Topological Feature Compression for Molecular Graph Neural Networks

Rahul Khorana
Department of Computing
Imperial College London
London, SW7 2AZ, UK
rahul.khorana24@imperial.ac.uk

Abstract

Recent advances in molecular representation learning have produced highly effective encodings of molecules for numerous cheminformatics and bioinformatics tasks. However, extracting general chemical insight while balancing predictive accuracy, interpretability, and computational efficiency remains a major challenge. In this work, we introduce a novel Graph Neural Network (GNN) architecture that combines compressed higher-order topological signals with standard molecular features. Our approach captures global geometric information while preserving computational tractability and human-interpretable structure. We evaluate our model across a range of benchmarks, from small-molecule datasets to protein-ligand interaction datasets, and demonstrate superior performance using a parameter-efficient architecture. We achieve the best performing results in both accuracy and robustness across almost all benchmarks. We open source all code [1].

1 Introduction

Machine learning has emerged as a powerful paradigm for modeling biochemical systems across various levels of chemical scales (Chandak et al. 2023; Somnath et al. 2021; Wang, Fu, et al. 2023; Wang, Li, Jin, et al. 2021; Wang, Li, Jiang, et al. 2019; Wang, Wang, Cao, et al. 2022). A predominant approach relies on one-dimensional, string-based representations such as SMILES and SELFIES, or derived fingerprints like ECFP. However, these representations fundamentally lack explicit encoding of 3D geometry and topology. This inherent limitation curtails their expressivity for structure-sensitive tasks crucial to drug discovery and materials science, including quantitative structure-activity relationship (QSAR) analysis, molecular docking, de novo drug design, compound identification, and molecular property prediction (Esders et al. 2025; Guha 2013; Hartenfeller and Schneider 2010; Keith et al. 2021; Khorana et al. 2024; Ma et al. 2011; McKinney et al. 2000; Williams 2008).

To imbue models with spatial and relational inductive biases, recent efforts have focused on geometrically-aware architectures, principally Graph Neural Networks (GNNs) and, more recently, higher-order structures like simplicial and cellular complexes (Bodnar 2023; Bodnar et al. 2021; Goh et al. 2022; Khorana et al. 2024; Verma et al. 2024; Wu, Yip, et al. 2023). Despite their enhanced representational power, these methods often present a challenging trade-off between model complexity and predictive performance. Paradoxically, on certain benchmark tasks, their performance, as measured by standard regression metrics (RMSE/MAE), can be surpassed by simpler, non-geometric baselines, highlighting issues of scalability, optimization, and generalization.

At the other end of the spectrum, methods grounded in first principles, such as Density Functional Theory (DFT), and surrogate Machine Learning Interaction Potentials (MLIPs) offer high-fidelity

[1] All code and results can be found on Github https://github.com/rahulkhorana/TFC-PACT-Net

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Science.

predictions by approximating quantum mechanical interactions (Butera 2024; Friederich et al. 2021; Huang et al. 2023). Nonetheless, their substantial computational cost, often scaling poorly with system size, renders them computationally infeasible. This establishes a clear trade-off in the field: a choice between computationally efficient but geometrically impoverished representations, geometrically aware yet poorly performing representations, and highly accurate but computationally prohibitive first-principles calculations. To address this challenge, graph neural networks (GNNs) have become a widely adopted tool (Hamilton et al. 2017; Kipf and Welling 2017; Veličković et al. 2018; Wang, Wang, Cao, et al. 2022; Xu et al. 2019; Zang et al. 2023). However, the conventional approach of using molecular graphs as input for GNNs may inadvertently constrain their potential for generating robust representations and performing accurate property prediction (Jiang, Xiao, et al. 2025).

In this study, we introduce PACTNET, a new graph neural network specifically designed to leverage compressed higher-order topological features from cellular representations. We create these enriched knowledge graphs by fusing these compressed features with knowledge graphs using our efficient cellular compression (ECC) algorithm. Our approach enables us to capture features representative of complex 3D molecular structures, producing robust embeddings, leading to improved property prediction performance on numerous benchmarks across many levels of chemical scale, from small molecules to complex biomolecules (protein-ligand complexes), and quantum properties. We demonstrate the empirically best performing approach in RMSE and MAE across all but two datasets. Our major contributions can be summarized as follows:

- **Novel Topological Feature Integration.** We introduce ECC, a method for augmenting molecular graphs with compressed features derived from higher-order cellular complexes. This technique creates a topologically-informed graph representation that enriches the node and edge features with multi-dimensional geometric and relational data. We demonstrate that this method provides a principled way to distill complex topological information into a standard graph structure, enhancing downstream model performance without requiring specialized higher-order architectures.

- **Computationally Efficient, Geometrically-Aware Molecular Embeddings.** We propose a new representation learning framework that resolves the prevailing trade-off between geometric fidelity and computational cost. By leveraging features from cellular complexes, our method generates embeddings that (1) retain crucial 3D structural information, overcoming the limitations of string representations; (2) demonstrate superior performance and robustness over standard GNNs across a diverse set of chemical tasks and scales; and (3) remain orders of magnitude more computationally efficient than first-principles methods like DFT.

- **A Novel Graph Neural Network.** We propose the PACTNET, a GNN that synergistically combines three distinct classes of features: (1) local neighborhood structure via principal neighborhood aggregation, (2) global higher-order topology via spectral features, and (3) node-level connectivity statistics via degree histograms. This multi-faceted aggregation scheme allows our resulting network, PACTNET, to capture a richer hierarchy of structural information than prior methods, demonstrably improving its expressive power and imbuing it with strong, chemically-relevant inductive biases.

To empirically validate our proposed methods, we conducted a comprehensive evaluation of PACTNET on seven benchmark datasets for molecular property prediction, encompassing diverse chemical properties and molecular scales. The performance of our model was benchmarked against a suite of well-established GNN architectures, including GCN (Kipf and Welling 2017), GAT (Veličković et al. 2018), GraphSAGE (Hamilton et al. 2017), and GIN (Xu et al. 2019). GIN is considered to be a gold standard benchmark in general graph based molecular learning (Fey et al. 2020). GCN and GAT are similarly considered to be competitive baseline models for graph based molecular learning (Jiang, Wu, et al. 2021).

Our results demonstrate that PACTNET establishes a new, high-performing baseline, significantly outperforming these widely-adopted models on five of the seven tasks with an empirical reduction in Root Mean Squared Error (RMSE).

2 Background & Related Work

2.1 Molecular Representations & Embeddings

Molecular machine learning faces a persistent trade-off between geometric fidelity, predictive accuracy, and computational cost. String-based encodings such as SMILES (Weininger 1988), SELFIES (Krenn et al. 2020), and fingerprint derivatives like ECFP (Rogers and Hahn 2010) are efficient but discard 3D structural information critical to molecular function (Liu, Wang, et al. 2021). Geometrically aware methods, ranging from 3D GNNs (Alhamoud et al. 2024; Godwin et al. 2021) to polyatomic complexes (Khorana et al. 2024), offer richer representations but often fail to surpass simpler baselines. At the other extreme, first-principles models such as MLIPs and Behler–Parrinello networks (Dusson et al. 2022; Lysogorskiy et al. 2021; Zuo et al. 2020) achieve high accuracy yet remain computationally prohibitive for large-scale screening.

Our framework seeks a middle ground, combining the geometric expressivity of topological methods with the predictive power of learned embeddings, while retaining practical scalability.

2.2 Graph Neural Networks

Operating directly on 2D molecular graphs, GNNs such as GCN (Kipf and Welling 2017), GraphSAGE (Hamilton et al. 2017), GAT (Veličković et al. 2018), and GIN (Xu et al. 2019) perform message passing, iteratively aggregating local neighborhoods (Zhong, Li, et al. 2023). This approach is limited by the Weisfeiler–Leman graph isomorphism test and struggles to capture long-range or global topological structure without deep, unstable architectures (Balcilar et al. 2021; Feng et al. 2022). Even specialized 3D models like PointNet++ (Qi et al. 2017), SchNet (Schütt et al. 2018), and ForceNet (Hu et al. 2021) address some of these issues but incur greater computational cost.

We address these limitations by enriching message passing with higher-order topological features, giving GNNs direct access to geometric and global information they cannot efficiently learn alone, improving robustness and expressivity without resorting to full 3D or first-principles methods.

3 Methods

In this section we present our PactNet neural network, the PACT-linear Layer (PACTLAYER), and ECC representation. We define the following key concepts below.

Definition 3.1. (**Molecular Graph**) A molecular graph (MG) is a structured representation of a molecule, namely a graph $\mathcal{G} = (V, E)$ where V is the vertex set and E the edge set. In the case of molecular graphs, V contains the atoms and E contains all bonds. Therefore providing a structured way in which to represent a molecule.

Definition 3.2. (**Knowledge Graph**) A knowledge graph (KG) is a structured representation of knowledge in which nodes are connected by relations or edges. Formally a directed knowledge graph is represented as a set of triples $\mathcal{T} = \{(h, r, t)_i\}_{i=1}^{n}$ where each triple contains a head entity h, tail entity t, and relation r connecting them. A knowledge graph can also be viewed as a graph $\mathcal{G} = (V, E)$.

Definition 3.3. (**Lifting Transformation**, Bodnar et al. (2021)) A cellular lifting transformation is a function $f : \mathcal{G} \to \mathcal{X}$ from the space of graphs $\mathcal{G}$ to the space of regular cell complexes $\mathcal{X}$ with the property that two graphs $\mathcal{G}_1, \mathcal{G}_2$ are isomorphic iff the cell complexes $f(\mathcal{G}_1)$, $f(\mathcal{G}_2)$ are isomorphic.

3.1 Topological Compression & ECC Algorithm

The need for compression Formally, we want to leverage Cell Complexes to provide higher order geometric information to our model. In areas such as materials design, quantum chemistry, and protein informatics the geometry of the molecule is of fundamental importance for property prediction and design (Krapp et al. 2024; Wang, Wang, Li, et al. 2024). However, cell complexes can be high dimensional and complex to learn over (Bodnar et al. 2021). In order to reduce the complexity, we compress relevant information that can be extracted from the cell complex.

