RELIABLE EVALUATION FOR SMALL MOLECULE MACHINE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Cross-validation (CV) allows us to assess the generalizability of machine learning models for small molecule property prediction. Yet, checking whether evaluation results carry over to the target application domain is non-trivial. We introduce a method for comparing any CV split of small molecules against any chosen domain of application. We find that existing splitting methods produce test sets far from the application domain and are prone to data leakage. Next, we introduce a new splitting method based on Maximum Common Edge Subgraph (MCES) distances, and show that resulting CV splits align much better with the application domain of biomolecules. We evaluate five baseline and nine foundation models on different splits, and find that the strongest baseline consistently outperforms all but four of the foundation models on commonly used experimental datasets. All models show a drop of about 0.06 and 0.1 AUROC for MCES splits, indicating that previous splits substantially overstate real-world performance. Notably, pretraining data of foundation models overlaps substantially with downstream evaluation datasets. In the future, our methods may help to reach reliable evaluations for other graph-based machine learning tasks.

1 INTRODUCTION

Small molecule machine learning has prospered during the last decade (Gilmer et al., 2017; Wu et al., 2018; Stokes et al., 2020). Previously, quantitative structure activity relationships (QSAR) models relied on physicochemical modeling, providing mechanistic insights into the chemical or biochemical processes. Machine learning models sacrifice mechanistic insight for much better performance in practice. Recent years have seen a growing number of models trained on large-scale data to learn representations of small molecules that capture chemical structure and properties, commonly referred to as *foundation models*. Datasets can be either from quantum chemistry computations or from experimental measurements made using authentic standards. In the following, we will concentrate on experimental datasets, measuring 'molecular properties' such as toxicity or solubility. Datasets usually contain below 10k compounds.

When there are computational limitations due to massive datasets and excessive training times, as for large language models, it is common practice to split the data into training, testing, and optionally validation portions. In a small or medium data regime, *cross-validation* (CV) is preferable. This guarantees that every sample is used for evaluation exactly once, and avoids results that hinge on a single fortunate split. Initially, models for small molecule machine learning were evaluated on uniform splits, somewhat imprecisely referred to as 'random splits'. Unfortunately, uniform splits usually result in a massive overestimation of model performance, in particular for experimental datasets: When creating an experimental dataset, we are mostly restricted to authentic standards that we can buy. Moreover, in synthetic chemistry, if we can synthesize some molecule A, we can usually also synthesize a molecule A' identical to A except for some minor change, such as the addition or removal of a hydroxy group. Consequently, small molecules in these datasets cluster; the overlap between any two datasets is usually much larger than expected by chance. Sorting A and A' into different folds results in data leakage (Kapoor & Narayanan, 2023) and, consequently, in a massive overestimation of method performance in application. To prevent this, we need a notion of distance between molecules that lets us detect similar ones and keep them within the same fold.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The Maximum Common Edge Subgraph (MCES) distance (Bokhari, 1981) agrees well with the chemical intuition of structural similarity between small molecules: In particular, the above-mentioned cases where two small molecules are identical except for a single hydroxy group correspond to the minimum non-zero distance of one. (Distance $\frac{1}{2}$ is chemically impossible for two molecular graphs.) Until recently, computing MCES distances was extremely demanding due to the underlying hardness of the problem; notably, the problem generalizes CLIQUE. In 2025, Kretschmer et al. (2025) introduced the *myopic MCES distance* where only small distances are computed exactly, whereas lower bounds are used for larger distances. This makes the computation feasible at the scales considered here.

Here, we propose a new method for crafting cross-validation folds, based on the MCES distance. Importantly, our take is not that our folds are 'harder' than previous splitting methods — which turns out to be true — but rather, that we can, for the first time, measure how close our cross-validation results are in comparison to our desired domain of application, resulting in a more reliable evaluation. We find that the new folds align substantially better with our application domain than those from other splitting methods. Unfortunately, for several datasets, it is impossible to establish how well resulting models will perform in application, due to massive clustering of compounds.

For foundation models, we evaluate whether these models can truly provide 'foundational' embeddings by using frozen weights. A pretrain/fine-tune approach results in better evaluation statistics, but again, has very limited informative value for performance in applications Praski et al. (2025). We train different classifiers using the latent representation of molecular structures: *Linear probing* (LP) implicitly assumes that the latent representations are linearly separable with regard to the downstream task; we argue that this is unlikely for complex biological and biochemical properties such as toxicity. Consequently, we also evaluate Multilayer Perceptron (MLP) probing and random forests (RF). Using the MCES split, all models show substantially worse performance than for uniform and scaffold splits (Table 1). Baseline models outperform the majority of foundation models, while both show a similar performance drop. Finally, we test for structural exposure, i.e., an overlap of molecular structures, during pretraining of foundation models, and observed substantial overlap between compounds in our evaluation sets and the foundation model pretraining data.

2 RELATED WORK

Structure-aware splits and benchmarks. It was observed that uniform splitting of small molecule datasets results in overestimation of performance (Sheridan, 2013). Other methods for deriving cross-validation folds were suggested over the years: One of the earlier such methods is based on scaffolds (Bemis & Murcko, 1996; Hu & Bajorath, 2010; Hu et al., 2011). Different scaffold definitions exist (Bemis & Murcko, 1996; Hu & Bajorath, 2010), resulting in slightly different folds. As an alternative, Butina (1999) suggested to cluster small molecules based on fingerprint Tanimoto distances in 1999. In both cases, resulting clusters can be distributed into the desired number of folds by, say, a greedy distribution algorithm. Guo et al. (2025) argued that scaffold folds are still 'too easy', in the sense that evaluation results are still far better than what is observed in practice. They proposed a new method based on Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) to cluster samples. Recently, Reimer et al. (2026) found that for many datasets, scaffold and UMAP splits are not more challenging than uniform splits. Additionally Joeres et al. (2025) introduced a method for two-dimensional splitting. In contrast, time-based splitting (Sheridan, 2013) is usually not applicable for these datasets, or small molecule machine learning in general. Previous ad hoc evaluations using MCES distances were introduced in (Kretschmer et al., 2024; Bushuiev et al., 2024; Gupta & Skinnider, 2026). In these works, either very small thresholds for MCES distances were chosen, or the resulting folds were highly imbalanced.

In 2018, Wu et al. (2018) introduced the MoleculeNet benchmark for small molecule property prediction, providing uniform and scaffold splits for ten experimental datasets. At present, this is the most widely used benchmark for this task. An alternative collection of datasets was later provided by Huang et al. (2021; 2022). Throughout this paper, we concentrate on those datasets where labels were experimentally determined. See Appendix A for a discussion of quantum chemistry datasets. Experimental datasets are restricted in size and distribution by the massive cost and labor of synthesizing small molecules and determining the experimental values.

Foundation models. Numerous so-called small molecule *foundation models* have been developed during the last decade (Rong et al., 2020; Fabian et al., 2020; Ahmad et al., 2022; Ross et al., 2022; Wang et al., 2022; Edwards et al., 2022; Irwin et al., 2022; Zhou et al., 2023; Seidl et al., 2023; Soares et al., 2025). These models provide a latent representation of molecular structure. Foundation models are pretrained on datasets substantially larger than the downstream tasks considered here; pretraining datasets often contain millions to billions of molecular structures. Given the limitations of experimental small molecule datasets considered here, foundation models hold the promise of improving the generalization power and transfer capability for a downstream task. To that end, all weights of the foundation model have to be kept frozen (static embeddings). Fine-tuning complex models will lead to better evaluation statistics, but may result in severe overfitting for the low to medium data regime considered here (Praski et al., 2025) and, consequently, undermine the goal of better generalization.

Complex foundation models and model architectures do not necessarily result in better performance. Sun et al. (2022) found that pretraining graph neural networks often yields no statistically significant gain over training from scratch. Deng et al. (2023) found limited benefits from using foundation models on small molecule datasets. Praski et al. (2025) evaluated 25 foundation models under scaffold splits; only a single foundation model performed significantly better than a baseline method.

3 MCES FOLDS

Given two graphs G and G', the *maximum common edge subgraph* problem is to find a graph H with as many edges as possible which is isomorphic to both a subgraph of G and a subgraph of G'. Exact MCES computation is hard, since the problem is NP-HARD. Let E, E', F be the edge sets of G, G', H, respectively. The MCES distance $d(G, G') = |E| + |E'| - 2|F|$ equals the minimum number of edge edit operations transforming G into G' (Bokhari, 1981). For molecular structures, we ignore all hydrogen atoms and focus on heavy atoms. Consequently, two molecular structures differing by a single hydroxy group (OH) are at distance one. Kretschmer et al. (2025) introduced the *myopic* MCES distance, where exact values are computed only up to a threshold T, whereas a lower bound is returned for values above T. In addition, different bond types are considered, so that double bonds can be transformed into single bonds, potentially resulting in non-integer distances from $\frac{1}{2}\mathbb{N}$ due to aromatic bonds.

3.1 THE UNBOUNDED ALGORITHM

In the first step of our algorithm, we cluster molecules; in the second step, we merge clusters into the desired number of folds k. For our initial clustering, we choose a threshold $\tau < T$, which is the targeted minimum cross-cluster distance. Whereas T bounds which individual distances are computed exactly, τ bounds which molecules end up in the same cluster. Let V be the molecules of a dataset and $n := |V|$. To generate our folds, we first compute MCES distances between all pairs of molecules in the dataset. Clearly, solving $\binom{n}{2}$ instances of an NP-hard problem is time-demanding. We speed up computations by computing the myopic MCES distance with threshold $T := \tau + \frac{1}{2}$: Kretschmer et al. (2025) observed an average running time of roughly 100 ms per instance and core for reasonable thresholds. These computations have to be carried out only once per experimental dataset, and are very fast in comparison to the time required to measure the dataset.

We now cluster molecules V based on the MCES distances using hierarchical clustering; this will later allow us to choose the largest clusters within our requirements. We want to maximize the minimum distance between folds, naturally corresponding to *single linkage* clustering. In contrast, UPGMA (unweighted pair group method with arithmetic mean) would not guarantee that cross-cluster distances are above the threshold. Single-linkage clustering results in a rooted tree (*dendrogram*) where nodes are labeled with the minimum distance of two samples from different subtrees below this node, the *height* of the node. We can compute this agglomerative clustering in time $O(n^2)$ (Müllner, 2013); in any case, this time is negligible in comparison to the computation of MCES distances, and we may also resort to naive cubic-time algorithms.

Let $G_\tau = (V, E_\tau)$ with $E_\tau := \{uv : d(u, v) < \tau\}$. Cutting the single-linkage dendrogram at height τ yields exactly the connected components of G_τ. For components $C_i \neq C_j$ and $u \in C_i$, $v \in C_j$ this implies $d(u, v) \geq \tau$. Consequently, $\tau < T$ implies that all computations relevant for clustering

are exact, meaning that the myopic MCES distance results in the same clustering as the exact MCES distance. Notably, using UPGMA instead of single linkage would thwart this argument, as we would average between exact MCES values up to T and inexact values above T.

Unfortunately, this simple method has a problem: As τ grows, G_τ gains edges until it merges into a single component. Hence, we cannot use the resulting degenerated clustering to construct k folds. Yet, problems start much earlier: Even if we end up with k or more clusters, we observe that one of the clusters often becomes very large, much larger than the maximum size of n/k that would result in a balanced split. Assume we want to enforce that all clusters have size below or equal to αn, for a certain $\alpha \in (0, 1]$. We want to determine the maximal MCES distance thresholds so that no cluster violates the condition. Consider a threshold $\tau \in \frac{1}{2}\mathbb{N}$. Let $s(\tau)$ be the size of the largest cluster for this threshold, that is, the largest connected component in G_τ. We define

$$\tau_*(\alpha) := \max\{\tau \in \tfrac{1}{2}\mathbb{N} \,:\, s(\tau)/n \leq \alpha\} \,.$$

This function is monotonically increasing, and takes values in $\frac{1}{2}\mathbb{N}$ due to $\tau_*(\alpha) \leq \max\{d(u, v)\}$.

Clearly, we can lower the threshold τ to the point where all cluster sizes are below the desired size threshold. In practice, we found that $\tau_*(\alpha)$ can be very small, even for moderate values of α. In part, this is due to a peculiarity of the MCES distance: The distance between 'very small' molecular structures is usually very small, too. For a molecular graph with only ten nodes, we do not need many edit operations to transform it into another molecular graph with ten nodes. Consequently, one of our clusters will often contain all 'very small' molecules. Experimental datasets usually contain a large fraction of such 'very small' molecules. In addition, we often observe 'chaining' of molecules, where even for a moderate choice of τ such as $\tau = 6$, the vast majority of molecules in the dataset are connected via chains of pairwise similar molecules and, hence, form a single connected component (Figs. 2 and 5). To ensure that those characteristics do not result in one fold becoming too large, we would have to substantially lower τ, which is not appropriate for the remaining molecules.

3.2 Variable-size splits

In the following, we assume for simplicity that k divides n. Let $(p_1, \ldots, p_k)$ be the split distribution, $0 < p_i < 1$ and $\sum_i p_i = 1$, where $p_i n$ is the size of fold i. A *same-size* split enforces $p_i = 1/k$, and a split is *variable-size* if it is not same-size. Variable-size splits come with certain advantages and disadvantages. We do CV to estimate the performance of our model, trained on all the available training data, if we apply it to new, unseen data. For a same-size 5-fold CV, we evaluate the performance of a model that has been trained on 80 % of the data, for every sample. We use the resulting statistics as a proxy of the performance of a model trained on 100 % of the data. Variable-size splits introduce additional variability into our evaluation: different models are trained on different fractions of the data, and assuming that more training data results in better performance, this translates into variability of our statistics. Next, models in a variable-size CV are trained using a smaller fraction of the training data on average than for a same-size split, namely, $1 - \sum_i p_i^2 \leq 1 - 1/k$. Consequently, if the size distribution varies substantially from a same-size split, evaluation statistics may no longer be a good proxy for the model trained on 100 % of the training data. On the positive side, variable-size splits allow us to better enforce the requirement that no sample should be 'too similar' to the training data. Additional care has to be taken when reporting the between-fold variance of statistics: Here, weighted mean and weighted variance must be computed.

3.3 Bounding the cluster size

In the following, we want to enforce that no cluster is larger than αn, for a chosen $\alpha \geq 1/k$. Clearly, this will also force our method to accept smaller distance thresholds; yet, these smaller distances will only be accepted locally, that is, for clusters that do exceed the threshold. Given the dendrogram from above, let $h(v)$ be the height and n_v the number of leaves below node v. Clearly, both can be computed via a tree traversal in linear time. We call v *valid* if $h(v) < \tau$ and $n_v \leq \alpha n$. In a second tree traversal, we output each node v that is valid but has an invalid parent. The leaves below every such v form one cluster. This can be carried out in time linear in the size of the tree and, hence, total time $O(n)$. Notably, for a fixed τ, decreasing α starting at 1.0 has no effect until αn drops below the size of the largest cluster of the unconstrained cut; below this point, cluster are devided into subclusters as α decreases.

3.4 Merging clusters into folds

In the end, we distribute resulting clusters into the desired number of folds using a greedy distribution algorithm: We sort clusters by size, from largest to smallest; we initialize k empty folds; we iterate over the clusters, and add the current cluster to the fold with the smallest cardinality. The greedy algorithm will try to 'even out' all folds, which usually results in folds of similar size, except for the largest one. If desired, other optimization criteria such as stratification or maximum average distance can be taken into account in this step.

It is possible that merging clusters results in folds that are larger than the largest cluster and, potentially, also larger than the size threshold αn: As a simple example, consider $k + 1$ clusters of equal size. Each cluster has size $1/(k + 1) < 1/k$; yet, the largest fold must have size $2/(k + 1) \gg 1/k$. In practice, we did not observe this behavior: As noted, we usually observe one very large cluster, whereas the size of subsequent clusters quickly decreases (Fig. 2).

4 Estimating the evaluation gap

Models often perform substantially worse in practice than in evaluations, due to differences in distributions between evaluation data and application domain. This is a particularly pressing problem for experimental small molecule datasets, as these are severely restricted by what can be synthesized and commercially acquired for a reasonable price. We use the distance between a query molecular structure and the training data, measured by the myopic MCES distance, as a proxy for the 'difficulty' of the sample (Sheridan et al., 2004; Sahigara et al., 2012). The MCES distance is a chemically well-founded measure of chemical dissimilarity; our experimental results below also clearly indicate that larger MCES distances indeed result in worse evaluation statistics, see also (Kretschmer et al., 2024; Bushuiev et al., 2024; Gupta & Skinnider, 2026). Doing so, allows us to investigate and compare the distributions of MCES distances and 'sample difficulties' for different cross-validation split.

More importantly, this method allows us to estimate 'sample difficulty' for an *arbitrary application domain* of the model. We assume that any model is trained with a certain application domain in mind. Throughout this paper, we focus on frequently used datasets containing experimental measurements of chemical properties (say, solubility) and bioactivity (say, toxicity). Here, a natural application domain is the entirety of all *small molecules of biological interest* (metabolites, drugs, and drug degradation products, synthetic molecules regularly observed in biological samples). Another application domain may be all small molecules that were previously tested for bioactivity, say, in the ChEMBL database (Zdrazil et al., 2024). Our analysis is agnostic to the choice of application domain; yet, it requires to *explicitly state the desired application domain*, avoiding implicit and vague assumptions about this domain.

For any chosen application domain, we can now compute the same distribution as for a CV split: Solely to speed up computations, we first choose a uniform subsample of molecules from the application domain set of appropriate size. Importantly, we do not have to know any labels for these molecular structures; the structures by themselves are sufficient. Then, for every molecule in the set, we compute the minimum MCES distance to any molecule in the complete training data. We can then compare this distribution with the distribution used in the cross-validation evaluation, both visually and by different statistics such as mean or median. Large shifts in distribution indicate that the performance of the model in evaluation is likely different from its performance on application.

5 Experimental setup

Datasets. We use ten datasets with experimental measurements from the MoleculeNet benchmark (Wu et al., 2018), both for classification and regression tasks. Datasets were downloaded from the repository of the ChemProp framework (Heid et al., 2024), `https://github.com/chemprop/chemprop`. We cleaned datasets, removing invalid molecules and doublets (Table 2 and Appendix E). See Appendix D for dataset descriptions and details.

