JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau

Article

Active Learning Exploration of Transition-Metal Complexes to Discover Method-Insensitive and Synthetically Accessible Chromophores

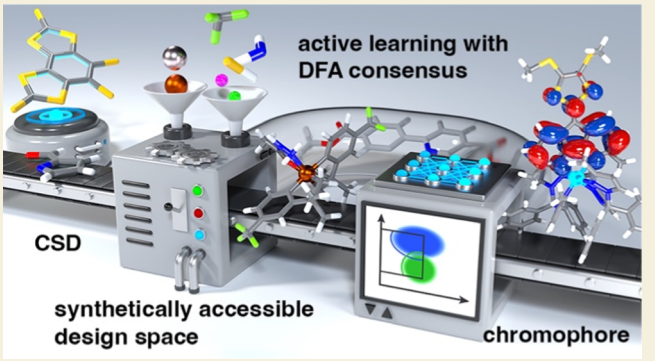
Chenru Duan, Aditya Nandy, Gianmarco G. Terrones, David W. Kastner, and Heather J. Kulik*

Cite This: *JACS Au* 2023, 3, 391−401

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Transition-metal chromophores with earth-abundant transition metals are an important design target for their applications in lighting and nontoxic bioimaging, but their design is challenged by the scarcity of complexes that simultaneously have well-defined ground states and optimal target absorption energies in the visible region. Machine learning (ML) accelerated discovery could overcome such challenges by enabling the screening of a larger space but is limited by the fidelity of the data used in ML model training, which is typically from a single approximate density functional. To address this limitation, we search for consensus in predictions among 23 density functional approximations across multiple rungs of "Jacob's ladder". To accelerate the discovery of complexes with absorption energies in the visible region while minimizing the effect of low-lying excited states, we use two-dimensional (2D)efficient global optimization to sample candidate low-spin chromophores from multimillion complex spaces. Despite the scarcity (i.e., ∼0.01%) of potential chromophores in this large chemical space, we identify candidates with high likelihood (i.e., >10%) of computational validation as the ML models improve during active learning, representing a 1000-fold acceleration in discovery. Absorption spectra of promising chromophores from time-dependent density functional theory verify that 2/3 of candidates have the desired excited-state properties. The observation that constituent ligands from our leads have demonstrated interesting optical properties in the literature exemplifies the effectiveness of our construction of a realistic design space and active learning approach.



KEYWORDS: *machine learning, transition-metal chromophore, active learning, chemical discovery, density functional theory*

INTRODUCTION

Transition-metal chromophores are an important design target because they play a key role in many chemical and biological processes ranging from natural light harvesting[1−3]and light-emitting technologies[4] to photocatalysis.[5,6] Due to the delicate interplay[7,8] required to tune complex properties, it is challenging to use a standard Edisonian approach[9] to simultaneously alter metal−ligand interactions, ligand field strength, electron-donating/withdrawing effects, and the relative energetic positioning between the ground- and excited-state potential energy surfaces. Therefore, computation has been used to facilitate the design of transition-metal chromophores. One example comes from the work of Dixon and co-workers,[10,11] where they identified a single Fe(II) complex as a potential luminophore among seven compounds, a prediction which was recently verified by an experimental study.[12] A notable exception to the small-scale, Edisonian approach is a recent work[13] that utilized high-throughput experiments to identify heteroleptic Ir(III)-based chromophores. Nevertheless, to facilitate scalable material design, transition-metal chromophores made with earth-abundant 3d metals with d^6 electron configurations are preferred relative to their state-of-the-art 4d and 5d metal (e.g., Ru(II) and Ir(III)) analogs.[7,8]

The combination of virtual high-throughput screening (VHTS)[14−22] and machine learning (ML)[23−29] shows great promise and has started to address combinatorial challenges in accelerating the design of functional molecules and materials. In this approach, a large set of materials or molecules are studied with density functional theory (DFT) to develop structure−property relationships.[20,30−36] Then, either supervised learning (i.e., forward) models[23−26,37] are trained to screen a large preconstructed design space or generative (i.e., inverse) models[38,39] are applied to obtain candidate molecules

Received: October 3, 2022
Revised: November 15, 2022
Accepted: November 16, 2022
Published: December 1, 2022

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

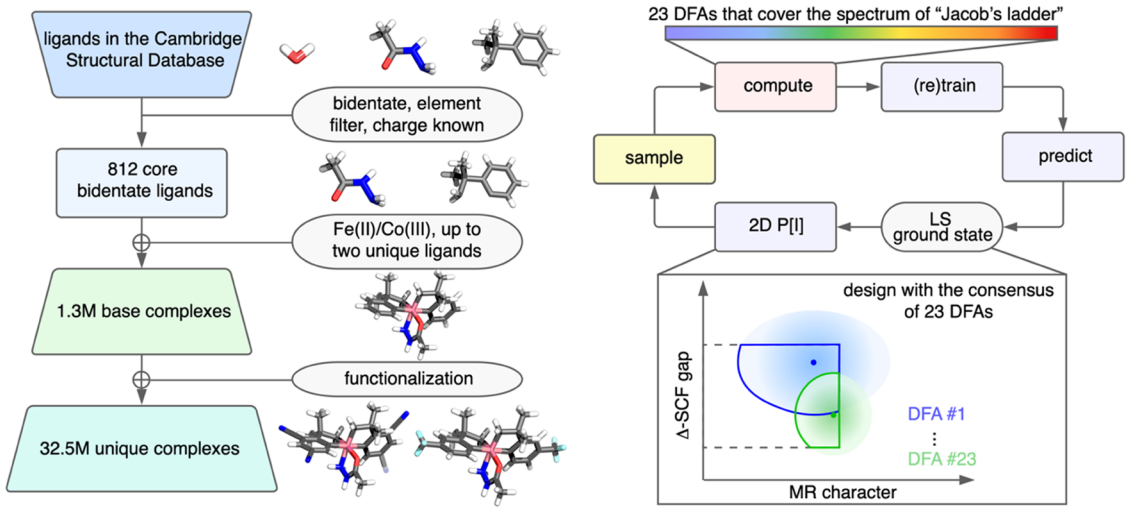


Figure 1. (Left) Hierarchical assembly of the 32.5M complex design space of transition-metal chromophores. All ligands in the CSD are first filtered to retain bidentate ligands with common elements, ≤25 heavy atoms, and known charge. The resulting 812 bidentate ligands are paired with either Fe(II) or Co(III), under the constraint that each complex has ≤2 unique ligand types, to form a design space of 1.3M base complexes. Lastly, these complexes are expanded to the full design space of 32.5M complexes with functionalization on the coordinating rings with a series of electron-donating or electron-withdrawing functional groups. (Right) Active learning for discovering DFA consensus-designed transition-metal chromophores. DFT simulations are performed with 23 DFAs that span multiple rungs of "Jacob's ladder", which are used to train independent ML models. These ML models are applied to predict the ground spin state and Δ-SCF gap of complexes for 23 DFAs, the MR character, and their corresponding uncertainties. These quantities are used to select complexes with low-spin (LS) ground states and to evaluate the 2D $P[I]$ of the design space to sample candidate complexes to compute in the next generation. The inset is an illustration of the ML prediction (solid dot), uncertainty (shaded area), and effective 2D $P[I]$ area (solid outline) for multiple DFAs (blue and green) with respect to a target zone (rectangle with dashed lines).

with targeted properties. A single-shot train-then-predict approach usually requires too much computational time for data generation and can be sensitive to how the compounds are selected for training. Therefore, active learning with Bayesian optimization[40−42] has been recognized as an attractive paradigm for balancing data acquisition in ML model training (i.e., exploration) and ML model-based prediction (i.e., exploitation) for chemical discovery,[43−46] demonstrating a 500-fold acceleration[47] compared to random search.

Despite the success of this active learning approach in many applications, there remain significant challenges that prevent experimental realization of the predictions yielded by computational workflows. First, the outcome depends on the density functional approximation (DFA) choice. A DFA that works well on certain systems may fail prominently on other systems due to the approximations made in the exchange−correlation functional.[17,48,49] When a single-DFA approach is used in VHTS, the DFA choice can lead to large biases in the data sets generated, which in turn biases the candidates the ML models recommend.[50] For transition-metal complexes (TMCs), in particular, the electronic structure is sometimes dominated by static correlation[51] that would make DFT error-prone, and predictions can be highly sensitive to DFA choice. Additionally, it is difficult to guarantee that the predicted lead molecules are synthesizable, despite the ability to add explicit constraints to ML models.[52−54] For TMCs, the synthesizability problem becomes multiplicative[55] (i.e., all ligands comprising a TMC need to be synthesizable and compatible with complex formation).

