

MIT Open Access Articles

Representations and strategies for transferable machine learning improve model performance in chemical discovery

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

Citation: Harper, Daniel R, Nandy, Aditya, Arunachalam, Naveen, Duan, Chenru, Janet, Jon Paul et al. 2022. "Representations and strategies for transferable machine learning improve model performance in chemical discovery." *The Journal of Chemical Physics*, 156 (7).

As Published: 10.1063/5.0082964

Publisher: AIP Publishing

Persistent URL: <https://hdl.handle.net/1721.1/145470>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution 4.0 International license



Representations and strategies for transferable machine learning improve model performance in chemical discovery

Cite as: J. Chem. Phys. **156**, 074101 (2022); <https://doi.org/10.1063/5.0082964>

Submitted: 20 December 2021 • Accepted: 25 January 2022 • Accepted Manuscript Online: 27 January 2022 • Published Online: 15 February 2022

Daniel R. Harper,  Aditya Nandy,  Naveen Arunachalam, et al.

COLLECTIONS

Paper published as part of the special topic on [Chemical Design by Artificial Intelligence](#)



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[Perspective on integrating machine learning into computational chemistry and materials science](#)

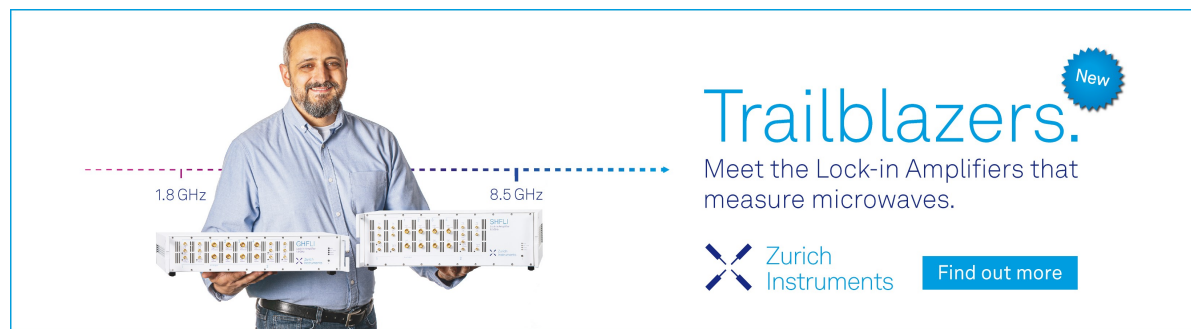
The Journal of Chemical Physics **154**, 230903 (2021); <https://doi.org/10.1063/5.0047760>


[ManyHF: A pragmatic automated method of finding lower-energy Hartree-Fock solutions for potential energy surface development](#)

The Journal of Chemical Physics **156**, 071101 (2022); <https://doi.org/10.1063/5.0080817>


[Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability](#)

The Journal of Chemical Physics **156**, 084104 (2022); <https://doi.org/10.1063/5.0079574>



Trailblazers. 

Meet the Lock-in Amplifiers that measure microwaves.

 Zurich Instruments [Find out more](#)

Representations and strategies for transferable machine learning improve model performance in chemical discovery

Cite as: J. Chem. Phys. 156, 074101 (2022); doi: 10.1063/5.0082964

Submitted: 20 December 2021 • Accepted: 25 January 2022 •

Published Online: 15 February 2022



View Online



Export Citation



CrossMark

Daniel R. Harper,^{1,2} Aditya Nandy,^{1,2}  Naveen Arunachalam,¹  Chenru Duan,^{1,2}  Jon Paul Janet,¹
and Heather J. Kulik^{1,a)} 

AFFILIATIONS

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

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

Note: This paper is part of the JCP Special Topic on Chemical Design by Artificial Intelligence.

a) Author to whom correspondence should be addressed: hjkulik@mit.edu. Tel.: 617-253-4584

ABSTRACT

Strategies for machine-learning (ML)-accelerated discovery that are general across material composition spaces are essential, but demonstrations of ML have been primarily limited to narrow composition variations. By addressing the scarcity of data in promising regions of chemical space for challenging targets such as open-shell transition-metal complexes, general representations and transferable ML models that leverage known relationships in existing data will accelerate discovery. Over a large set (~1000) of isovalent transition-metal complexes, we quantify evident relationships for different properties (i.e., spin-splitting and ligand dissociation) between rows of the Periodic Table (i.e., 3d/4d metals and 2p/3p ligands). We demonstrate an extension to the graph-based revised autocorrelation (RAC) representation (i.e., eRAC) that incorporates the group number alongside the nuclear charge heuristic that otherwise overestimates dissimilarity of isovalent complexes. To address the common challenge of discovery in a new space where data are limited, we introduce a transfer learning approach in which we seed models trained on a large amount of data from one row of the Periodic Table with a small number of data points from the additional row. We demonstrate the synergistic value of the eRACs alongside this transfer learning strategy to consistently improve model performance. Analysis of these models highlights how the approach succeeds by reordering the distances between complexes to be more consistent with the Periodic Table, a property we expect to be broadly useful for other material domains.

© 2022 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0082964>

I. INTRODUCTION

Machine learning (ML) models have been demonstrated as a powerful alternative to conventional computational high-throughput screening^{1,2} by greatly accelerating chemical discovery,^{3–8} e.g., with active learning.^{9–12} It is typically the case at the outset of a discovery campaign that a great deal is known from experiments, computation, and associated ML models about a highly localized region of chemical space, but the most promising directions for improvement lie outside this region. For example, many of the most active catalysts^{13,14} or single molecule magnets^{15,16} require the use of rare and toxic 4d or 5d metals, and identification

of principles¹⁷ for their replacement with 3d metals would be advantageous^{18,19} in a number of difficult small molecule catalytic conversions (e.g., H₂,^{20–25} CO₂,²⁶ and O₂).²⁷ Thus, ML-accelerated discovery requires models and representations that generalize well and strategies to exploit existing knowledge while data in new regions are scarce.

Despite the need for ML models that generalize across the Periodic Table,^{3,28–31} ML models have much more commonly been applied in narrow regions of chemical space, such as closed-shell organic (i.e., CHNOF-containing) molecules in the QM9 dataset³² with few exceptions.³³ For open-shell transition-metal chemistry, in particular, unique challenges are present in the generation of

sufficiently large datasets for ML model training while spanning a large range of metal centers, ligand chemistry, and the cooperative effect the metal/ligand identity play in the resulting spin or oxidation state of the complex.³⁴ While ML-model-accelerated computational discovery has targeted numerous transition-metal catalysts^{35–39} or functional complexes^{9,40} and related metal–organic materials (e.g., metal–organic frameworks^{41–43} and transition metal oxides⁴⁴), these models have also generally been restricted to a small number of transition metals or modest modifications of the ligands. While only preliminarily demonstrated on open-shell transition-metal complexes,^{45–47} there is evidence that electronic descriptors such as partial charges or bond valence⁴⁵ and molecular orbitals^{46,48,49} or other computed electronic properties^{50,51} are one path to training transferable ML models on small datasets. Sidestepping ML with direct, computationally demanding simulation [e.g., with density functional theory (DFT)] in large spaces of functional materials and catalysts is plagued by combinatorial challenges^{34,52–54} that can only be partly reduced through inverse design strategies (e.g., with alchemical derivatives^{53,55–57} or heuristics⁵⁸). Nevertheless, all such approaches that require an electronic structure calculation to be completed for property prediction or optimization hamper the large-scale exploration that is required for large-scale material discovery.

For transferable ML models, opportunities and challenges for either conventional 3D-structure-based^{59–63} or graph-based^{64–66} representations remain. Given the importance of this challenge, numerous groups have made effort to understand how to transfer information learned on small molecules to larger molecules (e.g., with persistent homology,⁶⁷ using a smooth overlap of atomic positions (SOAP) kernel or atom centered symmetry functions⁶⁸ or graph neural networks⁶⁹) and how to ensure transferability across different datasets⁷⁰ and to new solid state materials with unseen elements (e.g., using Euclidean neural networks⁷¹). Here, we focus on graph-based revised autocorrelations^{65,72,73} (RACs) that are no-cost products and differences of heuristic atomic properties (e.g., nuclear charge or Pauling electronegativity) on the molecular graph that have been demonstrated for transition metal chemistry. The 155-feature form of RACs (i.e., RAC-155) was used⁹ to train a multi-task artificial neural network (ANN) that accelerated identification of soluble and high redox potential transition metal complexes (TMCs) for redox flow batteries from nearly 3 M candidates in weeks instead of decades.⁹ In this or other discovery campaigns for which there is a reasonable degree of similarity between training and test data, RAC-155 and its feature-selected subsets have been used^{9,35,38,65,74–77} to train ML [e.g., kernel ridge regression (KRR) and ANN] models that exhibit good performance on modestly sized (~200 to 1000 data points) datasets of DFT properties for TMCs, including around 1–4 kcal/mol accuracy for adiabatic spin-splitting energies,^{65,76} bond lengths,^{65,74,78} ionization/redox potentials,⁶⁵ HOMO energies,⁷⁵ and catalytic properties.³⁸

Exemplary of how RAC-trained models have limited domains of applicability, Janet *et al.* observed⁷⁶ that introducing new elements (e.g., P or As) into an out-of-distribution test molecule uniquely challenged ANN models trained only on first-row, 3*d* open-shell transition-metal complexes with primarily 2*p* ligand chemistry. Conversely, introducing diverse metal coordination environments with a wider range of sampled elements into model training greatly improved ANN model generalization,⁷⁹ reducing mean absolute

errors by around 25%–30%. Because the distance in feature space (e.g., in KRR) or latent space (i.e., in the ANN) is a critical component, prediction by ML models of new materials with compositions not well supported by training data is universally challenging. Nevertheless, RACs and other widely used representations⁶⁰ employ the nuclear charge in the representation, exaggerating differences between rows of the Periodic Table with respect to differences in the group number.⁸⁰ While some heuristic properties in RACs (e.g., the Pauling electronegativity) more faithfully encode chemical similarity,⁶⁴ the nuclear-charge-based features are often emphasized in feature-selected subsets, highlighting their importance in model prediction.⁷⁴

Limitations in the representation will make it challenging for an ML model to learn relationships between isovalent materials with elements from different rows of the Periodic Table, despite clear relationships^{81–83} in both ligand field strength (e.g., 3*p* P > S and 2*p* N > O) and *d*-filling trends of spin-splitting energies⁸⁴ between 3*d* and 4*d* TMCs. One may anticipate that changing the row of the metal requires a change in the ligand chemical composition (e.g., for nitrogen-coordinating ligands vs phosphorus-coordinating ligands) as well. The existence of these structure–property relationships presents an opportunity for representation improvement. In this work, we tackle the challenge of improving the transferability of no-cost representations by introducing the group number as a heuristic into RACs, an approach previously demonstrated only in closed-shell organic molecules^{85,86} or in metallic alloys.^{33,87} We show how this always improves model performance and devise a synergistic strategy for leveraging large amounts of data containing elements from one row of the Periodic Table to improve the learning curves on materials belonging to a new target row. While demonstrated for the first time here on open shell 3*d*/4*d* learning tasks, we expect this approach to be broadly useful for other materials.

