Artificial Intelligence in the Life Sciences 2 (2022) 100042

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Research Article

An unsupervised computational pipeline identifies potential repurposable drugs to treat Huntington's disease and multiple sclerosis

Luca Menestrina, Maurizio Recanatini*

Department of Pharmacy and Biotechnology, Alma Mater Studiorum - University of Bologna, Via Belmeloro 6, 40126 Bologna, Italy

ARTICLE INFO

Keywords:
Computational drug discovery
Drug repurposing
Network pharmacology
Huntington's disease
Multiple sclerosis

ABSTRACT

Drug repurposing consists in identifying additional uses for known drugs and, since these new findings are built on previous knowledge, it reduces both the length and the costs of the drug development. In this work, we assembled an automated computational pipeline for drug repurposing, integrating also a network-based analysis for screening the possible drug combinations. The selection of drugs relies both on their proximity to the disease on the protein-protein interactome and on their influence on the expression of disease-related genes. Combined therapies are then prioritized on the basis of the drugs' separation on the human interactome and the known drug-drug interactions. We eventually collected a number of molecules, and their plausible combinations, that could be proposed for the treatment of Huntington's disease and multiple sclerosis. Finally, this pipeline could potentially provide new suggestions also for other complex disorders.

Introduction

Discovering a new drug and bringing it to the market is a process both money and time consuming. Instead, relying on established drugs, computational drug repositioning offers a valuable alternative approach for providing promising treatments for disorders without a cure [1,2]. In recent years, a plethora of computational approaches to drug repurposing have been proposed and applied to a wide variety of therapeutic areas [3]. Most of such approaches rely either on machine learning or on the traditional methods of computational drug design, even though some conceptually innovative ideas have brought to the light the possibility of taking new paths towards the prediction of potentially repurposable drugs. One of such ideas is based on a system view and takes the human protein-protein interactome as a reference network to quantify the relatedness between drugs and diseases by calculating the distance between drug targets and disease-associated proteins. This distance has been proposed as a suitable metrics to measure the "proximity" between drugs and diseases [4]. Recently, leveraging on the concept of drug-disease proximity [5], novel drug indications for the treatment of cardiovascular diseases [5,6], cancers [7], COVID-19 [8], Alzheimer's disease [9] have been proposed, demonstrating how a network-based approach could successfully assist the selection of drugs to be repurposed.

In this work, we assembled an automated computational pipeline by integrating a recently developed scheme to screen repurposable drugs that combines a network-based technique with an analysis of biological and experimental data [10,11], with a strategy for filtering all the possible drug combinations [6]. Initially, the procedure estimates the proximity between the disease-related proteins and the drug targets on the protein-protein interactome, performing a first selection of candidates. Then, only those drugs that significantly influence the expression of disease-related genes are considered plausible for repurposing. Finally, evaluating the separation of these drugs' targets on the human interactome and taking into consideration the known drug-drug interactions, combined therapies are prioritized. The workflow of the procedure is schematically illustrated in Fig. 1. The entire process is automated in order to reduce human intervention, thus accelerating the whole procedure and limiting execution errors.

We applied this pipeline to Huntington's disease (HD) and multiple sclerosis (MS) because, despite the fact that they are both neurological disorders, their different nature could represent a challenge for our strategy, and the outcomes could give us insights into its methodological strengths and limitations. HD, is reported as a typical monogenic disease, even though many other genes are known to influence its progression [12], while for MS a single genetic cause has not been found yet, probably because many factors play an important role in the etiology. Indeed, MS fits well the definition of complex disease to be considered in the framework of network medicine. On the other hand, HD was included in our study in order to test the capabilities of the proposed method in a case where different clinical phenotypes might be related to a disease module eventually influenced by genetic modifiers leading to different pathophysiological states [12–14].

* Corresponding author.
E-mail address: maurizio.recanatini@unibo.it (M. Recanatini).

https://doi.org/10.1016/j.ailsci.2022.100042
Received 4 May 2022; Received in revised form 23 June 2022; Accepted 20 July 2022
Available online 27 July 2022

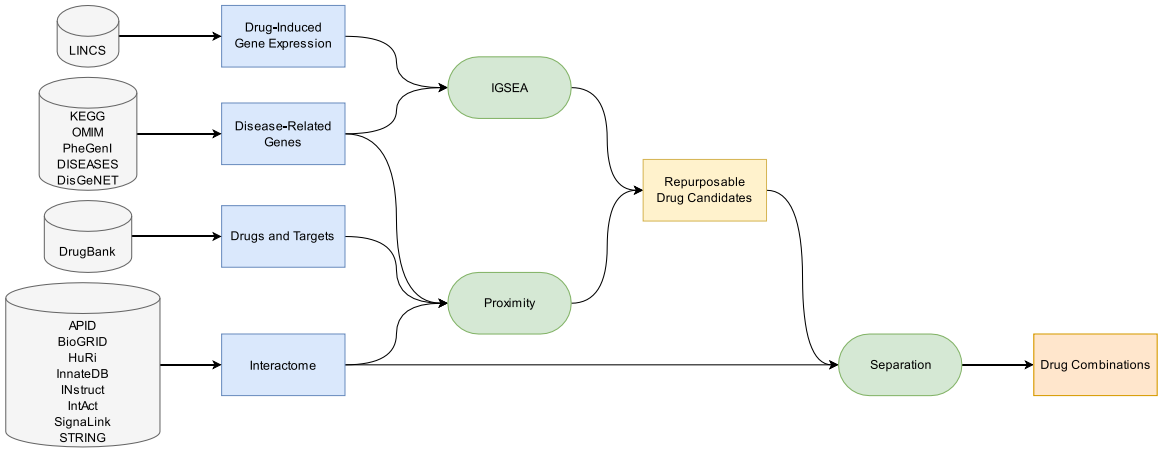

Fig. 1. Pipeline Flowchart. The flowchart shows the sources and the steps of the automated procedure to screen repurposable drug candidates and prioritize their combinations.

HD is the most common monogenic neurological disorder. The onset is typically in the early stage of adult life, and it is characterized by motor dysfunction, cognitive impairment, and neuropsychiatric features [12,15]. The autosomal dominant mutation that causes HD is located in the *HTT* gene, and it consists in a cytosine-adenosine-guanine trinucleotide repetition (CAG, encoding glutamine) leading to an over-expansion of the polyglutamine (polyQ) tail in the huntingtin protein. The mutated protein tends to aggregate and accumulate, forming inclusion bodies that have deleterious consequences for the neural cell. Both the inclusion bodies and the length of the CAG expansion are proven to play an important role in the development of the disease. The clearance of the first ones slows the HD progression, while the longer the CAG expansion, the earlier the disease may manifest [15]. The remaining uncertainty on the course of HD can be ascribed to other genetic differences in the genome of the patients [12,14].

MS is both the most frequent non-traumatic disabling disease in young adults [16] and the commonest demyelinating disease [17]. The etiology and the mechanism causing its worsening progression are still unclear, nevertheless it has been proven that a complex interplay of genetic and environmental factors is important [18,19]. The main known risk factors are smoking, childhood obesity, infection with the Epstein-Barr virus, and low vitamin D levels [19]. MS is generally viewed as a two-phases autoimmune disease, in which initially focal inflammatory processes cause a relapsing-remitting form of the disease, and subsequently demyelinating plaques (lesions resulted by the previous immune response) and oligodendrocyte damage lead to neurodegeneration and non-relapsing progressive course [17,19]. MS is commonly characterized by progressive spastic paraparesis, cognitive impairment, and sensory and cerebellar dysfunctions[19].

Both HD and MS are still lacking resolutive treatments [20,21], whose development needs a deeper knowledge of the underlying mechanisms [22]. To this aim, network-based models, as the ones we utilized in this study, could be adequate theoretical tools for investigating such multifactorial disorders. They would allow us to take into account the latent complex structure of these diseases without losing a comprehensive view [23]. Through the methodology presented here, we were able to collect a number of approved drugs and their plausible combinations that could be proposed for the treatment of HD and MS.

