Redefining the Game: MVAE-DFDPnet's Low-Dimensional Embeddings for Superior Drug-Protein Interaction Predictions

Liang-Yong Xia, Yu Wu, Longfei Zhao, Leying Chen, Shiyi Zhang, Mengdi Wang and Jie Luo

***Abstract*— Precisely predicting drug-protein interactions (DPIs) is pivotal for drug discovery and advancing precision medicine. A significant challenge in this domain is the high-dimensional and heterogeneous data characterizing drug and protein attributes, along with their intricate interactions. In our study, we introduce a novel deep learning architecture: the Multi-view Variational Auto-Encoder embedded within a cascade Deep Forest (MVAE-DFDPnet). This framework adeptly learns ultra-low-dimensional embedding for drugs and proteins. Notably, our t-SNE analysis reveals that two-dimensional embedding can clearly define clusters corresponding to diverse drug classes and protein families. These ultra-low-dimensional embedding likely contribute to the enhanced robustness and generalizability of our MVAE-DFDPnet. Impressively, our model surpasses current leading methods on benchmark datasets, functioning in significantly reduced dimensional spaces. The model's resilience is further evidenced by its sustained accuracy in predicting interactions involving novel drugs, proteins, and drug classes. Additionally, we have corroborated several newly identified DPIs with experimental evidence from the scientific literature. The code used to generate and analyze these results can be accessed from https://github.com/Macau-LYXia/MVAE-DFDPnet-V2.**

***Index Terms*— DPI; heterogeneous networks; multi-view; deep learning; ensemble learning; cascade deep forest**

I. Introduction

Over the past few decades, a myriad of computational methodologies for the identification of DPIs have been devised, substantially narrowing the search scope for potential drug and protein candidates. This advancement has markedly diminished the costs and boosted the efficiency of drug discovery and development processes. Typically, these methodologies fall into three distinct classes: ligand-based [1], molecular docking [2], and machine learning-based methods [3]–[7]. While the ligand-based approach relies on a sufficient number of known ligands for a given protein; the Molecular docking approach is limited to available 3D protein structures [8]. Conversely, machine learning-based methods have emerged as a highly promising avenue for predicting DPIs [9]–[12].

Over the years, the input features for machine learning-based approaches have evolved from pharmacogenetics, which comprises information on drugs (i.e., chemical structures) and proteins (i.e., encoding sequences) represented as feature vectors [13], [14], to similarity network-based input features [15], and more recently, the trend has shifted toward incorporating heterogeneous networks that integrate various biological and pharmacological data [5], [16], [17]. Such heterogeneous data sources provide a rich and inherently related information, offering a multi-view perspective for the prediction of novel DPIs. On one hand, incorporating these diverse data sources can enrich the feature set and potentially boost prediction accuracy. On the other hand, it also increases input dimensionality, posing challenges for subsequent analyses.

The application of deep learning techniques further enhances the power of drug-protein pair prediction models. DeepWalk [18] built a tripartite, heterogeneous network from biomedical linked datasets and utilized the network's node similarity for prediction. NeoDTI [17] used the neighborhood information of the network and learned topology preserving representations of drugs and proteins. deepDTnet [19] adopted a deep auto-encoder to learn high-quality features from heterogeneous networks and then applied positive-unlabeled matrix completion to predict new DPIs. AOPEDF [20] employed arbitrary-order proximity to derive a low-dimensional representation of drugs and proteins, and subsequently developed a cascade deep forest classifier to predict new interactions.

While these methods have each made notable strides in the field, there remains substantial room for improvement. A particularly challenging task lies in the construction of compact, low-dimensional embedding of drugs and proteins that are consistent across different types of drug/protein-related networks. The use of deep learning as a classifier is also challenging due to the need to fine-tune numerous hyperparameters. As such, the construction of low-dimensional drug and protein embedding, and the design of an effective deep learning-based pipeline are urgently required for efficient

This work is supported by Postdoctoral Research Foundation of China [2020M671125] and Shanghai Super Postdoctoral Incentive Schemes [2019257]. (Corresponding authors: Jie Luo; Shiyi Zhang.)

Liang-Yong Xia, Longfei Zhao, and Jie Luo are with National Engineering Research Center of Advanced Magnetic Resonance Technologies for Diagnosis and Therapy, School of Biomedical Engineering, Shanghai Jiao Tong University, 800 Dongchuan RD. Minhang District, Shanghai, China, 200240 (e-mail: xia2yin1234@gmail.com;longfei@sjtu.edu.cn;jieluo@sjtu.edu.cn).

Yu Wu and Mengdi Wang are with Center for Statistics and Machine Learning Department of Electrical Engineering Department of Computer Science, Princeton University, 41 Olden Street, New Jersey, USA, 08544. (e-mail: yuw@princeton.edu;mengdiw@princeton.edu).

Leying Chen and Shiyi Zhang are with School of Biomedical Engineering, Shanghai Jiao Tong University, 800 Dongchuan RD. Minhang District, Shanghai, China, 200240 (e-mail: chenleying@sjtu.edu.cn;zhangshiyi@sjtu.edu.cn).

and accurate DPIs prediction.

To address these issues, we develop MVAE-DFDPnet, a network-based framework for DPIs prediction that fuses a multi-view variational auto-encoder (MVAE) with a cascade deep forest (CDF). The strengths of MVAE-DFDPnet are twofold: Firstly, it effectively consolidates individual networks into a unified low-dimensional embedding representation; secondly, it employs a deep cascade forest classifier, as described by Zhou et al. [21], which delivers high-performance classification with significantly fewer hyperparameters than traditional deep neural networks. This deep forest allows for automatic complexity adjustment. The synergy between advanced data compression and a flexible classification approach culminates in a robust, sophisticated, and appreciably enhanced model for drug discovery.

We evaluate MVAE-DFDPnet on benchmark datasets and compare its performance against several state-of-the-art methods. Our analysis reveals that MVAE-DFDPnet outperforms its counterparts in predictive accuracy, achieving this with a reduced number of drug and protein embedding. We also evaluate the robustness of MVAE-DFDPnet using previously unseen drug-protein pairs, demonstrating its high performance under stringent conditions. Finally, we provide visualizations of the learned drug and protein embedding and validate novel interactions predicted by our model.

