A Density Functional Recommendation Approach for Accurate Predictions of Vertical Spin Splitting of Transition Metal Complexes

Chenru Duan [1 2] **Aditya Nandy** [1 2] **Heather J. Kulik** [1]

Abstract

Both conventional and machine learning-based density functional approximations (DFAs) have emerged as versatile approaches for virtual high-throughput screening and chemical discovery. To date, however, no single DFA is universally accurate for different chemical spaces. This DFA sensitivity is particularly high for open-shell transition-metal-containing systems, where strong static correlation may dominate. With electron density fitting and transfer learning, we build a DFA recommender that selects the DFA with the lowest expected error in a system-dependent manner. We demonstrate this recommender approach on the prediction of vertical spin-splitting energies (i.e., the electronic energy difference between the high-spin and low-spin state) of challenging transition metal complexes. This recommender yields relatively small errors (i.e., 2.1 kcal/mol) for transition metal chemistry and captures the distributions of the DFAs that are most likely to be accurate.

1. Introduction

Virtual high-throughput screening (Coley et al., 2020; Curtarolo et al., 2013) and machine learning (ML)-accelerated chemical discovery (Nandy et al., 2021; Keith et al., 2021) with approximate density functional theory (DFT) have started to address the combinatorial challenges in discovering and designing functional molecules and materials (Rosen et al., 2021). A density functional approximation (DFA), that works well on certain systems, however, may fail prominently on other systems due to the approximations made in the exchange-correlation functional (Cohen et al., 2012; Mardirossian & Head-Gordon, 2017). This DFA dependence is especially strong in open-shell transition metal chemistry, where strong static electron correlation may dominate (Janesko, 2021). To ensure high fidelity in chemical discovery, it is vital to maintain a rather uniform accuracy within data sets so that surrogate ML models are not biased (Duan et al., 2021).

Despite recent advances in using ML to improve the cost–accuracy trade-off for first-principles calculations (Dick & Fernandez-Serra, 2020; Qiao et al., 2020; Chen et al., 2021; Kirkpatrick et al., 2021), current ML density functionals still have severe limitations that curtail their practical use (Unke et al., 2021; Pederson et al., 2022). These DFAs have mostly been developed for and applied on narrow sets of closed-shell organic molecules and are often less transferable compared to conventional DFAs developed in the theoretical chemistry community over the past few decades. More importantly, these ML functionals only target the electronic energy of a geometry rather than other properties of chemical interest, such as those involving multiple electronic states.

Here, we take an alternate route. Instead of developing a new DFA, we leverage transfer learning (TL) to develop a recommender (McAnanama-Brereton & Waller, 2018) to select the best conventional DFA for a given system and a property of interest. Specifically, we aim to recommend a DFA that accurately evaluates the vertical spin splitting of transition metal complexes (TMCs). Our recommender has a mean absolute error (MAE) of 2.1 kcal/mol and identifies DFAs that are likely to be accurate and successfully selects a top-5 DFA two-thirds of the time.

2. Methods

2.1. Density fitting procedure

In Kohn–Sham (KS) DFT, it is known that the ground state energy of any interacting system is captured by a universal functional of the electron density (Hohenberg & Kohn, 1964; Kohn & Sham, 1965). In practice, the electron density ($\rho(r)$) is obtained from the occupied KS orbitals $\psi_i(r)$, expanded as a linear combination of the products of one-electron basis

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA [2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Correspondence to: Heather J. Kulik <hjkulik@mit.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

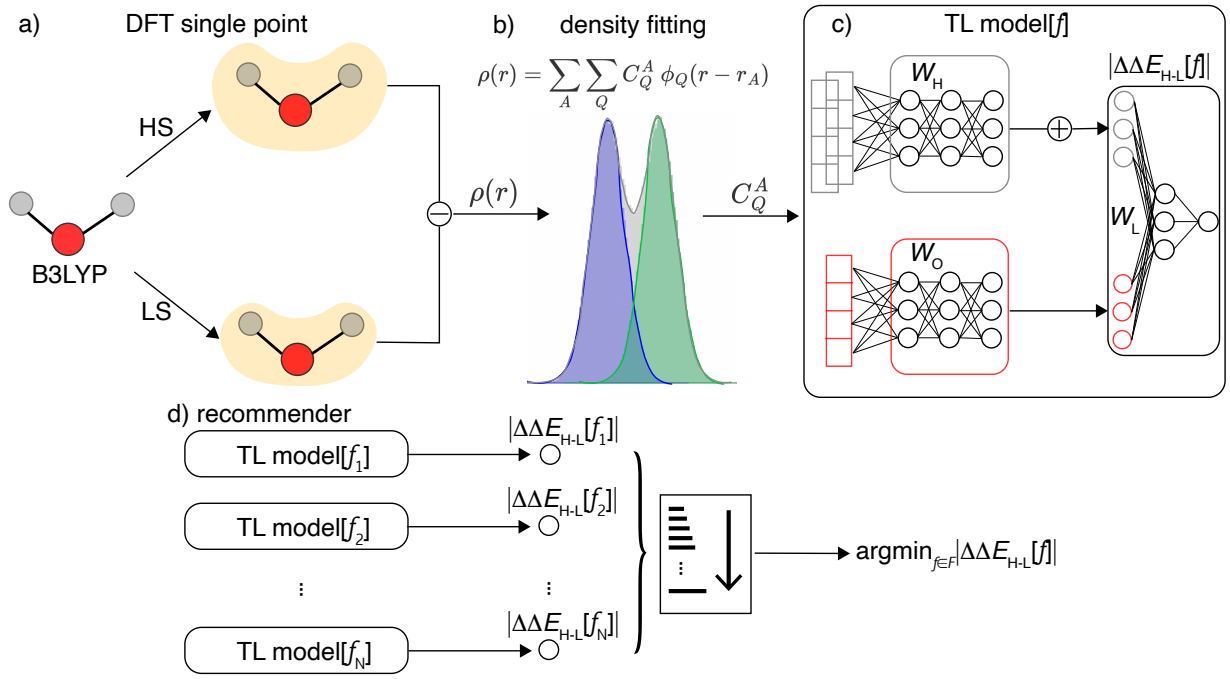

Figure 1. Workflow for the DFA recommender. a) B3LYP/def2-TZVP single-point energy calculations are performed on both the high-spin (HS) and low-spin (LS) states to obtain their electron density at B3LYP level. b) The difference of the electron densities between the HS and LS states is decomposed to each atom with a density fitting procedure. c) These coefficients are passed to a Behler–Parrinello-type neural network as a TL model to predict $|\Delta\Delta E_{\mathrm{H-L}}[f]|$ for each DFA f in our pool of 48 DFAs. Coefficients of different atoms but the same group in the periodic table share the same local network and weights (e.g., W_O for O (red) and W_H for H (gray)). The latent vector of each element are lastly concatenated and passed to a fully-connect network (W_L) for predicting $|\Delta\Delta E_{\mathrm{H-L}}[f]|$. d) The predicted $|\Delta\Delta E_{\mathrm{H-L}}[f]|$ are sorted, where the DFA that yields the lowest predicted $|\Delta\Delta E_{\mathrm{H-L}}[f]|$ is recommended.

