Proceedings of CONF-MLA Workshop: Mastering the Art of GANs: Unleashing Creativity with Generative Adversarial Networks
DOI: 10.54254/2755-2721/80/2024CH0052

Application of geometric deep learning in disease-gene relationship prediction

Ruirui Tan

Nanjing University, No. 1520, Taihu Avenue, Suzhou, China

1204446770@qq.com

Abstract. The correlation between diseases and genetic factors represents a pivotal challenge in the biomedical field, often conceptualized as a task of link prediction. Conventional methods employed for prediction, such as statistical models and various machine learning algorithms, frequently fall short in terms of accuracy when confronted with the intricacies of biological network data. These methods also tend to inadequately represent the complex relationships inherent in such data. In contrast, recent advances in geometric deep learning have introduced a powerful tool within artificial intelligence, particularly adept at processing non-Euclidean data structures. This study delves into the potential of leveraging geometric deep learning techniques to enhance the prediction of disease-gene associations. Initially, we conduct a thorough review of existing research related to link prediction, encompassing both traditional approaches and contemporary methods grounded in deep learning. Subsequently, we propose a geometric deep learning framework, incorporating Graph Convolutional Networks (GCN) and Graph Auto-Encoder (GAE), to develop and assess our predictive model. The results of our experiments demonstrate that the proposed geometric deep learning model surpasses conventional techniques in accurately predicting disease-gene associations. In conclusion, we evaluate the implications of our findings, discuss their practical applications in the biomedical domain, and suggest possible avenues for future research.

Keywords: Geometric Deep Learning, Link Prediction, Disease-Gene Relationship, Graph Convolutional Networks, Graph Auto-Encoder.

1. Introduction

The intricate relationship between genes and diseases forms a crucial foundation for comprehending the complexities of human health and the origins of various medical conditions. Identifying specific genes linked to particular diseases offers deep insights into the biological processes that drive these disorders. Understanding these relationships is crucial for enhancing diagnostic techniques and fostering the creation of more accurate and targeted treatments. This advancement is particularly significant in the realm of personalized medicine, where therapies are tailored according to an individual's unique genetic makeup. With the rapid progression of genomics and bioinformatics, the exploration of disease-gene associations has become increasingly vital in biomedical research. Nevertheless, the inherent complexity of biological systems continues to pose significant challenges in fully unraveling these relationships, underscoring the need for ongoing research and innovation in this area.

DOI: 10.54254/2755-2721/80/2024CH0052

Understanding the relationships between diseases and genes can be framed as a link prediction problem within a network analysis context. In this context, nodes in the network symbolize either diseases or genes, while the edges between them represent established associations. Common approaches include logistic regression, which models the probability of a link between nodes; support vector machines, which classify data into distinct categories; and random forests, which aggregate predictions from multiple decision trees to improve accuracy. Although these methods are effective in certain scenarios, they exhibit notable limitations. Firstly, many traditional techniques assume Euclidean space, which can be problematic when applied to the non-Euclidean structures prevalent in biological networks. The intricate topological relationships within these networks pose challenges for conventional models, which often fail to accurately capture and represent these complexities, leading to reduced performance, particularly when high accuracy is crucial. Secondly, these methods frequently rely on manually crafted feature extraction, which may not fully capture the richness and variability of biological data. In the realm of disease-gene predictions, effective feature extraction is essential, yet manual processes can overlook critical information or introduce biases. Moreover, biological networks encompass diverse types of data, including gene expression profiles, protein interactions, and genetic variations, which are often heterogeneous. Traditional models struggle to integrate and leverage these multifaceted data sources, further impairing their predictive accuracy. Additionally, the scalability and computational efficiency of traditional methods present significant challenges. As biological datasets expand, the computational demands and memory usage of these models increase substantially, complicating their application to large-scale networks. Traditional methods also lack adaptability, requiring model retraining as new data emerges or network structures evolve, which can be cumbersome in practical applications. Finally, traditional approaches often fall short in addressing the multi-scale and hierarchical nature of biological networks. These networks can span various levels, from molecular interactions to broader cellular processes, with each level potentially influencing disease mechanisms differently. Conventional methods often struggle to handle these multi-level data and perform inadequately in predicting relationships across different scales of biological organization.

