ODAMNet: A Python package to identify molecular relationships between chemicals and rare diseases using overlap, active module and random walk approaches

Morgane Térézol, Anaïs Baudot, Ozan Ozisik

SoftwareX 26 (2024) 101701

Contents lists available at ScienceDirect

SoftwareX

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

ODAMNet: A Python package to identify molecular relationships between chemicals and rare diseases using overlap, active module and random walk approaches

Morgane Térézol [a,*], Anaïs Baudot [a,b,c], Ozan Ozisik [a]

[a] Aix Marseille Univ, INSERM, MMG, Marseille, France
[b] Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
[c] CNRS, Marseille, France

ARTICLE INFO

Keywords:
Rare disease
Chemical
Environmental factor
Overlap analysis
Random walk with restart
Active module identification

ABSTRACT

Environmental factors are external conditions that can affect the health of living organisms. For a number of rare genetic diseases, an interplay between genetic and environmental factors is known or suspected. However, the studies are limited by the scarcity of patients and the difficulties in gathering reliable exposure information.

In order to aid in fostering research between environmental factors and rare diseases, we propose ODAMNet, a Python package to investigate the possible relationships between chemicals, which are a subset of environmental factors, and rare diseases. ODAMNet offers three different and complementary bioinformatics approaches for the exploration of relationships: overlap analysis, active module identification and random walk with restart. ODAMNet allows systematic analysis of chemical - rare disease relationships and generation of hypotheses for further investigation of effect mechanisms.

Metadata

C1	Current code version	v1.1.0
C2	Permanent link to code/ repository used for this code version	https://github.com/MOohTus/ODAMNet
C3	Permanent link to reproducible capsule	https://github.com/MOohTus/ODAMNet/tree/v1.1.0
C4	Legal code license	MIT License
C5	Code versioning system used	git
C6	Software code languages, tools and services used	python, DOMINO server, CTD, WikiPathways, NDEx
C7	Compilation requirements, operating environments and dependencies	python >= 3.9, multiXrank==0.1, requests, SPARQLWrapper, pandas, scipy, statsmodels, alive_progress, click_option_group, click, ndex2, networkx
C8	If available, link to developer documentation/manual	https://odamnet.readthedocs.io/en/latest/
C9	Support email for questions	morgane.terezol@univ-amu.fr

1. Motivation and significance

Environmental factors are external conditions that can affect the health of living organisms. These factors can be physical, biological, social, economic, or political [1]. They can play a significant role in the development and progression of genetic diseases. The relationship between genes and environment has been compared to the relationship between a loaded gun and its trigger [2].

The role of environmental factors have been reviewed for different diseases and disease groups, such as autism [3], inflammatory bowel disease [4,5], cardiovascular disease [6,7], congenital anomalies of the kidney and urinary tract [8,9], amyotrophic lateral sclerosis [10], idiopathic pulmonary fibrosis [11], and Legg–Calvé–Perthes Disease [12].

There are multiple methodological challenges in studying the interplay between environmental factors and genetic diseases. Environmental factors have large spatial and temporal heterogeneities; a person's activity patterns, residential changes and other conditions can modify the amount of exposure [13]. For the diseases that manifest later in life, relationships with an exposure that happened decades ago is hard to prove [10]. In the case of rare diseases, sample scarcity is adding

* Corresponding author.
E-mail address: morgane.terezol@univ-amu.fr (M. Térézol).

https://doi.org/10.1016/j.softx.2024.101701
Received 6 July 2023; Received in revised form 20 December 2023; Accepted 21 March 2024
Available online 1 April 2024

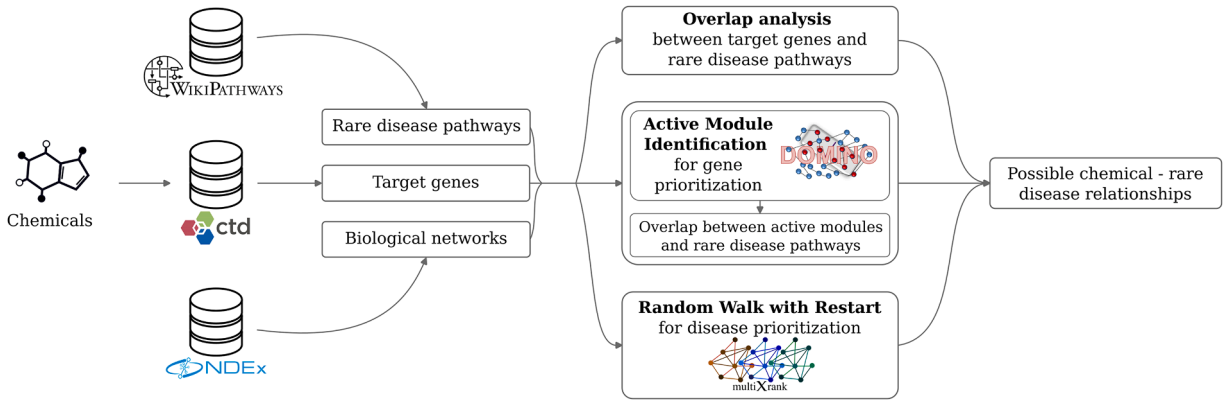

Fig 1. ODAMNet workflow.

another level of difficulty to the studies of environmental factor - disease relationships: low sample size and testing for multiple factors decrease the statistical power. Hence, limiting the environmental factors to be investigated is important, and this requires well supported hypotheses.

With this study, our aim is to provide a tool that can generate knowledge-based hypotheses regarding the relationships between chemicals, which are a subset of environmental factors, and the rare diseases. We previously investigated the role of vitamin A and vitamin D in the etiology of Congenital Anomalies of the Kidney and Urinary Tract (CAKUT) [14]. We explored the overlap between vitamin target genes and gene sets related to CAKUT. We observed significant enrichment of vitamin A target genes in CAKUT-related gene sets. Here, we propose ODAMNet ("Overlap, Diffusion, Active Module, Network"), a Python package that allows performing systematic analyses with i) multiple chemicals, ii) multiple rare diseases, and iii) multiple integrative bioinformatics approaches to explore the possible relationships between chemicals and rare diseases.

