Network Modeling Analysis in Health Informatics and Bioinformatics (2025) 14:8
https://doi.org/10.1007/s13721-025-00502-8

REVIEW ARTICLE

A survey on computational methods used for drug repositioning

Eslam Abushaaban[1] · **Reda Alhajj**[1,2,3]

Received: 20 July 2024 / Revised: 1 January 2025 / Accepted: 3 January 2025

Abstract
Drug repositioning offers a promising route to increase the efficacy and streamline the development of pharmaceutical treatments. With the high costs and extensive time frames associated with traditional drug development pathways, computational approaches have risen as pivotal alternatives. This paper surveys the latest network-based methodologies in computational drug repositioning, focusing on network analysis, machine learning, and deep learning methods. By examining the predictive accuracies of various methods across consistent datasets, we present a comparative analysis that reveals the strengths and potential of each category. Our findings highlight the evolution of computational strategies, emphasized by the growing complexity and predictive capabilities of recent models, especially those leveraging deep learning frameworks.

Keywords Classification · Data imbalance · Health care · Machine learning · Drug repositioning · Deep learning

1 Introduction

Drug development has varying costs, with expenses ranging from tens of millions to hundreds of millions of dollars. These costs cover everything from drug discovery to clinical trials and approval Sertkaya et al. (2016). They can be complex, as they include both direct clinical trial expenses and long-term capital investments for preclinical research. One study found combined costs of about $1.8 billion, with $870 million as out-of-pocket expenses Paul et al. (2010). A study conducted in 2020 found that the median expense of introducing a new drug to the market was estimated to be $985 million, while the average cost stood at $1.3 billion. This figure represents a significant reduction compared to earlier estimations, which had placed the cost at $2.8 billion Wouters et al. (2020). This process is not only costly but also requires a long time to be accomplished. It can take 12–15 years to complete all phases and license a new drug Hughes et al. (2011). Hence, an alternative approach was necessary to address the lengthy and costly nature of this process. Drug repositioning or repurposing emerged as a popular and more suitable approach in the pharmaceutical industry, proving highly successful over the years.

Drugs were repositioned in various phases of their development cycle, some were already approved and marketed before being repositioned, some were repositioned during clinical trials and others were repositioned after unsuccessful clinical trials. In the 1960s, zidovudine (AZT) was being developed for its anticancer potential but failed to prove any activity in the clinical trials and was shut down. After almost 20 years, reports showed that it can stop HIV replication in vitro and was approved as an anti-HIV drug later. This was the first successful case of drug repositioning Trivedi et al. (2020). One of the oldest and most famous examples of successful drug repositioning is Aspirin (acetylsalicylic acid) which is an approved and marketed analgesic drug but in the 1980s was repositioned as an antiplatelet aggregation drug at low doses only Monneret and Bohuon (2009). Another very famous drug repurposing example is sildenafil, a drug developed originally to treat Angina which refers to chest pain or discomfort experienced when a portion of the heart muscle receives an inadequate supply of oxygen-rich blood but during phase 1 trials it was found that it can cause sustained erection in healthy males. So, now it is one of the most successful drugs for erectile dysfunction after being repositioned Boolell et al. (1996).

✉ Reda Alhajj
alhajj@ucalgary.ca

[1] Department of Computer Engineering, Istanbul Medipol University, Istanbul, Turkey

[2] Department of Computer Science, University of Calgary, Alberta, Canada

[3] Department of Health Informatics, University of Southern Denmark, Odense, Denmark

Published online: 14 January 2025

To overcome the drawbacks of traditional methods, the computational methods started to be adopted to achieve faster and cheaper ways for drug repositioning Shim and Liu (2014). With the emergence of genome-wide association studies Sanseau et al. (2012), computational methods can be scalable to include drugs, diseases, genes, targets or proteins. Computational drug repositioning methods can be separated into many categories, but what concerns us here is the network analysis-based techniques Lotfi Shahreza et al. (2018). These techniques take advantage of the links found between drugs and diseases or targets in a network structure which can be very informative in the analysis process Morselli Gysi et al. (2021).

Initially, the primary focus of computational drug repositioning revolved around uncovering interactions between drugs and their molecular targets. The presence of a drug-target interaction was pivotal as it offered substantial insights into drug repositioning Yıldırım et al. (2007). This concept came from the belief that numerous drugs have interactions with many targets Hopkins (2008). The general idea for network analysis-based drug repositioning is based on the assumption that genes responsible for the same disease will be found Near to one another within a network Ata et al. (2021). And this idea can also be applicable in drug-disease networks. Heterogeneous networks -which are characterized by containing two or more types of nodes and their links- are mainly used in network-based approaches, where several features of drugs and diseases can be incorporated alongside genes, proteins or targets interactions.

With the rise of machine learning techniques popularity and their high accuracy results, they were adopted to create new models that can predict new unknown drug-disease relations with higher accuracy and more reliability. Some examples of machine learning techniques used: feature-based classification (Yang and Agarwal 2011; Wang et al. 2013; Oh et al. 2014), label propagation Zhang et al. (2018) and matrix factorization (Zhang et al. 2014; Xuan et al. 2019; Zhang et al. 2018). Then, the novel deep learning approaches were introduced with their superior performance in many areas like face recognition Vellal et al. (2021) and computational biology (Angermueller et al. 2016; Yue et al. 2020), there was a shift towards them. Multimodal deep autoencoders Hong et al. (2015), variational autoencoders Kingma et al. (2019) and graph convolutional networks Kipf and Welling (2016) are some of the deep learning techniques that showed increased potential recently.

This survey's primary objective is to summarize recent network analysis-based approaches in predicting drug-disease relations for computational drug repurposing and to fairly set side by side their predictive accuracies using consistent datasets. Here, we discuss three categories of network-based drug-disease association prediction: Network Analysis, Machine Learning and Deep Learning techniques.

2 Related work

In 2022, Yoonbee Kim et al. (2022) conducted a review on computational methods for predicting drug-disease relations divided into three groups: Graph Mining, Matrix Factorization and Deep Learning. They started by discussing the methods covered, talking about the general methodology and where the data used came from. They covered several methods throughout the three categories. For graph-mining algorithms, they used the MBiRW, TP-NRWRH, DR-IBRW, and BGMSDDA methods; for matrix factorization they used the DRRS, OMC, DRIMC, and MSBMF methods; and for deep learning, they used the deepDR, ANMF and LAGCN methods.

They aimed to compare these methods by testing their performance on a dataset created for this experiment. Initially, they established a heterogeneous network linking drugs, genes, and diseases. This network was constructed by computing similarities between drug pairs, disease pairs, and gene pairs. They computed drug-drug similarity using two distinct measures: one based on their chemical structures, labeled as network-CS, and the other based on the Anatomical Therapeutic Chemical (ATC) codes Olson and Singh (2017), labeled as network-ATC. Disease-disease and gene-gene similarities were assessed using a semantic similarity metric utilizing the Human Phenotype Ontology (HPO) Köhler et al. (2020) and the Gene Ontology (GO) Consortium et al. (2020) annotation datasets, respectively. Subsequently, they constructed a tripartite graph incorporating associations between drugs and diseases, drugs and genes, and diseases and genes. This tripartite graph was then integrated with the drug-drug, disease-disease, and gene-gene similarities to form a unified weighted heterogeneous network. In their experimentation, they utilized drug-disease associations from the C-dataset, an expansion of the gold standard dataset utilized in prior drug repositioning studies (F-dataset). They evaluated the prediction of drug-disease associations separately on the drug and disease sides. Prediction on the drug side aimed to forecast additional diseases that could be treated by each drug, while prediction on the disease side aimed to identify drugs with the potential to treat each disease. To do this evaluation, they used 10 fold cross-validation and compared AUC and AUPR scores of each method. They also adjusted the precision score calculations by using True Positive Rate and False Positive Rate instead of just using True Positive and False Positive values. This step was to overcome the imbalanced and sparse nature of the associations. After comparing the results they also performed robustness and efficiency tests to conclude their experiment. However, they didn't discuss the datasets used in the covered methods and their important effect on the model performance.