Materials and methods

The workflow of this study can be outlined in the following steps (Fig. 1): (1) collection of disease-related genes; (2) gene sets validation through enrichment analysis; (3) collection of drugs, targets, protein interaction data, and construction of protein-protein interactome; (4) computation of drug-disease proximity on the human protein-protein interactome; (5) estimation of drug-induced gene expression signature enrichment; (6) calculation of drug-drug separation on the human interactome. Except for the collection of the disease-related genes, each operation is performed by a Python 3 script, and the entire procedure is brought together and coordinated by a main file in the same programming language.

Collection of disease-related genes

For each of the two considered diseases, a set of related genes was retrieved from KEGG [24] (https://www.genome.jp/kegg/), OMIM [25] (https://www.omim.org/), PheGenI [26] (https://www.ncbi.nlm.nih.gov/gap/phegeni), DISEASES [27] (https://diseases.jensenlab.org/search), and DisGeNET [28] (https://www.disgenet.org/).

Briefly, for HD, 306 genes were retrieved from the KEGG Huntington Disease pathway “hsa05016”; 152 querying OMIM for “Huntington Disease”; 1 from the DISEASES database and 17 were those associated to “Huntington Disease” on DisGeNET and having an Evidence Index (https://www.disgenet.org/dbinfo#section36) of at least 0.95.

On the other hand, for MS, 160 genes were the result of querying OMIM for “Multiple Sclerosis”; 89 were collected from PheGenI with NHGRI (National Human Genome Research Institute) genome-wide association study as source and a p-value $< 1 \times 10^{-8}$; 5 were retrieved from the DISEASES database; 30 gathered from DisGeNET with the same conditions applied to HD.

The genes were mapped to official gene symbols taking advantage of the NCBI database and then combined.

Ontology (GO, HPO) enrichment analysis

Functional enrichment analysis is often employed to perform a preliminary analysis on an investigated gene set. Examining the Gene Ontology [29] (GO, http://geneontology.org/) and the Human Phenotype Ontology [30] (HPO, https://hpo.jax.org/) associations, we gained insights on biological processes, molecular functions, cellular components and phenotypes most frequently associated to those genes. We conducted the functional enrichment analysis using the Python library GOATOOLS [31] and considered significantly enriched only those terms with a false discovery rate (FDR, p-value corrected for multiple comparisons using the Benjamini-Hochberg procedure [32]) lower than 1×10^{-4}. We then looked at the first 20 terms ranked on the basis of their fold enrichment (computed as the ratio of the percentage of genes in the study set related to a specific term, divided by the corresponding percentage in the background, i.e., the entire human proteome).

Drugs and targets collection, gene expression profiles retrieval and protein-protein interactome construction

Drugs information was collected from DrugBank [33] (version 5.1.9). Only those molecules having at least one human protein as target are considered, obtaining 5 798 drugs and 2 755 corresponding targets.

Drug-induced gene expression profiles were retrieved from the Library of Integrated Network-based Cellular Signatures [34] (LINCS, profiles "GSE70138" and "GSE92742"), downloaded from Gene Expression Omnibus [35] (GEO, http://www.ncbi.nlm.nih.gov/geo). Due to the fact that we are inspecting neurological disorders, those signatures tested on neural cell lines (NEU, NPC, SHSY5Y) were examined for both diseases. Additionally, in order to consider disease-specific features, also muscular cell lines (SKB, SKL) were included for HD, and haematopoietic and lymphoid tissue cell lines (L60, JURKAT, NOMO1, PL21, SKM1, THP1, U937, WSUDLCL2) for MS. Furthermore, to guarantee maximum reliability of the results, only the data about the Best Inferred Genes (BING) in every dataset (drug signature) in these profiles was kept. The BING subset includes 978 landmark genes and 9 196 inferred genes, which are identified among the 12 328 genes in the L1000 assay by Subramanian et al. [36] evaluating the most reliable inference predictions.

Extensive interactions among proteins are a key factor in accomplishing many biological processes and functions. For this reason, we opted for a network-based approach to evaluate the correlation between drugs and diseases or drugs and other drugs. We built a human protein-protein interaction (PPI) network combining data from eight publicly available resources: Agile Protein Interactomes DataServer [37] (APID, http://cicblade.dep.usal.es:8080/APID/init.action), Biological General Repository for Interaction Datasets [38] (BioGRID, https://thebiogrid.org/), The Human Reference Interactome [39] (HuRI, http://www.interactome-atlas.org/), InnateDB [40] (https://www.innatedb.com/), INstruct [41] (http://instruct.yulab.org/), IntAct [42] (https://www.ebi.ac.uk/intact/home), SignaLink [43] (http://signalink.org/), and Search Tool for the Retrieval of Interacting Genes/Proteins [44] (STRING, https://string-db.org/). Supplementary Table 1 gives additional info about the interactions reported in the databases and the applied filters.

The retrieved interactions were then combined, obtaining a network (available in the Supplementary Information) consisting of 20 445 nodes (genes/proteins) and 1 125 173 edges (interactions). Consistency is granted by the fact that all listed proteins are mapped to official gene symbols taking advantage of the NCBI database. Since the protein-protein interactome is the supporting pillar of the whole procedure, we assessed its validity comparing the results of the entire analyses based on two other interactomes. The first rerun was carried out on the widely recognized interactome from Cheng et al. [5] (16 677 unique proteins and 243 603 experimentally confirmed protein–protein interactions). The second one was performed on a drastically restricted version of our own interactome (16 954 proteins and 246 080 interactions), in which only interactions from low throughput studies (listing less than 20 interactions) were included.

Network proximity

Proteins related to a specific disease are unlikely to be scattered throughout the interactome, rather, they tend to group together forming the so-called disease module [45]. The relationship between a drug and a disease could be estimated by means of an unsupervised and unbiased network-based approach [4], which quantifies the interplay of drug targets and disease-related genes measuring a network proximity. Here we used a recently modified version of such method [10] that includes a term (w) for taking into account the degree of the drug targets directly into the distance calculation. Given G, the set of disease-related genes; T, the set of drug targets; and $d(g,t)$, the shortest path length between nodes g ($g \in G$) and t ($t \in T$) in the human protein-protein interactome; the distance $d(G,T)$ between each drug and the disease was calculated as:

$$d(G,T) = \frac{1}{|T|}\sum_{t\in T}\min_{g\in G}(d(g,t)+w) \tag{1}$$

where w weights the targets based on their node degree in the interactome ($w = -\ln(D+1)$ if the target is related to the disease, $w = 0$ otherwise). D is the degree of the target in the PPI network.

Then, for each drug, the significance of its association to the investigated disease was assessed comparing the measured distance to that of a dummy reference distribution. This reference was obtained computing 10 000 times the distance ($d(G,R)$, defined by Eq. (1)) between the disease-related genes and randomly selected (from the human interactome) sets of proteins (R) matching the number of the drug targets. Since the degree of the drug targets is already taken into consideration in the distance calculation, the sampling of the randomly selected proteins is facilitated having to match only the number and not also the degree distribution of the drug targets. The mean $\mu_{d(G,R)}$ and standard deviation $\sigma_{d(G,R)}$ of the reference distribution were used to normalize the observed distance into a proximity value (z-score):

$$z(G,T) = \frac{d(G,T)-\mu_{d(G,R)}}{\sigma_{d(G,R)}} \tag{2}$$

Inverted gene set enrichment analysis

Starting from the hypothesis that effective drugs should be able to restore the healthy expression of genes deregulated by a disease, the drugs with signatures most enriched in disease-related genes should also be the most promising ones in treating such disease. In order to gain this knowledge, an Inverted Gene Set Enrichment Analysis [10] (IGSEA) on the datasets (drug signatures) of LINCS was performed, looking for the disease-related genes under study. For each analyzed dataset, the normalized enrichment score and the p-value (estimated comparing the enrichment score with those of a null distribution generated from 100 000 permutations) were computed for measuring the enrichment magnitude and its statistical significance, respectively. The resulted p-values were then corrected for multiple comparison using the Benjamini-Hochberg procedure [32], obtaining the FDR. If the dataset was significantly enriched (FDR < 0.25), the corresponding drug was considered a potential drug candidate.

