Journal of Translational Medicine

RESEARCH

Open Access

HEDDI-Net: heterogeneous network embedding for drug-disease association prediction and drug repurposing, with application to Alzheimer's disease

Yin-Yuan Su[1], Hsuan-Cheng Huang[1], Yu-Ting Lin[1], Yi-Fang Chuang[2,3,4], Sheh-Yi Sheu[1,5] and Chen-Ching Lin[1*]

Abstract

Background The traditional process of developing new drugs is time-consuming and often unsuccessful, making drug repurposing an appealing alternative due to its speed and safety. Graph neural networks (GCNs) have emerged as a leading approach for predicting drug-disease associations by integrating drug and disease-related networks with advanced deep learning algorithms. However, GCNs generally infer association probabilities only for existing drugs and diseases, requiring network re-establishment and retraining for novel entities. Additionally, these methods often struggle with sparse networks and fail to elucidate the biological mechanisms underlying newly predicted drugs.

Methods To address the limitations of traditional methods, we developed HEDDI-Net, a heterogeneous embedding architecture designed to accurately detect drug-disease associations while preserving the interpretability of biological mechanisms. HEDDI-Net integrates graph and shallow learning techniques to extract representative diseases and proteins, respectively. These representative diseases and proteins are used to embed the input features, which are then utilized in a multilayer perceptron for predicting drug-disease associations.

Results In experiments, HEDDI-Net achieves areas under the receiver operating characteristic curve of over 0.98, outperforming state-of-the-art methods. Rigorous recovery analyses reveal a median recovery rate of 73% for the top 100 diseases, demonstrating its efficacy in identifying novel target diseases for existing drugs, known as drug repurposing. A case study on Alzheimer's disease highlighted the model's practical applicability and interpretability, identifying potential drug candidates like Baclofen, Fluoxetine, Pentoxifylline and Phenytoin. Notably, over 40% of the predicted candidates in the clusters of commonly prescribed clinical drugs Donepezil and Galantamine had been tested in clinical trials, validating the model's predictive accuracy and practical relevance.

Conclusions HEDDI-NET represents a significant advancement by allowing direct application to new diseases and drugs without the need for retraining, a limitation of most GCN-based methods. Furthermore, HEDDI-Net provides detailed affinity patterns with representative proteins for predicted candidate drugs, facilitating an understanding of their physiological effects. This capability also supports the design and testing of alternative drugs that are similar to existing medications, enhancing the reliability and interpretability of potential repurposed drugs. The case study on Alzheimer's disease further underscores HEDDI-Net's ability to predict promising drugs and its applicability in drug repurposing.

Keywords Drug repurposing, Drug-disease associations, Heterogeneous embedding, Alzheimer's disease

*Correspondence:
Chen-Ching Lin
chenching.lin@nycu.edu.tw
Full list of author information is available at the end of the article

Background

As the complexity of diseases continues to grow, it is essential to study them from multiple perspectives to comprehensively understand their pathological mechanisms [1, 2]. However, developing safe and effective drugs is a time-consuming and costly endeavor, often spanning 10–15 years and costing an average of 2.6 billion US dollars [3, 4]. Moreover, less than 10% of new drugs are approved for clinical use after undergoing a series of lengthy drug design processes [5]. Consequently, drug repurposing—identifying new therapeutic applications for existing drugs beyond their original indications—offers a valuable alternative, particularly for diseases without known treatments [6]. The advantage of drug repurposing is rooted in the use of approved drugs that have already undergone rigorous animal testing and clinical trials as well as equipped with well-characterized safety profiles and documented side effects.

The accumulation of extensive, diverse databases and advancements in computer hardware and software has spurred the development of innovative computational methods in medical research [7]. Computational prediction methods for drug-disease associations generally fall into four categories: network propagation-based, matrix factorization- and completion-based, machine learning-based, and deep learning-based techniques [8–10]. These methods computationally identify previously undiscovered connections between diseases and approved drugs, significantly reducing time and expenses while offering potential candidates for further experimental validation. Notably, graph-based deep learning stands out as the current leader, acclaimed for its exceptional performance [8–11].

Graph-based deep learning methods employ multiple-layer artificial neural networks (ANNs) to learn patterns and relationships within diverse biological and biomedical networks encompassing drug and disease data. These techniques automatically derive data features, eliminating the need for manual feature engineering, and allowing the identification of complex, non-linear relationships between drugs and diseases. However, these methods mandate substantial data for effective training and can entail high computational costs. Moreover, interpreting the acquired models may present challenges, potentially hindering clinical integration. Nonetheless, deep learning-based methods have demonstrated encouraging outcomes, surpassing alternative methods in recent investigations. Examples include using random walking to explore the neighbor topology structures of diseases and drugs on different heterogeneous networks and constructing multi-layer convolutional network modules to learn the representative attributes for drug-disease node pairs, then aggregating topological and attribute representation to train predictive model from a multilayer neural network architecture [12]. Similar ideas have also been applied in discovering potential associations between miRNAs and diseases to investigate the molecular mechanisms and pathogenesis of complex diseases [13]. Other approaches involve graph convolutional networks (GCNs) and multilayer attention networks to encode the embeddings of diseases and drugs in inter- and intra-domain, then decoding to drug-disease association probability scores [14, 15]. Some methods even extract and integrate common and specific topologies and attributes in multiple heterogeneous networks and subnets from multi-sourced information of drugs and diseases via different types of graph convolutional auto-encoders, then adaptively integrating these representations for final association prediction [16]. Although these methods may capture complicated and multi-scale non-linear relationships between drug-disease pairs, most of them are limited to obtaining the predicted probabilities of drug-disease pairs. They are primarily focused on discovering unobserved links for known drugs and diseases from various heterogeneous sources without effectively presenting and explaining the biological functions and meanings contained in the used data.

To address the issue of limited interpretability and effectiveness of existing methods, we present a heterogeneous embedding drug and disease information network (HEDDI-Net) architecture that integrates shallow learning algorithms and graph theory into a deep learning model for drug-disease association prediction (Fig. 1). Our proposed method demonstrates excellent performance, achieving an area under the receiver operating characteristic curve of over 0.98 in the datasets obtained from Comparative Toxicogenomics Database (CTD) [17] after evaluation with tenfold cross-validation. Furthermore, it outperforms several state-of-the-art methods. The main contributions of this work can be summarized as follows:

- HEDDI-Net effectively connects the links of heterogeneous data, retaining data readability while achieving outstanding performance in both independent testing and comparison with other methods.
- The architecture retains high interpretability by utilizing specific proteins and diseases as features and training the model based on the affinity and similarity with the representative markers of each input pair (known drug and targeted disease).

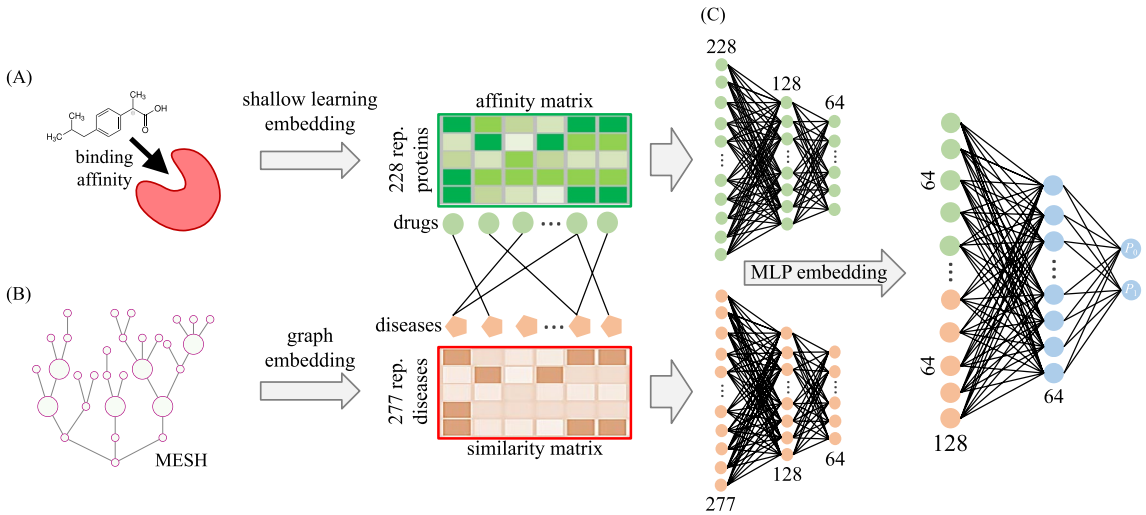

Fig. 1 HEDDI-Net workflow: protein-drug affinity, disease similarity, feature embedding, and association learning model. **A** Protein-drug affinity model. This component predicts the binding affinities between various drugs and representative proteins. The shallow learning embedding processes the binding affinity data to generate a comprehensive affinity matrix for the drugs. **B** Disease similarity model. This part calculates the semantic similarities between diseases using the MeSH hierarchical structure. It generates a similarity matrix by embedding the graph-based similarities, which helps in identifying representative diseases. **C** Association learning model. This deep learning model integrates the embeddings derived from the protein-drug affinity and disease similarity models. It concatenates these features to predict the association probabilities between drugs and diseases using a multi-layer perceptron (MLP) with dropout layers to ensure robust predictions

Importantly, unlike existing solutions, HEDDI-Net can be extended to new diseases or novel drugs without the need to modify the model.

