Transferable diversity – a data-driven representation of chemical space

Tim Gould,[1] Bun Chan,[2] Stephen G. Dale,[1,3] and Stefan Vuckovic[4]
[1)] *Queensland Micro- and Nanotechnology Centre, Griffith University, Nathan, Qld 4111, Australia*
[2)] *Graduate School of Engineering, Nagasaki University, Bunkyo 1-14, Nagasaki 852-8521, Japan*
[3)] *Institute of Functional Intelligent Materials, National University of Singapore, 4 Science Drive 2, Singapore 117544*
[4)] *Department of Chemistry, University of Fribourg, Fribourg, Switzerland.*[a)]

Transferability, especially in the context of model generalization, is a paradigm of all scientific disciplines. However, the rapid advancement of machine learned model development threatens this paradigm, as it can be difficult to understand how transferability is embedded (or missed) in complex models. A rigorous understanding of transferability in chemical representation remains an open problem. To this end, we introduce a transferability assessment tool and apply it to a controllable data-driven model for developing density functional approximations (DFAs), an indispensable tool in everyday chemistry research. We reveal that human intuition in the curation of training data introduces chemical biases that can hamper the transferability of data-driven DFAs. We use our transferability assessment tool to motivate *transferability principles*; one of which introduces the key concept of *transferable diversity*. Finally, we use transferability principles to propose data curation strategies for general-purpose machine learning models in chemistry.

I. INTRODUCTION

For the past half-century, Density Functional Theory (DFT)[2,3] has made an unparalleled impact across a range of scientific and engineering disciplines. Nowadays, this impact is greater than ever, as evidenced by the large portion of the world's supercomputing power being consumed by DFT simulations[4,5]. In recent years, machine learning (ML) is transforming nearly all scientific disciplines, and DFT is no exception[6,7]. Recent advancements in ML-based DFT[8] signal the beginning of a race to discover the DFT 'holy grail' or at least a highly effective surrogate thereof – holding promise to revolutionize the entire field of chemistry[9]. Building on this momentum, ML of density functional approximations (DFAs) is enabling rapid advances in the predictive quality of quantum chemistry, by enhancing the practical cost and quality benefits of DFT by empirical strategies based on "big data" training sets[10,11].

[a)] Electronic mail: stefan.vuckovic@unifr.ch

The assumption that a DFA is transferable is implicit in every DFA developed for general use, and this culture of universal density functionals has been readily adopted by the machine-learned DFA (ML-DFA) community. The data-driven approach of ML-DFAs allows for examination of the feedback loop between the ML-DFA training sets, and the overall ML-DFA performance post training. To conduct this analysis this work will introduce a transferability assessment tool that involves training an ML-DFA functional on a test set **A**, and assessing the performance of that functional on test set **B**, abbreviated to **B@A** (or [test set]@[training set]), more details given in Section II. Achieving high performance on **A@A** is often straightforward, as we can always increase model flexibility by adding more parameters. However, the true challenge lies in ensuring that the ML-DFA is transferable to **B** (i.e. **B@A**), meaning it genuinely learns rather than simply memorizes patterns in **A**. This task prompts a range of questions.

> First, a key and outstanding problem is how do we create **A** to *target* transferability of our ML-DFA model to a wide range of chemical physics?
>
> Is *more* always *more*? (i.e. does increasing the size of set **A** always improve **B@A**?)
>
> Can we quantify how difficult test set **B** is for a model trained on **A**? (e.g. can we quantify the intuition that training a model on atomisation energies of alkanes better predicts atomisation energies of alkenes than transition metal barrier heights?);
>
> Can we quantify the 'distance' or difficulty level between training set **A** and test set **B**?
>
> Does the inclusion of well-known or well-studied chemical structures in **A** limit the model's transferability to unseen chemistry?

After all, the ultimate goal of DFT simulations is not just to confirm and rationalize what we already know from experiments but to accurately predict (transfer to) unseen chemistry and unperformed experiments[9].

In using the *transferability assessment tool* (TAT) to explore the above questions, we show that simply expanding the number and/or type of chemical systems in a given training set is insufficient to improve an ML-DFA in general (Section 3). By contrast, we reveal three *transferability principles* that do embed transferability in a

