

GRAPHCLIFF: SHORT-LONG RANGE GATING FOR SUBTLE DIFFERENCES BUT CRITICAL CHANGES

Anonymous authors

Paper under double-blind review

ABSTRACT

Quantitative structure–activity relationship assumes a smooth relationship between molecular structure and biological activity. However, activity cliffs defined as pairs of structurally similar compounds with large potency differences break this continuity. Recent benchmarks targeting activity cliffs have revealed that classical machine learning models with extended connectivity fingerprints outperform graph neural networks. Our analysis shows that graph embeddings fail to adequately separate structurally similar molecules in the embedding space, making it difficult to distinguish between structurally similar but functionally different molecules. Despite this limitation, molecular graph structures are inherently expressive and attractive, as they preserve molecular topology. To preserve the structural representation of molecules as graphs, we propose a new model, GraphCliff, which integrates short- and long-range information through a gating mechanism. Experimental results demonstrate that GraphCliff consistently improves performance on both non-cliff and cliff compounds. Furthermore, layer-wise node embedding analyses reveal reduced over-smoothing and enhanced discriminative power relative to strong baseline graph models.

1 INTRODUCTION

Quantitative Structure–Activity Relationship (QSAR) is based on the premise that molecules with similar structures have similar biological activity. QSAR modeling plays a crucial role in drug discovery as it reduces the number of compounds that require experimental testing, thereby saving both cost and time. In particular, QSAR-guided drug discovery enables virtual screening for hit identification, lead optimization, and ADMET (absorption, distribution, metabolism, excretion, and toxicity) evaluation, thus streamlining the experimental workflow (Cherkasov et al., 2014). To support such virtual screening efforts, a wide range of machine learning and deep learning models have recently been developed to directly predict molecular properties and biological activities from molecular structures (Hu et al., 2019; Wang et al., 2022; Heid et al., 2023; Li et al., 2023; Qiao et al., 2025). However, there exists a class of cases that breaks the continuity of the typical structure–activity relationship, known as *activity cliffs*. Unlike the conventional assumption that structurally similar molecules exhibit similar activities, activity cliffs describe cases where minor structural differences lead to large and abrupt changes in activity. They are formally quantified as the ratio of the activity difference between two compounds to their distance in a given chemical space (Maggiora, 2006). In practical terms, activity cliffs are defined as pairs or groups of structurally similar compounds that are active against the same target protein but exhibit large potency differences (Stumpfe et al., 2019). Although analog groups corresponding to activity cliffs may deviate from general QSAR assumptions, they highlight the importance of local structural changes and provide valuable insight into processes such as hit-to-lead optimization and structural alert development (Stumpfe & Bajorath, 2012; Wedlake et al., 2019).

Motivated by activity cliffs’ importance in drug discovery, Van Tilborg et al. (2022) curated the MoleculeACE dataset from ChEMBL (Gaulton et al., 2012) and evaluated a wide range of models. The results revealed that machine learning models with extended connectivity fingerprints (ECFPs) consistently outperform deep learning approaches, with CNNs and LSTMs using SMILES providing moderate success, while transformer and GNNs generally underperformed. The strong performance of ECFPs can be attributed to their design, in which binary bit vectors represent radius-based substructures that are highly sensitive to chemical modifications (Rogers & Hahn, 2010). This represen-

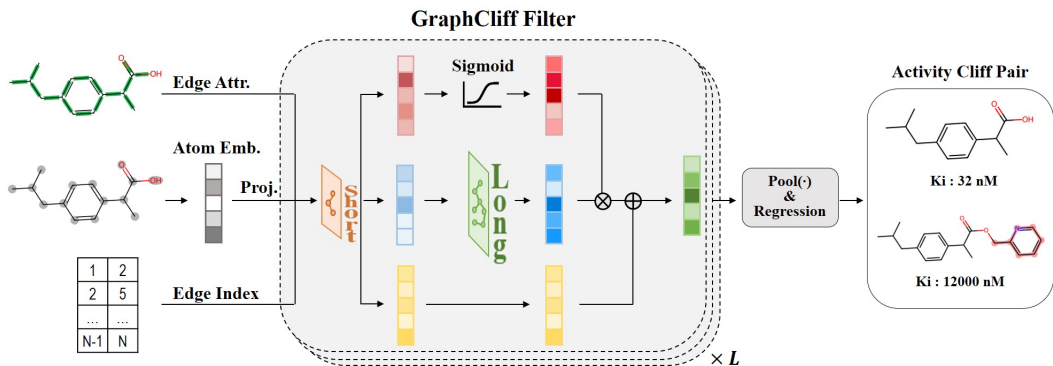


Figure 1: Overall architecture of GraphCliff.

tation introduces a strong inductive bias and low variance, expanding atom-centered neighborhoods within a fixed radius and hashing them into sparse vectors that suppress noise and yield stable encodings. In small-data regimes, such inductive bias allows ECFP-based models to generalize more reliably than flexible deep models. In contrast, GNNs introduce numerous parameters and high modeling flexibility, which increase variance under limited data (Baptista et al., 2022). Moreover, as layers deepen, node embeddings become homogenized due to Laplacian smoothing, leading to the over-smoothing phenomenon where fine-grained local distinctions vanish (Wu et al., 2023). This explains why LSTM and CNN models, which emphasize local structural changes, often perform better than transformer and GNNs. Nevertheless, molecular graphs inherently preserve rich structural information, where atoms are represented as nodes and bonds as edges, with extensions to 3D coordinates, charges, or bond orders directly incorporated (Kearnes et al., 2016). Unlike ECFPs that rely on predefined radius-based hashing, graph representations can adaptively capture complex topological patterns, stereochemistry, and long-range dependencies. The central challenge is therefore to design graph architectures that preserve the expressiveness of molecular graph structures while mitigating over-smoothing and achieving ECFP-level sensitivity to local patterns.

To confirm that GNNs have difficulty preserving the same level of local sensitivity as ECFPs, we performed an analysis based on the MoleculeACE results, comparing the ability of ECFPs and graph embeddings to capture local structural changes within activity cliff pairs. For ECFPs, each molecule in a cliff pair was represented as a 1024-dimensional fingerprint, and the dissimilarity between a pair of molecules was measured as $1 - \text{TanimotoSimilarity}(A, B)$, where A and B denote the ECFPs of the two molecules in the pair. For graph embeddings, we extracted embeddings from graph-based models for each molecule in a cliff pair and calculated the Euclidean distance (Liberti et al., 2014) between them. To ensure a fair comparison, we applied min-max normalization separately to the ECFP dissimilarities and graph embedding Euclidean distances, scaling each to the range $[0, 1]$. Appendix Figure 4 compares ECFP dissimilarities (x-axis) and graph embedding Euclidean distances (y-axis) for activity cliff pairs, with the diagonal line $y = x$ (red) serving as a reference. If the two measures were similar, the points would align closely with this line. However, most points lie below the diagonal for GCN, GAT, and MPNN, indicating that ECFP dissimilarities tend to be larger than the corresponding graph embedding distances. The fitted regression lines (green) further confirm this trend, with slopes below 1, indicating that ECFPs capture larger bit-level differences between cliff pairs. Thus, ECFPs are more sensitive to local structural changes than graph embeddings. The slopes computed for each model across all individual datasets are reported in Appendix Table 2.

These findings highlight a critical limitation of existing GNNs: despite their expressive capacity, they fail to preserve ECFP-level sensitivity to local structural changes. This limitation motivates the need for graph models that can preserve the inherent expressiveness of molecular graph structures while matching the local sensitivity of ECFPs. Therefore, we aimed to create a graph-based model that effectively integrates global context with local structural details. Similar efforts to combine local and global dependencies have also been explored in sequence modeling. StripedHyena2 (Ku et al., 2025) is a multi-hybrid sequence architecture that extends the original Hyena long convolution by introducing short explicit (SE), middle regularized (MR), and long implicit (LI) convolutional

components to jointly capture short-, middle-, and long-range dependencies. While StripedHyena2 operates on 1D token sequences, we adapt the same principle to molecular graphs by combining short-range message passing layers with long-range propagation modules. This design shares the same goal of capturing both short- and long-range information through a gating mechanism that selectively integrates local features with global context. Our contributions are as follows:

- We introduce **GraphCliff**, a novel graph neural architecture that explicitly integrates local structural details and global context through a gating mechanism over short- and long-range representations, with the explicit goal of overcoming the loss of local sensitivity and over-smoothing issues observed in existing GNNs.
- We provide extensive empirical evidence on the benchmark, demonstrating **consistent improvements on both non-cliff and activity cliff compounds**.
- We present a comprehensive analysis which shows that our model mitigates over-smoothing in node representations, yielding **more discriminative representations** than existing GNNs.

2 RELATED WORKS

Contextual dependencies at varying ranges Modeling dependencies across multiple contextual ranges is essential for tasks that require both fine-grained local detail and broad long-range coherence. StripedHyena2 addresses this challenge with a convolution-centric architecture that operates entirely on 1D convolutional modules optimized for sequence modeling. StripedHyena used the transformed Hyena block (Poli et al., 2023), which was converted into short-, middle-, and long-range variants that are sequentially connected to capture information across multiple scales. These variants are implemented as three convolutional operators with distinct receptive fields: Short-Explicit (SE), Medium-Regularized (MR), and Long-Implicit (LI), which are combined into sequential compositions such as SE-MR-LI. Each module is specialized to capture a different scale of interaction. SE focuses on local recall through short explicit filters and has been empirically shown to be particularly effective at capturing short-range dependencies. MR models medium-range interactions using regularized filters. LI aggregates information across the entire sequence via implicit long convolutions. This hierarchical design is particularly advantageous for ultra-long sequence domains such as genomic data, and has been shown to scale to contexts of up to one million tokens. To unify representations obtained at different scales, StripedHyena2 employs a learnable gating mechanism that adaptively balances short-, middle- and long-range features. The gating formulation allows the model to dynamically adjust the contribution of local versus global signals, thereby preserving critical short-range information while maintaining coherence across long-range contexts.

Activity cliff Stumpfe et al. (2019) established two key criteria to enable a systematic and quantitative investigation of activity cliffs. The first criterion concerns the structural similarity between two compounds, while the second considers the magnitude of their potency difference. To define analog groups, the authors employed the concept of Matched Molecular Pairs (MMPs), identifying pairs of molecules with single or multiple substitution sites that exhibit changes in potency (or ΔpKi) greater than 2. In another study, Van Tilborg et al. (2022) formalized the concept of activity cliffs by constructing a curated dataset from ChEMBL specifically designed for activity cliff analysis. Structural similarity was quantified using three complementary measures: (i) substructure similarity, computed as the Tanimoto coefficient on ECFPs to capture shared radial, atom-centered substructures between molecules, thereby reflecting global differences across their entire substructural composition, (ii) scaffold similarity, based on ECFPs computed on molecular scaffolds, to detect compounds differing in their core structures, and (iii) SMILES similarity, measured via Levenshtein distance, to account for character insertions, deletions, and translocations in the string representation of molecules. Activity cliffs were then defined as compound pairs with at least a 10-fold difference in Ki . The benchmark further evaluated a broad range of models, including deep learning approaches such as graph-based methods, Attentive Fingerprint (AFP) (Xiong et al., 2019), Graph Attention Networks (GAT) (Veličković et al., 2017), Graph Convolutional Networks (GCN) (Kipf, 2016), and Message Passing Neural Networks (MPNN) (Gilmer et al., 2017), as well as Convolutional Neural Networks (CNN) (Kimber et al., 2021), Long Short-Term Memory networks (LSTM) (Hochreiter & Schmidhuber, 1997), Multilayer Perceptrons (MLP), and Transformer (Vaswani et al., 2017). In addition, traditional machine learning algorithms were assessed,

including Gradient Boosting Machines (GBM) (Friedman, 2001), k-Nearest Neighbors (KNN) (Fix, 1985), Random Forests (RF) (Predictors, 1996), and Support Vector Machines (SVM) (Cristianini, 2000).