In this work, we apply an active learning approach to discover 3d^6 Fe(II)/Co(III) transition-metal chromophores in a design space with 32.5M TMCs. Specifically, we use efficient global optimization,[56] which takes the expected or probability of improvement as the criteria to determine the next points to sample in active learning (Figure 1). We address the outstanding challenge of synthesizability of candidate chromophores by carefully crafting the design space with constraints using synthetically accessible fragments and ligand symmetries in the Cambridge Structural Database (CSD). We avoid bias from DFA choice by applying a DFA consensus approach[50] that considers property evaluation as an ensemble of predictions from 23 DFAs that span multiple rungs of "Jacob's ladder".[57] Our active learning approach successfully identifies promising transition-metal chromophores and is estimated to exhibit a 1000-fold acceleration compared to random sampling. We reveal that Co(III) complexes with large, strong-field ligands with more saturated bonds are preferred as candidate transition-metal chromophores. By introducing electron-donating or electron-withdrawing functionalization on compounds and invoking electronic fine tuning (i.e., Hammett tuning effects[58]), we further enrich the number of potential transition-metal chromophores and verify our most promising candidates with time-dependent DFT (TDDFT) calculations.

RESULTS AND DISCUSSION

Design Space

We construct and explore a hypothetical design space of TMCs where all of the constituent fragments (i.e., metal ions and ligands) are synthetically accessible (Figure 1). We further constrain the TMCs in the space to contain three bidentate ligands (e.g., Fe(II)(bpy)$_3$) and restrict ourselves to d^6 Fe(II) or Co(III) metal centers based on the precedent of these metal/oxidation state combinations forming octahedral geometries that make support efficient chromophores. We limit the number of unique ligands in a complex to two to promote the likelihood of synthesizability. We started with 5173 CSD

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

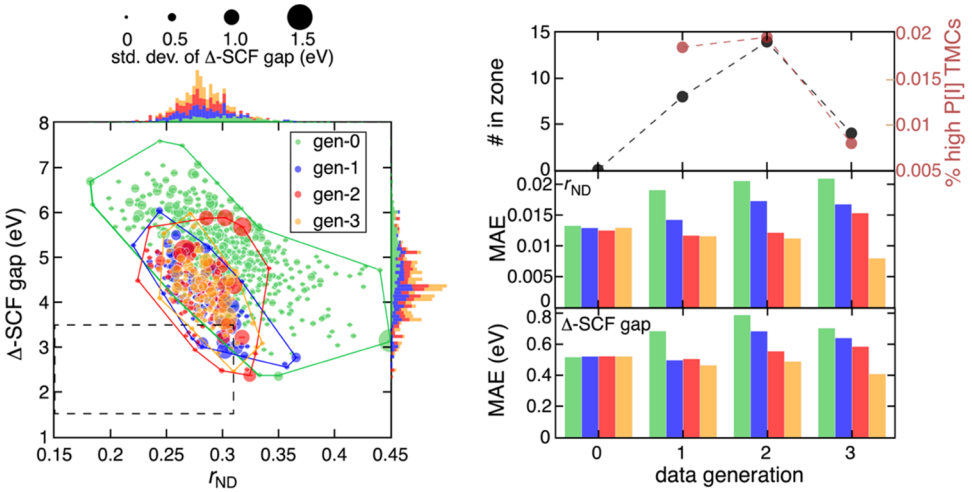


Figure 2. (Left) DFT-computed r_{ND} vs Δ-SCF gap for base complexes in gen-0 to gen-3. For each complex, the average Δ-SCF gap over all DFAs is shown as a circle sized by the corresponding standard deviation (std. dev.) over all DFAs. The range of values sampled in each generation is indicated by a convex hull. The target zone is shown as a rectangle with dashed lines. Normalized stacked marginal histograms for Δ-SCF gap and r_{ND} are also shown. (Right) The number of complexes in the target zone (black) and the percentage of the design space that has a 2D $P[I] > 1/3$ (brown) at each generation (top). The 2D $P[I]$ at gen-0 is not available as the ML models have only been trained after gen-0. MAE for r_{ND} (middle) and Δ-SCF gap (bottom) at all combinations of model generations (indicated by color) and data generation (indicated by the number on the x axis). At each generation, the ML models are trained on the combined training set of all previous generations and are tested on the set-aside test set of each generation separately. For example, the gen-2 model (blue bars) was trained on the combined training set of gen-0, gen-1, and gen-2 data. Generations are colored as follows throughout: green for gen-0, blue for gen-1, red for gen-2, and orange for gen-3. Gen-0 represents a k-medoid sampling of the 1.3M base TMC space.

ligands that we curated in previous work,[59] which included assigning the charges for these ligands. From this set, we selected bidentate ligands that contain common elements (i.e., H, B, C, N, O, F, Si, P, S, Cl, Br, and I) and ≤25 heavy (i.e., non-hydrogen) atoms, leaving a set of 812 ligands (Supporting Information). Combined with either Fe(II) or Co(III), the constraint of forming a complex with up to two unique bidentate ligands in an octahedral complex with three bidentate ligands produces 2 × 812 = 1624 homoleptic and 2 × 812 × 811 = 1,317,064 heteroleptic TMCs. We refer to these 1.3M TMCs as the "base complexes". Hammett tuning is a commonly adopted strategy in experiments to fine-tune the electronic properties of a complex by adding electron-donating or electron-withdrawing functional groups on conjugated rings. Here, we consider 3 distinct functionalization positions and 10 functional groups, expanding the final design space to 32.5M functionalized TMCs, which we refer to as the "functionalized complex space" (Figure 1; see details in the Exploring the Functionalized Design Space section). By adding these constraints to the design space, we expect that the candidate complexes discovered during the active learning process have a higher likelihood to be synthesizable.

Active Learning Procedure and Design Criteria

For most applications of transition-metal chromophores, the photoexcited state should have a lifetime that is sufficiently long, such that the resulting chemical potential can be redirected before it is lost to unproductive competing pathways, with a few exceptions such as photorelease reactions.[60] Correspondingly, it is advantageous to have the photoexcited electron populate a long-lived metal-ligand charge-transfer (MLCT) state and avoid low-lying metal-centered states that deactivate electron transfer from the expected photoexcited state. Therefore, we target complexes with low-spin (LS) ground states to increase the likelihood of MLCT states and to destabilize metal-centered states.[8] We also want target complexes to have weak multireference (MR) character. Avoiding high MR character has the benefit of avoiding complexes for which even a consensus DFT approach is likely to be inaccurate. It is possible to efficiently estimate MR character from fractional occupation number DFT as the contribution from nondynamical correlation[61,62] (i.e., r_{ND};[63] see the Methods section). When this value is low, we also anticipate a lack of deleterious low-lying electronic states. In addition, the absorption energy should fall within the wavelengths of the visible spectrum, ranging from 1.5 eV (825 nm) to 3.5 eV (350 nm). The absorption energy is estimated from Δ-SCF calculations,[64] which are more robust to DFA choice than the highest occupied molecular orbital (HOMO)−lowest unoccupied molecular orbital (LUMO) gap from orbital energies (see the Methods section).

We use efficient global optimization to sample TMCs with LS ground states in a target zone of [1.5, 3.5 eV] for Δ-SCF gap and [0, 0.307] for r_{ND} as candidate transition-metal chromophores (Figure 1; see details in the Methods section). To estimate the overall probability of each TMC residing within the target region, we compute the probability of improvement in the two-dimensional space (i.e., 2D $P[I]$) spanned by the Δ-SCF gap and r_{ND}. The 2D $P[I]$ score is amenable to the design goal of discovering complexes with a range of equally valid Δ-SCF gaps with modest r_{ND} values. For efficient global optimization, we largely followed the established protocols from our previous work[46,47] (Figure 1): (1) complexes in each new generation were selected by k-medoids sampling over the full design space, (2) we then used DFT to evaluate the Δ-SCF gap, r_{ND}, and the ground spin state of these complexes, (3) after combining the new data with data from previous generations, we retrained our ML models, and (4) lastly, we used the updated ML models to evaluate the ground spin state and 2D $P[I]$ of the whole complex space. This information about the design space is then fed into the

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

sampling process (i.e., step 1), closing the active learning loop. Importantly, because both the ground-state assignment and Δ-SCF gap are sensitive to DFA choice, we need to consider the variance of the results from different DFAs (Supporting Information Figures S1−S3). Therefore, we adopted our previous DFA consensus procedure,[50] where we considered an ensemble of 23 DFAs that cover the broad spectrum of "Jacob's ladder" of functionals to increase the robustness of our lead candidate chromophores to DFA choice (see the Methods section).

