STRUCTURE-BASED METABOLITE FUNCTION PREDICTION USING GRAPH NEURAL NETWORKS

Tancredi Cogne, Mariam Ait Oumelloul, Ali Saadat
School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{tancredi.cogne,mariam.aitoumelloul,ali.saadat}@epfl.ch

Janna Hastings
Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich
School of Medicine, University of St. Gallen
Swiss Institute of Bioinformatics
janna.hastings@uzh.ch

Jacques Fellay
School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
Biomedical Data Science Center, Lausanne University Hospital and University of Lausanne
jacques.fellay@epfl.ch

ABSTRACT

Being able to broadly predict the function of novel metabolites based on their structures has applications in systems biology, environmental monitoring and drug discovery. To date, machine learning models aiming to predict functional characteristics of metabolites have largely been limited in scope to predicting single functions, or only a small number of functions simultaneously. Using the Human Metabolome Database as a source for a wider range of functional annotations, we assess the feasibility of predicting metabolite functions more broadly, as defined by four elements, namely location, role, the process it is involved in, and its physiological effect. We evaluated three graph neural network architectures to predict available functional ontology terms. Among the models tested, the Graph Attention Network, incorporating embeddings from the pre-trained ChemBERTa model to predict the process metabolites are involved in, achieved the highest performance with an F1-score of 0.889 and a recall of 0.903. The model identified function-associated structural patterns within metabolite families, demonstrating the potential for interpretably predicting metabolite functions from structural information.

1 INTRODUCTION

Metabolites are the small molecules produced during metabolism that play essential roles in biochemical pathways in all living organisms. Understanding their functions is important to advance the fundamental understanding of molecular and cellular pathways and for a range of different application areas including environmental monitoring and drug discovery. Functions are commonly linked to structures in biochemistry, including for metabolites (1) (2), enabling the potential prediction of functional characteristics based on structures. This idea has been extensively explored for proteins, where the prediction of function based on structure is a longstanding grand challenge in computational biology (3)(4)(5)(6), and has seen great improvements in recent years with the rise of machine learning models such as DeepFRI (7), which predicts functional Gene Ontology (GO) terms based on the structure of proteins.

To our knowledge, unlike for proteins, for metabolites no machine learning model exists to predict a broad range of functions at the same time based on their structure. A possible explanation could be the lack of a gold-standard ontology for metabolites that matches the scale of the Gene Ontology (GO) for proteins. Various ontologies exist, such as the 'role' branch of the ChEBI (8) ontology or

ChemFOnt (9), but these are not as complete as GO or as comprehensive in terms of annotations. Nevertheless, chemical compounds have more stable and defined structures than proteins, which should represent a valuable source of information. An increase in the number and scale of publicly accessible metabolite databases, such as the Human Metabolome Database (HMDB)(10), gives rise to an opportunity to address this gap.

We aimed to develop a model to address the challenge of predicting metabolite function based on their structure. We hypothesize that the molecular structures of metabolites contain sufficient information to enable prediction of their functional metabolic characteristics as defined by their location, their role, the processes they are involved in, and their physiological effects (as per HMDB annotations). Based on the logical representation of molecules as graphs, we compare three different graph neural network architectures for this task, and furthermore show the ability of an attention-based model to detect the importance of certain chemical bonds for the function of molecules.

1.1 Related work

Advancements in protein function prediction have been driven by machine learning models that effectively leverage sequence and structural information. For example, DeepGO uses deep learning and protein sequence embeddings to predict protein function annotations, with its updated version, DeepGO-SE, incorporating a pretrained large language model (LLM) to predict Gene Ontology (GO) functions (11). Similarly, DeepFRI combines graph convolutional networks (GCNs) and protein contact maps to identify functional sites in protein structures. GCNFold (12) applies GCNs to protein structure graphs for functional site prediction, while TAWFN integrates convolutional neural networks (CNNs) and GCNs for enhanced protein function annotation (13).

While these advances highlight the power of machine learning in protein function prediction, no model has yet been developed to perform multi-label functional predictions specifically for metabolites. In the broader domain of chemical predictive modelling, QSAR (14) models link molecular structure to specific biological activity, for example models may predict binding affinity, inhibitory concentration, or toxicity using linear and non-linear methods. Other models have been developed to address related prediction tasks, such as DLMPM (15), which employs a latent factor model to identify disease-metabolite associations, and Deep-DRM, which applies graph deep learning techniques for the same purpose. GCNAT (16) combines GCNs with graph attention networks for predicting disease-metabolite associations. Expanding to molecular property prediction, Qu et al. (2024) (17) used GNNs to predict properties such as boiling points and mass spectra. Huckvale et al. (2024) (18) proposed a model to determine metabolite-pathway involvement using features of both metabolites and pathways. Additionally, Porokhin et al. (2023) (19) demonstrated GNN applications for site-of-metabolism prediction, while Glauer et al. (2024) (20) introduced a Transformer-based model to extend ChEBI's ontology (the structural branch rather than the role branch) by classifying unseen chemical structures.

