Computational and Structural Biotechnology Journal 19 (2021) 4559–4573

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A data-driven methodology towards evaluating the potential of drug repurposing hypotheses

Lucía Prieto Santamaría [a,b,c,*,1], Esther Ugarte Carro [a,1], Marina Díaz Uzquiano [a], Ernestina Menasalvas Ruiz [a,b], Yuliana Pérez Gallardo [c], Alejandro Rodríguez-González [a,b]

[a] *Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain*
[b] *ETS Ingenieros Informáticos, Universidad Politécnica de Madrid, 28660 Boadilla del Monte, Madrid, Spain*
[c] *Ezeris Networks Global Services S.L., 28028 Madrid, Spain*

ARTICLE INFO

Article history:
Received 20 May 2021
Received in revised form 8 July 2021
Accepted 3 August 2021
Available online 9 August 2021

Keywords:
Drug repurposing
Drug repositioning
Data integration
Disease understanding
DISNET knowledge base
Drug-disease validation

ABSTRACT

Drug repurposing has become a widely used strategy to accelerate the process of finding treatments. While classical *de novo* drug development involves high costs, risks, and time-consuming paths, drug repurposing allows to reuse already-existing and approved drugs for new indications. Numerous research has been carried out in this field, both *in vitro* and *in silico*. Computational drug repurposing methods make use of modern heterogeneous biomedical data to identify and prioritize new indications for old drugs. In the current paper, we present a new complete methodology to evaluate new potentially repurposable drugs based on disease-gene and disease-phenotype associations, identifying significant differences between repurposing and non-repurposing data. We have collected a set of known successful drug repurposing case studies from the literature and we have analysed their dissimilarities with other biomedical data not necessarily participating in repurposing processes. The information used has been obtained from the DISNET platform. We have performed three analyses (at the genetical, phenotypical, and categorization levels), to conclude that there is a statistically significant difference between actual repurposing-related information and non-repurposing data. The insights obtained could be relevant when suggesting new potential drug repurposing hypotheses.

1. Introduction

The process of giving new uses for already-existing and approved drugs is known after the name of drug repurposing or repositioning (DR) [1]. It has become a promising strategy, attracting attention to both the pharmaceutical and the research community, due to the notorious advantages over *de novo* drug discovery. *De novo* drug design is a tremendously time-consuming, costly, and difficult task. Generally, the process takes 17 years and costs an average of US$2.6 billion. Furthermore, only approximately 2.01% of all drug development candidates finally make it to the market as a successful treatment [2]. The current increase in bioinformatic knowledge and omics data, makes DR a potential alternative for drug discovery, reducing time, investment, and risks [3].

Throughout history, multiple DR cases have been discovered by serendipity. For instance, drugs with certain side effects that resulted to be indications for another diseases [4]. In the 1970s, Minoxidil was approved to treat hair loss when it was planned to treat hypertension. In the 1990s and 2000s, Sildenafil, which was aimed for the

Abbreviations: ACE, Angiotensin I Converting Enzyme; AHR, Aryl Hydrocarbon Receptor; ALK, Anaplastic Lymphoma Kinase; API, Application Programming Interface; CMap, Connectivity Map; COX-2, Cyclooxygenase 2; CUI, Concept Unique Identifier; DR, Drug Repurposing or Drug Repositioning; DRD3, Dopamine Receptor D3; ESR1, Estrogen Receptor 1; ESR2, Estrogen Receptor 2; FCGR2A, Fc Fragment Of IgG Receptor IIa; FCGR3A, Fc Fragment Of IgG Receptor IIIa; FCGR3B, Fc Fragment Of IgG Receptor IIIb; GDA, Gene Disease Association; ICD-10-CM, International Classification of Diseases, 10th revision, Clinical Modification; ID, Identifier; KDR, Kinase insert Domain Receptor; LTα, Lymphotoxin alpha; MeSH-PA, Medical Subject Headings – Pharmacological Action; ND, New Disease; NLM, National Library of Medicine; OD, Original Disease; PTGS2, Prostaglandin-endoperoxidase synthase 2; SM, Supplementary Material; SRD5A1, Steroid 5 Alpha-Reductase 1; SRD5A2, Steroid 5 Alpha-Reductase 2; TNFα, Tumour Necrosis Factor alpha; UMLS, Unified Medical Language System.

[*] Corresponding author.
E-mail address: alejandro.rg@upm.es (A. Rodríguez-González).

[1] The authors contributed equally to the manuscript.

https://doi.org/10.1016/j.csbj.2021.08.003
2001-0370/

treatment of angina, was repositioned for erectile dysfunction. Duloxetine, which was originally intended for major depressive disorder, was then used for stress urinary incontinence [2].

Identifying the right drug for an indication of interest with a high level of confidence is complicated. The approaches to generate DR hypotheses are divided into: (i) experimental or activity-based, (ii) computational or *in silico* and (iii) literature-based or existing knowledge methods. On the one hand, experimental strategies consist in testing drugs in assays. These methods are based on available comprehensive clinical compound databases, requiring an entire collection of drugs. They are laborious and expensive processes [5]. Binding assays to identify target interactions and phenotypic screening stand up in this category [6]. On the other hand, computational DR employs online public databases and bioinformatic tools to detect interactions between drugs, targets, and diseases. *In silico* strategies reduce investment and time, but need structural information of target proteins and drug-induced cell/disease phenotypes [7]. There are different methods covered by these approaches: retrospective clinical analysis, signature matching, molecular docking, Genome-Wide Association Studies (GWAS), pathway mapping or network analysis [6], data mining, and machine learning [5], among others. Because of their capacity to integrate multiple data sources, we point out network-based approaches, widely used since approximately ten years ago [3]. Networks can shed light on drugs, diseases, and targets modes of action and relationships, aiding to identify therapeutic potentials and uncover DR applications [8–11].

The principal objective of the current research is to propose a new methodology for validating DR cases through biomedical integrated data. That is, to identify clear differences in the phenotypic and genetic DR and non-DR data. This way, future hypothetical DR cases could be suggested and prioritized in the light of the named discovered differences. To achieve this objective, we have used biomedical-integrated data, in particular, the DISNET project knowledge base. DISNET is a large complex network that stores information about diseases, symptoms, genes, and drugs extracted from different public sources. The integration of this data can uncover novel patterns and associations, and lead to hypotheses for new DR case studies. A schematic illustration of DISNET 3 levels database is provided in Fig. 1.

The manuscript is organized as follows: Section 2 explains the materials and methods used for the validation analysis, Section 3 and Section 4 respectively present and discuss the results obtained, and Section 5 details the conclusions.

2. Materials and methods

To develop a complete methodology to evaluate the potential of new DR hypotheses by means of biomedical integrated data, a study has been carried out following the stages shown in Fig. 2. Firstly, a list of successful DR cases was generated from literature evidence (Section 2.1). Afterwards, data pre-processing was implemented to translate those DR cases to DISNET vocabularies (Section 2.2 and 2.3). Then, we analysed the genes (Section 2.4), symptoms (Section 2.5), and categories (Section 2.6) to conclude whether the DR cases presented significant differences with non-DR data. The two first analyses sought to study the therapeutic potential of the underlying mechanisms of action of genes and symptoms that were shared between the original and the new indications for the drug. The category analysis was carried out to find patterns within the disease and drug groups.

2.1. DR successful known cases

For this study, we selected some important DR cases to check if they showed clear differences with the other biomedical data. To choose those successful known DR cases, the following methodology was followed. The combination of “drug repositioning” and “examples” or “drug repositioning” AND “review” was searched in Google Scholar without including a time interval specification (dated 01/07/2021). The first 30 entries, ordered by relevance, were examined looking for publications that contained tables with DR examples, not adapted from previous revisions but original from those articles. The cases and publications must accomplish the following conditions:

- Cases should consist of “Drug name + Original condition + New condition” triples. A condition should be a concept that maps to a disease, but not, for example, to the role of the drug in the disease. Original and new indications must be clearly differentiated. The publications discarded due to the non-completion of this condition were [12,13].
- Publications should not only be focused on one type of disease or a specific condition (e.g., cancer or Arthritis Rheumatoid). More examples of dismissed works as they focused on particular conditions or types of conditions were [5,7,14–17].
- Drugs in the cases should be approved, not investigational. Some of the cases of [8,18] were discarded as they were not accepted.

The 30 first result entries obtained from querying Google Scholar can be found in Table 1. The publications in yellow met the necessary conditions to be used for the proposed analysis. *Ashburn et al. and Novac, N.*, are referenced in *Xue et al’s* table, so, eventually, they were not selected since the cases described in them were already included in the previous paper.

2.2. DISNET data acquisition and integration

The DISNET project incorporates biomedical disease knowledge from public textual and structured sources to include, among others, data regarding diseases, related symptoms, genes, drugs, drug targets [37]. All these data are organized in 3 levels: the phenotypical layer (containing mainly disease-symptoms associations), the biological layer (containing diseases’ associations to genes and proteins, among others), and the drugs layer (that stores drug-related data, including their associations to diseases and the drugs targets).

Since the very aim of the present work is to suggest a new methodology to validate the potential of new possible DR cases by means of integrated biomedical knowledge, we have worked with DISNET database information (diseases, symptoms, genes, drugs, drug targets and their relationships have been studied). Details regarding the typology of the data are in Table 2.

2.3. Pre-processing of DR cases

The pre-processing of repositioning cases needed to have literature concepts translated to the different codifications present in DISNET. DISNET is the result of querying and mining different data sources and, accordingly, it registers different vocabularies. Literature concepts were not normalized so they could not be found directly in DISNET. Therefore, it was necessary to translate them to DISNET’s vocabularies. For the analysis of the genes, symptoms and categories, only the DR cases having this information were kept. The complete process is summarized in Fig. 3.

From literature-extracted DR cases, we (i) obtained the UMLS (Unified Medical Language System) Concept Unique Identifier (CUI) identifying conditions: the original and new indications described in the literature were searched in UMLS’ API [43] to get their corresponding CUIs. The UMLS API retrieves CUIs when searching by code or term, having a properly-authenticated user.

