Explainable Drug Repurposing Approach From Biased Random Walks

Filippo Castiglione, Christine Nardini, Elia Onofri, Marco Pedicini, and Paolo Tieri

Abstract—Drug repurposing is a highly active research area, aiming at finding novel uses for drugs that have been previously developed for other therapeutic purposes. Despite the flourishing of methodologies, success is still partial, and different approaches offer, each, peculiar advantages. In this composite landscape, we present a novel methodology focusing on an efficient mathematical procedure based on gene *similarity scores* and *biased random walks* which rely on robust drug-gene-disease association data sets. The recommendation mechanism is further unveiled by means of the *Markov chain* underlying the random walk process, hence providing *explainability* about how findings are suggested. Performances evaluation and the analysis of a case study on *rheumatoid arthritis* show that our approach is accurate in providing useful recommendations and is computationally efficient, compared to the state of the art of drug repurposing approaches.

Index Terms—Drug repurposing, explainable artificial intelligence, network medicine, Markov chain, biased random walk

1 Introduction

Drug repurposing (DR) is emerging as an essential and potentially valuable undertaking to rapidly exploit existing and tested drugs for new uses, such as emerging and neglected diseases, as well as an alternative and convenient choice as opposed to the *de novo* drugs development. However, in the exemplary case of the ongoing COVID-19 pandemic, despite the high number of DR attempts (a non-exhaustive PubMed search performed on January 2022 reported more than 900 results using the keywords "covid" and "drug repurposing"), the effectiveness of DR still appeared to be low: out of over 400 drugs tested, just a few, and precisely four of them, were definitively shown to be effective, i.e., graded "A" (i.e., established effectiveness; endorsement by professional societies) [1], [2], [3]. The reported success rate of $\sim$1% is certainly not satisfactory [4] and calls for better use of the increasingly robust data and for the development of more efficient methods for DR algorithms and processes capable of making the most out of previous knowledge.

DR approaches fuse well also with the concept of Synergistic Drug Combinations, where the aim is to find two or more active pharmaceutical ingredients that co-op well to target multiple conditions (e.g., [5] and [6]).

In what follows, we explain some prominent examples of current DR processes representing the starting point of our study, and we highlight some salient related limitations.

1.1 Brief Exploration of Previous Relevant Works

The pioneering work of Keiser *et al.* [7] introduced the possibility to infer by computational means novel drugs for neglected diseases, and suggested general computational DR systems. Since then, various sorts of computational approaches exploiting different databases and different similarity criteria have been designed to tackle the problem. In what follows we briefly report, with no claim of completeness, some of the studies that are most relevant for the work here presented (see e.g., [8] for a more in-depth review).

Luo *et al.* [9] exploited the concept of similarity of drugs (chemical-based) and diseases (MeSH-based [10]) to build two networks, then they linked them together exploiting data from the databases DrugBank [11] and Online Mendelian Inheritance in Man (OMIM) [12], to generate 1933 drug-disease associations among 593 drugs and 313 diseases. Finally, a random walk approach was implemented simultaneously on the two similarity networks to rank drug-disease associations and propose predictions.

Nam *et al.* [13] proposed a method involving three steps: (i) drug network reconstruction using drug-target protein associations, (ii) network reinforcement, i.e., a machine learning approach providing *information augmentation* by using drug-drug interaction knowledge, and (iii) a recommendation step via the generation of a score computed by a graph-based semi-supervised learning procedure. The authors identified and recommended 11 novel drugs for the case study of vascular dementia through their approach.

Ozsoy *et al.* [14] carried out DR as a recommendation process in three steps: similarity evaluation, neighbour and disease selection, actually implementing a collaborative filtering with the possibility to integrate multiple data sources and multiple features. The method produced recommendations

- *Filippo Castiglione, Christine Nardini, and Paolo Tieri are with the Institute for Applied Computing of the National Research Council (CNR–IAC) of Italy, 00185 Rome, RM, Italy. E-mail: f.castiglione@iac.cnr.it, {christine.nardini, paolo.tieri}@cnr.it.*
- *Elia Onofri and Marco Pedicini are with the Department of Mathematics and Physics, Roma Tre University, 00146 Roma, RM, Italy. E-mail: {elia.onofri, marco.pedicini}@uniroma3.it.*

Manuscript received 1 March 2022; revised 1 July 2022; accepted 9 July 2022. Date of publication 15 July 2022; date of current version 3 April 2023. (Corresponding author: Elia Onofri.)
This article has supplementary downloadable material available at https://doi.org/10.1109/TCBB.2022.3191392, provided by the authors.
Digital Object Identifier no. 10.1109/TCBB.2022.3191392

based on the similarity and overlap between symptoms of the diseases and the effectiveness of the drugs, hence showing better performances compared to other methods available in the literature.

1.2 Limitations and Issues With Available Data

DR approaches exploit previous knowledge, data cross-linking and associations via database interoperability to infer *de novo* predictions and testable hypotheses. Such approaches leverage on many different large datasets developed during the last decades, when the global knowledge on drugs and diseases properties has considerably increased [15]. However, a fast growth like this often generates an inconsistency in nomenclatures, resulting in a persistent difficulty in fusing data coming from different sources [16].

Literature reports many attempts in setting up actual standards (e.g., the latest World Health Organization's ICD-11 [17]), however, different usages imply different requirements, not always addressed by all evolving standards. As such, for our purposes, we refer to the common practice in the scientific community operating on DR, that currently widely employs only a few *de facto* leading open-access reference sources and standards (i.e., the DrugBank knowledge base for drug-target associations [11] and the DisGeNet discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases [18]). We refer to more pragmatic reviews like [8] for an in-depth analysis of the available datasets.

1.3 Overview of the Proposed Approach

Building on previous works and with the aim of enhancing reliability and capabilities of preceding attempts, we here propose a Markov process-based similarity approach that exploits available data in the form of a knowledge graph [19] such as experimental drug-gene interactions, disease-gene associations, and drug-disease pharmacological indications.

Our approach consists of five distinctive features: (i) a careful data selection and nomenclature mapping, (ii) a database bipartite graph design, (iii) a BLAS-based data structure [20], (iv) an Ergodic Markov Process representation of the DR system, and (v) an explainable output.

A careful selection of the terms grants maximum consistency and therefore minimal loss of information, in particular when data sets are joined together. This also considerably reduces the effort spent manually curating the data sets. For this reason, we chose two *de facto* standards: `cas-number` (Chemical Abstracts Service Reference Number) for drugs [21] as provided by DrugBank, and the `UMLS` (Unified Medical Language System) identifier for diseases [22]. We converted all the involved entities from the various data sets through different dictionaries provided by MalaCards [23], OMIM [12], DrugBank [11], and many others and we assigned each drug/disease a unique integer identifier (used for evaluation purposes). This choice enables us to employ two widely used, highly curated, and up-to-date data sets – DrugBank [11] and DisGeNET [18] – granting nomenclature standardisation and interoperability.

The savvy mathematical formulation of the data sets as bipartite graphs allows the natural construction of useful entropy-inspired similarity measures that lays the foundations for our knowledge graph (see e.g., [24] where a similar approach is used to create a collaborative filtering for miRNA-disease associations).

We embed our graph structure in the BLAS environment, where structures are stored as sparse (Boolean) matrices. This solution yields fast and reliable performances based on matrix-matrix operation, like the ability to represent connections between entities as a sequence of multiplications.

We used the resulting knowledge graph with normalised connections to generate a *Markov-process-based* DR system. The underlying mathematical structure enforces the usage of the notion of *ergodicity*, therefore providing a mathematical proof of the stability of the recommendations with respect to small changes in the input data, see Section 3.2.

Finally, recommendations provided on the basis of a biased random walk make them self-explainable. Explainability in Artificial Intelligence methods (see [25]) like recommendation systems is particularly useful since it allows, e.g., to interpret the results and provide practical hints in testing, both features strongly demanded in recent trends.

We found that all the above-mentioned characteristics are crucial to providing effective, fast, usable, and reliable results for DR.

