

A deep learning architecture for metabolic pathway prediction

Mayank Baranwal, Abram Magner, Paolo Elvati, Jacob Saldinger, Angela Violi, and Alfred O. Hero

Supplementary Material

Dataset

A dataset of 6669 compounds belonging to one or more of these 11 constituent pathway classes was downloaded (February 2019) from the KEGG database: <https://www.genome.jp/kegg/pathway.html>. Figure 1 shows the distribution of compounds across 11 constituent pathway classes in KEGG database.

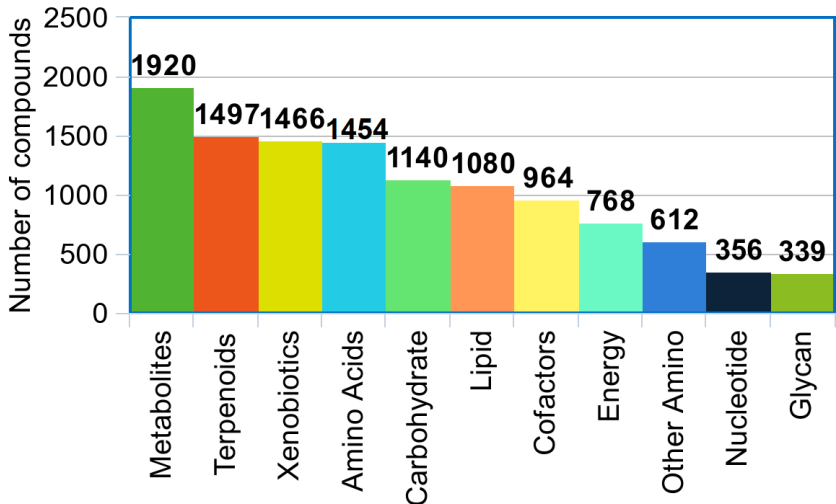


Figure 1: Distribution of 6669 compounds across multiple pathway classes in the KEGG database.

Feature Importance Analysis

We turn to the analysis of the relative importance of various features. The main findings are reported. First, the output of the GCN alone can achieve better performance than using only expert-chosen descriptors and MACCS fingerprints, so that the GCN architecture is a valuable addition to the chemical analysis toolbox. Second, our data-driven analysis of the relative importance of the expert-chosen descriptors and fingerprints (without the GCN output) is in agreement with previously reported experiments.

We use *Shapley additive explanations* (abbreviated *shap*) to estimate the average contribution of each feature to the classifier’s output in the presence of a uniformly random subset of other features. The average contribution is quantified for that feature. Unlike other methods, e.g., greedy feature selection, the *shap* approach computes an average importance in the presence of random subsets of other features. Moreover, in its most basic form, *shap* is the *only* feature importance analysis method that satisfies natural axioms [3]:

- **Efficiency:** The Shapley values for all features of a given data instance to be explained must sum to the difference between the value of the model output on that instance and the average value of the model output (taken over random instances from the data set). In other words, the difference between the prediction of a model on a given input and its average prediction is completely captured by the Shapley values of all of the features associated with that input.
- **Symmetry:** For a pair of features, if including either one or the other in the presence of an arbitrary set of other features results in the same model output, then the Shapley values of the two features are equal.

- **Dummy:** If a predictive model’s output is constant with respect to a given feature, regardless of the inclusion or exclusion of other features, then the Shapley value of that feature is 0.
- **Additivity:** If a predictive model output is a linear combination $c_1x_1 + \dots + c_kx_k$ of the outputs $\{x_i\}_{i=1}^k$ of k other models, then the Shapley value ϕ of a given feature for the combined model is given by $\phi = c_1\phi_1 + \dots + c_k\phi_k$, where ϕ_i is the Shapley value for the given feature with respect to the i th model. For instance, the random forest classifier computes an average output of the outputs of its constituent decision trees. The Shapley value for a given feature for the classifier can thus be expressed as an average of the Shapley values of that feature for each of the decision trees.

We apply the Shapley method to rank the importance of features in terms of their contribution to the performance of the classifier in the following settings:

- random forest with global feature inputs,
- GCN with feedforward network output and global feature inputs,
- random forest with GCN embeddings and global feature inputs.