3 METHODS

3.1 DATASETS

We utilized MoleculeACE as benchmark, which comprises curated compound–protein interaction data extracted from ChEMBL. Each dataset contains potency values (Ki) for compounds targeting a specific protein, where low-quality samples were removed during curation based on predefined criteria. In total, the benchmark includes 30 datasets, each corresponding to a distinct protein target, where Ki values serve as regression labels. Each of the 30 datasets in MoleculeACE is associated with a different protein target and can be used independently to assess model generalization across diverse biological contexts. In addition to MoleculeACE, we also employed the benchmark datasets introduced by Group (2023). The Low-Sample Size and Narrow Scaffold (LSSNS) datasets consist of small molecules built around highly conserved scaffolds, with each dataset containing ranging from a few dozen to slightly over one hundred compounds. These data were compiled from fragment-to-lead medicinal chemistry studies and the ChEMBL database. As the LSSNS collection does not provide pre-defined activity cliff annotations, we applied the same criteria used in MoleculeACE to annotate cliff molecules. Structural similarity between compounds was defined using three complementary metrics: substructure similarity, scaffold similarity, and SMILES string similarity. Two compounds were considered structurally similar if at least one of these similarity scores exceeded 0.9. If their Ki values differed by more than the predefined threshold (i.e., at least a 10-fold difference), the pair was regarded as exhibiting a significant potency change. Compound pairs that satisfied both criteria were designated as activity cliffs. In the benchmark, activity cliffs were encoded in a binary manner, indicating whether each compound belongs to a cliff, without explicitly enumerating pairs or groups. Train–test splits were constructed to preserve the overall ratio of activity cliffs across datasets. Additional details are provided in Appendix Tables 3 and 4.

3.2 SHORT- AND LONG-RANGE GATING

We take inspiration from StripedHyena2, which transforms the Hyena block into short-, middle-, and long-range variants that are sequentially connected to capture information across multiple scales. Extending this design principle to molecular graphs, we construct a graph architecture where conventional graph modules are adapted to process short- and long-range dependencies and their outputs are fused through a learnable gating mechanism. An overview of this architecture is illustrated in Figure 1, and the formulation of our model is presented in the following equations.

Atom encoding The model architecture begins by defining the initial node and edge features based on atom types, bond types, and other chemical descriptors. This is followed by an atom encoding stage. Each input node feature $x_i \in \mathbb{R}^{d_{\text{in}}}$ is transformed into a d -dimensional hidden representation via $h_i^{(0)} = \phi_{\text{atom}}(x_i) \in \mathbb{R}^d$, where ϕ_{atom} denotes an MLP followed by normalization and a nonlinear activation function.

GraphCliff filter Adopted from StripedHyena2, in which a single projection yields three separate components (input, gating, and output), our projection layer maps d -dimensional space into a $3d$ -dimensional space to facilitate decomposition into functionally distinct streams. Each filter layer processes short- and long-range information, and their outputs are subsequently integrated via a gating mechanism. The short-range filter adopts a GINE (Xu et al., 2018) message passing operator to capture local neighborhood interactions. In contrast, the long-range filter captures multi-hop dependencies within a single layer using Chebyshev polynomials (Hammond et al., 2011), thereby avoiding the need to stack multiple GNN layers. Recent work demonstrates that Chebyshev polynomials operate directly on the normalized Laplacian, propagating information across multiple hops without explicit edge rewiring or architectural modifications that distort the original topology (Hariri et al., 2025b).

At layer ℓ , given hidden node representations $h^{(\ell)} \in \mathbb{R}^{N \times d}$, we first apply normalization followed by a linear projection:

$$Z = h^{(\ell)} W, \quad Z \in \mathbb{R}^{N \times 3d}, \quad (1)$$

where $W \in \mathbb{R}^{d \times 3d}$ is a trainable projection matrix.

SHORT FILTER The short-range filter applies a GINE message passing operator to the projected features Z :

$$Z' = \text{GINE}(Z, E_{idx}, E_{attr}), \quad (2)$$

where E_{idx} denotes edge index and E_{attr} denotes edge features. The output $Z' \in \mathbb{R}^{N \times 3d}$ is then split along the feature dimension into three parts:

$$Z' = [x_2 \parallel x_1 \parallel v], \quad x_2, x_1, v \in \mathbb{R}^{N \times d}. \quad (3)$$

The GINE operator is defined as:

$$z'_i = \psi \left((1 + \epsilon) z_i + \sum_{j \in \mathcal{N}(i)} (z_j + \phi(e_{ij})) \right), \quad (4)$$

where e_{ij} denotes the edge attribute associated with the directed edge from source node j to target node i , ϕ is an MLP applied to edge attributes, ψ is a node-wise MLP, and ϵ is a learnable scalar parameter.

LONG FILTER To capture global context, we compute Chebyshev polynomials over the normalized adjacency matrix \hat{A} , applied to the short-path feature x_2 :

$$T_0 = x_2, \quad T_1 = \hat{A}x_2, \quad T_k = 2\hat{A}T_{k-1} - T_{k-2} \quad (k \geq 2). \quad (5)$$

The long-range module computes $\text{Long}(x_2) = \sum_{k=0}^K \alpha_k T_k$, where α_k are learnable coefficients.

GATED FUSION We combine short- and long-range information using a sigmoid gating function:

$$g = \sigma(x_1), \quad u = g \odot \text{Long}(x_2) + v, \quad (6)$$

where σ denotes the element-wise sigmoid function and \odot is the element-wise product. Gating mechanisms have been shown to alleviate over-smoothing in GNNs by adaptively regulating information flow (Xin et al., 2020), providing empirical support for our design. Finally, the filter layer output is updated via a residual connection as $h^{(\ell+1)} = u^{(\ell)} + h^{(\ell)}$.

We stack L GraphCliff filters sequentially, where the output of each layer serves as the input to the next:

$$h^{(\ell+1)} = \text{GraphCliffFilter}^{(\ell)}(h^{(\ell)}, E_{idx}, E_{attr}), \quad \ell = 0, \dots, L-1. \quad (7)$$

Pooling and regression Finally, we apply an attention-based graph pooling operation to adaptively select and aggregate informative nodes. Specifically, we employ SAGPool (Lee et al., 2019), and the resulting pooled representation is passed to a regression head to produce the final output corresponding to the target property. Formally, after obtaining the final layer node embeddings $h^{(L)}$, the graph-level representation is constructed as:

$$\hat{y} = \phi_{\text{reg}} \left(\text{SAGPool} \left(h^{(L)} \right) \right), \quad (8)$$

where the graph-level representation is obtained via SAGPool and subsequently passed to the regression MLP ϕ_{reg} .

Table 1: RMSE (\downarrow) and RMSE_{cliff} (\downarrow) values for each algorithm on six ChEMBL targets. The best results are highlighted in **bold**, and the second-best results are underlined.

Algorithm	Descriptor	CHEMBL1871 (Ki)	CHEMBL204 (Ki)	CHEMBL2147 (Ki)	CHEMBL228 (Ki)	CHEMBL239 (EC50)	CHEMBL244 (Ki)
GraphCliff	GRAPH	0.628 / 0.797	0.691 / 0.821	0.560 / 0.579	0.651 / 0.674	0.670 / 0.792	0.668 / 0.752
SVM	ECFP	0.665 / 0.873	0.723 / 0.859	0.576 / 0.580	0.662 / 0.676	0.678 / 0.819	0.715 / 0.797
Chemprop	GRAPH	0.704 / 0.919	0.811 / 0.851	0.649 / 0.639	0.670 / 0.695	0.819 / 0.827	0.726 / 0.797
MLP	ECFP	0.737 / 0.958	0.815 / 0.962	0.723 / 0.704	0.755 / 0.757	0.756 / 0.901	0.796 / 0.850
SCAGE (w/o 3D)	GRAPH	0.758 / 0.823	0.814 / 0.888	0.910 / 0.858	0.771 / 0.786	0.796 / 0.851	0.892 / 0.972
LSTM	SMILES	0.662 / 0.850	0.822 / 0.930	0.647 / 0.725	0.779 / 0.884	0.765 / 0.905	0.800 / 0.913
Transformer	TOKENS	0.809 / 1.073	1.098 / 1.255	0.903 / 0.893	0.908 / 0.979	0.910 / 1.032	1.078 / 1.071
GCN	GRAPH	0.769 / 1.009	1.056 / 1.201	0.840 / 0.825	0.958 / 1.000	0.906 / 1.024	1.075 / 1.060
CNN	SMILES	0.810 / 1.041	1.131 / 1.234	0.925 / 0.934	0.965 / 0.944	0.910 / 0.986	1.095 / 1.071
GAT	GRAPH	0.798 / 1.042	1.138 / 1.281	0.966 / 0.917	1.026 / 1.028	0.902 / 1.012	1.088 / 1.117
MPNN	GRAPH	1.058 / 1.154	1.458 / 1.581	1.025 / 0.934	1.000 / 1.015	1.288 / 1.481	1.660 / 1.557
MolCLR _{gen}	GRAPH	0.948 / 0.863	1.592 / 1.616	1.551 / 1.545	1.340 / 1.317	0.968 / 1.025	1.831 / 1.837
AFP	GRAPH	1.143 / 1.274	1.553 / 1.743	1.906 / 1.368	1.192 / 1.160	1.361 / 1.573	1.706 / 1.591
MolCLR _{gin}	GRAPH	1.077 / 1.020	1.689 / 1.709	1.226 / 1.015	1.459 / 1.364	1.054 / 1.112	1.849 / 1.837
Contextpred	GRAPH	1.647 / 1.687	2.012 / 2.269	1.295 / 1.857	1.663 / 1.803	1.893 / 2.168	1.922 / 2.056
KPGT	GRAPH	1.976 / 1.822	2.302 / 2.210	1.676 / 1.201	1.856 / 1.847	2.751 / 2.660	2.126 / 2.103

4 RESULTS

We evaluated our method on all 30 benchmark datasets provided by MoleculeACE. As baselines, we included the machine learning and deep learning models reported in the original MoleculeACE study: graph-based models (AFP, GAT, GCN, MPNN), SMILES-based models (CNN, LSTM, Transformer), and ECFP-based models (MLP, GBM, KNN, RF, SVM). We also incorporated additional models known for their strong performance in molecular property prediction tasks. **ContextPred** (Hu et al., 2019) is a pretraining method that learns to predict masked subgraphs using contextual information, thereby enhancing structural awareness. **MolCLR** (Wang et al., 2022) applies contrastive learning to molecular graphs, encouraging structurally similar molecules to be mapped closer in the learned embedding space. **Chemprop** (Heid et al., 2023) is based on a message passing neural network (MPNN) architecture that incorporates directed edge information (D-MPNN). **KPGT** (Li et al., 2023) is a knowledge-guided pretraining method designed to integrate domain-specific chemical insights into the representation learning process. **SCAGE** (Qiao et al., 2025) is a self-conformation-aware graph transformer that incorporates 3D geometric information and functional group tagging through multitask pretraining. We excluded the 3D atom-distances for fair comparison with our 2D graph setting. We used two evaluation metrics: root mean squared error (RMSE) and RMSE_{cliff}. RMSE is computed over all molecules in the test set and measures the overall accuracy of the predicted pKi values. In contrast, RMSE_{cliff} is calculated specifically on compounds identified as activity cliffs, thereby quantifying the prediction error on these particularly challenging and structure-sensitive samples.

