MoG-Bench: Robust Evaluations for CFMs

Anonymous authors
Paper under double-blind review

Abstract

Experimentally measuring the properties of novel molecules is expensive, slow, and hard to scale. Chemical Foundation Models (CFMs) provide a way to train models on small sets of experimental data, and make predictions for many more molecules. However, evaluations of these models in practical settings remains limited. CFMs have largely been evaluated on traditional property prediction datasets that do not test generalization to new molecules. To address these limitations, we develop the Molecular Generalization Benchmark (MoG-Bench), a set of twenty datasets spanning four categories: binding affinity, ADMET, functional activity, and screening. We make the following contributions: 1) careful curation and quality control for each dataset in MoG-Bench, including a predefined data split that estimates models' ability to generalize to chemically distinct compounds; 2) two variants, MoG-Bench-Medium Data (MD) and MoG-Bench-Low Data (LD), evaluating six state-of-the-art foundation model performance on standard supervised learning and few-shot learning, respectively; 3) practical guidance on how to maximize performance of these models and takeaways to guide future model development.

1 Introduction

Predicting molecules' properties is important to many industries (Fang et al., 2022; Cumming et al., 2013; van de Waterbeemd & Gifford, 2003; Wu et al., 2018). In many areas like pharmaceuticals, agriculture, and materials science, the limited availability of experimental data and the cost of collecting additional experimental data makes finding methods that can make accurate predictions for new compounds even more important. This challenge has led to a surge of recent work on Chemical Foundation Models (CFMs) (Singh et al., 2026; Cai et al., 2025; Rong et al., 2020; Adrian et al., 2025; Burns et al., 2025; Seidl et al., 2023; Rottach et al., 2026).

CFMs seek to learn a robust way of encoding molecules that can be used for many different property prediction tasks. CFMs are first trained on vast datasets of compounds without seeing experimental measurements, with various self-supervised learning objectives. These objectives have the goal of training the model to learn to represent underlying chemical rules and relationships between molecules. These models are then used to generate embeddings for datasets of molecules with labels for the properties a user is interested in. Once the embeddings have been generated for labeled molecules, a new model can be trained with the embeddings and the user's target labels. As long as the foundation model can generate embeddings for new compounds, the new supervised model can be used to make predictions for these new compounds. For tasks with small datasets—where fitting a large, more powerful machine learning model would be impossible—this promises to improve performance by borrowing the underlying patterns and chemical rules learned by the molecular foundation model. The performance of the new supervised model will be improved by the patterns the pretrained chemical foundation model learned from training on unlabeled data. Pretraining should allow CFMs to make accurate predictions for new compounds that are significantly different from the compounds the model was trained on. The flexibility of foundation models is especially appealing for making predictions for biologically active small molecules, where experimental data is limited and many different tasks are critical. However, this also means understanding how well the representations of foundation models perform on new molecules is critical to their application.

Even though a large number of CFMs have been developed, with some of them widely used, no rigorous benchmark has been developed for CFMs. Foundation models have instead been evaluated on a variety benchmarks developed for other purposes, leading to challenges in predicting their performance in real-world drug discovery and other practical applications. Previous work has shown

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.

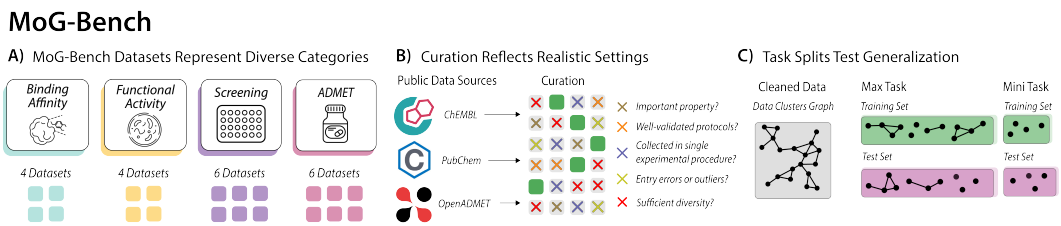

Figure 1: MoG-Bench consists of A) 20 molecular property prediction tasks across four categories: ADMET (6 tasks), screening (6 tasks), binding affinities (4 tasks) and functional activity (4 tasks). Each categories' tasks has been B) carefully curated and reviewed for real world relevance, and C) features a rigorous chemical diversity split.

that drug discovery task evaluations tend to overestimate the performance of machine learning models due to the similarity of compounds in training sets and test sets (Wu et al., 2018; Huang et al., 2021; Walters, 2023). Other work has suggested that depending on the evaluation condition and tasks considered, many chemical foundation models may fail to outperform simple baseline methods (Praski et al., 2025). Finally, many methods report performance on tasks that are not representative of real-world applications, such as predicting computationally calculated molecular properties or descriptors from quantum mechanics (Ramakrishnan et al., 2014).

To accurately evaluate the performance of a foundation model in the context of a real-world application such as drug discovery, one must address several challenges. Most available datasets for curating a CFM benchmark are contaminated by various artifacts. Publicly available data is often collected across many different labs using different methodologies, leading to high variability in measurements even for the same compound (Landrum & Riniker, 2024; Schiebroek et al., 2026; Walters, 2023). Finally, predictions for compounds very different from those that have already been measured are particularly important, because chemically similar compounds are likely to behave similarly.

Based on these observations, we introduce the **Mo**lecular **G**eneralization **Benchmark**, or **MoG-Bench** (Figure 1), a suite designed to rigorously evaluate the performance of CFMs for biologically active molecules. Our contributions are the following:

1. We carefully curate high-quality experimental datasets corresponding to 20 diverse tasks important for molecular property prediction and drug discovery, including prediction of binding affinity, functional activity, ADMET properties, and screening results. We perform rigorous quality control on this data, ensuring that strong model performance on MoG-Bench is representative of strong performance on real drug discovery tasks.
2. We employ a diverse train/test splitter (called MBDSplitter) that guarantees molecules in training sets are chemically very different from molecules in test sets. In addition, we ensure that the molecules we test on are diverse, and do not just belong to one chemical series, ensuring that final validation numbers are representative of models' ability to generalize to a wide range of out-of-distribution compounds.
3. Using MBDSplitter we split datasets into two sub-benchmarks, MoG-Bench-MD (Medium Data) and MoG-Bench-LD (Low Data). MoG-Bench-MD splits data into 3 folds, minimizing holdout in order to maximize total number of training and testing compounds. In MoG-Bench-LD, we deliberately decrease the number of training compounds to 24 to simulate a few-shot learning scenario for the foundation models. Only having a few dozen experimental measurements for compounds is very common in drug discovery, as collecting more experimental data is time- and resource-intensive. Useful models must be able to learn from a few labeled compounds and still be able to generalize to out-of-distribution data.
4. We evaluate six state-of-the-art foundation models and offer practical guidance on using these models, including using linear probes with frozen embeddings for downstream prediction, avoiding finetuning, and that cancellation of errors from using multiple molecular foundation models can help boost performance on tasks.

Table 1: MoG-Bench tasks: category, task name, description, and median split sizes over the three folds.