In the following paragraphs, for each setting we illustrate Shapley value estimates for the features in the multi-class setting via two diagrams: the first is a heatmap, where the intensity of the color at a given position indicates the Shapley value for the corresponding feature for predicting membership in the corresponding pathway class. The second gives cumulative information via a bar plot, where each feature has a corresponding bar, and the length of each bar gives the sum of the absolute Shapley values for that feature across all pathway classes.

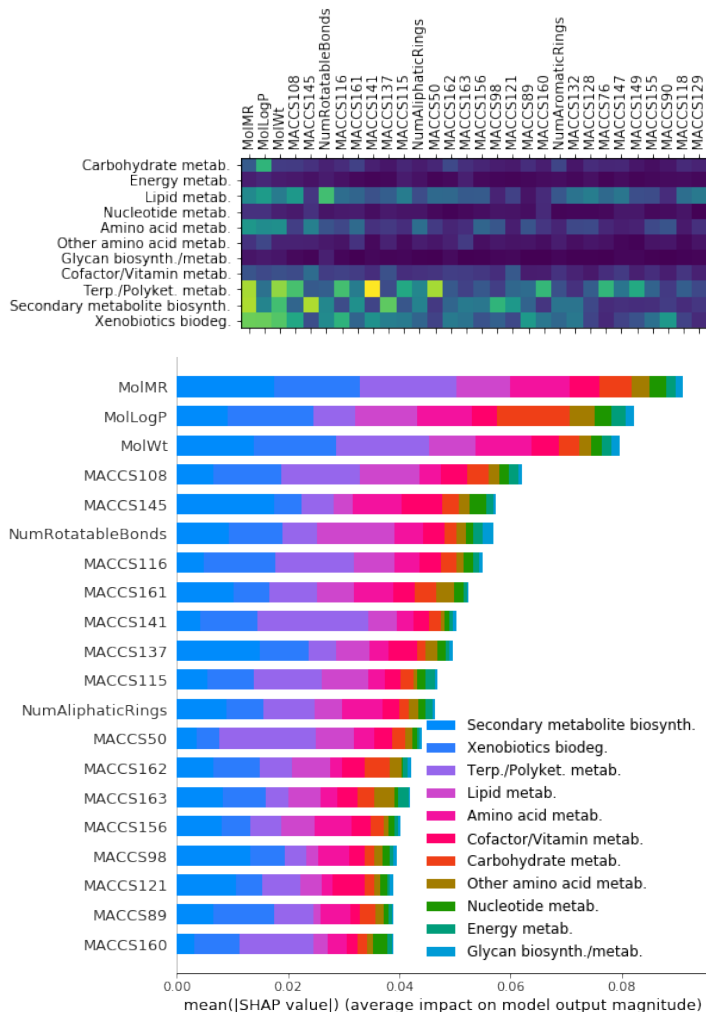


Figure 2: Top 20 Shapley values for Random Forest with global molecular features as input.