functions $\chi_\mu(r)$,

$$\rho(r) = \sum_i |\psi_i(r)|^2 = \sum_{\mu\nu} D_{\mu\nu}\chi_\mu(r)\chi_\nu(r) \tag{1}$$

where D is the density matrix and μ and ν are indices for one-electron basis functions. The electron density in Eq. 1, however, is not expressed in an atom-centered basis and thus cannot be directly used as features in neural networks. It is thus common to use density-fitting (DF) basis functions to re-write the electron density as an expansion of atom-centered densities,

$$\rho(r) = \sum_A \sum_Q C_Q^A \phi_Q(r - r_A) = \sum_A \rho_A(r) \tag{2}$$

where $\phi_Q(r - r_A)$ is the Qth DF basis function for atom A (Margraf & Reuter, 2021; Grisafi et al., 2019). However, C_Q^A contains elements resulting from DF basis sets where the angular momentum is nonzero ($L \neq 0$) and is thus not rotationally invariant. To obtain a rotationally invariant representation, we calculated the power spectrum of C_Q^A as the norm for each angular momentum L in the DF basis set.

$$p_L^A = \sum_{Q\in L} ||C_Q^A||^2 \tag{3}$$

Therefore, the p_L^A can be used as features to represent the chemical environment of atom A (Figure 1 (b)). For this procedure, we employ only the density obtained from B3LYP regardless of which functional is being studied in the TL models.

Here we considered the vertical spin-splitting energy ($\Delta E_{\mathrm{H-L}}$) as our property of interest, which is the electronic energy difference between the high-spin (HS) and low-spin (LS) states of open-shell TMCs. We focus on the spin-splitting energy because it identifies the quantum mechanical ground state, which is an essential property of an open-shell system (Nandy et al., 2021). Because our target property involves two distinct spin states in open-shell systems, we decomposed the difference between the HS and LS electron densities for both the majority spin and minority spin separately (Figure 1 (a)).

$$\begin{aligned} \rho^\alpha_{\mathrm{HS}}(r) - \rho^\alpha_{\mathrm{LS}}(r) &= \sum_A \sum_Q C_Q^{A^\alpha} \phi_Q(r - r_A) \\ \rho^\beta_{\mathrm{HS}}(r) - \rho^\beta_{\mathrm{LS}}(r) &= \sum_A \sum_Q C_Q^{A^\beta} \phi_Q(r - r_A) \end{aligned} \tag{4}$$

For an atom A, we obtained and concatenated the power spectra of $C_Q^{A^\alpha}$ and $C_Q^{A^\beta}$ as its features using Eq. 3. We

used the def2-universal-jkfit (Pritchard et al., 2019) as our DF basis set throughout in this work. The number of DF features for different atoms can vary due to their difference in the auxiliary basis functions used. Here, we zero-padded the DF features for all atoms to the maximum dimension of 116, which is the number of the DF basis set of the largest transition metal atom in a TMC.

2.2. Behler–Parrinello-type networks as transfer learning models

We built Behler–Parrinello-type networks upon the DF representation of the TMCs in this work (Behler & Parrinello, 2007). This fully-connected neural network used the DF representation of each atom as input,

$$X_A^l = \sigma(W_{A\in g}^l X_A^{l-1}) \tag{5}$$

where X_A^l is the representation of atom A at layer l, $W_{A\in g}^l$ is the lth-layer weights for the network of elements in group g, and σ is the activation function. Specifically, X_A^0 is the set of concatenated DF features of atom A (see Section 2.1). The last layer of the network outputs, X_A^n, are summed for each chemical element (e), $X_e^n = \sum_{A\in e} X_A^n$. These X_e^n of different elements are then concatenated and passed to a fully-connected neural network to obtain the final output (Figure 1 (c)).

The main differences between our model and the original Behler–Parrinello neural network are: 1) 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; 2) we use the same local network for chemical elements that are in the same group of the periodic table (e.g., O and S) to promote inter-row learning; 3) we keep the latent vector X_e^n for each element and use a neural network to obtain the final output because our final target is not an electronic energy of the ground state.

We adopted TL strategies and chose our target to be the absolute difference of vertical spin-splitting energies between the result from each DFA and a reference calculation ($|\Delta\Delta E_{\mathrm{H-L}}[f]|$, see Section 3.3). For each fully-connected neural network, we used three hidden layers and 96 neurons per layer. The shifted softplus activation function (i.e., $\sigma(x) = \mathrm{softplus}(x) - \log(2)$) is used throughout.

2.3. Recommender

One may expect that the best approach to build a recommender is to treat it as a multi-class classification problem, where the classes are different DFAs. We find, however, that many DFAs can perform similarly well in some cases, leading to noise in the labeling process of determining the "best" DFA (Appendix A). Therefore, we constructed separate TL models for each DFA (f) to predict $|\Delta\Delta E_{\mathrm{H-L}}[f]|$ from a pre-selected pool of DFAs (F). For a given system, we recommend the DFA, f_{rec}, that yields the lowest predicted $|\Delta\Delta E_{\mathrm{H-L}}[f]|$,

$$f_{\mathrm{rec}} = \mathrm{argmin}_{f\in F} |\Delta\Delta E_{\mathrm{H-L}}[f]| \tag{6}$$

When we evaluate the practical performance of the DFA recommender, we focus on the absolute error introduced by using f_{rec} relative to the reference method (i.e., $|\Delta\Delta E_{\mathrm{H-L}}[f_{\mathrm{rec}}]|$) and the actual ranking of f_{rec} among the pool of DFAs F.

3. Data sets

3.1. 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 states: quintet and singlet for d^6 Co(III)/Fe(II) and d^4 Mn(III)/Cr(II); sextet and doublet for d^5 Fe(III)/ Mn(II), and quartet and doublet for d^3 Cr(III) and d^7 Co(II). 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 (Appendix B). We allowed up to two unique ligands in a TMC and did not pose any constraints on ligand symmetry, leading to a hypothetical space of 24,480 TMCs. We randomly sampled 750 TMCs in this space as our starting data set.

