ENHANCING GRAPH NEURAL NETWORKS FOR MULTI-TARGET ACTIVITY PREDICTION IN LEAD OPTIMIZATION THROUGH MULTI-TASK LEARNING AND KNOWLEDGE DISTILLATION

ARTHUR CERVEIRA[1]; FREDERICO KREMER[2]; GABRIEL GOMES[3]; ULISSES CORRÊA[4]

[1]Universidade Federal de Pelotas – aacerveira@inf.ufpel.edu.br
[2]Universidade Federal de Pelotas – fred.s.kremer@gmail.com
[3]Universidade Federal de Pelotas – gagomes@inf.ufpel.edu.br
[4]Universidade Federal de Pelotas – ulisses@inf.ufpel.edu.br

1. INTRODUCTION

Artificial intelligence (AI) has been widely adopted in the discovery of novel drugs for treating a wide range of diseases (MAK, 2023). In particular, AI has proven crucial in the early stages of the drug discovery pipeline (DDP), where rapid and cost-effective identification of promising candidate molecules is essential (MAK, 2023). In this context, virtual screening (VS) enables the rapid evaluation of large virtual molecular libraries (GUIDOTTI, 2023). Compounds can be computationally screened through VS to predict their potential activity against biological targets. The top-ranked candidate molecules are then prioritized for further experimental validation, moving forward to subsequent phases of the DDP.

Drug targets, typically proteins associated with specific disease mechanisms, are fundamental to discovering candidate compounds (MAK, 2023). Pharmaceutical companies and researchers increasingly leverage machine learning (ML) models to predict the activity of molecules against these targets. In particular, Quantitative structure-activity relationship (QSAR) models are widely employed to relate molecular structures to their biological activities, facilitating the identification of high-potency compounds (MAK, 2023).

Notably, there are often strong similarities when training QSAR models for related targets, particularly within the same protein family or biological pathway. Some studies have exploited these similarities through multi-task learning (MTL) and knowledge distillation (KD), a paradigm where a single model is trained to predict activities across multiple targets simultaneously (MOON, 2022) (LU, 2023), achieving promising results by sharing information across related tasks.

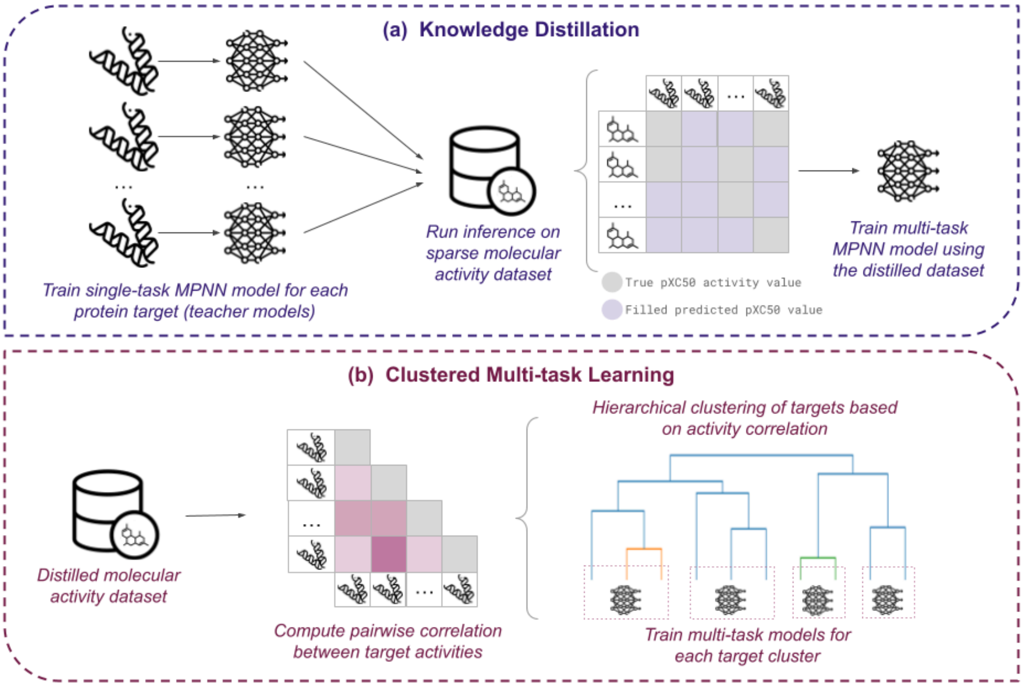
Modern neural network architectures, such as graph neural networks (GNNs), have achieved state-of-the-art performance across many QSAR modeling tasks (HEID, 2024). However, most works exploring MTL for molecular activity prediction have focused on classical ML algorithms (ROSENBAUM, 2013) or feed-forward neural networks (FFNs) (MOON, 2022) using precomputed molecular fingerprints as input, rather than leveraging graph-based representations.