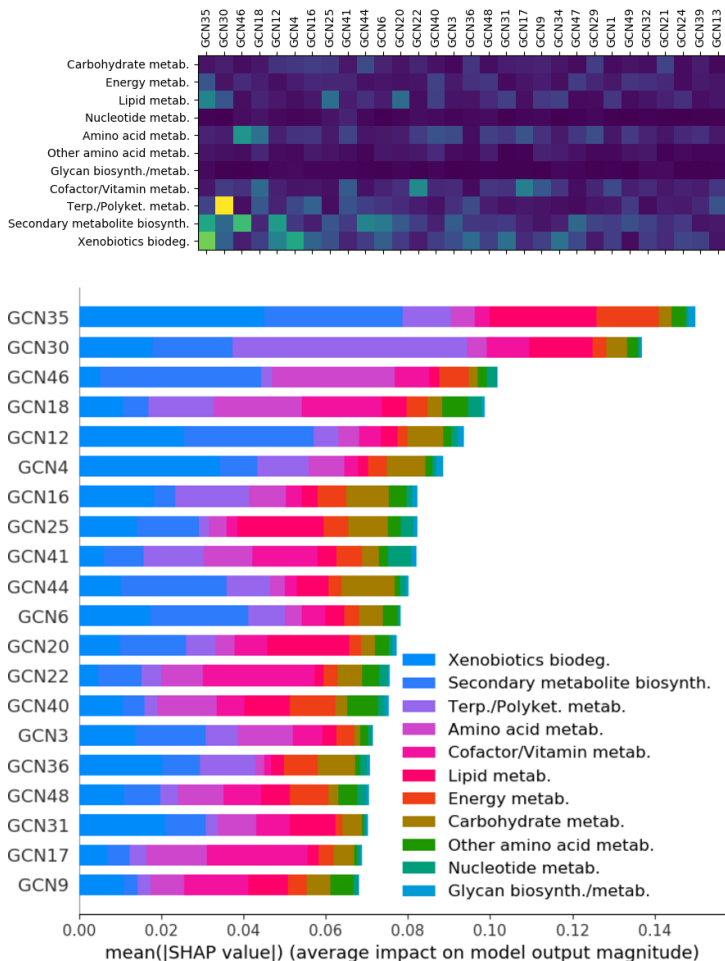


Figure 3: Top 20 Shapley values for GCN with global molecular features as input and feedforward network as output.

Feature Analysis For RF With Global Molecular Features.

In Figure 2, we plot Shapley value estimates for the top twenty global molecular in the multi-class setting. The features identified by the Shapley analysis correspond to factors that are known to influence small molecule metabolism. The most important identified feature is molar refractivity. This property correlates to the polarizability [13] of a molecule which has been shown to be an important property for predicting biological interactions [7] as it describes how a molecule’s charge distribution is affected by an applied electric field such as in solution [1]. The second most important feature is the octanol-water partition coefficient, which is a measure of the molecule lipophilicity. This finding is in agreement with past studies [10,11] that have identified $\log P$ as one of the most important factors in distinguishing molecule interaction potency, due to its correlation with substrate transport and binding properties [8]. While $\log P$ is indicative of the molecule’s propensity to reach and interact with the correct environment, it is not a specific measure for the interaction with a protein or enzymatic site. These interactions are primarily controlled by steric factors, and this general knowledge is matched by several of the top features, like the number of rotatable bonds, mass (as a rough estimator of the size), and several descriptors (e.g. MACCS 98, 121, 137, 145) related to rings, which are generally considered inflexible [5]. Beyond solubility and flexibility, the analysis indicates that other features are important, such as the presence of aromatic sites, as it promotes hydroxylation, which is key in many metabolic pathways [4,6], or the presence of methyl groups (MACCS 50, 108, 116, 141, 149, 160), which promote selectivity towards hydrophobic binding sites [12] and induce favorable binding conformations [2,9]. Finally, the presence of hydrogen bond donors and acceptors (MACCS 89, 136, 161) is the last group of features that was identified as relevant, in overall agreement with the work by Lipinski et al. [12].

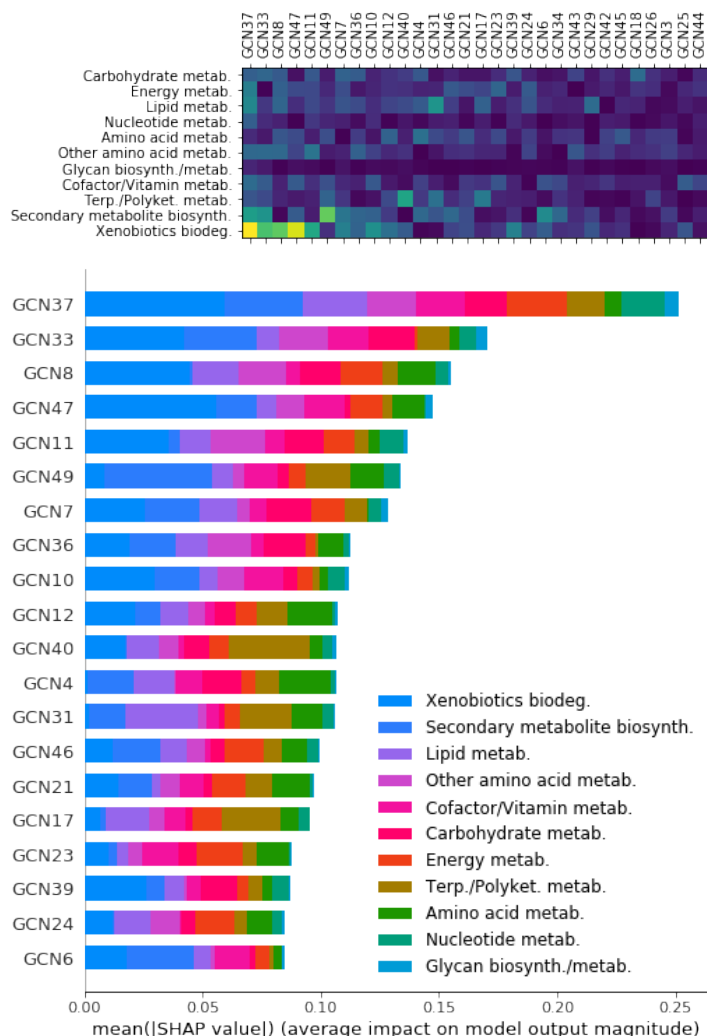


Figure 4: Top 20 Shapley values for Random Forest with GCN embeddings and global molecular features as input.