In 2023, Chunyan Ao et al. (2023) provided a review of computational techniques for predicting drug-disease relations categorized into four categories: neural network, matrix-based prediction, recommendation algorithms, and methods based on text analytics and language intelligence. They discussed the algorithm overview of the four groups and gave a brief summary about the various methods covered in each group. Then they delved into some detailed overview about the best method in each category. After that, they mentioned the advantages and disadvantages of each group.

Next, they designed an experiment to test a few chosen models from each category where they used the C-dataset to evaluate and compare these methods. They performed 10 fold cross-validation and relied on AUC and AUPR scores to measure and compare model performances. Then, they discussed some of the problems and important aspects regarding the drug-disease association prediction field. They mentioned the data imbalance problem alongside the feature fusion and scalability importance. However, in their comparison they made a general assumption about the whole category based on a small number of models, which can be biased or not very accurate. They also did not discuss the datasets used in the models.

3 Datasets review

Several datasets were used by the methods covered here, but most of the methods used a gold standard data known as F-dataset obtained from the PREDICT method Gottlieb et al. (2011) for either training or evaluation. This dataset is large and consists of 1933 relations between 313 diseases and 593 drugs. It includes drug data where drugs were extracted from DrugBank Wishart et al. (2008), and drug feature information, like targets, indications, and molecular input and side effects, were integrated from various sources, such as canonical simplified molecular input line entry specification (SMILES) Weininger (1988), Matador Günther et al. (2008) and KEGG DRUG Kanehisa and Goto (2000) databases, DCDB Liu et al. (2010) and SIDER Kuhn et al. (2010). It also contains biological interaction data like protein sequence, protein-protein interaction information and gene ontology (GO) annotations from UniProt Jain et al. (2009). Besides, it has clinical trials data and disease signature data from gene expression profiles associated with diseases Parkinson et al. (2009). Diseases were obtained from OMIM Hamosh (2002) and drug-disease associations were assembled.

Another dataset known as DN-dataset was obtained from the DrugNet method Martinez et al. (2015). It consists of a drug network containing 1490 drugs from DrugBank 3.0 Knox et al. (2010) with drug similarity based on ATC codes WHO (2023), a disease network containing 4517 diseases Built utilizing the disease ontology (DO) Schriml et al. (2018) and a protein network of 136867 unique interactions between 18107 proteins extracted from BioGRID 3.2 Chatr-Aryamontri et al. (2012). From these three networks drug-disease, disease-protein and drug-protein association networks were constructed.

Using a combination of the previous two datasets, a broader set known as C-dataset was constructed in the MBiRW method Luo et al. (2016). It comprises 663 drugs recorded in DrugBank, 409 diseases documented in the OMIM database, and 2352 established drug-disease associations. Drug similarity relies on chemical structure, while disease similarity is derived from phenotypes using MimMiner Kohn et al. (2006). Table 1 summarizes the reviewed datasets, detailing their structure, data sources, and the features derived from each.

The three datasets (F-Dataset, DN-Dataset, and C-Dataset) demonstrate several common strengths and limitations in their approach to drug-disease association prediction. Key strengths include the integration of multiple data sources and types, which provides comprehensive coverage of drugs, diseases, and their associations. This integration allows for the creation of diverse similarity measures, capturing various aspects of drug-drug and disease-disease relationships. The datasets also

Table 1 Dataset Review Summary

Dataset	Structure	Data Sources	Features Derived
F-dataset	1933 relations between 313 diseases and 593 drugs	DrugBank, OMIM, Matador, KEGG DRUG, DCDB, SIDER, UniProt, ArrayExpress	Drug features (targets, indications, molecular input, side effects), biological interactions, clinical trials data
DN-dataset	Drug network: 1490 drugs	DrugBank 3.0, ATC codes	Drug similarity based on ATC codes
	Disease network: 4517 diseases	Disease Ontology (DO)	disease similarity based on DO
	Protein network: 18107 proteins, 136867 interactions	BioGRID 3.2	protein-protein interactions
C-dataset	2352 associations between 663 drugs and 409 diseases	DrugBank, OMIM, MimMiner	Drug similarity based on chemical structure, disease similarity based on phenotypes

benefit from up-to-date information, including data from clinical trials and approved drugs. However, they share some limitations, such as potential biases towards well-studied drugs and diseases, incomplete data for certain entries, and the presence of noise in high-throughput data. The reliance on existing annotations and the lack of negative examples in some cases may limit the datasets' ability to capture the full spectrum of drug-disease relationships. Additionally, while the datasets excel in representing known associations, they may struggle with emerging or less-studied drug-disease pairs. Despite these limitations, the diverse approaches and data sources used across these datasets provide a robust foundation for drug repurposing and indication prediction research.

4 Review criteria

In conducting this survey, we employed the following criteria to select the computational methods for review:

- Relevance: We focused on methods specifically designed for drug-disease association prediction or drug repositioning.
- Novelty: We prioritized methods that introduced new approaches or significant improvements to existing techniques.
- Diversity: We aimed to cover a broad spectrum of approaches within each category (Network Analysis, Machine Learning, and Deep Learning) to provide a comprehensive overview of the field.
- Recency: While we included some seminal older works, we emphasized more recent methods (published within the last 5-7 years) to reflect current trends in the field.
- Performance: We selected methods that demonstrated competitive performance on benchmark datasets or showed promising results in their respective studies.

By applying these criteria, we aimed to provide a representative and informative survey of the most significant and innovative computational methods in drug-disease association prediction. We acknowledge that this selection may not be exhaustive, but we believe it captures the key developments and trends in the field.

5 Review of network analysis based methods

Network analysis in drug repositioning is characterized by its ability to integrate diverse biological data, identify disease-related molecular clusters, map drug-target interactions, and provide predictive modeling for new drug-disease relationships. By dissecting the complex architecture of biological networks-spanning genes, proteins, and metabolic pathways-this approach illuminates how drugs interact within the vast molecular landscape. Central to its utility is the ability to pinpoint critical nodes and pathways that play pivotal roles in disease processes, suggesting that drugs targeting these key points may be effective against other, mechanistically similar conditions. It offers a system-level understanding of diseases, facilitating the discovery of multi-target therapeutic strategies. Network analysis goes further by identifying disease-specific modules within the network, clusters of interconnected elements that are disproportionately affected by certain conditions, hinting at novel repositioning opportunities. Moreover, by analyzing network communities or clusters, it can reveal new therapeutic targets within tightly knit molecular conglomerates. This approach enhances the efficiency and cost-effectiveness of drug development by leveraging existing drugs and data, reducing the need for extensive safety trials. Network analysis is dynamic, evolving with new research and technologies, making it a potent tool for uncovering novel therapeutic uses for approved drugs. As summarized in Table 2, In this section, we go through the primary aspects associated with methods based on network analysis.

