MIT Open Access Articles

Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Duan, Chenru et al., "Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models." *Journal of Chemical Theory and Computation* 15, 4 (April 2019): 2331–2345 doi. 10.1021/acs.jctc.9b00057 ©2019 Authors

As Published: <https://dx.doi.org/10.1021/acs.jctc.9b00057>

Publisher: American Chemical Society (ACS)

Persistent URL: <https://hdl.handle.net/1721.1/128282>

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of Use: Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models

Chenru Duan^{1,2}, Jon Paul Janet¹, Fang Liu¹, Aditya Nandy^{1,2}, and Heather J. Kulik^{1,*}

¹*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA*

02139

²*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

ABSTRACT: High-throughput computational screening for chemical discovery mandates the automated and unsupervised simulation of thousands of new molecules and materials. In challenging materials spaces, such as open shell transition metal chemistry, characterization requires time-consuming first-principles simulation that often necessitates human intervention. These calculations can frequently lead to a null result, e.g., the calculation does not converge or the molecule does not stay intact during a geometry optimization. To overcome this challenge toward realizing fully automated chemical discovery in transition metal chemistry, we have developed the first machine learning models that predict the likelihood of successful simulation outcomes. We train support vector machine and artificial neural network classifiers to predict simulation outcomes (i.e., geometry optimization result and degree of $\langle S^2 \rangle$ deviation) for a chosen electronic structure method based on chemical composition. For these static models, we achieve an area under the curve of at least 0.95, minimizing computational time spent on non-productive simulations and therefore enabling efficient chemical space exploration. We introduce a metric of model uncertainty based on the distribution of points in the latent space to systematically improve model prediction confidence. In a complementary approach, we train a convolutional neural network classification model on simulation output electronic and geometric structure time series data. This dynamic model generalizes more readily than the static classifier by becoming more predictive as input simulation length increases. Finally, we describe approaches for using these models to enable autonomous job control in transition metal complex discovery.

1. Introduction

High-throughput computation is essential for advancing molecular¹⁻⁵ and materials⁶⁻¹² discovery and in data generation¹³⁻¹⁵ for machine learning property prediction models¹⁶⁻²¹ and potentials²²⁻²⁶. Such efforts mandate the automated^{7-8, 15, 27-28} simulation of thousands of new molecules and materials. While this task is mostly straightforward in ground state organic chemistry, it is far more challenging for open shell transition metal chemistry central to the development of new functional catalysts^{4, 29-36} and materials³⁷⁻⁴⁵. Characterization of open shell transition metal complexes requires time-consuming first-principles simulation, typically^{4, 29, 46-47} with density functional theory (DFT), because few semi-empirical theories⁴⁸ or force fields⁴⁹⁻⁵⁰ can describe the multiple accessible oxidation and spin states of transition metal complexes in a balanced manner. Such simulations typically require manual validation that the simulation was successful, e.g., if the structure optimized to the intended geometry^{15, 48} and if the electronic structure reflects the expected spin state. Automated approaches have only begun to be applied, e.g., for functional selection⁵¹, active space selection in wavefunction theory calculations⁵², or for interpreting reactivity⁵³⁻⁵⁵. If typical calculations fail manual validation checks, it at best enriches the knowledge of the researcher carrying out the calculation, who then repeats the procedure. Only through trial and error and growing expertise do individual success rates rise.

Although manually setting up and observing simulation behavior is an essential component of developing individual expertise in computational inorganic chemistry, high-throughput discovery requires a different approach. Here, even a slight improvement in success rate can have a dramatic impact: the high computational cost of each geometry optimization (i.e., dozens of hours per DFT geometry optimization), whether successful or not, can mean that the

time to discovery or obtaining predictive data-driven models is greatly slowed by failed calculations. Furthermore, when property prediction models are used in discovery mode to guide which molecules to simulate with DFT^{5, 15}, it is crucial that knowledge of what makes a feasible molecule for simulation is imparted to such a workflow.

To overcome this challenge toward realizing fully autonomous chemical discovery, it becomes necessary to develop computational tools that identify and prevent unsuccessful simulations. Here, we take a data driven approach, motivated by the recent success of machine learning (ML) models to map complex chemical composition relationships to properties for rapid property prediction^{18-21, 56-57} and classification^{56, 58-59}. We leverage the representations we have developed⁶⁰⁻⁶¹ for inorganic chemistry that have shown significant transferability⁶⁰ among key properties, such as metal-ligand bond length⁶⁰⁻⁶¹, spin state prediction⁶⁰⁻⁶¹, redox potential⁶⁰, and frontier orbital energetics¹⁵. Where this work differs from prior efforts is that our goal is to build models that predict the likelihood of success or failure of simulation outcomes as fundamental logic components in an automated simulation workflow. We anticipate our approach to be a useful complement to other machine learning approaches designed to accelerate simulation, e.g., through improving the optimization procedure⁶²⁻⁶⁴ or bootstrapping methods⁶⁵⁻⁶⁷.

The value of “failed experiments” has been previously demonstrated in experimental workflows for materials synthesis⁶⁸. Predicting simulation outcomes based on an appropriate heuristic molecular representation⁶⁰ requires collection of a large number of failed and successful simulations as training data prior to initial prediction with a “static” data driven model. A unique opportunity for first-principles simulation is the ease with which *in situ* measurements of simulation fitness may be obtained: electronic and geometric properties are trivial to compute at each step of a calculation, in contrast with the efforts required during experiments. Here, we thus

also explore “dynamic” data driven models that can learn from such information to, once trained, process simulation data to predict outcomes on-the-fly.

In addition to “static” and “dynamic” models for simulation outcome prediction, we develop a new uncertainty metric to assess model confidence^{24, 61, 69-73} especially in out-of-distribution predictions. The rest of this article is outlined as follows. In section 2, we provide the Methods of the work, including machine learning models, representations, and the simulated data sets. In section 3, we present results and discussion on static and dynamic classifiers that predict simulation outcomes as well as autonomous job control strategies using new uncertainty metrics. Finally, in section 4, we provide our conclusions and outlook.

2. Methods

2a. Representations for Machine Learning Models.

Static model representations. Revised autocorrelations (RACs) are selected as connectivity-only representations that require no precise geometric information and have previously demonstrated performance^{15, 60} in inorganic chemistry. RACs are sums of products and differences of five atom-wise heuristic properties, i.e., topology, identity, electronegativity, covalent radius, and nuclear charge, on the molecular graph. As motivated previously⁶⁰, we limit the maximum depth, d , of bond paths over which two atomic properties are correlated to three bonds. There are $30d+30$ product-based RACs (i.e., $6d+6$ for each property) that arise from differing starting points (e.g., metal-centered or ligand-centered). Ligand-centered RACs are averaged either over all equatorial or all axial ligands following the convention in Ref. 60, and any effect on model performance for lower symmetry compounds is discussed in Sec. 3a. After excluding identity-based and zero-depth difference RACs, there are $12d$ additional non-trivial difference RACs. From this total of 156 product and difference $d \leq 3$ RACs, 5 are constant

across the octahedral complexes studied. To this set, we add i) oxidation state and spin multiplicity to distinguish complexes with the same composition and connectivity and ii) total net ligand charge, as motivated by a good correlation with the geometry failure rate in optimizations, for a final feature set size of 154 (Supporting Information Figure S1).

Dynamic model representation. We select descriptors for the dynamic classifier that are directly derived from the evolving electronic and geometric structure during first-principles (i.e., DFT) geometry optimizations. Electronic-structure-derived descriptors are obtained from components of: i) Mulliken charges, ii) the bond order matrix⁷⁴, and iii) the matrix of the gradient of the energy in Cartesian coordinates. All descriptors were obtained from a developer version of TeraChem⁷⁵⁻⁷⁶, which was modified to write each property at every step of the geometry optimization. Descriptors include the metal Mulliken charge and the Mulliken charge of connecting atoms averaged separately for the equatorial and axial ligand sets. We include the metal bond valence of the bond order matrix and the bond order between the metal and connecting atoms averaged separately for the equatorial and axial ligand sets. We also incorporate the first four values from the singular value decomposition of both the full bond order matrix and a modified bond order matrix with diagonal elements set to zero. We extract the metal gradient norm, the gradient norm of connecting atoms averaged separately over equatorial and axial ligands, the maximum gradient norm on any atom, root mean square gradient, and three singular values of the gradient matrix. The maximum norm and three singular values are also obtained from the gradient matrix after its conversion into metal-centered internal coordinates. In total, there are 26 electronic-structure-derived features for each complex (Supporting Information Table S1). We also incorporated the eight geometric metrics of octahedral complex structure introduced in prior work¹⁵ that are used to define geometry optimization success and

failure in the classifier models (Supporting Information Table S2 and see Sec. 3). Including these metrics in the feature set guarantees correct classification for the optimization result at very long times approaching full convergence or failure of the calculation. The value for the normalization procedure for both types of dynamic descriptors is provided in Supporting Information Table S3. Both types of featurizations used in the text are provided for each complex as data in the Supporting Information.

For a geometry optimization trajectory, the total feature set size is $N_{\text{step}} \times N_{\text{feature}}$, where N_{step} is the number of optimization steps used as inputs and N_{feature} is the number of features extracted in each step (here, 34) of the optimization. Several values of N_{step} are used to train models, up to a maximum of 40 to limit computational time, as this largest N_{step} requires an average of 13 GPU hours (Supporting Information Figure S2).

2b. Machine Learning Model Parameters and Architectures.

Static classification models. Support vector machines (SVMs) and artificial neural networks (ANNs) are trained on RAC descriptors to form static models for geometry and spin classification. An SVM constructs hyperplanes that maximize the margin between distinctly labeled classes of training data. In kernel SVM, the input data is transformed into a higher-dimensional space by the so-called kernel trick, and the classification task is carried out in the kernel space. We trained an SVM with a Gaussian kernel using the scikit-learn package⁷⁷ and selected hyperparameters with Hyperopt⁷⁸. Two hyperparameters are required for the SVM, kernel width and the lower bound of the fraction of support vectors, which controls regularization⁷⁹ (Supporting Information Table S4). The validation set area under curve (AUC) was used as the metric for hyperparameter selection.

The ANN classifier flexibly transforms input features via a series of nonlinear functions

(i.e., layers) before entering the final linear classification layer⁸⁰ (Supporting Information Table S4). ANN models with two hidden layers were trained using the Keras software package⁸¹ with Tensorflow⁸² as the backend. We refined hyperparameters from Hyperopt⁷⁸ by trial and error using the validation set AUC as the metric (Supporting Information Table S4). To accelerate hyperparameter selection, we applied earlystopping by monitoring the validation set AUC with a patience of 300 epochs, and these hyperparameters were used to train the final model for 2,000 epochs.

For both static classifiers, data set preparation was as follows. We partitioned the entire *spectrochemical data set* (see Sec. 2c) into 80% train and 20% test. The initial training set was split further into 80% training (i.e., 64% of the total set) and 20% validation (i.e., 16% of the total set). During hyperparameter selection, models were trained on the training set and evaluated by the validation set. To obtain the out-of-sample performance, models were trained on the combination of the training and validation set and then evaluated on the out-of-sample test set.

Dynamic classification models. First applied on image classification and natural language processing⁸³⁻⁸⁵, convolutional neural networks (CNNs) are now frequently used in studying chemical systems⁸⁶⁻⁹³ because of their weight sharing and translational invariance. The ease of training and interpretability of CNNs are why we selected them for the time-series data in the geometry optimizations studied here rather than other models such as recurrent neural networks. We train a dynamic CNN-based classifier that consists of i) a one-dimensional convolutional layer along the axis of optimization steps, which convolves effects of all features at various steps, ii) one fully connected layer that further extracts information of each convolutional filter, and iii) an output layer for the final classification task (Supporting Information Figure S3). We use L2 regularization and dropout in both the convolutional layer and the fully connected layer,

and batch normalization⁹⁴ is used during model training. We implemented our dynamic model using the Keras software package⁸¹ with Tensorflow⁸² as the backend. As with the ANN static classifier, hyperparameters are selected using the validation AUC as the metric, during which earlystopping with 300-epoch patience is activated (Supporting Information Tables S5-S6).

2c. Simulation Details.