3.2. DFT geometry optimization

Because we are interested in vertical spin splitting, only one structure needs to be geometry optimized. In this case, we optimize only the HS state. For each HS complex, a DFT geometry optimization with the B3LYP (Becke, 1993; Stephens et al., 1994; Lee et al., 1988) global hybrid functional was carried out using a developer version of graphical processing unit (GPU)-accelerated electronic structure code TeraChem (Ufimtsev & Martinez, 2009; Seritan et al., 2021). The LANL2DZ effective core potential (Hay & Wadt, 1985) basis set was used for metals and the 6-31G* basis (Pritchard et al., 2019) for all other atoms. In all DFT geometry optimizations, level shifting (Saunders & Hillier, 1973) of 0.25 Ha on all virtual orbitals was employed. Initial geometries were assembled by molSimplify (Ioannidis et al., 2016) and optimized using the L-BFGS algorithm in translation rotation internal coordinates (TRIC) (Wang & Song, 2016) 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. Because all HS TMCs are open-shell, the unrestricted formalism was used for all geometry optimizations.

Geometry checks were applied to eliminate optimized struc-

tures that deviated from the expected octahedral shape following previously established metrics without modification (Nandy et al., 2018; Duan et al., 2019). 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 of $S(S+1)$ by $> 1\mu_B^2$. (Nandy et al., 2018; Duan et al., 2019). After these two filtering steps, we converged 452 HS TMCs with good octahedral geometries and electronic structures.

3.3. Single-point energy calculations with multiple DFAs and DLPNO-CCSD(T)

We followed the protocol introduced by Duan *et al.* (Duan et al., 2021) for the computation of HS and LS electronic energies with multiple DFAs for the 452 optimized TMCs using a developer version of Psi4 1.4 (Smith et al., 2020). 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 (Duan et al., 2021) were chosen to be evenly distributed among the rungs of "Jacob's ladder" (Perdew & Schmidt, 2001) (Appendix C). Practically, it has been observed that there is a nearly linear change of chemical properties (e.g., spin splitting) computed with a DFA at different percentages of Hartree–Fock (HF) exchange (Liu et al., 2019). Therefore, we sampled the HF exchange from 10% to 50% with an interval of 10% on five selected semi-local functionals (i.e., BLYP, PBE, SCAN, M06-L, and MN15-L). This procedure results in 25 additional DFAs (Appendix D). 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 benchmark for DFT (Mardirossian & Head-Gordon, 2017). Here, we used domain-based local pair natural orbital (DLPNO)-CCSD(T), 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 (Flaser et al., 2020). In addition, we expect our DFA recommender approach to be general and have similar accuracy if reference data is derived from higher-level theory (e.g., phaseless auxiliary field quantum Monte-Carlo (Shee et al., 2018)) or experiments in the future.

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 (Pritchard et al., 2019). We refer to the data set that contains vertical spin splitting ($\Delta E_{\mathrm{H-L}}$) computed from 48 DFAs and DLPNO-CCSD(T) as *VSS-452*.

3.4. Train/test 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 TL models, the hyperparameters were selected using HyperOpt (Bergstra et al., 2013) with 200 evaluations, with 60 points of the training set used as the validation set. All TL models were built with PyTorch (Paszke et al., 2019). All models were trained with the Adam optimizer up to 2000 epochs, using dropout and early stopping to avoid over-fitting.

4. Results

4.1. Performance of DFA recommender

We first demonstrate the performance of our TL models for predicting the difference in vertical spin-splitting energy obtained by a DFA and DLPNO-CCSD(T). The 48 TL models have mean absolute errors (MAEs) ranging from 2.3 kcal/mol to 3.4 kcal/mol, with a median MAE of 2.5 kcal/mol (Figure 2(a)). These MAEs are low considering the fact that the TL models were only trained on a small data set of 300 TMCs that contain diverse chemistry. In addition, these TL models are already at the level of transition metal chemical accuracy (i.e., a 3 kcal/mol uncertainty of experimental observations for TMCs (Jiang et al., 2012)) and have comparable performance to previous examples in the literature (Husch et al., 2021). Interestingly, the ranking of TL model performance (i.e., MAE) does not have the same order as that of the error of the underlying DFA relative to DLPNO-CCSD(T). For example, the DFA with the lowest TL MAE is a double-hybrid functional DSD-PBEB95-D3BJ, which has the fifth-lowest MAE relative to the reference calculation (Appendix F). MN15, which gives the highest TL MAE, only ranks 17th among the DFAs with highest MAEs relative to DLPNO-CCSD(T). For the set of functionals, the rank-ordering coefficient (i.e., Spearman's r) between DFAs ranked by TL model MAE and those ranked by DFA performance is 0.36. This observation suggests that a TL model does not necessarily perform better when the MAE of the DFA-derived MAE of the baseline DFA is smaller, posing an interesting question of how to select the best baseline method from which the transfer learning yields the lowest errors.

We then utilize the predicted $|\Delta\Delta E_{\mathrm{H-L}}[f]|$ for all 48 DFAs to recommend a DFA and evaluate the performance of the recommender (Section 2.3). The recommender achieves an MAE of 2.1 kcal/mol, better than the best of the 48 TL models. This MAE is only 2.5 times the theoretical lower bound

