nature portfolio

Peer Review File

Activity-cliff awareness enables robust graph learning for molecular property prediction

Corresponding Author: Professor Wanxiang Shen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Overall Assessment
This manuscript proposes a novel contrastive learning approach, ACANET, for molecular property prediction to address the challenge of activity cliffs, where structurally similar molecules show very different activities. By introducing an inductive bias on activity cliffs awareness to commonly used graph neural network models, ACANET shows improved performance compared to standard bioactivity prediction as well as strong performance in pharmacokinetic and safety-relevant properties in a mixed collection of datasets.
Overall, the manuscript is well organized with clear sections, nice figures and sufficient details on experiments and results. However, since there are still major concerns to be addressed, my recommendation would be Major Revisions.

Strengths
1. The manuscript addresses an important yet often overlooked issue, activity cliffs, in molecular property prediction.
2. The manuscript develops a novel contrastive learning approach, which combines regular regression loss and a triplet soft margin loss in metric learning to enhance a model's ability in distinguishing structurally similar molecules with differing activities. Moreover, the authors introduced three tunable hyperparameters (alpha, cu and cl) to adjust the AC-awareness extent, in case of specific needs for certain datasets.
3. The manuscript evaluates the proposed approach in a wide range of 52 datasets, which are further grouped into four tasks, namely 1). predicting activity in 9 low-sample size and narrow scaffold (LSSNS) datasets (regression) 2). 30 high-sample size and mixed scaffold (HSSMS) datasets (regression) 3). 3 matched molecular pair (MMP) datasets (classification) and 4) 10 ADMET datasets (delta prediction).

Major Concerns
1. Despite the efforts of addressing activity cliffs through a contrastive learning approach to distinguish molecules that have different activities with similar structures in the latent space, the definitions of 1) similar structures and 2) different activities are not clearly stated in the first place, which entails two major questions 1) which structures are considered to be similar, with the same Bemis-Murcko scaffold or achieves a Tanimoto similarity above some threshold? 2) how large an activity difference should be. The reason why I list this point as the first major concern is because without clear definitions, we would not know the applicable scenarios of the proposed ACANet.
1.1. To follow up on this point, please consider calculating the number of scaffolds and internal diversity for the LSSNS and HSSMS datasets. Also, consider reporting the distribution of properties in each dataset.
1.2. Additionally, it would be helpful to report the number of molecules observed with activity-cliffs issue. This can reflect the extent of activity-cliffs challenge in each dataset, which further justifies the selection of specific dataset for evaluation.
1.3. For instance, on Page 5 Line 151, a medium-size PPARδ is used to validate the effectiveness of AC-awareness. I wonder why this dataset was chosen in the first place and if the same observation persists in other datasets?
2. Next major concern is regarding model evaluation to address activity cliffs. Can you provide more details on the splitting method? Did you adopt the scaffold split or just random split? Since the proposed approach aims to address activity cliffs in molecular property prediction, a direct evaluation would be examining prediction performance on molecules with activity-cliffs issues at inference. However, it is unclear how activity cliffs are present in the training-validation-test sets for each dataset. In other words, if there are not sufficient molecules with activity-cliffs issues, would it still be fair to attribute the reduced RMSE by ACANet to AC awareness?
2.1. A follow-up question would be why ADMET datasets are included for evaluating ACANet? Unlike activity (specifically,

binding affinity), some PK and safety-related properties cannot be attributed to chemical structures alone. In this case, the assumption for activity/property-cliffs is undermined. Even with improvement in delta prediction, how to justify the contribution of AC awareness over model architectures and hyperparameters tuning, etc?
3. The third major concern is that, despite the shortcoming in addressing activity cliffs, learning a latent space based on chemical structures is still the basis for molecular representation learning such that property prediction can be made for unseen molecules. By introducing the AC awareness during model training, anchors-positives are pulled close and anchors-negatives are pushed away, as nicely shown in Figure 6. In this case, would this inductive bias related to activity labels interfere with the latent space viewed from the structural perspective? If yes, how to balance between the two perspectives to maximize the predictive power?
3.1. A related question in Figure 6 is that are the two triplets from the training set or test set? They seem to have the same scaffold from a visual check. It would be interesting to show examples from 1) both training set and test set and 2) different scaffolds to see if the latent space keeps organized 1) from training to inference and 2) across different chemical spaces.
4. The ablation study on the impact of cliff cutoffs and the awareness factor can be very helpful, if expanded to more datasets. Ideally, it would be of high value if some general rules can be established between tunable hyperparameters (alpha, cu and cl) for AC-awareness and datasets (e.g. degree of scaffold diversity, prevalence of activity cliffs as mentioned in Discussion) because then the value of ACANet can become evident.

Minor Concerns
1. Consider replacing “differing activities” when discussing activity cliffs because a very modest change in activity can always be called differing but they do not necessarily constitute an activity cliff.
2. The authors nicely pointed out in “Overview of datasets and tasks” (Page 4) that on LSSNS datasets are representative during the lead identification phase and MMP datasets are relevant in lead optimization phase. On the contrary, HSSMS datasets are more relevant to a different stage, mid-stage drug discovery. Here, please consider clarifying what it means by mid-stage drug discovery to alleviate confusion. Is it hit identification where different scaffolds are involved in a broader chemical space?
3. Please check throughout the manuscript to ensure all abbreviations are spelled in their first occurrence. For example, on Page 6 Line 164 “ACANet model uses PNA as its default backbone…”, there is no annotation for PNA until the Methods section.
4. On Page 21 Line 423, the Methods section has a title as”Online Methods”. Is there a specific reason why it is called “online” methods? Besides, the Methods section can be cleaned a bit. For example, consider removing numbering for subsections under “Activity cliff awareness (ACA) loss function” in Pages 23-25 and subsubsections under “Analysis of label incoherence across chemical and latent spaces” in Page 26. Additionally, consider numbering all equations in the manuscript.
5. Consider examining the use of “significantly” or “significant”. Usually in scientific publications, it usually means a result from a statistical test that is not likely to happen by chance. However, when describing difference in labels or losses, there are quite some sentences like “when their activities differ significantly” in Line 34-Line 35 (Page 2), “final MAE loss is not significantly different” in Line 155 (Page 5) and “achieved significantly lower RMSE values” in Line 162 (Page 5). Please consider using “significantly” or “significant” with more caution.
6. While reading the manuscript, I found that sometimes I would get confused by M (conditional triplets) and M’ (mined triplets). Consider using more straightforward abbreviations and provide more illustrations on M and M’ in the first place.
7. Consider adding the number of molecules for the overall test set and the cliff test set for the 30 HSSMS benchmark datasets in Figure 7 or provide them in Supplementary Information.
8. Throughout the experiments, there are cases1) where MAE loss or RMSE loss is used and 2) a certain dataset is used to illustrate the impact of AC awareness or related hyperparameters. It would be nice if there is a figure to summarize the experiment design. Also, consider elaborating more about why some experiments are designed in such a way in the Discussion section. For example, why Figures 3&4 focus on PPARδ whereas Figures 5&6 focus on BRAF.

(Remarks on code availability)
I checked the code repository and the authors have nicely documented their project in github.

Reviewer #2

(Remarks to the Author)
Summary

This manuscript proposes the novel framework for effective learning of activity-cliff (AC) awareness in modeling molecular activity. In general, ACs make molecular activity prediction challenging in traditional computational chemistry or quantitative structure-activity relationship (QSAR). This challenge arises because compounds with similar structural features can nonetheless display markedly different activities.

In this manuscript, AC-awareness was achieved by the novel contrastive learning-based graph neural network (ACNet) with triplet soft margin (TSM) loss between high-value activity cliff triplets (HV-ACTs). This TSM loss is defined on a group of three molecules; an anchor molecule, a positive and a negative molecule to an anchor, each. Based on the established knowledge, TSM loss reduces the distance between the latent feature of an anchor compound (A) and its corresponding positive (P) while increasing the distance between the anchor (A) and its negative (N). positive To evaluate this pipeline, the authors constructed four scenarios of activity cliffs across 52 molecular datasets, including ADMET property prediction. The ablation studies also showed the effectiveness of the proposed method.

Advantages

The authors clearly conveyed the importance of the potential challenge caused by activity-cliff in molecular modeling or drug development. Especially, traditional QSAR-based method may struggle to these HV-ACTs. In addition, the contrastive learning with TSM loss is appropriately designed in accordance with the motivation of this research and the basic property of activity cliff. The experiments were conducted in various situations and lots of datasets.

Limitations and questions

Although the authors validated superiority over various tasks compared with other methods, they may not be sufficient to explicitly validate the effectiveness of the propose method. Because 1) in regard to Figure 7, the lowest error values of ACANet may come from the well-optimized parameters in deep networks. How were other networks (PNA, GIN, MPNN, …) optimized for these comparisons? 2) Is ACANet clearly have an advantage over traditional (non-machine-learning-based) QSAR methods?

Questions on datasets and data embedding: 4) Is there any cases that lots of positives or negatives, corresponding to one anchor? 5) Is there any dataset having class or data imbalances? For example, is it possible that one anchor may have up to 50 positives and less than 5 negatives? If it can be possible, then what is the potential problem caused by this imbalance and how ACANet can handle?

Additionally, I would suggest that the authors should supplement more background about the definition and nature of activity cliffs, including that the semantic or functional similarity between anchor and its positive one is naturally close to that between anchor and its negative one.

Additional ask for the evaluation

1. In Figure 4, the number of HV-ACTs decreases during both training and inference process, when the model was trained with TSM losses. Is it possible to show several hard cases of HV-ACT which remains the last time of training or inference?
2. As I mentioned in Limitations and questions, is it possible to show another evaluation to validate advantages of ACANet over other QSAR methods? Either quantitative or case-study-based approach may be good.

(Remarks on code availability)
I can run the source code without any problem.

Reviewer #3

(Remarks to the Author)
In the manuscript entitled “Activity Cliff-Informed Contrastive Learning for Molecular Property Prediction” by Shen et al. a novel method for activity cliff awareness is introduced into a graph based deep learning model. Activity cliffs are an important aspect of drug discovery and it has lately been shown by numerous groups that even nonlinear deep learning models cannot accurately capture these effect. Here, a novel method is introduced in order to be able to reflect the activity cliff also in the latent space representation. The method is of high relevance, however before publication I have a number of comments which would show the real value of the inclusion of the AC awareness into the model.

Main
•1st paragraph: Nonadditivity has only been mentioned on the side in the introduction, however, this is basically the same effect analyzed in a different way. Also here, different people have been working on the topic, i.e.
o https://pubs.acs.org/doi/full/10.1021/acs.jcim.9b00631
o https://link.springer.com/article/10.1186/s13321-021-00525-z
o https://pubs.acs.org/doi/full/10.1021/acsomega.2c02738
o https://link.springer.com/article/10.1007/s10822-024-00566-0
•2nd paragraph, line 31: How is structurally similar defined?
•l 63: BRAF dataset – please add a reference
•Ll 67-68: “more informative molecular representation”: What do you mean here? In what sense are these representations more informative and for whom?
•L 71: ACANet first time mentioned, so it would be good to know what the abbreviation meant, short explanation if Net also stands for something or if this is just the name you gave it.
•Last paragraph in general: A lot of results are mentioned here already, I would prefer to have those in the results section instead of here
o Ll 66, 71, 72 especially

Results
•Line 86: How are the triplets identified? How are triplets assembles for test data, i.e. prospective use cases? This part needs more explanation.
•L 99: “number of mined HV-ACTs”: how is this number determined, is there a max/min you set in your methods – This needs further explanations.
o How is decided how many triplets are considered?
o What is the effect on the model accuracy?
o Extension: Can you use the results from the triplets for uncertainty estimates for the test compounds?
•Table S1: please indicate the number of activity cliffs, their percentage on the overall dataset. How did you split into

train/test data?
•Table S3 – How do you assign AC here, especially if you have more non cliff mmps than total compounds? Can compounds be in a cliff and not in a cliff pair?
•Table S4 – all tables are structured differently. Please try to have the same content as far as possible in all of them, i.e. number and percentage of AC, splitting into train and test...
•L 137: Can you always generate a triplet for any anchor compound? How do you define an activity cliff based on what similarity measure?
o L. 141 – again: what is structural similarity?
•L 154, Figure 3c: I would like to see the performance difference more differentiated: How does the model perform on additive data, how on cliff data points? Is the improvement really due to improved performance on cliff data?
•L 161: “external validation and test set” - How did you split the data into train/test/validation?
•Ll 167-169 “Performance improvements were observed across the board, with test RMSE reductions of 5.81%, 5.48%, 4.63%, 169 and 6.62%” – Do they improve on similar data points or are they improving in different chemical areas?
•Ll 188-189 Figure S2: The color scheme of the RMSD scale seems quite arbitrary and different for every plot. You could select 0.3 as a good value and ~0.5 as a bad one. Thus leading to a better comparison of the influence of the M parameter, which seems very dataset dependent.
•Ll 189-191: “This study highlights the importance of selecting appropriate cliff cut-off parameters to efficiently mine ACTs and enhance cliff awareness and model performance.” – What are your learning from this? Any recommendations on how to best set these parameters?
•ll 203-204: “Our results suggest that selecting appropriate cliff parameters and alpha can enable efficient cliff awareness” – Any indications or learning from this?
•Ll 221-222_ “Superior overall performance” - Superior sounds as if they are much better, but the performance of both looks rather on par. Please comment and/or reword to a more realistic wording.
•Ll 225 – full paragraph on “Using AC-awareness enhances activity-coherent latent molecular representations.” - I am not jet convinced that this method really learns about activity cliffs. Currently it seems more like the method is learning from what it already knows are activity cliffs. In order to see, what it is truly learning, it would need a different test dataset, which contains activity cliffs, where neither positive nor negative example are in the training data. I have not seen any such test data set in the manuscript.
•l 273 “RMSEcliff” – please explain what is meant with this, how is it calculated.
•Ll 286-287: “ACANet outperformed the SOTA SVM for 23 and 21 of the 30 datasets in RMSEall and RMSEcliff” - Are there any ACs that are only in the test set for which you could show performance improvement?
•L 313, Table S9 - Can you please comment on the significance of the differences in the presented results?
•Ll 330-331: “Moreover, ACANET achieved the lowest cross-validation RMSE across all other seven DMET datasets” - Out of which some are also on par with other methods. It would be interesting to get significance values here and otherwise reword it to better reflect the actual results.

Minor
l 113: “difficultThe” – missing space and point.
L 344: ref 21 duplicated – please check references for correctness

(Remarks on code availability)
I partially tested the code. I could install the package without further trouble - thanks for that! There is no code example with a given dataset to run ACANet immediately. This should be included to give people an easy start.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The reviewer appreciates the tremendous efforts that the authors have put to address all the major and minor concerns. The revised manuscript is greatly improved with much more clarity. The Supplementary Information also adds detailed explanations to the experiments.

Furthermore, the rebuttal file provides an *impressive*, in-depth discussion on addressing activity cliffs in molecular property prediction, which can be highly beneficial for future studies.

Given these considerations, my recommendation would be Accept. Kudos to the authors!

(Remarks on code availability)
The authors have nicely documented their project in github. I can install clsar and run ACANet from my end.

Reviewer #3

(Remarks to the Author)

Thank you for the comprehensive and well-executed revision and response. Interesting to see your preliminary exploration of uncertainty estimation - thanks for reporting it!

A few minor things:
- consistent first-use expansions (e.g., Activity-Cliff Awareness Network (ACANet); Principal Neighborhood Aggregation (PNA); GCN, GAT, GIN) in the Main text.
- Standardize phrasing and hyphenation (e.g., AC-aware; principal neighborhood aggregation vs principal neighbourhood aggregation) and RMSEall/RMSEcliff typography.
- I found a few typos/spacing (e.g., missing space after “space. By”; “network (GIN)”).
- Thanks for the explanation with the heatmap coloring. I would recommend adding a brief note in the Supplementary caption why heatmap color scales are dataset-specific to aid reader interpretation.

(Remarks on code availability)
I did not have time to retry it, but I looked into the changes the authors mentioned - thanks!

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

RESPONSE TO REVIEWER COMMENTS

OVERALL RESPONSE TO THE REVIEWERS: 2

RESPONSE TO REVIEWER #1: 4

MAJOR CONCERNS 5

MINOR CONCERNS 30

REVIEWER #1 (REMARKS ON CODE AVAILABILITY): 34

RESPONSE TO REVIEWER #2: 35

LIMITATIONS AND QUESTIONS 35

QUESTIONS ON DATASETS AND DATA EMBEDDING: 40

ADDITIONAL ASK FOR THE EVALUATION 42

REVIEWER #2 (REMARKS ON CODE AVAILABILITY): 49

RESPONSE TO REVIEWER #3: 50

MAIN 50

RESULTS 53

MINOR 70

REVIEWER #3 (REMARKS ON CODE AVAILABILITY): 70

Overall response to the reviewers:

We sincerely thank the reviewers for their careful evaluations and constructive suggestions. Guided by these comments, we have substantially revised the manuscript and Supplementary Information to clarify the mechanism of activity-cliff awareness (ACA), to strengthen the evidence on latent-space organisation, and to provide more systematic ablation studies. The key changes are summarised below.

Global performance on independent HSSMS test sets

We added a figure comparing RMSE for models with and without ACA across all HSSMS datasets that possess independent test sets. The ACA-aware models consistently achieve lower test-set RMSE, confirming that the gains are not confined to a few targets or to random splits, but generalise across a broad benchmark.

Latent-space analyses on training and test data

To make the effect of ACA on representation learning more transparent, we now provide PCA projections of compound embeddings on the HSSMS training set, coloured by pChEMBL values. Without ACA, the latent space shows fragmented and discontinuous activity landscapes, whereas with ACA the activity gradients become smooth and coherent, indicating fewer latent-space cliffs. In addition, a detailed JAK1 (CHEMBL2835_Ki) test-set case study combines tmap (structure space) with PCA of GNN embeddings, demonstrating that ACA reorganises the latent space so that anchors cluster with potency-matched positives and separate from negatives even across different scaffolds.

Label Incoherence Index (LII) and its temporal evolution

We introduce the Label Incoherence Index (LII) as a quantitative measure of local label coherence. The workflow first finds k-nearest neighbours (k = 10) in ECFP4 space or in latent spaces with/without ACA, and then computes LII from the pChEMBL labels of each compound and its neighbours. New figures track the evolution of LII over training epochs on both training and test sets. Across all datasets, ACA persistently reduces the latent-space LII below the ECFP4 baseline and the non-ACA model, providing a clear quantitative link between ACA and improved local label organisation.

Effect of structural-similarity gates on HV-ACT mining and performance

To address questions about how activity-cliff triplets are mined, we added experiments that vary the structural-similarity threshold used as a gate. For several HSSMS targets, we jointly report the number of mined high-value ACTs per epoch and the corresponding test RMSE. The results show that looser gates (lower similarity thresholds) yield many more informative ACTs and consistently lower RMSE, whereas very strict gates severely suppress ACT mining and degrade predictive accuracy. This clarifies the trade-off and explains why we avoid overly rigid structural constraints.

Ablation of an additional structure gate in ACANet

We further compare the original ACA-aware ACANet with a variant that includes an extra “traditional” structure gate. Across 30 single-scaffold HSSMS targets and multiple training-dynamics snapshots, the gated variant never outperforms the original ACANet and often

converges more slowly or to higher RMSE, while the non-ACA baseline remains the weakest. These results show that the improvements are driven by ACA itself rather than by adding more hand-crafted structural gates.

Text, figure and notation refinements

Finally, we have revised the main text and captions to standardise terminology (e.g. definition of activity-cliff triplets, HV-ACTs and LII), clearly distinguish training vs. test analyses, and explain all colour schemes and axes. This should make the methodology, experiments and conclusions easier to follow and to reproduce.

Taken together, these revisions directly address the reviewers' main concerns regarding mechanism, latent-space behaviour, and ablation, and we believe they substantially strengthen the manuscript.

Formatting of this response

In the detailed point-by-point replies:

Red bold text is used to highlight newly added or substantially revised key sentences, figures and tables in the revised manuscript.

Blue bold text is used to highlight previously existing key points, figures and tables that we explicitly refer back to in our responses.

Square brackets "[…]" are used to cross-reference items within this response package, such as specific response to reviewer comments, figures and tables.

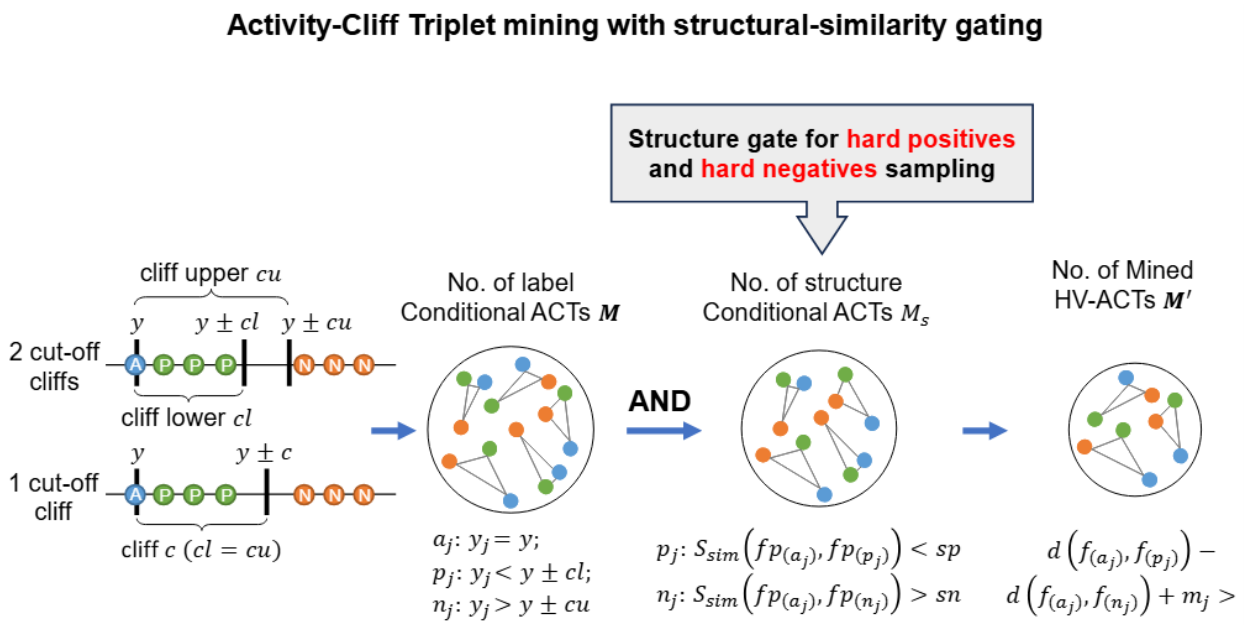
Response to Reviewer #1:

Overall Assessment
This manuscript proposes a novel contrastive learning approach, ACANET, for molecular property prediction to address the challenge of activity cliffs, where structurally similar molecules show very different activities. By introducing an inductive bias on activity cliffs awareness to commonly used graph neural network models, ACANET shows improved performance compared to standard bioactivity prediction as well as strong performance in pharmacokinetic and safety-relevant properties in a mixed collection of datasets. Overall, the manuscript is well organized with clear sections, nice figures and sufficient details on experiments and results. However, since there are still major concerns to be addressed, my recommendation would be Major Revisions.
Strengths
1. The manuscript addresses an important yet often overlooked issue, activity cliffs, in molecular property prediction.
2. The manuscript develops a novel contrastive learning approach, which combines regular regression loss and a triplet soft margin loss in metric learning to enhance a model's ability in distinguishing structurally similar molecules with differing activities. Moreover, the authors introduced three tunable hyperparameters (alpha, cu and cl) to adjust the AC-awareness extent, in case of specific needs for certain datasets.
3. The manuscript evaluates the proposed approach in a wide range of 52 datasets, which are further grouped into four tasks, namely 1). predicting activity in 9 low-sample size and narrow scaffold (LSSNS) datasets (regression) 2). 30 high-sample size and mixed scaffold (HSSMS) datasets (regression) 3). 3 matched molecular pair (MMP) datasets (classification) and 4) 10 ADMET datasets (delta prediction).

RESPONSE: Thank you for your positive evaluation and for recognizing the importance and novelty of our ACANet approach. We are glad that the manuscript's focus on addressing the challenge of activity cliffs in molecular property prediction was appreciated, and that the novel contrastive learning approach (ACANet) was well-received. We also acknowledge the importance of the experiments across a wide range of datasets and the introduction of the AC-awareness mechanism, which was a key strength of the manuscript.

In response to your major concerns below, in this revision we have extended ACALoss with an optional structural similarity-based triplet refining logic (see **Response Figure 1 / Supplementary Figure 10**, the "**Structural Gate**" for triplet mining). Concretely, we have added a detailed description of this extension in **Methods,** updated the algorithm pseudocode (**Supplementary Methods 1**), and included a new ablation study in **Results.** And release a new version of ACANet and loss (v2.0.0) on Pypi: https://pypi.org/project/clsar/2.0.0/.

We hope these additions clarify how structure-aware sampling interacts with activity-cliff mining and why label-only sampling remains the strongest baseline. Thank you again for highlighting this important aspect.



Response Figure 1 (new Supplementary Figure 10). Activity-cliff triplet mining with a traditional structure gate. Label-conditioned ACTs (M) are first obtained via activity cut-offs; a structural-similarity gate retains only hard positives ($S_{sim} < sp$) and hard negatives ($S_{sim} > sn$), yielding structure-conditioned ACTs (M_s) and, ultimately, the high-value triplet pool M' used for training.

Major Concerns

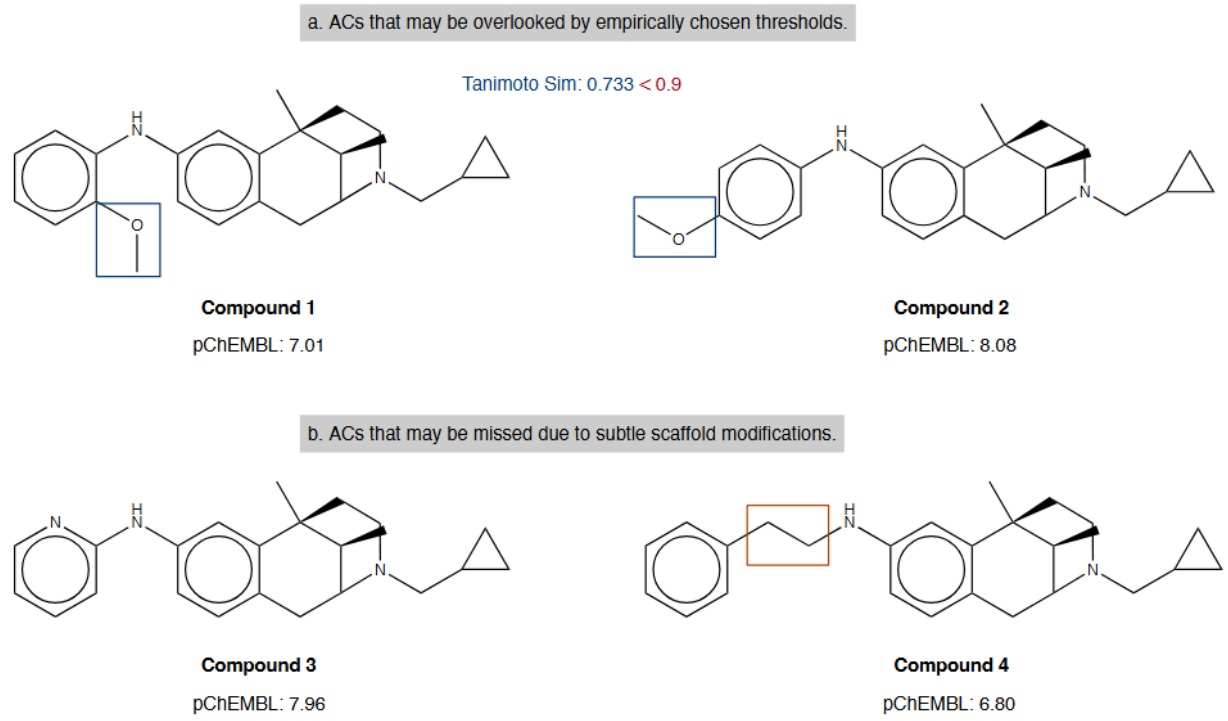
1. Despite the efforts of addressing activity cliffs through a contrastive learning approach to distinguish molecules that have different activities with similar structures in the latent space, the definitions of 1) similar structures and 2) different activities are not clearly stated in the first place, which entails two major questions 1) which structures are considered to be similar, with the same Bemis-Murcko scaffold or achieves a Tanimoto similarity above some threshold? 2) how large an activity difference should be. The reason why I list this point as the first major concern is because without clear definitions, we would not know the applicable scenarios of the proposed ACANet.

Response: Thank you for raising this important concern. Precisely defining activity cliffs (ACs), which requires consideration of both structural similarity and potency differences, is indeed a fundamental challenge. Prior studies have employed diverse criteria, and as summarized in **Response Table 1 (new Supplementary Table 15)**, there is currently no universally accepted definition of ACs.

Structural similarity. Traditional approaches typically rely on fingerprint-based similarity (e.g., Tanimoto or Jaccard coefficients) or scaffold-identity rules (e.g., Bemis–Murcko, RECAP). However, both strategies have notable limitations. Fingerprint-based methods depend on empirically selected thresholds (e.g., 0.85 or 0.9), which may miss cliffs caused by subtle but functionally critical modifications. Scaffold rules, on the other hand, cannot capture ACs that arise within a shared core scaffold (see **Response Figure 2 / Supplementary Figure 21**). Moreover, when these fixed rules are applied directly in QSAR, they lack adaptivity—failing to adjust dynamically during training and restricting coverage of the chemical space.

Response Table 1 (new Supplementary Table S15): Activity cliff definitions across different published studies.

Descriptor	Similarity	Potency difference	Study
MACCs	Tanimoto coeff. of 0.85	2-fold	Bajorath group. *Future Med. Chem.* **7**, 1565–1579 (2015); *F1000Res.* **2** (2013).
ECFPs	Tanimoto coeff. of 0.55	2-fold	Bajorath group. *Future Med. Chem.* **7**, 1565–1579 (2015); *F1000Res.* **2** (2013).
MMP	Substructure match	2-fold	Bajorath group. *Mol. Inform.* **35**, 181–191 (2016); *Future Med. Chem.* **7**, 1565–1579 (2015); *F1000Res.* **2** (2013); *J. Chem. Inf. Model.* **52**, 1138–1145 (2012).
MMP	Substructure match	1 or 2-fold	Pérez-Benito *et al.*, *J. Chem. Theory Comput.* **15**, 1884–1895 (2019).
Several 2D and 3D, including MACCS	Several similarity measures between 0.8–0.95	1 or 2-fold	Bajorath group. *J. Chem. Inf. Model.* **52**, 670–677 (2012).
Maximum common substructure	50% of common substructure atoms	1-fold	Jiménez-Luna. *J. Chem. Inf. Model.* **62**, 274–283 (2022).
Physicochemical properties	50% of max SALI score	–	Seebeck *et al.*, *ChemMedChem* **6**, 1630–1639 (2011).
MACCs, ECFP, TARIS, ROCS, Pharmacophore	SALI score	Variable	Medina-Franco *et al.*, *J. Chem. Inf. Model.* **49**, 477–491 (2009).
AP descriptor	Dice similarity of 0.7	Custom metric	Sheridan *et al.*, *J. Chem. Inf. Model.* **60**, 1969–1982 (2020).
ECFPs, Bemis–Murcko scaffolds, SMILES	Tanimoto coeff. and edit similarity of 0.9	1-fold	Van Tilberg *et al.*, *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
MMP	Hussain and Rea algorithm	2-fold	Junhui Park *et al.*, *J. Chem. Inf. Model.* **62**, 2341–2351 (2022).

