Advancements in Ligand-Based Virtual Screening through the Synergistic Integration of Graph Neural Networks and Expert-Crafted Descriptors

Yunchao (Lance) Liu

Department of Computer Science, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
yunchao.liu@vanderbilt.edu

Rocco Moretti

Department of Chemistry, Center for Structural Biology, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
rocco.moretti@vanderbilt.edu

Yu Wang

Department of Computer Science, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
yu.wang.1@vanderbilt.edu

Ha Dong

Department of Neural Science, Amherst College, 220 South Pleasant Street
Amherst, Massachusetts 01002, USA
hdong26@amherst.edu

Bailu Yan

Department of Biostatistics, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
bailu.yan@vanderbilt.edu

Bobby Bodenheimer

Department of Computer Science, Electrical Engineering and Computer Engineering, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
bobby.bodenheimer@vanderbilt.edu

Tyler Derr*

Department of Computer Science, Data Science Institute, Data Science Institute, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
tyler.derr@vanderbilt.edu

Jens Meiler*

Department of Chemistry, Center for Structural Biology, Vanderbilt University, 2201 West End Ave
Nashville, Tennessee 37235, USA
Institute of Drug Discovery, Leipzig University Medical School, Härtelstraße 16-18,
Leipzig, 04103, Germany
Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Humboldtstraße 25,
Leipzig, 04105, Germany
jens.meiler@vanderbilt.edu

* Corresponding Authors

The fusion of traditional chemical descriptors with Graph Neural Networks (GNNs) offers a compelling strategy for enhancing ligand-based virtual screening methodologies. A comprehensive evaluation revealed that the benefits derived from this integrative strategy vary significantly among different GNNs. Specifically, while GCN and SchNet demonstrate pronounced improvements by incorporating descriptors, SphereNet exhibits only marginal enhancement. Intriguingly, despite SphereNet's modest gain, all three models-GCN, SchNet, and SphereNet-achieve comparable performance levels when leveraging this combination strategy. This observation underscores a pivotal insight: sophisticated GNN architectures may be substituted with simpler counterparts without sacrificing efficacy, provided that they are augmented with descriptors. Furthermore, our analysis reveals a set of expert-crafted descriptors' robustness in scaffold-split scenarios, frequently outperforming the combined GNN-descriptor models. Given the critical importance of scaffold splitting in accurately mimicking real-world drug discovery contexts, this finding accentuates an imperative for GNN researchers to innovate models that can adeptly navigate and predict within such frameworks. Our work not only validates the potential of integrating descriptors with GNNs in advancing ligand-based virtual screening but also illuminates pathways for future enhancements in model development and application. Our implementation can be found at https://github.com/meilerlab/gnn-descriptor.

Keywords: artificial intelligence; graph neural network; virtual screening

1. Introduction

Virtual screening is a major way to supplement traditional high-throughput screening (HTS) for cost and time efficient drug discovery[1]. Two major branches of virtual screening exist: ligand-based, and structure-based. For the application of structure-based methods, detailed knowledge of the target's structure is essential, typically acquired through experimental methods such as X-ray crystallography or nuclear magnetic resonance (NMR). In cases where experimental data is lacking, computational predictions like homology modeling are employed to infer the three-dimensional configurations of targets. Recently, there are many AI-driven protein structure prediction tools available as well, such as AlphaFold [2], RosettaFold[3, 4], ESMFold[5].

This work focuses on the ligand-based method, for situations where the target structure remains unknown or cannot be computationally predicted. These methods depend on the knowledge of previously identified active compounds that bind to the target, leveraging this information to identify new potential drugs[6]. Even in the age that computational protein structure prediction tools are available, ligand-based approaches are needed for several reasons. First, while structure prediction tools have made remarkable progress, there are still limitations in their ability to accurately predict all protein structures, especially for proteins with highly dynamic regions and transient conformations. The ligand-based method does not require structural information, making it valuable for targets where high-quality structures are not available. Secondly, ligand-based methods can sometimes be faster and less resource-intensive than structure-based methods, especially in the early stages of drug discovery. They allow researchers to quickly screen vast chemical spaces or compound libraries to identify potential hits without detailed structural information.

Thirdly, some targets have multiple or flexible binding sites that can be challenging to characterize with structure-based methods alone. Ligand-based methods can help identify ligands that interact with such targets by leveraging data from known active compounds without relying on a fixed 3D structure.

Meanwhile, numerous studies applied GNNs to molecule-related tasks, given the intrinsic graph nature of molecules [7-13]. While some of those tasks achieve good results, several factors still make GNN for molecule representation learning challenging. First, data available for training in drug discovery campaigns is usually limited due to the high cost of experimental assays. Secondly, GNNs typically have difficulty learning molecular-level features due to their limited receptive field or learning non-additive molecular-level features such as total polar surface area. Thirdly, GNN intrinsically suffers from problems such as over-smoothing [14] and over-squashing [15] that introduce information loss in obtaining the global learned embedding from the atomic features.

As a solution, integrating the expert knowledge in the GNN workflow has become a new trend [16]. Expert knowledge can help supplement the data-hungry GNNs with prior knowledge to increase data efficiency and overcome intrinsic GNN shortcomings. One of the simplest ways to integrate expert knowledge is to combine the expert-crafted descriptors with GNN-learned representation through concatenation [17, 18]. However, while commonly used, a thorough evaluation of this concatenation strategy is lacking.

