A transferable recommender approach for selecting the best density functional approximations in chemical discovery

Chenru Duan, Aditya Nandy, Ralf Meyer, Naveen Arunachalam & Heather J. Kulik

Approximate density functional theory has become indispensable owing to its balanced cost–accuracy trade-off, including in large-scale screening. To date, however, no density functional approximation (DFA) with universal accuracy has been identified, leading to uncertainty in the quality of data generated from density functional theory. With electron density fitting and Δ-learning, we build a DFA recommender that selects the DFA with the lowest expected error with respect to the gold standard (but cost-prohibitive) coupled cluster theory in a system-specific manner. We demonstrate this recommender approach on the evaluation of vertical spin splitting energies of transition metal complexes. Our recommender predicts top-performing DFAs and yields excellent accuracy (about 2 kcal mol⁻¹) for chemical discovery, outperforming both individual Δ-learning models and the best conventional single-functional approach from a set of 48 DFAs. By demonstrating transferability to diverse synthesized compounds, our recommender potentially addresses the accuracy versus scope dilemma broadly encountered in computational chemistry.

Density functional theory (DFT) has become an indispensable companion to experiments as well as a primary tool in virtual high-throughput screening (VHTS) to generate large-scale computational data sets. Combined with machine learning (ML), these data sets have accelerated computational chemical discovery and revolutionized the process of scientific discovery in the physical and life sciences. All ML models, however, are limited by the quality of the training data, imposing strict requirements on the accuracy of the first-principles methods used in VHTS. In DFT, a density functional approximation (DFA) that works well on certain systems can fail catastrophically on other systems due to the approximations made in the exchange–correlation functional. This DFA dependence is particularly strong in open-shell transition metal chemistry, with many examples of compelling functional materials (for example, metal–organic frameworks) and catalytic reactions (such as C–H activation) that are dominated by static correlation, leading to high variability in the quality of DFT predictions.

Computational chemists face a choice of either carrying out large studies with one functional or careful but small-scale study. While a single DFA is typically selected to screen a large chemical space in VHTS, this approach can bias the generated computational data sets as well as any trained ML models or candidate materials from ML-accelerated discovery. With great care and effort that is incompatible with large-scale screening, researchers may alternately focus on a small set of molecules for which they can obtain experimental or accurate correlated wavefunction theory reference data against which to benchmark a DFA in a system- and property-dependent manner. The recommended DFA may also vary depending on the data set chosen for benchmarking. Not only limiting VHTS, current challenges in the system-specific nature of DFA accuracy also limit mechanistic enquiry by computational chemists aiming to reach ‘chemical accuracy’ in predictive regimes where current experimental reference data may be lacking.

1Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. 2Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. E-mail: hjkulik@mit.edu
To improve the fidelity of DFT-derived data sets, one may develop DFAs with increased accuracy and generalizability. Recent demonstrations using ML to develop neural network potentials\cite{3,4} and exchange-correlation functionals\cite{5,6,7} have shown promise in advancing electronic structure theory. However, they have severe limitations that curtail their practical use. These DFAs have primarily been developed for and applied to closed-shell organic molecules and are still less transferable relative to conventional DFAs developed in the theoretical chemistry community over the past few decades. These models also target the total electronic energy of a configuration of atoms rather than other properties of chemical interest, such as those involving multiple electronic states (for example, spin splitting). Instead of developing a new DFA, we leverage Δ-learning and introduce a regress-then-classify strategy\cite{8,9} to develop a recommender\cite{10} to select the best DFA for a given system and property of interest based on the electron density of the system. To demonstrate this approach, we recommend a DFA that most accurately evaluates the vertical spin splitting energy (ΔΔE_{H–L}) of 2.1 kcal mol\(^{-1}\), providing the accuracy required\cite{22} for the exploration of transition metal chemical space. Because the electron density is a fundamental property of the system, our recommender demonstrates excellent transferability, achieving similar accuracy for recommending DFAs on unseen experimentally synthesized TMCs that contain diverse chemistry. This recommender approach provides a general framework for method selection to improve data quality in VHTS and ML-accelerated discovery in computational sciences.

### Results

#### Overview of the DFA recommender

Electron densities lie at the core of Kohn–Sham (KS) DFT and can be used to derive any ground-state property of interest\cite{21}. However, it is less commonly used as a representation in ML models due to the non-local nature of KS orbitals and its cumbersome-to-fulfil translational and rotational symmetries when discretized to three-dimensional (3D) cubes. Here, we perform density fitting (DF) on the electron density\cite{12,13} of a TMC obtained from a DFT calculation using the global generalized gradient approximation hybrid functional B3LYP, a representative DFA.

The resulting coefficients of the DF basis functions, which preserve the required physical (translational, rotational and permutation) symmetries, are used as the TMC’s representation (Fig. 1a,b and DF procedure section). To make this approach general with respect to chemical properties that involve multiple electronic states (here, vertical spin splitting), we decompose the difference of the electron densities between two electronic states into each atom using a DF procedure. \(\phi_{Q} (r)\) and \(C_{A}^{Q}\) are the Qth DF basis function and coefficient for atom \(A\), respectively. \(p_{A}^{Q}\) is the power spectrum at angular momentum \(l\) for \(C_{A}^{Q}\). The colour (blue, green and grey) and shape of the density are only for demonstration purpose. These coefficients \((p_{A}^{Q})\) are used in a Behler–Parrinello-type neural network as a Δ-learning model to predict \(\Delta \Delta E_{H–L} (f)\) for each DFA \((f)\) in our pool of 48 DFAs. Coefficients of different atoms that are in the same group of the Periodic Table share the same local network and weights and are represented by the corresponding element colour (for example, \(W_{L}\) for O and S sharing the same weights \(W_{S}\)). The latent vectors of each element are finally concatenated and passed to a fully connected network \((W_{L})\) to predict \(\Delta \Delta E_{H–L} (f)\). The 48 predicted values of \(\Delta \Delta E_{H–L} (f)\) are then sorted, and we recommend the DFA that yields the lowest predicted \(\Delta \Delta E_{H–L} (f)\).