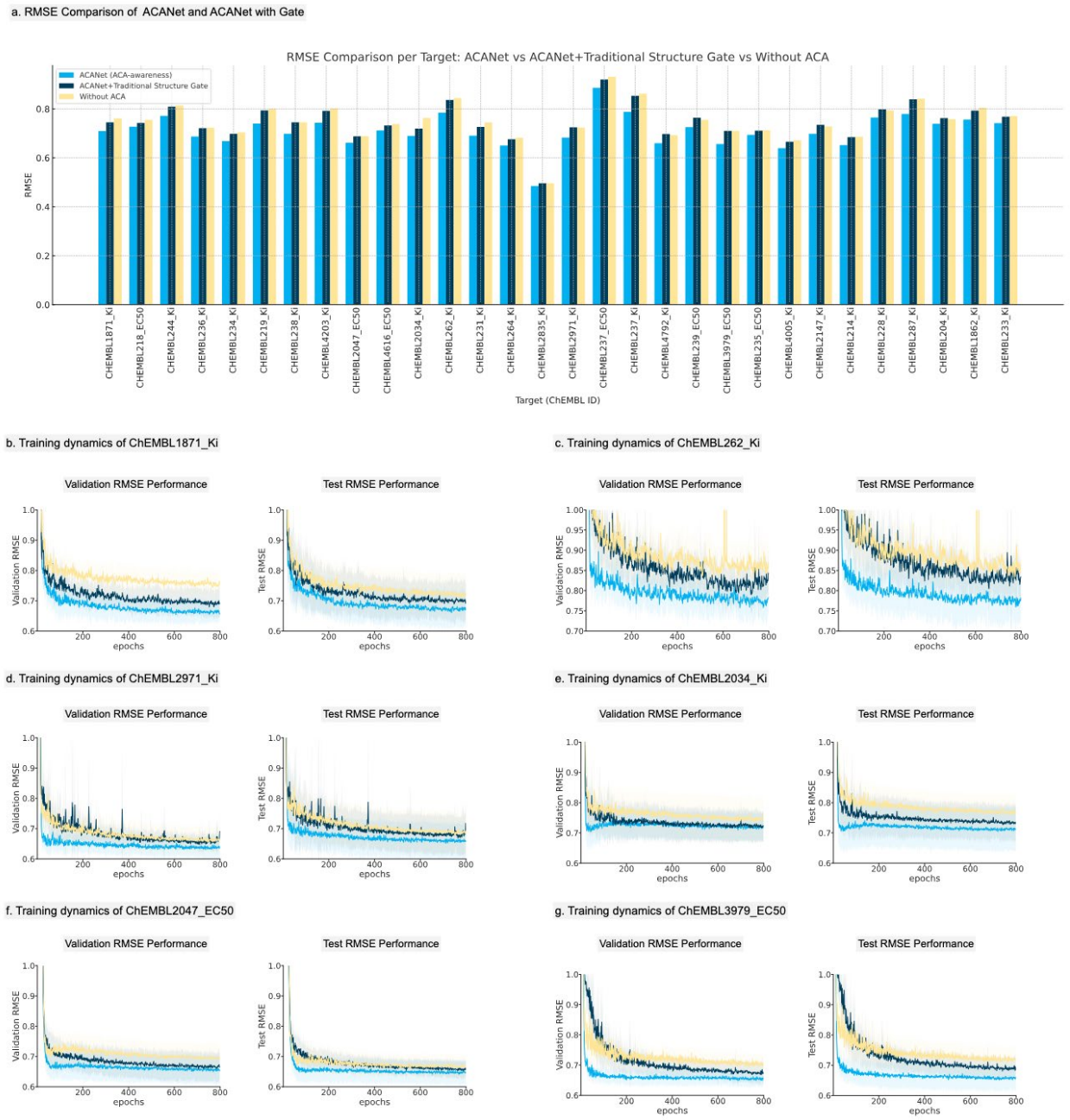


Response Figure 2 (new Supplementary Figure 21): Examples of activity cliffs (ACs) at the δ-opioid receptor (DOR) that elude traditional discriminative rules.

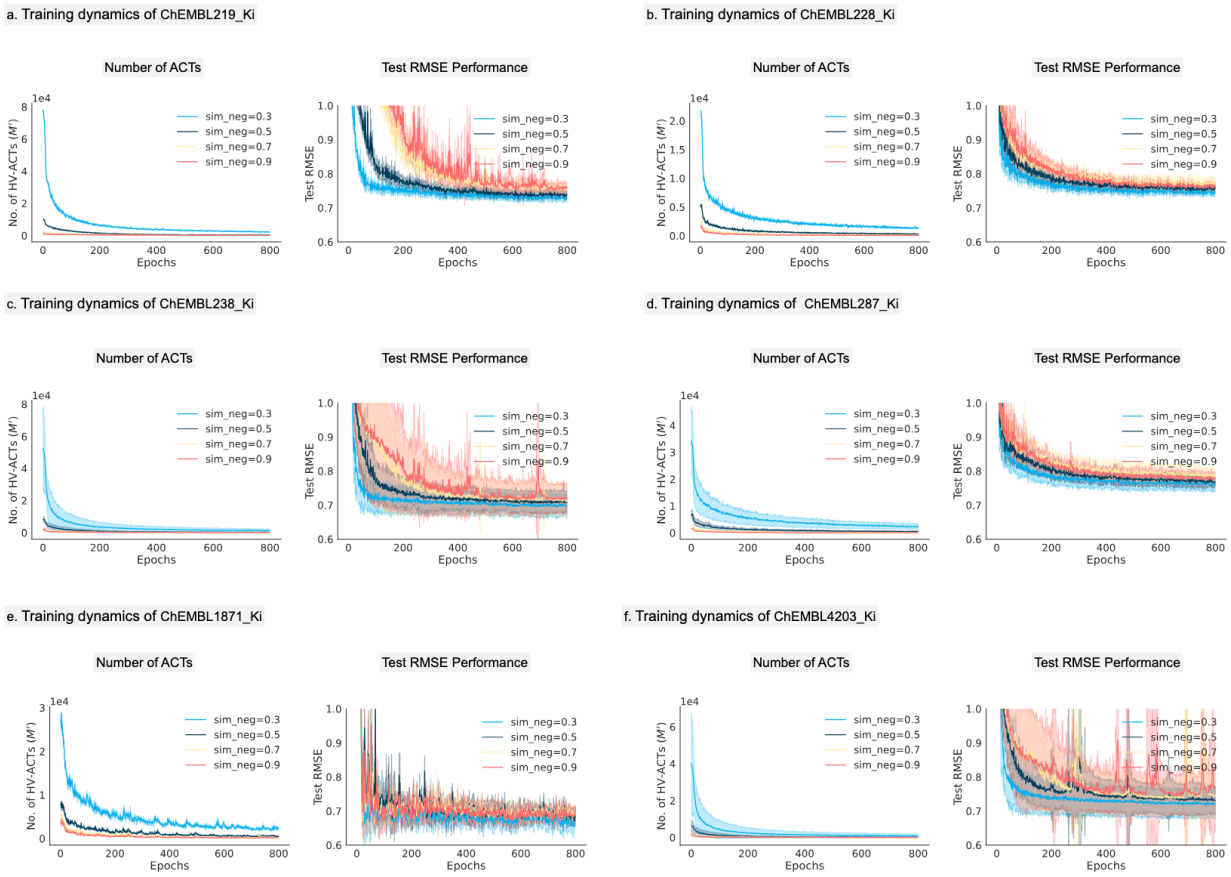
GNN-based similarity. In recent years, GNN-derived embeddings have been increasingly adopted as an alternative to traditional similarity measures (e.g., Coupry & Pogány, Yang et al.). By encoding atom- and bond-level features and propagating them through message passing, GNNs generate refined structural representations that capture subtle variations beyond static fingerprints. In ACANet, we deliberately avoided relying on **external hard cutoffs** (such as Tanimoto ≥ 0.9) to define structural similarity. Instead, we leveraged the fact that the early graph-convolution layers naturally embed molecular similarity into the latent representation $f(\cdot)$, where molecules with shared substructures cluster together. This embedding-based distance

thus serves as a flexible and adaptive similarity measure, rather than being constrained by rigid, predefined thresholds.

This strategy enables the model to uncover meaningful ACs that static fingerprint- or scaffold-based rules often miss. To validate this design choice, we implemented three classical structural gate criteria following Van Tilberg et al.—(i) Tanimoto similarity of ECFP fingerprints ≥ 0.9, (ii) Tanimoto similarity of Bemis–Murcko scaffolds ≥ 0.9, and (iii) SMILES edit similarity ≥ 0.9—and used them to constrain triplet selection (see **Response Figure 1; Supplementary Method 1**). As shown in **Response Figure 3 / Supplementary Figure 11**, our original method (latent-space similarity without hard gates) outperformed the gate-constrained variant. Further analysis of varying similarity thresholds (**Response Figure 4 / Supplementary Figure 12**) revealed that overly stringent gates reduce model performance, likely because rigid filters exclude structurally informative pairs. This suggests that ACANet's adaptive GNN-based approach captures a broader and more nuanced range of cliffs than traditional fixed-threshold strategies.



Response Figure 3 (new Supplementary Figure 11): **Comparison of ACANet variants with and without an additional structure gate.** (a) Across 30 single-scaffold HSSMS targets, the ACANet (ACA-aware, blue) bars are generally lower than the ACANet with traditional structure gate (dark green) bars, indicating that the extra gate does not improve predictive accuracy. (b–g) Training-dynamics snapshots for six example targets show the same pattern: the gated curves (dark green) remain above or overlap the ACA-aware curves (blue), while the non-ACA baseline (yellow) performs worst. Overall, adding the gate yields no benefit and can even hinder convergence, whereas the original ACANet retains the best and most stable performance.



Response Figure 4 (new Supplementary Figure 12): **Effect of structural-similarity gate on mined HV-ACTs and test RMSE.** (a–f) Training dynamics on six example HSSMS targets. Left panels: number of mined high-value ACTs (per epoch) under different similarity thresholds (sim_neg = 0.3, 0.5, 0.7, 0.9). Right panels: corresponding test-set RMSE curves. A looser gate (lower threshold, e.g., 0.3–0.5) yields substantially more HV-ACTs and consistently lower RMSE, whereas a stricter gate (e.g., 0.9) suppresses ACT mining and degrades predictive performance.

Potency differences. Activity cliffs are inherently defined by potency contrast, yet the magnitude of this contrast varies widely across targets (see **Response Figure 5 / Supplementary Figure 13**). While a 10-fold (1 log unit) difference is often used as a minimal criterion, more pronounced cliffs may involve 100-fold or greater shifts. Importantly, applying a fixed cutoff across all targets is problematic in regression-based QSAR tasks, where activity ranges differ substantially.

To accommodate this heterogeneity, we use **two tunable thresholds**: a **cliff upper** cu and a **cliff lower** cl . For a given anchor a with logged activity y^a:

- a **negative** candidate n forms an AC with aaa only if $|y^a - y^n| > cu$ **and** the structural-similarity condition is satisfied (if with structure similarity gate);
- a **positive** candidate p must satisfy $|y^a - y^p| < cl$.

Triplets (a, p, n) are valid only when **both** conditions hold with respect to the **same** anchor. This tunable design allows ACANet to **adapt to the potency landscape of each target**, rather than enforcing a one-size-fits-all rule.

Triplet sampling and optimization. We adopt a margin-based triplet loss that adapts to target-specific activity spreads. Let the margin be:

$$m = |y^a - y^n| - |y^a - y^p|,$$

where y denotes logged activity. A triplet contributes positive loss whenever

$$d(f_{(a)}, f_{(n)}) - d(f_{(a)}, f_{(p)}) < m,$$

i.e. when the negative sample is not sufficiently farther from the anchor than the positive by at least m. As training progresses, easily distinguishable triplets will be optimized soon, and once they have been correctly separated or brought close, they will no longer be over-optimized. This shifts the training focus toward the most challenging molecules in QSAR during later stages, preventing the model from overfitting to trivial triplets. This method offers an advantage over traditional approaches that are restricted to filtering a predetermined set of samples.

Why is this effective? Triplet sampling and optimization are effective because they reformulate AC detection as relative constraints in the latent space. Instead of minimizing only global prediction error, each triplet enforces that the anchor remains closer to similar-activity compounds than to dissimilar ones. This directly encodes the pairwise nature of ACs, which cannot be fully captured by regression losses alone. The adaptive margin ensures that steep cliffs enforce stronger separation while smaller differences impose weaker constraints, allowing the model to focus more attention on pharmacologically meaningful discontinuities. In addition, the self-paced nature of triplet optimization ensures that training progresses from easy to hard cases, enhancing robustness.

Importantly, the self-paced nature of triplet optimization means that easy triplets are quickly optimized and drop out of the loss, shifting training toward more challenging cases. This prevents overfitting to trivial pairs and creates a curriculum-like effect that enhances robustness. In the context of QSAR regression, this strategy complements global RMSE/MAE objectives: while traditional regression emphasizes overall accuracy, ACANet explicitly regularizes local discontinuities that are rare but clinically critical.

In summary, by combining GNN-derived similarity with tunable potency thresholds and a margin-based triplet loss, ACANet provides an explicit and adaptive framework for capturing activity cliffs. This approach avoids the rigidity of fixed fingerprint/scaffold rules, accommodates target-specific potency distributions, and ultimately achieves broader coverage and better generalization across QSAR tasks.

We hope this clarifies that ACANet does not lack clear definitions. On the contrary, both structural similarity and potency differences are formulated in an explicit and reproducible manner: similarity is encoded through GNN-based embeddings rather than arbitrary hard cutoffs, and potency contrasts are delineated by tunable thresholds cl and cu that adapt to the distribution of each target. This framework ensures that the scenarios where ACANet applies are well defined and broadly generalizable, while also being validated against classical threshold-based baselines. We believe this addresses the reviewer's concern about applicability

and demonstrates that ACANet offers a more rigorous and flexible definition of activity cliffs than fixed-rule approaches.



Response Figure 5 (new Supplementary Figure 13): Three-dimensional activity-landscape surfaces for 30 single-scaffold HSSMS datasets, generated by projecting 1024-bit Morgan fingerprints onto the first two PCA components and smoothing the activities with an RBF interpolator. Colour encodes activity: blue/green = lower, yellow/red = higher.

1.1. To follow up on this point, please consider calculating the number of scaffolds and internal diversity for the LSSNS and HSSMS datasets. Also, consider reporting the distribution of properties in each dataset.

RESPONSE: Thank you for this valuable suggestion. In the revision we have added quantitative diversity statistics—specifically, the average pairwise dissimilarity (1 - Tanimoto coefficient) between molecular fingerprints of all compounds, and reported the number of unique Bemis–Murcko scaffolds for every dataset in both the HSSMS ,LSSNS and ADMET benchmarks. These metrics are now presented in **Response Table 2 / Supplementary Table 1, Response Table 3**

/ Supplementary Table 2, Response Table 4 / Supplementary Table 4). We also provided data-distribution plots for each dataset in three benchmarks, which are now included in **Response Figure 6, Response Figure 7** and **Response Figure 8**.

Response Table 2 (new Supplementary Table S1): The 30 datasets of high-sample-size and mixed scaffolds (HSSMS) for molecular activity prediction. Datasets are the molecular-activity-prediction benchmark datasets from MoleculeACE. Diversity is calculated as the average Tanimoto distance between all molecular pairs within the dataset, while Scaffold denotes the number of unique Murcko scaffolds.

Dataset ID	Dataset Name (Target)	Assay Type	Target Type	Train / Test	% Cliff	Diversity	Num Scaffolds
CHEMBL1871_Ki	Androgen receptor (AR)	Ki	NR	659 (134)	24 (23)	0.603	206
CHEMBL218_EC50	Cannabinoid receptor 1 (CB1)	EC50	GPCR	1031 (208)	36 (36)	0.619	548
CHEMBL244_Ki	Coagulation factor X (FX)	Ki	Protease	3097 (621)	44 (43)	0.624	1250
CHEMBL236_Ki	Delta opioid receptor (DOR)	Ki	GPCR	2598 (521)	37 (37)	0.574	1249
CHEMBL234_Ki	Dopamine D_3 receptor (D3R)	Ki	GPCR	3657 (734)	39 (40)	0.683	1709
CHEMBL219_Ki	Dopamine D_4 receptor (D4R)	Ki	GPCR	1859 (374)	38 (38)	0.695	862
CHEMBL238_Ki	Dopamine transporter (DAT)	Ki	Other	1052 (213)	25 (25)	0.723	459
CHEMBL4203_Ki	Dual specificity protein kinase CLK4	Ki	Kinase	731 (149)	9 (9)	0.698	465
CHEMBL2047_EC50	Farnesoid X receptor (FXR)	EC50	NR	631 (128)	39 (39)	0.676	234
CHEMBL4616_EC50	Ghrelin receptor (GHSR)	EC50	GPCR	682 (139)	48 (49)	0.599	290
CHEMBL2034_Ki	Glucocorticoid receptor (GR)	Ki	NR	750 (152)	31 (31)	0.544	262
CHEMBL262_Ki	Glycogen synthase kinase-3 (GSK3)	Ki	Kinase	856 (173)	18 (18)	0.675	447
CHEMBL231_Ki	Histamine H_1 receptor (HRH1)	Ki	GPCR	973 (197)	23 (23)	0.698	536
CHEMBL264_Ki	Histamine H_3 receptor (HRH3)	Ki	GPCR	2862 (574)	38 (38)	0.736	1516
CHEMBL2835_Ki	Janus kinase 1 (JAK1)	Ki	Kinase	615 (126)	7 (8)	0.364	269
CHEMBL2971_Ki	Janus kinase 2 (JAK2)	Ki	Kinase	976 (197)	12 (13)	0.526	506
CHEMBL237_EC50	Kappa opioid receptor (KOR) agonism	EC50	GPCR	955 (193)	42 (42)	0.549	443
CHEMBL237_Ki	Kappa opioid receptor (KOR) inhibition	Ki	GPCR	2602 (521)	36 (36)	0.582	1231
CHEMBL233_Ki	Mu-opioid receptor (MOR)	Ki	GPCR	3142 (630)	35 (35)	0.596	1479
CHEMBL4792_Ki	Orexin receptor 2 (OX2R)	Ki	GPCR	1471 (297)	52 (52)	0.532	684
CHEMBL239_EC50	PPAR-α (PPARα)	EC50	NR	1721 (344)	41 (41)	0.665	689
CHEMBL235_EC50	PPAR-γ (PPARγ)	EC50	NR	2349 (470)	38 (38)	0.659	966
CHEMBL3979_EC50	PPAR-δ (PPARδ)	EC50	NR	1125 (225)	42 (42)	0.656	399
CHEMBL4005_Ki	PI3-kinase p110-α subunit (PIK3CA)	Ki	Transferase	960 (193)	37 (36)	0.529	385
CHEMBL2147_Ki	Serine/threonine-protein kinase PIM1	Ki	Kinase	1456 (294)	33 (33)	0.597	757
CHEMBL214_Ki	Serotonin 1A receptor (5-HT_{1A})	Ki	GPCR	3317 (666)	35 (35)	0.675	1475
CHEMBL228_Ki	Serotonin transporter (SERT)	Ki	Other	1704 (342)	35 (35)	0.709	728
CHEMBL287_Ki	Sigma opioid receptor (SOR)	Ki	Other	1328 (267)	35 (35)	0.748	730
CHEMBL204_Ki	Thrombin (F2)	Ki	Protease	2754 (553)	36 (36)	0.653	1216
CHEMBL1862_Ki	Tyrosine-protein kinase ABL1	Ki	Kinase	794 (161)	32 (32)	0.623	354

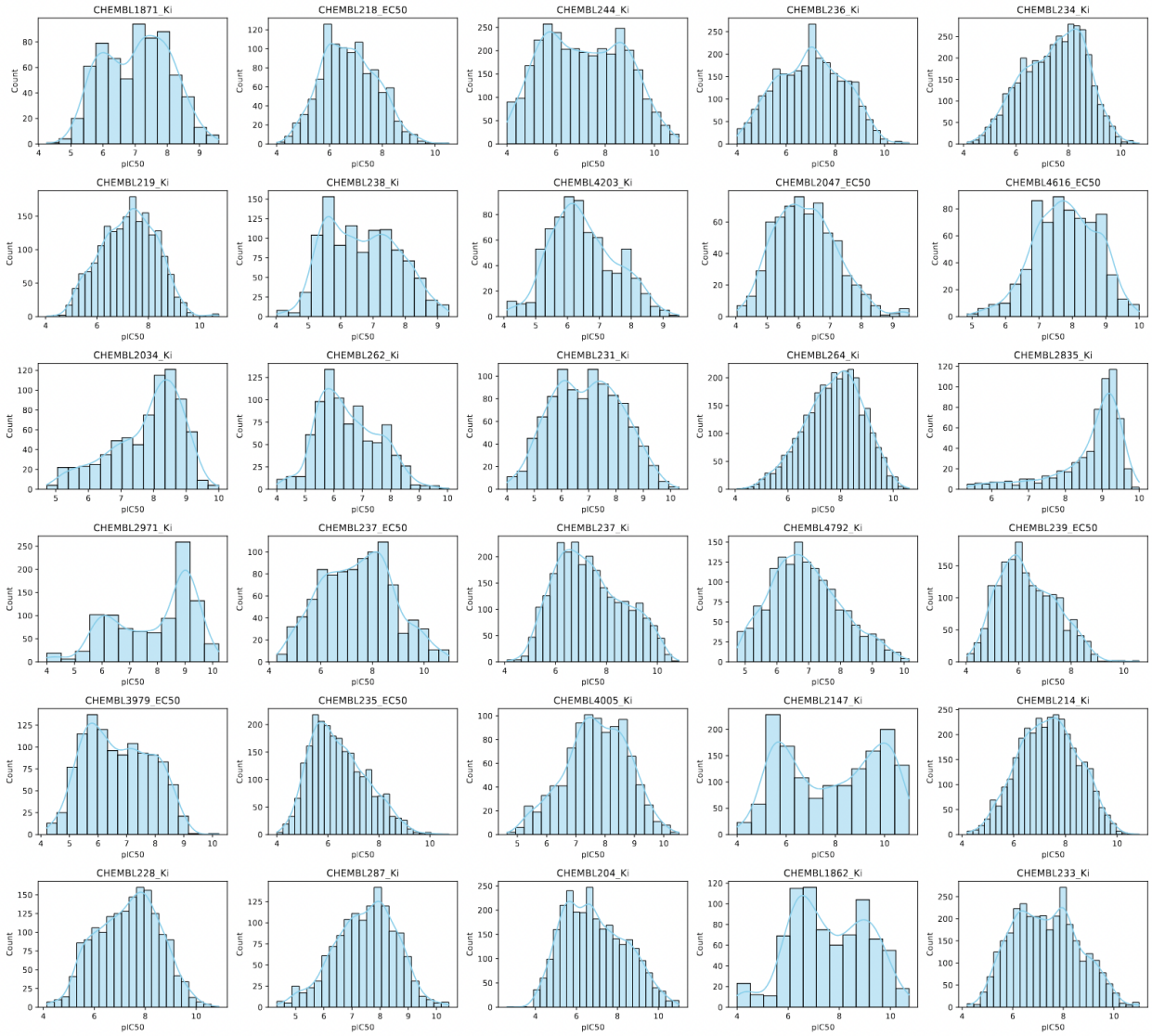
Response Table 3 (new Supplementary Table S2): The 9 datasets of low sample size and narrow scaffolds (LSSNS) for molecular activity prediction. % Cliff was identified using the same approach as MoleculeACE. Since 5-fold cross-validation was employed, we report the proportion calculated over the entire dataset. Diversity is calculated as the average Tanimoto distance between all molecular pairs within the dataset, while Scaffold denotes the number of unique Murcko scaffolds.

Dataset	Target Name	Assay type	Target type	Institution	Size	% Cliff	Diversity	Num Scaffolds
USP7	Ubiquitin carboxyl-terminal hydrolase 7	IC50	Protease	Almac Discovery Ltd., Centre for Precision Therapeutics [3]	45	95.6	0.342	16
RIP2	Serine/threonine-protein kinase RIPK2	IC50	Kinase	Genomics Institute of the Novartis Research Foundation [4]	46	91.3	0.346	14
PKCI	Protein kinase C iota	IC50	Kinase	Agency for Science, Technology and Research (A*STAR) [5]	48	64.6	0.650	34
PHGDH	D-3-phosphoglycerate dehydrogenase	IC50	Other Enzyme	Boehringer Ingelheim RCV GmbH & Co. KG [6]	51	98.0	0.377	8
PLK1	Serine/threonine-protein kinase PLK1	IC50	Kinase	Hefei University of Technology [7]	73	98.6	0.223	18
IDO1	Indoleamine 2,3-dioxygenase	IC50	Other Enzyme	NewLink Genetics Corporation [8]	78	92.3	0.421	36
RXFP1	Relaxin receptor 1	AC50	GPCR	The Florida International University Board of Trustees [9]	117	91.5	0.389	42
BRAF	Serine/threonine-protein kinase B-raf	IC50	Kinase	The Development Center for Biotechnology (DCB) [10]	128	96.1	0.308	14
MGLUR2	Metabotropic glutamate receptor 2	IC50	GPCR	Eisai R&D Management Co., Ltd. [11]	244	99.6	0.128	9

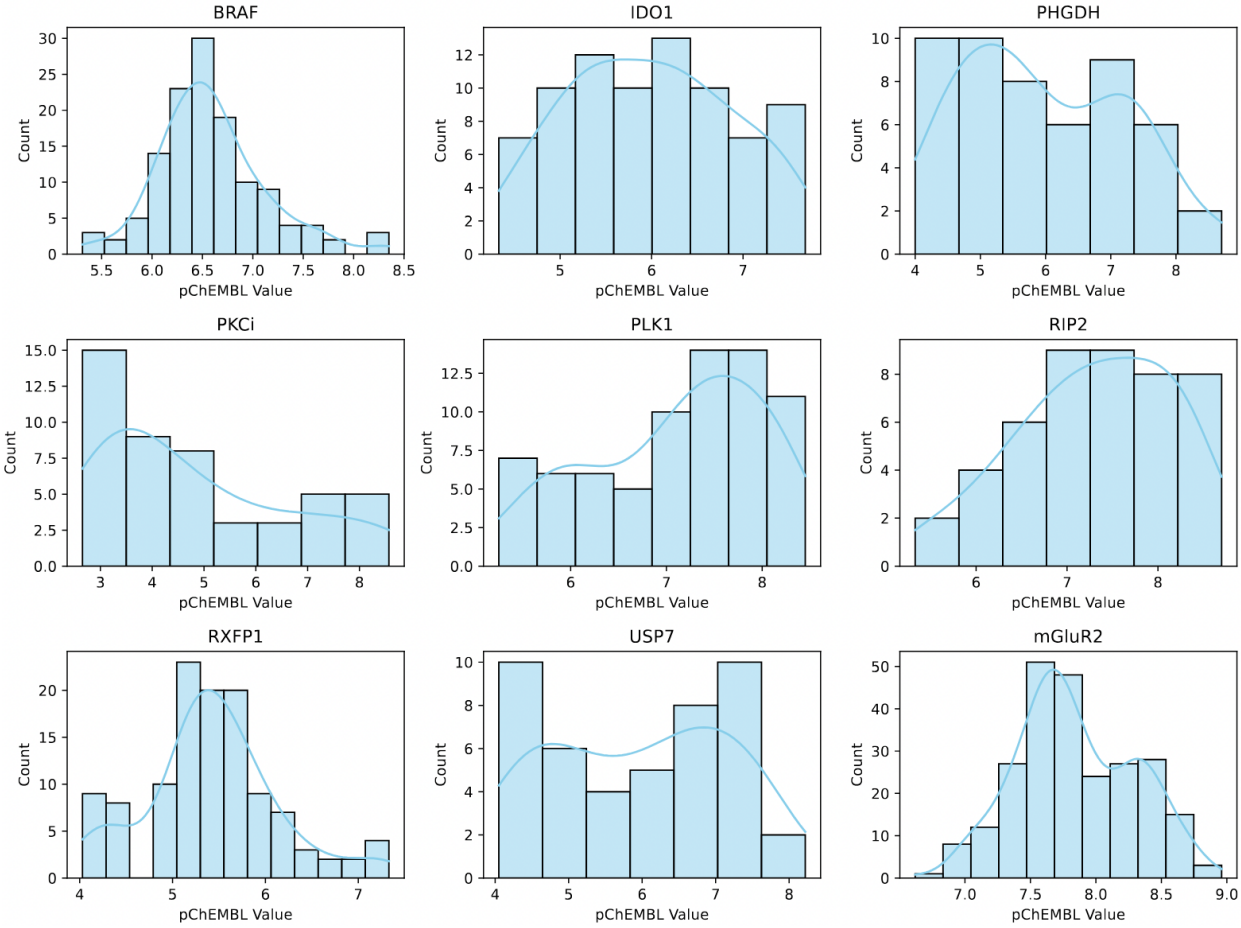
Response Table 4 (new Supplementary Table S4): The 10 ADMET datasets used for Δ-prediction and their structural statistics. % Cliff was identified using the same approach as

MoleculeACE. Since 5-fold cross-validation was employed, we report the proportion calculated over the entire dataset. Diversity is calculated as the average Tanimoto distance between all molecular pairs within the dataset, while Scaffold denotes the number of unique Murcko scaffolds.