Category	Task Name	Description	Train	Test
ADMET	MDR1	Measures membrane permeability	1,760	880
	LogD	Measures lipophilicity at a set concentration	6,007	33
	PPB	Percentage of drug that binds to blood plasma proteins	1,023	512
	Liver intrinsic clearance	Measures liver's metabolism	3,628	32
	Efflux	Ratio of a drug being effluxed vs. entering	3,002	50
	Cytotoxicity	Measures the potential of a drug to kill cells	267	133
Binding	gp160	Measures IC_{50} against HIV-1 envelope glycoprotein gp160	231	105
	RET	Measures K_d binding to the RET receptor tyrosine kinase	59	29
	ALPHA1B	Measures K_i binding to the alpha-1B adrenergic receptor	49	24
	SIGMA2	Measures K_i binding to the sigma-2 receptor (TMEM97)	49	25
Functional	AT1R	Measures beta-arrestin recruitment in AT1R	256	129
	B2AR	Dose-response assay against B2AR	114	57
	OATP1B1	Inhibition of sodium uptake in OATP1B1	1,042	492
	MRP2	Drug interactions with ATP-binding cassette transporter	115	58
Screening	MITF	Inhibition of microphthalmia-associated transcription factor	5,738	1,460
	MEX-5	Inhibition of MEX-5 RNA-binding protein	6,000	1,500
	Ubiquitin	Inhibition of the ubiquitin receptor	6,000	1,500
	HepG2	Inhibition of liver cancer cells	6,000	1,500
	NSCLC	Non-small-cell lung cancer cell growth inhibition assay	5,964	1,500
	SKN-1	Inhibition of SKN-1 receptor in *C. elegans*	6,000	1,500

Task sources: MDR1 (Fang et al., 2023); LogD, liver intrinsic clearance, and Efflux (Castellanos et al., 2026); PPB (Wenlock & Tomkinson, 2014); Cytotoxicity (Ketas & Moore, 2014); gp160 (Cai & Gochin, 2007); RET (Klaeger et al., 2017); ALPHA1B (ChEMBL, 2024); SIGMA2 (EUbOPEN, 2024); AT1R (EUbOPEN, 2023); B2AR (PubChem, d); OATP1B1 (De Bruyn et al., 2013); MRP2 (Pedersen et al., 2008); MITF (Wang et al., 2017); MEX-5 (PubChem, b; Pagano et al., 2007); Ubiquitin (PubChem, c); HepG2 (Abraham et al., 2020); NSCLC (PubChem, a; Alley et al., 1988); SKN-1 (PubChem, e).

2 The MoG-Bench Dataset

2.1 Task Selection

We focus on four main categories of molecular property prediction tasks that are representative of early-stage drug discovery: ADMET, screening, binding affinity, and functional activity. For each category, all tasks were manually curated and reviewed. These tasks were evaluated for practical relevance to real world molecular property prediction, and critically, carefully examined for the quality and robustness of the underlying experimental protocols used for data collection. A summary of the categories, tasks, and dataset sizes can be found in Table 1.

2.1.1 Screening

When finding a suitable drug candidate to treat a disease, medicinal chemists have an effectively unlimited choice of compounds to consider. Due to cost and time considerations, however, only a few compounds can be synthesized and experimentally tested. An attractive alternative is an *in silico* virtual screen, which uses a computational approach to predict the compounds that will have a desired activity profile. In cases where the specific target is known and is a protein, researchers may apply computational methods to predict the binding of molecules to the protein. In other cases where the target is not know or fully understood, researchers nonetheless might have data on compounds that were screened for a desired functional activity against cells or organisms in a disease state, and want to find additional compounds with improved activity profiles for that state. In either case, a machine learning model that can identify active compounds is highly relevant to the virtual screening task. We feature screening tasks for various types of targets, including known specific protein targets, functional targets at the cellular level, and targets at the organism level.

2.1.2 Binding Affinity

After a large screen, chemists are left with a few promising candidate molecules. Beyond just knowing if a molecule will bind or not, chemists want to know *how well* a molecule will bind. A tighter

binder means a molecule will work at a lower concentration and that less of it is required to have the intended effect, reducing risk of toxicity. Therefore, it is also important for a molecular foundation model to be able to predict strength of binding of these top molecules.

2.1.3 FUNCTIONAL ACTIVITY

Screening and binding affinity tasks aim to answer the question: does this molecule bind to the target, and how well? Functional activity tasks aim to answer: once the molecule is bound, what is the downstream effect? Functional activity assays therefore focus on measuring these effects, which can be several steps downstream of direct binding, and be parts of large biochemical pathways. Given a compound known to bind a protein, a functional assay might measure if the compound is an agonist, a compound that activates the protein target and stimulates the biochemical pathways that target controls, a partial agonist, or an antagonist, a compound that prevents the natural activation of the protein target and the downstream pathways it controls.

2.1.4 ADMET PROPERTY PREDICTION

ADMET (Absorption, Distribution, Metabolism, Excretion, Toxicity) are a set of properties of a drug that define how a drug moves through and interacts with different systems in the body, including any potential safety risks or harms that it could cause. While many chemical compounds may be capable of binding to a target protein, a candidate drug must also be able to be absorbed by the body, reach the tissues impacted by disease, and be retained long enough to have an effect. During this process it must also not have serious negative side-effects or cause harm to any of the other systems of the body, or otherwise its toxicity will preven its use. ADMET broadly covers all of these aspects of a drug. Because of this, medicinal chemists are especially interested in modeling the ADMET properties of a drug, which are often major factors in new compounds failing to get approved (Scotti & Scotti, 2025).

2.2 QUALITY CONTROL

When identifying and selecting data for tasks, it is often appealing to gather datasets that are as large as possible. For many of the tasks described in the previous sections, much larger datasets can be obtained by pooling results collected from multiple different sources. However, we require that all data for a specific molecular property prediction task come from a **single** experimental data source, or assay, as a requirement for our quality control process.

We believe this step is critical for several reasons. First, the experimental values for compounds for each of the tasks described are often very sensitive to the actual experimental protocol used to measure and collect the data. Combining data collected with different assay protocols, or even data from experiments conducted in different laboratories that may nominally be using the same protocol, often results in systematic error and high variability even in measurements for the same exact compound (Landrum & Riniker, 2024; Schiebroek et al., 2026; Kramer et al., 2012). This is especially problematic for publicly available data, which is often collected across many different labs using different methodologies, with little explicit documentation. The high amount of noise in combining assays limits our ability to understand and measure how well a model actually works.

Given our requirement that data comes from a single experimental source, we carefully manually reviewed all of the underlying publications, reports, and documentation available for each of the tasks we included in MoG-Bench. We verified the quality of the underlying experimental procedure and also monitored any abnormalities in the reported data, discarding any datasets or tasks that failed to pass these manual curation standards. In addition, we applied a rigorous computational quality control pipeline for further curation and processing. Those additional steps are listed in the Appendix.

2.3 DATA SPLITTING

When preparing molecular datasets for machine learning, the splitting procedure used to obtain train, validation, and test sets can have major impacts on model performance (Walters, 2023; Liu et al., 2024). That's because compounds that are chemically similar or that share significant chemical structure, are much more likely to also share similar properties, and cannot be treated as independent samples. Still, the two most common splitting procedures used for molecular datasets are **Random splits** and **Bemis-Murcko Scaffold** splits.