This work contributes to the field by comprehensively evaluating this commonly used strategy in a virtual screening setting using nine well-curated HTS datasets. We find that although this strategy is often effective, it is not always the case. Additionally, we discover that the combined GNNs show convergence of performance metrics, suggesting the potential interchangeability of sophisticated GNN architectures with simpler counterparts under this integrative strategy. Moreover, surprisingly we found that descriptors are fairly robust under the scaffold split scenario, which is often a more realistic setting in a drug discovery campaign. These findings prompt the need to examine the current integration strategies to understand their limitations, find better ways to integrate domain expert knowledge and provide a path for more advanced ligand-based virtual screening.

2. Results

2.1. *Concatenate descriptors with GNNs*

As shown in Figure 1, the concatenation strategy [17, 18] examined in this work proposes to train a neural network to predict activity by combining a GNN-derived molecular representation with the expert-crafted descriptors [19]. Specifically, for a representation h from the GNN, it is concatenated with the descriptor h_{dp}.

$$h = GNN(m)$$
$$\hat{p} = f([h||h_{dp}])$$

where m is the input molecular graph and, h_{dp} is a descriptor. $f(\cdot)$ is a classifier, usually a Multi-Layer-Perceptron (MLP). $\hat{p}$ is the predicted activity.

The model is trained by optimizing the binary cross entropy loss L:

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{p}) + (1 - y_i)\log(1 - \hat{p})$$

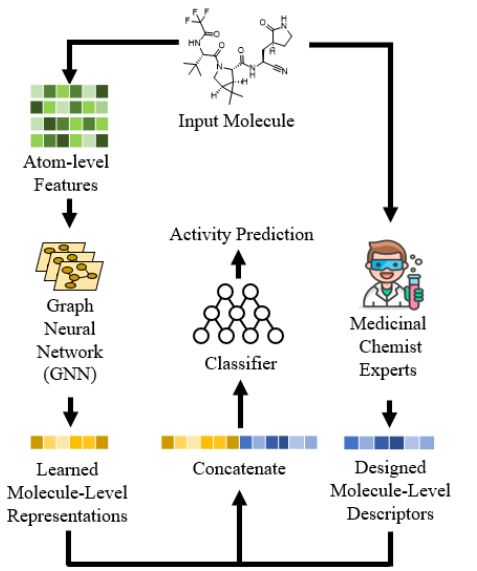


Figure 1. Overview of the investigated method. The learned molecular representation of GNN is concatenated with expert-crafted descriptors to enhance the predictive power.

In this work, we used three GNN models in our experiments: GCN [20], SchNet [11] and SphereNet [13]. We used the BioChemical Library (BCL) [21] to generate descriptors.

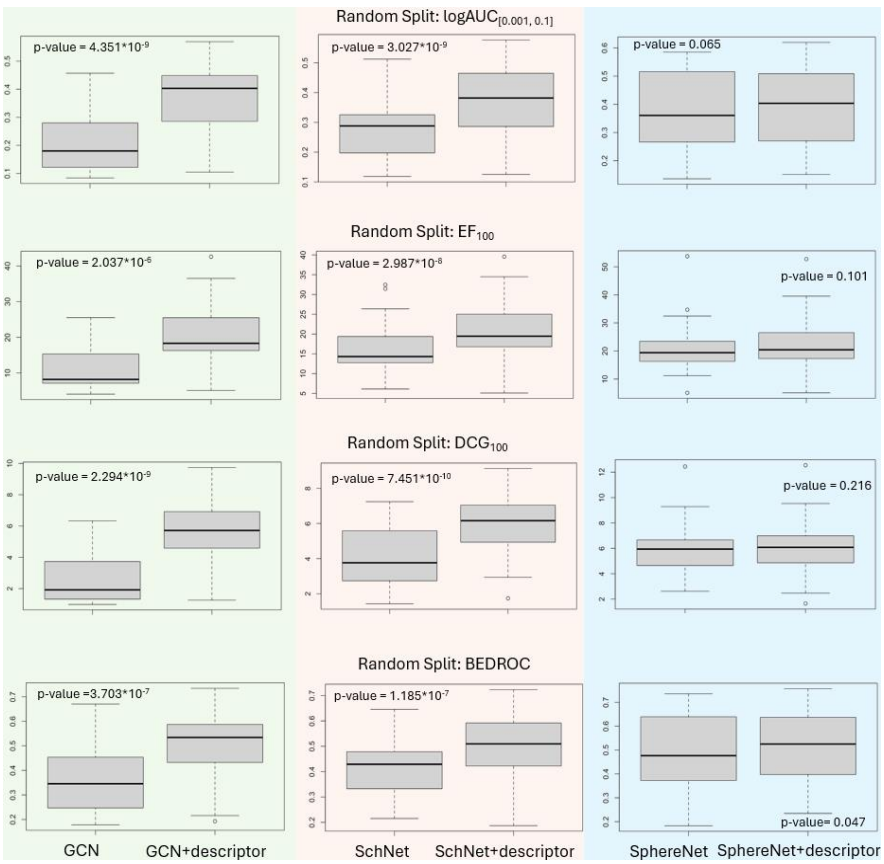


Figure 2. Random Split: Performance of three GNNs with their corresponding descriptor-integrated counterpart.

where n is the number of samples in a batch, and y_i is the experimentally determined active/inactive status of the *i*-th molecule.

2.2. *Effectiveness of the Concatenation Strategy Varies for GNNs with Random Split*

In Figure 2 the boxplots of model performances evaluated using four different metrics are shown (Experiments are detailed in Section 3 Method). The p-value is calculated using paired t-test [22].

The significant improvements observed in both GCN and SchNet models across four evaluation metrics highlight the investigated strategy's potential to facilitate the

identification of bioactive compounds in drug discovery. Although the benefits were less pronounced for the SphereNet model (as a bigger p-value is observed), the overall results advocate for the integration strategy's adoption as a valuable tool in computational chemistry.

