Similarity-Quantized Relative Difference Learning for Improved Molecular Activity Prediction

Karina Zadorozhny **Kangway V. Chuang** **Bharath Sathappan**
Ewan Wallace **Vishnu Sresht** **Colin A. Grambow**
Prescient Design, Genentech
South San Francisco, CA 94080
{zadorozhny.karina,grambow.colin}@gene.com

Abstract

Accurate prediction of molecular activities is crucial for efficient drug discovery, yet remains challenging due to limited and noisy datasets. We introduce Similarity-Quantized Relative Learning (SQRL), a learning framework that reformulates molecular activity prediction as relative difference learning between structurally similar pairs of compounds. SQRL uses precomputed molecular similarities to enhance training of graph neural networks and other architectures, and significantly improves accuracy and generalization in low-data regimes common in drug discovery. We demonstrate its broad applicability and real-world potential through benchmarking on public datasets as well as proprietary industry data. Our findings demonstrate that leveraging similarity-aware relative differences provides an effective paradigm for molecular activity prediction.

1 Introduction and Background

The ability to predict molecular activity is critical for small molecule drug discovery. However, experimental data for training machine learning models are often limited and noisy, making robust generalization challenging. While deep learning approaches have enabled rich representations from chemical structures [1–13], most focus on predicting absolute property values, ignoring valuable information in relationships between structurally similar molecules.

Drawing inspiration from medicinal chemists [14], who examine how specific structural modifications influence properties relative to a parent compound or a matched molecular pair [15], we introduce **S**imilarity-**Q**uantized **R**elative **L**earning (**SQRL**)—a training and evaluation framework that reformulates property prediction as learning relative differences between nearby compounds.

The key contributions of our work are:

- We introduce a robust **similarity-thresholded learning approach** that significantly enhances model performance by focusing on **predicting property differences between the most informative compound pairs**. This allows effective learning from limited and noisy data.
- Our analysis across molecular distance metrics and thresholds demonstrates that **similarity-aware dataset matching outperforms indiscriminate pairing of all inputs**, as it more effectively leverages local structural information.
- Extensive **benchmarking** demonstrating the benefits of SQRL **across diverse state-of-the-art network architectures** and multiple activity prediction datasets, including publicly available activity cliff prediction tasks and real-world industry datasets.

AI for New Drug Modalities at NeurIPS 2024.

2 Related Work

Molecular property prediction. Significant research has focused on methods for molecular property and activity prediction, including recent work on graph neural networks (GNNs) [1–7] along with chemical language models [8–12] that learn meaningful molecular representations from molecular data [13]. Despite these advances, simpler tree-based models often outperform more complex neural approaches due to limited data availability and inherent modeling challenges [16, 17].

Activity cliff prediction. Activity cliffs refer to pairs of molecules with high structural similarity but significantly different activity levels. Previous approaches have addressed this challenging problem by representing molecular graphs as images [18, 19], applying graph convolutional networks (GCNs) to matched molecular pairs [20], or using chemical reaction information [21]. Many of these methods have formulated the problem as a classification task, aiming to identify whether a given molecular pair exhibits an activity cliff [22] or as a standard regression task that relies on learning the discontinuous chemical space directly from the data [23]. Our approach focuses on predicting the difference in potency values between any pair of similar molecules, providing a more versatile solution that can also be applied to standard potency prediction.

Metric, similarity, and few-shot learning. Pairwise learning approaches have been widely adopted in fields like ranking, metric, and similarity learning [24, 25]. Notably, few-shot learning approaches have been developed for molecular property prediction for improving generalization in low-data regimes [26–28]. Additionally, pairwise data matching has proven effective in implicit guidance of generative models for drug design. [29].

Relative prediction. Recently, pairwise learning has been applied to regression tasks. Wetzel et al. used Siamese networks to predict differences between all data points in both supervised and unsupervised settings as a way of producing ensembles of predictions and uncertainty estimates [30, 31]. This approach has been extended to classification tasks for tree-based models [32]. Tynes et al. applied the concept of pairwise learning to computational chemistry, analyzing the performance of random forest models trained on all pairs of inputs points [33]. Similarly, Fralish et al. trained a D-MPNN model on paired compounds and observed improvements in ADME property predictions [34] and molecule selection in an active learning setting [35].

Unlike previous approaches that primarily focus on absolute property predictions or indiscriminate pairwise learning, our work introduces a similarity-thresholded framework that emphasizes learning from the most informative compound pairs.

3 Similarity-Thresholded Relative Representation

Problem formulation. We formulate the relative prediction task as follows. Given a dataset of molecular structures $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents molecule i and $y_i \in \mathbb{R}$ denotes its corresponding property value, our goal is to learn a function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that predicts the relative difference in property values between two molecules. Formally, for any pair of molecules (x_i, x_j), we aim to predict the relative difference $\Delta y_{ij} = y_i - y_j$.

Dataset matching. To train models on the relative prediction task, we construct a new dataset $\mathcal{D}_{\text{rel}}$ by considering pairs of molecules in the original dataset $\mathcal{D}$. To focus on the most informative comparisons, we restrict the pairs to those within a certain threshold of structural similarity, as measured by a predefined similarity metric (e.g., Tanimoto similarity). This allows models to learn from local differences in chemical space where relative changes in property values are most meaningful. We define $\mathcal{D}_{\text{rel}}$ as:

$$\mathcal{D}_{\text{rel}} = \{((x_i, x_j), \Delta y_{ij}) \mid x_i, x_j \in \mathcal{D}, d(x_i, x_j) \leq \alpha\} \tag{1}$$

where $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ is a distance function in the input space $\mathcal{X}$ and $\alpha \in \mathbb{R}_{>0}$ is a distance threshold. See Appendix A.4 for examples of paired structures.