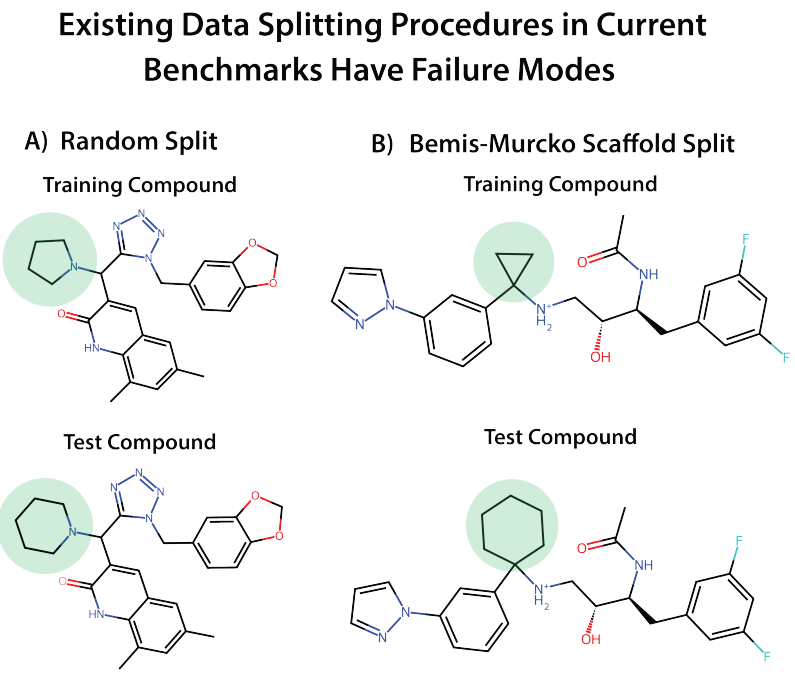

Figure 2: Issues with existing splitting methods. A) Random splits and B) Bemis-Murcko Scaffold (BMS) splits, which yield highly similar molecules appearing in both train and test.

Random splits can result in highly similar compounds being placed in train and test sets, as shown in Figure 2. While this may be acceptable for certain purposes, such as making predictions for a series of closely related compounds, this setup is not suitable for evaluating the generalization of CFMs, due to increased training data leakage that may benefit larger models. Bemis-Murcko scaffold splits are often used instead for this purpose. They are based around the concept of a scaffold in medicinal chemistry, which refers to a series of chemical compounds that result from the modification of a shared parent substructure. While this is a better fit for real drug development in theory, in practice the algorithm used for Bemis-Murcko scaffold splits is very rigid and does not always reflect the chemical intuition behind a scaffold split. In Figure 2 we highlight in a real pair of train-test compounds from an existing dataset how they share almost all substructure despite a train-test split.

To reduce such leakage and properly evaluate models, we introduce our diversity split, the MoG-Bench Diversity Splitter (MBDSplitter), which explicitly guarantees and enforces a maximum similarity between compounds placed in train and test sets of 0.4 Tanimoto similarity for Morgan fingerprints. This greatly reduces full substructures from being shared between train and test compounds. Such splitters already exist (Steshin, 2023; Ramsundar et al., 2019); however, MBDSplitter differs from these splitters as we place constraints on the molecules *within* test. It is possible (and highly probable for small datasets) that test sets are comprised entirely of a single cluster containing many similar compounds. While respecting the strict Tanimoto cutoff, the Hi split hinders foundation model evaluation. Evaluations on single chemical series can be strongly correlated and not representative of the model's performance on a diverse range of chemical space. Therefore, this necessitates the need for the MBDSplitter, which places constraints on the diversity of the test sets. More details on MBDSplitter can be found in the Appendix.

2.4 MoG-Bench Low Data

In addition to these tasks, we wanted to understand models' abilities on few shot learning. In the real world, experimental asssays may often only yield a handful of compounds. In fact, the original mog bench tasks have a lot more compounds than the standard experimental assay in chembl. To replicate real world assay envirnments, we introduce mog-bench-LD, where we resplit the existing tasks in the same methodology, but reduce the number of training compounds down to 2 dozen, while

still evaluating on as many compounds as possible. That way, we can evaluate foundation models' performance in these real world setting and uinderstand their few shot learning performance.

2.5 Metric Selection

The binding affinity, functional, and ADMET tasks all comprise continuous-valued properties of a molecule, such as its permeability, percent inhibition, and inhibition constant. We refer to these three task groups as the regression tasks. For regression tasks, all statistics are reported as Kendall's τ correlation (Kendall et al., 1938). We use Kendall's τ over Pearson's r and Spearman's ρ for several reasons. First, r requires normality in the data, is calculated on the raw absolute labels predicted by the model, and is highly susceptible to outliers. ρ relies on rankings instead of raw values, but is still susceptible to outliers–a compound incorrectly predicted as having a low rank would be penalized heavily by Spearman's squared distance. Therefore, we use Kendall's τ as it is more robust to outliers–it only relies on paired rankings, and not on ranking distance. Moreover, it has been shown to shrink confidence intervals, control Type I errors, and be robust on low data regimes (Kendall et al., 1938; Arndt et al., 1999; Croux & Dehon, 2010). While not all of our tasks are low data, we use τ for all regression tasks for consistency.

For screening tasks, the goal is to identify likely hits amongst a large number of compounds, most of which are inactive against a target (the "needle in a haystack" problem in drug discovery). A commonly used metric is LogAUC (Knight et al., 2022; Mysinger & Shoichet, 2010; Mysinger et al., 2012; Stein et al., 2021; Coleman et al., 2013; Lyu et al., 2019; Sterling & Irwin, 2015); however, the use of a cutoff parameter in the metric can cause some instability in LogAUC scores (Knight et al., 2022; Mysinger & Shoichet, 2010; Clark & Webster-Clark, 2008). To this end, we instead utilize the Enrichment Score (ES) metric introduced by Knight et al. (2022), which stabilizes the aforementioned cutoff parameter.

2.6 Comparison to Other Molecular Benchmarks

No previous benchmarks have been developed specifically for chemical foundation models, but a number of datasets have been developed for—and proven useful for—other purposes. We summarize the differences the differences between MoG-Bench and some of the most commonly used drug discovery benchmark sets here, in the context of benchmarking chemical foundation models. A full description of each of the benchmarks can be found in the Appendix.

Table 2: Comparison of molecular machine learning benchmarks. For each benchmark, we compare whether its tasks are computational or experimental, whether the experimental data were curated from a single assay, whether the molecules are diverse across train and test, and whether it supports few-shot learning, defined as having fewer than 30 training compounds.