The remainder of this paper is organised as follows. In Section 2 we describe the databases and datasets used, their cross-mappings, the related necessary integration, and their pre-processing. Section 3 introduces the approach used to exploit such data (Section 3.1), how to build the recommender system and how it behaves (Section 3.2). In Section 4 we discuss parameters tuning (Section 4.1) and compare the methodology's performances with four existing similar methods (Section 4.2). We provide further results in Section 5, by analysing a specific DR case study, that of rheumatoid arthritis (RA), a chronic autoimmune disease with complex aetiology and no cure to date. Finally, Section 6 provides a summary of the contribution and possible further developments.

2 Materials

The present DR approach makes use of several datasets and data sources for the reconstruction of a knowledge graph on which the recommendation system is built.

As presented in Section 1.2, we extracted drugs and diseases information from the two *de facto* standards DrugBank (*DB*) [11] and DisGeNET (*DGN*) [18] respectively. This helped us in finding many (five) different datasets to obtain drug–diseases association, namely: MalaCards (*MC*) [23], RepoDB (*RDB*) [26], iDrug (*ID*) [27], as well as data from Li and Lu article (referred to as *LL*) [28], and data from the Similarity-based LArge-margin learning of Multiple Sources DR framework (*SLAMS*) [29]. Since the latter four (described in Section 2.2) were employed by other DR published methodologies, we found them suitable to be used as benchmark datasets (see Section 4.2). On the contrary, we used the first three (described in Section 2.1) to build and tune our methodology (see Section 3).

Table 1 summarises the statistics of the seven data sets reporting the number of vertices and edges of the underlying graph while in the following we briefly account them. Their usage is summarised in Fig. 1.

TABLE 1
Datasets Structure

Name	Ref.	#Diseases	#Drugs	#Genes	#Connections
DB	[11]	–	13563	4118	20279
DGN	[18]	30293	–	26137	3261324
MC	[23]	31642	12240	–	544857
RDB	[26]	1229	1519	–	10563
ID	[27]	3966	1314	–	111481
LL	[28]	719	799	–	3250
SLAMS	[29]	406	305	–	3871

2.1 Main Data Sets and Data Sources

DB [11] DrugBank is a pharmaceutical knowledge base that enables major advances across the data-driven medicine industry. It is provided as a drug-oriented XML, where each drug is labelled by a DrugBank unique identifier `DB` (`DBxxxxx`). We used it to extract drug-gene relations (target gene polypeptides), drug names (and `cas-number`), and their market status (approved, illicit, experimental, ...) to provide further filtering on the final recommendations. We were able to extract 7,262 `cas-number` drugs (out of 13,563) connected to 4,118 Genes through 19,792 connections (from 1 to 305 connections per drug).

DGN [18] DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated with human diseases. It is provided as an SQLite DB built upon many domain, typological and associative tables. We used it to extract the relations between diseases and target genes, along with diseases' names and `UMLS` identifiers. We consider 30,170 Diseases (out of 30,293) connected to 21,671 (out of 26,137) genes through 1,135,045 connections (from 1 to 10,161 connections per disease).

MC [23] MalaCards is an integrated database of human maladies and their annotations, modelled on the architecture and richness of the popular GeneCards database of human genes. It provides the relations between drugs and diseases, already expressed as relations between `cas-number` for drugs and `UMLS` identifier for diseases. We also used it as a source of dictionaries for converting the test data sets. We filtered the entries, obtaining a total of 2088 `cas-number` (out of 12,240) and 7,009 `UMLS` identifiers (out of 31,642) connected via 495,060 links.

The diagram in Fig. 2 shows the links between the records of the data bases. We represented relations as highly sparse Boolean matrices and records as unit sparse vectors of the corresponding size.

As described, each data set is mainly used to extract the relations between two different kinds of entities; hence it can be interpreted as a bipartite graph $G_\circ = (V_\circ, E_\circ)$ where nodes $V_\circ$ represents entities (drugs, diseases, or genes) and edges $E_\circ$ enforces connections among them. For ease of notation, we denote each graph with the acronym of the corresponding data set, meaning we build the recommendation system through three of them:

1) A Drug-Gene graph $G_{DB} = (V_{DB}, E_{DB})$ based on *DB* data set
2) A Disease-Gene graph $G_{DGN} = (V_{DGN}, E_{DGN})$ based on *DGN* database
3) a Drug-Disease graph $G_{MC} = (V_{MC}, E_{MC})$ based on *MC* relations

The graphs, stored as sparse Boolean weighted adjacency matrices, are built by multiplying the matrices in Fig. 2.

2.2 Databases for Test and Performance Comparison

RDB [26] The RepoDB recommender system provides data set extracted from DrugCentral [30] and ClinicalTrials [31] originally thought as a benchmark database for drug repurposing systems testing. It counts 10,563 connections between 1,519 unique approved drugs (in *DB* format) and 1,229 unique diseases (in `UMLS` format). We filtered these data on the drugs and diseases

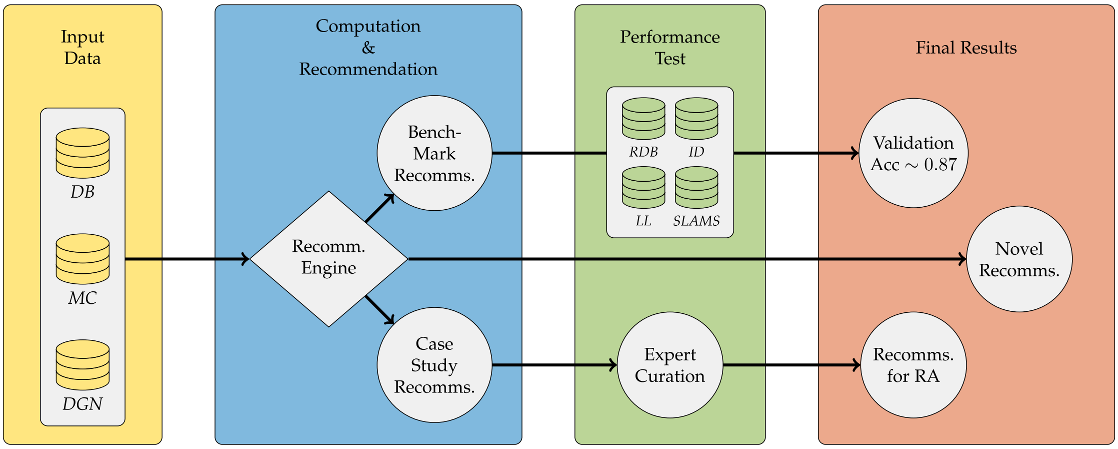

Fig. 1. The architecture of the proposed model. From left to right, the main datasets (see also Fig. 2) are used to build the recommendation engine (see also Fig. 3). The recommendation engine is validated against four benchmark datasets, hence obtaining an average accuracy of 0.87 (see also Table 3 and Fig. 5) and by a manual expert curation in the specific case of Rheumatoid Arthritis, hence obtaining promising recommendations (see also Table 4 and Fig. 6).

