), and redox potential (randF-38G, bottom right). Atom sizes are scaled relative to the number of descriptor dimensions involving that atom (divided into first shell, second shell and other), scaled, with iron kept the same size in all sets. The color bar and absolute percentages of electronic and topological descriptors, as defined in the main text, is shown in inset right.

We confirm our earlier observations<sup>308</sup> of locality, especially in spin-splitting with

randF-26/41: randF-49B and randF-38G both have more non-local (to the metal) and topological descriptors than randF-26/41.

For direct ligand connection atoms, 80% of the descriptors are electronic for randF-26/41, but only 52% are electronic for randF-49B and 50% for randF-38G, which reflects the inclusion of additional first-shell *T*- and *I*-based RACs (Figure 4-11). Moving to the second shell shows increased topological fraction across all feature sets while preserving the first shell trends, with second shell descriptors around 65% electronic for the spin-splitting-selected randF-26/41 but only 40% electronic for randF-38G. LASSO-28 has an even stronger electronic, proximal bias than randF-26, possibly explaining its poorer performance for redox potential prediction (Appendix B, Figure B-23). These observations suggest that overall ligand shape and size are more useful for prediction of redox potentials and bond lengths compared to spin splitting within the random forest model. These locality measures also highlight the features to be varied when collecting additional data in future work to enlarge the size of our redox data set and reach smaller ML prediction errors (e.g., 0.1 eV MUE) that would be beneficial for screening and discovery.

Inorganic chemical similarity is less well established than equivalent concepts in organic chemistry, so proximity of inorganic complexes in descriptor space can provide valuable chemical insight. Principal component analysis in the randF-38G feature set of the redox data set reveals simple, intuitive relationships between homoleptic complexes as well as the heteroleptic complexes that arise from interchanging ligands to convert between homoleptic data points (Figure 4-12). The homoleptic Fe(II/III) strong-field methylisocyanide complex with a carbon connecting atom is distant in the redox PCA space from either weaker field furan (oxygen connecting atom) or pyridine (nitrogen connecting atom) ligands. The higher relative distance between carbon and oxygen connecting-atom ligands is also consistent with our expectations

about ligand field effects (Figure 4-12). The heteroleptic complexes that are formed by substituting select axial or equatorial ligands in any of these homoleptic complexes fall in the PCA space on the straight lines that connect between these complexes. Analysis of complex distances in the descriptor space represented by the randF-38G feature set reveals intuitive relationships between inorganic complexes. In addition to machine learning property prediction, such feature sets then provide a path to mapping inorganic chemical space and identifying regions to study in order to identify new complexes similar to known complexes with desired properties.

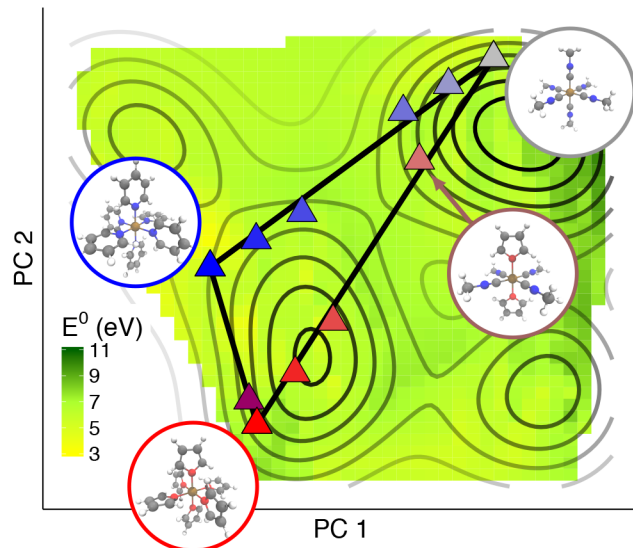


Figure 4-12: Simplified principal component analysis for the redox data set using the randF-38G feature set. The color map indicates redox potential (in eV, as indicated in inset color bar), and the contours represent data density (increasing from gray to black). Representative Fe(II/III) redox couples are indicated with triangles, colored according to the atomic identity of metal-coordinating atoms: nitrogen (blue), oxygen (red), and carbon (gray). Three reference homoleptic Fe complexes, pyridine, methylisocyanide, and furan, are indicated with inset ball and stick structures; these structures form the vertices of a triangle in the PCA space (solid black lines). Computed heteroleptic combinations colored according to the mixing of ligand identities in the PCA space fall along the legs of the triangle, and the location of  $\text{Fe}(\text{furan})_2(\text{misc})_4$  is indicated with an arrow and inset.

## 4.5 Conclusions

We have introduced a new series of revised autocorrelation (RACs) descriptors for machine learning of quantum chemical properties that extend prior ACs to incorporate modified starting points, scope over the molecule of interest, and incorporate differences of atomic properties. We first demonstrated superior performance of standard ACs on a large organic molecule test set, both showing the best yet performance for atomization energies based only on topological information, particularly when maximum topological distances were truncated at a modest maximum 3-bond distance.

We confirmed transferability of RACs from organic to inorganic chemistry with KRR model test set MUEs for the full RAC-155 set of 1 kcal/mol, in comparison to 15-20 $\times$  larger errors from Coulomb-matrix-derived descriptors and 2-3 $\times$  larger with our prior MCDL-25 set. We attribute this improvement to overestimation of size-dependence in CM descriptors and underestimation of distal effects in MCDL-25. LASSO or random forest feature selection yielded smaller subsets (LASSO-28 and randF-41, respectively) with improved or comparable sub- to 1-kcal/mol test MUEs. Restriction to a common set of descriptors identified by the three best feature selection tools yielded half as large spin-splitting errors (test MUE: 1.9 kcal/mol) compared to MCDL-25 with a still smaller 12 variable feature set. Both random forest as a feature selection tool and the spin-splitting-selected randF-26 showed the best combined transferability to bond length (0.005 Å test MUE) and redox potential (0.23 eV test MUE).

Random forest applied directly on bond length selected more topological features than for spin-splitting with equivalent locality bias. Selection based on redox potential data revealed redox potential to be both more non-local and more topological in



nature than spin-splitting or bond lengths. However, invariant data-clustering within the trained KRR model means that no improvement in KRR test errors was observed with redox-selected features for redox potentials and only modest improvement using bond-length selected features for bond length prediction. Overall, this work provides both a prescription for machine learning models capable of making accurate predictions of inorganic complex quantum-mechanical properties and provides insight into locality in transition metal chemistry structure-property relationships.



# Chapter 5

## Design of spin crossover materials with ANNs and DFT

Note: This chapter was originally published as “Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* 2018, 9, 1064–1071” and has been formatted for consistency. Supporting information provided with the manuscript, available online at <http://www.doi.org/10.1021/acs.jpcllett.8b00170>, has been placed in Appendix C.

### Chapter summary

In this chapter, genetic algorithm (GA) optimization is used to discover unconventional spin-crossover complexes in combination with efficient scoring from an artificial neural network (ANN) that predicts spin-state splitting of inorganic complexes. A compound space of over 5,600 candidate materials derived from eight metal/oxidation state combinations and a 32-ligand pool is explored. A new strategy for error-aware ML-driven discovery is introduced by limiting how far the GA travels away from the nearest ANN training points while maximizing property (i.e., spin-splitting) fitness, leading to discovery of 80% of the leads from full chemical space enumeration. Over a 51-complex subset, average unsigned errors (4.5 kcal/mol) are close to the ANN’s baseline 3 kcal/mol error, which indicates that generalization errors were efficiently controlled. By obtaining leads from the trained ANN within

seconds rather than days from a DFT-driven GA, this strategy demonstrates the power of ML for accelerating inorganic materials discovery.

## 5.1 Introduction

Although increases in computing power and efficient algorithms<sup>103–105,501–503</sup> have cemented first-principles screening<sup>6,99,108,472,474–476,482,489,504</sup> (e.g., with density functional theory or DFT) as a critical component of materials and chemical discovery<sup>14,99,109,472–478,505</sup>, further acceleration is needed to overcome the combinatorial challenge of the vast regions of unexplored chemical space<sup>22,90</sup>. With the increased availability of large training sets<sup>31</sup>, machine learning (ML) has emerged<sup>212,223,407,486,487</sup> as a tool to replace first-principles characterization, demonstrating improvement over linear structure-property relationships<sup>231,481</sup> and, where large data sets are available, predicting energies with an accuracy that approaches or exceeds the baseline accuracy of approximate DFT energetics<sup>137,488</sup>. ML models have excelled in design for narrow composition spaces (e.g., alloys<sup>231,481</sup> or phase stabilities<sup>506</sup>). Descriptors used in ML model training can have strong size and domain dependence<sup>24,25,79,250</sup> that restrict discovery to a specific size range and chemical composition. Following the successes of force field development<sup>429</sup>, group additivity<sup>431</sup>, and cheminformatics<sup>428,430</sup>, major ML-driven advances have been made in organic molecule design and discovery<sup>228,396</sup> where structure-property relationships are well-defined. Inorganic chemistry represents a challenging case where few<sup>432</sup> force fields are available, informatics approaches are less well-developed<sup>482,489,507–510</sup>, and properties of interest such as spin-state ordering or redox potential require robust first-principles characterization. Nevertheless, the enlarged chemical space afforded by inorganic chemistry motivates ML model development as a tool to accelerate discovery. We recently trained<sup>308</sup>

an artificial neural network (ANN) on 2,690 geometry optimized transition metal complexes to predict transition metal complex adiabatic high-spin to low-spin state splitting ( $\Delta E_{\text{H-L}}$ ) with root mean square error (RMSE) of 3 kcal/mol along with its Hartree-Fock exchange sensitivity and metal-ligand bond length. We selected 25 mixed continuous (i.e., oxidation state) and discrete (i.e., metal identity) local descriptors (MCDL-25) that focused on metal-proximal effects and demonstrated superior transferability over whole complex representations to the prediction on diverse molecules from experimental databases (Figure 5-1). This feature set was selected from seven candidate feature sets, as evaluated by retained features and errors with LASSO<sup>209</sup>, and the inclusion of discrete features was made possible by their compatibility with an ANN. As suggested by the success of ligand field theory<sup>197,434,435</sup>, our representation<sup>308,511</sup> is ideal for predicting inherently local, electronic properties such as spin state splitting.

Now we turn to the outstanding challenge of using ML models to enable chemical discovery in inorganic chemistry. An open question for the use of ML models<sup>25,79,308,481,487</sup> in discovery<sup>34,389,423,512</sup> is the manner in which we should optimally balance exploration of new compounds with model confidence. Although ML model predictions are of virtually no computational cost versus direct simulation, if the ML model lacks extrapolative power to previously unstudied complexes, then its utility for chemical space exploration will be limited. A second concern is the manner in which optimization is carried out in continuous, data driven representations versus discrete representations needed for characterization, e.g., by simulation.

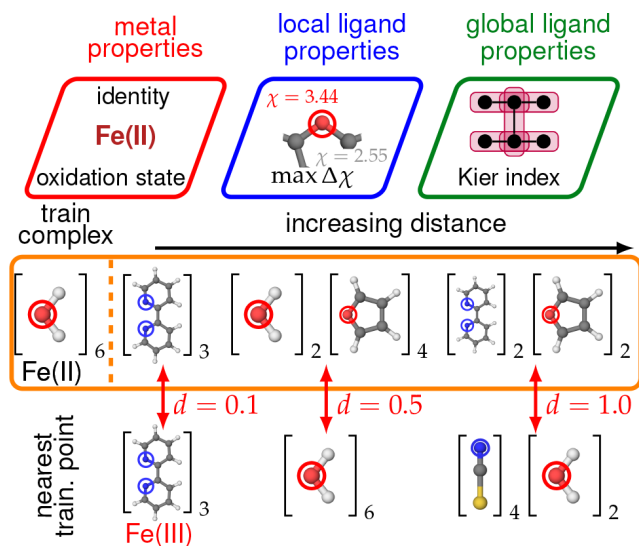


Figure 5-1: (top) Representative descriptors in MCDL-25: metal properties, metal-adjacent (i.e., local ligand properties), and global ligand properties. (bottom) Representative complexes including Fe(II)(H<sub>2</sub>O)<sub>6</sub> in training data and increasingly distant complexes from the training data (left to right): Fe(II)(bpy)<sub>3</sub>, Fe(II)(H<sub>2</sub>O)<sub>2</sub>(furan)<sub>4</sub>, and Fe(II)(bpy)<sub>2</sub>(furan)<sub>2</sub>. The closest training point and its distance is indicated below each complex.

In this work, we develop strategies for balancing exploration in inorganic chemical space with model confidence in a manner that makes leads obtained from a ML model amenable to straightforward validation by first-principles simulation, a necessary step towards automated, adaptive learning. We circumvent the secondary question of continuous optimization<sup>21,398,402,403,513</sup> by using widely used<sup>376,514–516</sup> genetic algorithms (GAs) on discrete ligand pools in combination with our ANN to discover unconventional spin-crossover complexes.

Spin-crossover complexes (SCOs) are defined by near-zero free-energy differences of high (H) and low (L) spin states ( $\Delta G_{H-L}$ ), with changes in spin in response to light/heat due to entropic differences<sup>52</sup>. This behavior makes SCOs compelling for potential applications, e.g., spintronics and sensing of light<sup>517,518</sup> or small molecules<sup>519–521</sup>. Conventional<sup>60,522–524</sup> Fe(II)/nitrogen SCOs are well-studied<sup>434,435,477</sup>, and design

rules for these complexes have been recently suggested<sup>525,526</sup>. In this work, we use GAs on a larger (i.e., several thousand) compound space to reveal both expected and unconventional SCOs as judged through adiabatic electronic energy differences (i.e.,  $|\Delta E_{\text{H-L}}| < 5$  kcal/mol, see Section 5.5).

## 5.2 Design methodology

We now will explore GA-driven strategies for discovering octahedral spin-crossover complexes from a chemical space comprised of metals in the original ANN (i.e., M(II/III) where M=Cr, Mn, Fe, or Co) with 32 unique ligands with varied denticity (i.e., 16 monodentate, 14 bidentate, and 2 tetradentate), direct connecting atom identity (CA, i.e., C, N, and O), and size (i.e., from 2 atoms in CO to 52 atoms in cyanoaceticporphyrin) (Appendix C, Table C.1 and Figure C-1). Taking into account ligand compatibility and the symmetry required by the ANN51 (i.e., one gene each for axial and equatorial ligand identity), these combinations produce a compound space of 5,664 (i.e., 708 ligand combinations  $\times$  8 metals) transition metal complexes ranging from 13 to 103 atoms in size (Appendix C, Table C.2). Of the 32 ligands, 14 were in the original set of ligands used to train the ANN, but only 113 of 5,664 compounds (2%) in the design space have been previously studied<sup>308</sup>.

For spin crossover complex discovery, our target is to minimize the spin-state splitting (i.e.,  $\Delta E_{\text{H-L}}$ ) obtained (e.g., with DFT or an ANN), using H-L definitions as in previous work<sup>308</sup> (see Section 5.5 and Appendix C, Table C.3). At each generation in the GA, compound spin-splitting fitness ( $F_s$ ) is evaluated as:

$$F_s = \exp \left[ - \left( \frac{\Delta E_{\text{H-L}}}{\Delta w_{\text{H-L}}} \right)^2 \right] \quad (5.1)$$

where  $\Delta w_{\text{H-L}}$  controls the decrease in fitness with increasing magnitude of  $\Delta E_{\text{H-L}}$ , chosen to be 15 kcal/mol to preserve  $F_s \sim 1.0$  for  $|\Delta E_{\text{H-L}}| < 5$  kcal/mol (Appendix C, Figure C-2). For the GA, we follow a similar strategy to Ref.<sup>14</sup>: starting from a pool of 20 randomly selected complexes, 21 generations are carried out with fitness evaluation, which includes 5 crossovers and random mutation probability ( $p_{\text{mut}}$ ) of 0.15 (i.e., of the metal or ligand genes, details in Appendix C, Text C.1). Differences from standard choices<sup>14,514</sup> are the reduced number of generations and a higher mutation probability to increase diversity<sup>515</sup>, both motivated by the modest compound space. We introduce a diversity control mode to further increase diversity (i.e., percentage of complexes with distinct gene combinations in the total pool) by raising  $p_{\text{mut}}$  by 0.5 when the diversity of a generation falls below 25% of the pool and reduce  $p_{\text{mut}}$  to the 0.15 default once diversity reaches at least 25%.

Evaluation of compound fitness during GA optimization with a trained ANN motivates consideration of uncertainty in model predictions. Beyond sometimes overconfident credible intervals<sup>459</sup> obtained from dropout, we have identified<sup>308</sup> large (i.e.,  $> 1.0$ ) Euclidean norm of the distance ( $d$ ) in normalized MCDL-25 descriptor space to training data to be a useful indicator of low ANN accuracy (Appendix C, Text C.1). MCDL-25 emphasizes the direct metal-ligand environment: Preserving an oxygen connecting atom but replacing water ligands with larger furan ligands (i.e., changing the  $\Delta\chi$  and topology from a truncated Kier index<sup>453</sup>) produces moderate distances (i.e.,  $d = 0.5$ , see Figure 5-1), whereas a changed CA in otherwise comparably structured ligands (i.e., imidazole vs furan) produces large distances (i.e.,  $d = 1.5$ ). Differences in oxidation state (e.g., Fe(II)(bpy)<sub>3</sub> vs Fe(III)(bpy)<sub>3</sub>:  $d = 0.1$ ) are closer in descriptor space than different metals (e.g., Fe(II)(furan)<sub>6</sub> vs Mn(II)(furan)<sub>6</sub>:  $d > 1.0$ ) (see Figure 5-1). Large distances in descriptor space can arise from substantial differences in all ligands, even when metal, oxidation state, and



direct CAs (i.e., all proximal features) are unchanged. Fe(II)(bpy)<sub>2</sub>(furan)<sub>2</sub> is distant (i.e.,  $d = 1.0$ ) from the closest ANN training point<sup>308</sup>, Fe(II)(NCS)<sub>4</sub>(H<sub>2</sub>O)<sub>2</sub>, due to differing denticity,  $\Delta\chi$ , and truncated Kier index<sup>453</sup>. Remote changes more than three bonds away from the metal-ligand bond<sup>511,527</sup> are neglected in the nearsighted descriptor set, so distinct compounds can be identical in MCDL-25 (see Appendix C, Table C.1).

Thus, using observations about the relationship between chemical differences and descriptor distances, we define our target discovery region for ANN scoring as  $0.3 \leq d \leq 1.0$ , to avoid both "discovery" of complexes too similar to training data and high-promise but very low confidence complexes. We introduce a modified fitness function ( $F_{s,d}$ ):

$$F_{s,d} = \exp \left[ - \left( \frac{\Delta E_{\text{H-L}}}{\Delta w_{\text{H-L}}} \right)^2 \right] \exp \left[ - \left( \frac{d}{d_{\text{opt}}} \right)^2 \right] \quad (5.2)$$

where in addition to a splitting fitness term, a penalty scaled by  $d_{\text{opt}}$  is set to discourage sampling compounds with a very large distance to the training data. To encourage compound discovery (i.e.,  $0.3 \leq d \leq 1.0$ ), we introduce a distance control mode that adapts the fitness function from eq. 5.1 to eq. 5.2 only if the average  $d$  of all complexes is large ( $d_{\text{av}} > 0.6$ , selected by trial and error) after a generation has been selected for fitness and reverts to eq. 5.1 if  $d_{\text{av}}$  falls below 0.6 in a subsequent generation. We have selected  $d_{\text{opt}} = 1.0$  in eq. 5.2 by trial and error to avoid overpenalizing discovery.

We compare four modes of spin crossover complex GA optimization using an ANN for scoring: i) distance control in the fitness function, ii) mutation-rate enhancement to encourage diversity, iii) both distance and diversity controls, and iv) a standard

spin-splitting fitness GA. A 21-generation GA run requires a little over 5 minutes to complete (limited by complex assembly with partial force field optimization for optional follow-up DFT study), whereas fitness evaluation with DFT single points at guessed geometries<sup>181</sup> would require 4 days (Appendix C, Text C.2). All molecules are built, scored, and evolved using an automated design extension to our molSimplify toolkit<sup>181</sup>, which is freely available online (Appendix C, Text C.3). As expected, the standard GA rapidly (i.e., within 5 generations) approaches a mean pool fitness of 1.0 through a dramatic drop in diversity corresponding to roughly one lead compound at the end of the GA run (Figure 5-2). Introducing diversity control improves the number of retained compounds, but diversity control or standard GA runs converge to high-distance/low-confidence leads ( $d_{av} \sim 1.0$ , Figure 5-2). Adjusting fitness evaluation from eq. 5.1 to eq. 5.2 with distance control reduces  $d_{av}$  to around 0.5 (Figure 5-2). Introducing a distance control comes at the cost of slightly reducing the mean spin-splitting only fitness of the retained ligands to around 0.8, a modest increase in  $|\Delta E_{H-L}|$  due to the exponential fitness function, and, interestingly, increases the pool diversity (Figure 5-2). Finally, combining both controls preserves the good features of both strategies: diversity of leads at the end of a GA run is highest overall, and mean distance to training set is unchanged from distance control (Figure 5-2). Incorporating diversity or both controls increases the number of distinct complexes sampled in the GA runs by 50% (150 vs. 100) over other modes and localizes retained hits to a narrow area of target distance and spin-splitting (Appendix C, Figures C-3–C-4).

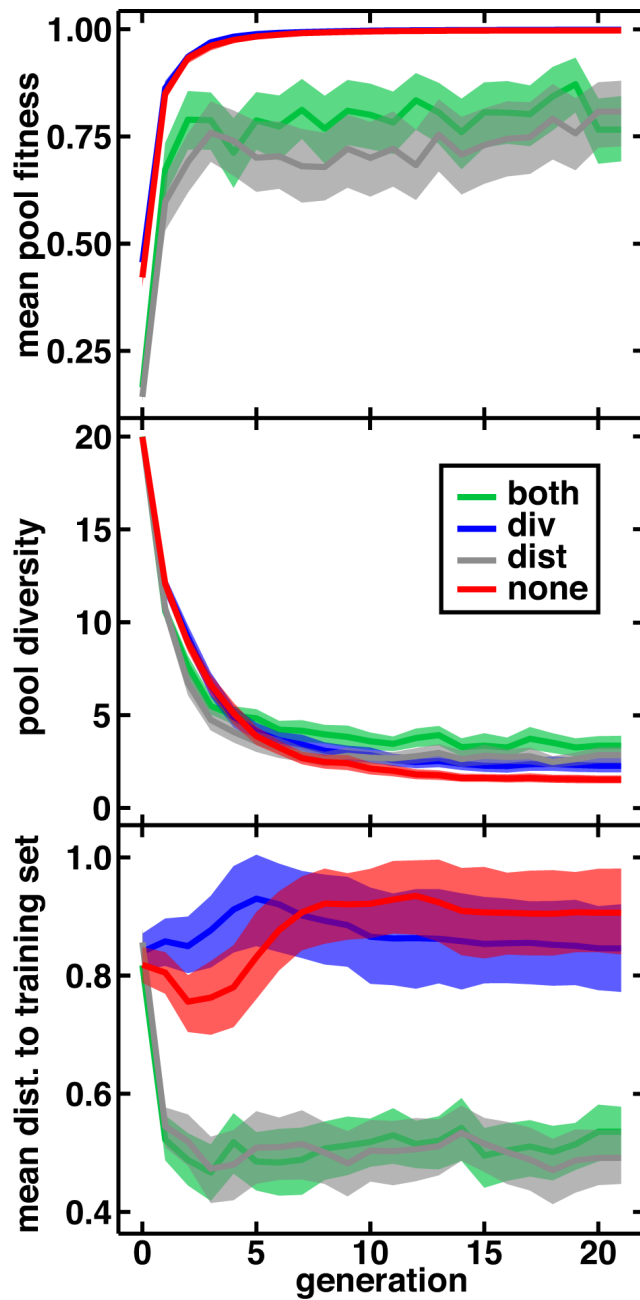


Figure 5-2: ANN GA runs with diversity and distance control (both, green), diversity control only (div, blue), distance only (dist, gray), and no controls on (none, red), as described in the main text: mean splitting-only pool fitness (top), diversity of the pool (middle), and mean distance to training data (bottom) with one standard deviation over 50 runs shown in translucent shading.

### 5.3 Results and discussion

Over 50 repeats, roughly half ( $\sim 2,800$ ) of the compound space is sampled by the standard GA, and the slight reduction ( $\sim 2,650$ ) in compounds sampled with distance control is compensated by combination with diversity control ( $\sim 3,300$ ) (Appendix C, Table C.4 and Figure C-5). We evaluated the full feasible design space with the ANN in a little over 7 hours on a standard desktop machine to identify the fraction of leads (i.e.,  $|\Delta E_{H-L}| < 5$  kcal/mol and  $0.3 \leq d \leq 1.0$ ) missed during these GA optimizations (Appendix C, Text C.2).

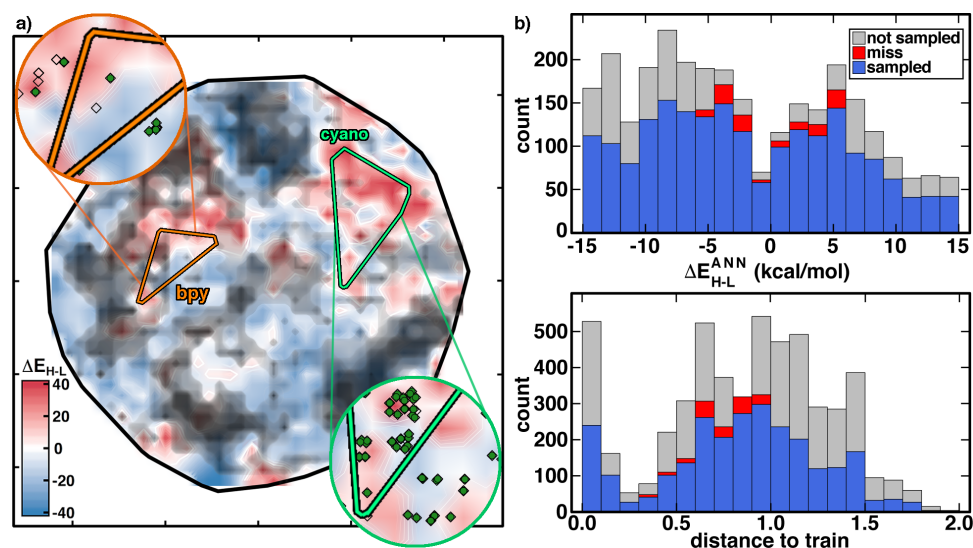


Figure 5-3: (a) t-SNE plot of the full compound space colored by  $\Delta E_{H-L}$  (in kcal/mol as indicated in inset color bar) with increasingly high distance-to-train regions indicated in darker shades of gray. The convex hulls of two families of Fe(II) with substituted-bpy ligands and substituted-cyano ligands are indicated by orange and bright green triangles, respectively. Insets show zooms to each of these regions with discrete hits in empty diamonds with sampled hits as filled dark green circles. (b) 1D histograms of the ANN-predicted  $\Delta E_{H-L}$  (top) and data distance to training data (bottom) using both controls in a stacked bar graph consisting of all sampled points (blue), non-sampled, non-hits (gray), and non-sampled hits (red)

Roughly 8% (474 complexes) of the constructed design space corresponds to our definition of lead compounds (Figure 5-3, pane a and Appendix C, Table C.4). Our recommended control strategy (both) recovers nearly 80% of the lead compounds, a substantial improvement over a standard GA or distance control. Most missed compounds are at larger ( $d > 0.5$ ) distances (Figure 5-3, pane b and Appendix C, Table C.4 and Figure C-6).

Dimensionality reduction<sup>528</sup> of the full compound space in continuous descriptors<sup>511</sup> similar to MCDL-25 reveals why it is challenging to ensure that a GA samples all compound leads (Figure 5-3, pane a and Appendix C, Text C.4). Although families of related complexes are reasonably well-clustered in this representation (i.e., most Fe(II) substituted-bpy and nearly all Fe(II) substituted-cyano complexes are sampled in small regions), variation of properties in this space is quite rough. Narrow target regions that correspond to SCOs are surrounded by non-leads, and several of these promising compound regions are in areas where the ANN confidence is low (black shading in 5-3, pane a).

Both distance- and diversity-controlled GA exploration provides a promising approach to reveal a large fraction of theoretical leads in compound space with an ANN. An additional concern is whether our distance control ensures reasonable reliability of the ANN-based fitness scoring. We quantify the ANN prediction accuracy over a randomly selected 51-complex subset (i.e., roughly 15%) of the 372 identified leads by fully geometry optimizing the high-spin and low-spin states (see Section 5.5 and Appendix C, Table C.5). Overall performance on these newly generated complexes is good, with mean unsigned error (MUE) of 4.5 kcal/mol, 40% (80%) of all compounds are predicted at or below  $1\times$  ( $1\times$ ) baseline error of the ANN<sup>308</sup> on a set-aside test set (Figure 5-4). Around  $\frac{2}{3}$  of ANN spin-crossover leads are validated (i.e.,  $|\Delta E_{H-L}| \leq 5$  kcal/mol) by DFT geometry optimization (Figure 5-4). Inclusion of solvent and

thermodynamic corrections, which were omitted from ANN training or our fitness function, reduces this fraction only slightly to around  $\frac{1}{2}$  of candidates (22 of 49, see Appendix C, Table C.5). Improvement upon this performance would likely require ANN training directly on  $\Delta G_{H-L}$  rather than shifting the fitness function because inclusion of solvent and thermodynamic corrections does not produce a systematic shift of  $\Delta G_{H-L}$  with respect to  $\Delta E_{H-L}$ . Unconventional, promising complexes (i.e., non-Fe(II)/N, with  $\Delta G_{H-L} \sim 1.5$  kcal/mol) identified by the ANN and confirmed with DFT  $\Delta G_{H-L}$  include Mn(II)(CNCH<sub>3</sub>)<sub>2</sub>(CO)<sub>4</sub> or Fe(II)(CO)<sub>2</sub>(NCS)<sub>4</sub>. Conventional<sup>60,522–524</sup> complexes (e.g., Fe(II)(phen)(en)<sub>2</sub> and Fe(III)(NCS)<sub>2</sub>(mebpy)<sub>2</sub>) are also captured (Appendix C, Table C.5).

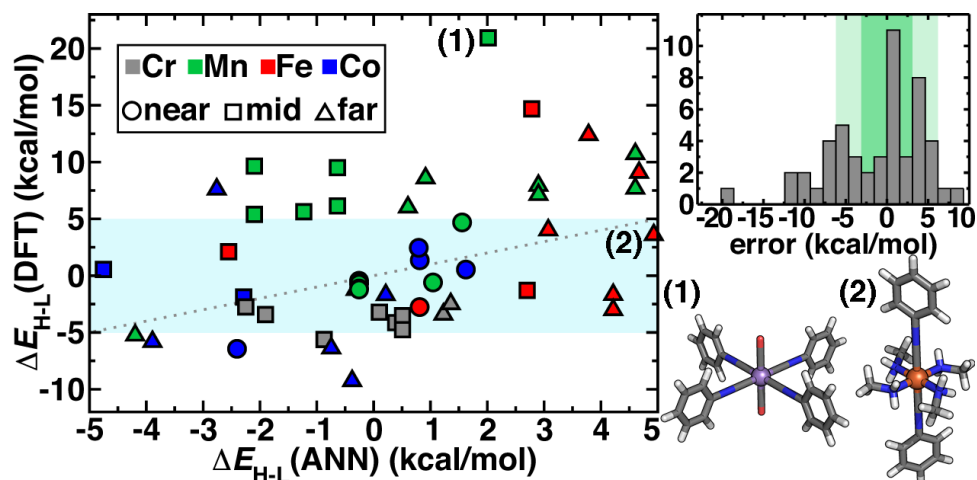


Figure 5-4: (left) B3LYP (DFT) geometry optimization  $\Delta E_{H-L}$  vs. ANN prediction distinguished by metal identity (Cr gray, Mn green, Fe red, and Co blue) and distance from closest training point (near, circles:  $< 0.5$ , mid, squares:  $0.5$  to  $0.75$ , and far, triangles:  $> 0.75$ ). A parity line is shown (gray, dotted), and the  $|\Delta E_{H-L}| \leq 5$  is shown in light blue. (right, top) Error histogram of ANN predictions with baseline error (medium green) and  $2\times$  baseline error (light green) regions shown. (right, bottom) Representative complexes corresponding to labels at left.

Categorizing distance to training data into near ( $d < 0.5$ ), mid ( $0.5 < d < 0.75$ ),

and far ( $0.75 < d < 1.0$ ) complexes reveals excellent prediction accuracy on the near subset (MUE = 1.5 kcal/mol) and non-monotonically worsening performance for mid (MUE = 6.24 kcal/mol) and far (MUE = 4.7 kcal/mol) subsets (Figure 5-4 and Appendix C, Table C.5 and Figure C-7). Good performance is obtained for far iron complexes, such as Fe(II)(CNPh)<sub>2</sub>(NH<sub>2</sub>CH<sub>3</sub>)<sub>4</sub> ( $d = 0.79$  and error: 1.3 kcal/mol, see 2 in Figure 5-4). A systematic underprediction of Mn complexes is apparent, with MUE for all Mn of 5.1 kcal/mol vs. 4.1 kcal/mol for remaining metals, despite comparable average distances over the two subsets ( $d_{av} = 0.65$  vs 0.63). The most notable example is Mn(II)(CO)<sub>2</sub>(CNPh)<sub>4</sub> ( $d = 0.51$  and error: -18.9 kcal/mol, see 1 in Figure 5-4). The closest training points<sup>308</sup> are homoleptic Mn(II)(CNCH<sub>3</sub>)<sub>6</sub> ( $\Delta E_{H-L} = 10$  kcal/mol) and Mn(II)(CO)<sub>6</sub> ( $\Delta E_{H-L} = -6.6$  kcal/mol), explaining why the ANN might predict the Mn(II)(CO)<sub>2</sub>(CNPh)<sub>4</sub> complex to have near degenerate spin states (i.e., by averaging these two compounds), even though the ANN can predict non-additive effects. The origin of this unexpected deviation is indicated by large (i.e.,  $> 2.5$  Å) Mn-CO distances in the DFT-optimized high spin complex compared to low-spin complexes (1.9 Å), suggesting electronic structure differences in this sampled heteroleptic compound absent from the homoleptic-focused training data. These observations at mid/far distances motivate adaptive retraining (i.e., to incorporate more heteroleptic combinations of weak and strong field ligands) for improved accuracy in using ANNs for discovery.

From a theoretical compound space of which only a fraction were likely spin crossover complexes, the ANN-GA results are enriched in the number of DFT-level spin crossover complexes by around an order of magnitude (Appendix C, Text C.2). Best estimates of a full geometry-optimization-driven GA run walltime are around 10 – 30 days, even with parallel evaluation of each generation (Appendix C, Text C.2 and Figures C-8–C-9). We considered alternatively using guessed<sup>181,308,482</sup> H and L geometries

to evaluate single point (SP)  $E_{\text{H-L}}$  with DFT, requiring around 4 days for a GA run (Appendix C, Text C.2 and Figures C-8 and C-10). Imbalanced effects in bond length prediction errors<sup>308</sup> on spin-state ordering means that the DFT-SP-GA performs worse than the ANN with MAEs of 11 kcal/mol and only 30% of compounds remaining spin-crossover complexes after geometry optimization (Appendix C, Table C.6).

A final consideration in computational discovery of spin crossover complexes is the strong dependence of spin-state ordering on functional<sup>59,194,196–199,435,437</sup> with few exceptions<sup>529</sup>, especially on admixture of Hartree-Fock (HF) exchange<sup>59,194,196–199,435</sup> due to differences in delocalization error between spin states<sup>530</sup>. Our ANN was trained on a range of HF exchange, making it possible to identify SCOs in a functional-dependent manner. Re-running the ANN GA with both controls at reduced 15% exchange (i.e., B3LYP\*<sup>198,199</sup> vs. 20% in B3LYP thus far) yields new candidates with weaker field ligands (e.g.,  $\text{Mn(III)(NH}_2\text{CH}_3)_4(\text{CNPyr})_2$  and  $\text{Fe(II)(ox)}_2(\text{CN})_2$   $E_{\text{H-L}} \sim 0\text{--}2$  kcal/mol), in line with our prior observations<sup>197,308</sup> (Appendix C, Figure C-11). Of leads predicted by the ANN, exchange sensitivity is predicted by the ANN to be lowest for Mn(III)/en ligand complexes or Co(II) complexes, and this kind of functional invariance could be a useful metric in future multiobjective optimization.

## 5.4 Conclusions

In conclusion, we have demonstrated an ML-driven strategy for accelerating SCO discovery with an ANN. By pairing our trained ML model with a strategy for controlling novelty of leads in the GA, we discover complexes sufficiently distinct from



training data but for which the ML model can still be suitably employed to make predictions. Using this approach, we have explored a space of  $> 5,500$  candidate materials generated from eight possible metal/oxidation state combinations and 32 possible ligands. Of over 51 representative spin-crossover complexes distinct from ANN training points, average unsigned errors (4.5 kcal/mol) are close to the ANN’s baseline 3.1 kcal/mol error. Two thirds of the discovered compounds, including unconventional complexes, are still considered spin-crossover candidates after full DFT geometry optimization. The largest errors can be avoided in future work by applying an even more conservative distance-control, using a series of independently trained ANNs, or enriching the data set with more heteroleptic compounds. This strategy demonstrates the power of ML for accelerating materials discovery through pre-screening vast chemical space. In future work, we will identify ways to exploit (instead of avoid) high-promise, low-confidence compounds for adaptive retraining of ML models during discovery. We expect this suite of ML models, discovery algorithms, and simulation automation software to be valuable for the optimization of key properties in inorganic chemistry.

## 5.5 Computational details

Single point energies and geometry optimizations were carried out with TeraChem<sup>105,447</sup> at the B3LYP<sup>152,157,158</sup> level of theory with LANL2DZ effective core potential<sup>172</sup> for all transition metals, bromine, and iodine and the 6-31G\* basis for the remaining atoms, as employed during ANN training<sup>308</sup>. Basis set dependence is observed to be small (Appendix C, Table C.7). Although their inclusion has been motivated<sup>192</sup>, vibrational or solvent contributions, which often have compensating effects<sup>192,482</sup> are neglected during fitness scoring by DFT or with the trained ANN (Appendix C,

Table C.5). On representative molecules, vibrational enthalpy and entropy corrections were obtained through calculation of the gas phase Hessian of each spin state. Solvent corrections were obtained from differences in solvation free energy on the gas phase geometries using COSMO<sup>205,499</sup> ( $\epsilon = 78.39$  and a cavity constructed from  $1.2 \times$  Bondi radii<sup>500</sup>).

# Chapter 6

## Uncertainty and extrapolation of ANNs for chemical discovery

Note: Sections 6.1–6.4 of this chapter were originally published as “Janet, J. P., Duan, C., Yang, T., Nandy, A., Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* 2019, 10, 7913–7922” and has been formatted for consistency. Supporting information provided with the manuscript, available online at <http://www.doi.org/10.1039/C9SC02298H>, has been placed in Appendix D. Section 6.5 was originally published as “Gugler, S., Janet, J. P., Kulik, H. J. Enumeration of de novo inorganic complexes for chemical discovery and machine learning. *Mol. Syst. Des. Eng.* 2019, Advance Article”, and includes an alternative strategy to improve generalization performance on the same test set for comparison.

### Chapter summary

Artificial neural networks have emerged as a complement to high-throughput screening, enabling characterization of new compounds in seconds instead of hours, but the promise large-scale chemical space exploration can only be realized if it is straightforward to identify when molecules and materials are outside the model’s domain of applicability. Established uncertainty metrics for neural network models are either costly to obtain (e.g., ensemble models) or rely on feature engineering (e.g., feature space distances), and each has limitations in estimating prediction errors for chemical

space exploration. This chapter develops new strategies to quantify the generalization ability of neural networks to chemistry that is dissimilar to their training data, based on the ANN-learned latent space. This metric serves as a low-cost, quantitative uncertainty description that works for both inorganic and organic chemistry. Calibrated probabilistic models based on these distances outperform widely used uncertainty metrics and are readily applied to models of increasing complexity at no additional cost. Tightening latent distance cutoffs systematically drives down predicted model errors below training errors, thus enabling predictive error control in chemical discovery or identification of useful data points for active learning.

## 6.1 Introduction

Machine learning (ML) models for property prediction have emerged<sup>231,279,423,427,511,531-533</sup> as powerful complements to high-throughput computation<sup>75,77,181,533-535</sup> and experiment<sup>536-538</sup>, enabling the prediction of properties in seconds rather than the hours to days that direct observations would require. Using large data sets, trained interpolative potentials<sup>213,252,254,539,540</sup> and property prediction models<sup>231,279,423,427,511,531-533</sup> have achieved chemical accuracy with respect to the underlying data<sup>212</sup>. Predictive models hold great promise in the discovery of new catalysts<sup>231,279,541,542</sup> and materials<sup>35,237,333,533,543-546</sup> by enabling researchers to overcome combinatorial challenges in chemical space exploration. While application of ML to chemical space exploration is increasingly becoming a reality, a key outstanding challenge remains in knowing in which regions of chemical space a trained ML model may be confidently applied<sup>547</sup>. While trained ML models are fast to deploy to large compound spaces, many models (e.g., artificial neural networks or ANNs) are typically trained only after acquisition of thousands<sup>31</sup> to millions<sup>32,252</sup> of data points. Quantitative uncertainty metrics are most critical in applications of active learning<sup>375,548</sup> where the model is improved by acquisition of selected data. Although some models (e.g., Gaussian process re-

gression) inherently provide estimates of model uncertainty<sup>288,549</sup>, quantitative uncertainty measures for models suited to handle large data sets (e.g., ANNs) remains an active area of research<sup>550-552</sup>.

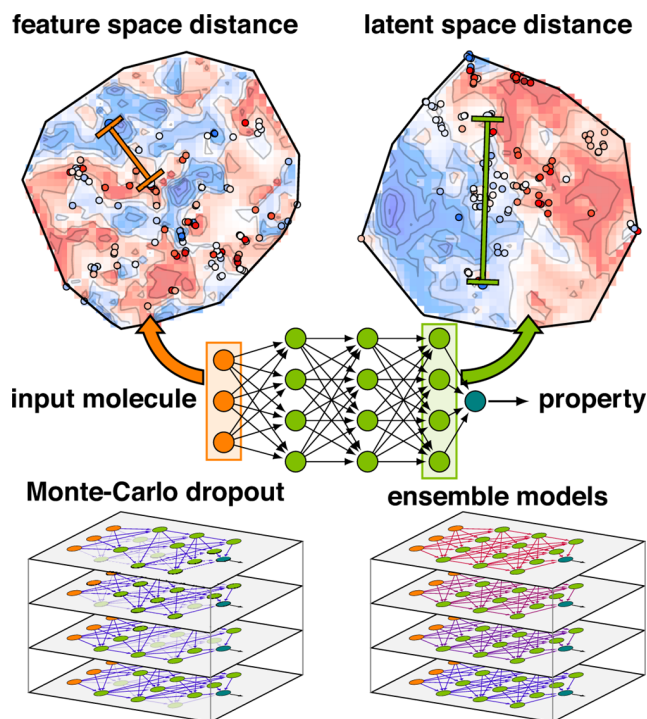


Figure 6-1: Schematic of an ANN annotated with the four uncertainty metrics considered in this work. Two points are compared in terms of their feature space distance (i.e., the difference between two points in the molecular representation) on a t-distributed stochastic neighbor embedding map<sup>528</sup> (t-SNE) of data in the input layer (top, left, annotations in orange) and the latent space distance (i.e., the difference between two points in the final layer latent space) on a t-SNE of the data in the last layer (top, right, annotations in green). The standard ANN architecture (middle) is compared at bottom for Monte-Carlo dropout (i.e., zeroed out nodes) and ensemble models (i.e., varied model weights) at bottom left and right.

One approach to estimating model uncertainty is to train an ensemble of identical architecture models on distinct partitions of training data to provide both a

mean prediction and associated variance (Figure 6-1). While widely employed in the chemistry community<sup>213,550,551,553,554</sup> ensembles increase the model training effort in proportion to the number of models used (typically an order of magnitude, Appendix D, Text D.1). Although this additional effort may be practical for some models (e.g., networks with only a few layers), the training effort becomes cost-prohibitive<sup>555</sup> during iterative retraining for active learning or for more complex models that are increasingly used in chemical discovery, such as those using many convolutional<sup>27,211</sup> or recurrent<sup>351,556</sup> layers. Thus, ensemble uncertainty estimates have been most frequently applied<sup>213,551</sup> in the context of simpler networks, especially in neural network potentials that are trained in a one-shot manner. A key failing of ensemble metrics is that with sufficient model damping (e.g., by  $L_2$  regularization), variance over ensemble models can approach zero<sup>553</sup> for compounds very distant from training data, leading to over-confidence in model predictions.

Another approach to obtain model-derived variances in dropout-regularized neural networks is Monte Carlo dropout (mc-dropout)<sup>459</sup> (Figure 6-1). In mc-dropout, a single trained model is run repeatedly with varied dropout masks, randomly eliminating nodes from the model (Appendix D, Text D.1). The variance over these predictions provides an effective credible interval with the modest cost of running the model multiple times rather than the added cost of model re-training. In transition metal complex discovery, we found that dropout-generated credible intervals provided a good estimate of errors on a set aside test partition but were over-confident when applied to more diverse transition metal complexes<sup>511,533</sup>. Consistent with the ensembles and mc-dropout estimates, uncertainty in ANNs can be interpreted by taking a Bayesian view of weight uncertainty where a prior is assumed over the distribution of weights of the ANN and then updated upon observing data, giving a distribution over possible models<sup>557</sup>. However, if the distribution of the new test data is distinct

from training data, as is expected in chemical discovery, this viewpoint on model uncertainty may be incomplete.

A final class of widely applied uncertainty metrics employs distances in feature space of the test molecule to available training data to provide an estimate of molecular similarity and thus model applicability. The advantages of feature space distances are that they are easily interpreted, may be rapidly computed, and are readily applied regardless of the regression model<sup>511,533,558</sup> (Figure 6-1). We used<sup>511,533</sup> high feature space distances to successfully reduce model prediction errors on retained points while still discovering new transition metal complexes. Limitations of this approach are that the molecular representation must be carefully engineered such that distance in feature space is representative of distances in property space, the relationship between distance cutoff and high property uncertainty must be manually chosen, and this metric cannot be applied to message-passing models that learn representations<sup>235,327</sup>.

A chief advantage of multi-layer neural network models over simpler ML models is that successive layers act to automatically engineer features, limiting the effect of weakly-informative features that otherwise distort distances in the feature space (Figure 6-1). Thus, for multi-layer ANNs, feature-based proximity can be very distinct from the intrinsic relationship between points in the model. Such ideas have been explored in generative modeling where distances in auto-encoded latent representations have informed chemical diversity<sup>228,307</sup> and in anomaly detection with separate models<sup>559,560</sup> (e.g., autoencoders<sup>561-563</sup> or nearest-neighbor classifiers<sup>564,565</sup>) have enabled detection of 'poisoned' input data<sup>566</sup>. However, the relationship between latent space properties and feature space properties has not been exploited or understood in the context of error estimation for property prediction (i.e., regression) ML models.

In this work, we propose the distance in latent space, i.e., the distance of a test point to the closest training set point or points in the final layer latent space, as a new uncertainty metric (Figure 6-1). The advantages of this approach are that i) it introduces no overhead into model training or evaluation, ii) it can work just as easily with both simple and complex ML models that have been used for chemical property prediction (e.g., hierarchical<sup>26</sup>, recurrent<sup>351,556</sup> or convolutional<sup>28,211,567-569</sup>), and iii) it naturally ignores distances corresponding to features to which the model prediction is insensitive, obviating the need for feature engineering to develop an estimate of test point proximity to prior training data. We show that these attributes yield superior performance over other metrics in chemical discovery.

## 6.2 Results and discussion

To demonstrate the advantages of the latent space distance metric in a quantitative fashion, we compare to three established uncertainty metrics. This assessment is particularly motivated by the nature of chemical discovery applications<sup>533</sup>, where data set sizes are often smaller and have more broadly varying chemistry than typical applications in neural network potentials<sup>213,551</sup> and in quantitative structure–property relationships in cheminformatics<sup>552,558</sup>. To mimic chemical discovery efforts, we train neural networks to predict transition metal complex spin state energetics<sup>511</sup> and test them on diverse transition metal complexes from experimental databases. To confirm the generality of our observations, we also compare uncertainty estimates for neural network models trained on a very small subset (i.e., 5%) of QM9<sup>31</sup>, a widely used<sup>26,212,223,330,486,487,570</sup> data set in organic chemistry ML.

For open-shell transition metal chemistry, we use 1901 equilibrium high (H)/low (L) spin splitting energies (i.e.,  $\Delta E_{\text{H-L}}$ ) for octahedral first-row transition metal



(i.e., M(II) or M(III) where M = Cr, Mn, Fe, or Co) complexes generated in prior work<sup>511,533</sup> using density functional theory (DFT). We use the previously introduced<sup>511</sup> full set of revised autocorrelation (RACs) descriptors (i.e., RAC-155) to train a fully connected ANN with three 200-node hidden layers (see Computational Details and Appendix D, Text D.2, Table D.1, and Figure D-1). RACs have been demonstrated for training predictive models of transition metal complex properties<sup>180,259,511,533</sup> including spin splitting, metal-ligand bond length, redox and ionization potentials, and likelihood of simulation success.

To mimic chemical discovery application of this model, we extracted a set of 116 octahedral, first-row transition metal complexes that have been characterized experimentally (i.e., from the Cambridge Structural Database or CSD<sup>184</sup>) as an out-of-sample test set (Figure 6-2 and Appendix D, Text D.2 and Figures D-2–D-5). We selected these CSD complexes to be intentionally distinct from training data, as is apparent from principal component analysis (PCA) in the RAC-1557 representation (Figure 6-2). Several complexes in the CSD test set fall outside the convex hull of the training data in the first two principal components (ca. 50% of the variance) and are distant from training data, as judged by the Euclidean distance in the full RAC-155 feature space (Figure 6-2 and Appendix D, Figure D-6). High distances are observed for complexes containing elements rarely present (e.g., an S/N macrocycle for a Co(II) complex, CSD ID: FATJIT) or completely absent from our training data (e.g., B in boronated dipyrazole ligands of the Fe(II) complex CSD ID: ECODIM and As in thioarsenite ligands in an Mn(II) complex, CSD ID: CEDTAJ) as well as ligand topologies (e.g., acrylamide axial ligands in an Mn(II) complex, CSD ID: EYUSUO) not present in training data (Figure 6-2).

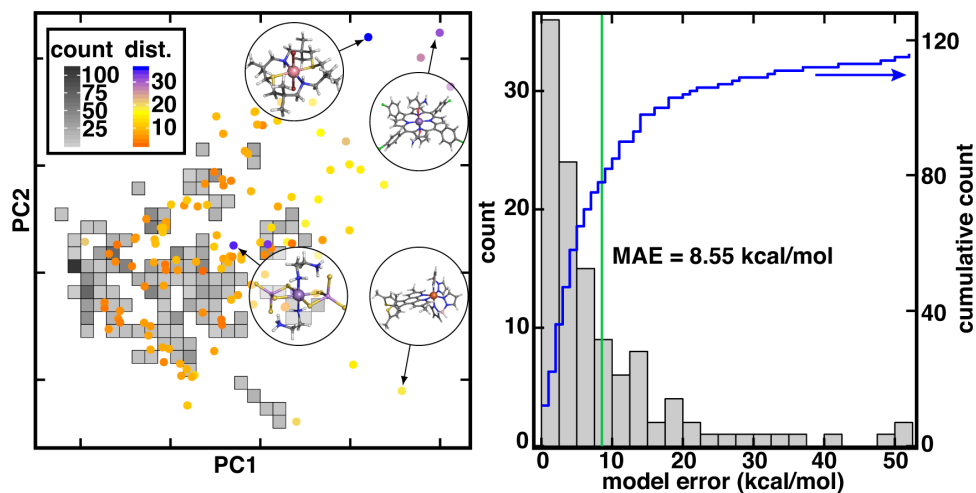


Figure 6-2: (left) Comparison of inorganic training and CSD test data in the dominant two principal components of the RAC-155 representation of the training data set. The density of training data is shown as gray squares shaded as indicated in inset count colorbar. CSD test data points are shown as circles colored by the 10-nearest-neighbor-averaged Euclidean distance in RAC-155 space, as shown in dist. inset color bar. Four representative high-distance structures are shown in circle insets in ball and stick representations: (top left inset, CSD ID: FATJIT) a Co(II) complex with S/N macrocycle and axial Br- ligands, (top right inset, CSD ID: EYUSUO) Mn(II) tetra-chlorophenyl-porphyrin with acrylamide axial ligands, (bottom left inset, CSD ID: CEDTAJ) a Mn(II) complex with thioarsenite ligands, and (bottom right inset, CSD ID: ECODIM) an Fe(II) complex with boronated dipyrazole and thiolated phenanthrene ligands. (right) Distribution of absolute CSD test set model errors for  $\Delta E_{H-L}$  (in kcal/mol, bins: 2.5 kcal/mol) with the MAE annotated as a green vertical bar and the cumulative count shown in blue according to the axis on the right.

Due to the distinct nature of the CSD test set from the original training data, the 8.6 kcal/mol mean absolute error (MAE) of the RAC-155 ANN on the CSD data set is much larger than the 1.5 kcal/mol training set MAE (Figure 6-2 and Appendix D, Table D.2). Use of ensemble- or mc-dropout-averaged predictions unexpectedly<sup>571</sup> worsens or does not improve test MAEs (ensemble: 9.0 kcal/mol; mc-dropout: 8.5 kcal/mol), which we attribute to noise in averaging due to the relatively heterogeneous training data (Appendix D, Figures D-7–D-9). The relative error increase on diverse data is consistent with our prior work where we achieved low errors on test

set partitions of 1 – 3 kcal/mol<sup>7511</sup> that increased<sup>308</sup> to around 10 kcal/mol on sets of diverse molecules (e.g., 35 molecules from a prior curation of the CSD). These observations held across feature sets<sup>308</sup> (e.g., MCDL-25<sup>308</sup> vs. RAC-155<sup>511</sup>) and model architectures (e.g., kernel ridge regression<sup>308,533</sup> vs. ANNs<sup>308,533</sup>) for  $\Delta E_{\text{H-L}}$  property prediction.

Despite the increase in MAE, errors are not uniformly high across the 116 molecules in our new CSD data set (Figure 6-2). A significant number (24 or 21%) of the complexes have errors within the 1.5 kcal/mol training MAE, a substantial fraction are within the 3 kcal/mol test set error described in prior work<sup>308</sup> (41or35%), and a majority (61or53%) have errors 5 kcal/mol or below (Figure 6-2). At the same time, a number of outlier compounds have very large absolute errors with 31 (27%) above 10 kcal/mol and 12(10%) above 20 kcal/mol (Figure 6-2). Large errors are due to both underestimation of  $\Delta E_{\text{H-L}}$  by the model (e.g., Fe(II) complex CSD ID: CEYSAA, model = -23.8 kcal/mol, DFT data = 26.6 kcal/mol) and overestimation (CSD ID: Mn(III) complex CSD ID: EYUSUO, model = 5.7 kcal/mol, DFT data = -46.4 kcal/mol, see Figure 6-2). Given the heterogeneity of observed errors, we apply uncertainty metrics to this data set with the aim to i) systematically drive down error on predicted data points by only making predictions within the model’s domain of applicability and ii) identify data points that should be characterized and incorporated to the model training set in an active learning setting.

For heavily engineered feature sets (i.e., MCDL-25<sup>308</sup>), we showed the Euclidean norm feature space distance to the closest training point could be used to control ANN errors in inorganic complex discovery<sup>533,543</sup>, typically limiting discovery MAEs to only slightly larger (i.e., 4–5 kcal/mol) than the original test MAE. This approach required that we select a cutoff over which distances were deemed too high, a quantity that can be sensitive to the nature of the feature set and the number of

nearest neighbors used in the average (Appendix D, Figures D-10–D-11). Averaging Euclidean norm distances in RAC-155<sup>511</sup> or a feature-selected subset<sup>180,511</sup> over the nearest (i.e., 1–10) neighbors in the training data and only predicting on points sufficiently close to training data systematically eliminates the highest error points (Appendix D, Figure D-11). Consistent with prior work<sup>308,533</sup>, this approach allows us to achieve sub-6 kcal/mol MAE on over half (64 of 116) points in the CSD set, but further improvement of predicted-data MAEs below 5 kcal/mol is not possible (Appendix D, Figure D-11).

In the large, non-engineered feature spaces typically used as input to neural networks, feature space distances may be insufficient for identifying when predictions lack support by data in the model. Thus, we turn to the latent space distance evaluated at the final output layer (Figure 6-1). Using high distances in latent space as the criterion for prediction uncertainty, we drive down MAEs on predicted data nearly monotonically, well below the 5 kcal/mol MAE that could be achieved using feature space distances (Appendix D, Figure D-11). This difference in performance is motivated by the distinct, higher effective dimensionality of the principal components in the latent space over the feature space (Appendix D, Figure D-6). With the distance in latent space as our guide, 76 points can be identified as falling within model domain of applicability (i.e., sub-6 kcal/mol MAE), and 3 kcal/mol MAE can be achieved on over 25% of the data (ca. 30 points), indicating a close relationship between high latent space distance and model error (Appendix D, Figures D-11–D-13). The distance in latent space has the added advantage of being less sensitive to the number of nearest neighbors over which the distance evaluation is carried out than feature space distances (Appendix D, Figure D-11). Our approach is general and not restricted to the distance in the latent space described here. In future work, we could move beyond potential ambiguities<sup>572</sup> in measuring high-dimensional similarity with

Euclidean distances and compare to alternatives, including averaged properties<sup>228</sup> or those that incorporate other geometric features of the latent data distribution.

Having confirmed that distances in latent space provide significant advantages over feature space distances at no additional cost, we also would like to consider the performance with respect to mc-dropout and ensemble-based uncertainty metrics (Appendix D, Figures D-14–D-15). To do so, we overcome the key inconvenience that the distance measure itself does not provide an error estimate in the units of the property being predicted. After model training, we calibrate the error estimate by fitting the predictive variance to a simple conditional Gaussian distribution of the error,  $\varepsilon$ , for a point at latent space distance,  $d$ :

$$\varepsilon(d) \sim \mathcal{N}(0, \sigma_1^2 + d\sigma_2^2) \quad (6.1)$$

where the error is assumed to be normally distributed with a baseline variance term  $\sigma_1^2$  and a growing term  $\sigma_2^2$ . Selection of  $\sigma_1^2$  and  $\sigma_2^2$  using a simple maximum likelihood estimator on a small subset (ca. 20 points) of the CSD test set is relatively robust, leading to property-derived uncertainties (Figure 6-3, Appendix D, Figure D-16 and Tables D.3–D.4). Over the 116-complex CSD test set, this latent-space derived metric spans a large 8 – 24 kcal/mol range and correlates as well to absolute model errors as do ensemble and mc-dropout standard deviation (std. dev.) metrics (Appendix D, Figure D-13).

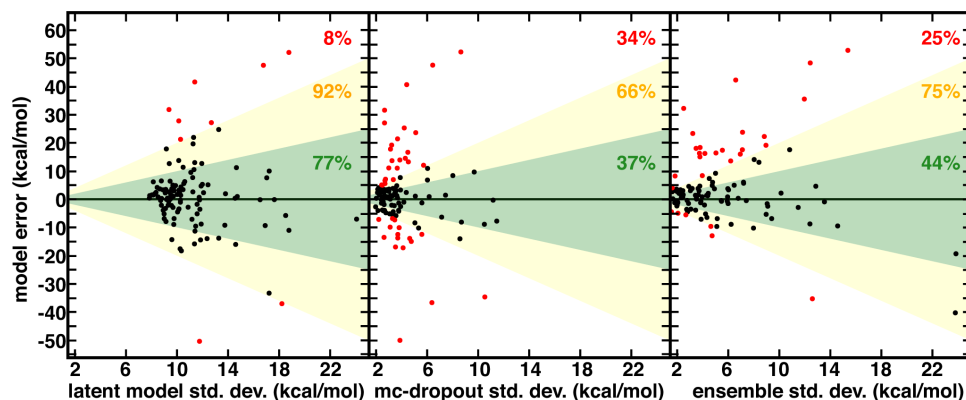


Figure 6-3: Relationship between spin-splitting ANN model errors (in kcal/mol) on a 116 molecule CSD set and three uncertainty metrics all in kcal/mol: latent model energetic, calibrated std. dev. (left), mc-dropout std. dev. (middle), and 10-model ensemble std. dev. (right). The translucent green region corresponds to one std. dev. and translucent yellow to two std. dev.. The points with model errors that lie inside either of these two bounds are shown in black, and the percentage within the green or yellow regions are annotated in each graph in green and yellow, respectively. The points outside two std. dev. are colored red, and the percentage of points in this group is annotated in each graph in red. Three points are omitted from the ensemble plot to allow for a consistent x-axis range.

Although not unique and depending on the training process of the model, the distance in latent space-derived energetic uncertainties provide a superior bound on high error points (Figure 6-3). Observed errors reside within one std. dev. in the majority (77%) of cases, and only a small fraction (8%) exceed two std. dev. ranges (Figure 6-3). In comparison, less than half of errors are within one std. dev. evaluated from the ensemble (44%) or mc-dropout (37%), and a significant fraction of errors exceed two std. dev. (23% and 34%, respectively, Figure 6-3). When the ensemble or mc-dropout uncertainty metrics are used as cutoffs to decide if predictions should be made, model over-confidence leads to inclusion of more high error (i.e.,  $> 12$  kcal/mol) points than when using the latent distance (Appendix D, Figure D-17). The ability to smoothly transition between high cutoffs where more points are characterized with the ML model (e.g., to achieve 8 kcal/mol MAE) vs. conser-

vative where the error is small (e.g., 2 kcal/mol) but only a minority of predictions are made is important for predictive control; here, the latent distance provides the more robust separation between these two regimes, thus enabling greater distinction between the two (Appendix D, Figure D-15).

There are numerous cases where both ensemble and mc-dropout are relatively confident on very high error points in comparison to latent distance. For example, an Fe(II) complex with ethanimine and alkanamine ligands (CSD ID: DOQRAC) is predicted erroneously by the model to be strongly high spin ( $\Delta E_{\text{H-L}_{ANN}} = -34.7$  kcal/mol vs.  $\Delta E_{\text{H-L}_{DFT}} = -1.4$  kcal/mol), but this point has a low std. dev. from the ensemble (4.3 kcal/mol) in comparison to a relatively high 17.2 kcal/mol std. dev. from the latent space distance. Conversely, there are no cases where the latent distance uncertainty is uniquely over-confident, but there are cases where all metrics are overconfident. For example, an Mn(II) complex with four equatorial water ligands and two axial, oxygen-coordinating 4-pyridinone ligands is expected by all metrics to be reasonably well predicted (std. dev. ensemble = 2.5 kcal/mol, mc-dropout = 2.7 kcal/mol, and latent space = 9.4 kcal/mol), but the DFT preference for the high-spin state is underestimated by the ANN ( $\Delta E_{\text{H-L}_{ANN}} = -45.5$  kcal/mol vs.  $\Delta E_{\text{H-L}_{DFT}} = -77.4$  kcal/mol). Although the latent distance error estimate does not bound all high error points predicted by the model, it provides a high fidelity, no cost uncertainty estimate for > 90% of the data.

o assess the generality of our observations on inorganic complexes for other chemical data sets, we briefly consider the approach applied to atomization energies computed with hybrid DFT (i.e., B3LYP<sup>152,157,158</sup>/6-31G<sup>573</sup>) for a set of organic (i.e., C, H, N, O, and F-containing) small molecules. The QM9 data set<sup>31</sup> consists of 134k organic molecules with up to 9 heavy atoms and has been widely used as a benchmark for atomistic machine learning model development<sup>212,223,486,487</sup> with the best models in

the literature reporting MAEs well below 1 kcal/mol<sup>25,26,212,330,486</sup>. As in previous work<sup>511</sup>, we employ standard autocorrelations (ACs)<sup>244</sup> that encode heuristic features<sup>22</sup> on the molecular graph and perform well (ca. 6 kcal/mol MAE) even on small (< 10%) training set partitions for QM9 atomization energies<sup>511</sup>, exceeding prior performance from other connectivity-only featurizations<sup>486</sup>. For this work, we trained a two-hidden layer residual ANN using AC features and passing the input layer forward in a ResNet-like architecture<sup>306</sup> to improve performance over a fully-connected architecture (Computational Details and Appendix D, Figure D-18 and Tables D.5–D.6). We use only 5% (6,614) of the data points for training, reserving the remaining 127k molecules for our test set to mimic chemical discovery in a single random partition, the choice of which does not influence overall performance (Appendix D, Table D.7). Baseline model performance for QM9 atomization energies with the ANN is improved over our prior work for both train (4.6 kcal/mol) and test (6.8 kcal/mol) MAE, with some further improvement of test MAE with an ensemble model (6.1 kcal/mol, see Appendix D, Tables D.7–D.8). A wide distribution of errors is observed with some outlier points such as hexafluoropropane (error = 120 kcal/mol) having very large errors for both the single and ensemble models (Appendix D, Figure D-19). For the residual ANN, the mc-dropout uncertainty has not been derived, and so we compare only the other three uncertainty metrics. We observe ensemble and latent space distance uncertainty metrics to have similar correlations to model errors and both to outperform feature space distance in this regard (Appendix D, Figure D-20). Selecting either the distance in latent space or ensemble uncertainty as a cutoff, we can systematically drive down MAEs on the predicted data fraction, and latent distance again provides superior control when error tolerance is low (Appendix D, Figure D-21). For example, setting a tolerance of 3.5 kcal/mol for the MAE leads to a pool of over 4200 points retained with the latent



space distance metric vs. few points (74) for the ensemble std. dev. (Appendix D, Figure D-21).

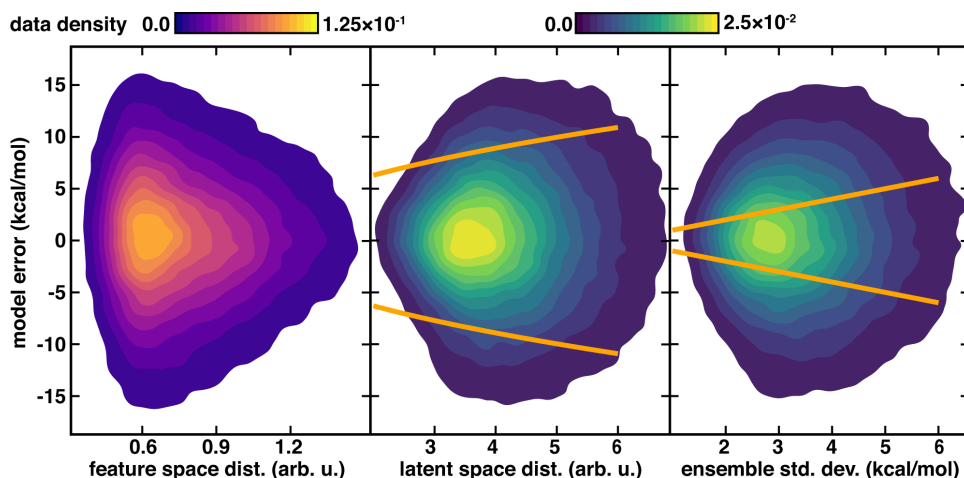


Figure 6-4: Model errors (in kcal/mol) for 127k QM9 atomization energy test points shown as contours as a function of uncertainty metrics. The three uncertainty metrics compared are: feature space distance (in arb. u., left, with top left color bar), latent space distance (in arb. u., middle, with top right color bar), and 10-model ensemble std. dev. (in kcal/mol, with top right color bar). One standard deviation cutoffs are shown as orange lines for the latent space distance from the calibrated error model (center) and directly from the ensemble (right).

We again observe that the AC feature space distance is a poor indicator of increasing model errors, with as many high error points occurring at low distances as at high distances (Figure 6-4). In contrast to feature space distance, ensemble std. dev. and latent distance both grow with increasing error (Figure 6-4). Calibration of the latent space distance to the output property enables direct comparison to ensemble uncertainties (Appendix D, Table D.9). As in the inorganic data set, the ensemble std. dev. values are overconfident, capturing a smaller amount (44%) of the errors within a single std. dev. in comparison to the distance in latent space (77%) metric (Figure 6-4 and Appendix D, Figure D-22). For the ensemble uncertainty, a signif-

icant fraction (28%) of points have errors larger than twice the std. dev., whereas only a small fraction (5%) do so for the distance in latent space (Figure 6-4 and Appendix D, Figure D-22).

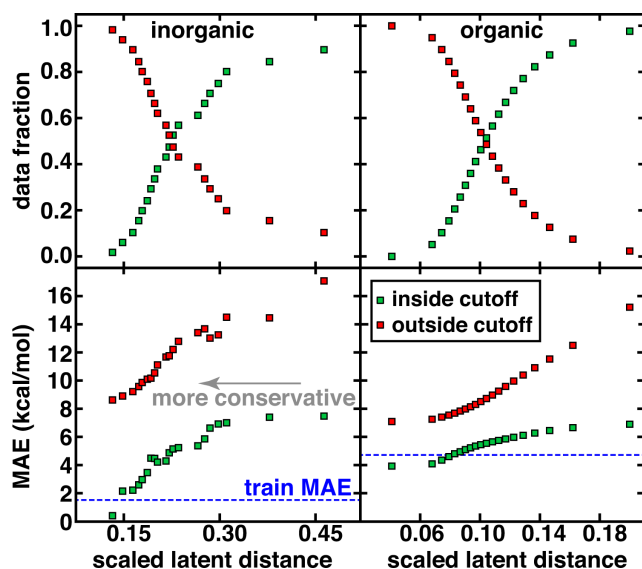


Figure 6-5: MAE for predicted points (inside cutoff, green squares) and those not predicted (outside cutoff, orange squares) compared to the training data MAE (blue horizontal dashed line) along with data fraction in each set (top) for the inorganic CSD test set (left) and organic QM9 set (right). The most distant point in the test set is scaled to have a latent distance of 1.0 for comparison across data sets but the x-axis range is then truncated to focus on the range of latent distance cutoffs that affect most of the data.

For both the CSD test set and the QM9 set, a systematic reduction in baseline error can be observed in a practical use case where the user adjusts the applied uncertainty metric to become more conservative (Figure 6-5). Smooth reductions in MAE on data inside the uncertainty cutoffs can be achieved across a wide range of latent distance cutoffs, with errors nearly monotonically approaching the training data MAE, which may be recognized as a qualitative lower bound on our test set error (Figure 6-5). Combining all error metrics to choose the most conservative result

does not improve upon the single latent space distance metric (Appendix D, Figure D-23). PCA or uniform manifold approximation and projection (UMAP)<sup>574</sup> analysis of the latent space distance indicates that a large number of the latent space dimensions are needed for error estimation (Appendix D, Figure D-24 and Table D.10). For either data set, at the point on which half of all possible predictions are made, predicted data MAE is less than half of that for the excluded points (Figure 6-5). The latent distance as a predictive estimate of uncertainty also shows promise for application in active learning, where a model is trained iteratively by acquiring data in regions of high model uncertainty. To mimic such an application in the context of inorganic chemistry, we returned to the CSD data set and identified the 10 least confident points based on the distance in latent space, retrained the ANN using the same protocol, and re-evaluated model MAE (Appendix D, Table D.11). Incorporating these data points during retraining reduced model errors from 8.6 to 7.1 kcal/mol, whereas simply removing these points only reduced model MAE to 7.7 kcal/mol (Appendix D, Table D.11). This effect is particularly significant considering the relatively small change in the number of data points (i.e., 10 added to 1901 or 0.5%) and an even larger reduction in root mean square error is observed (Appendix D, Table D.11). When compared to an ensemble or mc-dropout cutoff, selection of retraining points based on latent space distance results in the largest reduction in model MAE while only requiring retraining on ANN (Appendix D, Table D.11). Although we have focused on applications in chemical discovery with fully connected neural networks, application to other network architectures is straightforward. We trained convolutional neural networks for image classification tasks on two standard benchmarks, MNIST<sup>320</sup> and Fashion-MNIST<sup>575</sup>. Incorrectly classified images are observed at higher latent distances in both cases (Appendix D, Text D.3, Table D.12, and Figure D-25).

## 6.3 Conclusions

We have demonstrated on two diverse chemical data sets that the distance in the latent space of a neural network model provides a measure of model confidence that out-performs the best established metrics (i.e., ensembles) at no additional cost beyond single model training. The distance in latent space provides an improved approach to separating low- and high-confidence points, maximizing the number of retained points for prediction at low error to enable extrapolative application of machine learning models. We introduced a technique to calibrate latent distances that required only a small fraction of out-of-sample data, enabling conversion of this distance-based metric to error estimates on the property prediction. In doing so, > 90% of model errors were bounded within 2 std. dev. of latent distance estimates, in significant improvement beyond typically over-confident ensemble estimates. Like ensembles or mc dropout, the latent space distance could still be challenged by unstable models, such as those trained on highly discontinuous properties. The latent space distance metric is general beyond the examples demonstrated here and is expected to be particularly useful in complex architectures that are normally time-consuming and difficult to train or in active learning approaches where rapid, iterative model retraining may be needed.

## 6.4 Computational details

Neural networks were trained for this work with hyperparameters selected using Hyperopt<sup>222</sup> followed by manual fine-tuning in Keras<sup>576</sup> with the Tensorflow<sup>577</sup> backend

(Appendix D, Figure D-17 and Tables D.5 and D.13). The  $\Delta E_{\text{H-L}}$  energy evaluation protocol for inorganic chemistry training data and the curated CSD<sup>184</sup> test set used molSimplify<sup>181,533</sup> to automate hybrid (i.e., B3LYP<sup>152,157,158</sup>) DFT calculations, with more details provided in Appendix D, Text D.2. For the organic chemistry test, the QM9 atomization energy data set was obtained from the literature<sup>31</sup>. In all cases, we normalize the representations and properties to make the training data have zero mean and unit variance. For calculating ensemble properties, we employ 10 sub-models trained on 10-fold cross-validation splits of the training data. For mc-dropout, we used the same 8.25% dropout as in training with 100 realizations, and we employed maximum likelihood to optimize the baseline uncertainty parameter,  $\tau$  (Appendix D, Text D.1 and Table D.2). We do not apply mc-dropout to the organic test case because it has not been developed for residual-connectivity networks. For feature space distance, we measured Euclidean distance in the normalized feature space as indicated (e.g., RAC-155<sup>511</sup>) directly. For latent distances, we use the latent space after the last hidden layer, which has the dimensionality of the model (i.e., 200 for spin splitting, 120 for the organic model).

## 6.5 Addendum: improving generalization with smart data acquisition

In related work<sup>578</sup>, we created a database of the smallest feasible ligands consisting of only two heavy atoms (from CONPS), by enumerating all possibilities and scoring them using a series of metrics, resulting in a final set of an additional 343 transition metal complexes and associated DFT-computed  $\Delta E_{\text{H-L}}$  values, termed the ‘octahedral ligand database (OHLDB)’. The small size of these ligands facilitated affordable

simulation while the great diversity of near-metal coordination environments is intended to provide robust sampling of the space of possible inner coordination shells. According to insights gained from Chapter 4, the near-metal environment has a controlling effect on the spin state, and therefore it is hypothesized that inclusion of the OHLDB complexes in the ANN could increase its generalization ability, even on large CSD structures.

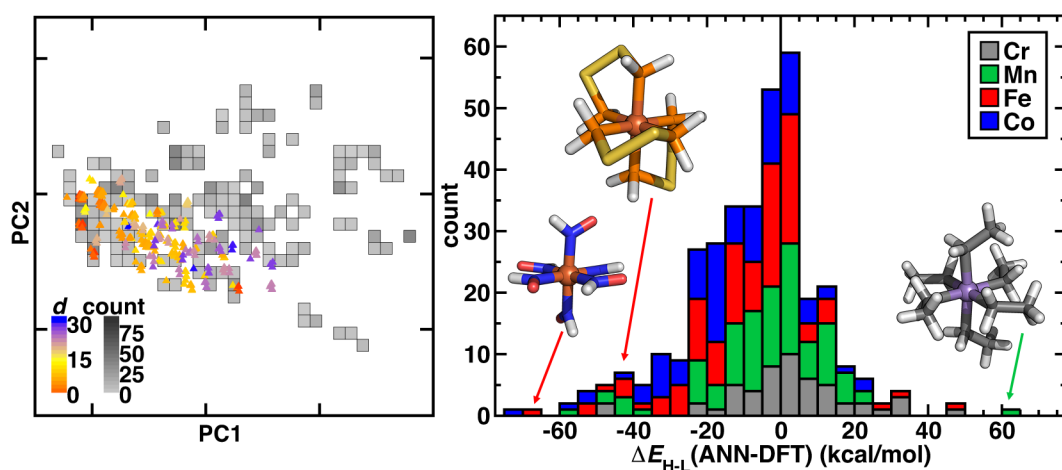


Figure 6-6: (left) Principal component analysis of new OHLDB data in the RAC-155 representation colored by Euclidean norm distance to available training data ( $d$ , colored according to inset colorbar) and overlaid on top of a 2D histogram of available data, with bins colored by count as indicated in grayscale colorbar. (right) Stacked histogram of errors (bin width: 5 kcal/mol) colored by metal type for the RAC-155/ANN prediction on OHLDB molecules with successful DFT  $\Delta E_{H-L}$  evaluations. Representative large error complexes are shown in the histogram inset (left to right): Fe(II)(HNO)<sub>6</sub>, Fe(II)(PH<sub>2</sub>SSPH<sub>2</sub>)<sub>3</sub>, and Mn(III)(CH<sub>2</sub>CH<sub>3</sub>)<sub>6</sub>.

From the successfully converged complexes that make up our curated OHLDB, we quantified the extent to which these systematically enumerated complexes reflected chemistry divergent from the 1901  $\Delta E_{H-L}$  values used for training ANNs in this chapter. To compare diversity in the chemical structures, we featurized each new

complex with the same RAC-155 representation. Although OHLDB complexes primarily lie within the convex hull of the first two principal components (PCs) in the RAC-155 representation, the overall Euclidean norm distance in feature space averaged over the ten nearest neighbors in existing data is quite large ( $> 20$ ) for a number of the complexes (Figure 6-6). The complexes indeed fall outside the convex hull of the pre-existing data but do so especially at higher PCs (i.e., 7–8), where the first eight PCs generally contain the vast majority of the variance (89%).

An alternative measure of data diversity is in property space, which we assessed first by determining if the previously trained ANN model could have predicted the  $\Delta E_{\text{H-L}}$  values exhibited by the OHLDB complexes (Figure 6-6). Overall, although a large number of complexes were well predicted, significant (e.g.,  $> 60$  kcal/mol) over- and underestimations of  $\Delta E_{\text{H-L}}$  are indicative of limited prior knowledge by the ANN (MAE = 14.3 kcal/mol) of the chemistry of the OHLDB complexes (Figure 6-6). Indeed, high error points are both chemically distinct and exhibit unexpected spin-state ordering, such as an Fe(II)(HNO)<sub>6</sub> complex ( $\Delta E_{\text{H-L}}$  ANN:  $-17.1$ , DFT: 50.1 kcal/mol), which contains an NO motif adjacent to the metal that had been absent from prior training complexes and is erroneously predicted by the ANN to be weak field in nature (Figure 6-6). Similarly, no phosphorus-coordinating metal complexes and few sulfur-containing ligands had been in training data, leading to large errors for an Fe(II) complex with bidentate PH<sub>2</sub>SSPH<sub>2</sub> ligands ( $\Delta E_{\text{H-L}}$  ANN:  $-27.8$ , DFT: 15.2 kcal/mol, Figure 6-6). Although phosphorus ligands are known to be low-spin directing, their absence from our training data means that accurate ANN predictions on such complexes cannot be expected. Finally, in some cases, the coordinating atom may be present in training data, but the chemistry is still unusual, as is the case for a strongly high-spin favoring Mn(III)(CH<sub>2</sub>CH<sub>3-</sub>)<sub>6</sub> complex (Figure 6-6). Although the ANN correctly predicts this complex to be high spin,

it cannot predict the strong high-spin stabilization observed in the DFT calculation ( $\Delta E_{\text{H-L}}$  ANN:  $-11.8$ , DFT:  $-72.0$  kcal/mol) for this saturated, negatively charged carbon ligand that is distinct from other C-coordinating ligands (e.g., CO) in our prior training data sets.