In ODAMNet, targets of the chemicals are retrieved from the Comparative Toxicogenomics Database (CTD) [15], and rare disease pathways are retrieved from WikiPathways [16]. We used three different and complementary bioinformatics approaches to integrate the data: overlap analysis, active module identification and random walk with restart. Of note, for the network-based approaches, i.e. active module identification and random walk with restart, the biological interaction networks are downloaded from the Network Data Exchange (NDEx) [17].

There are multiple resources that integrate biological data stored in different databases, including Pathway Commons [18], OmniPath [19,20], Hetionet [21], NeDRex [22], Drugst.one [23], BOCK [24], PrimeKG [25]. While Pathway Commons and OmniPath are specialized in molecular interactions, Hetionet, BOCK, and PrimeKG serve as extensive knowledge graphs, integrating highly heterogeneous data. These resources gather a wealth of information but leveraging this information is not straightforward and demands additional tools. In contrast, NeDRex and Drugst.one, akin to ODAMNet, focus on integrating data sources and analysis methods tailored to specific research questions, that is drug repurposing in this instance. A detailed comparison of ODAMNet with these tools can be found in the Impact section.

The novelty of ODAMNet stems from multiple factors. First of all, ODAMNet is developed for the systematic analysis of chemical - rare disease relationships in order to generate hypotheses. ODAMNet integrates the three most relevant databases for this purpose. ODAMNet does not use static data but works by querying the integrated databases. This, for example, allows harnessing up-to-date rare disease information available from WikiPathways, benefiting from the ongoing curation efforts. ODAMNet, with its extensive documentation and simple setup thanks to pip and conda, provides easy access to three complementary bioinformatics approaches. Furthermore, ODAMNet offers the flexibility to use any user-provided input dataset, allowing the application of the bioinformatics approaches to any data source and research question. Overall, ODAMNet extends the bioinformatics analysis ecosystem available in Python.

2. Software description

2.1. Software architecture

ODAMNet is a Python package for the investigation of chemical - rare disease relationships. It takes a list of chemicals as input and automatically retrieves the genes that are targeted by these chemicals from the Comparative Toxicogenomics Database (CTD) (Fig 1). Rare disease pathways are retrieved automatically from WikiPathways and networks are downloaded automatically from NDEx [17]. The user can also provide their own target genes, pathways of interest and biological networks. Then, ODAMNet can perform three different approaches. The first approach is an overlap analysis between the target genes and the rare disease pathways. The second approach is active module identification (AMI) using DOMINO [26], followed by an overlap analysis between the identified active modules and rare disease pathways. The third approach is a random walk with restart (RWR) using multiXrank [27].

ODAMNet is written in Python 3 and is designed primarily as a command line program for the Linux operating system. It can be installed using pip and a detailed documentation is available on https://odamnet.readthedocs.io/en/latest/, with descriptions of:

- the three integrative bioinformatics approaches
- the format of input and output files
- all input parameters
- two use-cases, in which the datasets are either automatically retrieved or provided by the user

2.2. Software functionalities

2.2.1. Data retrieval by queries

ODAMNet can retrieve the datasets required for an analysis, which are chemical target genes, rare disease pathways and biological networks, by querying the relevant databases.

For the retrieval of genes targeted by a list of chemicals, the required input is a list of MeSH identifiers (https://meshb.nlm.nih.gov/) that correspond to those chemicals. Chemical target genes which have been reported for human are automatically retrieved from CTD (http://ctdbase.org/tools/batchQuery.go) using HTTP requests. The user can select the type of chemical - target gene associations. By default, *directAssociation* parameter is set to "True" and returns target genes solely for the input chemicals. Alternatively, setting this parameter to "False" will

retrieve target genes not only for the input chemicals but also for any of its descendant chemicals. The user can also filter chemical - target gene associations based on the number of references, with the default being relationships supported by at least two references.

In the context of the European Joint Programme on Rare Diseases (EJP RD), there has been an extensive effort to curate rare disease pathways for the WikiPathways database (https://www.wikipathways.org/communities/rarediseases.html), resulting in more than 100 rare disease pathways. These rare disease pathways are composed of disease-associated genes and other genes that function in the mechanisms associated with the disease. This is well-suited for our study because we are using the information on chemicals' effects on the genes; genes without any genetic associations to the disease are also relevant. ODAMNet retrieves rare disease pathways and the genes they involve from WikiPathways using SPARQL queries (https://sparql.wikipathways.org/). All human pathways and associated genes are also retrieved to construct the background gene set that will be used in the statistical tests.

In ODAMNet active module identification approach, biological networks are downloaded automatically from NDEx (https://www.ndexbio.org/) using the NDEx2 Python Client [28] and the networks' universally unique identifiers (UUID). In the ODAMNet random walk with restart approach, multiple networks, including a bipartite network, are required. For this reason, ODAMNet does not perform the automatic download but provides assistive functions for downloading or building the necessary networks (please see Section 2.2.4).

2.2.2. Data input by the user

While the automatic retrieval of chemical targets and pathways from relevant databases is the default workflow, ODAMNet also allows the user to input custom target genes and gene sets such as annotation terms or pathways. In the case of using such custom gene sets, the user is expected to provide the background gene sets (as GMT files) that will be used in the statistical tests. It is possible to use gene sets from different databases (e.g. WikiPathways, Reactome [29], Gene Ontology (GO) [30,31]) in the same analysis; in this case, a background gene set for each database should also be provided. It should be noted that the identifier type of the genes must be consistent between the input files.

As stated in the previous section, in random walk with restart, ODAMNet does not download and use networks automatically. However ODAMNet provides assistive functions to download networks from NDEx. The user can also provide their own networks. The user can provide their own network in the active module identification approach, as well.

2.2.3. Integrative bioinformatics approaches

ODAMNet uses three approaches for finding the relationships between chemical target genes and rare disease pathways.

The first approach is an overlap analysis. It assesses if the chemical target genes are part of the rare disease pathways and applies a hypergeometric test to determine statistical significance. Benjamini-Hochberg method is then applied for the multiple testing correction.

