

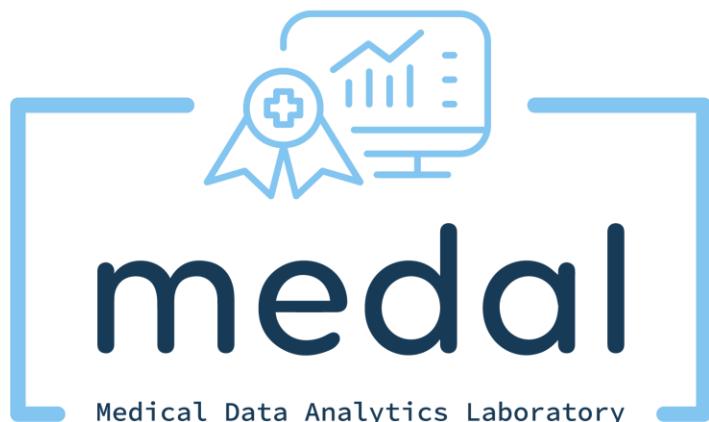


POLITÉCNICA

"Ingeniamos el futuro"



center for
biomedical
technology



<https://medal.ctb.upm.es/>
 @MEDAL_CTB

Disease understanding: Dealing with complex and unstructured big data in biomedical domain

February 7th, 2025

Madrid

Alejandro Rodríguez González

Lucía Prieto Santamaría

Belén Otero Carrasco

Last update: Feb 2025

How doesn't work this seminar?



Please.. Be participative!
If you have any doubt: ask!



POLITÉCNICA



Who are we?





Material

The material of the seminar will be available at Moodle next week.

Agenda

1. What is medical informatics?
2. Diagnosis problem
3. Medical knowledge: representation
4. Medical vocabularies and data sources
5. Medical knowledge: acquisition
6. Diagnosis knowledge acquisition
7. Human disease complex networks
8. Disease understanding
9. Drug repositioning

Medical informatics

What is Medical Informatics?



Medical informatics



Source: <https://www.linkedin.com/topic/health-informatics>

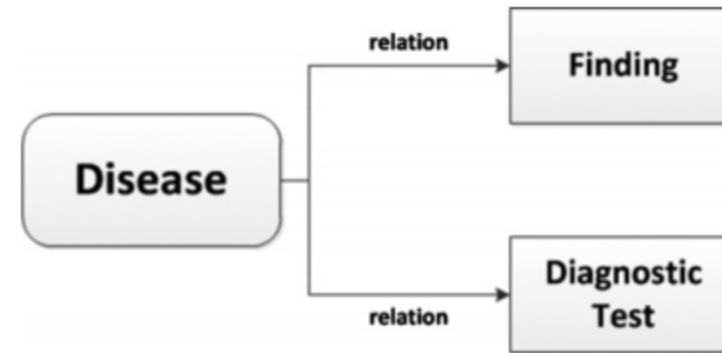
Diagnosis

How is performed the diagnosis in medicine?



Diagnosis

The model used is the DCM (Diagnosis Criterion Model):



Source: Rodríguez-González, A., Alor-Hernández, G. (2012). An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies. *Computers in Biology and Medicine*.

Diagnosis

The diagnosis of a disease is based in non-monotonic logic:

- A increase of the input findings..
- Decrease the resulting outputs.



Diagnosis

The diagnosis is mainly base in two elements:

- Findings: includes symptoms and signs.
- Diagnostic tests: tests that allows to confirm some hypothesis or abnormal conditions: blood test, urine tests, radiography, scanner, etc.

Diagnosis

Diagnosis is a difficult problem because:

1. Findings are usually very unspecific (same symptoms for different diseases). Only a few diseases contain pathognomonic findings (Koplik's spots in Measles, Aschoff bodies (nodules in heart) in rheumatic fever, etc.).

Diagnosis

Diagnosis is a difficult problem because:

2. Comorbidity: the presence of one or more additional disorders (or diseases) co-occurring with a primary disease or disorder. This is very usual in elderly people.

Diagnosis

Diagnosis is a difficult problem because:

3. The clinical manifestations of a disease are not a binary value. Prevalence of the findings is a factor to take into account.

Diagnosis

Diagnosis is a difficult problem because:

4. The findings are the manifestation of the external agent/mutation in a specific host: this manifestations depends on several genetic parameters, environmental factors, population, etc.



Diagnosis

How computer systems could for example help to the health sector:

My story in Thailand

Diagnosis



Source: Own elaboration

Diagnosis



Diagnosis

Diagnoses made in my case was mainly based on the prevalence of the diseases in Spain. The trip and the associated information was considered, but was not changing the decision-making process.

Primary care physicians didn't even take this information in consideration to send me to the hospital.

Diagnosis

Some of the diagnosis that were firstly tested included various types of diseases with high prevalence in Spain, not considering that the event happened in Thailand.

A DSS should take this into account to change the probability of the diagnosis.

Diagnosis

Ontology DDX

Archivo Ver Opciones Ayuda

Datos del paciente

Edad: Información no relevante

Sexo: Indiferente Hombre Mujer

Operaciones Transfusiones

Datos clínicos

Países visitados

Nombre
 Angola
 Antártida

Medicinas usadas

Nombre	Código
Hidralazina	NCM0010

Indicios médicos

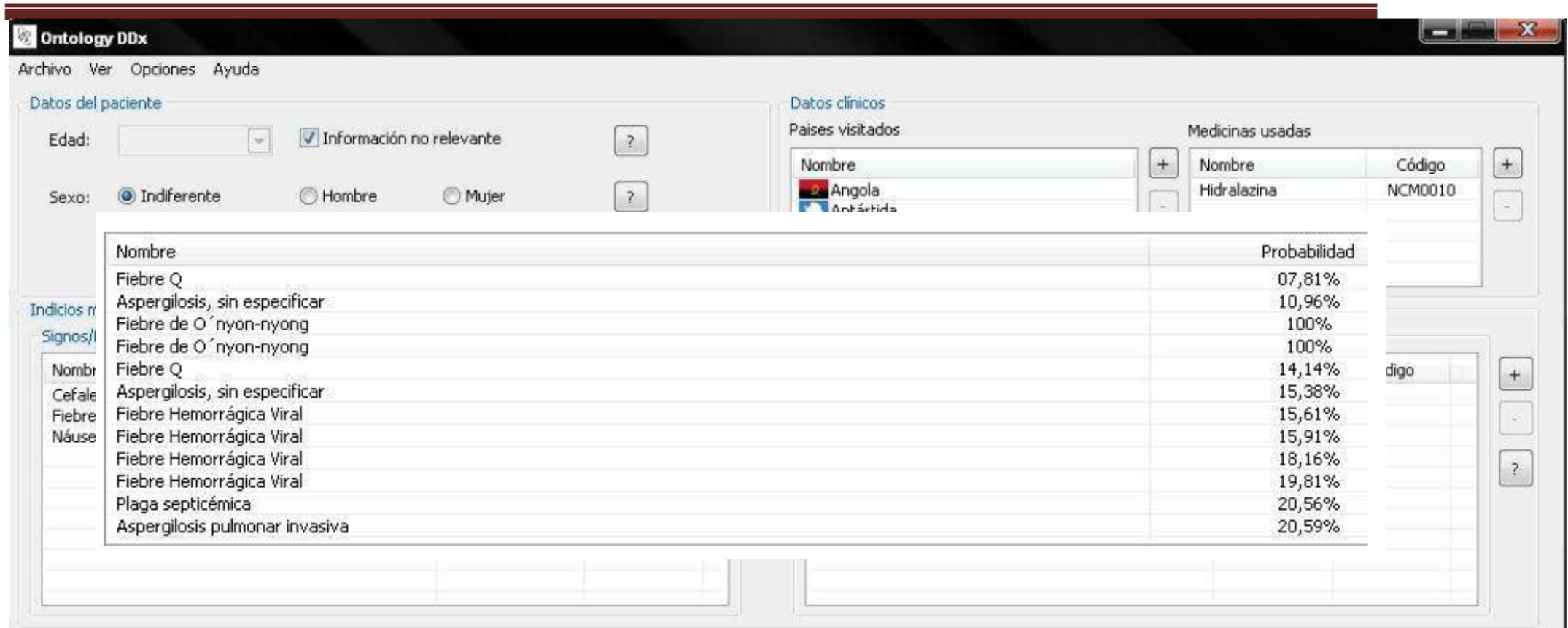
Signos/Enfermedades

Nombre	Peso	Código
Cefalea	Medio	R51
Fiebre sin especificar	Muy bajo	R50.9
Náusea y vómito	Alto	R11

Pruebas de laboratorio

Nombre	Peso	Código
--------	------	--------

Diagnosis



The screenshot shows the Ontology DDX software interface. At the top, there is a menu bar with Archivo, Ver, Opciones, and Ayuda. Below the menu, there are sections for Datos del paciente (Patient Data) and Datos clínicos (Clinical Data). In the Datos del paciente section, there are fields for Edad (Age), Sexo (Sex), and checkboxes for Información no relevante (Irrelevant information) and Paises visitados (Visited countries). In the Datos clínicos section, there are fields for Medicinas usadas (Used medicines) and Paises visitados (Visited countries). The main area displays a table of diagnostic results:

Nombre	Probabilidad
Fiebre Q	07,81%
Aspergilosis, sin especificar	10,96%
Fiebre de O'nyon-nyong	100%
Fiebre de O'nyon-nyong	100%
Fiebre Q	14,14%
Aspergilosis, sin especificar	15,38%
Fiebre Hemorrágica Viral	15,61%
Fiebre Hemorrágica Viral	15,91%
Fiebre Hemorrágica Viral	18,16%
Fiebre Hemorrágica Viral	19,81%
Plaga septicémica	20,56%
Aspergilosis pulmonar invasiva	20,59%

Diagnosis

AL

What is the most likely diagnosis if the symptoms are fever, headache, muscle pain and malaise?



The most likely diagnosis for these symptoms could be the flu (influenza).



AL

Ok, and what about if the patient fall into a river?



If a patient falls into a river and presents with the symptoms of fever, headache, muscle pain and malaise, the diagnosis could be hypothermia or pneumonia (if water was contaminated).

AL

Ok, and what about if the river was in Thailand?



If the river was in Thailand, additional diseases to consider as a possible diagnosis for a patient who falls into the river and presents with symptoms of fever, headache, muscle pain and malaise include: dengue fever, leptospirosis, and melioidosis.



AI in biomedicine: introduction

AL

We have a patient that just arrived from a trip to Thailand two days ago. Current symptoms are fever, headache, muscle pain, and malaise. He started with these symptoms around 48 hours ago. On the trip to Thailand, he mentions that he fell into a river when he was practicing rafting around 10 days ago. When he fell into the river he was about to drown, and he has some small injuries when he tried to stop being dragged from the river by putting their feet into the ground of the river. Apart from this, he mentions that he also did a tattoo and ate a scorpion.

What is the most likely diagnosis?



Based on the patient's symptoms and recent travel history, it is possible that the patient has contracted a tropical disease such as malaria, dengue fever, or chikungunya. Additionally, falling into a river could have exposed the patient to various waterborne illnesses such as leptospirosis or schistosomiasis.

The fact that the patient also got a tattoo and ate a scorpion is not directly related to the symptoms but may increase the likelihood of bacterial infections.

Without further information or diagnostic tests, it is difficult to provide a definitive diagnosis. Therefore, it is essential to evaluate the patient further, conduct diagnostic tests, and consult with a medical professional to determine the most likely diagnosis and appropriate treatment. The patient should seek immediate medical attention.

Source: Own elaboration

Diagnosis

The diagnosis process can be improved from a computational perspective.

Most of diagnosis systems are based on the following approaches:

1. Rule-based systems.
2. Probabilistic techniques.
3. Machine learning techniques.

Diagnosis

The creation of computational approaches to deal with the diagnosis needs two basic elements:

1. Knowledge bases with the clinical manifestations and diagnostic tests of a disease.
2. A way to represent this information (concepts and relations between them).

Medical knowledge: representation

How can we represent medical knowledge?



Medical knowledge: representation

It is a really complex problem because:

1. We have several kinds of concepts to represent.
2. Those concepts can be simple or complex concepts that can be related among them.
3. The concepts can have units.
4. The concepts can be negated.
5. Time is involved also.



Medical knowledge: representation

Small exercise: 5 minutes to think how to represent the following knowledge:

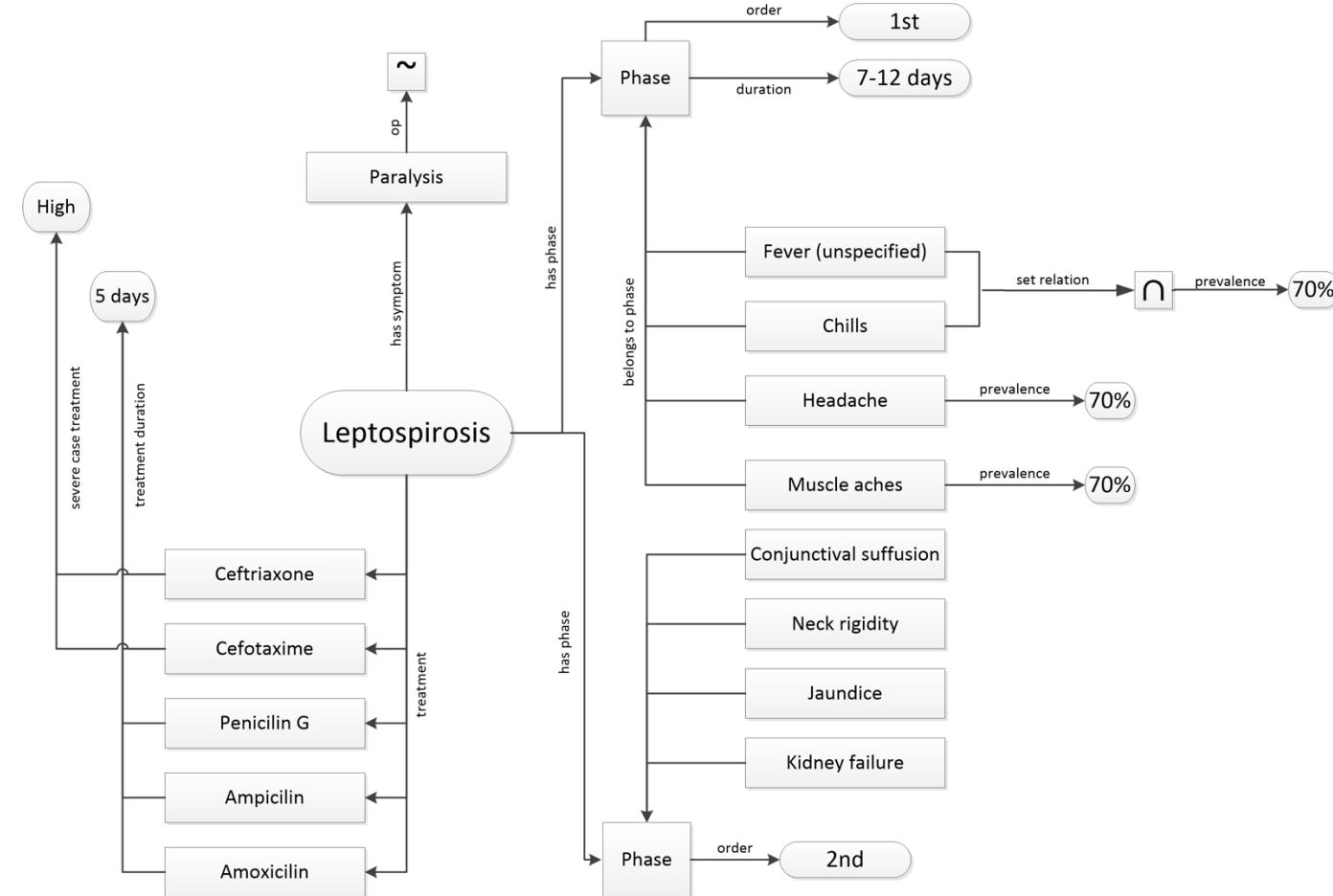
We want to make a representation about how to create a diagnosis system for the leptospirosis, a disease caused by leptospira bacteria. The incubation period is 7-12 days. In this phase shows symptoms similar to common cold: fever with chills (70%), headache (84%), muscle aches (90%). In the second phase other symptoms such as conjunctival suffusion, neck rigidity, jaundice, kidney failure. A serology could be made to confirm the presence of the bacteria in the blood. This disease never produces paralysis. Effective antibiotics include penicillin G, ampicillin, and amoxicillin. In more severe cases cefotaxime or ceftriaxone should be preferred. Treatment duration: 5 days.

Medical knowledge: representation

Entities to be represented in the example and problems associated:

- Symptoms: The symptoms can have subtypes: fever is very unspecific (low-fever, fever, hyperpyrexia, intermittent fever, continuous fever..?). Muscle aches: any kind of muscle involved? Negation: never produces paralysis.
- Prevalence: symptoms includes prevalence.
- Laboratory tests: serology.
- Temporal information: Information such as the incubation period or treatment duration.
- Treatment information: drugs to be used.

Medical knowledge: representation



Medical knowledge: representation

Graph/network based representations are optimal.

It is necessary to decompose the terms used trying to reach maximum atomicity.

As more atomicity we reach, more complexity can have the model but more difficult is to reach comprehensive definitions of concepts.



Medical knowledge: representation

Semantic technologies (Semantic Web) offers a good knowledge representation framework to create relationships between medical concepts.

Different ontologies could be created to represent “kinds” of concepts in biomedical domain.

Medical vocabularies and data sources

There are several vocabularies and ontologies that are widely used to represent biomedical knowledge.

Do you know any?



Medical vocabularies and data sources

One of the most important and largest existing vocabularies is SNOMED-CT (CT stands for “Clinical Terms”).

SNOMED-CT is a collection of medical terms that provides codes, synonyms, hierarchical classification and definitions. It is used in clinical documentation and reporting.



Medical vocabularies and data sources

SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world (Benson, 2010).

Although it is supposed to be an standard vocabulary, it has different versions depending on the country.

Benson, T. (2010). *Principles of Health Interoperability HL7 and SNOMED*. Springer.

Medical vocabularies and data sources

A couple of SNOMED-CT browsers:

- <http://browser.ihtsdotools.org/>
- <http://www.snomedbrowsers.com/>

Medical vocabularies and data sources

The screenshot shows a search interface for medical vocabularies. On the left, a search bar contains 'Influenza'. Below it, a table lists search results:

Result	Description
Influenza	Influenza (disorder)
Avian influenza	Avian influenza (disorder)
Influenza virus	Family Orthomyxoviridae (organism)
Influenzavirus B	Genus Influenzavirus B (organism)
Influenzavirus C	Genus Influenzavirus C (organism)
Balkan influenza	Q fever (disorder)
Influenza C virus	Influenza C virus (organism)
Influenza A virus	Influenza A virus (organism)
Influenza vaccine	Influenza virus vaccine (substance)
Influenza virus A	Influenza A virus (organism)
Influenza vaccine	Influenza virus vaccine (product)

On the right, a detailed view of the 'Influenza (disorder)' concept is shown. It includes:

- Parents:**
 - Disease due to Orthomyxoviridae (disorder)
 - Respiratory tract infection (disorder)
- Concept Details:**
 - Influenza (disorder) ★ 🔍**
 - SCTID: 6142004
 - 6142004 | Influenza (disorder) |
 - Influenza
 - Flu
 - Grippe
 - Influenza (disorder)
- Relationships:**
 - Causative agent → Family Orthomyxoviridae
 - Pathological process → Infectious process
 - Finding site → Structure of respiratory system
- Children (12):**
 - Bronchiolitis caused by influenza virus (disorder)
 - Healthcare associated influenza disease (disorder)
 - Influenza due to Influenza A virus (disorder)
 - Influenza due to Influenza virus, type B (disorder)



Medical vocabularies and data sources

ICD (International Classification of Diseases)

ICD is a classification promoted by WHO. Current version of ICD is ICD-10. It has several subsets such as CM (Clinical Modification) to classify diagnosis and reason for visits in health care settings or PCS (Procedure Coding System).

Medical vocabularies and data sources

ICD is mainly used for administrative uses but has been also used to describe clinical picture of a patient, report morbidity or for billing information.

ICD includes a section for mental and behavioural disorders developed alongside with American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders (DSM).



Medical vocabularies and data sources

ICD-10-CM Example

J09 Influenza due to certain identified influenza viruses

Excludes1: influenza due to other identified influenza virus (J10.-)
influenza due to unidentified influenza virus (J11.-)

J09.0 Influenza due to identified avian influenza virus

Avian influenza

Bird flu

Influenza A/H5N1

J09.01 Influenza due to identified avian influenza virus with pneumonia

Code also associated lung abscess, if applicable (J85.1)

J09.010 Influenza due to identified avian influenza virus with identified avian influenza pneumonia

J09.018 Influenza due to identified avian influenza virus with other specified type of pneumonia

Code also the specified type of pneumonia

J09.019 Influenza due to identified avian influenza virus with unspecified type of pneumonia

J09.02 Influenza due to identified avian influenza virus with other respiratory manifestations

Influenza due to identified avian influenza virus NOS

Prepared for AHLA – Institute on Medicare and Medicaid Payment Issues
March 21, 2013.

Page 8

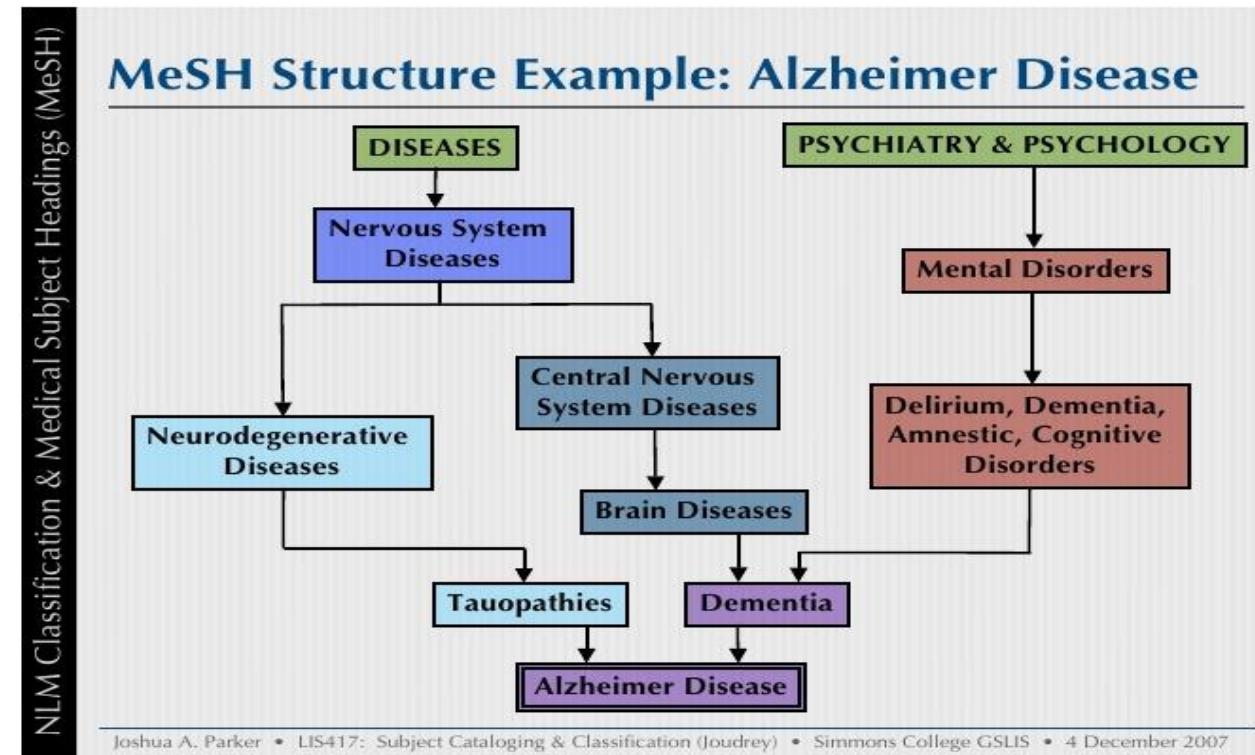
Source: <http://goo.gl/DBKoBi>

Medical vocabularies and data sources

MeSH (Medical Subject Headings) is a controlled vocabulary used with the aim of indexing journal articles and books in life sciences.

It is used by PubMed/MEDLINE article database as well as by ClinicalTrials.gov registry to classify which diseases are studied by trials registered at the page.

Medical vocabularies and data sources



Source: <http://goo.gl/UtjVah>

Medical vocabularies and data sources

Unified Medical Language System (UMLS) is a compendium of several controlled vocabularies in the biomedical sciences.

It provides mappings between the vocabularies, allowing to translate among various terminology systems.

Medical vocabularies and data sources

UMLS consists in the knowledge sources (databases) a set of software tools such as MetaMap or SemRep, among others.

The full list of sources used by UMLS is available at:

<https://www.nlm.nih.gov/research/umls/sourcerelease/docs/index.html>

Medical vocabularies and data sources

The concepts in UMLS are represented by a code named CUI (Concept Unique Identifier). However, there are several other codes regarding lexical terms, strings or atomic entities.

More information:

https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta005.html

A1412439	headaches (BI)
S1459113	headaches
A2882187	Headache (SNOMED)
A0066000	Headache (MeSH)
S0046854	Headache
L0018681	headache
A1641293	Cranial Pain (MeSH)
S1680378	Cranial Pain
L1406212	cranial pain
A0418053	HEAD PAIN CEPHALGIA (DxP)
S0375902	HEAD PAIN CEPHALGIA
L0290366	cephalgia head pain
C0018681	Headache

Medical vocabularies and data sources

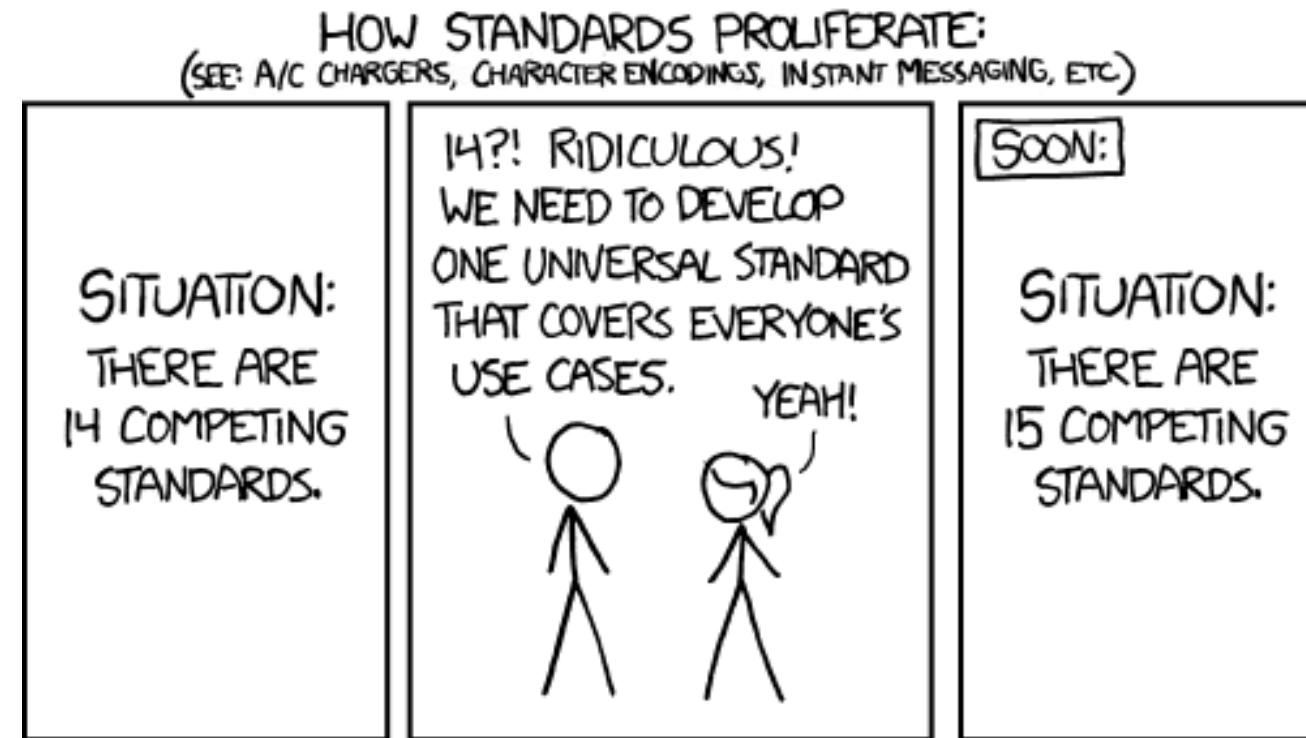
Another interesting sources with biomedical semantic vocabularies/ontologies and data sources are available at:

- [OBO-Foundry](#)
- [BioPortal](#)

Medical vocabularies and data sources

One of the main problems about medical vocabularies is the vast amount of standards to represent the information. Moreover, most of these standards not always offer direct mapping, creating a problem in terms of information exchange.

Medical vocabularies and data sources



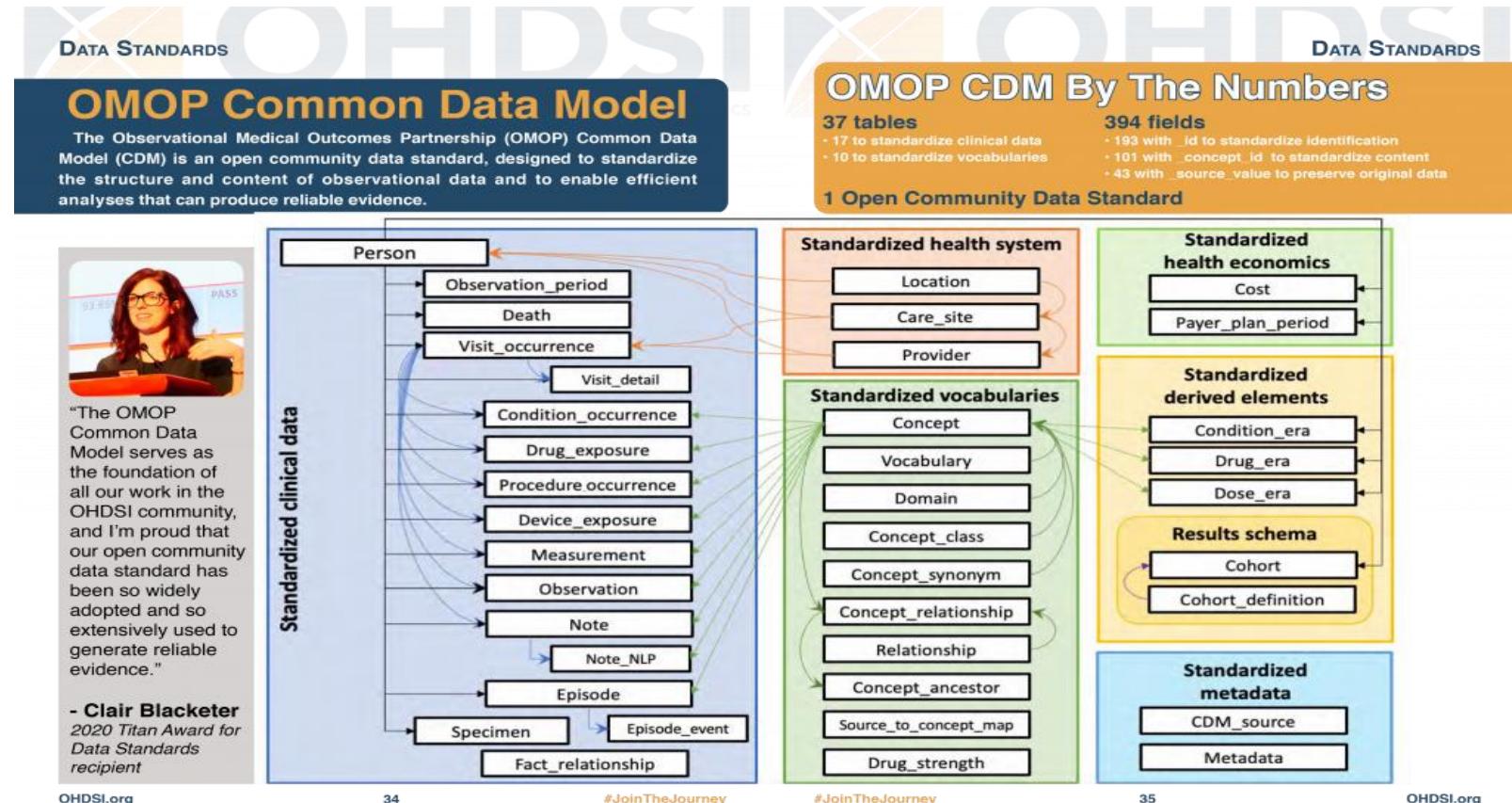
Source: <https://xkcd.com/927/>

Medical vocabularies and data sources

This is a problem indeed not only in research, but also in clinical practice. Two clear examples:

- Hospital EHR systems are usually independent, even in the same hospital.
- The problem increase when we talk about hospitals in different regions (same country), and even more among countries.
- Private and public health sector are not communicated.

Medical vocabularies and data sources



Source: <https://www.ohdsi.org/data-standardization/>

Medical knowledge acquisition

Medical knowledge acquisition refers to the set of techniques or procedures that allows to obtain structured knowledge regarding medical information in an automatic or semi-automatic way.



Medical knowledge acquisition

Generally, the extraction of the medical knowledge is done from sources such as texts (e.g.: clinical history (EHR); scientific literature), time series (ECG, EEG, ...) and images.

Texts contains information regarding medical data expressed in natural language. It is important to develop method and techniques that allows accuracy extraction of the knowledge from texts.

Diagnosis knowledge acquisition

Do you know of any resource that provides you a relation of the clinical manifestations of a set of diseases?



Diagnosis knowledge acquisition

We have some databases.. (OMIM, Orphanet, Disease DB).. but most of them are:

1. Difficult to access and handle.
2. Focused on genetic or rare diseases.
3. With incomplete information.

Diagnosis knowledge acquisition

Imagine that we want to create a database with all the infectious diseases and their respective symptoms. How can you get this information?



Diagnosis knowledge acquisition

The acquisition of diagnosis knowledge is a harder task. Several things should be taken into account:

1. For the same disease we can have several sources and each source can contain different information.
2. The information regarding the diagnosis of a disease could contain several types of information (remember leptospirosis example).

Diagnosis knowledge acquisition

The acquisition of diagnosis knowledge is a harder task. Several things should be taken into account:

3. The information expressed in natural language is hard to process (negations, inter-relationships, correct association of semantic type, ...).
4. It is necessary to extract the information accurately, but also it is important to create a good knowledge representation schema.

Big Data, diagnosis and knowledge

Ok. But what is the utility of having a list of the findings of a disease?





POLITÉCNICA

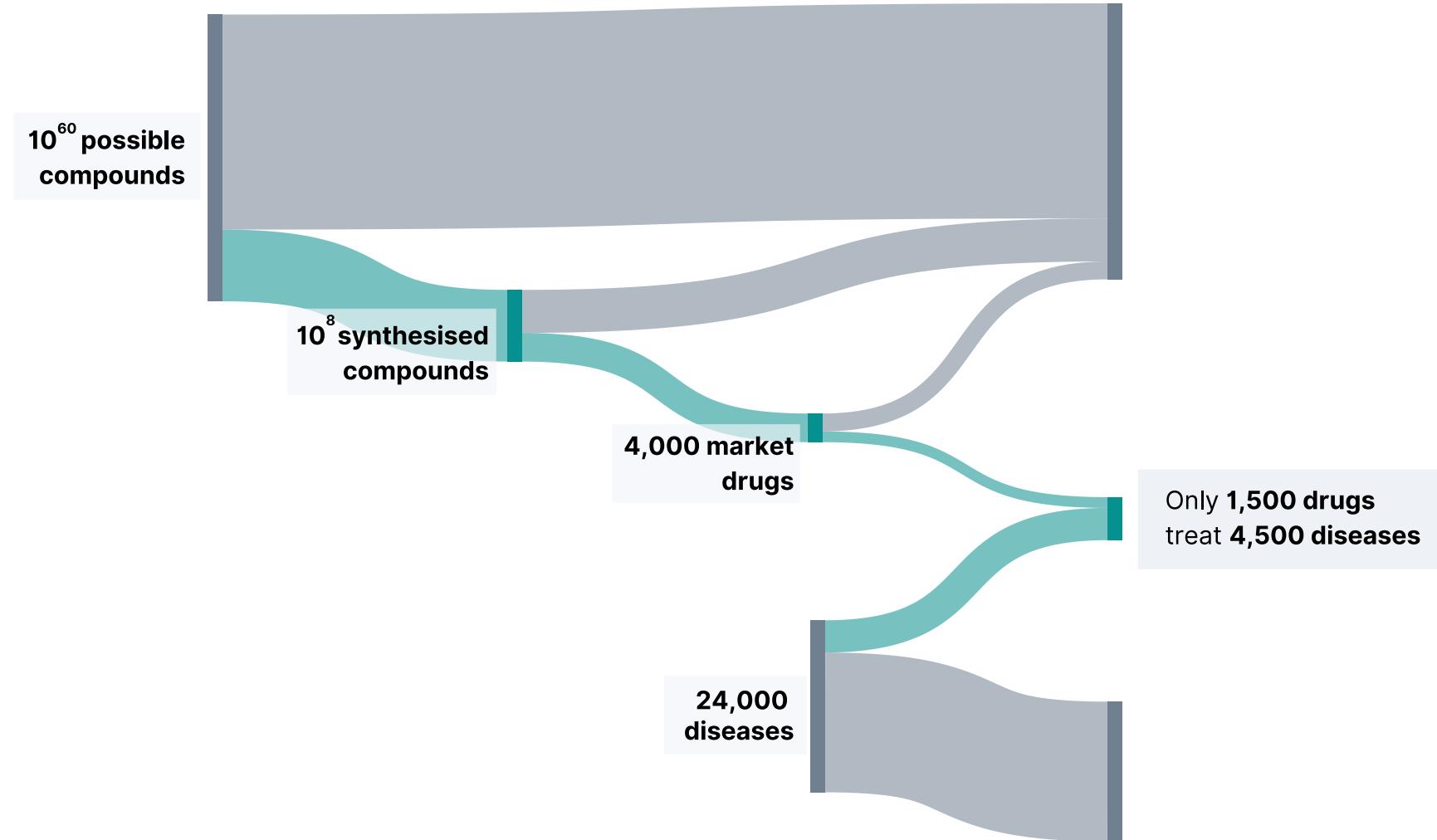


Can we use Big Data to find disease cures?