Active Learning on the 1.3M Base Complexes

We observe a strong negative linear correlation between Δ-SCF gap and r_{ND} for the 2000 TMCs sampled in the initial generation (gen-0), which introduces difficulties for identifying candidates with simultaneously low Δ-SCF gap and r_{ND} (Figure 2). This negative linear correlation exists because a small Δ-SCF gap generally suggests the existence of low-lying excited states, which would lead to high r_{ND} because MR character arises from near-degenerate occupied and virtual orbitals. In addition, we find that LS complexes often have stronger MR character and thus higher r_{ND} relative to their HS counterparts because they can access more configuration state functions[65] (Supporting Information Figure S4). The nature of this multiple-objective search for transition-metal chromophores suggests that TMCs that can fulfill all our design requirements will be scarce. Indeed, for the 2000 TMCs in gen-0, no complex with an LS ground state matches our target criteria with desirably low Δ-SCF gap and r_{ND} (Figure 2). To put this in context, the lack of suitable compounds in gen-0 suggests an extremely low probability, p, of a TMC residing in the target region: when p is as low as 0.030%, there would only be 1/3 chance of finding at least one target complex in 2000 random trials. Our ML models trained on gen-0 data also give a similarly conservative estimate that only 0.018% of TMCs have a 2D $P[I]$ > 1/3, i.e., have a one-third chance of simultaneously fulfilling the two design criteria (Figure 2).

Despite the initial absence of promising transition-metal chromophores, we used active learning to discover lead TMCs in the target zone. The distributions of the sampled points in gen-1 to gen-3 shift toward the target zone due to the identification of compounds that overcome the trade-off of the negative linear correlation between Δ-SCF gap and r_{ND} present in gen-0 (Figure 2). Although only 200 complexes are sampled during each subsequent generation, we discover numerous TMCs in the target zone once their DFT properties are explicitly computed. This enrichment is greatest in gen-2, where we identify 14 new TMCs that fulfill the design criteria, leading to a rather high (7%) lead conversion rate (i.e., number of leads over the number of samples). A conservative estimate using the binomial distribution shows that one would need to sample 200,000 TMCs randomly in the base complex space to produce 14 lead complexes, indicating our active learning approach achieves a 1000-fold acceleration relative to random sampling. In addition, the ML models improve systematically as active learning proceeds from gen-0 to gen-3, as exemplified by the reduction in relative MAEs of predicting the DFT results for the set-aside test data with each new model generation (Figure 2, Supporting Information Figure S5; see the Methods section).

After three generations of active learning, both the number of TMCs landing in the target zone (i.e., after validation with DFT) and the percentage of high 2D $P[I]$ complexes decrease, indicating that most candidate base TMCs have likely been identified, at the same time as ML model performance levels off on the 1.3M base complexes. Therefore, we used the gen-3 models to screen through the base complex design space to reveal chemical trends for the 2432 TMCs that have a reasonable probability of residing in the target zone (i.e., 2D $P[I]$ > 1/6). Here, we use a smaller cutoff for 2D $P[I]$ (i.e., 1/6 compared to 1/3) to retain a reasonable number of complexes for statistical analysis (Supporting Information Figure S6). From this set, we find that complexes with Co(III) and strong-field ligands (e.g., coordinating atom combinations of CC, CN, NP, and PP) are significantly enriched (Figure 3). This likely

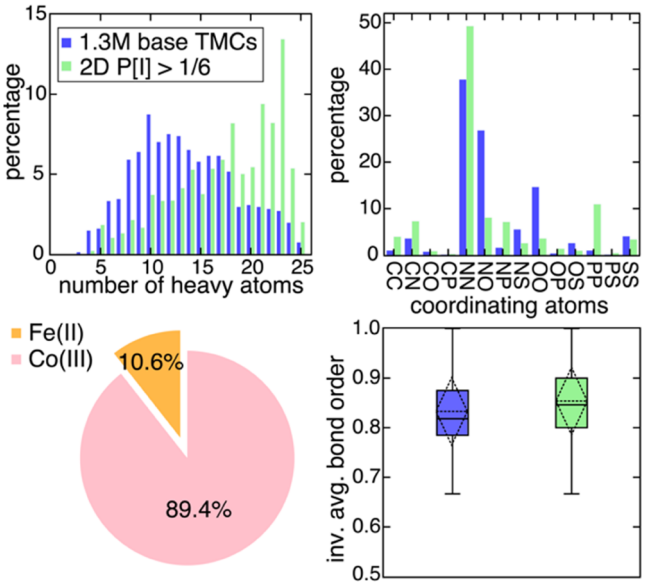


Figure 3. Comparison of property distributions of the 2432 complexes with 2D $P[I]$ > 1/6 evaluated by gen-3 ML models (green) and the 1.3M base complexes (blue). Bar plots for the average number of heavy atoms in the ligands involved in the complexes (top left) and their coordinating atom types (top right), a pie chart for the core metal (orange for Fe(II) and pink for Co(III), bottom left), and a box plot for inverse average (inv. avg.) bond order for the ligands (bottom right). For each box, the median is shown as a horizontal solid line, the mean and std. dev. are shown as a dashed diamond, and the two extrema are shown by the bar.

occurs due to our requirement of an LS ground state during the screening procedure (Figure 1). Because we prefer a small Δ-SCF gap, the complexes that are favored by 2D $P[I]$ tend to have large ligands, consistent with our previous observation that Δ-SCF gap has a negative linear correlation with complex size[65] (Figure 3). Lastly, we find that complexes with reasonable (i.e., >1/6) 2D $P[I]$ tend to consist of ligands that are more saturated, as measured by their increased inverse average bond order[59] (Figure 3). This trend can be understood by the fact that unsaturated ligands tend to contain higher MR character, and complex properties correlate (i.e., are additive) with those of their constituent ligands[59] (Supporting Information Figure S7). In general, we learn from our ML models that a complex with Co(III) and large, strong-field, and relatively saturated ligands would have an increased chance of being a transition-metal chromophore with the desired properties.

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

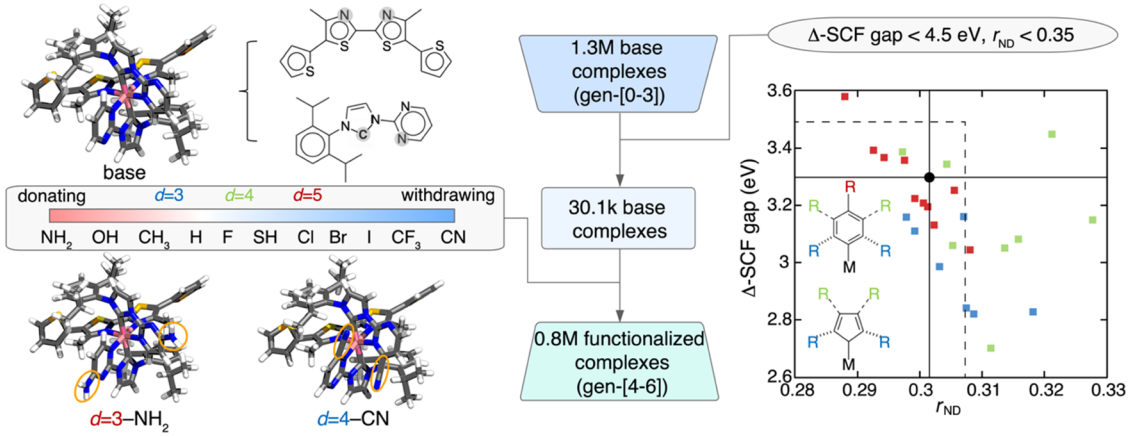


Figure 4. Procedure for constructing the functionalized TMC design space. (Middle) The 1.3M base complexes used in gen-0 to gen-3 are first filtered down to 30.1k base complexes that are predicted to have an LS ground state, an average Δ-SCF gap < 4.5 eV, and r_{ND} < 0.35, based on gen-3 ML models. These complexes are then functionalized on the coordinating rings with a chosen position (i.e., d = 3, 4, or 5) and functional group, enlarging the design space to 0.8M functionalized TMCs to be used in gen-4 to gen-6. (Left) Example of functionalizing the base complex $Co(III)(C_{19}H_{22}N_4)_2(C_{16}H_{12}N_2S_4)$. The base complex and corresponding ligands are shown at the top, where the coordinating atoms are shaded in gray on the skeleton structures. The functional groups used to perform Hammett tuning are shown in the middle. Two functionalized complexes (left with NH_2 at the d = 3 position and right with CN at the d = 4 position) are shown at the bottom. (Right) Average r_{ND} vs Δ-SCF gap for functionalized $Co(III)(C_{19}H_{22}N_4)_2(C_{16}H_{12}N_2S_4)$ at each possible position (blue for d = 3, green for d = 4, and red for d = 5) and functional group. The target zone is shown as a rectangle with dashed lines. The predicted properties of $Co(III)(C_{19}H_{22}N_4)_2(C_{16}H_{12}N_2S_4)$ are shown as a black circle intersected with solid lines. The insets show the functionalization positions for a six-membered and five-membered ring, respectively.