We present results for six datasets in Table 1, while the complete results across all 30 benchmark datasets are provided in Appendix Tables 5, 6, and 7 due to space limitations. Across these results, GraphCliff achieved the best overall performance, with particularly large improvements over other graph-based models in both general prediction tasks and activity cliff scenarios. Among graph-based models, **Chemprop** achieved the strongest performance. Its advantage can be attributed to its D-MPNN architecture, which incorporates bond directionality into message passing. This design enables the model to distinguish chemically distinct but structurally similar motifs, such as C=O versus O=C. Following closely, **SCAGE (w/o 3D)** also delivered competitive results. The presence of explicit functional group annotations guides the model to attend to chemically meaningful substructures. In contrast, **MolCLR** exhibited relatively weaker performance. While its contrastive learning objective promotes generalization by aligning embeddings of structurally similar molecules, it primarily captures global molecular similarity. This global bias may limit MolCLR’s sensitivity to functionally important substructures, thereby contributing to its higher RMSE and RMSE_{cliff} scores. **ContextPred**, based on a GIN (Xu et al., 2018) backbone, showed limited performance because it failed to capture contextual substructure information without pretraining. Performance improved when context-based pretraining was applied followed by fine-tuning. However, the learned structural context alone was still insufficient to fully address the challenges posed by activity-cliff compounds. Finally, **KPGT**, a knowledge-guided pretraining method that incorporates domain-specific chemical information such as pharmacophore patterns and functional groups, showed modest per-

formance compared to other models. This suggests that, despite being chemically informed, KPGT struggles to capture the subtle structure–activity discontinuities characteristic of activity cliff compounds. Overall, while several baselines demonstrated competitive performance, our results indicate that GraphCliff achieves consistently strong performance relative to prior approaches. This underscores the effectiveness of explicitly balancing local substructural sensitivity with global molecular context in addressing both general prediction tasks and activity cliff scenarios.

We also evaluated our approach on the nine datasets in LSSNS benchmark. As shown in Appendix Tables 8 and 9, GraphCliff does not uniformly dominate across all LSSNS protein targets. This outcome is expected, given that LSSNS was deliberately designed under low-data conditions. Each dataset is composed of only a few dozen to slightly over one hundred molecules and all built around narrow and highly conserved scaffolds. In such conditions, training a high-capacity graph neural network from scratch often leads to overfitting, unstable optimization, and limited generalization. To address this limitation, we investigated whether knowledge transfer from related, larger-scale datasets could provide more stable initialization. For each protein target in LSSNS, we identified a biologically similar target in the MoleculeACE. The results of this transfer-initialization strategy are reported in Tables 8 and 9. Across most protein targets, Transferred GraphCliff substantially reduced both RMSE and RMSE_{cliff} compared to training from scratch. For instance, in the PKC ϵ and mGluR2 tasks, transferred models achieved notable gains in predictive accuracy. While performance was not uniformly improved for every target due to imperfect biological similarity, the overall trend clearly demonstrates that leveraging prior knowledge from related MoleculeACE datasets mitigates the difficulties of learning in data-scarce scenarios. These findings suggest that transfer learning is a practically useful strategy for extending the applicability to real-world settings, where data availability is often limited. The mapping from each LSSNS target to its MoleculeACE counterpart is provided in the Appendix Table 10.

5 ANALYSIS

5.1 ABLATION STUDIES

We conducted ablation studies to assess the individual contributions of the short-range filter, long-range filter, and gating mechanism in our architecture. As shown in Appendix Table 11, removing the short-range filter caused the most substantial performance drop, underscoring its critical role in the model. The short-range filter captures essential one-hop message-passing information and serves multiple functions: feeding into the long-range filter, providing input to the gating mechanism, and contributing to the final sum fusion. These pathways ensure that localized chemical information is effectively preserved and propagated throughout the network. Removing the long-range filter also degraded performance, though to a lesser extent. Because it captures broader structural context up to three hops, its absence restricts the model’s ability to integrate global molecular features. The gating mechanism, while having the smallest standalone effect, still made a positive contribution through the adaptive combination of short- and long-range information. This mechanism proved more effective than naive feature summation and enhanced the model’s ability to balance local and global information. We evaluated different graph pooling strategies and found that other methods (mean, sum, max) exacerbate over-smoothing by uniformly aggregating indistinguishable node embeddings. To overcome this, we adopted SAGPool, which adaptively selects informative nodes based on learned importance scores. Empirically, SAGPool outperformed basic pooling methods, resulting in lower RMSE and RMSE_{cliff}, confirming that adaptive node selection helps preserve both local and global information.

As shown in Appendix Table 12, we investigated the effect of using different GNN architectures in the short- and long-range filters. Our default configuration employs **GINE** for the short-range filter and **Chebyshev polynomials** for the long-range filter. We replace either component with GCN, GAT, or GIN. GINE, which incorporates edge features into the message-passing process, is particularly beneficial for molecular graphs, as edge attributes such as bond type, aromaticity, and stereochemistry encode important chemical information. In contrast, GCN and GIN do not explicitly utilize bond features, limiting their expressiveness. For long-range propagation, Chebyshev polynomials outperformed stacked GNNs such as GIN and GAT. This improvement is likely due to its use of spectral polynomials, which efficiently encode multi-hop neighborhood information within a single layer. These results underscore the importance of selecting GNNs that are structurally aligned

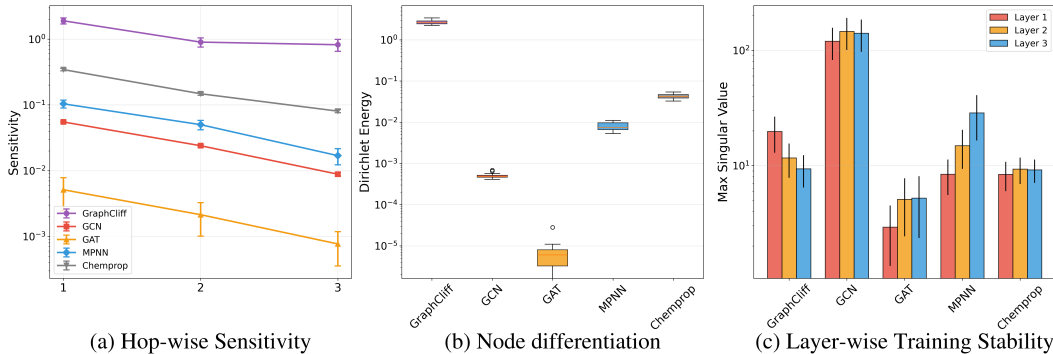


Figure 2: Comprehensive analysis of propagation dynamics and stability across models. (a) Hop-wise sensitivity, where higher values indicate stronger long-range information flow. (b) Dirichlet Energy measuring node differentiation, where higher values reflect better resistance to over-smoothing. (c) Layer-wise Jacobian singular values assessing gradient flow stability, where moderate values indicate robust propagation.

with the type of information local or global being modeled. Moreover, they demonstrate that efficient compression of global context is particularly advantageous for representing complex molecular graphs.

5.2 ANALYSIS OF OVER-SMOOTHING MITIGATION IN GRAPHCLIFF

Following the methodology of Hariri et al. (2025a), we assessed long-range information propagation by quantifying how perturbations to node u influence the output of a distant node v across k -hop neighborhoods. Specifically, we computed the sensitivity as $\|f(\mathbf{x}_{\text{perturbed}}) - f(\mathbf{x}_{\text{original}})\|/\epsilon$, where ϵ denotes the perturbation magnitude. This hop-wise sensitivity measures how local feature changes propagate through the graph to affect distant representations, thereby linking localized perturbations to global structural responses. In line with prior findings on spectral GNNs, Figure 2a illustrates that baseline GNNs such as GCN and GAT exhibited severe sensitivity decay, approaching values close to zero beyond 2-3 hops. In contrast, GraphCliff maintained stable and substantial sensitivity across multiple hops. This shows that GraphCliff effectively integrates local perturbations with global propagation, thereby capturing both fine-grained variations and long-range structural dependencies within molecular graphs.

To assess over-smoothing behavior, we adopted Dirichlet Energy analysis as recommended by Rusch et al. (2023), defining $E = \sum_{(i,j) \in \mathcal{E}} \|h_i - h_j\|^2$, where h_i denotes the embedding of node i , and the summation runs over all edges (i, j) in the graph. This measure reflects how well neighboring nodes remain distinguishable. In Figure 2b, higher values indicate preserved separability, whereas exponential decay toward zero signals progressive over-smoothing, with embeddings collapsing to near-identical representations. Such collapse restricts the model’s ability to balance local distinctiveness with global coherence. Our results reveal clear differences across models. GraphCliff maintains the highest Dirichlet Energy (average 2.6862), whereas traditional GNNs suffer from strong over-smoothing, with GAT showing severe degradation (6.1×10^{-6}), GCN and MPNN exhibiting moderate smoothing (0.0004, 0.008), and Chemprop achieving somewhat higher values (0.0430) but still falling far short of GraphCliff. These findings confirm that GraphCliff effectively alleviates the exponential convergence phenomenon highlighted in prior work and achieves a balanced integration of local and global information by preserving critical representational diversity.

We further examined gradient flow stability through the singular values of the Jacobian matrix $\mathbf{J} = \frac{\partial \text{output}}{\partial \text{input}}$ at each layer, following Hariri et al. (2025a). Stable propagation is associated with singular values close to one, while values approaching zero indicate vanishing gradients and excessively large values suggest gradient explosion. As shown in Figure 2c, GraphCliff exhibits consistent stability across all three layers, with maximum singular values of 19.75, 11.67, and 9.37, well within a moderate range that avoids both vanishing and exploding behaviors. In contrast, baseline models display problematic patterns, with GCN showing uncontrolled growth at deeper layers, GAT main-

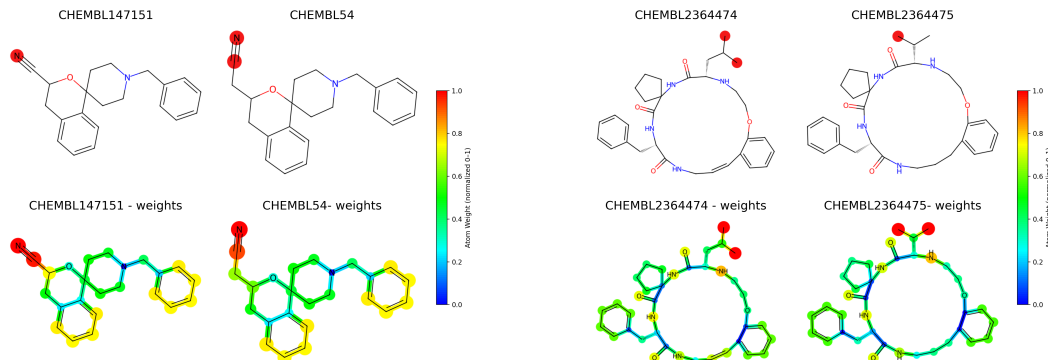


Figure 3: Visualization of a cliff pair with large functional divergence. Top: atoms responsible for the activity cliff (highlighted in red). Bottom: attention weights from the sigmoid gating vector $\sigma(x_1)$, with warmer colors indicating higher importance.

taining low values suggestive of limited expressiveness, MPNN suffering from high inter-layer variability, and Chemprop remaining relatively stable. Together with hop-wise sensitivity and Dirichlet Energy analysis, these results demonstrate that GraphCliff not only preserves molecular structure but also achieves stable propagation and training dynamics, enabling a balanced integration of local variations and global dependencies within molecular graphs.

5.3 QUALITATIVE ANALYSIS

To qualitatively assess whether our model identifies functionally relevant substructures, we visualized atom-level importance scores derived from the gating vector $\sigma(x_1) \in \mathbb{R}^{N \times d}$, where the sigmoid function assigns attention weights to each node. Specifically, we investigated whether atoms with high gating values align with those responsible for activity cliffs. Figure 3 shows two representative activity cliff pairs, where the top row highlights the difference atoms (shown in red) between the two compounds in each pair, and the bottom row visualizes atom importance scores obtained from the sigmoid-gated vector. The importance values are normalized between 0 and 1, and colored accordingly. We observe that atoms with the highest attention weights (indicated by warmer colors such as red and orange) are frequently aligned with the structural differences responsible for activity cliffs. This suggests that the gating mechanism successfully highlights substructures that are functionally discriminative, rather than relying solely on global molecular context. These results provide qualitative evidence that the gating path captures meaningful local information and contributes to the model’s robustness in handling activity cliff compounds.