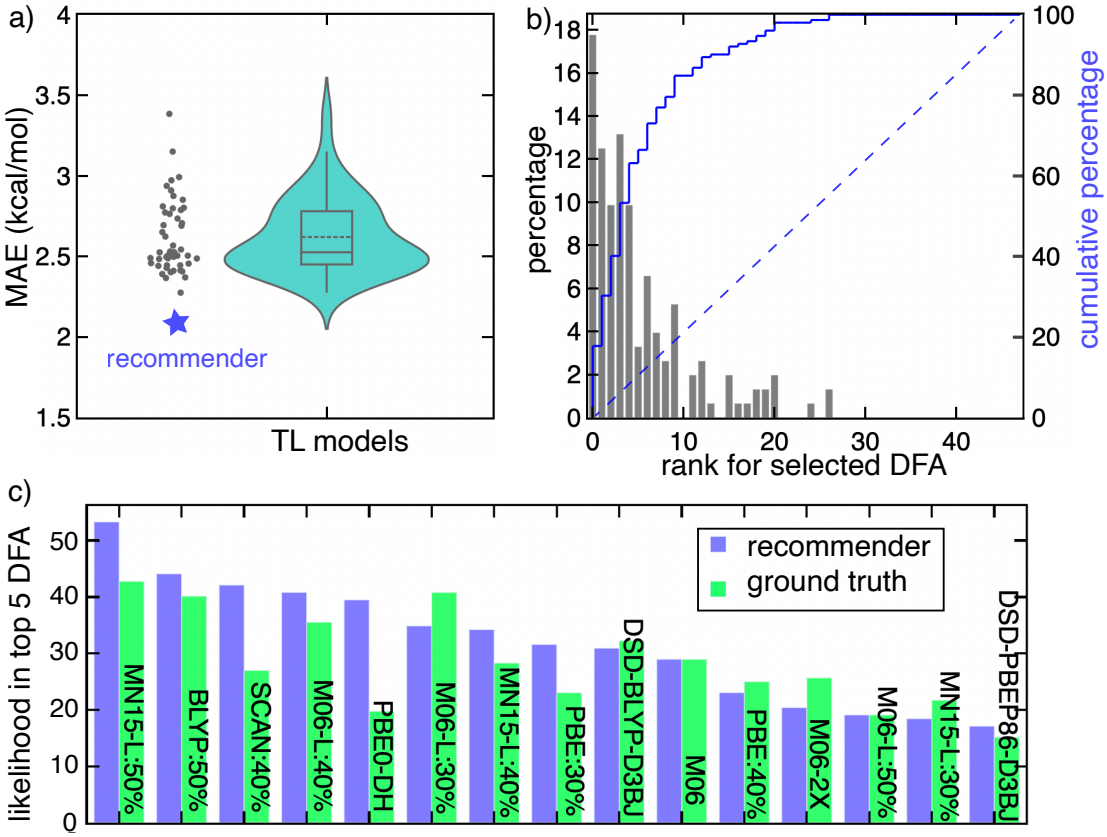

Figure 2. a) Violin plot for the MAE of 48 TL models for the prediction of $|\Delta\Delta E_{\mathrm{H-L}}|$ with a box indicating the mean (dotted line) and median (solid line). The MAEs for each TL model (black circle) and our recommender approach (blue star) are also shown. b) Normalized distribution of the rank of the DFA selected using our recommender approach, with the cumulative percentage (blue solid line) shown according to the axis on the right. The cumulative curve for a random guess (blue dashed line) is also shown. c) Bar plot for the percentage likelihood of a DFA residing in the top 5 choices suggested by ground truth (green) and our recommender approach (blue). The DFAs are sorted in a descending order of the predicted likelihood of the recommender. In all cases, the model performance is evaluated on the set-aside 152 test complexes of *VSS-452*.

(0.8 kcal/mol) that we obtain by assuming that the DFA that gives the lowest absolute error can always be selected (Appendix E). Our recommender outperforms random DFA selection (MAE=13.3 kcal/mol) by 6.5 fold. More importantly, even if an alternative recommender had prior knowledge to pick the single DFA with the lowest error in the *VSS-452* test set, its DFA-derived MAE (i.e., 6.2 kcal/mol) would be three times larger than that of our recommender. One distinct advantage of this recommender approach is that its performance is likely to improve systematically with increasing number of DFAs under consideration, despite using the same training data set and TL models. For example, the MAE of the recommender would achieve an MAE 3.0 kcal/mol if we had used the smaller set of 23 DFAs introduced in Duan *et al.* (Duan et al., 2021). There, its performance is even slightly worse than the median of the 48 TL models (2.5 kcal/mol). However, as we added the remaining HF-sampled DFAs, the MAE decreased to 2.1 kcal/mol, exceeding the accuracy of our best TL model of DSD-PBEB95-D3BJ (2.3 kcal/mol).

One distinct feature of our recommender approach is its system specificity, meaning that a DFA is selected based on the prior knowledge (here the electron density) of a system. As a result, our recommender can avoid selecting a DFA that is on average good, yet particularly bad on the given complex. For example, DSD-BLYP-D3BJ has the lowest DFA-derived MAE for $\Delta E_{\mathrm{H-L}}$ (i.e., 6.2 kcal/mol) against DLPNO-CCSD(T) but gives a relatively high absolute error of 9.2 kcal/mol on *cis* $\mathrm{Fe(III)(acetonitrile)_4(CO)_2}$. Our recommender, instead, selects M06-2X, which has a very small error of 0.1 kcal/mol on this compound, despite the fact the DFA-derived MAE of M06-2X is higher (i.e, 6.6 kcal/mol).

4.2. Statistical analysis of recommended DFAs

Next, we investigate the statistics of the recommended DFA. Surprisingly, we find a significant percentage (18%) of our

recommended DFAs actually produce the lowest absolute error of vertical splitting among the 48 candidate DFAs, although some TMCs might have multiple equivalently performing DFAs (Appendix A). Over 40% of our recommended DFAs rank in the top 3 among the 48 DFAs and two-thirds (i.e., 100 out of 152 complexes in the set-aside test set of *VSS-452*) of the recommended DFAs are within the top 5 DFAs relative to the ground truth (Figure 2 (b)). In less than 15% of cases, our approach recommends a DFA that is not in the top 10 out of 48 candidate DFAs. Interestingly, we get more favorable statistics on the ranking using only the 23 DFAs, despite a higher recommender MAE. With only 23 DFAs, 40% of our recommended DFAs are the best ones, 75% are within the top 3, 88% are within the top 5, and nearly none (i.e. 4%) fall out of the top 10 DFAs (Appendix G). This behavior is expected to be general in our recommender approach: With more candidate DFAs in the pool, it is more difficult to get favorable ranking statistics but easier to obtain a lower MAE for practical performance because there are more DFAs to choose from.

Lastly, we focus on DFAs that are within the top 5 choices, because they usually result in the accuracy required for transition metal chemistry (i.e., 3 kcal/mol, Appendix E). (Jiang et al., 2012). Out of the 48 DFAs, MN15-L with 50% HF exchange (i.e., MN15-L:50%) is most frequently among the top 5 DFAs, with a likelihood of 43% for the 152 set-aside test complexes (Figure 2 (c)). Correspondingly, our recommender identifies the same DFA to have the highest likelihood (53%) being in top 5. In addition, our recommender maintains the rank ordering of probable top-5 DFAs compared to the ground truth, leading to a Spearman's r of 0.96. This extremely high correspondence demonstrates that our DFA recommender is capable of identifying DFAs that are most likely to be accurate in a chemical space.

5. Conclusions

ML techniques have been adopted in computational chemistry to develop data-driven DFAs for accelerated property assessment within large design spaces. However, these ML DFAs are generally less transferable compared to conventional DFAs and have a strong focus on main group chemistry. Here, we take a different approach by framing the question as DFA recommendation and demonstrate our recommender approach on the challenging space of transition metal chemistry. Using a DF procedure, we decompose electron density into atom-centered features that are compatible with neural network architectures. We build TL models with these DF features, resulting in the accuracy required for transition metal chemistry. Exploiting a set of TL models built on different DFAs, we develop a recommender that selects a DFA in a system-specific manner. This recommender gives a high accuracy of 2.1 kcal/mol and selects a top-5 DFA two-thirds of the time. Our DFA recommender is also capable of identifying DFAs that are most likely to be accurate in a chemical space by preserving the rank ordering (Spearman's r of 0.96) of top-5 DFAs compared to the ground truth. We have demonstrated our recommender for selecting conventional DFAs, however it should be considered as a general approach for method selection in computational chemistry. We anticipate this recommender approach to continue to be useful in virtual high-throughput screening and chemical discovery with advances in both conventional and ML-based quantum chemistry methods.