Feature Analysis For GCN With Global Feature Inputs And Feed-Forward Network Output.

In Figure 3, we give the Shapley plots for GCN with global feature inputs and feed-forward network output. We see, first, that the most important features for the prediction task, in a cumulative sense, are the GCN embedding features. Furthermore, for most pathway classes, there is no single feature that is overwhelmingly important. This indicates, first, that in the presence of GCN embeddings, global molecular features are treated as redundant by the feed-forward network, and, second, that information that is useful in distinguishing pathway classes is spread across various learned shape features.

Feature Analysis For RF With GCN Embeddings And Global Features.

In Figure 4, we give the Shapley plots for random forest with GCN embeddings and global features as input. We may draw similar conclusions from these plots as for those in the case of the GCN with feedforward network output and global feature input. However, we note that *different* GCN embedding features are important for the random forest output, in comparison to the feed-forward network output. We next delve more deeply into the analysis and interpretation of the GCN embedding features. Among other insights, we will give evidence that at least some GCN embedding elements that are highly ranked according to the cumulative Shapley analysis are capturing information about the same graph structural properties.

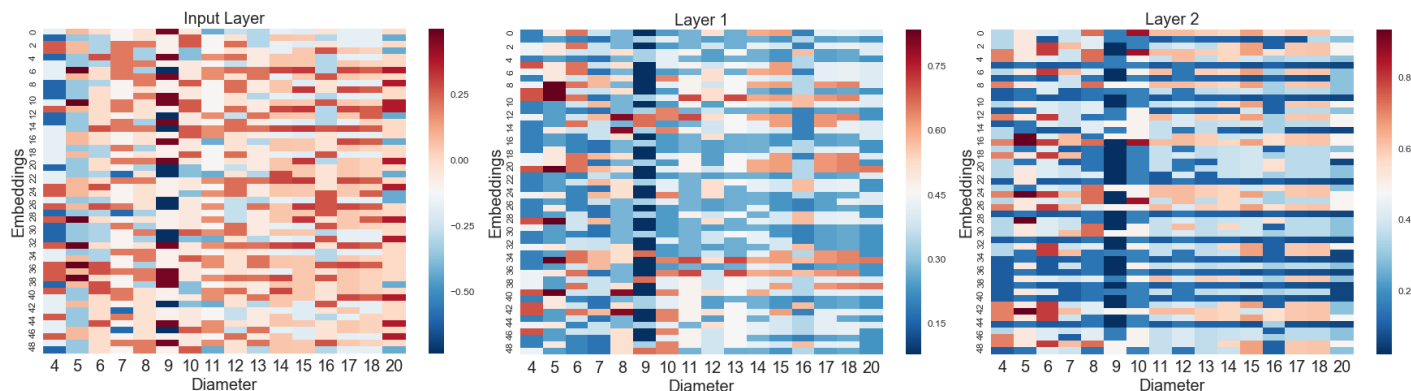


Figure 5: Layer-wise heatmaps of GCN embeddings of medoid molecules corresponding to different graph diameters.

Analysis Of GCN Embeddings

In addition to our exploration of the GCN embedding vectors in relation to global molecular features, we performed experiments to elucidate the interpretation of the embedding elements in terms of class-wise and purely graph-theoretic properties of the molecules.

The general pattern of the experiment was as follows: consider a graph property $f : G \rightarrow S$, where G is the space of molecular graphs, and S is an arbitrary range. For example, $f(G)$ may be the diameter of the graph G , in which case S is the set of non-negative integers. We partition the test dataset into the *level sets* of f , resulting in partition elements S_x , for each $x \in S$. When f is the diameter function, S_x is the medoid molecule in the test dataset with diameter exactly x .

