



Drug-Disease Association Analysis via Machine Learning on Extracted Features by Matrix Decomposition

Zahra Rafei, Seyedeh Fatemeh Hosseini, Behnam Yousefimehr^{} and Mehdi Ghatte^{*}

Abstract

Drug repurposing presents a cost-effective and time-efficient alternative to traditional drug discovery by identifying new therapeutic uses for existing medications. As biomedical data grows in scale and complexity, there is an increasing demand for predictive models that balance accuracy, interpretability, and computational efficiency. In this study, we systematically evaluate hybrid models that combine established matrix factorization techniques with machine learning regressors, with an emphasis on interpretable and lightweight models such as the Decision Tree Regressor. Using the widely adopted Fdataset, comprising 1,933 known associations between 593 drugs and 313 diseases, we demonstrate that several of these hybrid approaches achieve predictive performance comparable to or surpassing that of complex models like WNMFDMA, while significantly reducing memory usage and training time. Notably, our framework relies solely on the drug-disease association matrix, removing the dependency on auxiliary similarity data, which is often unavailable in real-world applications. Among the tested models, the NMF DecisionTreeRegressor offers the highest accuracy, making it ideal for accuracy-critical scenarios, while the Ridge model stands out for its efficiency and suitability for resource-constrained environments. To enhance transparency, we further apply LIME (Local Interpretable Model-Agnostic Explanations) to provide interpretable insights into model predictions. These findings highlight a practical and scalable framework for drug repurposing, particularly suited for environments with limited computational resources. Our approach supports the development of accessible, data-driven predictive tools that accelerate the transition from computational modeling to clinical application.

Keywords: Drug-disease association, Machine learning, Matrix decomposition, Hybrid models.

Corresponding author (E-mail: ghatte@aut.ac.rs)

2020 Mathematics Subject Classification: 68T01 , 68T09.

How to cite this article

Z. Rafei, S. F. Hosseini, B. Yousefimehr and M. Ghatee Drug-disease association analysis via machine learning on extracted features by matrix decomposition, *Math. Interdisc. Res.* **10** (3) (2025) 295-313.

1. Introduction

Drug repurposing, also referred to as drug repositioning, has emerged as an innovative and cost-effective strategy in pharmaceutical research, providing a practical alternative to traditional drug discovery [1]. Instead of developing new drugs entirely from scratch, this approach focuses on identifying new therapeutic uses for existing medications that have already undergone regulatory approval. Repurposing enables researchers to leverage prior safety and efficacy data, thereby accelerating development timelines and reducing associated costs [2, 3]. Additionally, successful examples such as Minoxidil and Sildenafil demonstrate how repurposed drugs can bring significant clinical and commercial impact [2, 4, 5].

With the growing availability of large-scale biomedical datasets, computational methods have become increasingly important for accelerating drug repurposing. In particular, machine learning techniques have demonstrated strong potential in predicting drug-disease associations by uncovering hidden patterns in complex biological data [6, 7]. Furthermore, numerical linear algebra methods have contributed to improving both the speed and accuracy of these predictions by enabling efficient low-rank approximations and matrix decompositions.

In this study, we focus on optimizing drug-disease association prediction using a combination of machine learning and numerical linear algebra-based techniques. To evaluate model performance, we employed a well-established benchmark dataset known as *Fdataset*, originally compiled by Gottlieb et al. [8], which includes 1933 known associations between 593 drugs and 313 diseases. This binary drug-disease adjacency matrix has been widely adopted for evaluating predictive models in drug repurposing research. Additionally, prior work such as WNMFDDA [9] has utilized auxiliary drug-drug and disease-disease similarity matrices computed from chemical structure and semantic clinical text to further guide association prediction. WNMFDDA represents a prominent state-of-the-art model for drug-disease association, achieving an AUC of 0.939 on the *Fdataset* under ten-fold cross-validation [9]. However, the model's high computational and memory demands, alongside the need for auxiliary similarity matrices, limit its practicality in real-world biomedical applications.

Academic Editor: Abbas Saadatmandi

Received 7 March 2025, Accepted 30 June 2025

DOI: 10.22052/MIR.2025.256471.1506

© 2025 University of Kashan



This work is licensed under the Creative Commons Attribution 4.0 International License.

Our previous work [10] demonstrated that simpler models like the Decision Tree Regressor can drastically reduce memory usage and execution time while maintaining competitive predictive accuracy. Nonetheless, we identified accuracy as an area for further improvement.

In this paper, we propose novel hybrid predictive methods that combine matrix factorization techniques (e.g., PCA, NMF, Alternating Projections, SGD) with machine learning models to achieve substantial improvements in both prediction accuracy and runtime. These methods reduce Mean Squared Error (MSE) by up to an order of magnitude compared to WNMFD, while also eliminating the need for external similarity matrices. This simplification not only improves computational efficiency but also broadens the applicability in clinical scenarios where such auxiliary data may be unavailable.

In summary, this study presents a set of scalable and resource-efficient hybrid models that advance the state of drug-disease association prediction. Extensive benchmarking using DrugBank [11] and OMIM [12] confirms the superiority of our methods across multiple evaluation criteria. These contributions support the broader goal of accelerating drug repurposing workflows while reducing resource requirements.

