The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Optimizing Drug-Disease Association Analysis: A Resource-Efficient Approach Using Numerical Linear Algebra and Machine Learning

Zahra Rafei* Seyedeh Fatemeh Hosseini† Behnam Yousefimehr ‡ Sajed Tavakkoli§ Mehdi Ghatee¶

Abstract

In today's data-driven world, the growing volume of information demands the development of models that balance accuracy and computational efficiency. Drug repurposing has emerged as a pivotal strategy in the pharmaceutical industry, enabling the identification of new therapeutic uses for existing drugs. However, with the increasing amount of available data, it is essential for researchers and industry stakeholders to create models that maintain predictive accuracy while minimizing computational costs. Our study builds on the state-of-the-art WNMFDDA (Weighted Graph Regularized Collaborative Non-negative Matrix Factorization for Drug-Disease Association Prediction) model, which is known for its high predictive accuracy. We focus on optimizing this approach by significantly reducing memory usage and computational time, achieving an 8-fold reduction in time cost and over a 400-fold decrease in storage cost during model training, all without compromising accuracy. Our findings show that several alternative methods can deliver performance metrics close to the reference model while substantially lowering both memory and computational requirements. This approach not only retains the accuracy of drug-disease association predictions but also enhances the efficiency of the drug repurposing process, enabling quicker transitions from research to clinical applications. By optimizing computational resources, this work provides a scalable and efficient solution for future drug discovery and repurposing efforts.

Keywords: Drug Repurposing, Machine Learning, Numerical Linear Algebra, Drug-Disease Association, Non-negative matrix factorization

*Department of Mathematics and Computer Science, Amirkabir University of Technology, zahra.rafei@aut.ac.ir

†Department of Mathematics and Computer Science, Amirkabir University of Technology, sf.hosseini@aut.ac.ir

‡Department of Mathematics and Computer Science, Amirkabir University of Technology, behnam.y2010@aut.ac.ir

§Department of Mathematics and Computer Science, Amirkabir University of Technology, sajedtavakoli@aut.ac.ir

¶Department of Mathematics and Computer Science, Amirkabir University of Technology, ghatee@aut.ac.ir , Corresponding author

1 Introduction

Drug repurposing, also known as drug repositioning, has become a clever and efficient strategy in the pharmaceutical industry for developing new treatments. Instead of the traditional route of discovering and creating brand-new drugs, this approach takes advantage of medications that are already approved and prescribed for specific diseases. What's fascinating is that these drugs might have untapped potential for treating other conditions we don't fully grasp yet. By leveraging existing clinical data and regulatory approvals, drug repurposing speeds up the research and development process[1]. This means we can significantly cut down on both the financial costs and the time usually required, since these drugs have already gone through initial safety and efficacy evaluations.

This method doesn't just save on production costs; it also opens up new avenues for tackling therapeutic challenges. By uncovering new uses for existing drugs, we minimize the risks tied to side effects and safety issues that often come up when developing new medications[2][3]. Think about drugs like Minoxidil and Sildenafil,they started out for different purposes but eventually found new therapeutic roles, adding significant value to the pharmaceutical industry[2][4][5].

Given the importance of this topic and the rise of advanced computational techniques especially in numerical linear algebra and the latest strides in machine learning, we have a real opportunity to make the drug repurposing process even better. By using machine learning methods, including algorithms based on neural networks, we can find hidden patterns in large clinical datasets and pull out relevant information [6, 7]. This makes drug repurposing more efficient. Specifically, numerical linear algebra computations help us identify drug-disease associations faster and more accurately, and they improve predictions about how effective a drug might be.

With this in mind, we examined a dataset extracted using various machine learning techniques to dig out the necessary information from drugs and diseases. To find a technique that offers solid accuracy at a lower cost compared to some existing studies, we evaluated and compared the WNMFDDA(weighted graph regularized collaborative non-negative matrix factorization

for drug-disease association prediction) [13] approach that is state of the art in this domain with various machine learning algorithms and numerical linear algebra-based methods. These were our main tools for optimizing and processing large datasets. Our aim was to pinpoint and introduce a method that excels in accuracy, speed, and memory efficiency compared to others. By applying these techniques, we can develop more efficient and precise computational methods for drug repurposing, ultimately slashing both research costs and the time needed to achieve positive clinical outcomes.

