Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 3227

Harnessing surrogate models for data-efficient predictive chemistry: descriptors *vs.* learned hidden representations

Guanming Chen and Thijs Stuyver *

Predictive chemistry often faces data scarcity, limiting the performance of machine learning (ML) models. This is particularly the case for specialized tasks such as reaction rate or selectivity prediction. A common solution is to use quantum mechanical (QM) descriptors—physically meaningful features derived from electronic structure calculations—to enhance model robustness in low-data regimes. However, computing these descriptors is costly. Surrogate models address this by predicting QM descriptors directly from molecular structure, enabling fast and scalable input generation for data-efficient downstream ML models. In this study, we compare two strategies for using surrogate models: one that feeds predicted QM descriptors into downstream models, and another that leverages the surrogate's internal hidden representations instead. Across a diverse set of chemical prediction tasks, we find that hidden representations often outperform QM descriptors, particularly when descriptor selection is not tightly aligned with the downstream task. Only for extremely small datasets or when using carefully selected, task-specific descriptors do the predicted values yield better performance. Our findings highlight that the hidden space of surrogate models captures rich, transferable chemical information, offering a robust and efficient alternative to explicit descriptor use. We recommend this strategy for building data-efficient models in predictive chemistry, especially when feature importance analysis is not a primary goal.

Received 9th June 2025
Accepted 13th September 2025

DOI: 10.1039/d5dd00256g

rsc.li/digitaldiscovery

Introduction

A common issue faced when designing data-driven models in chemistry is data scarcity.[1,2] For many specialized predictive tasks, *e.g.*, prediction of reaction rates, enantiomeric excess, and solvation energies, only limited amounts of relevant and accurate data can be mined from the literature, and high-throughput experimentation is technically challenging and/or prohibitively expensive.[3,4] As a consequence, datasets for these types of tasks typically only contain several hundred, up to a couple of thousand, data points at best. In such a data-limited regime, conventional machine learning (ML) algorithms tend to perform poorly.

One strategy to address the issue of data scarcity consists of representation engineering, *i.e.*, describing the molecules/reactions through a limited set of (carefully selected) informative, physically meaningful descriptors, so that a robust relationship between ML model input and output can be learned.[5–7] Quantum Mechanical (QM) descriptors are a particularly popular choice in this regard. Unfortunately, the calculation of such descriptors typically requires resource-intensive density functional theory (DFT) calculations,[8–12] which limits the applicability of this approach to big datasets, as well as to use cases where inference is expected at a high-throughput speed.

An alternative strategy that has been pioneered in recent years is to avoid the explicit calculation of QM descriptors, by predicting their values for unseen molecules and/or reactions with the help of a surrogate ML model.[1,13–16] Taking this approach, QM descriptors can be inferred on-the-fly, so that the generation of the input representation of the downstream model can be seamlessly integrated into a single end-to-end model. This aggregate model then rivals regular ML models in terms of inference speed and computational resource footprint, while enabling higher accuracy and increased robustness due to the physical information encoded in the intermediately predicted descriptors.

Of course, setting up such a surrogate model still requires an initial training dataset, constructed through high-throughput QM calculations. However, once generated, this data, as well as the resulting surrogate models, can be applied to – and repurposed for – different downstream prediction tasks. Consequently, a range of high-throughput QM datasets have been released in recent years. In addition to the prototypical QM9 dataset,[17] one of the earliest large-scale examples was the QMugs[18,19] dataset, which contains a wide range of quantum mechanically (QM) computed descriptors and properties for 665k biologically and pharmacologically relevant molecules

Ecole Nationale Supérieure de Chimie de Paris, Université PSL, CNRS, i-CLeHS, 75 005 Paris, France. E-mail: thijs.stuyver@chimieparistech.psl.eu

extracted from the ChEMBL[20] database. Other examples include the QM40 (ref. 21) dataset, which contains QM properties for 163k compounds extracted from the ZINC[22] database, the QCDGE[23] dataset, which contains both ground- and excited-state properties for 450k C, H, N, O, F containing compounds, the BDE-db[24] dataset, which contains QM descriptors for more than 200k organic radicals, and the tmQM[25] dataset, which contains properties for 86k transition metal complexes.

The growing availability and diversity of public QM descriptor datasets indicate that surrogate modeling will become increasingly accessible—and presumably more widely adopted as a consequence—in the years to come. As such, it is important to establish guidelines on how the QM information, captured in these datasets, can be leveraged optimally for downstream tasks.

Taking a closer look at the typical architecture of the surrogate models used so far,[1,14,15] one can conclude that they generally start from a SMILES string from which a molecular graph is deduced. The atomic vectors of this graph are subsequently embedded into a learned (hidden) representation, after which multiple feed-forward neural networks (FFNN), or readout functions, lead to the actual QM descriptors, that is, the surrogate model targets (Fig. 1).

Starting from this realization, a natural question arises: how does directly using the final hidden representation of the embedder in the surrogate model as input for the downstream model compare to the conventional surrogate model approach of introducing the predicted QM descriptors as the input, or as supplementary features, for the downstream ML model?

Note that the alternative hidden representation strategy, introduced above, is conceptually connected to a pre-training strategy: a model is first trained on an extensive descriptor dataset, after which the weights in the trained encoder are frozen, and the prediction heads, leading to the descriptors, are detached and replaced by pristine heads that lead to the targets of the downstream task.[26–30]

Intuitively, arguments can be devised in favor of a superior performance for either the descriptors or the hidden representation approach. On the one hand, the readout process to go from hidden representation to QM descriptors may result in information loss due to the compression of certain, useful, hidden features, *i.e.*, the learned hidden space may contain some features that are beneficial for downstream tasks, which are not transferred when QM descriptors are used as input for the latter model. On the other hand, the hidden representations themselves are usually high-dimensional (*e.g.*, 1200 dimensions for a single atomic representation), and hence they may be inherently non-linear, and/or may contain a high degree of redundancy (*i.e.*, many dimensions either correlate only very weakly with downstream targets, or are largely co-linear with some of the other hidden dimensions). As such, the high-dimensionality may make it difficult to fully leverage the