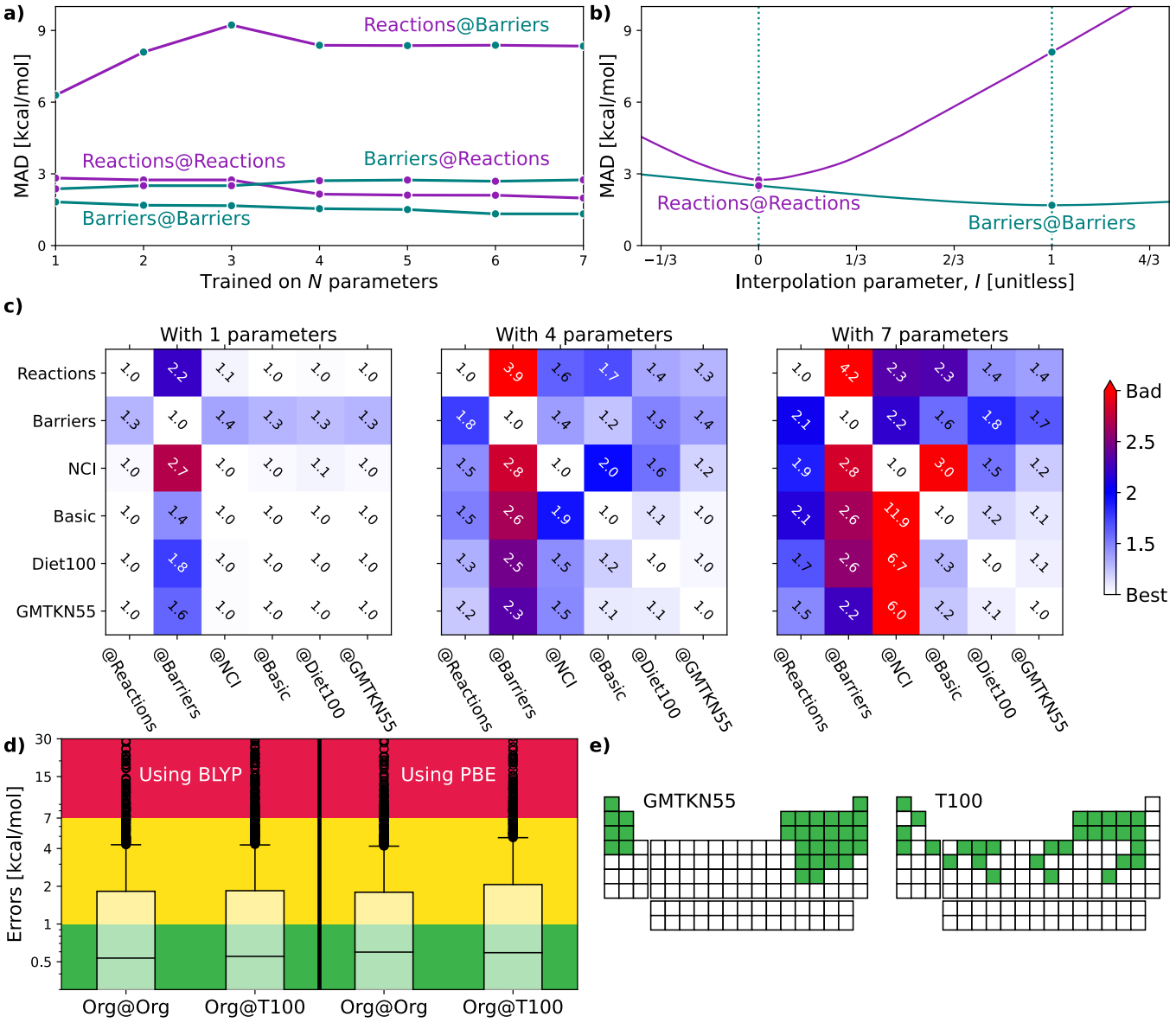

FIG. 1. **a)** Errors for XYZ-DFAs with 1–7 parameters applied to subsets covering reaction and barrier chemistry – line colour indicates the test set and dot colour the training set. **b)** Errors for XYZ_2 (2-parameter double hybrid DFA) along a linear interpolation path ($\alpha = I\alpha_{\mathbf{Barriers}} + (1-I)\alpha_{\mathbf{Reactions}}$ and similar for β) between the Reactions ($I = 0$) and Barriers ($I = 1$) minima. **c)** Transferability matrices between selected benchsets for XYZ_1, XYZ_4 and XYZ_7 (double hybrids with varying parameter number). **d)** Boxplots with XYZ_7 (one with BLYP and other with PBE *semilocal* parts) errors for a large organic database (**Org.**=**GMTKN55**[1]) with parameters trained on the whole database and on the **T100** benchset (designed from our transferability principles). **e)** Periodic tables showing the elements (green) included in **GMTKN55** and **T100**.

benchset, taken together, and that may therefore be used in the curation of better training benchsets. Most importantly, we introduce the concept of *transferable diversity* to our training set design – meaning we aim for our training set to yield good transferability to a diverse range of chemical behaviours. We use these principles to design the **T100** benchset (final part of Section 3). Ultimately, this work leaves us positioned to recommend a strategy, detailed in the Conclusions, for the development of new benchsets that are designed to embed transferability into ML-DFAs.

The following sections will delve into specific details. For now, it suffices to mention that as a controllable model, we use a *double-hybrid functional form*[12,13], defined by one[12] to seven[14] parameters. This model facilitates the construction of thousands of data-driven density functional approximations, effectively illustrating the utility and analytic power of our TAT. Some key findings of our study are presented in Fig. 1. Fig. 1(a) focuses on our model's efficacy in predicting reaction energies and

barrier heights – crucial for calculating thermodynamics and kinetics, respectively[1]. We train on reaction energies and test on barrier heights (**Barriers@Reactions**), and then reverse the sets (**Reactions@Barriers**). From Fig. 1(a) it is clear that our model excels in transferring from reaction energies to barrier heights (thermodynamic to kinetic parameters), but not the other way around. The reason for this asymmetry becomes apparent when we look at the shapes of the cost functions for our two-parameter model and compare the values at their respective minima to those at each other's minima, as shown in Fig. 1(b).

Fig. 1(c) introduces the transferability matrix $T_{\mathbf{B@A}}$, a unitless measure precisely defined as how well a given model trained on **A** performs for **B** (**B@A**) relative to the accuracy limit of that model for **A@A**. Unlike in Fig. 1(a), which focuses solely on the transferability between reaction energies and barrier heights, Fig. 1(c) includes multiple classes of organic chemical processes[1]. The matrix provides insights into: (i) transferability for each $T_{\mathbf{B@A}}$ pair; (ii) asymmetry in transferabilities, as shown by differences in $T_{\mathbf{B@A}}$ and $T_{\mathbf{A@B}}$ values; (iii) the rate at which transferability decreases with the increasing number of parameters for different **B@A** pairs; (iv) the chemical classes most transferable to and most transferable from. Transferability matrices are thus a key foundation of our TAT.

Fig 1(d) demonstrates that two different flavours of our seven-parameter model[14], trained on the **T100** benchset (of 100 processes carefully curated around transferability principles of reaction, elemental and transferable diversity), perform as well as their accuracy limits when tested on the extensive "general-main group thermochemistry, kinetics and noncovalent interactions" (GMTKN55) database of 1505 organic processes[1]. This confirms that transferability principles effectively enhance the model's applicability to larger datasets. Fig 1(e) further highlights the greater elemental diversity in our small **T100** compared to large **GMTKN55**, as it covers a far broader range of groups in the periodic table, despite being fifteen times smaller.

II. TRANSFERABILITY ASSESSMENT TOOL

To measure transferability from **A** to **B**, we introduce a two-set error $\mathrm{MAD}_{\mathbf{B@A}}$, which is the mean absolute deviation (MAD) on test set **B** for a DFA trained on **A**. We then formulate a unitless transferability matrix:

$$T_{\mathbf{B@A}} = \frac{\mathrm{MAD}_{\mathbf{B@A}} + \eta}{\mathrm{MAD}_{\mathbf{B@B}} + \eta} \geq 1 \, . \tag{1}$$

$\eta = 0.01$ kcal/mol regularizes results for small energies. By definition, $T_{\mathbf{A@A}} = 1$ (the case of perfect transferability) and minimization principles dictate that $T_{\mathbf{B@A}} \geq 1$, with larger values indicating poorer transferability. The $T_{\mathbf{B@A}}$ transferability matrix therefore quantifies the performance of a model trained on **A** when applied to **B**, normalized by the model's inherent accuracy limit for **B**.

To demonstrate our TAT, we use a double hybrid (DH) family of DFAs, called XYZ_p[14], where p is the number of empirical parameters varying from one[15] to seven[14] (see Methods for the functional forms). Varying the number of parameters lets us vary the level of empiricism, and thus emulate varying degrees of "machine learning". The DH form is chosen for its generality, as it sits at the top of the current DFA Jacob's ladder (a hierarchy of DFAs based on their mathematical complexity)[16,17]. This allows our DH forms to reduce to functional forms from lower rungs of the ladder during parameter optimization. We use Hartree-Fock (HF) orbitals to calculate all energy terms, to prevent uncontrolled error cancellation of *functional-* and *density-driven errors* when building data-driven DFAs[15,18].

We are now ready to apply the TAT to real data, for the purpose of revealing limitations of existing protocols, and uncovering key principles that enhance transferability and performance across diverse systems.

III. RESULTS

Before beginning a detailed analysis of transferability, consider a "minimally-empirical" approach in which a DFA is designed around several fundamental constraints, and then optimised over a small number of processes to determine any remaining parameters. The case of XYZ_3[19] we train here on **G21IP** is a prototypical example of such a strategy. The 3-parameter XYZ form approximately satisfies various constraints by construction[19], and training on the 21 ionisation potentials in the benchset **G21IP**[1,20] fills in the missing gaps.