---

**Fig. 1** Workflow for the DFA recommender. **a**, B3LYP/def2-TZVP single-point energy calculations are performed on both the HS and LS states to obtain their electron densities at the B3LYP level. **b**, The difference of the electron densities \((\rho (r))\) between the HS and LS states is decomposed into each atom using a DF procedure. \(\phi_{Q} (r)\) and \(C_{A}^{Q}\) are the Qth DF basis function and coefficient for atom \(A\), correspondingly. \(p_{A}^{Q}\) is the power spectrum at angular momentum \(l\) for \(C_{A}^{Q}\). The colour (blue, green and grey) and shape of the density are only for demonstration purpose. **c**, These coefficients \((p_{A}^{Q})\) are used in a Behler–Parrinello-type neural network as a Δ-learning model to predict \(\Delta \Delta E_{H–L} (f)\) for each DFA \((f)\) in our pool of 48 DFAs. Coefficients of different atoms that are in the same group of the Periodic Table share the same local network and weights and are represented by the corresponding element colour (for example, \(W_{L}\) for O and S sharing the same weights \(W_{S}\)). The latent vectors of each element are finally concatenated and passed to a fully connected network \((W_{L})\) to predict \(\Delta \Delta E_{H–L} (f)\). The 48 predicted values of \(\Delta \Delta E_{H–L} (f)\) are then sorted, and we recommend the DFA that yields the lowest predicted \(\Delta \Delta E_{H–L} (f)\).
errors on certain molecules, whereas some ‘worse’ DFAs can give more accurate energetics (Supplementary Fig. 2).

There are two approaches to evaluate the performance of the DFA recommender. The first is the absolute error achieved using the recommended DFA relative to the reference DLPNO-CCSD(T) method. This measure serves as a practical metric for evaluating the accuracy obtained by the recommender in VHTS. The second is the rank ordering of the recommended DFA among the pool of candidate DFAs. This statistical measure quantifies how well the recommender distinguishes top-performing candidate DFAs. Throughout our work, we will consider both.

Performance of Δ-learning models
We consider a set of 452 octahedral TMCs composed of 3d mid-row transition metals and small organic ligands in the spectrochemical series (VSS-452; Supplementary Fig. 2 and Data set construction section). We demonstrate the performance of our Δ-learning models for predicting the differences in $\Delta E_{\text{H-L}}$ obtained by a DFA and DLPNO-CCSD(T) for a pool of 48 DFAs that cover multiple runs of Jacob’s ladder. Additional information about each of the 48 DFAs tested, as well as references, can be found in Supplementary Tables 1 and 2. These 48 Δ-learning models have MAEs (as well as relative errors) ranging from 2.3 to 3.4 kcal mol$^{-1}$, with a median MAE of 2.5 kcal mol$^{-1}$ on the 152 set-aside test TMCs in VSS-452 (Fig. 2a). Here, we emphasize that the converged B3LYP density outperforms other geometry-based representations and the initial guess density in terms of the Δ-learning MAE (Supplementary Table 3). This superior performance is probably not restricted to B3LYP and would apply to other DFAs (for example, the range-separated meta-generalized gradient approximation hybrid, ωB97M-V). This observation encourages other DFAs (for example, the range-separated meta-generalized gradient approximation hybrid, ωB97M-V) and the associated DFA performance is probably not restricted to B3LYP and would apply to future applications of invariant representations of electron density, especially with equivariant model architectures. Considering the small set of diverse TMCs used in training, these Δ-learning MAEs are below the typical (3 kcal mol$^{-1}$) experimental uncertainties for TMC properties (Data set construction section). Although some DFAs have large (>30 kcal mol$^{-1}$) MAEs, all 48 Δ-learning models yield reasonably low MAEs, suggesting the general applicability of Δ-learning.

The ranking of MAEs of the Δ-learning models does not follow the same order as the error ranking of the underlying DFA results relative to DLPNO-CCSD(T). The dispersion-corrected double hybrid DSD-PBEP86-D3BJ has the lowest Δ-learning MAE but only the fifth-lowest DFA-derived MAE, whereas the meta-generalized gradient approximation hybrid M06-L has the highest MAE among 48 Δ-learning models despite ranking only 17th in its DFA-derived MAE (Supplementary Figs. 4 and 5). The rank-order coefficient (Spearman’s r) between the 48 DFAs ranked by Δ-learning model MAE and the rank by DFA-derived MAE is only 0.36, suggesting that Δ-learning performance is not necessarily better when the DFA-derived MAE of the conventional DFA is smaller.

Performance of the DFA recommender
We then utilize the predicted $|\Delta \Delta E_{\text{H-L}}(f)|$ from all 48 Δ-learning models to recommend a ‘best’ DFA for each compound. We compare $\Delta \Delta E_{\text{H-L}}$ obtained by the recommended DFA and the DLPNO-CCSD(T) reference to evaluate the performance of the recommender (see Single-point energy calculation section). The recommender achieves an MAE of 2.1 kcal mol$^{-1}$ on the set-aside test set of VSS-452, outperforming all 48 Δ-learning models. This MAE is only 2.5 times the theoretical lower bound (0.8 kcal mol$^{-1}$) that would be achieved if the lowest-error DFA were always selected (Supplementary Fig. 6). Our recommender outperforms random DFA selection (with an MAE of 13.3 kcal mol$^{-1}$) by a factor of 0.2. This performance is attributed to the recommender’s ability to distinguish between DFAs and select the most accurate one.