As an exemplary domain of application, we use a set of molecules of biological interest that were collected from 14 small molecule databases, then subsampled to the order 20k, see (Kretschmer et al., 2025) for details. The PubChem-standardized (Kim et al., 2016)

Simplified Molecular Input Line Entry System (SMILES) representations of these molecular structures were obtained from https://raw.githubusercontent.com/boecker-lab/myopic-mces-data/refs/heads/main/biostructures_20k.csv.

MCES computations. We computed myopic MCES distances between all molecular structures in each dataset using the Python package `myopic-mces` (Kretschmer et al., 2025) version 1.0.0. We chose $T = 30$, notably higher than $\tau = 10$, mainly to provide informative distribution plots. We enabled the option of choosing MCES bounds dynamically, meaning that a weak lower bound suffices for instances with a distance above the threshold, resulting in faster computations. Likewise, we employ a time limit of 5 min for exact computations; for instances where the computation does not finish within this time a lower bound is used, even when the distance is below the threshold. This applies to a very small number of computations: For example, only 0.022 % for all distances between the 20k subsample of biomolecules from Kretschmer et al. (2025). Doing so, $\tau < T$ would no longer guarantee exact MCES computations; given that we use $T = 30 \gg \tau = 10$ in our computations, this is irrelevant for all practical considerations. We encountered timeouts only for the HIV and Lipo datasets, with 66 (0.005 %) and 4 (0.01 %) inexact values below τ, respectively.

Splits. Uniform splits were computed using scikit-learn 1.9.0 (Pedregosa et al., 2011). Scaffold and Butina splits were computed using RDKit (version 2026.03.5, https://www.rdkit.org), greedily merging clusters. For scaffold splits we used Bemis-Murcko scaffolds. For UMAP splits we followed Guo et al. (2025): we embedded Extended Connectivity Fingerprints of diameter 4 (ECFP4, Rogers & Hahn (2010)) with UMAP (McInnes et al. (2020), umap-learn version 0.5.12, 30 neighbors), clustered the embeddings with Ward agglomerative clustering in scikit-learn, and greedily merged the resulting clusters into folds. We compute MCES splits with $\tau = 10$, following existing literature (Bushuiev et al., 2024; Kretschmer et al., 2025), for three different values of α, namely $\alpha \in \{0.3, 0.4, 0.5\}$. Here, $\alpha = 0.3$ corresponds to an 'almost same-size split'; differences of evaluation statistics to a hypothetical same-size split are presumably negligible. In contrast, $\alpha = 0.5$ allows one of the folds to contain half of the samples; for many tasks and training datasets, results should nevertheless be transferable to a model trained on all of the training data.

Foundation and baseline models. We evaluate five baseline and nine foundation models. We used a unified pipeline, in which the model is a configurable component, so that every model is trained and evaluated on exactly the same folds and comparisons remain fair. We evaluate on MCES splits as well as uniform and scaffold splits, as the latter two are most widely used. A detailed description of the foundation models can be found in Appendix C.

As baselines, we use random forests (RF) (Breiman, 2001) and XGBoost (XGB) (Chen & Guestrin, 2016) on ECFP4 fingerprints (Rogers & Hahn, 2010), a Graph Convolutional Network (GCN) (Kipf & Welling, 2017), a directed message-passing neural network (D-MPNN) (Yang et al., 2019; Heid et al., 2024) and a kNN-nearest neighbor (kNN) algorithm with $k = 5$ (Cover & Hart, 1967). As foundation models we evaluate Uni-Mol (Zhou et al., 2023), ChemBERTa-2 (Ahmad et al., 2022), CLAMP (Seidl et al., 2023), MolBERT (Fabian et al., 2020), MolCLR (Wang et al., 2022), MoLFormer (Ross et al., 2022), MolT5 (Edwards et al., 2022), SMI-TED (Soares et al., 2025), and GROVER (Rong et al., 2020). Since our aim is to compare splitting strategies rather than to maximize absolute predictive performance, we deliberately *do not* perform hyperparameter tuning and use no early stopping: any model-specific tuning budget would confound the comparison we are actually interested in.

For foundation models, we use the pretrained checkpoints released by the authors to generate embeddings for every molecule in the *downstream datasets*, and then train three different heads on top of them. First, a linear probe consisting of a single linear layer, which measures the raw quality of the embedding. Second, a small MLP, to test whether the embeddings require additional capacity to be exploited effectively. Third, an RF using the embedding as input; since we use the same RF on ECFP4 as a baseline, this isolates the effect of the representation from that of the learner.

Metrics and variance. For classification datasets, we report the area under the receiver operating characteristic curve (AUROC), along with its variance across folds. For multiclass datasets, the AUROC is macro-averaged over all classes. For regression datasets, we report the root mean squared

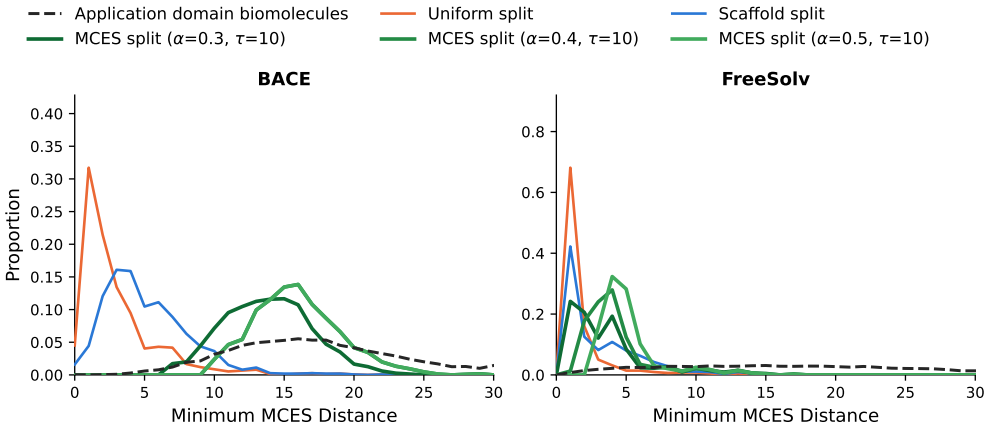

Figure 1: **MCES distance distributions.** Shown are the minimum MCES distances between test and train sets for uniform, scaffold, and MCES splits. In addition, we show minimum MCES distances between dataset and biomolecule reference. Shown are BACE and FreeSolv datasets as two illustrative cases, representing high and low achievable separation. For consistency, we show biomolecules as the application domain for BACE, see also Fig. 6. All distances rounded down; note the varying y-axis. For MCES splits we use $\tau = 10$ and $\alpha \in \{0.3, 0.4, 0.5\}$. Notably, some distances (for FreeSolv, *practical all distances*) are below the threshold $\tau = 10$ even for MCES splits. See Fig. 3 for all datasets.

error (RMSE). As MCES splits are not same-size, we compute the *weighted* mean and *weighted* standard deviation.

6 Results

We noted above that experimental datasets may contain large clusters of 'very small molecules'. In fact, even if we allow a cluster/fold to be as large as half the dataset size, we have to set the MCES distance threshold as low as 3 (FreeSolv dataset), see Table 3 and Fig. 2. We observe that this value is almost perfectly correlated with the median number of heavy (non-hydrogen) atoms of molecules in a dataset. This is in accordance with our claim that 'very small' molecules result in a large cluster with small MCES distances.

6.1 Achieved separation under different splits

As our method uses MCES distances to derive the split, it is not surprising that MCES splits 'outperform' uniform and scaffold splits according to the MCES distance distribution. Yet, MCES distances are a chemically meaningful measure of dissimilarity; hence, this analysis is nevertheless informative. In addition, we will see below that MCES splits indeed result in more difficult instances. We show the distributions of MCES distances achieved by different splitting methods in Fig. 1, and we report statistics of the distributions in Tabs. 4 and 5. See Appendix B for details on the distribution plots. Each split is summarized by the distribution of minimum MCES distances from every test molecule to its nearest training molecule. In addition, we compute the distribution of minimum MCES distances against the application domain of biomolecules. See Figs. 7 and 8 for the between-fold balance achieved by MCES splits for different values of α.

Across all ten datasets, we observe that distance distributions for the uniform split lie furthest from the application domain, those of the scaffold split are 'intermediate', whereas those of the MCES split are closest to the distance distribution of the application domain. This is also reflected when considering median and mean distances, where again, values for MCES splits are closest to the application domain of biomolecules. In absolute terms, even distances for the MCES split sometimes differ substantially from those of the application domain, in particular considering the long tail of the latter distributions. Yet, we argue that evaluations using the MCES split will result in statistics

Table 1: **AUROC values averaged over all seven classification datasets for all models and all splitting methods**. Using $\tau = 10$ and three values of α for our MCES split. The effective distance threshold can be smaller, see Table 3. Each entry is the mean over folds weighted by fold size, averaged over the seven classification datasets; $\pm$ is the corresponding weighted standard deviation over folds, likewise averaged over datasets. For foundation models we report the best-performing of three heads over the five splits (average). See Table 7 for all foundation model heads.

				MCES split		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.809 ±0.019	0.782 ±0.032	0.724 ±0.044	0.721 ±0.038	0.709 ±0.046
RF		0.802 ±0.022	0.776 ±0.026	0.714 ±0.051	0.708 ±0.048	0.698 ±0.058
GCN		0.787 ±0.026	0.767 ±0.029	0.706 ±0.054	0.702 ±0.047	0.685 ±0.063
D-MPNN		0.758 ±0.019	0.744 ±0.031	0.691 ±0.065	0.691 ±0.063	0.664 ±0.079
kNN		0.757 ±0.028	0.730 ±0.030	0.676 ±0.053	0.674 ±0.046	0.664 ±0.047
CLAMP	MLP	0.816 ±0.019	0.799 ±0.032	0.750 ±0.052	0.738 ±0.055	0.733 ±0.059
MolBERT	MLP	0.817 ±0.023	0.792 ±0.027	0.732 ±0.048	0.732 ±0.045	0.724 ±0.058
GROVER	MLP	0.820 ±0.019	0.799 ±0.027	0.739 ±0.053	0.733 ±0.053	0.721 ±0.058
SMI-TED	RF	0.814 ±0.014	0.789 ±0.031	0.734 ±0.040	0.716 ±0.054	0.700 ±0.069
MolT5	MLP	0.780 ±0.017	0.760 ±0.039	0.711 ±0.052	0.700 ±0.054	0.678 ±0.068
MoLFormer	MLP	0.755 ±0.021	0.728 ±0.043	0.678 ±0.052	0.670 ±0.057	0.655 ±0.061
ChemBERTa-2	MLP	0.786 ±0.022	0.765 ±0.037	0.698 ±0.050	0.695 ±0.047	0.685 ±0.060
Uni-Mol	MLP	0.759 ±0.024	0.736 ±0.037	0.673 ±0.061	0.658 ±0.066	0.651 ±0.063
MolCLR	RF	0.761 ±0.016	0.729 ±0.037	0.668 ±0.053	0.645 ±0.058	0.636 ±0.061

much closer to what we will observe when applying the model in practice. This is clearly dependent on the actual dataset: For SIDER, mean and median distance of the MCES split basically match those of the application domain; for HIV, mean and median are substantially smaller even for the MCES split with $\alpha = 0.5$. The remaining gap may be read as a measure of how far benchmark evaluations are from deployment conditions.

Finally, we use the visualization of Kretschmer et al. (2025) to map the folds of each split directly onto the biomolecule application domain (Fig. 4). These plots show not only how well the folds are separated from one another, but also how well each dataset covers the chosen application domain overall, see also Figs. 1 and 3 there.

6.2 Model performance across splits

We report evaluation statistics for the ten datasets in Tabs. 8 to 17, and condense the statistics for the seven classification datasets in Table 1. For each foundation model, we train three different heads (LP, MLP probing, RF) and report the one achieving the highest AUROC averaged over all datasets; the selected head is given in the second column and is used for every dataset and every split. We observe that LP heads usually perform worst (MolBERT on BACE is one exception, see Table 8), whereas MLP and RF take turns in achieving the best performance. Selecting the head on the same results we report gives foundation models a slight advantage over the baselines, which have no comparable degree of freedom, although the effect is limited to a single choice among three options per model. For all models, both baseline and foundation, we observe a clear drop in performance from uniform splits to scaffold splits to MCES splits. Among others, this confirms that MCES distance to the training data is a good proxy for difficulty of a prediction: It is undisputed that out-of-distribution predictions are hard, and it is also undisputed that MCES distances are a good distance measure for chemical dissimilarity. Reported AUROC for MCES splits drops as α increases. This is not necessarily due to higher MCES values, but potentially also to the larger imbalance of splits.

Across the seven classification datasets, foundation model embeddings do not perform substantially better than the classical baselines. Even with the head selected in their favour, only three of the nine foundation models (CLAMP, MolBERT and GROVER) beat XGB on all three MCES splits, and only four outperform RF; the fourth, SMI-TED, exceeds RF by 0.002 at $\alpha = 0.5$. We observe

similar trends for uniform and scaffold splits. Notably, the drop from scaffold splits to MCES splits is roughly twice the difference between foundation and baseline models. The latent representations of foundation models thus appear to offer little to no benefit on these downstream tasks. Next, reported AUROC for MCES splits is smaller than for uniform and scaffold splits, for more than 99 % of (dataset, model) pairs (Appendix F). This is an extremely strong indicator that larger MCES distances correspond to harder instances. We observe similar trends for the regression datasets (Tabs. 15 to 17).

6.3 Structural exposure vs. foundation model performance

Based on our analysis of the achieved separation and the application domain gap, we expected a gradual drop in performance across splitting methods. While Table 1 does show this trend, we also examined each model's performance per dataset separately. For HIV (Table 11), the performance of CLAMP did not change at all between the uniform, scaffold, and MCES splits, staying above 0.9 throughout. This raised the question of whether foundation model pretraining already included the very molecules we were trying to separate from one another. To answer this, we gathered all available pretraining datasets and compared InChIKeys (hashed IUPAC International Chemical Identifiers) against our evaluation datasets, see Table 6. More details can be found in Appendix C.

Every foundation model we evaluated showed structural exposure, ranging from 7.3 % to 90.9 % across all datasets. For CLAMP, we found that 82 % of the molecules in HIV were already present in its pretraining data. In Fig. 9 we tested whether overlap and performance are correlated on the seven classification datasets. While this points to a positive relationship, the evidence is not conclusive, with models such as MoLFormer showing the highest overall overlap, but low performance.

7 Conclusion

We have explored the possibility of performing a reliable evaluation for small molecule machine learning, in the sense that performance statistics have validity for real-world applications. Our approach is empirical, but may give rise to research that integrates these ideas into a theoretical framework. Beyond pure evaluation, our methods allow us to gain practical insights into whether a training dataset is suitable for an arbitrary, yet explicitly stated application domain. In fact, certain experimental datasets are potentially not suited for evaluating and comparing the performance of machine learning models. This is no surprise, as these datasets were usually not measured with the idea of being used as training and evaluation data for machine learning. To measure the dissimilarity of two MCES distributions, we could use metrics such as the Wasserstein metric. In practice, we argue that doing so results in a massive loss of information.

Evaluation results for foundation models are generally underwhelming. We observe some correlation between structural exposure and model performance, underlining that pretraining datasets should be chosen with a certain application domain in mind, too. Our evaluation of CLAMP suffers from data contamination, as we evaluate the model on tasks that were also used during pretraining (Seidl et al., 2023); consequently, our results for CLAMP should be taken with a grain of salt. From the foundation models, MolBERT and GROVER perform consistently better than all baselines, for all splits. Yet, under realistic splits, a considerable gap to practical applicability remains. Instead of scraping more datasets from databases and publication, we suggest a critical reassessment of existing datasets: Is a particular dataset suitable to compare the performance of different models? Notably, experimental datasets for small molecules will remain in a medium data regime, and waiting for more data is not an option. We argue that careful data curation is required, instead of adding even more datasets 'into the mix' not suited for this type of evaluation and comparison.

Reproducibility Statement

To support reproducibility, we release all computed MCES distances, the cleaned datasets, splits, and the code used to generate them at https://anonymous.4open.science/r/mces-split-72D6. Despite the algorithm for merging clusters into folds being deterministic, there is the possibility that two clusters have exactly the same size. In this case the processing order may depend on low-level implementation details. This may result in slightly different folds.

AI use statement

We used a general-purpose AI assistant (Claude, Anthropic) for three purposes. All text was written by the authors and polished with the help of AI, meaning we used it to find better phrasing, spot spelling mistakes, and fix grammar. Additionally, AI was used to polish our plotting scripts in order to change the presentation; all scripts were checked and reviewed by the authors for potential bugs or incorrect statements. Finally, we employed it as a review assistant, submitting our completed draft to identify remaining errors and shortcomings. We then evaluated its suggestions and incorporated those we considered appropriate.

We did not use generative AI tools to generate synthetic data sets; develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in the writing of proofs; propose or refine hypotheses; design or provide feedback on research methodology or experiments; implement methods; assist with translation; clean or reformat the dataset; or interpret results. Support for qualitative and thematic data analysis is not applicable to this quantitative study.

We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of generative AI.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards chemical foundation models, 2022.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *J Med Chem*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

S. H. Bokhari. On the mapping problem. *IEEE Trans Comput*, C-30(3):207–214, 1981. doi: 10.1109/TC.1981.1675756.

Leo Breiman. Random forests. *Mach Learn*, 45(1):5–32, 2001.

Nathan Brown, Marco Fiscato, Marwin H. S. Segler, and Alain C. Vaucher. GuacaMol: Benchmarking models for de novo molecular design. *J Chem Inf Model*, 59(3):1096–1108, 2019. doi: 10.1021/acs.jcim.8b00839.

Roman Bushuiev, Anton Bushuiev, Niek F. de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A. Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S. Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D. Mak, Soha Hassoun, Florian Huber, Justin J.J. van der Hooft, Michael A. Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS 2024)*, volume 37, pp. 110010–110027. Curran Associates, Inc., 2024. doi: 10.52202/079017-3491. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c6c31413d5c53b7d1c343c1498734b0f-Paper-Datasets_and_Benchmarks_Track.pdf.

Darko Butina. Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J Chem Inf Comput Sci*, 39(4):747–750, 1999. doi: 10.1021/ci9803381.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proc. of ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (KDD 2016)*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. doi: 10.1145/2939672.2939785.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi: 10.1109/TIT.1967.1053964.