At first sight, this seems like an effective strategy: it yields $\mathrm{MAD}_{\mathbf{GMTKN55@G21\text{-}IP}} = 1.91$ kcal/mol across the extensive **GMTKN55** organic benchset, not far from the optimal $\mathrm{MAD}_{\mathbf{GMTKN55@GMTKN55}} = 1.84$ kcal/mol achieved by full optimization of the three XYZ_3 parameters over **GMTKN55**. Using Eq. (1), we find a transferability matrix element of $T_{\mathbf{GMTKN55@G21IP}} = \frac{1.91+0.01}{1.84+0.01} = 1.04$, indicating **G21IP**'s high transferability to **GMTKN55**. This compares to, for example, $T_{\mathbf{GMTKN55@W4\text{-}11}} = 1.7$ obtained when training on the **W4-11**[1,21] set of atomization energies.

So what makes **G21IP** such an exceptionally good training set? We shall later see that the answer from our TAT is "nothing at all" and that this example calls for more transparency when it comes to selecting a training set in designing data-driven DFAs. But for more insightful answers, we must further analyze the nuances of transferability, which we will do through concrete examples in the following sections.

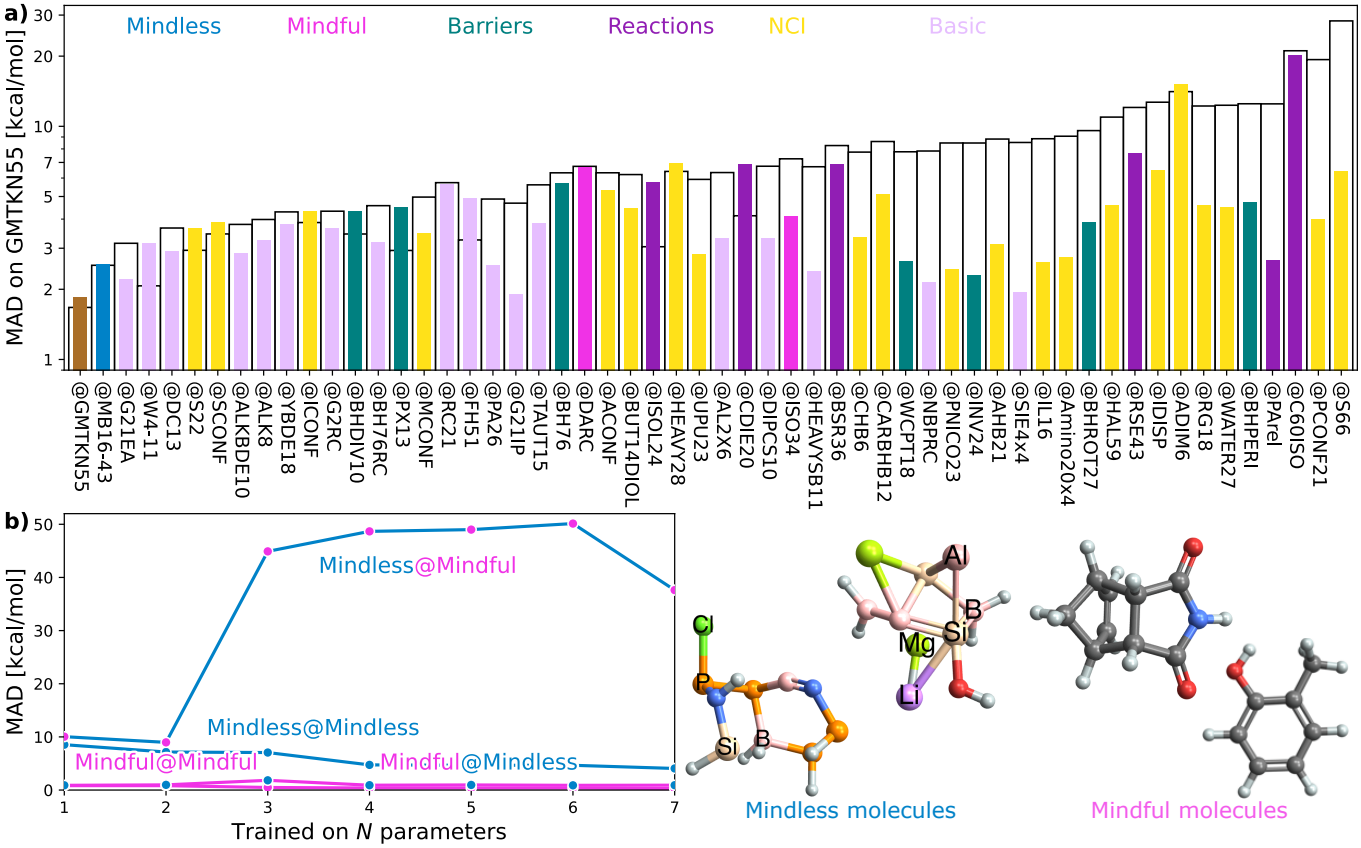

FIG. 2. Mean absolute deviation (MAD, log scale) for **GMTKN55@subset**, where **subset** is a subset of GMTKN55. Solid colours indicate XYZ_3, black lines indicate XYZ_7. The order reflects the MAD and absolute difference between XYZ_3 and XYZ_7. **b)** Errors for DFAs with 1–7 parameters applied to subsets covering mindless and mindful construction of benchmark set. Some example mindless and mindful molecules are shown at right.

Transferability concepts: motivation through organic chemistry

Our goal is motivate transferability principles that can be applied more broadly. As a first step, let us use the key concepts introduced in Section II to delve into the details of Fig. 1(a-c), focusing on transferability within the large GMTKN55 organic chemistry database.

Fig. 1(a) shows that training barrier heights (194 processes[1]) on reaction energies (243 processes[1]) performs nearly as well as training on barriers themselves. However, reaction energies perform poorly when trained on barriers, suggesting either barriers are easier to learn or that reactions are better for training. Fig. 1(b) explains this result and lets us pick the right conclusion for the case of a two-parameter XYZ_2 (the parameters being exact exchange fraction and MP2 correlation fractions). Errors in **Barriers** are rather insensitive to changes in parameters, meaning that picking sub-optimal parameters does not lead to major additional errors. Not so for errors in **Reactions**, where curvature is much sharper and, consequently, changing parameters rapidly worsens results. Therefore **Barriers** are easier to learn than **Reactions**.

The $T_{\mathbf{B@A}}$ transferability matrices in Fig. 1 for XYZ_1, XYZ_4, and XYZ_7 show how transferability rapidly worsens as the number of model parameters increases. In the 1-parameter case, many $T_{\mathbf{B@A}}$ values are close to 1.0, indicating high transferability. Conversely, in the 7-parameter model, numerous entries exceed 3, implying performance three times worse than optimal. The upper 4×4 block highlights transferabilities among four test subsets: **Reactions**, **Barriers**, **NCI**, and **Basic**[1] (everything else, such as atomization energies, ionization potentials, proton/electron affinities, etc.). The block reveals that **Reactions** is the most transferable training set, indicated by the smallest values in its column. Conversely, **Basic** appears to be the most challenging to transfer to, as evidenced by the largest values in its row. In the Supplementary Information (SI), we show $T_{\mathbf{B@A}}$ by further breaking down GMTKN55's subsets (Supp. Figs S6–S8). Interestingly, within XYZ_1, reaction sets are more transferable to barriers than different barrier sets are to each other (Supp. Fig. S6).

Furthermore, Fig. 1(c) already challenges the obvious, and so far dominant in data-driven DFA development, strategy of increasing the size of datasets. **Diet100** (with 100 processes) does a much better job as a training set

than any of the larger ($\sim$ 250 processes) 'chemistry' subsets; and performs nearly as well as **GMTKN55** at predicting **Reactions**, **Barriers** and **Basis**. Unfortunately, the way **Diet100** was constructed offers no useful insights for improving transferability principles, although it does convincingly confirm that quality is more important than quantity.