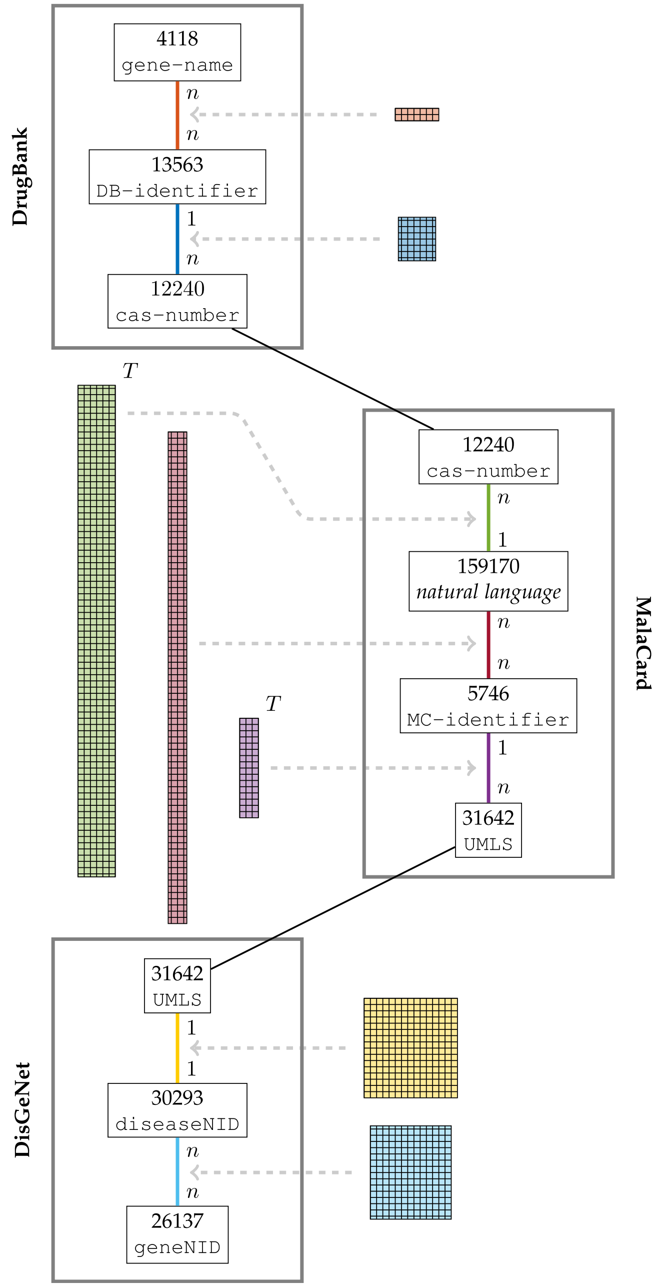

Fig. 2. The structure of the three data sets *DB*, *MC* and *DGN* (top to bottom). Data and connections are shown in the diagram by means of the sparse matrix representation adopted (coloured rectangles with proportional sizes), which grants an efficient and convenient solution to data storage and manipulation.

available in our system, therefore obtaining 2,860 links between 930 drugs and 556 diseases, 49% of which are not in *MC*.

ID [27] The iDrug recommender system provides a drug-diseases data set coming from Comparative Toxicogenomics Database [32] and the gold data set from PREDICT [33]. It is made of 111,481 links between 1,314 drugs (in *DB* format) and 3,966 diseases. The conversion in our nomenclature produced a total of 46,607 links between 813 drugs and 1,335 diseases, 84% of which are not in *MC*.

LL [28] The article by Li and Lu provides a sparse matrix representation of 3,250 links between 799 drugs and 719 diseases, both in natural language. Data are extracted from the National Drug File - Reference Terminology. We were able to extract 1,240 valid links between 673 drugs and 384 diseases, 22% of which are not in *MC*.

SLAMS [29] The SLAMS recommender system provides 3871 interactions between 355 drugs and 406 diseases, both stored in natural language. Data are extracted from *DB* and from the National Drug File - Reference Terminology as for [28]. After our filtering, we obtained 1,420 useful links between 242 drugs and 194 diseases, 77% of which are not in *MC*.

3 Methods

In the following, we describe the implementation of the recommendation engine, starting from how data gathered from the sources (described in Section 2) are used and providing details about the working parameters.

3.1 Assembling Available Data

We employed G_{DB} and G_{DGN} to extract similarities score σ between drugs and diseases respectively, hence obtaining two complete weighted graphs $G_{\text{drg}} = (V_{\text{drg}}, E_{\text{drg}})$ and $G_{\text{dis}} = (V_{\text{dis}}, E_{\text{dis}})$ with drugs/diseases as nodes and similarity scores σ as the weight of the edges. Then, G_{MC} provides a natural way to connect V_{drg} and V_{dis}, hence obtaining a single graph G. A pictorial representation of these steps can be found in Fig. 3 while a pseudo-code is provided in the supplementary materials, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2022.3191392.

We define $p(t)$ as the presence ratio of the gene t in the graph $G_{DB} = (V_{DB}, E_{DB})$, namely

$$p(t) = \frac{|\{u \in V \mid (t,u) \in E_{DB}\}|}{|E_{DB}|}. \tag{1}$$

Here, $p(t)$ is a probability distribution, i.e., $\sum_t p(t) = 1$, therefore, given a gene sets T, the Shannon Entropy

$$H(T) = -\sum_{t \in T} p(t) \cdot \log_2(p(t)), \tag{2}$$

is well defined. We define $V_{\text{drg}} = V_{DB} \cap V_{MC}$ as the set of drugs and we compare two given drugs $u, v \in V_{\text{drg}}$ by means of their target genes sets T_u and T_v as

$$\sigma(u,v) = \frac{2H(T_u \cap T_v)}{H(T_u) + H(T_v)}, \tag{3}$$

hence weighting each gene according to its relevance within the data set. In particular, $0 \le \sigma(u,v) \le 1$, where $\sigma(u,v) = 1$ if and only if $T_u = T_v$ and $\sigma(u,v) = 0$ if and only if $T_u \cap T_v = \emptyset$. We finally consider the drug similarity-weighted graph $G_{\text{drg}} = (V_{\text{drg}}, E_{\text{drg}})$ by defining the weight of the edge as

$$E_{\text{drg}}[u,v] = \begin{cases} \frac{\sigma(u,v)}{\sum_{w \in V_{\text{drg}}} \sigma(u,w)} & \text{if } \sigma(u,v) > 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

In particular, since $\sum_{v \in V_{\text{drg}}} E_{\text{drg}}[u,v] = 1 \; \forall u \in V_{\text{drg}}$ – namely, it is right-stochastic – then E_{drg} can be interpreted as a transition matrix.[1]

Following the same pattern on G_{DGN}, we define $G_{\text{dis}} = (V_{\text{dis}}, E_{\text{dis}})$ as the disease similarity-weighted graph, hence making diseases and drugs comparable by means of the similarity score. A pseudo-code summarisig and generalising the procedure used to build the similarity graph can be found in the Supplementary materials, available in the online supplemental material.

We later define two sets of edges $E_{\text{drg→dis}}$ and $E_{\text{dis→drg}}$ to connect V_{drg} to V_{dis} (and vice versa) as

$$E_{\text{drg→dis}}[u,v] = \begin{cases} \frac{1}{|\{w \in V_{\text{dis}} \,|\, (u,w) \in E_{MC}\}|} & \text{if } (u,v) \in E_{MC} \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

for $u \in V_{\text{drg}}$ and $v \in V_{\text{dis}}$, and

$$E_{\text{dis→drg}}[u,v] = \begin{cases} \frac{1}{|\{w \in V_{\text{drg}} \,|\, (u,w) \in E_{MC}\}|} & \text{if } (u,v) \in E_{MC} \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

for $u \in V_{\text{dis}}$ and $v \in V_{\text{drg}}$. Such edges generates two right-stochastic complete bipartite graphs on the vertex set $V = V_{\text{dis}} \cup V_{\text{drg}}$.[2] A summary of the approach is provided in the Supplementary materials by means of a pseudo-code, available in the online supplemental material.

The union of the four set of edges E_{dis}, E_{drg}, $E_{\text{drg→dis}}$, and $E_{\text{dis→drg}}$ yields a complete weighted graph in the vertices V, namely $G = (V, E)$. In order to preserve the transition matrix interpretation, we halve the weights of each of the edge sets and we set the weight $E[v,v]$ for each vertex $v \in V$ as $1 - \sum_{w \neq v} E[v,w]$,[3] namely

$$E[u,v] = \begin{cases} 1 - \sum_{w \neq v} E[u,w] & \text{if } u = v \\ E_{\text{drg}}[u,v] & \text{if } u,v \in V_{\text{drg}} \\ E_{\text{dis}}[u,v] & \text{if } u,v \in V_{\text{dis}} \\ E_{\text{drg→dis}}[u,v] & \text{if } u \in V_{\text{drg}}, v \in V_{\text{dis}} \\ E_{\text{dis→drg}}[u,v] & \text{if } u \in V_{\text{dis}}, v \in V_{\text{drg}} \end{cases}. \quad (7)$$

The transition matrix of the graph obtained via Equation (7) represents the base of our DR. In the following, we will denote it as $R^{(1)}$, since it holds all the information we know from the very first step of our DR.

3.2 The Core of the Recommendation System

In this section, we exploit the transition matrix property of $R^{(1)}$ to model our system as a recommender for the most probable destination of fixed-length biased random walks.

