Review article / https://doi.org/10.15446/rcciquifa.v53n2.114447

The random forest machine learning model performs better in predicting drug repositioning using networks: Systematic review and meta-analysis

Darlyn Juranny García Marín[a], Jerson Alexander García Zea[b]

Universidad EAFIT, Carrera 49, N° 7 Sur -50, Medellín, Antioquia, Colombia

E-mail adresses: [a]djgarciam@eafit.edu.co, [b]jagarciaz2@eafit.edu.co

Received: February 5, 2024
Corrected: March 10, 2024
Accepted: March 12, 2024

Summary

Introduction: The lengthy and costly process of drug development can be expedited through drug repositioning (DR), a strategy that identifies new therapeutic targets using existing products. Supervised machine learning (SML) models, incorporating interaction networks, offer a promising approach for DR. This study aims to systematically review and meta-analyze SML models predicting DR, identifying key characteristics influencing their performance. **Methodology:** A systematic review was conducted to identify SML models that used networks to predict DR, which were evaluated by comparing their performance through a random-effects meta-analysis. **Results:** 19 studies were included in the qualitative synthesis and 17 in the quantitative evaluation, The Random Forest (RF) model emerged as the predominant classifier (63%), yielding the highest performance in AUC ROC comparisons (overall value: 0.91, 95% CI: 0.86 – 0.96). Validation efforts in 18 studies confirmed the predictions of the SML models, affirming the proposed drugs. The incorporation of chemical structure in model training was found to enhance performance by aiding in prediction discrimination. **Conclusion:** SML models can predict DR, the RF model was the most widely used SML model with the best performance results, which underscores the potential use of FR models for predicting DR using network form biomedical information.

Keywords: Drug Repositioning, Drug development, Biological Networks, Machine Learning, Random Forest.

Resumen

El modelo de aprendizaje automático bosque aleatorio presenta un mejor desempeño para predecir el reposicionamiento de medicamentos usando redes: Revisión sistemática y Meta-análisis

Introducción: El proceso de investigación y desarrollo de fármacos se puede acelerar mediante el reposicionamiento de medicamentos (DR), una estrategia que identifica nuevos objetivos terapéuticos utilizando productos existentes. Los modelos de aprendizaje automático supervisado (SML), que incorporan redes de interacción, ofrecen un enfoque prometedor para DR. Este estudio tiene como objetivo revisar y meta-analizar sistemáticamente los modelos SML que predicen DR, identificando características clave que influyen en su desempeño. **Metodología:** Se realizó una revisión sistemática para identificar modelos SML que utilizaran redes para predecir DR, los cuales se evaluaron comparando su desempeño mediante un meta-análisis de efectos aleatorios. **Resultados:** Se incluyeron 19 estudios en la síntesis cualitativa y 17 en la evaluación cuantitativa. El modelo Bosque aleatorio surgió como el clasificador predominante (63%), obteniendo el mayor rendimiento en las comparaciones AUC ROC (valor general: 0,91, 95% IC: 0,86 – 0,96). Los esfuerzos de validación en 18 estudios confirmaron las predicciones de los modelos SML, afirmando los medicamentos propuestos. Se descubrió que la incorporación de estructura química en el entrenamiento de modelos mejora el rendimiento al ayudar en la discriminación de predicciones. **Conclusión:** Los modelos SML pueden predecir la DR, el modelo RF fue el modelo SML más utilizado con los mejores resultados de rendimiento, lo que resalta el uso potencial de modelos FR para predecir el DR utilizando redes de información biomédica.

Palabras clave: Reposicionamiento de medicamentos, Desarrollo de medicamentos, Redes biológicas, Aprendizaje automático, Bosque aleatorio.

Resumo

O modelo de aprendizado de máquina Floresta Aleatória apresenta melhor desempenho para prever o reposicionamento de medicamentos utilizando redes: Revisão Sistemática e Meta-análise

Introdução: O processo longo e custoso de desenvolvimento de medicamentos pode ser acelerado por meio do reposicionamento de medicamentos (DR), uma estratégia

que identifica novos alvos terapêuticos usando produtos existentes. Modelos de aprendizado de máquina supervisionado (SML), incorporando redes de interação, oferecem uma abordagem promissora para o DR. Este estudo tem como objetivo revisar sistematicamente e realizar meta-análises de modelos SML que preveem DR, identificando características-chave que influenciam seu desempenho. **Metodologia:** Foi realizada uma revisão sistemática para identificar modelos SML que usaram redes para prever DR, os quais foram avaliados comparando seu desempenho por meio de uma meta-análise de efeitos aleatórios. **Resultados:** 19 estudos foram incluídos na síntese qualitativa e 17 na avaliação quantitativa, o modelo Floresta Aleatória (RF) emergiu como o classificador predominante (63%), apresentando o melhor desempenho em comparações de AUC ROC (valor geral: 0,91, IC 95%: 0,86 - 0,96). Os esforços de validação em 18 estudos confirmaram as previsões dos modelos SML, afirmando os medicamentos propostos. A incorporação da estrutura química no treinamento do modelo mostrou-se capaz de melhorar o desempenho ao auxiliar na discriminação das previsões. **Conclusão:** Os modelos SML podem prever DR, o modelo RF foi o modelo SML mais amplamente utilizado com os melhores resultados de desempenho, o que destaca o potencial uso dos modelos FR para prever DR usando informações biomédicas de rede.

Palavras-chave: Reposicionamento de medicamentos, Desenvolvimento de medicamentos, Redes biológicas, Aprendizado de máquina, Floresta Aleatória.

Introduction

The research and development of new drugs constitute a lengthy and costly process. Given the varying complexities of therapeutic fields, it proves challenging to universally quantify the approximate cost in terms of both time and money for all drugs. Nevertheless, available data indicates that the investment required to bring a drug to market ranges from $161 million to $4.5 billion dollars [1] and can take 10 to 15 years [2]. Despite the substantial efforts invested in this process, approximately 90% of drugs experience failure during their clinical phase. This outcome leads to prolonged waiting periods for many patients, anticipating a successful approval process for a treatment molecule [3]. Given the extended timeline and a relatively low success rate, it becomes crucial to adopt strategies aimed at enhancing the efficiency and success of the drug development and approval process.

Drug repositioning (DR) is an approach that enables the identification of new therapeutic targets from products that are already known or currently in the market. This

approach facilitates an expedited drug approval process [4]. This characteristic makes DR an advantageous strategy compared to the traditional pharmaceutical product development process. Typically, for a repositioned drug, a significant portion of non-clinical research has already been explored. This includes aspects such as chemical analysis, manufacturing and control, animal toxicology, and clinical pharmacology. Consequently, it can progress more directly to the clinical phase, where there is already existing basic clinical information that could prove beneficial for the new treatment. This streamlined process has the potential to significantly reduce the overall development costs by 50 – 60%. Moreover, given that this medicine already possesses a known safety profile, the risk of failure in later stages of the process is significantly reduced [5].

DR can be conducted through experimental work where the drug is evaluated in the laboratory. However, it may not be the most optimal path to take initially in the research for drug repositioning. This is because it demands time, facilities, equipment, and personnel dedicated to experimentation, along with the physical product, to conduct the necessary assessments for each specific analysis [6]. This, added to the high number of molecules available on the market, makes it strategically better to have computational tools that initiate the DR process. In this regard, high-performance computing and artificial intelligence can help accelerate the identification of potential active substances for repositioning and reduce high failure rates [7].

Thanks to the different open data initiatives, the extensive pharmaceutical knowledge in the literature, large databases of diseases, drugs and adverse effects, it has been possible to develop computational tools for the repositioning of drugs [7]. Computational prediction between drugs and diseases has emerged as a crucial process in drug repositioning research. Advances in systems biology and interaction networks have facilitated the evolution of network pharmacology, transforming the paradigm of drug interaction. Initially perceived as a linear path "one drug – one therapeutic target – one disease", it has evolved into a network model: "Multiple Drug" network, involving multiple therapeutic targets and multiple diseases. This shift represents a fundamental change in our understanding of drug interactions [8].