Fortunately, GMTKN55 comprises 55 subsets, each representing (more-or-less) different types of chemistry. We can leverage this diversity to develop a better understanding of transferability and use it to create the **T100** set, explicitly engineered for high transferability, as hinted at in Fig. 1(d) and (e). We will revisit these two panels after elaborating on the essential principles that inform this set's design.

Transferability principle 1: Reduce human bias in the training set to achieve genuine reaction diversity

Consider a hypothetical experiment involving two distinct groups: chemistry students and art students. Given a molecular editor and specific drawing rules (e.g., use no more than 16 spheres in total and stick to the colors white, gray, blue, etc.), the optimized structures and benchmarked energies from their drawings would form the basis for two different empirical density functionals ('Art' and 'Chemistry' functionals). We will show that functionals trained on the art students' molecules would easily outperform those based on the chemistry students' designs. The latter group's chemical intuition is to blame, as it introduces unexpected biases in the data.

To begin, let us play a game where we optimize our DFA models for each of the 55 subsets within GMTKN55 and then assess how well each of the 55 resulting DFAs transfers to the full GMTKN55 database. Fig. 2(a) shows the key results from this game, displaying MADs for **GMTKN55@subset** from each of the 55 subsets, using 3- and 7-parameter models, XYZ_3 and XYZ_7. In most cases, MAD for XYZ_3 and XYZ_7 are vastly different, and even when they are not, MAD are very large. These indicators of poor transferability reflect the fact that different subsets capture different chemistry and do not represent the whole GMTKN55 in this specific transferability context.

Returning to our opening example, we see that **G21IP** performs well with XYZ_3 but poorly with XYZ_7 – its transferability is strongly influenced by the number of free parameters (Supp. Fig. S1 further highlights this point when both XYZ_3 and XYZ_7 are applied to noncovalent interactions). Indeed, **G21IP** is not unique in that regard – transferability for XYZ_7 is almost always worse than XYZ_3. Increasing parameters elevates the risk of overfitting, challenging us to identify datasets whose transferability remains robust despite additional parameters. While regularization strategies applied to a DFA form (through e.g., physical constraints) can enhance its transferability[22,23], our TAT has a different focus that complements this regularization strategy. Namely, Eq. (1) allows us to see how transferability varies with different training sets for *any* optimizable DFA form, enabling us to identify general principles for the design of training sets with improved transferability.

Transferability principle 1 is revealed by the standout performer in Fig. 2(a): MB16-43[1,24]. What is special about **MB16-43**? It is the only subset in GMTKN55 that is not biased toward chemical intuition or the limited chemical space it spans. Simply put, unlike the remaining 54 subsets, its structures have not been manually drawn by humans before undergoing geometry optimizations. Rather, MB16-43 avoids unnoticed human bias via "mindless" (more accurately, a clever random strategy) construction of molecules – we shall henceforth denote it as **Mindless** to stress this feature.

Fig. 2(b) shows that DFAs trained on **Mindless** (43 processes) predict good energies for a similarly-sized more **Mindful** (DARC+ISO34 with 48 processes covering Dies-Adler and isomerisation reaction energies[1]) selection of data. But, the reverse is not true – **Mindless@Mindful** has up to six-fold increases in errors compared to **Mindless@Mindless**. Our results thus confirm that mindless benchmarking achieves its goal of "[making] use of random elements constrained by systematic and controllable specifications to avoid unsystematic and uncontrolled criteria".[24] The small size of **Mindless** again stresses the importance of quality over quantity.

Furthermore, the transferability captured by **Mindless** is independent of both the **Mindful** dataset (Supp. Fig. S11) and the semilocal part of our functional (Supp. Fig. S12). We therefore see that **Mindless** captures *genuine* diversity of chemical interactions – i.e., it achieves transferability principle 1. In simpler terms, **Mindless** (art students) molecules yield far better functionals here than **Mindful** (chemistry students) ones.

Transferability principle 2: Ensure transferability to and from transition metal chemistry for elemental diversity

Modern technologies rely on most stable elements in the periodic table.[27] By contrast, two thirds of processes in GMTKN55 contain *only* C, H, N, O or F. This highlights a second limitation of the training data we have considered so far – a lack of elemental diversity. Improving elemental diversity is the most intuitive of the transferability principles, yet we shall see it still throws up some surprises.

As GMTKN55 completely excludes transition metals [Fig. 1(e) shows the elements of the periodic table that GMTKN55 covers], we turn to TMC151,[25] a 151-process benchset based around transition metal (TM) chemistry, which lets us introduce some inorganic chemistry into our game and supplement the results of GMTKN55. Despite the sparsity of TM benchmarking (151 versus 1505

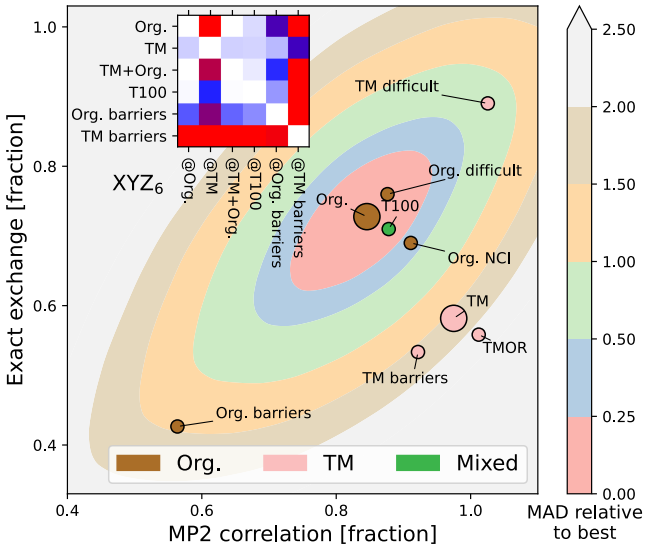

FIG. 3. Optimal values for the two-parameter model (markers) for organic (Org.=GMTKN55[1]) and transition metal (TM=TMC151[25]) processes, and subsets thereof (e.g., TMOR = metal-organic reactions[25,26]). Also shows the MAD (contours) of organic processes as a function of the two parameters, relative to the optimal value. **Inset**: XYZ_6 transferability matrix for selected Org. and TM sets.

processes) we are nonetheless able to develop an understanding of transferability between main group and TM chemistry.

Fig. 3 reveals that training on main group elements is not a good strategy for predicting transition metal chemistry, or vice versa, even in the simple XYZ_2 model. The optimal parameters for TM sets live in a different region of the parameter space compared to those for the main group sets. Transferability from TMC151 (denoted **TM** to stress its focus on transition metals) to GMTKN55 (denoted **Org.** to stress organic chemistry) is very poor, as can be seen from the contour plots (for XYZ_2) and inset transferability matrix (for XYZ_6). Simply adding the two sets (**TM+Org.**) improves results in general, but still has transferability issues for both **Org. Barriers** and **TM Barriers** (see inset).