Therefore, this work provides an assessment of GNN models in multi-task learning scenarios for molecular activity prediction. Our evaluation focuses on measuring the variation in performance between single-task and multi-task approaches while also benchmarking against classical ML models and tabular network baselines. We conduct a comprehensive analysis of molecular activity prediction using the practical lead optimization (Lo) benchmark splits (STESHIN, 2023). Our results highlight the strengths and limitations of graph-based architectures in multi-task settings and provide insights into their practical utility for improving the lead optimization phase of drug discovery workflows.

2. METHODS

Our proposed MTL approach comprises two main components: (1) a KD system to transfer knowledge from pre-trained target-specific models to construct dense multi-target activity prediction datasets, and (2) a task clustering approach based on the hierarchical clustering algorithm to identify groups of related tasks and exploit the similarities between them. Combining these two components allows us to exploit the similarities between targets to improve the learned latent space representations of molecular structures. Figure 1 depicts the KD system and clustering pipeline.

Figure 1. Proposed MTL approach depicting (a) the KD system and (b) the task clustering pipeline.


(a) Knowledge Distillation
Train single-task MPNN model for each protein target (teacher models)
Run inference on sparse molecular activity dataset
True pXC50 activity value
Filled predicted pXC50 value
Train multi-task MPNN model using the distilled dataset
(b) Clustered Multi-task Learning
Distilled molecular activity dataset
Compute pairwise correlation between target activities
Hierarchical clustering of targets based on activity correlation
Train multi-task models for each target cluster

The proposed knowledge distillation system builds a dense multi-target activity prediction dataset by distilling knowledge from pre-trained target-specific models. We can achieve this by training a single-task model for each considered target and using the model predictions as soft labels to train new supervised models. Only the missing activity annotations are replaced by the new predicted labels, while the existing activity values are kept in the dataset. This dense dataset is then used to train multi-task models. This process is applied to the training dataset to avoid data leakage when evaluating the model's performance on the test and validation datasets.

The task clustering approach is based on the hierarchical clustering algorithm using the correlation coefficient between the activity values of the targets as the distance metric. Using the distilled molecular activity dataset, we can construct a distance matrix between the targets by computing the Pearson correlation coefficient between activity values for each pair of targets. The distance metric is defined as 1 - Δ, where Δ represents the considered correlation coefficient. To identify groups of related targets, we apply the agglomerative clustering algorithm — a hierarchical clustering approach that uses a bottom-up approach to group the targets into a pre-defined number of clusters.

3. RESULTS AND DISCUSSION

We assess the GNN, FFN, and Bambu (classic ML baseline) ability to predict molecular activity for each predefined target in a filtered evaluation dataset. We consider the Lo benchmark split for evaluation. The metrics considered are the mean squared error (MSE), the mean absolute error (MAE), and the coefficient of determination (R^2). These metrics are computed for each model across all targets, with results reported as the mean across all targets. Lower MSE and MAE indicate better performance, while higher R^2 values are preferred.

Table 1 reports the results for all models trained and evaluated on the Lo split. Methods are grouped by model architecture and learning strategy. We consider five configurations for both GNN and FFN architectures: standard ST, MTL, clustered MTL, MTL with KD, and clustered MTL with KD (the proposed method). These are indicated with the Multi-Task, Clustered, and Distilled columns in the results table. Bambu, which only supports traditional ML algorithms, is excluded from MTL-based evaluations.

