KERMT: Improving Graph Foundation Models for Drug Discovery through Multitask Fine-Tuning, Data Scaling, and Accelerated Training

Matthew Adrian [1] **Yunsie Chung** [1] **Kevin Boyd** [2] **Saee Paliwal** [2] **Srimukh Prasad Veccham** [2] **Alan C. Cheng** [1]

Abstract

Chemical foundation models hold significant potential to improve prediction of critical properties in small molecule drug discovery by leveraging representations learned from self-supervised pretraining. We implement multitask fine-tuning in two pretrained graph transformer models, KPGT and KERMT (Kinetic grovER MultiTask), our enhanced version of the GROVER model. We benchmark the models on diverse on-target biology assay and off-target ADMET (absorption, distribution, metabolism, excretion, and toxicity) prediction tasks, and show that KERMT has improved predictive performance for ADMET endpoints with 2,000 or more training examples. Contrary to popular hypotheses that pretrained models show better performance in data-scarce regimes, KERMT improves most as dataset size increases. KERMT generalizes better than Chemprop, a graph neural network model that generally leads in the field, across all compound similarity bins and additional pretraining on chemical structures more similar to the downstream structures does not improve performance. To support future work, we release two multitask ADMET splits and an accelerated implementation of KERMT for scalable pretraining, fine-tuning, and inference.

1. Introduction

Predictive models are central to multi-property optimization in molecular drug design and drug discovery. While supervised deep learning architectures, particularly graph neural networks (GNNs), have improved predictive accuracy over traditional machine learning models by learning latent molecular representations directly from chemical graphs, their real world performance is constrained by amount of available labeled data, experimental variation in data, and data distribution shifts. (Cáceres et al., 2020; Chen et al., 2025)

Recent developments in chemical foundation models aim to address these limitations by leveraging large-scale, self-supervised pretraining on unlabeled molecular structures (Ross et al., 2022; Li et al., 2023; Rong et al., 2020). These models fall into two broad architectural categories: language-based models such as MolFormer (Ross et al., 2022) which process SMILES strings using standard NLP transformers and graph-based models which incorporate strong chemical inductive biases by treating atoms as nodes and bonds as edges (Duvenaud et al., 2015; Rong et al., 2020; Yang et al., 2019; Feinberg et al., 2020; Li et al., 2023). While graph-based approaches are considered more expressive for molecular topologies, scaling them for greater downstream generalization remains a challenge.

Simultaneously, multitask learning (MTL) has emerged as a powerful paradigm to improve model generalization by exploiting inductive transfer between tasks (Caruana, 1997). MTL has been extensively used in GNNs trained from scratch but has not been as widely applied to graph foundation models.

Here, we introduce KERMT (Kinetic grovER MultiTask) and show that multitask fine-tuning significantly amplifies the performance of pretrained graph transformers and results in overall improved predictive performance. KERMT is an enhanced and extended implementation of the GROVER model (Rong et al., 2020). Through extensive benchmarking on diverse datasets including on-target assay and ADMET (absorption, distribution, metabolism, excretion, and toxicity) data, we show that KERMT challenges the common hypothesis that pretraining strictly benefits low-data regimes. Instead, KERMT scales highly efficiently, outperforming state-of-the-art baselines in large-data scenarios. To enable practical deployment, we implement distributed-GPU training and accelerated graph featurization, reducing memory overhead and significantly improving throughput.

Our main contributions are as follows:

[1]Merck & Co., Inc., South San Francisco, CA, USA [2]BioNeMo, NVIDIA, Santa Clara, CA, USA. Correspondence to: Alan Cheng <alan.cheng@merck.com>, Srimukh Veccham <sveccham@nvidia.com>.

Workshop on Graph Foundation Models, *43rd International Conference on Machine Learning (ICML 2026)*, Seoul, South Korea.
Copyright 2026 by the author(s).

1. We introduce KERMT (Kinetic grovER MultiTask), a graph transformer with multitask fine-tuning and engineering improvements in pretraining and inference to make it practical for regular use in drug discovery settings.
2. We show that multitask fine-tuning significantly improves KERMT predictive performance over both single-task pretrained models and multitask non-pretrained GNNs.
3. Contrary to prevailing hypotheses, we find that KERMT's advantage over non-pretrained baselines grows with dataset size. For endpoints with over 2,000 training examples, KERMT shows equivalent or better performance in all cases. This suggests that larger pretrained models can benefit from the richer supervision signal of large labeled datasets during fine-tuning.
4. We provide scaling analyses over both dataset size and number of tasks, showing that KERMT's performance improves monotonically with both, while non-pretrained GNNs plateau.
5. We release KERMT as a practical, open-source tool at `https://github.com/NVIDIA-BioNeMo/KERMT`. We additionally release two new multitask ADMET benchmark datasets to enable standardized evaluation of future multitask models.

2. Methods

2.1. Datasets

Public and Private Datasets. We trained and evaluated all models on a combination of public and proprietary datasets.

For public data, we aggregated 25 ADMET endpoints from the literature (Gaulton et al., 2017), comprising over 114,000 unique molecules. This public dataset was partitioned using a K-means cluster split based on Morgan fingerprints to simulate prediction on novel chemical scaffolds. We also created a similar cluster split for a published ADMET dataset from Biogen (Fang et al., 2023) and benchmarked on public kinase potency datasets curated from BindingDB and other literature sources (Liu et al., 2024).

Our proprietary dataset, collected from Merck & Co., Inc. (Rahway, NJ, USA; known as MSD outside the United States and Canada), spans 30 key ADMET endpoints for over 800,000 unique molecules, including small molecules and cyclic peptides. For evaluation, we created two temporal splits: an "80-20 split" and a "1-year split", where only the test set contains molecules synthesized after April 2018 and January 2023, respectively. Single-task datasets were derived directly from these multitask splits to ensure consistent train/test sets across all comparisons.

New Public Multitask ADMET Benchmark and Downsampling. Recognizing the lack of suitable public benchmarks for multitask evaluation, we developed and are releasing two new benchmarks, a large multitask-appropriate ADMET benchmark ($\sim$120k molecules, 25 endpoints) with a cluster split and a multitask cluster split for the Biogen ADMET dataset. To facilitate experimental reproducibility and enable community validation of multitask models, we have made available the public datasets along with their corresponding train/validation/test splits. These resources are accessible through Figshare for download and use.

For our model scaling experiments, we systematically downsampled the proprietary ADMET training dataset by data size or number of tasks. Data size was reduced by varying percentages ranging from 0.2% to 10%, while the number of tasks was reduced from 30 down to 25, 16, or five representative endpoints. In all downsampling studies, the test set remained unchanged to ensure fair comparison.

2.2. Models and Evaluation

Model Architectures. We benchmarked three diverse pretrained models: KERMT, KPGT, and MoLFormer, alongside Chemprop (Heid et al., 2024; Graff et al., 2026), a D-MPNN architecture trained from scratch and considered a leading GNN model architecture for ADMET prediction (Chen et al., 2025; Castellanos et al., 2026). KERMT was pretrained on 11 million compounds using atom/bond and functional group prediction tasks. KPGT (Li et al., 2023) is a line graph transformer pretrained on 2 million molecules using a knowledge-guided strategy. MoLFormer is a SMILES-based transformer pretrained on nearly 100 million compounds using a masked language modeling objective.

KERMT, a graph transformer model based on the GROVER architecture, uses atom-based and bond-based message passing to learn meaningful chemical representation of molecules (Rong et al., 2020). The KERMT model is based on the GTransformer architecture and uses dynamic message passing neural networks to form building blocks of a transformer. The model was pretrained on two types of tasks: (1) node/edge-level classification tasks which identify the statistical properties of a k-hop local subgraph from the embedding of the node/edge and (2) graph-level motif classification which is a multi-label task that uses graph embeddings to identify the presence or absence of 85 predefined functional groups within the molecule. The model was pretrained on 11 million compounds from the ZINC15 (Sterling & Irwin, 2015) and ChEMBL (Gaulton et al., 2012). Unless otherwise specified, we evaluated this method using the published 48M parameter base model weights.

KPGT is a graph transformer that operates on a molecular line-graph representation. KPGT is pretrained using a knowledge-guided pretraining strategy involving three tasks: (1) prediction of masked/corrupted nodes and predic-

tion of masked K-node features, (2) prediction of RDKit fingerprints, and (3) prediction of 200 molecular descriptors generated with RDKit (Landrum, 2021). We used the published 100M parameter model checkpoint, pretrained on 2 million molecules from the ChEMBL29 dataset (Gaulton et al., 2017), for fine-tuning. For KPGT, we employed two fine-tuning schemes; the simple fine-tuning model that we refer to as "KPGT" was trained by simply allowing the encoder and FFN to update more freely, while the complex fine-tuning model that we refer to as "$\text{KPGT}_{\text{complex}}$" utilized the complex fine-tuning strategies presented in the original paper. These complex fine-tuning strategies include reinitialization of the last n layers of the encoder, layer-wise learning rate decay, and regularization constraining the fine-tuned weights to be close to the pretrained weights. The 'simple fine-tuning' KPGT results are shown in the main text while additional 'complex fine-tuning' $\text{KPGT}_{\text{complex}}$ results are available in the Appendix.

For KERMT and KPGT, we used the hyperparameter search package, Optuna (Akiba et al., 2019), to optimize hyperparameters on the single-task models trained using the internal data. These hyperparameters are reported in the Appendix. All parameters that are not described in the Appendix tables are set to their default values (Li et al., 2023; Rong et al., 2020). Only data excluding the test sets were used for hyperparameter optimization to prevent data leakage.

MoLFormer (Ross et al., 2022) is a SMILES-based language transformer model utilizing rotational positional embeddings. It uses a traditional masked language modeling objective for pretraining as introduced in BERT (Devlin et al., 2019). We evaluated this method using the public Huggingface model trained on 10% of the dataset that was used to train the XL version which was used for evaluation in their paper (Wolf et al., 2020). The full MoLFormer XL version is not available for general and commercial use. The 10% model was pretrained on nearly 100 million compounds from ZINC (Irwin & Shoichet, 2005) and PubChem (Kim et al., 2023) datasets. We fine-tuned the model based on optimized hyperparameters reported in the Appendix.

Model Fine-Tuning and Evaluation. We implemented multitask finetuning in KERMT and KPGT. We adapted the standard single-task fine-tuning process for multitask prediction by modifying the final feed-forward network to output predictions for all tasks simultaneously, while still updating the pretrained encoder and readout weights. Hyperparameters for each model were optimized using `Optuna` package (Akiba et al., 2019) on a subset of the proprietary training data. Model performance was primarily measured by the Pearson r^2 between predicted and experimental values on the test set. Final results are reported as the average of four independent training runs to ensure robustness.

Model Acceleration and Parallelization. To make KERMT practical for industrial-scale workflows, we implemented significant performance enhancements. The pretraining stage was parallelized using PyTorch's Distributed Data Parallel (DDP) scheme, enabling efficient scaling across multiple GPUs. For fine-tuning and inference, we accelerated the time-consuming molecular featurization step by integrating `cuik-molmaker`. This package enables batched, on-the-fly generation of molecular features directly on the GPU, which substantially reduced processing time and CPU memory usage. Our model acceleration and parallelization benchmarking was performed on a single node containing eight NVIDIA A100 GPUs and 16 AMD EPYC 7J13 CPUs

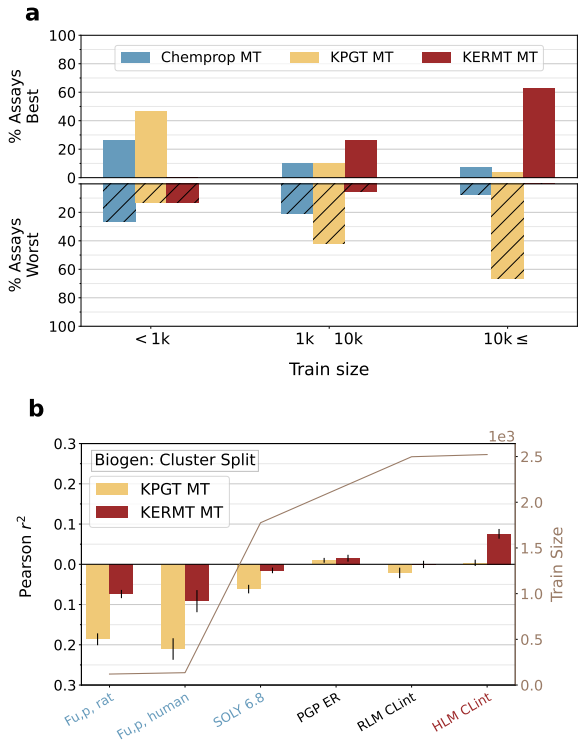


Figure 1. **ADMET assay performance comparison. (a)** Number of assays where each model statistically outperforms (N_{best}) or underperforms (N_{worst}), categorized by training data size. **(b)** Performance improvement of pretrained MT models relative to Chemprop MT on the cluster-split Biogen public dataset. The brown line indicates training data size, and colors denote the statistically superior model for each assay.