Methods

Study design

This study aims to develop and validate HEDDI-Net, a deep learning-based model designed to predict drug-disease associations and facilitate drug repurposing. HEDDI-Net integrates protein-drug affinity profiles and disease similarity measures to generate interpretable embeddings that enhance the model's predictive accuracy. The model's performance and benchmark comparison were evaluated using tenfold cross-validation on both large and small datasets, as well as balanced and imbalanced sample pairs. Metrics such as AUC, AUPR, recall, specificity, accuracy, and F1-score were used to comprehensively assess the model's predictive capabilities. Furthermore, to validate HEDDI-Net's robustness, we systematically removed top-ranked diseases and drugs along with their associations and evaluated the model's ability to recover these associations using the remaining data. A case study focusing on Alzheimer's disease was conducted to demonstrate the model's interpretability and practical applicability. Potential drug candidates were clustered based on their binding affinity profiles and compared with commonly prescribed Alzheimer's drugs, Donepezil and Galantamine.

These comprehensive approaches highlight HEDDI-Net's potential to predict novel drug-disease associations, offering a valuable tool for drug repurposing and accelerating the drug discovery process.

HEDDI-Net architecture: harnessing shallow learning and graph theory for drug-disease associations

In this study, we introduce the Heterogeneous Embedding Drug and Disease Information Network (HEDDI-Net), an innovative architecture that synergizes a shallow learning algorithm and graph theory within a deep learning framework to identify drug-disease associations. The shallow learning algorithm and graph theory are strategically employed to extract representative proteins and diseases as interpretable features, respectively. These features are subsequently integrated into a deep neural network model to predict drug-disease associations. Figure 1 provides an overview of the HEDDI-Net workflow.

To gather drug-related information, we collected drug-protein interactions and their binding affinities from the BindingDB database [18–20]. We developed a shallow learning model to predict the binding affinity between proteins and drugs, thereby defining representative proteins. The predicted binding affinity profiles of these representative proteins were then utilized as input features for the drugs in HEDDI-Net (Fig. 1A).

For disease representation, we constructed a hierarchical directed acyclic graph (DAG) using Medical Subject Headings (MeSH) disease descriptors, grounded in the

semantic definitions of diseases. We identified eigen-diseases within the MeSH DAG to serve as representative diseases and calculated the semantic similarity between eigen-diseases and target diseases to generate input features for the diseases in HEDDI-Net (Fig. 1B).

Lastly, we trained a deep neural network on known disease-drug associations. The drug and disease features, derived from protein-drug affinity and disease similarity profiles, were utilized as input nodes. The deep neural network, consisting of multiple fully connected layers, predicted the association probability for each drug-disease pair (Fig. 1C).

Chemical-disease associations for training and testing HEDDI-Net

We utilized chemical-disease relationships from the Comparative Toxicogenomics Database (CTD) [17] as drug-disease associations to train HEDDI-Net. Two types of chemical-disease associations were employed for training and evaluating our model:

1. Direct Evidence: This dataset included all association data with direct evidence, such as marker/mechanism and therapeutic relations, resulting in 71,187 chemical-disease associations between 6,074 chemicals and 2,802 diseases.
2. Therapeutic: This dataset comprised only therapeutic association data, deemed more reliable, with 26,789 drug-disease associations between 4,157 drugs and 2,149 diseases.

Feature embedding of drugs in HEDDI-Net

In HEDDI-Net, achieving model interpretability necessitated the utilization of protein binding affinity profiles to embed drug features. Drugs generally exhibit multiple binding sites for various proteins, resulting in diverse effects on molecular biology and pharmaceutical functions [21, 22]. To address the complexity inherent in these interactions, we employed the Support Vector Regression (SVR) algorithm to predict the affinity between different drugs and individual proteins, thereby encapsulating the binding characteristics of each drug. These predicted binding affinity profiles, particularly those with the representative proteins, were subsequently employed as input features for drugs in our drug-disease association prediction model.

To develop affinity prediction models (Aff) for proteins (PR) and drugs (DR), we collected information on drug-protein interactions and their binding affinities from Binding Database (BindingDB), a publicly accessible repository containing experimentally derived binding affinities of protein–ligand interactions. BindingDB comprises over 2.5 million binding data entries for more than 8,900 target proteins and 1.1 million drug-like small molecules up to mid-2022. We specifically selected human proteins with unique UniProt [23] entry names and drug compounds possessing structural information, PubChem CID [24], InChI key [25], and binding affinity to corresponding proteins. The binding affinity served as the dependent variable for SVR model prediction, while the structural information of drugs (F_1^{DR}) was used as the first set of features. We obtained the structural information from BindingDB and converted it into a 166-bit MACCS set [26] using RDKit [27]. The second set of features, the physicochemical properties of drug compounds (F_2^{DR}), was acquired from the ChEMBL [28, 29], a public database maintained by the European Bioinformatics Institute (EBI), using InChI keys. ChEMBL includes 2-D structures, calculated properties, and bioactivity data from primary scientific literature. We selected 17 physicochemical properties potentially relevant to binding affinity for training the regression model, as listed in Table S1. After filtering, 203,725 binding data entries between 1,289 proteins and 118,366 small molecules met the criteria and were used to establish affinity models in this study.

The affinity model of protein (PR) was trained using the following equation:

$$Aff_{DR}^{PR} = SVR(F_1^{DR}, F_2^{DR}) \tag{1}$$

where Aff_{DR}^{PR} represents the affinity value for drug DR with protein PR,$F_1^{DR} \in \mathbb{R}^{A^{PR} \times S^{DR}}$, and$F_2^{DR} \in \mathbb{R}^{A^{PR} \times P^{DR}}$. In this context, A^{PR} denotes the number of drugs with available affinity data for protein PR, S^{DR} is 166, which is the length of MACCS set, and P^{DR} is 17, which is the number of used physicochemical properties of drug compounds. The affinity profiles of drugs (I_{DR}) to the representative proteins (PR_r) were used as input features for drugs in the drug-disease association learning model, represented as:

$$I_{DR} = Aff_{DR}^{PR_r} \tag{2}$$

To ensure that the representative proteins effectively represent the relevant biological interactions, we implemented a rigorous filtering and evaluation process. This process was crucial for identifying the most informative proteins, thereby enhancing the model's predictive accuracy and interpretability. We initially filtered proteins with at least 15 small molecule drug affinity data, resulting in 645 proteins meeting the criteria. For each protein, we partitioned the affinity data into 70% for training and 30% for testing. The testing data was further utilized

to evaluate the performance of the protein-drug affinity model using Spearman's rank correlation coefficient (Spearman's ρ) [30].

To ensure that superior performance was not due to chance, we conducted 1,000 permutation tests for each model and calculated the z-score based on the permutations. The z-score equation [31] used to evaluate the performance of the protein-drug affinity model is:

$$\mathrm{z} = \frac{\rho - \mu_{perm}}{\sigma_{perm}} \tag{3}$$

where ρ represents the observed Spearman's ρ, μ_{perm} is the mean of the Spearman's ρ from the 1000 permutation tests, and σ_{perm} is the standard deviation of the Spearman's ρ from the 1000 permutation tests. This process (1000 permutation tests for z-score calculation) was repeated for 100 times by randomly dividing training and testing dataset. Accordingly, for each protein, 100 z-scores were calculated, and the median z-score was used to evaluate its model performance in predicting binding affinity of drugs.

We selected 228 proteins with a median z-score greater than or equal to 4 as representative proteins (PR_r) based on the testing performances. It is important to note that, following this selection, we used all available small molecule drug data to construct the affinity models for the representative proteins. Additionally, we applied MinMaxScaler to normalize the data for all affinity models, based on the physicochemical properties of all 118,366 drug-like small molecules. This normalization process was subsequently applied to the drug-disease association models, ensuring consistent scaling and facilitating the integration of affinity profiles into the overall predictive framework.

Feature embedding of diseases in HEDDI-Net

The Medical Subject Headings (MeSH) database (https://www.ncbi.nlm.nih.gov/mesh/) [32], a controlled vocabulary thesaurus maintained by the National Library of Medicine (NLM) and encompasses subject headings in MEDLINE/PubMed and other NLM databases, serves as a critical resource for indexing, cataloging, and searching biomedical and health-related information. MeSH descriptors are systematically organized in a hierarchical structure, ranging from general to specific disease terms across up to thirteen hierarchical levels, thereby forming a hierarchical directed acyclic graph (DAG). Based on the layered structure of the DAG, we posited that diseases sharing a greater number of descriptors would exhibit more common phenotypes and symptoms, impact similar physiological functions, and possess closely related molecular origins. This implies that such diseases may respond to similar therapeutic interventions or treatments.

In this study, we utilized a total of 4,933 diseases to construct the DAG tree and quantify the relationship between two diseases based on the MeSH descriptor 2022. We employed Wang's method [33] to calculate semantic similarity. Wang's method leverages the topological information of two nodes (MeSH descriptors) within the biomedical ontology tree and accounts for the varying contributions of each node. This approach is widely recognized for its efficacy in similarity calculations due to its comprehensive consideration of the hierarchical structure and node significance within the DAG.