In recent times, geometric deep learning has gained prominence as an advanced methodology for the analysis of data structures that are non-Euclidean, such as graphs. Unlike conventional graph-based techniques, geometric deep learning is adept at handling data that exists in irregular or complex domains, including networks and manifolds. This characteristic renders it particularly effective for tasks like predicting relationships between diseases and genes, as it is capable of capturing the complex, multi-dimensional nature of biological networks.

This study investigates the potential of employing geometric deep learning techniques for the prediction of disease-gene relationships. Initially, we conduct a thorough review of existing literature on link prediction methodologies, encompassing both traditional approaches and modern deep learning techniques. Following this, we present the dataset utilized for our experiments and detail the preprocessing procedures undertaken to prepare the data. Subsequently, we develop and implement geometric deep learning models, specifically focusing on Graph Convolutional Networks (GCNs) and Graph Autoencoders, and assess their effectiveness through a rigorous process of training and evaluation. In conclusion, we analyze the outcomes of our research, consider their implications for advancements in the biomedical domain, and propose directions for future studies.

2. Related Work

Link prediction is a crucial task aimed at identifying potential links within a network, with applications spanning social and biological networks. This task is approached through various methodologies, primarily categorized into statistical model-based and machine learning-based techniques.

Statistical model-based methods focus on developing statistical models that are pivotal in enhancing classification accuracy. These models estimate the probability of a connection between two nodes using a range of statistical metrics. A prominent method in this field is the Katz index, developed by Leo Katz in 1953. The Katz index employs a path-based approach to gauge the strength of potential links within a network. It achieves this by aggregating the weighted contributions of all possible paths that link two

DOI: 10.54254/2755-2721/80/2024CH0052

nodes, providing a comprehensive measure of their connectivity. The Katz index has proven particularly valuable in social network analysis, where it effectively captures complex network structures and relational dynamics [1]. In another influential study on link prediction within social networks, David Liben-Nowell, in 2004, examined several statistical features, including the number of common neighbors, the Jaccard coefficient, and the Adamic/Adar index. These indicators offer various perspectives on node relationships, enhancing the accuracy and depth of link prediction analyses. These indicators were utilized to forecast potential links by analyzing the network's topology. The findings indicated that more sophisticated proximity measures, including the Katz index, showed superior performance in predicting links compared to simpler methods. In particular, the Katz index demonstrated a prediction accuracy that was up to 50 times greater than random predictions, reflecting its ability to uncover valuable information embedded within the network's structure [2]. Another influential study in this area was conducted by Tao Zhou and colleagues in 2009. They introduced straightforward statistical models that utilize local features of nodes, including the count of shared neighbors, the lengths of the shortest paths between nodes, and the resource allocation index. Among these, the resource allocation index has proven particularly effective, achieving an Area Under the ROC Curve (AUC) score surpassing 0.9 in various real-world network datasets. This high AUC value highlights the index's strong performance in accurately predicting links within large and intricate networks, demonstrating its utility in practical applications [3].

Beyond statistical model-based approaches, traditional machine learning techniques for link prediction also emphasize the extraction of key features through manual selection. In 2006, Mohammad Al Hasan and colleagues conducted a study where they manually identified features such as Keyword Match, Sum of Neighbors Count, Sum of Papers Count, and the shortest path length for nodes within the network. They utilized these features to perform link prediction using several models, including Support Vector Machines (SVM) and k-Nearest Neighbors (KNN). Their methods demonstrated significant success, achieving a high prediction accuracy of 90.87% on the BIOBASE dataset and 83.18% on the DBLP dataset [4]. In 2010, Ryan N. Lichtenwalter and his team advanced the field by addressing several limitations of earlier supervised learning approaches. They focused on improving aspects such as the duration of network observations, the generalizability of existing models, variance reduction, the influence of topological configurations, and the effects of imbalance and sampling techniques. Their work led to the identification of new features that significantly improved prediction accuracy. They introduced an innovative flow-based prediction algorithm and a novel evaluation method, resulting in an over 30% enhancement in the Area Under the ROC Curve (AUC) compared to the most effective unsupervised methods tested [5]. Following this, in 2011, William Cukierski and his research team developed a link prediction approach using Random Forests, specifically applied to the Flickr social network dataset. They selected 94 distinct graph-based features, including Common Neighbors, Adar, and Jaccard coefficients, to serve as inputs for their classification models. By integrating a comprehensive set of features to capture various aspects of the network structure, they achieved an impressive AUC of 0.9695. This accomplishment earned them second place in the IJCNN 2011 Social Network Challenge and marked the highest performance for models applied to non-de-anonymized datasets [6].