References

Becke, A. D. Density-functional thermochemistry. iii. the role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. doi: 10.1063/1.464913. URL https://doi.org/10.1063/1.464913.

Behler, J. and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007. doi: 10.1103/PhysRevLett.98.146401. URL https://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

Bergstra, J., Yamins, D., and Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 115–123, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/bergstra13.html.

Chen, Y., Zhang, L., Wang, H., and E, W. Deepks: A comprehensive data-driven approach toward chemically accurate density functional theory. *Journal of Chemical Theory and Computation*, 17(1):170–181, 2021. doi: 10.1021/acs.jctc.0c00872. URL https://doi.org/10.1021/acs.jctc.0c00872. PMID: 33296197.

Cohen, A. J., Mori-Sánchez, P., and Yang, W. Challenges for density functional theory. *Chemical Reviews*, 112(1):289–320, 2012. doi: 10.1021/cr200107z. URL https://doi.org/10.1021/cr200107z. PMID: 22191548.

Coley, C. W., Eyke, N. S., and Jensen, K. F. Autonomous discovery in the chemical sciences part i: Progress. *Angewandte Chemie International Edition*, 59(51):22858–22893, 2020. doi: https://doi.org/10.1002/anie.201909987. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201909987.

Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S., and Levy, O. The high-throughput highway to computational materials design. *Nature Materials*, 12(3):191–201, Mar 2013. ISSN 1476-4660. doi: 10.1038/nmat3568. URL https://doi.org/10.1038/nmat3568.

Dick, S. and Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nature Communications*, 11(1): 3509, Jul 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17265-7. URL https://doi.org/10.1038/s41467-020-17265-7.

Duan, C., Janet, J. P., Liu, F., Nandy, A., and Kulik, H. J. Learning from failure: Predicting electronic structure calculation outcomes with machine learning models. *Journal of Chemical Theory and Computation*, 15(4):2331–2345, 2019. doi: 10.1021/acs.jctc.9b00057. URL https://doi.org/10.1021/acs.jctc.9b00057. PMID: 30860839.

Duan, C., Chen, S., Taylor, M. G., Liu, F., and 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. doi: 10.1039/D1SC03701C. URL http://dx.doi.org/10.1039/D1SC03701C.

Flaser, B. M., Guo, Y., Riplinger, C., Tuczek, F., and Neese, F. Detailed pair natural orbital-based coupled cluster studies of spin crossover energetics. *Journal of Chemical Theory and Computation*, 16(4):2224–2235, 2020. doi: 10.1021/acs.jctc.9b01109. URL https://doi.org/10.1021/acs.jctc.9b01109. PMID: 32196337.

Grisafi, A., Fabrizio, A., Meyer, B., Wilkins, D. M., Corminboeuf, C., and Ceriotti, M. Transferable machine-learning model of the electron density. *ACS Central Science*, 5(1):57–64, 2019. doi: 10.1021/acscentsci.8b00551. URL https://doi.org/10.1021/acscentsci.8b00551. PMID: 30693325.

Hay, P. J. and Wadt, W. R. Ab initio effective core potentials for molecular calculations. potentials for the transition metal atoms sc to hg. *The Journal of Chemical Physics*, 82(1):270–283, 1985. doi: 10.1063/1.448799. URL https://doi.org/10.1063/1.448799.

Hohenberg, P. and Kohn, W. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964. doi: 10.1103/PhysRev.136.B864. URL https://link.aps.org/doi/10.1103/PhysRev.136.B864.

Husch, T., Sun, J., Cheng, L., Lee, S. J. R., and Miller, T. F. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *The Journal of Chemical Physics*, 154 (6):064108, 2021. doi: 10.1063/5.0032362. URL https://doi.org/10.1063/5.0032362.

Ioannidis, E. I., Gani, T. Z. H., and Kulik, H. J. molsimplify: A toolkit for automating discovery in inorganic chemistry. *Journal of Computational Chemistry*, 37 (22):2106–2117, 2016. doi: https://doi.org/10.1002/jcc.24437. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.24437.

Janesko, B. G. Replacing hybrid density functional theory: motivation and recent advances. *Chem. Soc. Rev.*, 50: 8470–8495, 2021. doi: 10.1039/D0CS01074J. URL http://dx.doi.org/10.1039/D0CS01074J.

Jiang, W., DeYonker, N. J., Determan, J. J., and Wilson, A. K. Toward accurate theoretical thermochemistry of first row transition metal complexes. *The Journal of Physical Chemistry A*, 116(2):870–885, 2012. doi: 10.1021/jp205710e. URL https://doi.org/10.1021/jp205710e. PMID: 22107449.

Keith, J. A., Vassilev-Galindo, V., Cheng, B., Chmiela, S., Gastegger, M., Müller, K.-R., and Tkatchenko, A. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical Reviews*, 121(16):9816–9872, 2021. doi: 10.1021/acs.chemrev.1c00107. URL https://doi.org/10.1021/acs.chemrev.1c00107. PMID: 34232033.

Kirkpatrick, J., McMorrow, B., Turban, D. H. P., Gaunt, A. L., Spencer, J. S., Matthews, A. G. D. G., Obika, A., Thiry, L., Fortunato, M., Pfau, D., Castellanos, L. R., Petersen, S., Nelson, A. W. R., Kohli, P., Mori-SÃ¡nchez, P., Hassabis, D., and Cohen, A. J. Pushing the frontiers of density functionals by solving the fractional electron problem. *Science*, 374(6573):1385–1389, 2021. doi: 10.1126/science.abj6511. URL https://www.science.org/doi/abs/10.1126/science.abj6511.

Kohn, W. and Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140: A1133–A1138, Nov 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133.

Lee, C., Yang, W., and Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988. doi: 10.1103/PhysRevB.37.785. URL https://link.aps.org/doi/10.1103/PhysRevB.37.785.

Liu, F., Yang, T., Yang, J., Xu, E., Bajaj, A., and Kulik, H. J. Bridging the homogeneous-heterogeneous divide: Modeling spin for reactivity

in single atom catalysis. *Frontiers in Chemistry*, 7, 2019. ISSN 2296-2646. doi: 10.3389/fchem.2019.00219. URL https://www.frontiersin.org/article/10.3389/fchem.2019.00219.