The choice of α is critical, as it involves a trade-off between the quantity and relevance of generated pairs. We propose selecting α based on the distribution of distances in the training data, specifically by choosing a threshold smaller than the average pairwise distance. This approach can help form more informative pairs by focusing on molecules with greater structural similarity and relevance.

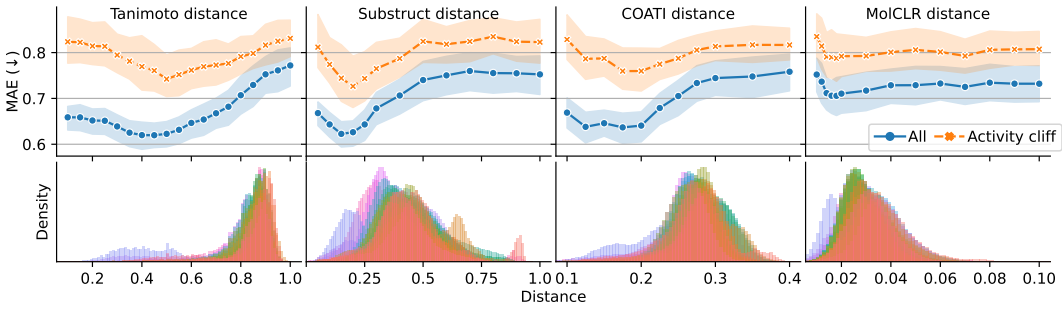

Figure 1: **Leveraging local structural information enhances predictive performance.** *Top*: Incorporating neighbors only up to a certain distance threshold α improves MAE (↓). *Bottom*: Pairwise distance distributions of training data (overlaid for all 30 MoleculeACE tasks) with greater skewness and kurtosis yield the best performance and a wider range of acceptable values of α.

Relative representation. We define $g : \mathcal{X} \to \mathbb{R}^d$ as a mapping function that converts a molecular compound from the input space $\mathcal{X}$ into a d-dimensional real-valued vector. This can be either a learnable model (such as a graph neural network), a pre-trained model, or a fixed molecular fingerprinting algorithm. Next, a machine learning model $f : \mathbb{R}^d \to \mathbb{R}$ uses the difference between molecular representations generated by g to predict the relative differences in properties.

We optimize parameters θ of f and g (if learnable) by minimizing:

$$\min_{\theta} \mathcal{L}(\theta) = \min_{\theta} \sum_{((x_i,x_j),\Delta y_{ij})\in\mathcal{D}\text{rel}} \ell\big(f\left(g(x_i) - g(x_j)\right), \Delta y_{ij}\big) \tag{2}$$

where ℓ is mean squared error loss. For a new molecule x_{new}, we compute the prediction $\hat{y}_{\text{new}}$ as:

$$\hat{y}_{\text{new}} = \frac{1}{n} \sum_{x_i \in \text{NN}_n(x_{\text{new}})} y_i + f\left(g(x_i) - g(x_{\text{new}})\right) \tag{3}$$

where $\text{NN}_n(x_{\text{new}})$ denotes the set of n molecules from the training data $\mathcal{D}$ that are nearest to x_{new} as determined by the distance function $d(x_i, x_{\text{new}})$. Unless otherwise specified, we set $n = 1$.

4 Experimental Results

4.1 Experimental setup

To understand the generality of SQRL, we conducted extensive evaluations across a diverse set of models and molecular activity datasets. Each model was trained to predict absolute property values directly (Standard) or to predict relative differences between selected molecule pairs (SQRL). More details about models, hyperparameter selection, and datasets can be found in Appendices A.1–A.4.

Models. We benchmarked baselines (RF [36], XGBoost [37], KNN) using Morgan fingerprints [38] with RDKit features [39], MLP with Morgan fingerprints, GNNs (AttentiveFP [4], GINE [40], PNA [41], MolCLR [42]), and transformer models (COATI [43], SAFE-GPT [44], and Uni-Mol [45]).

Distance metrics. Several distance metrics were used to create training data pairs and obtain the closest training molecules for inference. We evaluated Tanimoto distances between Morgan fingerprints [38], Tanimoto distances between substructure count vectors [46], as well as Euclidean distances using COATI [43], Uni-Mol, and MolCLR embeddings [42] (see A.5).

Datasets. We evaluated our approach on 30 activity prediction tasks (pEC_{50}/$\text{p}K_i$) for ChEMBL targets using the MoleculeACE dataset [23]. This dataset includes challenging activity cliff molecules (MoleculeACE-Cliff)—structurally similar compounds with large activity differences—providing a crucial test of local generalizability (see A.4). Additionally, we evaluated models on 5 proprietary drug discovery projects (Internal Targets) to assess the real-world applicability of our approach. Models are trained on single tasks and results are reported aggregated across tasks.

Table 1: **Spearman's ρ (↑) comparison of predictive performance.** Spearman's ρ with standard deviation across tasks for Standard and SQRL methods using Tanimoto distance with $\alpha = 0.7$. Uni-Mol was evaluated on a subset of tasks due to computational constraints. See Appendix A.1 for more details and MAE scores (Table 3).