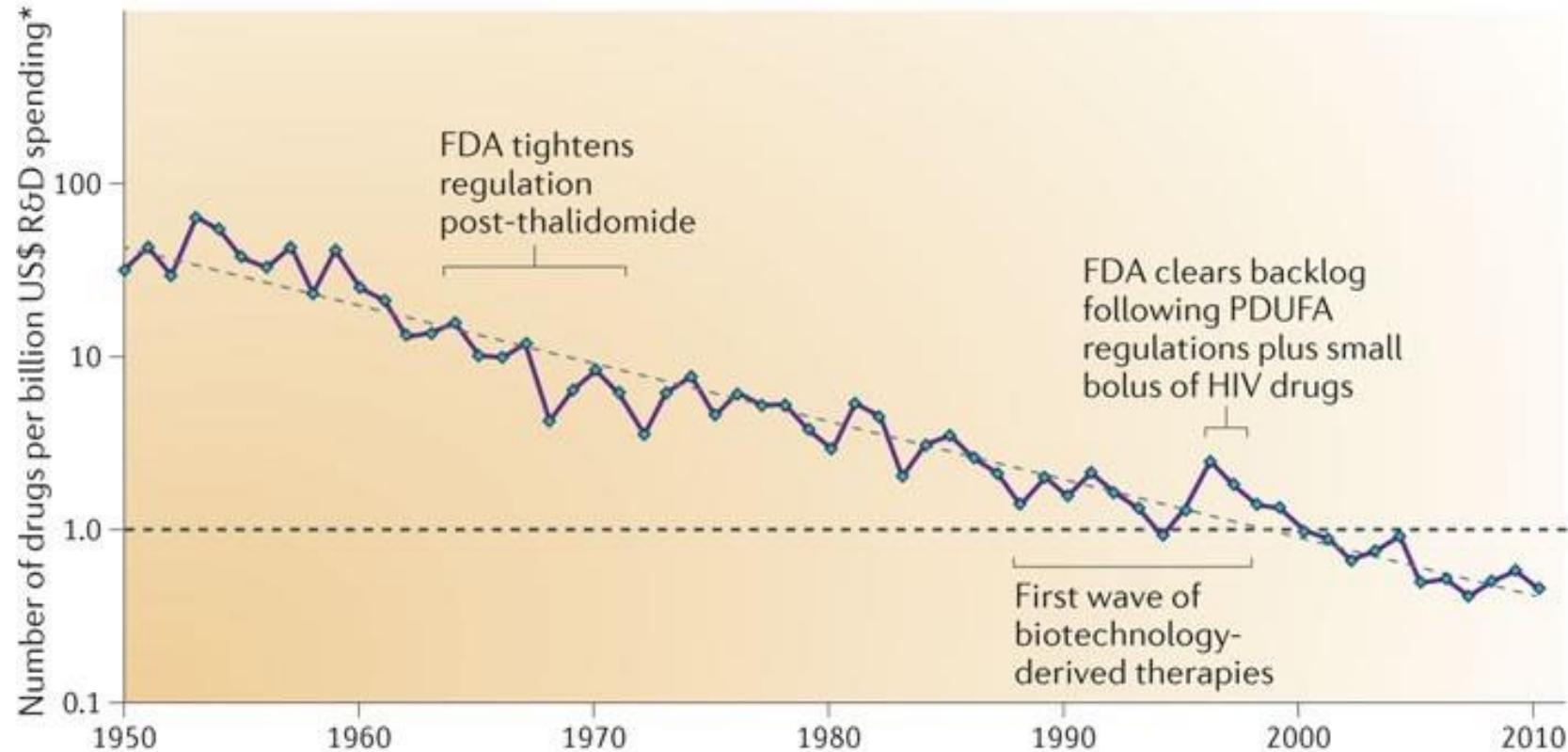
Disease understanding: Dealing with complex and unstructured big data in biomedical domain

Can we use Big Data to find disease cures? CTB



Can we use Big Data to find disease cures? CTB

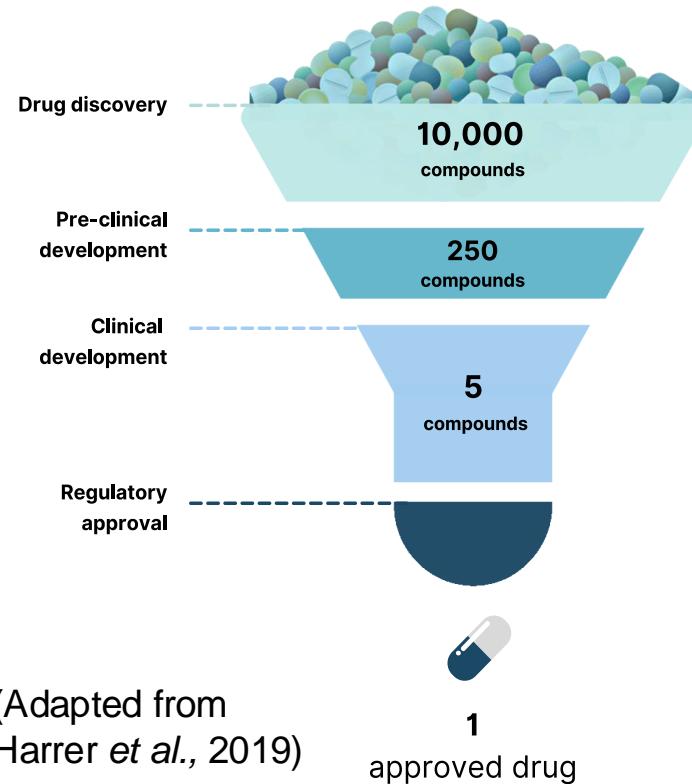
a Overall trend in R&D efficiency (inflation-adjusted)



(Scannell *et al.*, 2012)

Can we use Big Data to find disease cures? CTB

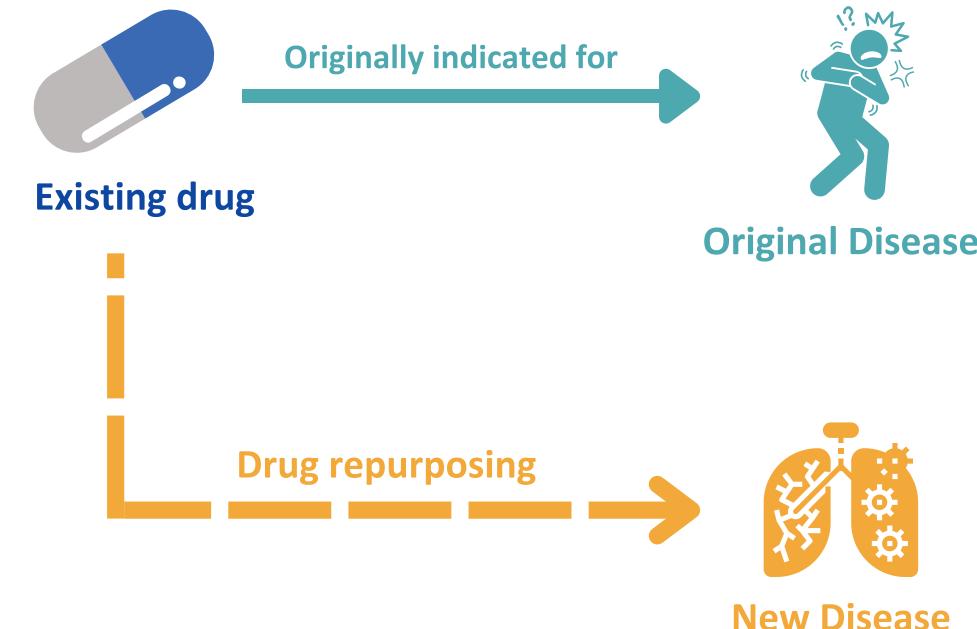
***De novo* drug design** is a tremendously time-consuming, costly, and difficult task.



Drug repurposing (DR) seeks to identify novel potential uses for existing drugs.

Potential alternative

- Reduce:
- Time
 - Cost
 - Risk



Can we use Big Data to find disease cures? CTB

A) De novo drug discovery



 10 - 15 years
 \$12 billion

B) Drug repurposing



 3 - 12 years
 ↓ 75%

(Adapted from Ashburn *et al.*, 2004)

Classical repurposing examples

Duloxetine

- Original indication: depression
- New indication: urinary incontinence

Retinoic acid

- Original indication: acne
- New indication: acute myeloid leukaemia

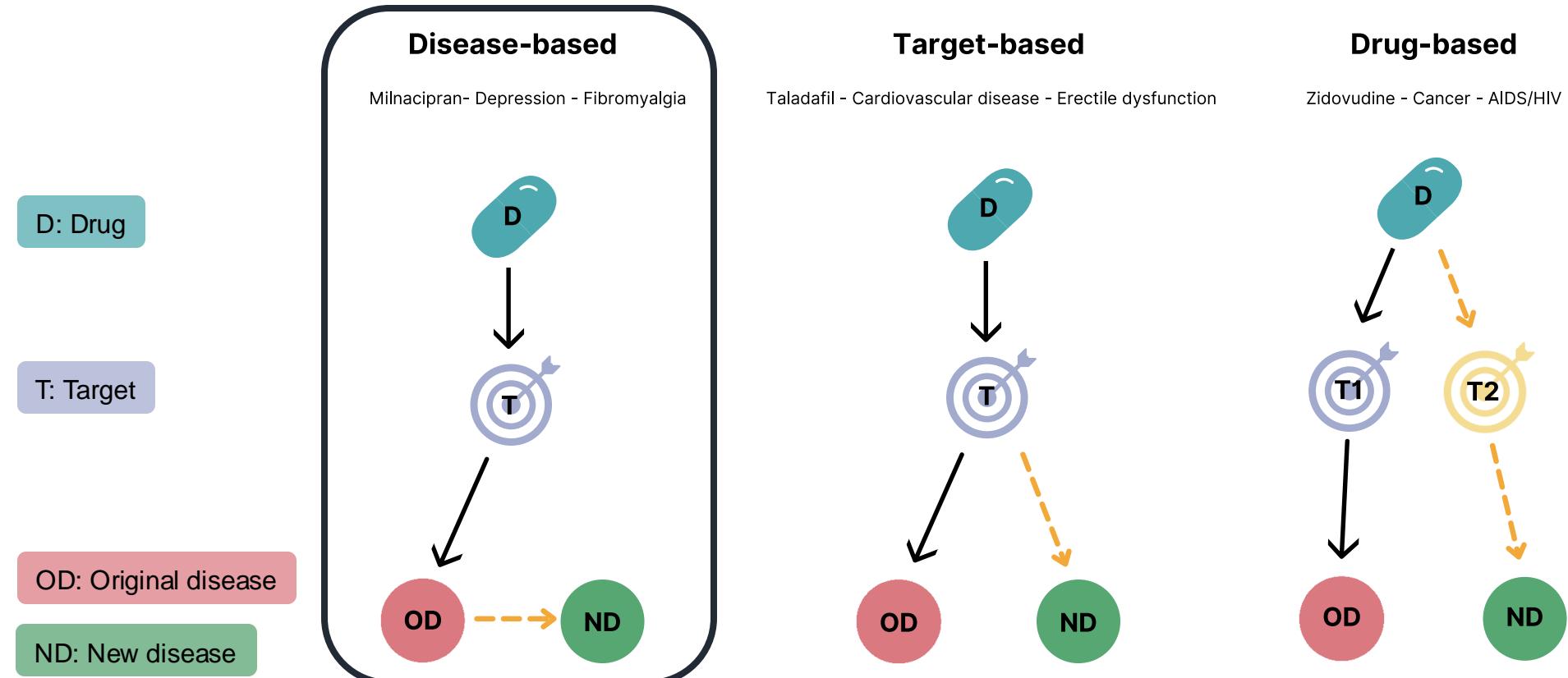
Sildenafil

- Original indication: angina pectoris
- New indication: erectile dysfunction

Thalidomide

- Original indication: anti-nausea
- New indication: leprosy

Strategies for repurosing



(Adapted from Parisi *et al.*, 2020)

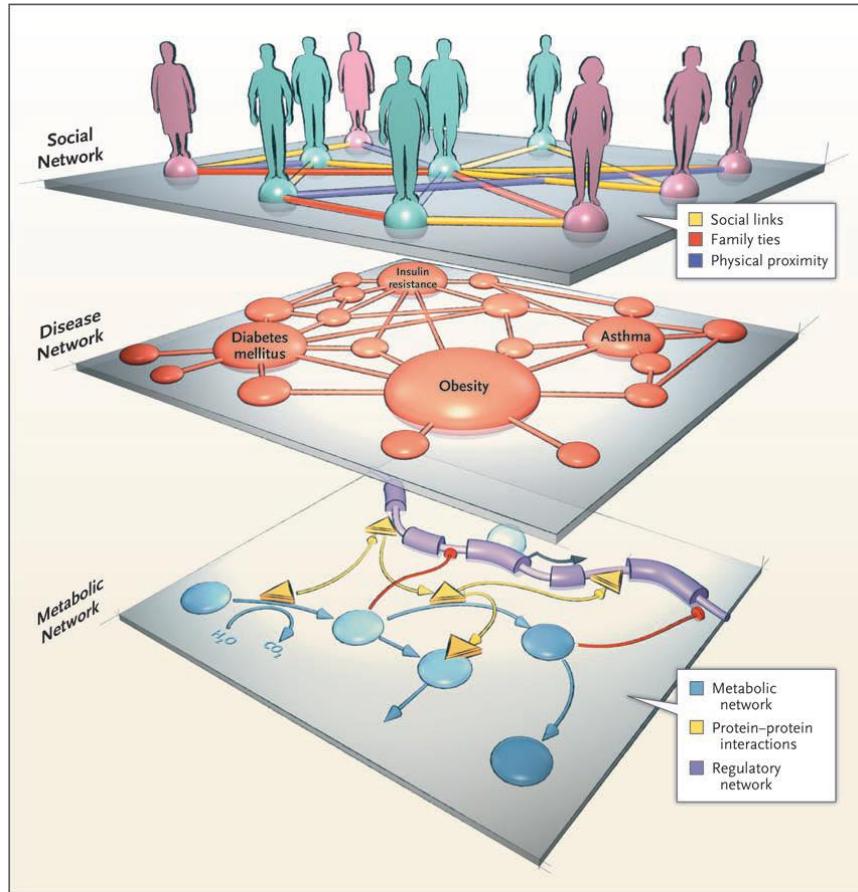


POLITÉCNICA



Network medicine

Network medicine



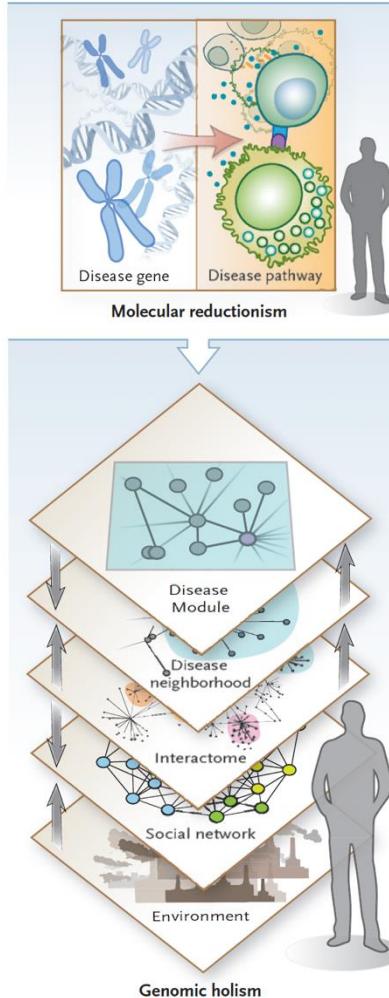
(Barabási, 2007)

Network Medicine – From Obesity to the “Diseasome”

New England Journal of Medicine, vol. 357, no. 4, pp. 404–407, Jul. 2007, doi: 10.1056/NEJMMe078114.

Network medicine offers a framework to explore systematically not only the molecular complexity of a particular disease but also the molecular relationships among apparently distinct (patho)phenotypes.

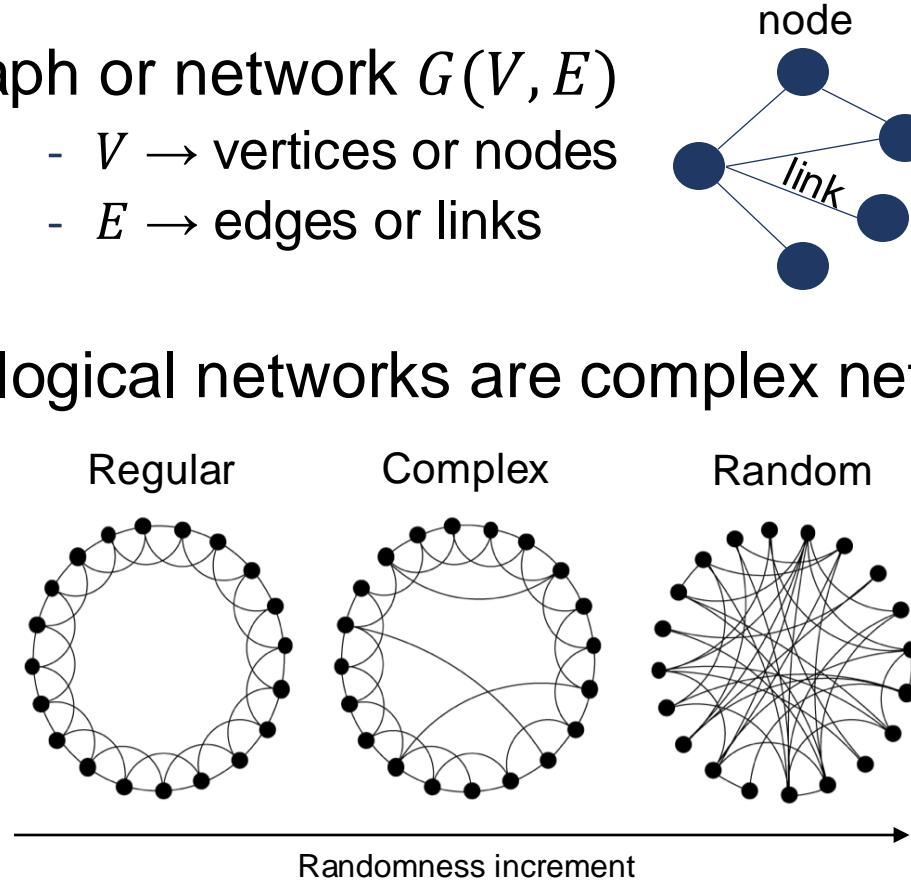
Network medicine



(Greene and Loscalzo, 2012)

Network medicine is based on **graph theory**

- Graph or network $G(V, E)$
 - $V \rightarrow$ vertices or nodes
 - $E \rightarrow$ edges or links
- Biological networks are complex networks



Network medicine

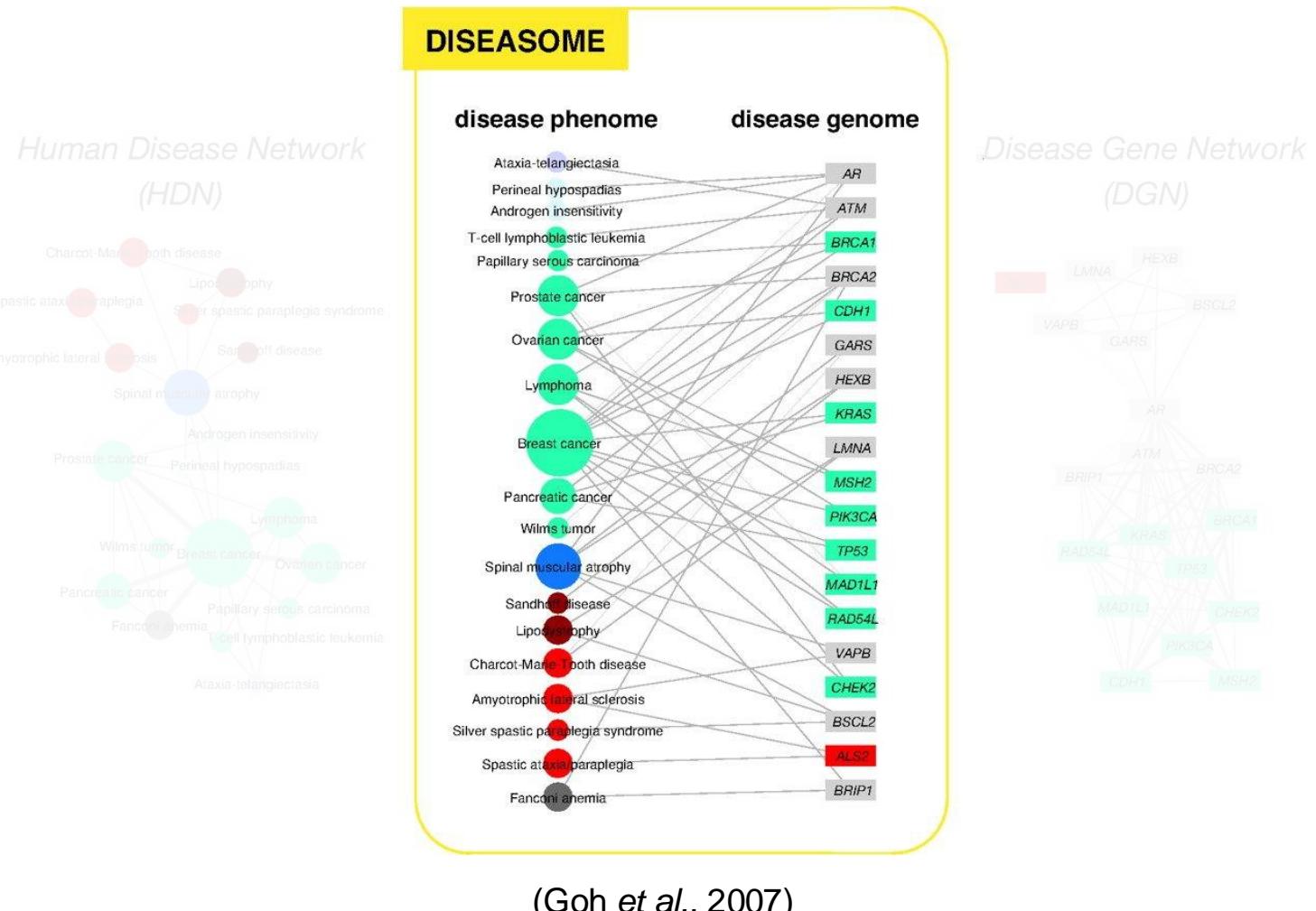
Two main objectives of network medicine

- Provide better understanding of human disease and its complexity
- Find better treatments

One of the pillars of network medicine → **Disease networks**

- Nodes: diseases
- Links: relationships between diseases

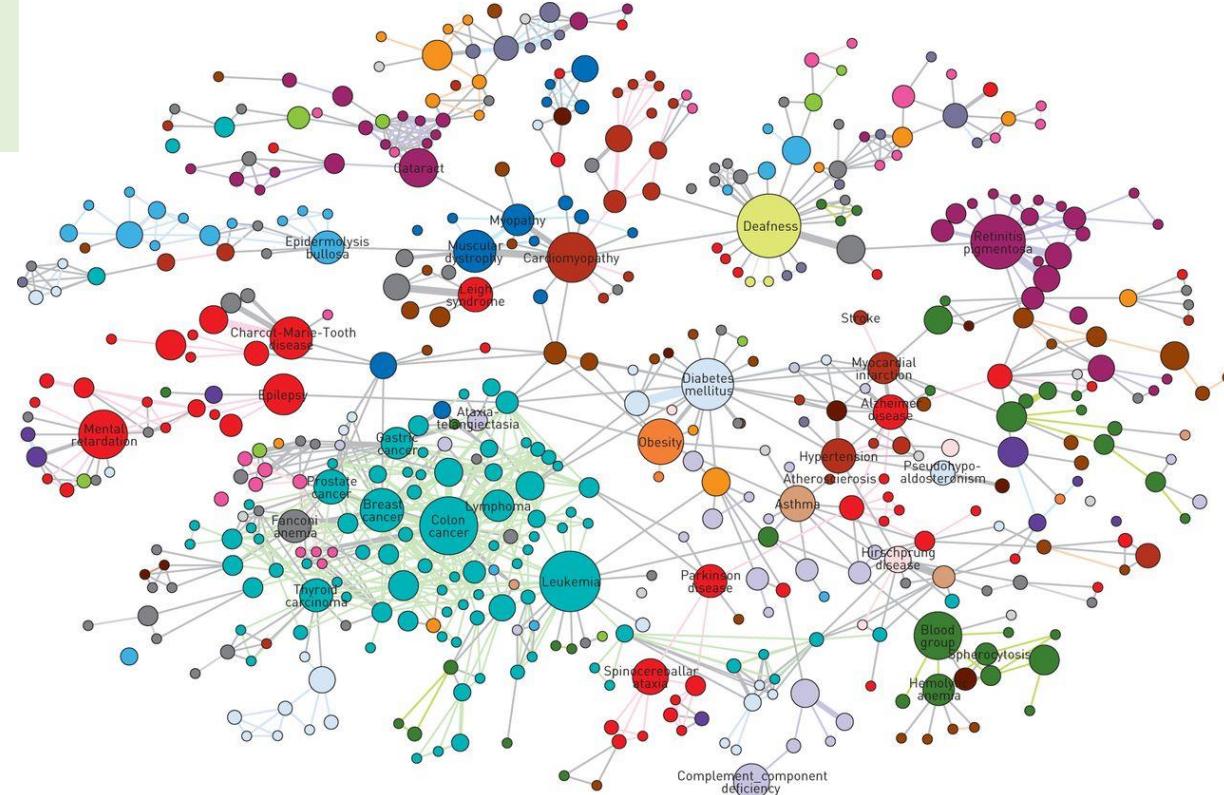
Network medicine



Network medicine

- Human Disease Network (HDN), 2007

Nodes: diseases
Links: genes



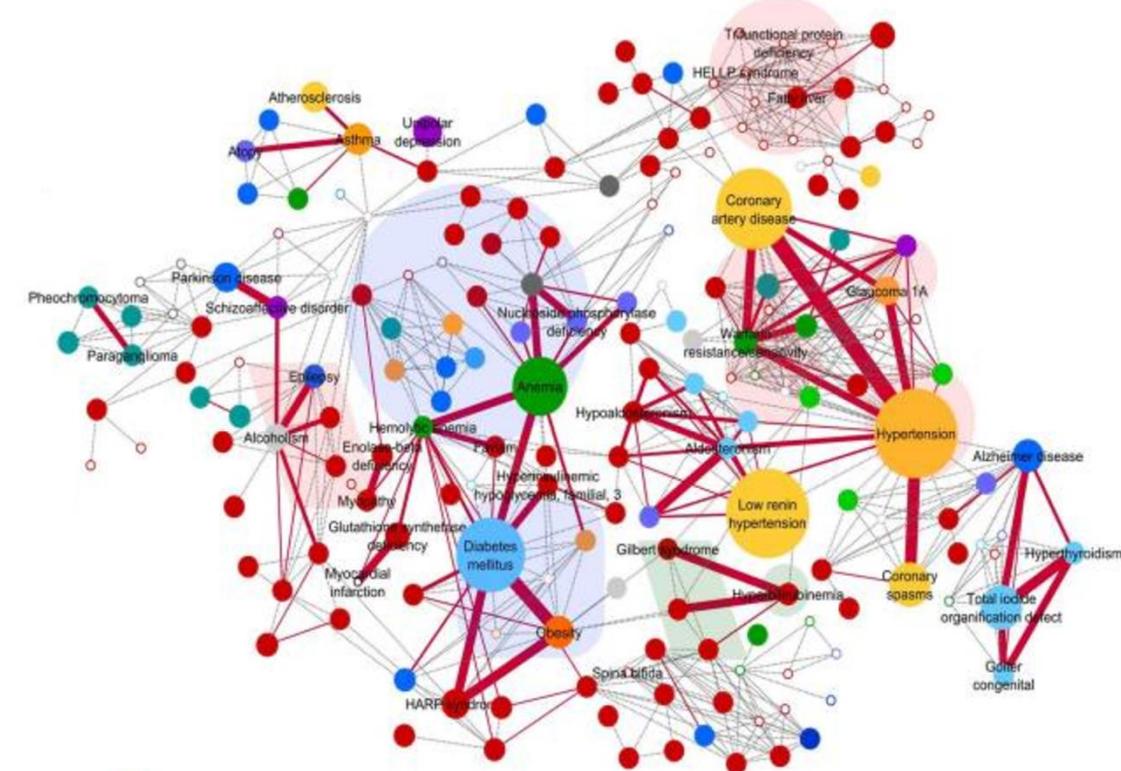
(Goh et al., 2007)

Network medicine

- Metabolism-based human Disease Network (MDN), 2008

Nodes: diseases

Links: enzymes



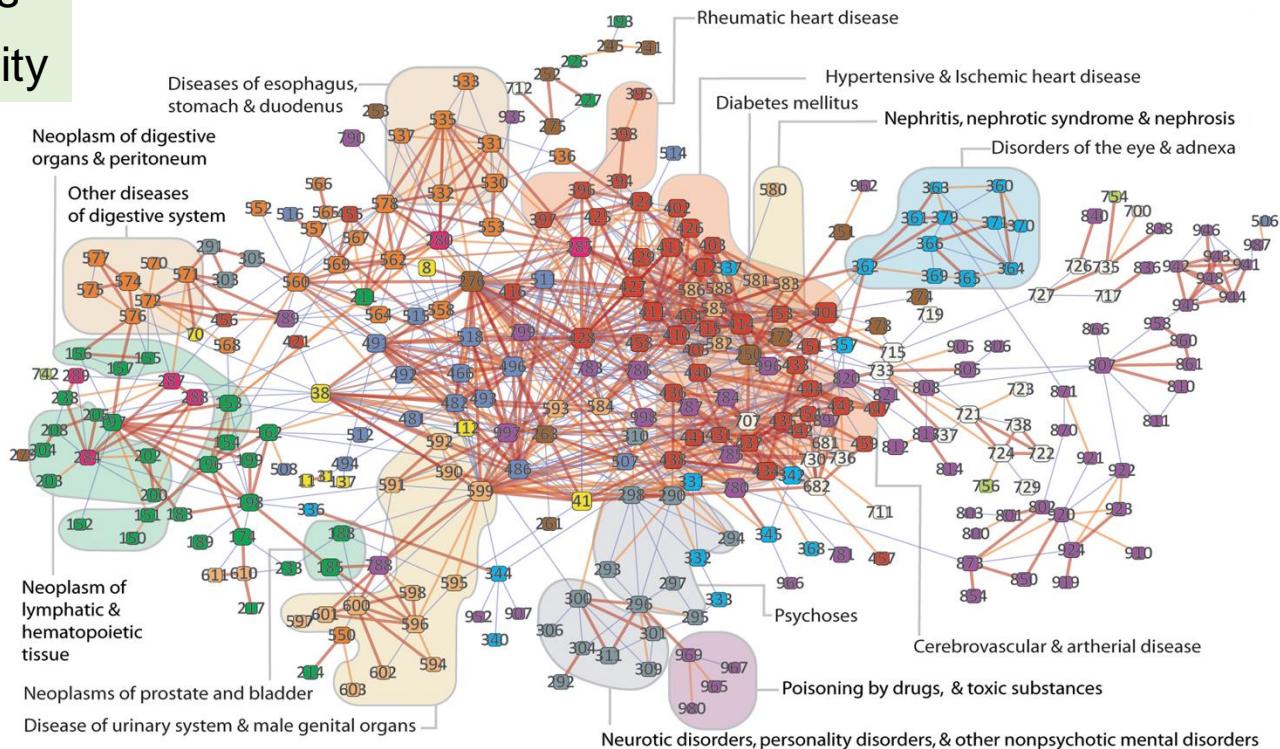
(Lee *et al.*, 2008)

Network medicine

- Phenotypic Disease Network (PDN), 2009

Nodes: diseases

Links: comorbidity



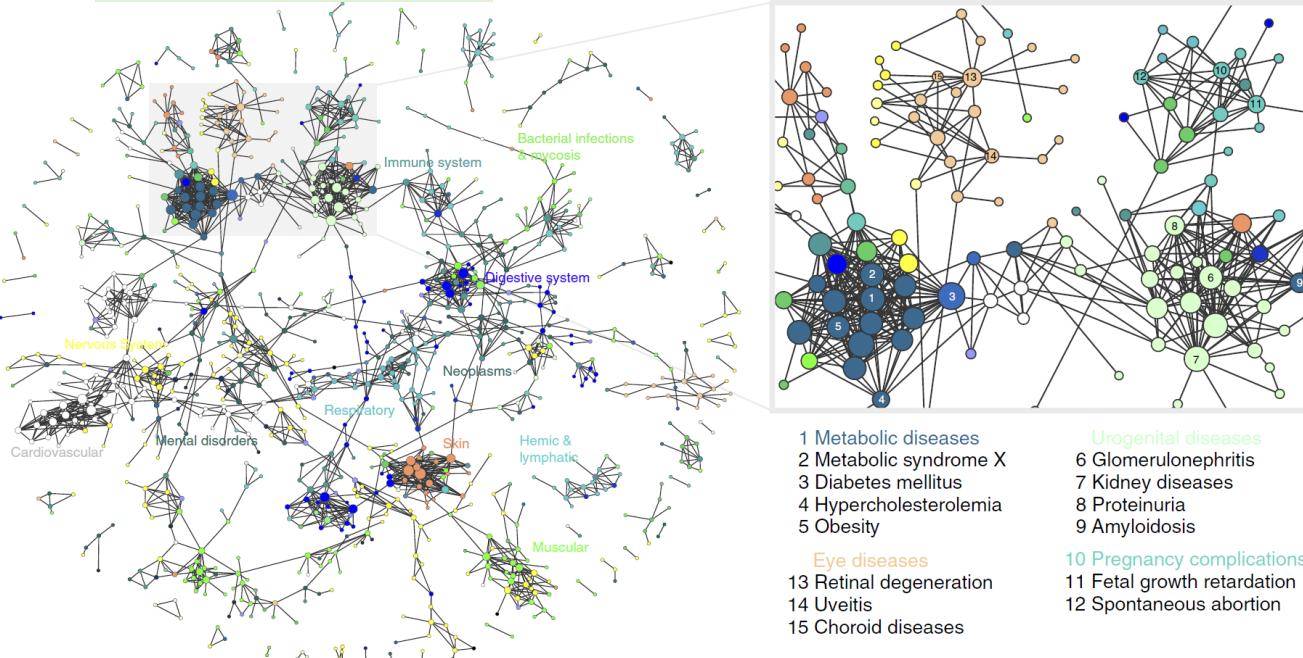
(Hidalgo *et al.*, 2009)

Network medicine

- Human Symptoms-Disease Network (HSDN), 2014

Nodes: diseases

Links: symptoms



(Zhou et al., 2014)

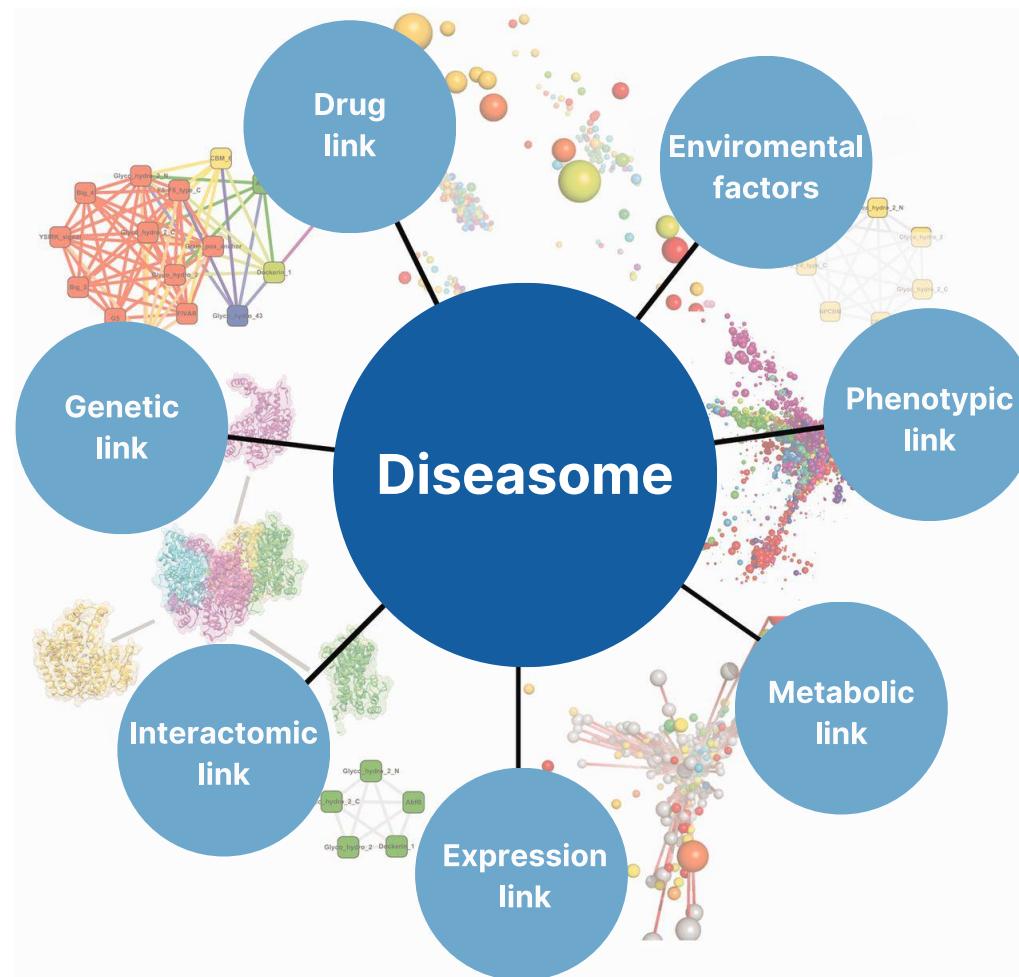
Given:

- Disease x – vector d_x of associated symptoms
- Disease y – vector d_y of associated symptoms