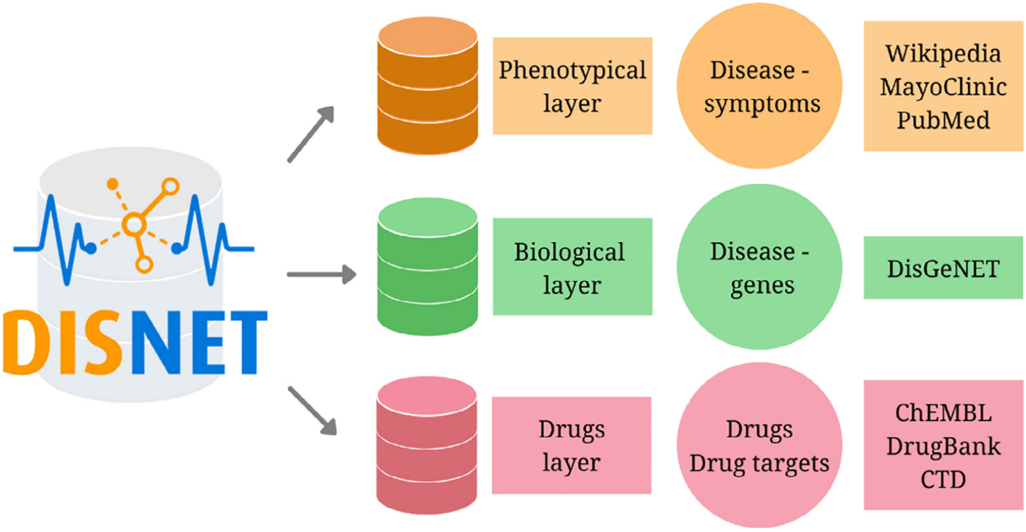

Fig. 1. DISNET's 3-level database. Data are integrated in 3 layers: the phenotypical (in orange), the biological (in green) and the drugs layer (in pink). The figure provides simplifications of the studied entities and data sources. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

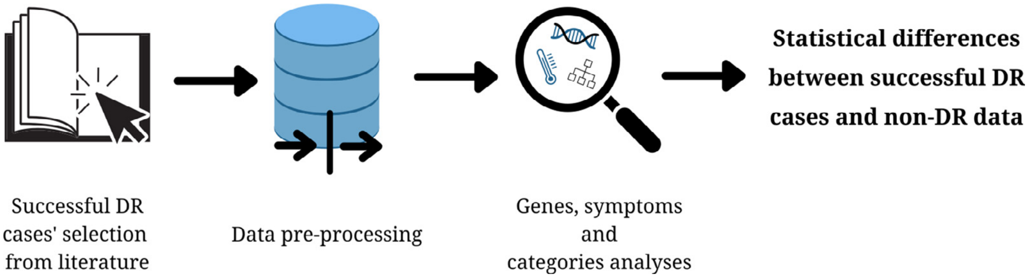
Statistical differences between successful DR cases and non-DR data
Successful DR cases' selection from literature
Data pre-processing
Genes, symptoms and categories analyses

Fig. 2. Pipeline for the evaluation of DR triples. DR cases selected from the literature were pre-processed and then used for genes, symptoms, and categories analyses, in order to confirm DISNET's capacity to find new uses for existing drugs.

In this study, the original and new literature concepts of each DR case were searched. An "exact searchType" query was made, retrieving only concepts that included a synonym that exactly matched the search term. This process was carried out for each publication to have the DR cases standardized (**SM** S2.UMLS_triples.xlsx, sheets "Xue et al.", "Jarada et al." and "Li et al."). At this point, (ii) the coincidences among them were checked and repetitions were discarded keeping the unique triples (**SM** S2.UMLS_triples.xlsx, sheet "Unique Triples").

We (iii) mapped this final list of normalized cases to DISNET (**SM** S3.DISNET_triples.xlsx, sheet "DISNET Vocabularies"). Those DR cases whose conditions corresponded to a disease in DISNET were selected. Therefore, only the ones with a DISNET ID (identifier) were maintained. CUIs allowed us to get gene and category information. And the DISNET IDs helped to obtain the symptoms. Finally, we (iv) deleted those cases without any information about genes, symptoms, or drug-targeted genes.

The pre-processing result was a list of "Drug – Original Disease (OD) – New Disease (ND)" triples (**SM** S3.DISNET_triples.xlsx, sheet "DISNET Final Triples"), namely DR triples. Each triple was composed of: (i) the drug information (its name in the article and in DISNET, and the ChEMBL ID); (ii) the OD information (its name in the article, CUI and CUI Name, and DISNET ID and name) and (iii) the ND information (its name in the article, CUI and CUI Name, and DISNET ID and name).

It is important to highlight the difference between DR cases and triples (Fig. 4). One DR case can be represented by multiple triples since one indication concept can be represented with different CUIs or DISNET IDs. All the combinations were considered for the present validation analysis. As there are different codifications in the DISNET knowledge base, different vocabularies allowed us to obtain the information needed: ChEMBL drug IDs led us to associated drug targets, disease CUIs led us to associated genes, and DISNET IDs led us to associated symptoms.

2.4. Analysis of the genes

One of the ideas underneath DR is that the two diseases involved in it may have some kind of relationship with the gene that encodes the target for the repurposed drug. This gene would be important in the molecular development, pathway, or etiology of a disease, this being one of the main reasons to repurpose a drug for a new disease. Pathway mapping is one of the current DR computational methods, being based on drug or disease networks [2]. This hypothesis has been the basis of the present analysis, depicted in Fig. 5.

We compared the genes shared between the original and new disease in the DR triples. We also extracted drug protein targets and their codifying genes. In this manner, we could detect if the diseases in each DR triple shared the gene targeted by the DR triple drug.

To measure and give weights to disease-gene associations (GDAs) we used DisGeNET scores. Those scores are in-house developed metrics reflecting how well established a particular

Table 1
Examination of the first 30 entries returned by Google Scholar's search engine to identify successful DR cases. We chose those publications that contained tables and met the established conditions. In dark yellow, the papers finally used for the validation analysis are indicated. In pale yellow, entries that accomplished the established requirements but were referenced in the previous ones are depicted. Entries are presented by order of appearance in Google Scholar.

Google Scholar entries	Table with DR examples	Valid conditions
Xue et al., 2018 [3]	X	X
Jarada et al., 2020 [8]	X	X
Lotfi Shahreza et al., 2018 [19]		
Brown and Patel, 2018 [20]		
Luo et al., 2021 [21]		
Novac, N., 2013 [22]	X	X
Hurle et al., 2013 [23]		
Ashburn et al., 2004 [1]	X	X
Rameshrad et al., 2020 [24]		
Turanli et al., 2018 [5]	X	
Corbett et al., 2012 [16]	X	
Ballard et al., 2020 [25]		
Lima et al., 2020 [15]	X	
Turanli et al., 2021 [14]	X	
Dudley et al., 2011 [26]		
Wilkinson and Pritchard, 2015 [27]		
Shim and Liu, 2014 [7]	X	
Masuda et al., 2020 [17]	X	
Liu et al., 2013 [12]	X	
Li et al., 2016 [28]		
Adasme et al., 2021 [29]		
Ma et al., 2013 [30]		
Oprea and Overington, 2015 [31]		
Nzila et al., 2011 [32]		
Yella et al., 2018 [33]	X	
Kharkar et al., 2014 [34]		
Li and Jones, 2012 [18]	X	X
Chen et al., 2015 [35]		
Yang and Agarwal, 2011 [36]		
Frail et al., 2015 [13]	X	

association is based on the current knowledge. Varying between 0 and 1, they give highest values to associations that are reported by several databases, by expert-curated resources, and with large numbers of supporting publications (https://www.disgenet.org/dbinfo). DisGeNET scores were extracted for "Original disease – Shared Target Gene" and "New disease – Shared Target Gene" associations, in those DR cases that shared target genes.

At this stage, the goal was to verify if the scores in such DR cases showed stronger associations than the average. That is, if target genes involved in DR cases had a stronger association with their OD and ND than the rest of DISNET GDAs. To assess the statistical significance of the difference between association score values of "OD – Shared Target Gene" and "ND – Shared Target Gene", and all the other GDAs in DISNET, a Welch T-Test has been used. The Welch T-Test is a two-sample location test that is used to test the hypothesis that two populations have equal means having the samples different variances and/or size [44].

Eventually, to confirm the differences between DR cases and non-DR data, 10,000 random "Drug – OD – ND" triples that shared the drug target gene were selected from all the possibilities present in DISNET. The mean of their GDA scores was calculated and compared with the GDA scores of the actual DR cases through a Welch T-Test. This way, we could assess the reliability of new DR hypothesis by using DISNET to check if they have higher GDA scores than the randomly generated triples. It is noteworthy that the generated triples share an association with a gene, which encodes the protein that will be target to a drug. Such pathway is one of the many that can be considered when generating new DR hypothesis. These steps to compare actual DR and DISNET randomly-generated triples are schematically shown in Fig. 6.

2.5. Analysis of the symptoms

At the phenotypical level, it has been proven that diseases that share symptoms may share "something more". Therefore, we have analyzed the phenotypical similarity between diseases, to understand whether two diseases involved in a DR case present higher values of similarity than the rest of the diseases. This is, if when looking at their symptoms, OD-ND pairs had a lower metric distance than other disease pairs in DISNET.

To calculate such distances between all disease combinations, we have built Boolean symptom vectors. For each disease, we have generated a vector in which each dimension represents the presence (1) or absence (0) of an associated symptom. Distance has been measured between all combinations of vectors by means of Jaccard index, which is computed as follows: $d_{Jaccard}(A,B) = 1 - \frac{|A\cap B|}{|A\cup B|}$, being A and B, the two symptom vectors related to the two different diseases. Values of this coefficient close to 0 represent two highly similar diseases phenotypical-wise, whilst values that approximate to 1 would illustrate a more distant disease pair.

A Welch T-Test was used to examine the statistical significance of the phenotypical distance difference in the two disease pair sets: DR OD – ND set and the rest of DISNET disease – disease (D – D) set.

2.6. Categorization patterns of the triples

Discovering patterns within the repositioning cases would help in the discovery of new uses of existing drugs and shed light on uncovered relationships. To this aim, different standards and classifications can help to generalize the multiple cases of DR. DISNET includes several drug and disease classifications. For this study, MeSH-PA Therapeutic Uses [45] and ICD-10-CM [46] were considered, respectively, and analysed in the different DR triples.

MeSH (Medical Subject Headings) is NLM's (National Library of Medicine) vocabulary to index PubMed and Medline publications by drugs, chemicals, diseases, and other biomedical processes. MeSH does not consistently link drugs to diseases or conditions in an explicit manner. Nonetheless, it counts with a relevant Pharmacological Action (PA) association which connects to therapeutic uses which could be mapped to diseases/conditions and processes [47].

ICD (International Classification of Diseases) is a global standard used to classify and monitor causes of injury and death and to identify health trends and statistics. DR ODs and NDs were classified according to ICD-10-CM (International Classification of Diseases, 10th revision, Clinical Modification, https://www.cdc.gov/nchs/icd/icd10cm.htm).

3. Results

Three main publications have provided the DR cases. Xue et al. and Jarada et al. included a review of the important sources, challenges, and opportunities of DR computational approaches. Li and Jones specifically focused on the potential of combining personalized medicine and DR. The collection of successful DR cases selected from the literature has been included in Supplementary Materials (SM) section (S1.Literature_cases.xlsx). DR cases consisted of the drug name, the condition (disease or symptom) for which the drug was indicated, and the new indication for which the drug was repositioned.

DR cases were pre-processed to perform the validation analysis. The final number of cases considered was 79, as shown in Fig. 7. Through the stages of the data pre-processing, we dismissed some cases. Once we obtained the CUIs of the DR diseases, the repeated cases were discarded. Only the unique cases among the three pub-

Table 2
Summary of DISNET's data typology. Entities and relationships are included. A description, DISNET layer, total count, the identifiers' nature, data sources and accessed date are provided.