We define a ℓ-length $R^{(1)}$-biased random walk over $\mathcal{V}$ as a sequence of $\ell + 1$ nodes

$$X = (\overbrace{X^{(0)}}^{\text{source}}, X^{(1)}, \ldots, \overbrace{X^{(\ell)}}^{\text{dest}}), \quad X^{(i)} \in \mathcal{V} \quad (8)$$

sampled according $R^{(1)}$, that is

$$\mathbb{P}\{X^{(i+1)} = v, \,|\, X^{(i)} = u\} = R^{(1)}[u,v]. \quad (9)$$

We define the ℓ recommendation $R^{(\ell)}$ as

$$R^{(\ell)}[u,v] = \mathbb{P}\{X^{(\ell)} = v, \,|\, X^{(0)} = u\} \quad (10)$$

that can be evaluated as

$$R^{(\ell)}[u,v] = \overbrace{R^{(1)} \times \ldots \times R^{(1)}}^{\ell\text{-times}}[u,v]. \quad (11)$$

Given a percentage of drugs $0 < \mathfrak{p} < 1$ we want to recommend, we define the drug recommendation system $\mathfrak{R}_\ell^{\mathfrak{p}} \equiv \mathfrak{R}(R^{(1)}, \ell, \mathfrak{p})$, as the map

$$\begin{array}{cccc} \mathfrak{R}_\ell^{\mathfrak{p}}: & V_{\text{dis}} & \rightarrow & V_{\text{drg}}^{p} \\ & u & \mapsto & \operatorname{argmax}_{v \in V_{\text{dis}}}^{p}(R^{(\ell)}[u,v]) \end{array}, \quad (12)$$

where $p = \lfloor \mathfrak{p} \cdot |V_{\text{drg}}| \rfloor$ ($\lfloor x \rfloor$ is the integer part of x) and $\operatorname{argmax}_\circ^p(f(\circ))$ are the pre-images $\circ$ of the first p maximum values of $f(\circ)$.

The definition of the recommendation system also makes the methodology self-explainable. In fact, given a recommendation r for a disease d, we can query the system to retrieve which paths contribute most for the novel connection itself, i.e., the heaviest paths $d = w_0 \; w_1 \ldots w_\ell = r$.

We can also define an analogous *disease* recommendation system $\mathfrak{R}_\ell^{\mathfrak{p}} : V_{\text{drg}} \rightarrow V_{\text{dis}}^{\mathfrak{p}}$ that, given a specific drug, unveils its relationship with other diseases. Due to the asymmetric nature of the argmax operator and of the matrix $R^{(1)}$, the recommendations generated by the two methodologies may differ i.e., $\mathfrak{R}_\ell^{\mathfrak{p}}$ is not symmetric. In fact, given a disease $d \in V_{\text{dis}}$, we can have $r \in \mathfrak{R}_\ell^{\mathfrak{p}}(d)$ for some drug $r \in V_{\text{drg}}$ while $d \notin \mathfrak{R}_\ell^{\mathfrak{p}}(r)$. In other words, it could happen that the DR does not recommend a disease d to be treated via a specific drug r even if r was recommended from the treatment of that specific disease d.

3.3 The Parameters

Assuming $R^{(1)}$ as fixed, there are two parameters in $\mathfrak{R}_\ell^{\mathfrak{p}}$: the length of the walk $\ell > 1$ and the percentage of recommendations $0 < \mathfrak{p} < 1$.

In order to provide an upper bound for ℓ, we notice that the biased random walk X is a memory-less process and therefore we tackle the task to determine the argmax as a *Markov process* problem $\mathcal{G}$. In particular, since $R^{(1)}$ is stochastic, the graph G itself forms the basis for a ℓ-step Markov chain (see [34] for a detailed description of Markov processes and Markov chains).

Without loss of generality, we may assume that $\mathcal{G}$ is irreducible, that is for all couples $u, v \in \mathcal{V}$, there exists $\ell > 0$ such that $R^{(\ell)}[u,v] > 0$. In fact, if it is not irreducible, then there exists a disjoint partition of $\mathcal{V}$ and therefore we can

[1] Actually, some of the rows could be empty if a drug has no common target genes with the others, hence making necessary the "otherwise" clause; we get rid of this problem, also causing the matrix not to be right-stochastic, in the next steps.

[2] See Footnote 1

[3] It is a good practice to reduce each positive entry $E[u,v]$ by a factor $o(\min\{e \in E \,|\, e > 0\})$ before evaluating $E[u,u]$. This operation gets rid of any numerical issue caused by rounding operations on small floating-point values, hence ensuring no negative value is assigned to $E[u,u]$.

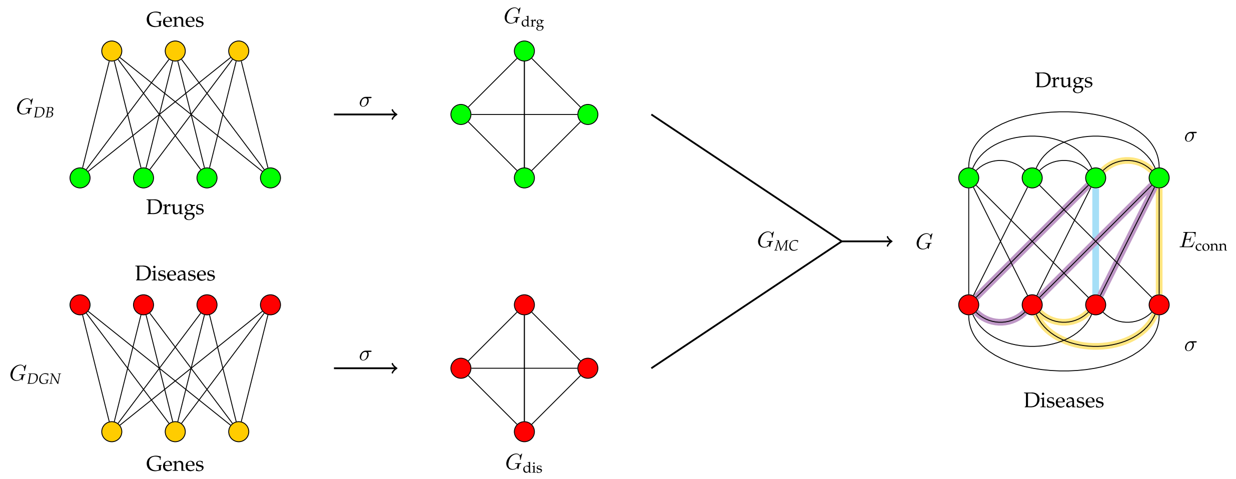

Fig. 3. The data sets *DB* and *DGN* (on the left) are processed according to the σ measure, yielding G_{drg} and G_{dis} respectively. These two graphs are then connected by E_{conn} edges, obtained from *MC* (with weight set to 1). The recommendation shown in cyan is obtained from $R^{(\ell)}$ that relies on biased random walks like yellow and purple ones.

split it into independent irreducible Markov processes and work on them individually.

According to (7), for any state $u \in \mathcal{V}$ we have $R^{(1)}[u,u] > 0$ and therefore $\mathcal{G}$ is positive recurrent and aperiodic. An irreducible, aperiodic and positive recurrent Markov Chain is said to be Ergodic. The Ergodic Theorem states

$$R^{(\ell)} = {R^{(1)}}^{\ell} \xrightarrow{\ell \to \infty} \Pi, \tag{13}$$

holds for every Ergodic Markov chain R and for some stationary distribution $\Pi = (\pi|\pi|\ldots|\pi)^T$, and, in general,

$$\forall \epsilon > 0,\ \exists \bar{\ell} < \infty \quad \text{s.t.} \quad \|\mathcal{E}^{(\bar{\ell})} - \Pi\| < \epsilon. \tag{14}$$

Then $\bar{\ell}$ is an upper bound for ℓ since $\mathfrak{R}_{\bar{\ell}}^{\mathfrak{p}}$ would give the same recommendation independently from the disease considered, hence being uninformative for repurposing scope.

However, the existence of a stationary distribution π enforces the stability of the method. In fact, slightly modifying $\mathcal{E}$ (i.e., partially modifying the initial conditions) we are expected to reach a *similar* π', i.e., $\|\pi - \pi'\|$ is negligible.

4 Performance Evaluation

In this section, we provide an analysis of our method by means of the commonly used performance indicators based on the confusion matrix obtained from the method's execution. We first analyse the method on its own, hence providing parameters tuning by means of ROC curves. Later, we apply our methodology to cross-validation datasets provided by other literature recommender systems.