Assuming a tree structure of disease DI is represented as $DAG_{DI} = (DI, T_{DI}, E_{DI})$, where T_{DI} is the set of all ancestor nodes of DI in DAG_{DI} (including the term DI), and E_{DI} is the set of corresponding links (semantic relations). The semantic value of disease DI, as the cumulative contribution of all nodes in DAG_{DI} to term DI, can be denoted as:

$$\mathrm{SV}(DI) = \sum_{t \epsilon T_{DI}} S_{DI}(t) \tag{4}$$

where $S_{DI}(t)$ is the semantic value of term t related to term DI, defined as:

$$S_{DI}(t) = \begin{cases} 1 & if\ t = DI \\ max\{\Delta \times S_{DI}(t') | t' \epsilon \text{ childeren of } t\} & if\ t \neq DI \end{cases} \tag{5}$$

where Δ is the semantic contribution factor for term t with its child term $t\prime$, which is chosen between 0 and 1 to reduce the contributions of ancestor nodes that are far from term DI. In this study, Δ is set to 0.5 based on recommendations from findings in the original literature [33]. The semantic similarity between disease DI_a and disease DI_b is calculated as:

$$\mathrm{Sim}(DI_a, DI_b) = \frac{\sum_{t \epsilon T_{DI_a} \cap T_{DI_b}} (S_{DI_a}(t) + S_{DI_b}(t))}{\mathrm{SV}(DI_a) + \mathrm{SV}(DI_b)} \tag{6}$$

where t represents the common ancestor nodes of disease DI_a and disease DI_b. More common ancestors tend to create higher semantic similarities. $S_{DI_a}(t)$ is the semantic value of term t related to term DI_a, and $S_{DI_b}(t)$ is the semantic value of term t related to term DI_b. The semantic similarity profiles of diseases DI to the representative diseases DI_r are used as the input features of diseases in the drug-disease association learning model and are expressed as:

$$I_{DI} = \mathrm{Sim}(DI, DI_r) \tag{7}$$

Building on the semantic similarity calculations, we aimed to identify representative diseases that could serve as key nodes within the MeSH tree. These representative diseases were defined as playing a crucial role in capturing the interconnectedness of various diseases

and enhancing the predictive accuracy of our model. The identification of these representative diseases is essential for developing a robust disease embedding framework within HEDDI-Net.

Eigenvector centrality [34, 35] is a measure of a node's importance in a network, accounting for the importance of the nodes to which it is connected. Within the MeSH tree structure, eigenvector centrality is utilized to identify the most representative diseases, or eigen-diseases, in the network. This approach assesses the importance of a disease by considering its connections to other influential diseases. A high eigenvector score indicates that a disease is well-connected to other diseases with high scores.

To apply this method, we first converted the MeSH Directed Acyclic Graph (DAG) into an undirected tree structure. We then used the eigenvector centrality method to identify the top 277 representative nodes (DI_r), based on their similarity distribution with CTD diseases (Supplementary Fig. S1). This selection process ensured that the most significant diseases, in terms of their network connectivity, were included as representative nodes.

We mathematically represented our MeSH tree structure as $G = \{V, E\}$, where V denotes vertices and E denotes edges. The adjacency matrix of the graph G was represented by A, with elements $A_{ij} = 1$ if there is a connection between vertices i and j, otherwise $A_{ij} = 0$. The centrality of vertex i was denoted by x_i, adjusted for this effect by making x_i proportional to the average of the centralities of i's network neighbors using the equation:

$$x_i = \frac{1}{\lambda} \sum_{j \in V \neq i} A_{ij} x_j \tag{8}$$

where λ is a constant. We defined the vector of centralities as $x = (x_1, x_2, \dots)$, and rewrote the equation in matrix form as $\lambda x = A \bullet x$. Thus, x is an eigenvector of the adjacency matrix A with eigenvalue λ. Supplementary Figure S2 shows the score distribution of eigenvector centrality for the MeSH tree. This distribution highlights the central nodes within the MeSH tree, which play a pivotal role in capturing the interconnectedness of diseases within the biomedical ontology. Identifying these central nodes enhances our understanding of disease relationships and supports the robust selection of representative diseases for subsequent analysis. This detailed eigenvector centrality analysis, combined with the semantic similarity measures, provides a comprehensive framework for embedding disease features within HEDDI-Net, thereby enhancing the predictive accuracy of our drug-disease association model.

Deep learning model to predict association probability

To predict the association probability between drugs and diseases, we proposed a deep learning model that concatenates embedded features derived from protein-drug affinity and disease similarity. Given a pair of drug i (DR_i) and disease j (DI_j), along with their respective affinity profile I_{DR_i} and similarity profile I_{DI_j} obtained from previously described feature embedding approaches, the drug embedding is defined as:

$$Z_{DR_i} = Drop_{0.25}(f_2(f_1(I_{DR_i}))) \tag{9}$$

where I_{DR_i} is a 228-dimensional vector recording the binding affinities of drug i to the 228 representative proteins, $f_1(\cdot)$ indicates the tanh activation function with 228 input dimensions and 128 output dimensions, while $f_2(\cdot)$ represents the tanh activation function with 128 input dimensions and 64 output dimensions. $Drop_{0.25}(\cdot)$ denotes a dropout layer with a rate of 0.25, randomly selecting neurons to be dropped out (as illustrated in the upper MLP of Fig. 1C).

Similarly, the disease embedding is defined as:

$$Z_{DI_j} = Drop_{0.25}(f_2(f_1(I_{DI_j}))) \tag{10}$$

where I_{DI_j} is a 277-dimensional vector consisting of the similarities of disease i with the 277 representative diseases. The function $f_1(\cdot)$ represents the tanh activation function with 277 input dimensions and 128 output dimensions, while $f_2(\cdot)$ represents the tanh activation function with 128 input dimensions and 64 output dimensions (as illustrated in the lower MLP of Fig. 1C).

The predicted association probability between the drug and disease, denoted as $\widehat{P}_{ij}$, is obtained through a neural network model that concatenates the embeddings of the protein-drug affinity and disease similarity models. Specifically, we define

$$\widehat{P}_{ij} = softmax(Drop_{0.5}(f_2(f_1(Z_{DR_i} \oplus Z_{DI_j})))) \tag{11}$$

where $\oplus$ indicates the concatenation operation, $f_1(\cdot)$ is a sigmoid function with 128 input dimensions and 64 output dimensions, and $f_2(\cdot)$ is a sigmoid function with 64 input dimensions and 2 output dimensions. $Drop_{0.5}(\cdot)$ represents a dropout layer with a 0.5 rate, and $softmax(\cdot)$ represents a nonlinear activation function (right MLP of Fig. 1C).

The model was optimized by minimizing the categorical cross-entropy loss function [36]

$$\mathcal{L} = -\sum_{(i,j) \in N_{train}} y_{i,j} log \widehat{P}_{ij}, + (1 - y_{ij}) log(1 - \widehat{P}_{ij}) \tag{12}$$

where N_{train} is the set of training pairs and $y_{i,j}$ indicates the real association label of (DR_i) – (DI_j) derived from CTD. When there is an actual association between DR_i and DI_j, $y_{i,j} = 1$; otherwise, $y_{i,j} = 0$. The optimization was performed using the Adam algorithm [37].

Experimental design and hyper-parameter settings

To evaluate the performance and generalizability of our drug-disease association prediction model, we employed tenfold cross-validation approach using direct evidence and therapeutic datasets from CTD [17]. Comprehensive evaluation metrics were used, encompassing ranking-based metrics such as the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPR). Additionally, threshold-based metrics including recall, specificity, accuracy (ACC), and F1-score were employed to assess the model's performance, thereby ensuring a robust evaluation while mitigating potential dataset biases.

To create a balanced dataset, we randomly selected an equal number of negative samples to match the number of positive samples. This process ensured an equal representation of positive and negative instances in the dataset, thereby reducing potential bias in the model's training and evaluation. To minimize misclassification risk, we carefully selected negative samples by randomly sampling drug-disease combinations from the therapeutic dataset while rigorously excluding any potential positive links. Specifically, we cross-referenced the CTD database, considering both curated and inferred chemical-disease associations. Curated associations, based on published literature, provide direct evidence of positive interactions, whereas inferred associations link drugs and diseases via shared gene interactions. Since drugs and diseases connected through common genes are more likely to represent true associations, we excluded these pairs from the negative sample set to avoid mislabeling potential positives as negatives. This approach strengthens the reliability of our model by reducing the likelihood that undiscovered positives are mistakenly included among negative samples, enhancing predictive accuracy.

Regarding the hyper-parameter settings in the deep learning model, we considered various combinations for the batch size and number of epochs. Specifically, we explored batch sizes ranging from {10,000, 11,000, 12,000, 13,000, 14,000, 15,000} and epochs ranging from {1,000, 1,100, 1,200, 1,300, 1,400, 1,500, 1,600}. Through extensive experimentation, we determined that the optimal settings were 1,600 epochs, a batch size of 11,000, and a learning rate of 10^{-3}. All results reported for the testing data were validated using retrained models based on the original training data, ensuring robustness and reliability of the model's performance.