In general, in the study of drug repositioning, it is assumed that chemical structures, target proteins, and even adverse effects enrich the information to establish new indications [9]. The chemical structure of a drug provides information based on its structural function, target proteins provide information on the direct effect of the drug at the molecular level, and adverse effects establish key points in relation to the undesired effect at the level of the drug phenotype [10]. Understanding the rules, patterns, and similarities within existing data facilitates comprehending the interaction between a drug and its efficacy in treating a disease. In the same vein, exploring these data sets

can lead to the discovery of new relationships. Studies have demonstrated that drugs sharing similar chemical structures, target proteins, or exhibiting comparable adverse effects have an increased probability of effectively treating the same disease [10, 11]. In this sense, the use of biological networks provides the opportunity to know those topological characteristics between nodes and edges that predict unknown associations between drugs and diseases [9].

For making predictions, Supervised Machine Learning (SML) models have been implemented. These models are trained using the characteristics of the network extracted from biomedical databases to measure similarities in conjunction with known drug-disease interactions. Through this approach, the model becomes capable of generating new potential candidate drugs for repositioning [12]. On the other hand, the use of SML is an efficient strategy that provides predictions on a larger scale, which broadens the scope of repositioning evaluation, allowing screening to be made to a greater number of drugs [13]. In addition, Machine Learning models are not limited to the total knowledge of the three-dimensional structure of chemical ligands and protein targets, which is the main disadvantage of the well-known molecular docking modeling [14].

The utilization of SML for suggesting drug candidates for repositioning based on network information has seen substantial growth. This presents an opportune scenario to synthesize information, discern strategic data pertaining to the models and algorithms employed, the origin of the data utilized, and identify potential performance moderators influencing the prediction of new therapeutic targets. In addition, to be able to establish the reliability of the different models based on their validation strategies in the candidate drugs for repositioning.

Thus, in this work, a systematic review and meta-analysis [15] was developed in order to synthesize the available information on the use of SML in proposing candidate drugs for repositioning, using network information and provide a tool for academia or industry with key synthesized information for the development of computational models for DR.

Methodology

The systematic review was conducted following the guidelines in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement 2020 [16].

Literature search strategy

A literature search was executed in English in the Scopus and PubMed databases, where studies up to April 2023 were included. The search syntax related the terms net-

works (Network*) and drug repurpose* joined with the AND operator, and the Boolean operator OR was used to include the synonyms. The complete search strategy for both databases can be reviewed in Supplementary Table 1. Employing search engine filters, articles in languages other than English, conference abstracts, and reviews were excluded.

Study selection and eligibility criteria

The initial screening of studies was conducted using the Rayyan platform [17]. Leveraging the tools provided by Rayyan, duplicates between the two databases were identified and removed. Two review authors independently conducted the title and abstract screening. To be included, a study had to explicitly indicate the use of at least one network model to predict drug repositioning (or any of its synonyms). Any discrepancies in this initial review were resolved through discussions among the review authors.

In a subsequent step, one of the authors categorized the studies that emerged from the initial screening based on the identified computational strategy, a classification that was verified by the second reviewer. The primary focus of interest was on SML models, encompassing any technique within this category employed for predicting drug repositioning through interaction networks. For encoding SML, the Python programming language's SciKit Learn library was utilized [18]. The models incorporated into this package served as a reference for classifying the studies.

The full-text review and data collection process was carried out in all studies using at least one MLS model. One author conducted this review, and the findings were corroborated by the second author to address any discrepancies. Data extraction involved populating a matrix in Microsoft Excel 2010, including author data, country, year of publication, objectives, model type, data source for model development, network type, performance metrics, details on whether the model was compared against other prediction models, and information on validation. At this stage, the reasons for exclusion were documented in each case.

The studies that were eligible in the qualitative analysis met the following criteria:

- The primary objective was to use networks to predict drug repositioning, employing at least one SML strategy.
- The study specified the name of the algorithm utilized.
- The developed model targeted more than one therapeutic specialty (screening) for medication repositioning.
- The developed model focuses on more than one drug.

- The data used for training the model was not veterinary.
- The study did not center on herbal medicines.
- The study did not center on traditional Chinese medicine.

The full article was accessible.

Studies were eligible in the quantitative analysis if they met the following criteria:

- Fulfill all the criteria mentioned above.
- Evaluate the performance of the model with available results.

Studies were excluded if they met the following criteria:

- The article did not meet the inclusion criteria.
- The article was in a language other than English.
- The article was a review.
- The article was a conference summary.
- The article was a book chapter.

Quality assessment of study

The quality of the publications was measured after the final selection process. The following list of questions was used to assess the credibility of the selected publications.

Question 1: Specify type of model, model-building procedures, and method for internal validation.

Question 2: Specify measures used to assess model performance and, to compare multiple models.

Question 3: Discuss the results with reference to performance and any other validation data.

Question 4: Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.

Each of the above questions was assessed with responses categorized as "Yes," "Partially," or "No." These responses were assigned numerical values: "Yes"=1, "Partially"=0.5 and "No"=0. For each selected study, its quality score was determined by summing the

scores of the responses to the four questions. This evaluation process was conducted by one of the authors of this article. The quality levels were classified as High (score = 4), Medium (2 ≤ score <4) and Low (score <2). Studies with scores greater than or equal to 3 were included in the final selection [19]. In cases where studies assessed performance using only one parameter, question two was marked "partially".

Performance measurement and meta-Analysis

The AUC-ROC (Area Under the Curve - Receiver Operator Characteristics) was chosen as the primary metric to assess the models' performance in the quantitative analysis, given its prevalent inclusion in nearly all articles (19 out of 17). While additional metrics were gathered, they were not incorporated into the quantitative analysis. In cases where a study reported results from multiple predictive models, performance parameters were collected exclusively for the model with the most favorable outcome.

The random effects meta-analysis model employing the DerSimonian-Laird model [20] was selected for the corresponding statistical analyses. This approach involved estimating the variability between studies (τ^2) and applying Cochrane's Q and I^2 tests to assess heterogeneity across studies. In the calculations, "n" was defined as the number of interactions during network construction. Statistical analysis was conducted in R Studio version 4.2.3 [21] using the metafor package.

Results

Article Selection for Synthesis

Following the application of search terms in the databases and the use of filter tools, reviews, conference papers, and articles not in English were excluded. The initial total of articles was 941 (PubMed) and 861 (Scopus). After removing duplicates across both databases, a total of 1057 articles were obtained and included in the initial screening process. Within this set, articles utilized at least one network model for predicting DR. These articles were categorized according to the computational strategy identified (see Table 1). Among these publications, we identified 44 articles that used SML models for the prediction of DR and were included in the full article review process. A comprehensive review of the 44 publications disclosed that 19 fulfilled all the inclusion criteria for synthesis. All 19 studies were incorporated in the qualitative synthesis and only 17 were included in the meta-analysis, with 2 studies [22, 23] being excluded. At this point, the reasons for exclusion were recorded and can be reviewed in Figure 1.

Table 1 Classification of articles by computational strategy.

Model Classification	Number of articles
Deep Learning (DL)	67
Supervised Machine Learning	44
Graph theory	21
Network based	20
Semi supervised machine learning	7
Undetermined	37

Model Categorization

The 196 studies were categorized into the six groups outlined in Table 1 and the frequency of model publication per year was determined (Figure 2). The initial publication appeared in 2009 with the most recent in 2023. DL emerged as the predominant modeling strategy, representing 34% of the total publications, and the 84% of its articles were published between 2020 and 2023, with the peak occurring in 2022, marking it as the year with the highest number of publications. This trend positions DL as the strategy that has experienced the most significant increase in the last three years. In contrast, SML constituting 22% of the articles, had publications spanning from 2013 to 2022, with representation in all years except 2015. Similarly, Network Based and Graph Theory models made their appearances in 2012 and 2013, respectively. While Figure 2 visually illustrates the distribution of studies up to 2023, it's crucial to note that the data for this year only includes articles published up to April.