Finally, ChemBERTa (21) can learn molecular representations using SMILES and was used in our model to improve the results. MolCLR (22) also aims to learn molecular representations in a self-supervised manner.

1.2 Contributions of our study

We propose a novel machine learning approach for predicting multiple metabolite functions based solely on chemical structure, addressing a significant unmet need in current metabolomics research, where increasingly large numbers of metabolite structures may be characterised in samples yet not yet be associated with any functional annotations.

We extract and filter a dataset of 3'278 metabolite structures and associated functions from the HMDB. We evaluated and compared three graph neural network (GNN) architectures: Graph Convolutional Network (GCN), Graph Isomorphism Network (GIN), and Graph Attention Network (GAT), while also assessing the effectiveness of ChemBERTa embeddings to augment these models. We highlight the attention-based model's ability to detect important molecular substructures by leveraging explainable AI techniques on the attention weights.

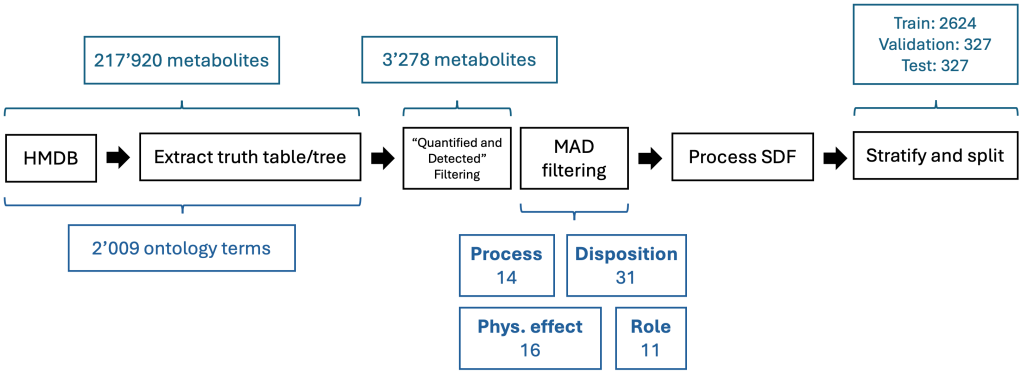

Figure 1: Data processing pipeline. Files from the HMDB were parsed to extract a truth table and a tree ontology. Metabolites were filtered by keeping only the ones with the label "Detected and quantified". Ontology terms were first filtered to keep only the terminal (without a child) nodes. A subset was selected using median absolute deviation filtering. Input/output information was extracted for this given set of metabolites and ontology terms. The data was stratified and split to tackle the unbalanced nature of the data set regarding possible outputs.

2 METHODS

2.1 DATABASE

The Human Metabolome DataBase (HMDB) includes 217'920 metabolites, each characterized by various attributes such as molecular weight and chemical structure.

2.2 METABOLITE FILTERING

Each metabolite in HMDB falls into one of four categories: "expected", "predicted", "detected but not quantified", or "detected and quantified". For each ontology term, we counted the number of metabolites associated with it, allowing us to calculate the standard deviation across different metabolite categories (Fig. 2) using:

$$\sigma_j = \sqrt{\frac{\sum_i (x_{i,j} - \bar{x}_j)}{N-1}}$$

Here, $x_{i,j}$ represents the binary value for metabolite i and ontology term j. $\bar{x}_j$ is the mean value of the binary values for ontology term j and N is the number of metabolites. This analysis was based on the assumption that if all metabolites are associated with a given term, or none are, the standard deviation would be zero, indicating a lack of informative value for the model. An important difference can be seen in the different distributions. Most ontology terms have a near zero standard deviation when looking only at "expected", "predicted", or "detected but not quantified" metabolites. This can be explained by the fact that most of them do not contain any ontology-related information and thus are set to 'False' yielding a very small standard deviation. Thus, we decided to focus on the "detected and quantified" (3'278 metabolites) category (Fig. 1) because it was the only one with enough ontology-related information to test our hypothesis.