2 Related Work

When it comes to predicting how drugs relate to diseases, researchers use a variety of methods, each with its own techniques to make the most of the available data. These methods rely on similarities between drugs and diseases to find new connections and improve drug repurposing or redesign efforts.

One example is the SAEROF(Sparse Auto-Encoder-Based Rotation Forest) model [9]. This approach combines a sparse autoencoder with Rotation Forest to pinpoint drug-disease links. It looks at how similar drug structures are and how diseases relate semantically, which helps make more accurate predictions. The downside, though, is that SAEROF is computationally heavy, meaning it takes a lot of processing power and time, especially with large or complex datasets.

Another method is the DDA-SKF(drug–disease associations prediction using similarity kernels fusion) model [10], which uses a similarity kernel fusion technique. By merging different similarity measures for drugs and diseases and applying the Laplacian Regularized Least Squares algorithm, this model performs well even when there's not much data. However, it struggles with accuracy if the similarity information it relies on is lacking or incorrect.

Deep learning approaches are also important in this area. Take the DCNN(Densely Connected Convolutional Networks) model [11], for instance. It uses a Dense Convolutional Neural Network with attention mechanisms to find drug-disease associations by spotting hidden patterns in the data. The challenge here is that DCNN requires a lot of training data and can be complex to set up, making it hard to use effectively with small or imbalanced datasets.

There are also models based on matrix factorization. The SCMFDD(similarity constrained matrix factorization) [12] is one such example, which uses similarity constraints during matrix factorization. It projects drugs and diseases into a lower-dimensional space and predicts their associations based on these similarities. However, SCMFDD doesn't perform as well with new or unseen data, reducing its accuracy in those cases.

More recently, the WNMFDDA [13] method has gained attention. This model combines non-negative matrix factorization with graph regularization to find potential drug-disease links. It starts by calculating similarities based on drug chemical structures and disease information, then uses a weighted K-nearest neighbors approach to rebuild association scores. Finally, it applies matrix factorization and graph regularization to predict new association. While WNMFDDA is accurate and mathematically robust, it needs a lot of data and memory, making it costly to compute. Plus, it's not very easy to interpret, so there's interest in finding methods that keep the accuracy high but use less memory and compute time while being easier to understand [13].

Table 1 provides an overview of some recent work in this domain.

Table 1: Summary of Techniques and Related Works

Ref.	Technique and Description
[9]	**SAEROF:** Sparse Auto-Encoder-Based Rotation Forest combines sparse autoencoders with Rotation Forest to enhance drug-disease predictions using structural and semantic relationships.
[10]	**DDA-SKF:** Drug–Disease Associations using Similarity Kernels Fusion merges similarity measures for drugs and diseases, applying the Laplacian Regularized Least Squares algorithm for association predictions, especially useful with limited data.
[11]	**DCNN:** Densely Connected Convolutional Networks with attention mechanisms identify hidden patterns in data to predict drug-disease associations, though it requires substantial training data.
[12]	**SCMFDD:** Similarity Constrained Matrix Factorization uses similarity constraints in matrix factorization to project drugs and diseases into lower-dimensional spaces for association prediction.
[13]	**WNMFDDA:** Weighted Graph Regularized Collaborative Non-negative Matrix Factorization combines graph regularization with matrix factorization to predict drug-disease associations based on chemical and disease similarities.

Overall, since figuring out drug-disease associations is still a relatively new challenge, all these methods are valuable. While creating new techniques is important, it might be more effective at first to fine-tune the existing ones to make them faster and less resource-heavy without losing accuracy. These different approaches all aim to better predict how drugs and diseases are connected, helping discover new uses for existing drugs.

3 Methodology

In this paper, we aimed to develop a model for accurately predicting drug-disease associations while minimizing computational resources. To achieve this, we compared the Weighted Graph Regularized Collaborative Non-negative Matrix Factorization for Drug-Disease Association Prediction (WNMFDDA), a state-of-the-art method known for its high accuracy, against various alternative approaches, including machine learning algorithms and numerical linear algebra techniques. Our objective was to identify methods that could achieve comparable accuracy with reduced memory usage and processing time. While WNMFDDA offers excellent predictive performance, it is computationally intensive. Thus, we explored other methods that could provide similar accuracy with greater efficiency.