Benchmark	**Diverse set of tasks**	**From one assay**	**Diverse molecules in train & test**	**Supports few-shot learning**
MoG-Bench	✓	✓	✓	✓
MoleculeNet	✓	✗	✗	✗
TDC	✓	✗	✗	✗
GEOM-Drugs	✓	✗	✗	✗
QM9	✗	✗	✗	✗
WelQRate	✗	✓	✗	✗
MoleculeACE	✗	✓	✗	✗
LoHi	✗	✗	✓	✗
FS-Mol	✗	✓	✗	✓

Sources: MoleculeNet (Wu et al., 2018); TDC (Huang et al., 2021); QM9 (Ramakrishnan et al., 2014); WelQRate (Liu et al., 2024); MoleculeACE (Van Tilborg et al., 2022); LoHi (Steshin, 2023); FS-Mol (Stanley et al., 2021).

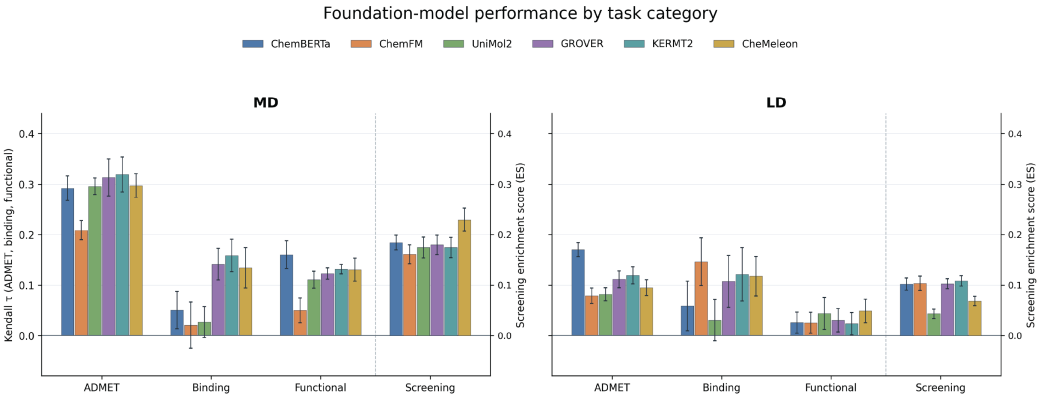

Figure 3: Performance of the six foundation models averaged within the 4 categories of tasks, on MD and LD. We observe a drop in performance on the LD tasks.

3 Application of MoG-Bench to Leading Foundation Models

We evaluate six state-of-the-art foundation models, **ChemBERTa-3** (Singh et al., 2026), **ChemFM** (Cai et al., 2025), **Uni-Mol2** (Ji et al., 2024), **GROVER** (Rong et al., 2020), **KERMT2** (Adrian et al., 2025), and **CheMeleon** (Burns et al., 2025), on MoG-Bench-MD and MoG-Bench-LD. We perform frozen embedding and finetuning evaluation for each model. In addition, we benchmark Mol-JEPA (Rottach et al., 2026) and CLAMP (Seidl et al., 2023); however, we note that these methods have been pretrained with labels on ChEMBL and PubChem, so we exclude them from comparisons here. Check the Appendix for their results. In order to facilitate fair comparisons in the future, we request that developers hold out the ChEMBL and PubChem assays included in the benchmark.

For frozen evaluation, we freeze the model weights and extract embeddings. We then train one of four probes–ridge linear/logistic regression, random forest (RF), XGBoost (XGB), or a multi-layer perceptron (MLP)–using the frozen embeddings. We perform hyperparameter tuning using "inner" validation folds, after which we train on the full training data and evaluate on the test data (an "outer" fold). Each fold is run with five different random seeds. Model performance on a task is aggregated across all that task's folds. We also run finetuning evaluations, using the finetuning recipes detailed in the models' documentation. Model performance using linear probes is shown in Firure 3.

Further details on benchmarking setup can be found in the Appendix. We present practical guidance and directions for future progress in the following sections.

3.1 Linear probes outperform Random Forests, XGBoost, and MLP probes

Figure 4 shows performance of the four probes on Medium Data and Low Data. On Medium Data, linear probes perform the best out of the four probes, followed by RF and MLP (essentially tied), and XGBoost. On Low Data, random forests nearly reach the performance of linear probes, implying that random forests can recover some of the nonlinear signal with few training compounds. MLP performance significantly worsens. Even with strong regularization and dropout, MLPs are likely overfitting to the small number of training compounds.

3.2 Linear probes perform better than finetuning

Linear probes also performed better than finetuning of these models. The training sets in MoG-Bench are quite small by foundation model standards (the largest Medium Data train set is only 5,000 compounds, and Low Data is only 24 compounds). These relatively small finetuning sets can cause foundation models to overfit to the compounds and perform poorly on out-of-distribution data (Kumar et al., 2022; Zhou et al., 2024). In contrast, linear heads are robust under limited data (Hoerl & Kennard, 1970; Bertinetto et al., 2019). Based on these results, we recommend that practitioners freeze the raw embeddings and train linear probes to extract the best performance out of these models.

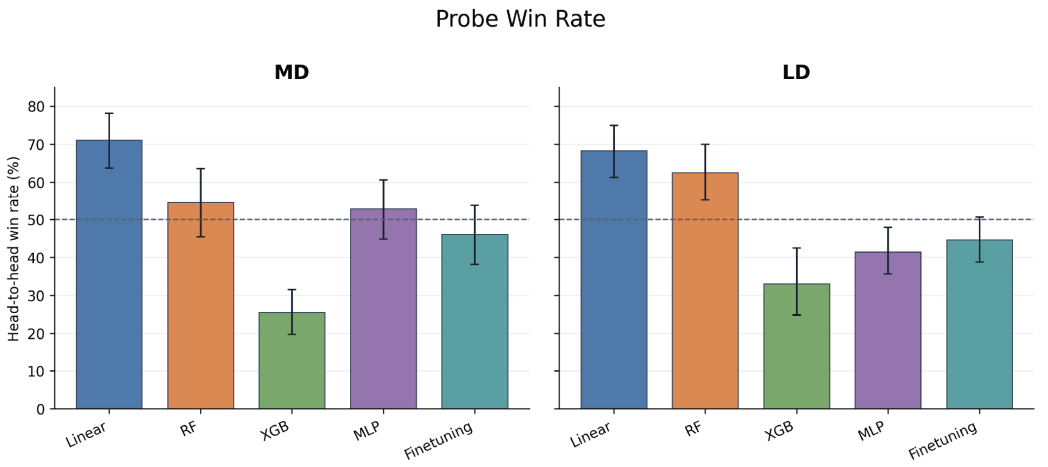

Figure 4: Performance of the four probes on MD and LD. Regardless of task or MD/LD, linear probes outperform the three other heads, and also outperform finetuning the models directly.