5.1 TL-HGBI (2014)

– Dataset: Drugs with integrated information about chemical structures and targets with the sequences of amino acids within their respective proteins. Disease phenotypes were used. Similarity profiles of drug-drug, disease-disease and target-target networks were calculated. Drug-target and drug-disease interaction network was also constructed. Drug-disease relations were acquired from F-dataset.
– Methodology: In 2013 Wang W et al. used drugs and targets to create a two-layer model Wang et al. (2013) that can also be used for drugs and diseases but can't be immediately extended to a three-layer model. So, they then proposed a three layer model Wang et al. (2014) that used drugs, diseases and targets to create a heterogeneous network. They used a network propagation iterative algorithm that calculates proximity scores between drugs and diseases and is utilized for ranking potential drugs. The algorithm updates scores based on the graph structure until convergence. For that, they constructed a heterogeneous network and calculated the intraconnection similarity between the same node type i.e drug-drug, disease-disease and target-target.

Table 2 Network Analysis Methods Overview

Method	Dataset	Data Source
TL-HGBI (2014)	5080 diseases, 1409 drugs and 3989 targets. 1461 associations between 549 drugs and 233 diseases, and 2098 associations between 554 drugs and 602 targets	DrugBank, SMILES, ENSEMBL, InterPro-BLAST, OMIM, Mesh, MimMiner
DrugNet (2015)	1490 drugs, 4517 diseases and 18107 proteins. 1008 drug-disease relations, 4026 drug-protein relations and 11658 disease-protein relations	DrugBank 3.0, ATC, DO, BioGRID 3.2, DGA
MBiRW (2016)	F-dataset and C-dataset: 2352 drug-disease associations between 663 drugs and 409 diseases	DrugBank, OMIM
TP-NRWRH (2016)	NA	SMILES, Mesh, OMIM, MimMiner
NTSIM (2017)	18416 drug-disease associations between 269 drugs and 598 diseases	CTD, DrugBank, KEEG, MeSH
DR-IBRW (2018)	718 drugs, 525 diseases and 2177 drug-disease associations	MeSH, PubChem, supporting Material from literature
BGMSDDA (2021)	F-dataset, C-dataset	DrugBank, OMIM

5.2 DrugNet (2015)

- Dataset: Created DN-dataset. Drug network with therapeutic chemical codes, disease network with disease annotations and protein network. Drug-disease, disease-protein and drug-protein interaction networks were also created.
- Methodology: Martinez et al. (2015) suggested a prioritization method based on network analysis. It was derived from the ProphNet method Martínez et al. (2014) which employs a propagation flow algorithm which prioritizes entities according to their interconnection in a complex network. They used drug, disease and protein datasets to construct a global graph that consists of two types of networks: networks representing interactions among elements of identical types (e.g., drug-drug) and networks representing interactions among elements of diverse types (e.g., drug-disease). The propagation algorithm measures the global distances within a network and then ranks the entities in the target network by iteratively computing scores for nodes in the target network.

5.3 MBiRW (2016)

- Dataset: Used the drug-disease associations from F-dataset and used DN-dataset for validation and testing. Also created C-dataset for further evaluation and validation.
- Methodology: Huimin Luo et al. (2016) proposed a novel approach that integrates a thorough similarity measure with a bi-directional random walk algorithm to predict possible drug repositioning. They used correlation analysis in order to overcome insufficient resemblances lacking informative value. They constructed a Drug-Disease Heterogeneous Network that connects drugs and diseases via known associations and applied the BiRW algorithm to prioritize candidate diseases for each drug based on network traversal probabilities.

5.4 TP-NRWRH (2016)

- Dataset: Drug network with chemical fingerprint and disease network with phenotypes. Disease-disease and drug-drug similarities were computed. Drug-disease association network was also created. Used F-dataset and C-dataset for evaluation.
- Methodology: Liu et al. (2016) suggested a method that adopts two pass random walk algorithm with restart on a heterogeneous network of drugs and diseases. They built the heterogeneous network using disease-disease and drug-drug similarity networks alongside the known drug-disease relation network. They used random walkers focused on drugs and diseases which combined provide complementary information about possible drug-disease relations by computing the mean probabilities of reaching specific nodes.

5.5 NTSIM (2017)

- Dataset: Drugs with integrated information about drug-drug interactions, substructures, enzymes, targets and pathways and diseases integrated with MeSH descriptors. Similarity profiles were calculated. Used F-dataset, LRSSL dataset dataset and TL-HGBI for validation and comparison.
- Methodology: Huang et al. (2020) suggested a novel computational technique that only uses known drug-disease relations to predict unknown associations. They also proposed Linear Neighborhood Similarity (LNS) which is a new similarity measure to calculate similarities in the drug-disease bipartite network based on asso-

ciation profiles. Then, the prediction of new associations is carried out using the label propagation process. The inference method based on network topological similarity combines both drug-drug similarity and disease-disease similarity based inference methods which results in better performance.

5.6 DR-IBRW (2018)

- Dataset: Drugs with integrated information about chemical fingerprints and diseases with integrated symptoms. Similarity profiles were calculated. Used F-dataset and C-dataset for validation.
- Methodology: Wang et al. (2019) proposed a novel method that applies a bi-random walk algorithm with restart on a heterogeneous network assigning quantified walk-length for each node individually. The proposed drug repurposing approach focuses on individual bi-random walks within the heterogeneous network. Unlike traditional methods that use fixed walk-lengths, DR-IBRW calculates individual walk-lengths for each node (drugs and diseases) based on their network topology and relationships. It uses similarity measures for diseases and drugs and the bi-random walk with restart procedure enables the prediction of probabilistic association between drugs and new diseases.

5.7 BGMSDDA (2021)

- Dataset: Used both F-dataset and C-dataset.
- Methodology: Xie et al. (2021) proposed a novel computational approach integrating drug-disease similarities to predict new associations. They employed Weighted K Nearest Known Neighbors (WKNKN) algorithm to enhance the inadequacy of the original drug-disease interaction matrix by inferring potential associations based on similarity. To capture relationships between drugs and diseases both linear and nonlinear, they used two similarity measures: Linear Neighborhood Similarity (LNS) and Gaussian Kernel Similarity. The core technique involves modeling drug-disease relationships as a bipartite graph and using a two-step diffusion method for predicting associations. This approach avoids the "cold start" problem and improves prediction accuracy.

each review section contains a table at the end for comparing the datasets used in details

6 Review on machine learning based methods

Machine learning in drug repositioning is characterized by its ability to recognize complex patterns in vast datasets, thereby offering insights driven by actual data and advanced analytics to pinpoint new applications for drugs. Its capacity to manage and interpret the ever-expanding volumes of data across the biomedical landscape enables a scalable approach to drug discovery. Machine learning can integrate various data types, supporting a comprehensive understanding of disease and drug mechanisms. It enables personalized drug repositioning by considering individual genetic variations, and streamlines the drug discovery process through automation and high-throughput analysis. These capabilities significantly enhance the efficiency and effectiveness of identifying repurposing opportunities for existing drugs, opening up new therapeutic avenues with heightened precision and reduced development times. As reported in Table 3, in this section, we go through the main aspects of the machine learning based methods.

