This review provides a comprehensive description of the conceptual foundation and computational developments in the field of **in silico** *repurposing. Furthermore, a generic modular description for repurposing workflows is described.*

Design of efficient computational workflows for *in silico* drug repurposing

Quentin Vanhaelen[1], Polina Mamoshina[1], Alexander M. Aliper[1], Artem Artemov[1], Ksenia Lezhnina[1], Ivan Ozerov[1], Ivan Labat[2] and Alex Zhavoronkov[1]

[1] Insilico Medicine Inc., Johns Hopkins University, ETC, B301, MD 21218, USA
[2] BioTime Inc., 1010 Atlantic Avenue, 102, Alameda, CA 94501, USA

Here, we provide a comprehensive overview of the current status of *in silico* repurposing methods by establishing links between current technological trends, data availability and characteristics of the algorithms used in these methods. Using the case of the computational repurposing of fasudil as an alternative autophagy enhancer, we suggest a generic modular organization of a repurposing workflow. We also review 3D structure-based, similarity-based, inference-based and machine learning (ML)-based methods. We summarize the advantages and disadvantages of these methods to emphasize three current technical challenges. We finish by discussing current directions of research, including possibilities offered by new methods, such as deep learning.

Introduction

Currently, pharmaceutical companies face a challenging economical and societal environment that requires them to continuously look for strategies to improve their capacities to develop original drugs at reduced cost [1,2]. Within this context, the pharmaceutical community considers that finding novel indications and targets for already existing drugs, a method called 'drug repurposing', first discussed by Ashburn and Thor in 2004 [3,4], can compensate for the lack of technical efficiency of the traditional drug discovery approaches that results in a high failure rate and continual decline in the number of new approved small-molecular entities released by pharmaceutical industry pipelines [5,6]. The major advantages of a drug-repurposing approach are that the preclinical, pharmacokinetic, pharmacodynamic and toxicity profiles of the drug are already known, reducing the risk of compound development. Thus, the compound can rapidly translate into Phase II and III clinical studies, resulting in a decreased development cost [6], a better return on investment and an accelerated development time [7]. Drug repurposing is also interesting from the point of view of intellectual property (IP) and patent protection, because patent protection for a new use of an existing drug whose composition of matter patents are still running can be obtained assuming that the new use is not covered and proven in the original

Quentin Vanhaelen earned his BSc and MSc in Physics at Université Libre de Bruxelles (ULB) in 2006. He earned his DEA in Sciences at ULB in 2007. He earned his PhD in Theoretical Physics (Grant FNRS–FRIA) at ULB in 2010. During his PhD studies, he developed an interest in questions related to regenerative medicine. He was a postdoctoral fellow at Ottawa Hospital Research Institute (OHRI), Ottawa, Canada from 2011 to 2013. He is now at Insilico Medicine Inc., where he is engaged in research and development devoted to improving *in silico* drug-repurposing algorithms.

Corresponding author: Vanhaelen, Q. (vanhaelen@insilicomedicine.com)

1359-6446/© 2016 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.drudis.2016.09.019

GLOSSARY

Bipartite graph A graph with two types of node (e.g. nodes for disease and nodes for drugs); edges (interactions) between nodes of the same type are prohibited.
Chemoinformatics The field of study of all aspects of the representation and use of chemical and biological information on computers.
Chemoproteomics The field of study linking chemicals to molecular targets with therapeutic indications.
Disease signature A relatively short list of genes associated with disease or drug effects derived either by manual curation or automated filtering from high-throughput experiments.
Genomics Sequencing, assembling and analyzing the function and structure of genomes.
New candidate Either a drug candidate compound that does not have any known targets or a target candidate protein that is not targeted by any drugs or compounds in the network of interactions investigated for repurposing purpose.
Phenomics Measures how the genetic and epigenetic effects affect the physical and biological traits in an organism.
Proteomics The large-scale study of proteins with an emphasis on their structures and functions.
Receiver operating curve (ROC) A quality measure obtained by computing the true positive rate (TPR) and false positive rate (FPR) at different thresholds.
Sensitivity The ratio of the successfully predicted experimentally verified drug–disease associations to the total experimentally verified drug–disease associations. It is mathematically defined as TP/(TP + FN), where TP and FN are the number of true positives and false negatives, respectively.
Specificity The percentage of the negative (unknown or random) drug–disease associations predicted by the algorithm among all negative drug–disease associations. The specificity is computed using the formula TN/(TN + FP), where TN and FP are the number of true negatives and false positives, respectively.

patents [8,9]. Furthermore, reusing already approved drugs can help to protect the original IP of the pharmaceutical company against competitor adjacency moves and can provide alternative models to outlicense some of its clinical drug candidates [10]. For example, the company can retain the original use rights to the drug, and outlicense the rights to the new indication. As a consequence of these opportunities, the initiatives of several governments for developing drug repositioning have also emerged. For example, in the USA, the National Centre for Advancing Translational Sciences (NCATS) has launched the Discovering New Therapeutic Uses for Existing Molecules Programme. In the UK, the Developmental Pathway Funding Scheme of the Medical Research Council (MRC) provides researchers with funding for repurposing clinical studies, whereas the Netherlands Organisation for Health Research and Development (ZonMw) has funded a project on the stimulation of drug rediscovery related to drug repositioning.

Drug repurposing is used to find alternative candidates to cure various types of disease. Moreover, there are high expectations regarding the use of repurposing approaches for addressing major health issues. Examples include Alzheimer's disease, for which several alternative candidates are at different development stages or already in clinical trials [11,12]. Other investigations have looked for alternative candidates for antiaging therapies [13]. Moreover, with only 5% of the oncology drugs that enter Phase I clinical trials being approved, there is great demand for new anticancer drugs and for cell- and target-based screening assays; thus, drug repurposing also attracts attention from the field of anticancer drug discovery [14,15]. Many known drugs, including metformin [16] and vitamin D [17], have been analyzed to identify potential anticancer properties. The advanced development stage and ongoing clinical trials of other alternative candidates are reviewed in [18]. Finally, drug-repurposing methods could help to find cures for orphan diseases [19]. Indeed, there are 400 million people worldwide affected by such diseases, but with current research and development costs, it is impossible to develop *de novo* therapies for each of the 5000–8000 orphan diseases identified so far [20,21].

Drug repurposing is performed either by using an experimental approach, called 'activity-based drug repositioning', or by making use of a specific computational method [18]. The latter approach, named '*in silico* drug repurposing', is one of the latest application areas of computational pharmacology, a larger field that encompasses *in silico*-based methods developed to investigate how drugs affect biological systems. From a technical perspective, the development of efficient algorithms for *in silico* drug repurposing is made possible by two technological trends [18,22]. First, the accumulation of various high-throughput data generated from different research areas, such as proteomics, genomics, chemoproteomics and phenomics. As a result, entire pathway maps, as well as data providing characterizations of disease phenotypes and drug profiles, are available. The second technological trend is the progress made in computational and mathematical sciences [23] that, combined with increasingly powerful computational resources, allows the development of not only repurposing algorithms, but also software for retrospective analysis as well as the maintenance of web-based databases, which are required for the gathering and classification of the experimental data [21,22,24–26].

Compared with activity-based repositioning techniques, *in silico* methods allow a faster repurposing process at a reduced cost. However, these methods require high-resolution structural information of targets as well as either disease and phenotype information or gene expression profiles of drugs, depending on the nature of the targets, making any of them strongly dependent on the availability of experimental data. Moreover, the biological significance of the putative targets predicted by the algorithm must also be assessed. This step necessitates supplementary experimental testing [4,22]. Regardless of these challenges, various directions of research have been followed by the scientific community and the arsenal of traditional methods relying on ligand-based [27] or receptor-based [28] approaches has been enriched with, for instance, network- and phenotypic-based inference algorithms [29,30]. These efforts to improve and extend the use of *in silico* repurposing techniques are also pursued by companies using state-of-the-art computational approaches for prioritizing existing candidates, performing targeted searches and identifying new targets for repurposing. Research and results include the identification of tricyclic antidepressants as inhibitors of small cell lung cancer by

NuMedii [31]; the development of monoclonal antibodies to innovative and therapeutic targets in oncology and autoimmune disease by Capella Biosciences; the development and use of a cloud-based drug discovery platform by TwoXar that aims to find unanticipated associations between drug and disease with a focus on therapeutic areas, including autoimmunology, oncology and neurology; and the development and use of parametric and artificially intelligent drug discovery and repurposing systems by Insilico Medicine [32,33].

Several authors have recently reviewed different aspects of *in silico* repurposing approaches. Hodos *et al.* [21] considered three aims and applications of computational pharmacology: prediction of drug–target interactions; application to drug repurposing; and prediction of side effects or adverse drug reactions. The description of these applications was supported by a presentation of the methods to measure and quantify the pharmacological space and by a description of the main databases and tools used for data processing. Some algorithms were described with an emphasis on their performances and drawbacks. Alaimo *et al.* [34] focused on the algorithmic aspects of *in silico* repurposing approaches, describing the different classes of method [27], followed by the mathematical foundations of network-based inference methods. Using the DT-hybrid algorithm as an example, they discussed several current issues of *in silico* repurposing. Here, we present a global description of the key properties of the main classes of *in silico* repurposing method [24] to show that such methods are organized as workflows of three modules devoted to specific tasks, namely, data processing, *in silico* generation of putative candidates for repurposing and validation of the predictions. Furthermore, we emphasize that, in addition to their specific advantages and disadvantages, repurposing methods share three technical issues: the inability to predict drug–target interactions involving target or drug for which no interaction is known, the high dependence of the *in silico* methods regarding the model parameters; and the dependency of the methods on data sets that are biased with respect to different aspects. This broad synthesis should provide the reader with a comprehensive overview of the most effective approaches for designing a repurposing workflow while emphasizing the main pitfalls to be avoided.