Octahedral transition metal complexes with Cr, Mn, Fe, and Co in +2 and +3 oxidation states were studied in high-spin (HS) and low-spin (LS) states defined as: quintet-singlet for d^6 Co(III)/Fe(II), sextet-doublet for d^5 Fe(III)/Mn(II), quintet-singlet for d^4 Mn(III)/Cr(II), and quartet-doublet for both d^3 Cr(III) and d^7 Co(II). The *spectrochemical data set* generated for this work consisted of these four metals in complex with 16 ligands from the experimental spectrochemical series⁹⁵ for a total of 3,622 complexes (Supporting Information Table S7). These octahedral complexes are formed by a single equatorial ligand type and one (2,136 complexes) or up to two (1,486 complexes) distinct axial ligands. From this set, 3,477 complexes were used in the static geometry classifier, 1,973 non-singlet complexes were used for the $\langle S^2 \rangle$ deviation classifier, and a 2,157 complex subset was used for training the dynamic classifier. To generate this data, DFT geometry optimizations with the B3LYP⁹⁶⁻⁹⁸ hybrid functional were carried out using the graphical-processing unit (GPU)-accelerated electronic structure code TeraChem⁷⁵⁻⁷⁶. The default definition of B3LYP in TeraChem employs the VWN1-RPA⁹⁹ form for the local density approximation correlation component. The LANL2DZ effective core potential¹⁰⁰ was used for metals and the 6-31G* basis for all other atoms. Besides singlet states, all other calculations were carried out in an unrestricted formalism that employed level shifting¹⁰¹. In those cases, virtual and occupied orbitals level shifts of 1.0 and 0.1 Ha, respectively, were employed to enable self-consistent field convergence to an unrestricted

solution.

All initial geometries were generated with molSimplify^{27, 102}, which uses OpenBabel^{28, 103} as a backend for organic structure generation and force field optimization, and calculations were automated by molSimplify Automatic Design (mAD)^{5, 15}, with up to five job resubmissions for 24 hour periods each. Geometry optimizations used the L-BFGS algorithm in translation rotation internal coordinates (TRIC)¹⁰⁴ as implemented in TeraChem to the default tolerances of 4.5×10^{-4} hartree/bohr for the maximum gradient and 1×10^{-6} hartree for the change in energy between steps. For the static classifier, a geometry check¹⁵ was carried out before each resubmission. Only the complexes that passed all geometric checks with loose cutoffs were resubmitted, and all other structures were labeled as bad geometries with no further resubmission attempts. For the dynamic classifier, the 40 steps were typically completed in the first submission. If a job finished in under 40 steps, the features of the last step were used to pad the descriptors to the appropriate dimension (see Sec. 2a). Full details of the composition of all simulated complexes are provided in the Supporting Information, and all optimization trajectories are also available for download¹⁰⁵.

For out-of-distribution testing of the static classifier, a *diverse data set* of 5,451 octahedral complexes^{5, 60-61, 102} was collected from prior work on spin splitting energy ML model training, spin-crossover complex design, and an enumerative exploration of transition metal complexes with small ligands. For out-of-distribution testing of the dynamic classifier, an *oxo data set* of 515 model catalysts was collected from parallel work in preparation in our lab. The simulation procedure for these data sets is described in prior work and largely identical to the new data generated here, but any relevant distinctions are provided in greater detail in Supporting Information Text S1. The *diverse data set* was selected for its large size, whereas the

oxo data set was chosen due to all structures having the electronic descriptor data needed by the dynamic classifier. The out-of-distribution data set properties and structures are provided in the Supporting Information, and optimization trajectories are available for download¹⁰⁵.

3. Results and Discussion

3a. Classification Task Label Definitions

We first define the labels used in the classification task for the automated detection of successful vs. unsuccessful electronic structure calculations. For geometry classification, we employ previously introduced¹⁵ geometric metrics to quantify if structural integrity of transition metal complexes is lost during geometry optimization. The metrics for an intact octahedral complex include: i) preservation of six-fold coordination, ii) minimal deviation of ligand-metal-ligand angles from expected values (i.e., 90° or 180°), iii) no more than moderate (ca. 0.4 Å) asymmetry between equatorial and axial metal-ligand bond lengths, and iv) preserved qualitative structure of the ligand, as judged through root mean square deviation (RMSD) and angles for linear ligands (Figure 1 and Supporting Information Table S2). In our data sets, converged geometry optimizations that do not violate any of these metrics are labeled as “good”, and ones that violate any of them are labeled as “bad”.

Figure 1. Examples of “good” (top) and “bad” (bottom) geometries (left) and $\langle S^2 \rangle$ deviations

(right). For the spin classifier labels, the spin density is shown with an isosurface of 0.01 e (orange for positive spin density, and green for negative spin density).

Calculations generated for static classification training data are carried out in up to five 24 hour periods, and after each 24 hour period we apply loose cutoffs (25-35% higher than standard cutoffs applied to converged structures, see Supporting Information Table S2). If these loose metrics are not met, the geometry is also labeled as “bad” and the calculation is not resubmitted (Supporting Information Table S8). Otherwise, these looser cutoffs allow for fluctuations in structure during geometry optimization from which the optimizer can be expected to recover upon restart: around 10% of converged “good” structures fall between loose and strict cutoffs for at least part of the optimization trajectory, and almost all (97%) of converged “bad” geometries exceed both strict and loose cutoffs (Supporting Information Table S9). When the maximum number of resubmissions has been reached, we apply the loose geometry metric cutoffs. We then label a data point as “bad” if the geometry fails the loose cutoffs or otherwise exclude the point from the data set (Supporting Information Table S8). Overall, the “good” (1,744 points) and “bad” (1,733 points) geometry labels form a balanced data set, with only a small fraction (126 points) of the original data set excluded due to slow calculations not reaching completion after five submissions (Supporting Information Table S8).

For electronic structure, we quantify the degree of deviation of the expectation value of the $\hat{S}^2$ operator from the assigned simulation value (i.e., $S(S+1)$) for all unrestricted (i.e., non-singlet in this work) calculations:

$$\langle \hat{S}^2 \rangle - S(S+1) > C \tag{1}$$

Although broken symmetry solutions can have utility in transition metal chemistry, we employ this metric as an indication of cases where the Kohn-Sham DFT solution may be ambiguous. For

fully converged calculations, we use $C = 1.0$ and label the electronic structure as “bad” if eqn. (1) is satisfied. This cutoff was chosen because it corresponds to one net electron occupation of a distinct spin up versus spin down orbital (Figure 1). We label any converged simulations that do not satisfy the inequality as “good”. During resubmission, loose tolerances of up to $C = 2.0$ are allowed for the eqn. (1) metric for reasons similar to that for the geometry classification task. At the end of the maximum number of resubmissions, unconverged simulations are labeled as “bad” only if they exceed the loose $\langle S^2 \rangle$ deviation cutoff. The fact that no singlets are included in this data set leads to a smaller 1,973-complex spin classifier training set than for the 3,477-complex geometry training set. The spin classifier data set is also less balanced, with more “good” (1,649 points) than “bad” (324 points) results (Supporting Information Table S8).

3b. No-cost Static Classification of Simulation Outcomes

Important questions to answer regarding the utility of ML models for avoiding non-productive simulations in chemical discovery are: i) whether the ML models can be trained on modest data sets to avoid requiring large numbers of failed calculations for model training, ii) if the ML model is sufficiently predictive to guide autonomous job control in chemical space exploration, and iii) whether ML model failure can be predicted with uncertainty measures or otherwise biased to control the effect of model inaccuracy on chemical discovery. To address the first point, we trained separate SVM and ANN classifier models to predict geometry optimization outcomes and $\langle S^2 \rangle$ deviations (see examples in Figure 1). For both types of classifiers, the “good” outcome is used as the positive class (i.e., 1) in the binary classification, the “bad” label is assigned as 0, and a standard 0.5 cutoff is used with “good” corresponding to values > 0.5 and “bad” for values ≤ 0.5 . The SVM correctly classifies most (84%) geometry optimization outcomes, with a range of probabilistic errors for the misclassified cases (Figure 2

and Table 1). For $\langle S^2 \rangle$ deviation, performance is even better (95% accuracy), with few good cases being misclassified, and a slightly higher rate of bad structures being labeled as good (Figure 2 and Table 1).

Figure 2. SVM classifier prediction probability and ROC for geometry (left) and $\langle S^2 \rangle$ deviation (right). All data points are grouped by actual good (blue circles) and bad (red squares) outcomes. The SVM probability ranges from 0.0 (bad) to 1.0 (good), and a standard threshold (0.5) is shown as a black dashed line. The ROC is indicated for the test set (green line) and a random guess (gray line).

Table 1. The accuracy and AUC of SVM and ANN models for both geometry and spin classification tasks, as judged on the validation set for hyperparameter selection and the set-aside test set.

Model	Validation accuracy	Validation AUC	Test accuracy	Test AUC
Geometry				
SVM	0.85	0.93	0.84	0.92
ANN	0.88	0.95	0.88	0.95
$\langle S^2 \rangle$ deviation				
SVM	0.95	0.98	0.95	0.98

ANN	0.95	0.98	0.96	0.97
-----	------	------	------	------

Rather than accuracy alone, which is sensitive to the data set balance¹⁰⁶, characteristics of the receiver operating curve (ROC), which relates true and false positive rates, provide a more universal measure of classifier performance¹⁰⁶⁻¹⁰⁷. The area under the curve (AUC) of the ROC for a perfect model is 1.0 and 0.5 for a random guess, regardless of the balance in the data set between the two labels. The SVM geometry and spin classifier test set AUC values of 0.92 and 0.98, respectively, indicate excellent discrimination of “good” and “bad” results (Figure 2 and Table 1). The ANN model improves upon the SVM geometry classifier with an 88% test set accuracy and AUC of 0.95 (Table 1). This performance is insensitive to the axial averaging in the ligand-centered RACs, with comparable AUCs for models trained and tested only on the axially-symmetric subset of the *spectrochemical data set* (Supporting Information Table S10). The more complex ANN spin classifier model performs comparably to the already good SVM spin classifier, likely due to the simplicity of the prediction task in which a few ligand variables correlate strongly to outcomes (Table 1 and Supporting Information Figure S4).

The geometry classification task appears more challenging due to the fact that many factors can dictate success or failure. These factors include path-dependent choices made by the optimization algorithm, i.e., a different result might arise from an alternative optimizer, code version, numeric precision, or varied optimization settings. For example, the failed optimization of a quintet Cr(II)(py)₄(SH₂)₂ complex that occurs when the two axial hydrogen sulfide ligands detach is correctly predicted by the ANN (0.17, “bad”) but not the SVM (0.85, “good”). This complex is moderately challenging because the equivalent singlet Cr(II) complex also has a “bad” structure but the quartet and doublet Cr(III) complexes are “good” (see data in Supporting Information). In the case of quintet Cr(II)(PH₃)₄(py)₂, two equatorial phosphine ligands detach

during optimization but both models predict the structure incorrectly as “good” (ANN: 0.73, SVM: 0.86). This prediction task is relatively difficult for both the SVM and ANN because the first four nearest neighbors in feature space to the complex, including singlet $\text{Cr(II)(PH}_3)_4(\text{py})_2$, all have a “good” geometry. In this same complex, the “good” $\langle S^2 \rangle$ deviation is correctly predicted by the ANN spin classifier. This spin classification task is simpler than the geometry classification because four out of the five nearest neighbors also have “good” spin classification labels. Alternatively, there are cases where the geometry is correctly classified, but the $\langle S^2 \rangle$ deviation is misclassified by the ANN. The doublet $\text{Mn(II)(NCO}^-\text{)}_6$ complex is predicted correctly by the ANN to have a “good” geometry but misclassified as “good” by the ANN spin classifier (SVM: 0.64) likely owing to close proximity of four nearest neighbors in RAC space with “good” spin labels.

To evaluate whether the accuracy of our models is sufficient for application in autonomous chemical space exploration, we assess the potential combinatorial performance of the spin and geometry ANN classifiers to predict the “simulation promise” of a transition metal complex. Here, the positive label is assigned only to structures that pass both the geometry and spin classification tasks (i.e., are “good”), otherwise either “bad” spin or geometry labels triggers a negative result (Figure 3). The confusion matrix of true and false positives and negatives reveals good combined accuracy of 88%, suggesting that the geometry classifier performance is still the limiting factor (Figure 3). We can also compare the frequency of false positives (FPs, 48 of 683, FP rate: 0.13), i.e., cases where a simulation is predicted to be successful that would fail and false negatives (FNs, 35 of 683, FN rate: 0.11), i.e., cases where a simulation is predicted to fail but would have succeeded (Figure 3). Although most FPs and FNs are due to geometry misclassification, more FPs (9 of 48) arise from spin misclassification than FNs (3 of 35, see

Supporting Information Table S11).