6.1 PREDICT (2011)

- Dataset: Created F-dataset. Drugs with integrated information about chemical structure, drug target, approved indications and side-effect info, gene ontology and clinical trials data. Diseases with information about gene

Table 3 ML Methods Overview

Method	Dataset	Data Source
PREDICT (2011)	1933 drug-disease associations between 593 drugs, 313 disease	DrugBank, OMIM, GO annotations, SIDER
PreDR (2013)	NA	DrugBank, PubChem, SIDER, KEGG BRITE, BRENDA, SuperTarget, OMIM, MimMiner
TMSDA (2014)	377 drugs, 80 diseases and 1295 drug-disease associations	DrugBank, OMIM, UniPort, CTD
SKMF (2016)	F-dataset	DrugBank, SMILES, SIDER, OMIM, MeSH, GO annotations
LRSSL (2017)	763 drugs and 681 diseases	PubChem, SIDER, InterPro, GO, UniPort
HED (2019)	Dataset I: 301 drugs, 1142 diseases and 4236 drug-disease associations Dataset II: F-Dataset	DrugBank, CTD

expression profiles associated with diseases. Similarity profiles were calculated.
– Methodology: Gottlieb et al. (2011) suggested an approach to predict new relations between drugs and diseases using similarity measures between drugs and diseases and assess the association evidence through a logistic regression model. The algorithm has three phases: (1) Development of measures for drug-drug and disease-disease similarity. (2) Creating classification features by combining drug and disease similarity measures. These features can indicate the similarity with the closest known association for each unknown drug-disease pair and then (3) employing a logistic regression model to automatically determine the weight of various features, resulting in a classification score.

6.2 PreDR (2013)

– Dataset: Drugs with integrated information about chemical structure, side-effect data and drug-target interactions. Diseases with phenotypic information. Similarity profiles for information in the drug network were calculated along with disease similarity profiles. Used F-dataset.
– Methodology: Wang et al. (2013) proposed a comprehensive predictor for drug repurposing which constructs drug similarity profiles based on the collected data then calculates similarity scores using appropriate distance metrics. They used a kernel function to integrate the different drug similarity profiles into a unified representation. Then they constructed the integrated similarity matrix by combining the individual similarity matrices through a weighted sum. To detect the new drug-disease relations, the integrated similarity matrix is used as the input for a linear SVM-based binary classification model to learn the decision boundary between positive and negative examples.

6.3 TMSDA (2014)

– Dataset: Integrative genetic network combining three different groups of gene and protein networks. Drugs with integrated information about their targets and also, diseases with integrated information about their susceptible genes.
– Methodology: Oh et al. (2014) suggested a new method named “Two Methods for Scoring Drug-disease Association” which uses scoring techniques for quantifying the relationships between drugs and diseases. This method consists of a two-stage prediction process: (1) Scoring Associations: Employing adjacency-based and module-distance-based inferences. The former relies on network proximity, while the latter identifies shared biological modules. (2) Characterizing Relationships via Features: Converting scores into features to represent drug-disease relationships and using these features in a prediction classifier. They used three classifiers -C4.5, MLP and RF- and compared their results.

6.4 SKMF (2016)

– Dataset: Drug-disease network incorporating characteristics like side effects, chemical structure, gene ontology, sequence alignment, protein-protein interactions (PPI) network and phenotype of both diseases and drugs. Drug-disease relations were acquired from F-dataset.
– Methodology: Moghadam et al. (2016) suggested a novel fusion method called “Scored Mean Kernel Fusion” that combines features using two fusion techniques: (1) Scored Mean Fusion, a new technique to assign scores to each feature based on their importance and combine features using aggregation operators to integrate diverse information from multiple sources. (2) Kernel Fusion which integrates drug and disease information into a single kernel function and applies a tensor product-based kernel to calculate similarity between drug-disease pairs. The scored mean fusion is used prior to the classification while the kernel fusion is used as the kernel function for the C-support vector classifier that is used for prediction of drug-disease associations.

6.5 LRSSL (2017)

– Dataset: Drugs with integrated information about chemical fingerprints and side effect information and diseases with integrated information about protein domains and gene ontology of targets.
– Methodology: Liang et al. (2017) proposed a new computational technique named "Laplacian Regularized Sparse Subspace Learning" which is a machine learning method that combines sparse learning with graph Laplacian regularization. It is used to identify important drug features and predict drug-disease associations. They integrated data from various sources in order to create drug feature profiles then applied feature selection techniques to identify relevant drug features. Then they utilized L1-regularized regression models to predict relationships between drugs and diseases. This allows the model to identify the most relevant features and their contributions to the predictions.

6.6 HED (2019)

– Dataset: Dataset I: drug and disease associations in the form of chemicals mapped to drugs. Dataset II: F-dataset.

– Methodology: Yang et al. (2019) proposed a new approach named "Heterogeneous network Embedding for Drug-disease association" which employs the metapath-2vec technique for node embedding in a heterogeneous network. They constructed a heterogeneous network with drug and disease similarities and represented drug-disease relations using adjacency matrices creating a comprehensive framework for modeling drug-disease interactions. Then they utilized the metapath2vec approach, which uses metapaths (sequences of node types) to guide random walks in the network. The random walk results are then used to generate feature vectors for nodes. These feature vectors were used as input for an SVM model to predict new drug-disease relations.

7 Review on deep learning methods

Deep learning's application in drug repositioning is distinguished by its sophisticated neural networks that navigate the complexity of biological data, capturing hidden patterns and nonlinear relationships crucial for identifying novel drug uses. It has the capability to autonomously learn and discern important features directly from unprocessed data, removing the necessity for manual feature engineering. Its success in managing high-dimensional inputs, such as molecular structures or genomic information, enables the exploration of drug-disease interactions. Deep learning algorithms are particularly proficient at synthesizing information from varied data types, ranging from structured biochemical properties to unstructured textual data in scientific literature, providing a complete view of potential repositioning opportunities. The accuracy of deep learning models, supported by architectures like CNNs and RNNs, lends confidence to the predictions of drug efficacy and safety profiles. Scalability is another hallmark, with the ability to extend analysis across vast datasets, facilitated by efficient parallel processing. As summarized in Table 4, in this section, we go through the main aspects of the deep learning based methods. The discussed methods didn't explicitly discuss the technical details of the architecture of their models. They explained the model and how it works and the overall flow of the methods and their integration.

7.1 deepDR (2019)

– Dataset: Drug-disease network integrated with nine feature networks: drug-drug and drug-target interactions, drug-side-effect relations, chemical and therapeutic similarities, drugs' target sequence similarities and biological process, cellular component and molecular function derived from Gene Ontology (GO).
– Methodology: Zeng et al. (2019) proposed a new technique to merge various information derived from various types of networks and predict novel drug-disease relations by utilizing collective variational autoencoder (cVAE). They collected diverse datasets to construct multiple networks integrating information from nine heterogeneous networks. They implemented a method based on random walk algorithm with restart (RWR) to the networks to capture structural information and computed transition probabilities between network vertices to reveal relationships between nodes. Then, they employed Network Fusion through a Multi-Modal Deep Autoencoder (MDA) to generate a low-dimensional feature representation that combines information from multiple networks. Finally, they utilized the features that were extracted from MDA as additional information for drugs and employed a cVAE model to predict new drug-disease relations by encoding and decoding drug-disease associations and drug features through shared networks.