In view of the extremely poor transferability of DFAs trained on TMs to Org., adding elemental diversity (e.g., molecules with $3d$ elements) to a main-group training set could ruin the good accuracy of DFAs for Org (further highlighted in Supp. Fig. S19). However, as we shall soon see, this risk is completely eliminated once the training set is diversified in a manner that explicitly favors transferability. Thus, what we seek in a training set is not just elemental diversity, as this can come with drawbacks. Instead, what we want in the training set and what we advocate for is a balance between genuine reaction diversity, elemental diversity and *transferable (chemical) diversity* – to be defined soon. **Mindless** gave us our first hint that human intuition may be counterproductive to such a goal. We will now proceed to show how it can be achieved more systematically.

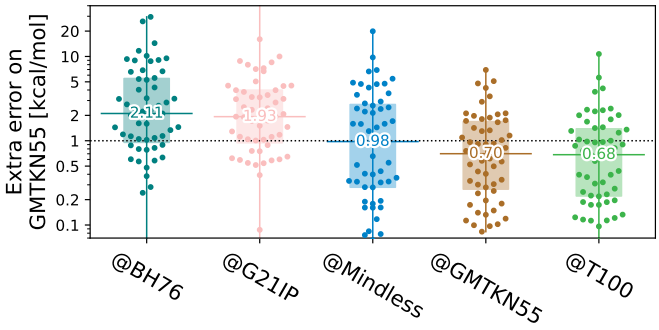

FIG. 4. Transferability energy (log scale – note magnitude of outliers) of the 55 GMTKN55 subsets trained on different benchsets, for a 7-parameter XYZ-DFA. Beeswarm plots show the 55 benchsets, horizontal lines and numbers indicate the median, boxes indicate the 1st–3rd quartiles.

Transferability principle 3: Transferable diversity at work in the T100 set

After adding some TM into the game, we are ready to return to the last two panels of Fig. 1, where we showed some results for our new benchset, **T100**. The most important feature of **T100** is that it is explicitly designed around *three* transferability principles: 1) it randomly selects chemical processes from **TM+Org.** to yield genuine reaction diversity; 2) it includes a bias in construction toward genuine elemental diversity; 3) it is optimized to improve average transferability in the XYZ_1, XYZ_4 and XYZ_7 functional forms, giving us a final ML-DFA that is explicitly designed to give good transferability. The principles behind the first two have already been discussed. Full details are in Methods and SI Sec. S2.

Importantly, the third design feature for **T100** provides an implicit definition of transferable diversity: a benchset has transferable diversity if an approach trained on it is transferable to (i.e. performs well on) other benchsets. Transferable diversity is therefore the property that "chemistries" are appropriately weighted or proportioned in the benchset, so as to improve predictions without accidental bias. **Mindless** has good transferable diversity, but less elemental diversity than **T100**.

The boxplots in Fig. 1(d) indicate that XYZ_7 trained solely on the 100 chemical processes in **T100** performs nearly as well as when trained on all 1505 **GMTKN55** processes. This holds for both the BLYP-based XYZ_7 model used in T100 creation; and a PBE-based XYZ_7 variant that has not been *seen* during the construction of T100. The differences between the two are described in Methods. Fig. 1(e) shows that **T100** covers a far *broader range* of periodic table groups than **GMTKN55**, despite the two containing similar *numbers* of elements. Figs 1(d,e) thus reveal the effectiveness of using transferability principles in data curation.

The results shown in Fig. 4 highlight that the T100 optimisation strategy has very useful consequences for the *transferability energy cost*,

$$\Delta\mathrm{MAD}_{\mathbf{B@A}} := \mathrm{MAD}_{\mathbf{B@A}} - \mathrm{MAD}_{\mathbf{B@B}} \geq 0\,, \quad (2)$$

(i.e. the increase in MAD energy caused by training on the 'wrong' **A**) which supplements the transferability matrix, $T_{\mathbf{B@A}}$. **B** is any of the 55 subsets of GMTKN55 while **A** (listed below each figure) is the training benchset, used to optimise XYZ_7. We see that both **BH76**[1] and our old friend **G21IP** provide poor training data, leading to excess errors of over 1 kcal/mol in 75% of subsets. Thus, the poor results of Figure 2(a) are not caused by a small number of outliers, but are systematic.

By contrast, **T100** actually *out performs* **GMTKN55** when applied to diverse organic chemistry – albeit as a consequence of our choice to sample by set. This is despite being optimized to balance transferability between main group and TM chemistry [remember the periodic tables for the two sets shown in Fig. 1(e)]. Indeed, 75% of benchsets are predicted to within 2 kcal/mol of their optimal (self-trained) values and 50% are within 0.7 kcal/mol.

TABLE I. MAD (kcal/mol) for different datasets (rows) of the XYG_7 functional trained on the datasets given in columns. Results shown for BLYP- and r^{2}SCAN-based XYZ_7.

Set	@Self	@T100	@Mindless	@Mindful
		BLYP		
S66	0.18	0.34	0.33	0.32
W4-11	2.58	4.58	6.85	57.38
Water27	0.08	0.82	4.82	6.08
BH76	1.41	3.70	3.11	4.96
OrgDiff	5.41	7.59	8.87	37.24
ISOL24	0.36	1.36	1.65	0.86
TMB	1.21	4.83	5.75	4.37
		r^{2}SCAN		
S66	0.21	0.41	0.36	0.71
W4-11	2.41	3.46	4.43	32.25
Water27	0.06	1.36	0.98	5.35
BH76	1.77	3.13	3.10	4.77
OrgDiff	6.11	7.89	7.70	18.06
ISOL24	0.51	2.17	1.52	0.94
TMB	1.85	5.06	5.50	5.65

Table I reports results for 7-parameter DFAs tested on a diverse list of example benchsets; and reveals that, $XYZ_7(\text{@}\mathbf{T100}) = 0.853E_\mathrm{x}^\mathrm{HF} - 0.024E_\mathrm{x}^\mathrm{LDA} + 0.161E_\mathrm{x}^\mathrm{B88} - 0.036E_\mathrm{c}^\mathrm{LDA} + 0.490E_\mathrm{c}^\mathrm{LYP} + 0.461E_\mathrm{c}^\mathrm{MP2_{ss}} + 0.749E_\mathrm{c}^\mathrm{MP2_{os}}$, introduces only modest errors compared to a very high target – the best possible result for each set (**@Self**, that is $\mathrm{MAD}_{\mathbf{B@B}}$). Interestingly, this DFA has more exact exchange and MP2 correlation than other double hybrids,[12,19,28] in part because we use HF orbitals as inputs[15]. High amounts of exact exchange and MP2 correlation also enable XYZ@T100 to give high accuracy for self-interaction-error (SIE) related problems which are typically challenging even for double hybrids[15] (see Figs S20 and S21 for further examples for the related SIE4x4 set). Going back to Fig. 4, training on mindless benchmarks (**@Mindless**) is a little worse on average, but still better than using **@Mindful** molecules. Results for r^{2}SCAN (with different optimal parameters) follow a similar trend.

The accuracy limit and focus on difficult cases

Finally, the TAT also lets us evaluate the accuracy limit of double hybrids – that is the **A@A** case, which is the best possible results for a specific kind of problem given the double hybrid functional form. We remind the reader that $XYZ_7(\mathbf{A})$ is optimized over all seven parameters, so represents the best possible pure (i.e. not range-separated) double hybrid for a given benchset **A**. Therefore, $\mathrm{MAD}_{\mathbf{A@A}}$ indicates the smallest possible error from our XYZ_7 double hybrid family and dictates its accuracy limit.