The second approach is based on active module identification. An active module is a connected subset of gene nodes in a biological interaction network relevant to the investigated condition. For the discovery of active modules that have high connectivity and contain a high number of target genes (considered as active genes), we use the DOMINO method [26] with the default parameters through its web server [32]. Following the active module identification, a pairwise overlap analysis is performed between all the identified active modules and all the rare disease pathways. Duplicate significant rare disease pathways are removed, keeping only the most significant ones. The biological interaction network used by DOMINO can be automatically downloaded from NDEx or provided by the user.

There are multiple reasons behind the selection of DOMINO for active module identification. First, in contrast to other methods that require p-values as input, DOMINO takes binary input; genes should be either "active" or "non-active", which is convenient in our case where we use a list of chemical target genes as input. Second, DOMINO provides a web server and an API to access it, which is again convenient for integration with ODAMNet. Third, DOMINO works fast. Finally, DOMINO has been shown to outperform its competitors in both the original DOMINO study [26] and in an independent study that evaluated active module identification methods by randomizing input networks [33].

The third approach is the random walk with restart (RWR) approach. RWR measures the proximity between given seed nodes (i.e., chemical target genes) and all the other nodes in the network, in a way analogous to a person that travels randomly on the connected nodes and sometimes teleports back to one of the seed nodes. RWR thereby considers both the network distance and the network topology. RWR has been widely used in research for disease-associated gene prediction and drug repurposing [34–41]. In the ODAMNet RWR approach, we add the rare disease pathways as nodes connected to their associated genes in the network. Then, RWR can find the rare disease pathway nodes that are proximal to the target genes, which are set as seeds. We use multiXrank v0.1 [27], a Python package that enables RWR on any kind of multilayer network. The biological networks are provided by the user. The input networks must include at least one network of genes/proteins, one rare disease pathways network (each node only connected to itself, disconnected otherwise) and a bipartite network connecting the rare disease pathway nodes to the genes in other networks. The rare disease pathways network and its corresponding bipartite network can be created using the *networkCreation* function available in ODAMNet. The user needs to provide a configuration file for multiXrank in which the seeds file and the network files are stated. multiXrank is run with the default parameters but these can be adjusted through the multiXrank configuration file. Please see our example configuration on https://github.com/MOohTus/ODAMNet/tree/v1.1.0, and multiXrank documentation at https://multixrank-doc.readthedocs.io and [27].

2.2.4. Assistive functions

ODAMNet provides two assistive functions to download and build networks. The *networkDownloading* function allows downloading biological networks from NDex. This function uses the NDEx2 Python Client and the networks' corresponding universally unique identifiers (UUID) provided by the user. The second function, *networkCreation*, allows the creation of the networks necessary for the ODAMNet RWR approach. It creates a rare disease pathways network (each node only connected to itself, disconnected otherwise) and its corresponding bipartite network which connects the rare disease pathway nodes to the genes.

3. Illustrative examples

Vitamin A is a fat-soluble compound that plays an essential role in vision, intercellular communication, mucin production, embryogenesis, cell growth, and cell differentiation [42]. Deficiency or excess of vitamin A might play a role in disease development. Vitamin A deficiency can cause ocular degeneration, diverse changes in epithelial tissues, immune deficits, and excessive mortality from childhood diseases [42]. Excess or systemic intake of vitamin A during pregnancy can cause a spectrum of malformations including ocular, pulmonary, cardiovascular, and urogenital birth defects [42]. In this context, we used ODAMNet to investigate the molecular relationships between vitamin A and rare diseases. We used the automatic data retrieval functions of ODAMNet and its three analysis approaches.

3.1. Data query results

As input, we gave ODAMNet a chemicals file that contains the vitamin A MeSH ID (D014801). ODAMNet automatically retrieved the genes targeted by vitamin A and its 9 descendant molecules from CTD (*directAssociation* parameter set to "False"), which resulted in 7765

Table 1
Top 10 of the 28 rare disease pathways significantly overlapping with vitamin A target genes.

Pathway ID	Pathway name	Pathway size	Intersection size	pAdjusted
WP5087	Malignant pleural mesothelioma	346	146	3.77e-24
WP4298	Acute viral myocarditis	85	45	9.38e-16
WP2447	Amyotrophic lateral sclerosis (ALS)	38	25	1.04e-11
WP5053	Development of ureteric collection system	60	28	2.61e-08
WP4879	Overlap between signal transduction pathways contributing to LMNA laminopathies	57	25	7.80e-07
WP5124	Alzheimer's disease	262	69	1.15e-06
WP3584	MECP2 and associated Rett syndrome	73	28	2.80e-06
WP2059	Alzheimer's disease and miRNA effects	275	70	2.60e-06
WP3995	Prion disease pathway	33	17	3.86e-06
WP4746	Thyroid hormones production and peripheral downstream signaling effects	92	31	1.26e-05

human target genes for 10 chemicals. Then, we kept the chemical - gene associations with at least 2 references. After filtering, we obtained 2143 target genes for vitamin A and its 6 descendants.

ODAMNet retrieved all pathways labeled as "rare disease" in WikiPathways, which resulted in 104 rare disease pathways. It also retrieved all human pathways to construct the background gene set. We obtained 1281 human pathways in total.

The protein-protein interaction (PPI) network (UUID: bfac0486-cefe-11ed-a79c-005056ae23aa) was downloaded from NDEx. This network is the fusion of three datasets (Lit-BM, Hi-Union [43] and APID [44]). It is composed of 15,390 nodes and 131,087 edges.

The molecular complexes network (UUID: 419ae651-cf05–11ed-a79c-005056ae23aa) was also downloaded from NDEx. This network is the fusion of two molecular complex databases (CORUM [45] and HuMap [46]). It is composed of 8497 nodes and 62,073 edges.

The last network downloaded from NDEx is the Reactome pathways network (UUID: b13e9620-cefd-11ed-a79c-005056ae23aa). This network was built based on data derived from Reactome protein-protein interaction data [29]. It is composed of 4598 nodes and 19,292 edges.

All results are available on https://github.com/MOohTus/ODAMNet/tree/v1.1.0. Queries were made on 7 September 2022.

3.2. Overlap analysis results

We first used the ODAMNet overlap analysis approach. ODAMNet retrieved the chemical target genes and pathways as described in Section 3.1 ODAMNet performed an overlap analysis between the 2143 target genes and the 104 rare disease pathways. We obtained a significant overlap (adjusted p-value ≤ 0.05) between target genes and 28 rare disease pathways (Table 1).

