REPOTRIAL

An in silico-based approach to improve the efficacy and precision of drug REPurpOsing TRIALs for a mechanism-based patient cohort with predominant cerebro-cardiovascular phenotypes

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

Deliverable D1.10
Refined diseasome and drugome 3.5

Work Package WP1
Data integration and in-silico trial prediction

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Document information

Grant Agreement Number: 777111				**Acronym: REPO-TRIAL**
Full title	An in silico-based approach to improve the efficacy and precision of drug REPurpOsing TRIALs for a mechanism-based patient cohort with predominant cerebro-cardiovascular phenotypes			
Topic	In-silico trials for developing and assessing biomedical products			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 February 2018	**Duration**	60 months	
Project URL	https://repo-trial.eu/			
EU Project Officer	Dusan Sandor, Programme Officer, European Commission			
Project Coordinator	Prof. Dr. Harald H.H.W. Schmidt, Universiteit Maastricht (UM)			
Deliverable	D1.10 Refined diseasome and drugome 3.5			
Work Package	WP1 Data integration and in-silico trial prediction			
Date of Delivery	**Contractual**	31/07/2022 (M54)	**Actual**	30/07/2022 (M54)
Nature	Other	**Dissemination Level**	PUBLIC	
Lead Beneficiary	13 UHAM			
Responsible Author(s)	Sepideh Sadegh (UHAM)			
	Andreas Maier (UHAM)			
Keywords	Diseasome, Drugome, Comorbiditome, Network medicine, Drug repurposing			

History of changes

Version	Date	Contributions	Contributors (name and institution)
V0.1	18/07/2022	First draft	Sepideh Sadegh (UHAM), Andreas Maier (UHAM)
V0.2	20/07/2022	Comments	Emre Guney (STALICLA), Cristian Nogales (UM)
V0.3	22/07/2022	Draft	Sepideh Sadegh (UHAM), Andreas Maier (UHAM)
V1	22/07/2022	Final version	Sepideh Sadegh (UHAM)
V1	30/07/2022	Approval	Harald Schmidt (UM)
V1	30/07/2022	Submission	Miriam Simon (concentris)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

Table of Content

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

1. Objectives of the deliverable based on the Description of Action (DoA)

The objectives of this deliverable are determined as follows:

1. **Refine** current gene-based diseasome by constructing multi-level disease-disease networks (**diseasome** v3.5) based on the public molecular data including disease-variant associations, disease-symptom, drug-indication as well as the comorbidity data from health records.
2. Build a **comorbidity-based diseasome** (**comorbiditome**), using clinical comorbidity data (common clinical co-appearance of two disease phenotypes) from health records.
3. Map networks to a unified disease ontology which is required for (1) and (2).
4. **Refine** the current target-based **drugome** by including drug-indication data.
5. Devise graph-based methods to analyse the (dis)similarity and accordance of different diseasomes and drugomes.
6. Implement the established graph-based methods in (5).
7. Investigate the impact of disease term granularity originated from different disease ontologies on the results of conducted network-based analyses.
8. Assess if the results of network medicine approaches in a broad sense are reliable when they are based on the data annotated with mechanistically inadequate disease definitions, i.e. phenotype-based disease definitions.

2. Executive Summary

- Methodology:

  A Diseasome/drugome is a network consisting of nodes, representing diseases/drugs, and edges, representing the relationships between diseases/drugs. The construction of diseasomes/drugomes is based on the bipartite networks of disease/drug-*T* relationships, where *T* is the associated data of different types. In the context of the REPO-TRIAL project, constructing multi-level diseasomes/drugomes in which the edges represent various types of similarity between diseases/drugs and assessing the conformity between them could provide further insights to pathomechanisms underlying complex diseases and repurposing candidates.

  For the task of network construction we used the NeDRexDB data (Sadegh *et al.*, 2021) integrated from public databases such as OMIM (Amberger *et al.*, 2019), DisGeNET (Piñero

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

et al., 2020), HPO (Köhler *et al.*, 2021), DrugBank (Wishart *et al.*, 2018), DrugCentral (Avram *et al.*, 2021), CTD (Davis *et al.*, 2021), IID (Kotlyar *et al.*, 2019), and UniProt (UniProt Consortium, 2019). Furthermore, for the construction of comorbiditome we used the health records of around 140K patients from the Estonian Biobank on the pairwise comorbidity of diseases. To continue with the analyses of the constructed networks, they all needed to be mapped into the same disease ID namespace. We decided to use the Monarch Disease Ontology (MONDO) as the primary identifier for diseases, as the mapping between MONDO and other identifiers (e.g., the Unified Medical Language Systems (UMLS), used by DisGeNET) is more complete than other disease identifiers. Since the source of comorbidity data (Estonian Biobank) uses ICD-10 codes, we mapped all the networks also to this namespace and repeated the same analyses. The disease ID mapping task was carried out by (1) using available mapping sources such as MONDO and OXO ontology mapping (https://www.ebi.ac.uk/spot/oxo/) and (2) a manual mapping effort by UHAM, UNEW and UM partners to check the validity of the existing mappings and adding the missing ones.

To check the similarity between the networks both on a global and local scale, we used two versions of Graph Edit Distance (GED) (Bunke and Allermann, 1983; Sanfeliu and Fu, 1983) measure using uniform and rank-based edge editing costs. We also used shortest path distance to evaluate the closeness of diseases pair-wise in the integrated PPI network (interactome). The hypothesis was that disease pairs with some levels of similarity based on the genetic signature, symptoms or comorbidity, are located closer in the interactome than the ones without similarity. In a similar fashion, we evaluated the closeness of drugs pair-wise in the PPI network with the hypothesis that drug pairs with some levels of similarity based on the target or indication, are located closer in the integrated PPI network than the ones without similarity. We also evaluated the closeness of drug-disease pairs in the interactome and the hypothesis was that the drug-disease pairs having drug-indication associations are located closer in the integrated interactome than the ones without such associations.

- Results:

  Apart from the gene-based approach, four new versions of diseasomes (variant-based, symptom-based, drug-based, comorbidity-based) and one new drugome (indication-based) were constructed and further systematically compared. All the constructed drugomes and diseasomes including the comorbiditome are publicly available on GitHub: https://github.com/repotrial/graphsimgt

  The networks can be loaded in any network visualization tool such as Cytoscape and be explored further.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

For all evaluated pairs of networks (both in MONDO and in ICD-10 namespace), we obtained smaller **global GEDs** for the original diseaseomes, drugomes, or drug-disease networks than for randomised counterparts, leading to empirical *P*-values which are significant at 0.001 level.

For **shortest path distance analyses** between disease-disease, drug-drug, and drug-disease pairs in the PPI-included networks of disease-gene-gene-disease, drug-protein-protein-drug, and disease-protein-protein-drug networks, we observed that shortest path distances are significantly shorter for node pairs that are directly connected by a link in the reference networks. In particular, the results show (1) that distances between diseases that are connected by edges in diseasomes constructed based on comorbidities, shared drugs, shared symptoms, or shared genetic variants are significantly shorter than distances between diseases without such edges; (2) that distances of disease-drug pairs with shared indication edges are significantly shorter than distances of disease-drug pairs without such edges; and (3) that distances between drug pairs with shared indication are significantly shorter than distances for drug pairs without shared indications. distances between disease genes, disease gene & drug targets and drug targets.

The overview of the results of the **local GED analyses** for different disease namespaces shows that the comparisons performed in ICD-10 namespace (at three-character level) led to more significant similarities than the ones performed in MONDO namespace. Furthermore, we computed local empirical *P*-values individually for nodes based on local GEDs. The fractions of significant local empirical *P*-values at 0.05 level show that, for a substantial fraction of disease nodes, local neighbourhoods of diseases are not preserved significantly better than expected by chance across the different diseasomes.

In sum, our global analyses provide solid evidence for the global validity of the network medicine paradigm while our local analyses only provide weak evidence for the local scale hypothesis, indicating the network medicine tends to produce locally blurred results.

- Distinctive features and progress beyond the state-of-the-art:
  While gene-based disease-disease similarity networks, so-called diseasomes, are not novel – for example, Goh et al. published such a network for the first time (Goh *et al.*, 2007) – our multifaceted version of the diseasome builds on the works of other researchers in the network medicine field by including various types of data in addition to genetic signatures, such as symptoms, comorbidity, and drugs. Comorbidity and symptom data are specifically very important since they can potentially capture environmental and lifestyle-influenced factors into the play. Whereas the target-based drug-drug network, so-called drugome was first published by Udrescu et al. (Udrescu *et al.*, 2020), our extended version of the drugome builds on previous works by integrating drug-indication data. To our knowledge, global and local

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

similarity analyses on the multitude of diseasomes (including comorbiditome), drugomes and disease-drug networks have not been done before. With our analyses, we also point out the caveats of using large-scale data in network medicine and propose a way forward by using more close-up methods where the study is focused on specific diseases and starts with molecular data for well-characterised patient cohorts.

3. Introduction (Challenge)

In network medicine, multi-level health data modelled as complex networks is mined to identify causal mechanisms of complex diseases. In other words, some network medicine approaches are based on the following intuitive idea: Assume that two diseases d_1 and d_2 share an unknown causal molecular mechanism. Then this joint mechanism should lead to similarities in health data associated with d_1 and d_2. And vice versa, i.e., if there are similarities in the data, the diseases share common mechanisms. For instance, we would expect that the diseases d_1 and d_2 have similar profiles of disease-associated genes, that they exhibit high comorbidity, that they lead to similar symptoms, and that they can be treated by similar drugs. To uncover the unknown mechanism, network medicine therefore makes the following core assumption (Goh *et al.*, 2007): If health data associated with two diseases d_1 and d_2 are sufficiently similar, this constitutes prior evidence for the conjecture that d_1 and d_2 share a causal molecular mechanism. Based on this assumption, some network medicine approaches generate complex networks (so-called “diseasomes”), where nodes are diseases and two diseases d_1 and d_2 are linked by an edge if their associated health data are sufficiently similar. The diseasomes are then mined for densely connected regions of interest, which are further investigated for (joint) disease mechanisms.

Since the pioneering articles by Goh et al. (Goh *et al.*, 2007) and Barabási et al. (Barabási, Gulbahce and Loscalzo, 2011), network medicine has developed into an increasingly mature research field with its own dedicated journals (Baumbach and Schmidt, 2018) and associations (Maron *et al.*, 2020). Various studies have generated evidence for the validity of its overall approach: For instance, Menche et al. (Menche *et al.*, 2015) demonstrated that disease-associated genes form so-called disease modules, i.e., highly connected subnetworks within protein-protein interaction (PPI) networks, and that biological and clinical similarity of two diseases results in significant topological proximity of these modules. In a similar vein, Iida et al. (Iida, Iwata and Yamanishi, 2020) showed that shared therapeutic targets or shared drug indications are correlated with high topological module proximity. Guney et al. (Guney *et al.*, 2016) and Cheng at al. (Cheng *et al.*, 2018) showed that the network-based separation between drug targets and disease modules is indicative of drug efficacy. Cheng et al. (Cheng, Kovács and Barabási, 2019) and Zhou et al. (Zhou *et al.*, 2020) found that

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777111.