II. APPROACH

A. Dataset construction

Mononuclear, octahedral TMCs were assembled from combinations of eight transition metals with ten small, common ligands to explore relationships among isovalent metal centers and metal-coordinating ligand atoms. While the electronic structure of some of these TMCs were discussed in Ref. 84, we reiterate the rationale for their curation in paired datasets for ML on isovalent properties here. The TMCs contain 3*d* or 4*d* mid-row transition metals (Cr/Mo, Mn/Tc, Fe/Ru, and Co/Rh) in +2 and +3 oxidation states (Fig. 1). We study the energetically accessible spin states for these metals, which correspond to low-spin (LS), intermediate-spin (IS), and high-spin (HS) states, where the IS and HS states differ by having two more or four more unpaired electrons, respectively, with respect to the singlet or doublet LS state. The earliest Cr(III)/Mo(III) and latest Co(II)/Rh(II) are studied in only two states, a LS doublet and an IS quartet (supplementary material, Table S1).

The ten common ligands were selected for their varying field strengths (i.e., from weak field halides to strong field carbonyl). Six (i.e., three pairs) of these ligands contain isovalent coordinating atoms that differ in the principal quantum number, e.g., 2*p* vs 3*p* NH₃ and PH₃ (Fig. 1). A large set of possible TMC isomers were assembled from up to two of these ligands (i.e., L₁ and L₂) in four arrangements: (i) homoleptic M(L₁)₆; (ii) trans M(L₁)₄(L₂)₂; (iii) cis

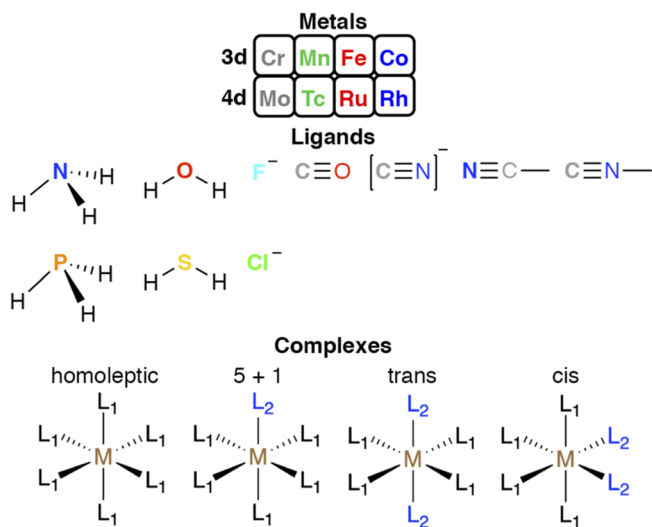


FIG. 1. Overview of eight metals and ten ligands used to form mononuclear, octahedral transition-metal complexes in the datasets in this work. (Top) 3d and 4d metals grouped by the row and column order in the Periodic Table. (Middle) Skeletal structures of the ten ligands with the metal-coordinating atom shown in bold. The six left-most ligands are grouped by isovalent 2p metal-coordinating atoms in the first row (i.e., N, O, and F) with 3p metal-coordinating atoms in the second row (P, S, and Cl). Four additional ligands that have carbon or nitrogen coordinating atoms are depicted at right. (Bottom) Symmetries of transition-metal complexes studied by assembling up to two ligand types (L_1 in black and L_2 in blue) in octahedral complexes: homoleptic ($L_1 = L_2$), 5 + 1 (a single axial L_2), and trans or cis placement of two L_2 ligands.

$M(L_1)_4(L_2)_2$; and (iv) 5 + 1 $M(L_1)_5(L_2)$ (Fig. 1). All homoleptic (i.e., 440) TMCs and trans or 5 + 1 (i.e., 3960 each) structures were set up for calculation with DFT, whereas only a subset (1,080) of the cis structures were set up (see Sec. III A and the supplementary material, Table S2). For the TMCs in this set, we compute (see Sec. III A) up to three gas-phase, adiabatic spin-splitting energies, i.e., between the LS and HS states, ΔE_{H-L} , as well as between the IS state and either the LS or the HS state (i.e., ΔE_{H-I} and ΔE_{I-L}), which were discussed in part in Ref. 84. In this work, for a subset of the TMCs, we now also compute a vertical ligand dissociation (LD) energy, ΔE_{LD} , the energy required to rigidly remove a single axial ligand (see Sec. III A).

From nearly ten thousand hypothetical TMCs, we obtain a smaller subset for testing the transferability of ML models by pairing isovalent species with differing principal quantum numbers via a filtering procedure that is more restrictive than that used in Ref. 84. For the isovalent metal pairing set, we compare computed properties of pairs of complexes that have metal centers in the same group of the Periodic Table but differ in their principal quantum number [e.g., 3d Fe(II) and 4d Ru(II)] with identical characteristics for all other TMC properties (e.g., spin multiplicity, L_1/L_2 ligands, and symmetry class). After accounting for data fidelity and convergence rates, we obtain over 1200 pairs of spin-splitting ΔE values depending on the spin states compared with the I-L set roughly twice as large as the H-L or H-I sets. The isovalent metal pairing set of ΔE_{LD} values across spin states contains over 1000 pairs (supplementary material, Table S3).

We also construct an isovalent ligand pairing set, which compares properties of pairs of TMCs that have ligands with metal-coordinating atoms that belong to the same group of the Periodic Table but differ in the principal quantum number (e.g., 2p N and 3p P). Because only six of the ten ligands belong to one of these 2p/3p pairs, heteroleptic TMCs could include between zero and two unique ligands from this subset. For the pairs in the isovalent ligand pairing set, we change all relevant ligands simultaneously while holding any incompatible ligands and all other TMC properties (e.g., metal and oxidation and spin state) fixed. Because we start from a smaller subset of TMCs that have the relevant ligands, isovalent ligand pairing subsets of spin-splitting and ligand dissociation energies are about half of the size of the isovalent metal pairing sets (supplementary material, Table S4). Despite the importance of machine learning to accelerate discovery in transition metal chemistry, our dataset here represents the first systematic approach to identify transferability of properties when changing both the metal center and ligand principal quantum numbers. The extent to which trends exist in this data will be assessed in Sec. IV where we show some trends exist but not enough to rely on simple models (i.e., linear models), thus motivating development of transferable ML representations (Sec. II B).

B. Representations

We employ Moreau–Broto autocorrelations (ACs)^{72,88} and revised AC variants (i.e., RACs),⁶⁵ which are representations formed by discrete operations on the heuristic properties of atoms in the molecular graph in analogy to continuous autocorrelations. A standard AC is evaluated over all of the n atoms in a molecule that are d bonds apart,

$$P_d = \sum_i^{\text{start scope}} \sum_j P_i P_j \delta(d_{ij}, d),$$

where P is the chosen property and P_i and P_j are the specific values of the i th and j th atoms, respectively, that are d_{ij} bonds apart. The five heuristic properties^{65,72,73,88,89} we have employed for transition-metal chemistry^{65,89} include the Pauling electronegativity, nuclear charge, covalent radius, topology, and the identity (i.e., 1) and produce a $5d + 5$ dimensional vector. ACs have achieved best-in-class performance for geometry-free descriptor predictions of atomization energies on subsets of QM9³² using KRR⁶⁵ or ANN models⁷⁶ with diminishing returns observed for depth cutoffs above three,⁶⁵ although higher depth cutoffs are sometimes used.^{73,89}

Extensions of ACs to enrich metal-local descriptors⁹⁰ in the feature vector led to the development of RACs, as first introduced in Ref. 65. RACs include centering (i.e., always including in the sum) the property evaluation on the metal (i.e., mc-RACs) or the ligand atom that coordinates the metal (i.e., lc-RACs). The lc-RACs are averaged separately over ligands of specific types, which were referred to as the scope in Ref. 65. In the octahedral complexes studied here, lc-RACs are averaged separately over the equatorial or axial ligands. RACs also introduced the difference of two atomic properties,

$$P'_d = \sum_i^{\text{start scope}} \sum_j (P_i - P_j) \delta(d_{ij}, d),$$

which can be non-trivially computed for metal-centered or ligand-centered RACs excluding the identity property for any $d > 0$ bond

paths. In total, the complete RAC set evaluated on the five original heuristic properties consists of $30d + 30$ product-based and $12d$ difference-based RACs for a total feature vector size of $42d + 30$. For a typical⁶⁵ cutoff of $d = 3$, this yields 156 possible features. For the mononuclear octahedral TMCs that are the focus of this work, excluding five constant features (i.e., corresponding to connectivity around the metal center) produces 151 features. In the original work, four additional descriptors of ligand denticity, oxidation state, spin state, and Hartree–Fock exchange fraction in the DFT functional were added, leading to the name RAC-155.⁶⁵ In this work, we always include the oxidation state (i.e., for 152 total features) and the spin state is included only for ΔE_{LD} (i.e., for 153 total features) but still refer to the set as RAC-155.

Because RAC heuristic properties depend on the properties of individual atoms, we revisit the extent to which such properties encode useful information about chemical similarity of distinct elements. Some, such as the Pauling electronegativity, encode close chemical similarity of distinct elements (e.g., χ : 1.88 for Co vs 1.90 for Tc). Still, the nuclear charge, Z , exaggerates the dissimilarity of elements with differing principal quantum number belonging to the same group (e.g., Z : 27 for Co vs 45 for Rh or 7 for N vs 15 for P). Thus, we expect low-complexity kernel models trained on RAC-155 to exaggerate differences in TMCs with metals or ligands from differing rows of the Periodic Table in comparison to those from the same row. To illustrate this, we carried out principal component analysis (PCA) on a subset (i.e., $d = 0$ or 1) of mc-RACs in RAC-155 applied to all isovalent metal pairing (i.e., with both $3d$ and $4d$ metals) complexes for which we have computed the ΔE_{I-L} property. For these TMCs, the first PC in this subset of RAC-155 primarily distinguishes the metal centers (Fig. 2). Not only are $3d$ and $4d$ metals very distant from each other, but the strong variation of the metal Z over the set in comparison to the other heuristics leads to the later $3d$ Co being the most proximal metal to the early $4d$ Mo.