		Description	DISNET layer	Count	Identifiers	Sources	Access date
Entities	Diseases	Data representing diseases	Phenotypical	9225	DISNET own identifiers	Wikipedia (https://www.wikipedia.org/)Mayo Clinic (https://www.mayoclinic.org/)PubMed (https://pubmed.ncbi.nlm.nih.gov/)	February 2018 – April 2021 (twice a month)
			Biological	24,314	UMLS CUIs	DisGeNET [38] (https://www.disgenet.org/)	May 2020
	Symptoms	Data representing symptoms and phenotypical effects	Phenotypical	2248	UMLS CUIs	Wikipedia Mayo Clinic PubMed	February 2018 – April 2021 (twice a month)
	Genes	Data representing genes	Biological	20,610	NCBI identifiers	DisGeNET	May 2020
	Proteins and Targets	Data representing proteins and drug targets	Biological	18,521	UniProt Accession Numbers	UniProt [39] (https://www.uniprot.org/)	May 2020
			Drugs	1594	ChEMBL identifiers	ChEMBL [40] (https://www.ebi.ac.uk/chembl/)	May 2020
	Drugs	Data representing drugs of different molecular types	Drugs	3944	ChEMBL identifiers	ChEMBL	May 2020
				2540	DrugBank identifiers	DrugBank [41] (https://www.drugbank.com/)	May 2020
Relationships	Disease - Symptom	Associations between diseases and their related symptoms	Phenotypical	211,362	–	Wikipedia Mayo Clinic PubMed	February 2018 – April 2021 (twice a month)
	Disease - Gene	Associations between diseases and their related genes	Biological	358,209	–	DisGeNET	May 2020
	Gene - Protein	Associations between genes and the proteins they encode	Biological	15,770		DisGeNET	May 2020
	Drug - Disease	Associations between diseases that are indications for drugs and drugs	Drugs	628,036	–	CTD [42] (http://ctdbase.org/)	May 2020
	Drug - Target	Associations between drugs and the targets to which they are directed	Drugs	7727	–	ChEMBL DrugBank	May 2020 December 2020

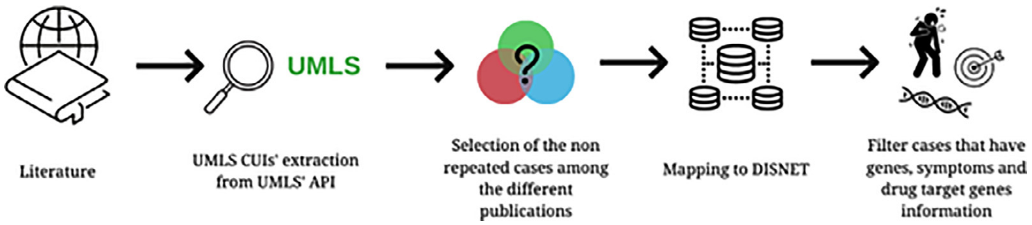
UMLS
Literature
UMLS CUIs' extraction from UMLS' API
Selection of the non repeated cases among the different publications
Mapping to DISNET
Filter cases that have genes, symptoms and drug target genes information

Fig. 3. Data pre-processing pipeline. Once the CUIs of the literature concepts were extracted, repeated cases were discarded. DISNET registers different codifications, thus, literature concepts needed to be translated to their vocabularies in order to carry out the validation analysis. Furthermore, diseases and drugs in the DR cases must fulfil other conditions: they should have at least one associated gene and symptom, and at least one associated target gene, respectively.

lications were used. We also deleted the cases that did not have the information needed for the analyses in the DISNET knowledge base (information about genes, symptoms or drug targets).

Moreover, the numbers of DR triples through pre-processing final stages are indicated in Table 3. As previously explained, one case could correspond to several triples on account of the different codifications present in DISNET.

To summarize, the current study was performed with 79 successful DR cases which were disaggregated in 247 different triples composed of combinations of 40, 34, and 53 different drugs, original and new diseases, respectively. Among the final 79 cases, every "OD – ND" pair shared genes except five. Furthermore, 46 pairs shared the gene that encodes the target protein of the corresponding repositioned drug.

To analyse the distribution of GDA scores in both DR cases and in DISNET in general, we have included Fig. 8. It represents the scores of the associations "OD – Shared Target Gene" and "ND – Shared Target Gene" from the actual DR cases, and the associations "DISNET Disease – Gene" from other DISNET diseases. Relative densities of the GDA scores in each disease set (ODs for "Original Diseases", NDs for "New Diseases" and DISNET Ds for all the other DISNET diseases not encompassed in the previous two) have been depicted along with their mean values (0.32 for ODs, 0.23 for NDs, and 0.14 for DISNET Ds).

As it can be observed in Fig. 8, ODs and NDs associations with their target genes show stronger connections than DISNET's average. The statistical significance of such difference between DR

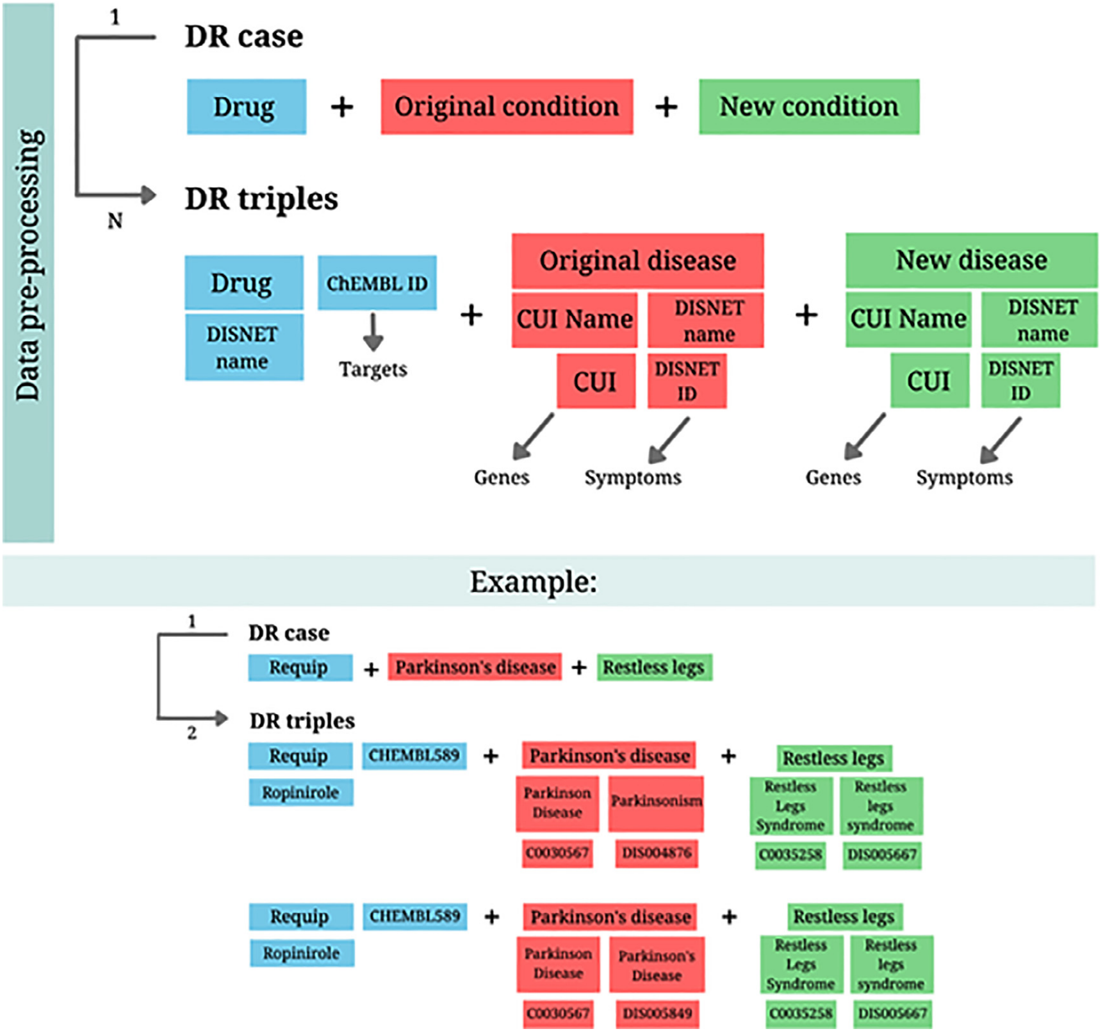

Fig. 4. DR cases vs DR triples. Through data pre-processing, DR cases are transformed to DR triples. One DR case can correspond to different DR triples due to the different codifications present in DISNET.

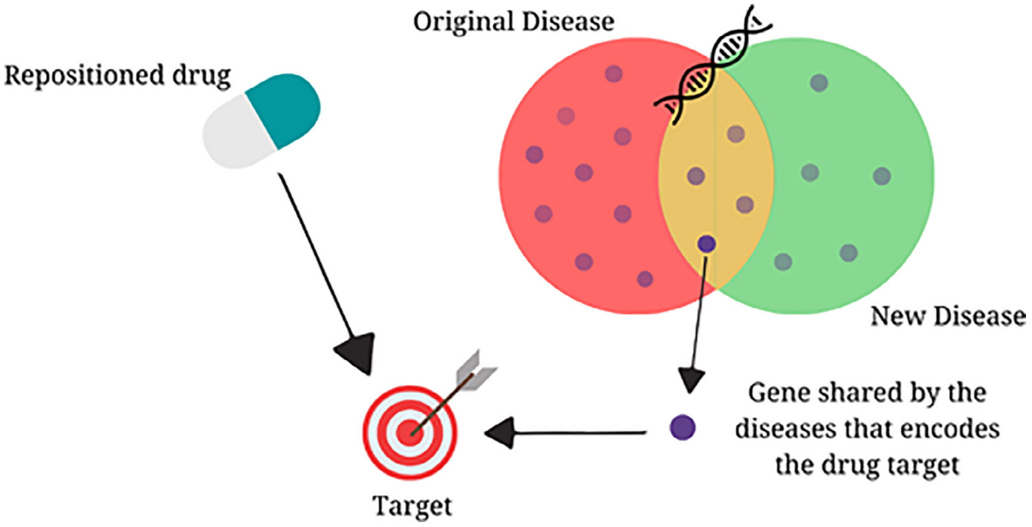

Fig. 5. Hypothesis for evaluating DR cases based on gene analysis. The original and new diseases involved in DR would share (within others) the gene that encodes the drug target.

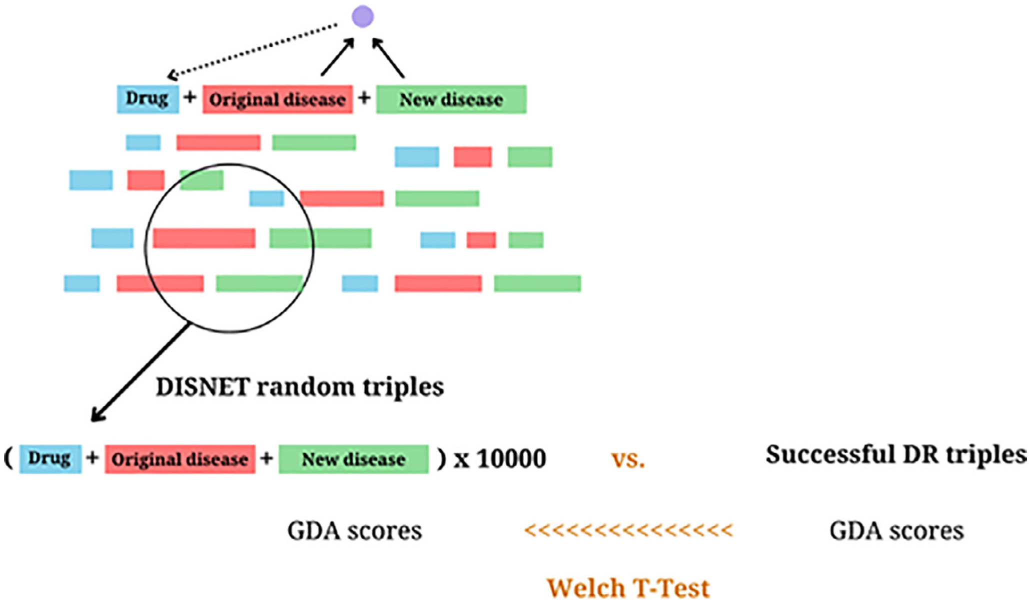

Fig. 6. Comparing GDA scores from DISNET randomly-selected vs actual DR triples. From all the possible "Drug – OD – ND" triples that share a drug target gene (in purple) allocated in DISNET, 10,000 were randomly selected to calculate the mean of their GDA scores. The results were compared with actual DR triples via Welch T-Test. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