ECC Algorithm Given a molecular graph G we use a Molecular-Lifting Transformation, which is a kind of lifting transformation as in definition 3.3.

Definition 3.4. (**Molecular-Lifting Transformation**) Formally, let M be a molecule and G_M be the corresponding molecular graph. Then we define a function $\mathcal{F}$ which sends $G_M \mapsto X_M$, where X_M is a cell complex. In particular, given a set of atoms in the molecule contained in $V_M := \{A_1, \ldots, A_n\}$ we determine information about the number of protons, neutrons and electrons for each atom, and use these as base-points X_M^0 (0-cells). We encapsulate these inside the individual atoms via attachment (1-cells). Then we connect the bonds (2-cells). Finally we consider induced cycles, chemical rings, and k-hop interactions (3-cells). Thus we end up with, for molecule M, the corresponding $\{X_M^0, X_M^1, X_M^2, X_M^3\}$. The resulting cell complex is termed X_M and skeleton preserving (isomorphic in the sense that $\mathcal{F}(X_M^2)$ corresponds to the molecular graph G).

This type of molecular lifting map is more complex than the scheme proposed by Bodnar et al. (2021), yet less complex and containing far less information than the representation developed by Khorana et al. (2024). The map $\mathcal{F}$ occupies an optimal middle ground, balancing representational complexity, geometric information, and computational cost.

Therefore, upon transforming molecule M to cell complex $\mathcal{F}(M)$, we can easily extract relevant topologically rich features. We compute the betti-numbers, take the eigen-decomposition of the chain matrix (termed spectral chains), accumulate the top-k eigenvalues of the Laplacians, compute the degree centrality over skeleta, and determine the all-pairs shortest path distance over X_M^2. The definitions of the chain matrix and Laplacians are provided in Ribando-Gros et al. (2024). Intuitively, one can think of the chain matrix as consisting of many sampled formal sums over the k-cells. Effectively, this is somewhat analogous to the notion of collecting many paths over cells in a matrix. Each path is like a random walk over nodes in a graph, but instead of nodes, one has cells. The Laplacians effectively provide information about the connectedness of neighboring cells. One can compute statistics in this context and reduce dimensionality further. A technique one may apply is mean aggregation as in Rahmani et al. (2021). All of these computed features are tensors that are concatenated and padded to uniform dimension. The resulting tensor is termed an ECC representation of molecule M. The algorithm is summarized in Figure 1.

3.2 GNN Architecture

Upon determining the molecular graph G_M for molecule M we construct our ECC representation. Then simultaneously, we enrich our graph by adding in features related to rotatable bonds, aromaticity, degree, charge and bond type. Additionally, we compute the degree histograms and embed them. Afterward we can apply convolutional layers, batch norm and pooling. Our choice of convolution is the principal neighborhood aggregation scheme developed by Corso et al. (2020). The complete architecture is summarized in Figure 1.

4 Experiments

We run a wide array of experiments across standard benchmark datasets and provide performance metrics (RMSE/MAE) across these datasets. The datasets we use are summarized in table 1. A deeper dive of these datasets can be found in Appendix B.

Table 1: Overview of datasets used for experimental validation. All datasets are sourced from the MoleculeNet, BindingDB, and QuanDB benchmark suites.

Dataset	Task Type	# Molecules	Target Property	Scale/Domain
ESOL	Regression	1,128	Water Solubility (log mol/L)	Biophysics (Small)
FreeSolv	Regression	642	Hydration Free Energy (kcal/mol)	Biophysics (Small)
Lipophilicity	Regression	4,200	Octanol/Water Distribution (logD)	Biophysics (Small)
Boiling Point	Regression	2,983	Boiling Point (°C)	Physical Chemistry
QM9	Regression	134k	Heat Capacity (Cv)	Quantum Mechanics
IC50	Regression	2,822	pIC50 (-log10 of IC50)	Pharmacology
BindingDB	Regression	4,614	Binding Affinity (Ki/Kd)	Pharmacology

Data Preprocessing All datasets were subjected to a standard preprocessing pipeline. To manage computational load, we created subsets by taking a random sample of at most 2000 molecules from

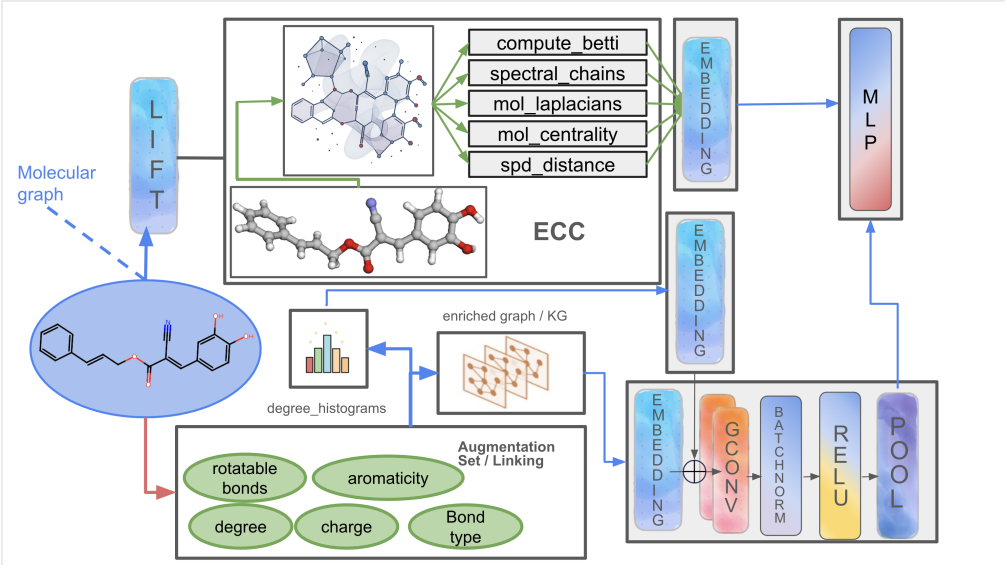

Figure 1: Overview of PactNet architecture and ECC representation. The model takes a molecular graph as input, extracts spectral, chemical, and structural features, enriches the graph, and processes it through embedding and graph convolution layers before prediction.

any dataset exceeding this size, and made these subsets publicly available for reproducibility. For the IC50 and BindingDB datasets, target values were transformed to pIC50, a standard practice in biochemical modeling (Kramer et al. 2012). For graph-based models (GNNs), molecular graphs were generated from SMILES strings using RDKit, following the protocol outlined by Dablander (2024) and Bento et al. (2020). Any molecules that failed parsing or contained null values were removed prior to experimentation.

Experimental Setup To estimate the generalization error of our entire learning procedure (including hyperparameter optimization), we employ a 5-FOLD NESTED CROSS-VALIDATION scheme (Zhong, Chalise, et al. 2023). A 5-fold partition was chosen to balance the trade-off between the bias and variance of the error estimate against the considerable computational expense of the nested procedure. The outer loop provides a nearly unbiased estimate of the true error, where each of the 5 folds serves as a hold-out validation set exactly once. The inner loop is used exclusively for hyperparameter selection on the remaining 4 folds.

Crucially, for all experiments, the outer-loop data splits were generated once using a fixed random seed (`random_state=42`) and reused for every model and baseline. This ensures that any observed performance differences are attributable to the models themselves, not to variance in the data partitioning, thus satisfying the pairing assumption for subsequent statistical tests.

Within each inner loop of the nested cross-validation, we performed automated hyperparameter optimization (HPO) on an 80/20 split of the inner-loop data. We utilized a Tree-structured Parzen Estimator (TPE) for optimization, implemented in the Optuna library, chosen for its demonstrated efficiency over random search (Akiba et al. 2019).

The HPO was configured to run for 15 trials, with the objective of minimizing the validation set's Mean Absolute Error (MAE). For fairness, a consistent hyperparameter search space was used for all GNN models, defined as follows:

- **Learning Rate (η):** Sampled from a log-uniform distribution over $[5 \times 10^{-4}, 1 \times 10^{-3}]$. This range was selected to ensure training stability, as preliminary experiments revealed that unconstrained searches led to highly erratic performance.
- **Hidden Dimensionality (d_h):** A categorical choice from $\{64, 128, 256\}$.
- **Batch Size (B):** A categorical choice from $\{32, 64\}$.