	MoleculeACE		**MoleculeACE-Cliff**		**Internal Targets**	
Model	Standard	SQRL	Standard	SQRL	Standard	SQRL
Baselines						
XGBoost	**0.79 ± 0.10**	0.76 ± 0.08	**0.72 ± 0.18**	0.64 ± 0.15	**0.73 ± 0.14**	0.65 ± 0.16
RF	**0.80 ± 0.09**	0.77 ± 0.07	**0.72 ± 0.18**	0.68 ± 0.14	**0.70 ± 0.13**	0.68 ± 0.14
KNN	0.66 ± 0.10	**0.67 ± 0.13**	0.57 ± 0.19	**0.59 ± 0.23**	**0.52 ± 0.14**	**0.52 ± 0.14**
MLP	0.32 ± 0.17	**0.73 ± 0.09**	0.23 ± 0.16	**0.62 ± 0.16**	0.39 ± 0.14	**0.59 ± 0.20**
GNNs						
AttentiveFP	0.52 ± 0.20	**0.77 ± 0.09**	0.43 ± 0.19	**0.67 ± 0.17**	0.49 ± 0.13	**0.66 ± 0.19**
GINE	0.33 ± 0.19	**0.76 ± 0.09**	0.29 ± 0.21	**0.68 ± 0.18**	0.40 ± 0.22	**0.61 ± 0.17**
PNA	0.51 ± 0.18	**0.72 ± 0.08**	0.41 ± 0.17	**0.61 ± 0.18**	0.52 ± 0.12	**0.64 ± 0.15**
MolCLR	0.35 ± 0.22	**0.77 ± 0.09**	0.28 ± 0.20	**0.66 ± 0.18**	0.39 ± 0.27	**0.68 ± 0.17**
Transformers						
COATI	0.69 ± 0.11	**0.74 ± 0.09**	0.59 ± 0.15	**0.64 ± 0.16**	0.52 ± 0.26	**0.61 ± 0.18**
Uni-Mol	0.26 ± 0.19	**0.69 ± 0.10**	0.19 ± 0.20	**0.57 ± 0.18**	0.40 ± 0.16	**0.51 ± 0.14**
SAFE-GPT	0.61 ± 0.20	**0.71 ± 0.08**	0.56 ± 0.22	**0.61 ± 0.14**	0.58 ± 0.14	**0.59 ± 0.15**

4.2 Leveraging local structural information improves learning

A key hypothesis of SQRL is that not all possible relative pairs are equally informative, and training on all pairwise comparisons as done previously [32–34] may overemphasize global relationships at the expense of local consistency. We analyzed pairwise distance distributions using various distance metrics and found that some distributions exhibit significant left-skew and/or high kurtosis (Figure 1 and Appendix Figure 5). We hypothesize that these highly similar compounds contain the most informative signal. To test this assumption, we trained MLP models on top of Morgan fingerprint features across a range of similarity thresholds for each of the metrics (Figure 1). We observed that incorporating neighbors up to a certain threshold provides a pronounced improvement in performance, but beyond this point, including more dissimilar pairs degrades performance. Moreover, performance is best for distance metrics with the desired distribution characteristics (e.g., Tanimoto and COATI). These results demonstrate the benefit of our similarity-thresholded approach: a smaller distance threshold α yields fewer but potentially more informative pairs, while a larger threshold increases pair count but may introduce less relevant comparisons. Our findings support the hypothesis that focusing on the most similar pairs effectively leverages local structural information (Appendix Figure 6), leading to improved predictive performance across various model architectures and datasets.

4.3 SQRL consistently improves predictive performance across all neural network architectures

To understand which model types benefit from this approach, we performed extensive benchmarking of molecular property prediction models trained using both the standard absolute prediction objective and SQRL. While the baseline models did not significantly benefit from relative training, all deep learning architectures exhibit consistent improvement of the Spearman rank correlation across all datasets when trained with SQRL (Table 1). We observe significant improvements for all GNN architectures and especially for the pre-trained transformer-based models—0.57 and 0.43 point improvement for COATI and Uni-Mol, respectively. Notably, substantial improvements are observed on the MoleculeACE-Cliff subset, highlighting SQRL's ability to capture fine-grained structural differences that significantly impact molecular properties. Evaluation on internal targets shows that the observed benefits are transferable to real-world scenarios and underscore the robustness and broad applicability of this approach.

The lack of improvement for XGBoost, RF, and KNN with SQRL is notable. This may indicate that the simple difference fingerprint representation is not sufficient for these models as it forces these models to only learn from substructures that differ between a pair of molecules without taking into account the rest of the molecular structure of both molecules. A higher fidelity representation, such as concatenating the full molecular fingerprints to the difference representation, may overcome this limitation. Even so, most modern deep learning methods do not outperform conventional baselines (with or without SQRL). The small size of the training data for each task likely contributes to this result. However, we expect that neural network approaches may outperform baselines with additional tuning (e.g., more rigorous hyperparameter selection, choosing a more optimal distance threshold, or redesigning the training objective).

5 Conclusions and Future Directions

Overall, our work demonstrates consistent improvements for neural networks in molecular activity prediction using SQRL, particularly in capturing molecular activity cliffs. Our method shows promise in learning from pairwise differences, potentially offering a more nuanced understanding of structure-activity relationships. The limitations of the current work include the assumption that meaningful distance metrics are available. Future research could focus on refining SQRL for applications where similarity measures are less well-defined or more challenging to establish.

References

[1] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12):1023–1032, 2021. URL https://doi.org/10.1038/s42256-021-00418-8.

[2] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1263–1272, 2017. URL https://proceedings.mlr.press/v70/gilmer17a/gilmer17a.pdf.

[3] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. URL https://doi.org/10.1021/acs.jcim.9b00237.

[4] Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of Medicinal Chemistry*, 63 (16):8749–8760, 2020. URL https://doi.org/10.1021/acs.jmedchem.9b00959.

[5] Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical message passing for 3D molecular graphs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=givsRXsOt9r.

[6] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf.

[7] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2022. URL https://arxiv.org/abs/2102.09844.

[8] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pages 429–436, New York, NY, USA, 2019. Association for Computing Machinery. URL https://doi.org/10.1145/3307339.3342186.

[9] Shion Honda, Shoi Shi, and Hiroki R. Ueda. SMILES Transformer: Pre-trained molecular fingerprint for low data drug discovery, 2019. URL https://arxiv.org/abs/1911.04738.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

[11] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

[12] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. URL https://doi.org/10.1038/s42256-022-00580-7.