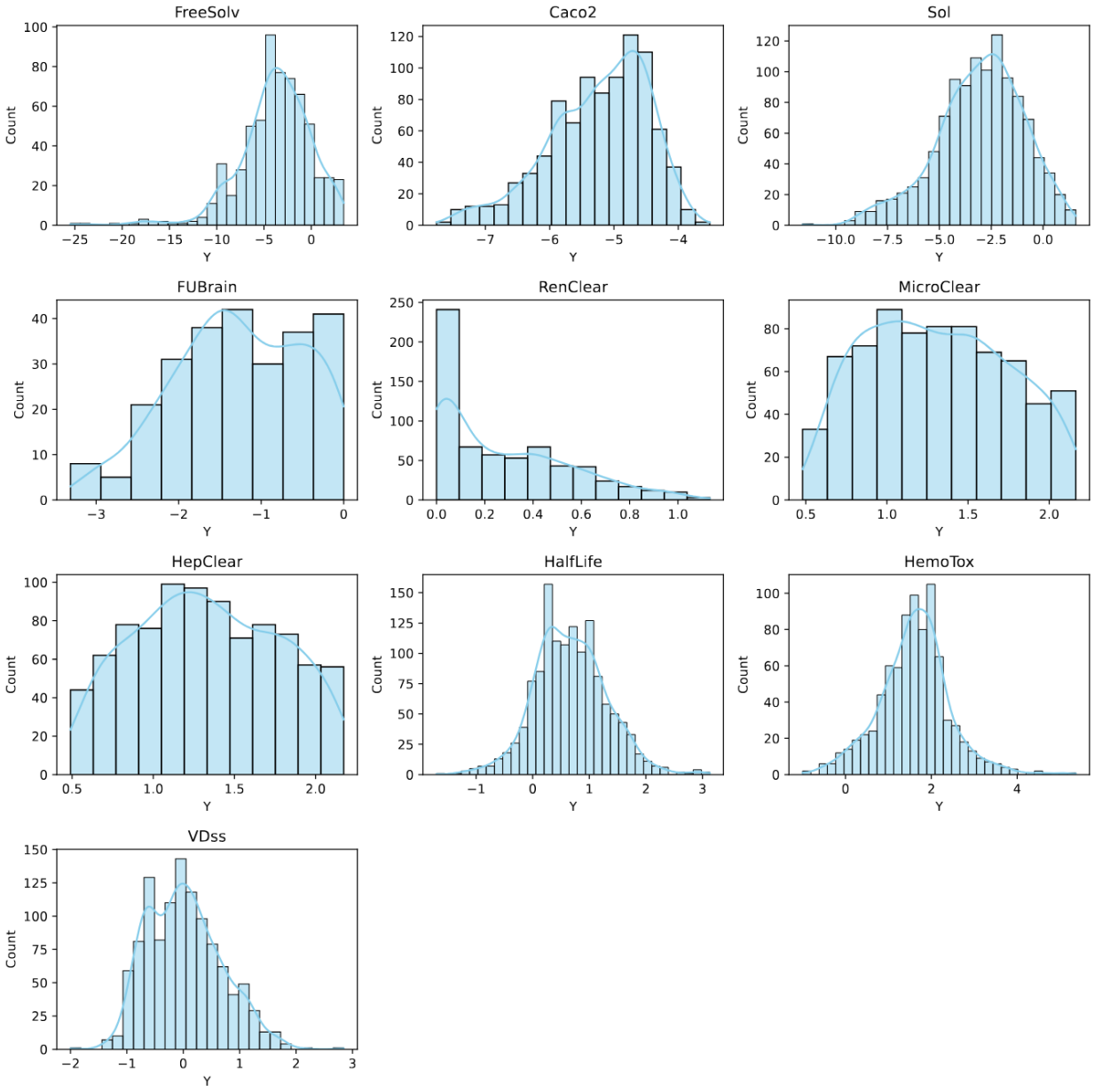
Property	Dataset	Units	Size	% Cliff	Diversity	Scaffolds
Absorption	Free Solvation	Experimental Hydration Free Energy in Water	642	92.5	0.933	63
	Caco2	Log(Papp)	910	30.8	0.741	488
	Aqueous Solubility	LogS	1128	80.2	0.920	269
Distribution	Fraction Unbound, Brain	Log(fu,brain)	253	15.4	0.724	200
	Volume of Distribution at Steady State	Log(Body/Blood Concentration in L/kg)	1130	20.1	0.730	778
Excretion	Renal Clearance	Log(CL_r)	636	5.0	0.740	447
Metabolism	Microsomal Clearance	Log(mL/min/kg cleared)	731	16.6	0.693	534
	Hepatic Clearance	Log(mL/min/kg cleared)	881	15.3	0.690	595
	Half-Life	Log(Half-Life in Hours)	1321	21.0	0.728	936
Toxicity	Hemolytic Toxicity	Log(HD_{50})	828	54.7	0.682	349



Response Figure 6: Label distribution of HSSMS benchmark.



Response Figure 7: Label distribution of LSSNS benchmark.



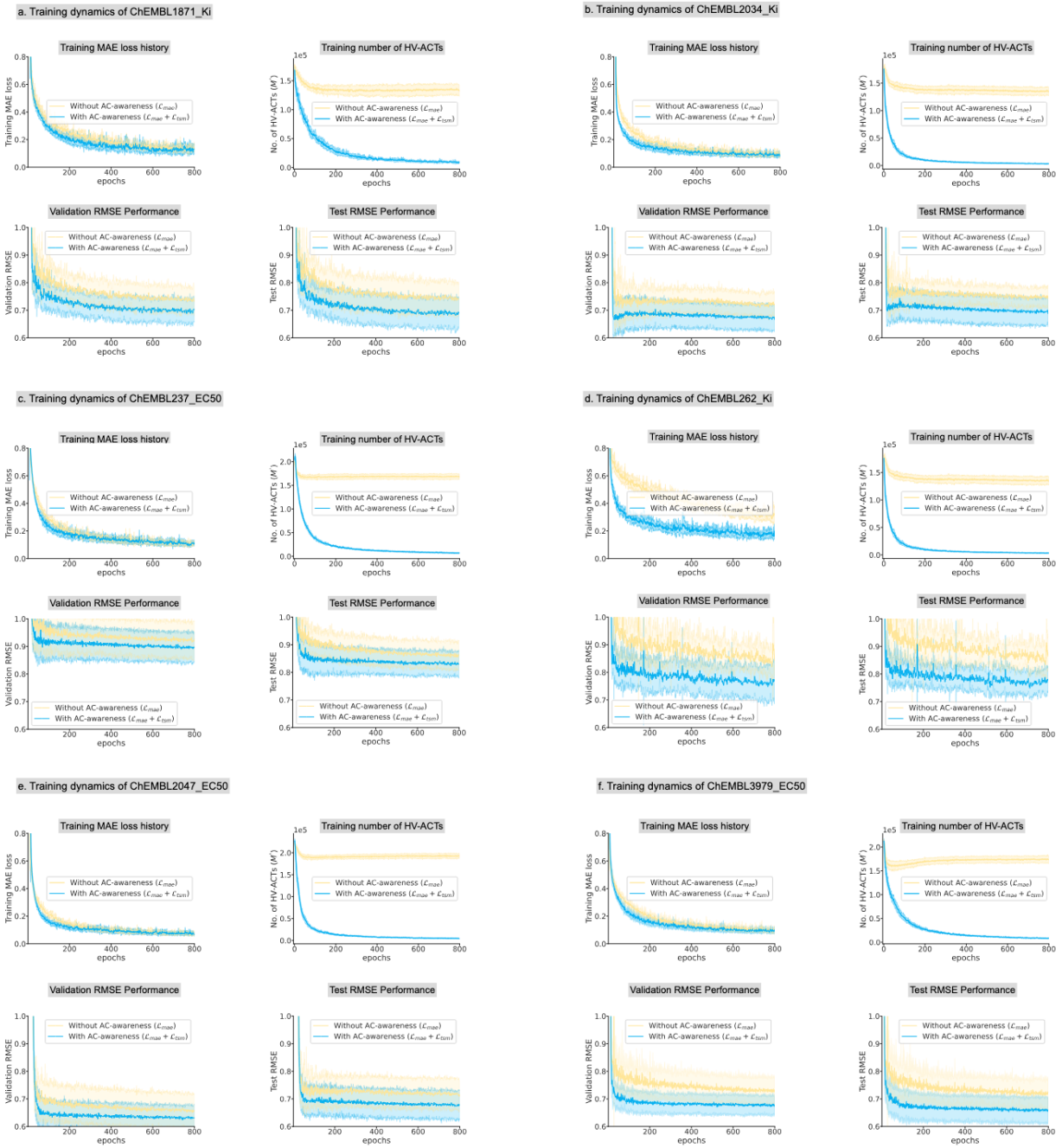
Response Figure 8: Label distribution of ADMET benchmark.

1.2. Additionally, it would be helpful to report the number of molecules observed with activity-cliffs issue. This can reflect the extent of activity-cliffs challenge in each dataset, which further justifies the selection of specific dataset for evaluation.

RESPONSE: Thank you for your suggestion. We have now also included this information in the updated **Response Table 2 / Supplementary Table S1** for clarity. Using the same activity cliff selection criteria as HSSMS, we have additionally calculated and reported the number and percentage of AC molecules for the LSSNS benchmark in **Response Table 3 / Supplementary Table S2**. Compared to HSSMS, it exhibits a notably higher proportion of activity cliffs—in 8 out of 9 datasets, more than 90% of the molecules are affected by the AC issue.

1.3. For instance, on Page 5 Line 151, a medium-size PPARδ is used to validate the effectiveness of AC-awareness. I wonder why this dataset was chosen in the first place and if the same observation persists in other datasets?

RESPONSE: Thank you for raising this point. As shown in **Response Table 2 / Supplementary Table 1**, we selected the PPARδ dataset because the proportion of AC molecules is relatively pronounced (about 42% of the molecules form AC pairs) in the HSSMS benchmark, and it also has a moderate dataset size (Train:1125/ Test:225), which makes it suitable for both rapid method validation and clear illustration. To verify that our observation is not dataset-specific, we have added additional examples (see **Response Figure 9 / Supplementary Figure 10**) and observed consistent trends across other datasets as well.



Response Figure 9 (new Supplementary Figure 10): Additional examples of ACANet training dynamics across datasets. Each row shows a target with four panels: training MAE loss, number of mined HV-ACTs, validation RMSE, and test RMSE. Curves compare the baseline (l_{mae}) and AC-aware training ($l_{mae} + l_{tsm}$); AC-aware runs converge to lower errors while mining more HV-ACTs. Each row shows a target with four panels: training MAE loss, number of mined HV-ACTs, validation RMSE, and test RMSE. Curves compare the baseline (l_{mae}) and AC-aware training ($l_{mae} + l_{tsm}$); AC-aware runs converge to lower errors while mining more HV-ACTs.

2. Next major concern is regarding model evaluation to address activity cliffs. Can you provide more details on the splitting method? Did you adopt the scaffold split or just random split? Since the proposed approach aims to address activity cliffs in molecular property prediction, a direct evaluation would be examining prediction performance on molecules with activity-cliffs issues at inference. However, it is unclear how activity cliffs are present in the training-validation-test sets for each dataset. In other words, if there are not sufficient molecules with activity-cliffs issues, would it still be fair to attribute the reduced RMSE by ACANet to AC awareness?

RESPONSE: We appreciate the reviewer's insightful comments. Ensuring that each subset contains a sufficient number of molecules with AC characteristics is indeed essential for conducting effective and fair AC-related evaluations.

For the **HSSMS benchmark**, the pre-defined train/test splits provided by van Tilborg *et al.* were directly adopted. In this benchmark, each molecule was first annotated as a cliff or non-cliff compound based on three structural similarity criteria (ECFP fingerprint Tanimoto ≥ 0.9, Bemis–Murcko scaffold similarity ≥ 0.9, or SMILES edit similarity ≥ 0.9) combined with a potency difference threshold of 10-fold. A compound was labeled as an AC if it had at least one analogue in the dataset meeting any of the similarity conditions with a potency gap greater than 10-fold. To construct the train/test sets, spectral clustering was applied on ECFP fingerprints to group structurally similar molecules into clusters, followed by stratified random sampling within clusters according to the cliff labels (80% training, 20% testing). This procedure preserves scaffold diversity and ensures that both subsets contain comparable proportions of AC and non-AC compounds. As shown in **Response Table 2 / Supplementary Table S1**, the benchmark already provides balanced splits that allow meaningful AC-aware evaluation.

For the **LSSNS benchmark**, the datasets are much smaller in size and inherently represent a narrow-scaffold scenario where most molecules share only a few scaffolds and thus ACs are highly prevalent. As indicated in **Response Table 3 / Supplementary Table S2**, more than 90% of molecules in eight out of nine datasets are involved in at least one AC. Given both the small sample sizes and the high prevalence of ACs, a **five-fold random cross-validation** strategy was adopted. Cross-validation not only maximizes the use of limited data but also ensures that sufficient AC compounds are present in every fold for reliable evaluation.

Together, these benchmark designs ensure that ACs are well represented in all training and test subsets, so that the observed performance improvements of ACANet can be attributed to its AC-aware learning mechanism rather than to sampling artifacts.

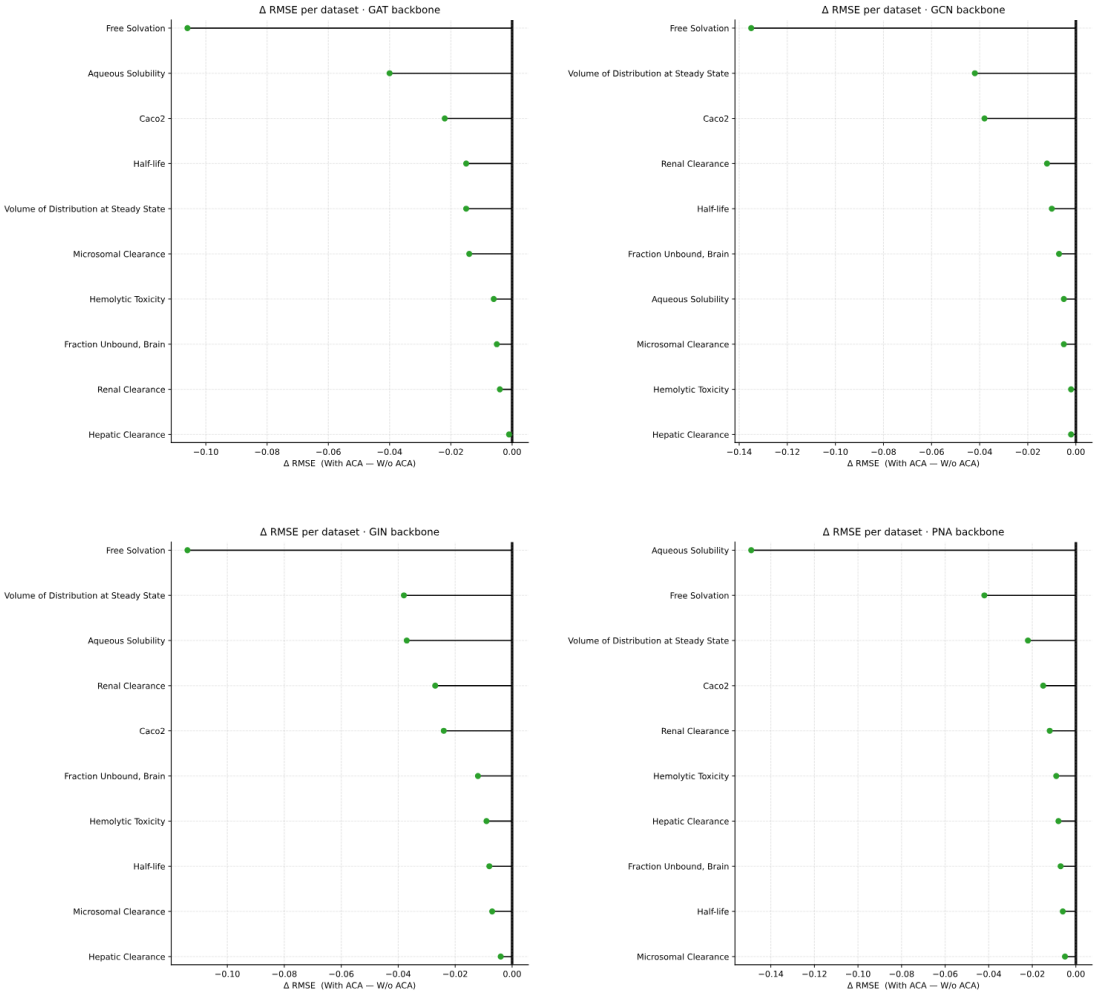
2.1. A follow-up question would be why ADMET datasets are included for evaluating ACANet? Unlike activity (specifically, binding affinity), some PK and safety-related properties cannot be attributed to chemical structures alone. In this case, the assumption for activity/property-cliffs

is undermined. Even with improvement in delta prediction, how to justify the contribution of AC awareness over model architectures and hyperparameters tuning, etc?

RESPONSE: We agree with the reviewer's point that some pharmacokinetic (PK) and safety-related properties in the ADMET dataset cannot be fully attributed to the chemical structure itself. We incorporated the ADMET dataset into the evaluation of ACANet based on the following considerations:

(1)The concept of generalized active cliff: Although PK and safety properties are not entirely dependent on molecular structure, numerous studies indicate that small structural differences can lead to significant changes in these properties and have treated this as a special “property cliff” (Sheridan et al., Ahsan et al., Abdellatif et al. , Yin et al.). Therefore, we included it in the evaluation to examine ACANet's ability to capture the "activity/property cliff" in a broad sense, thereby more comprehensively demonstrating the model's generalization ability.

(2) Verify the universality of AC sensing mechanism: Adding predictions of ADMET properties can help us evaluate the sensitivity of the deep learning model to changes in different types of properties. This further demonstrates that the AC awareness design concept is not limited to traditional activity prediction tasks but is also applicable to other important drug property prediction scenarios, demonstrating the universality of the methodology.



Response Figure 10 (new Supplementary Figure 15): ΔRMSE per dataset on ADMET across GAT, GCN, GIN, and PNA backbones under fixed ACA hyperparameters. All ACA-aware models are trained with the same default settings ($cu = 1, cl = 1, \alpha = 1$). For each dataset, the change in test error is shown as $\Delta RMSE = RMSE_{With\ ACA} - RMSE_{W/o\ ACA}$. Lollipop stems are centered at 0; green circles (left of 0) indicate improvement (lower RMSE with ACA), while the absence of red circles indicates that no dataset experienced RMSE degradation with ACA. Datasets in each panel are sorted by $\Delta RMSE$. ACA consistently outperforms the non-ACA baseline, showing superior predictive performance across all ADMET datasets.

In order to distinguish the contribution of the AC awareness mechanism itself from the influence of model architecture or hyperparameter tuning, we have conducted rigorous ablation studies in the experimental setup. As shown in **Response Figure 10 / Supplementary Figure 15**, by adjusting only the introduction of AC awareness across different backbones while keeping the hyperparameters fixed, we observed a significant improvement in model performance. These experimental results clearly demonstrate the significance of AC awareness beyond the underlying model architecture.

3. The third major concern is that, despite the shortcoming in addressing activity cliffs, learning a latent space based on chemical structures is still the basis for molecular representation learning such that property prediction can be made for unseen molecules. By introducing the AC awareness during model training, anchors-positives are pulled close and anchors-negatives are pushed away, as nicely shown in Figure 6. In this case, would this inductive bias related to activity labels interfere with the latent space viewed from the structural perspective? If yes, how to balance between the two perspectives to maximize the predictive power?

RESPONSE: We thank the reviewer for raising this important question about the potential tension between structure-based and activity-based perspectives in latent space learning. Indeed, learning a latent space grounded in chemical structures is fundamental to molecular representation learning and property prediction. The activity-cliff awareness (ACA) mechanism does not replace this foundation, but rather augments it with an additional inductive bias that emphasizes pharmacologically meaningful discontinuities.

Specifically, the ACA loss jointly optimizes (i) a standard regression loss that preserves global structure–activity trends and (ii) a triplet soft-margin loss that acts only on high-value activity cliff triplets (HV-ACTs). In these triplets, the anchor and positive compound are structurally close and similar in potency, while the anchor and negative compound are structurally close but very different in potency. By construction, this contrastive signal is sparse and local: it adjusts the relative geometry of such exceptional cases without disrupting the broader structural organization of the latent space. As shown in **Figure 6, Response Figure 13-15 / Extended Data Figure E1-E3,** and quantified by the label incoherence index (LII), the ACA loss reduces fragmentation and improves label coherence, meaning that the resulting latent space is both structurally meaningful and activity-consistent.

To address the reviewer's point on balancing the two perspectives, we emphasize that the ACA loss includes a tunable weighting factor α. This parameter controls the relative contribution of regression (structure-driven global fit) and triplet contrastive optimization (activity-driven local refinement). As shown in **Figure 5a** and **Figure 5b**, when there are no valid AC triplets or when α = 0, the model performance is significantly worse than when the ACA loss is included. In **Figure 5b**, by systematically exploring different α values through ablation studies and

comprehensive experiments, we were able to identify optimal settings that balance structural and label-driven objectives, thereby maximizing predictive performance. Thus, the method itself provides a principled mechanism to balance structural and activity information, ensuring that predictive power is maximized rather than compromised.

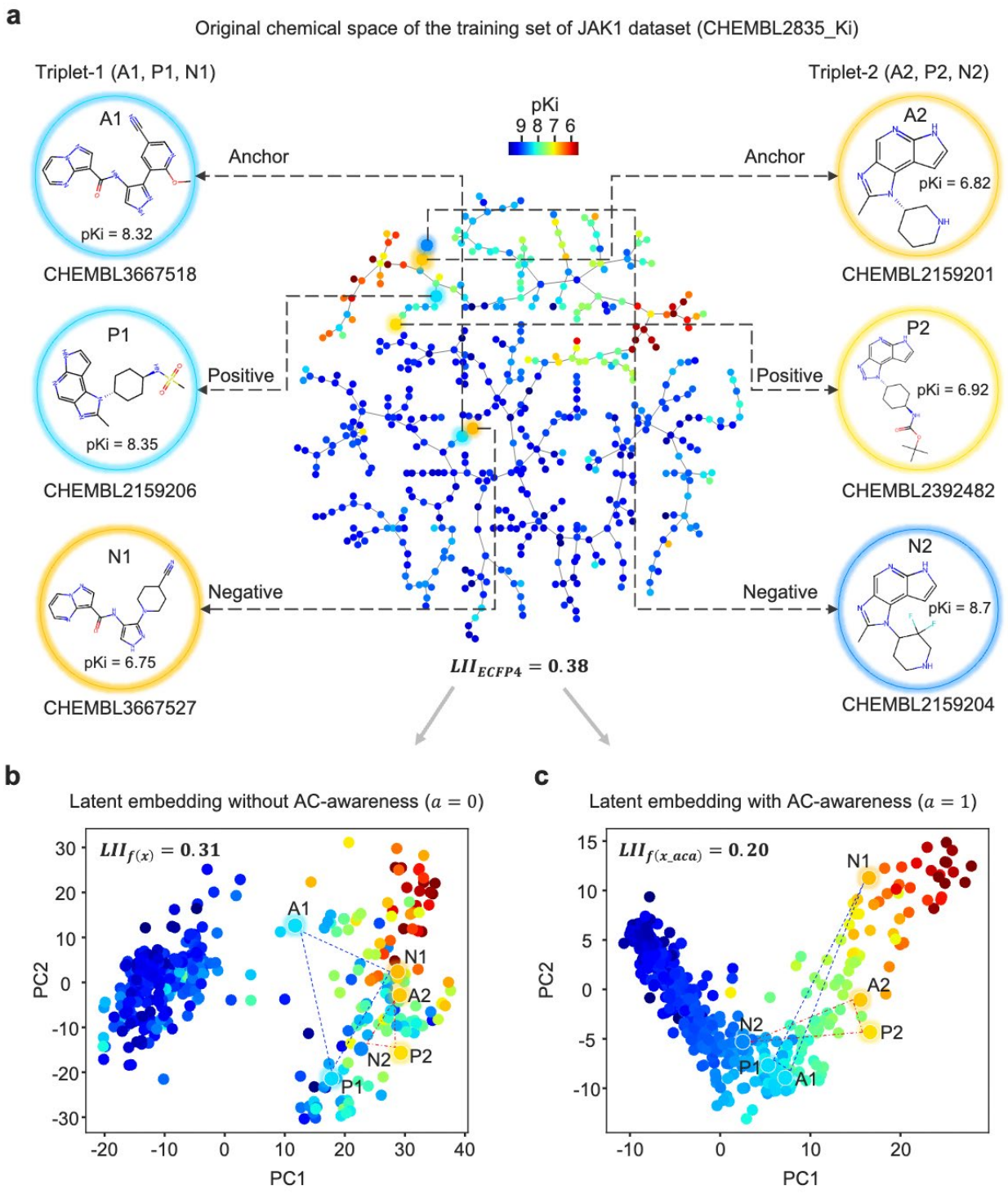
In sum, the inductive bias introduced by AC-awareness does not interfere with the structural perspective. Instead, it selectively corrects cases where structural similarity alone would be misleading, thereby aligning latent space organization more faithfully with the true structure–activity landscape.

3.1. A related question in Figure 6 is that are the two triplets from the training set or test set? They seem to have the same scaffold from a visual check. It would be interesting to show examples from 1) both training set and test set and 2) different scaffolds to see if the latent space keeps organized 1) from training to inference and 2) across different chemical spaces.

RESPONSE: We thank the reviewer for this careful observation and helpful suggestion. The two triplets originally shown in Figure 6 are indeed from the **training set** of the BRAF dataset. To address the reviewer's point, we have now extended our analysis with additional examples in the revision.

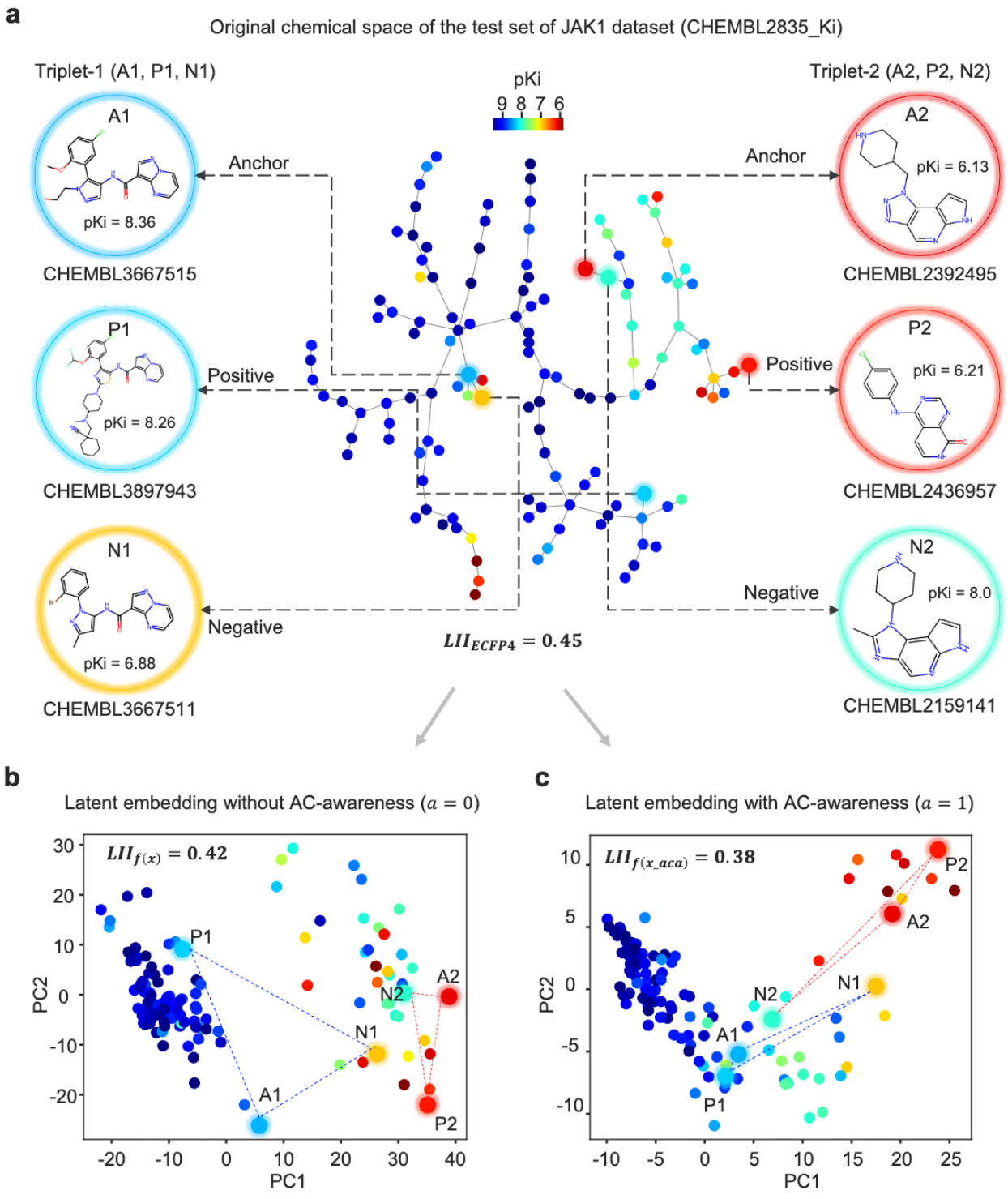
1. **Training vs. test sets:** We provide triplets drawn from both the training and test sets of the JAK1 dataset from the HSSMS collection (**Response Figure 11-12 / Extended Data Figures E4–E5**). These examples confirm that the organization of the latent space with AC-awareness is preserved not only during training but also at inference time on unseen molecules.
2. **Different scaffolds:** The new examples include compounds spanning different Bemis–Murcko scaffolds, demonstrating that AC-awareness maintains an activity-coherent latent space even across broader chemical series and chemical spaces.
3. **Quantitative evidence:** In addition, we computed and visualized the training history of the Local Label Incoherence (LII) across all 30 HSSMS datasets (**Response Figure 13-15 / Extended Data Figures E1–E3**). On both training and test sets, models equipped with AC-awareness consistently exhibited lower LII values compared to baselines, indicating a more coherent latent space across diverse datasets.

Together, these results demonstrate that ACANet generalizes well: (i) it preserves activity-coherent latent space organization both from training to inference, and (ii) it achieves robust clustering of compounds with similar activities across distinct scaffolds. This shows that the inductive bias introduced by AC-awareness not only corrects local activity cliffs but also enhances the global robustness of the latent space.



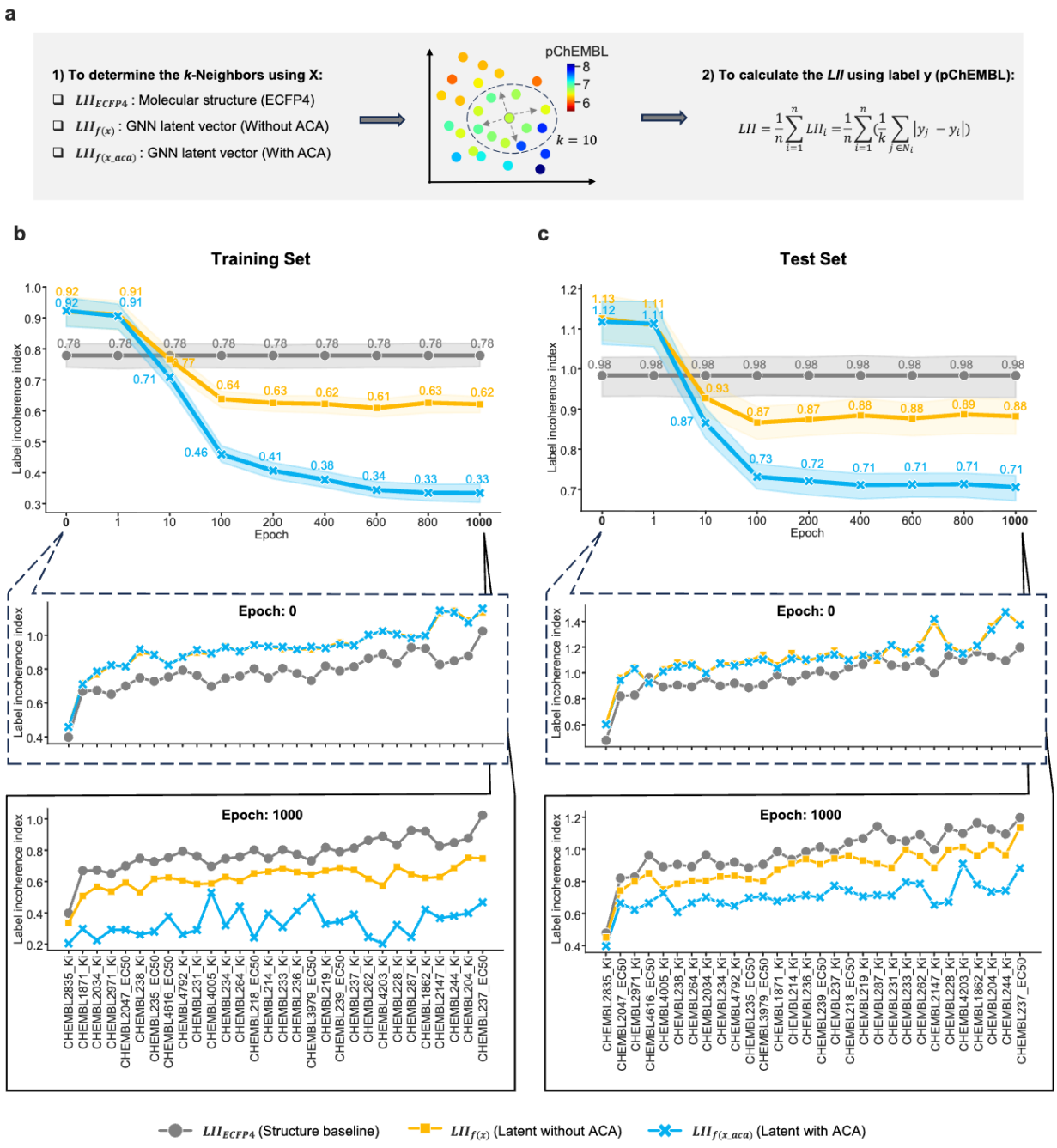
Response Figure 11 (new Extended Data Figure 4): Impact of activity-cliff awareness on latent organisation of the JAK1 training set (CHEMBL2835_Ki). (a) tmap visualisation of the test-set chemical space. All test compounds are positioned by ECFP4 similarity and coloured by pKi (blue → low, red → high). Two representative activity-cliff triplets are highlighted: A_1–P_1–N_1 (left) and A_2–P_2–N_2 (right). Each triplet comprises an anchor (A), a potency-matched positive (P), and a lower-potency negative (N); the two triplets belong to

different Bemis–Murcko scaffolds. The structure-based label-incoherence index is $LII_{ECFP4} = 0.38$. **(b) Latent embedding without AC awareness (α = 0).** Two-dimensional PCA of GNN embeddings trained without ACA shows that, for both scaffolds, anchors (A) remain closer to their negatives (N) than to their potency-matched positives (P), leaving scaffold and potency information entangled ($LII_{f(X)} = 0.31$). **(c) Latent embedding with AC awareness (α = 1).** Including the ACA loss reorganises the latent space: anchors cluster tightly with their positives and are clearly separated from their negatives, even for these two distinct scaffolds, reducing the index to $LII_{f(X_{ACA})} = 0.20$. This demonstrates that ACA improves local label coherence on the test set and effectively handles activity cliffs across different scaffolds.



