Briefings in Bioinformatics, 2023, 1–12

https://doi.org/10.1093/bib/bbad189

Problem Solving Protocol

MPI-VGAE: protein–metabolite enzymatic reaction link learning by variational graph autoencoders

Cheng Wang, Chuang Yuan, Yahui Wang, Ranran Chen, Yuying Shi, Tao Zhang, Fuzhong Xue, Gary J. Patti, Leyi Wei and Qingzhen Hou

Corresponding authors. Gary J. Patti, Department of Chemistry, Washington University in St. Louis, St. Louis, MO 63130, USA. Tel: +1-3149353512. E-mail: gjpattij@wustl.edu; Leyi Wei, School of Software, Shandong University, Jinan 250100, China. Tel: +86-18526645848. E-mail: weileyi@sdu.edu.cn; Qingzhen Hou, Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, China. Tel: +86-18853122360. E-mail: houqingzhen@sdu.edu.cn

Abstract

Enzymatic reactions are crucial to explore the mechanistic function of metabolites and proteins in cellular processes and to understand the etiology of diseases. The increasing number of interconnected metabolic reactions allows the development of *in silico* deep learning-based methods to discover new enzymatic reaction links between metabolites and proteins to further expand the landscape of existing metabolite–protein interactome. Computational approaches to predict the enzymatic reaction link by metabolite–protein interaction (MPI) prediction are still very limited. In this study, we developed a Variational Graph Autoencoders (VGAE)-based framework to predict MPI in genome-scale heterogeneous enzymatic reaction networks across ten organisms. By incorporating molecular features of metabolites and proteins as well as neighboring information in the MPI networks, our MPI-VGAE predictor achieved the best predictive performance compared to other machine learning methods. Moreover, when applying the MPI-VGAE framework to reconstruct hundreds of metabolic pathways, functional enzymatic reaction networks and a metabolite–metabolite interaction network, our method showed the most robust performance among all scenarios. To the best of our knowledge, this is the first MPI predictor by VGAE for enzymatic reaction link prediction. Furthermore, we implemented the MPI-VGAE framework to reconstruct the disease-specific MPI network based on the disrupted metabolites and proteins in Alzheimer's disease and colorectal cancer, respectively. A substantial number of novel enzymatic reaction links were identified. We further validated and explored the interactions of these enzymatic reactions using molecular docking. These results highlight the potential of the MPI-VGAE framework for the discovery of novel disease-related enzymatic reactions and facilitate the study of the disrupted metabolisms in diseases.

Keywords: metabolite, metabolite–protein interaction, enzymatic reaction, graph neural network, machine learning

INTRODUCTION

Characterizing enzymatic reactions is important to understand biochemical transformations, allosteric inhibition and protein signaling [1–3]. Enzymatic reactions start with interactions between metabolites and proteins that occur at the active site of enzymes and are building blocks for metabolic networks [4, 5]. Functional annotations and characterizations of enzymatic reactions pave the way to understanding the metabolic mechanisms and uncovering associations between metabolomics and diseases in biomedical research [6, 7]. The rapid advances in high-throughput metabolomics and proteomics technologies promote the systematic profiling of metabolites and proteins

Cheng Wang received his PhD in chemistry from The Ohio State University, USA. He is currently an assistant professor in School of Public Health at Shandong University, China. His research interests include bioinformatics and machine learning-based approach with applications to biomedical networks.
Chuang Yuan is currently a research assistant at Shandong University. He received his MS in biology from the University of Science and Technology of China. His research interests include neuroscience, cancer biology, bioinformatics and computational biology.
Yahui Wang is a PhD student in the Department of Chemistry at Washington University in St. Louis. Her research interests include biochemistry, mass spectrometry-based metabolomics and cancer metabolism.
Ranran Chen is a master graduate student in the School of Public Health at University of Shandong, China. Her research interests include biostatistics and bioinformatics.
Yuying Shi is a master graduate student in the School of Public Health at University of Shandong, China. Her research interests include biostatistics and bioinformatics.
Tao Zhang is a professor in the School of Public Health, Shandong University, China. His areas of research are longitudinal statistical methods of life course and dynamic multi-omics in chronic diseases.
Fuzhong Xue is a professor in the School of Public Health and is the head of National Institute of Health Data Science of China, Shandong University, China. His areas of research are theoretical and methodological research on healthcare big data, and the development of theoretical causal inference methods on big data.
Gary J. Patti is the Michael and Tana Powell Professor at Washington University in St. Louis, where he holds appointments in the Department of Chemistry and the Department of Medicine. He is also the senior director of the Center for Metabolomics and Isotope Tracing at Washington University. His research interests include metabolomics, bioinformatics, high-throughput mass spectrometry, environmental health, cancer and aging.
Leyi Wei received his PhD in computer science from Xiamen University, China. He is currently a professor in the School of Software at Shandong University, China. His research interests include machine learning and its applications to bioinformatics.
Qingzhen Hou received his PhD in the Centre for Integrative Bioinformatics VU (IBIVU) from Vrije Universiteit Amsterdam, the Netherlands. Since 2020, he has served as the head of Bioinformatics Center in National Institute of Health Data Science of China and Assistant Professor in School of Public Health, Shandong University, China. His areas of research are bioinformatics and computational biophysics.

Received: February 20, 2023. **Revised:** April 10, 2023. **Accepted:** April 27, 2023

[8–10]. The discovery of novel enzymatic reactions is essential to investigating their mechanistic roles in cellular metabolisms and disease progression. Since the interaction of enzyme and substrate, namely, enzymatic reaction link, is a prerequisite for an enzymatic reaction, accurate identification of metabolite–protein interactions (MPIs) would facilitate the discovery of new enzymatic reactions. Experimental approaches have been developed to systematically discover enzymatic reaction links by mapping MPIs in cells [11–13]. For example, 1678 interaction pairs between 20 designated metabolites and cellular proteins in *Escherichia coli* were experimentally determined by mass spectrometry [11]. High-resolution NMR relaxometry was recently developed to detect MPIs in biological fluids [14]. Although these experimental methods offer high reproducibility and accuracy to characterize enzymatic reaction links, the low binding affinity of MPIs and labor-intensive sample preparation hamper the process for large-scale enzymatic reaction screening.

Recently, a variety of machine learning-based methods have been developed to predict MPI computationally, while most of these methods focus on predicting the allosteric interaction instead of the likelihood of enzymatic reaction [15, 16]. It is imperative to develop efficient computational approaches to predict MPI for enzymatic reaction link identification. Thousands of enzymatic reactions have been cataloged in multiple metabolic databases such as KEGG, Reactome and PathBank [17–19]. These interconnected graph-based representations of enzymatic reactions generate genome-scale metabolic networks, while there are still no computational methods to fully explore the metabolic reaction networks for MPI prediction.

The prediction of MPI based on the enzymatic reaction network takes advantage of interconnected features of proteins and metabolites, which can be formulated as a so-called 'link prediction' computational problem [20, 21]. Graph neural networks (GNNs) integrate the graph topology and node/edge features and show superb performance of link prediction than traditional machine learning methods [20, 22, 23]. Graph neural networks have been widely applied to recognize the link in network properties of biomolecules ranging from protein structure prediction, protein–protein interaction networks and protein–RNA binding, to multi-omics disease studies [24–29]. However, there are no available GNN methods for MPI prediction within the enzymatic reaction network.

In this study, we constructed a heterogeneous network of metabolite–protein functional interaction networks from thousands of enzymatic reactions and developed a Variational Graph Autoencoders (MPI-VGAE) framework to predict enzymatic reaction links for different organisms. Ten organism-wise MPI networks were constructed and MPI prediction with the VGAE model was trained and optimized. By comparing with conventional similarity-based and graph-based methods, we demonstrated that the MPI-VGAE method outperformed other models for MPI prediction in different genome-scale MPI networks with the highest AUC and Average precision (AP) scores. The heterogeneous node features of metabolites and proteins and neighboring information were well incorporated into the MPI-VGAE framework via the feature transformation module. We applied MPI-VGAE framework to multiple scenarios, including the reconstruction of metabolic pathways, functional metabolic networks and homogeneous metabolic reaction networks. Finally, the MPI-VGAE framework was applied to study Alzheimer's disease and colorectal cancer to reconstruct the MPI network by hundreds of disrupted metabolites and proteins. MPI-VGAE could predict new potential enzymatic reactions, which were further investigated with the possible binding poses for several examples. The MPI-VGAE framework will facilitate the discovery of novel enzymatic reactions in biomedical research.