Figure 3. a) Illustration of the decision workflow for job control using both geometry and spin classifiers. Once the structure passes both classifiers, input files can be generated (e.g., by mAD) to run the job. b) The confusion matrix of positive (P) and negative (N) results colored by whether the predicted and actual result is the same (true) or different (false) for an out-of-sample 683 molecule test set. The top number in bold indicates the results of the combined static classifiers with a probability threshold for label assignment of 0.5 and the result in parentheses below is for the case where the probability threshold is set to 0.3.

Applying this model as-is in autonomous job control on the out-of-sample 683 complex test set would lead to 48% of all jobs being executed to cover 88% of the valid compound space, dramatically reducing computational cost by not carrying out 352 simulations without significantly reducing the amount of data acquired. In a resource-poor mode, the goal would be to avoid any failed simulations, whereas in a lead-poor mode, the goal would be to simulate all leads, even those that had a significant probability of failure. In this latter scenario, we can examine the effect of favoring FNs by reducing the prediction probability threshold to 0.3. This strategy reduces the number of FNs (from 35 to 14) at the cost of increasing FPs (from 48 to 74, see Figure 3). Examining the individual classification labels, spin-classifier-derived FNs are eliminated at this threshold, and 23 of the 26 new FPs are also due to the geometry being misclassified (Supporting Information Table S11). This threshold choice would increase chemical space exploration to 95% of the valid compounds by carrying out calculations on 55% of the space, where the increased number of simulated compounds are about half true leads and

the other half would fail (i.e., 21 and 26, see Figure 3).

Quantitative estimates of classification model prediction uncertainty should identify the model's domain of applicability. For the out-of-sample *spectrochemical data set* test partition, the ANN model prediction probability itself serves as a simple indicator of prediction uncertainty, with far more correct points at probability values close to 0 or 1 than mislabeled points (Supporting Information Figure S5). By excluding points with probability values far from 0 or 1, model accuracy can be systematically increased, and 98% accuracy can be reached by making predictions on only half of the out-of-sample test partition (Supporting Information Figure S6). However, there are challenging cases where the model can be quite confident and still wrong. The highly-charged, sextet $\text{Mn(II)(S}^{2-}\text{)}_4(\text{H}_2\text{O})_2$ complex is closest in feature space only to other “bad” geometries due to the strong role of ligand charge in optimization outcomes (Supporting Information Figure S4). Thus the model (ANN: 0.11) confidently predicts a “bad” geometry, but this complex optimizes smoothly to a “good” geometry instead.

In out-of-distribution applications, uncertainty estimates become essential. Applying the ANN geometry classifier trained on the *spectrochemical data set* to the out-of-distribution *diverse data set* leads to a much lower AUC (0.6) and accuracy (59%, Supporting Information Table S10). Practical autonomous control in chemical discovery on distinct complexes thus motivates quantitative measures of prediction uncertainty. However, in the out-of-distribution test, prediction probability cannot be used to improve upon the error, motivating investigation of alternative metrics (Supporting Information Figures S5-S6).

We considered the ensemble variance¹⁰⁸, which is widely-applied in prediction^{24, 69} and classifier models¹⁰⁹ as an uncertainty metric. However, ensemble variance exhibits the same limitation as the prediction probability: it performs well for out-of-sample test cases but cannot

be used to improve upon out-of-distribution accuracy (Supporting Information Figures S5-S6). Feature space distance metrics^{5, 61, 70, 72, 110-112} have previously been motivated as a potential uncertainty measure, but because we did not carry out feature selection⁶⁰ in this work, the utility of feature space distances for estimating prediction uncertainty is limited (Supporting Information Figure S7). We thus introduce an uncertainty measure that depends directly on the data distribution in the ANN latent space¹¹³, i.e., the space spanned by the last layer of neurons before the output layer.

For reliable predictions, the latent space i) should be smooth, with the closest compounds to a test point all having the same label, and ii) the test point should not be too distant in latent space to available training data. We quantify the latent space entropy (LSE) as a measure of these two properties. For a sample point i in class j , the same-class nearest neighbors contribute to a score evaluated as:

$$p_i^{(j)} = f(2) + \sum_{k \in j} f(d_{ik}) \quad (2)$$

where d_{ik} is the latent space Euclidean distance between the test point i and k th training data point normalized by the averaged 5-nearest-neighbor distance in the training data. The sum is truncated to include only the first 300 nearest neighbors to the test point. The $f(x)$ is a scoring function selected here to be:

$$f(x) = \exp(-x^2) \quad (3)$$

Thus, the first term, $f(2)$ in eqn. (2) acts to regularize the distance score by being the dominant term in $p_i^{(j)}$ when all neighbors of i in class j are distant but contributing little if numerous close neighbors are present. The score of each class is normalized:

$$\tilde{p}_i^{(j)} = \frac{p_i^{(j)}}{\sum_j p_i^{(j)}} \quad (4)$$

where J is the total number of classes (here, 2). The LSE is then calculated as the information entropy of these normalized scores:

$$s_i = -\sum_j \tilde{p}_i^{(j)} \ln(\tilde{p}_i^{(j)}) \quad (5)$$

obtained in practice as an average of 10 evaluations from a model trained on 90% of the data with 10-fold cross-validation. The LSE thus ranges from 0 for perfect confidence to $\ln(J)$ for no confidence (i.e., here $\ln(2) = 0.693$). This measure directly captures both disagreement between proximal latent space points as well as high latent space distance of the test point to all training data.

The LSE depends on a well-converged latent space representation obtained through careful model training. Changes in the latent space, as visualized¹¹³ through t-distributed stochastic neighbor embedding (t-SNE)¹¹⁴, and the LSE averaged over test points during model training demonstrate the useful properties of the LSE in a well-trained model (Figure 4). Samples that appear to be initially randomly distributed become more uniformly separated into a clear region of “bad” and “good” labeled structures in the latent space during model training. By definition, this training process therefore smoothly decreases the average test set LSE (Figure 4).

Figure 4. Changes in the average test set LSE (squares with black outline) with training epoch for an ANN geometry classifier. A t-SNE visualization of test set data is shown in inset colored by good (blue translucent circles) and bad (red translucent squares) geometries for the latent space of selected training epochs (0, 500, 1000, and 1500 as indicated by arrows to the respective shaded gray filled squares). A 10% randomized subset of the *spectrochemical test set* data is shown for clarity.

The LSE provides an adjustable metric to separate out the *diverse data set* test points for which the classification model trained on the *spectrochemical data set* is confident (Figure 5). By restricting predictions to points with more conservative (i.e., smaller) LSE values, the classifier makes fewer predictions but the accuracy on the predicted points inside the LSE cutoff rises smoothly. For an LSE cutoff of 0.35, predictions will be made for only 42% of the *diverse data points* but at an accuracy of 72% that significantly improves upon the model accuracy on the full *diverse data set* and with no significant effect on relative FP/FN rates (Figure 5 and Supporting Information Figure S8). The accuracy on the data outside the cutoff then drops to only 44%, and similar observations hold regardless of the LSE cutoff chosen for the analysis (Figure 5). When the cutoff is instead set by the ensemble model variance or prediction probability, no improvement in accuracy is observed for the inside cutoff partition of the *diverse data set* (Figure 5 and Supporting Information Figures S6 and S9). The quintet $\text{Mn(III)(py)}_4(\text{NCO}^-)(\text{O}_2^{2-})$ complex provides an example of how the LSE can provide a superior

performance as compared to ensemble variance metrics. This complex converges to a “good” geometry, but the ensemble prediction is “bad” (0.47) with a relatively low variance (0.05) despite spanning the decision boundary. The high 0.67 LSE of this complex would prevent prediction on this model and likely encapsulates the competing effects of a modest ligand total charge with potential Jahn-Teller effects for high-spin Mn(III) complexes that make label assignment challenging.

Figure 5. Prediction accuracy (top) and fraction of data inside the cutoff (bottom) with respect to LSE (left) and ensemble variance (right) cutoffs. Data accuracy and fraction inside the cutoff is shown in green and outside is shown in orange. The gray arrow indicates the direction of increasingly conservative cutoffs. The models were trained on the *spectrochemical data set* and tested on the *diverse data set*.

The LSE retains the good performance of ensemble and prediction probability metrics in the out-of-sample case, and applying the same 0.35 LSE cutoff increases the accuracy above 95% on two thirds of the original *spectrochemical* test set (Supporting Information Figure S10).

Reversing the nature of the train and test data sets to train on *diverse* data but test on the *spectrochemical* data set shows the transferability of good LSE performance, with 88% accuracy achieved at an LSE cutoff of 0.35 despite no significant improvement using other cutoff metrics (Supporting Information Figure S11). This superior performance of the LSE-tuned *diverse*-trained model over the *spectrochemical*-trained model is likely due to the larger and more varied training set in the former case. The LSE value may be tuned by the user to determine the degree of certainty that the static model must have in order to reject a simulation. While the classifier model will not be transferable to all new methods and simulation parameters, we expect that it should hold for small changes (e.g., to other hybrid or semi-local DFT functionals) that are not expected to significantly change simulation outcomes.

3c. Dynamic Classification of Simulation Outcomes

While the static classifier uses only the final result of a simulation to generate a prediction based on the molecular representation, a complementary approach is to use the time-dependent optimization data to identify “dynamic” signatures of simulation failure. The training set for the static model (i.e., chemical composition and final optimization outcome) depends sensitively on the simulation parameters used to generate that simulation. By training on properties generated during a geometry optimization, the dynamic model should instead learn what electronic and geometric descriptors predict a failed simulation outcome. To systematically investigate the relative merit of this complementary approach, our dynamic classifier models are trained on a mixture of electronic and geometric structure descriptors in lieu of RACs. These descriptors provide general information about the metal coordination environment, gradient, and structure (Supporting Information Tables S1-S2). At long times, the geometric descriptors should correspond exactly to the label assignment of the geometry classification task. Electronic

descriptors can be expected to provide a more indirect assessment, e.g., if the metal-ligand bond order drops to zero when a ligand becomes detached.

In building a dynamic model, we must choose how much simulation information is to be accumulated before a prediction is made because a larger number of optimization steps, N_{step} , should be expected to lead to more confident predictions but require more computational time. To evaluate such a potential trade-off, we built a series of independent CNN models trained and tested on the *spectrochemical data set* to predict geometry optimization outcomes using electronic and geometric descriptors (see Sec. 2a) obtained from $N_{\text{step}} = 2$ up to 40 (Figure 6). Dynamic model performance increases with N_{step} , but good out-of-sample performance (accuracy: 83%, AUC: 0.91) is observed even with the $N_{\text{step}} = 2$ model (Figure 6). An average of only 2 hours of GPU time is needed to obtain data on a new complex for model prediction, and its performance is only slightly worse than the static model (i.e., 88% accuracy and 0.95 AUC, see Table 1). The $N_{\text{step}} = 40$ model requires an average of 13 GPU hours prior to prediction but exhibits superior model performance (accuracy: 96%, AUC: 0.99) over the static classifier (Figure 6). Analysis of the 4% of simulations still misclassified by this model reveals cases that have distinct long-time outcomes from what is observed within the first 40 steps, i.e., structures that are “bad” at short time converging to “good” structures or vice versa (Supporting Information Figure S12).

Figure 6. AUC (top) and accuracy (bottom) for models trained at N_{step} values ranging from 2 to 40 (solid symbols) on the *spectrochemical* data set. Test set performance for a set-aside *spectrochemical* partition (black circles) and an out-of-distribution *oxo* data set (red diamonds) are shown alongside results obtained by resizing additional input lengths to the closest available model (translucent black circles and translucent red diamonds).