To introduce the reader to the key steps of *in silico* repurposing, we begin with an example of a repurposing workflow. The following section then generalizes the main steps of the repurposing process to encompass the main approaches currently used. The gathering and processing of the data as well as current limitations inherent to their use that must be taken into account when using them are covered. We then describe algorithms of each category (structure-based, similarity-based, inference-based and ML-based techniques), along with their main features as well as their advantages and disadvantages. We conclude this section with a description of the procedure for assessing the algorithms and its predictions. We end our review with a conclusion summarizing the key technical challenges to be addressed and different approaches suggested to address them.

Identification of fasudil as an alternative autophagy enhancer

To introduce the main steps of the *in silico* repurposing procedure, the method MANTRA, presented in [35], is used as an example and its application for identifying fasudil as a new autophagy enhancer serves as a case study. MANTRA belongs to the class of similarity-based methods. These methods use the intuitive notion that similar compounds have similar properties. In the case of MANTRA, alternative drug candidates are found by analyzing similarities between transcriptional responses of various types of tissue to the addition of drugs under different experimental conditions. The first step for developing such method is to assemble the data of interest. Here, the Connectivity Map (cMap) [36], a repository that contains 6100 genome-wide expression profiles obtained by treatment of five different human cell lines at different dosages with a set of 1309 different molecules, was used. One would want to represent the information contained in these data by using a drug network (DN) whose nodes are the drugs, as shown in Fig. 1, Module 1. By default, these nodes are connected to each other with edges of arbitrary length. From a biological point of view, one would want to interpret the length of the edge between two drugs as a function of the similarity between them.

Thus, the second step is to build a metric, called the similarity measure, for quantifying in terms of pairwise distance the similarity of the transcriptional response between two drugs. The procedure requires converting the transcriptional profiles obtained for each drug and tissues into a set of pairwise distances between drugs. This procedure can be described as follows [Fig. 1, Module 2(A)]. First, the lists of genes are ranked according to their differential expression following drug treatment, from the most upregulated to the most downregulated. Then, the ranked lists of genes obtained by treating cells with the same drug are merged in one single list using a rank-aggregation algorithm [37]. This is a three-step algorithm using a measure of the distance between two ranked lists (Spearman's Footrule [38]), the Borda Merging Method to merge two or more ranked lists [39], or the Kruskal algorithm to obtain a single ranked list from a set of lists in a hierarchical way [39]. The output is a single prototype ranked list (PRL) of genes for each drug. The PRLs are then used to compute pairwise distances. The distance between drugs A and B is computed using an optimal signature (i.e. a subset of the most differentially expressed genes in the corresponding PRLs of the two drugs). To assess the degree of similarity between the PRLs, the randomness in the distribution of the genes of the optimal signature of drug A along the PRL of drug B, and vice versa, is quantified using gene set enrichment analysis (GSEA) [40]. The two enrichment scores (one for the optimal signature of drug A and one for the optimal signature of drug B) are combined to compute the distance between A and B. If the number of pairwise distances is very high, the empirical probability distribution of these data is used to estimate a significance threshold for the distance (the upper bound of the 5% quartile of the empirical pdf, as shown in Fig. 1). The network is interpreted as follows. Drugs closely connected to, or neighbors of, another drug induce similar transcriptional responses and are assumed to share common mode of action (MoA). This interpretation is confirmed by investigating the topology of the network [41]. Indeed, gene ontology (GO) fuzzy-enrichment analysis of communities identified using the affinity propagation algorithm [42] confirmed that compounds belonging to the same community share similar MoA [Fig. 1, Module 2(B)]. Furthermore, drugs of a given community share similar ATC codes and common target genes.

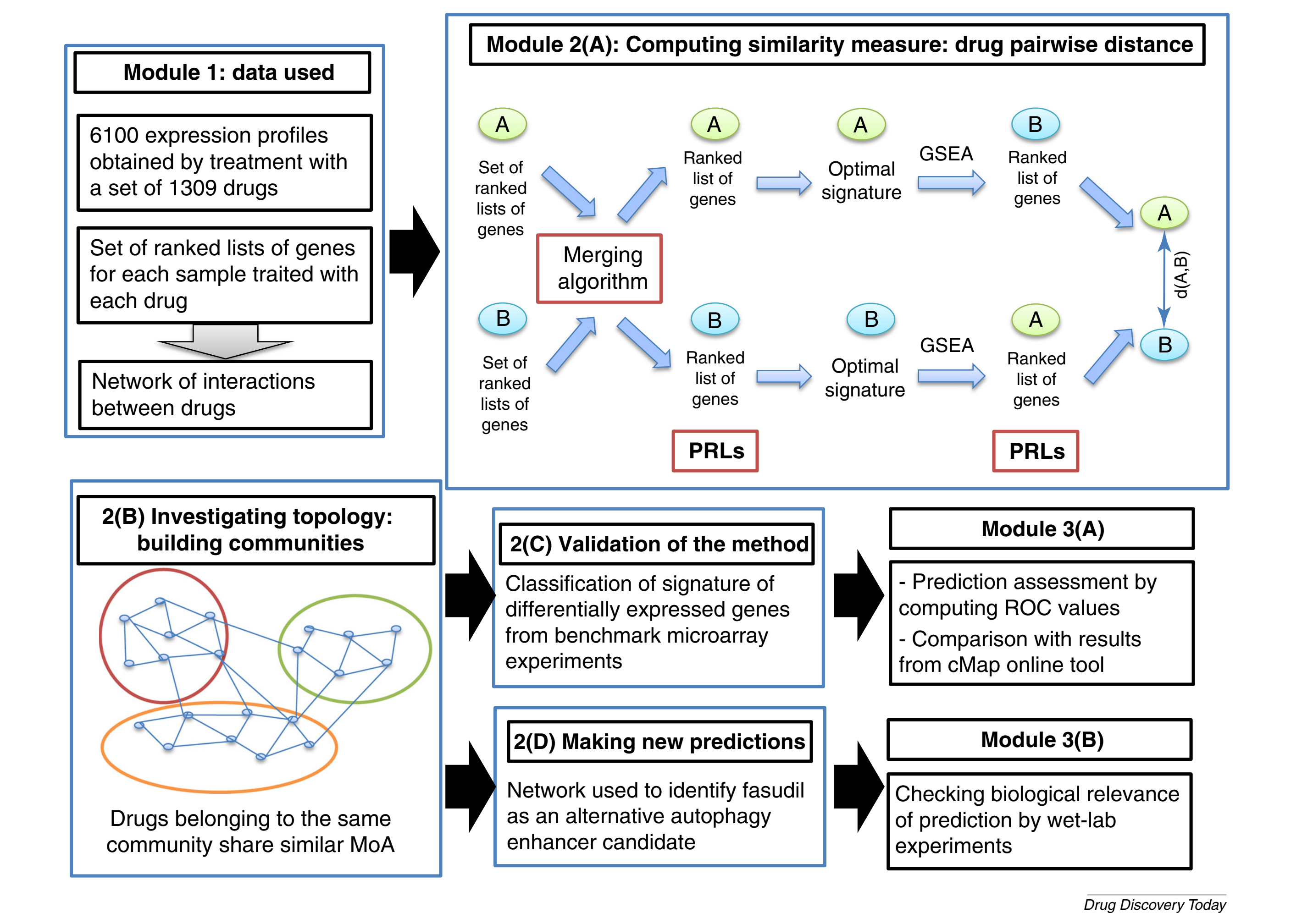

Module 1: data used
6100 expression profiles obtained by treatment with a set of 1309 drugs
Set of ranked lists of genes for each sample traited with each drug
Network of interactions between drugs
Module 2(A): Computing similarity measure: drug pairwise distance
A
Set of ranked lists of genes
Merging algorithm
A
Ranked list of genes
A
Optimal signature
GSEA
B
Ranked list of genes
A
d(A,B)
B
B
Set of ranked lists of genes
B
Ranked list of genes
B
Optimal signature
GSEA
A
Ranked list of genes
PRLs
PRLs
2(B) Investigating topology: building communities
Drugs belonging to the same community share similar MoA
2(C) Validation of the method
Classification of signature of differentially expressed genes from benchmark microarray experiments
Module 3(A)
- Prediction assessment by computing ROC values
- Comparison with results from cMap online tool
2(D) Making new predictions
Network used to identify fasudil as an alternative autophagy enhancer candidate
Module 3(B)
Checking biological relevance of prediction by wet-lab experiments
Drug Discovery Today

FIGURE 1

Flowchart of the MANTRA algorithm. Module 1: data sets are genome profiles of cell lines treated with different drugs. For each sample, a ranked list of genes is created. Module 2(A): ranked lists of genes are merged together and a single prototype ranked list (PRL) is associated with each drug. Then an optimal signature is built and pairwise distances with all other drugs are computed. Module 2(B): using a clustering algorithm, the community is computed. Module 2(C): benchmark microarray experiments are used to validate the method. Module 3(A): the method is validated on benchmark data sets by computing receiver operating curve (ROC) values. The method is then used to identify alternative enhancers of autophagy. Module 2(D): the network is used to identify alternative candidates. Module 3(B): the findings are validated by wet-lab experiments.

The final step of the development of the method is to quantify the reliability of its predictions. This is done by applying the method on benchmark data sets and by comparing the results with the ones of an already validated algorithm, in this case the cMap Online Tool [Fig. 1, Module 2(C)]. Concretely, a traditional signature of differentially expressed genes (a list of significant genes according to t test corrected with a false discovery rate) from microarray experiments is used to compare the classification results, by means of receiver operating curve (ROC) [43] analysis, with those obtained using the cMap online tool [Fig. 1, Module 3(A)]. cMap measures the signature-profile similarity by generating a signature from one profile and by using a nonparametric technique to assess the nonrandom distribution of these signatures in another ranked profile. The output is a list of drugs connected to each of the input signatures. The drugs that were predicted to be negatively connected to the input signature are filtered out, and each of the remaining drugs is considered a true positive if it belonged to at least one of four different reference golden standard sets. The reference sets included the counterpart of the tested drugs already present in the cMap. The drugs included in these sets are all those known to have the same MoA as the tested drugs. Overall, the DN approach performed comparably and sometimes better than the cMap classic online tool. The percentage of cases in which the first neighbor of a tested compound in the DN was a true positive was equal to 89% for the average distance. This value increases to 100% in cases where there is at least a true positive among the first two neighbors of each tested compound, for both the distances.