METHODS

Datasets and characteristics of metabolites and proteins

In enzymatic reactions, metabolite (substrate) interacts with protein (enzyme), which is denoted as an enzymatic reaction link in the present study. The aim of MPI-VGAE is to predict the enzymatic reaction link by integrating the molecular information of metabolites and proteins and the MPI network. To construct a MPI network, the metabolite–protein interactions were extracted from all metabolic pathways in PathBank. Each metabolic pathway contains metabolites and protein information. The metabolite–protein interaction networks were constructed and specified for ten organisms separately, such as *Homo sapiens*. In the MPI network, each metabolite and protein was modeled as a node and each interaction was modeled as an edge. The classes of proteins and metabolites were characterized and classified by using BRENDA and RefMet. The dataset curation process is depicted in Figure 1A. The homogenous metabolic reaction network was constructed based on the chemical reactions in the KEGG database. Metabolites in all pathways of different organisms in the KEGG database were extracted via KEGG API. In the metabolic reaction network, each metabolite was modeled as a node, and the metabolite reactants and products in each metabolic reaction were modeled as edges.

Feature representation of metabolites and proteins

The vectorized features of metabolites and proteins were generated as input for the graph-based neural network models. For metabolites, molecular fingerprints were used to represent the features of each metabolite. A molecular fingerprint uses a series of binary digits to indicate the presence or absence of a particular substructure in the molecule. Three popular molecular fingerprint representations were considered for comparison, including the Extended Connectivity Fingerprint (ECFP), MACCS keys and topological fingerprints [30, 31]. RDKit (https://www.rdkit.org/) was used to generate molecular fingerprints based on the SMILES string of each molecule, which yielded a binary vector with a fixed-length [32]. In addition, the fingerprints of metabolites were further normalized and transformed by using principal component analysis (PCA). Essentially, all fingerprints of 78 726 metabolites were fit by using PCA and transformed into a 1024-bit vector. For proteins, the raw amino-acid sequence was used to capture the feature information and converted into a vector with a length of 1024 based on a pre-trained SeqVec model and an ESM-1b transformer model [33, 34]. The SeqVec model is a representation learning model based on the language model ELMo, taken from natural language processing. ELMo creates embeddings in 0.03 s per protein sequence, on average. The state-of-the-art ESM-1b transformer protein language model is a deep contextual language model trained on 86 billion amino acids across 250 million protein sequences spanning evolutionary diversity. The generated features were used as input for graph neural network-based model training and optimization.

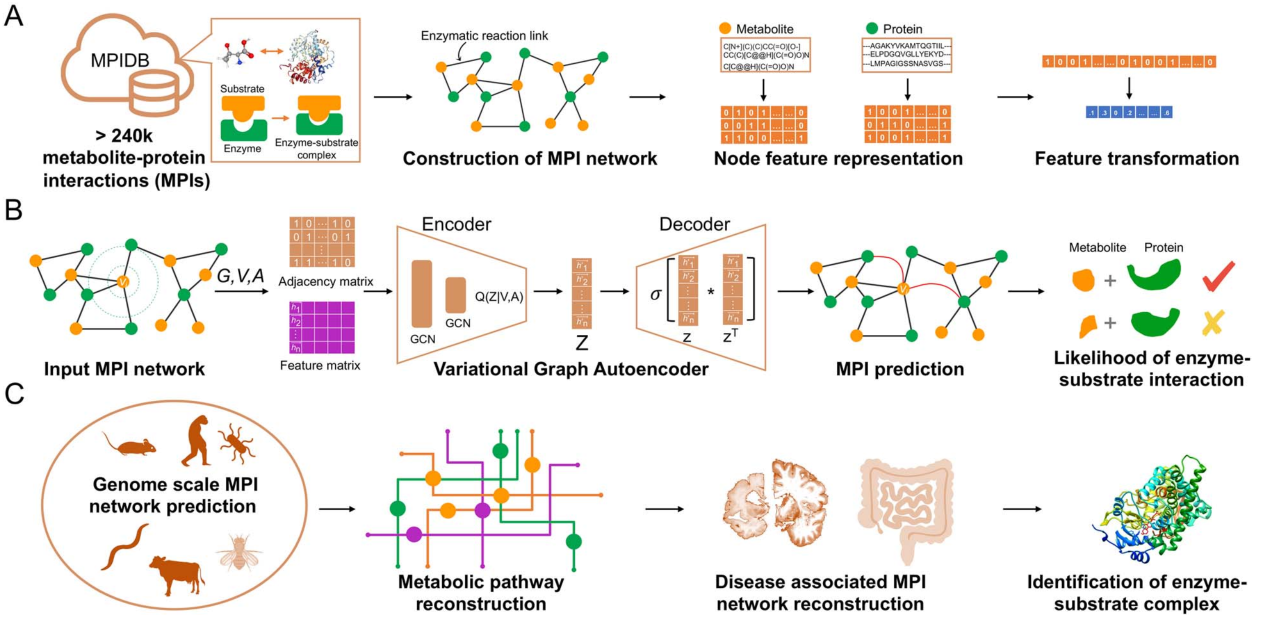

Figure 1. Overview of metabolite–protein interactions prediction by Variational Graph Autoencoder. Panel **A** depicts the construction of the metabolite–protein interaction (MPI) network and featurization of metabolites and proteins. Panel **B** depicts the Variational Graph Autoencoder algorithm to predict MPIs. Adjacency matrix and feature matrix were constructed based on the MPI network and explicit features of metabolites and proteins. The encoder module takes both the adjacency matrix and feature matrix by two graph convolutional layers, followed by the decoder for the reconstruction of the adjacency matrix. The likelihood of MPI is computed based on the latent embedding vectors. Panel **C** depicts the application scenarios of MPI-VGAE framework, including genome-scale MPI prediction, metabolic pathway reconstruction and reconstruction of disease-associated MPI networks.

Mathematical representation of metabolic reaction graph

The metabolite–protein interaction network and metabolic reaction network were represented as undirected graph $G = (V, A, X)$ with $N = |V|$ nodes. For each graph, G was represented by its adjacency matrix $A \in R^{N\times N}$. $A \in \{1, 0\}^{N\times N}$, where 1 denotes there is an edge (enzymatic reaction link) between a pair of nodes (metabolite/protein) and 0 otherwise. Given that the reaction direction was not considered in this study, the graphs were undirected and no weights between edges were specified. The adjacency matrix A was symmetric and unweighted. The feature matrix $X \in R^{N\times D}$ denotes the node features. The node features consist of both explicit and latent features. Explicit features are node attributes, such as the molecular fingerprints of metabolites and vectorized representation of proteins. Latent features are the matrix representations of the graph learned by the graph-embedding methods. These low-dimensional latent representations preserve the properties of the graph such as the local neighborhood of nodes. In the current study, the latent features were used in the graph-embedding models, including Node2vec and Variational Graph Autoencoders.