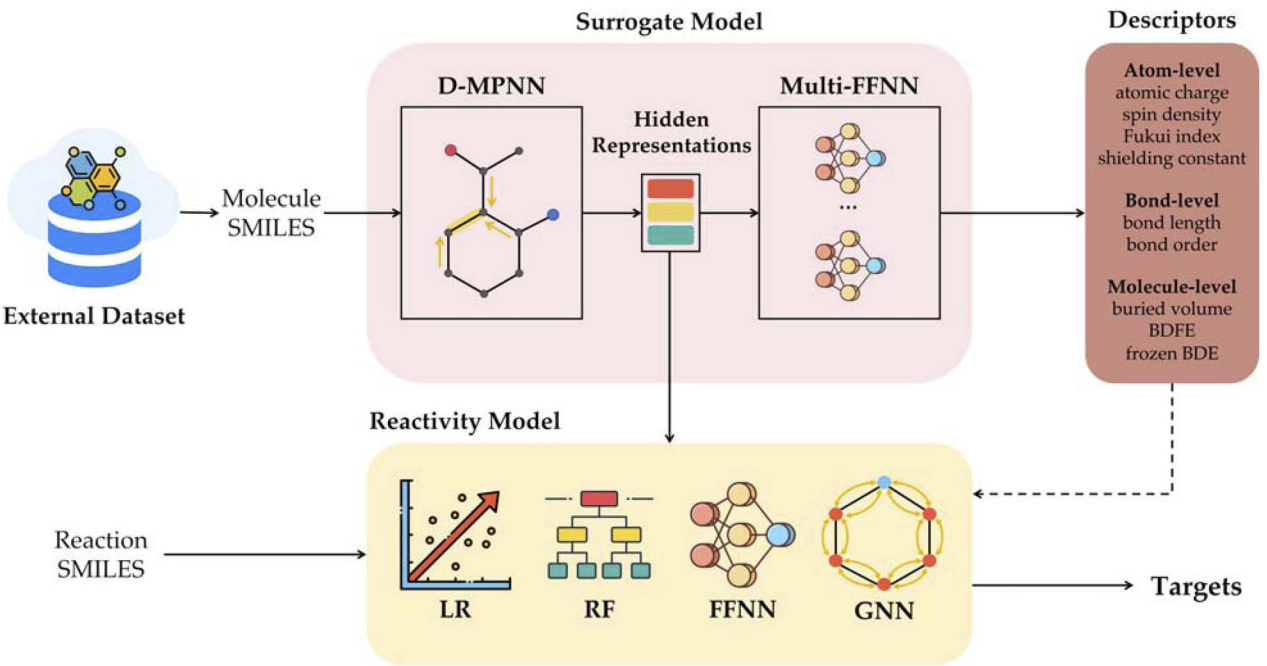

Fig. 1 Predictive chemistry with surrogate models: scheme of the surrogate model and the downstream reactivity prediction model architectures employed in this study, taking reaction prediction as an example. The surrogate model is trained with external quantum chemical descriptor datasets to generate informative representations on-the-fly. The SMILES representation of a molecule is first converted into a 2-D molecular graph, and then a (multi-layer) D-MPNN is employed to generate hidden representations for each atom and bond within this molecule, based on which multiple FFNNs, serving as readout functions, predict individual descriptors. The Reactivity Model uses LR, RF, FFNN or GNN as backbone and is augmented with the on-the-fly reaction representation generated by the surrogate model to predict reactivity targets. Previous studies leveraged the explicitly predicted descriptors, while in this study, the learned hidden representations of the surrogate model are utilized directly as input.

encoded information under sparse data conditions, which is typically the case for the downstream tasks for which we would want to use a surrogate model.

In this work, incorporating predicted QM descriptors, and directly utilizing the hidden space of the surrogate model, will be compared head-to-head for a selection of representative downstream tasks, mainly within the realm of chemical reactivity.

Our results suggest that only when a carefully engineered – and complete – set of descriptors is selected, the direct use of QM descriptor values may result in a superior performance of the downstream model. In most cases, however, and especially when descriptors are selected without too much attention to the underlying physics/chemistry of the downstream task (which is typically the case when descriptors are added to a predictive model),[1,14,31] leveraging the surrogate model's hidden space actually results in a better predictive performance – the discrepancy is often significant. As such, unless gaining qualitative insights through feature importance analysis is the main goal of the study,[32–34] we propose as a general recommendation to focus on the hidden space of surrogate models to enhance predictive models for downstream tasks, not the predicted descriptors themselves.

Methodology

Selected datasets

Here, we focus on three case studies, in which surrogate models for QM descriptor prediction have been successfully applied before.

First, we focus on the work by Alfonso-Ramos and co-workers[15] on hydrogen atom transfer (HAT) reactivity, where the aim consisted of accurately predicting activation energies ($\Delta G^{\ddagger}$) for a variety of small, downstream datasets. In this study, the previously mentioned BDE-db database was used to construct the surrogate model. The relevant QM descriptors were identified through a careful Valence Bond (VB) analysis for a generic HAT model reaction (see Section S1 in the SI for an in-depth discussion). Specifically, the following descriptors were mined from BDE-db: (1) partial charges and spin densities as atom-level descriptors; (2) buried volume, frozen bond dissociation energy (BDE), and bond dissociation free energy (BDFE) as molecule-level descriptors. The vector, obtained by concatenating these descriptors on both the reactant and product sides, contains 14 elements.

The specific downstream datasets considered are: (1) In-House, a dataset consisting of 1511 DFT computed gas-phase reaction profiles for pairs of organic compounds extracted from BDE-db; (2) Alkoxy,[35] consisting of computed reaction profiles for alkoxy radicals abstracting hydrogens from hydrocarbons and heterosubstituted compounds in an acetonitrile solution, where 238 reactions were pre-selected for training and validation and 60 reactions for testing; (3) Exp. alkoxy,[36] a small dataset of experimentally reported selectivities for 6 hydrocarbons by $CH_3O^{\bullet}$, and the corresponding 15 DFT-computed reaction barriers; (4) Photoredox HAT,[37] 564 photoredox-mediated HAT catalysis reactions with various allylic, propargylic, benzylic, aldehyde and alkyl substrates and O/N-based radical species extracted from a larger data set published by Hong and co-workers;[37] (5) Exp. cumyloxyl,[38] an experimental dataset containing 45 HAT reactions from C(sp^3)–H bonds by cumyloxyl radical; (6) Cytochrome P450,[39] consisting of 24 activation energies for HAT by the cytochrome P450 enzyme from organic compounds, where 6 reactions out of 24 are pre-selected for testing; (7) Atmospheric HAT,[40] consisting of 73 HAT reactions encountered in atmospheric chemistry and extracted from RMechDB.[41] Except for the In-House dataset, these sets were originally extracted from the literature by Alfonso-Ramos *et al.*,[15] with the aim to cover a wide range of HAT settings (*i.e.*, atmospheric, metabolic, and synthetic reactivity), while staying as much as possible within the scope of the surrogate model.

Secondly, we focus on the work by Guan *et al.*[1] on predicting experimentally observed regiochemistry for selective organic reactions extracted from the patent literature. In this case, the surrogate model was trained on an in-house generated dataset of QM descriptors computed at B3LYP/def2-SVP level-of-theory[42–44] for 136k organic molecules, containing C, H, O, N, P, S, F, Cl, Br, I, Si, B elements, curated from the ChEMBL and Pistachio[45] databases. For the selection of the descriptors, Guan *et al.* simply opted to compute a series of the most frequently used local reactivity indices, that is, (1) atomic charges, nucleophilic Fukui indices, electrophilic Fukui indices[46,47] and NMR[48] shielding constants as atomic descriptors; (2) bond lengths and bond orders as bond descriptors.

The specific downstream datasets considered in this work, are: (1) C–H, consisting of 2244 aromatic C–H functionalization reactions, (2) C–X, consisting of 1024 aromatic C–X substitution, and (3) others, consisting of all selective substitution reactions that do not fall in either of the previously mentioned categories (552 entries in total).

Finally, we focus on the work by Li *et al.*,[14] in which the surrogate model approach was applied to a broad range of downstream tasks. Here, the surrogate model was trained on an in-house generated dataset of QM descriptors computed at 6 different levels-of-theory for 65k organic molecules, extracted from a range of public databases.[20,49–54] We decided to focus on the descriptors computed at ωB97X-D/def2-SVP//GFN2-xTB level-of-theory.[44,55,56] 37 descriptors were considered in total, 13 atom-level descriptors (*e.g.*, NPA charges,[57] Parr functions,[58] NMR shielding constants, and valence orbital occupancies), 4 bond-level (*e.g.*, bond order, bond length, bonding electrons, and bond natural ionicity), and 20 molecule-level descriptors (*e.g.*, HOMO–LUMO gap, ionization potential, electron affinity, and dipole and quadrupole moments). In their work, Li *et al.*,[14] trained 3 separate surrogate models (1 for both atom-level and bond-level, and 2 for molecule-level descriptors). Here, because of the nature of the downstream tasks, we focus on the surrogate models that predict molecule-level descriptors.