Next, we considered the extent to which OHLDB data could be used to improve ML model predictions on the large, diverse CSD complexes used in this chapter by improving coverage of metal-local environments in the training data. Because the CSD complexes were chosen to be distinct from the 1901 complexes used in the training of the ANN, the CSD set  $\Delta E_{\text{H-L}}$  MAE of 8.6 kcal/mol was much poorer than set-aside test set errors (ca. 1–3 kcal/mol) or uncertainty-controlled, out-of-sample prediction errors (ca. 4.5 kcal/mol) obtained in the earlier sections of this chapter. Notably, very high  $\Delta E_{\text{H-L}}$  prediction errors, either due to over or underestimation, were observed on the order of 20–50 kcal/mol (Figure 6-7). Incorporating OHLDB data and retraining the ANN eliminated many of these highest error points and reduced CSD set average error to 6.7 kcal/mol (Figure 6-7). Despite the fact that most of the CSD complexes are much larger in size, significant improvements are observed for complexes that had metal-adjacent coordination environments present in the OHLDB but absent in our prior data, such as coordination by NO species.



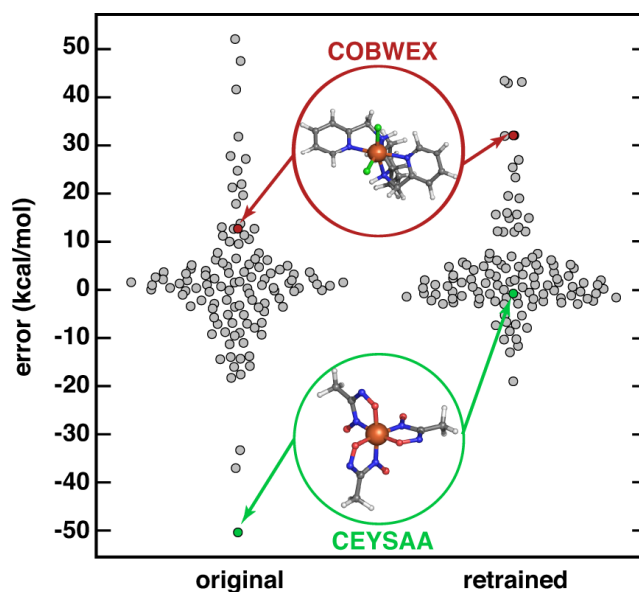


Figure 6-7: Swarm plot of RAC-155/ANN signed errors (in kcal/mol) on an out-of-sample 116 CSD structure data set (original, left) and after retraining with OHLDB data (retrained, right). The single most improved (CSD ID: CEYSAA) and worsened (CSD ID: COBWEX) points are shown in green and red insets, respectively, and have data points colored in the same manner.

In most cases model performance improved, but for select complexes model performance remained the same or worsened slightly in a manner that is not dependent on the metal center (CSD ID: COBWEX, Figure 6-7). Given that most of the CSD curated set is multidentate in nature, whereas the OHLDB is weighted toward monodentate ligands, further improvement could likely be achieved through continued systematic enumeration of a greater number of ligands of higher denticity.

Therefore, enriching machine learning models with the cheaply-acquired, targeted chemical variations in the OHLDB data improved machine learning model prediction performance on an out-of-sample CSD set even more than including a single round of active learning in Section 6.2.



# Chapter 7

## Multiobjective, multifidelity redox couple design

### Chapter summary

In this chapter, we apply a combination of first-principles simulations, multitask neural network surrogate models and uncertainty estimates based on latent space extrapolation to conduct multiobjective probabilistic optimization of transition metal complex redox couples. We are able to screen a combinatorially-assembled space of nearly 3 million unique complexes and identify leads that have both high redox potential and favorable solubility using active learning and a two dimensional expected-improvement scoring function that explicitly balances exploration of the space with extension of the Pareto frontier of these properties. Beginning with a set of diversity-orientated clusters, we perform four additional active learning cycles and are able to advance the Pareto frontier of the simulated materials with each cycle. Over the course of a few hundred DFT calculations we are able to identify promising ligands and functionalizations that simultaneously sample the extreme tails of the estimated redox potential and solubility distributions, representing an estimated 500-fold increase in sample efficiency relative to random search. This demonstrates the clear advantages in high-throughput screening afforded by data-driven methods. Additional details are provided in Appendix E.

## 7.1 Introduction

The proliferation of machine-learned ‘surrogate’ models that predict the outcome of atomistic simulations on new materials at greatly reduced cost has enormous potential to accelerate the search for new materials<sup>5,8</sup>. Such surrogate models have been extensively demonstrated in both data-rich applications such as organic chemistry<sup>24,25,27,213,451</sup> or bulk, crystalline materials<sup>75,80,231,324,389,407,579</sup>, helped in part by the availability of large databases of computed properties<sup>12,31,32,108</sup>, but also, increasingly, in data-scarce applications such as inorganic molecular systems<sup>308,533</sup> and catalysis<sup>232,279,542</sup>.

Rapid screening with these methods, combined with active learning, where the model is used to iteratively select informative points to add to the available data<sup>580,581</sup>, can help navigate very large chemical design spaces and identify lead materials from far more candidates than would be possible with simulation or experiment alone<sup>5,359</sup>. Some applications of machine-learning assisted virtual design include screening light emitting materials<sup>34</sup>, solar cells<sup>582</sup>, redox battery electrolytes<sup>583</sup>, spin-crossover materials<sup>543</sup> and perovskites<sup>237</sup>, catalysts<sup>232</sup> and drug candidates<sup>374,580</sup>.

The use of active learning can result in highly data-efficient surrogate models, decreasing requirements by an order-of-magnitude<sup>375</sup>. While for construction of a globally-accurate model, for example when preparing a neural network potential, it may be desirable to focus entirely on exploration<sup>254,375</sup>, in design problems we wish to produce a series of leads which satisfy our design criteria with as few evaluations as possible and do not care about accuracy of the ML model on poor-performing materials. Data-efficient methods that balance exploration with optimization have been successfully used to conduct one dimensional virtual materials optimization to identify crystals with targeted melting temperatures<sup>389</sup>, low-energy conformations<sup>387</sup>,

stable perovskites<sup>388</sup> and crystal structures<sup>584</sup>.

However, real materials are required to meet many different criteria simultaneously, necessitating multiobjective optimization. In chemical sciences, Pareto optimization has been previously applied to identify cost-activity tradeoffs in heterogeneous catalysis<sup>585</sup>, distillation<sup>586</sup> and continuous flow reactor operation<sup>400</sup>, with proposed techniques including evolutionary algorithms<sup>383</sup> and hierarchical methods<sup>587</sup>.

In this work, we will develop an active-learning assisted optimization algorithm to help identify transition metal complexes for redox flow batteries (RFB). RFBs are a grid-level energy storage technology<sup>64,65</sup>. They are able to decouple power delivery and cell capacity by holding the redox active species in liquid solution which is stored separately and pumped to a cell when required. This minimizes unintended discharge and makes RFBs economically-appealing for large, fixed energy storage applications<sup>66,67</sup>. Transition metal ions are good candidates for active materials in redox flow batteries due to their ability to exist stably in different oxidation states, which translates into good reversibility of the redox process and long service life. Complexation of these ions, for example with acetylacetonate- or bipyridine-based ligands<sup>68</sup>, can both tune their redox potential as well as prevent diffusion across the cell membrane. However, these complexes show modest solubility in the polar organic solvents that are desired for use in RFBs<sup>50</sup>. Recent experimental screening has demonstrated that distal (with respect to the metal) ligand functionalization can tune solubility (in acetonitrile) by orders of magnitude, while only weakly perturbing redox potential. This holds promise for design of high redox potential, highly-soluble complexes for use in RFBs by balanced ligand tuning. We will attempt to realize this objective by screening a multi-million complex design space for candidates that have both targeted redox potential and solubility properties using a combination of first-principles simulation and machine learning.

The remainder of this chapter is structured as follows: in Section 7.2.1 we will present our approach for first-principles screening of candidate complexes for RFBs, in Section 7.2.2 we will construct a multi-million complex design space over which to optimize, in Sections 7.2.3–7.2.4 we will describe the machine learning and probabilistic optimization framework used, in Section 7.3 we will present and discuss the optimization results over five generations and in Section 7.4 we will provide conclusions. A detailed description of the computational methods used is provided as Section 7.5.

## 7.2 Approach and methods

### 7.2.1 First-principles calculations

We seek to discover candidate redox couples for redox flow batteries by accelerating the screening of transition metal complexes with high  $\Delta G_{\text{solv}}$  values and good solubility in polar solvents. To facilitate screening a large compound space, we employ density functional theory (DFT). The open-shell character of first-row transition metal complexes results in a large number of possible one-electron redox processes. Consistent with prior work<sup>482,511</sup>, we compute the M(II/III) redox couple for the solvent-corrected ground state spin of the M(II) complex using a thermodynamic cycle approach. We used this spin state assignment to first compute the adiabatic, gas phase ionization potential (IP),  $\Delta E_{\text{III-II}}$ :

$$\Delta E_{\text{III-II}} = \Delta E_{\text{III}} - \Delta E_{\text{II}} \quad (7.1)$$

where  $E_{\text{II}}$  and  $E_{\text{III}}$  are the electronic energies of the gas phase, geometry optimized complexes in the LS or HS ground state of the M(II) complex and an M(III) complex that differs by single-electron removal (Appendix E, Table E.1). To obtain the solvent-corrected IP,  $\Delta G_{\text{solv}}$ , we adjust for the difference in M(II) and M(III) aqueous solvation free energies,  $\Delta G_{\text{s,water}}$ :

$$\Delta G_{\text{solv}} = \Delta E_{\text{III-II}} + \Delta G_{\text{s,water}}(\text{M(III)}) + \Delta G_{\text{s,water}}(\text{M(II)}) \quad (7.2)$$

where  $\Delta G_{\text{s,water}}$  is obtained as a single point energy on the relevant gas phase complex. Here, we neglected vibrational corrections that we incorporated in prior work<sup>511</sup> because they are small in magnitude but significantly increase the cost of computational screening (Appendix E, Figure E-1). Computational redox potentials computed with respect to a reference value (e.g., ferrocene/ferrocenium) at same level of theory have shown good agreement with experimental redox potentials<sup>207</sup>; however, we neglect this correction since it simply shifts all values by a constant factor and our focus is on relative trends. The accuracy of this protocol is estimated to be  $\sim 0.1$  eV relative to experimental measurements<sup>207</sup>.

We approximate the solubility of the transition metal complexes in the high dielectric solvents (i.e., water or polar, aprotic organic solvents such acetonitrile) favored for RFBs<sup>50,66</sup>. As a proxy for this quantity, we estimate the standard hydrophilicity (i.e., partition coefficient, logP) between octanol and water on the M(II) complex in its ground state spin as:

$$\log P = \log_{10} \frac{\Delta G_{\text{s,octanol}}}{\Delta G_{\text{s,water}}} \quad (7.3)$$

We note that logP is a powerful descriptor in phenomenological models for the experimental solubility for organic species<sup>588,589</sup> and can itself be well predicted by simple QSAR models based on molecular composition<sup>590</sup> and first-principles observables from implicit solvent calculations<sup>591</sup>. Larger logP values correspond to lower solubility in the target RFB solvents.

## 7.2.2 Design space

To construct our design space, we carried out stepwise ligand construction to generate nearly three million unique candidate redox couples (Figure 7-1). This approach is designed to provide both diversity in terms of ligand scaffolds as well as high data density for fine-tuning properties. Inspired by experimentally-accessible<sup>50,68,525,592</sup> ligands for transition metal complexes, we start with 38 unique five- or six-member heterocycles that have either a nitrogen or oxygen heteroatom to coordinate to the metal center (Appendix E, Figure E-2). These can serve as monodentate ligands directly and this range covers common ligand motifs including pyridine, imidazole, furan, thiazole, oxazole, pyrrole, as well as unsaturated variants thereof such as tetrahydrofuran and piperidine. We also form bidentate ligands from all possible heterocycle pairs, which combine with the monodentate heterocycles to create a set of around 800 unique core ligands (Appendix E, Text E.1). In order to tune solubility of the resulting complexes, we consider a set of 900 possible functionalizations, consisting of common functional groups in organic chemistry, such as methyl, carboxylic acid, amide, chloride and alcohol groups that can be added to the base ligands as substituents (Appendix E, Text E.1 and Figure E-3). We consider only adding a single unique functionalization to each base ligand resulting in approximately 700k final, chemically distinct ligands. The functionalizations are placed distal to the



metal connecting atoms as this type of modification has been experimentally observed<sup>50</sup> to modulate solubility for chromium redox couples with negligible impact on cyclic voltammetry. We restrict this study to homoleptic complexes formed from combinations of these ligands and first-row transition metal complexes Cr to Co. For simplicity, we consider only M(II) to M(III) redox processes, and we seek to identify lead complexes from this space with desired properties, i.e. high redox potential and high solubility in polar solvents.

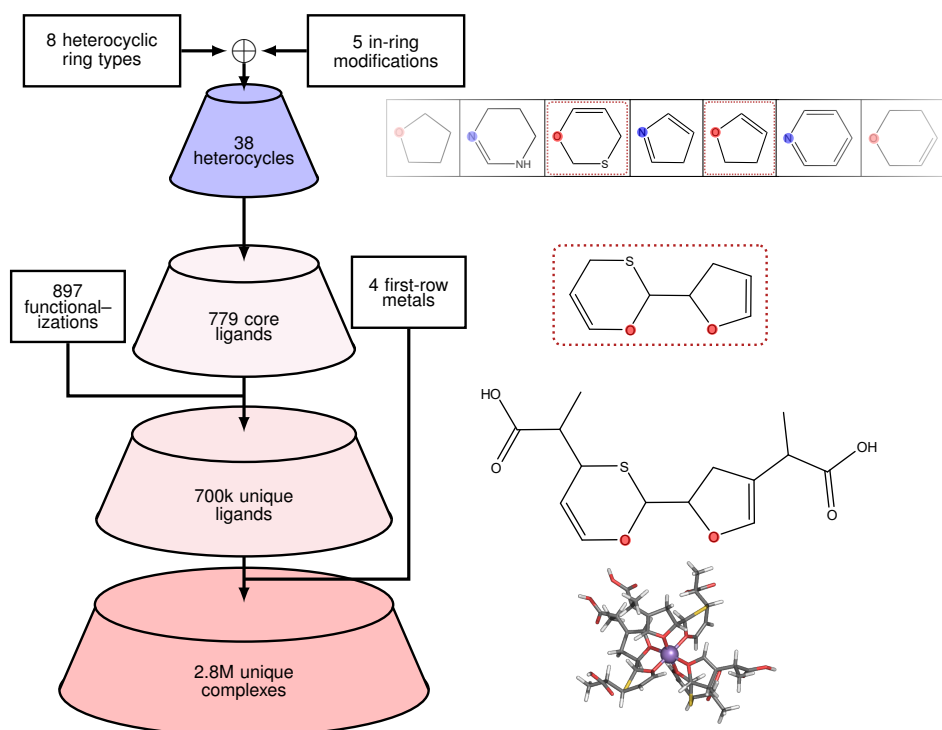


Figure 7-1: Design space of potential RFB candidates. Starting with 38 common heterocycles, core ligands are assembled from all possible combinations of one or two heterocycles, creating nearly 800 unique mono- and bi-dentate ligands that coordinate to the metal center with oxygen (red) or nitrogen (blue) atoms. These are functionalized distal from the metal with approximately 900 common chemical motifs and combined with first-row metals Cr-Fe to yield around 2.8M unique candidates.

### 7.2.3 Machine learning

To facilitate efficient optimization over this enormous space of candidates, we turn to machine learning methods, and active learning in particular. Machine learning on chemical systems depends inextricably<sup>223</sup> on the choice of numerical representation of the system of interest. Here, we use revised autocorrelations (RACs)<sup>511</sup>, a family of graph-based descriptors that were specially developed for inorganic molecular systems. RACs consist of a series of products and differences of atomic properties (e.g. nuclear charges, Pauling electronegativities) of atoms separated by a fixed number of bonds. RACs have shown good performance in predicting properties of transition metal complexes including spin state ordering<sup>511,593</sup>, metal ligand bonding<sup>180,511</sup> and redox and ionization potentials<sup>511</sup>, frontier orbital energies<sup>533</sup>, catalytic energies<sup>232</sup> and calculation outcomes<sup>259</sup> with both of kernel methods and ANNs. We use the full set of RAC-155 features<sup>511</sup>, with the exception of oxidation state, since each redox couple spans multiple oxidation states, and Hartree-Fock exchange, since all results here use 20% exact exchange. This produces a 153-dimensional representation. Since the representation is based on the molecular graph only, there is no need to compute 3D geometries for the entire design space in order to make predictions.

Here, we will primarily use a single, ‘multitask’, feed-forward ANN that predicts both endpoints ( $\Delta G_{\text{solv}}$  and logP) at the same time, though we provide comparisons with a pair of independent Gaussian process (GP) models or ANNs to predict these quantities separately. Multitask networks have shown superior predictive performance to separate, single-task models in some cases<sup>311</sup>.

In order to perform active learning, it is necessary to have a surrogate model that is both capable of predicting the properties of new complexes and providing an estimate of its own uncertainty. While the GP framework provides inherent uncertainty

estimates from the structure of the learned posterior distribution<sup>594</sup>, ANNs do not possess a similar automatic qualification of model applicability. We recently proposed a new method<sup>593</sup> for estimating this extrapolative uncertainty in chemical discovery by measuring how far a new potential complex lies from the available training in the ANN latent space, denoted  $d$  and computed as the average Euclidean distance to the 10-nearest training points in the final ANN latent space. We model this generalization error using a conditionally-Gaussian distribution:

$$\varepsilon(d) = \mathcal{N}(0, \sigma_1^2 + d\sigma_2^2) \quad (7.4)$$

where  $\sigma_1$  and  $\sigma_2$  are parameters obtained from maximum likelihood estimation on a small set of out-of-sample complexes. It was demonstrated that this metric provided as good, if not better, qualitative correlation with out-of-sample errors and superior quantitative results<sup>593</sup> compared to ensemble methods which are commonly used<sup>254,375</sup> for ANN uncertainty in chemistry, and so we use this framework to provide uncertainty estimates for our ANN models. In the case of the multitask ANN, there is only one latent distance for each complex, but  $\sigma_1$  and  $\sigma_2$  are fit independently for  $\Delta G_{\text{solv}}$  and  $\log P$ . We use 10% of available data, selected randomly, to calibrate these parameters.

Because the error distribution proposed in eq. 7.4 is Gaussian, regardless of whether the surrogate model in question is a multitask ANN or independent GP, the predicted distributions of the  $\Delta G_{\text{solv}}$  and  $\log P$  values for a new, potential complex  $x$ , are given by:

$$\begin{array}{l} \Delta G_{\text{solv}}(x) \\ \log P(y) \end{array} \sim \mathcal{N} \left( \begin{bmatrix} \hat{\mu}_{\Delta G_{\text{solv}}} \\ \hat{\mu}_{\log P} \end{bmatrix}, \begin{bmatrix} \hat{\sigma}_{\Delta G_{\text{solv}}}^2 & 0 \\ 0 & \hat{\sigma}_{\log P^2} \end{bmatrix} \right) \quad (7.5)$$

where  $\hat{\mu}_{\Delta G_{\text{solv}}}$  and  $\hat{\mu}_{\log P}$  are the predicted mean values for  $\Delta G_{\text{solv}}$  and  $\log P$  respectively, while  $\hat{\sigma}_{\Delta G_{\text{solv}}}^2$  and  $\hat{\sigma}_{\log P^2}$  are the effective variances ( $\sigma^2 = \sigma_1^2 + d\sigma_2^2$  for ANN models).

## 7.2.4 Multiobjective design framework

Equipped with machine-learned surrogate models and uncertainty estimates, we implement a probabilistic, active learning strategy<sup>385,396</sup> to identify promising candidates from the design space to simulate. This can be understood as defining an acquisition function<sup>221</sup>, which assigns a score to each potential element in the design space, and this function decides how to weight exploitation of the model (picking candidates with ‘good’ properties) and exploration (improving the model by incorporation of high-uncertainty points).

A popular acquisition function for optimization with probabilistic surrogate models is the expected improvement criterion<sup>372,386</sup>, formulated for a one-dimensional minimization as the expected decrease in the value of the objective value, or improvement,  $I$ :

$$I[\hat{y}(x)] = \max(y(x^*) - \hat{y}(x), 0) \quad (7.6)$$

at a point  $x$  with estimated value  $\hat{y}(x)$  distributed as  $p_x(\hat{y})$  (i.e. the distribution of model predictions for design  $x$ ), relative to the current best design,  $x^*$ , with known value  $y(x^*)$ ,

$$\mathbb{E}[I](x) = \int_{-\infty}^{\infty} I[\hat{y}(x)] p_x(\hat{y}) d\hat{y} \quad (7.7)$$

For Gaussian process surrogates,  $p_x(\hat{y})$  follows a Gaussian distribution around mean value  $\mu(x)$  with variance  $\sigma^2(x)$  that is different for each design. In such cases, eq. 7.7 can be integrated analytically over all possible values of  $\hat{y}$ , leading to a method known as Efficient Global Optimization<sup>372</sup> (EGO). In EGO, the acquisition function at every new point  $x$  is given in terms of the mean and standard deviations of the predictive distribution:

$$\mathbb{E}[I](x) = (y(x^*) - \mu(x)) \Phi\left(\frac{y(x^*) - \mu(x)}{\sigma(x)}\right) + \sigma(x) \phi\left(\frac{y(x^*) - \mu(x)}{\sigma(x)}\right) \quad (7.8)$$

where  $\Phi$  and  $\phi$  are the cumulative and distribution functions of the standard normal distribution. The first term encourages exploitation of the model and the second term favors exploration of high uncertainty points. EGO is the simplest Bayesian optimization method and has been widely applied in chemical design problems<sup>387–389,584</sup>, although alternatives exist, such evolutionary algorithms<sup>383</sup>, Thompson sampling<sup>396</sup>, optimal learning<sup>595</sup> and the Phoenix<sup>397</sup> algorithm, that have been suggested to have better sample efficiency or stability in chemistry problems.

In the RFB case, at least two important characteristics must be optimized simultaneously. Instead of lumping both  $\Delta G_{\text{solv}}$  and solubility estimates into a single scalar objective, we adopt a multiobjective optimization strategy that attempts to map the Pareto frontier of the design space. In Pareto optimization<sup>596</sup>, a potential new candidate is considered dominated if there is an existing point that is lower along both objective functions (gray portion of the distribution in Figure 7-2). The set of all points which are not dominated is referred to as the Pareto frontier, and represents all possible optimal tradeoffs between the design variables (black dashed in line Figure 7-2). In the context of this work, a complex is dominated if there is another complex with both a higher  $\Delta G_{\text{solv}}$  (here, lower negative  $\Delta G_{\text{solv}}$ ) and lower

logP value. We seek to map this frontier in our design space in order to understand how these variables are related. For each point on the our Pareto front, the  $\Delta G_{\text{solv}}$  cannot be improved without worsening solubility, or vice versa.

The expected improvement framework can be extended to multidimensional Pareto-optimization problems in a natural manner<sup>386,394</sup>. Instead of considering the one dimensional improvement, improvement is defined as the Euclidean distance a candidate lies beyond the current Pareto frontier (Figure 7-2). The total of the probability mass for a candidate that lies beyond the Pareto frontier defines  $\mathbb{E}[I](x)$  in direct analogy with eq. 7.7. By approximating the distance to the front by the distance to the nearest point of the frontier, analytical expressions for these integrals for independent Gaussian distributions were derived in Ref.<sup>394</sup>, which are generalizations of eq. 7.8. Due to the structure of our ANN uncertainty model<sup>593</sup>, we can directly apply these integrals to ANN surrogate models (see Section 7.2.3). We will utilize this approach to balance exploitation and exploration in ranking candidates from our design space and choose leads for simulation.

In order to accelerate the initial process, we begin by constructing surrogate models on a set of 235 precomputed  $\Delta G_{\text{solv}}$  values from prior studies<sup>308,511,543,593</sup> (Appendix E, Table E.2). However, we observed that this data was highly distinct from the proposed design space, as judged by the location of this data in the first two principal components using the RAC-155 representation (Figure 7-2, right). The pre-existing complexes are primarily monodentate and include connecting atom types (carbon) not present in the design space, and are also smaller and less symmetric on average, which likely limits the ability to extrapolate from these complexes to the design space (Appendix E, Figures E-4–E-5). Therefore, instead of using these models to select the initial generation of points, we performed k-medoids clustering to identify 300 most representative structures to simulate in the initial step. We observe these

clusters are generally well-representative of the design space (Appendix E, Text E.2, Table E.3, and Figures E-6–E-7). Of our initially-simulated clusters, we were able to obtain redox and logP for 107 distinct complexes that passed our automated checks for calculation geometry and wavefunction convergence after our time-limited calculation scheme, which is a result of a baseline success rate of 65% but requires three independent calculations to converge.

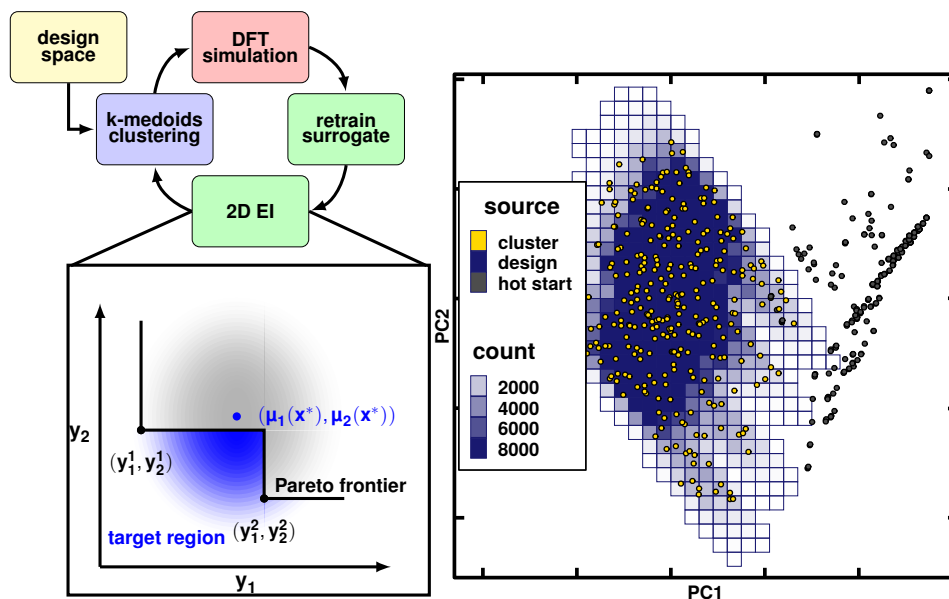


Figure 7-2: (Left) Illustration of the active learning workflow used in this work. DFT simulations are performed on cluster medoids which are used to iteratively train ML surrogate models. Thus, surrogate models score all possible candidates using 2D expected improvement, and the top scoring complexes are clustered and simulated to repeat the process. Inset: illustration of a Pareto set (black points) and frontier (dashed black line) for a 2D objective function. The distribution of values at trial point  $x$  is shown around the mean value,  $(\mu_1, \mu_2)$  and the probability mass below the frontier is shown in blue. (Right) 2D PCA based on RAC-155 representation of the design space, showing 2.8M candidates (gray, histogram filled by count), 300 cluster medoids (golden dots) and 235 existing data points (red dots).

We use this set to benchmark our initial models and select the most transferable model, which is then retrained with this additional data in order to yield the first ANN generation. This model is then used to estimate the multiobjective  $\mathbb{E}[I]$  for each as-yet-unexplored candidate in the design space. We then collect the 10k highest-ranked candidates and perform  $\mathbb{E}[I]$ -weighted k-medoids clustering to select a new batch of 100 leads for DFT simulation as the second generation. This data is used to retrain the model the process is repeated for a further 5 generations (Figure 7-2 and Appendix E, Figure E-8).

## 7.3 Results and discussion

### 7.3.1 Model selection and extrapolation from existing data

We compared the predictive ability of our multitask ANN with two alternative approaches: a pair of independent Gaussian process (GP) and single task ANN models which predict redox and logP separately. In all cases, we trained the models on 90% of our initial set of 235 existing complexes and evaluated their performance on a 10% subset held out as test data, and finally their ability to generalize to the 107 initial medoids from the design space. We were able to obtain mean absolute errors (MAEs) on test data of 0.3–0.4 eV of  $\Delta G_{\text{solv}}$  ( $\sim 6\%$  mean absolute percent error, MAPE), comparable to previous kernel methods on a simpler data<sup>511</sup>, and  $1.5\text{--}2.5 \times 10^{-4}$  for logP ( $\sim 0.5\%$  MAPE). The single task logP ANN outperforms the multitask ANN on test data while the reverse is true for redox, where the multitask ANN holds a small lead (Appendix E, Table E.4). Overall, errors on the uniformly selected test candidates are comparable. Redox errors are correlated between models, with Pear-



son  $r = 0.95$  for redox errors between multi- and single task ANNs (0.81 between multitask and GP). The logP errors are not as significantly correlated between models with  $r < 0.35$  between any pair (Appendix E, Table E.5).

Latent-distance based uncertainty allows superior control of generalization error compared with GP variance for redox prediction, where excluding the five highest uncertainty points leads to a reduction in MAE from 0.32 eV to 0.21 eV for the multitask ANN (c.f. 0.45 eV to 0.34 eV for GP standard deviation, see Appendix E, Figure E-9). The picture is less clear for logP for the best-performing single task model, but we note that it obtains similar training and test errors, and the latent distance approach only accounts for increases in error due to being unlike training data, and not baseline accuracy achieved on training data.

However, extrapolation to the generation 1 cluster medoids yields a different story – all generalization errors are substantially increased, with the redox errors at least doubling (14 – 30% MAPE, Figure 7-3). There is an order-of-magnitude increase for logP (7 – 30% MAPE, Appendix E, Figure E-10). We attribute this to the aforementioned differences in size, composition, ligand symmetry and denticity between the existing data and the design space, in line with recent observations<sup>545,552,597</sup> that uniform test partitions provide limited information about generalization capacity to different regions of chemical space. These differences also manifest in discrepancies in the distribution of the output properties between these sets, with mean (standard deviation) of calculated logP values of  $-4.10 \times 10^{-2}$  ( $1.78 \times 10^{-3}$ ) for the training data and  $4.51 \times 10^{-2}$  ( $2.66 \times 10^{-3}$ ) for the medoids. This corresponds to a magnitude difference of the means of the two distributions of around two standard deviations, indicating poor overlap in the range of output variables. The means (standard deviations) for the two  $\Delta G_{\text{solv}}$  distributions are 7.15 (1.38) and 5.62 (0.89) eV respectively for a similar difference (Appendix E, Figure E-11).

We observe similar (for redox) and stronger (for logP) correlation between errors from different models on this out-of-distribution set as compared to the test set (Appendix E, Table E.6). All uncertainty metrics agree that this data is consistently out-of-sample, with the minimum latent distance to training data (5.8 for multitask, average 7.5) being larger than the average for the uniformly-selected test data (5.6 for multitask ANN). Few points are predicted well, with nearly 60% (62 of 107) having redox errors  $> 0.5$  eV for the best multitask model, and no uncertainty metric is able to isolate a subset of complexes with comparable errors to the uniformly-selected test errors (Appendix E, Figure E-12). However, in all cases the multitask ANN provides the best generalization performance on this out-of-sample set, and so we select it as the model used to compute  $\mathbb{E}[I]$  and drive our active learning process.

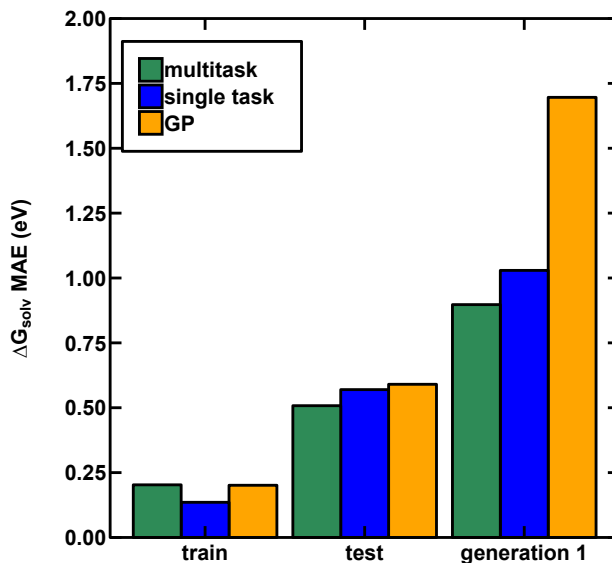


Figure 7-3: Mean absolute errors in eV for a multitask ANN (green), a single task ANN (blue) and a Gaussian process model (orange) when predicting DFT-derived  $\Delta G_{\text{solv}}$  potential on their training data, a uniform 10% test partition, and the first generation of out-of-distribution set of cluster medoids from the design space (generation 1).

### 7.3.2 Active learning process

We retrain our multitask ANN model incorporating the diversity-based cluster medoids and refer to this as the generation 1 model. These initial medoids also define an initial Pareto frontier consisting of six of 107 points (blue line in Figure 7-4), spanning six unique bidentate ligands coordinating complexes with two Fe, Co and Mn metal centers. We will only define Pareto frontiers based on DFT results, not ML predictions. The maximum  $\Delta G_{\text{solv}}$  on the initial frontier is 7.37 eV (with logP value of  $-4.38 \times 10^{-2}$ ), obtained for a large (145 atom) Mn complex with a ligand comprised of two oxygen-coordinating heterocycles. The minimum logP of  $-5.31 \times 10^{-2}$  (with redox value of 6.36 eV) is obtained for a smaller (82 atom) Fe complex with a ligand comprised of two nitrogen-coordinating heterocycles (insets in Figure 7-4). All other points lie between these two values.

We combine our ANN predictions with uncertainty estimates to calculate the probability that each candidate in the 2.8M complex design space will improve on this frontier ( $P[I]$ ), as well as the associated expected improvement ( $\mathbb{E}[I]$ ). While the probability of improvement is insensitive to the location of the proposed point on the frontier, the expected improvement weighs which region of the front might lead to the largest possible increase (Figure 7-4). We also observe the impact of approximating the distance to front by distance to the nearest point in the Pareto set introduced in Ref.<sup>394</sup>, with points equidistant between two existing points most highly scored ( $\mathbb{E}[I]$  response in Figure 7-4, right). We select 100 clusters from the top 10k candidates ranked based on  $\mathbb{E}[I]$  and perform DFT simulations. The selected 100 medoids have an average  $\mathbb{E}[I]$  value of 0.46, compared to 0.45 for the top 10k  $\mathbb{E}[I]$  candidates and 0.04 over the full space.

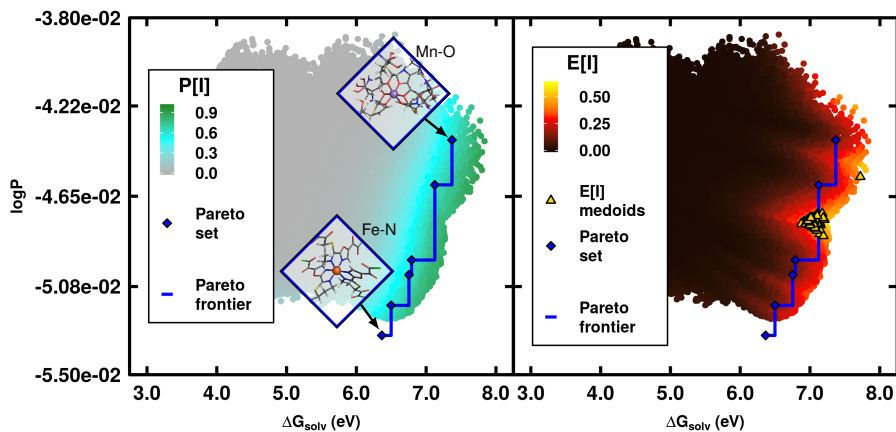


Figure 7-4: ANN predicted  $\log P$  and  $\Delta G_{\text{solv}}$  potential values for the 2.8M complex design space after incorporating generation 1 medoid data, with the Pareto frontier and DFT Pareto set shown in blue. The points are colored by the probability of improving on the front ( $P[I]$ , left) or expected improvement ( $\mathbb{E}[I]$ , right). The highest  $\Delta G_{\text{solv}}$  potential Mn-oxygen and lowest  $\log P$  Fe-nitrogen complexes are shown as insets (left) and the 100 selected medoids for the next generation are shown as triangles (right).

We use the outcome of these simulations to obtain generation 2 data and repeat the process to generate a further four sets of data, up to generation 5. The number of successful DFT calculations that pass all automated checks generally decreases as we repeat this process, leading to only 14 successful redox calculations in the generation 5 data (Appendix E, Table E.7). This may be a consequence of the algorithm seeking information about regions of the design space where there is no data and no data can be acquired, likely because the corresponding medoid calculation was also unsuccessful. Observing variations in the computed  $P[I]$  and  $\mathbb{E}[I]$  as more data is acquired, the algorithm generally becomes more pessimistic about further improvement in the frontier, with the average  $P[I]/\mathbb{E}[I]$  in the 100 cluster medoids points for generation

1 being 0.68/0.45 vs. 0.13/0.04 for generation 4 (Appendix E, Figure E-13), which motivates stopping the algorithm after 5 iterations.

The predictive power of the trained models can be assessed by calculating the ‘lookahead’ errors of each model on the data simulated in later generations, which is by definition excluded from training data. Since the lookahead error of a given model can only be evaluated on subsequently acquired data, the number of available lookahead errors for later models is smaller and we cannot evaluate lookahead error for the final generation 5 model at all. The generation 0 model, trained on pre-existing data, performs worst on tests, with the addition of the generation 1 data reducing the generation 1 redox lookahead MAE from 0.76 eV to 0.46 eV (Figure 7-5). On the final set of generation 5 data, the MAE is decreased from 1.64 eV to 0.41 eV moving from the generation 0 to the generation 2 model (including up to generation 2 data). However, adding the generation 3 or 4 data does not further improve this accuracy ( $\sim 0.42$  eV MAE), which remains higher than the training errors or held-out errors ( $\sim 0.2$ – $0.3$  eV, Appendix E, Table E.8). This is expected since  $\mathbb{E}[I]$  favors high-uncertainty points, all else being equal (e.g. eq. 7.8).

In addition to the lookahead errors, we simulate a further 300 complexes from the design space uniformly randomly to serve as a final, diverse test set. We obtain 122 converged results for these random complexes after automated screening. Adding the generation 1 model effectively halves the redox MAE of the generation 0 model from 0.70 eV to 0.41 eV, illustrating the efficacy of the k-medoids approach, but only modest accuracy increases are gained from the additional data sets down to 0.38 eV for the final generation 5 model (Figure 7-5). It is not expected that model performance will continually improve on this ‘global’ test set as more data is added, since the algorithm selectively enriches a particular region of the design space (i.e. those with high  $\Delta G_{\text{solv}}$  potential and low  $\log P$ ).

We observe a similar set of results for the logP data (Appendix E, Figure E-14), with incorporation of the first generation of data points reducing the lookahead MAE on the final generation 5 dataset by 60%, from  $5.98 \times 10^{-3}$  to  $2.36 \times 10^{-3}$ , which is further improved by additional data to  $1.77 \times 10^{-3}$  for the generation 4 model. The random test error shows a more rapid convergence, with the MAE dropping to  $1.71 \times 10^{-3}$  from  $2.93 \times 10^{-3}$  with the inclusion of the first generation data, and oscillating between  $1.59 \times 10^{-3}$  to  $1.64 \times 10^{-3}$  for the later models (Appendix E, Table E.9). For all logP models, there is consistently degraded generalization performance relative to  $\Delta G_{\text{solv}}$  models (i.e. test errors are an order of magnitude worse than train errors), which indicates overfitting and is possibly a result of the more serious mismatch between the pre-existing data and the design space (Appendix E, Figure E-11).

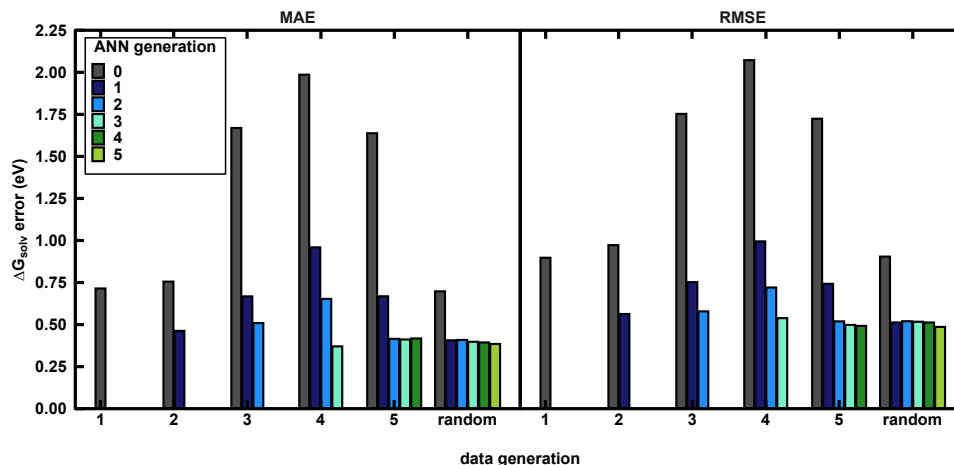


Figure 7-5: Lookahead and random-test MAE (left) and RMSE (right) error metrics for  $\Delta G_{\text{solv}}$  predictions in eV from sequential generations of multitask ANN models on the sequentially-gathered data subsequent to training each model and on a randomly-sampled set.

### 7.3.3 Analysis of leads

The generation 1 data is spread evenly across metal types since it originates from diversity orientated clustering. However, the metals selected based on  $\mathbb{E}[I]$  vary strongly across subsequent generations, initially favoring Co and Mn complexes, then Mn only complexes for generations 3 and 4, before incorporating more diverse metals again in generation 5 (Appendix E, Figure E-15). In the 194 total complexes sampled (including initial diversity-orientated medoids), there is one duplicated ligand, DbHe, an oxygen-coordinating bidentate comprised of one six-member 2H-pyran ring and one five-member, sulfur-containing 1,3-oxathiole ring, combined with a branching diaminochloropropanol (SMILES -C(=O)(C)C(N)(N)Cl) decoration that is selected with Co ( $\Delta G_{\text{solv}} = 6.17$  eV,  $\log P = -5.31 \times 10^{-2}$ ) as the metal in generation 2 and with Mn ( $\Delta G_{\text{solv}} = 6.10$  eV,  $\log P = -5.40 \times 10^{-2}$ ) as the metal in generation 4 (Appendix E, Figure E-16). In terms of the core ligands, i.e. ignoring functionalizations, the 194 complexes span 137 unique core ligands with the DbHe (the pyran-oxathiole combination) and the closely-related DeHe (an oxathiin, 4H-1,3-oxathiin, with the same oxathiole, adding a sulfur to the six member ring) the most frequently appearing, accounting for 12 observations each, 11 times with Mn as the metal, once with Co. Other frequently occurring core ligands are DeGb (5 observations, paring the oxathiin with 2,5-dihydrofuran) and DeHc (4 observations paring the oxathiin with 1,3-dioxole), leading to an abundance of five- and six member oxygen-coordinating heterocycles, particularly in later generations (Appendix E, Figures E-17–E-18). The combination of Mn metals and bidentate ligands comprised of these core heterocycles seem to consistently produce high  $\Delta G_{\text{solv}}$ , though  $\log P$  does not appear to vary as consistently with composite heterocycles, likely due to the impact of functionalizations (Appendix E, Figure E-19).

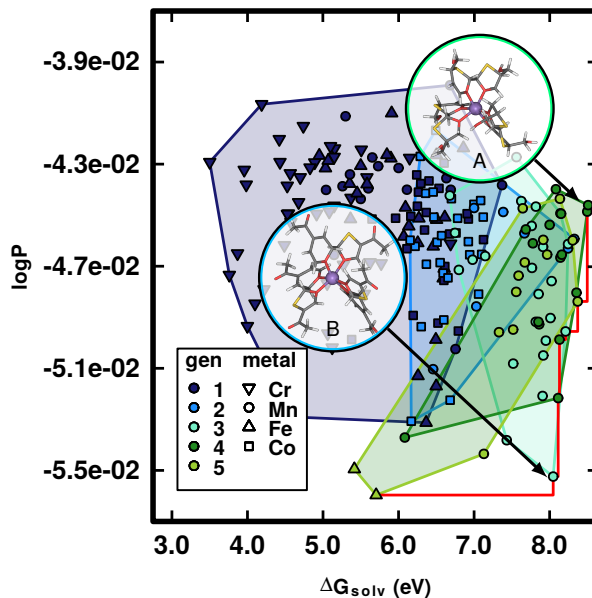


Figure 7-6:  $\Delta G_{\text{solv}}$  and logP values for complexes simulated during 5 generations of the design algorithm, colored by generation. Different symbols are used for different metals. The range of values sampled in each generation is enclosed with a convex hull. The final Pareto frontier is indicated by a red line, and the insets show renders of two Mn complexes (A and B) that are elements of the final Pareto set, as described in the text.

These repeatedly-occurring cases allow us to probe the impact of the functionalizations added to the core ligand. In the DeHe case, we obtained 12 Mn complexes with different functional groups attached to the ligand, with total system sizes for the complexes ranging from 88 to 124 atoms. The  $\Delta G_{\text{solv}}$  window of these 12 complexes from DFT is 7.33 eV (decoration SMILES: C(C)C(=C)) to 8.15 eV (decoration SMILES: CC(=C)O) with a mean of 7.86 eV (range 0.82eV). The computed logP values vary from a minimum of  $-4.99 \times 10^{-2}$  (decoration SMILES: C(=O)C(N)Cl) and a maxi-



mum of  $4.40 \times 10^{-2}$  (decoration SMILES: CC(N)), with a mean value of  $-4.70 \times 10^{-2}$  (range  $0.60 \times 10^{-2}$ ). Compared to the full range of DFT-computed values across the 194 complexes (5.00 eV for  $\Delta G_{\text{solv}}$  and  $-1.61 \times 10^{-2}$  for logP), this represents a larger variation in logP (37%) compared to  $\Delta G_{\text{solv}}$  (16%), highlighting the important role of the polar functionalizations in mediating solubility that has been observed experimentally<sup>50</sup>. However, the change in  $\Delta G_{\text{solv}}$  is not negligible and suggests that  $\Delta G_{\text{solv}}$  can be manipulated through chemical modification of the carbon site three bonds from the metal. Investigating the M(II) spin splitting energy variation between these 12 Mn complexes reveals a small change, with high spin being uniformly favored by 49 kcal/mol to 55 kcal/mol (47 kcal/mol to 52 kcal/mol without solvent effects), consistent with our prior observations that spin state ordering is less sensitive to changes in the ligand atoms distal from the metal<sup>511,533</sup>.

Considering the full range of functionalizations selected for the 194 complexes, we sample a total of 166 unique functionalizations, with seven functionalizations repeated across three complexes and 14 functionalizations repeated over two complexes (Appendix E, Table E.10). The seven most-frequently sampled functionalizations all feature highly-polar side chains, for example alcohol and carbonyl groups (6/7) or amine groups (4/7), with most (5/7) terminating in the only available halide, Cl. In total, 44 out of the total 166 functionalizations sampled feature this terminal chloride group, which is consistent with designing complexes that are preferentially soluble in polar media. Compared to the ranges of values sampled with fixed core ligand, the decoration exerts very weak control over  $\Delta G_{\text{solv}}$  but strongly constrains logP. For example, the decoration C(N)C(=O)Cl occurs three times (2 Mn and 1 Co complex), with  $\Delta G_{\text{solv}}$  ranging from 6.6 eV to 8.2 eV, but the corresponding logP values range from  $-4.91 \times 10^{-2}$  to  $-5.08 \times 10^{-2}$ . As another example, decoration C(N)(O)C(N)(O)Cl is also sampled three times (all Co complexes) with  $\Delta G_{\text{solv}}$  val-

ues ranging from 6.22 eV to 6.65 eV, and logP values ranging from  $-4.75 \times 10^{-2}$  to  $4.62 \times 10^{-2}$ .

In spite of the modest number of calculations added (87 converged complexes selected based on  $\mathbb{E}[I]$  over 5 generations), we observe that points beyond the Pareto frontier are successfully identified in each generation, improving on the best previous results with every iteration (Figure 7-6, Appendix E, Table E.11). The Pareto frontier is improved along both axis in different generations, for example an iron complex in generation 5 improves the minimum logP complex found (logP =  $-5.59 \times 10^{-2}$ ,  $\Delta G_{\text{solv}} = 5.7$  eV) over the previous best Mn complex that was established in generation 3 (logP =  $-5.55 \times 10^{-2}$ ,  $\Delta G_{\text{solv}} = 8.05$  eV). The final Pareto set consists of eight complexes, with two added in generation 5, four added in generation 4, and two added in generation 3 (Appendix E, Table E.12). This set consists of seven Mn complexes and one Fe complex. The maximum obtained  $\Delta G_{\text{solv}}$  is improved from 7.37 eV in generation 1 to 8.51 eV in generation 5 for an Mn complex with bidentate, oxygen-coordinating ligands comprised of 4H-1,3-oxathiin and 1,3-oxathiolan rings with hydroxymethyl decorations (complex A in Figure 7-6) – an increase of 15%. The minimum obtained logP is decreased from  $-5.30 \times 10^{-2}$  to  $-5.60 \times 10^{-2}$  for an Fe complex with a bidentate ligand containing the same oxathiin, this time combined with 2,5-dihydrofuran and possessing a large, very bulky halide-capped phenyl decoration. Not only are the optimal values improved independently, the compromises made on the frontier are improved; for example the logP value of the highest  $\Delta G_{\text{solv}}$  point in generation 5 is simultaneously improved from  $-4.38 \times 10^{-2}$  in generation 1 to  $-4.46 \times 10^{-2}$  in generation 5.

Analyzing this final set, it is apparent that seven Mn complexes all possess similar, large  $\Delta G_{\text{solv}}$  (mean 8.27 eV, higher than best generation 1 complex), with varied logP values ranging from  $-5.52 \times 10^{-2}$  to  $-4.46 \times 10^{-2}$ , which is only slightly higher

( $7.25 \times 10^{-4}$ ) than the Fe complex, which sacrifices 2.34 eV of  $\Delta G_{\text{solv}}$  for this marginal advantage. Instead, a Mn complex with an 2H-pyran and 1,3-oxathiole core ligand (coded as DbHe) and ketone (SMILES C(=O)C decoration), possessing a  $\Delta G_{\text{solv}}$  of 8.05 eV and logP of  $-5.52 \times 10^{-2}$ , seems to represent a better compromise (complex B in Figure 7-6).

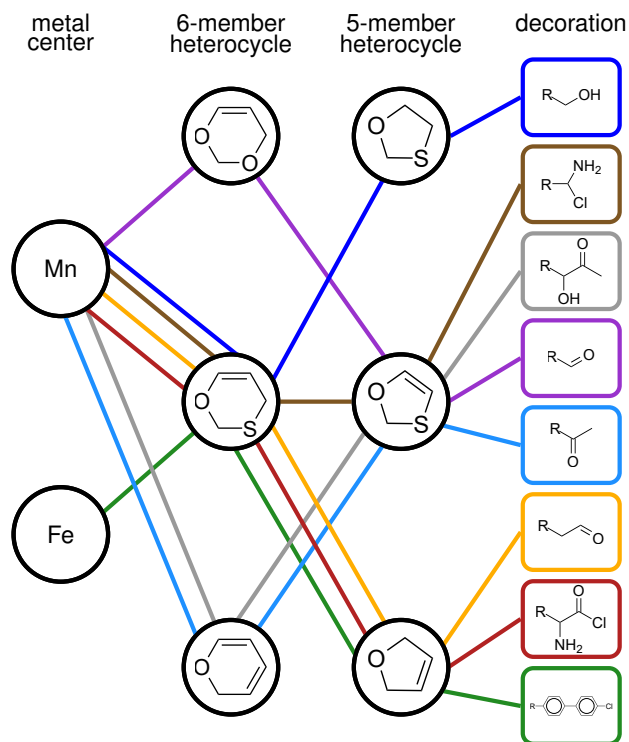


Figure 7-7: Composition of the eight complexes in the final Pareto set. Each complex consists of one metal center with bidentate ligands assembled from one 6-member and one 5-member ring, with a decoration attached symmetrically to both rings. Each complex is represented by a unique colored line.

In all cases, the complexes in this final set have bidentate, oxygen-coordinating ligands comprised of one six-member and one five-member core heterocycle (Figure 7-7).

All ligands include at least one sulfur heteroatom in one of the two heterocycles. In terms of core heterocycle composition, the oxathiin “De” heterocycle occurs 5 times, with the related 2H-pyran ring, which differs by replacing the sulfur heteroatom with a unsaturated carbon, occurring twice. The last complex features a 4H-1,3-dioxine heterocycle, replacing the sulfur heteroatom with a second oxygen. In terms of the five-member rings, 1,3-oxathiole occurs four times, while 2,5-dihydrofuran occurs three times and 1,3-oxathiolan once (Appendix E, Table E.12). These complexes are uniformly predicted to have high spin ground states by at least 25 kcal/mol, consistent with the oxygen ligands being weak field (Appendix E, Table E.11).

Therefore, these Mn complexes appear to be the most promising leads from our algorithmic approach. This is consistent with prior, unrelated and unguided DFT screening of transition metal complex redox potentials which identified Mn(II)/Mn(III) couples as the most promising first-row candidates<sup>69</sup>. Here, in contrast, the solubility of these complexes is simultaneously optimized.

In order to quantitatively assess the efficiency of the proposed approach, we compare the  $\Delta G_{\text{solv}}$  and logP values obtained for points in this Pareto set to those obtained through both the initial diversity-driven clustering (generation 1) and through random sampling (Figure 7-8, Appendix E, Figures E-20–E-21). For the  $\Delta G_{\text{solv}}$  values of the randomly-sampled data we obtain a mean value of 5.80 eV (5.62 eV for generation 1) with an empirical standard deviation of 0.89 eV (0.88 eV for generation 1). The best candidate (highest  $\Delta G_{\text{solv}}$ ) in the Pareto set lies  $\sim 3$  standard deviations above the mean for both distributions and nearly one standard deviation higher than the maximum values (0.7 standard deviations for random samples, 1.3 standard deviations for generation 1). For logP values, the mean and standard deviation for the randomly-sampled data are  $-4.45 \times 10^{-2}$  and  $2.75 \times 10^{-3}$  respectively ( $-4.45 \times 10^{-2}$  and  $2.63 \times 10^{-3}$  for generation 1), with the suggested compromise Mn complex (‘B’

in Figure 7-6) with highest solubility lying  $\sim 4.0$  standard deviations below the mean for either randomly-sampled or clustered data.

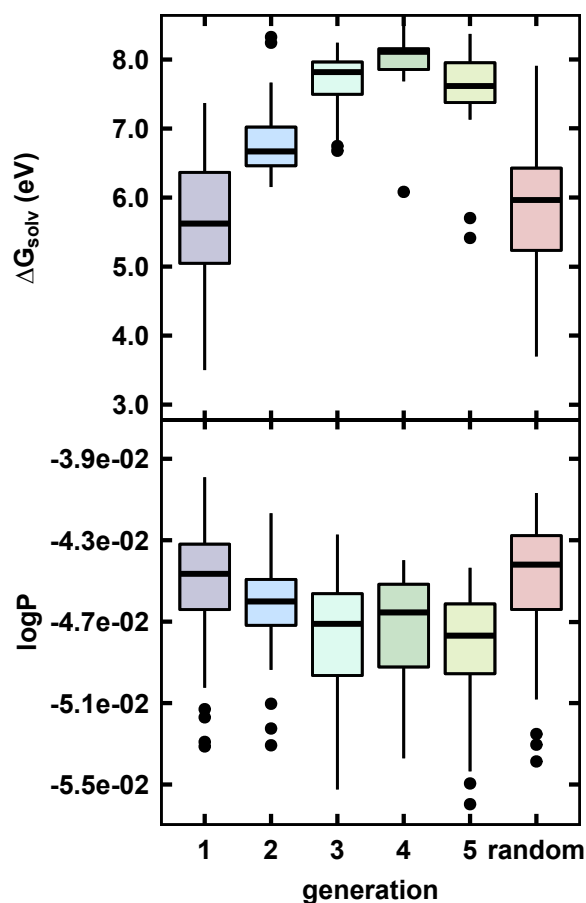


Figure 7-8: Tukey box-and-whiskers plot for  $\Delta G_{\text{solv}}$  (top) and  $\log P$  (bottom) estimated with DFT for complexes in sequential generations of the design algorithm and a uniform-randomly selected test set (indicated as ‘random’). The box indicates the interquartile range (IQR) of the data, while the line denotes the median value. The whiskers extend to the largest value closer than  $1.5 \times \text{IQR}$  and any points outside this range are denoted with black points.

Since our design approach identifies complexes that are outside the range of values

obtained from random sampling, we approximate the distribution of the randomly-sampled data with a normal distribution to allow us to estimate the quantiles of our candidates. This analysis reveals that, of the 84 complexes obtained through  $\mathbb{E}[I]$  based selection, 27 lie above the 99<sup>th</sup> percentile and 16 lie above the 99.5<sup>th</sup> percentile with respect to  $\Delta G_{\text{solv}}$  (13 below the 1<sup>st</sup> percentile and 12 below the 0.5<sup>th</sup> percentile with respect to  $\log P$ ). Considering both objectives together, we are able to identify 3 complexes that lie above the 99<sup>th</sup> percentile with respect to  $\Delta G_{\text{solv}}$  and simultaneously below the 1<sup>st</sup> percentile with respect to  $\log P$ , while our comprise candidate ('B' in Figure 7-6) is simultaneously above the 99.5<sup>th</sup> percentile and below the 0.5<sup>th</sup> respectively. Assuming the two properties are independent, this represents a  $\sim 500$  fold increase in sampling efficiency relative to random search. To probe the validity of these assumptions, we performed Shapiro-Wilk normality tests on these distributions and concluded that, while they are not normally distributed ( $p = 0.02/p = 2 \times 10^{-6}$  for redox/ $\log P$ ), the tails of the empirical data are significantly lighter than a normal distribution (i.e. the empirical data is more concentrated around the mean, Appendix E, E-21), suggesting that this analysis of acceleration is conservative. To investigate the independence assumption, we observe that there is little correlation between the randomly-sampled redox and  $\log P$  values (Pearson  $r = -0.22$ ).

## 7.4 Conclusions

In conclusion, we presented a multifidelity, multiobjective strategy for identifying promising transition metal redox couples. Leveraging geometry-free representations, we are able screen large, multi-million complex design spaces in minutes using ANNs. Our design space is combinatorially assembled from simple building blocks and con-

tains 2.8M unique complexes. We combine machine learned surrogate models and our recently-introduced latent space based measure of extrapolation uncertainty to conduct multiobjective probabilistic optimization in this design space, iteratively selecting leads for DFT simulation in a closed-loop active learning process. Our approach is able to advance the DFT-simulated Pareto frontier of redox potential and solubility with every generation of the calculations, identifying promising lead complexes that simultaneously are estimated to be soluble in polar solvents and have high redox potential from a few hundred DFT evaluations in total. We estimate that this represents at least a 500-fold enrichment over what could be obtained from random sampling and candidates that are simultaneously in the extreme tails of the estimated redox potential and logP distributions are discovered. We believe this approach would be broadly applicable to other chemical systems or objective targets, providing strong evidence for the practical enhancements that data-driven strategies can bring to virtual screening, even in challenging and data-scarce applications.

## 7.5 Computational details

### 7.5.1 First-principles methods

Two data sets are used: i) ‘hot start’ data from prior work<sup>308,511,543,593</sup> and ii) newly generated structures from the theoretical design space, both of which followed a similar protocol. The ‘hot start’ 235 transition metal complex data set contains an even distribution of Cr, Mn, Fe, and Co in predominantly heteroleptic complexes (Appendix E, Figures E-4–E-5 and Table E.2). Structures were generated with molSimplify<sup>181</sup>, which assembles complexes with ANN-predicted metal-ligand bond lengths<sup>180,308,482</sup> and uses OpenBabel<sup>77</sup> as a backend for ligand structure gener-

ation. All DFT geometry optimizations and single point energy calculations were performed with a developer version of the TeraChem<sup>105,447</sup> code. For all calculations, the B3LYP<sup>152,157,158</sup> hybrid GGA exchange-correlation functional was employed, which in TeraChem uses the VWN1-RPA<sup>149</sup> form for the LDA component of the correlation energy. The LANL2DZ effective core potential<sup>172</sup> was employed for transition metals and the 6-31G\* basis set was employed for the remaining atoms. Geometry optimizations were carried out with the TRIC<sup>498</sup> optimizer using default tolerances of  $4.5 \times 10^{-4}$  hartree/bohr for the maximum gradient and  $1 \times 10^{-6}$  hartree for the change in self-consistent field (SCF) energy between steps. Some ‘hot start’ data had employed the L-BFGS optimization using the DL-FIND<sup>449</sup> optimizer with the same thresholds.

Low spin (LS) and high spin (HS) state multiplicities for M(II)/M(III) ions grouped by nominal electron configuration were as follows: quintet-singlet for both *d4* Mn(III)/Cr(II) and *d6* Co(III)/Fe(II), sextet-doublet for *d5* Fe(III)/Mn(II), and quartet-doublet for both *d3* Cr(III) and *d7* Co(II). Unrestricted calculations were carried out except for singlet states, which were treated in the restricted formalism.

Level-shifting<sup>448</sup> values of 0.25 Ha for both virtual and occupied orbitals were applied to all complexes in the present work, a slight shift from the 1.0 and 0.1 Ha values, respectively, employed in ‘hot start’ data generation<sup>308,511,543,593</sup>. Single point energies were carried out with the conductor-like solvation model (COSMO)<sup>205</sup>, as implemented<sup>499,598</sup> in TeraChem, to implicitly model both octanol ( $\epsilon = 10.3$ ) and water ( $\epsilon = 78.39$ ) for all successful gas phase geometry optimizations. The solute cavity was built using Bondi’s van der Waals radii<sup>500</sup> for HCNOS atoms and 1.2 times the standard van der Waals radii for metals<sup>599</sup>.

Calculations were automatically submitted and monitored using molSimplify Automatic Design (mAD)<sup>533,543</sup>. The mAD workflow flags and excludes calculations that



remain unconverged after the default maximum job resubmissions or unrestricted calculations with deviations of the expectation of the  $\langle S \rangle^2$  operator of more than 1.0  $\mu\text{B}$  from the expected value. The mAD code also checks structures for intact geometries based on predefined criteria<sup>533</sup>. For the bidentate ligands studied in this work, we have loosened select angular and root mean square deviation (RMSD) thresholds from their defaults (Appendix E, Table E.13).

## 7.5.2 Machine learning methods

ANNs were trained with hyperparameters selected using Bayesian optimization<sup>390</sup> with Hyperopt<sup>222</sup> followed by manual fine-tuning in Keras<sup>576</sup> with the Tensorflow<sup>577</sup> backend. We independently optimized hyperparameters for 1500 iterations with 10% validation data for three ANN models: two independent single task networks for predicting  $\Delta G_{\text{solv}}$  and  $\log P$  and one multi-task network for predicting both quantities (Appendix E, Figure E-22 and Table E.14). We normalized all inputs and outputs to avoid inconsistency between the  $\Delta G_{\text{solv}}$  and  $\log P$  scales. At each stage, we held out a uniform-random 10% test set partition to calibrate the ANN uncertainty model<sup>593</sup> (Appendix E, Figure E-23). For ANN retraining, models were initialized with previously converged weights and using a more aggressive learning rate decay and a larger batch size for smoother optimization (Appendix E, Table E.14). We trained GP regression models to independently predict  $\Delta G_{\text{solv}}$  and  $\log P$  with single-parameter isotopic Gaussian covariance kernel models using kernlab<sup>467</sup> in R v.3.6.1<sup>456</sup>. Hyperparameters were optimized with a logspaced grid search using 10-fold CV, as implemented in CVST<sup>468</sup> (Appendix E, Table E.15 and Figure E-24). We use k-medoids clustering with the Cluster package<sup>600</sup> in R to select diverse candidates for DFT simulation.



# Chapter 8

## Concluding remarks

### 8.1 Conclusions

This thesis has demonstrated that machine-learned surrogate models can be combined with DFT simulations to identify transition metal complexes with targeted quantum mechanical properties more efficiently. The thesis began by establishing the first data-driven models for molecular inorganic chemistry, predicting the spin state ordering dependence on metal and ligand chemistry, successfully addressing the unique challenges in this area of chemical space. In the process, we identified existing representations for molecular organic chemistry are inefficient as they do not encode the unique metal-locality of properties of transition metal complexes, and showed that simple heuristic descriptors of the local metal environment based on chemical understanding of these systems provided superior performance. Combining these heuristic representations with simple ANNs models, the DFT spin state orderings of out-of-sample complexes could be predicted to around 3 kcal/mol accuracy with heuristic MCDL-25 descriptors (Chapter 3) or 1 kcal/mol for RACs (Chapter 4),

making correct qualitative ground-state assignment in effectively all cases, and only making errors in cases where the ground state is ambiguous.

Uncertainty with respect to DFT functional choice, an endemic issue for simulation of these systems, was addressed by training surrogate models on data sampled from DFT calculations with different levels of exact exchange and providing predictions of the unique, system-specific functional sensitivity of transition metal complexes, capturing variations across metal and ligand behavior. The ability to extrapolate from the Cambridge Structural Database (CSD) to out-of-distribution complexes was also investigated and was found to be highly variable, with many well-predicted ground states and a few large errors. This surrogate model predictive uncertainty was not correlated with organic chemical similarity metrics (i.e. Tanimoto distances), but could be well-predicted based on extrapolation distance in the proposed feature space, defined by the Euclidean distance between the proposed candidate and nearest example in training data.

The developed approach was also used to address the difficulty in initializing new simulations of unknown metal-ligand combinations in a spin and oxidation state dependent manner by predicting DFT equilibrium metal-ligand bond lengths. This capacity was added to the open source molSimplify toolkit, which combines these distance predictions with force field calculations on organic bonds to construct initial geometries. This enables future simulations to benefit from high quality initial geometries that were previously only available for organic systems.

A new framework for graph-based representations for transition metal complexes was developed that are capable of describing a full range of metal-local and global features while retaining chemical interpretability. In addition to improving predictive inference of trained surrogate models based on purely 2D information, the interpretability of these features allows for extraction of chemical insight from the thousands of com-

plexes used for training. A series of feature selection techniques was compared and selection based on random forests was best able to extract compact feature sets that showed good transferability across multiple tasks. Examining how the metal-local to metal-distal character of these feature sets varied for different prediction targets, providing insights into the relationship between ligand field and properties of the metal center. For example, spin state ordering appears to be strongly controlled by the immediate ligand environment (the first coordination shell) while the redox potential is more sensitive to distal ligand modifications.

This framework was applied to design SCOs from a space of five thousand candidates, exploiting the developed surrogate ANN and a newly-introduced evolutionary algorithm that could balance both model uncertainty (as captured by feature space extrapolation) and property optimization. The modified GA outperforms a naïve implementation and generates hundreds of ANN-predicted leads in a fraction of the time required for first-principles screening. We assessed a subset of these leads using DFT and identified that the majority are SCOs at a DFT level, even after accounting for additional free energy effects ignored in the training data. This approach provides at least an order-of-magnitude decrease in the number of first-principles calculations required to identify novel SCO chemistry.

Returning to the question of surrogate model uncertainty, feature space distances were found to be less well-correlated with extrapolation errors for high-dimensional feature spaces (as opposed to their good performance for simple, low-dimensional features). Therefore, chemical extrapolation in the latent space of learned ANN models was proposed as an alternative, and this was observed to provide a better qualitative description of out-of-distribution model confidence compared with feature space distances, ensemble averaging and dropout-based standard deviations. To provide an estimate of uncertainty in relevant units, we developed a simple probabilistic model

based on latent distance that can be calibrated with a small amount of out-of-sample data and gives good quantitative error bounds on both inorganic and organic datasets, as well as good results when used for active learning.

Finally, all of this machinery was combined to conduct multiobjective transition metal complex discovery for redox flow battery systems with high redox potential and favorable solubility. A combinatorial ligand-construction strategy based on experimentally accessible heterocycles was employed to enumerate a large design space of nearly 3M unique complexes, which was explored using a combination of multitask ANN, latent distance based probabilistic uncertainty and active learning based on a 2D Bayesian optimization algorithm. This method was able to extend the Pareto frontier of designs with every attempted iteration, identifying novel Mn-oxygen complexes that have desired properties at a DFT level. The method develops surrogate models that are accurate over the region of interest using a few hundred DFT calculations, and is highly efficient compared to random sampling, consistently enriching the most extreme quantiles of both redox and solubility simultaneously.

Taken together, the results of this thesis advance the field of virtual screening of transition metal complexes, addressing many idiosyncratic challenges occurring in this complicated but highly important region of chemical space. The methods developed in this thesis have been integrated into freely-available python packages that are able to select, execute and analyze DFT simulations of transition metal complexes automatically on remote computing resources, allowing new applications and search spaces to be incorporated easily and thus contributing significantly to the open source ecosystem of design tools for inorganic materials.

The developed techniques extend past purely inorganic systems and include of new values to quantify uncertainty in chemical discovery and active learning.

## 8.2 Future directions

The chemical optimization strategy developed in this thesis is expected to be transferable to other application areas for transition metal complexes, such as catalysis. While transition metal complexes play a central role in homogeneous catalysis, characterization of the catalytic ability of different metal and ligand combinations remains highly challenging. Therefore, the methods introduced here could potentially address some of the challenges in moving toward automatic design of homogeneous catalysts. In one recent application<sup>232</sup>, methods from this thesis were used to help identify catalytic iron-oxo complexes that violated intuitive scaling between frontier orbital properties and reaction thermodynamics, and it is hoped that continued work in this area will lead to the design of complexes with targeted reaction energies for important chemical transformations such as C–H bond activation and water splitting. Other systems that could potentially be treated by similar approaches are those that possess localized active regions such as metal-organic frameworks and single site catalysis.

While predicting DFT functional sensitivity is a valuable first step in understanding the applicability of relatively low-cost quantum chemistry for transition metal complexes, much more expensive multireference calculations are necessary to obtain reliable results for some systems. However, the cost and complexity of these methods is prohibitive for widespread application, so an ongoing effort in the Kulik group is focussed on identifying when it is necessary to use these methods, leveraging the data-driven framework developed here to predict the extent of multireference character on a system-specific basis. In a similar vein, many calculations on molecular inorganic systems are not successful due to pathological convergence of the geometry optimizer or the wavefunction, resulting in wasted computational resources. A

recent approach<sup>259</sup> using some of the methods developed in this thesis, has demonstrated that prediction of calculation outcomes coupled with live inspection of the wavefunction can be used to avoid or terminate unproductive calculations on the fly. More generally, one can envision an autonomous discovery workflow that can intelligently decide not only which complexes to investigate but also what method, from ANN surrogate to expensive simulation, to use in order to conduct transition metal complex design reliably, and manage convergence of the calculation dynamically to ensure the highest possibility of success.

Finally, the excitement around generative models may be disproportionate, but the potential to both conduct optimization directly in a continuous space and to move away from biases inherent in existing databases is very promising for the future of the rational molecular design. The trends in metal locality and the optimization approaches presented in thesis could be valuable in introducing generative models into inorganic chemistry.



# Bibliography

- [1] Kuntz, I. D.; Meng, E. C.; Shoichet, B. K. Structure-Based Molecular Design. *Acc. Chem. Res.* **1994**, *27*, 117–123.
- [2] von Lilienfeld, O. A.; Lins, R. D.; Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **2005**, *95*, 153002.
- [3] Boyd, D. B. *Rational Drug Design*; American Chemical Society, 1999; Chapter 22, pp 346–356.
- [4] Takeda, H.; Cometto, C.; Ishitani, O.; Robert, M. Electrons, Photons, Protons and Earth-Abundant Metal Complexes for Molecular Catalysis of CO<sub>2</sub> Reduction. *ACS Catal.* **2017**, *7*, 70–88.
- [5] Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; Amador-Bedolla, C.; Brabec, C. J.; Maruyama, B.; Persson, K. A.; Aspuru-Guzik, A. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **2018**, *3*, 5–20.
- [6] Nørskov, J. K.; Bligaard, T. The Catalyst Genome. *Angew. Chem. Int. Ed.* **2013**, *52*, 776–777.
- [7] Bogani, L.; Wernsdorfer, W. Molecular spintronics using single-molecule magnets. *Nat. Mater.* **2008**, *7*, 179–186.
- [8] Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)evolution. *ACS Cent. Sci.* **2018**, *4*, 144–152.
- [9] Balzani, V. Photochemical molecular devices. *Photochem. Photobiol. Sci.* **2003**, *2*, 459–476.

- [10] Hammer, B.; Nørskov, J. K. Theoretical surface science and catalysis—calculations and concepts. *Impact of Surface Science on Catalysis*. 2000; pp 71–129.
- [11] Kozuch, S.; Shaik, S. How to Conceptualize Catalytic Cycles? The Energetic Span Model. *Acc. Chem. Res.* **2011**, *44*, 101–110.
- [12] Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501–1509.
- [13] Ostrovsky, G. M.; Achenie, L. E. K.; Sinha, M. On the solution of mixed-integer nonlinear programming models for computer aided molecular design. *Comput. Chem.* **2002**, *26*, 645–660.
- [14] Shu, Y.; Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *J. Chem. Phys.* **2015**, *142*, 104104.
- [15] Santhanamoorthi, N.; Lo, C.-M.; Jiang, J.-C. Molecular Design of Porphyrins for Dye-Sensitized Solar Cells: A DFT/TDDFT Study. *J. Phys. Chem. Lett.* **2013**, *4*, 524–530.
- [16] Bignozzi, C. A.; Argazzi, R.; Boaretto, R.; Busatto, E.; Carli, S.; Ronconi, F.; Caramori, S. The role of transition metal complexes in dye sensitized solar devices. *Coord. Chem. Rev.* **2013**, *257*, 1472–1492.
- [17] Pepe, G.; Cole, J. M.; Waddell, P. G.; Perry, J. I. Rationalizing the suitability of rhodamines as chromophores in dye-sensitized solar cells: a systematic molecular design study. *Mol. Syst. Des. Eng.* **2016**, *1*, 416–435.
- [18] Venkatasubramanian, V.; Chan, K.; Caruthers, J. M. Computer-aided molecular design using genetic algorithms. *Comput. Chem. Eng.* **1994**, *18*, 833–844.
- [19] Baik, M.-H.; Friesner, R. A. Computing Redox Potentials in Solution: Density Functional Theory as A Tool for Rational Design of Redox Agents. *J. Phys. Chem. A* **2002**, *106*, 7407–7412.
- [20] Seh, Z. W.; Kibsgaard, J.; Dickens, C. F.; Chorkendorff, I.; Nørskov, J. K.; Jaramillo, T. F. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **2017**, *355*.

- [21] Keinan, S.; Therien, M. J.; Beratan, D. N.; Yang, W. Molecular Design of Porphyrin-Based Nonlinear Optical Materials. *J. Phys. Chem. A* **2008**, *112*, 12203–12207.
- [22] Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic Voyages into Uncharted Chemical Space Produce a Representative Library of All Possible Drug-Like Compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296–7303.
- [23] Zupan, J.; Gasteiger, J. Neural networks: A new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta* **1991**, *248*, 1–30.
- [24] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
- [25] Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.
- [26] Lubbers, N.; Smith, J. S.; Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **2018**, *148*, 241715.
- [27] Schütt, K.; Kindermans, P.; Felix, H. E. S.; Chmiela, S.; Tkatchenko, A.; Müller, K. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in Neural Information Processing Systems* 30. 2017; pp 991–1001.
- [28] Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757–1772.
- [29] Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. Found in Translation: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018**, *9*, 6091–6098.
- [30] LeCun, Y.; Bengio, Y.; Hinton, G. E. Deep learning. *Nature* **2015**, *521*, 436–444.
- [31] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.

- [32] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, 170193.
- [33] Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L. W.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **2012**, *58*, 227–235.
- [34] Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120.
- [35] Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **2018**, *4*.
- [36] Perkinson, C. F.; Tabor, D. P.; Einzinger, M.; Sheberla, D.; Utzat, H.; Lin, T.-A.; Congreve, D. N.; Bawendi, M. G.; Aspuru-Guzik, A.; Baldo, M. A. Discovery of blue singlet exciton fission molecules via a high-throughput virtual screening and experimental approach. *J. Chem. Phys.* **2019**, *151*, 121102.
- [37] Sanvito, S.; Oses, C.; Xue, J.; Tiwari, A.; Zic, M.; Archer, T.; Tozman, P.; Venkatesan, M.; Coey, M.; Curtarolo, S. Accelerated discovery of new magnets in the Heusler alloy family. *Sci. Adv.* **2017**, *3*.
- [38] Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D. Appl. Phys.* **2019**, *52*, 013001.
- [39] Zassinovich, G.; Mestroni, G.; Gladiali, S. Asymmetric hydrogen transfer reactions promoted by homogeneous transition metal catalysts. *Chem. Rev.* **1992**, *92*, 1051–1069.
- [40] Schilling, M.; Patzke, G. R.; Hutter, J.; Luber, S. Computational Investigation and Design of Cobalt Aqua Complexes for Homogeneous Water Oxidation. *J. Phys. Chem. C* **2016**, 7966–7975.
- [41] Cokoja, M.; Bruckmeier, C.; Rieger, B.; Herrmann, W. A.; Kühn, F. E. Transformation of Carbon Dioxide with Homogeneous Transition-Metal Catalysts: A Molecular Solution to a Global Challenge? *Angew. Chem. Int. Ed.* **2011**, *50*, 8510–8537.