All dynamic models with $N_{\text{step}} > 10$ outperform the static model on the *spectrochemical* test set, and more intermediate models would only require about six to seven hours of simulation time prior to prediction (Supporting Information Figure S2). To overcome only being able to make predictions at N_{step} values corresponding to trained models, we made predictions on resized features with linear interpolation¹¹⁵ at intermediate $N_{\text{step}} = 7$ to 37 in 5 step increments (Figure 6). Prediction performance at these intermediate values was comparable to the explicitly trained models, suggesting predictions can be carried out at any time during the initial 40 optimization steps (Figure 6).

To test transferability to out-of-distribution data, we evaluated the dynamic classifier on the *oxo* data set (see Sec. 2 and Supporting Information Text S1, Table S10, and Figures S13-S14). Although lower accuracy is observed on the *oxo* data set for the $N_{\text{step}} = 2$ model (accuracy: 78% and AUC: 0.71) in comparison to the *spectrochemical* test set, this performance is still better than that observed for the static classifier (Supporting Information Table S10). The model

performance on the *oxo data set* also becomes nearly comparable to the *spectrochemical* test set for $N_{\text{step}} \geq 20$ (e.g., $N_{\text{step}} = 40$, accuracy: 95% and AUC: 0.98, see Figure 6).

Here, we have focused on the geometry classification task due to the high performance of the static spin classifier model (see Sec. 3b), but good performance is also observed for a dynamic $\langle S^2 \rangle$ deviation model even at $N_{\text{step}} = 2$ (accuracy: 94% and AUC: 0.95, Supporting Information Figure S15). Dynamic performance exceeds the static model at $N_{\text{step}} = 20$ with AUC approaching 1.0 at $N_{\text{step}} = 40$ (Supporting Information Figure S15). The model also generalizes well to the *oxo data set* where performance is comparable to *spectrochemical* test set results across all N_{step} values even though the *oxo data set* contains spin and oxidation states absent from training data (Supporting Information Text S1 and Figure S15).

To assess the relative importance of electronic and geometric descriptors in the prediction task at different time points, we compared the models trained only on electronic or geometric descriptor subsets to the full model (Figure 7 and Supporting Information Figure S16 and Table S5). At short times ($N_{\text{step}}=2$), the electronic model performance is superior to the geometric model and comparable to the full model when tested on either the *spectrochemical* or *oxo data sets* (Figure 7 and Supporting Information Figure S16). The electronic model can predict “bad” geometry optimization outcomes at shorter timescales than the geometric model alone because electronic properties (e.g., the metal bond valence) can foretell poor health before the structure rearranges (Supporting Information Figure S17). At intermediate ($N_{\text{step}} = 15$) to long times ($N_{\text{step}} = 40$), the geometric model performance rises much faster than the electronic model, outperforming the latter (Figure 7). At these higher step numbers, the *oxo data set* differs from the *spectrochemical* test set by showing a larger gap between the full model performance and the geometry-only performance, suggesting that electronic descriptors are essential for model

generality (Figure 7).

Figure 7. AUC (top) and accuracy (bottom) for models trained at three N_{step} values: 2, 15, and 40 on the *spectrochemical data set* and then tested on the *spectrochemical* test set (left, solid bars) and *oxo data set* (right, translucent bars). Three types of models are shown: those trained on all descriptors (all, black bars), on only the geometric descriptors (geom, red bars), and only on the electronic descriptors (elec, blue bars).

To gain further insight into what guides dynamic model decisions¹¹⁶⁻¹¹⁷, we employed the gradient-weighted class activation map (Grad-CAM)¹¹⁷, which was developed to interpret CNN image recognition decisions. Grad-CAM uses the gradients of the prediction with respect to the final convolutional layer to highlight the regions of a feature set that contribute most to model prediction (Supporting Information Figure S18). Over the $N_{\text{step}} = 40$ model data set, Grad-CAM reveals the most important portion of the time series to model prediction (Figure 8 and Supporting Information Table S12). Correct predictions of “bad” optimization outcomes can arise from both early changes in the trajectory (step < 15) or from longer time dynamics (step > 20); in either case, Grad-CAM highlights the most essential features to predict the correct “bad” optimization outcome (Figure 8). Similarly, correct predictions of “good” optimization outcomes can arise due to smooth trajectories with no significant variation or due to short time fluctuations that do not interfere with the final outcome, in which case Grad-CAM de-emphasizes the earlier

portion of the time series data (Figure 8). These observations also suggest it would be challenging to obtain extremely high prediction accuracies from small N_{step} models. For example, a quintet $\text{Cr(II)(O}_2^{2-})_4(\text{CN}^-)_2$ complex reaches a “good” geometry optimization outcome at long times, which is correctly predicted by the $N_{\text{step}} = 30$ and 40 models (Supporting Information Figure S19). However, large fluctuations in descriptors (e.g., the metal bond valence) lead to erroneous “bad” predictions by models (e.g., $N_{\text{step}} = 2-10$) trained on less simulation data (Supporting Information Figure S19).

Figure 8. Four correctly predicted geometry optimization outcomes with scaled geometric (geom.) and electronic features (range from -1.0 in blue to 1.0 in red, as shown in inset colorbar) and the corresponding Grad-CAM (GCAM in inset, normalized from 0.0 in white to 1.0 in green and colored with inset color bar): a) “good” geometry that optimizes smoothly, b) a “good” geometry with fluctuations of both descriptor types in the first 25 steps, c) a “bad” geometry with large fluctuations of both descriptor types in the first 15 steps, and d) a “bad” geometry with

worsening geometry metrics starting after 20 steps. The normalized Grad-CAM shows the degree of focus of the model when assigning a label, with high focus corresponding to values near 1.0 (green). Predictions and Grad-CAM analysis are generated from the *spectrochemical*-trained dynamic classifier on $N_{\text{step}}=40$, and the test molecules are from out-of-sample test data. Feature normalization is discussed in the Supporting Information, and in the plots, feature ranges are further adjusted so that each feature reaches a value of 1.0 along the trajectory to add clarity to the feature variation.

3d. On-the-fly Simulation Management

Of the two model types we have developed, the static model will be preferable if it can make a confident prediction since it requires no simulation time prior to prediction, whereas the dynamic model should provide better prediction accuracy when less is known *a priori* about a transition metal complex. To identify if challenging cases for the static model remained challenging for the dynamic model, we separated both the *spectrochemical* test set and the *oxo data set* by static model prediction accuracy into the bottom 20% (i.e., least accurate) and the remaining top 80% (Supporting Information Table S13). For the set aside *spectrochemical* test data, the cases that were difficult to predict with the static classifier remain challenging to predict with the dynamic classifier (Figure 9 and Supporting Information Figures S20-S21). For example, the static classifier miscategorizes a homoleptic, low-spin Co(II) pyridine complex because other spin states of equivalent Co(py)₆ complexes converge to “good” geometries, and the dynamic classifier also fails because the ligand detachment occurs at long optimization times (Supporting Information Figure S22). Comparing the AUC of the two models reveals that the dynamic model generally outperforms the static model, especially for intermediate (i.e., $N_{\text{step}} > 10$) dynamic models (Figure 9). Unlike the *spectrochemical* test set, comparable dynamic model performance is observed for the bottom 20% and top 80% subsets of the *oxo data set* (Figure 9). In this out-of-distribution application that reflects chemical discovery use cases, the dynamic model AUC exceeds the static value for $N_{\text{step}} \geq 2$ when evaluated on the full data set or the

bottom 20% cases but only outperforms the static model on the top 80% at $N_{\text{step}} > 20$ (Figure 9 and Supporting Information Table S13).

Figure 9. Performance for models trained at N_{step} values ranging from 2 to 40 (circles) on the *spectrochemical* data set: AUC is reported over the full *spectrochemical* test set (spectro, left) and *oxo* data set (oxo, right). The data points for the full set (all, grey circles) are also divided into the bottom 20% (red circles) and remaining, top 80% (green circles) based on prediction accuracy of the static classifier. In all cases, a third order regression of the discrete model AUC with N_{step} is also shown. The static classifier AUCs are also shown as horizontal dashed lines for all data (grey) and the top 80% partition (green). The static classifier AUCs of the bottom 20% for spectro (0.17) and oxo (0.00) are not shown for clarity due to their location significantly outside of the range of the AUC of the dynamic classifier models.

Although the static and dynamic classifier models could in principle be used in combination, we focus on a specific application of the dynamic classifier in challenging out-of-distribution cases likely common in chemical space exploration where the static classifier is expected to perform poorly. We now apply the LSE we introduced for the static model in Sec. 3b as a prediction confidence measure for the dynamic model to enable predictive, out-of-distribution autonomous simulation control. In such a workflow, the geometric and electronic structure properties are fed to the dynamic classifier at each geometry update, and a decision about whether to stop the calculation due to a “bad” predicted optimization outcome would only be carried out if the data was inside the LSE cutoff for the dynamic model (Figure 10).

Otherwise, lack of confidence or a confident “good” label would allow the calculation to proceed (Figure 10). Such cutoffs can be chosen to be more discovery-focused (i.e., higher LSE cutoff) or more frugal with respect to compute time (i.e., low LSE cutoff).

Figure 10. Schematic of proposed on-the-fly dynamic simulation control: structures are initialized and updated (XYZ update) and checked for completion. If simulations are not completed, they are passed to the dynamic classifier, which makes a prediction only if the LSE is below a user-specified cutoff, C , and otherwise returns the calculation for more structure updates by the electronic structure code. If the LSE is below the user-specified cutoff, the dynamic classifier makes a prediction, returning “good” structures for additional updates by the code but discarding confidently predicted “bad” structures.

We tested our proposed protocol at a range of LSE thresholds (here, 0.15-0.30) for both the *oxo data set* and *spectrochemical* test set (Figure 11 and Supporting Information Figure S23). At a fixed LSE cutoff and short times (i.e., $N_{\text{step}} < 10$), the dynamic classifier is only confident on a small portion of the *oxo data set*. Although the model only makes predictions on 25% of the data at $N_{\text{step}} = 10$ (LSE cutoff = 0.3), a much higher accuracy of 95% is achieved than the 81% observed for all data with the same model (Figure 11). For tighter LSE cutoffs, even higher accuracy is achieved on the *oxo data set*, albeit on a smaller data fraction (Figure 11). As N_{step} increases, the model becomes increasingly confident in its predictions, resulting in predictions on the majority of the data set for all three LSE cutoffs being compared (Figure 11). At the same time, the accuracy remains consistently high by only making predictions when the model has accumulated enough time series descriptors to be confident. The same observations hold on the set aside *spectrochemical* test set, with greater confidence and therefore a higher fraction of the

data predicted at low N_{step} values (Supporting Information Figure S23). These observations point to the possibility of predicting simulation outcomes with $> 95\%$ accuracy on out-of-distribution complexes through iteratively improving dynamic model confidence.

Figure 11. Performance on the *oxo data set* of a dynamic classifier model trained on the *spectrochemical data set*. Accuracy (top) on the inside cutoff data fraction (bottom) is shown at three LSE cutoffs, 0.15 (green squares), 0.20 (blue diamonds), and 0.30 (gray triangles), and compared to the full data set (no cutoff, red circles).

4. Conclusions

We have trained ML models to predict simulation outcomes as a first step toward realizing fully autonomous computational discovery in transition metal chemistry. As representative indications of simulation health, we labeled simulations as having “good” or “bad” optimized structures and $\langle S^2 \rangle$ deviations. We showed that both SVM and ANN geometry and spin classifier models could be trained to predict these simulation outcomes. Because our static classifier models use heuristic, molecular graph-based representations, they do not require simulation to be carried out to make a prediction. With these models, we achieved accuracies near 90% and AUCs of at least 0.95. In a combinatorial spin and geometry classifier, this

translated in a representative chemical space exploration to simulating less than 50% of the theoretical space to characterize nearly 90% of all leads.

We quantified uncertainty of these model predictions to ensure that the model would be applicable in chemical discovery applications where new lead molecules might differ strongly from training data. We introduced an uncertainty metric especially useful in out-of-distribution predictions with classifier models, the latent space entropy (LSE), which quantifies the distance and distribution of a test point to available training data in latent space. We showed that by tightening the LSE cutoff the model accuracy increased smoothly, even for low baseline accuracy cases. Geometry and spin classifier ANN models broadly trained on all data presented in this work as well as the as well as the LSE metric to estimate prediction uncertainty are available from the molSimplify code repository¹¹⁸.