Results

Accurate prediction of drug-disease associations with HEDDI-Net, surpassing state-of-the-art models

To ensure the capability of HEDDI-Net in drug repurposing, we evaluated its effectiveness in predicting known drug-disease associations. We applied the HEDDI-Net architecture to build models for the CTD datasets: direct evidence and therapeutic datasets, separately. To ensure more reliable negative datasets, we generated balanced negative instances using unobserved associations between 4157 chemicals and 2149 diseases in the therapeutic dataset.

In the direct evidence dataset, which contains 71,187 drug-disease associations (positive samples), HEDDI-Net achieves a median AUC and AUPR greater than 0.982 in a tenfold cross-validation (Fig. 2A). For the therapeutic dataset, comprising 26,789 drug-disease associations, the median AUC and AUPR in the tenfold cross-validation are both over 0.986 (Fig. 2A). These results demonstrate that HEDDI-Net consistently and accurately predicts known drug-disease associations.

To better understand the contribution of drug and disease features to the model's performance, we evaluated the performance using either drug (Z_{DR}) or disease (Z_{DI}) embeddings alone. As shown in Fig. 2A, for the direct evidence dataset, the performance of Z_{DR} surpass that of Z_{DI}. Combining both Z_{DR} and Z_{DI} improves the tenfold median AUC and AUPR by 10.58% and 9.74%, respectively, compared to using only Z_{DR}. For the therapeutic dataset, there is no significant difference between the performance of Z_{DR} and Z_{DI}. However, combining both embeddings greatly improves the average median AUC by 18.18% and the average median AUPR by 17.18%.

These findings demonstrate that our model can generate accurate predictions using either drug or disease information alone when sufficient sample size is available. However, when dealing with smaller datasets, such as therapeutic dataset, combining both drug and disease features largely enhances the model's performance. This highlights the importance of using both features in scenarios with limited data. Furthermore, the results from using either drug or disease embeddings alone further validate that the representative proteins and diseases effectively capture the essential information of drug-disease pairs.

To further confirm HEDDI-Net's performance, we benchmarked it against four state-of-the-art approaches——DRHGCN [15], DRWBNCF [38], LAGCN [14], and HINGRL [39] (Supplementary Materials)——using a dataset of 18,416 known drug-disease associations involving 269 drugs and 598 diseases [40]. This dataset has been used in the studies of these four methods, but it had fewer drugs and diseases compared

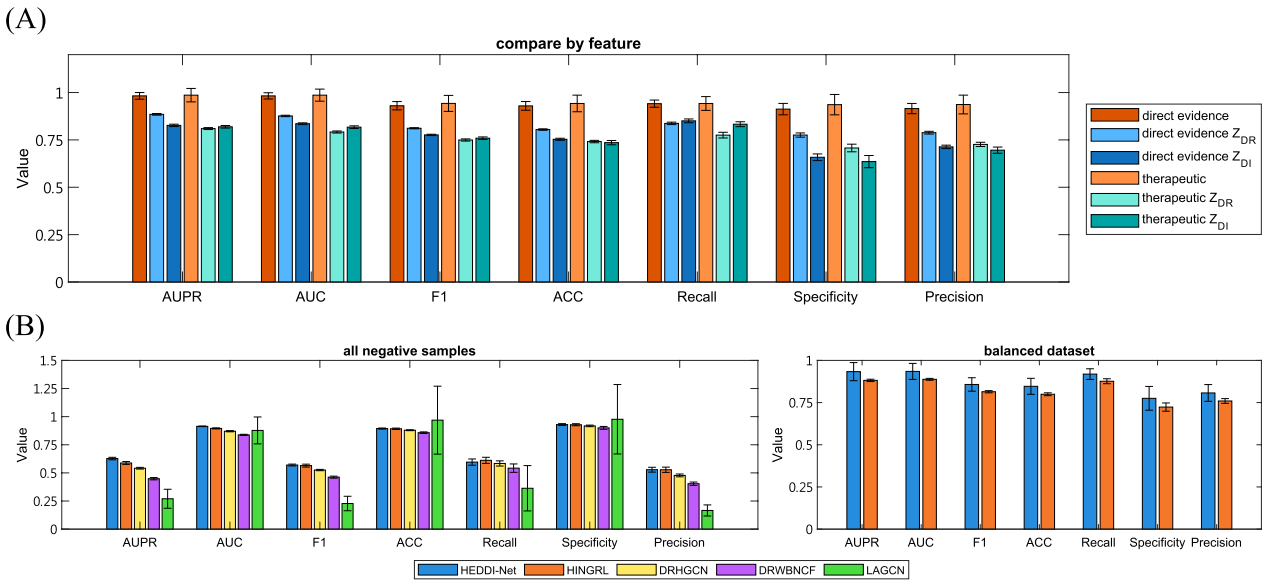

Fig. 2 Comparative performance analysis of HEDDI-Net and other state-of-the-art models. **A** Performance evaluation of HEDDI-Net using direct evidence and therapeutic datasets with different feature sets. The comparison is made between the direct evidence dataset (left) and the therapeutic dataset (right), with further analysis based on drug embedding (Z_{DR}) and disease embedding (Z_{DI}). Here, Z_{DR} represents the use of only drug embeddings (excluding disease embeddings and their corresponding MLP), while Z_{DI} represents the use of only disease embeddings (excluding drug embeddings and their corresponding MLP). **B** Comparison of HEDDI-Net with state-of-the-art models (HINGRL, DRHGCN, DRWBNCF, LAGCN) using all negative samples and a balanced dataset. The performance metrics are evaluated across various datasets to highlight HEDDI-Net's effectiveness in predicting drug-disease associations. The error bars represent standard deviations derived from tenfold cross-validation

to our dataset used for training and evaluating HEDDI-Net. To ensure a fair comparison, we retrained all models, including HEDDI-Net, on this smaller dataset by using 18,416 known and all unobserved drug-disease associations as positive and negative samples, respectively. For the four state-of-the-art approaches, we used hyperparameters that were either built-in or recommended by their respective studies. During HEDDI-Net training, we set the batch size to 13,000 and the number of epochs to 1,400 for all negative sample scenario.

HEDDI-Net demonstrates superior performance compared to all other methods in terms of AUPR, AUC, F1 score, and precision (left panel in Fig. 2B and Table S2). Additionally, HEDDI-Net secures the second-highest scores in ACC, recall, and specificity. Among the graph deep learning-based methods, HINGRL shows the second-best performance, highlighting the importance of incorporating biological information in predictive modeling. However, LAGCN exhibits the lowest overall performance while achieving the highest ACC and specificity. This suggests that LAGCN tends to generate a large number of negative predictions to enhance accuracy and specificity in the imbalanced dataset, negatively impacting recall and precision. Furthermore, we also compare the performance of HINGRL with HEDDI-Net on a balanced dataset. The results indicate that HEDDI-Net outperforms HINGRL in all evaluation metrics (right panel in Fig. 2B and Table S2). This further substantiates the effectiveness of our approach in predicting drug-disease associations, even with datasets containing fewer drugs and diseases.

In summary, HEDDI-Net has demonstrated exceptional predictive performance in identifying drug-disease associations, consistently outperforming state-of-the-art models. By leveraging both drug and disease features, HEDDI-Net ensures high accuracy even with smaller datasets. The model's robust performance unveils its potential for drug repurposing, thereby facilitating the drug discovery and development process.

High recovery rate of drug-disease associations: HEDDI-Net for drug repurposing

To evaluate HEDDI-Net's applicability in drug repurposing, we conducted a systematic procedure. We sequentially removed the top 100 drugs or diseases according to the number of associations they possessed in a descending order, along with their respective associations. For each removal, we used the rest diseases, drugs, and drug-disease associations to train a model, using the removed associations of the removed drugs (diseases) for independent validation.

Notably, we observed a median recovery rate of 73% for retrieving the removed drug-disease associations from the models of the top 100 diseases (Fig. 3A). This

result suggests that HEDDI-Net can effectively identify unknown target diseases for existing drugs, supporting its potential for drug repurposing. However, the median recovery rate is only 40% for retrieving the removed drug-disease association from the models of the top 100 drugs (Fig. 3B). This performance indicates that while our model shows promise in developing new drugs for existing diseases, further experimental validation is needed. Additionally, this analysis demonstrates the reliability and solidity of HEDDI-Net.

State-of-the-art graph-based deep learning approaches, although effective, require retraining to predict drug-disease associations for previously unseen drugs and diseases. In contrast, HEDDI-Net allows new drugs and diseases to be directly fed into the pre-trained model without the need for retraining, making it more practical than many advanced graph-based deep learning methods.

In summary, our robustness analysis highlights HEDDI-Net's ability to predict novel drug-disease associations, even in diverse conditions or unknown associations. This capability positions HEDDI-Net as a powerful tool for drug repurposing and drug discovery, emphasizing its practicality and efficiency in handling new data without retraining.