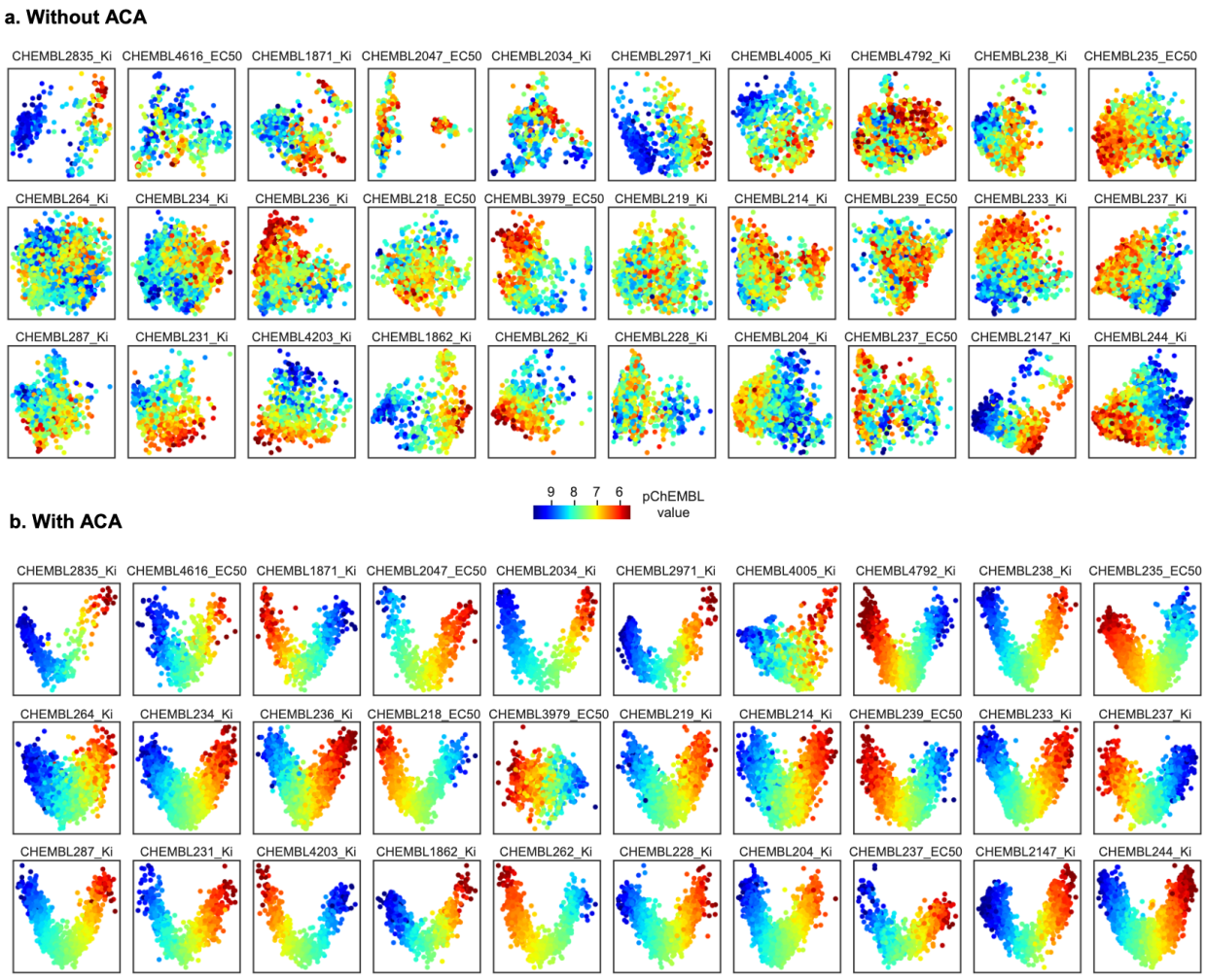
Response Figure 12 (new Extended Data Figure 5): Impact of activity-cliff awareness on latent organisation of the JAK1 test set (CHEMBL2835_Ki). (a) tmap visualisation of the test-set chemical space. All test compounds are positioned by ECFP4 similarity and coloured by pKi (blue → low, red → high). Two representative activity-cliff triplets are highlighted: A_1–P_1–N_1 (left) and A_2–P_2–N_2 (right). Each triplet comprises an anchor (A), a potency-matched positive (P), and a lower-potency negative (N); the two triplets belong to different Bemis–Murcko

scaffolds. The structure-based label-incoherence index is $LII_{ECFP4} = 0.45$. **(b) Latent embedding without AC awareness (α = 0).** Two-dimensional PCA of GNN embeddings trained without ACA shows that, for both scaffolds, anchors (A) remain closer to their negatives (N) than to their potency-matched positives (P), leaving scaffold and potency information entangled ($LII_{f(X)} = 0.42$). **(c) Latent embedding with AC awareness (α = 1).** Including the ACA loss reorganises the latent space: anchors cluster tightly with their positives and are clearly separated from their negatives, even for these two distinct scaffolds, reducing the index to $LII_{f(X_{ACA})} = 0.38$. This demonstrates that ACA improves local label coherence on the test set and effectively handles activity cliffs across different scaffolds. PCA projections of compound embeddings for the HSSMS training set.

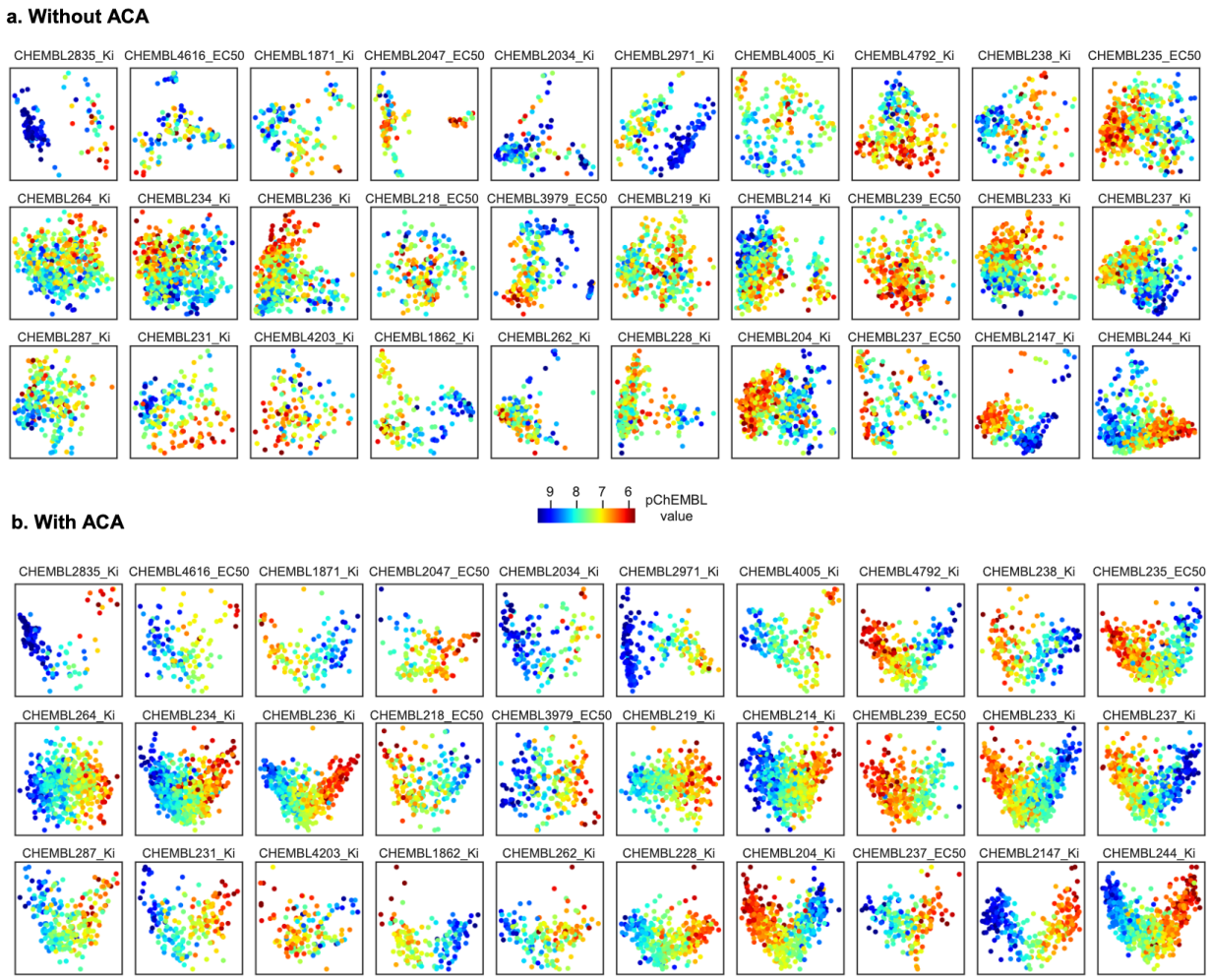


Response Figure 13 (new Extended Data Figure 1): Temporal evolution of the label-incoherence index (LII) in the HSSMS benchmark. Lower LII values indicate greater local label coherence. (a) LII calculation workflow – Step 1: for each compound, identify the $k = 10$ nearest neighbours using one of three representations: ECFP4 fingerprints, GNN latent vectors without ACA, or GNN latent vectors with ACA. Step 2: compute LII from the pChEMBL labels of the query compound and its neighbours. **(b) Training set** – Epoch-wise LII dynamics for the three representations (grey = ECFP4, yellow = latent without ACA, blue = latent with ACA), accompanied by per-dataset LII values at epoch 0 and epoch 100. **(c) Test set** – Corresponding LII dynamics for held-out compounds, with per-dataset snapshots at epoch 0

and epoch 100. Across all datasets, ACA persistently drives the latent-space LII below both the ECFP baseline and the non-ACA model.



Response Figure 14 (new Extended Data Figure 2)：PCA projections of compound embeddings for the HSSMS training set, coloured by pChEMBL value (blue → low, red → high). (a) Without ACA regularisation – Two-dimensional PCA embeddings from the baseline model display fragmented and discontinuous activity landscapes, indicative of latent-space activity cliffs. **(b) With ACA regularisation** – PCA embeddings learned with ACA show smoother and more coherent activity gradients, demonstrating improved preservation of structure–activity relationships and reduced cliff severity.



Response Figure 15 (new Extended Data Figure 3)：PCA projections of compound embeddings for the HSSMS test set, coloured by pChEMBL value (blue → low, red → high). (a) Without ACA regularisation – Two-dimensional PCA embeddings from the baseline model display fragmented and discontinuous activity landscapes, indicative of latent-space activity cliffs. **(b) With ACA regularisation** – PCA embeddings learned with ACA show smoother and more coherent activity gradients, demonstrating improved preservation of structure–activity relationships and reduced cliff severity.

4. The ablation study on the impact of cliff cutoffs and the awareness factor can be very helpful, if expanded to more datasets. Ideally, it would be of high value if some general rules can be established between tunable hyperparameters (alpha, cu and cl) for AC-awareness and datasets (e.g. degree of scaffold diversity, prevalence of activity cliffs as mentioned in Discussion) because then the value of ACANet can become evident.

RESPONSE: Thank you very much for this valuable suggestion. Following the reviewer's recommendation, we have substantially expanded the ablation study to all 30 HSSMS datasets, in addition to the LSSNS datasets already included in the previous version. Because an exhaustive two-dimensional grid search over the lower and upper cliff cutoffs would be extremely computationally expensive, we adopted a more efficient "T-shape" search strategy. Specifically, we first fixed the awareness factor alpha at its default value of 0.1 and varied the upper cutoff from 0.5 to 4.0 in steps of 0.5 while keeping the lower cutoff fixed at 0.1. After

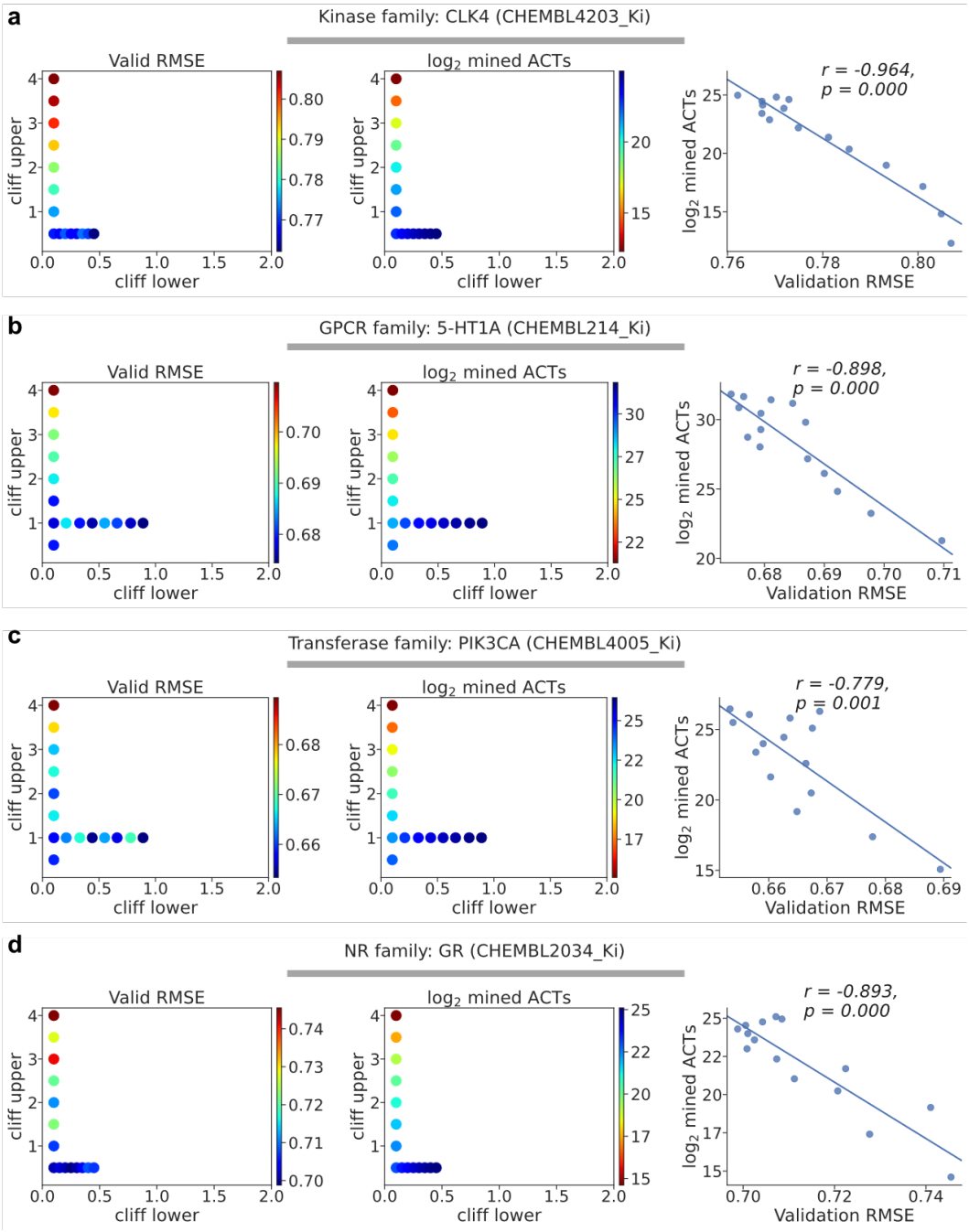
determining the best upper cutoff under this setting, we then varied the lower cutoff between 0.1 and that upper cutoff using eight evenly spaced values. Once the optimal pair of lower and upper cliff cutoffs was obtained, we selected the best alpha value through stratified five-fold cross-validation using a range including 0, 0.01, 0.05, 0.1, 0.2, 0.5 and 1.0. All results for the 30 HSSMS datasets are publicly available at: https://github.com/shenwanxiang/ACANet/tree/main/experiment/03_HSSMS_performance/benchmark_performance/mae_aca_opt_cliff

A clear and consistent finding emerges across all dataset families. **The predictive performance of the model (measured by RMSE) shows a strong negative correlation with the number of mined activity-cliff triplets.** In other words, datasets or hyperparameter combinations that allow the model to sample more informative triplets always lead to better predictive accuracy. This relationship is illustrated in **Response Figure 16**, where we show one representative dataset from each of the four target families (Kinase, GPCR, Transferase and Nuclear Receptor) in HSSMS. The trend is fully consistent with the patterns previously observed in LSSNS datasets as shown in **Response Figure 3 / Supplementary Figure 2.**

These results suggest a general rule that applies to both narrow-scaffold (LSSNS) and mixed-scaffold (HSSMS) datasets**: the lower and upper cliff cutoffs can be selected efficiently before model training simply by examining the number of feasible triplets under each cutoff combination**. The number of mined triplets is easy to compute and strongly predictive of the final model performance, which also highlights why the ACA loss leads to improved results — more informative triplets provide stronger contrastive supervision.

We further examined the optimized values of the awareness factor alpha across all datasets. In the LSSNS collection, seven out of nine datasets preferred an alpha value of 1.0. In the HSSMS collection, nine datasets also preferred alpha equal to 1.0, and eight additional datasets selected alpha equal to 0.5. This indicates that even in structurally diverse datasets, a relatively high weight on the contrastive component is often beneficial.

Overall, these expanded ablation studies not only generalize the findings to a larger set of datasets, but also reveal practical guidelines for selecting hyperparameters in ACANet. They show that both the cliff cutoffs and the awareness factor can be chosen in a data-driven manner and can be flexibly adapted to different dataset characteristics.



Response Figure 16: Correlation between the number of mined label-conditioned triplets (M′) and the validation RMSE across representative HSSMS datasets. Panels (a)–(d) show results for one dataset from each major target family: CLK4 (Kinase), 5-HT1A (GPCR), PIK3CA (Transferase), and GR (Nuclear Receptor). For each dataset, the cliff lower and cliff upper thresholds were scanned using the T-shape strategy, while the awareness factor α was kept fixed. The resulting number of mined ACTs M under each cliff-parameter combination is visualized alongside the corresponding validation RMSE obtained from three-fold cross-validation. The scatter plots on the right show the strong negative correlation between $\log_2(M)$ and the RMSE of the validation set, illustrating that models benefiting from a larger pool of informative triplets tend to achieve better predictive accuracy.

Minor Concerns

1. Consider replacing “differing activities” when discussing activity cliffs because a very modest change in activity can always be called differing but they do not necessarily constitute an activity cliff.

RESPONSE: We appreciate your suggestion. “differing activities” has been replaced with **“with a substantial potency difference”** in the revised manuscript.

2. The authors nicely pointed out in “Overview of datasets and tasks” (Page 4) that on LSSNS datasets are representative during the lead identification phase and MMP datasets are relevant in lead optimization phase. On the contrary, HSSMS datasets are more relevant to a different stage, mid-stage drug discovery. Here, please consider clarifying what it means by mid-stage drug discovery to alleviate confusion. Is it hit identification where different scaffolds are involved in a broader chemical space?

RESPONSE: Thank you for raising this point. We agree that the phrase “mid-stage drug discovery” may introduce ambiguity, and we have removed it from the revised manuscript. Instead, we now describe the two dataset groups directly in terms of **their chemotype diversity and modeling purpose**, which more accurately reflects how they are used in practice.

The LSSNS datasets consist of small, single-scaffold series and therefore correspond to the classical QSAR setting, where medicinal chemists explore local structure–activity relationships within a narrow chemical series. In contrast, the HSSMS datasets contain substantially larger numbers of compounds spanning multiple scaffold families. Such datasets are typically used in **machine-learning–based activity prediction across broader chemical space**, where the goal is to prioritize or rank molecules originating from diverse chemotypes, including those beyond a single lead series.

Thus, rather than implying a specific stage of drug discovery, we clarify that the HSSMS datasets simply represent **broader-chemotype, larger-scale prediction tasks**, where scaffold diversity is intentionally higher and models must generalize across different structural classes. This revision removes the ambiguity and highlights the intended role of HSSMS datasets as a complementary evaluation setting to the narrow-scaffold LSSNS datasets.

We have updated the corresponding paragraph (**Overview of datasets and tasks**) in the manuscript to reflect this clearer terminology.

3. Please check throughout the manuscript to ensure all abbreviations are spelled in their first occurrence. For example, on Page 6 Line 164 “ACANet model uses PNA as its default backbone…”, there is no annotation for PNA until the Methods section.

RESPONSE: Thank you for pointing this out. We have carefully checked the full manuscript to ensure that every abbreviation is spelled out at its first occurrence. In particular, the sentence on Page 6 has been revised to: “The ACANet model uses **principal neighbourhood aggregation (PNA)** as its default backbone, but we also evaluated three additional GNN backbones: **graph convolutional network (GCN)**, **graph attention network (GAT)**, and **graph isomorphism network (GIN)**.”
All other abbreviations throughout the manuscript have been reviewed and corrected accordingly.

4. On Page 21 Line 423, the Methods section has a title as"Online Methods". Is there a specific reason why it is called "online" methods?

RESPONSE: Thank you for the question. We use the term **"Online Methods"** to emphasize that activity-cliff triplets are **not precomputed for the entire dataset**, but are instead **generated on-the-fly within each mini-batch during training**. This online mining scheme avoids the prohibitive memory and time requirements of enumerating all possible triplets beforehand. More importantly, it allows the set of mined triplets to **adapt dynamically as the model learns**: once easier triplets are resolved, the training naturally shifts toward harder and more informative cases. This continual adaptation is central to ACANet's design and is why we describe the procedure as an online method.

Besides, the Methods section can be cleaned a bit. For example, consider removing numbering for subsections under "Activity cliff awareness (ACA) loss function" in Pages 23-25 and subsubsections under "Analysis of label incoherence across chemical and latent spaces" in Page 26. Additionally, consider numbering all equations in the manuscript.

RESPONSE: We appreciate the reviewer's suggestions regarding the organization of the Methods section. In the revised manuscript, we have streamlined the formatting to improve readability and consistency. Specifically, we have removed the internal numbering from subsections under **"Activity cliff awareness (ACA) loss function"** and from the subsubsections under **"Analysis of label incoherence across chemical and latent spaces."** The content now follows a clean hierarchical structure that is easier to navigate.

Additionally, we have reviewed all mathematical expressions and numbered the equations throughout the manuscript to ensure consistent referencing. These revisions help improve clarity and make cross-referencing between the text, Methods, and Supplementary Information more straightforward.

5. Consider examining the use of "significantly" or "significant". Usually in scientific publications, it usually means a result from a statistical test that is not likely to happen by chance. However, when describing difference in labels or losses, there are quite some sentences like "when their activities differ significantly" in Line 34-Line 35 (Page 2), "final MAE loss is not significantly different" in Line 155 (Page 5) and "achieved significantly lower RMSE values" in Line 162 (Page 5). Please consider using "significantly" or "significant" with more caution.

RESPONSE: We appreciate the reviewer's suggestion. In the revised manuscript, we have replaced the uses of significant and significantly that were not supported by statistical analyses with more appropriate alternatives (e.g., markedly, substantially). We have retained the term significant only in contexts where statistical significance was explicitly assessed, such as in **Supplementary Table 8, Supplementary Table 16** and **Supplementary Table 19**, to ensure accuracy and rigor in our descriptions.

6. While reading the manuscript, I found that sometimes I would get confused by M (conditional triplets) and M' (mined triplets). Consider using more straightforward abbreviations and provide more illustrations on M and M' in the first place.

Response: Thank you for pointing this out. We agree that the distinction between M and M' was not sufficiently clear. In the revision, we have clarified this in two ways. First, Figure 1 (panel b) and its caption now explicitly illustrate how the cliff thresholds define the set of **label-conditioned triplets** (M) and how the online mining step extracts from these the **mined high-value triplets** (M') that actually contribute to the loss.

Second, we introduce these descriptive terms at first occurrence and use them consistently throughout the manuscript:
• *M* = label-conditioned triplets (all triplets satisfying the cliff criteria),
• *M′* = mined high-value triplets (the subset with positive triplet-soft-margin loss).

These changes make the notation clearer and help readers understand how triplet conditioning and online mining operate within ACANet.

7. Consider adding the number of molecules for the overall test set and the cliff test set for the 30 HSSMS benchmark datasets in Figure 7 or provide them in Supplementary Information.

RESPONSE: Thank you for your suggestion. We have now included the number of molecules in the overall test set as well as the proportion of cliff molecules within it for each of the 30 HSSMS benchmark datasets. This information has been added to **Response Table 2 / Supplementary Table 1** for clarity and completeness.

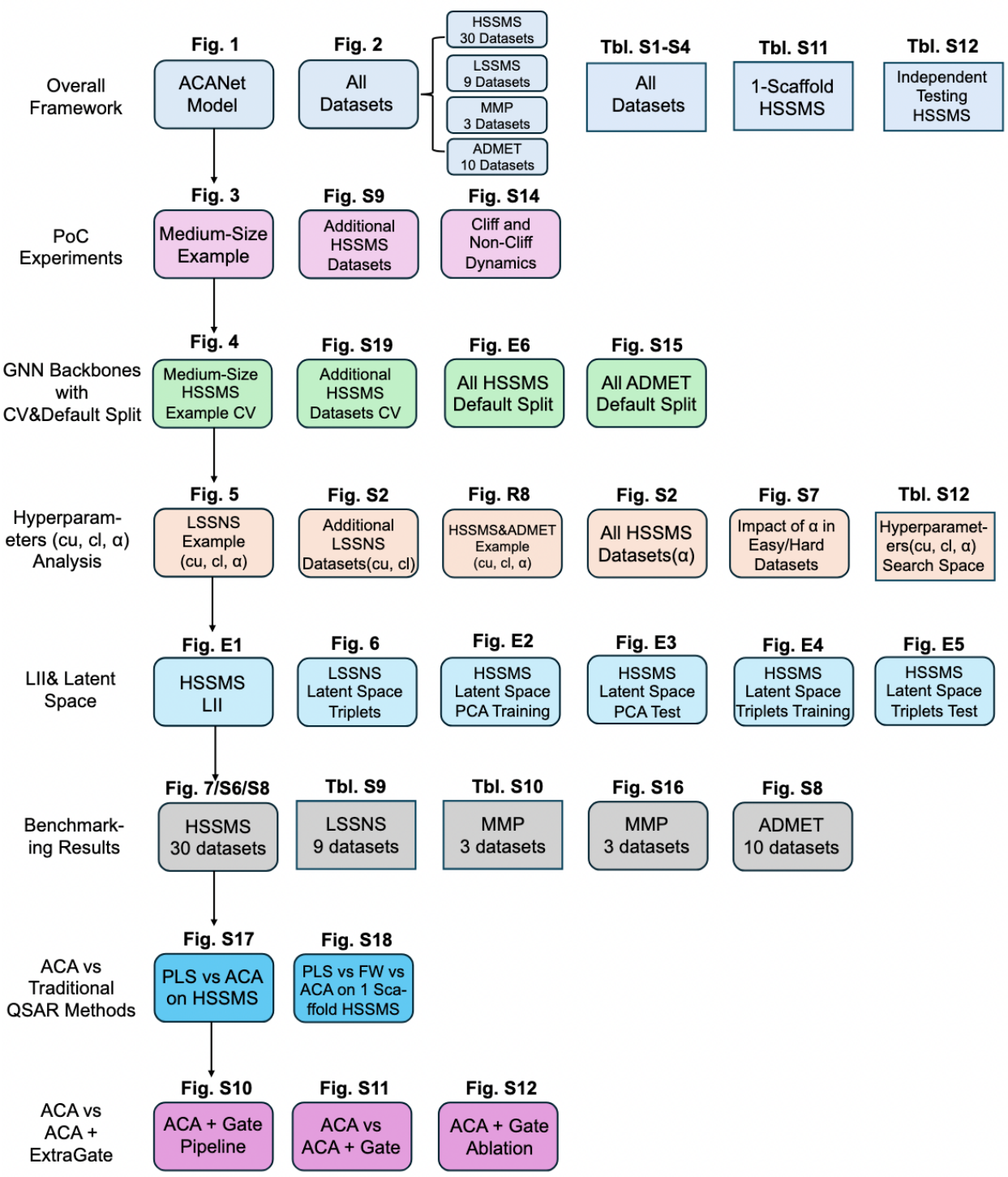
8. Throughout the experiments, there are cases1) where MAE loss or RMSE loss is used and 2) a certain dataset is used to illustrate the impact of AC awareness or related hyperparameters. It would be nice if there is a figure to summarize the experiment design. Also, consider elaborating more about why some experiments are designed in such a way in the Discussion section. For example, why Figures 3&4 focus on PPARδ whereas Figures 5&6 focus on BRAF.

RESPONSE: Thank you for the helpful suggestion. We have added a new **Response Figure 17** that summarizes the overall experimental design, including the overall ACANet framework, proof-of-concept experiments, backbone and hyperparameter analyses, latent-space/LII investigations, large-scale benchmarking across HSSMS/LSSNS/MMP/ADMET datasets, and comparisons between ACA-aware and traditional QSAR methods.

We also revised the **Discussion** section to explain the rationale behind choosing specific datasets for **Figures 3–6**.