6 CONCLUSION

In this work, we introduced GraphCliff, a novel graph neural architecture designed to address two key limitations of existing GNNs: the loss of local sensitivity and the tendency toward over-smoothing. By explicitly integrating local structural details with global context through a gating mechanism over short- and long-range representations, GraphCliff provides a more balanced and chemically meaningful representation of molecules. Our extensive evaluation on the MoleculeACE benchmark demonstrated that GraphCliff consistently achieves improved performance across both non-cliff and activity cliff compounds, highlighting its robustness in challenging prediction settings. Furthermore, our in-depth analysis confirmed that GraphCliff effectively alleviates node over-smoothing, yielding more discriminative representations than conventional GNNs. Taken together, these findings suggest that explicitly combining local and global information is a promising direction for molecular graph representation. Future research may build on this framework by further incorporating chemically informed descriptors, such as fingerprint-derived substructures, to bridge the gap between domain knowledge and learned graph representations.

7 LLM USAGE

Large language models (LLMs) were used in a limited assistive role during the preparation of this paper. LLMs were employed for grammar checking, rephrasing, and improving clarity of sentences. Some sentences were rephrased with the help of LLMs to improve readability, without altering the technical content. LLMs were occasionally used to identify relevant related work and papers, which were subsequently verified and selected by the authors.

REFERENCES

- Delora Baptista, João Correia, Bruno Pereira, and Miguel Rocha. Evaluating molecular representations in machine learning models for drug response prediction and interpretability. *Journal of Integrative Bioinformatics*, 19(3):20220006, 2022.
- Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010, 2014.
- N Cristianini. Shawe taylor j. *An Introduction to Support Vector Machines and Other Kernel based Learning Methods*, pp. 93–112, 2000.
- Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*, volume 1. USAF school of Aviation Medicine, 1985.
- Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.
- Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.
- Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. Pmlr, 2017.
- BIDD Group. Mpcd: Molecular property cliff dataset. <https://github.com/bidd-group/MPCD>, 2023. Accessed: 2025-09-25.
- David K Hammond, Pierre Vandergheynst, and Rémi Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.
- Ali Hariri, Álvaro Arroyo, Alessio Gravina, Moshe Eliasof, Carola-Bibiane Schönlieb, Davide Bacciu, Kamyar Azizzadenesheli, Xiaowen Dong, and Pierre Vandergheynst. Return of chebnet: Understanding and improving an overlooked gnn on long range tasks. *arXiv preprint arXiv:2506.07624*, 2025a.
- Ali Hariri, Álvaro Arroyo, Alessio Gravina, Moshe Eliasof, Carola-Bibiane Schönlieb, Davide Bacciu, Kamyar Azizzadenesheli, Xiaowen Dong, and Pierre Vandergheynst. Return of chebnet: Understanding and improving an overlooked gnn on long range tasks. *arXiv preprint arXiv:2506.07624*, 2025b.
- Esther Heid, Kevin P Greenman, Yunsie Chung, Shih-Cheng Li, David E Graff, Florence H Vermeire, Haoyang Wu, William H Green, and Charles J McGill. Chemprop: a machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1): 9–17, 2023.
- Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.
- Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

- Steven Kearnes, Kevin McCloskey, Marc Berndl, Vijay Pande, and Patrick Riley. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8): 595–608, 2016.
- Talia B Kimber, Maxime Gagnebin, and Andrea Volkamer. Maxsmi: maximizing molecular property prediction performance with confidence estimation using smiles augmentation and deep learning. *Artificial Intelligence in the Life Sciences*, 1:100014, 2021.
- TN Kipf. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- Jerome Ku, Eric Nguyen, David W Romero, Garyk Bixi, Brandon Yang, Anton Vorontsov, Ali Taghibakhshi, Amy X Lu, Dave P Burke, Greg Brockman, et al. Systems and algorithms for convolutional multi-hybrid language models at scale. *arXiv preprint arXiv:2503.01868*, 2025.
- Junhyun Lee, Inyeop Lee, and Jaewoo Kang. Self-attention graph pooling. In *International conference on machine learning*, pp. 3734–3743. pmlr, 2019.
- Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1):7568, 2023.
- Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications. *SIAM review*, 56(1):3–69, 2014.
- Gerald M Maggiora. On outliers and activity cliffs why qsar often disappoints, 2006.
- Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pp. 28043–28078. PMLR, 2023.
- L B Bagging Predictors. Bagging predictors. *Machine learning*, 24:123–140, 1996.
- Jianbo Qiao, Junru Jin, Ding Wang, Saisai Teng, Junyu Zhang, Xuetong Yang, Yuhang Liu, Yu Wang, Lizhen Cui, Quan Zou, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. *Nature Communications*, 16(1): 4382, 2025.
- David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
- T Konstantin Rusch, Michael M Bronstein, and Siddhartha Mishra. A survey on oversmoothing in graph neural networks. *arXiv preprint arXiv:2303.10993*, 2023.
- Dagmar Stumpfe and Jurgen Bajorath. Exploring activity cliffs in medicinal chemistry: miniperpective. *Journal of medicinal chemistry*, 55(7):2932–2942, 2012.
- Dagmar Stumpfe, Huabin Hu, and Jurgen Bajorath. Evolving concept of activity cliffs. *ACS omega*, 4(11):14360–14368, 2019.
- UniProt. D-3-phosphoglycerate dehydrogenase (phgdh). <https://www.uniprot.org/uniprotkb/O43175>, a. Accessed: 2025-09-22.
- UniProt. Receptor-interacting serine/threonine-protein kinase 2 (ripk2). <https://www.uniprot.org/uniprotkb/O43353>, b. Accessed: 2025-09-22.
- UniProt. Thrombin (f2). <https://www.uniprot.org/uniprotkb/P00734>, c. Accessed: 2025-09-22.
- UniProt. Coagulation factor x (f10). <https://www.uniprot.org/uniprotkb/P00742>, d. Accessed: 2025-09-22.
- UniProt. 5-hydroxytryptamine receptor 1a (htr1a). <https://www.uniprot.org/uniprotkb/P08908>, e. Accessed: 2025-09-22.

- UniProt. Serine/threonine-protein kinase pim1. <https://www.uniprot.org/uniprotkb/P11309>, f. Accessed: 2025-09-22.
- UniProt. Indoleamine 2,3-dioxygenase 1 (ido1). <https://www.uniprot.org/uniprotkb/P14902>, g. Accessed: 2025-09-22.
- UniProt. Serine/threonine-protein kinase b-raf (braf). <https://www.uniprot.org/uniprotkb/P15056>, h. Accessed: 2025-09-22.
- UniProt. Dopamine d4 receptor (drd4). <https://www.uniprot.org/uniprotkb/P21917>, i. Accessed: 2025-09-22.
- UniProt. Histamine h1 receptor (hrh1). <https://www.uniprot.org/uniprotkb/P35367>, j. Accessed: 2025-09-22.
- UniProt. Dopamine d3 receptor (drd3). <https://www.uniprot.org/uniprotkb/P35462>, k. Accessed: 2025-09-22.
- UniProt. Protein kinase c iota type. <https://www.uniprot.org/uniprotkb/P41743>, l. Accessed: 2025-09-22.
- UniProt. Glycogen synthase kinase-3 beta (gsk3b). <https://www.uniprot.org/uniprotkb/P49841>, m. Accessed: 2025-09-22.
- UniProt. Polo-like kinase 1 (plk1). <https://www.uniprot.org/uniprotkb/P53350>, n. Accessed: 2025-09-22.
- UniProt. Metabotropic glutamate receptor 2 (grm2). <https://www.uniprot.org/uniprotkb/Q14416>, o. Accessed: 2025-09-22.
- UniProt. Ubiquitin carboxyl-terminal hydrolase 7 (usp7). <https://www.uniprot.org/uniprotkb/Q93009>, p. Accessed: 2025-09-22.
- UniProt. Relaxin family peptide receptor 1 (rxfp1). <https://www.uniprot.org/uniprotkb/Q9HBX9>, q. Accessed: 2025-09-22.
- Derek Van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *Journal of chemical information and modeling*, 62(23): 5938–5951, 2022.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.
- Andrew J Wedlake, Maria Folia, Sam Piechota, Timothy EH Allen, Jonathan M Goodman, Steve Gutsell, and Paul J Russell. Structural alerts and random forest models in a consensus approach for receptor binding molecular initiating events. *Chemical Research in Toxicology*, 33(2):388–401, 2019.
- Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36:35084–35106, 2023.
- Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M Jose. Graph highway networks. *arXiv preprint arXiv:2004.04635*, 2020.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

A APPENDIX

A.1 GRAPH EMBEDDING VS. ECFPs DISTANCE ANALYSIS

Figure 4 illustrates the relationship between ECFP fingerprint dissimilarities (x-axis) and graph embedding Euclidean distances (y-axis) across different GNN architectures. If the two measures were aligned, points would concentrate along the diagonal $y = x$, but conventional GNNs (GCN, GAT, and MPNN) generally underestimate distances, placing most points below the diagonal. GraphCliff achieves a distribution closer to the reference line, indicating that its embeddings more faithfully reflect structural differences captured by ECFPs. The regression slopes reported in Table 2 quantify this trend, confirming that GraphCliff narrows the gap between graph- and fingerprint-based representations.

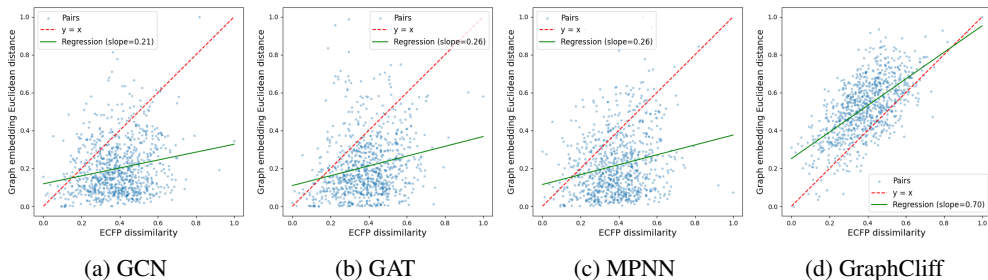


Figure 4: Comparison of graph embedding Euclidean distances with ECFP fingerprint dissimilarities across different models.

Table 2: Slopes of fitted regression lines comparing ECFP dissimilarities (x-axis) and graph embedding Euclidean distances (y-axis) for activity-cliff pairs across different GNN models.