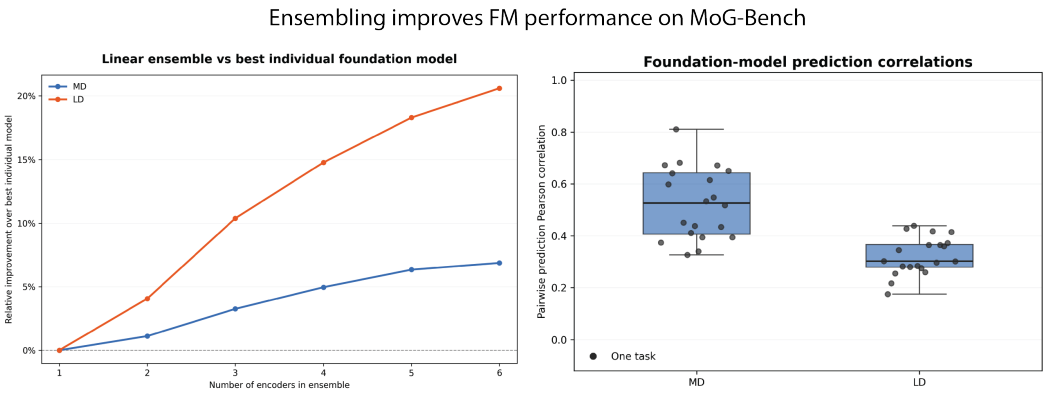

Figure 5: Panel A: we average predictions from 2, 3, 4, 5, and 6 foundation models using linear heads and see gains in performance relative to only using one foundation model. Panel B: In general, foundation model predictions tend to have moderate correlation, and slightly lower on LD. This implies the performance gains from ensembling, as the models learn differing but somewhat complementary patterns in the data.

3.3 Ensembling predictions boosts performance

We explored different ways to improve performance on MoG-Bench. In Figure 5A we investigate ensembling methods' predictions together. For this experiment, we take multiple foundation models' predictions with linear probes and average their individual molecule predictions together. We find a striking improvement in performance when ensembling. Ensembling yields large relative improvements in the Low Data regime, with a 20% gain over one foundation model when ensembling eight models. We suggest a possible reason for this success in Figure 5B. Within a task's folds, we compare the six models' predictions and calculate a Pearson correlation, which is then averaged across the folds. We find that the models show moderate correlation, implying that they share some similar learnings from pretraining, while still learning different things. LD tasks see a lower correlation, perhaps showing why LD gains more by ensembling.

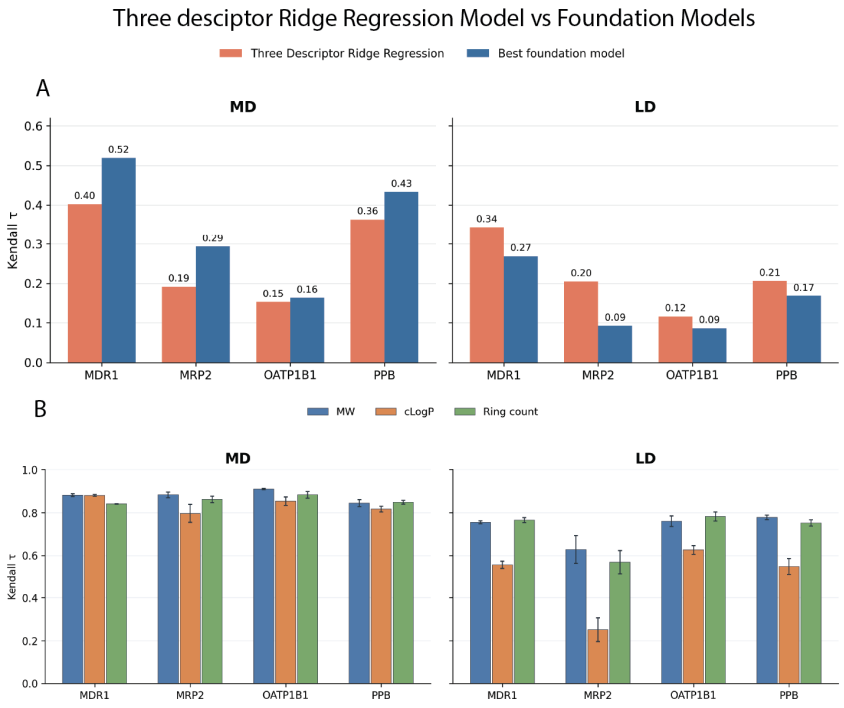

Figure 6: Panel A: We train a ridge regression model on three molecular descriptors on four MoG-Bench datasets and compare to foundation models on MD and LD. On MD, the average foundation model matches or outperforms the simple baseline. On LD, the ridge model outperforms the foundation models. Panel B: We train a ridge regression model on the four datasets to predict the three descriptors, and report Kendall tau. On MD and LD, models recover these descriptors.

3.4 Models perform well on Medium Data but need improvement on Low Data

We sought out to understand the performance of the foundation models on Medium versus Low Data regimes. We observe a drop in performance from Medium Data to Low Data (see Appendix); however, this is to be expected, as making predictions with less training data is much more difficult for models. All units are in Kendall τ correlation.

To better understand this accuracy drop, we trained a ridge regression model, labeled 3Ridge, using three molecular descriptors: molecular weight, LogP, and number of rings. We compare 3Ridge to foundation models with linear heads on four tasks: MDR1, MRP2, OATP1B1, and PPB (Figure 6A).

On Medium Data, the average foundation model outperforms 3Ridge. On Low Data, the models are unable to reach 3Ridge's performance. In Figure 6B, we show that these three descriptors are linearly accessible from the pretrained embeddings. We train a linear probe to predict the three descriptors from the embeddings, and show strong Kendall tau correlation for all three descriptors on both MD and LD. Descriptor information is contained within the embeddings, but is not easily organized for recovery by a probe with limited supervision. We identify this as a future direction for improvement.

4 Conclusion

We present MoG-Bench, a curated set of 20 tasks spanning 4 categories for molecular property prediction. We highlight how MoG-Bench contributes significant benchmarking and methodological improvements for the evaluation of CFMs. While current CFMs are promising, we identify generalization to robust, real world evaluation settings as the next frontier for model performance, and hope our efforts help inform future model development efforts.

REFERENCES

Mathew Abraham, K. Gagaring, Maria L. Martino, et al. Probing the open global health chemical diversity library for multistage-active starting points for next-generation antimalarials. *ACS Infectious Diseases*, 2020. doi: 10.1021/acsinfecdis.9b00482.

Matthew Adrian, Yunsie Chung, Kevin Boyd, Saee Paliwal, Srimukh Prasad Veccham, and Alan C Cheng. Multitask finetuning and acceleration of chemical pretrained models for small molecule drug property prediction. *arXiv preprint arXiv:2510.12719*, 2025.

Michael C. Alley et al. Feasibility of drug screening with panels of human tumor cell lines using a microculture tetrazolium assay. *Cancer Research*, 1988.

Stephan Arndt, Carolyn Turvey, and Nancy C Andreasen. Correlating and predicting psychiatric symptom ratings: Spearmans r versus kendalls tau correlation. *Journal of psychiatric research*, 33 (2):97–104, 1999.

Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.08136.

Jackson Burns, Akshat Zalte, and William Green. Descriptor-based foundation models for molecular property prediction. *arXiv preprint arXiv:2506.15792*, 2025.

Feiyang Cai, Katelin Zacour, Tianyu Zhu, Tzuen-Rong Tzeng, Yongping Duan, Ling Liu, Srikanth Pilla, Gang Li, and Feng Luo. Chemfm as a scaling law guided foundation model pre-trained on informative chemicals. *Communications Chemistry*, 2025.