Articles included in the synthesis/Study characteristics

Table 2 outlines the characteristics of the 19 studies. These were published between 2013 and 2022, 11 were conducted in China and the remainder in Thailand, Iran, Korea, the United Kingdom, Canada, and the United States.

To delineate the relationship between drugs and therapeutic targets, the 19 studies incorporated drug-target interaction databases. The most frequently utilized database was DrugBank (14 out of 19 articles), followed by the Kyoto Encyclopedia of Genes and Genomes (KEGG). In studies aimed at establishing the known relationship between drugs and diseases (drug-disease), the Comparative Toxicogenomics Database (CTD) was commonly employed. To determine the protein-disease relationship (Protein-disease), DisGeNET was implemented. When microRNAs and long noncoding RNAs were included, data from LncRNA2Target, LncRNADisease, lncRNASNP2, miRTarBase, and lncRNASNP2 were utilized. The protein-protein relationship was defined using the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING). In studies that incorporated adverse effects to establish relation-

ships, SIDER (Side effect Resource) was employed, enabling the tracing of adverse effects associated with medications.

To characterize and establish similarities between chemical structures, molecule/drug repositories such as FDA-approved drugs, DrugBank or PubChem were implemented. The similarity of the chemical structure of the drugs was made based on the chemical language SMIleS or in other cases the molecular descriptor MACCS (Molecular ACCess System).

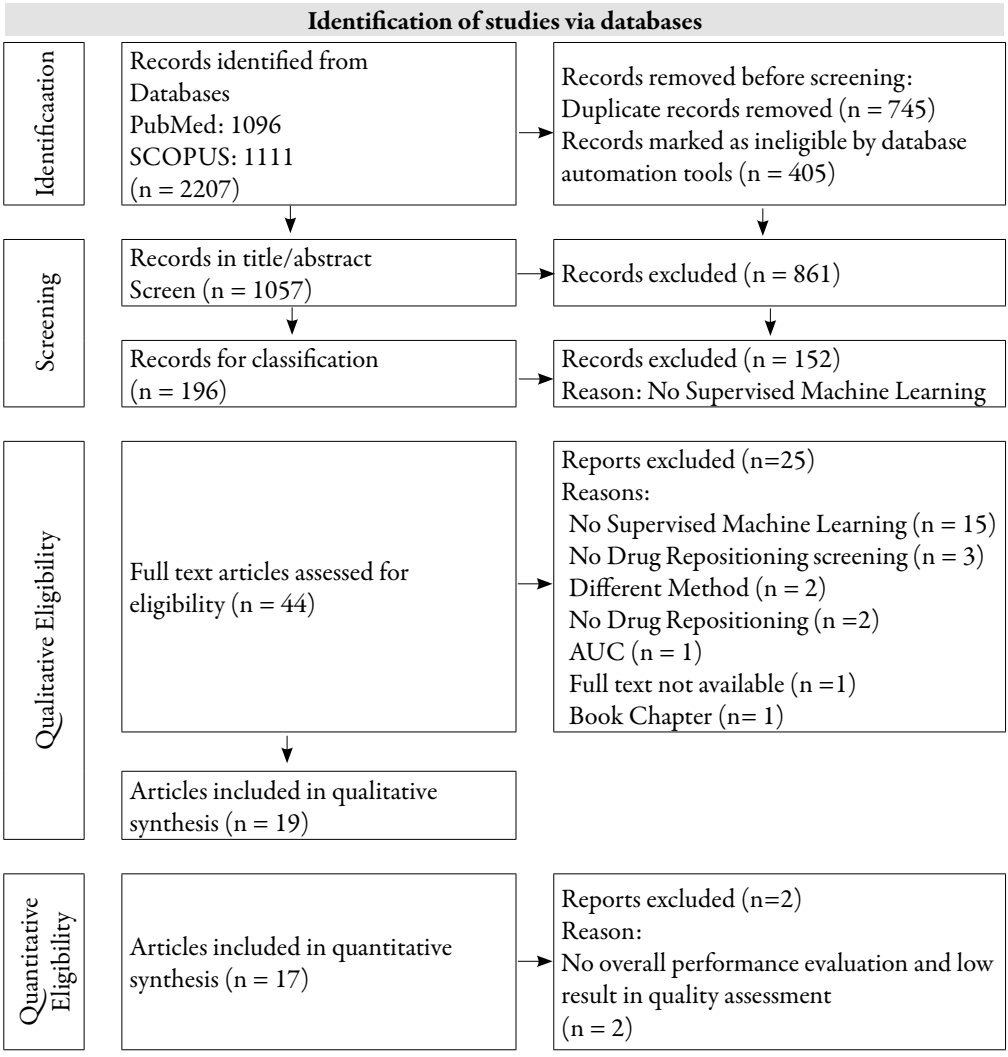

Figure 1. PRISMA Flow diagram. Papers identified in databases, title/abstract screened, read full text, and included in the synthesis. Reasons for exclusion are listed.

To establish relationships between genes and diseases, the studies utilized information available in Online Mendelian Inheritance in Man (OMIM) in some cases or DisGeNET (a database of gene-disease associations) in others, using them as compendiums of human genes and genetic disorders. For detailed information on the articles and their databases/libraries, see supplementary table 2.

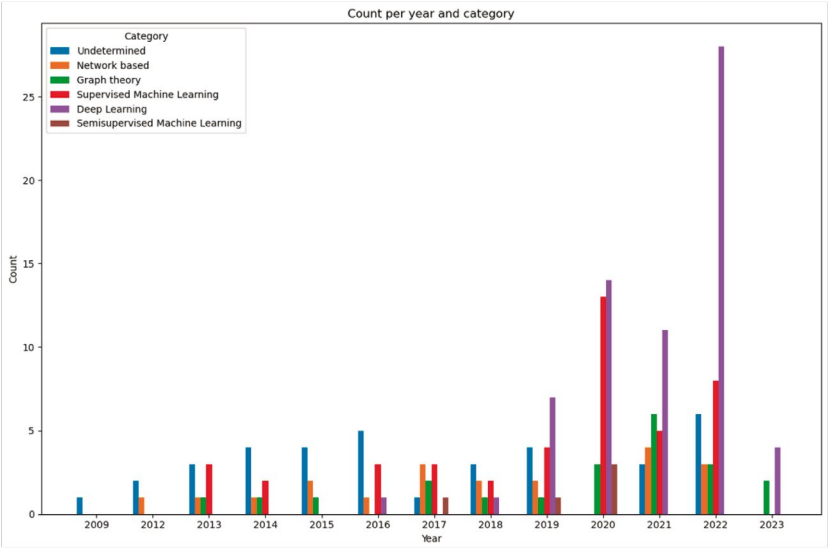

Figure 2. Categorization of studies resulting from screening that used at least one network model to predict drug repositioning.

In each study, considering the database utilized, various types of networks were constructed. The studies generated networks between drugs and targets with up to 9,400,000 interactions, 11 defined the interaction between drug and disease, 3 focused on the relationship between diseases and proteins. Additionally, 12 studies incorporated chemical structure in the development of the model. Moreover, in other instances, networks of interaction between genes and diseases or genes and drugs, as well as networks between adverse effects and drugs, were constructed. The strategy employed varied across studies; a single network was created to train the models, while in others, different networks were developed within the same study for training the model and establishing new indications for drugs (refer to Table 2 for more details).

Concerning the SML models, the most frequently employed classifier was Random Forest (RF) (63%), implemented in 12 studies. Other models included Support Vector Machine (SVM) (16%), logistic regression (11%), Gradient Boosting (5%) and XGBoost (5%). Each of the studies had a unique networking process to train the models and predict new therapeutic targets in medicines.