Table 1. Assessment of Baseline Models on Lo Dataset

	Multi-Task	Clustered	Distilled	MSE	MAE	R2
GNN	✓	✓	✓	**1.046 ± 0.6**	**0.742 ± 0.3**	**0.052 ± 0.2**
	✓		✓	1.113 ± 0.7	0.769 ± 0.3	0.017 ± 0.2
	✓	✓		1.099 ± 0.7	0.759 ± 0.3	0.017 ± 0.2
	✓			1.164 ± 0.7	0.785 ± 0.3	-0.063 ± 0.3
				1.077 ± 0.7	0.755 ± 0.3	0.031 ± 0.2
FFN	✓	✓	✓	2.841 ± 4.2	1.196 ± 0.9	-3.101 ± 8.7
	✓		✓	13.32 ± 7.0	3.318 ± 1.0	-15.22 ± 17.0
	✓	✓		2.626 ± 1.9	1.259 ± 0.5	-2.02 ± 3.4
	✓			14.48 ± 13.5	3.147 ± 1.9	-15.00 ± 40.9
				2.626 ± 1.8	1.256 ± 0.5	-1.938 ± 2.6
Bambu				1.155 ± 0.7	0.768 ± 0.3	-0.019 ± 0.2

The proposed MTL configuration applied to the GNN architecture achieved the best performance across all metrics: 1.046 for MSE, 0.742 for MAE, and 0.052 for R^2. The single-task GNN configuration was the next best performer, suggesting that the performance gains of MTL arise primarily from the combination of task clustering and distillation. FFN-based models performed worst across all metrics. This can be partially attributed to the fact that the original FFN implementation was designed for binary activity classification (active/inactive), whereas our task involves continuous activity prediction — a more complex and information-rich regression problem that poses greater challenges for simpler neural architectures. These results highlight the advantage of graph-based models like GNN for multi-target molecular activity prediction, particularly in regression settings. Lo benchmarks simulate significantly more difficult scenarios than the random split commonly used in ML evaluations. For instance, R^2 values are frequently negative, indicating performance worse than the null hypothesis (i.e., a horizontal line at the mean activity level). This performance decline is partially attributable to reduced dataset sizes resulting from the considered data split.

4. CONCLUSIONS

This work explored the application of MTL to enhance the performance of GNNs in multi-target molecular activity prediction tasks. Our approach combines task clustering with KD to better capture and exploit shared patterns across related biological targets. To ensure a realistic evaluation, we adopted the practical Lo data split that mimics real-world drug discovery scenario. We benchmarked our method against several baselines, including single-task and multi-task variants of GNNs, FFNs trained on molecular fingerprints, and traditional machine learning models implemented with Bambu. The proposed MTL configuration consistently improved the performance of graph-based models, achieving superior results in RS and Lo benchmarks. Our findings suggest that MTL is a promising strategy even in data-constrained settings. As a future direction, we plan to explore transfer learning techniques to further improve performance in such scenarios and evaluate whether GNNs can outperform simpler models in these scenarios. We provide all source code and datasets used in this study to support reproducibility and further research. This work serves as a foundation for the development of more effective approaches to multi-target molecular activity prediction.

5. REFERENCES

GUIDOTTI IL, NEIS A, MARTINEZ DP, SEIXAS FK, MACHADO K, KREMER FS. Bambu and its applications in the discovery of active molecules against melanoma. **J Mol Graph Model**. 2023.

HEID, E.; GREENMAN, K. P.; CHUNG, Y.; LI, S.-C.; GRAFF, D. E.; VERMEIRE, F. H.; WU, H.; GREEN, W. H.; McGILL, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. **Journal of Chemical Information and Modeling**, Washington, v.64, n.1, p.9-17, 2024.

LU, R.; WANG, J.; LI, P.; LI, Y.; TAN, S.; PAN, Y.; LIU, H.; GAO, P.; XIE, G.; YAO, X. Improving drug-target affinity prediction via feature fusion and knowledge distillation. **Briefings in Bioinformatics**, Oxford, v.24, n.3, p.bbad145, 2023.

MAK, K., WONG, Y., PICHIKA, M.: Artificial intelligence in drug discovery and development. **Drug Discovery and Evaluation: Safety and Pharmacokinetic Assays.** Hock, F., Pugsley, M. Springer, Cham, 2023.

MOON, C.; KIM, D. Prediction of drug–target interactions through multi-task learning. **Scientific Reports**, London, v.12, n.18323, p.1-12, 2022.

ROSENBAUM, L.; DÖRR, A.; BAUER, M. R.; et al. Inferring multi-target QSAR models with taxonomy-based multi-task learning. **Journal of Cheminformatics**, London, v.5, n.33, p.1-12, 2013.

STESHIN, S. Lo-Hi: Practical ML Drug Discovery Benchmark. In: OH, A.; NAUMANN, T.; GLOBERSON, A.; SAENKO, K.; HARDT, M.; LEVINE, S. (Ed.). Advances in **Neural Information Processing Systems. v.36**. Curran Associates, Inc., 2023. p.64526–64554.