We thus add to our original set of RAC heuristic properties the group number, G , which had been shown to be beneficial for increasing transferability in organic molecules.^{85,86} This property increases

the number of feature vector dimensions by $6d + 6$ product-based and $3d$ difference-based RACs. For the $d = 3$ cutoff used here, this adds 33 features to our original 152- or 153-dimensional set, and we refer to this set as eRAC-185. The G is identical for isovalent species (e.g., 9 for Co and Rh and 5 for N and P) with distinct principal quantum numbers. As expected, PCA in a subset (i.e., $d = 0$ or 1) of the new G -derived mc-RACs distinguishes TMCs by the metal G in PC1 and by the metal-coordinating ligand atoms in PC2 (Fig. 2 and supplementary material, Fig. S1). As a result, isovalent $3d$ and $4d$ metals (e.g., Co and Rh) will be identical in this feature set, and the full G set of features will order TMCs by increasing d filling and groups isovalent $3d$ and $4d$ metals close to each other (supplementary material, Figs. S2 and S3). Conversely, both PCA and average Euclidean distances over isovalent metal pairing complexes for the full RAC-155 set indicate large differences between isovalent $3d$ and $4d$ metal centers and do not reproduce expected trends from the Periodic Table (supplementary material, Figs. S2 and S3). Because the new features encode intuitive relationships, we expect that incorporating them into ML property predictions should be beneficial for learning tasks that span multiple rows in the Periodic Table. Despite this natural expectation, the extent to which including intuitive relationships is necessary to improve property predictions from models and in which data regimes this relationship is most needed to pass to the model will be addressed in Sec. IV.

III. METHODS

A. Electronic structure calculations

Some of the DFT data used for training in this work was previously published in Ref. 84, and all data generation steps followed an established protocol⁷⁵ for DFT training data for octahedral, mononuclear transition-metal complexes,⁴⁵ which we reiterate here. DFT geometry optimizations were performed using a development version of TeraChem v1.9.^{91,92} The B3LYP^{93–95} global hybrid functional was employed with the LANL2DZ⁹⁶ effective core potential for transition metals and the 6-31G* basis⁹⁷ for all other atoms. Singlet calculations were carried out in a spin-restricted formalism, while all other calculations were unrestricted. Level shifting⁹⁸ was employed to aid self-consistent field convergence with the majority-spin and minority-spin virtual orbitals each shifted by 0.25 Ha. Geometry optimizations were carried out in translation rotation internal coordinates⁹⁹ using the L-BFGS algorithm. Default tolerances of 4.5×10^{-4} hartree/bohr and 10^{-6} hartree were applied in the convergence criteria for the gradient and energy difference between steps, respectively.

Initial structures were generated with molSimplify,^{76,100,101} which uses Open Babel^{102,103} as a backend, and calculations were automated with molSimplify Automatic Design (mAD).^{40,75} The mAD calculations run for 24 h of wall time with up to five resubmissions. Geometric criteria⁷⁵ were applied after each resubmission and at tighter thresholds on the final structure (supplementary material, Table S5). We use the same geometric criterion values as in Ref. 75, which preserve connectivity and penalize excessive asymmetry, except we have tightened the criterion on metal–ligand distance asymmetry for homoleptic complexes (supplementary material, Table S5). Out of 9440 initiated $3d$ and $4d$ geometry optimizations, converged structures were obtained for 8729 TMCs, and 6480 TMCs satisfied all geometric criteria (supplementary material, Table S6).

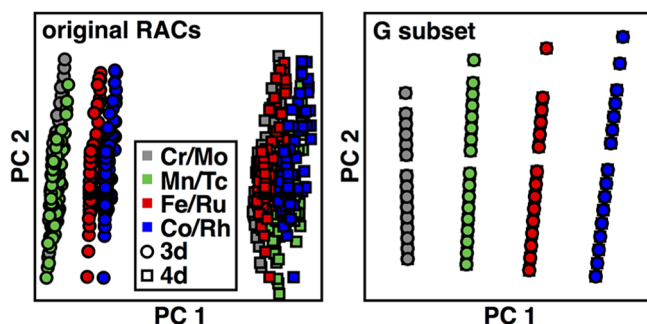


FIG. 2. PCA obtained using the subset of features in RAC-155 that are metal-centered and $d = 0$ or 1 (left) and using only the metal-centered RACs of the G atomic property with the same cutoff (right). The PCA was applied to 642 TMCs from the isovalent metal pairing dataset. The $3d$ TMCs (circles) and $4d$ TMCs (squares) are overlapping in the right PCA. Metals are colored according to their isovalent pairing: gray for Cr/Mo, green for Mn/Tc, red for Fe/Ru, and blue for Co/Rh, as indicated in the inset legend.

As in previous work,^{45,75,76} criteria based on electronic structure properties were also used to filter the dataset. For all open shell (i.e., non-singlet) complexes, calculations with deviations of $\langle S^2 \rangle$ from its expected $S(S + 1)$ value by $1 \mu_B^2$ or more were eliminated (196 runs) as were cases where the Mulliken spin on the metal was at least $1 \mu_B$ lower than the expected total spin (292 runs, [supplementary material](#), Table S6). After all filtering steps, 5992 octahedral $3d$ or $4d$ TMCs were retained for spin-splitting property prediction, but because evaluation of this property requires the convergence of multiple spin states, the final dataset sizes were smaller ([supplementary material](#), Table S7). The ΔE_{LD} property was only evaluated for trans or homoleptic complexes with neutral ligands, and filtering steps by the $\langle S^2 \rangle$ and Mulliken metal spin criteria were also applied to these complexes ([supplementary material](#), Table S8).

The retained TMCs were then processed through two additional steps to generate the datasets used in ML model training, a more restrictive procedure than that outlined in Ref. 84. First, TMCs were paired to construct the isovalent metal pairing and isovalent ligand pairing datasets, reducing the candidate TMC set sizes by a factor of two ([supplementary material](#), Tables S3 and S4). We then detected cases where the electronic structure was qualitatively distinct between isovalent pairs of complexes, as judged by the metal d -orbital occupations. The occupation of each d orbital was obtained from the NBO v6.0 package,³⁸ and a heuristic cutoff of $3e^-$ for the total difference summed over all d orbitals was used to exclude a pair of calculations ([supplementary material](#), Fig. S4). This procedure eliminates a small number (~ 10 to 100 for spin-splitting and 250 for ΔE_{LD}) of outliers ([supplementary material](#), Tables S3 and S4). The final isovalent metal pairing datasets contain ~ 300 to 600 pairs of spin-splitting (i.e., I-L, H-L, or H-I) ΔE values and over 1000 pairs of ΔE_{LD} values, and the final isovalent ligand pairing datasets are a little more than half the size of the isovalent metal pairing sets ([supplementary material](#), Tables S3 and S4). The total energies and structures of all TMCs converged in this work along with details of structures eliminated are provided in the [supplementary material](#).

B. Machine learning

We largely follow a previously developed protocol^{65,75} for training kernel ridge regression (KRR) models to predict properties in open-shell transition-metal chemistry with some modifications noted. KRR models with a Gaussian kernel were trained on both RAC-155 and eRAC-185 representations and isovalent metal pairing or isovalent ligand pairing data using the scikit-learn¹⁰⁴ software package. The pairs of complexes in the isovalent metal pairing or isovalent ligand pairing sets were randomly partitioned into 80% train and 20% validation. We use the 20% validation data to select hyperparameters and then report the errors on this set as a lower bound for the test error. The pairs are preserved during the random partitioning, e.g., if an isovalent metal pairing $3d$ complex is in the training set, its isovalent $4d$ counterpart will be as well. All inputs and outputs were normalized to have zero mean and unit variance on the training data.

Two types of models were trained: single-row models that only are trained on a subset of the training data (in practice, from 1 to 50 points) from the same row as the learning task and transfer models that are supplemented with all of the training data from the alternate row. Using the 324-pair of complex isovalent metal pairing set

for the $3d$ to $4d$ learning task as an example, models (see Sec. IV) were trained on a subset of the 259-complex 80% $4d$ -only training data (i.e., single-row models) or this same subset of points along with the 80% $3d$ data (i.e., transfer models). Exhaustive grid search from 10^{-12} to 10^2 with logarithmic spacing was used to select hyperparameters (i.e., kernel width and regularization strength). For the isovalent metal pairing $3d$ to $4d$ example, hyperparameters selected from ten-fold cross-validation errors on $3d$ data produce models with low test set mean unsigned error (MUE) on the in-distribution $3d$ complexes but high MUE on the out-of-distribution $4d$ test set (i.e., the isovalent pairs to the 20% $3d$ test set, [supplementary material](#), Fig. S5). Thus, we select hyperparameters for each KRR model on the 20% data from the same row as the learning task and hereafter refer to this as the validation MUE, which provides a lower bound of a true unseen test set model error ([supplementary material](#), Fig. S5).

Feature selection was employed as motivated by observations^{65,75} that it improves KRR model generalization in open-shell transition-metal chemistry. Feature-selected subsets were obtained in a one-shot fashion using LASSO¹⁰⁵ or random forest (RF)¹⁰⁶ as in some prior work.⁶⁵ Features with an importance greater than 1% were retained from RF. For LASSO feature selection, five values for the regularization strength were tested (10^{-2} , 10^{-1} , 10^0 , 10^1 , or 10^2), and all nonzero features were retained from the LASSO model with the lowest validation MUE. The overall feature set (i.e., from LASSO, RF, or the full feature set) that produced the lowest validation MUE was selected for training the KRR model. We report the MUE as the ensemble-average MUE and a credible interval from the ensemble standard deviation. A representative example of the selected hyperparameters, the feature selection method, and MUEs is provided in detail in the [supplementary material](#) (Table S9). All other model hyperparameters and feature sets are provided in the [supplementary material](#).

IV. RESULTS AND DISCUSSION

A. Correlations of properties between rows

Because our modified representations are motivated by the assumption that isovalent metals or ligands should exhibit similar properties and structure–property relationships, we first validate this expectation over the isovalent metal pairing and isovalent ligand pairing sets. For the isovalent $3d$ and $4d$ transition metals in the isovalent metal pairing set, we observe good correlations between the adiabatic gas-phase spin-splitting energies of the first- and second-row metals (Fig. 3). This correlation holds somewhat better ($R^2 = 0.9$) for the two-electron ΔE_{I-L} and four-electron ΔE_{H-L} than for the two-electron ΔE_{H-I} ($R^2 = 0.65$) values (Fig. 3 and [supplementary material](#), Table S10). As observed in prior work,⁸⁴ all $4d$ TMC DFT spin-splitting energies are consistently low-spin shifted with respect to equivalent $3d$ TMCs, especially for strong-field ligands (i.e., with positive $3d$ TMC ΔE values). For the ΔE_{I-L} and ΔE_{H-L} cases where the correlations hold best, the $4d$ TMC ΔE increases by around 1.4 kcal/mol for every 1 kcal/mol increase in the $3d$ TMC (Fig. 3 and [supplementary material](#), Table S10).