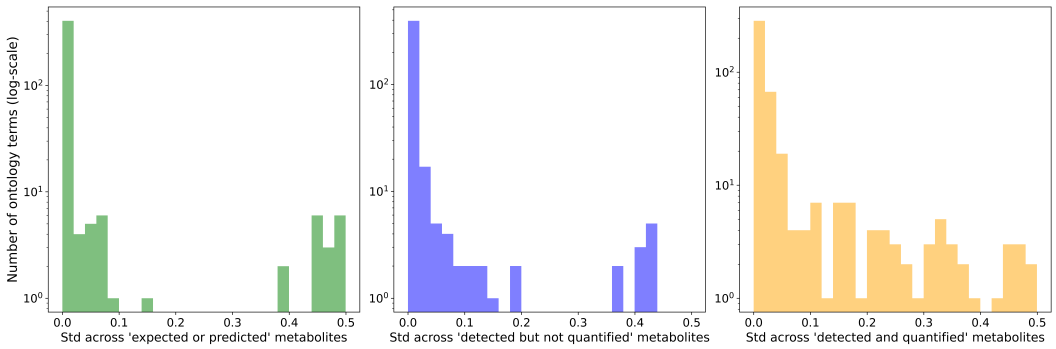

Figure 2: Histogram of standard deviation across the different categories of metabolites as described in the HMDB. Many ontology terms have a near zero standard deviation and were filtered out to retain only a small subset of ontology terms as outputs of the model. Distribution for all the 'expected' or 'predicted' metabolites (green), 'detected but not quantified' metabolites (blue), and 'detected and quantified' metabolites (yellow).

2.3 Function definition

The HMDB's functional hierarchical structure comprises 2,009 distinct ontology terms, all of which are categorized under one of four primary nodes:

- "Disposition" is defined by the HMDB as the "origin of a chemical, its location within an organism or its route of exposure". We will refer to this category as "location".
- "Role" is defined by the HMDB as the "purpose or function of a chemical, either naturally or as intended by humans".
- "Process" is defined by the HMDB as the "biological or chemical events, or a series thereof, leading to a known function or a known end product".
- "Physiological effect" is defined by the HMDB as the "measured or observed physiological effect on an organism resulting from its exposure to a chemical".

We used median absolute deviation (MAD) to filter which ontology terms were used as an output of the model. MAD is a robust statistical measure of variability and is sensitive to outliers. The threshold was selected with a modified Z-score M_i based on the similarity with a gamma distribution (23):

$$M_i = \frac{0.6745(s_i - s)}{\text{MAD}}$$

where s_i is the standard deviation of the ontology term i, s is the median standard deviation, and MAD is the median absolute deviation. Terms with an absolute $M_i > 3.5$ were selected as outputs of the model. After filtering, 14 child nodes remained in "Process", 31 in "Disposition", 16 in "Physiological effect", and 11 in "Role" (Fig. 1).

2.4 Data processing

Graph representation is commonly used for molecules (24) (25) (26) (27) (28) as a direct mapping can be made by using a node for each atom and an edge between two atoms for each bond. The input representation used in our model stores multiple pieces of information for each metabolite: 2D coordinates as given by HMDB and atomic number of each atom, the two atoms at the extremities of each bond, and the bond type (single, double, ...). One-hot encoding is commonly used for categories in machine learning to avoid arbitrary ordinal relationships and was used in this model to represent each possible atom. The 2D coordinates of each molecule were normalized and standardized given that one-hot encodings were also used as input of the model. Indeed, having both extreme coordinates and one-hot encodings as inputs would result in the model likely focusing only on the coordinates. The model predicts a binary value separately for each selected ontology term (multi-label classification). Even though some terms are related, the model treats all outputs as independent for simplification.

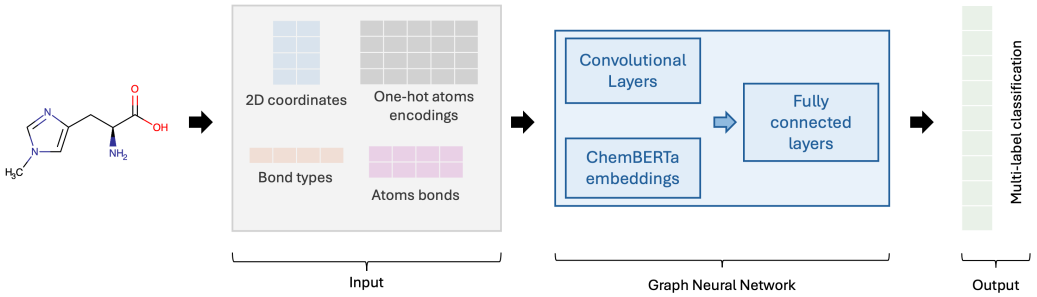

Figure 3: Visual representation of the model from input to the output. The input is composed of four arrays to describe all the parts of each molecule: 2D coordinates, atomic numbers, and bonds. The compared architectures all include convolutional layers followed by fully connected layers. Multi-label classification is used and a binary value is outputted for each selected ontology term.

