

Supplementary Information

Activity-Cliff Awareness Enables Robust Graph Learning for Molecular Property Prediction

Contents

Supplementary Notes	S3
Supplementary Note 1: AC awareness with additional structural gate	S3
Supplementary Note 2: Assessing ACA against classical quantitative SAR methods	S4
Supplementary Note 3: Stringent split strategies for validating the generalization of ACA	S4
Supplementary Figures	S5
Supplementary Figure 1: Chemical space distribution and mined triplets in LSSNS datasets	S5
Supplementary Figure 2: Mined triplets and validation RMSE in LSSNS datasets	S6
Supplementary Figure 3: Hyperparameter tuning on the HSSMS D4R dataset	S7
Supplementary Figure 4: Hyperparameter tuning on the ADMET Caco2 dataset	S7
Supplementary Figure 5: Effect of the ACA factor on HSSMS performance	S8
Supplementary Figure 6: Benchmark comparison on HSSMS datasets	S9
Supplementary Figure 7: ACA factor effects on easy and hard HSSMS cases	S10
Supplementary Figure 8: ADMET delta prediction performance	S11
Supplementary Figure 9: Additional examples of ACANet training dynamics	S12
Supplementary Figure 10: Activity-cliff triplet mining with a structure gate	S13
Supplementary Figure 11: Comparison of ACANet variants with and without a structure gate	S14
Supplementary Figure 12: Effects of structural-similarity gates on HV-ACTs and RMSE	S15
Supplementary Figure 13: Activity-landscape surfaces across HSSMS datasets	S16
Supplementary Figure 14: Training dynamics on cliff and non-cliff test sets	S17
Supplementary Figure 15: Backbone-agnostic ACA gains on ADMET datasets	S18
Supplementary Figure 16: MMP model performance with and without ACA	S19
Supplementary Figure 17: Comparison between PLS and ACANet on HSSMS	S19
Supplementary Figure 18: Comparison among PLS, Free-Wilson, and ACANet on single-scaffold HSSMS	S20
Supplementary Figure 19: Backbone-agnostic gains from AC-aware training	S20
Supplementary Figure 20: Independent-test performance on HSSMS datasets	S21
Supplementary Figure 21: Activity-cliff cases missed by traditional rules	S21

Supplementary Figure 22: Latent organisation of the JAK1 training set	S22
Supplementary Figure 23: Latent organisation of the JAK1 test set	S23
Supplementary Figure 24: Backbone-agnostic ACA gains on HSSMS datasets	S24

Supplementary Tables	S25
-----------------------------	------------

Supplementary Table 1: HSSMS dataset statistics	S25
Supplementary Table 2: LSSNS dataset statistics	S26
Supplementary Table 3: MMP activity-cliff classification datasets	S27
Supplementary Table 4: ADMET delta-prediction datasets	S27
Supplementary Table 5: ACANet graph featurization	S27
Supplementary Table 6: GNN backbone performance on PPAR δ	S28
Supplementary Table 7: LSSNS performance with MAE and MSE losses	S29
Supplementary Table 8: HSSMS performance with and without ACA	S30
Supplementary Table 9: MMP classification performance	S31
Supplementary Table 10: Model hyperparameter search space	S32
Supplementary Table 11: Single-scaffold HSSMS dataset statistics	S32
Supplementary Table 12: Independent-test HSSMS scaffold statistics	S33
Supplementary Table 13: Published activity-cliff definitions	S34
Supplementary Table 14: HSSMS scaffold-split performance with and without ACA	S35

Supplementary Notes

Supplementary Note 1: AC awareness with additional structural gate

In the AC mining stage, we provide an option to decide whether to use an additional traditional structural gate as the structural constraint. As shown in [Supplementary Figure 10](#), when the traditional structural gate is enabled, a candidate pair (x_a, x_j) is regarded as structurally similar must satisfy an additional structural requirement, namely that its structural similarity under any of the following traditional measures:

1. ECFP fingerprint Tanimoto ≥ 0.9 ,
2. Bemis–Murcko scaffold Tanimoto ≥ 0.9 ,
3. SMILES edit similarity ≥ 0.9 .

This method constrains all ACA-mined triplets to satisfy this condition. When the traditional gate is disabled, structural similarity between molecule pairs is determined solely by their distance in the GNN latent space. We provide a pseudocode description in [Supplementary Algorithm 1](#) to detail this procedure.

Supplementary Algorithm 1 Online mining of HV-ACTs with additional structural gate

Require: Batch $B = \{(x_i, y_i)\}$; encoder $f(\cdot; \theta)$; distance $d(\cdot, \cdot)$; cut-offs $cl < cu$

Require: Mode GATE $\in \{\text{NONE}, \text{TRADITIONAL}\}$

$\triangleright \text{NONE} \Rightarrow$ latent-only learning

Ensure: HV-ACT set M and total loss L_{total}

```
1:  $Z \leftarrow [f(x_i; \theta)]$   $\triangleright$  encode molecules to latent space
2:  $M \leftarrow \emptyset, S \leftarrow 0, C \leftarrow 0$   $\triangleright S$  sum of losses,  $C$  number of mined triplets with  $\ell > 0$ 
3: for all anchors  $a$  in  $B$  do
4:    $P_a \leftarrow \{j \neq a \mid |y_j - y_a| \leq cl\}$   $\triangleright$  label-conditioned positives
5:    $N_a \leftarrow \{j \neq a \mid |y_j - y_a| \geq cu\}$   $\triangleright$  label-conditioned negatives
6:   if GATE = TRADITIONAL then
7:      $P_a \leftarrow \{j \in P_a \mid \text{TRADITIONALSIMPASS}(x_a, x_j)\}$ 
8:      $N_a \leftarrow \{j \in N_a \mid \text{TRADITIONALSIMPASS}(x_a, x_j)\}$ 
9:   for all  $p \in P_a$  do
10:    for all  $n \in N_a$  do
11:       $m_{\text{APN}} \leftarrow |y_n - y_a| - |y_p - y_a|$   $\triangleright$  dynamic margin from activities
12:       $\ell_{\text{APN}} \leftarrow d(z_a, z_p) - d(z_a, z_n) + m_{\text{APN}}$ 
13:      if  $\ell_{\text{APN}} > 0$  then
14:         $M \leftarrow M \cup \{(a, p, n, \ell_{\text{APN}})\}; S \leftarrow S + \ell_{\text{APN}}; C \leftarrow C + 1$ 
15:      else
16:        continue  $\triangleright$  not high-value; contributes zero loss
17:  $L_{\text{total}} \leftarrow \begin{cases} 0, & C = 0 \\ S/C, & \text{otherwise} \end{cases}$ 
18: return  $(M, L_{\text{total}})$ 
```

To determine which strategy is more suitable for our model, we conducted comparative experiments to evaluate the ACA approach, the ACA variant with a traditional gate, and a non-ACA baseline. As shown in [Supplementary Figure 11a](#), both the latent-only ACA and the gated ACA outperform the non-ACA method, with the latent-only variant achieving the best results. We hypothesize that this is because certain AC phenomena cannot be captured by traditional similarity rules. As illustrated in [Supplementary Figure 21](#), in

some cases important ACs containing quantitative SAR signal may be overlooked due to heuristic thresholds and inherent limitations, which can restrict the model’s ability to learn from these ACs. To test this hypothesis, we examined performance under different levels of structural constraint ([Supplementary Figure 12](#)) by varying the classical similarity threshold from 0.3 to 0.9. We found that relaxing the traditional constraint increases the number of mined triplets and indeed improves model performance.

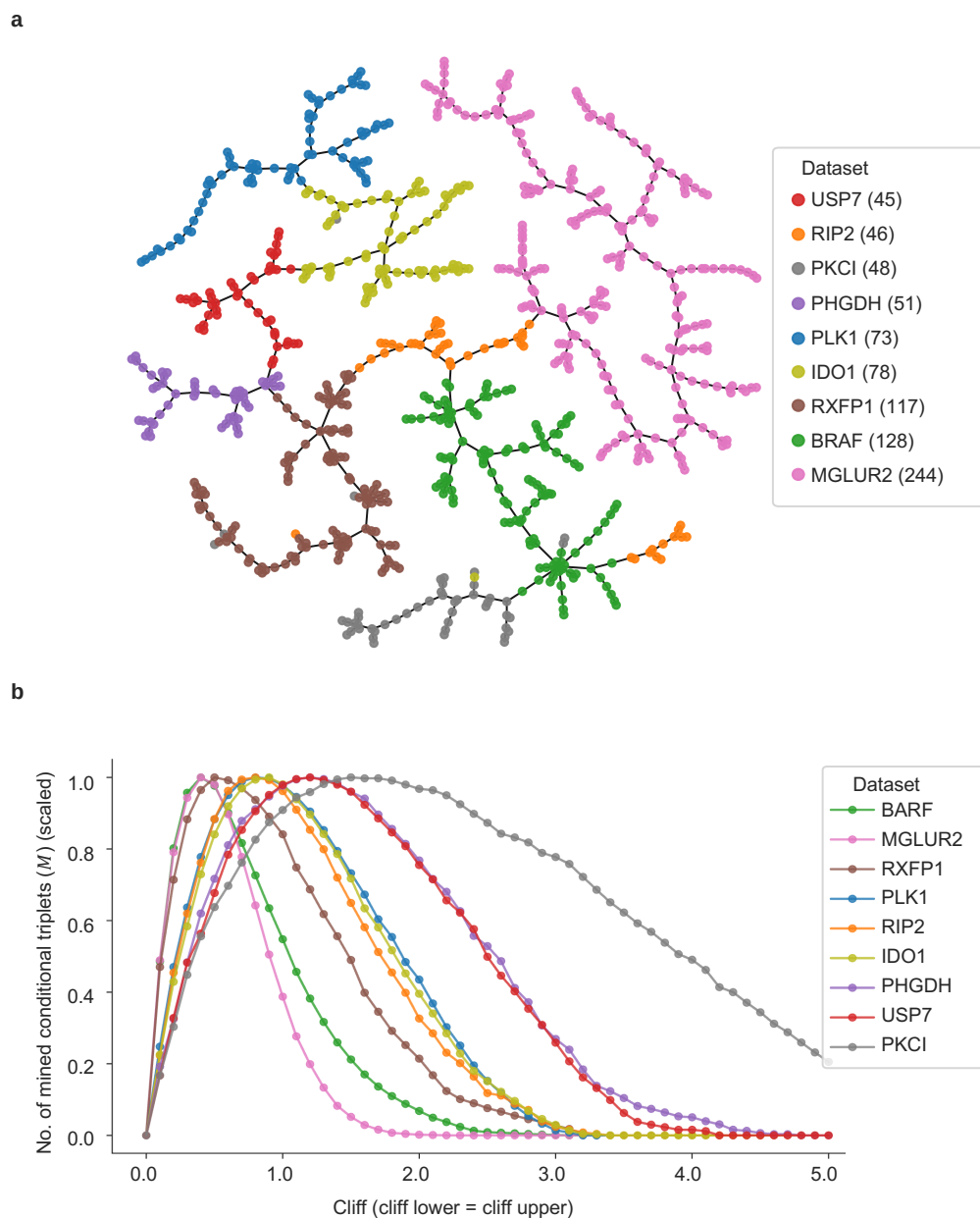
Supplementary Note 2: Assessing ACA against classical quantitative SAR methods

To evaluate whether ACA has advantages over traditional (non-machine-learning) quantitative SAR paradigms, we compared it against two canonical baselines: a scaffold-based Free–Wilson model¹ and a descriptor-based Partial Least Squares (PLS) regressor². For Free–Wilson, we adhered to its single-scaffold assumption by constructing an HSSMS subset containing only molecules with the most frequent Bemis–Murcko scaffold ([Supplementary Table 10](#)); for PLS, we used the full HSSMS set with standard topological/physicochemical descriptors. All methods used the original preprocessing and data splits. As shown in [Supplementary Figure 17](#) and [Supplementary Figure 19](#), ACA consistently and substantially outperforms both Free–Wilson and PLS across targets and metrics.

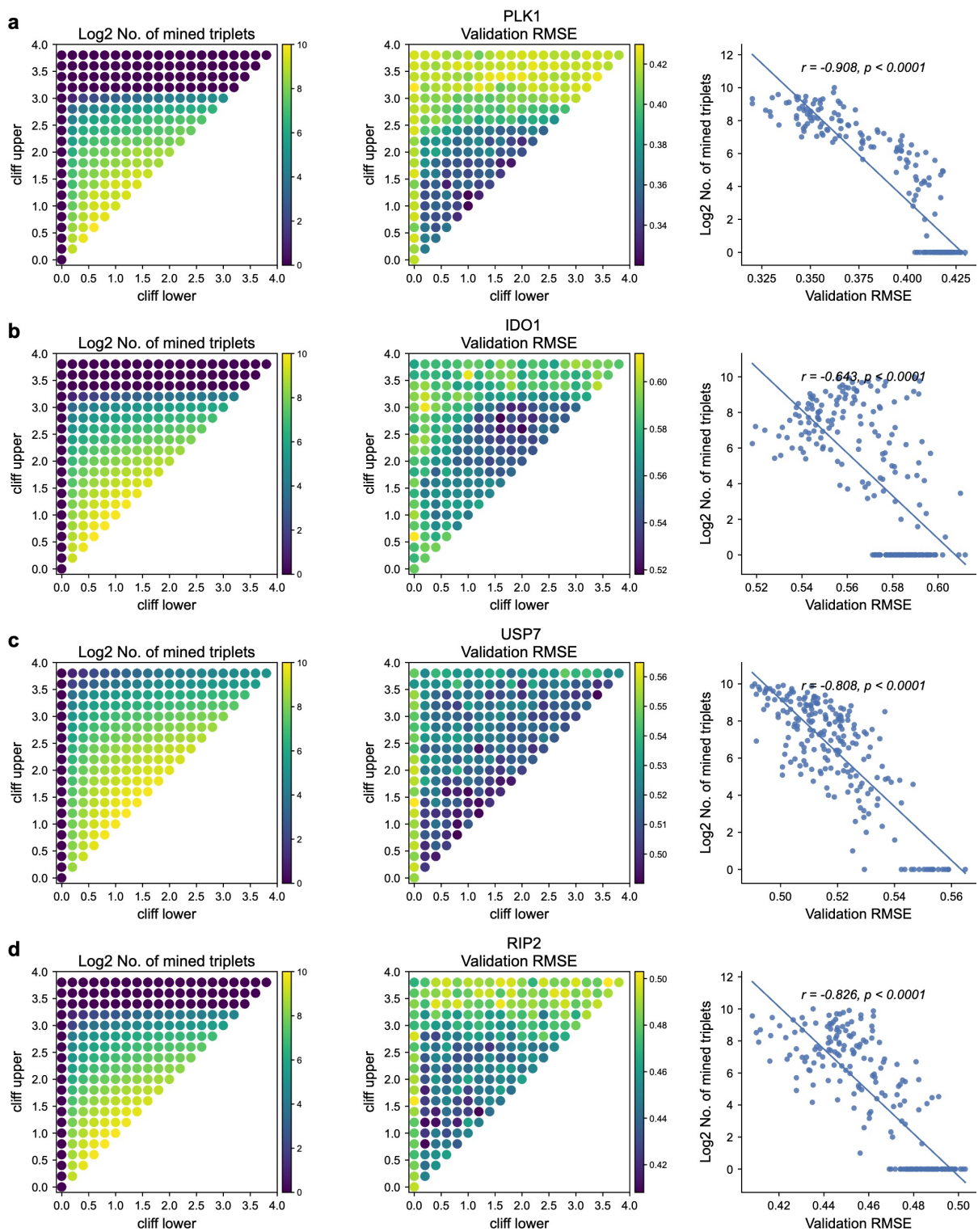
Supplementary Note 3: Stringent split strategies for validating the generalization of ACA

To test whether ACA can identify previously unseen cliffs, we first directly performed a Bemis–Murcko scaffold-based split: molecules were clustered by their scaffolds and 20% of scaffolds were assigned to the test set, with the remaining scaffolds used for training and validation. As summarised in [Supplementary Table 14](#), ACA achieves lower RMSE than the non-ACA baseline on all the 30 HSSMS targets under this setting, indicating that ACA can already generalize across different chemotypes. To construct an even more stringent independent test set, we further pruned the test partition by removing any molecule whose fingerprint Tanimoto, scaffold Tanimoto, or SMILES-edit similarity to any training compound exceeded 0.90, thereby striving to ensure that no test molecule could form an AC with the training set. This procedure yielded a test split in which, on average, 76.8% of molecules were annotated as cliffs ([Supplementary Table 12](#)). On this stringent split, ACA models achieved lower RMSE on 28 of 30 targets than their non-ACA counterparts ([Supplementary Figure 20](#)), suggesting that ACA can generalize to—and make use of—cliff information even when the relevant cliffs are encountered only at test time.