FDA-approved drug combinations are proximal to each other and to the modules of the targeted diseases in the interactome.

Although network medicine approaches have led to some new mechanistic insights into complex diseases (AbdulHameed *et al.*, 2014; Samokhin *et al.*, 2018; Wang and Loscalzo, 2018; Halu *et al.*, 2019), large-scale adoption in the biomedical sciences and major translational breakthroughs are still pending. As part of the concluding deliverable of WP1 for the REPO-TRIAL project, and in addition to constructing the refined diseasome and drugome, it was necessary to analyse these networks and investigate the reasons for this translational underperformance and propose ideas how to address this issue in the network medicine discipline. Here, we also investigated to which extent currently available data do indeed back the core assumption of network medicine.

In order to answer this question with quantitative means, we derive the following testable hypotheses from the core assumption (see “Methods” for an argument that these hypotheses indeed follow from the core assumption):

> **Global scale hypothesis:** For all disease association data types T_1 and T_2 that contain useful information about diseases, diseasomes G_1 and G_2 constructed based on T_1 and T_2 are pairwise more similar than expected by chance.
>
> **Local scale hypothesis:** For all disease association data types T_1 and T_2 that contain useful information about disease and any disease d_i appearing in the association data, the direct neighbourhood of d_i in the diseasomes G_1 and G_2 constructed based on T_1 and T_2 are pairwise more similar than expected by chance.

To test these two hypotheses, we constructed diseasomes based on (1) disease-gene associations, (2) disease-variant associations, (3) comorbidity data, (4) symptom data, and (5) drug-indication data. Moreover, we constructed drug-disease and drug-drug networks (so-called “drugomes”) based on drug-indication and drug-target data. We then compared all pairs of diseasomes, drugomes, and drug-disease networks both on a global and on a local scale, using customised versions of the graph edit distance (GED). We also evaluated how competing disease ontologies of different granularity affect the results, by carrying out the analyses using MONDO IDs (finer granularity) and ICD-10 three-character codes (coarser granularity) as node IDs in the constructed networks, respectively.

In line with the findings of the prior studies summarised above (Menche *et al.*, 2015; Guney *et al.*, 2016; Cheng *et al.*, 2018; Cheng, Kovács and Barabási, 2019; Iida, Iwata and Yamanishi, 2020; Zhou *et al.*, 2020), our analyses provide solid evidence for the global-scale hypothesis. However, they only partially support the local-scale hypothesis.

We hypothesise that one important reason for this “local blurriness” of network medicine is that the current symptom- and organ-based disease definitions largely do not reflect causal mechanisms (Nogales *et al.*, 2022). This leads to unspecific data on all scales, because in the data used as input

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777111.

by network medicine approaches, these mechanistically inappropriate disease definitions are employed for annotation purposes. We therefore advocate that, in order to deliver on the promises made by network medicine discipline, experts in this field should work hand in hand with biomedical researchers and together aim at a **mechanistically grounded disease ontology**.

4. Methodology

4.1. *Data integration*

As shown in Table 1, the data sources used to create the different networks use a range of competing disease vocabularies to refer to diseases. We hence had to map these vocabularies to a common namespace to be able to investigate network (dis-)similarities. The similarity analyses were performed in both MONDO (Monarch Disease Ontology) and ICD-10 namespaces and disease ID mapping was carried out *via* the two-step approach implemented in the NeDRex platform. First, MONDO contains mappings between its own disease vocabulary and various other vocabularies, including OMIM, MeSH, and ICD-10. Then, mappings between several vocabularies and ICD-10 could be achieved by mapping disease terms to MONDO, followed by mapping MONDO to ICD-10. For all pairwise analyses, the two compared networks were aligned before computing GEDs, i.e., only the nodes contained in both of them were taken into account.

Table 1. Data sources used for network construction.

Data source	Used disease vocabularies	Data type	Networks constructed from data source
HPO	OMIM, Orphanet (ORPHA)	Disease-symptom	Symptom-based diseasome
DisGeNET	Concept Unique Identifiers of Unified Medical Language System (UMLS CUI)	Disease-gene, disease-variant	Gene-based diseasome, variant-based diseasome, disease-gene-gene-disease network, drug-protein-protein-drug network, drug-protein-protein-disease network
OMIM	OMIM	Disease-gene	Gene-based diseasome, disease-gene-gene-disease network, drug-protein-protein-disease network
DrugCentral	SNOMED Clinical Terms (SNOMEDCT)	Drug-target, drug-indication	Target-based drugome, indication-based drugome and drug-disease network, drug-protein-protein-drug network, drug-protein-protein-disease network
DrugBank	–	Drug-target	Target-based drugome, drug-protein-protein-drug network, drug-protein-protein-disease network
CTD	MeSH	Drug-indication	Drug-disease network, indication-based drugome
IID	–	Protein-protein interaction	Disease-gene-gene-disease network, drug-protein-protein-drug network, drug-protein-protein-disease network

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

UniProt	–	Gene-protein	Drug-protein-protein-disease network
Estonian Biobank	ICD-10 (three- and four-character codes)	Comorbidity data	Comorbidity-based diseasome

Additionally, further data harmonisation steps were carried out: Since HPO contains both general and specific terms, we pruned the data by removing very general symptom terms, using the existing hierarchy in HPO. More specifically, we decomposed the generated hierarchical phenotype network into its levels and removed the terms from the top three levels.

The diagnoses in around 140K patients records available in the Estonian Biobank are encoded in ICD-10 namespace, and the records contain both three- and four-character ICD-10 codes. In order to generate uniform data, we therefore truncated and aggregated all four-character codes to the corresponding three-character level. Moreover, we removed diseases with incidence below five from the data, as well as the codes from the ICD-10 chapters XV (“Pregnancy, childbirth and the puerperium”), XVI (“Certain conditions originating in the perinatal period”), XVIII (“Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified”), XIX (“Injury, poisoning and certain other consequences of external causes”), XX (“External causes of morbidity and mortality”), XXI (“Factors influencing health status and contact with health services”), and XXII (“Codes for special purposes”).

We only used the curated gene-disease associations from DisGeNET and the associations based on text mining are not included. This is also explained in the D1.7 report, where we described the data integrated in NeDRexDB. The Drug-Disease relationships in CTD, which is a new addition to NeDRexDB, have a “Direct Evidence” attribute. We filtered the relationships based on the “Therapeutic” which means that the chemical has a known or potential therapeutic role in a disease. For DrugBank, drug-target interactions are marked as: Target, Enzymes, Carriers, Transporters. We only integrated those marked as Targets. For information about the version/retrieval dates of data sources see Supplementary Table 1.

4.2. *Network construction*

For network construction, some part of the data such as disease-gene, drug-indication, drug-target, gene-encoding-protein, and PPI data were obtained from the databases shown in Table 1, using the data access and mapping provided by the updated version of our NeDRex platform (Sadegh *et al.*, 2021). Disease-variant associations were directly obtained from DisGeNET.

Supplementary Table 2 shows the most important properties of all constructed networks. The comorbidity-based diseasome was constructed *via* ϕ-correlation. Let I_i denote the incidence of disease i and C_{ij} be the number of patients who were simultaneously diagnosed with diseases i and

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777111.

j. The comorbidity between the two diseases can be measured by $\phi_{ij} = \frac{C_{ij}N - I_i I_j}{\sqrt{I_i I_j (N-I_i)(N-I_j)}}$, where N is the total number of patient records ($N = 139{,}065$ for the Estonian Biobank data). When two diseases co-occur more frequently than expected by chance, we have $\phi_{ij} > 0$. We used one-tailed Fisher's exact test followed by Benjamini-Hochberg correction for multiple testing to determine the significance of comorbidity associations and connected two diseases by an edge if adjusted $P \leq 0.05$. Edge weights were defined using the ϕ-correlation, i.e., we set $w_{ij} = \phi_{ij}$ for diseases i and j with significant comorbidity association.

The indication- and target-based drugomes as well as the gene-, variant-, symptom-, and indication-based diseasomes were constructed based on the Jaccard index of the respective annotations. A_i denotes the set of annotations for a disease or drug i used as node in the network under construction (e.g., when constructing the gene-based diseasome, A_i is the set of all genes associated with disease i). We connected diseases i and j by an edge if $|A_i \cap A_j| \geq 1$ and defined the edge weights as $w_{ij} = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$. Disease nodes with $|A_i| = 0$ were removed from the networks, i.e., empty annotation sets were treated as missing data.

The bipartite indication-based drug-disease network was directly constructed from the data source, i.e., we connected a disease i with a drug j if i is an indication for j. For the bipartite target-based drug-disease network, we connected a disease i with a drug j if j targets a protein encoded by a gene associated to i. In both drug-disease networks, edges are unweighted. Finally, we constructed drug-protein-protein-disease networks where drugs are connected to their targets, experimentally validated PPIs from IID are used to connect proteins, and diseases are connected to proteins encoded by disease-associated genes.

All the constructed networks are available in the GitHub repo: https://github.com/repotrial/graphsimgt/tree/main/data/graphs

4.3. *Graph edit distance*

GED is a distance measure for attributed graphs. It is defined as the minimum cost of transforming a source graph $G_1 = (V_1, E_1)$ into a target graph $G_2 = (V_2, E_2)$ via elementary edit operations, i.e., by deleting, inserting, and substituting nodes and edges. Equivalently, GED can be defined as the minimum edit cost induced by a node map π from G_1 to G_2, where nodes maps $\pi \subseteq (V_1 \cup \{\epsilon_1\}) \times (V_2 \cup \{\epsilon_2\})$ are relations that cover all nodes $u \in V_1$ and $v \in V_2$ exactly once (ϵ_1 and ϵ_2 are dummy nodes that may be covered multiple times or left uncovered).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

We used a customised version of GED to compare the different diseasomes, drugomes, and drug-disease networks constructed as detailed in the previous section as well as their randomised counterparts. Since the networks were aligned before all pairwise comparisons, we had $V_1 = V_2 = V$ (node sets are identical) whenever comparing two networks. Consequently, we fixed π as the identity and computed GED as the sum of edge edit costs induced by the identity (the edge edit cost functions sub, del, and ins are explained below):

$$GED(G_1, G_2) = \sum_{uv \in E_1 \cap E_2} sub(uv) + \sum_{uv \in E_1 \setminus E_2} del(uv) + \sum_{uv \in E_2 \setminus E_1} ins(uv)$$

$GED(G_1, G_2)$ quantifies the global distance between the graphs G_1 and G_2. Since the node sets of G_1 and G_2 are identical in our analyses, it can be decomposed as

$GED(G_1, G_2) = \sum_{u \in V} \quad GED(G_1, G_2, u)/2,$

where $GED(G_1, G_2, u)$ is the local distance between the neighbourhood $N_1(u)$ of node u in G_1 and its neighbourhood $N_2(u)$ in G_2. The local distances are defined as follows:

$$GED(G_1, G_2, u) = \sum_{v \in N_1(u) \cap N_2(u)} sub(uv) + \sum_{v \in N_1(u) \setminus N_2(u)} del(uv) + \sum_{v \in N_2(u) \setminus N_1(u)} ins(uv)$$

Based on the local distances, we also computed cluster-level distances for a cluster of nodes $C \subseteq V$ as $GED(G_1, G_2, C) = \sum_{u \in C} \quad GED(G_1, G_2, u)/2$.