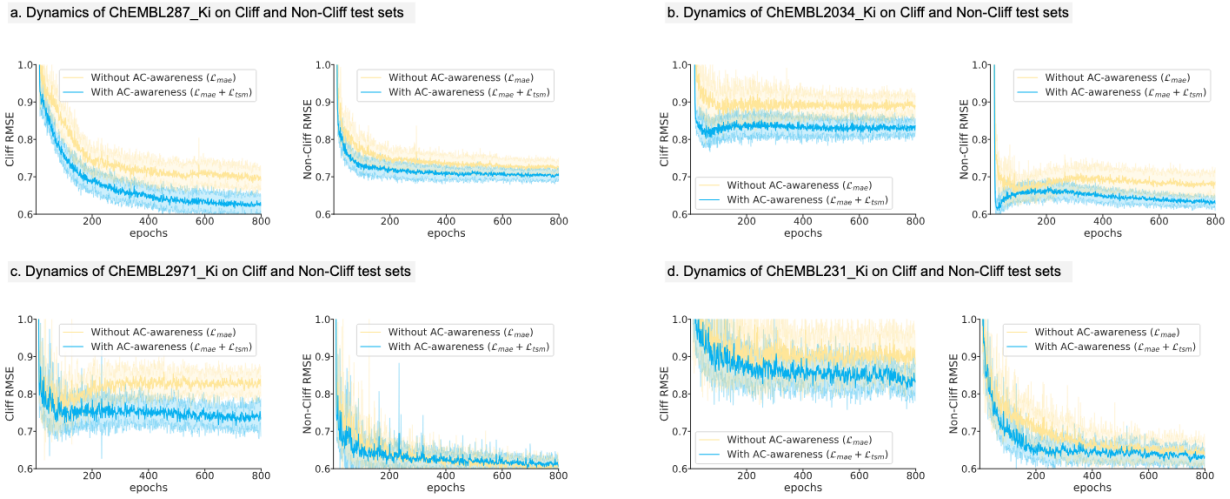
For **Figures 3 and 4**, we focused on **PPARδ** because it is a medium-sized HSSMS dataset with broad scaffold diversity and stable training dynamics. These characteristics make it ideal for clearly showing (i) the effect of AC-awareness on training trajectories, and (ii) fair comparisons across different GNN backbones. Although PPARδ is used as the illustrative example, we observe the same behavior in other datasets. In the revision, we added two additional HSSMS datasets—ChEMBL287_Ki (SOR) and ChEMBL2034_Ki (GR)—showing consistent results (**Response Figure 18 / Supplementary Figure 14**).

For **Figure 5**, **BRAF** was selected because it is a relatively small LSSNS dataset (n = 128), which allows efficient grid search over cliff parameters. Importantly, this choice reflects convenience rather than dataset specificity. We also extended the same ablation study to four additional LSSNS datasets—PLK1, IDO1, USP7, and RIP2—showing the same trend (**Response Figure 3 / Supplementary Figure 2**). In addition, we performed the T-shape cliff search on four HSSMS datasets selected from different target families (Kinase, GPCR, Transferase, and Nuclear Receptor), and the same pattern was observed (**Response Figure 16**).



Response Figure 17: Overview of the experimental and organizational structure of this study. The flowchart groups all main and supplementary figures/tables by topic, including the overall ACANet framework, proof-of-concept experiments, backbone and hyperparameter analyses, latent-space/LII investigations, large-scale benchmarking across HSSMS/LSSNS/MMP/ADMET datasets, and comparisons between ACA-aware and traditional QSAR methods.

For **Figure 6**, BRAF was originally chosen because it contains clear analog series and pronounced activity-cliff behavior, making it suitable for visualizing how AC-awareness reorganizes the latent embedding space and reduces label–structure incoherence. In the revision, we also added a parallel case study on the HSSMS JAK1 dataset—which has the lowest *LII* among all HSSMS datasets—and observed the same improvements in both training and test sets (**Response Figure 11-12 / Extended Data Figures 4–5**).



Response Figure 18 (new Supplementary Figure 14): Training dynamics comparing AC-aware training (blue) versus the baseline (yellow). (a–d) RMSE curves on activity-cliff and non-cliff test sets. AC-aware training consistently improves performance on activity-cliff data compared to the baseline.

Overall, these additional results demonstrate that the observations reported in **Figures 3–6** are **general phenomena**, not dataset-specific artifacts. The datasets highlighted in the main figures were chosen for clarity of illustration and computational convenience, and the expanded analyses confirm that AC-awareness consistently improves model behavior across diverse datasets.

Reviewer #1 (Remarks on code availability):

I checked the code repository and the authors have nicely documented their project in github.

RESPONSE: Thank you very much for reviewing our code repository and for your positive feedback on its documentation.

Response to Reviewer #2:

Summary
This manuscript proposes the novel framework for effective learning of activity-cliff (AC) awareness in modeling molecular activity. In general, ACs make molecular activity prediction challenging in traditional computational chemistry or quantitative structure-activity relationship (QSAR). This challenge arises because compounds with similar structural features can nonetheless display markedly different activities.
In this manuscript, AC-awareness was achieved by the novel contrastive learning-based graph neural network (ACNet) with triplet soft margin (TSM) loss between high-value activity cliff triplets (HV-ACTs). This TSM loss is defined on a group of three molecules; an anchor molecule, a positive and a negative molecule to an anchor, each. Based on the established knowledge, TSM loss reduces the distance between the latent feature of an anchor compound (A) and its corresponding positive (P) while increasing the distance between the anchor (A) and its negative (N). positive To evaluate this pipeline, the authors constructed four scenarios of activity cliffs across 52 molecular datasets, including ADMET property prediction. The ablation studies also showed the effectiveness of the proposed method.

Advantages
The authors clearly conveyed the importance of the potential challenge caused by activity-cliff in molecular modeling or drug development. Especially, traditional QSAR-based method may struggle to these HV-ACTs. In addition, the contrastive learning with TSM loss is appropriately designed in accordance with the motivation of this research and the basic property of activity cliff. The experiments were conducted in various situations and lots of datasets.

RESPONSE: We sincerely thank you for your thoughtful and encouraging evaluation of our work. We greatly appreciate your recognition of the challenges posed by activity cliffs in QSAR and drug discovery, the rationale behind our contrastive learning design with the triplet soft margin loss, and the comprehensiveness of our evaluation across diverse datasets, which we believe reinforces the robustness and significance of the proposed ACNet framework.

Limitations and questions

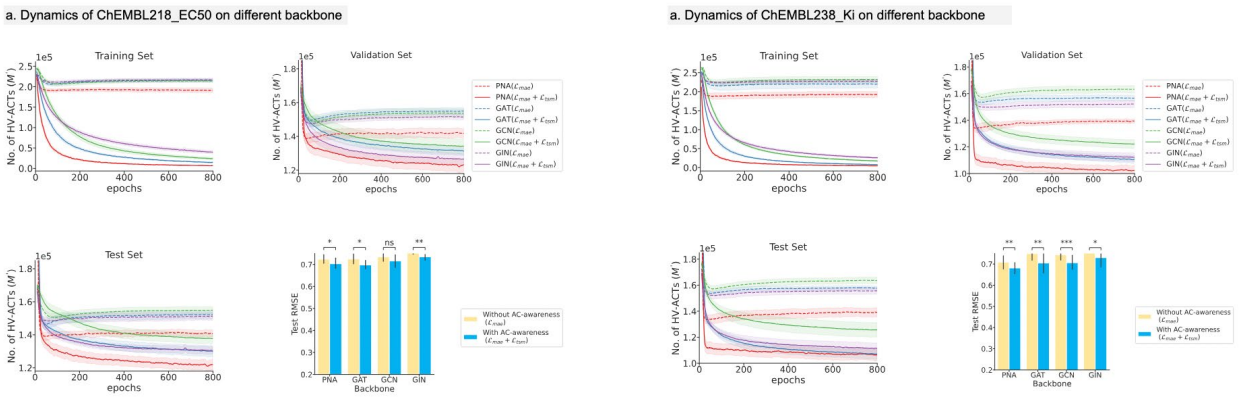
1. Although the authors validated superiority over various tasks compared with other methods, they may not be sufficient to explicitly validate the effectiveness of the propose method. Because 1) in regard to Figure 7, the lowest error values of ACANet may come from the well-optimized parameters in deep networks. How were other networks (PNA, GIN, MPNN, …) optimized for these comparisons?

RESPONSE: We thank you for raising this important point. We employed a parameter-search strategy that was as fair as possible across all models. Moreover, we validated the models using default parameters on the original dataset splits and further performed a 5-fold cross-validation to mitigate potential bias. In both scenarios, the models incorporating ACA consistently demonstrated significant performance improvements.

Uniform hyper-parameter optimisation across backbones.
All deep-learning baselines (PNA, GIN, MPNN, GCN, GAT, etc.) were tuned under an identical search protocol. We defined the same search grids for learning rate, batch size, number of layers and hidden width, and applied an grid search to every architecture. The explored ranges are summarised in the **Response Table 5 / Supplementary Table 10**.

Response Table 5 (new Supplementary Table S10)：Hyperparameter search space for each model.

Model	Hyperparameter	Search space
GCN	Learning rate (lr) Hidden atom features (ha) Num convolutional layers (nc) Multiset transformer nodes (ht), Predictor features (hp)	$\{5\times10^{-4}, 5\times10^{-5}, 5\times10^{-6}\}$ {32, 64, 128, 256, 512} {1, 2, 3, 4, 5} ht: {64, 128, 256, 512}, hp: {128, 256, 512}
GAT	GCN hyperparameters + variant *(see GCN)*	Same as GCN + GATv1 or GATv2 convolution
ACANet	GCN hyperparameters + AC-awareness params cu, cl, α	Same as GCN + $\{cu, cl, \alpha\}$ cu: {0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5}; cl: {0.1, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0, cu}; α: {0, 0.01, 0.05, 0.1, 0.2, 0.5, 1}
MPNN	Learning rate (lr) Hidden atom/edge features (ha/he) Message passing steps (sm) Multiset transformer nodes (ht) Predictor features (hp)	$\{5\times10^{-4}, 5\times10^{-5}, 5\times10^{-6}\}$ {32, 64, 128, 256} {1, 2, 3, 4, 5} {64, 128, 256, 512} {128, 256, 512}
AFP	Attentive layers (na), Timesteps (nt) Predictor features (hp) Learning rate (lr)	{1, 2, 3, 4, 5} {32, 64, 128, 256} $\{5\times10^{-4}, 5\times10^{-5}, 5\times10^{-6}\}$
FFNN (MLP)	Learning rate (lr) Hidden features (nh) Number of layers (nl)	$\{5\times10^{-4}, 5\times10^{-5}, 5\times10^{-6}\}$ {256, 512, 1024} {1, 2, 3, 4, 5}
LSTM	Pretraining Architecture Optimizer, lr Epochs	Next-character prediction on 36k SMILES BatchNorm → LSTM(1024) → LSTM(256) → BatchNorm Adam, 1×10^{-4} 100 (pretrain) + 100 (regression)
1D CNN	Learning rate (lr) Fully connected hidden (nh) Convolution kernel size (nk) Epochs	$\{5\times10^{-4}, 5\times10^{-5}, 5\times10^{-6}\}$ {128, 256, 512, 1024} {4, 8, 10} 500
Transformer (ChemBERTa)	Base model Fine-tuning Learning rate (lr) Epochs	Pretrained on 10M PubChem molecules Freeze all weights, replace pooling with FC regression head 5×10^{-4} 100

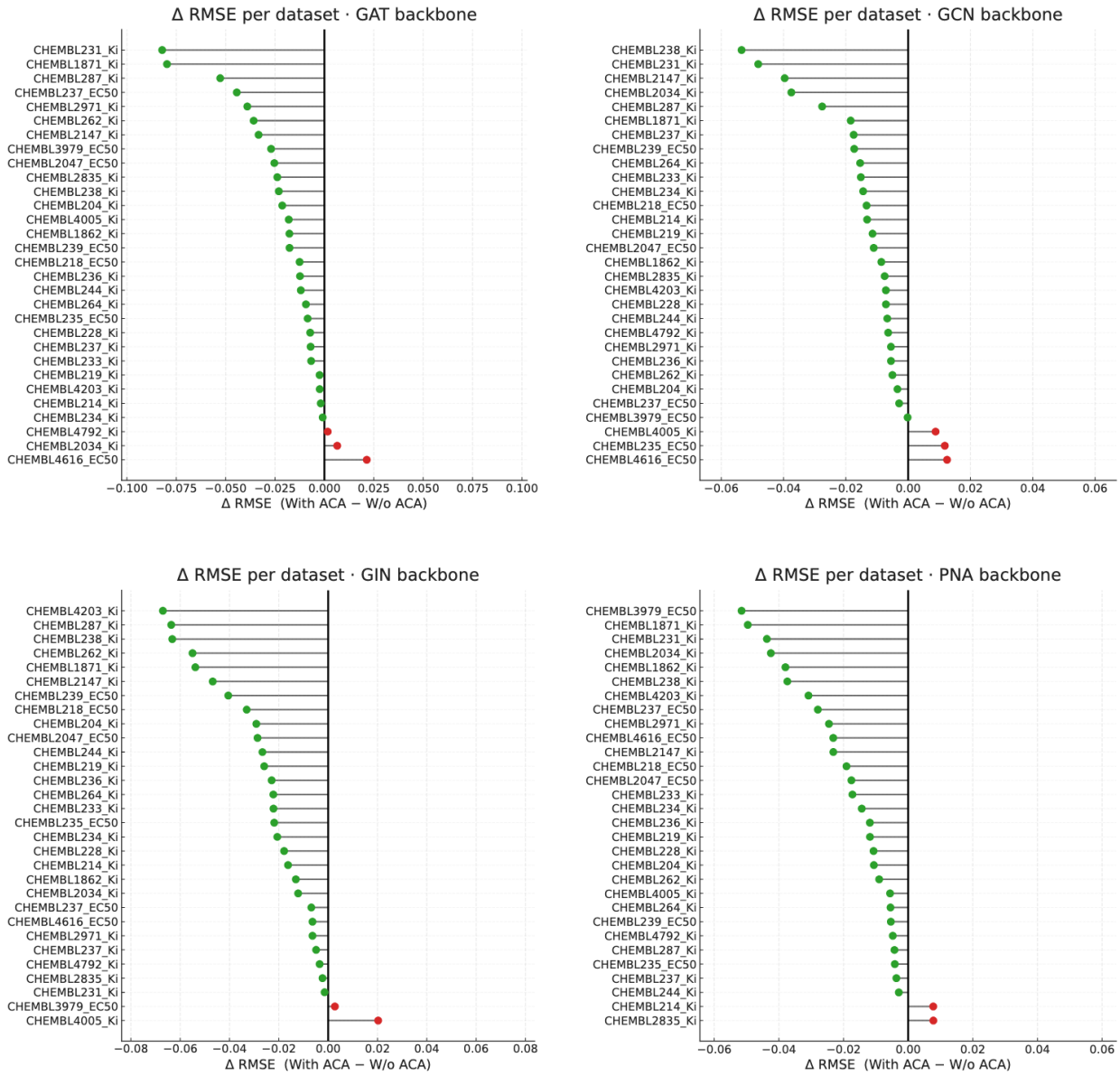


Response Figure 19 (new Supplementary Figure 19): Backbone-agnostic gains from AC-aware training. Training, validation, and test dynamics under fixed hyperparameters (5-fold cross-validation) for two representative targets. Line plots show the number of mined HV-ACTs over epochs; bar charts report test RMSE (mean±s.d.). Across PNA, GAT, GCN, and GIN backbones, adding the AC-aware loss ($l_{mae} + l_{tsm}$) consistently improves accuracy over the baseline (l_{mae}), with significance indicated above bars.

Backbone-agnostic ACA ablation.

To verify that ACA itself—and not merely improved hyper-parameter tuning—contributes to performance gains, we conducted evaluations using two approaches. First, using the original data splits, we applied the ACA loss to four representative backbone models (PNA, GIN, GCN, and GAT) with identical fixed ACA hyper-parameters (cu=1, cl=1, α=1). As shown in **Response Figure 19 / Supplementary Figure 19**, models with ACA consistently yielded clear performance improvements over the corresponding models without ACA under the default data

split. Additionally, to prevent potential bias arising from fixed dataset splits, we conducted unbiased validation using a 5-fold cross-validation strategy. As shownin **Response Figure 20 / Extended Data Figure 6**, models with ACA still achieved significant performance enhancements across all backbone models under the cross-validation split. These experiments demonstrate the intrinsic effectiveness of the ACA method.



Response Figure 20 (new Extended Data Figure 6): ΔRMSE per dataset on ADMET across GAT, GCN, GIN, and PNA backbones under fixed ACA hyperparameters. All ACA-aware models are trained with the same default settings ($cu = 1, cl = 1, \alpha = 1$). For each target, the change in test error is shown as $\Delta RMSE = RMSE_{With\ ACA} - RMSE_{W/o\ ACA}$. Lollipop stems are centered at 0; green circles (left of 0) indicate improvement (lower RMSE with ACA), while red circles (right of 0) indicate degradation. Datasets in each panel are sorted by $\Delta RMSE$. Overall, ACA yields markedly lower RMSE than the non-ACA baseline, demonstrating clear and consistent superiority.

Together, these results confirm that the observed benefits indeed stem from the ACA mechanism itself rather than from hyper-parameter selection alone.

2) Is ACANet clearly have an advantage over traditional (non-machine-learning-based) QSAR methods?

RESPONSE: We appreciate the reviewer's question. Yes, ACANet exhibits clear advantages over traditional (non-machine-learning-based) QSAR methods. To demonstrate this rigorously, we considered two major categories of traditional QSAR approaches:

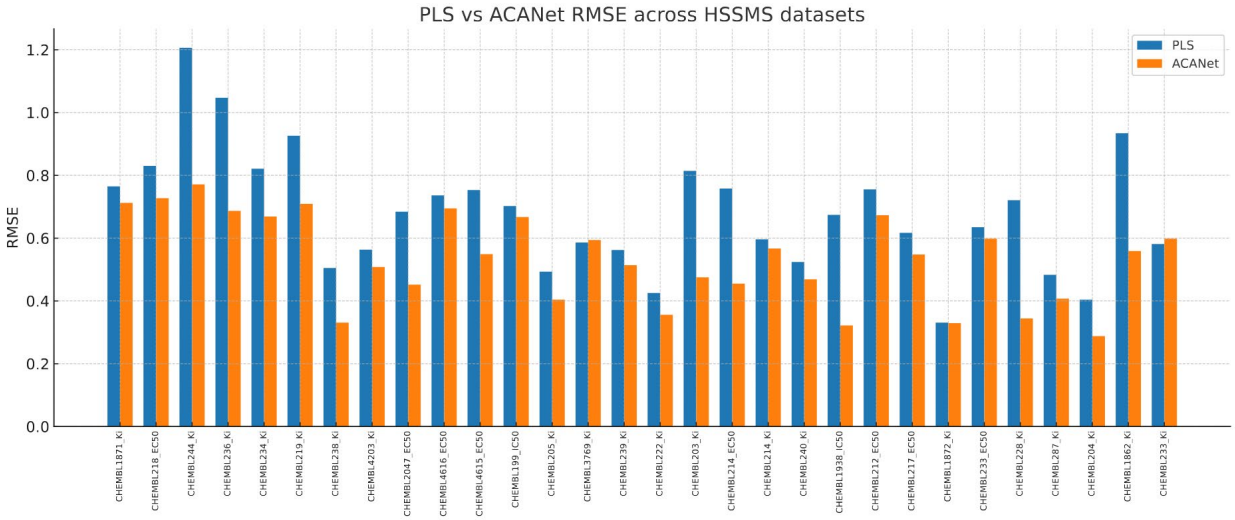
1. **Scaffold-based methods** (e.g., Hansch analysis and the Free-Wilson approach). These assume a single molecular scaffold; consequently, they perform poorly on datasets containing multiple scaffolds because a separate model must be trained for each scaffold, which limits generalizability.
2. **Descriptor-based statistical models** that rely on intrinsic molecular properties (e.g., topological and physicochemical descriptors). Such methods possess a more limited capacity to capture complex structure–activity relationships compared with modern deep-learning models.

For quantitative comparison, we selected one representative from each category—**Free-Wilson** (scaffold-based) and **Partial Least Squares (PLS)** (descriptor-based)—and benchmarked them against ACANet:

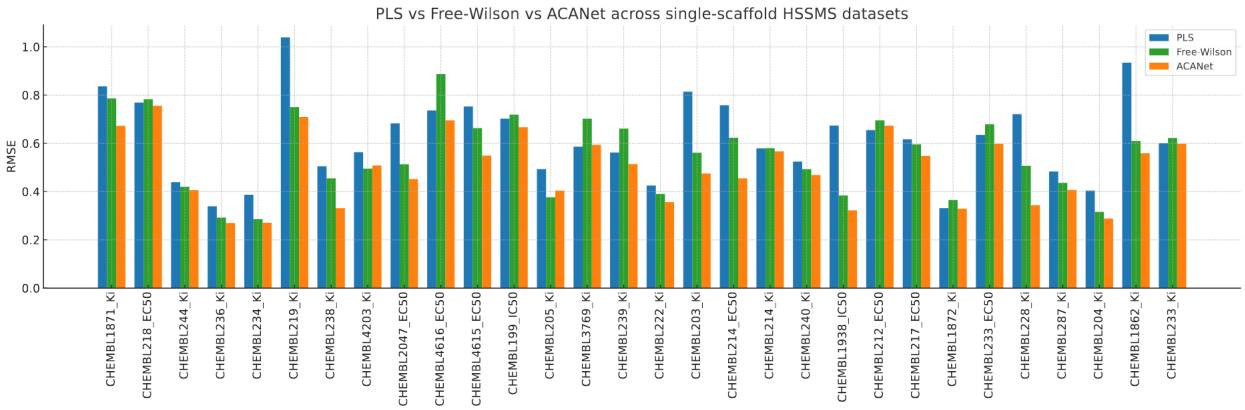
- **Free-Wilson:** We constructed a subset of the HSSMS dataset containing only molecules with the most frequent scaffold (see **Response Table 6 / Supplementary Table 11**)).
- **PLS:** We trained on the full, original HSSMS dataset.

Response Table 6 (new Supplementary Table S11): Datasets statistics across single-scaffold HSSMS

Dataset	Top_Scaffold	Train Num	Test Num
CHEMBL1871_Ki	O=c1ccc2ccccc2[nH]1	48	11
CHEMBL218_EC50	O=c1cc(NC2CCNCC2)c2cc(C(c3ccccc3)c3ccccc3)ccc2[nH]1	32	4
CHEMBL244_Ki	c1ccc(Oc2cccc(Oc3ccccc3)n2)cc1	75	15
CHEMBL236_Ki	c1ccc(C2CCN(C(c3ccccc3)c3ccccc3)CC2)cc1	47	9
CHEMBL234_Ki	c1cc2c(c(N3CCN(CCC4CCCCC4)CC3)n1)CCO2	32	7
CHEMBL219_Ki	O=C(CN1CCN(Cc2ccccc2)CC1)N1CCc2ccccc21	39	13
CHEMBL238_Ki	c1ccc(C2CC3CCC(C2)N3)cc1	51	15
CHEMBL4203_Ki	O=C(Nc1ccccn1)Nc1ccnc2ccccc12	15	3
CHEMBL2047_EC50	C1CCC2C(C1)CCC1C3CCCC3CCC21	31	10
CHEMBL4616_EC50	c1ccc(COCCc2nnn[nH]2)cc1	31	12
CHEMBL4615_EC50	Cc1ccc(C(=O)N2CCC(CC2)NC(=O)CN2C(=O)C(c3ccccc3)C3CCCN32)cc1	29	12
CHEMBL199_IC50	c1ccc(CNC(NC(=O)c2cccc[nH]2)C(C)C)cc1	60	15
CHEMBL205_Ki	O=c1c(OC2CCCC2)nc2ccccc2[nH]1	29	11
CHEMBL3769_Ki	N#Cc1cn([C@H]2CN(c3ccc(F)cc3)C(=O)O2)nn1	43	11
CHEMBL239_Ki	Cc1ccc(C(=O)NCCN2CCCC2)cc1	55	15
CHEMBL222_Ki	c1ccc2c(c1)cnn2C1CCCNC1	34	6
CHEMBL203_Ki	Cc1cccc(CNC(=O)C(C)N)cc1	58	14
CHEMBL214_EC50	Cn1cc(C2CCC(N3CCC(C(=O)N4CC=C5CCC(C)C5C4)CC3)CC2)cn1	35	8
CHEMBL214_Ki	Cn1cc(C2CCC(N3CCC(C(=O)N4CC=C5CCC(C)C5C4)CC3)CC2)cn1	57	13
CHEMBL240_Ki	c1ccc2c(c1)ccc(N1CCC(N2C(=O)NCC3CC3)CC1)c2	68	17
CHEMBL1938_IC50	c1ccccc(C(=O)N(C)C)cc1	46	12
CHEMBL212_EC50	Cc1ccc(C(=O)N2CCC[N+](C)(C)CC2)cc1	22	7
CHEMBL217_EC50	O=C(Nc1ccc(O)cc1)Nc1ccc(CN2CCOCC2)cc1	34	9
CHEMBL1872_Ki	Cc1ccc(CNc2nc3c(c(N)n2)OCC(O)C3O)cc1	36	9
CHEMBL233_EC50	COC(=O)c1ccc(OCCCN2CCCC2)cc1	43	11
CHEMBL228_Ki	Cc1ccc(CN(C)C(=O)c2ccc[nH]2)cc1	44	11
CHEMBL287_Ki	c1ccc2ccccc2c1	31	7
CHEMBL204_Ki	Cc1ccc(NC(=O)NC)cc1	37	8
CHEMBL1862_Ki	N#Cc1cc[nH]c(=O)[nH]1	41	10
CHEMBL233_Ki	COC(=O)c1ccc(OCCCN2CCCC2)cc1	44	10



Response Figure 21 (new Supplementary Figure 17): PLS vs. ACANet on HSSMS. Per-dataset RMSE comparison (lower is better) across the HSSMS benchmark shows ACANet consistently outperforming PLS on most targets.



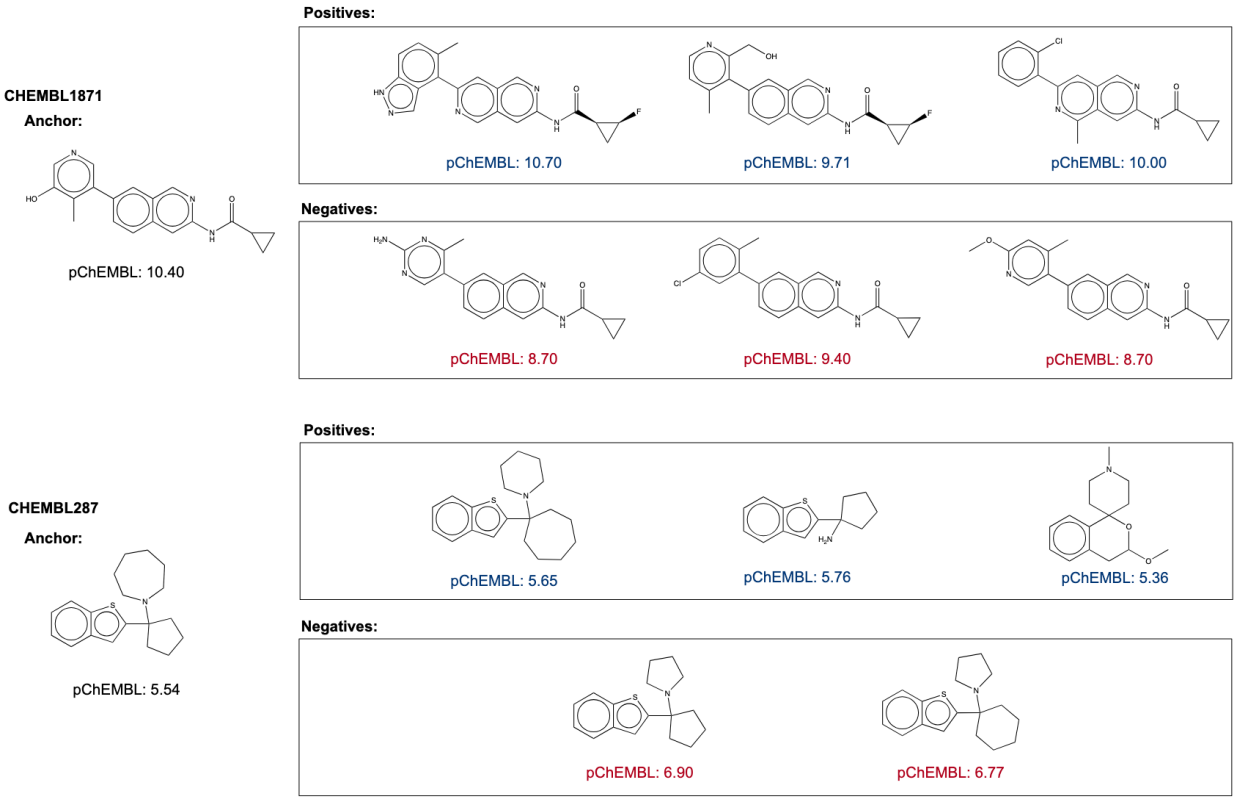
Response Figure 22 (new Supplementary Figure 18): PLS vs Free-Wilson vs ACANet on single-scaffold HSSMS datasets. Per-dataset RMSE (lower is better) comparing PLS, Free-Wilson, and ACANet. ACANet attains the lowest error on most targets.

As shown in **Response Figure 21 / Supplementary Figure 17** and **Response Figure 22 / Supplementary Figure 18**, ACANet consistently and substantially outperforms both the Free-Wilson and PLS models across all evaluated metrics, underscoring its superior predictive power and enhanced generalizability.

Questions on datasets and data embedding:

4) Is there any cases that lots of positives or negatives, corresponding to one anchor?

RESPONSE: Yes, such cases are possible. Within a single mini-batch, one anchor may correspond to multiple positive as well as multiple negative samples. We have included representative examples in **Response Figure 23** to illustrate an anchor associated with numerous positive and negative samples.



Response Figure 23: Examples of a single anchor associated with multiple positive and negative samples. Shown for two targets (CHEMBL1871 and CHEMBL287): each anchor (left) is paired with several positives (top) and negatives (bottom).

5) Is there any dataset having class or data imbalances? For example, is it possible that one anchor may have up to 50 positives and less than 5 negatives? If it can be possible, then what is the potential problem caused by this imbalance and how ACANet can handle?

RESPONSE: Thank you for raising this important point. Yes—depending on the label distribution of a dataset, it is indeed possible that an anchor may have many "positives" but only a few "negatives," or vice versa. This imbalance arises naturally from the activity distribution rather than from the ACANet framework itself. For instance, if the dataset contains many analogs with similar activities but only a few molecules exhibiting large activity differences, an anchor could have dozens of positives but only a small number of negatives.

In principle, such an imbalance could bias the triplet mining process if every candidate triplet were treated equally, because an overabundance of easy positives (or easy negatives) might overwhelm the more informative cases. However, ACANet mitigates these issues in several ways:

1) Online mining selects only high-value triplets (M′). Although an anchor may have many conditional positives/negatives (M), only those triplets whose soft-margin loss is positive (i.e., triplets that violate the desired ordering in latent space) are mined and used for optimization (M′). This means that even if the raw distribution is imbalanced, the model only focuses on the subset of hard and informative triplets, effectively normalizing the imbalance.

2) Self-balanced gradients.The soft-margin formulation encourages the model to ignore "easy" positives or negatives—these contribute zero loss and do not affect training. Thus, anchors with many easy positives do not distort optimization because those triplets simply vanish from M′.
3) Label-defined windows (cl, cu) naturally control imbalance. The cliff parameters act as a data-adaptive filter. In datasets where positives greatly outnumber negatives (or vice versa), cl and cu can be tuned so that the conditional ACT set (M) is neither too small nor overly dominated by a single class. As shown in our ablation studies, optimal cl and cu values differ across datasets, precisely because they adapt to dataset-specific imbalances.
4) Latent-space dynamics reduce imbalance over training. When AC-awareness is enabled, the number of high-value triplets (M′) decreases over time as the latent space becomes better structured. This effect is consistent across both narrow-scaffold (LSSNS) and mixed-scaffold (HSSMS) datasets, indicating that ACANet converges even under initially imbalanced sampling.