Variational graph autoencoders model

The variational graph autoencoders model is an unsupervised learning framework for graph-structured data using variational Bayesian methods. Here, we recapitulate the VGAE model and illustrate how it incorporates the metabolite and protein features for link prediction. The VGAE model first maps the nodes onto low-dimensional vector features by an encoder and uses a decoder to reconstruct the original graph topological information. The encoder of VGAE simultaneously incorporates both node structural information and attributes by two graph convolutional layers, as shown below:

$$\mathrm{GCN}\,(\mathrm{X,A}) = \tilde{\mathrm{A}}\mathrm{ReLU}\left(\tilde{\mathrm{A}}\mathrm{X}\mathrm{W}_0\right)\mathrm{W}_1 \tag{1}$$

$$\tilde{\mathrm{A}} = \mathrm{D}^{-\frac{1}{2}}\mathrm{A}\mathrm{D}^{-\frac{1}{2}} \tag{2}$$

where $\tilde{\mathrm{A}}$ is the symmetrically normalized adjacency matrix. The graph convolutional network (GCN) layer performs convolution on graphs to extract local substructure features for individual nodes. Then, it aggregates node-level features into a graph-level feature vector. The latent distribution is then produced by the encoder as follows:

$$\mu = \mathrm{GCN}_\mu\,(\mathrm{X,A}) \tag{3}$$

$$\log\sigma = \mathrm{GCN}_\sigma\,(\mathrm{X, A}) \tag{4}$$

$$\mathrm{Z} = \mu + \sigma * \epsilon \tag{5}$$

where $\epsilon \sim \mathcal{N}(0, \mathrm{I})$ and Z denotes the graph-embedding matrix. Then, the encoder is formulated as

$$\mathrm{q}\,(\,\mathrm{z_i}|\,\mathrm{X,A}) = \mathcal{N}\left(\mathrm{z_i}|\mu_\mathrm{i}, \mathrm{diag}\left({\sigma_\mathrm{i}}^2\right)\right) \tag{6}$$

The decoder decodes the embeddings by reconstructing the graph adjacency matrix:

$$\hat{\mathrm{A}}_{\mathrm{v,u}} = \mathrm{p}\left(\mathrm{A_{ij}} = 1|\mathrm{z_i}, \mathrm{z_j}\right) = \mathrm{S}\left(\mathrm{z_i^T z_j}\right) \tag{7}$$

where $\mathrm{S}(x)$ is the sigmoid function. The loss function is a combination of the reconstruction loss and the KL-divergence:

$$\mathcal{L} = \mathbb{E}_{q(Z|X,A)}\left[logp\,(A|Z) - KL\,[q\,(Z|X, A)||p(Z)\right] \tag{8}$$

The workflow of the VGAE model for metabolite–protein interaction prediction is shown in Figure 1B. First, a graph of metabolite–protein interactions is constructed. The features of metabolites and proteins are also embedded. Second, the VGAE model learns the embedding vectors of metabolites and proteins. Lastly, the interaction of a pair of metabolites and proteins is predicted based on the learned embedding vectors.

Model training and performance evaluation

During the model training, all the existing edges in the network were considered as positive examples while all non-existent edges in the network were considered as negative examples. The model input consists of a list of the edges with node attributes, i.e. features of protein and metabolite, the model output is the probability score of whether the edge exists or not. Due to the sparseness of the MPI and metabolic reaction network, the number of true positive examples was significantly less than the number of true negative examples. This would cause an imbalance problem during the training process. We applied the undersampling approach to balance the dataset by considering an equal number of positive and negative examples during the model training and testing progress. All the positive and negative edges in the original graph of each organism were split into training and testing datasets with a proportion of 80% and 20%, respectively. The training dataset was used for feature selection, model training and optimization. Five-fold cross-validation was used for the optimal feature selection and optimization of parameters. The model generates a prediction score for the edge probability in the node pair. To measure the performance of the model, the area under the curve (AUC), receiver operating characteristics (ROC) and precision–recall (PR) curve were used. The ROC curve was plotted with the true positive rate (TPR) against the false positive rate (FPR), where TPR is on the *y*-axis and FPR is on the *x*-axis. The higher the AUC, the better the performance of the model at predicting positive and negative links. The PR curve was constructed by calculating and plotting the precision and recall at a variety of thresholds. The Average Precision (AP), the weighted mean of precisions at each threshold where the weight is the increase in recall, was used to summarize the PR curve. The evaluation functions are listed as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{TPR} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{11}$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{12}$$

$$AP = \sum_{k=0}^{k=n-1} \left[Recalls(k) - Recalls\left(k+1\right) \right] * Precisions(k) \tag{13}$$

$$Recalls(n) = 0, Precisions(n) = 0, n = Number\ of\ thresholds \tag{14}$$

where TP stands for true positive (i.e. the model predicts a link exists between a pair of nodes and there is a reaction between a protein and a metabolite in the dataset), FP stands for false positive, TN stands for true negative (i.e. the model predicts no link exists between a pair of nodes and there is no reaction between a protein and a metabolite in the dataset) and FN stands for false negative.

Model comparison and evaluation

We compared the VGAE model with several baseline machine learning models, including similarity-based models, random walk-based models and graph-embedding models. The similarity-based measure is a type of unsupervised approach that computes the likelihood of each non-existing link from the similarity score of two nodes, such as Adamic–Adar and Preferential Attachment [35, 36]. Spectral clustering was also included in the baseline models [37]. For graph data, spectral clustering creates node representations by taking top *d* eigenvectors of the normalized Laplacian matrix of the graph. Then, the feature vector representations of nodes were used for computing the likelihood of a pair of nodes. Random walk-based methods learn node representations by generating 'node sequences' through random walks in graphs, as inspired by Natural Language Processing, which tries to learn word representations from sentences. Node2vec is a skip-gram-based approach to learning node embeddings from random walks within a given graph [38]. Graph embedding is a graph representation learning technique that converts graph data into vectors followed by generating a representation of nodes in a lower-dimensional space. Since the embeddings preserve the graph properties such as the local neighborhood of nodes, the likelihood of a link between two targeted nodes is computed based on the embeddings of the nodes. GraphSAGE is an inductive graph neural network model that incorporates node feature information to efficiently generate representations on large graph, which was used to as benchmark model to compare MPI-VGAE [39].

Molecular docking for metabolite–protein interactions

AutoDock Vina was implemented for protein–metabolite docking. We selected the predicted enzyme reactions of protein Cholesterol side-chain cleavage enzyme (CYP11A) and Aldo-keto reductase family 1 member C4 (AKR1C4) binding with 24-hydroxycholesterol and 27-hydroxycholesterol, respectively, as examples. For CYP11A, we used the protein structure from Protein Data Bank (PDB ID 3N9Z) as starting structure. The original binding ligands (22-hydroxycholesterol and adrenodoxin) of 3N9Z were depleted and the interacting sites between ligands and protein were calculated by distance measure (distance <6 Å). We then built the simulation box covering all interaction regions and perform protein–metabolite docking between CYP11A with 24-hydroxycholesterol or 27-hydroxycholesterol, respectively, using AutoDock Vina. The docking processes were run 10 times with random seeds and the conformation with lowest binding affinity was selected as the representing structure.

For protein AKR1C4, we downloaded the predicted structure from Alphafold2 Database (averaged pLDDT >90) and constructed a simulation box covering all proteins to find the most possible binding pose. The docking was also performed between AKR1C4 structure and 24-hydroxycholesterol or 27-hydroxycholesterol, respectively, for 10 times. The structure with lowest binding affinity was selected.

RESULTS

Overview of the MPI-VGAE framework

The complete MPI-VGAE framework is depicted in Figure 1. The genome-scale metabolite–protein interaction networks for different organisms were automatically curated based on thousands of

Table 1. Details of the metabolite–protein interaction network across a variety of organisms

Organisms	Number of nodes	Number of edges	Average degree	Number of proteins	Number of metabolites
Homo sapiens	2306	5786	5.02	855	1451
Mus musculus	1664	3826	4.60	469	1195
Rattus norvegicus	1823	4423	4.85	685	1138
Escherichia coli	1811	4288	4.73	440	1371
Bos taurus	1771	4240	4.79	646	1125
Arabidopsis thaliana	1286	2885	4.49	438	848
Drosophila melanogaster	1086	2568	4.73	374	712
Saccharomyces cerevisiae	1125	2629	4.67	308	817
Caenorhabditis elegans	1011	2442	4.83	322	689
Pseudomonas aeruginosa	1280	2935	4.59	328	952

metabolic pathways from PathBank, followed by manual screening to remove redundant nodes and edges. Explicit node features included molecular fingerprints of metabolites and sequence-based features of proteins, which formed a feature matrix. The Variational Graph Autoencoder model was trained and optimized by using the adjacency matrix and feature matrix from the MPI network. The encoder module consisted of double graph convolutional layers, followed by the decoder for the reconstruction of the adjacency matrix. The likelihood of MPI was computed based on the embedding vectors. The training hyperparameters of the MPI-VGAE model are provided in Table S1.