We used two types of edge edit cost functions, namely, uniform costs and costs based on normalised edge ranks. The uniform costs are defined by simply setting $sub(uv) = 0$ and $del(uv) = ins(uv) = 1$ for all edges uv. GED with uniform costs quantifies topological (dis-)similarity between two graphs but does not consider edge weights. Since edges are weighted in all compared diseasomes, we additionally defined edge edit costs based on normalised ranks. For this, we sorted the diseasomes' edges in increasing order with respect to their weights and then normalised the obtained ranks to the interval $[0,1]$ via division by the maximum rank. Let $r_1(uv)$ be the normalised rank of edge uv in diseasome G_1 and $r_2(uv)$ be its normalised rank in G_2. Then we defined the rank-based edit costs as $sub(uv) = |r_1(uv) - r_2(uv)|$, $del(uv) = r_1(uv)$, and $ins(uv) = r_2(uv)$. That is, substitutions are expensive if the involved edge's rank differs a lot in the two graphs and deletions and insertions are more expensive for high-ranked than for low-ranked edges. Uniform and rank-based edit costs led to similar results.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

4.4. *Statistical analyses based on graph edit distances*

Using GED, we tested the local- and the global-scale hypotheses as follows: For each pair G_1, G_2 of compared networks, we generated 1,000 randomised counterparts $G_1^1, \ldots, G_1^{1000}$ and $G_2^1, \ldots, G_2^{1000}$. For this, we used a random network generator which repeatedly swaps edges and non-edges to obtain randomised counterparts which exactly preserve the node degrees of the original networks. For each node u, we then computed $GED(G_1, G_2, u)$ as well as $GED(G_1^i, G_2^i, u)$ for each $i = 1, \ldots, 1000$ and also computed the global distances $GED(G_1, G_2)$ and $GED(G_1^i, G_2^i)$.

To test the global-scale hypothesis, we computed empirical *P*-values as$P = (1 + \sum_{i=1}^{1000} [GED(G_1, G_2) \geq GED(G_1^i, G_2^i)]) / (1 + 1000)$, where $[true] = 1$ and $[false] = 0$. To test the local-scale hypothesis, we used the one-sided Mann-Whitney U test to assess whether the local distances $\{GED(G_1, G_2, u) | u \in V\}$ for the original networks are significantly smaller than the local distances $\{GED(G_1^i, G_2^i, u) | u \in V, i = 1, \ldots, 1000\}$ for the randomised counterparts. Moreover, we computed node-specific local empirical *P*-values as $P(u) = (1 + \sum_{i=1}^{1000} [GED(G_1, G_2, u) \geq GED(G_1^i, G_2^i, u)]) / (1 + 1000)$ and cluster-level empirical *P*-values as $P(C) = (1 + \sum_{i=1}^{1000} [GED(G_1, G_2, C) \geq GED(G_1^i, G_2^i, C)]) / (1 + 1000)$, where $C \subseteq V$ is a cluster of nodes.

4.5. *Statistical analyses based on shortest path distances*

We carried out analyses based on shortest path distances between (1) all disease-disease pairs in a disease-gene-gene-disease network, (2) all drug-drug pairs in a drug-protein-protein-drug network, and (3) all disease-drug pairs in a disease-protein-protein-drug network. For each network, we split the multi-set of obtained distances into multi-sets X_0 and X_1, where X_1 contains the shortest path distances for all node pairs contained as edge in a reference network and X_0 contains all other shortest path distances. As reference networks, we used (1) drug-, symptom-, comorbidity-, and variant-based diseasomes, (2) a bipartite drug-indication network, and (3) an indication-based drug-drug network. We then used the one-sided Mann-Whitney *U* test to assess whether the shortest path distances contained in X_1 are significantly smaller than those contained in X_0.

4.6. *Implementation*

We have implemented all network analysis approaches underlying this work in a Python package called GraphSimQT ("graph similarity quantification tool"), which is freely available on GitHub (https://github.com/repotrial/graphsimqt). GraphSimQT uses graph-tool library for network handling and Scipy library for carrying out statistical tests and comes with all networks and scripts to reproduce

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

the results reported in this work. Moreover, GraphSimQT can be used to compare user-provided networks. Significance of comorbidity associations was evaluated using the Scipy implementation of Fisher's exact test and the stats models implementation of Benjamini-Hochberg multiple testing correction.

5. Results

We first constructed various diseasomes (including comorbiditome), drugomes, and drug-disease networks based on different data types. An overview of the used data types and derived networks is shown in Supplementary Figure 1A. The data bases used to construct the networks as well as some properties of the derived networks are shown in Table 1 and Supplementary table 2, respectively. Using GED, we then compared these networks in a pairwise manner both on a local scale, i.e. zoomed-in on individual disease or drug nodes, and on a global scale. In order to test the hypotheses introduced before, we generated 1000 permuted networks as randomised counterparts for each network. Network randomization (done with edge swap method) and computation of local and global GED are illustrated in Supplementary Figures 1B and 1C. We also investigated the impact of disease ontologies of different granularity on the similarity analyses of networks. To this end, where possible, we constructed the networks both in MONDO and in ICD-10 namespace (using three-character level codes). Since the pair-wise comorbidity data provided by Estonian Biobank was originally in ICD-10 codes and access to the patient level of data, which is necessary to map to another disease ID system before establishing the comorbidity edges, was not possible, the GED plots for analyses in MONDO is missing for the comparisons of comorbidity-based diseasome to other diseasomes.

We computed two different versions of GED using uniform and rank-based edge editing costs, respectively. Uniform edit costs discard the association strengths of the edges in the compared networks, i.e., the networks are considered unweighted; rank-based edit costs incorporate them by making it more expensive to delete or insert edges with strong associations or to substitute them by edges with weak associations. More details on disease ontology mapping, network construction, and GED computation can be found in the Methods Section.

5.1. Analyses of diseasomes and drugomes on the global scale

To test the global-scale hypothesis, we computed empirical *P*-values for each pair of networks based on global GEDs. For all evaluated pairs of networks (both in MONDO and in ICD-10 namespace), we obtained smaller global GEDs for the original diseaseomes, drugomes, or drug-disease networks than for randomised counterparts, leading to empirical *P*-values which are significant at 0.001 level. Differences between GEDs obtained for permuted and original networks are shown in Figures 1 and 2.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

Furthermore, we performed analyses based on shortest path distances between disease-disease, drug-drug, and drug-disease pairs in disease-gene-gene-disease, drug-protein-protein-drug, and disease-protein-protein-drug networks, where protein-protein and gene-gene links were obtained from PPIs. We then compared shortest path distances for node pairs which do and node pairs which do not have a link in different reference networks, using the Mann-Whitney U test. For all analyses, we observed that shortest path distances are **significantly shorter** for node pairs that are connected by a link in the reference networks (Figure 3). In particular, the results show (1) that distances between diseases that are connected by edges in diseasomes constructed based on comorbidities, shared drugs, shared symptoms, or shared genetic variants are significantly shorter than distances between diseases without such edges (Figure 3A–3D); (2) that distances of disease-drug pairs with shared indication edges are significantly shorter than distances of disease-drug pairs without such edges (Figure 3E); and (3) that distances between drug pairs with shared indication are significantly shorter than distances for drug pairs without shared indications (Figure 3F). In sum, our global analyses hence provide solid evidence for the global validity of the network medicine paradigm.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

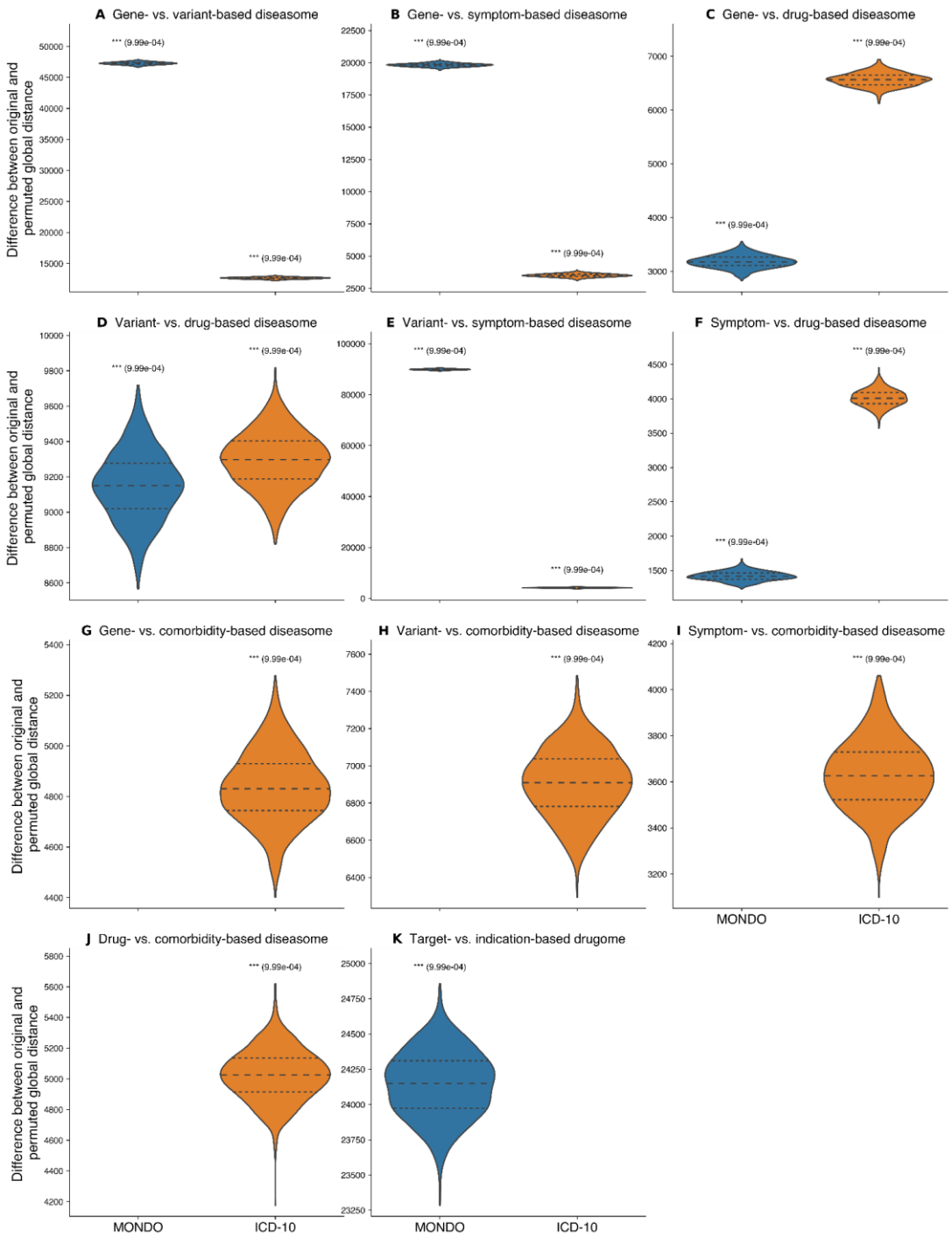


Figure 1. Pairwise similarities between networks, global distances, rank-based edge edit costs. Differences of global GEDs between the original and permuted network, and corresponding global empirical *P*-values. (A-F) Similarity between diseasomes in both MONDO and ICD-10 namespaces. (G-J) Comorbidity-based vs. other diseasomes in ICD-10 namespaces. (K) Target- vs. indication-based drugomes. All *P*-value results are at the lower limit for precision due to the 1,000 random networks used in the computation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