Future applications of invariant representations of electron density, especially with equivariant model architectures, are promising for the future of quantum chemistry. The Δ-learning approach can be extended to other electronic properties and reaction energies, providing a robust framework for the development of next-generation quantum chemistry software.
factor of 6.5. If we picked the single DFA with the lowest average error over VSS-452, this DFA-derived MAE (6.2 kcal mol\(^{-1}\)) with the dispersion-corrected double hybrid DSD-BLYP-D3BJ) would be three times the recommender error. With the \(\Delta\)-learning-based recommender, we select DFAs that give errors within the targeted 3 kcal mol\(^{-1}\) threshold in 77% of cases (Fig. 2b). For almost all complexes (94%), the recommender suggests a DFA with a better result than the single-best DFA approach MAE of 6.2 kcal mol\(^{-1}\). Our reference method, DLPNO-CCSD(T), may suffer from errors when the multireference character of a system is large\(^3\). As we remove complexes with strong multireference character from VSS-452, we observe a monotonically decreasing recommender MAE, suggesting the possibility of further improving the recommender by providing it with more accurate reference data (Supplementary Fig. 7).

One distinct advantage of the recommender is that its performance should improve with an increasing number of DFAs despite using the same number of training points. If we had the smaller pool of 23 candidate DFAs\(^2\), the recommender MAE would have been 3.0 kcal mol\(^{-1}\) (Supplementary Table 1). As we add the 25 DFAs that incorporate alternate fractions of Hartree–Fock (HF) exchange, the MAE reduces to 2.1 kcal mol\(^{-1}\) (Supplementary Table 2). Notably, we gain this additional accuracy without substantially increasing the computational cost, as only additional DFT single-point energies have to be computed.

The other feature of our recommender is its system specificity. Compared with the strategy of selecting DFAs on the basis of the statistically averaged performance over a benchmark data set\(^3\), the recommender chooses the DFA based on the chemistry (here, the electron density) of the system. Our recommender avoids selecting a DFA that performs well on average yet particularly poorly on the given complex. As an example of how the single DFA approach can fail while the recommender will succeed, DSD-BLYP-D3BJ has the lowest DFA-derived MAE for \(\Delta\Delta E_{\text{H-L}}[f]\) (6.2 kcal mol\(^{-1}\)) against DLPNO-CCSD(T) but gives a relatively high absolute error of 9.2 kcal mol\(^{-1}\) on Co(III)(C\(_2\)H\(_5\)N\(_2\))(PH\(_3\))\(_2\). Our recommender instead selects \(\omega\)B97X, which has a very small error of 0.1 kcal mol\(^{-1}\) on this compound, even though the DFA-derived MAE of \(\omega\)B97X is much higher (16.9 kcal mol\(^{-1}\)) and only ranks 37th out of 48 DFAs (Fig. 2c and Supplementary Fig. 4).

We investigated the statistics of the DFA recommender to determine when it selects the top-performing functional. We focus on DFAs that are within the top five choices because they usually achieve transition metal chemical accuracy\(^2\) (3.0 kcal mol\(^{-1}\)) and because multiple DFAs can achieve similar accuracy for a TMC (Supplementary Figs. 6 and 8). For two-thirds (100 out of 152) of the test set TMCs in VSS-452, the recommended DFAs are within the top five DFAs (Supplementary Fig. 9). In over 85% of cases, our recommender selects a DFA in the top-10 DFAs. We achieve even more favourable ranking statistics when using only the 23 DFA subset, despite a higher recommender MAE (Supplementary Table 1). Here, 88% of the DFAs selected by the recommender are within the top five, and almost all (that is, 96%) are within the top ten DFAs (Supplementary Fig. 10). This behaviour would be general in our recommender approach: with a larger DFA pool, ranking statistics are less favourable, but lower MAEs are easier to achieve because there are more DFAs to choose from.

We assess the DFAs that mostly frequently reside in the top five choices suggested by our recommender relative to the ground truth. Out of the 48 DFAs, the local meta-generalized gradient approximation functional MN15-L augmented with 50% HF exchange (MN15-L:50%) appears most frequently as a top five DFA, with a likelihood of 43% for the 152 set-aside test TMCs in VSS-452 (Fig. 2d). Our recommender identifies the same DFA to have the highest likelihood (53%) of being in the top five candidate DFAs. For each semi-local functional family (the generalized gradient approximations BLYP and PBE, or the...
than a single DFA because it can adjust recommendations based on the ligand field strength. The DFA recommender works better for such tasks, as its ligand field strength. The DFA recommender works better for such tasks, as its ligand field strength.

E accuracy of a DFA for predicting ΔE changes between TMCs in the set-aside test set of VSS-452 for each DFA. Regardless of the type of DFA, all 48 Δ-learning models have a much stronger focus on the metal-local versus the metal-non-local environments compared with an untrained model, in agreement with our previous work. The DFA recommender avoids selecting DFAs that have a large MAE (Fig. 4b).