Figure 1 shows flowchart of our methodology.

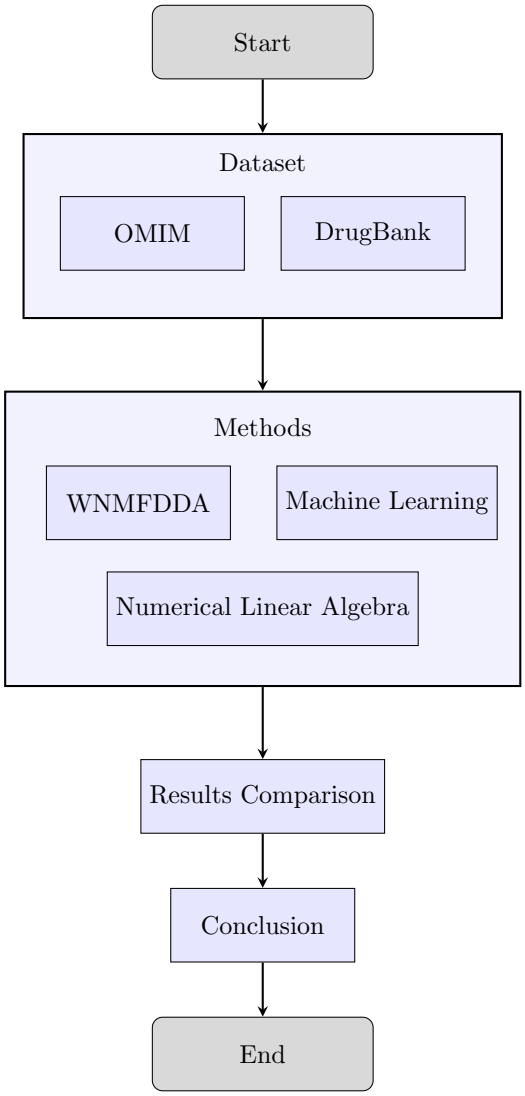

Figure 1: General Workflow of the Study

First, we examine the foundational method upon which the WNMFDDA is based to better understand its general principles. Although the WNMFDDA method consumes more time and memory than traditional NMF, its improved accuracy justifies these costs, especially since the resource usage figures are comparable. Therefore, these factors can be overlooked in pursuit of better model performance.

We conducted a comparative analysis of a framework of methods. Among these, the method that performed most closely to the baseline in terms of evaluation criteria was Decision Tree Regressor, which demonstrated both high accuracy and lower resource consumption compared to the WNMFDDA method. Additionally, Decision Tree Regressor offers superior interpretability over the baseline, a quality that enhances its appeal in interdisciplinary applications. Consequently, the decision tree regressor emerged as the most effective method for this case study.

3.1 Non-negative Matrix Factorization (NMF)

One of the key methods that formed the basis of the reference method studied is Non-negative Matrix Factorization (NMF) [16]. In this method, the drug-disease association matrix V is decomposed into two non-negative matrices W and H:

$$V \approx W \times H \tag{1}$$

where:

- V: Drug-disease association matrix
- W: Latent features of drugs
- H: Latent features of diseases

3.2 DecisionTreeRegressor

The Decision Tree Regressor [17] is a non-parametric regression method that predicts target values by learning simple decision rules inferred from the features. This model recursively partitions the data space into subsets to minimize prediction error, resulting in a tree-like structure of decisions. By using this approach, it can capture non-linear relationships without requiring feature scaling or transformation, making it suitable for various regression tasks. The model works by iteratively splitting the dataset into smaller groups based on specific feature thresholds, resulting in leaf nodes that contain samples with similar target values.

To evaluate the quality of a split, the Decision Tree Regressor uses criteria like reduction in mean squared error (MSE). The split is chosen to minimize the MSE for the resulting child nodes, effectively reducing the impurity of the partitions. This reduction in error at each split is known as the decrease in impurity or information

gain, and the tree-building process continues recursively until a stopping criterion is met (e.g., maximum depth, minimum number of samples per leaf).