Conventional link prediction techniques typically rely on network structural attributes, including the degree of individual nodes and the degree of similarity between neighboring nodes. These approaches, while foundational, encounter several challenges. Notably, the process of manually extracting meaningful features is both complex and labor-intensive, often resulting in models that are limited in depth. Additionally, these traditional techniques frequently struggle with accurately modeling intricate relationships within non-Euclidean data, which can lead to reduced performance when applied to complex biological networks.

In contrast, recent advancements have seen a shift towards deep learning-based approaches for link prediction. These modern methods typically involve a two-phase process: first, leveraging models such as Graph Neural Networks (GNNs) to generate low-dimensional embeddings for nodes within the network; second, using these embeddings as input for a classifier to predict potential links. This dual-

DOI: 10.54254/2755-2721/80/2024CH0052

step methodology allows for a more nuanced representation of network structures, effectively capturing complex relational features and improving performance across various benchmark datasets.

Geometric deep learning is a rapidly advancing area within artificial intelligence that excels in managing non-Euclidean data, such as graphs and manifolds. In the biomedical domain, geometric deep learning has shown considerable promise in tasks like predicting protein structures and designing drug molecules. Recent studies have produced noteworthy advancements in this field. For example, in 2022, Kanchan Jha and colleagues introduced a method utilizing Graph Neural Networks (GNNs) to forecast protein-protein interactions (PPI). Their research investigated how different node features influence model performance, employing GCN (Graph Convolutional Networks) and GAT (Graph Attention Networks) models alongside various feature extraction techniques, including embeddings from LSTM-based language models and BERT [7]. In the same year, Ziduo Yang and his team developed a deep multiscale graph neural network known as MGraphDTA, tailored for predicting drug-target binding affinity (DTA) with explainability. The MGraphDTA model demonstrated exceptional performance on the Human and C. elegans datasets, achieving precision scores of 0.955 and 0.980, and AUC scores of 0.983 and 0.991, respectively, significantly surpassing the accuracy of competing methods [8].

These developments highlight the advantages of geometric deep learning over conventional statistical models and traditional machine learning techniques that rely on manually extracted features. Unlike traditional methods, which require the manual specification of features, geometric deep learning approaches utilize models such as GCN and GAT to automatically learn and identify relevant features for classification tasks. This capability allows deep learning-based node encoding methods to effectively capture the complex characteristics of non-Euclidean data, leading to superior predictive performance compared to traditional approaches.

The application of geometric deep learning to the prediction of disease-gene relationships represents an innovative research avenue. By accurately modeling the intricate interactions between diseases and genes, geometric deep learning techniques hold the potential to enhance prediction accuracy and offer valuable insights for advancing biomedical research.

3. Dataset and Preprocessing

The dataset employed in this research is sourced from the Stanford Biomedical Network Dataset Collection, specifically the BioSNAP dataset [9-12]. This extensive dataset includes a variety of biomedical networks, such as interactions between proteins and between drugs and their targets, which are crucial for examining and predicting complex relationships in the field of biomedicine. The dataset features information on 519 distinct diseases and 7,294 genes, along with their various interconnections. These relationships are structured as a graph, where nodes represent diseases and genes, and the edges signify the associations between them.

Given that link prediction tasks often necessitate extensive network connection data for effective training, the challenge of acquiring and processing large-scale datasets can lead to considerable computational demands and resource consumption. To address these challenges while exploring link prediction techniques, we opted for a smaller-scale dataset for our experimental analysis. Using a more manageable dataset for initial experiments allows for enhanced research efficiency, easier model tuning based on preliminary outcomes, and sets a robust groundwork for further in-depth studies with larger datasets.