Interpreting the Δ-learning model and recommender predictions

We use a virtual adversarial attack to interrogate the Δ-learning model and the recommender. Using the predicted |ΔΔE^r^| from all 48 Δ-learning models, we use virtual adversarial attack to test the DFA recommender on a realistic task for exploring transition metal chemical space. Without seeing any TMCs in CSD-76, the 48 Δ-learning models have MAEs for |ΔΔE^r^| that range from 3.1 to 6.6 kcal mol^-1, with a median of 4.5 kcal mol^-1, which is less than twice the VSS-452 test set error (Fig. 5a). While there is some loss in accuracy, the Δ-learning transferability improves substantially over chemical-composition-based representations that often show increases larger than a factor of five in MAE on out-of-distribution CSD data, probably due to the use of electron density as inputs.

The chemistry of VSS-452 and CSD-76 varies drastically. While the former consists of small monodentate ligands, high-denticity ligands with distinct coordinating atoms are present in CSD-76. Therefore, it is no surprise that the top-performing DFAs will vary for the two data sets (Supplementary Table 5 and Supplementary Fig. 3, 13 and 16). Nevertheless, the recommender maintains the high ranking order (Spearman’s ρ = 0.95) of the top five DFAs relative to the ground truth on the CSD-76 set, demonstrating its great transferability to unseen chemistry. As in the case of VSS-452, the DFA recommender can also identify the top-performing DFAs and correctly predict the relative likelihood of a DFA to be accurate for CSD-76, even though the best DFAs now differ substantially (Fig. 5c). MN15-L:50%, the most probable (45%) DFA residing in the top five choices for VSS-452, has only a 17% likelihood of being in the top five for CSD-76. Meanwhile, the global meta-generalized gradient approximation hybrid M06, which is a top-five DFA only 30% of the time for VSS-452, becomes the most probable (56%) DFA residing in the top five choices for CSD-76. Despite the new data set, it still captures these trends well (Fig. 5c).

Transferability of the DFA recommender to diverse CSD complexes

A challenging test of the DFA recommender is its application to chemically distinct out-of-distribution complexes. For this purpose, we construct CSD-76, a set of 76 TMCs randomly sampled from the Cambridge Structural Database (CSD) that contain diverse ligand chemistry, symmetry and connectivity (Supplementary Fig. 13). Because these complexes have been experimentally synthesized and crystallized, they test the DFA recommender on a realistic task for exploring transition metal chemical space. Without seeing any TMCs in CSD-76, the 48 Δ-learning models have MAEs for |ΔΔE^r^| that range from 3.1 to 6.6 kcal mol^-1, with a median of 4.5 kcal mol^-1, which is less than twice the VSS-452 test set error (Fig. 5a). While there is some loss in accuracy, the Δ-learning transferability improves substantially over chemical-composition-based representations that often show increases larger than a factor of five in MAE on out-of-distribution CSD data, probably due to the use of electron density as inputs.

Using the predicted |ΔΔE^r^| from the 48 Δ-learning models, the recommender has an MAE of 3.0 kcal mol^-1, which again slightly outperforms the best Δ-learning model (3.1 kcal mol^-1) from the global meta-generalized gradient approximation hybrid M06-2X) and is only 1.5 times that of VSS-452 (Fig. 3a). If we insisted on using the single ‘best’ DFA benchmarked on the VSS-452 set (DSD-BLYP-D3BJ) for exploring CSD chemical space, we would have an MAE of 5.9 kcal mol^-1. Additional discussion about comparing the statistics of recommended DFAs for the out-of-distribution CSD-76 and the set-aside test set of VSS-452 can be found in Supplementary Figs. 14 and 15 and Section 1.

The chemistry of VSS-452 and CSD-76 varies drastically. While the former consists of small monodentate ligands, high-denticity ligands with distinct coordinating atoms are present in CSD-76. Therefore, it is no surprise that the top-performing DFAs will vary for the two data sets (Supplementary Table 5 and Supplementary Fig. 3, 13 and 16). Nevertheless, the recommender maintains the high ranking order (Spearman’s ρ = 0.95) of the top five DFAs relative to the ground truth on the CSD-76 set, demonstrating its great transferability to unseen chemistry. As in the case of VSS-452, the DFA recommender can also identify the top-performing DFAs and correctly predict the relative likelihood of a DFA to be accurate for CSD-76, even though the best DFAs now differ substantially (Fig. 5c). MN15-L:50%, the most probable (45%) DFA residing in the top five choices for VSS-452, has only a 17% likelihood of being in the top five for CSD-76. Meanwhile, the global meta-generalized gradient approximation hybrid M06, which is a top-five DFA only 30% of the time for VSS-452, becomes the most probable (56%) DFA to reside in the top five for CSD-76. Although the recommender does not have prior knowledge about the best-performing functionals for the new data set, it still captures these trends well (Fig. 5c).
The recommender also ‘intelligently’ avoids DFAs that only perform well on VSS-452 (for example, SCAN:40%, MN15-L:40% and M06-L:40%) and emphasizes DFAs that perform well on CSD-76 (for example, the range-separated global hybrid LRC-ωPBEh, the meta-generalized gradient approximation hybrid SCAN0 and the HF-exchange-augmented range-separated global hybrid LRC-ωPBEh, the meta-generalized gradient approximation PBE:20%).

We also tested our DFA recommender on a series of small linear TMCs (such as CoF₂) frequently investigated in benchmarking studies as well as isolated ions. Surprisingly, the DFA recommender is transferable (with an MAE of 2.9 kcal mol⁻¹) to these species despite the fact that TMCs in the training data are octahedral (Supplementary Table 6). Considering the relatively high cost of double hybrids (DHs), we re-evaluated the DFA recommender performance after excluding DHs and found that it does not deteriorate substantially (0.1–0.3 kcal mol⁻¹) because there are multiple DFAs that give comparably low errors on a complex (Supplementary Fig. 1 and Supplementary Table 7). These observations suggest that our DFA recommender can be reliably applied to explore diverse transition metal chemical spaces with high accuracy.