[13] Kangway V. Chuang, Laura M. Gunsalus, and Michael J. Keiser. Learning molecular representations for medicinal chemistry. *Journal of Medicinal Chemistry*, 63(16):8705–8722, 2020. URL https://doi.org/10.1021/acs.jmedchem.0c00385.

[14] Gareth Thomas. *Fundamentals of Medicinal Chemistry*. John Wiley & Sons, Chichester, England, 1st edition, March 2004.

[15] Ziyi Yang, Shaohua Shi, Li Fu, Aiping Lu, Tingjun Hou, and Dongsheng Cao. Matched molecular pair analysis in drug discovery: Methods and recent applications. *Journal of Medicinal Chemistry*, 66(7): 4361–4377, 2023. URL https://doi.org/10.1021/acs.jmedchem.2c01787.

[16] 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), 2023. URL https://doi.org/10.1038/s41467-023-41948-6.

[17] Jun Xia, Lecheng Zhang, Xiao Zhu, Yue Liu, Zhangyang Gao, Bozhen Hu, Cheng Tan, Jiangbin Zheng, Siyuan Li, and Stan Z. Li. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 64774–64792. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cc83e97320000f4e08cb9e293b12cf7e-Paper-Conference.pdf.

[18] Zhixiang Cheng, Hongxin Xiang, Pengsen Ma, Li Zeng, Xin Jin, Xixi Yang, Jianxin Lin, Yang Deng, Bosheng Song, Xinxin Feng, Changhui Deng, and Xiangxiang Zeng. MaskMol: Knowledge-guided molecular image pre-training framework for activity cliffs with pixel masking. *bioRxiv*, 2024. URL https://doi.org/10.1101/2024.09.04.611324.

[19] Javed Iqbal, Martin Vogt, and Jürgen Bajorath. Prediction of activity cliffs on the basis of images using convolutional neural networks. *Journal of Computer-Aided Molecular Design*, pages 1–8, 2021. URL https://doi.org/10.1007/s10822-021-00380-y.

[20] Junhui Park, Gaeun Sung, SeungHyun Lee, SeungHo Kang, and ChunKyun Park. ACGCN: Graph convolutional networks for activity cliff prediction between matched molecular pairs. *Journal of Chemical Information and Modeling*, 62(10):2341–2351, 2022. URL https://doi.org/10.1021/acs.jcim.2c00327.

[21] Dragos Horvath, Gilles Marcou, Alexandre Varnek, Shilva Kayastha, Antonio de la Vega de León, and Jürgen Bajorath. Prediction of activity cliffs using condensed graphs of reaction representations, descriptor recombination, support vector machine classification, and support vector regression. *Journal of Chemical Information and Modeling*, 56(9):1631–1640, 2016. URL https://doi.org/10.1021/acs.jcim.6b00359.

[22] Michael Dablander, Thomas Hanser, Renaud Lambiotte, et al. Exploring qsar models for activity-cliff prediction. *Journal of Cheminformatics*, 15:47, 2023. doi: 10.1186/s13321-023-00708-w. URL https://doi.org/10.1186/s13321-023-00708-w.

[23] 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. URL https://doi.org/10.1021/acs.jcim.2c01073.

[24] Brian Kulis. Metric Learning: A Survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2013. URL http://dx.doi.org/10.1561/2200000019.

[25] Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 1st edition, 2011. URL https://doi.org/10.1007/978-3-642-14267-3.

[26] Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS Central Science*, 3(4):283–293, 2017. URL https://doi.org/10.1021/acscentsci.6b00367.

[27] Daniel Vella and Jean-Paul Ebejer. Few-shot learning for low-data drug discovery. *Journal of Chemical Information and Modeling*, 63(1):27–42, 2023. URL https://doi.org/10.1021/acs.jcim.2c00779.

[28] 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 J. Vanschoren and S. Yeung, editors, *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/file/8d3bba7425e7c98c50f52ca1b52d3735-Paper-round2.pdf.

[29] Nataša Tagasovska, Vladimir Gligorijević, Kyunghyun Cho, and Andreas Loukas. Implicitly guided design with PropEn: Match your data to follow the gradient, 2024. URL https://arxiv.org/abs/2405.18075.

[30] Sebastian Johann Wetzel, Kevin Ryczko, Roger Gordon Melko, and Isaac Tamblyn. Twin neural network regression. *Applied AI Letters*, 3(4), 2022. URL http://dx.doi.org/10.1002/ail2.78.

[31] Sebastian J Wetzel, Roger G Melko, and Isaac Tamblyn. Twin neural network regression is a semi-supervised regression algorithm. *Machine Learning: Science and Technology*, 3(4):045007, 2022. URL http://dx.doi.org/10.1088/2632-2153/ac9885.

[32] Mohamed Karim Belaid, Maximilian Rabus, and Eyke Hüllermeier. Pairwise difference learning for classification, 2024. URL https://arxiv.org/abs/2406.20031.

[33] Michael Tynes, Wenhao Gao, Daniel J. Burrill, Enrique R. Batista, Danny Perez, Ping Yang, and Nicholas Lubbers. Pairwise difference regression: A machine learning meta-algorithm for improved prediction and uncertainty quantification in chemical search. *Journal of Chemical Information and Modeling*, 61(8): 3846–3857, 2021. URL https://doi.org/10.1021/acs.jcim.1c00670.

[34] Zachary Fralish, Ashley Chen, Paul Skaluba, and Daniel Reker. DeepDelta: predicting ADMET improvements of molecular derivatives with deep learning. *Journal of Cheminformatics*, 15(1):101, 2023. URL https://doi.org/10.1186/s13321-023-00769-x.

[35] Zachary Fralish and Daniel Reker. Finding the most potent compounds using active learning on molecular pairs. *Beilstein Journal of Organic Chemistry*, 20:2152–2162, 2024. URL https://doi.org/10.3762/bjoc.20.185.