Network separation

An important aspect in investigating drug combinations is to evaluate whether the two drug-target modules are overlapped (overlapping exposure) or separated (complementary exposure) on the human protein-protein interactome [6]. In the case of overlapping exposure, there is a higher similarity in chemical, biological, functional, and clinical profiles. The desired combinations, instead, are those with complementary exposure, both drugs being topologically and pharmacologically distinct. In the latter case, the two drugs synergistically cooperate in treating the disease, yet each one in its own way.

As we did for computing the drug-disease proximity, also for measuring drug-drug separation s_{AB} in drug combinations, we employed a network-based approach [6,45]:

$$s_{AB} = \langle d_{AB}\rangle - \frac{\langle d_{AA}\rangle + \langle d_{BB}\rangle}{2} \tag{3}$$

where A is the target module of one drug and B that of the other. Here, the mean shortest distances (calculated with Eq. (1) with the weight w fixed to 0) between the target modules of each drug ($\langle d_{AA}\rangle$ and $\langle d_{BB}\rangle$, computable only for drugs with at least two targets) are compared to the mean shortest distance between all possible A-B target pairs ($\langle d_{AB}\rangle$). When computing the distances between A-B target pairs, if a protein is targeted by both drugs, its distance is zero by definition. A drug combination exposure is deemed complementary if $s_{AB} \geq 0$, overlapping otherwise.

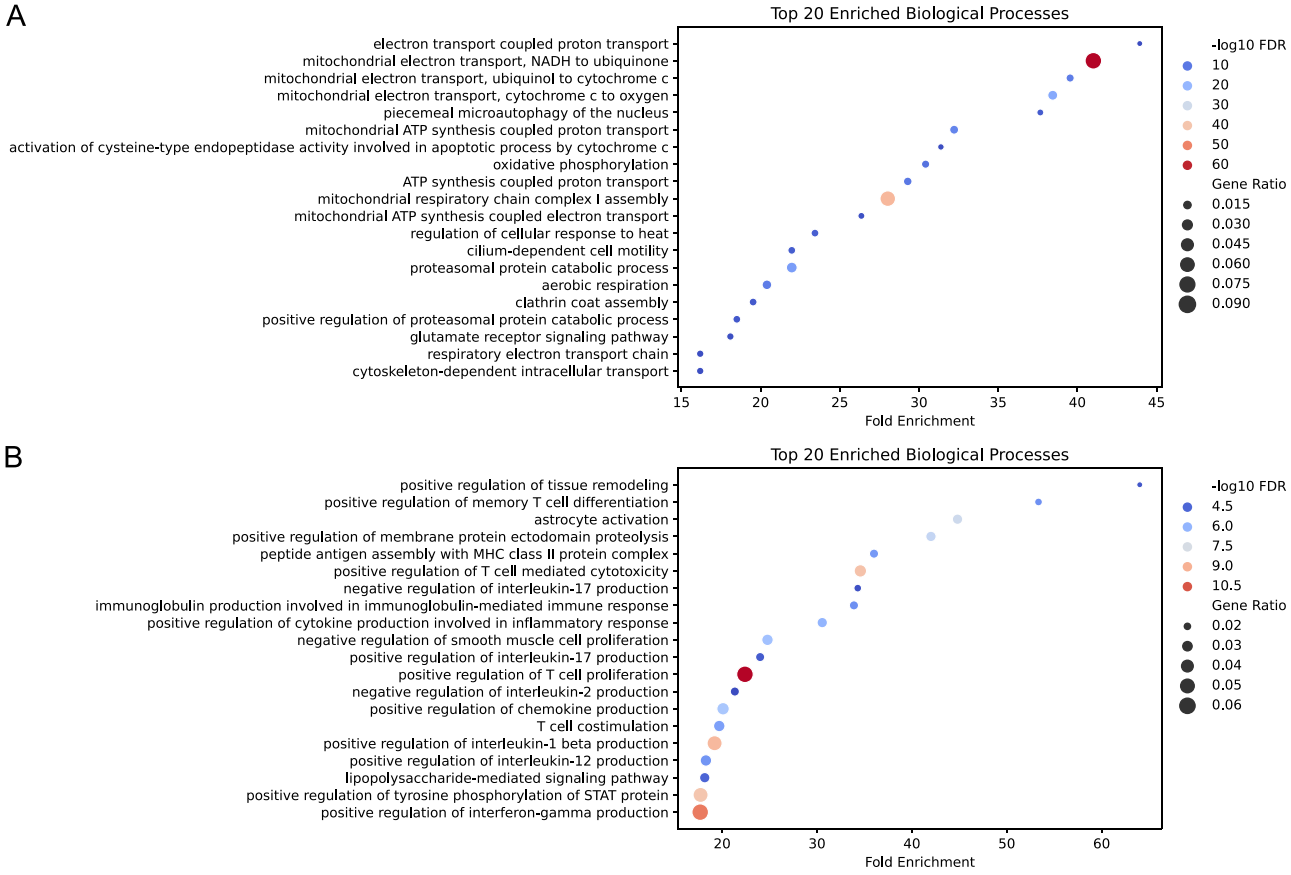

Fig. 2. Enriched Biological Processes. The bubbleplots display the top 20 most enriched Gene Ontology terms relative to biological processes for Huntington's disease (A) and multiple sclerosis (B). On the horizontal axis, the fold enrichment is shown. The color encodes the negative of the false discovery rate logarithm, and the size represents the gene ratio (computed as the ratio of the percentage of genes in the study set related to a specific term, divided by the corresponding percentage in the background, i.e., the entire human proteome).

Results

Computational framework

In this study, we automated the pipeline shown in Fig. 1 for screening repurposable drug candidates and prioritizing their combinations. In order to run, the script only requires the disease name, the disease-related genes, and the cell lines of interest as inputs. This procedure consists of: collecting, cleaning and organizing the source data (disease-related genes, drugs, targets, protein interactions, drug-induced gene expression signatures); identifying repurposable drug candidates evaluating both their proximity to the disease and their effect on the expression of the disease-related genes; screening the possible drug combinations on the basis of their relative exposure and known interactions. The output of the routine is a collection of tables (tab-separated values files) and plots, recording both intermediate and final results.

Compared to previous related works [10,11], such a systematic strategy should be more efficient and have an improved reproducibility thanks to the organization and standardization of both the overall study and results. Additionally, it takes a step forward since it evaluates also possible combined therapies.

The single steps and the outcomes of the application of the framework to HD and MS are presented and discussed in the following.

Disease-related genes collection and validation

We gathered the disease-related genes as described in the Methods section: this resulted in 451 and 217 genes associated to HD and MS, respectively. In order to evaluate whether these genes were representative of the investigated diseases, we performed an enrichment analysis on GO and HPO terms. This allowed us to check if the most enriched biological processes, molecular functions, cellular components and phenotypes were in accordance with previous knowledge.

Prior studies relate HD to dysfunctions in transcription, intracellular signaling, intracellular transport, endocytic recycling, and mitochondria [12]. This knowledge is consistent with the biological processes, cellular components and molecular functions that we found to be enriched (Fig. 2A and Supplementary Figure 1 A, B). The same holds for the phenotypes, which are associated to negativism, social and occupational deterioration, mitochondrial and nervous issues (Supplementary Figure 1 C) [12,15].

MS is an autoimmune disorder whose inflammatory infiltrates contain T-lymphocytes and B-cells, and leads to oligodendrocyte damage and demyelination [19]. This is coherent with the enriched terms in our analysis (Fig. 2B and Supplementary Figure 2).

The fact that the obtained results were confirmed by the literature suggested that the gathered genes were representative of the diseases under study.

Furthermore, as in Menche et al. [45], the disease modules were tested to be nonrandom gene aggregates. The size of the largest connected component of the disease module was compared to the size of the one obtained by randomly picking the proteins (matching the number of the disease-related genes) from the interactome (the comparison is shown in Supplementary Figure 3). For both diseases, the disease module resulted to be significantly larger than the random counterpart, allowing us to state that they cannot be attributed to a casual aggregation of genes.