In total, 16 downstream applications were considered in the original work by Li *et al.*[14] Here, we only consider a representative subselection of the corresponding datasets: (1) ESOL,[59] an experimental regression dataset containing water solubilities for 1127 molecules, (2) FreeSolv,[60] an experimental regression

dataset, consisting of 642 hydration free energy values, (3) QM9,[17] a regression dataset consisting of 12 (quantum chemically computed) energetic, electronic and thermodynamic properties, for 134k organic compounds with up to 9 heavy atoms, (4) HIV,[61] an experimental classification dataset indicating the ability to inhibit HIV replication for 41 127 molecules, (5) ClinTox,[62] an experimental classification dataset indicating the toxicity of 1477 compounds. The above datasets are also available through MoleculeNet.[62,63]

Note that the benchmarking datasets selected above for all three case studies are identical to those in the original studies,[1,14,15] and even the same data splits are adopted to ensure a fair comparison. Their summary is shown in Table 1.

Surrogate model architecture

In each of the three case studies mentioned above, a multi-task deep learning model architecture, derived from the original ChemProp model,[27] was used as the surrogate. As such, this is also the surrogate model architecture that we adopted here consistently.

An in-depth discussion of the ChemProp architecture can be found in a recent publication by Heid *et al.*[27] In short, the model starts by constructing a molecular graph from a SMILES input, after which the graph is passed through a (multi-layer) directed message passing neural network (D-MPNN) encoder. The D-MPNN iteratively aggregates and updates information from neighboring atoms and bonds and, as a result, encodes a molecule into separate atom- and bond-level embeddings. These representations are then passed on to multiple FFNN readout functions, where each FFNN is trained to predict one individual QM descriptor. For molecule-level descriptors, the atom-level vectors are first sum-pooled (or mean-pooled), before passing the final hidden representation/embedding to the respective FFNN readout functions. For globally constrained atom-level descriptors – *e.g.*, the sum of all the atomic charges within a neutral molecule should always be equal to zero – an attention-based correction is applied to ensure that this constraint is satisfied.

For the second[1] and third[14] case studies, the surrogate models were not retrained at any point, and hence, the hyperparameters selected in the corresponding original works were adopted here as well consistently. This means that the sizes of the atom- and molecule-level hidden representations are also fixed and pre-set – at 600 and 700/900, respectively.

For the first case study,[15] we also opted to use the original trained surrogate model throughout the first part of our analysis. Additionally, we also performed some tests where we retrained the model on only a subset of the QM descriptors (*vide infra*), but with the same hyperparameters. This means that the hidden atom- and molecule-level representations contain 1200 dimensions each by default. Furthermore, we also performed some tests where we systematically modulated the hidden dimension size h, of the surrogate. As evident from the discussion in the final section below, doing so does not affect the conclusions drawn significantly.

Extraction of the hidden representation of the surrogate as downstream model input

Before presenting the specific approach for the individual case studies, we aim to specify more clearly what is meant in general by "hidden representations" in our approach and where exactly these are extracted in the surrogate model architecture. As shown in Fig. 1, when a molecule is input into the surrogate model, it undergoes several message-passing steps within the D-MPNN encoder. This results in learned vector embeddings for atoms and bonds, with dimensionality determined by encoder settings. A molecule-level embedding is then obtained *via* sum-pooling of the atom-level vectors.

In the standard descriptor-based pipeline, these atom-, bond-, and molecule-level embeddings are passed through corresponding FFNNs to produce the respective (surrogate

Table 1 Summary of datasets used in this study

Dataset	Labels	Task
In-House[15]	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
Alkoxy[35]	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
Exp. alkoxy[36]	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
Photoredox HAT[37]	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
Exp. cumyloxyl[38]	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
Cytochrome P450 (ref. 39)	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
Atmospheric HAT[40]	$\Delta G^{\ddagger}$	HAT reactivity prediction (regression)
C–H, C–X, Others, All[1]	Primary product formed among enumerated regio-isomers	Regioselectivity prediction (classification)
ESOL[59]	Log(S)	Molecular property prediction (regression)
FreeSolv[60]	$\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$, $\Delta G_{\mathrm{hyd}}^{\mathrm{calc}}$	Molecular property prediction (two-task regression)
QM9 (ref. 17)	DFT-derived μ, α, $\varepsilon_{\mathrm{HOMO}}$, $\varepsilon_{\mathrm{LUMO}}$, ε_{gap}, $\langle R^2 \rangle$, zpve, C_{v}, U_0, U, H, G	Molecular property prediction (multi-task regression)
HIV[61]	Ability to inhibit HIV replication	Molecular property prediction (binary classification)
ClinTox[62]	Toxicity of drug candidates	Molecular property prediction (binary classification)

model-dependent) descriptors. These descriptors are then concatenated to form the input to the downstream model.

In our hidden representation approach, we directly extract and concatenate selected subsets of the learned atom-, bond-, and molecule-level embeddings—bypassing the FFNNs altogether—to construct alternative downstream model inputs. This custom selection approach is necessary to ensure a fair comparison: both approaches should only have access to the same underlying descriptor information. Since descriptor generation varies between case studies, the construction of the hidden representation inputs must be modulated accordingly.

In the original HAT reactivity study,[15] both (reactive) atom-level, and molecule-level, descriptors, respectively on the reactant and product side of the reaction, were extracted from the surrogate model as input for the downstream tasks. As such, we decided to concatenate both atom-level embeddings of the reactive atoms, *i.e.*, the sites undergoing a bonding change, and molecule-level embeddings in the hidden representation approach.

The general reaction scheme for HAT reactivity can be written as follows,

$$R_1\text{–}A^{\bullet}\text{–}H + R_2\text{–}B^{\bullet} \rightarrow R_1\text{–}A^{\bullet} + R_2\text{–}B^{\bullet}\text{–}H \quad (1)$$

The input representation we constructed for this case study thus consists of 6 concatenated vectors: the hidden representations of the four molecules R_1–$A^{\bullet}$–H, R_2–$B^{\bullet}$, R_1–$A^{\bullet}$, and R_2–$B^{\bullet}$–H, involved in the reaction, and the atom-level representations of the radical sites on reactant and product side respectively, *i.e.*, $B^{\bullet}$ in R_2–$B^{\bullet}$ and $A^{\bullet}$ in R_1–$A^{\bullet}$.

In their study on regioselectivity prediction, Guan *et al.*[1] focused exclusively on the effect of including (atom-level) QM descriptors from the two main reactive sites of the reactants in the downstream model. As such, we decided here to use a simple concatenation of the corresponding atom-level hidden representations of the reactive sites as the alternative input for the downstream model.

Since all the downstream tasks, considered in Li *et al.*'s work,[14] involved molecule-level properties, we simply extracted the hidden representations of the 2 molecule-level surrogate models they generated, concatenated them, and used them as the downstream model input.

Downstream model architectures

To enable optimal comparison between the two considered surrogate strategies, we adopted the downstream model architectures with associated hyperparameters, if applicable, from each corresponding original study.[1,14,15] To compare the effectiveness of both strategies, the root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2) have consistently been selected as the evaluation metrics for the regression tasks, and accuracy and ROC-AUC for the classification tasks.