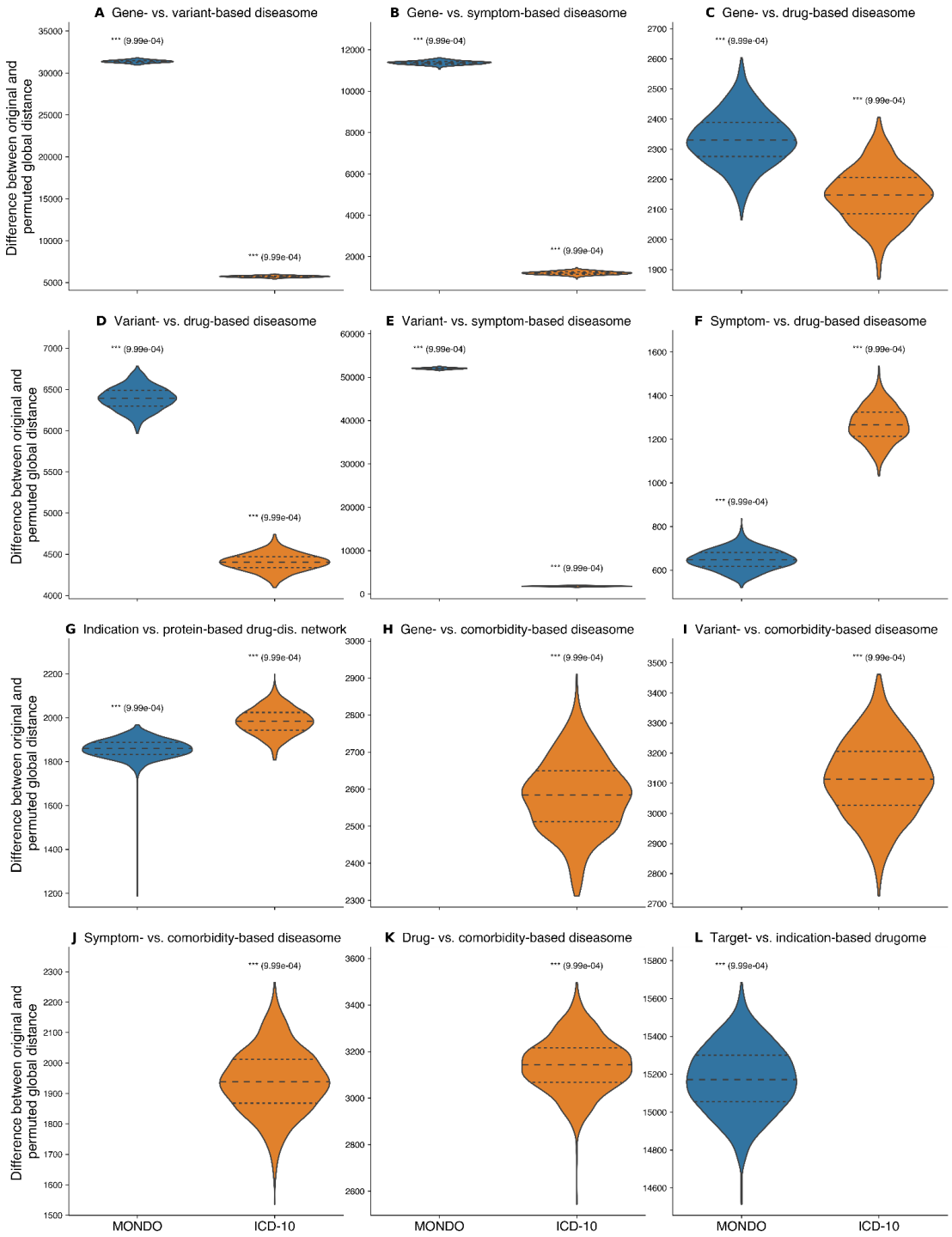


Figure 2. Pairwise similarities between networks, global distances, uniform edge edit costs. Differences of global GEDs between the original and permuted network, and corresponding global empirical *P*-values. (A-F) Similarities between diseasomes in both MONDO and ICD-10 namespaces. (G) Indication- vs. protein-based drug-disease network in both MONDO and ICD-10 namespaces. (H-K) Comorbidity-based vs. other diseasomes in ICD-10 namespace. (L) Target- vs. indication-based drugomes. All *P*-value results are at the lower limit for precision due to the 1,000 random networks used in the computation.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

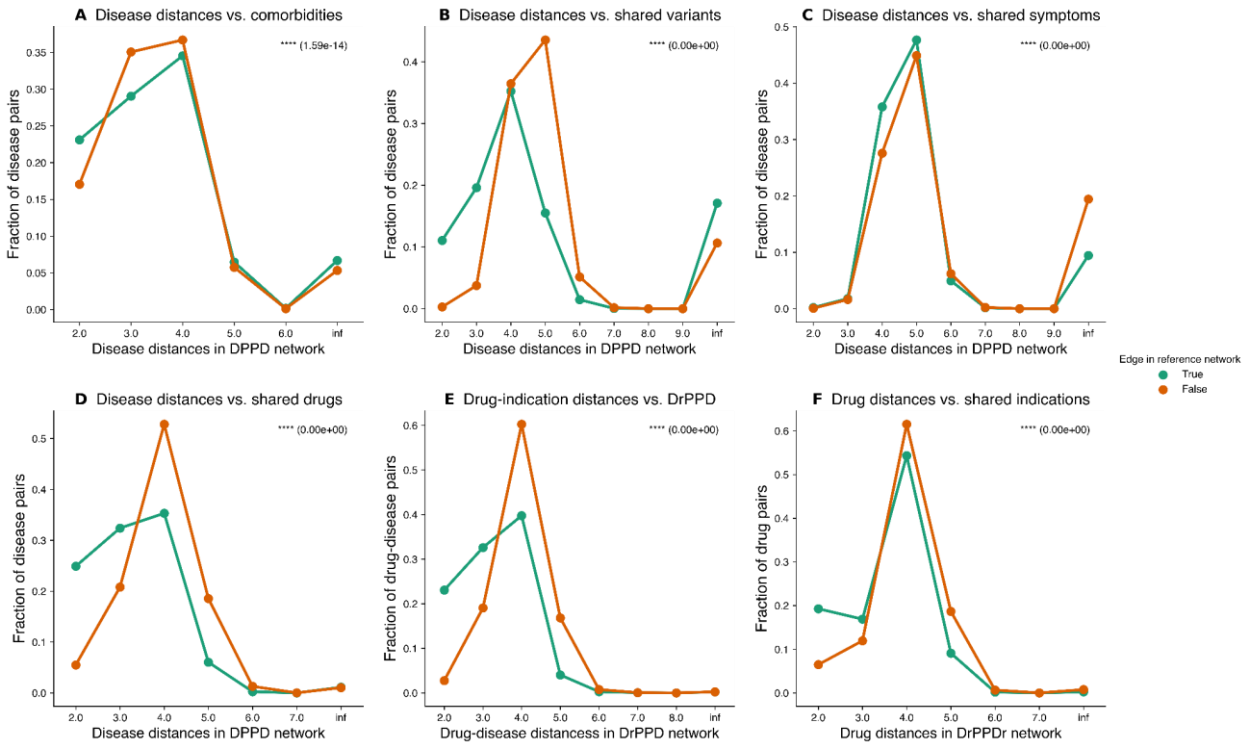


Figure 3. Shortest path distances for node pairs with and without edges in reference networks and the corresponding Mann-Whitney U *P*-values. (A-D) Disease distances in disease-protein-protein-disease (DPPD) network vs. different diseasomes as the reference network. (E) Drug-disease distances in drug-protein-protein-disease (DrPPD) network vs. drug-indication network as the reference network. (F) Drug distances in drug-protein-protein-drug (DrPPDr) network vs. indication-based drugome as the reference network. All networks are constructed in the MONDO namespace.

5.2. Analyses of diseasomes and drugomes on the local scale

To test the local-scale hypothesis, we computed *P*-values using the one-sided Mann-Whitney U test based on local GEDs to evaluate whether the local distances for the original networks are significantly smaller than the local distances for the permuted counterparts. Local GEDs of nodes obtained for the permuted and original networks and the corresponding Mann-Whitney U *P*-values are shown in Figures 4 and 5. The overview of the results of the local GED analyses in different namespaces shows that the comparisons performed in ICD-10 namespace (at three-character level) led to more significant similarities than the ones performed in MONDO namespace (Figures 6 and 7). As an example, the *P*-value computed from the local GEDs of drug-based vs. gene-based diseasomes in ICD-10 namespace is significant at 0.0001 level ($P \approx 7.1 \times 10^{-7}$), while it is not significant in MONDO namespace ($P \approx 0.071$).

The results of the Mann-Whitney U test for local GED analyses point out that we have more significant similarities in ICD-10 than in MONDO namespace. The results also suggest that variant-

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

based diseasomes have higher similarities with other diseasomes than gene-based diseasomes, considering both namespaces. By inspecting the *P*-values of drug nodes against disease nodes obtained from local similarity analyses of indication- vs. protein-based drug-disease network as well as *P*-values obtained from target- and indication-based drugome (significant at 0.001 level), we discovered that, in general, drug neighbourhoods are better preserved across the compared networks than disease neighbourhoods (Figure 6A, right panel).