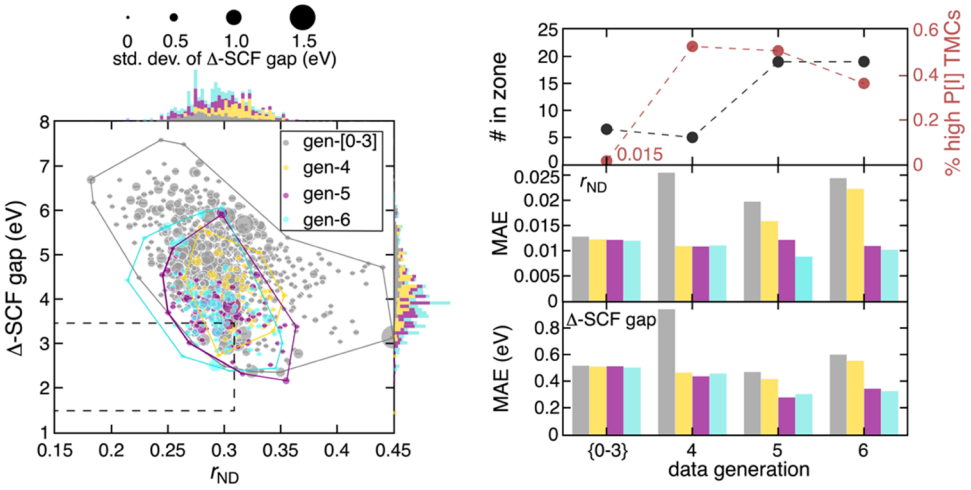


Figure 5. (Left) DFT-computed r_{ND} vs Δ-SCF gap for functionalized complexes from gen-4 to gen-6 (yellow for gen-4, purple for gen-5, and cyan for gen-6). The base complexes in gen-0 to gen-3 are combined as gen-[0−3] (gray). For each complex, the average Δ-SCF gap over all DFAs is shown as a circle scaled by the corresponding std. dev. of Δ-SCF gaps. The range of values sampled in each generation is indicated by a convex hull. The target zone is shown as a rectangle with dashed lines. Normalized stacked marginal histograms for the Δ-SCF gap and r_{ND} are also shown. (Right) The number of complexes in the target zone (black) and the percentage of the design space that has a 2D $P[I]$ > 1/3 (brown) at each generation (top), with the average shown for the combined gen-[0−3]. MAE for r_{ND} (middle) and Δ-SCF gap (bottom) at all combinations of model generations (indicated by color) and data generation (indicated on the x axis). At each generation, the ML models are trained on the combined training set of all previous generations and are tested on the set-aside test set of each generation separately. For the combined gen-[0−3], the MAEs are evaluated on the combined set-aside test sets from gen-0 to gen-3 using the gen-3 ML models. Gen-4 represents a k-medoid sampling of 200 TMCs on the 0.8M-complex functionalized TMC space.

Exploring the Functionalized Design Space

Hammett tuning, i.e., functionalization on conjugated rings, is a common procedure applied in experiments to fine-tune the electronic properties of a TMC without dramatically sacrificing its synthesizability.[66,67] We considered three possible functionalizable positions, categorized by the bond depth (d, i.e., the number of bonds that separate two atoms on the molecular graph) of the H atom on a ring with respect to the metal (Figure 4). For a six-membered ring, d = 3, 4, or 5 corresponds to the *ortho*, *meta*, and *para* positions, respectively. For a five-membered ring, only d = 3 or 4 are feasible. We consider a wide range of 10 electron-donating or electron-withdrawing functional groups (Figure 4). To retain good likelihood of synthesizability, we constrain the in silicofunctionalization procedure to consist of one unique functionalizable position and one unique functional group for a TMC and disallow any combinations with multiple functionalizable positions or functional group identities. Despite this constraint, the base design space is expanded by a factor of 25 after accounting for rings that are not functionalizable, leading to 32.5M TMCs

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

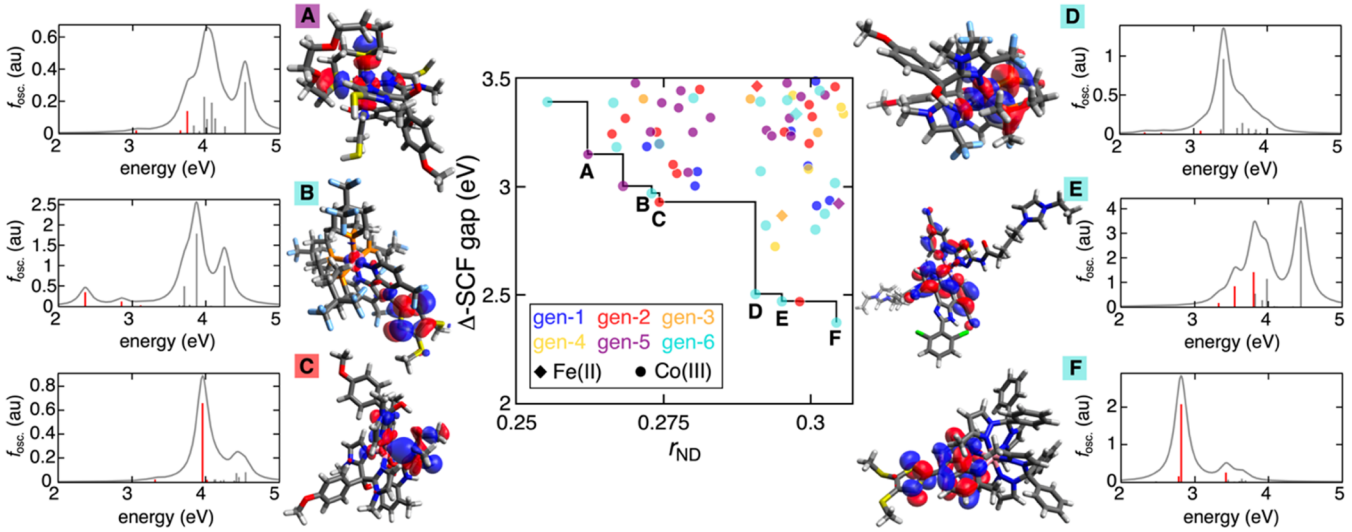


Figure 6. Sixty-nine TMCs sampled through the active learning process that have an LS ground state and land in the target zone computed by DFT, colored by generation and with unique symbols for each metal center (diamond for Fe(II) and circle for Co(III)). The trade-off of best r_{ND} values for a given Δ-SCF gap is indicated by the black solid lines. Six out of nine TMCs exhibiting this optimal trade-off are verified to have desired excited-state properties (i.e., an excited state lower than 3.5 eV and MLCT character) by TDDFT calculations. These six complexes are indicated by the letters **A**−**F**, where the absorption spectra and the orbital for the first bright state are shown. The lowest three absorption energies are colored red for better visibility due to the large variance among the oscillator strengths for different excitations (detailed information and geometries of complexes **A**−**F** are provided in the Supporting Information).

(Figure 1). However, the effect of Hammett tuning is not expected to be large enough to move all of the base complexes into the target zone, so those far enough from the target zone can be immediately discarded. For a representative Co(III) complex that resides in the target zone, we find that Hammett tuning can roughly tune the Δ-SCF gap by 1.0 eV and r_{ND} by 0.04 (Figure 4). Therefore, we use the gen-3 ML models to screen through the 1.3M base complexes and only keep TMCs with a predicted Δ-SCF gap < 4.5 eV and r_{ND} < 0.35 as candidates for Hammett tuning. These promising 30.1k base complexes lead to a design space of 0.8M functionalized TMCs for further exploration. Because our primary goal is to accelerate the discovery of promising transition-metal chromophores rather than identifying all of them in the design space, we expect this stepwise filtering of the design space to be beneficial.