Characteristics of metabolite–protein interaction networks

The details of metabolites and proteins in the MPI network are shown in Figure 2. For metabolites, fatty acyls (FA), organic acid (OA) and nucleic acid (NA) are the most prevalent metabolite classes, which account for 46% of the total (Figure 2A). Among the seven enzyme classes (i.e. oxidoreductase, ligase, hydrolase, isomerase, lyase, transferase and translocase), transferase and oxidoreductase are the two main enzyme categories in all metabolic pathways (Figure 2B), which suggest that most metabolic reactions involve group transfer reactions and oxidation–reduction reactions. The node degree represents the number of interconnected nodes in the MPI network. Based on the node degree distributions in the integrated MPI network of all organisms in Figure 2C and D, the median links of metabolites and proteins are 3 and 5, respectively. Among all metabolites, 55.4% have degrees more than three, suggesting that more than half of the metabolites participate in at least three enzymatic reactions. Among all proteins, 78.5% have degrees more than three, suggesting that most proteins may catalyze reactions that involve three or more unique metabolite substrates. There are 258 metabolites (11.2%) and 256 proteins (10.5%) that have degrees more than 10. It should be noted that the node degree distributions may vary across different organisms because of the variation of metabolic pathways among different organisms. The details of the metabolite–protein interaction network for each organism are summarized in Table 1. The MPI network was visualized by using the Fruchterman–Reingold layout. Figure 3 shows an example of the human MPI network and the characteristics of metabolites and proteins are provided in Figure S1. The top interconnected metabolites and proteins with the highest degree are annotated. Essential metabolites such as ATP, NAD and L-glutamic acid are the most interconnected metabolites in the MPI network. Cytochromes P450 family enzymes have the most diverse connections with metabolites. The detailed MPI network characteristics of each fold in each organism are provided in Table S2.

Selection of feature representations of metabolites and proteins

A notable advantage of GNNs is capable to combine the node attributes with the graph topological features. The molecular structures of metabolites and proteins are critical to determine the likelihood of the enzymatic reaction link. Encoding the molecular structure as the node attributes in the MPI network would enhance the performance of the VGAE model. Given that many types of numerical representations for structures of metabolites and proteins are available, we examined and selected the optimal feature representations for the MPI prediction of the VGAE model. Molecular fingerprints are a kind of fixed-length of binary vectors to represent the structures of small molecules. Importantly, they are rapid to generate and thus are selected as the representation of metabolites. Sequence-based feature extraction methods were used for proteins. To select the optimal representations of metabolites and proteins, we used different types of feature representations to train our VGAE model and compared the performances of MPI predictions as described previously. Here, we evaluated the models by different combinations of three molecular fingerprints and PCA-transformed molecular fingerprints and two protein embedding models (SeqVec and ESM-1b transformer). In the training datasets of *H. sapiens*, 5-fold cross-validation was performed and the results of each model are summarized in Table 2. For metabolite features, it was found that PCA-transformed molecular fingerprints have better performance than traditional molecular fingerprints. For proteins, the SeqVec model performed slightly better than the ESM-1b transformer. The combination of ECFP (PCA-transformed) of metabolites and SeqVec of proteins achieved the best result with an AUC score of 0.930 and an AP score of 0.938. Therefore, the ECFP (PCA-transformed) and SeqVec were selected to generate features of metabolites and proteins for model evaluation and application. In addition, to select the best PCA-transformed dimensionality, we further tested the MPI prediction performance with different dimensionality parameters, including 32, 64, 128, 256, 512 and 1024. The results are provided in Table S3. It is shown that the feature vector with 128-bit achieved the best MPI prediction performance. Therefore, it was selected as the final dimensionality of prediction in the MPI-VGAE framework.

VGAE model performance and evaluation on organism-wise MPI networks

After selecting the optimal feature representations of metabolites and proteins, we trained the MPI-VGAE framework across all ten organisms and evaluated it on the MPI network in the testing dataset (Figure S2). Figure 4 shows the performance of MPI-VGAE on metabolite–protein interaction network of *H. sapiens*. Figure 4A

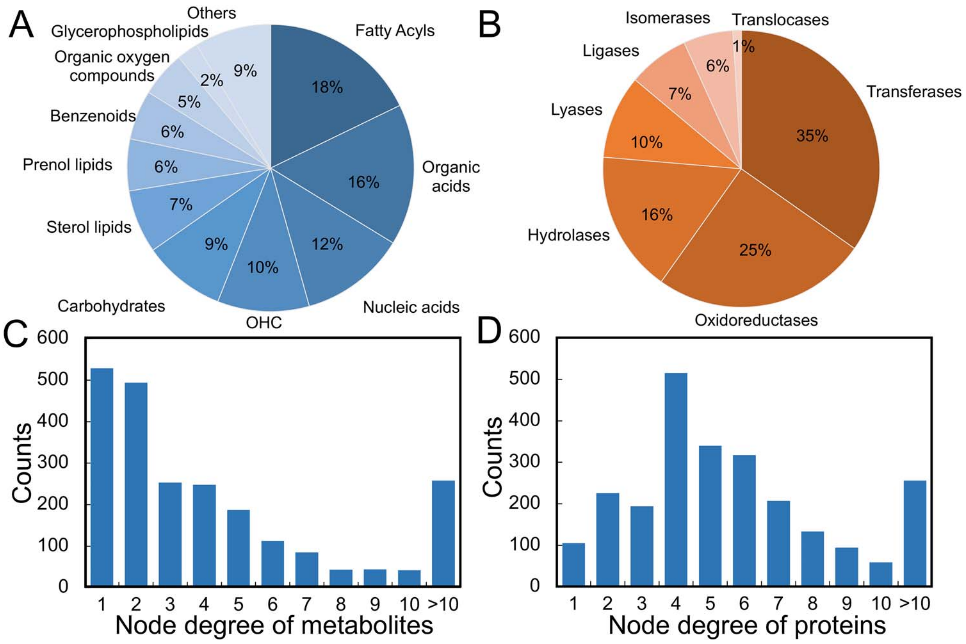

Figure 2. Characteristics of metabolites and proteins in the metabolite–protein interaction network. Panels **A** and **B** show the classes of metabolites and proteins in the metabolite–protein interaction network of all organisms. Panels **C** and **D** show the distributions of node degree in the metabolite–protein interaction network of all organisms.

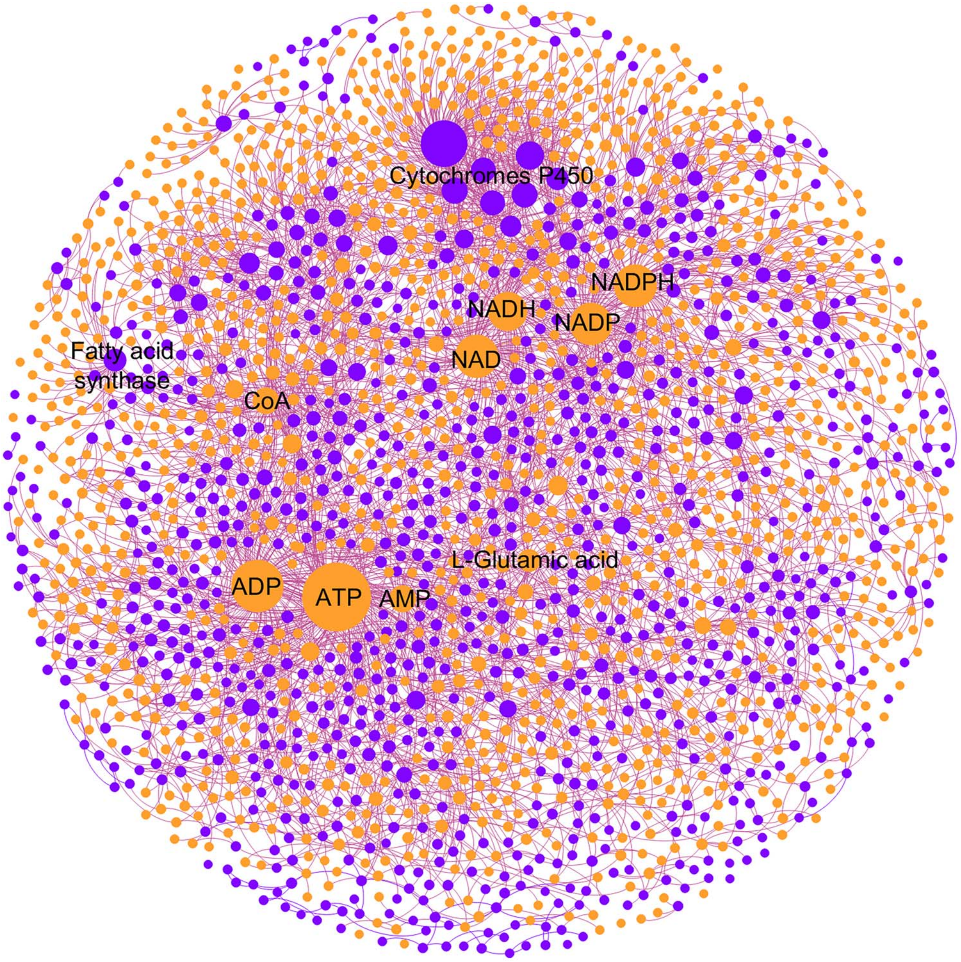

Figure 3. Metabolite–protein interaction network of *Homo sapiens*. The orange circle denotes a metabolite and the purple circle denotes a protein. The size of the circle is proportional to the node degree in the network. The top interconnected nodes are annotated on the graph.