Furthermore, we computed local empirical *P*-values individually for nodes based on local GEDs. The fractions of significant local empirical *P*-values at 0.05 level are shown in Figures 6B and 7B. By comparing the results in different namespaces, we observe that the fraction of diseases with small local empirical *P*-values is higher in ICD-10 namespace than in MONDO namespace. Moreover, our results show that, for a substantial fraction of disease nodes, local neighbourhoods of diseases are not preserved not only not significantly better but worse than expected by chance across the different diseasomes. The local-scale hypothesis hence seems to hold for some diseases, but does not hold at all for others. In follow-up analyses, we tried to identify patterns explaining these results, e.g., by assessing whether there are certain chapters of the ICD-10 disease ontology which are enriched with diseases with very small or very large empirical *P*-values. However, no clear patterns could be discovered, indicating that it is very hard to predict for which concrete diseases network medicine approaches can be expected to yield robust and reliable results.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

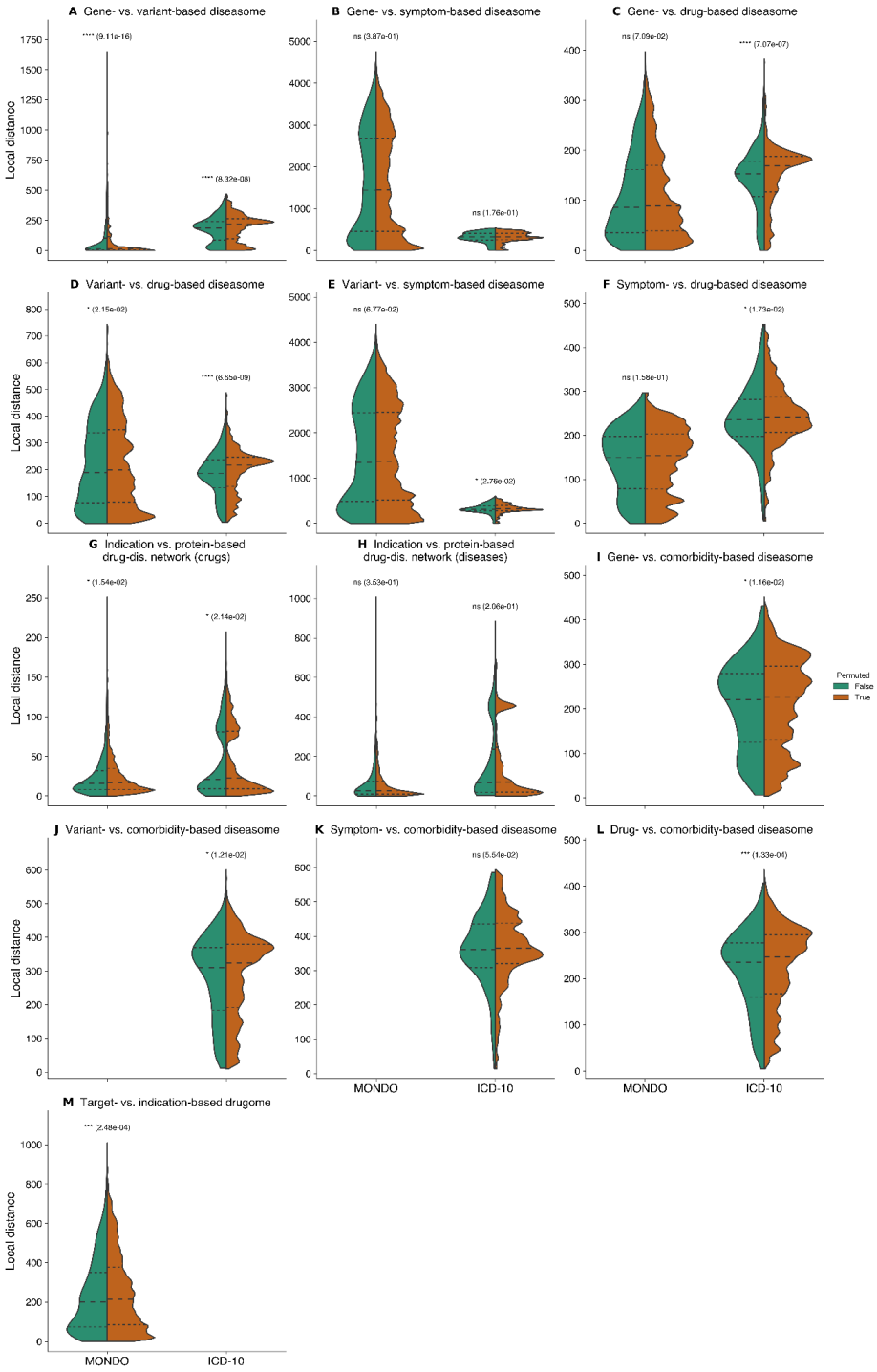


Figure 4. Pairwise similarities between networks, global view on local distances, uniform edge edit costs. Local GEDs (of all nodes) between a pair of networks for the original vs. permuted network and corresponding Mann-Whitney U *P*-values. (A-F) Similarities between diseasomes in both MONDO and ICD-10 namespaces. (G-H) Indication- vs. protein-based drug-disease network in both MONDO and ICD-10 namespaces (separately for drugs and diseases). (I-L) Comorbidity-based vs. other diseasomes in ICD-10 namespaces. (M) Target- vs. indication-based drugomes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

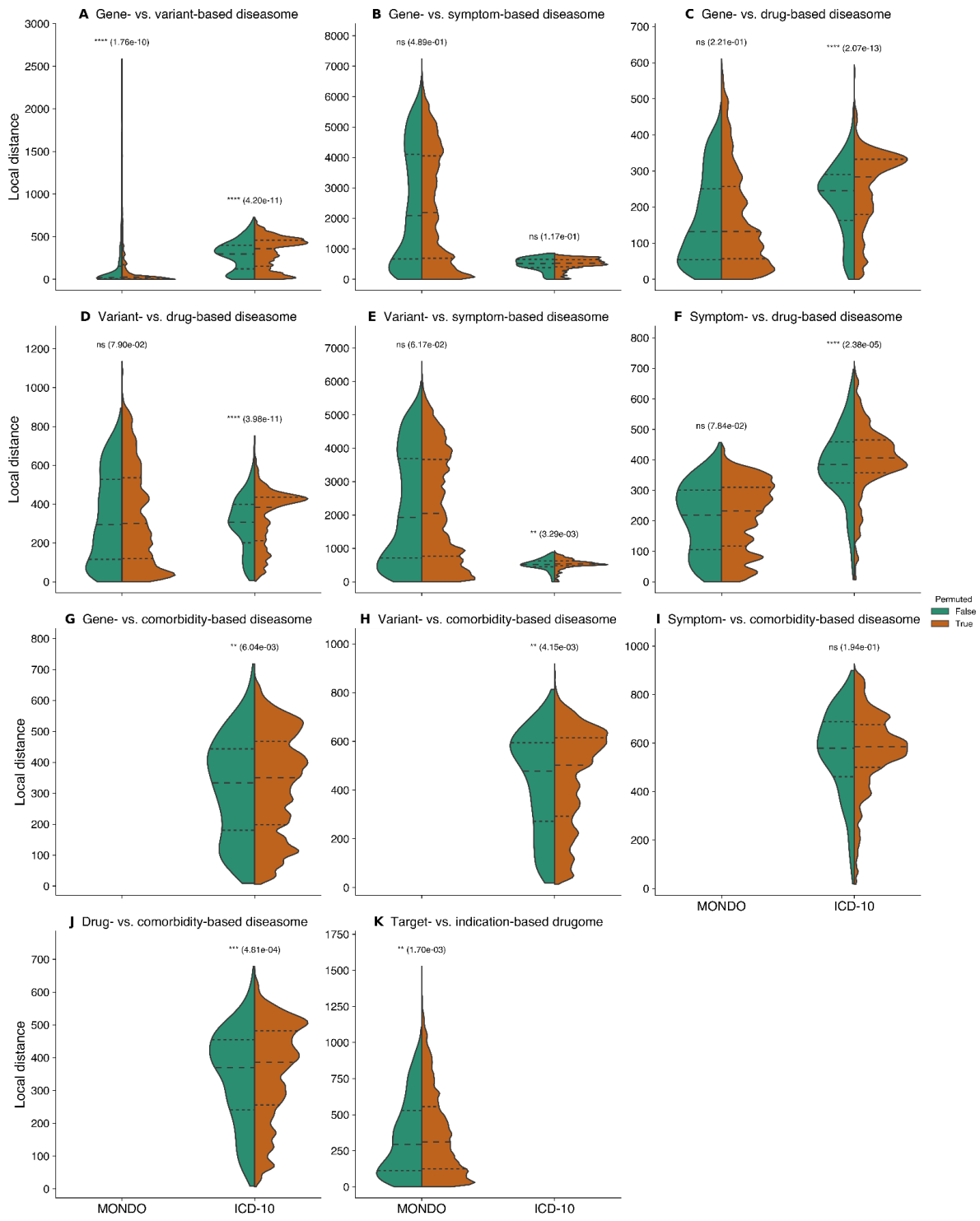


Figure 5. Pairwise similarities between networks, global view on local distances, rank-based edge edit costs. Local GEDs (of all nodes) between a pair of networks for the original vs. permuted network and corresponding Mann-Whitney U *P*-values. (A-F) Similarity between diseasomes in both MONDO and ICD-10 namespaces. (G-J) Comorbidity-based vs. other diseasomes in ICD-10 namespace. (K) target- vs. indication-based drugomes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

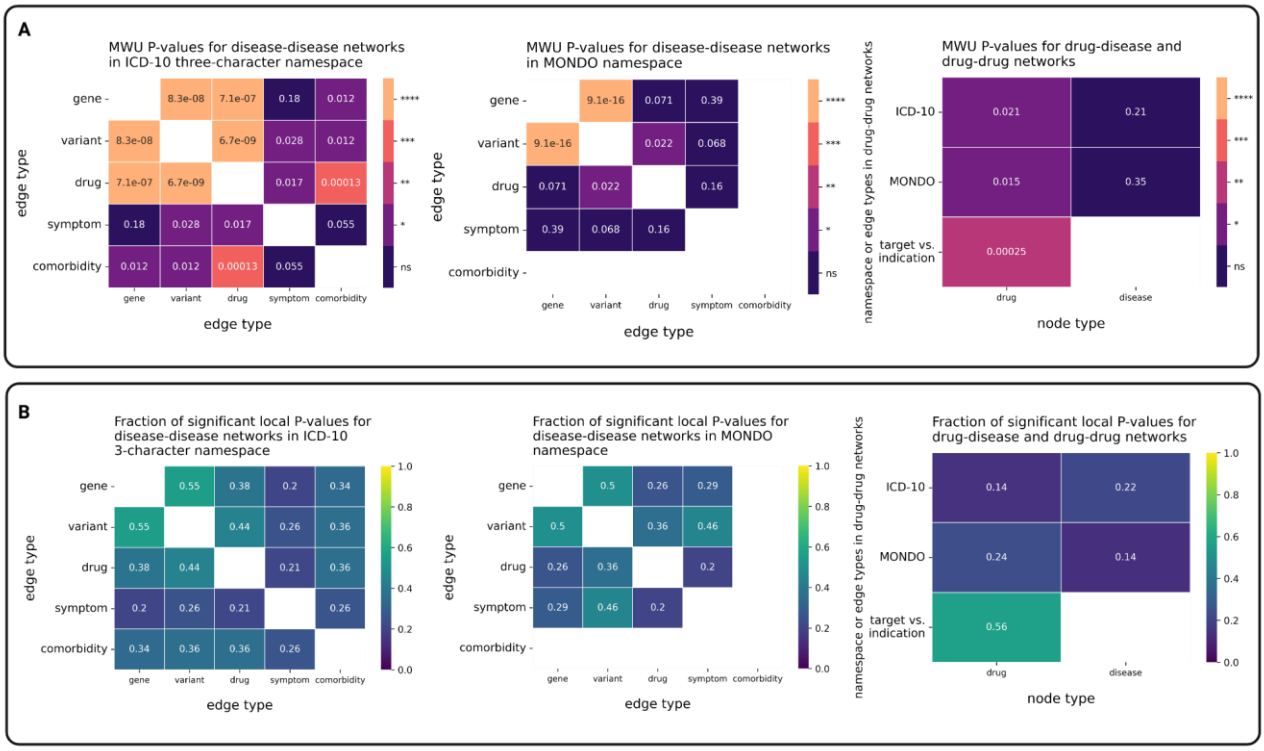


Figure 6. Overview of local scale analyses, uniform edge edit costs. (A) Mann-Whitney U *P*-values computed from local GEDs with the level of their significance. (B) Fraction of significant local empirical *P*-values at 0.05 level computed from local GEDs on a pair of networks for the original vs. permuted network.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