The command used to perform this overlap analysis is:

```
odamnet overlap     --chemicalsFile useCases/InputData/chemicalsFile.csv \
                    --directAssociation FALSE \
                    --nbPub 2 \
                    --outputPath useCases/OutputResults_useCase1/
```

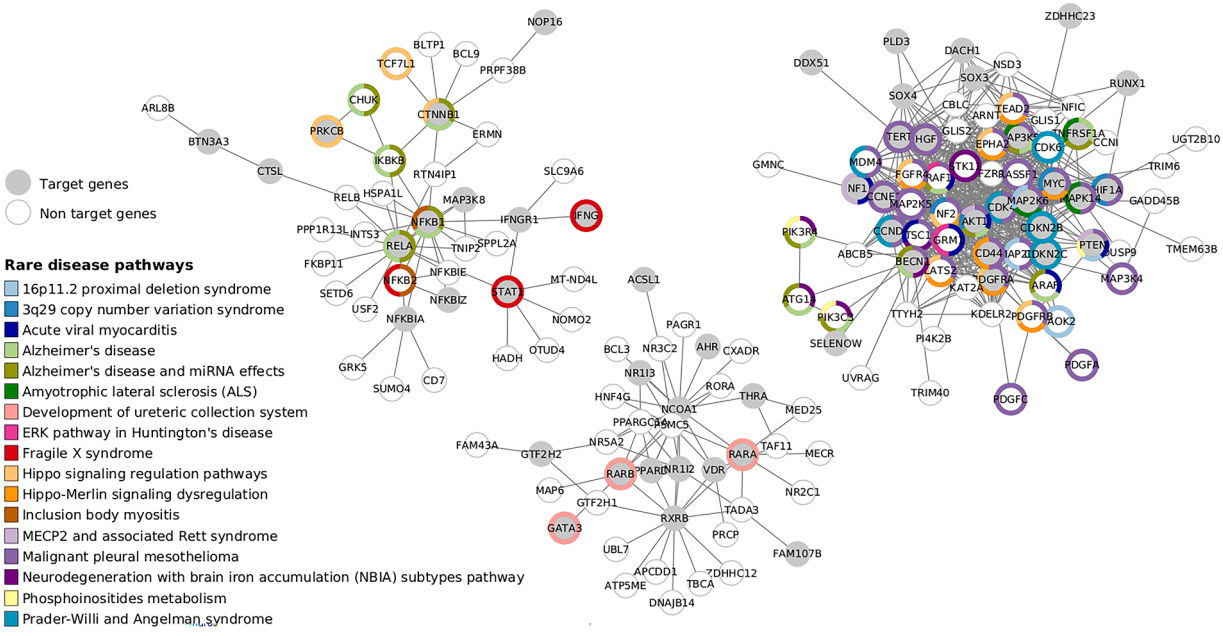

Fig 2. Active module identification with DOMINO followed by overlap analysis between genes in the module and the rare disease pathways. Visualization of 3 over 6 active modules identified by DOMINO. Target genes are in gray and non target genes are in white.

3.3. Active module identification results

To perform an active module identification approach, ODAMNet needs a chemicals file with the MeSH ID of vitamin A and the UUID of the PPI network. For the overlap analysis between the identified active modules and rare disease pathways, ODAMNet extracted pathways from WikiPathways (see Section 3.1 for more details about the query results). Over the 2143 target genes, 1937 were found in the PPI and used by DOMINO as active genes to find active modules. DOMINO found 12 active modules enriched in chemical target genes. Then, ODAMNet performed an overlap analysis between the 12 identified active modules and the 104 rare disease pathways. It found a significant overlap between 6 active modules and 19 rare disease pathways.

In Fig 2, we present 3 active modules as examples. We can observe that the topology of those active modules and the associated rare disease pathways vary. For instance, the active module on the right is highly connected and the genes are involved in many different rare disease pathways. The two other modules are less connected. The genes contained in the active module in the middle are involved only in the *Development of ureteric collection system* rare disease pathway.

The command used to perform this active module identification is:

```
odamnet domino      --chemicalsFile useCases/InputData/chemicalsFile.csv \
                    --directAssociation FALSE \
                    --nbPub 2 \
                    --netUUID bfac0486-cefe-11ed-a79c-005056ae23aa \
                    --outputPath useCases/OutputResults_useCase1/
```

3.4. Random walk with restart results

To run the random walk with restart (RWR) analysis, we provided the vitamin A MeSH ID to automatically retrieve target genes from CTD (see Section 3.1 for more details about the query results). We also provided a configuration file containing the path to the multilayer network. The multilayer network contains a gene multiplex network composed of three layers with the same nodes but different types of edges: the PPI network, the molecular complexes network and a network composed of Reactome pathways. We downloaded these networks from NDEx using the *networkDownloading* function. The multilayer network contains also a layer for the rare disease pathways network. This network layer contains disconnected rare disease pathway nodes (i.e., each node is only connected to itself), which are linked to the gene multiplex network by bipartite gene-disease associations. This rare disease pathways network and its corresponding bipartite network were created using the *networkCreation* function in ODAMNet.

MultiXrank used 2012 chemical target genes available in the multiplex network (over the 2143 genes retrieved from CTD). Then it calculated a RWR score for all the nodes of the multilayer network, which can be gene or rare disease pathway nodes. We selected the top 20 rare disease pathways nodes based on their RWR scores (Table 2).