Despite the apparent good linear correlation over the dataset, the MUE of the linear model for predicting $4d$ ΔE_{H-L} from its $3d$ TMC value is still large at around 10 kcal/mol, and the maximum unsigned error is even larger at 45 kcal/mol ([supplementary material](#), Table S10). While MUEs are reduced somewhat for ΔE_{H-I}

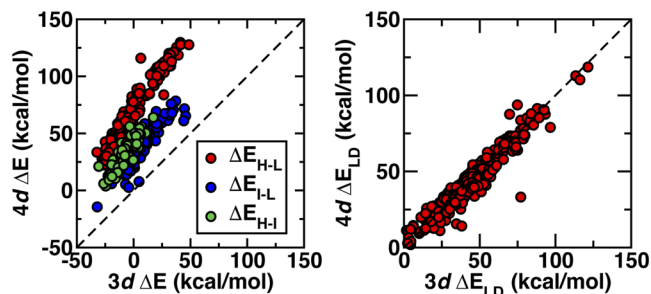


FIG. 3. Parity plot for DFT-calculated spin-splitting energies, ΔE (left, in kcal/mol), and DFT-calculated ligand dissociation energies, ΔE_{LD} (right, in kcal/mol), of Fe/Ru TMCs in the isovalent metal pairing dataset. The spin-splitting energy is colored with ΔE_{H-L} in red, ΔE_{I-L} in blue, and ΔE_{H-I} in green, and all data are shown in red at right for ΔE_{LD} . The parity line is shown as a dashed black line in both panes.

and ΔE_{I-L} (to ~ 7 kcal/mol), the range of values are also smaller for these properties. Each individual $3d$ or $4d$ metal/oxidation state spans a wide range of ΔE_{H-L} values (~ 100 kcal/mol) with metal-specific distributions that reflect changes in ligand field strength in our set (i.e., from weak-field halide to strong-field methylisocyanide) for both $3d$ and $4d$ TMCs (supplementary material, Fig. S6). When linear models are fit on each individual metal and oxidation state, the MUE of each linear model is somewhat reduced (e.g., ΔE_{H-L} 4–5 kcal/mol vs 10 kcal/mol for the full set) in part due to the fact that a smaller range of ΔE values is sampled (supplementary material, Figs. S7–S9 and Tables S10 and S11). Subdividing the data by metal/oxidation state and building individual linear relationships as a strategy for TMC design would, nevertheless, be challenged by variability in dataset sizes (supplementary material, Tables S11 and S12). Thus, even though these spin-splitting properties are related for metal pairs in the isovalent metal pairing set, their relationship is not trivial and could be expected to benefit from a flexible ML regression model. A predictive ML model would have the added benefit of not requiring the calculation of one property (e.g., $3d$ ΔE_{H-L}) to predict the other (e.g., $4d$ ΔE_{H-L}) that would be required for even a better linear fit.

For the ligand dissociation energy, ΔE_{LD} , introduced in this work, a single linear model obtains good correlation ($R^2 = 0.91$) between $3d$ and $4d$ TMCs over all metals, oxidation states, and spin states (Fig. 3). The ratio of $4d:3d$ ΔE_{LD} values (~ 0.93) is close to unity but indicates that ligands bind $4d$ TMCs somewhat more weakly than equivalent $3d$ TMCs (Fig. 3 and supplementary material, Table S13 and Fig. S10). Subdividing ligand dissociation energies by metal/oxidation state or spin provides limited benefit because subset comparisons generally have similar $4d:3d$ ratios and correlation coefficients to the overall set (supplementary material, Tables S13 and S14 and Fig. S11). Whether overall or in metal/oxidation/spin-state-specific subsets, the MUE of the linear fits is somewhat lower (4–5 kcal/mol) than was observed for spin-splitting energies (supplementary material, Tables S11 and S13).

The isovalent ligand pairing set consists of TMCs containing ligands (i.e., six of ten) that vary by isovalent (i.e., $2p$ vs $3p$) changes to the metal-coordinating atom of the ligand. Depending on the metal and oxidation state, the correlation of $2p$ and $3p$ spin-splitting

energies ranges from good ($R^2 > 0.9$) to poor ($R^2 < 0.1$), where poorer correlations correspond most typically to unchanged spin-splitting energies with the ligand substitution and with earlier (i.e., d^3-d^4) transition metals (supplementary material, Fig. S12 and Table S15). The influence of the $2p$ to $3p$ isovalent substitution on spin-splitting energies depends on the degree of change in ligand field strength (Fig. 4). A consistent trend can be observed over all d^5-d^7 metal centers in either row for the H_2O/H_2S and NH_3/PH_3 substitutions with a good correlation and low standard deviation (Fig. 4 and supplementary material, Fig. S13 and Table S16). For the majority of $4d$ TMCs, this change corresponds to an increase in spin-splitting energy, but there are some exceptions especially with acetonitrile ligands (supplementary material, Table S17). While average trends can be observed over the dataset, linear relationships differ as the oxidation and spin states being compared are varied. For example, a correlation that works well for Co(II) metal centers does not transfer to either Co(III) or Rh(II) (supplementary material, Table S15). This observation motivates our focus on development of machine learning models.

Conversely, the limited change in field strength among the halides means these TMCs have a poorer correlation for the substitution (supplementary material, Fig. S13 and Table S16). For ΔE_{LD} , trends are less clear, but the $2p$ to $3p$ substitution of two axial ligands (i.e., including the ligand undergoing dissociation) generally decreases the ligand dissociation energy (by 12–16 kcal/mol) for the H_2O/H_2S and NH_3/PH_3 pairs (supplementary material, Tables S17 and S18). As with the isovalent metal pairing set, there is a clear structure to the isovalent ligand pairing data indicating that properties of $2p$ complexes are related to their $3p$ counterparts, but the relationships depend more strongly on the group identity of the metal-coordinating atom.

We briefly summarize the structure–property relationships uncovered for our learning tasks to test ML model and representation generalization across the Periodic Table. For isovalent metal pairing spin-splitting energies, $3d/4d$ relationships are sensitive to the spin states and ligand field strengths being compared, whereas

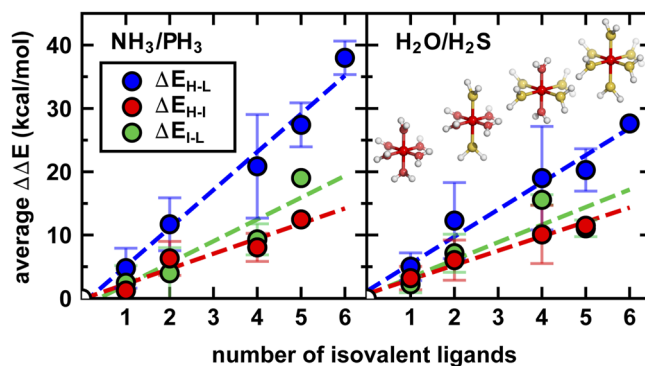


FIG. 4. Differences in DFT-calculated spin-splitting energies (ΔE_{H-L} in blue, ΔE_{H-I} in red, and ΔE_{I-L} in green in kcal/mol) with isovalent ligand swaps (NH_3 for PH_3 , left and H_2O for H_2S , right) in the isovalent ligand pairing dataset averaged over all d^5-d^7 $4d$ TMCs [i.e., Tc(II), Ru(III/II), and Rh(III/II)] where the spin-splitting energy is available and defined. The standard deviation is shown as an error bar along with best-fit lines also fit through (0,0). The majority of both $3d$ and $4d$ TMCs have a positive change in spin splitting from $2p$ to $3p$. Exceptions are listed in the supplementary material, Table S17.

isovalent metal pairing ΔE_{LD} values are comparable across $3d$ and $4d$ TMCs. Thus, conclusions about which features are essential for learning spin state ordering trends can be expected to differ from those based on ligand dissociation energy, highlighting the importance of comparing a range of properties when assessing featureizations. For isovalent ligand pairing spin-splitting energies, $3d/4d$ relationships are dependent on ligand field strength, and ligand dissociation energies are more clearly influenced by the principal quantum number of the dissociated axial ligands. Given the evidence for some relationships between rows of the Periodic Table, ML model training can be expected to benefit from the G heuristics present in eRAC-185 for both learning tasks that involve changing the metal or changing the ligand principal quantum number. Nevertheless, the relative extent to which these features improve property prediction requires further quantitative assessment (Sec. IV B).

B. ML property prediction across the periodic table

Based on the ability of linear models to predict isovalent metal pairing properties as well as some evidence of structure in the isovalent ligand pairing dataset, we expect that data from one row of the Periodic Table can be used to inform predictions on isovalent counterparts. To test this hypothesis, we construct a series of KRR models using both RAC-155 and the eRAC-185 representation. We adopt two strategies to test model generalization (see Sec. III B). While we train all models on a randomly selected and small subset (i.e., from 10 to 50) of training points from the same row as the 20% validation data, we also train “transfer” models that are supplemented with all available (i.e., the full 80%) training data from the alternate row. To unambiguously identify the training data included in each model, we adopt the notation *data-representation*, where the training *data* are only from the same row (S) as the test (here, also validation) data or transfer (T) data from the alternate row are also included and *representation* is the feature set (i.e., RACs or eRACs) used by the model. For example, a T-eRAC model applied to the $3d$ to $4d$ ΔE_{H-L} learning task will be provided with all available $3d$ ΔE_{H-L} data in the isovalent metal pairing set along with a small number (e.g., 10–50) of randomly selected $4d$ ΔE_{H-L} points from the isovalent metal pairing set and trained on eRAC-185 and feature-selected subsets (see Sec. III B).

We first focus on $3d$ to $4d$ prediction of ΔE_{H-L} and ΔE_{LD} in the isovalent metal pairing dataset to assess the relative performance of the possible data–representation combinations. Models incorporating data from both rows [i.e., T-(e)RACs] consistently outperform single-row models [i.e., S-(e)RACs], and simultaneously, eRAC-185-trained models have lower MUEs than models trained with RAC-155 (Fig. 5). The MUE is evaluated as the average error from an ensemble of 25 feature-selected models (see Sec. III B), and observed differences in model performance are typically larger in magnitude than the standard deviation of the errors from the ensemble. The benefit of eRAC-185 is most apparent when limited training data (i.e., <20 points) from the same row as the validation data are used. For example, prediction of ΔE_{LD} from the isovalent metal pairing dataset has a substantially lower MUE (by ~ 6.6 kcal/mol) with the T-eRAC model with 20 $4d$ points in the training set than with the S-eRAC model. When we increase the number of $4d$ training points modestly (i.e., to 50), the relative benefit of this additional data from another row is smaller (i.e., the MUE is lower by only ~ 3.5 kcal/mol) but still significant (supplementary material, Table S19).

To investigate if this observed maximum benefit of eRAC-185 when few examples from the same row as the validation data are available is a general phenomenon, we expanded our analysis to models trained to predict ΔE_{L-L} and ΔE_{H-L} with only 20 $4d$ data points. Indeed, we find that transfer models consistently outperform single-row models, and the use of eRAC-185 feature set is synergistic with this benefit (Fig. 6). The T-eRAC models consistently achieve the lowest overall MUEs for predicting all three spin-splitting properties, and those trained on eRAC-185 achieve larger reductions in MUE (~ 2 to 5 kcal/mol) than equivalent models training using RAC-155 (~ 1 to 2 kcal/mol). For predicting ΔE_{LD} , transfer models significantly improve over single-row models (by ~ 6 kcal/mol), but this improvement is observed for both RAC-155 and eRAC-185 representations (supplementary material, Table S19). We attribute the difference in ΔE_{LD} model errors to the observed lack of dependence on the principal quantum number of the metal for this property (see Sec. IV A). Comparing MUEs scaled by the property range in the dataset, T-RAC or T-eRAC models predict ΔE_{LD} with a scaled MUE roughly half of that of the S-eRAC model (0.04 vs 0.07, supplementary material, Table S20).