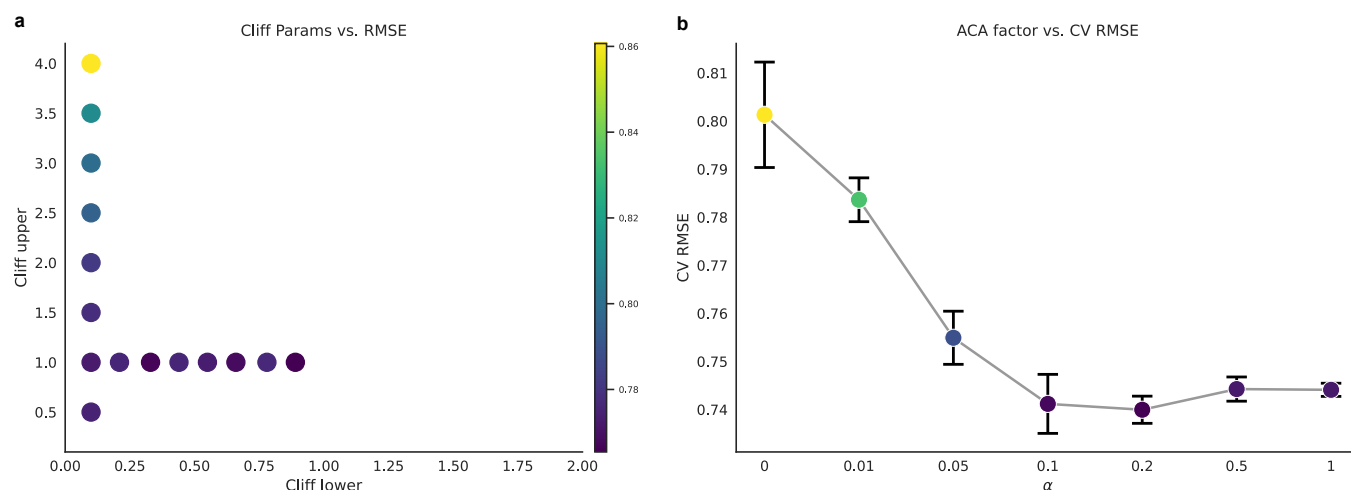
Supplementary Figures



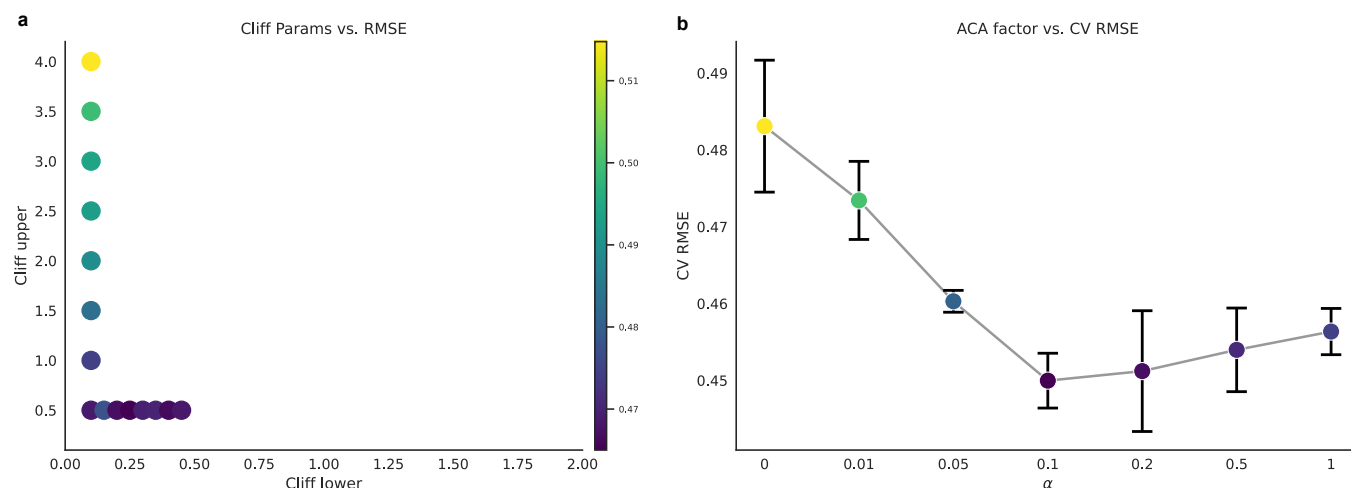
Supplementary Figure 1: Chemical space distribution of the nine LSSNS datasets and their number of mined triplets under different cliff parameters. (a) Chemical space distribution of the nine LSSNS datasets is illustrated, where the structural similarities are measured using the Tanimoto distance of ECFP fingerprints. (b) The number of mined conditional triplets (M) for each dataset is shown when the cliff lower parameter is equal to the cliff upper parameter. Source data are provided as a Source Data file.



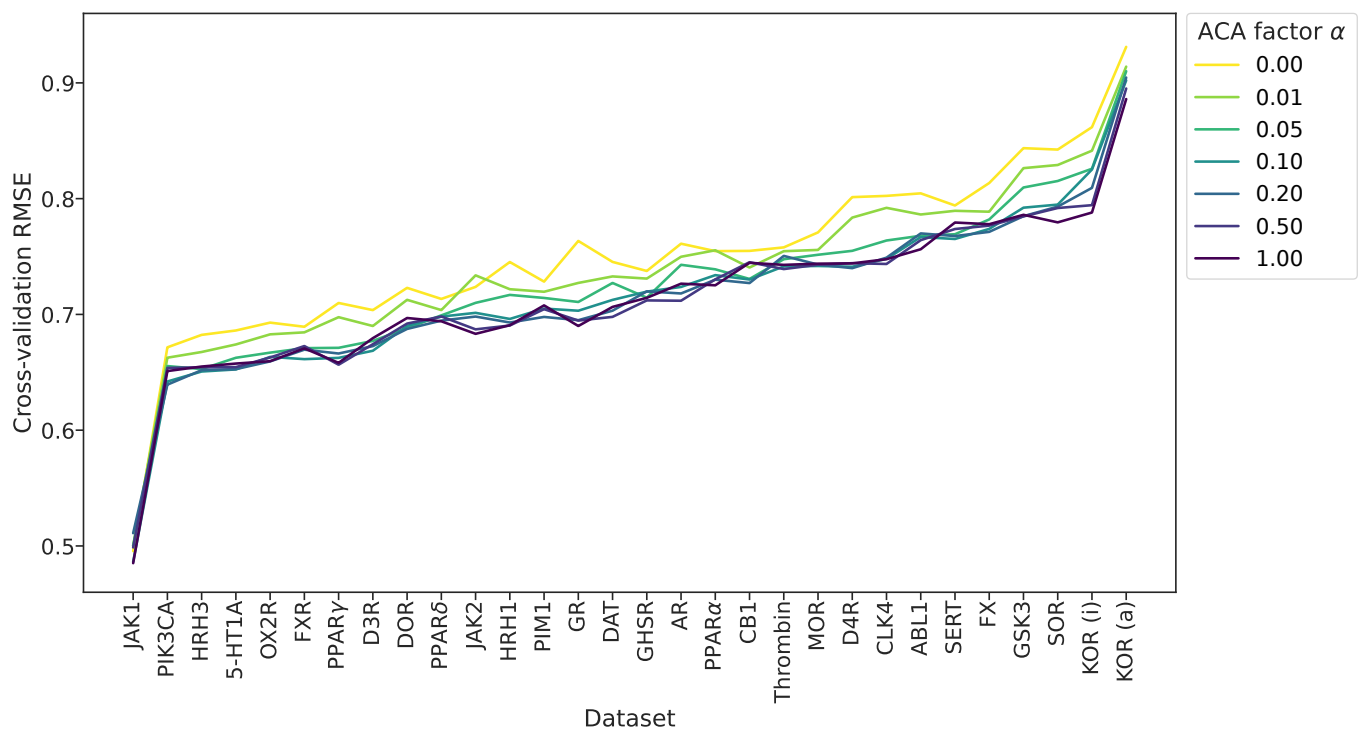
Supplementary Figure 2: Correlations between the number of mined conditional triplets (M) in the training set and validation RMSE performance. Shown are the ablation study results for four additional LSSNs datasets: PLK1 (a), IDO1 (b), USP7 (c), and RIP2 (d). The cliff lower and cliff upper thresholds were varied from 0 to 4 with a step size of 0.1, while the awareness factor α was fixed. The number of mined conditional activity cliff triplets (ACTs), M , under each parameter setting is shown, with purple dots indicating settings for which no conditional ACTs were mined ($M = 0$). Validation RMSE values were obtained from five-fold cross-validation, and the correlation between the $\log_2$ -transformed number of mined label-conditioned ACTs and validation RMSE is shown. Because the absolute RMSE ranges differ across datasets, the heatmap color scale is set separately for each dataset to preserve within-dataset contrast and avoid visual compression under a single global scale. Source data are provided as a Source Data file.



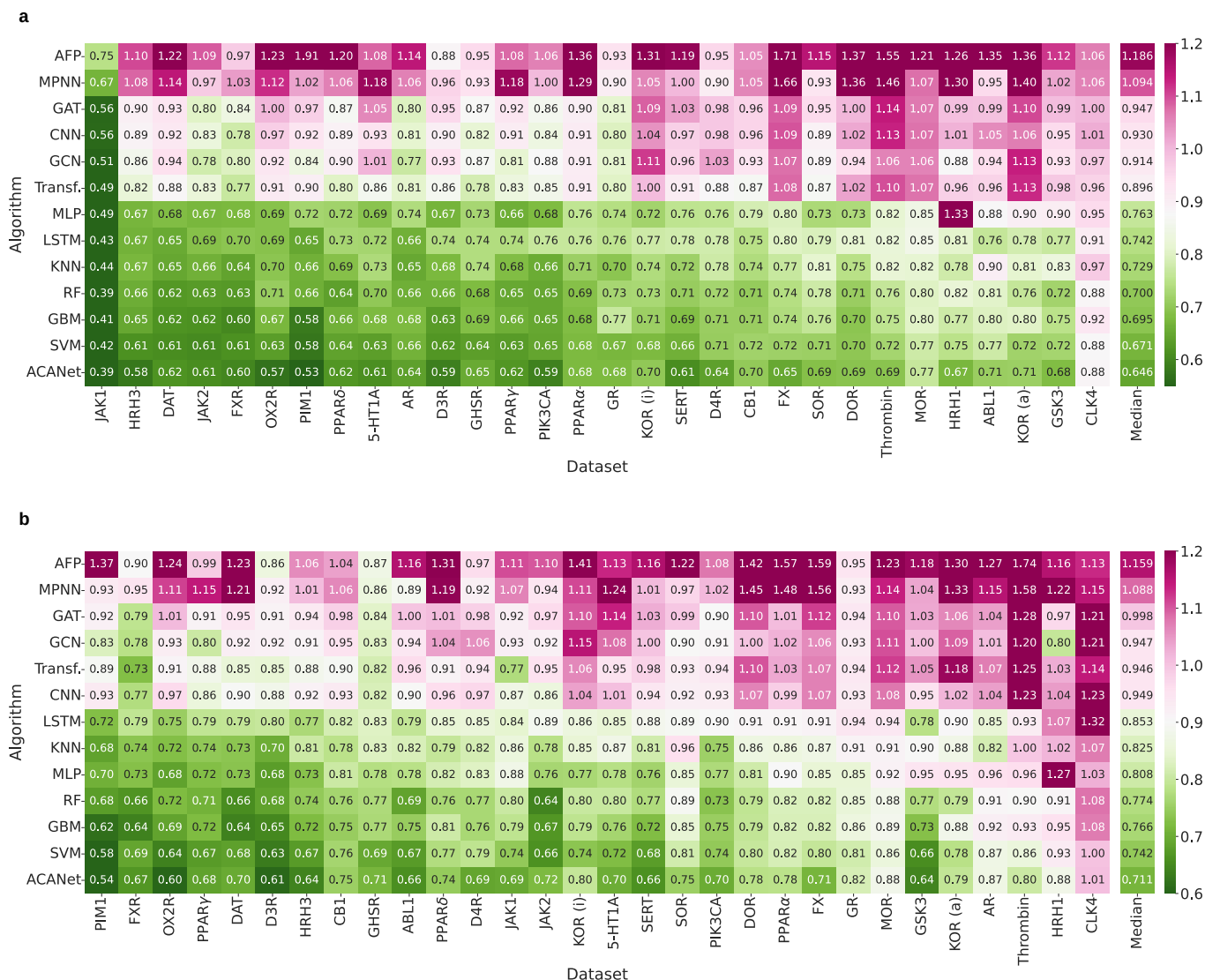
Supplementary Figure 3: Example of hyperparameter tuning for cliff lower, cliff upper, and ACA factor α on the training set of the D4R dataset from the HSSMS collection. (a) Example of “T-shaped” grid search for cliff lower and cliff upper. Specifically, cliff lower was first fixed at 0.1, and cliff upper was varied from 0.5 to 4.0 with a step size of 0.5. Then, using the best cliff upper value, cliff lower was varied from 0.1 to the selected cliff upper value with 8 steps. (b) Optimization of the ACA factor α under the selected cliff lower and cliff upper parameters. The ACA factor α was searched over [0, 0.01, 0.05, 0.1, 0.2, 0.5, 1]. The RMSE from nested fivefold cross-validation (CV) was used as the model performance metric. Error bars indicate the standard deviation of the best smoothed CV RMSE across repeated runs. Source data are provided as a Source Data file.



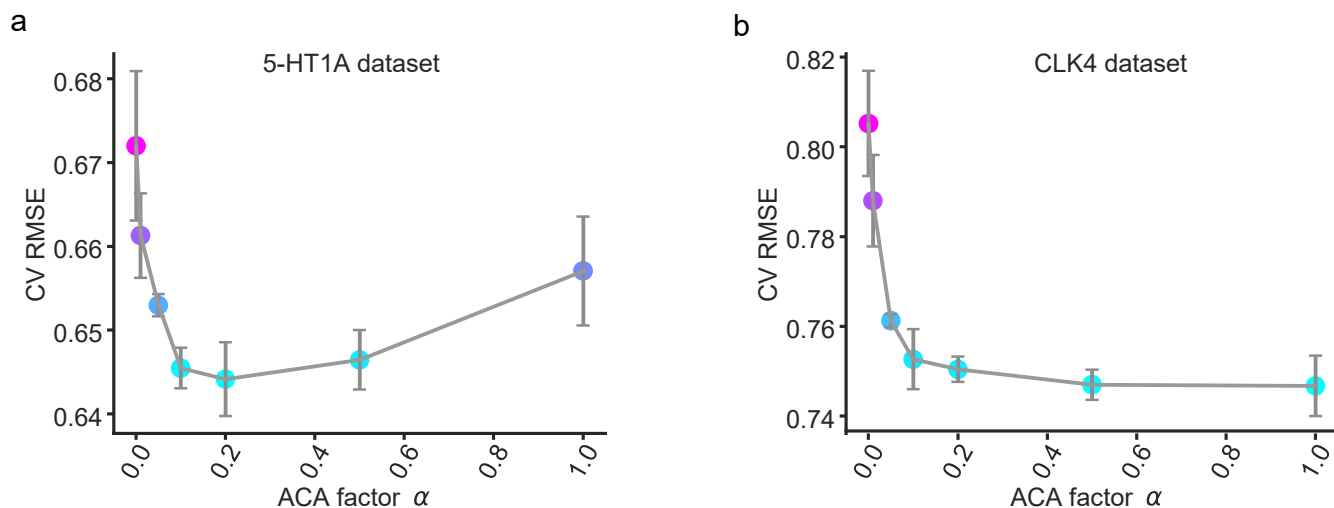
Supplementary Figure 4: Example of hyperparameter tuning for cliff lower, cliff upper, and ACA factor α on the training set of the Caco2 permeability prediction dataset from the ADMET collection. (a) Example of “T-shaped” grid search for cliff lower and cliff upper. Specifically, cliff lower was first fixed at 0.1, and cliff upper was varied from 0.5 to 4.0 with a step size of 0.5. Then, using the best cliff upper value, cliff lower was varied from 0.1 to the selected cliff upper value with 8 steps. (b) Optimization of the ACA factor α under the selected cliff lower and cliff upper parameters. The ACA factor α was searched over [0, 0.01, 0.05, 0.1, 0.2, 0.5, 1]. The RMSE from nested fivefold cross-validation (CV) was used as the model performance metric. Error bars indicate the standard deviation of the best smoothed CV RMSE across repeated runs. Source data are provided as a Source Data file.