In our study, we undertook several preprocessing steps on the original dataset, including mapping nodes and edges. Initially, we extracted the node column from the dataset and assigned a unique integer index to each node, starting from a predefined offset. This resulted in a mapping dictionary linking node names to integer indices. Simultaneously, we created a reverse mapping to associate integer indices with the original node IDs, forming a dictionary that maps integer indices back to node names. We then processed the source and target node columns from the dataset, converting the original node names into their corresponding integer indices using the previously established mapping. This conversion yielded a tensor-form edge index list. To complete the process, we incorporated reverse edge indices and combined both index lists to construct an undirected graph edge index. To streamline data processing,

DOI: 10.54254/2755-2721/80/2024CH0052

we developed a Data object containing the number of nodes, edge indices, and node features, which were all initialized to ones. Additionally, we calculated the degree of each node and sorted them in descending order. These preprocessing steps are designed to facilitate subsequent data analysis and usage, ensuring a well-structured foundation for further investigation.

4. Model Design

We developed a geometric deep learning framework utilizing Graph Convolutional Networks (GCNs). The initial step involved encoding the nodes using the Node2Vec technique, which provided an initial node feature matrix, denoted as A. This matrix A was then input into a GCN-based encoder, which transformed it into a lower-dimensional feature matrix Z. Subsequently, matrix Z was processed by a decoder to generate a reconstructed node feature matrix, A'. The similarity between the original matrix A and the reconstructed matrix A' was computed to refine and optimize the encoder model. During the testing phase of the Graph Autoencoder (GAE) model, the features of pairs of nodes were concatenated and used for binary classification to ascertain the presence of an undirected edge between them. This procedure was essential for assessing the GAE model's performance in link prediction tasks.

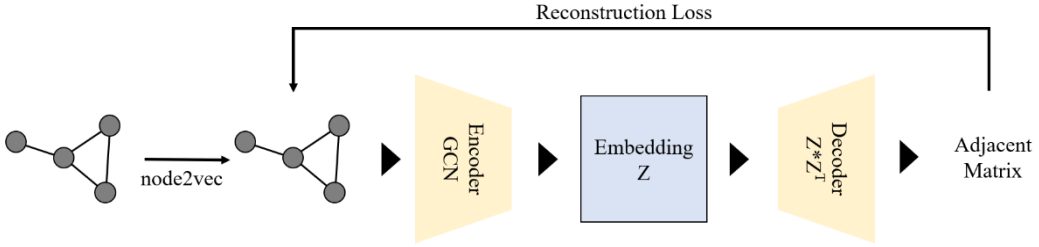

Figure 1. Model Architecture.

For the hyperparameter configuration of our model, we set the feature dimension to 100, the hidden layer size of the GCN encoder to 200, and the output channel count to 50. The model was trained for 100 epochs, with dropout techniques applied to mitigate overfitting. To evaluate the model's efficacy, we employed several metrics, including Average Precision, the convergence rate of Training Loss, and the Area Under the Curve (AUC). These metrics were used to validate the model's performance in predicting disease-gene relationships and ensure its effectiveness.

5. Result

We evaluated the GCN-GAE model's performance using a test dataset, and the results from our experiments were highly promising. The model achieved a notable Area Under the Curve (AUC) score of 0.9586 and an Average Precision (AP) score of 0.9449, while the loss value was recorded at 1.1140. These metrics highlight the GCN-GAE model's effectiveness in addressing the link prediction task. The high AUC and AP scores indicate the model's strong ability to accurately predict links within the network, demonstrating its overall robustness and reliability in handling the link prediction challenge.

DOI: 10.54254/2755-2721/80/2024CH0052

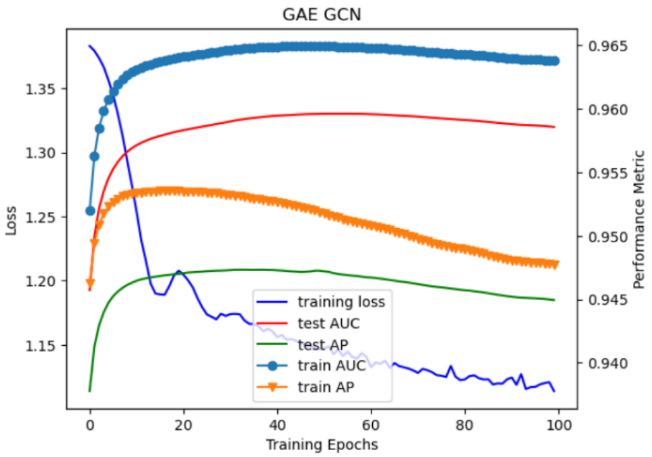

Figure 2. Training Dynamics of the GAE-GCN Model.