II. Materials and Methods

A. Dataset

We integrated heterogeneous bio-networks as multi-view data inputs, encompassing four distinct entity types (Drugs, Proteins, Diseases, and Sideeffect) and 15 different types of associations; further details are provided in Table I. Our task was to infer unknown DPIs within a network comprising 732 Food and Drug Administration (FDA)-approved drugs and 1,915 unique proteins. Within this network, 4,978 known DPIs are labeled as positive samples, with a corresponding number of randomly-selected non-interacting (or 'unknown') pairs labeled as negative samples. For more details on dataset collection, readers are referred to the recent works by Luo et al. [5] and Zeng et al. [19].

TABLE I: Summary of heterogeneous biological networks as multi-view data.

Category	View	Association	Drug	Potein	Disease	Sideeffect	No.of.Egde	Database
Drug	1	Drug-Drug	732	-	-	-	132768	[22]
	2	Drug-Disease	732	-	440	-	1208	[22], [23]
	3	Drug Sideeffect	732	-	-	12904	263805	[24]
	4	Drug chemical similarity	732	-	-	-	118578	[25]
	5	Drug therapeutic similarity	732	-	-	-	240825	[26]
	6	Drug sequence similarity	732	-	-	-	215902	[22]
	7	Biological Processes similarity	732	-	-	-	243812	[27]
	8	Cellular Component similarity	732	-	-	-	271192	[27]
	9	Molecular Function similarity	732	-	-	-	244226	[27]
Protein	1	Protein-Protein	-	1915	-	-	112468	[28]
	2	Protein-Disease	-	1915	440	-	23080	[29]
	3	Protein sequence similarity	-	1915	-	-	1705532	[30]
	4	Biological Processes similarity	-	1915	-	-	1611052	[27]
	5	Cellular Component similarity	-	1915	-	-	1665635	[27]
	6	Molecular Function similarity	-	1915	-	-	1653812	[27]

B. Methods

In this paper, we propose a novel network-based method termed MVAE-DFDPnet, which achieves substantial feature compression along with a flexible prediction mechanism. MVAE-DFDPnet takes multi-view data as input. The biological interaction network of each view is preprocessed into a probabilistic co-occurrence (PCO) matrix [19], then calculate a shifted positive pointwise mutual information (PPMI) matrix by following Bullinaria and Levy [31]. The next phase involves inputting a drug/protein's high-dimensional feature vector into the MVAE to produce a reduced-dimensionality embedding. Lastly, a CDF is then employed to predict DPIs using the concatenated embedding of drug-protein pairs. The output is a binary indicator that denotes the presence or absence of an interaction between a specific drug-protein pair. We sequentially present the stages of our methodology in Fig. 1, progressing from left to right.

*1) **Data preprocessing**:* Our model utilizes heterogeneous biological networks as inputs, which are subjected to a standardized preprocessing routine independently to ensure consistency.

The random surfing model, which we employ to construct the PCO matrix, is an adaptation of the PageRank algorithm, originally developed by Google's founders to rank webpages in search engine results [32], [33]. We have adapted this model to elucidate the interconnections among various biological entities. In a biological context, the 'links' might signify the interactions or associations between these entities. To construct the PCO matrix, we start by assuming a random surfer navigating the biological network. The 'surfer' randomly moves from one entity to another, following the edges or 'links'. The probability that the surfer ends up on a particular node or entity is calculated. This process is repeated for all entities in the network, resulting in the PCO matrix. Subsequently, we convert the PCO matrix into a PPMI matrix, a technique widely employed in natural language processing and other domains to delineate semantic correlations among elements such as words, or in our context, biological entities like genes, proteins, or diseases, within a network. The PPMI matrix, informed by the PCO matrix, emerges as a potent instrument for elucidating complex, non-linear interdependencies and sparse associations among biological entities. Lastly, we apply matrix decomposition to the PPMI matrix, deriving novel, lower-dimensional representations of the network. This step may reveal concealed structures or patterns, thus offering fresh perspectives on the biological system under investigation. The details of data preprocessing steps are described as below:

(i) PCO matrix: Let $G = (V, E)$ denote the network which contains vertices V and edges E. Suppose the vertex set V is sorted and has n_1 biological entities from category 1 and n_2 biological entities from category 2. It means the size of V is $|V| = n_1 + n_2$, and the first n_1 (last n_2) vertices belong categories 1(2). The edge set E contains undirected binary interaction information, i.e. $E = \{(v_i, v_j)|v_i, v_j \in V, \exists$ a known interaction between v_i and $v_j\}$. The info of E can be also represented as an adjacency matrix $\mathbf{A} \in \{0, 1\}^{|V|\times|V|}$, where the entry $\mathbf{A}_{i,j} = 1$ if $(v_i, v_j) \in E$ and $\mathbf{A}_{i,j} = 0$ otherwise. We define the operation $g(\cdot)$ that transforms a network's adjacency matrix into a PCO matrix $g(\mathbf{A}) \in \mathbb{R}^{|V|\times|V|}$:

$$g(\mathbf{A}) = \sum_{t=1}^{T} \mathbf{P}_t,$$

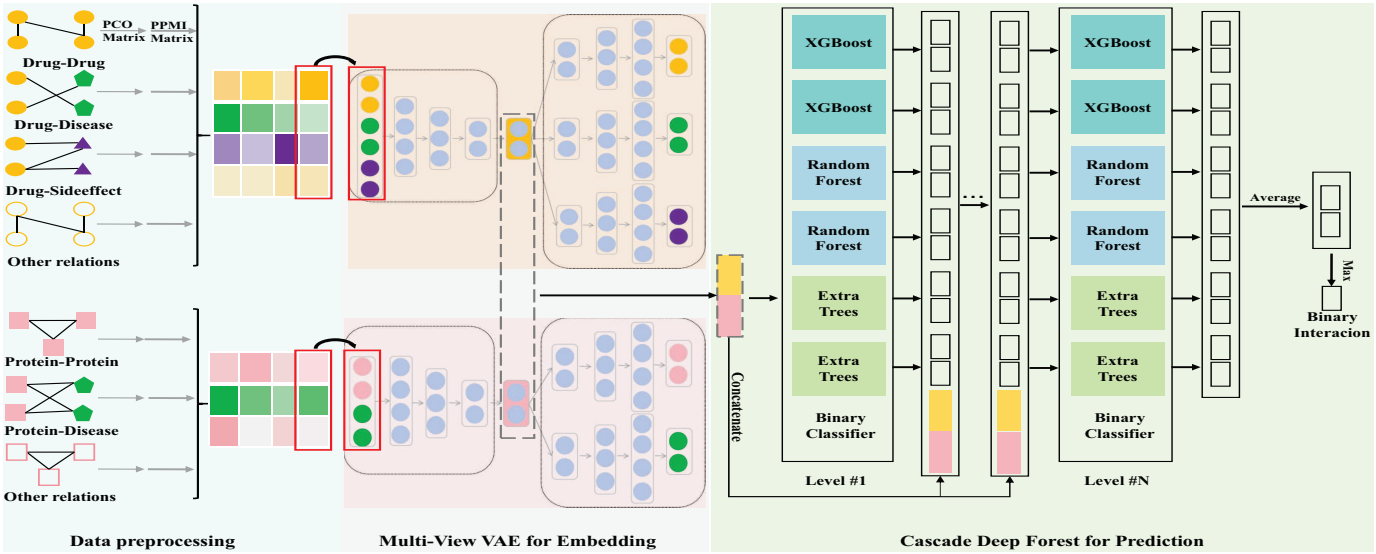


Fig. 1: **The schematic flowchart of the MVAE-DFDPnet pipeline.** This flowchart elucidates the three key stages: **I. Data preprocessing** involves the integration of diverse sources of drug and protein-related information to construct unique, heterogeneous interaction networks for each view; the creation of a PCO matrix using a random surfing model, capturing the intricate network topological information of the drugs and proteins; Subsequent conversion of PCO into PPMI matrix. **II.MVAE** embeds each column of drug or protein features into a low-dimensional vector that retains the ability to reconstruct the original topological properties present within every interaction network. **III. CDF** infers a DPI as a binary value based on the concatenated learned embedding of the drug-protein pair.

where

$$\mathbf{P}_t = \alpha \mathbf{P}_{t-1}\mathbf{A} + (1-\alpha)\mathbf{P}_0 \; and \; \mathbf{P}_0 = \mathbf{I}_{|V|}.$$

The i-th row of $\mathbf{P}_0$ is one-hot vector meaning that we start a random walk from the vertex v_i. In each iteration process, the random surfing process will continue with probability of α, and there is a 1-α probability to return to the original vertex and restart this process. So the i-th row of $\mathbf{P}_t$ is the probabilistic distribution of appearance at some vertex after t transitions starting from v_i. Therefore, $g(\mathbf{A})$ collects the co-occurrence information between vertices during the number of steps T in the random walk.

(ii) PPMI matrix: We are using the exact same formula in [19]; here we write this formula in our notations. Given a network represented by an adjacency matrix $\mathbf{A}$ for the graph $\mathbf{G}$, where $\boldsymbol{A_{i,j}}$ indicates the strength of the relationship between nodes v_i and v_j, a PCO matrix $g(\mathbf{A})$ is first computed, which often normalizes or thresholds the adjacency matrix to emphasize significant connections.

We define another entry-wise operation $h(\cdot)$ that transforms a PCO matrix $g(\mathbf{A})$ into a PPMI matrix $h(g(\mathbf{A})) \in \mathbb{R}^{|V|\times|V|}$. For each pair of nodes (v_i, v_j) in the network, the PPMI value $h(g(\mathbf{A}))_{i,j}$ is calculated as the maximum of the pointwise mutual information and zero, which can be mathematically expressed as:

$$h(g(\mathbf{A}))_{i,j} = \max\left(\log\left(\frac{g(\mathbf{A})_{i,j}\cdot\sum_{k=1}^{|V|}\sum_{l=1}^{|V|} g(\mathbf{A})_{k,l}}{\sum_{v=1}^{|V|}\left(\sum_{k=1}^{|V|} g(\mathbf{A})_{k,v}\right)\cdot\left(\sum_{l=1}^{|V|} g(\mathbf{A})_{v,l}\right)}\right), 0\right)$$

Here, the numerator represents the joint probability of observing both v_i and v_j together (normalized by their individual probabilities in the network), while the denominator corresponds to the product of their marginal probabilities. By applying the log-likelihood ratio and taking the maximum with zero, PPMI effectively removes negative associations and emphasizes stronger, more statistically significant relationships between the nodes.

(iii) Data integration: Now we have a PPMI matrix $h(g(\mathbf{A}))$ retaining the network topology. The i-th row of matrix $h(g(\mathbf{A}))$ contains the information of how the entity v_i interacts with all other entities in this network. Since we focus on only one category or part of biological entities to embed, we select the rows associated with the entities we are interested in. Let $\mathbf{X}$ denote the sub-matrix, which consists of selected rows of the PPMI matrix $h(g(\mathbf{A}))$.

We integrate the network information for drug/protein across multiple views by stacking matrices. Suppose we focus on K^d-view data for n^d drugs, there are K^d networks $\{G_k^d = (V_k^d, E_k^d)\}_{k\in[K^d]}$ generating PPMI submatrices with selected rows of n^d drug only $\{\mathbf{X}_k^d\}_{k\in[K^d]}$. Here $\mathbf{X}_k^d \in \mathbb{R}^{n^d\times|V_k^d|}$, and the superscription d means drug. So the integrated multi-view drug-related data are represented as $\boldsymbol{X}^d \in \mathbb{R}^{n^d\times\sum_{k=1}^{K^d}|V_k^d|}$:

$$\boldsymbol{X}^d = [\mathbf{X}_1^d, \mathbf{X}_2^d, ..., \mathbf{X}_{K^d}^d], \tag{1}$$

We consider the i^{th} column of $\boldsymbol{X}^{d^\top}$ as the feature vector of drug i to be embedded by multi-view VAE, like the column framed by a red line in the Fig. 1.

Similarly, we have the superscription p for protein. We preprocess and then integrate K^p-view data of n^p proteins into $\boldsymbol{X}^p \in \mathbb{R}^{n^p\times\sum_{k=1}^{K^p}|V_k^p|}$:

$$\boldsymbol{X}^p = [\mathbf{X}_1^p, \mathbf{X}_2^p, ..., \mathbf{X}_{K^p}^p], \tag{2}$$

We have the j^{th} column of $\boldsymbol{X}^{p\top}$ as the feature vector of protein j to be fed into the next module.