Table 4 DL Methods Overview

Method	Dataset	Data Source
deepDR (2019)	1519 drugs, 1229 diseases and 6677 drug-disease connections	DrugBank, repoDB, MeSH, UMLS
HNRD (2019)	F-dataset, C-dataset, DN-dataset	DrugBank, OMIM, GO annotations, SIDER
LAGCN (2021)	18416 drug-disease associations between 269 drugs and 598 diseases and another 6244 therapeutic associations	CTD, DrugBank, MeSH
SNF-NN (2021)	867 drugs, 803 diseases, and 8684 drug-disease associations	DrugBank, repoDB
HINGRL (2022)	18416 drug-disease associations between 269 drugs and 598 diseases, 11107 drug-protein associations between 969 drugs and 613 proteins, 25087 protein-disease associations between 832 proteins and 692 diseases	CTD, DrugBank, DisGeNET
LBMFF (2023)	18416 drug-disease associations between 269 drugs and 598 diseases, 673665 full-text scientific literature	CTD, DrugBank, SIDER (2023), SMILES, MeSH

7.2 HNRD (2019)

- Dataset: Used F-dataset. Also used C-dataset and DN-dataset for validation.
- Methodology: Yingdong et al. (2019a) suggested a neural network based technique that incorporates neighborhood information in a heterogeneous network. The methodology starts by representing drugs and diseases as nodes in a heterogeneous network. It aggregates features for each node from its neighboring nodes. Feature representations are learned using a neural network with activation functions like ReLU, to learn structural and topological information, resulting in lower-dimensional vectors. The methodology treats the prediction task as a matrix completion problem, aiming to fill missing associations in drug-disease matrices. A neural network is trained to minimize the difference between reconstructed and initial matrices, using projection matrices to extract principal features. Gradient descent is employed for training, taking advantage of differentiable operations. The model can predict drug-disease associations based on learned representations.

7.3 LAGCN (2021)

- Dataset: Used SCMFDD dataset which consists of a drug network integrated with information about targets, enzymes, drug-drug interactions, pathways and substructures and disease network. Similarity profiles were calculated. Used deepDR dataset for validation and comparison.
- Methodology: Yingdong et al. (2019b) proposed a novel approach that employs GCN to learn node embeddings in a heterogeneous network. GCN aggregates information from neighboring nodes to update node embeddings. They integrated drug-diseases associations and drug and disease similarities into a heterogeneous network. They introduced a new computational technique called "Layer Attention Graph Convolutional Network". It extends GCN with an attention mechanism, which adapts to the importance of embeddings from different layers, capturing structural information at various orders. To reconstruct the drug-disease interaction matrix, they utilized a bilinear decoder. It uses drug and disease embeddings to predict association scores. The model is trained using weighted cross-entropy loss with Adam optimizer. Node dropout and regular dropout techniques are applied to prevent overfitting.

7.4 SNF-NN (2021)

- Dataset: Created SND dataset which consists of three distinct types of information, drug-disease interaction information, drug-related similarity information, and disease-related similarity information. The drug-related similarity information was collected from 10 different networks, while the disease-related similarity information was collected from 14 different networks. Similarity profiles were calculated and other datasets -namely C-dataset and LRSSL dataset- were used for validation.
- Methodology: Jarada et al. (2021) proposed an integrative approach that uses Similarity Network Fusion and Neural Network deep learning to predict new unknown drug-disease associations with improved accuracy. This methodology calculates the similarity measures that are used to quantify the drug-disease relationships in a heterogeneous network. Then, it uses a similarity selection process to choose the most relevant informative similarity measures. After that it uses a non-linear process to update the highly informative similarity matrices resulted previously into one that captures the shared information across matrices. It then uses a feed forward neural network to capture the information flow through the network where the hidden layers number and the choice of activation functions affects the model performance.

7.5 HINGRL (2022)

- Dataset: Adopted B-dataset which consists of three types of biological networks, as drug-disease, disease-protein and drug-protein relations. Used SCMFDD dataset for drug-disease relations along with drug-protein and disease-protein relation networks. Also used F-dataset integrated with drug-protein and disease-protein relations for validation.
- Methodology: Zhao et al. (2021) proposed a new approach that utilizes network topology and biological knowledge regarding drugs and diseases for the purpose of drug repositioning. First, they combined different biological networks (drug-disease, disease-protein and drug-protein) to create a Heterogeneous Information Network (HIN). These networks are enriched with biological knowledge about drugs and diseases. Then, they used autoencoders to reduce the dimensionality of drug and disease attributes, helping to capture essential information while reducing noise. They then applied Heterogeneous Network Representation Learning where the DeepWalk algorithm captures network topology information, generating continuous vector representations that preserve connectivity patterns in the HIN. Finally, they combined the reduced biological information and network representations into an integrated feature matrix.

A Random Forest model was then applied to predict unknown drug-disease relations using these features.

7.6 LBMFF (2023)

- Dataset: SCMFDD dataset integrated with drug chemical structures, drug-side-effect interactions, drug-target interactions and diseases tree numbers. Similarity profiles were calculated. In addition, full-text scientific literature where drugs or diseases from the benchmark dataset appeared in the titles or abstracts. This extensive literature was used as a corpus for the semantic similarity calculations. Also, B-dataset and TL-HGBI dataset were used for validation.
- Methodology: Kang et al. (2023) suggested a new method called Literature Based Multi-Feature Fusion (LBMFF) to predict drug-disease relations. This method used not only integration of various heterogeneous biological interactions like (drugs, diseases, side effects and targets), but also semantic embeddings extraction from scientific literature. For that, they applied feature extraction to capture the features from the different interaction networks like (drugs, diseases, drug-disease, drug-side-effect and drug-target). Also, they used literature semantic similarity based on BERT which is pre-trained on a vast scientific literature corpus and fine-tuned for semantic similarity between drugs and diseases. Then, they applied a similarity matrix fusion process by optimizing fusion coefficients for the different similarity matrices to combine them effectively. After that they constructed an association feature representation matrix based on drug-disease associations and the fused similarity matrices. Finally, they predicted the unknown associations using GCN where the drug-disease relations adjacency matrix is introduced into a GCN encoder to extract embeddings for drugs and diseases. An attention mechanism is then added to adaptively adjust different GCN layers' weights. Sigmoid activation is used in the GCN decoder to predict drug-disease relations.

8 Result comparison

Our approach to comparing methods across different categories (Network Analysis, Machine Learning, and Deep Learning) varied due to the diverse nature of datasets and evaluation methods used in the original studies:

- Network Analysis and Machine Learning Methods: For these categories, we initially attempted to unify the evaluation by using scores obtained from evaluating methods on a gold standard dataset (F-dataset). However, due to the variability in implementation details and the potential for reproduction discrepancies, we ultimately decided to use the published AUC scores from the original papers for consistency.
- Deep Learning Methods: Given the wide range of curated and standard datasets used in deep learning approaches, a unified evaluation was not feasible. Therefore, we relied on the published AUC scores reported by the authors of each method.

It's important to note that we did not re-run the methods ourselves due to the complexity and computational resources required to replicate each study accurately. Instead, we compiled and compared the reported results from the original publications. This approach allows for a broad overview of method performance but may be subject to variations in evaluation protocols across different studies. We acknowledge that this approach has limitations, particularly in ensuring direct comparability across all methods. However, it provides a general landscape of performance across different approaches in the field of drug-disease association prediction.

Table 5 Network Analysis Methods Comparison

Method	Algorithm	Direct Comparison	AUC Scores
TL-HGBI (2014)	Network propagation		0.915
DrugNet (2015)	Prioritization method that implements network propagation algorithm	PREDICT	0.779
MBiRW (2016)	Bi-directional random walk	DrugNet	0.917
TP-NRWRH (2016)	Two path random walk with restart	DrugNet, MBiRW	0.939
NTSIM (2017)	Propagation and Linear neighborhood similarity is used	TL-HGBI, LRSSL, PREDICT	0.928
DR-IBRW (2018)	Bi-random walk with restart with quantification of individual walk length for each node	MBiRW	**0.955**
BGMSDDA (2021)	Two part graph diffusion	MBiRW, DrugNet	0.939

8.1 Network analysis methods comparison

The comparison detailed in Table 5, delves into Network-based methods, outlining the algorithms, direct comparisons, and corresponding AUC scores. The purpose is to provide an in-depth analysis of how different methods leverage network structures for drug-disease association predictions. This elucidates the effectiveness of algorithms in utilizing network information and their comparative performance.

the results section also has several tabels to compare the results obtained from the review

To compare network analysis based methods, we tried to unify the evaluation of these methods and used the scores obtained by evaluating the methods on the gold standard dataset. We found that DR-IBRW is the best performing method with AUC of 0.955. We also noticed that overall the random walk based methods tend to perform better with some variations due to the application of the algorithm. The random walk with restart seems to be more effective and produce more robust results.