John S. Delaney. ESOL: Estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci*, 44(3):1000–1005, 2004. doi: 10.1021/ci034243x.

Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. A systematic study of key elements underlying molecular property prediction. *Nat Commun*, 14(1):6395, 2023. doi: 10.1038/s41467-023-41948-6.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pp. 375–413. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.26.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020.

Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The ChEMBL database in 2017. *Nucleic Acids Res*, 45(D1):D945–D954, 2017. doi: 10.1093/nar/gkw1074.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proc. of International Conference on Machine Learning (ICML 2017)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 2017.

Qianrong Guo, Saiveth Hernandez-Hernandez, and Pedro J. Ballester. UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines. *J Cheminf*, 17(1):94, 2025. doi: 10.1186/s13321-025-01039-8.

Vishu Gupta and Michael A. Skinnider. Confronting spurious evaluations of computational methods in small molecule mass spectrometry, 2026.

Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res*, 44(D1):D1214–9, 2016. doi: 10.1093/nar/gkv1031.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A machine learning package for chemical property prediction. *J Chem Inf Model*, 64(1):9–17, 2024. doi: 10.1021/acs.jcim.3c01250.

Ye Hu and Jürgen Bajorath. Scaffold distributions in bioactive molecules, clinical trials compounds, and drugs. *ChemMedChem*, 5(2):187–190, 2010. doi: 10.1002/cmdc.200900419.

Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath. Lessons learned from molecular scaffold analysis. *J Chem Inf Model*, 51(8):1742–1753, 2011. doi: 10.1021/ci200179y.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. In *Proc. of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4c56ff4ce4aaf9573aa5dff913df997a-Paper-round1.pdf.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Artificial intelligence foundation for therapeutic science. *Nat Chem Biol*, 18(10):1033–1036, 2022. ISSN 1552-4469. doi: 10.1038/s41589-022-01131-2.

Ruili Huang, Menghang Xia, Dac-Trung Nguyen, Tongan Zhao, Srilatha Sakamuru, Jinghua Zhao, Sampada A. Shahane, Anna Rossoshek, and Anton Simeonov. Tox21challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. *Front Environ Sci*, 3, 2016. doi: 10.3389/fenvs.2015.00085.

Jaime Huerta-Cepas, François Serra, and Peer Bork. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*, 33(6):1635–1638, 2016. doi: 10.1093/molbev/msw046.

John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model*, 52(7):1757–1768, 2012.

Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: A pre-trained transformer for computational chemistry. *Mach Learn Sci Technol*, 3(1):015022, 2022. doi: 10.1088/2632-2153/ac3ffb.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nat Commun*, 16(1):3337, 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-58606-8.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. doi: 10.1016/j.patter.2023.100804.

Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. PubChem substance and compound databases. *Nucleic Acids Res*, 44:D1202–D1213, 2016. doi: 10.1093/nar/gkv951.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proc. of International Conference on Learning Representations (ICLR 2017)*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Fleming Kretschmer, Eva-Maria Harrieder, Michael Witting, and Sebastian Böcker. Times are changing but order matters: Transferable prediction of small molecule liquid chromatography retention times, 2024.

Fleming Kretschmer, Jan Seipp, Marcus Ludwig, Gunnar W. Klau, and Sebastian Böcker. Coverage bias in small molecule machine learning. *Nat Commun*, 16(1):554, 2025. doi: 10.1038/s41467-024-55462-w.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Res*, 44:D1075–D1079, 2016. doi: 10.1093/nar/gkv1075.

Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *J Chem Inf Model*, 52(6):1686–1697, 2012. doi: 10.1021/ci300124c.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction, 2020.

David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*, 47(D1):D930–D940, 2019. doi: 10.1093/nar/gky1075.

David L. Mobley and J. Peter Guthrie. Freesolv: A database of experimental and calculated hydration free energies, with input files. *J Comput Aided Mol Des*, 28(7):711–720, 2014. ISSN 1573-4951. doi: 10.1007/s10822-014-9747-x.

Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. *J Stat Softw*, 53(9):1–18, 2013. doi: 10.18637/jss.v053.i09.

Paul A. Novick, Oscar F. Ortiz, Jared Poelman, Amir Y. Abdulhay, and Vijay S. Pande. SWEETLEAD: an in silico database of approved drugs, regulated chemicals, and herbal isolates for computer-aided drug discovery. *PLoS One*, 8(11), 2013. doi: 10.1371/journal.pone.0079568.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *J Mach Learn Res*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning, 2025.

Bryn Marie Reimer, Vy Do, and Anna G. Green. Chemical space separation: Scaffold and UMAP splits are closer to random than we think, 2026.

Ann M. Richard, Richard S. Judson, Keith A. Houck, Christopher M. Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihae Yang, James Rathman, Matthew T. Martin, John F. Wambaugh, Thomas B. Knudsen, Jayaram Kancherla, Kamel Mansouri, Grace Patlewicz, Antony J. Williams, Stephen B. Little, Kevin M. Crofton, and Russell S. Thomas. ToxCast chemical landscape: Paving the road to 21st century toxicology. *Chem Res Toxicol*, 29(8):1225–1251, 2016. doi: 10.1021/acs.chemrestox.6b00135.

Ann M. Richard, Ruili Huang, Suramya Waidyanatha, Paul Shinn, Bradley J. Collins, Inthirany Thillainadarajah, Christopher M. Grulke, Antony J. Williams, Ryan R. Lougee, Richard S. Judson, Keith A. Houck, Mahmoud Shobair, Chihae Yang, James F. Rathman, Adam Yasgar, Suzanne C. Fitzpatrick, Anton Simeonov, Russell S. Thomas, Kevin M. Crofton, Richard S. Paules, John R. Bucher, Christopher P. Austin, Robert J. Kavlock, and Raymond R. Tice. The tox21 10k compound library: Collaborative chemistry advancing toxicology. *Chem Res Toxicol*, 34(2):189–216, 2021. doi: 10.1021/acs.chemrestox.0c00264.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J Chem Inf Model*, 50(5): 742–754, 2010. doi: 10.1021/ci100050t.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Proc. of Advances in Neural Information Processing Systems (NeurIPS 2020)*, volume 33, pp. 12559–12571. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/94aef38441efa3380a3bed3faf1f9d5d-Paper.pdf.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nat Mach Intell*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Faizan Sahigara, Kamel Mansouri, Davide Ballabio, Andrea Mauri, Viviana Consonni, and Roberto Todeschini. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, 17(5):4791–4810, 2012. ISSN 1420-3049. doi: 10.3390/molecules17054791.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *Proc. of International Conference on Machine Learning (ICML 2023)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 30458–30490, 2023. URL https://proceedings.mlr.press/v202/seidl23a.html.

Robert P. Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J Chem Inf Model*, 53(4):783–790, 2013. ISSN 1549-9596. doi: 10.1021/ci400084k.

Robert P. Sheridan, Bradley P. Feuston, Vladimir N. Maiorov, and Simon K. Kearsley. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J Chem Inf Comput Sci*, 44(6):1912–1928, 2004. ISSN 0095-2338. doi: 10.1021/ci049782w.

Eduardo Soares, Emilio Vital Brazil, Victor Shirasuna, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. An open-source family of large encoder-decoder foundation models for chemistry. *Commun Chem*, 2025. doi: 10.1038/s42004-025-01585-0.

Teague Sterling and John J. Irwin. ZINC 15 – ligand discovery for everyone. *J Chem Inf Model*, 55 (11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559.

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M. Donghia, Craig R. MacNair, Shawn French, Lindsey A. Carfrae, Zohar Bloom-Ackermann, Victoria M. Tran, Anush Chiappino-Pepe, Ahmed H. Badran, Ian W. Andrews, Emma J. Chory, George M. Church, Eric D. Brown, Tommi S. Jaakkola, Regina Barzilay, and James J. Collins. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. ISSN 0092-8674. doi: 10.1016/j.cell.2020.01.021. URL https://doi.org/10.1016/j.cell.2020.01.021.

Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *J Chem Inf Model*, 56(10):1936–1949, 2016. doi: 10.1021/acs.jcim.6b00290.

Ruoxi Sun, Hanjun Dai, and Adams Wei Yu. Does GNN pretraining help molecular representation? In *Proc. of Advances in Neural Information Processing Systems (NeurIPS 2022)*, volume 35, pp. 12096–12109, 2022. doi: 10.52202/068431-0879. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4ec360efb3f52643ac43fda570ec0118-Paper-Conference.pdf.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nat Mach Intell*, 4(3):279–287, 2022. doi: 10.1038/s42256-022-00447-x.

Mark Wenlock and Nicholas Tomkinson. Experimental in vitro DMPK and physicochemical data on a set of publicly disclosed compounds. https://doi.org/10.6019/CHEMBL3301361, 2015.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: a benchmark for molecular machine learning. *Chem Sci*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *J Chem Inf Model*, 59(8):3370–3388, 2019. doi: 10.1021/acs.jcim.9b00237.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J. Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F. Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A. Patrícia Bento, Melissa F. Adasme, Peter Monecke, Gregory A. Landrum, and Andrew R. Leach. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res*, 52(D1):D1180–D1192, 2024. doi: 10.1093/nar/gkad1004.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A universal 3d molecular representation learning framework. In *Proc. of International Conference on Learning Representations (ICLR 2023)*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

APPENDIX

A QUANTUM CHEMISTRY VS. EXPERIMENTAL DATASETS

Besides datasets where labels were determined experimentally, there exists a second type of small molecule datasets: Quantum chemistry (QC) datasets may contain 100k and more small molecules. Different from experimental datasets, we can determine the QC objective function for an arbitrary small molecule structure. Consequently, we are only limited by the amount of compute power we are willing to invest when generating the dataset. More importantly, we can uniformly sample from the space of small molecules we are interested in, when constructing the dataset. Hence, no involved CV is required to ensure that evaluation results are similar to what is observed in practice. Unfortunately, the resulting machine learning model is not supposed to predict any physical or chemical "truth"; rather, the model simply has to predict what QC computations would give us, anyways. Consequently, the job of the machine learning model is merely to save computation time.

Here, we concentrate on datasets containing experimental data. Some datasets provide direct measurements of emergent chemical or physical properties of the small molecules, such as water solubility or lipophilicity. Most experimental datasets provide measurements for complex biological traits such as "toxicity" or "bioactivity". Yet, a compound may be "not toxic" simply because it has low water solubility, or because the biological uptake of the compound is low. Consequently, trying to predict any such complex biological property using a single machine learning model, is a somewhat ill-defined problem. On the positive side, predicting experimental properties is clearly a worthwhile endeavor that goes far beyond saving computation time: Notably, experimental datasets are often small (usually below 10k molecules) mainly because measuring the experimental data is very expensive. A powerful model allows us to estimate these values "for free". Even more important, experimental measurements are restricted to those molecules that synthetic chemistry can (relatively cheaply) generate. Finally, legal and safety requirements may make it challenging or impossible to measure experimental data for certain molecules: Say, cocaine, or highly toxic compounds. All of these restrictions do not apply when estimating values via a machine learning model.

B MCES DISTRIBUTION PLOTS

Myopic MCES distances, as defined by Kretschmer et al. (2025), can take half-integer values such as 7.5. For our distribution plots, we round down non-integer values. Notably, non-integer values are the exception, not the rule.

To visualize the similarity between the train and test portions of datasets, or alternatively compare to the desired domain of application, we show distributions of minimum MCES distances. We use myopic MCES distances, which require specifying a threshold T above which lower bounds are used as approximations of exact MCES distances. Furthermore, to ensure that the myopic MCES distance is a metric, a constraint is used preventing exact computations of values greater than the threshold. Both thresholding and the constraint lead to an overrepresentation of minimum MCES distances at the threshold, resulting in a very visible 'spike'. Notably, a spike is only clearly visible for *minimum* distances, not for the general distribution of distances, as shown in Fig. 4 from (Kretschmer et al., 2025). The overrepresentation is not confined to T itself, but extends into the bins immediately above it. Consequently, a correction that treats only T as affected redistributes the excess into values that are themselves artifacts, which relocates the spike to $T + 1$ and amplifies it rather than removing it. We therefore treat a contaminated band $[T,\, T + k]$, and use $k = 1$ throughout.

This spike is merely an artifact, and distracts from the distributions themselves. We therefore chose to artificially smooth out the spike using the following steps, artificially modifying the proportion values displayed on the y-axis. The underlying distances are left unchanged, and no reported quantity is derived from the smoothed curves.

Let $p(x)$ denote the displayed proportion at a minimum distance of x. First, we linearly interpolate the proportion values across the band from the neighboring values $p(T-1)$ and $p(T+k+1)$, which we assume to be unaffected, yielding interpolated values $\tilde{p}(x)$ for $x \in [T,\, T + k]$.

We want to distribute the overestimated part of the band, the total difference $X = \sum_{x=T}^{T+k} \big(p(x) - \tilde{p}(x)\big)$, across all values greater than the band. For this, we compute the sum of proportion values

greater than $T+k$, including half of the interpolated band, $S = \sum_{x>T+k} p(x) + \frac{1}{2}\sum_{x=T}^{T+k} \tilde{p}(x)$. We multiply all values greater than $T+k$ with $c = \frac{S+X}{S}$. Finally, we interpolate the band again from $p(T-1)$ and the updated $p(T+k+1)$. For $k=0$, this reduces to redistributing the excess at T alone. Distributions in which the band shows no excess over the interpolation are left unchanged.

C Model descriptions

No model was tuned, and no early stopping was used (Sec. 5). All models handle multi-task datasets by ignoring missing labels.

C.1 Baseline models.

RF (Breiman, 2001) is trained on 2048-bit ECFP4 fingerprints (Rogers & Hahn, 2010) with 300 trees of unrestricted depth, fitting one forest per task.

XGB (Chen & Guestrin, 2016) is trained on the same fingerprints with 300 boosted trees of maximum depth 6, a learning rate of 0.05 and row and column subsampling of 0.8, fitting one model per task.

kNN (Cover & Hart, 1967) predicts from the $k=5$ training molecules closest in Tanimoto distance on the same fingerprints, averaging their labels weighted by distance.

GCN (Kipf & Welling, 2017) learns directly from the molecular graph, using two graph convolutional layers of size 64 followed by a dense layer of size 128 and dropout ($p=0.1$). It is trained for 50 epochs with a batch size of 256 and a learning rate of 10^{-3}.

D-MPNN (Yang et al., 2019) passes messages along directed bonds of the molecular graph; we use Chemprop 2 (Heid et al., 2024) with a depth-3 message-passing network of hidden size 300, normalized aggregation and a single-layer feed-forward predictor of size 300. It is trained for 50 epochs with a batch size of 256 and dropout ($p=0.1$) in the message-passing layers, using a learning rate warmed up over two epochs to 10^{-3} and decayed to 10^{-4}.

C.2 Foundation model heads.

The linear probe is a single linear layer. The MLP has two hidden layers of 256 and 128 units with ReLU (rectified linear unit) activations and dropout ($p=0.1$) after each, followed by a linear output layer. The RF head uses 300 trees of unrestricted depth and fits one forest per task. The linear probe and the MLP are trained for 50 epochs with a batch size of 256, using Adam with a learning rate of 10^{-3}, weight decay of 10^{-4} and cosine learning rate annealing. We minimize binary cross-entropy for classification and mean squared error for regression. Regression targets are standardized using training-fold statistics only.

C.3 Foundation models.

ChemBERTa-2 (Ahmad et al., 2022) has multiple checkpoints, of which we embed ChemBERTa-77M-MTR: it was shown to be the most effective among the variants pretrained on the 77M PubChem corpus (Kim et al., 2016). This checkpoint was pretrained using multi-task regression (MTR), a self-supervised objective where the model predicts molecular properties that can be computed from the SMILES string. The 77M corpus itself was not released; we scan a third-party 78-shard archive of it, whose first ten shards concatenate byte-for-byte into the officially released 10M subset, while the remaining 68 shards (87 % of the molecules) are unverified.

Uni-Mol (Zhou et al., 2023) was pretrained on approximately 19 million unique molecules (10 three-dimensional conformers plus one two-dimensional representation each; approximately 209 million 3D conformations) drawn from 18 commercial vendor catalogs together with molecules collected from ZINC (Irwin et al., 2012) and ChEMBL (Mendez et al., 2019). Objectives use atom types, coordinates, and pairwise distances. InChIKey overlap in Table 6 is matched against the official unique-SMILES dump after subtracting the 100k validation holdout, leaving 19.05M training molecules.

MoLFormer (Ross et al., 2022) was pretrained by masked language modeling on more than 1.1 billion canonical SMILES (approximately 111 million PubChem and >1 billion ZINC); we embed the MoLFormer-c3-1.1B checkpoint. InChIKey overlap against the redistributed 1.1B ZINC+PubChem SMILES file is in Table 6. Of all corpora we examine, this one has the largest overlap with our evaluation data: it contains 90.9 % of the evaluation molecules.

GROVER (Rong et al., 2020) collected approximately 11 million molecules from ZINC15 (Sterling & Irwin, 2015) and ChEMBL (Mendez et al., 2019), held out 10 %, and pretrained on roughly 10 million molecular graphs with contextual atom/bond prediction and 85 functional-group motifs. The exact 11M ZINC15/ChEMBL list and its random pretraining split were not released, and the authors report no selection criterion for either source, so no exact corpus can be reconstructed. We therefore scan both source databases: a 10M ZINC15 sample and that sample united with ChEMBL 26 (March 2020). Those two scans are the ends of the GROVER range in Table 6, 0.3–63.6 %, which differ by more than 200× because the proportion drawn from each source is unreported. On our evaluation sets, the ZINC sample contributes almost nothing beyond ChEMBL: 169 of its 210 hits are also ChEMBL hits, so it adds only 41 molecules. The range therefore mainly shows how much of ChEMBL is in GROVER's 11M. If it took ChEMBL as a whole, its overlap ranks fifth of the nine corpora we examine, between Uni-Mol and ChemBERTa-2.