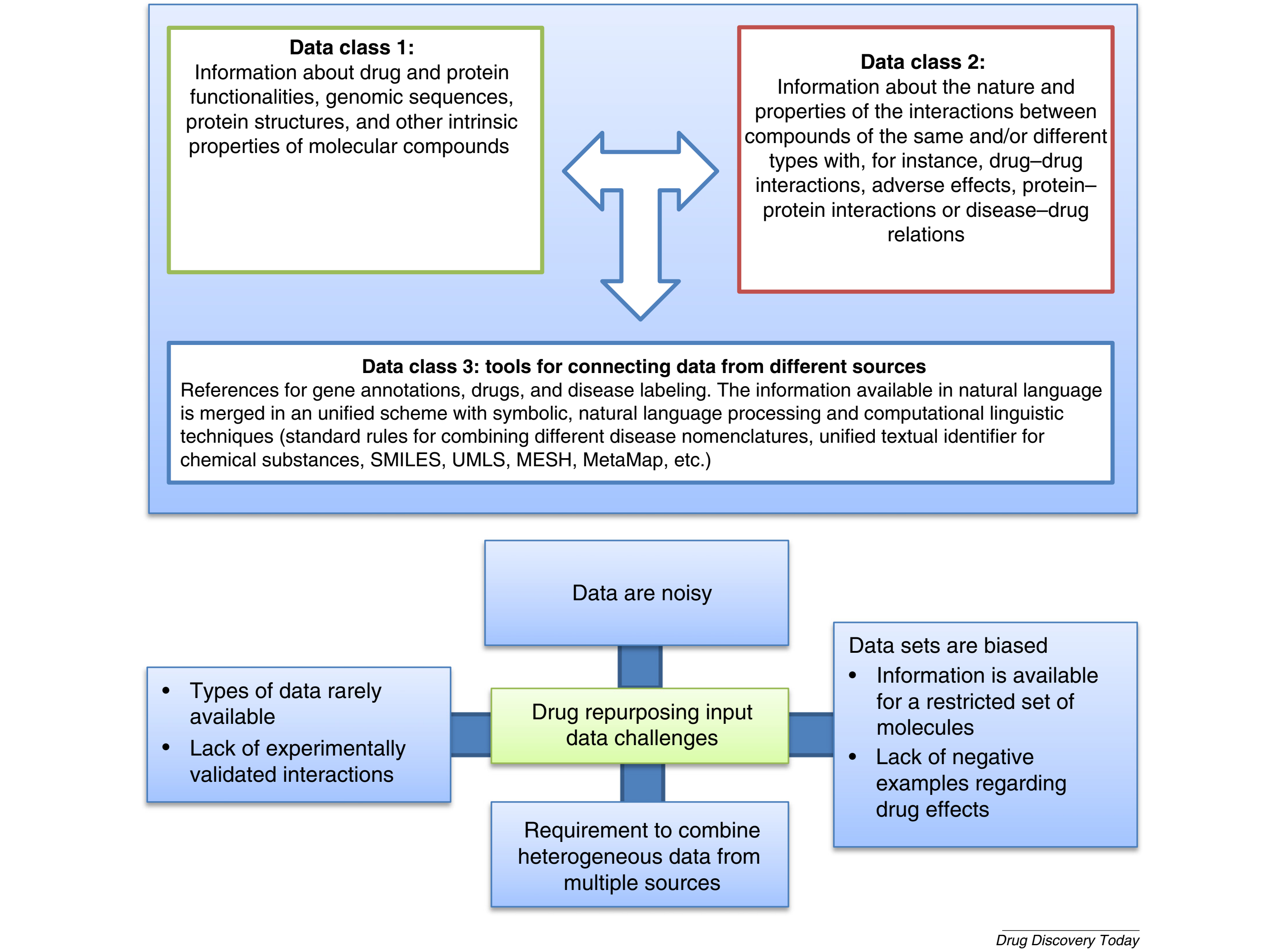
The MANTRA algorithm and its associated DN have been used on different case studies [35], including finding alternative drug candidates that could enhance autophagy. In practice, the DN was screened for drugs similar to 2-deoxy-D-glucose (2DOG), a molecule with the ability to induce autophagy [44]. 2DOG was found in a community with other molecules including, in increasing order of distance, fasudil, sodium-phenylbutirate, tamoxifen, arachidonyltrifluoromethane and novobiocin. Fasudil was the closest drug to 2DOG, whereas tamoxifen is another known autophagy inducer [45]. A supplementary analysis was performed by analyzing the distances of 2DOG from the other compounds in the DN independently of the community they belonged to. Again, in order of similarity, fasudil appeared to be the closest compound to 2DOG and, therefore, could be a suitable candidate as new autophagy enhancer. To test the validity of this hypothesis, the effect of fasudil on the induction of the autophagic pathway was experimentally tested by evaluating the LC3-II levels in wild-type human fibroblasts treated with fasudil, and other experiments using HeLa cells confirmed the findings [Fig. 1, Module 3(B)]. The fact that fasudil has the ability to enhance autophagy was not previously known. Thus, this example illustrates that drug repurposing can also lead to unexpected observations in drugs, contributing to our fundamental biological understanding.

Technical characteristics of *in silico* repurposing workflows

Despite the various methods and data types available, the different steps emphasized in the previous section are common to all *in silico* methods. Generally, the key modules of these methods are structured as shown in Fig. 2. Here, provide a description of the technical characteristics of each module.

Module 1: integration of data

Although the method presented above integrates only one type of data, many recent methods combine different types of data to improve their predictive power. Indeed, as emphasized by Hodos [21], the generation of more accurate and biologically relevant predictions relies on the capacity of the methods to capture as many characteristics of the systemic drug–target interaction


Data class 1: Information about drug and protein functionalities, genomic sequences, protein structures, and other intrinsic properties of molecular compounds

Data class 2: Information about the nature and properties of the interactions between compounds of the same and/or different types with, for instance, drug–drug interactions, adverse effects, protein–protein interactions or disease–drug relations

Data class 3: tools for connecting data from different sources References for gene annotations, drugs, and disease labeling. The information available in natural language is merged in an unified scheme with symbolic, natural language processing and computational linguistic techniques (standard rules for combining different disease nomenclatures, unified textual identifier for chemical substances, SMILES, UMLS, MESH, MetaMap, etc.)

Data are noisy

- Types of data rarely available
- Lack of experimentally validated interactions

Drug repurposing input data challenges

Data sets are biased
- Information is available for a restricted set of molecules
- Lack of negative examples regarding drug effects

Requirement to combine heterogeneous data from multiple sources

Drug Discovery Today

FIGURE 2

Modular organization of the drug repurposing pipeline. Module 1: assembly of the data sets. The nodes of the network of interactions represent the compounds, whereas the edges represent the interactions occurring between the compounds. Module 2: all algorithms that generate lists of potential candidates for repurposing are based on simple assumptions to define the similarities. The algorithms are classified into four categories: (i) 3D structure-based; (ii) similarity-based; (iii) inference-based; and (iv) ML-based methods. Module 3: using benchmark data sets, the ability of the algorithm to make reliable predictions is assessed by computing quality measures. Literature-based searches and alternative computational methods using text-mining techniques can be used to obtain partial confirmation of the *in silico* predictions (comparison against orthogonal databases, text-mining methods for mapping connected diseases onto known MeSH terms using the MeSH disease tree), but definitive validation of the biological relevance of the predictions must be done by wet-lab experiments.