3. Results

3.1. Chemical foundation model benchmarking

We benchmark three diverse pretrained encoder approaches against the leading widely-used GNN standard, Chemprop (Yang et al., 2019; Castellanos et al., 2026). The pretrained models include two graph-based transformers, KPGT (Li et al., 2023) and KERMT, and a language-based model,

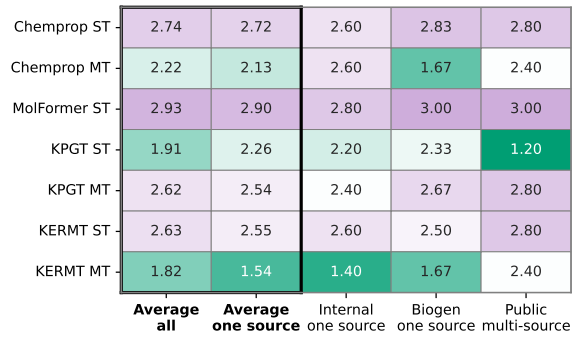


Figure 2. **ADMET model comparison summary using win rates.** A heatmap visualizing average scores (win=1, tie=2, loss=3) across 16 assays from three datasets. The columns compare performance on single-source datasets versus all datasets combined. See Appendix Figs. 3 and 4 for source data and calculations.

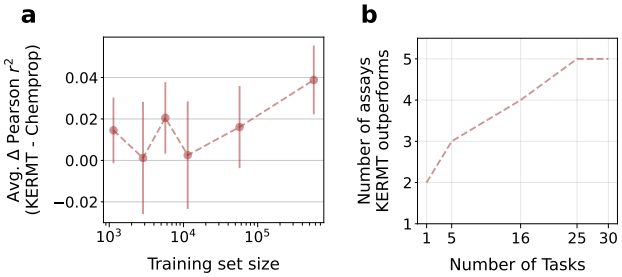


Figure 3. **Scaling of KERMT and Chemprop MT models with data size and task number.** All results are from the internal ADMET dataset with a temporal split. **(a)** Average performance improvement of KERMT relative to Chemprop MT as training set size increases. **(b)** Number of assays where KERMT outperforms Chemprop versus the number of multitask tasks.

MoLFormer (Ross et al., 2022). KERMT is an enhanced and extended implementation of the GROVER model(Rong et al., 2020). We assess the models as single-task (ST) models and then develop multitask (MT) fine-tuning for the two highest performers, KPGT-MT and KERMT.

Our evaluation utilizes several ADMET and on-target potency datasets. The ADMET data sets include a large internal proprietary dataset from Merck & Co., Inc. (Rahway, NJ, USA), a data set created from public sources, and a public Biogen dataset (see Methods for details). For on-target potency, we compiled 16 datasets from public sources and internal data. To ensure rigorous evaluation, we used temporal splits for the proprietary datasets—simulating real-world prospective prediction—and cluster-based splits for the public datasets.(Sheridan, 2013)

3.1.1. Performance on Industrial and Public ADMET Datasets

We use two single-source industrial ADMET benchmarks, an internal one from Merck & Co., Inc. (Rahway, NJ, USA) and an open-sourced one from Biogen. On the internal dataset, which provides a large-scale, single-source benchmark, all five graph-based models outperformed the language-based MoLFormer, underscoring the value of chemical graph structure as an inductive bias (Fig. 2. Among single task (ST) models, KERMT-ST and KPGT were the top performers. For KPGT, we found that 'simple fine-tuning' consistently outperforms 'complex fine-tuning' in our work (Appendix Tables 7-9) and consistently show the best KPGT results in the main text. KERMT was the clear winner, outperforming or tying for the best performance on all five representative assays we tested (Fig. 2 and Appendix Fig. 6). Multitasking provided a significant boost to KERMT's performance across the board, whereas KPGT saw little benefit, and even performance degradation on one assay. When the multitask models were benchmarked on 30 internal ADMET endpoints, KERMT was statistically superior in 18, and was the best model overall (Appendix Fig. 7).

On the Biogen dataset, KERMT performs equivalently or outperforms on datasets with at least 2,000 training examples. Overall, KERMT is a top model along with KERMT ST and Chemprop MT (Fig. 2) despite the smaller datasets in the Biogen dataset (2,500 or less) compared to the internal dataset.

On the curated public ADMET dataset that is overall smaller and aggregated from multiple different sources, both KERMT and Chemprop MT performed well. KPGT ST also showed strong performance, which may be attributed to its pretraining on ChEMBL29, a dataset with significant chemical space overlap with the public ADMET data. A key trend emerged from these public sets: for assays with at least 1,000 data points, KERMT consistently matched or exceeded the performance of Chemprop MT (Fig. 1) and consistently outperformed for assays with over 2,000 data points, similar to what we observed with the internal and Biogen data sets.

3.1.2. KERMT Outperforms on Large-Scale Datasets

A combined analysis across 16 diverse ADMET assays confirms that KERMT is the top-performing model, especially on large, consistent datasets from a single source (Fig. 2).

For on-target potency modeling, where datasets are typically smaller, the benefits of pretraining were more modest. Both multitask KPGT and KERMT offered performance improvements over Chemprop MT on datasets with more than 2,000 data points (Appendix Fig. 9), comparable to what was found for ADMET datasets.

3.2. Scaling Data and Tasks

To understand the relationship between data size and performance, we fine-tuned models on progressively smaller subsets of the internal ADMET data. We found that KERMT's performance advantage over Chemprop MT increases with dataset size (Fig. 3a). The performance gap widens significantly as data size grows, with KERMT showing clear superiority on datasets with over 60,000 total data points. This may be due to KERMT's much larger model capacity ($\sim$51M parameters vs. $\sim$5M for Chemprop), which allows it to better leverage large volumes of labeled data when learning. To support this argument, we generated a random initialized KERMT model. With fine-tuning on the Biogen dataset, the randomly initialized KERMT consistently underperforms for moderate to large datasets and performs equivalently to pre-trained KERMT for small datasets that have $<$500 training examples (Appendix Fig. 10).

We also found that KERMT's performance scales favorably with the number of tasks in a multitask model. While Chemprop's performance plateaus after about five tasks, KERMT continues to show improvement as more tasks are added (Figs. 3b and 11).

3.3. Chemical space generalization

To assess chemical space generalization, we constructed subsets of the test set with different Tanimoto similarity thresholds to the training set and compared Chemprop MT and KERMT. In principle, pretraining may improve generalization to molecules that are less similar to the training data. As expected, both models perform better as train–test similarity increases (Fig. 4a). Across all similarity regimes, KERMT consistently outperforms Chemprop MT (Fig. 4b–c).

3.4. Effect of pretraining on more similar data

We next explored how pretraining on internal molecules that are more chemically similar to the downstream tasks affects performance. We pretrained KERMT on approximately 800K internal structures, with training only performed with the training-set molecules, and found similar downstream results to the published base model on both the full test set and the cyclic peptide subset (Appendix Fig. 12). The base model still slightly outperformed the internally pretrained model on more tasks than vice versa, suggesting that greater chemical overlap alone is not sufficient; instead, better pretraining objectives or broader and more diverse pretraining corpora may be needed to improve transfer.

3.5. Improving the Speed and Efficiency of KERMT

A practical challenge for deploying large models like KERMT in drug discovery is their computational cost and resources needed for pretraining and training. To address this, we implemented two key optimizations. First, we parallelized the pretraining stage using PyTorch Distributed Data Parallel (DDP) and achieved a scaling efficiency of 86% on an eight-GPU node (Fig. 5a). Second, we integrated `cuik-molmaker` into the dataloader to accelerate the on-the-fly featurization of molecular graphs. This resulted in a 2.2x speedup in fine-tuning and a 2.9x speedup in inference, along with a 34% reduction in CPU memory usage (Fig. 5b). These accelerations make the regular updating and deployment of KERMT models feasible in an industrial setting.

4. Discussion and Limitations

In this work, we demonstrate that multitask fine-tuning of pretrained chemical models with KERMT, our enhanced and extended implementation of GROVER, yields performance improvement on datasets with at least 2,000 training examples, with increasing advantage for larger datasets. We introduced multitask capabilities in both KERMT and KPGT, and rigorously evaluated them against the widely used GNN standard, Chemprop (ST and MT), on a wide range of internal and public ADMET and potency datasets. To facilitate their practical use, we accelerated KERMT's algorithms and have open-sourced the code at `https://github.com/NVIDIA-BioNeMo/KERMT`.

Our findings challenge a common assumption that pretrained models primarily excel in low-data regimes. Instead, our results show that KERMT's advantage grows with the scale of the data (Fig. 1) and that KERMT's advantage over non-pretrained models like Chemprop is most pronounced in large-data regimes, outperforming in over 25% of tasks when trained with at least 2,000 training examples and outperforming in 60% of tasks when trained with over 10,000 training examples (Fig. 1). This makes it exceptionally well-suited for ADMET prediction, where large datasets are common. KPGT performs well in single-task and smaller-data settings, making it a good choice for on-target potency modeling and suggesting that different pretraining strategies favor different downstream settings. A key strength of KERMT is that it consistently performs as well as or better than Chemprop MT across nearly all scenarios.

In scaling the number of tasks in the multitask model, we found that KERMT is more effective than the non-pretrained GNN, Chemprop MT, at capturing inter-task correlations, likely benefiting from the general representations learned during pretraining. Our chemical space generalization analysis further supports KERMT's robustness. Across chemical similarity bins, KERMT outperformed Chemprop MT at all similarity levels, including less similar test compounds. This suggests that KERMT learns transferable representations rather than relying only on close chemical analogs.

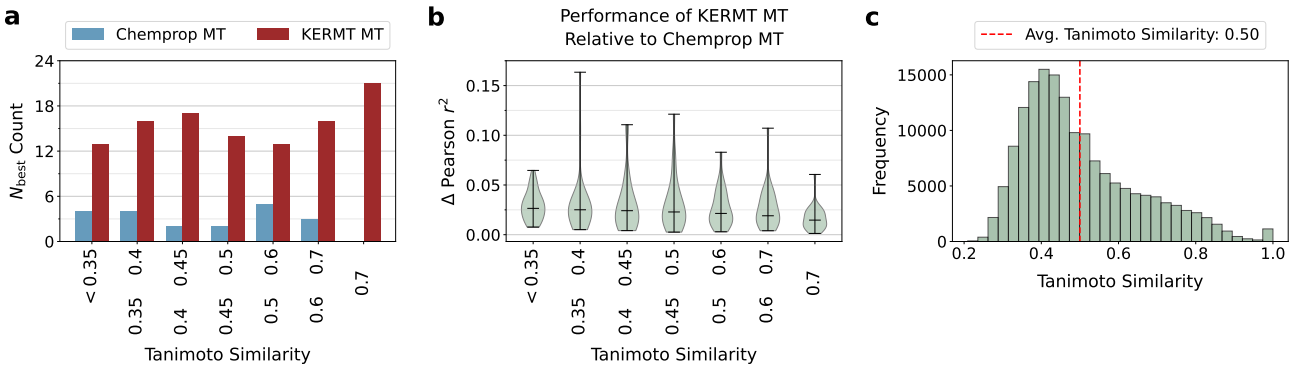


Figure 4. **Multitask model performance for Chemprop and KERMT across a range of Tanimoto similarity scores.** All results are evaluated on the internal ADMET dataset with an 80:20 temporal split. Tanimoto similarities were calculated using Morgan fingerprints (r=2) as implemented in RDKit (Landrum, 2021). **a** Number of assays on which each model statistically outperforms the other for each similarity range. **b** Violin plot showing the distribution of performance improvement for KERMT relative to Chemprop MT. The horizontal line in each violin indicates the median. In **a** and **b**, all 30 assays from the internal dataset are used. **c** Histogram of Tanimoto similarity scores for the test compounds.