- [42] Wang, Z. J.; Clary, K. N.; Bergman, R. G.; Raymond, K. N.; Toste, F. D. A supramolecular approach to combining enzymatic and transition metal catalysis. *Nat. Chem.* **2013**, *5*, 100.
- [43] Wiestner, M. J.; Ulmann, P. A.; Mirkin, C. A. Enzyme Mimics Based Upon Supramolecular Coordination Chemistry. *Angew. Chem. Int. Ed.* **2011**, *50*, 114–137.
- [44] Lyons, J. E.; Ellis, P. E.; Durante, V. A. Active Iron Oxo Centers for the Selective Catalytic Oxidation of Alkanes. Structure-Activity and Selectivity Relationships in Heterogeneous Catalysis. 1991; pp 99–116.
- [45] Joergensen, K. A. Transition-metal-catalyzed epoxidations. *Chem. Rev.* **1989**, *89*, 431–458.
- [46] Rohde, J.-U.; In, J.-H.; Lim, M. H.; Brennessel, W. W.; Bukowski, M. R.; Stubna, A.; Münck, E.; Nam, W.; Que, L. Crystallographic and Spectroscopic Characterization of a Nonheme Fe(IV)=O Complex. *Science* **2003**, *299*, 1037–1039.
- [47] Sahara, G.; Ishitani, O. Efficient Photocatalysts for CO<sub>2</sub> Reduction. *Inorg. Chem.* **2015**, *54*, 5096–5104.
- [48] Rao, H.; Schmidt, L. C.; Bonin, J.; Robert, M. Visible-light-driven methane formation from CO<sub>2</sub> with a molecular iron catalyst. *Nature* **2017**, *548*, 74.
- [49] Wäckerlin, C.; Donati, F.; Singha, A.; Baltic, R.; Decurtins, S.; Liu, S.-X.; Rusponi, S.; Dreiser, J. Excited Spin-State Trapping in Spin Crossover Complexes on Ferroelectric Substrates. *J. Phys. Chem. C* **2018**, *122*, 8202–8208.
- [50] Suttill, J. A.; Kucharyson, J. F.; Escalante-Garcia, I. L.; Cabrera, P. J.; James, B. R.; Savinell, R. F.; Sanford, M. S.; Thompson, L. T. Metal acetylacetonate complexes for high energy density non-aqueous redox flow batteries. *J. Mater. Chem. A* **2015**, *3*, 7929–7938.
- [51] Swart, M.; Gruden, M. Spinning around in Transition-Metal Chemistry. *Acc. Chem. Res.* **2016**, *49*, 2690–2697.
- [52] Bousseksou, A.; Molnár, G.; Matouzenko, G. Switching of Molecular Spin States in Inorganic Complexes by Temperature, Pressure, Magnetic Field and Light: Towards Molecular Devices. *Eur. J. Inorg. Chem.* **2004**, *2004*, 4353–4369.

- [53] Murray, K. S.; Oshio, H.; Real, J. A. Spin-Crossover Complexes. *Eur. J. Inorg. Chem.* **2013**, *2013*, 577–580.
- [54] Molnár, G.; Salmon, L.; Nicolazzi, W.; Terki, F.; Bousseksou, A. Emerging properties and applications of spin crossover nanomaterials. *J. Mater. Chem. C* **2014**, *2*, 1360–1366.
- [55] Bomben, P. G.; Robson, K. C. D.; Koivisto, B. D.; Berlinguette, C. P. Cyclometalated ruthenium chromophores for the dye-sensitized solar cell. *Coord. Chem. Rev.* **2012**, *256*, 1438–1450.
- [56] Pepe, G.; Cole, J. M.; Waddell, P. G.; Griffiths, J. R. D. Molecular engineering of fluorescein dyes as complementary absorbers in dye co-sensitized solar cells. *Mol. Syst. Des. Eng.* **2016**, *1*, 402–415.
- [57] Schröder, D.; Shaik, S.; Schwarz, H. Two-State Reactivity as a New Concept in Organometallic Chemistry. *Acc. Chem. Res.* **2000**, *33*, 139–145.
- [58] Harvey, J. N.; Poli, R.; Smith, K. M. Understanding the reactivity of transition metal complexes involving multiple spin states. *Coord. Chem. Rev.* **2003**, *238–239*, 347–361.
- [59] Reiher, M. Theoretical Study of the Fe(phen)<sub>2</sub>(NCS)<sub>2</sub> Spin-Crossover Complex with Reparametrized Density Functionals. *Inorg. Chem.* **2002**, *41*, 6928–6935.
- [60] Halcrow, M. A. Structure: function relationships in molecular spin-crossover complexes. *Chem. Soc. Rev.* **2011**, *40*, 4119–4142.
- [61] Zhao, X.-H.; Zhang, S.-L.; Shao, D.; Wang, X.-Y. Spin Crossover in [Fe(2-Picolylamine)<sub>3</sub>]<sup>2+</sup> Adjusted by Organosulfonate Anions. *Inorg. Chem.* **2015**, *54*, 7857–7867.
- [62] Mikolasek, M. et al. Complete Set of Elastic Moduli of a Spin-Crossover Solid: Spin-State Dependence and Mechanical Actuation. *J. Am. Chem. Soc.* **2018**, *140*, 8970–8979.
- [63] Boonprab, T.; Lee, S. J.; Telfer, S. G.; Murray, K. S.; Phonsri, W.; Chastanet, G.; Collet, E.; Trzop, E.; Jameson, G. N. L.; Harding, P.; Harding, D. J. The First Observation of Hidden Hysteresis in an Iron(III) Spin-Crossover Complex. *Angew. Chem. Int. Ed.* **2019**, *58*, 11811–11815.

- [64] de León, C. P.; Frías-Ferrer, A.; González-García, J.; Szánto, D. A.; Walsh, F. C. Redox flow cells for energy conversion. *J. Power Sources* **2006**, *160*, 716–732.
- [65] Dunn, B.; Kamath, H.; Tarascon, J.-M. Electrical Energy Storage for the Grid: A Battery of Choices. *Science* **2011**, *334*, 928–935.
- [66] Skyllas-Kazacos, M.; Chakrabarti, M. H.; Hajimolana, S. A.; Mjalli, F. S.; Saleem, M. Progress in Flow Battery Research and Development. *J. Electrochem. Soc.* **2011**, *158*, R55–R79.
- [67] Weber, A. Z.; Mench, M. M.; Meyers, J. P.; Ross, P. N.; Gostick, J. T.; Liu, Q. Redox flow batteries: a review. *J. Appl. Electrochem.* **2011**, *41*, 1137.
- [68] Matsuda, Y.; Tanaka, K.; Okada, M.; Takasu, Y.; Morita, M.; Matsumura-Inoue, T. A rechargeable redox battery utilizing ruthenium complexes with non-aqueous organic electrolyte. *J. Appl. Electrochem.* **1988**, *18*, 909–914.
- [69] Kirby, F.; Burnea, B.; Shi, H.; Ko, K. C.; Lee, J. Y. Reduction potential tuning of first row transition metal M(III)/M(II) (M=Cr, Mn, Fe, Co, Ni) hexadentate complexes for viable aqueous redox flow battery catholytes: A DFT study. *Electrochim. Acta* **2017**, *246*, 156–164.
- [70] Popov, I. A.; Mehio, N.; Chu, T.; Davis, B. L.; Mukundan, R.; Yang, P.; Batista, E. R. Impact of Ligand Substitutions on Multielectron Redox Properties of Fe Complexes Supported by Nitrogenous Chelates. *ACS Omega* **2018**, *3*, 14766–14778.
- [71] Popov, I. A.; Davis, B. L.; Mukundan, R.; Batista, E. R.; Yang, P. Catalyst-Inspired Charge Carriers for High Energy Density Redox Flow Batteries. *Front. Phys.* **2019**, *6*, 141.
- [72] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 675–679.
- [73] Cramer, C. J.; Truhlar, D. G. Density functional theory for transition metals and transition metal chemistry. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757–10816.
- [74] Reiher, M. A Theoretical Challenge: Transition-Metal Compounds. *CHIMIA Int. J. Chem.* **2009**, *63*, 140–145.

- [75] Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218–226.
- [76] Landrum, G. RDKit: Open-source cheminformatics. <http://www.rdkit.org>.
- [77] O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
- [78] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.
- [79] Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.
- [80] Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
- [81] Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.
- [82] Peplow, M. Organic synthesis: The robo-chemist. *Nature* **2014**, *512*, 20–22.
- [83] Houben, C.; Lapkin, A. A. Automatic discovery and optimization of chemical processes. *Curr. Opin. Chem. Eng.* **2015**, *9*, 1–7.
- [84] Fitzpatrick, D. E.; Battilocchio, C.; Ley, S. V. Enabling Technologies for the Future of Chemical Synthesis. *ACS Cent. Sci.* **2016**, *2*, 131–138.
- [85] Nikolaev, P.; Hooper, D.; Webber, F.; Rao, R.; Decker, K.; Krein, M.; Poleski, J.; Barto, R.; Maruyama, B. Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2016**, *2*, 16031.



- [86] Pendleton, I. M.; Cattabriga, G.; Li, Z.; Najeeb, M. A.; Friedler, S. A.; Norquist, A. J.; Chan, E. M.; Schrier, J. Experiment Specification, Capture and Laboratory Automation Technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Commun.* **2019**, *9*, 846–859.
- [87] Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **2017**, *17*, 97.
- [88] Senkan, S. M. High-throughput screening of solid-state catalyst libraries. *Nature* **1998**, *394*, 350–353.
- [89] Li, Z.; Najeeb, M. A.; Alves, L.; Sherman, A.; Parrilla, P. C.; Pendleton, I. M.; Zeller, M.; Schrier, J.; Norquist, A. J.; Chan, E. Robot-Accelerated Perovskite Investigation and Discovery (RAPID): 1. Inverse Temperature Crystallization. 2019.
- [90] Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.
- [91] Becke, A. D. Perspective: Fifty years of density-functional theory in chemical physics. *J. Chem. Phys.* **2014**, *140*, 18A301.
- [92] Hammer, B.; Nørskov, J. K. Electronic factors determining the reactivity of metal surfaces. *Surf. Sci.* **1995**, *343*, 211–220.
- [93] Goodpaster, J. D.; Bell, A. T.; Head-Gordon, M. Identification of Possible Pathways for C-C Bond Formation during Electrochemical Reduction of [CO<sub>2</sub>]: New Theoretical Insights from an Improved Electrochemical Model. *J. Phys. Chem. Lett.* **2016**, 1471–1477.
- [94] Siegbahn, P. E. M.; Borowski, T. Modeling Enzymatic Reactions Involving Transition Metals. *Acc. Chem. Res.* **2006**, *39*, 729–738.
- [95] Kulik, H. J.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J. Ab Initio Quantum Chemistry for Protein Structures. *J. Phys. Chem. B* **2012**, *116*, 12501–12509.
- [96] Jensen, K. P.; Ryde, U. How O<sub>2</sub> Binds to Heme: Reasons for Rapid Binding and Spin Inversion. *J. Biol. Chem.* **2004**, *279*, 14561–14569.
- [97] Nørskov, J. K.; Rossmeisl, J.; Logadottir, A.; Lindqvist, L.; Kitchin, J. R.; Bligaard, T.; Jónsson, H. Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode. *J. Phys. Chem. B* **2004**, *108*, 17886–17892.

- [98] Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy Environ. Sci.* **2014**, *7*, 698–704.
- [99] Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **2009**, *1*, 37–46.
- [100] Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014**, *6*, 1044–1048.
- [101] Medford, A. J.; Vojvodic, A.; Hummelshøj, J. S.; Voss, J.; Abild-Pedersen, F.; Studt, F.; Bligaard, T.; Nilsson, A.; Nørskov, J. K. From the Sabatier principle to a predictive theory of transition-metal heterogeneous catalysis. *J. Catal.* **2015**, *328*, 36–42.
- [102] Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. Accelerating Resolution-of-the-Identity Second-Order Moller-Plessest Quantum Chemistry Calculations with Graphical Processing Units. *J. Phys. Chem. A* **2008**, *112*, 2049–2057.
- [103] Ufimtsev, I. S.; Martínez, T. J. Quantum Chemistry on Graphical Processing Units. 1. Strategies for Two-Electron Integral Evaluation. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.
- [104] Ufimtsev, I. S.; Martínez, T. J. Quantum Chemistry on Graphical Processing Units. 2. Direct Self-Consistent-Field Implementation. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.
- [105] Ufimtsev, I. S.; Martínez, T. J. Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619–2628.
- [106] Song, C.; Wang, L.-P.; Martínez, T. J. Automated Code Engine for Graphical Processing Units: Application to the Effective Core Potential Integrals and Gradients. *J. Chem. Theory Comput.* **2016**, *12*, 92–106.

- [107] Andermatt, S.; Cha, J.; Schiffmann, F.; VandeVondele, J. Combining Linear-Scaling DFT with Subsystem DFT in Born–Oppenheimer and Ehrenfest Molecular Dynamics Simulations: From Molecules to a Virus in Solution. *J. Chem. Theory Comput.* **2016**, *12*, 3214–3227.
- [108] Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
- [109] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241–2251.
- [110] Sarker, P.; Harrington, T.; Toher, C.; Oses, C.; Samiee, M.; Maria, J.-P.; Brenner, D. W.; Vecchio, K. S.; Curtarolo, S. High-entropy high-hardness metal carbides discovered by entropy descriptors. *Nat. Commun.* **2018**, *9*, 4980.
- [111] Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276–285.
- [112] Lewis, R. A.; Wood, D. Modern 2D QSAR for drug discovery. *WIREs Comput. Mol Sci.* **2014**, *4*, 505–522.
- [113] Cohen, N.; Benson, S. W. Estimation of heats of formation of organic compounds by additivity methods. *Chem. Rev.* **1993**, *93*, 2419–2438.
- [114] Gordon, M. S.; Fedorov, D. G.; Pruitt, S. R.; Slipchenko, L. V. Fragmentation Methods: A Route to Accurate Calculations on Large Systems. *Chem. Rev.* **2012**, *112*, 632–672.
- [115] Collins, M. A. Ab initio lattice dynamics of nonconducting crystals by systematic fragmentation. *J. Chem. Phys.* **2011**, *134*, 164110.
- [116] Verlet, L. Computer “Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98–103.

- [117] Tuckerman, M. E.; Martyna, G. J. Understanding Modern Molecular Dynamics: Techniques and Applications. *J. Phys. Chem. B* **2000**, *104*, 159–178.
- [118] Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024–10035.
- [119] Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *J. Phys. Chem. B* **2010**, *114*, 2549–2564.
- [120] Robertson, M. J.; Qian, Y.; Robinson, M. C.; Tirado-Rives, J.; Jorgensen, W. L. Development and Testing of the OPLS-AA/M Force Field for RNA. *J. Chem. Theory Comput.* **2019**, *15*, 2734–2742.
- [121] Levine, I. N. *Quantum Chemistry*; Pearson Prentice Hall, 2009.
- [122] Szabo, A.; Ostlund, N. S. *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation, 1989.
- [123] Born, M.; Oppenheimer, R. Zur Quantentheorie der Molekeln. *Ann. Phys.* **1927**, *389*, 457–484.
- [124] Helgaker, T.; Jørgensen, P.; Olsen, J. *Mol. Electron. Theory*; John Wiley & Sons, Ltd: Chichester, UK, 2000.
- [125] Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
- [126] Stewart, J. J. P. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221–264.
- [127] Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys. Rev. B* **1995**, *51*, 12947–12957.
- [128] Koskinen, P.; Mäkinen, V. Density-functional tight-binding for beginners. *Comput. Mater. Sci.* **2009**, *47*, 237–253.
- [129] Čížek, J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell–Type Expansion Using Quantum–Field Theoretical Methods. *J. Chem. Phys.* **1966**, *45*, 4256–4266.

- [130] Čížek, J. *Advances in Chemical Physics*; John Wiley & Sons, Ltd, 1969; pp 35–89.
- [131] Kümmel, H. G. A Biography of the Coupled cluster method. *Int. J. Mod. Phys. B* **2003**, *17*, 5311–5325.
- [132] Knowles, P. J.; Handy, N. C. A new determinant-based full configuration interaction method. *Chem. Phys. Lett.* **1984**, *111*, 315–321.
- [133] Sherrill, C. D.; Schaefer, H. F. The Configuration Interaction Method: Advances in Highly Correlated Approaches. *Advances in Quantum Chemistry*. 1999; pp 143–269.
- [134] Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618–622.
- [135] Head-Gordon, M.; Pople, J. A.; Frisch, M. J. MP2 energy evaluation by direct methods. *Chem. Phys. Lett.* **1988**, *153*, 503–506.
- [136] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289–320.
- [137] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The  $\Delta$ -Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087–2096.
- [138] Burke, K. Perspective on density functional theory. *J. Chem. Phys.* **2012**, *136*, 150901.
- [139] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864–B871.
- [140] Eschrig, H. *The fundamentals of density functional theory*; Springer, 1996.
- [141] Smargiassi, E.; Madden, P. A. Orbital-free kinetic-energy functionals for first-principles molecular dynamics. *Phys. Rev. B* **1994**, *49*, 5220–5226.
- [142] Gavini, V.; Bhattacharya, K.; Ortiz, M. Quasi-continuum orbital-free density-functional theory: A route to multi-million atom non-periodic DFT calculation. *J. Mech. Phys. Solids* **2007**, *55*, 697–718.

- [143] Yao, K.; Parkhill, J. Kinetic Energy of Hydrocarbons as a Function of Electron Density and Convolutional Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 1139–1147.
- [144] Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133–A1138.
- [145] Burke, K. *The ABC of DFT*; Department of Chemistry, University of California, 2007.
- [146] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- [147] Ceperley, D. M.; Alder, B. J. Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **1980**, *45*, 566–569.
- [148] Ortiz, G.; Ballone, P. Correlation energy, structure factor, radial distribution function, and momentum distribution of the spin-polarized uniform electron gas. *Phys. Rev. B* **1994**, *50*, 1391–1405.
- [149] Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- [150] Jones, R. O.; Gunnarsson, O. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.* **1989**, *61*, 689–746.
- [151] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098–3100.
- [152] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.
- [153] Perdew, J. P. Accurate Density Functional for the Energy: Real-Space Cutoff of the Gradient Expansion for the Exchange Hole. *Phys. Rev. Lett.* **1985**, *55*, 1665–1668.
- [154] Perdew, J. P.; Kurth, S.; Zupan, A.; Blaha, P. Accurate Density Functional with Correct Formal Properties: A Step Beyond the Generalized Gradient Approximation. *Phys. Rev. Lett.* **1999**, *82*, 2544–2547.

- [155] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.
- [156] Perdew, J. P.; Schmidt, K. Jacob’s ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings* **2001**, *577*, 1–20.
- [157] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Chem. Phys.* **1994**, *98*, 11623–11627.
- [158] Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648.
- [159] Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- [160] Ernzerhof, M.; Scuseria, G. E. Assessment of the Perdew–Burke–Ernzerhof exchange-correlation functional. *J. Chem. Phys.* **1999**, *110*, 5029–5036.
- [161] Bauschlicher, C. W. A comparison of the accuracy of different functionals. *Chem. Phys. Lett.* **1995**, *246*, 40–44.
- [162] Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. Validation of Exchange-Correlation Functionals for Spin States of Iron Complexes. *J. Phys. Chem. A* **2004**, *108*, 5479–5483.
- [163] Gerber, I. C.; Ángyán, J. G. Hybrid functional with separated range. *Chem. Phys. Lett.* **2005**, *415*, 100–105.
- [164] Vydrov, O. A.; Scuseria, G. E. Assessment of a long-range corrected hybrid functional. *J. Chem. Phys.* **2006**, *125*, 234109.
- [165] Refaely-Abramson, S.; Baer, R.; Kronik, L. Fundamental and excitation gaps in molecules of relevance for organic photovoltaics from an optimally tuned range-separated hybrid functional. *Phys. Rev. B* **2011**, *84*, 075144.
- [166] Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003**, *118*, 8207–8215.

- [167] Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **2006**, *125*, 224106.
- [168] Pribram-Jones, A.; Gross, D. A.; Burke, K. DFT: A Theory Full of Holes? *Annu. Rev. Phys. Chem.* **2015**, *66*, 283–304.
- [169] Grimme, S. Semiempirical GGA-type density functional constructed with a long-range dispersion correction. *J. Comput. Chem.* **2006**, *27*, 1787–1799.
- [170] Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- [171] Grimme, S. Density functional theory with London dispersion corrections. *WIREs Comput. Mol. Sci.* **2011**, *1*, 211–228.
- [172] Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *J. Chem. Phys.* **1985**, *82*, 270–283.
- [173] Shriver, D. F.; Atkins, P. W. *Inorganic chemistry.*, 3rd ed.; W.H. Freeman and Co., 1999.
- [174] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
- [175] Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and other widely available force fields for conformational energies and for intermolecular-interaction energies and geometries. *J. Comput. Chem.* **1999**, *20*, 730–748.
- [176] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100–D1107.
- [177] Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.



- [178] Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750.
- [179] Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.
- [180] Janet, J. P.; Liu, F.; Nandy, A.; Duan, C.; Yang, T.; Lin, S.; Kulik, H. J. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorg. Chem.* **2019**, *58*, 10592–10606.
- [181] Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106–2117.
- [182] Rappe, A. K.; Colwell, K. S.; Casewit, C. J. Application of a universal force field to metal complexes. *Inorg. Chem.* **1993**, *32*, 3438–3450.
- [183] Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr., Sect. B* **2002**, *58*, 380–388.
- [184] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B* **2016**, *72*, 171–179.
- [185] Gaggioli, C. A.; Stoneburner, S. J.; Cramer, C. J.; Gagliardi, L. Beyond Density Functional Theory: The Multiconfigurational Approach To Model Heterogeneous Catalysis. *ACS Catal.* **2019**, *9*, 8481–8502.
- [186] Song, S.; Kim, M.-C.; Sim, E.; Benali, A.; Heinonen, O.; Burke, K. Benchmarks and Reliable DFT Results for Spin Gaps of Small Ligand Fe(II) Complexes. *J. Chem. Theory Comput.* **2018**, *14*, 2304–2311.
- [187] Jiang, W.; DeYonker, N. J.; Determan, J. J.; Wilson, A. K. Toward Accurate Theoretical Thermochemistry of First Row Transition Metal Complexes. *J. Phys. Chem. A* **2012**, *116*, 870–885.
- [188] Roos, B. O. *Advances in Chemical Physics*; John Wiley & Sons, Ltd, 1987; pp 399–445.
- [189] Roos, B. O.; Andersson, K.; Fülcher, M. P.; Malmqvist, P.-å.; Serrano-Andrés, L.; Pierloot, K.; Merchán, M. *Advances in Chemical Physics*; John Wiley & Sons, 1996; pp 219–331.

- [190] Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12*, 1760–1771.
- [191] Bao, J. J.; Dong, S. S.; Gagliardi, L.; Truhlar, D. G. Automatic Selection of an Active Space for Calculating Electronic Excitation Spectra by MS-CASPT2 or MC-PDFT. *J. Chem. Theory Comput.* **2018**, *14*, 2017–2025.
- [192] Mortensen, S. R.; Kepp, K. P. Spin Propensities of Octahedral Complexes From Density Functional Theory. *J. Phys. Chem. A* **2015**, *119*, 4041–4050.
- [193] Coskun, D.; Jerome, S. V.; Friesner, R. A. Evaluation of the Performance of the B3LYP, PBE0, and M06 DFT Functionals, and DBLOC-Corrected Versions, in the Calculation of Redox Potentials and Spin Splittings for Transition Metal Containing Systems. *J. Chem. Theory Comput.* **2016**, *12*, 1121–8.
- [194] Ganzenmüller, G.; Berkaine, N.; Fouqueau, A.; Casida, M. E.; Reiher, M. Comparison of density functionals for differences between the high- ( $5T_{2g}$ ) and low- ( $1A_{1g}$ ) spin states of iron(II) compounds. IV. Results for the ferrous complexes  $[\text{Fe}(\text{L})(\text{'NHS}_4)]$ . *J. Chem. Phys.* **2005**, *122*, 234321.
- [195] Kepenekian, M.; Calborean, A.; Vetere, V.; Le Guennic, B.; Robert, V.; Maldivi, P. Toward Reliable DFT Investigations of Mn-Porphyrins through CASPT2/DFT Comparison. *J. Chem. Theory Comput.* **2011**, *7*, 3532–3539.
- [196] Droghetti, A.; Alfe, D.; Sanvito, S. Assessment of density functional theory for iron(II) molecules across the spin-crossover transition. *J. Chem. Phys.* **2012**, *137*, 124303.
- [197] Ioannidis, E. I.; Kulik, H. J. Towards quantifying the role of exact exchange in predictions of transition metal complex properties. *J. Chem. Phys.* **2015**, *143*, 034104.
- [198] Reiher, M.; Salomon, O.; Hess, B. A. Reparameterization of hybrid functionals based on energy differences of states of different multiplicity. *Theor. Chem. Acc.* **2001**, *107*, 48–55.
- [199] Salomon, O.; Reiher, M.; Hess, B. A. Assertion and validation of the performance of the B3LYP\* functional for the first transition metal row and the G2 test set. *J. Chem. Phys.* **2002**, *117*, 4729–4737.
- [200] Siig, O. S.; Kepp, K. P. Iron(II) and Iron(III) Spin Crossover: Toward an Optimal Density Functional. *J. Phys. Chem. A* **2018**, *122*, 4208–4217.

- [201] Liu, F.; Yang, T.; Yang, J.; Xu, E.; Bajaj, A.; Kulik, H. J. Bridging the Homogeneous-Heterogeneous Divide: Modeling Spin for Reactivity in Single Atom Catalysis. *Front. Chem.* **2019**, *7*, 219.
- [202] Radoń, M.; Gąssowska, K.; Szklarzewicz, J.; Broclawik, E. Spin-State Energetics of Fe(III) and Ru(III) Aqua Complexes: Accurate ab Initio Calculations and Evidence for Huge Solvation Effects. *J. Chem. Theory Comput.* **2016**, *12*, 1592–1605.
- [203] Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R. Calculation of One-Electron Redox Potentials Revisited. Is It Possible to Calculate Accurate Potentials with Density Functional Methods? *J. Phys. Chem. A* **2009**, *113*, 6745–6750.
- [204] Paulsen, H.; Duelund, L.; Winkler, H.; Toftlund, H.; Trautwein, A. X. Free Energy of Spin-Crossover Complexes Calculated with Density Functional Methods. *Inorg. Chem.* **2001**, *40*, 2201–2203.
- [205] Klamt, A.; Schüürmann, G. COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799–805.
- [206] Hughes, T. F.; Friesner, R. A. Correcting Systematic Errors in DFT Spin-Splitting Energetics for Transition Metal Complexes. *J. Chem. Theory Comput.* **2011**, *7*, 19–32.
- [207] Konezny, S. J.; Doherty, M. D.; Luca, O. R.; Crabtree, R. H.; Soloveichik, G. L.; Batista, V. S. Reduction of Systematic Uncertainty in DFT Redox Potentials of Transition-Metal Complexes. *J. Phys. Chem. C* **2012**, *116*, 63490–6356.
- [208] Vapnik, V. *The nature of statistical learning theory*; Springer-Verlag New York, 2000.
- [209] Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction, 2<sup>nd</sup> Edition*; Springer series in statistics; Springer, 2009.
- [210] Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* **2001**, *16*, 199–231.

- [211] Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
- [212] Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255–5264.
- [213] Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203.
- [214] Artrith, N.; Morawietz, T.; Behler, J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B* **2011**, *83*, 153101.
- [215] Morawietz, T.; Behler, J. A Density-Functional Theory-Based Neural Network Potential for Water Clusters Including van der Waals Corrections. *J. Phys. Chem. A* **2013**, *117*, 7356–7366.
- [216] Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436–9444.
- [217] Li, H.; Collins, C.; Tanha, M.; Gordon, G. J.; Yaron, D. J. A Density Functional Tight Binding Layer for Deep Learning of Chemical Hamiltonians. *J. Chem. Theory Comput.* **2018**, *14*, 5764–5776.
- [218] Zilian, D.; Sotriffer, C. A. SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2013**, *53*, 1923–1933.
- [219] Vapnik, V. N.; Chervonenkis, A. Y. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications* **1971**, *16*, 264–280.
- [220] Efron, B.; Tibshirani, R. Improvements on Cross-Validation: The 632+ Bootstrap Method. *J. Am. Stat. Assoc.* **1997**, *92*, 548–560.

- [221] Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances in Neural Information Processing Systems* 25. 2012; pp 2951–2959.
- [222] Bergstra, J.; Cox, D. D.; Yamins, D. Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. *Proceedings of the 12<sup>th</sup> Python in science conference* **2013**, 13–20.
- [223] Huang, B.; von Lilienfeld, O. A. Communication: Understanding molecular representations in machine learning: The role of uniqueness and target similarity. *J. Chem. Phys.* **2016**, *145*, 161102.
- [224] Polishchuk, P. Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639.
- [225] Willatt, M. J.; Musil, F.; Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **2018**, *20*, 29661–29668.
- [226] Imbalzano, G.; Anelli, A.; Giofr , D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730.
- [227] Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.
- [228] G mez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hern andez-Lobato, J. M.; S nchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.
- [229] Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10*, 370–377.
- [230] Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *Advances in Neural Information Processing Systems* 30. 2017; pp 2607–2616.

- [231] Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H. Machine-Learning-Augmented Chemisorption Model for CO<sub>2</sub> Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528–3533.
- [232] Nandy, A.; Zhu, J.; Janet, J. P.; Duan, C.; Getman, R. B.; Kulik, H. J. Machine Learning Accelerates the Discovery of Design Rules and Exceptions in Stable Metal–Oxo Intermediate Formation. *ACS Catal.* **2019**, *9*, 8243–8255.
- [233] Fernandez, M.; Boyd, P. G.; Daff, T. D.; Aghaji, M. Z.; Woo, T. K. Rapid and Accurate Machine Learning Recognition of High Performing Metal Organic Frameworks for CO<sub>2</sub> Capture. *J. Phys. Chem. Lett.* **2014**, *5*, 3056–3060.
- [234] Bleiziffer, P.; Schaller, K.; Riniker, S. Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* **2018**, *58*, 579–590.
- [235] Duvenaud, D. K.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* 28. 2015; pp 2224–2232.
- [236] Pilia, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **2017**, *129*, 156–163.
- [237] Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **2018**, *9*, 3405.
- [238] Calle-Vallejo, F.; Martínez, J. I.; García-Lastra, J. M.; Sautet, P.; Loffreda, D. Fast Prediction of Adsorption Properties for Platinum Nanocatalysts with Generalized Coordination Numbers. *Angew. Chem. Int. Ed.* **2014**, *53*, 8316–8319.
- [239] Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
- [240] Wiener, H. Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons. *J. Am. Chem. Soc.* **1947**, *69*, 2636–2638.

- [241] Kier, L. B.; Hall, L. H.; Murray, W. J.; Randić, M. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971–1974.
- [242] Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular Connectivity V: Connectivity Series Concept Applied to Density. *J. Pharm. Sci.* **1976**, *65*, 1226–1230.
- [243] Kier, L. B.; Hall, L. H. Molecular Connectivity VII: Specific Treatment of Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806–1809.
- [244] Broto, P.; Moreau, G.; Vandycke, C. Molecular structures: perception, auto-correlation descriptor and SAR studies. *Eur. J. Med. Chem.* **1984**, *19*, 71–78.
- [245] Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A. A. Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation. *arXiv e-prints* **2018**, *abs/1811.02633*.
- [246] Goh, G. B.; Hodas, N. O.; Siegel, C.; Vishnu, A. SMILES2Vec: An Interpretable General-Purpose Deep Neural Network for Predicting Chemical Properties. *arXiv e-prints* **2017**, *abs/1712.02034*.
- [247] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.
- [248] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.
- [249] Thomas, N.; Smidt, T.; Kearnes, S.; Yang, L.; Li, L.; Kohlhoff, K.; Riley, P. Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds. *arXiv e-prints* **2018**, arXiv:1802.08219.
- [250] Montavon, G.; Hansen, K.; Fazli, S.; Rupp, M.; Biegler, F.; Ziehe, A.; Tkatchenko, A.; von Lilienfeld, A.; Müller, K. Learning Invariant Representations of Molecules for Atomization Energy Prediction. *Advances in Neural Information Processing Systems* 25. 2012; pp 449–457.
- [251] Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Katja; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.

- [252] Yao, K.; Herr, J. E.; Toth, D.; Mckintyre, R.; Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **2018**, *9*, 2261–2269.
- [253] Unke, O. T.; Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *Journal of Chemical Theory and Computation* **2019**, *15*, 3678–3693.
- [254] Behler, J. Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
- [255] Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
- [256] Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027–1044.
- [257] Amat, L.; Carbó-Dorca, R.; Ponec, R. Simple Linear QSAR Models Based on Quantum Similarity Measures. *J. Med. Chem.* **1999**, *42*, 5169–5180.
- [258] Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360*, 186–190.
- [259] Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J. Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *J. Chem. Theory Comput.* **2019**, *15*, 2331–2345.
- [260] Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.
- [261] Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.
- [262] Lei, X.; Medford, A. J. Design and analysis of machine learning exchange-correlation functionals via rotationally invariant convolutional descriptors. *Phys. Rev. Mater.* **2019**, *3*, 063801.
- [263] Trefethen, L. N.; Bau, D. *Numerical Linear Algebra*; SIAM, 1997.
- [264] Egan, W. J.; Merz, K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867–3877.



- [265] Livingstone, D. J.; Salt, D. W. Judging the Significance of Multiple Linear Regression Models. *J. Med. Chem.* **2005**, *48*, 661–663.
- [266] Guo, J.-Y.; Minko, Y.; Santiago, C. B.; Sigman, M. S. Developing Comprehensive Computational Parameter Sets To Describe the Performance of Pyridine-Oxazoline and Related Ligands. *ACS Catal.* **2017**, *7*, 4144–4151.
- [267] Robinson, S. G.; Yan, Y.; Hendriks, K. H.; Sanford, M. S.; Sigman, M. S. Developing a Predictive Solubility Model for Monomeric and Oligomeric Cyclopropenium-Based Flow Battery Catholytes. *J. Am. Chem. Soc.* **2019**, *141*, 10171–10176.
- [268] Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018**, *9*, 2398–2412.
- [269] Kaneko, H. Discussion on Regression Methods Based on Ensemble Learning and Applicability Domains of Linear Submodels. *J. Chem. Inf. Model.* **2018**, *58*, 480–489.
- [270] Williams, C. K. I. Learning Kernel Classifiers. *J. Am. Stat. Assoc.* **2003**, *98*, 489–490.
- [271] Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel Methods in Machine Learning. *Ann. Stat.* **2008**, *36*, 1171–1220.
- [272] Schölkopf, B.; Herbrich, R.; Smola, A. J. A Generalized Representer Theorem. *Computational Learning Theory 14*. Berlin, Heidelberg, 2001; pp 416–426.
- [273] Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2001.
- [274] Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.
- [275] Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*.

- [276] Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci.* **2019**, *116*, 3401–3406.
- [277] Bogojeski, M.; Brockherde, F.; Vogt-Maranto, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Efficient prediction of 3D electron densities using machine learning. *arXiv e-prints* **2018**, arXiv:1811.06255.
- [278] Noh, J.; Back, S.; Kim, J.; Jung, Y. Active learning with non-ab initio input features toward efficient CO<sub>2</sub> reduction catalysts. *Chem. Sci.* **2018**, *9*, 5152–5159.
- [279] Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077.
- [280] Kamath, A.; Vargas-Hernández, R. A.; Krems, R. V.; Carrington, T.; Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.* **2018**, *148*, 241702.
- [281] Huang, B.; Anatole von Lilienfeld, O. The DNA of chemistry: Scalable quantum machine learning with amons. *arXiv e-prints* **2017**, arXiv:1707.04146.
- [282] Williams, C. K. I.; Rasmussen, C. E. *Gaussian processes for machine learning*; MIT press Cambridge, MA, 2006.
- [283] Szlachta, W. J.; Bartók, A. P.; Csányi, G. Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B* **2014**, *90*, 104108.
- [284] Nguyen, T. T.; Székely, E.; Imbalzano, G.; Behler, J.; Csányi, G.; Ceriotti, M.; Götz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, 241725.
- [285] Fujikake, S.; Deringer, V. L.; Lee, T. H.; Krynski, M.; Elliott, S. R.; Csányi, G. Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures. *J. Chem. Phys.* **2018**, *148*, 241714.

- [286] Denzel, A.; Kästner, J. Gaussian process regression for geometry optimization. *J. Chem. Phys.* **2018**, *148*, 094114.
- [287] Proppe, J.; Gugler, S.; Reiher, M. Gaussian Process-Based Refinement of Dispersion Corrections. *arXiv e-prints* **2019**, arXiv:1906.09342.
- [288] Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14*, 5238–5248.
- [289] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **2015**, *61*, 85–117.
- [290] Ivakhnenko, A. G.; Lapa, V. G. *Cybernetics and forecasting techniques*; Modern analytic and computational methods in science and mathematics; Elsevier, 1967.
- [291] Ivakhenko, A. G. The Group Method of Data of Handling ; A rival of the method of stochastic approximation. *Soviet Automatic Control* **1968**, *13*, 43–55.
- [292] Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408.
- [293] Kleene, S. C. Representation of Events in Nerve Nets and Finite Automata. Automata Studies. (AM-34). Princeton, 1956; pp 3–42.
- [294] Hebb, D. O. *The Organization of Behavior: A Neuropsychological Theory*; Taylor & Francis, 1949.
- [295] McCulloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* **1943**, *5*, 115–133.
- [296] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI* **2019**, *1*.
- [297] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25. 2012; pp 1097–1105.
- [298] Deng, J.; Dong, W.; Socher, R.; Li, L.-j.; Li, K.; Fei-fei, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. 2009.

- [299] Graves, A.; Schmidhuber, J. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Advances in Neural Information Processing Systems* 21. 2009; pp 545–552.
- [300] Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. A Novel Connectionist System for Unconstrained Handwriting Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 855–868.
- [301] LeCun, Y.; Bengio, Y. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* **1995**, *3361*, 1995.
- [302] Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.
- [303] Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press, 2016; <http://www.deeplearningbook.org>.
- [304] Hahnloser, R. H.; Sarpeshkar, R.; Mahowald, M. A.; Douglas, R. J.; Sejung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature* **2000**, *405*, 947–51.
- [305] Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Fort Lauderdale, FL, USA, 2011; pp 315–323.
- [306] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. 2016; pp 770–778.
- [307] Iovanac, N. C.; Savoie, B. M. Improved Chemical Prediction from Scarce Data Sets via Latent Space Enrichment. *J. Phys. Chem. A* **2019**, *123*, 4295–4302.
- [308] Janet, J. P.; Kulik, H. J. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.* **2017**, *8*, 5137–5152.
- [309] St. John, P. C.; Phillips, C.; Kemper, T. W.; Wilson, A. N.; Guan, Y.; Crowley, M. F.; Nimlos, M. R.; Larsen, R. E. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **2019**, *150*, 234111.
- [310] Zubatyuk, R.; Smith, J. S.; Leszczynski, J.; Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Sci. Adv.* **2019**, *5*.

- [311] Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.
- [312] Glorot, X.; Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS. 2010; pp 249–256.
- [313] Montavon, G., Orr, G. B., Müller, K., Eds. *Neural Networks: Tricks of the Trade - Second Edition*; Lecture Notes in Computer Science; Springer, 2012; Vol. 7700.
- [314] Auer, P.; Herbster, M.; Warmuth, M. K. Exponentially many local minima for single neurons. Advances in Neural Information Processing Systems 8. 1996; pp 316–322.
- [315] Dauphin, Y.; Pascanu, R.; Gulcehre, C.; Cho, K.; Ganguli, S.; Bengio, Y. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *arXiv e-prints* **2014**, arXiv:1406.2572.
- [316] Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 3<sup>rd</sup> International Conference on Learning Representations. 2015.
- [317] Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. *arXiv e-prints* **2012**, *abs/1212.5701*.
- [318] Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **1980**, *36*, 193–202.
- [319] LeCun, Y.; Boser, B. E.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W. E.; Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551.
- [320] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
- [321] Maturana, D.; Scherer, S. VoxNet: A 3D Convolutional Neural Network for real-time object recognition. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2015; pp 922–928.
- [322] Kiranyaz, S.; Ince, T.; Gabbouj, M. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 664–675.

- [323] Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.
- [324] Jørgensen, M. S.; Mortensen, H. L.; Meldgaard, S. A.; Kolsbjerg, E. L.; Jacobsen, T. L.; Sørensen, K. H.; Hammer, B. Atomistic structure learning. *J. Chem. Phys.* **2019**, *151*, 054111.
- [325] Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-developed QSAR/QSPR Models. *arXiv e-prints* **2017**, *abs/1706.06689*.
- [326] Staker, J.; Marshall, K.; Abel, R.; McQuaw, C. M. Molecular Structure Extraction from Documents Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1017–1029.
- [327] Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. Proceedings of the 34<sup>th</sup> International Conference on Machine Learning. 2017; pp 1263–1272.
- [328] Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.
- [329] Matlock, M. K.; Dang, N. L.; Swamidass, S. J. Learning a Local-Variable Model of Aromatic and Conjugated Systems. *ACS Cent. Sci.* **2018**, *4*, 52–62.
- [330] Jørgensen, P. B.; Jacobsen, K. W.; Schmidt, M. N. Neural Message Passing with Edge Updates for Predicting Properties of Molecules and Materials. *arXiv e-prints* **2018**, *abs/1806.03146*.
- [331] Dai, H.; Dai, B.; Song, L. Discriminative Embeddings of Latent Variable Models for Structured Data. Proceedings of the 33<sup>rd</sup> International Conference on Machine Learning. 2016; pp 2702–2711.
- [332] Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

- [333] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.
- [334] Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. 2<sup>nd</sup> International Conference on Learning Representations. 2014.
- [335] Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J. Comput. Chem.* **2018**, *39*, 2079–2102.
- [336] Wang, W.; Gómez-Bombarelli, R. Coarse-Graining Auto-Encoders for Molecular Dynamics. *arXiv e-prints* **2018**, arXiv:1812.02706.
- [337] Doersch, C. Tutorial on Variational Autoencoders. *arXiv e-prints* **2016**, *abs/1606.05908*.
- [338] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.
- [339] Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; Bengio, Y. Generative Adversarial Nets. *Advances in Neural Information Processing Systems* 27. 2014; pp 2672–2680.
- [340] Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. 4<sup>th</sup> International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. 2016.
- [341] Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv e-prints* **2017**, *abs/1701.07875*.
- [342] Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharmaceutics* **2017**, *14*, 3098–3104.
- [343] Simonovsky, M.; Komodakis, N. GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. *Artificial Neural Networks and Machine Learning - ICANN* 27. 2018; pp 412–422.

- [344] Jin, W.; Barzilay, R.; Jaakkola, T. S. Junction Tree Variational Autoencoder for Molecular Graph Generation. Proceedings of the 35<sup>th</sup> International Conference on Machine Learning. 2018; pp 2328–2337.
- [345] Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59*, 43–52.
- [346] Guimaraes, G. L.; Sanchez-Lengeling, B.; Farias, P. L. C.; Aspuru-Guzik, A. Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models. *arXiv e-prints* **2017**, *abs/1705.10843*.
- [347] Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for De-Novo Drug Design. *arXiv e-prints* **2017**, *abs/1711.10907*.
- [348] Maziarka, L.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Warchol, M. Mol-CycleGAN – a generative model for molecular optimization. *arXiv e-prints* **2019**, *abs/1902.02119*.
- [349] Sutton, R. S.; McAllester, D. A.; Singh, S. P.; Mansour, Y. Policy Gradient Methods for Reinforcement Learning with Function Approximation. Advances in Neural Information Processing Systems 12. 1999; pp 1057–1063.
- [350] Yu, L.; Zhang, W.; Wang, J.; Yu, Y. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. *arXiv e-prints* **2016**, *abs/1609.05473*.
- [351] Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.
- [352] Cao, N. D.; Kipf, T. MolGAN: An implicit generative model for small molecular graphs. *arXiv e-prints* **2018**, *abs/1805.11973*.
- [353] Popova, M.; Shvets, M.; Oliva, J.; Isayev, O. MolecularRNN: Generating realistic molecular graphs with optimized properties. *arXiv e-prints* **2019**, *abs/1905.13372*.
- [354] You, J.; Liu, B.; Ying, Z.; Pande, V. S.; Leskovec, J. Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation. Advances in Neural Information Processing Systems 31. 2018; pp 6412–6422.
- [355] Kusner, M. J.; Paige, B.; Hernández-Lobato, J. M. Grammar Variational Autoencoder. Proceedings of the 34<sup>th</sup> International Conference on Machine Learning. 2017; pp 1945–1954.



- [356] Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. 6<sup>th</sup> International Conference on Learning Representations. 2018.
- [357] Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. *arXiv e-prints* **2019**, *abs/1905.13741*.
- [358] Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Nikolenko, S. I.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv e-prints* **2018**, *abs/1811.12823*.
- [359] Jensen, K. F.; Coley, C. W.; Eyke, N. S. Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem. Int. Ed.* **2019**,
- [360] Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649–663.
- [361] Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous discovery in the chemical sciences part II: Outlook. *Angew. Chem. Int. Ed.* **2019**, *0*.
- [362] Harper, P. M.; Gani, R. A multi-step and multi-level approach for computer aided molecular design. *Comput. Chem. Eng.* **2000**, *24*, 677–683.
- [363] Krishnan, S.; Sharma, V.; Singh, P.; Ramprasad, R. Dopants in lanthanum manganite: Insights from first-principles chemical space exploration. *J. Phys. Chem. C* **2016**, *120*, 22126–22133.
- [364] Pearlman, D. A.; Murcko, M. A. CONCEPTS: New dynamic algorithm for de novo drug suggestion. *J. Comput. Chem.* **1993**, *14*, 1184–1193.
- [365] Ishchenko, A. V.; Shakhnovich, E. I. SMall Molecule Growth 2001 (SMoG2001): An Improved Knowledge-Based Scoring Function for Protein–Ligand Interactions. *J. Med. Chem.* **2002**, *45*, 2770–2780.
- [366] Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825–24836.

- [367] Blaschke, T.; Olivecrona, M.; Engkvist, O.; Bajorath, J.; Chen, H. Application of Generative Autoencoder in De Novo Molecular Design. *Mol. Inf.* **2018**, *37*, 1700123.
- [368] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.
- [369] Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194–1204.
- [370] Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Panté, N.; Hancock, R. E. W.; Cherkasov, A. Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. *J. Med. Chem.* **2009**, *52*, 2006–2015.
- [371] Peherstorfer, B.; Willcox, K.; Gunzburger, M. Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization. *SIAM Rev.* **2018**, *60*, 550–591.
- [372] Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Global Optim.* **1998**, *13*, 455–492.
- [373] Forrester, A. I. J.; Sóbester, A.; Keane, A. J. Multi-fidelity optimization via surrogate modelling. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **2007**, *463*, 3251–3269.
- [374] Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458–465.
- [375] Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.
- [376] Chu, Y.; Heyndrickx, W.; Occhipinti, G.; Jensen, V. R.; Alsberg, B. K. An Evolutionary Algorithm for de Novo Optimization of Functional Transition Metal Compounds. *J. Am. Chem. Soc.* **2012**, *134*, 8885–8895.
- [377] Jensen, J. H. A graph-based genetic algorithm and generative model/Monte Carlo tree search for the exploration of chemical space. *Chem. Sci.* **2019**, *10*, 3567–3572.

- [378] Froemming, N. S.; Henkelman, G. Optimizing core-shell nanoparticle catalysts with a genetic algorithm. *J. Chem. Phys.* **2009**, *131*, 234103.
- [379] Ong, Y. S.; Nair, P. B.; Keane, A. J. Evolutionary Optimization of Computationally Expensive Problems via Surrogate Modeling. *AIAA Journal* **2003**, *41*, 687–696.
- [380] Glen, R. C.; Payne, A. W. R. A genetic algorithm for the automated generation of molecules within constraints. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 181–202.
- [381] Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T. Genetic algorithms for computational materials discovery accelerated by machine learning. *npj Comput. Mater.* **2019**, *5*, 46.
- [382] Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multi-objective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **2002**, *6*, 182–197.
- [383] Bradford, E.; Schweidtmann, A. M.; Lapkin, A. Efficient multiobjective optimization employing Gaussian processes, spectral sampling and a genetic algorithm. *J. Global Optim.* **2018**, *71*, 407–438.
- [384] Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63.
- [385] Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148–175.
- [386] Forrester, A. I. J.; Keane, A. J. Recent advances in surrogate-based optimization. *Prog. Aerosp. Sci.* **2009**, *45*, 50–79.
- [387] Carr, S.; Garnett, R.; Lo, C. BASC: Applying Bayesian Optimization to the Search for Global Minima on Potential Energy Surfaces. Proceedings of The 33<sup>rd</sup> International Conference on Machine Learning. New York, New York, USA, 2016; pp 898–907.
- [388] Herbol, H. C.; Hu, W.; Frazier, P.; Clancy, P.; Poloczek, M. Efficient search of compositional space for hybrid organic-inorganic perovskites via Bayesian optimization. *npj Comput. Mater.* **2018**, *4*, 51.

- [389] Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of compounds for machine-learning prediction of physical properties. *Phys. Rev. B* **2017**, *95*, 144110.
- [390] Snoek, J.; Rippel, O.; Adams, R. P. Scalable Bayesian Optimization Using Deep Neural Networks. *Int. Conf. Mach. Learn.* **2015**, 0–3.
- [391] Ginsbourger, D.; Le Riche, R.; Carraro, L. In *Computational Intelligence in Expensive Optimization Problems*; Tenne, Y., Goh, C.-K., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2010; pp 131–162.
- [392] Frazier, P.; Powell, W.; Dayanik, S. The Knowledge-Gradient Policy for Correlated Normal Beliefs. *INFORMS J. Comput.* **2009**, *21*, 599–613.
- [393] Scott, W.; Frazier, P. I.; Powel, W. B. The Correlated Knowledge Gradient for Simulation Optimization of Continuous Parameters using Gaussian Process Regression. *SIAM J. Optim.* **2011**, *21*, 996–1026.
- [394] Keane, A. J. Statistical Improvement Criteria for Use in Multiobjective Design Optimization. *AIAA Journal* **2006**, *44*, 879–891.
- [395] Thompson, W. R. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* **1933**, *25*, 285–294.
- [396] Hernández-Lobato, J. M.; Requeima, J.; Pyzer-Knapp, E. O.; Aspuru-Guzik, A. Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space. *arXiv e-prints* **2017**,
- [397] Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenix: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134–1145.
- [398] Qin, J.; Khaira, G. S.; Su, Y.; Garner, G. P.; Miskin, M.; Jaeger, H. M.; de Pablo, J. J. Evolutionary pattern design for copolymer directed self-assembly. *Soft Matter* **2013**, *9*, 11467–11472.
- [399] Khaira, G. S.; Qin, J.; Garner, G. P.; Xiong, S.; Wan, L.; Ruiz, R.; Jaeger, H. M.; Nealey, P. F.; de Pablo, J. J. Evolutionary Optimization of Directed Self-Assembly of Triblock Copolymers on Chemically Patterned Substrates. *ACS Macro Lett.* **2014**, *3*, 747–752.

- [400] Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives. *Chem. Eng. J.* **2018**, *352*, 277–282.
- [401] Kuhn, C.; Beratan, D. N. Inverse Strategies for Molecular Design. *J. Phys. Chem.* **1996**, *100*, 10595–10599.
- [402] Weymuth, T.; Reiher, M. Gradient-driven molecule construction: An inverse approach applied to the design of small-molecule fixating catalysts. *Int. J. Quantum Chem.* **2014**, *114*, 838–850.
- [403] Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228–3232.
- [404] Hu, X.; Beratan, D. N.; Yang, W. A gradient-directed Monte Carlo approach to molecular design. *J. Chem. Phys.* **2008**, *129*, 064102.
- [405] Rinderspacher, B. C.; Andzelm, J.; Rawlett, A.; Dougherty, J.; Beratan, D. N.; Yang, W. Discrete Optimization of Electronic Hyperpolarizabilities in a Chemical Subspace. *J. Chem. Theory Comput.* **2009**, *5*, 3321–3329.
- [406] Pyzer-Knapp, E. O.; Li, A., Kewei and Aspuru-Guzik Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Adv. Funct. Mater.* **2015**, *25*, 6495–6502.
- [407] Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **2014**, *89*, 094104.
- [408] Li, L.; Snyder, J. C.; Pelaschier, I. M.; Huang, J.; Niranjana, U.-N.; Duncan, P.; Rupp, M.; Müller, K.-R.; Burke, K. Understanding machine-learned density functionals. *Int. J. Quantum Chem.* **2016**, *116*, 819–833.
- [409] Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058–1073.
- [410] Behler, J. Representing Potential Energy Surfaces by High-Dimensional Neural Network Potentials. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.

- [411] Lorenz, S.; Groß, A.; Scheffler, M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. *Chem. Phys. Lett.* **2004**, *395*, 210–215.
- [412] Prudente, F. V.; Neto, J. J. S. The fitting of potential energy surfaces using neural networks. Application to the study of the photodissociation processes. *Chem. Phys. Lett.* **1998**, *287*, 585–589.
- [413] Mones, L.; Bernstein, N.; Csanyi, G. Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression. *J. Chem. Theory Comput.* **2016**, *12*, 5100–5110.
- [414] Mills, K.; Spanner, M.; Tamblyn, I. Deep learning and the Schrödinger equation. *Phys. Rev. A* **2017**, *96*, 042113.
- [415] Snyder, J. C.; Rupp, M.; Hansen, K.; Blooston, L.; Müller, K.-R.; Burke, K. Orbital-free bond breaking via machine learning. *J. Chem. Phys.* **2013**, *139*, 224104.
- [416] Yao, K.; Herr, J. E.; Parkhill, J. The many-body expansion combined with neural networks. *J. Chem. Phys.* **2017**, *146*, 014106.
- [417] Hase, F.; Valleau, S.; Pyzer-Knapp, E.; Aspuru-Guzik, A. Machine learning exciton dynamics. *Chem. Sci.* **2016**, *7*, 5139–5147.
- [418] Li, Z.; Kermode, J. R.; De Vita, A. Molecular Dynamics with On-the-Fly Machine Learning of Quantum-Mechanical Forces. *Phys. Rev. Lett.* **2015**, *114*, 096405.
- [419] Botu, V.; Ramprasad, R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. *Int. J. Quantum Chem.* **2015**, *115*, 1074–1083.
- [420] Pilia, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine learning bandgaps of double perovskites. *Sci. Rep.* **2016**, *6*, 19375.
- [421] Mannodi-Kanakkithodi, A.; Pilia, G.; Huan, T. D.; Lookman, T.; Ramprasad, R. Machine Learning Strategy for Accelerated Design of Polymer Dielectrics. *Sci. Rep.* **2016**, *6*, 20952.
- [422] Huan, T. D.; Mannodi-Kanakkithodi, A.; Ramprasad, R. Accelerated materials property predictions and design using motif-based fingerprints. *Phys. Rev. B* **2015**, *92*, 014106.

- [423] Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **2013**, *3*, 2810.
- [424] Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **2016**, *93*, 115104.
- [425] Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci.* **2016**, 201602375.
- [426] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- [427] De, S.; Bartk, A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids Across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 1–18.
- [428] Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2013**, *57*, 3186–3204.
- [429] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- [430] Kubinyi, H. QSAR and 3D QSAR in drug design. Part 1: Methodology. *Drug Discovery Today* **1997**, *2*, 457–467.
- [431] Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O’neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. Additivity rules for the estimation of thermochemical properties. *Chem. Rev.* **1969**, *69*, 279–324.
- [432] Deeth, R. J. The ligand field molecular mechanics model and the stereoelectronic effects of d and s electrons. *Coord. Chem. Rev.* **2001**, *212*, 11–34.
- [433] Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **2014**, *89*, 205118.
- [434] Ashley, D. C.; Jakubikova, E. Ironing out the Photochemical and Spin-Crossover Behavior of Fe (II) Coordination Compounds with Computational Chemistry. *Coord. Chem. Rev.* **2017**, *337*, 97–111.

- [435] Bowman, D. N.; Jakubikova, E. Low-Spin versus High-Spin Ground State in Pseudo-Octahedral Iron Complexes. *Inorg. Chem.* **2012**, *51*, 6011–6019.
- [436] Gani, T. Z. H.; Kulik, H. J. Where Does the Density Localize? Convergent Behavior for Global Hybrids, Range Separation, and DFT+U. *J. Chem. Theory Comput.* **2016**, *12*, 5931–5945.
- [437] Ioannidis, E. I.; Kulik, H. J. Ligand-Field-Dependent Behavior of meta-GGA Exchange in Transition-Metal Complex Spin-State Ordering. *J. Phys. Chem. A* **2017**, *121*, 874–884.
- [438] Huang, W.; Xing, D.-H.; Lu, J.-B.; Long, B.; Schwarz, W. H. E.; Li, J. How Much Can Density Functional Approximations (DFA) Fail? The Extreme Case of the FeO<sub>4</sub> Species. *J. Chem. Theory Comput.* **2016**, *12*, 1525–1533.
- [439] Stewart, J. J. P. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Model.* **2013**, *19*, 1–32.
- [440] Shen, L.; Wu, J.; Yang, W. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 4934–4946.
- [441] Kulik, H. J. Perspective: Treating electron over-delocalization with the DFT+U method. *J. Chem. Phys.* **2015**, *142*, 240901.
- [442] Sutton, J. E.; Guo, W.; Katsoulakis, M. A.; Vlachos, D. G. Effects of correlated parameters and uncertainty in electronic-structure-based chemical kinetic modelling. *Nat. Chem.* **2016**, *8*, 331–337.
- [443] Simm, G. N.; Reiher, M. Systematic Error Estimation for Chemical Reaction Energies. *J. Chem. Theory Comput.* **2016**,
- [444] Walker, E.; Ammal, S. C.; Terejanu, G. A.; Heyden, A. Uncertainty Quantification Framework Applied to the Water–Gas Shift Reaction over Pt-Based Catalysts. *J. Phys. Chem. C* **2016**, *120*, 10328–10339.
- [445] Létard, J.-F.; Guionneau, P.; Goux-Capes, L. *Spin Crossover in Transition Metal Compounds III*; Springer, 2004; pp 221–249.
- [446] Kramida, A.; Ralchenko, Y.; Reader, J.; NIST ADS Team, NIST Atomic Spectra Database (version 5.7.1). 2019; <https://doi.org/10.18434/T4W30F>, [accessed March 14, 2017].



- [447] PetaChem, L. L. C. TeraChem v1.9. 2015; <http://www.petachem.com>.
- [448] Saunders, V. R.; Hillier, I. H. A “Level–Shifting” Method for Converging Closed Shell Hartree–Fock Wave Functions. *Int. J. Quantum Chem.* **1973**, *7*, 699–705.
- [449] Kaštner, J.; Carr, J. M.; Keal, T. W.; Thiel, W.; Wander, A.; Sherwood, P. DL-FIND: An Open-Source Geometry Optimizer for Atomistic Simulations. *J. Phys. Chem. A* **2009**, *113*, 11856–11865.
- [450] Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Model.* **1996**, *36*, 128–136.
- [451] Gastegger, M.; Marquetand, P. High-Dimensional Neural Network Potentials for Organic Reactions and an Improved Training Algorithm. *J. Chem. Theory Comput.* **2015**, *11*, 2187–2198.
- [452] Hageman, J. A.; Westerhuis, J. A.; Frhauf, H. W.; Rothenberg, G. Design and Assembly of Virtual Homogeneous Catalyst Libraries - Towards in Silico Catalyst Optimisation. *Adv. Synth. Catal.* **2006**, *348*, 361–369.
- [453] Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109–116.
- [454] Gastegger, M.; Kauffmann, C.; Behler, J.; Marquetand, P. Comparing the Accuracy of High-Dimensional Neural Network Potentials and the Systematic Molecular Fragmentation Method: A Benchmark Study for All-Trans Alkanes. *J. Chem. Phys.* **2016**, *144*, 194110.
- [455] Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22.
- [456] R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2014.
- [457] Larochelle, H.; Bengio, Y.; Louradour, J.; Lamblin, P. Exploring Strategies for Training Deep Neural Networks. *J. Mach. Learn. Res.* **2009**, *10*, 1–40.
- [458] Aiello, S.; Kraljevic, T.; Maj, P. *H2O: R Interface for H2O.*; Report, 2015.
- [459] Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33<sup>nd</sup> International Conference on Machine Learning. 2016; pp 1050–1059.

- [460] Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
- [461] Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *arXiv e-prints* **2012**, 1–18.
- [462] Bengio, Y. *Neural Networks: Tricks of the Trade*; Springer, 2012; pp 437–478.
- [463] Candel, A.; Parmar, V.; LeDell, E.; Arora, A. Deep learning with H2O. *H2O.ai* **2016**,
- [464] Recht, B.; Ré, C.; Wright, S. J.; Niu, F. Hogwild: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. *Advances in Neural Information Processing Systems* 24. 2011; pp 693–701.
- [465] Kingston, G. B.; Lambert, M. F.; Maier, H. R. Bayesian Training of Artificial Neural Networks Used for Water Resources Modeling. *Water Resour. Res.* **2005**, *41*, 1–11.
- [466] Secchi, P.; Zio, E. Quantifying uncertainties in the estimation of safety parameters by using bootstrapped artificial neural networks. *Ann. Nucl. Energy* **2008**, *35*, 2338–2350.
- [467] Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab – An S4 Package for Kernel Methods in R. *J. Stat. Softw.* **2004**, *11*, 1–20.
- [468] Krueger, T.; Panknin, D.; Braun, M. Fast cross-validation via sequential testing. *J. Mach. Learn. Res.* **2015**, *16*, 1103–1155.
- [469] Therneau, T.; Atkinson, B.; Ripley, B.; Ripley, M. B. Rpart: Recursive Partitioning and Regression Trees. 2015; <https://cran.r-project.org/package=rpart>, [accessed March, 14 2017].
- [470] Hughes, T. F.; Harvey, J. N.; Friesner, R. A. A B3LYP-DBLOC Empirical Correction Scheme for Ligand Removal Enthalpies of Transition Metal Complexes: Parameterization Against Experimental and CCSD(T)-F12 Heats of Formation. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7724–38.
- [471] Bajusz, D.; Racz, A.; Heberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* **2015**, *7*, 20.

- [472] Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I. B.; Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **2006**, *5*, 909–913.
- [473] Jensen, P. B.; Bialy, A.; Blanchard, D.; Lysgaard, S.; Reumert, A. K.; Quaade, U. J.; Vegge, T. Accelerated DFT-based design of materials for ammonia storage. *Chem. Mater.* **2015**, *27*, 4552–4561.
- [474] Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding nature’s missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **2010**, *22*, 3762–3767.
- [475] Jain, A.; Hautier, G.; Moore, C. J.; Ong, S. P.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295–2310.
- [476] Hautier, G.; Miglio, A.; Ceder, G.; Rignanese, G.-M.; Gonze, X. Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat. Commun.* **2013**, *4*, 2292.
- [477] Bowman, D. N.; Bondarev, A.; Mukherjee, S.; Jakubikova, E. Tuning the Electronic Structure of Fe(II) Polypyridines via Donor Atom and Ligand Scaffold Modifications: A Computational Study. *Inorg. Chem.* **2015**, *54*, 8786–8793.
- [478] Eckert, H.; Bojorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
- [479] Su, X.; Kulik, H. J.; Jamison, T. F.; Hatton, T. A. Anion-Selective Redox Electrodes: Electrochemically Mediated Separation with Heterogeneous Organometallic Interfaces. *Adv. Funct. Mater.* **2016**, *26*, 3394–3404.
- [480] Li, H.; Leung, K.-S.; Wong, M.-H.; Ballester, P. J. Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets. *Mol. Inf.* **2015**, *34*, 115–126.
- [481] Li, Z.; Ma, X.; Xin, H. Feature engineering of machine-learning chemisorption models for catalyst design. *Catal. Today* **2017**, *280*, 232–238.
- [482] Janet, J. P.; Gani, T. Z. H.; Steeves, A. H.; Ioannidis, E. I.; Kulik, H. J. Leveraging Cheminformatics Strategies for Inorganic Discovery: Application to Redox Potential Design. *Ind. Eng. Chem. Res.* **2017**, *56*, 4898–4910.

- [483] Devillers, J.; Domine, D.; Guillon, C.; Bintein, S.; Karcher, W. Prediction of partition coefficients ( $\log P_{\text{oct}}$ ) using autocorrelation descriptors. *SAR QSAR Environ. Res.* **1997**, *7*, 151–172.
- [484] Broto, P.; Devillers, J. *Autocorrelation of properties distributed on molecular graphs*; Kluwer Academic Publishers: Dordrecht, The Netherlands, 1990.
- [485] Puzyn, T.; Leszczynski, J.; Cronin, M. T. *Recent advances in QSAR studies: methods and applications*; Springer Science & Business Media, 2010; Vol. 8.
- [486] Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant size descriptors for accurate machine learning models of molecular properties. *J. Chem. Phys.* **2018**, *148*, 241718.
- [487] Yao, K.; Herr, J. E.; Brown, S. N.; Parkhill, J. Intrinsic Bond Energies from a Bonds-in-Molecules Neural Network. *J. Phys. Chem. Lett.* **2017**, *8*, 2689–2694.
- [488] Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. *arXiv e-prints* **2017**,
- [489] Gani, T. Z. H.; Ioannidis, E. I.; Kulik, H. J. Computational Discovery of Hydrogen Bond Design Rules for Electrochemical Ion Separation. *Chem. Mater.* **2016**, *28*, 6207–6218.
- [490] Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
- [491] Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **2007**, *23*, 2507–2517.
- [492] Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. Royal Stat. Soc.: Series B (Methodological)* **1996**, *58*, 267–288.
- [493] Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. Royal Stat. Soc.: Series B (Statistical Methodology)* **2005**, *67*, 301–320.
- [494] Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
- [495] Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236.
- [496] Kuhn, M. Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* **2008**, *28*, 1–26.

- [497] Liaw, A.; Wiener, M. Classification and Regression by randomForest. 2002; <https://CRAN.R-project.org/doc/Rnews/>.
- [498] Wang, L.-P.; Song, C. Geometry optimization made simple with translation and rotation coordinates. *J. Chem. Phys.* **2016**, *144*, 214108.
- [499] Liu, F.; Luehr, N.; Kulik, H. J.; Martínez, T. J. Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *J. Chem. Theory Comput.* **2015**, *11*, 3131–3144.
- [500] Bondi, A. van der Waals volumes and radii. *J. Chem. Phys.* **1964**, *68*, 441–451.
- [501] Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; John Wiley & Sons, Inc.: New Jersey, 2007; Vol. 23; pp 1–82.
- [502] Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials. *Theor. Chem. Acc.* **1997**, *97*, 119–124.
- [503] Eichkorn, K.; Treutler, O.; Öhm, H.; Häser, M.; Ahlrichs, R. Auxiliary basis sets to approximate Coulomb potentials. *Chem. Phys. Lett.* **1995**, *240*, 283–290.
- [504] Kim, J. Y.; Steeves, A. H.; Kulik, H. J. Harnessing Organic Ligand Libraries for First-Principles Inorganic Discovery: Indium Phosphide Quantum Dot Precursor Design Strategies. *Chem. Mater.* **2017**, *29*, 3632–3643.
- [505] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191–201.
- [506] Huang, S.-D.; Shang, C.; Zhang, X.-J.; Liu, Z.-P. Material discovery by combining stochastic surface walking global optimization with a neural network. *Chem. Sci.* **2017**, *8*, 6327–6337.
- [507] Tortorella, S.; Marotta, G.; Cruciani, G.; De Angelis, F. Quantitative structure-property relationship modeling of ruthenium sensitizers for solar cells applications: novel tools for designing promising candidates. *RSC Adv.* **2015**, *5*, 23865–23873.

- [508] Cruz, V. L.; Martinez, S.; Ramos, J.; Martinez-Salazar, J. 3D-QSAR as a tool for understanding and improving single-site polymerization catalysts. a review. *Organometallics* **2014**, *33*, 2944–2959.
- [509] Fey, N.; Orpen, A. G.; Harvey, J. N. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus (III)-donor ligands and the metal–phosphorus bond. *Coord. Chem. Rev.* **2009**, *253*, 704–722.
- [510] Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative structure–property relationship modeling of diverse materials properties. *Chem. Rev.* **2012**, *112*, 2889–2919.
- [511] Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.
- [512] Kim, C.; Pilania, G.; Ramprasad, R. From organized high-throughput data to phenomenological theory using machine learning: the example of dielectric breakdown. *Chem. Mater.* **2016**, *28*, 1304–1311.
- [513] Keinan, S.; Hu, X.; Beratan, D. N.; Yang, W. Designing molecules with optimal properties using the linear combination of atomic potentials approach in an AM1 semiempirical framework. *J. Phys. Chem. A* **2007**, *111*, 176–181.
- [514] Leardi, R. Genetic algorithms in chemistry. *J. Chromatogr. A* **2007**, *1158*, 226–233.
- [515] Leardi, R. Genetic algorithms in chemometrics and chemistry: a review. *J. Chemom.* **2001**, *15*, 559–569.
- [516] Venkatraman, V.; Abburu, S.; Alsberg, B. K. Artificial evolution of coumarin dyes for dye sensitized solar cells. *Phys. Chem. Chem. Phys.* **2015**, *17*, 27672–27682.
- [517] Decurtins, S.; Gütlich, P.; Köhler, C. P.; Spiering, H.; Hauser, A. Light-induced excited spin state trapping in a transition-metal complex: The hexa-1-propyltetrazole-iron (II) tetrafluoroborate spin-crossover system. *Chem. Phys. Lett.* **1984**, *105*, 1–4.
- [518] Hauser, A. *Spin Crossover in Transition Metal Compounds II*; Springer, 2004; pp 155–198.

- [519] Reed, D. A.; Xiao, D. J.; Gonzalez, M. I.; Darago, L. E.; Herm, Z. R.; Grandjean, F.; Long, J. R. Reversible CO Scavenging via Adsorbate-Dependent Spin State Transitions in an Iron (II)–Triazolate Metal–Organic Framework. *J. Am. Chem. Soc.* **2016**, *138*, 5594–5602.
- [520] Groizard, T.; Papior, N.; Le Guennic, B.; Robert, V.; Kepenekian, M. Enhanced Cooperativity in Supported Spin-Crossover Metal–Organic Frameworks. *J. Phys. Chem. Lett.* **2017**, *8*, 3415–3420.
- [521] Neville, S. M.; Halder, G. J.; Chapman, K. W.; Duriska, M. B.; Moubaraki, B.; Murray, K. S.; Kepert, C. J. Guest tunable structure and spin crossover properties in a nanoporous coordination framework material. *J. Am. Chem. Soc.* **2009**, *131*, 12106–12108.
- [522] Gütlich, P. Spin crossover in iron(II)-complexes. Metal Complexes. Berlin, Heidelberg, 1981; pp 83–195.
- [523] Gütlich, P.; Hauser, A. Thermal and light-induced spin crossover in iron (II) complexes. *Coord. Chem. Rev.* **1990**, *97*, 1–22.
- [524] Bousseksou, A.; Molnár, G.; Salmon, L.; Nicolazzi, W. Molecular spin crossover phenomenon: recent achievements and prospects. *Chem. Soc. Rev.* **2011**, *40*, 3313–3335.
- [525] Phan, H.; Hrudka, J. J.; Igimbayeva, D.; Lawson Daku, L. M.; Shatruk, M. A Simple Approach for Predicting the Spin State of Homoleptic Fe (II) Tris-diimine Complexes. *J. Am. Chem. Soc.* **2017**, *139*, 6437–6447.
- [526] Rodríguez-Jiménez, S.; Yang, M.; Stewart, I.; Garden, A. L.; Brooker, S. A Simple Method of Predicting Spin State in Solution. *J. Am. Chem. Soc.* **2017**, *139*, 18392–18396.
- [527] Fias, S.; Heidar-Zadeh, F.; Geerlings, P.; Ayers, P. W. Chemical transferability of functional groups follows from the nearsightedness of electronic matter. *Proc. Natl. Acad. Sci.* **2017**, *114*, 11633–11638.
- [528] Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
- [529] Wilbraham, L.; Verma, P.; Truhlar, D. G.; Gagliardi, L.; Ciofini, I. Multiconfiguration Pair-Density Functional Theory Predicts Spin-State Ordering in Iron

- Complexes with the Same Accuracy as Complete Active Space Second-Order Perturbation Theory at a Significantly Reduced Computational Cost. *J. Phys. Chem. Lett.* **2017**, *8*, 2026–2030.
- [530] Gani, T. Z. H.; Kulik, H. J. Unifying Exchange Sensitivity in Transition-Metal Spin-State Ordering and Catalysis through Bond Valence Metrics. *J. Chem. Theory Comput.* **2017**, *13*, 5443–5457.
- [531] Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 1668–1673.
- [532] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.
- [533] Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J. Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Ind. Eng. Chem. Res.* **2018**, *57*, 13973–13986.
- [534] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.
- [535] Martínez, T. J. Ab initio reactive computer aided molecular design. *Acc. Chem. Res.* **2017**, *50*, 652–656.
- [536] Caruthers, J. M.; Lauterbach, J. A.; Thomson, K. T.; Venkatasubramanian, V.; Snively, C. M.; Bhan, A.; Katare, S.; Oskarsdottir, G. Catalyst design: knowledge extraction from high-throughput experimentation. *J. Catal.* **2003**, *216*, 98–109.
- [537] Katare, S.; Caruthers, J. M.; Delgass, W. N.; Venkatasubramanian, V. An intelligent system for reaction kinetic modeling and catalyst design. *Ind. Eng. Chem. Res.* **2004**, *43*, 3484–3512.
- [538] Corma, A.; Díaz-Cabanas, M. J.; Moliner, M.; Martínez, C. Discovery of a new catalytically active and selective zeolite (ITQ-30) by high-throughput synthesis techniques. *J. Catal.* **2006**, *241*, 312–318.



- [539] Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.
- [540] Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **2017**, *3*, e1603015.
- [541] Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine learning for heterogeneous catalyst design and discovery. *AIChE J.* **2018**, *64*, 2311–2323.
- [542] Kitchin, J. R. Machine learning in catalysis. *Nat. Catal.* **2018**, *1*, 230–232.
- [543] Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery with Machine Learning: Simulated Evolution of Spin Crossover Complexes with an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.
- [544] Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T. Accelerated Discovery of Large Electrostrains in BaTiO<sub>3</sub>-Based Piezoelectrics Using Active Learning. *Adv. Mater.* **2018**, *30*, 1702884.
- [545] Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **2018**, *3*, 819–825.
- [546] He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *J. Phys. Chem. Lett.* **2018**, *9*, 4562–4569.
- [547] Kailkhura, B.; Gallagher, B.; Kim, S.; Hiszpanski, A.; Yong-Jin Han, T. Reliable and Explainable Machine Learning Methods for Accelerated Material Discovery. *arXiv e-prints* **2019**,
- [548] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547.

- [549] Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **2017**, *8*, 14621.
- [550] Musil, F.; Willatt, M. J.; Langovoy, M. A.; Ceriotti, M. Fast and Accurate Uncertainty Estimation in Chemical Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 906–915.
- [551] Peterson, A. A.; Christensen, R.; Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978–10985.
- [552] Liu, R.; Wallqvist, A. Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies Out-of-Domain Compounds. *J. Chem. Inf. Model.* **2019**, *59*, 181–189.
- [553] Cortés-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Prediction Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2018**,
- [554] Morais, C. L. M.; Lima, K. M. G.; Martin, F. L. Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines. *Anal. Chim. Acta* **2018**,
- [555] Huang, G.; Li, Y.; Pleiss, G.; Liu, Z.; Hopcroft, J. E.; Weinberger, K. Q. Snapshot Ensembles: Train 1, get M for free. *arXiv e-prints* **2017**, *abs/1704.00109*.
- [556] Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
- [557] Neal, R. M. *Bayesian learning for neural networks*; Springer Science & Business Media, 2012; Vol. 118.
- [558] Liu, R.; Glover, K. P.; Feasel, M. G.; Wallqvist, A. General Approach to Estimate Error Bars for Quantitative Structure–Activity Relationship Predictions of Molecular Activity. *J. Chem. Inf. Model.* **2018**, *58*, 1561–1575.
- [559] Li, X.; Li, F. Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics. *arXiv e-prints* **2017**, *abs/1612.07767*.
- [560] Metzen, J. H.; Genewein, T.; Fischer, V.; Bischoff, B. On Detecting Adversarial Perturbations. 5<sup>th</sup> International Conference on Learning Representations. 2017.

- [561] Gu, S.; Rigazio, L. Towards Deep Neural Network Architectures Robust to Adversarial Examples. *arXiv e-prints* **2014**,
- [562] Zhou, C.; Paffenroth, R. C. Anomaly Detection with Robust Deep Autoencoders. Proceedings of the 23<sup>rd</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017; pp 665–674.
- [563] Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; Langs, G. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. *Information Processing in Medical Imaging* 25. 2017; pp 146–157.
- [564] Jiang, H.; Kim, B.; Guan, M. Y.; Gupta, M. R. To Trust Or Not To Trust A Classifier. *Advances in Neural Information Processing Systems* 31. 2018; pp 5546–5557.
- [565] Papernot, N.; McDaniel, P. D. Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *arXiv e-prints* **2018**, *abs/1803.04765*.
- [566] Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; Srivastava, B. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. *Workshop on Artificial Intelligence Safety 2019 co-located with the Thirty-Third AAAI Conference on Artificial Intelligence 2019 (AAAI-19)*. 2019.
- [567] Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513–530.
- [568] Gomes, J.; Ramsundar, B.; Feinberg, E. N.; Pande, V. S. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv e-prints* **2017**, *abs/1703.10603*.
- [569] Xie, T.; Grossman, J. C. Hierarchical visualization of materials space with graph convolutional neural networks. *J. Chem. Phys.* **2018**, *149*, 174111.
- [570] Gubaev, K.; Podryabinkin, E. V.; Shapeev, A. V. Machine learning of molecular properties: Locality and active learning. *J. Chem. Phys.* **2018**, *148*, 241727.
- [571] Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123–140.

- [572] Aggarwal, C. C.; Hinneburg, A.; Keim, D. A. On the Surprising Behavior of Distance Metrics in High Dimensional Space. Database Theory – ICDT 2001. 2001; pp 420–434.
- [573] Ditchfield, R.; Hehre, W. J.; Pople, J. A. Self-Consistent Molecular Orbital Methods. 9. Extended Gaussian-type basis for molecular orbital studies of organic molecules. *J. Chem. Phys.* **1971**, *54*, 724.
- [574] McInnes, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints* **2018**, *abs/1802.03426*.
- [575] Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv e-prints* **2017**, *abs/1708.07747*.
- [576] Chollet, F., et al. Keras. <https://keras.io>, 2015.
- [577] Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015; <https://www.tensorflow.org/>, Software available from tensorflow.org.
- [578] Gugler, S.; Janet, J. P.; Kulik, H. J. Enumeration of de novo inorganic complexes for chemical discovery and machine learning. *Mol. Syst. Des. Eng.* **2019**, Advance Article.
- [579] Gossett, E.; Toher, C.; Oses, C.; Isayev, O.; Legrain, F.; Rose, F.; Zurek, E.; Carrete, J.; Mingo, N.; Tropsha, A.; Curtarolo, S. AFLOW-ML: A RESTful API for machine-learning predictions of materials properties. *Comput. Mater. Sci.* **2018**, *152*, 134–145.
- [580] Warmuth, M. K.; Liao, J.; Rättsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
- [581] Cohn, D.; Atlas, L.; Ladner, R. Improving generalization with active learning. *Machine Learning* **1994**, *15*, 201–221.
- [582] Jørgensen, P. B.; Mesta, M.; Shil, S.; García Lastra, J. M.; Jacobsen, K. W.; Thygesen, K. S.; Schmidt, M. N. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **2018**, *148*, 241735.

- [583] Kim, S.; Jinich, A.; Aspuru-Guzik, A. MultiDK: A Multiple Descriptor Multiple Kernel Approach for Molecular Discovery and Its Application to Organic Flow Battery Electrolytes. *J. Chem. Inf. Model.* **2017**, *57*, 657–668.
- [584] Toyoura, K.; Hirano, D.; Seko, A.; Shiga, M.; Kuwabara, A.; Karasuyama, M.; Shitara, K.; Takeuchi, I. Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides. *Phys. Rev. B* **2016**, *93*, 054112.
- [585] Andersson, M. P.; Bligaard, T.; Kustov, A.; Larsen, K. E.; Greeley, J.; Johannessen, T.; Christensen, C. H.; Nørskov, J. K. Toward computational screening in heterogeneous catalysis: Pareto-optimal methanation catalysts. *J. Catal.* **2006**, *239*, 501–506.
- [586] Miranda-Galindo, E. Y.; Segovia-Hernández, J. G.; Hernández, S.; Gutiérrez-Antonio, C.; Briones-Ramírez, A. Reactive Thermally Coupled Distillation Sequences: Pareto Front. *Ind. Eng. Chem. Res.* **2011**, *50*, 926–938.
- [587] Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: enabling hierarchy based multi-objective optimization for self-driving laboratories. *Chem. Sci.* **2018**, *9*, 7642–7655.
- [588] Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000–1005.
- [589] Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **2008**, *48*, 220–232.
- [590] Palmer, D. S.; O’Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.
- [591] Sarmah, P.; Deka, R. C. DFT-based QSAR and QSPR models of several cis-platinum complexes: solvent effect. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 343–354.
- [592] Chen, Y.-W. D.; Santhanam, K. S. V.; Bard, A. J. Solution Redox Couples for Electrochemical Energy Storage: I Iron (III)-Iron(II) Complexes with O-Phenanthroline and Related Ligands. *J. Electrochem. Soc.* **1981**, *128*, 1460–1467.