8.2 Machine learning methods comparison

Table 6 offers a detailed examination of ML methods, encompassing algorithm specifics, direct comparisons, and AUC scores. This analysis aims to shed light on applying traditional machine learning methods in drug-disease association prediction. By exploring various algorithms and their performance metrics, researchers can discern the efficacy of ML-based approaches.

To compare machine learning-based methods, we attempted to replicate our approach from the network analysis comparisons. We used scores obtained by evaluating the methods on the gold standard dataset. We found that all the methods used in the comparison performed very closely despite the application of various algorithms. We noticed that LRSSL and TMSDA (RF) had slightly better performance than other methods. Also, the methods that used SVM had different scores depending on the adaptation of the algorithm. SKMF used scored mean kernel fusion with SVM which resulted in better performance than the other SVM based methods.

8.3 Deep learning methods comparison

Table 7 focuses on DL methods, presenting algorithm details, direct comparisons, and AUC scores per dataset. This comparison aims to highlight the evolving landscape of deep learning techniques in predicting drug-disease relations. It provides insights into the utilization of complex models and their performance across different datasets.

To compare the deep learning based methods, we couldn't follow what we applied for the previous two comparisons because the methods in this comparison used a wide range of curated and standard datasets. Therefore, we used the scores published by the methods to gain an overview of their performance. We found that representation learning based methods had the best performance with HNRD as the best performing method with AUC score of 0.942 followed by HINGRL with a score of 0.936.

Table 6 ML Methods Comparison

Method	Algorithm	Direct Comparison	AUC Scores
PREDICT (2011)	Logistic regression		0.90
PreDR (2013)	SVM	PREDICT	0.865
TMSDA (2014)	Random Forest (RF), MLP, C4.5	PREDICT	**0.917**
SKMF (2016)	Scored mean kernel fusion with SVM	PREDICT, PreDR	0.91
LRSSL (2017)	Laplacian regularized sparse space learning, L1 regularized regression	TL-HGBI, PreDR	**0.917**
HED (2019)	Metapath2vec with SVM		0.899

Table 7 DL Methods Comparison

Method	Algorithm	Direct Comparison	AUC Scores
deepDR (2019)	Network fusion via multi-modal deep autoencoder then using collective variational autoencoder		0.855
HNRD (2019)	Representation learning with deep learning model of neural network	MBiRW, DrugNet	**0.942**
LAGCN (2021)	GCN	deepDR, TL-HGBI	0.890
SNF-NN (2021)	Similarity network fusion with neural network application	MBiRW	0.867
HINGRL (2022)	Graph representation learning with Deepwalk	LAGCN, deepDR	0.936
LBMFF (2023)	Combined drug-disease embeddings with semantic features and applied GCN	LAGCN	0.882

Table 8 Direct Comparison Matrix

Methods	Network Analysis						Machine Learning				Deep Learning				
	DrugNet	MBiRW	TP-NRW	NTSIM	DR-IBRW	BGMSDA	PreDR	TMSDA	SKMF	LRSLL	HNRD	LAGCN	SNF-NN	HINGRL	LBMFF
TL-HGBI			X	X						X		X			
DrugNet		X	X			X					X				
MBiRW					X	X					X		X		
PREDICT	X			X			X	X	X						
PreDR							X		X	X					
LRSSL				X											
deepDR												X		X	
LAGCN														X	X

Table 9 Superior Methods

Category	Superior Method	AUC Score
Network Analysis	DR-IBRW (2018)	0.955
Machine Learning	LRSSL (2017)	0.917
Deep learning	HNRD (2019)	0.942

8.4 Direct comparison matrix

The Direct Comparison Matrix reported in Table 8 provides a visual representation of how different methods were compared, highlighting the authors' benchmarking against previous studies. This matrix serves as a strength for evaluating the contributions of each work, emphasizing the importance of contextualizing new methodologies within the landscape of existing research. By offering a quick overview of the relationships between various approaches, it aids in understanding the evolution of methodologies over time. This comprehensive comparison strengthens the assessment of each method by considering its performance relative to established benchmarks.

The matrix is organized into three main categories: Network Analysis (net), Machine Learning (ML), and Deep Learning (DL). Each row and column in the matrix corresponds to a specific method, while the cells indicate whether a direct comparison was made between the respective methods. The symbols in the matrix cells indicate whether a direct comparison was made between the corresponding methods. This matrix aids in comprehending the comparative performance of different approaches, facilitating insights into the landscape of drug-disease association prediction methods.

8.5 Identification of superior methods

The Superior Methods shown in Table 9 includes across categories distil key findings. This table helps in identifying methods that have demonstrated superior performance, providing valuable insights for researchers and practitioners in the selection of suitable methods for drug-disease association predictions.

8.6 Grouped comparison of AUC scores by dataset used

Table 10 groups AUC scores for different methods based on the dataset used. This provides a clear comparison of performance on various datasets, helping to identify methods that excel in specific contexts.

Based on the comparison table provided, there are several observations and relations that can be drawn regarding the computational methods and their AUC scores:

Table 10 Grouped Comparison of AUC Scores by Dataset Used

Dataset	Method	AUC Score
LRSSL Dataset	TL-HGBI (net-2014)	0.901
	MBiRW (net-2016)	0.816
	LRSSL (ML-2017)	0.917
	NTSIM (net-2017)	0.902
	SNF-NN (DL-2021)	**0.936**
TL-HGBI Dataset	TL-HGBI (net-2014)	0.915
	NTSIM (net-2017)	**0.961**
	LAGCN (DL-2021)	0.916
	LBMFF (DL-2023)	0.916
DN-Dataset	DrugNet (net-2015)	0.955
	MBiRW (net-2016)	0.958
	HNRD (DL-2019)	**0.97**
C-Dataset	DrugNet (net-2015)	0.804
	MBiRW (net-2016)	0.934
	TP-NRWRH (net-2016)	0.954
	DRIBRW (net-2018)	**0.964**
	HNRD (DL-2019)	0.950
	BGMSDA (net-2021)	0.954
B-Dataset	TL-HGBI (net-2014)	0.740
	deepDR (DL-2019)	0.857
	LAGCN (DL-2021)	**0.890**
	HINGRL (DL-2022)	0.884
	LBMFF (DL-2023)	0.882

– Dataset Variability: The performance of methods varies across different datasets, which suggests that dataset characteristics (such as the size, feature types, and underlying distribution) have a significant impact on the effectiveness of each method.

– Temporal Trend: Newer methods, especially those utilizing deep learning (DL), generally tend to have higher AUC scores. This might indicate that advancements in algorithmic approaches and computational power have led to more sophisticated models that can capture complex patterns more effectively.

8.7 Evolution of performance on F-dataset

The F-dataset stands out as a pivotal choice for evaluating the evolution of performance across various methods used to predict drug-disease relations. This dataset has been consistently employed by a diverse range of approaches, including network analysis (net), machine learning (ML), and deep learning (DL) techniques. The use of the F-dataset is grounded in its widespread use across multiple studies, rendering it a suitable benchmark for tracking the progress of methodologies over time.