The command to download the PPI network used in the random walk with restart is:

```
odamnet networkDownloading      --netUUID bfac0486-cefe-11ed-a79c-005056ae23aa
                                --networkFile PPI_network.gr
                                --simple True
```

The command to create rare disease pathway network and its corresponding bipartite network is:

```
odamnet networkCreation     --networksPath useCases/InputData/multiplex/2/ \
                            --networksName WP_RareDiseasesNetwork_fromRequest.gr \
                            --bipartitePath useCases/InputData/bipartite/ \
                            --bipartiteName
Bipartite_WP_RareDiseases_geneSymbols_fromRequest.gr \
                            --outputPath useCases/OutputResults_useCase1
```

The command to perform this random walk with restart is:

```
odamnet multixrank   --chemicalsFile useCases/InputData/chemicalsFile.csv \
                     --directAssociation FALSE \
                     --nbPub 2 \
                     --configPath useCases/InputData/config_minimal_useCase1.yml \
                     --networksPath useCases/InputData/ \
                     --seedsFile useCases/InputData/seeds.txt \
                     --sifFileName resultsNetwork_useCase1.sif \
                     --top 10 \
                     --outputPath useCases/OutputResults_useCase1/
```

3.5. Comparison of the results obtained with the three integrative bioinformatics approaches

To summarize and compare the results from the three approaches, we used orsum [47], a Python package for filtering and integrating enrichment analysis results obtained from multiple studies (Fig 3). We observed that some rare disease pathways are identified as related to vitamin A by all three approaches, e.g. *Malignant pleural mesothelioma* and *Acute viral myocarditis*. Some others are specific to one or two approaches. For instance, *Male infertility* is retrieved with the overlap analysis and the random walk with restart analysis but not with the active module identification approach.

4. Impact

ODAMNet offers a versatile and comprehensive framework for exploring the relationships between chemicals and rare diseases through its integrative approach that combines three databases and three distinct and complementary bioinformatics approaches. These three bioinformatics approaches work at different levels of knowledge: while overlap analysis focuses on the direct involvement of the chemical target genes in the rare disease pathways, active module identification considers both the target genes and the non-target genes in interaction with rare disease pathways. On the other hand, RWR finds the proximity of chemical target genes to the disease pathway nodes using more biological information through a multilayer network.

ODAMNet can be used in multiple ways to generate hypotheses for chemical - rare disease relationships. When investigating a specific chemical, this chemical can be tested for its relationships with multiple pathways, e.g. all the rare disease pathways from WikiPathways. Conversely, when investigating a specific disease, a set of chemicals and disease associated pathways can be determined and tests can be performed to assess their relationships. Furthermore, when the relationship between a chemical and a disease is already known, ODAMNet enables a focused analysis on the shared genes. For all these analyses, we advise the close examination of the target genes, active modules (when active module identification is used), and the rare disease pathways to understand the possibly perturbed mechanisms.

The automatic retrieval of chemical target genes, rare disease pathways and biological networks from relevant databases is the default workflow for ODAMNet. This ensures the usage of up-to-date data and enhances ease-of-use. ODAMNet also accepts direct custom input of target genes, gene sets and biological networks. This has multiple advantages. First of all, it allows reproducibility; target genes and rare disease pathways from specific versions of CTD and WikiPathways can be stored and used in ODAMNet. Second, even if the external resources become temporarily unavailable, ODAMNet can still be run with the previously saved data. And last, this functionality extends the possible use-cases of ODAMNet, as any gene list and gene set data can be incorporated.

NeDRex and Drugst.one are two tools developed for drug repurposing that have similar strategies with ODAMNet (Table 3). These tools integrate multiple databases and provide different bioinformatics analysis methods. The main difference between ODAMNet and the two other tools is the targeted research question. ODAMNet is developed to study chemical - rare disease relationships and for this purpose it integrates the three most relevant databases. ODAMNet also has the advantage of allowing custom data input.

Table 2
Top 10 rare disease pathways identified by the ODAMNet RWR approach for vitamin A.

Pathway ID	Pathway name	RWR score
WP5087	Malignant pleural mesothelioma	2.85e-03
WP4673	Male infertility	9.02e-04
WP2059	Alzheimer's disease and miRNA effects	7.76e-04
WP5124	Alzheimer's disease	7.76e-04
WP4298	Acute viral myocarditis	7.69e-04
WP3584	MECP2 and associated Rett syndrome	6.03e-04
WP4746	Thyroid hormones production and peripheral downstream signaling effects	5.83e-04
WP4549	Fragile X syndrome	5.77e-04
WP5224	2q37 copy number variation syndrome	5.66e-04
WP4657	22q11.2 copy number variation syndrome	5.62e-04