*2) **MVAE for Embedding**:* Without loss of generality, we give as an example MVAE for drug embeddings, exactly the same model for protein embeddings. MVAE takes one preprocessed high-dimensional drug feature as input, e.g. drug i's feature $[\boldsymbol{X}^d]_i \in \mathbb{R}^{\sum_{k=1}^{K^d}|V_k^d|}$ which consists of K^d vectors corresponding to each drug-related view. Firstly, the input feature $[\boldsymbol{X}^d]_i$ will be embedded into a low-dimensional latent feature $Z_i^d \in \mathbb{R}^{r^d}$ by a fully-connected neural network. Then the latent feature Z_i^d will be reconstructed back to K^d vectors of each view respectively by one fully-connected neural network, $\{[\mathbf{X}_k^d]_i\}_{k\in[K^d]}$. An ideal embedding should be able to recover raw input vectors of every view. So we formulate the loss function as:

$$\mathcal{L} = \sum_{i=1}^{n^d}\sum_{k=1}^{K^d} f([\boldsymbol{X}_k^d]_i, [\widehat{\boldsymbol{X}_k^d}]_i) + KL(q(Z_i^d|[\boldsymbol{X}_k^d]_i)||p(Z_i^d)) \tag{3}$$

where we choose binary cross-entropy as the reconstruction loss f, and we use the Kullback-Leibler divergence term to regularize the difference between the learnt latent distribution and the prior distribution.

Finally, we learn a collection of drug embedding $\{Z_i^d\}_{i\in n^d}$, one for each drug. Similarly we have $\{Z_j^p \in \mathbb{R}^{r^p}\}_{j\in n^p}$ as protein embedding.

*3) **CDF for Prediction**:* The CDF uses ensemble and multiple-layer strategies to harness the strengths of various forest classifiers, which contribute to a powerful and robust predictive model. The diversity of the classifiers and the multi-grained scanning strategy allow the model to capture complex patterns in the data, making it suitable for a wide range of prediction tasks, including those with high-dimensional features and complex interactions.

Therefore, we use a CDF to predict DPIs. Specifically, we feed a concatenated pair embedding of drug i and protein j, i.e. $[Z_i^d, Z_j^p] \in \mathbb{R}^{n^d+n^p}$, into L ensemble layers then output the final binary results or the score between zero and one.

We include different types of binary forest classifiers to encourage diversity beyond ensembling; each ensemble layer consists of two XGBoost [34], two Random Forests [35] and two Extra Trees [36].

Each binary forest classifier outputs two non-negative values summing up as 1. Additionally, the cascade deep forest boosts the prediction performance by emphasizing the initial input, i.e. the drug-protein pair embedding. It means every layer after the first one takes all outputs of classifiers in the previous layer, along with the drug-protein pair embedding, as input. The cascade deep forest also boosts by deepening. The number of layers N is determined adaptively; during training, we stop adding layers when there is no noticeable decrease in the loss value. Finally, we average the outputs of the last layer to get two non-negative values summing up as 1. These two values mean the score of 'interacting' and 'not interacting' respectively.

Therefore, we take the 'interacting or not' result associated with the maximum score as the final binary prediction between drug i and protein j.

*4) **Time complexity of proposed MVAE-DFDPnet**:* With reduced embedding dimensions, we may achieve smaller number of neurons in each layer in the time complexity of MVAE structure; while getting smaller maximum number of splits considered per feature, and smaller average depth of the tree in the CDF structure, thus improve overall model time efficiency. Detailed analysis in Supplementary materials Section A.

III. Results

A. Low-dimensional embedding reveals latent drug/protein families

The reconstruction loss in each view (as shown in Table II) suggests that MVAE preserves information effectively. Visualization of the low-dimensional embedding with t-distributed stochastic neighbor embedding algorithm (t-SNE) [37], [38] reveals that MVAE captures drugs/proteins information and successfully separate them into distinct clusters. This is exemplified by the 14 types of drugs by Anatomical Therapeutic Chemical (ATC)-based classification in Fig. 2a, and four classical protein families (i.e. G-protein-coupled receptors (GPCRs), kinases, nuclear receptors (NRs), ion channels (ICs), and others.) [39] in Fig. 2b. Taken together, t-SNE analysis demonstrates that the MVAE learned embedding not only greatly reduce dimensionality, but may also capture underlying biological associations within drugs and proteins.

TABLE II: MVAE reconstruction loss (binary Cross-entropy) across different data views.

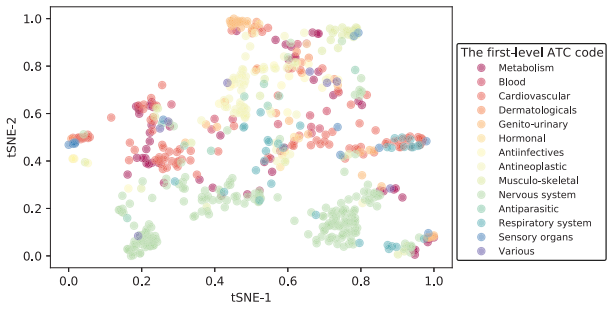
Drug	View	1	2	3	4	5	6	7	8	9
	Value	0.0811	0.0691	0.0853	0.0317	0.0790	0.0968	0.0978	0.0903	0.1242
Protein	View	1	2	3	4	5	6			
	Value	0.0616	0.0417	0.0192	0.0561	0.0606	0.0586			

B. MVAE-DFDPnet outperforms baseline methods

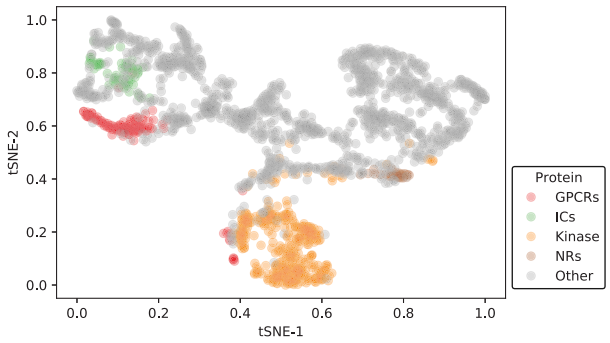
We compare MVAE-DFDPnet with the following baseline methods:

1. KBMF2K uses kernelized Bayesian matrix factorization with twin kernels for prediction [40].
2. DTINet learns low-dimensional vector representations from heterogeneous data and then applies inductive matrix completion for prediction [5].
3. NeoDTI utilizes the neighborhood information of the network for prediction [17].
4. deepDTnet obtains low-dimensional vector representations auto-encoder algorithm by and utilizes positive-unlabeled matrix completion algorithm for prediction [19]
5. AOPEDF uses an arbitrary-order proximity and a cascade deep forest classifier to infer new interactions [20] .