Fig. 5 explores the accuracy limits of double hybrid functional forms by showing the distribution of absolute errors for various benchsets, with a focus on difficult cases[25,29]. It reports a selection of optimal (self-optimized **A@A** cases) and non-optimal (**A@B** cases) DFAs, to reveal that the overwhelming majority of organic processes can be predicted with good (< 1 kcal/mol; chemical) or ok (1–7 kcal/mol; useful) accuracy, so long as they are trained on a good reference benchset (here, **Org.**=GMTKN55 or **T100**).

But, Fig. 5 also reveals that difficult cases, particularly in transition metals, remain elusive. A quarter (24%) of difficult organic (**OrgDiff**)[29] and half (53%) of difficult transition metal (**TMDiff**)[25] processes exceed acceptable error margins, even with the optimal DFAs. Supp. Fig. S22 reveals that errors cannot be explained by spin-contamination or low-quality benchmarks. Despite generally excellent performance on main group chemistry, current DFA strategies are simply not ready to address true chemical diversity (mechanism and elements) with standard functional types even when using ingredients from all rungs of Jacob's ladder[16,17]. Moreover, DFAs trained on these difficult cases perform poorly on the full **Org.**, especially compared to the almost "best case scenario" of **T100** as a training set.

There is a plus side, however, as difficult cases for DFAs are often also difficult cases for the (very expensive) creation of benchmarking data. The accuracy limit suggests that benchmark quality (and thus cost) may therefore *carefully* be relaxed in some difficult cases.

IV. DISCUSSION AND CONCLUSIONS

Despite involving only very simple mathematics, we have seen that the transferability assessment tool (TAT), especially when applied to the XYZ_p protocol, provides a wealth of analytic information about the training and

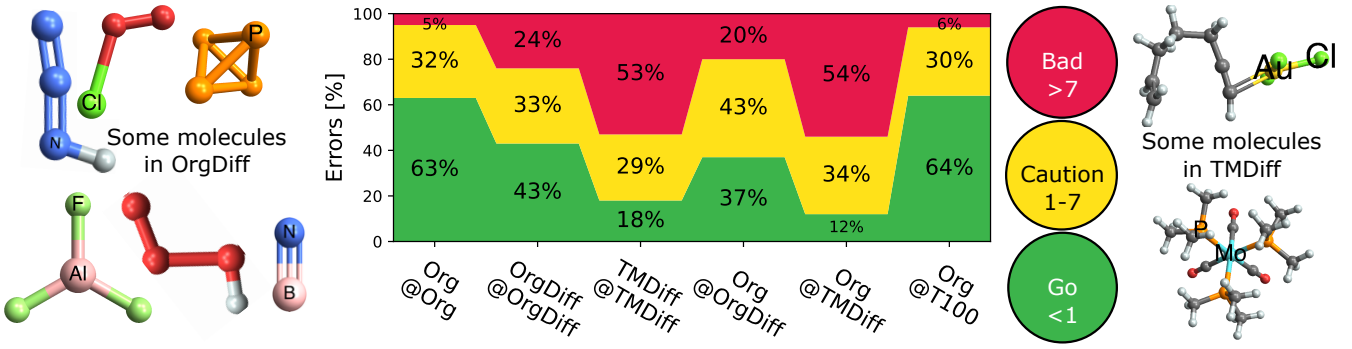

FIG. 5. Fraction of proccesses with good (< 1 kcal/mol), ok (1–7 kcal/mol) and bad (> 7 kcal/mol) errors, $\mathrm{MAD}_{\mathbf{B@A}}$. Includes selected optimal ($\mathbf{B} = \mathbf{A}$) and suboptimal ($\mathbf{B} \neq \mathbf{A}$) combinations. Some example difficult molecules are illustrated to the left (organic) and right (TM).

testing of data-driven DFAs. We can use it to determine what chemistry is hard to learn, what kinds of processes are useful to train on, and to answer many of the questions posed in the introduction. Transferability thus provides an alternative conceptual framework for understanding chemical diversity.

The main conclusion from our work is that following *transferability principles* in data curation is crucial for the construction of general-purpose models in chemistry. Thus, a training benchset should capture genuine chemical and elemental diversity; in such proportions within the benchset that they improve transferability (i.e. with good *transferable diversity*). The evidence presented here therefore suggests the following strategy for better construction, optimization and refinement of benchsets that can be used to train complex, data-driven DFAs:

1. Human input/bias should be reduced in the creation of training (and test) sets, in favour of randomness in chemical construction;

2. Elemental diversity of training sets should be improved, possibly via lower quality benchmarks;

3. Training sets and DFAs should be optimized and refined with an explicit bias toward improving transferability, by testing transferability matrices during their construction.

Our work has revealed that both **Mindless** (= **MB16-43**, Figs 2 and 4) and **T100** (Figs 1, 3–5) make large steps in the right direction: **Mindless** eschews predetermined chemistry and **T100** captures diversity and transferability, both by design. The mindless strategy can be (i) adapted to other cases (e.g., mindless ionization potential or barrier height benchsets); (ii) further extended by introducing randomness in the selection of *mindless potential energy surface* points, which are not confined to local minima; (iii) be biased toward elemental and transferable diversity [as done for **T100**, eq. (5) below] to construct entirely new benchsets.

The success of **Mindless** and **T100** as training sets is also reminiscent of historical successes of DFAs *trained on unrealistic chemical physics* (e.g. homogeneous electron gas[30] or Hooke's atom[31]) in describing *realistic chemistry*. These DFAs are informed by a philosophy that training should be done using areas of chemical space that would typically be *unseen* by the test results. The TAT and transferability principles therefore let us extend this philosophy beyond paradigmatic cases or intuition – which becomes vitally important for machine-learned DFAs, where better interpolation on chemistry *seen* in training risks poorer extrapolation to (prediction of) chemistry *unseen* in training.

It is also worth stressing that the TAT may be applied to any empirical model, and especially those for which the level of empiricism can be controlled. This includes models based on wave function theories (at one extreme) and machine learning of 'classical' energies from molecular geometries (at the other extreme). Work along these lines should be pursued.

Finally, it is important to note that transferability principles are important to consider even for models that explicitly target a specific type of chemistry problem (e.g. DFAs optimized for organic barriers or materials chemistry). Despite their narrower goals, such approaches implicitly assume that the training benchset contains sufficient diversity to enable predictions of similar problems; and that the diversity is appropriately weighted. The low transferability between subsets of **Barriers** reveals that these assumptions are not guaranteed. Embedding transferable diversity into training benchsets, even for narrowly-focussed problems, enables higher confidence in their predictive reliability.