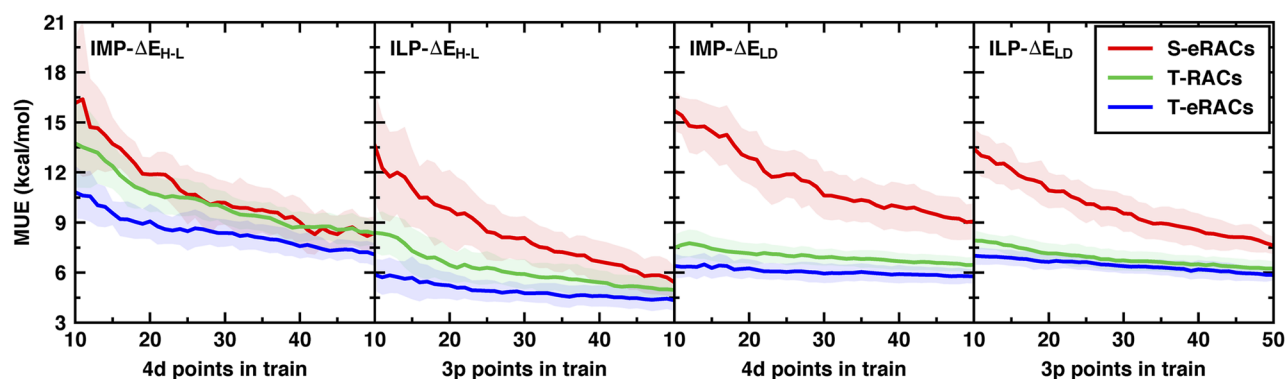


FIG. 5. Mean unsigned error (MUE in kcal/mol) for the prediction of ΔE_{H-L} (left) and ΔE_{LD} (right) properties for the isovalent metal pairing (IMP) $3d$ to $4d$ learning task and isovalent ligand pairing (ILP) $2p$ to $3p$ learning task with addition of training data as indicated on the x axis. Transfer-learning models (i.e., that contain all available $3d$ or $2p$ data) with eRAC-185 (blue) and RAC-155 (green) are shown alongside a single-row model (i.e., that only contains $4d$ or $3p$ data) trained on eRAC-185 (red). Shaded areas indicate the standard deviation of the ensemble of 25 models, and the solid line represents the average of this ensemble.

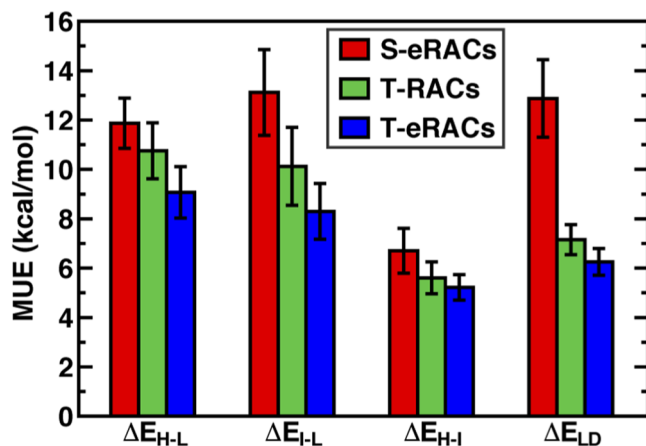


FIG. 6. Mean unsigned error (MUE in kcal/mol) for the prediction of 4d TMC properties with only 20 4d data points with eRAC-185 (red) or by including 3d data with 20 4d data points and using either the RAC-155 (green) or eRAC-185 (blue) on the isovalent metal pairing dataset. The colored bars represent the average from an ensemble of 25 feature-selected KRR models, and the error bars are the standard deviation of the ensemble.

Similar improvements are observed on scaled MUE values for the spin-splitting properties (i.e., from as high as 0.1 to as low as 0.05, [supplementary material](#), Table S20). While the lowest scaled errors from the transfer models are slightly larger than those obtained on models with the same metals in train and test data, they still represent a significant improvement.

To probe the source of benefit of the inclusion of the first 20 inter-row data points, we developed an additional ML model training procedure in which we grouped inter-row data by metal in order to preferentially add one metal at a time to the training set. When the training data includes only a single 4d metal (i.e., Mo, Tc, Ru, or Rh) for ΔE_{I-L} , no improvement in the model is observed ([supplementary material](#), Fig. S14). Rather than reducing MUEs with more data, the MUEs are instead high and constant at ~ 15 kcal/mol when one to 15 training points of this type are added ([supplementary material](#), Table S21). If data points are added from multiple additional 4d metals, performance instead improves significantly. In this case, MUEs for predicting ΔE_{I-L} reduce to ~ 11 kcal/mol when training data include two 4d metals ([supplementary material](#), Table S21). After adding fewer than 20 points, the T-eRAC model with knowledge of two 4d metals alongside all 3d metals outperforms the S-eRAC model (MUE: 13 kcal/mol) that was trained on 20 random points that included all 4d metals (Fig. 6 and [supplementary material](#), Fig. S15). The MUE is reduced further when a third 4d metal (i.e., to 7.8 kcal/mol) and all 4d metals (i.e., to 6.2 kcal/mol) are included in the training set ([supplementary material](#), Table S21). Thus, we expect that inter-row models perform best when examples of each new metal type are included in the training data by giving the model balanced information about the relationship among all groups of the Periodic Table studied in the validation set. The significance of the identified sample size of 20 inter-row data points maximizing benefits in transfer models is likely attributable to reaching a sample that includes sufficient diversity of isovalent metals or ligands. This result is non-trivial because at such small set sizes, the model has information about all metals but not in combination with

ligands and thus must infer the cooperativity of metal and ligand field effects.

Returning to the eRAC-185 representation, we tested whether its observed benefit in the isovalent metal pairing dataset 3d to 4d learning task is general to both the reverse learning task (i.e., 4d to 3d isovalent metal pairing) as well as to changes in ligand chemistry (i.e., 2p to 3p isovalent ligand pairing and 3p to 2p isovalent ligand pairing). Across all of these learning tasks, a combination of transfer models and eRAC-185 reduces validation MUEs for most properties with results and improvements comparable to those observed for the isovalent metal pairing 3d to 4d learning task ([supplementary material](#), Fig. S15). In all learning tasks and properties compared, isovalent ligand pairing 2p to 3p and 3p to 2p ΔE_{H-I} are the only cases where RAC-155-trained models outperform eRAC-185, but these differences are small (0.3 kcal/mol) and within the standard deviation (0.6 kcal/mol) of the 25-model ensembles ([supplementary material](#), Table S22). As in the isovalent metal pairing dataset, the benefit of inter-row, transfer models trained to predict ΔE_{LD} in the isovalent ligand pairing set (~ 6 kcal/mol) is observed to be independent of the representation (i.e., RAC-155 or eRAC-185) chosen. For spin-splitting energy prediction in the isovalent ligand pairing set, we again observe that eRAC-185 improves inter-row models with T-eRAC models outperforming single-row models by a slightly larger margin (~ 2 to 6 kcal/mol) than T-RAC models (~ 2 to 4 kcal/mol). Taken together, we conclude that the use of eRAC-185 in combination with a transfer model trained on a small set of examples from the new learning task provides a near-universal benefit and never significantly degrades the accuracy in comparison to a standard ML model. The converse is thus also true and an important conclusion from our study. Specifically, while some transferability could have been expected given the relationship of properties when changing the metal or ligand identity (Sec. IV A), once sufficient data are available, the benefit of feature engineering diminishes although it never worsens the learning performance. We expect these conclusions to be broadly true across other material classes, but to understand what chemical trends are learned by the models, we carry out further analysis of selected features next in Sec. IV C.

C. Feature analysis reveals learned chemical trends

To investigate the source of the improved performance observed when using eRAC-185, we evaluate the averaged, down-selected feature space. For each model in the ensemble, features are selected independently (i.e., from one-shot LASSO or random forest 1% cutoff, see Sec. III B), so this analysis is on the normalized feature contribution averaged over the KRR models in the ensemble. Features containing the *G* atomic property are frequently selected in inter-row models for predicting both spin-splitting energies and ΔE_{LD} (see the [supplementary material](#) for all features and models). For 3d to 4d prediction of ΔE_{I-L} (i.e., on the isovalent metal pairing set), *G* eRACs involving the metal center, atoms one bond-path away from the metal, and atoms two bond-paths away from the metal are all selected (Fig. 7). This suggests that the T-eRAC models achieve performance improvements by positioning both metals and ligands closer to their inter-row counterparts in the feature space than standard RACs. In contrast, for 3d to 4d prediction of ΔE_{LD} , *G* features involving the metal are rarely selected (Fig. 7). This analysis mirrors our observation that the expanded representation did not lead

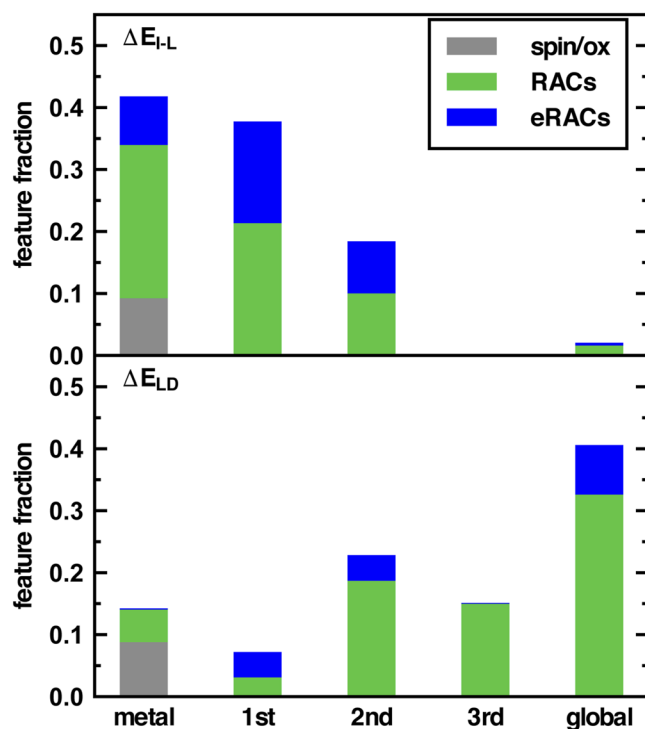


FIG. 7. Feature fraction grouped by the most metal-distant atom in the feature (i.e., metal, first, second, third, or global) and shown as a stacked bar plot. Colors within each feature fraction correspond to whether the features are obtained from RACs (green), the eRAC-specific G features (blue), or features encoding the metal center's spin and oxidation (ox) state (gray). The selected features are obtained from their normalized weight in an ensemble of 25 models (i.e., T-RACs or T-eRACs) trained to predict $4d$ properties using $3d$ data along with 20 $4d$ data points.

to significant improvements for a ΔE_{LD} inter-row, transfer model (see Sec. IV B) due to the relative lack of dependence of the ligand dissociation on the principal quantum number of the metal.