Cosine similarity between diseases is computed as

$$\cos(d_x, d_y) = \frac{\sum_i d_{x,i} d_{y,i}}{\sqrt{\sum_i d_{x,i}^2} \sqrt{\sum_i d_{y,i}^2}}$$

Network medicine



(Adapted from Goh and Choi, 2012)

Disease understanding: Dealing with complex and unstructured big data in biomedical domain



POLITÉCNICA



DISNET platform



Disease understanding: Dealing with complex and unstructured big data in biomedical domain



POLITÉCNICA



DISNET platform

*Drug repurposing and **D**ISease
understanding through complex **N**ETworks
creation and analysis*



<https://disnet.ctb.upm.es/>



Data-driven drug
repurposing
applying graph
neural networks



RTI2018-094576-A-I00

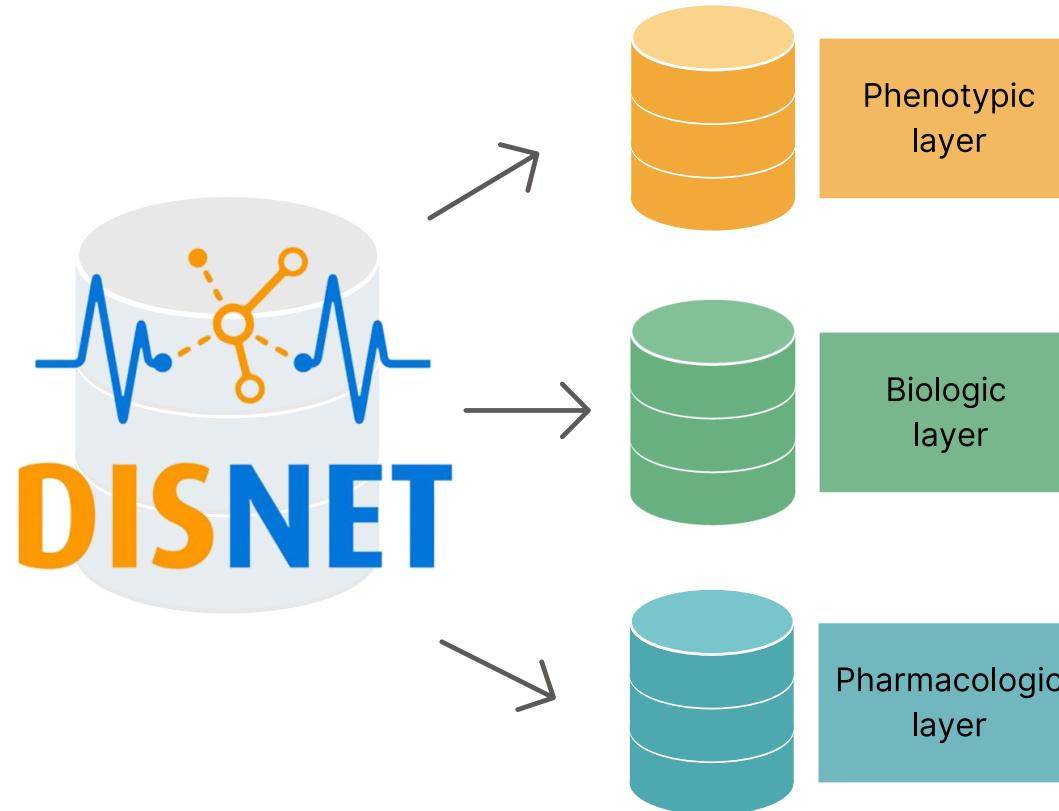


IND2019/TIC-17159



PID2021-122659OB-I00

DISNET platform

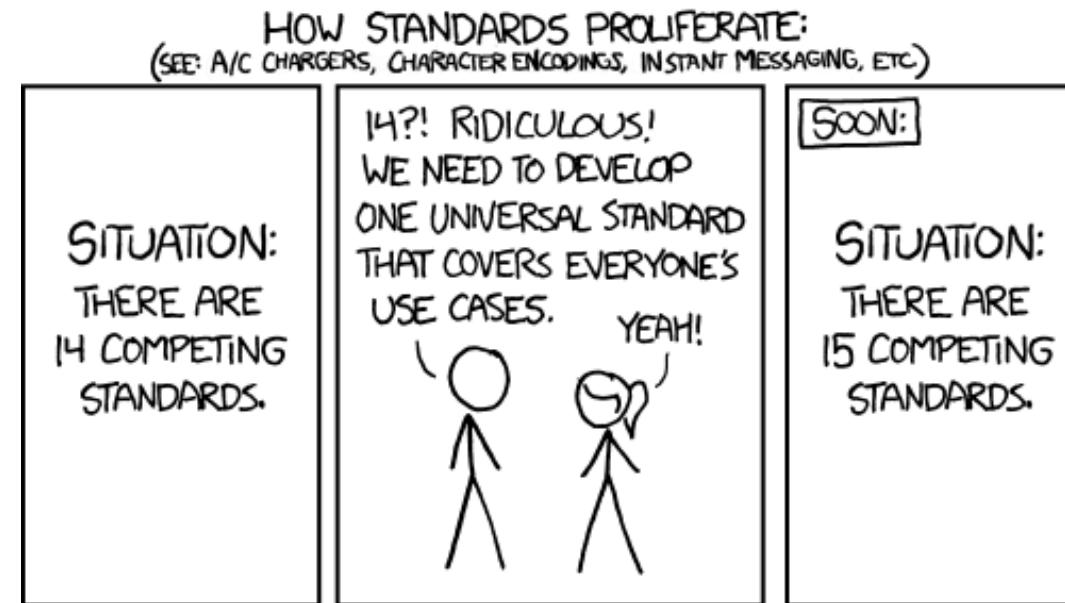


Node type	Link type
Disease Symptom	Disease - Symptom
Disease Gene Protein Genetic variant Biological pathway Marker ncRNA	Disease - Gene Disease - Variant Disease - Pathway Disease - Marker Disease - ncRNA Gene - Protein Protein - Protein (PPIs) Gene - Pathway
Disease Drug Drug target Phenotype effects	Disease - Drug Drug - Drug (DDIs) Drug - Target Drug - Phenotype (Indication) Drug - Phenotype (Side effect)

Integrating biomedical knowledge

Medical standard vocabularies and terminologies

“Finite collection of terms used to represent as unambiguously as possible medical meaning”



Source: <https://xkcd.com/927/>

Integrating biomedical knowledge

Medical standard vocabularies and terminologies



Leading healthcare terminology, worldwide



MVX



Clinical
Classifications
Software (CCS) for
ICD-10-CM/PCS



HCPCS
Level II
2017
Professional



DISNET platform

Layer	Node type	Identifiers from	Link type	Data source
Phenotypic	Disease Symptom	UMLS UMLS	Disease - Symptom	Wikipedia + MayoClinic + PubMed
Biologic	Disease Gene Protein Genetic variant Biological pathway Marker ncRNA	UMLS Entrez UniProt Entrez WikiPathways MarkerDB HMDD + LncRNADisease	Disease - Gene Disease - Variant Disease - Pathway Disease - Marker Disease - ncRNA Gene - Protein Protein - Protein (PPIs) Gene - Pathway	DisGeNET DisGeNET WikiPathways MarkerDB HMDD + LncRNADisease Entrez NextProt WikiPathway
Pharmacologic	Disease Drug Drug target Phenotype effects	UMLS ChEMBL ChEMBL UMLS	Disease - Drug (Therapeutic) Disease - Drug (Marker) Drug - Drug (DDIs) DDIs - Phenotype Drug - Target Drug - Phenotype (Indication) Drug - Phenotype (Side effect)	CTD CTD DrugBank TWOSIDES ChEMBL ChEMBL SIDER

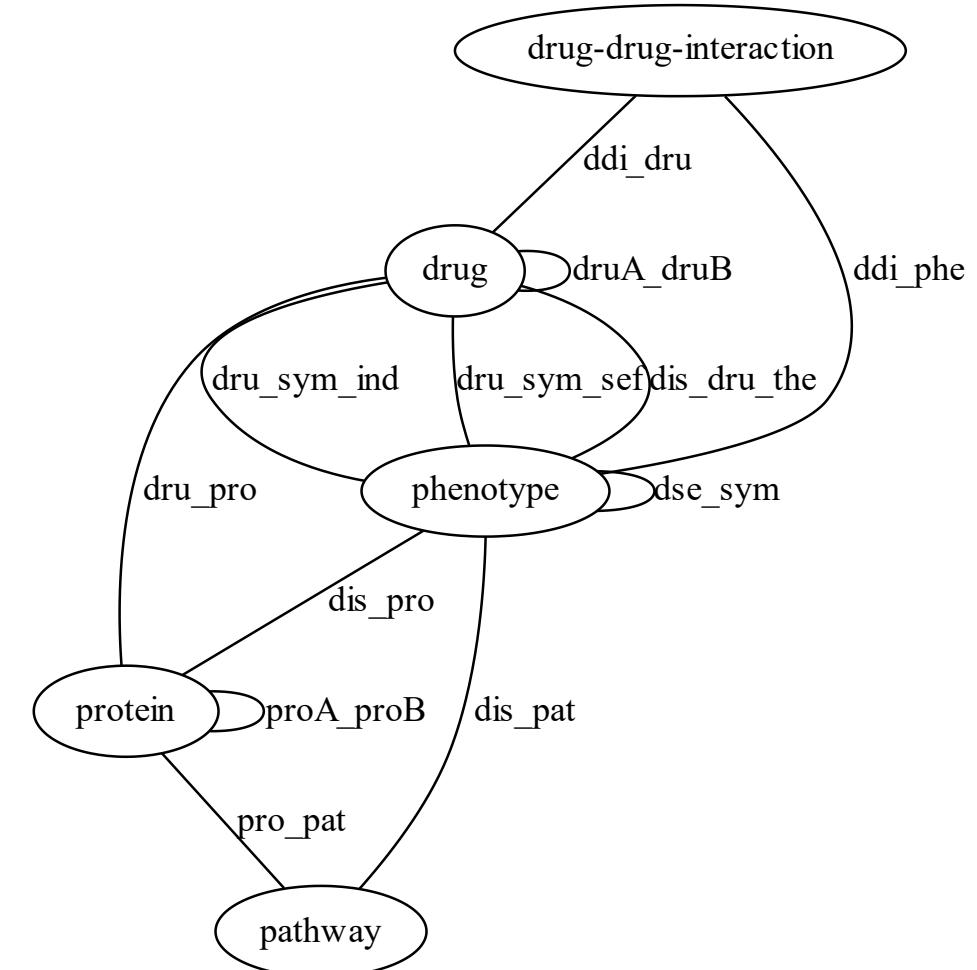
DISNET platform

Layer	Node type	Count	Link type	Count
Phenotypic	Disease	9,715	Disease - Symptom	201,541
	Symptom	1,999		
Biologic	Disease	24,314	Disease - Gene	358,209
	Gene	20,610	Disease - Variant	210,498
	Protein	18,521	Disease - Pathway	664
	Genetic variant	117,337	Disease - Marker	2,763
	Biological pathway	1,105	Disease - ncRNA	49,711
	Marker	618	Gene - Protein	15,770
	ncRNA	19,913	Protein - Protein (PPIs)	240,585
			Gene - Pathway	10,736
Pharmacologic	Disease	8,366	Disease - Drug (Therapeutic)	56,717
	Drug	3,944	Disease - Drug (Marker)	104,317
	Drug target	5,627	Drug - Drug (DDIs)	614,665
	Phenotype effects	8,374	DDIs - Phenotype	85,152
			Drug - Target	7,902
			Drug - Phenotype (Indication)	10,475
			Drug - Phenotype (Side effect)	52,168

DISNET knowledge graph

Number of nodes: 153,747

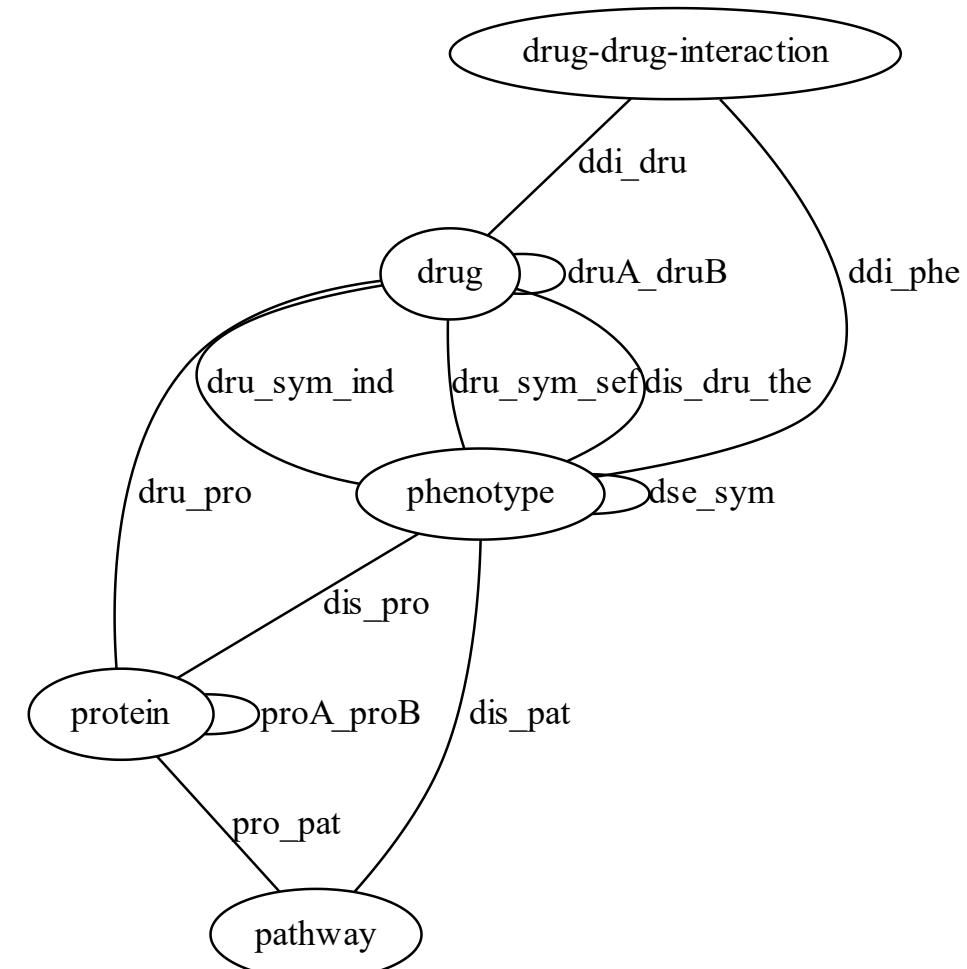
- Phenotype: 30,731
- Drug: 3,944
- Pathway: 1,105
- Protein: 18,521
- DDI: 99,446



DISNET knowledge graph

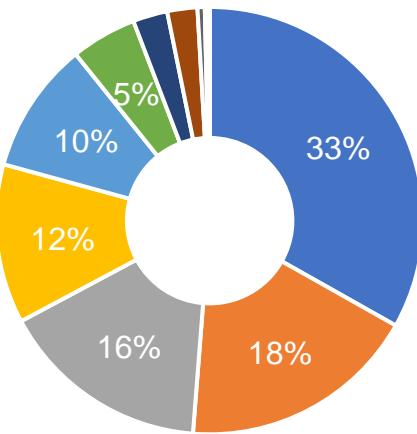
Number of edges: 1,996,658

- Disease-Protein: 360,985
- Disease-Drug (therapeutic): 52,179
- Disease-Symptom: 318,550
- Disease-Pathway: 424
- Drug-Drug: 662,281
- Drug-Protein: 5,946
- Drug-Symptom (side effect): 45,516
- Drug-Symptom (indication): 863
- Protein-Protein: 240,585
- Protein-Pathway: 10,991
- DDI-Phenotype: 99,446
- DDI-Drug: 198,892

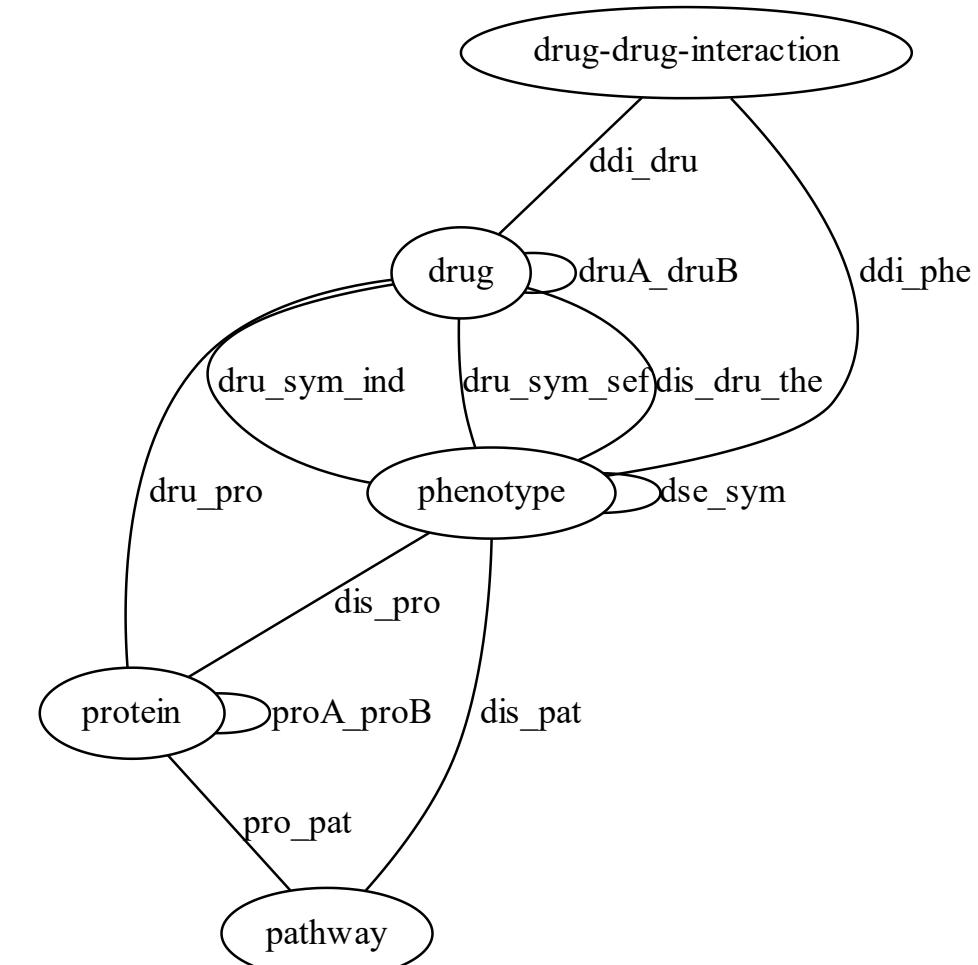


DISNET knowledge graph

Distribution of the number of edges



- Drug-Drug (33.17 %)
- Disease-Protein (18.08 %)
- Disease-Symptom (15.95 %)
- Protein-Protein (12.05 %)
- DDI-Drug (9.96 %)
- DDI-Phenotype (4.98 %)
- Disease-Drug (therapeutic) (2.61 %)
- Drug-Symptom (side effect) (2.28 %)
- Protein-Pathway (0.55 %)
- Drug-Protein (0.30 %)
- Drug-Symptom (indication) (0.04 %)
- Disease-Pathway (0.02 %)



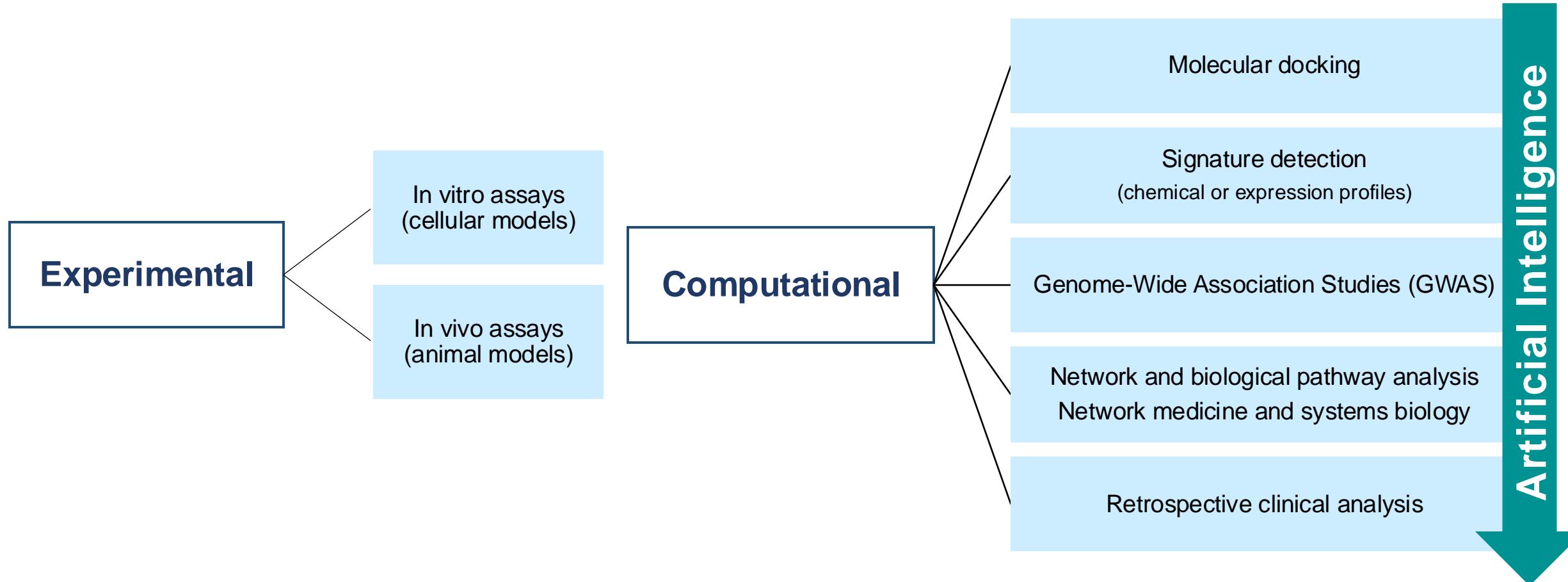


POLITÉCNICA



Drug repurposing

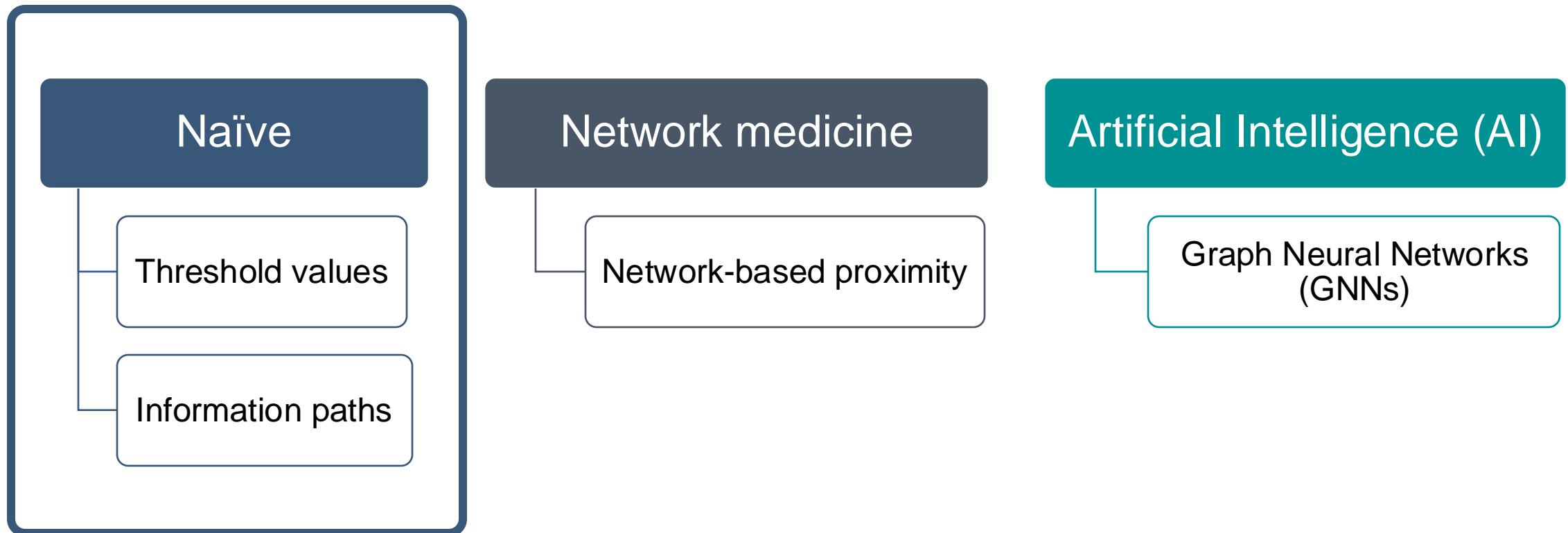
Drug repurposing



(Adapted from Low *et al.*, 2020)

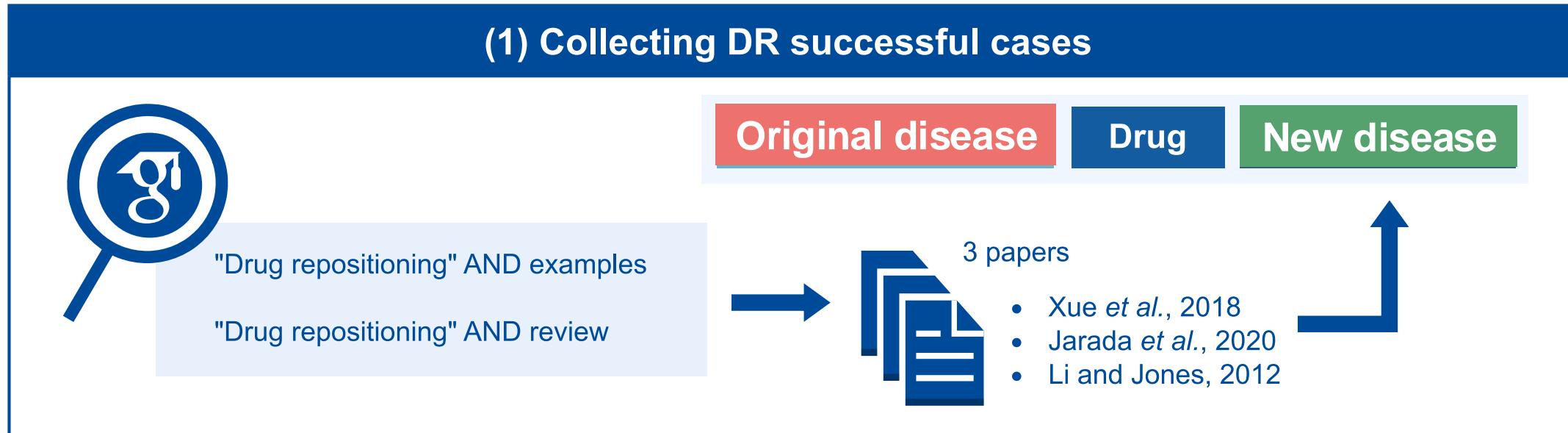
Drug repurposing

Some of the drug repurposing methodologies that we implement



Drug repurposing – Threshold values

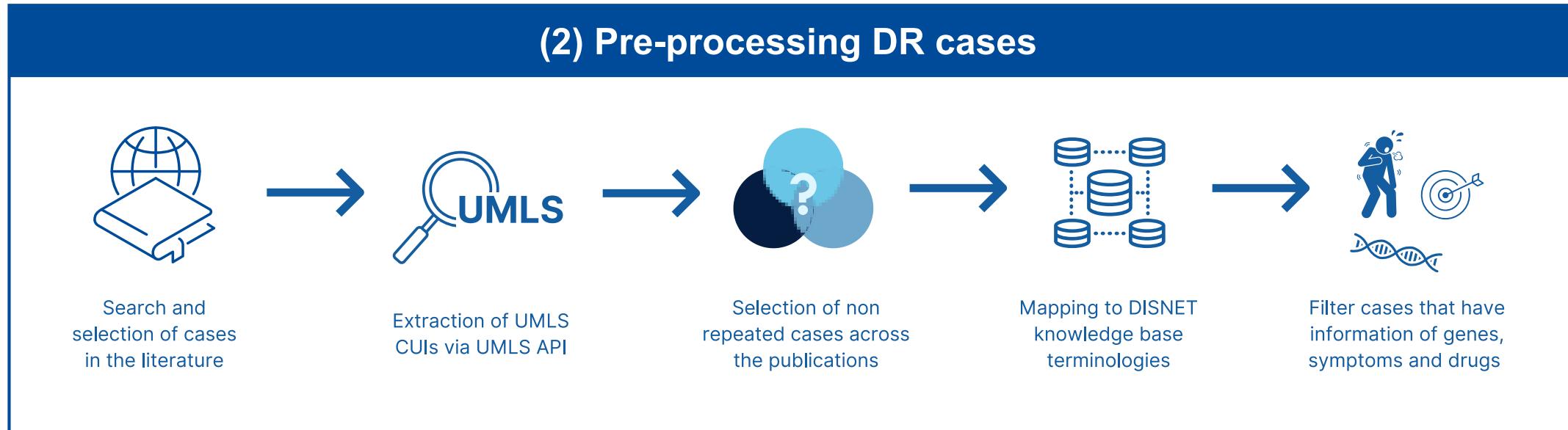
Uncovering the differences between DR and non-DR data



(Adapted from Prieto Santamaría *et al.*, 2021)

Drug repurposing – Threshold values

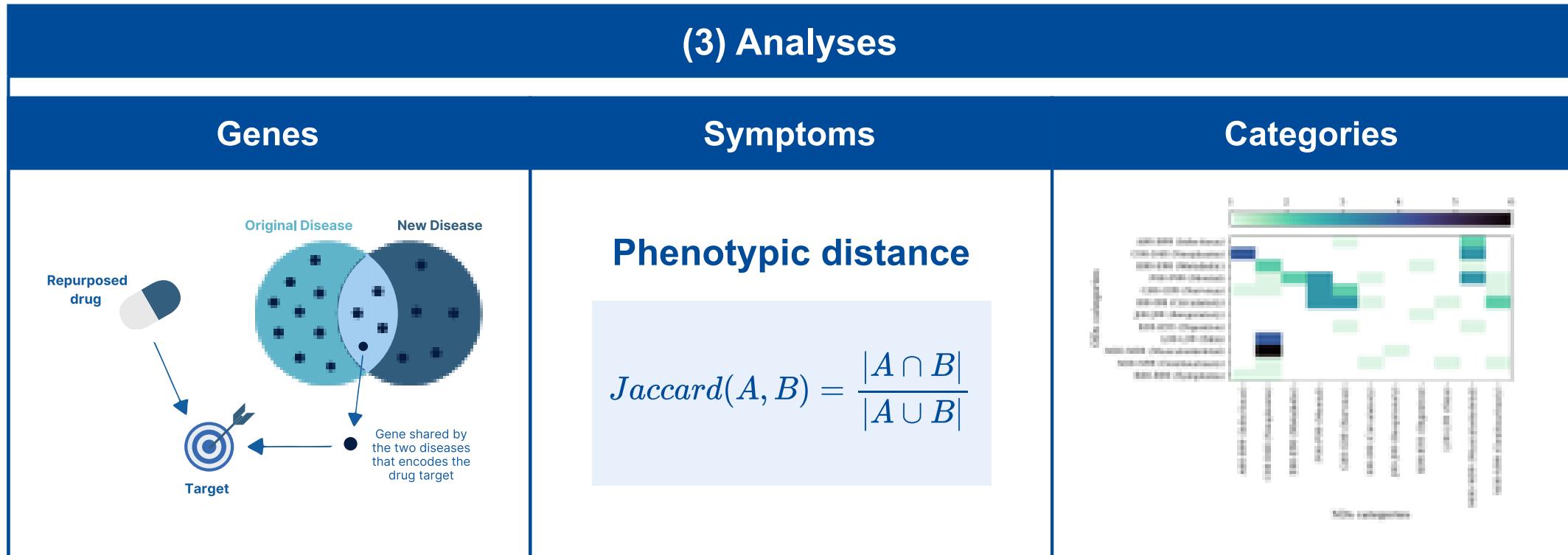
Uncovering the differences between DR and non-DR data



(Adapted from Prieto Santamaría et al., 2021)

Drug repurposing – Threshold values

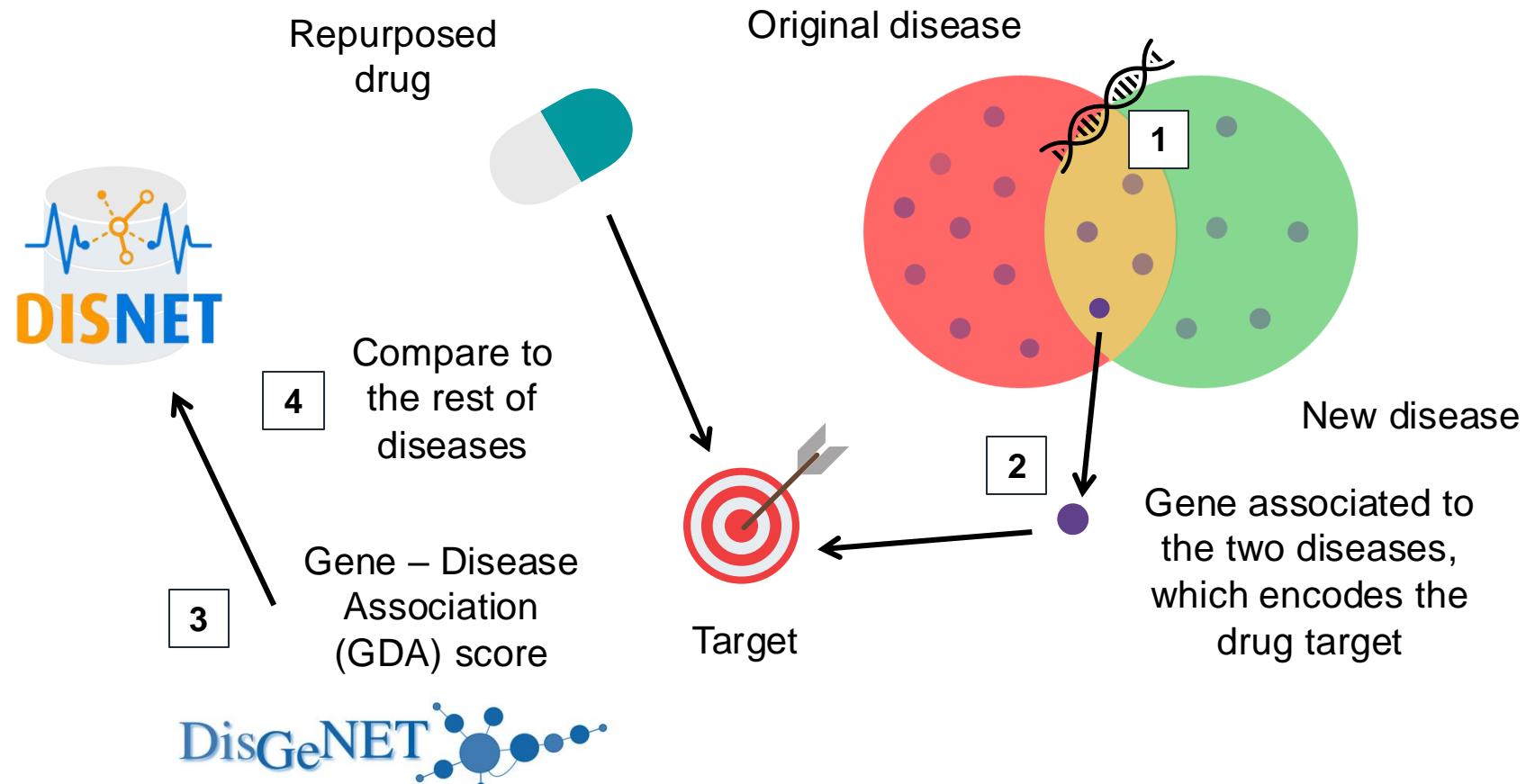
Uncovering the differences between DR and non-DR data



(Adapted from Prieto Santamaría et al., 2021)

Drug repurposing – Threshold values

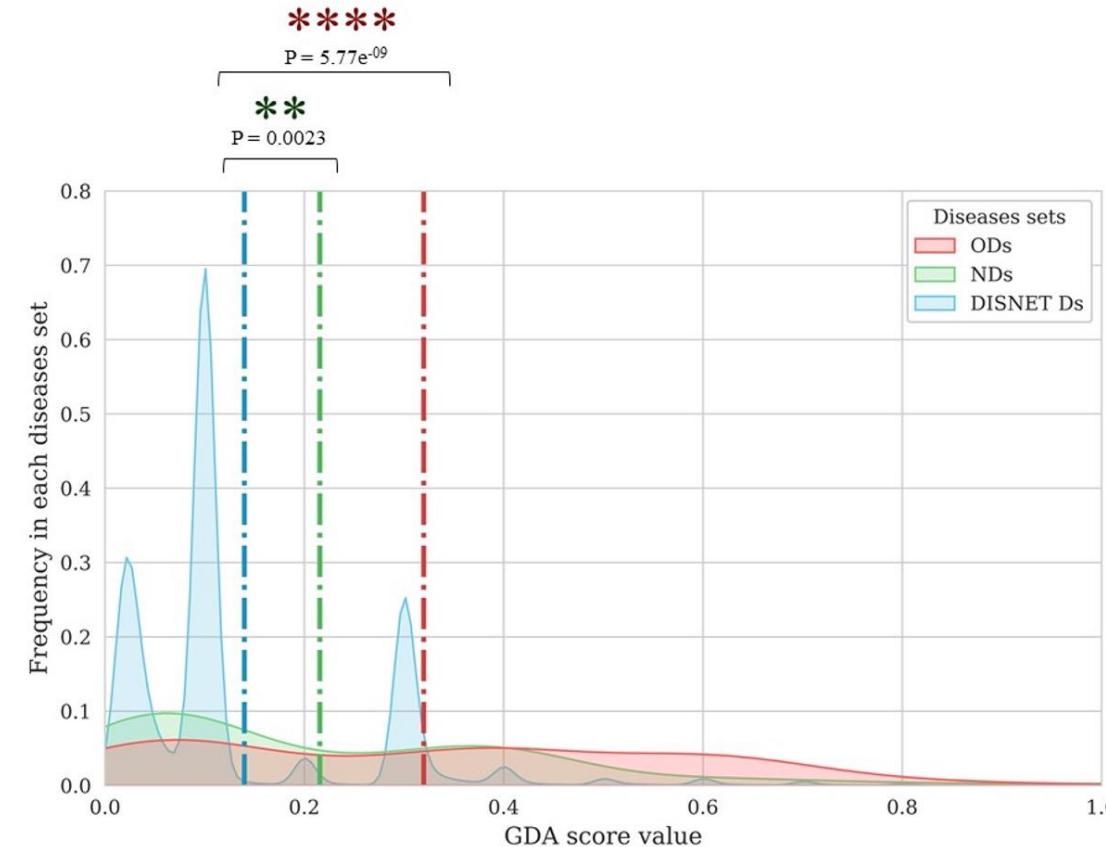
Uncovering the differences between DR and non-DR data



