Towards Learning Activity Cliff-Aware Molecular Representations

Anonymous Authors[1]

Abstract

Current deep learning based methods for molecular property prediction show pronounced shortcomings when predicting molecular properties in the presence of activity cliffs (AC): pairs of structurally similar molecules with significant differences in potency. We investigate how inductive biases of increasing complexity, from simple Multilayer Perceptrons (MLPs) to self-supervised models, impact the learning of representations from Extended-connectivity Fingerprints (ECFPs). Leveraging the Matched Molecular Pair (MMP) abstraction, we explore various pre-training schemes designed to capture AC relationships. While simple models remain competitive, we show extensive differences and avenues for potential improvement in performance across different inductive bias choices and pre-training strategies, paving the way for AC-aware and consequently, chemically robust model design. Code available online at [Footnote to be inserted after review process].

1. Introduction

Despite widespread adoption in chemical modeling, deep learning methods do not yet show a clear advantage in predicting molecular properties from chemical structure over classical methods, especially in the presence of Activity Cliffs (ACs) (Mayr et al., 2018), (van Tilborg et al., 2022). Pervasive in most popular datasets, these molecules present a difficult problem of molecular representation; we expect structurally similar molecules to exhibit similar bioactivity properties. However, in the case of ACs, a structural change at a single atomic position between a pair of otherwise identical molecules is enough to induce abrupt changes in biochemical activity. In addition to the former, due to the immense size and diversity of chemical compound space, accounting for activity cliffs across datasets and chemically relevant tasks is a formidable challenge (Deng et al., 2023). This warrants a shift from exhaustive labeling or clustering of specific molecular phenomena, ACs being a notable example, towards increasingly generalizing, coarser-grained approaches that capture subtle differences in molecular representation, while preserving the broader notion of "standard" chemical function (Jiang et al., 2021). A simple, deep learning-based, representation-inductive bias combination that has shown consistent performance across the chemical space is the use of Extended-connectivity Fingerprints (ECFPs), expert-designed descriptors, with Multi-Layer Perceptrons (MLPs), known for their flexibility as universal function approximators (Steshin, 2023). Their combined success highlights the potential of merging domain knowledge with adaptable learning architectures to achieve robust property prediction.

While various inductive bias and representation pairs across levels of abstraction have been explored, none specifically focus on the disparate activity-structure representation phenomenon of ACs. In the broader molecular property prediction domain, contextual enrichment of representations and substructure aware losses have been shown to enhance prediction (Schimunek et al., 2023), (Amara et al., 2023). Inspired by these approaches, we center our present work on the Matched Molecular Pair (MMP) abstraction, where molecules differ at a single atomic position and may or may not consequently exhibit AC-like properties. We leverage the "Mix" subset of the ACNet dataset for MMPs to integrate pre-training as we climb in the model complexity space (Zhang et al., 2023). We also explore the use of self-supervised methods, building upon successful applications in related chemical domains, to further enhance our models' ability to capture ACs (Magar et al., 2022), (Lin, 2023).

Contributions concretely, we make the following contributions:

- We empirically demonstrate the effectiveness of multiple pre-training schemes across chemical data regimes with varying AC prevalence. We assess the impact of pre-training objectives and model architectures on downstream AC prediction performance.
- We compare multiple loss functions that operate exclu-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

sively on latent or unmodified representations derived from ECFPs. This approach preserves a valid molecular view throughout the process, without fragmenting into substructures, while still incorporating information about ACs into the model.

- We extend the traditional contrastive loss to consider both the agreement between reconstructed ECFPs and their original molecular structures and the differences between substructures in molecular pairs, conclusive to ACs. This novel loss function, the SiamACLoss, explicitly encourages the model to learn representations that are sensitive to AC relationships.

2. Methodology

Our exploration is based on two central assumptions:

1. **Activity cliffs, defined by the Matched Molecular Pair abstraction, provide a sufficient augmentation.** MMPs in AC settings introduce a controlled "noising" operation at the differing position, while the conserved scaffold acts as a stable reference point. This creates two alternative views of a chemical structure, capturing the inherent variability associated with ACs. Such augmentation is often crucial in the proper functioning of semi and self-supervised methods.

2. **Working with latent representations of ECFPs allow for transfer to the broader chemical space while capturing AC relationships** We focus on AC relationships in the latent space, avoiding direct modification of ECFPs to preserve molecular validity throughout the training process.