Table 2: Detailed performance of PactNet on the QM9 dataset. Our model achieves the lowest validation and test RMSE and MAE, outperforming all other baselines. The units are in cal $\text{mol}^{-1}\text{K}^{-1}$.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
QM9	**PactNet (ECC)**	$\mathbf{0.999 \pm 0.099}$	$\mathbf{0.659 \pm 0.069}$	$\mathbf{1.0480 \pm 0.1805}$	$\mathbf{0.6510 \pm 0.0815}$
	GAT (ECFP)	2.490 ± 0.149	1.826 ± 0.051	2.4370 ± 0.2845	1.7870 ± 0.1620
	GAT (SELFIES)	1.898 ± 0.072	1.493 ± 0.048	1.8170 ± 0.1360	1.4210 ± 0.1135
	GAT (SMILES)	1.896 ± 0.071	1.488 ± 0.052	1.7820 ± 0.1350	1.4030 ± 0.1105
	GCN (ECFP)	2.483 ± 0.171	1.811 ± 0.085	2.4260 ± 0.2380	1.8390 ± 0.1615
	GCN (SELFIES)	2.358 ± 0.096	1.871 ± 0.078	2.1550 ± 0.1510	1.7270 ± 0.1290
	GCN (SMILES)	2.423 ± 0.063	1.926 ± 0.046	2.1260 ± 0.1575	1.6980 ± 0.1265
	GIN (ECFP)	2.583 ± 0.113	1.913 ± 0.066	2.5470 ± 0.2480	1.9270 ± 0.1660
	GIN (SELFIES)	1.883 ± 0.081	1.489 ± 0.068	1.8120 ± 0.1300	1.4350 ± 0.1060
	GIN (SMILES)	1.876 ± 0.063	1.492 ± 0.058	1.8120 ± 0.1395	1.4190 ± 0.1085
	SAGE (ECFP)	2.516 ± 0.165	1.836 ± 0.086	2.4860 ± 0.2580	1.8310 ± 0.1640
	SAGE (SELFIES)	1.561 ± 0.031	1.206 ± 0.029	1.4880 ± 0.1180	1.1560 ± 0.0915
	SAGE (SMILES)	1.552 ± 0.050	1.209 ± 0.034	1.4290 ± 0.1120	1.0890 ± 0.0905

Individual models were trained using the Adam optimizer with default parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) to minimize the Mean Absolute Error (Kingma and Ba 2014). Training was conducted for a maximum of 200 epochs. An early stopping criterion was implemented to mitigate overfitting to the inner-loop validation set. Specifically, we monitored the validation MAE and terminated training if no improvement was observed for a patience of 20 epochs. This patience value was chosen to be large enough to allow the model to escape shallow local minima without risking significant overfitting. Upon termination, the model parameters that yielded the lowest validation MAE during that run were restored for the evaluation on the outer-loop validation fold. The experimental setup is summarized in Algorithm 1 which can be found in Appendix A. Compute cost is discussed in Appendix F.

5 Results

In this section, we present and analyze the main experimental results. To maintain focus, we provide a detailed synthesis for a representative subset of the foundational datasets. These tasks highlight the key performance trends across the selected datasets. The comprehensive results for all seven datasets are provided in Appendix D. We report that PactNet+ECC achieves the best validation set MAE, validation set RMSE, test set MAE, and test set RMSE across five of the seven selected datasets. For the other two datasets (BindingDB, IC50) PactNet+ECC test set performance is within 5% of the best performing model. On the BindingDB dataset PactNet+ECC achieves the best validation RMSE and validation MAE. On the IC50 dataset PactNet+ECC achieves the best validation RMSE.

5.1 Summary Table

In this section, we provide the summary table for the QM9 benchmark dataset. Note that across all dataset benchmark tables we adopt the standard of reporting validation metrics as mean $\pm$ standard deviation from 5-fold nested cross-validation. The global test set metrics are reported as the mean $\pm$ the half-width of the 95% confidence interval from a non-parametric bootstrap. The best validation set and test set RMSE and MAE for each dataset is highlighted in **bold**. We delve into the statistical analysis of these results in the next subsection. All seven summary tables are provided in Appendix D. Moreover a deep dive into each dataset can be found in the Appendix B.

5.2 Statistical Analysis

To rigorously evaluate the performance of our proposed model, PACTNET, against existing baselines, we employ a robust statistical validation framework. Our primary evaluation metric is the Root Mean Squared Error (RMSE), obtained through a 5-fold nested cross-validation procedure as detailed in Algorithm 1. This nested CV provides a near unbiased estimate of each model's performance on the five outer folds, yielding a vector of five RMSE and MAE scores for each model being compared. To

determine if the observed performance improvements of PACTNET are statistically significant, we conduct a multi-stage analysis. We justify our choices of our test in Appendix C.

Design Suppose we have a fixed prediction task and dataset $\mathcal{D}$. We partition our dataset into D_{trainval} and D_{test} as in Algorithm 1. Then let L be a loss function, either MAE or RMSE. Model development and HPO are confined to D_{trainval} where we perform nested K-fold cross-validation as per the conventions in Zhong, Chalise, et al. (2023). After selection, the chosen model M is refit with early stopping on all of D_{trainval} and evaluated once on D_{test}. The reported values include a 95% confidence interval determined via non-parametric bootstrap. Test-set metrics are reported descriptively and are not used for the formal within-dataset hypothesis tests. As discussed by Bengio and Grandvalet (2004), Nested Cross Validation mitigates selection bias in point estimation but does not render outer folds independent, which matters for valid inference (Bengio and Grandvalet 2004; Cawley and Talbot 2010; Varma and Simon 2006). However, it is still more rigorous than the alternative, namely vanilla Cross Validation (Cawley and Talbot 2010; Varma and Simon 2006).

Estimand and hypotheses Fix a competitor j and the designated control model c. Let $\ell_i^{(m)}$ denote the outer-fold validation loss on fold $i \in \{1, \ldots, K\}$ for model $m \in \{c, j\}$. Define the paired fold-wise contrast

$$d_i^{(L)} \;\equiv\; \ell_i^{(j)} - \ell_i^{(c)}, \qquad \bar{d}^{(L)} \;\equiv\; \frac{1}{K}\sum_{i=1}^{K} d_i^{(L)} \tag{1}$$

as is convention; see Nadeau and Bengio (1999), who define the per-split loss difference $L_{A\text{–}B}(j,i) = L_A(j,i) - L_B(j,i)$ and base inference on the mean of such differences across splits with a corrected variance, and Dietterich (1998).

By construction, $d_i^{(L)} > 0$ indicates the control attains smaller loss than the competitor on fold i, and $\bar{d}^{(L)} > 0$ indicates an average advantage. Our one-sided hypothesis for each loss L is

$$H_0 : \mathbb{E}\Big[\bar{d}^{(L)}\Big] \le 0 \qquad \text{vs} \qquad H_1 : \mathbb{E}\Big[\bar{d}^{(L)}\Big] > 0. \tag{2}$$

This is the classic one-sided, one-sample (paired) t-test on the mean fold-wise difference (Bouckaert and Frank 2004; Nadeau and Bengio 1999). Classic constructions that justify the same paired-per-split paradigm include Alpaydın (1999) and Dietterich (1998).

Nadeau-Bengio corrected t-statistic (primary, within-dataset) To account for K-fold dependence we adopt the Nadeau-Bengio correction (Nadeau and Bengio 1999). Let

$$s^2 \equiv \frac{1}{K-1}\sum_{i=1}^{K}\Big(d_i^{(L)} - \bar{d}^{(L)}\Big)^2, \qquad \rho_0 \equiv \frac{1}{K-1}. \tag{3}$$

The corrected standard error and test statistic are

$$\widehat{\mathrm{SE}}_{\mathrm{NB}} \equiv \sqrt{\Big(\tfrac{1}{K} + \rho_0\Big)\, s^2} = \sqrt{\Big(\tfrac{1}{K} + \tfrac{1}{K-1}\Big)\, s^2}, \qquad t_{\mathrm{NB}} \equiv \frac{\bar{d}^{(L)}}{\widehat{\mathrm{SE}}_{\mathrm{NB}}}. \tag{4}$$

We reference t_{NB} to a Student distribution with $\nu = K - 1$ degrees of freedom and report the one-sided upper-tail p-value

$$p = \Pr\left(T_\nu \ge t_{\mathrm{NB}}\right). \tag{5}$$

Because the alternative (control superiority) is pre-specified, we avoid post-hoc tail selection. A conservative NB-style interval for $\bar{d}^{(L)}$ may be reported as

$$\bar{d}^{(L)} \pm t_{1-\alpha/2,\nu}\, \widehat{\mathrm{SE}}_{\mathrm{NB}}. \tag{6}$$

The correction is intentionally conservative at small K (e.g., $\rho_0 = 1/4$ when $K = 5$) and effect sizes are considered.

Multiplicity Within a dataset the control is compared to $M - 1$ competitors. We control the family-wise error rate (FWER) at level α via Holm's sequentially rejective procedure (Holm 1979), applied separately to the MAE family and to the RMSE (or MSE) family. Holm's method is uniformly more powerful than Bonferroni and provides strong FWER control without restrictive dependence assumptions.

Reporting For each competitor we report the one-sided p-value, and Holm-adjusted p-value for both MAE and RMSE. Additionally we provide the raw mean difference in RMSE, $t_{NB-RMSE}$, mean difference in MAE, t_{NB-MAE}, and 95% confidence intervals for NB RMSE and NB MAE. Note that with $K = 5$ the degrees of freedom are small and the NB inflation is conservative. This is a deliberate choice that privileges calibrated size over power when the dependence structure is only partially observable (Bengio and Grandvalet 2004; Nadeau and Bengio 1999). The resulting values from our statistical tests for all datasets are given in Appendix E.

5.3 Results and Discussion

Table 3: Summary of statistical analysis across all datasets.

Dataset	***p*-value** (< 0.05)	**Holm *p*-value** (< 0.05)	**Conclusion (PactNet)**
QM9	✓	✓	Statistically Superior
ESOL	✓	✓	Statistically Superior
BOILINGPOINT	✓	✓	Statistically Superior
LIPOPHIL	✓	✓	Statistically Superior
FREESOLV	✓	✓	Statistically Superior
BINDINGDB	—	—	Statistical Tie
IC50	—	—	Statistical Tie

Datasets: QM9, ESOL, BOILINGPOINT, LIPOPHIL, FREESOLV Based on the analysis summarized in Table 3, our model, PactNet, demonstrates statistically significant, superior performance. The observed p-values and Holm-corrected p-values are highly significant, leading to a confident rejection of the null hypothesis. This provides strong statistical evidence of PactNet's superiority over the compared baseline models across these five datasets. Furthermore, the empirical results consistently show that PactNet achieves the lowest Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) values on both the validation and test sets for every dataset (Appendix D). This consistent and statistically significant improvement solidifies our conclusion that PactNet is a superior model compared to the commonly used or gold-standard baselines.

Datasets: BINDINGDB, IC50 As detailed in Table 3, PactNet demonstrates performance statistically equivalent to the best-performing models on these two datasets. Although the statistical analysis did not find a significant difference that would allow us to confidently reject the null hypothesis, the empirical results are compelling. The differences in observed RMSE and MAE values across both validation and test sets are less than 5%. This negligible margin of error indicates that for practical applications, PactNet's performance is indistinguishable from the top-performing baselines on these specific benchmarks. Therefore we determine the outcome is a statistical tie or perhaps negligibly worse.