Due to the small size and the unbalanced nature of the dataset, the split between training and test sets had to be carefully done. Indeed, many metabolites share similar outputs. To avoid their uneven grouping in the training (or test) sets, the metabolites were stratified based on output labels with a ratio of 0.9/0.1.

2.5 Architectures

We used three graph architectures using convolutions and compared their performance: Graph Convolutional Network (GCN), Graph Isomorphism Network (GIN) and Graph Attention Network (GAT) (29) (30). GCNs extend the concept of convolution from grid-like data (such as images) to graph data, allowing the aggregation of feature information from neighboring nodes. This approach effectively captures local graph structure and node features. GINs are designed to be powerful for graph isomorphism, making them capable of distinguishing a wide variety of graph structures. They achieve this by using a multi-layer perceptron (MLP) to aggregate node features, enhancing their discriminative power. GATs introduce an attention mechanism to GNNs, enabling nodes to assign different importance weights to their neighbors. This allows for more flexible and expressive feature aggregation, potentially improving performance on tasks where certain neighbors have more influence than others. Each architecture is made of 2 convolutional layers followed by a fully connected layer. The Adam optimizer with the binary cross entropy (BCE) loss is used for all the models. Sum aggregation is used to aggregate the features in order to do graph classification.

2.6 Embedding calculation

To improve model performances, pre-trained models are commonly used to include extra information. We tested this hypothesis with the pre-trained model ChemBERTa (31), which is inspired by the BERT (32) large language model (LLM) applied to molecules. This chemical model was trained on 77M unique SMILES annotations from PubChem (33). For each part of the SMILES, an embedding, i.e. a vector capturing important semantic and structural features of the molecule, is obtained. By averaging these embeddings, we obtained a vector representation of each molecule which is added before the fully connected layer of the model (Fig. 3).

2.7 Metrics

Due to the limited size of the dataset used (3'278 metabolites) and the small proportion of metabolites with specific ontology terms, the model is susceptible to generate a high number of false negatives. Accuracy is thus not a suitable metric to evaluate the model. We used instead the macro F1-score, which combines precision and recall. The true positive rate (TPR), also known as recall, was also used since we were mainly interested in knowing which ontology terms are associated with a given metabolite.

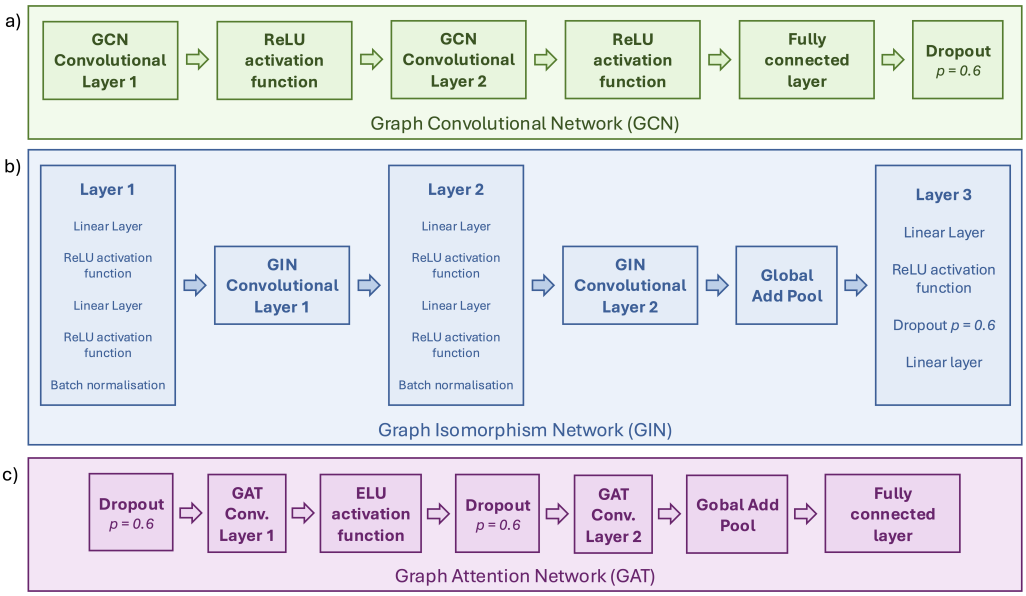

Figure 4: Summary of the tested architectures. a) Graph Convolutional Network (GCN) b) Graph Isomorphism Network (GIN) c) Graph Attention Network (GAT).