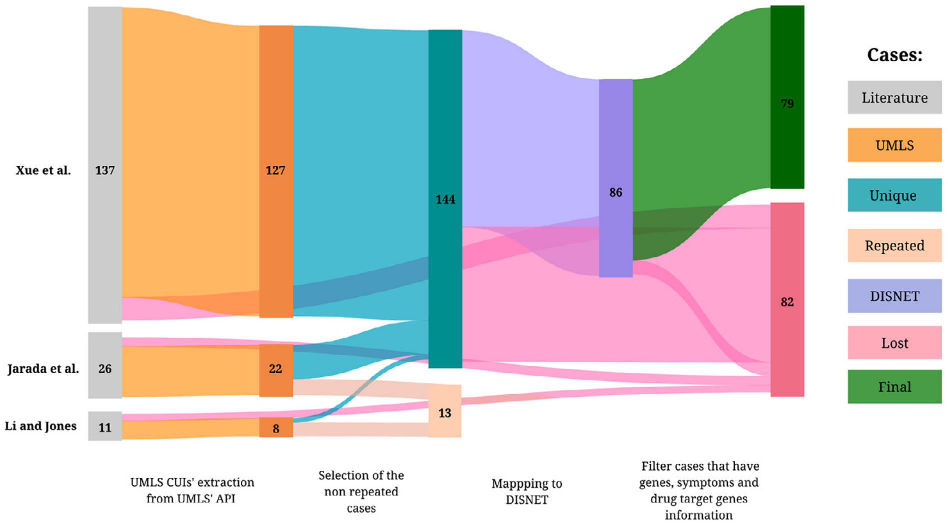

Fig. 7. Evolution of the successful DR cases through the pre-processing. In grey, Xue et al. included 137 cases, Jarada el at., [26], and Li and Jones, [11]. In orange, we searched the CUIs of literature concepts (127, 22, 8) and dropped those without them. In blue, DR cases were normalized, and we selected non-repeated ones (144). In nude, we deleted 13 repetitions. In purple, we mapped the unique cases to DISNET (16) and, finally, in green, picked those with genes, symptoms, and drug target genes data to fulfil the validation analysis (79). In pink, those cases lost through the stages have been indicated (82). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Evolution of successful DR triples through the pre-processing. The unique number of drugs, original and new diseases, and the total number of triples are indicated in each stage starting from the unique ones to the final. Header colours correlate with stages in Fig. 7.

	Unique	DISNET	Final
Triples	789	367	247
Drugs	69	42	40
ODs	68	37	34
NDs	91	58	53

and DISNET's mean GDA score was confirmed by a Welch T-Test for each pair of distributions.

The obtained result (p-values <0.05) confirmed that null hypotheses (that is, "means are equal") were false. Since higher values of the score illustrate stronger gene-disease relationships, associations of DR diseases with shared drug target gene seemed to be more intense than the others in DISNET, given that the means DR GDA scores were greater and statistically significant.

A total of 10,000 "Drug – OD – ND" combinations of diseases sharing associations to a target gene connected to the

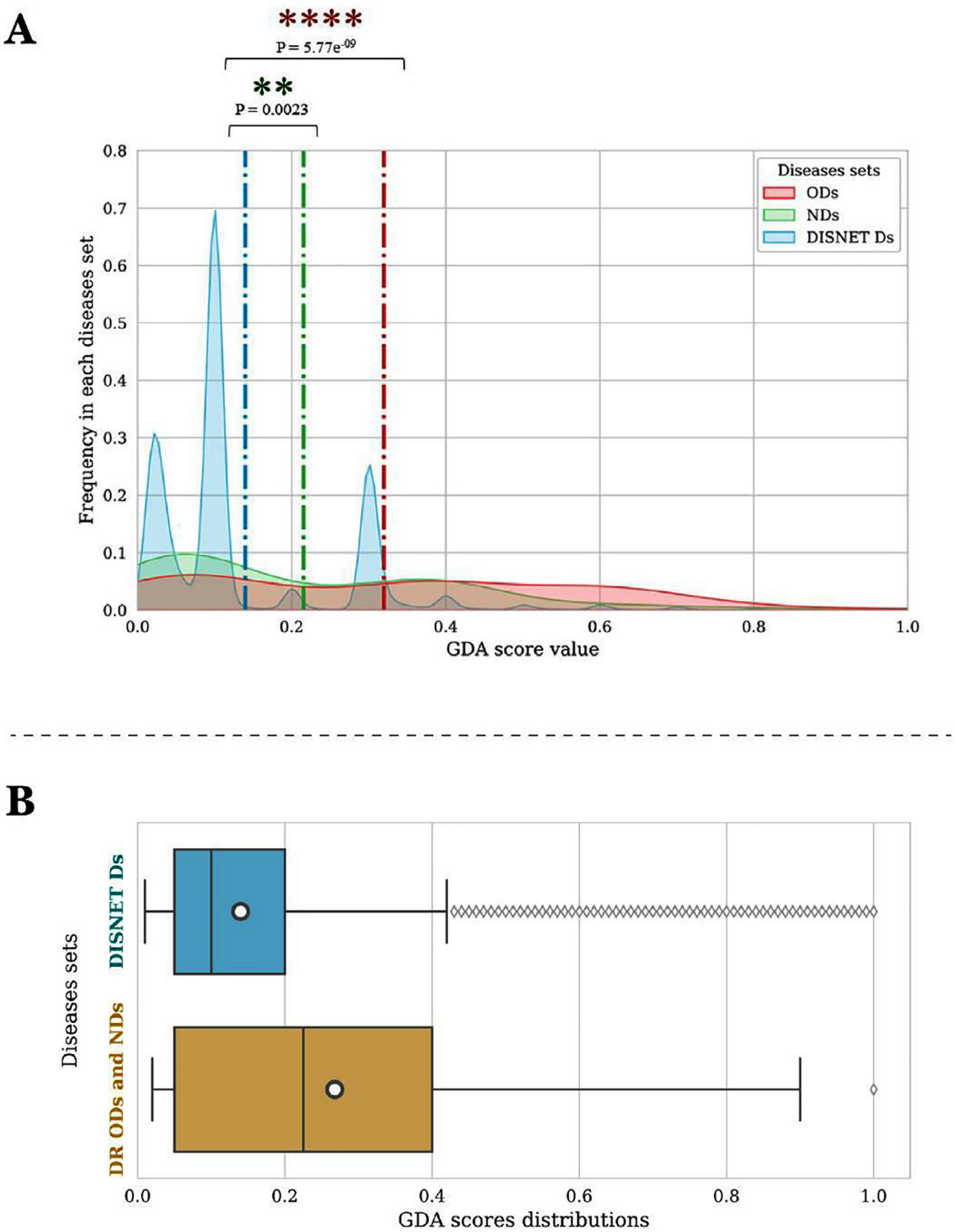

Fig. 8. GDA scores' distributions. [A] Comparing GDA scores' relative densities in each disease set. In red, GDA scores distributions of "Original Disease – Shared Target Gene" associations. In green, "New Disease – Shared Target Gene" associations. In blue, all the rest of "DISNET Disease – Gene" associations. The vertical lines indicate the mean for each group: 0.32, 0.22, and 0.14, respectively. Probability density functions have been generated by means of Kernel Density Estimation (KDE). [B] Comparing DISNET vs DR diseases GDA scores distributions. DISNET diseases (DISNET Ds) are coloured in blue and DR diseases (DR ODs and NDs) in gold. Medians are represented with vertical black lines in the boxes and means with white circles. Outlier points are depicted by diamonds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

drug, were randomly generated from DISNET data. GDA score distributions in both sets of triples have been represented in Fig. 9.

In both parts of Fig. 9, it can be observed that actual DR triples presented higher GDA scores than randomly generated ones. To prove the statistical significance of such a difference, the T-Tests depicted were performed separating ODs and NDs. Both DR ODs and NDs had relevant greater GDAs than DISNET random ODs and NDs, yet ODs seemed to present higher GDA scores than NDs.

From a symptomatology point of view, we have studied the phenotypic similarity of the repositioning diseases compared to DISNET's. Among the 79 DR cases, all the OD – ND pairs had symptoms in common except two. Distance distributions of OD – ND pairs and the rest of DISNET's "Disease – Disease" (D – D) pairs represented in Fig. 10.

The repositioning diseases appeared to be phenotypically more similar than the rest of DISNET's, as shown in Fig. 10. Diseases that share symptoms and that are, thereby, phenotypically more similar, would show distance values closer to 0. We confirmed that DR disease-pairs were phenotypically more alike with a statistical significance than the rest of disease-pairs in DISNET by a Welch T-Test (obtaining a p-value <0.05).

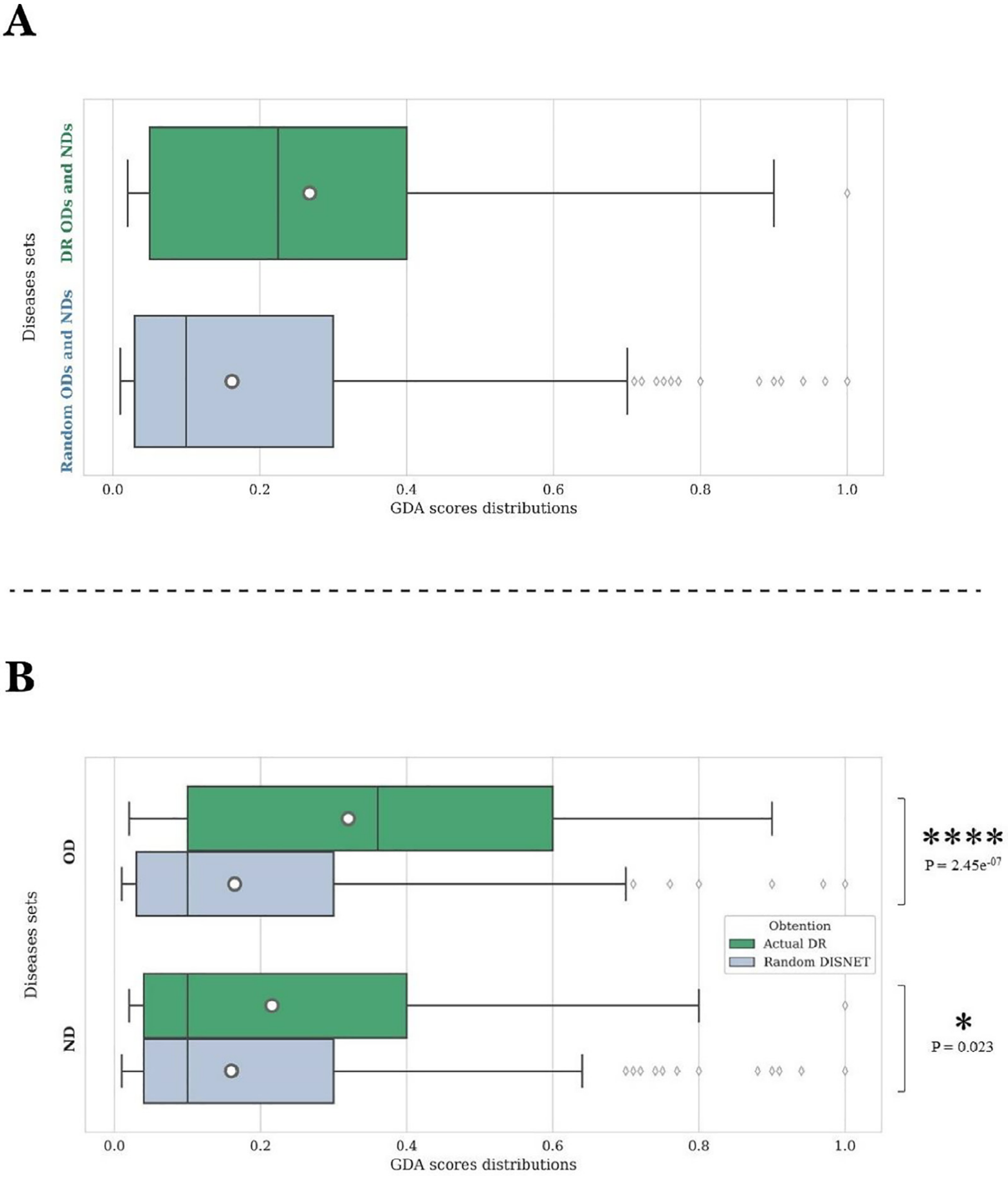

Fig. 9. GDA score distributions in actual DR and random DISNET triples. [A] Comparing actual DR ODs and NDs vs randomly-selected DISNET ODs. [B] Comparing ODs vs NDs distinguishing actual DR triples vs randomly-selected DISNET triples. ODs and NDs diseases sets are differentiated in the Y axis. In both subfigures, actual DR triples are coloured in green and random DISNET triples are coloured in grey. Medians are represented with vertical black lines in the boxes and means with white circles. Outlier points are depicted by diamonds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