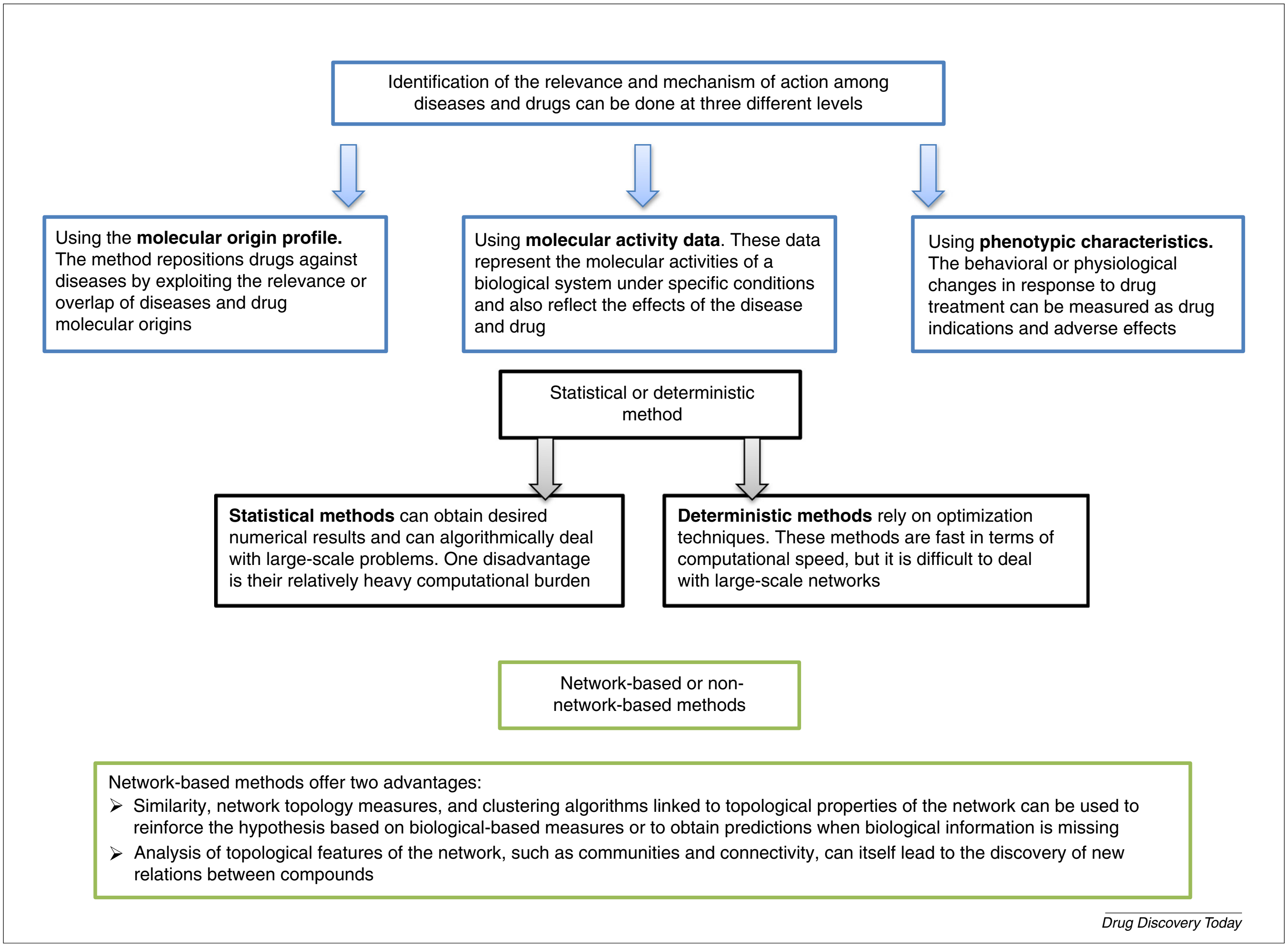

Identification of the relevance and mechanism of action among diseases and drugs can be done at three different levels
Using the **molecular origin profile.** The method repositions drugs against diseases by exploiting the relevance or overlap of diseases and drug molecular origins
Using **molecular activity data**. These data represent the molecular activities of a biological system under specific conditions and also reflect the effects of the disease and drug
Using **phenotypic characteristics.** The behavioral or physiological changes in response to drug treatment can be measured as drug indications and adverse effects
Statistical or deterministic method
Statistical methods can obtain desired numerical results and can algorithmically deal with large-scale problems. One disadvantage is their relatively heavy computational burden
Deterministic methods rely on optimization techniques. These methods are fast in terms of computational speed, but it is difficult to deal with large-scale networks
Network-based or non-network-based methods
Network-based methods offer two advantages:
➢ Similarity, network topology measures, and clustering algorithms linked to topological properties of the network can be used to reinforce the hypothesis based on biological-based measures or to obtain predictions when biological information is missing
➢ Analysis of topological features of the network, such as communities and connectivity, can itself lead to the discovery of new relations between compounds
Drug Discovery Today

FIGURE 3

Data types used for drug repurposing and technical issues. Data used for *in silico* drug repurposing are categorized into three main classes. Class 1 is used to gather information about intrinsic properties of molecular compounds. Class 2 gives information about the characteristics of the interactions between compounds. Class 3 provides tools for gathering all these data types together in an efficient way. Current data sets and databases have several limitations and bias. These limitations can complicate the use of *in silico* methods and introduce various biases into the network topology of known interactions.

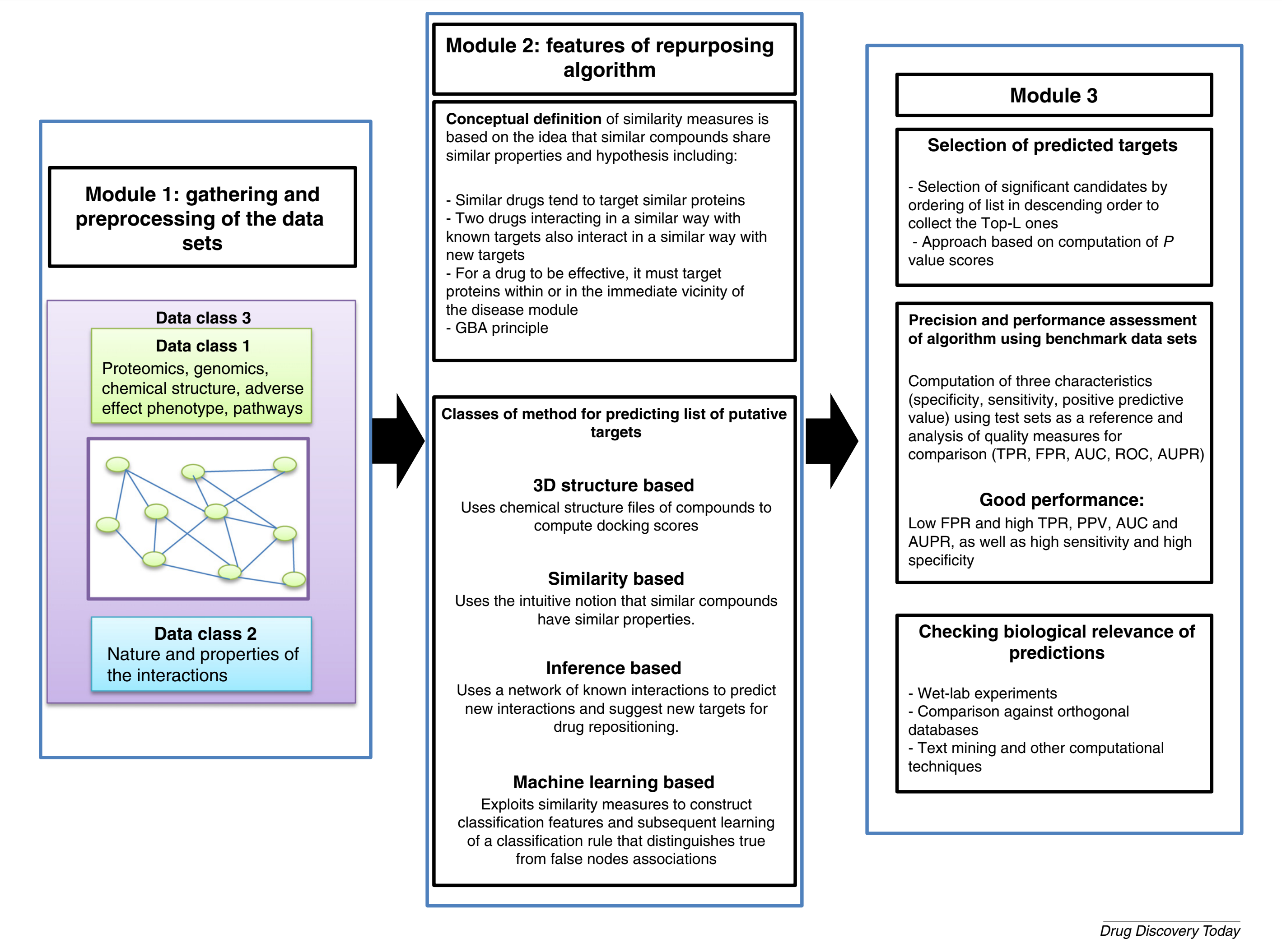
scheme as possible, essentially by combining data of different origins. As explained in Fig. 3 Top and Table S1 in the supplemental information online, although they cover a range of different types of biological information, databases can be classified into three main classes. The two first classes contain information about compounds and interaction properties, whereas the class three contains tools for integrating data of multiple origins within a unified nomenclature scheme. Using these web-based databases and software for data processing that is now available, it is possible to design sophisticated algorithms for investigating interactions not only between diseases and genes [30,46], diseases and drugs [4,47–51], drugs and genes [29,35,52], but also by combining interactions between diseases, drugs and proteins [53] or the associated genes [19,54–56]. Other methods look at the complex interplay between adverse effects of the drugs and targets [57] (Table S2 in the supplemental information online). However, although many methods perform their analysis at the molecular level, several approaches work at a larger scale by investigating how the activation of an entire biological pathway is affected by the addition of drugs. Indeed, signaling pathways form a network with many crosstalks that are responsible for drug adverse effects, cancer resistance [58], or common activation of pathways under a given perturbation [59]. Pathway-based approaches for drug repurposing provide as an output a prioritized list of drug-induced pathways that can be assembled as a database for further analysis. Examples of such drug-induced pathways database are presented in [60,61], while, in [62], an inference-based drug–target pathway prediction method is implemented.

However, as summarized in Fig. 3, current data sets and databases suffer from several biases and imperfections. Given that computational repurposing methods are dependent on the availability and quality of these data, many of these technical imperfections affect the validation process or intervene during the repurposing process and can reduce the ability of a method to elaborate a prioritized list of putative targets. Methods have been suggested to correctly introduce the topology of networks of known interactions [63] or to take into account the fact that most data used as an input usually contain not only many reliable positive examples (i.e. a drug is effective against a disease), but also many less high-confidence negative examples [64]. Other methods have been suggested to address specific issues ([65–67]; reviewed in [68]).

Module 2: algorithms for identifying candidates for repurposing

Current algorithms are classified into four categories: 3D structure-based, similarity-based, network inference-based and ML-based methods [24]. In addition to this classification, repurposing methods are characterized by three generic properties, as described in Fig. 4. First, the level at which the interactions between the compounds are considered. Second, the type of computational approach used (i.e. stochastic or deterministic). Third, the method can be network based or not depending on whether it explicitly uses topology properties to gain additional information about the interactions.