The prediction for any input is obtained by traversing the tree from the root to a leaf node, where the predicted value is typically the mean target value of the samples in that leaf:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \tag{2}$$

where:

- $\hat{y}$: Predicted value for a given input
- y_i: Actual target values within a leaf node
- N: Number of samples in the leaf node

The ability of Decision Tree Regressors to capture complex, non-linear relationships without needing extensive pre-processing is a significant advantage. However, one of the potential drawbacks is the tendency to overfit, particularly with deep trees. Regularization techniques, such as setting a maximum depth or minimum number of samples per leaf, are commonly employed to prevent overfitting and improve generalization.

The computational complexity of building a Decision Tree Regressor is influenced by the number of samples N and the number of features d. At each node, the algorithm evaluates all possible splits across all features, which requires $O(d \cdot N \log N)$ operations. The factor $N \log N$ comes from sorting the data at each split. If the tree has T terminal nodes (leaf nodes), the overall training complexity becomes $O(d \cdot N \log N \cdot T)$, as this process is repeated at each level of the tree.

For prediction, the complexity is $O(\text{depth})$, where "depth" is the depth of the tree. In the worst case, this depth can be $O(N)$, resulting in a prediction time complexity of $O(N)$. However, with appropriate regularization (e.g., setting a maximum depth), the prediction complexity can often be reduced to $O(\log N)$.

Given that the Decision Tree Regressor algorithm constructs a tree by selecting splits that maximize the reduction of mean squared error (MSE) at each node and offers high interpretability, it can be effectively utilized in applications like drug redesign, where clear interpretations are essential for diverse stakeholders. Additionally, the model's structure allows for easy visualization and analysis of the decision-making process, providing insights into which features are most influential in the prediction outcomes.

4 Results and Experiments

4.1 Dataset

For evaluating these methods, we used a validated dataset containing 1933 confirmed associations between 593 drugs and 313 diseases. These data were extracted from public sources, DrugBank [14] and Online Mendelian Inheritance in Man (OMIM) [15], and are recognized as the gold standard for predicting drug-disease associations. The dataset included molecular features of drugs and clinical information about diseases. After removing duplicate pairs, the final dataset was used for experiments.

4.2 Model Comparison and Evaluation

To evaluate the performance of the models and compare them, we used the Mean Squared Error (MSE) metric. MSE measures the average squared difference between the actual and predicted values, providing an indication of how close predictions are to the actual results. It is calculated using the formula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3}$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of data points. Lower MSE values indicate better model performance.

For this purpose, we chose the WNMFDDA method as the reference model and compared the output of other models against this method. The WNMFDDA model serves as a baseline for high accuracy, but it is resource-intensive in terms of memory and execution time. Other methods, like matrix factorization-based approaches and certain machine learning algorithms, were able to approximate the results of the WNMFDDA model while requiring significantly fewer resources.

The results show that some alternative methods, such as matrix factorization-based methods and certain machine learning algorithms, can achieve results close to the reference model while using less time and memory. However, we excluded certain methods, like Linear Regression, because they were not accurate enough. These models failed to capture the associations between drugs and diseases as effectively as other models, leading to poorer performance. Specifically, we removed the regression method due to its inaccuracy and the CUR Decomposition method due to its high processing time and inefficiency compared to other methods.

The comparisons are visualized through graphs, including time charts, memory usage charts, and line graphs, providing a detailed view of the trade-offs between time, memory usage, and prediction accuracy.

Table 2 and the results presented in Figures 2, 3 and 4 compare the performance of various methods for

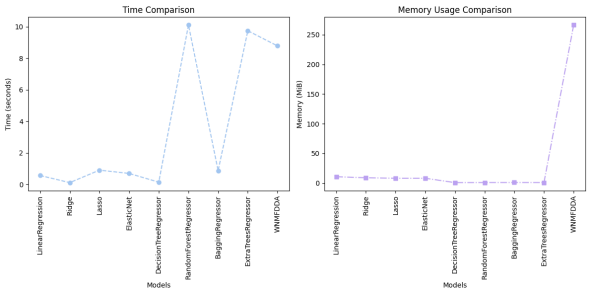

Figure 2: comparison of Time and Memmory usage of WNMFDDA and Numerical Algebras methods