4.1 Parameters Tuning

As pointed out in Section 3.3, the recommendation system $\mathfrak{R}_{\ell}^{\mathfrak{p}}$ can be tuned by means of two parameters:

- The length of the Markov Process $1 < \ell \ll \bar{\ell}$ (discrete)
- The recommendations percentage $0 < \mathfrak{p} < 1$ (continuous)

One of the most used tools in parameter tuning is the Receiver Operating Characteristic (ROC) curve (true-positive rate versus false-positive rate) constructed via the confusion matrix: at the varying of the parameters.

		$\mathfrak{R}_{\ell}^{\mathfrak{p}}$	
		Recomm.	Not Recomm.
G T	Recomm.	Correctly Recomm.	Wrongly Not Recomm.
	Not Recomm.	Novel Recomm.	Not Recomm.

The objective is to reduce the false-positive rate. As it is common practice in the recommender systems literature, even if little or no *a priori* information on the ground truth (GT) negatives is available (the absence of a given recommendation might be due to its effectiveness not being assessed, yet), it is assumed that only a small fraction of the negatives are actually to be recommended.

For small values of ℓ, we can then build a ROC curve on the parameter $\mathfrak{p}$, varying on $0 \leq \mathfrak{p} \leq 1$. Fig. 4 represents such ROC curves with $1 < \ell \leq 5$. The corresponding area under the curve (AUC) values approximated via the Newton-Cotes formula are reported in Table 2.

Results are listed with two indicators: in *unweighted* rates are evaluated with regard to the complete data set (without taking into account the difference between the drugs); *weighted* rates, on the contrary, are first evaluated regarding the drugs individually and then averaged amongst them.

Fewer known-recommendations are, however, more unreliable in the system as highlighted by the fact that Weighted AUC values are smaller than Unweighted ones. In fact, they have a stronger negative impact on the mean accuracy than those with more recommendations available.

According to ROCs, the best choice is $\ell = 3$. Recommendation $(u,v) \in V_{\text{drg}} \times V_{\text{dis}}$ are therefore built employing the following patterns:

- $u \mapsto u' \mapsto u'' \mapsto v$
- $u \mapsto u' \mapsto v'' \mapsto v$

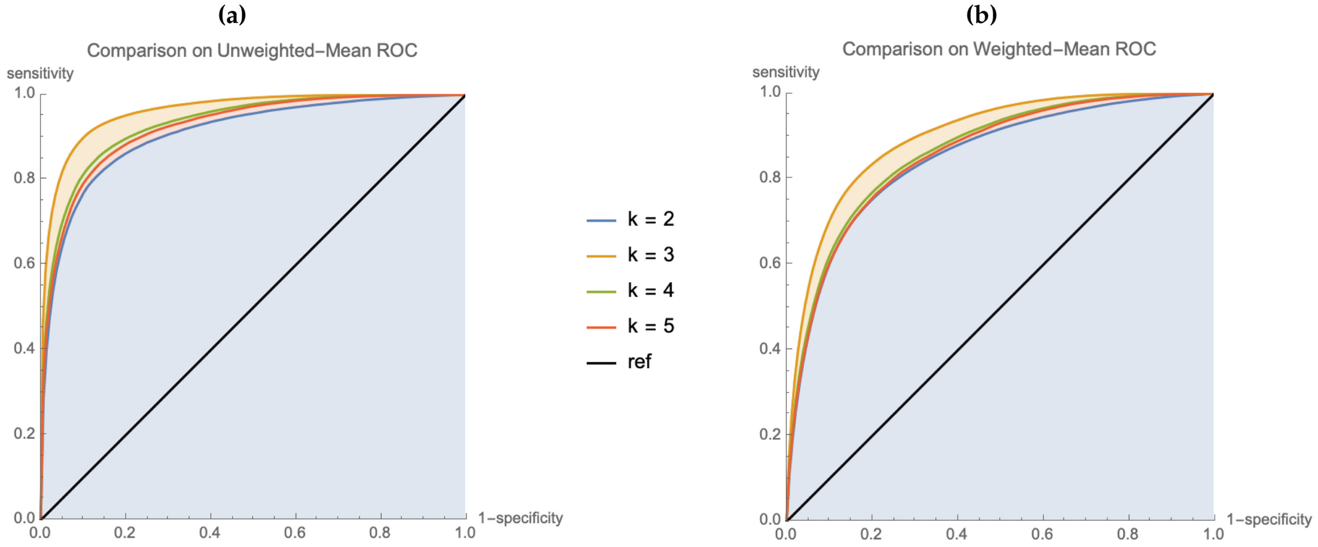

Fig. 4. The ROC curves obtained by the recommendation system with various path lengths ℓ. The axis represents false- and true-positive ratios. Curves are sampled with rate of $\Delta\mathfrak{p} = 0.01$. Numeric values of the AUC (integral of the curve) are reported in Table 2. **(a)** true- and false-positive ratios are evaluated with regards to the complete data-set. **(b)** true- and false-positive ratios are first evaluated for each drug independently and then the average is computed.

- $u \mapsto v' \mapsto u'' \mapsto v$
- $u \mapsto v' \mapsto v'' \mapsto v$

where $u,' u'' \in V_{\text{drg}}$ and $v,' v'' \in V_{\text{dis}}$ and u, u' and u'' might be the same state as well as v, v' and v''.

For $\ell = 3$, we evaluate the best value of $\mathfrak{p}$ as the parameter generating the furthest point from the reference diagonal

$$\underset{\text{params}}{argmax}\{\text{sensitivity} + \text{specificity}\}, \tag{15}$$

hence, obtaining $\mathfrak{p} = 12.5\%$. In our setting, such a value corresponds to 213 recommendations.

Analogous results are obtained also with leave-one-out 10-fold cross-validation tests, where the AUC results in a mean value of 0.930174 and 0.862815 for weighted and unweighted approaches respectively, both with a standard deviation of 10^{-4}.

4.2 Comparison With Other Methods

To provide more insightful information, we compare our method against the four benchmark data sets *RDB*, *ID*, *LL*, and *SLAMS* introduced in Section 2. To keep recommendations consistent with our method, we restricted such data sets by filtering the diseases and the drugs available in our training and then we collect the result regarding their connections. In situations like these ones, two main approaches are possible: (i) a gentle re-tuning of the method with the new data set in order to prove the model's flexibility to different data or (ii) a direct usage of the data set as a cross-validation set. While the first approach consists in performing parameter analysis again on the novel dataset, hence 'forgetting' the original set of data and therefore taking the best results of the methodology over the benchmarks, the second approach – the one we adopted – tests the benchmarks over the parameter gathered from the original dataset. In this sense, we compared the recommendations our method proposes with the ground truth obtained from the four restricted data sets. However, to keep the environment of the test as close to reality as possible, we did not restrict our method's recommendation set to the drugs available in each benchmark data set, meaning that we made our method infer no information about the validation set it was tested against; if this was the case, in fact, sensitivity and specificity values would grow above $(0.99, 0.9)$ due to the high number of recommendations, hence being uninformative.

TABLE 2
AUC at the Varying of ℓ

ℓ	unweighted mean AUC	weighted mean AUC
2	0.908442	0.847814
3	0.961536	0.897780
4	0.931462	0.863560
5	0.923588	0.856419

In Fig. 5 we report, with respect to the Weighted-mean ROC curve, the true- and false-positive rates achieved on

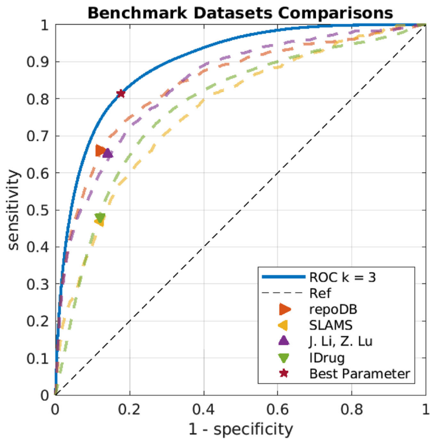

Fig. 5. Plot of the benchmark datasets indicators (from Table 3) w.r.t. the ROC curve achieved by $\mathfrak{R}_3^{\mathfrak{p}}$ (Fig. 4(b)). The best setting $\mathfrak{p} = 12.5\%$ is also plotted as a star.