Table 2. Characteristics of the included studies

ID	Reference	Study Location	Year	Algorithm	Network	Models for Comparison (Name if available)	Model validation
2	Wang *et al.* (2013) [10]	China	2013	SVM	1,933 Interactions among 593 drugs and 313 diseases	PREDICT	ClinicalTrials.gov
4	Cao *et al.* (2014) [24]	China	2014	RF	5127 Drug target interactions	BGL, BLM, NetLapRLS, RLS	Literature searching
5	Qabaja *et al.* (2014) [22]	Canada	2014	LR	2343 genes and 22 diseases Network, 2343 genes and 406 drugs Network	-	Literature searching
8	Moghadam *et al.* (2016) [25]	iran	2016	SVM	1933 Drugs diseases interactions	PReDICT, a- PreDR	ClinicalTrials.gov
12	Lee and Yoon (2018) [23]	Korea	2018	RF	4248 Drugs and targets interactions, 7083 drugs and diseases interactions	PREDICT, PreDR	ClinicalTrials.gov and PubMed
16	Wei *et al.* (2019) [9]	China	2019	SVM	5868 Side effect for 1430 drugs, 19.906 drug target interactions	-	ClinicalTrials.gov and PubMed
18	Chen *et al.* (2020) [26]	China	2020	RF	105546 Drug target interactions	PREDICT, TL-HGBI, MBiRW, LRSSL, SCMFDD, a- EMP-SVD	-
20	Liu *et al.* (2020) [12]	China	2020	RF	4642 Drug protein interactions, 1365 disease protein interactions, 1827 drug disease interactions	HGBI, LDB, TL-HGB, a- Drug Net	Literature searching

(Continued)

ID	Reference	Study Location	Year	Algorithm	Network	Models for Comparison (Name if available)	Model validation
21	Fahimian *et al.* (2020) [27]	Iran	2020	RF	9400000 Drug disease interactions	MBiRW, DTINet, HGBI, NBI, a- JUST	A Cell toxicity assay to assess the effectiveness of the repurposed drug on breast cancer stage II (HER2 cell line)
23	Zhou *et al.* (2020) [28]	China	2020	RF	1933 Drug disease interactions	PREDICT	Literature searching
25	Gilvary *et al.* (2020) [29]	United States	2020	Gradient Boosting	22,399 Protein coding genes, 6,679 drugs	DDR, NEDD, NRLMF, DTINet, CMF, BLM-NII, a-NetLapRLS	Literature searching
28	Yue and Because (2021) [30]	China	2021	RF	90 Nuclear receptors, 1476 Ion channels, 635 GPCR, 2926 Enzymes, 9881 drugs	Pathima Nusrath Hameed *et al.*, Lei Chen *et al.*, methods, SPACE	Literature searching
29	Yang *et al.* (2021) [31]	China	2021	LR	8969 Substructure drug target domain interactions, 8053 indication drug target domain interactions	Lee and Yoon, Wu *et al.* methods	Literature searching
31	Kitsiranuwat *et al.* (2021) [32]	Thailand	2021	RF	14264 Drug disease interactions	DNILMF, DT-Hybrid, DDR	ClinicalTrials.gov, PubMed and AACT data base
32	Amiri-Souri *et al.* (2022) [33]	United kingdom	2022	XGBoost	2057 Real negative drug target interactions, 1721 Positive interaction	LAGCN, DTINet, a-deepDR	Drug-protein docking simulation and literature searching

(Continued)

ID	Reference	Study Location	Year	Algorithm	Network	Models for Comparison (Name if available)	Model validation
33	Zhao *et al.* (2022) [34]	China	2022	RF	18416 Drug disease interactions, 11107 drug protein interactions, 25.087 protein disease interactions	DeepDR, DTINet, GIPAE, a- HINGRL	Literature searching
34	Zhang *et al.* (2022) [35]	China	2022	RF	18416 Drug disease interactions, 5898 Disease protein interactions, 3110 Drug protein interactions	-	Literature searching
35	Kitsiranuwat *et al.* (2022) [36]	Thailand	2022	RF	27,683 Drug protein interactions	SCMFDD, LNS	Literature searching
36	Jiang and Huang (2022) [37]	China	2022	RF	6528 Drug target interactions	PREDICT, TL-HGBI, MBiRW, LRSSL, SCMFDD, a- EMP-SVD	Literature searching

RF– Random Forest / SVM – Support Vector Machine / LR – Logistic Regression

For the prediction of new drug targets, accuracy was assessed using AUC ROC ranging between 0.83 and 0.99 for the RF model and 0.88 to 0.90 for the SVM model.

Of the 19 articles selected after the full review of the study, 17 met all the final requirements to be included in the quantitative evaluation, while articles [22] and [23] were excluded. Despite the fact that all 17 articles presented model performance measures, there was no homogeneity in the performance parameters evaluated. The only common parameter in all 17 articles was the AUC ROC; Accuracy, F Measure, Sensitivity, and Precision were reported in 10 articles, Matthews' correlation coefficient was reported in 5, and specificity and AUPR were reported in 3 (see Table 3). Additionally, out of the 17 articles, all performed cross-validation, and 15 of 17 studies compared their prediction models with other models reported in the literature. Among these, 13 made the comparison by calculating the AUC ROC.

Table 3. Model Performance Overview

ID	AUC ROC	Acc	F measure	MCC	Its	Spe	For	AUPR	Cross validation / folds
2	0.902	0.823	0.822	-	0.847	-	0.808	-	Yes/10
4	0.986	0.948	-	-	0.965	0.930	-	-	Yes/5
8	0.88	0.83	0.82	0.66	-	-	-	-	Yes/10
16	0.903	0.853	0.779	-	-	-	0.778	-	Yes/10
18	0.920	0.854	-	0.713	0.796	0.912	0.900	-	Yes/5
20	0.966	0.919	0.918	0.840	0.939	-	0.899	-	Yes/5
21	0.83	-	0.72	-	0.79	-	0.66	-	Yes/10
23	0.923	-	-	-	-	-	-	-	Yes/10
25	0.841	-	-	-	-	-	-	-	Yes/5
28	0.998	-	-	-	-	-	-	-	Yes/10
29	0.876	-	-	-	-	-	-	-	Yes/-
31	0.879	0.967	1.00	-	-	-	-	-	Yes/5
32	0.940	-	0.928	-	0.928	-	0.927	0.894	Yes/10
33	0.876	-	0.794	0.589	0.796	-	0.793	0.866	Yes/10
34	0.963	0.900	0.900	-	0.897	-	0.903	-	Yes/10
35	0.938	0.857	0.866	-	0.928	-	0.812	0.932	Yes/5
36	0.879	0.798	-	0.596	0.800	0.796	0.797	-	Yes/5

AUC ROC: Area Under the Curve - Receiver Operator Characteristics / Acc: Accuracy / MCC: Matthews Correlation Coefficient / Sen: Sensitivity / Spe: Specificity / Pre: Precision/ AUPR: Area Under Precision-Recall Curve.

After evaluating the performance of the developed models, tests were conducted to validate the veracity of the predictions, and since the models aimed to predict new indications for the drugs, confirming the effectiveness of the newly established relationships became crucial. Out of the 19 studies, 18 presented validation results for their models. The most commonly employed strategy was consulting the literature to identify whether the new indication for the drug(s) had been reported at any time. In this exercise, the most cited sources of information were ClinicalTrials.gov and PubMed, in other cases the source consulted was not specified. In all the studies that used this way of validating the candidate drugs for repositioning, they presented citations of clinical studies or other references where this new indication for repositioned drugs had already been mentioned in the scientific literature. Finally, 2 studies presented different strategies to literature searching [27, 33], in the case of Amiri-Souri *et al.* (2022) [33], an *in vitro* study was done and in the Fahimian *et al.* (2020) [27] work, a simulation with molecular docking was performed (reference to Table 2 for more details).

Assessing the quality of the evidence

10 studies were rated as high quality, 8 as medium quality and 1 as low quality. For those studies that only assessed performance focused on a single metric, they were assigned "Partially" in question 2. In cases where there were no validation results, "Partially" was assigned in question 4. Since studies with a score equal to or greater than 3 were included in the quantitative assessment, 2 studies were excluded at this point in the process. The full result of the quality assessment of the studies can be found in Supplementary Table 3.