2. Related work

Various computational approaches have been developed to predict drug-disease associations, leveraging diverse data representations and machine learning architectures. These methods aim to identify novel therapeutic opportunities by analyzing biomedical data through different algorithmic lenses.

One of the early methods is Sparse Auto-Encoder-Based Rotation Forest-(SAEROF) [13], which combines sparse autoencoders with Rotation Forests to model structural and semantic patterns between drugs and diseases. Although SAEROF improves prediction accuracy by learning richer representations, it incurs high computational overhead due to its deep network components. Another approach is Drug-Disease Association using Similarity Kernel Fusion-(DDA-SKF) [14], which integrates multiple drug and disease similarity matrices using a kernel-based fusion strategy. It employs Laplacian Regularized Least Squares to make predictions and demonstrates robustness in data-scarce settings. However, its performance deteriorates when similarity information is incomplete or noisy.

Deep learning-based methods have also gained traction in drug-disease association prediction. A key example is the Densely Connected Convolutional Networks(DCNN) model [15], which utilizes a densely connected convolutional neural network with attention mechanisms to uncover hidden patterns in drug-disease relationships. Despite its ability to capture complex interactions, DCNN requires extensive training data and involves intricate model configurations. These factors make it challenging to implement effectively, particularly when dealing with small or imbalanced datasets.

Matrix factorization-based techniques offer another promising avenue. The Similarity Constrained Matrix Factorization(SCMFDD) model [16] applies similarity constraints to matrix factorization, projecting drugs and diseases into a lower-dimensional space to facilitate association predictions. However, a key limitation of SCMFDD is its reduced performance when encountering new or previously unseen data, which negatively affects its generalization capability.

More recent developments have focused on integrating graph-based reasoning and representation learning. For instance, Reinforcement Symmetric Metric Learning and Graph Convolution Network(RSML-GCN) [17] introduces a reinforcement learning framework with symmetric metric learning over heterogeneous graphs. By modeling topological patterns and feedback-driven associations, it improves drug repositioning accuracy, though at the cost of increased computational complexity. Another notable approach is Semantic Graph and Function Similarity Representation for Drug–Disease Association Prediction(SFRLDDA) [18], which employs semantic graphs and biological function similarity to generate enriched feature spaces. This method improves prediction robustness, but relies on complex feature engineering and domain-specific resources.

Collaborative Filtering and Multiple Kernel Graph Attention Network for Drug-Disease Association Prediction(CFMKGATDDA) [19] combines collaborative filtering with multi-kernel attention mechanisms on heterogeneous graphs. This hybrid model effectively captures multi-scale interactions, but its layered architecture introduces significant memory and training overhead. Also, the Weighted Graph Regularized Collaborative Non-negative Matrix Factorization for Drug-Disease Association Prediction(WNMFDDA) model [9] has emerged as a state-of-the-art approach. This method integrates non-negative matrix factorization with graph regularization to uncover potential drug-disease links. It first calculates similarity scores based on drug chemical structures and disease characteristics, refines association scores using a weighted K-nearest neighbors approach, and then applies matrix factorization and graph regularization for prediction. While WNMFDDA demonstrates high predictive accuracy and robust mathematical foundations, it demands significant computational resources, particularly in terms of memory and processing time. Additionally, its complexity makes it less interpretable, highlighting the need for alternative methods that maintain high accuracy while improving efficiency and interpretability [9]. Table 1 provides a comprehensive summary of techniques used for drug–disease association prediction.

Table 1: Summary of Drug–Disease Association Prediction Techniques.

Ref.	Approach	Benefits	Limitations
[13]	SAEROF : Combines sparse autoencoders with Rotation Forest to model structural and semantic patterns.	Richer representations, improve prediction accuracy	High computational overhead
[14]	DDA-SKF : Integrates multiple similarity matrices using kernel fusion and Laplacian Regularized Least Squares.	Performs well on sparse datasets	Sensitive to incomplete or noisy similarity inputs
[15]	DCNN : Applies densely connected convolutional networks with attention to capture hidden drug-disease patterns.	Powerful for complex drug-disease relationships	Requires large, balanced datasets and intricate configurations, Intricate configurations
[16]	SCMFDD : Uses similarity-constrained matrix factorization to project data into lower dimensions.	Effective for structured prediction	Weak generalization to novel or unseen data
[17]	RSML-GCN : Leverages reinforcement symmetric metric learning with graph convolution over heterogeneous networks.	High accuracy via topological modeling and feedback mechanisms	Increased computational demands
[18]	SFRLDDA : Utilizes semantic graphs and biological function similarity to construct enriched feature representations.	Offers robustness	Requires intensive feature engineering and expert resources
[19]	CFMKGATDDA : Integrates collaborative filtering and multiple kernel graph attention networks for multi-scale interaction modeling.	High prediction performance	Substantial memory usage and training time
[9]	WNMFDDA : Combines weighted KNN, graph regularization, and non-negative matrix factorization using similarity scores.	High accuracy and mathematical robustness	Complex and resource-intensive, limited interpretability

In contrast to the above approaches, our proposed hybrid models integrate matrix factorization techniques (such as NMF, PCA, or SGD) with lightweight regressors like Decision Tree, yielding comparable or even superior predictive performance while drastically reducing computational resources. Specifically, our methods reduce memory consumption and cut training time compared to WNMFDDA. Furthermore, they eliminate the need for external drug or disease similarity matrices, making them highly practical for clinical scenarios with limited data availability or computational constraints.