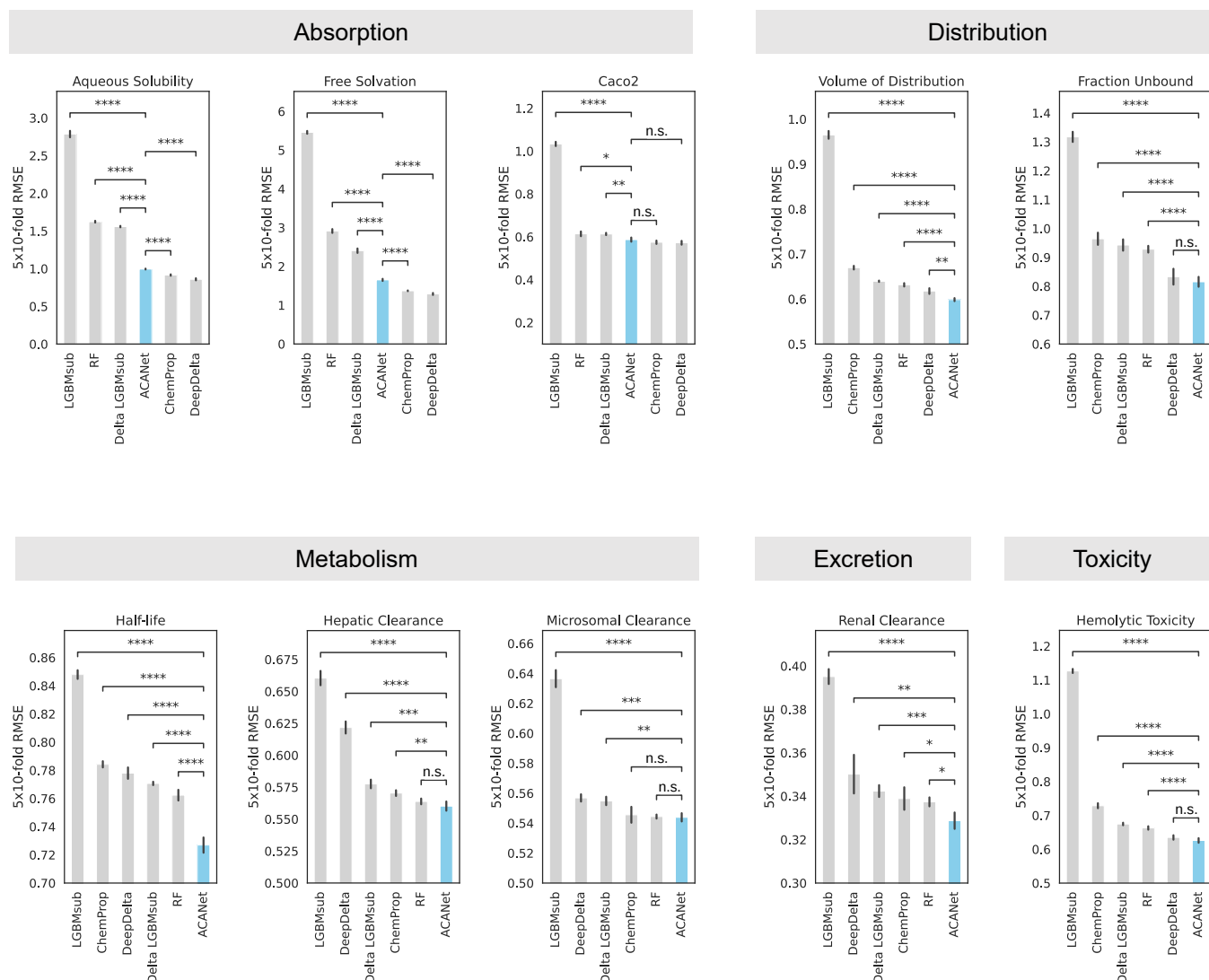
Supplementary Figure 5: Impact of ACA factor α on model performance for 30 HSSMS datasets The evaluation metric used in this analysis is the RMSE of nested fivefold cross-validation, which provides insights into the predictive accuracy of the model. By varying the α value, we can observe how it influences the performance of the model across different datasets. When α equals zero (the yellow line), it indicates the absence of ACA in the model. Source data are provided as a Source Data file.



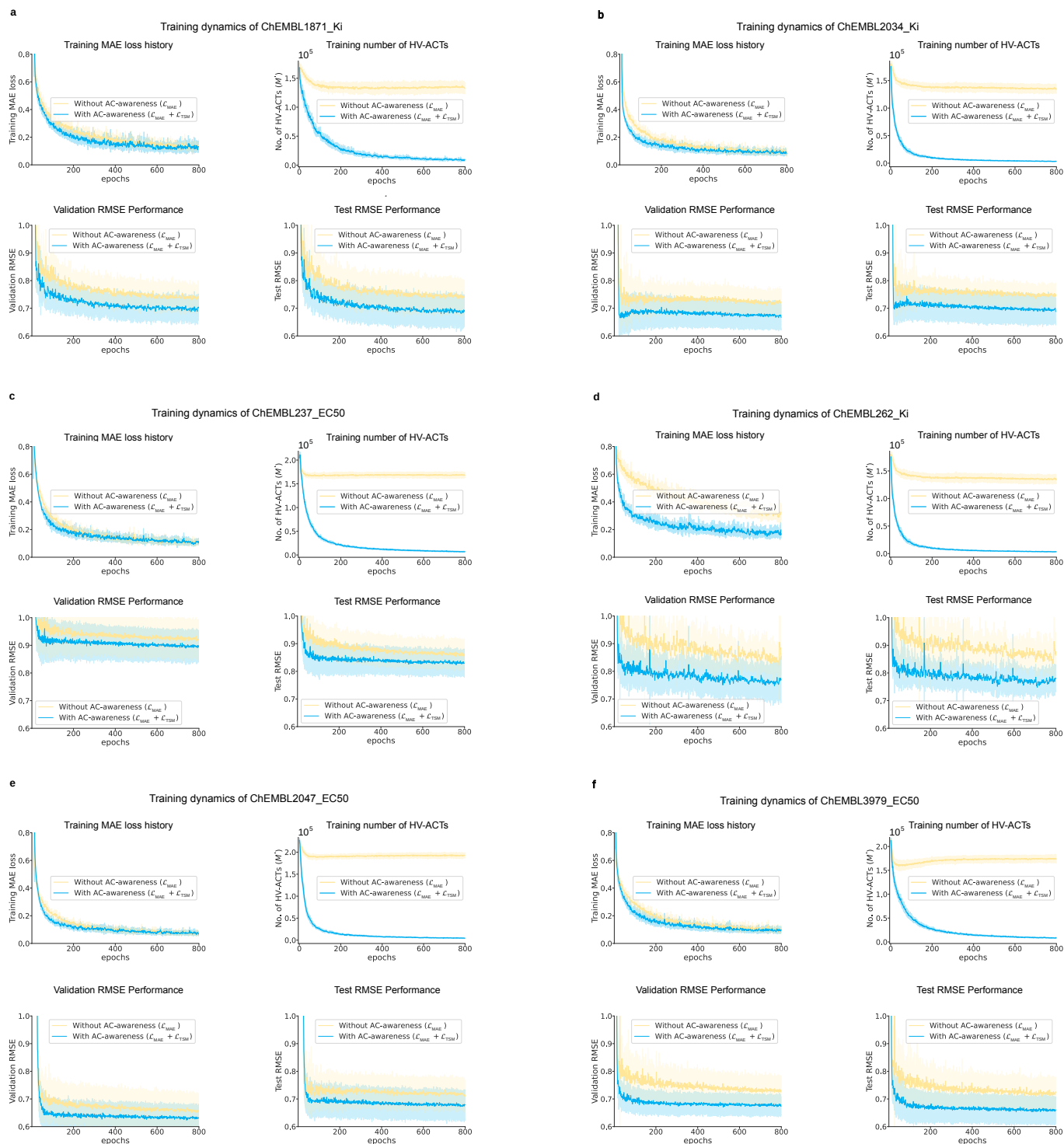
Supplementary Figure 6: Performance comparison of ACANet with off-the-shelf ML models for molecular activity prediction on 30 HSSMS benchmark datasets. (a) the RMSE performance on the entire test set (RMSE_{all}) and (b) the RMSE performance on the cliff-specific test set (RMSE_{cliff}). Source data are provided as a Source Data file.



Supplementary Figure 7: Effect of the ACA factor α on the RMSE performance of the easy case 5-HT1A dataset and hard case CLK4 dataset for molecular activity prediction. (a) Effect of ACA factor α on 5-HT1A performance with fixed cliff lower and cliff upper. Nested cross-validation RMSE was reported with increasing values of α (0., 0.02, 0.05, 0.1, 0.2, 0.5, and 1.0). (b) Effect of ACA factor α on CLK4 with fixed cliff lower and cliff upper. Error bars represent the standard deviation of ten repeat runs of nested five-fold cross-validation. Source data are provided as a Source Data file.

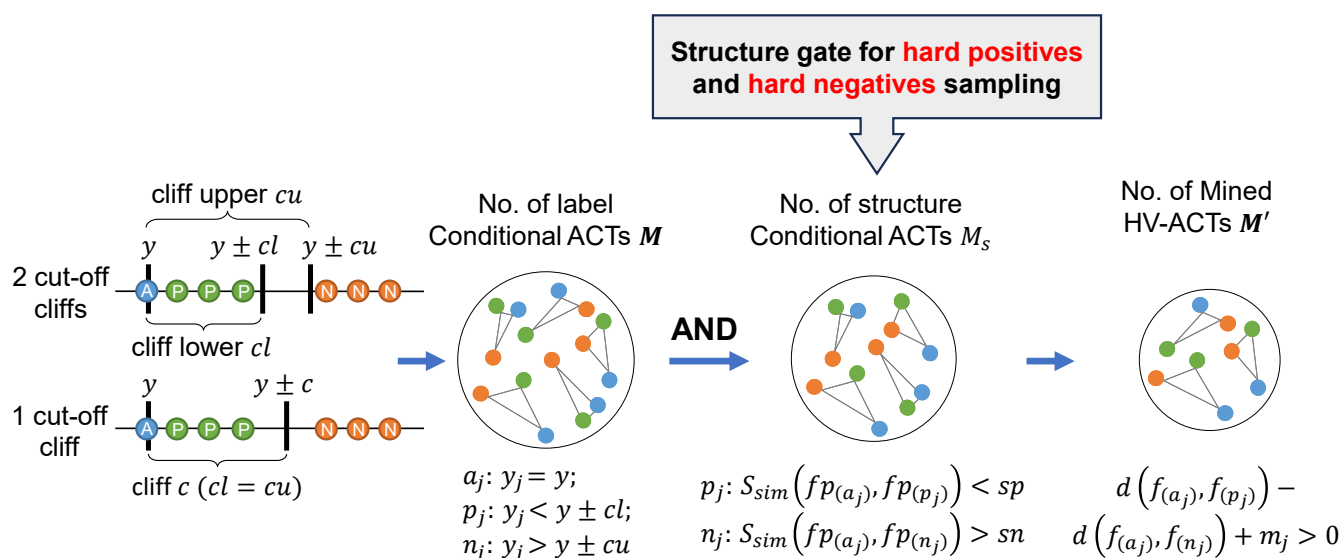


Supplementary Figure 8: RMSE performance ACA infused ACANet on ADMET delta prediction. The 10 ADMET delta prediction datasets are obtained from the DeepDelta paper³. For Random Forest (RF) and Light Gradient Boosting Machine (LGB) models, ECFP fingerprints are used. ChemProp, DeepDelta, and ACANet use 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 $\pm$ SD over five cross-validation repeats; asterisks indicate paired *t*-test significance versus ACANet (**** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. = not significant). Source data are provided as a Source Data file.

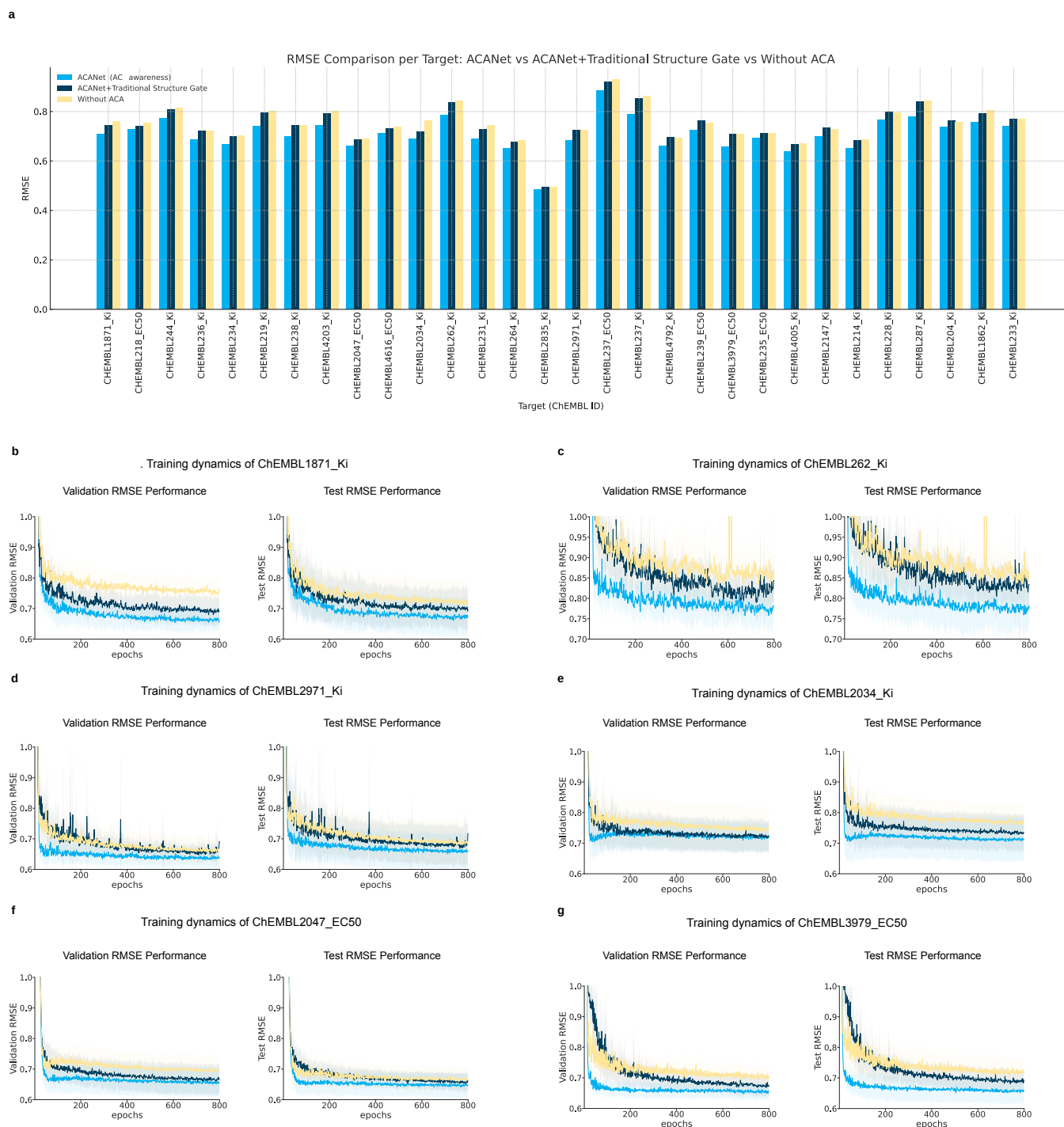


Supplementary Figure 9: Additional examples of ACANet training dynamics across datasets. (a-f) Each figure shows a target with four panels: training MAE loss, number of mined HV-ACTs, validation RMSE, and test RMSE. Curves compare the baseline (ℓ_{MAE}) and AC-aware training ($\ell_{\text{MAE}} + \ell_{\text{TSM}}$); AC-aware runs converge to lower errors while mining more HV-ACTs. Source data are provided as a Source Data file.