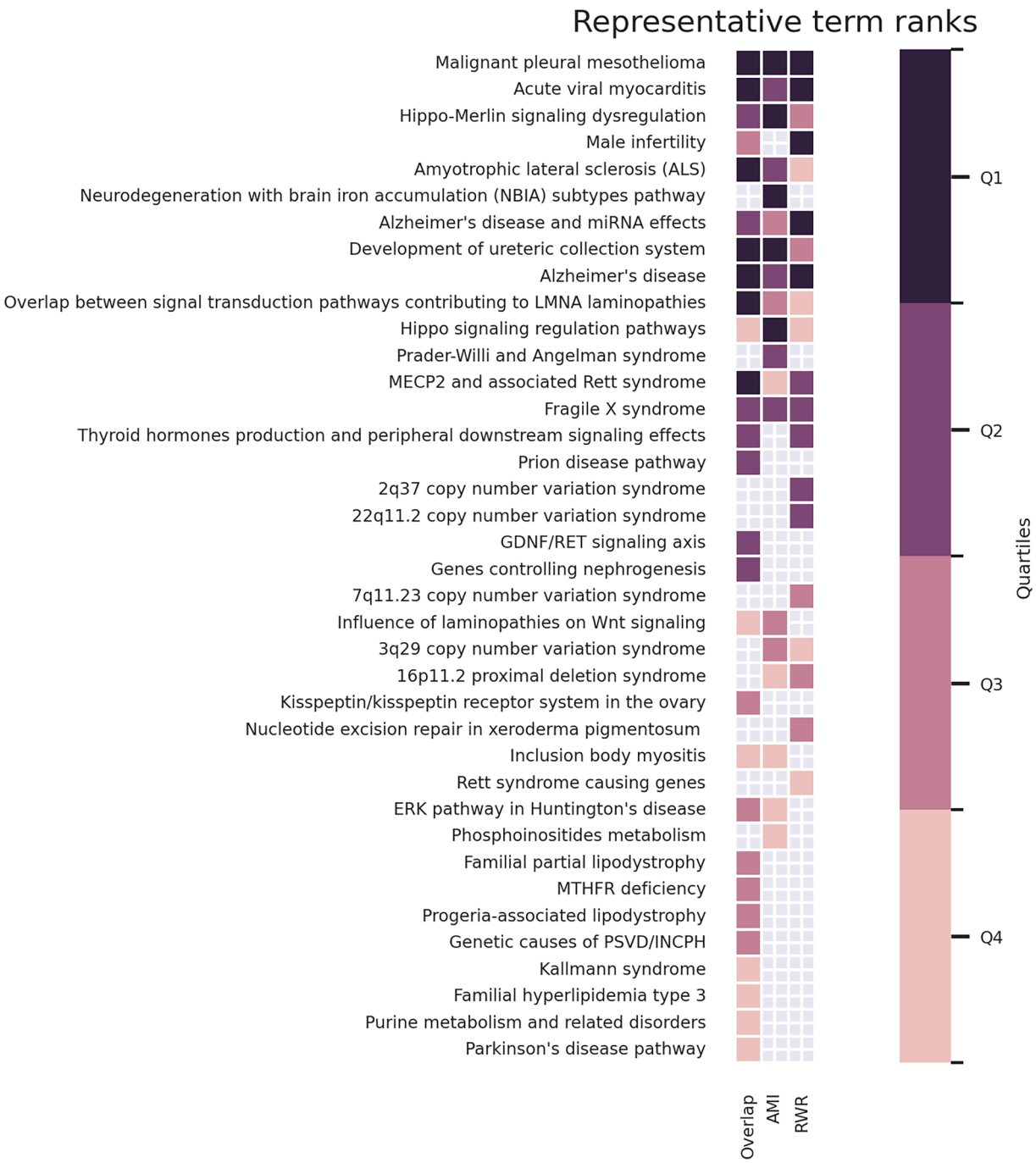

Fig 3. Summarization of rare disease pathways that are found to be related to vitamin A using overlap analysis (Overlap), active module identification (AMI) and random walk with restart (RWR) analysis. This heatmap is created by orsum [47].

Table 3
Comparison of ODAMNet with NeDRex and Drugst.One.

	ODAMNet	NeDRex	Drugst.One
Research question	Chemical - rare disease relationships	Drug repurposing	Drug repurposing
Integrated databases	CTD, WikiPathways, NDEx	OMIM [48], DisGeNET [49], UniProt [50], NCBI gene info [51], IID [52], MONDO [53], DrugBank [54], Reactome, DrugCentral [55]	APID, BioGRID [61], ChEMBL [62], CTD, DGIdb [63], DisGeNET, DrugBank, DrugCentral, GTEx [64], IID, IntAct [65], NeDRex, OMIM, STRING [66]
Module identification algorithms	DOMINO	BiCoN [56], DIAMoND [57], ROBUST [58], MuST [59]	Multi-Level Steiner Trees [67], KeyPathwayMiner [68]
Network extension algorithms	(For ranking of rare disease pathways) Random walk with restart with multilayer network support (multiXrank)	(For ranking of drugs) TrustRank [60], Closeness centrality	(For ranking of drugs or drug targets) TrustRank, Harmonic Centrality, Degree Centrality, Betweenness Centrality, Network Proximity [69]
Enrichment analysis	Embedded code for hypergeometric test	–	(Outlinks to) g:Profiler [70], DIGEST [71], NDEx integrated query
Custom PPI network	Yes	No	Yes
Custom annotation data	Yes (input GMT file)	No	Yes (as provided by the outlinked enrichment tools)
Access	Python package	Cytoscape app, RESTful api, Neo4j endpoint	Python package, web interface, web plugin

5. Conclusions

Environmental factors, in particular chemical substances, are known or suspected to be playing a role in multiple diseases. There is growing knowledge on chemical - gene/protein interactions and rare disease molecular mechanisms. We designed ODAMNet to bridge the gap between these knowledge sources. Using three different and complementary approaches, ODAMNet can generate knowledge-based hypotheses for chemical - rare disease relationships and their underlying mechanisms. Overall, ODAMNet's comprehensive approach, flexibility in analysis design, and utilization of up-to-date or custom input data should allow its usage in a wide variety of contexts.

Fundings

MT received funding from the European Union's Horizon 2020 research and innovation programme under the EJP RD COFUND-EJP No 825575. OO received funding from the «Priority Research Programme on Rare Diseases» of the French Investments for the Future Programme.

CRediT authorship contribution statement

Morgane Térézol: Methodology, Software, Visualization, Writing – original draft, Writing – review & editing. **Anaïs Baudot:** Methodology, Project administration, Supervision, Writing – review & editing. **Ozan Ozisik:** Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

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

Data availability

All codes and data are available on https://github.com/MOohTus/ODAMNet.

Acknowledgments

We thank the members of the European Joint Programme on Rare Diseases for their support and discussions. We thank Nazli Sila Kara for being one of the first users of ODAMNet and giving much useful feedback. We thank Cécile Beust who created the biological networks and uploaded them to NDEx.