Summary. While activity-distribution–induced imbalance among positives and negatives is common in real QSAR datasets, it does not harm ACANet in practice. The online high-value mining strategy, together with the soft-margin formulation and adjustable cliff parameters, ensures that training automatically emphasizes the most informative triplets and remains robust even when the raw triplet distribution is highly imbalanced.

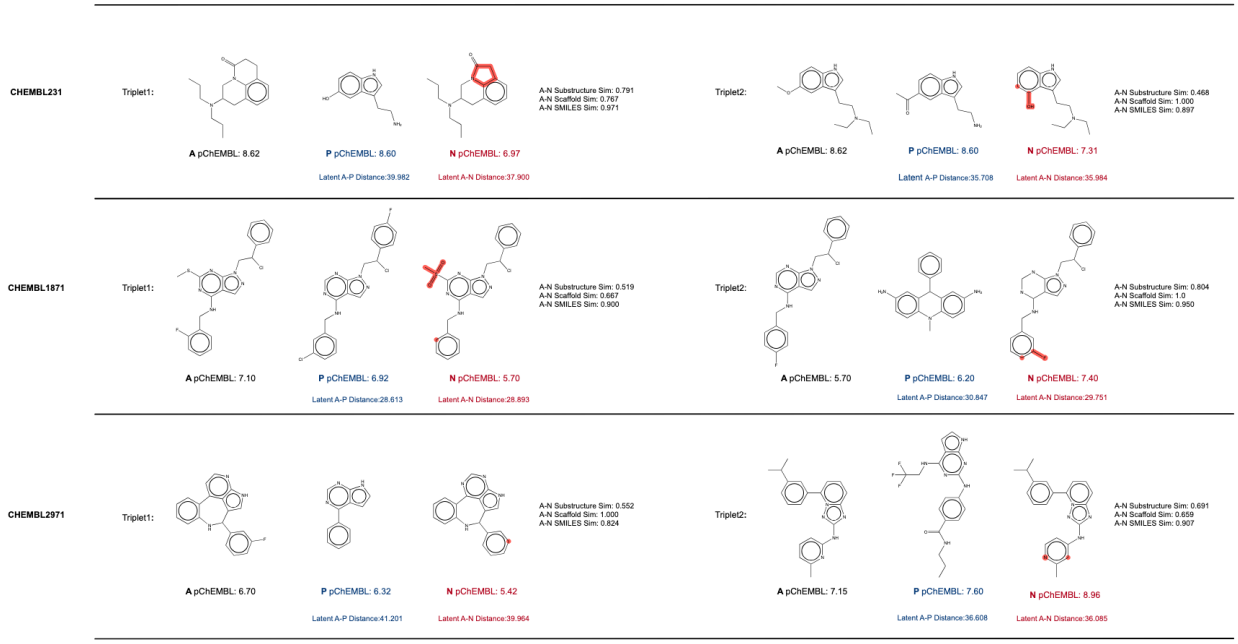
Additionally, I would suggest that the authors should supplement more background about the definition and nature of activity cliffs, including that the semantic or functional similarity between anchor and its positive one is naturally close to that between anchor and its negative one.

RESPONSE: Thank you for your suggestion. We have added more background on activity cliffs in the main section and provided additional details regarding their definition in the Methods section.

Additional ask for the evaluation

1. In Figure 4, the number of HV-ACTs decreases during both training and inference process, when the model was trained with TSM losses. Is it possible to show several hard cases of HV-ACT which remains the last time of training or inference?

RESPONSE: Of course. As shown in **Response Figure 24**, we provide examples of HV-ACTs that remain at the final stage of training for three different targets.



Response Figure 24: Examples of the final mined triplets. For three representative targets, we show the last triplets (A/P/N) at the end of training with their pChEMBL values and pairwise A--P/A--N distances; highlighted marks indicate key substructure changes and the reported scaffold/SMILES similarities.

These triplets at the final stage typically exhibit large structural differences between the anchor and the positive despite their similar activities, and vice versa. This explains why these samples still require optimization at the final stage.

2. As I mentioned in Limitations and questions, is it possible to show another evaluation to validate advantages of ACANet over other QSAR methods? Either quantitative or case-study-based approach may be good.

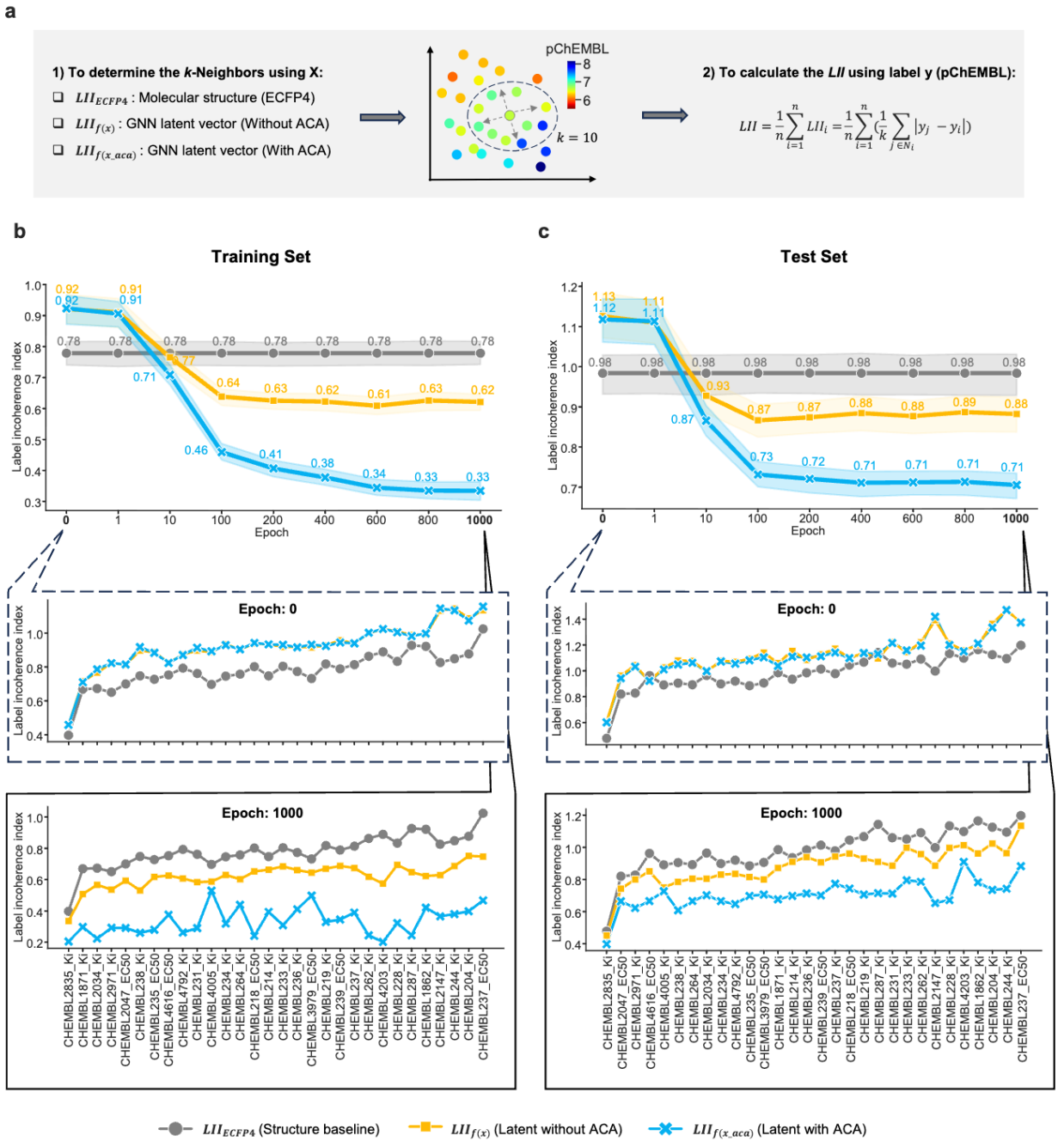
RESPONSE: Of course. In addition to the performance improvements discussed previously, we also employed the Label Incoherence Index (LII) as an independent evaluation criterion. LII measures how closely a model's feature representation aligns with the true distribution of its labels. It is calculated as

$$LII = \frac{1}{n} \sum_{1}^{n} \frac{1}{k} \left(\sum_{j \in N_i} |y_j - y_i| \right),$$

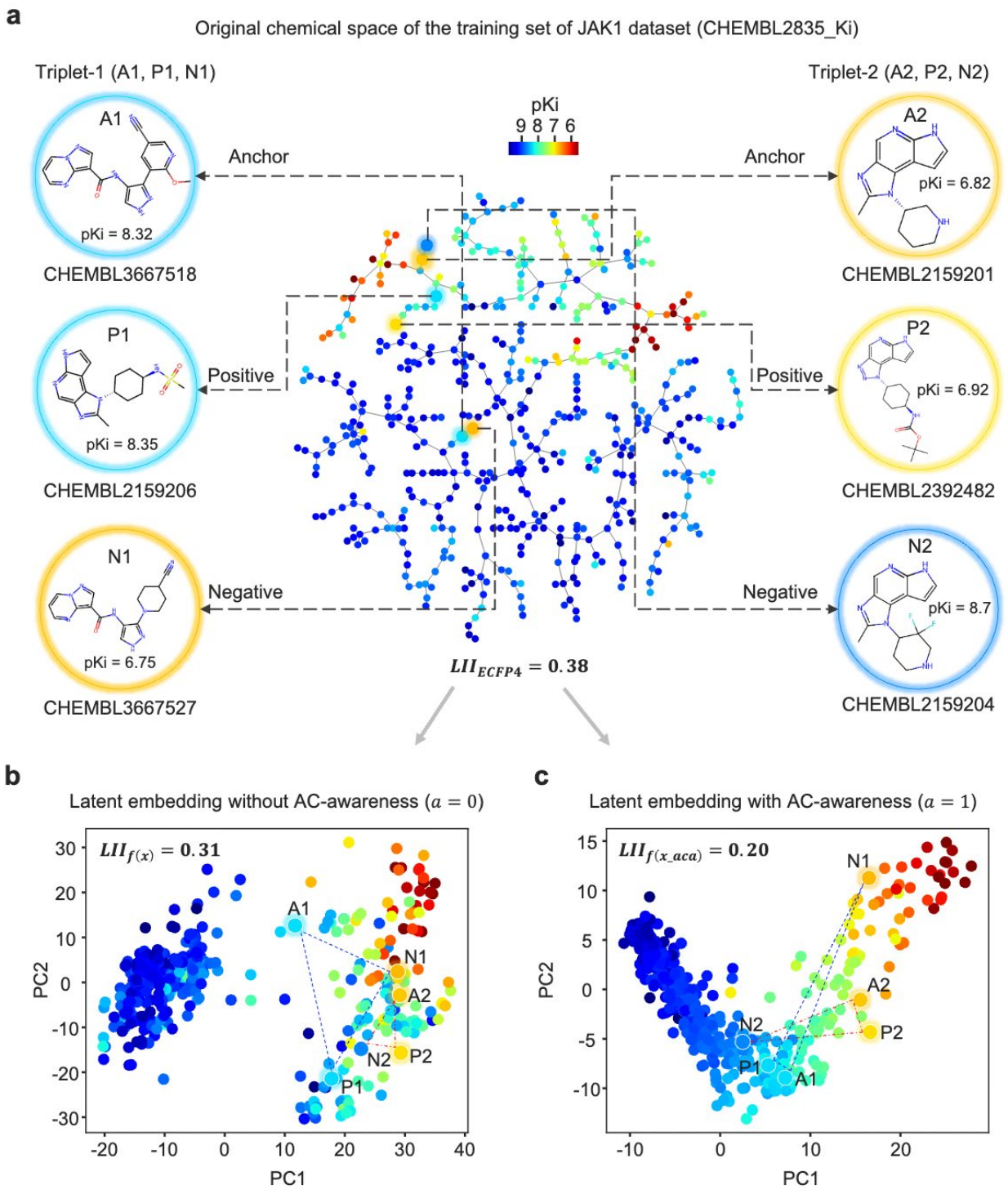
where N_i is the set of k nearest neighbors of molecule i in the feature space and y_i is its measured activity. A lower LII means that the feature distribution is consistent with the molecules' true properties, which is critical for molecular visualization, clustering, and the overall interpretability and trustworthiness of QSAR models. We computed LII for three molecular representations:

- ECFP4 fingerprints
- GNN embeddings without ACA
- ACANet embeddings (GNN with ACA)

Response Figure 13 / Extended Data Figure 1 shows that incorporating ACA markedly lowers LII compared with both ECFP4 and the vanilla GNN across all datasets, indicating a closer match between the learned feature distribution and true molecular activity. **Response Figure 11-12 / Extended Data Figures 4–5** illustrate that ACANet cleanly separates compounds involved in activity cliffs, whereas the non-ACA representation does not. In the two representative triplet examples shown, the ACA-trained GNN clearly separates activity-cliff (AC) molecules in feature space while clustering molecules with similar activities together. In contrast, the non-ACA feature distribution is disorganized and inconsistent with the true activity landscape, failing to differentiate AC pairs clearly in both triplets. These findings demonstrate that ACANet produces more coherent and robust feature representations.

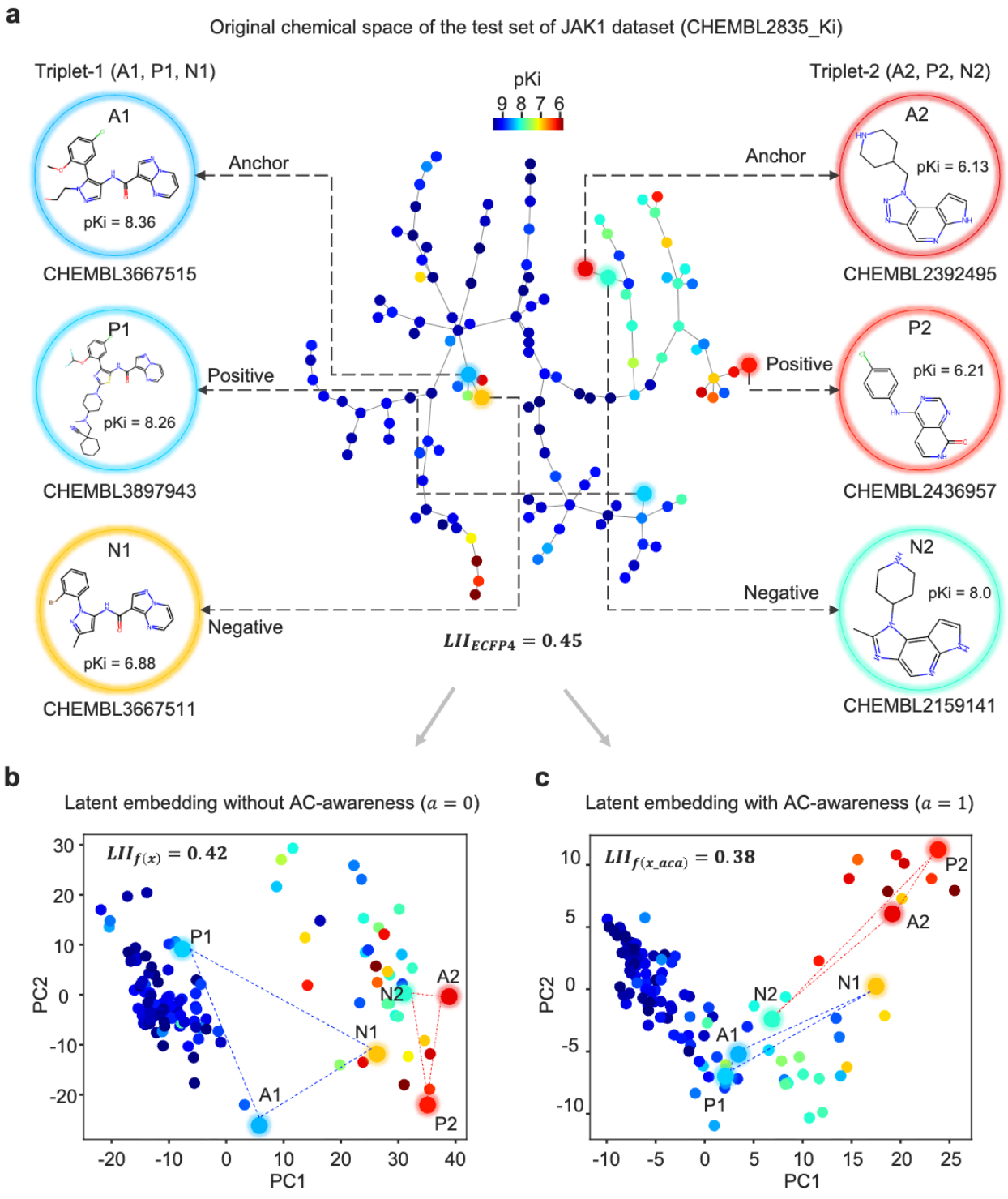


[Response Figure 13 (new Extended Data Figure 1)]: Temporal evolution of the label-incoherence index (LII) in the HSSMS benchmark. Lower LII values indicate greater local label coherence. (a) LII calculation workflow – Step 1: for each compound, identify the $k = 10$ nearest neighbours using one of three representations: ECFP4 fingerprints, GNN latent vectors without ACA, or GNN latent vectors with ACA. Step 2: compute LII from the pChEMBL labels of the query compound and its neighbours. **(b) Training set** – Epoch-wise LII dynamics for the three representations (grey = ECFP4, yellow = latent without ACA, blue = latent with ACA), accompanied by per-dataset LII values at epoch 0 and epoch 100. **(c) Test set** – Corresponding LII dynamics for held-out compounds, with per-dataset snapshots at epoch 0 and epoch 100. Across all datasets, ACA persistently drives the latent-space LII below both the ECFP baseline and the non-ACA model.



[Response Figure 11 (new Extended Data Figure 4)]: Impact of activity-cliff awareness on latent organisation of the JAK1 training set (CHEMBL2835_Ki). (a) tmap visualisation of the test-set chemical space. All test compounds are positioned by ECFP4 similarity and coloured by pKi (blue → low, red → high). Two representative activity-cliff triplets are highlighted: A_1–P_1–N_1 (left) and A_2–P_2–N_2 (right). Each triplet comprises an anchor (A), a potency-matched positive (P), and a lower-potency negative (N); the two triplets belong to

different Bemis–Murcko scaffolds. The structure-based label-incoherence index is $LII_{ECFP4} = 0.38$. **(b) Latent embedding without AC awareness (α = 0).** Two-dimensional PCA of GNN embeddings trained without ACA shows that, for both scaffolds, anchors (A) remain closer to their negatives (N) than to their potency-matched positives (P), leaving scaffold and potency information entangled ($LII_{f(X)} = 0.31$). **(c) Latent embedding with AC awareness (α = 1).** Including the ACA loss reorganises the latent space: anchors cluster tightly with their positives and are clearly separated from their negatives, even for these two distinct scaffolds, reducing the index to $LII_{f(X_{ACA})} = 0.20$. This demonstrates that ACA improves local label coherence on the test set and effectively handles activity cliffs across different scaffolds.



[Response Figure 12 (new Extended Data Figure 5)]: Impact of activity-cliff awareness on latent organisation of the JAK1 test set (CHEMBL2835_Ki). (a) tmap visualisation of the test-set chemical space. All test compounds are positioned by ECFP4 similarity and coloured by pKi (blue → low, red → high). Two representative activity-cliff triplets are highlighted: A_1–P_1–N_1 (left) and A_2–P_2–N_2 (right). Each triplet comprises an anchor (A), a potency-matched positive (P), and a lower-potency negative (N); the two triplets belong to different Bemis–Murcko

scaffolds. The structure-based label-incoherence index is $LII_{ECFP4} = 0.45$. **(b) Latent embedding without AC awareness (α = 0).** Two-dimensional PCA of GNN embeddings trained without ACA shows that, for both scaffolds, anchors (A) remain closer to their negatives (N) than to their potency-matched positives (P), leaving scaffold and potency information entangled ($LII_{f(X)} = 0.42$). **(c) Latent embedding with AC awareness (α = 1).** Including the ACA loss reorganises the latent space: anchors cluster tightly with their positives and are clearly separated from their negatives, even for these two distinct scaffolds, reducing the index to $LII_{f(X_{ACA})} = 0.38$. This demonstrates that ACA improves local label coherence on the test set and effectively handles activity cliffs across different scaffolds. PCA projections of compound embeddings for the HSSMS training set.

Reviewer #2 (Remarks on code availability):

I can run the source code without any problem.

RESPONSE: Thank you very much for confirming the reproducibility of our code. We greatly appreciate your effort in reviewing the implementation and are glad to hear that it runs smoothly.

Response to Reviewer #3:

In the manuscript entitled “Activity Cliff-Informed Contrastive Learning for Molecular Property Prediction” by Shen et al. a novel method for activity cliff awareness is introduced into a graph based deep learning model. Activity cliffs are an important aspect of drug discovery and it has lately been shown by numerous groups that even nonlinear deep learning models cannot accurately capture these effect. Here, a novel method is introduced in order to be able to reflect the activity cliff also in the latent space representation. The method is of high relevance, however before publication I have a number of comments which would show the real value of the inclusion of the AC awareness into the model.

RESPONSE: We sincerely thank you for the thoughtful evaluation of our work and for highlighting the importance of activity cliffs in drug discovery. We greatly appreciate your recognition of the relevance of our proposed method and the novelty of incorporating activity-cliff awareness into graph neural networks. We also acknowledge your comments and agree that addressing them will further demonstrate the value of AC-awareness in our model. In the revised manuscript, we have carefully considered your feedback and will provide additional clarifications, experiments, and discussions as necessary to strengthen the contribution and impact of our work.

Main

•1st paragraph: Nonadditivity has only been mentioned on the side in the introduction, however, this is basically the same effect analyzed in a different way. Also here, different people have been working on the topic, i.e.

o https://pubs.acs.org/doi/full/10.1021/acs.jcim.9b00631

o https://link.springer.com/article/10.1186/s13321-021-00525-z

o https://pubs.acs.org/doi/full/10.1021/acsomega.2c02738

o https://link.springer.com/article/10.1007/s10822-024-00566-0

RESPONSE: We thank you for highlighting the important link between activity cliffs (ACs) and non-additivity in molecular modelling. We fully agree that ACs and non-additive effects are intimately related, as both arise when simple additive assumptions fail. We appreciate the reviewer’s recommended literature, which offers valuable insights into this subject.

In the revised manuscript, we have added a dedicated discussion of non-additivity, clarified its connection to ACs, and cited all the suggested references.

Although the two concepts are related, our work focuses on explicitly integrating AC information into deep-learning frameworks, thereby offering a practical and effective approach to improve QSAR performance and addressing concerns raised in previous studies.

•2nd paragraph, line 31: How is structurally similar defined?

RESPONSE: In our method, structural similarity is defined in the GNN latent space, where we employ a soft margin $M = y_j^A - y_j^N - \left|y_j^A - y_j^P\right|$ as the threshold to determine whether two molecules are considered similar, as shown in **Figure 1c**. We also conducted comparision and ablation experiments with three traditional string-based similarity metrics, including the

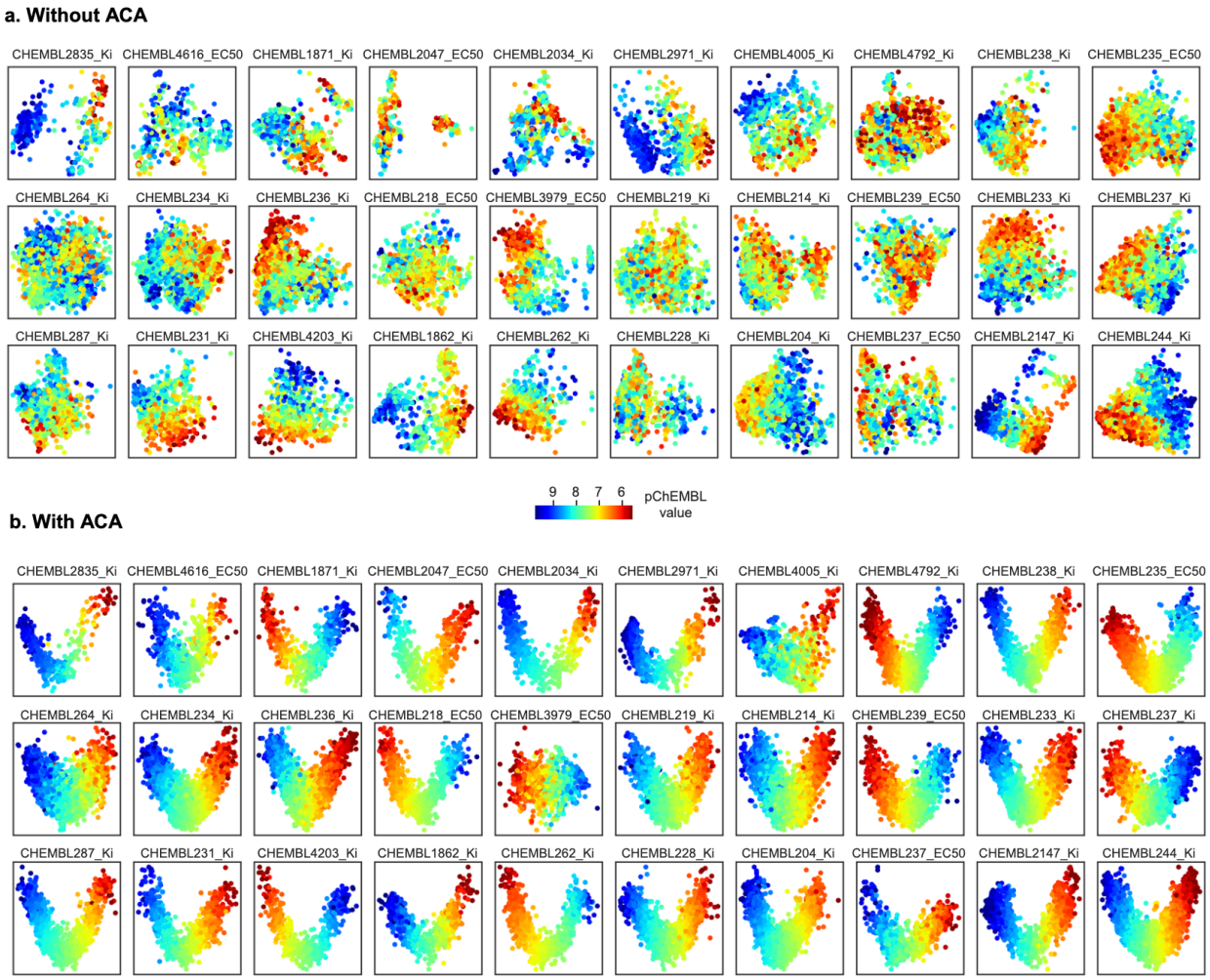
Tanimoto coefficient computed from molecular fingerprints, Tanimoto coefficient computed from scaffold fingerprints and SMILES edit similarity to serve as alternative structural-similarity criteria. Additional details can be found in our **[Response to Reviewer 1, Major Concern 1].**

•l 63: BRAF dataset – please add a reference

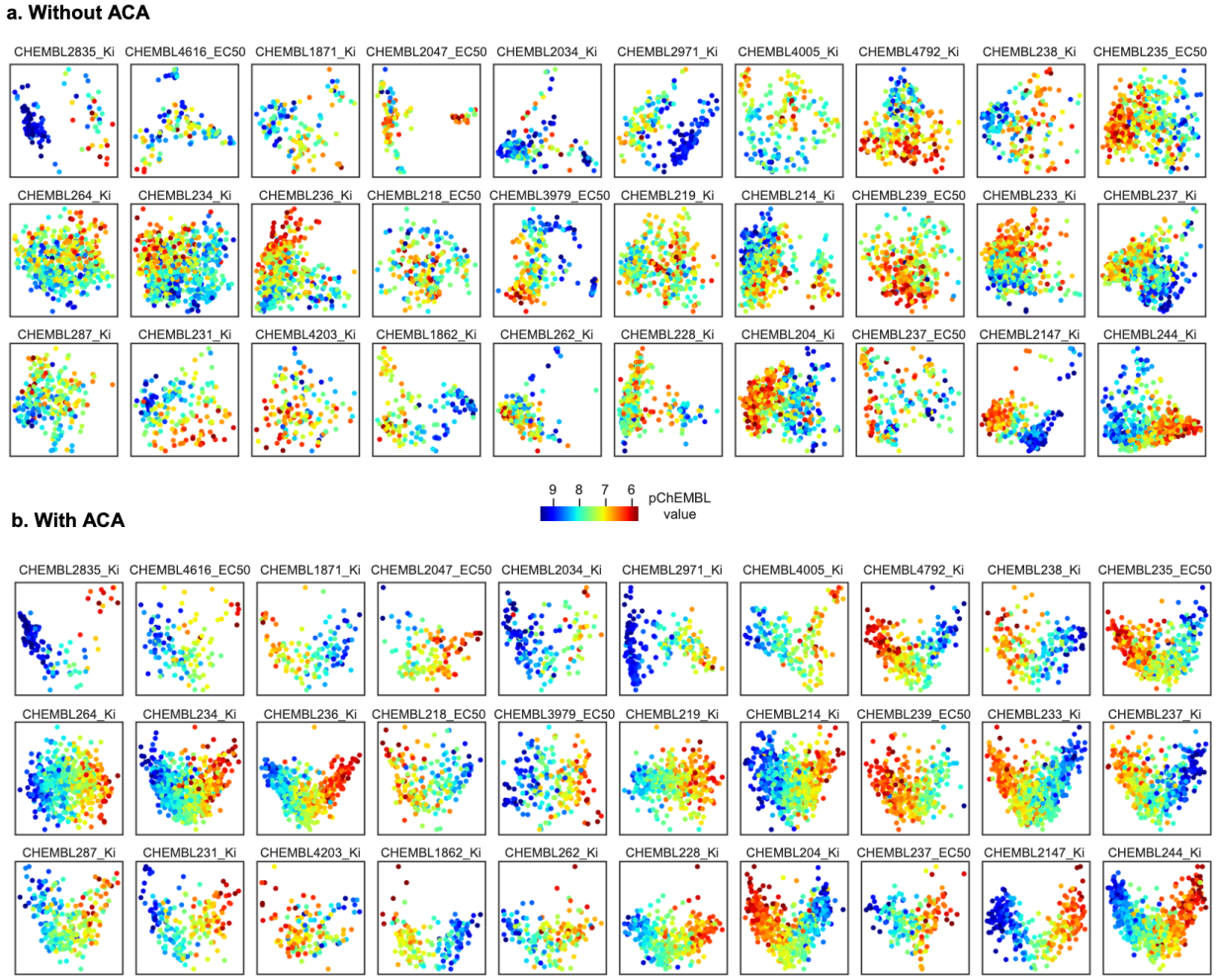
RESPONSE: Thank you for pointing this out. We have added the appropriate reference for the BRAF dataset in the revised manuscript.

•Ll 67-68: “more informative molecular representation”: What do you mean here? In what sense are these representations more informative and for whom?

RESPONSE: Thank you for requesting clarification on this point. By “more informative molecular representation,” we mean that, unlike standard methods, which directly learn representations solely from molecular property labels, our approach incorporates additional information derived from ACTs via an auxiliary ACT loss. This encourages the model to learn representations where AC samples are effectively distinguished, thereby creating a more structured and discriminative latent space. Such a latent space not only enhances predictive accuracy for molecular properties but is also valuable for related representation-based tasks, such as molecular clustering and visualization (see **Response Figure 11,12,14,15 / Extended Data Figure 2,3,4,5**).