Table 1: Hyper-parameters

Layers	Features	Heads	Dimension	Epochs	Threshold	Learning Rate	Weight Decay
2	64	8	32	20	0.5	0.005	0.001

2.8 Model comparison and hyper-parameters

We performed a model comparison using five-fold cross validation to select the best architecture and related parameters for each level 1 node. The model's general structure can be seen in Fig. 3 and Fig. 4. All the models were trained using the same hyper-parameters that are listed in table 1.

2.9 Attention nodes

The GAT architecture offers a class called "Explainer" which allowed us to easily interpret some of the results of our study. The GAT architecture we used consists of eight heads which are composed of attention nodes. An attention node has the advantage, as compared to a regular node, of giving a different importance (or weight) to each of its neighbors. These different weights are useful for classification and pattern detection. Our model uses the max attention explainer algorithm, which selects the biggest weight across the different heads for a given node or bond. Min-max scaling was used across each metabolites' weights to see more distinct patterns.

3 Results

3.1 Model comparison

Overall, we obtained better results for models including the pre-trained embeddings (Tables 2, 3). Using ChemBERTa embeddings improved the F1-score by an average of 7.96% and the recall by 5.88%. The best architecture depends on the function category (location, role, process, and physiological effect) and the use or not of ChemBERTa embeddings. The GAT architecture using ChemBERTa embeddings predicting "Process" nodes yielded the best macro F1-score (0.889). The GIN architecture without ChemBERTa embeddings predicting "Disposition" nodes yielded the best recall (0.960). The "Physiological effect" nodes yielded poorer results with a best F1-score of 0.393.

Table 2: Metrics evaluated on the three different architectures without using ChemBERTa embeddings.

METRIC	CATEGORY	GCN	GIN	GAT
F1-Score	Disposition	0.624 ± 0.003	$\mathbf{0.645 \pm 0.004}$	0.601 ± 0.003
	Role	0.811 ± 0.008	$\mathbf{0.822 \pm 0.009}$	0.706 ± 0.012
	Process	0.834 ± 0.009	$\mathbf{0.875 \pm 0.007}$	0.823 ± 0.009
	Physiological Effect	0.278 ± 0.006	$\mathbf{0.287 \pm 0.002}$	0.147 ± 0.050
Recall	Disposition	0.918 ± 0.013	$\mathbf{0.960 \pm 0.019}$	0.903 ± 0.008
	Role	$\mathbf{0.834 \pm 0.009}$	0.831 ± 0.014	0.776 ± 0.014
	Process	0.841 ± 0.011	$\mathbf{0.895 \pm 0.005}$	0.830 ± 0.014
	Physiological Effect	$\mathbf{0.278 \pm 0.003}$	0.277 ± 0.002	0.157 ± 0.050

Table 3: Metrics evaluated on the three different architectures using ChemBERTa embeddings.

METRIC	CATEGORY	GCN	GIN	GAT
F1-Score	Disposition	$\mathbf{0.682 \pm 0.004}$	0.659 ± 0.007	0.672 ± 0.002
	Role	0.874 ± 0.007	0.857 ± 0.008	$\mathbf{0.877 \pm 0.006}$
	Process	0.883 ± 0.009	0.862 ± 0.014	$\mathbf{0.889 \pm 0.007}$
	Physiological Effect	$\mathbf{0.393 \pm 0.028}$	0.293 ± 0.003	0.368 ± 0.055
Recall	Disposition	0.880 ± 0.018	$\mathbf{0.942 \pm 0.006}$	0.939 ± 0.007
	Role	0.910 ± 0.020	0.874 ± 0.012	$\mathbf{0.928 \pm 0.007}$
	Process	$\mathbf{0.904 \pm 0.020}$	0.867 ± 0.015	0.903 ± 0.005
	Physiological Effect	$\mathbf{0.416 \pm 0.061}$	0.287 ± 0.005	0.353 ± 0.053