6 Conclusion

Thus, we have shown that our approach delivers empirically and statistically significant improvements on most benchmarks. It strikes a promising balance, avoiding the computational cost of first-principles methods and the representational limits of purely 2D or string-based models, while retaining geometric expressivity and scalability. We have also introduced an algorithm for molecular embeddings that is efficient to compute, geometrically aware, and robust. Future work will focus on scaling our architecture to larger biomolecules, incorporating richer topological invariants or equivariant layers, and investigating the training dynamics of models built on our ECC embeddings. The main limitations of our study are its evaluation on a finite set of benchmarks with relatively small datasets, and the fact that our representation is not well suited for tasks requiring extremely fine-grained physical detail, such as quantum-level interaction modeling. These results strongly suggest that our framework can serve as a scalable and versatile foundation for the next generation of molecular machine learning models.

References

[1] Takuya Akiba et al. "Optuna: A next-generation hyperparameter optimization framework". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[2] Kumail Alhamoud et al. "Leveraging 2D molecular graph pretraining for improved 3D conformer generation with graph neural networks". In: *Computers & Chemical Engineering* 183 (2024), p. 108622.

[3] E Alpaydın. "Combined 5× 2 cv F test for comparing supervised classification learning classifiers". In: *Neural Computation* 11 (1999), pp. 1975–1982.

[4] Muhammet Balcilar et al. "Breaking the limits of message passing graph neural networks". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 599–608.

[5] Yoshua Bengio and Yves Grandvalet. "No unbiased estimator of the variance of k-fold cross-validation". In: *Journal of machine learning research* 5.Sep (2004), pp. 1089–1105.

[6] A Patrícia Bento et al. "An open source chemical structure curation pipeline using RDKit". In: *Journal of Cheminformatics* 12.1 (2020), p. 51.

[7] Cristian Bodnar. "Topological deep learning: graphs, complexes, sheaves". PhD thesis. 2023.

[8] Cristian Bodnar et al. "Weisfeiler and lehman go topological: Message passing simplicial networks". In: *International conference on machine learning*. PMLR. 2021, pp. 1026–1037.

[9] Remco R Bouckaert and Eibe Frank. "Evaluating the replicability of significance tests for comparing learning algorithms". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2004, pp. 3–12.

[10] Valeria Butera. "Density functional theory methods applied to homogeneous and heterogeneous catalysis: a short review and a practical user guide". In: *Physical Chemistry Chemical Physics* 26.10 (2024), pp. 7950–7970.

[11] Gavin C Cawley and Nicola LC Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation". In: *The Journal of Machine Learning Research* 11 (2010), pp. 2079–2107.

[12] Payal Chandak, Kexin Huang, and Marinka Zitnik. "Building a knowledge graph to enable precision medicine". In: *Scientific Data* 10.1 (2023), p. 67.

[13] Gabriele Corso et al. "Principal neighbourhood aggregation for graph nets". In: *Advances in neural information processing systems* 33 (2020), pp. 13260–13271.

[14] Markus Dablander. "Investigating graph neural networks and classical feature-extraction techniques in activity-cliff and molecular property prediction". In: *arXiv preprint arXiv:2411.13688* (2024).

[15] Janez Demšar. "Statistical comparisons of classifiers over multiple data sets". In: *Journal of Machine learning research* 7.Jan (2006), pp. 1–30.

[16] Thomas G Dietterich. "Approximate statistical tests for comparing supervised classification learning algorithms". In: *Neural computation* 10.7 (1998), pp. 1895–1923.

[17] Genevieve Dusson et al. "Atomic cluster expansion: Completeness, efficiency and stability". In: *Journal of Computational Physics* 454 (2022), p. 110946.

[18] Rob Eisinga et al. "Exact p-values for pairwise comparison of Friedman rank sums, with application to comparing classifiers". In: *BMC bioinformatics* 18.1 (2017), p. 68.

[19] Malte Esders et al. "Analyzing atomic interactions in molecules as learned by neural networks". In: *Journal of Chemical Theory and Computation* 21.2 (2025), pp. 714–729.

[20] Jiarui Feng et al. "How powerful are k-hop message passing graph neural networks". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 4776–4790.

[21] Matthias Fey, Jan-Gin Yuen, and Frank Weichert. "Hierarchical inter-message passing for learning on molecular graphs". In: *arXiv preprint arXiv:2006.12179* (2020).

[22] Pascal Friederich et al. "Machine-learned potentials for next-generation matter simulations". In: *Nature Materials* 20.6 (2021), pp. 750–761.

[23] Jonathan Godwin et al. "Simple gnn regularisation for 3d molecular property prediction & beyond". In: *arXiv preprint arXiv:2106.07971* (2021).

[24] Christopher Wei Jin Goh, Cristian Bodnar, and Pietro Lio. "Simplicial attention networks". In: *arXiv preprint arXiv:2204.09455* (2022).

[25] Rajarshi Guha. "On exploring structure–activity relationships". In: *In silico models for drug discovery* (2013), pp. 81–94.

[26] Will Hamilton, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs". In: *Advances in neural information processing systems* 30 (2017).

[27] Markus Hartenfeller and Gisbert Schneider. "De novo drug design". In: *Chemoinformatics and computational chemical biology* (2010), pp. 299–323.

[28] Sture Holm. "A simple sequentially rejective multiple test procedure". In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[29] Weihua Hu et al. "Forcenet: A graph neural network for large-scale quantum calculations". In: *arXiv preprint arXiv:2103.01436* (2021).

[30] Bing Huang, Guido Falk von Rudorff, and O Anatole von Lilienfeld. "The central role of density functional theory in the AI age". In: *Science* 381.6654 (2023), pp. 170–175.

[31] Dejun Jiang, Zhenxing Wu, et al. "Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models". In: *Journal of cheminformatics* 13.1 (2021), p. 12.

[32] Pengcheng Jiang, Cao Xiao, et al. "Bi-level contrastive learning for knowledge-enhanced molecule representations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 1. 2025, pp. 352–360.

[33] John A Keith et al. "Combining machine learning and computational chemistry for predictive insights into chemical systems". In: *Chemical reviews* 121.16 (2021), pp. 9816–9872.

[34] Rahul Khorana, Marcus Noack, and Jin Qian. "Polyatomic Complexes: A topologically-informed learning representation for atomistic systems". In: *arXiv preprint arXiv:2409.15600* (2024). URL: https://arxiv.org/abs/2409.15600.

[35] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[36] Thomas N. Kipf and Max Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *International Conference on Learning Representations*. 2017. URL: https://openreview.net/forum?id=SJU4ayYgl.

[37] Christian Kramer et al. "The experimental uncertainty of heterogeneous public K i data". In: *Journal of medicinal chemistry* 55.11 (2012), pp. 5165–5173.

[38] Lucien F Krapp et al. "Context-aware geometric deep learning for protein sequence design". In: *Nature Communications* 15.1 (2024), p. 6273.

[39] Mario Krenn et al. "Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation". In: *Machine Learning: Science and Technology* 1.4 (2020), p. 045024.

[40] Shengchao Liu, Hanchen Wang, et al. "Pre-training molecular graph representation with 3d geometry". In: *arXiv preprint arXiv:2110.07728* (2021).

[41] Tiqing Liu, Yuhmei Lin, et al. "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities". In: *Nucleic acids research* 35.suppl_1 (2007), pp. D198–D201.

[42] Yury Lysogorskiy et al. "Performant implementation of the atomic cluster expansion (PACE) and application to copper and silicon". In: *npj computational materials* 7.1 (2021), p. 97.

[43] Dik-Lung Ma, Daniel Shiu-Hin Chan, and Chung-Hang Leung. "Molecular docking for virtual screening of natural product databases". In: *Chemical science* 2.9 (2011), pp. 1656–1665.

[44] James D McKinney et al. "The practice of structure activity relationships (SAR) in toxicology". In: *Toxicological Sciences* 56.1 (2000), pp. 8–17.

[45] Claude Nadeau and Yoshua Bengio. "Inference for the generalization error". In: *Advances in neural information processing systems* 12 (1999).

[46] Charles Ruizhongtai Qi et al. "Pointnet++: Deep hierarchical feature learning on point sets in a metric space". In: *Advances in neural information processing systems* 30 (2017).

[47] Mostafa Rahmani, Rasoul Shafipour, and Ping Li. "Non-local feature aggregation on graphs via latent fixed data structures". In: *2021 55th Asilomar Conference on Signals, Systems, and Computers*. IEEE. 2021, pp. 1551–1557.

[48] Kandethody M Ramachandran and Chris P Tsokos. *Mathematical statistics with applications in R*. Academic Press, 2020.

[49] Emily Ribando-Gros et al. “Combinatorial and hodge Laplacians: similarities and differences”. In: *SIAM Review* 66.3 (2024), pp. 575–601.

[50] David Rogers and Mathew Hahn. “Extended-connectivity fingerprints”. In: *Journal of chemical information and modeling* 50.5 (2010), pp. 742–754.

[51] Kristof T Schütt et al. “Schnet–a deep learning architecture for molecules and materials”. In: *The Journal of chemical physics* 148.24 (2018).

[52] Vignesh Ram Somnath, Charlotte Bunne, and Andreas Krause. “Multi-scale representation learning on proteins”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 25244–25255.

[53] Sudhir Varma and Richard Simon. “Bias in error estimation when using cross-validation for model selection”. In: *BMC bioinformatics* 7.1 (2006), p. 91.

[54] Petar Veličković et al. “Graph Attention Networks”. In: *International Conference on Learning Representations*. 2018. URL: https://openreview.net/forum?id=rJXMpikCZ.