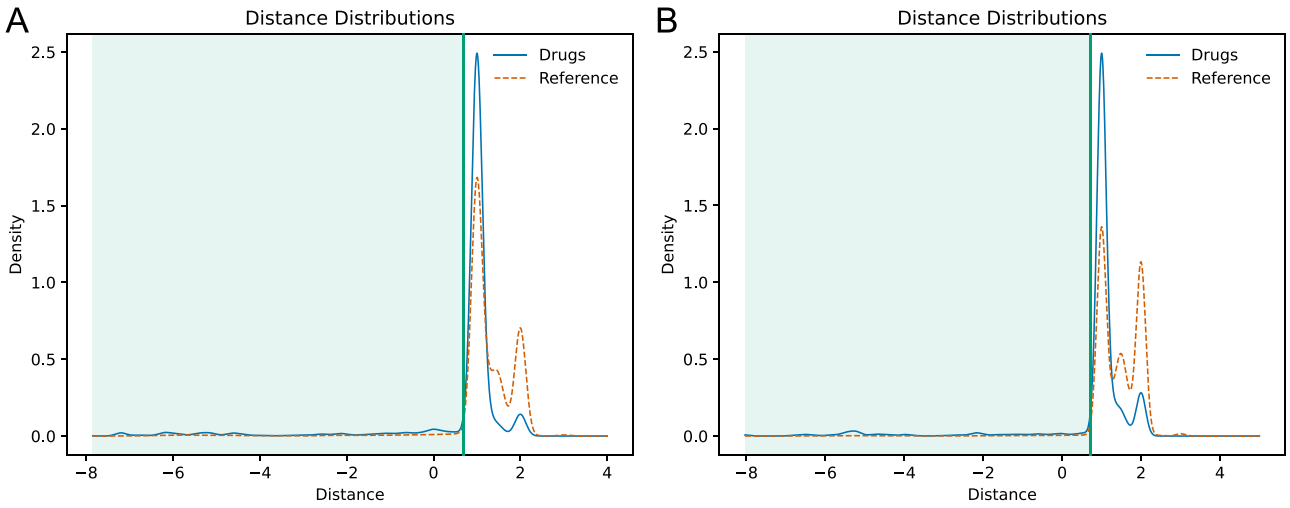

Fig. 3. Distance Distributions. The distribution of the distance between drug targets and disease-related proteins (solid blue line) compared to that of a reference collection (dashed orange line), for Huntington's disease (A) and multiple sclerosis (B). On the vertical axis, the kernel density estimation of the distribution is shown. The plot is divided into two parts by the chosen distance threshold (green line, see Methods).

Repurposable drugs selection

The network-based proximity analysis, leveraging on the potential of a system view, could suggest valuable drugs able to interfere with the disease molecular determinants in a non-trivial way (i.e., not only directly targeting disease-related genes). The idea behind this method is that drugs proximal to the disease module should be more effective than distant ones, as shown by Guney et al. in an extensive analysis that considered known diseases and disease-associated genes, as well as drugs and their targets [4]. Following Peng's protocol [10], the procedure compares the distribution of the distances between drug targets and disease-related proteins to that of a reference collection (see Methods and Fig. 3). For both diseases, it was possible to identify a distance value below which the two density curves (drugs and reference) drop dramatically. In particular, the reference density assumes negligible values for distances below this point (Fig. 3, green part of the plot). We elected such distance value (Fig. 3, vertical green line) as the threshold to discriminate drugs associated to the diseases. These distances are 0.68 and 0.71 (corresponding to proximity: −0.53 and −0.98) for HD and MS, respectively. From this analysis, 685 (11.8%) out of the 5798 drugs collected from DrugBank were considered significantly proximal medicaments for HD, and 475 (8.2%) for MS.

In order to evaluate the impact of a drug on the disease, we examined the effect of its administration on the expression of the disease-related genes in relevant cell lines (see Methods). We pursued this objective by performing an Inverted Gene Set Enrichment Analysis (IGSEA) on 896 drugs, observed in 6 212 LINCS expression datasets for HD and 960 drugs in 5 579 datasets for MS. This analysis resulted in 843 and 600 significantly enriched drugs, for HD and MS respectively.

The drugs that were both significantly enriched and proximal to the disease were deemed to be repurposable drug candidates: 138 for HD and 38 for MS (Supplementary Tables 2, 3). The interactions between the MS-related-genes, the drug targets, and the repurposable drug candidates are visualized in Fig. 4 (and Supplementary Figure 4 for HD), showing how drugs can be related to the disease through their targets.

Unfortunately, only a small portion of the proximal drugs has data in the LINCS database (21.9% for HD and 13.7% for MS). Even though the IGSEA analysis increases the reliability of the results, it dramatically reduces the number of molecules that can be investigated and possibly proposed. This has to be taken into account when evaluating the outcomes of the study.

To be more confident on the pool of predicted repurposable drugs, we replicated the entire procedure using three different interactomes. Our original one and the two networks used to validate it differ both in size and in listed interactions (see Methods). Despite these differences, the repurposable drug sets suggested for both investigated diseases resulted fairly consistent. In the case of HD, 138 drugs were prioritized based on the original interactome, 110 on Cheng's one, 133 on our restricted interactome. It is noteworthy that all the molecules retrieved from the two smaller interactomes are included among those of the first one. A very similar conclusion could be drawn for MS, for which the procedure predicted 39 drugs with the large interactome, 26 with Cheng's one, and 29 with the severely constrained version of our interactome.

Drug combinations

Combined therapies and multi-targeted agents have proven to offer significant advantages over monotherapy, presenting higher efficacies and less adverse reactions [6,46]. Due to combinatorial explosion, however, it is generally not feasible to test all theoretically possible associations. For this reason, we adopted a recent methodology proposed by Cheng et al. [6], which is based on the estimation of target neighborhoods separation on the human protein-protein interactome. Taking advantage of that, the investigated combinations may be screened on the basis of the pharmacological relationship between drugs (see Methods). Additionally, we looked in DrugBank for approved associations and interactions that increase the risk or severity of adverse effects. In this way, we ended up having an assortment of plausible combinations listed in Supplementary Tables 4, 5 and identifiable in the annotated heatmaps of Fig. 5 for HD and of Supplementary Figure 5 for MS.

Discussion

Protein-Protein interactomes

For both diseases, the repurposable drug pools predicted using the three interactomes are in reasonable concordance. However, it is significant that the execution of the pipeline on our interactome, compared to the same procedure on Cheng's interactome, improves the prediction adding 7 drugs with evidence from clinical trials, 9 from in vivo studies, 1 from in vitro experiments for HD, and 5 clinically tested drugs and 1 investigated in an animal model for MS.