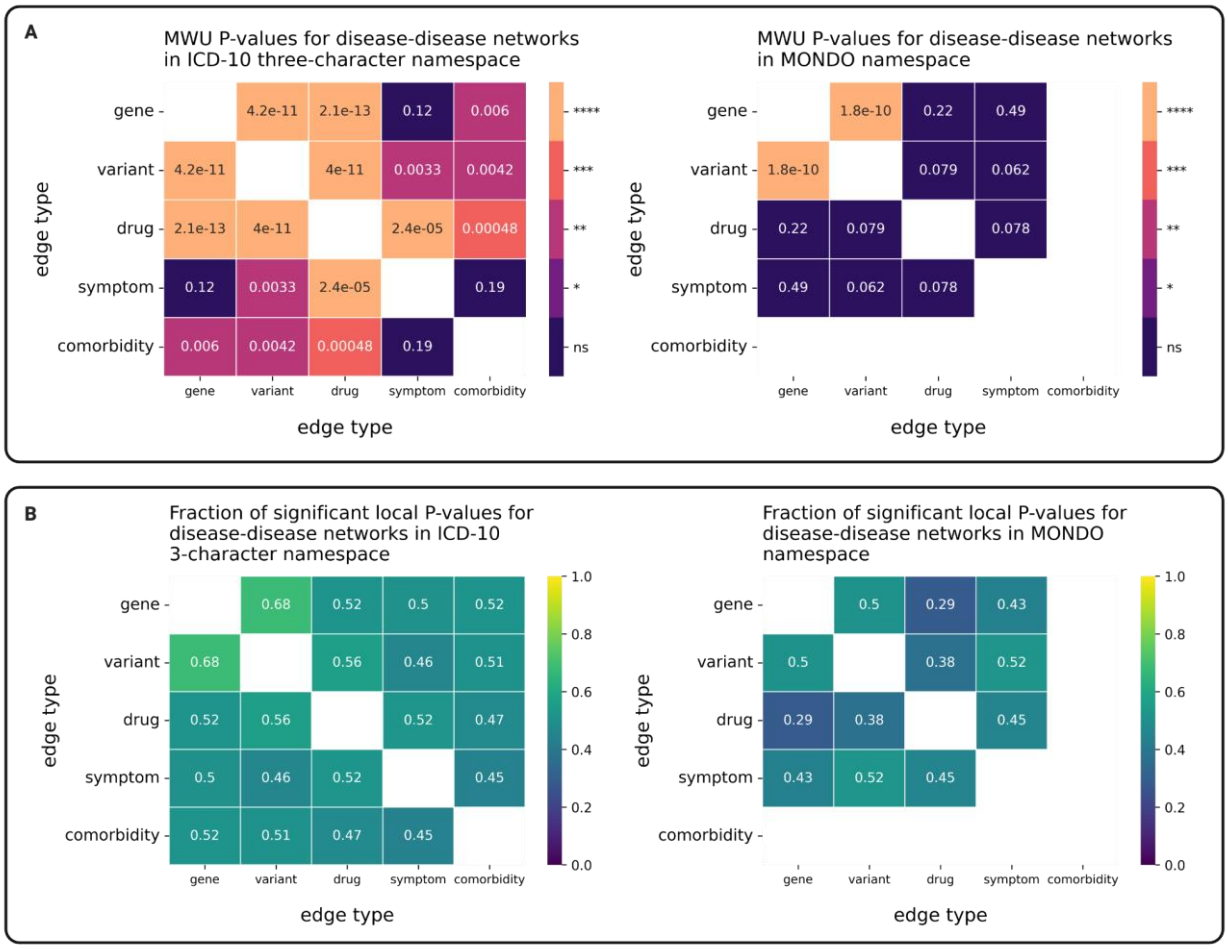


Figure 7. Overview of local scale analyses, rank-based edge edit costs. (A) Mann-Whitney U *P*-values computed from local GEDs with the level of their significance. (B) Fraction of significant local empirical *P*-values at 0.05 level computed from local GEDs on a pair of networks for the original vs. permuted network.

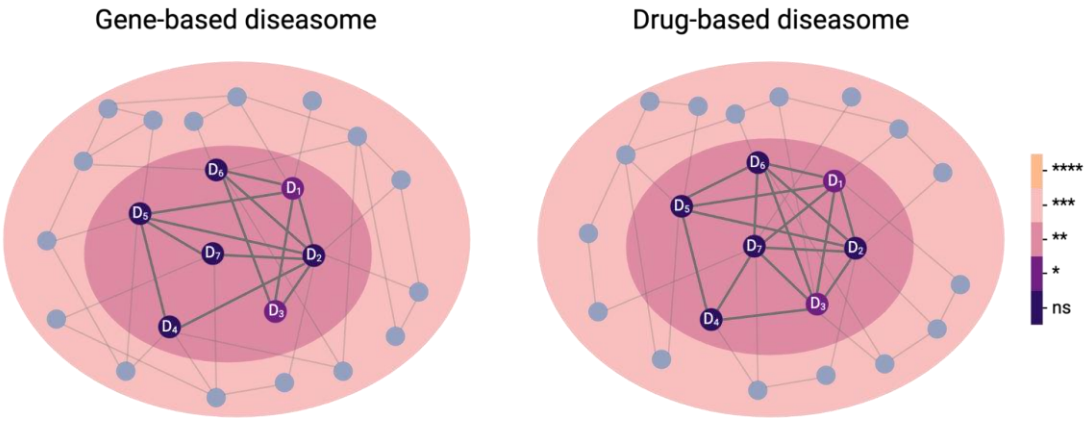
When using MONDO terms (i.e., individual diseases) as nodes in the diseasomes, only the comparisons between gene- and variant-based diseasomes consistently (with respect to both uniform and rank-based edit costs) led to smaller local distances in the original networks than in their randomised counterparts. No other comparisons in the MONDO namespace yielded significant *P*-values for both uniform and rank-based edit costs. When using ICD-10 three-character codes (which denote disease clusters rather than individual diseases), around 50% of all computed *P*-values are significant at 0.001 level.

5.3. Neurodegenerative diseases as case example

Here, we visualise the phenomenon of local blurriness in network medicine with a small example. We compiled a list of “neurodegenerative diseases” from the MONDO disease hierarchy. From ths list, we kept those for which we have nodes in the aligned gene- and drug-based diseasomes. This

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777111.

led to a cluster of seven neurodegenerative diseases which are highly connected in both diseasomes. Figure 8 shows this cluster, together with the contained diseases' local empirical *P*-values obtained from the comparison of gene- and drug-based diseasomes in MONDO space, the global empirical *P*-value, as well as the cluster-level empirical *P*-value. While only two local empirical *P*-values are significant at 0.05 level, the cluster-level and global empirical *P*-values are significant at levels 0.01 and 0.001, respectively.



Neurodegenerative diseases	Local empirical *P*-value
D_1: Alzheimer's disease	3.5964e-2
D_2: Parkinson disease	3.12687e-1
D_3: Multiple sclerosis	2.8971e-2
D_4: Motor neuron disorder	5.77423e-1
D_5: Amyotrophic lateral sclerosis	2.25774e-1
D_6: Huntington's disease	1.07892e-1
D_7: Cerebellar ataxia	7.07293e-1

Cluster-level empirical *P*-value

2.997e-3

Global empirical *P*-value

9.99e-4

Figure 8. Blurred results for neurodegenerative diseases at the local level. The colour gradient visualises local, global, and cluster-level empirical *P*-values obtained from the comparison of gene- and drug-based diseasomes in MONDO namespace. The gene-based diseasome was constructed based on disease-gene association data and two diseases were connected by an edge if they share at least one disease associated gene. The drug-based diseasome was constructed based on drug-indication data integrated and two diseases were connected by an edge if they share at least one indicated drug.

6. Open issues

6.1. Differing disease ontologies

While there are vast amounts of datasets online that contain useful information about diseases such as genetic associations, comorbidities, and symptoms, each of these datasets may use different disease ontologies to describe their associations. The ontologies have different degrees of granularity, and are generated in different ways and for different purposes. Consequently, integration of disease terms is a mammoth task that involves losing large swathes of data due to unmappable

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

terms (see Figure 9 for the levels of completeness of disease ontology mappings underlying this work).

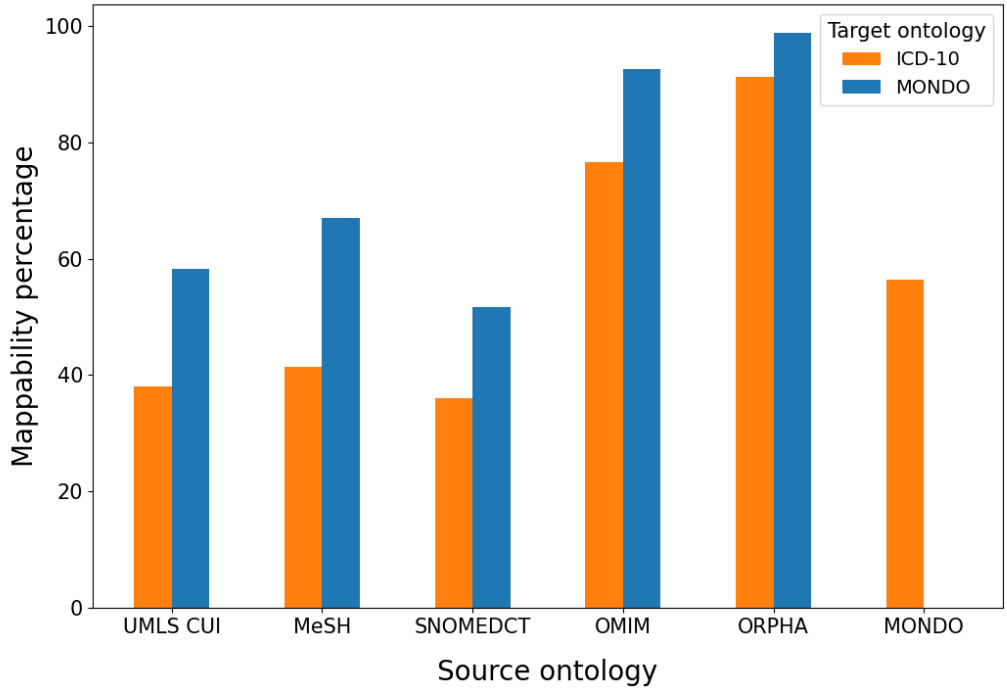


Figure 9. Levels of completeness of disease ontology mappings underlying this article. For each source-target ontology pair, mappability is computed as the percentage of terms in the source ontology used in this study that could be mapped to a term in the target ontology.