Since we created a new design space with functionalized complexes that our ML models have not seen before, we expected that the 2D $P[I]$ computed based on the trained model predictions and uncertainties would not be able to directly guide the exploration of candidate chromophores. Therefore, we repeated the *k*-medoids sampling that we performed in gen-0 but this time limited to the new 0.8M-complex functionalized design space, selecting a set of 200 complexes. Indeed, we find that the gen-3 models have significantly higher errors in the predictions of the *k*-medoids sampled gen-4 data (Figure 5). However, the ML models improve quickly after retraining on the functionalized complexes in gen-4. Therefore, we expect the 2D $P[I]$ to regain its predictive power and undertook two generations of active learning using the 2D $P[I]$ criterion. During these two generations, the ML models achieve MAEs that are comparable to those on the base complexes (Figure 5). Because we have already isolated a promising fraction of the functionalized TMC space, both the number of TMCs landing in the target zone and the percentage of high (i.e., >1/3) 2D $P[I]$ complexes increased relative to the previous three generations (Figure 5). At both gen-5 and gen-6, 19 (i.e., 10%) of the sampled functionalized TMCs are verified as candidate transition-metal chromophores by DFA consensus. This number greatly surpasses the average (i.e., 6) and the maximum (i.e., 14) in the previous generations. More importantly, the sampled functionalized complexes in gen-5 and gen-6 further expand the convex hull in the 2D property space, with their distributions shifted toward the target zone. We find that functionalization (i.e., Hammett tuning) indeed pushes more complexes into the target zone by fine-tuning their electronic properties. For example, the d = 3 CH_3 functionalization of the base complex Co(III)$(N_2C_{16}H_{12}S_4)$-$(N_4C_{18}H_{16})_2$ (Δ-SCF gap = 2.83 eV, r_{ND} = 0.300) leads to complex **F** (Δ-SCF gap = 2.39 eV, r_{ND} = 0.305), which lowers the Δ-SCF gap with a better r_{ND} value compared to other base complexes sampled in the active learning process (Figure 6 and Supporting Information Table S2). This observation showcases the effectiveness of our strategy for using Hammett tuning to further enrich a pool of candidate chromophores and improve their electronic properties. Because we find the Hammett tuning strategy to be very effective in improving upon the base complexes, it would be of interest to investigate the effect of Hammett tuning on different metal and oxidation state combinations to make earth-abundant metals more promising for light-harvesting applications in future work.

To verify that complexes discovered through the active learning process have the desired excited-state properties of transition-metal chromophores, we computed the excited-state properties with TDDFT of lead chromophores with the lowest r_{ND} for representative Δ-SCF gap values in the target zone from all seven generations (see the Methods section). Using this approach, we verified that six out of nine of our lead complexes have the desired transition energy in the visible region (i.e., <3.5 eV) in their absorption spectra (Figure 6). In addition, the orbitals for the first bright state (i.e., the hole corresponding to the lowest excitation) are all delocalized on the ligands instead of localized on the metal, which suggests

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

that the photoexcitation process involves an MLCT state (Figure 6). This observation is surprising because we did not set design objectives that involve explicit excited-state calculations of our complexes during active learning. Still, we achieve high likelihood (i.e., 67%) of obtaining lead complexes with promising excited-state properties by carefully crafting the ground-state design objectives and using DFA consensus. The remaining three complexes have excited states with energy < 3.5 eV but are "dark" states (i.e., with small oscillator strengths), which could not be foreseen by the ground-state calculations that we performed during the active learning procedure (Supporting Information Table S3). If we had not required a consensus low Δ-SCF gap (i.e., <3.5 eV) and LS ground state, the likelihood of obtaining leads with promising excited-state properties would have been much lower, at around 15% (Supporting Information Table S4). We would like to note that due to the complex electronic structures of transition-metal complexes, the current TDDFT protocol may still not be accurate enough to verify the lead complexes.

Since all of the bidentate ligands considered are synthetically accessible, we propose these complexes as promising candidates for experimental verification. The quality of the leads selected by the consensus approach is expected to be better than from a standard single-functional screen because the functionals from different rungs agree, and previous work has shown that DFT consensus more frequently produces leads similar to those that are experimentally validated.[50] Moreover, although complexes **A**−**F** are not in the CSD, they all contain ligands present in other synthesized compounds that have demonstrated photoinduced properties (Supporting Information Table S5). For example, despite the fact that complex **F** has not been characterized experimentally, one of its constituent ligands (i.e., 4′,5′-diaza-9′-[4,5-bis(methylthio)-1,3-dithiol-2-ylidene]fluorene) has been studied[68] for its interesting nonlinear optical properties in Co, Cu, and Cd complexes (Figure 6). These observations showcase the power of our strategy for identifying experimentally relevant candidate transition-metal chromophores that were potentially missed by previous experimental exploration due to the combinatorial challenges in chemical discovery.

CONCLUSIONS

We applied efficient global optimization with a two-dimensional probability of improvement criterion to discover potential $3d^6$ Fe(II)/Co(III) transition-metal chromophores in a design space of 32.5M compounds that simultaneously fulfill three design objectives: a low-spin ground state, a Δ-SCF gap corresponding to an electronic transition in the visible region of the electromagnetic spectrum, and weak multireference character. We avoid common biases that arise from density functional approximation (DFA) choice in virtual high-throughput screening and machine learning (ML)-accelerated chemical discovery by applying a DFA consensus approach that considers the property evaluations from 23 DFAs that span multiple rungs of "Jacob's ladder". We also addressed the challenge of synthesizability for computationally designed functional molecules by constraining the design space construction to ligands that are synthetically accessible and symmetry classes that are easy to access in experiments. Compounds discovered through this active learning workflow therefore have a higher likelihood to be synthesizable and predicted properties are expected to be of higher fidelity (i.e., more robust to changes in DFA choice).

Despite the scarcity of potential transition-metal chromophores in our design space, judged by the fact that no compounds in the initial 2000 samples landed in our target objective zone, our active learning process gradually shifts the distributions of sampled compounds toward the target zone and successfully identifies many leads. A conservative estimate suggests that our active learning approach achieves a 1000-fold acceleration relative to random sampling. Interrogation of our ML models revealed that Co(III) complexes with large, strong-field, and relatively saturated ligands are preferred as candidate transition-metal chromophores. To fine-tune their electronic properties, we used Hammett tuning by functionalization of the base complexes, which further increased the number of complexes that satisfied the design criteria. Lastly, we performed time-dependent density functional theory calculations on the nine most promising leads. We found that six of the nine compounds demonstrated the desired excited-state properties with metal−ligand charge-transfer states and contain ligands that have been previously studied experimentally due to their interesting optical properties. We expect our strategy for design space construction and our DFA consensus-enhanced active learning workflow to be broadly useful in discovering candidate molecules and materials that are more synthesizable and computationally robust in transition-metal chemical space.

METHODS

DFT Calculation Details

All initial geometries were generated using molSimplify,[69,70] where initial ligand geometries were derived from the crystal structures of transition-metal complexes containing the ligands (Supporting Information). DFT geometry optimizations were carried out using TeraChem,[71] as automated by molSimplify[69,70] with a 24 h wall time per run with up to five resubmissions. These calculations used the B3LYP[72−74] global hybrid functional with the LACVP* basis set, which corresponds to the LANL2DZ[75] effective core potential for transition metals (i.e., Fe, Co) and heavier elements (i.e., I or Br) and the 6-31G* basis for all remaining elements. These geometries were optimized using the L-BFGS algorithm in translation rotation internal coordinates (TRICs)[76] to the default tolerances of 4.5×10^{-4} hartree/bohr for the maximum gradient and 10^{-6} hartree for the energy change between steps. All HS (i.e., quintet) states were calculated with an unrestricted formalism and LS (i.e., singlet) states with a restricted formalism. In all calculations, a level shifting of 0.25 Ha was employed between the occupied and virtual orbitals. Geometry checks were applied to eliminate optimized structures that deviated from the expected octahedral shape following previously established metrics without modification.[77] Open-shell structures were also removed from the data set following established protocols if the expectation value of the S^2 operator deviated from its expected value[77] of $S(S + 1)$ by >1 μ_B^2 (Supporting Information Table S6). Although we did not include dispersion corrections during the geometry optimization, we would like to note that dispersion corrections are helpful for large complexes and will be considered in our future work.

For optimized TMCs, we followed our established protocol[50] for the calculation of the Δ-SCF gap with multiple DFAs using a developer version of Psi4 1.4.[78] We adopted a consistent spin state convention:[50] we removed a majority-spin (i.e., spin-up) electron from the N-electron reference for the $N-1$-electron calculation and added a minority-spin (i.e., spin-down) electron for the $N + 1$-electron case. The Δ-SCF gap is then computed as $2 \times E[N] - (E[N - 1] + E[N + 1])$. In this workflow, the converged wave function obtained from the B3LYP geometry optimization was used as an initial guess for the single-point energy calculations with other DFAs, thus maximizing the correspondence of the converged electronic state among all DFAs and also reducing the computational cost. We use 23

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

DFAs as in our previous work[50] that were chosen to be evenly distributed among the rungs of "Jacob's ladder"[79] (Supporting Information Table S7). While the calculation of gaps from TDDFT might be preferable, challenges associated with implementation for all 23 functionals studied in this work as well as the higher computational cost motivated our focus on Δ-SCF gaps.

We evaluated the r_{ND} diagnostic[61,62,80] by performing finite-temperature DFT[81] calculations using TeraChem.[71] Specifically, we followed a literature recommendation[61,62,80] to use a temperature of 9000 K for B3LYP. Here, we evaluated fractional occupation numbers (FONs) from a broadened distribution (i.e., with Fermi−Dirac statistics).