[55] Yogesh Verma, Amauri H Souza, and Vikas Garg. “Topological neural networks go persistent, equivariant, and continuous”. In: *arXiv preprint arXiv:2406.03164* (2024).

[56] Jutharath Voraprateep. “Robustness of Wilcoxon signed-rank test against the assumption of symmetry”. PhD thesis. University of Birmingham, 2013.

[57] Hanchen Wang, Tianfan Fu, et al. “Scientific discovery in the age of artificial intelligence”. In: *Nature* 620.7972 (2023), pp. 47–60.

[58] Hongwei Wang, Weijiang Li, Xiaomeng Jin, et al. *Chemical-Reaction-Aware Molecule Representation Learning*. 2021. arXiv: 2109.09888 [cs.LG]. URL: https://arxiv.org/abs/2109.09888.

[59] Xiaofeng Wang, Zhen Li, Mingjian Jiang, et al. “Molecule property prediction based on spatial graph embedding”. In: *Journal of chemical information and modeling* 59.9 (2019), pp. 3817–3828.

[60] Yusong Wang, Tong Wang, Shaoning Li, et al. “Enhancing geometric representations for molecules with equivariant vector-scalar interactive message passing”. In: *Nature Communications* 15.1 (2024), p. 313.

[61] Yuyang Wang, Jianren Wang, Zhonglin Cao, et al. “Molecular contrastive learning of representations via graph neural networks”. In: *Nature Machine Intelligence* 4.3 (2022), pp. 279–287.

[62] David Weininger. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”. In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[63] Antony J Williams. “Public chemical compound databases”. In: *Current Opinion in Drug Discovery and Development* 11.3 (2008), p. 393.

[64] Hanrui Wu, Andy Yip, et al. “Simplicial complex neural networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.1 (2023), pp. 561–575.

[65] Zhenqin Wu, Bharath Ramsundar, et al. “MoleculeNet: a benchmark for molecular machine learning”. In: *Chemical science* 9.2 (2018), pp. 513–530.

[66] Keyulu Xu et al. “How Powerful are Graph Neural Networks?” In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=ryGs6iA5Km.

[67] Zhijiang Yang et al. “QuanDB: a quantum chemical property database towards enhancing 3D molecular representation learning”. In: *Journal of cheminformatics* 16.1 (2024), p. 48.

[68] Xuan Zang, Xianbing Zhao, and Buzhou Tang. “Hierarchical molecular graph self-supervised learning for property prediction”. In: *Communications Chemistry* 6.1 (2023), p. 34.

[69] Yi Zhong, Prabhakar Chalise, and Jianghua He. “Nested cross-validation with ensemble feature selection and classification model for high-dimensional biological data”. In: *Communications in statistics-simulation and computation* 52.1 (2023), pp. 110–125.

[70] Zhiqiang Zhong, Cheng-Te Li, and Jun Pang. “Hierarchical message-passing graph neural networks”. In: *Data Mining and Knowledge Discovery* 37.1 (2023), pp. 381–408.

[71] Yunxing Zuo et al. “Performance and cost assessment of machine learning interatomic potentials”. In: *The Journal of Physical Chemistry A* 124.4 (2020), pp. 731–745.

A Experimental Algorithm

Algorithm 1: Experimental Design: Nested CV & Final Evaluation

Require : Dataset $D = (\mathcal{X}, \mathcal{Y})$, split into D_{trainval} and D_{test}
Require : Number of outer folds $K = 5$
Ensure : Unbiased validation metrics $(\mu_{\text{val}}, \sigma_{\text{val}})$
Ensure : Final test metrics with confidence intervals $\text{Metrics}_{\text{test}}$

▷ `Part 1: Unbiased Performance Estimation`
$(\mu_{\text{val}}, \sigma_{\text{val}}) \leftarrow$ `NestedCVEvaluation(`$D_{trainval}, K$`);`
▷ `Part 2: Final Model Training and Testing`
$\text{Metrics}_{\text{test}} \leftarrow$ `FinalModelEvaluation(`$D_{trainval}, D_{test}$`);`

Function `NestedCVEvaluation(`$D_{trainval}, K$`):`
  fold_metrics $\leftarrow [\,]$;
  for $k \leftarrow 1$ **to** K **do**
    $(D_{\text{train}}^{(k)}, D_{\text{val}}^{(k)}) \leftarrow k$-th fold of D_{trainval};
    $\theta_k^* \leftarrow$ `FindBestHyperparameters(`$D_{train}^{(k)}$`)` ▷ `Inner loop for HPO`
    $S_k \leftarrow$ Scaler().fit($D_{\text{train}}^{(k)}$);
    $M_k \leftarrow$ `TrainModel(`$D_{train}^{(k)}, \theta_k^*, S_k$`);`
    $\text{metric}_k \leftarrow$ `Evaluate(`$M_k, D_{val}^{(k)}, S_k$`);`
    Append metric_k to fold_metrics;
  return (mean(fold_metrics), std(fold_metrics));

Function `FinalModelEvaluation(`$D_{trainval}, D_{test}$`):`
  $\theta_{\text{final}}^* \leftarrow$ `FindBestHyperparameters(`$D_{trainval}$`);`
  $S_{\text{final}} \leftarrow$ Scaler().fit(D_{trainval});
  $M_{\text{final}} \leftarrow$ `TrainModel(`$D_{trainval}, \theta_{final}^*, S_{final}$`);`
  $y_{\text{true}}, y_{\text{pred}} \leftarrow$ `Evaluate(`$M_{final}, D_{test}, S_{final}$`);`
  $\text{metrics}_{\text{test}} \leftarrow$ `BootstrapCI(`y_{true}, y_{pred}`);`
  return $\text{metrics}_{\text{test}}$;

B Dataset Deep Dive

These tasks were selected from the MoleculeNet, BindingDB, and QuanDB benchmark sets (Liu, Lin, et al. 2007; Wu, Ramsundar, et al. 2018; Yang et al. 2024). MoleculeNet is a comprehensive and commonly used set of benchmark datasets for molecular property prediction. It was released as part of the DeepChem library and contains datasets from across quantum mechanics, physical chemistry, biophysics and physiology (Wu, Ramsundar, et al. 2018). The QM9 dataset is a comprehensive dataset which provides geometric, electronic and related thermodynamic properties. All molecules are modeled using density functional theory with the B3LYP functional and 6-31G basis set. The ESOL dataset is a small dataset consisting of water solubility data for compounds. Note that solubility is a property the molecule and not its conformers. FreeSolv provides experimentally determined hydration free energy of small molecules in water. Lipophilicity is an important feature of drug-like molecules that affects membrane permeability and solubility. The Lipophilicity dataset contains such molecules as well as their experimental octanol/water distribution. The BindingDB dataset contains experimentally determined protein-ligand binding affinities for numerous protein targets including isoforms and mutational variants Liu, Lin, et al. (2007). Affinity data across proteins is of key interest in computer-aided drug design for screening for drug candidates (Liu, Lin, et al. 2007). The QuanDB benchmark suite was designed for quantum chemical property prediction and materials discovery (Yang et al. 2024). It consists of a diverse array of organic molecular compounds wherein each molecule is small namely having less than 40 atoms. The Boiling Point dataset contains numerous molecules and their corresponding boiling point between $[-100°C, 100°C]$. The IC50 dataset contains small molecules and the concentration of those molecules which inhibits a biological

process by 50%. This quantity is crucial for assessing bioactivity Yang et al. 2024. This information is summarized along with the scale of each dataset and units in Table 1.

C Statistical Test Justification

Justification of test choice Outer-fold validation losses under nested CV are (near) unbiased for the post-selection performance but are statistically dependent: training subsets overlap and validation folds partition the same finite sample. There is, in general, no universally unbiased estimator of the variance of K-fold CV (Bengio and Grandvalet 2004). Consequently, fold-wise tests that ignore dependence such as Wilcoxon signed-rank, and classical paired t do not achieve nominal size (Ramachandran and Tsokos 2020). Moreover, Wilcoxon also presumes symmetry of paired differences (Voraprateep 2013). We deliberately avoid Friedman-type omnibus rank tests across folds, as these presume observations in different blocks are independent (Demšar 2006; Eisinga et al. 2017). This assumption does not apply to CV folds from a single dataset namely D_{trainval} (Demšar 2006).

D Benchmark Tables

Table 4: Detailed performance of PactNet on the QM9 dataset. Our model achieves the lowest validation and test RMSE and MAE, outperforming all other baselines. The units are in cal $\text{mol}^{-1}\text{K}^{-1}$.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
QM9	**PactNet (ECC)**	$\mathbf{0.999 \pm 0.099}$	$\mathbf{0.659 \pm 0.069}$	$\mathbf{1.0480 \pm 0.1805}$	$\mathbf{0.6510 \pm 0.0815}$
	GAT (ECFP)	2.490 ± 0.149	1.826 ± 0.051	2.4370 ± 0.2845	1.7870 ± 0.1620
	GAT (SELFIES)	1.898 ± 0.072	1.493 ± 0.048	1.8170 ± 0.1360	1.4210 ± 0.1135
	GAT (SMILES)	1.896 ± 0.071	1.488 ± 0.052	1.7820 ± 0.1350	1.4030 ± 0.1105
	GCN (ECFP)	2.483 ± 0.171	1.811 ± 0.085	2.4260 ± 0.2380	1.8390 ± 0.1615
	GCN (SELFIES)	2.358 ± 0.096	1.871 ± 0.078	2.1550 ± 0.1510	1.7270 ± 0.1290
	GCN (SMILES)	2.423 ± 0.063	1.926 ± 0.046	2.1260 ± 0.1575	1.6980 ± 0.1265
	GIN (ECFP)	2.583 ± 0.113	1.913 ± 0.066	2.5470 ± 0.2480	1.9270 ± 0.1660
	GIN (SELFIES)	1.883 ± 0.081	1.489 ± 0.068	1.8120 ± 0.1300	1.4350 ± 0.1060
	GIN (SMILES)	1.876 ± 0.063	1.492 ± 0.058	1.8120 ± 0.1395	1.4190 ± 0.1085
	SAGE (ECFP)	2.516 ± 0.165	1.836 ± 0.086	2.4860 ± 0.2580	1.8310 ± 0.1640
	SAGE (SELFIES)	1.561 ± 0.031	1.206 ± 0.029	1.4880 ± 0.1180	1.1560 ± 0.0915
	SAGE (SMILES)	1.552 ± 0.050	1.209 ± 0.034	1.4290 ± 0.1120	1.0890 ± 0.0905