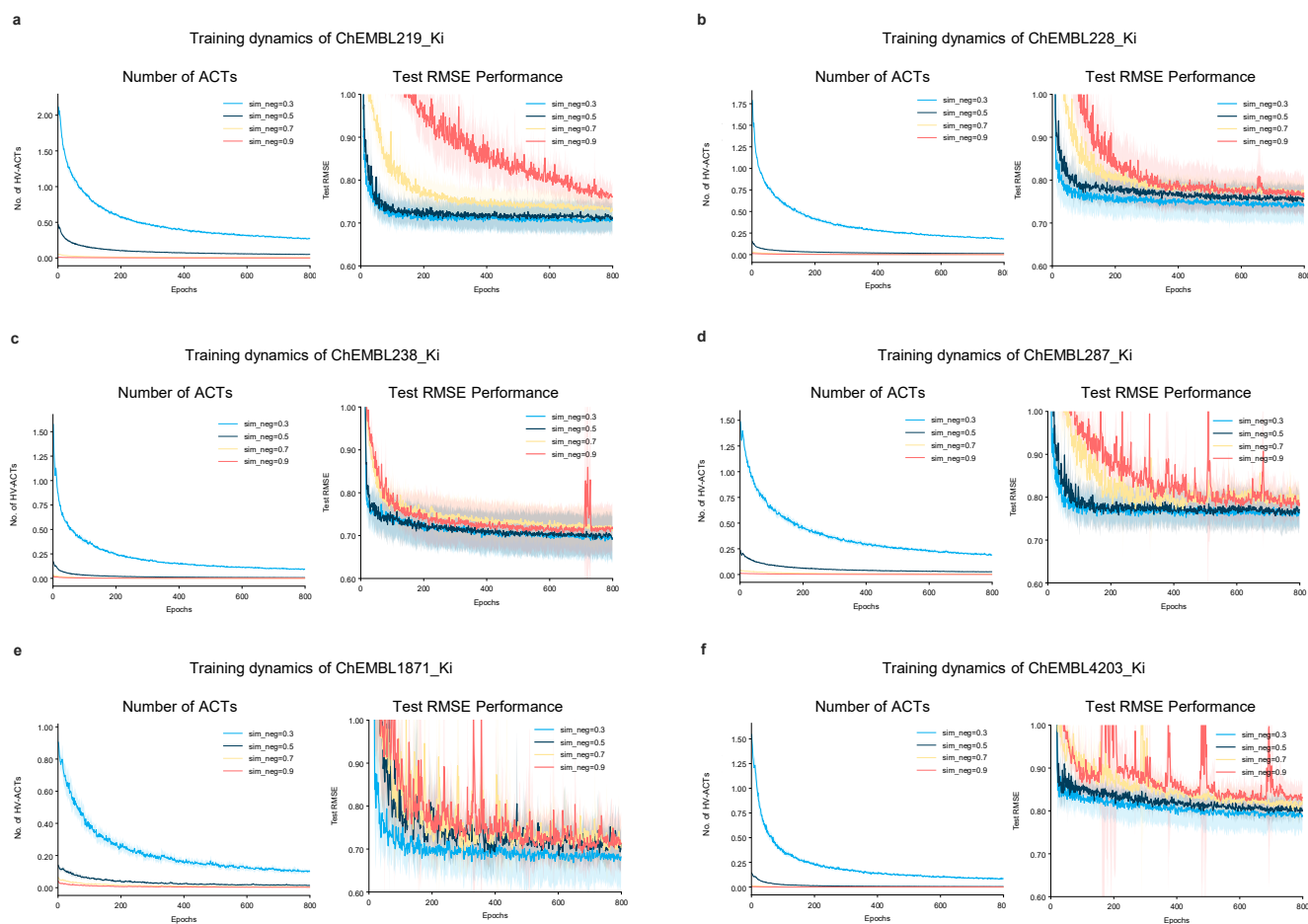
Activity-Cliff Triplet mining with structural-similarity gating



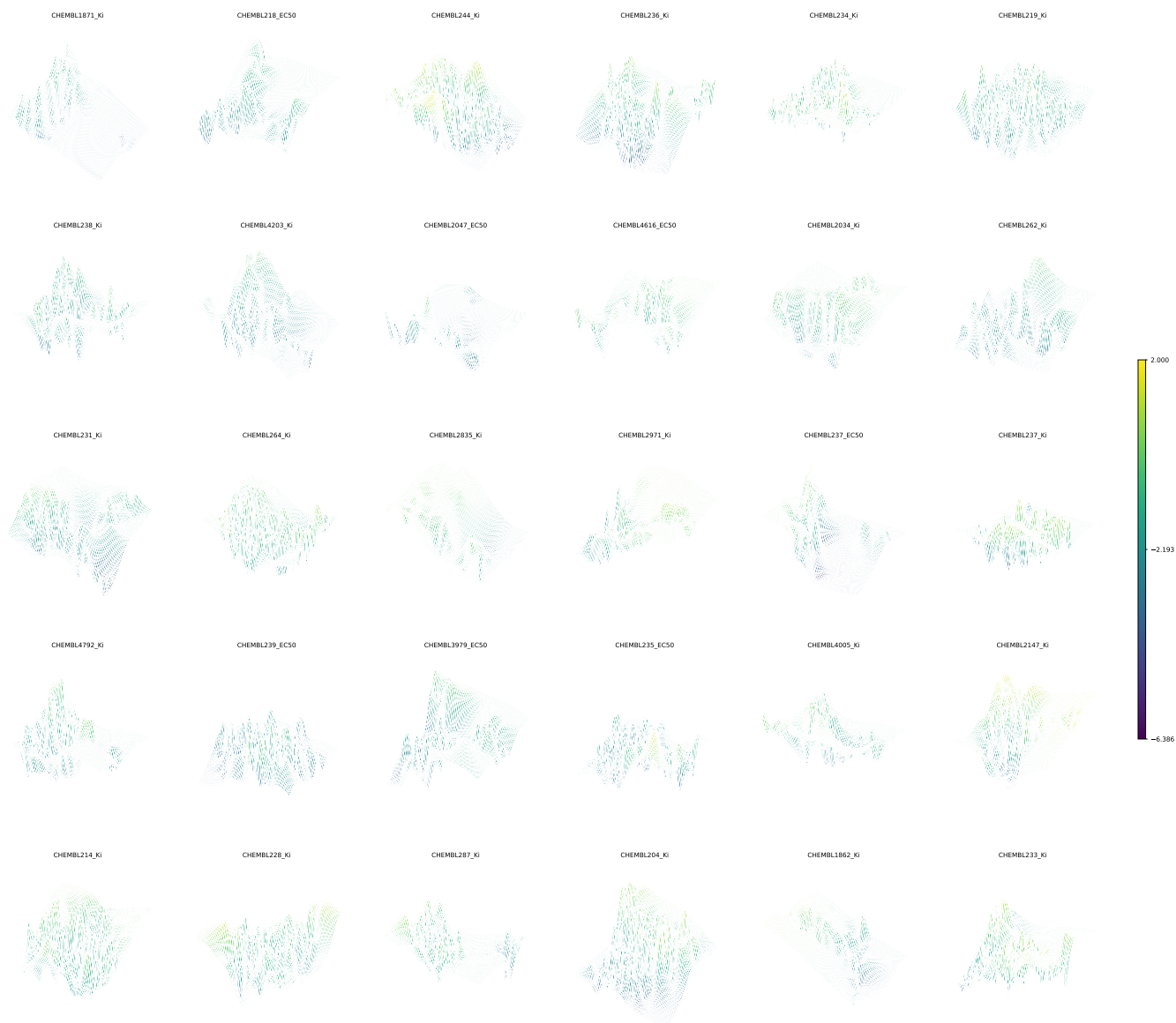
Supplementary Figure 10: Activity-cliff triplet mining with an additional structure gate. Label-conditioned ACTs (M) are first obtained via activity cut-offs; a structural-similarity gate retains only hard positives ($S_{sim} < s_p$) and hard negatives ($S_{sim} > s_n$), yielding structure-conditioned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



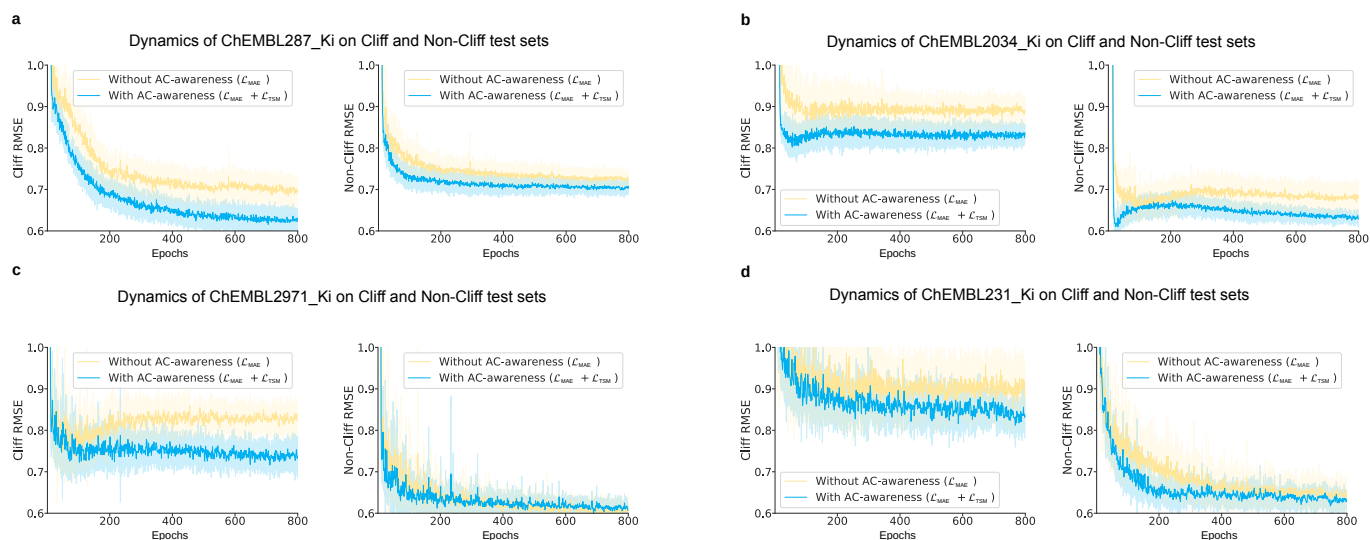
Supplementary Figure 11: Comparison of ACANet variants with and without an additional structure gate. (a) Across 30 single-scaffold HSSMS targets, the cyan bars denote ACANet (ACA-aware), which are generally lower than the dark blue bars denoting ACANet + traditional structure gate. (b–g) Training-dynamics snapshots for six example targets show the same pattern: the dark blue gated curves remain above or overlap the cyan ACA-aware curves, while the light yellow non-ACA baseline performs worst. Overall, although adding the structure gate can surpass the non-ACA baseline, the original latent-distance-based ACANet retains the best and most stable performance. Source data are provided as a Source Data file.



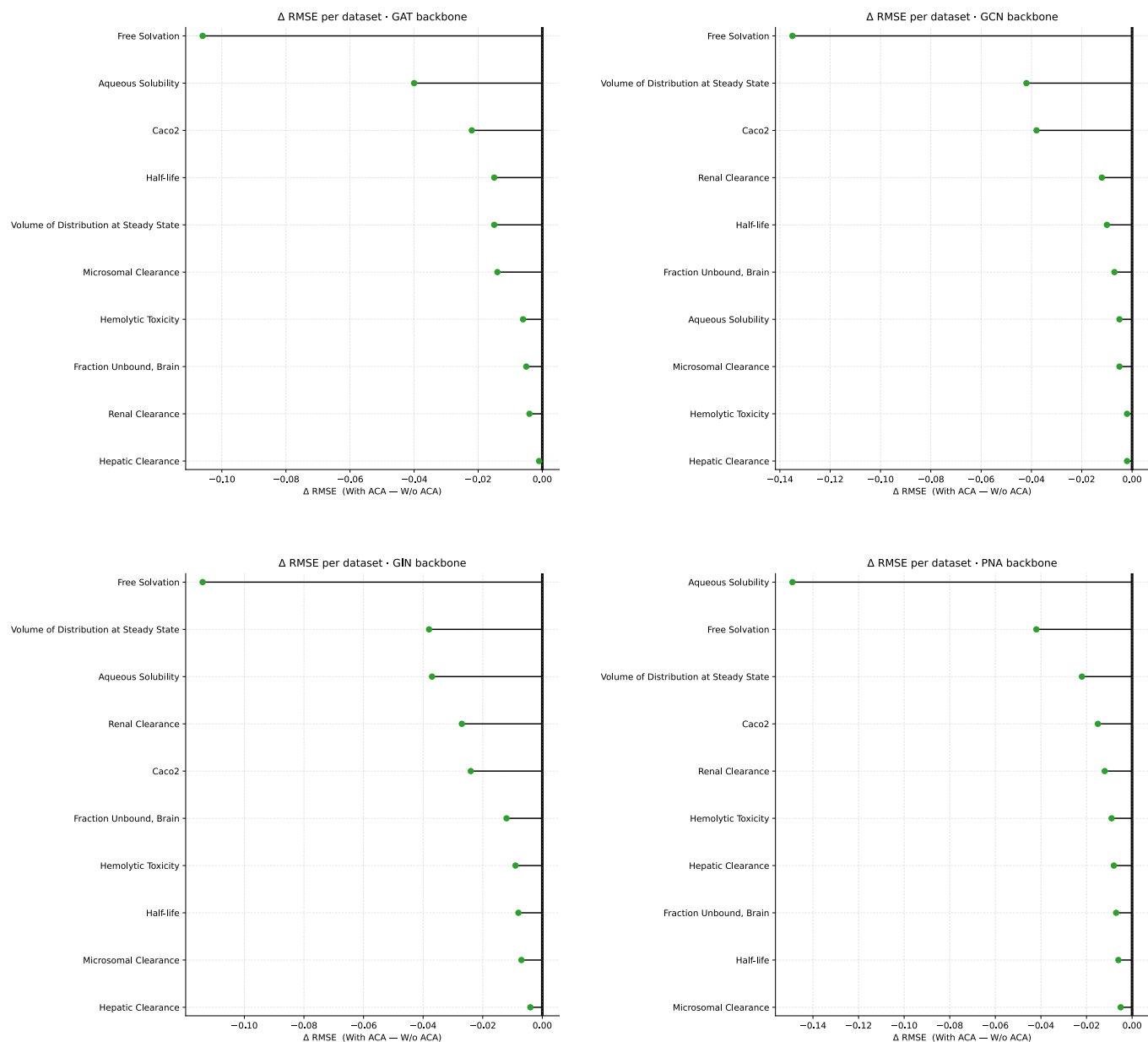
Supplementary Figure 12: Effect of structural-similarity gate on mined HV-ACTs and test RMSE. (a–f) Training dynamics on six example HSSMS targets. For each target, the number of mined high-value ACTs per epoch is reported in units of 10^5 , and the corresponding test-set RMSE curves are shown under different similarity thresholds ($\text{sim_neg} = 0.3, 0.5, 0.7, 0.9$). 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. Source data are provided as a Source Data file.