3. Methodology

In this study, we aimed to develop a model that accurately predicts drug-disease associations while optimizing computational efficiency, building upon our previous work [10]. To this end, we evaluated the state-of-the-art Weighted Graph Regularized Collaborative Non-negative Matrix Factorization for Drug-Disease Association Prediction (WNMFDDA), known for its high predictive accuracy, alongside various alternative approaches, including machine learning algorithms and numerical linear algebra techniques. Our objective was to find methods that maintain comparable accuracy but with reduced memory and processing requirements.

While WNMFDDA demonstrates strong predictive performance, it is computationally intensive due to its reliance on matrix decomposition. Hybrid models like the **NMF DecisionTreeRegressor** similarly require more memory and sometimes longer execution times, but can offer superior accuracy. If accuracy is the

primary concern, the NMF DecisionTreeRegressor is recommended. However, for use in resource-constrained settings, simpler models such as **Ridge** provide a better balance by offering fast execution and minimal memory usage with competitive accuracy.

Figure 1 illustrates the overall methodology followed in this study.

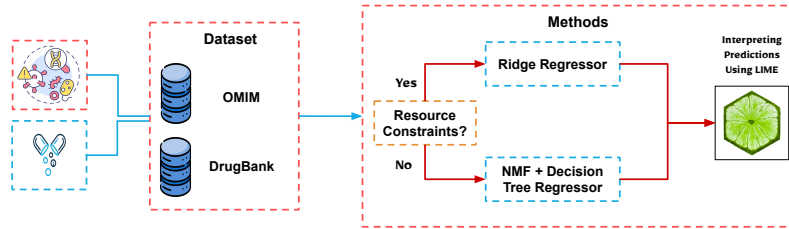


Figure 1: Flowchart illustrating the overall methodology followed in this study.

3.1 Foundational analysis of WNMFD DA

To gain a comprehensive understanding of the WNMFD DA method, we first examine its underlying principles. While WNMFD DA demands greater computational time and memory than traditional NMF, its improved accuracy makes these trade-offs justifiable. Moreover, the resource consumption remains within a comparable range, allowing us to prioritize improved model performance over computational efficiency.

A comparative evaluation was conducted across multiple methodologies. Among them, the Decision Tree Regressor exhibited performance closest to the baseline while maintaining high accuracy and reduced computational demands compared to WNMFD DA. Additionally, it offers superior interpretability, making it particularly advantageous for interdisciplinary applications. As a result, the Decision Tree Regressor emerged as the most effective model in this case study.

3.2 Non-negative matrix factorization (NMF)

A fundamental technique that serves as the basis for the reference method examined is Non-negative Matrix Factorization (NMF) [20]. This approach decomposes the drug-disease association matrix V into two non-negative matrices W and H , as follows:

$$V \approx W \times H, \quad (1)$$

where:

- V : Drug-disease association matrix
- W : Latent features representing drugs

- H : Latent features representing diseases

The time complexity of Non-negative Matrix Factorization (NMF) is $O(k \cdot mnr)$, where m and n are the dimensions of the input matrix, r is the target rank, and k is the number of iterations. Each iteration involves matrix multiplications and element-wise updates for the factor matrices $W \in \mathbb{R}^{m \times r}$ and $H \in \mathbb{R}^{r \times n}$.

The space complexity is $O(mr + nr)$, which corresponds to the memory required to store the two factor matrices. In our experiments, we set the rank $r = 50$ and limited the number of iterations to balance computational cost and approximation quality. This setup provides an efficient, low-rank representation while keeping the memory footprint manageable.

3.3 Decision tree regressor

The Decision Tree Regressor [21] is a non-parametric regression model that predicts target values by learning a set of hierarchical decision rules. It recursively partitions the data space into smaller subsets to minimize prediction errors, forming a tree-like structure of decisions. This method effectively captures non-linear relationships without requiring feature scaling or transformation, making it highly adaptable for diverse regression tasks.

The model operates by iteratively dividing the dataset into subsets based on feature thresholds, leading to leaf nodes containing samples with similar target values. The quality of a split is determined by the reduction in MSE, ensuring that each partition reduces impurity. This process continues recursively until predefined stopping criteria, such as maximum tree depth or minimum samples per leaf are met.

For any given input, predictions are generated by traversing the decision tree from the root to a leaf node. The predicted value \hat{y} is computed as the mean of target values in that leaf:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N y_i, \quad (2)$$

where:

- \hat{y} : Predicted output
- y_i : Actual target values in the leaf node
- N : Number of samples in the leaf node

The Decision Tree Regressor is particularly advantageous due to its ability to model complex, non-linear relationships while maintaining interpretability. However, it is susceptible to overfitting, especially when deep trees are constructed. To mitigate this, regularization techniques such as limiting tree depth or enforcing a minimum number of samples per leaf are commonly employed.

The computational complexity of training a Decision Tree Regressor is a function of the number of samples N and features d . Evaluating all potential splits at each node requires $O(d \cdot N \log N)$ operations due to the sorting process. If the tree has T terminal nodes, the overall complexity increases to $O(d \cdot N \log N \cdot T)$, as this operation is repeated across all tree levels. For prediction, the complexity is proportional to the depth of the tree. In the worst case, this depth can be $O(N)$, leading to a prediction complexity of $O(N)$. However, with proper regularization, the depth can often be constrained to $O(\log N)$, improving efficiency.