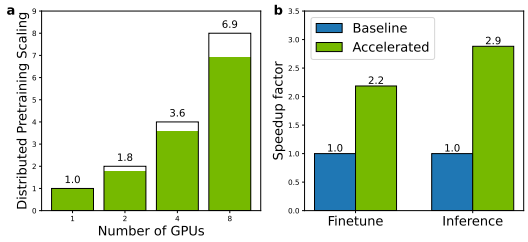


Figure 5. **KERMT model parallelization and acceleration.** **a** Scaling efficiency of pretraining KERMT model parallelized across multiple GPUs. **b** Acceleration of KERMT model for fine-tuning and inference using `cuik-molmaker` for batched molecular graph generation in dataloader.

We hypothesize that KERMT's superior performance stems from its larger model capacity, which enables it to fully exploit the rich information in large, multitask datasets. The general representations learned during pretraining provide a more effective starting point for multitask fine-tuning, facilitating the learning of a shared representation that captures complex inter-task relationships.

To validate this, we are separating pretraining from the larger model capacity through random initialization of KERMT. Initial results on the Biogen dataset, which is the smallest of the three datasets, show that random initialization of KERMT degrades performance below both pretrained KERMT and Chemprop MT (Appendix Fig. 10). We are expanding this validation to our public and internal datasets.

The disparity in how KERMT and KPGT benefit from multitasking is likely due to differences in their GNN architecture and pretraining tasks. KERMT's dual-trunk GTransformer architecture may offer greater flexibility, while KPGT's use of a "knowledge node" for each pretraining task might limit the regularization effect of multitasking. We recognize that this work validates KERMT against several leading baselines but does not compare it to more recent graph foundation chemistry models, such as UniMol2, and look to extend our benchmarking.

Our architecture and acceleration work with KERMT combines the inductive bias of graph neural networks, the learning capacity of large-parameter models, and the regularization benefits of multitasking. The promising performance of KERMT in the recent OpenADMET competition, where it was used in five of the top 20 submissions (out of 370+ total submissions), validates its potential (Castellanos et al., 2026) and we advocate for the continued development of multitask benchmarks to further advance this class of models.

Impact Statement

We have developed multitask chemistry foundation models that significantly enhances the accuracy of ADMET property prediction. This improvement enables the production of more successful molecules from generative models and empowers drug discovery teams to better prioritize candidate designs. These advancements will lead to faster development of therapeutics and, ultimately, significant societal benefits through improved disease management and patient outcomes. As with all methods that accelerate drug discovery, the techniques presented here could be applied towards the discovery of harmful molecules. Responsible use of molecular prioritization methods in practice requires appropriate domain oversight, which is outside the scope of this work.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: a next-generation hyperparameter optimization framework. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2623–2631, 2019. doi: 10.1145/3292500.3330701.

Cáceres, E. L., Tudor, M., and Cheng, A. C. Deep learning approaches in predicting ADMET properties. *Future Medicinal Chemistry*, 12(22):1995–1999, 2020. ISSN 17568927. doi: 10.4155/fmc-2020-0259.

Caruana, R. Multitask learning. *Machine Learning*, 28: 41–75, 1997. ISSN 14680319. doi: 10.1111/j.1468-0319.1995.tb00042.x.

Castellanos, M., MacDermott-Opeskin, H., Ainsley, J., and Walters, P. Lessons learned from the OpenADMET - ExpansionRx blind challenge. https://openadmet.ghost.io, February 2026. Accessed: 2026-05-07.

Chen, J., Chung, Y., Tynan, J., Cheng, C., Yang, S., and Cheng, A. C. Data scaling and generalization insights for medicinal chemistry deep learning models. *Journal of Chemical Information and Modeling*, 2025. ISSN 1549960X. doi: 10.1021/acs.jcim.5c00538.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1(Mlm): 4171–4186, 2019.

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

Fang, C., Wang, Y., Grater, R., Kapadnis, S., Black, C., Trapa, P., and Sciabola, S. 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. ISSN 1549960X. doi: 10.1021/acs.jcim.3c00160.

Feinberg, E. N., Joshi, E., Pande, V. S., and Cheng, A. C. Improvement in ADMET prediction with multitask deep featurization. *Journal of Medicinal Chemistry*, 63(16): 8835–8848, 2020. doi: 10.1021/acs.jmedchem.9b02187.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40(D1):1100–1107, 2012. ISSN 03051048. doi: 10.1093/nar/gkr777.

Gaulton, A., Hersey, A., Nowotka, M. L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 2017. ISSN 13624962. doi: 10.1093/nar/gkw1074.

Graff, D. E., Morgan, N. K., Burns, J. W., Doner, A. C., Li, B., Li, S.-C., Manu, J., Menon, A., Pang, H.-W., Wu, H., Zalte, A. S., Zheng, J. W., Coley, C. W., Green, W. H., and Greenman, K. P. Chemprop v2: An efficient, modular machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 66(1):28–33, 2026. doi: 10.1021/acs.jcim.5c02332. URL https://doi.org/10.1021/acs.jcim.5c02332. PMID: 41453060.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, jan 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01250. URL https://pubs.acs.org/doi/10.1021/acs.jcim.3c01250.

Irwin, J. J. and Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005. ISSN 15499596. doi: 10.1021/ci049714+.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. ISSN 13624962. doi: 10.1093/nar/gkac956.

Landrum, G. rdkit/rdkit:2021_09_2 (q3 2021). 2021.

Li, H., Zhang, R., Min, Y., Ma, D., Zhao, D., and Zeng, J. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 2023. doi: 10.1038/s41467-023-43214-1.

Liu, T., Hwang, L., Burley, S. K., Nitsche, C. I., Southan, C., Walters, W. P., and Gilson, M. K. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Research*, 53(November 2024):1633–1644, 2024. URL https://chemrxiv.org/engage/chemrxiv/article-details/66e360db12ff75c3a12b3a00.

Rong, Y., Bian, Y., Xu, T., Xie, W., and Wei, Y. Self-Supervised Graph Transformer on Large-Scale Molecular Data. (NeurIPS):1–13, 2020.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Payel, D. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790, 2013. ISSN 1549960X. doi: 10.1021/ci400084k.

Sterling, T. and Irwin, J. J. ZINC 15 - ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015. ISSN 1549960X. doi: 10.1021/acs.jcim.5b00559.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Transformers: state-of-the-art natural language processing. *EMNLP 2020 - Conference on Empirical Methods in Natural Language Processing, Proceedings of Systems Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3353–3600, 2019. doi: 10.1021/acs.jcim.9b00237.

A. ADMET comparison details

A.1. Performance differences across seven model types

In addition to the large proprietary industrial dataset (Fig. 6a), we curated available public ADMET datasets as benchmarks to assess the performance of the seven methods. To better represent real-world performance, we used a cluster split for comparisons using these datasets. Temporal splits would be even more realistic but timing data was not available for these datasets.

For eleven diverse endpoints, both Chemprop MT and KERMT models performed well, with both being at the top or tied for top for 73% (8 of 11 endpoints) (Fig. 6b-c). Among just the single task models, KPGT performed the best and is either at the top or tied for top across all 11 endpoints. Due to the small test sizes in the public datasets, many models are tied in performance ranking once statistical significance is accounted for.

We then assessed just the multitask models on a larger set of 25 public assays and found that KPGT MT outperforms Chemprop MT in about 50% of tasks, while KERMT outperforms Chemprop MT in about 40% of tasks (Fig. 7d). However, for nearly all assays with at least 500 datapoints, KERMT outperforms or performs equivalently to Chemprop MT, with average improvements of 12% (Table 1). For assays with under 550 datapoints, KPGT MT generally equals or outperforms Chemprop MT, though the low absolute performance in most of these low-data cases limit the practical impact.

The public dataset consists of data aggregated from multiple sources. To understand performance on single-source data that is public, we performed a multiple comparison analysis using repeated measures ANOVA followed by the Tukey HSD test for all methods and all six assays from the Biogen public dataset (Fig. 6b). We find that KERMT outperforms or performs equivalently to the best model in all six assays. Chemprop MT, the next best performing model on the Biogen benchmark, outperforms or performs equivalently to the best model in five out of six assays. Interestingly, for the two small datasets, rat and human $F_{u,p}$ with 120-135 training data, KERMT does not outperform but does perform equivalently to the best model, Chemprop MT.

Similar to the proprietary data benchmarks, KPGT does not improve from multitasking in the public datasets (Fig. 7e). In fact, KPGT decreases in performance for $> 50\%$ of the assays across both the public and Biogen datasets.

A.2. Details on pre-trained model hyperparameters

Table 1. KERMT common hyperparameters

Parameter	Value
Max learning rate	1E-04
Final learning rate	2E-05
Dropout	0
Self-attention heads	8
Coefficient of disagreement loss	0.15
Drop edge ratio	0.2
FFN depth	3
FFN hidden size	700

Table 2. KPGT common hyperparameters

Parameter	Value
Learning rate	2.74E-05
Dropout	0
Weight decay	0
LLRD decay rate	0.95 for complex; (1 for simple)
Reinit top-n layers	3 for complex; (0 for simple)
L^2-SP weight	0.02 for complex; (0 for simple)

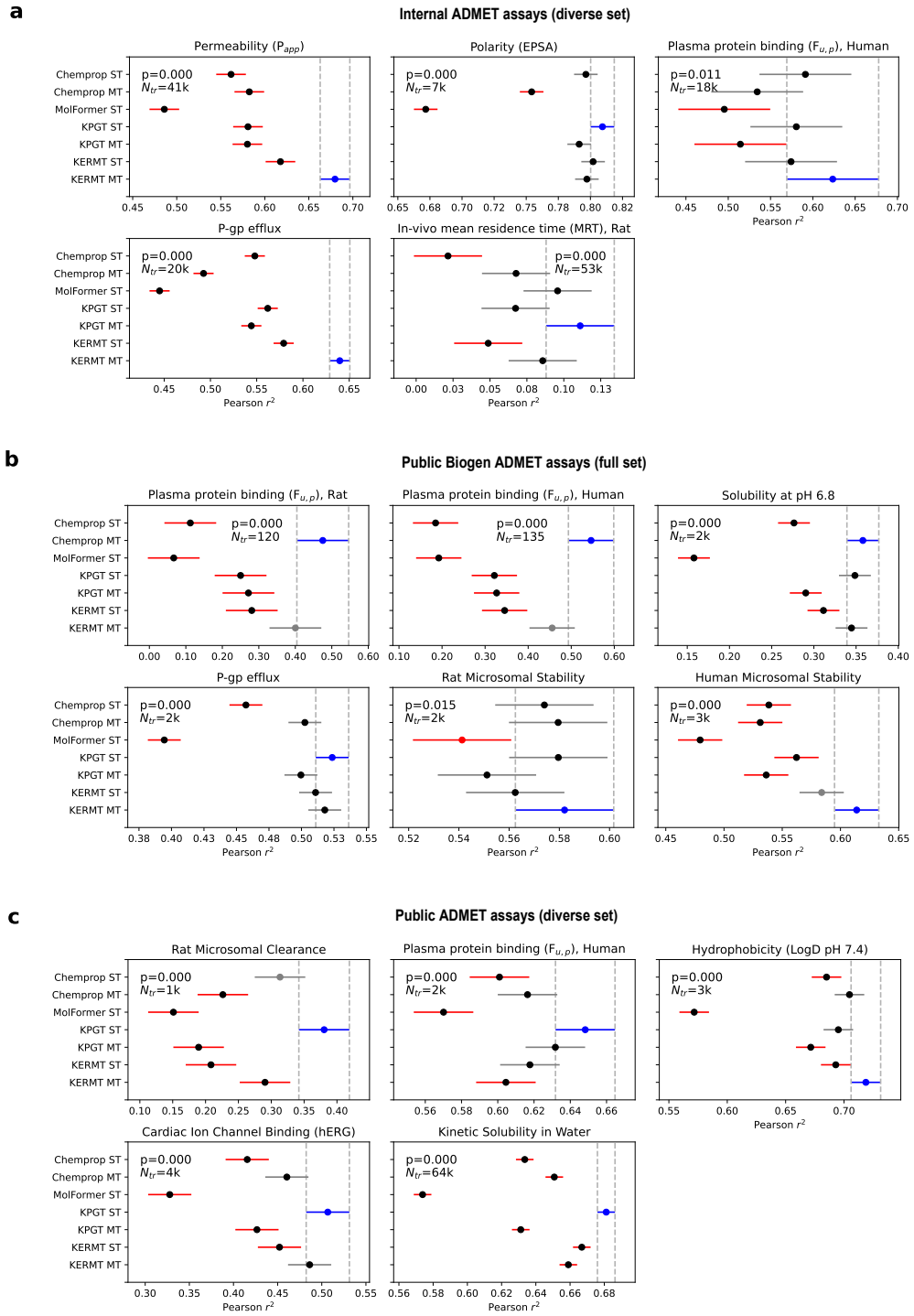


Figure 6. **Model comparisons on a diverse set of ADMET assays**. **a** A diverse set of proprietary internal ADMET datasets, **b** the public Biogen dataset, and **c** a diverse set of public datasets. A Tukey HSD statistical analysis is used to determine significant performance differences for single-task (ST) and multitask (MT) models. For each plot, the model in blue has the highest mean Pearson r^2, models in gray are statistically equal to the top performing model, and models in red are statistically worse than the top model. The dashed vertical lines indicate the confidence intervals for the best method.