We test different hyper parameters (Supplementary Table S1 and S2) of MVAE and evaluate the MVAE-DFDPnet model performance using the area under ROC curve (AUROC) and the area under the precision-recall curve (AUPR). Since negative samples are many more compared to positive samples in our data, we randomly sampled the negative samples to



(a) Drug



(b) Protein

Fig. 2: Visualization of the learned drug and protein embedding via t-SNE [37], [38]. The visualizations were created based on embeddings learned through MVAE-DFDPnet. **(a)** Drugs color-coded according to the first level of the ATC-based classification (http://www.whocc.no/atc/). **(b)** Proteins color-coded by their corresponding drug target families.

reach ratio of 1:1 in each test. All methods are performed 10 times of random 5-fold cross-validation and computed the average performance. Fig. 3 shows the average AUROC and AUPR for each method. The data indicate that MVAE-DFDPnet surpasses leading-edge methods such as KBMF2K, DTINet, NeoDTI, deepDTnet, and AOPEDF. Notably, MVAE-DFDPnet achieves impressive results with drug-protein embeddings of merely four dimensions (AUROC = 0.973 and AUPR = 0.974, as shown in Table III using a pairwise train-test split), outstripping most prior methods. As the dimensionality of the embeddings increases to 200 and 2,000, MVAE-DFDPnets performance further enhances, yielding an accuracy on par with the top-performing method, AOPEDF (AUROC = 0.975 and AUPR = 0.974), while utilizing a significantly reduced dimensional space for embeddings (200 compared to AOPEDFs 1,650). The result of compared AOPEDF with different dimensions is in Supplementary Table S3.

C. *Robustness and generalizability of MVAE-DFDPnet*

Random splitting of drug-protein pairs for testing may result in the inclusion of drugs or proteins in the test set that have been seen during training, potentially leading to an overestimation of the model's performance, and overly optimistic conclusions. To address this issue, we adjust sample

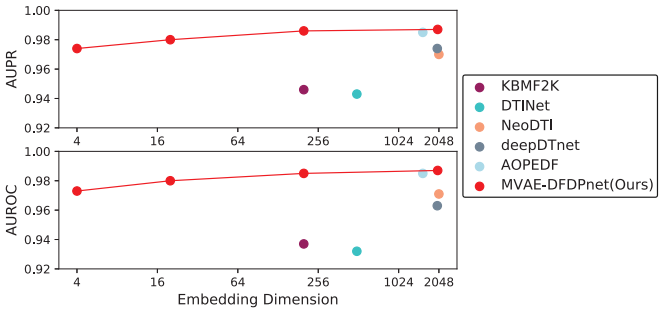


Fig. 3: Performance comparison of MVAE-DFDPnet against State-of-the-art methods at various embedding dimensions.

TABLE III: Evaluation performance of MVAE-DFDPnet compared with state-of-art methods. The associated variance is in parentheses.

Method	Embedding Dimension	AUROC	AUPR
KBMF2K	200	0.937	0.946
DTINet	500	0.932	0.943
NeoDTI	2048	0.971	0.97
deepDTnet	2000	0.963	0.974
AOPEDF	1650	0.975	0.974
MVAE-DFDPnet	4	0.975(0.008)	0.976(0.008)
	20	0.982(0.006)	0.984(0.008)
	200	0.985(0.006)	0.986(0.008)
	2000	**0.986(0.006)**	**0.986(0.006)**

splitting to test the MVAE-DFDPnet method on entirely novel drugs, novel proteins, or both novel scenarios (Fig. 4a). The AUROC and AUPR remain at 0.9 even when both the protein and the drug in the test set are entirely new (Fig. 4b). This demonstrates the robustness of MVAE-DFDPnet in predicting DPIs for new drugs or proteins, even without prior knowledge of the drug or protein.

Further, drugs belonging to the same class might be structurally and functionally similar, therefore excluding an entire drug class in training data would pose an even more challenging task for DPI prediction. We test MVAE-DFDPnet in predicting DPIs of each ATC drug class while excluding the given class of drugs from training data (Fig. 4c), resulting in most ATC class AUROC and AUPR scores of >0.9 indicating good generalizability across most drug class. Note that the absence of antineoplastic drugs in training data has a significant impact on DPI prediction. The number of drugs in each ATC drug class (Fig. 4d) is not associated with their impact on the DPI prediction.

D. *MVAE-DFDPnet reveals novel DPIs*

We validate novel DPIs with DGIdb database [41]; DGIdb supports 1,705 known + 1,379 novel DPIs. Among top 60 DGIdb-validated predictions in Supplementary Table S4, we find that many of them can also be supported by the previously known experimental or clinical evidence in the literature [42]–[51].

We visualize the top 100 novel DPIs between 30 drugs and 40 proteins. As shown in Fig. 5, kinases and their drug interactions tend to form a cluster that is separated from those of GPCR and others. In the kinase network, the interactions