To discover patterns within the DR cases, we extracted the drug and disease categories of the final set (Fig. 11). Drug categories from *MeSH-PA Therapeutic Uses* are shown in Fig. 11[A]. Out of the 40 different drugs in the final studied triples, 32 had an associated categorization. One drug may belong to several categories. The most repeated classes among repositioned drugs were "Central Nervous System Agents" (22%), "Antineoplastic Agents" (16%) and "Antirheumatic Agents" (14%). The 247 DR final triples included a total of 34 different ODs and 53 different NDs. Fig. 11[B] presents the ICD-10-CM category frequencies of OD and ND, separately. "F00-F99 (Mental and behavioural disorders)" (17%), "I00-I99 (Diseases of the circulatory system)" (16%) and "G00-G99 (Diseases of the nervous system)" (13%) are the most frequent classes for ODs, in that order. NDs are principally classified after "C00-D48 (Neoplasms)" (25%), followed by "M00-M99 (Diseases of the musculoskeletal system and connective tissue)" (16%) and "F00-F99 (Mental and behavioural disorders)" (14%).

Moreover, in Fig. 12 we show the class frequency of OD – ND pairs in DR triples. The most recurrent case has been seen to be of a drug initially used to treat a "M00-M99 (Diseases of the musculoskeletal system and connective tissue)" disease that is repositioned to a "C00-D48 (Neoplasms)". Most of the drugs that are related to this OD – ND category pattern belong to the "Antirheumatic Agents" MeSH-PA Therapeutic Uses class. Other relevant patterns would be "C00-D48 (Neoplasms)" – "A00-B99 (Certain infectious and parasitic diseases)" and "L00-L99 (Diseases of the skin and subcutaneous tissue)" – "C00-D48 (Neoplasms)".

4. Discussion

In this work, we have proposed a complete methodology to analyse the suitability of new potential DR cases by means of biomedical data integration. For such validation, we have first collected a series of previously demonstrated cases of successful DR

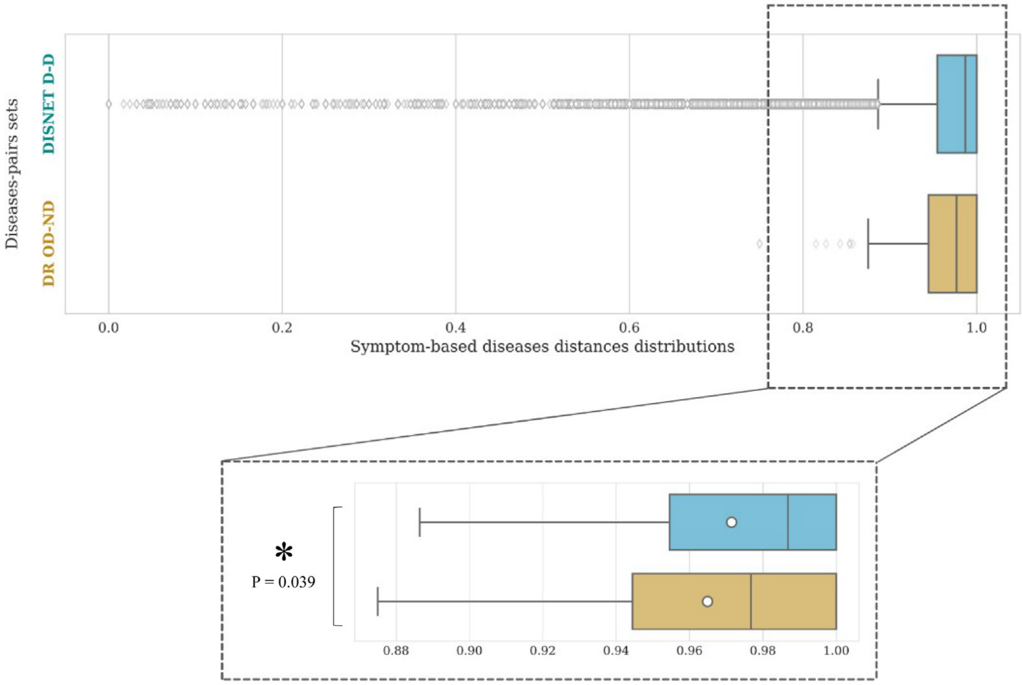

Fig. 10. Distribution of symptom-based disease distance in each disease-pair set. DISNET disease pairs are coloured in blue, whilst diseases involved in DR are coloured in gold. The entire distribution range is represented in the upper part of the figure. Diamonds depict outlier points. The lower part zooms in the region of the X axis where both distributions are enclosed, discarding outliers. Medians are represented with vertical black lines in the boxes and means with white circles. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

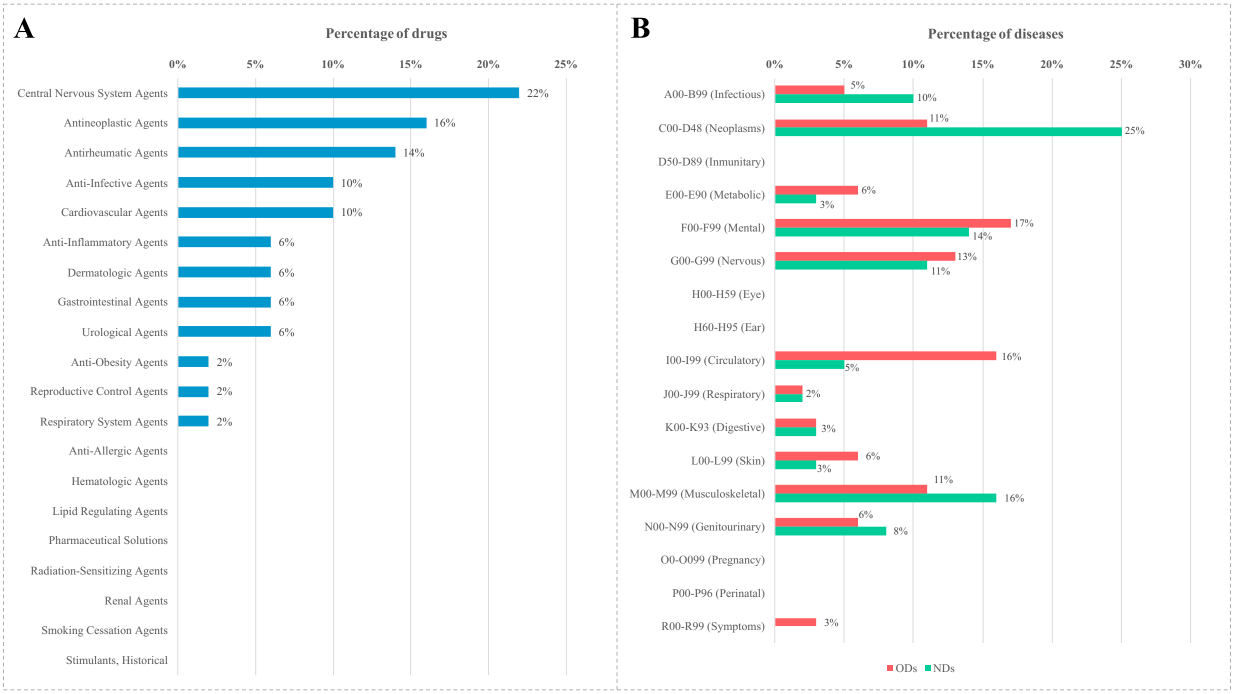

Fig. 11. Categories of the drugs and diseases in DR triples. [A] MeSH-PA Therapeutic Uses categories of the repositioned drugs. In the Y-axis, classes are represented. In the X-axis, the percentage of drugs in each group is indicated. [B] ICD-10-CM categories of ODs and NDs. In the Y-axis, disease classes are represented, while X-axis indicates the percentage of diseases in each group. In red, the frequency of ODs categories from the final DR triples. In green, NDs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

from the literature. Once normalized to DISNET codifications, we have analysed 3 aspects of our data: the genes, the symptoms, and the categories. We will start this section discussing the results obtained in each of them.

The **gene analysis** has provided us a clear perspective of the distributions of GDA scores when comparing different disease sets. On the one hand, we have compared GDA scores of DR ODs and NDs vs GDA scores of all DISNET diseases. The results (both the represen-

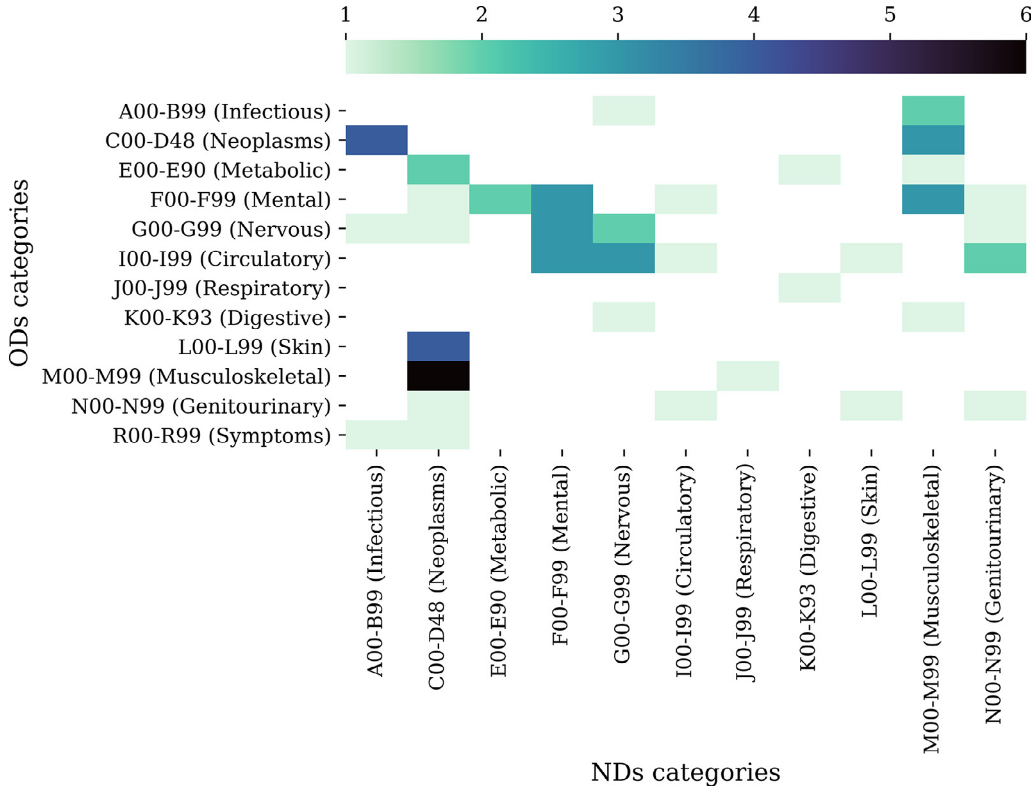

Fig. 12. ICD-10-CM category frequencies of OD – ND pairs in DR triples. Heatmap showing from which type of disease and to which type a drug is most frequently repositioned.

tations and statistical tests in Fig. 8) have suggested that diseases that participate in known DR cases tend to have greater associations with the target gene of the repurposed drug, than the rest of GDAs in DISNET. That is, when it comes to suggest new DR cases, we would rather first pose hypotheses that involve diseases whose GDA scores with the drug target gene are higher than the average.