- [593] Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **2019**, *10*, 7913–7922.
- [594] Rasmussen, C. E. *Summer School on Machine Learning*; Springer, 2003; pp 63–71.
- [595] Tallorin, L.; Wang, J.; Kim, W. E.; Sahu, S.; Kosa, N. M.; Yang, P.; Thompson, M.; Gilson, M. K.; Frazier, P. I.; Burkart, M. D.; Gianneschi, N. C. Discovering de novo peptide substrates for enzymes using machine learning. *Nat. Commun.* **2018**, *9*, 5253.
- [596] Ngatchou, P.; Zarei, A.; El-Sharkawi, A. Pareto Multi Objective Optimization. Proceedings of the 13<sup>th</sup> International Conference on, Intelligent Systems Application to Power Systems. 2005; pp 84–91.
- [597] Wallach, I.; Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **2018**, *58*, 916–932.
- [598] Liu, F.; Sanchez, D. M.; Kulik, H. J.; Martínez, T. J. Exploiting graphical processing units to enable quantum chemistry calculation of large solvated molecules with conductor-like polarizable continuum models. *Int. J. Quantum Chem.* **2019**, *119*, e25760.
- [599] Batsanov, S. S. Van der Waals Radii of Elements. *Inorg. Mater.* **2001**, *37*, 871–885.
- [600] Maechler, M.; Rousseeuw, P.; Struyf, A.; Hubert, M.; Hornik, K. cluster: Cluster Analysis Basics and Extensions. 2019; <https://cran.r-project.org/package=cluster>.
- [601] Markwell, A. J.; Pratt, J. M.; Shaikjee, M. S.; Toerien, J. G. The Chemistry of Vitamin B12. Part 28. Crystal structure of Dicyanocobyrinic Acid Heptamethyl Ester and its Interaction with Alcohols: the Effects of Hydrogen Bonding to Co-ordinated Cyanide. *J. Chem. Soc., Dalt. Trans.* **1987**, 1349.
- [602] Grüning, B.; Holze, G.; Jenny, T. A.; Nesvadba, P.; Gossauer, A.; Ernst, L.; Sheldrick, W. S. Structure and Reactivity of Xanthocorrinoids. Part II. Influence of the *c*-Acetic-Acid Chain on the Course of the Hydroxylation of the Corrin Chromophore by Oxygen in the Presence of Ascorbic Acid. *Helv. Chim. Acta* **1985**, *68*, 1754–1770.

- [603] Janas, Z.; Sobota, P.; Lis, T. Interaction Of Tin Chlorides With Iron, Chromium And Vanadium Chlorides In Tetrahydrofuran. Crystal Structures Of  $[\text{Fe}_2(\mu\text{-Cl})_3(\text{thf})_6][\text{SnCl}_5(\text{thf})]$ ,  $[\text{Sn}_2(\mu\text{-OH})_2\text{Cl}_6(\text{thf})_2] \cdot 2\text{thf}$  and  $\text{trans-}[\text{CrCl}_2(\text{thf})_4][\text{SnCl}_5(\text{thf})]$ . *J. Chem. Soc., Dalt. Trans.* **1991**, 2429–2434.
- [604] Moubaraki, B.; Ley, M.; Benlian, D.; Sorbier, J. P. Structure and Single-Crystal Conductivity Measurements of Oxidized 3D Metal Phthalocyanines. Bis-Chloro Derivatives of Chromium(III), Iron(III) and Cobalt(III). *Acta Crystallogr., Sect. C Cryst. Struct. Commun.* **1990**, *46*, 379–385.
- [605] Oba, Y.; Mochida, T. Thermal Properties And Crystal Structures Of Cobalt(III)–Cyclam Complexes With The Bis(Trifluoromethanesulfonyl)Amide Anion (Cyclam=1,4,8,11-Tetraazacyclotetradecane). *Polyhedron* **2015**, *99*, 275–279.
- [606] Musaev, F.; Nadzhafov, G. N.; Amiraslanov, I. R.; Mamedov, K. S.; Movsumov, E. M. X-Raying Investigation into Complexes of n-Aminobenzoic Acid with Metals. *Zh. Strukt. Khim.* **1979**, *20*, 1075.
- [607] Janczak, J.; Kubiak, R. Stereochemistry And Properties Of The M(II)N(Py) Coordination Bond In The Low-Spin Dipyridinated Iron(II) And Cobalt(II) Phthalocyanines. *Inorg. Chim. Acta* **2003**, *342*, 64–76.
- [608] Ton, T. M. U.; Tejo, C.; Tania, S.; Chang, J. W. W.; Chan, P. W. H. Iron(II)-Catalyzed Amidation of Aldehydes with Iminoiodinanes at Room Temperature and under Microwave-Assisted Conditions. *J. Org. Chem.* **2011**, *76*, 4894–4904.
- [609] Mashiko, T.; Reed, C. A.; Haller, K. J.; Kastner, M. E.; Scheidt, W. R. Thioether Ligation in Iron-Porphyrin Complexes: Models for Cytochrome C. *J. Am. Chem. Soc.* **1981**, *103*, 5758–5767.
- [610] Hai, H.; Wang, H.; Jin, W. Y.; An, X. D.; Huang, W. G.; Zhang, S. H. Synthesis, Crystal Structures, and Properties of Three Complexes of 5-(pyridin-2-ylmethoxy) Isophthalic Acid. *Synth. React. Inorg., Met.-Org., Nano-Met. Chem.* **2015**, *45*, 1870–1874.
- [611] Nishijo, J.; Judai, K.; Numao, S.; Nishi, N. Chromium Acetylide Complex Based Ferrimagnet and Weak Ferromagnet. *Inorg. Chem.* **2009**, *48*, 9402–9408.

- [612] Yan, Y.-L.; Miller, M. T.; Cao, Y.; Cohen, S. M. Synthesis of Hydroxypyrrone- and Hydroxythiopyrrone-Based Matrix Metalloproteinase Inhibitors: Developing a Structure–Activity Relationship. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 1970–1976.
- [613] Aruffo, A. A.; Santarsiero, B. D.; Schomaker, V.; Lingafelter, E. C. Bis(phenylmethanethiolato)(2,3,9,10-tetramethyl-1,4,8,11-tetraaza-1,3,8,10-cyclotetradecatetraene)iron(III) hexafluorophosphate,  $C_{28}H_{38}FeN_4S^{2+} \cdot PF_6^-$ . *Acta Crystallogr., Sect. C Cryst. Struct. Commun.* **1984**, *40*, 1693–1695.
- [614] Liang, Q.-Q.; Liu, Z.-Y.; Yang, E.-C.; Zhao, X.-J. 1,2,4-Triazole Controlled Cd(II)/Mn(II) Complexes with Discrete Mononuclear, Polymeric 1D Chain, and 2D Layer Motifs. *Z. Anorg. Allg. Chem* **2009**,
- [615] Simonov, Y. A.; Dvorkin, A. A.; Bulgak, I. I.; Starish, M. P.; Batir, D. G. . *Koord. Khim.* **1979**, *5*, 1883.
- [616] Mossin, S.; Sørensen, H. O.; Weihe, H. Trans-bis(cyano- $\kappa$ C)(1,4,8,11-tetraazacyclotetradecane- $\kappa$ 4N)manganese(III) Perchlorate, a Low-Spin Manganese(III) Complex. *Acta Crystallogr., Sect. C Cryst. Struct. Commun.* **2002**, *58*.
- [617] Capilla, A. V.; Aranda, R. A.; Gomez-Beltran, F. Trans-diaqua-bis (1, 2-cyclohexanediamine) nickel (II) chloride ( $Ni(C_6H_{14}N_2)_2(H_2O)_2Cl_2$ ). *Cryst. Struct. Commun.* **1980**, *9*.
- [618] Pfluger, C. E.; Haradem, P. S. Coordination Sphere Geometry Of Tris(Acetylacetonato)Metal(II) Complexes: The Crystal And Molecular Structure Of Tris(1,1,1,5,5,5-Hexafluoroacetylacetonato)Iron(III). *Inorg. Chim. Acta* **1983**, *69*, 141–146.
- [619] Huxel, T.; Riedel, S.; Lach, J.; Klingele, J. Iron(II) and Nickel(II) Complexes of N-Alkylimidazoles and 1-Methyl-1H-1, 2, 4-Triazole: X-ray Studies, Magnetic Characterization, and DFT Calculations. *Z. Anorg. Allg. Chem* **2012**, *638*, 925–934.
- [620] Nfor, E. N.; Asobo, P. F.; Nenwa, J.; Nfor, O. N.; Njapba, J. N.; Njong, R. N.; Offiong, O. E. Nickel (II) and Iron (II) Complexes with Azole Derivatives: Synthesis, Crystal Structures and Antifungal Activities. *Int. J. Inorg. Chem.* **2013**, *2013*, 1–6.



- [621] Setifi, F.; Ota, A.; Ouahab, L.; Golhen, S.; Yamochi, H.; Saito, G. Charge Transfer Salts of BO with Paramagnetic Isothiocyanato Complex Anions: (BO)[M(isoq)<sub>2</sub>(NCS)<sub>4</sub>]; M=Cr(III) or Fe(III), isoq=isoquinoline and BO=Bis(ethylenedioxo)tetrathiafulvalene. *J. Solid State Chem.* **2002**, *168*, 450–456.
- [622] Marlin, D. S.; Olmstead, M. M.; Mascharak, P. K. Reaction of ( $\mu$ -Oxo)diiron(III) Core with CO<sub>2</sub> in N-Methylimidazole: Formation of Mono( $\mu$ -carboxylato)( $\mu$ -oxo)diiron(III) Complexes with N-Methylimidazole as Ligands. *Inorg. Chem.* **2003**, *42*, 1681–1687.
- [623] Patra, R.; Chaudhary, A.; Ghosh, S. K.; Rath, S. P. Modulation of Metal Displacements in a Saddle Distorted Macrocyclic Ligand: Synthesis, Structure, and Properties of High-Spin Fe(III) Porphyrins and Implications for the Hemoproteins. *Inorg. Chem.* **2008**, *47*, 8324–8335.
- [624] Bang, E.; Monsted, O. Chromium (III) Complexes of Macrocyclic Ligands. I. Crystal Structure of trans-Bis (0-carbamato)(1, 4, 8, 11-tetraazacyclotetradecane) chromium (III) Perchlorate Sesquihydrate. *Acta Chem. Scand. A* **1982**, 353.
- [625] Huxel, T.; Demeshko, S.; Klingele, J. 2-Amino-5-(2-pyridyl)-thiadiazole as Bidentate Ligand. *Z. Anorg. Allg. Chem.* **2015**, *641*, 1711–1717.
- [626] Ebralidze, I. I.; Leitus, G.; Shimon, L. J. W.; Wang, Y.; Shaik, S.; Neumann, R. Structural Variability in Manganese(II) Complexes of N,N-bis(2-pyridinylmethylene) Ethane (and Propane) Diamine Ligands. *Inorg. Chim. Acta* **2009**, *362*, 4713–4720.
- [627] Mossin, S.; Sørensen, H. O.; Weihe, H.; Glerup, J.; Søtofte, I. Manganese (III) Cyclam Complexes with Aqua, Iodo, Nitrito, Perchlorato and Acetic Acid/acetato Axial Ligands. *Inorg. Chim. Acta* **2005**, *358*, 1096–1106.
- [628] Sakiyama, H.; Mitsuhashi, R.; Mikuriya, M. Pseudo-S<sub>6</sub> Complex Cations of a Hexakis-N-methylformamide Nickel(II) Complex. *X-Ray Struct. Anal. Online* **2015**, *31*, 45–46.
- [629] Liu, Y.; Xu, D.; Liu, J. Synthesis And Crystal Structure Of Tetraimidazole(Diaqua)-Manganese(II) Terephthalate. *J. Coord. Chem.* **2001**, *54*, 175–181.

- [630] Zhao, H.-Y.; Ma, J.-J.; Yang, X.-D.; Yang, M.-L. Synthesis, Crystal Structure, and Antibacterial Activity of a New Cobalt(II) Complex Containing Imidazole as Ligand. *Synth. React. Inorg., Met.-Org., Nano-Met. Chem.* **2016**, *46*, 45–50.
- [631] Nishijo, J.; Enomoto, M. Synthesis, Structure And Magnetic Properties Of [Cr-cyclam(Cc-<sub>6</sub>-Methoxynaphthalene)<sub>2</sub>](Tcnq)N(1,2-Dichloroethane) (N=1, 2). *Inorg. Chim. Acta* **2015**, *437*, 59–63.
- [632] Hatlevik, Ø.; Arif, A. M.; Miller, J. S. Synthesis and Characterization of Hexakis(acetonitrile)chromium(III) tetrafluoroborate, [Cr(III)(NCMe)<sub>6</sub>][BF<sub>4</sub>]<sub>3</sub>. A Nonaqueous Cr(III) Source. *J. Phys. Chem. Solids* **2004**, *65*, 61–63.
- [633] Kaufman, L.; Rousseeuw, P. J. *Finding groups in data: an introduction to cluster analysis*; John Wiley & Sons, 2009; Vol. 344.

# Appendix A

## Surrogate models for transition metal complexes

#### Text A.1: Estimation of $\tau$

We determine a representative value of  $\tau$  by maximizing the log predictive likelihood of the corresponding GP based on the training data, which is a measure of how likely the observed data are under the GP, and is approximated<sup>459</sup> by

$$\log p(\mathbf{y}(\mathbf{x}_n) | \mathbf{x}_n, \mathbf{X}, \mathbf{Y}) \approx \log \left[ \sum_{j=1}^J e^{-\frac{1}{2}\tau \|\tilde{\mathbf{y}}(\mathbf{x}_n) - \tilde{\mathbf{y}}_j(\mathbf{x}_n)\|_2^2} \right] - \log J - \frac{1}{2} \log 2\pi - \frac{1}{2} \log \tau^{-1} \quad (\text{A.1})$$

Here, we have only scalar output and we use the training data to optimize eq. D.4 with respect to  $\tau$  numerically. We use  $J = 100$  repeats, as in the network itself. The determined values of  $\tau$ , based on the respective training data, are 0.4 for predicting the splitting energy, 0.07 for predicting the HF exchange sensitivity, and 10000 for the metal-ligand distances respectively. The magnitude of these numbers is close to the training errors observed:  $\sqrt{0.4^{-1}} \sim 2.5$ ;  $0.4^{-1} \sim 1.6$ ;  $\sqrt{0.007^{-1}} \sim 12$ ;  $0.4^{-1} \sim 0.01$ . These numbers represent the estimated inherent variance in the training data that limits the accuracy that could be expected from the trained networks.

#### Text A.2: Use of Coulomb matrix descriptor

We compare the descriptors proposed here to the Coulomb matrix descriptor, which has previously<sup>79,250</sup> been correlated with various molecular properties for a number of organic molecule data sets. In order to allow comparison of complexes with differing numbers of atoms, we pad all matrices with zeros to a size of  $151 \times 151$ , necessitating  $\mathcal{O}(10^4)$  elements per compound. We sort the rows and columns of the matrices in order to obtain indexing-invariant representations and use KRR with an exponential kernel and the matrix  $L_1$  norm as a distance metric as in Ref<sup>79</sup>. The complexes in our training data range in size from 7 to 151 atoms, but have a mean and median size of 38 and 29 atoms respectively. This large skew toward smaller complexes means that most of the descriptor elements are zero, and this may make learning good model parameters difficult. For example, the  $L_1$  distances between the sorted matrix representations of the small Fe(III)(CN)<sub>6</sub> complex and two large complexes, Fe(III)(tbuc)<sub>6</sub> and Fe(III)(pisc)<sub>6</sub> are very similar (36.85 to 36.88 where the range of distances spans  $\sim 20 - 60$ ), despite pisc being a similar strong C-connecting ligand and tbuc being a much weaker O-connecting ligand. We train and test on the same data as used in the other methods, but because the Coulomb matrix representation does not encode any functional-dependent information, we also provide a comparison against only B3LYP data (as opposed to varying HFX fractions).

### Text A.3: Testing ANN performance in molSimplify

In order to assess if the ANN can assist automated structure design, we used it to predict bond lengths instead of using the metal-ligand bond length database integrated into our structure generation toolbox, molSimplify<sup>181</sup>. We selected four of the original benchmark structures where molSimplify was found not to reduce RMS gradient error relative to simple force fields. Further details about the test cases are in the original paper. We project the negative of the energy gradient on the metal and connection atoms at the initial geometry onto the vector joining them, as explained in Figure S32, and use this as a measure of how close to an equilibrium bond length the initial geometry is. Note that a negative value for  $g$  means the bond would shrink in a steepest descent step, while a positive value means that it would lengthen. Large magnitudes indicate the bonds are far from equilibrium.

We achieve reductions in the absolute magnitude of  $g$  by 54–90% for bidentate cases and 7% for the monodentate case. We note that the reductions in the metal-ligand projected gradient do not necessarily correspond to reductions in the RMS gradient, which considers contributions from all atoms. In the Cr(bipy)<sub>3</sub> case, the RMS gradient is reduced by 30%, but it is unchanged or marginally higher in the other cases. This may be explained by considering the signs of the projected gradient, which show that the ANN universally reduces the metal-ligand bond length relative to original structure. This brings the bidentate ligands closer to the metal center and hence closer to each other, and we observe that the dominant contribution to the RMS gradient is from other atoms in the ligand structure. This could possibly be improved by training a similar ANN on the bite angles.

### Text A.4: Molecular descriptors for CSD compounds

The poor correlation ( $R^2 = 0.1$ ) between the Tanimoto dissimilarity (for CSD and training ligands) and the prediction error can be understood by considering that the molecular fingerprint is insensitive to the arrangement of groups in the ligand, so two ligands might appear similar in the Tanimoto metric because they both contain certain groups, but this does not ensure that the same groups are coordinating to the metal center. The descriptors used in this work strongly suggest that the immediate metal environment determines behavior of the complex, and so this highlights a specific difficulty in translating established ideas from organic molecular similarity analysis to transition metal systems.

Table A.1: Ligand properties

Number	ID	Name	Denticity	Charge	Connection	max $\delta\chi$	Bond Order	Truncated Kier
1	cl	chloride	1	1-	Cl	0	0	0
2	scn	thiocyanate	1	1-	S	0.03	2	2
3	pisc	t-butylphenyl isocyanide	1	0	C	-0.49	3	2.25
4	misc	Methyl isocyanide	1	0	C	-0.49	3	2
5	cn	cyanoate	1	1-	C	-0.49	3	0
6	co	carbonyl	1	0	C	-0.89	3	0
7	ncs	isothiocyanate	1	1-	N	0.49	2	2
8	ammo	ammonia	1	0	N	0.84	1	0
9	bipy	2,2-bipyridine	2	0	N	0.49	2	4.297
10	phen	phenanthroline	2	0	N	0.49	2	3.868
11	en	ethylenediamine	2	0	N	0.84	1	3
12	porphyrin	porphyrin	4	2-	N	0.49	2	6.958
13	h2o	water	1	0	O	1.24	1	0
14	acac	acetylacetonate	2	0	O	0.89	2	3.10
15	tbuc	t-butyl catecholate	2	2-	O	0.89	1	2.52
16	ox	oxalate	2	2-	O	0.89	2	2.22

Number	ID	SMILES
1	cl	[Cl-]
2	scn	[S-]C#N
3	pisc	CC(C)(C)C1=CC=C(C=C1)[N+]#[C-]
4	misc	C[N+]#[C-]
5	cn	[C-]#N
6	co	CO
7	ncs	[N-]=C=S
9	bipy	C1ccnc(c1)c2cccn2
10	phen	C1=CC2=CC=C3C=CC=NC3=C2N=C1
11	en	NCCN
12	porphyrin	[NH]1C2=CC3=NC(=CC4=CC=C([NH]4)C=C5C=CC(=N5)C=C1C=C2)C=C3
13	h2o	O
14	acac	CC(=O)CC(=O)C
15	tbuc	CC(C)(C)C1=CC(=C([O-])C=C1)[O-]
16	ox	[O-]C(=O)C([O-])=O

## Text A.5: Dissimilarity metrics for LS/HS bond length prediction

Using the same dissimilarity metrics that were employed to evaluate reliability of spin-state splitting, correlations between HS bond distance error and proximity to test data is smaller for both the HS bond distances ( $R^2 = 0.0, 0.1$  and  $0.2$  for the Tanimoto similarity metric, Pearson, and Euclidean distances, respectively) and LS bond distances ( $R^2 = 0$  for all metrics). However, we do observe that four of the five large (i.e.,  $> 0.1$  Å) HS bond distance errors have a minimum Euclidean distance greater than 1.0, supporting the use of this heuristic for evaluating prediction reliability. Bond length errors are generally smaller for LS states compared to HS states, with only two cases (tests 26 and 30) greater than  $0.1$  Å. We observe an overall correlation between the low spin bond distance prediction inaccuracy and poor splitting energy prediction, but bond lengths may still be well-predicted when spin-state splittings are not (e.g.,  $0.006$ – $0.03$  Å errors in LS bond distances for the cyclams).

Table A.2: Core homoleptic ligands

metal	oxidation	number of converged & included HFX values per ligand										total
		acac	bipy	c2h3s	cn	co	en	h2o	ncs	nh3	ox	
co	2	7	7	7	0	7	7	7	7	7	7	63
co	3	7	7	7	7	7	7	7	7	7	7	70
fe	2	7	0	7	7	7	7	7	7	7	7	63
fe	3	7	7	7	7	7	7	7	7	7	7	70
mn	2	7	7	7	0	7	7	7	7	7	5	61
mn	3	7	7	7	7	7	7	7	0	7	7	63
cr	2	7	7	7	0	7	7	7	0	7	7	56
cr	3	7	7	7	7	7	7	7	7	0	7	63
ni	2	7	7	7	5	7	7	7	7	7	7	68
total (core)		83 complexes, 577 HFX values, 1154 geometries										
total (additional)		111 complexes, 768 HFX values, 1536 geometries										
total (all)		194 complexes, 1345 HFX values, 2690 geometries										

Table A.3: List of test structures from the CSD, CSD IDs, metal and oxidation state along with a short note for each and references to the original observations

number	CSD ID	Metal	Ox.	Con. Atom (Ax./Equit)	Note	Ref.
1	CODZAW10	Co	2	C/N	dicyano-cobyrinate	601
2	DORLEB	Co	2	C/N	tetrapyrrole	602
3	KOCLET	Cr	3	Cl/O	tetrahydrofuran dichloride	603
4	KEHXEA	Co	3	Cl/N	pthalocyanine	604
5	ZUNSEI	Co	3	Cl/N	cyclam	605
6	ABZACO10	Co	2	O/O	benzoate	606
7	LUWPAU	Fe	2	N/N	pthalocyanine	607
8	TPYFEC04	Fe	2	Cl/N	tetrapyridine	608
9	TPPSFE10	Fe	3	S/N	substitued porphyrin	609
10	FUMJOO	Mn	2	O/O	large oxygen ligand	610
11	BUHKIA	Ni	2	O/O	sulfoxide diphenyl- propandionate	231
12	SUMLET	Cr	3	C/N	cyclam	611
13	BOSDIX	Fe	3	O/O	bidentate oxygen	612
14	BMADFE10	Fe	3	S/N	cyclam	613
15	AHAVUB	Mn	2	N/O	dinitrobenzoate	614
16	DOCNFE	Fe	2	C/N	benzylidioximate	615
17	AFAROO	Mn	3	C/N	cyclam	616
18	AHDNIC	Ni	2	O/N	cyclam	617
19	BUPTAH	Fe	3	O/O	bidentate oxygen	618
20	DEDKII	Fe	2	O/N	nitrogen rings and oxy- gen	619
21	PUTHIX	Fe	2	N/N	bidentate nitrogen	61
22	AGIZEX	Fe	2	N/N	bidentate nitrogen	620
23	AKAGEY	Fe	3	N/N	large monodentate ni- trogen	621
24	GUWYUS	Fe	3	N/N	monodentate nitrogen ligand	622
25	EGILOW	Fe	3	O/N	heavily substituted porphyrin	623
26	BINPET	Cr	3	O/N	cyclam, oxygen	624
27	JUSCIL	Fe	2	N/N	bidentate nitrogen	625
28	BULVAG	Mn	2	O/N	salen-like	626
29	FIXGID	Mn	3	O/N	cyclam	627
30	KUSKEQ	Ni	2	O/O	monodentate oxygen	628
31	AGUWEE	Mn	2	O/N	cyclic nitrogen	629
32	DMAZCO03	Co	2	N/N	imidazole	630
33	YUJCIQ	Cr	3	C/N	cyclam	631
34	ACALEW	Cr	3	N/N	monodentate nitrogen	632
35	AKAGAU	Cr	3	N/N	monodentate nitrogen	621



Table A.4: Relaxed geometry optimization tolerances for some CSD structures .

number	spin	energy tolerance, au
HFX = 0.20		
2	LS	$1.05 \times 10^{-5}$
2	HS	$5 \times 10^{-5}$
1	LS	$5 \times 10^{-6}$

Table A.5: Excluded structures due to spin contamination

metal	ox	axlig	eqlig	aHF	metal	ox	axlig	eqlig	aHF
cr	3	nh3	nh3	all (7)	fe	2	tbuc	tbuc	0.30
mn	3	pisc	h2o	0.30	mn	2	ox	ox	0.25, 0.30
cr	3	tbuc	tbuc	0, 0.05					
	total			5 complexes, 13 HFX values					

Table A.6: Excluded structures due to geometric breakup

metal	ox	axlig	eqlig	aHF	metal	ox	axlig	eqlig	aHF
co	3	nh3	co	0	ni	2	cn	cn	0,0.05
mn	3	ncs	ncs	all (7)	cr	2	ncs	ncs	all (7)
mn	2	cn	cn	all (7)	co	2	cn	cn	all (7)
cr	2	cn	cn	all (7)	cr	2	ncs	pisc	0
fe	3	co	scn	0,0.05	fe	2	ncs	pisc	0.25
ni	2	nh3	cn	all (7)	ni	2	scn	ox	all (7)
ni	2	h2o	h2o	all (7)					
	total			13 complexes, 63 HFX values					

Table A.7: List of input space descriptors and the normalization constants used in the ANNs. For a given variable  $x$ , the normalization is  $\tilde{x} = \frac{x-c}{f}$ .

name	unit	$c$	$f$
split energy	kcal/mol	-54.19	142.71
HFX sensitivity	kcal/mol.HFX	-174.20	161.58
ls min bond	Å	1.8146	0.6910
hs min bond	Å	1.8882	0.6956
oxidation state		2	1
$a_{HF}$		0	0.3
axlig charge		-2	2
eqlig charge		-2	2
axlig dent		1	1
eqlig dent		1	3
mdelen		-5.34	12.54
maxmd		-0.89	2.09
axlig bo		0	3
eqlig bo		0.00	3
axlig ki		0.00	4.29
eqlig ki		0.00	6.96

Table A.8: Variable Selection for  $\Delta E_{\text{HS-LS}}$ : set **a**. Values are given for regularized and unregularized coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	7.13	1.14
metalcr	-23.3	-17.3
metalfe	-8.43	-2.76
metalmn	-25.4	-19
metalni	-29.8	-23.9
ox	6.73	6
alpha	-69.4	-64.4
axligbipy	1.82	3.48
axligc2h3ns	29.2	24.4
axligcl	-9.54	-4.77
axligcn	11.7	12.5
axligco	8.79	4.07
axligen	-4.65	0
axligh2o	-10.3	-9.9
axligncs	-10.2	-4.6
axlignh3	4.86	0.21
axligox	-22	-0.42
axligphen	-1.86	0
axligpisc	1.23	0
axligporphyrin	0	0
axligscn	-1.19	0
axligtbuc	-8.8	0
eqligbipy	9.07	0
eqligc2h3n	-1.19	$1.37 \times 10^{-5}$
eqligcl	-6.67	-2.43
eqligcn	8.06	10.2
eqligco	11.1	12.7
eqligen	10.1	0.284
eqligh2o	-4.94	-4.31
eqligncs	-3.36	-0.298
eqlignh3	-0.946	0
eqligox	9.04	0
eqligphen	15.8	6.99
eqligpisc	34.2	31.2
eqligporphyrin	35.3	13.2
eqligscn	-4.83	0
eqligtbuc	-4.26	-1.58
axlig charge	$-9.92 \times 10^{-2}$	5.36
eqlig charge	-2.29	0
axlig dent	4.05	0
eqlig dent	-7.86	0
MSE	199	213

Table A.9: Variable Selection for  $\Delta E_{\text{HS-LS}}$ : set **b**. Values are given for regularized and unregularized linear coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	20.6	18.3
metalcr	-22.3	-17.9
metalfe	-7.03	-2.25
metalmn	-24	-19.1
metalni	-28.6	-23.6
ox	6.63	6.1
alpha	-69.5	-65
axlig charge	3.55	2
eqlig charge	-0.613	0.107
axlig dent	6.83	0.999
eqlig dent	-0.633	0
axlig connectCl	-12.9	-9.59
axlig connectN	-10.2	-6.14
axlig connectO	-17.1	-12.4
axlig connectS	-4.6	-2.39
eqlig connectCl	-23.1	-21.2
eqlig connectN	-18.2	-16.1
eqlig connectO	-23.6	-21.5
eqlig connectS	-21.6	-18.8
axlig natoms	-0.235	0
eqlig natoms	0.525	0.414
MSE	218	227

Table A.10: Variable Selection for  $\Delta E_{\text{HS-LS}}$ : set **c**. Values are given for regularized and unregularized linear coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	26.4	13.6
metalcr	-22.5	-18.6
metalfe	-7.1	-3.18
metalmn	-24	-20
metalni	-28.9	-24.7
ox	6.6	6.17
alpha	-69.4	-66.1
axlig charge	3.57	2.34
eqlig charge	-1.22	0.702
axlig dent	8.6	0
eqlig dent	-0.841	0
axlig connectCl	-17.3	-7.43
axlig connectN	-19	-0.818
axlig connectO	-28.9	-4.73
axlig connectS	-9.04	0
eqlig connectCl	-26	-16.8
eqlig connectN	-24.6	-5.78
eqlig connectO	-31.5	-8.28
eqlig connectS	-24.6	-14
axlig natoms	-0.299	0
eqlig natoms	0.568	0.399
$\Sigma \Delta\chi$	-0.203	-1.95
min $\Delta\chi$	7.08	0
max $\Delta\chi$	5.43	-1.33
MSE	216	230

Table A.11: Variable Selection for  $\Delta E_{\text{HS-LS}}$ : set **d**. Values are given for regularized and unregularized linear coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	26.1	14.8
metalcr	-22.5	-18.9
metalfe	-7.1	-3.49
metalmn	-24.1	-20.3
metalni	-28.8	-25.1
ox	6.58	6.2
alpha	-69.4	-66.4
axlig charge	3.69	2.39
eqlig charge	-1.3	0.595
axlig dent	7.92	0.115
eqlig dent	-0.48	0
axlig connectCl	-15.2	-7.78
axlig connectN	-15.2	-1.6
axlig connectO	-23.9	-5.75
axlig connectS	-7.03	0
eqlig connectCl	-27.3	-17.8
eqlig connectN	-26.9	-7.23
eqlig connectO	-34.5	-10.1
eqlig connectS	-25.8	-15.1
axlig natoms	-0.277	0
eqlig natoms	0.556	0.407
max $\Delta\chi$	-1.79	-1.47
$\Sigma \Delta\chi$	2	-1.68
MSE	216	228

Table A.12: Variable selection for  $\Delta E_{\text{HS-LS}}$ : set **e**. Values are given for regularized and unregularized linear coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	32.7	19.1
metalcr	-22.2	-19.8
metalfe	-7.72	-4.74
metalmn	-24.1	-21.5
metalni	-29.6	-26.4
ox	6.6	6.38
alpha	-69.3	-67.3
axlig charge	5.12	2.83
eqlig charge	-3.68	0
axlig dent	6.49	1.89
eqlig dent	0.769	0.283
axlig connectCl	-16.7	-10.8
axlig connectN	-13.6	-4.97
axlig connectO	-18.1	-8.77
axlig connectS	0.368	0
eqlig connectCl	-26.5	-19.7
eqlig connectN	-30.7	-14
eqlig connectO	-46.1	-20.4
eqlig connectS	-37.8	-21.2
axlig natoms	$-4.3 \times 10^{-2}$	0
eqlig natoms	0.32	0.378
axlig bo	3.64	1.39
eqlig bo	-5.96	-1.78
$\Sigma \Delta\chi$	2.25	-0.591
max $\Delta\chi$	-2.41	-1.46
MSE	213	221

Table A.13: Variable selection for  $\Delta E_{\text{HS-LS}}$ : set **f**. Values are given for regularized and unregularized linear coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	27.8	17.8
metalcr	-22.4	-19.5
metalfe	-6.83	-3.71
metalmn	-24.3	-21.1
metalni	-29.9	-26
ox	6.69	6.27
alpha	-69.4	-66.9
axlig charge	2.38	2.11
eqlig charge	-0.265	0.992
axlig dent	11.3	2.18
eqlig dent	-2.95	0.846
axlig connectCl	-17.1	-9.14
axlig connectN	-14.4	-4.04
axlig connectO	-25.7	-9.6
axlig connectS	-1.94	-0.457
eqlig connectCl	-28.6	-21.5
eqlig connectN	-33.9	-14.6
eqlig connectO	-39.4	-17.7
eqlig connectS	-35.4	-22
axlig ki	-2.65	-0.343
eqlig ki	4	1.78
max $\Delta\chi$	-1.19	-0.721
$\Sigma \Delta\chi$	2.48	-0.868
MSE	218	227



Table A.14: Variable selection for  $\Delta E_{\text{HS-LS}}$ : set **g**. Values are given for regularized and unregularized linear coefficients and MSE in kcal<sup>2</sup>/mol<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	35.2	21.6
metalcr	-22.2	-19.6
metalfe	-7.5	-4.28
metalmn	-24.4	-21.5
metalni	-30.1	-26.5
ox	6.63	6.32
alpha	-69.3	-67.1
axlig charge	4.04	2.75
eqlig charge	-2.84	0
axlig dent	10.1	2.19
eqlig dent	0.186	2.29
axlig connectCl	-17.7	-10.4
axlig connectN	-13.9	-4.73
axlig connectO	-22	-8.81
axlig connectS	0.214	0
eqlig connectCl	-27.7	-21.3
eqlig connectN	-34	-16.3
eqlig connectO	-47.9	-23.7
eqlig connectS	-41.9	-25.8
axlig bo	2.05	1.24
eqlig bo	-5.52	-2.71
axlig ki	-1.66	0
eqlig ki	1.97	0.925
max $\Delta\chi$	-2.08	-1.09
$\Sigma \Delta\chi$	2.41	-0.622
MSE	215	223

Table A.15: Variable selection for  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$ : set **a**. Values are given for regularized and unregularized coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HFX<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-167	-167
metalcr	21.7	21.2
metalfc	-26.3	-26.3
metalmn	-15.2	-15.2
metalni	14.9	14.2
ox	28.2	27.9
eqligbipy	-20.7	-16.6
eqligc2h3ns	-30.5	-30.1
eqligcl	-2.95	-0.898
eqligcn	-6.33	-4.87
eqligco	-18.4	-14.8
eqligen	-4.02	-1.31
eqligh2o	8.22	11.8
eqligncs	1.03	2.08
eqlignh3	4.88	7.82
eqligox	-0.272	-1.45
eqligphen	-1.28	0
eqligpisc	-31.6	-27.5
eqligporphyrin	8.4	1.34
eqligscn	6.26	6.82
eqligtbuc	10.8	7.53
axligbipy	15.5	10.2
axligc2h3ns	-1.39	-0.326
axligcl	6.15	0.329
axligcn	-9.2	-14
axligco	-21.8	-24.1
axligen	17.4	13.5
axligh2o	15	11.8
axligncs	7.92	2.13
axlignh3	4.47	2.07
axligox	11	0.491
axligphen	-10.2	-11.8
axligpisc	-16.3	-18.9
axligporphyrin	0	0
axligscn	2.97	-1.65
axligtbus	8.32	0
eqlig charge	2.31	0
axlig charge	-1.12	-4.04
eqlig dent	-1.75	0
axlig dent	$-8.15 \times 10^{-2}$	0
alpha	5.72	5.42
MSE	422	423

Table A.16: Variable selection for  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$ : set **b**. Values are given for regularized and unregularized linear coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HF<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-189	-186
metalcr	21.9	21.7
metalfe	-26.9	-26.7
metalmn	-16.3	-16.1
metalni	13.8	13.3
ox	28.6	28.2
eqlig charge	-0.266	-0.359
axlig charge	-3.79	-3.49
eqlig dent	0.179	0
axlig dent	-7.42	-6.37
eqlig natoms	-0.47	-0.452
axlig natoms	$2.13 \times 10^{-2}$	0
eqlig connectCl	10.5	9.61
eqlig connectN	17.6	17
eqlig connectO	19.3	18.9
eqlig connectS	20.1	19.2
axlig connectCl	18.6	18.1
axlig connectN	24.5	23.8
axlig connectO	29.7	28.8
axlig connectS	18.1	17.7
alpha	4.89	4.58
MSE	472	472

Table A.17: Variable selection for  $\frac{\partial \Delta_{\text{EH-L}}}{\partial a_{\text{HF}}}$ : set **c**. Values are given for regularized and unregularized linear coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HF<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-170	-167
(Intercept)	0	0
metalcr	20.8	20.9
metalfe	-27.6	-27.2
metalmn	-17.1	-16.8
metalni	13.1	12.8
ox	28.6	28.4
eqlig charge	-2.06	-1.99
axlig charge	-3.98	-3.44
eqlig dent	2.38	0
axlig dent	0.717	0
eqlig natoms	-0.385	-0.325
axlig natoms	-0.144	-0.161
eqlig connectCl	-2.56	0.619
eqlig connectN	-9.37	-2.65
eqlig connectO	-13.9	-5.88
eqlig connectS	6.76	9.66
axlig connectCl	6.83	7.52
axlig connectN	-0.514	2.22
axlig connectO	-4.21	0
axlig connectS	5.29	6.36
alpha	4.21	3.69
$\Sigma \Delta \chi$	1.76	1.02
max $\Delta \chi$	14.4	13.1
min $\Delta \chi$	14.2	13
MSE	449	450

Table A.18: Variable selection for  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$ : set **d**. Values are given for regularized and unregularized linear coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HF<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-168	-166
metalcr	20.8	20.9
metalfe	-27.4	-27.2
metalmn	-17.2	-17
metalni	13.2	12.8
ox	28.6	28.4
eqlig charge	-2.24	-2.05
axlig charge	-3.61	-3.24
eqlig dent	0.574	-0.291
axlig dent	-1.65	-1.14
eqlig natoms	-0.388	-0.347
axlig natoms	-0.108	-0.147
eqlig connectCl	-4.31	0
eqlig connectN	-12.4	-6.75
eqlig connectO	-18.8	-11.6
eqlig connectS	5.36	8.2
axlig connectCl	10.9	10.6
axlig connectN	7.03	8.26
axlig connectO	5.61	7.47
axlig connectS	9.34	9.41
alpha	4.17	3.55
$\Sigma \Delta \chi$	5.86	4.97
max $\Delta \chi$	0.542	0.725
MSE	451	452

Table A.19: Variable selection for  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$ : set **e**. Values are given for regularized and unregularized linear coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HF<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-190	-187
metalcr	20.7	20.6
metalfe	-26.5	-26.4
metalmn	-16.9	-16.8
metalni	13.9	13.5
ox	28.6	28.3
eqlig charge	0.243	0
axlig charge	-4.29	-3.67
eqlig dent	0.362	-1.06
axlig dent	-3.11	-1.84
eqlig natoms	-0.252	-0.172
axlig natoms	$-6.34 \times 10^{-2}$	-0.159
eqlig connectCl	-4.69	-0.473
eqlig connectN	-9.57	-3.89
eqlig connectO	-8.03	-0.743
eqlig connectS	16.6	19.5
axlig connectCl	12.8	12.3
axlig connectN	8.55	9.11
axlig connectO	7.62	8.08
axlig connectS	10.1	9.26
alpha	5.12	4.43
axlig bo	$5 \times 10^{-1}$	0
eqlig bo	5.57	5.77
$\Sigma \Delta \chi$	5.62	4.81
max $\Delta \chi$	1.4	1.45
MSE	444	445

Table A.20: Variable selection for  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$ : set **f**. Values are given for regularized and unregularized linear coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HFX<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-192	-122
metalcr	20.9	17.9
metalfe	-27.7	-21.6
metalmn	-16.8	-10.8
metalni	14.3	1.98
ox	28.5	19.8
eqlig charge	-1.58	-0.317
axlig charge	-3.9	-2.2
eqlig dent	9.74	0
axlig dent	-0.587	0
eqlig connectCl	-1.05	0
eqlig connectN	-0.537	0
eqlig connectO	-5.67	0
eqlig connectS	16.9	0
axlig connectCl	15.5	0
axlig connectN	13.8	0
axlig connectO	15.2	0
axlig connectS	11.4	0
n alpha	5.18	0
axlig ki	-0.177	0
eqlig ki	-4.11	0
max $\Delta\chi$	0.434	0
$\Sigma\Delta\chi$	3.91	4.27
MSE	453	553

Table A.21: Variable selection for  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$ : set **g**. Values are given for regularized and unregularized linear coefficients and MSE in HF kcal<sup>2</sup>/mol<sup>2</sup>HF<sup>2</sup>

	unregularized	$\mathcal{L}_1$ regularization
(Intercept)	-192	-187
metalcr	20.9	20.8
metalfc	-27.7	-27.5
metalmn	-16.8	-16.6
metalni	14.3	13.8
ox	28.5	28.2
eqlig charge	-1.58	-1.68
axlig charge	-3.9	-3.66
eqlig dent	9.74	7.56
axlig dent	-0.587	0
eqlig connectCl	-1.05	0
eqlig connectN	-0.537	0
eqlig connectO	-5.67	-3.96
eqlig connectS	16.9	15.7
axlig connectCl	15.5	14.3
axlig connectN	13.8	13.3
axlig connectO	15.2	14.4
axlig connectS	11.4	11.3
alpha	5.18	4.64
axlig ki	-0.177	-0.364
eqlig ki	-4.11	-3.37
max $\Delta\chi$	0.434	0.522
$\Sigma\Delta\chi$	3.91	3.83
MSE	453	454



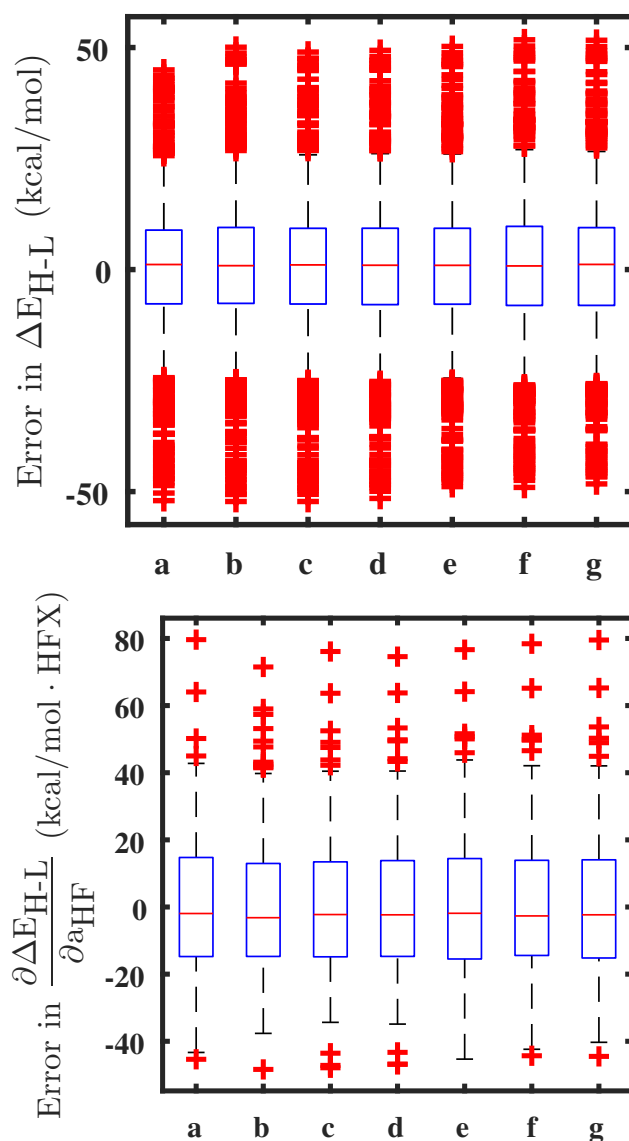


Figure A-1: Comparison of errors for different descriptor sets for a regularized linear effects model predicting  $\Delta E_{H-L}$  in kcal/mol (left) and  $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$  in kcal/molHF (right). Set **a** includes the metal properties and full ligand identity and number of atoms. Set **b** replaces ligand identity with the identity of connection atom only, while set **c** adds information from the sum, maximum and minimum ligand  $\delta\chi$  to set **b**. Set **d** is the same as set **c** but excludes the minimum  $\delta\chi$ . Set **e** adds in bond order information with an MSE, while set **f** replaces the ligand size metric with our truncated index. Set **g** represents our final set, and includes the same descriptors from **f** and adds bond order information.

Table A.22: Optimal hyperparameter selection for KRR and SVR models found via a grid search and 10-fold cross-validation. Parameters were selected using a Cartesian grid search in  $[10^{-7}, 10^4]$  for the regularization weights,  $[10^{-4}, 10^3]$  for the exponential kernel correlation length and  $[0.1, 0.9]$  for  $\nu$ .

	$\sigma$ (kernel lengthscale)	$\lambda/C$ (KRR/SVR) (regularization weight)	$\nu$ (SV fraction)
KRR (set <b>g</b> )	1	$10^{-4}$	
SVR (set <b>g</b> )	1	100	0.75
KRR (sorted Coulomb Matrix, B3LYP)			
KRR (sorted Coulomb Matrix, B3LYP)	316	0.01	

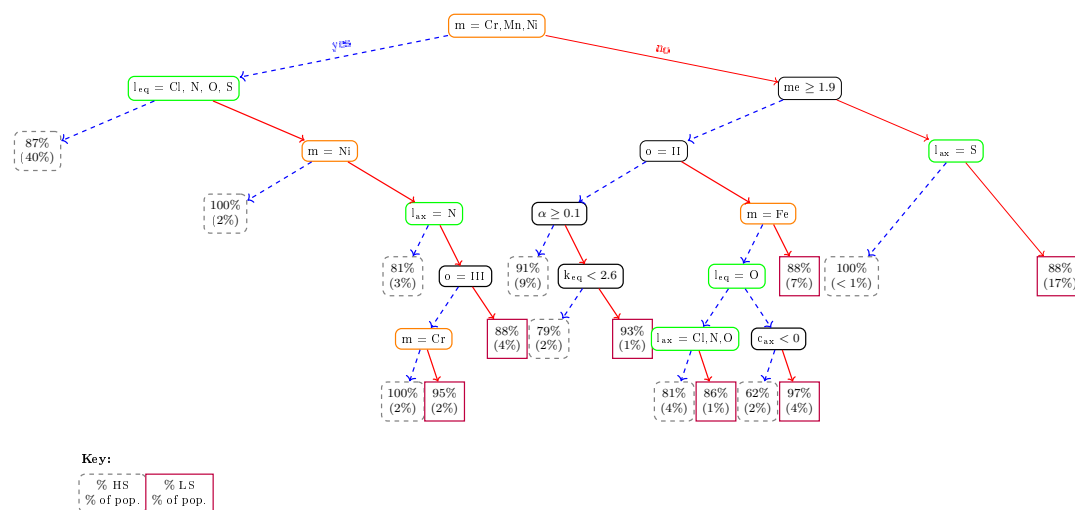


Figure A-2: Binary ground state classification tree for heteroleptic compounds. M indicates metal identity, l ligand connection atom, o oxidation state,  $\alpha$  the fraction of HF exchange and me the sum of  $\delta\chi$  values across ligands. The first line in each leaf node is the percent of elements in that leaf that have the indicated ground state, and the second line indicates the percentage of the total heteroleptic population in each leaf node. Dashed blue arrows indicate yes, solid red arrows indicate no.

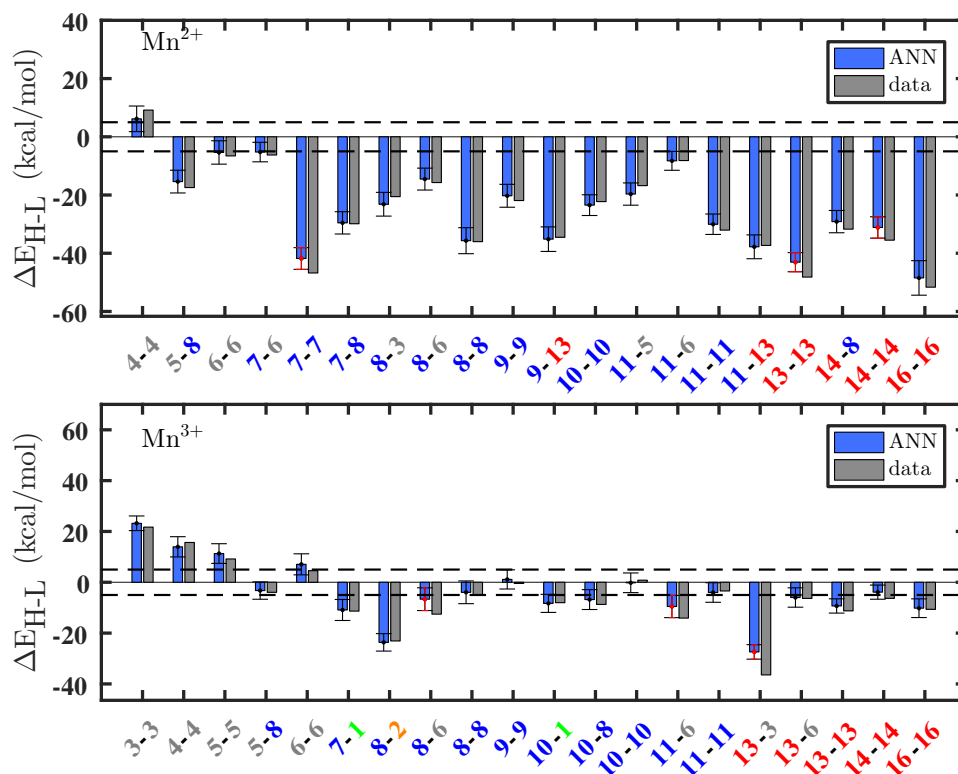


Figure A-3: Model predictions of  $\Delta E_{H-L}$  and data for Mn using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

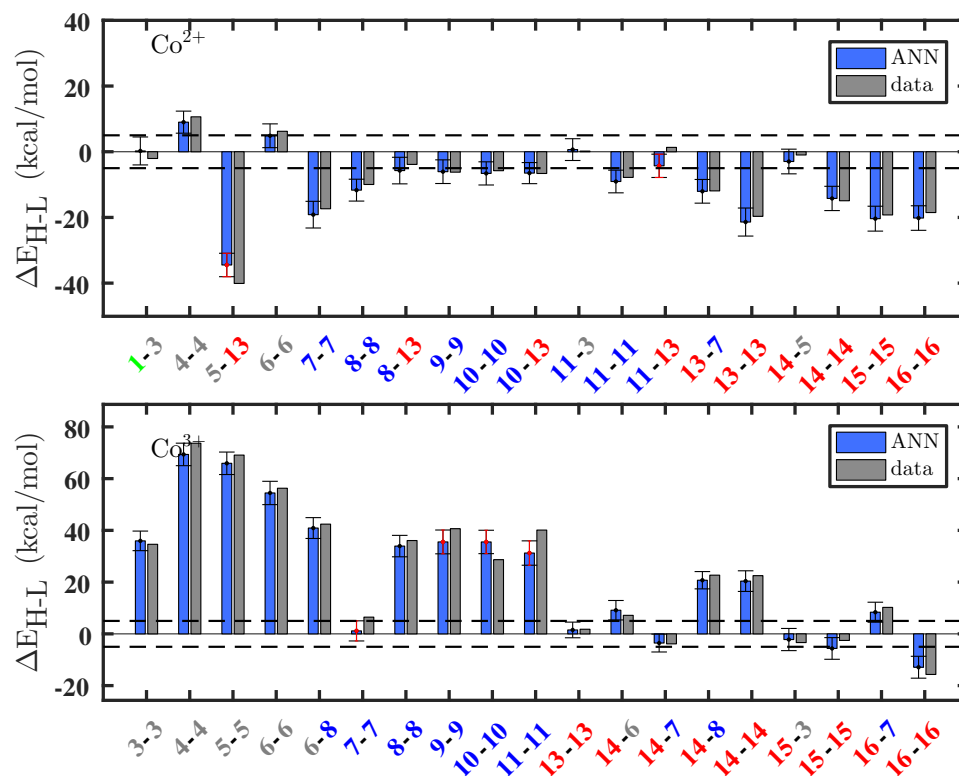


Figure A-4: Model predictions of  $\Delta E_{\text{HS-LS}}$  and data for Co using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

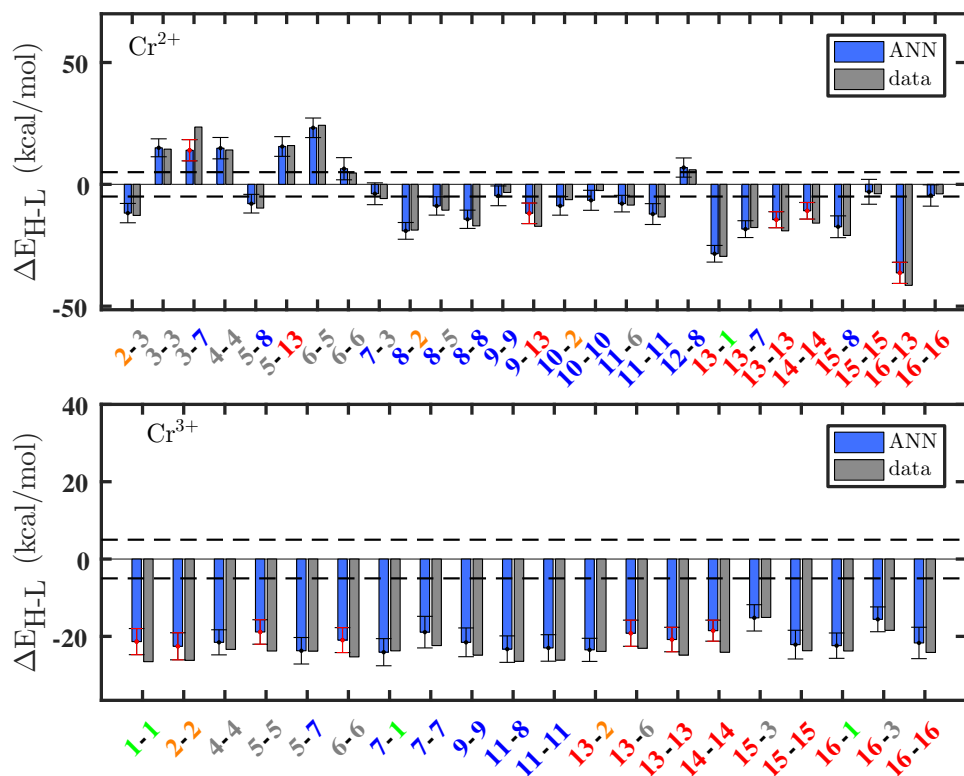


Figure A-5: Model predictions of  $\Delta E_{\text{H-L}}$  and data for Cr using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction..

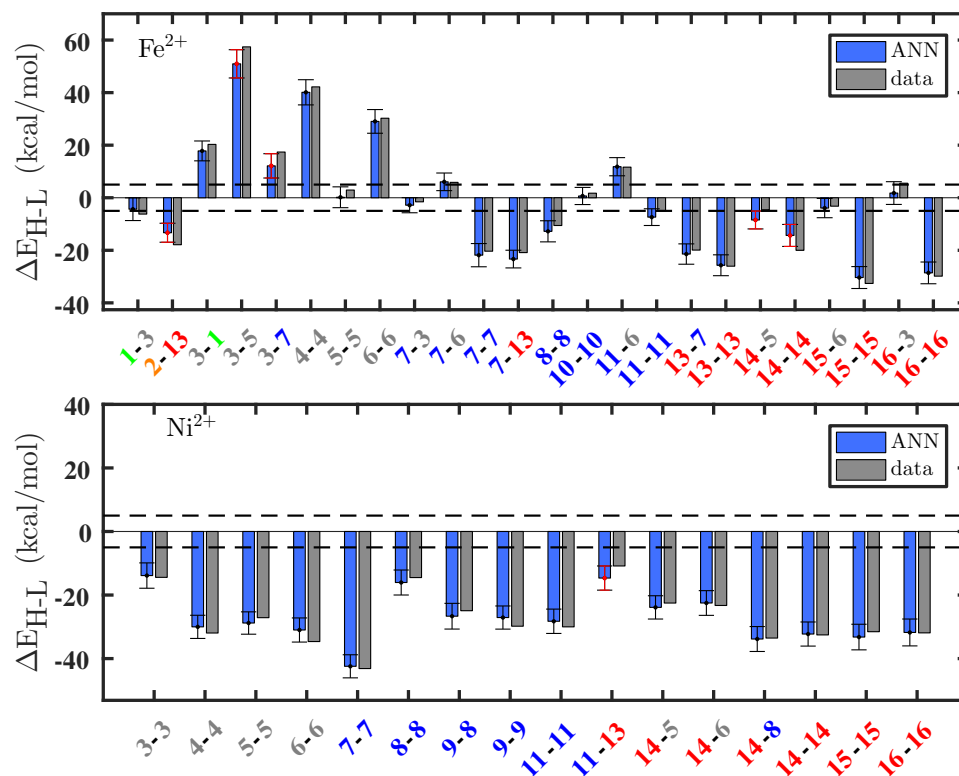


Figure A-6: Model predictions of  $\Delta E_{H-L}$  and data for Fe(II) (top) and Ni(II) (bottom) using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction..

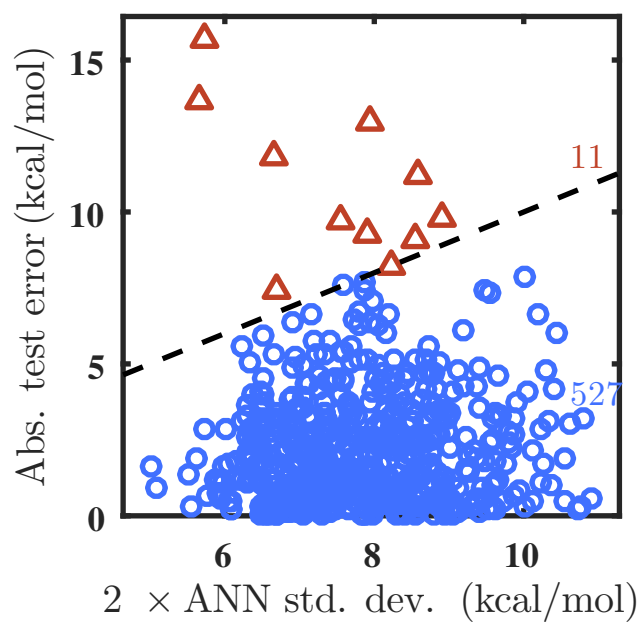


Figure A-7: Parity plot for  $\pm 2$  standard deviation from the mean prediction and absolute prediction error for test case  $\Delta E_{H-L}$  prediction using ANN. All units are kcal/mol. The black line is  $y = x$ .

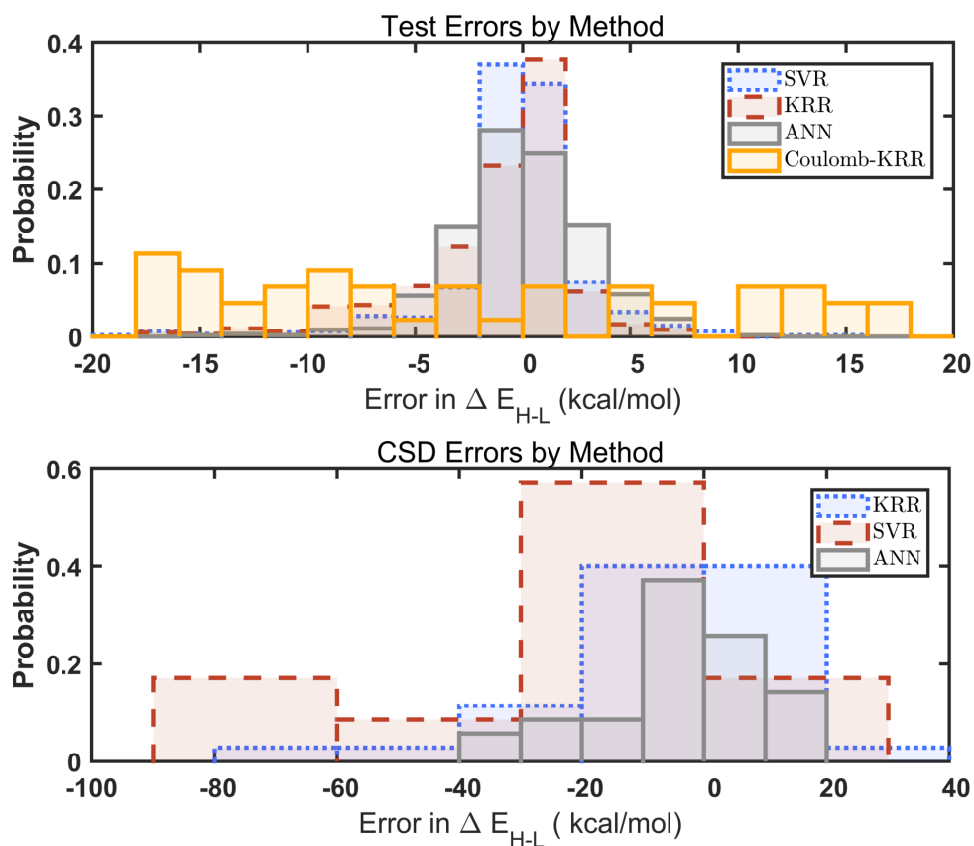


Figure A-8: Normalized error histogram for HF = 0.2 (B3LYP) test data (top) and CSD structures (bottom), comparing ANN, KRR and SVR models using descriptor set  $\mathbf{g}$ , as well as a KRR model using the Coulomb matrix descriptor (trained on B3LYP data only).



Table A.23: RMS prediction errors for  $\Delta E_{\text{H-L}}$  on test data using a deep ANN in kcal/mol for test data divided by metal and oxidation state. The number of test cases is indicated in parentheses

Species	RMS Test error	Min. Abs. Test Error	Max. Abs. Test Error
Cr(II)	3.3 (86)	0.03	11.2
Cr(III)	2.5 (50)	0.06	6.3
Mn(II)	2.8 (57)	0.01	7.5
Mn(III)	2.1 (60)	0.04	5.6
Fe(II)	3.7 (60)	0.07	13.0
Fe(III)	3.0 (71)	0.09	7.9
Co(II)	2.8 (58)	0.07	11.8
Co(III)	4.5 (57)	0.02	15.7
Ni(II)	2.5 (39)	0.09	5.8

Table A.24: Average HF exchange sensitivity values, in kcal/mol.HFX, for homoleptic compounds with C, N and O ligands grouped by metal, oxidation state and ligand connecting atom

Metal	Oxidation state	Ligand connection		
		C	N	O
Cr	II	-82	-55	-59
	III	-20	-20	-18
Mn	II	-167	-92	-87
	III	-58	-44	-32
Fe	II	-164	-76	-59
	III	-118	-74	-63
Co	II	-106	-48	-45
	III	-98	-53	-46
Ni	II	-65	-57	-52

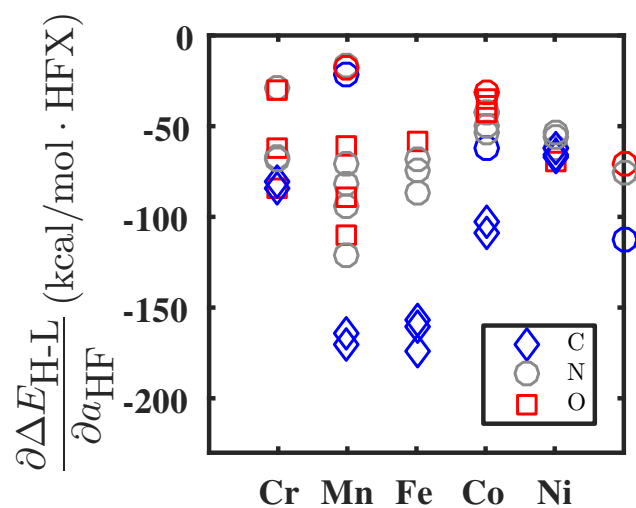


Figure A-9: Scatter plot of HFX sensitivity in kcal/mol · HFX for homoleptic M(II) complexes, colored by connection atom, for homoleptic (II) complexes with C (gray), N (blue) and O (square) ligands.

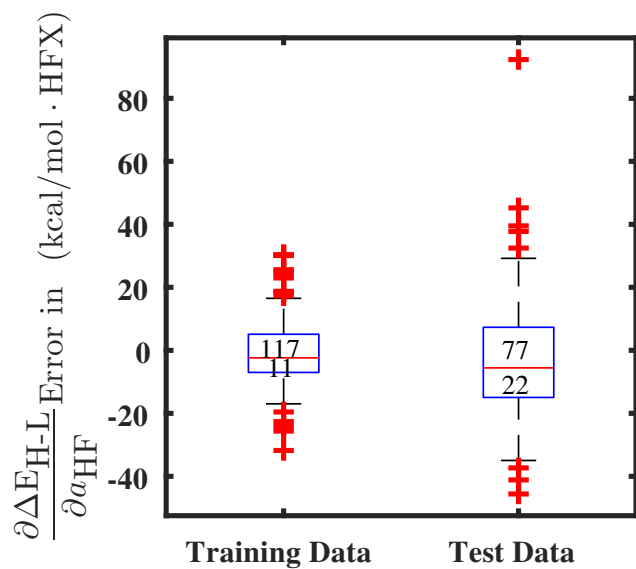


Figure A-10: Error boxplot for regression of  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$  using an ANN, showing training and test data comparison. The top number indicates the number of trials, while the bottom indicates the RMSE.

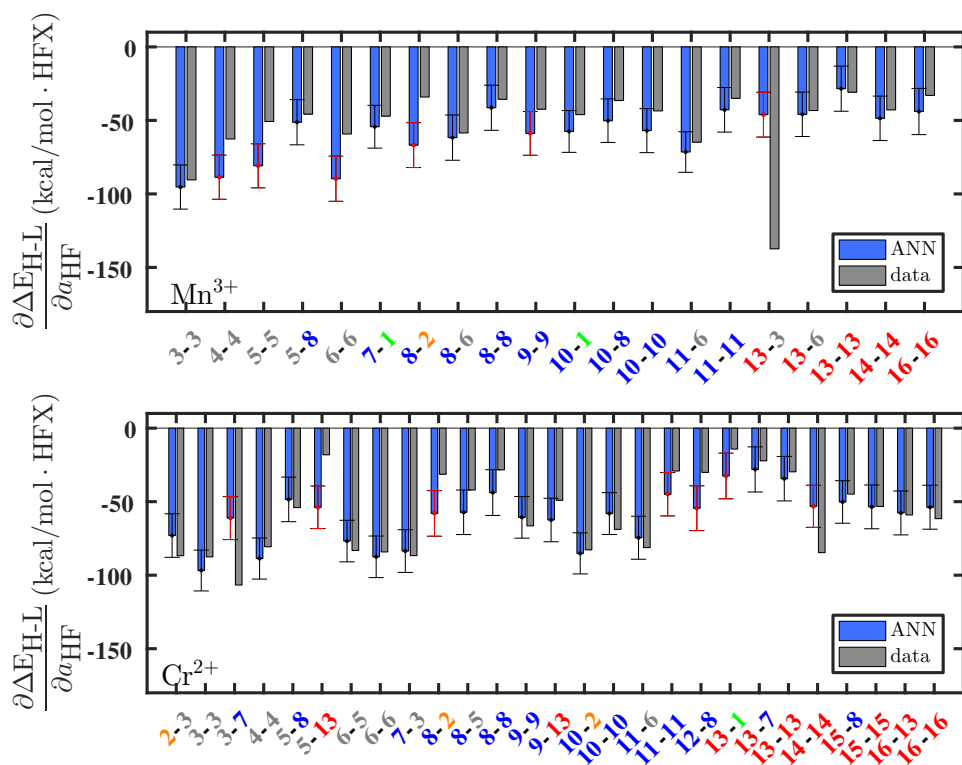


Figure A-11: Model predictions of  $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$  and data for Co using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

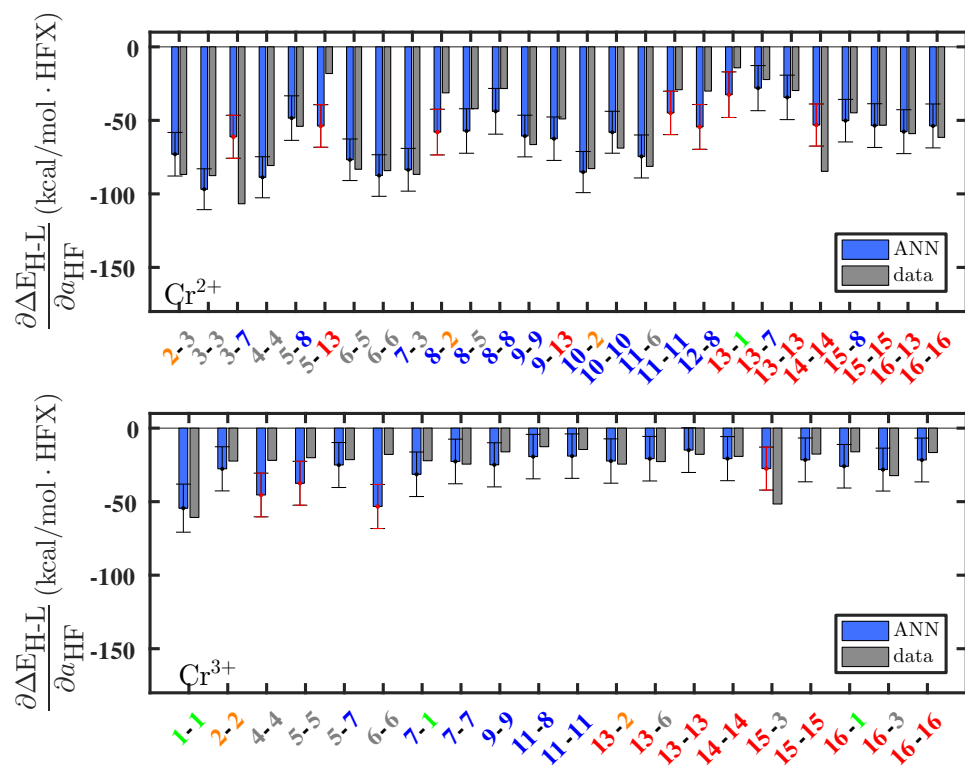


Figure A-12: Model predictions of  $\frac{\partial \Delta E_{\text{H-L}}}{\partial a_{\text{HF}}}$  and data for Cr using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

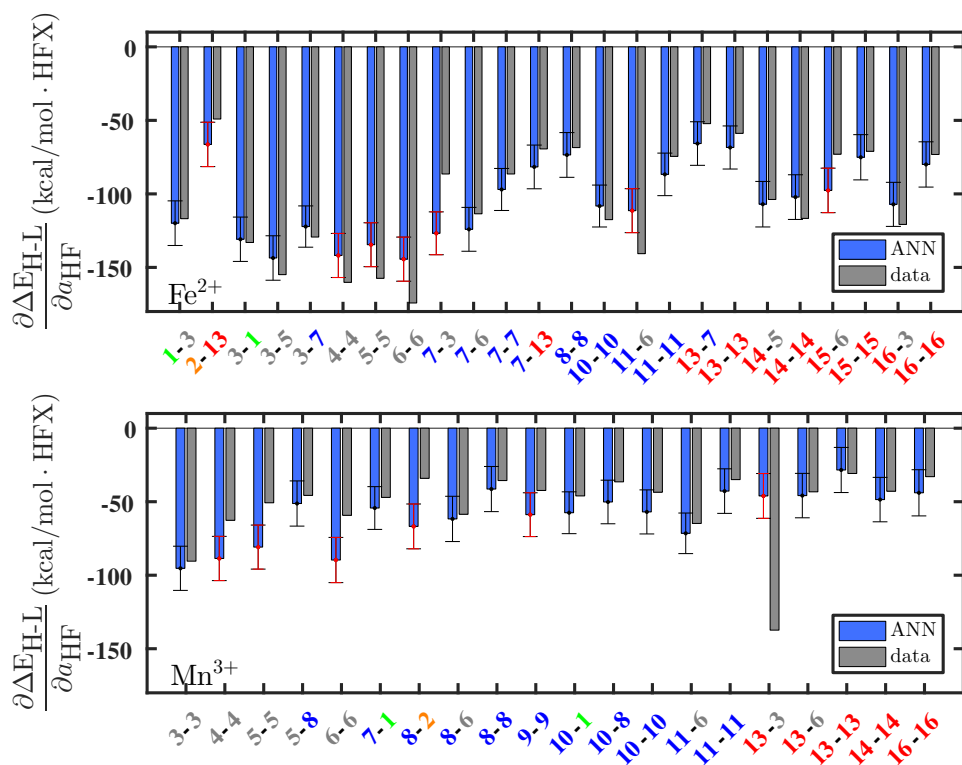


Figure A-13: Model predictions of  $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$  and data for low-spin Fe(II) (top) and Ni(II) (bottom) using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

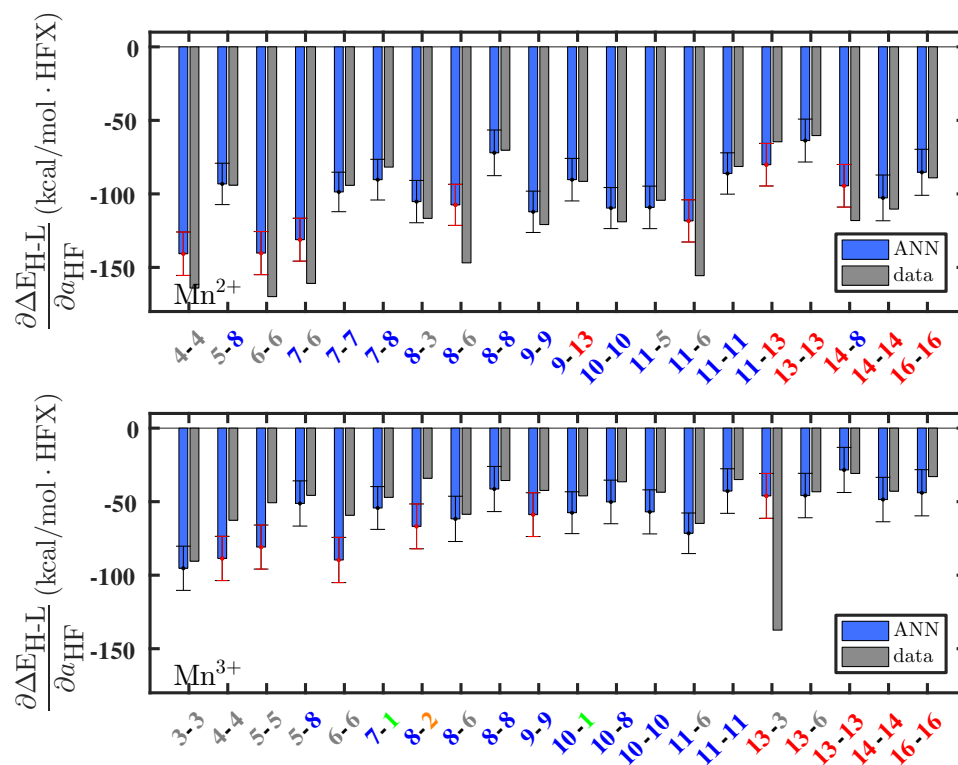


Figure A-14: Model predictions of  $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$  and data for Mn using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

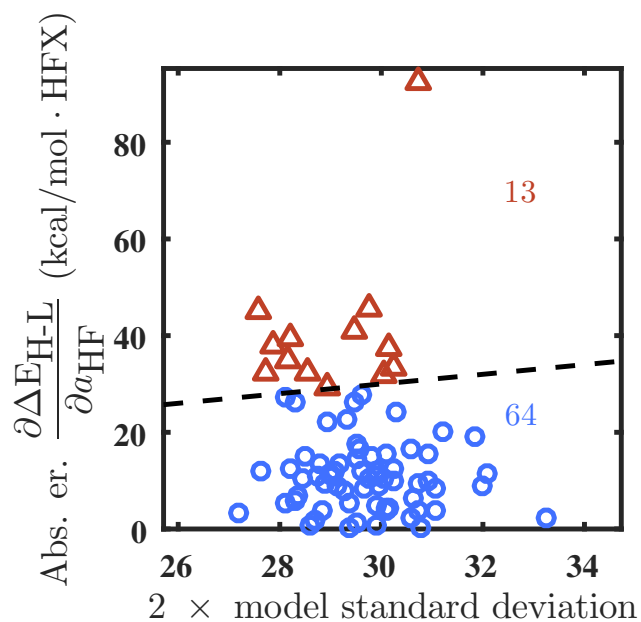


Figure A-15: Parity plot for  $\pm 2$  standard deviation from the mean prediction and absolute prediction error for test case  $\frac{\partial \Delta E_{H-L}}{\partial a_{HF}}$  prediction using ANN. All units are kcal/mol. The black line is  $y = x$ .

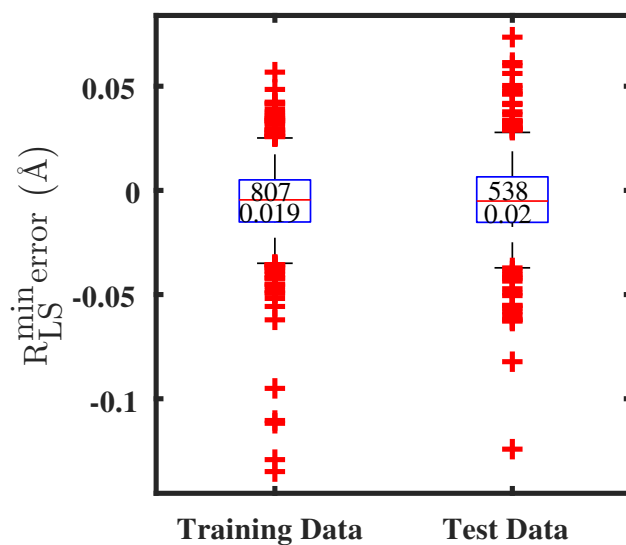


Figure A-16: Error boxplot for regression of  $R_{LS}^{\min}$  using an ANN, showing training and test data comparison. The top number indicates the number of trials, while the bottom indicates the RMSE.