Table 3. MoLFormer common hyperparameters

Parameter	Value
Learning rate	3E-05
Dropout	0.1
FFN depth	2
FFN hidden size	768

A.3. Details of performance on Public and Proprietary ADMET datasets

On the proprietary ADMET dataset, the performance improvement from multitask training is visualized in Fig. 7c. The KERMT model shows the most significant improvement from multitasking and does not degrade in performance on any assay. In contrast, KPGT benefits little from multitasking and even shows reduced performance for the human $F_{u,p}$ assay. Across all 30 internal ADMET endpoints, KERMT is statistically superior in 18 assays, whereas KPGT MT is best in only one.

When assessing multitask models on 25 public assays (Fig. 7d), we found that for assays with at least 500 data points, KERMT outperforms or performs equivalently to Chemprop MT. For smaller assays with under 550 data points, KPGT MT generally performs better.

Finally, Fig. 7e demonstrates that, similar to the proprietary benchmarks, KPGT does not benefit from multitasking on the public datasets and shows a decrease in performance for over 50% of the assays.

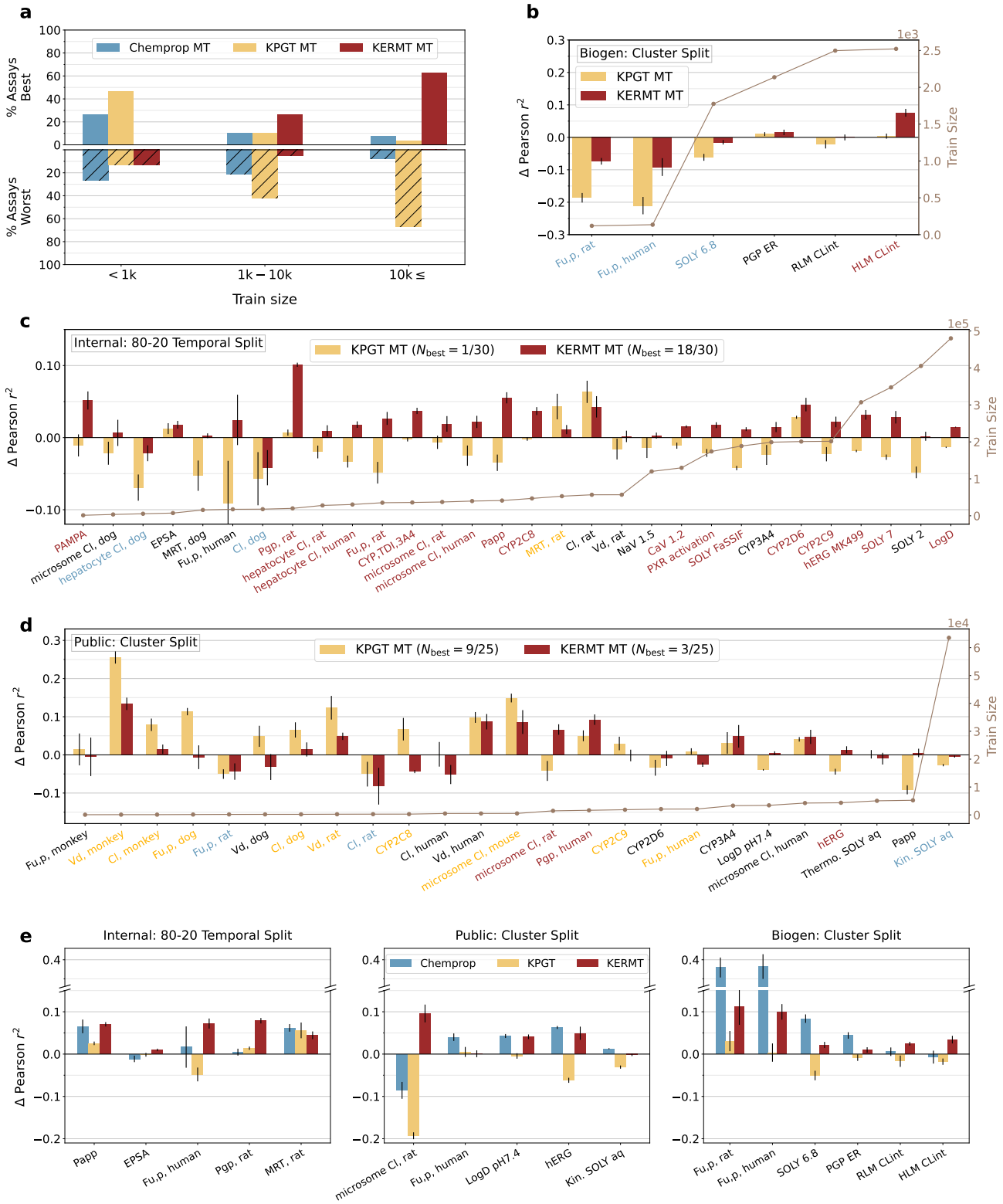


Figure 7. **Comparison of model performance on ADMET assay predictions. a** Number of assays the model statistically outperforms (N_{best}) or underperforms (N_{worst}) for different size ranges of assay training data, covering all datasets. **b** - **d** Performance improvement of pretrained MT models compared to Chemprop MT model. The plots show the results on all available assays for the Biogen, internal, and public datasets, respectively. Brown solid lines represent the number of training data for each assay. A colored assay name (blue: Chemprop, yellow: KPGT, red: KERMT) represents the model which performs statistically better as indicated by the Pearson r^2 being higher than the upper error limit of other models. Black labeled assays indicate that two or three models performed statistically the same. N_{best} represents the number of assays the model statistically outperforms. **e** Performance improvement of multitask (MT) models compared to single-task (ST) models.

Table 4. Pearson r^2 on internal ADMET data (80-20 temporal split). In parentheses are % improvement relative to Chemprop MT.

Assay Name	Train Size	Chemprop MT	KPGT MT	KERMT MT
PAMPA	1.4k	0.596 ± 0.026	0.585 ± 0.012 (-1.85%)	0.647 ± 0.011 (8.56%)
microsome Cl, dog	3.6k	0.528 ± 0.027	0.506 ± 0.018 (-4.17%)	0.534 ± 0.015 (1.14%)
hepatocyte Cl, dog	5.6k	0.543 ± 0.021	0.474 ± 0.016 (-12.71%)	0.521 ± 0.009 (-4.05%)
EPSA	7.4k	0.805 ± 0.011	0.817 ± 0.007 (1.49%)	0.822 ± 0.002 (2.11%)
MRT, dog	15.8k	0.190 ± 0.019	0.138 ± 0.042 (-27.37%)	0.193 ± 0.024 (1.58%)
Fu,p, human	17.6k	0.641 ± 0.089	0.550 ± 0.040 (-14.20%)	0.666 ± 0.025 (3.90%)
Cl, dog	18.0k	0.253 ± 0.045	0.196 ± 0.029 (-22.53%)	0.211 ± 0.013 (-16.60%)
Pgp, rat	20.1k	0.582 ± 0.003	0.589 ± 0.008 (1.20%)	0.683 ± 0.008 (17.35%)
hepatocyte Cl, rat	28.1k	0.376 ± 0.009	0.356 ± 0.017 (-5.32%)	0.385 ± 0.011 (2.39%)
hepatocyte Cl, human	30.6k	0.405 ± 0.007	0.372 ± 0.010 (-8.15%)	0.423 ± 0.012 (4.44%)
Fu,p, rat	35.4k	0.624 ± 0.021	0.575 ± 0.018 (-7.85%)	0.650 ± 0.004 (4.17%)
CYP,TDI,3A4	36.1k	0.288 ± 0.007	0.285 ± 0.004 (-1.04%)	0.325 ± 0.006 (12.85%)
microsome Cl, rat	37.4k	0.577 ± 0.012	0.571 ± 0.013 (-1.04%)	0.596 ± 0.012 (3.29%)
microsome Cl, human	39.9k	0.580 ± 0.017	0.554 ± 0.012 (-4.48%)	0.601 ± 0.008 (3.62%)
Papp	41.3k	0.657 ± 0.014	0.622 ± 0.012 (-5.33%)	0.712 ± 0.005 (8.37%)
CYP2C8	47.0k	0.408 ± 0.009	0.407 ± 0.005 (-0.25%)	0.445 ± 0.004 (9.07%)
MRT, rat	53.0k	0.085 ± 0.012	0.128 ± 0.037 (50.59%)	0.096 ± 0.014 (12.94%)
Cl, rat	56.9k	0.393 ± 0.041	0.457 ± 0.022 (16.28%)	0.436 ± 0.031 (10.94%)
Vd, rat	57.0k	0.514 ± 0.017	0.498 ± 0.018 (-3.11%)	0.516 ± 0.010 (0.39%)
NaV 1.5	120.1k	0.278 ± 0.008	0.263 ± 0.022 (-5.40%)	0.280 ± 0.010 (0.72%)
CaV 1.2	129.8k	0.281 ± 0.006	0.270 ± 0.007 (-3.91%)	0.296 ± 0.003 (5.34%)
PXR activation	174.2k	0.380 ± 0.009	0.359 ± 0.013 (-5.53%)	0.397 ± 0.004 (4.47%)
SOLY FaSSIF	188.8k	0.302 ± 0.003	0.260 ± 0.006 (-13.91%)	0.314 ± 0.008 (3.97%)
CYP3A4	199.2k	0.456 ± 0.026	0.432 ± 0.012 (-5.26%)	0.470 ± 0.014 (3.07%)
CYP2D6	200.9k	0.258 ± 0.006	0.286 ± 0.006 (10.85%)	0.304 ± 0.014 (17.83%)
CYP2C9	201.6k	0.340 ± 0.012	0.317 ± 0.010 (-6.76%)	0.361 ± 0.004 (6.18%)
hERG MK499	307.2k	0.469 ± 0.010	0.450 ± 0.011 (-4.05%)	0.500 ± 0.004 (6.61%)
SOLY 7	347.5k	0.584 ± 0.015	0.556 ± 0.012 (-4.79%)	0.612 ± 0.003 (4.79%)
SOLY 2	405.4k	0.694 ± 0.017	0.645 ± 0.008 (-7.06%)	0.695 ± 0.007 (0.14%)
LogD	480.2k	0.880 ± 0.001	0.866 ± 0.002 (-1.59%)	0.895 ± 0.001 (1.70%)

Table 5. Pearson r^2 of multitask methods on cluster split public ADMET data. In parenthesis are % improvement versus Chemprop MT.