Application of HEDDI-Net in Alzheimer's drug discovery

To further substantiate the applicability of our model, we conducted a case study on the relationship between Alzheimer's disease (AD) and the candidate drugs identified by HEDDI-Net. AD, an unstoppable and irreversible brain disorder, is the primary contributor to dementia among the elderly population. Previous research suggests that the primary pathological hallmarks of AD include extracellular plaques constituted by amyloid-β (Aβ) and intracellular neurofibrillary tangles (NFT) comprised of tau protein [41]. Unfortunately, the current medical landscape offers no definitive remedy for AD, leaving available therapeutic interventions restricted to alleviating symptoms, with no capacity to alter the underlying progression of the disease.

In our case study, we trained 100 models using therapeutic drug-disease associations. Each model encompassed all positive data and an equal number of randomly selected negative samples. Subsequently, we tested all chemicals in the dataset for their association with AD. For each model, we retained predictions for chemicals if the positive probability exceeded 0.8, ensuring a high confidence level in the results. To determine the collective consensus of predicted drugs, we summarized the probability values of all 100

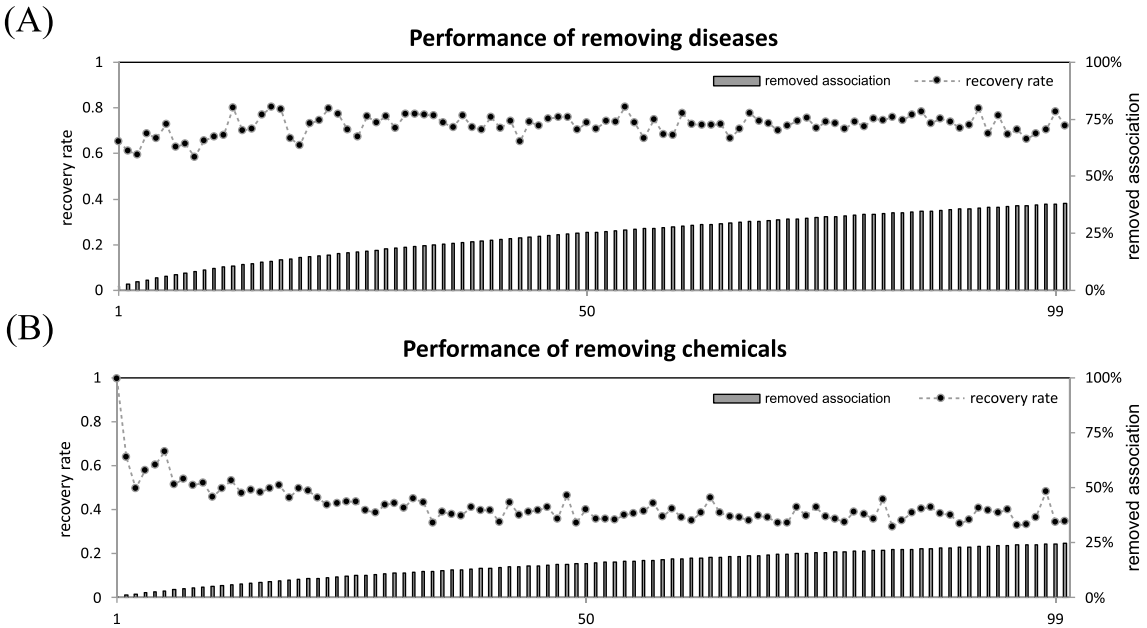

Fig. 3 HEDDI-Net's capacity to recover drug-disease associations for drug repurposing. **A** Disease association removal and recovery. Associations were removed based on the number of associations for each disease. Diseases within the direct evidence dataset were first ranked by their number of associations in descending order. The top 100 diseases, along with all their associations, were systematically eliminated. This process allowed us to evaluate the model's capacity to recover the removed associations by considering the remaining data. **B** Drug association removal and recovery. Similar to the disease analysis, drugs within the direct evidence dataset were ranked by their number of associations. The top 100 drugs, along with all their associations, were systematically removed. The model's ability to recover these missing associations was then assessed using the remaining data. These two procedures demonstrated the robustness of HEDDI-Net in recovering critical drug-disease relationships

models, aiming to eliminate bias stemming from the randomness of the sampling process. We identified the predicted drugs, which are not associated with AD in CTD, achieving a cumulative sum of positive probability exceeding 50 as candidates, totaling 58 chemicals. This threshold indicates a consistent and robust association with AD on the predictions derived from the 100 models.

To delve deeper into the potential of these 58 candidate drugs, we explored their relationships to two commonly prescribed medications——Donepezil and Galantamine——using the proximity analysis. Specifically, we investigated the similarity of binding affinity profiles to the 228 representative proteins between the candidates and the two prescribed medications. We then performed the conventional hierarchical clustering to identify the candidates clustered with Donepezil, Galantamine, or both. Interestingly, Donepezil and Galantamine were separated into two different clusters. We therefore assigned the predicted candidates into three groups——donepezil, galantamine, and others——according to their proximity to the two prescribed drugs (Fig. 4 and Table 1).

According to the records of ClinicalTrials.gov from NIH, we found that 8 out of 19 (42%) candidates in the donepezil cluster, 8 out of 17 (47%) candidates in the galantamine cluster, and 9 out of 22 (41%) candidates in the others cluster have been tested in clinical trials (Table 1). Briefly, more than 40% of predicted candidates have been recognized as possessing promising potential in treating AD by the pharmaceutical industry. These results further emphasize the predictive accuracy and practicability of HEDDI-Net in real world applications. Moreover, it provides a demonstration of how to apply HEDDI-Net to predict candidate chemicals for drug repurposing in practice.

Moreover, we investigated those predicted drugs that have not been tested in clinical trials to demonstrate the interpretability of HEDDI-Net. That is, by studying the target proteins, we can interpret why those predicted drugs clustered with the two known prescribed medications, and assess the repurposing potential of the predicted candidate drugs. In the donepezil cluster (Fig. 4), besides those predicted drugs that have been tested in clinical trials, HEDDI-Net also identified Baclofen and Fluoxetine as promising candidate drugs for the treatment of AD. Baclofen, a GABA-ergic agonist, has demonstrated potential neuroprotective effects by modulating neuroinflammation and reducing neuronal excitotoxicity, which are critical in preventing neuronal damage and cognitive decline associated with AD [42, 43]. Fluoxetine, a selective serotonin reuptake inhibitor (SSRI), not only improves mood and cognitive function by increasing serotonin levels but also exhibits neuroprotective properties and reduces neuroinflammation [44, 45]. These drugs address key pathological aspects of AD, including neuroinflammation, psychiatric symptoms, and neurodegeneration, thereby supporting their potential repurposing for this condition. Moreover, by investigating the drug feature embedding of HEDDI-Net, we discovered that Baclofen and Fluoxetine may target ITGB3 (ITB3 protein): their predicted targeted protein with the second

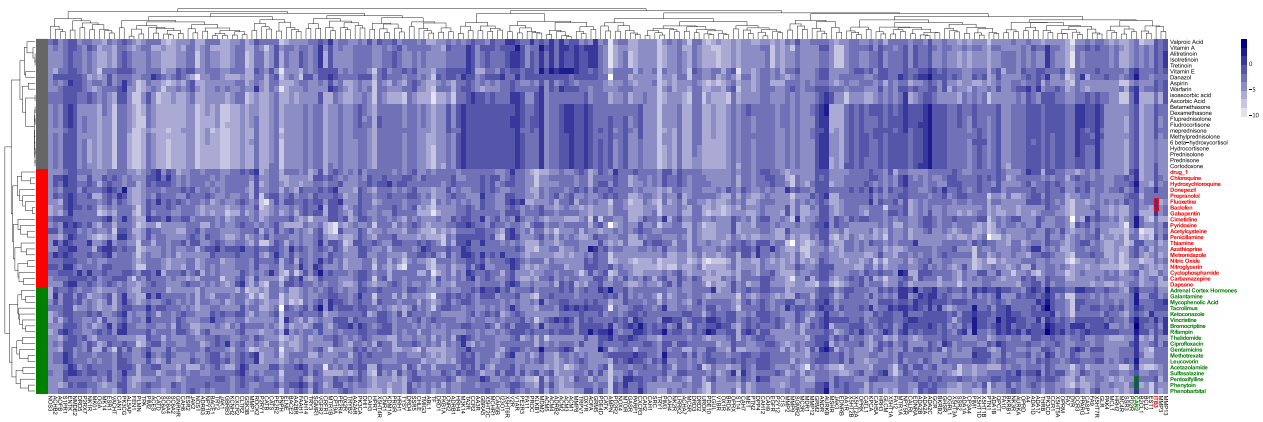

Fig. 4 HEDDI-Net's applicability and interpretation: Clustering drugs candidates with commonly prescribed drugs in Alzheimer's disease by affinity. The selected high-confidence drug-like chemicals (rows) were clustered based on the similarity of their binding affinity profiles to the representative proteins (columns). The color bar represents the strength of binding affinity. According to their proximity to two commonly prescribed drugs for Alzheimer's disease, Donepezil and Galantamine, the chemicals were clustered into three groups: the donepezil (colored red), the galantamine (colored green), and others clusters. The proteins selected to display the interpretability of HEDDI-Net were colored to correspond with their respective group. Among the chemicals, drug_1 represents N-(oxo-5,6-dihydrophenanthridin-2-yl)-N, N-dimethylacetamide hydrochloride