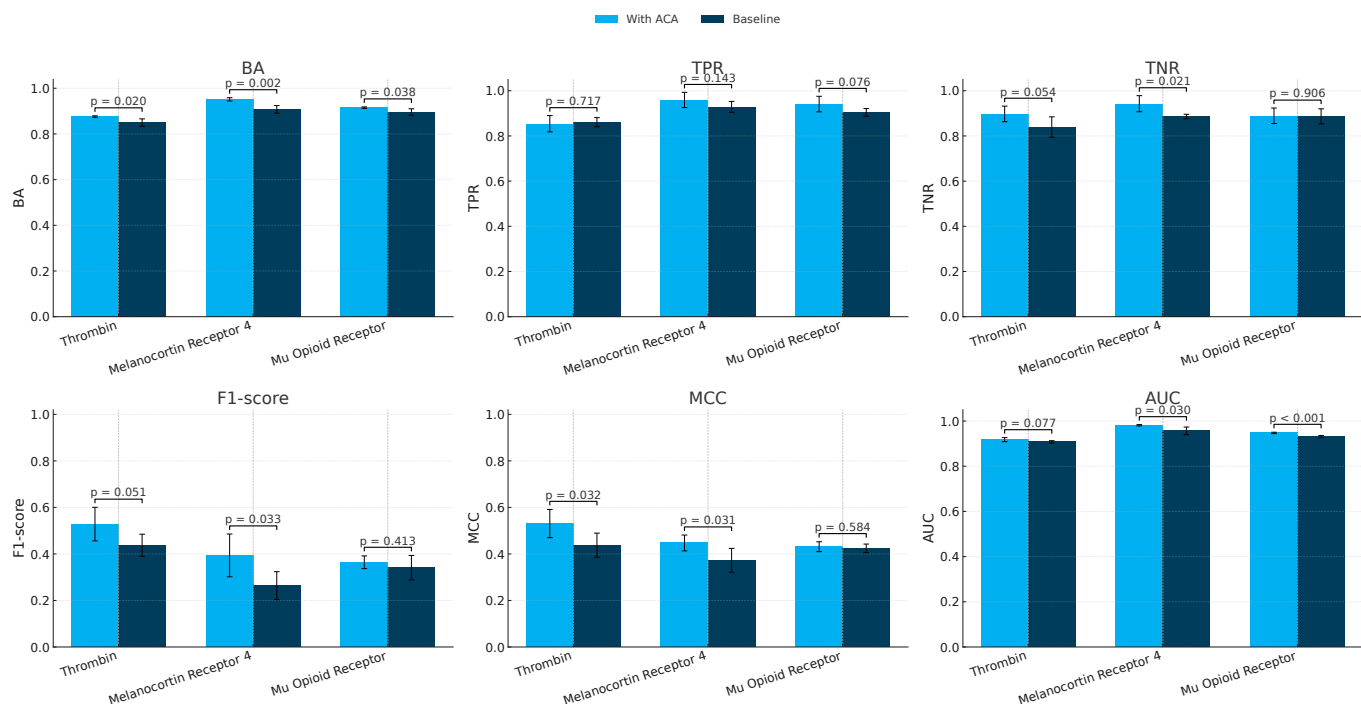
Supplementary Figure 13: Three-dimensional activity-landscape surfaces across 30 HSSMS datasets. For each dataset, molecules were represented by 1024-bit Morgan fingerprints and projected onto the first two principal components. The activity surface was generated by radial basis function (RBF) interpolation over the two-dimensional PCA space. A shared colour scale is used across all panels, where dark blue/purple indicates lower activity values and yellow indicates higher activity values. Source data are provided as a Source Data file.



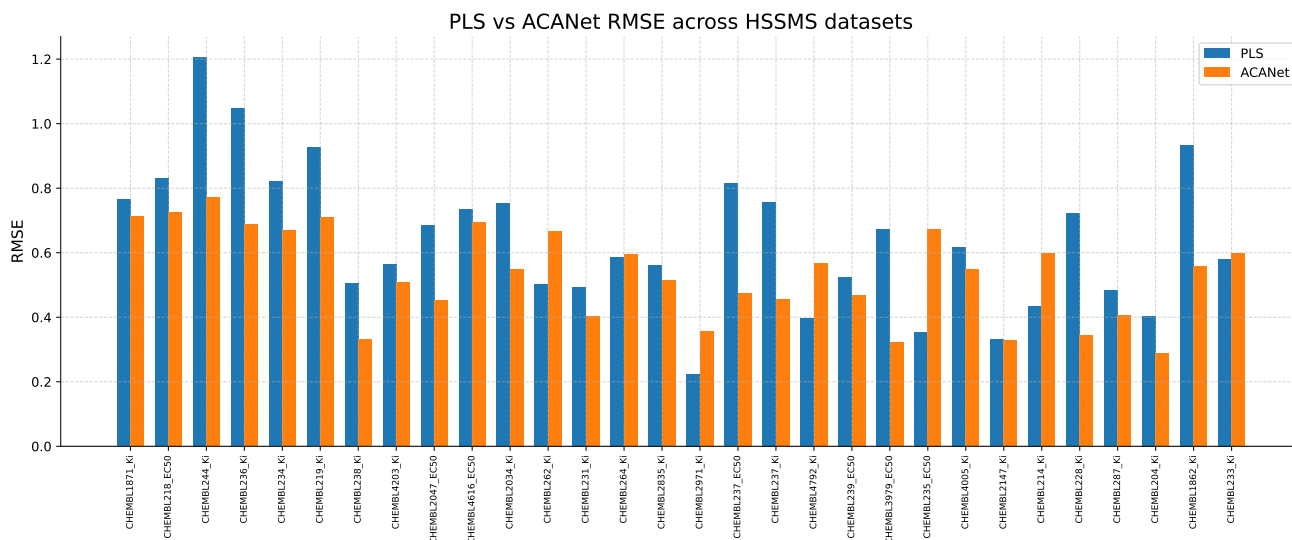
Supplementary Figure 14: Training dynamics comparing AC-aware training versus the baseline. (a–d) RMSE curves on activity-cliff and non-cliff test sets. AC-aware training consistently improves performance on activity-cliff and non-cliff data compared to the baseline.



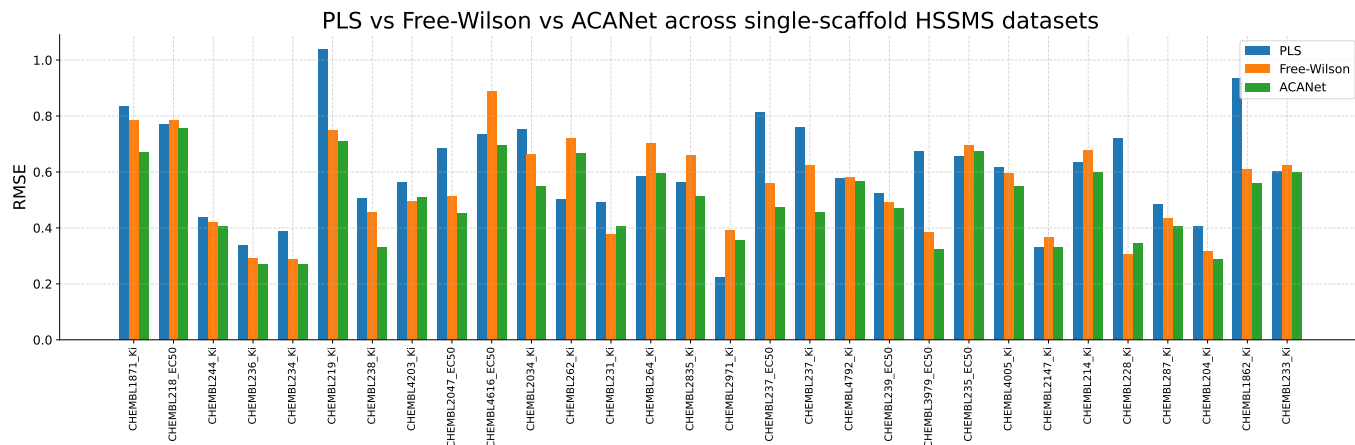
Supplementary Figure 15: ΔRMSE per dataset on ADMET across GAT, GCN, GIN, and PNA backbones under fixed ACA hyperparameters. All ACA-aware models are trained with the same default settings ($c_u = 1$, $c_l = 1$, $\alpha = 1$) and compared against their non-ACA counterparts. For each target, the change in test error is shown as $\Delta \text{RMSE} = \text{RMSE}_{\text{With ACA}} - \text{RMSE}_{\text{W/o ACA}}$. Lollipop stems are centered at 0; green circles indicate improvement ($\Delta \text{RMSE} < 0$, corresponding to lower RMSE with ACA), and no red circles are present, indicating that no dataset experienced RMSE degradation with ACA. Datasets in each panel are sorted by ΔRMSE . Under this fixed configuration, ACA consistently outperforms the non-ACA baseline, showing superior predictive performance across all ADMET targets. Source data are provided as a Source Data file.



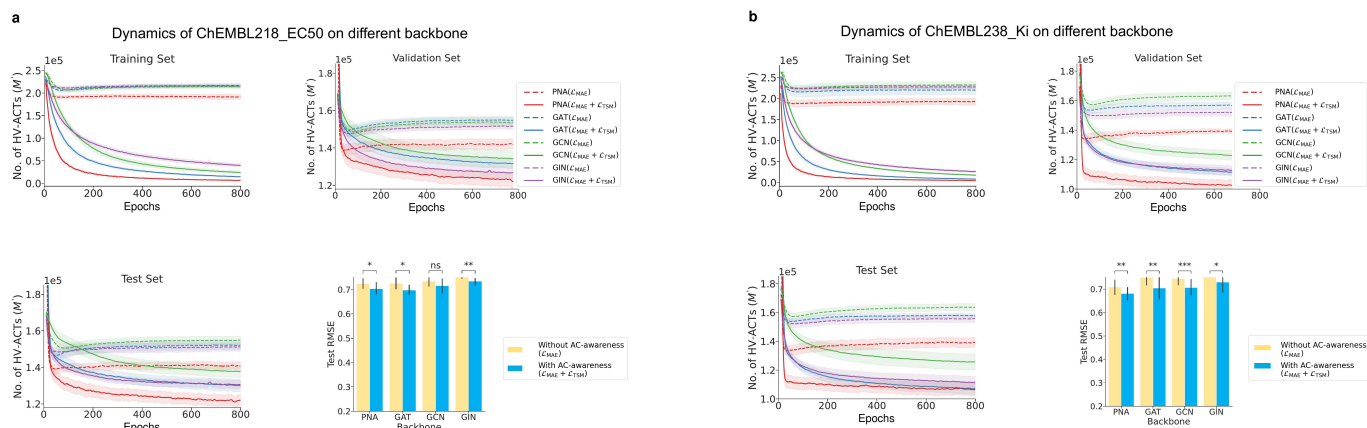
Supplementary Figure 16: Performance of MMP models with and without ACA. Across three datasets, per-metric bars (BA, TPR, TNR, F1-score, MCC, AUC) show mean $\pm$ s.d. over 5×5 CV (five random seeds, five folds). *p*-values (Welch's *t*-test) are annotated above paired bars. AC-aware training generally yields lower error and higher classification scores. Source data are provided as a Source Data file.



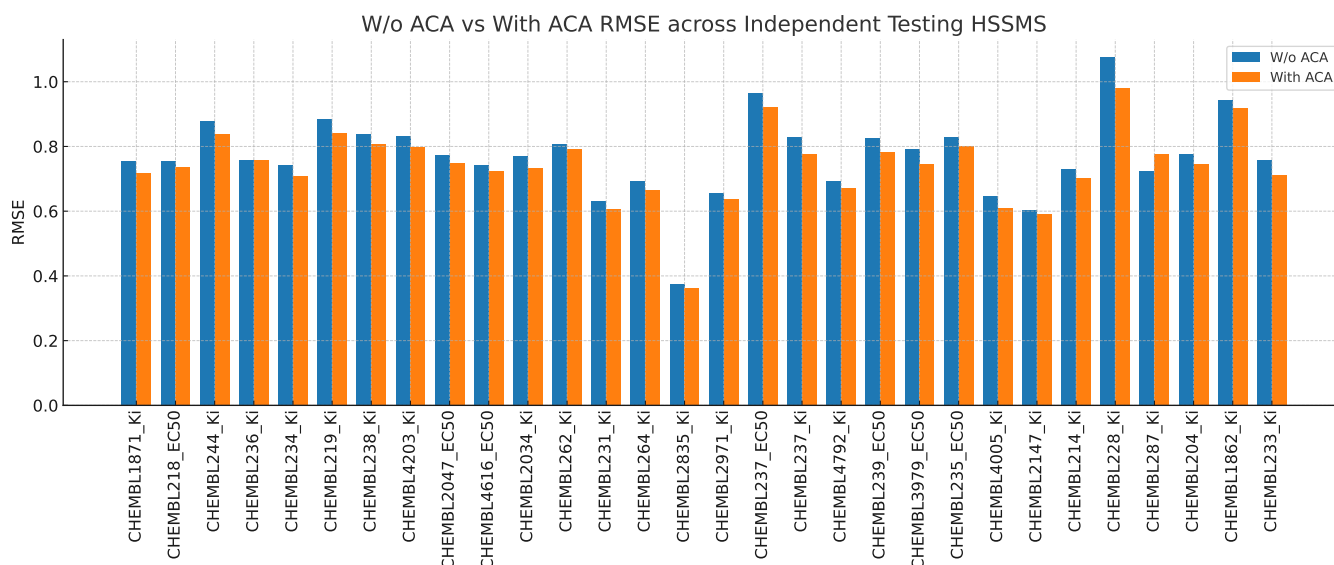
Supplementary Figure 17: PLS (blue) vs. ACANet (orange) on HSSMS. Per-dataset RMSE comparison (lower is better) across the HSSMS benchmark shows ACANet consistently outperforming PLS on most targets. Source data are provided as a Source Data file.



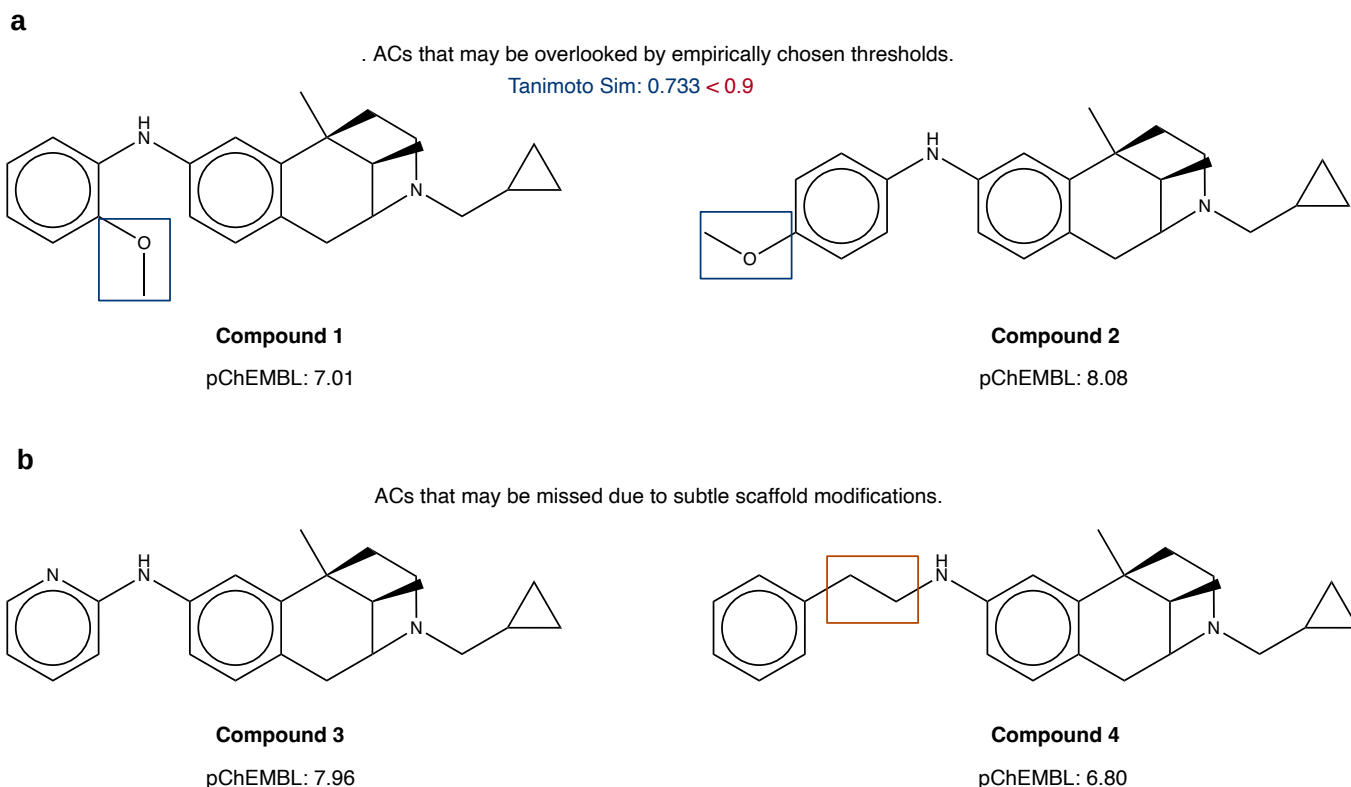
Supplementary Figure 18: PLS (blue) vs Free-Wilson (green) vs ACANet (orange) on single-scaffold HSSMS datasets. Per-dataset RMSE (lower is better) comparing PLS, Free-Wilson, and ACANet. ACANet attains the lowest error on most targets. Source data are provided as a Source Data file.