Assay Name	Train Size	Chemprop MT	KPGT MT	KERMT MT
Fu,p, monkey	64	0.369 ± 0.072	0.383 ± 0.017 (3.79%)	0.364 ± 0.056 (-1.36%)
Vd, monkey	84	0.163 ± 0.004	0.418 ± 0.029 (156.44%)	0.297 ± 0.033 (82.21%)
Cl, monkey	87	0.067 ± 0.026	0.146 ± 0.027 (117.91%)	0.082 ± 0.020 (22.39%)
Fu,p, dog	129	0.350 ± 0.038	0.463 ± 0.047 (32.29%)	0.344 ± 0.030 (-1.71%)
Fu,p, rat	167	0.334 ± 0.032	0.283 ± 0.009 (-15.27%)	0.290 ± 0.033 (-13.17%)
Vd, dog	187	0.266 ± 0.051	0.315 ± 0.022 (18.42%)	0.234 ± 0.038 (-12.03%)
Cl, dog	194	0.095 ± 0.053	0.160 ± 0.029 (68.42%)	0.109 ± 0.025 (14.74%)
Vd, rat	240	0.169 ± 0.017	0.293 ± 0.057 (73.37%)	0.218 ± 0.010 (28.99%)
Cl, rat	263	0.189 ± 0.065	0.139 ± 0.014 (-26.46%)	0.107 ± 0.035 (-43.39%)
CYP2C8	275	0.129 ± 0.028	0.196 ± 0.037 (51.94%)	0.085 ± 0.025 (-34.11%)
Cl, human	500	0.233 ± 0.052	0.235 ± 0.013 (0.86%)	0.182 ± 0.024 (-21.89%)
Vd, human	506	0.257 ± 0.034	0.355 ± 0.008 (38.13%)	0.344 ± 0.022 (33.85%)
microsome Cl, mouse	527	0.223 ± 0.009	0.372 ± 0.019 (66.82%)	0.309 ± 0.060 (38.57%)
microsome Cl, rat	1.5k	0.252 ± 0.039	0.210 ± 0.025 (-16.67%)	0.318 ± 0.021 (26.19%)
Pgp, human	1.7k	0.170 ± 0.023	0.220 ± 0.024 (29.41%)	0.262 ± 0.016 (54.12%)
CYP2C9	1.9k	0.266 ± 0.016	0.295 ± 0.023 (10.90%)	0.264 ± 0.028 (-0.75%)
CYP2D6	2.1k	0.336 ± 0.025	0.302 ± 0.030 (-10.12%)	0.327 ± 0.019 (-2.68%)
Fu,p, human	2.1k	0.658 ± 0.002	0.668 ± 0.017 (1.52%)	0.632 ± 0.013 (-3.95%)
CYP3A4	3.3k	0.287 ± 0.033	0.319 ± 0.040 (11.15%)	0.336 ± 0.030 (17.07%)
LogD pH7.4	3.4k	0.738 ± 0.008	0.699 ± 0.011 (-5.28%)	0.742 ± 0.012 (0.54%)
microsome Cl, human	4.2k	0.429 ± 0.025	0.470 ± 0.014 (9.56%)	0.476 ± 0.012 (10.96%)
hERG	4.4k	0.506 ± 0.013	0.462 ± 0.019 (-8.70%)	0.520 ± 0.025 (2.77%)
Thermo. SOLY aq	5.1k	0.782 ± 0.022	0.783 ± 0.002 (0.13%)	0.772 ± 0.013 (-1.28%)
Papp	5.2k	0.501 ± 0.018	0.410 ± 0.008 (-18.16%)	0.506 ± 0.010 (1.00%)
Kin. SOLY aq	63.5k	0.689 ± 0.002	0.662 ± 0.007 (-3.92%)	0.684 ± 0.004 (-0.73%)

Table 6. Pearson r^2 values of multitask methods on Biogen ADMET data with the cluster split. Values in parenthesis represent the percent improvement relative to Chemprop MT.

Assay Name	Train Size	Chemprop MT	KPGT MT	KERMT MT
Fu,p, rat	120	0.501 ± 0.046	0.315 ± 0.020 (-37.13%)	0.427 ± 0.038 (-14.77%)
Fu,p, human	135	0.572 ± 0.059	0.362 ± 0.019 (-36.71%)	0.481 ± 0.040 (-15.91%)
SOLY 6.8	1.8k	0.380 ± 0.012	0.318 ± 0.013 (-16.32%)	0.365 ± 0.008 (-3.95%)
Pgp MDCK	2.1k	0.519 ± 0.012	0.529 ± 0.007 (1.93%)	0.535 ± 0.009 (3.08%)
microsome Cl, rat	2.5k	0.602 ± 0.011	0.580 ± 0.020 (-3.65%)	0.602 ± 0.011 (0.00%)
microsome Cl, human	2.5k	0.558 ± 0.019	0.562 ± 0.007 (0.72%)	0.634 ± 0.010 (13.62%)

Table 7. Public ADMET data benchmarking of single-task and multitask methods (Pearson r^2) on the cluster split

Model	microsome Cl, rat n_{train}=1k	$F_{u,p}$, human n_{train}=2k	LogD n_{train}=3k	hERG binding n_{train}=4k	Kinetic SOLY, aq n_{train}=64k
MolFormer ST	0.168 ± 0.027	0.593 ± 0.013	0.596 ± 0.006	0.353 ± 0.025	0.602 ± 0.002
Chemprop ST	0.338 ± 0.041	0.618 ± 0.017	0.695 ± 0.011	0.443 ± 0.013	0.676 ± 0.005
KPGT ST_{complex}	0.337 ± 0.006	0.641 ± 0.007	0.624 ± 0.004	0.434 ± 0.010	0.619 ± 0.002
KPGT ST	0.403 ± 0.013	0.662 ± 0.009	0.703 ± 0.011	0.524 ± 0.015	0.692 ± 0.001
KERMT ST	0.223 ± 0.028	0.631 ± 0.008	0.702 ± 0.006	0.470 ± 0.014	0.686 ± 0.002
Chemprop MT	0.252 ± 0.039	0.658 ± 0.002	0.738 ± 0.008	0.506 ± 0.013	0.689 ± 0.002
KPGT MT	0.210 ± 0.025	0.668 ± 0.017	0.699 ± 0.011	0.462 ± 0.019	0.662 ± 0.007
KERMT MT	0.318 ± 0.021	0.632 ± 0.013	0.742 ± 0.012	0.520 ± 0.025	0.684 ± 0.004

Table 8. Biogen ADMET data benchmarking of single-task and multitask methods (Pearson r^2) on the cluster split

Model	$F_{u,p}$, rat n_{train}=120	$F_{u,p}$, human n_{train}=135	SOLY 6.8 n_{train}=2k	P-gp MDCK n_{train}=2k	microsome Cl, rat n_{train}=2k	microsome Cl, human n_{train}=3k
MolFormer ST	0.119 ± 0.023	0.269 ± 0.035	0.181 ± 0.010	0.415 ± 0.007	0.570 ± 0.015	0.510 ± 0.004
Chemprop ST	0.121 ± 0.085	0.190 ± 0.012	0.296 ± 0.027	0.475 ± 0.009	$\mathbf{0.596 \pm 0.014}$	0.566 ± 0.022
KPGT ST_{complex}	0.269 ± 0.043	0.291 ± 0.031	0.337 ± 0.003	$\mathbf{0.520 \pm 0.007}$	0.548 ± 0.013	0.505 ± 0.004
KPGT ST	0.284 ± 0.064	0.358 ± 0.062	$\mathbf{0.369 \pm 0.012}$	$\mathbf{0.538 \pm 0.012}$	$\mathbf{0.597 \pm 0.013}$	0.580 ± 0.009
KERMT ST	0.314 ± 0.064	0.381 ± 0.018	0.343 ± 0.009	$\mathbf{0.525 \pm 0.009}$	$\mathbf{0.577 \pm 0.017}$	0.600 ± 0.019
Chemprop MT	$\mathbf{0.501 \pm 0.046}$	$\mathbf{0.572 \pm 0.059}$	$\mathbf{0.380 \pm 0.012}$	$\mathbf{0.519 \pm 0.012}$	$\mathbf{0.602 \pm 0.011}$	0.558 ± 0.019
KPGT MT	0.315 ± 0.020	0.362 ± 0.019	0.318 ± 0.013	$\mathbf{0.529 \pm 0.007}$	$\mathbf{0.580 \pm 0.020}$	0.562 ± 0.007
KERMT MT	$\mathbf{0.427 \pm 0.038}$	$\mathbf{0.481 \pm 0.040}$	$\mathbf{0.365 \pm 0.008}$	$\mathbf{0.535 \pm 0.009}$	$\mathbf{0.602 \pm 0.011}$	$\mathbf{0.634 \pm 0.010}$

Table 9. Internal ADMET data benchmarking of single-task and multitask methods (Pearson r^2) on the 80-20 temporal split

Model	P_{app} n_{train}=41k	EPSA n_{train}=7k	$F_{u,p}$, human n_{train}=18k	Pgp, rat n_{train}=20k	MRT, rat n_{train}=53k
MolFormer ST	0.506 ± 0.014	0.679 ± 0.009	0.534 ± 0.032	0.468 ± 0.010	$\mathbf{0.115 \pm 0.011}$
Chemprop ST	0.591 ± 0.021	$\mathbf{0.816 \pm 0.007}$	$\mathbf{0.625 \pm 0.019}$	0.578 ± 0.014	0.023 ± 0.010
KPGT ST_{complex}	0.496 ± 0.004	0.795 ± 0.003	0.574 ± 0.015	0.513 ± 0.004	0.058 ± 0.012
KPGT ST	0.596 ± 0.004	$\mathbf{0.819 \pm 0.002}$	0.598 ± 0.012	0.575 ± 0.004	0.072 ± 0.004
KERMT ST	0.641 ± 0.012	0.813 ± 0.005	0.594 ± 0.007	0.605 ± 0.006	0.052 ± 0.013
Chemprop MT	0.657 ± 0.014	0.805 ± 0.011	$\mathbf{0.641 \pm 0.089}$	0.582 ± 0.003	$\mathbf{0.085 \pm 0.012}$
KPGT MT	0.622 ± 0.012	$\mathbf{0.817 \pm 0.007}$	0.550 ± 0.040	0.589 ± 0.008	$\mathbf{0.128 \pm 0.037}$
KERMT MT	$\mathbf{0.712 \pm 0.005}$	$\mathbf{0.822 \pm 0.002}$	$\mathbf{0.666 \pm 0.025}$	$\mathbf{0.683 \pm 0.008}$	$\mathbf{0.096 \pm 0.014}$

B. Details of on-target potency comparison

Table 10. Pearson r^2 values of single-task methods on on-target potency data. Values in parenthesis represent the percent improvement relative to Chemprop ST.