V. METHODS

A. XYZ DFAs

All XYZ_p functionals considered in this work have the same fundamental functional form,

$$\begin{aligned} E_{\mathrm{xc}} =& a_1 E_{\mathrm{x}}^{\mathrm{HF}} + a_2 E_{\mathrm{x}}^{\mathrm{LDA}} + a_3 E_{\mathrm{x}}^{(\mathrm{m})\mathrm{GGA}} \\ &+ a_4 E_{\mathrm{c}}^{\mathrm{LDA}} + a_5 E_{\mathrm{c}}^{(\mathrm{m})\mathrm{GGA}} + a_6 E_{\mathrm{x}}^{\mathrm{MP2_{ss}}} + a_7 E_{\mathrm{x}}^{\mathrm{MP2_{os}}}, \end{aligned} \tag{3}$$

where $E_{\mathrm{x(c)}}$ indicate exchange (correlation) energy approximations, $E_{\mathrm{x}}^{\mathrm{HF}}$ is the exact HF exchange energy and $E_{\mathrm{c}}^{\mathrm{MP2_{ss(os)}}}$ indicate the same-spin and opposite-spin parts of the MP2 energy. $E_{\mathrm{x}}^{(\mathrm{m})\mathrm{GGA}}$ and $E_{\mathrm{c}}^{(\mathrm{m})\mathrm{GGA}}$ denote GGA or meta-GGA exchange and correlation.

The DFA of Eq. 3 is thus defined by a seven-component vector, $\vec{a}$. XYZ_7 allows flexible choice of all seven components. For $\mathrm{XYZ}_{p<7}$, the components of the vector are determined by the following rules:

- $p = 1$: Choose exact exchange fraction, α, and set $a_1 := \alpha$, $a_2 := a_4 := 0$, $a_3 := 1 - \alpha$, $a_5 := 1 - \alpha^2$, $a_6 := a_7 := \alpha^2$
- $p = 2$: Choose exact exchange fraction, α, and MP2 fraction, β, and set $a_1 := \alpha$, $a_2 := a_4 := 0$, $a_3 := 1 - \alpha$, $a_5 := 1 - \beta$, $a_6 := a_7 := \beta$;
- $p = 3$: Choose free a_1, a_3 and a_6, and set $a_2 := a_4 := 0$, $a_5 := 1 - a_6$, $a_7 := a_6$;
- $p = 4$: Choose free a_1, a_2, a_3 and a_6, and set $a_4 := 0$, $a_5 := 1 - a_6$, $a_7 := a_6$;
- $p = 5$: Choose all except $a_4 := 0$ and $a_7 := a_6$;
- $p = 6$: Choose all except $a_7 := a_6$.

Unless otherwise specified, throughout this work we use Becke's (B88)[32] exchange GGA and Lee, Yang and Parr's (LYP)[33] for $E_{\mathrm{x}}^{(m)\mathrm{GGA}}$ and $E_{\mathrm{c}}^{(m)\mathrm{GGA}}$, respectively (BLYP). The optimal DFA for set **A** is then defined via,

$$\mathrm{XYZ}_p(\mathbf{A}) = \arg\min_{\mathrm{XYZ}_p} \mathrm{MAD}(\mathrm{XYZ}_p \text{ on } \mathbf{A}) \tag{4}$$

where XYZ_p indicates all possible variants of Eq. (3) consistent with the number, p, of parameters (using BLYP as GGAs); and MAD(DFA on **set**) indicates the mean absolute deviation of energies computed using DFA, averaged across all processes in **set**. We thereby obtain, $\mathrm{MAD}_{\mathbf{B@A}} := \mathrm{MAD}(\mathrm{XYZ}_p(\mathbf{A}) \text{ on } \mathbf{B})$

The results for two other combinations – PBE exchange + PBE correlation[30]; and r^{2}SCAN exchange + r^{2}SCAN correlation[34] – are given in the SI. The main conclusions of our work do not change once we replace the BLYP-based GGAs with their PBE-/r^{2}SCAN-based counterparts in Eq. 3.

B. Computational details

All HF and DFT calculations were conducted with Orca 5.0.0[35]. We used def2-QZVPPD for GMTKN55 and def2-QZVP for TMC151. For costly cases, def2-QZVP(P) or def2-TZVP(P) were used. Further details are in Sec. S1 of the SI. Orbitals were computed using unrestricted Hartree-Fock (UHF) theory in all cases.

C. Special benchmark sets

Mostly we use the categories from GMTKN55 and TMC151 or preexisting subsets (e.g. Diet100[36]). We also have some special benchset (and aliases to stress important features):

Mindless is an alias for MB16-43[1,24], to stress its most important feature;

Mindful combines DARC and ISO34 sets[1]; chosen to represent chemical intuition-based counterpart of **Mindless**;

Org. is an alias for GMTKN55, to stress its focus on organic chemistry;

Org. difficult=**OrgDiff** is the P30-5 'poison' subset of GMTKN55, from Ref. 29;

Org. X indicates a subset from GMTKN55;

TM is an alias for TMC151, to stress its focus on transition metal chemistry;

TM difficult=**TMDiff** is a subset of TMC151 composed of TMD + two MOR41 reactions + six TMB barriers, all identified as difficult in Ref. 25;

TM X indicates a subset from TMC151;

TM+Org. is the combination of GMTKN55 and TMC151;

T100 is a subset of **TM+Org.** designed around transferable diversity principles (see below).

To construct **T100** we first 'mindlessly' breed twenty "pretty transferable" (denoted $\mathbf{PT}_{1\ldots20}$) subsets of the combined GMTKN55 and TMC151 (**TM+Org.**) benchset, each with 100 processes. Survival is dictated by a genetic approach similar to that used to construct Diet sets, with breeding success based on transferability of XYZ_7.[36] Full details are in Section S2 of the SI. Then, we obtain **T100** by selecting the best one, using:

$$\mathbf{T100} = \arg\min_{\mathbf{PT}_k} \left[\tfrac{1}{3} \sum_{p\in 1,4,7} \bar{T}_p(\mathbf{PT}_k) - 0.03 N_{\mathrm{el}}(\mathbf{PT}_k) \right]. \tag{5}$$

Here, $\bar{T}_p(\mathbf{PT}_k) = \frac{1}{58} \sum_{\mathbf{B}\in\mathbf{TM+Org.}} T_{\mathbf{B@PT}_k;\mathrm{XYZ}_p}$ is the average transferability from $\mathbf{PT}_k$ to all 58 subsets of

GMTKN55 and TMC151, using XYZ_p. Averaging over $p \in 1, 4, 7$ helps to avoid 'accidental' transferability for any specific number of parameters. Biasing to a larger number, $N_{\text{el}}(\mathbf{PT}_k)$, of unique elements in $\mathbf{PT}_k$ helps to avoid over-representation of main group chemistry, which is 10 times more common than TM chemistry in **TM+Org.**.

We use BLYP (Becke exchange[32] and Lee-Yang-Parr correlation[33]) in Eq. (3) for both the breeding and optimisation stages, which means the transferable diversity of **T100** is biased toward BLYP. In principle, other functional choices might lead to other sets. Nevertheless, Supp. Fig. S23 reveal that training PBE- and r^2SCAN-based XYZ_p on BLYP's **T100** gives them transferability similar to DFAs trained on the full GMTKN55 benchset. **T100** also works for a different functional form – that of B3LYP,[37] which excludes MP2 contributions entirely (see Supp. Fig. S24). It follows that transferable diversity features of **T100** are largely independent of functional form choice.