**Discussion**

DFT has become indispensable in both mechanistic study and accelerated, automated chemical discovery. Its accuracy, however, can depend on the combination of the DFA choice and system under study. The single-DFA approach widely used in VHTS leads to bias in data acquisition, and expert knowledge and heuristics cannot be expected to be predictive across large chemical spaces. Our DFA approach has the promise of overcoming these limitations by identifying the best choice of DFA as chemical composition is varied.

The current recommender approach has a few limitations. First, since it uses the B3LYP electron density as inputs, this recommender is not ‘zero cost’, and its advantage is therefore greatest in learning tasks where a property prediction with ‘beyond DFT’ accuracy is needed. The use of ML-predicted density, semi-empirical densities or guessed densities (for example, superpositions of atomic potentials) can reduce the cost further. Second, a DFA may not be universally accurate across all the properties (for example, band gap versus spin splitting energy) of a system. Sometimes, we may also want to keep the DFA choice the same for a set of systems (for example, in catalysis). Generalization of the current recommender may include redefining the loss function in the Δ-learning models to explicitly encode all relevant objectives. A natural extension to solid-state materials is possible, where the chief modifications would be the use of an orbital localization scheme (for example, Wannier functions) or a localized basis set for feature extraction, and an alternative technique, such as the Green’s function method based on the $G\beta$ method, for example, the screened Coulomb interaction $W_s$ to acquire high-quality reference data. Lastly, due to the use of the differences between electron density as well as Behler–Parrinello-type neural networks, our DFA recommender is not directly applicable to properties involving multiple nuclear coordinates for adiabatic spin splitting or systems with unseen chemical elements (that is, from distinct groups from the Periodic Table). Future work is needed to extend the recommender to these general applications.
Once trained, this recommender approach does not introduce additional computational cost when combined with existing DFT-based VHTS workflows that natively output an optimized geometry and electron density of a molecule. Therefore, it can be used in conjunction with VHTS to improve the quality of the data generated at no additional cost. Our recommender approach is not restricted to predicting a single electronic energy of a molecule and thus can be generalized to more complex applications such as catalysis. Although we demonstrate the recommender to select from conventional DFAs, it is a general approach for method selection, including among semi-empirical theories, ML-derived DFAs or wavefunction theories. By enabling computational chemists to side-step the trade-off between research scope and accuracy, we expect this recommender approach to be broadly useful alongside continuing advances in the methods available in the computational sciences.

Methods

DF procedure

In KS DFT, it is known that the ground-state energy of any interacting system is captured by a universal functional of the electron density23,37. In practice, the electron density is obtained from the occupied KS orbital wavefunction \( \psi(r) \), expanded as a linear combination of the products of one-electron basis functions \( \chi(r) \),

\[
\rho(r) = \sum_i |\chi_i(r)|^2 = \sum \mu_i D_{\mu_i} \chi_\mu (r) \chi_i (r),
\]

where \( D \) is the density matrix, \( \mu \) and \( \nu \) are indices for one-electron basis functions and \( i \) is the index of each KS orbital. The electron density in equation (1), however, is not expressed in an atom-centred basis and thus cannot be used directly as a representation in neural networks. Thus, it is common to use DF basis functions to rewrite the electron density as an expansion of atom-centred densities

\[
\rho(r) = \sum_A \sum Q c_{Q}^{A} \phi_{Q}(r-r_{A}) = \sum_A \rho_{A}(r),
\]

where \( \phi_{Q}(r-r_{A}) \) is the Qth DF basis function for atom A and \( c_{Q}^{A} \) is the coefficient for that basis function. However, \( c_{Q}^{A} \) contains elements resulting from DF basis functions where the angular momentum \( L \) is nonzero (\( L \neq 0 \)) and is thus not rotationally invariant but equivariant. To obtain a rotationally invariant representation, we calculated the power spectrum, \( p_{A}^{L} \), of \( c_{Q}^{A} \) as the norm for each angular momentum \( L \) in the DF basis set.

\[
p_{A}^{L} = \sum_{Q} \frac{1}{|C_{Q}^{A}|} \left| c_{Q}^{A} \right|^2.
\]

Behler–Parrinello-type neural networks for \( \Delta \)-learning

We built Behler–Parrinello-type neural networks using the DF representation of the TMCs in this work39. These fully connected neural networks used the DF representation of each atom as inputs,

\[
X_{A}^{l} = \sigma(W_{A}^{l} X_{A}^{l-1}),
\]

where \( X_{A}^{l} \) is the representation of atom A at layer \( l \), \( W_{A}^{l} \) is the \( l \)th-layer weights for the network of elements in group \( g \) and \( \sigma \) is the activation function. Specifically, \( X_{A}^{l} \) is the set of concatenated DF features of atom A (DF procedure section). The last layer of network outputs, \( X_{c}^{n} \) are summed for each chemical element \( c \):

\[
X_{c}^{n} = \sum_{A} X_{A}^{l}.
\]

These \( X_{c}^{n} \) of different elements are then concatenated and passed to a fully connected neural network to obtain the final output (Fig. 1c).

Our model has three main differences from the original Behler–Parrinello neural network. First, we replace the symmetry functions that describe the local geometric environment of an atom therein by the DF representation, which is derived from the electron density and is thus a more transferable representation. Second, we use the same local network for chemical elements that are in the same group of the Periodic Table (for example, O and S) to promote inter-row learning20. Lastly, we keep the latent vector \( X_{c}^{n} \) for each element and use a neural network to obtain the final output because our final target is not a single electronic energy of the ground state.