TABLE 3
Benchmark Performances Indicators

Dataset	Shared recoms	Accuracy	Sensitivity	Specificity	AUC
RDB	51%	0.875031	0.660791	0.875903	0.8417
ID	16%	0.873807	0.486318	0.881105	0.7811
LL	78%	0.852050	0.646048	0.853405	0.8415
SLAMS	23%	0.873245	0.469907	0.875486	0.7701

the benchmark datasets according to the parameters extracted in Section 4. We also collect the corresponding numeric values in Table 3 along with the Accuracy boolean indicator and the area under the curve of the underlying ROC curve.

As it is expected, better sensitivity and specificity results are obtained on *LL* and *RDB* since they share more than 50% of the recommendation with our training set *MC*. If we consider *SLAMS* and *ID*, while having less than $^1/_4$ of common data with *MC*, they achieve anyway a comparable Accuracy and Specificity. Such comparable results provide a further clue of our method's stability.

5 Case Study: Rheumatoid Arthritis

In addition to the performances assessed in Section 4 we here explore in more detail the results on an exemplar non-communicable disease (NCD), i.e., rheumatoid arthritis (RA), a worldwide threat associated with spreading chronic inflammation [35]. NCDs provoke, in fact, more than 44 million deaths per year [36], [37] and it is estimated that RA, amongst the others, affects around 1% of the world population [38]. RA's incidence and representativeness make it an interesting case study to explore the relevance of the results offered by the recommender system also being its aetiology complex and not fully elucidated, since it includes both genetic and environmental factors [39], [40].

Standard therapy (true positives) for RA was extracted from the 2021 update of the American College of Rheumatology (ACR, [41]). Then, the relevance of the findings towards potential clinical translation was manually curated for the top-ranking novel results, using two sources of knowledge to deepen the information retrieved by each entry. The first source is represented by PubMed [42], the most comprehensive medical database. The second one is the well-known repository ClinicalTrials [31] collecting concluded and ongoing clinical trials. In both cases, the search terms were represented by the disease (rheumatoid arthritis) and the drug, as recovered by the recommender system.

The top ten novel (w.r.t. the dataset) recommendation are reported in Table 4. Score, drug name and CAS number are the output of the recommender system, while clinical trials and related publications are the results of our manual curation of the findings. Recommendations are divided according to the associated evidence (publication or clinical study), into three sets: (i) *approved:* that are the ones already employed in the literature (see [41]) but not in our training set *MC*, (ii) *potential*, that are the experimentally promising according to recent publications (reported in the table as well), and (iii) *open*, that are the ones supported by no hints of recent/ongoing related investigation.

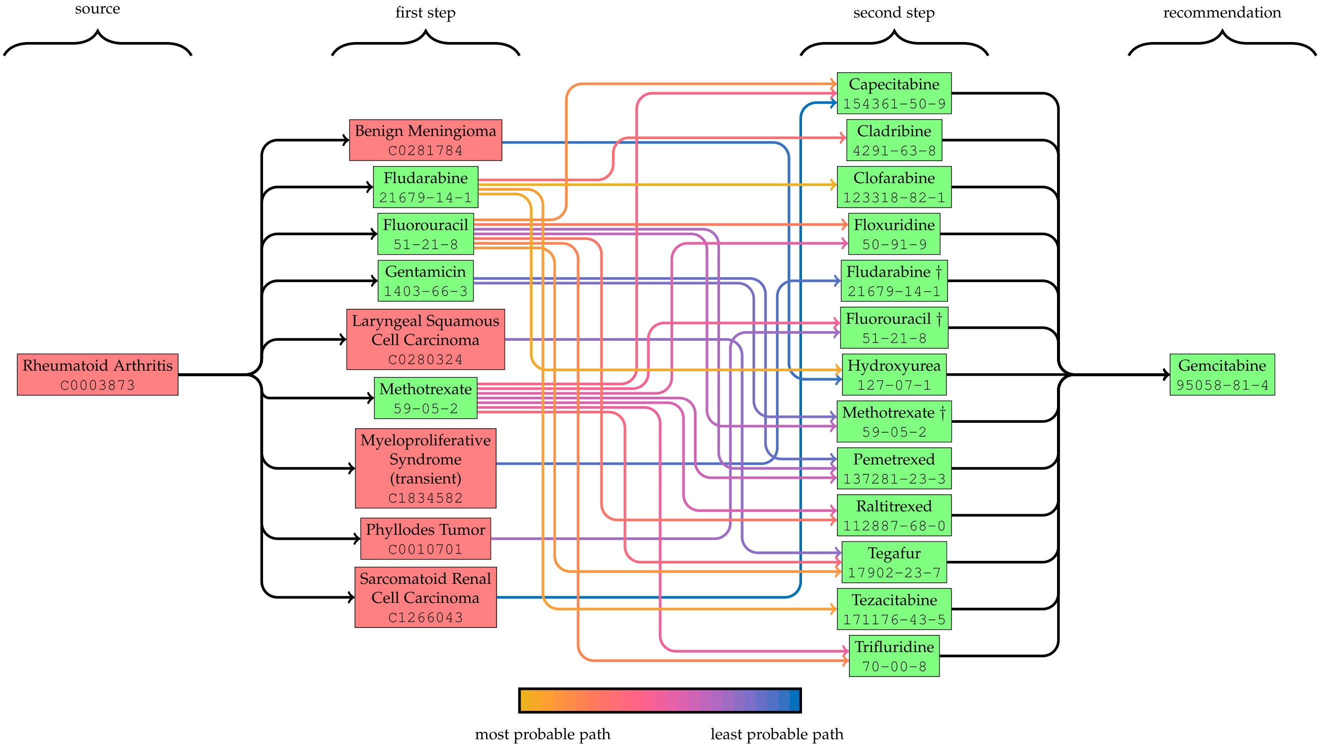

Fig. 6. Depiction of the recommendation process of the drug Gemcitabine in the case study of RA (top 25 path contributions shown, coloured from yellow, as the most probable path, to blue, as the least probable one). From left to right, the disease RA ("source") is linked in the first step of the process with similar diseases (in red) according to σ scores and approved drugs (in green) through known associations. Then, a second step is performed and a different set of drugs is reached (some of the nodes are returning from the first step, marked with †). Finally, the paths reach Gemcitabine with a third (and last, $\ell = 3$) step, generating the recommendation.

TABLE 4
Best Novel Recommendations for RA

Score [%]	Drug Name	CAS-Number	# Trials	Relevance
3.883	Dihydrofolic Acid	4033-27-6	0	open
3.3545	Folic acid	59-30-3	307	approved
3.0627	Cyclophosphamide	50-18-0	7	approved
2.1424	Gemcitabine	95058-81-4	0	potential [43]
2.0943	Doxorubicin	23214-92-8	0	open
2.0700	Tretinoin	302-79-4	2	open
1.9034	Farletuzumab	896723-44-7	0	potential [44]
1.8320	Cisplatin	15663-27-1	0	open
1.7145	Lenalidomide	191732-72-6	0	potential [45]
1.7101	Docetaxel	114977-28-5	0	open

Approved drugs offer a measure of the validation of the system itself. *Potential* results, appear to be mostly identified by drugs whose validation in clinical trials is not (yet) engaged, but whose usage in animal models (mostly collagen-induced arthritis – CIA – rodents) has offered recent, promising results. This is the case for Gemcitabine (95058-81-4) [43] and Lenalidomide (191732-72-6) [45], whose identification by our system is particularly interesting due to their recent developments (i.e., after 2017).

Further inspection can be carried out thanks to the explainability of the output provided by the Markov Process representation of the DR system. As an example, Fig. 6 reports the 25 paths of length $\ell = 3$ that are contributing most to Gemcitabine novel recommendation, hence giving hints on the connections found. The connection with highest value is {RA → Fludarabine → Clofarabine → Gemcitabine}. More in general, all these 25 connections share the same few drugs (14) and diseases (5) with an interesting relevance in Fludarabine (CAS 21679-14-1, present in five paths), Fluorouracil (CAS 51-21-8, present in nine paths), and Methotrexate (CAS 59-05-2, present in nine paths); these paths consists in fact in 21 out of the 25 total (do note that such drugs are present in the same path twice). All of these anti-neoplastic drugs can be found at a distance of one from RA, meaning that they are present in *MC* dataset.