Meta-analysis

A total of 17 studies were included in the quantitative assessment after removing two studies due to quality assessment outcome and lack of overall model performance. The studies were evaluated using a random-effects model and compared based on their AUC ROC (95% CI), the summary of the results can be seen in Figure 3. We assessed inter-study variability τ^2 = 0.0103 (SE 0.0104) and Cochrane's Q=223772.7841, p-val <.0001, I^2 = 99.99% for heterogeneity.

The AUC ROC is measured on the scale 0 – 1, the closer to 1, the better the classifier [38]. The overall analysis of the results obtained from the meta-analysis showed that the mean AUC ROC for the included studies was 0.91 (95% CI 0.86 – 0.96). The highest AUC ROC result was obtained by Yue and He (2021) [30] with a value of 1.00 who implemented an RF model. The model with the lowest accuracy was Fahimian *et al.* (2020) [27] with 0.83 AUC ROC, which also used RF for prediction. Regard-

ing the studies that used SVM, the highest accuracy value was obtained by Wang *et al*. (2013) [10] and Wei *et al.* (2019) [9] with 0.90 and the lowest for Moghadam *et al.* (2016) [25] with a result of 0.88 in the meta-analysis. The model that implemented the highest number of interactions to build the network was Fahimian *et al.* (2020) [27], which obtained a result of 0.83. In contrast, the models implemented by Moghadam *et al*. (2016) [25], Wang *et al*. (2013) [10] and Zhou *et al*. (2020) [28] generated the lowest number of network drug-target interactions (1933) and showed a value for AUC ROC of 0.88, 0.90 and 0.92 respectively.

The study developed by Yue and He (2021) [30] obtained the highest result (1.00 AUC ROC) among the models that included databases that established a relationship between Drug–Target and Drug–Disease. Among the studies that used the Protein-Disease relationship in the construction of interactions, the highest result was AUC ROC of 0.96 [35]. As for those that included the Protein-Protein ratio, Kitsiranuwat *et al.* (2022) [36] obtained 0.94 as the highest value. The chemical structure of the drugs was integrated into the development of 12 models. Among these, Cao *et al.* (2014) [24], achieved the highest accuracy (0.99). Additionally, the association between genes and diseases was used in 10 articles, Liu *et al.* (2020) [12] obtained 0.97 in this category. Wang *et al*. (2013) [10] and Wei *et al*. (2019) [9] with 0.90, integrated adverse effects in drugs into the development of the model. Furthermore, both Jiang and Huang (2022) [37] and Chen *et al.* (2020) [26] were the only studies that included RNA and obtained values of 0.88 and 0.92 for AUC ROC, respectively.

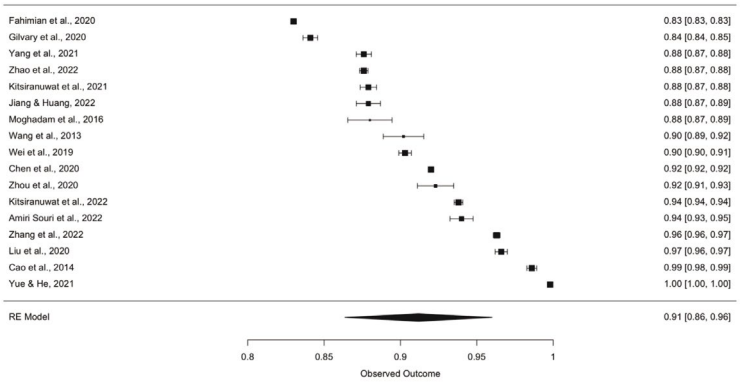

Figure 3. Forest plot comparing the papers by AUC ROC, in a random effects model. Overview of meta-analysis under a random effects model, comparing the AUC ROC of SML studies for drug repositioning. The data is organized in ascending order.

Discussion

Generally, in drug repositioning studies, it is assumed that chemical structures, target proteins, and even adverse effects enrich and provide valuable information for the identification of new indications [9]. Drugs with similar structures have been shown to have a high probability of treating the same disease [11] and in conjunction with the integration of biological networks offers the opportunity to understand the topological characteristics between nodes and edges, predicting previously unknown associations between drugs and diseases [9].

In this context, studies aimed at predicting drug repositioning use biomedical databases that relate information from the target to the drug. These databases are combined with various sources of information to build the network, measure similarities, and obtain the characteristics needed to train the prediction model. These predictions are based on the idea that drugs or diseases with similar topological network properties may be functionally related. Figure 2 illustrates the increase in studies predicting drug repositioning using interaction networks from 2009 to 2022. It highlights the increase in studies implementing SML from 2013 to 2020, with continuity in publications during 2021 and 2022. At this point, it is opportune to conduct a systematic review of the available information on these prediction models that implement SML methods. These models are designed to identify new candidate indications for drug repositioning, taking into account the various types of relationships between biomolecules and known drug-disease interactions. In addition, a quantitative evaluation of these models through a meta-analysis was proposed.

Considerable efforts have been devoted to demonstrating that drugs with similar chemical structure, similar adverse effects, and drugs that target the same target protein can treat the same specific disease. In a previous study, Wang *et al*. (2013) [10], showed that adopting this perspective favors the process of drug repositioning. It highlights that an SML model designed for repositioning can be trained using information from databases, such as chemical structures, target proteins, or adverse effects. In addition, various data, such as genes, expression profiles [22], microRNAs and long noncoding RNAs [26, 37], have proven to be valuable in the construction of these models. However, when evaluating performance, it has been observed that merging all these resources during the training process significantly improves the model compared to the individual use of information [10, 25]. Therefore, it can be considered that making a comprehensive characterization of drugs at multiple levels (chemical structure, target protein/gene or adverse effects) can substantially expand and improve the quality of predictions in drug repositioning.

Regarding the development of the models, the predominant approach in the reviewed studies consisted of the creation of interaction networks, the identification of their characteristics and the incorporation of this information into SML models, addressing the problem as a classification task. In this way, it can be determined whether a drug might be linked to a new disease, based on available training information. In several models developed, it has been shown that including the networks significantly improves the overall performance of the model. In these cases, network profiles provide valuable additional information compared to databases that do not incorporate interaction profiles [11, 24, 26]. This finding suggests that the consideration of network characteristics enriches the predictive capacity of the models, offering a more complete perspective in the identification between drugs and diseases.

In this study, papers were considered in the quantitative assessment if they reported at least AUC ROC. Some studies used a variety of performance measures (see Table 3), while others limited to report only the AUC ROC. The quantitative analysis was conducted using a meta-analysis, using a random-effects model that considers the variability both intra-study and between-studies. For this purpose, the measure τ^2 was calculated, which provides an estimate of the variance in effects between studies, while I^2 describes the percentage of variability attributable to heterogeneity, assessing the extent to which the studies agree with each other [39]. In our case, a high value for heterogeneity was obtained, which can be manifested through statistical uncertainty or random variability, which could be derived from methodological diversity due to different strategies in the construction of the network, variations in training data, differences in data sources, the definition of similarities in interactions and the use of different SML models.

In the resulting quantitative analysis, it was observed that the most frequently used classifier was RF, being not only the most common alternative, but also the model with the highest accuracy measured by the AUC ROC in the general summary of the meta-analysis (see Figure 3). In fact, the four highest values of this performance measure, ranging from 0.96 to 1.00, corresponded to models developed with RF. Additionally, studies comparing RF performance with other supervised learning models found that this strategy was shown to have more consistent performance and prediction [26, 37]. This finding becomes relevant when making large-scale predictions about the association between drugs and diseases. It is plausible that the preference for RF implementation in various studies is due to its suitability for larger datasets, as this method excels in the detection of Out-Of-Bag errors, the proximity between features, and the handling of unbalanced datasets [40].

In this sense, the model with the highest AUC ROC was the one developed by Yue and He (2021) [30], however, this study only reported the AUC ROC as a performance measure, other parameters such as sensitivity or specificity were not provided, and had a quality assessment classified as medium. In addition, other databases such as genes, adverse effects or chemical structures were not explored to create the network or establish similarities, which could have enriched the model. This fact was demonstrated by Cao *et al*. (2014) [24], where in their study, the drug-target relationship can be influenced by characteristics of the chemical structure in relation to the structural and physicochemical properties of the targets, greatly helping the discrimination of predictions. Similarly, studies that included adverse effects as part of the data to build the network, generated better predictive results in relation to the other data sources commonly used as a drug-target [10, 25].