There are three rationales for this approach. First, data available for training in drug discovery campaigns is usually limited due to the high cost of experimental assays. The expert-crafted descriptors supplement GNNs with prior knowledge, i.e., descriptors that worked well in virtual screening in the past, which reduces the need for GNNs to learn that knowledge from a large amount of data. Secondly, GNNs typically have difficulty learning molecular-level features due to their limited receptive field or learning non-additive molecular-level features such as total polar surface area. On the other hand, molecular-level descriptors provide global features directly. Thirdly, GNN intrinsically suffers from problems such as over-smoothing [14] and overs-quashing [15] that introduce information loss in obtaining the global learned embedding from the atomic features. Meanwhile, the descriptors extract the molecular features directly and circumvent information loss, complementing GNN-learned embeddings.

2.3. *All Descriptor-integrated GNNs Converge to Similar Performance with Random Split*

The analysis undertaken in this study revealed significant insight regarding the investigated strategy’s performance. Initially, the GNNs—each with its intrinsic computational complexities and capabilities—demonstrated disparate levels of efficacy. However, upon the integration of descriptors, a notable convergence in their performance metrics was observed, spanning all four evaluated metrics. As shown in Figure 2, SphereNet and SchNet, are more advanced GNNs compared with GCN. Yet, when these advanced GNNs were coupled with descriptors, the resultant performance was not just enhanced but aligned closely with that of their simpler counterparts GCN.

This intriguing outcome underscores the potency of the integration strategy in equalizing the performance landscape among GNN architectures. By integrating expert-crafted descriptors through the integration approach, even less complex GNN models could elevate their predictive accuracies to levels akin to those of more complex GNNs. Essentially, the integration strategy acts as a performance catalyst, diminishing the gaps between GNN models of varying complexities and facilitating a more uniform field of competition. Such findings highlight the potential of combining deep learning techniques with established domain knowledge, suggesting a reevaluation of the necessity for complex GNNs in scenarios where their simpler counterparts can achieve comparable outcomes through integration with descriptors.

2.4. *Expert-crafted Descriptor Still Outperforms Most GNNs Using Scaffold Split*

Besides random split, we also conducted experiments on scaffold split. This is a realistic scenario because medicinal chemists often need to determine the activity of structures substantially different from those in the known training set. They seek these structural differences for various reasons, such as avoiding patented structures, finding simpler synthetic routes, improving compound properties etc [23].

As expected, the overall performance under the scaffold split decreased compared with that under the random split. This decrease is due to the greater difficulty in predicting the performance of structures significantly different from the training set, as the data distribution differs between training and testing. However, as shown in Figure 3, the results from the scaffold split evaluation solidify the potential of the integration strategy in enhancing the performance of various GNN architectures for ligand-based virtual screening. The combined GNN-derived molecular representations with descriptors, improve the identification and prioritization of active compounds (Although outliers exist, which is consistent with our results for random split that the effectiveness of this strategy varies).

Most interestingly, we found that the descriptors alone outperform many GNNs. In some cases, it even outperforms the integrated-version GNNs. We hypothesize that this could result from the fact that deep learning-based methods are more easily overfit to the training data and therefore will perform worse than the expert-crafted ones when the data distribution is shifted. This finding prompts us to reconsider whether data-driven methods alone, despite their growing popularity, are the best approach for real-world drug discovery