Dataset	GCN	GAT	MPNN	GraphCliff
CHEMBL1862_Ki	0.121	0.508	0.718	0.564
CHEMBL1871_Ki	-0.282	0.259	0.261	0.733
CHEMBL2034_Ki	-0.055	0.388	0.533	0.433
CHEMBL204_Ki	0.2	0.188	0.138	0.684
CHEMBL2047_EC50	-0.047	0.401	0.543	0.541
CHEMBL214_Ki	0.158	0.095	0.328	0.686
CHEMBL2147_Ki	0.535	0.398	0.65	0.821
CHEMBL218_EC50	-0.028	-0.095	0.429	0.572
CHEMBL219_Ki	0.051	0.23	0.318	0.742
CHEMBL228_Ki	0.234	0.156	0.489	0.635
CHEMBL231_Ki	0.207	0.431	0.389	0.672
CHEMBL233_Ki	0.144	0.44	0.366	0.628
CHEMBL234_Ki	0.034	0.001	0.138	0.75
CHEMBL235_EC50	0.129	0.127	0.18	0.733
CHEMBL236_Ki	0.029	-0.037	0.073	0.55
CHEMBL237_EC50	0.222	0.144	0.226	0.67
CHEMBL237_Ki	0.153	0.305	0.383	0.577
CHEMBL238_Ki	0.02	0.056	0.302	0.556
CHEMBL239_EC50	-0.071	-0.077	-0.018	0.527
CHEMBL244_Ki	0.209	0.084	0.152	0.701
CHEMBL262_Ki	0.513	-0.135	-0.301	0.473
CHEMBL264_Ki	0.365	0.241	0.379	0.655
CHEMBL2835_Ki	0.557	0.172	0.398	0.924
CHEMBL287_Ki	0.255	-0.01	0.02	0.643
CHEMBL2971_Ki	0.627	-0.062	0.216	0.555
CHEMBL3979_EC50	0.12	0.549	0.549	0.477
CHEMBL4005_Ki	0.237	0.134	0.307	0.549
CHEMBL4203_Ki	0.459	0.084	0.417	0.708
CHEMBL4616_EC50	0.008	-0.065	-0.027	0.667
CHEMBL4792_Ki	0.047	-0.073	0.123	0.573

A.2 DATASET METADATA

Tables 3 and 4 summarize the datasets used in our experiments. Table 3 provides general statistics of the MoleculeACE benchmark datasets, including the number of compounds, activity labels, and target proteins. Table 4 reports metadata of the LSSNS datasets, which consist of small sample sizes with narrow scaffold diversity.

Table 3: Statistics of MoleculeACE datasets corresponding to the ChEMBL targets used in this study.

Dataset	ChEMBL ID	Type	Target name	Receptor Class	Train compounds (Train cliff)	Test compounds (Test cliff)	Total compounds (cliff)
CHEMBL1862_Ki	CHEMBL1862	Ki	Tyrosine-protein kinase ABL1	Kinase	633 (202)	161 (51)	794 (253)
CHEMBL1871_Ki	CHEMBL1871	Ki	Androgen Receptor	NR	525 (126)	134 (31)	659 (157)
CHEMBL2034_Ki	CHEMBL2034	Ki	Glucocorticoid receptor	NR	598 (183)	152 (47)	750 (230)
CHEMBL2047_EC50	CHEMBL2047	EC50	Farnesoid X receptor	NR	503 (195)	128 (50)	631 (245)
CHEMBL204_Ki	CHEMBL204	Ki	Thrombin	Protease	2201 (790)	553 (199)	2754 (989)
CHEMBL2147_Ki	CHEMBL2147	Ki	Serine/threonine-protein kinase PIM1	Kinase	1162 (387)	294 (98)	1456 (485)
CHEMBL214_Ki	CHEMBL214	Ki	Serotonin 1a receptor	GPCR	2651 (917)	666 (230)	3317 (1147)
CHEMBL218_EC50	CHEMBL218	EC50	Cannabinoid receptor 1	GPCR	823 (292)	208 (75)	1031 (367)
CHEMBL219_Ki	CHEMBL219	Ki	Dopamine D4 receptor	GPCR	1485 (572)	374 (143)	1859 (715)
CHEMBL228_Ki	CHEMBL228	Ki	Serotonin transporter	Other	1362 (479)	342 (120)	1704 (599)
CHEMBL231_Ki	CHEMBL231	Ki	Histamine H1 receptor	GPCR	776 (178)	197 (46)	973 (224)
CHEMBL233_Ki	CHEMBL233	Ki	u-opioid receptor	GPCR	2512 (889)	630 (222)	3142 (1111)
CHEMBL234_Ki	CHEMBL234	Ki	Dopamine D3 receptor	GPCR	2923 (1150)	734 (291)	3657 (1441)
CHEMBL235_EC50	CHEMBL235	EC50	PPAR gamma	NR	1879 (703)	470 (178)	2349 (881)
CHEMBL236_Ki	CHEMBL236	Ki	Delta opioid receptor	GPCR	2077 (772)	521 (193)	2598 (965)
CHEMBL237_EC50	CHEMBL237	EC50	Kappa opioid receptor	GPCR	762 (319)	193 (81)	955 (400)
CHEMBL237_Ki	CHEMBL237	Ki	Kappa opioid receptor	GPCR	2081 (753)	521 (188)	2602 (941)
CHEMBL238_Ki	CHEMBL238	Ki	Dopamine transporter	Other	839 (209)	213 (54)	1052 (263)
CHEMBL239_EC50	CHEMBL239	EC50	PPAR alpha	NR	1377 (568)	344 (141)	1721 (709)
CHEMBL244_Ki	CHEMBL244	Ki	Coagulation factor X	Protease	2476 (1080)	621 (270)	3097 (1350)
CHEMBL262_Ki	CHEMBL262	Ki	GSK-3 beta	Kinase	683 (127)	173 (31)	856 (158)
CHEMBL264_Ki	CHEMBL264	Ki	Histamine H3 receptor	GPCR	2288 (865)	574 (219)	2862 (1084)
CHEMBL2835_Ki	CHEMBL2835	Ki	Janus kinase 1	Kinase	489 (36)	126 (10)	615 (46)
CHEMBL287_Ki	CHEMBL287	Ki	Sigma opioid receptor	Other	1061 (371)	267 (93)	1328 (464)
CHEMBL2971_Ki	CHEMBL2971	Ki	Janus kinase 2	Kinase	779 (95)	197 (25)	976 (120)
CHEMBL3979_EC50	CHEMBL3979	EC50	PPAR delta	NR	900 (373)	225 (94)	1125 (467)
CHEMBL4005_Ki	CHEMBL4005	Ki	PI3K p110-alpha subunit	Transferase	767 (281)	193 (70)	960 (351)
CHEMBL4203_Ki	CHEMBL4203	Ki	CLK4	Kinase	582 (51)	149 (13)	731 (64)
CHEMBL4616_EC50	CHEMBL4616	EC50	Ghrelin receptor	GPCR	543 (262)	139 (68)	682 (330)
CHEMBL4792_Ki	CHEMBL4792	Ki	Orexin receptor 2	GPCR	1174 (610)	297 (153)	1471 (763)

843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875

Table 4: Statistics of LSSNS datasets corresponding to the protein targets used in this study.

Dataset	ChEMBL ID	Type	Target name	Receptor Class	Train compounds (Train cliff)	Test compounds (Test cliff)	Total compounds (cliff)
USP7	CHEMBL4251701	–	Ubiquitin carboxyl-terminal hydrolase 7	Protease	36 (19)	9 (5)	45 (24)
RIP2	CHEMBL4266012; CHEMBL4130524	–	Serine/threonine-protein kinase RIPK2	Kinase	36 (16)	10 (4)	46 (20)
PKC ι	CHEMBL4184321	–	Protein kinase C iota	Kinase	38 (12)	10 (3)	48 (15)
PHGDH	CHEMBL4373702	–	D-3-phosphoglycerate dehydrogenase	Other Enzyme	40 (13)	11 (3)	51 (16)
PLK1	CHEMBL4406868; CHEMBL4138231	–	Serine/threonine-protein kinase PLK1	Kinase	58 (22)	15 (6)	73 (28)
IDO1	CHEMBL4364294	–	Indoleamine 2,3-dioxygenase	Other Enzyme	62 (34)	16 (9)	78 (43)
RXFP1	CHEMBL3714716	–	Relaxin receptor 1	GPCR	93 (52)	24 (14)	117 (66)
BRAF	CHEMBL3638563	–	Serine/threonine-protein kinase B-raf	Kinase	102 (27)	26 (7)	128 (34)
mGluR2	CHEMBL3886984	–	Metabotropic glutamate receptor 2	GPCR	195 (92)	49 (23)	244 (115)

A.3 ALL RESULTS OF 30 DATASETS IN MOLECULEACE

Tables 5, 6, and 7 report the complete results across all 30 protein–ligand datasets included in MoleculeACE. Each table includes the performance of baseline machine learning models, graph-based neural networks, and our proposed GraphCliff. The results include both RMSE and $\text{RMSE}_{\text{cliff}}$, allowing a direct comparison of overall predictive accuracy and sensitivity to activity cliffs.

930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962Table 5: RMSE / RMSE_{cliff} (Part 1/3) with best (**bold**) and second-best (underlined) highlighted.