Regarding the databases used in the development of the prediction models for DR, all studies incorporated databases that linked drugs to their targets. In addition, some authors explored the inclusion of information to understand the topology and similarities of their network. For example, Cao *et al*. (2014) [24], developed the model with the highest AUC ROC (0.99) among those who evaluated the chemical structures of drugs. This performance measure outperforms other studies that included additional information, such as genes and diseases, like Liu *et al.* (2020) [12] with AUC ROC of 0.97 or the protein-disease interaction [35] with an AUC ROC of 0.96; protein-protein [36] with an AUC ROC of 0.94; the inclusion of RNA [26] with an AUC ROC of 0.92, or the integration of adverse drug effects by Wang *et al.* (2013) [10] and Wei *et al.* (2019) [9], both with AUC ROC of 0.90. The observation that the inclusion of chemical structures obtained a higher AUC ROC value may be linked to what Cao *et al.* (2014) [24] defined. In this study, they suggest that the drug-target relationship may be influenced by characteristics of the chemical structure and its relationship with the properties of the targets, which helps in the discrimination of predictions.

A limitation of models that used supervised learning to make predictions is their need for both positive and negative training data. However, most data sources allow positive therapeutic relationships between drugs and diseases to be established but lack the ability to determine negative relationships between them. Therefore, one of the main challenges when defining a drug repositioning model lies in obtaining reliable negative data for training. To address this issue, many models chose to take unlabeled data or unrelated pairs (drug-targets) as negative examples. However, this strategy may introduce some possible positive drug-disease pairs into negative samples, generating noisy training data and decreasing the reliability of predictions. In contrast, studies such as that by Amiri-Souri *et al.* (2022) [33] and Liu *et al*. (2020) [12] adopted a more careful and realistic strategy when selecting negative training data, which contributed

to improved model performance compared to strategies that took unlabeled data randomly. Despite these advances, the study by Liu *et al.* (2020) [12] emphasizes the need to consider even more reliable strategies when employing supervised learning models in drug repositioning.

Regarding the validation of the results of the models, the most common strategy was the confirmation of predictions in scientific databases, where the main purpose was to identify references where the relationship between the repositioned drug and its new disease had previously been established. In some cases, the literature validation took a more focused approach to the study of the repositioned drug. This included the exploration of genes related to the disease, and the review of information related to the metabolic pathways in which the drug might be involved. This approach made it possible to determine whether the repositioned drug could be associated with the same pathways that lead to the treatment of the disease [10]. In addition, another strategy was to compare the new drug with those currently used to treat the disease, looking for similarities at the molecular or metabolic level to evaluate its potential effectiveness in treatment [10].

In some cases, the validations went beyond the literature review. An example is the study by Fahimian *et al*. (2020) [27], which carried out a structural comparative analysis between new molecules repositioned to treat breast cancer and those already approved for this indication. In addition, they conducted an *in vitro* study to demonstrate the efficacy of the newly repositioned drug in the treatment of this pathology [27]. Another highlight is the study by Amiri-Souri *et al*. (2022) [33], which performed a simulation with molecular docking and an exhaustive literature review to validate the new interactions between target proteins and drugs predicted by the SML model for cancer treatment.

In short, the application of an integrative approach, including multiple sources of biomolecules and validation of results using various methodologies, not only proposes promising candidates to treat various diseases, but also plays an important role in the understanding of effective therapies that influence multifactorial diseases. These models, apart from predicting new indications, can contribute significantly to clinical pharmacogenomics research, defining relationships between drugs–gene and disease–gene [22, 26]. Likewise, the repositioning of drugs can be very useful for the development of knowledge regarding adverse effects since, by discovering possible new indications for medicines, it is easier to conduct studies more focused on the mechanisms of action of drugs, reducing the gap between medical indications and the understanding of the effects of drugs [10].

Conclusions

This systematic review and meta-analysis of the published literature to predict drug repositioning by using SML models presents relevant aspects that may be useful for future studies in this field. This review suggests that SML models can predict the repositioning of drugs with high performance values, which underscores the potential use of RF in DR. It was identified that the inclusion of chemical structures in the development of the model can improve performance, which allows us to suggest the evaluation of this parameter in future studies. It is suggested to explore how the drug repositioning prediction could contribute to better understand multifactorial diseases and adverse drug effects in future studies. It should be noted that the RF model was the most widely used SML model to predict drug repositioning and with the best performance results compared to AUC ROC. It recognizes that it is important to define strategies to define negative samples when training SML models so as not to incur the increased risk of false negatives. Finally, it could be interesting to study the DL models developed to predict new indications for medications, given that its publications have considerably increased since 2019.

Acknowledgments

We would like to thank EAFIT University and the Master's Degree in Biosciences program. We would like to thank Professor Nicolas Pinel for his guidance on this work.

Conflict of interest

The authors report that they have no conflict of interest.

References

1. M. Schlander, K. Hernandez-Villafuerte, C.Y. Cheng, J. Mestre-Ferrandiz, M. Baumann, How much does it cost to research and develop a new drug? A systematic review and assessment, *Pharmacoeconomics*, **39**, 1243-1269 (2021). Doi: https://doi.org/10.1007/s40273-021-01065-y

2. PhRMA, *Research and Development Policy Framework*. URL: https://phrma.org/policy-issues/Research-and-Development-Policy-Framework, accessed May 2023.

3. D. Sun, W. Gao, H. Hu, S. Zhou, Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, **12**(7), 3049-3062 (2022). Doi: https://doi.org/10.1016/j.apsb.2022.02.002

4. H. Luo, M. Li, M. Yang, F.X. Wu, Y. Li, J. Wang, Biomedical data and computational models for drug repositioning: A comprehensive review, *Briefings in Bioinformatics*, **22**(2), 1604-1619 (2021). Doi: https://doi.org/10.1093/bib/bbz176

5. M. Rudrapal, S.J. Khairnar, A.G. Jadhav, Drug repurposing (DR): An emerging approach in drug discovery, in: Badria, F.A., *Drug Repurposing - Hypothesis, Molecular Aspects and Therapeutic Applications*, IntechOpen Limited, London, 2020. Doi: https://doi.org/10.5772/intechopen.93193

6. B. Turanli, M. Grøtli, J. Boren, J. Nielsen, M. Uhlen, K.Y. Arga, A. Mardinoglu, Drug repositioning for effective prostate cancer treatment, *Frontiers in Physiology*, **9**, 500 (2018). Doi: https://doi.org/10.3389/fphys.2018.00500

7. J. Li, S. Zheng, B. Chen, A.J. Butte, S.J. Swamidass, Z. Lu, A survey of current trends in computational drug repositioning, *Briefings in Bioinformatics*, **17**(1), 2-12 (2016). Doi: https://doi.org/10.1093/bib/bbv020

8. J.L. Medina-Franco, M.A. Giulianotti, G.S. Welmaker, R.A. Houghten, Shifting from the single- to the multitarget paradigm in drug discovery, *Drug Discovery Today*, **18**(9-10), 495-501 (2013). Doi: https://doi.org/10.1016/j.drudis.2013.01.008

9. X. Wei, Y. Zhang, Y. Huang, Y. Fang, Predicting drug–disease associations by network embedding and biomedical data integration, *Data Technologies and Applications*, **53**(2), 217-229 (2019). Doi: https://doi.org/10.1108/dta-01-2019-0004

10. Y. Wang, S. Chen, N. Deng, Y. Wang, Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data, *PLoS One*, **8**(12), e78518 (2013). Doi: https://doi.org/10.1371/journal.pone.0078518

11. C. Gilvary, J. Elkhader, N. Madhukar, C. Henchcliffe, M.D. Goncalves, O. Elemento, A machine learning and network framework to discover new indications for small molecules, *PLoS Computational Biology*, **16**(8), e1008098 (2020). Doi: https://doi.org/10.1371/journal.pcbi.1008098

12. J. Liu, Z. Zuo, G. Wu, Link prediction only with interaction data and its application on drug repositioning, *IEEE Transactions in NanoBioscience*, **19**(3), 547-555 (2020). Doi: https://doi.org/10.1109/TNB.2020.2990291

13. H. Ding, I. Takigawa, H. Mamitsuka, S. Zhu, Similarity-based machine learning methods for predicting drug–target interactions: A brief review, *Briefings in Bioinformatics*, **15**(5), 734-747 (2014). Doi: https://doi.org/10.1093/bib/bbt056

14. M.L. Shahreza, N. Ghadiri, S.R. Mousavi, J. Varshosaz, J.R. Green, A review of network-based approaches to drug repositioning, *Briefings in Bioinformatics*, **19**(5), 878-892 (2018). Doi: https://doi.org/10.1093/bib/bbx017

15. A.B. Haidich, Meta-analysis in medical research, *Hippokratia*, **14**(Suppl. 1), 29-37 (2010). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/, accessed May 2023.