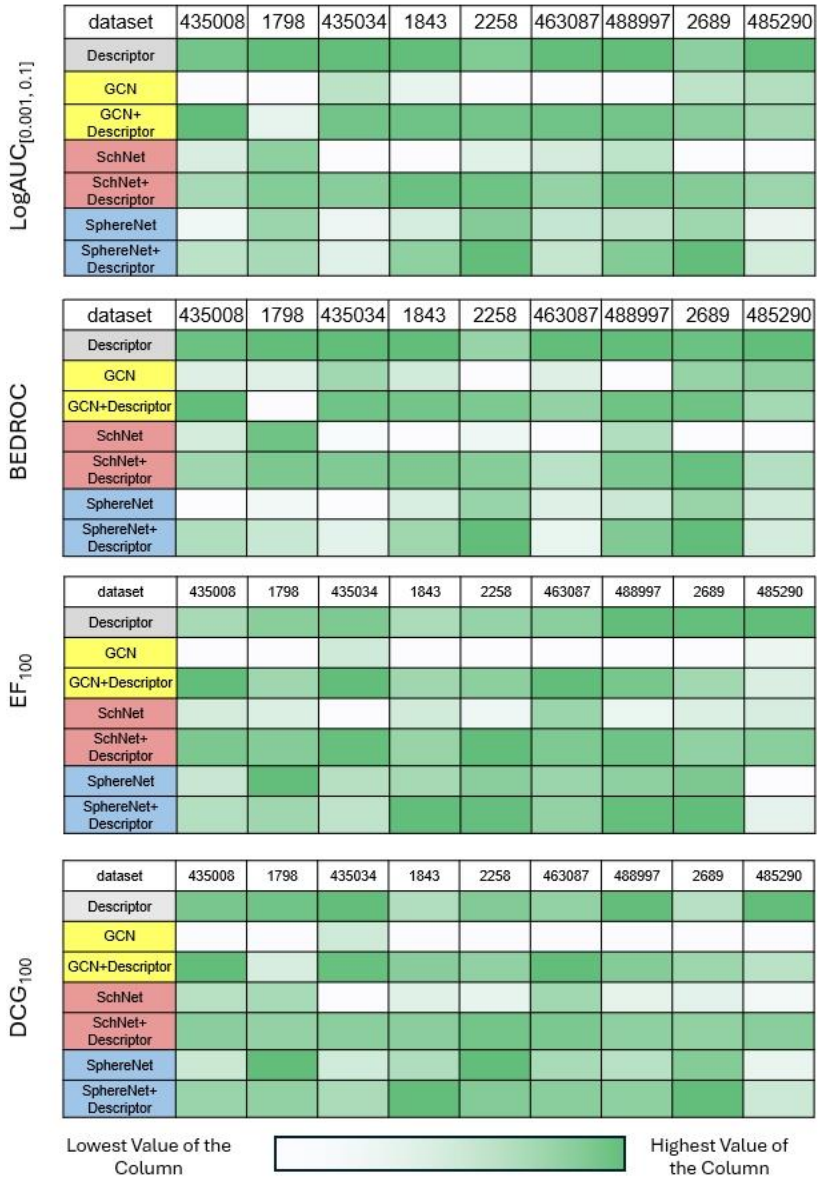


Figure 3. Scaffold Split: Performance of different models. The concatenation strategy still enhances the GNNs for most cases. Notably, descriptors perform better than many models across different metrics, especially salient in logAUC$_{[0.001, 0.1]}$ and BEDROC.

campaigns. Moreover, this also shows that even when coupled with descriptors, the performance of the integrated model may decrease and not always offer benefits. Finally, this finding emphasizes the need for developing better frameworks that integrate domain knowledge for improved predicted power under scaffold split scenarios.

3. Method

3.1. *Datasets*

We validate the effectiveness of the proposed strategy via nine well-curated high-throughput screening (HTS) datasets. To avoid issues with experimental artifacts and high false positive rates [24], for the validation of our strategy, we chose datasets carefully curated [25] from high throughput screens in the PubChem database [26]. Only datasets with robust secondary validation of compounds were considered. Datasets details are shown in Table 1.

SMILES from the datasets were converted to SDF files using Open Babel [27]. Standardized 3D coordinates are generated using Corina [28]. Molecules are further filtered with atom type validity and duplicates with the BioChemical Library (BCL) [21].

Random split is used for the experiments, and each dataset is split into 80% for training and 20% for testing. Because preliminary results and previous literature [19] have shown that dropout can help avoid overfitting and the number of known active compounds is limited, we take the model from the last training epoch instead of the one from early stopping determined by validation performance. Multiple splits are used to prove the robustness of the proposed strategy.

3.2. *Evaluation Metric*

1. Logarithmic Receiver-Operating-Characteristic Area Under the Curve with the False Positive Rate in the range [0.001, 0.1] (**$\mathbf{logAUC_{[0.001,0.1]}}$**)

Ranged logAUC [29] is used because only a small percentage of molecules predicted with high activity can be selected for experimental tests in consideration of cost in a real-world drug discovery campaign [24]. This high decision cutoff corresponds to the left side of the Receiver-Operating-Characteristic (ROC) curve, i.e., those False Positive Rates (FPRs) with small values. Also, because the threshold cannot be predetermined, the area under the curve is used to consolidate all possible thresholds within a certain small FPR range. Finally, the logarithm is used to bias towards smaller FPRs. Following prior work [19], we choose to use $\mathrm{logAUC}_{[0.001,0.1]}$. A perfect classifier achieves a $\mathrm{logAUC}_{[0.001,0.1]}$ of 1, while a random classifier reaches a $\mathrm{logAUC}_{[0.001,0.1]}$ of around 0.0215, as shown below:

$$\frac{\int_{0.001}^{0.1} x \,\mathrm{d}\log_{10} x}{\int_{0.001}^{0.1} 1 \,\mathrm{d}\log_{10} x} = \frac{\int_{-3}^{-1} 10^{u} \mathrm{d}u}{\int_{-3}^{-1} 1 \mathrm{d}\, u} \approx 0.0215$$

2. Boltzmann-enhanced discrimination of receiver operating characteristic (**BEDROC**)

BEDROC [25] is a metric that evaluates the early recognition ability of a given model. It prioritizes the identification of active compounds early in the ranked list. BEDROC ranges from 0 to 1, where a score closer to 1 indicates better performance in recognizing active compounds early in the list.

3. Enrichment factor with cutoff 100 (**EF_{100}**)

Enrichment factor [26] is often used metric in virtual screening. It measures how well a screening method can increase the proportion of active compounds in a selection set, compared to a random selection set. Here we select the top 100 compounds as the selection set. And the EF_{100} can be defined as follows.

$$EF_{100} = \frac{n_{100}/N_{100}}{n/N}$$

Where n_{100} is the number of true active compounds in the ranked top 100 predicted compounds given by the model, N_{100} is the number of compounds in the top 100 predicted compounds (i.e., 100), n is the number of active compounds in entire dataset, N is the number of compounds in the entire dataset. It is essentially a measure of the method's ability to "enrich" the set of compounds for further testing.

A random selection set receives a EF_{100} of 1. If no true active compounds are in the top 100 compounds, the EF_{100} becomes 0.

4. Discounted cumulative gain with cutoff 100 (**DCG_{100}**)

DCG [27] is a measure of ranking quality often used in web search. In a web search, it is obvious that a method is better when it positions highly relevant documents at the top of the search results. Virtual screening has a similar evaluation logic where we desire the active molecules to appear at the top of the selection set.