Now, for each medoid molecule and each fixed GCN layer $\ell \in \{0, 1, 2\}$, we compute the GCN embedding vector $v_{G,\ell}$ of the graph after the ℓ th layer. In general, the interpretation of this is as follows: as deeper layers of the GCN are explored, if there are many elements of embedding vectors that are correlated for each particular x , then this indicates a (potentially complicated) functional relationship between the activation of those embedding elements and the value $f(G)$. In such a case, we will say that the embedding elements are *sensitive* to f . We can visually depict the results of this experiment using layer-wise heatmaps (Figure 5) for the medoid molecules shown in Figure 9. Each of these medoid molecules has a different *graph diameter*. Each row in the heatmap corresponds to a given embedding element index, and each column corresponds to a diameter value. It can be observed that the embeddings corresponding to diameters 14, 15, 17 and 18 appear to be highly correlated. Interestingly, the medoid molecules corresponding to these diameter values are also structurally very similar, indicating a potential mapping of graphical structures to GCN embeddings. This experiment was repeated for other graph properties (radius and average degree), and a similar trend was observed (see Figure 8).

Figure 7 depicts sensitivity of few GCN embeddings in the final layer to graph diameters. It can be observed from that the embedding indices, such as, 3, 17, 25 and 30, which are also ranked highly by the Shapley analysis, appear to be sensitive to diameter. On the other hand, embedding indices, such as, 10, 28, 41 and 45, which do not contribute significantly towards classification task, appear to be insensitive to graph diameters.

We also perform a class-wise partition of the test dataset and depict the GCN embeddings corresponding to each of the 11 metabolic pathway classes averaged over all molecules belonging to the same class in Figure 6. Unlike in the previous experiment, no two classes seem to have significantly correlated embeddings in this case. This is critical from the point of view of classification problem, since molecules from different classes must have significantly different embedding vectors in order to be differentiated easily from one another. All of this gives evidence that the GCN embedding reflects information about global graph properties (such as the graph diameter) and that these play an important role in the classification of the molecules into pathway classes in both the feed-forward neural network and random forest output cases.

Class-Wise Performance Of Multi-GCN Classifier

Table 1 (here) shows the class-wise confusion matrices for the multi-class GCN classifier evaluated on a held-out test set comprising of 1334 molecules. Note that in multiclass tasks, labels are binarized under a one-vs-rest

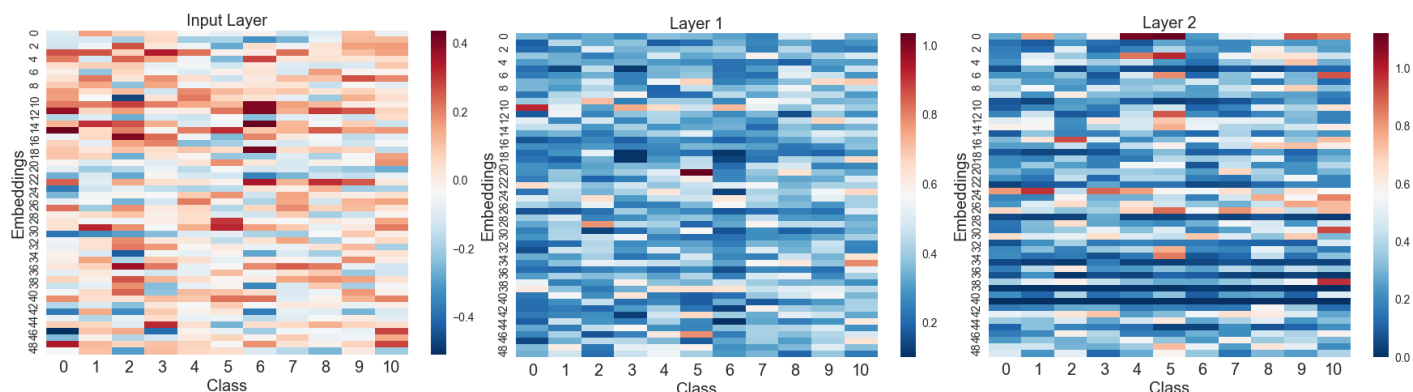


Figure 6: Layer-wise heatmaps of GCN embeddings of molecules belonging to different metabolic pathway classes. Note that there is sufficient variability among embeddings coming from two different classes, and thus GCN makes it easier for a classifier, such as, RF to be more effective at the classification task.