For the first case study,[15] an ensemble of 4 FFNNs was consistently set as the model architecture, and the optimal number of layers, hidden size, learning rate and dropout rate of these FFNNs were determined for the In-House dataset through a grid search. The resulting settings were subsequently transferred across the different downstream tasks.

For the second case study,[1] the original models for the descriptor-based strategy were directly transferred without any alterations, *i.e.*, a 3-layer FFNN, respectively with 500, 250, and 125 nodes, was selected for each of the downstream datasets.

For the third case study,[14] we selected a 3-layer FFNN, each with 1000 nodes, as the downstream model architecture. Note that for the corresponding baseline model, *i.e.*, the one that takes the predicted descriptors as input, a radial basis function (RBF) expansion ($n = 50$) was performed first, as this turned out to improve the performance of this model compared to just selecting directly the descriptor values themselves (see Section S2 and S3 in the SI for more details and comparisons, respectively).

A more in-depth discussion about the downstream models can be found in S4 of the SI.

Results and discussion

Hidden representations *versus* descriptors for HAT reactivity prediction

Table 2 presents the difference between using predicted descriptors as input for the downstream reactivity model and directly using the hidden representations from the surrogate model for the HAT case study.[15] The hidden representations input, in combination with an FFNN downstream model, readily outperforms the descriptor-based baseline on 4 out of 7 downstream datasets, already underscoring the utility of the former approach.

In line with our intuition outlined in the Introduction, we observe that for the two smallest datasets considered, with only 15 and 24 data points, respectively, the descriptor approach slightly outperforms the hidden representation approach. This suggests that in these extreme situations, the high dimensionality of the downstream model input (which in this case amounts to $6 \times 1200 = 7200$ dimensions, *vide supra*) may indeed negatively affect the performance. However, even if this is indeed the root cause of this observation, it is clearly an effect that vanishes rapidly as dataset size increases: already for the datasets with 45 and 73 datapoints respectively, *i.e.*, Exp. cumyloxyl and Atmospheric HAT, the hidden representation takes over as the most performant downstream model input in our analysis.

Next to the extremely small downstream datasets, there is only one other example for which we observe that the descriptors outperform the hidden representation, namely the biggest of them all, *i.e.*, the In-House dataset.

We hypothesize that the reason the descriptors perform so well for this specific case is that the surrogate QM descriptor model was, in fact, specifically set up with this dataset, consisting of organic HAT reactions run in the gas phase, in mind as the main downstream application in the original publication.[15] More specifically, the constructed VB model, from which the descriptors to include in the surrogate were selected, neglected solvent/environment effects altogether, and the calculations performed to construct the BDE-db database[24] – the

Table 2 Comparison between descriptor-based and hidden representation-based methods in HAT activation energy prediction, where the best results are in bold. For the descriptor-based baseline method, the best model architecture for the downstream reactivity model is indicated. For the hidden representation-based method, an ensemble of 4 FFNNs has consistently been used. For the In-House, Exp. cumyloxyl and Atmospheric HAT datasets, standard 10-fold cross validation (CV) was employed to report the results, while for Alkoxy and Cytochrome P450 datasets, certain datapoints were pre-selected for the test set, and the rest were randomly split into training and validation sets, following 10-fold CV; Standard 5-fold CV is employed for Photoredox HAT. For the evaluation of Exp. alkoxy, the reactivity model trained on Alkoxy was used without further fine-tuning. These settings are in accordance with Alfonso-Ramos *et al.*'s original study.[15] The mean and standard deviation are calculated across 5 different random seeds $s \in [0, 1, 2, 3, 4]$. The p-value of the paired t-tests between descriptor-based and hidden representation-based RMSE: for In-House, Exp. alkoxy, Photoredox HAT, and Exp. cumyloxyl, $p < 0.001$; for Cytochrome P450 and Atmospheric HAT, $p < 0.05$; for Alkoxy, $p = 0.14$. An additional baseline comparison based on Morgan fingerprints can be found in S5 of the SI

		Descriptor-based				Hidden representation-based			
Dataset	Size	Best model	RMSE (kcal mol^{-1})	MAE (kcal mol^{-1})	R^2	RMSE (kcal mol^{-1})	MAE (kcal mol^{-1})	R^2
In-House	1511	FFNNs	**2.74 ±0.01**	**1.96 ±0.00**	**0.85 ±0.00**	3.35 ± 0.02	2.40 ± 0.02	0.77 ± 0.00
Alkoxy	298	RF	1.30 ± 0.03	1.12 ± 0.03	0.78 ± 0.01	**1.26 ±0.03**	**1.03 ±0.02**	**0.79 ±0.01**
Exp. alkoxy	15	RF	**1.29 ±0.03**	**0.97 ±0.02**	**0.62 ±0.02**	1.61 ± 0.02	1.25 ± 0.02	0.41 ± 0.02
Photoredox HAT	564	RF	1.36 ± 0.00	0.90 ± 0.01	0.92 ± 0.00	**0.84 ±0.02**	**0.60 ±0.01**	**0.97 ±0.00**
Exp. cumyloxyl	45	FFNNs	0.79 ± 0.04	0.66 ± 0.04	0.37 ± 0.16	**0.55 ±0.04**	**0.45 ±0.03**	**0.70 ±0.10**
Cytochrome P450	24	FFNNs	**1.09 ±0.08**	**0.95 ±0.11**	**0.47 ±0.08**	1.28 ± 0.01	1.14 ± 0.02	0.27 ± 0.02
Atmospheric HAT	73	FFNNs	2.30 ± 0.12	1.88 ± 0.08	0.62 ± 0.12	**2.08 ±0.10**	**1.60 ±0.10**	**0.75 ±0.04**

training data for the surrogate model – were also performed in the gas-phase (and at the same level-of-theory). Finally, the reactions in the In-House dataset were constructed by combining radicals and molecules from the same distribution as the BDE-db dataset, so one can expect that the QM descriptors predicted by the surrogate model are particularly appropriate and accurate for this dataset.

This stands in contrast to the other downstream datasets, where measurements/calculations were either performed in a solvent (or even in an enzymatic) environment, and/or molecules were involved that are (mostly) out-of-distribution with respect to the BDE-db database.

As a first test to verify this hypothesis, *i.e.*, that descriptors were able to outperform the hidden representation for the In-House dataset only because a (close to) "ideal" representation had been designed, we probed what would happen if we retrain the surrogate model with only part of the QM descriptors, identified as relevant in the VB analysis. Specifically, we kept the other configurations unchanged by re-using the source code and training set from Alfonso-Ramos *et al.*[15] and considered two alternative versions of the surrogate model where only the atom-level/molecule-level QM descriptors were kept as targets (Fig. 2).

As expected, we observe that when re-training the surrogate model with only atom-level descriptors, the performance of the descriptor-based model, evaluated on the In-House dataset, becomes significantly worse than the corresponding hidden representation-based model, due to a sharp drop in the accuracy of the former. Interestingly, also for the only other gas phase HAT reactivity dataset, Atmospheric HAT, we observe a major loss in model performance upon exclusion of the explicitly predicted molecule-level/atom-level descriptors, accentuating the superior performance of the hidden representation strategy for this dataset even further. Similar, though somewhat less pronounced, results are also obtained when re-training exclusively with molecule-level descriptors.