This outcome seems to suggest that injecting more input data in the procedure (still maintaining high reliability standards) leads to in-

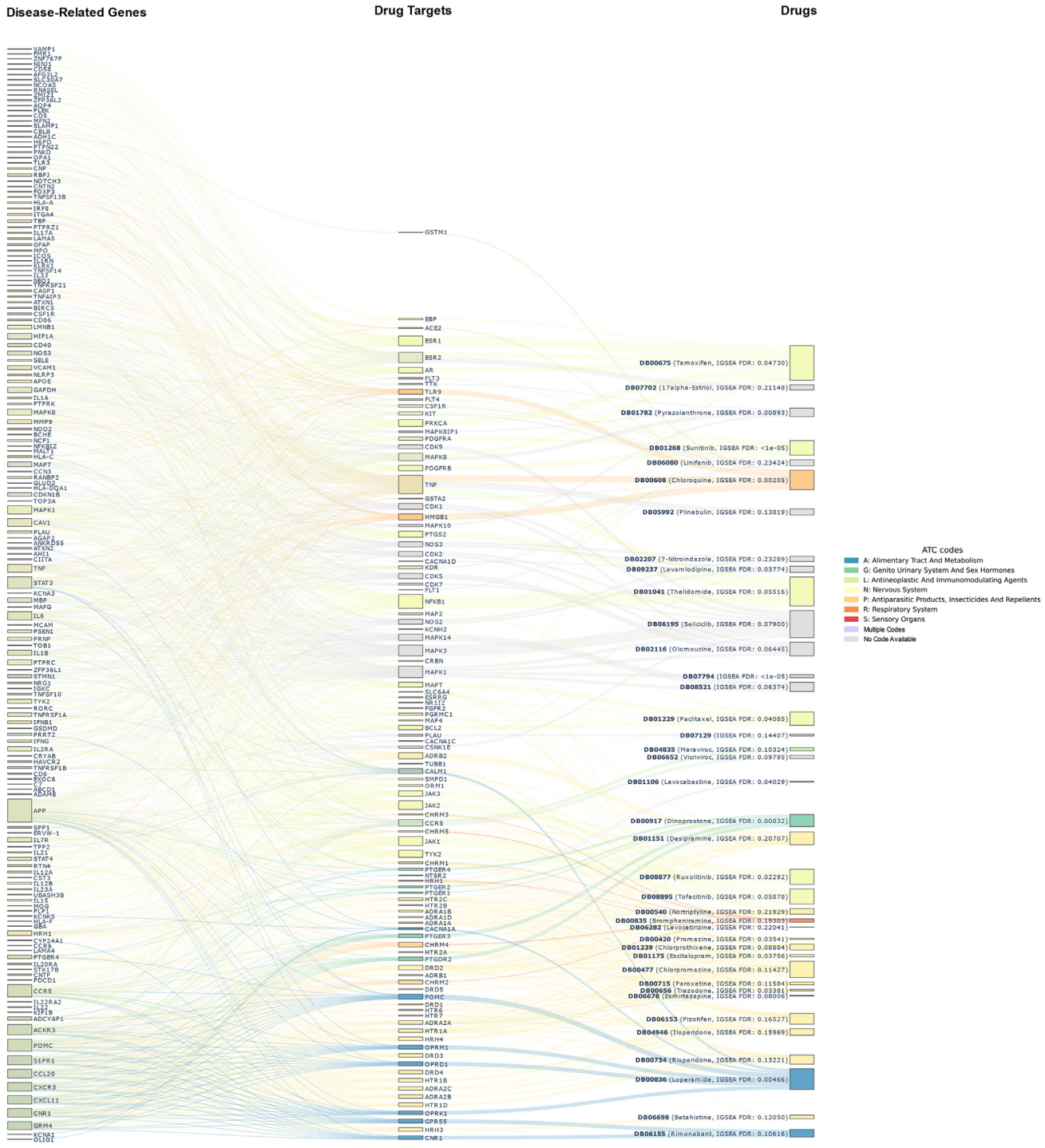

Fig. 4. Multiple Sclerosis Gene-Target-Drug Network. The Sankey diagram illustrates the interconnections between disease-related genes, drug targets, and drugs. Each drug (right column) is connected to its reported targets (middle column), which, in turn, are proximal on the human interactome to some of the disease-associated proteins (left column). Drugs are colored by the respective ATC code, and the FDR of the IGSEA analysis (see Methods) is reported in the label.

creased performance, which is perfectly in line with the Big Data perspective [47].

Repurposable drugs

Among the drugs selected to tackle HD (138), several (17) have been clinically tested and suggested, many show strong evidence from in vivo tests (35) or promising results from in vitro assays (9). All the references are reported in the Supplementary Table 2, the most noticeable examples being selisistat [48], lisuride [49], valproic acid [50], and risperidone [20].

Selisistat was found to be safe, well tolerated, and capable of reaching a plasma concentration compatible with the SirT1 inhibition, which has been shown to restore transcriptional dysregulation in models of HD [48]. Lisuride is able to induce a temporary yet significant improvement in the motor performance of patients with hyperkinesia caused by HD [49]. Valproic acid was shown to be a possible alternative treatment for HD patients suffering from myoclonic hyperkinesia [50]. Risperidone

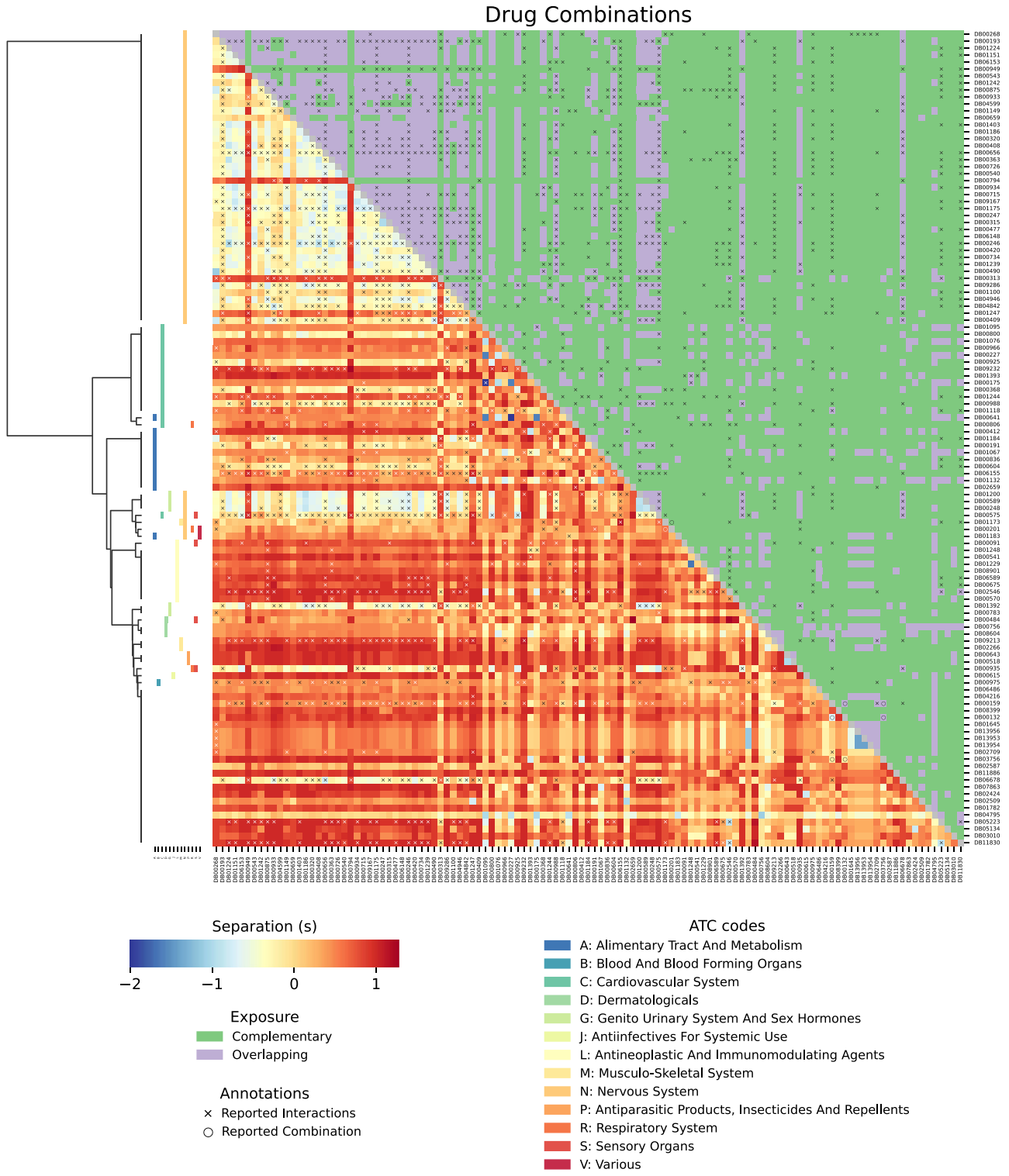

Fig. 5. Huntington's Disease Drug Combinations. The annotated heatmap provides info about possible combinations of the selected drugs. A combination is marked with × if an interaction is reported in DrugBank, and with ○ if it is present in an approved formulation. The lower-left part of the heatmap shows the separation of the inspected drugs, color coded from blue (no separation) to red (strongly separated). The upper-right portion, instead, displays the kind of exposure: violet if overlapping and green if complementary. At the leftmost part, the ATC codes of the drugs are reported along with a dendrogram of their hierarchical clustering.

has beneficial effects in the treatment of psychiatric manifestations and stabilization of motor symptoms in patients with HD [20].