ACKNOWLEDGEMENTS

SV acknowledges funding from the SNSF Starting Grant project (TMSGI2_211246). TG was supported by an Australian Research Council (ARC) Discovery Project (DP200100033) and Future Fellowship (FT210100663). SD was supported by an Australian Research Council (ARC) Discovery Project (DP200100033) and by the Ministry of Education, Singapore, under its Research Centre of Excellence award to the Institute for Functional Intelligent Materials, with Project No. EDUNC-33-18-279-V12. BC acknowledges research funding from Japan Society for the Promotion of Science (22H02080) and generous grants of computer time from the RIKEN Information Systems Division (Q23266), Japan.

AUTHOR CONTRIBUTIONS

SV conceived the transferability presented here and carried out most computations. TG and SV worked together on analysis (including coding) and writing. SD helped with chemical insights. BC helped with insights into benchmarking and computation.

[1] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, and S. Grimme, "A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions," Phys. Chem. Chem. Phys. **19**, 32184–32215 (2017).

[2] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Phys. Rev. **136**, B864–B871 (1964).

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

[4] C. D. Sherrill, D. E. Manolopoulos, T. J. Martinez, and A. Michaelides, "Electronic structure software," J. Chem. Phys. **153** (2020), 10.1063/5.0023185.

[5] S. Vuckovic, A. Gerolin, T. J. Daas, H. Bahmann, G. Friesecke, and P. Gori-Giorgi, "Density functionals based on the mathematical structure of the strong-interaction limit of dft," WIREs Comput Mol SciScience **13** (2022), 10.1002/wcms.1634.

[6] B. Kalita, L. Li, R. J. McCarty, and K. Burke, "Learning to approximate density functionals," Acc. Chem. Res. **54**, 818–826 (2021).

[7] R. Pederson, B. Kalita, and K. Burke, "Machine learning and density functional theory," Nat Rev Phys **4**, 357–358 (2022).

[8] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, "Pushing the frontiers of density functionals by solving the fractional electron problem," Sci **374**, 1385–1389 (2021).

[9] B. Huang, G. F. von Rudorff, and O. A. von Lilienfeld, "The central role of density functional theory in the AI age," Sci **381**, 170–175 (2023).

[10] O. A. von Lilienfeld, K.-R. Müller, and A. Tkatchenko, "Exploring chemical compound space with quantum-based machine learning," Nat Rev Chem **4**, 347–358 (2020).

[11] O. A. von Lilienfeld and K. Burke, "Retrospective on a decade of machine learning for chemical discovery," Nat. Commun. **11** (2020), 10.1038/s41467-020-18556-9.

[12] S. Grimme, "Semiempirical hybrid density functional with perturbative second-order correlation," J. Chem. Phys. **124** (2006), 10.1063/1.2148954.

[13] J. M. L. Martin and G. Santra, "Empirical double-hybrid density functional theory: A 'third way' in between WFT and DFT," Isr. J. Chem. **60**, 787–804 (2019).

[14] I. Y. Zhang and X. Xu, "Exploring the limits of the XYG3-type doubly hybrid approximations for the main-group chemistry: The xDH@b3lyp model," J. Phys. Chem. Lett. **12**, 2638–2644 (2021).

[15] S. Song, S. Vuckovic, E. Sim, and K. Burke, "Density sensitivity of empirical functionals," J. Phys. Chem. Lett. **12**, 800–807 (2021).

[16] J. P. Perdew, "Jacob's ladder of density functional approximations for the exchange-correlation energy," in *AIP Conference Proceedings* (AIP, 2001).

[17] S. Hammes-Schiffer, "A conundrum for density functional theory," Sci **355**, 28–29 (2017).

[18] E. Sim, S. Song, S. Vuckovic, and K. Burke, "Improving results by improving densities: Density-corrected density functional theory," J. Am. Chem. Soc. **144**, 6625–6639 (2022).

[19] Y. Zhang, X. Xu, and W. A. Goddard, "Doubly hybrid density functional for accurate descriptions of nonbond interactions, thermochemistry, and thermochemical kinetics," Proc. Natl. Acad. Sci. **106**, 4963–4968 (2009).

[20] L. A. Curtiss, K. Raghavachari, G. W. Trucks, and J. A. Pople, "Gaussian-2 theory for molecular energies of first- and second-row compounds," J. Chem. Phys. **94**, 7221–7230 (1991).

[21] A. Karton, S. Daon, and J. M. Martin, "W4-11: A high-confidence benchmark dataset for computational thermochemistry derived from first-principles w4 data," Chem. Phys. Lett. **510**, 165–178 (2011).

[22] J. Hollingsworth, L. Li, T. E. Baker, and K. Burke, "Can exact conditions improve machine-learned density functionals?" J. Chem. Phys. **148** (2018), 10.1063/1.5025668.

[23] R. Nagai, R. Akashi, and O. Sugino, ""mindless" DFT benchmarking," J. Chem. Theory Comput. **5**, 993–1003 (2009).

[24] M. Korth and S. Grimme, ""mindless" DFT benchmarking," J. Chem. Theory Comput. **5**, 993–1003 (2009).

[25] B. Chan, P. M. W. Gill, and M. Kimura, "Assessment of DFT methods for transition metals with the TMC151 compilation of data sets and comparison with accuracies for main-group chemistry," J. Chem. Theory Comput. **15**, 3610–3622 (2019).

[26] S. Dohm, A. Hansen, M. Steinmetz, S. Grimme, and M. P. Checinski, "Comprehensive thermochemical benchmark set of re-

alistic closed-shell metal organic reactions," J. Chem. Theory Comput. **14**, 2596–2608 (2018).
[27]M. G. Taylor, D. J. Burrill, J. Janssen, E. R. Batista, D. Perez, and P. Yang, "Architector for high-throughput cross-periodic table 3d complex building," Nature Communications **14** (2023), 10.1038/s41467-023-38169-2.
[28]S. Kozuch and J. M. L. Martin, "Spin-component-scaled double hybrids: An extensive search for the best fifth-rung functionals blending DFT and perturbation theory," J. Comput. Chem. (2013), 10.1002/jcc.23391.
[29]T. Gould and S. G. Dale, "Poisoning density functional theory with benchmark sets of difficult systems," Phys. Chem. Chem. Phys. **24**, 6398–6403 (2022).
[30]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865–3868 (1996).
[31]S. Śmiga, F. D. Sala, P. Gori-Giorgi, and E. Fabiano, "Self-consistent implementation of kohn–sham adiabatic connection models with improved treatment of the strong-interaction limit," J. Chem. Theory Comput. **18**, 5936–5947 (2022).
[32]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior," Phys. Rev. A **38**, 3098–3100 (1988).
[33]C. Lee, W. Yang, and R. G. Parr, "Development of the colle-salvetti correlation-energy formula into a functional of the electron density," Phys. Rev. B **37**, 785–789 (1988).
[34]J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, "Accurate and numerically efficient r^2scan meta-generalized gradient approximation," J. Phys. Chem. Lett. **11**, 8208–8215 (2020).
[35]F. Neese, "Software update: The orca program system – version 5.0," WIREs Computational Molecular Science **12** (2022), 10.1002/wcms.1606.
[36]T. Gould, "'diet GMTKN55' offers accelerated benchmarking through a representative subset approach," Phys. Chem. Chem. Phys. **20**, 27735–27739 (2018).
[37]A. D. Becke, "A new mixing of Hartree-Fock and local density-functional theories," J. Chem. Phys. **98**, 1372–77 (1993).