2.1. Exploring inductive biases of increasing complexity

Inspired by earlier work showing that neural networks learn statistics of increasing complexity, we gradually increased model parametrization and introduced additional inductive biases to assess their effectiveness in capturing activity cliffs (Refinetti et al., 2022), (Tamura et al., 2023). Pre-training methods were trained to minimize the validation loss. Model training was stopped early if no improvement was obtained after ten non-consecutive epochs. All methods involving pre-training are compressed down to a 256-dimension latent vector to have a fixed point for posterior MLP evaluation.

2.2. MLP based methods

Our exploration begins with a baseline of MLPs using radius 4 ECFPs of varying sizes as input. We progressively incorporate more complex architectures with an increasing number of parameters, pre-training and normalization.

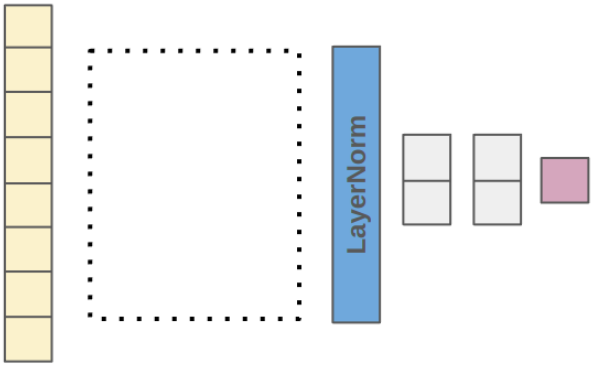

Figure 1. The HotSwapEncoderMLP: general MLP used to evaluate all obtained pre-training embeddings. Starting from an input layer of fingerprint size 2048, a frozen, pre-trained encoder is then placed directly afterwards. The obtained embeddings are then layer normalized and passed through a 2 layer MLP to obtain a final molecular activity regression or AC label classification prediction.

- **Varying fingerprint size:** We increase the input parameter space training a simple MLP baseline with 256, 1024 and 2048 initial ECFP size.

- **Fine-Grained Encoder:** A stepwise-halving linear layer encoder gradually reduces ECFP size down to a 256-dimensional latent representation. This assesses whether a more complex encoder captures chemical nuance in greater detail than raw fingerprints.

- **Pre-training:** Models are pre-trained on randomized single molecules from the ACNet dataset using classification and reconstruction training objectives, using the same MLP setup as in the former step and a dataset composed of MMPs loaded sequentially in a random fashion, without specifying a paired representation.

- **Layer Normalization:** We investigate the effect of layer normalization on the learned, pre-trained embeddings before transferring to the final MLP for property prediction (Ba et al., 2016).

2.3. Activity-cliff based methods

Next, we incorporate inductive biases that explicitly leverage the pairwise nature of ACs using the MMP abstraction in a classification setting.

- **Joint MLP:** ACNet-obtained paired molecular representations are concatenated and trained with classification or reconstruction objectives, using the stepwise-halving encoder.

- **Siamese networks:** Networks with shared weights are introduced to learn joint embeddings of molecular pairs with both the Manhattan and Cosine distances as an association metric.

2.4. AC latent guided methods

Finally, we further explore models that operate uniquely on latent representations or reconstructions, across different supervision settings.

- **Siamese autoencoders:** Each molecule in an MMP is passed through an identical encoder-decoder architecture, processed independently in the same forward pass. The symmetric loss is then computed between the losses obtained from evaluating input-reconstruction pairs with Binary cross entropy (BCE).

- **SiamACLoss:** A Siamese autoencoder is trained using a novel loss, combining reconstruction loss with a contrastive term that encourages similarity for non-AC pairs and dissimilarity for AC pairs.

- **Negative cosine similarity:** Minimizes the negative cosine similarity (NCS) between reconstructions as a training objective for a Siamese autoencoder, aiming to push known AC pairs apart in latent space, given their dissimilar activities.

- **SimSiam:** A simple, self-supervised approach using Siamese networks with a stop gradient operation trained exclusively on positive pairs (Chen & He, 2020).

- **Positive set training:** Inspired by the former, we assess the impact of using only known AC pairs during pre-training for Siamese autoencoder methods and SimSiam. See Table 1.

Table 1. Amount of MMPs in the "Mix" subset of the ACNet dataset by known AC pairs.