References

[1] Gerstman BB. Epidemiology kept simple: an introduction to traditional and modern epidemiology. 3rd ed. Ltd: John Wiley & Sons; 2013.
[2] Olden K, Wilson S. Environmental health and genomics: visions and implications. Nat Rev Genet 2000;1:149–53. https://doi.org/10.1038/35038586.
[3] Chaste P, Leboyer M. Autism risk factors: genes, environment, and gene-environment interactions. Dialogues Clin Neurosci 2012;14:281–92. https://doi.org/10.31887/DCNS.2012.14.3/pchaste.
[4] Abegunde AT, Muhammad BH, Bhatti O, Ali T. Environmental risk factors for inflammatory bowel diseases: evidence based literature review. World J Gastroenterol 2016;22:6296–317. https://doi.org/10.3748/wjg.v22.i27.6296.
[5] Ananthakrishnan AN, Bernstein CN, Iliopoulos D, Macpherson A, Neurath MF, Ali RAR, et al. Environmental triggers in IBD: a review of progress and evidence. Nat Rev Gastroenterol Hepatol 2018;15:39–49. https://doi.org/10.1038/nrgastro.2017.136.
[6] Bhatnagar A. Environmental determinants of cardiovascular disease. Circ Res 2017;121:162–80. https://doi.org/10.1161/CIRCRESAHA.117.306458.
[7] Münzel T, Sørensen M, Lelieveld J, Hahad O, Al-Kindi S, Nieuwenhuijsen M, et al. Heart healthy cities: genetics loads the gun but the environment pulls the trigger. Eur Heart J 2021;42:2422–38. https://doi.org/10.1093/eurheartj/ehab235.
[8] Nicolaou N, Renkema KY, Bongers EMHF, Giles RH, Knoers NVAM. Genetic, environmental, and epigenetic factors involved in CAKUT. Nat Rev Nephrol 2015;11:720–31. https://doi.org/10.1038/nrneph.2015.140.
[9] Murugapoopathy V, Gupta IR. A primer on congenital anomalies of the kidneys and urinary tracts (CAKUT). Clin J Am Soc Nephrol 2020;15:723–31. https://doi.org/10.2215/CJN.12581019.
[10] Oskarsson B, Horton DK, Mitsumoto H. Potential environmental factors in amyotrophic lateral sclerosis. Neurol Clin 2015;33:877–88. https://doi.org/10.1016/j.ncl.2015.07.009.
[11] Meltzer EB, Noble PW. Idiopathic pulmonary fibrosis. Orphanet J Rare Dis 2008;3:8. https://doi.org/10.1186/1750-1172-3-8.
[12] Rodríguez-Olivas AO, Hernández-Zamora E, Reyes-Maldonado E. Legg-Calvé-Perthes disease overview. Orphanet J Rare Dis 2022;17:125. https://doi.org/10.1186/s13023-022-02275-z.
[13] Zheng Y, Zhen Z, Pearson T, Zhao J, Hu H, Prosperi M. Design and methodology challenges of environment-wide association studies: a systematic review. Environ Res 2020;183:109275. https://doi.org/10.1016/j.envres.2020.109275.
[14] Ozisik O, Ehrhart F, Evelo CT, Mantovani A, Baudot A. Overlap of vitamin A and vitamin D target genes with CAKUT-related processes. F1000Res 2021;10:395. https://doi.org/10.12688/f1000research.51018.2.
[15] Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative toxicogenomics database (CTD): update 2023. Nucleic Acids Res 2023;51:D1257–62. https://doi.org/10.1093/nar/gkac833.
[16] Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res 2021;49:D613–21. https://doi.org/10.1093/nar/gkaa1024.
[17] Pratt D, Chen J, Welker D, Rivas R, Pillich R, Rynkov V, et al. NDEx, the network data exchange. Cell Syst 2015;1:302–5. https://doi.org/10.1016/j.cels.2015.10.001.
[18] Rodchenkov I, Babur O, Luna A, Aksoy BA, Wong JV, Fong D, et al. Pathway commons 2019 update: integration, analysis and exploration of pathway data. Nucleic Acids Res 2019:gkz946. https://doi.org/10.1093/nar/gkz946.