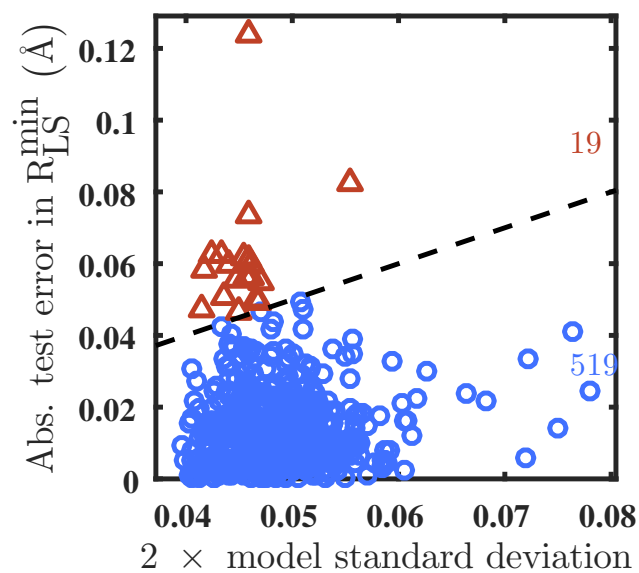


Figure A-17: Parity plot for  $\pm 2$  standard deviation from the mean prediction and absolute prediction error for test case  $R_{LS}^{\min}$  prediction using ANN. All units are Å. The black line is  $y = x$ .



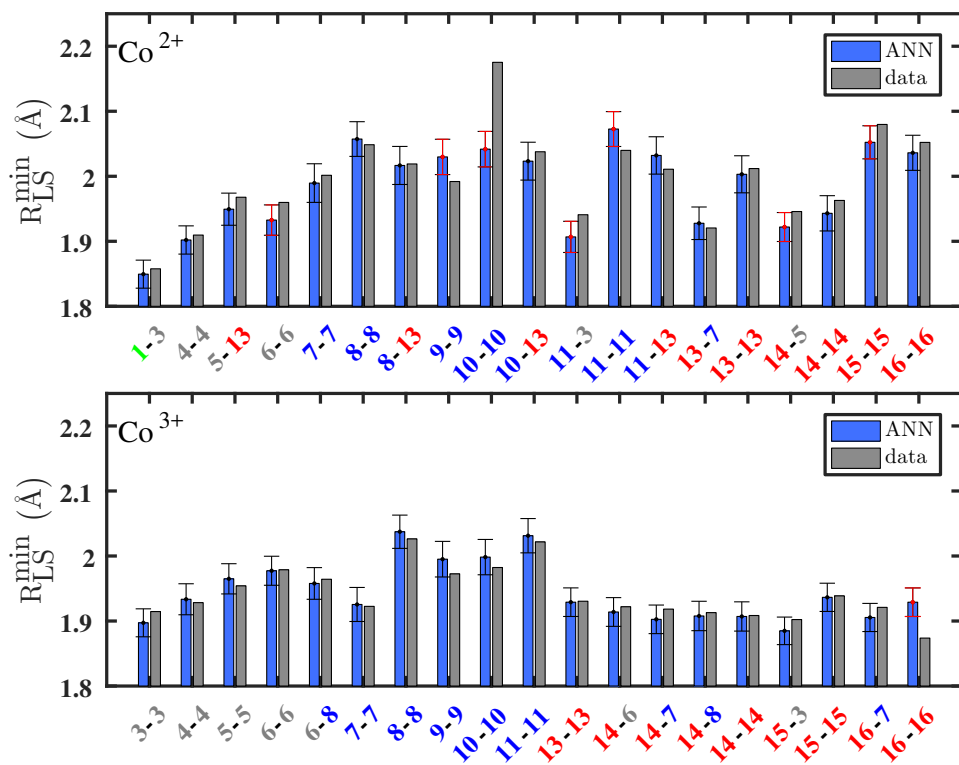


Figure A-18: Model predictions of  $R_{\text{LS}}^{\text{min}}$  and data for low-spin Co using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

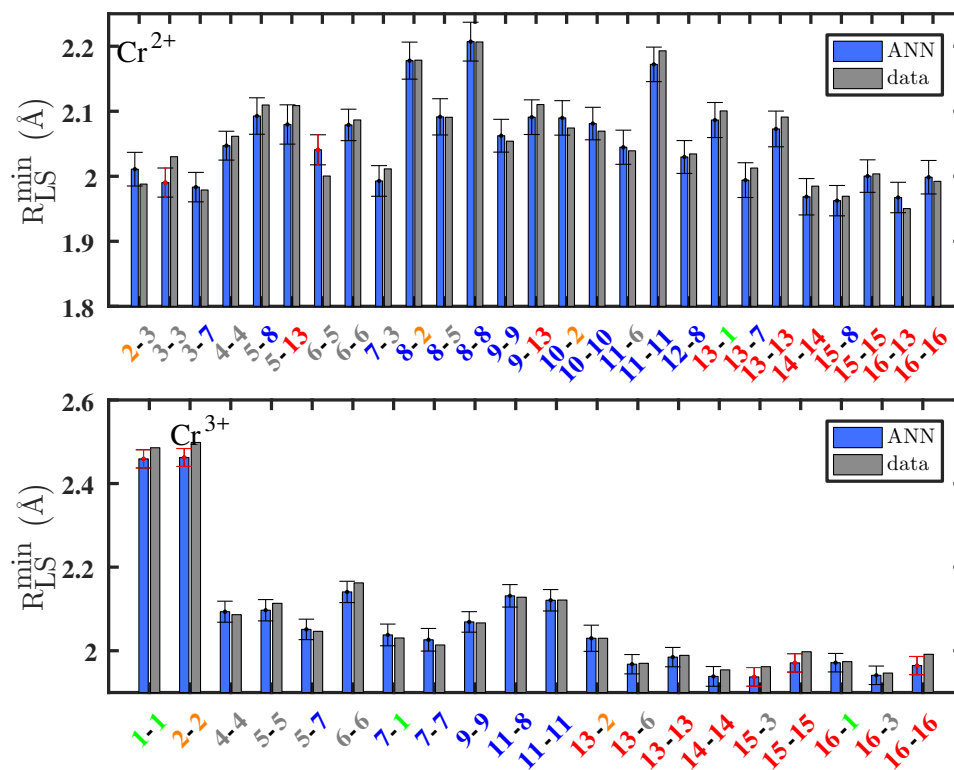


Figure A-19: Model predictions of  $R_{\text{LS}}^{\text{min}}$  and data for low-spin Cr using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

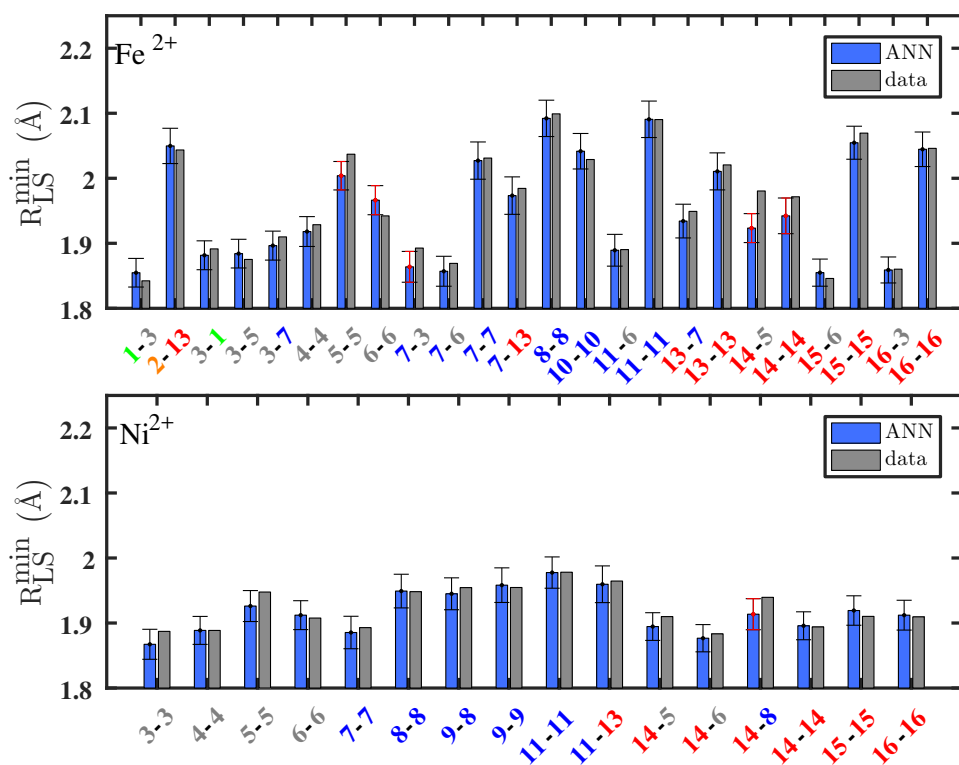


Figure A-20: Model predictions of  $R_{LS}^{\min}$  and data for low-spin Fe(II) (top) and Ni(II) (bottom) using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

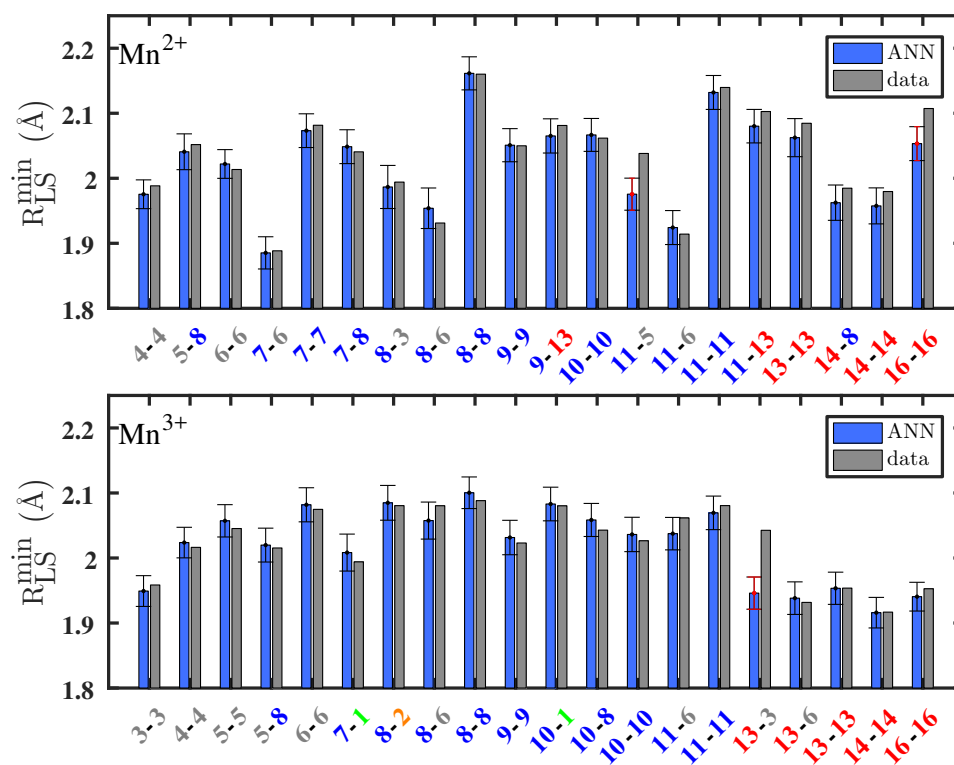


Figure A-21: Model predictions of  $R_{LS}^{min}$  and data for low-spin Mn using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

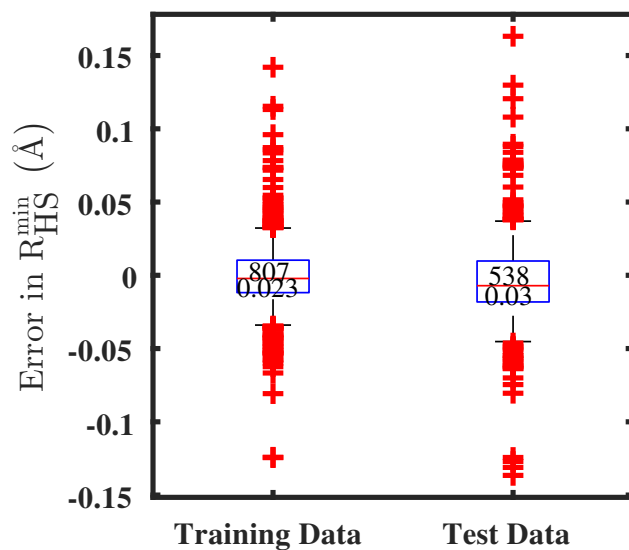


Figure A-22: Error boxplot for regression of  $R_{\text{HS}}^{\text{min}}$  using an ANN, showing training and test data comparison. The top number indicates the number of trials, while the bottom indicates the RMSE.

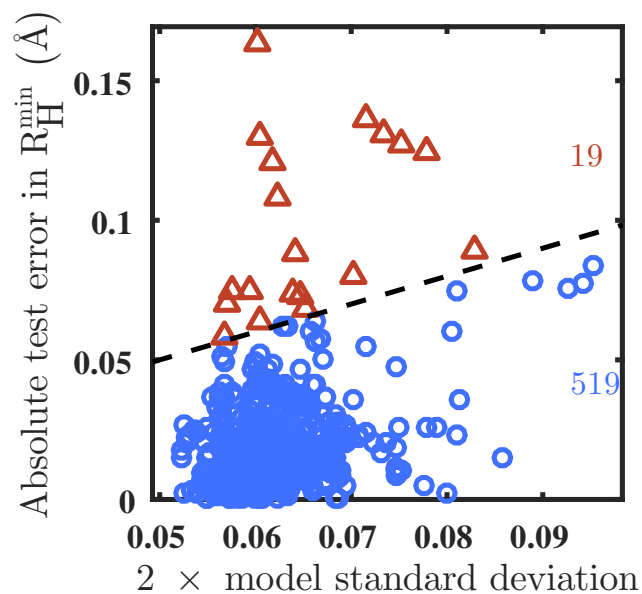


Figure A-23: Parity plot for  $\pm 2$  standard deviation from the mean prediction and absolute prediction error for test case  $R_{\text{HS}}^{\text{min}}$  prediction using ANN. All units are Å. The black line is  $y = x$ .

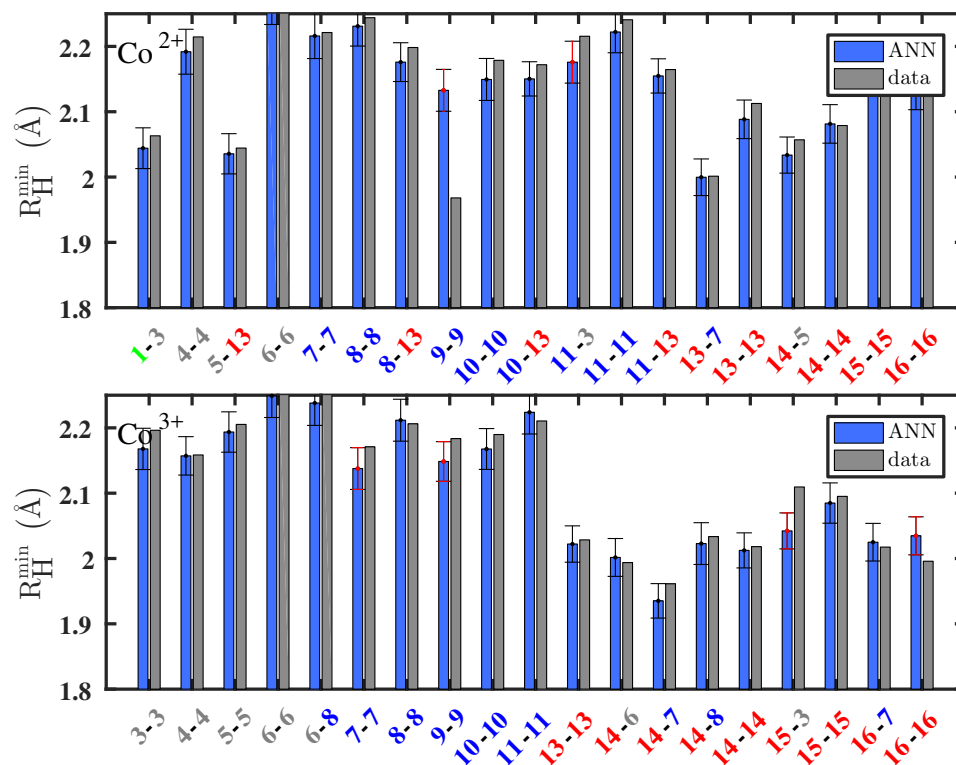


Figure A-24: Model predictions of  $R_{\text{HS}}^{\text{min}}$  and data for high-spin Co using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

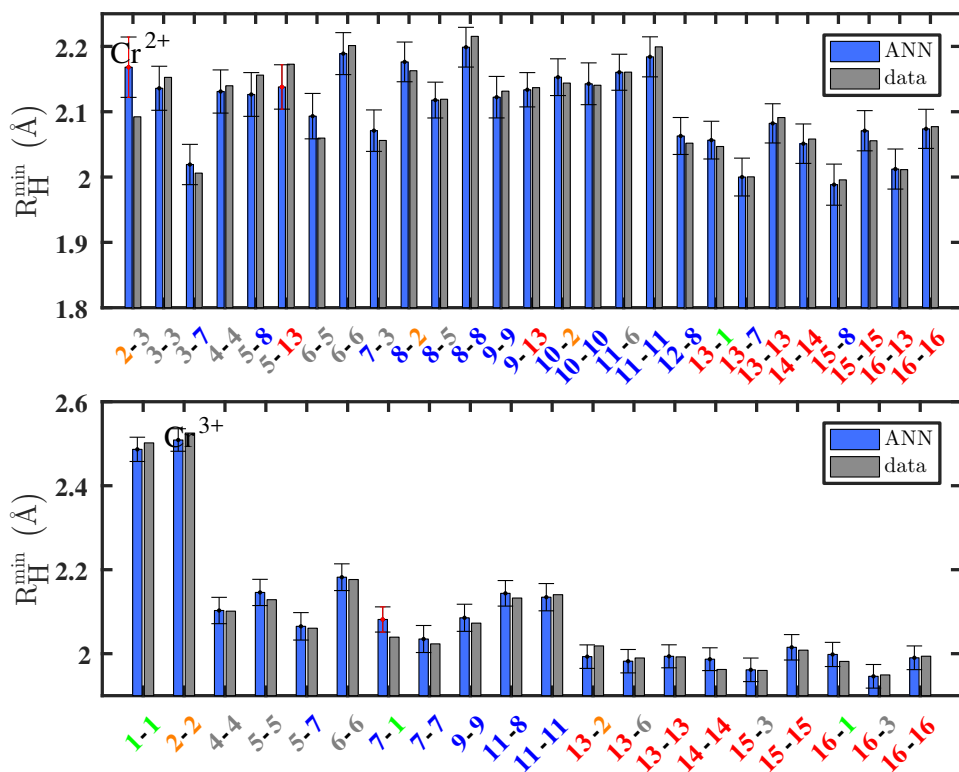


Figure A-25: Model predictions of  $R_{\text{HS}}^{\text{min}}$  and data for high-spin Cr using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

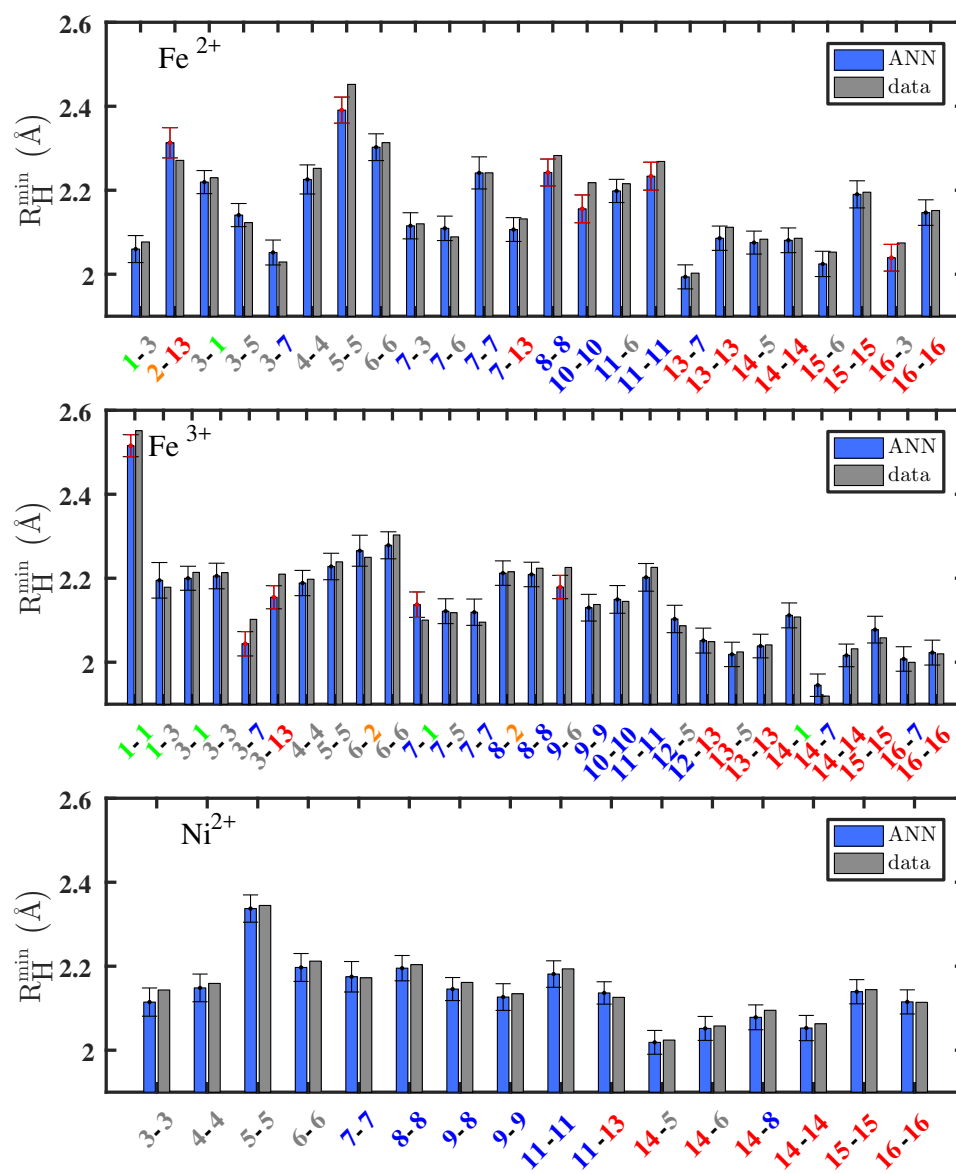


Figure A-26: Model predictions of  $R_{HS}^{\min}$  and data for high-spin Fe(II) (top), Fe(III) (middle) and Ni(II) (bottom) using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.



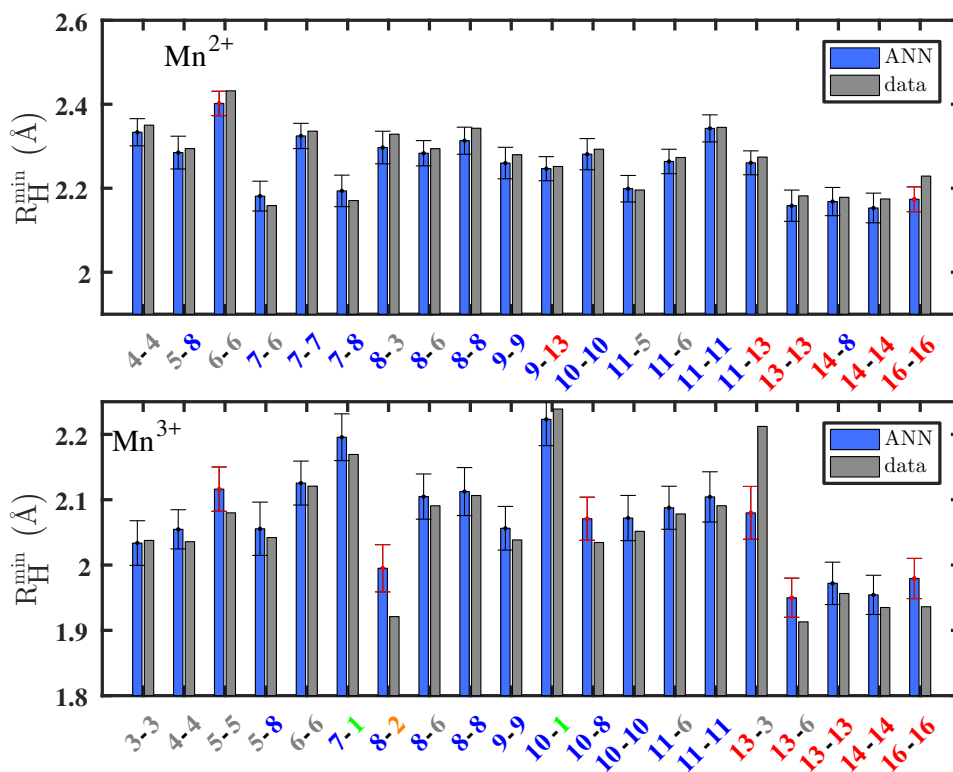


Figure A-27: Model predictions of  $R_{HS}^{\min}$  and data for high-spin Mn using an ANN. The ligands are described by two numbers indicating the equatorial first and then the axial, color coded by ligand identity (green for halogen, gray for carbon, blue for nitrogen, and red for oxygen). The error bars represent an estimated  $\pm 1$  standard deviation from the mean prediction.

Table A.25: RMSEs and MUEs (in Å) for minimum metal-ligand bond length prediction on test data by metal and oxidation state for both min(RLS) and min(RHS). The number of test cases is indicated in parentheses.

Species	RMSE (Å)		MUE (Å)	
	min(RLS)	min(RHS)	min(RLS)	min(RHS)
Cr(II)	0.02 (68)	0.02 (78)	0.02	0.01
Cr(III)	0.02 (54)	0.02 (54)	0.01	0.01
Mn(II)	0.02 (65)	0.03 (49)	0.02	0.02
Mn(III)	0.02 (60)	0.05 (59)	0.01	0.03
Fe(II)	0.02 (64)	0.03 (59)	0.01	0.03
Fe(III)	0.02 (84)	0.03 (88)	0.01	0.02
Co(II)	0.03 (55)	0.04 (53)	0.02	0.03
Co(III)	0.02 (52)	0.02 (51)	0.02	0.02
Ni(II)	0.01 (36)	0.02 (47)	0.01	0.01

Table A.26: molSimplify initial structure projected (g) gradients and RMS gradients with and without preliminary ANN assisted bond lengths.

name		g (kcal/Å)	RMS grad. (Hartree/Bohr)
Co(acac) <sub>3</sub>	default	-52	0.0487
	ANN	-5	0.0490
Cr(bipy) <sub>3</sub>	default	-48	0.0237
	ANN	29	0.0161
Fe(acac) <sub>3</sub>	default	-50	0.0429
	ANN	23	0.0437
Mn(misc) <sub>3</sub>	default	-57	0.0459
	ANN	-53	0.0454

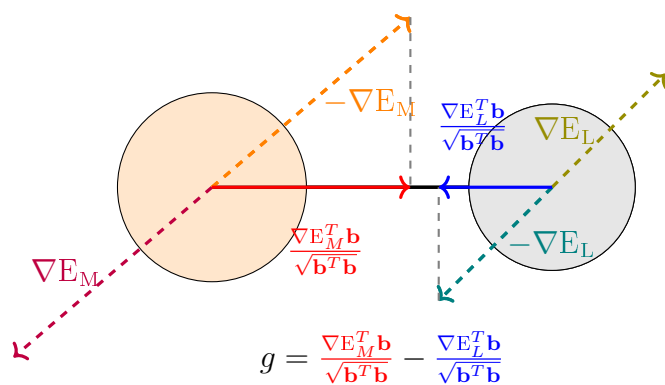


Figure A-28: Definition of bond projected gradient,  $g$ , used to estimate the closeness of an initial geometry to equilibrium. The projected gradient is the scalar difference between the component of the negative energy gradient projected into the vector joining nuclear positions of the metal (larger orange circle) and the ligand (smaller grey circle).

Table A.27: Spin splitting energy predictions and data in kcal/mol, for the CSD test structures. For each complex, values from DFT using B3LYP and ANN predictions are shown, along with the standard deviation of the ANN model.

Test number	ANN	DFT/B3LYP	ANN Std. Dev.	Error
1	21.80	29.06	3.29	-7.81
2	21.90	14.20	3.28	7.70
3	-26.50	-23.90	3.25	-2.66
4	28.70	34.10	3.65	-5.35
5	32.90	35.10	3.46	-2.18
6	-17.00	-11.50	3.87	-5.53
7	12.60	35.2	4.25	-22.60
8	-13.80	-16.40	4.14	2.61
9	30.60	0.26	3.77	30.40
10	-44.50	-40.90	3.27	-3.61
11	-31.20	-15.20	3.74	-16.00
12	4.50	-23.2	5.29	27.70
13	-13.10	-16.70	4.00	3.62
14	28.30	30.70	4.60	-2.38
15	-61.00	-45.10	3.08	-15.9
16	44.40	69.70	5.06	-25.30
17	6.92	3.51	3.92	3.41
18	-19.50	-6.35	3.25	-13.10
19	-11.50	-3.14	4.27	-8.32
20	-21.90	-15.00	3.69	-6.93
21	-5.78	-1.69	3.23	-4.10
22	-11.70	-8.49	3.91	-3.19
23	-4.20	-5.36	3.52	1.16
24	6.44	2.57	3.46	3.87
25	9.44	-0.23	4.73	9.67
26	-7.91	-24.40	4.00	16.40
27	0.54	3.59	3.24	-3.04
28	-13.40	-36.90	5.56	23.50
29	-6.85	-23.90	3.30	17.10
30	-15.60	-17.20	3.76	1.63
31	-28.20	-38.30	3.26	10.10
32	-10.00	-8.98	3.60	-1.03
33	4.50	-23.40	5.29	27.90
34	-22.00	-24.00	3.29	2.01
35	-13.90	-22.40	4.03	8.55

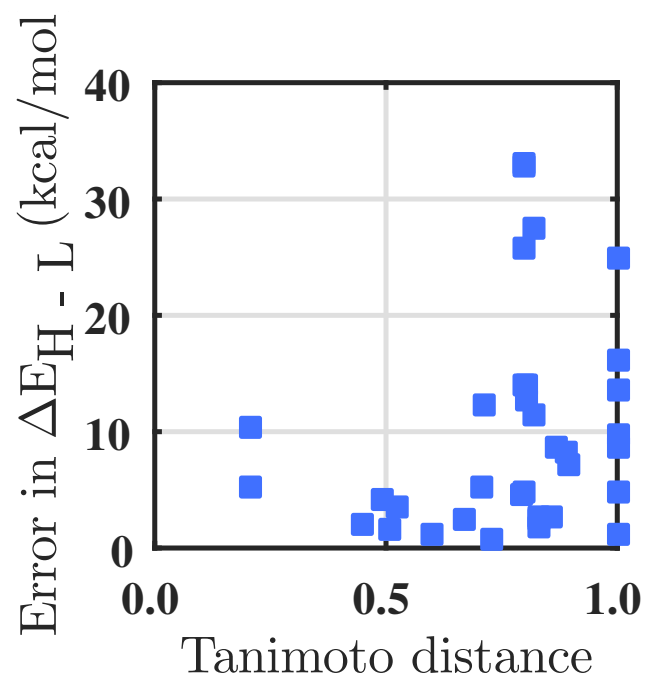


Figure A-29: Dissimilarity metrics for CSD data: errors in spin energy predictions for CSD structures are on the y-axis in kcal/mol and the minimum Tanimoto/FP2 dissimilarity metric (1 - the Tanimoto index) between the CSD ligands and the training ligands is shown on the x-axis. A value of 1 indicates no matches with the FP2 fingerprint.

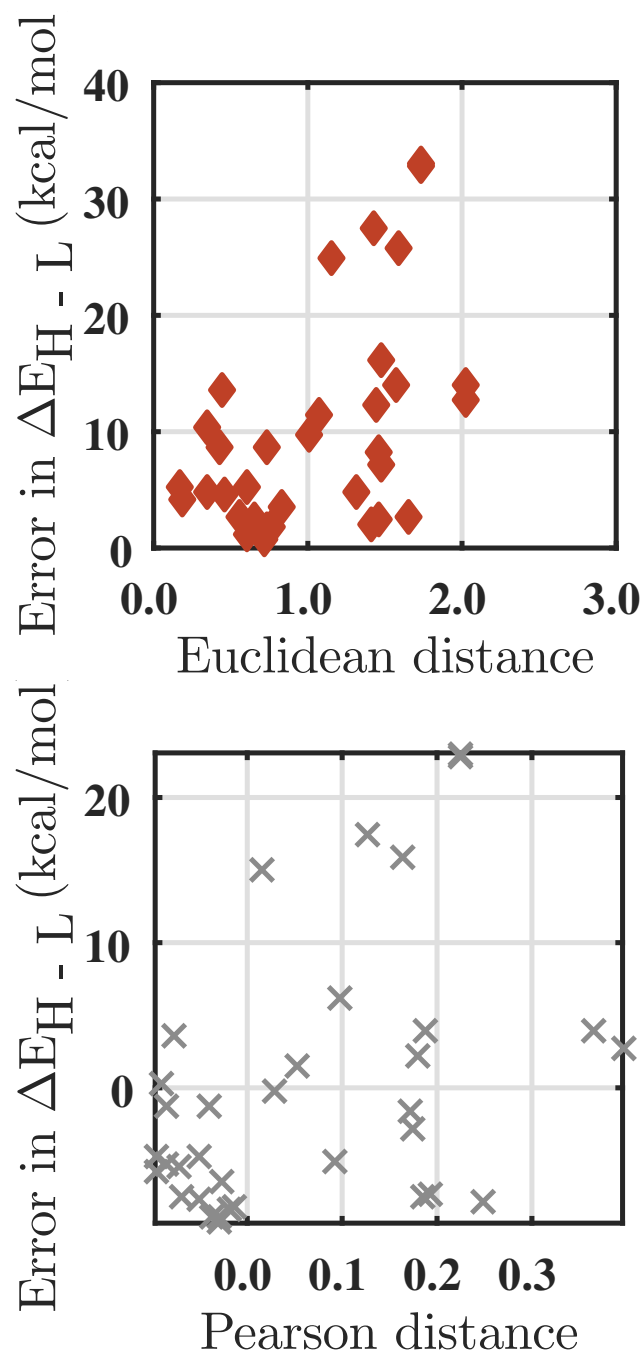


Figure A-30: Comparison of dissimilarity metrics for CSD data: errors in spin energy predictions for CSD structures are on the y-axis in kcal/mol and the Euclidean (left, red) and uncentered Pearson distances (gray, right) between the CSD structure and its nearest representation in dimensionless descriptor space is shown on the x-axis.

Table A.28: LS bond distance predictions and data in Å, for the CSD test structures. For each complex, values from DFT using B3LYP and ANN predictions are shown, along with the standard deviation of the ANN model.

Test number	ANN	DFT/B3LYP	ANN Std. Dev.	Error
testno	nn	data	variance	error
1	2.01	1.95	0.02	0.06
2	2.01	1.90	0.02	0.11
3	1.97	2.06	0.03	-0.10
4	1.98	1.95	0.03	0.04
5	1.98	2.00	0.03	-0.02
6	1.99	1.97	0.03	0.03
7	1.97	1.96	0.02	0.01
8	2.02	2.07	0.03	-0.05
9	1.99	2.03	0.03	-0.03
10	2.02	2.03	0.03	-0.01
11	1.89	1.99	0.03	-0.10
12	2.03	2.04	0.02	-0.02
13	1.91	1.89	0.03	0.02
14	2.02	1.95	0.03	0.07
15	2.04	2.08	0.03	-0.05
16	1.99	1.91	0.03	0.09
17	2.04	2.01	0.03	0.03
18	1.94	2.00	0.02	-0.06
19	1.90	1.88	0.03	0.02
20	2.08	2.05	0.03	0.03
21	2.08	2.04	0.02	0.03
22	2.05	1.98	0.02	0.07
23	1.94	1.91	0.03	0.03
24	2.05	2.03	0.02	0.02
25	1.98	1.99	0.03	-0.01
26	2.00	1.94	0.03	0.06
27	2.04	1.97	0.02	0.06
28	2.01	2.08	0.02	-0.07
29	1.99	2.04	0.3	-0.05
30	1.88	2.04	0.03	-0.17
31	2.10	2.09	0.02	0.01
32	1.99	2.02	0.03	-0.03
33	2.03	2.04	0.02	-0.01
34	2.08	2.01	0.02	0.07
35	2.16	1.98	0.02	0.19

Table A.29: HS bond distance predictions and data in Å, for the CSD test structures. For each complex, values from DFT using B3LYP and ANN predictions are shown, along with the standard deviation of the ANN model.

Test Number	ANN	DFT/B3LYP	ANN Std. Dev.	Error
1	2.06	1.95	0.02	0.11
2	2.06	1.87	0.02	0.16
3	1.94	2.07	0.02	-0.12
4	2.02	2.01	0.02	0.016
5	2.05	2.13	0.02	-0.08
6	2.02	2.06	0.02	-0.04
7	2.06	2.03	0.02	0.03
8	2.09	2.32	0.02	-0.23
9	2.00	2.05	0.02	-0.04
10	2.15	2.12	0.03	0.03
11	1.96	2.03	0.02	-0.07
12	1.98	2.06	0.02	-0.08
13	2.02	1.94	0.02	0.09
14	2.00	2.08	0.02	-0.08
15	2.21	2.22	0.02	-0.01
16	2.07	1.91	0.02	0.16
17	2.06	2.01	0.02	0.06
18	2.14	2.13	0.02	0.01
19	2.02	1.98	0.02	0.04
20	2.11	2.19	0.02	-0.08
21	2.23	2.05	0.02	0.187
22	2.09	2.07	0.02	0.02
23	2.12	2.03	0.02	0.10
24	2.24	2.16	0.02	0.08
25	1.99	2.02	0.02	-0.02
26	2.03	1.96	0.02	0.07
27	2.15	2.17	0.02	-0.02
28	2.13	2.26	0.02	-0.13
29	2.03	2.07	0.02	-0.03
30	2.12	2.08	0.02	0.04
31	2.28	2.25	0.02	0.03
32	2.19	2.22	0.02	-0.03
33	1.98	2.06	0.02	-0.09
34	2.09	2.02	0.02	0.07
35	2.21	1.99	0.02	0.22



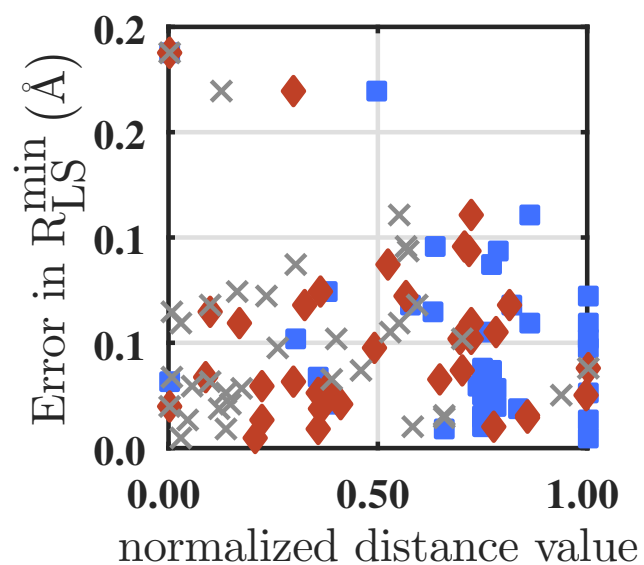


Figure A-31: Comparison of dissimilarity metrics for CSD data: errors in LS bond length prediction for CSD structures are shown on the y-axis in Å, and three normalized dissimilarity metrics are compared on the x-axis: the Tanimoto/FP2 dissimilarity metric between the CSD ligands and the training ligands (blue circles), and the Euclidean (red diamonds) and uncentered Pearson distances (gray crosses) between the CSD structure and its nearest representation in dimensionless descriptor space.

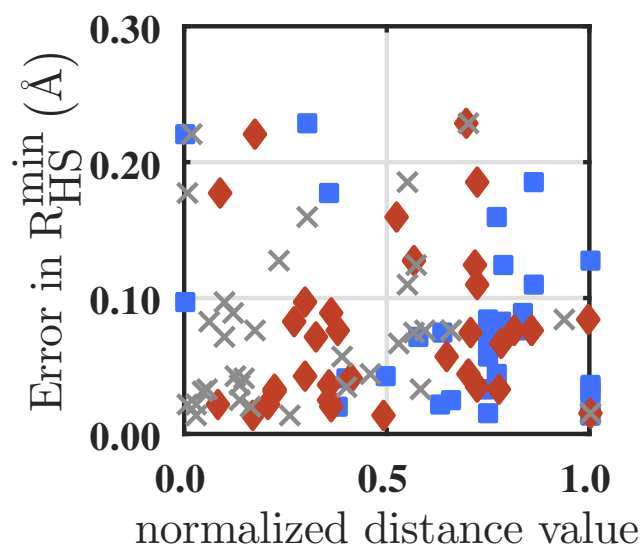


Figure A-32: Comparison of dissimilarity metrics for CSD data: errors in HS bond length prediction for CSD structures are shown on the y-axis in Å, and three normalized dissimilarity metrics are compared on the x-axis: the Tanimoto/FP2 dissimilarity metric between the CSD ligands and the training ligands (blue circles), and the Euclidean (red diamonds) and uncentered Pearson distances (gray crosses) between the CSD structure and its nearest representation in dimensionless descriptor space.

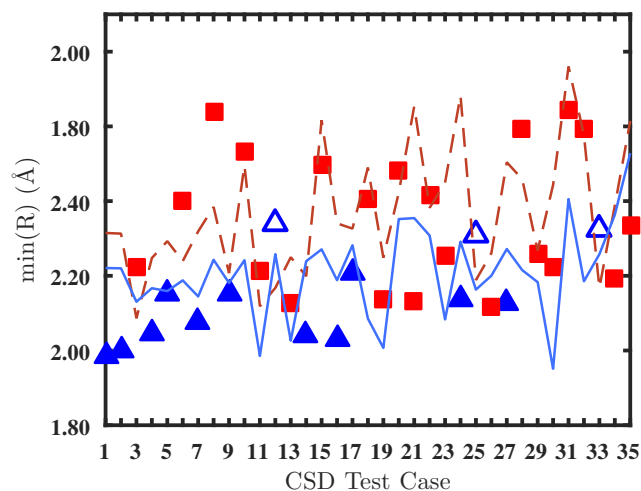


Figure A-33: Comparison of measured CSD bond distances in the crystal phase, represented by symbols (red squares for high-spin or blue triangles for low-spin based on DFT assignment at  $a_{\text{HF}}=0.20$ ) with the ANN predicted HS (red line) and LS (blue line) bond distances.

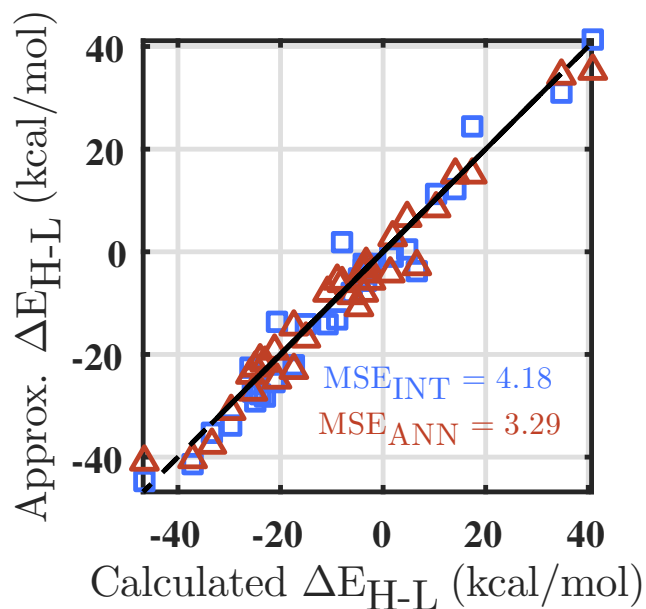


Figure A-34: Parity plot comparing prediction of  $\Delta E_{\text{HS-LS}}$  in kcal/mol. The x-axis is the DFT value at  $a_{\text{HF}} = 0.20$ , while the y-axis is the predicted value. The blue squares are obtained by interpolating from the values at  $a_{\text{HF}} = 0.00$  using the predicted slopes. The red triangles are the matching test cases (where these were tested at  $a_{\text{HF}} = 0.20$ ).



## Appendix B

### Mapping transition metal complex space for machine learning

### Text B.1: Kernel ridge regression

KRR is essentially linear regression in an expanded space of virtual features<sup>209</sup>, which are constructed from non-linear transformations of the original feature space. KRR is feasible, even for infinite-dimensional virtual feature spaces (e.g., Hilbert spaces), as only easy to compute inner products between virtual feature space elements are needed. We use a general exponential kernel form<sup>209</sup>:

$$y_{krr}(\mathbf{x}_j) = \sum_{i=1}^N \alpha_i \exp\left(-\sigma \|\mathbf{x}_j - \mathbf{x}_i\|_l^l\right)$$

Here,  $l = 1$  corresponds to a Laplacian kernel,  $l = 2$  corresponds to a Gaussian kernel, and  $\alpha$  are coefficients that are learned during training by minimizing the square error between model outputs and training points. The inverse correlation length,  $\sigma$ , is a hyperparameter that we estimate using 10-fold cross-validation and a grid search approach in the range of  $1 \times 10^{-8}$  to 0 or expanded as needed if the optimal value is found to lie on the boundaries of this interval. We traverse the interval in 50 logarithmically spaced increments. Use of an isotopic kernel (i.e., fixed  $\sigma$  for all descriptors) simplifies hyperparameter selection during training.  $L_2$  regularization is employed to prevent overfitting, and the regularization parameter is selected along with  $\sigma$  by simultaneous cross-validation.

### Text B.2: Gaussian vs. Laplacian kernels

Using *3d*-ACs, we observed Laplacian kernel errors on QM9 atomization energies of 19.0 kcal/mol training on 1k random molecules and testing on the remainder. The comparable Gaussian kernel value is 8.80 kcal/mol.

Table B.1: Comparison of learning rates for maximum depth three autocorrelation ( $3d$ -AC) and Coulomb eigenspectrum (CM-ES) descriptors tested on QM9 atomization energies (in kcal/mol). Increasing set sizes of randomly selected molecules are selected for training, and the remainder of the 134k molecule set are used for testing. Convergence of KRR hyperparameters  $\lambda$  and  $\sigma$  are also shown. Values are averaged over five training data samples, and one standard deviation on test MUE is given in parenthesis.

Training points	1000	4000	8000	16000
$3d$ -AC				
train MUE	7.80	6.73	5.99	5.81
test MUE	8.80	7.19	6.54	6.10
	(0.42)	(0.14)	(0.09)	(0.07)
$\sigma$	1.00E-06	3.16E-05	1.00E-04	1.00E-04
$\lambda$	1.00E-11	1.00E-11	1.00E-11	1.00E-11
CM-ES				
train MUE	1.60	23.03	20.70	19.05
test MUE	32.19	30.35	27.66	25.17
	(0.61)	(0.37)	(0.12)	(0.11)
$\sigma$	1.00E-006	7.20E-07	1.00E-07	1.00E-07
$\lambda$	1.00E-008	1.00E-08	1.00E-09	1.00E-09

Table B.2: Comparison of learning rates for maximum depth three autocorrelation ( $3d$ -AC) descriptors tested on QM9 dipole moments in Debye. Increasing set sizes of randomly selected molecules are selected for training, and the remainder of the 134k molecule set are used for testing. Convergence of KRR hyperparameters  $\lambda$  and  $\sigma$  are also shown. Values are averaged over five training data samples, and one standard deviation on test MUE is given in parenthesis. Results from Ref. <sup>486</sup> also shown.

Training points	1000	4000	8000	16000
$3d$ -AC				
train MUE	0.84	0.79	0.77	0.77
test MUE	0.88	0.82	0.80	0.79
	(0.015)	(0.004)	(0.003)	(0.003)
$\sigma$	1.00E-06	2.15E-05	5.99E-05	5.99E-05
$\lambda$	1.67E-11	2.78E-10	2.78E-10	2.78E-10
$12^N P3^B 4^B$ <sup>486</sup>				
test MUE	0.86	0.76	0.68	0.63

### Text B.3: Timing for KRR-RFE models

Hyperparameter re-optimization is needed during KRR as the dimension of the feature space is changing. Conducting a  $25 \times 25$  hyperparameter search for training a KRR model using RAC-155 on a workstation (a 4 core Intel Core i7-4820K) takes  $\sim 3.3$  minutes when running in parallel. In order to conduct RFE, it is necessary to conduct 155 such runs to determine the first feature to remove, which takes  $\sim 8$  hours. The number of steps required for full KRR is  $\frac{n(n+1)}{2} \approx 12k$ . While each subsequent step is accelerated by the smaller feature space, the cost of KRR is dominated by the complexity kernel matrix which scales with the (invariant) training space dimension. Therefore, assuming a constant hyperparameter search time, this gives a run time of over  $\sim 27$  days. The solution is to use a more crude hyperparameter search grid and to center to the value of the hyperparameter search around the previous values. Unfortunately, even after experimenting with different grids, we were unable to reduce the search fidelity sufficiently without producing discontinuous jumps in hyperparameter selection. This produces large jumps in the RFE mean CV error. This results in unreliable feature elimination behavior and substantially different KRR performance once the grid fidelity is increased for final training.

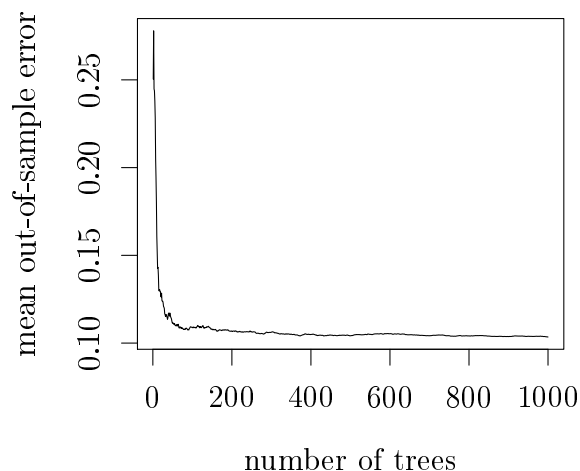


Figure B-1: Out-of-sample mean squared error (MSE) for spin-splitting energy prediction by random forest model on the *spin-splitting* data set in normalized units as a function of number trees in the random forest model.



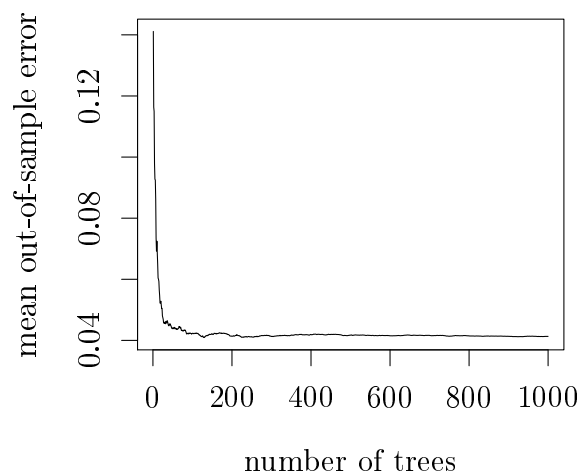


Figure B-2: Out-of-sample mean squared error (MSE) for bond length prediction by random forest model on the *spin-splitting* data set in normalized units as a function of number trees in the random forest model.

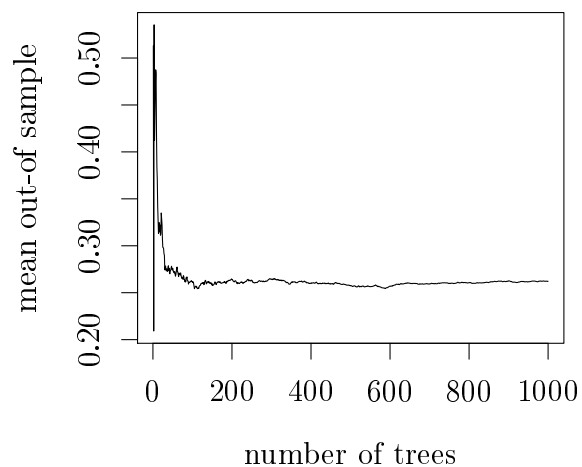


Figure B-3: Out-of-sample mean squared error (MSE) for ionization potential prediction by the random forest model on the *redox* data set in normalized units as a function of number trees in the random forest model.

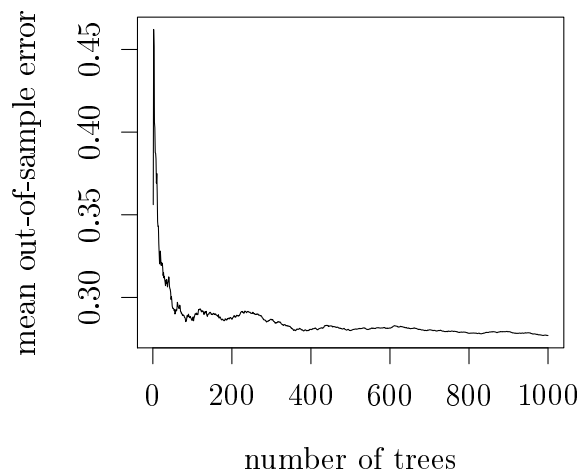


Figure B-4: Out-of-sample mean squared error (MSE) for redox potential prediction by the random forest model on the *redox* data set in normalized units as a function of number trees in the random forest model.

Table B.3: Hyperparameters selected by cross-validation for spin splitting energy prediction on different feature sets, including varying random forest (randF) cut-off values.

	$\sigma$	$\lambda$
RAC-155	1.39E-03	7.20E-12
LASSO-28	5.18E-03	3.73E-10
UV-86	4.64E-02	5.18E-08
randF-41 (1%)	6.11E-02	3.73E-07
randF-26 (2%)	1.10E-02	1.00E-08
randF-34 (1.5%)	7.85E-03	7.74E-09
randF-18 (3%)	7.85E-03	2.78E-09
RFE-43	1.93E-02	1.00E-09
C-12	4.94E-02	2.15E-07

Table B.4: Hyperparameters selected by cross-validation for bond length prediction on different feature sets.

	$\sigma$	$\lambda$
RAC-155	1.93E-02	5.18E-08
LASSO-28	7.20E-02	3.73E-07
randF-41	1.93E-02	1.93E-08
randF-26	1.93E-02	1.93E-08
LASSO-83B	1.93E-02	1.93E-08
randF-48B	7.20E-02	1.00E-06
randF-49B	1.93E-02	5.18E-08
C-12	1.93E-02	3.73E-10

Table B.5: Hyperparameters selected by cross-validation for ionization potential prediction on different feature sets.

	$\sigma$	$\lambda$
RAC-155	7.54E-03	7.20E-05
LASSO-28	4.94E-02	1.53E-05
randF-41	7.54E-03	1.15E-04
randF-26	2.33E-02	2.02E-04
LASSO-19I	3.39E-02	6.25E-04
randF-28I	2.33E-02	6.25E-04
C-12	3.39E-02	7.20E-05

Table B.6: Hyperparameters selected by cross-validation for redox potential prediction on different feature sets.

	$\sigma$	$\lambda$
RAC-155	7.54E-03	4.94E-05
LASSO-28	3.39E-02	2.12E-05
randF-41	1.10E-02	7.54E-05
randF-26	2.33E-02	1.15E-04
randF-38G	3.39E-02	7.54E-05
LASSO-19G	4.94E-02	1.15E-04
C-12	3.73E-04	1.53E-10

Table B.7: Spin splitting data set ligands from previous work<sup>308</sup>.

Number	ID	SMILES
1	cl	[Cl-]
2	scn	[S-]C#N
3	pisc	CC(C)(C)C1=CC=C(C=C1)[N+]#[C-]
4	misc	C[N+]#[C-]
5	cn	[C-]#N
6	co	CO
7	ncs	[N-]=C=S
9	bipy	C1ccnc(c1)c2cccn2
10	phen	C1=CC2=CC=C3C=CC=NC3=C2N=C1
11	en	NCCN
12	porphyrin	[NH]1C2=CC3=NC(=CC4=CC=C([NH]4)C=C5C=CC(=N5)C=C1C=C2)C=C3
13	h2o	O
14	acac	CC(=O)CC(=O)C
15	tbuc	CC(C)(C)C1=CC(=C([O-])C=C1)[O-]
16	ox	[O-]C(=O)C([O-])=O

Table B.8: Redox data set ligands from previous work<sup>482</sup>.

Number	SMILES
1	n1c(noc1N)c1ncccc1
2	c1(nn[nH]n1)c1ncccc1
3	c1(nn[nH]c1)c1ncccc1
4	n1c(nccc1N)c1ncccc1
5	c1(nn[nH]c1)c1ncc(cc1)C
6	n1c([nH]nc1N)c1ncccc1
7	[nH]1nc2c(c1)nc(n2)c1ncccc1
8	c1(nn[nH]c1)c1ncccc1C
9	c1(nn[nH]c1)c1nc(ccc1)C
10	n1c2c3ncccc3ccc2ccc1
11	o1c(=O)[nH]nc1c1cccc1
12	o1nc(c2ncccc2)cc1CO
13	n1c(nc(nc1N)N)c1ncccc1
14	O=C1c2c(nccc2)c2ncccc12
15	Clc1nc2c3ncccc3ccc2cc1
16	o1nc(nc1c1oncc1)c1ncccc1
17	Oc1nc(c2nc(O)ccc2)ccc1
18	[O-][N+](=O)c1nc([nH]c1[N+](=O)[O-])c1ncccc1
19	o1c(nc(c1N)C#N)c1ncccc1
20	c1(nc(nc(c1)N)c1ncccc1)O
21	c1(cccc(n1)Cl)c1n[nH]cn1
22	c1(ccccn1)c1n[nH]cn1
23	c1(cccc(n1)C)c1n[nH]cn1
24	c1(ccnc(n1)N)c1ncccc1
25	C(=O)(CN=[N+]=[N-])Nc1ccc(cc1)F
26	ONC(=N)[C](N)C
27	C(CN)N
28	N[C](C(=O)NO)Cc1[nH]enc1
29	[nH]1c(=O)[nH]cc(c1=O)NC(=O)CN
30	C(=C)(C(=O)NC)NC(=O)C
31	C(C(C(=O)O)N)NC(=O)OC
32	o1ence1
33	n1oc2c(c1)cccc2
34	O=C=NC
35	NC
36	C(#N)O
37	[O-][N+]#N
38	N#C
39	N(=C)O
40	N#CC=CCC
41	ClCCC#N

Table B.9: Redox/ionization spin multiplicities. The M(II) ground state determines the selected M(III) spin state for comparison.

	M(II) ground state	M(II) spin	M(III) spin
Cr	LS	3	2
	HS	5	4
Mn	LS	2	3
	HS	6	5
Fe	LS	1	2
	HS	5	6
Co	LS	2	1
	HS	4	5

Table B.10: Features included in the RAC-155 descriptor set.

1	ox	41	$f_{eq}S_0$	81	$lc_{eq}S_2$	121	$lc_{ax}\chi'_2$
2	$\alpha_{HF}$	42	$f_{eq}S_1$	82	$lc_{eq}S_3$	122	$lc_{ax}\chi'_3$
3	ax. denticity	43	$f_{eq}S_2$	83	$mc_{all}\chi_0$	123	$lc_{ax}Z'_1$
4	eq. denticity	44	$f_{eq}S_3$	84	$mc_{all}\chi_1$	124	$lc_{ax}Z'_2$
5	$f_{ax}\chi_0$	45	$lc_{ax}\chi_0$	85	$mc_{all}\chi_2$	125	$lc_{ax}Z'_3$
6	$f_{ax}\chi_1$	46	$lc_{ax}\chi_1$	86	$mc_{all}\chi_3$	126	$lc_{ax}T'_1$
7	$f_{ax}\chi_2$	47	$lc_{ax}\chi_2$	87	$mc_{all}Z_0$	127	$lc_{ax}T'_2$
8	$f_{ax}\chi_3$	48	$lc_{ax}\chi_3$	88	$mc_{all}Z_1$	128	$lc_{ax}T'_3$
9	$f_{ax}Z_0$	49	$lc_{ax}Z_0$	89	$mc_{all}Z_2$	129	$lc_{ax}S'_1$
10	$f_{ax}Z_1$	50	$lc_{ax}Z_1$	90	$mc_{all}Z_3$	130	$lc_{ax}S'_2$
11	$f_{ax}Z_2$	51	$lc_{ax}Z_2$	91	$mc_{all}I_2$	131	$lc_{ax}S'_3$
12	$f_{ax}Z_3$	52	$lc_{ax}Z_3$	92	$mc_{all}I_3$	132	$lc_{eq}\chi'_1$
13	$f_{ax}I_0$	53	$lc_{ax}I_1$	93	$mc_{all}T_1$	133	$lc_{eq}\chi'_2$
14	$f_{ax}I_1$	54	$lc_{ax}I_2$	94	$mc_{all}T_2$	134	$lc_{eq}\chi'_3$
15	$f_{ax}I_2$	55	$lc_{ax}I_3$	95	$mc_{all}T_3$	135	$lc_{eq}Z'_1$
16	$f_{ax}I_3$	56	$lc_{ax}T_0$	96	$mc_{all}S_0$	136	$lc_{eq}Z'_2$
17	$f_{ax}T_0$	57	$lc_{ax}T_1$	97	$mc_{all}S_1$	137	$lc_{eq}Z'_3$
18	$f_{ax}T_1$	58	$lc_{ax}T_2$	98	$mc_{all}S_2$	138	$lc_{eq}T'_1$
19	$f_{ax}T_2$	59	$lc_{ax}T_3$	99	$mc_{all}S_3$	139	$lc_{eq}T'_2$
20	$f_{ax}T_3$	60	$lc_{ax}S_0$	100	$f_{all}\chi_0$	140	$lc_{eq}T'_3$
21	$f_{ax}S_0$	61	$lc_{ax}S_1$	101	$f_{all}\chi_1$	141	$lc_{eq}S'_1$
22	$f_{ax}S_1$	62	$lc_{ax}S_2$	102	$f_{all}\chi_2$	142	$lc_{eq}S'_2$
23	$f_{ax}S_2$	63	$lc_{ax}S_3$	103	$f_{all}\chi_3$	143	$lc_{eq}S'_3$
24	$f_{ax}S_3$	64	$lc_{eq}\chi_0$	104	$f_{all}Z_0$	144	$mc_{all}\chi'_1$
25	$f_{eq}\chi_0$	65	$lc_{eq}\chi_1$	105	$f_{all}Z_1$	145	$mc_{all}\chi'_2$
26	$f_{eq}\chi_1$	66	$lc_{eq}\chi_2$	106	$f_{all}Z_2$	146	$mc_{all}\chi'_3$
27	$f_{eq}\chi_2$	67	$lc_{eq}\chi_3$	107	$f_{all}Z_3$	147	$mc_{all}Z'_1$
28	$f_{eq}\chi_3$	68	$lc_{eq}Z_0$	108	$f_{all}I_0$	148	$mc_{all}Z'_2$
29	$f_{eq}Z_0$	69	$lc_{eq}Z_1$	109	$f_{all}I_1$	149	$mc_{all}Z'_3$
30	$f_{eq}Z_1$	70	$lc_{eq}Z_2$	110	$f_{all}I_2$	150	$mc_{all}T'_1$
31	$f_{eq}Z_2$	71	$lc_{eq}Z_3$	111	$f_{all}I_3$	151	$mc_{all}T'_2$
32	$f_{eq}Z_3$	72	$lc_{eq}I_1$	112	$f_{all}T_0$	152	$mc_{all}T'_3$
33	$f_{eq}I_0$	73	$lc_{eq}I_2$	113	$f_{all}T_1$	153	$mc_{all}S'_1$
34	$f_{eq}I_1$	74	$lc_{eq}I_3$	114	$f_{all}T_2$	154	$mc_{all}S'_2$
35	$f_{eq}I_2$	75	$lc_{eq}T_0$	115	$f_{all}T_3$	155	$mc_{all}S'_3$
36	$f_{eq}I_3$	76	$lc_{eq}T_1$	116	$f_{all}S_0$		
37	$f_{eq}T_0$	77	$lc_{eq}T_2$	117	$f_{all}S_1$		
38	$f_{eq}T_1$	78	$lc_{eq}T_3$	118	$f_{all}S_2$		
39	$f_{eq}T_2$	79	$lc_{eq}S_0$	119	$f_{all}S_3$		
40	$f_{eq}T_3$	80	$lc_{eq}S_1$	120	$lc_{ax}\chi'_1$		

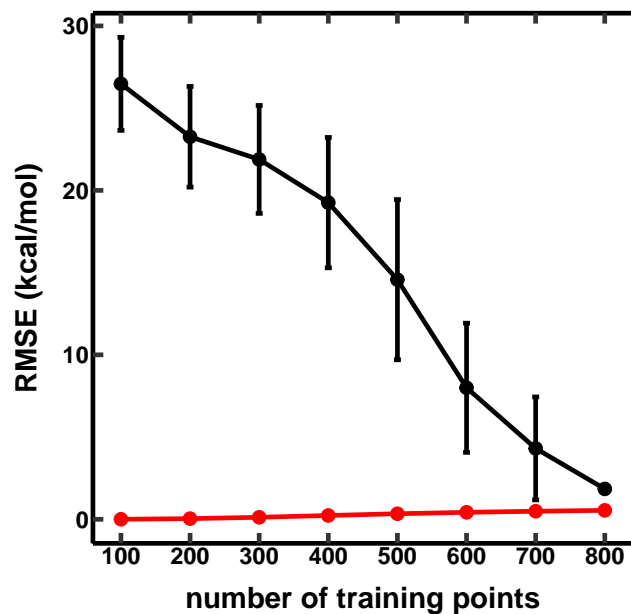


Figure B-5: Learning rates for KRR models trained on randomly drawn samples of 60% of of dataset 1 and tested on 40% remaining points. The points show mean spin splitting energy test (black) and train (red) RMSE values (in kcal/mol) from 10 samples and the confidence intervals indicate one standard deviation for test RMSE at different training set sizes using RAC-155.

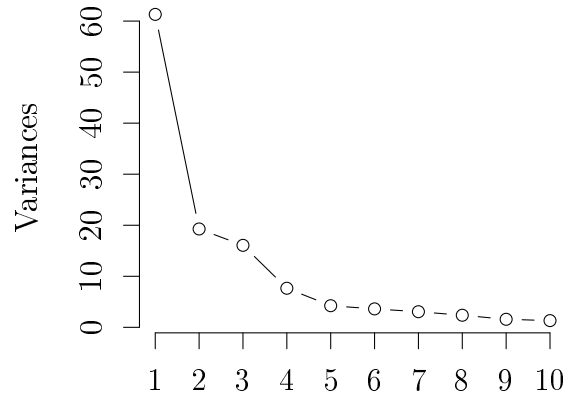


Figure B-6: PCA variance die-off with RAC-155 principal components (arb. units).

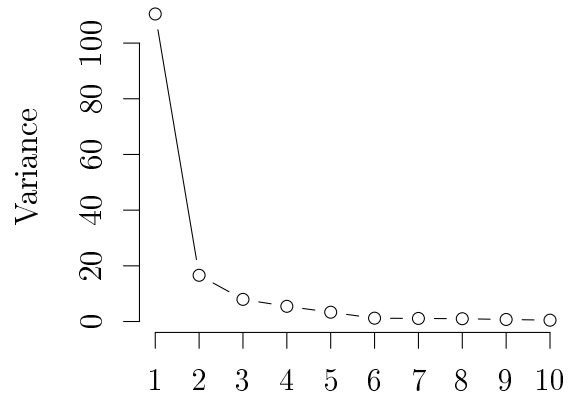


Figure B-7: PCA variance die-off with CM-ES principal components (arb. units).



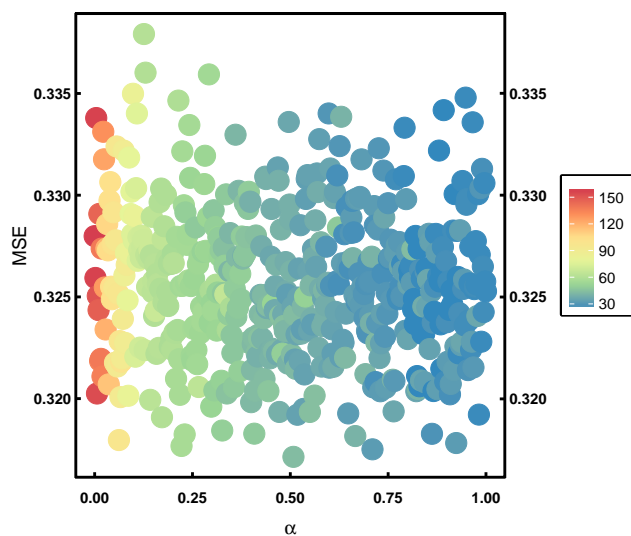


Figure B-8: Error response of elastic net model on spin splitting energy with varying  $\alpha$  parameter from 0 (ridge regression) to 1 (LASSO). Colors indicate the number of selected variables at each value.

Table B.11: Features included in the LASSO-28 descriptor set.

1	ox	11	$lc_{eq}Z_1$	21	$lc_{ax}Z'_3$
2	$\alpha_{HF}$	12	$mc_{all}\chi_0$	22	$lc_{ax}T'_2$
3	eq. denticity	13	$mc_{all}\chi_2$	23	$lc_{ax}S'_3$
4	$f_{ax}Z_0$	14	$mc_{all}Z_0$	24	$lc_{eq}\chi'_2$
5	$f_{ax}I_3$	15	$mc_{all}Z_1$	25	$mc_{all}\chi'_1$
6	$f_{ax}T_3$	16	$mc_{all}Z_3$	26	$mc_{all}\chi'_2$
7	$f_{eq}Z_2$	17	$mc_{all}S_0$	27	$mc_{all}\chi'_3$
8	$lc_{ax}\chi_0$	18	$f_{all}\chi_2$	28	$mc_{all}S'_1$
9	$lc_{ax}Z_2$	19	$lc_{ax}\chi'_3$		
10	$lc_{ax}T_3$	20	$lc_{ax}Z'_1$		

Table B.12: Features included in the UV-86 descriptor set.

1	ox	23	$f_{eq}S_2$	45	$mc_{all}T_1$	67	$f_{all}S_3$
2	$\alpha_{HF}$	24	$f_{eq}S_3$	46	$mc_{all}T_3$	68	$lc_{ax}\chi'_1$
3	ax. denticity	25	$lc_{ax}\chi_0$	47	$mc_{all}S_0$	69	$lc_{ax}\chi'_2$
4	eq. denticity	26	$lc_{ax}\chi_1$	48	$f_{all}\chi_0$	70	$lc_{ax}Z'_1$
5	$f_{eq}\chi_0$	27	$lc_{ax}Z_0$	49	$f_{all}\chi_1$	71	$lc_{ax}T'_1$
6	$f_{eq}\chi_1$	28	$lc_{ax}Z_1$	50	$f_{all}\chi_2$	72	$lc_{ax}S'_1$
7	$f_{eq}\chi_2$	29	$lc_{ax}Z_2$	51	$f_{all}\chi_3$	73	$lc_{eq}\chi'_1$
8	$f_{eq}\chi_3$	30	$lc_{ax}I_1$	52	$f_{all}Z_0$	74	$lc_{eq}\chi'_2$
9	$f_{eq}Z_0$	31	$lc_{ax}T_0$	53	$f_{all}Z_1$	75	$lc_{eq}\chi'_3$
10	$f_{eq}Z_1$	32	$lc_{eq}\chi_0$	54	$f_{all}Z_2$	76	$lc_{eq}Z'_1$
11	$f_{eq}Z_2$	33	$lc_{eq}\chi_1$	55	$f_{all}Z_3$	77	$lc_{eq}T'_3$
12	$f_{eq}Z_3$	34	$lc_{eq}Z_0$	56	$f_{all}I_0$	78	$lc_{eq}S'_1$
13	$f_{eq}I_0$	35	$lc_{eq}Z_1$	57	$f_{all}I_1$	79	$mc_{all}\chi'_1$
14	$f_{eq}I_1$	36	$lc_{eq}Z_2$	58	$f_{all}I_2$	80	$mc_{all}\chi'_2$
15	$f_{eq}I_2$	37	$lc_{eq}I_1$	59	$f_{all}I_3$	81	$mc_{all}Z'_1$
16	$f_{eq}I_3$	38	$lc_{eq}T_3$	60	$f_{all}T_0$	82	$mc_{all}Z'_2$
17	$f_{eq}T_0$	39	$mc_{all}\chi_0$	61	$f_{all}T_1$	83	$mc_{all}T'_1$
18	$f_{eq}T_1$	40	$mc_{all}\chi_1$	62	$f_{all}T_2$	84	$mc_{all}T'_2$
19	$f_{eq}T_2$	41	$mc_{all}Z_0$	63	$f_{all}T_3$	85	$mc_{all}S'_1$
20	$f_{eq}T_3$	42	$mc_{all}Z_1$	64	$f_{all}S_0$	86	$mc_{all}S'_2$
21	$f_{eq}S_0$	43	$mc_{all}Z_2$	65	$f_{all}S_1$		
22	$f_{eq}S_1$	44	$mc_{all}I_2$	66	$f_{all}S_2$		

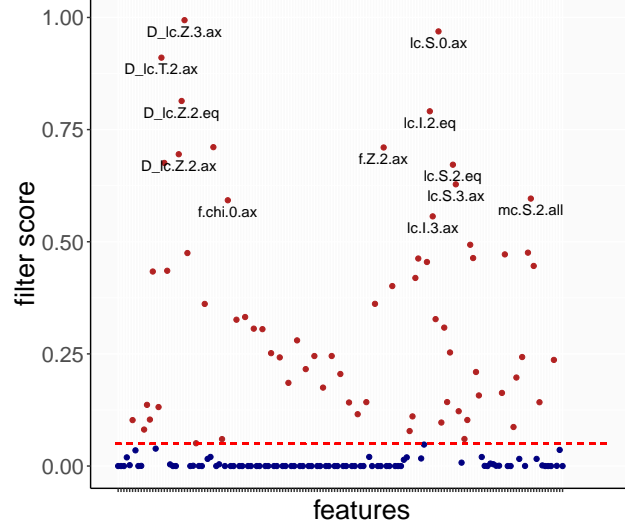


Figure B-9: Univariate filter scores for spin splitting energy. The 5% cutoff is indicated by a dashed red line, and rejected points are colored red. Some features with a score of more than 50% are labeled.

Table B.13: Features included in the RFE-43 descriptor set.

1	ox	12	$f_{eq}S_2$	23	$lc_{eq}S_0$	34	$f_{all}Z_0$
2	$\alpha_{HF}$	13	$f_{eq}S_3$	24	$lc_{eq}S_1$	35	$f_{all}I_2$
3	ax. denticity	14	$lc_{ax}\chi_0$	25	$mc_{all}\chi_1$	36	$f_{all}T_0$
4	eq. denticity	15	$lc_{ax}I_2$	26	$mc_{all}Z_1$	37	$f_{all}S_1$
5	$f_{ax}Z_2$	16	$lc_{ax}T_0$	27	$mc_{all}Z_3$	38	$f_{all}S_2$
6	$f_{ax}I_1$	17	$lc_{ax}S_0$	28	$mc_{all}T_3$	39	$lc_{ax}\chi'_1$
7	$f_{ax}I_2$	18	$lc_{ax}S_1$	29	$mc_{all}S_2$	40	$lc_{ax}S'_2$
8	$f_{ax}T_2$	19	$lc_{ax}S_2$	30	$mc_{all}S_3$	41	$lc_{eq}T'_2$
9	$f_{eq}\chi_0$	20	$lc_{ax}S_3$	31	$f_{all}\chi_0$	42	$lc_{eq}T'_3$
10	$f_{eq}\chi_2$	21	$lc_{eq}\chi_0$	32	$f_{all}\chi_1$	43	$mc_{all}Z'_1$
11	$f_{eq}T_0$	22	$lc_{eq}I_3$	33	$f_{all}\chi_3$		

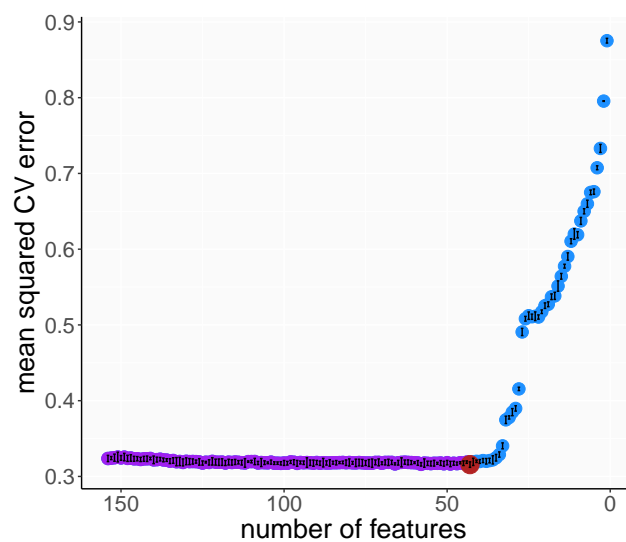


Figure B-10: RFE mean-CV error in normalized units with number of features retained with inner MLR model for spin splitting energy. Results are the mean value averaged over five repeats, and one standard deviation is shown with an error bar. The global minimum at 43 features is highlighted with a red circle.

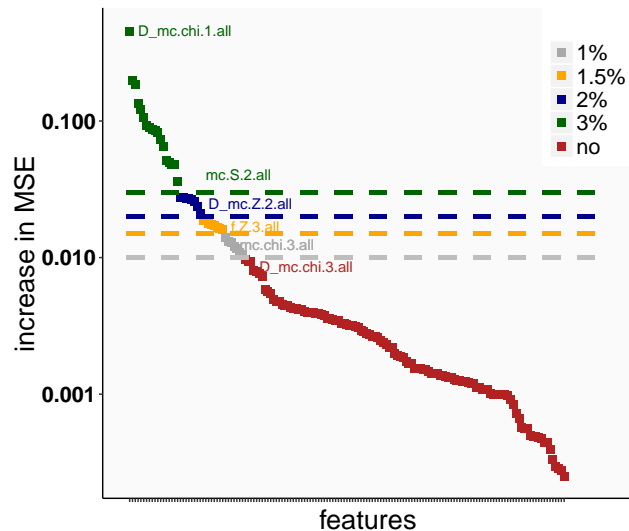


Figure B-11: Out-of-sample mean squared error (MSE) increase for spin splitting energy prediction by random forest model on the *spin-splitting* data set in normalized units. Lines indicate 3% (green), 2% (blue), 1.5% (yellow) 1% (gray) cutoff values, along with the resulting included variables that are indicated in the main text as randF-26 (green + blue squares) and randF-41 (green + blue + yellow + gray squares). Never-selected points are colored red. Some representative features are labeled.

Table B.14: Features included in the randF-41 descriptor set.