We adopted \( \Delta \)-learning strategies and chose our target to be the absolute difference of vertical spin splitting energies between the result from each DFA \( f \) and a reference calculation \( (|\Delta \Delta E_{\text{HS-LS}}|/f)| \). For each fully connected neural network, we used three hidden layers and 96 neurons per layer. The shifted softplus activation function, \( \sigma(x) = \text{softplus}(x) - \log(2) \) is used throughout. An alternative approach is to first learn the signed error \( (|\Delta \Delta E_{\text{HS-LS}}|/f) \) then take the absolute value. As we would like the recommender to pick a DFA that gives the lowest absolute error compared with the reference, we choose the direct approach instead. In practice, these two approaches show negligible differences in performance (Supplementary Table 8).

Recommender

We constructed separate \( \Delta \)-learning models for each DFA \( f \) to predict \( |\Delta \Delta E_{\text{HS-LS}}(f)| \) from a pre-selected pool of DFAs \( f \). For a given system, we recommend the DFA \( f_{\text{rec}} \) that yields the lowest predicted \( |\Delta \Delta E_{\text{HS-LS}}(f)| \)

\[
f_{\text{rec}} = \arg\min_{f} |\Delta \Delta E_{\text{HS-LS}}(f)|.
\]
to the reference method ($\langle \Delta \Delta E_{\text{ref}}(f) \rangle$) and the actual ranking of $f_{\text{rec}}$ among the pool of DFAs.

**Data set construction**

Mononuclear octahedral TMCs with Cr, Mn, Fe and Co in oxidation states II and III were studied in their HS and LS state: quintet and singlet for $d^6$ Co(II)/Fe(II) and $d^8$ Mn(III)/Cr(II), sextet and doublet for $d^8$ Fe(III)/Mn(II) and quartet and doublet for $d^5$ Cr(III) and $d^7$ Co(II). For VSS-452, we used 20 monodentate ligands from both the spectrochemical series and common organic ligands to obtain properties of complexes with ligand fields ranging from weak to strong (Supplementary Fig. 2). We allowed up to two unique ligands in a TMC and did not apply any constraints on ligand symmetry. Together with eight metal–oxidation state combination and 20 ligands, this procedure for assembling TMCs leads to a hypothetical space of 24,480 TMCs (8 × 20 × 160 homoepitic and 8 × 20 × 19 × 8 = 24,320 heteroleptic). We used k-medoids sampling to obtain 750 TMCs in this space as our starting data set. To test the transferability and practical usefulness of our recommender, we collected 100 experimentally synthesized TMCs with diverse ligand chemistry and connectivity from CSD as the starting point for CSD-76.

**DFT geometry optimization**

Since we are interested in vertical spin splitting, only one structure needs to be geometry optimized. In this case, we chose to optimize only the HS state. For each HS complex, a DFT geometry optimization with the B3LYP[40,41] global hybrid functional was carried out using a developer version of the graphical processing unit-accelerated electronic structure code TerraChem[42,43] version 1.9. The LANL2DZ effective core potential[42] basis set was used for metals, and the 6-31G* basis[38] for all other atoms. In all DFT geometry optimizations, level shifting[45] of 0.25 Ha on all virtual orbitals was employed. Initial geometries were assembled by molSimplify[46] for VSS-452 and were adopted from the crystal structure deposited in the CSD for CSD-76. These geometries were optimized using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm in translation rotation internal coordinates[47] to the default tolerances of $4.5 \times 10^{-4}$ hartree bohr$^{-1}$ for the maximum gradient and $10^{-8}$ hartree for the energy change between steps. Because all HS TMCs are open shell, the unrestricted formalism was used for all geometry optimizations. Although DFT can have large errors on energy evaluations, it has been shown that DFT still produces reasonable optimized TMC geometries[48]. Here, we choose B3LYP for geometry optimizations due to its widespread use in computational chemistry.

Geometry checks were applied to eliminate optimized structures that deviated from the expected octahedral shape following previously established metrics without modification[49]. 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[50] of $S(S + 1)$ by >1 Bohr magneton squared. After these two filtering steps, we converged 452 HS TMCs for VSS-452 and 76 HS TMCs for CSD-76 with good octahedral geometries and electronic structures (Supplementary Table 9).

**Single-point energy calculations**

We followed our established protocol for the computation of HS and LS electronic energies with multiple DFAs for the optimized TMCs using a developer version of Ps4 version 1.4 (ref. 25). In this workflow, the converged wavefunction obtained from the B3LYP geometry optimization was used as the 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.

The range of 23 DFAs used in the development of the protocol[41] were chosen to be evenly distributed among the rungs of Jacob’s ladder[26] (Supplementary Table 1). Practically, it has been observed that there is a nearly linear change of chemical properties (for example, spin splitting) computed with a DFA at different fractions of HF exchange[51,52]. Therefore, we sampled the HF exchange from 10% to 50% with an interval of 10% on five selected semi-local functionals (BLYP, PBE, SCAN, M06-1 and MN15-L). This procedure results in 25 additional DFAs (Supplementary Table 2). Combined with the original 23 DFAs, we have a final pool of 48 DFAs in total.