[19] Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. Nat Methods 2016;13:966–7. https://doi.org/10.1038/nmeth.4077.
[20] Türei D, Valdeolivas A, Gul L, Palacio-Escat N, Klein M, Ivanova O, et al. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol. Syst. Biol. 2021;17:e9923. https://doi.org/10.15252/msb.20209923.
[21] Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife 2017;6:e26726. https://doi.org/10.7554/eLife.26726.
[22] Sadegh S, Skelton J, Anastasi E, Bernett J, Blumenthal DB, Galindez G, et al. Network medicine for disease module identification and drug repurposing with the NeDRex platform. Nat Commun 2021;12:6848. https://doi.org/10.1038/s41467-021-27138-2.
[23] Maier A., Hartung M., Abovsky M., Adamowicz K., Bader G.D., Baier S., et al. Drugst.One - A plug-and-play solution for online systems medicine and network-based drug repurposing. ArXiv [Preprint] 2023.
[24] Renaux A, Terwagne C, Cochez M, Tiddi I, Nowé A, Lenaerts T. A knowledge graph approach to predict and interpret disease-causing gene interactions. BMC Bioinform 2023;24:324. https://doi.org/10.1186/s12859-023-05451-5.
[25] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data 2023;10:67. https://doi.org/10.1038/s41597-023-01960-3.
[26] Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. Mol Syst Biol 2021;17:e9593. https://doi.org/10.15252/msb.20209593.
[27] Baptista A, Gonzalez A, Baudot A. Universal multilayer network exploration by random walk with restart. Commun Phys 2022;5(1):170. https://doi.org/10.1038/s42005-022-00937-9.
[28] Pillich RT, Chen J, Churas C, Liu S, Ono K, Otasek D, et al. NDEx: accessing network models and streamlining network biology workflows. Curr Protoc 2021;1:e258. https://doi.org/10.1002/cpz1.258.
[29] Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res 2022;50:D687–92. https://doi.org/10.1093/nar/gkab1028.
[30] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 2000;25:25–9. https://doi.org/10.1038/75556.
[31] Ontology Consortium Gene, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The gene ontology knowledgebase in 2023. Genetics 2023;224:iyad031. https://doi.org/10.1093/genetics/iyad031.
[32] Levi H, Rahmanian N, Elkon R, Shamir R. The DOMINO web-server for active module identification analysis. Bioinformatics 2022;38:2364–6. https://doi.org/10.1093/bioinformatics/btac067.
[33] Lazareva O, Baumbach J, List M, Blumenthal DB. On the limits of active module identification. Brief Bioinform 2021;22:bbab066. https://doi.org/10.1093/bib/bbab066.
[34] Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet 2008;82:949–58. https://doi.org/10.1016/j.ajhg.2008.02.013.
[35] Li Y, Patra JC. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. Bioinformatics 2010;26:1219–24. https://doi.org/10.1093/bioinformatics/btq108.
[36] Erten S, Bebek G, Ewing RM, Koyutürk MDA DA. Degree-aware algorithms for network-based disease gene prioritization. BioData Min 2011;4:19. https://doi.org/10.1186/1756-0381-4-19.
[37] Smedley D, Köhler S, Czeschik JC, Amberger J, Bocchini C, Hamosh A, et al. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics 2014;30:3215–22. https://doi.org/10.1093/bioinformatics/btu508.
[38] Jiang R. Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol 2015;7:214–30. https://doi.org/10.1093/jmcb/mjv008.
[39] Zhang H, Ferguson A, Robertson G, Jiang M, Zhang T, Sudlow C, et al. Benchmarking network-based gene prioritization methods for cerebral small vessel disease. Brief Bioinform 2021;22:bbab006. https://doi.org/10.1093/bib/bbab006.
[40] Luo H, Wang J, Li M, Luo J, Ni P, Zhao K, et al. Computational drug repositioning with random walk on a heterogeneous network. IEEE ACM Trans Comput Biol Bioinf 2019;16:1890–900. https://doi.org/10.1109/TCBB.2018.2832078.
[41] Ulgen E, Ozisik O, Sezerman OU. PANACEA: network-based methods for pharmacotherapy prioritization in personalized oncology. Bioinformatics 2023;39:btad022. https://doi.org/10.1093/bioinformatics/btad022.
[42] Solomons N.W. Vitamin A. In: Erdman JW, Macdonald IA, Zeisel SH, editors Present knowledge in nutrition. 10th ed. 2012, p. 149–84. doi:10.1002/9781119946045.ch11.
[43] Luck K, Kim DK, Lambourne L, Spirohn K, Begg BE, Bian W, et al. A reference map of the human binary protein interactome. Nature 2020;580:402–8. https://doi.org/10.1038/s41586-020-2188-x.
[44] Alonso-López D, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M, et al. APID database: redefining protein-protein interaction experimental evidences and binary interactomes. Database 2019;2019:baz005. https://doi.org/10.1093/database/baz005 (Oxford).
[45] Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res 2019;47:D559–63. https://doi.org/10.1093/nar/gky973.
[46] Drew K, Wallingford JB, Marcotte EM. hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. Mol Syst Biol 2021;17:e10016. https://doi.org/10.15252/msb.202010016.
[47] Ozisik O, Térézol M, Baudot A. orsum: a Python package for filtering and comparing enrichment analyses using a simple principle. BMC Bioinform 2022;23:293. https://doi.org/10.1186/s12859-022-04828-2.
[48] Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype–gene relationships. Nucleic Acids Res 2019;47:D1038–43. https://doi.org/10.1093/nar/gky1151.
[49] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2019:gkz1021. https://doi.org/10.1093/nar/gkz1021.
[50] The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15. https://doi.org/10.1093/nar/gky1049.
[51] Maglott D. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 2004;33:D54–8. https://doi.org/10.1093/nar/gki031.
[52] Kotlyar M, Pastrello C, Malik Z, Jurisica I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. Nucleic Acids Res 2019;47:D581–9. https://doi.org/10.1093/nar/gky1037.
[53] Mungall CJ, McMurry JA, Köhler S, Balhoff JP, Borromeo C, Brush M, et al. The Monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res 2017;45:D712–22. https://doi.org/10.1093/nar/gkw1128.
[54] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res 2018;46:D1074–82. https://doi.org/10.1093/nar/gkx1037.
[55] Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL, Stathias V, et al. DrugCentral 2018: an update. Nucleic Acids Res 2019;47:D963–70. https://doi.org/10.1093/nar/gky963.
[56] Lazareva O, Canzar S, Yuan K, Baumbach J, Blumenthal DB, Tieri P, et al. BiCoN: network-constrained biclustering of patients and omics data. Bioinformatics 2021;37:2398–404. https://doi.org/10.1093/bioinformatics/btaa1076.
[57] Ghiassian SD, Menche J, Barabási AL. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLoS Comput Biol 2015;11:e1004120. https://doi.org/10.1371/journal.pcbi.1004120.
[58] Bernett J, Krupke D, Sadegh S, Baumbach J, Fekete SP, Kacprowski T, et al. Robust disease module mining via enumeration of diverse prize-collecting Steiner trees. Bioinformatics 2022;38:1600–6. https://doi.org/10.1093/bioinformatics/btab876.
[59] Sadegh S, Matschinske J, Blumenthal DB, Galindez G, Kacprowski T, List M, et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. Nat Commun 2020;11:3518. https://doi.org/10.1038/s41467-020-17189-2.
[60] Gyöngyi Z, Garcia-Molina H, Pedersen J. Combating Web Spam with TrustRank. In: Proceedings of the VLDB Conference. Elsevier; 2004. p. 576–87. https://doi.org/10.1016/B978-012088469-8.50052-8.
[61] Oughtred R, Rust J, Chang C, Breitkreutz B, Stark C, Willems A, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 2021;30:187–200. https://doi.org/10.1002/pro.3978.
[62] Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 2019;47:D930–40. https://doi.org/10.1093/nar/gky1075.
[63] Freshour SL, Kiwala S, Cotto KC, Coffman AC, McMichael JF, Song JJ, et al. Integration of the Drug–Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. Nucleic Acids Res 2021;49:D1144–51. https://doi.org/10.1093/nar/gkaa1084.
[64] Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, et al. A Novel Approach to High-Quality Postmortem Tissue Procurement: the GTEx Project. Biopreserv Biobank 2015;13:311–9. https://doi.org/10.1089/bio.2015.0032.
[65] Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, Broackes-Carter F, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2014;42:D358–63. https://doi.org/10.1093/nar/gkt1115.
[66] Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res 2023;51:D638–46. https://doi.org/10.1093/nar/gkac1000.
[67] Ahmed R, Angelini P, Sahneh FD, Efrat A, Glickenstein D, Gronemann M, et al. Multi-level Steiner trees. ACM J Exp Algorithmics 2019;24:1–22. https://doi.org/10.1145/3368621.
[68] List M, Alcaraz N, Dissing-Hansen M, Ditzel HJ, Mollenhauer J, Baumbach J. KeyPathwayMinerWeb: online multi-omics network enrichment. Nucleic Acids Res 2016;44:W98–104. https://doi.org/10.1093/nar/gkw373.
[69] Guney E, Menche J, Vidal M, Barábasi AL. Network-based *in silico* drug efficacy screening. Nat Commun 2016;7:10331. https://doi.org/10.1038/ncomms10331.
[70] Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res 2023;51:W207–12. https://doi.org/10.1093/nar/gkad347.
[71] Adamowicz K, Maier A, Baumbach J, Blumenthal DB. Online in silico validation of disease and gene sets, clusterings or subnetworks with DIGEST. Brief Bioinform 2022;23:bbac247. https://doi.org/10.1093/bib/bbac247.