	AC	Not AC	Total MMPs
Full set	16,607	261,760	278,367
Positive set	16,607	0	16,607

The contrastive loss is given by (Chopra et al., 2005)

$$L = \frac{1}{2N}\sum_{i=1}^{N}\left[y_i d_i^2 + (1-y_i)\max(0, m-d_i)^2\right]$$

Which we then extend into the SiamACLoss, as shown in Equation 4, given by:

Table 2. Hyperparameters for baseline and evaluation MLPs.

Hyperparameter	Value / Description
Task	Classification, Regression
ECFP radius	4
Input Features	2048, 1024, 256
Hidden Features	100
Hidden Layers	2
Output Features	1
Dropout	0.2
Layer Activation	ReLU
Optimizer	Adam
Learning Rate	0.001
Batch Size	128
Scheduler	ReduceLROnPlateau
Factor	0.1
Patience	10
Loss	BCEWithLogitsLoss, RMSE

Table 3. Characteristics of the chosen subset of MoleculeACE ChEMBL IDs for evaluations.

ID	Description	Abbreviation
234	Most Molecules	Max mol
2835	Fewest Molecules	Min mol
4616	Most Cliff Partners	Max AC
4203	Fewest Cliff Partners	Min AC
2047	Highest SMILES Similarity	SMILES
264	Highest Scaffold Similarity	Scaffold
4792	Highest Substructure Similarity	Sub

$$L_{\text{SiamAC}} = L_{\text{recon1}} + L_{\text{recon2}} + \lambda L_{\text{con}} \tag{1}$$

$$L_{\text{recon1}} = \text{BCEWithLogitsLoss}(\text{recon1}, x_1) \tag{2}$$

$$L_{\text{recon2}} = \text{BCEWithLogitsLoss}(\text{recon2}, x_2) \tag{3}$$

$$L_{\text{con}} = \frac{1}{2N}\sum_{i=1}^{N}\left[y_i\|\text{recon1}_i - \text{recon2}_i\|^2 + (1-y_i)\max(0, m - \|\text{recon1}_i - \text{recon2}_i\|)^2\right] \tag{4}$$

Where λ is a hyperparameter that can be tuned to change the influence of the contrastive term.

3. Experiments

We evaluate the learned representations on a 7 dataset subset of the broader 30 ChEMBL subsets contained in the MoleculeACE benchmark. These are chosen for their variety of AC-relevant characteristics across molecular similarity schemes and target data. See Table 3. We train a 2-layer, 100-unit MLP on these subsets using the learned embeddings through the aforementioned pre-training schemes, due

Table 4. $RMSE_{cliff}$ values across models in order of increasing complexity across 7 chosen MoleculeACE subsets

Model	Max mol	Min mol	Max AC	Min AC	SMILES	Scaffold	Sub
MLP 256	1.3683	1.3119	0.9737	1.2556	0.9959	1.3225	1.5282
MLP 1024	1.4005	1.3496	0.9917	1.1565	1.0563	1.3454	1.4834
MLP 2048	1.4005	1.3540	0.9805	1.1943	1.2072	1.3239	1.5106
halfstepMLP 1024	1.3972	**1.1961**	0.9807	1.2207	1.1139	1.3335	1.4909
halfstepMLP 2048	1.3820	1.2385	0.9741	1.1440	1.1319	1.3054	1.5596
PT MLP 2048	1.4005	1.3540	0.9805	1.1943	1.2072	1.3239	1.5106
PT AE MLP 2048	1.3825	1.3299	0.9934	1.1689	1.0192	1.3414	1.4514
PT AE MLP ln 2048	1.3954	1.3225	0.9515	1.15	0.9490	1.3316	1.4629
Joint 1024	1.3640	1.3912	0.9416	1.2506	1.1065	1.3858	1.5491
Joint AE 1024	1.3574	1.3536	0.9770	1.3646	1.0553	1.3538	1.5111
Siamese Manhattan	1.3578	1.2081	0.9560	1.3204	1.3900	1.1396	1.8156
Siamese Cosine	1.3466	1.3585	0.9112	1.3035	1.2737	**1.0724**	1.4777
Siamese AE Naive	1.4005	1.3540	0.9805	1.1943	1.2072	1.3239	1.5106
Siamese AE SiamAC	1.3649	1.2192	1.0167	1.1779	1.1099	1.3408	1.5262
Siamese AE SiamAC +	1.3564	1.3517	1.0570	1.1824	1.1007	1.3306	1.4923
Siamese AE NCS	1.0616	1.2880	**0.8797**	1.1508	**0.8975**	1.1980	**1.2565**
Siamese AE NCS +	**1.0593**	1.4185	0.8841	**1.1380**	0.8984	1.1959	1.2769
SimSiam	1.2900	1.3178	0.9579	1.2897	1.0797	1.3150	1.4694
SimSiam +	1.3231	1.3200	0.9670	1.1818	1.1359	1.3296	1.4184