To overcome potential limitations in a simulation prediction model based only on composition, we introduced a complementary dynamic geometry classifier. In this dynamic model, the time series of electronic and geometric structure data generated during the calculation were used as training data for a convolutional neural network. At short times, electronic structure contributions were indispensable, but, at longer times, geometric information enabled prediction to higher accuracy than had been achieved by the static model. In both out-of-sample and out-of-distribution test cases, high performance (AUC: 0.9) could be obtained based on data generated in the first six hours of a geometry optimization.

Finally, we proposed an approach for job control that merged these models with the LSE uncertainty metric. If the static model could not confidently make a prediction, geometry optimizations were passed to the dynamic model, which made a prediction about whether to terminate the simulation only when the dynamic model had sufficient confidence. As the

dynamic model became increasingly confident with more time series data, this enabled very high accuracy (>95%) over all predicted data both in- and out-of-distribution. Thus our machine learning models provide a path to fully autonomous control of typically time-intensive inorganic chemistry simulations. Next steps will be to build new decision logic that encodes other methodological decisions that experts typically make during the course of a computational study.

ASSOCIATED CONTENT

Supporting Information.

Default and loose cutoffs for the geometry check of octahedral complexes; statistics on the default/loose cutoffs after and during the DFT geometry optimizations; electronic structure features used in the dynamic classification; hyperparameters for the static SVM, ANN and dynamic classifiers; example architectures of dynamic classifier; list of ligands for the *spectrochemical* series data set; descriptions for the *diverse* and *oxo* data set; comparison among three uncertainty metrics on various training and test sets; examples of misclassified trajectories by the 40-step dynamic classifier; feature space distribution of *spectrochemical* series and *oxo* data set; performance of the dynamic classifier on $\langle S^2 \rangle$; AUC of the dynamic classifier with the full, electronic-only and geometry metrics only feature sets; example trajectories of typical complexes and their performance with various classifiers; demonstration of Grad-CAM on a hummingbird picture; statistics and accuracy of top 80% and bottom 20% subsets of *spectrochemical* series and *oxo* data set; performance of the job control system on the *spectrochemical* series data set; list of complexes used in Grad-CAM visualization; normalization procedure in the dynamic classifier. (PDF)

Complete list of all spectrochemical, oxo, and diverse data set compounds, featurizations, spin or geometry labels, and model predictions. (ZIP)

The Supporting Information is available free of charge on <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*email: hjkulik@mit.edu phone: 617-253-4584

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors acknowledge primary support by DARPA grant D18AP00039 for the generation of the classifier models. The automated electronic structure calculations and representation developments were supported by the Office of Naval Research under grant numbers N00014-17-1-2956 and N00014-18-1-2434. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. Training data generated in this work used the XStream computational resource, supported by the National Science Foundation Major Research Instrumentation program (ACI-1429830), as well as other resources from computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

References

1. Shu, Y.; Levine, B. G., Simulated Evolution of Fluorophores for Light Emitting Diodes.

- J. Chem. Phys.* **2015**, *142*, 104104.
2. Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T., et al., Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120-+.
 3. Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R., Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613-1623.
 4. Vogiatzis, K. D.; Polynski, M. V.; Kirkland, J. K.; Townsend, J.; Hashemi, A.; Liu, C.; Pidko, E. A., Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities. *Chem. Rev.* **2018**.
 5. 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.
 6. Curtarolo, S.; Hart, G. L.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O., The High-Throughput Highway to Computational Materials Design. *Nat. Mater.* **2013**, *12*, 191-201.
 7. Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O., AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218-226.
 8. 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.
 9. Nørskov, J. K.; Bligaard, T., The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013**, *52*, 776-777.
 10. Han, S.; Huang, Y.; Watanabe, T.; Dai, Y.; Walton, K. S.; Nair, S.; Sholl, D. S.; Meredith, J. C., High-Throughput Screening of Metal–Organic Frameworks for CO₂ Separation. *ACS Comb. Sci.* **2012**, *14*, 263-267.
 11. Wilmer, C. E.; Leaf, M.; Lee, C. Y.; Farha, O. K.; Hauser, B. G.; Hupp, J. T.; Snurr, R. Q., Large-Scale Screening of Hypothetical Metal–Organic Frameworks. *Nat. Chem.* **2012**, *4*, 83-89.
 12. Witman, M.; Ling, S.; Anderson, S.; Tong, L.; Stylianou, K. C.; Slater, B.; Smit, B.; Haranczyk, M., In Silico Design and Screening of Hypothetical Mof-74 Analogs and Their Experimental Synthesis. *Chem. Sci.* **2016**, *7*, 6263-6272.
 13. 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.
 14. Gossett, E.; Toher, C.; Oses, C.; Isayev, O.; Legrain, F.; Rose, F.; Zurek, E.; Carrete, J.; Mingo, N.; Tropsha, A., AFLOW-MI: A Restful Api for Machine-Learning Predictions of Materials Properties. *Comput. Mater. Sci.* **2018**, *152*, 134-145.
 15. 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.
 16. 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.
 17. 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.
 18. De, S.; Bartok, A. P.; Csanyi, G.; Ceriotti, M., Comparing Molecules and Solids across

- Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754-13769.
19. 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.
 20. Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R., Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3*, 2810.
 21. 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.
 22. Yao, K.; Herr, J. E.; Toth, D. W.; Mckintyre, R.; Parkhill, J., The Tensormol-0.1 Model Chemistry: A Neural Network Augmented with Long-Range Physics. *Chem. Sci.* **2018**, *9*, 2261-2269.
 23. Behler, J., Perspective: Machine Learning Potentials for Atomistic Simulations. *J. Chem. Phys.* **2016**, *145*, 170901.
 24. 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.
 25. 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.
 26. 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.
 27. 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.
 28. 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.
 29. Grajciar, L.; Heard, C. J.; Bondarenko, A. A.; Polynski, M. V.; Meeprasert, J.; Pidko, E. A.; Nachtigall, P., Towards Operando Computational Modeling in Heterogeneous Catalysis. *Chem. Soc. Rev.* **2018**, *47*, 8307-8348.
 30. Arockiam, P. B.; Bruneau, C.; Dixneuf, P. H., Ruthenium(II)-Catalyzed C-H Bond Activation and Functionalization. *Chem. Rev.* **2012**, *112*, 5879-5918.
 31. Prier, C. K.; Rankic, D. A.; MacMillan, D. W. C., Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. *Chem. Rev.* **2013**, *113*, 5322-5363.
 32. Rouquet, G.; Chatani, N., Catalytic Functionalization of C(Sp(2))-H and C(Sp(3))-H Bonds by Using Bidentate Directing Groups. *Angew. Chem., Int. Ed.* **2013**, *52*, 11726-11743.
 33. Schultz, D. M.; Yoon, T. P., Solar Synthesis: Prospects in Visible Light Photocatalysis. *Science* **2014**, *343*, 985-+.
 34. Shaffer, D. W.; Bhowmick, I.; Rheingold, A. L.; Tsay, C.; Livesay, B. N.; Shores, M. P.; Yang, J. Y., Spin-State Diversity in a Series of Co(I) Pnp Pincer Bromide Complexes. *Dalton Trans.* **2016**, *45*, 17910-17917.
 35. Tsay, C.; Yang, J. Y., Electrocatalytic Hydrogen Evolution under Acidic Aqueous Conditions and Mechanistic Studies of a Highly Stable Molecular Catalyst. *J. Am. Chem. Soc.* **2016**, *138*, 14174-14177.
 36. Lan, Z.; Sharada, S. M., Computational Strategies to Probe CH Activation in Dioxo-Dicopper Complexes. *Phys. Chem. Chem. Phys.* **2018**, *20*, 25602-25614.

37. 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.
38. 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.
39. Yella, A.; Lee, H. W.; Tsao, H. N.; Yi, C. Y.; Chandiran, A. K.; Nazeeruddin, M. K.; Diau, E. W. G.; Yeh, C. Y.; Zakeeruddin, S. M.; Gratzel, M., Porphyrin-Sensitized Solar Cells with Cobalt (II/III)-Based Redox Electrolyte Exceed 12 Percent Efficiency. *Science* **2011**, *334*, 629-634.
40. Czerwieniec, R.; Yu, J. B.; Yersin, H., Blue-Light Emission of Cu(I) Complexes and Singlet Harvesting. *Inorg. Chem.* **2011**, *50*, 8293-8301.
41. Dias, F. B.; Bourdakos, K. N.; Jankus, V.; Moss, K. C.; Kamtekar, K. T.; Bhalla, V.; Santos, J.; Bryce, M. R.; Monkman, A. P., Triplet Harvesting with 100% Efficiency by Way of Thermally Activated Delayed Fluorescence in Charge Transfer Oled Emitters. *Adv. Mater.* **2013**, *25*, 3707-3714.
42. Kuttipillai, P. S.; Zhao, Y. M.; Traverse, C. J.; Staples, R. J.; Levine, B. G.; Lunt, R. R., Phosphorescent Nanocluster Light-Emitting Diodes. *Adv. Mater.* **2016**, *28*, 320-326.
43. Leidl, M. J.; Kuchle, F. R.; Mayer, H. A.; Wesemann, L.; Yersin, H., Brightly Blue and Green Emitting Cu(I) Dimers for Singlet Harvesting in Oleds. *J. Phys. Chem. A* **2013**, *117*, 11823-11836.
44. Linfoot, C. L.; Leidl, M. J.; Richardson, P.; Rausch, A. F.; Chepelin, O.; White, F. J.; Yersin, H.; Robertson, N., Thermally Activated Delayed Fluorescence (TadF) and Enhancing Photoluminescence Quantum Yields of Cu-I(Diimine)(Diphosphine)(+) Complexes-Photophysical, Structural, and Computational Studies. *Inorg. Chem.* **2014**, *53*, 10854-10861.
45. Zink, D. M.; Bachle, M.; Baumann, T.; Nieger, M.; Kuhn, M.; Wang, C.; Kloppe, W.; Monkowius, U.; Hofbeck, T.; Yersin, H., et al., Synthesis, Structure, and Characterization of Dinuclear Copper(I) Halide Complexes with Pan Ligands Featuring Exciting Photoluminescence Properties. *Inorg. Chem.* **2013**, *52*, 2292-2305.
46. Cramer, C. J.; Truhlar, D. G., Density Functional Theory for Transition Metals and Transition Metal Chemistry. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757-10816.
47. Harvey, J. N.; Poli, R.; Smith, K. M., Understanding the Reactivity of Transition Metal Complexes Involving Multiple Spin States. *Coord. Chem. Rev.* **2003**, *238*, 347-361.
48. Minenkov, Y.; Sharapa, D. I.; Cavallo, L., Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. *J. Chem. Theory Comput.* **2018**, *14*, 3428-3439.
49. 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.
50. Rappé, A. K.; Casewit, C. J.; Colwell, K.; Goddard III, W. A.; Skiff, W., Uff, a Full Periodic Table Force Field for Molecular Mechanics and Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024-10035.
51. McAnanama-Brereton, S.; Waller, M. P., Rational Density Functional Selection Using Game Theory. *J. Chem. Inf. Model.* **2017**, *58*, 61-67.
52. Stein, C. J.; Reiher, M., Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12*, 1760-1771.
53. Vaucher, A. C.; Reiher, M., Molecular Propensity as a Driver for Explorative Reactivity

Studies. *J. Chem. Inf. Model.* **2016**, *56*, 1470-1478.