Ma, H., Narayanaswamy, A., Riley, P., and Li, L. Evolving symbolic density functionals, 2022. URL https://arxiv.org/abs/2203.02540.

Mardirossian, N. and Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics*, 115(19):2315–2372, 2017. doi: 10.1080/00268976.2017.1333644. URL https://doi.org/10.1080/00268976.2017.1333644.

Margraf, J. T. and Reuter, K. Pure non-local machine-learned density functional theory for electron correlation. *Nat. Comm.*, 12:344, Jan 2021. doi: 10.1038/s41467-020-20471-y. URL https://doi.org/10.1038/s41467-020-20471-y.

McAnanama-Brereton, S. and Waller, M. P. Rational density functional selection using game theory. *Journal of Chemical Information and Modeling*, 58(1):61–67, 2018. doi: 10.1021/acs.jcim.7b00542. URL https://doi.org/10.1021/acs.jcim.7b00542. PMID: 29257687.

Nandy, A., Duan, C., Janet, J. P., Gugler, S., and Kulik, H. J. Strategies and software for machine learning accelerated discovery in transition metal chemistry. *Industrial & Engineering Chemistry Research*, 57(42):13973–13986, 2018. doi: 10.1021/acs.iecr.8b04015. URL https://doi.org/10.1021/acs.iecr.8b0401,.

Nandy, A., Duan, C., Taylor, M. G., Liu, F., Steeves, A. H., and Kulik, H. J. Computational discovery of transition-metal complexes: From high-throughput screening to machine learning. *Chemical Reviews*, 121(16):9927–10000, 2021. doi: 10.1021/acs.chemrev.1c00347. URL https://doi.org/10.1021/acs.chemrev.1c00347. PMID: 34260198.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL https://dl.acm.org/doi/10.5555/3454287.3455008.

Pederson, R., Kalita, B., and Burke, K. Machine learning and density functional theory, 2022. URL https://arxiv.org/abs/2205.01591.

Perdew, J. P. and Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings*, 577(1):1–20, 2001. doi: 10.1063/1.1390175. URL https://aip.scitation.org/doi/abs/10.1063/1.1390175.

Pritchard, B. P., Altarawy, D., Didier, B., Gibson, T. D., and Windus, T. L. New basis set exchange: An open, up-to-date resource for the molecular sciences community. *Journal of Chemical Information and Modeling*, 59(11):4814–4820, 2019. doi: 10.1021/acs.jcim.9b00725. URL https://doi.org/10.1021/acs.jcim.9b00725. PMID: 31600445.

Qiao, Z., Welborn, M., Anandkumar, A., Manby, F. R., and Miller, T. F. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of Chemical Physics*, 153(12):124111, 2020. doi: 10.1063/5.0021955. URL https://doi.org/10.1063/5.0021955.

Rosen, A. S., Iyer, S. M., Ray, D., Yao, Z., Aspuru-Guzik, A., Gagliardi, L., Notestein, J. M., and Snurr, R. Q. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021. ISSN 2590-2385. doi: https://doi.org/10.1016/j.matt.2021.02.015. URL https://www.sciencedirect.com/science/article/pii/S2590238521000709.

Saunders, V. R. and Hillier, I. H. A "level–shifting" method for converging closed shell hartree–fock wave functions. *International Journal of Quantum Chemistry*, 7(4):699–705, 1973. doi: https://doi.org/10.1002/qua.560070407. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.560070407.

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 Jr., J. W., Song, C., Titov, A. V., Ufimtsev, I. S., Wang, L.-P., and Martínez, T. J. Terachem: A graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics. *WIREs Computational Molecular Science*, 11(2):e1494, 2021. doi: https://doi.org/10.1002/wcms.1494. URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1494.

Shee, J., Arthur, E. J., Zhang, S., Reichman, D. R., and Friesner, R. A. Phaseless auxiliary-field quantum monte carlo on graphical processing units. *Journal of Chemical Theory and Computation*, 14(8):4109–4121, 2018. doi: 10.1021/acs.jctc.8b00342. URL https://doi.org/10.1021/acs.jctc.8b00342. PMID: 29897748.

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., Sokolov, A. Y., Patkowski, K., DePrince, A. E., Bozkaya, U., King, R. A., Evangelista, F. A., Turney, J. M., Crawford, T. D., and Sherrill, C. D. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *The Journal of Chemical Physics*, 152(18):184108, 2020. doi: 10.1063/5.0006002. URL https://doi.org/10.1063/5.0006002.

Stephens, P. J., Devlin, F. J., Chabalowski, C. F., and Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *The Journal of Physical Chemistry*, 98(45): 11623–11627, 1994. doi: 10.1021/j100096a001. URL https://doi.org/10.1021/j100096a001.

Ufimtsev, I. S. and Martinez, T. J. Quantum chemistry on graphical processing units. 3. analytical energy gradients, geometry optimization, and first principles molecular dynamics. *Journal of Chemical Theory and Computation*, 5(10):2619–2628, 2009. doi: 10.1021/ct9003004. URL https://doi.org/10.1021/ct9003004. PMID: 26631777.

Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., and Müller, K.-R. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111. URL https://doi.org/10.1021/acs.chemrev.0c01111. PMID: 33705118.

Wang, L.-P. and Song, C. Geometry optimization made simple with translation and rotation coordinates. *The Journal of Chemical Physics*, 144(21):214108, 2016. doi: 10.1063/1.4952956. URL https://doi.org/10.1063/1.4952956.

6. Reply to reviewers' comments

6.1. Previous related works

Most of the efforts of applying machine learning on DFT have been made on developing new, either numerical (Dick & Fernandez-Serra, 2020; Qiao et al., 2020; Chen et al., 2021; Kirkpatrick et al., 2021) or symbolic (Ma et al., 2022) density functional approximations (DFAs). The only exception is is the work from McAnanama-Brereton and Waller (McAnanama-Brereton & Waller, 2018), where they adopted game theory set up for selecting a combination of DFA and basis set with an attempt to reach a balance between computational cost and accuracy. However, they only used chemical compositions as the inputs for their models and studied relatively small chemical systems.

6.2. Model architecture: Why Behler–Parrinello but not equivariant?

Since both the density features inputs (p_L^A in Eq. 3) and vertical spin splitting differences ($|\Delta\Delta E_{\mathrm{H-L}}|$) outputs are already scalars (i.e., invariant), using equivalent architectures will not generally improve model performance. However, it would be interesting to directly start with C_Q^A in Eq. 2 and build equivariant graph neural networks to predict $|\Delta\Delta E_{\mathrm{H-L}}|$ because C_Q^A contains higher-order tensors that correspond to basis functions with $L > 1$. It would be an interesting direction to explore in future study.