To calculate DCG, a simpler version metric named cumulative gain (CG) [27] is introduced below. CG is the sum of the relevance value of a compound in the selection set. In our case, a true active compound receives a relevance value of 1, while a true inactive compound receives a relevance value of 0. So, the CG with cutoff 100 (CG_{100}) equals the number of true active compounds in the top 100 compounds, i.e.,

$$CG_{100} = \sum_{i=1}^{100} y_i$$

It can be observed that CG_{100} is unaffected by changes in the ordering of compounds. DCG hence aims to penalize a true active molecule appearing lower in the selection set by logarithmically reducing the relevance value proportional to the predicted rank of the compound, i.e.,

$$DCG_{100} = \sum_{i=1}^{100} y_i / \log_2(i+1)$$

Table 1. Dataset statistics.

Protein Target Class	PubChem AID	Protein Target	Total Molecules	Active Molecules
GPCR	435008	Orexin1 Receptor	218,156	233
	1798	M1 Muscarinic Receptor Agonists	61,832	187
	435034	M1 Muscarinic Receptor Antagonists	61,755	362
Ion Channel	1843	Potassium Ion Channel Kir2.1	301,490	172
	2258	KCNQ2 Potassium Channel	302,402	213
	463087	Cav3 T-type Calcium Channels	100,874	703
Transporter	488997	Choline Transporter	302,303	252
Kinase	2689	Serine/Threonine Kinase 33	319,789	172
Enzyme	485290	Tyrosyl-DNA Phosphodiesterase	341,304	281

3.3. *Baseline Models*

We used three GNN models in our experiments: GCN [20], SchNet [11] and SphereNet [13]. We used the BCL [21] to generate traditional QSAR descriptors. Following previous examples [19, 30], we use the optimal descriptors where 391-element molecular-level features are generated. We provide a brief introduction to each of the models and the BCL below.

GCN extends the concept of convolution from regular, grid-like data (such as images) to graphs, which have arbitrary structures. GCNs work by aggregating information from a node's neighbors (potentially the node itself) to learn a representation of each node that captures both its features and local topology.

SchNet is a GNN designed for processing 3D molecules. The core design is continuous filters that are capable of handling unevenly spaced data, particularly, atoms. It also contains blocks that model interactions between atoms in a molecule.

SphereNet incorporates unique spherical message passing (SMP) for processing 3D molecules. The is encoded in a spherical coordinate system consisting of distance, angle and torsion. The SMP then uses the spherical coordinate system for the message passing process.

BCL is an application-based, open-source software package that integrates traditional small molecule cheminformatics tools with machine learning-based quantitative structure-activity/property relationship (QSAR/QSPR) modeling. It is designed to facilitate various cheminformatics tasks such as computing chemical properties, estimating druglikeness etc. It serves as a valuable resource for researchers in the computer-aided drug discovery field by providing a modular toolkit that supports the integration of cheminformatics and machine learning tools into their research workflows.

4. Future Work

In future work, we plan to expand our investigation by incorporating a broader array of GNN architectures and descriptor sets. This expansion will allow us to evaluate the generalizability and scalability of our integrative approach across a wider spectrum of computational models and chemical descriptor libraries.

We aim to explore advanced GNN models that may offer distinct advantages in capturing molecular features and interactions, potentially leading to improved predictive performance in virtual screening tasks. By comparing a diverse range of GNN architectures, we can better understand the nuances of how different models interact with various descriptor sets, and identify optimal combinations that maximize screening efficacy and accuracy.

Additionally, we intend to experiment with an expanded set of expert-crafted descriptors, including those that capture more intricate chemical and physical properties of molecules. This will enhance our ability to assess the impact of different types of descriptors on the performance of GNNs in virtual screening. By systematically evaluating the contribution of each descriptor type, we can refine our integration strategies to leverage the strengths of both GNNs and traditional chemical descriptors effectively.

Ultimately, our goal is to develop a comprehensive framework that can adapt to the evolving landscape of drug discovery, accommodating new advances in machine learning and cheminformatics.

5. Conclusion

Our study has rigorously evaluated the impact of integrating expert-crafted descriptors with GNNs and demonstrated that this integrative approach can significantly enhance the predictive power of virtual screening processes. Notably, the use of descriptors in conjunction with GNN architectures like GCN and SchNet has led to substantial improvements in identifying bioactive compounds.

In addition, The convergence in performance metrics across different GNN models, when supplemented with descriptors, suggests the potential for simpler GNN architectures to achieve results comparable to their more complex counterparts within this integrative framework. This finding underscores the viability of leveraging traditional knowledge and computational simplicity to advance the state-of-the-art in virtual screening.

Furthermore, our experiments with scaffold split scenarios revealed the robustness of descriptors, often outperforming combined GNN-descriptor models. This highlights the enduring value of expert knowledge in the face of evolving computational techniques and stresses the necessity for future models to effectively integrate this knowledge to enhance predictive power in realistic drug discovery settings.

In conclusion, our study serves as a compelling demonstration of how the synergistic integration of GNNs and expert-crafted descriptors can significantly advance the field of ligand-based virtual screening. As we move forward, it is imperative that we continue to explore and refine these integrative strategies, with the aim of developing more sophisticated and effective tools for drug discovery. The journey towards optimizing virtual screening methodologies is far from complete, but our work provides a significant step forward, offering a blueprint for future research in this dynamic and evolving field.

Acknowledgments

Yunchao (Lance) Liu acknowledges that the Nvidia Academic Hardware Grant provides an A6000 GPU for speeding up the computation. Yunchao (Lance) Liu thanks Holy Gagnon for inspiring the discussion of the results.

Funding Information

Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 HL122010, R01 DA046138. J.M. is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J.M. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB1423, project number 421152132 and through SPP 2363 for financial support.

Conflict of Interest

Authors have no conflict of interest to declare.

Appendix A. Descriptor Features

The descriptor sets used in this study are from [19]. There are 391 elements of features in total. Each signed 2D autocorrelation (2DA [31]) contains 32 bins. Each signed 3D autocorrelation (3DA [31]) contains 60 bins. See the original paper [19] for individual feature naming details.