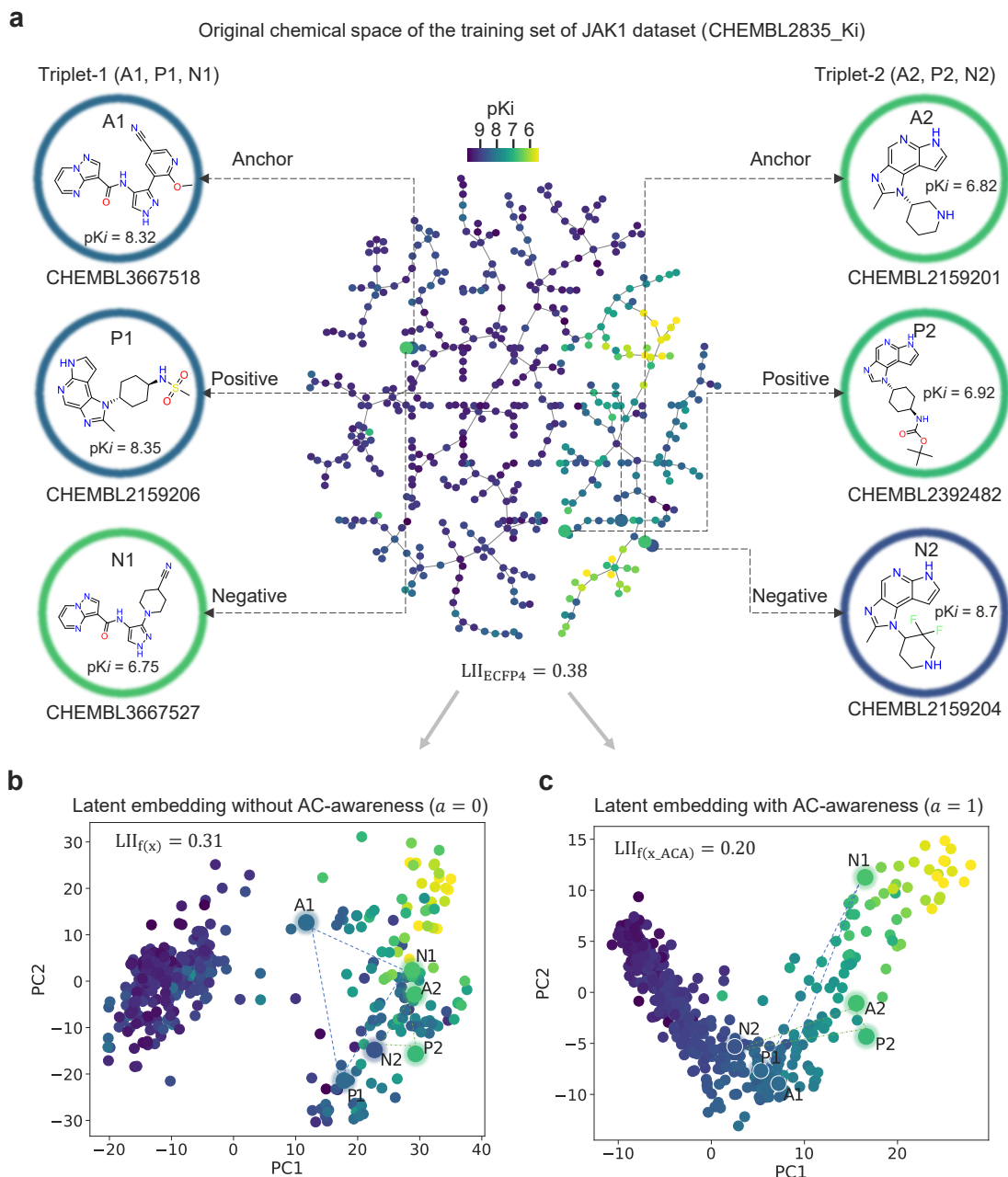
Supplementary Figure 19: More examples of backbone-agnostic gains from AC-aware training. (a, b) Training/validation/test dynamics with a 6:2:2 split repeated across 10 random seeds for two representative targets. Line plots show the number of mined HV-ACTs over epochs; bar charts report test RMSE (mean $\pm$ s.d.). Across PNA, GAT, GCN, and GIN backbones, adding the ACA loss ($\ell_{MAE} + \ell_{TSM}$) consistently improves accuracy over the baseline (ℓ_{MAE}), with statistical significance indicated above the bars. Source data are provided as a Source Data file. Source data are provided as a Source Data file.



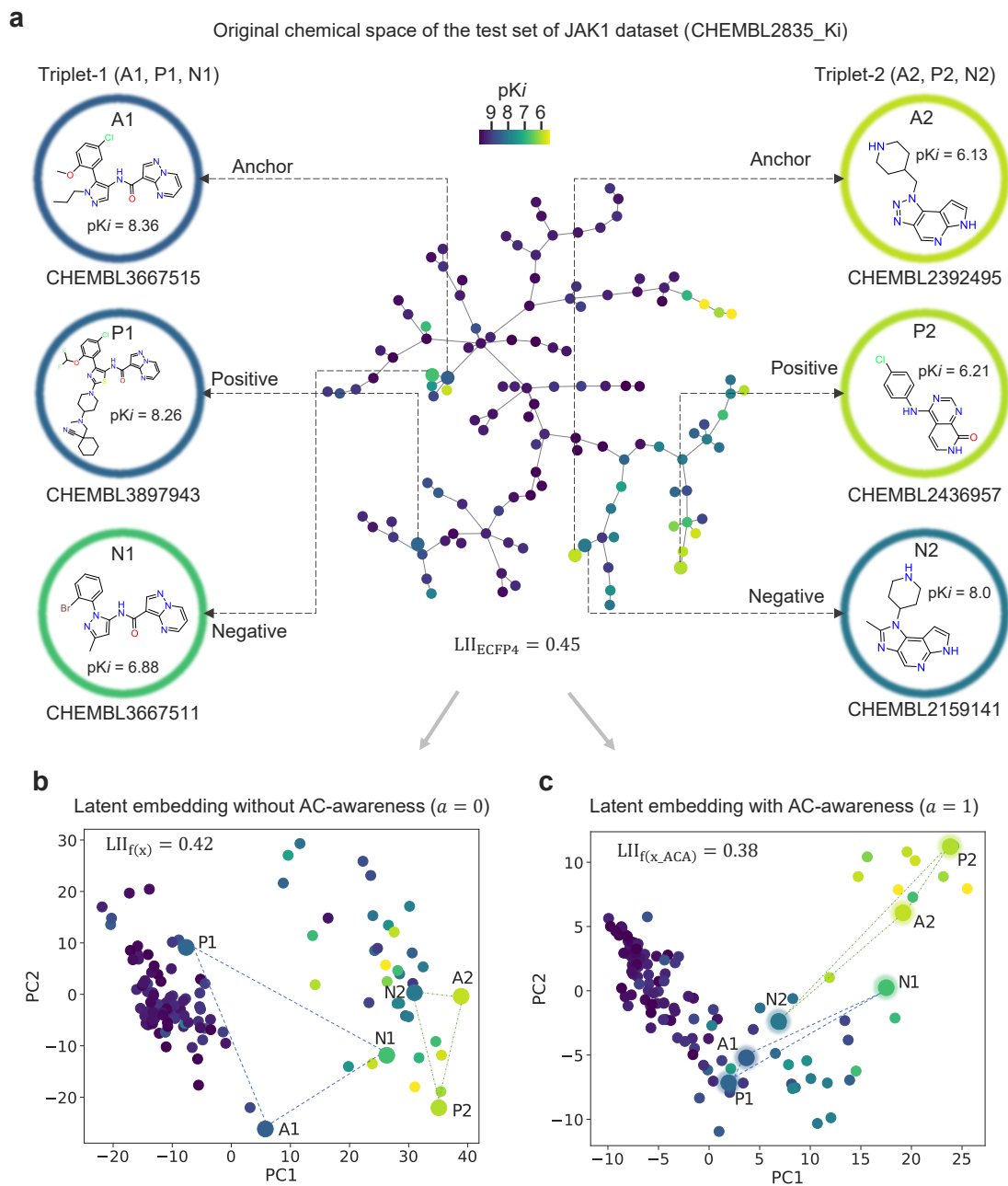
Supplementary Figure 20: ACA-aware models consistently reduce error on independent HSSMS test sets. Test-set RMSE for models trained without ACA (blue) and with ACA (orange) on HSSMS datasets that possess independent test partitions. Each pair of bars corresponds to one target, with lower bars indicating better performance. Across almost all targets, the ACA-aware variant attains lower RMSE than the non-ACA baseline, suggesting improved generalisation to unseen compounds. Source data are provided as a Source Data file.



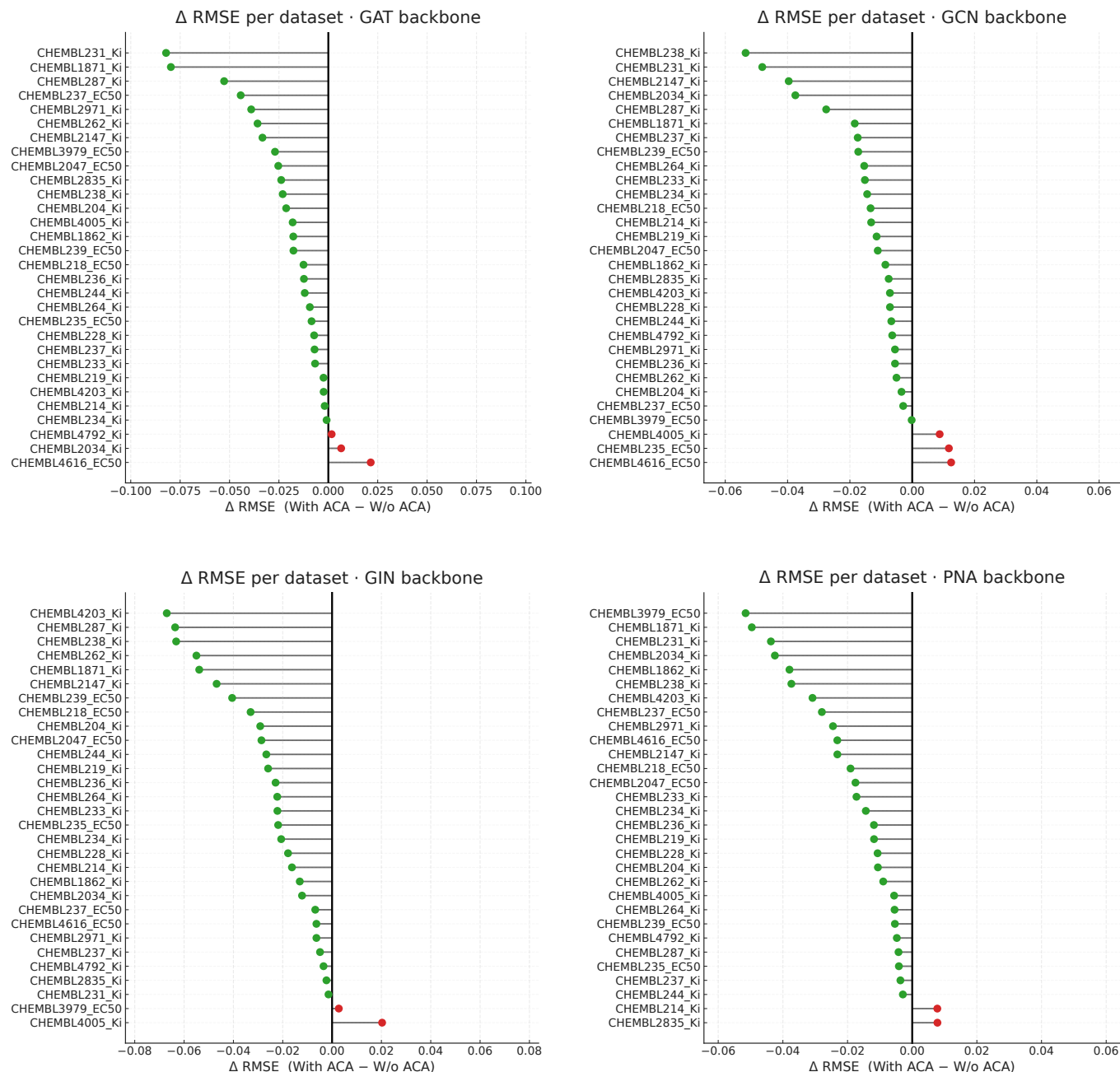
Supplementary Figure 21: Examples of activity cliffs (ACs) in the HSSMS δ -opioid receptor (DOR) dataset that elude traditional rules. (a) AC pair that would be overlooked by an empirically chosen ECFP Tanimoto threshold ($0.733 < 0.9$) despite a large potency difference. (b) AC pair that may be missed due to subtle scaffold modifications, where small changes in the highlighted region lead to pronounced activity shifts.



Supplementary Figure 22: Impact of activity-cliff awareness on latent organisation of the JAK1 training set (ChEMBL2835_Ki). (a) tmap visualisation of the original chemical space. All training compounds are laid out according to ECFP4 similarity and coloured by pK_i (blue $\rightarrow$ low, red $\rightarrow$ high). Two representative activity-cliff triplets are highlighted: A₁–P₁–N₁ and A₂–P₂–N₂, each comprising an anchor (A), a potency-matched positive (P), and a lower-potency negative (N) drawn from distinct Bemis–Murcko scaffolds. The structure-based label-incoherence index is $LII_{ECFP4} = 0.38$. (b) Latent embedding without AC awareness ($\alpha = 0$). Two-dimensional PCA of GNN embeddings trained without ACA shows that, for both scaffolds, anchors (A) remain close to their negatives (N) while their potency-matched positives (P) lie farther away, leaving scaffold and potency information entangled ($LII_{f(x)} = 0.31$). (c) Latent embedding with AC awareness ($\alpha = 1$). Including the ACA loss reorganises the latent space: anchors cluster tightly with their positives and are clearly separated from their negatives, even across different scaffolds, reducing the index to $LII_{f(x_{ACA})} = 0.20$. This illustrates ACA’s ability to disentangle scaffold identity from potency and to improve local label coherence. Source data are provided as a Source Data file.



Supplementary Figure 23: 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 pK_i (blue $\rightarrow$ low, red $\rightarrow$ high). Two representative activity-cliff triplets are highlighted: A_1 – P_1 – N_1 and A_2 – P_2 – N_2 . 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 ($\alpha = 0$). 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 ($\alpha = 1$). 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. Source data are provided as a Source Data file.



Supplementary Figure 24: Δ RMSE per dataset on HSSMS across GAT, GCN, GIN, and PNA backbones under fixed ACA hyperparameters. All ACA models are trained with the same default settings ($cu = 1$, $cl = 1$, $\alpha = 1$), and compared to their non-ACA counterparts. For each target, the change in test error is shown as $\Delta\text{RMSE} = \text{RMSE}_{\text{With ACA}} - \text{RMSE}_{\text{W/o ACA}}$. Lollipop stems are centered at 0; green circles indicate improvement (lower RMSE with ACA), while red circles indicate degradation. Datasets in each panel are sorted by ΔRMSE . Even under this fixed configuration, ACA yields markedly lower RMSE than the non-ACA baseline across most targets, demonstrating a robust, backbone-agnostic advantage. Source data are provided as a Source Data file.