Inspecting the drugs screened for MS (38), we obtained a comparable outcome: 7 of them are clinically studied and 9 experimented on animal models. All the evidence is listed in the Supplementary Table 3. Most of the drugs in clinical trials aim to alleviate the symptoms, while the only one we found to be capable of reducing relapses is Escitalopram [51] for which there is evidence suggesting it may be an effective and well-tolerated treatment for preventing stress-related relapses in women with MS [51].

Examining the Anatomical Therapeutic Chemical (ATC) codes of the repurposable drugs, the first thing to notice is the predictable prevalence of drugs associated to the ATC code N (Nervous System) for both diseases. Apart from this, the most common codes for HD repurposable drugs are C (Cardiovascular) and L (Antineoplastic and Immunomodulating Agents). The first group is mainly represented by statins, used to cope with the cholesterol impairment typical of HD patients [52]. The immunomodulating agents are principally immunosuppressants and histone deacetylases inhibitors, the last ones aimed at recovering from the histone hypoacetylation common in neurological disorders [53].

For MS, instead, the second most frequent code is L (Antineoplastic and Immunomodulating Agents). Some relevant examples are ruxolitinib, paclitaxel, tamoxifen, and thalidomide, which are capable of attenuating experimental autoimmune encephalomyelitis and of inducing remyelination [54–58].

Drug combinations

Observing the obtained results (depicted as annotated heatmaps in Fig. 5 for HD and Supplementary Figure 5 for MS) it is interesting to highlight that drugs that do not have ATC codes associated to them are also those with few (or nothing at all) reported interactions. This suggests that they are not sufficiently characterized and additional studies on them are needed before further consideration.

The collected plausible combinations are numerous, but the association of orphenadrine (DB01173) and caffeine (DB00201) for HD deserves to be highlighted. These molecules are present along with acetylsalicylic acid (ASA) in an FDA approved formulation for muscular pain relief. This medication is noteworthy for many reasons. First of all, pain is a known issue in HD and could be an important non-motor symptom [59,60] thus, its treatment should not be neglected. Furthermore, orphenadrine showed to be effective in preventing neurotoxicity in rats with a chemically-induced condition that mimics the histological and neurochemical features of HD [61]. Additionally, low dosages of caffeine showed to be beneficial in HD animal models [62]. Finally, ASA was included in the formulation for relieving pain and decreasing swelling. Even though ASA was proximal to HD, it was not included in our results because its data was not available in LINCS for the investigated cell lines. However, it is actually profitable for the present aim, since it showed to prevent protein aggregation in several neurodegen-

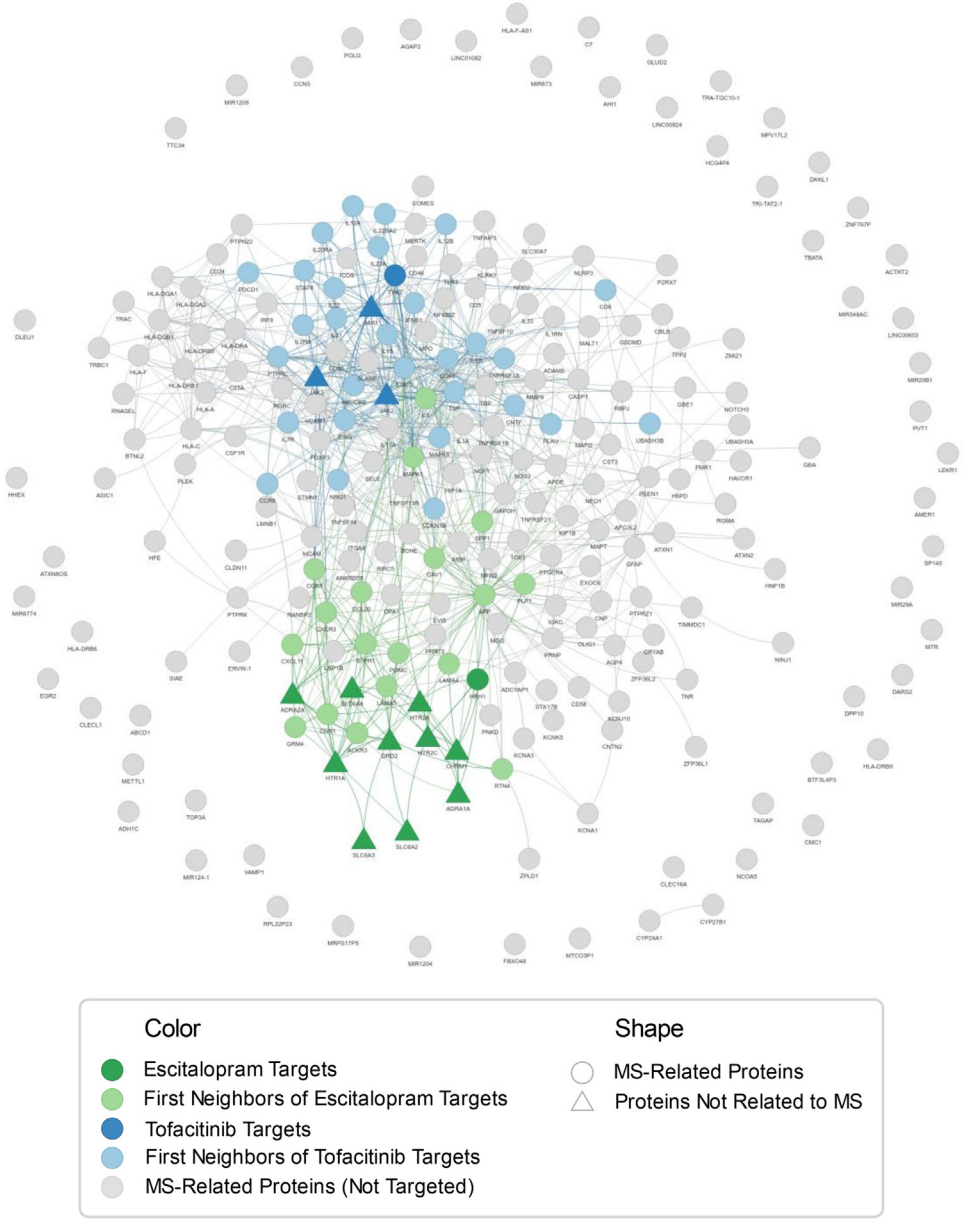

Fig. 6. Escitalopram and Tofacitinib Complementary Exposure. The network displays the proteins associated to MS (circles) and highlights those targeted by escitalopram and tofacitinib (dark green and dark blue, respectively). Targets that are not related to MS are indicated as triangles. In order to better illustrate the influence of these two molecules given by the tight interconnection of the proteome, the first neighbors of the drug targets are depicted in a lighter color (light green for neighbors of targets of escitalopram, and light blue tofacitinib's ones).

erative diseases [63]. Further assessments are needed, but this could be an interesting point where to start.