[36] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. URL https://doi.org/10.1023/A:1010933404324.

[37] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. URL http://doi.acm.org/10.1145/2939672.2939785.

[38] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. URL https://doi.org/10.1021/ci100050t.

[39] Greg Landrum. RDKit: Open-source cheminformatics software, 2016. URL https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

[40] Weihua Hu*, Bowen Liu*, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJlWWJSFDH.

[41] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Velickovic. Principal neighbourhood aggregation for graph nets. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. URL https://proceedings.neurips.cc/paper/2020/file/99cad265a1768cc2dd013f0e740300ae-Paper.pdf.

[42] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022. URL http://dx.doi.org/10.1038/s42256-022-00447-x.

[43] Benjamin Kaufman, Edward C. Williams, Carl Underkoffler, Ryan Pederson, Narbe Mardirossian, Ian Watson, and John Parkhill. COATI: Multimodal contrastive pretraining for representing and traversing chemical space. *Journal of Chemical Information and Modeling*, 64(4):1145–1157, 2024. URL https://doi.org/10.1021/acs.jcim.3c01753.

[44] Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C Lim, and Prudencio Tossou. Gotta be SAFE: A new framework for molecular design, 2023. URL https://arxiv.org/abs/2310.10773.

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

[46] Hans-Christian Ehrlich and Matthias Rarey. Systematic benchmark of substructure search in molecular graphs - from Ullmann to VF2. *Journal of Cheminformatics*, 4(1):13, 2012. URL https://doi.org/10.1186/1758-2946-4-13.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

A Appendix

A.1 Models

We used the following models to evaluate the effectiveness of the SQRL approach:

Baselines. All baseline models were trained on top of Morgan count fingerprints of size 2048, radius 2, and including chirality. For RF, XGBoost, and KNN, RDKit features from [39] were concatenated to Morgan fingeprints.

- **Random Forest (RF)**: An ensemble learning method using decision trees. Scikit-learn implementation was used [47] with default parameters.
- **XGBoost**: A gradient boosting framework, optimized for efficiency and performance. Implementation from Ref. 37 was used with default parameters.
- **k-Nearest Neighbors (KNN)**: A non-parametric method based on the similarity between data points. Scikit-learn implementation was used [47] with $k = 1$ to compare with our relative setting where we evaluate with respect to the closest training data point.
- **Multi-Layer Perceptron (MLP)**: A standard feedforward neural network.

Graph neural networks. GNNs were trained end-to-end, with their learned representations followed by MLP layers for standard or relative predictions.

- **AttentiveFP**: A graph neural network that uses graph attention mechanisms to capture atomic interactions [4].
- **GINE**: Graph Isomorphism Network with Edge features, enhancing the model's ability to capture bond information [40].
- **PNA**: Principal Neighbourhood Aggregation, a GNN architecture designed to be stable under permutations [41]. The implementation used in this paper follows Ref. 28.
- **MolCLR**: A pre-trained GNN using contrastive learning for molecular representation. In our experiments, we did not freeze the MolCLR layers, allowing them to be fine-tuned along with the rest of the model, according to the procedure of the original authors [42].

Transformer-based models. For transfomer-based models, we utilized the pre-trained embeddings as fixed feature extractors, followed by MLP for task-specific predictions.

- **SAFE-GPT**: A chemical language model trained on linear molecular notation which has been adapted for autoregressive tasks [44].

- **Uni-Mol**: A 3D-aware transformer model trained on molecular conformations using SE(3) equivariant operations [45]. Due to computationally expensive conformation generation,
- **COATI**: A multi-modal generative model combining 2D and 3D molecular information through contrastive learning [43]. To obtain embeddings for our applications, the text encoder part of the model was used (Barlow_Closed).

Due to the computationally expensive conformer generation step, Uni-Mol-SQRL was evaluated on the following subset of MoleculeACE tasks (23/30): `CHEMBL2971_Ki`, `CHEMBL2835_Ki`, `CHEMBL219_Ki`, `CHEMBL228_Ki`, `CHEMBL238_Ki`, `CHEMBL1862_Ki`, `CHEMBL218_EC50`, `CHEMBL231_Ki`, `CHEMBL235_EC50`, `CHEMBL287_Ki`, `CHEMBL2147_Ki`, `CHEMBL2047_EC50`, `CHEMBL4203_Ki`, `CHEMBL2034_Ki`, `CHEMBL1871_Ki`, `CHEMBL4792_Ki`, `CHEMBL244_Ki`, `CHEMBL234_Ki`, `CHEMBL239_EC50`, `CHEMBL262_Ki`, `CHEMBL4616_EC50`, `CHEMBL3979_EC50`, `CHEMBL4005_Ki`.

A.2 Hyperparameter selection

See Table 2 for optimized hyperparameters used for each model in a standard and SQRL setting.

Table 2: Model hyperparameters.