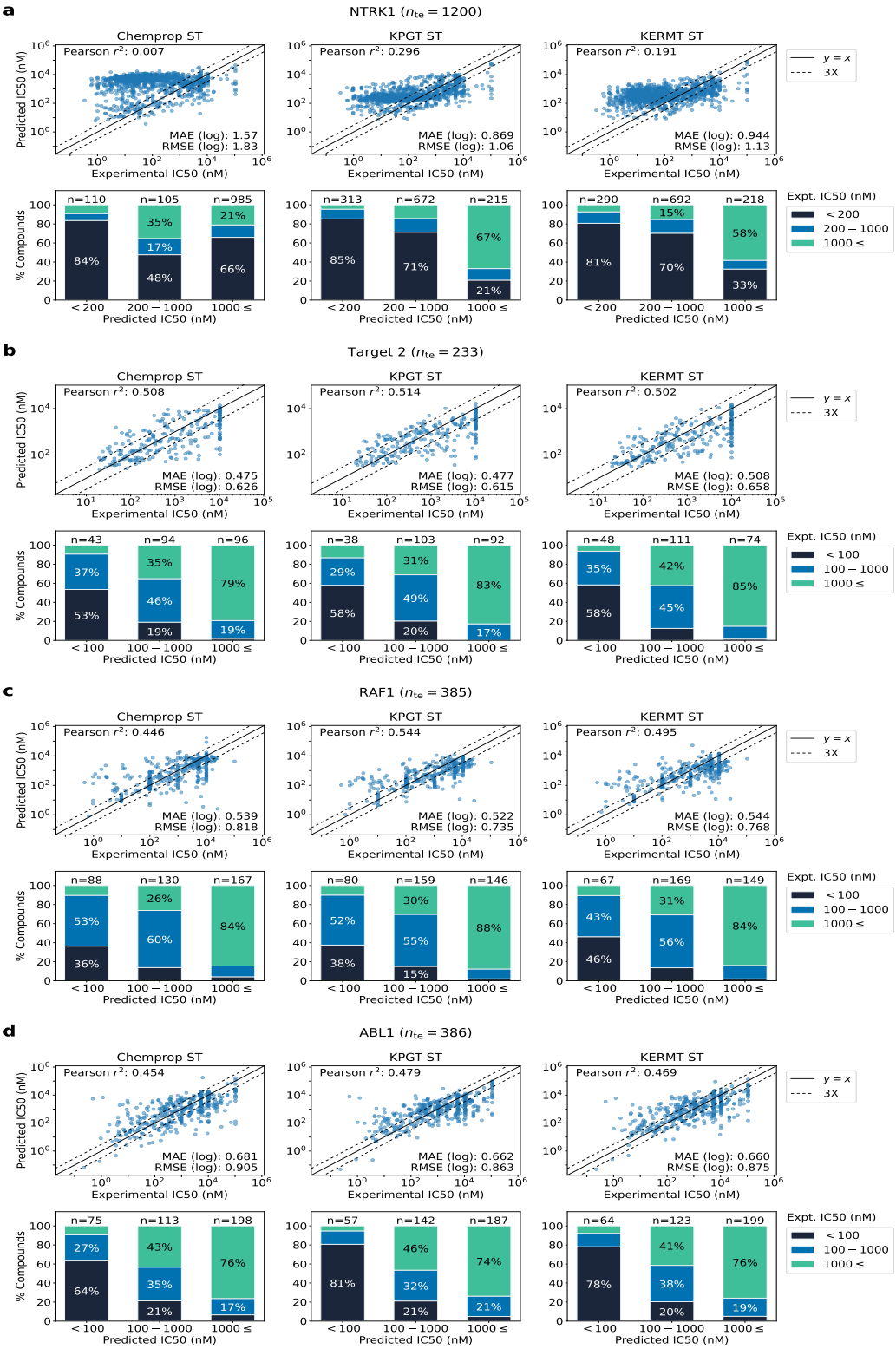
Target Name	Train Size	Chemprop ST	KPGT ST	KERMT ST	Boltz-2
Target 1	595	0.328 ± 0.055	0.410 ± 0.033 (25.00%)	0.338 ± 0.103 (3.05%)	0.036 ± 0.013 (-89.02%)
NTRK1	696	0.007 ± 0.009	0.296 ± 0.082 (4128.57%)	0.191 ± 0.029 (2628.57%)	0.316 ± 0.014 (4414.29%)
Target 2	930	0.508 ± 0.010	0.514 ± 0.022 (1.18%)	0.502 ± 0.022 (-1.18%)	0.073 ± 0.012 (-85.63%)
RAF1	1.0k	0.446 ± 0.015	0.544 ± 0.020 (21.97%)	0.495 ± 0.005 (10.99%)	0.170 ± 0.056 (-61.88%)
ABL1	1.7k	0.454 ± 0.024	0.479 ± 0.022 (5.51%)	0.469 ± 0.021 (3.30%)	0.153 ± 0.031 (-66.30%)
ERBB2	1.8k	0.417 ± 0.027	0.419 ± 0.022 (0.48%)	0.352 ± 0.016 (-15.59%)	0.088 ± 0.031 (-78.90%)
PIK3CG	1.9k	0.219 ± 0.041	0.308 ± 0.031 (40.64%)	0.236 ± 0.013 (7.76%)	0.457 ± 0.071 (108.68%)
MAPK14	2.6k	0.386 ± 0.031	0.435 ± 0.014 (12.69%)	0.410 ± 0.022 (6.22%)	0.333 ± 0.074 (-13.73%)
PIK3CA	2.6k	0.213 ± 0.053	0.189 ± 0.017 (-11.27%)	0.163 ± 0.013 (-23.47%)	0.496 ± 0.065 (132.86%)
BRAF	3.4k	0.487 ± 0.015	0.570 ± 0.018 (17.04%)	0.494 ± 0.025 (1.44%)	0.549 ± 0.001 (12.73%)
HPK1	3.5k	0.250 ± 0.064	0.293 ± 0.070 (17.20%)	0.285 ± 0.044 (14.00%)	0.285 ± 0.022 (14.00%)
KDR	5.5k	0.359 ± 0.016	0.370 ± 0.010 (3.06%)	0.341 ± 0.012 (-5.01%)	0.068 ± 0.015 (-81.06%)
EGFR_{kin}	5.8k	0.542 ± 0.007	0.585 ± 0.020 (7.93%)	0.573 ± 0.010 (5.72%)	0.121 ± 0.002 (-77.68%)
BTK	7.5k	0.181 ± 0.006	0.298 ± 0.022 (64.64%)	0.303 ± 0.012 (67.40%)	0.318 ± 0.025 (75.69%)
EGFR_{bdb}	7.8k	0.214 ± 0.040	0.291 ± 0.026 (35.98%)	0.312 ± 0.015 (45.79%)	0.000 ± 0.000 (-100.00%)
FGFR1	9.2k	0.627 ± 0.016	0.693 ± 0.005 (10.53%)	0.676 ± 0.007 (7.81%)	0.343 ± 0.026 (-45.30%)

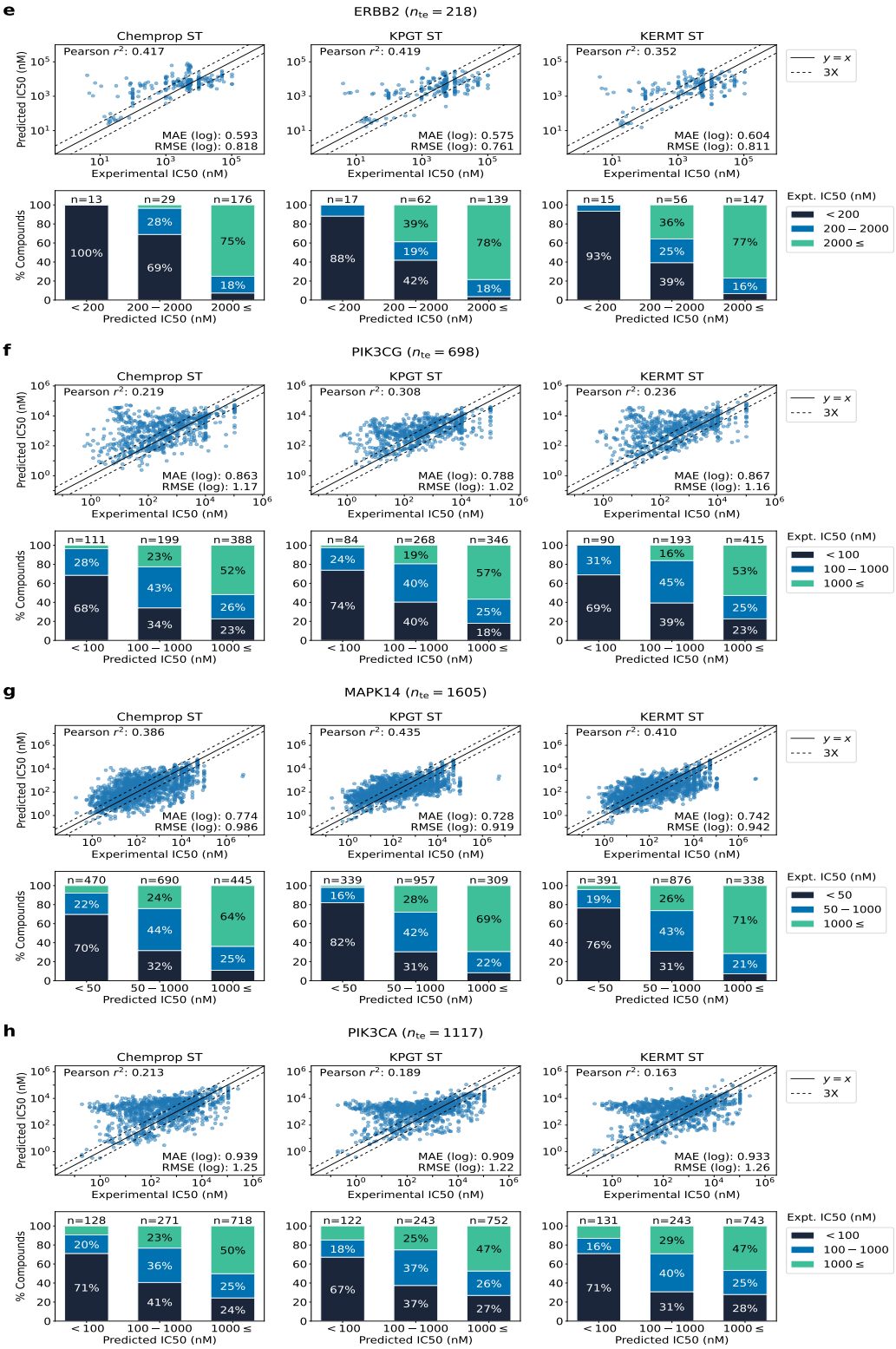
Table 11. Pearson r^2 values of multitask methods on kinase potency data. Values in parenthesis represent the percent improvement relative to Chemprop MT.

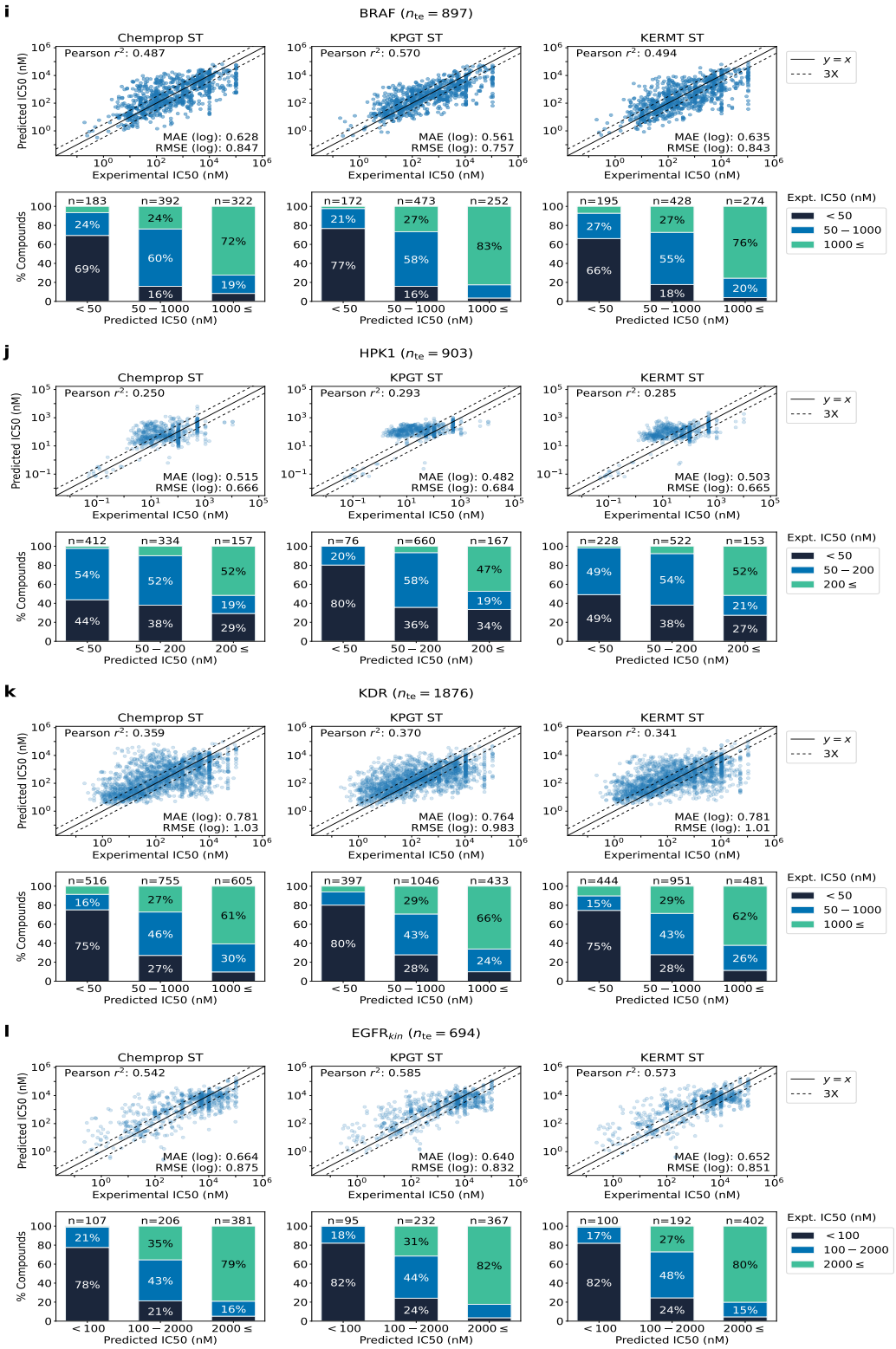
Target Name	Train Size	Chemprop MT	KPGT MT	KERMT MT
NTRK1	696	0.103 ± 0.053	0.126 ± 0.046 (22.33%)	0.131 ± 0.056 (27.18%)
RAF1	1.0k	0.433 ± 0.048	0.512 ± 0.029 (18.24%)	0.533 ± 0.021 (23.09%)
ABL1	1.7k	0.521 ± 0.024	0.476 ± 0.017 (-8.64%)	0.459 ± 0.019 (-11.90%)
ERBB2	1.8k	0.538 ± 0.100	0.418 ± 0.060 (-22.30%)	0.412 ± 0.017 (-23.42%)
PIK3CG	1.9k	0.310 ± 0.061	0.328 ± 0.027 (5.81%)	0.276 ± 0.043 (-10.97%)
MAPK14	2.6k	0.386 ± 0.051	0.444 ± 0.010 (15.03%)	0.436 ± 0.029 (12.95%)
PIK3CA	2.6k	0.163 ± 0.071	0.302 ± 0.025 (85.28%)	0.163 ± 0.013 (0.00%)
BRAF	3.4k	0.456 ± 0.035	0.593 ± 0.009 (30.04%)	0.499 ± 0.014 (9.43%)
KDR	5.5k	0.326 ± 0.010	0.348 ± 0.005 (6.75%)	0.380 ± 0.013 (16.56%)
EGFR	5.8k	0.595 ± 0.025	0.559 ± 0.022 (-6.05%)	0.571 ± 0.010 (-4.03%)