3.2 Cell membrane composition

To assess the biological relevance of our model, we focused on lipids and their presence in cell membranes. Given that lipids constitute the primary components of cell membranes, the model should demonstrate an understanding of this relationship. We therefore trained a model following the same methodology as described earlier, specifically designed to classify whether a given metabolite is found in the cell membrane or not. This approach enabled us to develop a more specialized model and identify patterns in its learning process. In the test set for this model, 203 metabolites are labeled as lipids by the HMDB, and 201 of those are found in the cell membrane. The model correctly predicted cell membrane localization for 194 of these metabolites, highlighting its accuracy in capturing this specific relationship (Fig. 5). Metabolites HMDB0000634 and HMDB0001452, which are both fatty acyls, were the only misclassified lipids for the cell membrane category. The various possible locations of fatty acyls could explain the more challenging nature of the predictions of the model for this family of lipids.

3.3 Interpretability

The use of attention nodes allowed us to look at the importance of certain bonds in the classification process. Patterns were found across triacylglycerols, which are a family of lipids mostly found in the cell membrane. Triacylglycerols are made of three fatty acyls and linked by a glycerol known as the head of the lipid. The most important element for the classification of these metabolites as cell-membrane metabolites appeared to be their head (Fig. 5). The bonds are colored using the attention weights meaning that a darker bond is more important for classification. The darker pattern shows the importance of bonds near the head of the molecule compared to the fatty acyls. The bonds that are the most important in the classification of the majority of triacylglycerols seem to be less important in the correct classification of metabolite HMDB0010445.

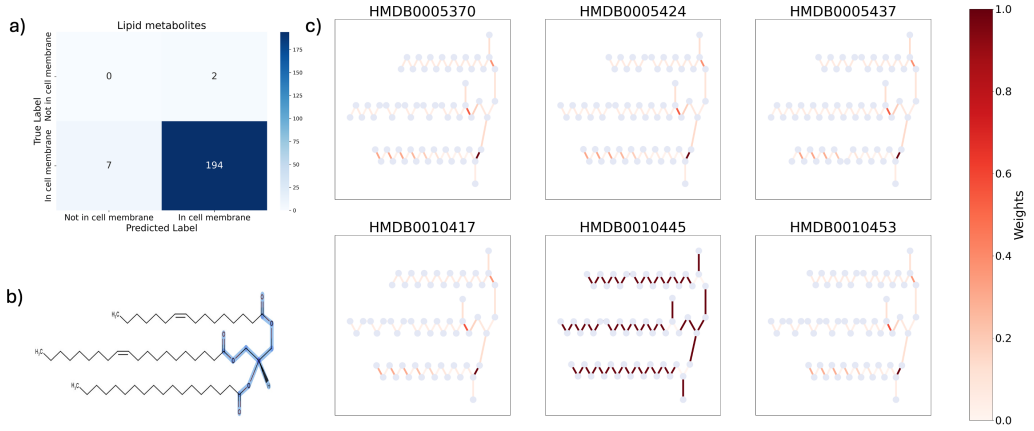

Figure 5: Interpretability analysis for the ontology term "Cell Membrane". a) Heat map for the confusion matrix of "Cell membrane" predictions for lipid metabolites. b) Visual representation of metabolite HMDB0005424 which belongs in the triacylglycerol family. The head of the metabolite, weighted important for the classification process, is highlighted. c) Visual representation of triacylglycerol metabolites present in the test set with colored bonds based on the weight attributed by the model. Darker bonds are more important in the classification process.

4 Discussion

We built a graph neural network model to predict multiple metabolite functions (location, role, process, and physiological effect) based on molecular structures. We compared various commonly used architectures, and the pre-trained ChemBERTa embeddings improved the results in the vast majority of the categories. The small number of output nodes for "Process" and "Role" compared to the "Disposition" nodes could in part explain the higher performance on these categories in general, as multi-label prediction generally becomes more challenging as the number of labels increases. The model underperformed on the "Physiological Effect" task compared to the others, which could be because measurable physiological effects might be driven by differences in metabolite concentrations rather than just their presence and structure, on top of many additional possible confounders. In the future, incorporating concentration data could improve the model's accuracy.

No clear distinction in the HMDB is made between missing information and a given metabolite not having the function represented by a specific ontology term – there are no true negatives annotated in the dataset. This, combined with the general sparsity of annotations, contributes to the model having a large number of false negative results.