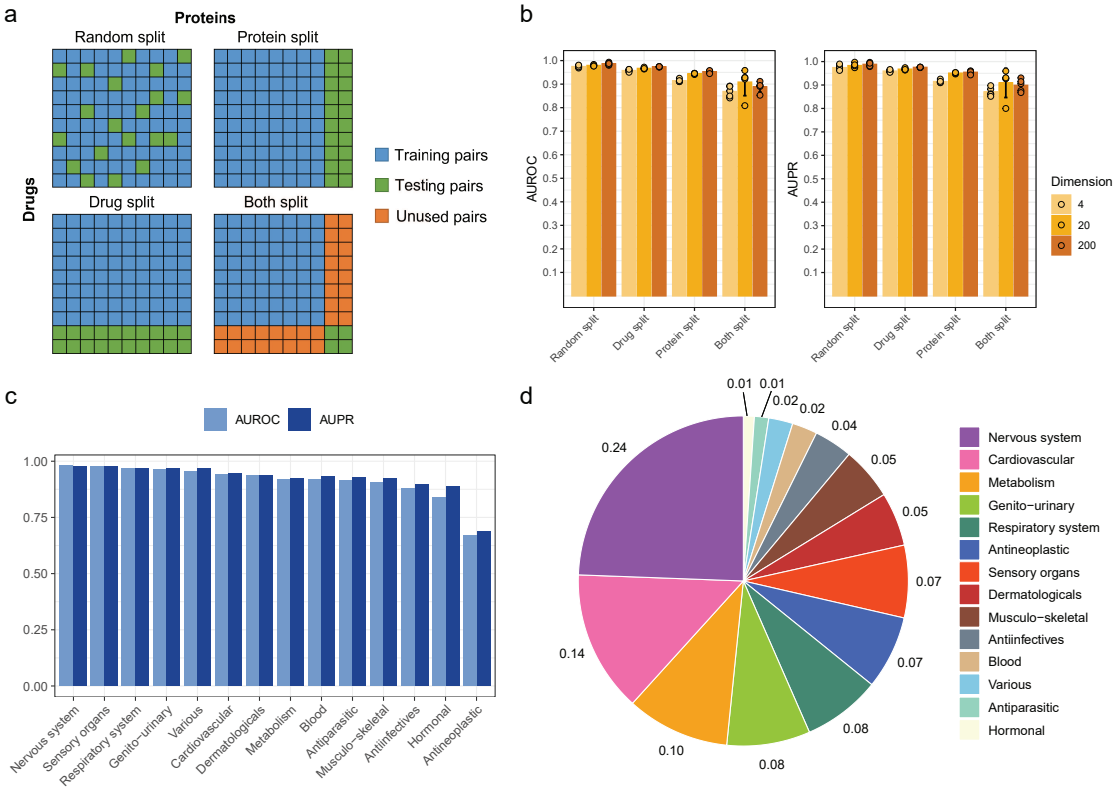


Fig. 4: The performance evaluation results of MVAE-DFDPnet on several challenging scenarios. (a) The schematic diagram of splitting drug-protein pairs into training and testing sets. (b) The performance of MVAE-DFDPnet to accurately predict the drug-protein interaction of completely novel drugs, proteins, or both novel. (c) The performance of the MVAE-DFDPnet in predicting novel drugs of ATC-based classifications. (d) The proportion of drugs in each ATC classification in this study.

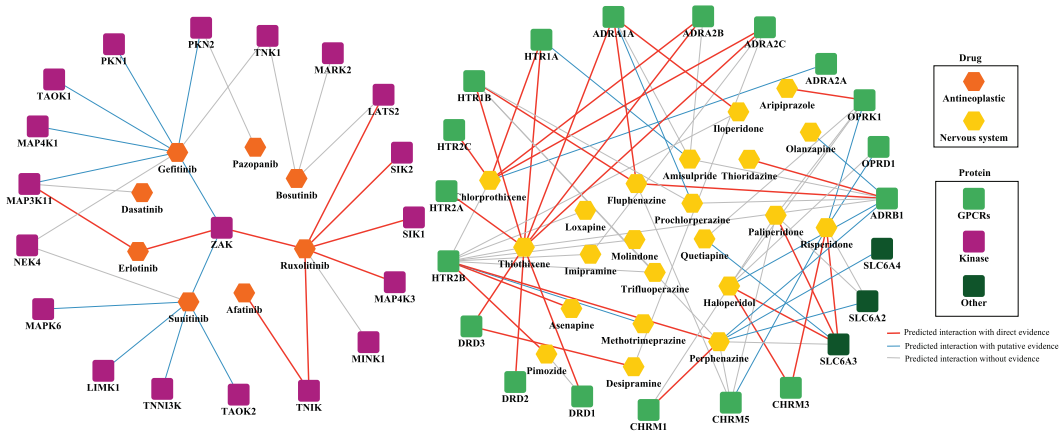


Fig. 5: Top 100 novel DPIs predicted by MVAE-DFDPnet. This figure showcases the top 100 novel DPIs as predicted by our MVAE-DFDPnet model. The drugs are grouped by the first ATC-level code, in this case, Antineoplastic and Nervous system. Many novel DPI predictions are corroborated by experimental data, database entries, or clinical findings reported in the literature. DPIs with direct evidence available are marked by red lines, those with putative evidence available are marked by blue lines; while DPIs lacking such evidence are represented by grey lines.

are between multiple tyrosine kinase inhibitors and kinases, while the majority of the interactions are supported by previously reported experimental evidence in the literature. For example, Sunitinib is a multi-target tyrosine kinase inhibitor, indicated for the treatment of renal cell carcinoma and imatinib-resistant gastrointestinal stromal tumor. LIMK1 is an important member of LIMK/cofilin signaling and participates in actin reorganization, cell migration, and tube formation. One research found that the indolin-2-one derivatives potently inhibit the LIMK1/cofilin signaling pathways, while sunitinib is a representative drug that emerged from indolin-2-one [52]. The prediction between Sunitinib and LIMK1 merit

further studies, which may point to therapy that target cancer invasive behavior. Similarly, EGFR-mutant non-small cell lung cancer is known to respond to EGFR inhibitors such as Gefitinib. A study identified several genetic determinants of EGFR TKI sensitivity through genome-wide CRISPR-Cas9 screening. Researchers found that sgRNA targeting PKN2 strongly sensitized HCC827 cancer cells to gefitinib treatment. Another synergic gene RIC8A was found attenuating YAP signaling, which might be modulated via both LATS1/2 [53]. This demonstrates the potential insights that can be gained from our novel DPI predictions.

E. *Case study: antiepileptic drugs*

We attempt to focus on the drugs in our dataset that fall under the ATC classification system N03A (Nervous system/antiepileptics). Antiepileptic drugs (AEDs) are typically prescribed for chronic, long-term use in patients with epilepsy, often extending over several years [54]. The prediction scores for these drugs are generally lower than those of the top 100 interactions previously discussed. Our MVAE-DFDPnet model was able to computationally identify 440 potential interactions linking 117 proteins with 12 AEDs. We further illustrated the top 100 interactions encompassing 10 AEDs and 36 proteins in Fig. 6. The potential for these predicted interactions to be supported by existing research seems to correlate positively with their prediction scores. It is noteworthy that ion channel modulators such as phenytoin, clonazepam, and vigabatrin exhibited unique interaction profiles, indicating diverse mechanisms of action. Furthermore, our model predicted a wide range of interactions between most of the AEDs and an array of GPCRs. This prediction could shed light on the ongoing discussion around the therapeutic potential of GPCRs for treating acquired epilepsy [55]. Thus, these findings could guide the development of new AEDs or therapeutic strategies and be utilized to repurpose existing drugs for acquired epilepsy treatment, potentially accelerating the drug development process.