Lin Cai and Miriam Gochin. A novel fluorescence intensity screening assay identifies new low-molecular-weight inhibitors of the gp41 coiled-coil domain of human immunodeficiency virus type 1. *Antimicrobial Agents and Chemotherapy*, 2007. doi: 10.1128/AAC.00150-07.

M. Castellanos, H. MacDermott-Opeskin, J. Ainsley, and P. Walters. Lessons learned from the openadmet–expansionrx blind challenge. Zenodo, 2026.

Jon Chambers, Mark Davies, Anna Gaulton, Anne Hersey, Sameer Velankar, Robert Petryszak, Janna Hastings, Louisa Bellis, Shaun McGlinchey, and John P Overington. Unichem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of cheminformatics*, 5(1):3, 2013.

ChEMBL. Drugmatrix in vitro pharmacology data. ChEMBL Document CHEMBL1909046, 2024. URL https://www.ebi.ac.uk/chembl/explore/document/CHEMBL1909046.

Robert D. Clark and Daniel J. Webster-Clark. Managing bias in roc curves. *Journal of Computer-Aided Molecular Design*, 22(3–4):141–146, 2008. doi: 10.1007/s10822-008-9181-z.

Ryan G. Coleman, Michael Carchia, Teague Sterling, John J. Irwin, and Brian K. Shoichet. Ligand pose and orientation transfer for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 53(3):647–659, 2013. doi: 10.1021/ci3005747.

Christophe Croux and Catherine Dehon. Influence functions of the spearman and kendall correlation measures. *Statistical methods & applications*, 19(4):497–515, 2010.

John G. Cumming, Andrew M. Davis, Sorel Muresan, Markus Haeberlein, and Hongming Chen. Chemical predictive modelling to improve compound quality. *Nature Reviews Drug Discovery*, 12: 948–962, 2013. doi: 10.1038/nrd4128.

Tom De Bruyn, Gerard Jan van Westen, Adriaan P. Ijzerman, et al. Structure-based identification of oatp1b1/3 inhibitors. *Molecular Pharmacology*, 2013. doi: 10.1124/mol.112.084152.

Jianyuan Deng, Zhibo Yang, Hehe Wang, Iwao Ojima, Dimitris Samaras, and Fusheng Wang. A systematic study of key elements underlying molecular property prediction. *Nature Communications*, 14(1):6395, 2023.

EUbOPEN. Gpcr results for eubopen chemogenomics library 3. ChEMBL Document CHEMBL5209801, 2023. URL https://www.ebi.ac.uk/chembl/explore/document/CHEMBL5209801.

EUbOPEN. Eubopen chemogenomics library – pdsp secondary binding. ChEMBL Document CHEMBL5442175, 2024. URL https://www.ebi.ac.uk/chembl/explore/document/CHEMBL5442175.

Cheng Fang, Ye Wang, Richard Grater, Sudarshan Kapadnis, Cheryl Black, Patrick Trapa, and Simone Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4:127–134, 2022. doi: 10.1038/s42256-021-00438-4.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://doi.org/10.1080/00401706.1970.10488634.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

John J Irwin, Khanh G Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R Wong, Munkhzul Khurelbaatar, Yurii S Moroz, John Mayfield, and Roger A Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of chemical information and modeling*, 60(12):6065–6073, 2020.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, et al. Uni-mol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*, 2024.

Roman Joeres, David B Blumenthal, and Olga V Kalinina. Data splitting to avoid information leakage with datasail. *Nature communications*, 16(1):3337, 2025.

Maurice G Kendall et al. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

T. Ketas and J. P. Moore. Mmv malaria box hiv and cytotoxicity screening. ChEMBL Document CHEMBL3301472, 2014.

Stephanie Klaeger, Stefan Heinzlmeir, Michael Wilhelm, et al. The target landscape of clinical kinase drugs. *Science*, 358(6367), 2017. doi: 10.1126/science.aan4368.

Ian Scott Knight, Slava Naprienko, and John J Irwin. Enrichment score: a better quantitative metric for evaluating the enrichment capacity of molecular docking models. *arXiv preprint arXiv:2210.10905*, 2022.

Christian Kramer, Tuomo Kalliokoski, Peter Gedeck, and Anna Vulpetti. The experimental uncertainty of heterogeneous public K_i data. *Journal of Medicinal Chemistry*, 55(11):5165–5173, 2012. doi: 10.1021/jm300131x.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Gregory A Landrum and Sereina Riniker. Combining ic50 or k i values from different sources is a source of significant noise. *Journal of chemical information and modeling*, 64(5):1560–1567, 2024.

Yunchao Liu, Ha Dong, Xin Wang, Rocco Moretti, Yu Wang, Zhaoqian Su, Jiawei Gu, Bobby Bodenheimer, Charles D Weaver, Jens Meiler, et al. Welqrate: defining the gold standard in small molecule drug discovery benchmarking. *Advances in Neural Information Processing Systems*, 37: 53222–53236, 2024.

Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tetiana Chepurnyna, Nastassia V. Kogoi, Artas A. Rustamov, Ying S. Huang, Allison Z. Zhou, Keith R. Roy, Michael D. Blum, Fangyu Ding, Jian Zou, Tanja Drača, Matthew R. Bradley, Qi Chen, Bradley J. Roth, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019. doi: 10.1038/s41586-019-0917-9.

Michael M Mysinger and Brian K Shoichet. Rapid context-dependent ligand desolvation in molecular docking. *Journal of chemical information and modeling*, 50(9):1561–1573, 2010.

Michael M. Mysinger, Michael Carchia, John J. Irwin, and Brian K. Shoichet. Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594, 2012. doi: 10.1021/jm300687e.

Joseph M. Pagano, Brian M. Farley, Laura M. McCoig, et al. Molecular basis of rna recognition by the embryonic polarity determinant mex-5. *Journal of Biological Chemistry*, 2007. doi: 10.1074/jbc.M700079200.

Jan M. Pedersen, Pär Matsson, Christel A. Bergström, et al. Prediction and identification of drug interactions with the human atp-binding cassette transporter multidrug-resistance associated protein 2. *Journal of Medicinal Chemistry*, 2008. doi: 10.1021/jm7015683.

Mateusz Praski, Jakub Adamczyk, and Wojciech Czech. Benchmarking pretrained molecular embedding models for molecular representation learning. *arXiv preprint arXiv:2508.06199*, 2025.

PubChem. Nci human tumor cell line growth inhibition assay: A549/atcc non-small cell lung cell line. PubChem BioAssay AID 19, a. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/19.

PubChem. Fluorescent polarization homogeneous dose response hts to identify inhibitors of mex-5 binding to tcr-2. PubChem BioAssay AID 1960, b. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/1960.

PubChem. Ubc13 polyubiquitin inhibitors measured by tr-fret. PubChem BioAssay AID 493155, c. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/493155.

PubChem. Dose response for hts for beta-2ar agonists via fap method from cp1. PubChem BioAssay AID 588763, d. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/588763.

PubChem. Identification of caenorhabditis elegans skn-1 inhibitors by fluorescence-ratio assay. PubChem BioAssay AID 624474, e. URL https://pubchem.ncbi.nlm.nih.gov/bioassay/624474.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

Florian Rottach, Sebastian Schieferdecker, William Rudman, Randall Balestriero, and Carsten Eickhoff. Mol-jepa: A multimodal joint embedding predictive architecture for molecules. *arXiv preprint arXiv:2608.22642*, 2026.