Algorithm	Descriptor	CHEMBL1862 (Ki)	CHEMBL1871 (Ki)	CHEMBL2034 (Ki)	CHEMBL2047 (EC50)	CHEMBL204 (Ki)	CHEMBL2147 (Ki)	CHEMBL214 (Ki)	CHEMBL218 (EC50)	CHEMBL219 (Ki)	CHEMBL228 (Ki)
GraphCliff	GRAPH	0.781 / 0.674	0.628 / 0.797	0.747 / 0.858	0.599 / 0.602	0.696 / 0.833	0.560 / 0.579	0.621 / 0.726	0.697 / 0.781	0.664 / 0.743	0.651 / 0.674
SVM	ECFP	<u>0.774 / 0.674</u>	0.665 / 0.873	0.674 / 0.813	0.614 / 0.687	<u>0.723 / 0.859</u>	<u>0.576 / 0.580</u>	<u>0.634 / 0.724</u>	0.719 / 0.761	0.710 / 0.788	<u>0.662 / 0.676</u>
GBM	ECFP	0.798 / 0.747	0.678 / 0.922	0.767 / 0.864	<u>0.602 / 0.640</u>	0.753 / 0.932	0.581 / 0.616	0.678 / 0.761	0.712 / 0.746	0.712 / <u>0.765</u>	0.686 / 0.723
RF	ECFP	0.805 / 0.686	0.660 / 0.906	0.727 / 0.852	0.628 / 0.665	0.763 / 0.899	0.662 / 0.676	0.700 / 0.801	0.706 / 0.760	0.723 / 0.765	0.709 / 0.773
Chemprop	GRAPH	0.815 / 0.693	0.704 / 0.919	0.775 / 0.886	0.693 / 0.720	<u>0.811 / 0.851</u>	0.649 / 0.639	0.660 / 0.825	0.758 / 0.794	<u>0.692 / 0.774</u>	0.670 / 0.695
KNN	ECFP	0.898 / 0.822	<u>0.650 / 0.817</u>	0.696 / 0.913	0.642 / 0.735	0.821 / 0.995	0.662 / 0.682	0.733 / 0.874	0.735 / 0.785	0.775 / 0.816	0.717 / 0.811
GBM	MACCS	0.860 / 0.778	<u>0.686 / 0.904</u>	0.723 / 0.857	0.670 / 0.702	0.800 / 0.958	0.791 / 0.777	0.730 / 0.857	0.715 / 0.766	0.816 / 0.862	0.770 / 0.807
RF	MACCS	0.874 / 0.844	0.703 / 0.904	0.700 / 0.828	0.676 / 0.707	0.815 / 0.971	0.806 / 0.793	0.747 / 0.887	0.669 / 0.712	0.807 / 0.838	0.769 / 0.810
MLP	ECFP	0.878 / 0.781	0.737 / 0.958	0.742 / 0.849	0.677 / 0.728	0.815 / 0.962	0.723 / 0.704	0.693 / 0.780	0.785 / 0.806	0.756 / 0.832	0.755 / 0.757
SVM	MACCS	0.890 / 0.859	0.670 / 0.891	<u>0.681 / 0.860</u>	0.695 / 0.736	0.806 / 0.983	0.794 / 0.806	0.746 / 0.889	0.703 / <u>0.729</u>	0.820 / 0.884	0.745 / 0.792
SCAGE (w/o 3D)	GRAPH	0.875 / <u>0.677</u>	0.758 / 0.823	<u>0.745 / 0.825</u>	0.714 / 0.605	0.814 / 0.888	0.910 / 0.858	0.779 / 0.834	0.738 / 0.783	0.876 / 0.863	0.771 / 0.786
LSTM	SMILES	0.761 / 0.793	0.662 / 0.850	0.755 / 0.944	0.696 / 0.790	0.822 / 0.930	0.647 / 0.725	0.723 / 0.851	0.748 / 0.818	0.780 / 0.855	0.779 / 0.884
KNN	MACCS	1.049 / 0.868	0.705 / 0.941	0.768 / 0.910	0.706 / 0.689	0.891 / 1.043	0.944 / 0.867	0.799 / 0.902	0.707 / 0.756	0.873 / 0.888	0.778 / 0.836
RF	PHYSCHEM	0.909 / 0.939	0.753 / 1.058	0.705 / 0.835	0.814 / 0.900	1.105 / 1.161	0.972 / 0.916	0.909 / 0.989	0.823 / 0.800	0.930 / 0.970	0.921 / 0.917
GBM	PHYSCHEM	0.943 / 0.903	0.724 / 1.048	0.729 / 0.867	0.813 / 0.910	1.145 / 1.211	0.949 / 0.894	0.895 / 0.988	0.829 / 0.826	0.950 / 0.980	0.907 / 0.923
KNN	PHYSCHEM	1.001 / 0.960	0.797 / 1.064	0.761 / 0.901	0.766 / 0.845	1.115 / 1.169	1.053 / 0.923	0.971 / 1.049	0.883 / 0.832	0.970 / 0.985	0.989 / 0.934
Transformer	TOKENS	0.961 / 0.964	0.809 / 1.073	0.804 / 0.938	0.767 / 0.725	1.098 / 1.255	0.903 / 0.893	0.862 / 0.952	0.869 / 0.895	0.878 / 0.940	0.908 / 0.979
SVM	PHYSCHEM	0.966 / 0.920	0.803 / 1.066	0.763 / 0.910	0.715 / 0.667	1.205 / 1.313	0.966 / 0.907	0.957 / 0.965	1.026 / 1.005	0.960 / 1.009	0.930 / 0.891
GCN	GRAPH	0.942 / 0.942	0.769 / 1.009	0.810 / 0.928	0.797 / 0.781	1.056 / 1.201	0.840 / 0.825	1.007 / 1.084	0.928 / 0.952	1.026 / 1.055	0.958 / 1.000
CNN	SMILES	1.049 / 0.900	0.810 / 1.041	0.798 / 0.933	0.776 / 0.774	1.131 / 1.234	0.925 / 0.934	0.931 / 1.007	0.958 / 0.934	0.977 / 0.973	0.965 / 0.944
GAT	GRAPH	0.987 / 1.001	0.798 / 1.042	0.809 / 0.941	0.840 / 0.790	1.138 / 1.281	0.966 / 0.917	1.051 / 1.137	0.957 / 0.983	0.979 / 0.982	1.026 / 1.028
RF	WHIM	0.864 / 0.886	0.882 / 1.104	0.782 / 0.877	0.833 / 0.767	1.258 / 1.360	1.042 / 0.996	1.026 / 1.080	0.884 / 0.894	0.989 / 1.022	1.050 / 1.041
GBM	WHIM	0.873 / 0.850	0.903 / 1.104	0.770 / 0.855	0.853 / 0.732	1.266 / 1.381	1.003 / 0.961	1.048 / 1.089	0.891 / 0.916	1.033 / 1.075	1.079 / 1.020
KNN	WHIM	0.986 / 0.849	0.881 / 1.069	0.805 / 0.929	0.841 / 0.833	1.365 / 1.482	1.141 / 1.013	1.069 / 1.112	0.921 / 0.919	1.061 / 1.086	1.076 / 1.017
SVM	WHIM	1.028 / 0.966	0.857 / 1.088	0.842 / 0.964	0.851 / 0.825	1.311 / 1.436	1.103 / 1.002	1.060 / 1.067	0.891 / 0.925	1.041 / 1.074	1.095 / 1.036
MPNN	GRAPH	0.948 / 0.888	1.058 / 1.154	0.905 / 0.927	1.030 / 0.951	1.458 / 1.581	1.025 / 0.934	1.183 / 1.243	1.053 / 1.062	0.903 / 0.919	1.000 / 1.015
MolCLR _{gcn}	GRAPH	0.990 / 1.110	0.948 / 0.863	1.303 / 1.315	0.781 / 0.757	1.592 / 1.616	1.551 / 1.545	1.009 / 1.039	1.100 / 1.106	0.987 / 0.943	1.340 / 1.317
MolCLR _{gcn} ^{pretrained}	GRAPH	1.068 / 1.194	0.925 / 0.832	1.292 / 1.305	0.767 / 0.739	1.557 / 1.579	1.905 / 2.063	1.050 / 1.076	1.072 / 1.118	0.984 / 0.940	1.340 / 1.329
MolCLR _{gin} ^{pretrained}	GRAPH	1.051 / 1.176	1.077 / 1.020	1.510 / 1.550	0.784 / 0.764	1.689 / 1.709	1.226 / 1.015	1.144 / 1.183	1.100 / 1.106	1.022 / 0.991	1.459 / 1.364
AFP	GRAPH	1.347 / 1.158	1.143 / 1.274	0.931 / 0.949	0.970 / 0.902	1.553 / 1.743	1.906 / 1.368	1.083 / 1.134	1.046 / 1.040	0.952 / 0.966	1.192 / 1.160
MolCLR _{gin}	GRAPH	1.051 / 1.176	1.077 / 1.020	1.454 / 1.492	0.784 / 0.764	1.689 / 1.709	1.226 / 1.015	1.144 / 1.183	1.070 / 1.059	1.022 / 0.991	1.459 / 1.364
Contextpred ^{pretrained}	GRAPH	1.351 / 1.813	1.714 / 1.766	1.415 / 1.412	1.654 / 2.007	1.957 / 2.202	1.266 / 1.872	1.501 / 1.593	1.726 / 1.947	1.661 / 1.695	1.494 / 1.657
Contextpred	GRAPH	1.371 / 1.817	1.647 / 1.687	1.475 / 1.453	1.740 / 2.061	2.012 / 2.269	1.295 / 1.857	1.680 / 1.768	1.583 / 1.778	1.686 / 1.742	1.663 / 1.803
KPGT	GRAPH	1.668 / 1.465	1.976 / 1.822	1.420 / 1.594	2.837 / 2.694	2.302 / 2.210	1.676 / 1.201	1.802 / 1.780	2.198 / 2.099	2.121 / 2.099	1.856 / 1.847
KPGT ^{pretrained}	GRAPH	1.627 / 1.396	1.995 / 1.885	1.416 / 1.549	2.848 / 2.757	2.288 / 2.182	1.615 / 1.130	1.781 / 1.777	2.330 / 2.305	1.963 / 1.974	1.830 / 1.847

995
994
993
992
991
990
989
988
987
986
985
984
983
982
981
980
979
978
977
976
975
974
973
972
971
970
969
968
967
966
965
964
963

Table 6: RMSE / RMSE_{cliff} (Part 2/3) with best (**bold**) and second-best (underlined) highlighted.