IV. Conclusion and Future Work

In this study, we present a novel deep learning architecture, MVAE-DFDPnet, tailored for DPIs prediction. This framework integrates a multi-view variational autoencoder with a cascading deep forest classifier, offering a refined and potent method for inferring novel DPIs.Our experimental results demonstrate the superiority of MVAE-DFDPnet in DPI prediction, outperforming existing state-of-the-art techniques while utilizing a significantly reduced dimensionality of drug-protein embeddings, with good generability and robustness. Future endeavors may include the integration of additional biological data sources to augment the model's input dataset.Further optimization and enhancement of the MVAE-DFDPnet components could yield even more accurate and efficient DPI predictions. Lastly, empirical validation of the DPI predictions in laboratory settings would reinforce the practical and scientific merits of our approach.

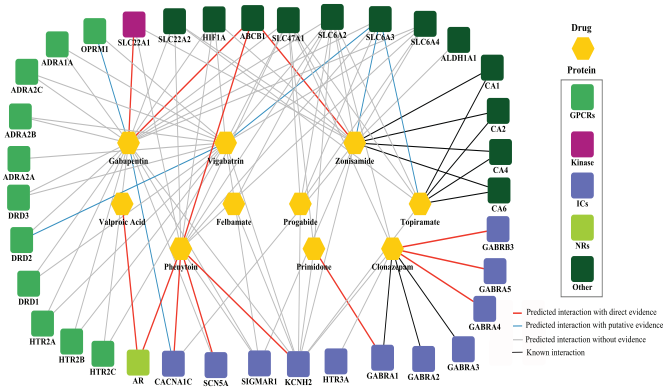


Fig. 6: Antiepileptic drug (AED)-related DPI network featuring the top 100 novel DPIs predicted by MVAE-DFDPnet. Drugs classified under the N03A antiepileptics category are represented by yellow hexagons. Proteins are depicted as squares, each color-coded according to their respective protein families. DPIs are differentiated by colored lines, each of which represents the type or availability of supporting evidence.

References

[1] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nat. Biotechnol.*, vol. 25, no. 2, pp. 197–206, 2007.

[2] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "Autodock4 and autodocktools4: Automated docking with selective receptor flexibility," *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, 2009.

[3] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinform.*, vol. 25, no. 18, pp. 2397–2403, 2009.

[4] L.-Y. Xia, Z.-Y. Yang, H. Zhang, and Y. Liang, "Improved prediction of drug–target interactions using self-paced learning with collaborative matrix factorization," *J. Chem. Inf. Model.*, vol. 59, no. 7, pp. 3340–3351, 2019.

[5] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nat. Commun.*, vol. 8, no. 1, pp. 1–13, 2017.

[6] I. Lee and H. Nam, "Sequence-based prediction of protein binding regions and drug–target interactions," *J. Cheminform.*, vol. 14, no. 1, pp. 1–15, 2022.

[7] Y.-C. Li, Z.-H. You, C.-Q. Yu, L. Wang, L. Wong, L. Hu, P.-W. Hu, and Y.-A. Huang, "Ppaedti: personalized propagation auto-encoder model for predicting drug-target interactions," *IEEE J. Biomed. Health. Inform.*, vol. 27, no. 1, pp. 573–582, 2022.

[8] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey," *Brief. Bioinform.*, vol. 20, no. 4, pp. 1337–1357, 2019.

[9] X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin, and Y. Zhang, "Drug–target interaction prediction: databases, web servers and computational models," *Brief. Bioinform.*, vol. 17, no. 4, pp. 696–712, 2016.

[10] Q. Ye, C.-Y. Hsieh, Z. Yang, Y. Kang, J. Chen, D. Cao, S. He, and T. Hou, "A unified drug–target interaction prediction framework based on knowledge graph and recommendation system," *Nat. Commun.*, vol. 12, no. 1, pp. 1–12, 2021.

[11] Y. Wu, M. Gao, M. Zeng, J. Zhang, and M. Li, "Bridgedpi: a novel graph neural network for predicting drug–protein interactions," *Bioinform.*, vol. 38, no. 9, pp. 2571–2578, 2022.

[12] H. Atas Guvenilir and T. Doğan, "How to approach machine learning-based prediction of drug/compound–target interactions," *J. Cheminform.*, vol. 15, no. 1, pp. 1–36, 2023.

[13] Y. Xue, Z.-R. Li, C. W. Yap, L. Z. Sun, X. Chen, and Y. Z. Chen, "Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 5, pp. 1630–1638, 2004.

[14] D.-S. Cao, S. Liu, Q.-S. Xu, H.-M. Lu, J.-H. Huang, Q.-N. Hu, and Y.-Z. Liang, "Large-scale prediction of drug–target interactions using protein sequences and drug topological structures," *Anal. Chim. Acta*, vol. 752, pp. 1–10, 2012.

[15] Y. Tabei, E. Pauwels, V. Stoven, K. Takemoto, and Y. Yamanishi, "Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers," *Bioinform.*, vol. 28, no. 18, pp. i487–i494, 2012.

[16] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinform.*, vol. 30, no. 20, pp. 2923–2930, 2014.

[17] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "Neodti: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," *Bioinform.*, vol. 35, no. 1, pp. 104–111, 2019.

[18] N. Zong, H. Kim, V. Ngo, and O. Harismendy, "Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations," *Bioinform.*, vol. 33, no. 15, pp. 2337–2344, 2017.

[19] X. Zeng, S. Zhu, W. Lu, Z. Liu, J. Huang, Y. Zhou, J. Fang, Y. Huang, H. Guo, L. Li *et al.*, "Target identification among known drugs by deep learning from heterogeneous networks," *Chem. Sci.*, vol. 11, no. 7, pp. 1775–1797, 2020.

[20] X. Zeng, S. Zhu, Y. Hou, P. Zhang, L. Li, J. Li, L. F. Huang, S. J. Lewis, R. Nussinov, and F. Cheng, "Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest," *Bioinform.*, vol. 36, no. 9, pp. 2805–2812, 2020.

[21] Z.-H. Zhou and J. Feng, "Deep forest: towards an alternative to deep neural networks," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 3553–3559.

[22] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda *et al.*, "Drugbank 5.0: a major update to the drugbank database for 2018," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[23] S. Avram, C. G. Bologa, J. Holmes, G. Bocci, T. B. Wilson, D.-T. Nguyen, R. Curpan, L. Halip, A. Bora, J. J. Yang *et al.*, "Drugcentral 2021 supports drug discovery and repositioning," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1160–D1169, 2021.

[24] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The sider database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2016.

[25] M. Quirós, S. Gražulis, S. Girdzijauskaitė, A. Merkys, and A. Vaitkus, "Using smiles strings for the description of chemical connectivity in the crystallography open database," *J. Cheminform.*, vol. 10, no. 1, pp. 1–17, 2018.