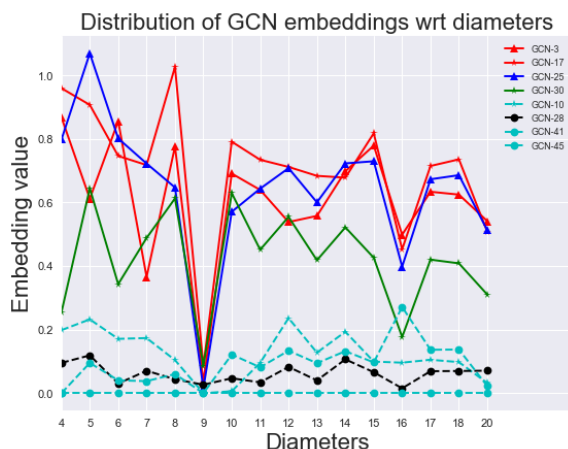


Figure 7: Sensitivity of selected GCN embeddings with respect to graph diameters. Features, such as, GCN-3, GCN-17, GCN-25 and GCN-30, which are identified as important by Shapley analysis, are sensitive to variation in graph diameters. On the other hand, features, such as, GCN-10, GCN-28, GCN-41 and GCN-45, which do not contribute significantly towards classification task, appear insensitive to variation in graph diameters.

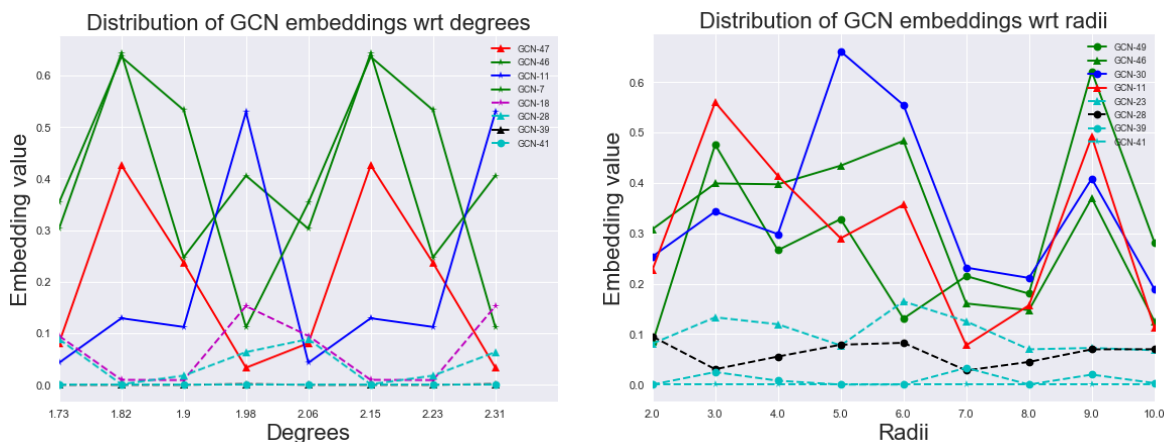


Figure 8: Sensitivity of selected GCN embeddings with respect to (a) average graph degrees, (b) graph radii. (a) Features, such as, GCN-47, GCN-48, GCN-11 and GCN-7, which are identified as important by Shapley analysis, are sensitive to variation in average graph degrees. On the other hand, features, such as, GCN-18, GCN-28, GCN-39 and GCN-41, which do not contribute significantly towards classification task, appear insensitive to variation in average graph degrees. (b) Features, such as, GCN-49, GCN-46, GCN-30 and GCN-11, which are identified as important by Shapley analysis, are sensitive to variation in graph radii. On the other hand, features, such as, GCN-23, GCN-28, GCN-39 and GCN-41, which do not contribute significantly towards classification task, appear insensitive to variation in graph radii.