We analyzed other learning tasks (i.e., beyond $3d$ to $4d$ learning of ΔE_{I-L} and ΔE_{LD}) to confirm that the benefit of using eRAC-185 was due to the presence of G features in feature-selected subsets. Prior to feature selection, the G features comprise 18% of the initial eRAC-185 set, and they are frequently retained after feature selection regardless of the learning task or property (supplementary material, Fig. S16). For the ΔE_{I-L} and ΔE_{H-I} properties where G -based features were most beneficial in model performance (see Sec. IV B), G contributions are enriched (to $\sim 23\%$ of features) after feature selection. For ΔE_{LD} and ΔE_{H-I} , cases where G -based features provided less benefit, they are retained at a comparable frequency ($\sim 18\%$ of the down-selected features) to others in the set, confirming that these features are neither particularly detrimental nor beneficial (supplementary material, Table S23).

As in previous work,^{38,65,75,77} we also analyzed the selected features to reveal qualitative chemical trends in the dataset. We select ΔE_{I-L} $3d$ to $4d$ learning as a representative property for spin-splitting energy trends (supplementary material, Fig. S16). Features retained by models trained on ΔE_{I-L} are strongly metal-local with 80% corresponding to the metal or its immediate coordinating atoms. Since this result recapitulates observations from prior work⁶⁵ that

emphasized the metal-local nature of spin-splitting energy prediction, the tailored representation preserves known trends (Fig. 7 and supplementary material, Table S24). We next analyze the features selected for predicting ΔE_{LD} , which is a property for which we had not previously trained ML models. In this case, the retained features are 79% distal (i.e., with information from atoms two or more bonds away) from the metal (Fig. 7 and supplementary material, Table S24). The high weight of ligand-based features and more global information is consistent with our observations of reduced dependence on the principal quantum number of the metal for ΔE_{LD} prediction (Fig. 3). Thus, feature analysis suggests that it is possible to design ligand binding strength (e.g., for catalysis) in a metal-row-independent fashion (e.g., predicting $4d$ properties from data obtained on $3d$ TMCs). Conversely, because the spin state is dictated by both the metal center and only the immediate coordinating atoms of the ligand, inter-row transfer models trained on eRACs and minimal but chemically diverse TMCs⁷⁹ should enable accelerated prediction of $4d$ TMC spin-splitting in cases where only $3d$ TMC data are available.

Because the property prediction obtained from a KRR model depends on a combination of feature-space distances and the kernel width, the relative distances between TMCs in a representation are critical for understanding model performance. We thus quantify the Euclidean norm distance between complexes with differing metals in feature-selected subspaces that have been averaged and normalized over the 25-model ensembles used for property prediction. In the features selected in T-RAC models, these Euclidean distances do not represent what we would expect from the arrangement of elements on the Periodic Table (supplementary material, Table S25). For example, in the $3d$ to $4d$ prediction of ΔE_{I-L} , Cr is positioned closer to Fe and Co in the feature space than it is to the more chemically similar Mn. Furthermore, Cr is comparatively distant to all of the $4d$ metals, artificially reducing its influence on predicting their properties (Fig. 8). In contrast, the features retained in T-eRAC models rearrange the feature space to more closely match intuition from the Periodic Table with Cr in closest proximity to Mn and Mn in closest proximity to Cr and Fe. The $3d$ and $4d$ metal pairs are also shifted closer together in T-eRAC models. For example, $3d$ Cr is closer to isovalent $4d$ Mo than to a dissimilar $3d$ metal such as Co (Fig. 8). The rearrangement of the feature space is observed for other metals with $3d/4d$ pairs shifted an average of 13% closer in T-eRAC models in comparison to equivalent T-RAC models (supplementary material, Table S25). These results suggest that the reduced prediction error when using eRAC-185 (see Sec. IV B) can be attributed to this representation encoding chemical trends more efficiently.

To demonstrate the value of using the eRAC representation in chemical discovery, we apply transfer learning models with both RACs and eRACs to new out of sample sets of TMCs. For the out of sample sets, we construct $4d$ TMCs from monodentate OHLDB⁷⁹ ligands from previous work. The OHLDB ligands consist of up to two heavy (i.e., C, N, O, P, or S) atoms in combination with hydrogen atoms enumerated in an exhaustive fashion and scored against heuristic rules.⁷⁹ The goal of the OHLDB set was to generate a set of ligands distinct from more commonly studied ligands and thereby to enable chemical discovery of novel metal-local environments. We now evaluate our best-performing IMP ΔE_{H-I} transfer learning models on never-before-studied $4d$ TMCs of OHLDB ligands as a representative chemical discovery example. All

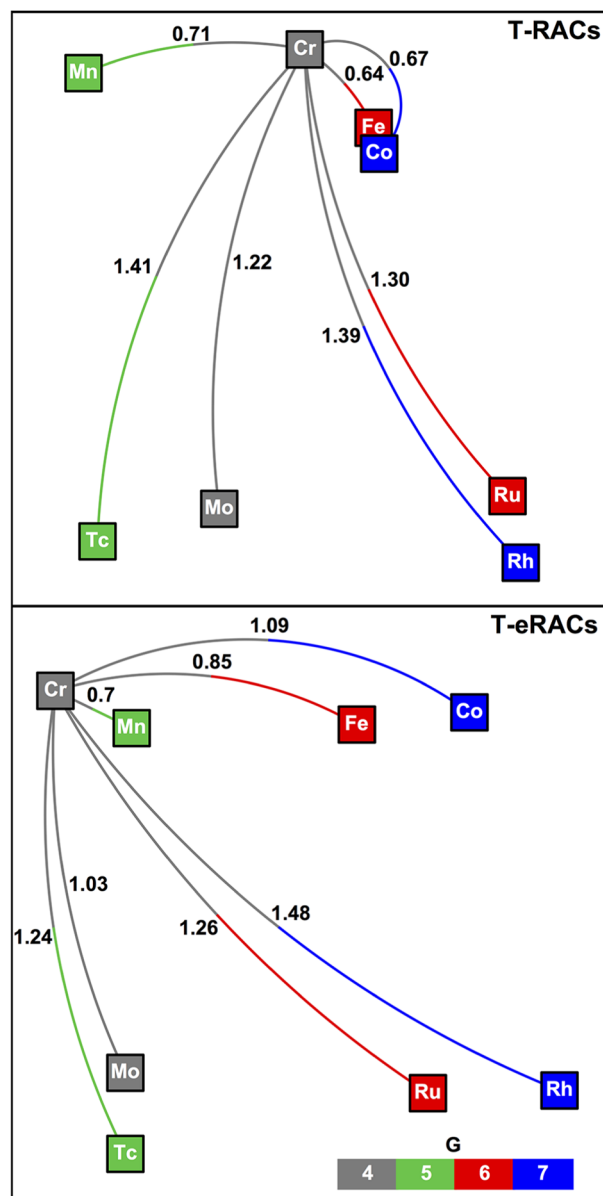


FIG. 8. Distribution of average, normalized distances of TMCs with the feature-selected subsets of T-RACs (top) or T-eRACs (bottom) from the ensemble of 25 KRR models trained to predict $4d \Delta E_L$ based on $3d$ data and 20 $4d$ examples (i.e., transfer models) visualized in a two-dimensional projection of the feature space. The average distances between Cr TMCs and TMCs with other metal centers are labeled. Elements are colored by their G values to highlight isoivalent $3d/4d$ pairs according to inset legend. The feature-space distances were normalized such that the average distance between any two TMCs is equal to one, and the weight of features selected from the ensemble was also normalized.

structures are provided in the [supplementary material](#). For consistency with our established approach, the best IMP transfer models were trained with $3d$ data and 20 seeded $4d$ data points (i.e., all from the initial ligand set) and applied to 70 new $4d$ homoleptic

OHLDB complexes. Prediction errors over this set are systematically higher for complexes containing $3p$ atoms two bonds distant from the metal, none of which are present in any of the training data ([supplementary material](#), Figs. S17 and S18). We thus focus our analysis on the subset of 52 data points that exclude more metal-non-local $3p$ atoms in ligands. Comparison of the T-RAC model and T-eRAC models for this subset emphasizes the advantage of using the group number as a descriptor ([supplementary material](#), Fig. S19). Despite having the same number of seeded $4d$ data points, the T-eRAC model predicts double the number of points (i.e., 26 vs 14) to lower than 10 kcal/mol error ([supplementary material](#), Fig. S19). The T-eRAC model also predicts all points with a lower overall MUE (21 kcal/mol vs 27 kcal/mol). For example, the T-eRAC model correctly predicts Tc(III)(methylamine)₆ to have a HS ground state with a modest error (i.e., 5 kcal/mol), whereas the T-RAC model error is over three times larger (19 kcal/mol, [supplementary material](#), Fig. S19). We emphasize that OHLDB ligands are particularly challenging for our models to predict because they represent a set of ligands designed to increase diversity from more commonly studied ligands. Thus, the T-eRAC approach is expected to be most beneficial when starting to collect data in spaces where chemistry is most distinct. We thus conclude that while many ML tasks have relied on large datasets and narrow chemical compositions, this work has highlighted the minimum new data needed to develop transferable ML models across a wider range of chemical compositions. While demonstrated here on mononuclear transition metal complexes with $3d$ and $4d$ metals and $2p$ and $3p$ ligands, we plan to next extend this approach to other heavier (i.e., $5d$ and $4f$) elements as well as more diverse ligand chemistry. In all such cases, this work has revealed opportunities in the small data limit to carry out chemical exploration once a minimum amount of data is used in combination with transferable ML representations.

V. CONCLUSIONS

For many materials design challenges in transition metal chemistry, it is common to find one element that works well for a chemical property or a transformation, even if it is preferred to use another due to price or scarcity of the element. Furthermore, when carrying out ML-accelerated chemical discovery efforts, one may start with a specific range of chemical compositions but then aim to change course. It is not known *a priori* how much prior data can be used in these cases. It is also unknown the extent to which chemical observations from one element (e.g., $3d$ vs $4d$) for either the metal or the ligand (e.g., $2p$ vs $3p$) within the same group of the Periodic Table should be transferable. Intuitively, chemical relationships between neighboring elements in the Periodic Table are expected, so efficient strategies for ML-driven discovery should leverage representations that preserve these similarities. To thus tackle these challenges, we carried out the first study of the extent to which ML models can best leverage transferability of properties between $3d$ and $4d$ transition metal complexes with $2p$ and $3p$ ligands.