The category *open*, having no current support in the literature, can represent both (expected) noise offered by such an automated system building on a meaningful but limited data base or the most innovative opportunities. In this sense, the explainability of the recommendation represents a significant added value, offering pharmacologists and rheumatologists a clear starting point for further consideration as well as the feasibility of (early) translation into clinical testing.

6 Discussion and Conclusion

Drug recommendation processes encompass experimental and computational approaches, with the latter potentially leveraging on computational resources as well as algorithmic advances, especially in the area of machine learning [46]. The power and flexibility of such tools created high expectations in the field of *personalised medicine*, a twenty years-long effort in the evolution of the medical paradigm [47], whose latest interpretations rely on machine learning for applications ranging from diagnosis to therapy. In this setting, however, high-level transparency and accountability are mandatory, hence promoting explainable machine learning methods [48].

In this perspective, we here proposed a similarity-based approach that exploits established associative knowledge on experimental drug-gene molecular interactions, disease-drug and disease-gene associations, and pharmacological (disease-drug) indications to build a predictive DR system.

The way the proposed approach is built grants several advantages. Amongst the others, one is the standardisation, interoperability, quality and up-to-date information obtained thanks to the choice of robust, binary-type data. Moreover, the use of associative binary data (i.e., drug-disease, gene-disease, drug-gene) allows to overcome the problem of the trustworthiness and sensibility to experimental settings of quantitative data (such as e.g., gene expression data, valuable and extremely refined but also sensitive to noise [49]), and to take advantage of prior knowledge characterised by high robustness, reliability and stability. Another key feature is represented by the generation of explainable prediction, which can be obtained, once again, since the choice made here to integrate several binary associative data sets from reliable sources (DrugBank, DisGeNET, MalaCards) allowed to build a robust and well-grounded set of data that the recommender system is able to work on with a transparent approach. The output is hence completely traceable, from the processing of the initial input sources (diseases, drugs) up to the formulation of the drug repurposing proposal, hence promoting it as an explainable artificial intelligence method.

On the other hand, typical issues of associative data generate DR shortcomings, including noisy recommendations (e.g., false positive interactions) or incompleteness in the data sets (false negative interactions). Here, the incorporation of sources containing *negative* data (e.g., experimentally validated non-interacting pairs such as the Negatome for protein-protein interactions [50]) or important pharmacological side effects that impair drug usage (e.g., data from SIDER [51]) may help in refining the output for practical, clinical use. Multiple, redundant or complementary data sources (e.g., the Drug Repurposing Knowledge Graph [52]) will also help in this direction to cover missing information as well as possible.

In future implementations, we plan to extend the capability of managing information about drug–drug interactions, hence performing a step forward in the problem of synergistic drug combinations while tackling the repurposing proposals issues. To this scope, we plan to include the resources for drug synergy (e.g., DrugCombDB [53]) and adversary effects (e.g., from DrugBank).

We also plan to include micro-RNA interactions, a growing field of interest in diagnostic and therapeutic pharmacology (see [54]), to further strengthen our knowledge-graph; in fact, miRNA plays an important role similar to target genes both for what concerns diseases (see [24]) and drugs (see [55]).

Acknowledgments

This work has been carried out under the auspices of the Gruppo Nazionale per la Fisica Matematica (GNFM) of Istituto Nazionale di Alta Matematica (INdAM).

Author Contributions

EO executed the study, provided the mathematical background, developed the code to perform the analysis and validation of the method and wrote the first draft. FC provided the initial idea and supervised the work. CN suggested the use-case, supervised its analysis and performed its validation. MP supported with the mathematical formalisation and the writing of the code. PT provided support in data acquisition and problem formulation. All authors contributed to the writing of the manuscript, revised it, and read and approved the final version.

References

[1] Repurposing drugs to treat COVID-19: Interview with David Fajgenbaum, Apr. 2021. [Online]. Available: REF:ascodaily.libsyn.com

[2] P. Venkatesan, "Repurposing drugs for treatment of COVID-19," *Lancet Respir. Med.*, vol. 9, no. 7, Jul. 2021, Art. no. e63.

[3] Corona project. Oct. 2021. [Online]. Available: https://cdcn.org/corona/

[4] A. D. Hingorani *et al.*, "Improving the odds of drug development success through human genomics: Modelling study," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019, Art. no. 18911.

[5] X. Chen, B. Ren, M. Chen, Q. Wang, L. Zhang, and G. Yan, "NLLSS: Predicting synergistic drug combinations based on semi-supervised learning," *PLoS Comput. Biol.*, vol. 12, no. 7, 2016, Art. no. e1004975.

[6] C. Zhang and G. Yan, "Synergistic drug combinations prediction by integrating pharmacological data," *Synthetic Syst. Biotechnol.*, vol. 4, no. 1, pp. 67–72, 2019.

[7] M. J. Keiser *et al.*, "Predicting new molecular targets for known drugs," *Nature*, vol. 462, no. 7270, pp. 175–181, Nov. 2009.

[8] X. Chen *et al.*, "Drug–target interaction prediction: Databases, web servers and computational models," *Brief. Bioinf.*, vol. 17, no. 4, pp. 696–712, 2016.

[9] H. Luo *et al.*, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics*, vol. 32, no. 17, pp. 2664–2671, May 2016. [Online]. Available: http://github.com//bioinfomaticsCSU/MBiRW

[10] C. E. Lipscomb, "Medical subject headings (mesh)," *Bull. Med. Library Assoc.*, vol. 88, no. 3, 2000, Art. no. 265.

[11] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Res.*, vol. 46, pp. D1074–D1082, Jan. 2018.

[12] J. S. Amberger and A. Hamosh, "Searching online mendelian inheritance in man (OMIM): A knowledge base of human genes and genetic phenotypes," *Curr. Protoc. Bioinf.*, vol. 58, pp. 1.2.1–1.2.12, Jun. 2017.

[13] Y. Nam, M. Kim, H.-S. Chang, and H. Shin, "Drug repurposing with network reinforcement," *BMC Bioinf.*, vol. 20, no. 13, Jul. 2019, Art. no. 383.

[14] M. G. Ozsoy, T. Özyer, F. Polat, and R. Alhajj, "Realizing drug repositioning by adapting a recommendation system to handle the process," *BMC Bioinf.*, vol. 19, no. 1, Apr. 2018, Art. no. 136.

[15] Z. Tanoli, U. Seemab, A. Scherer, K. Wennerberg, J. Tang, and M. Vähä-Koskela, "Exploration of databases and methods supporting drug repurposing: A comprehensive survey," *Brief. Bioinf.*, vol. 22, no. 2, pp. 1656–1678, Feb. 2020.

[16] P. Tieri and C. Nardini, "Signalling pathway database usability: Lessons learned," *Mol. Biosyst.*, vol. 9, no. 10, pp. 2401–2407, Oct. 2013.

[17] W. H. Organization, "ICD-11 reference guide: International classification of diseases for mortality and morbidity statistics," 2022. [Online]. Available: GUIDE:icd.who.int

[18] J. Piñero *et al.*, "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D845–D855, Nov. 2019.

[19] L. Ehrlinger and W. Wöss, "Towards a definition of knowledge graphs," *Proc. SEMANTICS 2016: Posters Demos Track*, vol. 1695, pp. 1–4, 2016.

[20] L. S. Blackford *et al.*, "An updated set of basic linear algebra subprograms (BLAS)," *ACM Trans. Math. Softw.*, vol. 28, no. 2, pp. 135–151, 2002.

[21] CAS registry [online], American Chemical Society, 2022. [Online]. Available: FAQ:cas.org

[22] O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Res.*, vol. 32, no. Database, pp. D267–270, Jan. 2004.

[23] N. Rappaport *et al.*, "MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D877–D887, Nov. 2016.

[24] C. Gu *et al.*, "Network-based collaborative filtering recommendation model for inferring novel disease-related miRNAs," *RSC Adv.*, vol. 7, pp. 44 961–44 971, 2017.

[25] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42 200–42 216, 2020.

[26] A. S. Brown and C. J. Patel, "A standard database for drug repositioning," *Sci. Data*, vol. 4, no. 1, pp. 1–7, 2017.