In direct contrast to the results for the descriptor-based strategy, the hidden-space input representations appear remarkably robust across the board, retaining an almost constant performance, regardless of which targets are selected for the surrogate model. As such, using hidden representations from the surrogate appears to be a safer option, especially when the set of QM descriptors has been selected in a suboptimal manner, *i.e.*, when major sources of variation in the downstream target are not covered by the prediction targets of the surrogate, and/or many of the descriptors are not causally linked with the variation of this target. An interpretation for this phenomenon through visualization can be found in S6 of the SI.

It should be emphasized that only in the case of the two very small downstream datasets—Exp. alkoxy and Cytochrome P450—do descriptors consistently yield better performance, regardless of whether all descriptors or only the atom-level/molecule-level subsets are considered. Additional experiments underscore that this effect is indeed exclusively linked to dataset size: when the amount of training data is progressively reduced for slightly larger datasets (where hidden representations normally outperform descriptor-based ones), the performance gap between the two strategies systematically narrows. In the extreme low-data regime (on the order of 50 samples), the compact descriptor representation tends to surpass hidden representations for these datasets as well (see Section S7 of the SI for details).

Finally, we would also like to note that the results presented above, for the full set of descriptors, are independent of the size of the hidden representation in the surrogate model. In Fig. 4, we show that, while the errors for both strategies are modulated somewhat by changing the hidden size, the overall trends, *i.e.*, whether the hidden representation-based or the descriptor-based strategy is the most performant, remain consistent per individual dataset. More results, along with re-training details, can be found in S8 of the SI.

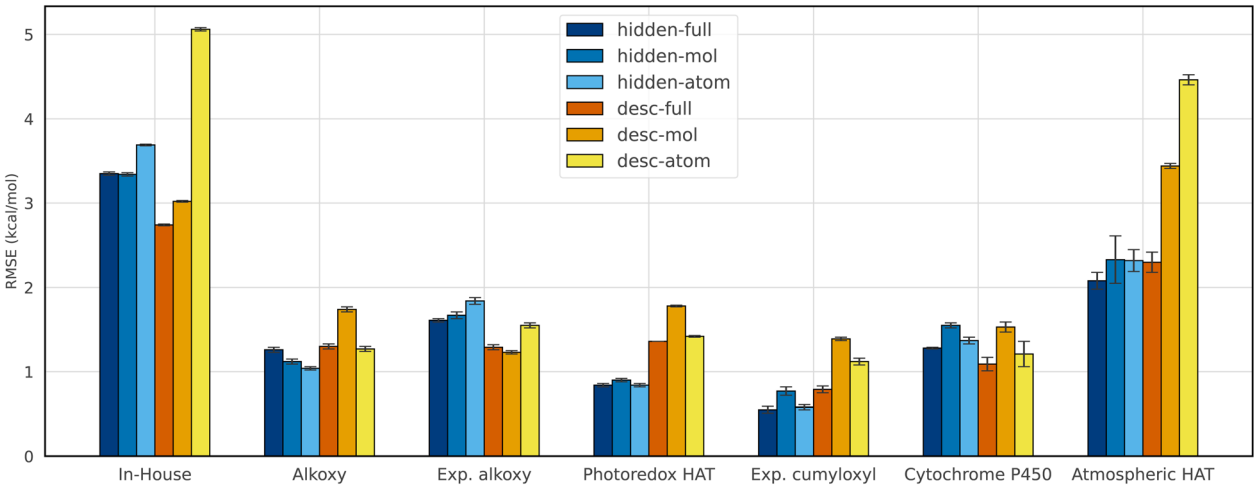

Fig. 2 Performance comparison between the strategy based on hidden representations and that based on descriptors when retraining the surrogate model using either the full set of descriptors (hidden-full and desc-full) or only the atom-level/molecule-level descriptors (hidden-atom/hidden-mol and desc-atom/desc-mol). The RMSE value (lower RMSE indicates better performance) of our proposed approach based on hidden space of the surrogate model is represented in blue, while that of the baseline is in orange/yellow.

Hidden representations *versus* descriptors for selectivity prediction

To further validate the conclusions drawn so far, we considered the datasets from the second case study. As discussed before, Guan *et al.*[1] did not select their set of QM descriptors based on an in-depth analysis. Instead, they simply opted to work with a selection of the most popular local QM descriptors used in the literature. As such, based on our preceding analysis, one would expect that the hidden representation strategy could potentially outperform the descriptor strategy as well.

Gratifyingly, this is exactly what we observe (see Table 3). For each of the downstream datasets considered, the hidden representation-based models achieve a 4–5% higher accuracy than the corresponding descriptor-based models. Interestingly, our hidden representation-based models even outperform the elaborate graph neural networks that integrate the surrogate model predicted QM descriptors, ml-QM-GNN, developed specifically for these datasets by Guan *et al.*[1] (except for the C–X dataset, where the two methods perform nearly the same). This observation further strengthens our hypothesis.

Table 3 10-fold cross validation for top-1 success rate of predicting the major reaction. The best results are in bold. The performance of predicted-descriptor-augmented GNN (ml-QM-GNN), hidden-representation-based FFNN (hidden-FFNN) and predicted-descriptor-based FFNN (desc-FFNN) are compared on three distinct datasets (C–H, C–X and Others) and their merged one (All). The *p*-values of the paired *t*-tests between desc-FFNN and hidden-FFNN are all below 0.001 across all datasets

Dataset	Size	Hidden-FFNN	Desc-FFNN	ml-QM-GNN
C–H	2244	**92.7% ± 0.3%**	88.9% ± 0.1%	92.4% ± 0.2%
C–X	1024	**96.5% ± 0.3%**	91.5% ± 0.6%	**96.6% ± 0.4%**
Others	552	**95.4% ± 0.4%**	91.5% ± 0.4%	94.5% ± 0.8%
All	3820	**94.0% ± 0.3%**	89.7% ± 0.4%	93.5% ± 0.2%

We should note here that it is not entirely inconceivable that, through the selection of a more complete set of descriptors (and taking for example solvent effects into account during their calculation/prediction), the hidden representation-based model could still be outperformed by a descriptor-based one. For the current selection, however, there is clearly no substantial benefit of using predicted descriptors.

Final validation on non-reactivity datasets

As a final confirmation of our hypothesis, we also considered our third case study.[14] In this case, the downstream tasks are not related to chemical reactivity prediction, and descriptor selection has been performed in a completely unprincipled manner, *i.e.*, a set of 20 popular/easy-to-compute molecule-level descriptors was selected, without any consideration of their physical connection to the respective downstream targets. As such, if our hypothesis is correct, the trends observed so far should become particularly pronounced here: we expect the hidden representation-based strategy to outperform the descriptor-based strategy by a significant margin. This is exactly what we observe in practice (Table 4). For all datasets, the hidden representation-based approach handily outperforms the descriptor-based approach. In the case of the regression tasks, the difference in accuracy can amount to over a factor of two.

Remarkably, despite the simplicity of the adopted downstream model architecture, the hidden representation strategy is even competitive with the elaborate QM-augmented graph neural network (ml-QM-GNN) approach for all datasets except QM9. In the case of QM9, the discrepancy between the hidden-FFNN and the ml-QM-GNN result is presumably due to a limited number of sub-tasks involving intensive properties. Indeed, the hidden-representation input based on molecular descriptors appears ill-suited to model properties that increase with molecule size (see Section S3 in the SI).