Given its capability to improve interpretability while optimizing prediction accuracy, the Decision Tree Regressor is particularly well-suited for applications such as drug repurposing. Its structured decision-making process provides valuable insights into feature importance, making it a valuable tool for interdisciplinary research and practical deployment.

To improve reproducibility, we provide a detailed description of how Non-negative Matrix Factorization (NMF) and machine learning algorithms are integrated in our approach. Specifically, the drug-disease association matrix is first decomposed via NMF to obtain low-dimensional latent features for drugs and diseases. These latent representations are subsequently used as inputs to various machine learning algorithms, such as Decision Tree Regressors, to predict drug-disease associations more accurately and efficiently. This integration leverages the dimensionality reduction capability of NMF with the predictive power of machine learning. Moreover, our methodology is generalizable across the broader field of biomedical data analysis, particularly in tasks involving large, sparse, and noisy matrices where interpretability and computational efficiency are critical. The proposed framework can easily be adapted to other applications such as gene-disease association prediction, protein-protein interaction analysis, or any matrix-based biomedical prediction problem, facilitating reproducible, scalable, and interpretable discoveries.

3.4 Ridge regressor

Ridge Regression [22] is a linear model that addresses multicollinearity by introducing a regularization term to the ordinary least squares (OLS) objective function. It minimizes the residual sum of squares while penalizing the magnitude of the regression coefficients, thereby reducing model variance and preventing overfitting [22]. The Ridge objective function is defined as:

$$\min_{\mathbf{w}} \left\{ \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2 \right\}, \quad (3)$$

where:

- \mathbf{x}_i : Feature vector for the i^{th} instance
- y_i : Actual target value for the i^{th} instance

- \mathbf{w} : Weight vector (model parameters)
- λ : Regularization strength (controls the penalty on coefficient size)

The addition of the L_2 penalty term $\lambda\|\mathbf{w}\|_2^2$ discourages large coefficients, leading to a more robust model that generalizes better on unseen data. Unlike Lasso, Ridge does not perform feature selection but instead shrinks all coefficients toward zero uniformly.

Ridge Regression is particularly well-suited for high-dimensional problems where the number of features may exceed the number of samples, as is common in biomedical datasets. It can be efficiently solved using matrix algebra through closed-form solutions:

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (4)$$

The training complexity is dominated by the matrix inversion operation, which is $O(d^3)$ for d features. However, this can be significantly reduced through efficient numerical solvers and is generally much lower than iterative methods used in tree-based models. Although Ridge Regression assumes a linear relationship between features and targets, its simplicity and low variance make it competitive for many tasks. In our experiments, the Ridge model demonstrated excellent runtime and memory efficiency while achieving acceptable prediction accuracy, making it ideal for deployment in resource-constrained environments.

3.5 Local interpretable model-agnostic explanations (LIME)

Interpretability is critical in healthcare applications, where understanding the rationale behind model predictions can improve trust, support clinical decision-making, and aid in identifying potential biases. To improve the transparency of our regression model, we employed Local Interpretable Model-Agnostic Explanations (LIME) [23], a widely adopted method for explaining individual predictions of black-box models.

LIME generates local surrogate models that approximate the behavior of the original model in the vicinity of a specific instance. For our analysis, we applied LIME to a trained model to interpret the predictions for selected test samples. The method identifies and ranks the most influential features contributing to each prediction.

4. Results and experiments

4.1 Dataset

To evaluate the performance of the proposed methods, we utilized a well-established benchmark dataset known as *Fdataset*, originally compiled by Gottlieb et al. [8]. This dataset is widely regarded for its reliability in assessing drug-disease association prediction techniques. It comprises 1933 confirmed associations between 593

unique drugs and 313 distinct diseases, forming a binary drug-disease adjacency matrix $A \in \mathbb{R}^{n_d \times n_{dis}}$, where $n_d = 593$ and $n_{dis} = 313$. In this matrix, each entry $A_{ij} = 1$ indicates a confirmed association between drug i and disease j , whereas $A_{ij} = 0$ represents no known association.

The dataset was curated from authoritative public biomedical resources, specifically *DrugBank* [11] and the *Online Mendelian Inheritance in Man (OMIM)* database [12], both of which are considered gold standards in the field of drug repurposing and disease gene discovery. It includes detailed molecular characteristics of drugs alongside clinical and semantic information related to diseases. To ensure data quality and consistency, duplicate entries were removed during preprocessing. In addition to the adjacency matrix, we derived drug and disease similarity information to improve the predictive modeling. Drug similarity was computed based on chemical structure, using 2D molecular fingerprints generated from SMILES strings. The *Chemical Development Kit* [24] was employed for fingerprint generation, and the *Tanimoto coefficient* [25] was used to calculate the pairwise similarity between drugs. On the other hand, disease similarity was assessed using *MimMiner* [26], a tool that applies text mining techniques to clinical descriptions in the OMIM database to estimate semantic relatedness between diseases.