The choice of the disease ontology has the potential to dramatically affect the results of downstream analyses: While our results largely support the local-scale hypothesis for ICD-10 three-character disease terms, the same analyses carried out in MONDO namespace led to only few significant results. At the same time, for most analysis tasks, the choice of the disease ontology is dictated by the format of the data and, thus, often impossible to change without losing information at the time of analysis. The ontologies used to annotate disease-associated data must hence be viewed as confounders which are very difficult if not impossible to control for.

6.2. Mechanistically inadequate disease ontologies

Currently used disease ontologies are not only discordant, but also mechanistically inadequate: Disease names are variable and non-standardized, often reflecting the person who coined the disease term (e.g., “Alzheimer’s disease”), areas in the body that are affected (e.g., “kidney stones”) or symptoms of the disease (e.g., “irritable bowel syndrome”). This leads to data that is blurred, as diseases that should be separately defined based on their mechanistic pathways are being aggregated together, e.g., due to symptom or organ commonality. This blurriness not only has

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777111.

severe clinical ramifications (patients with mechanistically distinct diseases receive the same untargeted treatment), but also makes it very challenging to mine disease-associated data for pathomechanisms *via* network medicine approaches (Nogales *et al.*, 2022). Since such analyses often require case-*versus*-control or subtype annotations as input, it is very difficult to obtain meaningful results if the employed disease definitions are too unspecific.

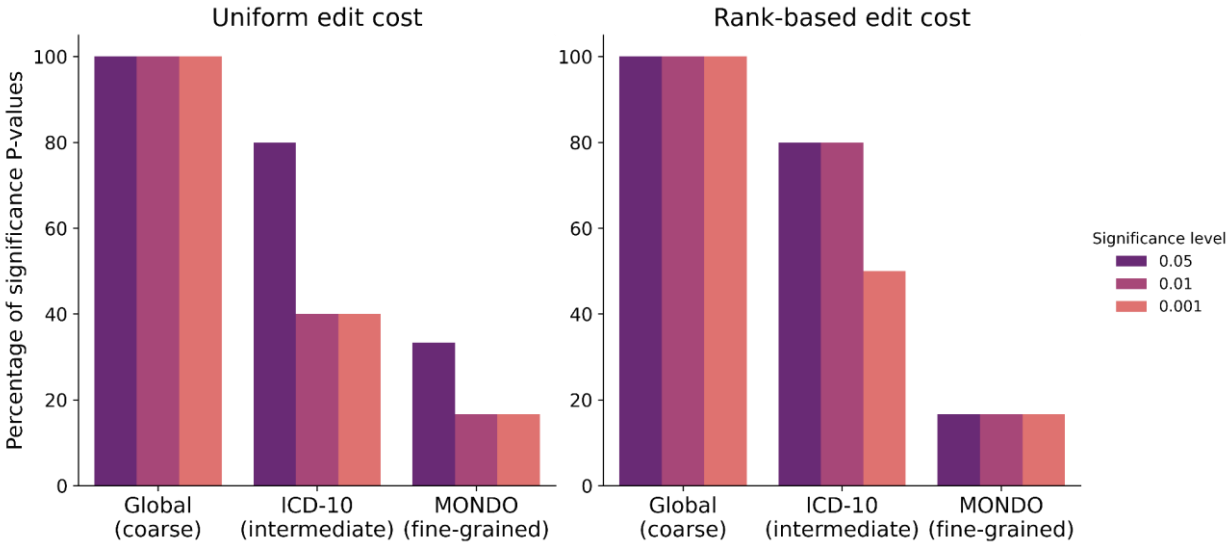


Figure 10. Effect of disease term granularity on results of GED-based analyses.

The results presented in this study, where drugome comparisons have led to more significant results on a local level than diseasome comparisons, are evidence that network-based analyses yield more targeted and reliable results when the underlying annotations are well-defined (such as in drug ontologies). Comparing the results of the GED-based analyses for full diseasomes (global analyses) with those obtained for analyses based on local GEDs in diseasomes with ICD-10 three-character codes and MONDO terms as nodes, respectively, further highlights the detrimental effect of local blurriness in currently used disease definitions: The higher the resolution of the analysis, the less significant the obtained *P*-values (see Figure 10).

6.3. Potential limitation of the comorbiditome

The constructed comorbiditome is not a real representative across different ethnicities. However, building such a complete database to cover all ethnicities, age groups and both genders is almost impossible because it requires the integration of numerous health record data of different countries (which is outside the scope of REPO-TRIAL), to balance out ethnic, gender and age imbalances within the data.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

6.4. Outlook

As mentioned earlier all the constructed networks are available in graph format files that can be loaded in any network visualisation tools like Cytoscape and explored further. Most of the data and networks are also available via the NeDRex-Web tool for further advanced and guided analyses (see deliverable D1.9). We have plans to make the comorbiditome data also available through the online user interface and provide features and functionalities for further exploration of comorbiditome in combination with other types of data. To make the most efficient use of time until the end of the project, we are currently collecting input from the potential users of NeDRex before implementing the new features based on users' needs.

7. Deviations (if applicable)

Not applicable.

8. Conclusion

The initial goal of this deliverable was to create a refined diseasome and drugome, which was achieved by using multitude disease/drug association data types such as genetic variant, symptom, comorbidity, and drug-indication data. This task was only possible by expert Disease ID mapping and curation to allow multi-scale diseasome construction. Making these information-rich networks available as pre-built graphs, allows for further investigations by the tools we developed for the NeDRex platform in the framework of REPO-TRIAL. Implementing additional network-based methods such as global and local GED methods in a Python package as part of the NeDRex platform equips experts to run further customised analyses.

Our global analyses provide solid evidence for the global-scale hypothesis and therefore global validity of the network medicine paradigm while our local analyses only provide weak evidence for the local scale hypothesis, indicating the network medicine tends to produce locally blurred results.

Our similarity analyses results indicate that the most prominent reason for the observed translational underperformance of network medicine is that disease-associated data is blurred at a local level due to inadequate disease definitions. The most obvious causes are the lack of a mechanism-based disease ontology, as well as the fact that data is annotated using a plethora of discordant and therefore partly unmappable disease ontologies. We are hence faced with a chicken-and-egg problem: Network medicine aims at uncovering pathomechanisms underlying complex diseases, but in order to reach this objective, it seems that we need mechanistically adequate disease definitions to begin with.

To escape this dilemma, we suggest the following way forward: Firstly, unsupervised network medicine methods are needed, which not only return candidate pathomechanisms but at the same

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

time de novo cluster patients into subgroups and hence do not rely on potentially misleading priorly available case-versus-control or subtype annotations. While few such approaches exist (Larsen, Schmidt and Baumbach, 2020; Lazareva *et al.*, 2020; Zolotareva *et al.*, 2020), most existing pathomechanism mining methods still rely on case-versus-control annotations (List *et al.*, 2016; Batra *et al.*, 2017) or lists of genes associated with a (potentially ill-defined) disease (Ghiassian, Menche and Barabási, 2015; Levi, Elkon and Shamir, 2021; Bernett *et al.*, 2022).

Secondly – and more importantly – we are convinced that the dilemma described above can only be resolved if bioinformaticians and data scientists working in the field of network medicine closely collaborate with researchers from the biomedical sciences and jointly analyse molecular as well as deep phenotype data for the same patients. In such a collaborative setup, a positive feedback loop could emerge, where initial hypotheses about disease subtypes and their underlying pathomechanisms are formulated based on the analysis of molecular data, further refined using deep phenotyping (e.g., histological images, blood-derived biomarkers, etc.) and expert knowledge of the clinicians, and finally validated in preclinical studies (e.g., gain- or loss-of-function studies). For this to happen, however, several hurdles have to be overcome: Bioinformaticians have to be willing to engage in close-up analyses of specific biomedical questions and to understand the involved biology. Biomedical scientists have to acknowledge the critical importance of data science to their fields and be willing to share data generated in their institutions. And, last but not least, protocols have to be implemented that allow joint analyses of molecular and deep phenotype data while respecting data protection regulations such as the General Data Protection Regulation of the European Union (Cohen and Nissim, 2020).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

9. References

AbdulHameed, M.D.M. *et al.* (2014) 'Systems level analysis and identification of pathways and networks associated with liver fibrosis', *PloS one*, 9(11), p. e112193.

Amberger, J.S. *et al.* (2019) 'OMIM.org: leveraging knowledge across phenotype-gene relationships', *Nucleic acids research*, 47(D1), pp. D1038–D1043.

Avram, S. *et al.* (2021) 'DrugCentral 2021 supports drug discovery and repositioning', *Nucleic acids research*, 49(D1), pp. D1160–D1169.

Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011) 'Network medicine: a network-based approach to human disease', *Nature Reviews Genetics*, pp. 56–68. doi:10.1038/nrg2918.

Batra, R. *et al.* (2017) 'On the performance of de novo pathway enrichment', *NPJ systems biology and applications*, 3, p. 6.

Baumbach, J. and Schmidt, H.H.H. (2018) 'The End of Medicine as We Know It: Introduction to the New Journal, *Systems Medicine*', *Systems Medicine*, pp. 1–2. doi:10.1089/sysm.2017.28999.jba.

Bernett, J. *et al.* (2022) 'Robust disease module mining via enumeration of diverse prize-collecting Steiner trees', *Bioinformatics* [Preprint]. doi:10.1093/bioinformatics/btab876.

Bunke, H. and Allermann, G. (1983) 'Inexact graph matching for structural pattern recognition', *Pattern Recognition Letters*, pp. 245–253. doi:10.1016/0167-8655(83)90033-8.

Cheng, F. *et al.* (2018) 'Network-based approach to prediction and population-based validation of in silico drug repurposing', *Nature communications*, 9(1), p. 2691.

Cheng, F., Kovács, I.A. and Barabási, A.-L. (2019) 'Network-based prediction of drug combinations', *Nature communications*, 10(1), p. 1197.

Cohen, A. and Nissim, K. (2020) 'Towards formalizing the GDPR's notion of singling out', *Proceedings of the National Academy of Sciences of the United States of America*, 117(15), pp. 8344–8352.