Table 2. Result of metabolite–protein interaction prediction of *Homo sapiens* using different representations of protein and metabolite features

Metabolite feature (molecular fingerprint)	Protein feature	AUC score	AP score
ECFP	SeqVec	0.921 ± 0.002	0.937 ± 0.001
ECFP (PCA)	SeqVec	0.930 ± 0.003	0.938 ± 0.002
MACCS	SeqVec	0.915 ± 0.005	0.928 ± 0.008
MACCS (PCA)	SeqVec	0.921 ± 0.002	0.934 ± 0.001
Topological	SeqVec	0.882 ± 0.013	0.894 ± 0.011
Topological (PCA)	SeqVec	0.920 ± 0.001	0.938 ± 0.004
ECFP	ESM-1b	0.914 ± 0.005	0.931 ± 0.004
ECFP (PCA)	ESM-1b	0.922 ± 0.007	0.937 ± 0.004
MACCS	ESM-1b	0.908 ± 0.001	0.924 ± 0.001
MACCS (PCA)	ESM-1b	0.917 ± 0.001	0.933 ± 0.003
Topological	ESM-1b	0.866 ± 0.007	0.878 ± 0.001
Topological (PCA)	ESM-1b	0.912 ± 0.010	0.933 ± 0.004

and B shows the details of training and testing result MPI-VGAE on the metabolite–protein interaction network of *H. sapiens*. The vectorized input features and graph embeddings of metabolites and proteins were visualized by a non-linear dimensionality technique t-distributed stochastic neighbor embedding (t-SNE). t-SNE maps high-dimensional features to low-dimensional ones by reserving information during dimension reduction. Figure 4C and D shows the t-SNE visualization of feature representations and graph embeddings of metabolites and proteins of *H. sapiens* in the VGAE model. The input features of metabolites and proteins were clearly separated as shown in Figure 4C. Since the VGAE model weighted the neighborhood node attributes in the MPI network, the graph embeddings of adjacent metabolites and proteins in the MPI network were very close (Figure 4D). Figure 4E and F shows the comparison of ROC and PR curves between MPI-VGAE and other machine learning models on metabolite–protein interactions of integrated MPI of *H. sapiens*. The VGAE model embedded with structural information of metabolites and protein sequence properties obtained the highest performance (AUC: 0.915, AP: 0.931), which suggests that the node molecular features greatly enhance the performance in predicting the likelihood of MPI. Table 3 and Table S4, S5, S6 show all the performance of ROC and AP scores by multiple machine learning models on different organisms. The MPI-VGAE with structural information performed the best across all 10 organisms. Compared with other similarity-based and graph-based methods, the VGAE model with node attributes boosted the performance up to 11%, which achieved AUC scores from 0.787 to 0.924, and AP scores from 0.827 to 0.942 for the 10 organisms. To improve the performance of MPI prediction, we have integrated the protein–protein interaction (PPI) and metabolite–metabolite interaction (MMI) information into the MPI network. By adding more connectivity information within the MPI network, the MPI-VGAE is capable to make use of the neighboring information of protein–protein interaction and metabolite–metabolite interaction. For *H. sapiens*, the prediction performance is improved with a ROC score of 0.949 and an AP score of 0.958 on the test dataset consisting of positive and negative MPI edges. Due to the variance of the number of metabolic pathways among different organisms, the node degree distributions may vary across different organisms. Minor fluctuations of performance existed in different organisms due to the variation of the MPI network. For instance, the MPI network of *Arabidopsis thaliana* has the lowest average node degree and the performance of MPI prediction was the worst among all organisms. A possible reason is that the sparseness of the MPI network affects the prediction performance of MPI.

Reconstruction of metabolic pathways and functional MPI networks

Next, we evaluated the ability of the MPI-VGAE models on metabolite–protein interaction prediction in complicated biological systems. The VGAE model was applied to reconstruct the metabolic pathways followed by the reconstruction of functional metabolic networks. We selected the pathways with more than five real metabolite–protein interaction pairs and generated an equal number of negative pairs randomly by considering the metabolites/proteins other than true MPI pairs in the testing dataset. There are 402 metabolic pathways covering most functional classes of metabolic pathways, such as amino acid metabolism, carbohydrate metabolism and energy metabolism. Figure 5A–C shows the distribution of the AUC scores and AP scores of the reconstructed metabolic pathways by the VGAE model. The MPI-VGAE achieved average AUC and AP scores of 0.928 and 0.939, respectively. For instance, there are 201 MPIs in purine metabolism. Our model could successfully predict 100% of the MPIs. For the negative MPIs randomly generated, the VGAE model also accurately predicted 84% of the MPIs as negative cases.

Based on the functional type of each metabolic pathway, five classes of metabolic networks were constructed, including the metabolic pathway network in biological systems, disease-associated metabolic pathway network, drug action metabolism, drug metabolism and protein/metabolite signaling metabolic pathways. For each functional metabolic network, 80% of the positive enzymatic reaction links and an equal number of negative enzymatic reaction links were used for VGAE model training and optimization, and 20% of the randomized selected positive and negative enzymatic reaction links were used for testing. The details of the functional metabolic networks are summarized in Table 4. The AUC score and AP score ranged from 0.883 to 0.909, and 0.899 to 0.924, showing that the performance of the VGAE model is stable across the different functional metabolic networks.

Prediction of metabolite–metabolite interaction by VGAE

The metabolic reaction network is a homogenous network consisting of chemical reactions between metabolites. We evaluated the effectiveness of the VGAE model to predict the metabolite–metabolite interactions in the metabolic reaction network. The metabolic reaction network was constructed based on the KEGG REACTION database, which consisted of