After confirming the strong inter-row structure–property relationships for spin-splitting energies and ligand dissociation in isoivalent $3d/4d$ metals and $2p/3p$ ligands in large sets (~ 1000 pairs) of data, we identified a strategy for increasing ML model transferability. We introduced eRAC-185, a tailored representation that blends together the monotonically increasing nuclear charge with

the periodic in nature group number heuristics. We showed how the addition of the *G* heuristic property in eRACs altered feature-space distances to encode a more intuitive degree of similarity among isovalent elements essential for kernel-based ML models.

Next, we demonstrated the synergistic value of eRACs alongside a transfer learning approach to leverage inter-row data in sets containing isovalent metals or ligand-coordinating atoms in transition-metal chemistry. To assess the potential improved performance of eRAC-185, we trained KRR models using data from either only the same row as the prediction task or both rows of the Periodic Table. While including data from the alternate row always reduced prediction errors, the eRAC-185 model errors were most substantially improved in the limit of very low data (~20 points). The model error improvement was strongest when the relationship being learned deviated most from parity. This highlights how transferable representations will be essential to improve ML model generalization from one row of the Periodic Table to another especially when carrying out chemical discovery where limited prior knowledge is available.

To identify differences in models trained on eRAC-185, we analyzed the distance and heuristic atomic properties retained during feature selection. Feature-space distances of TMCs grouped and averaged by metals confirmed that eRAC-185 subsets faithfully represented trends expected from the Periodic Table, causing elements from the same or adjacent group to influence KRR predictions most strongly. These representations, models, and approaches can be leveraged to accelerate the discovery of earth-abundant catalysts or materials from known materials featuring scarcer elements. We expect our strategies for inter-row design and expanding transferability to be broadly useful to other material challenges across the Periodic Table.

SUPPLEMENTARY MATERIAL

See the [supplementary material](#) for spin multiplicity definitions; subset of combinations studied for cis TMCs; statistics for filtering and pairing structures as well as 3d/4d comparisons to obtain properties; PCA of RAC and eRAC representations; comparison of the Euclidean distance between metals by the featurization type; geometry metrics used for geometric filtering; number of compounds failed at each filtering step with reasons; details of NBO d orbital check elimination; cross-validation heat maps for hyperparameter selection; KRR model hyperparameters by the model number; linear models for inter-row property prediction for the isovalent metal pairing dataset, overall and broken down by metal; histograms of property distributions for spin-splitting energy; inter-row parity plots for properties by the metal and oxidation state; ligand dissociation energy distribution broken down by the metal and oxidation state; frequencies of ligand dissociation energy by metal, oxidation, and spin state in the isovalent metal pairing and isovalent ligand pairing datasets; parity plots for spin-splitting energy in the isovalent metal pairing set; changes in the property by ligand and mixing for three spin-splitting energy properties; linear models for spin-splitting energies in isovalent ligand pairing dataset split by metal, oxidation state, and type of spin-splitting energy; change in spin-splitting energy as a function of ligand mixing in the isovalent ligand pairing dataset; quantification of ligand substitution in the isovalent ligand pairing dataset quantified by the ligand type;

differences in ligand dissociation energy for compounds that vary only in equatorial ligands or only in axial ligands; quantification of test set errors for KRR model ensembles for various inter-row learning tasks; model performance by ordered metal addition for 4d metals; quantification of errors with ordered metal addition of 4d metals; mean unsigned errors for various inter-row learning schemes over the isovalent metal pairing and isovalent ligand pairing datasets; stacked bar charts for the feature character for different properties obtained during each scheme of inter-row learning and tabulation of the corresponding feature character by the property and bond depth; quantification of feature space distances for feature selected subspaces of 3d/4d models in the isovalent metal pairing dataset (PDF); structures and total energies of all 3d and 4d TMCs studied in this work; reasons complexes were not included in the dataset; spin-splitting values for all pairs of 3d/4d TMCs; ligand dissociation energies for 3d/4d TMCs; total energies of all TMCs; OHLDB out of sample predictions; and KRR model performance and selected features and ensemble quantification (ZIP).

ACKNOWLEDGMENTS

The authors acknowledge initial partial support in data generation steps and machine learning representation development (for A.N.) by the U.S. Department of Energy under Grant No. DE-SC0012702. This work was also supported by the Office of Naval Research under Grant Nos. N00014-17-1-2956 and N00014-18-1-2434 (D.R.H. and J.P.J.), Grant No. N00014-20-1-2150 (C.D., D.R.H., and N.A.), and DARPA Grant No. D18AP00039 (for N.A., C.D., and A.N.). Ligand dissociation studies were supported by the National Science Foundation under Grant No. CBET-1704266 (for A.N.). A.N. and N.A. were partially supported by a National Science Foundation Graduate Research Fellowship under Grant No. 1122374. H.J.K. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund and an AAAS Marion Milligan Mason Award, which supported this work. This work was carried out, in part, using computational resources from the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation Grant No. ACI-1548562. The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts of interest to disclose.

Author Contributions

D.R.H. and A.N. contributed equally to this work.

DATA AVAILABILITY

The data that support the findings of this study are available within the article and its [supplementary material](#).

REFERENCES

- ¹S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, *Nat. Mater.* **12**(3), 191–201 (2013).
- ²A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, *APL Mater.* **1**(1), 011002 (2013).

- ³B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, *Phys. Rev. B* **89**(9), 094104 (2014).
- ⁴P. O. Dral, *J. Phys. Chem. Lett.* **11**(6), 2336–2347 (2020).
- ⁵J. P. Janet, C. Duan, A. Nandy, F. Liu, and H. J. Kulik, *Acc. Chem. Res.* **54**(3), 532–545 (2021).
- ⁶K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, *Nature* **559**(7715), 547–555 (2018).
- ⁷A. C. Mater and M. L. Coote, *J. Chem. Inf. Model.* **59**(6), 2545–2559 (2019).
- ⁸B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel, and C. Sutton, *AIChE J.* **64**(7), 2311–2323 (2018).
- ⁹J. P. Janet, S. Ramesh, C. Duan, and H. J. Kulik, *ACS Cent. Sci.* **6**, 513–524 (2020).
- ¹⁰H. C. Herbol, W. Hu, P. Frazier, P. Clancy, and M. Poloczek, *npj Comput. Mater.* **4**(1), 51 (2018).
- ¹¹J. M. Hernández-Lobato, J. Requeima, E. O. Pyzer-Knapp, and A. Aspuru-Guzik, in paper presented at the Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, 2017.
- ¹²K. Tran and Z. W. Ulissi, *Nat. Catal.* **1**(9), 696–703 (2018).
- ¹³W. J. Kerr, G. J. Knox, M. Reid, and T. Tuttle, *Chem. Sci.* **12**(19), 6747–6755 (2021).
- ¹⁴L. Duan, F. Bozoglian, S. Mandal, B. Stewart, T. Privalov, A. Llobet, and L. Sun, *Nat. Chem.* **4**(5), 418–423 (2012).
- ¹⁵J. M. Zadrozny, M. Atanasov, A. M. Bryan, C.-Y. Lin, B. D. Reken, P. P. Power, F. Neese, and J. R. Long, *Chem. Sci.* **4**(1), 125–138 (2013).
- ¹⁶P. C. Bunting, M. Atanasov, E. Damgaard-Møller, M. Perfetti, I. Crassee, M. Orlita, J. Overgaard, J. van Slageren, F. Neese, and J. R. Long, *Science* **362**(6421), eaat7319 (2018).
- ¹⁷I. Vollmer, S. Ould-Chikh, A. Aguilar-Tapia, G. Li, E. Pidko, J.-L. Hazemann, F. Kapteijn, and J. Gascon, *J. Am. Chem. Soc.* **141**(47), 18814–18824 (2019).
- ¹⁸R. M. Bullock, J. G. Chen, L. Gagliardi, P. J. Chirik, O. K. Farha, C. H. Hendon, C. W. Jones, J. A. Keith, J. Klosin, and S. D. Minter, *Science* **369**(6505), eaac3183 (2020).
- ¹⁹P. Chirik and R. Morris, *Acc. Chem. Res.* **48**(9), 2495 (2015).
- ²⁰G. A. Filonenko, R. van Putten, E. J. M. Hensen, and E. A. Pidko, *Chem. Soc. Rev.* **47**(4), 1459–1483 (2018).
- ²¹W. Zuo, A. J. Lough, Y. F. Li, and R. H. Morris, *Science* **342**(6162), 1080–1083 (2013).
- ²²M. R. Friedfeld, H. Zhong, R. T. Ruck, M. Shevlin, and P. J. Chirik, *Science* **360**(6391), 888–893 (2018).
- ²³C. M. Klug, A. J. P. Cardenas, R. M. Bullock, M. O'Hagan, and E. S. Wiedner, *ACS Catal.* **8**(4), 3286–3296 (2018).
- ²⁴A. Le Goff, V. Artero, B. Jousset, P. D. Tran, N. Guillet, R. Métayé, A. Fihri, S. Palacin, and M. Fontecave, *Science* **326**(5958), 1384–1387 (2009).
- ²⁵T. N. Huan, R. T. Jane, A. Benayad, L. Guetaz, P. D. Tran, and V. Artero, *Energy Environ. Sci.* **9**(3), 940–947 (2016).
- ²⁶I. Azcarate, C. Costentin, M. Robert, and J.-M. Savéant, *J. Am. Chem. Soc.* **138**(51), 16639–16644 (2016).
- ²⁷D. J. Martin, B. Q. Mercado, and J. M. Mayer, *Sci. Adv.* **6**(11), eaaz3318 (2020).
- ²⁸A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, *Sci. Adv.* **3**(12), e1701816 (2017).
- ²⁹C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, *Chem. Mater.* **31**(9), 3564–3572 (2019).
- ³⁰F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld, *J. Chem. Phys.* **148**(24), 241717 (2018).
- ³¹O. A. Von Lilienfeld, *Angew. Chem., Int. Ed. Engl.* **57**(16), 4164–4169 (2018).
- ³²R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. Von Lilienfeld, *Sci. Data* **1**, 140022 (2014).
- ³³X. Zheng, P. Zheng, and R.-Z. Zhang, *Chem. Sci.* **9**(44), 8426–8432 (2018).
- ³⁴K. D. Vogiatzis, M. V. Polynski, J. K. Kirkland, J. Townsend, A. Hashemi, C. Liu, and E. A. Pidko, *Chem. Rev.* **119**, 2453–2523 (2018).
- ³⁵P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik, and D. Balcells, *Chem. Sci.* **11**(18), 4584–4601 (2020).
- ³⁶B. Meyer, B. Sawatlon, S. Heinen, O. A. von Lilienfeld, and C. Corminboeuf, *Chem. Sci.* **9**(35), 7069–7077 (2018).
- ³⁷M. Cordova, M. D. Wodrich, B. Meyer, B. Sawatlon, and C. Corminboeuf, *ACS Catal.* **10**(13), 7021–7031 (2020).
- ³⁸A. Nandy, J. Zhu, J. P. Janet, C. Duan, R. B. Getman, and H. J. Kulik, *ACS Catal.* **9**, 8243–8255 (2019).
- ³⁹S. M. Maley, D.-H. Kwon, N. Rollins, J. C. Stanley, O. L. Sydora, S. M. Bischof, and D. H. Ess, *Chem. Sci.* **11**(35), 9665–9674 (2020).
- ⁴⁰J. P. Janet, L. Chan, and H. J. Kulik, *J. Phys. Chem. Lett.* **9**, 1064–1071 (2018).
- ⁴¹S. M. Moosavi, A. Nandy, K. M. Jablonka, D. Ongari, J. P. Janet, P. G. Boyd, Y. Lee, B. Smit, and H. J. Kulik, *Nat. Commun.* **11**, 4068 (2020).
- ⁴²G. S. Fanourgakis, K. Gkagkas, E. Tylanakis, and G. E. Froudakis, *J. Am. Chem. Soc.* **142**(8), 3814–3822 (2020).
- ⁴³Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, and A. Aspuru-Guzik, *Nat. Mach. Intell.* **3**(1), 76–86 (2021).
- ⁴⁴M. Eckhoff, K. N. Lausch, P. E. Blöchl, and J. Behler, *J. Chem. Phys.* **153**(16), 164107 (2020).
- ⁴⁵C. Duan, J. P. Janet, F. Liu, A. Nandy, and H. J. Kulik, *J. Chem. Theory Comput.* **15**(4), 2331–2345 (2019).
- ⁴⁶T. Husch, J. Sun, L. Cheng, S. J. R. Lee, and T. F. Miller, *J. Chem. Phys.* **154**(6), 064108 (2021).
- ⁴⁷M. D. Wodrich, B. Sawatlon, M. Busch, and C. Corminboeuf, *Acc. Chem. Res.* **54**(5), 1107–1117 (2021).
- ⁴⁸M. Welborn, L. Cheng, and T. F. Miller III, *J. Chem. Theory Comput.* **14**(9), 4772–4779 (2018).
- ⁴⁹A. M. Chang, J. G. Freeze, and V. S. Batista, *Chem. Sci.* **10**(28), 6844–6854 (2019).
- ⁵⁰B. J. Bucior, A. S. Rosen, M. Haranczyk, Z. Yao, M. E. Ziebel, O. K. Farha, J. T. Hupp, J. I. Siepmann, A. Aspuru-Guzik, and R. Q. Snurr, *Cryst. Growth Des.* **19**(11), 6682–6697 (2019).
- ⁵¹K. Mikami, *Polymer* **203**, 122738 (2020).
- ⁵²D. J. Durand and N. Fey, *Chem. Rev.* **119**(11), 6561–6594 (2019).
- ⁵³T. Weymuth and M. Reiher, *Int. J. Quantum Chem.* **114**(13), 823–837 (2014).
- ⁵⁴M. Foscatto and V. R. Jensen, *ACS Catal.* **10**(3), 2354–2377 (2020).
- ⁵⁵J. G. Freeze, H. R. Kelly, and V. S. Batista, *Chem. Rev.* **119**(11), 6595–6612 (2019).
- ⁵⁶D. Xiao, L. A. Martini, R. C. Snoeberger III, R. H. Crabtree, and V. S. Batista, *J. Am. Chem. Soc.* **133**(23), 9014–9022 (2011).
- ⁵⁷A. M. Chang, B. Rudshiteyn, I. Warnke, and V. S. Batista, *Inorg. Chem.* **57**(24), 15474–15480 (2018).
- ⁵⁸M. Bergeler, G. N. Simm, J. Proppe, and M. Reiher, *J. Chem. Theory Comput.* **11**(12), 5712–5722 (2015).
- ⁵⁹M. Fernandez, N. R. Trefiak, and T. K. Woo, *J. Phys. Chem. C* **117**(27), 14095–14105 (2013).
- ⁶⁰M. Rupp, A. Tkatchenko, K. R. Müller, and O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**(5), 058301 (2012).
- ⁶¹A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**(18), 184115 (2013).
- ⁶²K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko, *J. Phys. Chem. Lett.* **6**(12), 2326–2331 (2015).
- ⁶³J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**(14), 146401 (2007).
- ⁶⁴D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, *arXiv:1509.09292* (2015).
- ⁶⁵J. P. Janet and H. J. Kulik, *J. Phys. Chem. A* **121**(46), 8939–8954 (2017).
- ⁶⁶E. Burello and G. Rothenberg, *Adv. Synth. Catal.* **347**(15), 1969–1977 (2005).
- ⁶⁷J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas, and K. D. Vogiatzis, *Nat. Commun.* **11**(1), 3230 (2020).
- ⁶⁸A. Grisafi, A. Fabrizio, B. Meyer, D. M. Wilkins, C. Corminboeuf, and M. Ceriotti, *ACS Cent. Sci.* **5**(1), 57–64 (2019).
- ⁶⁹Z. Liu, L. Lin, Q. Jia, Z. Cheng, Y. Jiang, Y. Guo, and J. Ma, *J. Chem. Inf. Model.* **61**(3), 1066–1082 (2021).
- ⁷⁰A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, *J. Chem. Phys.* **152**(4), 044107 (2020).
- ⁷¹Z. Chen, N. Andrejevic, T. Smidt, Z. Ding, Q. Xu, Y. T. Chi, Q. T. Nguyen, A. Alatas, J. Kong, and M. Li, *Adv. Sci.* **8**(12), 2004214 (2021).
- ⁷²G. Moreau and P. Broto, *Nouv. J. Chim.* **4**(6), 359–360 (1980).