1	$mc_{all}\chi'_1$	12	$lc_{eq}Z'_1$	23	$f_{all}\chi_3$	34	$mc_{all}S_3$
2	ox	13	$lc_{eq}\chi'_1$	24	$f_{all}S_0$	35	$mc_{all}\chi_3$
3	$mc_{all}Z'_1$	14	$lc_{eq}Z_0$	25	$mc_{all}\chi_2$	36	$lc_{ax}Z_0$
4	$mc_{all}Z_1$	15	$lc_{eq}\chi_0$	26	$f_{all}Z_1$	37	$mc_{all}Z'_3$
5	$mc_{all}S'_1$	16	$mc_{all}\chi_1$	27	$f_{all}Z_3$	38	$lc_{ax}\chi_0$
6	$mc_{all}S_1$	17	$mc_{all}\chi'_2$	28	$f_{all}Z_2$	39	$mc_{all}S'_3$
7	$mc_{all}Z_0$	18	$mc_{all}S_2$	29	$f_{all}\chi_0$	40	$f_{all}I_3$
8	$mc_{all}Z_2$	19	$mc_{all}Z'_2$	30	$f_{all}Z_0$	41	$lc_{ax}Z'_1$
9	$mc_{all}S_0$	20	$f_{all}S_1$	31	$f_{all}\chi_1$		
10	$mc_{all}\chi_0$	21	$f_{all}S_2$	32	$f_{all}S_3$		
11	$\alpha_{HF}$	22	$mc_{all}S'_2$	33	$f_{all}\chi_2$		

Table B.15: Features included in the randF-26 descriptor set.

1	$^{mc}_{all}\chi'_1$	11	$\alpha_{HF}$	21	$^f_{all}S_2$
2	ox	12	$^{lc}_{eq}Z'_1$	22	$^{mc}_{all}S'_2$
3	$^{mc}_{all}Z'_1$	13	$^{lc}_{eq}\chi'_1$	23	$^f_{all}\chi_3$
4	$^{mc}_{all}Z_1$	14	$^{lc}_{eq}Z_0$	24	$^f_{all}S_0$
5	$^{mc}_{all}S'_1$	15	$^{lc}_{eq}\chi_0$	25	$^{mc}_{all}\chi_2$
6	$^{mc}_{all}S_1$	16	$^{mc}_{all}\chi_1$	26	$^f_{all}Z_1$
7	$^{mc}_{all}Z_0$	17	$^{mc}_{all}\chi'_2$		
8	$^{mc}_{all}Z_2$	18	$^{mc}_{all}S_2$		
9	$^{mc}_{all}S_0$	19	$^{mc}_{all}Z'_2$		
10	$^{mc}_{all}\chi_0$	20	$^f_{all}S_1$		

Table B.16: KRR error metrics for different random forest importance cutoffs on spin splitting energy in kcal/mol, showing test mean unsigned (MUE) and test and train root mean square (RMSE) errors.

cutoff		dim	train RMSE	test RMSE	test MUE
1.0%	randF-41	41	0.40	1.87	1.01
1.5%	randF-34	34	1.19	2.15	1.30
2.0%	randF-26	26	1.18	2.12	1.28
3.0%	randF-18	18	4.68	5.80	3.10

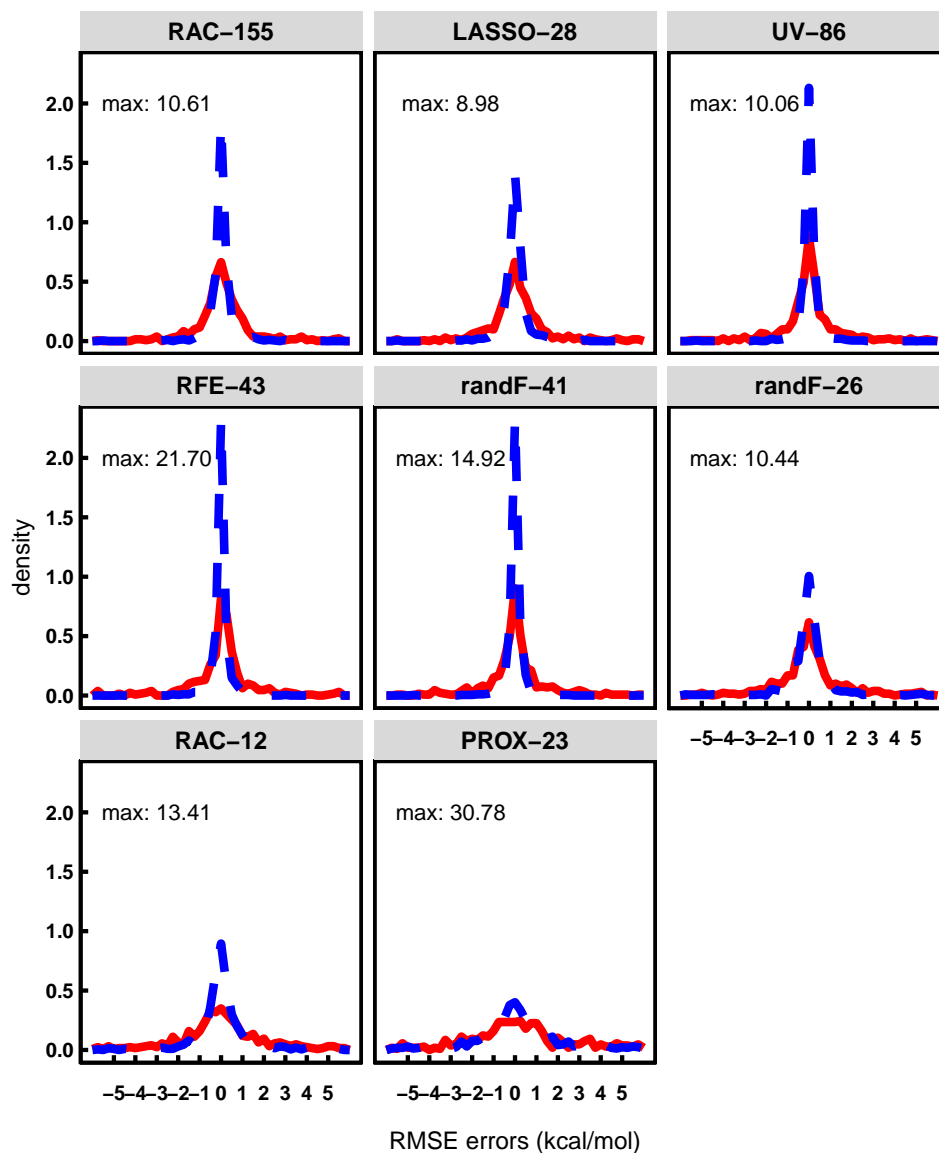


Figure B-12: Test and train KRR error distributions in kcal/mol for predicting spin splitting energy with different feature sets, indicating the maximum absolute test error for each feature set. Blue dashed lines indicate training errors and red lines indicate test errors.

Table B.17: Features included in the common set of all descriptors selected by the feature selection methods, RAC-12.

1	ox	7	$\frac{mc}{all}S_0$
2	$\alpha_{HF}$	8	$\frac{f}{all}\chi_2$
3	$\frac{lc}{ax}\chi_0$	9	$\frac{lc}{ax}Z'_1$
4	$\frac{mc}{all}\chi_0$	10	$\frac{mc}{all}\chi'_1$
5	$\frac{mc}{all}Z_0$	11	$\frac{mc}{all}\chi'_2$
6	$\frac{mc}{all}Z_1$	12	$\frac{mc}{all}S'_1$

Table B.18: Features in the proximal-only PROX-23 set.

1	ox	9	$\frac{lc}{ax}T_0$	17	$\frac{mc}{all}T_1$
2	$\alpha_{HF}$	10	$\frac{lc}{eq}T_0$	18	$\frac{mc}{all}S_0$
3	ax. denticity	11	$\frac{lc}{ax}S_0$	19	$\frac{mc}{all}S_1$
4	eq. denticity	12	$\frac{lc}{eq}S_0$	20	$\frac{mc}{all}T'_1$
5	$\frac{lc}{ax}\chi_0$	13	$\frac{mc}{all}\chi_0$	21	$\frac{mc}{all}\chi'_1$
6	$\frac{lc}{eq}\chi_0$	14	$\frac{mc}{all}\chi_1$	22	$\frac{mc}{all}Z'_1$
7	$\frac{lc}{ax}Z_0$	15	$\frac{mc}{all}Z_0$	23	$\frac{mc}{all}S'_1$
8	$\frac{lc}{eq}Z_0$	16	$\frac{mc}{all}Z_1$		



Table B.19: Features in the bond-length-selected LASSO-83B descriptor set.

1	ox	22	$lc_{ax}S_1$	43	$f_{all}chi_2$	64	$lc_{ax}T'_3$
2	$\alpha_{HF}$	23	$lc_{eq}\chi_0$	44	$f_{all}\chi_3$	65	$lc_{ax}S'_3$
3	ax. denticity	24	$lc_{eq}\chi_1$	45	$f_{all}Z_0$	66	$lc_{eq}\chi'_1$
4	eq. denticity	25	$lc_{eq}Z_0$	46	$f_{all}Z_1$	67	$lc_{eq}\chi'_3$
5	$f_{ax}Z_0$	26	$lc_{eq}Z_1$	47	$f_{all}Z_2$	68	$lc_{eq}Z'_2$
6	$f_{ax}Z_2$	27	$lc_{eq}Z_2$	48	$f_{all}Z_3$	69	$lc_{eq}T'_2$
7	$f_{ax}I_0$	28	$lc_{eq}I_1$	49	$f_{all}I_0$	70	$lc_{eq}T'_3$
8	$f_{ax}I_3$	29	$lc_{eq}T_0$	50	$f_{all}I_1$	71	$lc_{eq}S'_1$
9	$f_{ax}T_3$	30	$lc_{eq}T_3$	51	$f_{all}I_2$	72	$lc_{eq}S'_2$
10	$f_{eq}\chi_0$	31	$lc_{eq}S_1$	52	$f_{all}I_3$	73	$lc_{eq}S'_3$
11	$f_{eq}\chi_1$	32	$lc_{eq}S_2$	53	$f_{all}T_2$	74	$mc_{all}\chi'_2$
12	$f_{eq}\chi_2$	33	$mc_{all}\chi_0$	54	$f_{all}T_3$	75	$mc_{all}\chi'_3$
13	$f_{eq}\chi_3$	34	$mc_{all}\chi_1$	55	$f_{all}S_0$	76	$mc_{all}Z'_1$
14	$f_{eq}Z_1$	35	$mc_{all}\chi_2$	56	$f_{all}S_1$	77	$mc_{all}Z'_2$
15	$f_{eq}Z_2$	36	$mc_{all}\chi_3$	57	$f_{all}S_2$	78	$mc_{all}Z'_3$
16	$f_{eq}Z_3$	37	$mc_{all}Z_1$	58	$f_{all}S_3$	79	$mc_{all}T'_1$
17	$lc_{ax}\chi_0$	38	$mc_{all}Z_2$	59	$lc_{ax}\chi'_1$	80	$mc_{all}T'_2$
18	$lc_{ax}Z_0$	39	$mc_{all}Z_3$	60	$lc_{ax}\chi'_2$	81	$mc_{all}S'_1$
19	$lc_{ax}Z_2$	40	$mc_{all}S_0$	61	$lc_{ax}\chi'_3$	82	$mc_{all}S'_2$
20	$lc_{ax}T_0$	41	$mc_{all}S_2$	62	$lc_{ax}Z'_1$	83	$mc_{all}S'_3$
21	$lc_{ax}T_3$	42	$f_{all}\chi_0$	63	$lc_{ax}T'_2$		

Table B.20: Features in the randF-49B descriptor set, which was obtained by adding  $\alpha_{HF}$  to the randF-48B set obtained from random forest applied to bond length data in the *spin-splitting* data set, as described in the main text.

1	$^{mc}_{all}Z'_1$	14	$^f_{all}\chi_3$	27	$^f_{all}\chi_0$	40	$^f_{all}Z_2$
2	$^{mc}_{all}S'_1$	15	$^{mc}_{all}T'_1$	28	$^{mc}_{all}T_1$	41	$^{lc}_{eq}S_1$
3	$^{mc}_{all}Z_1$	16	$^{mc}_{all}\chi_2$	29	$^f_{all}S_0$	42	$^{mc}_{all}\chi'_3$
4	$^{mc}_{all}Z_0$	17	$^{mc}_{all}Z'_2$	30	$^{mc}_{all}\chi_1$	43	$^f_{all}S_3$
5	$^{mc}_{all}S_1$	18	$^f_{all}Z_0$	31	$^{lc}_{ax}T_1$	44	$^{lc}_{ax}\chi_1$
6	$^{mc}_{all}S'_2$	19	$^{lc}_{ax}\chi'_1$	32	$^{lc}_{eq}S_0$	45	$^f_{all}S_2$
7	$^{mc}_{all}\chi_0$	20	$^f_{all}\chi_1$	33	$^{lc}_{eq}\chi_0$	46	$^{mc}_{all}Z_3$
8	OX	21	$^{mc}_{all}Z_2$	34	$^{lc}_{ax}T_0$	47	$^{lc}_{ax}S'_1$
9	$^{mc}_{all}S_0$	22	$^f_{all}Z_1$	35	$^f_{all}\chi_2$	48	$^{lc}_{ax}Z_0$
10	$^{mc}_{all}S_2$	23	$^f_{all}Z_3$	36	$^f_{all}S_1$	49	$\alpha_{HF}$
11	$^{mc}_{all}\chi'_2$	24	$^{lc}_{eq}Z_0$	37	$^{lc}_{ax}I_1$		
12	$^{mc}_{all}T'_2$	25	$^{mc}_{all}I_2$	38	$^{lc}_{eq}S'_2$		
13	$^{lc}_{ax}Z'_1$	26	$^{mc}_{all}\chi'_1$	39	$^{mc}_{all}S_3$		

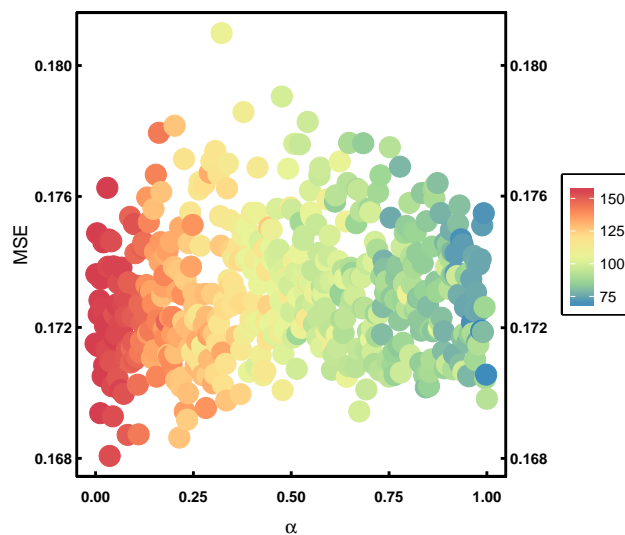


Figure B-13: Error response of elastic net model on bond lengths with varying  $\alpha$  parameter from 0 (ridge regression) to 1 (LASSO). Colors indicate the number of selected variables at each value.

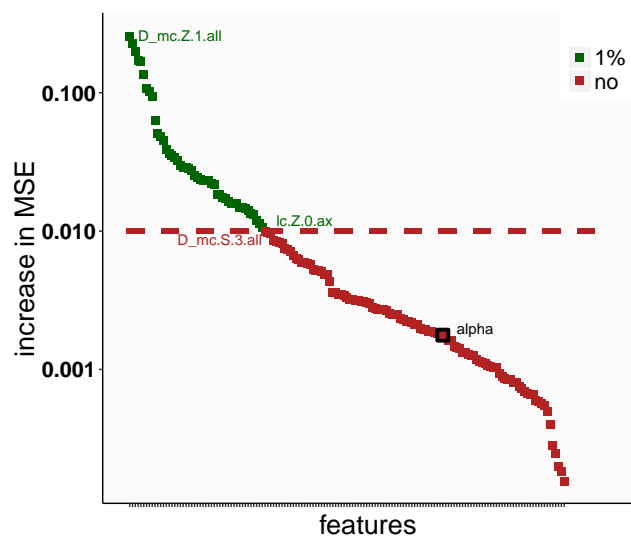


Figure B-14: Out-of-sample mean squared error (MSE) increase for bond length prediction by the random forest model on *spin-splitting* data set in normalized units. Lines indicate 1% (red) cutoff values. The points in the selected for the randF-51B set described in the main text (1% cutoff) are shown in green. Never-selected points are colored red. Some representative features are labeled, including the fraction of Hartree-Fock exchange, labeled as alpha.

Table B.21: Neighborhoods for Fe(III)(pisc)<sub>6</sub> in different feature sets. The neighborhood is defined as the 10 closest structurally-unique complexes in the *spin-splitting* data set. These complexes are sorted by Euclidean distance in the normalized feature space. Here, pisc = phenyl isocyanide, tbuc = t-butylphenyl catecholate, phen = phenanthroline. Hartree-Fock exchange re-sampling (i.e., the closest complex at any HF exchange value is shown once but not more than once for more exchange values) and test/train partitioning is ignored in this analysis (i.e., neighbors can come from test or train). The set of ten closest complexes for randF-41 and randF-49B are identical and only have 2 minor order inversions (shown in italics) for neighbors of comparable distance, while in six cases the order is unchanged (indicated in bold). The set is largely different for RAC-12 and PROX-23, as described in the main text.

randF-41		randF-49B		
1	<b>Co(3)_ax_pisc_eq_pisc</b>	1.1	<b>Co(3)_ax_pisc_eq_pisc</b>	1.1
2	<b>Ni(2)_ax_pisc_eq_pisc</b>	3.0	<b>Ni(2)_ax_pisc_eq_pisc</b>	3.1
3	<b>cr(2)_ax_pisc_eq_pisc</b>	3.2	<b>cr(2)_ax_pisc_eq_pisc</b>	3.2
4	<i>Mn(3)_ax_pisc_eq_pisc</i>	4.5	<i>Fe(3)_ax_NCS_eq_pisc</i>	4.4
5	<i>Fe(3)_ax_NCS_eq_pisc</i>	4.7	<i>Mn(3)_ax_pisc_eq_pisc</i>	4.5
6	<b>Fe(2)_ax_NCS_eq_pisc</b>	5.1	<b>Fe(2)_ax_NCS_eq_pisc</b>	4.8
7	<b>Fe(2)_ax_CN_eq_pisc</b>	5.4	<b>Fe(2)_ax_CN_eq_pisc</b>	5.2
8	<b>cr(2)_ax_NCS_eq_pisc</b>	5.7	<b>cr(2)_ax_NCS_eq_pisc</b>	5.5
9	<i>Fe(3)_ax_H2O_eq_pisc</i>	6.2	<i>Co(3)_ax_pisc_eq_tbuc</i>	6.4
10	<i>Co(3)_ax_pisc_eq_tbuc</i>	6.6	<i>Fe(3)_ax_H2O_eq_pisc</i>	6.7
RAC-12		PROX-23		
1	<b>Co(3)_ax_pisc_eq_pisc</b>	1.0	Fe(3)_ax_misc_eq_misc	0.0
2	Fe(3)_ax_NCS_eq_pisc	2.1	Fe(3)_ax_pisc_eq_pisc	0.0
3	Fe(2)_ax_CN_eq_pisc	2.5	Fe(3)_ax_CN_eq_CN	0.0
4	Co(3)_ax_pisc_eq_tbuc	2.7	Co(3)_ax_misc_eq_misc	1.1
5	Ni(2)_ax_pisc_eq_pisc	2.8	Co(3)_ax_CN_eq_CN	1.1
6	Fe(2)_ax_NCS_eq_pisc	2.9	Co(3)_ax_pisc_eq_pisc	1.1
7	cr(2)_ax_pisc_eq_pisc	3.0	Co(3)_ax_co_eq_co	1.1
8	Fe(3)_ax_phen_eq_phen	3.1	Fe(3)_ax_NCS_eq_pisc	1.5
9	Co(3)_ax_phen_eq_phen	3.2	Fe(2)_ax_co_eq_co	2.0
10	Co(2)_ax_pisc_eq_en	3.3	Fe(2)_ax_misc_eq_misc	2.0

Text B.4: Prediction and feature selection for ionization potential

For ionization potential, a single outlier in the redox data set, a homoleptic Fe-nitrogen complex with three bidentate 2-(4,5-dinitro-1H-imidazol-2-yl)pyridinyl ligands from prior work (Figure B-19), skews predictions (Figure B-18 and Table B.24). Eliminating this 2.0 eV IP underestimation reduces RAC-151 RMSE and MUE to 0.46 (3% error of the mean) and 0.35 eV (2% error of the mean), respectively, and we exclude this outlier from any further gas phase IP discussion (Table B.24). Gas phase IPs in this data set range from 2.3 to 20.6 eV with a mean of 14.4 eV. Comparison of gas phase IP errors for feature subsets selected on spin-splitting (LASSO-28, randF-41, and randF-26) reveals comparable RMSE performance of all three (0.57 eV), but better MUE performance for the smaller LASSO-28 and randF-26 sets at around 0.4 eV. Random forest trained on IP data (suffix "I") from the redox data set retain 28 features (i.e., randF-28I). Again, we observe no improvement in using generalized elastic net over LASSO (Figure B-15) and LASSO selects only 19 features (LASSO-19I, Table B.22). Both of these ionization-potential trained feature sets perform consistently with and without the outlier with RMSEs of  $\sim 0.7$  eV, but are not competitive with the other sets once the outlier is removed from consideration (Table B.24). The proximal-only PROX-23 feature sets produce 3x larger errors than any of the other sets, suggesting relevance of non-local effects, while the poor performance of the distal/middle focused randF-28I/LASSO-19I sets motivate the inclusion of at least  $\sim 25\%$  proximal features (Figure B-17).

Table B.22: Features in the LASSO selected on ionization potential in *redox* data set, LASSO-19I.

1	$\begin{smallmatrix} f \\ ax \end{smallmatrix} \chi_0$	11	$\begin{smallmatrix} lc \\ ax \end{smallmatrix} \chi'_3$
2	$\begin{smallmatrix} f \\ eq \end{smallmatrix} \chi_0$	12	$\begin{smallmatrix} lc \\ ax \end{smallmatrix} T'_3$
3	$\begin{smallmatrix} lc \\ ax \end{smallmatrix} I_3$	13	$\begin{smallmatrix} lc \\ ax \end{smallmatrix} S'_1$
4	$\begin{smallmatrix} lc \\ eq \end{smallmatrix} \chi_3$	14	$\begin{smallmatrix} lc \\ ax \end{smallmatrix} S'_2$
5	$\begin{smallmatrix} lc \\ eq \end{smallmatrix} I_3$	15	$\begin{smallmatrix} lc \\ ax \end{smallmatrix} S'_3$
6	$\begin{smallmatrix} lc \\ eq \end{smallmatrix} S_3$	16	$\begin{smallmatrix} lc \\ eq \end{smallmatrix} T'_3$
7	$\begin{smallmatrix} mc \\ all \end{smallmatrix} Z_0$	17	$\begin{smallmatrix} lc \\ eq \end{smallmatrix} S'_1$
8	$\begin{smallmatrix} mc \\ all \end{smallmatrix} Z_1$	18	$\begin{smallmatrix} lc \\ eq \end{smallmatrix} S'_3$
9	$\begin{smallmatrix} f \\ all \end{smallmatrix} \chi_0$	19	$\begin{smallmatrix} mc \\ all \end{smallmatrix} S'_1$
10	$\begin{smallmatrix} f \\ all \end{smallmatrix} I_0$		

Table B.23: Features in the random forest selected on ionization potential descriptor set, randF-28I.

1	$\overset{f}{all}\chi_1$	11	$\overset{f}{all}T_1$	21	$\overset{lc}{eq}I_3$
2	$\overset{f}{all}\chi_0$	12	$\overset{lc}{eq}T_1$	22	$\overset{lc}{eq}T'_2$
3	$\overset{f}{all}\chi_2$	13	$\overset{lc}{ax}T_1$	23	$\overset{lc}{ax}T'_3$
4	$\overset{f}{all}T_0$	14	$\overset{f}{ax}\chi_0$	24	$\overset{f}{all}\chi_3$
5	$\overset{mc}{all}T_3$	15	$\overset{f}{eq}\chi_0$	25	$\overset{lc}{eq}S_3$
6	$\overset{f}{all}I_0$	16	$\overset{f}{all}T_3$	26	$\overset{mc}{all}Z_3$
7	$\overset{f}{all}I_2$	17	$\overset{lc}{ax}T'_2$	27	$\overset{f}{all}I_3$
8	$\overset{f}{all}I_1$	18	$\overset{f}{all}S_0$	28	$\overset{mc}{all}S_3$
9	$\overset{lc}{ax}\chi_3$	19	$\overset{lc}{eq}T'_3$		
10	$\overset{lc}{eq}\chi_3$	20	$\overset{f}{all}S_1$		

Table B.24: KRR train and test prediction error for ionization potential with variable sets from spin-state splitting feature selection (LASSO-28, randF-41, and randF-26) including the common set among all spin splitting feature selection methods (RAC-12), from redox-based feature selection (LASSO-19I and randF-28I), and both the full set (RAC-155) and the proximal-only subset (PROX-23). Train and test root mean-squared errors (RMSE) and test mean unsigned errors (MUE) are reported in eV. Test set values are also given once a single outlier is excluded.

	train	test		test, excluding outlier	
	RMSE	RMSE	MUE	RMSE	MUE
RAC-155	0.25	0.53	0.38	0.46	0.35
LASSO-28	0.07	0.62	0.44	0.55	0.40
randF-41	0.80	0.64	0.48	0.57	0.44
randF-26	0.84	0.57	0.40	0.55	0.38
LASSO-19I	0.57	0.67	0.45	0.57	0.40
randF-28I	0.72	0.71	0.59	0.72	0.60
C-12	1.06	0.79	0.55	0.77	0.53
PROX-23	1.72	1.62	1.37	1.63	1.37

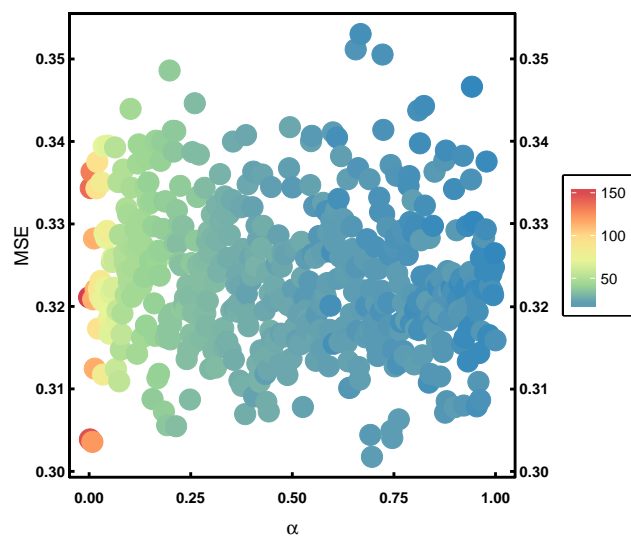


Figure B-15: Error response of elastic net model on ionization potential with varying  $\alpha$  parameter from 0 (ridge regression) to 1 (LASSO). Colors indicate the number of selected variables at each value.

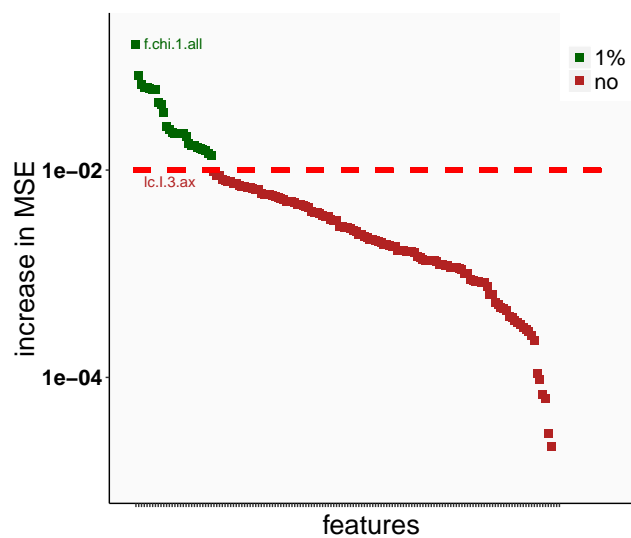


Figure B-16: Out-of-sample mean squared error (MSE) increase for ionization potential prediction by random forest model on the *redox* data set in normalized units. Lines indicate 1% (red) cutoff values and corresponding points which correspond to randF-29I (green) feature sets in the main text. Never-selected points are colored red. Some representative features are labeled.

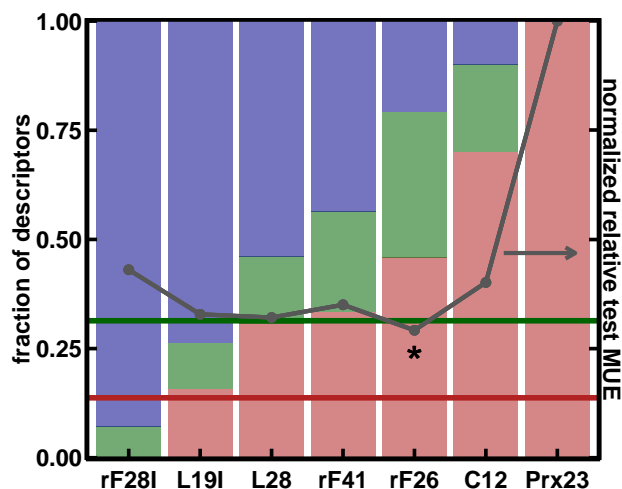


Figure B-17: Fraction of selected descriptors that are proximal (red), middle (green), or distal (blue), as defined in the main text and depicted in main text Figure 4-3. All subsets are compared against a RAC-155 reference (dark red and green horizontal lines). The normalized relative test set ionization potential MUE from a KRR model is shown in dark grey for each subset, and the lowest test MUE is indicated with an asterisk. Sets are sorted left to right with decreasing distal fraction: random forest on ionization potential (rf29I), LASSO on ionization potential (L19I) or spin-splitting (L28), random forest on spin-splitting (1%, rF41 or 2%, rF26), spin-splitting common set (C12), and proximal-only (Prx23). HF exchange and oxidation state are not used in any models.



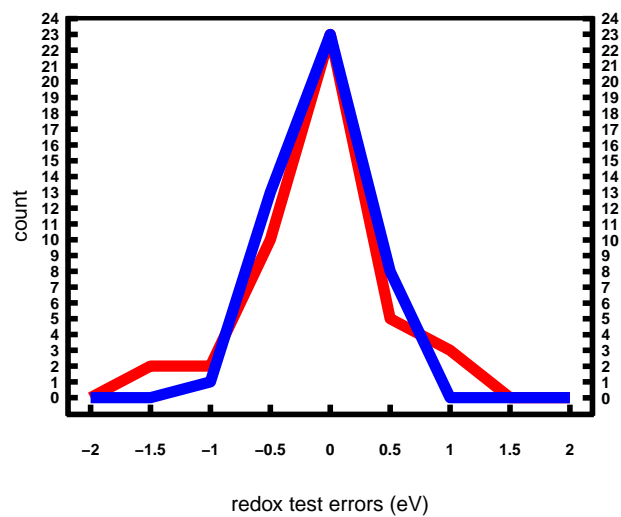


Figure B-18: Histogram of prediction errors for RAC-155 on ionization (red) and redox (blue) potentials in eV for an identical test set, as described in the main text.

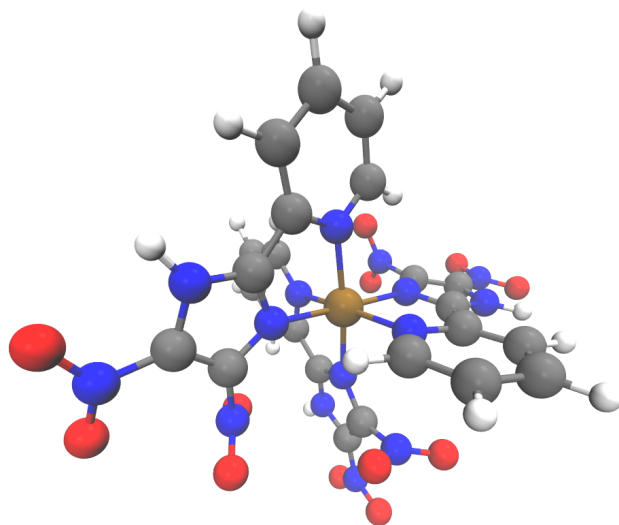


Figure B-19: Outlier iron-nitrogen complex for ionization potential prediction but not redox potential prediction.

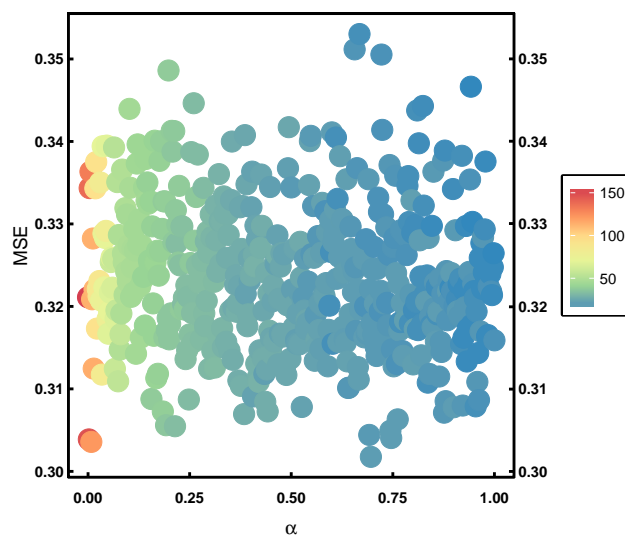


Figure B-20: Error response of elastic net model on redox potential with varying  $\alpha$  parameter from 0 (ridge regression) to 1 (LASSO). Colors indicate the number of selected variables at each value.

Table B.25: Features in the LASSO trained on redox potential descriptor set, LASSO-19G.

1	$lcI_3$	11	$lcS'_1$
	$axI_3$		$axS'_1$
2	$lcZ_1$	12	$lcS'_2$
	$eqZ_1$		$axS'_2$
3	$lcI_3$	13	$lcS'_3$
	$eqI_3$		$axS'_3$
4	$lcS_3$	14	$lc\chi'_2$
	$eqS_3$		$eq\chi'_2$
5	$mcZ_0$	15	$lcS'_1$
	$allZ_0$		$eqS'_1$
6	$mcS_0$	16	$lcS'_3$
	$allS_0$		$eqS'_3$
7	$mcS_1$	17	$mc\chi'_2$
	$allS_1$		$all\chi'_2$
8	$f\chi_0$	18	$mcZ'_1$
	$all\chi_0$		$allZ'_1$
9	$fI_0$	19	$mcS'_2$
	$allI_0$		$allS'_2$
10	$lc\chi'_3$		
	$ax\chi'_3$		

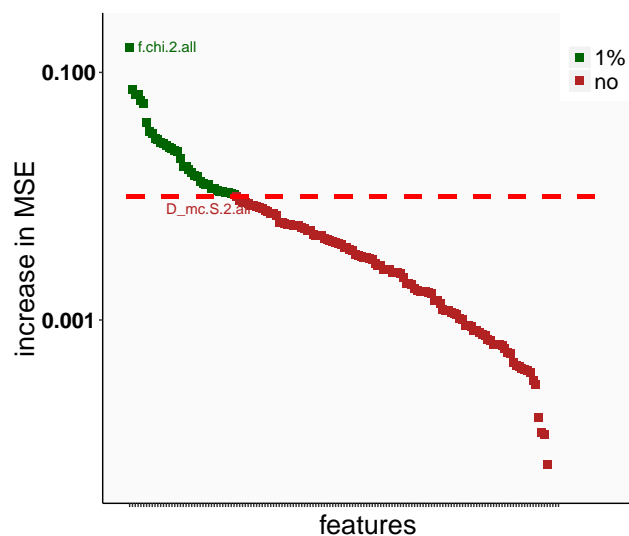


Figure B-21: Out-of-sample mean square error (MSE) increase for redox prediction by random forest model on redox data set in normalized units. Lines indicate 1% (red) cutoff values and corresponding points which correspond to randF-38G (green) feature sets in the main text. Never-selected points are colored red. Some representative features are labeled.

Table B.26: Features in the random forest on redox potential data descriptor set, randF-38G.

1	$f_{all}\chi_2$	11	$lc_{eq}T_1$	21	$lc_{eq}\chi_3$	31	$lc_{ax}T'_3$
2	$f_{all}I_3$	12	$f_{all}T_3$	22	$lc_{eq}Z'_2$	32	$mc_{all}Z'_2$
3	$f_{all}T_2$	13	$f_{all}I_1$	23	$mc_{all}S_2$	33	$f_{all}\chi_0$
4	$f_{all}\chi_1$	14	$mc_{all}T_2$	24	$lc_{ax}\chi_3$	34	$mc_{all}\chi'_1$
5	$mc_{all}Z'_1$	15	$lc_{ax}T_1$	25	$mc_{all}Z_2$	35	$f_{all}Z_1$
6	$f_{all}T_1$	16	$lc_{eq}Z_3$	26	$f_{eq}\chi_0$	36	$lc_{ax}S_3$
7	$mc_{all}Z_0$	17	$f_{all}Z_0$	27	$lc_{eq}Z'_1$	37	$f_{eq}\chi_1$
8	$f_{all}\chi_3$	18	$f_{all}I_0$	28	$mc_{all}Z_1$	38	$f_{all}Z_2$
9	$f_{all}T_0$	19	$lc_{eq}S_3$	29	$mc_{all}\chi'_2$		
10	$f_{all}I_2$	20	$lc_{eq}T'_3$	30	$mc_{all}T_3$		

Table B.27: Comparison of redox potential predictions using RAC-155 modeled by a one-electron ionization from the ground state of the reduced species (GS), high-spin state of the reduced species (HS), or low spin state of the reduced species (LS). Note the number of points is different in each case owing to missing values from some high-spin structures. Train and test partitions are divided to be roughly 80-20 in all cases and the GS set is selected to be comparable in size to the HS or LS cases.

	GS	LS	HS
train RMSE (eV)	0.23	0.16	0.29
test RMSE (eV)	0.41	0.46	0.35
test MUE (eV)	0.31	0.32	0.27
train size	148	157	143
test size	37	39	35

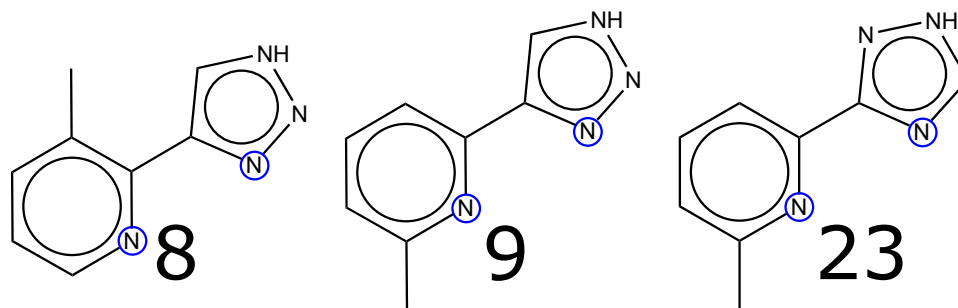


Figure B-22: Three triazolyl-pyridine ligands from the *redox* data set with different substituents and orientations: ligand 8 (SMILES:c1(nn[nH]c1)c1ncccc1C), ligand 9 (SMILES:c1(nn[nH]c1)c1nc(ccc1)C) and ligand 23 (SMILES:c1(cccc(n1)C)c1n[nH]cn1).

Text B.5: Effect of triazolyl-pyridine substituents on similarity in feature sets

We calculated the Euclidean distance in normalized feature space between the complexes with ligands 8, 9 and 23 (Figure B-22) and determined the nearest-neighbors for ligand 9 complex with randF-26, randF-41, and randF-38G. These ligand 8, 9, and 23 complexes have redox potentials of 5.5 eV, 6.1 eV, and 6.0 eV, respectively. Using randF-41 or randF-26, the nearest neighbor to the ligand 9 complex is the ligand 8 complex, with a distance of 0.38 and 0.03 for randF-26, whereas the complex with ligand 23 is at a distance of 1.24 or 0.93 under randF-26. With randF-38G, the nearest neighbor to ligand 9 is ligand 23, with the 9–23 distance being 1.11 compared to the 8–9 distance  $\sim 4.8$ . This large difference in intercomplex distances is caused by the distal metal-centered topological descriptors in randF-38G that relate ligand 9 and ligand 23 complexes by the position of the substituent.

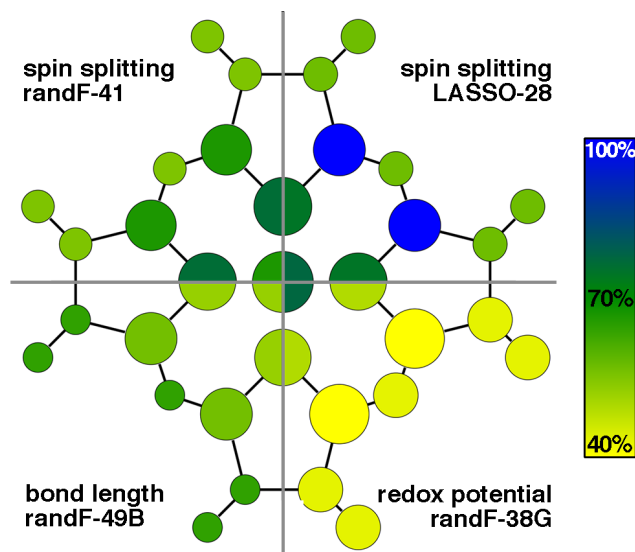


Figure B-23: Schematic of relative proximity and electronic (blue) or topological (yellow) of feature sets on an iron-porphyrin complex. Feature sets are designated by their training data: spin splitting (randF-41 and LASSO-28, top), bond length (randF-49B, bottom left), and redox potential (randF-38G, bottom right). Atom sizes are scaled relative to the number of descriptor dimensions involving that atom (divided into first shell, second shell and other), scaled, with iron kept the same size in all sets. The color bar and absolute percentages of electronic and topological descriptors, as defined in the main text, is shown in inset right.



## Appendix C

### Design of spin crossover materials with ANNs and DFT

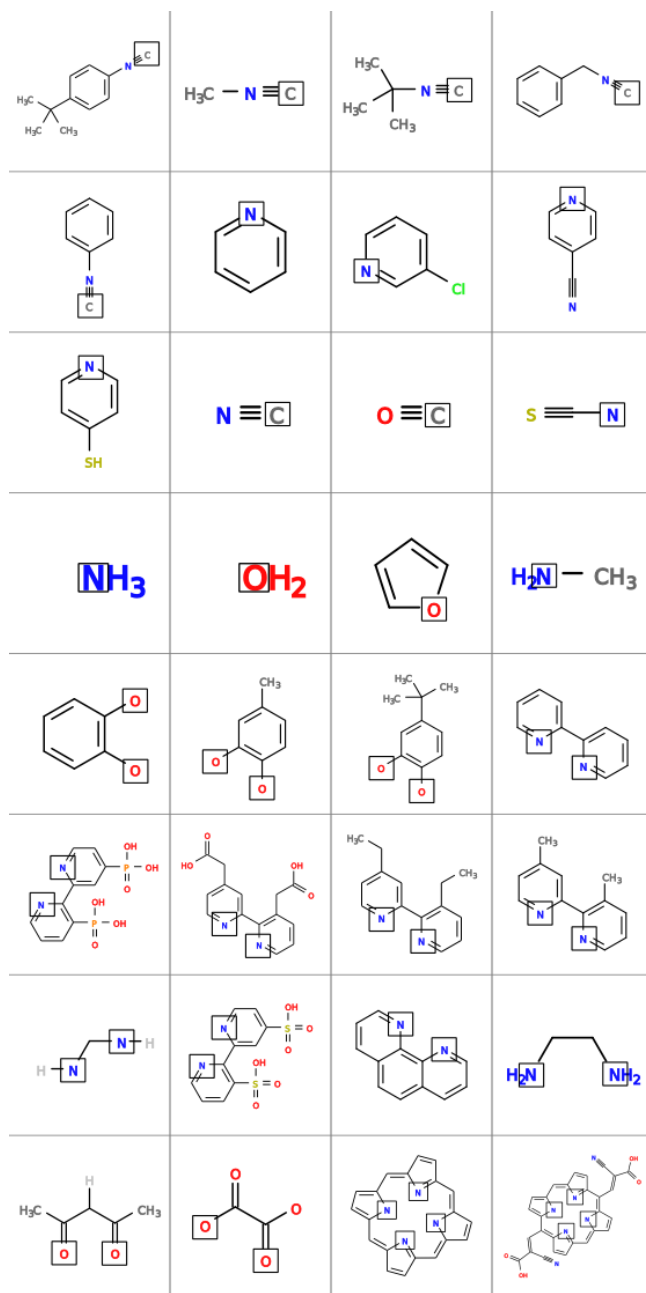


Figure C-1: Structures of design space ligands. The metal-coordinating atom is surrounded by a box in each case and ligands are ordered as in Table C.1.



Table C.1: Design space of ligands used in current work. Ligands occurring in ANN training data<sup>308</sup> are shown in boldface. The ANN pool also included Cl<sup>-</sup> and SCN ligands which are not included in the design space. Equivalent ligands are coded as "EC" with those with the same letter in that column corresponding to cases that are seen as identical by the MCDL-25 descriptor set.

index	name	dent.	n <sub>at</sub>	CA	EC	SMILES
1	<b>pisc</b>	<b>1</b>	<b>25</b>	<b>C</b>	<b>a</b>	<b>[C][N]c1ccc(C(C)(C)C)cc1</b>
2	<b>misc</b>	<b>1</b>	<b>6</b>	<b>C</b>		<b>[C][N]C</b>
3	tbisc	1	15	C		[C][N]C(C)(C)C
4	benzisc	1	16	C		[C][N]C1ccccc1
5	phenisc	1	13	C	a	[C][N]c1ccccc1
6	pyridine	1	11	N		c1ccncc1
7	chloropyr	1	11	N	b	c1c(ccc1)Cl
8	cyanopyr	1	12	N	b	c1(ccncc1)C#N
9	thiopyr	1	12	N	b	c1(ccncc1)S
10	<b>cyanide</b>	<b>1</b>	<b>2</b>	<b>C</b>		<b>[C]#N</b>
11	<b>carbonyl</b>	<b>1</b>	<b>2</b>	<b>C</b>		<b>[C]=O</b>
12	<b>isothiocyanate</b>	<b>1</b>	<b>3</b>	<b>N</b>		<b>[N]C#[S]</b>
13	<b>ammonia</b>	<b>1</b>	<b>4</b>	<b>N</b>	c	<b>N</b>
14	<b>water</b>	<b>1</b>	<b>3</b>	<b>O</b>		<b>O</b>
15	furan	1	9	O		o1cccc1
16	methylamine	1	7	N	c	CN
17	cat	2	12	O	d	[O]c1c(cccc1)[O]
18	mec	2	15	O	d	[O]c1c(cc(cc1)C)[O]
19	<b>tbuc</b>	<b>2</b>	<b>24</b>	<b>O</b>	d	<b>[O]c1c(cc(C(C)(C)C)cc1)[O]</b>
20	<b>bpy</b>	<b>2</b>	<b>20</b>	<b>N</b>		<b>n1ccccc1c1ncccc1</b>
21	phosacidbpy	2	30	N	e	n1ccc(cc1c1nccc(c1)P(=O)(O)O)P(=O)(O)O
22	aceticacidbpy	2	32	N	e	n1ccc(cc1c1nccc(c1)CC(=O)O)CC(=O)O
23	ethbpy	2	32	N	e	n1ccc(cc1c1nccc(c1)CC)CC
24	mebpy	2	26	N	e	n1ccc(cc1c1nccc(c1)C)C
25	diaminomethyl	2	7	N		[NH]C[NH]
26	sulfacidbpy	2	28	N	e	n1ccc(cc1c1nccc(c1)S(=O)(=O)O)S(=O)(=O)O
27	<b>phen</b>	<b>2</b>	<b>22</b>	<b>N</b>		<b>c1cc2ccc3ccncc3c2nc1</b>
28	<b>en</b>	<b>2</b>	<b>12</b>	<b>N</b>		<b>NCCN</b>
29	<b>acac</b>	<b>2</b>	<b>14</b>	<b>O</b>		<b>O=C(C)[CH]C(=O)C</b>
30	<b>ox</b>	<b>2</b>	<b>6</b>	<b>O</b>		<b>C(=O)([O])C(=O)[O]</b>
31	<b>porph</b>	<b>4</b>	<b>36</b>	<b>N</b>	f	<b>N1C2=CC3=NC(=CC4=CC=C(N4)C=C5C=CC(=N5)C=C1C=C2)C=C3</b>
32	cyanoaceticporph	4	52	N	f	N1C2C=CC1=C(c1[n]c(cc1)C=C1N=C(C(=c3[n]c(=C2)cc3)C=C(C(=O)O)C#N)C=C1)C=C(C(=O)O)C#N

Table C.2: Allowed ligand combinations.

class	allowed ax	allowed eq	total
monodentates	16	16	256
bidentates	14	14	196
monodenate ax. + bidentate eq.	16	14	224
monodenate ax. + tetradentate eq.	16	2	32

Table C.3: Spin multiplicities for each metal and oxidation state.

		M(II) spin	M(III) spin
Cr	LS	3	2
	HS	5	4
Mn	LS	2	3
	HS	6	5
Fe	LS	1	2
	HS	5	6
Co	LS	2	1
	HS	4	5

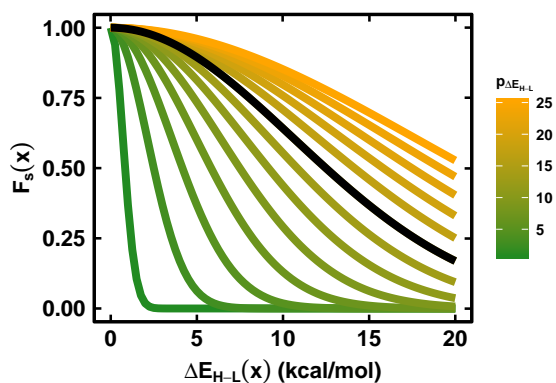


Figure C-2: Response of fitness function  $F_s(x)$  to changing splitting energy,  $\Delta E_{H-L}$  for different values of the parameter  $p_{\Delta E_{H-L}}$ . The default value  $p_{\Delta E_{H-L}} = 15$  is shown in black.

### Text C.1: Standard GA parameters

The following conditions were used for the 50 repeats of ANN-GA under each control condition as well as the DFT-SP-GA:

A **pool size of 20**, which are chosen randomly in each repeat run by randomly selecting a metal, oxidation state and ligand from their respective pools with equal probability. Incompatible ligand combinations (e.g. porphyrin and a bidentate) are rejected, as are duplicates, and this is continued until the pool is full. This pool size affects the maximum number of simultaneous evaluations needed and hence the parallelizability of the run, and is similar to that used in Ref<sup>14</sup> and in the typical range for evolutionary algorithms<sup>514</sup>.

The GA was run for **21 generations**. This is considerably shorter than Ref.<sup>14</sup>, where 200 generations are used. However, we observe that the average diversity of complexes in the pool falls sharply by this stage (see Figure 5-2, main text), which is possibly due to the much smaller design space used ( $\sim 5600$  compared to  $1.3 \times 10^6$  in Ref.<sup>14</sup>). Therefore, we conduct more repeat runs with random starting populations instead of running longer generations which will only conduct localized candidate searches which can be sparse.

A **mutation probability of 0.15** was used. This is on the high end of the range tested in Ref.<sup>14</sup> and higher than the recommendations in Ref.<sup>514</sup>. The general impact of mutation rate is to drive exploration of the design space at the expense of fitness<sup>515</sup>, but again our choice is motivated by the practical observation that high fitness and low diversity situations arise rapidly in this design space (see the main text Figure 5-2).

5 crossovers were used, meaning that at most 25% of the pool at each stage are combinations of parents, a similar ratio to Ref.<sup>14</sup>. The split and distance parameters ( $p_{\Delta E_{H-L}}$  and  $p_d$  in Equations C.1 and C.2) are set to 15 kcal/mol and 1.0 respectively:

$$F_s(x) := \exp \left[ - \left( \frac{\Delta E_{H-L}(x)}{p_{\Delta E_{H-L}}} \right)^2 \right] \quad (C.1)$$

This function depends only a parameter  $p_{\Delta E_{H-L}}$ , which determines how aggressively the fitness value decays as  $|\Delta E_{H-L}|$  increases (see Figure C-2). We take a default value of  $p_{\Delta E_{H-L}} = 15$  kcal/mol. In our modified form of (C.1), the splitting + distance fitness function:

$$F_{s+d}(x) := \exp \left[ - \left( \frac{\Delta E_{H-L}(x)}{p_{\Delta E_{H-L}}} \right)^2 \right] \exp \left[ - \left( \frac{d(x)}{p_d} \right)^2 \right] \quad (C.2)$$

we set  $p_d = 1$  by default. Since our original publication<sup>308</sup>, we have changed from a  $[0, 1]$  descriptor normalization (dividing by the maximum value) to a normalization to unit variance, meaning descriptors are divided by the standard deviation of that descriptor on training data. This change alters the scale of the normalized distances, so we apply a constant scaling factor of 3 to all distances in order to better match our previous scale.

---

**Algorithm 1 Basic Genetic Algorithm based on based on**<sup>14</sup>

---

**Require:**  $x \in X$  are possible allowed complexes,  $F(x) \rightarrow [0, 1]$  fitness function,

$N_{gen}, N_{pool}, N_{cross} \in \mathbb{N}_0, p_{mut} \in [0, 1]$   
 1:  $X_{pool} \leftarrow N_{pool}$  random combinations  $x_i, f_i \leftarrow F(x_i), i = 0$   
 2: **while**  $i \leq N_{gen}$  **do**  
 3:    $X_{select} \leftarrow \emptyset$   
 4:   **while**  $|X_{select}| \leq |X_{pool}|$  **do**  
 5:     choose random  $x_i \in X_{pool}, \eta \leftarrow \mathcal{U}(0, 1)$   
 6:     **if**  $f_i \geq \eta$  **then**  
 7:        $X_{select} \leftarrow X_{select} \cup x_i$   
 8:     **end if**  
 9:   **end while**  
 10:   **for**  $j = 1$  to  $N_{cross}$  **do**  
 11:     choose random  $x_i, x_j \in X_{select}$ , and perform crossover operation  
 12:   **end for**  
 13:   **for**  $j = 1$  to  $N_{pool}$  **do**  
 14:      $\eta \leftarrow \mathcal{U}(0, 1)$   
 15:     **if**  $p_{mut} \geq \eta$  **then**  
 16:       perform mutation operation  $x_j$   
 17:     **end if**  
 18:   **end for**  
 19:    $f_i \leftarrow F(x_i) \forall x_i \in X_{select}$   
 20:    $X_{pool} \leftarrow$  fittest  $N_{pool}$  from  $X_{pool} \cup X_{select}$   
 21: **end while**

---

## Text C.2: Timing information for DFT and ANN

ANN Timing: The combinatorial design space of 5664 complexes (11328 geometries) is amenable to full enumeration, which took 7.25 hours (2.3 secs/geometry) using molSimplify on a standard workstation. The main component of the run time is the constrained force field optimization of each complex. A typical GA run requires 40x2.3s steps to start followed by additional runs at each generation, producing a usual ANN-GA runtime of around 5 minutes total.

DFT Timing: All DFT timing data is collected on local computing resources, with DFT calculations conducted on single NVidia Geforce 970 GTX consumer-grade GPU.

The mean time to complete one DFT-SP evaluation is  $\sim 1.0$  hours (Figure C-8), and if all evaluations took the mean time one SP-GA run would complete in 21 hours assuming sufficient parallel resources (40 GPUs peak, being one dedicated GPU for running the high and low spin geometry optimization calculations in parallel). However, we observe that the mean time to complete one SP-GA in practice is  $\sim 100$  hours (Figure C-10) because each step is limited by the slowest calculation in that generation, requiring  $\frac{100}{21} \approx 5$  hours per GA generation. Based on the mean run time, evaluation of the full design space at the SP level would require  $\sim 460$  GPU-days.

Based on a mean time of 13 hours per DFT geometry optimization (GO, Figure C-9), one GA of 21 generations would require  $\approx 11$  days based on DFT geometry optimizations with sufficient parallel resources (40 GPUs peak as above). However, as above the SP-GA runs take  $\frac{5}{1.5} = 3.3$  times longer than expected due to the long tail in run times (Figure C-8). Assuming the same scaling applies to geometry optimizations, GO-GA runs would take 33 days with optimal parallel resources. In practice adaptations could be made to GA algorithm to prevent these bottlenecks but even removing the dead-time waiting entirely the computational costs of geometry optimizations remain prohibitively expensive. Conducting geometry optimizations on the full design space based on a mean time would require  $\sim 6000$  GPU-days.

### Text C.3: Explanation of mAD syntax

All of the genetic algorithms presented in this work were executed by molSimplify Automatic Design (mAD) <https://github.com/hjkgrp/AutomaticDesign>, an open source python module that uses molSimplify(<https://github.com/hjkgrp/molSimplify>) to conduct inorganic molecular design. The only requirement to use mAD with the ANN model is a working molSimplify installation.

mAD runs require two steps: run creation using the **-new** command and run execution using the **-resume** command. A default instance of the neural network guided genetic algorithm can be run as follows:

```
$ mad -new
$ mad -resume GA_run -reps 21
```

This will perform the default 21 generations of evolution using the ligands list used in this work along with the default parameters given in Text C.1 including both control types, and should complete in 5-7 minutes on a standard workstation. The parameters of the run can be extensively customized and run parameters are stored in an input file that can be passed to mAD using **-new**. mAD can also create, submit and monitor TeraChem DFT jobs on remote resources. Full instructions are available in the readme file <https://github.com/hjkgrp/AutomaticDesign/blob/master/molSimplifyAD/readme.docx>.

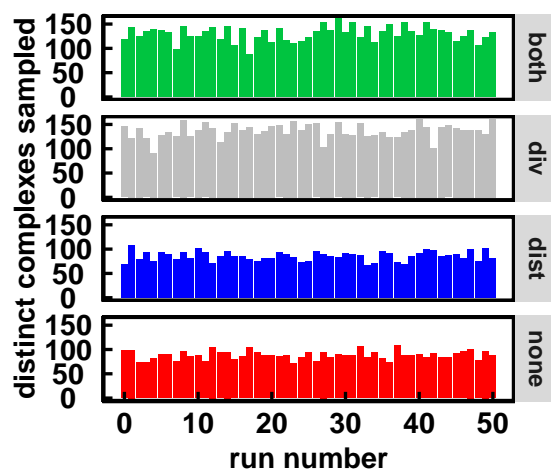


Figure C-3: Number of unique complexes sampled during 50 repeats of a 21-generation GA using an ANN under different control schemes.

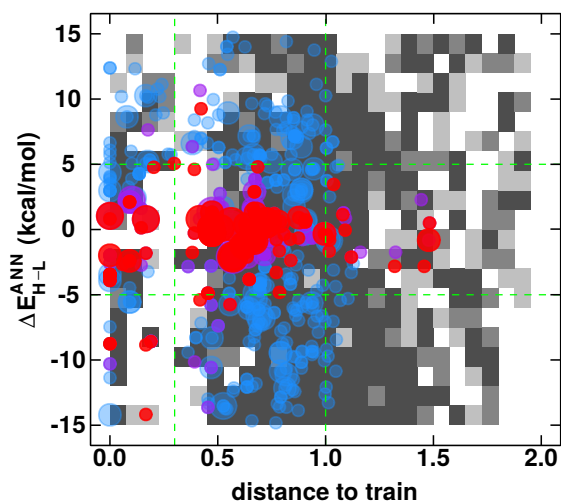


Figure C-4: 2D histogram showing the density of points in the full space partitioned by ANN-predicted spin splitting energy and distance to nearest training point (the frequency of complexes in each bin are shown in gray squares), with initial, middle and final retained complexes over 50 repeats of the GA with ‘both’ control shown in increasingly opaque circles for generations 0, 10 and 20 in blue, purple, and red, respectively. The size of the circles indicates the frequency at which a given point appears in the retained set.

Table C.4: Number of complexes and hits (points with  $|\Delta E_{H-L}(ANN)| \leq 5$  and  $\text{dist} \in [0.3, 1]$ ) sampled or missed during 50 repeats of the ANN-GA using different control modes

	complexes sampled	hits sampled	% hits missed
both	3297	372	21.5
dist	2639	300	36.7
div	3400	379	20.0
none	2794	327	31.0

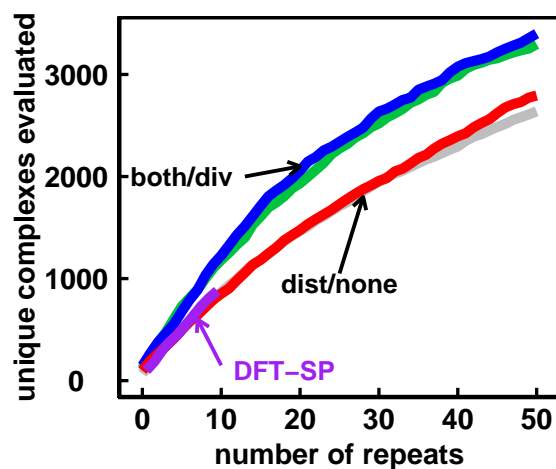


Figure C-5: Number of unique complexes sampled during 50 repeats of a 21-generation GA using an ANN under different control schemes

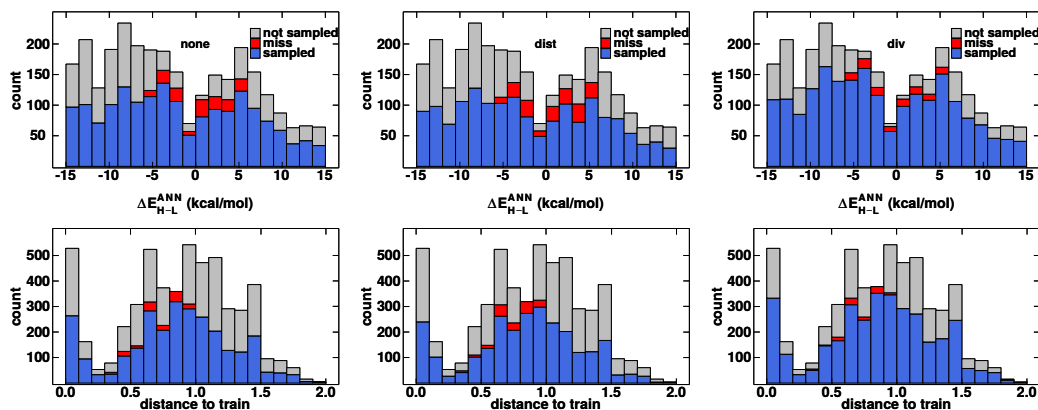


Figure C-6: Number of complexes explored by 50 repeats of the ANN-GA shown distributed by predicted splitting energy (top row) and distance (bottom row), for no control (left), distance control (middle) and diversity control (right). Both controls are shown in the main text. Sampled complexes are shown in blue, unsampled compounds in blue, and unsampled hits (points with  $|\Delta E_{H-L}^{ANN}| \leq 5$  and  $dist \in [0.3, 1]$ ) are marked in red.



#### Text C.4: Dimensionality reduction with t-SNE

In order to plot how the predicted spin splitting energy and distance to training parameters vary throughout the  $\sim 5600$  element design space, we describe the complexes with a fully continuous 41-descriptor set previously shown to be comparable in nature to the MCDL-25 set<sup>511</sup>. To visualize this space, we use t-stochastic network embedding<sup>528</sup> (t-SNE), a visualization technique that attempts to preserve the distribution of pairwise distances in the full space in a 2D representation. We use a perplexity value of 20, but the result is not sensitive to small ( $5\pm$ ) perturbations about this value. The t-SNE assigns each complex a unique set of coordinates in arbitrary units based on structure alone, and we color these points by predicted splitting energy and use linear smoothing to draw the surface. In order to represent uncertainty, as measured by the distance to training data, we superimpose increasingly large black regions set to 10% opacity corresponding to distances greater than 0.8 to 1.5 in 0.1 increments. Convex hulls are determined based on the 2D t-SNE representation using the R function *chull*, and consist of minimal linear enclosures of all points with only the specific ligand class.

Table C.5: Comparison of ANN predictions (ANN) and DFT energies on initial geometries (single points, SP) or optimized geometries (GO) all at the B3LYP/LACVP\* level of theory. Two thermochemical calculations did not converge and are marked with ‘-’.

metal	ox	ax. lig.	eq. lig.	$\Delta E_{\text{H-L}}$ (kcal/mol)			$\Delta G_{\text{H-L}}$	dist. to train
				GO	SP	ANN	(kcal/mol)	
Cr	2	2	13	-3.52	-2.45	0.51	-1.03	0.72
Mn	3	20	28	-1.22	75.17	-0.26	-3.70	0.47
Mn	3	24	28	-0.72	15.60	-0.26	-4.54	0.47
Mn	3	22	22	-0.65	14.81	-0.26	-5.81	0.47
Mn	3	16	5	9.51	12.64	-0.64	10.46	0.66
Cr	2	2	16	-4.76	-17.75	0.51	-7.72	0.72
Mn	3	13	3	5.40	10.44	-2.10	14.07	0.57
Mn	3	13	2	5.63	12.21	-1.23	11.91	0.62
Mn	3	21	28	-1.13	14.47	-0.26	-4.69	0.47
Fe	2	27	28	-2.76	2.03	0.81	-3.25	0.42
Co	2	27	20	-6.44	-1.32	-2.40	-10.65	0.09
Co	3	14	14	2.45	-9.52	0.79	12.74	0.17
Mn	3	16	3	9.65	13.31	-2.10	6.22	0.57
Mn	3	8	3	8.59	80.27	0.91	5.23	0.88
Cr	2	3	16	-5.62	-17.51	-0.87	-7.83	0.67
Cr	2	30	29	-3.22	15.12	0.10	-1.53	0.51
Mn	3	27	27	-0.60	16.40	1.05	-4.39	0.00
Mn	3	13	5	6.14	15.34	-0.64	14.38	0.66
Co	2	2	7	-1.69	-141.38	0.21	-1.48	0.91
Mn	3	13	4	6.03	26.72	0.60	18.73	0.78
Co	2	3	8	-1.23	37.03	-0.32	-0.98	0.91
Mn	3	23	28	-0.46	13.22	-0.26	-2.37	0.47
Co	2	3	28	0.54	-100.07	1.62	0.08	0.14
Mn	2	2	11	4.69	4.85	1.56	1.43	0.48
Co	2	14	28	1.35	-13.59	0.81	-0.53	0.00
Co	2	15	23	-1.88	-87.51	-2.29	0.32	0.54
Cr	2	1	9	-3.42	5.53	-1.90	-10.69	0.72
Fe	3	11	12	9.08	-4.58	4.67	1.26	0.94
Mn	2	11	5	20.94	17.29	2.02	11.54	0.50
Mn	3	8	2	7.93	79.01	2.90	9.04	0.91
Fe	3	12	20	-3.01	-71.83	4.22	0.38	0.87
Fe	3	1	8	4.01	-29.62	3.07	19.32	0.89
Mn	3	6	2	7.13	90.77	2.90	8.84	0.91
Cr	2	5	7	-2.75	13.08	-2.25	-	0.72
Fe	3	3	24	14.69	-44.10	2.78	9.38	0.68
Fe	3	1	24	12.40	-38.24	3.78	14.11	0.77
Fe	2	2	12	2.11	-36.53	-2.55	8.37	0.56
Mn	3	6	5	10.72	28.60	4.61	4.04	0.94
Mn	3	6	1	7.68	35.72	4.61	-	0.94
Cr	2	19	29	-4.13	13.46	0.39	-3.81	0.63
Fe	2	3	29	-1.29	-0.58	2.70	-7.12	0.51
Co	2	13	2	7.61	-17.04	-2.76	5.43	0.81
Cr	2	1	13	-3.41	-3.08	1.23	-1.97	0.76
Co	2	4	29	-5.81	47.55	-3.89	-9.83	0.93
Fe	2	5	16	3.59	-82.41	4.93	3.27	0.79
Co	2	16	29	-9.26	-34.21	-0.38	-12.87	0.81
Mn	3	12	20	-5.24	47.39	-4.20	-5.09	0.84
Fe	3	12	24	-1.67	-0.11	4.22	-2.04	0.87
Co	2	15	16	0.55	-97.45	-4.76	1.93	0.53
Cr	2	4	16	-2.49	-21.07	1.36	-5.45	0.81
Co	2	2	29	-6.38	-45.64	-0.75	-10.54	0.84

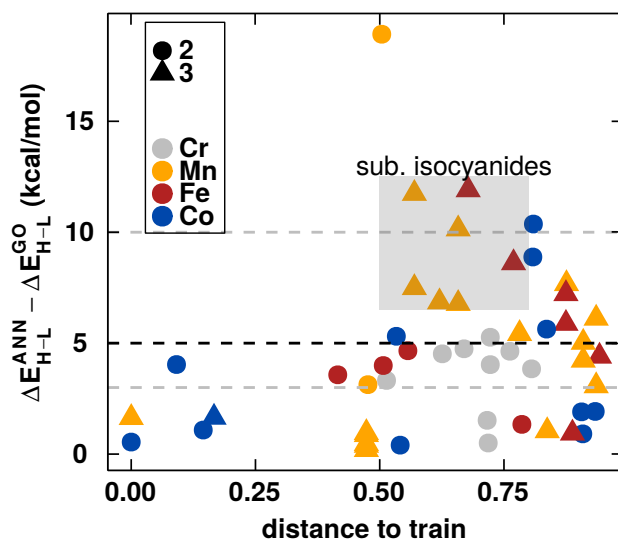


Figure C-7: Plot of error in ANN predictions with respect to DFT optimization compared to distance to closest training point for 51 molecule subset. Isocyanides are shaded in a gray square, excluding the single Mn(II) outlier.

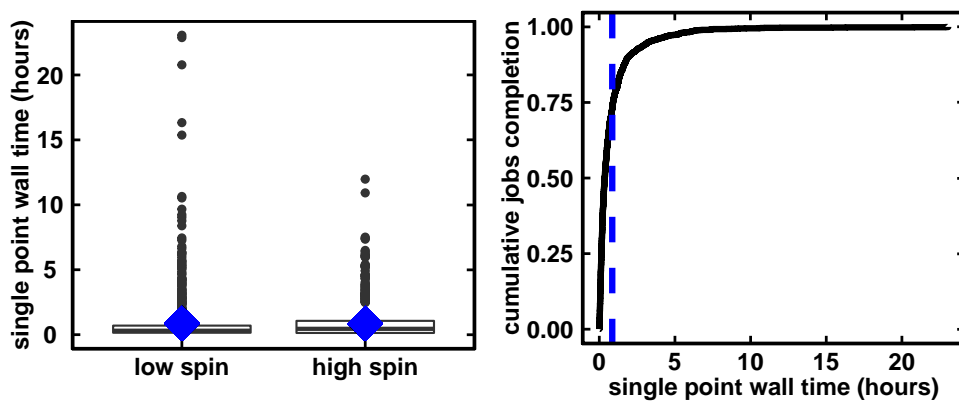


Figure C-8: Wall time in hours (left) and cumulative completion rates (right) for all DFT single point energies conducted during 10 repeat runs of 21 generations of the genetic algorithm. The mean time for high- and low-spin complexes are indicated with diamonds, the median time is indicated by horizontal bar and the interquartile range is boxed (left) while the mean overall time is indicated with a dashed vertical line (right)

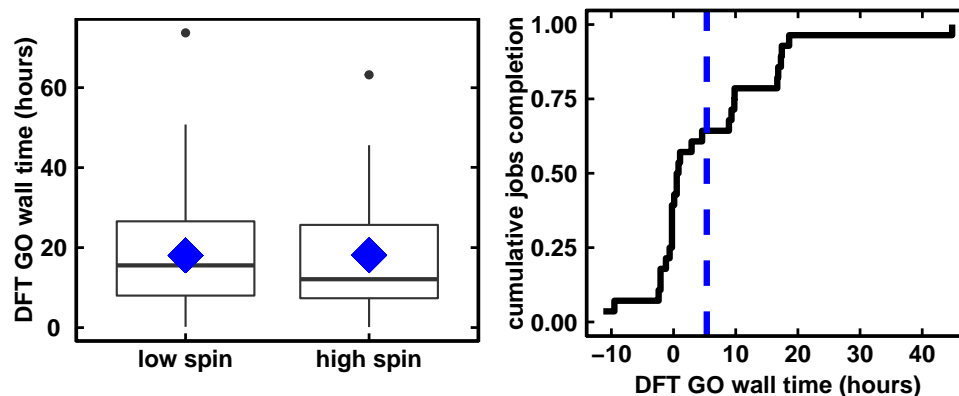


Figure C-9: Wall time in hours (left) and cumulative completion rates (right) for all DFT geometry optimizations conducted on 51 leads from the ANN-driven genetic algorithm. The mean time for high- and low-spin complexes are indicated with diamonds, while the median time is indicated by horizontal bar and the interquartile range is boxed (left) while the mean overall time is indicated with a dashed vertical line (right)

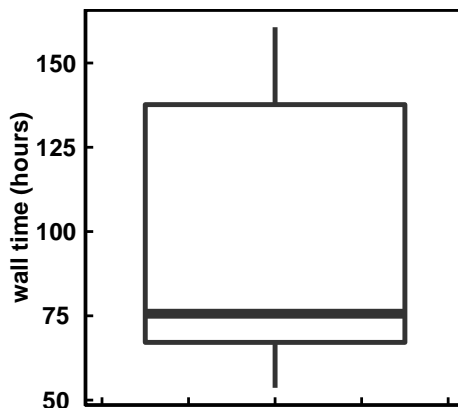


Figure C-10: Wall time in hours for 10 repeat runs of 21 generations of the genetic algorithm based on DFT-SP evaluation . The median time of 76 hours is indicated with a bar. The interquartile range is boxed.

Table C.6: Evaluation of DFT-SP-GA hits with full geometry optimization

metal	ox	ax. Lig.	eq. lig.	$\Delta E_{H-L}$ (kcal/mol)		error (kcal/mol)
				SP	GO	
Cr	2	9	24	-3.69	-11.06	7.38
Co	3	11	12	1.00	0.12	0.89
Fe	3	11	30	-1.91	4.92	-6.83
Fe	3	10	29	2.47	9.77	-7.30
Fe	3	10	28	-1.54	25.07	-26.61
Fe	3	2	12	-1.63	3.67	-5.30
Fe	3	14	30	-0.40	-4.98	4.58
Co	2	23	22	-0.52	-5.58	5.06
Fe	3	7	3	-0.16	26.96	-27.13
Co	3	12	21	1.21	-0.35	1.56
Co	3	12	19	-1.30	-10.17	8.87
Cr	2	9	27	-1.65	-10.20	8.55
Co	3	10	8	-1.57	44.63	-46.20
Co	3	12	7	0.20	7.32	-7.12
Fe	3	7	11	-1.89	11.87	-13.76
Co	2	24	20	-0.69	-6.05	5.36
Co	3	12	12	-0.94	6.82	-7.76
Fe	3	6	1	-0.63	18.83	-19.46
Fe	3	5	13	1.66	4.03	-2.37
Cr	2	4	24	-2.36	0.16	-2.52
Fe	3	3	12	-2.68	4.01	-6.70
Mn	2	23	27	-0.65	-22.03	21.38
Co	2	14	3	1.02	12.72	-11.70
Cr	2	14	5	-2.14	-2.34	0.21
Cr	2	7	24	1.01	-5.84	6.85
Fe	3	7	5	0.24	20.40	-20.16
Fe	3	14	31	-1.26	6.18	-7.45
Fe	3	1	13	0.48	4.05	-3.57
Fe	3	30	30	1.66	-11.79	13.44
Fe	3	2	13	0.09	13.85	-13.76
Cr	2	9	28	2.15	-11.13	13.27
Fe	3	16	1	-1.92	18.64	-20.56
Fe	3	8	1	1.69	18.00	-16.31
Cr	2	5	6	1.79	-3.91	5.70
Cr	2	14	1	1.06	-2.25	3.31
Mn	2	29	27	-0.27	-23.67	23.40
Co	3	9	22	2.59	8.66	-6.07
Co	3	12	17	-2.75	-12.29	9.55
Co	2	10	3	2.05	23.38	-21.34
Fe	3	5	12	-0.34	3.92	-4.27
Co	3	14	16	1.82	24.61	-22.79
Co	2	27	22	-0.57	-5.25	4.68
Fe	3	7	4	-1.22	17.74	-18.96
MAE (kcal/mol)						11.49

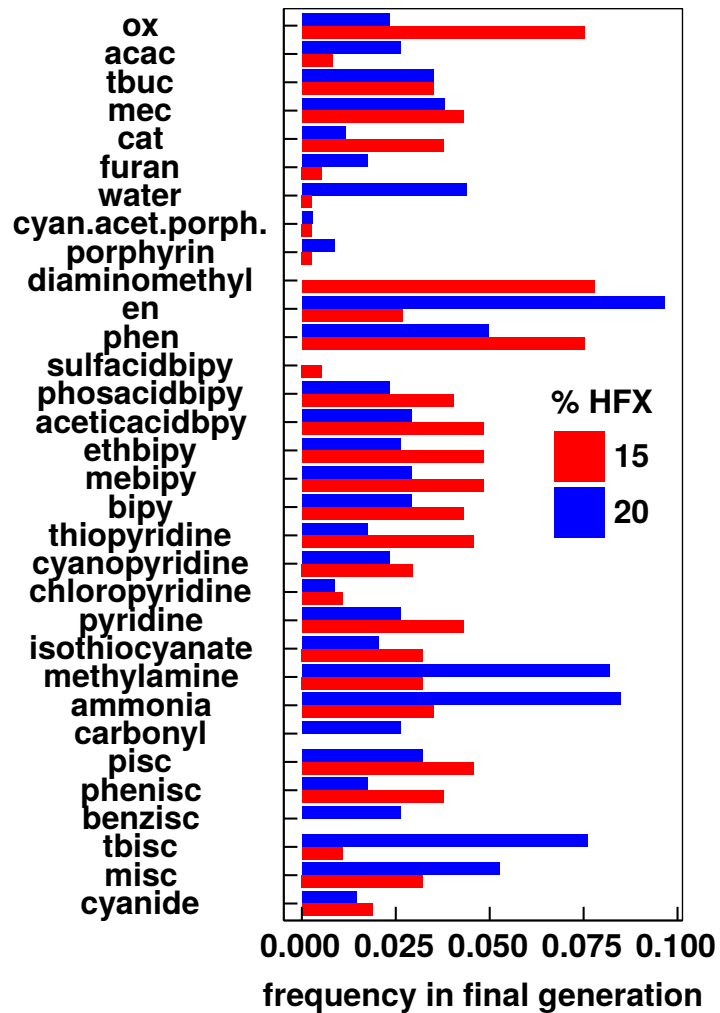


Figure C-11: Frequency of occurrence of ligands (in either axial or equatorial position) in ANN GA leads, defined as final population after 21 generations across 50 repeats with 'both' control, 15% and 20% exact exchange

Table C.7: Evaluation of high-spin to low-spin splitting with different basis sets. In all cases, metal atoms use the LANL2DZ effective core potentials.

metal	ox	eq. lig.	ax. lig.	$\Delta E_{\text{H-L}}$ (kcal/mol)		absolute change (kcal/mol)
				6-31g*	6-31g+*	
Cr	2	13	2	-3.52	-4.61	1.08
Co	2	28	14	1.35	1.91	0.56
Mn	3	28	20	-1.22	-2.14	0.92
Mn	3	28	24	-0.72	-1.47	0.74
Mn	2	11	2	4.69	4.22	0.47
Mn	3	28	23	-0.46	-1.21	0.75
Mn	3	5	16	9.51	10.76	1.26
Cr	2	16	2	-4.76	-5.76	1.01
Mn	3	2	8	7.93	6.97	0.96
Mn	3	2	13	5.63	6.57	0.95
Mn	3	28	21	-1.13	-1.11	0.02
Fe	2	28	27	-2.76	-3.62	0.85
Co	3	14	14	2.45	0.51	1.94
Co	2	28	3	0.54	-3.59	4.13
Mn	3	3	16	9.65	7.99	1.65
Cr	2	29	30	-3.22	-0.03	3.19
Mn	3	3	13	5.40	4.60	0.80
Mn	3	2	6	7.13	6.72	0.41
mean abs. difference (kcal/mol)						1.21





## Appendix D

# Uncertainty and extrapolation of ANNs for chemical discovery

### Text D.1: Details of ensemble and mc-dropout

Ensembles: One common approach to assign uncertainty estimates to predictions from data-driven models is to generate an ensemble of  $J$  different models. The mean of the predictions of these models is used as the predicted value at a new point and the variance in these predictions is used as a metric for model confidence. If  $x^*$  is a new trial point, and  $\sigma_{x^*}$  is the standard deviation associated with this prediction, the ensemble prediction is given as:

$$\bar{y}(x^*) = \frac{1}{n_{\text{ens}}} \sum_{j=1}^{n_{\text{ens}}} \hat{y}_j(x^*) \quad (\text{D.1})$$

with a variance of:

$$\sigma^2_{x^*} = \frac{1}{n_{\text{ens}}} \sum_{j=1}^{n_{\text{ens}}} (\bar{y}(x^*) - \hat{y}_j(x^*))^2 \quad (\text{D.2})$$

The prediction mean could be expected to have lower generalization error with respect to individual models. Typically, ensembles are generated by partitioning data to generate submodels, where each is trained on distinct subsets of data. Detection of uncertain points with ensemble models relies on the submodels being incorrect in different ways (i.e., high variance), which can occur when the model is evaluated for molecules dissimilar to training examples, where the behavior is only weakly constrained.