CLAMP (Seidl et al., 2023) is pretrained on molecule–bioassay pairs from the PubChem18 dump. For MoleculeNet linear probing they drop every pretraining molecule whose InChIKey matches a downstream *test* molecule. The exclusion list is not published, but the authors supplied the underlying compound lists, and we reconstruct and verify the removal. The exclusion is defined on test sets only, about a tenth of each dataset, and the policy does not strip MoleculeNet training or validation molecules, so the pooled overlap only falls from 80.3 % to 70.6 % (Table 6).

MolBERT (Fabian et al., 2020) is pretrained using masked language modeling plus prediction of 200 normalized RDKit descriptors. Pretraining uses the GuacaMol corpus (Brown et al., 2019), 1,591,378 molecules curated from ChEMBL 24 (Gaulton et al., 2017) by salt removal, charge neutralization and element and length filters, split 80/5/15. Only the training part is shown to the model, and it is that published file we scan (1,273,104 SMILES), so MolBERT's column in Table 6 measures optimizer-visible molecules rather than bounding them from above. At 49.8 % it sits mid-field, although it is by far the smallest corpus we scan.

MolCLR (Wang et al., 2022) was pretrained contrastively on the 9,999,918 PubChem molecules at a 95/5 split; the released loader applies an unseeded shuffle, so which 5 % was held out of the released run is not reconstructable, and we therefore scan the complete file.

MolT5 (Edwards et al., 2022) continues T5.1.1 pretraining on an equal mixture of C4 text and 100M ZINC15 SMILES (Sterling & Irwin, 2015) and fine-tunes on ChEBI-20 molecule–description pairs (Hastings et al., 2016). We use the `molt5-large-smiles2caption` checkpoint, which includes this fine-tuning, so we scan both molecule sources. The ZINC stream is not assembled by MolT5: it reuses the corpus of Chemformer (Irwin et al., 2022), which drew approximately 100 million SMILES at random from the approximately 1.5 billion molecules of ZINC15 under reactivity, purchasability, molecular weight $\leq$ 500 Da and LogP $\leq$ 5, then split them randomly 99/0.5/0.5. The training data is public; we scan the 98,999,985-molecule train split. Because Chemformer kept only approximately 7 % of the candidate pool, the ZINC stream alone covers only 2.9 % of our evaluation molecules; this reflects the subsampling rather than any exclusion of our evaluation molecules, as its filters admit 79.5 % of them. Most of MolT5's overlap instead comes from the 26,407 molecules of the ChEBI-20 train split, the part used for fine-tuning: together, the two sources cover 12.6 % of our evaluation molecules. This share is low only because of HIV (3.0 %); for every other dataset except BACE and Lipo it is between 27 % and 38 %.

SMI-TED (Soares et al., 2025) was pretrained on 91M curated PubChem molecules, obtained from 113M PubChem SMILES by deduplication, RDKit canonicalization, a validity check, and a 202-token length cap. The authors did not release that 91M list. We therefore rerun the public training filter on the dump shipped with MoLFormer, which leaves 95,719,882 unique SMILES. The authors' 95/5/0 split leaves 90.9M for encoder training, i.e., the stated 91M. We report InChIKey overlap against the full unique list (Table 6), which therefore bounds encoder-training membership from above: 88.0 % of our evaluation molecules, just below MoLFormer's 90.9 %.

D DATASET DESCRIPTION

For classification, we consider the following datasets:

- The *BACE* (β-site amyloid precursor protein cleaving enzyme) dataset comprises 1,513 human BACE-1 inhibitors collected from scientific literature (Subramanian et al., 2016). Arguably, ChEMBL is a more natural application domain than biomolecules for BACE. Yet, for consistency throughout this paper, we generally show distribution plots for biomolecules.
- The *BBBP* (blood-brain barrier penetration) dataset consists of 2,039 molecules curated from scientific literature discussing blood-brain barrier penetration (Martins et al., 2012).
- The *ClinTox* dataset compares approved drugs from the SWEETLEAD database (Novick et al., 2013) with drugs that failed clinical trials for toxicity. ClinTox contains 1,478 molecular structures with two near-complementary labels. One label measures toxicity ('clinical trial toxicity') and the other non-toxicity ('FDA approved').
- The *HIV* (human immunodeficiency virus) dataset contains data from the Drug Therapeutics Program AIDS (acquired immunodeficiency syndrome) Antiviral Screen, with 41,127 molecular structures. We follow Wu et al. (2018) and use 'confirmed active' and 'confirmed moderately active' as the positive class, and 'confirmed inactive' as the negative class.
- The *SIDER* dataset collects marketed drugs and adverse drug reactions from the Side Effect Resource (Kuhn et al., 2016). It contains 1,427 molecular structures and side effects grouped in 27 system organ classes as labels.
- *ToxCast* is an extended dataset from the Tox21 program that includes toxicology data based on *in vitro* high-throughput screening of 8,576 compounds and 617 binary classification endpoints (Richard et al., 2016).
- The *Tox21* dataset consists of samples from 12 toxicological experiments of *in vitro* bioassays and a total of 7,831 molecular structures (Huang et al., 2016; Richard et al., 2021).

Notably, ToxCast and Tox21 have a very large overlap in molecular structures: 6,362 molecular structures are found in both datasets.

For regression tasks we consider these datasets:

- The *ESOL* (estimated aqueous solubility) dataset contains 1,128 compounds with water solubility data (Delaney, 2004), and is sometimes referred to as '*Delaney*'.
- The *FreeSolv* dataset contains experimental hydration free energies for 642 small, neutral molecules in water. It is a single-task regression dataset (Mobley & Guthrie, 2014).
- The *Lipophilicity* dataset, *Lipo* for short, provides lipophilicity data for 4,200 compounds (Wenlock & Tomkinson, 2015).

As with most evaluations, we leave out datasets PubChem BioAssay (PCBA) and Maximum Unbiased Validation (MUV). These are in no way homogeneous datasets, but were rather scraped from diverse high-throughput screening experiments. They exhibit a particularly high class imbalance (Wu et al., 2018). In principle, these datasets contain 90k and 400k molecular structures; yet, due the class imbalance, the number of available structures for each label is substantially smaller.

E DATASET CLEANING

We standardized molecular structures for all datasets using the PubChem standardization web service (Kim et al., 2016). A small number of structures failed standardization; the corresponding rows in the datasets were removed (Table 2). Second, we removed duplicate rows in the following way: For 'true' duplicates with the same compound structure and same target labels we kept one instance. Rows with different labels for the same structure were removed completely. We removed 20 (ClinTox), 11 (BBBP), 9 (Lipo), 6 (ESOL), 2 (Tox21), 2 (ToxCast), 0 (FreeSolv, BACE, SIDER) conflicts. The number in (Table 2) is a combination of those conflicts and true duplicates. Third, 72 entries across the MoleculeNet datasets are single atoms or collections of atoms without covalent

bonds, such as potassium iodide. These occur in Tox21 (20), SIDER (31), ToxCast (15), HIV (3) and FreeSolv (3), and were excluded. Removed structures account for at most 2.2 % of molecules and 1.5 % of positive labels in any affected dataset; no individual dataset lost more than 5 % of its positives.

As Wu et al. (2018) already noted, several datasets within MoleculeNet are heavily imbalanced, with only SIDER, BACE being roughly balanced label-wise. BBBP, in contrast, has over 70 % positive labels, while ToxCast, Tox21, ClinTox and HIV sit at the lower end, ranging from 8.23 % down to 3.47 %. In addition, the datasets are small: the majority contain fewer than 10k molecules, and only HIV exceeds 40k.

F LARGE MCES DISTANCES CORRESPOND TO HARD INSTANCES

Comparing MCES distance distribution between CV splits and some application domain, is based on the idea that large MCES distances to the training data, indeed correspond to hard instances. This trend is rather obvious from Table 1, but we decided to investigate the question in more detail.

One may want to measure the correlation between MCES distance to the training data, and the achieved performance on different datasets. Yet, we argue that this analysis is somewhat misleading: Each sample in itself holds a certain 'hardness', in the extreme case due to a wrongly labeled sample. Now, grouping together different samples solely based on the distance to the training data, will inevitable result in massive artifacts.

We argue that evaluating on different CV folds is in fact a much better statistic for this question: In every k-fold CV, each sample is considered exactly once as a query for evaluation, and is present in the training data for $k - 1$ out of k queries.

Beyond Table 1, we tested how often any model on any dataset reaches a higher AUROC for the uniform or scaffold split, than for any of the three MCES splits (Tabs. 8 to 14). Clearly, we do not want to consider cases where a classifier does not perform better than random: For example, using the MolT5 foundation model with linear probing on the ClinTox dataset reaches AUROCs between 0.380 and 0.454. There are 208 pairs (dataset, model) where at least one of the reported AUROCs of the five splits is at least 0.6. Of these, only two (dataset BBBP, model MolCLR-LP; dataset HIV, model SMI-TED-LP) have uniform or scaffold split AUROCs higher than any of the MCES split AUROCs. Notably, both cases are linear probing of a foundation model. As noted above, linear separation of a complex biochemical property is highly unlikely.

SUPPLEMENTARY FIGURES AND TABLES

Table 2: **Dataset cleaning**. Column 'standardized' loss (number in brackets) is the number of molecules for which PubChem standardization failed; 'deduplicated' loss is the number of exact duplicate structures and doublets; 'valid' loss is the number of molecules with no edges, e.g., single atoms and inorganic salts.

dataset	published	standardized	deduplicated	valid
BACE	1,513	1,510 (−3)	1,510	1,510
BBBP	2,039	2,037 (−2)	1,956 (−81)	1,956
ClinTox	1,478	1,474 (−4)	1,434 (−40)	1,434
HIV	41,127	40,938 (−189)	40,935 (−3)	40,932 (−3)
SIDER	1,427	1,426 (−1)	1,426	1,395 (−31)
ToxCast	8,576	8,565 (−11)	8,558 (−7)	8,543 (−15)
Tox21	7,831	7,810 (−21)	7,801 (−9)	7,781 (−20)
ESOL	1,128	1,126 (−2)	1,109 (−17)	1,109
FreeSolv	642	640 (−2)	640	637 (−3)
Lipo	4,200	4,194 (−6)	4,176 (−18)	4,176

Table 3: **Dataset statistics.** Column 'heavy' is the median heavy atom count, that is, the number of non-hydrogen atoms of a molecule. Datasets sorted by heavy atoms. We report the maximum MCES distances so that no cluster of the unbounded method is larger than αn, for different values of α. We observe that the first four datasets in the table require extremely small values for τ to guarantee even a large $\alpha = 0.5$. Next, $s(10)/n$ is the relative cluster size of the largest cluster if we enforce $\tau = 10$.

dataset	n	heavy	$\tau_*(0.2)$	$\tau_*(0.3)$	$\tau_*(0.4)$	$\tau_*(0.5)$	$s(10)/n$
FreeSolv	637	8	1	1	1	3	0.962
ESOL	1,109	12	2	3	3	4	0.877
ToxCast	8,543	16	3	3	3	4.5	0.814
Tox21	7,781	16	3	3	3.5	4.5	0.819
HIV	40,932	23	5	6	6	7	0.743
BBBP	1,956	23	6	6	7	8	0.705
ClinTox	1,434	23	6	7	8	9	0.520
SIDER	1,395	25	7	8	9.5	10	0.453
Lipo	4,176	27	7	8	9	10	0.429
BACE	1,510	33	7	7	11	13	0.385

Table 4: **MCES distance statistics of different splits for BACE, BBBP, ClinTox, HIV, and SIDER.** For comparison, we also show these statistics for the application domain of biomolecules. Distances are the minimum MCES distance from each test molecule to its training set. Column '$d = 0$' counts molecules with *identical* molecular graph to a molecule in the training set; such pairs differ only in stereochemistry. Note that the reported mean and median may be lower bounds, since values above $T = 30$ are also lower bounds.

Distribution	Median dist.	Mean dist.	$d = 0$	$d < 3$	$d < 6$
BACE					
Biomolecules	18.5	20.99	0.0 %	0.1 %	1.1 %
Uniform	2.0	3.09	4.5 %	57.7 %	84.6 %
Scaffold	4.25	5.23	1.6 %	18.1 %	60.5 %
Butina	5.0	5.41	1.2 %	24.7 %	59.7 %
UMAP	13.0	12.84	0.0 %	0.3 %	6.2 %
MCES $\alpha = 0.3$	14.0	13.98	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.4$	16.0	16.28	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.5$	16.0	16.28	0.0 %	0.0 %	0.0 %
BBBP					
Biomolecules	14.0	16.53	0.2 %	1.8 %	9.6 %
Uniform	4.0	5.67	1.6 %	34.2 %	59.8 %
Scaffold	6.0	7.39	0.0 %	11.8 %	41.7 %
Butina	5.0	6.56	0.0 %	23.2 %	52.1 %
UMAP	9.0	10.33	0.0 %	6.1 %	24.7 %
MCES $\alpha = 0.3$	10.0	10.82	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.4$	10.0	11.41	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.5$	11.0	11.74	0.0 %	0.0 %	0.0 %
ClinTox					
Biomolecules	14.0	16.13	0.2 %	1.9 %	9.5 %
Uniform	6.0	7.47	6.4 %	25.7 %	49.9 %
Scaffold	8.0	9.58	1.4 %	8.9 %	32.6 %
Butina	7.0	9.09	1.0 %	15.3 %	40.4 %
UMAP	11.0	12.77	0.3 %	3.3 %	16.5 %
MCES $\alpha = 0.3$	11.0	12.75	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.4$	12.0	13.61	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.5$	13.0	14.17	0.0 %	0.0 %	0.0 %
HIV					
Biomolecules	17.0	17.39	0.8 %	3.5 %	18.4 %
Uniform	4.0	5.21	0.8 %	34.6 %	64.5 %
Scaffold	6.0	7.15	0.1 %	9.7 %	42.6 %
Butina	5.0	6.24	0.1 %	23.4 %	54.7 %
UMAP	9.0	10.38	0.0 %	2.5 %	18.3 %
MCES $\alpha = 0.3$	9.0	9.88	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.4$	9.0	10.24	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.5$	10.0	10.91	0.0 %	0.0 %	0.0 %
SIDER					
Biomolecules	14.0	15.69	0.2 %	1.6 %	8.5 %
Uniform	7.0	9.38	1.4 %	20.7 %	42.9 %
Scaffold	9.0	11.29	1.4 %	9.0 %	26.7 %
Butina	8.0	10.83	0.5 %	13.8 %	35.6 %
UMAP	12.0	19.04	0.1 %	2.7 %	13.4 %
MCES $\alpha = 0.3$	13.0	14.87	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.4$	14.0	15.49	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.5$	14.0	15.53	0.0 %	0.0 %	0.0 %

Table 5: **MCES distance statistics of different splits for ToxCast, Tox21, ESOL, FreeSolv, and Lipo.** Column descriptions as in Table 4. Note that the reported mean and median may be lower bounds, since values above $T = 30$ are also lower bounds.

Distribution	Median dist.	Mean dist.	$d = 0$	$d < 3$	$d < 6$
ToxCast					
Biomolecules	11.0	12.93	1.0 %	8.4 %	22.7 %
Uniform	2.0	4.28	8.3 %	54.1 %	73.7 %
Scaffold	5.0	6.05	2.0 %	26.9 %	58.3 %
Butina	3.0	5.04	1.4 %	46.9 %	69.1 %
UMAP	5.0	7.50	0.1 %	25.0 %	50.6 %
MCES $\alpha = 0.3$	6.0	7.42	0.0 %	0.0 %	49.6 %
MCES $\alpha = 0.4$	6.0	7.95	0.0 %	0.0 %	35.3 %
MCES $\alpha = 0.5$	7.0	8.77	0.0 %	0.0 %	10.7 %
Tox21					
Biomolecules	11.0	12.96	1.0 %	8.2 %	22.4 %
Uniform	2.0	4.43	4.1 %	51.9 %	73.2 %
Scaffold	5.0	6.02	1.0 %	26.0 %	58.4 %
Butina	3.0	4.98	1.1 %	45.8 %	68.9 %
UMAP	5.0	7.47	0.0 %	23.8 %	50.4 %
MCES $\alpha = 0.3$	6.0	7.37	0.0 %	0.0 %	49.4 %
MCES $\alpha = 0.4$	6.0	7.88	0.0 %	0.0 %	34.9 %
MCES $\alpha = 0.5$	7.0	8.68	0.0 %	0.0 %	11.1 %
ESOL					
Biomolecules	16.0	18.46	0.2 %	3.0 %	11.4 %
Uniform	1.0	3.13	0.4 %	66.9 %	83.6 %
Scaffold	4.0	4.93	0.0 %	36.4 %	65.7 %
Butina	2.0	3.32	0.0 %	64.6 %	81.2 %
UMAP	4.0	5.66	0.0 %	41.0 %	59.4 %
MCES $\alpha = 0.3$	6.0	7.02	0.0 %	0.0 %	48.3 %
MCES $\alpha = 0.4$	6.0	7.02	0.0 %	0.0 %	48.3 %
MCES $\alpha = 0.5$	7.0	7.90	0.0 %	0.0 %	17.5 %
FreeSolv					
Biomolecules	20.5	26.22	0.1 %	2.2 %	8.7 %
Uniform	1.0	1.84	0.6 %	84.8 %	94.3 %
Scaffold	2.0	3.21	0.0 %	54.8 %	81.8 %
Butina	1.0	1.92	0.0 %	82.7 %	93.9 %
UMAP	2.0	2.86	0.0 %	65.9 %	87.9 %
MCES $\alpha = 0.3$	3.0	3.71	0.0 %	43.0 %	84.6 %
MCES $\alpha = 0.4$	4.0	4.33	0.0 %	18.1 %	84.5 %
MCES $\alpha = 0.5$	5.0	5.18	0.0 %	0.0 %	76.9 %
Lipo					
Biomolecules	14.0	18.69	0.2 %	1.3 %	8.0 %
Uniform	4.0	5.36	3.3 %	36.9 %	62.1 %
Scaffold	7.0	7.52	0.0 %	7.8 %	36.6 %
Butina	7.0	7.42	0.0 %	16.5 %	41.9 %
UMAP	13.0	13.33	0.0 %	0.8 %	7.0 %
MCES $\alpha = 0.3$	12.0	13.06	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.4$	13.0	13.67	0.0 %	0.0 %	0.0 %
MCES $\alpha = 0.5$	13.5	14.01	0.0 %	0.0 %	0.0 %