Model	Linear Sizes	GNN Layers	Dropout	Learning Rate	Batch Size	Other
MLP	[256, 256]	-	0.0	1e-4	128	-
MLP-SQRL	[512, 256]	-	0.2	1e-5	64	-
AttentiveFP	[256]	3	0.2	1e-4	128	timesteps: 2
AttentiveFP-SQRL	[128]	3	0.0	1e-3	256	timesteps: 4
GINE	[128]	4	0.0	1e-4	64	-
GINE-SQRL	[256, 128]	5	0.0	1e-3	64	-
PNA	[128]	4	0.0	1e-4	64	-
PNA-SQRL	[128]	8	0.0	1e-5	256	-
MolCLR	[512]	-	0.0	1e-4	128	-
MolCLR-SQRL	[128]	-	0.0	1e-4	128	-
COATI	[256]	-	0.0	1e-5	64	-
COATI-SQRL	[256, 128]	-	0.0	1e-3	128	-
SAFE-GPT	[256, 128]	-	0.0	1e-3	32	-
SAFE-GPT-SQRL	[256, 128]	-	0.0	1e-4	128	-
Uni-Mol	[256, 128]	-	0.0	1e-3	128	-
Uni-Mol-SQRL	[128]	-	0.0	1e-4	128	-

A.3 MoleculeACE Dataset

The MoleculeACE dataset provided by van Tilborg et al. is curated from ChEMBL v29 and contains potency measurements for 30 targets. Activity cliff molecules were defined as pairs of molecules with greater than 90% substructure, scaffold, or SMILES similarity and greater than 10-fold activity difference. Figure 2 shows the number of training samples for each task in the MoleculeACE dataset along with the number of activity cliff molecules present in each task. The train-test split was performed by clustering molecules into 5 clusters and stratified splitting using the activity cliff label.

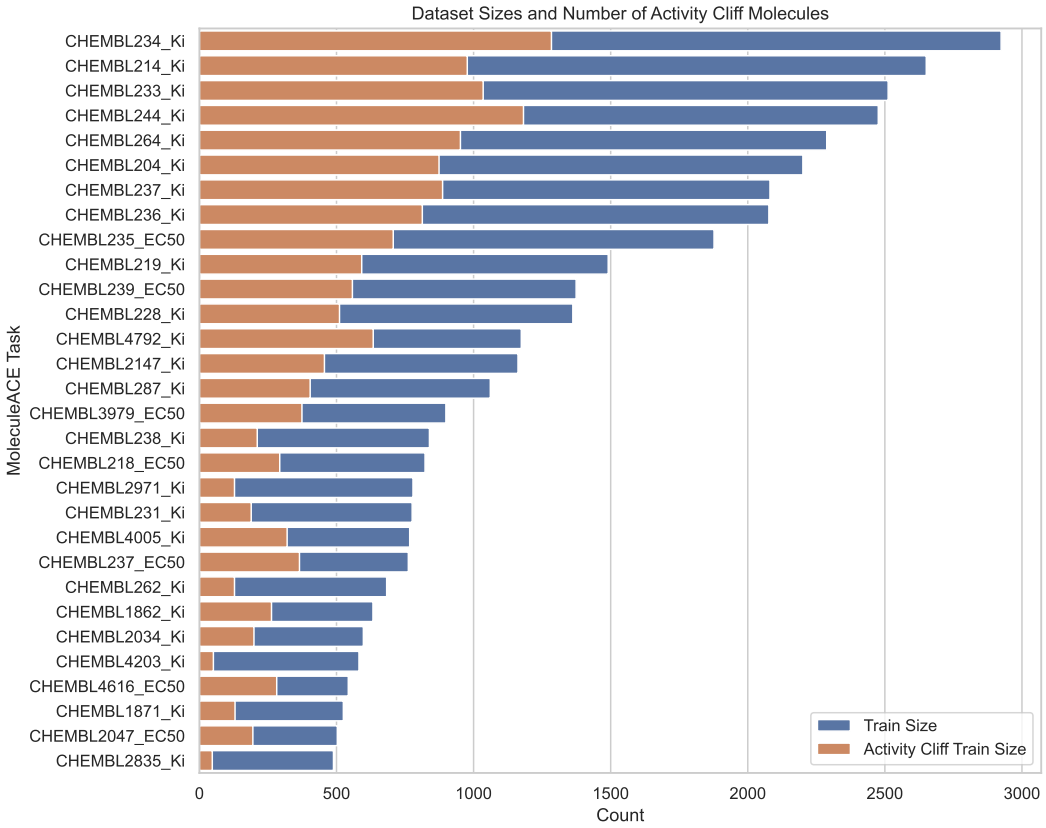

Figure 2: Training data sizes for each task in MoleculeACE.