Supplementary Tables

Supplementary Table 1: The 30 datasets of high-sample-size and mixed scaffolds (HSSMS) for molecular activity prediction. Datasets are the molecular-activity-prediction benchmark datasets from MoleculeACE⁴. 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	Full / 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 β (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 (PPAR α)	EC50	NR	1721 (344)	41 (41)	0.665	689
CHEMBL235_EC50	PPAR- γ (PPAR γ)	EC50	NR	2349 (470)	38 (38)	0.659	966
CHEMBL3979_EC50	PPAR- δ (PPAR δ)	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

Supplementary Table 2: The 9 datasets of low sample size and narrow scaffolds (LSSNS) for molecular activity prediction. % Cliff was identified using the same approach as MoleculeACE⁴. 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 ⁵	45	95.6	0.342	16
RIP2	Serine/threonine-protein kinase RIPK2	IC50	Kinase	Genomics Institute of the Novartis Research Foundation ⁶	46	91.3	0.346	14
PKCI	Protein kinase C iota	IC50	Kinase	Agency for Science, Technology and Research (A*STAR) ⁷	48	64.6	0.650	34
PHGDH	D-3-phosphoglycerate dehydrogenase	IC50	Other Enzyme	Boehringer Ingelheim RCV GmbH & Co. KG ⁸	51	98.0	0.377	8
PLK1	Serine/threonine-protein kinase PLK1	IC50	Kinase	Hefei University of Technology ⁹	73	98.6	0.223	18
IDO1	Indoleamine 2,3-dioxygenase	IC50	Other Enzyme	NewLink Genetics Corporation ¹⁰	78	92.3	0.421	36
RXFP1	Relaxin receptor 1	AC50	GPCR	The Florida International University Board of Trustees ¹¹	117	91.5	0.389	42
BRAF	Serine/threonine-protein kinase B-raf	IC50	Kinase	The Development Center for Biotechnology (DCB) ¹²	128	96.1	0.308	14
MGLUR2	Metabotropic glutamate receptor 2	IC50	GPCR	Eisai R&D Management Co., Ltd. ¹³	244	99.6	0.128	9

Supplementary Table 3: The 3 matched molecular pair (MMP) datasets of activity cliff classification¹⁴.

Dataset	total compounds	total MMPs	MMP-Cliffs	MMP-nonCliffs
Thrombin	3171	5751	317	4408
Mu opioid receptor	3625	8725	219	7097
Melanocortin receptor 4	1858	7169	111	5750

Supplementary Table 4: The 10 ADMET datasets used for Δ -prediction³ and their structural statistics. % Cliff was identified using the same approach as MoleculeACE⁴. 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	Endpoint/units	Size	% Cliff	Diversity	Scaffolds
Absorption	Free Solvation	Hydration free energy in water, kcal mol ⁻¹	642	92.5	0.933	63
	Caco2	log(P_{app})	910	30.8	0.741	488
	Aqueous Solubility	log S	1128	80.2	0.920	269
Distribution	Fraction Unbound, Brain	log($f_{u,brain}$)	253	15.4	0.724	200
	Volume of Distribution at Steady State	log($V_{d,ss}$; L kg ⁻¹)	1130	20.1	0.730	778
Excretion	Renal Clearance	log(CL_r)	636	5.0	0.740	447
Metabolism	Microsomal Clearance	log(clearance; mL min ⁻¹ kg ⁻¹)	731	16.6	0.693	534
	Hepatic Clearance	log(clearance; mL min ⁻¹ kg ⁻¹)	881	15.3	0.690	595
	Half-Life	log($t_{1/2}$; h)	1321	21.0	0.728	936
Toxicity	Hemolytic Toxicity	log(HD ₅₀)	828	54.7	0.682	349

Supplementary Table 5: The graph featurization in default ACANet. The 39 node features and 10 edge features are used¹⁵.

Atomic features		Description
1-16	Atomic symbols, one-hot vector of [B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te, I, At, metal]	
17-22	Node degrees, one-hot vector of [0,1,2,3,4,5]	
23	Formal charge	
24	Number of radical electrons	
25-30	Hybridization, a one-hot vector of [sp, sp2, sp3, sp3d, sp3d2, other]	
31	Whether the atom is part of an aromatic system	
32-36	Number of connected hydrogens, a one-hot vector of [0,1,2,3,4]	
37	Whether the atom is a chiral center	
38-39	Chirality type, a one-hot vector of [R, S]	
Bond features		Description
1-4	Bond type, a one-hot vector of [single, double, triple, aromatic]	
5	Whether the bond is conjugated	
6	Whether the bond is in a ring	
7-10	Stereo, a one-hot vector of [StereoNone, StereoAny, StereoZ, StereoE]	

Supplementary Table 6: Validation and Test RMSE Performance of ACANet with Different GNN Backbones on PPAR δ dataset (n = 1125) of HSSMS. The table presents the Root Mean Square Error (RMSE) performance of four GNN backbones—GCN, GIN, GAT, and PNA—on both validation and test datasets. The validation set was used to select the best model during training, and the test RMSE and validation RMSE of the selected models are reported. The results are averaged over 10 different random seeds, with each seed having 10 repeated data splits into training, validation, and test sets at a 60%, 20%, 20% ratio on the PPAR δ dataset (n = 1125). Comparisons are made between models without ACA ($a = 0$) and with ACA ($a = 1$). Incorporating ACA consistently improves the performance across all backbones, as reflected by the lower RMSE values. Statistical significance of the improvements is indicated by *: ($*=p<0.05$, $**=p<0.01$, $***=p<0.001$, $****=p<0.0001$).

Backbone	without ACA ($a = 0$)	with ACA ($a = 1$)	signif.	Improvement (%)
Validation RMSE performance				
GCN	0.691 $\pm$ 0.039	0.664$\pm$0.030	*	↓ 3.88%
GIN	0.690 $\pm$ 0.039	0.658$\pm$0.050	*	↓ 4.63%
GAT	0.659 $\pm$ 0.033	0.627$\pm$0.025	**	↓ 4.93%
PNA	0.689 $\pm$ 0.056	0.633$\pm$0.054	***	↓ 8.10%
Avg.	0.682 $\pm$ 0.015	0.645$\pm$0.018	***	↓ 5.39%
Test RMSE performance				
GCN	0.698 $\pm$ 0.053	0.666$\pm$0.041	*	↓ 4.63%
GIN	0.718 $\pm$ 0.055	0.670$\pm$0.047	**	↓ 6.62%
GAT	0.692 $\pm$ 0.046	0.654$\pm$0.041	***	↓ 5.48%
PNA	0.687 $\pm$ 0.039	0.647$\pm$0.030	**	↓ 5.81%
Avg.	0.689 $\pm$ 0.013	0.659$\pm$0.011	***	↓ 5.63%

Supplementary Table 7: Comparison of model performance (Test RMSE of 5FCV) with and without ACA on the LSSNS datasets using MAE and MSE as the regression loss. The table presents the test RMSE results of a study comparing the performance of ACANet models with and without the ACA loss on the nine LSSNS datasets (MGLUR2, BRAF, PLK1, RXFP1, PHGDH, USP7, IDO1, RIP2, and PKCI) using MAE (squared = False) and MSE (squared = True) as the regression loss. For each dataset, the table reports the average test RMSE (five-fold cross-validation with 10 times repeats) and standard deviation of the models without and with ACA loss, along with the corresponding cliff lower and cliff upper values. The statistical significance of the performance improvement of the ACA model is indicated by asterisks (*= $p<0.05$, **= $p<0.01$, ***= $p<0.001$, ****= $p<0.0001$) or "n.s." if not significant. The best loss parameters of cliff lower (cl), cliff upper (cu), and ACA factor alpha (α) are provided for each dataset.

Dataset	Without ACA	With ACA	Signif.	Cliffs (cl, cu)	alpha (α)	Improvement (%)
$\mathcal{L}_{\text{REG}} = \text{MAE (squared = False)}$						
MGLUR2	0.321(± 0.008)	0.297(± 0.006)	****	(0.10, 0.50)	1.0	↓ 7.54%
BRAF	0.377(± 0.027)	0.348(± 0.021)	*	(0.10, 1.00)	1.0	↓ 7.54%
PLK1	0.501(± 0.026)	0.457(± 0.014)	**	(0.45, 0.50)	1.0	↓ 8.88%
RXFP1	0.555(± 0.031)	0.525(± 0.009)	*	(0.15, 0.50)	0.1	↓ 5.37%
PHGDH	0.604(± 0.048)	0.534(± 0.020)	**	(0.10, 1.00)	1.0	↓ 11.57%
USP7	0.576(± 0.023)	0.551(± 0.028)	*	(0.10, 2.50)	0.1	↓ 4.33%
IDO1	0.735(± 0.034)	0.716(± 0.024)	n.s.	(1.08, 4.00)	1.0	↓ 2.65%
RIP2	0.866(± 0.025)	0.752(± 0.041)	****	(0.34, 2.00)	1.0	↓ 13.16%
PKCI	0.840(± 0.061)	0.811(± 0.047)	n.s.	(0.21, 1.00)	1.0	↓ 3.37%
Avg.	0.597(± 0.189)	0.555(± 0.177)	-	-	-	↓ 7.16%
$\mathcal{L}_{\text{REG}} = \text{MSE (squared = True)}$						
MGLUR2	0.332(± 0.019)	0.285(± 0.005)	****	(0.45, 1.50)	0.2	↓ 13.94%
BRAF	0.425(± 0.042)	0.337(± 0.017)	****	(0.10, 1.50)	0.2	↓ 20.61%
PLK1	0.553(± 0.024)	0.453(± 0.017)	****	(0.66, 1.00)	1.0	↓ 18.21%
RXFP1	0.636(± 0.036)	0.542(± 0.010)	****	(0.70, 2.50)	0.1	↓ 14.76%
PHGDH	0.670(± 0.086)	0.558(± 0.049)	**	(1.52, 2.00)	0.5	↓ 16.82%
USP7	0.637(± 0.024)	0.613(± 0.033)	*	(0.10, 3.00)	0.1	↓ 3.73%
IDO1	0.789(± 0.033)	0.689(± 0.016)	****	(0.33, 1.00)	1.0	↓ 12.65%
RIP2	1.099(± 0.237)	0.707(± 0.037)	***	(0.10, 1.00)	1.0	↓ 35.72%
PKCI	1.112(± 0.078)	0.904(± 0.045)	***	(0.70, 2.50)	0.5	↓ 18.65%
Avg.	0.695(± 0.269)	0.565(± 0.192)	-	-	-	↓ 17.23%

Supplementary Table 8: Comparison of model performance with and without ACA on 30 HSSMS datasets. The table presents the RMSE performance of 30 HSSMS datasets during nested cross-validation on the training set (The optimized cliff-awareness factor α values are shown in [Supplementary Figure 5](#)), with MAE as the regression loss. The improvements achieved with ACA are shown in the last column.

Dataset ID	Dataset Name (Target)	Without ACA	With ACA	Improvement
CHEMBL1862_Ki	ABL1	0.805 $\pm$ 0.005	0.756 $\pm$ 0.005	↓ 6.00%
CHEMBL1871_Ki	AR	0.761 $\pm$ 0.009	0.712 $\pm$ 0.012	↓ 6.47%
CHEMBL2034_Ki	GR	0.763 $\pm$ 0.025	0.690 $\pm$ 0.004	↓ 9.62%
CHEMBL2047_EC50	FXR	0.689 $\pm$ 0.008	0.661 $\pm$ 0.002	↓ 4.06%
CHEMBL204_Ki	Thrombin	0.758 $\pm$ 0.004	0.739 $\pm$ 0.002	↓ 2.47%
CHEMBL2147_Ki	PIM1	0.728 $\pm$ 0.008	0.698 $\pm$ 0.004	↓ 4.19%
CHEMBL214_Ki	5-HT1A	0.686 $\pm$ 0.002	0.652 $\pm$ 0.003	↓ 4.90%
CHEMBL218_EC50	CB1	0.755 $\pm$ 0.011	0.727 $\pm$ 0.002	↓ 3.68%
CHEMBL219_Ki	D4R	0.801 $\pm$ 0.011	0.740 $\pm$ 0.003	↓ 7.66%
CHEMBL228_Ki	SERT	0.794 $\pm$ 0.007	0.765 $\pm$ 0.004	↓ 3.64%
CHEMBL231_Ki	HRH1	0.745 $\pm$ 0.009	0.690 $\pm$ 0.004	↓ 7.36%
CHEMBL233_Ki	MOR	0.771 $\pm$ 0.005	0.742 $\pm$ 0.001	↓ 3.76%
CHEMBL234_Ki	D3R	0.704 $\pm$ 0.004	0.669 $\pm$ 0.003	↓ 4.99%
CHEMBL235_EC50	PPAR γ	0.713 $\pm$ 0.004	0.694 $\pm$ 0.002	↓ 2.71%
CHEMBL236_Ki	DOR	0.723 $\pm$ 0.004	0.687 $\pm$ 0.003	↓ 4.91%
CHEMBL237_EC50	KOR (a)	0.931 $\pm$ 0.024	0.886 $\pm$ 0.008	↓ 4.84%
CHEMBL237_Ki	KOR (i)	0.862 $\pm$ 0.003	0.788 $\pm$ 0.002	↓ 8.56%
CHEMBL238_Ki	DAT	0.745 $\pm$ 0.009	0.698 $\pm$ 0.008	↓ 6.37%
CHEMBL239_EC50	PPAR α	0.755 $\pm$ 0.003	0.725 $\pm$ 0.001	↓ 3.91%
CHEMBL244_Ki	FX	0.814 $\pm$ 0.005	0.771 $\pm$ 0.004	↓ 5.19%
CHEMBL262_Ki	GSK3	0.844 $\pm$ 0.007	0.785 $\pm$ 0.006	↓ 6.99%
CHEMBL264_Ki	HRH3	0.682 $\pm$ 0.006	0.651 $\pm$ 0.003	↓ 4.64%
CHEMBL2835_Ki	JAK1	0.496 $\pm$ 0.001	0.485 $\pm$ 0.003	↓ 2.10%
CHEMBL287_Ki	SOR	0.842 $\pm$ 0.007	0.780 $\pm$ 0.006	↓ 7.46%
CHEMBL2971_Ki	JAK2	0.724 $\pm$ 0.022	0.683 $\pm$ 0.001	↓ 5.60%
CHEMBL3979_EC50	PPAR δ	0.710 $\pm$ 0.009	0.657 $\pm$ 0.002	↓ 7.51%
CHEMBL4005_Ki	PIK3CA	0.672 $\pm$ 0.006	0.639 $\pm$ 0.007	↓ 4.80%
CHEMBL4203_Ki	CLK4	0.802 $\pm$ 0.006	0.744 $\pm$ 0.006	↓ 7.33%
CHEMBL4616_EC50	GHSR	0.738 $\pm$ 0.009	0.712 $\pm$ 0.004	↓ 3.45%
CHEMBL4792_Ki	OX2R	0.693 $\pm$ 0.004	0.659 $\pm$ 0.006	↓ 4.84%
Average	-	0.757 $\pm$ 0.085	0.705 $\pm$ 0.069	↓ 6.59%

Supplementary Table 9: Comparison of model performance on 3 MMP datasets of activity cliff classification (Bold represents the best performance). The table compares the performance of GNN-based ACGCN models equipped with and without ACA against fingerprint-based methods including XGBoost, LightGBM, RF, and MLP (DNN).

Dataset	Method	Feature	Evaluation Metrics					
			BA	TPR	TNR	F1-score	MCC	AUC
Thrombin	ACGCN (with ACA)	Duvenaud et al.	0.874	0.857	0.891	0.506	0.510	0.923
	ACGCN (without ACA)	Duvenaud et al.	0.862	0.845	0.879	0.490	0.487	0.906
	DNN	MACCS	0.830	0.758	0.901	0.487	0.477	0.885
		ECFP	0.800	0.706	0.893	0.444	0.425	0.872
	LightGBM	MACCS	0.816	0.736	0.897	0.464	0.449	0.898
		ECFP	0.817	0.736	0.899	0.470	0.454	0.892
	XGBoost	MACCS	0.827	0.794	0.861	0.429	0.426	0.902
		ECFP	0.819	0.765	0.874	0.440	0.431	0.893
	RF	MACCS	0.835	0.805	0.865	0.440	0.439	0.899
		ECFP	0.823	0.773	0.872	0.439	0.432	0.892
Melanocortin Receptor 4	ACGCN (with ACA)	Duvenaud et al.	0.959	0.955	0.963	0.491	0.550	0.978
	ACGCN (without ACA)	Duvenaud et al.	0.899	0.916	0.882	0.365	0.451	0.936
	DNN	MACCS	0.844	0.763	0.925	0.363	0.398	0.917
		ECFP	0.858	0.767	0.948	0.343	0.394	0.935
	LightGBM	MACCS	0.875	0.802	0.948	0.343	0.394	0.935
		ECFP	0.848	0.775	0.922	0.268	0.331	0.917
	XGBoost	MACCS	0.851	0.739	0.964	0.352	0.444	0.920
		ECFP	0.826	0.715	0.936	0.290	0.338	0.890
	RF	MACCS	0.877	0.829	0.926	0.304	0.371	0.915
		ECFP	0.867	0.806	0.928	0.308	0.370	0.918
Mu Opioid Receptor	ACGCN (with ACA)	Duvenaud et al.	0.915	0.909	0.920	0.405	0.462	0.947
	ACGCN (without ACA)	Duvenaud et al.	0.889	0.907	0.871	0.334	0.373	0.929
	DNN	MACCS	0.806	0.668	0.945	0.386	0.385	0.882
		ECFP	0.820	0.711	0.929	0.262	0.315	0.891
	LightGBM	MACCS	0.836	0.763	0.910	0.325	0.366	0.919
		ECFP	0.844	0.822	0.867	0.269	0.327	0.911
	XGBoost	MACCS	0.847	0.794	0.899	0.318	0.365	0.922
		ECFP	0.838	0.764	0.912	0.332	0.372	0.912
	RF	MACCS	0.859	0.850	0.868	0.288	0.347	0.911
		ECFP	0.837	0.783	0.892	0.313	0.357	0.897

Supplementary Table 10: Hyperparameter search space for each model.