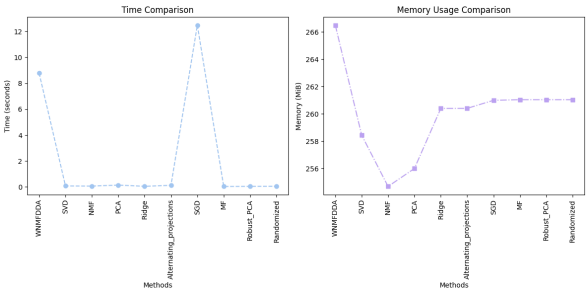

Figure 3: comparison of Time and Memmory usage of WNMFDDA and Machine Learning methods

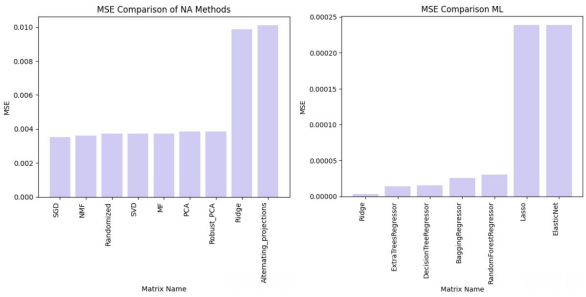

Figure 4: Comparison of MSE relative to the WNMFDDA method

Table 2: Performance Metrics for Baselines

Method Name	**Memory Usage (MiB)**	**MSE**	**Time (s)**
WNMFDDA [13]	266.50	-	8.7815
Singular Value Decomposition (SVD)	258.45	0.0037	0.0757
Non-negative Matrix Factorization (NMF)	254.69	0.0036	0.0647
Principal Component Analysis (PCA)	256.01	0.0038	0.1430
Ridge Regression	260.41	0.0099	0.0482
Alternating Projections	260.41	0.0101	0.1231
Stochastic Gradient Descent (SGD)	261.00	0.0035	12.4861
Matrix Factorization (MF)	261.04	0.0037	**0.0350**
Robust PCA	261.04	0.0038	0.0440
Randomized Matrix Factorization	261.04	0.0037	0.0536
Probabilistic Matrix Factorization (PMF)	261.20	0.0004	224.0376
Lasso	8.03	0.0003	0.9046
ElasticNet	8.02	0.0003	0.6957
Ridge	8.84	**0.0001**	0.1134
Random Forest Regressor	0.82	**0.0001**	10.1102
Bagging Regressor	0.87	**0.0001**	0.8674
Extra Trees Regressor	0.79	**0.0001**	9.7420
Decision Tree Regressor	**0.66**	**0.0001**	0.1410

predicting drug-disease associations from different aspects, including accuracy (using the MSE metric), memory consumption, and execution time, against the WNMFDDA method.

The WNMFDDA method is considered a baseline model due to its high accuracy and ability to identify complex relationships. This method utilizes a combination of matrix factorization and graph-based settings to uncover intricate patterns in drug and disease data. However, its significant memory consumption (266.50 MiB) and relatively high execution time (nearly 9 seconds) indicate that despite its desirable accuracy, its

computational cost is substantial, making it possibly less optimal in cases where lower processing resources and faster speed are required.

Nonetheless, several alternative methods have performed well with similar accuracy but with lower memory and time consumption compared to WNMFDDA, as shown in the second table and the corresponding figures. For example, Decision Tree Regressor achieve predict close to that of WNMFDDA and, in some cases, even lower MSE values. These models, by reducing processing resource consumption, are suitable choices for scenarios where model accuracy is important but less memory and time are required. The Decision Tree Regressor provides one of the best balances between accuracy and efficiency, as can be seen in Figure 3. Figure 4 clearly shows that this method has been able to reach an accuracy close to the baseline method using fewer processing resources and even demonstrating better performance in scenarios with limited resources by reducing MSE.

This comparison indicates that while WNMFDDA remains a standard method with high accuracy in drug-disease predictions, methods such as Decision Tree Regressor proven to be more efficient in terms of time and memory consumption. This allows us to use models with lower costs in situations where time and processing resources are of high importance without significantly compromising prediction accuracy.