Table 1 Predicted candidate drugs in clinical trials

Chemical Name	ID	study type	Phase	trail status	Last update	Cluster
Cyclophosphamide	NCT00013650	Interventional	Phase 1	Completed	2017–07-02	Donepezil
Nitric Oxide	NCT03451591	Interventional	Phase 2 & 3	Completed	2022–08–25	
Gabapentin	NCT00018291	Interventional	NA	Completed	2009–01–21	
	NCT03082755	Interventional	Phase 4	Unknown	2022–05–18	
Dapsone	NCT05894954	Interventional	Phase 3	Recruiting	2024–05–16	
Hydroxychloroquine	NCT05894954	Interventional	Phase 3	Recruiting	2024–05–16	
Metronidazole	NCT05894954	Interventional	Phase 3	Recruiting	2024–05–16	
Thiamine	NCT06223360	Interventional	Phase 2	Recruiting	2024–04–30	
	NCT02292238	Interventional	Phase 2	Completed	2022–06–28	
Acetylcysteine	NCT04740580	Interventional	Early Phase 1	Recruiting	2024–03–21	
	NCT04044131	Interventional	Phase 2	Completed	2022–08-08	
	NCT01370954	Observational	NA	Completed	2013–05–09	
Not in a trail (11)	Azathioprine, Baclofen*, Carbamazepine, Chloroquine, Cimetidine, Fluoxetine*, drug_1, Nitroglycerin, Penicillamine, Propranolol, Pyridoxine					
Bromocriptine	NCT04413344	Interventional	Phase 1 & 2	Completed	2022–04–07	Galantamine
Methotrexate	NCT04571697	Observational	NA	Completed	2021–10–18	
Ketoconazole	NCT00860275	Interventional	Phase 1	Completed	2011–1–25	
	NCT00931073	Interventional	Phase 1	Completed	2009–11–18	
Acetazolamide	NCT05443308	Observational	NA	Recruiting	2022–07-05	
Ciprofloxacin	NCT06185543	Interventional	Phase 2	Recruiting	2024–01–18	
Thalidomide	NCT01094340	Interventional	Phase 2 & 3	Unknown	2012–08-08	
Rifampin	NCT00715858	Interventional	Phase 3	Unknown	2009–02–04	
	NCT00439166	Interventional	Phase 3	Completed	2018–03–19	
	NCT00692588	Observational	NA	Completed	2011–04–06	
Tacrolimus	NCT04263519	Interventional	Phase 2	Withdrawn	2021–09–27	
Not in a trail (9)	Adrenal Cortex Hormones, Gentamicins, Leucovorin, Mycophenolic Acid, Pentoxifylline*, Phenobarbital, Phenytoin*, Sulfasalazine, Vincristine					
Ascorbic Acid	NCT00117403	Interventional	Phase 1	Completed	2009–04-03	others
Prednisone	NCT00000178	Interventional	Phase 3	Completed	2005–06–24	
Warfarin	NCT00827034	Interventional	Phase 1	Completed	2018–10–16	
	NCT00689637	Interventional	Phase 1	Completed	2009–07-02	
	NCT00726726	Interventional	Phase 1	Completed/	2008–11-05	
Valproic Acid	NCT01729598	Interventional	Early Phase 1	Completed	2019–10-09	
	NCT00071721	Interventional	Phase 3	Completed	2014–09–25	
	NCT00088387	Interventional	Phase 2	Completed	2008–03–04	
	NCT00208819	Interventional	Phase 4	Completed	2013–11–13	
	NCT00375557	Interventional	Phase 4	Withdrawn	2015–05–27	
Vitamin E	NCT00235716	Interventional	Phase 3	Completed	2014–07–23	
	NCT00040378	Observational	NA	Completed	2018–03–14	
	NCT00000173	Interventional	Phase 3	Completed	2009–12-11	
	NCT00056329	Interventional	Phase 3	Unknown	2012–05-04	
	NCT01594346	Interventional	Phase 3	Completed	2012–05–09	
	NCT00117403	Interventional	Phase 1	Completed	2009–04-03	
	NCT01320527	Interventional/	Phase 2/	Completed/	2016–03-03	
Aspirin	NCT05894954	Interventional	Phase 3	Recruiting	2024–05–16	
Hydrocortisone	NCT05894954	Interventional	Phase 3	Recruiting	2024–05–16	
Isotretinoin	NCT01560585	Interventional	Phase 1 & 2	Terminated	2022–06–15	
Tretinoin	NCT02439099	Observational	NA	Unknown	2021–08–09	
Not in a trail (13)	6 beta-hydroxycortisol, Alitretinoin, Betamethasone, Cortodoxone, Danazol, Dexamethasone, Fludrocortisone, Fluprednisolone, Isoascorbic acid, Meprednisone, Methylprednisolone, Prednisolone, Vitamin A					

drug_1: N-(oxo-5,6-dihydrophenanthridin-2-yl)-N, N-dimethylacetamide hydrochloride

* indicates that the drug is discussed in the article

highest binding affinity. ITGB3 is involved in neuroinflammatory processes, which are significant in AD [46]. Activated microglia, mediated by ITGB3, contribute to the neuroinflammation seen in AD, exacerbating disease progression [47, 48]. Accordingly, Baclofen's and Fluoxetine's interaction with ITGB3 could help modulate these inflammatory pathways, reducing neuroinflammation and protecting neuronal integrity. These mechanisms align with the established pathophysiological processes in AD, elucidating the potential of Baclofen and Fluoxetine in treating this neurodegenerative disorder.

In the galantamine cluster (Fig. 4), besides those predicted drugs that have been tested in clinical trials, our model has identified Pentoxifylline and Phenytoin as promising candidate drugs for the treatment of AD. Pentoxifylline, known for improving blood flow and reducing blood viscosity, also exhibits anti-inflammatory and neuroprotective properties. By enhancing cerebral blood flow and reducing neuroinflammation, Pentoxifylline could help mitigate some of the neurodegenerative processes in AD [49, 50]. Phenytoin, another anticonvulsant, similarly stabilizes neuronal membranes and decreases excitability, potentially mitigating hyperexcitability and excitotoxicity observed in Alzheimer's pathology [51]. Collectively, these drugs address key aspects of AD, including neuroinflammation, impaired cerebral blood flow, and excitotoxicity, supporting their potential repurposing for this condition. In our model, we discovered that Pentoxifylline and Phenytoin may all target CAPN2 (Calpain 2, CAN2 protein), which is their predicted targeted protein with the second highest binding affinity. Calpain 2 is implicated in neurodegenerative processes through its role in tau pathology and amyloid-beta (Aβ) production. Abnormal activation of Calpain 2 leads to the hyperphosphorylation of tau and the formation of neurofibrillary tangles, as well as increased Aβ production and aggregation, which are hallmark features of AD [52, 53]. By targeting Calpain 2, Pentoxifylline and Phenytoin could potentially reduce tau hyperphosphorylation and Aβ accumulation, thereby mitigating neurodegeneration. The neuroprotective effects of these two drugs, combined with their ability to stabilize neuronal activity and reduce neuroinflammation, make them compelling candidates for the treatment of AD, as suggested by our deep learning model.

In summary, this case study on Alzheimer's disease underscores HEDDI-Net's predictive capability in identifying candidate drugs for repurposing. With over 40% of the predicted candidates having been tested in clinical trials, the model's accuracy and practical relevance are evident, highlighting HEDDI-Net's potential for real-world applications. Importantly, HEDDI-Net not only identifies promising candidates but also interprets why these drugs can treat AD by investigating their target proteins. By elucidating the key processes in which the candidate drugs' target proteins are involved in AD, HEDDI-Net provides valuable insights into their repurposing potential. These achievements emphasize the model's effectiveness in predicting drug-disease associations and its utility in the drug discovery and development process.

Discussion

In this study, we present HEDDI-Net, a deep learning architecture designed to predict drug-disease associations with high accuracy, outperforming existing methods and reaffirming its effectiveness in this research domain. HEDDI-Net enhances prediction performance and interpretability by integrating heterogeneous embedded features from both disease and drug information. Accordingly, a critical aspect of model construction involves determining representative proteins and diseases, as these selections significantly impact the model's performance. To identify the optimal set of representative proteins and evaluate their influence on prediction accuracy, we conducted experiments using different thresholds for the median z-score of the protein affinity models. Specifically, we examined three thresholds of median z-score $\geq 3, \geq 4$, and ≥ 5, analyzing their respective performances on the direct evidence dataset and the therapeutic dataset.

The results indicate that there are no substantial differences in model performance among the thresholds of 3, 4, and 5 for both datasets (Table S3). This similarity in performance suggests that the model is robust to a range of z-score thresholds, indicating that the representative proteins selected at these thresholds capture the relevant biological information effectively. However, a threshold of 4 exhibited slightly superior performance compared to thresholds 3 and 5. Notably, when the threshold was set to 4, the variations in evaluation metrics were minimized, resulting in the most stable outcomes. This stability implies that a median z-score threshold of 4 provides a balanced selection of representative proteins instead of outliers, optimizing the trade-off between including relevant proteins and excluding noise. Consequently, we selected a median z-score greater than or equal to 4 as the threshold for choosing representative proteins. This threshold was used to compute the affinities between drugs and proteins, which served as the drug input features (I_{DR}) for the model. By excluding proteins with limited affinity data (< 15 drugs), our strategy effectively eliminates proteins that are less frequently accessed and may lack relevance to drug-disease associations.