[Response Figure 14 (new Extended Data Figure 2)]: PCA projections of compound embeddings for the HSSMS training set, coloured by pChEMBL value (blue → low, red → high). (a) Without ACA regularisation – Two-dimensional PCA embeddings from the baseline model display fragmented and discontinuous activity landscapes, indicative of latent-space activity cliffs. **(b) With ACA regularisation** – PCA embeddings learned with ACA show smoother and more coherent activity gradients, demonstrating improved preservation of structure–activity relationships and reduced cliff severity.



[Response Figure 15 (new Extended Data Figure 3)]: PCA projections of compound embeddings for the HSSMS test set, coloured by pChEMBL value (blue → low, red → high). (a) Without ACA regularisation – Two-dimensional PCA embeddings from the baseline model display fragmented and discontinuous activity landscapes, indicative of latent-space activity cliffs. **(b) With ACA regularisation** – PCA embeddings learned with ACA show smoother and more coherent activity gradients, demonstrating improved preservation of structure–activity relationships and reduced cliff severity.

•L 71: ACANet first time mentioned, so it would be good to know what the abbreviation meant, short explanation if Net also stands for something or if this is just the name you gave it.

RESPONSE: Thank you for your suggestion. We have revised the sentence accordingly to: **"Furthermore, AC-awareness network (ACANet) models outperform molecular fingerprint-based models on 70%–76% of the HSSMS benchmark datasets."**

•Last paragraph in general: A lot of results are mentioned here already, I would prefer to have those in the results section instead of here o LI 66, 71, 72 especially

RESPONSE: Thank you for your suggestion. We have revised this part and moved the detailed descriptions of the results to the Results section.

Results

•Line 86: How are the triplets identified? How are triplets assembles for test data, i.e. prospective use cases? This part needs more explanation.

RESPONSE: We apologize for not clearly explaining the process of assembling the activity-cliff triplets, which may have caused confusion. AC triplets consist of an anchor molecule, a positive molecule, and a negative molecule. Suppose we have a batch containing N molecules. Each molecule can serve as an anchor to identify triplets. We identify negative samples for each anchor by selecting molecules within the batch whose activity differences from the anchor exceed the predefined cliff upper threshold, while also satisfying structural similarity criteria (e.g., embedding-based or traditional string-based similarity criteria as detailed in our **[Response to Review 1, Major Concerns 1]**). Such pairs are considered anchor–negative pairs, that is, AC pairs. Conversely, positive samples are molecules whose activity differences from the anchor are below the cliff lower threshold. The identified anchor, positive, and negative samples then constitute an activity-cliff triplet, and our goal is to optimize the model based on these triplets.

During inference, ACANet does not perform the ACT construction process required during training, nor can we obtain ground-truth labels to construct ACTs at inference time. Instead, the property-prediction network, which has already been optimized with ACTs during training, can be directly employed for molecular property prediction.

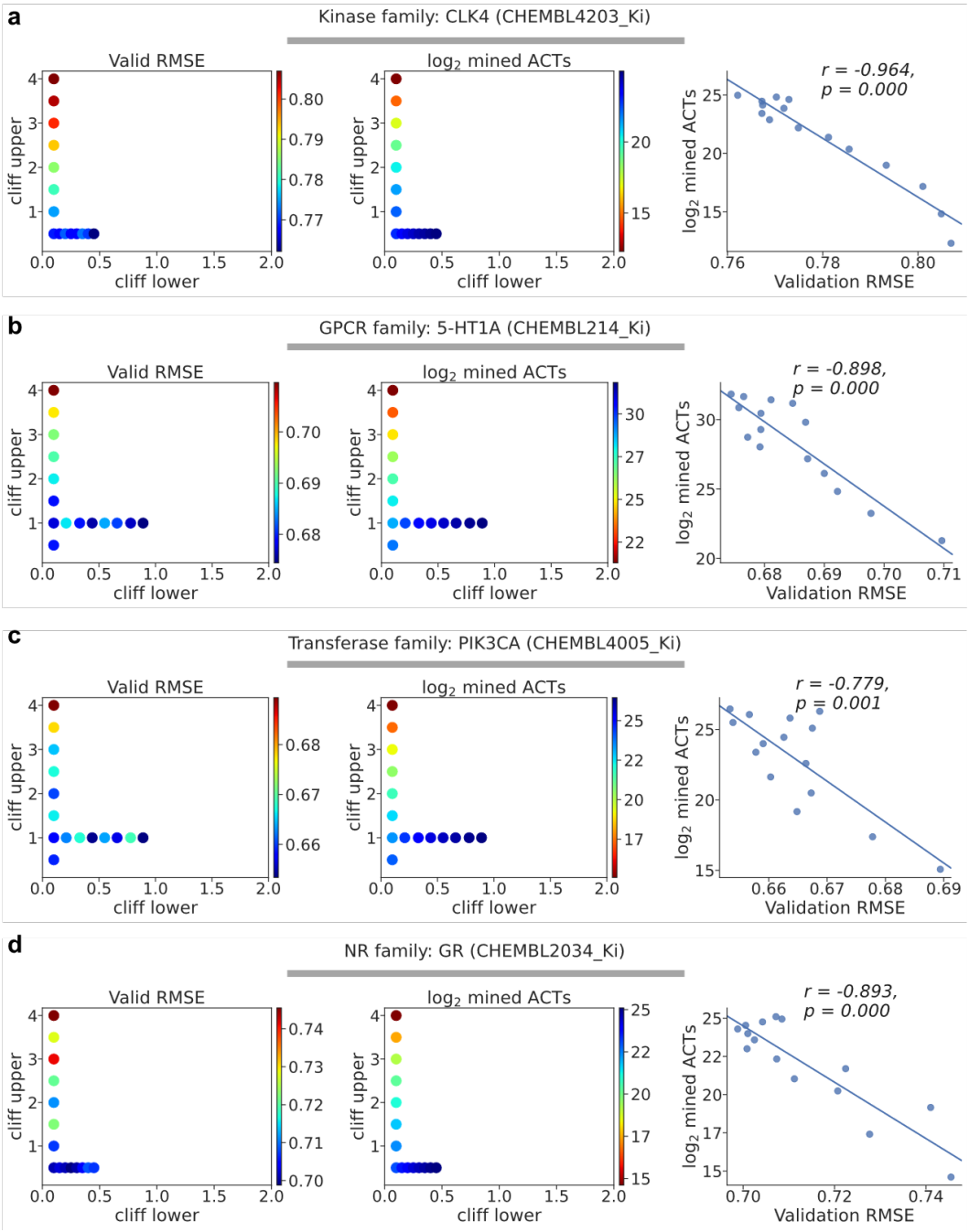
•L 99: "number of mined HV-ACTs": how is this number determined, is there a max/min you set in your methods – This needs further explanations. o How is decided how many triplets are considered? o What is the effect on the model accuracy? o Extension: Can you use the results from the triplets for uncertainty estimates for the test compounds?

RESPONSE: Thank you very much for raising these valuable questions and suggestions. In our approach, we do not impose explicit maximum or minimum constraints on the number of mined HV-ACTs. Instead, this number naturally emerges from the set of valid ACTs formed by all anchors within a batch.

(1) Specifically, as detailed in our **[Response to Line 86]**, given a batch of (M) molecules, each molecule (M_i) can serve as an anchor. If the numbers of corresponding positive and negative samples for (M_i) are (P_i) and (N_i), respectively, then the total number of ACTs mined in this batch is calculated as $K = \sum_{i=1}^{N} P_i \times N_i$.

(2) We have investigated the relationship between the number of ACTs and model accuracy. As shown in **Figure 5a**, **Supplementary Figure 2 and Response Figure 16**, there is a clear positive correlation between the number of ACTs and the model's performance.

(3) We appreciate your insightful suggestion regarding uncertainty estimation using ACTs-related information. Indeed, molecules significantly influenced by ACs could potentially increase model training difficulties, leading to heightened uncertainty, particularly epistemic uncertainty. However, accurately quantifying molecular uncertainty is a complex and challenging task.



[Response Figure 16]: Correlation between the number of mined label-conditioned triplets (M′) and the validation RMSE across representative HSSMS datasets. Panels (a)–(d) show results for one dataset from each major target family: CLK4 (Kinase), 5-HT1A (GPCR), PIK3CA (Transferase), and GR (Nuclear Receptor). For each dataset, the cliff lower and cliff upper thresholds were scanned using the T-shape strategy, while the awareness factor α was kept fixed. The resulting number of mined ACTs M under each cliff-parameter combination is visualized alongside the corresponding validation RMSE obtained from three-fold cross-validation. The scatter plots on the right show the strong negative correlation between $\log_2(M)$ and the RMSE of the validation set, illustrating that models benefiting from a larger pool of informative triplets tend to achieve better predictive accuracy.

Response Table 7: Performance of MC-Dropout with AC-related features vs. vanilla MC-Dropout across benchmark datasets.

Dataset	MC-DropoutWith Feature	MC-Dropout	Improvement
CHEMBL1871_Ki	0.331834	0.331765	0.000069
CHEMBL218_EC50	0.165836	0.165836	0.000000
CHEMBL244_Ki	0.055880	0.048674	0.007206
CHEMBL236_Ki	0.052974	0.036591	0.016383
CHEMBL234_Ki	0.027610	0.057853	-0.030243
CHEMBL219_Ki	0.190045	0.075719	0.114326
CHEMBL238_Ki	0.084341	0.048021	0.036320
CHEMBL4203_Ki	0.109530	0.089190	0.020340
CHEMBL2047_EC50	0.117296	0.117295	0.000001
CHEMBL4616_EC50	0.037142	0.211543	-0.174401
CHEMBL2034_Ki	0.073341	0.062908	0.010433
CHEMBL262_Ki	0.174530	0.160674	0.013856
CHEMBL231_Ki	0.126748	0.152686	-0.025938
CHEMBL264_Ki	0.193582	0.031690	0.161892
CHEMBL2835_Ki	0.170952	0.084738	0.086214
CHEMBL2971_Ki	0.012431	0.050309	-0.037878
CHEMBL237_EC50	0.092533	0.131072	-0.038539
CHEMBL237_Ki	0.130490	0.046943	0.083548
CHEMBL4792_Ki	0.223961	0.052689	0.171272
CHEMBL239_EC50	0.196063	0.049726	0.146337
CHEMBL3979_EC50	0.113353	0.113218	0.000136
CHEMBL235_EC50	0.101740	0.058198	0.043543
CHEMBL4005_Ki	0.023411	0.042864	-0.019453
CHEMBL2147_Ki	0.048285	0.235041	-0.186756
CHEMBL214_Ki	0.270683	-0.007870	0.278553
CHEMBL228_Ki	0.104643	0.104643	0.000000
CHEMBL287_Ki	0.045151	0.086145	-0.040993
CHEMBL204_Ki	0.118009	0.086413	0.031597
CHEMBL1862_Ki	0.079195	0.083552	-0.004356
CHEMBL233_Ki	0.066053	0.038165	0.027888

To preliminarily investigate this idea, we conducted tests on the HSSMS benchmark using two straightforward uncertainty measures. We first applied standard Monte-Carlo (MC) Dropout with a dropout ratio of 0.1. For each test sample (x), we performed 50 stochastic forward passes, yielding predictions y_t'. The mean and variance of the predictions are computed as:

$$y^{avg}(x) = \frac{1}{T}\sum_1^T y_t',\ \sigma_{MC}^2(x) = \frac{1}{T}\sum_1^T \left(y_t' - y^{avg}(x)\right)^2$$

This variance $\sigma_{MC}^2(x)$ serves as a baseline uncertainty estimate. To incorporate AC-triplet information, we define two anchor-specific metrics.

(i) **Risk**, the number of valid AC triplets that can be formed with anchor x: $risk(x) = \left|\{m - [d(x,n) - d(x,p)] > 0, \{p,n\}\}\right|$, which reflects, in part, how strongly the anchor is influenced by activity cliffs.

(ii) **Anchor-density score**, the average distance from the anchor embedding to its (k) nearest neighbours in the training latent space:$dens(x) = \frac{1}{k}\sum_{r=1}^{k} d\big(z(x), z(r)\big)$.

We reserved one-fifth of the training data as a support set $\{(x_i, y_i)\}$ and computed the residual error as:

$$e_i' = |y'(x_i) - y_i| - \sigma_{MC}^2(x_i)$$

This residual error was regressed onto the two aforementioned features $risk(x_i)$ and $dens(x_i)$ using a non-negative linear model:

$$e_i' = w_1 \cdot risk(x_i) + w_2 \cdot dens(x_i)$$

Consequently, our final fused uncertainty estimation is defined as:

$$U(x) = \sigma_{MC}^2(x) + w_1 \cdot risk(x) + w_2 \cdot dens(x)$$

Finally, we evaluated the quality of the uncertainty estimates by computing Spearman's rank correlation between each method's predicted uncertainty and the observed RMSE. Although neither approach has yet shown a strong positive correlation, the variant that incorporates the AC triplet-based features $risk(x)$ and $dens(x)$shows a modest improvement over MC Dropout alone as shown in **Response Table 7**. After incorporating the additional features, the Spearman correlation between predicted uncertainty and absolute error improved on 19 of the 30 benchmark datasets, yielding only an average gain of 0.023 over the baseline. At present, this slight change—achieved merely by adding AC-related features to a simple uncertainty-estimation method—may not fully reflect its ability to enhance uncertainty estimation. We will continue to explore and evaluate more advanced strategies in our future work.

•Table S1: please indicate the number of activity cliffs, their percentage on the overall dataset. How did you split into train/test data?

RESPONSE: Thank you for the valuable suggestion. We have now included the percentage of activity cliffs (AC) for each dataset, as shown in **Response Table 3 / Supplementary Table S2**. Notably, eight out of the nine datasets in the LSSNS benchmark consist of more than 90 % AC molecules.

[Response Table 3 (new Supplementary Table S2)]: The 9 datasets of low sample size and narrow scaffolds (LSSNS) for molecular activity prediction. % Cliff was identified using the same approach as MoleculeACE. Since 5-fold cross-validation was employed, we report the proportion calculated over the entire dataset. Diversity is calculated as the average Tanimoto distance between all molecular pairs within the dataset, while Scaffold denotes the number of unique Murcko scaffolds.

Dataset	Target Name	Assay type	Target type	Institution	Size	% Cliff	Diversity	Num Scaffolds
USP7	Ubiquitin carboxyl-terminal hydrolase 7	IC50	Protease	Almac Discovery Ltd., Centre for Precision Therapeutics [3]	45	95.6	0.342	16
RIP2	Serine/threonine-protein kinase RIPK2	IC50	Kinase	Genomics Institute of the Novartis Research Foundation [4]	46	91.3	0.346	14
PKCI	Protein kinase C iota	IC50	Kinase	Agency for Science, Technology and Research (A*STAR) [5]	48	64.6	0.650	34
PHGDH	D-3-phosphoglycerate dehydrogenase	IC50	Other Enzyme	Boehringer Ingelheim RCV GmbH & Co. KG [6]	51	98.0	0.377	8
PLK1	Serine/threonine-protein kinase PLK1	IC50	Kinase	Hefei University of Technology [7]	73	98.6	0.223	18
IDO1	Indoleamine 2,3-dioxygenase	IC50	Other Enzyme	NewLink Genetics Corporation [8]	78	92.3	0.421	36
RXFP1	Relaxin receptor 1	AC50	GPCR	The Florida International University Board of Trustees [9]	117	91.5	0.389	42
BRAF	Serine/threonine-protein kinase B-raf	IC50	Kinase	The Development Center for Biotechnology (DCB) [10]	128	96.1	0.308	14
MGLUR2	Metabotropic glutamate receptor 2	IC50	GPCR	Eisai R&D Management Co., Ltd. [11]	244	99.6	0.128	9

Consequently, we employed a five-fold random cross-validation scheme to split each dataset into training and test sets, which effectively ensures that every subset contains a sufficient number of activity cliffs and enhances the robustness and reliability of our evaluation results.

•Table S3 – How do you assign AC here, especially if you have more non cliff mmps than total compounds? Can compounds be in a cliff and not in a cliff pair?

RESPONSE: Thank you for the question. Yes. A compound can appear in both cliff and non-cliff pairs. With N compounds, up to N(N-1)/2 pairs can be formed, so the number of MMPs can exceed the compound count.

For comparision, we follow the procedure used in the ACGCN to assign AC: MMPs are generated by the Hussain-Rea algorithm, which fragments each molecule, indexes the fragments, and keeps only pairs that differ at a single substructure site by no more than 20% of heavy atoms. Each pair is labeled according to its potency difference: pairs showing a 100-fold or greater gap (ΔpKi ≥ 2) are tagged as MMP-Cliffs, pairs with a 10-fold or smaller gap (ΔpKi ≤ 1) are tagged as "non-cliffs," and pairs falling in between these thresholds are discarded.

•Table S4 – all tables are structured differently. Please try to have the same content as far as possible in all of them, i.e. number and percentage of AC, splitting into train and test...

Response Table 4 (new Supplementary Table S4): The 10 ADMET datasets used for Δ-prediction and their structural statistics. % Cliff was identified using the same approach as MoleculeACE. Since 5-fold cross-validation was employed, we report the proportion calculated over the entire dataset. Diversity is calculated as the average Tanimoto distance between all molecular pairs within the dataset, while Scaffold denotes the number of unique Murcko scaffolds.

Property	Dataset	Units	Size	% Cliff	Diversity	Scaffolds
Absorption	Free Solvation	Experimental Hydration Free Energy in Water	642	92.5	0.933	63
	Caco2	Log(Papp)	910	30.8	0.741	488
	Aqueous Solubility	LogS	1128	80.2	0.920	269
Distribution	Fraction Unbound, Brain	Log(fu,brain)	253	15.4	0.724	200
	Volume of Distribution at Steady State	Log(Body/Blood Concentration in L/kg)	1130	20.1	0.730	778
Excretion	Renal Clearance	Log(CL_r)	636	5.0	0.740	447
Metabolism	Microsomal Clearance	Log(mL/min/kg cleared)	731	16.6	0.693	534
	Hepatic Clearance	Log(mL/min/kg cleared)	881	15.3	0.690	595
	Half-Life	Log(Half-Life in Hours)	1321	21.0	0.728	936
Toxicity	Hemolytic Toxicity	Log(HD_{50})	828	54.7	0.682	349

RESPONSE: Thank you for the suggestion. We have revised **Response Table 4 / Supplementary Table 4** to include the number and percentage of activity-cliff compounds, dataset diversity, and the number of unique scaffolds, so that its contents are consistent with the statistics reported for HSSMS and LSSNS. Because the ADMET benchmark is evaluated using 5-fold random cross-validation, we report overall counts and proportions rather than separate train/test splits.

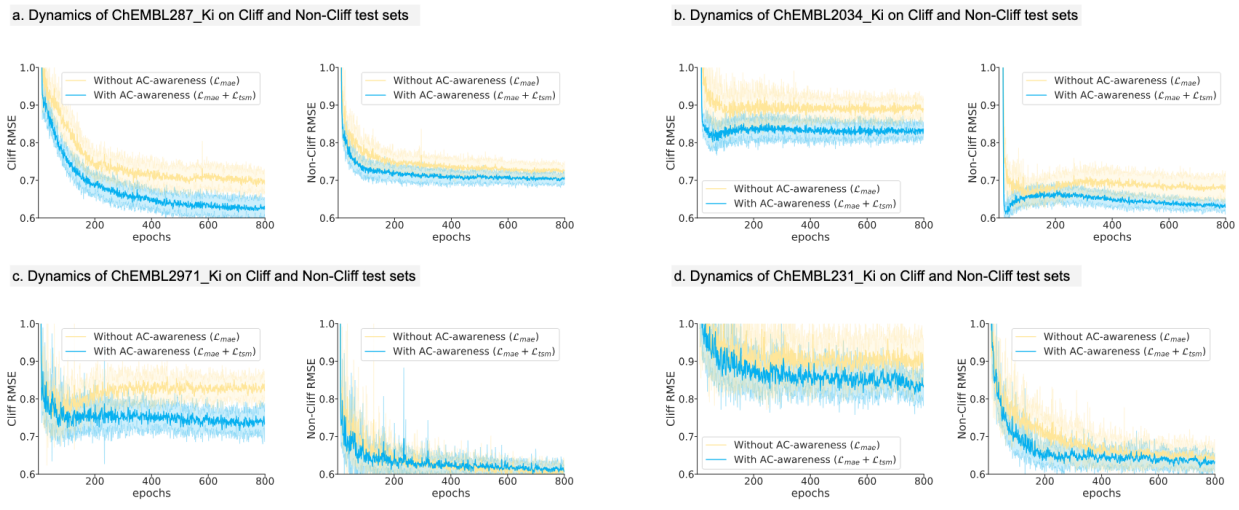
•L 137: Can you always generate a triplet for any anchor compound? How do you define an activity cliff based on what similarity measure?
o L. 141 – again: what is structural similarity?

RESPONSE: Thank you for the question. A triplet is generated for an anchor compound only when the mini-batch contains at least one positive and one negative molecule that meet both the structural-similarity and activity-difference criteria. So if no such molecules are present, the anchor does not contribute a triplet.

In our method, , leveraging the GNN's capacity to capture structural information, structural similarity is defined as the distance between compounds in the GNN-encoded latent space, with a soft $M = \left|y_j^A - y_j^N\right| - \left|y_j^A - y_j^P\right|$. We have also incorporated traditional string-based similarity metrics as gate, including (i) molecular-fingerprint Tanimoto similarity, (ii) Bemis–Murcko scaffold Tanimoto similarity, and (iii) SMILES edit similarity, as detailed in our **[Response to Reviewer 1, Major Concern 1].**

•L 154, Figure 3c: I would like to see the performance difference more differentiated: How does the model perform on additive data, how on cliff data points? Is the improvement really due to improved performance on cliff data?

RESPONSE: Thank you for the question. Similar to **Figure 3c**, we tracked test performance separately for cliff and non-cliff samples throughout training.



Response Figure 25 (new Supplementary Figure 17): Training dynamics comparing AC-aware training (blue) versus the baseline (yellow). (a–d) RMSE curves on activity-cliff and

non-cliff test sets. AC-aware training consistently improves performance on activity-cliff data compared to the baseline.

As shown in **Response Figure 25 / Supplementary Figure 17,** ACANet improves performance on both subsets, with broader and more sustained gains on the cliff samples.

•L 161: “external validation and test set” - How did you split the data into train/test/validation?

RESPONSE: In the HSSMS benchmark we employed the data-splitting protocol established in prior work. Specifically, for each molecule, we evaluated three tradtional similarity metrics—(1) Substructure similarity, (2) Scaffold similarity, and (3) SMILES similarity (see **[Response to Reviewer 1, Major Concern 1]**). If the dataset contained any other molecule with similarity ≥ 0.90 to the current molecule and the activity difference between the pair exceeded one order of magnitude, we labeled the molecule as a cliff mol; otherwise, it was labeled non-cliff.

We then applied spectral clustering to extended-connectivity fingerprints (ECFPs) to group structurally related molecules. Within each cluster we carried out stratified random sampling based on the cliff labels, allocating 80% of the data to the training set and 20% to the test set. During model development, 20% of the training set was further reserved as a validation set. This procedure guarantees that the training and test splits contain comparable proportions of cliff molecules and preserve similar structural diversity.

•Ll 167-169 “Performance improvements were observed across the board, with test RMSE reductions of 5.81%, 5.48%, 4.63%, 169 and 6.62%” – Do they improve on similar data points or are they improving in different chemical areas?

Response Table 8 (new Supplementary Table 14): Comparison of model performance with and without AC-awareness on 30 HSSMS datasets (scaffold-based split). The table reports RMSE under nested cross-validation on the training set with a scaffold-based split. MAE was used as the regression loss. Improvements are shown in the last column.

Dataset ID	Dataset Name (Target)	Without AC-awareness	With AC-awareness	Improvement
CHEMBL1871_Ki	AR	0.756 ± 0.005	**0.719 ± 0.005**	↓ 4.86%
CHEMBL218_EC50	CB1	0.753 ± 0.006	**0.737 ± 0.004**	↓ 2.15%
CHEMBL244_Ki	FX	0.879 ± 0.006	**0.839 ± 0.004**	↓ 4.62%
CHEMBL236_Ki	DOR	0.756 ± 0.005	**0.749 ± 0.006**	↓ 0.96%
CHEMBL234_Ki	D3R	0.741 ± 0.004	**0.707 ± 0.006**	↓ 4.56%
CHEMBL219_Ki	D4R	0.886 ± 0.005	**0.842 ± 0.006**	↓ 4.95%
CHEMBL238_Ki	DAT	0.838 ± 0.004	**0.808 ± 0.005**	↓ 3.55%
CHEMBL4203_Ki	CLK4	0.831 ± 0.007	**0.797 ± 0.004**	↓ 4.08%
CHEMBL2047_EC50	FXR	0.774 ± 0.005	**0.750 ± 0.005**	↓ 3.14%
CHEMBL4616_EC50	GHSR	0.741 ± 0.005	**0.722 ± 0.005**	↓ 2.56%
CHEMBL2034_Ki	GR	0.770 ± 0.004	**0.732 ± 0.004**	↓ 4.96%
CHEMBL262_Ki	GSK3	0.807 ± 0.007	**0.792 ± 0.005**	↓ 1.83%
CHEMBL231_Ki	HRH1	0.630 ± 0.005	**0.606 ± 0.002**	↓ 3.81%
CHEMBL264_Ki	HRH3	0.694 ± 0.004	**0.665 ± 0.005**	↓ 4.24%
CHEMBL2835_Ki	JAK1	0.374 ± 0.002	**0.361 ± 0.002**	↓ 3.46%
CHEMBL2971_Ki	JAK2	0.656 ± 0.004	**0.636 ± 0.004**	↓ 3.07%
CHEMBL237_EC50	KOR (a)	0.966 ± 0.006	**0.922 ± 0.005**	↓ 4.51%
CHEMBL237_Ki	KOR (i)	0.829 ± 0.006	**0.775 ± 0.005**	↓ 6.49%
CHEMBL4792_Ki	OX2R	0.693 ± 0.005	**0.672 ± 0.004**	↓ 3.04%
CHEMBL239_EC50	PPAR(α)	0.825 ± 0.005	**0.781 ± 0.004**	↓ 5.31%
CHEMBL3979_EC50	PPAR(γ)	0.793 ± 0.006	**0.746 ± 0.006**	↓ 5.90%
CHEMBL235_EC50	PPAR(δ)	0.830 ± 0.005	**0.802 ± 0.006**	↓ 3.36%
CHEMBL4005_Ki	PIK3CA	0.646 ± 0.004	**0.611 ± 0.005**	↓ 5.53%
CHEMBL2147_Ki	PIM1	0.602 ± 0.004	**0.589 ± 0.005**	↓ 2.11%
CHEMBL214_Ki	5-HT1A	0.730 ± 0.005	**0.702 ± 0.005**	↓ 3.90%
CHEMBL228_Ki	SERT	1.076 ± 0.008	**0.980 ± 0.007**	↓ 8.93%
CHEMBL287_Ki	SOR	0.723 ± 0.004	**0.776 ± 0.003**	↑ 7.35%
CHEMBL204_Ki	Thrombin	0.778 ± 0.005	**0.744 ± 0.004**	↓ 4.32%
CHEMBL1862_Ki	ABL1	0.943 ± 0.007	**0.917 ± 0.004**	↓ 2.69%
CHEMBL233_Ki	MOR	0.759 ± 0.004	**0.710 ± 0.004**	↓ 6.41%
Average	-	0.769 ± 0.124	**0.740 ± 0.114**	↓ 3.81%

RESPONSE: In the initial data split, we applied spectral clustering over the ECFPs of all molecules to group structurally similar compounds into clusters, and then sampled within each cluster to ensure a consistent structural distribution between training and test sets. Therefore, the reported improvements can be regarded as gains on structurally similar data.

To further examine whether ACA method also improves performance on structurally dissimilar compounds, we re-split the HSSMS datasets using a scaffold-based partitioning strategy, where 20% of the scaffolds were randomly selected and only molecules with these scaffolds were included in the test set, which guarantees dissimilarity between the training and test sets. As shown in **Response Table 8 / Supplementary Table 14**, ACA method still achieved better performance than the method without ACA. These results demonstrate that ACANet provides improvements on both structurally similar and dissimilar data points.

•Ll 188-189 Figure S2: The color scheme of the RMSD scale seems quite arbitrary and different for every plot. You could select 0.3 as a good value and ~0.5 as a bad one. Thus

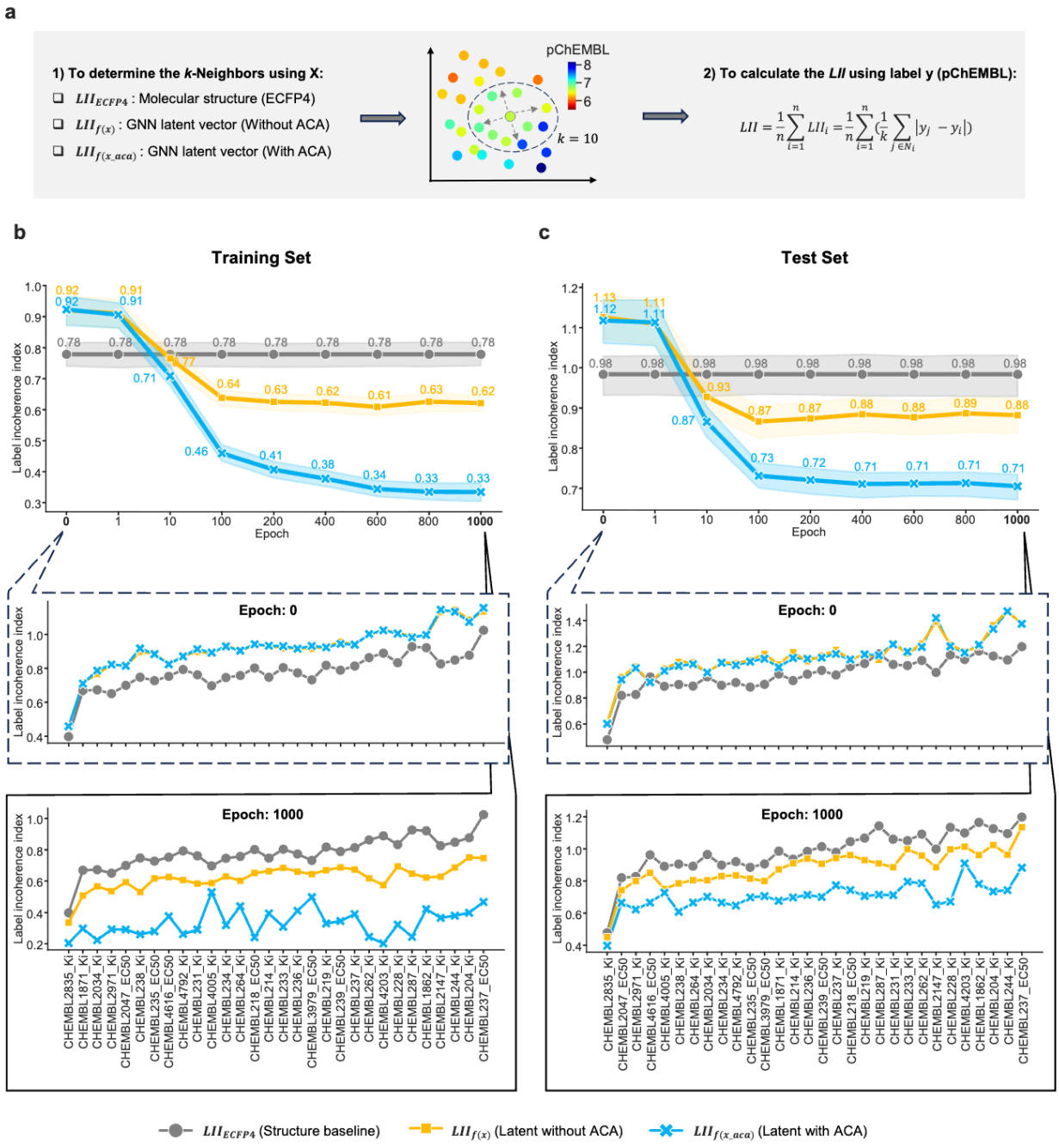
leading to a better comparison of the influence of the M parameter, which seems very dataset dependent.