Although these classifications hold for many algorithms, a direction of research for improving the efficiency of these methods is to combine features of different algorithms, leading to the implementation of more complex hybrid methods [21,34]. Nevertheless, a common feature of these algorithms is that they rely on simple assumptions to define similarity measures that are used as quantitative metrics to identify alternative candidates and targets. As an output, these algorithms provide a list of candidates matching a set of predefined criteria. Although a straightforward way for selecting the most significant candidates is to order them in descending order and to collect the Top-L ones, a more objective approach based on the computation of *P* value scores is preferable


Module 1: gathering and preprocessing of the data sets
Data class 3
Data class 1
Proteomics, genomics, chemical structure, adverse effect phenotype, pathways
Data class 2
Nature and properties of the interactions
Module 2: features of repurposing algorithm
Module 3

Conceptual definition of similarity measures is based on the idea that similar compounds share similar properties and hypothesis including:

- Similar drugs tend to target similar proteins
- Two drugs interacting in a similar way with known targets also interact in a similar way with new targets
- For a drug to be effective, it must target proteins within or in the immediate vicinity of the disease module
- GBA principle

Classes of method for predicting list of putative targets

3D structure based
Uses chemical structure files of compounds to compute docking scores

Similarity based
Uses the intuitive notion that similar compounds have similar properties.

Inference based
Uses a network of known interactions to predict new interactions and suggest new targets for drug repositioning.

Machine learning based
Exploits similarity measures to construct classification features and subsequent learning of a classification rule that distinguishes true from false nodes associations

Selection of predicted targets

- Selection of significant candidates by ordering of list in descending order to collect the Top-L ones
- Approach based on computation of *P* value scores

Precision and performance assessment of algorithm using benchmark data sets

Computation of three characteristics (specificity, sensitivity, positive predictive value) using test sets as a reference and analysis of quality measures for comparison (TPR, FPR, AUC, ROC, AUPR)

Good performance:
Low FPR and high TPR, PPV, AUC and AUPR, as well as high sensitivity and high specificity

Checking biological relevance of predictions

- Wet-lab experiments
- Comparison against orthogonal databases
- Text mining and other computational techniques

Drug Discovery Today

FIGURE 4

Three generic characteristics of drug-repurposing methods. Analyzing the interactions between drugs and diseases is done at different levels. The repurposing task is performed through the screening of different types of data using either one or several similarity measures (chemical similarity, molecular activity similarity, or adverse effect similarity), molecular docking, shared molecular pathology and/or topological properties of the network of interactions. The algorithmic method is either deterministic or stochastic. In the case of large-scale projects, a stochastic approach is favored because it can handle large data sets more easily despite a more important computational requirement. Algorithms are divided between non-network based and network-based methods. 'Network-based' methods use the information embedded inside the network topology to discover unknown drug–target relations. By contrast, the so-called 'non-network'-based methods cannot use topology-based similarities and, thus, are more dependent on the quality and accessibility of experimental data.

[34]. This approach requires the calculation of a supplementary similarity taking into account the function of the targets, for instance using ontological terms. This similarity is used to build correlation measures between subsets of targets, and evaluate, for each drug, which subset of predicted targets has a similarity unexpectedly high correlation with respect to the validated targets. The *P* value is used to provide a quality score for the association between predicted and validated targets of a single drug.

3D structure-based methods

3D structure-based methods make predictions of interactions by mining, for example, the chemical–protein interactome. These methods use the chemical structure files of the compounds to compute docking scores [28,69,70]. Hence, in [70], a docking program was used to calculate the binding energy between an uploaded molecule and other library drugs. A second algorithm is then used that utilizes the docking scores to compute association scores between the uploaded molecule and each library drug. An advantage of these methods is that the interaction can be analyzed with respect to the structural properties. Nevertheless, docking algorithms are computationally expansive and rely on structural files, which are not easily available.

Similarity-based methods

In addition to the approach described above, many similarity measures have been implemented using biological, chemical, or topological properties of the targets, drugs and known interactions. Performance and prediction power vary according to the similarities used and, generally, the accuracy of similarity-based methods improves with the amount of data available. However, current results show that not all similarity measures are equal regarding the type of information they have access to. For instance, topology- or network-based similarities do not give information regarding the drug MoA. For that reason, algorithms combining different similarity measures are advantageous, although such methods require the use of different data types. In [49], a disease–disease, drug–drug and disease–drug network was assembled by matching molecular profiles of disease and drug expression profiles. Two methods are used to compute the similarity for the pairs of genomic profiles. The first is based on correlations that measure the profile–profile similarity by calculating the Pearson correlation of the cyber-T *t*-statistic values from two profiles. The second method is based on the concept of enrichment and follows the procedure described in [36]. In [29], a combination of two similarity measures was implemented: (i) a chemical similarity measure based on the relations between terms related to the drugs annotated with distinct but closely related terms; and (ii) a phenotypic adverse effect similarity using the observation that there is a correlation between adverse effect similarity and the likelihood that two drugs share a protein target. Both similarities are applied to infer common target between two drugs. Results obtained showed that the two methods combined are more sensitive than when applied separately. In [71], a bipartite network of drugs and pharmaceutical compounds was built and a statistics-based chemoinformatics approach was developed to predict new off-targets. The core of the algorithm, the similarity ensemble approach (SEA) [72,73], relies on the chemical similarities between drugs and targets defined by its ligands to compare targets by the similarity of the ligands that bind to them, expressed as expectation values. Newly predicted off-targets are assumed to have a biological relevance if they meet at least one of the three following criteria: (i) the new targets contribute to the primary activity of the drug; (ii) they mediate drug adverse effects; or (iii) they are unrelated by sequence, structure and function to the canonical targets. The network-based method developed in [19] is based on a new proximity measure that combines six different topological measures and uses topological structures called 'disease modules'. A disease module is formed by genes associated with a given disease [46]. The authors hypothesized that a drug is effective again a disease if it targets proteins in the close vicinity of the related disease module. The proximity measure performs better than six of the most common similarities. Furthermore, this method is capable of taking into account the elevated number of interactions of targets and, as a result, it is not biased regarding either the number of targets a drug has or their degrees; however, this improvement requires access to disease genes, drug targets and drug-diseased annotations.

Inference-based methods

Inference-based methods use *a priori* knowledge about known interactions, referred to as the 'training set', to predict new interactions and suggest new targets for repurposing. In [47], two inference methods based solely on topology measures were applied to predict drug–disease associations. Following the work of Zhou *et al.* [74], the problem is formulated as recommending diseases for a drug by mining data on the properties of a drug–disease bipartite network of experimentally verified drug–disease associations. In [52], following a methodology derived from network theory [74], three methods based on different similarity measures were implemented: (i) a network-based similarity; (ii) a drug-based approach using the hypothesis that, if a drug interacts with a target, then other drugs similar to the drug will be recommended to the target; and (iii) a target-based method, whose basic idea is that, if a drug interacts with a target, then the drug will be recommended to other targets with similar sequences to the target. The results obtained give an advantage to the network similarity-based algorithm. In [62], a protein complex-based Bayesian factor analysis was developed that modeled the chemical–genetic profiles using protein complexes to infer, by Bayesian inference, the MoA of drugs on protein complexes. The DT-hybrid algorithm [75] improves the method of Cheng [52] by using a similarity matrix to directly plug the domain-dependent biological knowledge into the model. The similarity matrix is obtained as a linear combination of a structure similarity matrix and a target similarity matrix. This method performs better for the prediction of biologically significant interactions and outperforms the methods presented in [52,74] in recovering deleted links. Nevertheless, although the additional biological knowledge increases the performance and improves the numerical precision, the supplementary parameter introduced in the similarity matrix leads to practical complications because its optimal value depends on the characteristics of the data sets and an *a priori* analysis is required for its selection.

ML-based algorithms

Finally, ML-based algorithms exploit similarity measures to construct classification features and subsequent learning of a classification rule that distinguishes true from false node associations. Several ML methods have been published and, on average, their performance and prediction power is improved by integrating additional algorithmic approaches for dealing with the three challenges. As for the other category of complex algorithms, their design varies depending on the data sets used. For example, new targets are predicted in [76] using multiple-category Bayesian models trained on chemogenomics databases, whereas, in [77], the authors used a ML method to investigate the extent to which chemical features of small molecules can reliably be associated with significant changes in gene expression. A review of the network-based ML models and their use for the prediction of compound–target interactions both in target-based and phenotype-based drug discovery applications has been published elsewhere [26]. PREDICT is an example of a ML-based method for predicting novel associations between drugs and diseases [48]. Using a set of known drug–disease associations constructed from multiple sources as a training set, the algorithm ranks additional drug–disease associations based on their similarity to the known associations. For this step, five drug–drug similarity measures and two types of disease–disease similarity measure are constructed. The association scores calculated on pairs of these similarity measures are used by a logistic regression algorithm to construct classification features and subsequent learning of a classification rule that helps to identify new drug–disease associations. An advantage of this method compared with others presented in [78] is that it can be applied to novel molecules with no indication information. However, it requires experimentally verified negative drug–disease associations to proceed. In [79], Yamanishi *et al.* investigated new interactions for four different drug–target classes, using the Kernel Regression Method (KRM).

In this supervised learning method, the biological information is integrated within a 'pharmacological space' by combining chemical (drugs) and genomic (targets) spaces. A drug–target interaction network is constructed for each protein class using a bipartite graph representation. Then, a regression model is developed between the combined chemical structure and amino acid sequence-based similarity spaces and the pharmacological space. The putative drugs and targets are mapped into the pharmacological space using this regression model and new interactions are predicted by connecting drugs and targets that are closer than a threshold in the pharmacological space. More recently, Dai *et al.* [51] suggested a matrix factorization model taking advantage of the richness of interaction data to detect potential drug–disease associations rather than following, similar to many others [35,48,80,81], the usual approach of computing and matching drug and disease profiles. The method works in two steps. First, a gene interaction network is constructed and topology information is extracted from this genomic space by computing a gene closeness metric. Using this information, low-rank feature vectors are retrieved from the gene interaction network by using eigenvalue decomposition. Then, feature vectors of drugs and diseases are obtained from drug–gene interactions and disease–gene interactions, respectively. Second, the matrix factorization model is generated and used to approximate known associations between drugs and diseases. The model provides an estimate of the possibility of association between one given drug and disease. After this training phase, the model can be used to predict novel drug indications. Although the incorporation of topology information allows this method to perform better than others [82,83] when association information of drugs or diseases is rare, it remains limited by the availability of drug–gene interactions and disease–gene interactions that are required for an accurate measurement of feature vectors.