We performed linear-response TDDFT calculations with the Tamm−Dancoff approximation using ωB97X-D/def2-TZVP, a method and basis set combination chosen based on the recommendation of a recent benchmark study,[82] in Psi4 1.4.[78] We used a polarizable continuum implicit solvent model with water (ε = 80) as the solvent. Because we focus most on the lowest few excitations, only the first 30 states were computed. We broadened simulated spectra using Lorentzian functions, and we considered only excited states with significant oscillator strength (i.e., f_{osc} > 0.01 au).

Active Learning Details

We use efficient global optimization with a 2D probability of improvement ($P[I]$) criterion to sample TMCs with LS ground states in a target zone of [1.5, 3.5 eV] for Δ-SCF gap and [0, 0.307] for r_{ND} as candidate transition-metal chromophores (Figure 1). The 2D $P[I]$ is used to estimate the overall model probability (i.e., total area from the prediction and its model uncertainty) of residing within the target region. At odds with multiobjective optimization where the goal is to minimize or maximize quantities, the 2D $P[I]$ score employed here is amenable to the design goal of discovering complexes with a range of equally valid Δ-SCF gaps with modest r_{ND} values. The cutoff of 0.307 for r_{ND} was chosen based on our previous work[65] as a distinguishing cutoff for TMCs with weak vs strong MR character.

For efficient global optimization, we largely followed the established protocols from our previous work:[46,47] complexes in each new generation were selected by *k*-medoid sampling over the full design space. We then used DFT to evaluate the Δ-SCF gap, r_{ND}, and ground spin state of these complexes. After combining the new data with data from previous generations, we retrained our ML models. Lastly, we used the updated ML models to evaluate the ground spin state and 2D $P[I]$ of the complexes. We selected 2000 TMCs with the highest 2D $P[I]$ and performed *k*-medoid sampling to obtain the 200 complexes from the medoids as candidates for DFT simulation in the next generation. Importantly, because both the ground-state assignment and Δ-SCF gap are sensitive to DFA choice, we adopted a DFA consensus procedure,[50] where we considered an ensemble of 23 DFAs that cover the broad spectrum of "Jacob's ladder" of functionals to increase the robustness of our lead candidate chromophores (Supporting Information Figures S1−S3). Specifically, we only retained complexes when a majority of DFAs (i.e., 70% or >16 of 23) predict the complex to have an LS ground state (Supporting Information Table S1). During the evaluation of 2D $P[I]$, we consider 23 ML models separately trained on each DFA, from which the Δ-SCF gap and its corresponding model uncertainty (i.e., from a calibrated distance in latent space[83]) are estimated. The r_{ND} is computed only from an ML model trained on a single DFA (i.e., B3LYP) because trends in r_{ND} values have been shown to be insensitive to the functional once calibrated.[80] On the other hand, the Δ-SCF gap is averaged over the 23 models due to its relatively high DFA sensitivity (Supporting Information Figure S3). The resulting 23 2D $P[I]$ values derived from the r_{ND} and Δ-SCF gap values are averaged to rank and sample TMCs in the next generation.

ML Models

As in our prior work, we use extended revised autocorrelations[84,85] (eRACs) as descriptors for all our machine learning models. The eRAC features are sums of products and differences of six atom-wise heuristic properties (i.e., topology, identity, electronegativity, covalent radius, nuclear charge, and group number in the periodic table) on the 2D molecular graph. As motivated previously on large TMCs,[46] we truncated eRACs at the maximum bond depth of four to ignore direct interactions of any pairs of atoms that are >4 bonds away. We also eliminated RACs that were invariant (i.e., standard deviation of zero) over the mononuclear octahedral transition-metal complexes. We used metal oxidation state and total ligand charge of a complex as two additional features. Because we would like to discover transition-metal chromophores with an LS ground state with certain ranges of Δ-SCF gap and r_{ND}, we built ML models to predict these three properties. Specifically, we built (i) a classification model to predict whether a complex fulfills the consensus LS condition (i.e., >16 DFAs categorize the ground state to be LS), (ii) a regression model to predict r_{ND} of an LS complex, and (iii) 23 separate models to predict the Δ-SCF gap of an LS complex for each DFA (Supporting Information Table S1). In our workflow, we first used the ground-state classification model to filter out complexes that do not satisfy the consensus LS condition. We used the energy from both the HS and LS optimizations of a complex as training data for the model to determine its ground spin state. On the contrary, only the LS calculation was used for building the ML models that predict Δ-SCF and r_{ND}. For the 23 Δ-SCF gap models, we adopted our established workflow[50] to fine-tune the 22 non-B3LYP models initialized by the weights of the B3LYP model to avoid randomness in the weight initialization and to increase the consistency between ANN models trained with DFT data derived from different DFAs.

During each generation of the active learning, we partitioned the data using a random 80/20% train/test split and used 20% of the training data (i.e., 16% overall) as the validation set. As in our prior work,[46] all ANN models were trained using Keras[86] with a Tensorflow[87] backend and Hyperopt[88] for hyperparameter selection for gen-0 data (Supporting Information Table S8). For all other generations, the models were only fine-tuned with a reduced learning rate (i.e., 10^{-5}) on the combined training set of all previous generations. All ANN models were trained with the Adam optimizer up to 2000 epochs, and dropout, batch normalization, and early stopping were applied to avoid overfitting.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.2c00547.

Data .csv files for all DFT-computed complexes, base complexes with high 2D $P[I]$, and complexes that reside in the target zone; geometries for 812 ligands and DFT-optimized complexes; ML models for Δ-SCF gap and r_{ND} regression and ground spin state classification (ZIP)

Histogram for average ground-state spin; Δ-SCF gap computed at different DFAs and spin states; r_{ND} of optimized structures at LS and HS states; ground-state labeling with DFA consensus and AUC of ML classification models; comparison of ligands' r_{ND}; summary of the first three excited states for lead complexes that have "dark" states; Δ-SCF gap and r_{ND} for the functionalized counterpart of complex **F**; ligands involved in CSD complexes with photoinduced properties; summary of the filtering statistics during active learning; summary of 23 DFAs; and range of hyperparameters sampled for ANN models (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Heather J. Kulik − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Chemistry, Massachusetts Institute of Technology, Cambridge,*

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401

Massachusetts 02139, United States; orcid.org/0000-0001-9342-0191; Email: hjkulik@mit.edu

Authors

Chenru Duan − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-2592-4237

Aditya Nandy − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-7137-5449

Gianmarco G. Terrones − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0001-5360-165X

David W. Kastner − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States; Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-7766-4249

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.2c00547

Author Contributions

CRediT: **Chenru Duan** conceptualization, data curation, investigation, methodology, software, validation, visualization, writing-original draft, writing-review & editing; **Aditya Nandy** data curation, visualization, writing-review & editing; **Gianmarco G. Terrones** data curation, visualization, writing-review & editing; **David W. Kastner** data curation, visualization, writing-review & editing; **Heather J. Kulik** conceptualization, funding acquisition, investigation, project administration, supervision, writing-review & editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing, Office of Basic Energy Sciences, via the Scientific Discovery through Advanced Computing (SciDAC) program as well as by the Office of Naval Research under grant numbers N00014-18-1-2434 and N00014-20-1-2150. C.D. was partially supported by a seed fellowship from the Molecular Sciences Software Institute under NSF grant OAC-1547580. A.N. and D.W.K were partially supported by a National Science Foundation Graduate Research Fellowship under Grants #1122374 and Grant #1745302, respectively. H.J.K. holds a Sloan Fellowship in Chemistry, which supported this work. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing HPC resources that have contributed to the research results reported within this paper. The authors thank Adam H. Steeves for providing a critical reading of the manuscript.