6.3. Why do we call our models transfer learning models?

We call our model transfer learning (TL) models simply because we are learning a property difference (i.e., the absolute difference between a DFA f and our reference results from DLPNO-CCSD(T), $|\Delta\Delta E_{\mathrm{H-L}}[f]|$). To be more precise, one can also refer our learning task as Δ-learning.

A. Standard deviation of the absolute error for the top-5 DFAs

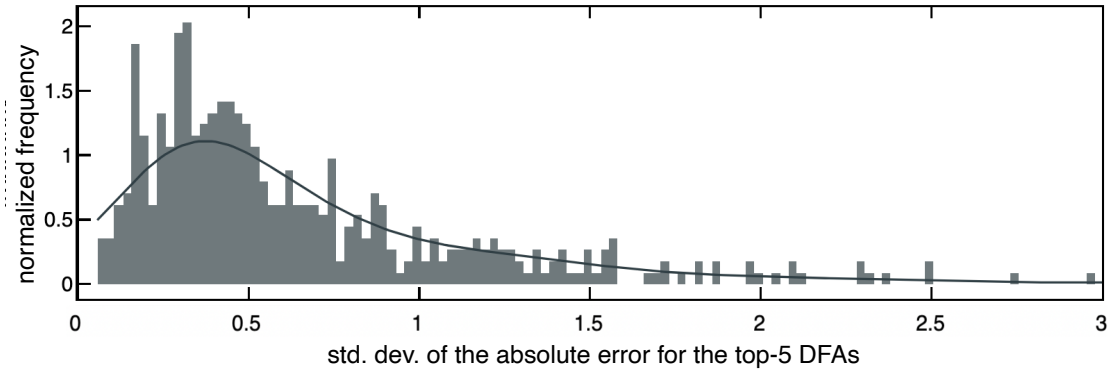

Figure A1. Normalized distributions of standard deviation (std. dev.) of $|\Delta\Delta E_{\mathrm{H-L}}|$ for the top-5 DFAs. A kernel density estimate (black) is also shown. It can be observed that the difference of $|\Delta\Delta E_{\mathrm{H-L}}|$ for the top-5 DFAs can be small (i.e to 1 kcal/mol) for many TMCs.

B. 20 ligands used in this work

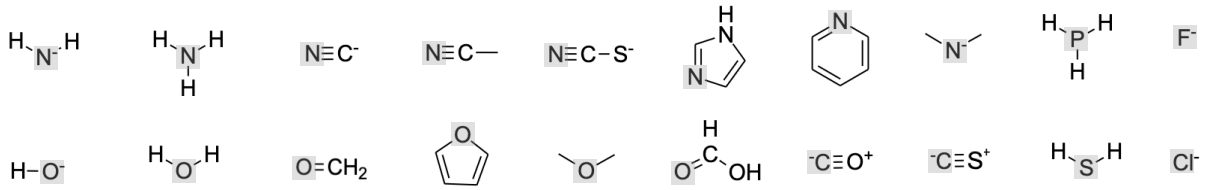

Figure A2. The 20 small ligands resembling those from the spectrochemical series that are considered in this work. The coordinating atom is shaded in gray for each ligand.

C. Summary of properties of the original 23 DFAs

Table A1. Summary of 23 functionals in the original work of Duan *et al.* (Duan et al., 2021), including their rungs on "Jacob's ladder" of DFT, Hartree–Fock (HF) exchange fraction, long-range correction (LRC) range-separation parameter (bohr^{-1}), MP2 correlation fraction, and whether empirical (i.e., D3) dispersion correction is included.

DFA	type	exchange type	HF exchange type	LRC RS parameter (bohr^{-1})	MP2 correlation	D3 dispersion
BP86	GGA	GGA	–	–	–	No
BLYP	GGA	GGA	–	–	–	No
PBE	GGA	GGA	–	–	–	No
TPSS	meta-GGA	meta-GGA	–	–	–	No
SCAN	meta-GGA	meta-GGA	–	–	–	No
M06-L	meta-GGA	meta-GGA	–	–	–	No
MN15-L	meta-GGA	meta-GGA	–	–	–	No
B3LYP	GGA hybrid	GGA	0.200	–	–	No
B3P86	GGA hybrid	GGA	0.200	–	–	No
B3PW91	GGA hybrid	GGA	0.200	–	–	No
PBE0	GGA hybrid	GGA	0.250	–	–	No
ωB97X	RS hybrid	GGA	0.158	0.300	–	No
LRCω-PBEh	RS hybrid	GGA	0.200	0.200	–	No
TPSSh	meta-GGA hybrid	meta-GGA	0.100	–	–	No
SCAN0	meta-GGA hybrid	meta-GGA	0.250	–	–	No
M06	meta-GGA hybrid	meta-GGA	0.270	–	–	No
M06-2X	meta-GGA hybrid	meta-GGA	0.540	–	–	No
MN15	meta-GGA hybrid	meta-GGA	0.440	–	–	No
B2GP-PLYP	double hybrid	GGA	0.650	–	0.360	No
PBE0-DH	double hybrid	GGA	0.500	–	0.125	No
DSD-BLYP-D3BJ	double hybrid	GGA	0.710	–	1.000	Yes
DSD-PBEB95-D3BJ	double hybrid	GGA	0.660	–	1.000	Yes
DSD-PBEP86-D3BJ	double hybrid	GGA	0.690	–	1.000	Yes

D. Summary of properties of the additional DFAs

Table A2. Summary of the additional 25 functionals compared to Duan *et al.* (Duan et al., 2021), including their rungs on "Jacob's ladder" of DFT, Hartree–Fock (HF) exchange fraction, long-range correction (LRC) range-separation parameter (bohr^{-1}), MP2 correlation fraction, whether empirical (i.e., D3) dispersion correction is included.