Davis, A.P. *et al.* (2021) 'Comparative Toxicogenomics Database (CTD): update 2021', *Nucleic acids research*, 49(D1), pp. D1138–D1143.

Ghiassian, S.D., Menche, J. and Barabási, A.-L. (2015) 'A DIseAse MOdule Detection (DIAMOnD) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome', *PLoS computational biology*, 11(4), p. e1004120.

Goh, K.-I. *et al.* (2007) 'The human disease network', *Proceedings of the National Academy of Sciences*, pp. 8685–8690. doi:10.1073/pnas.0701361104.

Guney, E. *et al.* (2016) 'Network-based in silico drug efficacy screening', *Nature communications*, 7, p. 10331.

Halu, A. *et al.* (2019) 'Exploring the cross-phenotype network region of disease modules reveals concordant and discordant pathways between chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis', *Human molecular genetics*, 28(14), pp. 2352–2364.

Iida, M., Iwata, M. and Yamanishi, Y. (2020) 'Network-based characterization of disease-disease relationships in terms of drugs and therapeutic targets', *Bioinformatics* , 36(Suppl_1), pp. i516–

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

i524.

Köhler, S. *et al.* (2021) 'The Human Phenotype Ontology in 2021', *Nucleic acids research*, 49(D1), pp. D1207–D1217.

Kotlyar, M. *et al.* (2019) 'IID 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species', *Nucleic acids research*, 47(D1), pp. D581–D589.

Larsen, S.J., Schmidt, H.H.H. and Baumbach, J. (2020) '*De Novo* and Supervised Endophenotyping Using Network-Guided Ensemble Learning', *Systems Medicine*, pp. 8–21. doi:10.1089/sysm.2019.0008.

Lazareva, O. *et al.* (2020) 'BiCoN: Network-constrained biclustering of patients and omics data', *Bioinformatics* [Preprint]. doi:10.1093/bioinformatics/btaa1076.

Levi, H., Elkon, R. and Shamir, R. (2021) 'DOMINO: a network-based active module identification algorithm with reduced rate of false calls', *Molecular systems biology*, 17(1), p. e9593.

List, M. *et al.* (2016) 'KeyPathwayMinerWeb: online multi-omics network enrichment', *Nucleic acids research*, 44(W1), pp. W98–W104.

Maron, B.A. *et al.* (2020) 'A global network for network medicine', *npj Systems Biology and Applications*. doi:10.1038/s41540-020-00143-9.

Menche, J. *et al.* (2015) 'Disease networks. Uncovering disease-disease relationships through the incomplete interactome', *Science*, 347(6224), p. 1257601.

Nogales, C. *et al.* (2022) 'Network pharmacology: curing causal mechanisms instead of treating symptoms', *Trends in pharmacological sciences*, 43(2), pp. 136–150.

Piñero, J. *et al.* (2020) 'The DisGeNET knowledge platform for disease genomics: 2019 update', *Nucleic acids research*, 48(D1), pp. D845–D855.

Sadegh, S. *et al.* (2021) 'Network medicine for disease module identification and drug repurposing with the NeDRex platform', *Nature communications*, 12(1), p. 6848.

Samokhin, A.O. *et al.* (2018) 'NEDD9 targets to promote endothelial fibrosis and pulmonary arterial hypertension', *Science translational medicine*, 10(445). doi:10.1126/scitranslmed.aap7294.

Sanfeliu, A. and Fu, K.-S. (1983) 'A distance measure between attributed relational graphs for pattern recognition', *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 353–362. doi:10.1109/tsmc.1983.6313167.

Udrescu, L. *et al.* (2020) 'Uncovering New Drug Properties in Target-Based Drug-Drug Similarity Networks', *Pharmaceutics*, 12(9). doi:10.3390/pharmaceutics12090879.

UniProt Consortium (2019) 'UniProt: a worldwide hub of protein knowledge', *Nucleic acids research*, 47(D1), pp. D506–D515.

Wang, R.-S. and Loscalzo, J. (2018) 'Network-Based Disease Module Discovery by a Novel Seed Connector Algorithm with Pathobiological Implications', *Journal of molecular biology*, 430(18 Pt A), pp. 2939–2950.

Wishart, D.S. *et al.* (2018) 'DrugBank 5.0: a major update to the DrugBank database for 2018', *Nucleic acids research*, 46(D1), pp. D1074–D1082.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

Zhou, Y. *et al.* (2020) 'Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2', *Cell discovery*, 6, p. 14.

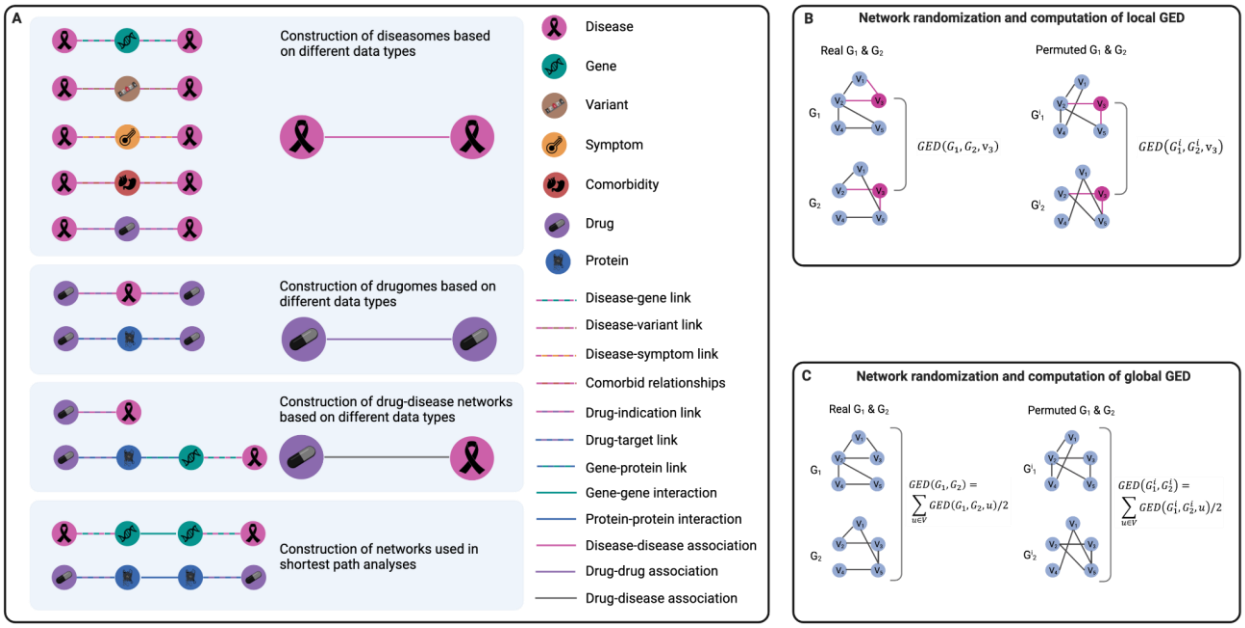
Zolotareva, O. *et al.* (2020) 'Identification Of Differentially Expressed Gene Modules In Heterogeneous Diseases', *Bioinformatics* [Preprint]. doi:10.1093/bioinformatics/btaa1038.

10. Table of acronyms and definitions

BIOCRATES	Biocrates Life Sciences AG
concentris	concentris research management GmbH
CC	Connected Component
GED	Graph Edit Distance
ICD	International Classification of Diseases
HMPC	Mucke Hermann
LCC	Largest Connected Component
MHH	Medizinische Hochschule Hannover
MLU	Martin-Luther-Universität Halle-Wittenberg
MONDO	Monarch Disease Ontology
UHAM	University of Hamburg
UKE	Universitätsklinikum Essen
UM	Universiteit Maastricht
UNEW	University of Newcastle upon Tyne
WP	Work package

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

11. Other supporting documents / figures / tables



Supplementary Figure 1. Overview of compared networks and graph edit distance computation. (A) We compared five different types of disease-disease networks (diseasomes), two different types of drug-drug networks (drugomes), and two different types of drug-disease networks. (B) Node dissimilarity across different networks was quantified by the local graph edit distance (GED) of their neighbourhoods. (C) Networks were compared globally via global GED, obtained by summing up the local GEDs of the individual nodes.

Supplementary Table 1. Version/retrieval dates of the used data sources.

Database	Date obtained (version, if known)	Edges/association contributed
OMIM	2020-03-10*	Gene-[associated with]-Disorder
IID	2020-02-11 (v2018-11)	Protein-[interacts with]-Protein
HPO	2022-02-14	Disorder-[has]-Phenotype
DrugBank	2020-02-11 (v5.1.5)	Drug-[has target]-Protein
DisGeNET*	2019-12-02 (v6.0)	Gene-[associated with]-Disorder
DrugCentral	2020-05-16	Drug-[has target]-Protein Drug-[has indication]-Disorder

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

CTD	2021-06	Drug-[has indication]-Disorder
UniProt	2020-02-11	Gene-[encoded by]-Protein Protein-[is isoform of]-Protein
Estonian Biobank	2020-04	Disease-[codiagnosed]-Disease

* Only curated gene-disease associations from DisGeNET are integrated.

Supplementary Table 2. Properties of constructed networks. Drugomes were constructed only in MONDO namespace and the comorbidity-based diseasome only in ICD-10 namespace (since Estonian Biobank uses ICD-10 codes, mapping the comorbidities to the finer-grained MONDO namespace was impossible).

Network	# Nodes	# Edges	# CCs	Size of LCC	Density	# Isolated nodes
MONDO namespace						
Gene-based diseasome	8,284	118,726	1,571	6,598	0.00346	1,483
Variant-based diseasome	8,759	677,241	1,401	7,162	0.01766	1,250
Symptom-based diseasome	9,926	10,982,517	43	9,883	0.22296	41
Drug-based diseasome	1,772	129,788	42	1,724	0.08271	36
Target-based drugome	5,878	272,010	164	5,619	0.01575	108
Indication-based drugome	2,249	194,959	42	2,195	0.07712	34
Protein-based drug-disease network	1,629 dr 607 dis	37,360	1	2,236	0.03778	0
Indication-based drug-disease network	2,249 dr 1,772 dis	15,800	1	4,021	0.00396	0
Drug-protein-protein-disease	7,473 dis 5,878 dr 24,757 pr	422,617	52	37,967	0.00066	0
Disease-gene-gene-disease	8,284 dis 18,113 g	371,385	810	24,768	0.00118	0
ICD-10 namespace (3-character codes)						
Gene-based diseasome	755	45,671	28	728	0.16045	27
Variant-based diseasome	894	101,554	22	872	0.25441	20
Symptom-based diseasome	735	150,618	3	733	0.37165	2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.

Drug-based diseasome	714	60,832	8	705	0.23899	5
Comorbidity-based diseasome	1,114	122,030	1	1,114	0.19684	0
Target-based drug-disease network	1,588 dr 382 dis	63,690	1	1,970	0.10499	0
Indication-based drug-disease network	1,950 dr 714 dis	14,299	8	2,640	0.01027	0

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111.