(Adapted from Prieto Santamaría et al., 2021)

Drug repurposing – Threshold values

Uncovering the differences between DR and non-DR data

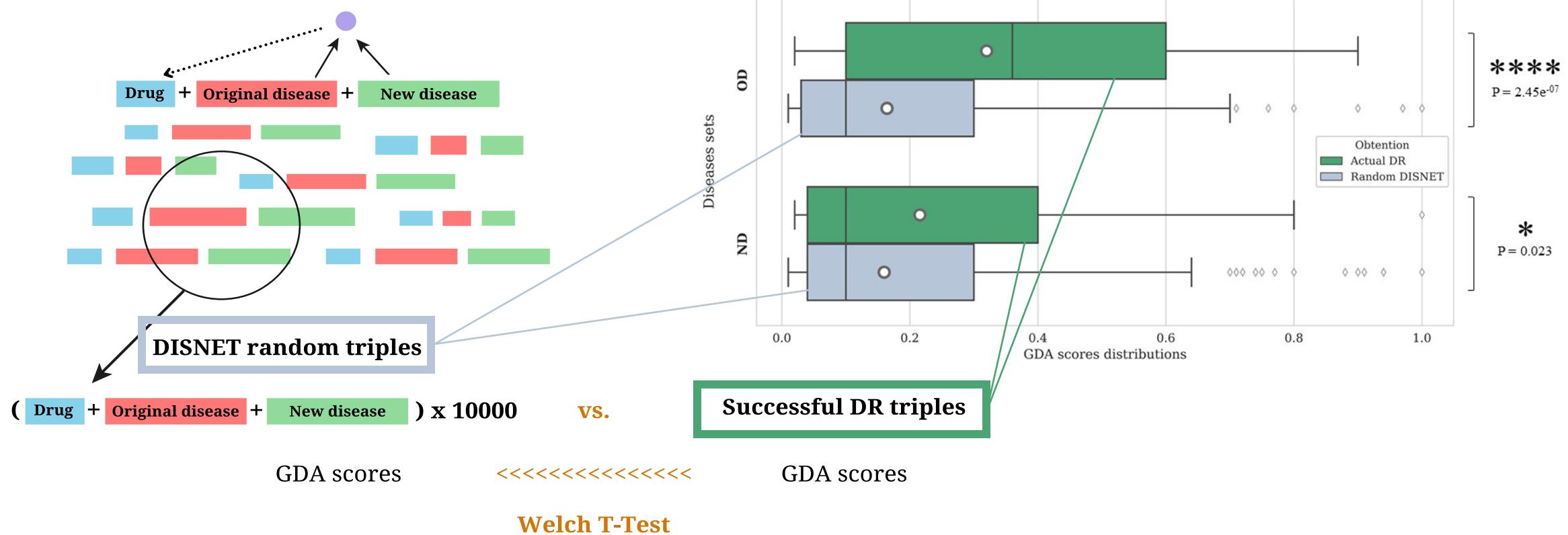


(Adapted from Prieto Santamaría *et al.*, 2021)

Disease understanding: Dealing with complex and unstructured big data in biomedical domain

Drug repurposing – Threshold values

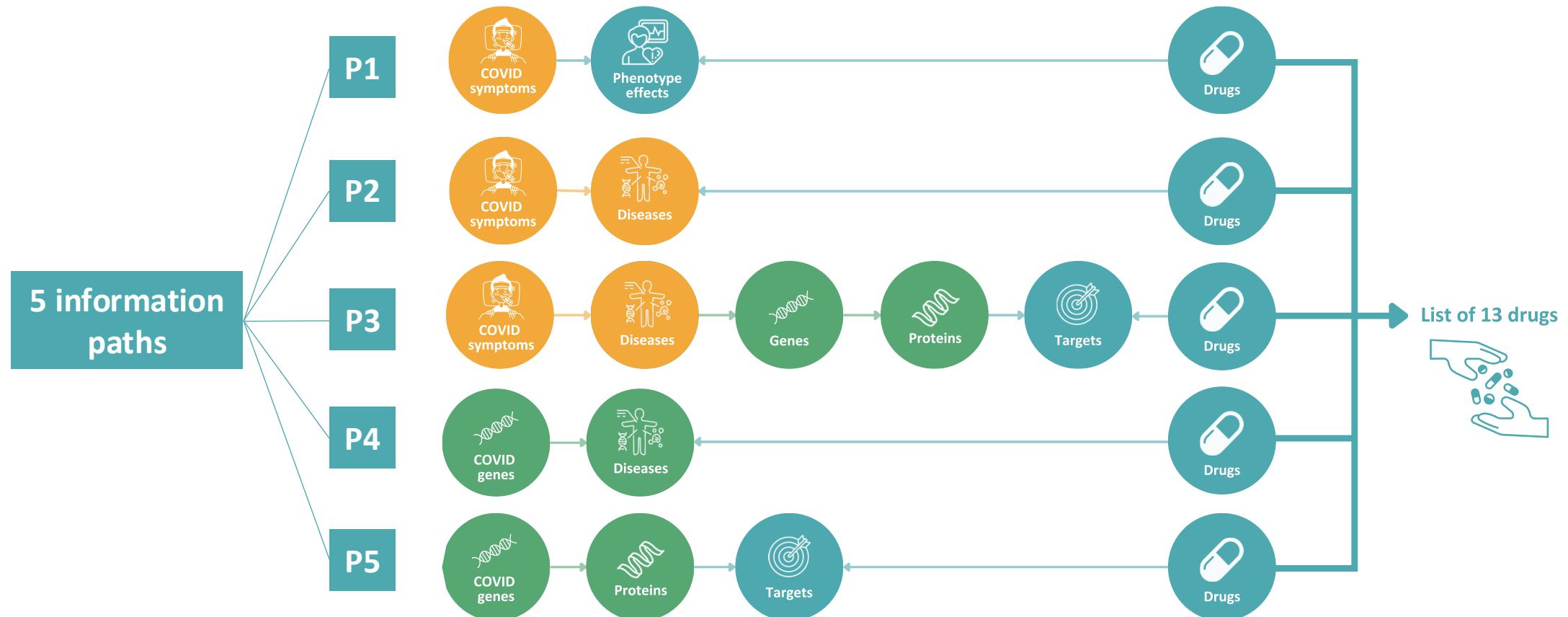
Uncovering the differences between DR and non-DR data



(Adapted from Prieto Santamaría et al., 2021)

Drug repurposing – Information paths

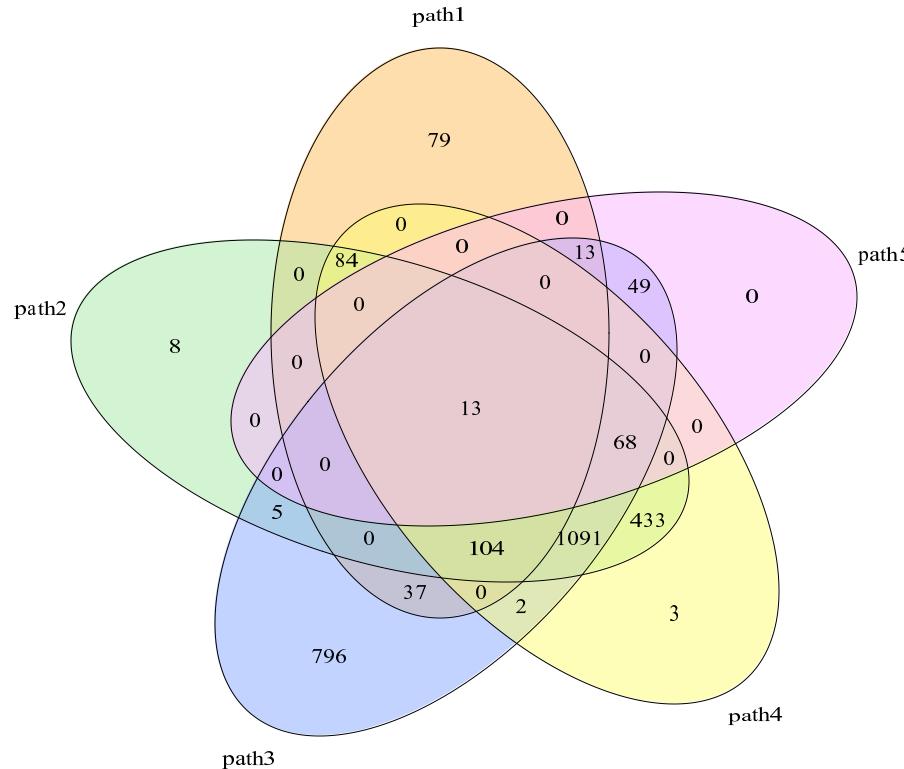
Suggesting repurposable candidates to treat **COVID-19**



(Adapted from Prieto Santamaría et al., 2022)

Drug repurposing – Information paths

Suggesting repurposable candidates to treat **COVID-19**



(Prieto Santamaría *et al.*, 2022)

Drug name	MeSH-PA therapeutic uses	COVID-19 associated target
ALDESLEUKIN	Antineoplastic Agents, Anti-Infective Agents	IL2RA
CANDESARTAN CILEXETIL	Cardiovascular Agents	AGTR1
CEFAZOLIN	Anti-Infective Agents	IL2
ENALAPRIL	Cardiovascular Agents	ACE
EPINEPHRINE	Cardiovascular Agents, Respiratory System Agents	TNF
EVEROLIMUS	Antineoplastic Agents	MTOR
HYDROXYCHLOROQUINE	Antirheumatic Agents, Anti-Infective Agents	ACE2
LOSARTAN	Cardiovascular Agents	AGTR1
MINOCYCLINE	Anti-Infective Agents	IL1B
RAMIPRIL	Cardiovascular Agents	ACE
SIROLIMUS	Antineoplastic Agents, Anti-Infective Agents	MTOR
SITAGLIPTIN	-	DPP4
VILDAGLIPTIN	-	DPP4

Drug repurposing – Information paths

Suggesting repurposable candidates to treat **COVID-19**

Aldesleukin 

Cefazolin 

Epinephrine 

Everolimus 

Sirolimus 

Hydroxychloroquine 

Losatran 

Candesartan 

Enalapril 

Ramipril 

Minocycline 

Stagliptin 

Vidagliptin 

Scientific reviews 

In-vitro studies 

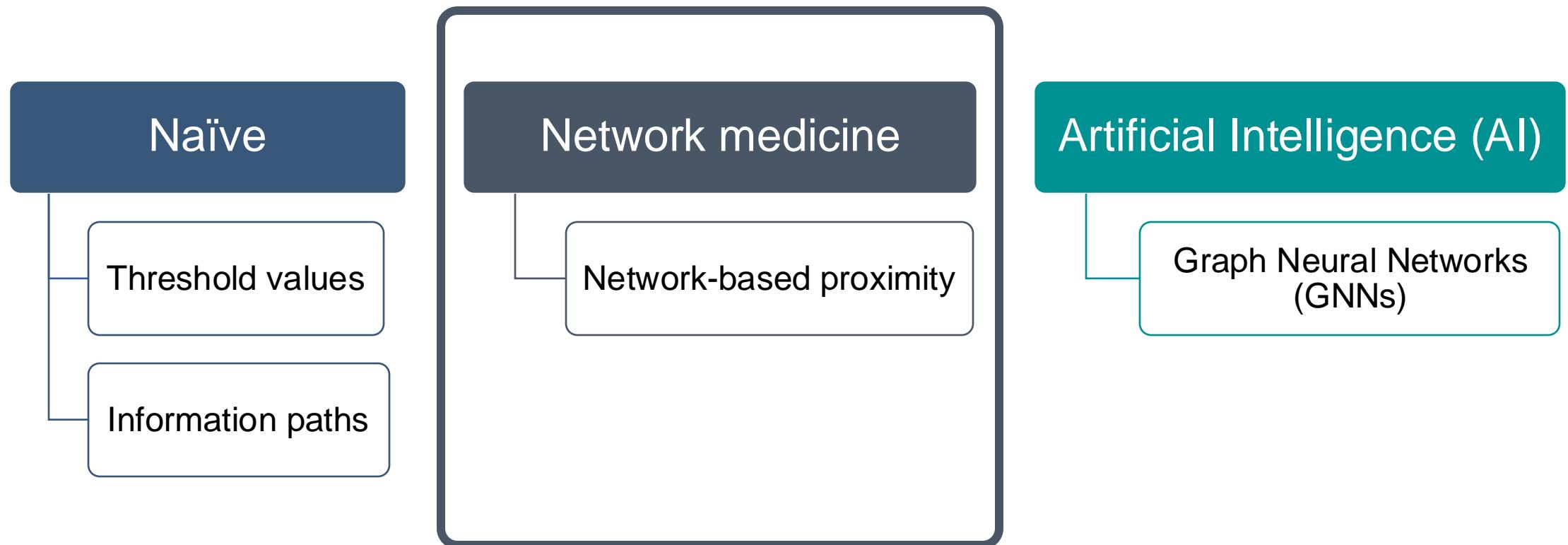
Computational studies 

Clinical trials 

(Prieto Santamaría *et al.*, 2022)

Drug repurposing

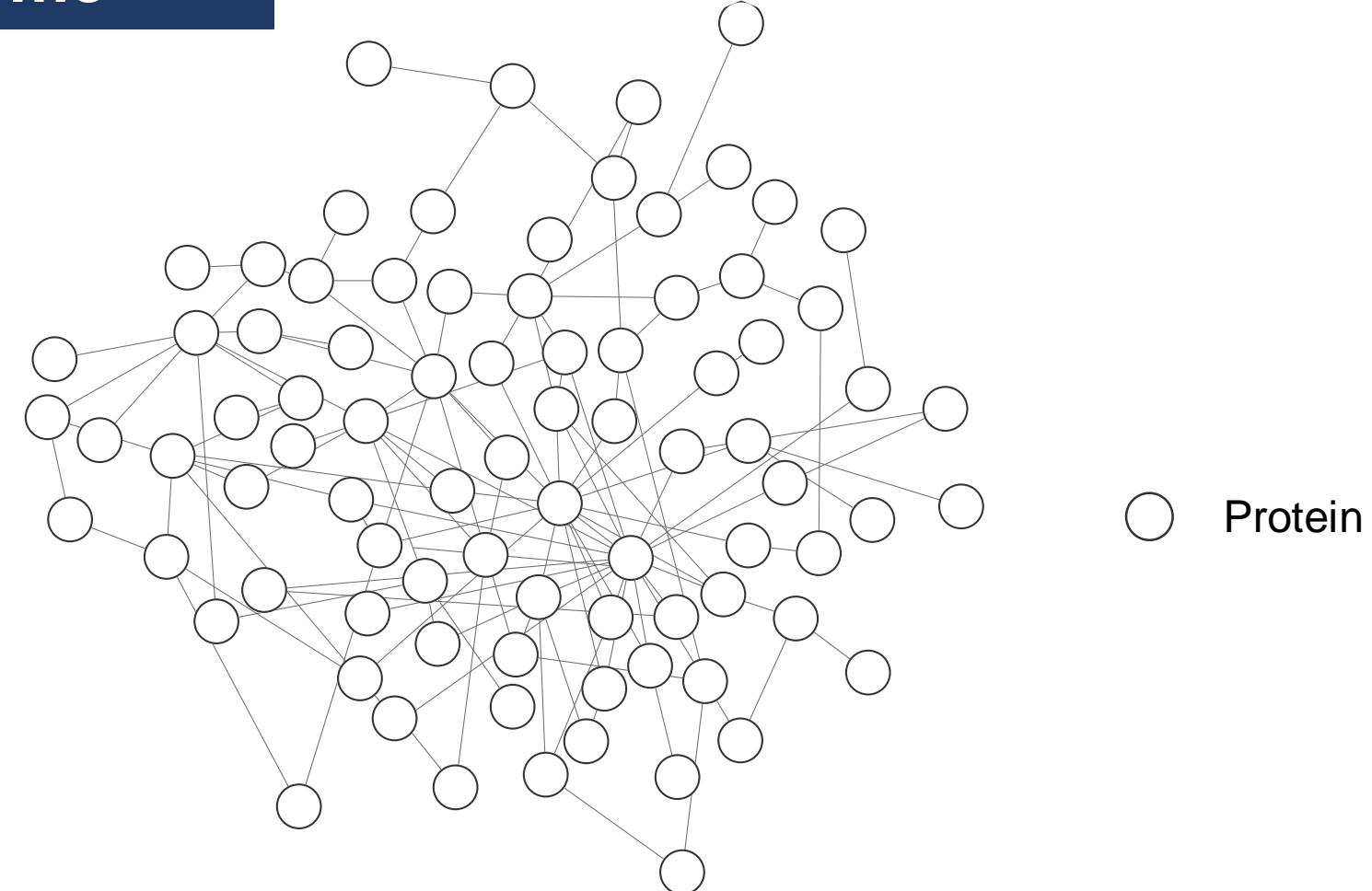
Some of the drug repurposing methodologies that we implement



Drug repurposing – Network medicine

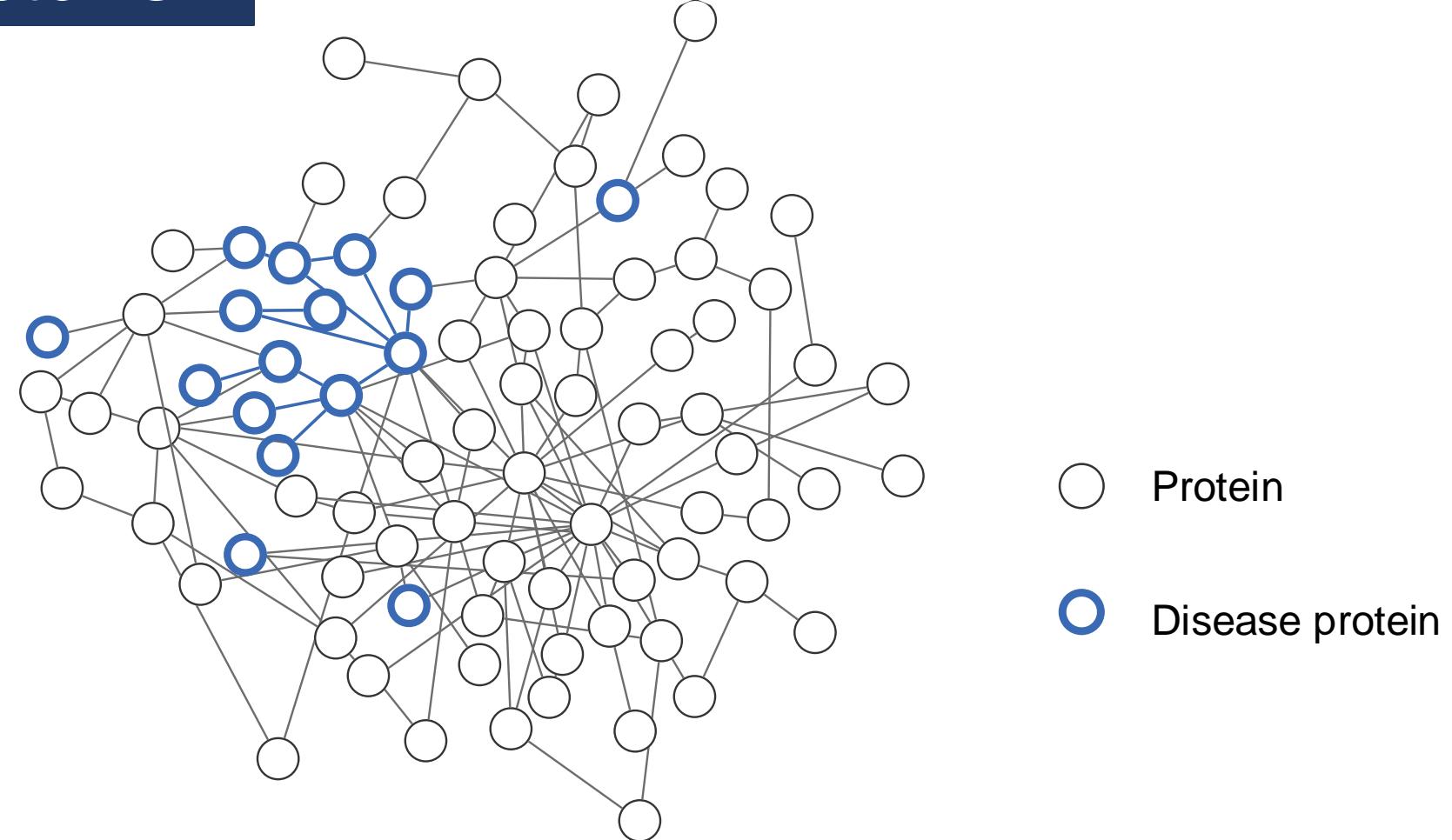
Interactome

Protein – Protein
Interaction (PPI) network



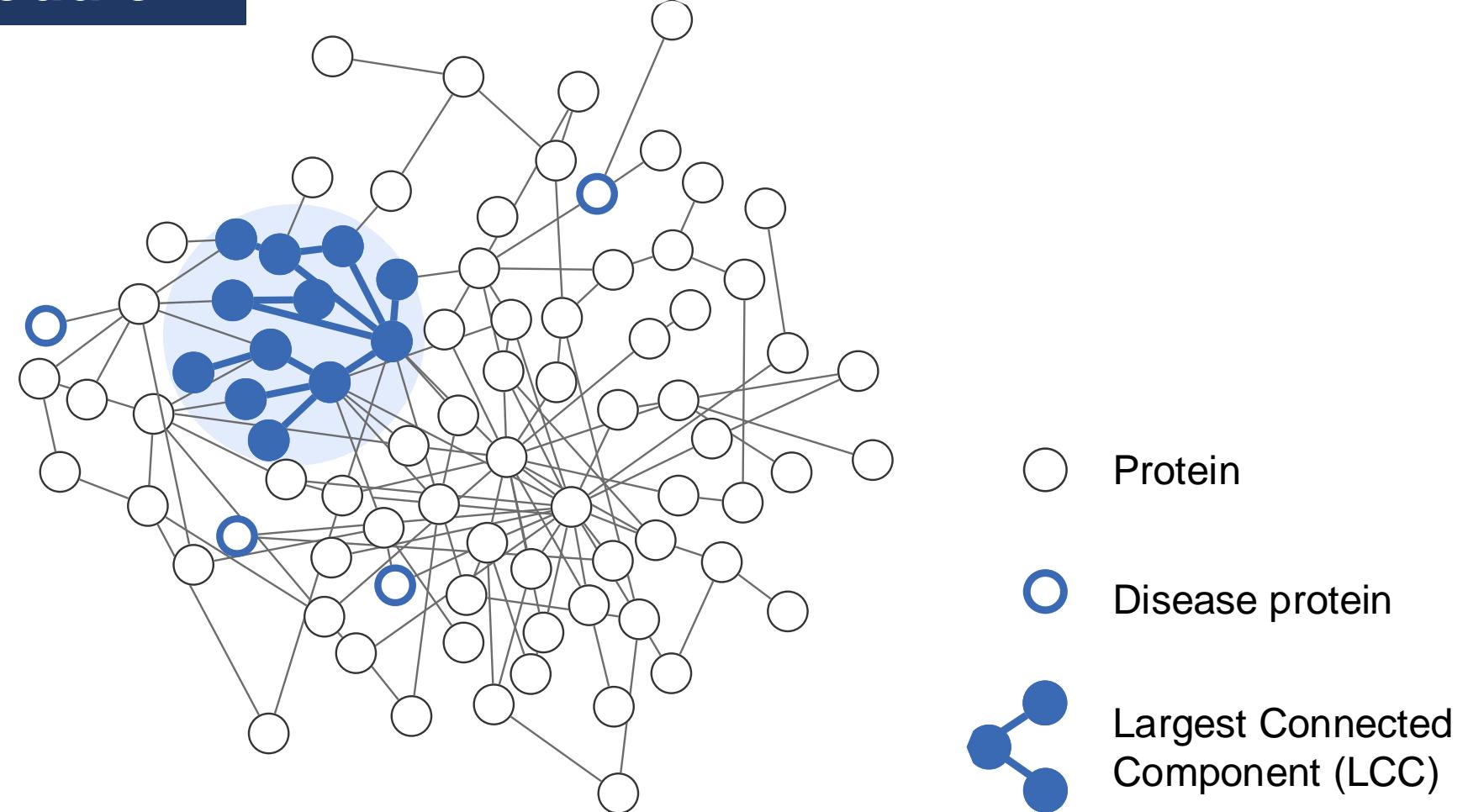
Drug repurposing – Network medicine

Disease proteins



Drug repurposing – Network medicine

Disease module



Drug repurposing – Network medicine

Statistical validation of the disease module

- We compare the actual LCC size (S) with the LCC size of 1,000 random sets used as controls
 - To obtain them, we preserve the degree distribution of the original disease proteins
 - We obtain a z-score (significant values > 1.65)

$$z(S) = \frac{S - \mu_S}{\sigma_S}$$

S : size of the disease module formed by the LCC

(Menche *et al.*, 2015)

Distance between disease and drugs

- Drugs already being used for treating a disease (we have focused on Schizophrenia)
→ “Treatment”
- Rest of the drugs (where to look for repurposing opportunities)
→ “Unknown”

Drug repurposing – Network medicine

Distance between disease and drugs

Closest distance

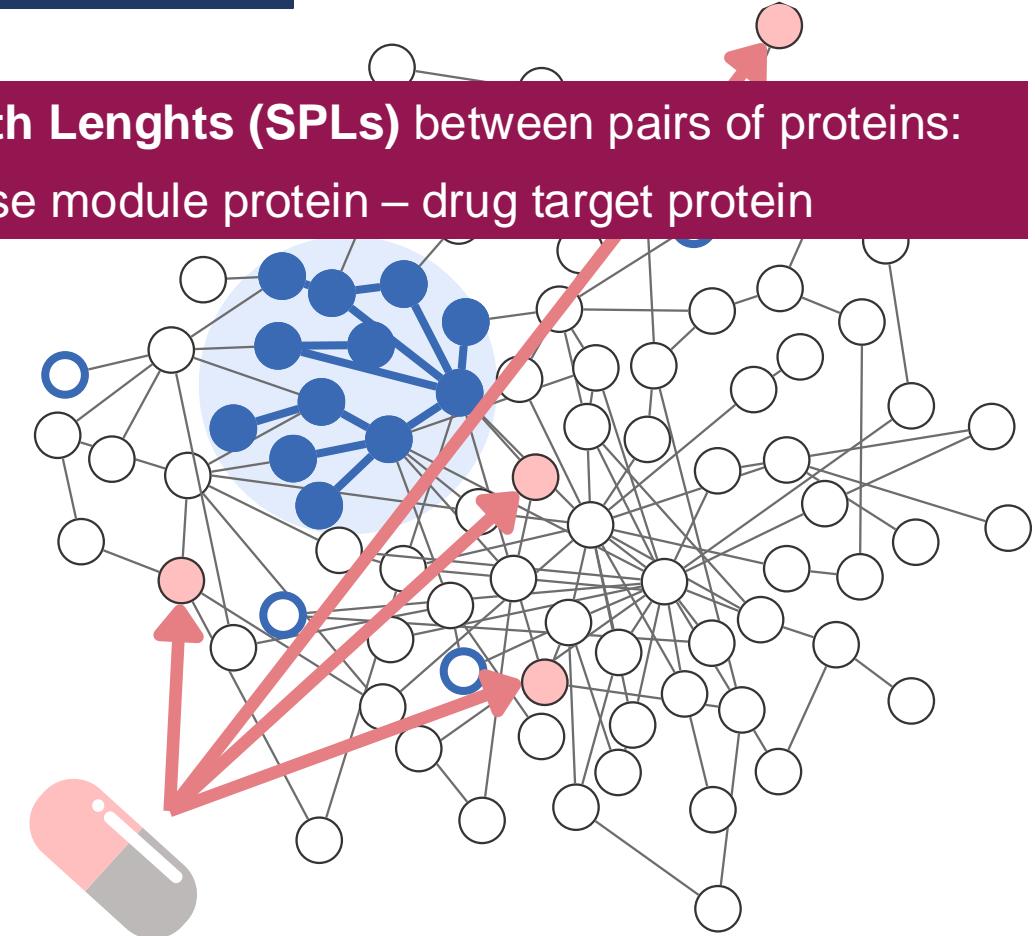
$$d_c(S, T) = \frac{1}{\|T\|} \times \sum_{t \in T} \min_{s \in S} d(s, t)$$

S: set of proteins in disease module

T: set of protein targets of the drugs

(Guney et al., 2016)

Shortest Path Lengths (SPLs) between pairs of proteins:
disease module protein – drug target protein



Drug repurposing – Network medicine

Distance between disease and drugs

Proximity z-score

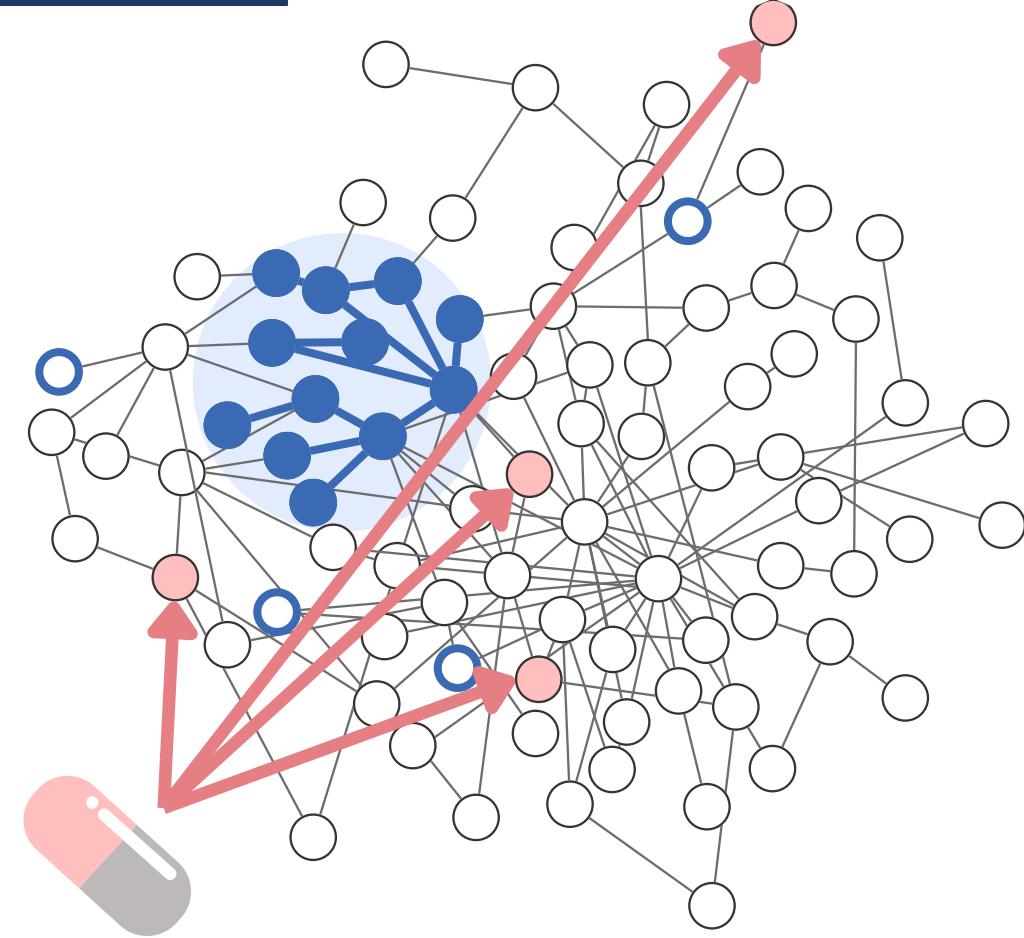
$$z(S, T) = \frac{d_c(S, T) - \mu_{d_c(S, T)}}{\sigma_{d_c(S, T)}}$$

S: set of proteins in disease module

T: set of protein targets of the drugs

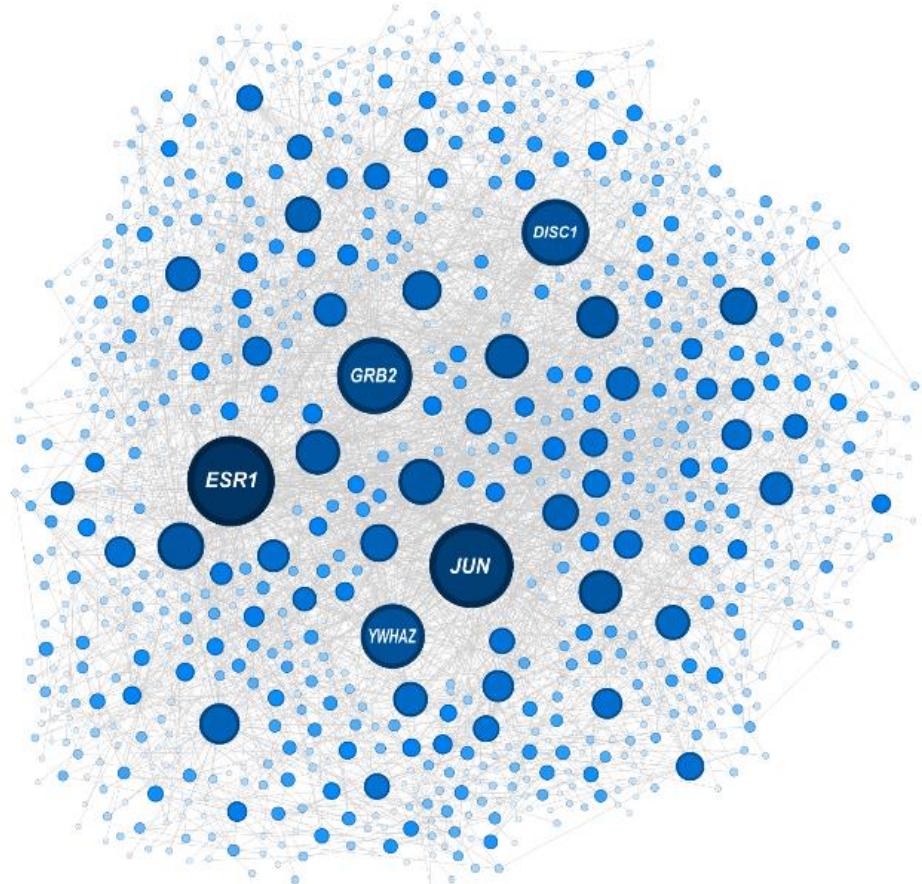
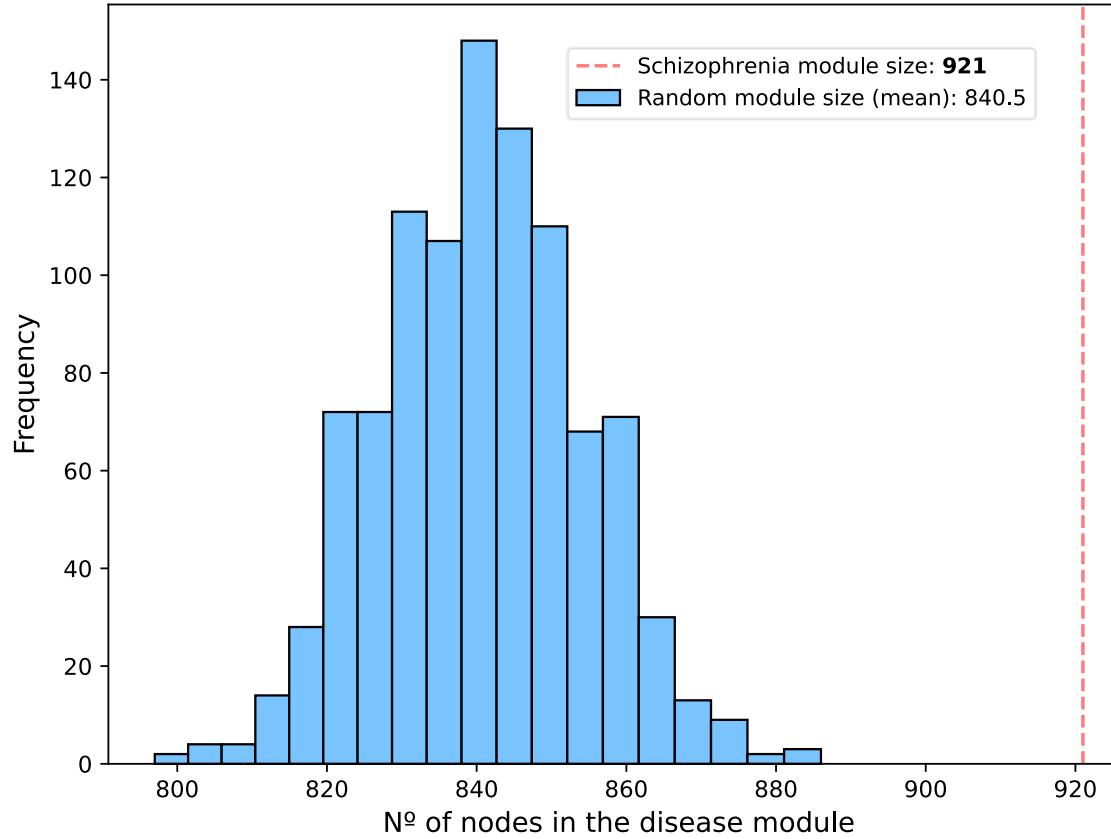
(Guney et al., 2016)