These three components, the binary drug-disease association matrix, the drug similarity matrix, and the disease similarity matrix served as the foundation for our predictive model. Specifically, they were integrated as inputs to the *WNMFDDA* framework, which was designed to infer potential novel associations between drugs and diseases.

4.1.1 Implementation and preprocessing details

To ensure reproducibility and fair evaluation, all experiments were implemented in Python using the scikit-learn machine learning library. The dataset was preprocessed to remove duplicate entries and ensure consistent formatting across matrices. The binary drug-disease association matrix was used as the primary input for all models. For hybrid methods, input matrices were first reduced using matrix factorization techniques (e.g., PCA, NMF), and the resulting low-dimensional representations were used to train Decision Tree regressors.

For machine learning and hybrid models, five-fold cross-validation was applied. The full dataset was randomly partitioned into five equal subsets, with each subset used once as the test set while the remaining four were used for training. This process was repeated five times, and the average performance metrics were reported. This approach provided a robust and unbiased estimate of model generalization performance.

In contrast, classical matrix factorization methods that are not trained in a supervised fashion (e.g., SVD, NMF, PCA) were evaluated using full-matrix reconstruction. Since these models do not involve learning from training data, we computed the MSE between the reconstructed and original matrices over the entire

dataset.

Hyperparameters were selected based on preliminary experimentation and standard values reported in related works. For Decision Tree regressors, we used a maximum tree depth of 10 and a minimum of 5 samples per leaf to mitigate overfitting. For matrix factorization methods, the number of latent components (rank) was fixed at 50, which balanced model expressiveness and computational efficiency. In the Ridge Regression baseline, the regularization parameter α was set to 0.1.

4.2 Model comparison and evaluation

To assess and compare model performance, we utilized the MSE metric, which quantifies the average squared difference between actual and predicted values. A lower MSE value signifies greater predictive accuracy. The MSE is calculated using the following formula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (5)$$

where:

- y_i : Actual values
- \hat{y}_i : Predicted values
- n : Number of data points

To establish a benchmark, we selected the WNMFDAA method as the reference model and compared the performance of alternative approaches against it. While WNMFDAA offers superior accuracy, it is computationally demanding in terms of memory usage and processing time. Alternative models, including matrix factorization based techniques and certain machine learning algorithms, demonstrated comparable performance to WNMFDAA while requiring significantly fewer computational resources.

Our analysis revealed that some alternative models could approximate the performance of WNMFDAA with reduced computational overhead. However, models such as Linear Regression were excluded due to their insufficient accuracy in capturing drug–disease associations. In addition, CUR matrix decomposition (Columns, U matrix, Rows), a low rank approximation technique that reconstructs the original matrix using selected actual columns and rows, may be useful in certain applications. However, in our case, it introduced substantially higher computational and memory overhead compared to other evaluated methods. As a result, it was excluded from the final comparison due to its inefficiency for the specific requirements of our study.

Table 2 presents a comparative analysis of different single methods used to predict drug–disease associations. The comparison is based on several factors, including accuracy (measured using the MSE metric), memory usage, and execution time, with the WNMFDAA method serving as a benchmark.

Table 2: Performance metrics for single methods.

Method Name	Memory Usage (MiB)	MSE	Time (s)
WNMFDDA [9]	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

WNMFDDA is considered a baseline due to its high accuracy and ability to capture complex relationships in drug-disease data. It leverages a combination of matrix factorization and graph-based techniques to identify intricate patterns. However, its computational demands are relatively high, requiring 266.50 MiB of memory and nearly 9 seconds for execution. While it excels in accuracy, these resource-intensive requirements make it less suitable for scenarios where efficiency and speed are critical.

In contrast, several alternative methods offer comparable accuracy while consuming less memory and processing time. For instance, the Decision Tree Regressor achieves performance close to WNMFDAA, sometimes even surpassing it with lower MSE values. By optimizing resource usage, these models present viable alternatives in cases where accuracy is essential, but efficiency is also a priority.

Overall, this comparison underscores that while WNMFDAA remains a strong standard due to its accuracy, models like the Decision Tree Regressor provide a more resource-efficient alternative. This enables the deployment of lower-cost models in time-sensitive and resource-limited settings without significantly sacrificing prediction performance.

Table 3 compares hybrid and classical machine learning methods based on prediction accuracy (Mean MSE with 95% confidence intervals), execution time, and memory consumption. It includes a selected set of models ranging from classical ML algorithms to hybrid approaches.

Among the hybrid models, **Alternating Projections DecisionTreeRegressor** is the most time-efficient, with a runtime of only **0.1104s**, while **Probabilistic Matrix Factorization DecisionTreeRegressor** is the most computationally expensive, requiring **234.97s**. In terms of memory usage, all hybrid methods consume approximately **291–293 MiB**, reflecting the overhead of matrix factorization

and latent feature construction.

In contrast, classical machine learning models are considerably more lightweight. The **Decision Tree Regressor** has the lowest memory footprint at only **0.66 MiB**, followed by **Extra Trees (0.79 MiB)** and **Random Forest (0.82 MiB)**. While some ensemble models like **Random Forest** and **Extra Trees** incur higher runtimes (**9 – 10s**), methods like **Ridge Regression** maintain excellent efficiency, completing in just **0.1134s** with minimal memory usage (**8.84 MiB**).