On the other hand, we have compared actual DR ODs and NDs vs DISNET randomly selected triples. The aim here was to prove whether, going a step further and imposing the condition of the two DISNET diseases sharing the drug target gene, the differences between the two sets were still significant. The results obtained in terms of visual representations and statistical tests (Fig. 9) have suggested so, meaning that diseases implicated in actual known DR cases have again a stronger connection to the drug target gene than randomly generated triples. Even when the diseases from random triples share the drug target, real DR cases still show significantly higher GDA scores.

It is worth mentioning that ODs have shown slightly stronger linkage to drug target genes than NDs in the two parts of the gene analysis. This may be due to the nature of drug development: the indication to which the drug was originally developed for, could be more highly related to the drug target than the disease to which the drug was repositioned. Be that as it may, it can also be relevant when bringing on the new DR hypotheses. We will keep this in mind and consider that the GDA scores between ODs and the drug target gene may be greater than for NDs.

Regarding the **symptom analysis**, we have studied the phenotypical resemblance of disease pairs in two sets (DR OD – ND and DISNET D – D), with the purpose of determining whether diseases involved in a repositioning process presented higher similarity than the rest of DISNET diseases. We have employed Jaccard distance to measure symptom-sharing between diseases: such metric coefficients range from 0 to 1, giving lower values to those diseases that phenotypically relate the most. For this reason, resulting Jaccard coefficient distributions can provide insights (both via graphic visualization and statistical test in Fig. 10) of the relationships stablished between DR and not-DR diseases. We can observe that those diseases participating in DR cases have a distribution moderately shifted to the left, that is, closer to 0 and thus phenotypically more similar. The statistical significance of the difference between the two sets was confirmed, suggesting that, when generating a new DR hypothesis, those ODs – NDs with higher phenotypic similarity should be acknowledged at first.

In order to identify patters in the categorization of triples and understand which classes predominate along ODs, NDs and drugs in repositioning cases, we have carried out a **category analysis**. ODs and NDs have been classified with ICD-10-CM, and drugs, according to MeSH-PA Therapeutic Uses. We have seen that in the final DR triples used for the study there has been a trend: most repositioned drugs are classified after “Central Nervous System Agents”. However, this class is the most frequent one in DISNET, consistent with the results obtained, and consequently not as informative as other examples. For instance, “Antirheumatic Agents”, which are not as frequent in DISNET, are presented in third place in Fig. 11, suggesting that this kind of drugs may be important in repurposing processes. Furthermore, in most repeated category patterns of OD – ND pairs (Musculoskeletal – Neoplasms, Fig. 12), the majority of drugs that are repurposed from one indication class to another, are antirheumatic treatments, stressing the relevance of this category.

There are some non-represented classes both for drugs and diseases. In the case of repurposed drugs, triples did not include “Anti-Allergic”, “Hematologic”, “Lipid Regulating”, “Pharmaceutical”, “Radiation-Sensitizing”, “Renal” and “Smoking Cessation” agents. In the case of diseases, “D50-D89 (Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism)", "H00-H59 (Diseases of the eye and adnexa)", "H60-H95 (Diseases of the ear and mastoid process)", "O00-O099 (Pregnancy, childbirth and the puerperium)", "P00-P96 (Certain conditions originating in the perinatal period)" had no representation in DR triples. Nevertheless, this also provides useful information: some classes are less likely to participate in a DR case, and this should be kept in mind when putting forward new hypothesis.

The novelty of the proposed work lies in the fact that it unveils the differences at the genetic, phenotypical and categorization levels between those data that are related to actual known cases that have been successful in DR and other data not yet acknowledged to be related to DR. Presenting the evidenced differences between DR and non-DR data, would allow us in a future to evaluate and prioritize cases when generating new DR hypotheses. To generate such new DR hypotheses multiple methods could be used (some examples can be found at [10,11,48–53]). While other popular DR methods place their attention in either disease, gene or drug signatures [54], the results obtained in the present work would aid in the task of selecting those possible candidates in terms of providing a confidence statistical score (i.e., a p-value), which depicts the potential of the hypothetical case both from a genetic and phenotypic point of view. Moreover, we have displayed the patterns in drugs and diseases (original and new indications) classes in those cases known to be successful repositioning stories.

To exemplify some of the most representative DR cases, we will now discuss them and their information from DISNET. Examining the 46 cases (59 triples) that shared drug target genes, further literature research was done for 8 of them as described in Table 4. In SM section (S7.DR_validation_summary.xlsx, sheet "DR Cases Numbers"), an extensive table with all the cases has been included. The selected cases had GDA scores higher than 0.13 (DISNET's average), and they also shared symptoms. DISNET stores information about the gene that encodes the target of the drug, as well as the type of action exerted by the drug on the target. Below, we have explained the role of such gene in each DR case. We have also included the information regarding each drug provided (when possible) by the Connectivity Map (CMap) [54] web application, CLUE Repurposing tool. The categories of these top 8 DR cases are shown in Table 5.

Celecoxib is an inhibitor of cyclooxygenase 2 (COX-2). COX-2 is an enzyme responsible for inflammation and pain [55]. **Rheumatoid Arthritis** (RA) and **breast cancer** shared the gene that encodes this enzyme, among 510 other genes, and a mean of 4.83 symptoms. Both are associated with COX-2 with a 0.4 GDA score, significantly higher than DISNET's average, denoting a strong relationship. In RA its inhibition reduces inflammation processes and has analgesic activity without adverse upper gastrointestinal tract effects [56]. Furthermore, COX-2 expression is associated with angiogenesis and lymph node metastasis in human breast cancer [57]. Celecoxib may stop the growth of tumour cells by blocking the enzymes and stopping blood flow to the tumour [58]. This drug, classified as an antirheumatic agent, is repositioned from a musculoskeletal disease to a neoplasm (Table 4), being an example of the most recurrent type of DR case, as shown previously in Fig. 12.

Etanercept is a tumour necrosis factor-alpha (TFN-α) inhibitor. TFN-α competitively binds to a proinflammatory cytokine and prevents interactions with its cell surface receptors. **RA** and **asthma** have in common 15 symptoms and 295 genes. They share five genes that encode targets of Etanercept: Fc fragment of IgG receptor IIa (FCGR2A), IIIa (FCGR3A), and IIIb (FCGR3B), lymphotoxin alpha (LT-α), and tumour necrosis factor (TNF-α). Only the association with the last one exceeds DISNET's average with GDA scores of 0.7 and 0.4, respectively. In RA, excessive production of TFN-α, in the synovial fluid and the serum, causes chronic inflammation, tissue damage, and immoderate keratinocyte proliferation [59]. Otherwise, in patients with severe asthma, TFN-α high levels have been found in bronchial biopsies and induced sputum [60]. The mechanism behind these observations has not been fully enlightened but it could be caused by a direct effect of TNF-α on airway smooth muscle or by the release of the cysteinyl-leukotrienes LT_{C4} and LT_{D4} [61]. Etanercept is englobed in the most recurrent MeSH-PA category: "Antirheumatic Agents" (22%).

Finasteride is a steroid 5-alpha-reductase (SRD5A) inhibitor that blocks the conversion of testosterone to dihydrotestosterone. Normal and abnormal growth of the prostate is dependent on the presence of hormones and growth factors pointing out dihydrotestosterone. The number of symptoms and genes in common between **benign prostatic hyperplasia** (BPH) and **hair loss** is lower: 3.33 and 6. At any rate, the gene that encodes SRD5A2 is shared, being the GDA scores 0.4 (OD) and 0.34 (ND). In BHP, the inhibition of steroid 5-alpha-reductases decreases the prostate size, thereby reducing the risk of acute urinary retention [62]. In the case of hair loss, dihydrotestosterone is the cause for androgenetic alopecia so the ingest of Finasteride promotes scalp hair growth and prevents further hair loss [63]. From a category standpoint, this DR case is not that common, since it is formed by a urological agent (6%), a genitourinary disease, and a skin disease.

Infliximab is a monoclonal antibody against TNF-α. **Crohn's Disease** and **RA** share 8 symptoms and 11 genes being TNF-α one of them. Both GDA scores have a value of 0.4. Crohn's Disease is characterized by segmental transmural inflammation and granulomatous lesions of the intestinal mucosa, and this drug appeared to be useful in patients with fistulas. In RA, infliximab binds with high affinity to both soluble and transmembrane TNF being able to reduce synovial inflammation, bone resorption, and cartilage degradation [64]. Infliximab is an antirheumatic agent repositioned from a digestive disease to a musculoskeletal one, the second more frequent type of ND.

Leflunomide is an agonist of aryl hydrocarbon receptor (AhR). AhR is a ligand-activated transcription factor that controls the toxicity and activity of dioxins and related chemicals. It plays an important role in cellular growth, differentiation processes, [65] and immune diseases. **RA** and **prostate cancer** are related by 9 symptoms and 36.4 genes. They are both associated with AhR with a GDA score of 0.37. AhR activation contributes to several aspects of Rheumatoid Arthritis pathogenesis: differentiation into Th17 cells from naïve T-cells; inflammation, angiogenesis, and cartilage destruction; production of proinflammatory cytokines; and osteoclastogenesis [66]. On the other hand, AhR has been shown to act as a tumour suppressant in animal models of cancer such as prostate and liver cancers, inhibiting the proliferation of cells through different mechanisms [67]. This DR case is once again, of the most frequent types of repositioning: from a musculoskeletal disease to a neoplasm. The drug is not classified under a MeSH-PA Therapeutic Uses category.

Perindopril is an angiotensin-converting enzyme (ACE) inhibitor. **Hypertension** (HT) and **Alzheimer's Disease** (AD) have in common 13 symptoms and 211 genes, being one ACE. The GDA scores are 0.6. Its inhibition improves endothelial dysfunction and prevents cardiac remodelling, being indicated for HT [68]. Otherwise, ACE is overexpressed in the hippocampus, frontal cortex, and caudate nucleus of AD patients. The treatment with perindopril has slowed down the rate of cognitive decline in patients that also have HT [69]. "Circulatory – Nervous" are quite recurring cases in DR.

Raloxifene is a selective estrogen receptor modulator (SERM) that works as an estrogen agonist in bone, and as an antagonist in breast and uterine tissues. **Breast cancer** and **osteoporosis** are connected by 4 symptoms, and estrogen receptors 1 (ESR1) and 2

Table 4
Top 8 DR cases that share drug target genes. The GDA scores are higher than DISNET's mean (0.13) for both ODs and NDs. Cases with more than one gene in common have the one with the score above the mean underlined. Since one DR case can have many DR triples, columns "N° of symptoms in common" and "N° of genes in common" represent the mean number of symptoms and genes per DR case. Information regarding CLUE Repurposing is also provided when possible.