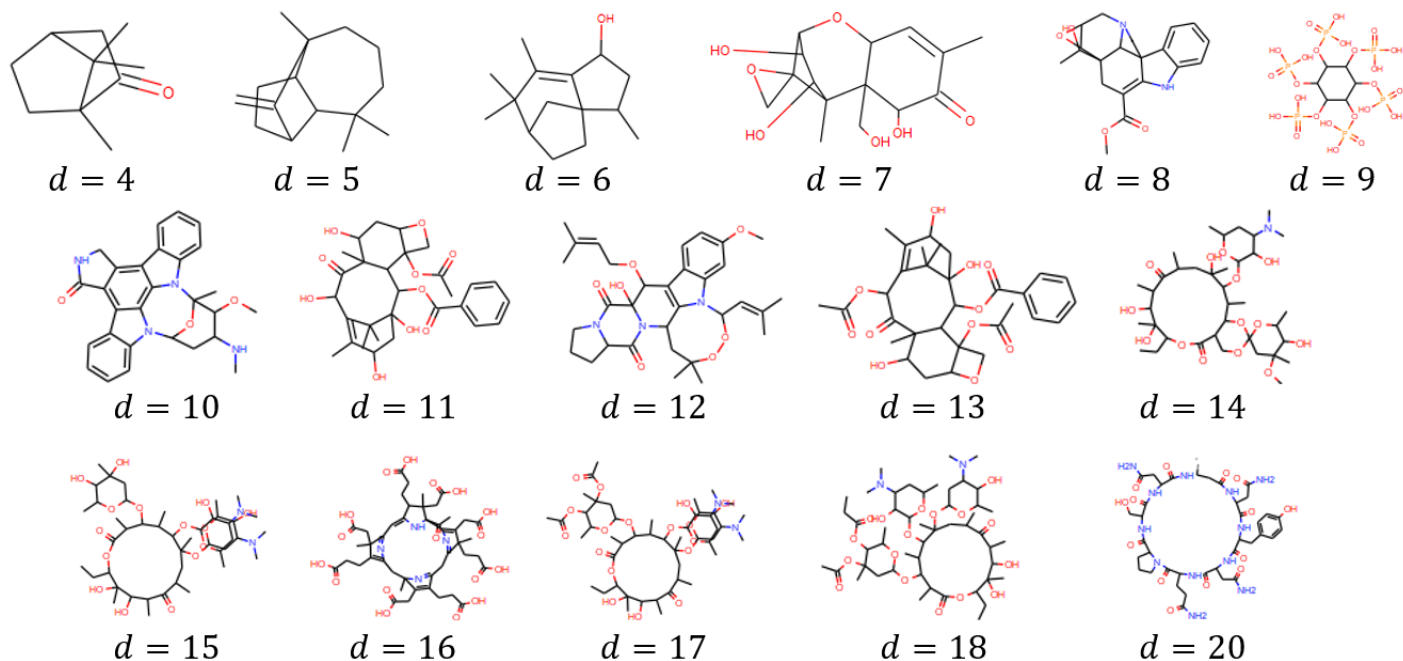


Figure 9: Medoid molecules corresponding to different graph diameter values: Note that the molecules corresponding to diameters 14, 15, 17 and 18 are structurally similar, and hence their GCN embeddings appear to be highly correlated in Figure 5.

Class	Scores	
	Confusion matrix	
<i>Carbohydrate</i>	1074	16
	13	231
<i>Energy</i>	1156	15
	13	150
<i>Lipid</i>	1090	24
	9	211
<i>Nucleotide</i>	1258	6
	7	63
<i>Amino acid</i>	1005	31
	26	272
<i>Other amino acids</i>	1199	12
	10	113
<i>Glycan</i>	1262	4
	6	62
<i>Cofactors/Vitamins</i>	1104	31
	20	179
<i>Terpenoids</i>	982	24
	18	310
<i>Sec. metabolites</i>	922	30
	34	348
<i>Xenobiotics</i>	1016	20
	27	271

Table 1: Class-wise confusion matrix of multi-class, mixed-membership GCN classifier

way, as in a mixed membership setup compounds can belong to multiple pathways. When calculating class-wise multilabel confusion matrix, the count of true negatives (TNs) for class i is $C_{i,0,0}$, false negatives (FNs) is $C_{i,1,0}$, true positives (TPs) is $C_{i,1,1}$ and false positives (FPs) is $C_{i,0,1}$.

Of these 1334 molecules, there are 958 molecules which belong to only one pathway class. (1) provides the standard 11×11 confusion matrix corresponding to the 11 identified metabolic pathway classes.

$$\begin{bmatrix}
 45 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
 0 & 1 & 109 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 12 & 2 & 0 & 0 & 0 & 0 & 0 & 1 \\
 1 & 0 & 1 & 0 & 35 & 0 & 0 & 1 & 0 & 1 & 2 \\
 0 & 0 & 0 & 0 & 1 & 14 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 & 1 & 0 & 0 & 46 & 1 & 3 & 0 \\
 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 185 & 2 & 1 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 2 & 188 & 2 \\
 1 & 0 & 2 & 0 & 0 & 0 & 0 & 3 & 0 & 2 & 140
 \end{bmatrix}. \quad (1)$$

Additionally, the class-wise accuracy, precision, recall and MCC scores are shown in Table 2.