In future work, we plan to tune the various hyper-parameters of the architecture, such as the number of convolutional layers, to improve the performance of the model. An additional next possible step to improve the performance could be to use exact ChemBERTa embeddings for each part of the metabolite instead of using mean embeddings as is currently done in our model. In this case, the model would use as input and as node feature for each atom and bond the exact atom or bond embedding as given by ChemBERTa.

SMILES is commonly used in computational chemistry as an input for models. We decided to directly use coordinates here on the hypothesis that it would yield better results. A next step would be to perform an ablation study to evaluate this hypothesis and in the case where coordinates do not significantly improve results, a new model could be made using SMILES as input. This would also solve the previously mentioned limitation as it would allow for easier integration of the ChemBERTa embedding of each part of the molecule.

Kengkanna et al. (2024) (34) introduced novel molecular graph representations by leveraging graph reduction techniques to enhance the capture of chemical substructures, including functional groups, chemical fragments, and pharmacophoric features. A promising avenue for future research lies in utilizing these graph reduction methods to provide varying levels of detail about key characteristics

relevant to compound property identification and interaction profiling, thereby potentially improving metabolites function prediction.

Our model establishes the feasibility of predicting functional ontology terms for metabolites based on their structural information. The availability of larger and more comprehensive datasets will be crucial for advancing the accuracy and applicability of machine learning techniques in this domain.

5 Data availability

All the code is available at `https://github.com/TancrediCogne/MetaboliteGNN` and the data needed to run the files can be downloaded directly from the HMDB website.

6 Competing interests

No competing interest is declared.

7 Author contributions statement

M.A.O conceived the experiment, T.C conducted the experiment. T.C, M.A.O, and A.S analysed the results. T.C, M.A.O, A.S, J.H, and J.F wrote and reviewed the manuscript.

References

[1] I. Nobeli, H. Ponstingl, E.B. Krissinel, and J.M. Thornton. A structure-based anatomy of the e.coli metabolome. *Journal of Molecular Biology*, 334(4):697–719, December 2003.

[2] V. Hatzimanikatis, C. Li, J.A. Ionita, and J.L. Broadbelt. Metabolic networks: enzyme function and metabolite structure. *Current Opinion in Structural Biology*, 14(3):300–306, June 2004.

[3] L.M. Cruz, S. Trefflich, V.A. Weiss, and M.A.A. Castro. Protein function prediction. In M. Kaufmann, C. Klinger, and A. Savelsbergh, editors, *Functional Genomics*, volume 1654 of *Methods in Molecular Biology*. Humana Press, New York, NY, 2017.

[4] M. Kulmanov, M.A. Khan, and R. Hoehndorf. Deepgo: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, February 2018.

[5] T. Zhao, Y. Hu, and L. Cheng. Deep-drm: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Briefings in Bioinformatics*, 22(4), July 2021.

[6] B. Lai and J. Xu. Accurate protein function prediction via graph attention networks with predicted structure information. *Briefings in Bioinformatics*, 23(1), January 2022.

[7] V. Gligorijević, P.D. Renfrew, and T. et al. Kosciolek. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12:3168, 2021.

[8] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(suppl1):D344–D350, January 2008.

[9] D.S. Wishart, S. Girod, H. Peters, E. Oler, J. Jovel, Z. Budinski, R. Milford, V.W. Lui, Z. Sayeeda, R. Mah, W. Wei, H. Badran, E. Lo, M. Yamamoto, Y. Djoumbou-Feunang, N. Karu, and V. Gautam. Chemfont: The chemical functional ontology resource. *Nucleic Acids Research*, 51(D1):D1220–D1229, 2023.

[10] D.S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B.L. Lee, M. Berjanskii, R. Mah, M. Yamamoto, J. Jovel, C. Torres-Calzada, M. Hiebert-Giesbrecht, V.W. Lui, D. Varshavi, D. Varshavi, D. Allen, D. Arndt, N. Khetarpal, A. Sivakumaran, K. Harford, S. Sanford, K. Yee, X. Cao, Z. Budinski, J. Liigand, L. Zhang, J. Zheng,

R. Mandal, N. Karu, M. Dambrova, H.B. Schiöth, R. Greiner, and V. Gautam. Hmdb 5.0: the human metabolome database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, January 2022.

[11] Maxat Kulmanov, Francisco Guzmán-Vega, and Robert Hoehndorf. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence*, 6(2):220–228, 2024.