Algorithm	Descriptor	CHEMBL231 (Ki)	CHEMBL233 (Ki)	CHEMBL234 (Ki)	CHEMBL235 (EC50)	CHEMBL236 (Ki)	CHEMBL237 (EC50)	CHEMBL237 (Ki)	CHEMBL238 (Ki)	CHEMBL239 (EC50)	CHEMBL244 (Ki)
GraphCliff	GRAPH	0.708 / 0.866	<u>0.786</u> / 0.880	0.618 / 0.632	0.650 / 0.754	<u>0.697</u> / 0.785	0.708 / 0.767	<u>0.705</u> / 0.783	0.597 / 0.729	0.670 / 0.792	0.668 / 0.752
SVM	ECFP	0.750 / 0.932	0.774 / 0.859	<u>0.622</u> / <u>0.632</u>	<u>0.640</u> / 0.774	<u>0.698</u> / 0.798	<u>0.720</u> / <u>0.782</u>	0.677 / 0.735	<u>0.610</u> / 0.681	<u>0.678</u> / <u>0.819</u>	<u>0.715</u> / <u>0.797</u>
GBM	ECFP	0.774 / 0.955	0.802 / 0.886	0.629 / 0.649	0.663 / 0.806	0.696 / 0.794	0.803 / 0.880	0.713 / 0.789	0.618 / 0.638	0.682 / 0.821	0.736 / 0.821
RF	ECFP	0.822 / 0.907	0.801 / 0.880	0.660 / 0.683	0.638 / <u>0.764</u>	<u>0.711</u> / <u>0.792</u>	0.762 / 0.793	0.729 / 0.799	0.625 / <u>0.656</u>	0.689 / 0.825	0.741 / 0.823
Chemprop	GRAPH	0.760 / 0.817	0.799 / <u>0.844</u>	0.687 / 0.653	0.708 / 0.799	0.799 / 0.883	0.767 / 0.839	0.743 / 0.800	0.673 / 0.736	0.819 / 0.827	0.726 / 0.797
KNN	ECFP	0.776 / 1.020	0.816 / 0.912	0.675 / 0.699	0.688 / 0.795	0.746 / 0.865	0.810 / 0.876	0.736 / 0.851	0.647 / 0.735	0.714 / 0.865	0.767 / 0.874
GBM	MACCS	0.754 / <u>0.759</u>	0.846 / 0.913	0.719 / 0.724	0.713 / 0.860	0.839 / 0.931	0.831 / 0.904	0.799 / 0.886	0.678 / 0.700	0.721 / 0.841	0.799 / 0.879
RF	MACCS	<u>0.739</u> / 0.828	0.828 / 0.906	0.744 / 0.750	0.705 / 0.823	0.828 / 0.957	0.849 / 0.911	0.814 / 0.927	0.702 / 0.708	0.735 / 0.868	0.846 / 0.917
MLP	ECFP	1.334 / 1.272	0.845 / 0.916	0.669 / 0.676	0.718 / 0.818	0.733 / 0.810	0.902 / 0.950	0.722 / <u>0.765</u>	0.684 / 0.732	0.756 / 0.901	0.796 / 0.850
SVM	MACCS	0.783 / 0.837	0.868 / 0.953	0.739 / 0.729	0.696 / 0.838	0.850 / 0.938	0.832 / 0.885	0.779 / 0.873	0.669 / 0.682	0.718 / 0.853	0.818 / 0.870
SCAGE (w/o 3D)	GRAPH	0.914 / 0.784	0.824 / 0.800	0.799 / 0.811	0.694 / 0.788	0.799 / 0.929	0.967 / 0.960	0.762 / 0.822	0.726 / 0.681	0.796 / 0.851	0.892 / 0.972
LSTM	SMILES	0.809 / 1.070	0.850 / 0.942	0.738 / 0.797	0.727 / 0.847	0.812 / 0.905	0.783 / 0.903	0.774 / 0.862	0.654 / 0.793	0.765 / 0.905	0.800 / 0.913
KNN	MACCS	0.837 / 0.959	0.852 / 0.911	0.782 / 0.758	0.750 / 0.868	0.896 / 1.046	0.962 / 0.950	0.825 / 0.943	0.708 / 0.735	0.791 / 0.895	0.902 / 0.947
RF	PHYSCHEM	0.908 / 0.732	1.019 / 0.996	0.887 / 0.859	0.795 / 0.892	0.989 / 0.998	0.966 / 0.893	0.961 / 0.980	0.903 / 0.851	0.884 / 1.032	1.116 / 1.119
GBM	PHYSCHEM	0.953 / 0.827	1.050 / 1.037	0.883 / 0.846	0.821 / 0.911	1.014 / 1.016	0.999 / 0.942	0.961 / 0.969	0.902 / 0.836	0.874 / 1.023	1.105 / 1.125
KNN	PHYSCHEM	1.045 / 0.900	1.037 / 1.030	0.911 / 0.898	0.874 / 0.938	1.032 / 1.028	0.998 / 0.930	0.969 / 0.955	0.939 / 0.853	0.874 / 0.991	1.138 / 1.142
Transformer	TOKENS	0.964 / 1.028	1.072 / 1.118	0.863 / 0.848	0.801 / 0.914	1.024 / 1.102	1.126 / 1.184	0.996 / 1.062	0.880 / 0.852	0.910 / 1.032	1.078 / 1.071
SVM	PHYSCHEM	0.994 / 0.892	1.159 / 1.130	0.944 / 0.889	0.913 / 1.006	1.115 / 1.223	1.085 / 1.071	1.020 / 1.044	1.001 / 0.955	0.964 / 1.055	1.160 / 1.116
GCN	GRAPH	0.878 / 0.797	1.056 / 1.106	0.934 / 0.919	0.901 / 1.039	0.942 / 1.000	1.132 / 1.094	1.112 / 1.152	0.937 / 0.925	0.906 / 1.024	1.075 / 1.060
CNN	SMILES	1.008 / 1.044	1.073 / 1.080	0.898 / 0.881	0.893 / 0.962	1.018 / 1.067	1.061 / 1.022	1.040 / 1.040	0.917 / 0.897	0.910 / 0.986	1.095 / 1.071
GAT	GRAPH	0.991 / 0.970	1.066 / 1.099	0.950 / 0.912	0.869 / 1.012	1.002 / 1.100	1.103 / 1.064	1.085 / 1.098	0.928 / 0.946	0.902 / 1.012	1.088 / 1.117
RF	WHIM	0.953 / 0.939	1.132 / 1.148	0.969 / 0.902	1.004 / 1.102	1.118 / 1.163	1.302 / 1.314	1.120 / 1.140	0.994 / 0.963	0.998 / 1.078	1.249 / 1.207
GBM	WHIM	0.959 / 0.962	1.147 / 1.158	0.999 / 0.912	1.000 / 1.101	1.132 / 1.179	1.357 / 1.350	1.125 / 1.129	0.991 / 0.949	1.048 / 1.133	1.287 / 1.230
KNN	WHIM	1.007 / 0.909	1.195 / 1.216	1.017 / 0.944	0.978 / 1.079	1.150 / 1.214	1.319 / 1.319	1.174 / 1.187	1.047 / 1.012	1.019 / 1.100	1.274 / 1.217
SVM	WHIM	0.956 / 0.900	1.150 / 1.191	0.987 / 0.922	0.992 / 1.082	1.139 / 1.200	1.307 / 1.299	1.175 / 1.222	0.973 / 0.990	1.016 / 1.135	1.305 / 1.277
MPNN	GRAPH	1.305 / 1.223	1.074 / 1.138	0.959 / 0.922	1.058 / 1.194	1.364 / 1.454	1.402 / 1.334	1.053 / 1.109	1.142 / 1.208	1.288 / 1.481	1.660 / 1.557
MolCLR _{gcn}	GRAPH	1.232 / 1.240	1.253 / 1.284	1.278 / 1.292	1.045 / 1.072	1.356 / 1.341	1.111 / 1.140	1.381 / 1.412	1.157 / 1.293	0.968 / 1.025	1.831 / 1.837
MolCLR _{gcn} ^{pretrained}	GRAPH	1.395 / 1.420	1.229 / 1.256	1.326 / 1.335	1.028 / 1.045	1.315 / 1.296	1.209 / 1.252	1.338 / 1.365	1.168 / 1.307	0.944 / 0.997	1.897 / 1.930
MolCLR _{gin} ^{pretrained}	GRAPH	1.643 / 1.659	1.299 / 1.360	1.347 / 1.357	1.313 / 1.391	1.414 / 1.382	1.089 / 1.103	1.345 / 1.390	1.198 / 1.336	1.054 / 1.112	1.849 / 1.837
AFP	GRAPH	1.262 / 1.156	1.211 / 1.233	0.885 / 0.864	1.202 / 1.308	1.370 / 1.423	1.361 / 1.304	1.310 / 1.411	1.216 / 1.225	1.361 / 1.573	1.706 / 1.591
MolCLR _{gin}	GRAPH	1.643 / 1.659	1.314 / 1.330	1.347 / 1.357	1.313 / 1.391	1.414 / 1.382	1.089 / 1.103	1.437 / 1.483	1.198 / 1.336	1.054 / 1.112	1.849 / 1.837
Contextpred ^{pretrained}	GRAPH	1.975 / 2.135	1.811 / 1.908	1.349 / 1.489	1.615 / 1.862	1.483 / 1.713	1.665 / 1.766	1.564 / 1.710	1.609 / 2.037	1.836 / 2.100	1.894 / 2.011
Contextpred	GRAPH	1.882 / 1.998	1.816 / 1.903	1.467 / 1.598	1.781 / 1.989	1.664 / 1.906	1.707 / 1.818	1.677 / 1.827	1.666 / 2.056	1.893 / 2.168	1.922 / 2.056
KPGT	GRAPH	2.605 / 2.622	1.670 / 1.656	1.676 / 1.708	2.670 / 2.636	2.154 / 2.057	1.691 / 1.762	2.158 / 2.094	2.389 / 2.240	2.751 / 2.660	2.126 / 2.103
KPGT ^{pretrained}	GRAPH	2.277 / 2.423	1.856 / 1.824	1.656 / 1.665	2.672 / 2.670	2.133 / 2.020	1.789 / 1.823	1.854 / 1.836	2.436 / 2.329	2.753 / 2.686	2.067 / 2.023

Table 7: RMSE / RMSE_{cliff} with best (**bold**) and second-best (underlined) highlighted.

Algorithm	Descriptor	CHEMBL262 (Ki)	CHEMBL264 (Ki)	CHEMBL2835 (Ki)	CHEMBL287 (Ki)	CHEMBL2971 (Ki)	CHEMBL3979 (EC50)	CHEMBL4005 (Ki)	CHEMBL4203 (Ki)	CHEMBL4616 (EC50)	CHEMBL4792 (Ki)
GraphCliff	GRAPH	0.752 / 0.702	0.619 / 0.671	0.396 / 0.795	0.706 / 0.798	0.615 / 0.778	0.623 / 0.654	0.617 / 0.712	0.900 / 1.177	0.634 / 0.719	0.635 / 0.651
SVM	ECFP	<u>0.724</u> / 0.656	0.615 / 0.674	0.420 / 0.743	0.714 / 0.812	0.605 / 0.659	<u>0.629</u> / <u>0.674</u>	<u>0.646</u> / 0.742	0.880 / 1.001	<u>0.635</u> / <u>0.692</u>	0.633 / 0.638
GBM	ECFP	0.750 / 0.727	0.649 / 0.722	0.405 / 0.789	0.759 / 0.847	0.616 / 0.667	0.660 / 0.722	0.647 / 0.748	0.919 / 1.075	0.686 / 0.768	0.674 / 0.687
RF	ECFP	0.721 / 0.775	0.659 / 0.742	0.388 / 0.802	0.776 / 0.891	0.630 / 0.643	0.650 / 0.708	0.648 / <u>0.732</u>	<u>0.882</u> / 1.081	0.682 / 0.770	0.709 / 0.721
Chemprop	GRAPH	0.868 / 1.028	0.637 / 0.652	0.433 / 0.762	<u>0.709</u> / 0.715	0.745 / 0.953	0.711 / 0.770	0.709 / 0.808	1.003 / 1.484	0.704 / 0.795	0.675 / 0.713
KNN	ECFP	0.834 / 0.899	0.674 / 0.805	0.436 / 0.858	0.810 / 0.957	0.663 / 0.782	0.684 / 0.739	0.656 / 0.754	0.972 / 1.074	0.740 / 0.828	0.695 / 0.724
GBM	MACCS	0.809 / 0.878	0.696 / 0.790	0.481 / 0.926	0.788 / 0.836	0.658 / <u>0.646</u>	0.661 / 0.703	0.676 / 0.790	0.984 / 1.424	0.715 / 0.795	0.756 / 0.793
RF	MACCS	0.885 / 0.884	0.738 / 0.831	0.437 / 0.824	0.789 / 0.833	0.637 / 0.656	0.696 / 0.727	0.701 / 0.845	0.929 / 1.327	0.717 / 0.772	0.787 / 0.827
MLP	ECFP	0.904 / 0.948	0.672 / 0.731	0.488 / 0.876	0.733 / 0.852	0.674 / 0.764	0.661 / 0.724	0.680 / 0.769	0.947 / 1.027	0.727 / 0.778	0.691 / 0.682
SVM	MACCS	0.834 / 0.959	0.720 / 0.813	0.464 / 0.765	0.738 / 0.789	0.657 / 0.699	0.673 / 0.715	0.723 / 0.844	0.982 / 1.467	0.717 / 0.780	0.749 / 0.780
SCAGE (w/o 3D)	GRAPH	0.923 / 0.845	0.705 / 0.738	0.505 / 0.694	0.801 / 0.813	0.745 / 0.726	0.921 / 0.894	0.705 / 0.770	1.024 / <u>1.016</u>	0.740 / 0.686	0.781 / 0.859
LSTM	SMILES	0.767 / 0.781	0.665 / 0.767	0.431 / 0.840	0.791 / 0.894	0.689 / 0.886	0.740 / 0.790	0.764 / 0.900	0.907 / 1.318	0.739 / 0.831	0.691 / 0.750
KNN	MACCS	0.917 / 1.129	0.770 / 0.891	0.467 / 0.882	0.842 / 0.924	0.732 / 0.671	0.707 / 0.749	0.766 / 0.873	1.031 / 1.502	0.711 / 0.782	0.863 / 0.886
RF	PHYSCHEM	0.863 / 0.865	0.851 / 0.886	0.502 / 0.891	0.784 / <u>0.787</u>	0.815 / 0.797	0.873 / 0.819	0.800 / 0.903	1.002 / 1.455	0.816 / 0.840	0.844 / 0.826
GBM	PHYSCHEM	0.875 / 0.912	0.878 / 0.916	0.539 / 0.905	0.806 / 0.815	0.859 / 0.888	0.867 / 0.800	0.782 / 0.863	1.014 / 1.538	0.828 / 0.852	0.846 / 0.822
KNN	PHYSCHEM	0.906 / 0.936	0.879 / 0.900	0.491 / 0.824	0.849 / 0.838	0.777 / 0.752	0.955 / 0.843	0.820 / 0.841	0.983 / 1.326	0.838 / 0.857	0.910 / 0.856
Transformer	TOKENS	0.976 / 1.052	0.822 / 0.882	0.485 / 0.772	0.869 / 0.927	0.826 / 0.954	0.834 / 0.880	0.855 / 0.941	0.959 / 1.145	0.784 / 0.817	0.912 / 0.911
SVM	PHYSCHEM	0.949 / 1.002	0.909 / 0.917	0.413 / 0.640	0.818 / 0.809	0.950 / 0.855	0.919 / 0.884	0.799 / 0.852	1.005 / 1.221	0.841 / 0.826	0.870 / 0.870
GCN	GRAPH	0.934 / 1.004	0.855 / 0.910	0.505 / 0.926	0.886 / 0.900	0.781 / 0.917	0.812 / 0.805	0.875 / 0.909	0.975 / 1.211	0.867 / 0.831	0.923 / 0.926
CNN	SMILES	0.948 / 0.953	0.890 / 0.915	0.560 / 0.871	0.891 / 0.921	0.831 / 0.861	0.907 / 0.859	0.838 / 0.928	1.013 / 1.231	0.819 / 0.821	0.967 / 0.966
GAT	GRAPH	0.994 / 1.032	0.896 / 0.936	0.555 / 0.924	0.947 / 0.994	0.803 / 0.966	0.923 / 0.914	0.861 / 0.901	1.004 / 1.208	0.873 / 0.835	1.004 / 1.014
RF	WHIM	0.929 / 1.023	0.936 / 0.970	0.478 / 0.744	0.917 / 1.024	0.750 / 0.826	0.996 / 0.938	0.885 / 0.905	0.997 / 1.110	0.912 / 0.888	1.040 / 1.022
GBM	WHIM	0.936 / 1.026	0.970 / 1.021	0.503 / 0.791	0.941 / 1.032	0.781 / 0.821	1.037 / 0.975	0.882 / 0.940	1.034 / 1.252	0.951 / 0.904	1.040 / 1.034
KNN	WHIM	0.913 / 0.868	0.974 / 1.009	0.534 / 0.901	0.974 / 1.073	0.810 / 0.831	1.019 / 0.978	0.929 / 0.997	1.070 / 1.305	0.885 / 0.846	1.052 / 1.020
SVM	WHIM	0.898 / 0.993	0.974 / 1.010	0.512 / 0.803	0.946 / 1.043	0.867 / 0.975	1.045 / 1.020	0.901 / 0.940	1.025 / 1.169	0.910 / 0.900	1.091 / 1.097
MPNN	GRAPH	1.021 / 1.036	1.082 / 1.012	0.668 / 1.067	0.927 / 0.973	0.973 / 0.945	1.183 / 1.145	0.998 / 1.016	1.056 / 1.149	0.935 / 0.860	1.122 / 1.114
MolCLR _{gcn}	GRAPH	1.211 / 1.107	1.081 / 1.025	1.178 / 0.983	0.882 / 0.862	1.789 / 1.789	1.066 / 0.943	1.072 / 1.054	1.067 / 1.488	0.902 / 0.816	1.225 / 1.221
MolCLR _{gcn} ^{pretrained}	GRAPH	1.220 / 1.090	1.068 / 1.030	1.343 / 1.182	0.873 / 0.850	1.818 / 1.840	1.053 / 0.921	1.045 / 1.043	1.073 / 1.502	0.899 / 0.816	1.348 / 1.356
MolCLR _{gin} ^{pretrained}	GRAPH	1.174 / 1.066	1.081 / 1.025	1.066 / 0.902	0.873 / 0.850	1.232 / 1.102	0.987 / 0.828	1.121 / 1.132	1.058 / 1.487	0.909 / 0.824	1.498 / 1.499
AFP	GRAPH	1.116 / 1.184	1.102 / 1.062	0.747 / 1.106	1.149 / 1.215	1.091 / 1.101	1.080 / 0.990	1.059 / 1.079	1.062 / 1.131	0.947 / 0.872	1.233 / 1.238
MolCLR _{gin}	GRAPH	1.174 / 1.066	1.176 / 1.126	1.066 / 0.902	1.002 / 0.988	1.232 / 1.102	0.987 / 0.828	1.121 / 1.132	1.034 / 1.452	0.960 / 0.857	1.498 / 1.499
Contextpred ^{pretrained}	GRAPH	1.981 / 2.061	1.661 / 1.561	1.542 / 1.025	1.516 / 1.572	1.764 / 1.627	1.623 / 1.867	1.557 / 1.494	1.737 / 2.027	1.238 / 1.332	1.706 / 1.783
Contextpred	GRAPH	1.971 / 2.059	1.550 / 1.469	1.692 / 1.182	1.316 / 1.386	1.484 / 1.434	1.748 / 1.987	1.598 / 1.511	1.904 / 2.085	1.604 / 1.648	1.722 / 1.801
KPGT	GRAPH	2.534 / 2.663	1.514 / 1.589	0.607 / <u>0.682</u>	1.473 / 1.556	1.337 / 1.428	2.142 / 2.054	1.506 / 1.590	2.536 / 2.640	1.475 / 1.469	2.304 / 2.211
KPGT ^{pretrained}	GRAPH	2.625 / 2.711	1.369 / 1.465	0.572 / <u>0.747</u>	1.632 / 1.677	1.332 / 1.457	2.379 / 2.339	1.527 / 1.635	2.575 / 2.729	1.384 / 1.402	2.191 / 2.151