Table 6: **Structural exposure between pretraining and evaluation data**. Column n is the number of unique keys per dataset. GROVER's column is a range, not an exact measurement. MolT5's column includes the ChEBI-20 train split used for fine-tuning. The last row pools all evaluation molecules showing the per-molecule overlap rate.

	n	ChemBERTa-2	CLAMP	GROVER	MolBERT	MolCLR	MoLFormer	MolT5	SMI-TED	Uni-Mol
BACE	1,510	39.5	20.7	0.0–29.9	25.8	5.0	34.4	0.8	34.4	29.7
BBBP	1,956	22.8	30.8	0.5–83.3	45.4	2.5	69.3	27.5	63.8	84.0
ClinTox	1,434	45.0	21.0	0.6–45.7	17.0	5.5	81.1	36.2	48.0	46.0
HIV	40,932	67.3	82.3	0.2–48.3	49.4	8.7	97.8	3.0	97.6	55.7
SIDER	1,395	32.8	56.6	0.4–71.9	50.1	3.4	82.6	35.7	78.2	72.7
ToxCast	8,543	36.7	62.8	0.4–94.6	49.5	5.1	85.7	27.9	79.2	95.9
Tox21	7,781	42.0	59.7	0.5–91.6	57.2	6.1	87.0	30.5	79.0	92.5
ESOL	1,109	59.2	72.0	0.2–77.3	62.6	9.6	98.8	37.9	98.4	89.6
FreeSolv	637	23.9	67.8	0.0–77.4	53.5	3.6	96.4	37.4	90.6	95.4
Lipo	4,176	39.2	50.3	0.3–98.2	57.7	5.1	75.2	12.1	73.0	98.2
weighted avg.	69,467	55.5	70.6	0.3–63.6	49.8	7.3	90.9	12.6	88.0	68.6

Table 7: **AUROC values averaged over all seven classification datasets for all models and all splitting methods**. Using $\tau = 10$ and three values of α for our MCES split. Each entry is the mean over folds weighted by fold size, averaged over the seven classification datasets; $\pm$ is the corresponding weighted standard deviation over folds, likewise averaged over datasets. Foundation models are shown with each of the three heads separately.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.809 ±0.019	0.782 ±0.032	0.724 ±0.044	0.721 ±0.038	0.709 ±0.046
RF		0.802 ±0.022	0.776 ±0.026	0.714 ±0.051	0.708 ±0.048	0.698 ±0.058
GCN		0.787 ±0.026	0.767 ±0.029	0.706 ±0.054	0.702 ±0.047	0.685 ±0.063
D-MPNN		0.758 ±0.019	0.744 ±0.031	0.691 ±0.065	0.691 ±0.063	0.664 ±0.079
kNN		0.757 ±0.028	0.730 ±0.030	0.676 ±0.053	0.674 ±0.046	0.664 ±0.047
CLAMP	LP	0.759 ±0.018	0.749 ±0.026	0.701 ±0.052	0.692 ±0.049	0.691 ±0.047
	MLP	0.816 ±0.019	0.799 ±0.032	0.750 ±0.052	0.738 ±0.055	0.733 ±0.059
	RF	0.819 ±0.016	0.797 ±0.027	0.733 ±0.057	0.722 ±0.063	0.717 ±0.057
MolBERT	LP	0.796 ±0.019	0.780 ±0.028	0.737 ±0.039	0.731 ±0.038	0.721 ±0.043
	MLP	0.817 ±0.023	0.792 ±0.027	0.732 ±0.048	0.732 ±0.045	0.724 ±0.058
	RF	0.814 ±0.017	0.790 ±0.025	0.735 ±0.038	0.724 ±0.043	0.710 ±0.053
GROVER	LP	0.792 ±0.021	0.771 ±0.026	0.720 ±0.059	0.715 ±0.058	0.707 ±0.056
	MLP	0.820 ±0.019	0.799 ±0.027	0.739 ±0.053	0.733 ±0.053	0.721 ±0.058
	RF	0.788 ±0.016	0.748 ±0.033	0.675 ±0.053	0.669 ±0.058	0.655 ±0.067
SMI-TED	LP	0.686 ±0.026	0.673 ±0.044	0.645 ±0.067	0.634 ±0.065	0.626 ±0.062
	MLP	0.763 ±0.020	0.750 ±0.035	0.716 ±0.053	0.707 ±0.054	0.693 ±0.050
	RF	0.814 ±0.014	0.789 ±0.031	0.734 ±0.040	0.716 ±0.054	0.700 ±0.069
MolT5	LP	0.653 ±0.029	0.639 ±0.044	0.614 ±0.060	0.591 ±0.083	0.579 ±0.089
	MLP	0.780 ±0.017	0.760 ±0.039	0.711 ±0.052	0.700 ±0.054	0.678 ±0.068
	RF	0.791 ±0.015	0.761 ±0.043	0.693 ±0.051	0.676 ±0.066	0.659 ±0.078
MoLFormer	LP	0.613 ±0.028	0.598 ±0.041	0.576 ±0.073	0.565 ±0.072	0.564 ±0.077
	MLP	0.755 ±0.021	0.728 ±0.043	0.678 ±0.052	0.670 ±0.057	0.655 ±0.061
	RF	0.746 ±0.021	0.706 ±0.046	0.640 ±0.065	0.617 ±0.077	0.603 ±0.087
ChemBERTa-2	LP	0.682 ±0.023	0.669 ±0.037	0.637 ±0.065	0.621 ±0.062	0.617 ±0.065
	MLP	0.786 ±0.022	0.765 ±0.037	0.698 ±0.050	0.695 ±0.047	0.685 ±0.060
	RF	0.777 ±0.026	0.745 ±0.039	0.670 ±0.063	0.655 ±0.071	0.641 ±0.078
Uni-Mol	LP	0.734 ±0.021	0.716 ±0.038	0.660 ±0.056	0.649 ±0.054	0.641 ±0.054
	MLP	0.759 ±0.024	0.736 ±0.037	0.673 ±0.061	0.658 ±0.066	0.651 ±0.063
	RF	0.735 ±0.019	0.709 ±0.032	0.647 ±0.058	0.638 ±0.061	0.628 ±0.062
MolCLR	LP	0.659 ±0.029	0.647 ±0.051	0.614 ±0.060	0.607 ±0.058	0.597 ±0.059
	MLP	0.694 ±0.028	0.679 ±0.042	0.643 ±0.060	0.637 ±0.057	0.625 ±0.061
	RF	0.761 ±0.016	0.729 ±0.037	0.668 ±0.053	0.645 ±0.058	0.636 ±0.061

Table 8: **AUROC on BACE for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3. Note that the resulting splits of 0.4 and 0.5 are identical and therefore also the performance.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.899 ±0.014	0.879 ±0.018	0.736 ±0.020	0.727 ±0.025	0.727 ±0.025
RF		0.894 ±0.018	0.887 ±0.017	0.767 ±0.024	0.756 ±0.034	0.756 ±0.034
GCN		0.813 ±0.031	0.811 ±0.043	0.629 ±0.106	0.615 ±0.098	0.615 ±0.098
D-MPNN		0.799 ±0.017	0.781 ±0.020	0.620 ±0.147	0.590 ±0.142	0.590 ±0.142
kNN		0.872 ±0.023	0.862 ±0.025	0.705 ±0.032	0.708 ±0.046	0.708 ±0.046
CLAMP	LP	0.763 ±0.008	0.715 ±0.053	0.577 ±0.102	0.552 ±0.080	0.552 ±0.080
	MLP	0.867 ±0.014	0.848 ±0.028	0.664 ±0.123	0.585 ±0.129	0.585 ±0.129
	RF	0.882 ±0.015	0.860 ±0.019	0.642 ±0.062	0.568 ±0.118	0.568 ±0.118
MolBERT	LP	0.872 ±0.013	0.866 ±0.019	0.791 ±0.041	0.787 ±0.034	0.787 ±0.034
	MLP	0.889 ±0.018	0.874 ±0.015	0.709 ±0.084	0.756 ±0.090	0.756 ±0.090
	RF	0.896 ±0.022	0.885 ±0.022	0.786 ±0.031	0.776 ±0.054	0.776 ±0.054
GROVER	LP	0.825 ±0.016	0.801 ±0.028	0.637 ±0.112	0.620 ±0.117	0.620 ±0.117
	MLP	0.861 ±0.016	0.841 ±0.021	0.646 ±0.100	0.623 ±0.111	0.623 ±0.111
	RF	0.878 ±0.018	0.862 ±0.004	0.638 ±0.085	0.614 ±0.097	0.614 ±0.097
SMI-TED	LP	0.732 ±0.032	0.727 ±0.071	0.692 ±0.089	0.683 ±0.037	0.683 ±0.037
	MLP	0.810 ±0.023	0.785 ±0.048	0.702 ±0.053	0.670 ±0.053	0.670 ±0.053
	RF	0.873 ±0.004	0.865 ±0.017	0.761 ±0.031	0.717 ±0.061	0.717 ±0.061
MolT5	LP	0.712 ±0.045	0.669 ±0.085	0.599 ±0.110	0.531 ±0.144	0.531 ±0.144
	MLP	0.844 ±0.005	0.811 ±0.047	0.686 ±0.038	0.665 ±0.064	0.665 ±0.064
	RF	0.882 ±0.012	0.851 ±0.033	0.694 ±0.013	0.679 ±0.062	0.679 ±0.062
MoLFormer	LP	0.715 ±0.033	0.689 ±0.088	0.586 ±0.123	0.562 ±0.088	0.562 ±0.088
	MLP	0.818 ±0.018	0.774 ±0.049	0.643 ±0.082	0.612 ±0.074	0.612 ±0.074
	RF	0.836 ±0.020	0.792 ±0.047	0.621 ±0.109	0.545 ±0.119	0.545 ±0.119
ChemBERTa-2	LP	0.783 ±0.027	0.757 ±0.054	0.683 ±0.078	0.660 ±0.095	0.660 ±0.095
	MLP	0.874 ±0.017	0.848 ±0.014	0.688 ±0.063	0.676 ±0.109	0.676 ±0.109
	RF	0.871 ±0.022	0.852 ±0.030	0.733 ±0.021	0.695 ±0.064	0.695 ±0.064
Uni-Mol	LP	0.767 ±0.035	0.734 ±0.035	0.635 ±0.062	0.614 ±0.057	0.614 ±0.057
	MLP	0.826 ±0.008	0.791 ±0.017	0.645 ±0.082	0.638 ±0.080	0.638 ±0.080
	RF	0.802 ±0.024	0.775 ±0.027	0.654 ±0.056	0.635 ±0.067	0.635 ±0.067
MolCLR	LP	0.683 ±0.058	0.661 ±0.092	0.550 ±0.150	0.538 ±0.098	0.538 ±0.098
	MLP	0.709 ±0.060	0.689 ±0.101	0.572 ±0.133	0.540 ±0.113	0.540 ±0.113
	RF	0.858 ±0.020	0.833 ±0.022	0.677 ±0.064	0.529 ±0.119	0.529 ±0.119

Table 9: **AUROC on BBBP for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\boldsymbol{\alpha = 0.3}$	$\boldsymbol{\alpha = 0.4}$	$\boldsymbol{\alpha = 0.5}$
XGB		0.922 ±0.008	0.884 ±0.030	0.802 ±0.075	0.810 ±0.065	0.790 ±0.068
RF		0.924 ±0.009	0.892 ±0.031	0.797 ±0.105	0.795 ±0.085	0.771 ±0.104
GCN		0.890 ±0.026	0.875 ±0.019	0.790 ±0.087	0.770 ±0.079	0.786 ±0.082
D-MPNN		0.850 ±0.034	0.827 ±0.046	0.802 ±0.062	0.780 ±0.077	0.775 ±0.088
kNN		0.904 ±0.015	0.887 ±0.029	0.789 ±0.088	0.809 ±0.069	0.793 ±0.074
CLAMP	LP	0.849 ±0.028	0.843 ±0.048	0.778 ±0.096	0.757 ±0.112	0.746 ±0.118
	MLP	0.912 ±0.019	0.898 ±0.038	0.843 ±0.088	0.819 ±0.092	0.816 ±0.090
	RF	0.925 ±0.018	0.899 ±0.029	0.848 ±0.088	0.842 ±0.063	0.833 ±0.064
MolBERT	LP	0.905 ±0.016	0.880 ±0.025	0.816 ±0.097	0.791 ±0.092	0.786 ±0.098
	MLP	0.932 ±0.021	0.901 ±0.023	0.832 ±0.064	0.800 ±0.065	0.801 ±0.085
	RF	0.928 ±0.014	0.911 ±0.023	0.823 ±0.096	0.798 ±0.104	0.802 ±0.110
GROVER	LP	0.895 ±0.015	0.870 ±0.033	0.832 ±0.101	0.816 ±0.072	0.800 ±0.077
	MLP	0.920 ±0.012	0.900 ±0.030	0.842 ±0.061	0.839 ±0.039	0.826 ±0.063
	RF	0.910 ±0.029	0.873 ±0.045	0.816 ±0.069	0.787 ±0.081	0.783 ±0.092
SMI-TED	LP	0.813 ±0.041	0.811 ±0.058	0.738 ±0.130	0.715 ±0.149	0.714 ±0.141
	MLP	0.849 ±0.021	0.838 ±0.058	0.793 ±0.115	0.770 ±0.121	0.760 ±0.128
	RF	0.921 ±0.019	0.897 ±0.034	0.818 ±0.090	0.788 ±0.107	0.776 ±0.123
MolT5	LP	0.810 ±0.042	0.811 ±0.051	0.770 ±0.110	0.745 ±0.136	0.744 ±0.121
	MLP	0.887 ±0.007	0.860 ±0.045	0.795 ±0.106	0.748 ±0.140	0.745 ±0.144
	RF	0.910 ±0.017	0.885 ±0.040	0.802 ±0.112	0.764 ±0.136	0.758 ±0.139
MoLFormer	LP	0.772 ±0.041	0.769 ±0.040	0.729 ±0.110	0.714 ±0.120	0.714 ±0.113
	MLP	0.884 ±0.014	0.860 ±0.041	0.802 ±0.093	0.774 ±0.097	0.771 ±0.108
	RF	0.878 ±0.023	0.858 ±0.037	0.772 ±0.114	0.755 ±0.116	0.760 ±0.109
ChemBERTa-2	LP	0.837 ±0.033	0.823 ±0.036	0.794 ±0.095	0.762 ±0.095	0.757 ±0.090
	MLP	0.910 ±0.009	0.884 ±0.031	0.816 ±0.060	0.794 ±0.065	0.788 ±0.081
	RF	0.916 ±0.021	0.881 ±0.035	0.765 ±0.118	0.741 ±0.129	0.736 ±0.128
Uni-Mol	LP	0.853 ±0.023	0.848 ±0.036	0.789 ±0.087	0.758 ±0.099	0.757 ±0.102
	MLP	0.876 ±0.011	0.862 ±0.030	0.788 ±0.072	0.754 ±0.103	0.732 ±0.105
	RF	0.875 ±0.025	0.856 ±0.030	0.772 ±0.091	0.761 ±0.093	0.754 ±0.102
MolCLR	LP	0.760 ±0.046	0.757 ±0.054	0.771 ±0.108	0.762 ±0.085	0.757 ±0.070
	MLP	0.792 ±0.035	0.782 ±0.044	0.780 ±0.091	0.775 ±0.083	0.758 ±0.093
	RF	0.893 ±0.019	0.867 ±0.028	0.809 ±0.065	0.790 ±0.065	0.786 ±0.073

Table 10: **AUROC on ClinTox for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.855 ±0.069	0.852 ±0.057	0.792 ±0.075	0.808 ±0.048	0.766 ±0.078
RF		0.794 ±0.075	0.783 ±0.043	0.698 ±0.095	0.708 ±0.064	0.674 ±0.115
GCN		0.832 ±0.085	0.834 ±0.033	0.782 ±0.066	0.830 ±0.041	0.755 ±0.104
D-MPNN		0.810 ±0.032	0.806 ±0.061	0.748 ±0.097	0.794 ±0.100	0.715 ±0.133
kNN		0.707 ±0.113	0.698 ±0.056	0.680 ±0.109	0.667 ±0.054	0.633 ±0.051
CLAMP	LP	0.653 ±0.054	0.655 ±0.032	0.607 ±0.083	0.589 ±0.068	0.625 ±0.034
	MLP	0.798 ±0.074	0.746 ±0.091	0.734 ±0.084	0.732 ±0.088	0.718 ±0.090
	RF	0.756 ±0.060	0.720 ±0.074	0.636 ±0.183	0.648 ±0.165	0.645 ±0.122
MolBERT	LP	0.818 ±0.054	0.807 ±0.063	0.782 ±0.056	0.788 ±0.024	0.752 ±0.070
	MLP	0.837 ±0.060	0.826 ±0.064	0.782 ±0.078	0.791 ±0.041	0.758 ±0.094
	RF	0.838 ±0.038	0.815 ±0.059	0.780 ±0.044	0.765 ±0.034	0.720 ±0.072
GROVER	LP	0.856 ±0.079	0.845 ±0.031	0.821 ±0.078	0.835 ±0.082	0.818 ±0.071
	MLP	0.892 ±0.062	0.888 ±0.026	0.866 ±0.062	0.873 ±0.057	0.837 ±0.051
	RF	0.774 ±0.030	0.729 ±0.046	0.663 ±0.068	0.673 ±0.075	0.621 ±0.103
SMI-TED	LP	0.647 ±0.053	0.628 ±0.044	0.625 ±0.139	0.614 ±0.124	0.588 ±0.089
	MLP	0.895 ±0.051	0.883 ±0.047	0.875 ±0.081	0.882 ±0.063	0.857 ±0.051
	RF	0.904 ±0.039	0.896 ±0.045	0.857 ±0.051	0.846 ±0.049	0.813 ±0.099
MolT5	LP	0.485 ±0.062	0.487 ±0.060	0.482 ±0.093	0.488 ±0.077	0.497 ±0.080
	MLP	0.820 ±0.061	0.801 ±0.058	0.761 ±0.084	0.775 ±0.048	0.714 ±0.097
	RF	0.801 ±0.047	0.781 ±0.068	0.721 ±0.070	0.712 ±0.066	0.690 ±0.077
MoLFormer	LP	0.380 ±0.060	0.380 ±0.030	0.408 ±0.147	0.421 ±0.098	0.454 ±0.106
	MLP	0.740 ±0.078	0.691 ±0.073	0.648 ±0.069	0.659 ±0.117	0.619 ±0.095
	RF	0.678 ±0.057	0.619 ±0.066	0.562 ±0.057	0.531 ±0.081	0.487 ±0.119
ChemBERTa-2	LP	0.436 ±0.050	0.443 ±0.051	0.441 ±0.148	0.435 ±0.114	0.448 ±0.098
	MLP	0.726 ±0.082	0.734 ±0.074	0.645 ±0.065	0.656 ±0.045	0.630 ±0.090
	RF	0.670 ±0.090	0.647 ±0.066	0.550 ±0.179	0.540 ±0.172	0.475 ±0.177
Uni-Mol	LP	0.692 ±0.042	0.685 ±0.054	0.593 ±0.120	0.583 ±0.094	0.553 ±0.078
	MLP	0.693 ±0.089	0.681 ±0.084	0.608 ±0.150	0.579 ±0.120	0.551 ±0.104
	RF	0.694 ±0.028	0.648 ±0.064	0.594 ±0.120	0.589 ±0.118	0.578 ±0.086
MolCLR	LP	0.591 ±0.047	0.593 ±0.098	0.560 ±0.035	0.559 ±0.072	0.543 ±0.076
	MLP	0.699 ±0.041	0.692 ±0.042	0.654 ±0.042	0.663 ±0.035	0.644 ±0.060
	RF	0.716 ±0.040	0.696 ±0.059	0.627 ±0.100	0.645 ±0.073	0.629 ±0.070