Table 2. Features used in the descriptor set. Originally used in [19].

Scalar Features	Signed 2D Autocorrelation	Signed 3D Autocorrelation
Weight	Atom_SigmaCharge	Atom_SigmaCharge
HbondDonor	Atom_Vcharge	Atom_Vcharge
HbondAcceptor	IsHTernary	IsHTernary
LogP	Atom_IsInAromaticRingTernary	Atom_IsInAromaticRingTernary
TotalCharge		
NRotBond		
NaromaticRings		
Nrings		
TopologicalPolarSurfaceArea		
Girth		
BondGrith		
MaxRingSize		
Limit(MinRingSize, max=8, min=0)		
MoleculeSum(Atom_InAromaticRingIntersection),		
MoleculeSum(Atom_InRingIntersection)		
MoleculeStandardDeviation(Atom_Vcharge)		
MoleculeStandardDeviation(Atom_SigmaCharge)		
MoleculeMax(Atom_Vcharge)		
MoleculeMin(Atom_Vcharge)		

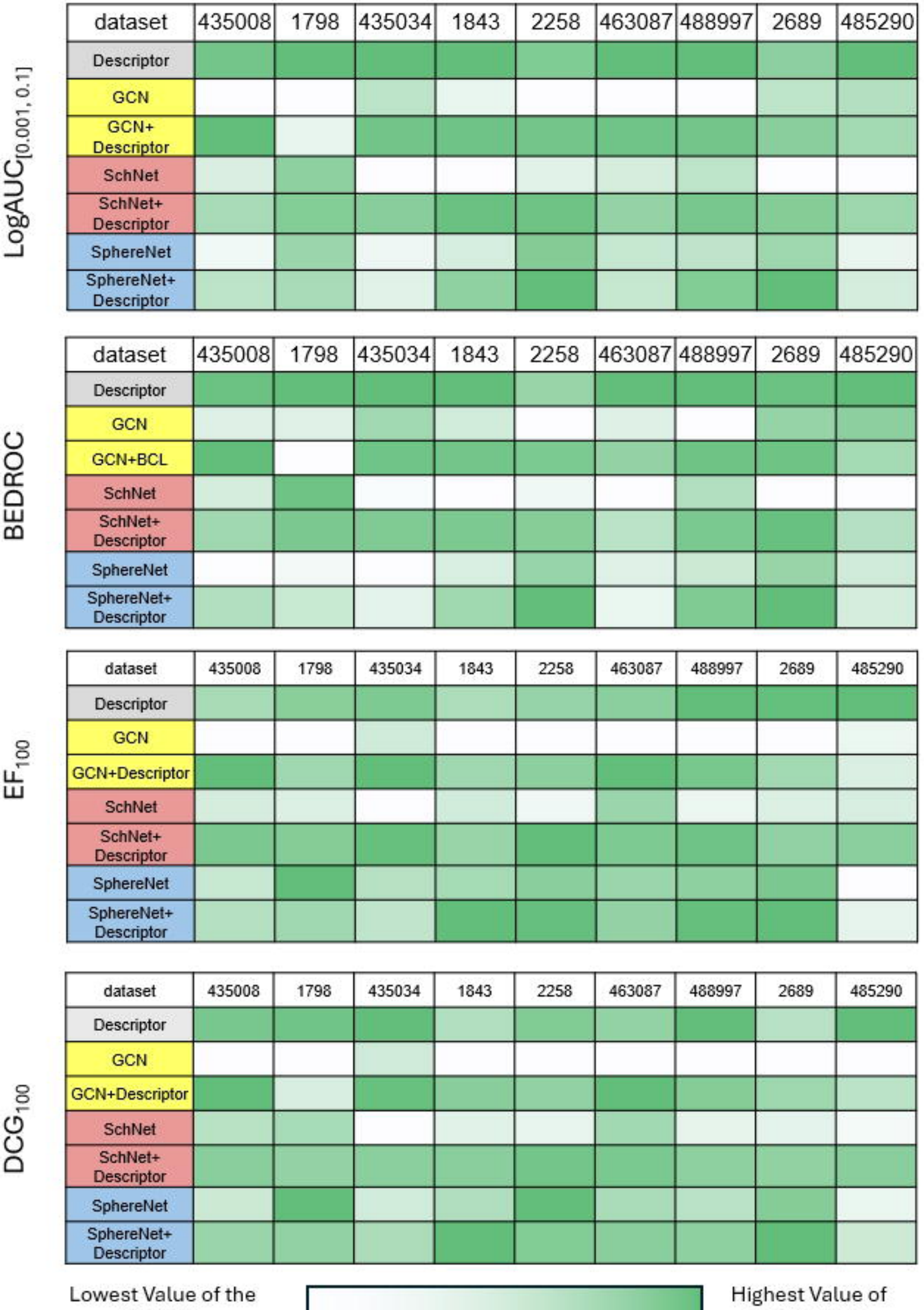
MoleculeMax(Atom_SigmaCharge)
MoleculeMin(Atom_SigmaCharge)
MoleculeSum(Abs(Atom_Vcharge))
MoleculeSum(Abs(Atom_SigmaCharge)