54. Bergeler, M.; Simm, G. N.; Proppe, J.; Reiher, M., Heuristics-Guided Exploration of Reaction Mechanisms. *J. Chem. Theory Comput.* **2015**, *11*, 5712-5722.
55. 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.
56. 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.
57. Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H., Machine-Learning-Augmented Chemisorption Model for CO₂ Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528-3533.
58. 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.
59. Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E., Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chemistry of Materials* **2017**, *29*, 9436-9444.
60. 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.
61. Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chem. Sci.* **2017**, *8*, 5137-5152.
62. Denzel, A.; Kästner, J., Gaussian Process Regression for Geometry Optimization. *J. Chem. Phys.* **2018**, *148*, 094114.
63. Koistinen, O.-P.; Dagbjartsdóttir, F. B.; Ásgeirsson, V.; Vehtari, A.; Jónsson, H., Nudged Elastic Band Calculations Accelerated with Gaussian Process Regression. *J. Chem. Phys.* **2017**, *147*, 152720.
64. Carr, S.; Garnett, R.; Lo, C. In *Basc: Applying Bayesian Optimization to the Search for Global Minima on Potential Energy Surfaces*, ICML, 2016; pp 898-907.
65. Wu, J.; Shen, L.; Yang, W., Internal Force Corrections with Machine Learning for Quantum Mechanics/Molecular Mechanics Simulations. *J. Chem. Phys.* **2017**, *147*, 161732.
66. 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*, 2087-96.
67. 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.
68. Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J., Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73.
69. Peterson, A. A.; Christensen, R.; Khorshidi, A., Addressing Uncertainty in Atomistic Machine Learning. *Phys. Chem. Chem. Phys.* **2017**, *19*, 10978-10985.
70. Liu, R.; Wang, H.; Glover, K. P.; Feasel, M. G.; Wallqvist, A., Dissecting Machine-Learning Prediction of Molecular Activity: Is an Applicability Domain Needed for Quantitative Structure-Activity Relationship Models Based on Deep Neural Networks? *J. Chem. Inf. Model.* **2018**.
71. Gal, Y.; Ghahramani, Z. In *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning*, international conference on machine learning, 2016; pp 1050-1059.

72. Balachandran, P. V.; Xue, D. Z.; Theiler, J.; Hogden, J.; Lookman, T., Adaptive Strategies for Materials Design Using Uncertainties. *Sci. Rep.* **2016**, *6*.
73. Luque Ruiz, I.; Gómez-Nieto, M. Á., Study of Data Set Modelability: Modelability, Rivality, and Weighted Modelability Indexes. *J. Chem. Inf. Model.* **2018**, *58*, 1798-1814.
74. Glendening, E. D.; Weinhold, F., Natural Resonance Theory: II. Natural Bond Order and Valency. *J. Comput. Chem.* **1998**, *19*, 610-627.
75. Petachem, L. Petachem. <http://www.petachem.com> (accessed July 1, 2018).
76. Ufimtsev, I. S.; Martinez, 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.
77. Scikit Learn. <https://scikit-learn.org> (accessed Jan 17, 2019).
78. Bergstra, J. C., D. D.; Yamins, D., Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms *Proceedings of the 12th Python in science conference* **2013**, 13-20.
79. Chang, C. C.; Lin, C. J., Libsvm: A Library for Support Vector Machines. *Acm Transactions on Intelligent Systems and Technology* **2011**, *2*.
80. Hornik, K.; Stinchcombe, M.; White, H., Multilayer Feedforward Networks Are Universal Approximators. *Neural Netw.* **1989**, *2*, 359-366.
81. Keras. <https://keras.io/> (accessed Jan 17, 2019).
82. Tensorflow. <https://www.tensorflow.org> (accessed Jan 17, 2019).
83. Rawat, W.; Wang, Z. H., Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **2017**, *29*, 2352-2449.
84. LeCun, Y.; Bengio, Y.; Hinton, G., Deep Learning. *Nature* **2015**, *521*, 436-444.
85. Simonyan, K.; Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* **2014**.
86. Zheng, X.; Zheng, P.; Zhang, R.-Z., Machine Learning Material Properties from the Periodic Table Using Convolutional Neural Networks. *Chem. Sci.* **2018**, *9*, 8426.
87. Chen, X.; Jorgensen, M. S.; Li, J.; Hammer, B., Atomic Energies from a Convolutional Neural Network. *J. Chem. Theory Comput.* **2018**, *14*, 3933-3942.
88. 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.
89. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E., Neural Message Passing for Quantum Chemistry. *arXiv preprint arXiv:1704.01212* **2017**.
90. Schutt, K.; Kindermans, P. J.; Sauceda, H.; Chmiela, S.; Tkatchenko, A.; Muller, K. R., Schnet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions. *Advances in Neural Processing Systems* **2018**, 992-1002.
91. Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Gómez- Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P., Convolutional Networks on Graphs for Learning Molecular Fingerprints. *Advances in Neural Information Processing Systems* **2015**, 2215-2223.
92. Goh, G. B.; Hodas, N. O.; Vishnu, A., Deep Learning for Computational Chemistry. *J. Comput. Chem.* **2017**, *38*, 1291-1307.
93. Wu, Z. Q.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V., Moleculenet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513-530.
94. Ioffe, S. S., Christain, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167, 2015* **2015**.

95. Tsuchida, R., Absorption Spectra of Co-ordination Compounds. I. *Bull. Chem. Soc. Jpn.* **1938**, *13*, 388-400.
96. 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.
97. Becke, A. D., Density-Functional Thermochemistry. III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648-5652.
98. 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. Phys. Chem.* **1994**, *98*, 11623-11627.
99. 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.
100. 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.
101. Saunders, V. R.; Hillier, I. H., Level-Shifting Method for Converging Closed Shell Hartree-Fock Wave-Functions. *Int. J. Quantum Chem.* **1973**, *7*, 699-705.
102. 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.
103. O'Boyle, N. M.; Morley, C.; Hutchison, G. R., Pybel: A Python Wrapper for the Openbabel Cheminformatics Toolkit. *Chem. Cent. J.* **2008**, *2*, 5.
104. Wang, L.-P.; Song, C., Geometry Optimization Made Simple with Translation and Rotation Coordinates. *J. Chem. Phys.* **2016**, *144*, 214108.
105. Geometry Optimization Trajectory Files.
<http://hjkgrp.mit.edu/sites/default/files/classifier/trajectories.zip> (accessed January 12, 2019).
106. Huang, J.; Ling, C. X., Using Auc and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering* **2005**, *17*, 299-310.
107. Fawcett, T., An Introduction to Roc Analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861-874.
108. Tebaldi, C.; Knutti, R., The Use of the Multi-Model Ensemble in Probabilistic Climate Projections. *Philos. Trans. R. Soc., A* **2007**, *365*, 2053-2075.
109. 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**.
110. Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T., Accelerated Search for Materials with Targeted Properties by Adaptive Design. *Nat. Commun.* **2016**, *7*, 11241.
111. 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.
112. Liu, R.; Wallqvist, A., Molecular Similarity-Based Domain Applicability Metric Efficiently Identifies out-of-Domain Compounds. *J Chem Inf Model* **2018**.
113. Cubuk, E. D.; Malone, B. D.; Onat, B.; Waterland, A.; Kaxiras, E., Representations in Neural Network Based Empirical Potentials. *J. Chem. Phys.* **2017**, *147*, 024104.
114. van der Maaten, L.; Hinton, G., Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* **2008**, *9*, 2579-2605.
115. Scipy. <https://www.scipy.org> (accessed Jan 17, 2019).

116. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A., Learning Deep Features for Discriminative Localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2016**, 2921-2929.
117. Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D., Grad-Cam: Visual Explanations from Deep Networks via Gradient-Based Localization. *arXiv preprint arXiv:1610.02391* **2016**.
118. molSimplify Github Page. <https://github.com/hjkgrp/molSimplify> (accessed January 21, 2019).

for Table of Contents only

DuanClassifier.pdf (2.94 MiB)

[view on ChemRxiv](#) • [download file](#)

Supporting Information for

Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models

Chenru Duan^{1,2}, Jon Paul Janet¹, Fang Liu¹, Aditya Nandy^{1,2}, and Heather J. Kulik^{1,*}

¹Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

²Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139

Contents

Figure S1 Geometry failure correlation with total net ligand charge	Page S3
Table S1 Electronic features for the dynamic classifier model	Page S3
Table S2 Default and loose geometric metric checks for octahedral complexes	Page S4
Table S3 Data normalization quantities for the dynamic classifier	Page S4
Figure S2 Computational cost characteristics for geometry optimizations	Page S5
Table S4 Hyperparameters for the static ANN and SVM classifiers	Page S6
Figure S3 The architecture of the dynamical classifier	Page S7
Table S5 Hyperparameters for the geometry classification dynamic classifier	Page S8
Table S6 Hyperparameters for the $\langle S^2 \rangle$ deviation dynamic classifiers	Page S8
Table S7 List of ligands in the spectrochemical series data set	Page S9
Text S1 Summary of diverse and oxo data sets	Page S10
Table S8 A summary of job status and labels for the spectrochemical series data	Page S11
Table S9 Statistics on strict/loose cutoffs during optimizations	Page S11
Table S10 Static classifier performance on oxo, diverse, and strongly symmetric sets	Page S12
Figure S4 Fraction of “bad” spin labeled compounds by ligand charge or identity	Page S12
Table S11 Source of combinatorial performance errors	Page S12
Figure S5 Prediction error with probability and ensemble uncertainty metrics	Page S13
Figure S6 Uncertainty metric-based performance on set aside and diverse data sets	Page S13
Figure S7 Prediction error with RAC feature space metrics	Page S14
Figure S8 Analysis of FPR and FNR with tightened LSE cutoffs	Page S14
Figure S9 Properties with LSE or probability for spectrochemical train and diverse test	Page S15
Figure S10 Properties with LSE for spectrochemical train and test	Page S16
Figure S11 Properties with LSE, probability, ensemble for diverse train spectro test	Page S17
Figure S12 Examples of misclassified optimization trajectories for dynamic classifier	Page S17
Figure S13 RAC space distance of spectro test and oxo data to spectro training data	Page S18
Figure S14 Dynamic electronic descriptor distributions for spectro and oxo data	Page S18
Figure S15 Performance of the dynamic spin classifier model	Page S19
Figure S16 AUC performance for total, electronic, and geometric dynamic models	Page S19
Figure S17 Example of quick failure prediction from electronic structure	Page S20
Figure S18 An example of Grad-CAM in image recognition	Page S20
Table S12 Properties of complexes used in Grad-CAM visualization	Page S20
Figure S19 Descriptor trajectory of case predicted correctly only at large N_{step}	Page S21
Table S13 Details about 20/80 partition	Page S21

Figure S20	Accuracy for easy/challenging spectrochemical cases	Page S21
Figure S21	Geometry deviations for easy and challenging spectrochemical cases	Page S22
Figure S22	Easy and challenging optimization paths for static/dynamic classifiers	Page S22
Figure S23	Performance of dynamic job control approach on spectrochemical test data	Page S23
References		Page S23

Figure S1. The fraction of failed geometry optimizations (i.e., failure rate) for each total summed net ligand charge over the transition metal complex in the spectrochemical data set.

Table S1. Complete list of 26 electronic structure features computed for the dynamic classifier model. The label axcon means axial connecting atoms, eqcon equatorial connecting atoms. Mulliken charge, C , is obtained over the metal or averaged values. Bond order properties include the bond valence, BV, bond order, BO, averaged as indicated, and singular values (SV) of the bond order matrix (BO) or one with the diagonal (diag) set to 0. Gradient matrix properties include the gradient norm (G), on the metal, averaged over equatorial and axial connecting ligands, maximum norm for the full matrix, root mean square of the matrix, and 3 singular values of the singular value decomposition. The superscript “mc” refers to the metal-centered internal coordinate representation of the gradient. Its max value and three singular decomposition values are used as variables.

From Mulliken charge					
C_{metal}	$C_{\text{eqcon}}^{\text{avrg}}$	$C_{\text{axcon}}^{\text{avrg}}$			
From bond order matrix					
BV_{metal}	$BO_{\text{eqcon}}^{\text{avrg}}$	$BO_{\text{axcon}}^{\text{avrg}}$	SV_1^{BO}	SV_2^{BO}	SV_3^{BO}
SV_4^{BO}	$SV_1^{BO, \text{diag}=0}$	$SV_2^{BO, \text{diag}=0}$	$SV_3^{BO, \text{diag}=0}$	$SV_4^{BO, \text{diag}=0}$	
From gradient matrix					
G_{metal}	$G_{\text{eqcon}}^{\text{avrg}}$	$G_{\text{axcon}}^{\text{avrg}}$	G_{max}	G_{RMS}	SV_1^{Grad}
SV_2^{Grad}	SV_3^{Grad}	$G_{\text{max}}^{\text{Grad, mc}}$	$SV_1^{\text{Grad, mc}}$	$SV_2^{\text{Grad, mc}}$	$SV_3^{\text{Grad, mc}}$

Table S2. The eight standard and loose (in parentheses) geometry check metric cutoffs used for an octahedral transition metal complex. These include coordination number (CN), angular deviations between the *i*th and *j*th connecting atoms, C, and the metal, and maximum difference between equatorial and axial metal-ligand bonds (i.e., Δd). The angular ligand distortion quantities are only computed for ligands that are expected to be linear. Standard geometry check metrics are the same as in Ref. 1.