$z(S, T) < -0.15 \rightarrow$ proximal drugs



Drug repurposing – Network medicine

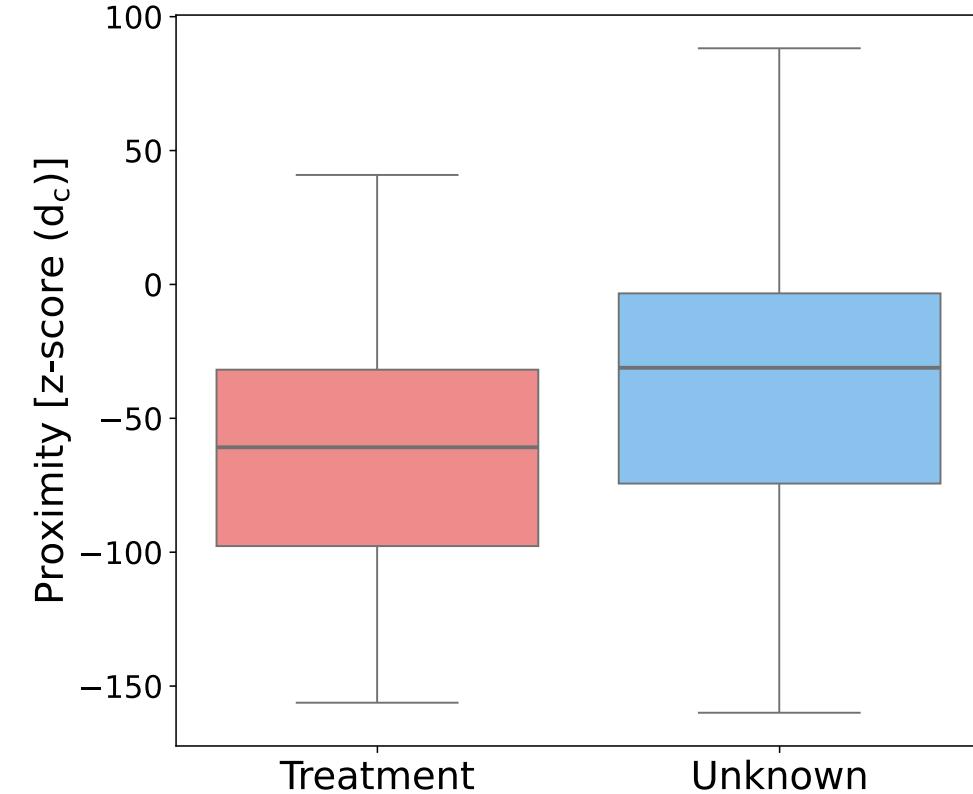
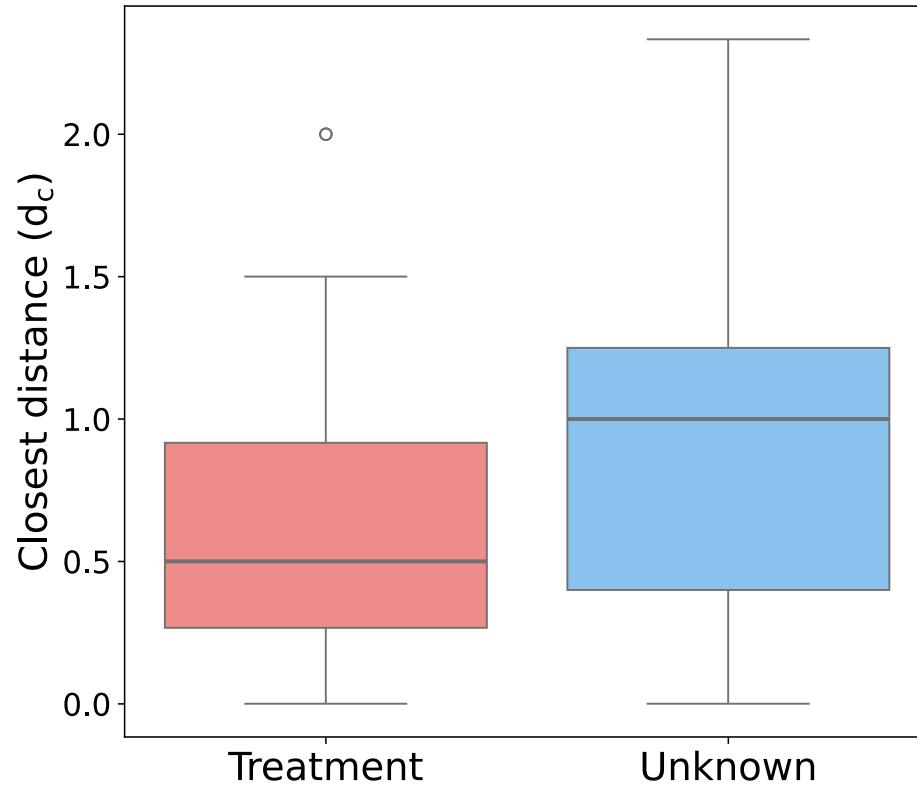
Schizophrenia module



(Marín Tercero *et al.*, 2024)

Drug repurposing – Network medicine

Distance and proximity



(Marín Tercero *et al.*, 2024)

Drug repurposing – Network medicine

Identification of drug repurposing candidates

Proximity

726 “unknown” proximal drugs

(proximity value ≤ -0.15)

Distance

113 proximal “unknown” drugs were filtered

(d_c between the first quartile and the median of the closest distance of “treatment” drugs of schizophrenia)

Targets as DGE

78 genes with significant differential expression

Targets in co-expression

69 genes presented a significant correlation with their co-expression module

14 final drugs

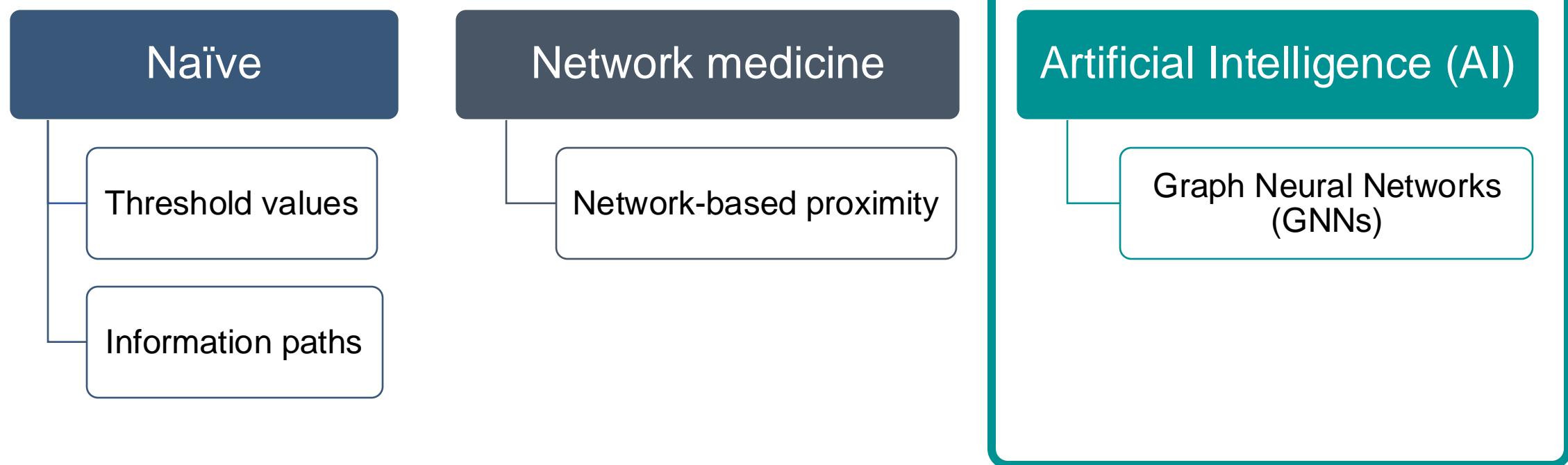
Drug repurposing – Network medicine

Identification of drug repurposing candidates

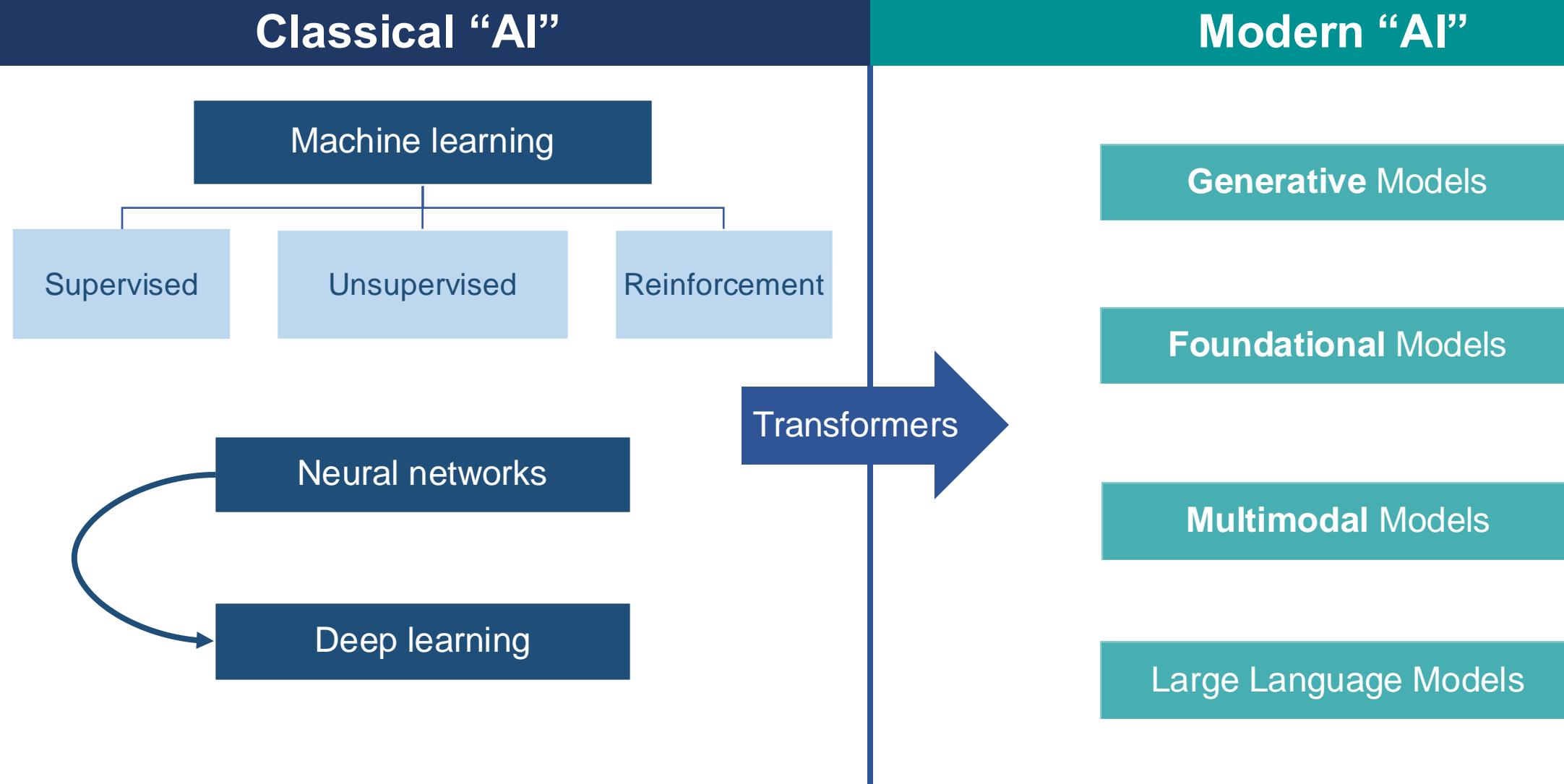
Drug Name	MeSH-PA	MOA	Proximity
Nefazodone	Antidepressive agents	Norepinephrine and serotonin uptake inhibitors	-115.58
Caffeine	Anti-Inflammatory agents, Antimutagenic agents	Enzyme inhibitors	-103.52
Imatinib	Antineoplastic agents	Protein kinase inhibitors	-94.88
Doxepin	Antidepressive agents	Norepinephrine and serotonin uptake inhibitors	-90.84
Terfenadine	Histamine H1 antagonists	Histamine H1 receptor antagonists	-87.51
Tamoxifen	Antineoplastic agents, Bone density conservation agents	Selective estrogen receptor modulators	-85.95
Glutamine	-	Biological macromolecular activity	-78.57
Flufenamic acid	Anti-Inflammatory agents	-	-77.57
Procainamide	Anti-Arrhythmia agents	Unknown cellular or molecular interaction	-75.76
Alfuzosin	Urological agents	Adrenergic alpha-antagonists	-74.37
Minaprine	Antidepressive agents	-	-65.01
Flecainide	Anti-Arrhythmia agents	Ion channel interactions	-51.72
Sirolimus	Anti-Bacterial agents, Antifungal agents	Protein kinase inhibitors	-22.77
Amsacrine	Antineoplastic agents	-	-8.87

Drug repurposing

Some of the drug repurposing methodologies that we implement



Drug repurposing – Artificial Intelligence



Classical AI in the task of repurposing

Machine Learning

Supervised Learning

- Training models on labeled datasets of known **drug-target pairs** to predict interactions for unexplored drug candidates.
- Using gene expression or proteomic profiles to classify diseases and suggest drugs targeting similar **molecular signatures**.
- Predicting the safety and **side effects** of repurposed drugs using patient data or chemical properties.

Combined approach with text mining from PubMed abstract and supervised models (gradient boosting, logistic regression, random forests and LASSO) for predicting psoriasis drug candidates (73).



Unsupervised Learning

- Grouping drugs** with similar chemical structures, side effects, or transcriptional responses to infer shared therapeutic potential.
- Identifying **subtypes of diseases** based on omics data or patient profiles, revealing new opportunities to match existing drugs to more specific conditions.
- Using clustering techniques to uncover **pathways or gene modules** in networks that link diseases to known drug targets.

UMAP and K-Means to cluster diseases based on gene expression, and predict 22 drug candidates for body myositis, polymyositis, and dermatomyositis (94).



Reinforcement Learning

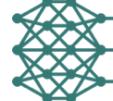
- Simulating and learning optimal **drug combinations** by rewarding strategies that maximize therapeutic effects while minimizing toxicity.
- Developing **personalized treatment schedules** where drug actions are evaluated and adjusted iteratively, based on patient responses.



Integration of transformer models to learn rich contextual embeddings of chemical compounds and biological targets with reinforcement learning algorithms to optimize repurposing predictions for neglected diseases.

Deep Learning

- Gene expression modeling** by using deep neural networks to analyze transcriptional responses to drugs, identifying candidates that can reverse disease-associated expression profiles.
- Applying convolutional neural networks to high-content imaging data to detect cell phenotypes indicative of a drug's efficacy for new diseases in the context of a **phenotypic screening**.
- Using deep learning architectures to **combine omics and clinical data** in order to predict the most promising drug candidates for specific diseases.



BiFusion, a convolutional network model, integrates heterogeneous information to predict novel candidate drugs for Parkinson's disease and breast carcinoma (111).

Modern AI in the task of repurposing

Large Language Models

- Biomedical knowledge extraction and hypothesis generation**
- Extracting repurposing insights by analyzing **vast amounts** of biomedical literature, patents, and clinical trial **data**.
- Generating new hypotheses by identifying **overlooked semantic relationships** between drugs, pathways, and diseases.
- Supporting clinical translation by summarizing drug efficacy, side effects, and trial outcomes for **informed decision-making**.



DrugChat combines a ChatGPT-like large language model with a graph neural network and an adapter to process molecular structures, in order to engage users in multi-turn, interactive discussions about compound inputs, and aiding in the process of finding connections to diseases (119).

Generative Models

- Hypothesis generation and molecule optimization**
- Designing **new drug candidates** by generating molecular structures optimized for binding to disease-related targets.
- Optimizing **existing drugs** by modifying their chemical properties to improve efficacy, reduce toxicity, or enhance bioavailability.
- Accelerating **virtual screening** by rapidly generating and testing large libraries of potential drug compounds *in silico*.



ChatGPT to identify repurposing candidates for Alzheimer's disease, predicting promising drugs through iterative queries, and validating the top suggestions using demographic characteristics and the EHR (116).

Multimodal Models

- Cross-domain data fusion and personalized medicine**
- Combining **molecular, clinical, and imaging data** to provide a holistic view of drug effects across different biological systems.
- Predicting drug responses by integrating **omics data, patient records, and phenotypic screening** results to refine repurposing candidates.
- Stratifying patient populations to identify subgroups that may benefit from specific repurposed drugs based on **multi-source data fusion**.



Integration of chemical structure data and gene expression perturbations to successfully identify clusters of small molecules and dysregulated genes that predict potential COVID-19 treatments with minimal side effects.

Foundational Models

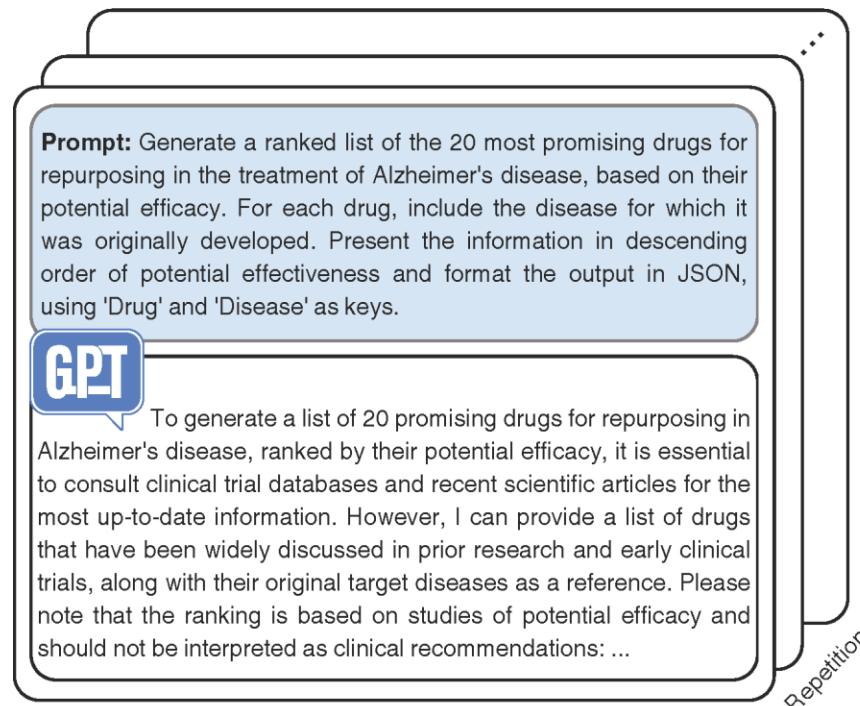
- Generalized knowledge and data integration**
- Learning **transferable drug-disease relationships** from large-scale biomedical data, enabling repurposing predictions even for rare diseases.
- Integrating **multimodal biomedical data** (e.g., genomics, proteomics, clinical records) to uncover hidden links between drugs and diseases.
- Enhancing prediction accuracy by leveraging **pre-trained embeddings** of drugs, diseases, and biological entities for better similarity assessments.



TxGNN, a graph foundation model based on graph neural networks, predicts indications and contraindications across challenging disease areas and neglected diseases (113).

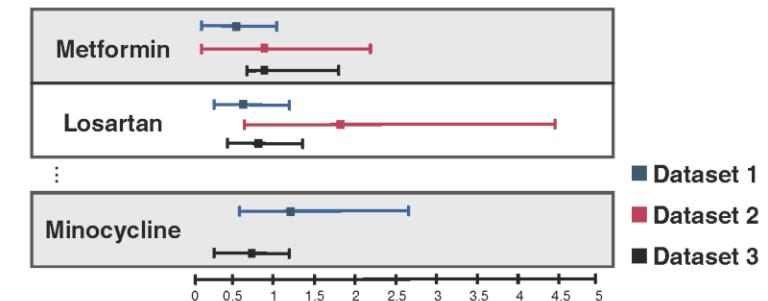
Drug repurposing – Artificial Intelligence

Conversational AI



```
[{"Drug": "Metformin", "Disease": "Diabetes Type 2"}, {"Drug": "Losartan", "Disease": "Hypertension"}, {"Drug": "Minocycline", "Disease": "Bacterial infections"}, {"Drug": "Simvastatin", "Disease": "Dyslipidemia"}, {"Drug": "Pioglitazone", "Disease": "Diabetes Type 2"}, {"Drug": "Rapamycin", "Disease": "Prevention of organ transplant rejection"}, ...]
```

Hazard ratio of associations between ChatGPT-suggested drug repurposing candidates and Alzheimer's disease risk in different clinical datasets



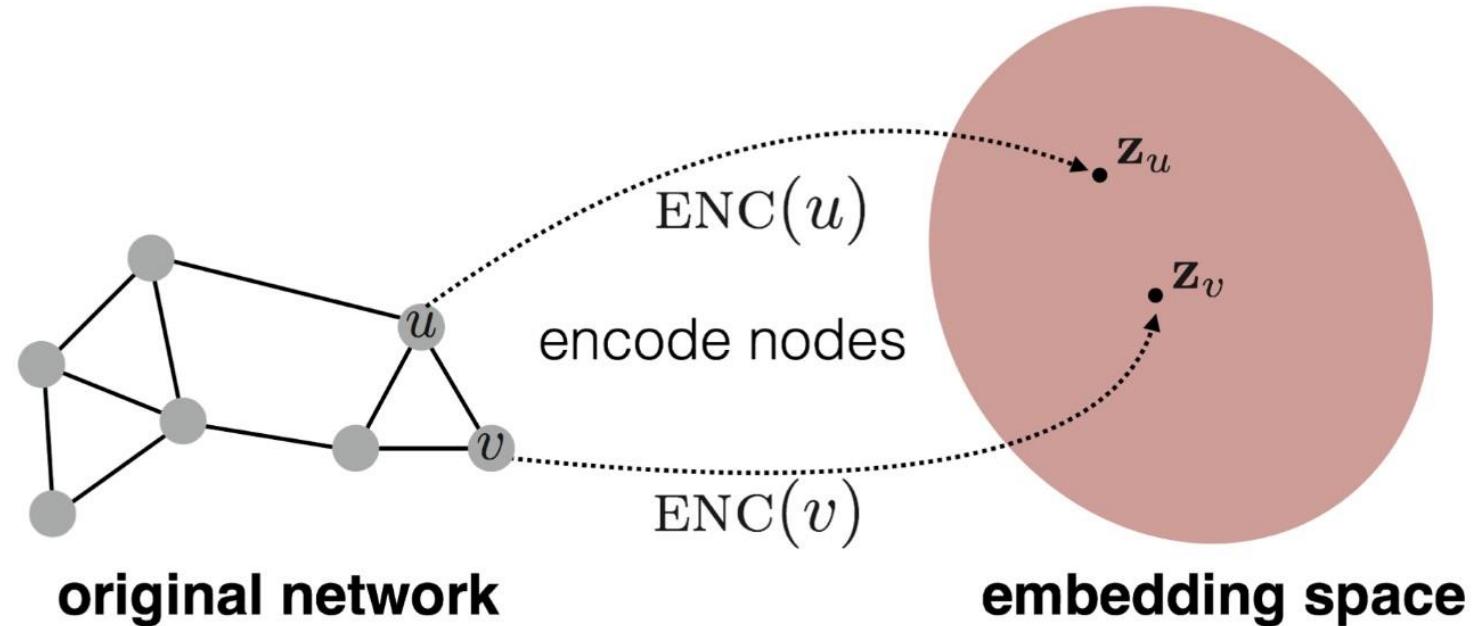
(Credits of this figure to
Andrea Álvarez Pérez)

(Adapted from Yan et al., 2024)

Drug repurposing – Artificial Intelligence

Graph Neural Networks (GNNs)

Neural Networks on graphs, non-Euclidean data

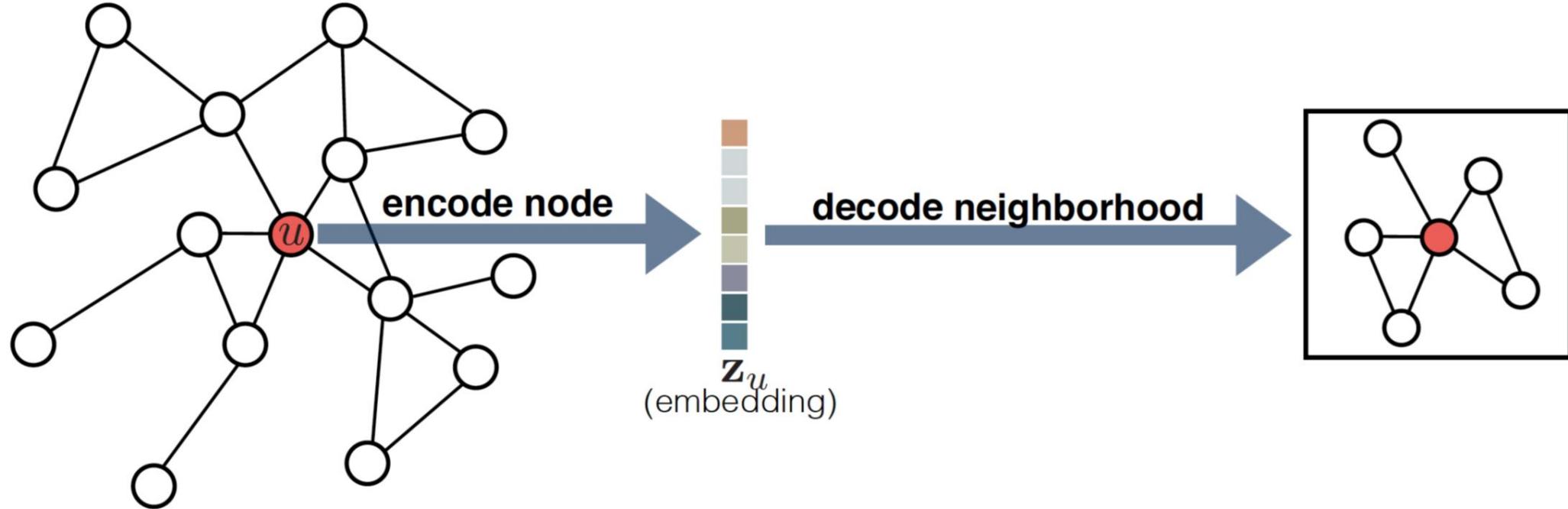


(Hamilton, 2020)

Drug repurposing – Artificial Intelligence

Graph Neural Networks (GNNs)

Neural Networks on graphs, non-Euclidean data

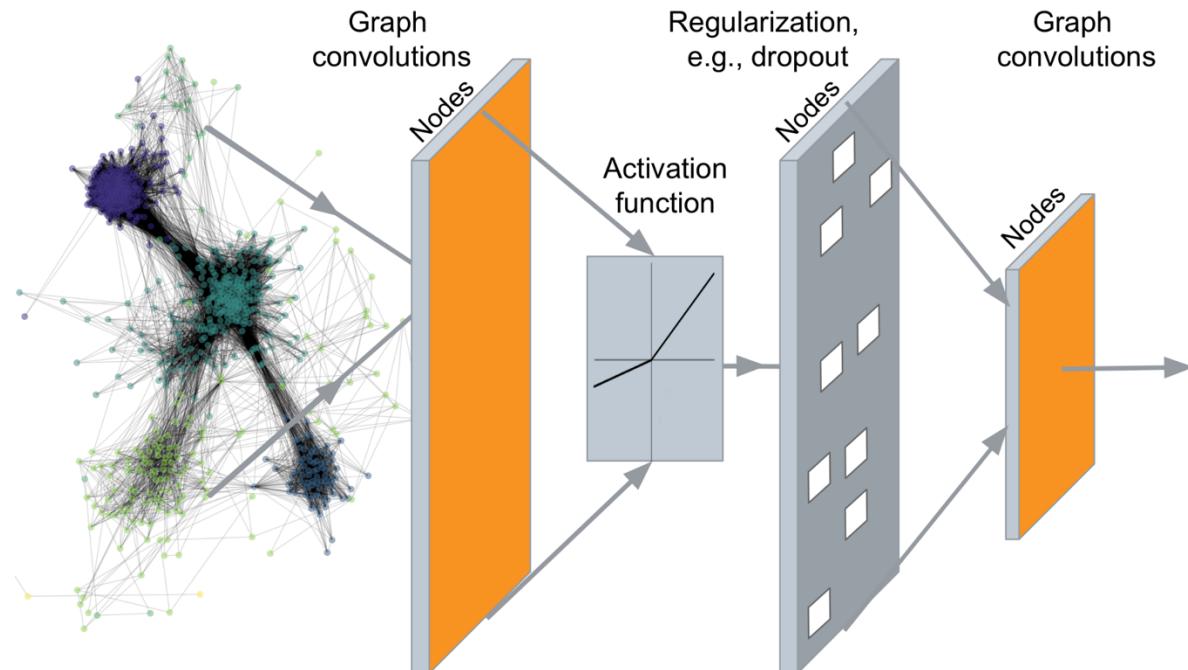


(Hamilton, 2020)

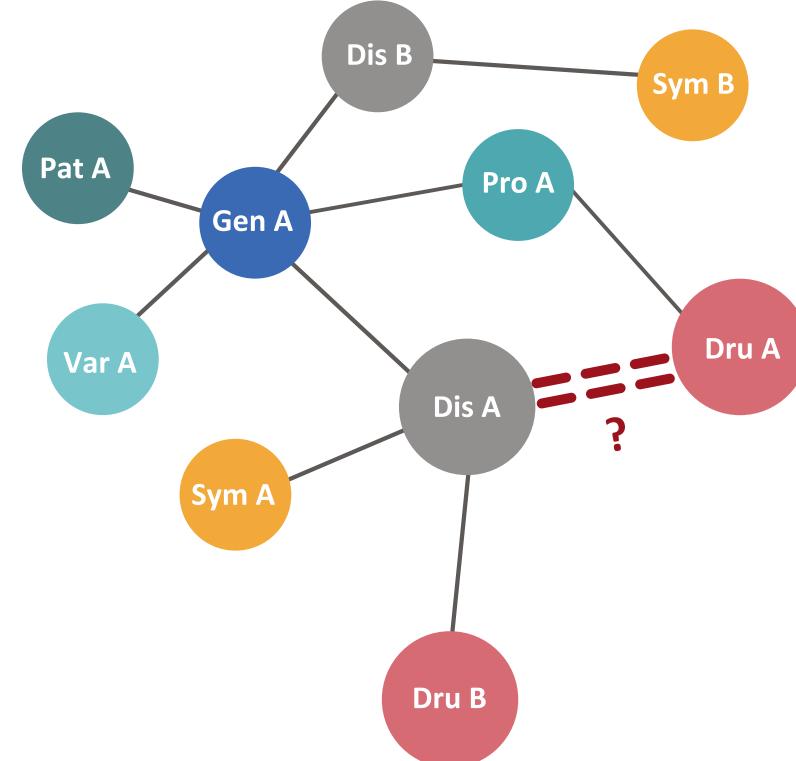
Drug repurposing – Artificial Intelligence

Graph Neural Networks (GNNs)

Neural Networks on graphs, non-Euclidean data

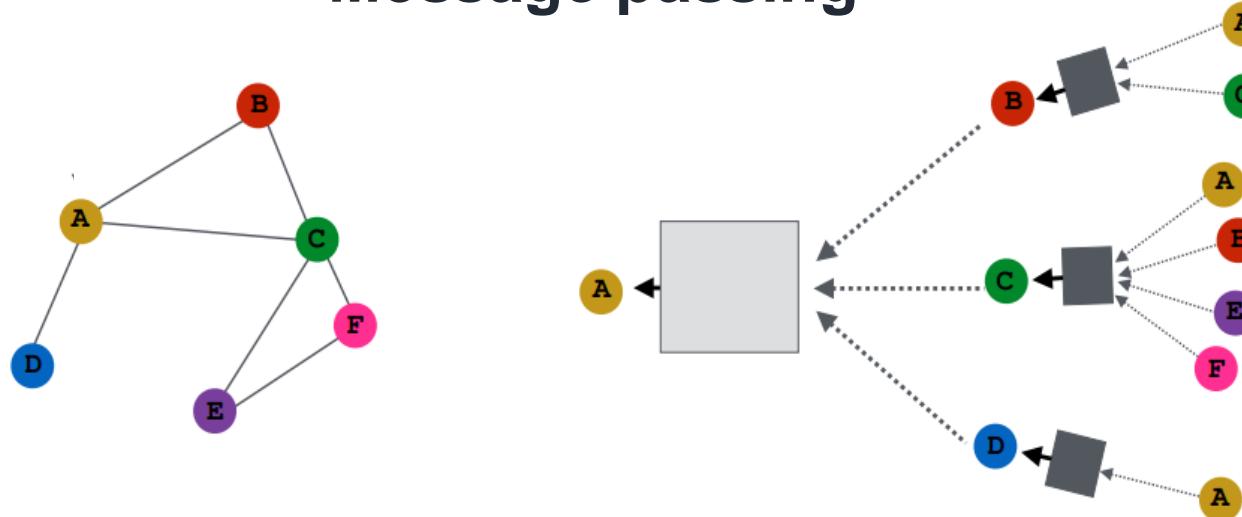


(Leskovec, 2021)



Graph Neural Networks (GNNs)

Message passing



(Leskovec, 2021)

$$h_u^{k+1} = \text{UPDATE}^{k+1} \left(h_u^k, \text{AGGREGATE}^{k+1} (\{h_v^k, \forall v \in N(u)\}) \right)$$

$$h_u^{k+1} = \text{UPDATE}^{k+1} (h_u^k, m_{N(u)}^k)$$

Drug repurposing – Artificial Intelligence

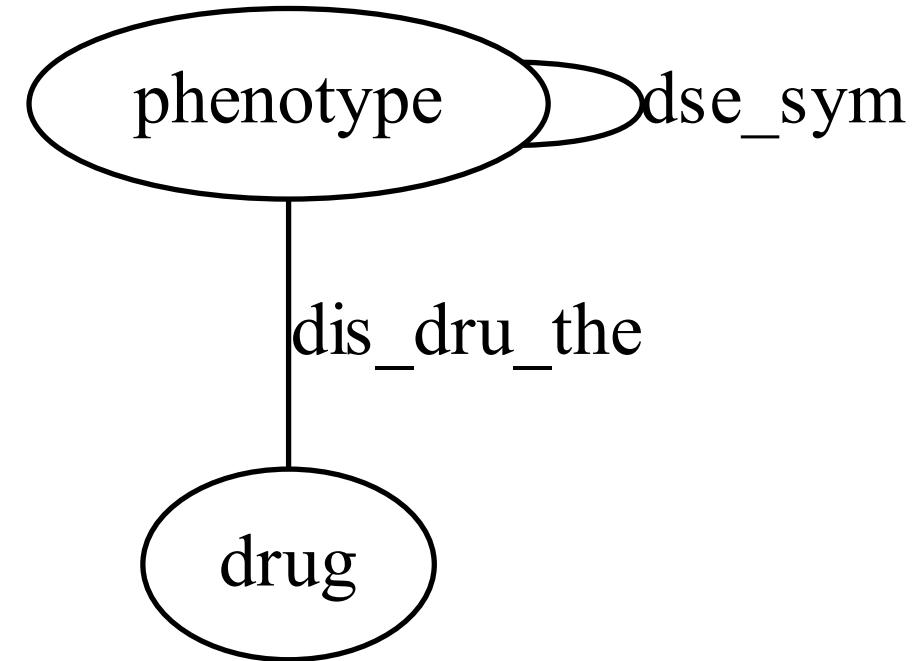
Our GNN-models

- **REDIRECTION:** dRug rEpurposing Disnet lInk pREdiCTION (Ayuso-Muñoz *et al.*, 2022)
- **BEHOR:** Bidirectional Edge and Hyperparameter Optimized REDIRECTION (Ayuso-Muñoz *et al.*, 2023a)
- **DMSR:** Drug Molecular Structure REDIRECTION (Ayuso-Muñoz *et al.*, 2023b)
- **DRAGON:** Drug Repurposing via Graph Neural Networks with Drug and Protein Embeddings as Features (Artiñano-Muñoz *et al.*, 2024)

Drug repurposing – Artificial Intelligence

Our GNN-models: REDIRECTION

- Link prediction proof of concept of disease – drug link type
- Unweighted links
- No hyperparameter optimization
- Simple graph
 - Nodes: 34,673
 - Phenotype: 30,729
 - Drug: 3,944
 - Links: 366,151
 - Disease-Symptom: 313,972
 - Disease-Drug: 52,179

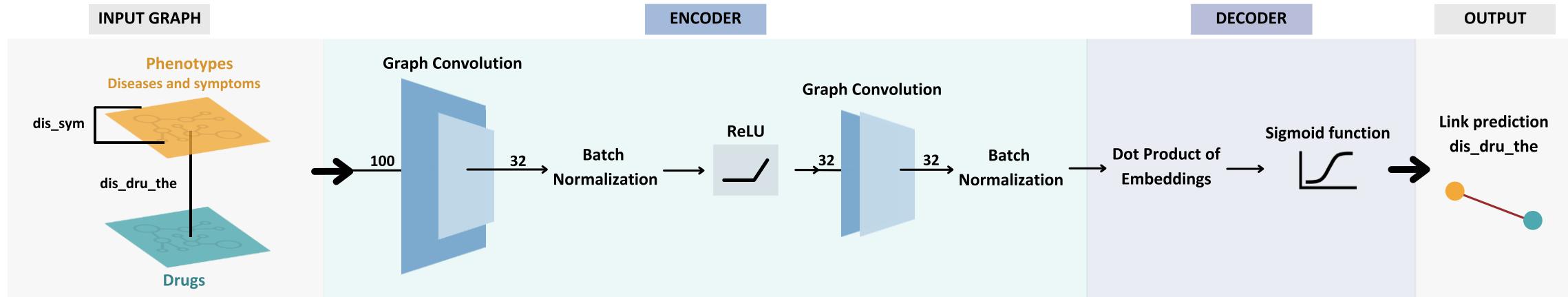


(Ayuso-Muñoz *et al.*, 2022)

Drug repurposing – Artificial Intelligence

Our GNN-models: REDIRECTION

- Link prediction proof of concept of disease – drug link type
- Unweighted links
- No hyperparameter optimization
- Simple graph