Furthermore, to ensure that the 228 representative proteins in the affinity matrix effectively capture disease

relevance, we conducted an analysis using gene-disease associations from the CTD. Among 1,893 therapeutic associations (572 diseases, 755 genes), we found that 69 of these representative proteins were targets across 155 diseases. Additionally, in 34,047 direct evidence associations (5,853 diseases, 9,098 genes), 212 proteins were targets for 800 diseases. This overlap confirms that our selected proteins are significantly associated with disease targets, including therapeutically relevant ones, thereby enhancing the biological relevance and interpretability of HEDDI-Net's predictions. These findings validate that the selected proteins effectively represent a wide array of diseases, supporting HEDDI-Net's strength in accurately inferring drug-disease associations.

Moreover, to address potential uncertainties in the layered predictions used to construct the affinity matrix, we implemented several rigorous measures. First, we trained the binding affinity prediction models on experimentally validated protein-drug interaction data from BindingDB and selected representative proteins with robust binding data. Additionally, we conducted 100 hold-out validation tests, where each protein's data was split into training and testing sets, to assess prediction stability. We also applied a stringent z-score filtering criterion to identify and retain only high-confidence proteins. For each protein, a median z-score was calculated from 1,000 permutation tests, measuring prediction accuracy against random chance. Proteins with a median z-score ≥ 4, indicating statistically significant prediction performance, were selected for the final matrix. This z-score threshold helped systematically exclude proteins with low or inconsistent accuracy, minimizing noise in downstream predictions. Together, repeated hold-out validation and z-score filtering ensure HEDDI-Net's affinity matrix remains robust and accurate, addressing layered prediction concerns and supporting the reliability of the model's outputs.

Shifting the focus to representative diseases, we recognize that certain rare disorders may have limited semantic connections with other diseases due to their unique characteristics and low connectivity within the MeSH DAG tree. Of the 2997 diseases considered, 59 had no semantic similarity with any of the 277 representative diseases. This subset primarily comprised conditions related to drug or substance abuse, alcoholism, or animal diseases. These conditions often face limitations regarding direct drug treatment and are usually addressed through medical management approaches. This further underscores that our strategy for including representative diseases can exclude non-targeted diseases, ensuring that the model focuses on relevant and targetable conditions.

We also recognized that using MeSH similarity alone for disease analysis has limitations, particularly for representing complex diseases with overlapping symptoms or unique pathophysiology. Despite these challenges, MeSH remains valuable for structuring disease relationships due to its well-established, hierarchical organization, widely used across biomedical databases, including the Comparative Toxicogenomics Database (CTD). Furthermore, to address potential gaps in MeSH similarity, we selected representative diseases based on eigenvector centrality within the MeSH network. This approach prioritizes diseases with high connectivity and broad representational value, enhancing the robustness of similarity profiles by focusing on diseases with strong relational ties and reducing the impact of underrepresented diseases. It also helps account for the complexity of diseases with overlapping symptoms through biomedical hierarchy relationships, thereby supporting HEDDI-Net's reliability.

Additionally, to investigate the effect of different classifiers on drug-disease association prediction, we replaced the deep learning-based (DL) association probability prediction model with other classifiers, including logistic regression (LR), linear support vector classification (SVC), SVC with radial basis function (RBF) kernel, random forest (RF), and extreme gradient boosting (XGB) [54]. The results, as presented in Table S4, demonstrated that the DL model consistently outperformed all other classifiers across all evaluation metrics for both direct evidence and therapeutic datasets. Notably, XGB exhibited the second-best performance, while SVC with RBF kernel yielded similar results to RF on average. Conversely, LR and linear SVC performed the poorest among the evaluated classifiers. These findings indicate that relying solely on the linear relationship between drug and disease features is insufficient to accurately distinguish drug-disease associations. This suggests that capturing complex, nonlinear relationships is crucial for improving prediction accuracy.

Furthermore, we conducted the same analysis on the benchmark dataset. As shown in Table S4, classifiers such as XGB, RF, and SVC with RBF kernels, which are capable of capturing nonlinear relationships, exhibited superior performance. The performance gap among these classifiers and the linear models also increased substantially with decreasing data size. In contrast, the DL architecture maintained a relatively stable and excellent predictive ability across different dataset sizes. These results highlight the generality and robustness of DL in our proposed model. The superior performance of the DL model, particularly in handling varying dataset sizes, underscores its capability to effectively capture the intricate patterns and relationships inherent in drug-disease associations. This robustness makes the DL approach

particularly well-suited for applications in drug repurposing and predictive modeling in biomedical research.

However, the DL models often function as "black-box" models, posing challenges in feature extraction and interpretation. This lack of transparency makes it difficult to explain physiological mechanisms and hinders their application in advanced research. In contrast, our method introduces representative proteins and diseases, allowing predicted outcomes to be aligned on a common scale and facilitating easily interpretable associations. This feature enhances the model's practical usability across various research and application domains. In order to further substantiate the interpretability of our model, we conducted a case study focusing on the relationship between Alzheimer's disease (AD) and the candidate drugs identified by our model. By incorporating representative biological entities, our approach provides clear insights into the underlying mechanisms of drug-disease interactions. This not only aids in explaining the physiological basis of predictions but also supports advanced research applications by offering a transparent framework for hypothesis generation and validation. The case study on AD underscores the model's capability to identify potential therapeutic candidates, highlighting its value in drug repurposing and biomedical research.

Besides, a key advantage of HEDDI-Net is its ability to incorporate new drugs or diseases without re-establishing or retraining. This flexibility stems from predefined representative embeddings for drugs and diseases, creating stable feature spaces that allow seamless integration of novel entities. HEDDI-Net's feature space is derived from binding affinity profiles with 228 representative proteins for drugs and semantic similarity profiles with 277 representative diseases for diseases, capturing essential interactions. When introducing a new drug or disease, we only need to calculate its embeddings using interaction profiles with these representative proteins or diseases. Specifically, MACCS fingerprints and physicochemical properties predict binding affinities for new drugs, while MeSH-based semantic similarity profiles locate new diseases within the established feature space. This approach enables HEDDI-Net to accommodate new entities without modifying model architecture or parameters. Unlike traditional GCNs that require retraining with new nodes, HEDDI-Net decouples drug and disease embeddings from the core association model, ensuring efficient and scalable integration of novel entities. This adaptability underscores HEDDI-Net's utility for real-world applications in drug repurposing and disease exploration.

Conclusions

Taken together, we propose HEDDI-Net as a stable and applicable resource for pinpointing potential targets in drug repurposing. Beyond this, our approach provides invaluable biological insights that can be leveraged for further investigation. The model's inherent adaptability and seamless integration capabilities position it as an ideal candidate for tackling emerging diseases or pioneering drug discoveries in future endeavors. In summary, HEDDI-Net demonstrates a strong combination of efficacy, resilience, and applicability, solidifying its role as an indispensable tool in expediting the intricate processes of drug discovery and development. Its unique ability to deliver both high predictive performance and interpretability ensures its continued relevance and utility in the rapidly evolving landscape of biomedical research.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05938-6.

Supplementary material 1.

Acknowledgements
We would like to express our gratitude to Dr. Yu-Chao Wang and Cho-Yi Chen at National Yang Ming Chiao Tung University, and Dr. Hsueh-Fen Juan at National Taiwan University for providing us with valuable comments that have helped improve this work.

Author contributions
Yin-Yuan Su: Conceptualization, Methodology, Investigation, Visualization, Writing—original draft, Writing—review & editing. Hsuan-Cheng Huang: Funding acquisition, Project administration, Supervision. Yu-Ting Lin: Methodology. Yi-Fang Chuang: Investigation, Writing—review & editing. Sheh-Yi Sheu: Supervision. Chen-Ching Lin: Conceptualization, Methodology, Investigation, Visualization, Funding acquisition, Project administration, Supervision, Writing—review & editing.

Funding
This work was supported by National Science and Technology Council in Taiwan (NSTC 109–2221-E-010 -014 -MY3 and NSTC 112–2221-E-A49 -106 -MY3) and Ministry of Health and Welfare in Taiwan (MOHW112-TDU-B-222–124013 and MOHW111-TDU-B-221–114007).

Availability of data and materials
All datasets used for training, testing, and validating the HEDDI-Net were derived from sources in the public domain and listed in this article. The source code of HEDDI-Net will be shared on request to the corresponding author for review process. After published, the source code of HEDDI-Net will be deposited to GitHub and freely downloaded.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Biomedical Informatics, National Yang Ming Chiao Tung University, Taipei, Taiwan. [2]School of Medicine, National Yang Ming Chiao Tung University, Taipei, Taiwan. [3]Institute of Public Health, National Yang Ming Chiao Tung University, Taipei, Taiwan. [4]Department of Psychiatry, Far Eastern Memorial Hospital, New Taipei, Taiwan. [5]Department of Life Science and Institute of Genome Science, National Yang-Ming University, Taipei, Taiwan.