<u>Coordination number</u>			
CN			
6 (6)			
<u>Shape of the first coordination sphere</u>			
$\text{avg}(\Delta\theta(C_i-M-C_j))$	$\text{max}(\Delta\theta(C_i-M-C_j))$	$\text{max}(\Delta d)$	$\text{max}(\Delta d_{eq})$
12° (16°)	22.5° (27.0°)	1.00 Å (1.25 Å)	0.35 Å (0.45 Å)
<u>Ligand distortion</u>			
$\text{max}(\text{RMSD})$	$\text{avg}(\Delta\theta(M-A-B))$	$\text{max}(\Delta\theta(M-A-B))$	
0.30 Å (0.40 Å)	20° (30°)	28° (40°)	

Table S3. Standardization scheme used in the dynamic classifier. In the geometry classification task, each electronic structure feature *x* (Table S1) is normalized by its mean ($\langle x \rangle$) and standard deviation (δx) across all complexes in the training set at $N_{\text{step}} = 0$, $\tilde{x} = (x - \langle x \rangle) / \delta x$. Each geometry metric feature (*y*) is scaled by its corresponding strict cutoff value (y_{cut}), $\tilde{y} = y / y_{\text{cut}}$. In the $\langle S^2 \rangle$ deviation task, the feature of $\langle S^2 \rangle$ deviation is normalized by its mean and standard deviation across all complexes at $N_{\text{step}} = 0$, while the binary feature of whether $\langle S^2 \rangle$ deviation > 1 ($\langle S^2 \rangle$ label) is scaled by a cutoff of 1. Values used in normalization are shown below.

Feature	Mean	Standard deviation
C_{metal}	-0.557	1.618
$C_{\text{eqcon}}^{\text{avg}}$	-0.290	0.428
$C_{\text{axcon}}^{\text{avg}}$	-0.293	0.357
BV_{metal}	3.912	1.174
$BO_{\text{eqcon}}^{\text{avg}}$	0.586	0.269
$BO_{\text{axcon}}^{\text{avg}}$	0.598	0.273
SV_1^{BO}	6.280	1.058
SV_2^{BO}	5.313	1.379
SV_3^{BO}	4.786	1.632
SV_4^{BO}	4.443	1.801
$SV_1^{\text{BO,diag}=0}$	2.820	0.536
$SV_2^{\text{BO,diag}=0}$	2.657	0.553
$SV_3^{\text{BO,diag}=0}$	2.310	0.714
$SV_4^{\text{BO,diag}=0}$	2.262	0.721
G_{metal}	0.020	0.021
$G_{\text{eqcon}}^{\text{avg}}$	0.054	0.061
$G_{\text{axcon}}^{\text{avg}}$	0.059	0.053
G_{max}	0.213	0.509
G_{RMS}	0.067	0.071

Feature	Mean	Standard deviation
SV_1^{Grad}	0.325	0.723
SV_2^{Grad}	0.138	0.178
SV_3^{Grad}	0.074	0.105
$G_{\text{max}}^{\text{Grad,mc}}$	0.223	0.509
$SV_1^{\text{Grad,mc}}$	0.361	0.718
$SV_2^{\text{Grad,mc}}$	0.141	0.178
$SV_3^{\text{Grad,mc}}$	0.084	0.113
$\Delta\langle S^2 \rangle$	0.618	1.057
	Cutoff	
$\text{avg}(\Delta\theta(C_i\text{-M-}C_j))$	12°	
$\text{max}(\Delta\theta(C_i\text{-M-}C_j))$	22.5°	
$\text{max}(\Delta d)$	1.00 Å	
$\text{max}(\Delta d_{eq})$	0.35 Å	
$\text{max}(\text{RMSD})$	0.30 Å	
$\text{avg}(\Delta\theta(\text{M-A-B}))$	20°	
$\text{max}(\Delta\theta(\text{M-A-B}))$	28°	
Geometry label	1	
$\langle S^2 \rangle$ label	1

Figure S2. Computational cost characteristics vs. step of the geometry optimization process: (left) average cumulative GPU hour wall time and (right) time required for that step of the geometry optimization process. Calculations are averaged over 3,622 complexes averaging 37 atoms in size on a single GeForce GTX 780/GeForce GTX 970 card using the simulation parameters described in the main text.

Table S4. A summary of hyperparameters that are selected for use in the ANN (top) and SVM (bottom) static classifier. For the ANN, the hidden size is a tuple of numbers, which represent the number of neurons at each hidden/intermediate layer. Those hidden layers employ relu activation functions, and the last layer of the ANN employs a sigmoid activation function prior to a linear output layer. The linear output layer is where any latent space analysis is carried out. The learning rate and decay are hyperparameters for the Adam optimizer, which was employed. Batch normalization is used during the ANN training.

	ANN hyperparameters	
Task	Geometry	$\langle S^2 \rangle$ deviation
Hidden size	(200, 200)	(300, 300)
Dropout rate	0.3	0.35
L2 regularization	2.5e-3	5e-3
Learning rate	5e-5	5e-5
Decay	1e-4	1e-4
Batch size	50	50
	SVM hyperparameters	
Task	Geometry	$\langle S^2 \rangle$ deviation
Nu	0.0865	0.3595
Inverse kernel width(γ)	0.0020	0.0226

Figure S3. The model architecture of the dynamic classifier with a data input as 40 steps of optimization as an example. This model plot is generated with Keras².

Table S5. A summary of hyperparameters that are used in the dynamic classifiers for the geometry classification task trained on different numbers of steps of geometry optimization. Note that L2-regularization is used in both the convolutional and the dense layer. Models with different feature sets (full, geometry metrics only, electronic structure information only) use the same set of hyperparameters as no improvement was found by tuning them for the individual feature sets.

Step of optimization	2	5	10	15	20	30	40
Kernel size	2	3	4	4	5	6	6
Stride	1	1	1	1	1	2	2
Filter	35	35	35	35	45	45	45
Dropout rate (convolutional layer)	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Dense neurons	50	50	50	50	50	100	100
Dropout rate (dense layer)	0.1	0.1	0.1	0.1	0.1	0.2	0.2
L2 regularization	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4	1e-4
Learning rate	2e-5	2e-5	2e-5	2e-5	2e-5	1.5e-5	1.5e-5
Decay	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5	1e-5
Batch size	256	256	256	256	256	256	256

Table S6. A summary of hyperparameters that are used in the $\langle S^2 \rangle$ deviation dynamic classifiers trained on different numbers of steps of geometry optimization. Note that L2-regularization is used in both the convolutional and the dense layer.

Step of optimization	2	5	10	15	20	30	40
Kernel size	2	3	4	4	5	6	6
Stride	1	1	1	1	1	2	2
Filter	35	35	35	35	45	45	45
Dropout rate (convolutional layer)	0.3	0.3	0.3	0.3	0.3	0.3	0.3
Dense neurons	50	50	50	50	50	100	100
Dropout rate (dense layer)	0.2	0.2	0.2	0.2	0.2	0.3	0.3
L2 regularization	3e-4	3e-4	3e-4	3e-4	3e-4	3e-4	3e-4
Learning rate	5e-5	5e-5	5e-5	5e-5	5e-5	5e-5	5e-5
Decay	1e-6	1e-6	1e-6	1e-6	1e-6	1e-6	1e-6
Batch size	50	50	50	50	50	50	50

Table S7. The list of ligands used in the spectrochemical series dataset, as indicated by the smiles string and connecting (heavy atom) for forming the complex read linearly from the SMILES string.

Ligand index	SMILES string	Connecting atom	Chemical formula
0	[O-][O-]	1	O ₂ ²⁻
1	[I-]	1	I ⁻
2	N#C[O-]	1	CNO ⁻
3	[OH-]	1	OH ⁻
4	N#[C-]	1	CN ⁻
5	[OH2]	1	OH ₂
6	N#CC	1	C ₂ H ₃ N
7	c1ccncc1	4	C ₅ H ₁₁ N
8	[NH3]	1	NH ₃
9	c3c(P(c1ccccc1)c2ccccc2)cccc3	3	C ₁₈ H ₁₅ P
10	C=O	2	CH ₂ O
11	[C-]#[O+]	1	CO
12	[S--]	1	S ²⁻
13	[F-]	1	F ⁻
14	[PH3]	1	PH ₃
15	[SH2]	1	SH ₂

Text S1. Summary description of diverse and oxo data sets used to test spectrochemical-series-trained static and dynamic classifiers.

Diverse data set. The diverse data set consists of 5,451 octahedral transition metal (5,167 used in the geometry classification) that were collected in prior work for training models to predict spin-splitting energy³, redox potential⁴, and full enumeration of small homoleptic transition metal complexes defined as those with two or fewer heavy atoms per monodentate ligand⁵. Points excluded in the geometry classification task were unconverged after 5 resubmissions but did not pass the loose cutoffs and are not labeled. All octahedral complexes are with Cr, Mn, Fe and Co in M(II) and M(III) oxidation states with identical equatorial ligands but up to two distinct axial ligands. The 33 ligands from complexes studied in Refs. ³⁻⁴ included acetylacetone, aceticacidbipyridine, ammonia, benzylisocyanide, bipyridine, carbonyl, catecholate, chloride, 3-chloropyridine, cyanide, cyanoaceticporphyrin, 4-cyanopyridine, ethylenediamine, ethbipyridine, furan, isothiocyanate, methylbipyridine, 4-methyl-o-catecholate, methylamine, methylisocyanide, oxalate, phen, phenylisocyanide, phosacidbipyridine, t-butyl-phenylisocyanide, porphine, pyrrole, pyridine, sulfacidbipyridine, t-butyl-isocyanide, 4-t-butyl-1,2-benzoquinone, thiocyanate, thiopyridine, and water were used in complexes with up to two distinct axial ligands. From Ref. 5, 274 additional monodentate ligands were used for the construction of 3,529 homoleptic complexes (3,310 used in the geometry classification task). These complexes were constructed with the same metal, oxidation, and spin state combination as in the earlier set. A list of their SMILES strings are provided in the Supporting Information ZIP file. High spin and low spin state definitions are the same as those described for the spectrochemical data set in the main text, and the other settings for simulations are also the same.

Oxo data set. The oxo data set consists of 515 complexes, 506 of which are used in the geometry classification, and 382 are used in $\langle S^2 \rangle$ deviation task. Points excluded in the geometry classification task were unconverged after 5 resubmissions but did not pass the loose cutoffs and are not labeled. These octahedral complexes consist of a Mn or Fe metal center with equatorial symmetry and differing axial ligands because one axial ligand as either an oxo (275 of 515) or hydroxyl (240 of 515). The remaining ligands consist of up to two of the following: acac, ammonia, carbonyl, chloride, cyanide, ethylenediamine, isothiocyanate, oxalate, phen, phenylisocyanide, porphyrin, tbutylisocyanide, thiocyanate, or water. Complexes with oxo have a metal in the oxidation state M(IV) or M(V), and complexes with hydroxyl are in an oxidation state M(III) or M(IV). That is, a formal charge of the metal of +2 or +3 is assigned prior to accounting for the ligand charge (i.e., hydroxyl withdraws one electron, oxo withdraws two). Mn(III)/Fe(IV) and Fe(III) are studied in high (i.e., quintet and sextet, respectively), intermediate (i.e., triplet and quartet), or low spin states (i.e., singlet and doublet). Mn(V) and Mn(IV)/Fe(V) are only studied in a single high (i.e., triplet and quartet, respectively) and low spin (i.e., singlet and doublet) state. This data set is generated with the same settings as the spectrochemical series data set described in the main text except that a semi-empirical dispersion (i.e., D3) correction⁶ is used during the geometry optimization.

A list of all structures for both data sets is provided in the Supporting Information zip file.

Table S8. Job status and the size of data for both the geometry and $\langle S^2 \rangle$ deviation classification at different metrics values and convergence status for the 3,622 complexes in the *spectrochemical data set*. A job status of 1 means a “good” geometry/not spin contaminated, 0 with a “bad” geometry/spin contaminated, and calculations with job status -1 were excluded from the data set as we do not have enough information to label them from slow optimizations.