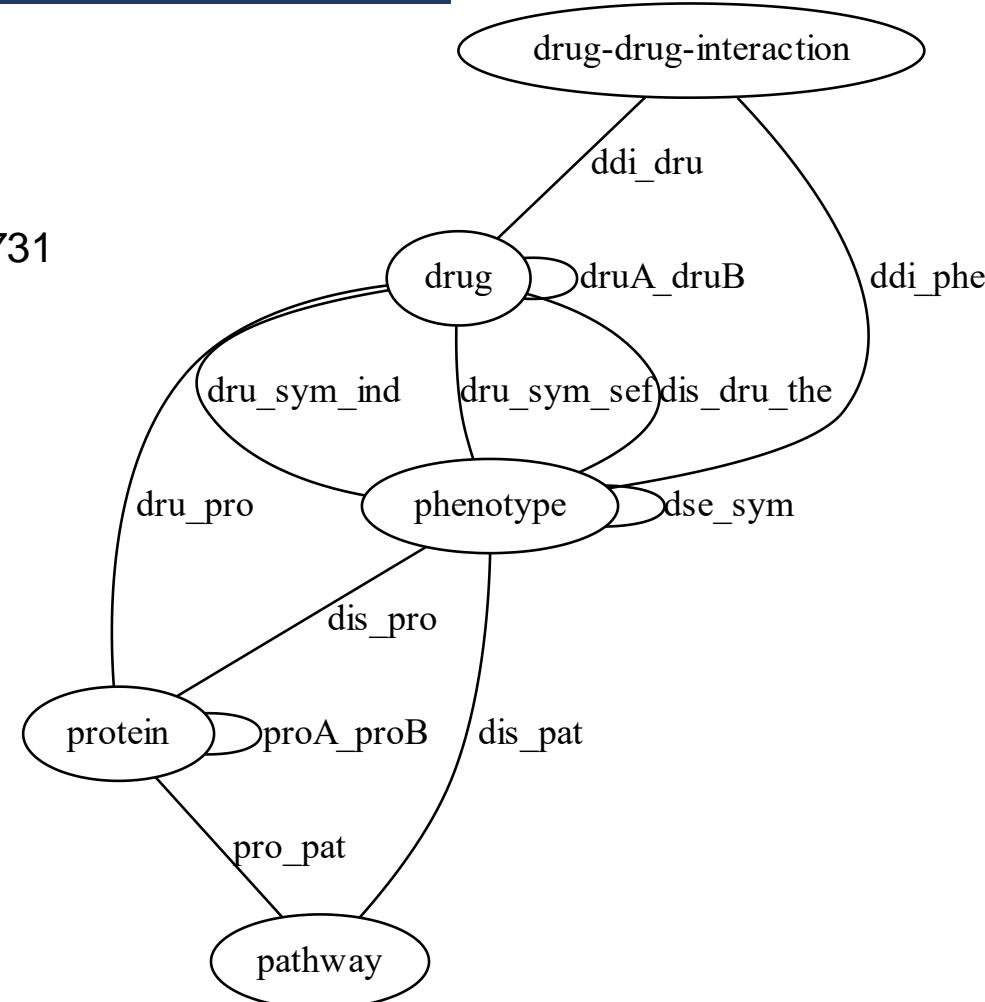


(Ayuso-Muñoz *et al.*, 2022)

Drug repurposing – Artificial Intelligence

Our GNN-models: BEHOR

- Upgraded version of REDIRECTION
- Complex graph
- Bidirectional links
- Hyperparameter optimization
- Baseline comparison
 - DeepWalk
 - Node2Vec
 - NetMF
 - Role2Vec
- Nodes: 153,747
 - Phenotype: 30,731
 - Drug: 3,944
 - Pathway: 1,105
 - Protein: 18,521
 - DDI: 99,446

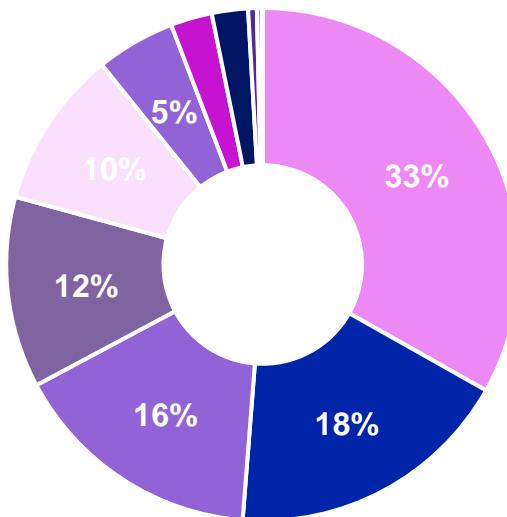


(Ayuso-Muñoz *et al.*, 2023)

Drug repurposing – Artificial Intelligence

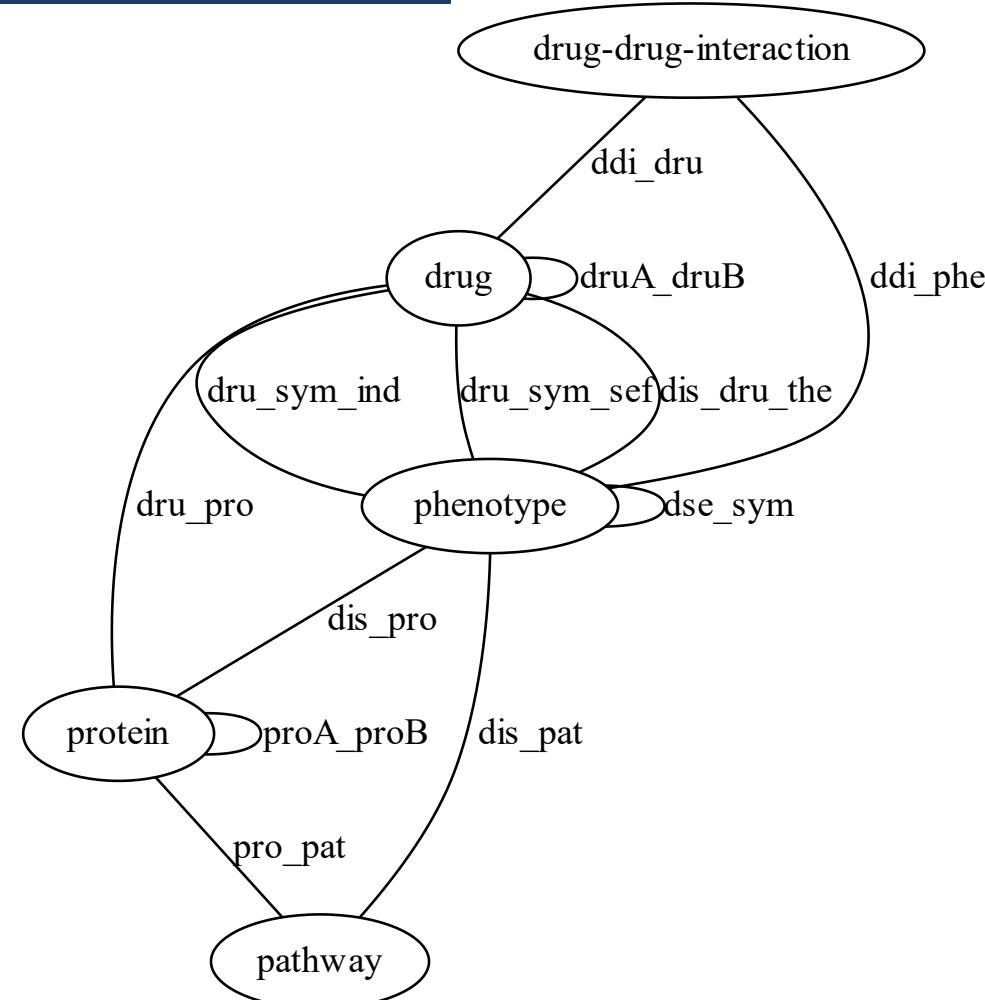
Our GNN-models: BEHOR

Number of links distribution



- Drug-Drug (33.17 %)
- Disease-Protein (18.08 %)
- Disease-Symptom (15.95 %)
- Protein-Protein (12.05 %)
- DDI-Drug (9.96 %)
- DDI-Phenotype (4.98 %)
- Disease-Drug (therapeutic) (2.61 %)
- Drug-Symptom (side effect) (2.28 %)
- Protein-Pathway (0.55 %)
- Drug-Protein (0.30 %)
- Drug-Symptom (indication) (0.04 %)
- Disease-Pathway (0.02 %)

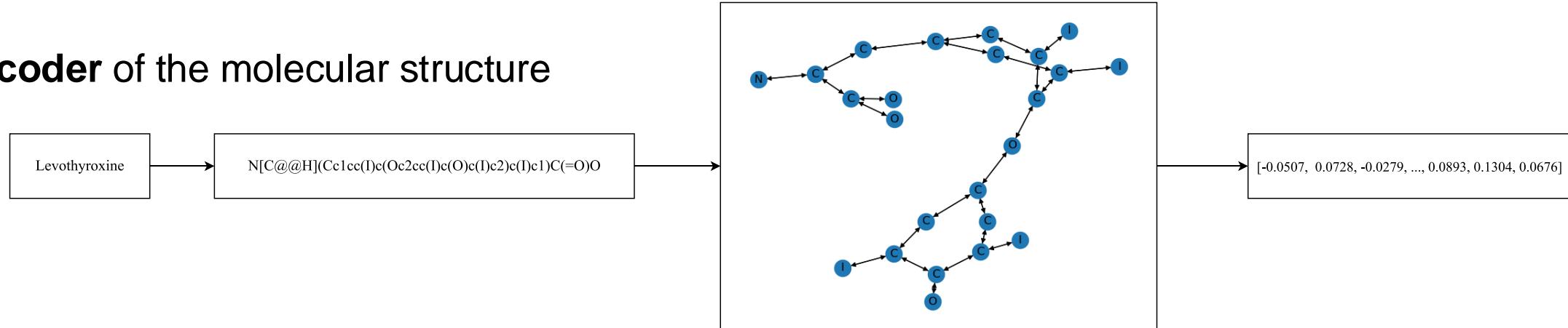
(Ayuso-Muñoz *et al.*, 2023)



Drug repurposing – Artificial Intelligence

Our GNN-models: DMSR

- Built from BEHOR
 - It incorporates drug molecular structures as initial vectors
 - Both with simple and complex graph
 - Hyperparameter tuning
-
- **Encoder of the molecular structure**

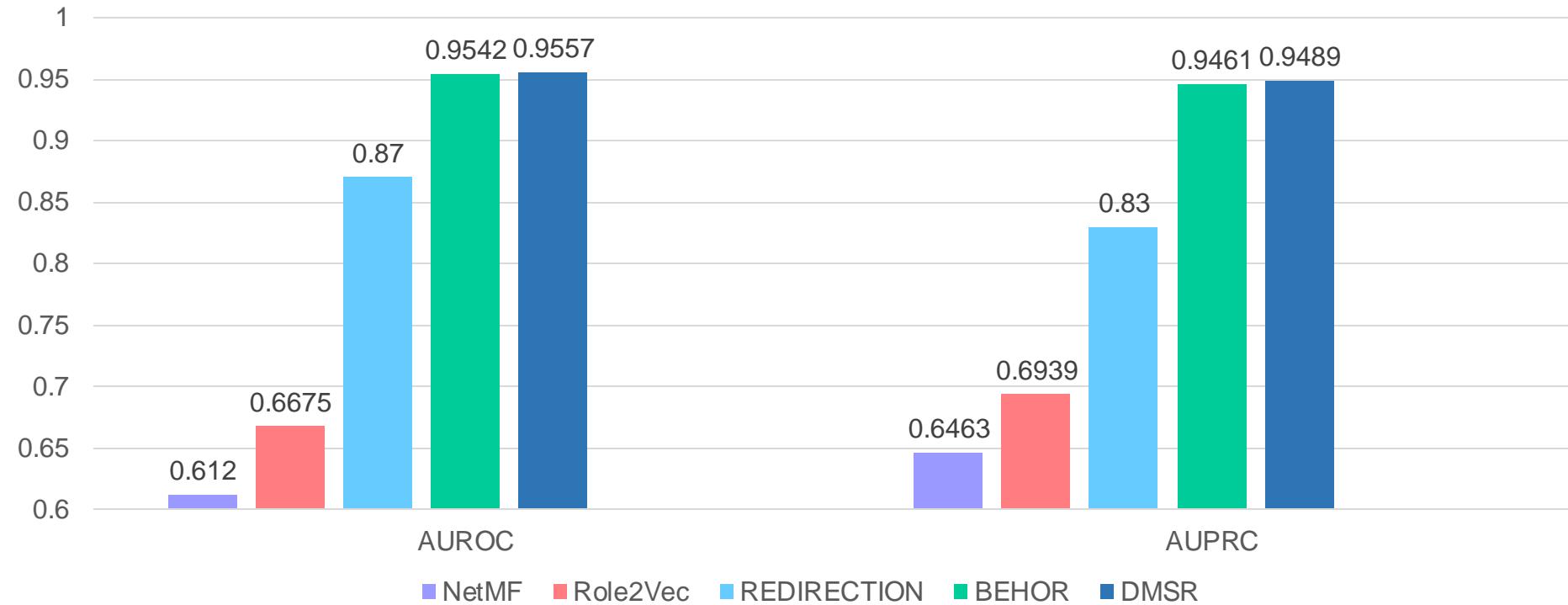


(Ayuso-Muñoz *et al.*, 2023)

Drug repurposing – Artificial Intelligence

Evaluation in simple graph

AUROC and AUPRC comparison in different models

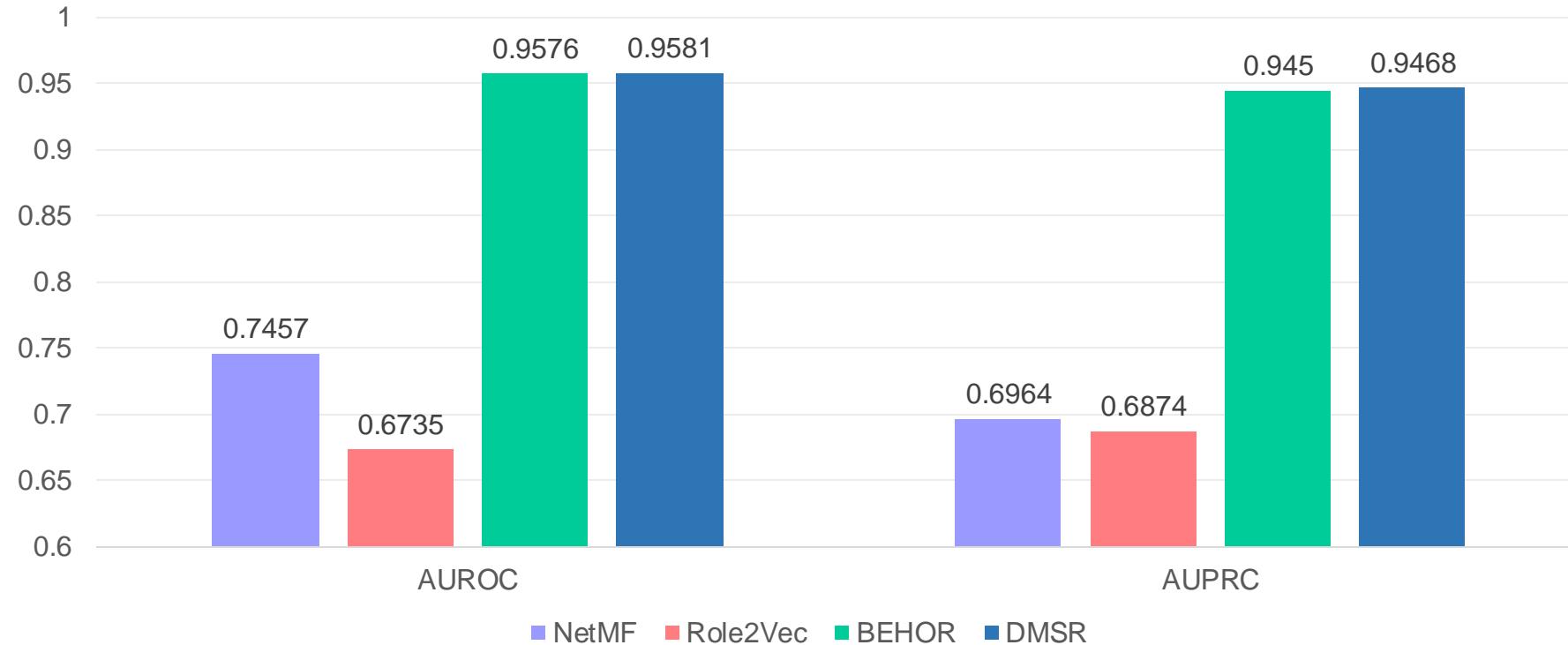


(Prieto-Santamaría *et al.*, 2023)

Drug repurposing – Artificial Intelligence

Evaluation in complex graph

AUROC and AUPRC comparison in different models

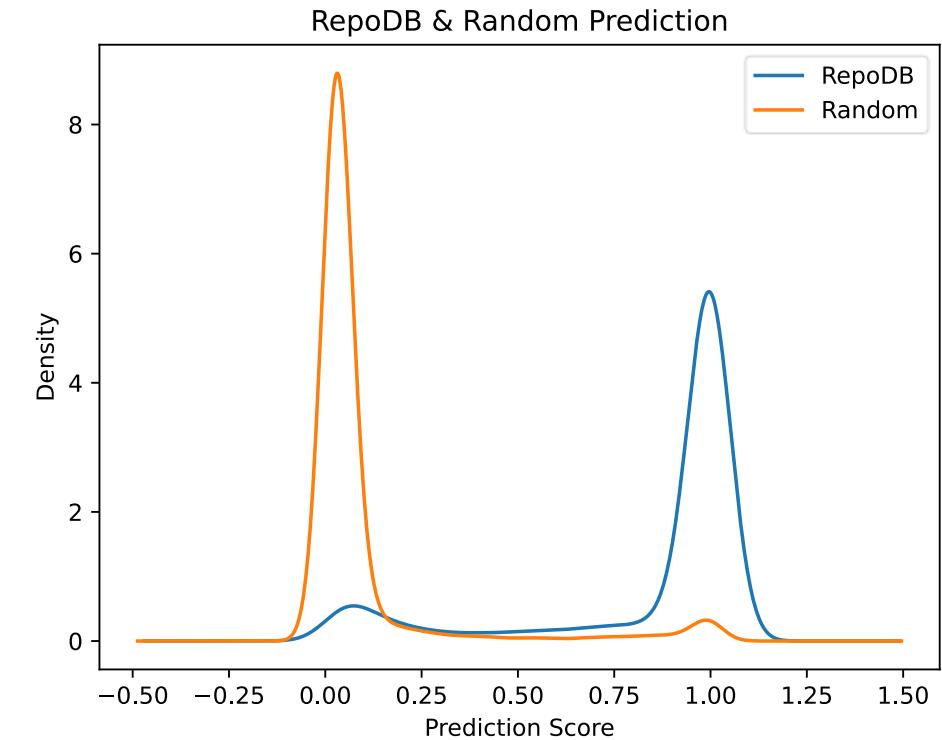
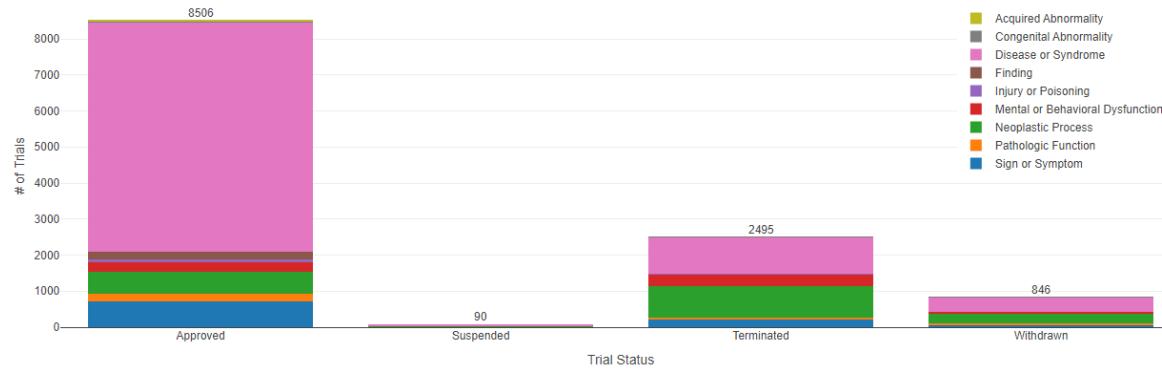


(Prieto-Santamaría *et al.*, 2023)

Drug repurposing – Artificial Intelligence

Testing DMSR with RepoDB

repoDB   
Drug Repositioning Database



(Ayuso-Muñoz *et al.*, 2023)

Drug repurposing – Artificial Intelligence

Examples of predictions

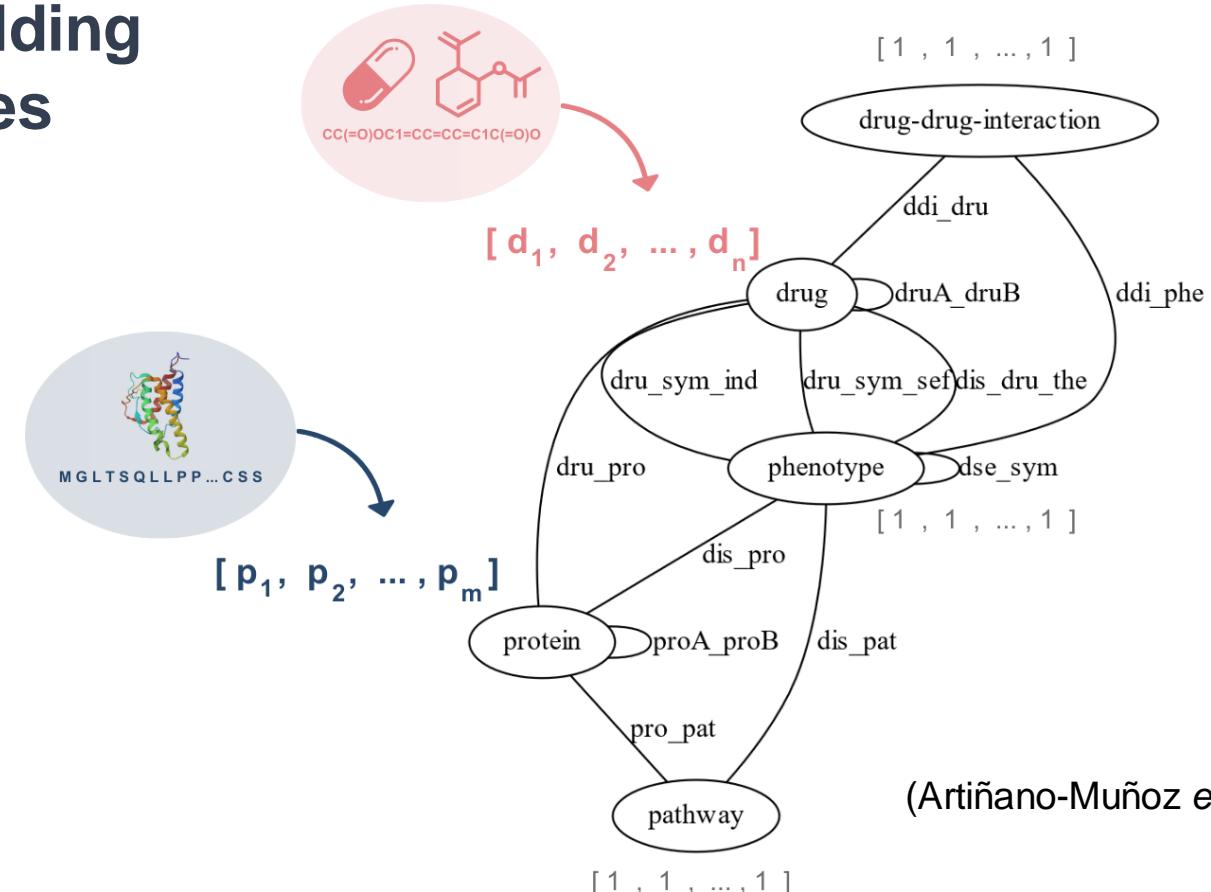
Disease UMLS CUI	Disease Name	Drug CHEMBL ID	Drug Name	Clinical Trials
C0018802	Congestive Heart Failure	CHEMBL1201668	Nesiritide	554
		CHEMBL1201866	Liraglutide	23
C0003873	Rheumatoid Arthritis	CHEMBL960	Leflunomide	78
		CHEMBL1789941	Ruxolitinib	4
C0149721	Left Ventricular Hypertrophy	CHEMBL3137301	Sacubitril	8

(Ayuso-Muñoz *et al.*, 2023)

Drug repurposing – Artificial Intelligence

Our GNN-models: DRAGON

**Drug and protein embedding
as initial node features**

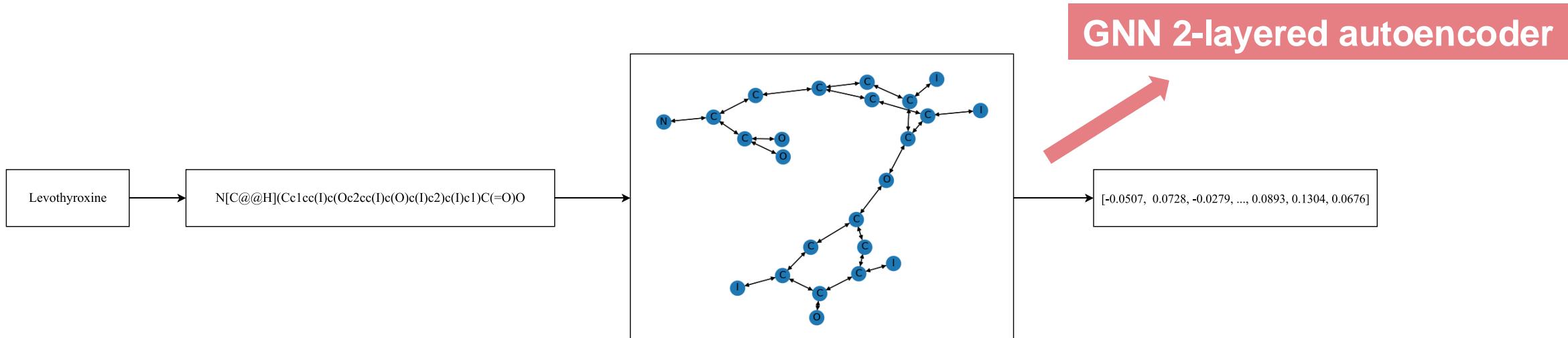


(Artiñano-Muñoz *et al.*, 2024)

Drug repurposing – Artificial Intelligence

Drug
embeddings

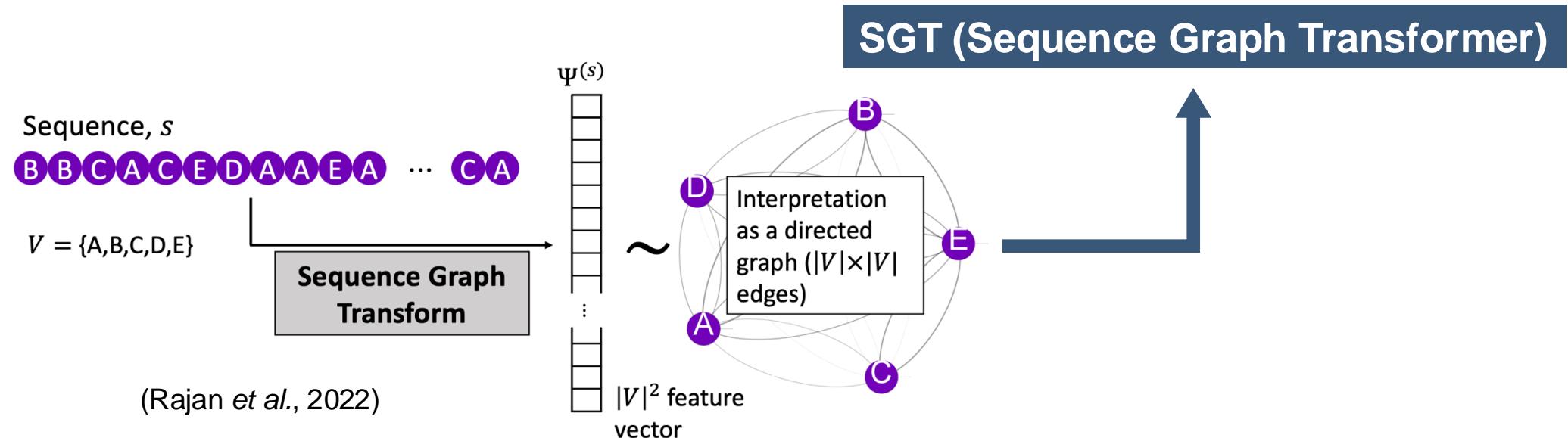
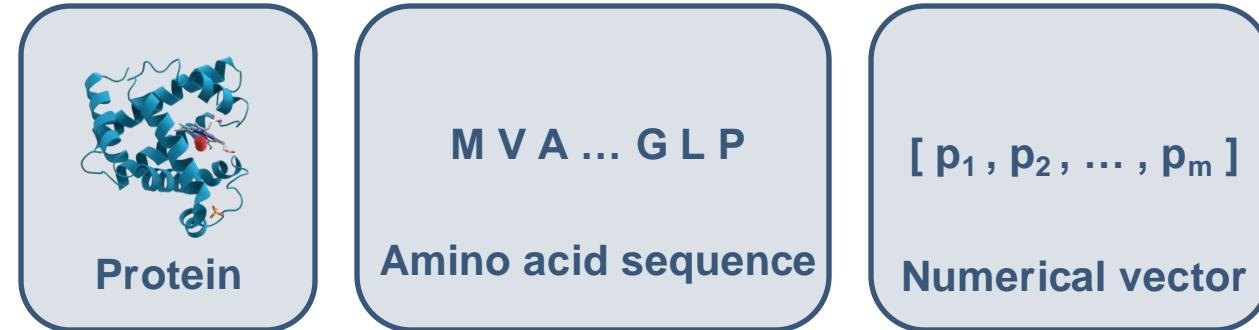
Our GNN-models: DRAGON



Drug repurposing – Artificial Intelligence

Protein
embeddings

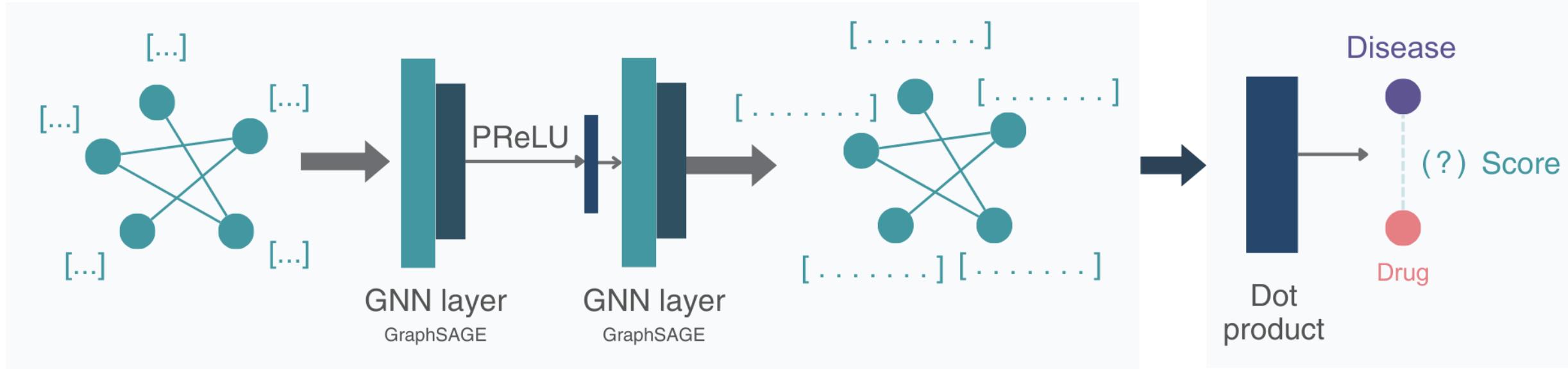
Our GNN-models: DRAGON



Drug repurposing – Artificial Intelligence

Our GNN-models: DRAGON

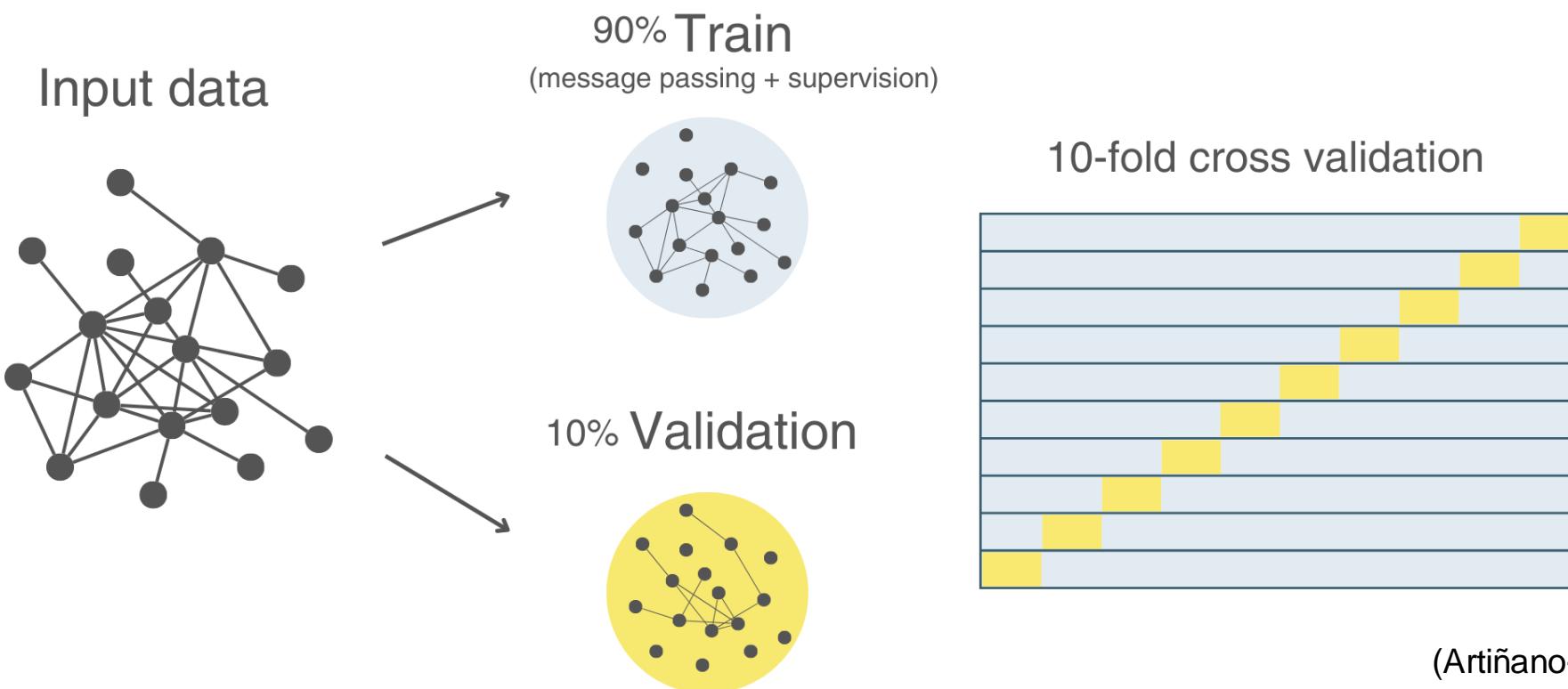
Architecture



(Artiñano-Muñoz *et al.*, 2024)

Our GNN-models: DRAGON

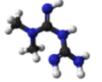
Model building



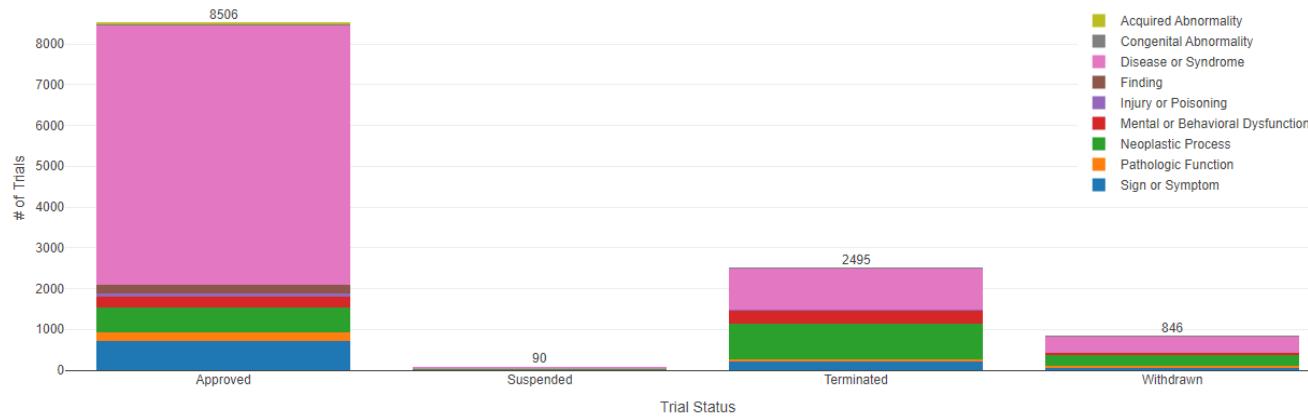
Drug repurposing – Artificial Intelligence

Our GNN-models: DRAGON

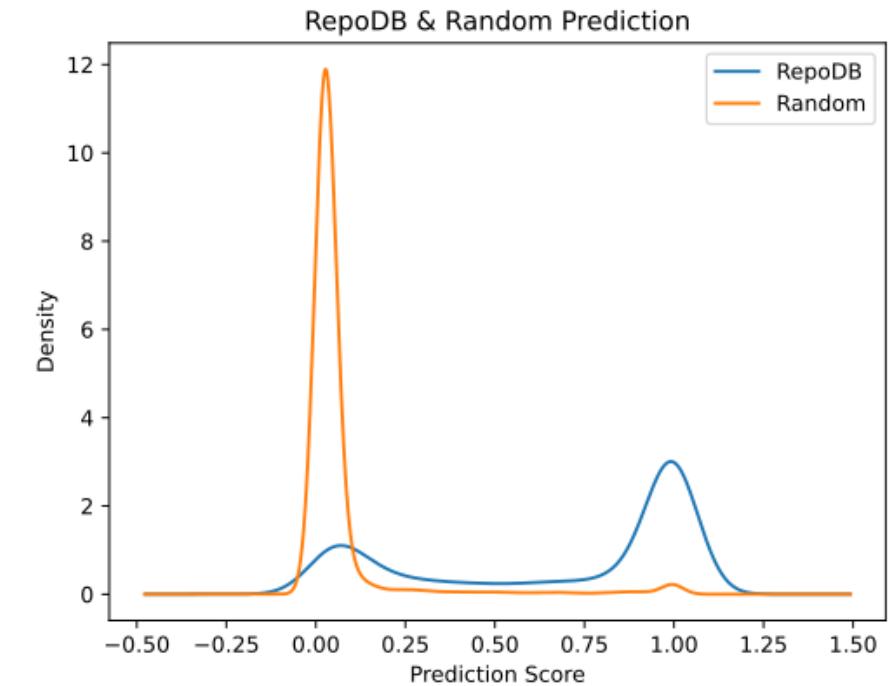
Evaluation

repoDB   

Drug Repositioning Database



(RepoDB, <https://unmtid-shinyapps.net/shiny/repodb/>)