Monte-Carlo dropout: A lower cost framework for deriving uncertainty estimates for dropout regularized neural networks has recently been suggested<sup>459</sup> in analogy to Gaussian processes. In practice, this entails running the model  $J$  times with the dropout mask kept on, removing random nodes from the network each time. The average of these predictions are used as in the case with ensembles. The predictive uncertainty is estimated from:

$$\sigma^2_{x^*} = \frac{1}{J} \sum_{j=1}^J (\bar{y}(x^*) - \hat{y}_j(x^*))^2 + \tau^{-1} I \quad (\text{D.3})$$

This expression differs from the ensemble expression by also including a learned baseline uncertainty term,  $\tau^{-1}$ , which must be estimated from training data. In comparison to ensemble models, the cost of this approach is lower because the model only needs to be trained once. For mc-dropout, we determine a representative value of  $\tau$  by maximizing the log predictive likelihood of the corresponding GP based on the training data. This is a measure of how likely the observed data are under the GP, and is approximated<sup>459</sup> by

$$\log p(\mathbf{y}(\mathbf{x}_n) | \mathbf{x}_n, \mathbf{X}, \mathbf{Y}) \approx \log \left[ \sum_{j=1}^J e^{-\frac{1}{2}\tau \|\bar{\mathbf{y}}(\mathbf{x}_n) - \hat{\mathbf{y}}_j(\mathbf{x}_n)\|_2^2} \right] - \log J - \frac{1}{2} \log 2\pi - \frac{1}{2} \log \tau^{-1} \quad (\text{D.4})$$

In the application here (i.e., for the fully connected spin splitting neural network), we have scalar output and we use the training data to optimize equation D.4 with respect to  $\tau$  numerically. We use  $J = 100$  repeats, as in the network itself. The determined value of  $\tau$  based on the training data is  $3.6 \times 10^8$  in dimensionless units.

Table D.1: Ligand identity and occurrence among 654 unique metal-ligand combinations in the inorganic complex training set. Occurrence sums over all instances of the ligand in either axial site and the equatorial site. SMILES are given in the final column with the connection atom(s) shown in red.

	Ligand	Cumulative total	SMILES	Charge	Formula
1	misc	293	<b>C</b> [N]#[C]	0	C <sub>2</sub> H <sub>3</sub> N
2	water	292	<b>O</b>	0	H <sub>2</sub> O
3	carbonyl	275	<b>CO</b>	0	CO
4	pyr	267	<b>c</b> 1ccncc1	0	C <sub>5</sub> H <sub>5</sub> N
5	furan	168	<b>o</b> 1cccc1	0	C <sub>4</sub> H <sub>4</sub> O
6	ammonia	91	<b>N</b>	0	NH <sub>3</sub>
7	pisc	64	CC(C)(C) <b>C</b> 1=CC=C(C=C1)[N]#[ <b>C</b> ]	0	(CH <sub>3</sub> ) <sub>3</sub> CC <sub>6</sub> H <sub>4</sub> NC
8	isothiocyanate	57	[ <b>N</b> -]=C=S	-1	NCS <sup>-</sup>
9	cyanide	52	[ <b>C</b> -]#N	-1	CN <sup>-</sup>
10	en	42	<b>NCCN</b>	0	NCH <sub>2</sub> CH <sub>2</sub> N
11	acac	38	CC(= <b>O</b> )C=C( <b>-O-</b> )C	-1	C <sub>5</sub> H <sub>8</sub> O <sub>2</sub> <sup>-</sup>
12	chloride	36	<b>Cl</b>	-1	Cl <sup>-</sup>
13	phen	35	C1=CC2=CC=C3C=CC= <b>N</b> C3=C2N=C1	0	C <sub>12</sub> H <sub>8</sub> N <sub>2</sub>
14	ox	28	[ <b>O</b> -]C(=O)C([ <b>O</b> -])=O	-2	C <sub>2</sub> O <sub>4</sub> <sup>2-</sup>
15	tbuc	27	CC(C)(C) <b>C</b> 1=CC(=C([ <b>O</b> -])C=C1)[ <b>O</b> -]	-2	(CH <sub>3</sub> ) <sub>3</sub> CC <sub>6</sub> H <sub>3</sub> O <sub>2</sub> <sup>2-</sup>
16	bipy	26	<b>C</b> 1ccn <b>c</b> (c1)c2cccc <b>n</b> 2	0	C <sub>10</sub> H <sub>8</sub> N <sub>2</sub>
17	tbisc	22	[ <b>C</b> ]#[N]C(C)(C)C	0	(CH <sub>3</sub> )CCN
18	etesacac	21	<b>O</b> =C(C)/C(=C(\[O])\C)/CC(= <b>O</b> )OCC	-1	C <sub>9</sub> H <sub>13</sub> O <sub>4</sub> <sup>-</sup>
19	cat	18	[ <b>O</b> ]c1c(cccc1)[ <b>O</b> ]	-2	C <sub>6</sub> H <sub>4</sub> O <sub>2</sub> <sup>2-</sup>
20	met hylamine	18	<b>NC</b>	0	NH <sub>2</sub> CH <sub>3</sub>
21	phenacac	18	C1=CC=C(C=C1)C(= <b>O</b> )C C(= <b>O</b> )C2=CC=CC=C2	-1	(C <sub>6</sub> H <sub>5</sub> CO) <sub>2</sub> [CH] <sub>1</sub> <sup>-</sup>
22	phenisc	14	[ <b>C</b> ][N]c1cccc1	0	C <sub>6</sub> H <sub>5</sub> NC
23	pyrrole	12	C1=C[ <b>N</b> ]C=C1	-1	C <sub>4</sub> H <sub>4</sub> N <sup>-</sup>
24	cyanopyr	10	c1(ccnc1)C#N	0	NCC <sub>5</sub> H <sub>4</sub> N
25	benzisc	8	[ <b>C</b> ][N]Cc1cccc1	0	C <sub>6</sub> H <sub>5</sub> CH <sub>2</sub> NC
26	mebpy	8	<b>n</b> 1ccc(cc1 <b>n</b> cccc(c1)C)C	0	C <sub>12</sub> H <sub>12</sub> N <sub>2</sub>
27	porphyrin	7	[ <b>N</b> -]1C2=CC3= <b>N</b> C(=CC4=CC=C([ <b>N</b> -] ]4)C=C5C=CC(= <b>N</b> 5)C=C1C=C2)C=C3	-2	C <sub>20</sub> H <sub>12</sub> N <sub>4</sub> <sup>2-</sup>
28	ethbpy	4	<b>n</b> 1ccc(cc1 <b>n</b> cccc(c1)CC)CC	0	C <sub>14</sub> H <sub>16</sub> N <sub>2</sub>
29	phosacidbpy	4	<b>n</b> 1ccc(cc1 <b>n</b> cccc(c1)P(=O)(O)O)P(=O) (O)O	0	C <sub>10</sub> P <sub>2</sub> O <sub>6</sub> H <sub>10</sub>
30	aceticacidbpy	2	<b>n</b> 1ccc(cc1 <b>n</b> cccc(c1)CC(=O)O)CC(=O)O	0	C <sub>14</sub> H <sub>14</sub> O <sub>4</sub> N <sub>2</sub>
31	chloropyr	2	c1c(cnc1)Cl	0	ClC <sub>5</sub> H <sub>4</sub> N
32	mec	2	[ <b>O</b> -]c1c(cc(c1)C)[ <b>O</b> -]	2-	CH <sub>3</sub> C <sub>6</sub> H <sub>4</sub> O <sub>2</sub> <sup>2-</sup>
33	thiopyr	1	c1(cc <b>n</b> cc1)S	0	SC <sub>5</sub> H <sub>4</sub> N

## Text D.2: Simulation details for inorganic complexes and CSD test set

In this work, we primarily use 1901 spin splitting energies from DFT data sets generated over several prior works<sup>308,511,533,543</sup> to train new machine learning models. We also generate new DFT data on a 116-molecule CSD data set. We concisely summarize some of the details of these efforts here but refer the reader to the original work for more detail. 788 of the compounds are from Ref.<sup>308</sup>, 286 of the compounds are from Ref.<sup>511</sup>, 19 of the compounds are from Ref.<sup>543</sup>, 87 of the compounds had revised spin states first published in Ref.<sup>533</sup>, and 721 of the compounds had not been previously published, including revised spins for compounds from previous sets.

Despite originating from several original sources, a consistent workflow has been employed, with distinctions noted as follows. The molSimplify<sup>181</sup> toolkit was used to generate octahedral transition metal complex structures from a pool of organic ligands common in inorganic chemistry (listed in Table D.1) with enforced equatorial symmetry but allowing up to two distinct axial ligands. DFT geometry optimizations were then carried out using TeraChem<sup>105</sup> with the B3LYP hybrid DFT functional, varying the fraction of Hartree-Fock (HF) exchange from its default 20% value in 5% increments over the range of 0-30% HF exchange. Thus, the 1901 data points corresponds to 564 unique chemical structures, with additional repeats at varied exchange fractions. The LANL2DZ effective core potential was employed for transition metals and heavy elements (i.e., Br) with the 6-31G\* basis for all other atoms. The effect of using a modest basis set, which enables larger data set generation for ML models, was found to be limited in prior work on the relative energies of interest<sup>482</sup>. The metals studied throughout were Cr, Mn, Fe, and Co in M(II) and M(III) oxidation states. The high-spin/low-spin definitions used to calculate the adiabatic electronic energy spin splitting were: quintet-singlet for both  $d^4$  Mn(III)/Cr(II) and  $d^6$  Co(III)/Fe(II), sextet-doublet for  $d^5$  Fe(III)/Mn(II), and quartet-doublet for both  $d^3$  Cr(III) and  $d^7$  Co(II). These spin states are a revision from initial work<sup>308</sup> that employed a triplet ground state for Cr(II) and Mn(III).

All open-shell complexes (i.e., all non-singlets) are treated with spin-unrestricted DFT with virtual and occupied orbitals level-shifted<sup>448</sup> by 1.0 and 0.1 Ha. respectively, to aid convergence to an unrestricted solution. Geometry optimizations were conducted for 788 cases with DL-FIND<sup>449</sup> in Cartesian coordinates. The protocol was shifted to employ the TRIC (translation rotation internal coordinates)<sup>498</sup> optimizer for the 1113 most recent cases. Both optimizers are available in TeraChem, and the same default tolerances were employed of  $4.5 \times 10^{-4}$  hartree/bohr for the maximum gradient and  $1 \times 10^{-6}$  hartree for the change in energy between steps.

Prior to their use in model training, structures are filtered and removed if they fail metrics of quality geometries we recently introduced<sup>533</sup>. Specifically, these metrics include preserved coordination number of 6 with reasonable bond lengths and no ligand distortions. Additionally we removed any complexes with large (i.e.,  $1.0 \mu_B$  or larger) deviation of  $\langle S^2 \rangle$  from the expected value based on the assigned spin.

For the CSD data set, we searched for diverse octahedral transition metal complexes with M(II)/M(III) M = Cr, Mn, Fe, or Co transition metals. For the geometry optimization, the same method, basis, and optimization approach was followed. Geometry checks and  $\langle S^2 \rangle$  deviations were used to eliminate structures. Additionally, we manually screened the collected points to exclude any that were duplicates within each other as judged through comparable connectivity but differing accession codes. We also removed those that were duplicates of data in the original data set, as judged through the assigned connectivity in RAC-155. As an additional constraint, we filtered out any complexes with evidence of ligand non-innocence. Specifically, we computed the Mulliken spin of the metal center and discarded complexes with Mulliken spin that was more than  $1.0 \mu_B$  less than the expected spin from the overall spin assigned to the complex.

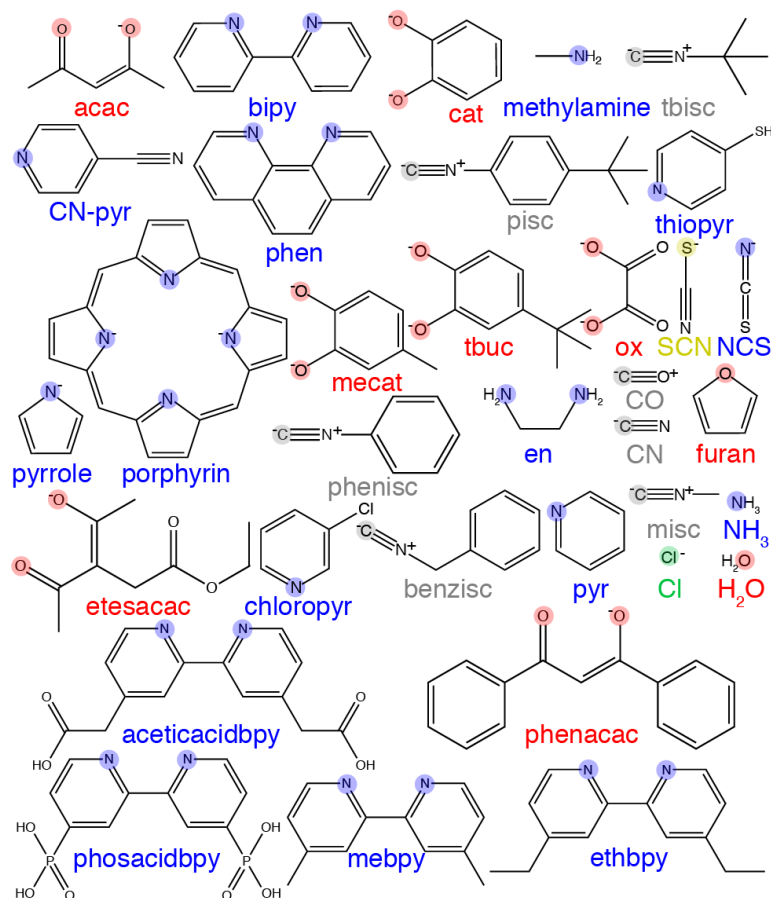


Figure D-1: Ligands used to train inorganic complex spin splitting ANN, metal connection atoms highlighted, with the highlight corresponding to the element: oxygen in red, nitrogen in blue, chlorine in red, carbon in gray, and sulfur in yellow. Charges are also shown on relevant atoms.

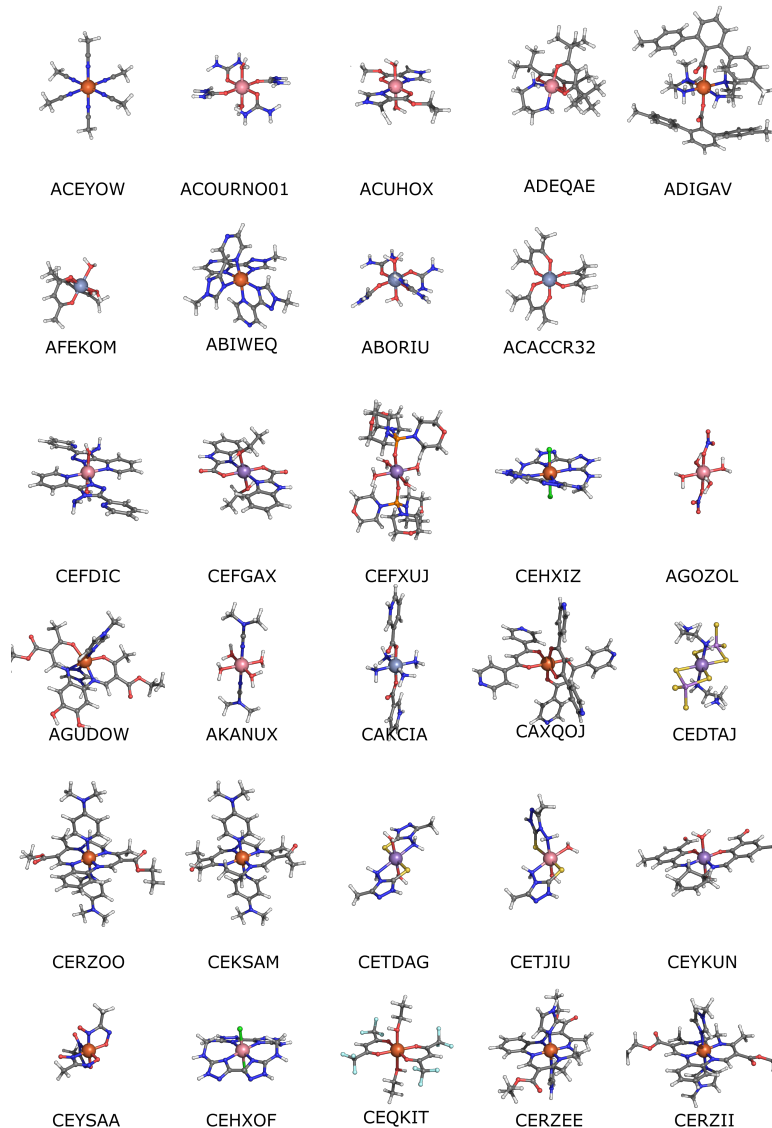


Figure D-2: Visualization of CSD structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

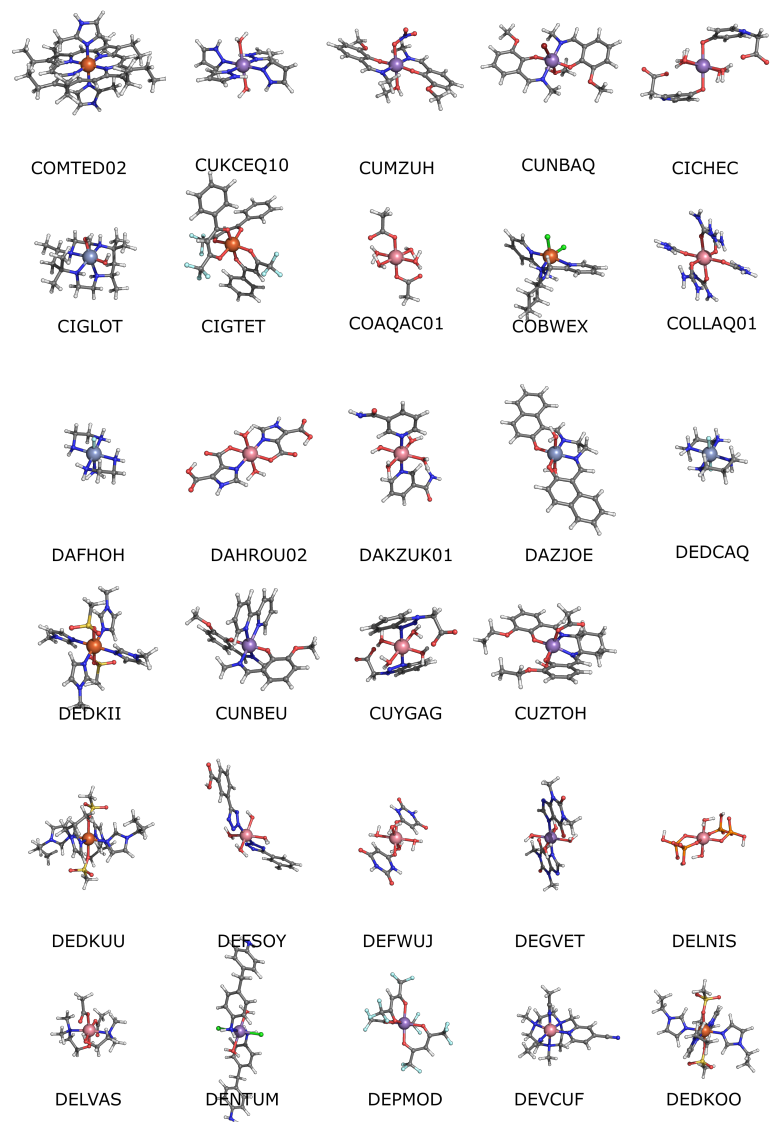


Figure D-3: Visualization of CSD structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

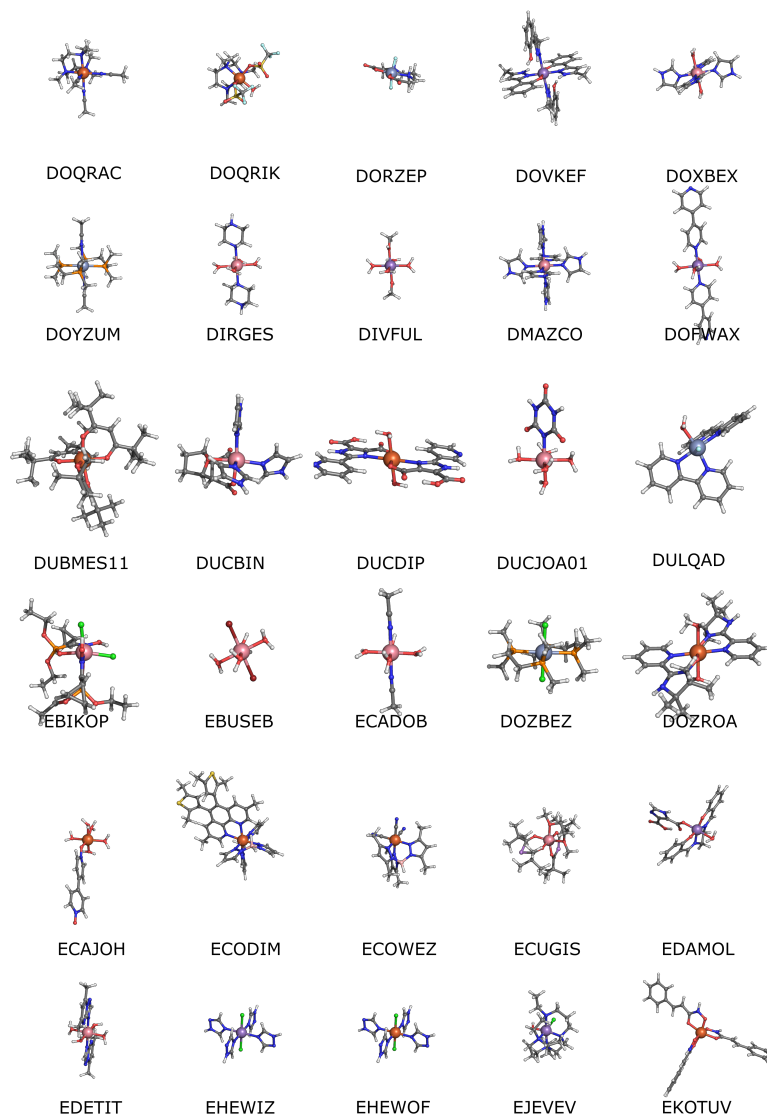


Figure D-4: Visualization of CSD structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.



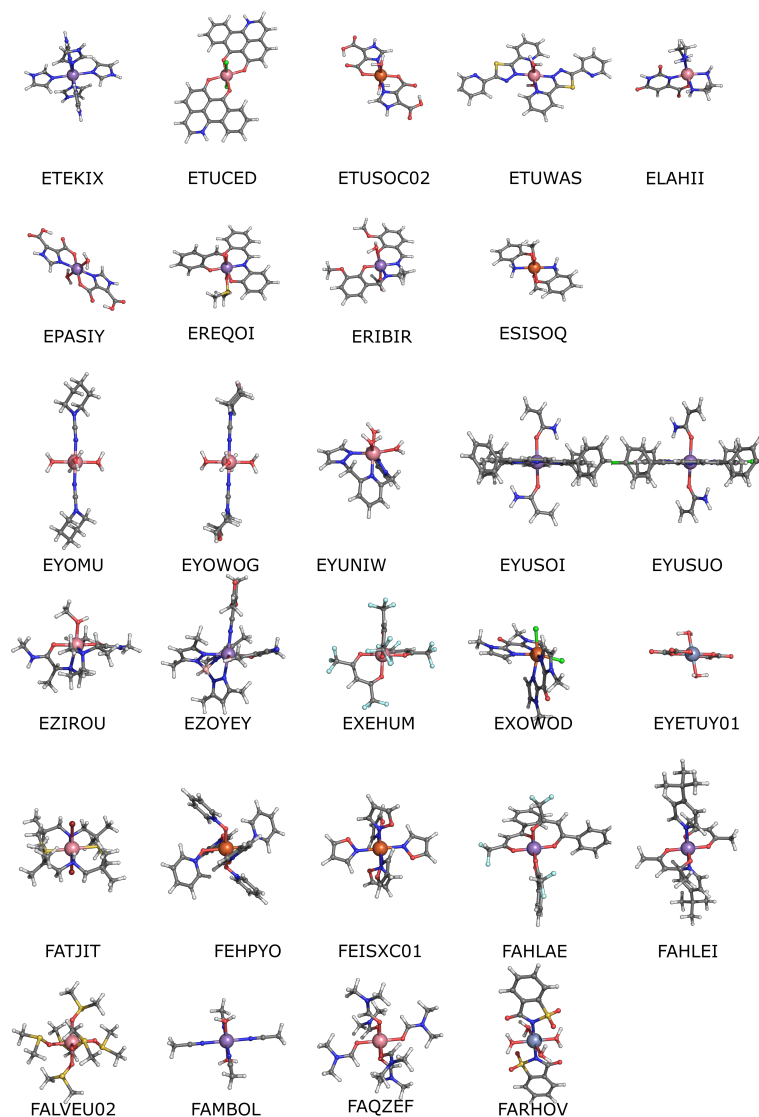


Figure D-5: Visualization of CSD structures used in this work at DFT-optimized ground spin states. CSD accession codes shown below each structure. Non-metal atoms are colored as follows: carbon is gray, hydrogen is white, nitrogen is blue, oxygen is red, chlorine is green, bromine is rust, fluorine is cyan, sulfur is yellow, phosphorous is orange, boron is pink and arsenic is purple. Metal centers are shown as large spheres and colored as follows: iron is orange, manganese is purple, cobalt is pink and chromium is metallic blue.

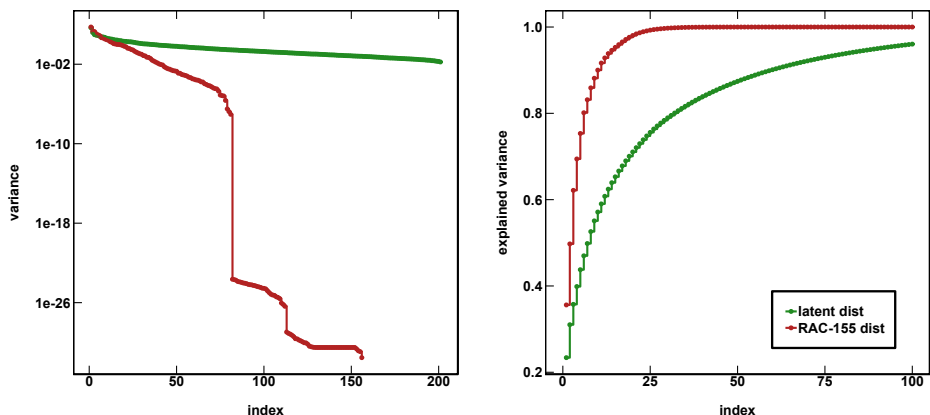


Figure D-6: Decay of variance (left) and cumulative relative explained variance of dimensions from principal component analysis of 1901 inorganic training points with RAC-155 representation (red) and the final model latent space (green).

Table D.2: Mean absolute error (MAE) and root-mean square error (RMSE) metrics for inorganic spin splitting ANN on training data and out of sample CSD prediction task. Errors are shown from a single model, the average of an ensemble of 10 models and the average of 100 Monte-Carlo dropout realizations of the single model. All error units are kcal/mol. This performance is comparable to a similar test in which we trained on 1400 transition metal complexes with the MCDL-25 descriptor set in a 2-hidden layer ANN. In that work<sup>308</sup>, we studied a set of 35 CSD test structures. In those cases, we observed an increase from 2.5 kcal/mol test set MAE to 9.78 kcal/mol MAE and 13.26 kcal/mol RMSE on the 35 CSD test structures.

model	training MAE	CSD MAE	CSD RMSE
		(kcal/mol)	
single ANN	1.52	8.55	13.61
10-model ensemble	-	8.95	14.76
100-model mc-dropout	-	8.53	13.45

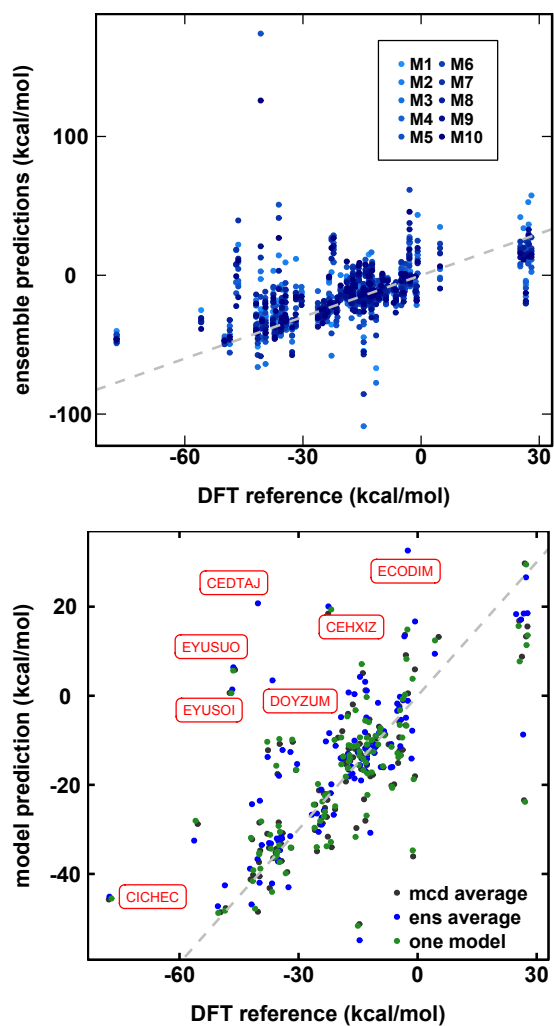


Figure D-7: Parity plots of DFT-calculated splitting energy of CSD structures and predictions from a 10-model ensemble (left) and a single model (green), the average of the 10-model ensemble (blue) and the average of 100 mc-dropout realizations (charcoal) compared (right). The parity line is shown as a dashed gray line, while the CSD codes for high error ( $\geq 30$  kcal/mol) points are shown in red. All units are kcal/mol.

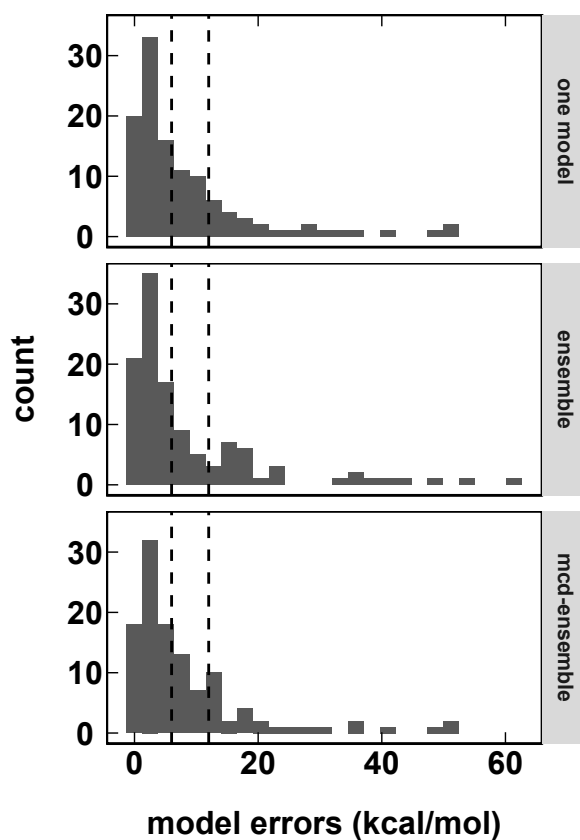


Figure D-8: Distribution of a errors for CSD prediction task from a single model, the average of an ensemble of 10 models and the average of 100 Monte-Carlo dropout realizations of the single model. Dashed vertical lines show nominal tolerances of 6 and 12 kcal/mol. All error units are kcal/mol.

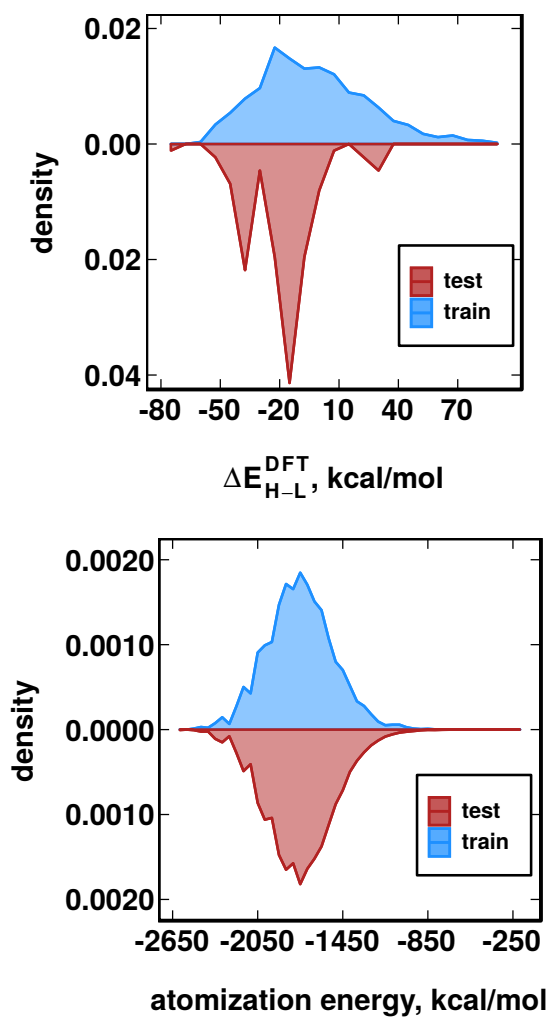


Figure D-9: Comparison of train and test distributions for inorganic spin splitting task with 1900 training points and CSD test data (top) and QM9 atomization energy task with uniform random 5% training data and the remaining 95% used as test (bottom). All units are kcal/mol.

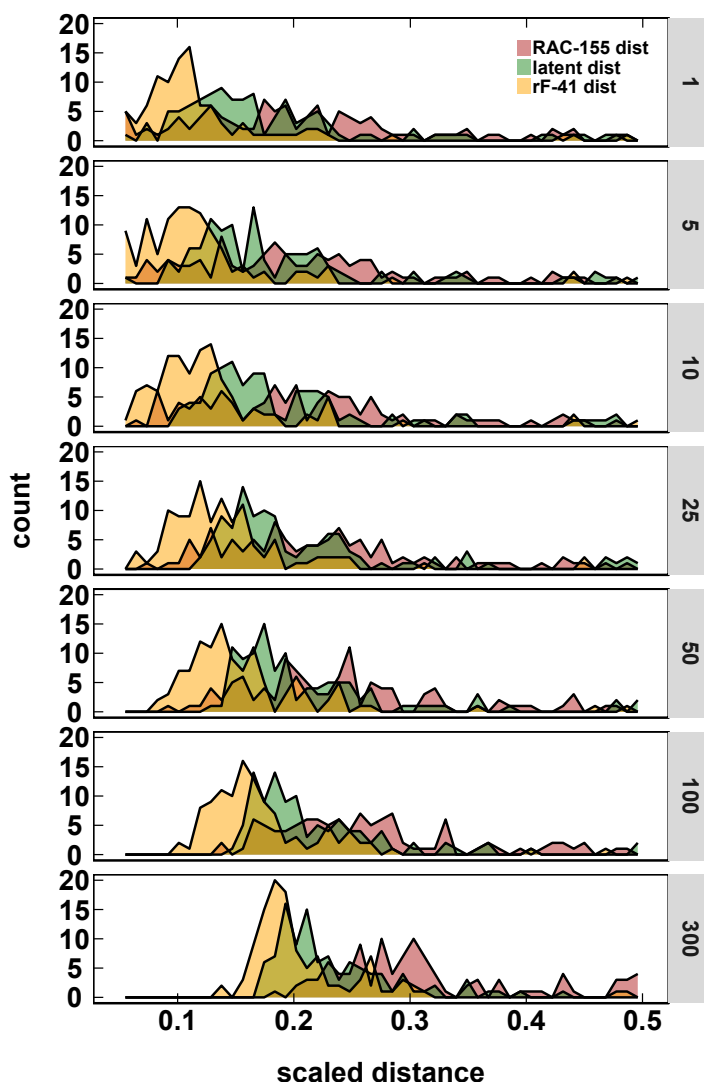


Figure D-10: Distribution of average distance to nearest training data as a function of number of neighbors over which the distance is averaged for 1 to 300 neighbors (as labeled on each graph) for the CSD prediction task, showing three different distance metrics: RAC-155, random forest 41-feature subset of RAC-155 (rF-41), and latent. Distances are normalized to  $[0, 1]$  for comparison and truncated to the region  $[0, 0.5]$ . Similarity of complexes in feature space (e.g., the simple Euclidean distance in feature space or a cheminformatic similarity metric such as the Tanimoto distance) can be measured to the nearest training point or averaged over multiple training points. Using nearest neighbor data only is likely sensitive to outlier training data, whereas using all training data will likely overestimate distances for new molecules supported by a relatively small amount of training data. Although we previously found good success in both using a single nearest neighbor or over 5-10 nearest neighbors, we now compare potential effects of nearest neighbor averaging on distance distributions. Feature space distances may not be a good proxy for chemical similarity and this approach also ignores automatic feature-engineering that occurs in complex models (e.g., multi-layer neural networks). Furthermore, high-dimensional feature spaces may contain weakly informative features that can "pollute" isotropic distance metrics.

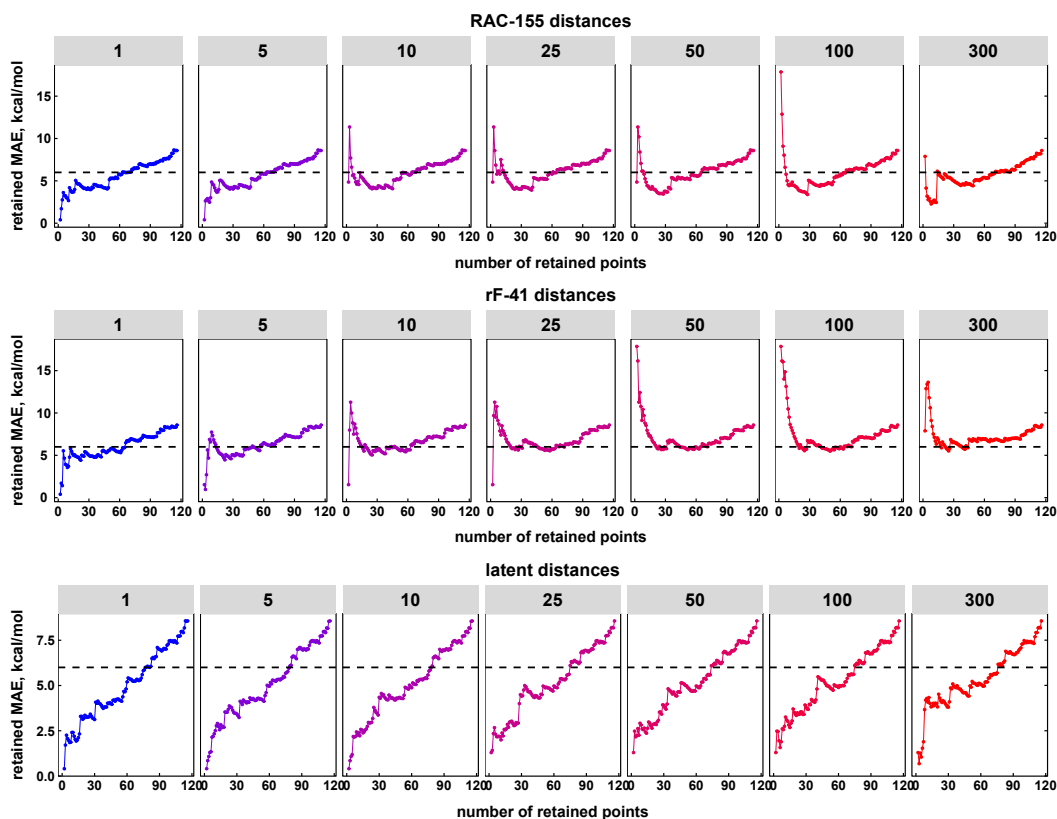


Figure D-11: Mean absolute spin splitting error (MAE) as a function of number of retained points for thresholds set using different distances: RAC-155 (top), 41-feature subset of RAC-155 selected with random forest (i.e., rF-41, middle), and latent (bottom), averaged over different numbers of nearest neighbors from 1 to 300 (in panels). Depending on how conservatively the boundary between trustworthy chemical space and untrustworthy chemical space is set, we include more or less test data. We therefore consider using each distance and the number of neighbors it is averaged over as a decision boundary and examine how error of retained points varies. Using feature space distances, the effect of nearest neighbors used in the average is most significant for highly conservative decisions that retain less than 20 of the 116 CSD cases. Feature space distances are generally poor at effectively classifying low error points. For intermediate data retention, feature-space-derived models are less sensitive to number of nearest neighbors and general in agreement with each other. Latent space distance shows the least nearest neighbor dependence. Distances are normalized to  $[0, 1]$  for comparison. The horizontal black line represents a nominal error tolerance of 6 kcal/mol.

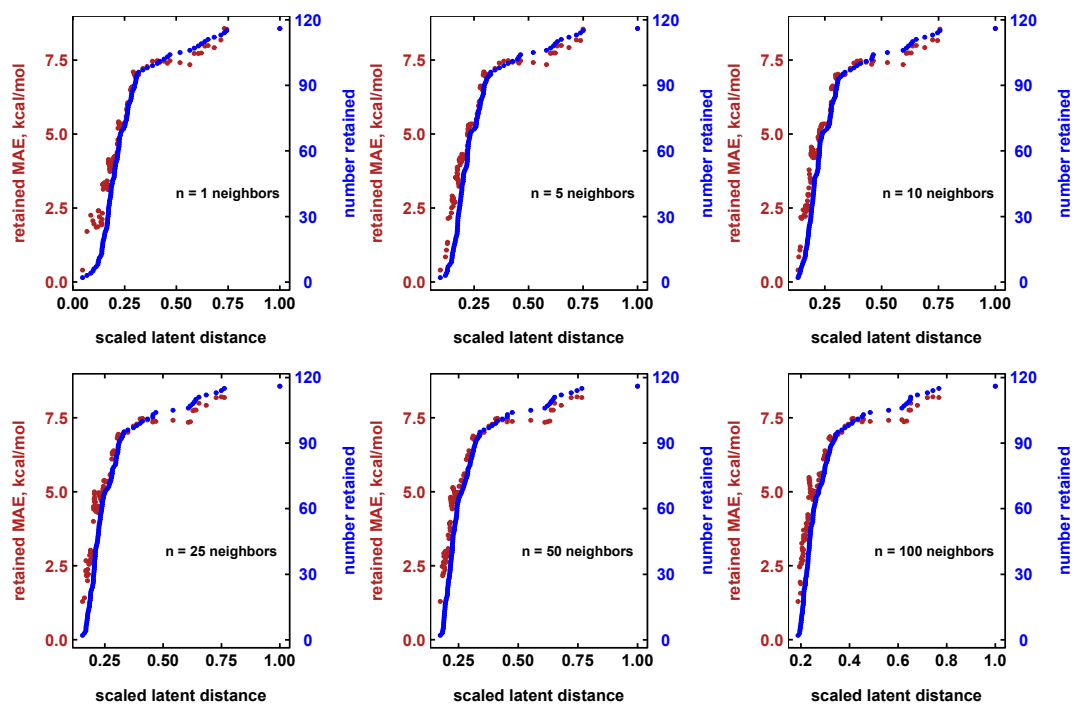


Figure D-12: Mean absolute spin splitting error (MAE) on retained data and number of retained candidates as a function of threshold latent distance to nearest training points, averaged over 1 to 100 nearest neighbors. Distances are normalized to  $[0, 1]$  for comparison.



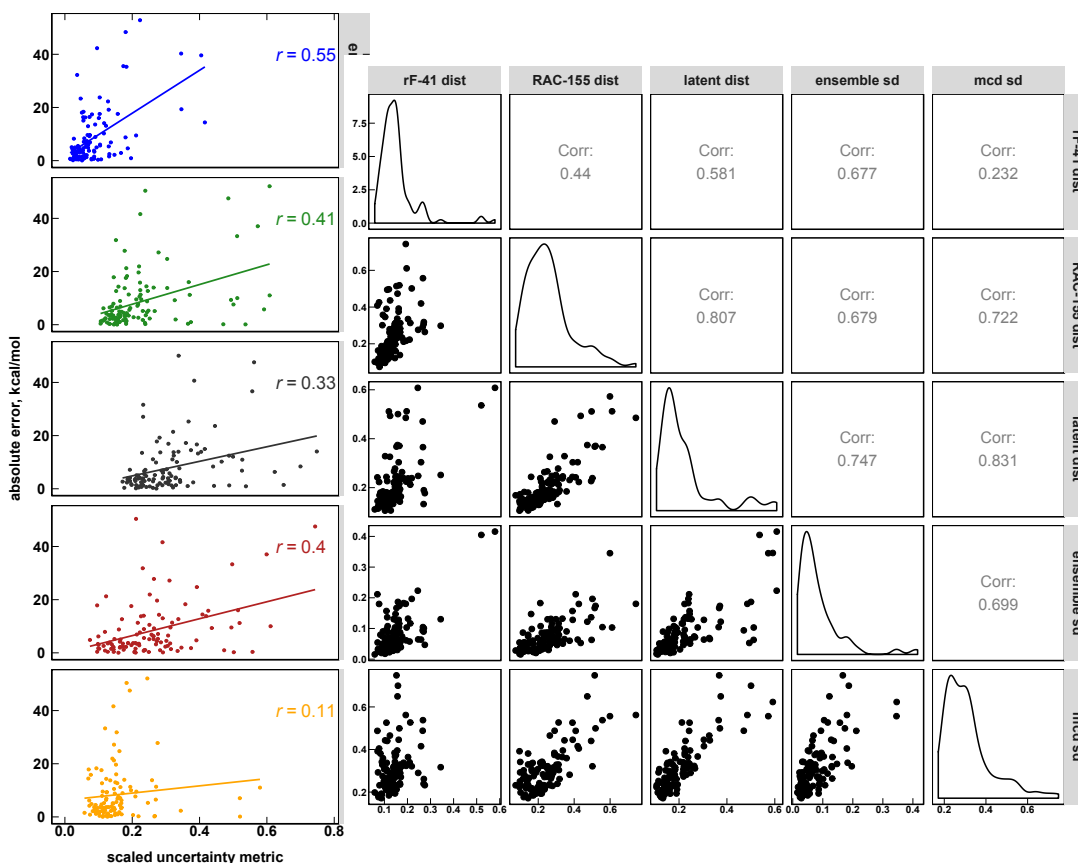


Figure D-13: Correlation between different uncertainty metrics (panels) and absolute model errors on CSD data (left), showing the correlation coefficient inset along with best fit line and (right) showing all pairwise cross-correlations and distributions of uncertainty metrics. Metrics shown are the standard deviations from 10-model ensemble, 10-neighbor average latent distance, standard deviation of of 100 mc-dropout realizations, 10-neighbor average feature space distance using RAC-155 and rF-41 representations. All units are kcal/mol and all metrics are normalized to  $[0, 1]$  for comparison. We truncate the plot at 0.75 to remove the few outlying points at extreme distances for clarity, excluding 1 ensemble point, 1 latent distance point, 7 mc-dropout points, 6 RAC-155 distance points, and 2 rF-41 points.

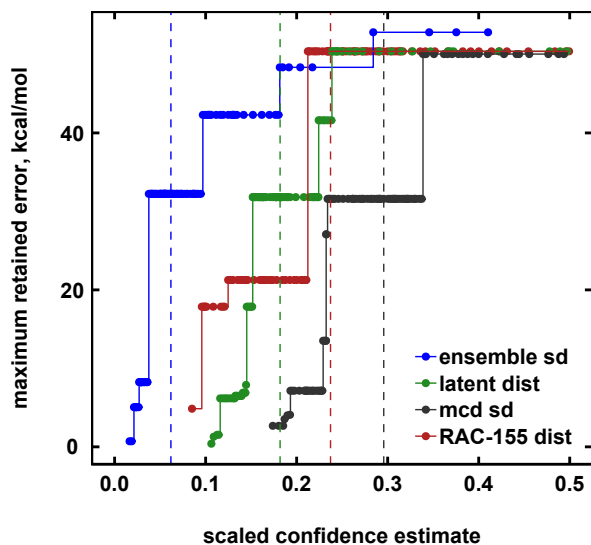


Figure D-14: Variation in the maximum ANN error (in kcal/mol) for retained points on CSD data as a function of thresholds in different uncertainty metrics, showing that the largest errors can be effectively avoided by truncating with respect to latent distance and ensemble metrics but not raw distances. Compared metrics are the 10-neighbor average distance to training data in both feature (RAC-155) and latent spaces, the standard deviation of a 10-model ensemble and the standard deviation of a 100 realizations of a mc-dropout ensemble. All metrics are normalized to  $[0, 1]$  for comparison. Vertical lines indicate the median of each scaled metric.

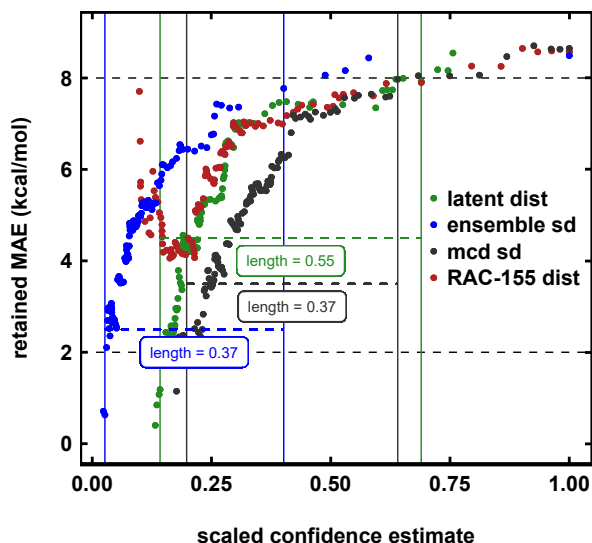


Figure D-15: Variation in the mean absolute error (MAE, in kcal/mol) from ANN models for retained points on CSD data as a function of thresholds in different uncertainty metrics, showing that 1) average retained errors can be controlled with all metrics and 2) different metrics show different sharpness in response to changing thresholds, as indicated by the annotation showing the length of the interval from MAE= 2 kcal/mol to MAE= 8 kcal/mol with a horizontal line. The interval for each model is also marked with solid vertical lines. Metrics compared are the 10-neighbor average distance in both feature (RAC-155) and latent spaces, the standard deviation of a 10-model ensemble and the standard deviation of 100 realizations of a mc-dropout ensemble. All metrics are normalized to  $[0, 1]$  for comparison. Annotation is not provided for RAC-155 owing to non-monotonic behavior at low thresholds.

Table D.3: CSD accession codes for points used to calibrate latent-distance uncertainty model.

ABORIU	ADEQAE	AGUDOW	CAKCIA	CEFDIC
CERZEE	CICHEC	CIGTET	COAQAC01	COMTED02
DEDKII	DEFWUJ	DUCBIN	EBUSEB	ECADOB
ECOWEZ	EKOTUV	ELAHII	EZIROU	FEHPYO

Table D.4: Values for  $\sigma_1$  and  $\sigma_2$  in latent-distance uncertainty model calibrated using maximum likelihood estimation on 5 different random samples of 20 CSD points. The bold values in the first row indicate those used in the rest of this work, corresponding to accession codes given in Table D.3.

repeat	$\sigma_1$ (kcal/mol)	$\sigma_2$
<b>1</b>	<b><math>4.57 \times 10^{-9}</math></b>	<b>3.20</b>
2	$2.24 \times 10^{-8}$	3.12
3	$1.61 \times 10^{-8}$	2.95
4	$8.93 \times 10^{-9}$	3.22
5	$2.58 \times 10^{-8}$	4.16

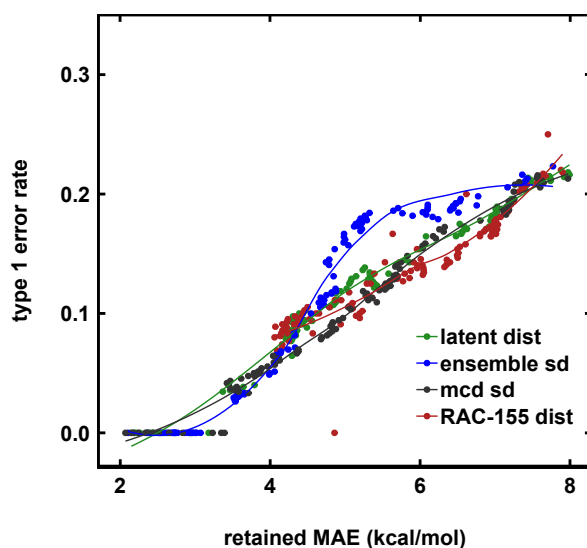


Figure D-16: Comparison of type I error rate, defined as the fraction of retained points with absolute errors  $> 12$  kcal/mol from ANN models, as a function of the mean absolute retained error when setting thresholds in different uncertainty metrics. Compared metrics are the 10-neighbor average distance to training data in both feature (RAC-155) and latent spaces, the standard deviation of a 10-model ensemble and the standard deviation of 100 realizations of mc-dropout. A smoothing spline is shown for each metric. Higher error rates are observed for 10-model ensemble for retained MAEs between 5 and 7 kcal/mol.

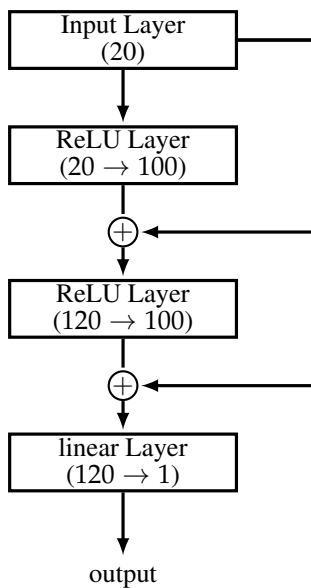


Figure D-17: Neural network architecture used for QM9 prediction task, showing two fully-connected layers with input pass-through connections. The size of each mapping is shown in parentheses under the layer name. The  $\oplus$  symbol represents concatenation. Dropout and batch normalization are applied to the ReLU layers.

Table D.5: Hyperparameters and topology for organic atomization energy ANN on QM9 benchmark.

parameter	value
layer 1 size	100
layer 2 size	100
activation function	relu
learning rate	0.00033
optimizer	adam
$\beta_1$	0.9945
$\beta_2$	0.9936
decay	0
dropout (all hidden)	0.053
batch size	128
epochs	800
$L^2$ regularization	1.32818E-8
semibatch normalization	yes
early stopping	none

Table D.6: Comparison of single-model performance of QM9 atomization ANNs with two hidden layers of 100 nodes, using residual architecture (original), without residual links at the same hyperparameters and without residual links after reoptimizing hyperparameters using hyperopt. The reoptimized hyperparameters are the same as in Table D.5 except for learning rate = 0.00196,  $\beta_1 = 0.9694$ ,  $\beta_2 = 0.9779$ , decay= 0,  $L^2$  regularization =  $2.33317 \times 10^{-9}$ .

model	training RMSE	test MAE	test RMSE
		(kcal/mol)	
original	6.24	6.79	9.97
no residual links	18.32	15.28	19.60
hyperparameter reoptimized	6.97	8.58	11.80

Table D.7: Repetition test showing train and test mean absolute errors (MAE) for atomization energy prediction on QM9 data using 100 different 5% training data samples. In all cases, all points not in the training set are used as test. Average and standard deviations are given at the end of the table.

	train MAE (kcal/mol)	test MAE (kcal/mol)		train MAE (kcal/mol)	test MAE (kcal/mol)		train MAE (kcal/mol)	test MAE (kcal/mol)		train MAE (kcal/mol)	test MAE (kcal/mol)
0	4.84	7.06	1	4.83	7.10	2	4.60	6.95	3	4.68	6.84
4	4.37	6.73	5	4.44	6.76	6	4.44	6.89	7	4.77	7.11
8	4.17	6.75	9	4.57	6.71	10	4.55	6.9	11	4.82	6.94
12	4.44	6.80	13	4.53	6.79	14	4.59	7.04	15	4.55	6.92
16	4.80	7.15	17	4.84	7.0	18	4.32	6.72	19	4.69	7.03
20	4.39	6.77	21	5.00	7.08	22	4.98	7.0	23	5.05	6.94
24	4.52	6.93	25	4.62	6.94	26	4.63	6.79	27	4.34	6.55
28	4.54	6.96	29	4.50	6.79	30	4.99	7.03	31	4.30	6.69
32	4.56	7.02	33	4.70	6.83	34	4.78	6.87	35	4.50	6.69
36	4.59	7.02	37	4.27	6.78	38	4.43	6.96	39	4.34	6.84
40	4.59	6.96	41	4.83	6.87	42	4.46	6.68	43	4.82	7.13
44	4.59	6.99	45	4.72	6.84	46	4.38	6.7	47	4.63	7.07
48	4.52	6.97	49	4.81	6.93	50	4.49	6.95	51	4.41	6.73
52	4.38	6.81	53	5.57	6.95	54	4.30	6.67	55	5.08	7.05
56	4.35	6.81	57	4.80	7.02	58	4.65	6.96	59	4.32	6.78
60	4.41	6.87	61	4.73	7.20	62	4.76	7.21	63	4.77	7.08
64	4.23	6.63	65	4.76	6.86	66	4.42	6.82	67	4.49	6.91
68	4.77	6.96	69	4.41	6.81	70	5.06	7.19	71	4.85	7.23
72	4.54	6.78	73	4.52	6.80	74	4.57	6.96	75	4.56	7.00
76	4.46	6.77	77	4.99	7.11	78	4.63	6.79	79	4.26	6.84
80	4.63	7.01	81	4.48	6.75	82	4.55	6.86	83	4.68	6.90
84	4.31	6.60	85	4.42	6.85	86	4.53	6.86	87	4.30	6.68
88	4.34	6.84	89	4.24	6.65	90	4.33	6.58	91	4.57	6.85
92	4.34	6.77	93	4.54	7.07	94	4.36	6.8	95	4.65	7.07
96	4.56	6.79	97	4.73	6.85	98	4.66	6.89	99	5.08	7.10
average train MAE = 4.59 kcal/mol						average test MAE = 6.89 kcal/mol					
sd train MAE = 0.23 kcal/mol						sd test MAE = 0.15 kcal/mol					

Table D.8: Mean absolute error (MAE) and root-mean square error (RMSE) metrics for QM9 atomization energy ANN trained on a random 5% of data tested on the remaining 127217 points. Errors are shown from a single model and the average of an ensemble of 10 models. All error units are kcal/mol.

model	training RMSE	test MAE	test RMSE
	(kcal/mol)		
single ANN	6.24	6.79	9.97
10-model ensemble	-	6.13	9.14

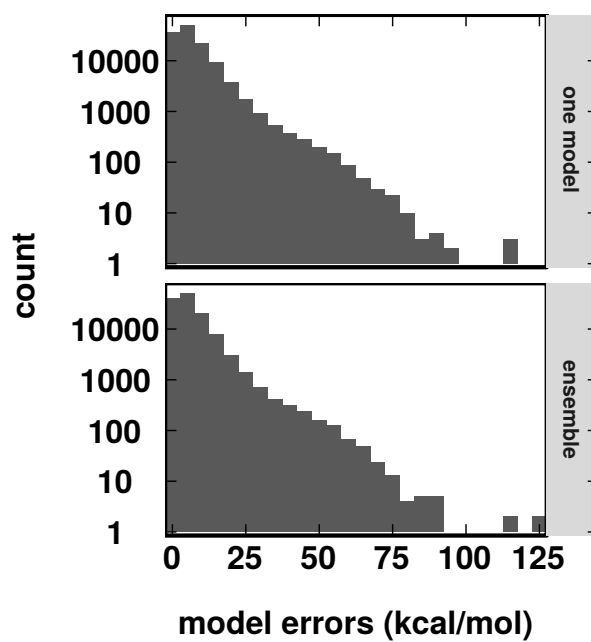


Figure D-18: Distribution of a errors for QM9 atomization prediction task from a single model and the average of an ensemble of 10 models. All error units are kcal/mol and counts are shown on a log y-axis. The maximum error for a single model is 119.97 kcal/mol and 124.10 kcal/mol for the ensemble model. These large errors are observed on the SMILES strings FC(F)(F)CC(F)(F)F (hexafluoropropane) and CC1N2C3C4=CCC13C24 (a cyclic tertiary amine), respectively.



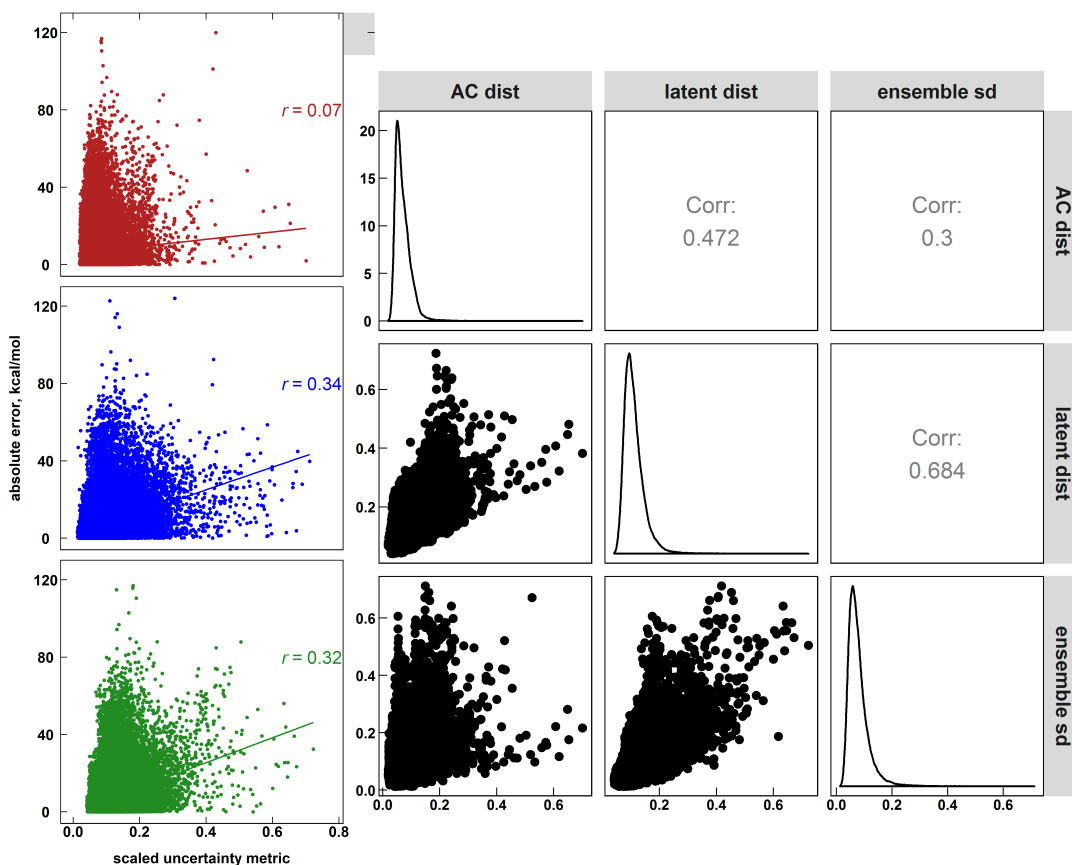


Figure D-19: Correlation between different uncertainty metrics (panels) and absolute model errors on QM9 atomization energy test data (left), showing the correlation coefficient inset along with best fit line and (right) showing all pairwise cross-correlations and distributions of uncertainty metrics. Metrics shown are the standard deviations from 10-model ensemble and 10-neighbor average latent distance and 10-neighbor average feature space distance using AC representations. All units are kcal/mol and all metrics are normalized to  $[0, 1]$  for comparison. We truncate the plot at 0.75 to remove the few outlying points at extreme distances for clarity, excluding 8 ensemble point, 3 latent distance point and 6 AC distance points.

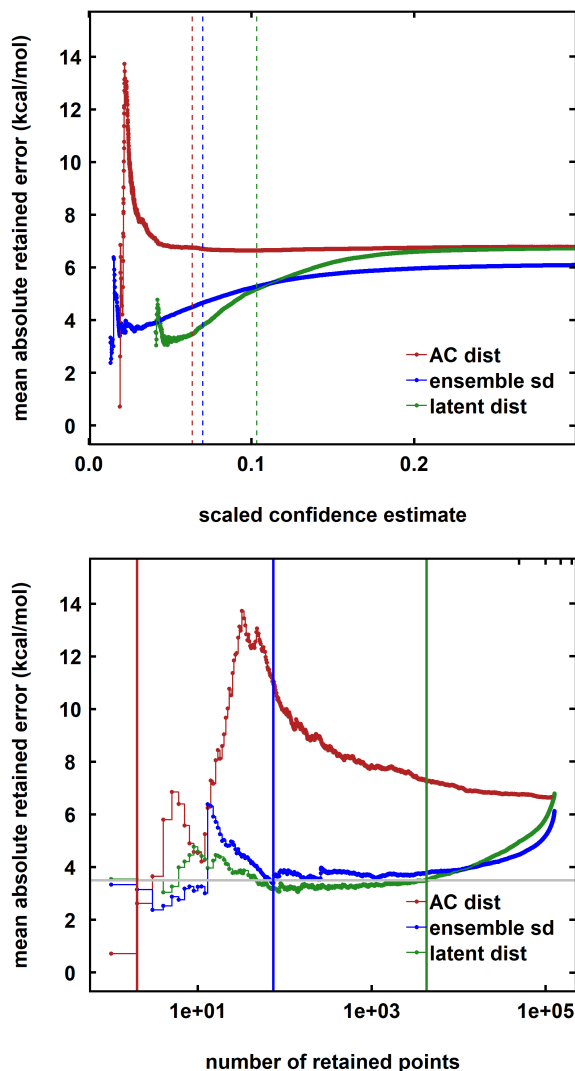


Figure D-20: Variation in the mean absolute error (MAE, in kcal/mol) on retained points from ANN models on QM9 atomization energy data as a function of (left) the thresholds in different uncertainty metrics and (right) the number of points retained. The metrics compared are the 10-neighbor average distance in both feature (AC) and latent spaces and the standard deviation of a 10-model ensemble. We also plot the maximum number of retained points before the retained MAE (right) crosses a 3.5 kcal/mol threshold (horizontal gray line) with solid vertical lines at 2, 74 and 4299 points for AC distances, ensembles and latent distances respectively. All metrics are normalized to  $[0, 1]$  for comparison. Dashed vertical lines show the median of each metric (left).

Table D.9: Values for  $\sigma_1$  and  $\sigma_2$  in latent-distance uncertainty model calibrated using maximum likelihood estimation on different numbers of random samples of QM9 test points. For each number of points, we present the mean and standard deviation over 10 random samples. The bold values in the last row indicate the single sample with 500 points used in the rest of this work. Thus, the conclusion is that only 500 points from  $> 120k$  are needed to calibrate parameters, indicating the proposed model learns this mapping easily from sparse data.

# of points	$\sigma_1$		$\sigma_2$	
	mean	std (kcal/mol)	mean	std
100	0.325	0.0053	15.70	0.528
500	$1.71 \times 10^{-7}$	$3.43 \times 10^{-7}$	4.54	0.313
1000	$1.39 \times 10^{-7}$	$2.20 \times 10^{-7}$	4.40	0.120
5000	$8.08 \times 10^{-7}$	$1.28 \times 10^{-6}$	4.41	0.0787
10000	$2.04 \times 10^{-7}$	$3.23 \times 10^{-7}$	4.45	0.0446
25000	$6.93 \times 10^{-7}$	$1.49 \times 10^{-6}$	4.45	0.0284
50000	$7.02 \times 10^{-7}$	$1.84 \times 10^{-6}$	4.45	0.0289
100000	$3.36 \times 10^{-7}$	$6.61 \times 10^{-7}$	4.46	0.011
<b>500</b>		<b><math>1.79 \times 10^{-6}</math></b>		<b>4.45</b>

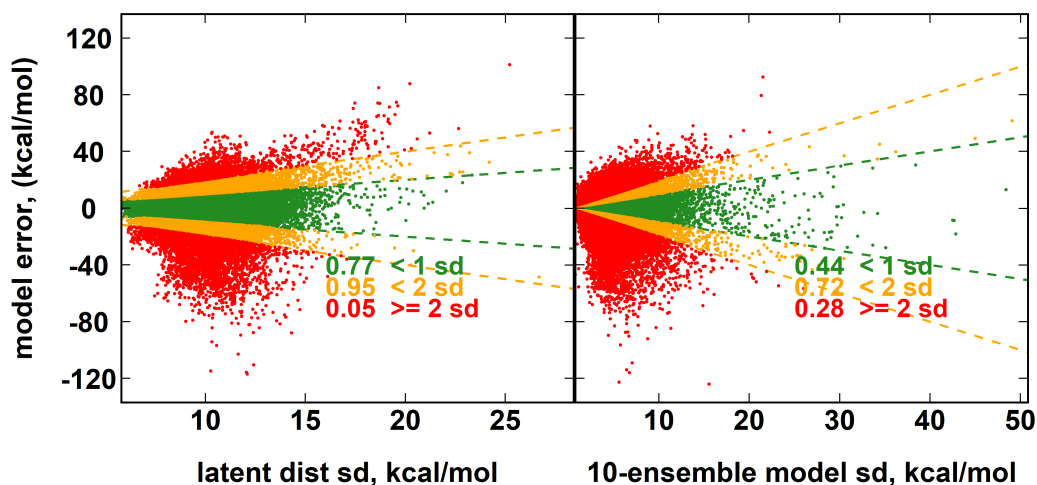


Figure D-21: Relationship between model errors and different uncertainty metrics for QM9 atomization energy ANN on test-set points. Standard deviations from calibrated latent distance model using 500 points (left) and 10-model ensemble (right) are compared, with points lying in one (two) sd colored green (yellow). Points outside two sd are colored red. Dotted lines indicate one and two standard deviations. Error units are kcal/mol.

Table D.10: CSD active learning experiment: mean absolute error (MAE) and root-mean square error (RMSE) metrics for out-of-sample CSD prediction task with a single model for the original 116 points, after removing the 10 points with lowest model confidence determined with different metrics and then after retraining with the 10 excluded points. Uncertainty metrics compared are the latent distance of 10-model ensemble metrics. All errors are in kcal/mol.

original single ANN		data selection method	10 removed		retrained	
MAE	RMSE		MAE	RMSE	MAE	RMSE
(kcal/mol)			(kcal/mol)			
8.55	13.61	latent distance	7.73	12.22	7.10	10.62
		10-model ensemble	7.61	11.65	7.56	10.87
		mcd-ensemble	7.57	11.79	7.46	11.39

Table D.11: Hyperparameters and topology for inorganic spin splitting ANN.

parameter	value
layer 1 size	200
layer 2 size	200
layer 3 size	200
activation function	relu
learning rate	0.00163
optimizer	sgd
momentum	0.998
decay	0.0015719
Nesterov acceleration	yes
dropout (all hidden)	0.0825
batch size	128
epochs	2000
$L^2$ regularization	7.101148E-14
semibatch normalization	yes
early stopping	none

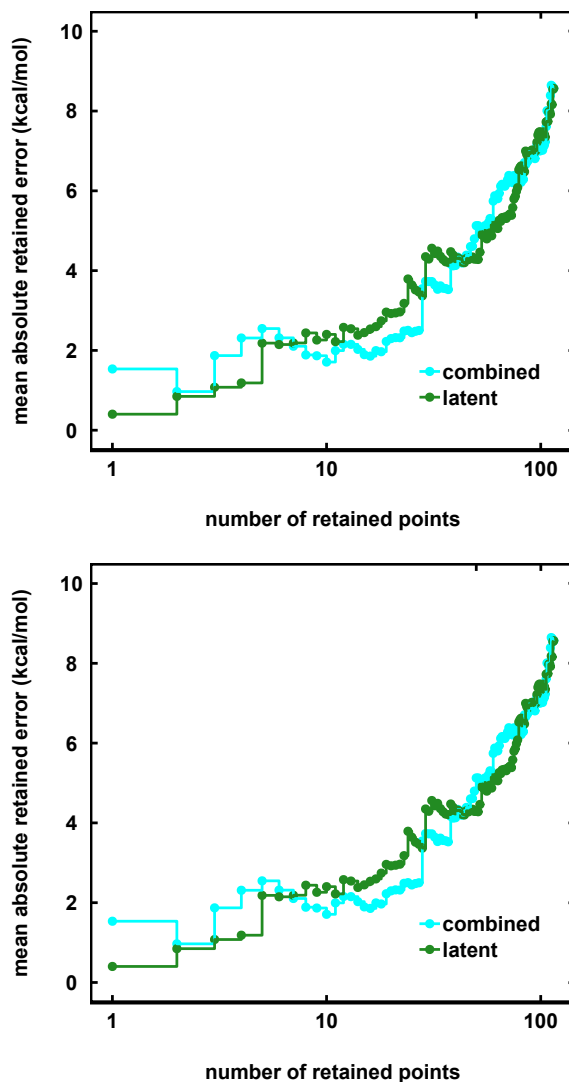


Figure D-22: Variation in the mean absolute error (MAE, in kcal/mol) on retained points from ANN models on CSD splitting energy prediction task as a function of (left) the thresholds in different uncertainty metrics and (right) the number of points retained. The metrics compared are the *minimum of the combination of the 10-neighbor average distance in latent space and the standard deviation of a 10-model ensemble and 100 realizations mc-dropout* and as well as the 10-neighbor average distance in latent space alone. All metrics are normalized to  $[0, 1]$  for comparison. Dashed vertical lines show the median of each metric (left). Errors are taken from single-ANN predictions only. It is apparent the minimum of the combined metrics can provide marginally better error control over some of the range, though latent distances alone perform better or equivalent for retained MAE values  $\gtrsim 4.00$  kcal/mol.

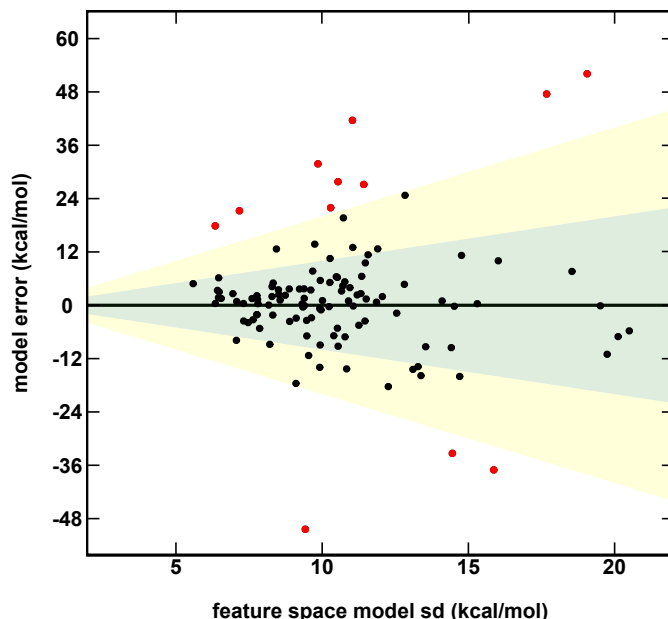


Figure D-23: Relationship between spin-splitting ANN model errors (in kcal/mol) on a 116 molecule CSD set and a calibrated distance-based uncertainty model using feature space distances. The model is fit using eq. (1) from the main text with  $\sigma_1 = 0$ ,  $\sigma_2 = 3.42$ . The translucent green region corresponds to one std. dev. and translucent yellow to two std. dev.. The points with model errors that lie inside either of these two bounds are shown in black, and the percentage within the green or yellow regions are annotated in each graph in green and yellow, respectively. The points outside two std. dev. are colored red.

Table D.12: Spearman (rank) correlation between distances from CSD points and nearest training data in the full spin splitting ANN latent space and low dimensional spaces from principal component analysis (PCA) and uniform manifold approximation (UMAP). In all cases the dimensionality reduction is conducted based on training data only.

method	Spearman correlation
2D PCA	0.32
4D PCA	0.68
16D PCA	0.93
32D PCA	0.97
2D UMAP	0.32
8D UMAP	0.17

### Text D.3: Application to MNIST and Fashion-MNIST classification task

In order to test the application of our proposed method to other tasks, we consider two standard benchmark test image classification tasks, MNIST<sup>320</sup> and Fashion-MNIST<sup>575</sup>. Both consist of 60k training and 10k test grayscale images of size  $28 \times 28$  pixels divided into 10 classes. We use a convolutional neural network (CNN) with the same hyperparameters for each task, trained with cross entropy loss and no explicit regularization (Table D.13).

Training on MNIST gives a train/test accuracy (top-1) of 100.00%/99.06% (0/94 errors), while training on Fashion-MNIST gives a train/test accuracy (top-1) of 99.87%/91.51% (77/849 errors).

As before, we average the distance of each test point to the nearest 10 training points to generate a confidence metric (Figure D-25). Comparison of the distribution of correctly and incorrectly classified points reveals a shift towards high distance for the incorrectly classified points, with an increase in mean distance of the incorrect points of 66.64% for MNIST and 11.90% for Fashion-MNIST. We perform a Mann–Whitney test to estimate if the difference in distances is significant and find  $p = 9.3 \times 10^{-47}$  and  $p = 1.12 \times 10^{-36}$  for MNIST and Fashion-MNIST respectively, although in both cases the number incorrect samples is low. This suggests that the methods proposed could be applied to other types of the neural networks (CNNs), datasets (images) and tasks (classification).