REFERENCES

(1) Wegeberg, C.; Wenger, O. S. Luminescent First-Row Transition Metal Complexes. *JACS Au* **2021**, *1*, 1860−1876.
(2) Wenger, O. S. Photoactive Complexes with Earth-Abundant Metals. *J. Am. Chem. Soc.* **2018**, *140*, 13522−13533.
(3) Ponseca, C. S.; Chábera, P.; Uhlig, J.; Persson, P.; Sundström, V. Ultrafast Electron Dynamics in Solar Energy Conversion. *Chem. Rev.* **2017**, *117*, 10940−11024.
(4) Yam, V. W.-W.; Au, V. K.-M.; Leung, S. Y.-L. Light-Emitting Self-Assembled Materials Based on D8 and D10 Transition Metal Complexes. *Chem. Rev.* **2015**, *115*, 7589−7728.
(5) Ardo, S.; Meyer, G. J. Photodriven Heterogeneous Charge Transfer with Transition-Metal Compounds Anchored to Tio2 Semiconductor Surfaces. *Chem. Soc. Rev.* **2009**, *38*, 115−164.
(6) Proppe, A. H.; Li, Y. C.; Aspuru-Guzik, A.; Berlinguette, C. P.; Chang, C. J.; Cogdell, R.; Doyle, A. G.; Flick, J.; Gabor, N. M.; van Grondelle, R.; Hammes-Schiffer, S.; Jaffer, S. A.; Kelley, S. O.; Leclerc, M.; Leo, K.; Mallouk, T. E.; Narang, P.; Schlau-Cohen, G. S.; Scholes, G. D.; Vojvodic, A.; Yam, V. W.-W.; Yang, J. Y.; Sargent, E. H. Bioinspiration in Light Harvesting and Catalysis. *Nat. Rev. Mater.* **2020**, *5*, 828−846.
(7) Kjær, K. S.; Kaul, N.; Prakash, O.; Chabera, P.; Rosemann, N. W.; Honarfar, A.; Gordivska, O.; Fredin, L. A.; Bergquist, K. E.; Haggstrom, L.; Ericsson, T.; Lindh, L.; Yartsev, A.; Styring, S.; Huang, P.; Uhlig, J.; Bendix, J.; Strand, D.; Sundstrom, V.; Persson, P.; Lomoth, R.; Warnmark, K. Luminescence and Reactivity of a Charge-Transfer Excited Iron Complex with Nanosecond Lifetime. *Science* **2019**, *363*, 249−253.
(8) McCusker, J. K. Electronic Structure in the Transition Metal Block and Its Implications for Light Harvesting. *Science* **2019**, *363*, 484−488.
(9) Braun, J. D.; Lozada, I. B.; Kolodziej, C.; Burda, C.; Newman, K. M. E.; van Lierop, J.; Davis, R. L.; Herbert, D. E. Iron(Ii) Coordination Complexes with Panchromatic Absorption and Nanosecond Charge-Transfer Excited State Lifetimes. *Nat. Chem.* **2019**, *11*, 1144−1150.
(10) Dixon, I. M.; Alary, F.; Boggio-Pasqua, M.; Heully, J.-L. Reversing the Relative 3mlct−3mc Order in Fe(Ii) Complexes Using Cyclometallating Ligands: A Computational Study Aiming at Luminescent Fe(Ii) Complexes. *Dalton Trans.* **2015**, *44*, 13498−13503.
(11) Dixon, I. M.; Khan, S.; Alary, F.; Boggio-Pasqua, M.; Heully, J. L. Probing the Photophysical Capability of Mono and Bis-(Cyclometallated) Fe(Ii) Polypyridine Complexes Using Inexpensive Ground State Dft. *Dalton Trans.* **2014**, *43*, 15898−15905.
(12) Leis, W.; Argüello Cordero, M. A.; Lochbrunner, S.; Schubert, H.; Berkefeld, A. A Photoreactive Iron(Ii) Complex Luminophore. *J. Am. Chem. Soc.* **2022**, *144*, 1169−1173.
(13) DiLuzio, S.; Mdluli, V.; Connell, T. U.; Lewis, J.; VanBenschoten, V.; Bernhard, S. High-Throughput Screening and Automated Data-Driven Analysis of the Triplet Photophysical Properties of Structurally Diverse, Heteroleptic Iridium(Iii) Complexes. *J. Am. Chem. Soc.* **2021**, *143*, 1179−1194.
(14) Shu, Y. N.; Levine, B. G. Simulated Evolution of Fluorophores for Light Emitting Diodes. *J. Chem. Phys.* **2015**, *142*, No. 104104.
(15) Gómez-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.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W. L.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120−1127.
(16) 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.
(17) 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.* **2019**, *119*, 2453−2523.
(18) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354−2377.
(19) 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.
(20) 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.
(21) Nørskov, J. K.; Bligaard, T. The Catalyst Genome. *Angew. Chem., Int. Ed.* **2013**, *52*, 776−777.
(22) Baiardi, A.; Grimmel, S. A.; Steiner, M.; Türtscher, P. L.; Unsleber, J. P.; Weymuth, T.; Reiher, M. Expansive Quantum Mechanical Exploration of Chemical Reaction Paths. *Acc. Chem. Res.* **2022**, *55*, 35−43.
(23) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B* **2014**, *89*, No. 094104.
(24) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336−2347.
(25) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.
(26) Chen, A.; Zhang, X.; Zhou, Z. Machine Learning: Accelerating Materials Development for Energy Storage and Conversion. *InfoMat* **2020**, *2*, 553−576.
(27) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, No. 143001.
(28) Sauceda, H. E.; Gálvez-González, L. E.; Chmiela, S.; Paz-Borbón, L. O.; Müller, K.-R.; Tkatchenko, A. Bigdml—Towards Accurate Quantum Machine Learning Force Fields for Materials. *Nat. Commun.* **2022**, *13*, No. 3733.
(29) Hermann, J.; Schätzle, Z.; Noé, F. Deep-Neural-Network Solution of the Electronic Schrödinger Equation. *Nat. Chem.* **2020**, *12*, 891−897.
(30) Smith, J. S.; Isayev, O.; Roitberg, A. E. Ani-1, a Data Set of 20 Million Calculated Off-Equilibrium Conformations for Organic Molecules. *Sci. Data* **2017**, *4*, No. 170193.
(31) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, No. 011002.
(32) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. Aflow: An Automatic Framework for High-Throughput Materials Discovery. *Comput. Mater. Sci.* **2012**, *58*, 218−226.
(33) Pizzi, G.; Cepellotti, A.; Sabatini, R.; Marzari, N.; Kozinsky, B. Aiida: Automated Interactive Infrastructure and Database for Computational Science. *Comput. Mater. Sci.* **2016**, *111*, 218−230.
(34) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; Palizhati, A.; Sriram, A.; Wood, B.; Yoon, J.; Parikh, D.; Zitnick, C. L.; Ulissi, Z. Open Catalyst 2020 (Oc20) Dataset and Community Challenges. *ACS. Catal.* **2021**, *11*, 6059−6072.
(35) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J.; Zemel, R.; Zhang, S.AlChemy: A Quantum Chemistry Dataset for Benchmarking AI Models. 2019, arXiv:1906.09427. arXiv.org e-Print archive. https://doi.org/10.48550/arXiv.1906.09427.
(36) Jia, H.; Nandy, A.; Liu, M.; Kulik, H. J. Modeling the Roles of Rigidity and Dopants in Single-Atom Methane-to-Methanol Catalysts. *J. Mater. Chem. A* **2022**, *10*, 6193−6203.
(37) Janet, J. P.; Duan, C.; Nandy, A.; Liu, F.; Kulik, H. J. Navigating Transition-Metal Chemical Space: Artificial Intelligence for First-Principles Design. *Acc. Chem. Res.* **2021**, *54*, 532−545.
(38) Walters, W. P.; Murcko, M. Assessing the Impact of Generative Ai on Medicinal Chemistry. *Nat. Biotechnol.* **2020**, *38*, 143−145.
(39) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernandez-Lobato, J. M.; Sanchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(40) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; Freitas, N. Taking the Human out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148−175.
(41) Snoek, J.; Larochelle, H.; Adams, R. P.Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*; Pereira, F.; Burges, C. J.; Bottou, L.; Weinberger, K. Q., Eds.; Curran Associates, Inc., 2012; Vol. *25*, pp 2951−2959.
(42) Bradford, E.; Schweidtmann, A. M.; Lapkin, A. Efficient Multiobjective Optimization Employing Gaussian Processes, Spectral Sampling and a Genetic Algorithm. *J. Global Optim.* **2018**, *71*, 407−438.
(43) Tallorin, L.; Wang, J.; Kim, W. E.; Sahu, S.; Kosa, N. M.; Yang, P.; Thompson, M.; Gilson, M. K.; Frazier, P. I.; Burkart, M. D.; Gianneschi, N. C. Discovering De Novo Peptide Substrates for Enzymes Using Machine Learning. *Nat. Commun.* **2018**, *9*, No. 5253.
(44) Herbol, H. C.; Hu, W.; Frazier, P.; Clancy, P.; Poloczek, M. Efficient Search of Compositional Space for Hybrid Organic−Inorganic Perovskites Via Bayesian Optimization. *npj Comput. Mater.* **2018**, *4*, No. 51.
(45) Yuan, R.; Liu, Z.; Balachandran, P. V.; Xue, D.; Zhou, Y.; Ding, X.; Sun, J.; Xue, D.; Lookman, T. Accelerated Discovery of Large Electrostrains in Batio3-Based Piezoelectrics Using Active Learning. *Adv. Mater.* **2018**, *30*, No. 1702884.
(46) Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. New Strategies for Direct Methane-to-Methanol Conversion from Active Learning Exploration of 16 Million Catalysts. *JACS Au* **2022**, *2*, 1200−1213.
(47) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513−524.
(48) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289−320.
(49) Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315−2372.
(50) Duan, C.; Chen, S.; Taylor, M. G.; Liu, F.; Kulik, H. J. Machine Learning to Tame Divergent Density Functional Approximations: A New Path to Consensus Materials Design Principles. *Chem. Sci.* **2021**, *12*, 13021−13036.
(51) Janesko, B. G. Replacing Hybrid Density Functional Theory: Motivation and Recent Advances. *Chem. Soc. Rev.* **2021**, *50*, 8470−8495.
(52) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part Ii: Outlook. *Angew. Chem. Int. Ed.* **2020**, *59*, 23414−23436.
(53) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020**, *60*, 5714−5723.
(54) Thakkar, A.; Chadimova, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J. L. Retrosynthetic Accessibility Score (Rascore) - Rapid Machine Learned Synthesizability Classification from Ai Driven Retrosynthetic Planning. *Chem. Sci.* **2021**, *12*, 3339−3349.