to their consistent performance across initial fingerprint sizes after Autoencoder compression (Ilnicka & Schneider, 2023). See Figure 1 for a schematic and Table 2 for the evaluation hyperparameters. We consider RMSE, $RMSE_{cliff}$, and AUROC as performance metrics (van Tilborg et al., 2022). RMSE and AUROC follow their traditional formulations. $RMSE_{cliff}$ is a version of the RMSE metric that is extended to exclusively take into account molecules with known AC partners as follows, as proposed in MoleculeACE (van Tilborg et al., 2022). See Equation 5.

$$\text{RMSE}_{\text{cliff}} = \sqrt{\frac{\sum_{j=1}^{n_c}(\hat{y}_j - y_j)^2}{n_c}} \tag{5}$$

Where $\hat{y}_j$ is the predicted regression activity value of the jth compound, y_j the reported experimental value and n_c represents the total number of activity cliff compounds considered.

4. Results

Our results demonstrate that model performance varies significantly across tasks and datasets, with AC-centered inductive biases and pre-training schemes showing a pronounced advantage in the regression setting, while the distinction is less clear in the classification case. Notably, the dataset with highest substructure similarity, ChEMBL 4792 is the most challenging across settings. While the addition of layer normalization was only marginally advantageous in an MLP setting, it was retained in all subsequent AC methods.

Given that $RMSE_{cliff}$ is an appropriate, challenging proxy for RMSE values our analysis is centered on the obtained $RMSE_{cliff}$ and AUROC values (van Tilborg et al., 2022). The obtained values corresponding to each metric for all models across the 7 ChEMBL subsets can be consulted in Tables 4 and 5, respectively.

Molecular Property Prediction through regression

The Siamese AE NCS models, in both their Full and Positive set training variants achieve the lowest $RMSE_{cliff}$ values, with the exception of datasets with the fewest total molecules or highest scaffold similarity. Learning representations based on exclusively positive AC pairs is sufficient for the evaluation subsets with the most total molecules or when ACs have the least amount of cliff partners. Although these methods still perform relatively well in the case of high scaffold similarity, a direct association with the siamese encoder utilizing cosine similarity as an association function is preferred. Due to the variability in performance observed across siamese encoder pre-training methods, representations learned through direct association seem especially susceptible to the properties of each downstream evaluation subset. The positive-only bias is less helpful with fewer total molecules, where the smoother HalfstepMLP1024 encoder shows superior performance. This suggests that training on both positive and negative AC pairs allows to learn representations that incorporate general, non-AC specific chemical knowledge, which is beneficial in low-data settings.

AC Identification through binary classification

Downstream classification, involving the identification of

Table 5. AUROC values across models in order of increasing complexity across 7 chosen MoleculeACE subsets