(Artiñano-Muñoz *et al.*, 2024)

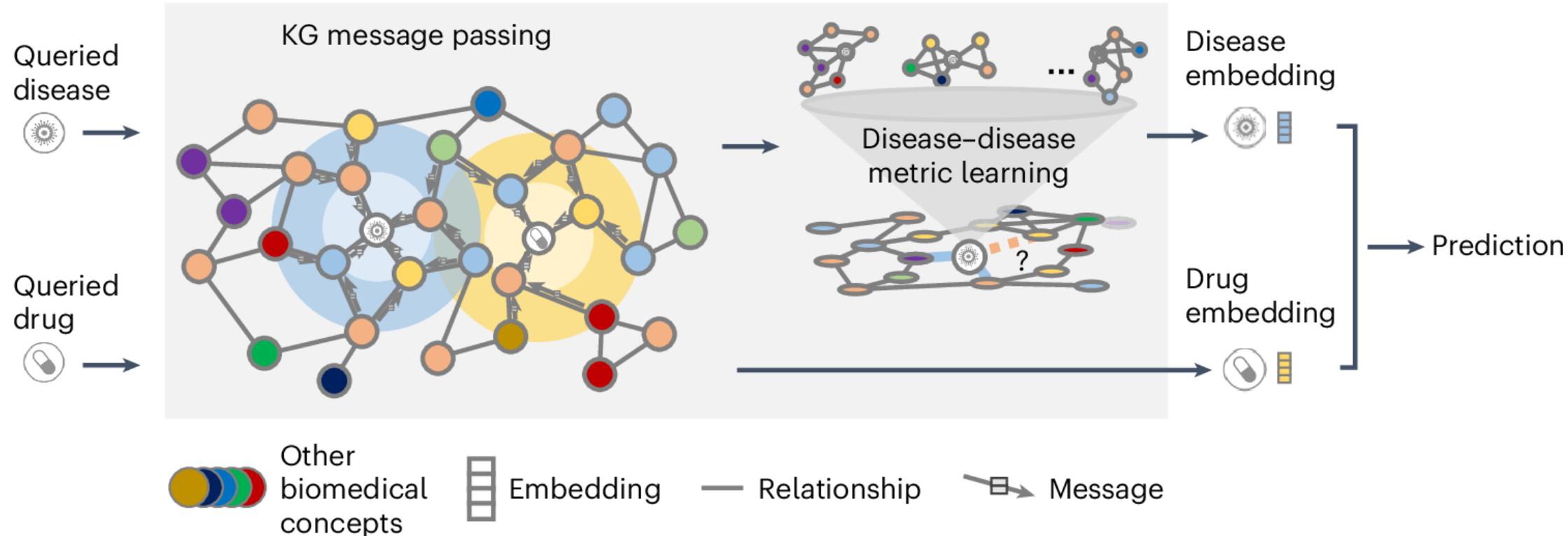
Drug repurposing – Artificial Intelligence

Our GNN-models: Comparison

Input embedding	Model	PR-AUC Mean ± SD
None	DMSR	0.917 ± 0.0047
Drug	DMSR	0.921 ± 0.0036
Protein	DMSR	0.933 ± 0.0117
Drug and protein	DRAGON	0.945 ± 0.0026

Drug repurposing – Artificial Intelligence

Other GNN-based models: TxGNN

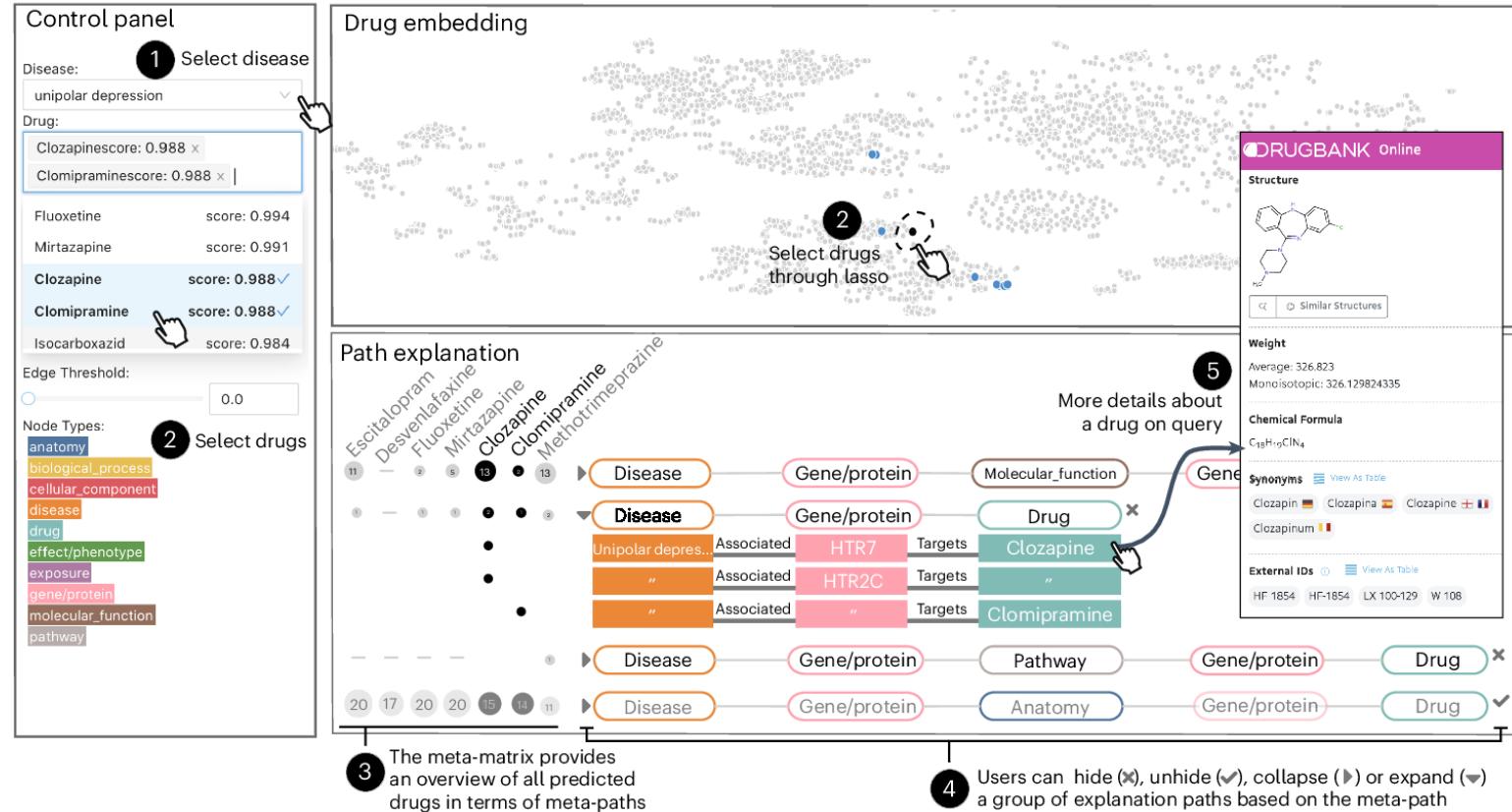


(Huang *et al.*, 2024)

Drug repurposing – Artificial Intelligence

Other GNN-based models: TxGNN

<http://txgnn.org/>



(Huang *et al.*, 2024)



POLITÉCNICA



New nosological models



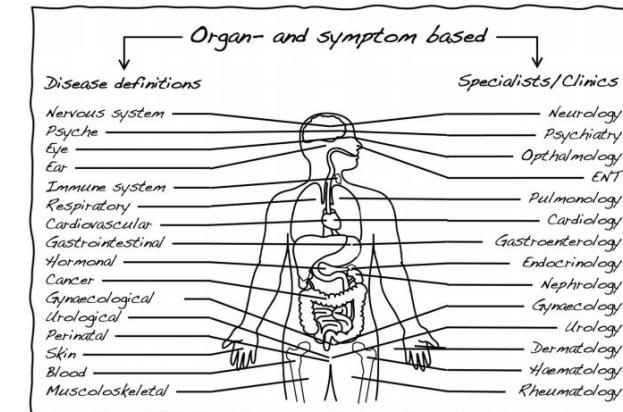
Disease understanding: Dealing with complex and unstructured big data in biomedical domain

138

New nosological models

Nosology

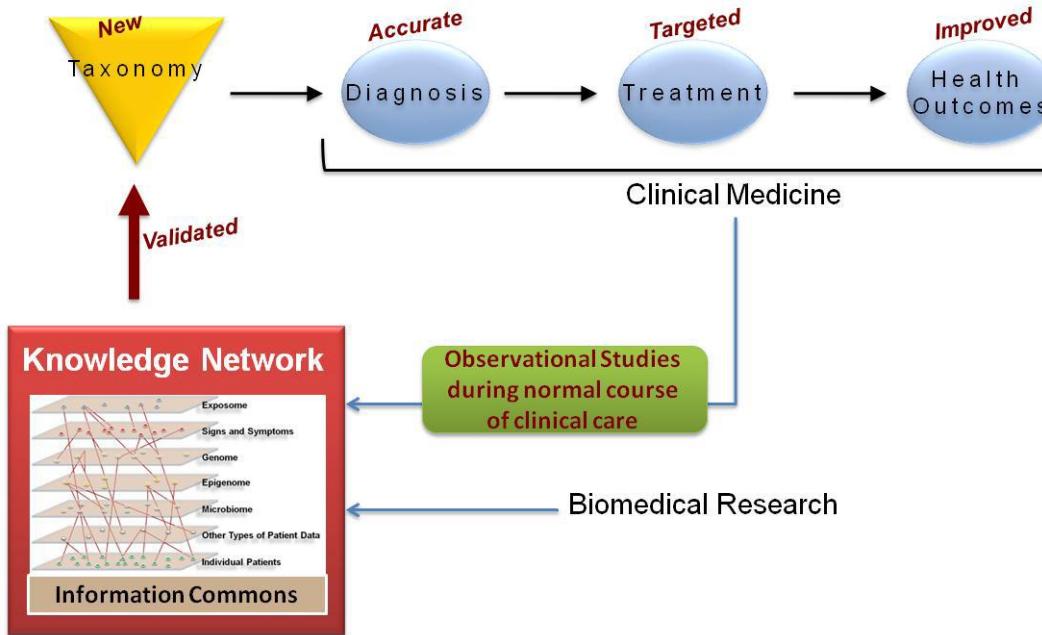
- Branch of medicine in charge of studying the classification of diseases
- Important for correct diagnosis, prevention and treatment
- Multiple approaches
 - Geographical
 - Epidemiological
 - Legal
 - **Organ-based medicine**



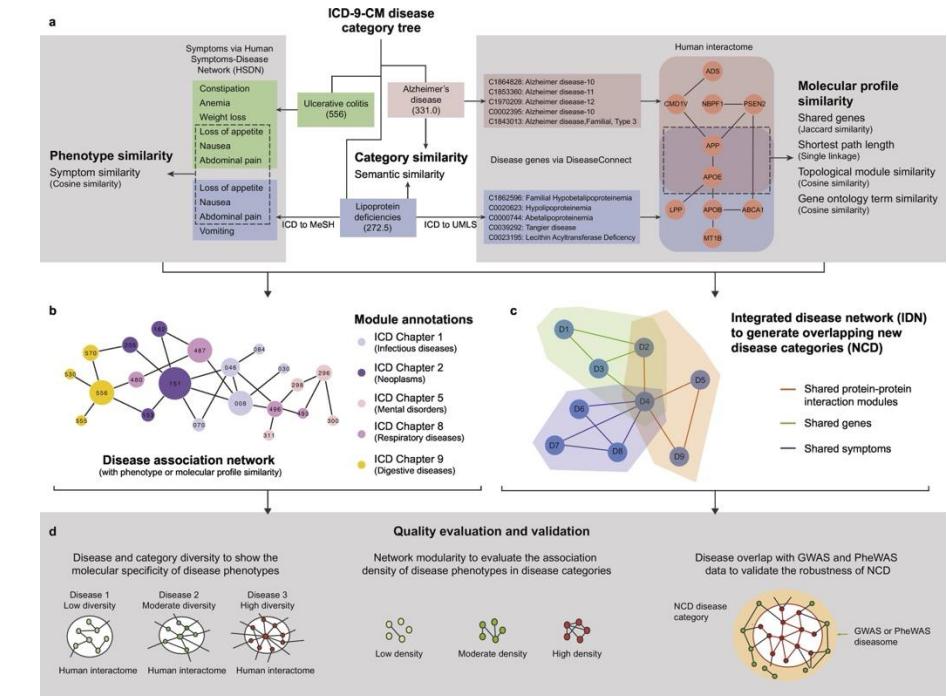
(Schmidt, 2022)

New nosological models

The need of a modern taxonomy of human disease



(U.S. National Academy of Sciences, 2011)



(Zhou et al., 2018)

New nosological models

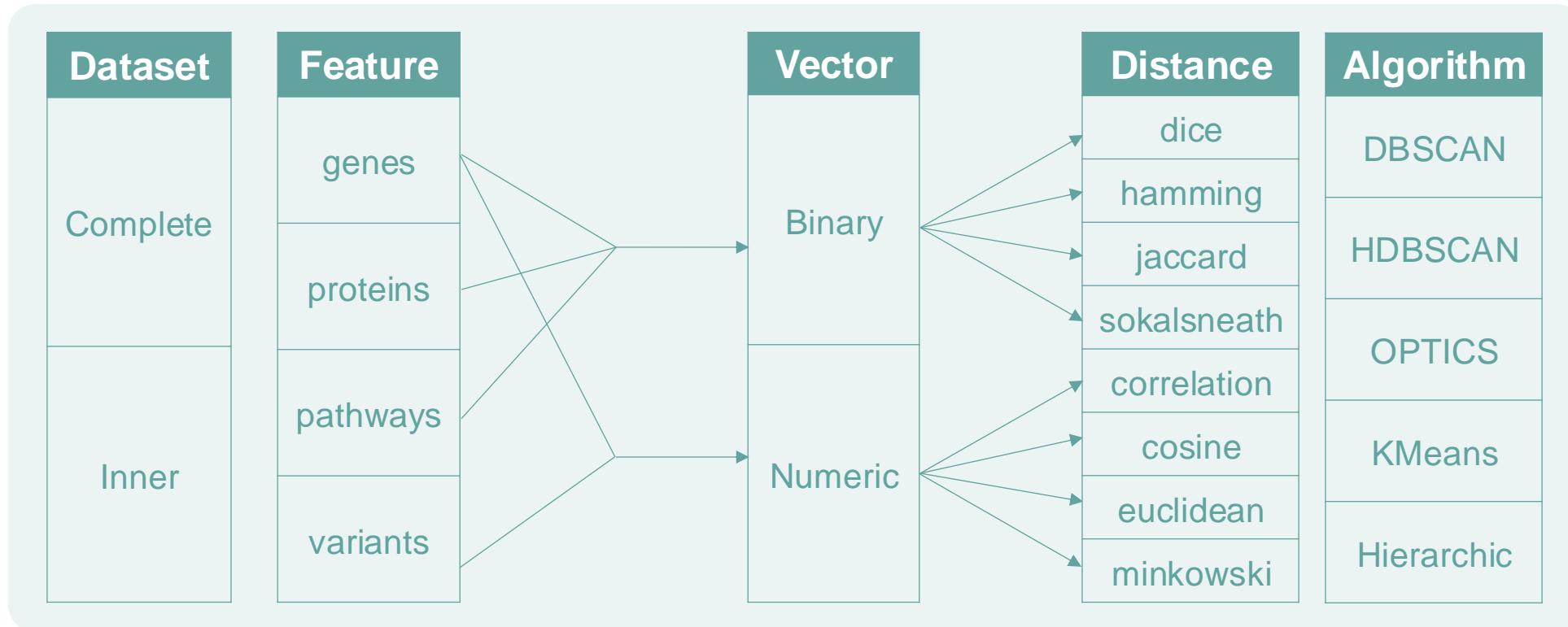
With the data that we had collected...

... any idea on how could we generate new nosological models?



New nosological models

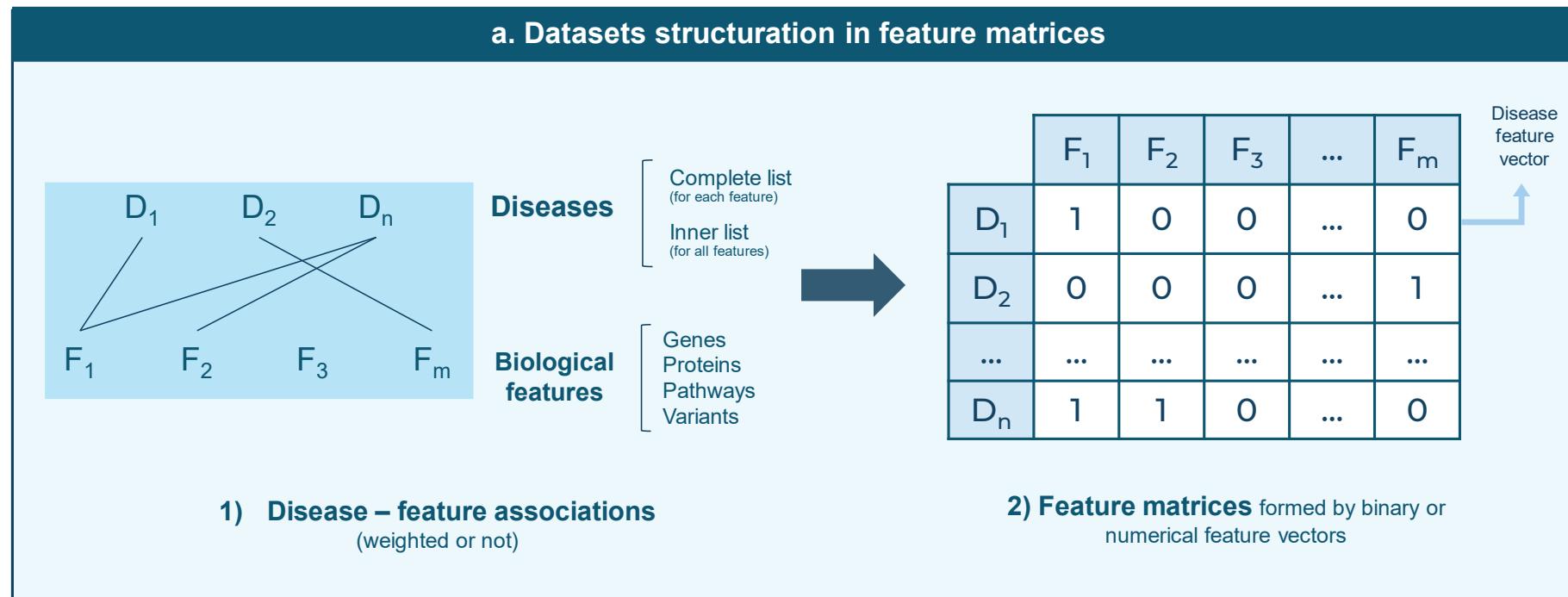
Generating nosological models following a clustering analysis pipeline



(Prieto-Santamaría *et al.*, 2021)

New nosological models

Generating nosological models following a clustering analysis pipeline



(Prieto-Santamaría *et al.*, 2021)

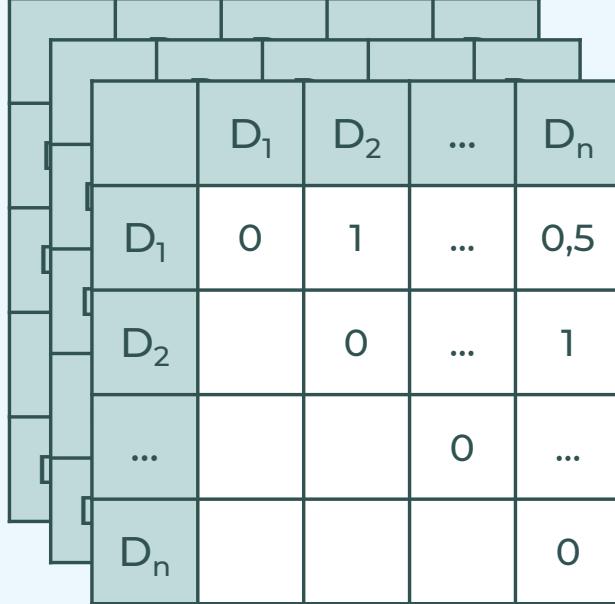
New nosological models

Generating nosological models following a clustering analysis pipeline

b. Computing distance matrices

Metrics

- **Binary vectors:**
 - dice
 - hamming
 - jaccard
 - sokalsneath
- **Numeric vectors:**
 - correlation
 - cosine
 - euclidean
 - minkowski



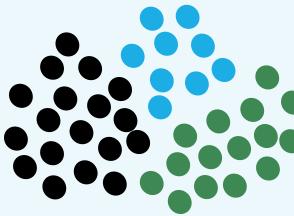
	D ₁	D ₂	...	D _n
D ₁	0	1	...	0,5
D ₂		0	...	1
...			0	...
D _n				0

(Prieto-Santamaría *et al.*, 2021)

New nosological models

Generating nosological models following a clustering analysis pipeline

c. Clustering implementation and evaluation	
Clustering algorithms	Evaluation
- DBSCAN	- Number of clusters
- HDBSCAN	- Noise
- OPTICS	- Silhouette
- KMeans	- Calinski-Harabasz
- Hierarchical	- Davies-Bouldin
	- SSE



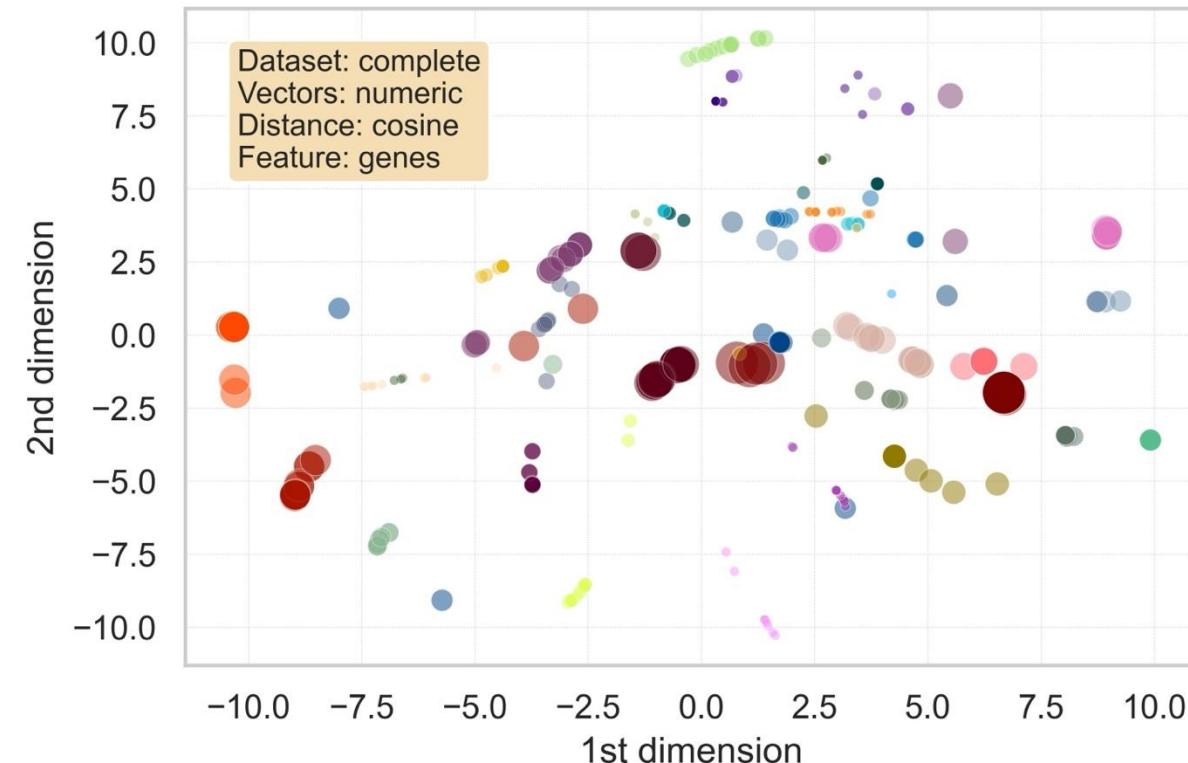
(Prieto-Santamaría *et al.*, 2021)

New nosological models

Generating nosological models following a clustering analysis pipeline

- Visualization of one of the clustering nosological models (obtained by OPTICS)

(Ankerst et al., 1999)

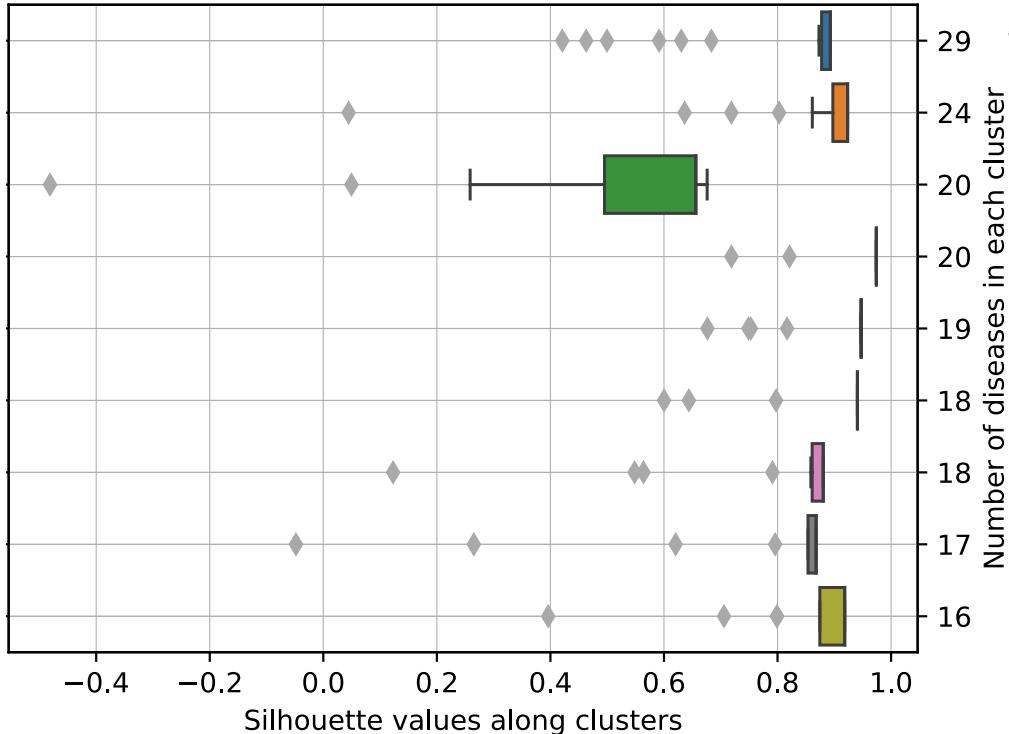


(Prieto-Santamaría et al., 2021)

New nosological models

Generating nosological models following a clustering analysis pipeline

- Distribution of Silhouette coefficient in the clusters formed in the previous clustering model

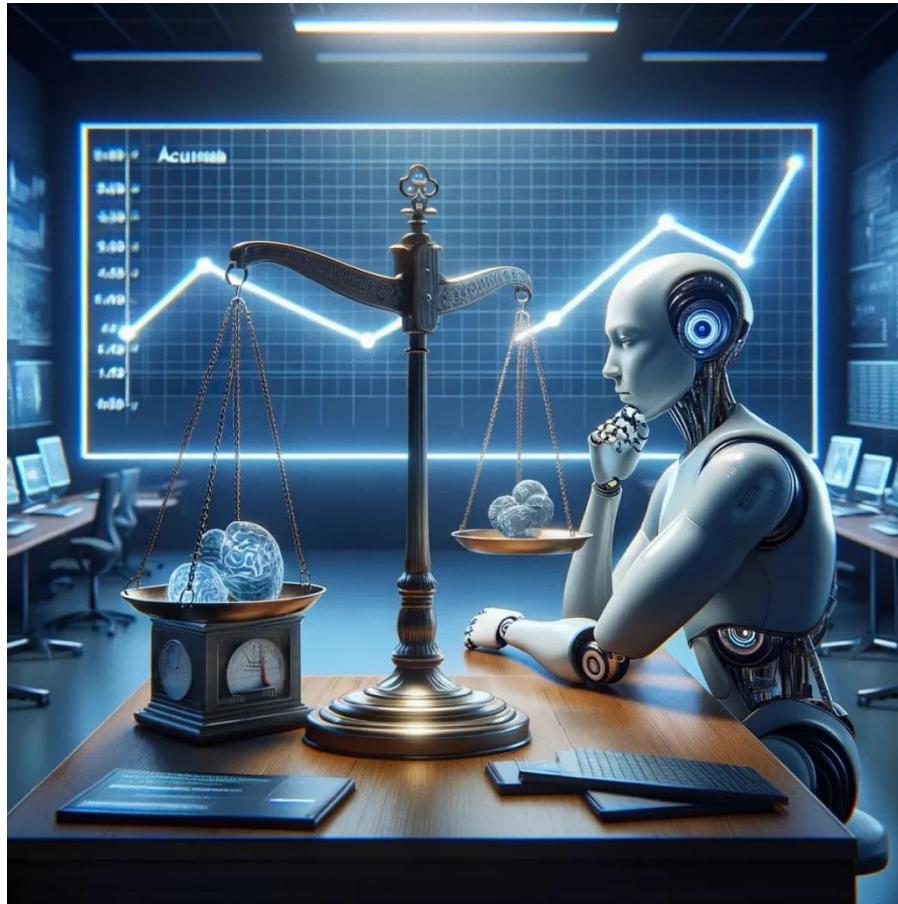


- ACTH Syndrome
- Adrenal Gland Hyperfunction
- Arthritis, Gouty
- Facial paralysis
- Hypernatremia
- Cerebral Palsy
- Subaortic stenosis
- Opsoclonus-Myoclonus Syndrome
- Sacroiliitis
- Kinsbourne Syndrome
- Proopiomelanocortin Deficiency

MeSH categories

- Neoplasms
- Endocrine
- Musculoskeletal
- Nervous
- Nutritional and metabolic
- Cardiovascular
- Eye

Bias in AI



- Selection / Data Bias
- Algorithmic Bias
- Confirmation Bias
- Automation Bias

<https://www.psychologytoday.com/sg/blog/a-hovercraft-full-of-eels/202402/is-bias-in-ai-necessarily-a-problem>



Does biological sex influence drug response?

Clinical trials

Female participation in clinical trials throughout history has been very limited.

Results of males were applied to females

Females were excluded for various reasons:

- Probability of pregnancy
- Breastfeeding
- Hormonal changes :
 - Menstrual cycle
 - Menopause



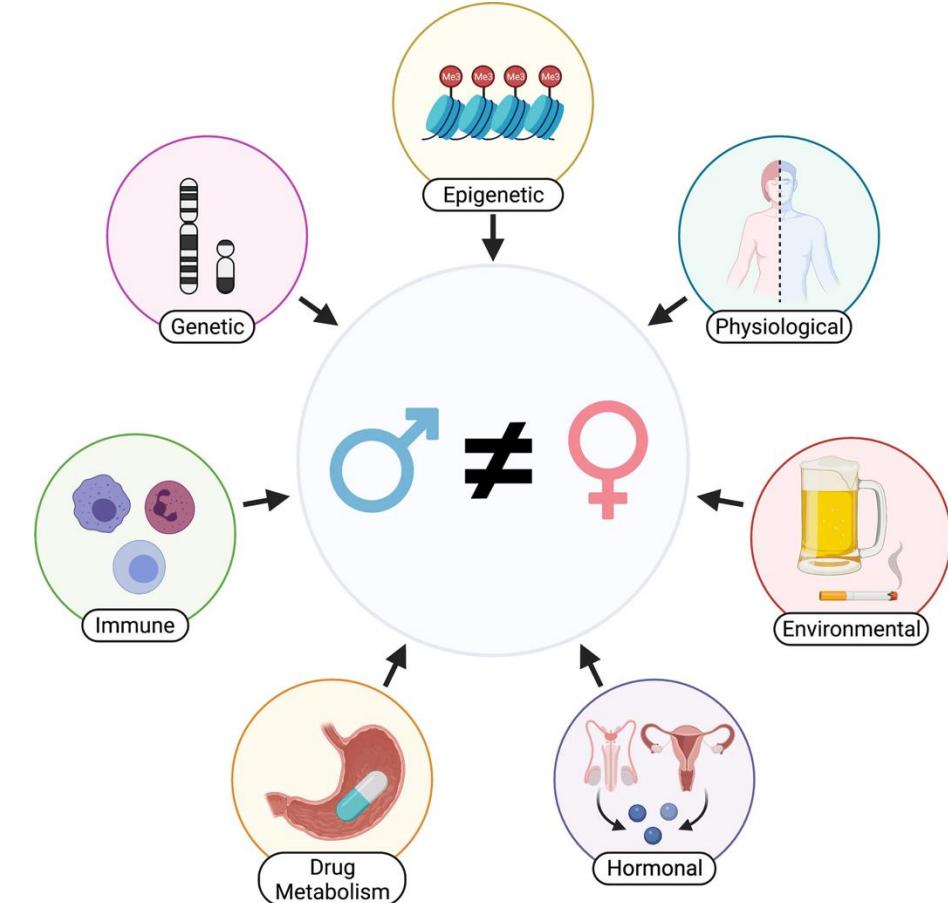
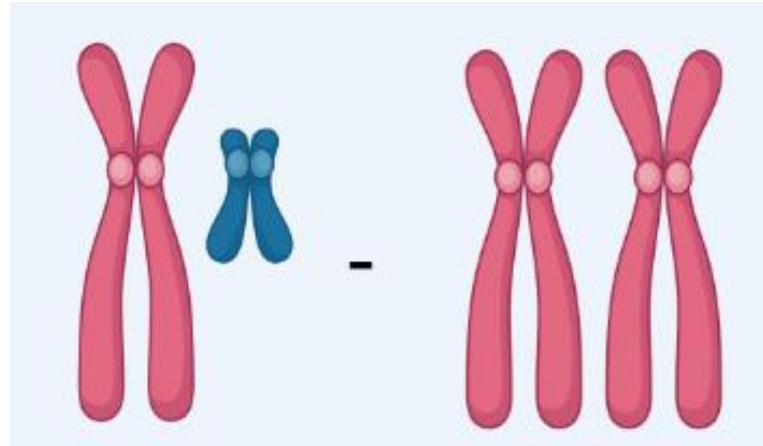
Unknown Adverse Drug
Reactions (ADRs)



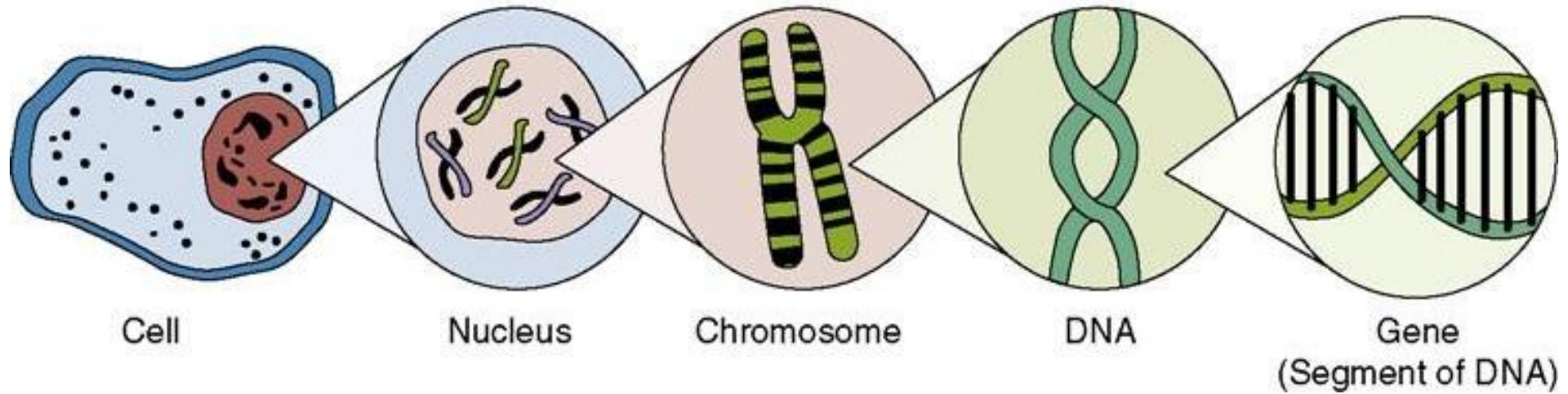
Pre-clinical studies

76% male cell-based
75% animal studies

Sex chromosomes



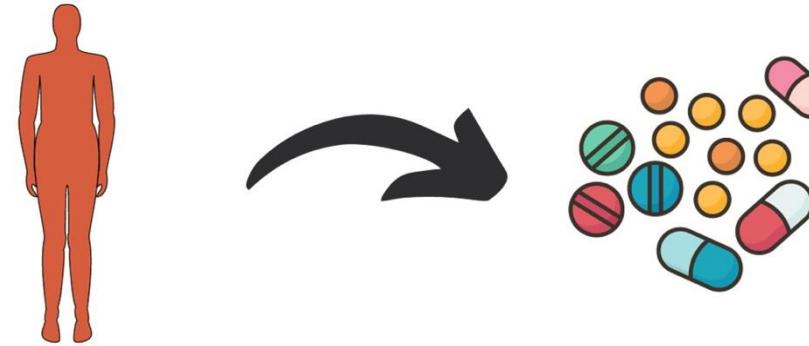
Genes expression



<http://www.cancermoonshotlund.com/index.php/6-what-are-genes-dna-and-proteins/>

Pharmacokinetics and Pharmacodynamics

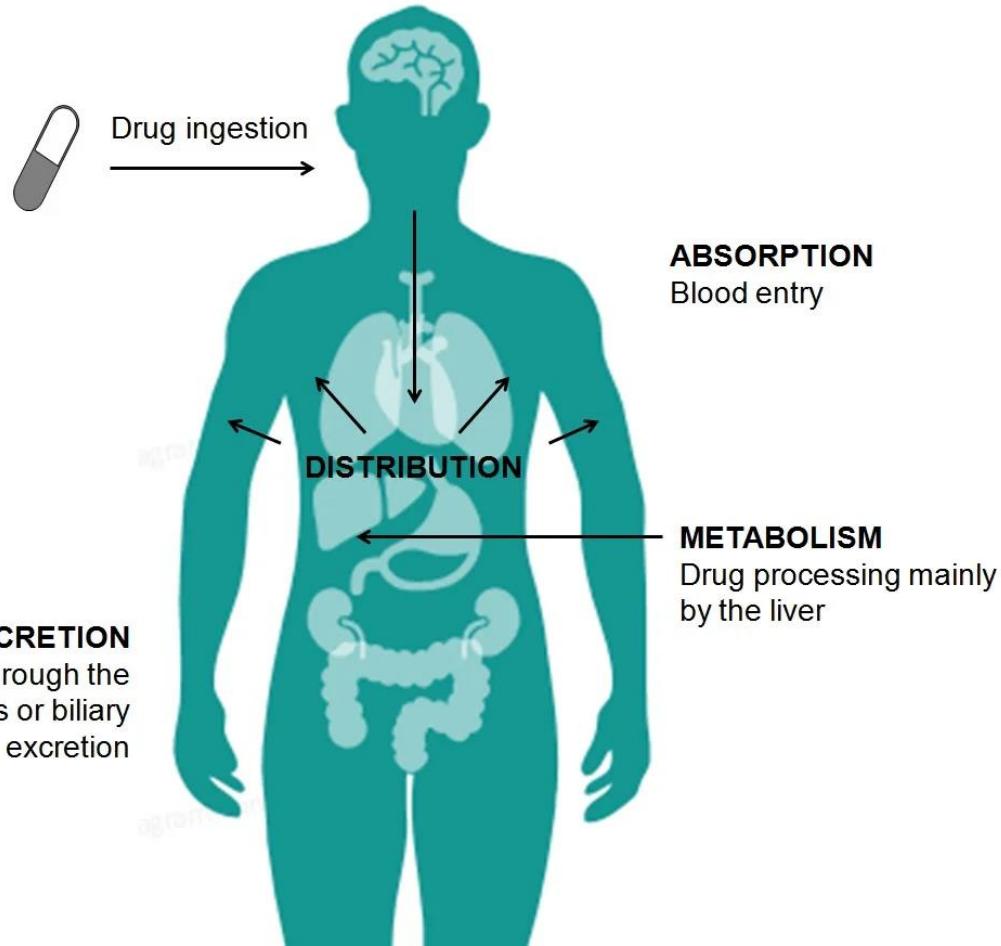
Pharmacokinetics. It is defined as the part of pharmacology that studies the processes that the organism produces on the drug from its administration to its elimination.