16. M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J.M. Grimshaw, A. Hróbjartsson, *et al.*, PRISMA 2020 statement: an updated guideline for reporting systematic reviews, *BMJ*, **372**, 71 (2021). Doi: https://doi.org/10.1136/bmj.n71

17. M. Ouzzani, H. Hammady, Z. Fedorowicz, A. Elmagarmid, Rayyan -A web and mobile app for systematic reviews, *Systematic Reviews*, **5**, 210 (2016). Doi: https://doi.org/10.1186/S13643-016-0384-4

18. Supervised learning — SciKit-Learn 1.4.0 documentation. URL: https://scikit-learn.org/stable/supervised_learning.html, accessed January, 2024.

19. M.I. Azeem, F. Palomba, L. Shi, Q. Wang, Machine learning techniques for code smell detection: A systematic literature review and meta-analysis, *Information and Software Technology*, **108**, 115-138 (2019). Doi: https://doi.org/10.1016/j.infsof.2018.12.009

20. R. DerSimonian, N. Laird, Meta-analysis in clinical trials, *Controlled Clinical Trials*, **7**(3), 177-188 (1986). Doi: https://doi.org/10.1016/0197-2456(86)90046-2

21. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2023. URL: https://lib.stat.cmu.edu/R/CRAN/doc/manuals/r-devel/fullrefman.pdf, accessed January, 2024.

22. A. Qabaja, M. Alshalalfa, E. Alanazi, R. Alhajj, Prediction of novel drug indications using network driven biological data prioritization and integration, *Journal of Cheminformatics*, **6**, 1 (2014). Doi: https://doi.org/10.1186/1758-2946-6-1

23. T. Lee, Y. Yoon, Drug repositioning using drug-disease vectors based on an integrated network, *BMC Bioinformatics*, **19**, 446 (2018). Doi: https://doi.org/10.1186/S12859-018-2490-X

24. D.-S. Cao, L.-X. Zhang, G.-S. Tan, Z. Xiang, W.-B. Zeng, Q.-S. Xu, A.F. Chen, Computational prediction of drug target interactions using chemical, biological, and network features, *Molecular Informatics*, **33**(10), 669-681 (2014). Doi: https://doi.org/10.1002/minf.201400009

25. H. Moghadam, M. Rahgozar, S. Gharaghani, Scoring multiple features to predict drug disease associations using information fusion and aggregation, *SAR and QSAR in Environment Research*, **27**(8), 609-628 (2016). Doi: https://doi.org/10.1080/1062936X.2016.1209241

26. Z.-H. Chen, Z.-H. You, Z.-H. Guo, H.-C. Yi, G.-X. Luo, Y.-B. Wang, Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model, *Frontiers in Bioengineering and Biotechnology*, **8**, 338 (2020). Doi: https://doi.org/10.3389/fbioe.2020.00338

27. G. Fahimian, J. Zahiri, S.S. Arab, R.H. Sajedi, RepCOOL: Computational drug repositioning via integrating heterogeneous biological networks, *Journal of Translational Medicine*, **18**, 375 (2020). Doi: https://doi.org/10.1186/S12967-020-02541-3

28. R. Zhou, Z. Lu, H. Luo, J. Xiang, M. Zeng, M. Li, NEDD: A network embedding based method for predicting drug-disease associations, *BMC Bioinformatics*, **21**, 387 (2020). Doi: https://doi.org/10.1186/S12859-020-03682-4

29. C. Gilvary, J. Elkhader, N. Madhukar, C. Henchcliffe, M.D. Goncalves, O. Elemento, A machine learning and network framework to discover new indications for small molecules, *PLoS Computational Biology*, **16**(8), e1008098 (2020). Doi: https://doi.org/10.1371/journal.pcbi.1008098

30. Y. Yue, S. He, DTI-HeNE: A novel method for drug-target interaction prediction based on heterogeneous network embedding, *BMC Bioinformatics*, **22**, 418 (2021). Doi: https://doi.org/10.1186/S12859-021-04327-W

31. J. Yang, D. Zhang, L. Liu, G. Li, Y. Cai, Y. Zhang, H. Jin, X. Chen, Computational drug repositioning based on the relationships between substructure–indication, *Briefings in Bioinformatics*, **22**(4), bbaa348 (2021). Doi: https://doi.org/10.1093/bib/bbaa348

32. S. Kitsiranuwat, A. Suratanee, K. Plaimas, Multi-data aspects of protein similarity with a learning technique to identify drug-disease associations, *Applied Sciences* (Basel), **11**(7), 2914 (2021). Doi: https://doi.org/10.3390/app11072914

33. E. Amiri-Souri, R. Laddach, S.N. Karagiannis, L.G. Papageorgiou, S. Tsoka, Novel drug-target interactions via link prediction and network embedding, *BMC Bioinformatics*, **23**, 121 (2022). Doi: https://doi.org/10.1186/S12859-022-04650-W

34. B.-W. Zhao, L. Hu, Z.-H. You, L. Wang, X.-R. Su, HINGRL: predicting drug–disease associations with graph representation learning on heterogeneous information networks, *Briefings in Bioinformatics*, **23**(1), bbab515 (2022). Doi: https://doi.org/10.1093/bib/bbab515

35. M.-L. Zhang, B.-W. Zhao, X.-R. Su, Y.-Z. He, Y. Yang, L. Hu, RLFDDA: a meta-path based graph representation learning model for drug–disease association prediction, *BMC Bioinformatics*, **23**, 516 (2022). Doi: https://doi.org/10.1186/S12859-022-05069-Z

36. S. Kitsiranuwat, A. Suratanee, K. Plaimas, Integration of various protein similarities using random forest technique to infer augmented drug-protein matrix for enhancing drug-disease association prediction, *Science Progress*, **105**(3), 1-30 (2022). Doi: https://doi.org/10.1177/00368504221109215

37. H. Jiang, Y. Huang, An effective drug-disease associations prediction model based on graphic representation learning over multi-biomolecular network, *BMC Bioinformatics*, **23**, 9 (2022). Doi: https://doi.org/10.1186/S12859-021-04553-2

38. A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**(7), 1145-1159 (1997). Doi: https://doi.org/10.1016/S0031-3203(96)00142-2

39. J.J. Deeks, J.P.T. Higgins, D.G. Altman, Chapter 10: Analysing data and undertaking meta-analyses, Cochrane Training: Online Learning. URL: https://training.cochrane.org/handbook/current/chapter-10#section-10-10, accessed: January 2024.