References

-
- [1] I. Banik and M. N. Roy. Study of solute–solvent interaction of some bio-active solutes prevailing in aqueous ascorbic acid solution. *Journal of Molecular Liquids*, 169:8–14, 2012.

Class	Scores (%)			
	Accuracy	Precision	Recall	MCC
<i>Carbohydrate</i>	98.09 \pm .46	93.36 \pm 2.3	95.87 \pm .78	91.91 \pm .78
<i>Energy</i>	97.64 \pm .54	89.27 \pm 3.1	90.32 \pm 3.9	91.30 \pm 2.3
<i>Lipid</i>	98.24 \pm .46	93.75 \pm .68	96.13 \pm 2.0	94.65 \pm .67
<i>Nucleotide</i>	99.10 \pm .37	93.25 \pm 2.9	90.49 \pm 6.5	94.95 \pm 3.1
<i>Amino acid</i>	96.21 \pm .43	91.28 \pm 1.3	92.34 \pm 1.6	90.93 \pm .82
<i>Other amino acids</i>	98.13 \pm .62	90.61 \pm 3.8	91.13 \pm 2.0	89.87 \pm 1.9
<i>Glycan</i>	99.40 \pm .21	90.06 \pm 2.6	97.49 \pm 1.7	96.67 \pm 1.1
<i>Cofactors/Vitamins</i>	95.99 \pm .75	84.93 \pm 2.2	89.33 \pm 2.5	86.03 \pm 2.1
<i>Terpenoids</i>	96.47 \pm .50	91.51 \pm 1.4	93.30 \pm 1.6	90.26 \pm .81
<i>Sec. metabolites</i>	95.54 \pm .14	91.75 \pm 1.2	92.40 \pm .90	89.39 \pm .78
<i>Xenobiotics</i>	96.40 \pm .55	91.67 \pm 2.3	92.08 \pm 2.6	90.42 \pm 1.3

Table 2: Class-wise performance analysis of multi-class GCN classifier

- [2] E. J. Barreiro, A. E. Kmmmerle, and C. A. M. Fraga. The methylation effect in medicinal chemistry. *Chemical Reviews*, 111(9):5215–5246, 2011.
- [3] S. Cohen, E. Ruppim, and G. Dror. Feature selection based on the shapley value. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence, IJCAI’05*, pages 665–670, San Francisco, CA, USA, 2005. Morgan Kaufmann Publishers Inc.
- [4] B. M. El-Haj, S. B. Ahmed, M. A. Garawi, and A. H. S. Linking aromatic hydroxy metabolic functionalization of drug molecules to structure and pharmacologic activity. *Molecules*, 23(9), 2018.
- [5] T. J. A. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design*, 15(5):411–428, 2001.
- [6] G. Fuchs. Anaerobic metabolism of aromatic compounds. *Annals of the New York Academy of Sciences*, 1125(1):82–99, 2008.
- [7] C. Hansch, W. E. Steinmetz, A. J. Leo, S. B. Mekapati, A. Kurup, and D. Hoekman. On the role of polarizability in chemicalbiological interactions. *Journal of Chemical Information and Computer Sciences*, 43(1):120–125, 2003.
- [8] A. J. Leo and C. Hansch. Role of hydrophobic effects in mechanistic qsar. *Perspectives in Drug Discovery and Design*, 17:1–25, 1999.
- [9] C. S. Leung, S. S. F. Leung, J. Tirado-Rives, and W. L. Jorgensen. Methyl effects on proteinligand binding. *Journal of Medicinal Chemistry*, 55(9):4489–4500, 2012.
- [10] D. F. Lewis, M. N. Jacobs, and M. Dickins. Compound lipophilicity for substrate binding to human p450s in drug metabolism. *Drug Discovery Today*, 9(12):530–537, 2004.
- [11] D. F. V. Lewis, B. G. Lake, and M. Dickins. Quantitative structure-activity relationships (qsars) in inhibitors of various cytochromes p450: The importance of compound lipophilicity. *Journal of Enzyme Inhibition and Medicinal Chemistry*, 22(1):1–6, 2007.
- [12] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1):3 – 25, 1997.
- [13] S. A. Wildman and G. M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999.