Model	Max mol	Min mol	Max AC	Min AC	SMILES	Scaffold	Sub
MLP 256	0.7977	0.8856	0.8032	0.8733	0.8326	0.8274	0.7443
MLP 1024	0.7917	0.8850	0.8400	0.9417	0.8490	0.8405	0.7672
MLP 2048	0.8071	**0.9251**	0.8477	0.9395	**0.8967**	0.8441	0.7955
halfstepMLP 1024	0.7994	0.9074	**0.8479**	0.8908	0.8818	0.8520	0.7648
halfstepMLP 2048	**0.8235**	0.8972	0.8460	0.8931	0.8454	0.8640	0.7726
PT MLP 2048	0.6303	0.8387	0.6660	0.4859	0.6354	0.6233	0.5705
PT AE MLP 2048	0.7627	0.8856	0.7808	0.6567	0.8587	0.8221	0.7154
PT AE MLP ln 2048	0.7804	0.9006	0.7916	0.7212	0.8564	0.8215	0.7330
Joint 1024	0.7999	0.8434	0.8273	**0.9429**	0.8726	0.8411	0.7121
Joint AE 1024	0.7281	0.8795	0.8305	0.8572	0.8715	**0.8479**	0.7734
Siamese Manhattan	0.5633	0.5453	0.5621	0.4774	0.4823	0.5088	0.5391
Siamese Cosine	0.4771	0.6644	0.4161	0.5158	0.6010	0.5021	0.5004
Siamese AE Naive	0.8071	**0.9251**	0.8477	0.9395	**0.8967**	0.8441	0.7955
Siamese AE SiamAC	0.7896	0.8747	0.7740	0.8258	0.8597	0.8392	0.7455
Siamese AE SiamAC +	0.7783	0.8618	0.8331	0.8162	0.8649	0.8157	**0.8230**
Siamese AE NCS	0.5475	0.8713	0.6025	0.5752	0.5572	0.5940	0.5212
Siamese AE NCS +	0.5000	0.8741	0.6335	0.5741	0.5382	0.5540	0.4498
SimSiam	0.6528	0.8244	0.7161	0.6154	0.7651	0.7690	0.6880
SimSiam +	0.7306	0.8788	0.8024	0.7585	0.8600	0.7924	0.7207

ACs through binary labels favors large fingerprint sizes and comparatively larger, parameter-heavy networks. The simple, larger MLP variants excel. Pre-training starts being beneficial when there are fewer AC partners included per molecule in the downstream dataset or in the challenging high scaffold similarity dataset, where joint classification methods leverage the concatenated joint representation. Pre-training generally shows no distinct advantage in this setting, where a naive Siamese AE equals the performance of the aforementioned methods, likely due to it capturing an expressive general representation of chemical space. Pre-training is superior in one particular scenario: The SiamAC loss performs the best in the high substructure similarity dataset. The contrastive term seems to be an asset when AC relationships are determined by the Tanimoto coefficient on ECFPs, which is designed to capture "global" differences between molecules by considering similarities between the entire set of substructures they're composed of (Cereto-Massagué et al., 2015).

5. Conclusion

In this work we demonstrate that pre-training models while explicitly accounting for the structural correspondence in activity cliffs through the matched molecular pair abstraction improves downstream regression and classification performance. Minimizing the negative cosine similarity loss as a training objective in unsupervised regimes, particularly when using siamese autoencoders, effectively models structurally similar compounds with dissimilar activities. Different similarity values pose distinct challenges, with increasing difficulty in the order of datasets with the highest SMILES, Scaffold and Substructure similarity. Datasets containing more mean cliff partners in a sample favors regression, while high SMILES similarity seems to favor classification. Methods without extensive pre-training remain performant across various schemes, notably, when dataset sizes are relatively large. Our findings provide insights for incorporating AC-aware components into model design, ultimately improving molecular property prediction that accounts for this challenging class of compounds.

6. Limitations

No exhaustive hyperparameter tuning was performed to keep training settings across modelling choices as similar as possible. A less challenging random split of 80:10:10 of the "Mix" subset was considered, as opposed to the proposed *target split*, originally proposed by the ACNet authors (Zhang et al., 2023). In consequence, the learned representations may not fully model AC-relevant features. Finally, our study is limited to seven subsets, which is not fully representative of chemical compound space.

7. Future work

Exploring and interpreting the latent spaces obtained by different pre-training strategies, metric learning, informed target splits and the inclusion of target information could all prove to be fruitful avenues for AC-aware model enhancement. Additionally, employing rich and diverse featurization

schemes beyond ECFPs and extending the models into similarly conceptualized practical settings such as targeted lead optimization could prove to be insightful.

Acknowledgements

We would like to thank the anonymous reviewers of the LXAI workshop at ICML 2024 and the MoML 2024 attendees for their valuable feedback.

References

Amara, K., Rodríguez-Pérez, R., and Jiménez-Luna, J. Explaining compound activity predictions with a substructure-aware loss for graph neural networks. *J. Cheminform.*, 15(1):67, July 2023.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. July 2016.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, January 2015.

Chen, X. and He, K. Exploring simple siamese representation learning. November 2020.

Chopra, S., Hadsell, R., and LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pp. 539–546 vol. 1, 2005. doi: 10.1109/CVPR.2005.202.

Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D., and Wang, F. A systematic study of key elements underlying molecular property prediction. *Nat. Commun.*, 14(1): 6395, October 2023.