By analyzing the performance metrics, particularly the AUC scores, of different methods on the F-dataset, we gain valuable insights into the evolving landscape of predicting drug-disease relations as shown in Fig. 1. This examination allows us to discern trends, identify advancements, and evaluate the effectiveness of novel techniques in comparison to established methodologies.

We can notice that performance varies widely, with some methods showing significant peaks in AUC scores, indicative of methodological breakthroughs. Also, deep learning methods display a greater range of performance, possibly reflecting the nuanced adaptation of these models to specific datasets.

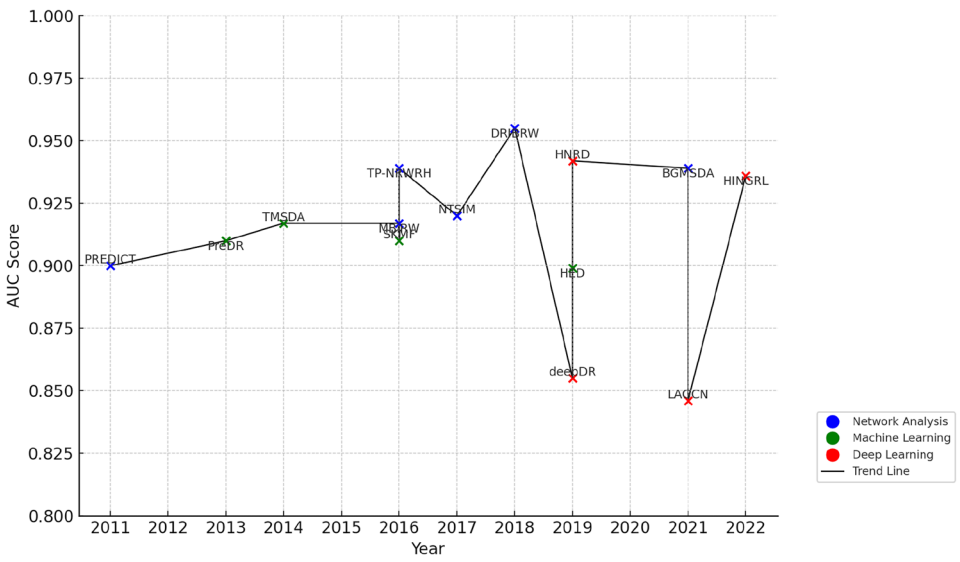

Fig. 1 Evolution of the Performance on F-dataset

9 Limitations

Machine learning depends heavily on the input data. In this setting, the data is limited and cannot fulfill the assumption of accurately representing the characteristics of drugs and diseases. Another limitation is the imbalanced nature of the data, which can cause false predictions. On the other hand, deep learning is characterized by its advanced learning capabilities, which can be very helpful in transforming the source data into feature representation. However, it needs a huge amount of data to leverage these learning abilities, but the data in this setting is sparse, and this leads to the problem of overfitting.

10 Conclusion

The survey of computational methods in drug repositioning has revealed a dynamic interplay between methodological innovation and the increasing complexity of biological data. The performance analysis across different datasets confirms that while no single method uniformly outperforms the rest, the trends towards integration of diverse data types and the adoption of advanced learning models are clear. Network Analysis-based methods offer a robust framework for understanding the intricate relationships in biological systems, machine learning methods excel at pattern recognition and scalability, and deep learning approaches have shown promising results, particularly in handling high-dimensional data and feature extraction. Based on our comprehensive review, we propose the following recommendations for future research and practical applications:

1) Data Integration and Standardization:
   - Develop standardized datasets and benchmarks to facilitate fair comparisons between methods.
   - Focus on integrating diverse data types (e.g., genomic, proteomic, and clinical data) to enhance prediction accuracy.
2) Interpretability and Explainability:
   - Prioritize the development of interpretable models, especially in deep learning approaches, to provide insights into the biological mechanisms underlying predicted associations.
   - Explore techniques such as attention mechanisms and feature importance analysis to enhance model interpretability.
3) Transfer Learning and Domain Adaptation:
   - Investigate transfer learning techniques to leverage knowledge from data-rich domains to improve predictions in domains with limited data.
   - Develop methods that can adapt to new drugs or diseases without requiring complete retraining.
4) Personalized Medicine:
   - Extend current methods to incorporate patient-specific data for personalized drug repositioning.
   - Investigate approaches that can account for genetic variations and individual patient characteristics in predictions.

Author contributions Conceptualization, EA and RA; analysis, EA and RA; methodology, EA and RA.; project administration, RA.; supervision RA; validation, EA and RA.; writing-original draft, EA.; writing-review and editing, EA and RA. All authors have read and agreed to the published version of the manuscript.

Funding Not applicable

Availability of data and materials Not applicable

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Ethical approval and consent to participate Not applicable.

Consent for publication Not applicable

References

Angermueller C, Pärnamaa T, Parts L, Stegle O (2016) Deep learning for computational biology. Mol Syst Biol 12(7):878

Ao C, Xiao Z, Guan L, Yu L (2023) Computational approaches for predicting drug-disease associations: a comprehensive review

Ata SK, Wu M, Fang Y, Ou-Yang L, Kwoh CK, Li X-L (2021) Recent advances in network-based methods for disease gene prediction. Briefings Bioinform 22(4):bbaa303

Boolell M, Allen MJ, Ballard SA, Gepi-Attee S, Muirhead GJ, Naylor AM, Osterloh IH, Gingell C (1996) Sildenafil: an orally active type 5 cyclic gmp-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. Int J Impot Res 8(2):47–52

Chatr-Aryamontri A, Breitkreutz B-J, Heinicke S, Boucher L, Winter A, Stark C, Nixon J, Ramage L, Kolas N, O'Donnell L, Reguly T, Breitkreutz A, Sellam A, Chen D, Chang C, Rust J, Livstone M, Oughtred R, Dolinski K, Tyers M (2012) The biogrid interaction database: 2013 update. Nucleic Acids Res 41:11

Consortium T, including, Acencio M, Kuiper M (2020) The gene ontology resource: enriching a gold mine. Nucleic Acids Res vol. 49

Gottlieb A, Stein G, Ruppin E, Sharan R (2011) Predict: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 7:496

Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales E, Gewiess A, Jensen L, Schneider R, Skoblo R, Russell R, Bourne P, Bork P, Preissner R (2008) Supertarget and matador: resources for exploring drug-target relationships. Nucleic Acids Res 36:D919-22

Hamosh A (2002) Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 30:52–55

Hong C, Yu J, Wan J, Tao D, Wang M (2015) Multimodal deep autoencoder for human pose recovery. IEEE Trans Image Process 24(12):5659–5670

Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4(11):682–690

Huang F, Qiu Y, Li Q, Liu S, Ni F (2020) Predicting drug-disease associations via multi-task learning based on collective matrix factorization. Front Bioeng Biotechnol 8:218

Hughes JP, Rees S, Kalindjian SB, Philpott KL (2011) Principles of early drug discovery. Br J Pharmacol 162(6):1239–1249

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek B, Martin M, Mcgarvey P, Gasteiger E (2009) Infrastructure for the life sciences: design and implementation of the uniprot website. BMC Bioinform 10:05

Jarada T, Rokne J, Alhajj R (01 2021) Snf-nn: Computational method to predict drug-disease interactions using similarity network fusion and neural networks. BMC Bioinform 22

Kanehisa M, Goto S (2000) Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30

Kang H, Hou L, Gu Y, Lu X, Li J, Li Q (2023) Drug-disease association prediction with literature based multi-feature fusion. Front Pharmacol 14:1205144

Kim Y, Jung Y-S, Park J-H, Kim S-J, Cho Y-R, (2022) Drug-disease association prediction using heterogeneous networks for computational drug repositioning. Biomolecules 12:1497