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

Supplementary Table 11: Dataset 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(Oc2ccccc(Oc3ccccc3)n2)cc1	75	15
CHEMBL236_Ki	c1ccc(C2CCN(C(c3ccccc3)c3ccccc3)CC2)cc1	47	9
CHEMBL234_Ki	c1cc2c(c(N3CCN(CCC4CCCC4)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(Nc1ccccc1)Nc1ccnc2ccccc12	15	3
CHEMBL2047_EC50	C1CCC2C(C1)CCC1C3CCCC3CCC21	31	10
CHEMBL4616_EC50	c1ccc(COCCc2nnn[nH]2)cc1	31	12
CHEMBL2034_Ki	Cc1ccc(C(=O)N2CCC(CC2)NC(=O)CN2C(=O)C(c3ccccc3)C3CCCN32)cc1	29	12
CHEMBL262_Ki	c1ccc(CNC(NC(=O)c2ccccc[nH]2)C(C)C)cc1	60	15
CHEMBL231_Ki	O=c1c(OC2CCCC2)nc2ccccc2[nH]1	29	11
CHEMBL264_Ki	N#Cc1cn([C@H]2CN(c3ccc(F)cc3)C(=O)O2)nn1	43	11
CHEMBL2835_Ki	Cc1ccc(C(=O)NCCN2CCCC2)cc1	55	15
CHEMBL2971_Ki	c1ccc2c(c1)cnn2C1CCCN1	34	6
CHEMBL237_EC50	Cc1cccc(CNC(=O)C(C)N)cc1	58	14
CHEMBL237_Ki	Cn1cc(C2CCC(N3CCC(C(=O)N4CC=C5CCC(C)C5C4)CC3)CC2)cn1	35	8
CHEMBL4792_Ki	Cn1cc(C2CCC(N3CCC(C(=O)N4CC=C5CCC(C)C5C4)CC3)CC2)cn1	57	13
CHEMBL239_EC50	c1ccc2c(c1)ccc(N1CCC(N2C(=O)NCC3CC3)CC1)c2	68	17
CHEMBL3979_EC50	c1cccc(C(=O)N(C)C)cc1	46	12
CHEMBL235_EC50	Cc1ccc(C(=O)N2CCC[N+](C)(C)CC2)cc1	22	7
CHEMBL4005_Ki	O=C(Nc1ccc(O)cc1)Nc1ccc(CN2CCOCC2)cc1	34	9
CHEMBL2147_Ki	Cc1ccc(CNc2nc3c(c(N)n2)OCC(O)C3O)cc1	36	9
CHEMBL214_Ki	COC(=O)c1ccc(OCN2CCNCC2)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(OCN2CCCC2)cc1	44	10

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
CHEMBL2047_EC50	450	131	76.0%	0.697
CHEMBL4616_EC50	509	118	77.3%	0.664
CHEMBL2034_Ki	543	147	77.6%	0.666
CHEMBL262_Ki	590	198	77.2%	0.640
CHEMBL231_Ki	690	205	71.5%	0.699
CHEMBL264_Ki	2081	552	75.3%	0.709
CHEMBL2835_Ki	438	128	70.8%	0.679
CHEMBL2971_Ki	712	186	78.1%	0.634
CHEMBL237_EC50	696	183	78.6%	0.705
CHEMBL237_Ki	1871	523	78.4%	0.699
CHEMBL233_Ki	2270	621	74.7%	0.703
CHEMBL4792_Ki	1114	239	77.9%	0.680
CHEMBL239_EC50	1231	352	76.2%	0.681
CHEMBL235_EC50	1733	428	76.5%	0.677
CHEMBL3979_EC50	804	231	70.0%	0.706
CHEMBL4005_Ki	728	155	75.1%	0.630
CHEMBL2147_Ki	1034	306	81.6%	0.650
CHEMBL214_Ki	2397	655	74.7%	0.671
CHEMBL228_Ki	1182	386	82.5%	0.640
CHEMBL287_Ki	1049	173	78.0%	0.700
CHEMBL204_Ki	2255	499	75.2%	0.621
CHEMBL1862_Ki	576	218	83.9%	0.531

Supplementary Table 13: Activity cliff definitions across different published studies. Part of the results are from MoleculeACE⁴.

Descriptor	Similarity	Potency difference	Study
MACCS	Tanimoto coeff. of 0.85	2-fold	Bajorath group. <i>Future Med. Chem.</i> 7 , 1565–1579 (2015) ¹⁶ ; <i>F1000Res.</i> 2 (2013) ¹⁷ .
ECFPs	Tanimoto coeff. of 0.55	2-fold	Bajorath group. <i>Future Med. Chem.</i> 7 , 1565–1579 (2015) ¹⁶ ; <i>F1000Res.</i> 2 (2013) ¹⁷ .
MMP	Substructure match	2-fold	Bajorath group. <i>Mol. Inform.</i> 35 , 181–191 (2016) ¹⁸ ; <i>Future Med. Chem.</i> 7 , 1565–1579 (2015) ¹⁶ ; <i>F1000Res.</i> 2 (2013) ¹⁷ ; <i>J. Chem. Inf. Model.</i> 52 , 1138–1145 (2012) ¹⁹ .
MMP	Substructure match	1 or 2-fold	Pérez-Benito <i>et al.</i> , <i>J. Chem. Theory Comput.</i> 15 , 1884–1895 (2019) ²⁰ .
Several 2D and 3D, including MACCS	Several similarity measures between 0.8–0.95	1 or 2-fold	Bajorath group. <i>J. Chem. Inf. Model.</i> 52 , 670–677 (2012) ²¹ .
Maximum common substructure	50% of common substructure atoms	1-fold	Jiménez-Luna. <i>J. Chem. Inf. Model.</i> 62 , 274–283 (2022) ²² .
Physicochemical properties	50% of max SALI score	–	Seebeck <i>et al.</i> , <i>ChemMedChem</i> 6 , 1630–1639 (2011) ²³ .
MACCS, ECFP, TARIS, ROCS, Pharmacophore	SALI score	Variable	Medina-Franco <i>et al.</i> , <i>J. Chem. Inf. Model.</i> 49 , 477–491 (2009) ²⁴ .
AP descriptor	Dice similarity of 0.7	Custom metric	Sheridan <i>et al.</i> , <i>J. Chem. Inf. Model.</i> 60 , 1969–1982 (2020) ²⁵ .
ECFPs, Bemis–Murcko scaffolds, SMILES	Tanimoto coeff. and edit similarity of 0.9	1-fold	Van Tilberg <i>et al.</i> , <i>J. Chem. Inf. Model.</i> 62 , 5938–5951 (2022) ⁴ .
MMP	Hussain and Rea algorithm	2-fold	Junhui Park <i>et al.</i> , <i>J. Chem. Inf. Model.</i> 62 , 2341–2351 (2022) ^{14,26} .

Supplementary Table 14: Comparison of model performance with and without ACA 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 ACA	With ACA	Improvement
CHEMBL1871_Ki	AR	0.756 $\pm$ 0.005	0.719 $\pm$ 0.005	↓ 4.86%
CHEMBL218_EC50	CB1	0.753 $\pm$ 0.006	0.737 $\pm$ 0.004	↓ 2.15%
CHEMBL244_Ki	FX	0.879 $\pm$ 0.006	0.839 $\pm$ 0.004	↓ 4.62%
CHEMBL236_Ki	DOR	0.756 $\pm$ 0.005	0.749 $\pm$ 0.006	↓ 0.96%
CHEMBL234_Ki	D3R	0.741 $\pm$ 0.004	0.707 $\pm$ 0.006	↓ 4.56%
CHEMBL219_Ki	D4R	0.886 $\pm$ 0.005	0.842 $\pm$ 0.006	↓ 4.95%
CHEMBL238_Ki	DAT	0.838 $\pm$ 0.004	0.808 $\pm$ 0.005	↓ 3.55%
CHEMBL4203_Ki	CLK4	0.831 $\pm$ 0.007	0.797 $\pm$ 0.004	↓ 4.08%
CHEMBL2047_EC50	FXR	0.774 $\pm$ 0.005	0.750 $\pm$ 0.005	↓ 3.14%
CHEMBL4616_EC50	GHSR	0.741 $\pm$ 0.005	0.722 $\pm$ 0.005	↓ 2.56%
CHEMBL2034_Ki	GR	0.770 $\pm$ 0.004	0.732 $\pm$ 0.004	↓ 4.96%
CHEMBL262_Ki	GSK3	0.807 $\pm$ 0.007	0.792 $\pm$ 0.005	↓ 1.83%
CHEMBL231_Ki	HRH1	0.630 $\pm$ 0.005	0.606 $\pm$ 0.002	↓ 3.81%
CHEMBL264_Ki	HRH3	0.694 $\pm$ 0.004	0.665 $\pm$ 0.005	↓ 4.24%
CHEMBL2835_Ki	JAK1	0.374 $\pm$ 0.002	0.361 $\pm$ 0.002	↓ 3.46%
CHEMBL2971_Ki	JAK2	0.656 $\pm$ 0.004	0.636 $\pm$ 0.004	↓ 3.07%
CHEMBL237_EC50	KOR (a)	0.966 $\pm$ 0.006	0.922 $\pm$ 0.005	↓ 4.51%
CHEMBL237_Ki	KOR (i)	0.829 $\pm$ 0.006	0.775 $\pm$ 0.005	↓ 6.49%
CHEMBL4792_Ki	OX2R	0.693 $\pm$ 0.005	0.672 $\pm$ 0.004	↓ 3.04%
CHEMBL239_EC50	PPAR(α)	0.825 $\pm$ 0.005	0.781 $\pm$ 0.004	↓ 5.31%
CHEMBL3979_EC50	PPAR(δ)	0.793 $\pm$ 0.006	0.746 $\pm$ 0.006	↓ 5.90%
CHEMBL235_EC50	PPAR(γ)	0.830 $\pm$ 0.005	0.802 $\pm$ 0.006	↓ 3.36%
CHEMBL4005_Ki	PIK3CA	0.646 $\pm$ 0.004	0.611 $\pm$ 0.005	↓ 5.53%
CHEMBL2147_Ki	PIM1	0.602 $\pm$ 0.004	0.589 $\pm$ 0.005	↓ 2.11%
CHEMBL214_Ki	5-HT1A	0.730 $\pm$ 0.005	0.702 $\pm$ 0.005	↓ 3.90%
CHEMBL228_Ki	SERT	1.076 $\pm$ 0.008	0.980 $\pm$ 0.007	↓ 8.93%
CHEMBL287_Ki	SOR	0.723 $\pm$ 0.004	0.776 $\pm$ 0.003	↑ 7.35%
CHEMBL204_Ki	Thrombin	0.778 $\pm$ 0.005	0.744 $\pm$ 0.004	↓ 4.32%
CHEMBL1862_Ki	ABL1	0.943 $\pm$ 0.007	0.917 $\pm$ 0.004	↓ 2.69%
CHEMBL233_Ki	MOR	0.759 $\pm$ 0.004	0.710 $\pm$ 0.004	↓ 6.41%
Average	-	0.769 $\pm$ 0.124	0.740 $\pm$ 0.114	↓ 3.81%

Supplementary references

1. Free, S. M. & Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **7**, 395–399 (1964).
2. Abdi, H. Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdiscip. Rev. Comput. Stat.* **2**, 97–106 (2010).
3. Fralish, Z., Chen, A., Skaluba, P. & Reker, D. DeepDelta: Predicting pharmacokinetic improvements of molecular derivatives with deep learning (2023).
4. Van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
5. O'Dowd, C. R. *et al.* Identification and structure-guided development of pyrimidinone based USP7 inhibitors. *ACS Med. Chem. Lett.* **9**, 238–243 (2018).
6. Haile, P. A. *et al.* Identification of quinoline-based RIP2 kinase inhibitors with an improved therapeutic index to the hERG ion channel. *ACS Med. Chem. Lett.* **9**, 1039–1044 (2018).
7. Kwiatkowski, J. *et al.* Fragment-based drug discovery of potent protein kinase c iota inhibitors. *J. Med. Chem.* **61**, 4386–4396 (2018).
8. Weinstabl, H. *et al.* Intracellular trapping of the selective phosphoglycerate dehydrogenase (PHGDH) inhibitor BI-4924 disrupts serine biosynthesis. *J. Med. Chem.* **62**, 7976–7997 (2019).
9. Zhan, M.-M. *et al.* Design, synthesis, and biological evaluation of novel highly selective polo-like kinase 2 inhibitors based on the tetrahydropteridin chemical scaffold. *Eur. J. Med. Chem.* **143**, 724–731 (2018).
10. Kumar, S. *et al.* Discovery of clinical candidate (1*r*, 4*r*)-4-((*r*)-2-((*s*)-6-fluoro-5*h*-imidazo [5, 1-*a*] isoindol-5-yl)-1-hydroxyethyl) cyclohexan-1-ol (navoximod), a potent and selective inhibitor of indoleamine 2, 3-dioxygenase 1. *J. Med. Chem.* **62**, 6705–6733 (2019).
11. AgoulNIK, A. I., AgoulNIK, I. U., Hu, X. & Marugan, J. Synthetic non-peptide low molecular weight agonists of the relaxin receptor 1. *Br. J. Pharmacol.* **174**, 977–989 (2017).
12. Huang, J.-J., Liao, C.-B. & Chen, P.-N. Quinazoline compounds as kinase inhibitors (2014). US Patent 8,785,459.
13. Takaishi, M. *et al.* Tetrahydroimidazo (1, 5-*d*)[1, 4] oxazepine derivative (2016). US Patent 9,458,176.
14. Park, J., Sung, G., Lee, S., Kang, S. & Park, C. ACGCN: graph convolutional networks for activity cliff prediction between matched molecular pairs. *J. Chem. Inf. Model.* **62**, 2341–2351 (2022).
15. Xiong, Z. *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2019).
16. Stumpfe, D. & Bajorath, J. Monitoring global growth of activity cliff information over time and assessing activity cliff frequencies and distributions. *Future Med. Chem.* **7**, 1565–1579 (2015).
17. Hu, Y., Stumpfe, D. & Bajorath, J. Advancing the activity cliff concept. *FI000Res.* **2**, 199 (2013).
18. Dimova, D. & Bajorath, J. Advances in activity cliff research. *Mol. Inform.* **35**, 181–191 (2016).

19. Hu, X., Hu, Y., Vogt, M., Stumpfe, D. & Bajorath, J. MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* **52**, 1138–1145 (2012).
20. Pérez-Benito, L., Casajuana-Martin, N., Jiménez-Rosés, M., van Vlijmen, H. & Tresadern, G. Predicting activity cliffs with free-energy perturbation. *J. Chem. Theory Comput.* **15**, 1884–1895 (2019).
21. Hu, Y. & Bajorath, J. Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. *J. Chem. Inf. Model.* **52**, 670–677 (2012).
22. Jiménez-Luna, J., Skalic, M. & Weskamp, N. Benchmarking molecular feature attribution methods with activity cliffs. *J. Chem. Inf. Model.* **62**, 274–283 (2022).
23. Seebeck, B., Wagener, M. & Rarey, M. From activity cliffs to target-specific scoring models and pharmacophore hypotheses. *ChemMedChem* **6**, 1630–1639 (2011).
24. Medina-Franco, J. L. *et al.* Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **49**, 477–491 (2009).
25. Sheridan, R. P. *et al.* Experimental error, kurtosis, activity cliffs, and methodology: what limits the predictivity of quantitative structure-activity relationship models? *J. Chem. Inf. Model.* **60**, 1969–1982 (2020).
26. Hussain, J. & Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **50**, 339–348 (2010).