Table 11: **AUROC on HIV for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.828 ±0.016	0.798 ±0.021	0.770 ±0.067	0.747 ±0.071	0.746 ±0.085
RF		0.833 ±0.023	0.801 ±0.012	0.772 ±0.070	0.755 ±0.076	0.756 ±0.084
GCN		0.824 ±0.018	0.786 ±0.023	0.750 ±0.060	0.734 ±0.072	0.712 ±0.095
D-MPNN		0.808 ±0.014	0.784 ±0.031	0.742 ±0.066	0.736 ±0.072	0.735 ±0.069
kNN		0.802 ±0.014	0.760 ±0.024	0.735 ±0.075	0.719 ±0.076	0.717 ±0.082
CLAMP	LP	0.916 ±0.007	0.915 ±0.011	0.915 ±0.009	0.911 ±0.017	0.907 ±0.023
	MLP	0.932 ±0.006	0.932 ±0.015	0.928 ±0.010	0.929 ±0.014	0.926 ±0.014
	RF	0.900 ±0.005	0.890 ±0.018	0.891 ±0.010	0.885 ±0.017	0.881 ±0.023
MolBERT	LP	0.792 ±0.020	0.775 ±0.019	0.749 ±0.043	0.733 ±0.058	0.731 ±0.059
	MLP	0.814 ±0.017	0.778 ±0.024	0.732 ±0.065	0.714 ±0.059	0.702 ±0.076
	RF	0.815 ±0.013	0.773 ±0.013	0.734 ±0.056	0.712 ±0.074	0.694 ±0.094
GROVER	LP	0.795 ±0.016	0.774 ±0.016	0.755 ±0.051	0.738 ±0.067	0.733 ±0.072
	MLP	0.810 ±0.017	0.782 ±0.034	0.740 ±0.083	0.713 ±0.095	0.702 ±0.108
	RF	0.804 ±0.012	0.752 ±0.016	0.691 ±0.063	0.678 ±0.079	0.658 ±0.093
SMI-TED	LP	0.687 ±0.030	0.678 ±0.012	0.682 ±0.050	0.669 ±0.081	0.657 ±0.101
	MLP	0.769 ±0.012	0.752 ±0.024	0.750 ±0.042	0.735 ±0.058	0.730 ±0.064
	RF	0.799 ±0.006	0.754 ±0.017	0.714 ±0.053	0.706 ±0.076	0.685 ±0.098
MolT5	LP	0.680 ±0.025	0.667 ±0.021	0.667 ±0.053	0.653 ±0.079	0.634 ±0.106
	MLP	0.794 ±0.019	0.768 ±0.018	0.744 ±0.071	0.730 ±0.076	0.708 ±0.099
	RF	0.797 ±0.006	0.751 ±0.023	0.688 ±0.079	0.670 ±0.089	0.638 ±0.118
MoLFormer	LP	0.676 ±0.021	0.659 ±0.039	0.659 ±0.056	0.642 ±0.079	0.626 ±0.102
	MLP	0.786 ±0.016	0.758 ±0.027	0.732 ±0.058	0.724 ±0.060	0.703 ±0.088
	RF	0.746 ±0.015	0.700 ±0.024	0.656 ±0.075	0.647 ±0.082	0.628 ±0.111
ChemBERTa-2	LP	0.743 ±0.018	0.727 ±0.022	0.713 ±0.068	0.694 ±0.077	0.687 ±0.086
	MLP	0.795 ±0.022	0.758 ±0.021	0.702 ±0.078	0.707 ±0.061	0.691 ±0.068
	RF	0.799 ±0.015	0.753 ±0.015	0.696 ±0.077	0.678 ±0.085	0.665 ±0.119
Uni-Mol	LP	0.740 ±0.020	0.715 ±0.030	0.696 ±0.068	0.684 ±0.083	0.676 ±0.087
	MLP	0.764 ±0.024	0.720 ±0.030	0.694 ±0.071	0.663 ±0.103	0.670 ±0.093
	RF	0.705 ±0.029	0.682 ±0.005	0.642 ±0.078	0.631 ±0.091	0.613 ±0.111
MolCLR	LP	0.646 ±0.009	0.623 ±0.022	0.613 ±0.074	0.599 ±0.089	0.580 ±0.109
	MLP	0.709 ±0.016	0.675 ±0.019	0.661 ±0.076	0.640 ±0.084	0.629 ±0.099
	RF	0.778 ±0.012	0.729 ±0.026	0.686 ±0.069	0.679 ±0.073	0.663 ±0.090

Table 12: **AUROC on SIDER for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.663 ±0.010	0.633 ±0.023	0.594 ±0.028	0.588 ±0.033	0.590 ±0.034
RF		0.662 ±0.020	0.634 ±0.021	0.599 ±0.026	0.589 ±0.041	0.595 ±0.032
GCN		0.603 ±0.010	0.580 ±0.025	0.561 ±0.014	0.548 ±0.010	0.549 ±0.007
D-MPNN		0.560 ±0.017	0.560 ±0.011	0.551 ±0.032	0.556 ±0.016	0.544 ±0.017
kNN		0.658 ±0.017	0.629 ±0.018	0.602 ±0.031	0.592 ±0.044	0.591 ±0.030
CLAMP	LP	0.586 ±0.013	0.581 ±0.016	0.555 ±0.034	0.558 ±0.028	0.557 ±0.033
	MLP	0.606 ±0.014	0.592 ±0.017	0.562 ±0.027	0.570 ±0.026	0.568 ±0.034
	RF	0.664 ±0.009	0.644 ±0.018	0.605 ±0.030	0.602 ±0.038	0.606 ±0.032
MolBERT	LP	0.631 ±0.018	0.612 ±0.026	0.573 ±0.016	0.572 ±0.021	0.575 ±0.008
	MLP	0.649 ±0.030	0.619 ±0.029	0.576 ±0.006	0.575 ±0.013	0.578 ±0.018
	RF	0.651 ±0.018	0.627 ±0.015	0.582 ±0.012	0.582 ±0.010	0.580 ±0.013
GROVER	LP	0.631 ±0.013	0.608 ±0.022	0.560 ±0.029	0.568 ±0.029	0.573 ±0.022
	MLP	0.646 ±0.014	0.618 ±0.026	0.578 ±0.027	0.584 ±0.037	0.582 ±0.033
	RF	0.663 ±0.010	0.636 ±0.025	0.594 ±0.027	0.598 ±0.026	0.600 ±0.024
SMI-TED	LP	0.576 ±0.012	0.562 ±0.037	0.537 ±0.018	0.547 ±0.015	0.551 ±0.011
	MLP	0.575 ±0.013	0.568 ±0.023	0.544 ±0.019	0.551 ±0.027	0.555 ±0.010
	RF	0.654 ±0.016	0.631 ±0.023	0.588 ±0.016	0.589 ±0.016	0.591 ±0.015
MolT5	LP	0.583 ±0.014	0.575 ±0.025	0.560 ±0.008	0.564 ±0.020	0.556 ±0.019
	MLP	0.600 ±0.008	0.595 ±0.035	0.572 ±0.031	0.569 ±0.019	0.557 ±0.023
	RF	0.644 ±0.014	0.626 ±0.029	0.586 ±0.024	0.584 ±0.019	0.580 ±0.027
MoLFormer	LP	0.580 ±0.016	0.572 ±0.031	0.558 ±0.025	0.558 ±0.021	0.557 ±0.016
	MLP	0.589 ±0.008	0.584 ±0.026	0.564 ±0.025	0.570 ±0.018	0.558 ±0.003
	RF	0.636 ±0.015	0.615 ±0.028	0.580 ±0.019	0.577 ±0.016	0.578 ±0.015
ChemBERTa-2	LP	0.599 ±0.018	0.587 ±0.022	0.563 ±0.023	0.564 ±0.009	0.569 ±0.019
	MLP	0.643 ±0.008	0.629 ±0.037	0.593 ±0.040	0.591 ±0.020	0.591 ±0.027
	RF	0.645 ±0.016	0.618 ±0.026	0.573 ±0.015	0.574 ±0.004	0.578 ±0.013
Uni-Mol	LP	0.606 ±0.012	0.596 ±0.019	0.552 ±0.020	0.555 ±0.026	0.559 ±0.016
	MLP	0.615 ±0.017	0.607 ±0.040	0.552 ±0.025	0.562 ±0.027	0.566 ±0.024
	RF	0.616 ±0.013	0.608 ±0.015	0.560 ±0.029	0.561 ±0.033	0.559 ±0.024
MolCLR	LP	0.583 ±0.019	0.577 ±0.028	0.554 ±0.013	0.557 ±0.024	0.550 ±0.025
	MLP	0.562 ±0.019	0.563 ±0.018	0.547 ±0.026	0.555 ±0.032	0.544 ±0.020
	RF	0.640 ±0.008	0.617 ±0.024	0.584 ±0.024	0.582 ±0.027	0.580 ±0.032

Table 13: **AUROC on Tox21 for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.816 ±0.008	0.779 ±0.046	0.741 ±0.033	0.741 ±0.015	0.725 ±0.018
RF		0.815 ±0.003	0.777 ±0.034	0.738 ±0.027	0.730 ±0.026	0.714 ±0.026
GCN		0.828 ±0.010	0.793 ±0.038	0.765 ±0.026	0.754 ±0.016	0.726 ±0.031
D-MPNN		0.786 ±0.007	0.766 ±0.033	0.720 ±0.025	0.721 ±0.018	0.672 ±0.055
kNN		0.749 ±0.004	0.693 ±0.046	0.657 ±0.033	0.661 ±0.027	0.644 ±0.029
CLAMP	LP	0.827 ±0.009	0.815 ±0.016	0.783 ±0.024	0.779 ±0.022	0.768 ±0.023
	MLP	0.861 ±0.006	0.849 ±0.022	0.810 ±0.016	0.813 ±0.010	0.806 ±0.014
	RF	0.861 ±0.001	0.839 ±0.028	0.802 ±0.015	0.803 ±0.019	0.791 ±0.018
MolBERT	LP	0.832 ±0.004	0.811 ±0.033	0.773 ±0.004	0.770 ±0.015	0.750 ±0.011
	MLP	0.839 ±0.010	0.805 ±0.024	0.777 ±0.016	0.778 ±0.016	0.763 ±0.015
	RF	0.840 ±0.003	0.813 ±0.031	0.764 ±0.012	0.762 ±0.008	0.738 ±0.008
GROVER	LP	0.829 ±0.004	0.807 ±0.031	0.771 ±0.023	0.768 ±0.016	0.752 ±0.015
	MLP	0.855 ±0.007	0.825 ±0.030	0.798 ±0.020	0.798 ±0.009	0.783 ±0.015
	RF	0.803 ±0.009	0.745 ±0.065	0.701 ±0.048	0.712 ±0.027	0.697 ±0.034
SMI-TED	LP	0.715 ±0.011	0.689 ±0.067	0.653 ±0.017	0.635 ±0.010	0.622 ±0.015
	MLP	0.765 ±0.014	0.751 ±0.037	0.716 ±0.029	0.707 ±0.042	0.660 ±0.016
	RF	0.822 ±0.004	0.781 ±0.056	0.742 ±0.017	0.717 ±0.041	0.685 ±0.052
MolT5	LP	0.716 ±0.010	0.693 ±0.054	0.658 ±0.024	0.608 ±0.074	0.563 ±0.094
	MLP	0.815 ±0.015	0.793 ±0.044	0.759 ±0.011	0.753 ±0.010	0.716 ±0.022
	RF	0.805 ±0.005	0.760 ±0.077	0.718 ±0.030	0.694 ±0.053	0.655 ±0.074
MoLFormer	LP	0.640 ±0.017	0.609 ±0.050	0.569 ±0.036	0.546 ±0.064	0.523 ±0.075
	MLP	0.784 ±0.009	0.762 ±0.056	0.711 ±0.008	0.714 ±0.008	0.694 ±0.023
	RF	0.771 ±0.006	0.720 ±0.079	0.676 ±0.050	0.659 ±0.071	0.627 ±0.085
ChemBERTa-2	LP	0.752 ±0.008	0.726 ±0.059	0.675 ±0.025	0.654 ±0.032	0.625 ±0.049
	MLP	0.827 ±0.011	0.794 ±0.064	0.762 ±0.027	0.762 ±0.010	0.748 ±0.018
	RF	0.814 ±0.009	0.768 ±0.069	0.727 ±0.010	0.712 ±0.022	0.700 ±0.022
Uni-Mol	LP	0.782 ±0.010	0.755 ±0.062	0.709 ±0.018	0.705 ±0.007	0.692 ±0.016
	MLP	0.814 ±0.011	0.785 ±0.046	0.747 ±0.010	0.740 ±0.009	0.730 ±0.010
	RF	0.770 ±0.008	0.738 ±0.055	0.680 ±0.011	0.667 ±0.014	0.651 ±0.024
MolCLR	LP	0.731 ±0.016	0.711 ±0.044	0.669 ±0.018	0.652 ±0.018	0.633 ±0.022
	MLP	0.748 ±0.012	0.725 ±0.056	0.685 ±0.022	0.681 ±0.026	0.676 ±0.028
	RF	0.779 ±0.009	0.732 ±0.062	0.693 ±0.032	0.690 ±0.024	0.676 ±0.022

Table 14: **AUROC on ToxCast for the five splitting strategies**. Higher is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$
XGB		0.680 ±0.008	0.652 ±0.028	0.632 ±0.008	0.626 ±0.009	0.621 ±0.011
RF		0.690 ±0.008	0.658 ±0.022	0.628 ±0.011	0.621 ±0.011	0.621 ±0.010
GCN		0.718 ±0.002	0.688 ±0.022	0.666 ±0.016	0.660 ±0.017	0.653 ±0.021
D-MPNN		0.693 ±0.011	0.682 ±0.015	0.656 ±0.027	0.658 ±0.017	0.618 ±0.048
kNN		0.610 ±0.008	0.581 ±0.015	0.567 ±0.005	0.560 ±0.010	0.562 ±0.014
CLAMP	LP	0.718 ±0.005	0.714 ±0.005	0.691 ±0.017	0.695 ±0.017	0.684 ±0.021
	MLP	0.732 ±0.003	0.729 ±0.011	0.708 ±0.015	0.716 ±0.028	0.714 ±0.037
	RF	0.744 ±0.006	0.727 ±0.004	0.706 ±0.012	0.706 ±0.021	0.698 ±0.023
MolBERT	LP	0.722 ±0.007	0.706 ±0.013	0.677 ±0.016	0.674 ±0.019	0.667 ±0.021
	MLP	0.759 ±0.006	0.739 ±0.008	0.713 ±0.025	0.714 ±0.030	0.707 ±0.029
	RF	0.733 ±0.007	0.710 ±0.009	0.678 ±0.015	0.671 ±0.014	0.660 ±0.018
GROVER	LP	0.713 ±0.007	0.693 ±0.024	0.663 ±0.018	0.659 ±0.022	0.653 ±0.017
	MLP	0.755 ±0.009	0.734 ±0.018	0.703 ±0.020	0.699 ±0.024	0.693 ±0.024
	RF	0.684 ±0.006	0.640 ±0.032	0.624 ±0.011	0.621 ±0.024	0.614 ±0.023
SMI-TED	LP	0.630 ±0.008	0.617 ±0.019	0.587 ±0.029	0.576 ±0.038	0.565 ±0.037
	MLP	0.676 ±0.009	0.672 ±0.006	0.633 ±0.031	0.636 ±0.017	0.622 ±0.028
	RF	0.723 ±0.008	0.698 ±0.027	0.659 ±0.024	0.652 ±0.029	0.634 ±0.035
MolT5	LP	0.586 ±0.006	0.571 ±0.015	0.562 ±0.026	0.545 ±0.049	0.529 ±0.061
	MLP	0.702 ±0.004	0.690 ±0.024	0.662 ±0.021	0.657 ±0.022	0.643 ±0.030
	RF	0.702 ±0.006	0.675 ±0.034	0.640 ±0.032	0.629 ±0.039	0.612 ±0.051
MoLFormer	LP	0.528 ±0.008	0.509 ±0.011	0.525 ±0.014	0.513 ±0.037	0.512 ±0.042
	MLP	0.684 ±0.007	0.669 ±0.030	0.643 ±0.026	0.639 ±0.023	0.626 ±0.035
	RF	0.673 ±0.009	0.637 ±0.042	0.617 ±0.034	0.607 ±0.051	0.598 ±0.053
ChemBERTa-2	LP	0.625 ±0.008	0.616 ±0.016	0.590 ±0.017	0.579 ±0.012	0.572 ±0.017
	MLP	0.726 ±0.006	0.709 ±0.019	0.684 ±0.015	0.679 ±0.022	0.671 ±0.024
	RF	0.721 ±0.010	0.692 ±0.029	0.648 ±0.023	0.645 ±0.024	0.635 ±0.026
Uni-Mol	LP	0.699 ±0.004	0.679 ±0.028	0.649 ±0.019	0.644 ±0.013	0.636 ±0.018
	MLP	0.724 ±0.006	0.707 ±0.016	0.675 ±0.016	0.673 ±0.020	0.667 ±0.028
	RF	0.682 ±0.005	0.659 ±0.026	0.628 ±0.019	0.622 ±0.012	0.608 ±0.022
MolCLR	LP	0.620 ±0.008	0.607 ±0.017	0.579 ±0.022	0.585 ±0.018	0.579 ±0.010
	MLP	0.641 ±0.013	0.626 ±0.016	0.602 ±0.030	0.606 ±0.027	0.587 ±0.017
	RF	0.665 ±0.004	0.626 ±0.036	0.603 ±0.016	0.596 ±0.023	0.590 ±0.019