- ⁷³A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang, and D. N. Beratan, *J. Am. Chem. Soc.* **135**(19), 7296–7303 (2013).
- ⁷⁴J. P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, S. Lin, and H. J. Kulik, *Inorg. Chem.* **58**, 10592–10606 (2019).
- ⁷⁵A. Nandy, C. Duan, J. P. Janet, S. Gugler, and H. J. Kulik, *Ind. Eng. Chem. Res.* **57**(42), 13973–13986 (2018).
- ⁷⁶J. P. Janet, C. Duan, T. Yang, A. Nandy, and H. J. Kulik, *Chem. Sci.* **10**, 7913–7922 (2019).
- ⁷⁷F. Liu, C. Duan, and H. J. Kulik, *J. Phys. Chem. Lett.* **11**, 8067–8076 (2020).
- ⁷⁸M. G. Taylor, T. Yang, S. Lin, A. Nandy, J. P. Janet, C. Duan, and H. J. Kulik, *J. Phys. Chem. A* **124**(16), 3286–3299 (2020).
- ⁷⁹S. Gugler, J. P. Janet, and H. J. Kulik, *Mol. Syst. Des. Eng.* **5**, 139–152 (2020).
- ⁸⁰Z. Allahyari and A. R. Oganov, *J. Phys. Chem. C* **124**(43), 23867–23878 (2020).
- ⁸¹A. Sundermann, M. Reiher, and W. W. Schoeller, *Eur. J. Inorg. Chem.* **1998**(3), 305–310.
- ⁸²M. Mastalir, M. Glatz, N. Gorgas, B. Stöger, E. Pittenauer, G. Allmaier, L. F. Veiros, and K. Kirchner, *Chem. Eur. J.* **22**(35), 12316–12320 (2016).
- ⁸³H. Tang, E. N. Brothers, and M. B. Hall, *Inorg. Chem.* **56**(1), 583–593 (2017).
- ⁸⁴A. Nandy, D. B. K. Chu, D. R. Harper, C. Duan, N. Arunachalam, Y. Cytter, and H. J. Kulik, *Phys. Chem. Chem. Phys.* **22**, 19326–19341 (2020).
- ⁸⁵C. Duan, F. Liu, A. Nandy, and H. J. Kulik, *J. Chem. Theory Comput.* **16**, 4373–4387 (2020).
- ⁸⁶C. Duan, F. Liu, A. Nandy, and H. J. Kulik, *J. Phys. Chem. Lett.* **11**, 6640–6648 (2020).
- ⁸⁷X. Li, R. Chiong, and A. J. Page, *J. Phys. Chem. Lett.* **12**, 5156–5162 (2021).
- ⁸⁸P. Broto, G. Moreau, and C. Vandycke, *Eur. J. Med. Chem.* **19**(1), 71–78 (1984).
- ⁸⁹J. P. Janet, T. Z. H. Gani, A. H. Steeves, E. I. Ioannidis, and H. J. Kulik, *Ind. Eng. Chem. Res.* **56**(17), 4898–4910 (2017).
- ⁹⁰J. P. Janet and H. J. Kulik, *Chem. Sci.* **8**, 5137–5152 (2017).
- ⁹¹S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, S. I. L. Kokkila-Schumacher, N. Luehr, J. W. Snyder, Jr., C. Song, A. V. Titov, and I. S. Ufimtsev, *J. Chem. Phys.* **152**(22), 224110 (2020).
- ⁹²S. Seritan, C. Bannwarth, B. S. Fales, E. G. Hohenstein, C. M. Isborn, S. I. Kokkila-Schumacher, X. Li, F. Liu, N. Luehr, and J. W. Snyder, Jr., *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **11**(2), e1494 (2021).
- ⁹³A. D. Becke, *J. Chem. Phys.* **98**(7), 5648–5652 (1993).
- ⁹⁴C. Lee, W. Yang, and R. G. Parr, *Phys. Rev. B* **37**, 785–789 (1988).
- ⁹⁵P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, *J. Phys. Chem.* **98**(45), 11623–11627 (1994).
- ⁹⁶P. J. Hay and W. R. Wadt, *J. Chem. Phys.* **82**(1), 270–283 (1985).
- ⁹⁷R. Ditchfield, W. J. Hehre, and J. A. Pople, *J. Chem. Phys.* **54**(2), 724 (1971).
- ⁹⁸V. R. Saunders and I. H. Hillier, *Int. J. Quantum Chem.* **7**(4), 699–705 (1973).
- ⁹⁹L.-P. Wang and C. Song, *J. Chem. Phys.* **144**(21), 214108 (2016).
- ¹⁰⁰E. I. Ioannidis, T. Z. H. Gani, and H. J. Kulik, *J. Comput. Chem.* **37**, 2106–2117 (2016).
- ¹⁰¹Kulik Group, <https://molsimplify.mit.edu> (last accessed February 3, 2022).
- ¹⁰²N. M. O’Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, *J. Cheminf.* **3**, 33 (2011).
- ¹⁰³N. M. O’Boyle, C. Morley, and G. R. Hutchison, *Chem. Cent. J.* **2**, 5 (2008).
- ¹⁰⁴F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- ¹⁰⁵R. Tibshirani, *J. R. Stat. Soc. B* **58**, 267–288 (1996).
- ¹⁰⁶L. Breiman, *Mach. Learn.* **45**(1), 5–32 (2001).