Table 4 10-fold cross validation comparison between FFNN based on the hidden space of molecule-level surrogate models (hidden-FFNN) and FFNN based on predicted molecule-level descriptors (desc-FFNN), where the better ones are in bold. The performance of ml-QM-GNN (Chemprop augmented with atom-level, bond-level and molecule-level descriptors) reported in the original paper[14] is added here for reference. For regression datasets ESOL, FreeSolv, and QM9, RMSE values are reported (geometric mean of RMSE is adopted for multitask data); for binary classification datasets HIV and ClinTox, ROC-AUC values are reported. For performance on individual targets in the multitask dataset, and the results for desc-FFNN using the raw predicted descriptors without RBF expansion, see Section S3 of the SI. The p-values of the paired t-tests between desc-FFNN and hidden-FFNN are all below 0.00001 across all datasets

Dataset	Size	Hidden-FFNN	Desc-FFNN	ml-QM-GNN
ESOL	1127	**0.646 ±0.008**	1.127 ± 0.008	0.539 ± 0.047
FreeSolv	642	**0.937 ±0.019**	2.358 ± 0.049	0.89 ± 0.16
QM9	133 885	**0.711 ±0.003**	1.901 ± 0.004	0.111 ± 0.004
HIV	41 127	**0.815 ±0.005**	0.714 ± 0.001	0.823 ± 0.029
ClinTox	1477	**0.892 ±0.007**	0.695 ± 0.008	0.871 ± 0.058

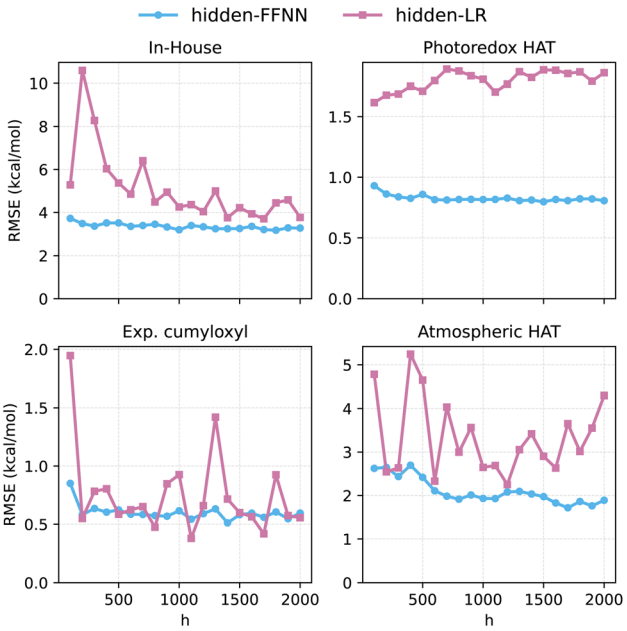

Fig. 3 Test RMSE of models based on hidden space with varying hidden sizes h (surrogate model trained with the full set of descriptors) for a selection of downstream datasets. The performance of FFNNs is compared to that of LR models.

Analyzing the linearity and roughness of the respective representations

Up to this point, we have provided empirical evidence that when downstream model performance is the main consideration, the hidden representation-based strategy generally outperforms the explicit (predicted) descriptor-based strategy.

In this final section, we will attempt to analyze the characteristics of the hidden spaces for the downstream tasks, with a particular focus on their respective linearity and roughness, in the hope of gaining some more insights into this observed behavior.

As noted by Tkatchenko and co-workers,[64] among others, to learn non-linearities, in particular interactions between the dimensions of a feature space, a sufficient number training points is typically needed. As such, one would naively expect that the smaller the size of the downstream dataset, the smaller the benefit of using an FFNN over a linear model.

To probe this, we compared the performance of our hidden-FFNN to its linear regression analog, hidden-LR, for a range of hidden space sizes ($h \in [100, 200, 300, \ldots, 2000]$), in the surrogate model for the HAT case study[15] (Fig. 3). Remarkably, while the performance of hidden-FFNN is relatively stable across h values, the errors achieved by hidden-LR fluctuate significantly; for some values of h, both downstream model architectures reach a similar performance, while for others, the hidden-FFNN outperforms hidden-LR by a significant margin. Note that this observation is valid for all the different downstream datasets. What this suggests is that the extent of linearity is not an inherently fixed property of the learned representation, but that the FFNN is equally good at dealing with either the fairly linear, as well as the non-linear, versions, regardless of the downstream dataset size. This implies that the hidden space inherently carries high-quality, readily exploitable information: the non-linearities that may emerge in the hidden space are generally not overly complex and easily learnable.

Interestingly, we also consistently observe that, among the dimensions of the hidden representation, there tends to be no significant co-linearity, *i.e.*, most of the dimensions are completely independent (*cf.* Section S9 of the SI), further underscoring the apparent richness of the learned representation in the surrogate model.

Finally, we also took a look at the roughness of the respective input feature spaces of the downstream models. Roughness metrics aim to quantify the "modellability" of structure–activity relationships (SAR) within a dataset, where rougher landscapes contain a greater number of large target property differences between molecules/reactions that are close in feature (or hidden) space. Such large property jumps across adjacent datapoints are commonly known as activity cliffs and increase the challenge of training a performant ML model.[65]

We selected an advanced, recently proposed roughness index ROGI-XD[66] to quantify the relationship between prediction targets and input feature spaces, with higher ROGI-XD values indicating rougher landscapes. Specifically, in the ROGI-XD approach, a dataset is progressively coarse-grained, and the evolution of the standard deviation of the targets throughout this coarse-graining process is tracked. In other words, the dataset is iteratively subdivided into clusters of increasing size, and at every instance, in each cluster, the labels of the individual samples are replaced with the average label for that cluster. Finally, the integral over the standard deviation of the (averaged) labels across cluster sizes is taken, resulting in the final roughness metric. When labels change only gradually across the feature space, the change in the standard deviation throughout the coarse-graining process will be limited, so that