Table 12. Performance comparison (Pearson r^2) across ten kinase targets and six methods (Mean ± SD). Bold indicates the best method and methods with statistically overlapping 95% confidence intervals (assuming n=5 runs per method).

Target	Train Size	Chemprop ST	KPGT ST	KERMT ST	Chemprop MT	KPGT MT	KERMT MT
NTRK1	696	0.007±0.009	**0.296±0.082**	**0.191±0.029**	0.103±0.053	0.126±0.046	0.131±0.056
RAF1	1000	0.446±0.015	**0.544±0.020**	0.495±0.005	0.433±0.048	**0.512±0.029**	**0.533±0.021**
ABL1	1700	0.454±0.024	0.479±0.022	0.469±0.021	**0.521±0.024**	**0.476±0.017**	0.459±0.019
ERBB2	1800	0.417±0.027	**0.419±0.022**	0.352±0.016	**0.538±0.100**	**0.418±0.060**	**0.412±0.017**
PIK3CG	1900	0.219±0.041	0.308±0.031	0.236±0.013	0.310±0.061	**0.328±0.027**	0.276±0.043
MAPK14	2600	0.386±0.031	**0.435±0.014**	0.410±0.022	0.386±0.051	**0.444±0.010**	**0.436±0.029**
PIK3CA	2600	0.213±0.053	0.189±0.017	0.163±0.013	0.163±0.071	**0.302±0.025**	0.163±0.013
BRAF	3400	0.487±0.015	**0.570±0.018**	0.494±0.025	0.456±0.035	**0.593±0.009**	0.499±0.014
KDR	5500	0.359±0.016	0.370±0.010	0.341±0.012	0.326±0.010	0.348±0.005	**0.380±0.013**
EGFR	5800	0.542±0.007	**0.585±0.020**	**0.573±0.010**	**0.595±0.025**	0.559±0.022	**0.571±0.010**


a
NTRK1 ($n_{te} = 1200$)
Chemprop ST
Pearson r^2: 0.007
MAE (log): 1.57
RMSE (log): 1.83
KPGT ST
Pearson r^2: 0.296
MAE (log): 0.869
RMSE (log): 1.06
KERMT ST
Pearson r^2: 0.191
MAE (log): 0.944
RMSE (log): 1.13
$y = x$
3X
Predicted IC50 (nM)
Experimental IC50 (nM)
% Compounds
n=110 n=105 n=985
84% 35% 17% 48% 21% 66%
n=313 n=672 n=215
85% 71% 67% 21%
n=290 n=692 n=218
81% 70% 15% 33% 58%
< 200
200 − 1000
1000 ≤
Expt. IC50 (nM)
b
Target 2 ($n_{te} = 233$)
Chemprop ST
Pearson r^2: 0.508
MAE (log): 0.475
RMSE (log): 0.626
KPGT ST
Pearson r^2: 0.514
MAE (log): 0.477
RMSE (log): 0.615
KERMT ST
Pearson r^2: 0.502
MAE (log): 0.508
RMSE (log): 0.658
n=43 n=94 n=96
53% 37% 19% 46% 35% 19% 79%
n=38 n=103 n=92
58% 29% 20% 49% 31% 17% 83%
n=48 n=111 n=74
58% 35% 45% 42% 85%
< 100
100 − 1000
1000 ≤
c
RAF1 ($n_{te} = 385$)
Chemprop ST
Pearson r^2: 0.446
MAE (log): 0.539
RMSE (log): 0.818
KPGT ST
Pearson r^2: 0.544
MAE (log): 0.522
RMSE (log): 0.735
KERMT ST
Pearson r^2: 0.495
MAE (log): 0.544
RMSE (log): 0.768
n=88 n=130 n=167
36% 53% 60% 26% 84%
n=80 n=159 n=146
38% 52% 15% 55% 30% 88%
n=67 n=169 n=149
46% 43% 56% 31% 84%
d
ABL1 ($n_{te} = 386$)
Chemprop ST
Pearson r^2: 0.454
MAE (log): 0.681
RMSE (log): 0.905
KPGT ST
Pearson r^2: 0.479
MAE (log): 0.662
RMSE (log): 0.863
KERMT ST
Pearson r^2: 0.469
MAE (log): 0.660
RMSE (log): 0.875
n=75 n=113 n=198
64% 27% 21% 35% 43% 17% 76%
n=57 n=142 n=187
81% 21% 32% 46% 21% 74%
n=64 n=123 n=199
78% 20% 38% 41% 19% 76%


e
ERBB2 ($n_{te} = 218$)
Chemprop ST
Pearson r^2: 0.417
MAE (log): 0.593
RMSE (log): 0.818
KPGT ST
Pearson r^2: 0.419
MAE (log): 0.575
RMSE (log): 0.761
KERMT ST
Pearson r^2: 0.352
MAE (log): 0.604
RMSE (log): 0.811
Predicted IC50 (nM)
Experimental IC50 (nM)
y = x
3X
% Compounds
n=13 n=29 n=176
100% 69% 28% 18% 75%
n=17 n=62 n=139
88% 42% 19% 39% 18% 78%
n=15 n=56 n=147
93% 39% 25% 36% 16% 77%
< 200
200 − 2000
2000 ≤
Predicted IC50 (nM)
Expt. IC50 (nM)
f
PIK3CG ($n_{te} = 698$)
Chemprop ST
Pearson r^2: 0.219
MAE (log): 0.863
RMSE (log): 1.17
KPGT ST
Pearson r^2: 0.308
MAE (log): 0.788
RMSE (log): 1.02
KERMT ST
Pearson r^2: 0.236
MAE (log): 0.867
RMSE (log): 1.16
n=111 n=199 n=388
68% 34% 23% 28% 43% 23% 26% 52%
n=84 n=268 n=346
74% 40% 40% 24% 19% 18% 25% 57%
n=90 n=193 n=415
69% 39% 23% 31% 45% 16% 25% 53%
< 100
100 − 1000
1000 ≤
g
MAPK14 ($n_{te} = 1605$)
Chemprop ST
Pearson r^2: 0.386
MAE (log): 0.774
RMSE (log): 0.986
KPGT ST
Pearson r^2: 0.435
MAE (log): 0.728
RMSE (log): 0.919
KERMT ST
Pearson r^2: 0.410
MAE (log): 0.742
RMSE (log): 0.942
n=470 n=690 n=445
70% 32% 22% 44% 24% 25% 64%
n=339 n=957 n=309
82% 31% 42% 16% 28% 22% 69%
n=391 n=876 n=338
76% 31% 43% 19% 26% 21% 71%
< 50
50 − 1000
1000 ≤
h
PIK3CA ($n_{te} = 1117$)
Chemprop ST
Pearson r^2: 0.213
MAE (log): 0.939
RMSE (log): 1.25
KPGT ST
Pearson r^2: 0.189
MAE (log): 0.909
RMSE (log): 1.22
KERMT ST
Pearson r^2: 0.163
MAE (log): 0.933
RMSE (log): 1.26
n=128 n=271 n=718
71% 41% 24% 20% 36% 25% 23% 50%
n=122 n=243 n=752
67% 37% 27% 18% 37% 26% 25% 47%
n=131 n=243 n=743
71% 31% 28% 16% 40% 25% 29% 47%
< 100
100 − 1000
1000 ≤


i
BRAF ($n_{te} = 897$)
Chemprop ST
Pearson r^2: 0.487
MAE (log): 0.628
RMSE (log): 0.847
KPGT ST
Pearson r^2: 0.570
MAE (log): 0.561
RMSE (log): 0.757
KERMT ST
Pearson r^2: 0.494
MAE (log): 0.635
RMSE (log): 0.843
$y = x$
3X
Predicted IC50 (nM)
Experimental IC50 (nM)
% Compounds
n=183 n=392 n=322
n=172 n=473 n=252
n=195 n=428 n=274
Expt. IC50 (nM)
< 50
50 − 1000
1000 ≤
Predicted IC50 (nM)
j
HPK1 ($n_{te} = 903$)
Chemprop ST
Pearson r^2: 0.250
MAE (log): 0.515
RMSE (log): 0.666
KPGT ST
Pearson r^2: 0.293
MAE (log): 0.482
RMSE (log): 0.684
KERMT ST
Pearson r^2: 0.285
MAE (log): 0.503
RMSE (log): 0.665
n=412 n=334 n=157
n=76 n=660 n=167
n=228 n=522 n=153
Expt. IC50 (nM)
< 50
50 − 200
200 ≤
k
KDR ($n_{te} = 1876$)
Chemprop ST
Pearson r^2: 0.359
MAE (log): 0.781
RMSE (log): 1.03
KPGT ST
Pearson r^2: 0.370
MAE (log): 0.764
RMSE (log): 0.983
KERMT ST
Pearson r^2: 0.341
MAE (log): 0.781
RMSE (log): 1.01
n=516 n=755 n=605
n=397 n=1046 n=433
n=444 n=951 n=481
Expt. IC50 (nM)
< 50
50 − 1000
1000 ≤
l
EGFR$_{kin}$ ($n_{te} = 694$)
Chemprop ST
Pearson r^2: 0.542
MAE (log): 0.664
RMSE (log): 0.875
KPGT ST
Pearson r^2: 0.585
MAE (log): 0.640
RMSE (log): 0.832
KERMT ST
Pearson r^2: 0.573
MAE (log): 0.652
RMSE (log): 0.851
n=107 n=206 n=381
n=95 n=232 n=367
n=100 n=192 n=402
Expt. IC50 (nM)
< 100
100 − 2000
2000 ≤