RESPONSE: We thank the reviewer for raising this point and apologise for any confusion. We respectfully note that the colour scheme was not chosen arbitrarily. Because the difficulty and intrinsic characteristics of the datasets differ markedly, their RMSE ranges vary substantially—for example, JAK1 spans 0.39–0.75, whereas CLK4 spans 0.88–1.06. Consequently, the same absolute RMSE does not signify equivalent predictive performance across datasets.

We therefore adaptively rescale the colour map for each dataset to the minimum and maximum RMSE values observed across all grid-search settings. Specifically, we apply BoundaryNorm-based normalisation so that each heat-map is rendered over its own dynamic range. When we experimented with a fixed, global colour scale, many datasets collapsed toward one end of the spectrum, obscuring meaningful differences between parameter settings and greatly reducing interpretability.

•Ll 189-191: "This study highlights the importance of selecting appropriate cliff cut-off parameters to efficiently mine ACTs and enhance cliff awareness and model performance." – What are your learning from this? Any recommendations on how to best set these parameters?

RESPONSE: Thank you for this important question. In real-world scenarios, structure-activity landscapes differ markedly across targets—some exhibit relatively smooth, shallow activity gradients (e.g., **ChEMBL2835_EC50** in **Response Figure 13 / Extended Data Figure 1** and **Response Figure 5 / Supplementary Figure 13**), whereas others contain sharp activity cliffs (e.g., **ChEMBL204_Ki** in **Response Figure 13 / Extended Data Figure 1** and **Response Figure 5 / Supplementary Figure 13**). Consequently, a single, fixed AC threshold is unlikely to serve as a universal standard or be optimal for QSAR prediction models.



[Response Figure 13 (new Extended Data Figure 1)]: Temporal evolution of the label-incoherence index (LII) in the HSSMS benchmark. Lower LII values indicate greater local label coherence. (a) LII calculation workflow – Step 1: for each compound, identify the $k = 10$ nearest neighbours using one of three representations: ECFP4 fingerprints, GNN latent vectors without ACA, or GNN latent vectors with ACA. Step 2: compute LII from the pChEMBL labels of the query compound and its neighbours. **(b) Training set** – Epoch-wise LII dynamics for the three representations (grey = ECFP4, yellow = latent without ACA, blue = latent with ACA), accompanied by per-dataset LII values at epoch 0 and epoch 100. **(c) Test set** – Corresponding LII dynamics for held-out compounds, with per-dataset snapshots at epoch 0

and epoch 100. Across all datasets, ACA persistently drives the latent-space LII below both the ECFP baseline and the non-ACA model.



[Response Figure 5 (new Supplementary Figure 13)]: Three-dimensional activity-landscape surfaces for 30 single-scaffold HSSMS datasets, generated by projecting 1024-bit Morgan fingerprints onto the first two PCA components and smoothing the activities with an RBF interpolator. Colour encodes activity: blue/green = lower, yellow/red = higher.

Our experiments (see **Figure 5** and **Supplementary Figure 2**) demonstrate that the optimal "cliff-upper" and "cliff-lower" values are target-dependent; there is no single cliff cutoff value that achieves the best performance across all datasets. We hypothesize that appropriate cutoffs should match the steepness of the target's landscape, thereby yielding a sufficient number of effective AC triplets, as a larger ACT set is strongly correlated with lower RMSE (see **Figure 5**).

•ll 203-204: “Our results suggest that selecting appropriate cliff parameters and alpha can enable efficient cliff awareness” – Any indications or learning from this?

RESPONSE: We conducted a hyperparameter analysis to reveal the interplay between the AC cut-off thresholds (cliff-upper(cu) and cliff-lower(cl)) and the parameter α. As shown in **Figure 5, Supplementary Figure 2** and **Response Figure 16**, inappropriate cliff threshold selection drastically reduces the number of ACTs to nearly zero, resulting in very poor model performance. However, simply maximizing the ACT count by adopting overly lenient thresholds is also inadvisable, as it floods the training set with numerous trivial triplets, hindering the model’s ability to learn decisive structure-activity relationships embedded within ACs. Empirically, we find that maintaining ACT yield within approximately 5–15% of all anchor pairs achieves the most reliable performance.

Beyond threshold selection, the choice of α is also critical. **Figures 5c, Supplementary Figure 5** illustrate that setting α=0 (i.e., disabling the ACA loss) consistently yields inferior results. Additionally, optimal α values differ among datasets (**see Supplementary Figure 7**).

•Ll 221-222_ “Superior overall performance” - Superior sounds as if they are much better, but the performance of both looks rather on par. Please comment and/or reword to a more realistic wording.

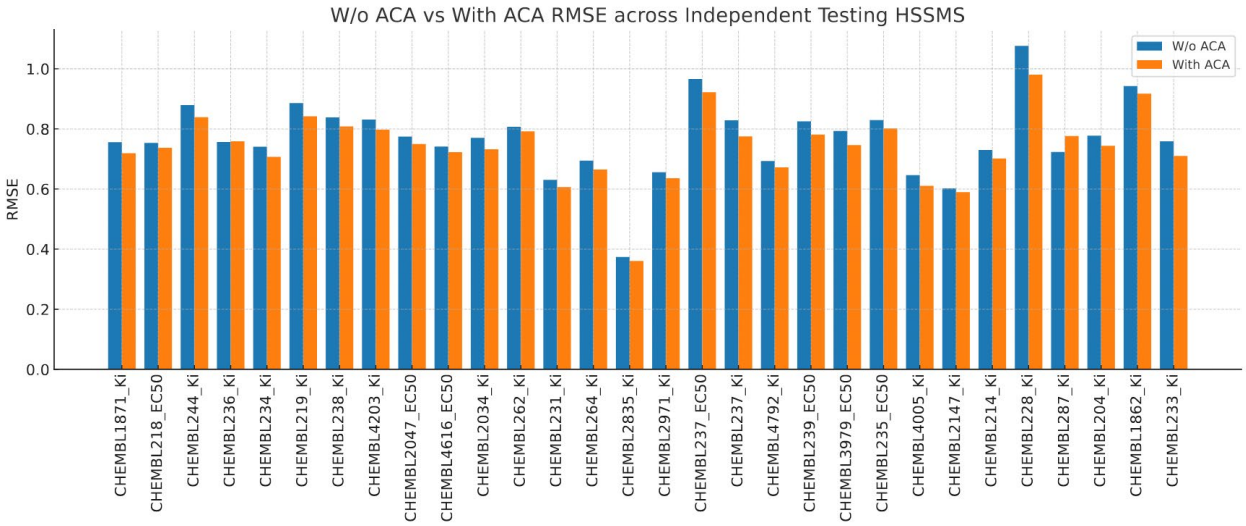
RESPONSE: Thank you for pointing this out. We agree that the term “superior” may overstate the magnitude of the improvement in some cases. Our results show modest but consistent gains over strong traditional baselines such as SVM and RF, and substantially larger gains over earlier deep-learning models. In the revision we rephrased the sentence to **“ACANet achieves the best overall performance, with consistent improvements over strong traditional baselines and much larger gains relative to previous deep-learning methods.”**

•Ll 225 – full paragraph on “Using AC-awareness enhances activity-coherent latent molecular representations.” - I am not jet convinced that this method really learns about activity cliffs. Currently it seems more like the method is learning from what it already knows are activity cliffs. In order to see, what it is truly learning, it would need a different test dataset, which contains activity cliffs, where neither positive nor negative example are in the training data. I have not seen any such test data set in the manuscript.

Response Table 9 (new Supplementary Table 12): Summary statistics of the newly curated HSSMS dataset generated via scaffold-based splitting and filtering for independent testing

Dataset	Train Size	Test Size	% Cliff	Test Diversity
CHEMBL1871_Ki	566	59	72.9%	0.595
CHEMBL218_EC50	810	94	63.8%	0.618
CHEMBL244_Ki	2459	323	78.3%	0.644
CHEMBL236_Ki	2048	248	80.2%	0.516
CHEMBL234_Ki	2906	326	69.6%	0.692
CHEMBL219_Ki	1437	251	79.3%	0.707
CHEMBL238_Ki	773	230	85.7%	0.677
CHEMBL4203_Ki	571	149	57.7%	0.677
CHEMBL202_Ki	658	191	76.0%	0.697
CHEMBL204_EC50	1744	406	77.3%	0.664
CHEMBL1914_Ki	1887	509	77.6%	0.666
CHEMBL199_Ki	1806	605	77.2%	0.640
CHEMBL4857_Ki	955	284	71.5%	0.699
CHEMBL240_Ki	961	255	75.3%	0.709
CHEMBL241_Ki	1324	387	70.8%	0.679
CHEMBL217_Ki	1104	288	78.1%	0.634
CHEMBL216_Ki	746	196	78.6%	0.705
CHEMBL227_Ki	1651	462	78.4%	0.699
CHEMBL210_Ki	1209	331	74.7%	0.703
CHEMBL251_Ki	1205	258	77.9%	0.680
CHEMBL264_Ki	1645	470	76.2%	0.681
CHEMBL224_Ki	2084	514	76.5%	0.677
CHEMBL4396_Ki	872	250	70.0%	0.706
CHEMBL232_Ki	2767	591	75.1%	0.630
CHEMBL4943_Ki	587	174	81.6%	0.650
CHEMBL1865_Ki	1331	364	74.7%	0.671
CHEMBL2188_Ki	944	308	82.5%	0.640
CHEMBL204_Ki	2255	499	75.2%	0.621
CHEMBL1862_Ki	576	218	83.9%	0.531
CHEMBL233_Ki	2593	549	79.1%	0.591

RESPONSE: We thank the reviewer for this valuable question. To rigorously evaluate ACANet's ability to detect previously unseen activity cliffs, we devised a scaffold-held-out split. First, all molecules were clustered by their Bemis–Murcko scaffolds, and 20 % of the scaffolds were randomly reserved for testing; every compound containing these scaffolds was excluded from training. We then removed any test molecule whose fingerprint Tanimoto similarity, scaffold Tanimoto similarity, or SMILES-edit similarity to any training compound exceeded 0.90, thereby strictly ensuring that no test molecule could form an activity cliff with the training set.



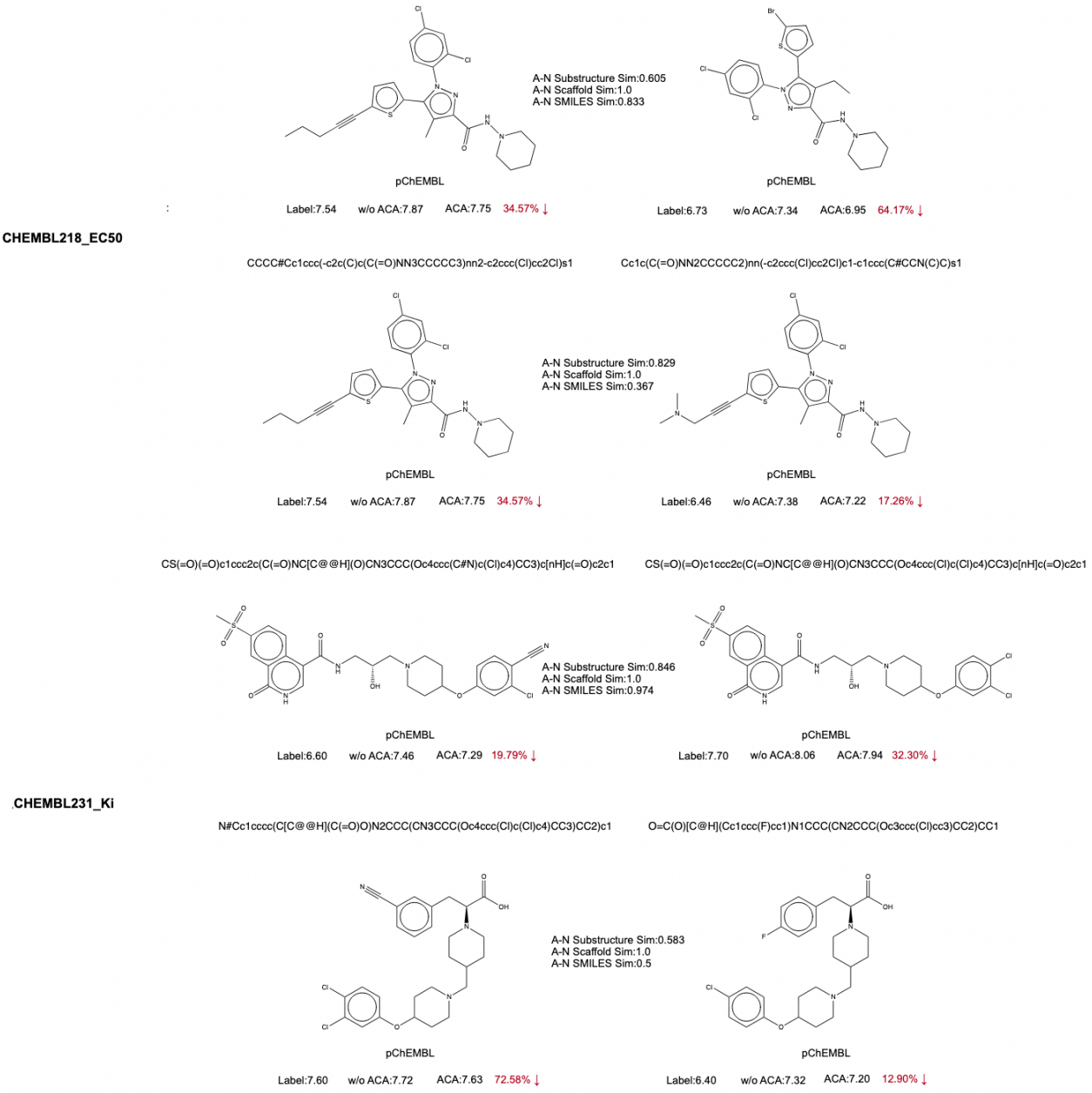
Response Figure 26 (new Supplementary Figure 20): Test-set RMSE comparison between models without ACA (W/o ACA) and with ACA (With ACA) across HSSMS datasets that possess independent test sets.

This protocol yielded a test partition in which, on average, 76.8 % of molecules were annotated as cliffs (**Response Table 9 / Supplementary Table 12**). When evaluated on this stringent split, models trained with the ACA loss achieved lower RMSE on 28 of 30 targets relative to their non-ACA counterparts (**Response Figure 26 / Supplementary Figure 20**), demonstrating that ACANet can generalise to—and capitalise on—cliff information even when the relevant cliffs are encountered only at test time.

•l 273 “RMSEcliff” – please explain what is meant with this, how is it calculated.

RESPONSE: In our study, RMSE*cliff* is the root-mean-square error calculated exclusively for those test compounds labelled as cliff molecules. In HSSMS , a compound is designated a cliff if the dataset contains at least one other molecule with ≥ 0.90 similarity—according to any of three traditional measures (fingerprint Tanimoto similarity, scaffold Tanimoto similarity, or SMILES-edit similarity)—and a potency difference of at least one order of magnitude (a ten-fold change) (see our **[Response to Line 161]**), following Van Tilborg . We compute RMSE*cliff* solely on this subset to gauge the model’s performance in these challenging regions of chemical space.

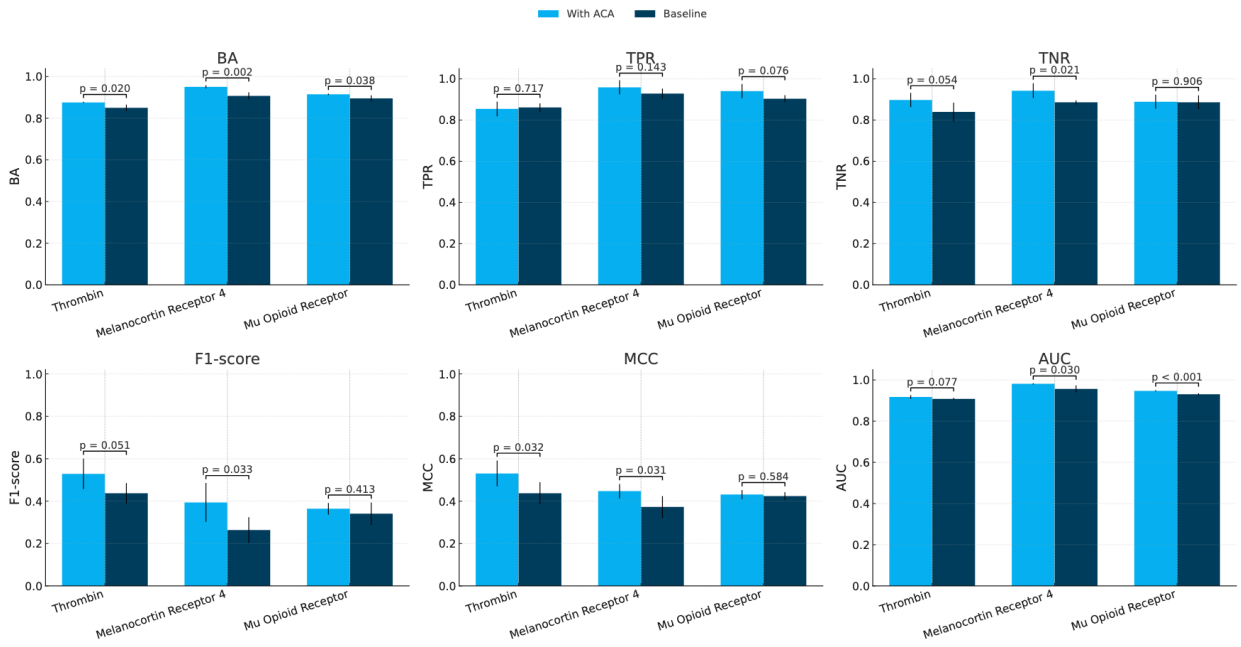
•Ll 286-287: “ACANet outperformed the SOTA SVM for 23 and 21 of the 30 datasets in RMSEall and RMSEcliff” - Are there any ACs that are only in the test set for which you could show performance improvement?



Response Figure 27: ACA-aware gains on test-only activity cliffs. For two representative targets (CHEMBL218_EC50 and CHEMBL231_Ki), each panel shows test-set AC pairs with substructure/scaffold/SMILES similarities and model scores. Compared with the baseline, the ACA-aware model yields markedly lower errors on these unseen ACs (red percentages), highlighting improved generalization to cliff cases that do not appear in training.

RESPONSE: Yes. **Response Figure 27** highlights several representative AC examples that occur exclusively in the test set. For each case we quantify the benefit of ACANet by the relative error reduction $\frac{|y'_{base}-y|-|y'_{ACA}-y||}{|y'_{base}-y|}$, where y'_{ACA} denotes the ACANet prediction, y'_{base} denotes the prediction of the non-AC-aware model and y denotes the label.

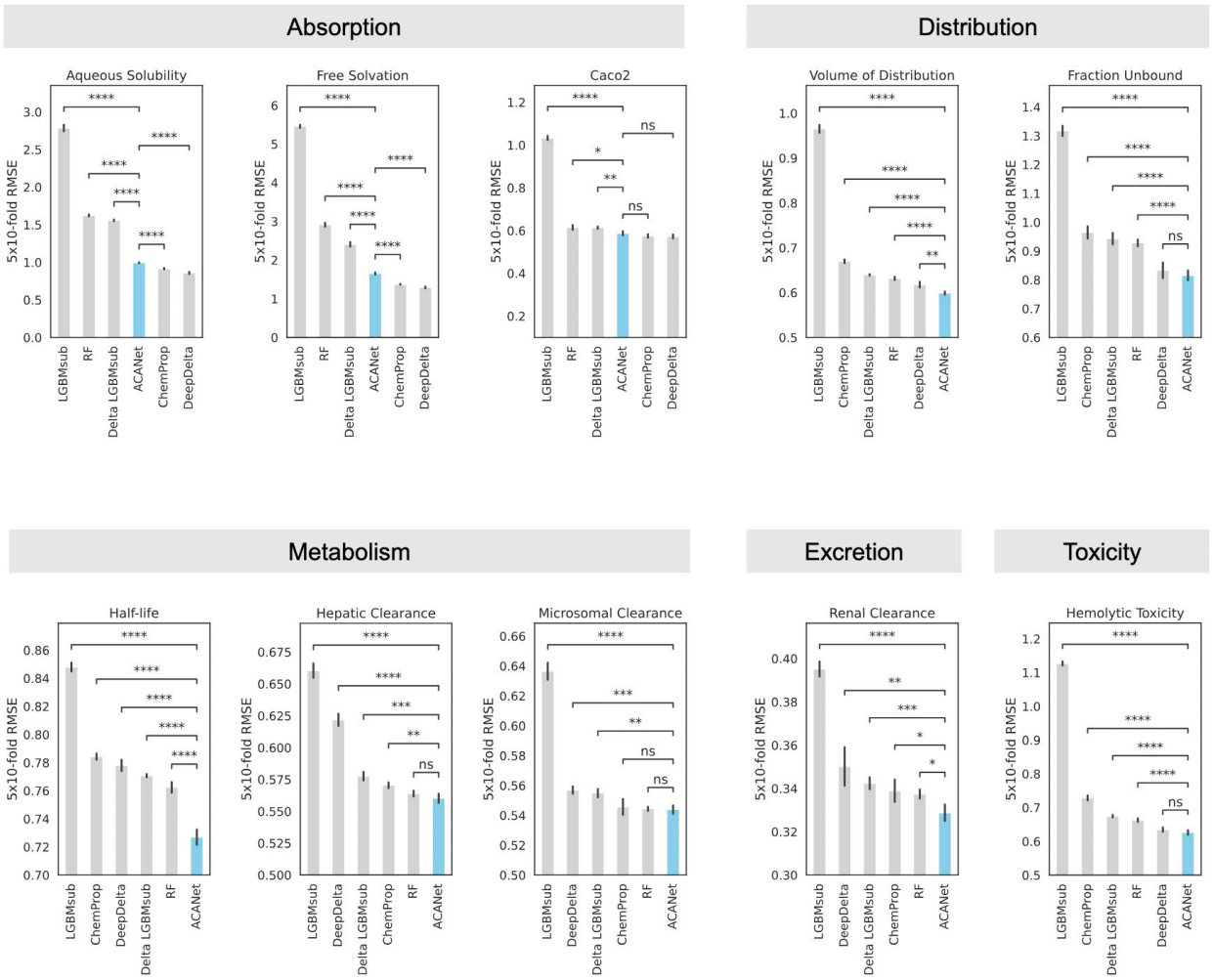
•L 313, Table S9 - Can you please comment on the significance of the differences in the presented results?



Response Figure 28 (new Supplementary Figure 16): Performance of MMP models with and without AC-awareness. Across three datasets, per-metric bars (BA, TPR, TNR, F1-score, MCC, AUC) show mean±s.d. over 5×5 CV (five random seeds, five folds). p-values (Welch's t-test) are annotated above paired bars; AC-aware training generally yields lower error and higher classification scores.

RESPONSE: In **Table S9**, the GNN-based ACGCN model with AC-awareness consistently outperforms all baselines. We performed five-fold cross-validation using five random seeds and applied a t-test at α = 0.05. As shown in **Response Figure 28 / Supplementary Figure 16**, most of the p-value bars are beneath the 0.05 threshold. The tests confirm significance for Thrombin (BA, MCC, AUC), for Melanocortin-4 Receptor (BA, TNR, F1, MCC, AUC), and for Mu-Opioid Receptor (BA, TPR, F1, MCC, AUC).

•Ll 330-331: "Moreover, ACANET achieved the lowest cross-validation RMSE across all other seven DMET datasets" - Out of which some are also on par with other methods. It would be interesting to get significance values here and otherwise reword it to better reflect the actual results.



Response Figure 28 (new Supplementary Figure 8): RMSE performance AC-awareness infused ACANet on ADMET delta prediction. The 10 ADMET delta prediction datasets are obtained from the DeepDelta paper 1. For Random Forest (RF) and Light Gradient Boosting Machine(LGB) models, ECFP fingerprints are used. ChemProp, DeepDelta, and ACANet use graph-based features. The RF, LGBsub,ChemProp, and ACANet models process a single molecule to predict absolute property values, which are then subtracted to calculate property differences between two molecules. In contrast, the Delta LGBsub and DeepDelta models are trained to predict the relative differences between two molecules. The model performance was evaluated by the RMSE of 10-fold cross-validation, repeated 5 times using different seeds. The error bars represent the standard deviation across the 5 repetitions.Bars show the mean ± SD over five cross-validation repeats; asterisks indicate paired t-test significance versus ACANet (** p <0.0001, ***p <0.001, ** p<0.01, * p <0.05, ns = not significant).**

RESPONSE: Thank you for your suggestion. We have added t-test p-values (see **Response Figure 28 / Supplementary Figure 8**) to indicate whether RMSE differences are statistically significant (**** p < 0.0001, *** p < 0.001, ** p < 0.01, * p < 0.05, ns = not significant). The sentence now reads:

“Moreover, ACANet achieved the lowest cross-validation or statistically comparable RMSE across all other seven DMET datasets—two distribution datasets, one excretion dataset, three metabolism datasets, and one toxicity dataset—as shown in Figure S27.”

Minor

l 113: “difficultThe” – missing space and point. L 344: ref 21 duplicated – please check references for correctness

RESPONSE: Thank you for spotting these issues. We have inserted the missing space in “difficultThe” and removed the duplicate entry for reference 21; all references have been rechecked for consistency in the revised manuscript.

Reviewer #3 (Remarks on code availability):

I partially tested the code. I could install the package without further trouble - thanks for that! There is no code example with a given dataset to run ACANet immediately. This should be included to give people an easy start.

RESPONSE: Thank you for trying out the code and for the helpful feedback. We have provided example code in the form of Jupyter notebooks under the ACANet experiment directory. In addition, we have added a lightweight quick-start notebook, quick_start.ipynb, which includes one of the preprocessed datasets and allows users to easily run ACANet right away.

RESPONSE TO REVIEWER COMMENTS

Overall response to the reviewers:

We sincerely thank the editor and reviewers for their careful evaluation of our revised manuscript. We are grateful to Reviewer #1 for recommending acceptance and for the positive comments on the clarity of the revised manuscript, the expanded Supplementary Information, the detailed rebuttal discussion, and the code availability. We also thank Reviewer #3 for the constructive final suggestions and for acknowledging the comprehensive revision and the preliminary uncertainty-estimation analysis.

In response to the remaining minor comments, we performed a focused final check of the Main text and Supplementary Information. The major final revisions are summarized below.

- We harmonized first-use expansions of key abbreviations, including Activity-Cliff Awareness Network (ACANet), Principal Neighborhood Aggregation (PNA), Graph Convolutional Network (GCN), Graph Attention Network (GAT), and Graph Isomorphism Network (GIN).
- We standardized phrasing and hyphenation, including consistent use of "AC-aware" and "Principal Neighborhood Aggregation", and unified the typography of $RMSE_{all}$ and $RMSE_{cliff}$ throughout the manuscript.
- We corrected the remaining typographical and spacing issues, including the missing space after "space. By" and the spacing in "network (GIN)".
- We added a brief note to the relevant Supplementary caption explaining why heatmap color scales are dataset-specific, so that readers can interpret within-dataset color contrasts more clearly.

Response to Reviewer #1:

Reviewer #1 (Remarks to the Author): *The reviewer appreciates the tremendous efforts that the authors have put to address all the major and minor concerns. The revised manuscript is greatly improved with much more clarity. The Supplementary Information also adds detailed explanations to the experiments. Furthermore, the rebuttal file provides an impressive, in-depth discussion on addressing activity cliffs in molecular property prediction, which can be highly beneficial for future studies. Given these considerations, my recommendation would be Accept. Kudos to the authors!*

RESPONSE: We sincerely thank you for your encouraging comments and for recommending acceptance. We are very pleased that you found the revised manuscript clearer and that the expanded Supplementary Information helped explain the experiments in more detail. We also appreciate your positive feedback on our discussion of activity cliffs in molecular property prediction, and we hope that these additional analyses will be useful for future studies in this area.

Reviewer #1 (Remarks on code availability): *The authors have nicely documented their project in github. I can install clsar and run ACANet from my end.*

RESPONSE: We sincerely thank you for checking the code and for confirming that the clsar package can be installed and that ACANet can be run successfully on your end. We are glad that the documentation was helpful. We will continue to maintain the GitHub repository and package release to ensure that the code remains consistent with the final version of the manuscript.

Response to Reviewer #3:

Reviewer #3 (Remarks to the Author): *Thank you for the comprehensive and well-executed revision and response. Interesting to see your preliminary exploration of uncertainty estimation - thanks for reporting it!*

RESPONSE: We sincerely thank you for your positive assessment of our revision and response. We are glad that you found the revision comprehensive and well executed, and we appreciate your encouraging comment on our preliminary exploration of uncertainty estimation. We have carefully addressed the remaining minor points, as detailed below.

Minor comments

Comment 1: *Consistent first-use expansions (e.g., Activity-Cliff Awareness Network (ACANet); Principal Neighborhood Aggregation (PNA); GCN, GAT, GIN) in the Main text.*

RESPONSE: We thank you for pointing this out. We have carefully checked the first appearances of key terms in the Main text and revised the first-use expansions to ensure consistent definitions, including but not limited to Activity-Cliff Awareness Network (ACANet), Principal Neighborhood Aggregation (PNA), Graph Convolutional Network (GCN), Graph Attention Network (GAT), and Graph Isomorphism Network (GIN). Subsequent occurrences now use the corresponding abbreviations consistently.

Comment 2: *Standardize phrasing and hyphenation (e.g., AC-aware; principal neighborhood aggregation vs principal neighbourhood aggregation) and $RMSE_{all}$/$RMSE_{cliff}$ typography.*

RESPONSE: Thank you for pointing this out. We standardized the manuscript to use "AC-aware" consistently and unified the spelling as "Principal Neighborhood Aggregation" throughout. We also revised the typography of $RMSE_{all}$ and $RMSE_{cliff}$ to ensure consistent formatting across the Main text, figure legends, and Supplementary Information.

Comment 3: *I found a few typos/spacing (e.g., missing space after "space. By"; "network (GIN)").*

RESPONSE: Thank you for pointing this out. We have corrected these typographical and spacing issues, including the missing space after "space. By" and the spacing in "network (GIN)". We also performed an additional proofreading pass to correct related minor spacing and typographical inconsistencies.

Comment 4: *Thanks for the explanation with the heatmap coloring. I would recommend adding a brief note in the Supplementary caption why heatmap color scales are dataset-specific to aid reader interpretation.*

RESPONSE: We agree that this clarification will help readers interpret the heatmaps more accurately. We have added a brief note to the relevant Supplementary caption, e.g., Supplementary Figure 2, explaining that the heatmap color scales are set separately for each dataset to preserve within-dataset contrast and avoid visually compressing datasets with different absolute RMSE ranges under a single global scale.

Reviewer #3 (Remarks on code availability): *I did not have time to retry it, but I looked into the changes the authors mentioned - thanks!*

RESPONSE: We thank the reviewer for looking into the code changes. We appreciate the reviewer's time and have kept the updated documentation and implementation available in the project repository to facilitate future installation and reproduction.