[27] H. Chen, F. Cheng, and J. Li, "iDrug: Integration of drug repositioning and drug-target prediction via cross-network embedding," *PLoS Comput. Biol.*, vol. 16, no. 7, 2020, Art. no. e1008040. [Online]. Available: https://github.com/Case-esaC/iDrug

[28] J. Li and Z. Lu, "A new method for computational drug repositioning using drug pairwise similarity," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2012, pp. 1–4.

[29] P. Zhang, P. Agarwal, and Z. Obradovic, "Computational drug repositioning by ranking and integrating multiple data sources," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discov. Databases*, 2013, pp. 579–594.

[30] Drugcentral [internet], Division of Translational Informatics at University of New Mexico, Oct. 2021. [Online]. Available: https://drugcentral.org

[31] Clinical trials [internet], "Bethesda (MD): National library of medicine (US)," 2021. [Online]. Available: https://www.clinicaltrials.gov/

[32] A. P. Davis *et al.*, "Comparative toxicogenomics database (CTD): Update 2021," *Nucleic Acids Res.*, vol. 49, no. D1, pp. D1138–D1143, 2021.

[33] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalised medicine," *Mol. Syst. Biol.*, vol. 7, no. 1, 2011, Art. no. 496.

[34] D. A. Levin and Y. Peres, *Markov Chains and Mixing Times*. Providence, Rhode Island: American Mathematical Society, 2017.

[35] M. G. Maturo, M. Soligo, G. Gibson, L. Manni, and C. Nardini, "The greater inflammatory pathway-high clinical potential by innovative predictive, preventive, and personalized medical approach," *EPMA J.*, vol. 11, no. 1, pp. 1–16, Mar. 2020.

[36] D. Abegunde and A. Stanciole, "An estimation of the economic impact of chronic noncommunicable diseases in selected countries," World Health Organization, Department of Chronic Diseases and Health Promotion, vol. 2006, 2006. [Online]. Available: https://www.researchgate.net/publication/253888397_An_estimation_of_the_economic_impact_of_chronic_noncommunicable_disease_in_selected_countries

[37] A. Saha and G. Alleyne, "Recognizing noncommunicable diseases as a global health security threat," *Bull. World Health Org.*, vol. 96, no. 11, 2018, Art. no. 792.

[38] Y. Xu and Q. Wu, "Prevalence trend and disparities in rheumatoid arthritis among us adults, 2005–2018," *J. Clin. Med.*, vol. 10, no. 15, 2021, Art. no. 3289.

[39] G. B. Rogers, “Germs and joints: The contribution of the human microbiome to rheumatoid arthritis,” *Nature Med.*, vol. 21, no. 8, pp. 839–841, 2015.

[40] Y. Okada *et al.*, “Genetics of rheumatoid arthritis contributes to biology and drug discovery,” *Nature*, vol. 506, no. 7488, pp. 376–381, 2014.

[41] L. Fraenkel *et al.*, “2021 american college of rheumatology guideline for the treatment of rheumatoid arthritis,” *Arthritis Care Res. (Hoboken)*, vol. 73, no. 7, pp. 924–939, Jul. 2021.

[42] Pubmed [internet], “Bethesda (MD): National library of medicine (US),” 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/

[43] A. F. Dağli *et al.*, “Antiinflammatory and antioxidant effects of gemcitabine in collagen-induced arthritis model,” *Turk J. Med. Sci.*, vol. 47, no. 3, pp. 1037–1044, Jun. 2017.

[44] Y. Feng *et al.*, “A folate receptor beta-specific human monoclonal antibody recognizes activated macrophage of rheumatoid patients and mediates antibody-dependent cell-mediated cytotoxicity,” *Arthritis Res. Ther.*, vol. 13, no. 2, Apr. 2011, Art. no. R59.

[45] B. Lopez-Millan *et al.*, “Therapeutic effect of the immunomodulatory drug lenalidomide, but not pomalidomide, in experimental models of rheumatoid arthritis and inflammatory bowel disease,” *Exp. Mol. Med.*, vol. 49, no. 2, Feb. 2017, Art. no. e290.

[46] S. Pushpakom *et al.*, “Drug repurposing: Progress, challenges and recommendations,” *Nature Rev. Drug Discov.*, vol. 18, no. 1, pp. 41–58, Jan. 2019.

[47] C. Nardini *et al.*, “The evolution of personalized healthcare and the pivotal role of european regions in its implementation,” *Future Med.*, vol. 18, pp. 283–294, Apr. 2021.

[48] E. Tjoa and C. Guan, “A survey on explainable artificial intelligence (XAI): Toward medical XAI,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.

[49] J. M. Raser and E. K. O’Shea, “Noise in gene expression: Origins, consequences, and control,” *Science*, vol. 309, no. 5743, pp. 2010–2013, Sep. 2005.

[50] P. Blohm *et al.*, “Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis,” *Nucleic Acids Res.*, vol. 42, no. Database, pp. 396–400, Jan. 2014.

[51] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The SIDER database of drugs and side effects,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–1079, Jan. 2016.

[52] V. N. Ioannidis *et al.*, “DRKG - Drug repurposing knowledge graph for COVID-19,” 2020.

[53] H. Liu, W. Zhang, B. Zou, J. Wang, Y. Deng, and L. Deng, “DrugCombDB: A comprehensive database of drug combinations toward the discovery of combinatorial therapy,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D871–D881, Jan. 2020, doi: 10.1093/nar/gkz1007.

[54] X. Chen, N.-N. Guan, Y.-Z. Sun, J.-Q. Li, and J. Qu, “MicroRNA-small molecule association identification: From experimental results to computational models,” *Brief. Bioinf.*, vol. 21, no. 1, pp. 47–61, 2020.

[55] X. Chen, C. Zhou, C.-C. Wang, and Y. Zhao, “Predicting potential small molecule–miRNA associations based on bounded nuclear norm regularization,” *Brief. Bioinf.*, vol. 22, no. 6, 2021, Art. no. bbab328.

Filippo Castiglione received the degree in computer science from the University of Milan, Italy, and the PhD degree in scientific computing from the University of Cologne, Germany. He was a visiting researcher in various centers (IBM T.J. Watson Research Center, the Department of Molecular Biology of Princeton University, the Harvard Medical School, and the Institute for Medical Bio-Mathematics in Tel Aviv) working with the interface between computation and biology. He has coordinated several EU projects related to health informatics. He currently works as a research director with the National Research Council of Italy and also teaches Computational Biology and Machine Learning with the University of RomaTre.

Christine Nardini received the PhD degree in electronics and informatics from the University of Bologna, Italy. She is currently working as a scientist with the National Research Council of Italy (CNR), Institute for Applied Mathematics, Rome, Italy. Her research interests include systems and computational biology. She is a member of the editorial board of *PLoS One* and *BioNanoScience*, Springer Publication and associate editor for *BMC Bioinformatics*.

Elia Onofri received the bachelor’s degree in mathematics, and the master’s degree in computational sciences from Roma Tre University, Rome, Italy, in 2018 and 2020, respectively. Currently, he is working toward the PhD degree in mathematics with Roma Tre University, under a collaboration with the National Research Council of Italy. His research fields concern applied mathematics in cryptography, machine learning, biology and forecasting simulations.

Marco Pedicini. received the PhD degree in mathematics (Logic and Theoretical Computer Science) from Paris 7 University. He is an associate professor with the Department of Mathematics and Physics of Roma Tre University. Before coming to Roma Tre he was a CNR researcher. His research is on topics with the intersection of theoretical computer science and pure mathematics: logic in computer science, computational number theory, cryptography, and computational methods for systems biology. On behalf of De Cifris, he coordinates the special interest group CifrisCloud.

Paolo Tieri a physicist by training, deals with network medicine, computational and systems biology and bioinformatics. He has been a post-doc researcher with the Interdepartmental Center ‘L. Galvani’ for Biocomplexity and with the Immunology section of the University of Bologna from 2003 to 2011, a visiting scientist with the Dana Farber Cancer Institute, Harvard Medical School, Boston (2005), visiting scientist with the CAS-MPG Partner Institute for Computational Biology, Shanghai (2012-2013). In 2011, he joined the IAC Institute for Applied Computing of the CNR National Research Council in Rome, and currently an adjunct professor of network medicine with the Sapienza University of Rome.

Open Access funding provided by ‘Università degli Studi Roma Tre’ within the CRUI CARE Agreement