Finally, a specific class of methods, called bipartite local models (BLMs), using similarity measures in the forms of kernels, has been developed [84]. The advantage of these methods is that they allow the incorporation of multiple sources of information for performing predictions [85]. The BLM can be summarized as follows [86]. The detection of drug–target interactions is done first by constructing a training comprising two classes: (i) all the known targets of the drug under investigation except the target of interest; and (ii) the targets for which no interaction with the drug is known *a priori*. Second, using the available genomic kernel for the targets, a support vector machine (SVM) that discriminates between the two classes is constructed. This model is used to predict the label of the target and to determine whether the considered drug–target pair shares an interaction. Using the chemical structure kernel, the procedure is applied with the roles of drugs and targets reversed and the two results are combined. BLM has also been investigated by van Laarhoven *et al.* [87]. His implementation differs in that the Gaussian kernel was constructed solely on the use of the topology information and by using regularized least squares (RLS) classifiers rather than SVM. The method works as follows. A bipartite network of drugs and targets constructed from known drug–target interactions is used to generate the interaction profiles from which a Gaussian interaction profile (GIP) kernel is constructed. The predictive power is improved by combining the GIP kernel with a kernel representation of chemical structure similarity between compounds and sequence similarity between proteins. These interaction profiles are used as feature vectors for two types of RLS classifier. It was concluded that the method provides more accurate results when the GIP kernel is combined with the chemical and genomic kernels, in particular for small data sets. Furthermore, it was noted that the sequence similarity for targets is more informative than the chemical similarity for drugs. Nevertheless, despite these promising results, the authors pointed out that the method is sensitive to inherent biases contained in the training data and that it can only be applied to detect new interactions for a target or a drug for which at least one interaction is already known. Interestingly, Mei and coauthors have released a method called BLM-NII [84], which combines a BLM with a procedure called 'neighbor-based interaction profile inferring' (NII), designed to tackle the inability to deliver predictions for drug and target that are new, a technical issue called here the 'new candidate problem' of BLM. The NII procedure extends the classifier to incorporate the capacity of learning from neighbors into the original BLM method. Comparisons with previous methods demonstrate the capacity of BLM-NII to predict interactions between new drug candidates and new target candidates with high reliability.

Module 3: validation of the predictions

Once implemented, an algorithm for drug repurposing should undergo a procedure to assess its ability to make relevant and accurate predictions (Fig. 2, Module 3). This procedure requires benchmark data sets to which the algorithm is applied. These are obtained from reliable sources, such as clinical trials and DrugBank, or specific case studies specifically designed for that purpose. The accuracy of the results is measured using a set of metrics designed to assess the reliability and accuracy of the predictions. In addition to the ROC, other metrics and quality measures can be computed. A straightforward method is to compute the values of area under the ROC curve (AUC) [19,47,79,87]. However, the performance of the algorithm can also be evaluated by computing characteristics such as specificity, sensitivity and positive predictive value (PPV) [34,79]. Furthermore, the recall, which provides information on the capacity of the algorithm to find the real unknown interactions, and the precision, which indicates the ability to discern biologically relevant interactions from untrue ones, can also be computed to draw the precision-recall curve [34,88]; that is, the plot of the ratio of true positives among all positive predictions for each given recall rate. The area under this curve (AUPR) provides an assessment of how well predicted scores of true interactions are separated from predicted scores of true non-interactions [84,87]. In the case of methods such as inference-based and ML-based methods containing multiple parameters whose values must be fixed, the validation procedure includes a first step called 'training', during which the algorithm is used on a part of the benchmark data set to find the parameter values that optimize the algorithm performances. When the parameters are fixed, the validation itself, which aims to test the ability of the algorithm to generalize on different data sets using the same parameter setting, is performed using the remaining data sets [47,87]. Finally, when a new method is implemented or new features are added to an already existing one, it is worth comparing the performances of the new method with already established ones using identical benchmark data sets. This step enables us to understand at which extent and in which context the new method provides better predictions. When the validation gives satisfying results, the algorithm can be used for discovering new relations between drugs, diseases and candidates for drug repurposing.

Once potential candidates are identified, the biological significance of the finding must be assessed. A first literature search can be performed to find evidence supporting the computational predictions. This was the method chosen in [89] for assessing predictions suggesting that the antiasthma drug pranlukast has anticancer metastasis activity, and in [90] for the suggested repositioning of cardiovascular drugs to parasitic diseases and for checking the prediction that the cancer-related kinase PIK3CG is a novel target of resveratrol. However, we recommend that wet-lab experiments are performed to confirm the suitability of the candidates. Examples of successful validations include: repurposing for early- and late-stage non-small cell lung cancer [54]; identification of an application of a hypertension drug, benzthiazide, as a potential agent to induce lung cancer cell death [53]; prediction of the antiulcer drug cimetidine as a candidate therapeutic in the treatment of lung adenocarcinoma [50]; and repositioning of the anticonvulsant topiramate for inflammatory bowel disease [55]. Nevertheless, in some cases, the predictions are not followed by experimental validation and, thus, must be considered with caution. This was the case in [91] with the finding of potential candidates among hypotension-related drugs that could be used for lowering blood pressure and in [70] with the prediction of new drug–drug associations for rosiglitazone and the repositioning of antipsychotics as anti-infectives. If these first tests are successful, the candidates could go through different development stages and, ultimately, reach clinical trials.

Concluding remarks and future perspectives

The different classes of *in silico* repurposing algorithm are attractive approaches for identifying alternative candidates. Nevertheless, as summarized in Fig. 5, they face technical issues.

The first issue concerns the dependency of *in silico* repurposing procedures with respect to the availability and characteristics of data sets. Given the current technical limitations of data sets, one could conclude that methods that reduce the need for data sets should be more adapted, but the progress made in developing more efficient methods relies on the use of data of multiple sources. Given that the dependence of a method on several types of data can limit its use in a range of practical situations, it is important to combine data widely available with similarity measures that have better predictive power.

The second issue is that the most elaborated algorithms use parameters for which optimal values are not easy to establish. For example, in the case of standard gene enrichment-based methods, empirical findings suggest that drug signatures established with too few genes lead to lower specificity and sensitivity. Furthermore, performances vary depending on the method used for the selection of the genes. Investigations suggest that gene selections based on fold change in combination with a greater *P* value threshold are more reliable than those based on *P* value or fold change alone [49]. For ML methods, the learning rate has a marginal effect, whereas the regularization coefficient has an important influence [51]. Furthermore, obtaining the statistical significance of the retained candidates from the list of targets to identify true positive requires the calculation of *P* values whose cut-offs vary from one study to another, affecting the final result. To summarize, reducing the dependency of the methods on the parameters and the parameter dependency with respect to the characteristics of data sets, as well as developing a systematic approach to determining their optimal values, might help the systematic use of these methods. A suggested strategy is to test the model for different set of values and choose the optimal parameter values according to the performance of AUC or other quality measures [51].

Finally, standard repurposing algorithms are often limited for making predictions involving candidates for which no interaction is known [34,84] and existing methods must be adapted to overcome this limitation. For instance, the DT-hybrid method [75] is an improvement of an inference-based method and the BLM-NII method is an enhanced version of the BLM. In the case of algorithms based on topology similarity, adding other similarity measures can improve their predictive power [47,52].

Although efficient hybrid algorithms can be elaborated with a combination of different approaches or by integrating methods using different information [49] (Table 1), another direction of development relies on completely new computational approaches. For instance, DL methods could overcome several limitations encountered by the standard ML methods. Indeed, although recent developments with ML methods are promising, it is not obvious that they could address all the remaining issues. Thus, DL methods could be the next move for improving the efficiency of repurposing techniques, for instance, for integrating biomedical data, which are relatively small and complex. The modern DL techniques include powerful approaches with deep architecture, called deep neural networks (DNNs) that are applied