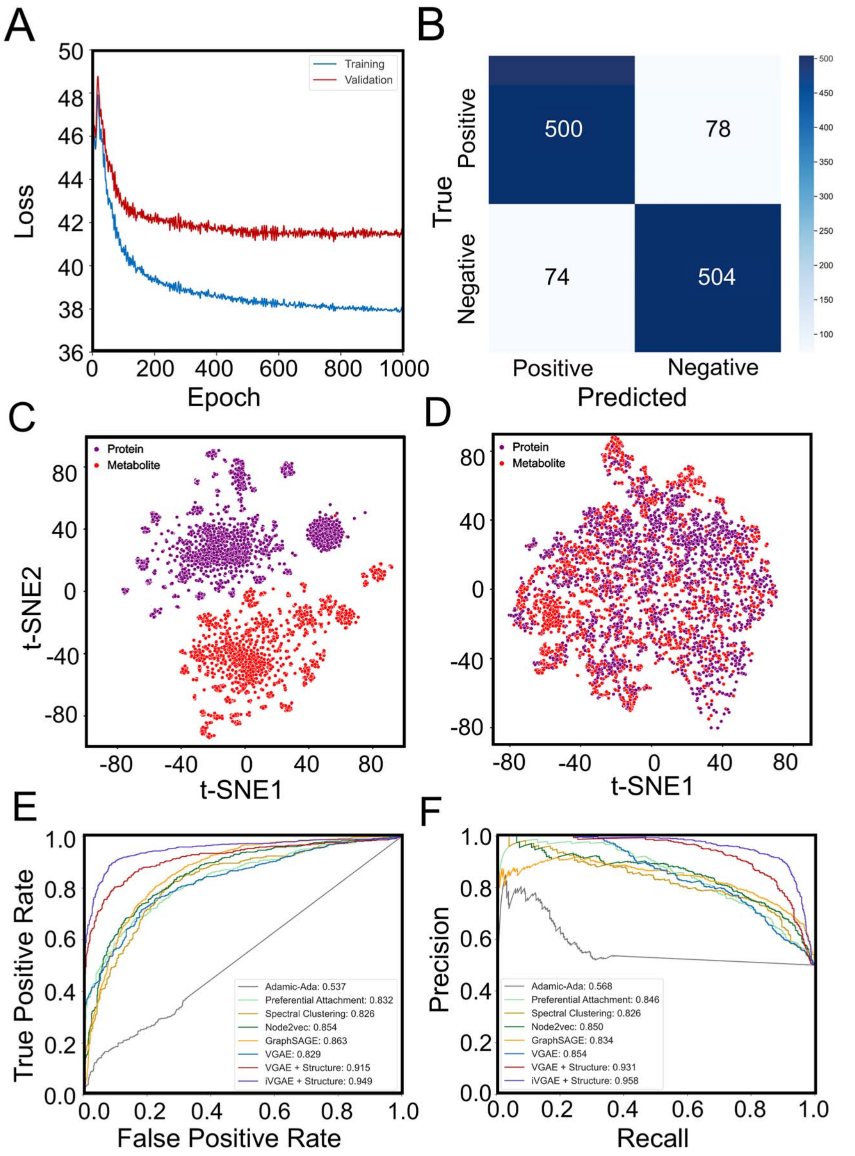

Figure 4. Performance of MPI-VGAE on metabolite–protein interaction network of *Homo sapiens*. Panel **A** shows the training and validation loss versus epoch in the MPI-VGAE. Panel **B** shows the confusion matrix on the metabolite–protein interaction network of *H. sapiens*. Panel **C** shows the t-SNE visualization of explicit feature representation by ECFP molecular fingerprints of metabolites and SeqVec features of proteins. Panel **D** shows the t-SNE visualization of graph embeddings of metabolites and proteins by VGAE. Panels **E** and **F** show the ROC and PR curves of MPI prediction by different machine learning models on the test dataset of the metabolite–protein interaction network of *H. sapiens*. The score of 'iVGAE+structure' denotes the performance of MPI-VGAE when protein–protein interaction and metabolite–metabolite interaction information are added to the MPI network.

Table 3. The AUC scores of metabolite–protein interaction prediction by different machine learning models on a variety of organisms

Organism	**Spectral clustering**	**Adamic–Adar**	**Preferential attachment**	**GraphSAGE**	**Node2vec**	**VGAE (no structure)**	**VGAE (structure)**
Homo sapiens	0.842	0.521	0.829	0.862	0.880	0.845	0.915
Mus musculus	0.766	0.472	0.816	0.867	0.793	0.814	0.859
Rattus norvegicus	0.824	0.573	0.821	0.821	0.831	0.788	0.895
Escherichia coli	0.741	0.484	0.753	0.780	0.728	0.725	0.787
Bos taurus	0.803	0.570	0.794	0.815	0.844	0.779	0.896
Arabidopsis thaliana	0.785	0.351	0.767	0.794	0.821	0.754	0.797
Drosophila melanogaster	0.748	0.335	0.793	0.751	0.796	0.773	0.851
Saccharomyces cerevisiae	0.763	0.326	0.778	0.823	0.825	0.781	0.835
Caenorhabditis elegans	0.746	0.330	0.817	0.844	0.824	0.788	0.850
Pseudomonas aeruginosa	0.705	0.416	0.845	0.853	0.748	0.810	0.816

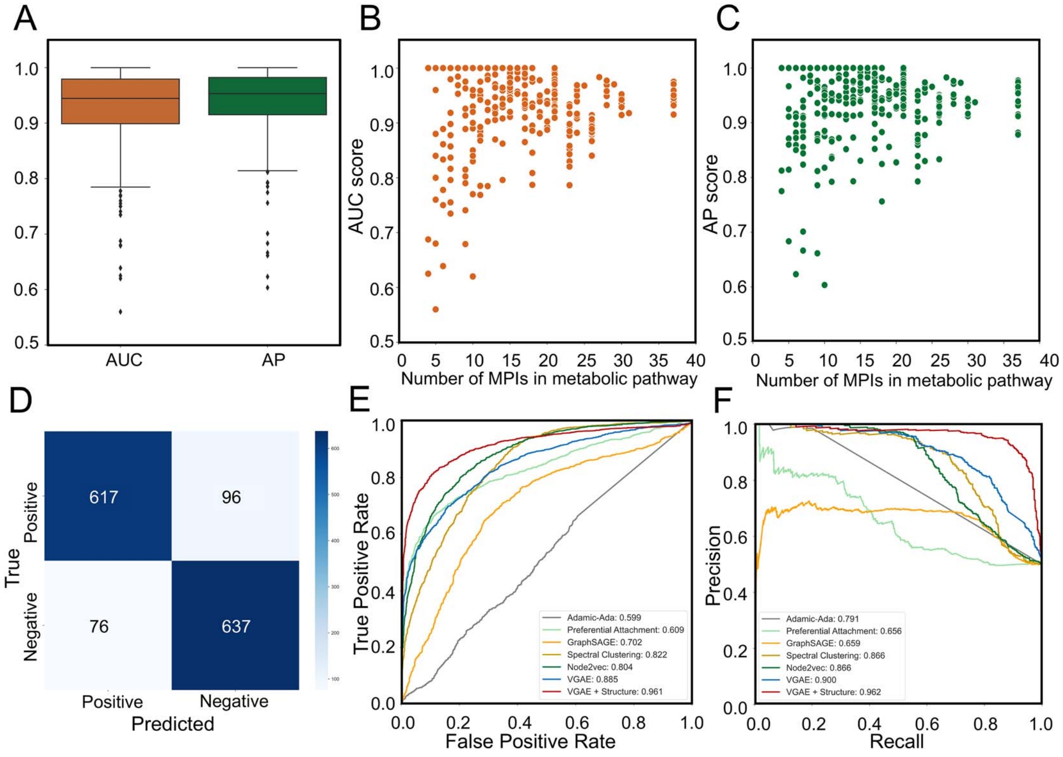

Figure 5. Reconstruction of metabolic pathways by the MPI-VGAE framework and applications of MPI-VGAE to KEGG metabolic reactions. Panel **A** shows the box plot of AUC and AP scores of metabolic pathway reconstruction by the VGAE model. To evaluate the metabolic pathway reconstruction performance of the VGAE model, 402 metabolic pathways from KEGG database were included. Panels **B** and **C** show the scatter plot of AUC and AP scores versus the number of MPIs in the metabolic pathway. Panel **D** shows the confusion matrix on the metabolic reaction network. Panels **E** and **F** show ROC and PR curve of metabolic reaction prediction by different machine learning models.

Table 4. Results of metabolite-protein interaction prediction based on the functional classification of MPI network

MPI network	Number of Nodes	Number of Edges	Average Degree	AUC score	AP score
Disease	1558	4791	6.15	0.890	0.907
Metabolic process	5055	17,506	6.93	0.883	0.901
Drug action	640	1499	6.68	0.909	0.924
Drug metabolism	336	937	5.58	0.883	0.899
Cellular signaling	479	827	3.45	0.909	0.927

2343 nodes with 6889 edges. We compared the performance of metabolic reaction prediction by different machine learning models and the result is shown in Figure 5D–F. By including the molecular structural information of metabolites, the VGAE model achieved an AUC score of 0.964 and an AP score of 0.962, which outperformed other similarity-based or embedding-based models such as ELP. Therefore, by embedding the structural information of nodes, the VGAE model shows excellent prediction performance in both homogeneous and heterogeneous biological networks of different organisms.