[12] Enbin Yang, Hao Zhang, Zinan Zang, Zhiyong Zhou, Shuo Wang, Zhen Liu, and Yuanning Liu. Gcnfold: A novel lightweight model with valid extractors for rna secondary structure prediction. *Computers in Biology and Medicine*, 164:107246, 2023.

[13] Lu Meng and Xiaoran Wang. Tawfn: a deep learning framework for protein function prediction. *Bioinformatics*, 40(10):btae571, 2024.

[14] Thereza A. Soares, Ariane Nunes-Alves, Angelica Mazzolari, Fiorella Ruggiu, Guo-Wei Wei, and Kenneth Merz. The (re)-evolution of quantitative structure–activity relationship (qsar) studies propelled by the surge of machine learning methods. *Journal of Chemical Information and Modeling*, 62(22):5317–5320, 2022. PMID: 36437763.

[15] Y. Wang, L. Juan, J. Peng, T. Wang, T. Zang, and Y. Wang. Explore potential disease related metabolites based on latent factor model. *BMC Genomics*, 23(Suppl 1):269, 2022.

[16] F. Sun, J. Sun, and Q. Zhao. A deep learning method for predicting metabolite–disease associations via graph neural network. *Briefings in Bioinformatics*, 23(4), July 2022.

[17] Chen Qu, Barry I. Schneider, Anthony J. Kearsley, Walid Keyrouz, and Thomas C. Allison. Applying graph neural network models to molecular property prediction using high-quality experimental data. *Artificial Intelligence Chemistry*, 2(1):100050, 2024.

[18] Erik D. Huckvale and Hunter N. B. Moseley. Predicting the pathway involvement of metabolites based on combined metabolite and pathway features. *Metabolites*, 14(5), 2024.

[19] Vasiliy Porokhin, Li-Ping Liu, and Soha Hassoun. Using graph neural networks for site-of-metabolism prediction and its applications to ranking promiscuous enzymatic products. *Bioinformatics*, 39(3):btad089, March 2023.

[20] Martin Glauer, Adel Memariani, Fabian Neuhaus, Till Mossakowski, and Janna Hastings. Interpretable ontology extension in chemistry. *Semantic Web*, 15(4):937–958, 2024.

[21] S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv*, 2020. arXiv:2010.09885, Last revised 23 October 2020.

[22] Y. Wang, J. Wang, Z. Cao, et al. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279–287, 2022.

[23] Yuval Itan, Liang Shang, Bertrand Boisson, Etienne Patin, Alexandre Bolze, Mariana Moncada-Vélez, Emma Scott, Mary Jane Ciancanelli, Fernando G Lafaille, Janet G Markle, Ruben Martinez-Barricarte, Simone J de Jong, Xi-Feng Kong, Patrick Nitschke, Abdelali Belkadi, Jacinta Bustamante, Anne Puel, Stephanie Boisson-Dupuis, Peter D Stenson, Joseph G Gleeson, David N Cooper, Lluis Quintana-Murci, Jean-Michel Claverie, Shen-Ying Zhang, Laurent Abel, and Jean-Laurent Casanova. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences*, 112(44):13615–13620, Nov 2015.

[24] Y. Song, S. Chang, J. Tian, W. Pan, L. Feng, and H. Ji. A comprehensive comparative analysis of deep learning based feature representations for molecular taste prediction. *Foods*, 12:3386, 2023.

[25] S. Raghunathan and U.D. Priyakumar. Molecular representations for machine learning applications in chemistry. *Quantum Chemistry*, 122(7):April 5, 2022.

[26] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, October 2017.

[27] A. Daigavane, B. Ravindran, and G. Aggarwal. Understanding convolutions on graphs. September 2021.

[28] L. Colliandre. Molecular graphs as input for neural networks. February 2020.

[29] D.C. Elton, Z. Boukouvalas, M.D. Fugea, and P.W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design and Engineering*, 4:828–849, 2019.

[30] B. Sanchez-Lengeling, E. Reif, A. Pearce, and A.B. Wiltschko. A gentle introduction to graph neural networks. September 2021.

[31] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models, 2022.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2019. arXiv:1810.04805, Last revised 24 May 2019.

[33] Sunghwan Kim, Jie Chen, Tiejun Cheng, Andrey Gindulyte, Jianfeng He, Shan He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bastian Yu, Ludmila Zaslavsky, Jian Zhang, and Evan E Bolton. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109, January 2019.

[34] A. Kengkanna and M. Ohue. Enhancing property and activity prediction and interpretation using multiple molecular graph representations with mmgx, 2024.