Kingma DP, Welling M et al (2019) An introduction to variational autoencoders. Found Trends® Mach Learn 12(4): 307–392

Kipf TN, Welling M (2016) Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907

Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart D (2010) Drugbank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res 39:D1035-41

Köhler S, Gargano M, Matentzoglu N, Carmody L, Lewis-Smith D, Vasilevsky N, Danis D, Balagura G, Baynam G, Brower A, Callahan T, Chute C, Est J, Galer P, Ganesan S, Griese M, Haimel M, Pazmandi J, Hanauer M, Robinson P (2020) The human phenotype ontology in 2021. Nucleic Acids Res

Kohn K, Aladjem M, Weinstein J, Pommier Y (2006) Molecular interaction maps of bioregulatory networks: a general rubric for systems biology. Mol Biol Cell 17:1–13

Kuhn M, Campillos M, Letunic I, Jensen L, Bork P (2010) A side effect resource to capture phenotypic effects of drugs. Mol Syst Biol 6:343

Liang X, Zhang P, Yan L, Fu Y, Peng F, Qu L, Shao M, Chen Y, Chen Z (2017) "Lrssl: Predict and interpret drug-disease associations based on data integration using sparse subspace learning. Bioinformatics (Oxford, England), vol. 33

Liu Y, Hu B, Fu C, Chen X (2010) Dcdb: drug combination database. Bioinformatics (Oxford, England) 26:587–8

Liu H, Song Y, Guan J, Luo L, Zhuang Z (2016) Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. BMC Bioinform 17:269–277

Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR (2018) A review of network-based approaches to drug repositioning. Brief Bioinform 19(5):878–892

Luo H, Wang J, Li M, Luo J, Peng X, Wu F-X, Pan Y (2016) Drug repositioning based on comprehensive similarity measures and Bi-random walk algorithm. Bioinformatics 32(17):2664–2671

Martínez V, Cano C, Blanco A (2014) Prophnet: a generic prioritization method through propagation of information. BMC Bioinform 15:1–13

Martinez V, Navarro C, Cano C, Fajardo W, Blanco A (2015) Drugnet: network-based drug-disease prioritization by integrating heterogeneous data. Artif Intell Med 63(1):41–49

Moghadam H, Rahgozar M, Gharaghani S (2016) Scoring multiple features to predict drug disease associations using information fusion and aggregation. SAR QSAR Environ Res 27:1–20

Monneret C, Bohuon C (2009) Fabuleux hasards: histoire de la découverte de médicaments. EDP Sciences, L'Editeur

Morselli Gysi D, Do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, Ghiassian SD, Patten J, Davey RA, Loscalzo J et al (2021) Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proc Natl Acad Sci 118(19):e2025581118

Oh M, Ahn J, Yoon Y (2014) A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. PLoS ONE 9(10):e111668

Oh M, Ahn J, Yoon Y (2014) A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. PLoS ONE 9:e111668

Olson T, Singh R (2017) Predicting anatomic therapeutic chemical classification codes using tiered learning. BMC Bioinform 18:06

Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner T, Rezwan F, Sharma A, Williams E, Brazma A (2009) Arrayexpress update-from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37:01

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL (2010) How to improve r &d productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 9(3):203–214

Sanseau P, Agarwal P, Barnes MR, Pastinen T, Richards JB, Cardon LR, Mooser V (2012) Use of genome-wide association studies for drug repositioning. Nat Biotechnol 30(4):317–320

Schriml L, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, Felix V, Jeng L, Bearer C, Lichenstein R, Bisordi K, Campion N, Hyman B, Kurland D, Oates C, Kibbey S, Sreekumar P, Le C, Giglio M, Greene C (2018) Human disease ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 47:11

Sertkaya A, Wong H-H, Jessup A, Beleche T (2016) Key cost drivers of pharmaceutical clinical trials in the united states. Clin Trials 13(2):117–126

Shim JS, Liu JO (2014) Recent advances in drug repositioning for the discovery of new anticancer drugs. Int J Biol Sci 10(7):654

Trivedi J, Mohan M, Byrareddy SN (2020) Drug repurposing approaches to combating viral infections. J Clin Med 9(11):3777

Vellal AD, Sirinukunwattan K, Kensler KH, Baker GM, Stancu AL, Pyle ME, Collins LC, Schnitt SJ, Connolly JL, Veta M et al (2021) Deep learning image analysis of benign breast disease to identify subsequent risk of breast cancer. JNCI Cancer Spectr 5(1):pkka119

Wang K, Sun J, Zhou S, Wan C, Qin S, Li C, He L, Yang L (2013) Prediction of drug-target interactions for drug repositioning only based on genomic expression similarity. PLoS Comput Biol 9(11):e1003315

Wang W, Yang S, Li J (2013) Drug target predictions based on heterogeneous graph inference. Pac Symp Biocomput Pac Symp Biocomput 18:53–64

Wang Y, Chen S, Deng N, Wang Y (2013) Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS ONE 8:e78518

Wang W, Yang S, Zhang X, Li J (2014) Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30(20):2923–2930

Wang Y, Guo M, Ren Y, Jia L, Yu G-X (2019) Drug repositioning based on individual bi-random walks on a heterogeneous network. BMC Bioinform 20:12

Weininger D (1988) Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36

WHO collaborating centre for drug statistics methodology, ATC classification index with DDD, (2023)

Wishart D, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) Drugbank: a knowledgebase for drugs, drug actions and drug targets. Nucl Acids Res 36:D901-6

Wouters OJ, McKee M, Luyten J (2020) Estimated research and development investment needed to bring a new medicine to market, 2009–2018. JAMA 323(9):844–853

Xie G, Li J, Gu G, Sun Y, Lin Z, Zhu Y, Wang W (2021) Bgmsdda: a bipartite graph diffusion algorithm with multiple similarity integration for drug-disease association prediction. Mol Omics 17:10

Xuan P, Cao Y, Zhang T, Wang X, Pan S, Shen T (2019) Drug repositioning through integration of prior knowledge and projections of drugs and diseases. Bioinformatics 35(20):4108–4119

Yang L, Agarwal P (2011) Systematic drug repositioning based on clinical side-effects. PLoS ONE 6(12):e28025

Yang K, Zhao X, Waxman D, Zhao X-M (2019) Predicting drug-disease associations with heterogeneous network embedding. Chaos (Woodbury, N.Y.) 29:123109

Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) Drug-target network. Nat Biotechnol 25(10):1119–1126

Yingdong W, Deng G, Zeng N, Song X, Zhuang Y (2019) Drug-disease association prediction based on neighborhood information aggregation in neural networks. IEEE Access 7:50 581-50 587

Yingdong W, Deng G, Zeng N, Song X, Zhuang Y (2019) Drug-disease association prediction based on neighborhood information aggregation in neural networks. IEEE Access 7:50 581-50 587

Yue X, Gutierrez BJ, Sun H (2020) Clinical reading comprehension: a thorough analysis of the emrqa dataset. arXiv preprint arXiv:1609.02907

Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F (05 2019) "Deepdr: A network-based deep learning approach to in silico drug repositioning. Bioinformatics (Oxford, England), 35

Zhang P, Wang F, Hu J (2014) Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. 2014: 1258

Zhang W, Yue X, Huang F, Liu R, Chen Y, Ruan C (2018) Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. Methods 145:51–59

Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, Liu F (2018) Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinform 19:1–12

Zhao B-W, Hu L, You Z-H, Wang L, Su X (12 2021) Hingrl: predicting drug-disease associations with graph representation learning on heterogeneous information networks. Briefings Bioinform 23