Carl C. G. Schiebroek, Gregory A. Landrum, and Sereina Riniker. Balancing data quantity and quality: Evaluating curation strategies for bioactivity prediction in lead optimization. *Journal of Chemical Information and Modeling*, 66(13):7446–7452, 2026. doi: 10.1021/acs.jcim.6c01018.

Luciana Scotti and Marcus Tullius Scotti. Adme properties in drug delivery, 2025.

Philipp Seidl, Andreu Vall, Sepp Hochreiter, and Günter Klambauer. Enhancing activity prediction models in drug discovery with the ability to understand human language. In *International conference on machine learning*, pp. 30458–30490. PMLR, 2023.

Riya Singh, Aryan Amit Barsainyan, Rida Irfan, Connor Joseph Amorin, Stewart He, Tony Davis, Arun Thiagarajan, Shiva Sankaran, Seyone Chithrananda, Walid Ahmad, et al. Chemberta-3: an open source training framework for chemical foundation models. *Digital Discovery*, 2026.

Megan Stanley, John Bronskill, Krzysztof Maziarz, Hubert Misztela, Jessica Lanini, Marwin Segler, Nadine Schneider, and Marc Brockschmidt. FS-Mol: A few-shot learning dataset of molecules. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/8d3bba7425e7c98c50f52ca1b52d3735-Abstract-round2.html.

Reed M. Stein, Yekaterina Yang, Trent E. Balius, Maureen J. O'Meara, Jiankun Lyu, Jessica Young, Kimberly Tang, Brian K. Shoichet, and John J. Irwin. Property-unmatched decoys in docking benchmarks. *Journal of Chemical Information and Modeling*, 61(2):699–714, 2021. doi: 10.1021/acs.jcim.0c00598.

Teague Sterling and John J. Irwin. Zinc 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. doi: 10.1021/acs.jcim.5b00559.

Simon Steshin. Lo-hi: Practical ml drug discovery benchmark. *Advances in Neural Information Processing Systems*, 36:64526–64554, 2023.

Prudencio Tossou, Cas Wognum, Michael Craig, Hadrien Mary, and Emmanuel Noutahi. Real-world molecular out-of-distribution: Specification and investigation. *Journal of Chemical Information and Modeling*, 64(3):697, 2024.

Han van de Waterbeemd and Eric Gifford. ADMET in silico modelling: towards prediction paradise? *Nature Reviews Drug Discovery*, 2:192–204, 2003. doi: 10.1038/nrd1032.

Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23): 5938–5951, 2022.

W Patrick Walters. We need better benchmarks for machine learning in drug discovery. 2023.

Jun Wang, Peng Fang, Peter Chase, et al. Development of an hts-compatible assay for discovery of melanoma-related microphthalmia transcription factor disruptors using alphascreen technology. *SLAS Discovery*, 2017. doi: 10.1177/1087057116675274.

M. Wenlock and N. Tomkinson. Experimental in vitro dmpk and physicochemical data on a set of publicly disclosed compounds. ChEMBL Document CHEMBL3301361, 2014.

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. *Chemical science*, 9(2):513–530, 2018.

Ziyi Zhou, Liang Zhang, Yuanxi Yu, Banghao Wu, Mingchen Li, Liang Hong, and Pan Tan. Enhancing efficiency of protein language models with minimal wet-lab data through few-shot learning. *Nature Communications*, 15:5566, 2024. doi: 10.1038/s41467-024-49798-6. URL https://www.nature.com/articles/s41467-024-49798-6.

5 APPENDIX

5.1 DATA QUALITY CONTROL

We complete the following additional steps in order to computationally filter and prepare our datasets.

1. Canonicalize all SMILES strings, and drop any molecules that fail RDKit parsing - this is done to prevent invalid molecules from contaminating the train/test sets. Canonicalization yields one unique string per structure so downstream deduplication operations are well-defined.
2. Remove small artifacts disconnected from the larger molecule. Salts, counterions, and solvents may be present in the raw assay data or result from initial data processing steps. Removing these artifacts yields the actual drug-like molecule that is likely actually being measured in the assay and nothing else.
3. Deduplicate molecules. If a molecule appears more than once in the data, we delete all occurrences of it. This prevents molecules with differing labels due to errors in data entry from contaminating the training data.
4. Filter out any molecules that weigh fewer than 150 Daltons or greater than 1000 Daltons. The 150–1000 Da range covers nearly all drug-like molecules, and prevents polypeptides from leaking into the data.

5.2 DESCRIPTION OF BENCHMARKED MODELS

We used MoG-Bench to evaluate both state-of-the-art chemical foundation models and widely used simple baseline models, as described below.

ChemBERTa-3 (Singh et al., 2026) The latest iteration in the ChemBERTa family of models, which treats molecules as a stream of tokens based on the SMILES string for each compound. This model is trained using a masked-language-modeling objective, using SMILES strings from the ZINC20 database (Irwin et al., 2020).

ChemFM (Cai et al., 2025) A foundation model trained on recovering masked SMILES strings, similar to ChemBERTa-3. ChemFM is trained on the smaller yet more diverse UniChem (Chambers et al., 2013) dataset, and is three times bigger than ChemBERTa-3.

CheMeleon (Burns et al., 2025) A foundation model built on top of Chemprop that has been pre-trained to predict Mordred fingerpints through message passing.

5.3 OTHER ADVANCED SPLITTING ALGORITHMS

DataSAIL (Joeres et al., 2025) DataSAIL optimizes an ILP that encourages the lowest possible average Tanimoto similarity between all pairs across train and test; it does not place any compounds into holdout. As a result, it yields the largest dataset sizes of the advanced splitters (along with Maximum Dissimilarity), but data leakage still exists. Because no compounds are held out, there can often be cases of highly similar datasets where compounds with Tanimoto ≥ 0.70 can end up across train and test

Maximum Dissimilarity (MD) (Tossou et al., 2024) MD sorts pairs by dissimilarity, and then iteratively partitions them into train and test. Similar to DataSAIL, it does not place any compounds into holdout, and has similar successes and failures. Every molecule is retained; however, there is still leakage. To truly ensure no train-test leakage, some molecules must be placed into holdout.

Hi (Steshin, 2023) The Hi splitter has a strict Tanimoto cutoff at 0.4, ensuring that no molecules across train and test violates this boundary. Hi does this by solving an ILP that respects this constraint while also minimizing the total number of holdout compounds. While the Hi splitter places compounds in test, it has stricter leakage rules compared to DataSAIL and MD.