Application to the metabolic pathway network reconstruction in Alzheimer's disease and colorectal cancer

Alzheimer's disease is a progressive neurologic disorder that is the most common form of dementia. Many proteomics and metabolomics approaches have been conducted to investigate the disrupted functions of proteins and dysregulated metabolisms. We mapped the disease-associated proteins and metabolites that were induced from DisGeNET database and HMDB databases. For Alzheimer's disease, 65 proteins and 86 metabolites were mapped to the MPI network. MPI-VGAE was applied to predict the likelihood of enzymatic reaction links of all 5590 metabolite–protein interaction pairs. Eleven known existing pairs of metabolite–protein interactions such as adenosine and purine nucleoside phosphorylase exist in the MPI network of Alzheimer's disease. MPI-VGAE is able to pinpoint 10 out of 11 known MPIs in enzymatic reactions accurately. In addition, MPI-VGAE predicts eight additional enzymatic reactions with high confidence score. For instance, Cholesterol side-chain cleavage enzyme (CYP11A) has highly confident interaction with 24-hydroxycholesterol. Since Cytochrome P450s (CYPs) play critical roles in cholesterol homeostasis, many CYPs were disrupted in the cholesterol metabolism and transport in AD that has been investigated by previous studies [40–43]. Due to the high structural similarity of CYPs and cholesterol derivatives, the enzymatic reaction initiated by the metabolite and protein interaction could be altered. The molecular docking simulates the interaction details between Cholesterol side-chain cleavage enzyme (CYP11A) binding with 24-hydroxycholesterol (Figure 6B) and 27-hydroxycholesterol (Figure S3A). We also show the molecular docking results of protein–ligand structures of protein Aldo-keto reductase family 1 member C4 (AKR1C4) binding with 27-hydroxycholesterol (Figure 6C) and 24-hydroxycholesterol (Figure S3B).

Colorectal cancer is the third most prevalent, and second most deadly, cancer worldwide. Integrated analysis of proteomics and metabolomics has been performed to study the proteomic and metabolic alterations in a variety of samples such as tissues and plasma. In total, 81 proteins and 216 metabolites were gathered from DisGeNET and HMDB databases that were identified to be related to colorectal cancer. MPI-VGAE was applied to predict all

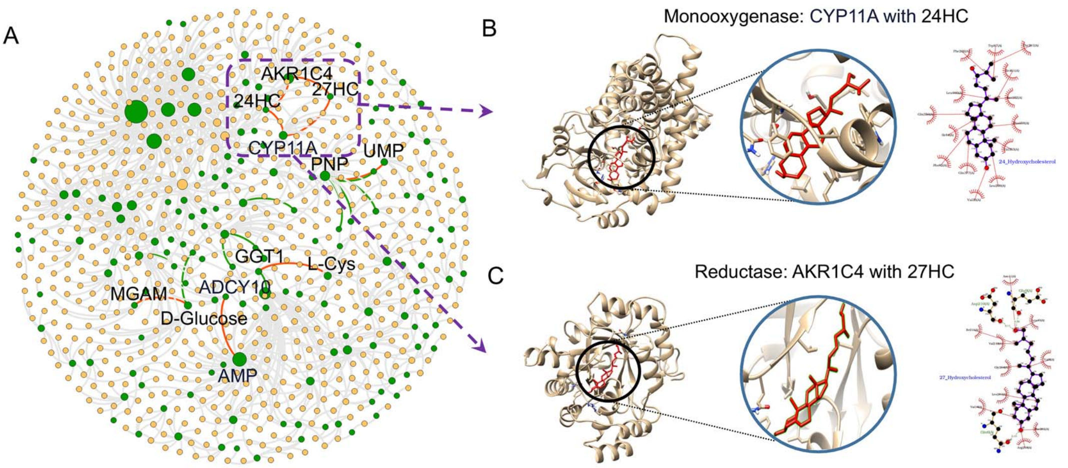

Figure 6. Reconstruction of MPI network of Alzheimer's disease by the MPI-VGAE framework using 65 proteins and 86 metabolites. Panel **A** shows the reconstructed MPI network. The green dots denote the disrupted proteins and metabolites in Alzheimer's disease. The orange dots denote the normal proteins and metabolites that have enzymatic reactions with the disrupted molecules. The circle size is proportional to the node degree in the MPI network. The edges denote the interaction between proteins and metabolite (gray edge: known interaction between normal proteins and metabolites, green edge: known interaction between disrupted proteins and metabolites, red edge: predicted interaction between disrupted proteins and metabolites). Panel **B** shows the molecular docking result between Cholesterol side-chain cleavage enzyme (CYP11A) binding with 24-hydroxycholesterol (24HC) (binding energy:-9.6 kcal/mol). Panel **C** shows the molecular docking result between Aldo-keto reductase family 1 member C4 (AKR1C4) binding with 27-hydroxycholesterol (27HC) (binding energy: −7.6 kcal/mol).

the possible enzymatic reaction links based on all 17 496 pairs. MPI-VGAE predicts 37 out 44 known enzymatic reaction links accurately. Figure S4 shows the reconstructed MPI network of enzymatic reaction link prediction by MPI-VGAE for colorectal cancer. The highly confident MPI pairs predicted by MPI-VGAE for Alzheimer's disease and colorectal cancer are summarized in Tables S7 and S8. In sum, the applications to the reconstruction of enzymatic reaction links for Alzheimer's disease and colorectal cancer demonstrate the efficiency and capability of MPI-VGAE to discover new disease-related enzymatic reactions and metabolic pathways.

DISCUSSION

In this study, we have developed a graph neural network-based method to identify metabolite–protein interactions based on the MPI network. To explore the best feature representations of metabolites and proteins, we compared different combinations of feature extraction approaches based on the MPI prediction performances by the VGAE model. As shown in Table 2, the best performance was obtained by using the combination of protein features from SeqVec and metabolite features from ECFP (PCA-transformed). All the AUC–ROCs were above 0.91 except the combination with topological features of metabolites, which indicated our predictors were rather robust with different features. The implementation of the PCA method to transform the molecular fingerprints further improved the prediction performances, e.g. ESM-1b with topological features (AUC 0.788 versus 0.914). Given that PCA had learned large-scale molecular fingerprints over 78 000 metabolites, the PCA-transformed molecular fingerprints reserved both the features of metabolites and the feature variance between metabolites. In addition, the PCA-transformed molecular fingerprints were no longer binary with 0 or 1 values, which reduced the chance of gradient vanishing during the training process. Interestingly, the combination with protein features by SeqVec shows slightly better performance than features by ESM-1b transformer methods. SeqVec is more computationally efficient and widely used for the fast feature extraction of proteins. It generated the same vector length for proteins with different sequences, which might ignore the size-dependent properties of proteins.

We also tested the MPI-VGAE model to predict metabolite–protein interactions in different organisms. Our approach obtained the best performances across all organisms compared with other methods. Moreover, MPI-VGAE shows stable performance when reconstructing different functional MPI networks. As shown in Table 4, among all five categories of functional MPI networks, the AUC scores are always better than 0.88 and AP scores reach above 0.89. This indicates that the MPI-VGAE method could be able to predict the MPI network involved in the major functional classes.

A notable feature of MPI-VGAE is the capability to reconstruct the MPI network of specific disease based on a list of the disrupted metabolites and proteins. Furthermore, MPI-VGAE predict the highly likely new enzymatic reaction links occurring among the metabolites and proteins, which will facilitate the understanding of disrupted metabolisms in diseases. For Alzheimer's disease and colorectal cancer, MPI-VGAE identified a few potential new enzymatic reaction links such as CYP11A and 24-hydroxycholesterol.

Although our MPI-VGAE could achieve the best performance among all the methods, there are a couple of limitations to our study. Prediction of MPI facilitates the identification of enzyme–substrate interaction that paves the way to discover novel enzymatic reactions. The complete enzymatic reactions from substrates to products catalyzed by enzymes await further investigation by experimental measurements. The imbalanced distribution of positive and negative edges in the network would hamper the accurate prediction of true positive edges (Figure S5, S6). For example, high specificity but low precision would be obtained if the model predicts all MPIs as negative. To reduce the effect of the imbalanced dataset on the VGAE model, we applied the downsampling approach to generate the training and testing dataset and used AUC and AP scores for performance measures. Hopefully, with the rapidly increasing discovery of protein and metabolite interactions, the imbalanced dataset issue in the MPI network will be alleviated.