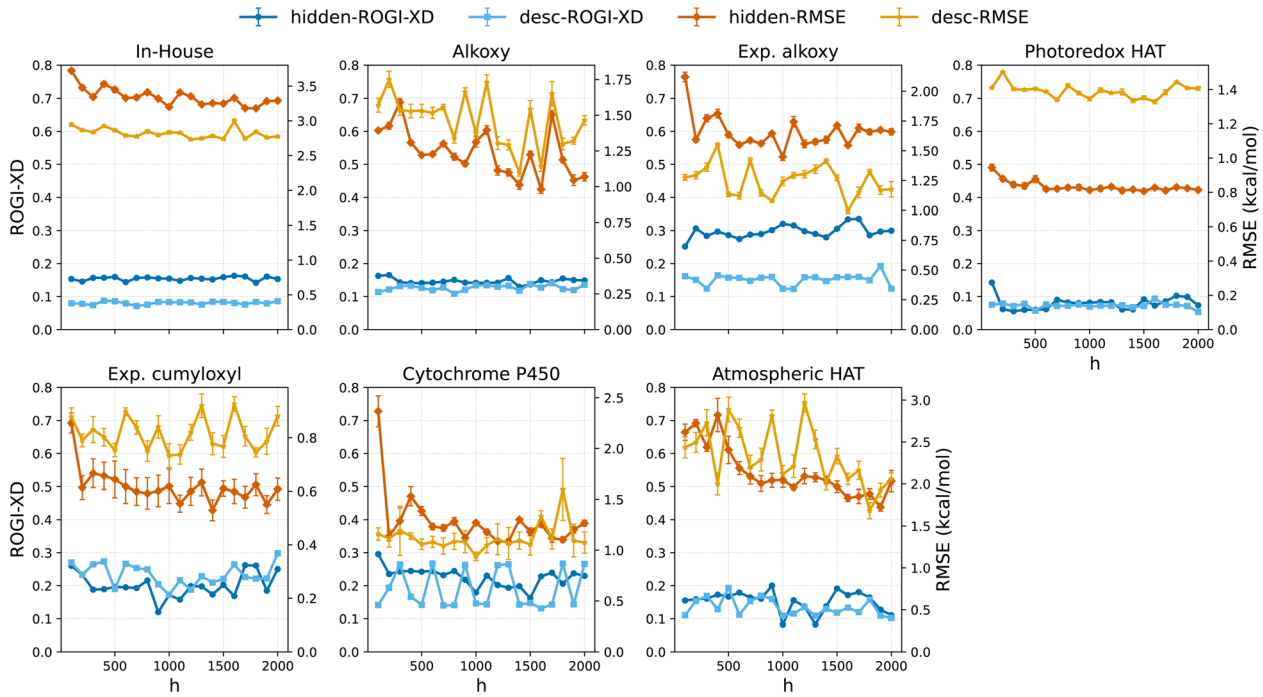

Fig. 4 Comparison between ROGI-XD values and model prediction performance (RMSE) based on hidden space or the predicted full set of descriptors. ROGI-XD labels (blue): "hidden-ROGI-XD" (darker blue) refers to the ROGI-XD computed on hidden representations, while "desc-ROGI-XD" (lighter blue) refers to the ROGI-XD computed on descriptor-based features. RMSE labels (orange): "hidden-RMSE" (darker orange) corresponds to the performance of hidden-FFNN from previous experiments, while "desc-RMSE" (lighter orange) refers to the best performance achieved using descriptor-based methods. Theoretically, lower ROGI-XD values indicate a smoother feature space that is more favorable for fitting a machine learning model, and thus tend to be associated with better prediction performance. Note that performance for both hidden space-based and predicted descriptor-based strategies across different hidden sizes is shown here (see Section S8 of the SI for more results).

the ROGI-XD value will be small. As such, this metric provides an assessment of the feature/hidden space's quality, and it has previously been demonstrated that this quantity correlates well with the performance of ML models.

As evident from Fig. 4, we observe that for the In-House and Exp. alkoxy datasets, the descriptor-based method results in a smoother landscape, and this agrees with the higher performance of the descriptor-based strategy for these downstream datasets. Curiously, for most of the other datasets, the ROGI-XD values are quite similar for both strategies, though the hidden representation-based model tends to outperform the descriptor-based one as indicated in Table 2 and Fig. 2. The reason for this seemingly asymmetric behavior may be that roughness metrics such as ROGI-XD presumably become somewhat misleading when very large hidden spaces are considered, due to the presence of many potentially "unproductive" dimensions (see Section S10 in the SI for a toy model analysis). In other words, for feature spaces with many dimensions, ROGI-XD values can be artificially inflated compared to more compact feature spaces.

Overall, it appears that despite ROGI-XD's limitations, differences in roughness do seem to explain, to a reasonable extent, the observed trends in model performance. Nonetheless, an unequivocal causal explanation for the trends observed throughout this work is still lacking. This will be the focus of future research.

Conclusions

This study aimed to compare two strategies for incorporating surrogate quantum chemical models into predictive chemistry workflows in the data-limited regime: one based on explicitly predicted QM descriptors, and one based on the learned hidden representations of the surrogate model. Across a broad set of downstream tasks—spanning reactivity, selectivity, and molecular property prediction—the hidden representation-based approach consistently outperformed the descriptor-based alternative in most scenarios. Descriptor-based models only showed an advantage when the descriptors were carefully selected and closely aligned with the target property, which is uncommon in practice. In contrast, the hidden representations provided a more flexible and data-efficient alternative, capturing nuanced chemical information without requiring manual feature selection. These representations also proved robust to changes in surrogate model architecture and descriptor targets, and could be effectively leveraged even in low-data regimes.

We acknowledge, however, a key trade-off: while hidden representations offer superior performance, they lack the transparency of explicit descriptors, making feature attribution and physical interpretability more difficult. As such, practitioners may still prefer explicit descriptors when mechanistic insight, feature importance analysis, or human-understandable

model behavior is of central importance—particularly in hypothesis generation or experimental design.

Nonetheless, when predictive accuracy is the primary concern, our results strongly support the use of hidden representations from well-trained surrogate models. Their consistent outperformance across diverse applications suggests that, in most practical scenarios, they represent the more effective choice for enhancing downstream models in predictive chemistry.

Author contributions

G. C.: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, visualization, writing. T. S.: conceptualization, supervision, funding acquisition, writing.

Conflicts of interest

There are no conflicts to declare.

Data availability

All the code and data used in this work (3 case studies and the synthetic experiments for ROGI-XD) are openly available in Zenodo at **https://doi.org/10.5281/zenodo.17100503**. The repository is maintained at **https://github.com/chimie-paristech-CTM/Hidden_vs_Desc**. For training and test sets of the surrogate model in the first case study, see **https://figshare.com/projects/Hydrogen_atom_transfer_reactions/188007** by Alfonso-Ramos *et al.*[15]

Supplementary information is available. See DOI: **https://doi.org/10.1039/d5dd00256g**.

Acknowledgements

G. C. thanks the China Scholarship Council (CSC) for a PhD fellowship (No. 202406020083). T. S. acknowledges the French National Agency for Research (ANR) for a CPJ Grant (ANR-22-CPJ1-0093-01).