This figure provides a detailed view of the performance metrics for both the GAE (Graph Autoencoder) and GCN (Graph Convolutional Network) models during their training phases. The horizontal axis of the graph tracks the number of training epochs, while the left vertical axis denotes the "Loss" values, and the right vertical axis displays the "Performance Metric" values. Different colored curves track training loss, AUC, and Average Precision (AP) on the test set. The blue curve representing training loss exhibits a rapid decrease in the initial epochs, suggesting effective optimization. Simultaneously, the curves representing the AUC (Area Under the Curve) and AP (Average Precision) metrics exhibit a steady increase and ultimately reach a stable plateau at higher values. This trend reflects the model's improving accuracy and consistent performance over time, showcasing its reliability in making precise predictions.

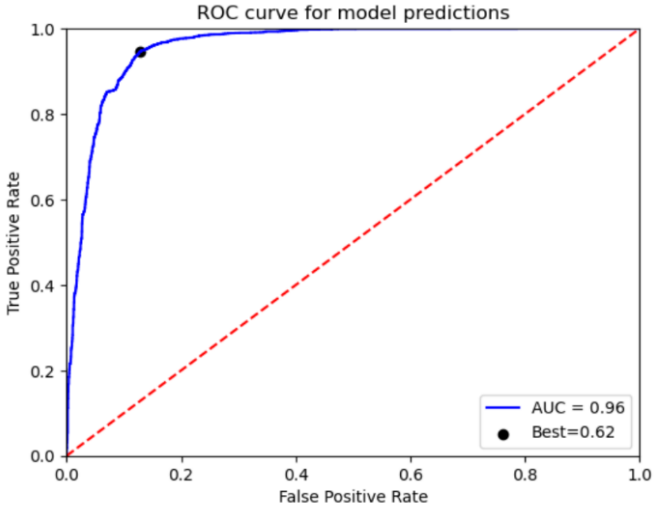

Figure 3. AUC.

This figure presents the Receiver Operating Characteristic (ROC) curve used to evaluate the classification model's performance. The x-axis denotes the "False Positive Rate," while the y-axis indicates the "True Positive Rate." The blue curve represents the model's actual performance, with an AUC of 0.96, highlighting its strong discriminatory ability. The red dashed line marks the baseline for random guessing. The point on the curve denotes the optimal classification threshold, with an accuracy of 0.62, suggesting that this threshold offers the best performance for the model.

DOI: 10.54254/2755-2721/80/2024CH0052

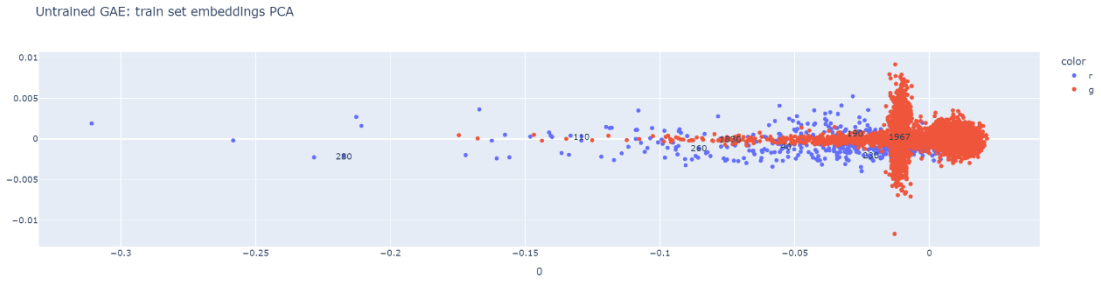

Figure 4. Untrained GAE.

This figure depicts the embeddings generated by the GAE model before training. The axes represent two principal components of the embedding vectors. The visualization reveals that the embeddings from the untrained model show significant overlap among points from different classes, making it challenging to distinguish between them. This overlap indicates that the untrained model has not yet learned to capture the underlying structure of the data effectively.