(55) Duan, C.; Nandy, A.; Kulik, H. J. Machine Learning for the Discovery, Design, and Engineering of Materials. *Annu. Rev. Chem. Biomol. Eng.* **2022**, *13*, 405−429.
(56) Jones, D. R.; Schonlau, M.; Welch, W. J. Efficient Global Optimization of Expensive Black-Box Functions. *J. Global Optim.* **1998**, *13*, 455−492.
(57) Perdew, J. P.; Schmidt, K. In *"Jacob's Ladder of Density Functional Approximations for the Exchange-Correlation Energy" in Density Functional Theory and Its Application to Materials*; VanDoren, V.; VanAlsenoy, C.; Geerlings, P., Eds.; American Institute of Physics: Melville, 2001; Vol. *577*.
(58) Nam, W. Synthetic Mononuclear Nonheme Iron−Oxygen Intermediates. *Acc. Chem. Res.* **2015**, *48*, 2415−2423.
(59) Duan, C.; Ladera, A. J.; Liu, J. C. L.; Taylor, M. G.; Ariyarathna, I. R.; Kulik, H. J. Exploiting Ligand Additivity for Transferable Machine Learning of Multireference Character across Known Transition Metal Complex Ligands. *J. Chem. Theory Comput.* **2022**, *18*, 4836−4845.
(60) Vuilleumier, J.; Gaulier, G.; De Matos, R.; Ortiz, D.; Menin, L.; Campargue, G.; Mas, C.; Constant, S.; Le Dantec, R.; Mugnier, Y.; Bonacina, L.; Gerber-Lemaire, S. Two-Photon-Triggered Photorelease of Caged Compounds from Multifunctional Harmonic Nanoparticles. *ACS Appl. Mater. Interfaces* **2019**, *11*, 27443−27452.
(61) Ramos-Cordoba, E.; Salvador, P.; Matito, E. Separation of Dynamic and Nondynamic Correlation. *Phys. Chem. Chem. Phys.* **2016**, *18*, 24015−24023.
(62) Ramos-Cordoba, E.; Matito, E. Local Descriptors of Dynamic and Nondynamic Correlation. *J. Chem. Theory Comput.* **2017**, *13*, 2705−2711.
(63) Kesharwani, M. K.; Sylvetsky, N.; Kohn, A.; Tew, D. P.; Martin, J. M. L. Do Ccsd and Approximate Ccsd-F12 Variants Converge to the Same Basis Set Limits? The Case of Atomization Energies. *J. Chem. Phys.* **2018**, *149*, No. 154109.
(64) Ziegler, T.; Rauk, A.; Baerends, E. J. On the Calculation of Multiplet Energies by the Hartree-Fock-Slater Method. *Theor. Chim. Acta* **1977**, *43*, 261−271.
(65) Liu, F.; Duan, C.; Kulik, H. J. Rapid Detection of Strong Correlation with Machine Learning for Transition-Metal Complex High-Throughput Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 8067−8076.
(66) Fatur, S. M.; Shepard, S. G.; Higgins, R. F.; Shores, M. P.; Damrauer, N. H. A Synthetically Tunable System to Control Mlct Excited-State Lifetimes and Spin States in Iron(Ii) Polypyridines. *J. Am. Chem. Soc.* **2017**, *139*, 4493−4505.
(67) Mukherjee, S.; Torres, D. E.; Jakubikova, E. Homo Inversion as a Strategy for Improving the Light-Absorption Properties of Fe(Ii) Chromophores. *Chem. Sci.* **2017**, *8*, 8115−8126.
(68) Zhu, Q.-Y.; Lu, W.; Zhang, Y.; Bian, G.-Q.; Gu, J.; Lin, X.-M.; Dai, J. Syntheses, Crystal Structures, and Optical Properties of Metal Complexes with 4′,5′-Diaza-9′-(4,5-Disubstituted-1,3-Dithiol-2-Ylidene)Fluorene Ligands. *Eur. J. Inorg. Chem.* **2008**, *2008*, 230−238.
(69) 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.
(70) KulikGroup. Molsimplify Documentation, 2020. http://molsimplify.mit.edu (accessed June 24, 2021).
(71) Seritan, S.; Bannwarth, C.; Fales, B. S.; Hohenstein, E. G.; Isborn, C. M.; Kokkila-Schumacher, S. I. L.; Li, X.; Liu, F.; Luehr, N.; Snyder, J. W., Jr.; Song, C.; Titov, A. V.; Ufimtsev, I. S.; Wang, L.-P.; Martínez, T. J. Terachem: A Graphical Processing Unit-Accelerated Electronic Structure Package for Large-Scale Ab Initio Molecular Dynamics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021**, *11*, No. e1494.
(72) Becke, A. D. Density-Functional Thermochemistry. Iii. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(73) 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.
(74) 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. A* **1994**, *98*, 11623−11627.
(75) 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.
(76) Wang, L.-P.; Song, C. Geometry Optimization Made Simple with Translation and Rotation Coordinates. *J. Chem. Phys.* **2016**, *144*, No. 214108.
(77) Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J. Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *J. Chem. Theory Comput.* **2019**, *15*, 2331−2345.
(78) Smith, D. G. A.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Di Remigio, R.; Alenaizan, A.; James, A. M.; Lehtola, S.; Misiewicz, J. P.; Scheurer, M.; Shaw, R. A.; Schriber, J. B.; Xie, Y.; Glick, Z. L.; Sirianni, D. A.; O'Brien, J. S.; Waldrop, J. M.; Kumar, A.; Hohenstein, E. G.; Pritchard, B. P.; Brooks, B. R.; Schaefer, H. F., 3rd; Sokolov, A. Y.; Patkowski, K.; DePrince, A. E., 3rd; Bozkaya, U.; King, R. A.; Evangelista, F. A.; Turney, J. M.; Crawford, T. D.; Sherrill, C. D. Psi4 1.4: Open-Source Software for High-Throughput Quantum Chemistry. *J. Chem. Phys.* **2020**, *152*, No. 184108.
(79) Perdew, J. P.; Schmidt, K. *Density Functional Theory and Its Application to Materials*; Melville: NY, 2001; Vol. *577*, pp 1−20.
(80) Grimme, S.; Hansen, A. A Practicable Real-Space Measure and Visualization of Static Electron-Correlation Effects. *Angew. Chem., Int. Ed.* **2015**, *54*, 12308−12313.
(81) Weinert, M.; Davenport, J. W. Fractional Occupations and Density-Functional Energies and Forces. *Phys. Rev. B* **1992**, *45*, 13709−13712.
(82) Liang, J.; Feng, X.; Hait, D.; Head-Gordon, M. Revisiting the Performance of Time-Dependent Density Functional Theory for Electronic Excitations: Assessment of 43 Popular and Recently Developed Functionals from Rungs One to Four. *J. Chem. Theory Comput.* **2022**, *18*, 3460−3473.
(83) Janet, J. P.; Duan, C.; Yang, T. H.; Nandy, A.; Kulik, H. J. A Quantitative Uncertainty Metric Controls Error in Neural Network-Driven Chemical Discovery. *Chem. Sci.* **2019**, *10*, 7913−7922.
(84) 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.
(85) Harper, D. R.; Nandy, A.; Arunachalam, N.; Duan, C.; Janet, J. P.; Kulik, H. J. Representations and Strategies for Transferable Machine Learning Improve Model Performance in Chemical Discovery. *J. Chem. Phys.* **2022**, *156*, No. 074101.
(86) Chollet, F. K. https://keras.io (accessed June 24, 2021).
(87) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jozefowicz, R.; Jia, Y.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Schuster, M.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://www.tensorflow.org (accessed Nov 8, 2022).
(88) Bergstra, J.; Yamins, D.; Cox, D. D., *Proceedings of the 12th Python in Science Conference*, 2013; pp 13−20.

https://doi.org/10.1021/jacsau.2c00547
JACS Au 2023, 3, 391−401