CCSD(T) has been treated as the ‘gold standard’ for quantum chemistry and is frequently used as a benchmark for DFT. Here, we used DLPNO-CCSD(T) (with T0 perturbative triple correction[53]), which is a proxy for canonical CCSD(T), as our reference method due to the sufficient accuracy of DLPNO-CCSD(T) on TMCs and the high computational cost of canonical CCSD(T) for a large data set[54]. In addition, we expect our DFA recommender approach to be general and have similar accuracy if reference data are derived from a higher-level theory (for example, phaseless auxiliary field quantum Monte Carlo[55]) or experiments in the future. Both DFT and DLPNO-CCSD(T) single-point energies for all non-singlet states were calculated with an unrestricted formalism and for singlet states with a restricted formalism. All single-point energy calculations were performed with a balanced polarized triple-zeta basis set def2-TZVP[56] using ORCA version 4.2.1. We performed B3LYP and DLPNO-CCSD(T) calculations with the def2-TZVP and def2-QZVP basis set on select small TMCs to verify that basis set truncation effects were -1 kcal mol$^{-1}$ (Supplementary Table 4). However, we acknowledge that our choice of reference in this work is limited by available computing power and resources.

**Training/testing partition and model training**

We randomly partitioned VSS-452, with 300 points (66%) as the training set and 152 (34%) points as the set-aside test set. For all $\Lambda$-learning models, the hyperparameters were selected using HyperOpt[57] with 200 evaluations, with 60 points of the training set used as the validation data. All $\Lambda$-learning models were built with PyTorch[58]. All models were trained with the Adam optimizer up to 2,000 epochs, using dropout and early stopping to avoid over-fitting. We treated CSD-76 as the out-of-distribution test set, and thus no points in CSD-76 were used during the model training procedure. All models and Python scripts to reproduce results reported in this work can be found[59] at https://doi.org/10.5281/zenodo.7350957 and https://github.com/hjkgrp/dfa_recommender.

**Statistics and reproducibility**

No statistical method was used to predetermine the sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments or outcome assessment.

**Data availability**

Source data for Figs. 2–5 are available with this manuscript. All structures and energies used to train the models along with the trained machine learning models are available at ref. 60 (https://doi.org/10.5281/zenodo.7350957).

**Code availability**

All models and Python scripts to reproduce results reported in this work can be found at Ref. 64 (https://doi.org/10.5281/zenodo.7350957).

**References**

1. Coley, C. W., Eype, N. S. & Jensen, K. F. Autonomous discovery in the chemical sciences part I: progress. Angew. Chem. Int. Ed. Engl. 59, 22858–22893 (2020).
2. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. 1, 011002 (2013).
3. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. Nature 559, 547–555 (2018).
Carleo, G. et al. Machine learning and the physical sciences. Rev. Mod. Phys. 91, 045002 (2019).

Nandy, A. et al. Computational discovery of transition-metal complexes: from high-throughput screening to machine learning. Chem. Rev. 121, 9927–10000 (2021).

Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for density functional theory. Chem. Rev. 112, 289–320 (2012).

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. 115, 2315–2372 (2017).

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. 12, 13021–13036 (2021).

Loipersberger, M., Cabral, D. G. A., Chu, D. B. K. & Head-Gordon, M. Mechanistic insights into Co and Fe quaterpyridine-based CO2 reduction catalysts: metal–ligand orbital interaction as the key driving force for distinct pathways. J. Am. Chem. Soc. 143, 744–763 (2021).

Zhang, D. Y. & Truhlar, D. G. Spin splitting energy of transition metals: a new, more affordable wave function benchmark method and its use to test density functional theory. J. Chem. Theory Comput. 16, 4416–4428 (2020).

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. 120, 143001 (2018).

Smith, J. S., Isaev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem. Sci. 8, 3192–3203 (2017).

Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat. Commun. 13, 2453 (2022).

Sauceda, H. E. et al. BIGDML—towards accurate quantum machine learning force fields for materials. Nat. Commun. 13, 3733 (2022).

Dick, S. & Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. Nat. Commun. 11, 3509 (2020).

Kirkpatrick, J. et al. Pushing the frontiers of density functionals by solving the fractional electron problem. Science 374, 1385–1389 (2021).

Li, L. et al. Kohn-sham equations as regularizer: building prior knowledge into machine-learned physics. Phys. Rev. Lett. 126, 036401 (2021).

Ma, H., Narayanaswamy, A., Riley, P. & Li, L. Evolving symbolic density functionals. Sci. Adv. 8, eabq0279 (2022).

Herrmann, J., Schätzle, Z. & Noé, F. Deep-neural-network solution of the electronic Schrödinger equation. Nat. Chem. 12, 891–897 (2020).

Kauwe, S. K., Graser, J., Murdock, R. & Sparks, T. D. Can machine learning find extraordinary materials? Comput. Mater. Sci. 174, 109498 (2020).

McAnamana-Breton, S. & Waller, M. P. Rational density functional selection using game theory. J. Chem. Inf. Model. 58, 61–67 (2018).

Jiang, W., DeYonker, N. J., Determan, J. J. & Wilson, A. K. Toward accurate theoretical thermochemistry of first row transition metal complexes. J. Phys. Chem. A 116, 870–885 (2012).

Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. Phys. Rev. 140, A1133–A1138 (1965).

Margraf, J. T. & Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. Nat. Commun. 12, 344 (2021).

Grisafi, A. et al. Transferable machine-learning model of the electron density. ACS Cent. Sci. 5, 57–64 (2019).

Frénay, B. & Verleysen, M. Classification in the presence of label noise: a survey. IEEE Trans. Neural Netw. Learn. Syst. 25, 845–869 (2013).

Floser, B. M., Guo, Y., Riplinger, C., Tuczek, F. & Neese, F. Detailed pair natural orbital-based coupled cluster studies of spin crossover energetics. J. Chem. Theory Comput. 16, 2224–2235 (2020).