Ilnicka, A. and Schneider, G. Compression of molecular fingerprints with autoencoder networks. *Molecular Informatics*, June 2023. doi: 10.1002/minf.202300059. URL https://doi.org/10.1002/minf.202300059.

Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1), February 2021. doi: 10.1186/s13321-020-00479-8. URL https://doi.org/10.1186/s13321-020-00479-8.

Lin, J. Y.-Y. Molsiam: Simple siamese self-supervised representation learning for small molecules. In *NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development*, 2023.

Magar, R., Wang, Y., and Barati Farimani, A. Crystal twins: self-supervised learning for crystalline material property prediction. *Npj Comput. Mater.*, 8(1), November 2022.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.*, 9 (24):5441–5451, 2018.

Refinetti, M., Ingrosso, A., and Goldt, S. Neural networks trained with SGD learn distributions of increasing complexity. 2022.

Schimunek, J., Seidl, P., Friedrich, L., Kuhn, D., Rippmann, F., Hochreiter, S., and Klambauer, G. Context-enriched molecule representations improve few-shot drug discovery. 2023.

Steshin, S. Lo-Hi: Practical ML drug discovery benchmark. 2023.

Tamura, S., Miyao, T., and Bajorath, J. Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. *J. Cheminform.*, 15 (1):4, January 2023.

van Tilborg, D., Alenicheva, A., and Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of Chemical Information and Modeling*, 62(23):5938–5951, December 2022. doi: 10.1021/acs.jcim.2c01073. URL https://doi.org/10.1021/acs.jcim.2c01073.

Zhang, Z., Zhao, B., Xie, A., Bian, Y., and Zhou, S. Activity cliff prediction: Dataset and benchmark, 2023.

Appendix A

Dataset descriptions were obtained based on the metastudy included in MoleculeACE's supplementary information, available in Table S4 of the original publication's supporting information (van Tilborg et al., 2022). An aggregate of relevant information for our analyses can be observed in Table 6.

Table 6. Specific cliff partner and mean max similarity values for the chosen MoleculeACE subsets. Train/Test and Cliffs represent total molecules per subset. Training set values provided unless otherwise specified.

ID	Mean partners	Test partners	Sub	Scaffold	SMILES	Train/Test	Cliffs
234	2.73	24	0.81	0.95	0.95	2923/734	1150/291
2835	1.43	0	0.82	0.91	0.96	489/126	36/10
4616	5.51	0	0.82	0.94	0.96	543/139	262/68
4203	1.25	0	0.67	0.93	0.92	582/149	51/13
2047	2.96	2	0.81	0.94	0.97	503/128	195/50
264	2.82	17	0.81	0.96	0.95	2288/574	865/219
4792	2.37	15	0.84	0.92	0.96	1174/297	610/153

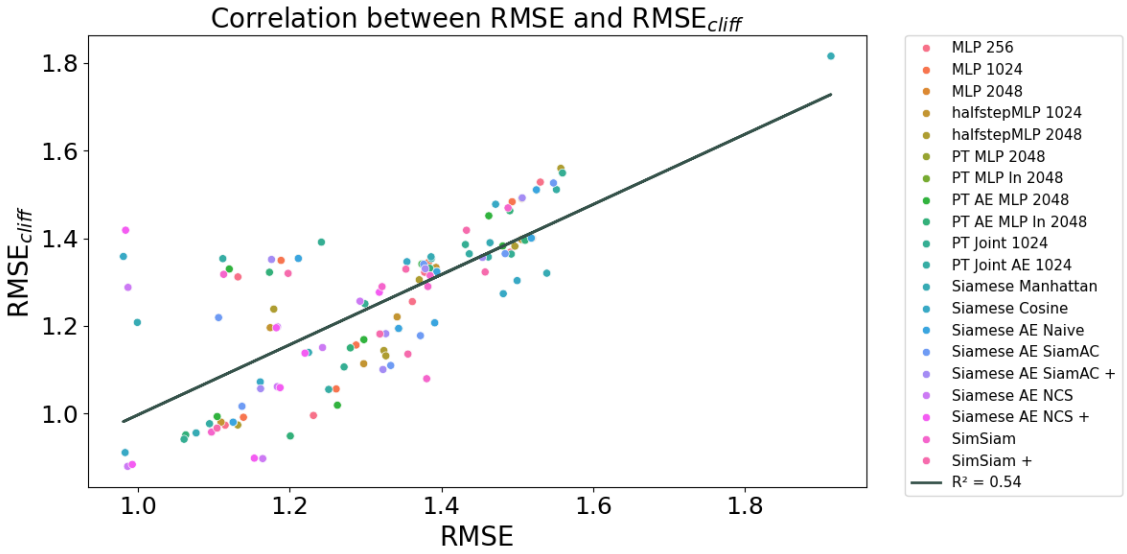

Figure 2. Correlation between the obtained RMSE and RMSE$_{\text{cliff}}$ metrics on downstream performance across the seven chosen MoleculeACE subsets per model.