5 Conclusion

This study aimed to findan efficient model for predicting drug-disease associations, balancing accuracy and computational resources. While the Weighted Graph Regularized Collaborative Non-negative Matrix Factorization (WNMFDDA) demonstrated high accuracy, its significant computational demands limit its practical applicability.

In contrast, alternative methods like the Decision Tree Regressor achieved comparable MSE with reduced memory and execution time, making them more suitable for resource-constrained environments. Our findings indicate that selecting the right model based on accuracy and resource availability can enhance performance in drug-disease prediction tasks. Future research should explore hybrid models that combine the strengths of these approaches for even better efficiency and accuracy.

References

[1] J.K. Yella, S. Yaddanapudi, Y. Wang, A.G. Jegga. Changing trends in computational drug repositioning. *Pharmaceuticals*, 11(2):57, 2018.

[2] T.T. Ashburn and K.B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004.

[3] N. Nosengo. Can you teach old drugs new tricks? *Nature*, 534(7607):314–316, 2016.

[4] A.I. Graul, L. Sorbera, P. Pina, M. Tell, E. Cruces, E. Rosa, et al. The year's new drugs & biologics-2009. *Drug News Perspectives*, 23(1):7–36, 2010.

[5] D. Sardana, C. Zhu, M. Zhang, R.C. Gudivada, L. Yang, A.G. Jegga. Drug repositioning for orphan diseases. *Briefings in Bioinformatics*, 12(4):346–356, 2011.

[6] H. Yang, I. Spasic, J.A. Keane, G. Nenadic. A text mining approach to the prediction of disease status from clinical discharge summaries. *Journal of the American Medical Informatics Association*, 16(4):596–600, 2009.

[7] X. Chen, G.-Y. Yan. Semi-supervised learning for potential human microRNA-disease associations inference. *Scientific Reports*, 4:5501, 2014.

[8] M.N. Wang, X.J. Xie, Z.H. You, et al. A weighted nonnegative matrix factorization approach to predict potential associations between drug and disease. *Journal of Translational Medicine*, 20:552, 2022.

[9] H.J. Jiang, Z.H. You, K. Zheng, Z.H. Chen. Predicting of drug-disease associations via sparse auto-encoder-based rotation forest. In *Intelligent Computing Methodologies: 15th International Conference, ICIC 2019, Nanchang, China, August 3–6, 2019, Proceedings, Part III*, volume 15, pages 369–380. Springer International Publishing, 2019.

[10] C.Q. Gao, Y.K. Zhou, X.H. Xin, H. Min, P.F. Du. DDA-SKF: predicting drug–disease associations using similarity kernel fusion. *Frontiers in Pharmacology*, 12:784171, 2022.

[11] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[12] W. Zhang, X. Yue, W. Lin, et al. Predicting drug-disease associations by using similarity constrained matrix factorization. *BMC Bioinformatics*, 19:233, 2018.

[13] M.N. Wang, X.J. Xie, Z.H. You, et al. A weighted nonnegative matrix factorization approach to predict potential associations between drug and disease. *Journal of Translational Medicine*, 20:552, 2022.

[14] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(suppl_1):D668–D672, 2006.

[15] A. Hamosh, A.F. Scott, J.S. Amberger, C.A. Bocchini, V.A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl_1):D514–D517, 2005.

[16] D. Lee, H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[17] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen. Classification and Regression Trees. *Wadsworth International Group*, 1984.

[18] S. Wold, K. Esbensen, P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37–52, 1987.

[19] E.J. Candes, X. Li, Y. Ma, J. Wright. Robust Principal Component Analysis? *Journal of the ACM*, 58(3):1–37, 2011.

[20] H. Kang, L. Hou, Y. Gu, X. Lu, J. Li, Q. Li. Drug–disease association prediction with literature based multi-feature fusion. *Frontiers in Pharmacology*, 14, 2023.

[21] D. Belete, B.M. Jacobs, C. Simonet, et al. Association Between Antiepileptic Drugs and Incident Parkinson Disease. *JAMA Neurology*, 80(2):183–187, 2023.