	Converged			Unconverged		
	job status	geometry data set size	$\langle S^2 \rangle$ deviation data set size	job status	geometry data set size	$\langle S^2 \rangle$ deviation data set size
metrics < strict cutoffs	1 “good”	1744	1649	-1	126	660
strict cutoffs < metrics < loose cutoffs	0 “bad”	26	111	-1	19	323
metrics > loose cutoffs	0 “bad”	690	34	0 “bad”	1017	179

Table S9. A summary of total number of jobs in several categories for the 3,622 complex *spectrochemical data set* by their final status along with the percentage of jobs by their final status and when they exceed strict or loose cutoffs of the geometry check in their optimization trajectories. For each category of either converged/unconverged or good/bad geometry, the values sum up to 100%. Loose cutoffs were chosen so that calculations that converge with good geometries generally don’t pass them and thus at resubmission time, calculations that pass the loose cutoff can be identified as bad.

Category	Number of jobs	Always stay below strict cutoffs (%)	Pass strict cutoffs and never reach loose cutoffs (%)	Pass loose cutoffs (%)
Converged, with good geometry	1744	89	9	2
Converged, with bad geometry	716	0	3	97
Unconverged, with good geometry	126	97	3	0
Unconverged, with bad geometry	1036	0	0	100

Table S10. A summary of the accuracy and AUC of the ANN static classifier on the spectrochemical data set, a strongly symmetric spectrochemical data set subset (i.e., only two unique ligands, as indicated in the table), diverse data set, and oxo data set. The ANN static classifier is trained on the spectrochemical series (spectro) training set in all cases, except the spectro strongly symmetric case refers to both train and test on that subset.

	spectro	spectro ($M(L_1)_4(L_2)_2$)	diverse	oxo
accuracy	0.88	0.89	0.57	0.75
AUC	0.95	0.96	0.59	0.66

Figure S4. The percentage of “bad” (i.e., $\langle S^2 \rangle$ deviation > 1.0) spin results (i.e., “failure rate”) grouped the net ligand charge (left) and the identity of the equatorial ligand (right) of a complex. The equatorial ligand index corresponds to values indicated in Table S6.

Table S11. Source of misclassification in the combined classifier: false negative (i.e., predicted failed calculation that succeeds) or false positive (i.e., predicted success that fails) based on whether the geometry flag is wrong (i.e., 0 for FN, 1 for FP) or the spin deviation flag is wrong (i.e., 0 for FN, 1 for FP). Most incorrect classifications arise from the geometry due to the lower accuracy of those models.

	Geometry	$\langle S^2 \rangle$ deviation	Both
Threshold=0.5			
False negative	32	3	0
False positive	39	9	0
Threshold=0.3			
False negative	14	0	0
False positive	62	12	0

Figure S5. The prediction error in the geometry classifier vs. uncertainty metrics of using the prediction probability directly and ensemble models for (left) train and test on the *spectrochemical data set* or (right) out-of-distribution training on *spectrochemical data set* and testing on the *diverse data set*. Correctly classified geometries are shown as blue circles and incorrectly classified geometries are shown as red squares. For the probability (top), the metric is chosen as $0.5-|\text{probability}-0.5|$ such that the predictions of 0 or 1 have lower uncertainties. For the ensemble models (bottom), the prediction standard deviation is used as the uncertainty metric.

Figure S6. ANN prediction accuracy versus the ratio of data included for three uncertainty metrics: latent space entropy (LSE) (black), prediction probability (dark green), and ensemble variance (light green) for: train on *spectrochemical data set* and test on the set aside test partition of the *spectrochemical data set* (left), train on the *spectrochemical data set* and test on the

diverse data set (middle), train on the *diverse data set* and test on the *spectrochemical data set* (right).

Figure S7. The prediction error in the geometry classifier vs. RAC space distance for (left) train and test on the *spectrochemical data set* or (right) out-of-distribution training on *spectrochemical data set* and testing on the *diverse data set*. Correctly classified geometries are shown as blue circles and incorrectly classified geometries are shown as red squares.

Figure S8. False negative (blue) and false positive (red) rates as a function of latent space entropy cutoff (x-axis) for spectrochemical train and test (left), spectrochemical train and diverse test (middle), and diverse train and spectrochemical test (right).

Figure S9. Prediction accuracy (top) and fraction of data retained (bottom) with respect to latent space entropy (LSE) (left) and probability (right) cutoffs. Data accuracy and fraction inside the cutoff (i.e., would be predicted) is shown in green and outside (i.e., would not be predicted) is shown in orange. Moving from right to left on the x -axis makes the model more conservative. The models were trained on the *spectrochemical data set* and tested on the *diverse data set*.

Figure S10. Properties with respect to the latent space entropy (LSE) for the *spectrochemical series data test set*, where models are trained on the *spectrochemical series data set*. (top) Prediction errors for each sample, both correctly classified (green circles) and misclassified (orange squares). (middle) Prediction accuracy for data that are included to make predictions as they are within the LSE cutoff (blue) and those that are excluded (gray). (bottom) The ratio of data included in predictions below the LSE cutoff.

Figure S11. Prediction accuracy (top) and fraction of data retained (bottom) with respect to latent space entropy (LSE) (left), ensemble variance (middle), and probability (right) cutoffs. Data accuracy and fraction inside the cutoff (i.e., would be predicted) is shown in green and outside (i.e., would not be predicted) is shown in orange. Moving from right to left on the x -axis makes the model more conservative. The models were trained on the *diverse data set* and tested on the *spectrochemical data set*.

Figure S12. Examples of geometry optimizations that are difficult to correctly classify with information only from the first 40 steps of geometry optimization. Both a false negative (left, i.e., “good” geometry predicted to be “bad”) and false positive (right, i.e., “bad” geometry predicted to be “good”) examples are provided. The quantity that appears good or bad is different for the two plots: the y-axis for the false negative example is the maximum difference of metal-ligand bond lengths and the y-axis for the false positive is the maximum angle deviation of the first coordination shell. The dynamic model prediction probabilities of being a good geometry are shown at steps 10, 20, 30 and 40.

Figure S13. Distance distribution in RAC space of spectrochemical series test data (red) and oxo data (blue) with respect to the spectrochemical series training data. Both spectrochemical series test data and oxo data are normalized by the spectrochemical series training data distance before the distance measurement.

Figure S14. Distribution of dynamic descriptors features at step 0 (left) and 40 (right) for the spectrochemical series test data (red) and the oxo data set (blue). Both the metal bond valence (BO_metal, top) and the averaged axial ligand bond order with metal (BO_axlig_avrg, bottom) are shown.

Figure S15. Accuracy and AUC for the $\langle S^2 \rangle$ deviation task for both the spectrochemical series test data (black square) and oxo test data (blue circle). The model is trained on the spectrochemical series training data. The feature set contains the 26 electronic structure descriptors (Table S1), the numerical $\langle S^2 \rangle$ deviation, and a binary number of whether the $\langle S^2 \rangle$ deviation is larger than 1.

Figure S16. AUC versus the step of optimization on both spectrochemical series test data (left) and out-of-distribution oxo data (right). Performance of models trained on the total feature set (black squares) are compared to a model trained on features only containing the geometry descriptors (blue circle) or only the electronic structure descriptors (orange triangle).

Figure S17. An example (sextet $\text{Mn(II)(CN}^-\text{)}_4(\text{NH}_3)(\text{F}^-)$) of a geometry optimization trajectory that is difficult to be detected by geometry metrics within 20 steps but is easier to be identified by the electronic structure information. The y axes are selected as the representative dimension of the complex from either a geometric (left, the maximum difference of metal-ligand bond lengths in a complex) or an electronic structure (right, metal bond valence) point of view.

Figure S18. An example of the application of the Gradient-weighted Class Activation Map (Grad-CAM) on a hummingbird image with the VGG16 model⁷ loaded from keras². Both the original image (left) and an image on which the Grad-CAM is superimposed (right) are shown. The Grad-CAM highlights the tail, the end of the wing and beak (circled), which are essential to the image classification task.

Table S12. Complexes used in the main text with the description of their optimization path as well as the final model prediction.

Case	Complex	Optimization type
(a)	(HS) $\text{Cr(III)(PH}_3\text{)}_6$	“good”, smooth/fast
(b)	(LS) $\text{Cr(II)(CH}_2\text{O)}_4(\text{H}_2\text{O})(\text{F}^-)$	“good”, fluctuating in the beginning
(c)	(LS) $\text{Co(III)(I}^-)_4(\text{OH}^-)(\text{CN}^-)$	“bad”, early detection/fast
(d)	(HS) $\text{Cr(III)(F}^-)_4(\text{OH}^-)(\text{C}_2\text{H}_3\text{N})$	“bad”, slow evolution

Figure S19. Metal bond valence time series data for a quintet $\text{Cr(II)(O}_2^{2-})_4(\text{CN}^-)_2$ complex that converges to a “good” structure with dynamic ANN classifier model overlaid for the $N_{\text{step}} = 2, 5, 10, 20, 30$, and 40 models. Only $N_{\text{step}} = 30$ and 40 definitively predict a correct “good” structure.

Table S13. Accuracy and AUC of the ANN static classifier top 80%, bottom 20%, and overall for both the spectrochemical (spectro) data set as well as the oxo data set. The low AUC on the bottom 20% indicates worse than random guess performance.

	Top 80%	Bottom 20%	Overall
spectro			
accuracy	1.00	0.36	0.88
AUC	1.00	0.17	0.95
oxo			
accuracy	0.94	0.00	0.75
AUC	0.97	0.00	0.66

Figure S20. Accuracy of the ANN static classifier of the overall dataset (black circles) and that separated into the bottom 20% most difficult to predict (blue circles) and the top 80% easiest to predict (orange triangles). The accuracy of the spectrochemical series test data is on the left and the oxo test on the right. The ANN static classifier is trained on the spectrochemical series

training data.

Figure S21. Averaged geometry deviation with number of optimization steps (up to 20) separated into the bottom 20% most difficult to predict (blue circles) and top 80% easiest to predict (orange triangles). These results are reported for training and testing on the spectrochemical data set. The left plot shows structures that should be classified as bad geometries and the right plot shows structures that should be classified as good: the difficult, bad structures have slower growing bad geometry metrics and the difficult, good structures have larger deviations than the easier good structures.

Figure S22. Examples of geometry optimization trajectories that are either easy (left, LS Fe(II)(O₂)₄(NNCCH₃)(OCH₂)) or difficult (right, LS Co(II)(pyr)₆) to be correctly classified by both the static and dynamic classifier. The y-axis is the maximum difference of metal-ligand bond lengths in a complex (max(del_BL)) selected for the most representative visualization of the changes in structure. The dynamic classifier prediction probabilities of being a good geometry are shown at steps 10, 20, 30, and 40.

Figure S23. Performance on the *spectrochemical* test set of a dynamic classifier model trained on the *spectrochemical* data set. Accuracy (left) on the included data fraction (right) is shown at three LSE cutoffs, 0.15 (blue circles), 0.20 (red triangles), and 0.30 (green diamonds), and compared to the full data set (no cutoff, which is the same as LSE cutoff = $\ln(2)$, black squares).

References

1. Nandy, A.; Duan, C.; Janet, J. P.; Gugler, S.; Kulik, H. J., Strategies and Software for Machine Learning Accelerated Discovery in Transition Metal Chemistry. *Industrial & Engineering Chemistry Research* **2018**, *57* (42), 13973-13986.
2. <https://keras.io/>.
3. Janet, J. P.; Kulik, H. J., Predicting Electronic Structure Properties of Transition Metal Complexes with Neural Networks. *Chemical Science* **2017**, *8*, 5137-5152.
4. 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* (46), 8939-8954.
5. Gugler, S.; Janet, J. P.; Kulik, H. J., Enumerating the Small Inorganic Complex Universe for Machine Learning and Discovery. *in preparation* **2019**.
6. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J Chem Phys* **2010**, *132* (15), 154104.
7. Simonyan, K.; Zisserman, A., Very Deep Convolutional Networks for Large-Scale Image Recognition. <https://arxiv.org/abs/1409.1556> **2014**.

SIClassifier_v5.pdf (5.28 MiB)

[view on ChemRxiv](#) • [download file](#)

Other files

data_set.zip (35.83 MiB)

[view on ChemRxiv](#) • [download file](#)