Table 15: **RMSE on ESOL for the five splitting strategies**. Lower is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3. Note that the resulting splits of 0.3 and 0.4 are identical and therefore also the performance.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\boldsymbol{\alpha = 0.3}$	$\boldsymbol{\alpha = 0.4}$	$\boldsymbol{\alpha = 0.5}$
XGB		1.147 ±0.068	1.521 ±0.345	1.789 ±0.330	1.789 ±0.330	1.922 ±0.445
RF		1.190 ±0.086	1.651 ±0.424	1.930 ±0.356	1.930 ±0.356	2.011 ±0.376
GCN		1.445 ±0.172	1.630 ±0.293	1.617 ±0.258	1.617 ±0.258	2.066 ±0.169
D-MPNN		0.919 ±0.087	0.986 ±0.115	1.117 ±0.233	1.117 ±0.233	1.576 ±0.596
kNN		1.266 ±0.055	1.781 ±0.476	2.213 ±0.552	2.213 ±0.552	2.283 ±0.615
	LP	1.605 ±0.037	1.631 ±0.394	1.724 ±0.520	1.724 ±0.520	2.043 ±0.313
CLAMP	MLP	0.830 ±0.075	0.961 ±0.080	0.982 ±0.090	0.982 ±0.090	1.060 ±0.081
	RF	0.903 ±0.082	1.095 ±0.211	1.188 ±0.236	1.188 ±0.236	1.306 ±0.215
	LP	0.722 ±0.037	0.834 ±0.101	0.939 ±0.101	0.939 ±0.101	1.019 ±0.100
MolBERT	MLP	0.582 ±0.033	0.700 ±0.098	0.749 ±0.111	0.749 ±0.111	0.864 ±0.028
	RF	0.799 ±0.075	1.003 ±0.143	1.129 ±0.156	1.129 ±0.156	1.139 ±0.119
	LP	0.895 ±0.114	1.096 ±0.207	1.430 ±0.417	1.430 ±0.417	1.340 ±0.242
GROVER	MLP	0.673 ±0.043	0.950 ±0.257	1.118 ±0.153	1.118 ±0.153	1.358 ±0.439
	RF	0.918 ±0.060	1.369 ±0.424	1.547 ±0.277	1.547 ±0.277	1.934 ±0.689
	LP	1.993 ±0.031	1.975 ±0.417	2.057 ±0.549	2.057 ±0.549	2.308 ±0.315
SMI-TED	MLP	0.982 ±0.067	1.023 ±0.199	1.238 ±0.333	1.238 ±0.333	1.199 ±0.166
	RF	0.830 ±0.065	0.972 ±0.158	1.178 ±0.229	1.178 ±0.229	1.253 ±0.188
	LP	1.483 ±0.058	1.498 ±0.299	1.558 ±0.434	1.558 ±0.434	1.836 ±0.331
MolT5	MLP	0.746 ±0.051	0.832 ±0.054	0.937 ±0.136	0.937 ±0.136	1.060 ±0.108
	RF	0.873 ±0.058	1.003 ±0.125	1.139 ±0.220	1.139 ±0.220	1.314 ±0.175
	LP	1.773 ±0.030	1.817 ±0.314	1.945 ±0.451	1.945 ±0.451	2.197 ±0.400
MoLFormer	MLP	0.927 ±0.032	1.055 ±0.123	1.206 ±0.247	1.206 ±0.247	1.439 ±0.350
	RF	1.191 ±0.051	1.437 ±0.250	1.587 ±0.284	1.587 ±0.284	1.915 ±0.574
	LP	1.014 ±0.081	1.052 ±0.126	1.105 ±0.179	1.105 ±0.179	1.293 ±0.227
ChemBERTa-2	MLP	0.669 ±0.056	0.727 ±0.051	0.919 ±0.195	0.919 ±0.195	1.027 ±0.144
	RF	0.795 ±0.070	0.925 ±0.094	1.187 ±0.266	1.187 ±0.266	1.164 ±0.071
	LP	1.003 ±0.091	1.087 ±0.072	1.153 ±0.173	1.153 ±0.173	1.266 ±0.140
Uni-Mol	MLP	0.765 ±0.047	0.877 ±0.039	0.956 ±0.100	0.956 ±0.100	1.067 ±0.070
	RF	1.037 ±0.069	1.153 ±0.103	1.267 ±0.240	1.267 ±0.240	1.473 ±0.218
	LP	1.804 ±0.113	1.830 ±0.447	1.897 ±0.486	1.897 ±0.486	2.105 ±0.344
MolCLR	MLP	1.377 ±0.063	1.457 ±0.230	1.650 ±0.292	1.650 ±0.292	1.806 ±0.340
	RF	1.082 ±0.071	1.296 ±0.173	1.527 ±0.170	1.527 ±0.170	1.722 ±0.376

Table 16: **RMSE on FreeSolv for the five splitting strategies**. Lower is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\boldsymbol{\alpha = 0.3}$	$\boldsymbol{\alpha = 0.4}$	$\boldsymbol{\alpha = 0.5}$
XGB		1.663 ±0.393	2.010 ±0.467	2.161 ±0.608	2.408 ±0.656	2.703 ±0.865
RF		2.039 ±0.541	2.385 ±0.543	2.608 ±0.674	2.867 ±0.832	3.056 ±1.030
GCN		2.634 ±0.343	2.797 ±0.346	3.060 ±1.058	3.456 ±0.933	3.510 ±1.139
D-MPNN		2.282 ±0.170	2.750 ±0.577	2.709 ±0.737	2.777 ±0.694	2.787 ±0.663
kNN		2.232 ±0.513	2.798 ±0.379	2.880 ±0.751	3.049 ±0.974	3.314 ±1.184
	LP	3.422 ±0.609	3.508 ±0.456	3.599 ±0.775	3.761 ±1.178	3.906 ±1.191
CLAMP	MLP	1.810 ±0.443	2.386 ±0.269	2.318 ±0.657	2.499 ±0.669	2.900 ±0.563
	RF	2.207 ±0.568	2.413 ±0.164	2.573 ±0.734	2.852 ±0.913	3.400 ±0.871
	LP	1.649 ±0.386	2.074 ±0.421	2.063 ±0.832	2.106 ±0.713	2.225 ±0.863
MolBERT	MLP	1.084 ±0.235	1.445 ±0.205	1.645 ±0.870	1.691 ±0.758	1.820 ±0.903
	RF	1.474 ±0.583	1.719 ±0.085	1.990 ±0.728	2.331 ±0.918	2.565 ±0.805
	LP	1.431 ±0.252	1.767 ±0.141	1.832 ±0.440	1.900 ±0.476	2.512 ±0.776
GROVER	MLP	1.120 ±0.197	1.593 ±0.146	1.700 ±0.676	1.774 ±0.502	2.220 ±0.569
	RF	1.518 ±0.290	1.834 ±0.277	2.218 ±0.528	2.337 ±0.720	2.767 ±0.666
	LP	3.733 ±0.605	3.789 ±0.470	3.930 ±0.757	4.066 ±1.191	4.161 ±1.237
SMI-TED	MLP	2.660 ±0.465	3.309 ±0.757	2.956 ±0.783	2.912 ±0.850	3.116 ±0.582
	RF	1.743 ±0.339	2.101 ±0.426	2.250 ±0.464	2.657 ±0.733	2.772 ±0.686
	LP	3.220 ±0.579	3.291 ±0.451	3.321 ±0.724	3.438 ±1.080	3.575 ±1.035
MolT5	MLP	1.610 ±0.321	1.983 ±0.315	1.939 ±0.678	2.009 ±0.747	2.015 ±0.673
	RF	1.892 ±0.267	2.074 ±0.227	2.280 ±0.768	2.465 ±0.798	2.669 ±0.887
	LP	3.384 ±0.597	3.455 ±0.487	3.508 ±0.726	3.637 ±1.124	3.814 ±1.090
MoLFormer	MLP	1.904 ±0.404	2.273 ±0.363	2.207 ±0.728	2.279 ±0.806	2.400 ±0.773
	RF	2.218 ±0.485	2.596 ±0.434	2.641 ±0.721	2.882 ±0.979	3.496 ±0.911
	LP	2.863 ±0.554	3.053 ±0.464	3.007 ±0.707	3.156 ±1.058	3.337 ±0.918
ChemBERTa-2	MLP	1.278 ±0.269	1.566 ±0.289	1.659 ±0.578	1.809 ±0.663	1.912 ±0.832
	RF	1.854 ±0.403	2.290 ±0.398	2.447 ±0.643	2.654 ±0.844	2.997 ±0.870
	LP	2.233 ±0.421	2.533 ±0.318	2.415 ±0.739	2.351 ±0.927	2.506 ±0.897
Uni-Mol	MLP	1.717 ±0.274	1.948 ±0.333	2.132 ±0.762	2.149 ±0.575	2.452 ±0.757
	RF	2.364 ±0.492	2.414 ±0.256	2.669 ±0.677	2.816 ±0.943	3.006 ±0.901
	LP	3.676 ±0.594	3.717 ±0.467	3.856 ±0.761	4.013 ±1.188	4.139 ±1.266
MolCLR	MLP	3.063 ±0.579	3.511 ±0.642	3.415 ±0.786	3.600 ±1.123	3.869 ±1.097
	RF	2.082 ±0.450	2.412 ±0.357	2.928 ±0.695	3.093 ±0.728	3.477 ±0.801

Table 17: **RMSE on Lipo for the five splitting strategies**. Lower is better. Each entry is the weighted mean over folds, ± the weighted standard deviation over folds. Foundation models are shown with each of the three heads separately. MCES splits use $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

				MCES		
Model	**Head**	**Uniform**	**Scaffold**	$\boldsymbol{\alpha = 0.3}$	$\boldsymbol{\alpha = 0.4}$	$\boldsymbol{\alpha = 0.5}$
XGB		0.823 ±0.017	0.886 ±0.036	1.032 ±0.078	1.063 ±0.091	1.082 ±0.101
RF		0.823 ±0.019	0.903 ±0.050	1.067 ±0.089	1.097 ±0.103	1.105 ±0.113
GCN		0.760 ±0.034	0.828 ±0.063	0.970 ±0.063	0.974 ±0.116	1.007 ±0.165
D-MPNN		0.750 ±0.032	0.776 ±0.039	0.830 ±0.045	0.904 ±0.052	0.887 ±0.079
kNN		0.910 ±0.027	0.994 ±0.020	1.163 ±0.106	1.179 ±0.098	1.183 ±0.117
CLAMP	LP	0.956 ±0.016	0.965 ±0.056	0.992 ±0.062	1.011 ±0.100	1.013 ±0.101
	MLP	0.698 ±0.007	0.754 ±0.022	0.869 ±0.060	0.900 ±0.043	0.913 ±0.069
	RF	0.789 ±0.012	0.857 ±0.043	0.964 ±0.043	0.970 ±0.075	0.982 ±0.067
MolBERT	LP	0.724 ±0.017	0.751 ±0.013	0.846 ±0.054	0.877 ±0.029	0.879 ±0.037
	MLP	0.605 ±0.015	0.661 ±0.028	0.805 ±0.084	0.820 ±0.038	0.846 ±0.047
	RF	0.760 ±0.017	0.823 ±0.023	0.936 ±0.015	0.965 ±0.048	0.967 ±0.081
GROVER	LP	0.793 ±0.036	0.821 ±0.052	0.892 ±0.022	0.935 ±0.085	0.943 ±0.070
	MLP	0.672 ±0.031	0.715 ±0.038	0.826 ±0.051	0.870 ±0.094	0.889 ±0.086
	RF	0.806 ±0.016	0.882 ±0.043	0.995 ±0.068	1.020 ±0.090	1.022 ±0.111
SMI-TED	LP	1.183 ±0.029	1.185 ±0.080	1.186 ±0.086	1.193 ±0.105	1.190 ±0.109
	MLP	0.923 ±0.008	0.934 ±0.045	0.980 ±0.028	1.002 ±0.050	1.001 ±0.075
	RF	0.862 ±0.022	0.925 ±0.053	1.021 ±0.042	1.035 ±0.062	1.036 ±0.082
MolT5	LP	1.049 ±0.023	1.056 ±0.063	1.074 ±0.072	1.085 ±0.093	1.084 ±0.112
	MLP	0.839 ±0.022	0.874 ±0.035	0.996 ±0.028	1.024 ±0.056	1.049 ±0.125
	RF	0.877 ±0.014	0.937 ±0.053	1.047 ±0.061	1.059 ±0.079	1.064 ±0.105
MoLFormer	LP	1.089 ±0.021	1.097 ±0.058	1.108 ±0.069	1.120 ±0.094	1.117 ±0.110
	MLP	0.899 ±0.008	0.942 ±0.037	1.021 ±0.036	1.051 ±0.072	1.061 ±0.110
	RF	0.977 ±0.011	1.018 ±0.056	1.078 ±0.044	1.085 ±0.066	1.085 ±0.103
ChemBERTa-2	LP	0.929 ±0.020	0.946 ±0.035	0.997 ±0.046	1.030 ±0.057	1.030 ±0.079
	MLP	0.710 ±0.024	0.765 ±0.035	0.914 ±0.040	0.936 ±0.064	0.964 ±0.062
	RF	0.834 ±0.010	0.904 ±0.041	1.006 ±0.036	1.036 ±0.041	1.033 ±0.093
Uni-Mol	LP	0.982 ±0.014	1.000 ±0.040	1.052 ±0.064	1.057 ±0.076	1.054 ±0.098
	MLP	0.889 ±0.025	0.940 ±0.049	1.056 ±0.061	1.091 ±0.082	1.098 ±0.111
	RF	1.001 ±0.015	1.028 ±0.052	1.087 ±0.075	1.087 ±0.083	1.089 ±0.095
MolCLR	LP	1.146 ±0.025	1.151 ±0.085	1.154 ±0.074	1.169 ±0.095	1.167 ±0.095
	MLP	1.038 ±0.020	1.048 ±0.056	1.062 ±0.043	1.077 ±0.066	1.081 ±0.083
	RF	0.875 ±0.022	0.933 ±0.049	1.031 ±0.028	1.051 ±0.061	1.048 ±0.093

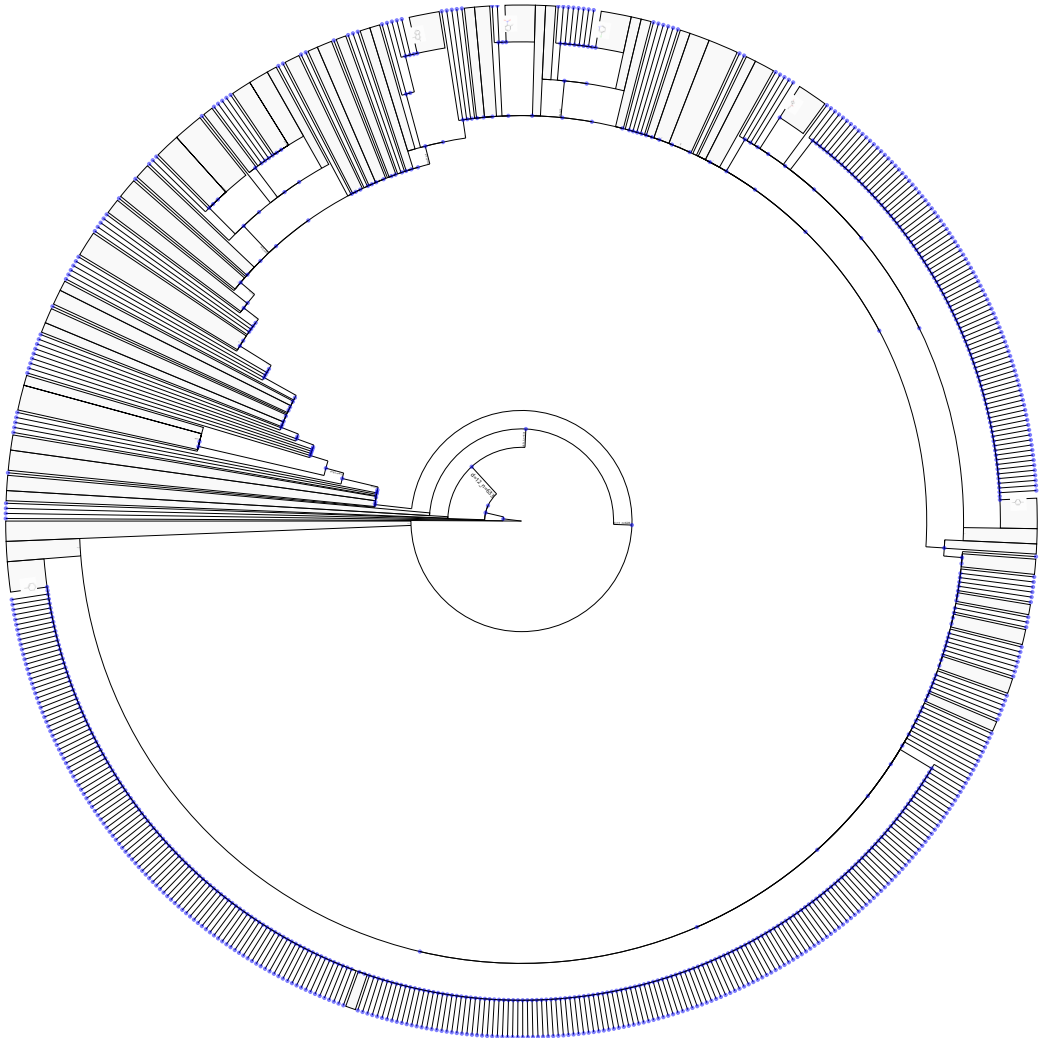

Figure 2: **Dendrogram of the FreeSolv dataset, showing MCES distance single linkage clusters.** Notably, the largest cluster consists of a long chain of small molecules, a typical chaining effect of single linkage, which prevents choosing a suitable cut-off τ. We use the Environment for Tree Exploration (ETE) toolkit (Huerta-Cepas et al., 2016) for visualization. An interactive version is provided as Jupyter Notebook in the project repository.