Table 5: Detailed performance of PactNet on the ESOL dataset. Our model achieves the lowest validation and test RMSE and MAE, outperforming all other baselines. The units are in $\log$ mol/L.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
ESOL	**PactNet (ECC)**	$\mathbf{0.681 \pm 0.032}$	$\mathbf{0.508 \pm 0.026}$	$\mathbf{0.8290 \pm 0.1480}$	$\mathbf{0.5930 \pm 0.0725}$
	GAT (ECFP)	1.226 ± 0.070	0.929 ± 0.051	1.1730 ± 0.1300	0.8790 ± 0.1060
	GAT (SELFIES)	1.170 ± 0.117	0.902 ± 0.082	0.9850 ± 0.1210	0.7440 ± 0.0890
	GAT (SMILES)	1.085 ± 0.125	0.834 ± 0.087	1.1400 ± 0.1200	0.8710 ± 0.0945
	GCN (ECFP)	1.223 ± 0.057	0.932 ± 0.048	1.1710 ± 0.1220	0.8880 ± 0.0960
	GCN (SELFIES)	1.279 ± 0.114	0.977 ± 0.085	1.2750 ± 0.1530	0.9490 ± 0.1070
	GCN (SMILES)	1.240 ± 0.172	0.958 ± 0.125	1.2970 ± 0.1690	0.9460 ± 0.1125
	GIN (ECFP)	1.155 ± 0.051	0.878 ± 0.030	1.1070 ± 0.1190	0.8500 ± 0.0915
	GIN (SELFIES)	1.247 ± 0.172	0.940 ± 0.138	1.3940 ± 0.1890	0.9990 ± 0.1235
	GIN (SMILES)	1.196 ± 0.082	0.908 ± 0.068	1.3370 ± 0.1570	0.9960 ± 0.1170
	SAGE (ECFP)	1.218 ± 0.076	0.922 ± 0.058	1.1870 ± 0.1175	0.8960 ± 0.1040
	SAGE (SELFIES)	1.055 ± 0.116	0.802 ± 0.080	0.9970 ± 0.1215	0.7460 ± 0.0840
	SAGE (SMILES)	1.068 ± 0.113	0.819 ± 0.079	1.0890 ± 0.1380	0.8140 ± 0.0940

Table 6: Detailed performance of PactNet on the LIPOPHIL dataset. Our model achieves the lowest validation and test RMSE and MAE, outperforming all other baselines.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
LIPOPHIL	**PactNet (ECC)**	$\mathbf{0.717 \pm 0.027}$	$\mathbf{0.531 \pm 0.023}$	$\mathbf{0.7500 \pm 0.0505}$	$\mathbf{0.5460 \pm 0.0335}$
	GAT (ECFP)	0.875 ± 0.025	0.666 ± 0.021	0.8630 ± 0.0500	0.6580 ± 0.0385
	GAT (SELFIES)	1.037 ± 0.052	0.833 ± 0.044	1.0810 ± 0.0460	0.8810 ± 0.0415
	GAT (SMILES)	1.031 ± 0.044	0.825 ± 0.034	1.0690 ± 0.0505	0.8600 ± 0.0445
	GCN (ECFP)	0.866 ± 0.024	0.663 ± 0.021	0.8380 ± 0.0495	0.6380 ± 0.0360
	GCN (SELFIES)	1.070 ± 0.041	0.863 ± 0.033	1.1060 ± 0.0460	0.9060 ± 0.0420
	GCN (SMILES)	1.075 ± 0.045	0.869 ± 0.033	1.1100 ± 0.0470	0.9060 ± 0.0450
	GIN (ECFP)	0.829 ± 0.036	0.620 ± 0.036	0.8080 ± 0.0500	0.6050 ± 0.0355
	GIN (SELFIES)	1.062 ± 0.065	0.848 ± 0.054	1.0850 ± 0.0530	0.8710 ± 0.0430
	GIN (SMILES)	1.072 ± 0.043	0.862 ± 0.036	1.0730 ± 0.0495	0.8650 ± 0.0410
	SAGE (ECFP)	0.865 ± 0.021	0.663 ± 0.022	0.8500 ± 0.0480	0.6510 ± 0.0365
	SAGE (SELFIES)	0.947 ± 0.038	0.750 ± 0.027	1.0300 ± 0.0615	0.8120 ± 0.0435
	SAGE (SMILES)	0.955 ± 0.041	0.763 ± 0.035	1.0160 ± 0.0630	0.7970 ± 0.0435

Table 7: Detailed performance of PactNet on the FREESOLV dataset. Our model achieves the lowest validation and test RMSE and MAE, outperforming all other baselines.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
FREESOLV	**PactNet (ECC)**	$\mathbf{1.313 \pm 0.111}$	$\mathbf{0.857 \pm 0.064}$	$\mathbf{1.4390 \pm 0.4430}$	$\mathbf{0.8560 \pm 0.1925}$
	GAT (ECFP)	1.980 ± 0.227	1.324 ± 0.137	2.5360 ± 0.7505	1.7260 ± 0.3340
	GAT (SELFIES)	2.787 ± 0.361	2.059 ± 0.334	3.6720 ± 0.7910	2.7010 ± 0.4400
	GAT (SMILES)	2.777 ± 0.373	2.070 ± 0.305	3.7270 ± 0.8130	2.7220 ± 0.4425
	GCN (ECFP)	2.005 ± 0.238	1.309 ± 0.147	2.5380 ± 0.8070	1.6090 ± 0.3315
	GCN (SELFIES)	3.486 ± 0.216	2.546 ± 0.097	3.7730 ± 0.8485	2.7260 ± 0.4650
	GCN (SMILES)	3.381 ± 0.240	2.486 ± 0.184	3.8800 ± 0.8570	2.8550 ± 0.4605
	GIN (ECFP)	1.737 ± 0.128	1.180 ± 0.160	2.1720 ± 0.5940	1.4270 ± 0.2865
	GIN (SELFIES)	3.427 ± 0.170	2.553 ± 0.133	3.8140 ± 0.8350	2.7930 ± 0.4465
	GIN (SMILES)	3.454 ± 0.199	2.528 ± 0.123	3.6900 ± 0.7775	2.6760 ± 0.4545
	SAGE (ECFP)	1.894 ± 0.162	1.285 ± 0.150	2.3650 ± 0.6605	1.5960 ± 0.3140
	SAGE (SELFIES)	2.498 ± 0.429	1.851 ± 0.389	3.7790 ± 0.8365	2.7620 ± 0.4305
	SAGE (SMILES)	2.703 ± 0.378	2.029 ± 0.290	3.7890 ± 0.8225	2.8020 ± 0.4540

Table 8: Detailed performance of PactNet on the BOILINGPOINT dataset. Our model achieves the lowest validation and test RMSE and MAE, outperforming all other baselines.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
BOILINGPT	**PactNet (ECC)**	$\mathbf{49.089 \pm 2.425}$	$\mathbf{33.893 \pm 1.592}$	$\mathbf{50.6350 \pm 6.4695}$	$\mathbf{33.2600 \pm 3.6895}$
	GAT (ECFP)	56.831 ± 1.794	43.534 ± 1.438	53.2660 ± 4.5465	40.3050 ± 3.2910
	GAT (SELFIES)	55.602 ± 1.883	42.244 ± 1.894	56.0480 ± 5.4735	40.4770 ± 3.9400
	GAT (SMILES)	55.240 ± 2.739	41.337 ± 2.469	60.7780 ± 5.1350	45.3330 ± 4.0420
	GCN (ECFP)	57.010 ± 1.529	43.315 ± 1.254	54.6140 ± 4.5240	41.7490 ± 3.6545
	GCN (SELFIES)	61.200 ± 1.848	47.294 ± 1.035	62.3820 ± 5.2935	46.8430 ± 4.1720
	GCN (SMILES)	61.586 ± 1.292	47.662 ± 1.277	59.0360 ± 4.8840	43.9270 ± 3.7900
	GIN (ECFP)	58.744 ± 1.789	45.098 ± 2.000	57.3840 ± 5.0375	43.2210 ± 3.6790
	GIN (SELFIES)	58.629 ± 0.867	44.500 ± 0.975	57.8800 ± 4.8570	42.4450 ± 3.8065
	GIN (SMILES)	59.282 ± 2.271	45.235 ± 2.034	57.1610 ± 4.9560	42.1930 ± 3.7810
	SAGE (ECFP)	57.161 ± 1.253	43.714 ± 0.873	55.2450 ± 4.4265	42.1960 ± 3.4290
	SAGE (SELFIES)	54.345 ± 2.264	40.948 ± 1.797	53.0400 ± 5.7445	37.4580 ± 3.5905
	SAGE (SMILES)	54.284 ± 3.661	40.291 ± 3.086	53.0110 ± 5.5890	36.5630 ± 3.7300