Drug	OD	ND	N° of symptoms in common	N° of genes in common	Target gene/s shared by the DR diseases	CLUE Repurposing
Celecoxib	**Rheumatoid Arthritis**	**Breast cancer**	4.83	510	Prostaglandin-endoperoxide synthase 2 (**PTGS2/COX-2**)	Indications for celecoxib: osteoarthritis, rheumatoid arthritis, ankylosing spondylitis, primary dysmenorrhea.
Etanercept	**Rheumatoid Arthritis**	**Asthma**	15	295	Fc fragment of IgG receptor IIa (**FCGR2A**), IIIa (**FCGR3A**) and IIIb (**FCGR3B**), lymphotoxin alpha (**LT-α**) and tumour necrosis factor (**TNF-α**)	–
Finasteride	**Benign prostatic hyperplasia**	**Hair loss**	3.33	6	Steroid 5 alpha-reductases 1 (**SRD5A1**) and 2 (**SRD5A2**)	Indication for finasteride: androgenetic alopecia
Infliximab	**Crohn's Disease**	**Alzheimer's Disease**	8	111	Tumour necrosis factor (**TNF-α**)	–
Leflunomide	**Rheumatoid Arthritis**	**Prostate cancer**	9	364	Aryl hydrocarbon receptor (**AhR**)	Indication for leflunomide: rheumatoid arthritis
Perindopril	**Hypertension**	**Alzheimer's Disease**	13	211	Angiotensin I converting enzyme (**ACE**)	Indications for perindopril: hypertension, myocardial infarction, coronary artery disease (CAD)
Raloxifene	**Breast cancer**	**Osteoporosis**	4	147	Estrogen receptors 1 (**ESR1**) and 2 (**ESR2**)	Indications for raloxifene: osteoporosis, breast cancer
Requip	**Parkinson Disease**	**Restless leg syndrome**	16	11	Dopamine receptor D3 (**DRD3**)	Indications for requip: Parkinson's Disease, restless leg syndrome

Table 5
Categories of the top 8 DR cases. Drug MeSH-PA therapeutic uses classification and disease ICD-10-CM categories.

Drug (MeSH-PA Therapeutic Uses category)	OD (ICD-10-CM category)	ND (ICD-10-CM category)
Celecoxib (Antirheumatic Agents)	**Rheumatoid Arthritis** (M00-M99: Musculoskeletal)	**Breast cancer** (C00-D48: Neoplasms)
Etanercept (Antirheumatic Agents)	**Rheumatoid Arthritis** (M00-M99: Musculoskeletal)	**Asthma** (J00-J99: Respiratory)
Finasteride (Urological Agents)	**Benign prostatic hyperplasia** (N00-N99: Genitourinary)	**Hair loss** (L00-L99: Skin)
Infliximab (Antirheumatic Agents)	**Crohn's Disease** (K00-K93: Digestive)	**Rheumatoid Arthritis** (M00-M99: Musculoskeletal)
Leflunomide (–)	**Rheumatoid Arthritis** (M00-M99: Musculoskeletal)	**Prostate cancer** (C00-D48: Neoplasms)
Perindopril (Cardiovascular Agents)	**Hypertension** (I00-I99: Circulatory)	**Alzheimer's Disease** (G00-G99: Nervous)
Raloxifene (–)	**Breast cancer** (C00-D48: Neoplasms)	**Osteoporosis** (M00-M99: Musculoskeletal)
Requip (Central Nervous System Agents)	**Parkinson's Disease** (G00-G99: Nervous)	**Restless leg syndrome** (G00-G99: Nervous)

(ESR2) genes, among 145 others. Raloxifene's ingest reduces the risk of fractures, improves the lipid profile, protects the breast, and provides uterine safety. It simulates the effects of estrogens on bone, increasing bone density [70]. Otherwise, it inhibits estrogen-dependent proliferation of human breast cancer cells [71]. Raloxifene is not classified under MeSH-PA Therapeutic Uses. This OD and ND combination has representation in DR.

Requip, also known as ropinirole, is a non-ergoline dopamine agonist. Dopamine receptors are G protein-coupled receptors that participate in motor activity, regulation, and several neurological disorders [72]. Ropinirole has the highest affinity for D3 receptors, which are concentrated in the limbic areas of the brain and may be responsible for some neuropsychiatric effects [73]. **Parkinson's Disease** (PD) and **restless leg syndrome** (RLS) share the highest number of symptoms: 16. Among the genes in common, they share the dopamine receptor D3 with a GDA score of 0.2. This drug reliefs PD and RLS symptoms by stimulating dopamine receptors, even though its exact mechanism of action is still unknown. Central Nervous System Agents are the most frequent drugs in DR (22%). Repositioning between nervous diseases has been repeated too.

5. Conclusions

The current manuscript has presented a new methodological pipeline for the potential generation of new DR hypotheses by means of integrating biomedical knowledge. The main conclusion of this work is that such type of data (in particular, we have used DISNET integrated data) could be considered and used to suggest novel potential repurposing cases. We mainly state this conclusion because: i) actual well-known DR cases show significant differences with other non-DR data regarding their gene and symptom similarities; and ii) DISNET provides known DR-related information important to the repurposing process (e.g., the drug target gene related to DR-involved diseases).

Other conclusions derive from the previous. The analysis of genes, symptoms and categories can provide hints on which DR cases should be prioritized or given more attention. The gene analysis has confirmed that the diseases participating in DR processes present higher associations with drug target genes than the rest of DISNET GDAs. The symptom analysis has demonstrated that diseases involved in a DR case are phenotypically more similar (in terms of shared symptoms) than the rest of DISNET phenotypical disease relationships. And the category analysis has suggested that the classes of repositioned drugs and DR diseases frequently follow some specific patterns. These three analyses will allow us to build better hypothetical DR cases in the future through DISNET knowledge. The differences shown between all DISNET data and actual DR data are key to discern those novel hypotheses.

Nonetheless, we have identified some limitations in the present analysis. The main one would be the difficulties found when integrating and interoperating with heterogeneous biomedical data coming from many different sources. The entities in our data differ in identification codes, hindering the discovery of plausible DR hypothesis. Although DISNET already works as a good integrating platform and has been successful for the proposed methodology, we are planning on semantizating DISNET's biomedical information in order to improve data integration. We have also detected that not having a disease – symptom association score (analogous to GDA scores) might signify a drawback. If we could weight such relationships, finer results could be drawn. Moreover, we have to state some potential drawbacks that should be considered when employing data-driven techniques for DR. Certain drugs might act through a different target that is not observed when treating the original indication. In these cases, some candidates could be overlooked and not prioritized, as the new and original indication would not share targets. Or even the definition of the target could be incorrect in some cases. More insights could be obtained by using tools as Connectivity Map analyses. In addition, an experimental *in vitro* or *in vivo* validation would be needed afterwards, in order to further test the prioritized candidates for the new indications.

Other future lines of this study would include expanding our set of known DR triples by covering other sources. These could be both scientific publications data and/or direct DR databases, as long as they satisfy the aforementioned requirements. Once proposed this approach for evaluating DR cases, the next step that we would like to implement, would be to directly use DISNET for suggesting new indications for already-existing drugs. That is, targeting drug repurposing.

CRediT authorship contribution statement

Lucía Prieto Santamaría: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Visualization. **Esther Ugarte Carro:** Methodology, Software, Writing - original draft, Visualization. **Marina Díaz Uzquiano:** Writing - original draft, Writing - review & editing, Visualization. **Ernestina Menasalvas Ruiz:** Writing - review & editing. **Yuliana Pérez Gallardo:** Writing - review & editing, Supervision, Funding acquisition. **Alejandro Rodríguez-González:** Conceptualization, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work is a result of the project “DISNET (Creation and analysis of disease networks for drug repurposing from heterogeneous data sources applied to rare diseases)”, that is being developed under grant “RTI2018-094576-A-I00” from the Spanish Ministerio de Ciencia, Innovación y Universidades. LPS's work is supported by “Programa de fomento de la investigación y la innovación (Doctorados Industriales)” from Comunidad de Madrid (grant IND2019/TIC-17159).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.08.003.

References

[1] Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004, 3(8):673-3, doi:10.1038/nrd1468.

[2] Low ZY, Farouk IA, Lal SK. Drug repositioning: new approaches and future prospects for life-debilitating diseases and the COVID-19 pandemic outbreak, Viruses, 2020, 12(9):1058, doi:10.3390/v12091058.

[3] Xue H, Li J, Xie H, Wang Y. Review of drug repositioning approaches and resources. Int J Biol Sci 2018;14(10):1232–44. https://doi.org/10.7150/ijbs.24612.

[4] Naylor S, Schonfeld JM. Therapeutic drug repurposing, repositioning and rescue - Part I: Overview. Drug Discov World. 2014;16:49-62.

[5] Turanli B, Grøtli M, Boren J, Nielsen J, Uhlen M, Arga KY, et al. Drug repositioning for effective prostate cancer treatment. Front Physiol 2018;9. https://doi.org/10.3389/fphys.2018.00500.

[6] Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2019;18(1):41–58. https://doi.org/10.1038/nrd.2018.168.

[7] Shim JS, Liu JO. Recent advances in drug repositioning for the discovery of new anticancer drugs. Int J Biol Sci 2014;10(7):654–63. https://doi.org/10.7150/ijbs.9224.

[8] Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. J Cheminformat 2020;12(1):46. https://doi.org/10.1186/s13321-020-00450-7.

[9] Fiscon G, Conte F, Farina L, Paci P, Marz M. SAveRUNNER: a network-based algorithm for drug repurposing and its application to COVID-19. PLOS Comput Biol. 2021;17(2):e1008686. https://doi.org/10.1371/journal.pcbi.1008686.

[10] Fiscon G, Conte F, Amadio S, Volonté C, Paci P. Drug repurposing: a network-based approach to amyotrophic lateral sclerosis, Neurotherapeutics, Published online May 13, 2021. doi:10.1007/s13311-021-01064-z.

[11] Morselli Gysi D, do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proc Natl Acad Sci USA 2021;118(19). https://doi.org/10.1073/pnas.2025581118.

[12] Liu Z, Fang H, Reagan K, Xu X, Mendrick DL, Slikker W, et al. In silico drug repositioning – what we need to know. Drug Discov Today 2013;18(3-4):110–5. https://doi.org/10.1016/j.drudis.2012.08.005.

[13] Frail DE, Brady M, Escott KJ, et al. Pioneering government-sponsored drug repositioning collaborations: progress and learning. Nat Rev Drug Discov. 2015;14(12):833-841. doi:10.1038/nrd4707.

[14] Turanli B, Altay O, Borén J, Turkez H, Nielsen J, Uhlen M, et al. Systems biology based drug repositioning for development of cancer therapy. Semin Cancer Biol 2021;68:47–58. https://doi.org/10.1016/j.semcancer.2019.09.020.

[15] Lima WG, Brito JCM, Overhage J, Nizer WS da C. The potential of drug repositioning as a short-term strategy for the control and treatment of COVID-19 (SARS-CoV-2): a systematic review. Arch Virol. 2020;165(8):1729-1737. doi:10.1007/s00705-020-04693-5.

[16] Corbett A, Pickett J, Burns A, Corcoran J, Dunnett SB, Edison P, et al. Drug repositioning for Alzheimer's disease. Nat Rev Drug Discov 2012;11(11):833–46. https://doi.org/10.1038/nrd3869.