DFA	type	exchange type	HF exchange type	LRC RS parameter (bohr^{-1})	MP2 correlation	D3 dispersion
BLYP:10%	GGA hybrid	GGA	0.100	–	–	No
BLYP:20%	GGA hybrid	GGA	0.200	–	–	No
BLYP:30%	GGA hybrid	GGA	0.300	–	–	No
BLYP:40%	GGA hybrid	GGA	0.400	–	–	No
BLYP:50%	GGA hybrid	GGA	0.500	–	–	No
PBE:10%	GGA hybrid	GGA	0.100	–	–	No
PBE:20%	GGA hybrid	GGA	0.200	–	–	No
PBE:30%	GGA hybrid	GGA	0.300	–	–	No
PBE:40%	GGA hybrid	GGA	0.400	–	–	No
PBE:50%	GGA hybrid	GGA	0.500	–	–	No
SCAN:10%	meta-GGA hybrid	meta-GGA	0.100	–	–	No
SCAN:20%	meta-GGA hybrid	meta-GGA	0.200	–	–	No
SCAN:30%	meta-GGA hybrid	meta-GGA	0.300	–	–	No
SCAN:40%	meta-GGA hybrid	meta-GGA	0.400	–	–	No
SCAN:50%	meta-GGA hybrid	meta-GGA	0.500	–	–	No
M06-L:10%	meta-GGA hybrid	meta-GGA	0.100	–	–	No
M06-L:20%	meta-GGA hybrid	meta-GGA	0.200	–	–	No
M06-L:30%	meta-GGA hybrid	meta-GGA	0.300	–	–	No
M06-L:40%	meta-GGA hybrid	meta-GGA	0.400	–	–	No
M06-L:50%	meta-GGA hybrid	meta-GGA	0.500	–	–	No
MN15-L:10%	meta-GGA hybrid	meta-GGA	0.100	–	–	No
MN15-L:20%	meta-GGA hybrid	meta-GGA	0.200	–	–	No
MN15-L:30%	meta-GGA hybrid	meta-GGA	0.300	–	–	No
MN15-L:40%	meta-GGA hybrid	meta-GGA	0.400	–	–	No
MN15-L:50%	meta-GGA hybrid	meta-GGA	0.500	–	–	No

E. Theoretical bound of recommender approach with 48 DFAs

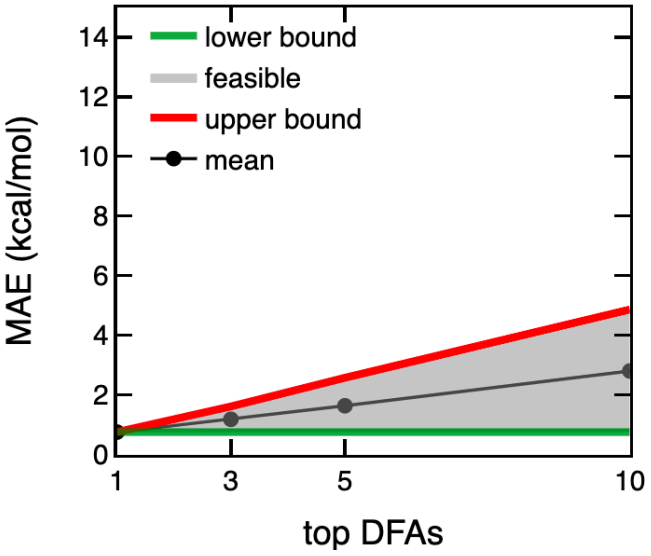

Figure A3. MAE of $|\Delta\Delta E_{\mathrm{H-L}}|$ constrained by only selecting a random top-n DFA (black dots and solid line) among the 48 DFAs. The lower bound (green line), upper bound (red line), and the feasible area (gray shaded) are also shown when the DFA selection is constrained in top-n choices.

F. MAEs derived from the 48 DFAs compared to DLPNO-CCSD(T)

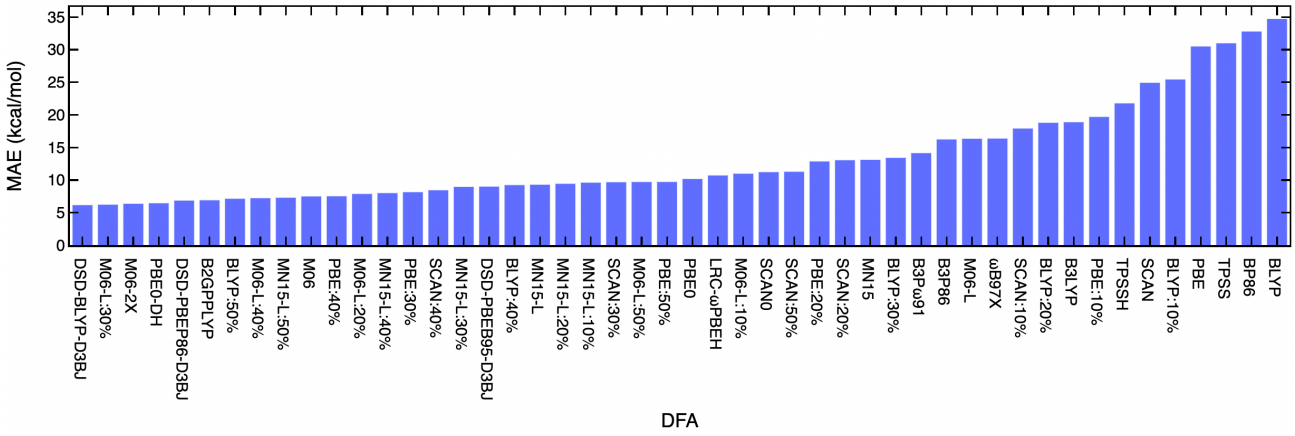

Figure A4. MAE of $|\Delta\Delta E_{\mathrm{H-L}}|$ for the 48 DFAs of the *VSS-452* set. The DFAs are sorted in an ascending order of their raw MAEs.

G. Recommender rank-ordering performance with 23 DFAs

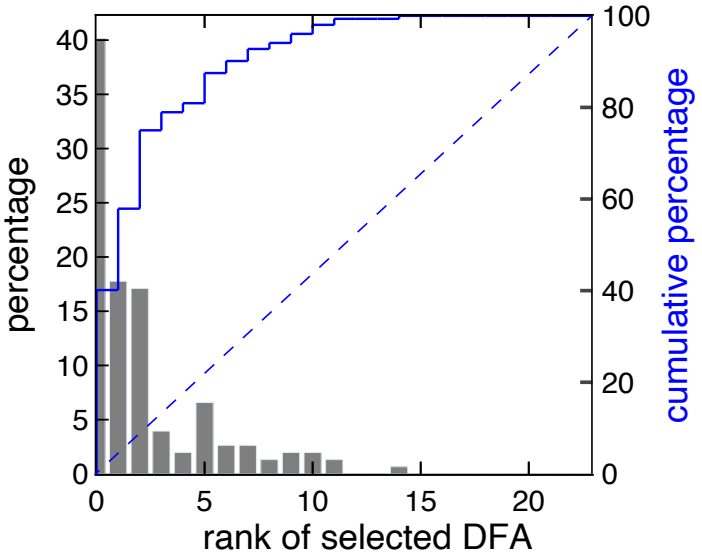

Figure A5. Normalized distribution of the rank for selected DFA using our recommender approach, with the cumulative percentage (blue solid line) shown according to the axis on the right. The cumulative curve for a random guess (blue dashed line) is also shown.