A.4 Examples of molecular pairs

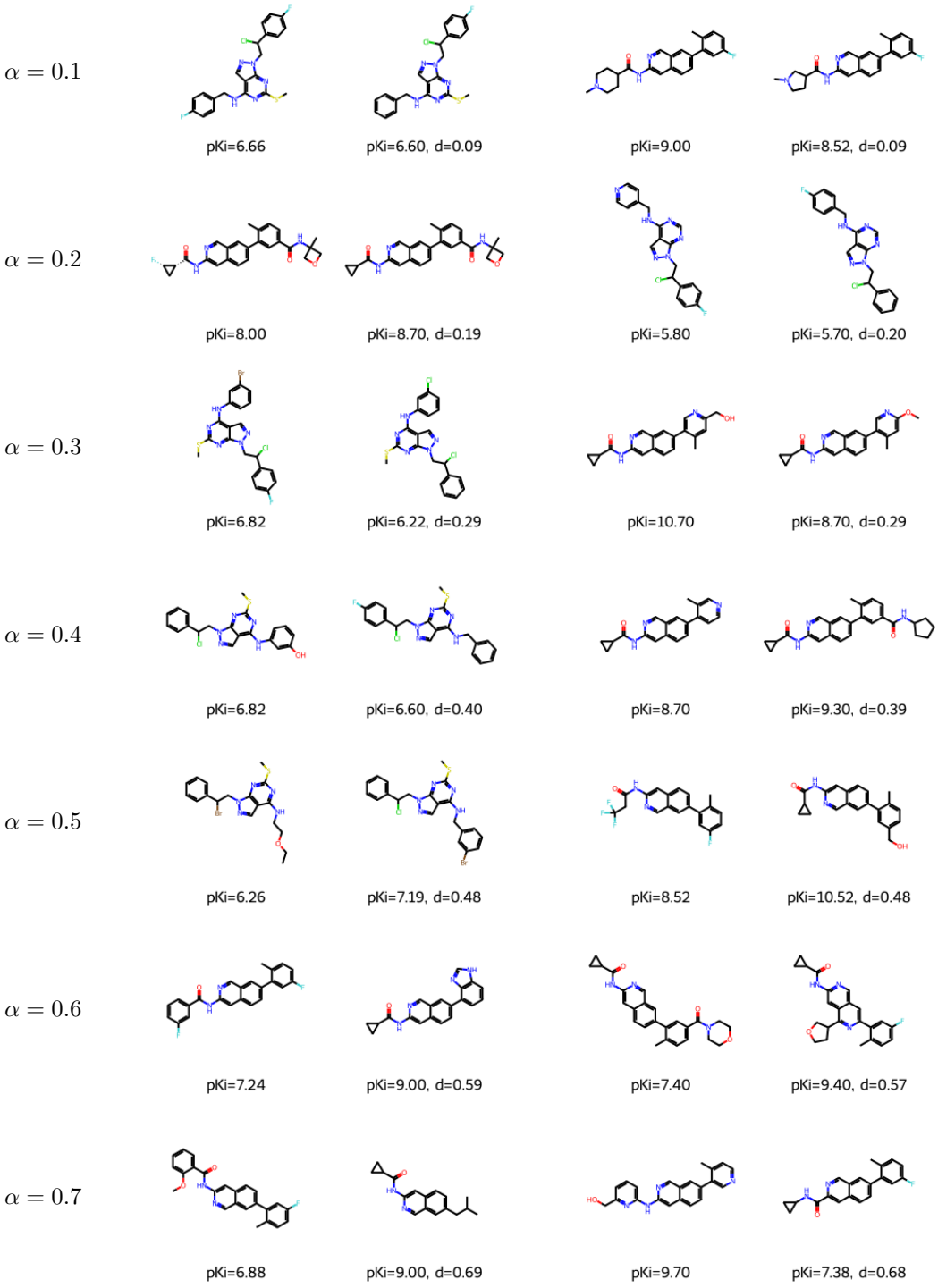

Figure 3: Molecular pairs obtained by the data matching procedure described in Section 3 at different Tanimoto distance thresholds α for MoleculeACE task `CHEMBL1862_Ki`.

Figure 4: Molecular pairs of *activity cliff molecules* obtained by the data matching procedure described in Section 3 at different Tanimoto distance thresholds α for MoleculeACE task `CHEMBL1862_Ki`.

A.5 Additional results

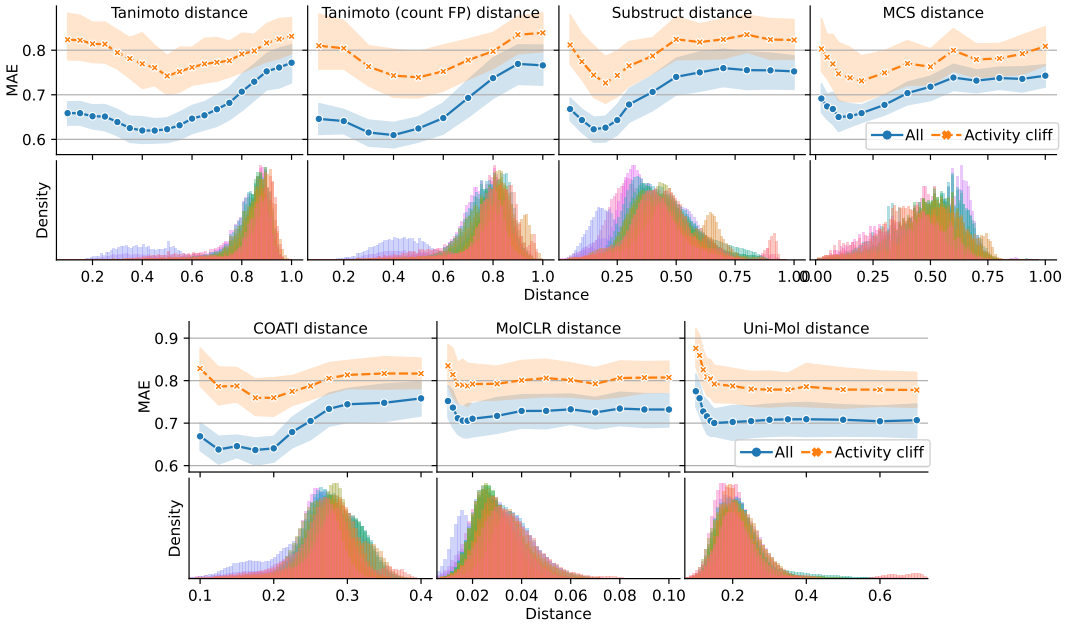

Figure 5: **Leveraging local structural information enhances predictive performance.** MAE (↓) as a function of distance threshold α for several additional distance metrics compared to Figure 1, as well as pairwise distance distributions for each metric. *Tanimoto*: Tanimoto (Jaccard) distance between binary Morgan fingerprints. *Tanimoto (count FP)*: Tanimoto (Jaccard) distance between count-based Morgan fingerprints. *Substruct*: Tanimoto (Jaccard) distance between substructure count vectors using a list of 1242 predefined substructures from Ehrlich and Rarey [46]. *MCS*: Distance metric based on maximum common substructure (MCS) defined as $1 - 2N_{\text{MCS}}/(N_i + N_j)$ where N_{MCS} is the number of atoms in the MCS, N_i is the number of atoms in molecule i, and N_j is the number of atoms in molecule j. *COATI [43], MolCLR [42], Uni-Mol [45]*: Euclidean distances between neural network embeddings obtained with these pre-trained models.