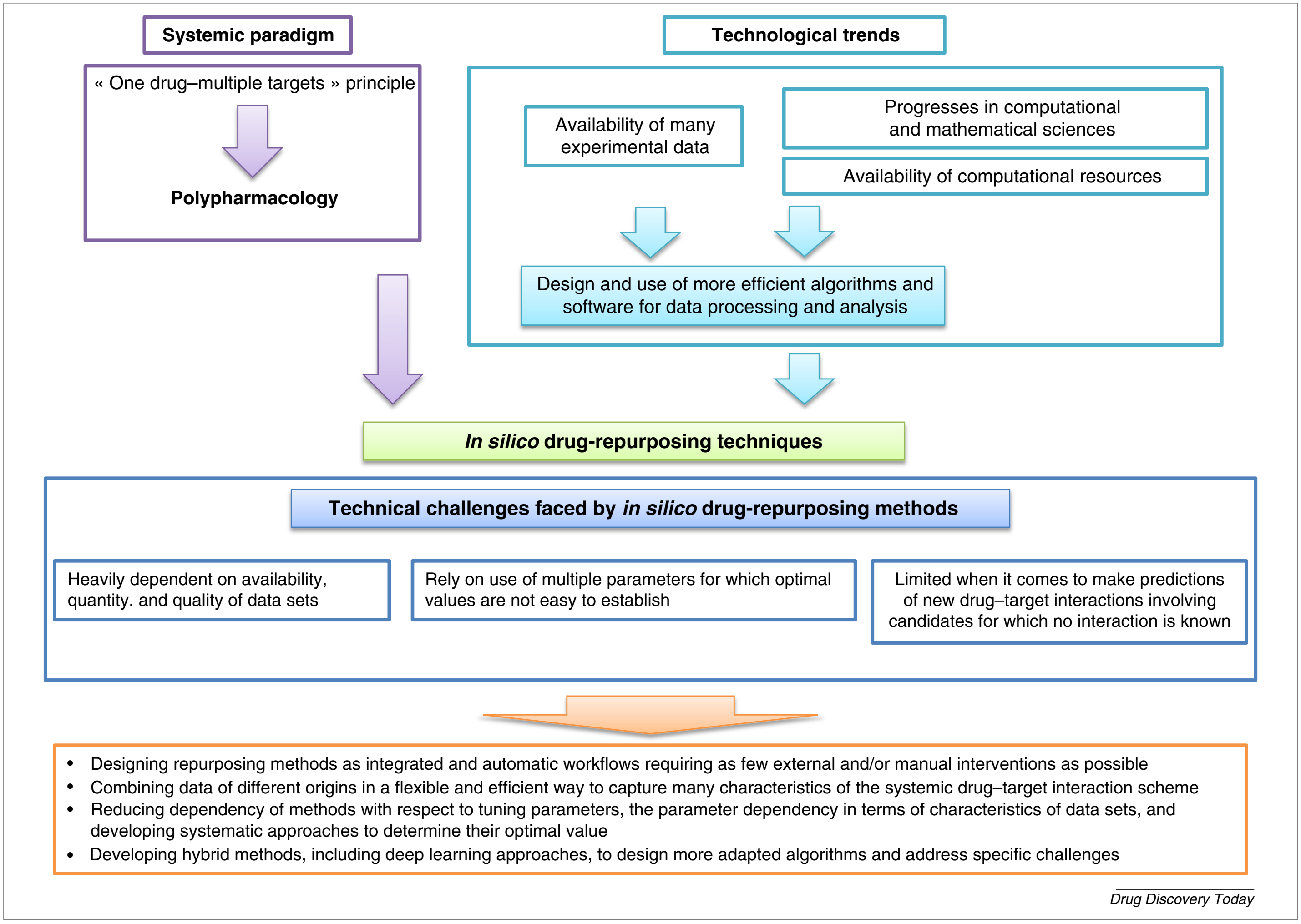

Systemic paradigm
« One drug–multiple targets » principle
Polypharmacology
Technological trends
Availability of many experimental data
Progresses in computational and mathematical sciences
Availability of computational resources
Design and use of more efficient algorithms and software for data processing and analysis
In silico drug-repurposing techniques
Technical challenges faced by *in silico* drug-repurposing methods
Heavily dependent on availability, quantity. and quality of data sets
Rely on use of multiple parameters for which optimal values are not easy to establish
Limited when it comes to make predictions of new drug–target interactions involving candidates for which no interaction is known
Designing repurposing methods as integrated and automatic workflows requiring as few external and/or manual interventions as possible
Combining data of different origins in a flexible and efficient way to capture many characteristics of the systemic drug–target interaction scheme
Reducing dependency of methods with respect to tuning parameters, the parameter dependency in terms of characteristics of data sets, and developing systematic approaches to determine their optimal value
Developing hybrid methods, including deep learning approaches, to design more adapted algorithms and address specific challenges
Drug Discovery Today

FIGURE 5

Foundation, technical challenges and directions of research for improving the drug-repurposing paradigm. The systemic paradigm and technological progress made in computational sciences are the cornerstones of drug-repurposing methods. Nevertheless, despite significant progress, the current algorithms still face three main technical challenges: (i) various technical limitations of the data sets can limit the predictive power; (ii) many sophisticated methods depend on free parameters whose fitting is tedious because it can depend on external factors that are not easy to control; and (3) algorithms are sometimes limited when it comes to make predictions for drugs or targets without any known interactions. Different solutions have been tested with more or less success and further possibilities offered by deep learning (DL) algorithms should allow significant progresses.

TABLE 1

The main characteristics and features of three of the most efficient methods currently available for *in silico* drug repurposing

Method type	Characteristics	Features	Resources	Refs
Similarity based	Uses a proximity measure combined with disease module identification on a network of drug–disease interactions	A representative example of a network-based method relying only on the use of a combination of topological measures. The proposed proximity measure outperforms other topology measures and the method is able to handle a large number of targets and interactions		[19]
Inference based	Uses a combination of structure similarity and target similarity matrices on a bipartite network of drug–target interactions	An example of how the inclusion, via drug and target similarities, of biological knowledge into the formalism of an inference-based method can improve the reliability, biological relevance and accuracy of the predictions. It illustrates the flexibility of the approach to combine various sources of information	R package DT-Hybrid-NBI	[75]
ML based	Uses a drug–target bipartite graph; interactions are deduced by training a classifier exploiting interaction information, and drug and target similarities; it is able to make predictions for drugs without known interactions	The latest improved version of the initial BLM. The addition of the algorithm NII allows the prediction of interactions between new drug candidates and new target candidates with high reliability	BLM-NII	[84]

for unlabeled and labeled data analysis, such as image, voice and language recognition [92]. They outperform ML methods, such as random forest or SVM, in training on quantitative structure–activity relation descriptors (QSAR) and for predicting various physical and chemical properties [93]. However, although DL methods could operate with several types of data for drug discovery and development, such as structural data, chemical descriptors, or transcriptomics data, and DNNs have been applied for modeling drug–target interactions using structural data [94], they are still underestimated in biomedical application [95]. This situation should evolve as new areas of applications emerge. For instance, it is now possible to predict the harmful potential of the compounds based on their raw structure using recursive or convolutional neural networks [96,97]. This is of particular interest in drug discovery for identifying well-designed and effective compounds that have toxic properties and DL-based approaches have proved to be effective for predicting such toxicity issues [98]. Furthermore, DNNs have already been applied for finding drug–target interactions using chemical structures and known interactions and promising results have been obtained [99,100]. However, DNNs come with technical issues. For example, the lack of theoretical foundation and the related lack of understanding of the method functionalities should be clarified. These issues are known for making the quality control and implementation of the results more complicated. Moreover, attempts were realized to address these issues with, for example, the TREPAN algorithms for extraction decision trees from hidden layers [101].

Conflict of interest

Q.V., P.M., A.M.A., A.A., K.L., I.O. and A.Z. are affiliated with Insilico Medicine, a company developing parametric and artificially intelligent drug discovery systems.

Acknowledgements

Authors thank the anonymous referees whose useful comments and suggestions helped improve this article.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.drudis.2016.09.019.

References

1 Tollman, P. *et al.* (2011) Identifying R&D outliers. *Nat. Rev. Drug Discov.* 10, 653–654

2 Scannell, J.W. *et al.* (2012) Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* 1, 191–200

3 Ashburn, T.T. and Thor, K.B. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673–683

4 Dudley, J.T. *et al.* (2011) Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.* 12, 303–311

5 Munos, B.H. and Chin, W.W. (2011) How to revive breakthrough innovation in the pharmaceutical industry. *Sci. Transl. Med.* 3, 89cm16

6 Mignani, S. *et al.* (2016) Why and how have drug discovery strategies in pharma changed? What are the new mindsets?. *Drug Discov. Today* 21, 239–249

7 Novac, N. (2013) Challenges and opportunities of drug repositioning. *Trends Pharmacol. Sci.* 34, 267–272

8 Mucke, H.A.M. (2016) Drug repurposing patent applications October–December 2015. *Assay Drug Dev. Technol.* 14, 308–312

9 Naylor, S. *et al.* (2015) Therapeutic drug repurposing, repositioning, and rescue: Part III. Market exclusivity using intellectual property and regulatory pathways. *Drug Discov. World* 62–69

10 Mucke, H.A.M. and Mucke, E. (2015) Sources and targets for drug repurposing: landscaping transitions in therapeutic space. *Drug Repurpos. Rescue Reposition.* 1, 22–27

11 Yarchoan, M. and Arnold, S.E. (2014) Repurposing diabetes drugs for brain insulin resistance in Alzheimer disease. *Diabetes* 63, 2253–2261

12 Mucke, H.A.M. (2016) Drug repurposing for vascular dementia: overview and current developments. *Future Neurol.* 11, 215–225

13 Snell, W.T. *et al.* (2016) Repurposing FDA-approved drugs for anti-aging therapies. *Biogerontology* http://dx.doi.org/10.1007/s10522-016-9660-x Published online August 2, 2016

14 Shumei, K. *et al.* (2015) Challenges and perspective of drug repurposing strategies in early phase clinical trials. *Oncoscience* 2, 576–580

15 Rutika, R. *et al.* (2016) Tumor deconstruction as a tool for advanced drug screening and repositioning. *Pharmacol. Res.* 111, 815–819

16 Heckman-Stoddard, B.M. *et al.* (2016) Repurposing old drugs to chemoprevention: the case of metformin. *Semin. Oncol.* 43, 123–133

17 Gilbert, D.C. (2016) Repurposing Vitamin D as an anticancer drug. *Clin. Oncol.* 28, 36–41

18 Shim, J.S. and Liu, J.O. (2014) Recent advances in drug repositioning for the discovery of new anticancer drugs. *Int. J. Biol. Sci.* 10, 654–663

19 Guney, E. *et al.* (2016) Network-based in silico drug efficacy screening. *Nat. Commun.* 7, 10331

20 Kaplan, W. *et al.* (2013) *Priority Medicines for Europe and the World Update 2013.* WHO

21 Hodos, R.A. *et al.* (2016) In silico methods for drug repurposing and pharmacology. *WIREs Syst. Biol. Med.* 8, 186–210

22 Wu, Z. *et al.* (2013) Network-based drug repositioning. *Mol. BioSyst.* 9, 1268–1281

23 Zou, J. *et al.* (2013) Advanced systems biology methods in drug discovery and translational biomedicine. *BioMed Res. Int.* 2013, 742835

24 Prathipati, P. and Mizuguchi, K. (2016) Systems biology approaches to a rational drug discovery paradigm. *Curr. Top. Med. Chem.* 16, 1009–1025