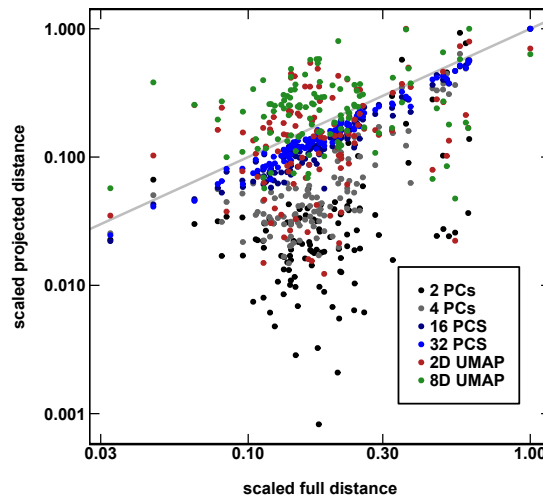


Figure D-24: Distance to nearest training point for CSD points in the full spin splitting ANN latent space and low dimensional spaces from principal component analysis (PCA) and uniform manifold approximation (UMAP). In all cases the dimensionality reduction is conducted based on the training data only. All distances are normalized by the largest value and the gray line shows parity. 472



Table D.13: Hyperparameters and topology for image classification CNN

parameter	value
layer 1	64 filter $3 \times 3$ 2D convolution
layer 2	32 $3 \times 3$ 2D convolution
layer 3	64 unit dense
layer 4	64 unit dense
layer 5	10 unit softmax
activation function	relu
learning rate	0.01
optimizer	adam
$\beta_1$	0.9
$\beta_2$	0.999
decay	0
dropout (all hidden)	none
batch size	128
epochs	50
$L^2$ regularization	none
semibatch normalization	no
early stopping	none

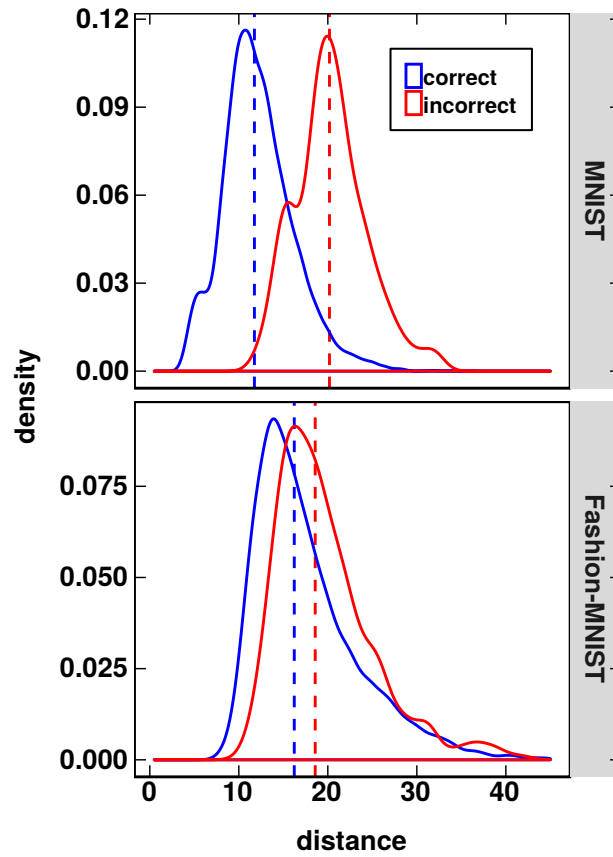


Figure D-25: Comparison of kernel density estimates for 10-neighbor average latent distance to training data for image classification task using a CNN, showing correctly (blue) and incorrectly (red) classified points for MNIST (top) and Fashion-MNIST (bottom) benchmarks. The dashed vertical lines represent the median values for each curve.

## Appendix E

### Multiobjective, multifidelity redox couple design

Table E.1: Spin multiplicities selected for ionization processes. The M(II) ground state evaluated from solvent-corrected gas phase geometry optimizations is determined first. Once that is identified, the oxidation process is evaluated by a single electron ionization from that M(II) ground state with spin multiplicities as follows for the M(III) state chosen based on the electron configurations of the M(II) ion also indicated below.

	M(II)/M(III) electron configuration	M(II) ground state	M(II) spin	M(III) spin
Cr	$[Ar]3d^4/[Ar]3d^3$	LS	1	2
		HS	5	4
Mn	$[Ar]3d^5/[Ar]3d^4$	LS	2	1
		HS	6	5
Fe	$[Ar]3d^6/[Ar]3d^5$	LS	1	2
		HS	5	6
Co	$[Ar]3d^7/[Ar]3d^6$	LS	2	1
		HS	4	5

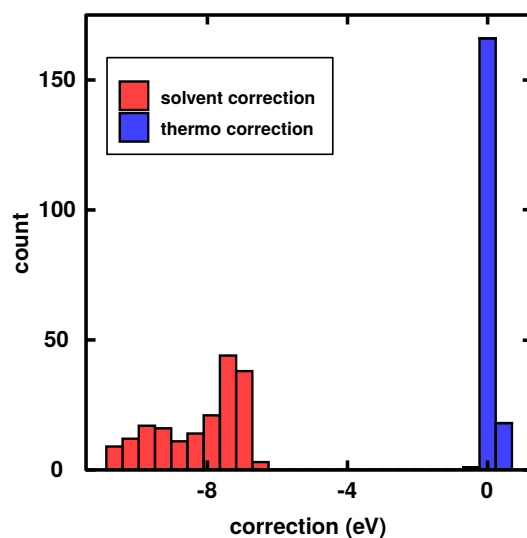


Figure E-1: Comparison of solvent (red) and thermodynamic (blue) free energy corrections to DFT calculated ionization potentials. Data on transition metal complexes taken from Ref<sup>511</sup>.

### Text E.1: Description of hierarchical ligand construction approach

A series of hierarchical rules were used to combinatorially enumerate a large space of monodentate and bidentate ligands. These consist of i) heterocycle and metal-coordinating atom selection, ii) modification of non-metal-coordinating portions of the ring, iii) (optional) ring fusion to form symmetric or asymmetric bidentate ligands, and iv) hierarchical functionalization of a single ring site. Specific rules are as follows:

- Eight core five- or six-membered rings are first defined (labels: A-D for six-membered rings, E-H for five-member rings), each with O- or N-atom coordination to the metal defined as position 1 on the ring. In the O-atom cases (C, D, G, and H), the oxygen always forms single bonds with neighboring carbon atoms in positions 2 and 6, whereas for the N-atom cases (A, B, E, and F), nitrogen forms one single bond and one double bond (between positions 1 and 2) in all heterocycles. Pairs of rings of the same size and coordinating atom are distinguished by the presence or absence of a double bond at the C( $n-1$ )-C( $n$ ) position.
- Five modifications to the core rings (labels: a-e) are carried out centered on C3 and influence the C2-C3-C4 bonds for both five- and six-membered rings (A-H) to form two-letter-coded ring structures (i.e., Aa-He). These modifications of C2-C3-C4 include: two single bonds with a carbon atom (a), a single (2-3) and double (3-4) bond (b), insertion of an oxygen atom at the C3 site (c), insertion of an amine nitrogen (NH) at the C3 site (d), or insertion of a sulfur atom at the C3 site (e). In some cases, modification with a does not modify the core heterocycle (e.g., Aa is A unchanged).
- From these combinations, Fb or Hb would create a C center with two double bonds at both 3-4 and 4-5. After these two combinations are excluded, we obtain a total of 38 modified, core ring structures that can be used as scaffolds for forming functionalized monodentate ligands.
- Bidentate ligands are also formed by joining the two rings at the 2nd position to form a 2-2' bond. Both symmetric and asymmetric bidentate ligands are formed through every possible pairwise permutation of the 38 unique core rings for a total of  $\binom{38+2-1}{2} = 741$  bidentate ligand scaffolds.
- Both monodentate and bidentate ligands are then functionalized hierarchically at the C4 site for both five- and six-membered rings through four steps described next.

- Methylene or phenyl groups are added either once or twice at the C4 site (referred to as B<sub>1</sub> and B<sub>2</sub> for the first and second addition). In the case of methylene group addition, the two hydrogen atoms can be further replaced with a functional group (i.e., S<sub>1</sub> or S<sub>2</sub>) either by a doubly bonded functional group (=CH<sub>2</sub>, =O) or by any two singly bonded groups (-CH<sub>3</sub>, -H, -OH, and -NH<sub>2</sub>). For phenyl group addition, this functionalization step is not carried out (i.e., S<sub>1</sub> = H). These changes are referred to as S<sub>1</sub> when they occur on the B<sub>1</sub> functional group chain or S<sub>2</sub> for B<sub>2</sub>. In total, each B<sub>1</sub> or B<sub>2</sub> site can have 13 chemical identities: 1 phenyl group, 2 doubly-bonded S<sub>1</sub> choices, 4 symmetric singly-bonded S<sub>1</sub> choices, and 6 asymmetric singly-bonded S<sub>1</sub> choices. Combining all possible choices of B<sub>1</sub>-S<sub>1</sub> with B<sub>2</sub>-S<sub>2</sub>, as well as the cases where only B<sub>1</sub>-S<sub>1</sub> are generated, yields  $14 \times 13 = 182$  functional group cores.
- The final modification to the functional group cores is to terminate them with a singly-bonded functional group modification (i.e., T) at either the para position on the phenyl group or on a terminal hydrogen not involved in S<sub>1</sub> or S<sub>2</sub> in the case of the methylene groups. The five terminating functional groups are H, CH<sub>3</sub>, OH, NH<sub>2</sub>, and Cl, and their combination with the other functional group modifications produces  $182 \times 5 = 910$  total functional groups. This corresponds to 897 unique functional groups after accounting for identical cases (i.e., no B<sub>2</sub>, T = CH<sub>3</sub> is the same as methylene B<sub>2</sub> and T = H).
- In total, 741 bidentate and 38 monodentate ligand scaffolds combined with 897 functional groups produces 698,763 unique ligands distributed over our design space. After combination with 4 metal identities, the total design space contains 2,795,052 complexes.
- A note on stereochemistry: since the representation used for this design exercise is based on the molecular graph, we do not distinguish stereoisomers at a design level. When performing DFT simulations, an isomer is chosen with decoration groups placed cis, where applicable.

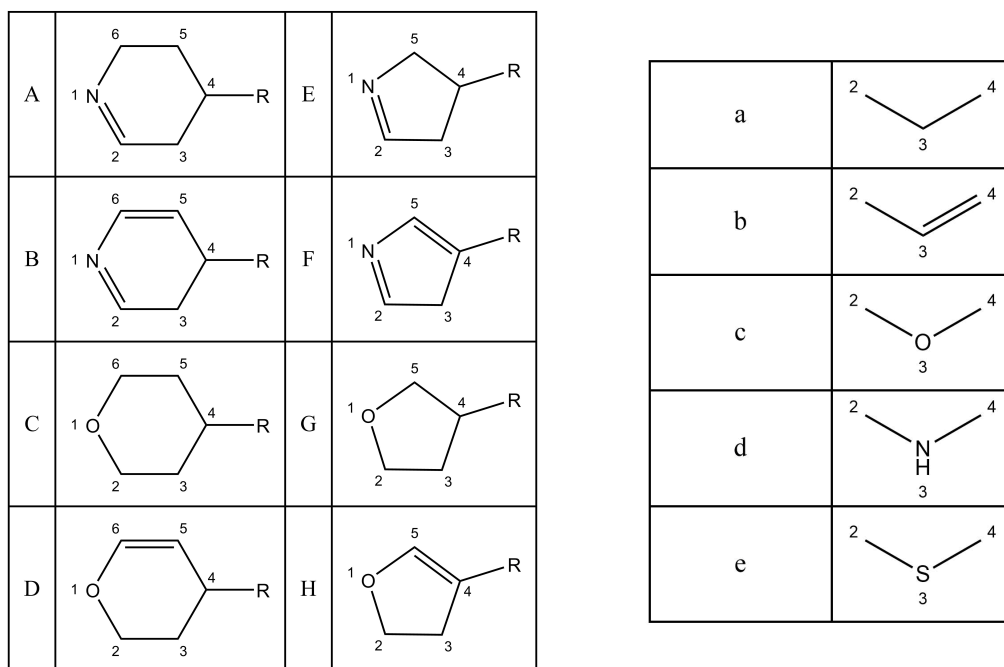


Figure E-2: Structures of eight core heterocycles (A-H, left) and corresponding C2-C3-C4 modifications (a-e, right) used to generate 38 monodentate ligands or for subsequent fusion to form bidentate scaffolds.

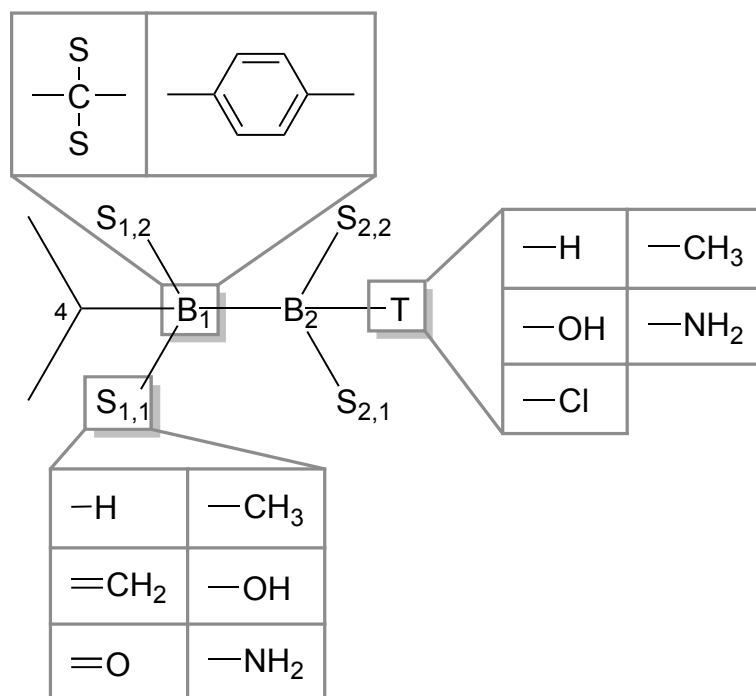


Figure E-3: Approach for functionalizing rings at the R group position centered on C4. The backbone of the functional group consists of two sites, B<sub>1</sub> and B<sub>2</sub>, which form a linear chain and can either be alkyl carbons or phenyl groups. Hydrogen atoms are then replaced symmetrically with so-called sidechain functionalization at the first backbone group site (i.e., S<sub>1</sub>) or second backbone group site (i.e., S<sub>2</sub>). In practice, this only is done for the alkyl chains, not the phenyl groups. Both alkyl and phenyl groups can then have a third terminating functional group added, denoted T.

Table E.2: Original sources of ‘hot start’ data. In 123 cases from Refs.<sup>308,511,543</sup>, revised low spin state definitions for Cr(III) and Mn(II) (i.e., singlet instead of triplet) were published in Ref<sup>593</sup> and used where relevant.

# complexes	description	original ref.
177	ligands CO, pyridine, water, furan, and methyl isocyanide	511
20	spectrochemical series complexes	308
33	spin crossover complex discovery leads	543
5	diverse complexes from uncertainty quantification	593



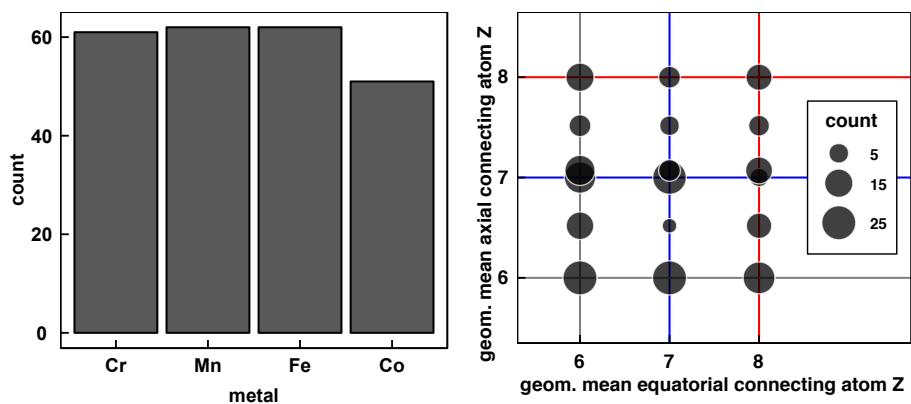


Figure E-4: Characteristics of 235 transition metal complex ‘hot start’ data points. (left) Distribution of metal centers in complexes. (Right) Comparison of geometric mean of the connecting atom nuclear charge,  $Z$ , for equatorial and axial ligands. The total number of data points in each combination is indicated according to the inset legend. The integer nuclear charges for carbon (gray), nitrogen (blue) and oxygen (red) are indicated with lines. The slightly shifted circles near 7 correspond to combinations of carbon and oxygen. For homoleptic complexes, the axial and equatorial connecting atom elements are the same, and for some of the heteroleptics, the complexes will be the same as well.

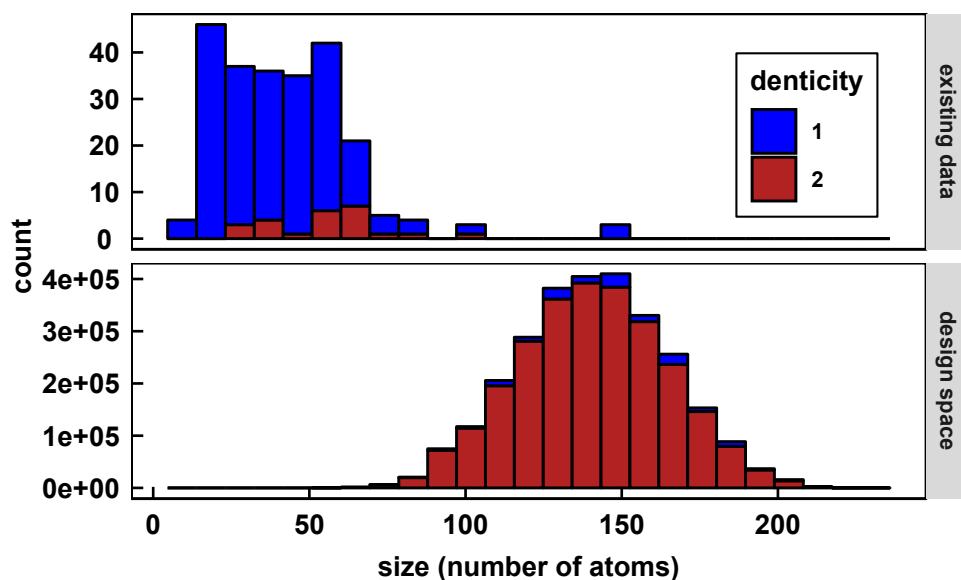


Figure E-5: Complex size (i.e., number of atoms) distribution distinguished by monodentate and bidentate ligands for ‘hot start’ data (top) in comparison to design space (bottom).

## Text E.2: Details of initial sample of constructed design space from clustering

The 698,763 ligands are combined with four transition metals (i.e., Cr, Mn, Fe, and Co) to produce a design space of 2,795,052 (i.e.,  $\sim 2.8$  million) transition metal complexes. The transition metal complexes are an average of 141 atoms in size (minimum: 55, maximum: 235 atoms) and evenly divided over the four transition metals by definition (Table E.5). In order to select a number of initial candidates to simulate, we obtained 300 cluster centroids from the design space using  $k$ -medoids clustering in the RAC-155 representation. We employed the clustering large applications (CLARA) algorithm<sup>633</sup>, as implemented in the R package ‘Cluster’<sup>600</sup>. We used Euclidean feature space (i.e., RAC-155) distance as the dissimilarity metric for clustering. The resulting clusters ranged from 738 to 35,466 complexes in size with an average size of 9,317 (Figure E-6). We judged cluster quality by the isolation ratio, which is the maximum in-cluster dissimilarity divided by the average distance between centroids. Most clusters (i.e., 285 of 300) have good characteristics, with isolation ratios below 2.5, whereas a minority (15 of 300) exhibit large intra-cluster dissimilarity (Figure E-7). Silhouette analysis assigns each cluster a score ranging from  $-1$  to  $1$  representing how well separated each cluster is, with  $1$  being perfect and  $0.5$  being well separated. Silhouette values near zero arise when data is on the border of two or more clusters. When accounting for cluster size, we observe that the majority of data is assigned to a cluster with a score of at least  $0.25$  (Figure E-6). The full design space characteristics are preserved in the clustered set of 300 unique complexes, including an even distribution over metals as well as the range of complex sizes (Table E.5).

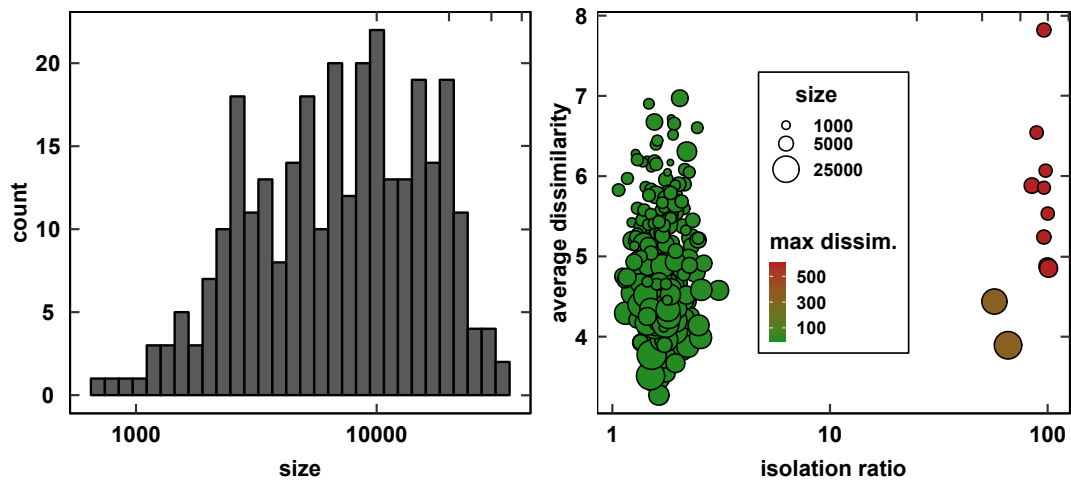


Figure E-6: Characteristics of 300 clusters obtained with k-medoids clustering using CLARA. (left) Unnormalized distribution of number of complexes in the clusters (i.e., size). (right) Dissimilarity, size, and isolation ratio of clusters. The dissimilarity is defined by the Euclidean distance to the medoid in RAC-155 feature space (i.e., lower means a cluster with less dissimilarity). The isolation ratio is defined as the maximum in-cluster dissimilarity divided by the average distance between centroids (i.e., lower values indicated good separation). Complexes are colored by their largest single in-cluster dissimilarity and the symbols are sized by the number of complexes in the cluster, as indicated by inset legend.

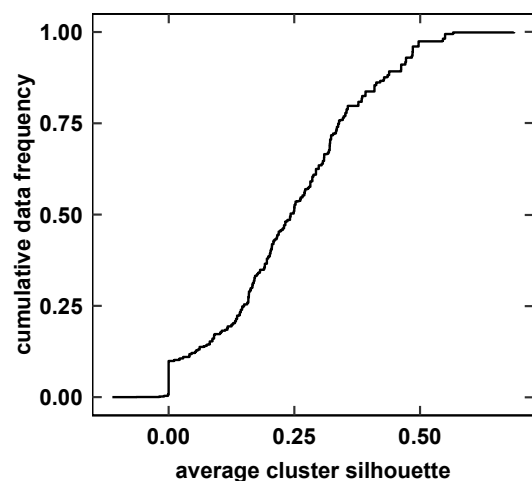


Figure E-7: Silhouette analysis for points in initial 300 clusters. From left to right on the x-axis shows the proportion of the total design space that is assigned to clusters with a given silhouette score. Scores have a theoretical range from  $-1$  (worst) to  $1$  (best). The practical range observed across this data set is predominantly from around  $0.0$  to just under  $0.75$ . The value  $0.5$  corresponds to good separation between clusters.

Table E.3: Design space and cluster population analysis. The total number of ligands, complexes, and fraction of each metal type is compared for the whole design space vs. those in the resulting clustered subset. The number of atoms in the assembled complex is shown across both the full design space and clustered subset, including the minimum and maximum complex sizes as well as the average (avg.) and standard deviation (std. dev.) of the complex size.

set	# distinct		metal fraction	# atoms			
	ligands	complexes	Cr/Mn/Fe/Co	min	max	avg.	std. dev.
design space	698,763	2,795,052	0.25/0.25/0.25/0.25	55	235	141.0	23.54
clustered set	300	300	0.27/0.26/0.21/0.25	73	217	142.4	27.50

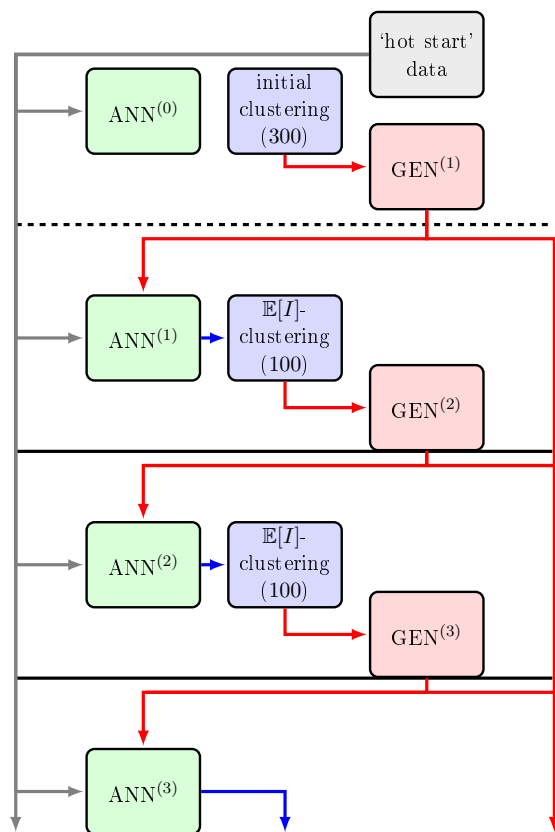


Figure E-8: Schematic of active learning workflow shown for four initial generations. The ‘hot start’ data is used to train an initial ANN, ANN<sup>0</sup>. Initial, diversity-orientated clustering is used to extract 300 medoids from the design space which are simulated to give the generation 1 data, GEN<sup>1</sup>. This data is combined with the hot start data and used to train the ANN<sup>1</sup> model, which is used to calculate  $\mathbb{E}[I]$  for the design space, and the top 10k complexes are used to extract 100 high- $\mathbb{E}[I]$  leads for simulation, GEN<sup>2</sup>. The results of these DFT calculations are combined with the previous datasets and used to train the next generation ANN, ANN<sup>2</sup>. This model is used to recalculate the  $\mathbb{E}[I]$  scores for the design space, and the process is repeated. DFT calculations are shown in red, model training steps are shown in green and clustering steps are shown in blue. The generation the model or data belong to is indicated in a superscripted number.

Table E.4: Prediction error metrics of initially trained models. The mean absolute error (MAE) and root-mean-square error (RMSE) are reported for different machine learning (ML) models that predict redox potential (in eV) and logP: single-task ANNs, multitask ANN, and a Gaussian-kernel gaussian process, or GP, model. Errors are reported on both train and a 10% uniformly-selected test partition from ‘hot start’ data used for initial ML model training. Errors are also reported on 107 complexes converged from generation 1 cluster medoids as an out-of-distribution set. The lowest errors across the three models for a given data partition are indicated in bold.

		multitask ANN	single-task ANN	GP
logP				
train	MAE	1.45E-04	1.45E-04	<b>1.05E-04</b>
	RMSE	2.46E-04	2.53E-04	<b>1.40E-04</b>
test	MAE	2.40E-04	<b>1.49E-04</b>	1.74E-04
	RMSE	3.13E-04	<b>1.90E-04</b>	2.59E-04
generation 1	MAE	<b>3.04E-03</b>	5.92E-03	1.31E-02
	RMSE	<b>3.74E-03</b>	6.81E-03	1.62E-02
redox (eV)				
train	MAE	0.15	<b>0.11</b>	0.13
	RMSE	0.20	<b>0.14</b>	0.20
test	MAE	<b>0.32</b>	0.36	0.44
	RMSE	<b>0.51</b>	0.57	0.59
generation 1	MAE	<b>0.71</b>	0.80	1.47
	RMSE	<b>0.90</b>	1.03	1.70

Table E.5: Relationship (i.e., linear correlation coefficient) between absolute test errors evaluated on 10% uniformly-selected test data partition across several ML models for both redox and logP. The ML models compared are a multitask ANN that predicts both properties and independent single-task ANNs as well as a Gaussian-kernel Gaussian process (GP) model. By definition, the linear correlation of model errors for the model with itself is 1.0.

		redox			logP		
		GP	multitask	single task	GP	multitask	single task
redox	GP	1.00	0.81	0.79	0.64	0.14	0.44
	multitask		1.00	0.95	0.71	0.20	0.39
	single task			1.00	0.70	0.26	0.42
logP	GP				1.00	-0.08	0.28
	multitask					1.00	0.34
	single task						1.00

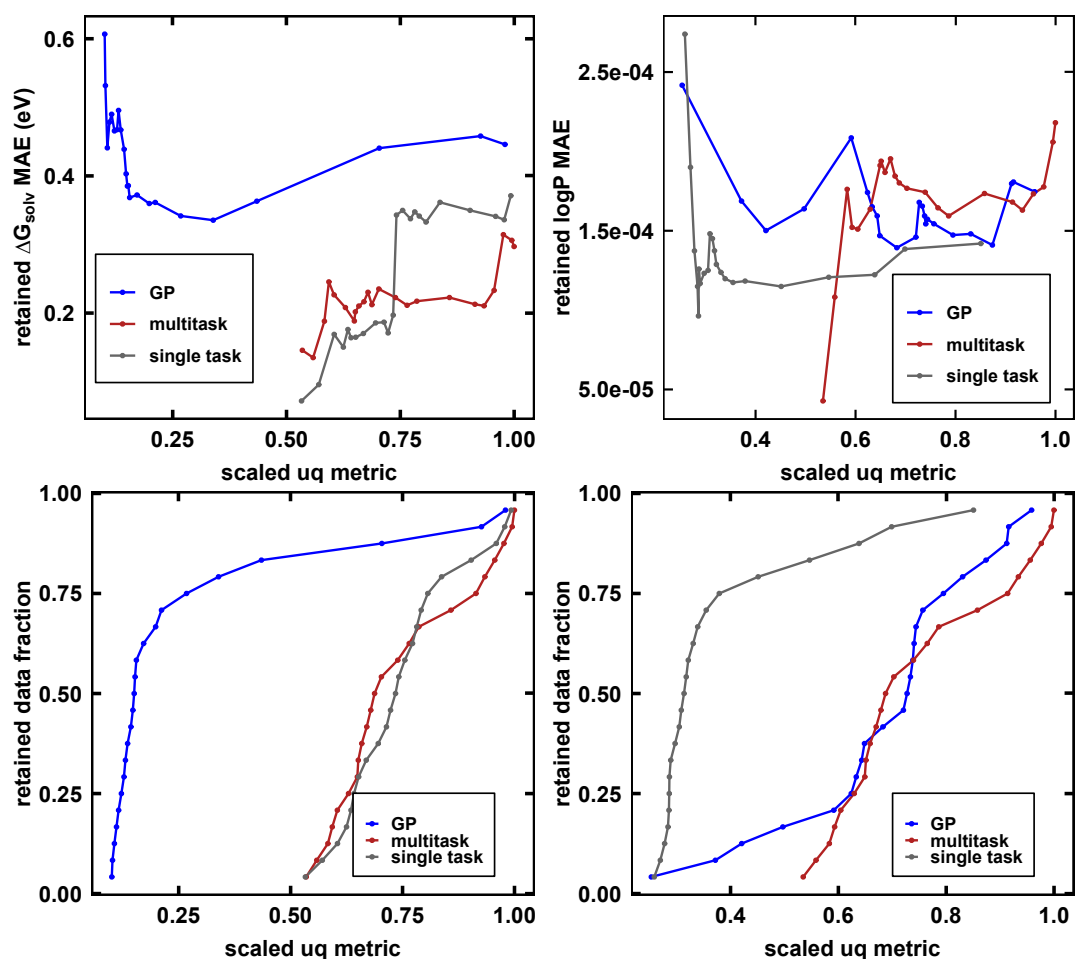


Figure E-9: Mean absolute error (MAE) on retained data with lower uncertainty than a given threshold as a function of this threshold (top) and showing the fraction of retained complexes over which the MAE is obtained at that threshold (bottom) for redox potential (eV, left) and logP (right) errors on uniformly-selected test data. Three approaches are compared: independent Gaussian processes (GP), a pair of single task ANNs and a multitask ANN. The GPs use their posterior standard deviations as uncertainty metrics while the ANNs use the average distance to the ten nearest training points in their final latent spaces as in Ref.<sup>593</sup>. All uncertainty measures are normalized to have a maximum of 1 for comparison.

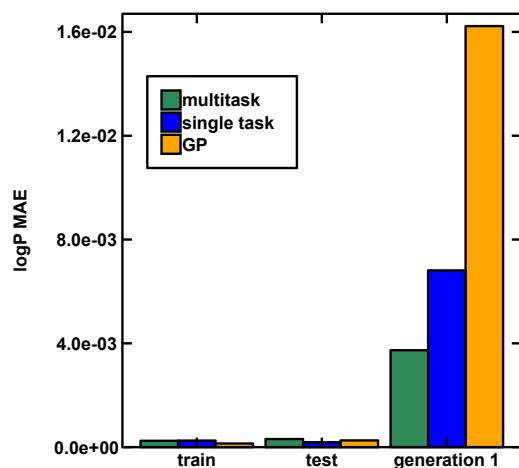


Figure E-10: Comparison of mean absolute errors for logP predictions. The models shown are a multitask ANN (green), a single task ANN (blue), and a Gaussian process model (GP, orange). Errors are reported on training data (train), a uniform 10% test partition (test), and the generation 1, out-of-distribution set, which corresponds to 107 converged results from cluster medoids of the full space (medoids on the x-axis of the chart).

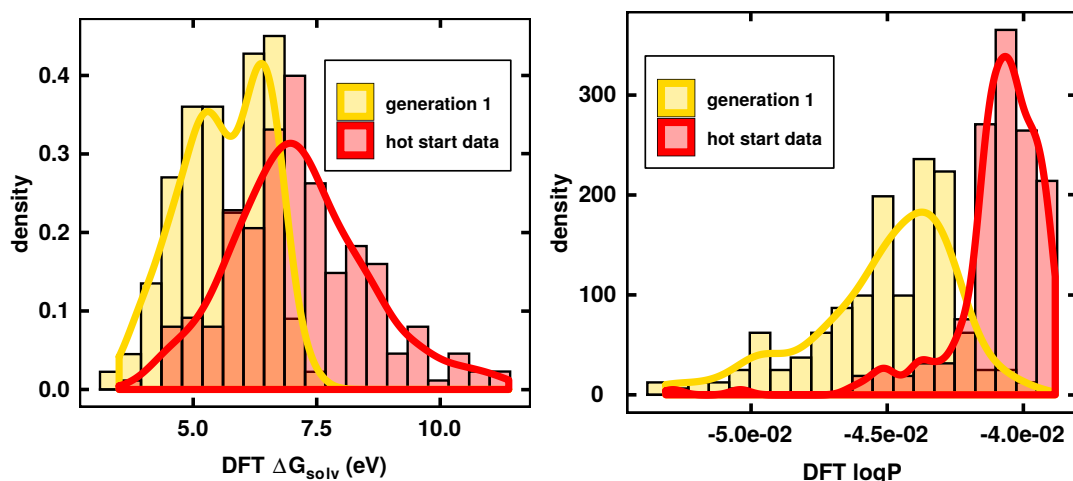


Figure E-11: Comparison of DFT-predicted  $\Delta G_{\text{solv}}$  (left) and logP (right) distributions for hot start data (235 cases, red) and generation 1 data (107 cases, yellow) shown as both a histogram (bars) and with a smooth kernel density estimate (lines).



Table E.6: Relationship (i.e., linear correlation coefficient) between absolute errors of model trained on ‘hot start’ data and evaluated on out-of-distribution, generation 1 set for both redox and logP. The out-of-distribution set consists of 107 complexes obtained from converged results on 300 initial cluster medoids of the full space. The ML models compared are a multitask ANN that predicts both properties and independent single-task ANNs as well as a Gaussian-kernel Gaussian process (GP) model. By definition, the linear correlation of model errors for the model with itself is 1.0.

		redox			logP		
		GP	multitask	single task	GP	multitask	single task
redox	GP	1.00	0.60	0.82	-0.04	-0.14	0.01
	multitask		1.00	0.83	0.02	-0.10	-0.05
	single task			1.00	-0.07	-0.09	-0.09
logP	GP				1.00	0.41	0.52
	multitask					1.00	0.57
	single task						1.00

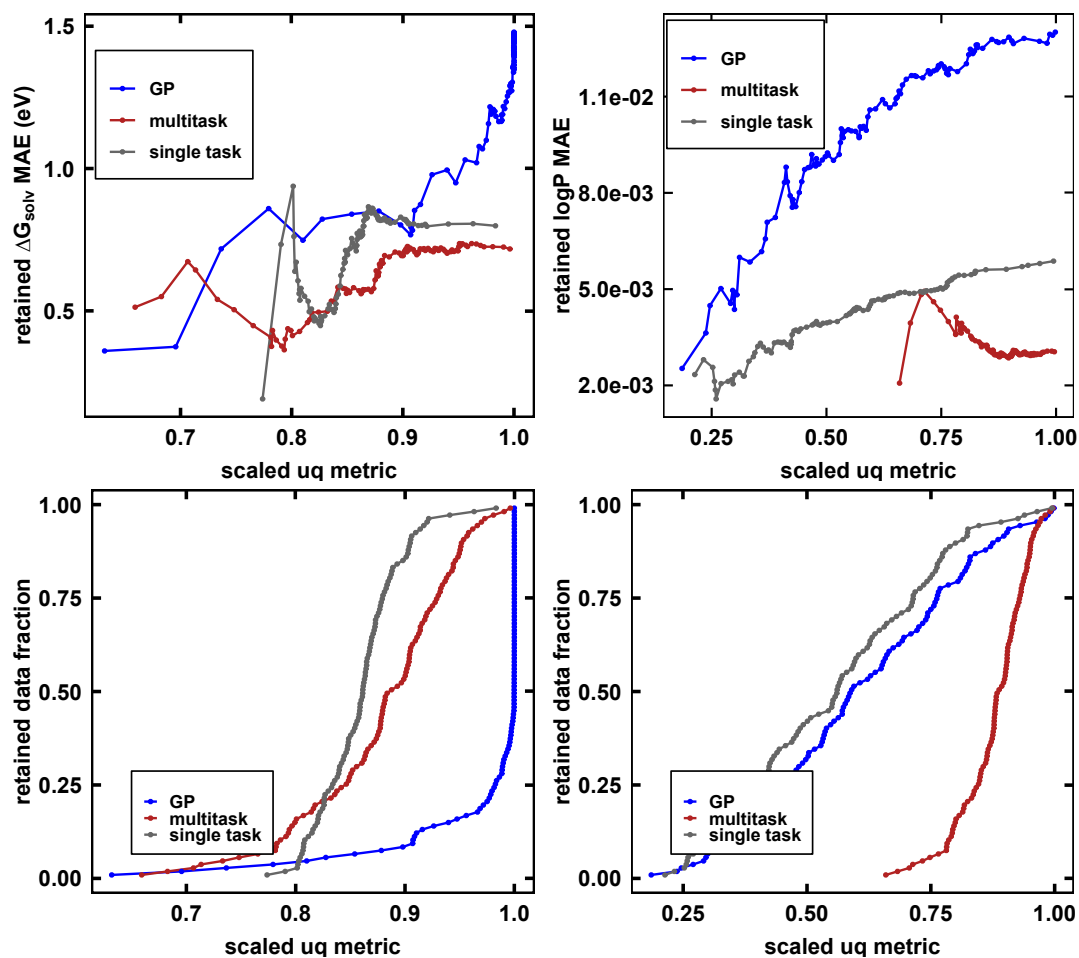


Figure E-12: Mean absolute error (MAE) on retained data with lower uncertainty than a given threshold as a function of this threshold (top) or the fraction of retained complexes at that threshold (bottom) for redox potential (eV, left) and logP (right) errors on out-of-distribution medoid data. Three approaches are compared: independent Gaussian processes (GP), a pair of single task ANNs and a multitask ANN. The GPs use their posterior standard deviations as uncertainty metrics while the ANNs use the average distance to the ten nearest training points in their final latent spaces as in Ref.<sup>593</sup>. All uncertainty measures are normalized to have a maximum of 1 for comparison.

Table E.7: Data sets used in this work: initial step, name of data set, number of complexes converged in the data set, total attempted, and additional details on the data set.

step	name	number	attempted	details
0	hot start	235	N/A	data from previous studies
1	generation 1	107	300	selected from k-medoids of full space
2	generation 2	34	100	selected from k-medoids of top 10k 2D EI leads
3	generation 3	24	100	selected from k-medoids of top 10k 2D EI leads
4	generation 4	15	100	selected from k-medoids of top 10k 2D EI leads
5	generation 5	14	100	selected from k-medoids of top 10k 2D EI leads
-	random test	122	300	uniform random selection

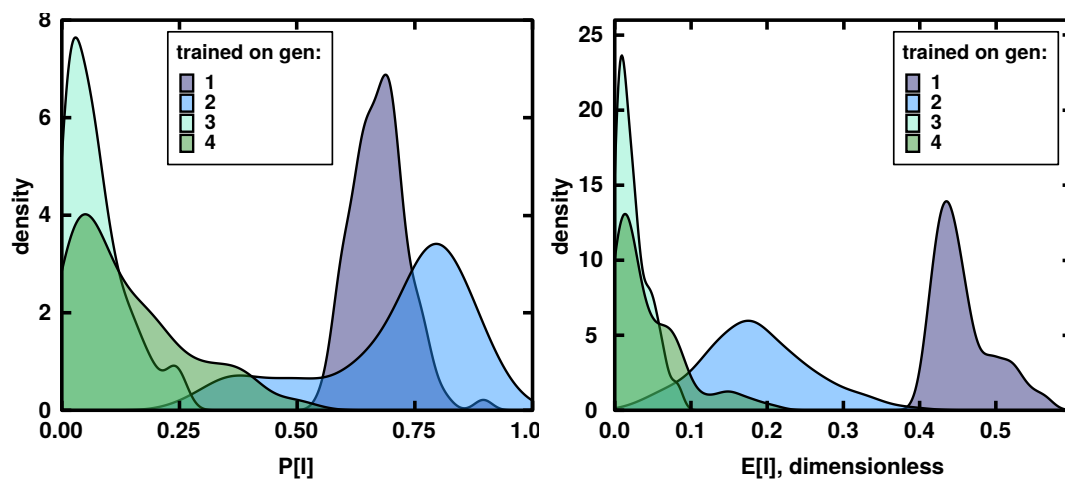


Figure E-13: Evolution of  $E[I]$  and  $P[I]$  distributions over different generations of active learning, colored as indicated in inset legend by the generation used to train the model. Distributions are estimated by kernel density estimates for the 100 selected medoids at each generation.

Table E.8: Mean absolute error (MAE) and root-mean-square error (RMSE) (both in eV) for  $\Delta G_{\text{solv}}$  prediction on training data (train), 10% uniform held out data (held out), lookahead errors to subsequently-acquired data sets, and performance on randomly-selected test data (random test).

data set	model generation	MAE (eV)	RMSE (eV)
train	0	0.15	0.20
	1	0.14	0.19
	2	0.14	0.21
	3	0.15	0.22
	4	0.15	0.22
	5	0.24	0.33
held out	0	0.32	0.51
	1	0.32	0.48
	2	0.19	0.26
	3	0.18	0.25
	4	0.22	0.30
	5	0.36	0.49
lookahead errors			
generation 1	0	0.71	0.90
generation 2	0	0.76	0.97
	1	0.46	0.56
generation 3	0	1.67	1.75
	1	0.67	0.75
	2	0.51	0.58
generation 4	0	1.99	2.07
	1	0.96	0.99
	2	0.65	0.72
	3	0.37	0.54
generation 5	0	1.64	1.72
	1	0.67	0.74
	2	0.42	0.52
	3	0.41	0.50
	4	0.42	0.49
random test	0	0.70	0.90
	1	0.41	0.51
	2	0.41	0.52
	3	0.40	0.52
	4	0.39	0.51
	5	0.38	0.49

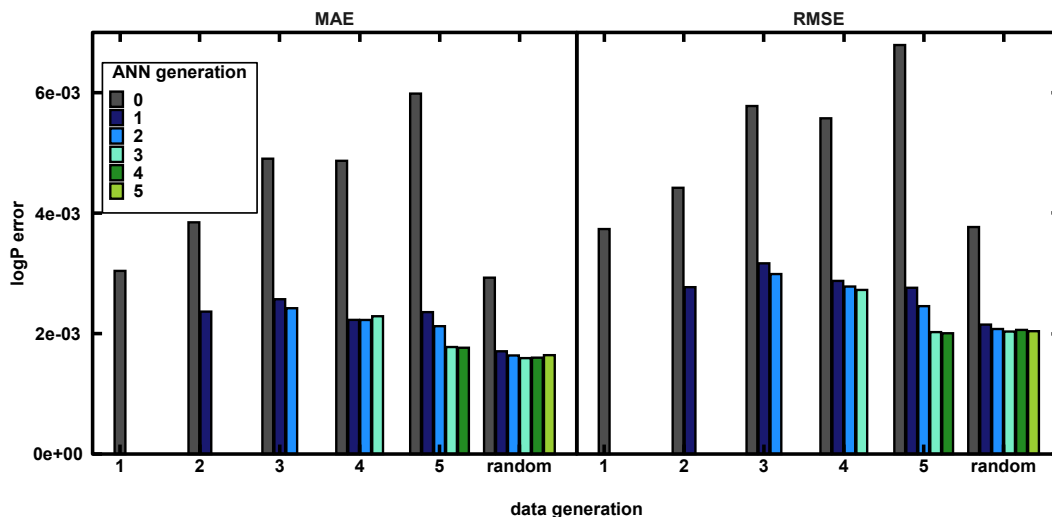


Figure E-14: Lookahead and random-test MAE (left) and RMSE (right) error metrics for logP predictions from sequential generations of multitask ANN models on all data sets gathered subsequent to training each model.

Table E.9: Mean absolute error (MAE) and root-mean-square error (RMSE) (both in eV) for logP prediction on training data (train), 10% uniform held out data (held out), lookahead errors to subsequently-acquired data sets, and performance on randomly-selected test data (random test).

data set	model generation	MAE	RMSE
train	0	1.45E-04	2.46E-04
	1	2.51E-04	3.80E-04
	2	3.73E-04	6.99E-04
	3	4.40E-04	8.56E-04
	4	4.58E-04	8.90E-04
	5	6.78E-04	1.10E-03
held out	0	2.40E-04	3.13E-04
	1	1.10E-03	1.64E-03
	2	5.29E-04	7.50E-04
	3	4.07E-04	7.64E-04
	4	5.37E-04	8.07E-04
	5	7.15E-04	1.29E-03
lookahead errors			
generation 1	0	3.04E-03	3.74E-03
generation 2	0	3.85E-03	4.42E-03
	1	2.36E-03	2.77E-03
generation 3	0	4.90E-03	5.78E-03
	1	2.57E-03	3.17E-03
	2	2.42E-03	2.99E-03
generation 4	0	4.87E-03	5.57E-03
	1	2.23E-03	2.87E-03
	2	2.23E-03	2.78E-03
	3	2.29E-03	2.72E-03
generation 5	0	5.98E-03	6.79E-03
	1	2.36E-03	2.76E-03
	2	2.12E-03	2.46E-03
	3	1.78E-03	2.02E-03
	4	1.77E-03	2.01E-03
random test	0	2.93E-03	3.77E-03
	1	1.71E-03	2.15E-03
	2	1.64E-03	2.08E-03
	3	1.59E-03	2.03E-03
	4	1.60E-03	2.06E-03
	5	1.64E-03	2.04E-03

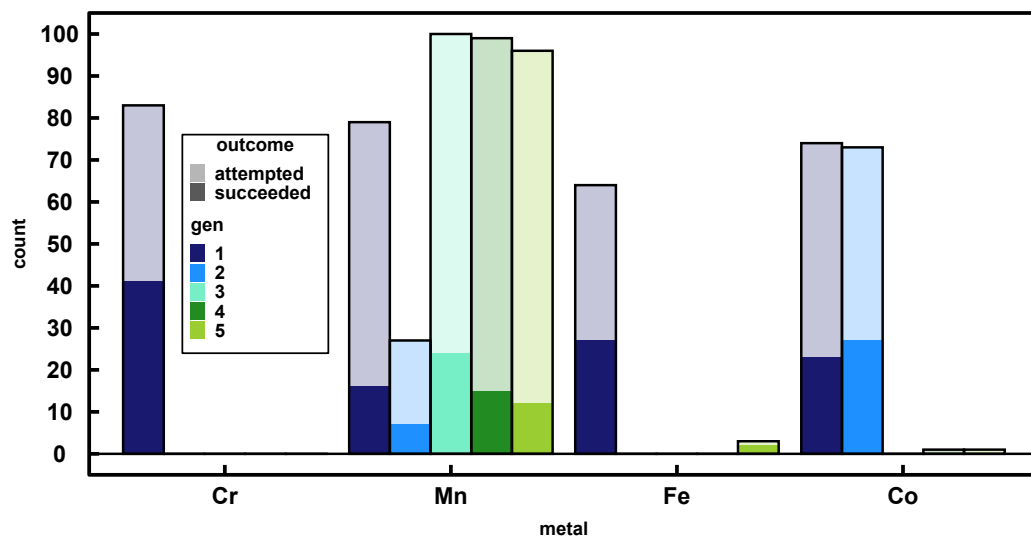


Figure E-15: Distribution of metals in the medoids selected for simulation (light) and those result in successful redox potential calculations (solid) across generations of the optimization procedure.

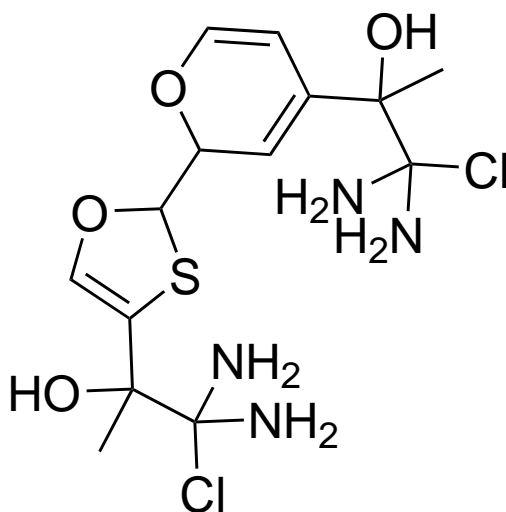


Figure E-16: Skeleton structure of a bidentate ligand selected for complexation with both Co and Mn during design space exploration. The oxygen atoms on the upper left of the skeleton structure both coordinate the relevant metal.

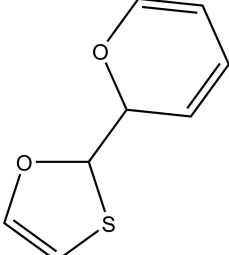
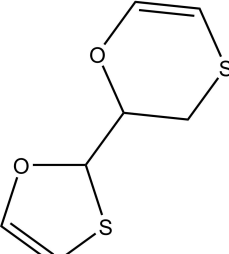
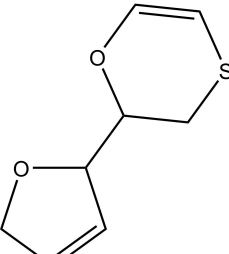
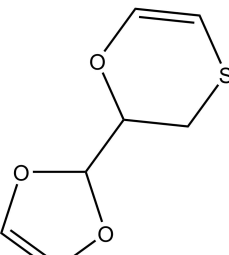
DbHe	
DeHe	
DeGb	
DeHc	

Figure E-17: Most commonly selected unfunctionalized bidentate ligands selected by the multiobjective optimization. Both the skeleton structure and associated naming convention are indicated, with the ring at the top corresponding to the first two-letter code and the ring at the bottom corresponding to the second two-letter code.



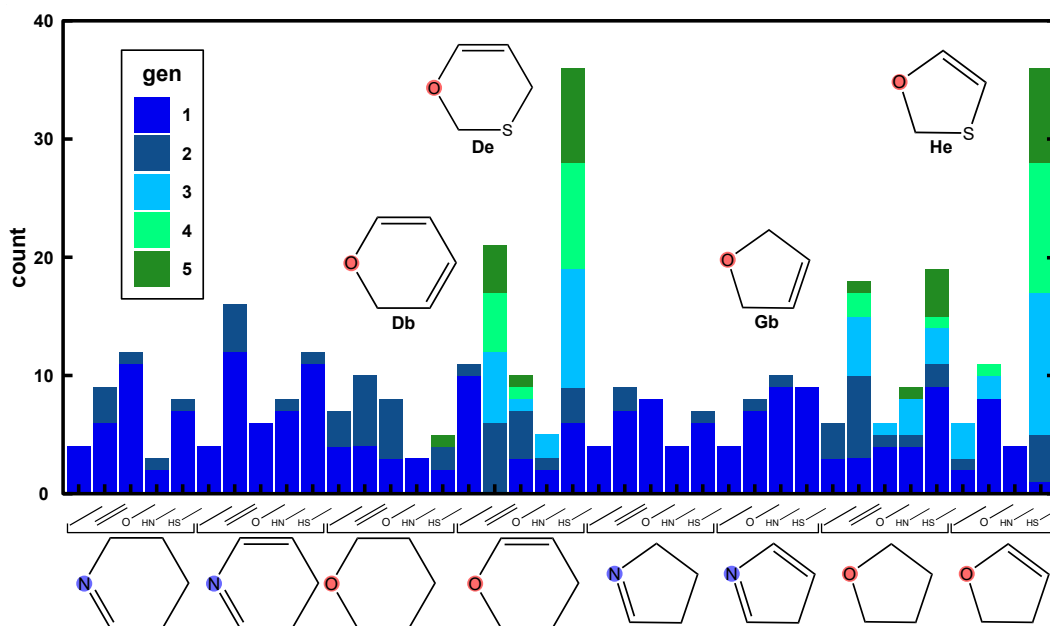


Figure E-18: Histogram of base heterocycle occurrence in ligands during five generations of the active learning process. Bidentate ligands are broken down into their constituent heterocycles and both are counted. A representative heterocycle is shown below each group of similar heterocycles, with the detailed functionalization of the heterocycles indicated on the x-axis. Illustrations of the most common motifs are inset beside their counts.

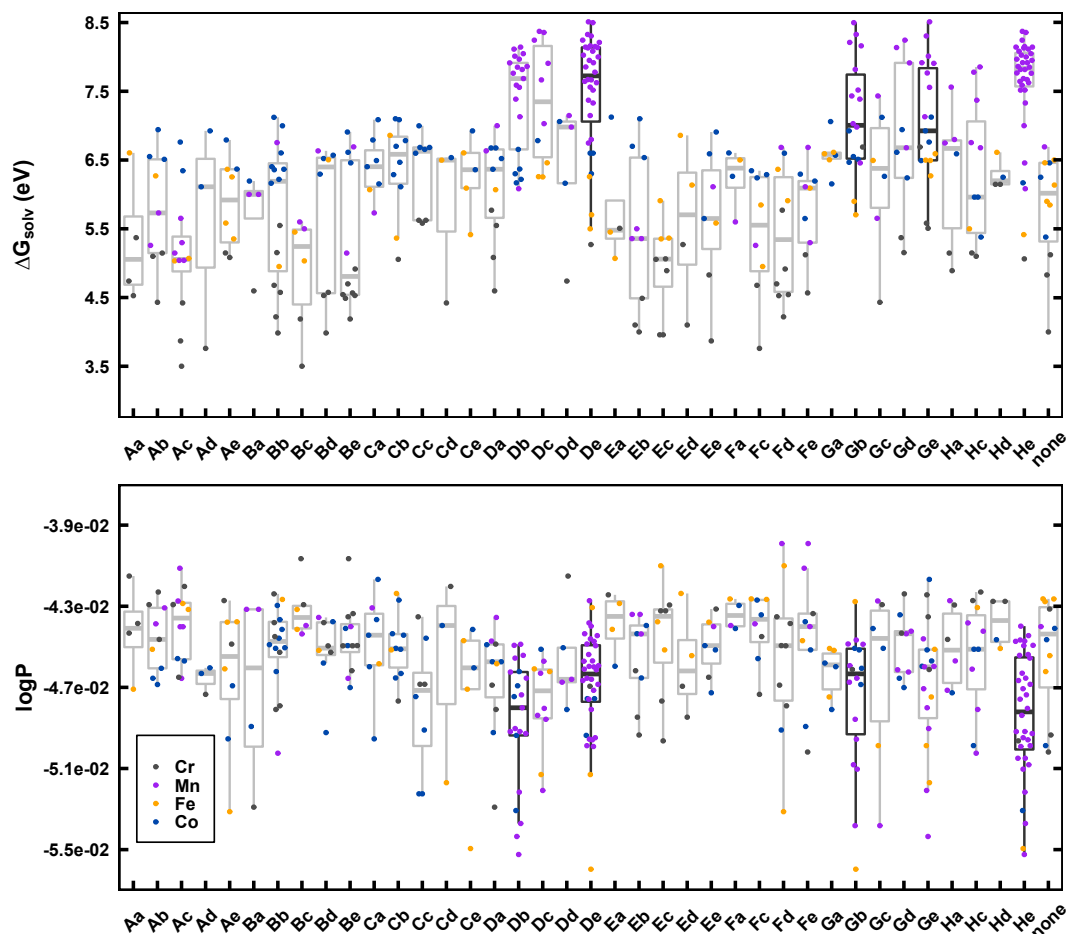


Figure E-19: Distribution of computed properties for  $\Delta G_{\text{solv}}$  (top, in eV) and  $\log P$  (bottom) for all calculated (i.e., with DFT) complexes. Both a box-and-whiskers plot (light gray) is shown with swarm plot overlay (points colored by metal, as indicated in inset). The data is further distinguished according to the x-axis by the unfunctionalized heterocycle for each ligand, where bidentate heterocycles appear twice, one for each heterocycle that makes up the bidentate, whereas monodentate ligands are indicated by the "none" column. Better ligands correspond to higher  $\Delta G_{\text{solv}}$  but lower  $\log P$  on each plot, and the box-and-whiskers plots for heterocycles that constitute ligands with at least one complex in the top 1%  $\Delta G_{\text{solv}}$  or lowest 1%  $\log P$  are emphasized in black.

Table E.10: The most commonly-selected functional groups on ligands from the multi-objective optimization sorted by their frequency and indicated by SMILES strings.

SMILES	frequency
<chem>C(N)C(N)Cl</chem>	3
<chem>C(N)C(=O)Cl</chem>	3
<chem>C(N)(O)C(N)(O)Cl</chem>	3
<chem>C(O)(C)C(=O)Cl</chem>	3
<chem>C(O)C(N)(O)Cl</chem>	3
<chem>C(O)C(=O)C</chem>	3
<chem>C(=O)C(O)(C)O</chem>	3
<chem>C(=C)</chem>	2
<chem>C(C)C(=C)O</chem>	2
<chem>C(C)Cl</chem>	2
<chem>C(=C)C(O)(C)O</chem>	2
<chem>C(=C)N</chem>	2
<chem>CC(=O)</chem>	2
<chem>CC(=O)C</chem>	2
<chem>CC(O)(O)Cl</chem>	2
<chem>C(N)C(=O)C</chem>	2
<chem>C(N)O</chem>	2
<chem>C(O)(C)C(N)(N)Cl</chem>	2
<chem>C(=O)C(N)Cl</chem>	2
<chem>C(=O)C(N)(O)N</chem>	2
<chem>C(=O)CO</chem>	2

Table E.11: Properties of complexes at the Pareto front after each generation (Pareto gen.) including the generation the complex originates from (source gen.), metal center, ligand identity following naming convention,  $\Delta G_{\text{solv}}$  (in eV), and  $\log P$ . The spin-state splitting (i.e.,  $\Delta E_{\text{H-L}}$ , in kcal/mol), and the resulting selected M(II) ground state (G.S.) for the redox process is also indicated as well as the size of the complex in # atoms (# at.). The mean of the properties of the Pareto front after each generation are also indicated.

Pareto gen.	source gen.	metal	ligand	$\Delta G_{\text{solv}}$ (eV)	$\log P$	$\Delta E_{\text{H-L}}$ (kcal/mol)	M(II) G.S.	size (# at.)
1	1	Fe	AeFd_896_896	6.36	-5.31E-02	62.55	LS	82
	1	Fe	CdGe_102_102	6.50	-5.17E-02	-24.20	HS	172
	1	Mn	BbHc_833_833	6.75	-5.02E-02	-35.63	HS	106
	1	Co	AeCa_142_142	6.79	-4.95E-02	-7.98	HS	118
	1	Co	EaGe_442_442	7.12	-4.60E-02	3.08	LS	118
	1	Mn	DeHc_460_460	7.37	-4.38E-02	-51.51	HS	145
		mean		6.82	-4.91E-02	-8.95		123.5
2	1	Fe	AeFd_896_896	6.36	-5.31E-02	62.55	LS	82
	2	Co	CcCc_275_275	6.68	-5.22E-02	-23.72	HS	151
	1	Mn	BbHc_833_833	6.75	-5.02E-02	-35.63	HS	106
	1	Co	AeCa_142_142	6.79	-4.95E-02	-7.98	HS	118
	2	Mn	DcGb_832_832	7.03	-4.86E-02	-48.78	HS	79
	2	Co	DdGa_48_48	7.06	-4.81E-02	-16.38	HS	136
	2	Co	CbEb_854_854	7.10	-4.65E-02	-4.31	HS	109
	2	Mn	DcHe_834_834	8.24	-4.63E-02	-53.46	HS	97
		mean		7.15	-4.90E-02	-19.57		118
3	3	Mn	DbHe_829_829	8.05	-5.52E-02	-53.58	HS	88
	3	Mn	DeGb_551_551	8.21	-4.96E-02	-50.28	HS	106
	2	Mn	DcHe_834_834	8.24	-4.63E-02	-53.46	HS	97
	3	Mn	DeGd_884_884	8.24	-4.61E-02	-51.06	HS	124
	2	Mn	DeGb_790_790	8.33	-4.59E-02	-48.43	HS	136
		mean		7.89	-4.87E-02	-46.45		110.2
4	3	Mn	DbHe_829_829	8.05	-5.52E-02	-53.58	HS	88
	4	Mn	DbHe_341_341	8.11	-5.22E-02	-56.31	HS	112
	4	Mn	DeHe_487_487	8.13	-4.99E-02	-53.54	HS	88
	3	Mn	DeGb_551_551	8.21	-4.96E-02	-50.28	HS	106
	4	Mn	DcHe_887_887	8.36	-4.80E-02	-54.42	HS	97
	4	Mn	DeGb_64_64	8.50	-4.49E-02	-47.66	HS	94
	4	Mn	DeGe_2_2	8.51	-4.46E-02	-51.17	HS	88
		mean		8.27	-4.92E-02	-52.42		94.5
5	5	Fe	DeGb_82_82	5.70	-5.60E-02	-28.42	HS	184
	3	Mn	DbHe_829_829	8.05	-5.52E-02	-53.58	HS	88
	4	Mn	DbHe_341_341	8.11	-5.22E-02	-56.31	HS	112
	4	Mn	DeHe_487_487	8.13	-4.99E-02	-53.54	HS	88
	3	Mn	DeGb_551_551	8.21	-4.96E-02	-50.28	HS	106
	5	Mn	DcHe_828_828	8.37	-4.84E-02	-57.53	HS	73
	4	Mn	DeGb_64_64	8.50	-4.49E-02	-47.66	HS	94
	4	Mn	DeGe_2_2	8.51	-4.46E-02	-51.17	HS	88
		mean		7.95	-5.01E-02	-49.81		106.5

Table E.12: Final, complexes along the Pareto front after generation 5. The source generation (source gen.), metal center, and ligand identity following naming convention are all indicated. The DFT-computed values of  $\Delta G_{\text{solv}}$  (in eV) and logP are also provided. Chemical structures of the component heterocycles as well as the SMILES string for relevant functional groups (F.G.) are provided.

source gen.	metal	ligand	$\Delta G_{\text{solv}}$ (eV)	logP	heterocycle 1	heterocycle 2	F.G. (SMILES)
5	Fe	DeGb_82_82	5.70	-5.60E-02			<chem>c1ccc(c2ccc(Cl)cc2)cc1</chem>
3	Mn	DbHe_829_829	8.05	-5.52E-02			<chem>C(=O)C</chem>
4	Mn	DbHe_341_341	8.11	-5.22E-02			<chem>C(O)C(=O)C</chem>
4	Mn	DeHe_487_487	8.13	-4.99E-02			<chem>C(N)Cl</chem>
3	Mn	DeGb_551_551	8.21	-4.96E-02			<chem>C(N)C(=O)Cl</chem>
5	Mn	DcHe_828_828	8.37	-4.84E-02			<chem>C(=O)</chem>
4	Mn	DeGb_64_64	8.50	-4.49E-02			<chem>CC(=O)</chem>
4	Mn	DeGe_2_2	8.51	-4.46E-02			<chem>CO</chem>

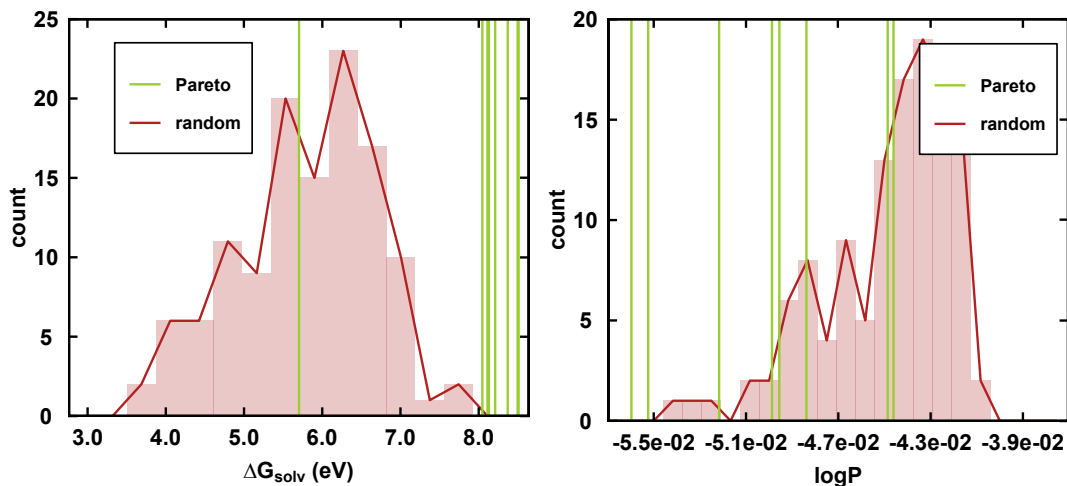


Figure E-20: Distributions of redox potentials (in eV, left) and log P values (right) for a subset of randomly-sampled complexes (red). Vertical green lines represent the respective values of the eight complexes in the final set that describes the Pareto front.

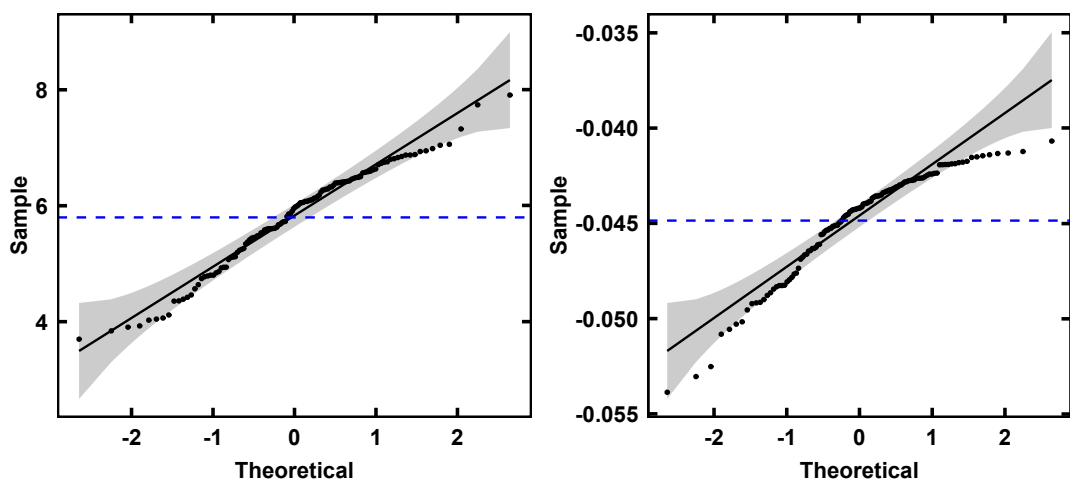


Figure E-21: Quantile-quantile plot comparing the distributions of redox potential (left) and logP (right) from the set of randomly sampled complexes (points) to a normal distribution (black line). The y-axis corresponds to the absolute value, and the x-axis corresponds to the distribution of points. The shaded region represents 95% confidence for a normal distribution, and the mean values in the sampled points are indicated by horizontal blue lines.

Table E.13: Revised geometry check tolerances used to screen complexes in this work compared to the originally-proposed values<sup>533</sup>. Tight thresholds are used on final geometries and loose tolerances are used to determine if an incomplete optimization should be resubmitted.

metric	setting	original	revised
maximum ligand connection atom angular deviation	tight	30°	22.5°
	loose	35°	27°
average ligand connection atom angular deviation	tight	15°	12°
	loose	20°	16°
RMSD difference from initial structure	tight	3.0 Å	0.3 Å
	loose	4.0 Å	0.4 Å

Table E.14: ANN model hyperparameters selected for single-task networks for logP and redox potential as well as a multitask network that predicts both logP and redox potential. The range of hyperparameters considered during optimization is shown at left. Parameters used during retraining on newly acquired data are indicated in blue. All other parameters are otherwise kept the same.

	range	logP	redox	multitask
activation	ReLU, tanh	ReLU	tanh	tanh
hidden layers	1,2,3	2	3	2
nodes	50,100,200,300	50	100	100
bypass	True, False	True	False	False
residual	True, False	True	False	False
semibatch	True, False	False	True	True
decay rate	-, 0.5, 0.75, 0.9	-	-	- (0.5)
decay interval	-,10,50,75,100,200	-	-	- (200)
dropout rate	[0,0.10]	0.01	0.012	0.092
epochs	500,1000,1500	1000	1500	1500
L <sub>2</sub> reg.	exp([-16,-8])	7.0e-11	4.8e-14	1.05e-11
batch size	[16, 32, 64, 128]	128	16	16 (64)
learning rate	exp([-5, -2])	0.003	0.004	0.004
early stopping on?	True, False	True	False	False
early stopping patience	[01,50,100,500]	100	-	-
early stopping min $\delta$	[0.0005, 0.001]	0.00075	-	-
momentum	[0, 1.0]	0.96	0.50	0.76
nesterov	True, False	True	False	False

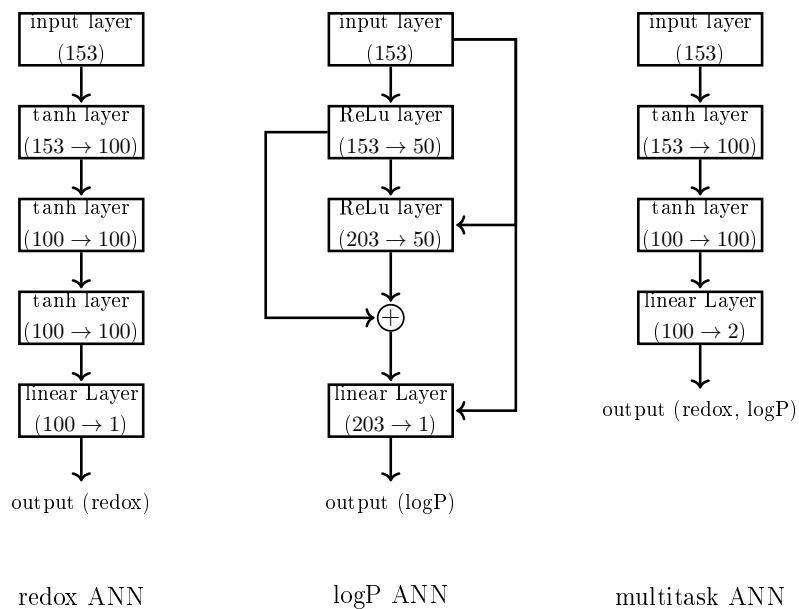


Figure E-22: Neural network architectures for single task prediction of redox potential (left) and logP (center) along with a multitask network for simultaneous prediction of both logP and redox potential (right). The number of input nodes in the input layer is the same for all models (153), and the number of input and output nodes is indicated in inset of each hidden or output layer. As shown schematically, the logP single task ANN is optimized to contain skip connections.



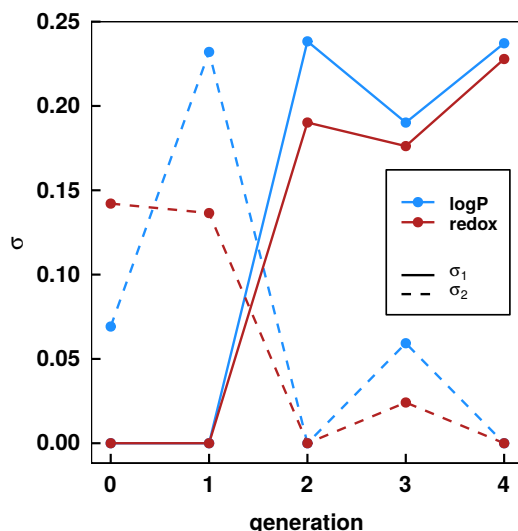


Figure E-23: The distribution of  $\sigma_1$  (solid lines) and  $\sigma_2$  (dashed lines) for maximum likelihood estimation for error bars. The two parameters are computed from 10% of test data from the multitask ANN for redox (red) and logP (blue) at each of 5 generations starting with generation 0 of EI-guided design. The parameters are fit to dimensionless errors with the model  $\varepsilon(d) \sim \mathcal{N}(0, \sigma_1^2 + d\sigma_2^2)$ .

Table E.15: Gaussian process (GP) hyperparameters including the inverse kernel width,  $\gamma$ , and regularization strength/GP baseline variance,  $\lambda$ . Hyperparameters were selected through 10-fold cross validation (CV) for predicting logP and redox independently. The hyperparameters are shown for five generations (0 through 4). The mean CV error in dimensionless units is given for each row.

$\gamma$	$\lambda$	CV error	generation	task
5.18E-03	3.24E-03	0.33	0	redox
5.18E-03	1.15E-02	0.27	1	redox
5.18E-03	7.54E-03	0.29	2	redox
7.20E-03	7.54E-03	0.29	3	redox
5.18E-03	4.94E-03	0.29	4	redox
3.73E-06	3.56E-08	0.21	0	logP
5.18E-04	3.24E-03	0.26	1	logP
7.20E-06	4.50E-07	0.29	2	logP
1.93E-05	8.69E-06	0.31	3	logP
5.18E-06	6.87E-07	0.29	4	logP

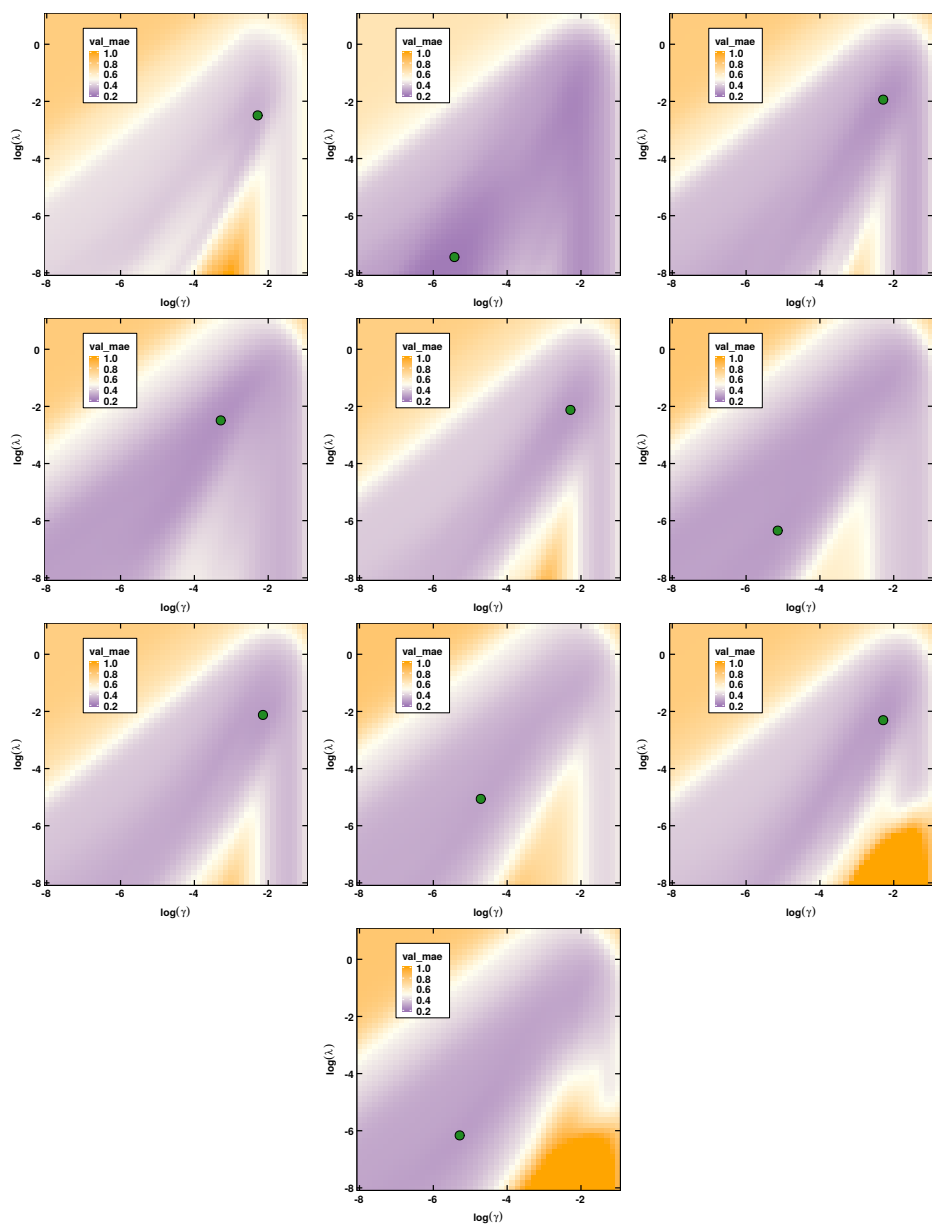


Figure E-24: Error surface showing cross-validation (CV) mean absolute error (MAE) for 10-fold CV used to selected GP hyperparameters  $\gamma$  and  $\lambda$  across generations 0 to 4 (left to right, top to bottom) grouped by redox and then logP (i.e., row 1, left to right: redox gen-0, logP gen-0, and redox gen-1; row 2, left to right: logP gen-1, redox gen-2, and logP gen-2; row 3, left to right: redox gen-3, logP gen-3, and redox gen-4; and row 4: logP gen-4). The green circles indicate the minimum values reported in Table E.15.