[26] I. Mandric, J. Rotman, H. T. Yang, N. Strauli, D. J. Montoya, W. Van Der Wey, J. R. Ronas, B. Statz, D. Yao, V. Petrova *et al.*, "Profiling immunoglobulin repertoires across multiple human tissues using rna sequencing," *Nat. Commun.*, vol. 11, no. 1, pp. 1–14, 2020.

[27] "The gene ontology resource: enriching a gold mine," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, 2021.

[28] D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork *et al.*, "The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D605–D612, 2021.

[29] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 52–55, 2002.

[30] "Uniprot: the universal protein knowledgebase in 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D480–D489, 2021.

[31] J. A. Bullinaria and J. P. Levy, "Extracting semantic representations from word co-occurrence statistics: A computational study," *Behav Res Methods*, vol. 39, pp. 510–526, 2007.

[32] T. H. Haveliwala, "Topic-sensitive pagerank," in *Proc. 11th Int. Conf. WWW*, 2002, pp. 517–526.

[33] M. A. Farooqi, M. A. Ashraf, and M. U. Shaukat, "Google page rank site structure strategies for marketing web pages," *J. Comput. Biomed. Inform.*, vol. 2, no. 02, pp. 140–157, 2021.

[34] A. Ogunleye and Q.-G. Wang, "Xgboost model for chronic kidney disease diagnosis," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 17, no. 6, pp. 2131–2140, 2019.

[35] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2012.

[36] C. Désir, C. Petitjean, L. Heutte, M. Salaun, and L. Thiberville, "Classification of endomicroscopic images of the lung based on random subwindows and extra-trees," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 9, pp. 2677–2683, 2012.

[37] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.

[38] D. Kobak and P. Berens, "The art of using t-sne for single-cell transcriptomics," *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[39] Q. Cheng, F. A. Lopez, C. Duran, C. Camarillo, T. I. Oprea, and S. Schurer, "The ontology reference model for visual selectivity analysis in drug-target interactions," in *IEEE. Int. Conf. Bioinformatics. Biomed.* IEEE, 2017, pp. 2091–2097.

[40] M. Gönen, "Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization," *Bioinform.*, vol. 28, no. 18, pp. 2304–2310, 2012.

[41] S. L. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. J. Song, M. Griffith, O. L. Griffith, and A. H. Wagner, "Integration of the drug–gene interaction database (dgidb 4.0) with open crowdsource efforts," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1144–D1151, 2021.

[42] D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka *et al.*, "Chembl: towards direct deposition of bioassay data," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D930–D940, 2019.

[43] M. Whirl-Carrillo, E. M. McDonagh, J. Hebert, L. Gong, K. Sangkuhl, C. Thorn, R. B. Altman, and T. E. Klein, "Pharmacogenomics knowledge for personalized medicine," *Clin. Pharmacol. Ther.*, vol. 92, no. 4, pp. 414–417, 2012.

[44] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan *et al.*, "Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1031–D1041, 2020.

[45] J. F. Armstrong, E. Faccenda, S. D. Harding, A. J. Pawson, C. Southan, J. L. Sharman, B. Campo, D. R. Cavanagh, S. P. Alexander, A. P. Davenport *et al.*, "The iuphar/bps guide to pharmacology in 2020: extending immunopharmacology content and introducing the iuphar/mmv guide to malaria pharmacology," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1006–D1021, 2020.

[46] M. Rask-Andersen, S. Masuram, and H. B. Schiöth, "The druggable genome: evaluation of drug targets in clinical trials suggests major shifts in molecular class and indication," *Annu. Rev. Pharmacol. Toxicol.*, vol. 54, pp. 9–26, 2014.

[47] Z. Tanoli, Z. Alam, M. Vähä-Koskela, B. Ravikumar, A. Malyutina, A. Jaiswal, J. Tang, K. Wennerberg, and T. Aittokallio, "Drug target commons 2.0: a community platform for systematic analysis of drug–target interaction profiles," *Database*, vol. 2018, 2018.

[48] M. Rask-Andersen, M. S. Almén, and H. B. Schiöth, "Trends in the exploitation of novel drug targets," *Nat. Rev. Drug Discov.*, vol. 10, no. 8, pp. 579–590, 2011.

[49] S. E. Patterson, R. Liu, C. M. Statz, D. Durkin, A. Lakshminarayana, and S. M. Mockus, "The clinical trial landscape in oncology and connectivity of somatic mutational profiles to targeted therapies," *Hum. Genomics.*, vol. 10, no. 1, pp. 1–13, 2016.

[50] M. Griffith, N. C. Spies, K. Krysiak, J. F. McMichael, A. C. Coffman, A. M. Danos, B. J. Ainscough, C. A. Ramirez, D. T. Rieke, L. Kujan *et al.*, "Civic is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer," *Nat. Genet.*, vol. 49, no. 2, pp. 170–174, 2017.

[51] Y. Gloor, C. Czarnetzki, F. Curtin, B. Gil-Wey, M. R. Tramèr, and J. A. Desmeules, "Genetic susceptibility toward nausea and vomiting in surgical patients," *Front. genet.*, vol. 12, p. 816908, 2022.

[52] J. Guo, M. Zhu, T. Wu, C. Hao, K. Wang, Z. Yan, W. Huang, J. Wang, D. Zhao, and M. Cheng, "Discovery of indolin-2-one derivatives as potent pak4 inhibitors: structure-activity relationship analysis, biological evaluation and molecular docking study," *Bioorg. Med. Chem.*, vol. 25, no. 13, pp. 3500–3511, 2017.

[53] H. Zeng, J. Castillo-Cabrera, M. Manser, B. Lu, Z. Yang, V. Strande, D. Begue, R. Zamponi, S. Qiu, F. Sigoillot *et al.*, "Genome-wide crispr screening reveals genetic modifiers of mutant egfr dependence in human nsclc," *Elife*, vol. 8, p. e50223, 2019.

[54] J. Anderson and C.-C. Moor, "Anti-epileptic drugs: a guide for the non-neurologist," *Clin. Med.*, vol. 10, no. 1, p. 54, 2010.

[55] Y. Yu, D. T. Nguyen, and J. Jiang, "G protein-coupled receptors in acquired epilepsy: Druggability and translatability," *Prog. Neurobiol.*, vol. 183, p. 101682, 2019.