40. M. Zakariah, Classification of large datasets using Random Forest Algorithm in various applications: Survey, *International Journal of Engineering and Innovative Technology (IJEIT)*, **4**(3), 2277-3754 (2008). URL: https://faculty.ksu.edu.sa/sites/default/files/classification_of_large_datasets_using_random.pdf

Supplementary Archive

Supplementary Table 1. Full search syntax for each database

PubMed	
Search	Syntax
#1	("Network*"[Title/Abstract])
#2	(("drug repurpos*"[Title/Abstract]) OR ("drug redirecting"[Title/Abstract]) OR ("drug reposition*"[Title/Abstract]) OR ("drug retasking"[Title/Abstract]) OR ("drug reprofiling"[Title/Abstract]) OR ("drug retargeting"[Title/Abstract]) OR ("drug relocation"[Title/Abstract]) OR ("Drug re-profiling"[Title/Abstract]) OR ("drug re-tasking"[Title/Abstract]) OR ("drug rescue"[Title/Abstract]) OR ("indication expansion"[Title/Abstract]) OR ("indication switching"[Title/Abstract]) OR ("drug rescuing"[Title/Abstract]) OR ("drug recycling"[Title/Abstract]) OR (drug redirection[Title/Abstract]) OR ("therapeutic switching"[Title/Abstract]) OR ("Novel drug us*"[Title/Abstract]) OR ("novel drug rediscovery"[Title/Abstract]) OR ("Novo drug us*"[Title/Abstract]) OR ("novo drug rediscovery"[Title/Abstract]))
#3	#1 and #2
SCOPUS	
Search	Syntax
#1	TITLE-ABS ("Network*")
#2	TITLE-ABS (("drug reposition*") OR ("drug repurpos*") OR ("drug redirecting") OR ("drug reposition*") OR ("drug retasking") OR ("drug reprofiling") OR ("drug retargeting") OR ("drug relocation") OR ("Drug re-profiling") OR ("drug re-tasking") OR ("drug rescue") OR ("indication expansion") OR ("indication switching") OR ("drug rescuing") OR ("drug recycling") OR (drug redirection) OR ("therapeutic switching") OR ("Novel drug us*") OR ("novel drug rediscovery") OR ("Novo drug us*") OR ("novo drug rediscovery"))
#3	#1 and #2

Supplementary Table 2. Databases and libraries used in the articles.

ID	Drug–Target	Drug–Disease	Protein–Disease	Protein–Protein	Side effects	Chemical structure	Gen - Disease	RNA
2	KEGG BRITE/ Brenda Supertarget/ DrugBank	-	-	-	CIDER	PubChem	OMIM	-
4	The Meaning of Meaning / Meaning	-	-	-	-	MACCS	-	-
5	DrugBank	-	-	-	-	-	OMIM, Reactome database STRING	-
8	-	ClinicalTrials.gov / DrugBank/ OMIM	-	-	CIDER	SMILES	-	-
12	DrugBank	CTD	-	-	-	-	DisGeNet BioCarta/ Reactome/ PID/ KEGG	-
16	DrugBank	mimMiner/ FDA approved drugs	-	-	CIDER	SMILES	OMIM, mimMiner	-
18	DrugBank	CTD	DisGeNET	STRING	-	-	-	LncRNA2Target/ LncRNADisease / lncRNASNP2/ miRTarBase/ HMDD / lncRNASNP2
20	DrugBank	PREDICT	-	-	-	-	OMIM	-
21	DrugBank	-	-	IntAct	-	-	OMIM/CTD/ DisGeNET/ COXPRESdb	-
23	-	PREDICT/ MeSH	-	-	-	SMILES	-	-

(*Continued*)

ID	Drug–Target	Drug–Disease	Protein–Disease	Protein–Protein	Side effects	Chemical structure	Gen - Disease	RNA
25	DrugBank	FDA approved drugs/ DrugBank/ PubChem	-	-	-	SMILES	MSigDB	-
28	Yamanishi	Olayan	-	-	-	-	-	-
29	DrugBank/ Uniprot	-	-	-	-	PubChem	-	-
31	DrugBank	CTD	-	STRING	-	-	DisGeNET	-
32	The Meaning / ChEMBL	-	-	-	-	MACCS	-	-
33	DrugBank	CTD	DisGeNET	-	-	SMILES	-	-
34	DrugBank	CTD/ MeSH	DisGeNET	-	-	SMILES	-	-
35	DrugBank/ KEGG	CTD	-	STRING	-	DD.chem.data	DisGeNET	-
36	DrugBank	SCMFDD-S/ MeSH	-	STRING	-	SMILES	DisGeNET	LncRNA2Target/ LncRNADisease/ lncRNASNP2/ miRTarBase/ HMDD/ lncRNASNP2

KEGG: Kyoto Encyclopedia of Genes and Genomes / Brenda: the enzyme database / Yamanishi: enzymes, ion channels, G-protein-coupled receptors and, nuclear receptors [Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug–target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics*, 24(13), i232–i240 (2008). Doi: https://doi.org/10.1093/bioinformatics/btn162] / Uniprot: Universal Protein Resource / CTD: Comparative Toxicogenomics Database / Olayan: FDA-approved drugs and human target proteins [R.S. Olayan, H. Ashoor, V.B. Bajic, DDR: Efficient computational method to predict drug-target interactions using graph mining and machine learning approaches, *Bioinformatics*, 34(7), 1164–1173 (2018). Doi: https://doi.org/10.1093/bioinformatics/btx731] / PREDICT [A. Gottlieb, G.Y. Stein, E. Ruppin, R. Sharan, PREDICT: A method for inferring novel drug indications with application to personalized medicine, *Molecular Systems Biology*, 7, 496 (2011). Doi: https://doi.org/10.1038/msb.2011.26] / SCMFDD-S: similarity constrained matrix factorization method for the drug-disease association prediction data set collected by Zhang *et al.* [W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, F. Liu, Predicting drug-disease associations by using similarity constrained matrix factorization, *BMC Bioinformatics*, 19, 233 (2018). Doi: https://doi.org/10.1186/S12859-018-2220-4] / IntAct: Molecular Interaction Database / SIDER: Side Effect Resource / PID: the Pathway Interaction Database / MACCS: Molecular ACCess System / SMILES: Simplified Molecular Input Line Entry System / OMIM: Online Mendelian Inheritance in Man / STRING: Search Tool for the Retrieval of Interacting Genes-Proteins / MeSH: Medical Subject Headings / HMDD: the Human microRNA Disease Database.

Supplementary Table 3. Quality of evidence assessment

ID	1. Specify type of model, model-building procedures, and method for internal validation.		2. Specify measures used to assess model performance and, to compare multiple models.		3. Discuss the results with reference to performance and any other validation data.		4. Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.		Total
	Concept	Score	Concept	Score	Concept	Score	Concept	Score	
2	Yes	1	Yes	1	Yes	1	Yes	1	4
4	Yes	1	Yes	1	Yes	1	Yes	1	4
5	Partially	0,5	No	0	No	0	Partially	0,5	1
8	Yes	1	Yes	1	Yes	1	Yes	1	4
12	Yes	1	Partially	0,5	Partially	0,5	Partially	0,5	2,5
16	Yes	1	Yes	1	Yes	1	Yes	1	4
18	Yes	1	Yes	1	Yes	1	No	0	3
20	Yes	1	Yes	1	Yes	1	Yes	1	4
21	Yes	1	Yes	1	Partially	0,5	Partially	0,5	3
23	Yes	1	Partially	0,5	Yes	1	Yes	1	3,5
25	Yes	1	Partially	0,5	Yes	1	Yes	1	3,5
28	Yes	1	Partially	0,5	Yes	1	Yes	1	3,5
29	Partially	0,5	Partially	0,5	Yes	1	Yes	1	3
31	Yes	1	Yes	1	Yes	1	Yes	1	4
32	Yes	1	Yes	1	Yes	1	Yes	1	4
33	Yes	1	Yes	1	Yes	1	Yes	1	4
34	Yes	1	Yes	1	Yes	1	Yes	1	4
35	Yes	1	Yes	1	Yes	1	Partially	0,5	3,5
36	Yes	1	Yes	1	Yes	1	Yes	1	4

How to cite this article

D.J. García-Marín, J.A. García-Zea, The random forest machine learning model performs better in predicting drug repositioning using networks: Systematic review and meta-analysis, *Rev. Colomb. Cienc. Quim. Farm.*, 53(2), 354-384 (2024). https://doi.org/10.15446/rcciquifa.v53n2.114447