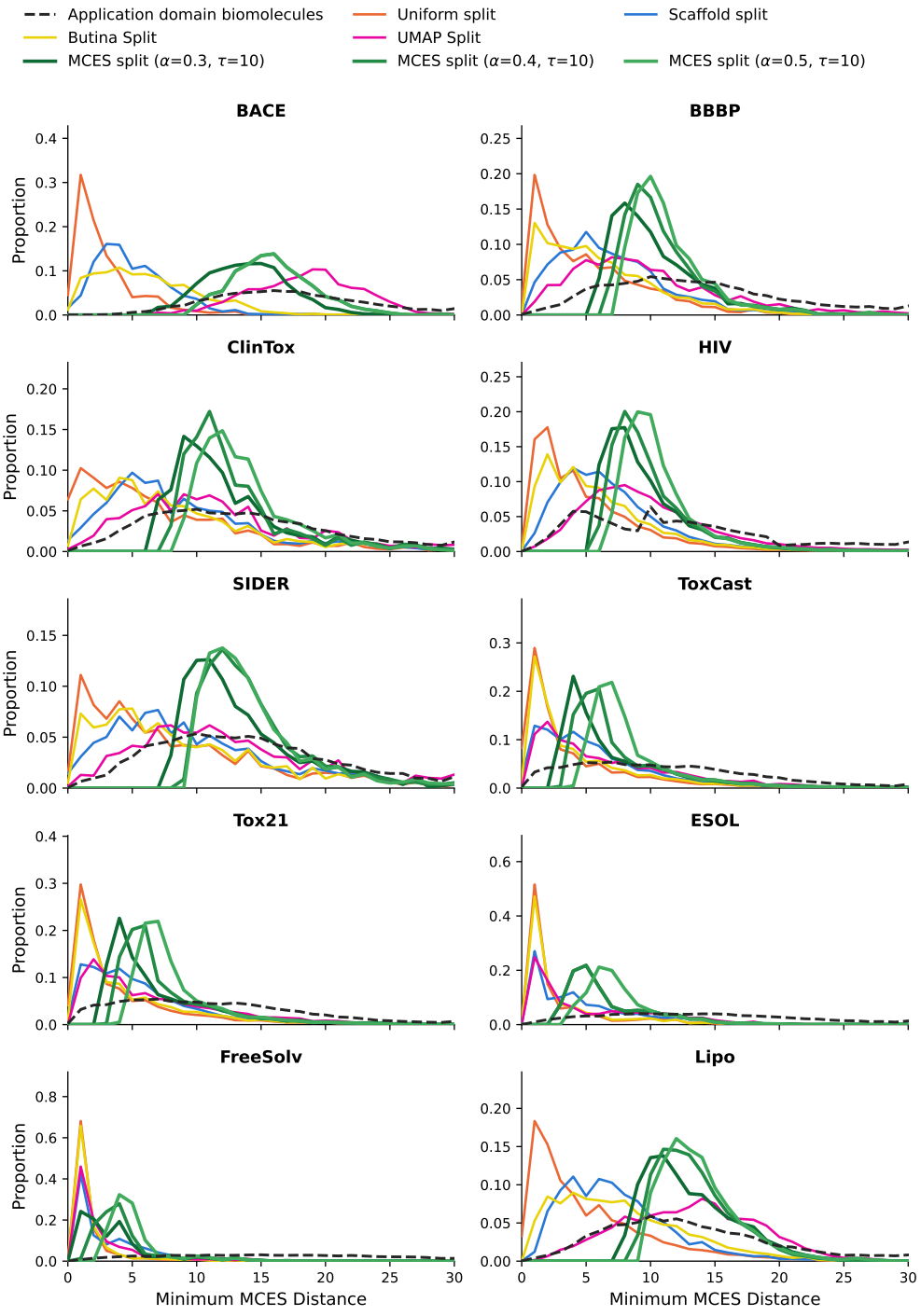

Figure 3: **MCES distance distributions.** Shown are the minimum MCES distances between test and train sets for uniform, scaffold, Butina, UMAP and MCES splits. In addition, we show minimum MCES distances between dataset and biomolecule reference. All distances rounded down; note the varying y-axis. For MCES splits, we use $\tau = 10$ and $\alpha \in \{0.3, 0.4, 0.5\}$. The effective distance threshold can be smaller, see Table 3. For BACE, $\alpha \in \{0.4, 0.5\}$ results in the same distribution; for ESOL this is the case for $\alpha \in \{0.3, 0.4\}$. For certain datasets such as FreeSolv, integrating all distributions in one plot is hard, see also Fig. 1.

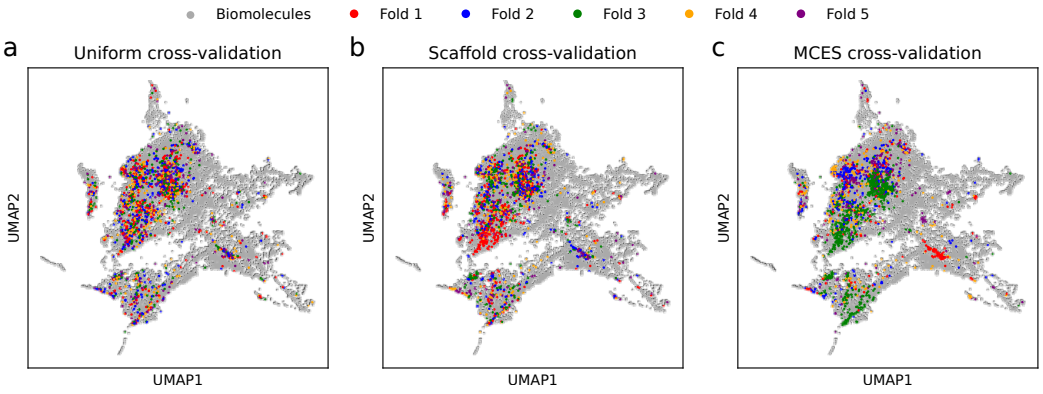

Figure 4: **UMAP plot for BBBP showing three different split methods**. MCES cross-validation folds increase separation between structurally similar molecules, as visualized via Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020). We use the BBBP dataset as example; the subsample of 20k biomolecules is used as the background distribution, mirroring the desired application domain, onto which UMAP embeddings are projected, as described in Kretschmer et al. (2025). We show (a) uniform, (b) scaffold, and (c) MCES split, with 5-fold cross-validation.

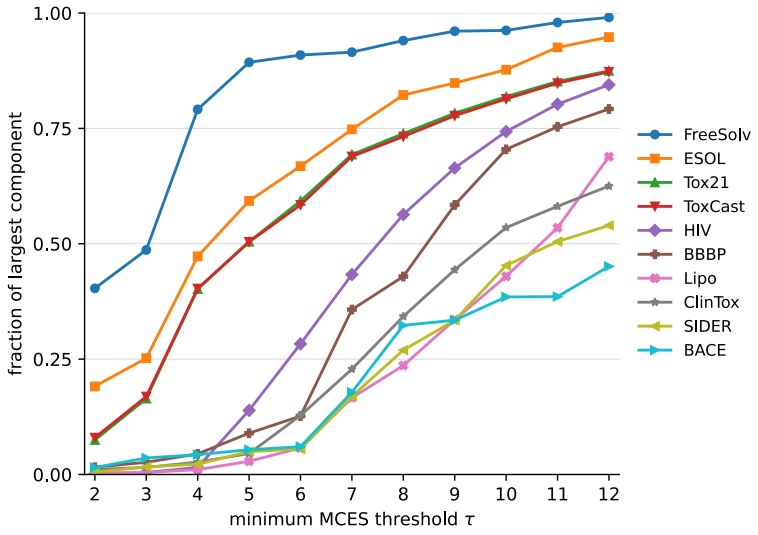

Figure 5: **Size of the largest cluster for different datasets and distance thresholds.** Largest connected component, as a fraction of n, produced by single-linkage clustering as a function of τ. Two groups of datasets emerge: For the first group, this fraction stays small up to $\tau \approx 6$ before growing gradually; for the second, the fraction grows sharply starting at $\tau = 3$.

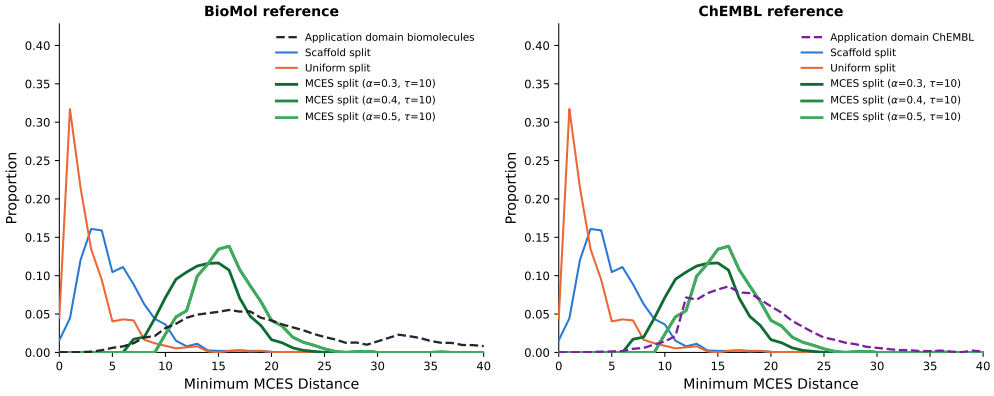

Figure 6: **MCES distance distributions.** Shown are the minimum MCES distances between test and train sets for uniform, scaffold, and MCES splits for the BACE dataset. Here, we compare biomolecules as application domain to small molecules from the ChEMBL database, as the arguably more natural application domain. For ChEMBL, we use the subsample of 20,000 structures from Kretschmer et al. (2025), see Supplementary Fig. 11 there for details. All distances rounded down; note the varying y-axis. For MCES splits we use $\tau = 10$ and $\alpha \in \{0.3, 0.4, 0.5\}$. Notably, some distances are below the threshold $\tau = 10$ even for MCES splits.

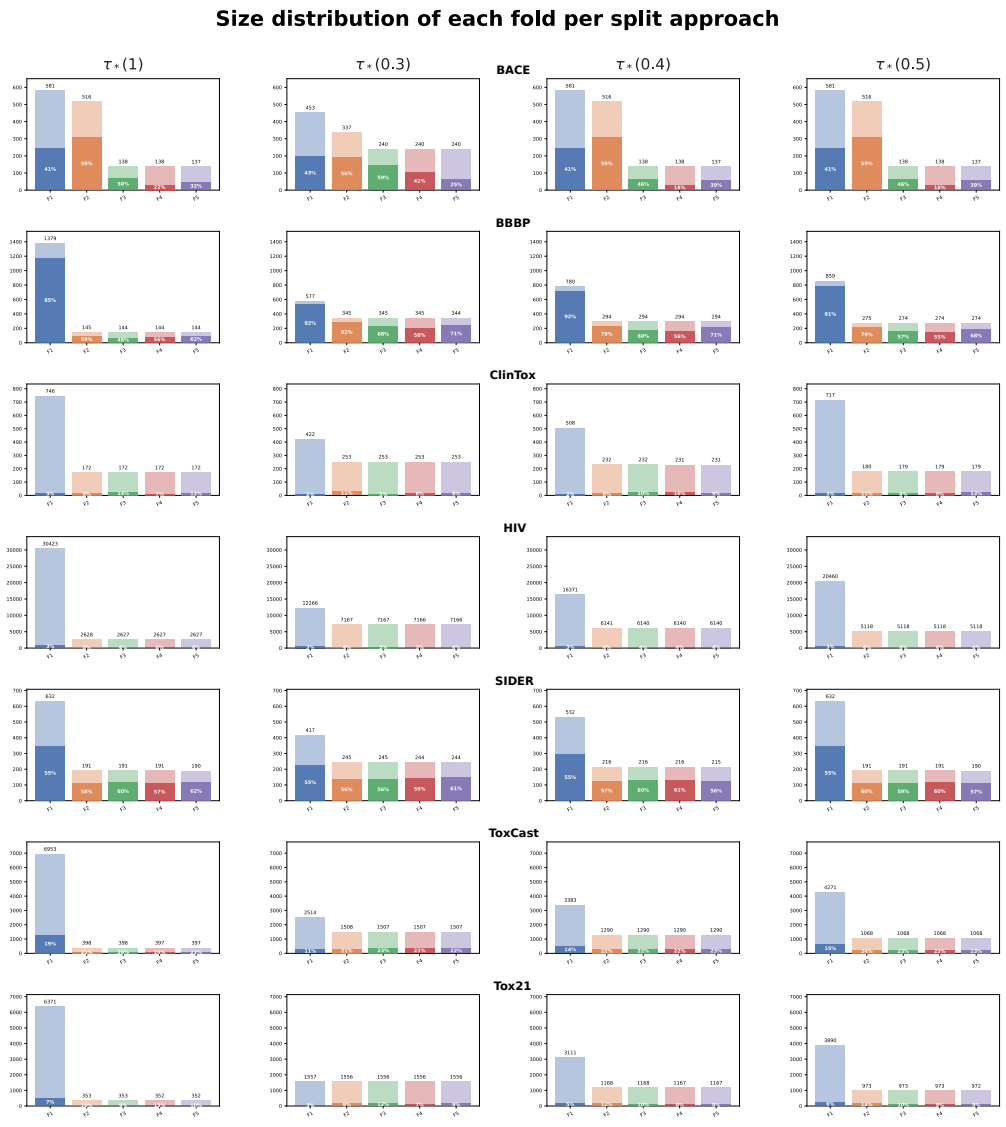

Figure 7: **Comparison of different MCES splits for classification datasets, with respect to how balanced the folds are**. Each bar shows the size of one fold, with the shaded portion giving the positive-label fraction. Shown are all classification datasets with $\alpha \in \{0.3, 0.4, 0.5, 1\}$ and $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

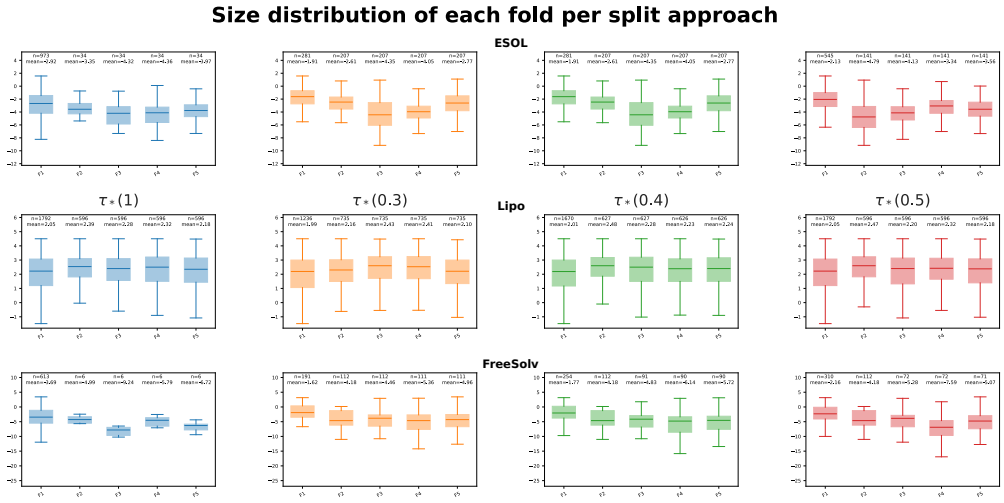

Figure 8: **Comparison of different MCES splits for regression datasets, with respect to how balanced the folds are**. Each box shows the label distribution of one fold, annotated with fold size and mean. Shown are all regression datasets with $\alpha \in \{0.3, 0.4, 0.5, 1\}$ and $\tau = 10$. The effective distance threshold can be smaller, see Table 3.

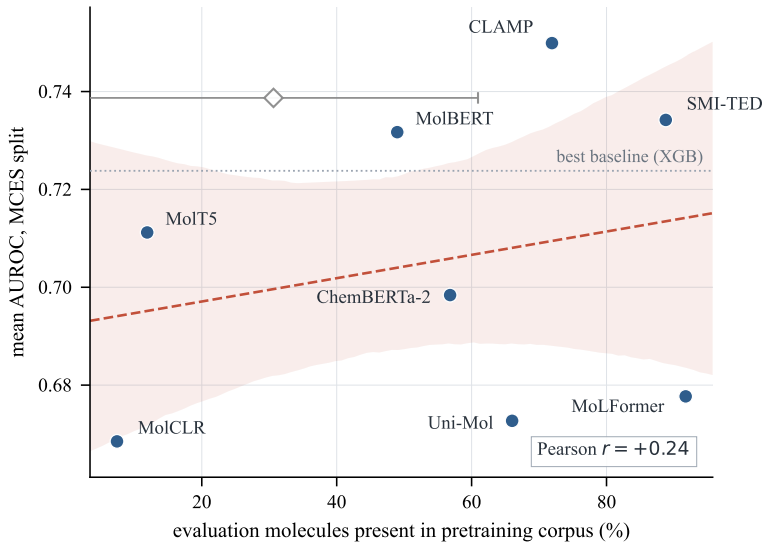

Figure 9: **Structural exposure between pretraining and evaluation data against model effectiveness**. Both axes use the seven classification datasets. The horizontal axis is the per-molecule InChIKey overlap pooled over BACE, BBBP, ClinTox, HIV, SIDER, Tox21 and ToxCast. The vertical axis is the $\alpha = 0.3$ column of Table 1: mean AUROC under the MCES split at $\tau = 10$, using the same probe head per foundation model as that table. Dashed line is an ordinary least squares fit, shaded band a bootstrap 90 % interval. GROVER (open diamond) is a bracket on this same pool rather than a measured value and is excluded from the fit.