25 Lavecchia, A. and Cerchia, C. (2016) In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov. Today* 21, 288–298

26 Cichonska, A. *et al.* (2015) Identification of drug candidates and repurposing opportunities through compound–target interaction networks. *Expert Opin. Drug Discov.* 10, 1333–1345

27 Gonzalez-Daz, H. *et al.* (2011) Mind-best: web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of G3PDH protein from *Trichomonas gallinae*. *J. Proteome Res.* 10, 1698–1718

28 Xie, L. *et al.* (2011) Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. *PLoS Comput. Biol.* 7, e1002037

29 Campillos, M. *et al.* (2008) Drug target identification using side-effect similarity. *Science* 321, 263–266

30 Pacini, C. *et al.* (2013) DvD: an R/Cytoscape pipeline for drug repurposing using public repositories of gene expression data. *Bioinformatics* 29, 132–134

31 Jahchan, N.S. *et al.* (2013) A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. *Cancer Discov.* 3, 1364–1377

32 Putin, E. *et al.* (2016) Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging* 8, 1021–1033

33 Aliper, A. *et al.* (2016) Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* 13, 2524–2530

34 Alaimo, S. *et al.* (2016) Recommendation techniques for drug–target interaction prediction and drug repositioning. Data mining techniques for the life sciences. *Methods Mol. Biol.* 1415, 441–462

35 Iorio, F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. U. S. A.* 107, 14621–14626

36 Lamb, J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313, 1929–1935

37 Iorio, F. *et al.* (2009) Identifying network of drug mode of action by gene expression profiling. *J. Comput. Biol.* 16, 241–251

38 Diaconis, P. and Graham, R. (1977) Spearman's footrule as a measure of disarray. *J. R. Stat. Soc.* 39, 262–268

39 Lin, S. *et al.* (2010) Space oriented rank-based data integration. *Stat. Appl. Genet. Mol. Biol.* 9, Article 20

40 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102, 15545–15550

41 Newman, M.E. (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U. S. A.* 103, 8577–8582

42 Frey, B.J. and Dueck, D. (2007) Clustering by passing messages between data points. *Science* 315, 972–976

43 Gribskov, M. and Robinson, N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* 20, 25–33

44 Ravikumar, B. *et al.* (2003) Raised intracellular glucose concentrations reduce aggregation and cell death caused by mutant huntingtin exon 1 by decreasing mTOR phosphorylation and inducing autophagy. *Hum. Mol. Genet.* 12, 985–994

45 de Medina, P. *et al.* (2009) Tamoxifen and AEBS ligands induced apoptosis and autophagy in breast cancer cells through the stimulation of sterol accumulation. *Autophagy* 5, 1066–1067

46 Menche, J. *et al.* (2015) Uncovering disease–disease relationships through the incomplete interactome. *Science* 347, 1257601

47 Chen, H. *et al.* (2015) Network-based inference methods for drug repositioning. *Comput. Math. Methods Med.* 2015, 130620

48 Gottlieb, A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496

49 Hu, G. and Agarwal, P. (2009) Human disease–drug network based on genomic expression profiles. *PLoS ONE* 4, e6536

50 Sirota, M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* 3, 96ra77

51 Dai, W. *et al.* (2015) Matrix factorization-based prediction of novel drug indications by integrating genomic space. *Comput. Math. Methods Med.* 2015, 275045

52 Cheng, F. *et al.* (2011) Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* 8, e1002503

53 Lee, H.S. *et al.* (2012) Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Syst. Biol.* 6, 80

54 Huang, C.H. *et al.* (2014) Drug repositioning discovery for early- and late-stage non-small-cell lung cancer. *BioMed Res. Int.* 2014, 193817

55 Dudley, J.T. *et al.* (2011) Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl. Med.* 3 96ra76

56 Jin, L. *et al.* (2014) Drug-repurposing identified the combination of Trolox C and Cytisine for the treatment of type 2 diabetes. *J. Transl. Med.* 12, 153

57 Kuhn, M. *et al.* (2013) Systematic identification of proteins that elicit drug side effects. *Mol. Syst. Biol.* 9, 663

58 Dorel, M. *et al.* (2015) Network-based approaches for drug response prediction and targeted therapy development in cancer. *Biochem. Biophys. Res. Commun.* 464, 386–391

59 Smith, S.B. *et al.* (2012) Identification of common biological pathways and drug targets across multiple respiratory viruses based on human host gene expression analysis. *PLoS ONE* 7, e33174

60 Zeng, H. *et al.* (2015) Drug-Path: a database for drug-induced pathways. *Database* 2015, bav061

61 Pan, Y. *et al.* (2014) Pathway analysis for drug repositioning based on public database mining. *J. Chem. Inf. Model.* 54, 407–418

62 Han, S. and Kim, D. (2008) Inference of protein complex activities from chemical-genetic profile and its applications: predicting drug–target pathways. *PLoS Comput. Biol.* 4, e1000162

63 Yildirim, M.A. *et al.* (2007) Drug–target network. *Nat. Biotechnol.* 25, 1119–1126

64 Liu, H. *et al.* (2015) Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 31, i221–i229

65 Sanseau, P. *et al.* (2012) Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* 30, 317–320

66 Mullen, J. *et al.* (2016) An integrated data driven approach to drug repositioning using gene–disease associations. *PLOS ONE* 11, e0155811

67 Cheung, W.A. *et al.* (2013) Compensating for literature annotation bias when predicting novel drug–disease relationships through Medical Subject Heading Over representation Profile (MeSHOP) similarity. *BMC Med. Genomics* 6 (Suppl. 2), S3

68 Chen, B. and Butte, A.J. (2016) Leveraging big data to transform target selection and drug discovery. *Clin. Pharmacol. Ther.* 99, 285–297

69 Siragusa, L. *et al.* (2016) Comparing drug images and repurposing drugs with BioGPS and FLAPdock: the thymidylate synthase case. *ChemMedChem* 11, 1–15

70 Luo, H. *et al.* (2011) DRAR-CPI: a server for identifying drug repositioning potential and adverse drug reactions via the chemical-protein interactome. *Nucleic Acids Res.* 39, W492–W498

71 Keiser, M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature* 462, 175–181

72 Keiser, M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206

73 Hert, J. *et al.* (2008) Quantifying the relationships among drug classes. *J. Chem. Inf. Model.* 48, 755–765

74 Zhou, T. *et al.* (2010) Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. U. S. A.* 107, 4511–4515

75 Alaimo, S. *et al.* (2013) Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 29, 2004–2008

76 Nidhi *et al.* (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* 46, 1124–1133

77 Fernald, G.H. and Altman, R.B. (2013) Using molecular features of xenobiotics to predict hepatic gene expression response. *J. Chem. Inf. Model.* 53, 2765–2773

78 Hansen, N.T. *et al.* (2009) Generating genome-scale candidate gene lists for pharmacogenomics. *Clin. Pharmacol. Ther.* 86, 183–189

79 Yamanishi, Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240

80 Li, J. and Lu, Z. (2013) Pathway-based drug repositioning using causal inference. *BMC Bioinform.* 14 (Suppl. 16), S3

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

82 Rendle, S. (2012) Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.* 3, 1–22

83 Chen, T. *et al.* (2012) SVDFeature: a toolkit for feature-based collaborative filtering. *J. Mach. Learn. Res.* 13, 3619–3622

84 Mei, J.P. *et al.* (2013) Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 29, 238–245

85 Schlkopf, B. *et al.* eds (2004) *Kernel Methods in Computational Biology*, MIT Press

86 Bleakley, K. and Yamanishi, Y. (2009) Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403

87 van Laarhoven, T. *et al.* (2011) Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 27, 3036–3043

88 Raghavan, V.V. *et al.* (1989) A critical investigation of recall and precision as measures of retrieval system performance. *ACM Trans. Inf. Syst.* 7, 205–229

89 Zhao, S. and Li, S. (2012) A co-module approach for elucidating drug-disease associations and revealing their molecular basis. *Bioinformatics* 28, 955–961

90 Daminelli, S. *et al.* (2012) Drug repositioning through incomplete bi-cliques in an integrated drug–target–disease network. *Integr. Biol.* 4, 778–788

91 Wang, K. *et al.* (2016) Opportunities for web-based drug repositioning: searching for potential antihypertensive agents with hypotension adverse events. *J. Med. Internet Res.* 18, e76

92 Oquab, M. *et al.* (2014) Learning and transferring mid-level image representations using convolutional neural networks. *Proc. 2014 IEEE Conf. Computer Vision Pattern Recogni.* pp. 1717–1724

93 Ma, J. *et al.* (2015) Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* 55, 263–274

94 Wang, C. *et al.* (2014) Pairwise input neural network for target–ligand interaction prediction. *2014 IEEE Int. Conf. Bioinformatics Biomed.* pp. 67–70

95 Mamoshina, P. *et al.* (2016) Applications of deep learning in biomedicine. *Mol. Pharm.* 13, 1445–1454

96 Xu, Y. *et al.* (2015) Deep learning for drug-induced liver injury. *J. Chem. Inf. Model.* 55, 2085–2093

97 Hughes, T.B. *et al.* (2015) Modeling epoxidation of drug-like molecules with a deep machine learning network. *ACS Cent. Sci.* 1, 168–180

98 Mayr, A. *et al.* (2016) DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* Published online February 2, 2016 http://dx.doi.org/10.3389/fenvs.2015.00080

99 Ramsundar, B. *et al.* (2015) *Massively Multitask Networks for Drug Discovery*. arXiv:1502.02072

100 Dahl, G.E. *et al.* (2014) *Multi-task Neural Networks for QSAR Predictions*. arXiv:1406.1231

101 Karim, A. and Zhou, S. (2015) *X-TREPAN: A Multi-class Regression and Adapted Extraction of Comprehensible Decision Tree in Artificial Neural Networks*. arXiv:1508.07551