A.4 ALL RESULTS OF NINE DATASETS IN LSSNS

Tables 8 and 9 present the complete results on all nine LSSNS datasets. We report both RMSE and $\text{RMSE}_{\text{cliff}}$ for baseline machine learning models, graph-based neural networks, and GraphCliff. These results provide a detailed view of model performance in small-sample, narrow-scaffold regimes, highlighting the challenges posed by limited data diversity and the relative robustness of different approaches. Table 10 provides a mapping between LSSNS targets and similar MoleculeACE datasets, enabling cross-dataset comparison and transfer evaluation.

Table 8: Comparison of performance (RMSE) across protein targets in LSSNS.

Algorithm	Descriptor	USP7	RIP2	PKC ι	PHGDH	PLK1	IDO1	RXFP1	BRAF	mGluR2
GCN	GRAPH	0.5419	0.7355	0.8603	1.0886	0.6306	0.6485	0.6747	0.4889	0.4228
GAT	GRAPH	0.5062	0.7870	0.8039	0.4396	0.4873	0.6595	0.6333	0.5551	0.4412
AFP	GRAPH	0.5080	0.7947	1.1758	0.4010	0.5048	0.6555	0.4803	0.4771	0.3578
MPNN	GRAPH	0.5833	0.7779	0.9875	1.1800	0.5036	0.6287	0.6349	0.4436	0.4558
SVM	ECFP	0.5350	0.5787	0.8224	0.6174	0.5026	0.7858	0.4181	0.4120	0.2927
MLP	ECFP	0.5082	0.5875	0.8188	0.6865	0.4293	0.7047	0.4234	0.3778	0.3260
GraphCliff	GRAPH	0.5181	0.4824	1.8550	1.2460	0.6134	0.7354	0.6909	0.4557	0.4463
Transferred GraphCliff	GRAPH	0.3409	0.5476	0.6478	–	0.4837	–	0.6537	0.4550	0.2884

Table 9: Comparison of performance ($\text{RMSE}_{\text{cliff}}$) across protein targets in LSSNS.

Algorithm	Descriptor	USP7	RIP2	PKC ι	PHGDH	PLK1	IDO1	RXFP1	BRAF	mGluR2
GCN	GRAPH	0.4338	0.6928	1.4547	1.2228	0.5635	0.6335	0.8078	0.7702	0.5221
GAT	GRAPH	0.5759	0.6655	1.3094	0.6524	0.4489	0.7711	0.7475	0.8501	0.5134
AFP	GRAPH	0.6049	0.7494	1.9475	0.6833	0.4461	0.7975	0.5126	0.6437	0.3853
MPNN	GRAPH	0.4499	0.6920	1.6844	1.2934	0.4217	0.5570	0.7389	0.6052	0.5605
SVM	ECFP	0.6939	0.6234	1.3064	0.8643	0.4099	0.9141	0.5158	0.6290	0.3117
MLP	ECFP	0.5824	0.6667	1.3662	0.9428	0.4078	0.8142	0.5173	0.4697	0.3434
GraphCliff	GRAPH	0.5322	0.5665	1.4120	1.1136	0.5711	0.7090	0.8589	0.6860	0.5380
Transferred GraphCliff	GRAPH	0.4350	0.4818	0.9259	–	0.4885	–	0.8121	0.6950	0.3515

Table 10: Mapping between LSSNS protein targets and similar MoleculeACE datasets.

LSSNS Target	Class	Similar MoleculeACE datasets (Class)
USP7	Protease (cysteine protease) (UniProt, p)	CHEMBL204_Ki (Thrombin, serine protease) (UniProt, c) CHEMBL244_Ki (Factor X, serine protease) (UniProt, d)
RIP2	Kinase (Ser/Thr kinase) (UniProt, b)	CHEMBL2147_Ki (PIM1, Ser/Thr kinase) (UniProt, f) CHEMBL262_Ki (GSK3 β , Ser/Thr kinase) (UniProt, m)
PKC ι	Kinase (Ser/Thr kinase) (UniProt, l)	CHEMBL2147_Ki (PIM1, Ser/Thr kinase) (UniProt, f) CHEMBL262_Ki (GSK3 β , Ser/Thr kinase) (UniProt, m)
PHGDH	Other enzyme (oxidoreductase) (UniProt, a)	–
PLK1	Kinase (Ser/Thr kinase) (UniProt, n)	CHEMBL2147_Ki (PIM1, Ser/Thr kinase) (UniProt, f) CHEMBL262_Ki (GSK3 β , Ser/Thr kinase) (UniProt, m)
IDO1	Other enzyme (oxidoreductase) (UniProt, g)	–
RXFP1	GPCR (Class A) (UniProt, q)	CHEMBL214_Ki (5-HT1A, class A) (UniProt, e) CHEMBL219_Ki (D4, class A) (UniProt, i) CHEMBL231_Ki (Histamine H1, class A) (UniProt, j) CHEMBL234_Ki (D3, class A) (UniProt, k)
BRAF	Kinase (Ser/Thr kinase) (UniProt, h)	CHEMBL2147_Ki (PIM1, Ser/Thr kinase) (UniProt, f) CHEMBL262_Ki (GSK3 β , Ser/Thr kinase) (UniProt, m)
mGluR2	GPCR (Class C) (UniProt, o)	CHEMBL214_Ki (5-HT1A, class A) (UniProt, e) CHEMBL219_Ki (D4, class A) (UniProt, i) CHEMBL231_Ki (Histamine H1, class A) (UniProt, j) CHEMBL234_Ki (D3, class A) (UniProt, k)

A.5 ABLATION STUDY

Table 11 summarizes the effects of ablating short- and long-range filters, gating, and pooling strategies. Removing any component leads to substantial performance degradation, while replacing SAGPool with simple pooling further increases error, underscoring the importance of each design choice. Table 12 reports the performance of different GNN variants used in the short- and long-range filter components. Across all tested combinations, the configuration with GINE as the short-range filter and Chebyshev polynomials as the long-range operator consistently achieved the best performance in terms of both RMSE and $\text{RMSE}_{\text{cliff}}$.

Table 11: Performance comparison across different module ablations and pooling methods.

Short	Long	Gating	Pooling	RMSE	ΔRMSE (%)	$\text{RMSE}_{\text{cliff}}$	$\Delta\text{RMSE}_{\text{cliff}}$
O	O	O	SAGPool	0.673	–	0.766	–
O	O	–	SAGPool	0.725	+7.7%	0.798	+4.2%
O	–	O	SAGPool	0.856	+27.2%	0.933	+21.8%
–	O	O	SAGPool	1.288	+91.3%	1.287	+68.0%
O	–	–	SAGPool	1.001	+48.6%	1.038	+35.6%
–	O	–	SAGPool	1.327	+97.2%	1.314	+71.6%
–	–	O	SAGPool	1.361	+102.2%	1.286	+67.9%
O	O	O	Max	0.811	+20.5%	0.871	+13.7%
O	O	O	Mean	0.874	+29.9%	0.950	+24.0%
O	O	O	Sum	0.963	+43.1%	1.024	+33.7%

Table 12: Comparison of different GNN types used in the short- and long-range filters. Bold rows correspond to our default configuration (Short:GINE + Long:Chebyshev), which achieved the best overall performance.

Short	Long	RMSE	$\text{RMSE}_{\text{cliff}}$	Short	Long	RMSE	$\text{RMSE}_{\text{cliff}}$
GCN	GCN	0.713	0.798	GAT	GCN	0.695	0.786
	GIN	0.712	0.791		GIN	0.703	0.784
	GAT	0.692	0.780		GAT	0.706	0.794
	Chebyshev	0.724	0.819		Chebyshev	0.689	0.778
GIN	GCN	0.704	0.792	GINE	GCN	0.715	0.803
	GIN	0.710	0.799		GIN	0.694	0.774
	GAT	0.699	0.795		GAT	0.696	0.778
	Chebyshev	0.688	0.777		Chebyshev	0.673	0.766