Table 9: Detailed performance of PactNet on the IC50 dataset. Our model achieves competitive validation and test RMSE and MAE, outperforming all other baselines.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
IC50	**PactNet (ECC)**	$\mathbf{0.756 \pm 0.037}$	0.596 ± 0.018	0.7500 ± 0.0640	0.6060 ± 0.0430
	GAT (ECFP)	0.768 ± 0.038	0.596 ± 0.016	0.7570 ± 0.0730	0.5890 ± 0.0475
	GAT (SELFIES)	0.781 ± 0.050	0.609 ± 0.020	0.7400 ± 0.0670	0.5880 ± 0.0435
	GAT (SMILES)	0.782 ± 0.049	0.608 ± 0.019	0.7410 ± 0.0710	0.5810 ± 0.0450
	GCN (ECFP)	0.764 ± 0.036	0.596 ± 0.019	0.7720 ± 0.0735	0.5950 ± 0.0505
	GCN (SELFIES)	0.782 ± 0.051	0.607 ± 0.022	0.7450 ± 0.0695	0.5890 ± 0.0435
	GCN (SMILES)	0.782 ± 0.051	0.607 ± 0.020	0.7410 ± 0.0695	0.5860 ± 0.0435
	GIN (ECFP)	0.782 ± 0.035	0.605 ± 0.013	0.7640 ± 0.0735	0.5930 ± 0.0485
	GIN (SELFIES)	0.783 ± 0.051	0.610 ± 0.020	0.7410 ± 0.0680	0.5900 ± 0.0450
	GIN (SMILES)	0.783 ± 0.051	0.609 ± 0.020	0.7400 ± 0.0735	0.5860 ± 0.0440
	SAGE (ECFP)	0.764 ± 0.036	$\mathbf{0.592 \pm 0.014}$	0.7840 ± 0.0775	0.6040 ± 0.0485
	SAGE (SELFIES)	0.782 ± 0.051	0.610 ± 0.021	$\mathbf{0.7360 \pm 0.0700}$	0.5840 ± 0.0450
	SAGE (SMILES)	0.782 ± 0.050	0.608 ± 0.020	0.7420 ± 0.0735	$\mathbf{0.5820 \pm 0.0445}$

Table 10: Detailed performance of PactNet on the BINDINGDB dataset. Our model achieves competitive validation and test RMSE and MAE, outperforming all other baselines.

Dataset	Model (Rep.)	Val RMSE	Val MAE	Test RMSE	Test MAE
BINDINGDB	**PactNet (ECC)**	$\mathbf{1.479 \pm 0.034}$	$\mathbf{1.211 \pm 0.030}$	1.7710 ± 0.2420	1.3650 ± 0.1080
	GAT (ECFP)	1.512 ± 0.055	1.235 ± 0.053	1.7750 ± 0.2480	1.3360 ± 0.1105
	GAT (SELFIES)	1.488 ± 0.029	1.227 ± 0.023	1.7820 ± 0.2440	1.3600 ± 0.1145
	GAT (SMILES)	1.491 ± 0.028	1.230 ± 0.020	1.7540 ± 0.2395	1.3560 ± 0.1100
	GCN (ECFP)	1.508 ± 0.058	1.231 ± 0.056	1.7620 ± 0.2620	1.3220 ± 0.1125
	GCN (SELFIES)	1.492 ± 0.031	1.231 ± 0.023	1.7700 ± 0.2485	1.3590 ± 0.1110
	GCN (SMILES)	1.495 ± 0.031	1.234 ± 0.023	1.8240 ± 0.2750	1.3740 ± 0.1210
	GIN (ECFP)	1.520 ± 0.050	1.253 ± 0.044	1.7830 ± 0.2580	1.3540 ± 0.1135
	GIN (SELFIES)	1.494 ± 0.030	1.231 ± 0.025	1.7550 ± 0.2410	1.3450 ± 0.1095
	GIN (SMILES)	1.494 ± 0.032	1.230 ± 0.027	$\mathbf{1.7440 \pm 0.2420}$	$\mathbf{1.3390 \pm 0.1070}$
	SAGE (ECFP)	1.506 ± 0.054	1.235 ± 0.046	1.7720 ± 0.2610	1.3450 ± 0.1140
	SAGE (SELFIES)	1.493 ± 0.031	1.234 ± 0.024	1.9120 ± 0.3785	1.3760 ± 0.1220
	SAGE (SMILES)	1.493 ± 0.030	1.232 ± 0.024	1.7870 ± 0.2450	1.3810 ± 0.1105

E Statistical Summary Tables

E.1 QM9

Table 11: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `QM9`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gcn_selfies	1.212	48.842	1.143	1.281	6e-06
pactnet vs gcn_smiles	1.267	25.966	1.132	1.403	7.2e-05
pactnet vs gin_ecfp	1.254	22.974	1.103	1.406	0.000106
pactnet vs gat_smiles	0.829	22.089	0.725	0.934	0.000112
pactnet vs gat_ecfp	1.167	20.716	1.011	1.324	0.000125
pactnet vs gin_selfies	0.831	20.866	0.720	0.941	0.000125
pactnet vs gin_smiles	0.834	19.948	0.718	0.950	0.000125
pactnet vs sage_ecfp	1.177	17.804	0.993	1.361	0.000146
pactnet vs gcn_ecfp	1.153	15.862	0.951	1.354	0.000185
pactnet vs gat_selfies	0.835	13.358	0.661	1.008	0.000272
pactnet vs sage_selfies	0.548	9.634	0.390	0.706	0.000649
pactnet vs sage_smiles	0.551	7.743	0.353	0.748	0.000749

Table 12: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `QM9`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gcn_smiles	1.424	30.029	1.292	1.556	4.4e-05
pactnet vs gcn_selfies	1.359	19.070	1.161	1.556	0.000245
pactnet vs gin_ecfp	1.584	16.790	1.322	1.846	0.000369
pactnet vs gat_ecfp	1.491	16.167	1.235	1.747	0.000385
pactnet vs sage_ecfp	1.516	14.189	1.220	1.813	0.000573
pactnet vs gcn_ecfp	1.484	12.955	1.166	1.802	0.000717
pactnet vs gat_smiles	0.897	12.286	0.694	1.100	0.000756
pactnet vs gin_selfies	0.884	11.020	0.661	1.107	0.000964
pactnet vs gin_smiles	0.877	10.341	0.641	1.112	0.000987
pactnet vs gat_selfies	0.899	8.979	0.621	1.177	0.00128
pactnet vs sage_selfies	0.562	6.329	0.315	0.808	0.00319
pactnet vs sage_smiles	0.553	6.171	0.304	0.802	0.00319

E.2 Lipophilicity

Table 13: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `lipophil`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gcn_selfies	0.332	35.043	0.305	0.358	2.4e-05
pactnet vs gcn_smiles	0.338	32.601	0.309	0.366	2.9e-05
pactnet vs sage_selfies	0.219	15.687	0.180	0.257	0.000482
pactnet vs gin_smiles	0.331	15.062	0.270	0.392	0.00051
pactnet vs gat_selfies	0.301	14.563	0.244	0.359	0.000517
pactnet vs sage_smiles	0.232	13.138	0.183	0.281	0.000678
pactnet vs gat_smiles	0.293	11.414	0.222	0.365	0.00101
pactnet vs gcn_ecfp	0.132	10.909	0.098	0.165	0.00101
pactnet vs sage_ecfp	0.132	10.861	0.098	0.166	0.00101
pactnet vs gat_ecfp	0.134	9.321	0.094	0.174	0.00111
pactnet vs gin_selfies	0.317	7.518	0.200	0.434	0.00168
pactnet vs gin_ecfp	0.089	3.640	0.021	0.157	0.011

Table 14: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `lipophil`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gcn_selfies	0.353	20.048	0.304	0.402	0.000219
pactnet vs gcn_smiles	0.358	17.665	0.302	0.415	0.000332
pactnet vs gin_smiles	0.355	13.628	0.283	0.427	0.000839
pactnet vs gat_selfies	0.320	11.120	0.240	0.400	0.00167
pactnet vs gat_ecfp	0.158	8.365	0.106	0.210	0.00259
pactnet vs gat_smiles	0.314	8.565	0.212	0.416	0.00259
pactnet vs gcn_ecfp	0.149	8.517	0.100	0.197	0.00259
pactnet vs gin_selfies	0.345	6.855	0.205	0.485	0.00259
pactnet vs sage_ecfp	0.148	9.091	0.103	0.193	0.00259
pactnet vs sage_selfies	0.230	9.643	0.164	0.296	0.00259
pactnet vs sage_smiles	0.238	9.620	0.170	0.307	0.00259
pactnet vs gin_ecfp	0.112	4.077	0.036	0.189	0.00757

E.3 FreeSolv

Table 15: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `freesolv`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gin_smiles	1.671	19.269	1.430	1.912	0.000256
pactnet vs gcn_selfies	1.690	16.561	1.406	1.973	0.000428
pactnet vs gin_selfies	1.697	15.350	1.390	2.004	0.000525
pactnet vs gcn_smiles	1.629	11.909	1.249	2.009	0.00128
pactnet vs gat_ecfp	0.467	7.425	0.292	0.641	0.00702
pactnet vs sage_smiles	1.173	6.812	0.695	1.651	0.0085
pactnet vs gat_smiles	1.213	6.224	0.672	1.754	0.00931
pactnet vs gcn_ecfp	0.452	6.376	0.255	0.649	0.00931
pactnet vs gat_selfies	1.202	5.427	0.587	1.817	0.0112
pactnet vs sage_ecfp	0.428	4.617	0.171	0.686	0.0149
pactnet vs sage_selfies	0.995	3.864	0.280	1.709	0.0181
pactnet vs gin_ecfp	0.323	2.856	0.009	0.637	0.0231

Table 16: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `freesolv`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gin_selfies	2.113	11.592	1.607	2.619	0.0019
pactnet vs gcn_selfies	2.172	11.171	1.632	2.712	0.00193
pactnet vs gin_smiles	2.141	11.296	1.615	2.667	0.00193
pactnet vs sage_ecfp	0.581	10.341	0.425	0.737	0.00222
pactnet vs gcn_smiles	2.067	9.301	1.450	2.684	0.00297
pactnet vs gat_ecfp	0.667	4.538	0.259	1.075	0.0146
pactnet vs gat_selfies	1.474	5.882	0.778	2.169	0.0146
pactnet vs gat_smiles	1.463	5.799	0.763	2.164	0.0146
pactnet vs gcn_ecfp	0.691	5.112	0.316	1.067	0.0146
pactnet vs gin_ecfp	0.424	5.376	0.205	0.643	0.0146
pactnet vs sage_selfies	1.185	4.426	0.442	1.928	0.0146
pactnet vs sage_smiles	1.390	5.690	0.712	2.068	0.0146