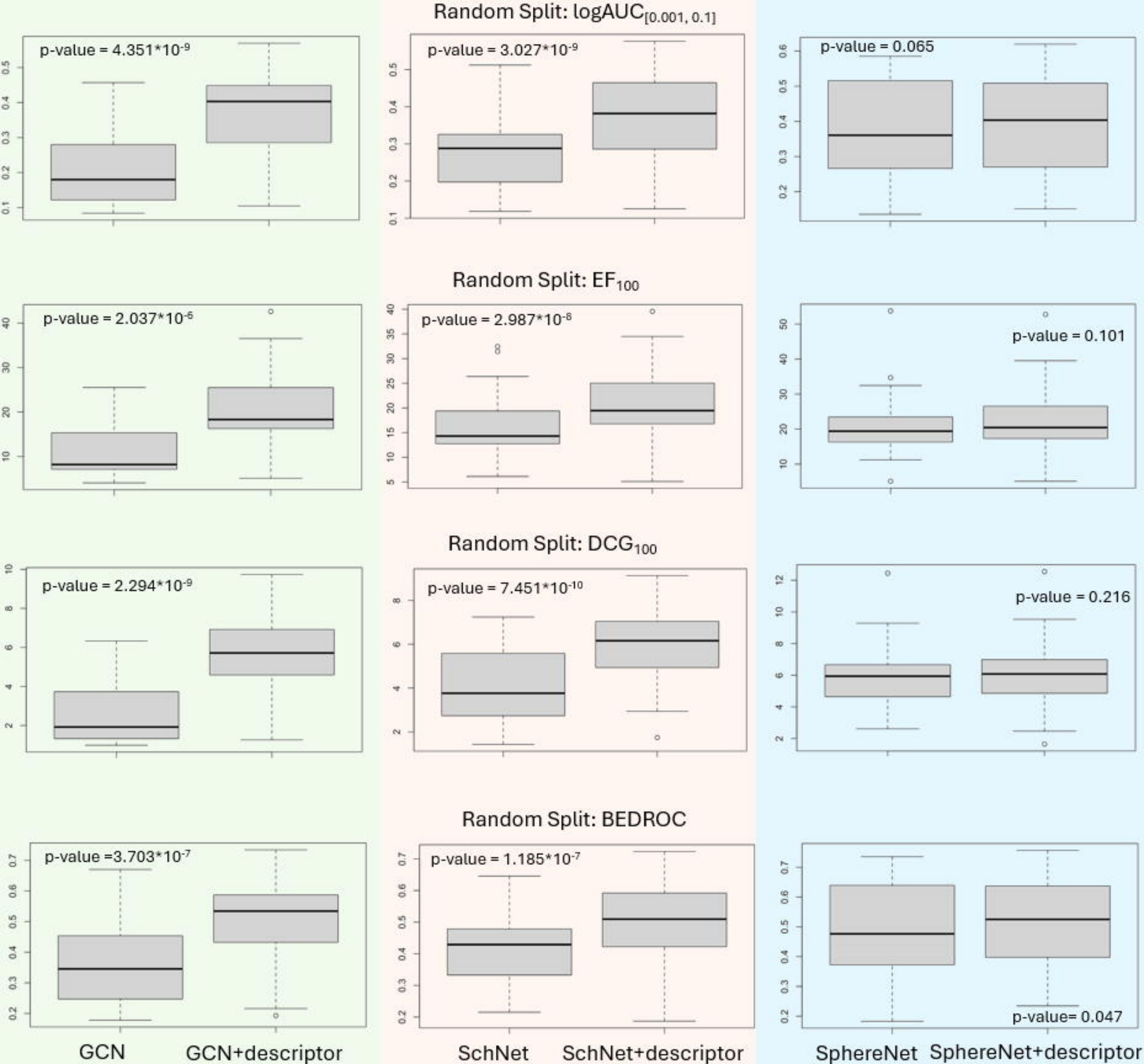
References

1. Sliwoski, G., S. Kothiwale, J. Meiler, and E.W. Lowe, Jr., *Computational methods in drug discovery.* Pharmacol Rev, 2014. **66**(1): p. 334-95.
2. Jumper, J., R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021.
3. Baek, M., I. Anishchenko, I. Humphreys, Q. Cong, D. Baker, and F. DiMaio, *Efficient and accurate prediction of protein structure using RoseTTAFold2.* bioRxiv, 2023: p. 2023.05. 24.542179.
4. Baek, M., F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, J. Wang, Q. Cong, L.N. Kinch, and R.D. Schaeffer, *Accurate prediction of protein structures and interactions using a three-track neural network.* Science, 2021. **373**(6557): p. 871-876.
5. Lin, Z., H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, and Y. Shmueli, *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.
6. Ripphausen, P., B. Nisius, and J. Bajorath, *State-of-the-art in ligand-based virtual screening.* Drug discovery today, 2011. **16**(9-10): p. 372-376.
7. Klicpera, J., J. Groß, and S. Günnemann, *Directional message passing for molecular graphs.* arXiv preprint arXiv:2003.03123, 2020.
8. Zhang, Y., *An In-depth Summary of Recent Artificial Intelligence Applications in Drug Design.* arXiv preprint arXiv:2110.05478, 2021.
9. Wang, L., Y. Liu, Y. Lin, H. Liu, and S. Ji, *ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs.* arXiv preprint arXiv:2206.08515, 2022.
10. Liu, Y., Y. Wang, O.T. Vu, R. Moretti, B. Bodenheimer, J. Meiler, and T. Derr, *Interpretable Chirality-Aware Graph Neural Network for Quantitative Structure Activity Relationship Modeling in Drug Discovery.* bioRxiv, 2022: p. 2022.08. 24.505155.
11. Schütt, K.T., H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, *Schnet– a deep learning architecture for molecules and materials.* The Journal of Chemical Physics, 2018. **148**(24): p. 241722.