Perdew, J. P. & Schmidt, K. Jacob’s ladder of density functional approximations for the exchange-correlation energy. Density Funct. Theory Its Application Mater. 577, 1–20 (2001).

Harper, D. R. et al. Representations and strategies for transferable machine learning improve model performance in chemical discovery. J. Chem. Phys. 156, 074101 (2022).

Duan, C., Liu, F., Nandy, A. & Kulik, H. J. Data-driven approaches can overcome the cost-accuracy trade-off in multireference diagnostics. J. Chem. Theory Comput. 16, 4373–4387 (2020).

Lehtola, S. Assessment of initial guesses for self-consistent field calculations. Superposition of atomic potentials: simple yet efficient. J. Chem. Theory Comput. 15, 1593–1604 (2019).

Mauer, L. R., Bursch, M., Grimm, S. & Hansen, A. Assessing density functional theory for chemically relevant open-shell transition metal reactions. J. Chem. Theory Comput. 17, 6134–6151 (2021).

Miyato, T., Maeda, S. I., Koyama, M. & Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 41, 1979–1993 (2019).

Janet, J. P. & Kulik, H. J. Resolving transition metal chemical space: feature selection for machine learning and structure-property relationships. J. Phys. Chem. A 121, 8939–8954 (2017).

Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge structural database. Acta Crystallogr. B 72, 171–179 (2016).

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. 10, 7913–7922 (2019).

Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. Phys. Rev. 136, 8864–8871 (1964).

Pritchard, B. P., Altarawy, D., Didier, B., Gibson, T. D. & Windus, T. L. New basis set exchange: an open, up-to-date resource for the molecular sciences community. J. Chem. Inf. Model. 59, 4814–4820 (2019).

Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys. Rev. Lett. 98, 146401 (2007).

Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. J. Chem. Phys. 98, 5648–5652 (1993).

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. 98, 11623–11627 (1994).

Seritan, S. et al. TeraChem: a graphical processing unit-activated electronic structure package for large-scale ab initio molecular dynamics. WIREs Comput. Mol. Sci. 11, e1494 (2021).

Ufimtsev, I. S. & Martinez, T. J. Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics. J. Chem. Theory Comput. 5, 2619–2628 (2009).

Hay, P. J. & Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. J. Chem. Phys. 82, 299–310 (1985).

Saunders, V. R. & Hillier, I. H. A “level-shifting” method for converging closed shell Hartree–Fock wave functions. Int. J. Quant. Chem. 7, 699–705 (1973).
46. Ioannidis, E. I., Gani, T. Z. H., & Kulik, H. J. molSimplify: a toolkit for automating discovery in inorganic chemistry. J. Comput. Chem. 37, 2106–2117 (2016).
47. Wang, L.-P. & Song, C. Geometry optimization made simple with translation and rotation coordinates. J. Chem. Phys. 144, 214108 (2016).
48. Finney, B. A., Chowdhury, S. R., Kirkvold, C., & Vlaisavljevich, B. CASPT2 molecular geometries of Fe(II) spin-crossover complexes. Phys. Chem. Chem. Phys. 24, 1390–1398 (2022).
49. 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. 15, 2331–2345 (2019).
50. Smith, D. O. A. et al. PSI4 1.4: open-source software for high-throughput quantum chemistry. J. Chem. Phys. 152, 184108 (2020).
51. Liu, F. et al. Bridging the homogeneous–heterogeneous divide: modeling spin for reactivity in single atom catalysis. Front. Chem. 7, 219 (2019).
52. Reiher, M. Theoretical study of the Fe(phen)_2(NCS)_2 spin-crossover complex with reparametrized density functionals. Inorg. Chem. 41, 6928–6935 (2002).
53. Shee, J., Arthur, E. J., Zhang, S., Reichman, D. R. & Friesner, R. A. Phaseless auxiliary-field quantum monte carlo on graphical processing units. J. Chem. Theory Comput. 14, 4109–4121 (2018).
54. Bergstra, J., Yamins, D. & Cox, D. D. HyperOpt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. In Proceedings of the 12th Python in Science Conference, 13, 20 (2013).
55. Pytorch https://pytorch.org/ (2022).
56. Duan, C., Nandy, A., Meyer, R., Arunachalam, N. & Kulik, H. J. A transferable recommender approach for selecting the best density functional approximations in chemical discovery. Zenodo https://doi.org/10.5281/zenodo.7350957 (2022).

Acknowledgements
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 program (R.M.) as well as by the Office of Naval Research under grant nos. N00014-18-1-2434 (A.N.) and N00014-20-1-2150 (C.D.). N.A. was partially supported by the U.S. Department of Energy under grant no. DE-NA0003965. C.D. was partially supported by a seed fellowship from the Molecular Sciences Software Institute under NSF grant OAC-1547580. A.N. and N.A. were partially supported by a National Science Foundation Graduate Research Fellowship under grants nos. 1122374 and 1745302, respectively. The authors thank A. H. Steeves for a critical reading of the manuscript.

Author contributions
C.D.: conceptualization, methodology, software, validation, investigation, data curation, writing of original draft, review and editing, and visualization. A.N.: data curation, software, and review and editing. R.M.: data curation, software, validation, and review and editing. N.A.: software, and review and editing. H.J.K.: conceptualization, supervision, project administration, funding acquisition, and review and editing.

Competing interests
The authors declare no competing financial interest.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-022-00384-0.

Correspondence and requests for materials should be addressed to Heather J. Kulik.

Peer review information Nature Computational Science thanks Jan Hermann, Stefan Vuckovic and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Handling editor: Kaitlin McCardle, in collaboration with the Nature Computational Science team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022