Received: 1 August 2024 Accepted: 3 December 2024
Published online: 01 February 2025

References

1. Pavlopoulos GA, Kontou PI, Pavlopoulou A, Bouyioukos C, Markou E, Bagos PG. Bipartite graphs in systems biology and medicine: a survey of methods and applications. Gigascience. 2018;7(4):1–31.
2. Clark C, Rabl M, Dayon L, Popp J. The promise of multi-omics approaches to discover biological alterations with clinical relevance in Alzheimer's disease. Front Aging Neurosci. 2022;14:1065904.
3. Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S. Advancing drug discovery via artificial intelligence. Trends Pharmacol Sci. 2019;40(8):592–604.
4. Emmert-Streib F, Tripathi S, Simoes RdM, Hawwa AF, Dehmer M. The human disease network. Syst Biomed. 2013;1(1):20–8.
5. Tamimi NA, Ellis P. Drug development: from concept to marketing! Nephron Clin Pract. 2009;113(3):c125-131.
6. Rao N, Poojari T, Poojary C, Sande R, Sawant S. Drug repurposing: a shortcut to new biological entities. Pharm Chem J. 2022;56(9):1203–14.
7. Wu Z, Wang Y, Chen L. Network-based drug repositioning. Mol Biosyst. 2013;9(6):1268–81.
8. Luo H, Li M, Yang M, Wu FX, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform. 2021;22(2):1604–19.
9. Kim Y, Jung YS, Park JH, Kim SJ, Cho YR. Drug-disease association prediction using heterogeneous networks for computational drug repositioning. Biomolecules. 2022;12(10):1497.
10. Jung YS, Kim Y, Cho YR. Comparative analysis of network-based approaches and machine learning algorithms for predicting drug-target interactions. Methods. 2022;198:19–31.
11. Zhou Y, Wang F, Tang J, Nussinov R, Cheng F. Artificial intelligence in COVID-19 drug repurposing. Lancet Digit Health. 2020;2(12):e667–76.
12. Zhang H, Cui H, Zhang T, Cao Y, Xuan P. Learning multi-scale heterogenous network topologies and various pairwise attributes for drug-disease association prediction. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbac009.
13. Zhong T, Li Z, You ZH, Nie R, Zhao H. Predicting miRNA-disease associations based on graph random propagation network and attention network. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab589.
14. Yu Z, Huang F, Zhao X, Xiao W, Zhang W. Predicting drug-disease associations through layer attention graph convolutional network. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbaa243.
15. Cai L, Lu C, Xu J, Meng Y, Wang P, Fu X, Zeng X, Su Y. Drug repositioning based on the heterogeneous information fusion graph convolutional network. Brief Bioinform. 2021. https://doi.org/10.1093/bib/bbab319.
16. Gao L, Cui H, Zhang T, Sheng N, Xuan P. Prediction of drug-disease associations by integrating common topologies of heterogeneous networks and specific topologies of subnets. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab467.
17. Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative toxicogenomics database (CTD): update 2023. Nucleic Acids Res. 2023;51(D1):D1257–62.
18. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016;44(D1):D1045-1053.
19. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK: BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. Nucleic Acids Res 2007, 35(Database issue):D198–201.
20. Chen X, Lin Y, Gilson MK. The binding database: overview and user's guide. Biopolymers. 2001;61(2):127–41.
21. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006;5(12):993–6.
22. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol. 2008;4(11):682–90.
23. UniProt C. UniProt: the Universal protein knowledgebase in 2023. Nucleic Acids Res. 2023;51(D1):D523–31.
24. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373–80.
25. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier. J Cheminform. 2015;7:23.
26. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. J Chem Inf Modeling. 1989;29(2):97–101.
27. Landrum G: RDKit: open-source cheminformatics software. 2016.
28. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrian-Uhalte E, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017;45(D1):D945–54.
29. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B *et al*: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012, 40(Database issue):D1100-1107.
30. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15(1):72–101.
31. z-Score. In: Encyclopedia of Public Health. Edited by Kirch W. Dordrecht: Springer Netherlands; 2008: 1484–1484.
32. Petersen AM, Rotolo D, Leydesdorff L. A triple helix model of medical innovation: supply, demand, and technological capabilities in terms of medical subject headings. Res Policy. 2016;45(3):666–81.
33. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics. 2010;26(13):1644–50.
34. Bonacich P. Factoring and weighting approaches to status scores and clique identification. J Math Sociol. 1972;2(1):113–20.
35. Diallo SY, Lynch CJ, Gore R, Padilla JJ. Identifying key papers within a journal via network centrality measures. Scientometrics. 2016;107(3):1005–20.
36. Murphy KP. Probabilistic machine learning : an introduction. Cambridge: The MIT Press; 2022.
37. Kingma DP, Ba JJC: Adam: A Method for Stochastic Optimization. 2014, abs/1412.6980.
38. Meng Y, Lu C, Jin M, Xu J, Zeng X, Yang J. A weighted bilinear neural collaborative filtering approach for drug repositioning. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab581.
39. Zhao BW, Hu L, You ZH, Wang L, Su XR. HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. Brief Bioinform. 2022. https://doi.org/10.1093/bib/bbab515.
40. Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, Liu F. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics. 2018;19(1):233.
41. Zhang H, Wei W, Zhao M, Ma L, Jiang X, Pei H, Cao Y, Li H. Interaction between Abeta and Tau in the pathogenesis of Alzheimer's disease. Int J Biol Sci. 2021;17(9):2181–92.
42. Park JY, Park J, Baek J, Chang JW, Kim YG, Chang WS. Long-term results on the suppression of secondary brain injury by early administered low-dose baclofen in a traumatic brain injury mouse model. Sci Rep. 2023;13(1):18563.
43. Pilipenko V, Narbute K, Beitnere U, Rumaks J, Pupure J, Jansone B, Klusa V. Very low doses of muscimol and baclofen ameliorate cognitive deficits and regulate protein expression in the brain of a rat model of streptozocin-induced Alzheimer's disease. Eur J Pharmacol. 2018;818:381–99.
44. Chang KA, Kim JA, Kim S, Joo Y, Shin KY, Kim S, Kim HS, Suh YH. Therapeutic potentials of neural stem cells treated with fluoxetine in Alzheimer's disease. Neurochem Int. 2012;61(6):885–91.
45. Zhou CN, Chao FL, Zhang Y, Jiang L, Zhang L, Fan JH, Wu YX, Dou XY, Tang Y. Fluoxetine delays the cognitive function decline and synaptic changes in a transgenic mouse model of early Alzheimer's disease. J Comp Neurol. 2019;527(8):1378–87.
46. Heneka MT, Carson MJ, El Khoury J, Landreth GE, Brosseron F, Feinstein DL, Jacobs AH, Wyss-Coray T, Vitorica J, Ransohoff RM,

et al. Neuroinflammation in Alzheimer's disease. Lancet Neurol. 2015;14(4):388–405.
47. Ivanova M, Belaya I, Kucharikova N, de Sousa MI, Saveleva L, Alatalo A, Juvonen I, Thind N, Andres C, Lampinen R, et al. Upregulation of Integrin beta-3 in astrocytes upon Alzheimer's disease progression in the 5xFAD mouse model. Neurobiol Dis. 2024;191: 106410.
48. Rehman R, Miller M, Krishnamurthy SS, Kjell J, Elsayed L, Hauck SM, Olde Heuvel F, Conquest A, Chandrasekar A, Ludolph A, et al. Met/HGFR triggers detrimental reactive microglia in TBI. Cell Rep. 2022;41(13): 111867.
49. Elseweidy MM, Mahrous M, Ali SI, Shaheen MA, Younis NN. Pentoxifylline as add-on treatment to donepezil in copper sulphate-induced Alzheimer's disease-like neurodegeneration in rats. Neurotox Res. 2023;41(6):546–58.
50. Black RS, Barclay LL, Nolan KA, Thaler HT, Hardiman ST, Blass JP. Pentoxifylline in cerebrovascular dementia. J Am Geriatr Soc. 1992;40(3):237–44.
51. Dhikav V. Can phenytoin prevent Alzheimer's disease? Med Hypotheses. 2006;67(4):725–8.
52. Kurbatskaya K, Phillips EC, Croft CL, Dentoni G, Hughes MM, Wade MA, Al-Sarraj S, Troakes C, O'Neill MJ, Perez-Nievas BG, et al. Upregulation of calpain activity precedes tau phosphorylation and loss of synaptic proteins in Alzheimer's disease brain. Acta Neuropathol Commun. 2016;4:34.
53. Chen HH, Liu P, Auger P, Lee SH, Adolfsson O, Rey-Bellet L, Lafrance-Vanasse J, Friedman BA, Pihlgren M, Muhs A, et al. Calpain-mediated tau fragmentation is altered in Alzheimer's disease progression. Sci Rep. 2018;8(1):16725.
54. Chen T, Guestrin C: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: association for computing machinery; 2016: 785–794.

Publisher's Note