12. Klicpera, J., S. Giri, J.T. Margraf, and S. Günnemann, *Fast and uncertainty-aware directional message passing for non-equilibrium molecules.* arXiv preprint arXiv:2011.14115, 2020.
13. Liu, Y., L. Wang, M. Liu, X. Zhang, B. Oztekin, and S. Ji, *Spherical message passing for 3d graph networks.* arXiv preprint arXiv:2102.05013, 2021.
14. Chen, D., Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun. *Measuring and relieving the over-smoothing problem for graph neural networks from the topological view*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.
15. Alon, U. and E. Yahav, *On the bottleneck of graph neural networks and its practical implications.* arXiv preprint arXiv:2006.05205, 2020.
16. Zhong, Z., A. Barkova, and D. Mottin, *Knowledge-augmented graph machine learning for drug discovery: A survey from precision to interpretability.* arXiv preprint arXiv:2302.08261, 2023.
17. Wu, Z., D. Jiang, C.-Y. Hsieh, G. Chen, B. Liao, D. Cao, and T. Hou, *Hyperbolic relational graph convolution networks plus: a simple but highly efficient QSAR-modeling method.* Briefings in Bioinformatics, 2021. **22**(5).
18. Yang, K., K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, and R. Barzilay, *Analyzing Learned Molecular Representations for Property Prediction.* Journal of Chemical Information and Modeling, 2019. **59**(8): p. 3370-3388.
19. Mendenhall, J. and J. Meiler, *Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout.* Journal of computer-aided molecular design, 2016. **30**(2): p. 177-189.
20. Kipf, T.N. and M. Welling *Semi-Supervised Classification with Graph Convolutional Networks*. arXiv e-prints, 2016. arXiv:1609.02907.
21. Benjamin P. Brown, O.V., Alexander R. Geanes, Sandeepkumar Kothiwale, Mariusz 4 Butkiewicz, Edward W. Lowe, Jr., Ralf Mueller, Richard Pape, Jeffrey Mendenhall, and Jens Meiler, *Introduction to the BioChemical Library (BCL): An application-based open-source toolkit for integrated cheminformatics and machine learning in computer-aided drug discovery.* 2022.
22. Hsu, H. and P.A. Lachenbruch, *Paired t test.* Wiley StatsRef: statistics reference online, 2014.
23. Böhm, H.-J., A. Flohr, and M. Stahl, *Scaffold hopping.* Drug discovery today: Technologies, 2004. **1**(3): p. 217-224.
24. Butkiewicz, M., Y. Wang, S.H. Bryant, E.W. Lowe, Jr., D.C. Weaver, and J. Meiler, *High-Throughput Screening Assay Datasets from the PubChem Database.* Chemical informatics (Wilmington, Del.), 2017. **3**(1): p. 1.
25. Butkiewicz, M., E.W. Lowe, Jr., R. Mueller, J.L. Mendenhall, P.L. Teixeira, C.D. Weaver, and J. Meiler, *Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database.* Molecules, 2013. **18**(1): p. 735-56.
26. Kim, S., J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E.E. Bolton, *PubChem 2019 update: improved access to chemical data.* Nucleic acids research, 2019. **47**(D1): p. D1102-D1109.
27. O'Boyle, N.M., M. Banck, C.A. James, C. Morley, T. Vandermeersch, and G.R. Hutchison, *Open Babel: An open chemical toolbox.* Journal of cheminformatics, 2011. **3**(1): p. 1-14.

28. Gasteiger, J., C. Rudolph, and J. Sadowski, *Automatic generation of 3D-atomic coordinates for organic molecules.* Tetrahedron Computer Methodology, 1990. **3**(6): p. 537-547.
29. Mysinger, M.M. and B.K. Shoichet, *Rapid Context-Dependent Ligand Desolvation in Molecular Docking.* Journal of Chemical Information and Modeling, 2010. **50**(9): p. 1561-1573.
30. Vu, O., J. Mendenhall, D. Altarawy, and J. Meiler, *BCL::Mol2D-a robust atom environment descriptor for QSAR modeling and lead optimization.* J Comput Aided Mol Des, 2019. **33**(5): p. 477-486.
31. Sliwoski, G., J. Mendenhall, and J. Meiler, *Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign.* Journal of computer-aided molecular design, 2016. **30**(3): p. 209-217.


LogAUC[0.001, 0.1]
dataset
435008
1798
435034
1843
2258
463087
488997
2689
485290
Descriptor
GCN
GCN+ Descriptor
SchNet
SchNet+ Descriptor
SphereNet
SphereNet+ Descriptor
BEDROC
dataset
435008
1798
435034
1843
2258
463087
488997
2689
485290
Descriptor
GCN
GCN+BCL
SchNet
SchNet+ Descriptor
SphereNet
SphereNet+ Descriptor
EF100
dataset
435008
1798
435034
1843
2258
463087
488997
2689
485290
Descriptor
GCN
GCN+Descriptor
SchNet
SchNet+ Descriptor
SphereNet
SphereNet+ Descriptor
DCG100
dataset
435008
1798
435034
1843
2258
463087
488997
2689
485290
Descriptor
GCN
GCN+Descriptor
SchNet
SchNet+ Descriptor
SphereNet
SphereNet+ Descriptor
Lowest Value of the Column
Highest Value of the Column


Random Split: logAUC[0.001, 0.1]
p-value = 4.351*10-9
p-value = 3.027*10-9
p-value = 0.065
Random Split: EF100
p-value = 2.037*10-6
p-value = 2.987*10-8
p-value = 0.101
Random Split: DCG100
p-value = 2.294*10-9
p-value = 7.451*10-10
p-value = 0.216
Random Split: BEDROC
p-value =3.703*10-7
p-value = 1.185*10-7
p-value= 0.047
GCN
GCN+descriptor
SchNet
SchNet+descriptor
SphereNet
SphereNet+descriptor