A reasonable hypothesis for treating MS might be an association of two drugs sufficiently separated from each other as escitalopram (DB01175) and tofacitinib (DB08895) or ruxolitinib (DB08877), capable of affecting complementary parts of the disease module. Fig. 6 shows the network of the interactions among proteins associated to MS (all circles) and highlights those targeted by escitalopram and tofacitinib (dark green and dark blue, respectively). Among these targets, two of them, namely HRH1 for escitalopram (dark green circle) and TYK2 for tofacitinib (dark blue circle), belong to the MS disease module, while those that are not directly related to MS are depicted as triangles (maintaining the same color coding). In order to better illustrate the influence on the disease module of the two drugs in terms of protein-protein interactions, the first neighbors of the drug targets are colored lighter (light green for neighbors of targets of escitalopram, and light blue for tofacitinib's ones). It can be seen that overall the targets of both escitalopram and tofacitinib or their first neighbors can influence a reasonable part of the disease module without redundantly interfering with the same MS proteins. In fact, our analysis showed that these drugs are proximal to MS and significantly influence proteins associated to this disease. Additionally, no interactions between them have been reported in DrugBank. Moreover, we found experimental evidence supporting this inference. Escitalopram is a selective serotonin re-uptake inhibitor (ATC code: N, Nervous) that in humans proved to prevent stress-related relapses [51]. Tofacitinib and ruxolitinib showed promising effects in animal models: the first one enhancing remyelination and improving myelin integrity [64], and the second one ameliorating the severity of the disease [54]. Furthermore, they are Janus kinase (JAK) inhibitors (ATC code: L, Antineoplastic and Immunomodulating Agents) and the JAK/STAT pathway is aberrantly activated in MS [21,65].

In the other drug combinations, which are sufficiently separated (see Methods, green on the heatmaps) and for which no adverse interactions are reported (not annotated with an × in the heatmaps), valuable clues for polypharmacological interventions could be found. A working hypothesis might be to choose two drugs tackling different aspects of a disease, for instance featuring distinct ATC codes.

Limitations

Despite our best efforts, this study is not exempt from some shortcomings that are common in data analysis, and regard mainly the data availability and quality. This could have led us to miss some promising compounds and, at the same time, it may compromise some of the analyses.

A complete characterization of all available drugs and human proteins is surely not at hand, and this has repercussions on many aspects of the study, like, e.g., the human protein-protein interactome construction, drug association to biological processes, cellular components, molecular functions and phenotypes, and drug induced gene expression profiles retrieval. Only sometimes, this issue could be partially mitigated by an extensive integration of data from a wider variety of databases. Noteworthy, puzzling examples could be the drug-target association and the availability of expression data in LINCS. The number of targets associated to a specific drug could considerably depend on the amount of research carried out on that medicine rather than on the actual biological interactions it has. This influences the drug-disease proximity evaluation. Additionally, as stated above, the LINCS database does not provide expression profiles for all the drugs selected by network proximity, limiting by far the choice space for drug repurposing.

Furthermore, if the knowledge we have about drugs is incomplete, the one we have on their combination is even sparser. This, obviously, affects our ability to screen and judge plausible associations.

Moreover, it could be argued that, even though the drug-disease proximity is evaluated with a rigorous geometrical approach, the choice of the distance threshold we use for discriminating drug efficacy is quite discretionary.

Conclusions

Here, we extended an unsupervised computational framework for drug repurposing with a network-based analysis for screening the possible drug combination therapies. Applying this pipeline to HD and MS, we identified several repurposable drug candidates, some of which have already been studied in humans. Eventually, we ended up with 138 potential drugs for HD and 38 for MS. Their plausible combinations are numerous, but this work can help to prioritize them. While these results are exploratory and should be experimentally verified before further consideration, they could provide valuable clues for improving the management of HD and MS.

Finally, this pipeline demonstrated to be effective on both investigated diseases, even though they have a different nature. For this reason, it could potentially provide new suggestions also for other complex disorders.

Data and code availability

The most relevant results are included in this article (and its supplementary information files). The whole generated data is publicly available from the GitHub repository https://github.com/LucaMenestrina/UnsupervisedComputationalFrameworkForDrugRepurposing, as well as the full code for the collection, building and analysis. A detailed reference of the source data is provided in the file "data/sources/sources.json" of the aforementioned repository (for every database are reported: name, version, license, employed files, URL and date of access).

Declaration of Competing Interest

The authors declare no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2022.100042.

References

[1] Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 2018;18:41–58. doi:10.1038/nrd.2018.168.
[2] Talevi A, Bellera CL. Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics. Expert Opin Drug Discov 2020;15:397–401. doi:10.1080/17460441.2020.1704729.
[3] Choudhury C, Arul Murugan N, Priyakumar UD. Structure-based drug repurposing: traditional and advanced AI/ML-aided methods. Drug Discov Today 2022. doi:10.1016/j.drudis.2022.03.006.
[4] Guney E, Menche J, Vidal M, Barábasi AL. Network-based in silico drug efficacy screening. Nat Commun 2016;7:10331. doi:10.1038/ncomms10331.
[5] Cheng F, Desai RJ, Handy DE, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun 2018;9:2691. doi:10.1038/s41467-018-05116-5.
[6] Cheng F, Kovács IA, Barabási AL. Network-based prediction of drug combinations. Nat Commun 2019;10:1197. doi:10.1038/s41467-019-09186-x.
[7] Cheng F, Lu W, Liu C, et al. A genome-wide positioning systems network algorithm for in silico drug repurposing. Nat Commun 2019;10:3476. doi:10.1038/s41467-019-10744-6.
[8] Zhou Y, Hou Y, Shen J, et al. A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for COVID-19. PLOS Biol 2020;18:e3000970. doi:10.1371/JOURNAL.PBIO.3000970.
[9] Fang J, Zhang P, Zhou Y, et al. Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease. Nat Aging 2021;1:1175–88. doi:10.1038/s43587-021-00138-z.