Overall, hybrid models typically require more memory and, in some cases, longer execution times due to their reliance on matrix decomposition techniques. However, they can offer superior predictive accuracy, as demonstrated by the **NMF DecisionTreeRegressor**. If higher accuracy is the priority, the NMF DecisionTreeRegressor is the recommended choice. Conversely, for deployment in resource-constrained environments, the **Ridge** model is more suitable due to its minimal memory usage and fast execution time.

Compared to the baseline WNMFDFA, which has a memory usage of 266.50 MiB and a runtime of 8.7815 s, our method *NMF DecisionTreeRegressor* improves results by reducing runtime by approximately 1.75 times and achieving a lower MSE of 0.000025. Notably, *Ridge* shows substantial improvements, using about 30.15 times less memory and running approximately 77.46 times faster, while maintaining a low MSE of 0.000028.

Table 3: Performance summary for hybrid & ML methods.

Method Name	Mean MSE	95% CI	Memory Usage (MiB)	Time (s)
NMF DecisionTreeRegressor	0.000025	(0.000002 – 0.000048)	291.40	5.0142
PCA DecisionTreeRegressor	0.000175	(0.000063 – 0.000288)	292.01	3.0903
SVD DecisionTreeRegressor	0.000222	(0.000044 – 0.000400)	292.02	3.6566
Truncated SVD DecisionTreeRegressor	0.000182	(0.000042 – 0.000322)	292.12	1.8524
Randomized SVD DecisionTreeRegressor	0.000155	(0.000089 – 0.000220)	292.12	1.9543
Alternating Projections DecisionTreeRegressor	0.000099	(0.000051 – 0.000147)	292.68	0.1104
Stochastic Gradient Descent DecisionTreeRegressor	0.000097	(0.000042 – 0.000152)	292.86	55.6646
Probabilistic Matrix Factorization DecisionTreeRegressor	0.000116	(0.000046 – 0.000185)	292.94	234.9735
Lasso	0.000243	(0.000127 – 0.000359)	8.03	0.9046
ElasticNet	0.000243	(0.000127 – 0.000359)	8.02	0.6957
Ridge	0.000028	(0.000003 – 0.000054)	8.84	0.1134
RandomForestRegressor	0.000088	(0.000037 – 0.000139)	0.82	10.1102
BaggingRegressor	0.000085	(0.000027 – 0.000143)	0.87	0.8674
ExtraTreesRegressor	0.000071	(0.000033 – 0.000110)	0.79	9.7420
DecisionTreeRegressor	0.000083	(0.000035 – 0.000131)	0.66	0.1410

4.3 Model interpretability using LIME

Figure 2 illustrates an example LIME explanation for one representative instance from the test set. The visualization highlights the top contributing features and indicates whether they increased or decreased the predicted value. This local analysis enables clinicians and researchers to better understand the model’s reasoning on a case-by-case basis.

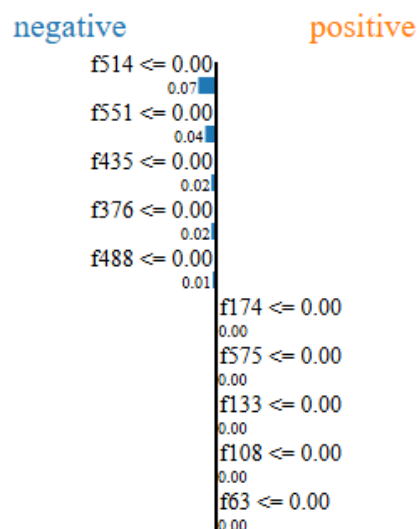


Figure 2: LIME explanation for a test sample. Features contributing positively or negatively to the prediction are shown, along with their relative impact.

4.4 Noise sensitivity analysis

To assess the robustness of the regression model against noisy data, we performed a noise sensitivity analysis. Gaussian noise was added to both the feature matrix and the target values at varying levels: 0%, 5%, 10%, 15%, 20%, 25%, and 30%. At each noise level, the model was retrained, and the MSE was computed on a test set split from the noisy data.

The results are presented in Figures 3 and 4, showing the relationship between noise level and prediction error. This analysis provides insights into the model's stability under noisy conditions, highlighting the importance of data quality when deploying models in real-world scenarios.

5. Conclusion

This study investigates the balance between predictive accuracy and computational efficiency in drug-disease association prediction. Instead of proposing a new model, we systematically evaluate combinations of well-established matrix factorization techniques with machine learning regressors, emphasizing lightweight and interpretable models such as the Decision Tree Regressor.

Our findings reveal that several of these hybrid approaches can match or surpass the predictive accuracy of more complex methods like WNMFDAA, while significantly reducing both memory usage and runtime. Importantly, our frame-

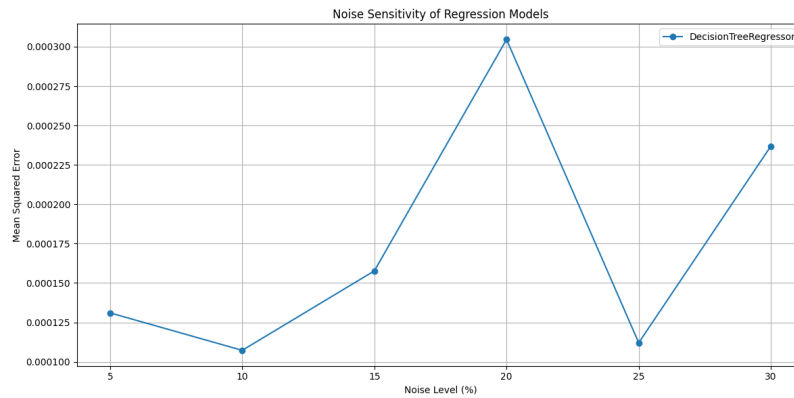


Figure 3: Noise sensitivity of Decision Tree regressor model.