Notes and references

1 Y. Guan, C. W. Coley, H. Wu, D. Ranasinghe, E. Heid, T. J. Struble, L. Pattanaik, W. H. Green and K. F. Jensen, *Chem. Sci.*, 2021, **12**, 2198–2208.
2 A. Gangwal, A. Ansari, I. Ahmad, A. K. Azad and W. M. A. W. Sulaiman, *Comput. Biol. Med.*, 2024, **179**, 108734.
3 A. Nandy, C. Duan and H. J. Kulik, *Curr. Opin. Chem. Eng.*, 2022, **36**, 100778.
4 N. Casetti, J. E. Alfonso-Ramos, C. W. Coley and T. Stuyver, *Chem.–Eur. J.*, 2023, **29**, e202301957.
5 L. C. Gallegos, G. Luchini, P. C. St. John, S. Kim and R. S. Paton, *Acc. Chem. Res.*, 2021, **54**, 827–836.
6 J. M. Crawford, C. Kingston, F. D. Toste and M. S. Sigman, *Acc. Chem. Res.*, 2021, **54**, 3136–3148.
7 L. M. Sigmund, M. Assante, M. J. Johansson, P.-O. Norrby, K. Jorner and M. Kabeshov, *Chem. Sci.*, 2025, **16**, 5383–5412.
8 T. Gensch, G. dos Passos Gomes, P. Friederich, E. Peters, T. Gaudin, R. Pollice, K. Jorner, A. Nigam, M. Lindner-D'Addario, M. S. Sigman, *et al.*, *J. Am. Chem. Soc.*, 2022, **144**, 1205–1217.
9 K. Jorner, T. Brinck, P.-O. Norrby and D. Buttar, *Chem. Sci.*, 2021, **12**, 1163–1175.
10 A. M. Żurański, J. Y. Wang, B. J. Shields and A. G. Doyle, *React. Chem. Eng.*, 2022, **7**, 1276–1284.
11 S. H. Newman-Stonebraker, S. R. Smith, J. E. Borowski, E. Peters, T. Gensch, H. C. Johnson, M. S. Sigman and A. G. Doyle, *Science*, 2021, **374**, 301–308.
12 T. Stuyver and C. W. Coley, *Chem.–Eur. J.*, 2023, **29**, e202300387.
13 T. Stuyver and C. W. Coley, *J. Chem. Phys.*, 2022, **156**, 084104.
14 S.-C. Li, H. Wu, A. Menon, K. A. Spiekermann, Y.-P. Li and W. H. Green, *J. Am. Chem. Soc.*, 2024, **146**, 23103–23120.
15 J. E. Alfonso-Ramos, R. M. Neeser and T. Stuyver, *Digital Discovery*, 2024, **3**, 919–931.
16 B. C. Haas, M. A. Hardy, S. S. SV, K. Adams, C. W. Coley, R. S. Paton and M. S. Sigman, *Digital Discovery*, 2025, **4**, 222–233.
17 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, *Sci. Data*, 2014, **1**, 140022.
18 C. Isert, K. Atz, J. Jiménez-Luna and G. Schneider, *Sci. Data*, 2022, **9**, 273.
19 R. M. Neeser, C. Isert, T. Stuyver, G. Schneider and C. W. Coley, *Chem. Data Collect.*, 2023, **46**, 101040.
20 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2011, **40**, D1100–D1107.
21 A. Madushanka, R. T. Moura and E. Kraka, *Sci. Data*, 2024, **11**, 1376.
22 J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad and R. G. Coleman, *J. Chem. Inf. Model.*, 2012, **52**, 1757–1768.
23 Y. Zhu, M. Li, C. Xu, *et al.*, *Sci. Data*, 2024, **11**, 948.
24 P. C. St. John, Y. Guan, Y. Kim, B. D. Etz, S. Kim and R. S. Paton, *Sci. Data*, 2020, **7**, 244.
25 D. Balcells and B. B. Skjelstad, *J. Chem. Inf. Model.*, 2020, **60**, 6135–6146.
26 J. Xia, Y. Zhu, Y. Du and S. Z. Li, *arXiv*, 2022, preprint, arXiv:2210.16484, DOI: **10.48550/arXiv.2210.16484**.
27 E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green and C. J. McGill, *J. Chem. Inf. Model.*, 2024, **64**, 9–17.
28 S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2020, preprint, arXiv:2010.09885, DOI: **10.48550/arXiv.2010.09885**.
29 R. Irwin, S. Dimitriadis, J. He and E. J. Bjerrum, *Mach. learn.: sci. technol.*, 2022, **3**, 015022.
30 N. Shoghi, A. Kolluru, J. R. Kitchin, Z. W. Ulissi, C. L. Zitnick and B. M. Wood, *arXiv*, 2023, preprint, arXiv:2310.16802, DOI: **10.48550/arXiv.2310.16802**.
31 X. Li, S.-Q. Zhang, L.-C. Xu and X. Hong, *Angew. Chem., Int. Ed.*, 2020, **59**, 13253–13259.
32 T. Stuyver and C. W. Coley, *J. Chem. Phys.*, 2022, **156**, 084104.

33 J. P. Janet and H. J. Kulik, *J. Phys. Chem. A*, 2017, **121**, 8939–8954.
34 D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186–190.
35 S. Ma, S. Wang, J. Cao and F. Liu, *ACS Omega*, 2022, **7**, 34858–34867.
36 Q. An, Z. Wang, Y. Chen, X. Wang, K. Zhang, H. Pan, W. Liu and Z. Zuo, *J. Am. Chem. Soc.*, 2020, **142**, 6216–6226.
37 L.-C. Yang, X. Li, S.-Q. Zhang and X. Hong, *Org. Chem. Front.*, 2021, **8**, 6187–6195.
38 M. Salamone, M. Galeotti, E. Romero-Montalvo, J. A. van Santen, B. D. Groff, J. M. Mayer, G. A. DiLabio and M. Bietti, *J. Am. Chem. Soc.*, 2021, **143**, 11759–11776.
39 P. W. Gingrich, J. B. Siegel and D. J. Tantillo, *Chem.: Methods*, 2022, **2**, e202100108.
40 M. Tavakoli, Y. T. T. Chiu, P. Baldi, A. M. Carlton and D. Van Vranken, *J. Chem. Inf. Model.*, 2023, **63**, 1114–1123.
41 M. Tavakoli, Y. T. T. Chiu, P. Baldi, A. M. Carlton and D. Van Vranken, *J. Chem. Inf. Model.*, 2023, **63**, 1114–1123.
42 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
43 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.
44 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
45 NextMove Software, *Pistachio*, **https://www.nextmovesoftware.com/pistachio.html**.
46 P. Fuentealba, P. Pérez and R. Contreras, *J. Chem. Phys.*, 2000, **113**, 2544–2551.
47 P. Geerlings, F. De Proft and W. Langenaeker, *Chem. Rev.*, 2003, **103**, 1793–1874.
48 R. Verma and C. Hansch, *Chem. Rev.*, 2011, **111**, 2865–2899.
49 L. Ruddigkeit, R. Van Deursen, L. C. Blum and J.-L. Reymond, *J. Chem. Inf. Model.*, 2012, **52**, 2864–2875.
50 *Tox21 Challenge*, **https://tripod.nih.gov/tox21/challenge/**.
51 K. L. Dionisio, K. Phillips, P. S. Price, C. M. Grulke, A. Williams, D. Biryol, T. Hong and K. K. Isaacs, *Sci. Data*, 2018, **5**, 180125.
52 J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.
53 D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, *Nucleic Acids Res.*, 2006, **34**, D668–D672.
54 B. A. Koscher, R. B. Canty, M. A. McDonald, K. P. Greenman, C. J. McGill, C. L. Bilodeau, W. Jin, H. Wu, F. H. Vermeire, B. Jin, T. Hart, T. Kulesza, S.-C. Li, T. S. Jaakkola, R. Barzilay, R. Gómez-Bombarelli, W. H. Green and K. F. Jensen, *Science*, 2023, **382**, eadi1407.
55 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
56 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
57 A. E. Reed, R. B. Weinstock and F. Weinhold, *J. Chem. Phys.*, 1985, **83**, 735–746.
58 L. R. Domingo, P. Pérez and J. A. Sáez, *RSC Adv.*, 2013, **3**, 1486–1494.
59 J. S. Delaney, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1000–1005.
60 D. L. Mobley and J. P. Guthrie, *J. Comput.-Aided Mol. Des.*, 2014, **28**, 711–720.
61 National Cancer Institute Developmental Therapeutics Program, *AIDS Antiviral Screen Data*, **https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data**.
62 Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chem. Sci.*, 2018, **9**, 513–530.
63 Pande Group at Stanford, *MoleculeNet*, **https://moleculenet.org/**.
64 A. E. A. Allen and A. Tkatchenko, *Constructing Effective Machine Learning Models for the Sciences: A Multidisciplinary Perspective*, 2022, **https://arxiv.org/abs/2211.11680**.
65 M. Aldeghi, D. E. Graff, N. Frey, J. A. Morrone, E. O. Pyzer-Knapp, K. E. Jordan and C. W. Coley, *J. Chem. Inf. Model.*, 2022, **62**, 4660–4671.
66 D. E. Graff, E. O. Pyzer-Knapp, K. E. Jordan, E. I. Shakhnovich and C. W. Coley, *Digital Discovery*, 2023, **2**, 1452–1460.