CONCLUSION

In this work, we present the variational graph autoencoder method to predict metabolite–protein interactions based on the MPI network. When incorporating the node attributes of metabolites and proteins, the performance of MPI-VGAE achieved the highest AUC and AP scores in different genome-scale enzymatic reaction networks. The MPI-VGAE framework also showed stable and excellent performance in reconstructing the metabolic pathways and functional MPI network. To the best of our knowledge, this is the first time that VGAE has been applied in the MPI network for efficient enzymatic reaction prediction. By applying MPI-VGAE to identify the enzymatic reaction link between metabolites and proteins in Alzheimer's disease and colorectal cancer, our method could not only find experiment-proved interactions, but also predict novel and reliable metabolite reactions which could be crucial for mechanistic investigation of disease and drug target discovery. We believe that the method will greatly assist the discovery of novel disease-related enzymatic reactions and pave the way for genome-scale metabolic pathway reconstruction by graph neural network approaches.

Key Points

- A comprehensive metabolite–protein interaction (MPI) database was developed that covers genome-scale heterogeneous networks with thousands of enzymatic reactions across 10 organisms.
- An enzymatic reaction link prediction method called Metabolite–Protein Interaction prediction by Variational Graph Autoencoders (MPI-VGAE) was developed and achieved the best performance compared with existing machine learning methods by encoding molecular features of metabolites and proteins.
- The MPI-VGAE framework has been applied in a variety of scenarios, including the reconstruction of metabolic pathways, functional enzymatic reaction networks, metabolic reaction networks, the reconstruction and discovery of novel enzymatic reaction links in Alzheimer's disease and colorectal cancer.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

National Institutes of Health (R35ES028365 to G.J.P.); Young Scholars Program of Shandong University (21320082164070 to C.W., 21320082064101 to Q.H.); Shandong Natural Science Foundation (ZR2022QB152 to C.W.); Shandong Province Key R&D Program (2021SFGC0504 to Q.H.); National Key Research and Development Program of China (2020YFC2008802 to Q.H.).

DATA AVAILABILITY AND IMPLEMENTATION

The MPI-VGAE framework and datasets are publicly accessible on GitHub https://github.com/mmetalab/mpi-vgae.

REFERENCES

1. Nagel ZD, Klinman JP. A 21st century revisionist's view at a turning point in enzymology. *Nat Chem Biol* 2009;**5**:543–50.
2. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. *Nat Rev Mol Cell Biol* 2019;**20**:353–67.
3. Bludau I, Aebersold R. Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nat Rev Mol Cell Biol* 2020;**21**:327–40.
4. Gao J, Ma S, Major DT, *et al.* Mechanisms and free energies of enzymatic reactions. *Chem Rev* 2006;**106**:3188–209.
5. Reuveni S, Urbakh M, Klafter J. Role of substrate unbinding in Michaelis-Menten enzymatic reactions. *Proc Natl Acad Sci* 2014;**111**:4391–6.
6. Luzarowski M, Vicente R, Kiselev A, *et al.* Global mapping of protein-metabolite interactions in Saccharomyces cerevisiae reveals that Ser-Leu dipeptide regulates phosphoglycerate kinase activity. *Commun Biol* 2021;**4**:181.
7. Milanesi R, Coccetti P, Tripodi F. The regulatory role of key metabolites in the control of cell Signaling. *Biomolecules* 2020;**10**:862.
8. Keshishian H, Burgess MW, Specht H, *et al.* Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry. *Nat Protoc* 2017;**12**:1683–701.
9. Alseekh S, Aharoni A, Brotman Y, *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat Methods* 2021;**18**:747–56.
10. Wang C, He L, Li D-W, *et al.* Accurate identification of unknown and known metabolic mixture components by combining 3D NMR with Fourier transform ion cyclotron resonance tandem mass spectrometry. *J Proteome Res* 2017;**16**:3774–86.
11. Diether M, Nikolaev Y, Allain FH, Sauer U. Systematic mapping of protein-metabolite interactions in central metabolism of Escherichia coli. *Mol Syst Biol* 2019;**15**:e9008.
12. Henry CS, DeJongh M, Best AA, *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 2010;**28**:977–82.
13. Piazza I, Kochanowski K, Cappelletti V, *et al.* A map of protein-metabolite interactions reveals principles of chemical communication. *Cell* 2018;**172**:358–72.e23.
14. Wang Z, Pisano S, Ghini V, *et al.* Detection of metabolite–protein interactions in complex biological samples by high-resolution Relaxometry: toward Interactomics by NMR. *J Am Chem Soc* 2021;**143**:9393–404.
15. Zhao T, Liu J, Zeng X, *et al.* Prediction and collection of protein–metabolite interactions. *Brief Bioinform* 2021;**22**.
16. Faulon J-L, Misra M, Martin S, *et al.* Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor. *Bioinformatics* 2007;**24**:225–33.
17. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 2000;**28**:27–30.
18. Fabregat A, Jupe S, Matthews L, *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2018;**46**:D649–55.
19. Wishart DS, Li C, Marcu A, *et al.* PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Res* 2020;**48**:D470–8.
20. Zhang MH, Chen YX. Link prediction based on graph neural networks. *Adv Neural Inf Proces Syst* 2018;**31**.
21. Kazemi SM, Poole D. SimplE embedding for link prediction in knowledge graphs. *Adv Neural Inf Proces Syst* 2018;**31**.

22. You J, Ying R, Leskovec J. Position-aware graph neural networks. *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California, PMLR, 2019;**97**:7134–43.
23. Long Y, Wu M, Liu Y, *et al.* Pre-training graph neural networks for link prediction in biomedical networks. *Bioinformatics* 2022;**38**: 2254–62.
24. Lu Y, Guo Y, Korhonen A. Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics* 2017;**18**:39.
25. Nasiri E, Berahmand K, Rostami M, Dabiri M. A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. *Comput Biol Med* 2021;**137**:104772.
26. Masuda A, Kawachi T, Ohno K. Rapidly growing protein-centric technologies to extensively identify protein-RNA interactions: application to the analysis of co-transcriptional RNA processing. *Int J Mol Sci* 2021;**22**:5312.
27. Bakker OB, Aguirre-Gamboa R, Sanna S, *et al.* Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses. *Nat Immunol* 2018;**19**:776–86.
28. Wang C, Kurgan L. Survey of similarity-based prediction of drug-protein interactions. *Curr Med Chem* 2020;**27**:5856–86.
29. Ryu JY, Kim HU, Lee SY. Deep learning improves prediction of drug-drug and drug-food interactions. *Proc Natl Acad Sci* 2018;**115**:E4304–11.
30. Durant JL, Leland BA, Henry DR, Nourse JG. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* 2002;**42**: 1273–80.
31. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J Chem Inf Comput Sci* 1985;**25**:64–73.
32. Landrum G. RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 2013;8.
33. Heinzinger M, Elnaggar A, Wang Y, *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 2019;**20**:723.
34. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**:e2016239118.
35. Adamic LA, Adar E. Friends and neighbors on the web. *Soc Networks* 2003;**25**:211–30.
36. Zeng S. Link prediction based on local information considering preferential attachment. *Physica A: Statistical Mechanics and its Applications* 2016;**443**:537–42.
37. Symeonidis P, Mantas N. Spectral clustering for link prediction in social networks with positive and negative links. *Soc Netw Anal Min* 2013;**3**:1433–47.
38. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 2016, 855–64.
39. Hamilton WL, Ying R, Leskovec J. Inductive representation learning on large graphs. *Adv Neural Inf Proces Syst* 2017;**30**.
40. Baloni P, Funk CC, Yan J, *et al.* Metabolic network analysis reveals altered bile acid synthesis and metabolism in Alzheimer's disease. *Cell Reports Medicine* 2020;**1**:100138.
41. Di Paolo G, Kim T-W. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nat Rev Neurosci* 2011;**12**: 284–96.
42. Martins IJ, Berger T, Sharman MJ, *et al.* Cholesterol metabolism and transport in the pathogenesis of Alzheimer's disease. *J Neurochem* 2009;**111**:1275–308.
43. Fonseca ACRG, Resende R, Oliveira CR, Pereira CMF. Cholesterol and statins in Alzheimer's disease: current controversies. *Exp Neurol* 2010;**223**:282–93.