[17] Masuda T, Tsuruda Y, Matsumoto Y, Uchida H, Nakayama KI, Mimori K. Drug repositioning in cancer: the current situation in Japan. Cancer Sci 2020;111(4):1039–46. https://doi.org/10.1111/cas.14318.

[18] Li YY, Jones SJ. Drug repositioning for personalized medicine. Genome Med 2012;4(3):27. https://doi.org/10.1186/gm326.

[19] Lotfi Shahreza M, Ghadiri N, Mousavi SR, Varshosaz J, Green JR. A review of network-based approaches to drug repositioning. Brief Bioinform. 2018;19(5):878–92. https://doi.org/10.1093/bib/bbx017.

[20] Brown AS, Patel CJ. A review of validation strategies for computational drug repositioning. Brief Bioinform 2018;19(1):174–7. https://doi.org/10.1093/bib/bbw110.

[21] Luo H, Li M, Yang M, Wu F-X, Li Y, Wang J. Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform, 2021;22(2):1604-19, doi:10.1093/bib/bbz176.

[22] Novac N. Challenges and opportunities of drug repositioning. Trends Pharmacol Sci 2013;34(5):267–72. https://doi.org/10.1016/j.tips.2013.03.004.

[23] Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 2013;93(4):335–41. https://doi.org/10.1038/clpt.2013.1.

[24] Rameshrad M, Ghafoori M, Mohammadpour AH, Nayeri MJD, Hosseinzadeh H. A comprehensive review on drug repositioning against coronavirus disease 2019 (COVID19). Naunyn Schmiedebergs Arch Pharmacol 2020;393(7):1137–52. https://doi.org/10.1007/s00210-020-01901-6.

[25] Ballard C, Aarsland D, Cummings J, et al. Drug repositioning and repurposing for Alzheimer disease. Nat Rev Neurol 2020;16(12):661–73. https://doi.org/10.1038/s41582-020-0397-4.

[26] Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. Brief Bioinform 2011;12(4):303–11. https://doi.org/10.1093/bib/bbr013.

[27] Wilkinson GF, Pritchard K. In vitro screening for drug repositioning. J Biomol Screen 2015;20(2):167–79. https://doi.org/10.1177/1087057114563024.

[28] Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Brief Bioinform 2016;17(1):2–12. https://doi.org/10.1093/bib/bbv020.

[29] Adasme MF, Parisi D, Sveshnikova A, Schroeder M. Structure-based drug repositioning: potential and limits. Semin Cancer Biol 2021;68:192–8. https://doi.org/10.1016/j.semcancer.2020.01.010.

[30] Ma D-L, Chan DS-H, Leung C-H. Drug repositioning by structure-based virtual screening. Chem Soc Rev. 2013;42(5):2130-2141. doi:10.1039/C2CS35357A.

[31] Oprea TI, Overington JP. Computational and practical aspects of drug repositioning. ASSAY Drug Dev Technol 2015;13(6):299–306. https://doi.org/10.1089/adt.2015.29011.tiodrrr.

[32] Nzila A, Ma Z, Chibale K. Drug repositioning in the treatment of malaria and TB. Future Med Chem 2011;3(11):1413–26. https://doi.org/10.4155/fmc.11.95.

[33] Yella JK, Yaddanapudi S, Wang Y, Jegga AG. Changing trends in computational drug repositioning. Pharmaceuticals 2018;11(2):57. https://doi.org/10.3390/ph11020057.

[34] Kharkar PS, Warrier S, Gaud RS. Reverse docking: a powerful tool for drug repositioning and drug rescue. Future Med Chem 2014;6(3):333–42. https://doi.org/10.4155/fmc.13.207.

[35] Chen H, Zhang H, Zhang Z, Cao Y, Tang W. Network-based inference methods for drug repositioning. Comput Math Methods Med 2015;2015:1–7. https://doi.org/10.1155/2015/130620.

[36] Yang L, Agarwal P, Csermely P. Systematic drug repositioning based on clinical side-effects. PLoS ONE 2011;6(12):e28025. https://doi.org/10.1371/journal.pone.0028025.

[37] Lagunes-García G, Rodríguez-González A, Prieto-Santamaría L, Valle EPG del, Zanin M, Menasalvas-Ruiz E. DISNET: a framework for extracting phenotypic disease information from public sources. PeerJ. 2020;8:e8580. doi:10.7717/peerj.8580.

[38] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. Published online November 4, 2019:gkz1021. doi:10.1093/nar/gkz1021.

[39] UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49(D1):D480-D489. doi:10.1093/nar/gkaa1100.

[40] ChEMBL: towards direct deposition of bioassay data | Nucleic Acids Research | Oxford Academic. Accessed May 10, 2021. https://academic.oup.com/nar/article/47/D1/D930/5162468.

[41] Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074-D1082. doi:10.1093/nar/gkx1037.

[42] Comparative Toxicogenomics Database (CTD): update 2021 | Nucleic Acids Research | Oxford Academic. Accessed May 10, 2021. https://academic.oup.com/nar/article/49/D1/D1138/5929242.

[43] UMLS REST API Home Page. Accessed April 5, 2021. https://documentation.uts.nlm.nih.gov/rest/home.html.

[44] Lu Z, Yuan K-H. Encyclopedia of research design, Welch's t test. N. J. Salkind, editor. Thousand Oaks, CA; 2010. 1620–1623.

[45] RxClass Overview. Accessed March 23, 2021. https://rxnav.nlm.nih.gov/RxClassIntro.html.

[46] Classification of Diseases (ICD). Accessed April 6, 2021. https://www.who.int/standards/classifications/classification-of-diseases.

[47] Sharp ME. Toward a comprehensive drug ontology: extraction of drug-indication relations from diverse information sources. J Biomed Semant 2017;8(1):2. https://doi.org/10.1186/s13326-016-0110-0.

[48] Zong N, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. Bioinformatics. 2017;33(15):2337-2344. doi:10.1093/bioinformatics/btx160.

[49] Martínez V, Navarro C, Cano C, Fajardo W, Blanco A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. Artif Intell Med 2015;63(1):41–9. https://doi.org/10.1016/j.artmed.2014.11.003.

[50] Emon MA, Domingo-Fernández D, Hoyt CT, Hofmann-Apitius M. PS4DR: a multimodal workflow for identification and prioritization of drugs based on pathway signatures. BMC Bioinf 2020;21(1):231. https://doi.org/10.1186/s12859-020-03568-5.

[51] Malas TB, Vlietstra WJ, Kudrin R, et al. Drug prioritization using the semantic properties of a knowledge graph. Sci Rep. 2019;9(1):6281. https://doi.org/10.1038/s41598-019-42806-6.

[52] Qabaja A, Alshalalfa M, Alanazi E, Alhajj R. Prediction of novel drug indications using network driven biological data prioritization and integration. J Cheminformat 2014;6(1):1. https://doi.org/10.1186/1758-2946-6-1.

[53] Hu G, Agarwal P, Jordan IK. Human disease-drug network based on genomic expression profiles. PLoS ONE 2009;4(8):e6536. https://doi.org/10.1371/journal.pone.0006536.

[54] Lamb J, Crawford ED, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease, Science. 2006;313(5795):1929-1935. doi:10.1126/science.1132939.

[55] Lipfert P, Seitz R, Arndt JO. Studies of local anesthetic action on natural spike activity in the aortic nerve of cats. Anesthesiology 1987;66(2):210–3. https://doi.org/10.1097/00000542-198702000-00016.

[56] Simon LS, Weaver AL, Graham DY, et al. Anti-inflammatory and upper gastrointestinal effects of celecoxib in rheumatoid arthritis A randomized controlled trial. JAMA. 1999;282(20):1921-1928. doi:10.1001/jama.282.20.1921.

[57] Jana D, Sarkar DK, Ganguly S, et al. Role of cyclooxygenase 2 (COX-2) in prognosis of breast cancer. Indian J Surg Oncol 2014;5(1):59–65. https://doi.org/10.1007/s13193-014-0290-y; Jana D.

[58] Arun B, Goss P. The role of COX-2 inhibition in breast cancer treatment and prevention. Semin Oncol. 2004;31:22-29. doi:10.1053/j.seminoncol.2004.03.042.

[59] Goffe B, Cather JC. Etanercept: An overview. J Am Acad Dermatol. 2003;49(2, Supplement):105-111. doi:10.1016/mjd.2003.554.

[60] Morjaria JB, Chauhan AJ, Babu KS, Polosa R, Davies DE, Holgate ST. The role of a soluble TNFα receptor fusion protein (etanercept) in corticosteroid refractory asthma: a double blind, randomised, placebo controlled trial. Thorax 2008;63(7):584–91. https://doi.org/10.1136/thx.2007.086314.

[61] Berry M, Brightling C, Pavord I, Wardlaw A. TNF-α in asthma. Curr Opin Pharmacol 2007;7(3):279–82. https://doi.org/10.1016/j.coph.2007.03.001.

[62] Tacklind J, Fink HA, Macdonald R, Rutks I, Wilt TJ. Finasteride for benign prostatic hyperplasia. Cochrane Database Syst Rev 2010. https://doi.org/10.1002/14651858.CD006015.pub3.

[63] McClellan KJ, Markham A. Finasteride. Drugs 1999;57(1):111–26. https://doi.org/10.2165/00003495-199957010-00014.

[64] Perdriger A. Infliximab in the treatment of rheumatoid arthritis. Biol Targets Ther. 2009;3:183–91.

[65] O'Donnell EF, Saili KS, Koch DC, et al. The anti-inflammatory drug leflunomide is an agonist of the aryl hydrocarbon receptor. PLoS ONE 2010;5(10):. https://doi.org/10.1371/journal.pone.0013128e13128.

[66] Nguyen NT, Nakahama T, Nguyen CH, et al. Aryl hydrocarbon receptor antagonism and its role in rheumatoid arthritis. J Exp Pharmacol. 2015;7:29-35. doi:10.2147/JEP.S63549; Nguyen NT, Nakahama T, Nguyen CH, et al. Aryl hydrocarbon receptor antagonism and its role in rheumatoid arthritis.

[67] Zhang C, Chu M. Leflunomide: A promising drug with good antitumor potential. Biochem Biophys Res Commun. 2018;496(2):726–30. https://doi.org/10.1016/j.bbrc.2018.01.107.

[68] Yamada K, Uchida S, Takahashi S, Takayama M, Nagata Y, Suzuki N, et al. Effect of a centrally active angiotensin-converting enzyme inhibitor, perindopril, on cognitive performance in a mouse model of Alzheimer's disease. Brain Res 2010;1352:176–86. https://doi.org/10.1016/j.brainres.2010.07.006.

[69] Ohrui T, Tomita N, Sato-Nakagawa T, Matsui T, Maruyama M, Niwa K, et al. Effects of brain-penetrating ACE inhibitors on Alzheimer disease progression. Neurology 2004;63(7):1324–5. https://doi.org/10.1212/01.WNL.0000140705.23869.E9.

[70] Carretero M. El raloxifeno en el tratamiento de la osteoporosis posmenopáusica. Offarm. 2003;22(1):134-136.

[71] Balfour JA, Goa KL. Raloxifene: Drugs Aging. 1998;12(4):335-341. doi:10.2165/00002512-199812040-00006.

[72] Del Campo N, Chamberlain SR, Sahakian BJ, Robbins TW. The roles of dopamine and noradrenaline in the pathophysiology and treatment of attention-deficit/hyperactivity disorder. Biol Psychiatry. 2011;69(12):e145-157. doi:10.1016/j.biopsych.2011.02.036.

[73] Shill HA, Stacy M. Update on ropinirole in the treatment of Parkinson's disease. Neuropsychiatr Dis Treat. 2009;5:33-36.