Table 7. Difference between the top performing AC aware method and the top performing MLP based method. Negative values mean AC-aware model exhibits a lower value and thus perform better for RMSE based metrics, the inverse for AUROC.

Model	Max mol	Min mol	Max AC	Min AC	SMILES	Scaffold	Sub
RMSE$_{\text{cliff}}$	-0.309	0.012	-0.0718	-0.006	-0.0984	-0.233	-0.1949
RMSE	-0.2973	-0.1396	-0.0799	-0.0599	-0.0475	-0.2096	-0.1698
AUROC	-0.0164	0	-0.0002	0.0012	0	-0.0161	0.0275

Table 8. Top performing model per dataset per metric

	$\mathbf{RMSE_{cliff}}$	RMSE	AUROC
Max mol	Siamese AE NCS +	Siamese AE NCS +	Halfstep MLP 2048
Min mol	Halfstep MLP 1024	Siamese Cosine	MLP 2048
Max AC	Siamese AE NCS	Siamese Cosine	Halfstep MLP 1024
Min AC	Siamese AE NCS +	Siamese AE NCS +	Joint 1024
SMILES	Siamese AE NCS	Siamese AE NCS +	Siamese AE Naive
Scaffold	Siamese Cosine	Siamese Cosine	Halfstep MLP 2048
Sub	Siamese AE NCS	Siamese AE NCS	Siamese AE SiamAC+

Table 9. RMSE across models and training objectives per dataset.

Model	Max mol	Min mol	Max AC	Min AC	SMILES	Scaffold	Sub
MLP 256	1.4908	1.1317	1.1149	1.3617	1.2313	1.3779	1.5305
MLP 1024	1.5198	1.1886	1.1390	1.2875	1.2613	1.3831	1.4933
MLP 2048	1.5186	1.2114	1.1254	1.3436	1.3912	1.3940	1.5253
halfstepMLP 1024	1.5058	1.1743	1.1095	1.3416	1.2976	1.3929	1.5061
halfstepMLP 2048	1.4969	1.1790	1.1315	1.3242	1.3268	1.3708	1.5575
PT MLP 2048	1.5186	1.2114	1.1254	1.3436	1.3912	1.3940	1.5253
PT MLP ln 2048	1.5186	1.2114	1.1254	1.3436	1.3912	1.3940	1.5253
PT AE MLP 2048	1.4809	1.1203	1.1044	1.2978	1.2629	1.3743	1.4625
PT AE MLP ln 2048	1.5104	1.1733	1.0629	1.2800	1.2008	1.3847	1.4902
Joint 1024	1.4922	1.2418	1.0607	1.2992	1.2717	1.4317	1.5597
Joint AE 1024	1.4619	1.1118	1.0943	1.4368	1.2514	1.3868	1.5517
Siamese Manhattan	1.3864	0.9991	1.0766	1.5390	1.4641	1.2249	1.9137
Siamese Cosine	1.3546	**0.9807**	**0.9830**	1.4999	1.4816	**1.1612**	1.4715
Siamese AE Naive	1.5186	1.2114	1.1254	1.3436	1.3912	1.3940	1.5253
Siamese AE SiamAC	1.4841	1.1062	1.1371	1.3724	1.3332	1.3767	1.5478
Siamese AE SiamAC +	1.4542	1.1760	1.1616	1.3267	1.3230	1.3787	1.5066
Siamese AE NCS	**1.1836**	0.9868	0.9863	1.2433	1.1644	1.1839	**1.2927**
Siamese AE NCS +	1.1873	0.9836	0.9924	**1.2201**	**1.1533**	1.1823	1.3182
SimSiam	1.3822	1.1129	1.0969	1.3218	1.3806	1.3852	1.4879
SimSiam +	1.4577	1.1979	1.1042	1.3189	1.3559	1.3532	1.4331