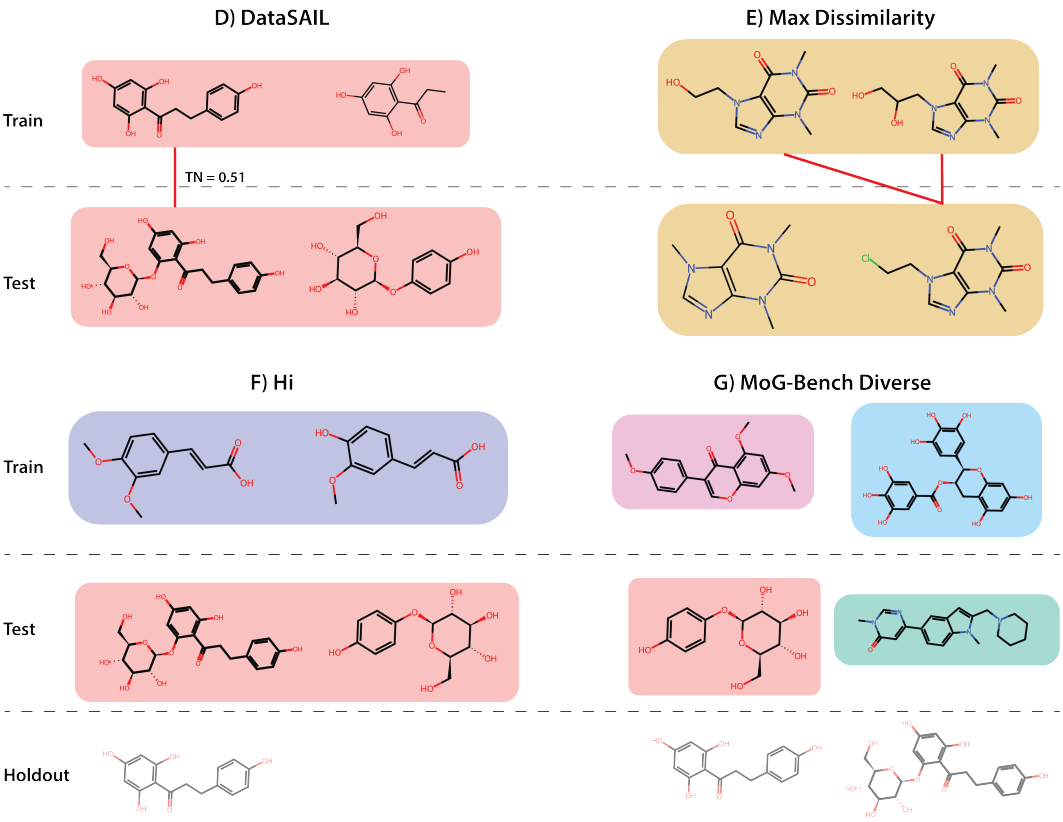

Figure 7: Comparisons between MBDSplitter and DataSAIL, Hi, and Maximum Dissimilarity Splitters. DataSAIL, Hi, Maximum Dissimilarity (MD), and MoG-Bench Diverse (MBD) Splitters yield more dissimilar molecules compared to Random and BMS. Despite this constraint, train and test folds are not always diverse (lower panel). Colored boxes correspond to molecules being in the same cluster. DataSAIL and MD do not constrain the diversity of train or test, so models can be tested on series of highly similar molecules (A and B). In addition, since neither utilizes holdout, there is still some data leakage. Hi splitter enforces a strict Tanimoto cutoff, but suffers from the same issue: no test set diversity is enforced, leading to models being evaluated on highly similar chemical series. The MBD split enforces a test set diversity, ensuring that compounds being tested on are diverse, and not just part of a single cluster.

5.4 MBDSplitter

Let a test set S be comprised of many clusters s_k. A molecule is in a cluster s_i if all compounds are connected by an edge or set of edges; an edge between two compounds is drawn iff they have greater than or equal to 0.40 Tanimoto of their Morgan fingerprints.

$$s_1, \ldots, s_K, \quad n = |S| = \sum_{k=1}^{K} s_k$$

We define the Largest Cluster Percentage (LCP) and the normalized Shannon entropy ($\widehat{H}$) of S.

$$\mathrm{LCP}(S) = \frac{\max_k s_k}{n}$$

$$\widehat{H}(S) = 1 - \frac{\sum_{k=1}^{K} s_k \log s_k}{n \log n}$$

The LCP is the fraction of the total test set that the largest cluster occupies; $\widehat{H}$ is the normalized cluster Shannon entropy: a value close to 0 suggests that all compounds come from one cluster, and a value of 1 suggests that every compound belongs to its own cluster. We add these constraints to our MILP to generate our data splits. We set the maximum LCP of any test split to 30%, and attempt to maximize $\widehat{H}$ while not losing too many compounds to holdout. We provide more detail in the Appendix.

Appendix Figure 7 shows a visual depiction of the differences between MBDSplitter versus DataSAIL, MD, and Hi.

5.5 Other Molecular Benchmarks

QM9 (Ramakrishnan et al., 2014) Although often used to benchmark machine learning models for drug discovery and chemical foundation models, this benchmark focuses on computationally calculated quantum mechanical properties. These properties are not directly relevant to drug discovery, in part because they are properties of individual conformations of molecules that can adopt many conformations. Instead, MoG-Bench uses experimentally determined datasets of molecular properties that are important in drug discovery.

MoleculeNet (Wu et al., 2018) MoleculeNet provides consolidated access to a collection of datasets, including QM9. Some of these tasks are relevant to drug discovery, while others are not. MoleculeNet datasets have errors including invalid chemical structures, inconsistent representations, and unclear endpoints (Deng et al., 2023; Walters, 2023) that we check for and remove in MoG-Bench. MoleculeNet employs hierarchical, random, and scaffold-based splits (Deng et al., 2023; Walters, 2023), all of which sometimes lead to inclusion of similar molecules in the train and tests sets. See our earlier discussion on data splits.

Therapeautic Data Commons (TDC) (Huang et al., 2021) Similarly to MoleculeNet, TDC also offers access to a collection of various datasets. TDC is impacted by similar data quality issues and benchmarking challenges. Also, similarly to MoleculeNet, it employs a training and test split methodology that allows for greater leakage than that used by MoG-Bench.

WelQRate (Liu et al., 2024) WelQRate is a curated dataset of high-throughput screening data that has been carefully processed and cleaned so that all chemical structures are valid, and all targets for the screens are therapeutically relevant. While we used some of the same criteria to curate our datasets, we report different metrics and focus on a training and test split methodology that minimizes leakage compared to WelQRate, which uses a scaffold split. MoG-Bench also includes much more diverse datasets reflecting a wide variety of chemical properties of interest.

MoleculeACE (Van Tilborg et al., 2022) This benchmark focuses on activity cliffs, which are cases where molecules that are structurally very similar to each other have very different molecular properties, like binding affinity. MoG-Bench focuses instead on generalization, and making predictions for molecules with chemical structures substantially different from training compounds. MoG-Bench also includes much more diverse datasets.

LoHi (Steshin, 2023) This benchmark focuses on lead optimization and hit identification. LoHi benchmarks models on their ability to distinguish actives from a large pool of compounds (hit identification) and their ability to select top binders amongst a set of promising candidates. This benchmark also utilizes a difficult train/test split; however, data is curated from ChEMBL across multiple different assays, which can lead to batch effects.

FS-Mol (Stanley et al., 2021) FS-Mol tests a model's few-shot learning performance, namely if information from previous assays can help predict activities in new assays with limited information.