Table 3: **Mean Absolute Error (MAE) (↓) comparison of predictive performance.** MAE with standard deviation across tasks for Standard and SQRL methods using Tanimoto distance with $\alpha = 0.7$. **Bold values** indicate the better performing method (lower MAE) for each model and dataset. Uni-Mol was evaluated on a smaller subset of tasks due to computational constraints (see Appendix A.1.)

	MoleculeACE		**MoleculeACE-Cliff**		**Internal Targets**	
Model	Standard	SQRL	Standard	SQRL	Standard	SQRL
Baselines						
XGBoost	**0.54 ± 0.07**	0.59 ± 0.08	**0.64 ± 0.12**	0.71 ± 0.13	**0.41 ± 0.18**	0.55 ± 0.19
RF	**0.54 ± 0.07**	0.57 ± 0.08	**0.63 ± 0.10**	0.68 ± 0.12	**0.41 ± 0.18**	0.50 ± 0.21
KNN	**0.69 ± 0.12**	0.77 ± 0.13	**0.83 ± 0.15**	0.88 ± 0.18	**0.71 ± 0.37**	0.95 ± 0.35
MLP	0.94 ± 0.22	**0.67 ± 0.10**	0.96 ± 0.20	**0.79 ± 0.12**	0.67 ± 0.41	**0.65 ± 0.20**
GNNs						
AttentiveFP	0.80 ± 0.12	**0.58 ± 0.11**	0.86 ± 0.12	**0.65 ± 0.16**	0.56 ± 0.24	**0.50 ± 0.15**
GINE	1.14 ± 0.24	**0.64 ± 0.09**	1.23 ± 0.32	**0.72 ± 0.14**	0.75 ± 0.13	**0.47 ± 0.19**
PNA	0.82 ± 0.21	**0.65 ± 0.09**	0.85 ± 0.18	**0.75 ± 0.14**	0.63 ± 0.27	**0.58 ± 0.26**
MolCLR	0.91 ± 0.12	**0.62 ± 0.09**	0.95 ± 0.13	**0.72 ± 0.14**	0.62 ± 0.39	**0.50 ± 0.21**
Transformers						
COATI	0.71 ± 0.10	**0.65 ± 0.09**	**0.74 ± 0.20**	0.75 ± 0.14	**0.59 ± 0.30**	0.61 ± 0.20
Uni-Mol	0.94 ± 0.26	**0.69 ± 0.09**	0.92 ± 0.22	**0.77 ± 0.14**	0.62 ± 0.41	**0.49 ± 0.48**
SAFE-GPT	0.76 ± 0.12	**0.68 ± 0.11**	**0.78 ± 0.10**	0.80 ± 0.13	**0.55 ± 0.22**	0.60 ± 0.26

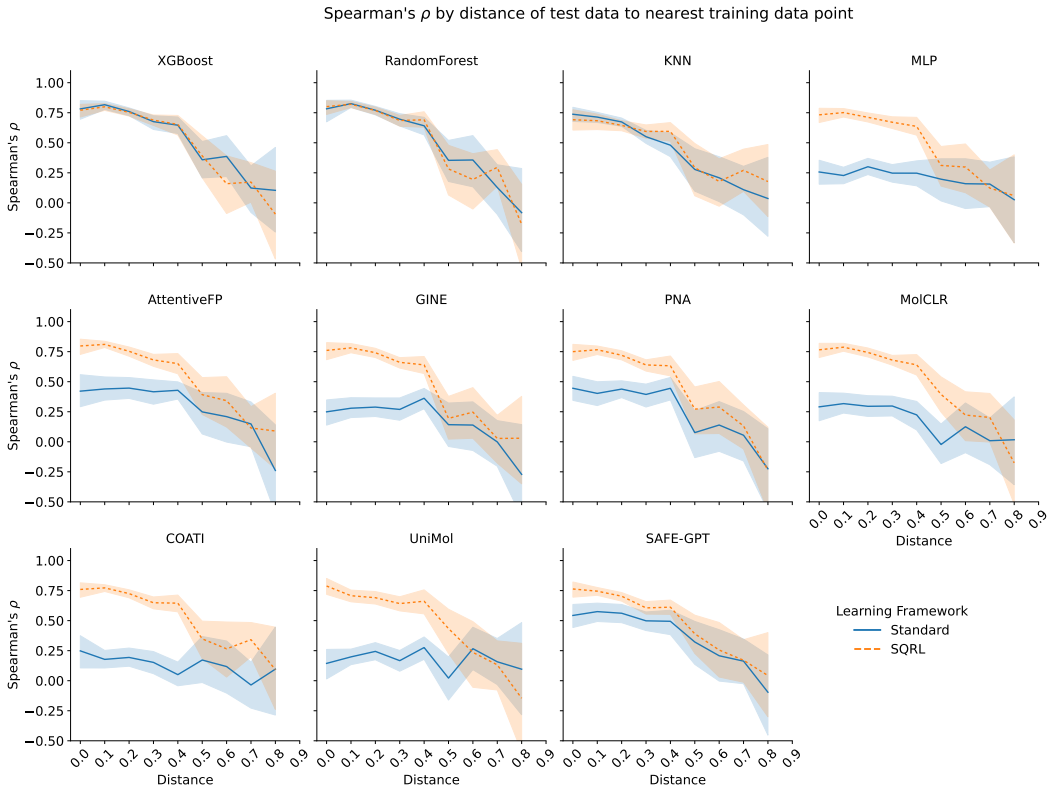

Figure 6: **SQRL enhances local molecular consistency.** Spearman's rank correlation coefficient (↑) plotted as a function of the distance between test points and their nearest neighbors in the training set. SQRL-trained models that benefit from this training strategy demonstrate the most significant performance gains for test points with close neighbors, while generally maintaining comparable performance to standard-trained models for more distant points. Models in this plot were trained with the distance threshold $\alpha = 0.7$ using Tanimoto distance.