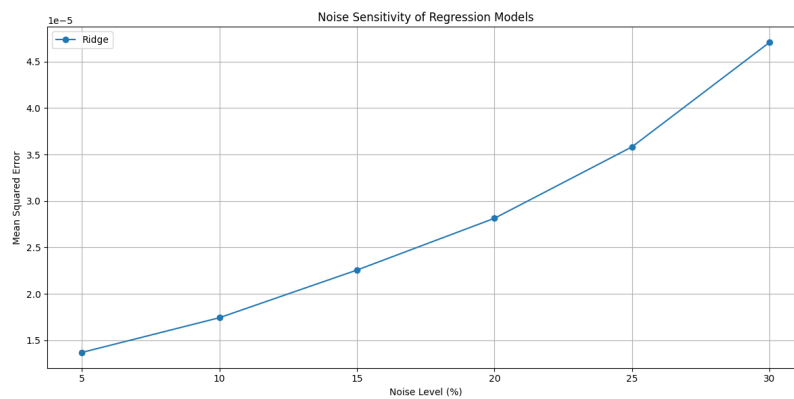


Figure 4: Noise sensitivity of Ridge regressor model.

work relies solely on the drug–disease association matrix, eliminating the need for auxiliary similarity data, which is often unavailable or inconsistent in practical applications.

Among the evaluated methods, the NMF combined with DecisionTreeRegressor delivers the highest predictive accuracy, albeit with increased memory consumption. Conversely, the Ridge model provides a highly efficient alternative, offering competitive performance with minimal resource requirements. This suggests a practical guideline: choose NMF DecisionTreeRegressor when maximizing accuracy is paramount, and opt for the Ridge model when deploying in resource-limited environments.

To improve model transparency, we also apply LIME (Local Interpretable Model-agnostic Explanations), offering valuable insights into the predictions and

supporting informed decision-making. Despite these advantages, the study has limitations. It is based on a single benchmark dataset, which may affect the generalizability of the results. Additionally, challenges such as data sparsity and limited capacity of the models to capture complex biological relationships remain.

Nonetheless, our framework presents a practical, scalable, and interpretable solution for drug repurposing, especially suited for settings with constrained computational resources. Compared to the baseline WNMFDAA, our approaches show notable improvements: the NMF DecisionTreeRegressor reduces runtime by approximately 1.75 times while improving predictive accuracy, and the Ridge model achieves substantial resource savings, using roughly 30 times less memory and running nearly 77 times faster, all while maintaining competitive accuracy.

Future work could integrate richer biomedical data to further enhance performance. Moreover, incorporating temporal or longitudinal datasets may allow modeling of dynamic drug–disease interactions. Ultimately, validating these computational predictions across diverse datasets and through experimental or clinical studies will be critical for their translation into therapeutic applications.

Conflicts of Interest. The authors declare that they have no conflicts of interest regarding the publication of this article.

References

- [1] J. K. Yella, S. Yaddanapudi, Y. Wang and A. G. Jegga, Changing trends in computational drug repositioning, *Pharmaceuticals* **11** (2018) #57, <https://doi.org/10.3390/ph11020057>.
- [2] T. T. Ashburn and K. B. Thor, Drug repositioning: identifying and developing new uses for existing drugs, *Nat. Rev. Drug Discov.* **3** (2004) 673 – 683, <https://doi.org/10.1038/nrd1468>.
- [3] N. Nosengo, Can you teach old drugs new tricks? *Nature* **534** (2016) 314–316, <https://doi.org/10.1038/534314a>.
- [4] A. I. Graul, P. Pina, M. Tracy and L. Sorbera, The year’s new drugs and biologics 2019, *Drugs Today* **56** (2020) #47, <https://doi.org/10.1358/dot.2020.56.1.3129707>.
- [5] D. Sardana, C. Zhu, M. Zhang, R. C. Gudivada, L. Yang and A. G. Jegga, Drug repositioning for orphan diseases, *Brief. Bioinform.* **12** (2011) 346 – 356, <https://doi.org/10.1093/bib/bbr021>.
- [6] H. Yang, I. Spasic, J. A. Keane and G. Nenadic, A text mining approach to the prediction of disease status from clinical discharge summaries, *J. Am. Med. Inform. Assoc.* **16** (2009) 596 – 600, <https://doi.org/10.1197/jamia.M3096>.