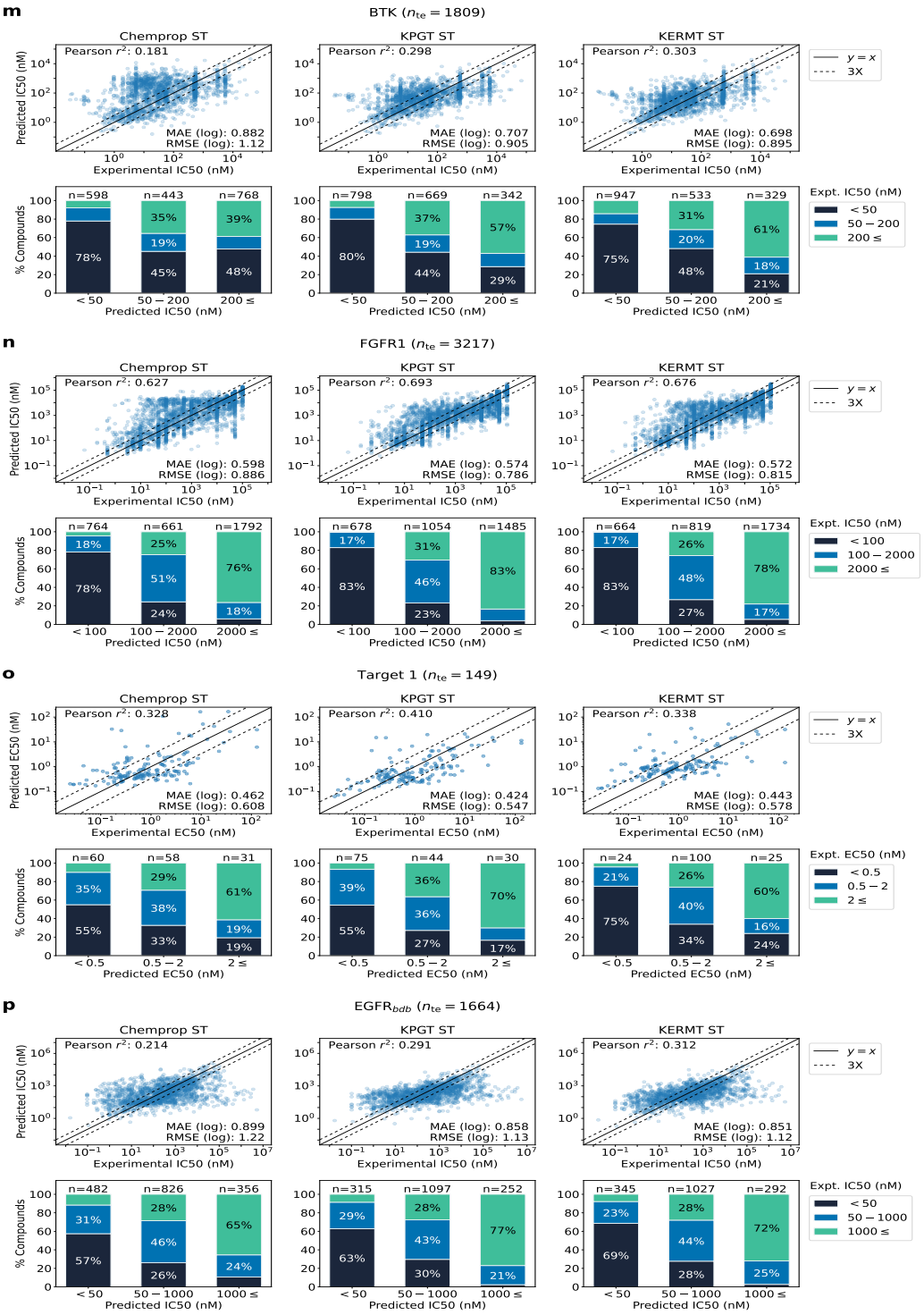


Figure 8. **Single-task model performance on on-target potency predictions.** Parity plots (top) and categorical bar plots (bottom) are displayed for each model. In the parity plot, MAE and RMSE represent mean absolute error and root mean squared error in $\log_{10}(\mathrm{nM})$ unit. The percentage in the categorical bar plot indicates the percentage of the predicted bin (x-axis) that belongs to each experimental bin (colored legend)

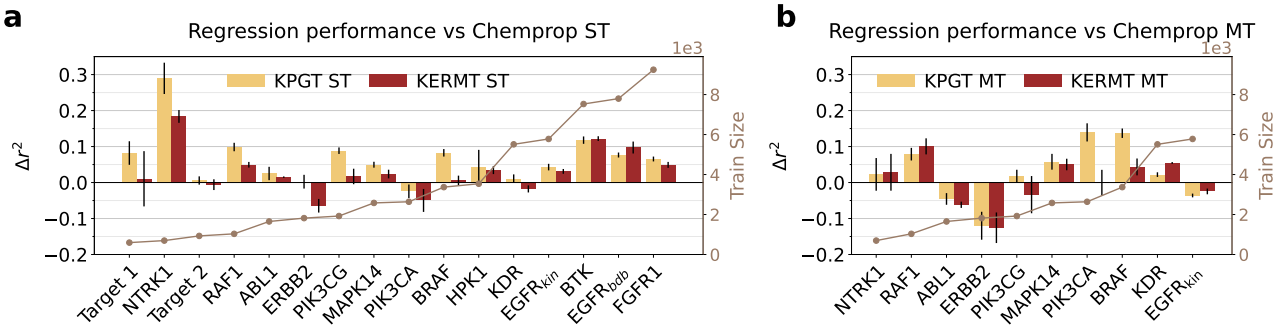


Figure 9. **Model performance for low-data, on-target potency prediction. a**, Single-task regression performance comparison of KPGT ST and KERMT ST relative to Chemprop ST based on Pearson r^2, with training data sizes, ranging from 0.6K to 9.2K, indicated in brown. **b**, multitask kinase potency model regression performance comparison of KPGT MT and KERMT MT relative to Chemprop MT based on Pearson r^2, with training data size indicated in brown. **c**, Categorical performance for 16 targets based on weighted F1 score. **d**, Categorical performance for 10 targets based on weighted F1 score.

C. Details of Data and Task Scaling Behavior

In Fig. 10, we benchmark randomly initialized KERMT. In Fig. 11a-c, we provide details on the behavior of KERMT performance relative to Chemprop MT and in Fig. 11d we provide detailed analysis on five diverse ADMET endpoints evaluated on the internal dataset.

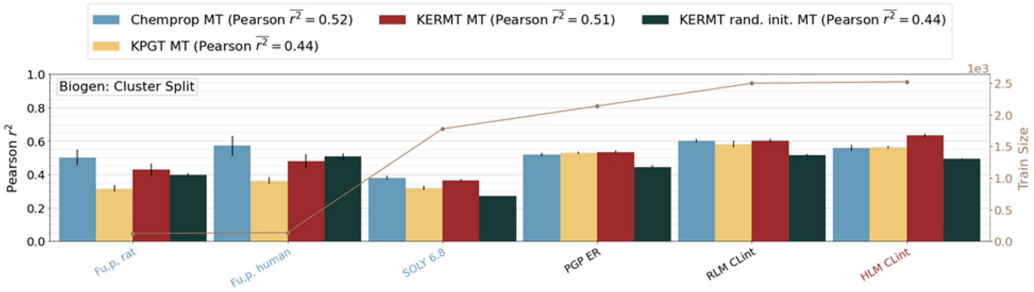


Figure 10. Benchmarking randomly initialized KERMT against Chemprop, KERMT, and KPGT multitask models.

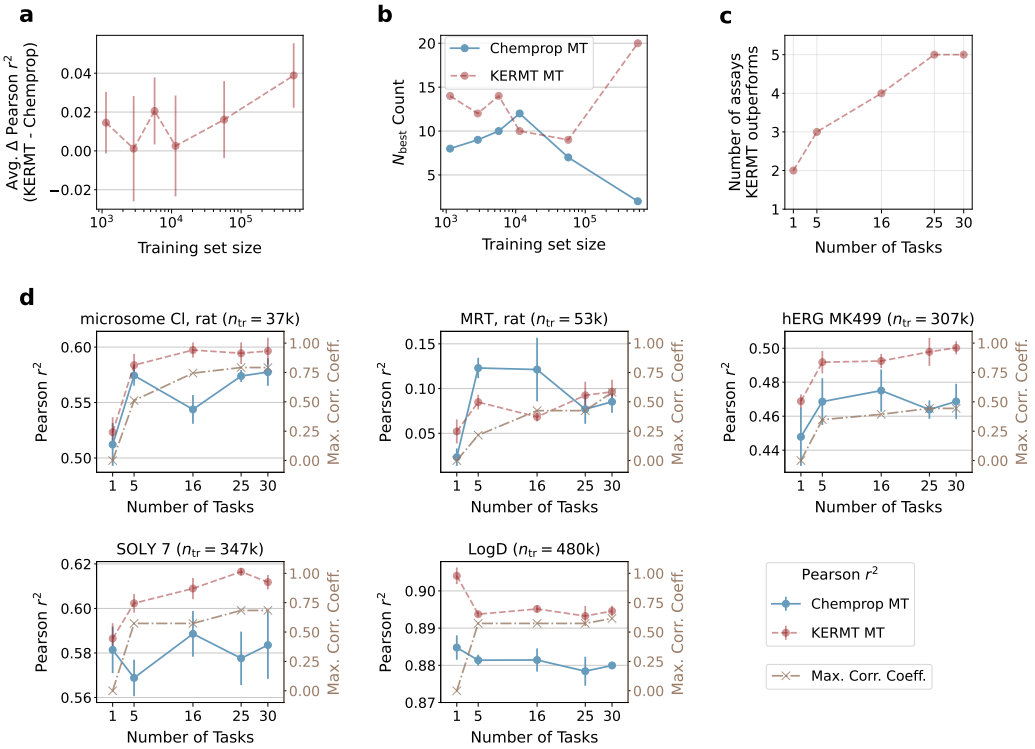


Figure 11. **Scaling of Chemprop multitask and KERMT models with training set size and number of tasks.** Evaluation on the internal ADMET dataset with an 80-20 temporal split. **a** Training set size influence on performance improvement by KERMT relative to Chemprop MT based on the 10 assays with the most performance difference between KERMT and Chemprop models. **b** Number of assays each model statistically outperforms for different training set sizes. **c** Number of tasks influence on performance improvement by KERMT relative to Chemprop MT **d** Pearson r^2 of each model on five diverse ADMET assays across different numbers of tasks for multitask models. A brown dash-dotted line with cross ('x') markers represents the maximum value of the absolute Spearman correlation coefficients between the selected assay and all other assays in each multitask dataset.

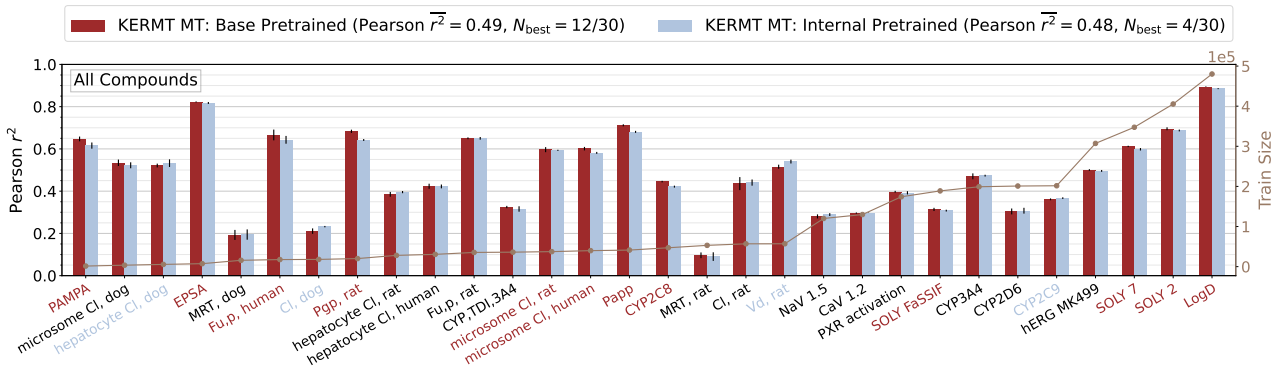


Figure 12. **Comparison of downstream performance after fine-tuning the KERMT base model (from Rong et al. (Rong et al., 2020) and trained on public ChEMBL and ZINC compounds) and the KERMT model pretrained on the internal compound data.** All results are evaluated on the internal ADMET dataset with the 80-20 temporal split. Model performance shown is on all test set compounds. The brown solid line represents the number of training data for each assay. A colored assay name represents the model which performs statistically better. Pearson $\overline{r^2}$ represents the average performance over all assays, and N_{best} represents the number of assays the model statistically outperforms.