E.4 ESOL

Table 17: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `ESOL`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gin_ecfp	0.370	46.410	0.348	0.392	8e-06
pactnet vs gat_ecfp	0.421	17.654	0.355	0.487	0.000333
pactnet vs gcn_ecfp	0.424	16.532	0.353	0.495	0.000392
pactnet vs sage_ecfp	0.414	16.233	0.343	0.484	0.000392
pactnet vs gat_selfies	0.393	7.609	0.250	0.537	0.00569
pactnet vs gcn_selfies	0.469	7.753	0.301	0.637	0.00569
pactnet vs gin_smiles	0.399	7.851	0.258	0.541	0.00569
pactnet vs sage_smiles	0.310	7.347	0.193	0.427	0.00569
pactnet vs gat_smiles	0.326	6.032	0.176	0.476	0.00762
pactnet vs gcn_smiles	0.449	5.213	0.210	0.689	0.00969
pactnet vs gin_selfies	0.431	4.552	0.168	0.694	0.00969
pactnet vs sage_selfies	0.293	5.035	0.132	0.455	0.00969

Table 18: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset ESOL; control pactnet. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gin_ecfp	0.475	19.913	0.409	0.541	0.000225
pactnet vs gcn_ecfp	0.543	17.259	0.456	0.630	0.000364
pactnet vs gat_ecfp	0.545	14.954	0.444	0.646	0.000583
pactnet vs sage_ecfp	0.538	13.238	0.425	0.650	0.000847
pactnet vs gin_smiles	0.515	10.397	0.377	0.652	0.00193
pactnet vs gcn_selfies	0.598	7.790	0.385	0.811	0.00513
pactnet vs gat_selfies	0.490	6.728	0.288	0.692	0.00763
pactnet vs sage_smiles	0.388	5.838	0.203	0.572	0.0107
pactnet vs gat_smiles	0.404	5.405	0.196	0.612	0.0113
pactnet vs gcn_smiles	0.559	5.047	0.252	0.867	0.0113
pactnet vs gin_selfies	0.566	4.778	0.237	0.896	0.0113
pactnet vs sage_selfies	0.374	4.407	0.138	0.609	0.0113

E.5 BoilingPoint

Table 19: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset boilingpoint; control pactnet. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gcn_selfies	13.401	9.426	9.454	17.348	0.00424
pactnet vs gcn_smiles	13.768	6.842	8.181	19.356	0.0131
pactnet vs gin_ecfp	11.204	6.414	6.355	16.054	0.0131
pactnet vs gin_selfies	10.606	6.218	5.871	15.342	0.0131
pactnet vs gin_smiles	11.342	6.554	6.537	16.147	0.0131
pactnet vs sage_ecfp	9.821	6.713	5.759	13.883	0.0131
pactnet vs gcn_ecfp	9.421	5.680	4.816	14.027	0.0142
pactnet vs gat_ecfp	9.641	5.181	4.474	14.807	0.0165
pactnet vs sage_selfies	7.055	4.786	2.962	11.147	0.0175
pactnet vs gat_selfies	8.351	3.356	1.441	15.260	0.0426
pactnet vs gat_smiles	7.444	3.000	0.554	14.335	0.0426
pactnet vs sage_smiles	6.398	2.356	-1.141	13.937	0.0426

Table 20: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset boilingpoint; control pactnet. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gcn_selfies	12.111	7.335	7.527	16.695	0.011
pactnet vs gcn_smiles	12.497	6.176	6.879	18.114	0.0192
pactnet vs gin_ecfp	9.655	5.861	5.081	14.229	0.0211
pactnet vs sage_ecfp	8.072	5.138	3.710	12.433	0.0306
pactnet vs gcn_ecfp	7.921	4.721	3.263	12.579	0.0366
pactnet vs gat_ecfp	7.742	4.407	2.865	12.619	0.0407
pactnet vs gin_selfies	9.540	4.336	3.431	15.649	0.0407
pactnet vs gin_smiles	10.193	3.937	3.004	17.381	0.0425
pactnet vs gat_selfies	6.513	3.322	1.069	11.956	0.0587
pactnet vs gat_smiles	6.151	2.813	0.081	12.222	0.0722
pactnet vs sage_selfies	5.256	2.449	-0.703	11.215	0.0722
pactnet vs sage_smiles	5.195	1.562	-4.040	14.429	0.0967

E.6 BindingDB

Table 21: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `bindingdb`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs sage_smiles	0.021	3.758	0.005	0.036	0.119
pactnet vs sage_selfies	0.022	3.236	0.003	0.041	0.175
pactnet vs gat_selfies	0.015	2.085	-0.005	0.035	0.386
pactnet vs gcn_selfies	0.020	1.929	-0.009	0.049	0.386
pactnet vs gcn_smiles	0.023	2.334	-0.004	0.050	0.386
pactnet vs gin_selfies	0.020	2.365	-0.003	0.043	0.386
pactnet vs gin_smiles	0.019	2.214	-0.005	0.042	0.386
pactnet vs gat_smiles	0.018	1.546	-0.015	0.051	0.425
pactnet vs gin_ecfp	0.042	1.671	-0.028	0.112	0.425
pactnet vs gat_ecfp	0.023	0.918	-0.048	0.094	0.431
pactnet vs gcn_ecfp	0.019	0.743	-0.053	0.092	0.431
pactnet vs sage_ecfp	0.024	1.227	-0.030	0.077	0.431

Table 22: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `bindingdb`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gat_ecfp	0.033	1.563	-0.026	0.093	0.825
pactnet vs gat_selfies	0.009	0.737	-0.025	0.043	0.825
pactnet vs gat_smiles	0.013	0.937	-0.025	0.050	0.825
pactnet vs gcn_ecfp	0.029	1.386	-0.030	0.088	0.825
pactnet vs gcn_selfies	0.013	1.029	-0.023	0.049	0.825
pactnet vs gcn_smiles	0.016	1.282	-0.018	0.050	0.825
pactnet vs gin_ecfp	0.041	1.853	-0.021	0.103	0.825
pactnet vs gin_selfies	0.016	1.775	-0.009	0.040	0.825
pactnet vs gin_smiles	0.015	1.620	-0.011	0.042	0.825
pactnet vs sage_ecfp	0.027	1.377	-0.027	0.081	0.825
pactnet vs sage_selfies	0.015	1.480	-0.013	0.042	0.825
pactnet vs sage_smiles	0.014	1.785	-0.008	0.037	0.825

E.7 IC50

Table 23: MAE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `ic50`; control `pactnet`. Positive Δ (competitor $-$ control) favors control. Holm controls FWER.

Comparison	ΔMAE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gat_ecfp	-0.000	-0.023	-0.037	0.036	1
pactnet vs gat_selfies	0.013	1.179	-0.017	0.043	1
pactnet vs gat_smiles	0.012	1.159	-0.017	0.041	1
pactnet vs gcn_ecfp	-0.000	-0.008	-0.023	0.023	1
pactnet vs gcn_selfies	0.011	0.840	-0.024	0.046	1
pactnet vs gcn_smiles	0.011	1.028	-0.019	0.041	1
pactnet vs gin_ecfp	0.009	0.825	-0.021	0.038	1
pactnet vs gin_selfies	0.014	1.507	-0.012	0.039	1
pactnet vs gin_smiles	0.013	1.296	-0.014	0.040	1
pactnet vs sage_ecfp	-0.004	-0.463	-0.030	0.021	1
pactnet vs sage_selfies	0.014	1.206	-0.018	0.045	1
pactnet vs sage_smiles	0.012	1.050	-0.020	0.043	1

Table 24: RMSE: NB-corrected one-sided tests (outer folds, $K = 5$) on dataset `ic50`; control `pactnet`. Positive Δ (competitor − control) favors control. Holm controls FWER.

Comparison	ΔRMSE	t_{NB}	CI_{low}	CI_{high}	p_{Holm}
pactnet vs gin_ecfp	0.026	5.608	0.013	0.039	0.0298
pactnet vs gat_ecfp	0.012	1.099	-0.018	0.042	1
pactnet vs gat_selfies	0.025	1.203	-0.032	0.081	1
pactnet vs gat_smiles	0.025	1.238	-0.031	0.082	1
pactnet vs gcn_ecfp	0.007	0.693	-0.022	0.037	1
pactnet vs gcn_selfies	0.026	1.277	-0.031	0.083	1
pactnet vs gcn_smiles	0.026	1.282	-0.030	0.082	1
pactnet vs gin_selfies	0.027	1.374	-0.027	0.081	1
pactnet vs gin_smiles	0.027	1.338	-0.029	0.082	1
pactnet vs sage_ecfp	0.007	0.789	-0.019	0.034	1
pactnet vs sage_selfies	0.026	1.239	-0.032	0.085	1
pactnet vs sage_smiles	0.026	1.244	-0.031	0.082	1

F Compute Cost

All experiments were performed using an AWS M8G.4XLARGE (CPU) instance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide all statistical evidence and rigorous experimental design to demonstrate all performance based claims. Claims of topological feature integration are justified and provided in the methods section. Our learning method is computationally efficient and our GNN is novel.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See discussion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No explicit proofs provided.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Full experimental algorithm provided in the appendix and discussed extensively in the main experiment section. Additionally, all code is open source.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Open sourced code and all datasets. See github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Absolutely, that and more see statistical analysis section of results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the appendix yes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We were ethical undoubtedly and have read the guidelines.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: No societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No such risk posed.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Absolutely.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is open source and clearly documented with readme files.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: No human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Research does not involve LLMs in any way/shape/form.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.