-
- [7] X. Chen and G. -Y. Yan, Semi-supervised learning for potential human microRNA-disease associations inference, *Scientific reports* **4** (2014) #5501, <https://doi.org/10.1038/srep05501>.
- [8] A. Gottlieb, G. Y. Stein, E. Ruppin and R. Sharan, PREDICT: a method for inferring novel drug indications with application to personalized medicine, *Mol. Syst. Biol.* **7** (2011) #496, <https://doi.org/10.1038/msb.2011.26>.
- [9] M. -N. Wang, X. -J. Xie, Z. -H. You, D. -W. Ding and L. Wong, A weighted non-negative matrix factorization approach to predict potential associations between drug and disease, *J. Transl. Med.* **20** (2022) #552, <https://doi.org/10.1186/s12967-022-03757-1>.
- [10] Z. Rafei, S. F. Hosseini, B. Yousefimehr, S. Tavakkoli and M. Ghatee, Optimizing drug-disease association analysis: a resource-efficient approach using numerical linear algebra and machine learning, *Proceedings of the First International Conference on Machine Learning and Knowledge Discovery (MLKD 2024)* (2024) 131 – 138.
- [11] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang and J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, *Nucleic Acids Res.* **34** (2006) D668 – D672, <https://doi.org/10.1093/nar/gkj067>.
- [12] A. Hamosh, A. F. Scott, J. Amberger, D. Valle and V. A. McKusick, Online mendelian inheritance in man (OMIM), *Hum. Mutat.* **15** (2000) 51 – 61, [https://doi.org/10.1002/\(SICI\)1098-1004\(200001\)15:1<57::AID-HUMU12>3.0.CO;2-G](https://doi.org/10.1002/(SICI)1098-1004(200001)15:1<57::AID-HUMU12>3.0.CO;2-G).
- [13] H. -J. Jiang, Z. -H. You, K. Zheng and Z. -H. Chen, Predicting of drug-disease associations via sparse auto-encoder-based rotation forest, *In International Conference on Intelligent Computing*, Springer (2019) 369 – 380, https://doi.org/10.1007/978-3-030-26766-7_34.
- [14] C. -Q. Gao, Y. -K. Zhou, X. -H. Xin, H. Min and P. -F. Du, DDA-SKF: predicting drug-disease associations using similarity kernel fusion, *Front. Pharmacol.* **12** (2022) #784171, <https://doi.org/10.3389/fphar.2021.784171>.
- [15] G. Huang, Z. Liu, L. Van Der Maaten and K. Q. Weinberger, Densely connected convolutional networks, *In Proceedings of the IEEE conference on computer vision and pattern recognition* (2017) 4700 – 4708.
- [16] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang and F. Liu, Predicting drug-disease associations by using similarity constrained matrix factorization, *BMC Bioinformatics* **19** (2018) #233, <https://doi.org/10.1186/s12859-018-2220-4>.

- [17] H. Luo, C. Zhu, J. Wang, G. Zhang, J. Luo and C. Yan, Prediction of drug-disease associations based on reinforcement symmetric metric learning and graph convolution network, *Front. Pharmacol.* **15** #1337764, <https://doi.org/10.3389/fphar.2024.1337764>.
- [18] B. -W. Zhao, X. -R. Su, Y. Yang, D. -X. Li, G. -D. Li, P. -W. Hu, Y. -G. Zhao and L. Hu, Drug-disease association prediction using semantic graph and function similarity representation learning over heterogeneous information networks, *Methods* **220** (2023) 106 – 114, <https://doi.org/10.1016/j.ymeth.2023.10.014>.
- [19] V. T. Nguyen, D. H. Vu, T. K. P. Pham and T. H. Dang, CFMKGATDDA: A new collaborative filtering and multiple kernel graph attention network-based method for predicting drug-disease associations, *Intelligence-Based Medicine*, **11** (2025) #100194, <https://doi.org/10.1016/j.ibmed.2024.100194>.
- [20] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* **401** (1999) 788 – 791, <https://doi.org/10.1038/44565>.
- [21] L. Breiman, J. Friedman, R. A. Olshen and C. J. Stone, Classification and regression trees, Chapman and Hall/CRC, 2017.
- [22] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics* **12** (1970) 55 – 67.
- [23] M. T. Ribeiro, S. Singh and C. Guestrin, " Why should I trust you?" explaining the predictions of any classifier, *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016) 1135 – 1144.
- [24] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann and E. Willighagen, The Chemistry development kit (CDK): An open-source java library for chemo- and bioinformatics, *J. Chem. Inf. Comput. Sci.* **43** (2003) 493 – 500, <https://doi.org/10.1021/ci025584y>.
- [25] T. T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*, International Business Machines Corporation, New York, 1958.
- [26] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner and J. A. Leunissen, A text-mining analysis of the human phenotype, *Eur. J. Hum. Genet.* **14** (2006) 535 – 542, <https://doi.org/10.1038/sj.ejhg.5201585>.

Zahra Rafei

Department of Mathematics and Computer Science,
Amirkabir University of Technology,
Tehran, I. R. Iran
e-mail: zahra.rafei@aut.ac.ir

Seyedeh Fatemeh Hosseini

Department of Mathematics and Computer Science,
Amirkabir University of Technology,
Tehran, I. R. Iran
e-mail: sf.hosseini@aut.ac.ir

Behnam Yousefimehr

Department of Mathematics and Computer Science,
Amirkabir University of Technology,
Tehran, I. R. Iran
e-mail: behnam.y2010@aut.ac.ir

Mehdi Ghatee

Department of Mathematics and Computer Science,
Amirkabir University of Technology,
Tehran, I. R. Iran
e-mail: ghatee@aut.ac.ir