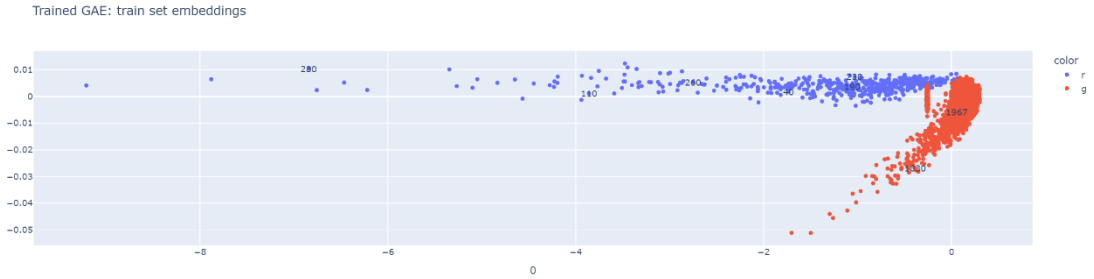

Figure 5. Trained GAE.

The visualization shown depicts the embeddings produced by the GAE model after it has been fully trained, applied to the training set. In this visualization, different classes are represented by distinct colors, and the axes correspond to two principal components of the embedding vectors. The image illustrates that the trained GAE model effectively differentiates between various classes, reflecting its strong classification capabilities within the embedding space. This separation indicates that the model successfully captures the underlying data structure, distinguishing it from other classes.

The comparative analysis of these images clearly highlights the effectiveness of the trained GAE model in data classification and structural representation. The trained model outperforms the untrained version, which struggles to separate classes, underscoring the progress achieved through training.

The exceptional performance of the geometric deep learning model can be attributed to its advanced capability in capturing the intricate topological and structural characteristics of the disease-gene relationship network. Unlike traditional methods, the GCN-GAE model combines graph convolution with graph autoencoders, thereby enhancing the interpretability of the encoded features and learning more nuanced and distinctive representations. This advanced approach leads to significantly improved prediction accuracy compared to conventional techniques.

6. Review and Implications

This research explores the use of geometric deep learning methodologies to predict and analyze relationships between diseases and genes. We designed and implemented a geometric deep learning model leveraging Graph Convolutional Networks (GCNs) and tested its performance using publicly available datasets. The experimental outcomes reveal that our model surpasses conventional link prediction approaches in terms of accuracy, convergence speed, and AUC scores, demonstrating its superior predictive capabilities.

DOI: 10.54254/2755-2721/80/2024CH0052

These results highlight the potential for advancing artificial intelligence applications within the biomedical domain. Our findings open up exciting avenues for further research, including the exploration of geometric deep learning in additional biomedical contexts, such as predicting drug targets and analyzing disease mechanisms. Moreover, we aim to explore the integration of geometric deep learning with other technologies, such as natural language processing, to drive further innovations in interdisciplinary research.

References

[1] Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika, 18(1), 39-43.

[2] Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. In Proceedings of the twelfth international conference on Information and knowledge management (pp. 556-559).

[3] Zhou, T., Lü, L., & Zhang, Y. C. (2009). Predicting missing links via local information. The European Physical Journal B, 71, 623-630.

[4] Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. J. (2006). Link prediction using supervised learning. In SDM06: Workshop on link analysis, counter-terrorism and security (pp. 798-805).

[5] Lichtenwalter, R. N., Lussier, J. T., & Chawla, N. V. (2010). New perspectives and methods in link prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 243-252).

[6] Cukierski, W., Hamner, B., & Yang, B. (2011). Graph-based features for supervised link prediction. In Proceedings of the 2011 International joint conference on neural networks (pp. 1237-1244). IEEE.

[7] Jha, K., Saha, S., & Singh, H. (2022). Prediction of protein–protein interaction using graph neural networks. Scientific Reports, 12(1), 8360.

[8] Yang, Z., Zhong, W., Zhao, L., & Li, J. (2022). MGraphDTA: Deep multiscale graph neural network for explainable drug–target binding affinity prediction. Chemical Science, 13(3), 816-833.

[9] Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., & Hamosh, A. (2014). OMIM.org: Online Mendelian Inheritance in Man, an online catalog of human genes and genetic disorders. Nucleic Acids Research, 43(D1), D789-D798.

[10] Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, T., & Mattingly, C. J. (2017). The Comparative Toxicogenomics Database: update 2017. Nucleic Acids Research, 45(D1), D972-D978.

[11] Piñero, J., Bravo, Á., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., ... & Furlong, L. I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Research, 45(D1), D833-D839.

[12] MINER: Gigascale multimodal biological network. (2017). Stanford SNAP Group. GitHub Repository. Retrieved from https://github.com/snap-stanford/miner