[10] Peng Y, Yuan M, Xin J, Liu X, Wang J. Screening novel drug candidates for Alzheimer's disease by an integrated network and transcriptome analysis. Bioinformatics 2020;36:4626–32. doi:10.1093/bioinformatics/btaa563.
[11] Quan P, Wang K, Yan S, et al. Integrated network analysis identifying potential novel drug candidates and targets for Parkinson's disease. Sci Rep 2021;11:13154. doi:10.1038/s41598-021-92701-2.
[12] Bates GP, Dorsey R, Gusella JF, et al. Huntington disease. Nat Rev Dis Prim 2015;1:15005. doi:10.1038/nrdp.2015.5.
[13] Loscalzo J, Kohane I, Barabasi AL. Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. Mol Syst Biol 2007;3:124. doi:10.1038/MSB4100163.
[14] Wright GEB, Black HF, Collins JA, et al. Interrupting sequence variants and age of onset in Huntington's disease: clinical implications and emerging therapies. Lancet Neurol 2020;19:930–9. doi:10.1016/S1474-4422(20)30343-4.
[15] Finkbeiner S, disease Huntington's. Cold Spring Harb Perspect Biol 2011;3:a007476. doi:10.1101/cshperspect.a007476.
[16] Kobelt G, Thompson A, Berg J, Gannedahl M, Eriksson J. New insights into the burden and costs of multiple sclerosis in Europe. Mult Scler 2017;23:1123–36. doi:10.1177/1352458517694432.
[17] Leray E, Moreau T, Fromont A, Edan G. Epidemiology of multiple sclerosis. Rev Neurol (Paris) 2016;172:3–13. doi:10.1016/j.neurol.2015.10.006.
[18] Ramagopalan SV, Dobson R, Meier UC, Giovannoni G. Multiple sclerosis: risk factors, prodromes, and potential causal pathways. Lancet Neurol 2010;9:727–39. doi:10.1016/S1474-4422(10)70094-6.
[19] Dobson R, Giovannoni G. Multiple sclerosis – a review. Eur J Neurol 2019;26:27–40. doi:10.1111/ene.13819.
[20] Duff K, Beglinger LJ, O'Rourke ME, et al. Risperidone and the treatment of psychiatric, motor, and cognitive symptoms in Huntington's disease. Ann Clin Psychiatry 2008;20:1–3. doi:10.1080/10401230701844802.
[21] Hamid KM, Isiyaku A, Kalgo MU, Yahaya IS, Mirshafiey A, Lodges in JAK-STAT. Multiple sclerosis: pathophysiology and therapeutic approach overview. Open Access Libr J 2017;4:e3492. doi:10.4236/oalib.1103492.
[22] Gitler AD, Dhillon P, Shorter J. Neurodegenerative disease: models, mechanisms, and a new hope. DMM Dis Model Mech 2017;10:499–502. doi:10.1242/dmm.030205.
[23] Recanatini M, Cabrelle C. Drug research meets network science: where are we? J Med Chem 2020;63:8653–66. doi:10.1021/acs.jmedchem.9b01989.
[24] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61. doi:10.1093/nar/gkw1092.
[25] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res 2019;47:D1038–43. doi:10.1093/nar/gky1151.
[26] Ramos EM, Hoffman D, Junkins HA, et al. Phenotype-genotype integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur J Hum Genet 2014;22:144–7. doi:10.1038/ejhg.2013.96.
[27] Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. Methods 2015;74:83–9. doi:10.1016/j.ymeth.2014.11.020.
[28] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020;48:D845–55. doi:10.1093/nar/gkz1021.
[29] Carbon S, Douglass E, Good BM, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021;49:D325–34. doi:10.1093/nar/gkaa1113.
[30] Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021. Nucleic Acids Res 2021;49:D1207–17. doi:10.1093/nar/gkaa1043.
[31] Klopfenstein DV, Zhang L, Pedersen BS, et al. GOATOOLS: a python library for gene ontology analyses. Sci Rep 2018;8:10872. doi:10.1038/s41598-018-28948-z.
[32] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B. 1995;57:289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.
[33] Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. doi:10.1093/nar/gkx1037.
[34] Lamb J, Crawford ED, Peck D, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006;313:1929–35. doi:10.1126/science.1132939.
[35] Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets - Update. Nucleic Acids Res 2013;41:D991–5. doi:10.1093/nar/gks1193.
[36] Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 2017;171:1437–52. doi:10.1016/j.cell.2017.10.049.
[37] Alonso-López Di, Campos-Laborie FJ, Gutiérrez MA, et al. APID database: redefining protein-protein interaction experimental evidences and binary interactomes. Database 2019;2019:baz005. doi:10.1093/database/baz005.
[38] Oughtred R, Rust J, Chang C, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 2021;30:187–200. doi:10.1002/pro.3978.
[39] Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8. doi:10.1038/s41586-020-2188-x.
[40] Breuer K, Foroushani AK, Laird MR, et al. InnateDB: systems biology of innate immunity and beyond - Recent updates and continuing curation. Nucleic Acids Res 2013;41:D1228–33. doi:10.1093/nar/gks1147.
[41] Meyer MJ, Das J, Wang X, Yu H. INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics 2013;29:1577–9. doi:10.1093/bioinformatics/btt181.
[42] Orchard S, Ammari M, Aranda B, et al. The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63. doi:10.1093/nar/gkt1115.
[43] Csabai L, Fazekas D, Kadlecsik T, et al. SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks. Nucleic Acids Res 2022;50:D701–9. doi:10.1093/nar/gkab909.
[44] Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019;47:D607–13. doi:10.1093/nar/gky1131.
[45] Menche J, Sharma A, Kitsak M, et al. Uncovering disease-disease relationships through the incomplete interactome. Science 2015;347:1257601. doi:10.1126/science.1257601.
[46] Jia J, Zhu F, Ma X, et al. Mechanisms of drug combinations: interaction and network perspectives. Nat. Rev. Drug Discov. 2009;8:111–28. doi:10.1038/nrd2683.
[47] Zhu H. Big data and artificial intelligence modeling for drug discovery. Annu Rev Pharmacol Toxicol 2020;60:573–89. doi:10.1146/annurev-pharmtox-010919-023324.
[48] Süssmuth SD, Haider S, Landwehrmeyer GB, et al. An exploratory double-blind, randomized clinical trial with selisistat, a SirT1 inhibitor, in patients with Huntington's disease, Br. J Clin Pharmacol 2015;79:465–76. doi:10.1111/bcp.12512.
[49] Frattola L, Albizzati MG, Alemani A, et al. Acute treatment of Huntington's chorea with lisuride. J Neurol Sci 1983;59:247–53. doi:10.1016/0022-510X(83)90042-4.
[50] Saft C, Lauter T, Kraus PH, Przuntek H, Andrich JE. Dose-dependent improvement of myoclonic hyperkinesia due to valproic acid in eight Huntington's Disease patients: a case series. BMC Neurol 2006;6:11. doi:10.1186/1471-2377-6-11.
[51] Mitsonis CI, Zervas IM, Potagas CM, et al. Effects of escitalopram on stress-related relapses in women with multiple sclerosis: an open-label, randomized, controlled, one-year follow-up study. Eur Neuropsychopharmacol 2010;20:123–31. doi:10.1016/j.euroneuro.2009.10.004.
[52] Karasinska JM, Hayden MR. Cholesterol metabolism in Huntington disease. Nat Rev Neurol 2011;7:561–72. doi:10.1038/nrneurol.2011.132.
[53] Shukla S, Tekwani BL. Histone deacetylases inhibitors in neurodegenerative diseases, neuroprotection and neuronal differentiation. Front Pharmacol 2020;11:537. doi:10.3389/FPHAR.2020.00537/BIBTEX.
[54] Hosseini A, Gharibi T, Mohammadzadeh A, et al. Ruxolitinib attenuates experimental autoimmune encephalomyelitis (EAE) development as animal models of multiple sclerosis (MS). Life Sci 2021;276:119395. doi:10.1016/j.lfs.2021.119395.
[55] Moscarello MA, Mak B, Nguyen TA, et al. Paclitaxel (Taxol) attenuates clinical disease in a spontaneously demyelinating transgenic mouse and induces remyelination. Mult Scler J 2002;8:130–8. doi:10.1191/1352458502ms776oa.
[56] Gonzalez GA, Hofer MP, Syed YA, et al. Tamoxifen accelerates the repair of demyelinated lesions in the central nervous system. Sci Rep 2016;6:1–13. doi:10.1038/srep31599.
[57] Rankin KA, Mei F, Kim K, et al. Selective estrogen receptor modulators enhance cns remyelination independent of estrogen receptors. J Neurosci 2019;39:2184–94. doi:10.1523/JNEUROSCI.1530-18.2019.
[58] Contino-Pépin C, Parat A, Patinote C, et al. Thalidomide derivatives for the treatment of neuroinflammation. ChemMedChem 2010;5:2057–64. doi:10.1002/CMDC.201000326.
[59] Underwood M, Bonas S, Dale M, Disease Huntington's. Prevalence and psychological indicators of pain. Mov Disord Clin Pract 2017;4:198–204. doi:10.1002/MDC3.12376.
[60] Sprenger GP, Roos RAC, van Zwet E, et al. The prevalence of pain in Huntington's disease in a large worldwide cohort. Park Relat Disord 2021;89:73–8. doi:10.1016/j.parkreldis.2021.06.015.
[61] Pubill D, Verdaguer E, Canudas AM, et al. Orphenadrine prevents 3-nitropropionic acid-induced neurotoxicity in vitro and in vivo. Br J Pharmacol 2001;132:693–702. doi:10.1038/sj.bjp.0703869.
[62] Kolahdouzan M, Hamadeh MJ. The neuroprotective effects of caffeine in neurodegenerative diseases. CNS Neurosci Ther 2017;23:272–90. doi:10.1111/cns.12684.
[63] Ayyadevara S, Balasubramaniam M, Kakraba S, et al. Aspirin-mediated acetylation protects against multiple neurodegenerative pathologies by impeding protein aggregation. Antioxidants Redox Signal 2017;27:1383–96. doi:10.1089/ars.2016.6978.
[64] Günaydın C, Önger ME, Avcı B, et al. Tofacitinib enhances remyelination and improves myelin integrity in cuprizone-induced mice. Immunopharmacol Immunotoxicol 2021;43:790–8. doi:10.1080/08923973.2021.1986063.
[65] Benveniste EN, Liu Y, McFarland BC, Qin H. Involvement of the Janus kinase/signal transducer and activator of transcription signaling pathway in multiple sclerosis and the animal model of experimental autoimmune encephalomyelitis. J Interf Cytokine Res 2014;34:577–88. doi:10.1089/jir.2014.0012.