Pharmacodynamics. It is defined as the part of pharmacology that studies the effects that the drug produces on the organism.



Absorption: Gastric pH



Absorption:

Females have a higher gastric pH than males, which may increase the absorption of certain compounds.

<https://www.laxxonmedical.com/post/what-is-pharmacokinetics-four-stages-in-a-drug-traveling-within-the-body>

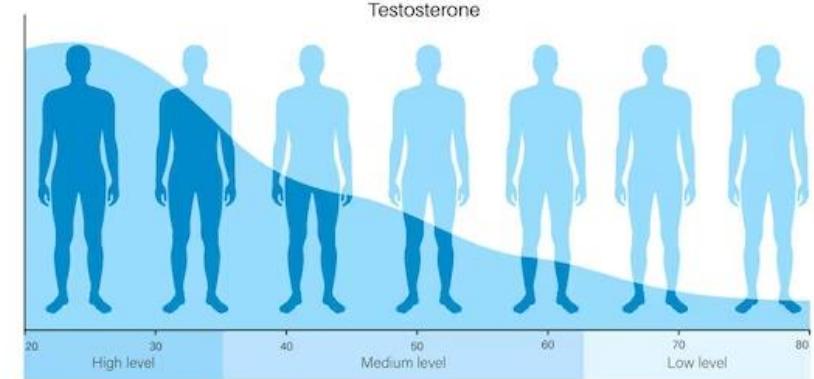
Metabolism: Sex hormones

Liver enzymes are altered by elevated estrogen and progesterone levels

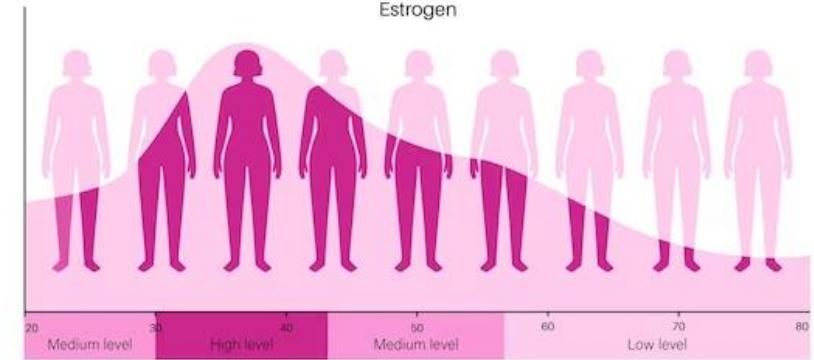


Drug accumulation

SEX HORMONE PRODUCTION BY AGE

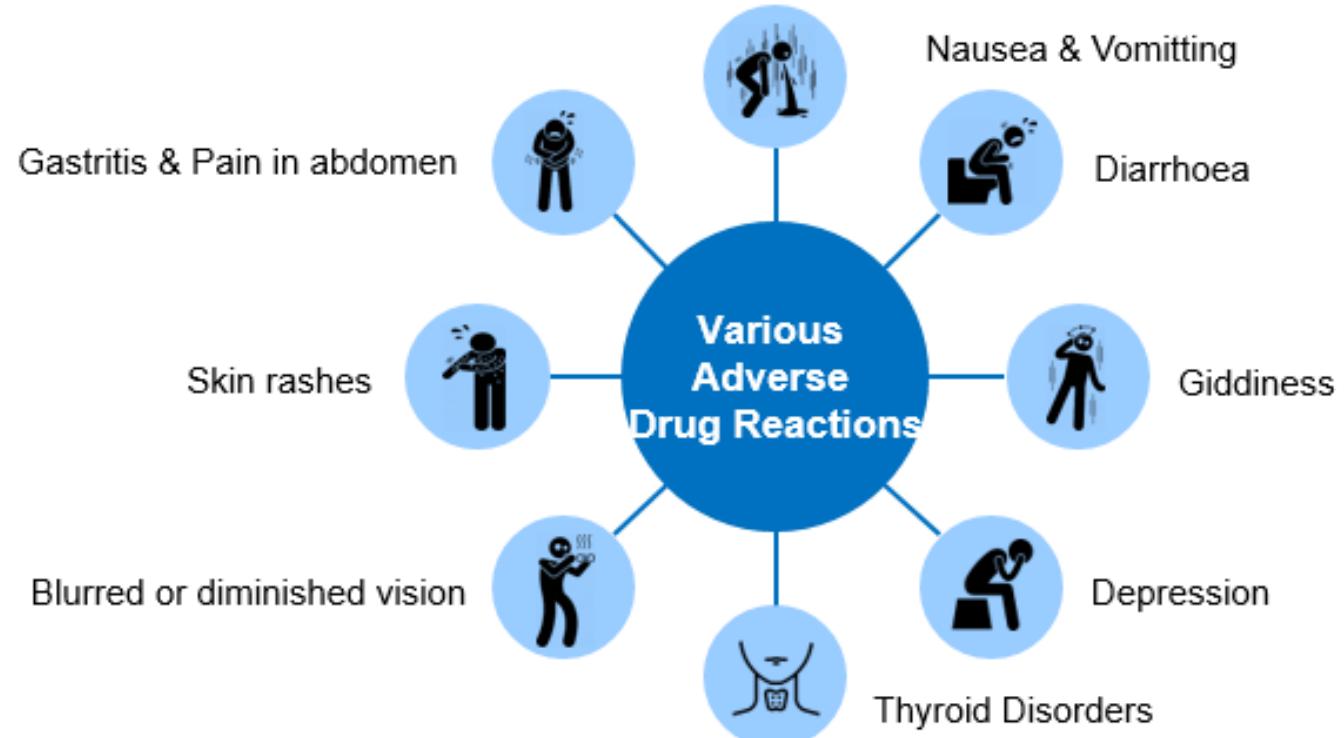


Estrogen



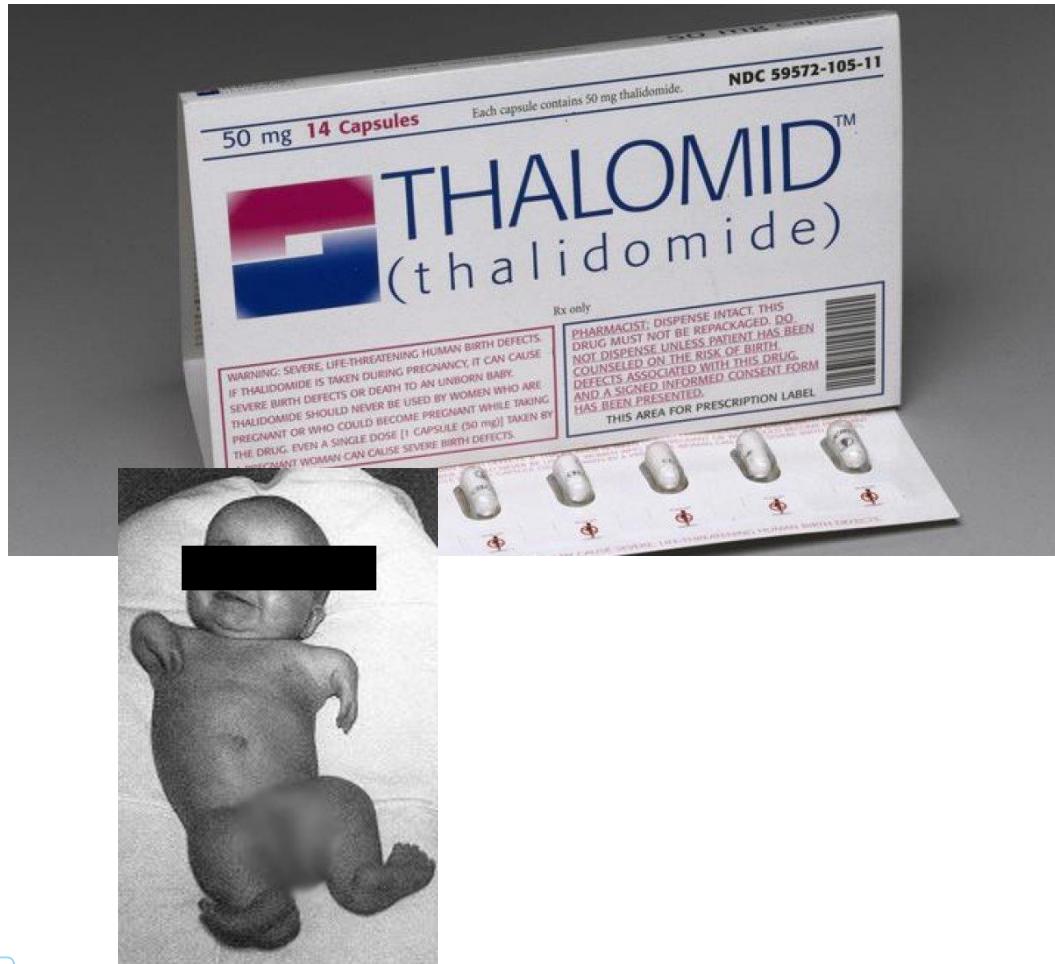
What are the side effects?

Side effects, also known as adverse reactions, are unwanted effects that are possibly related to a drug.



<https://ntep.in/node/605/CP-adverse-drug-reactions>

THALIDOMIDE

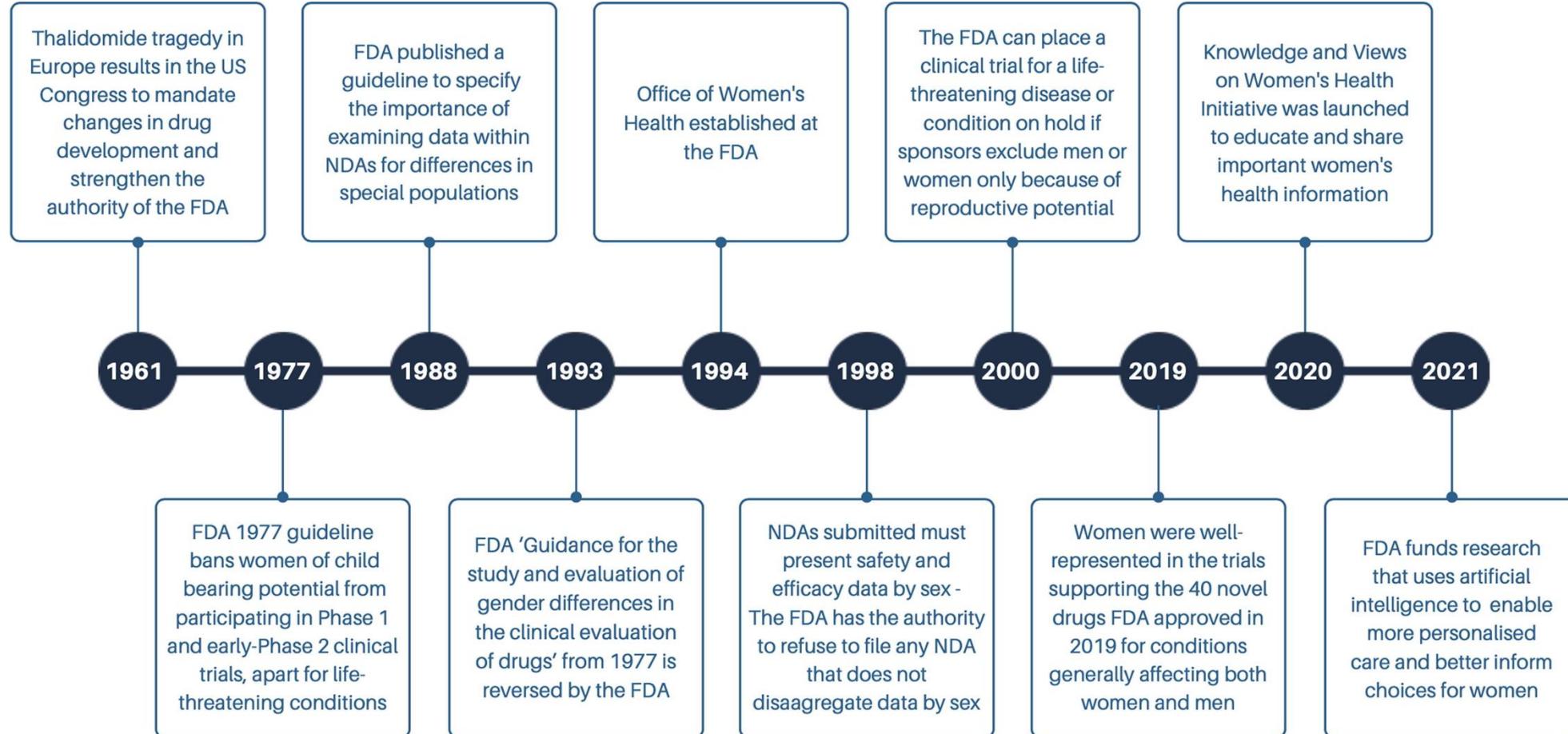


- Developed in 1953
- Launched in 1956
- Used as a sedative and for morning sickness in pregnant women.
- In 1961, side effects began to be reported in newborns

Congenital malformations
Phocomelia

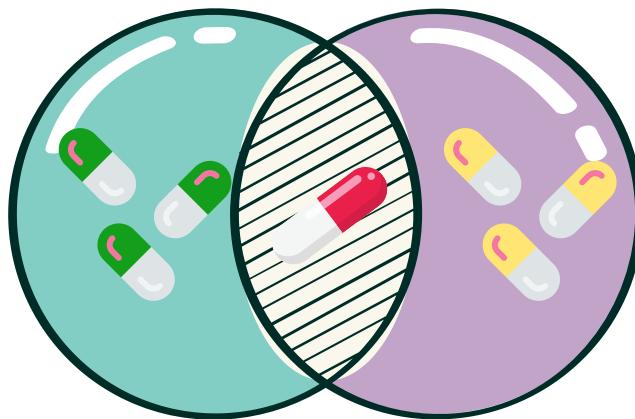
- Thalidomide was withdrawn from the market

Change of mind



Motivation

Objective: Build a molecular model to detect the sex divergent anticancer drug effects including adverse drug reactions and variable drug efficacy by sex.



Genome



Tissue



Phenotype

Part 1: Do small-molecule anticancer drugs alter sex-biased gene expression?
- in sex chromosomes
- in sex-biased autosomes

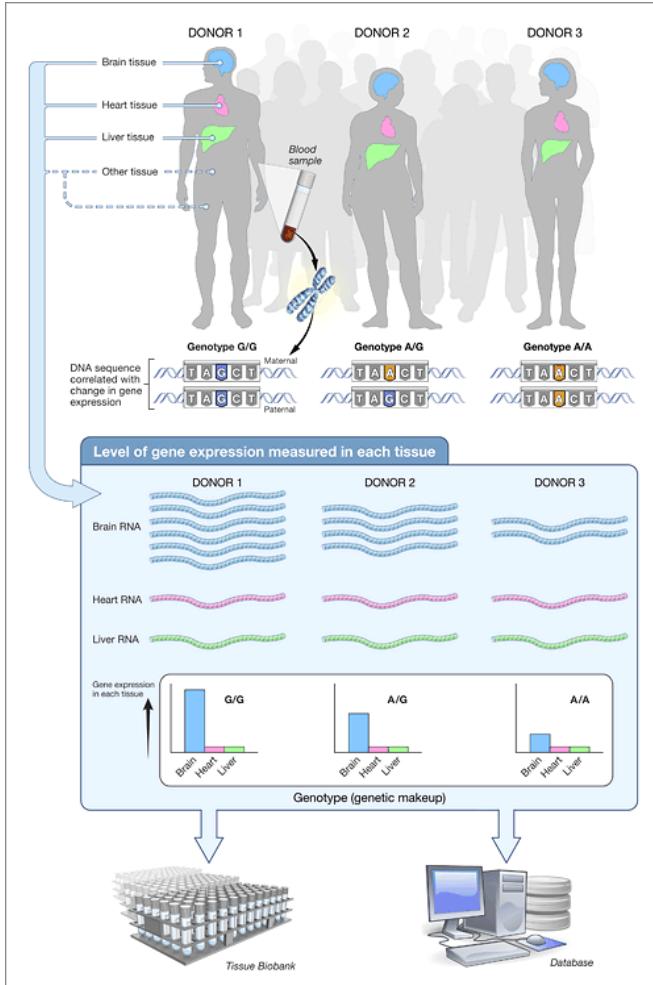
Part 2: Do small-molecule anticancer drugs alter tissue-specific genes?

Part 3: Do drug that modulate sex and tissue-specific genes induce sex-specific adverse drug reactions?



Sex-biased genes by genome and tissues when applying cancer drugs

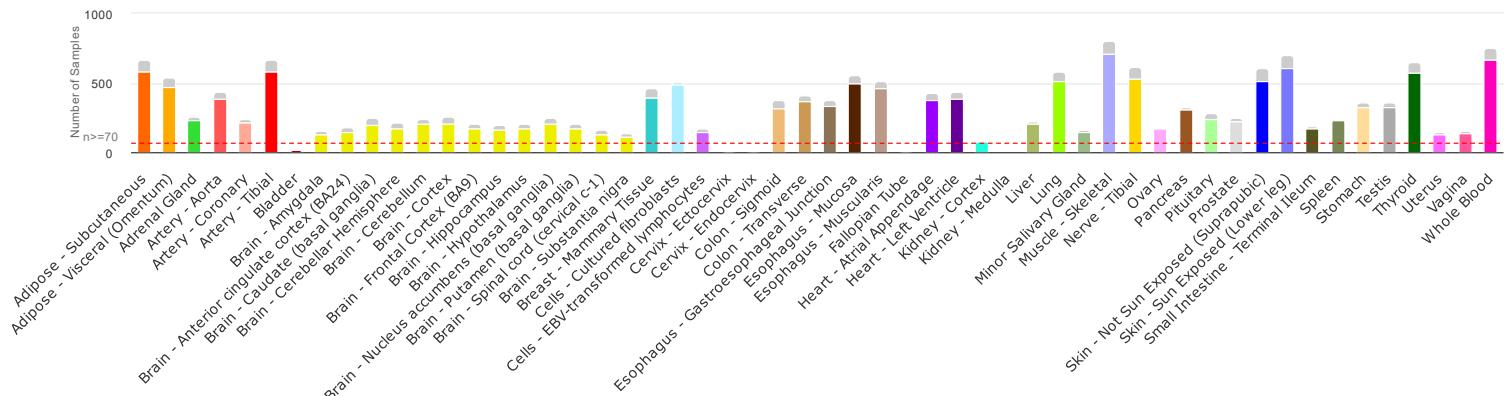
Data integration - Sex-biased cancer drugs



V8 Sample Info

V8 Release	# Tissues	# Donors	# Samples
Total	54	948	17382
With Genotype	54	838	15253
Has eQTL Analysis*	49	838	15201

* Number of samples with genotype ≥ 70

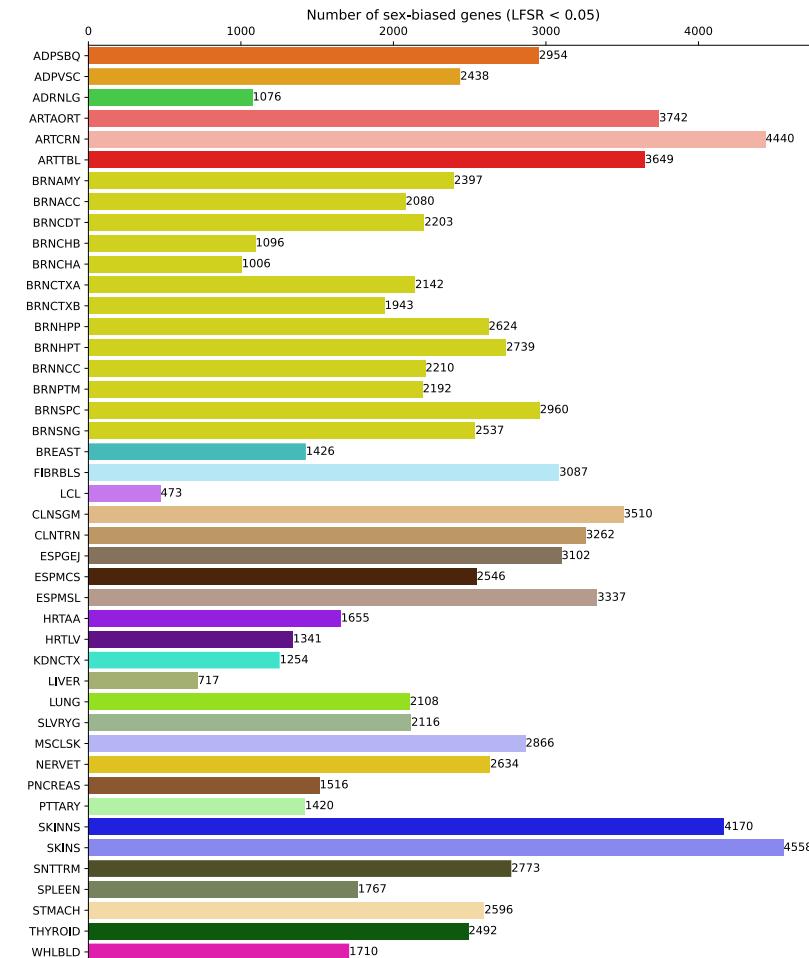


Data integration - Sex-biased cancer drugs



Sex-biased genes
by 44 tissues

GTEx_Analysis_v8_sbgenes



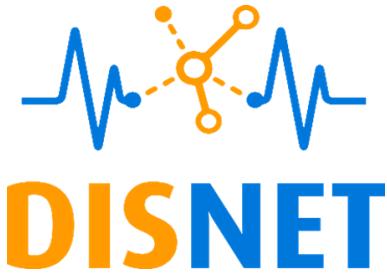
The impact of sex on gene expression across human tissues
(Oliva et al. 2020)

Data integration - Sex-biased cancer drugs



ConnectivityMap

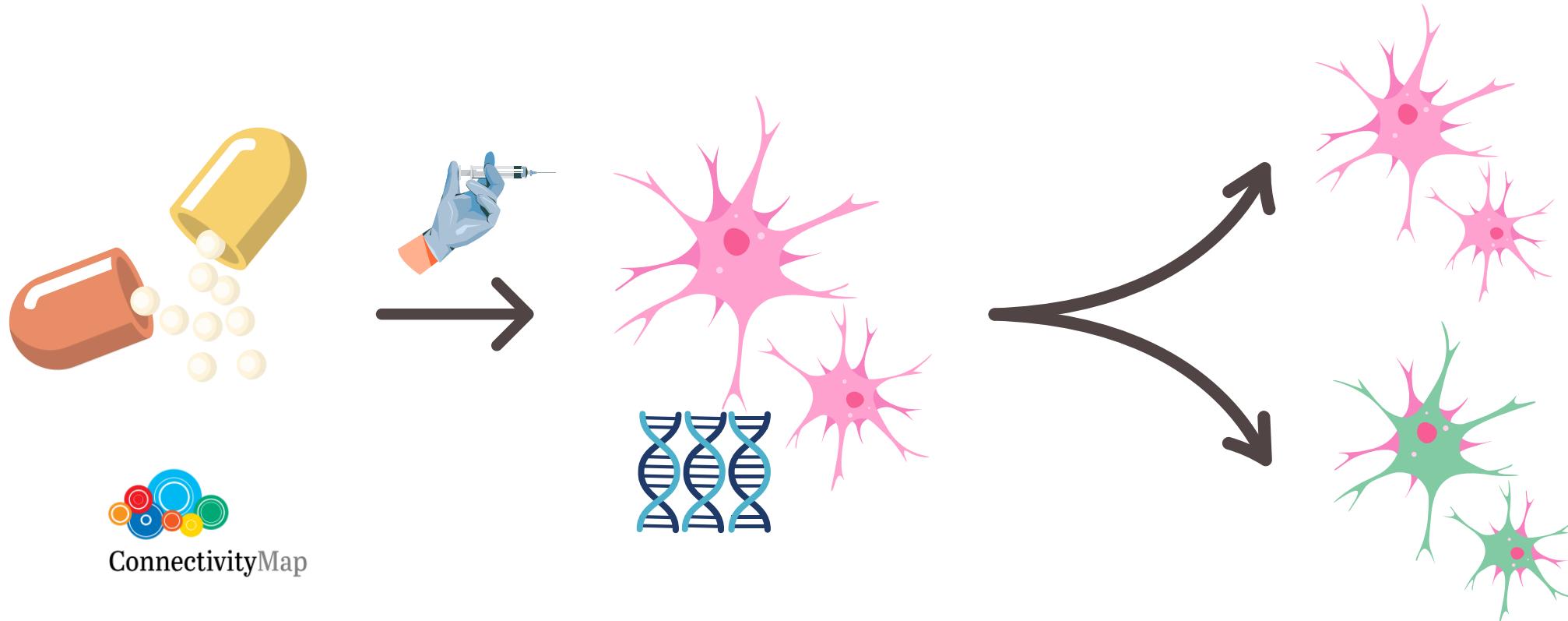
Touchstone - Level 5 - ModZScore



ATC_Code classification
(Antineoplastic and
immunomodulating
agents)

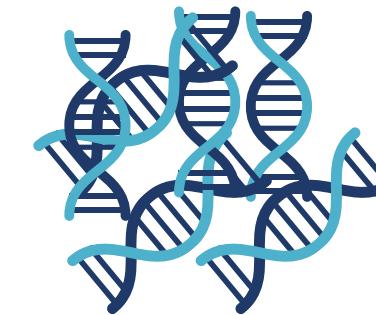
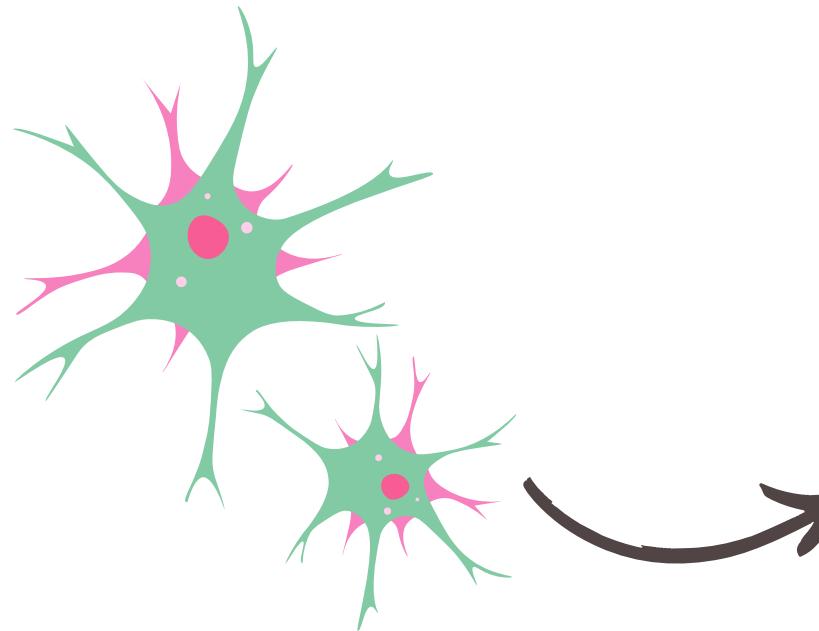
95
cancer drugs

Methodology – Sex-biased cancer drugs

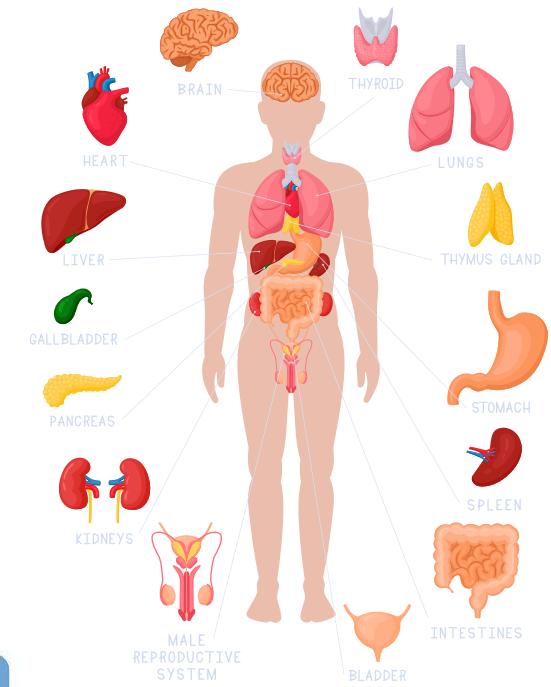


Mod Z-score > 2
Mod Z-score < -2

Methodology – Sex-biased cancer drugs



Sex-biased

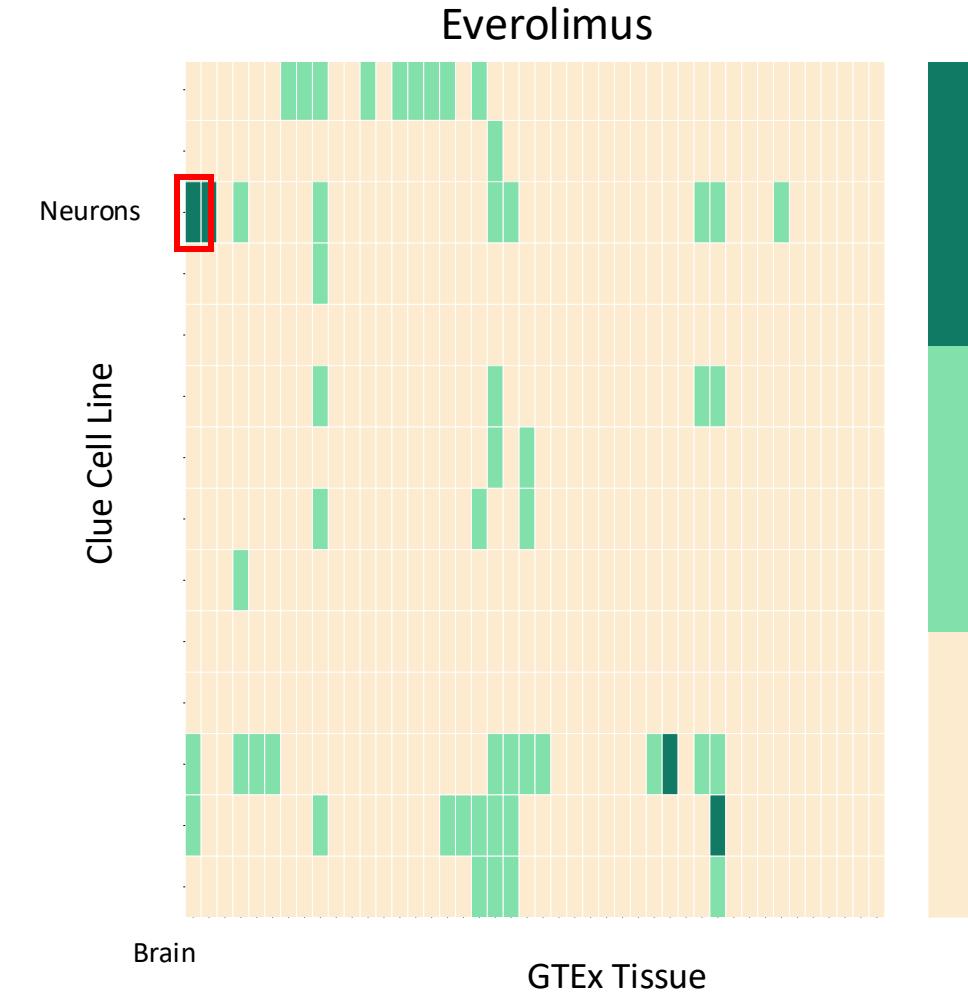


Results - Sex-biased cancer drugs

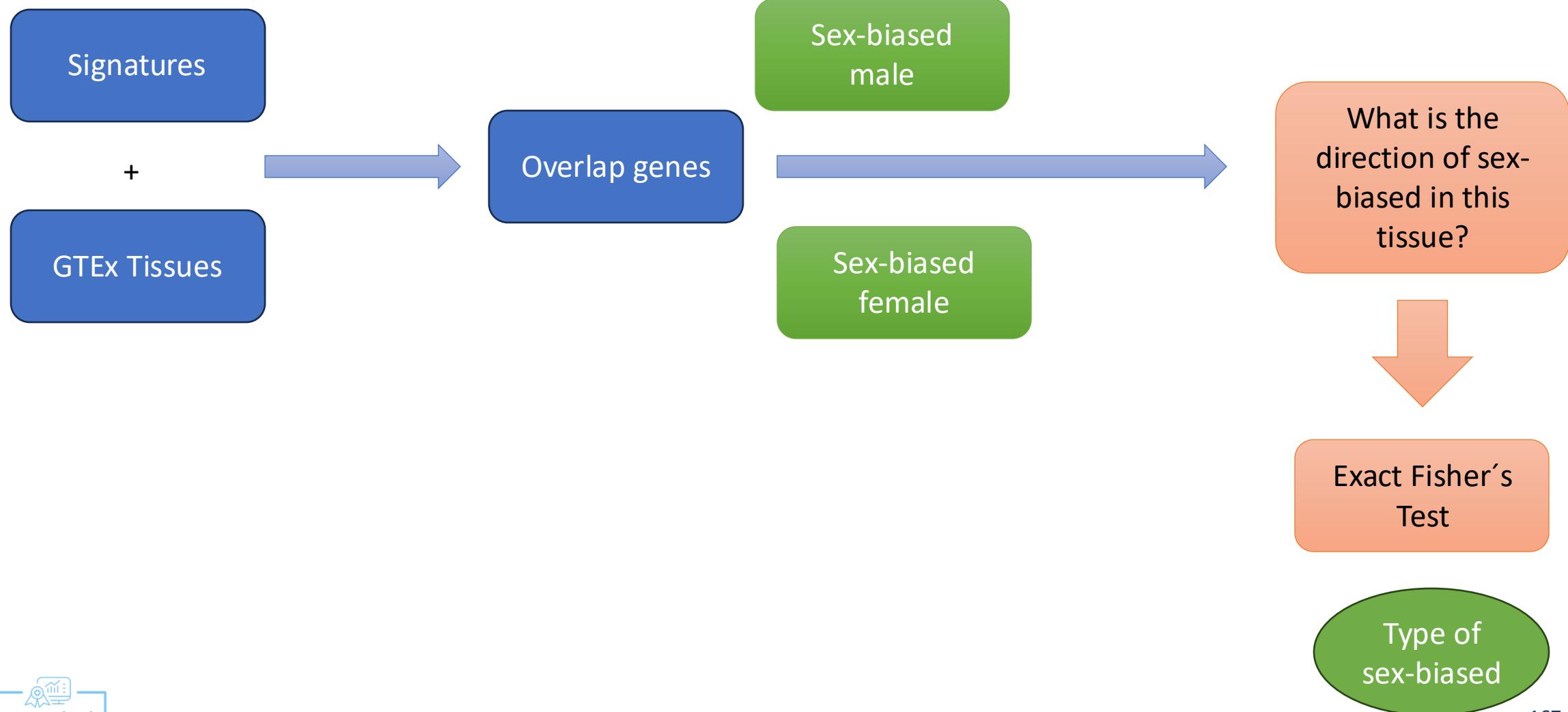
95 different anticancer drugs have been studied and all of them have a sex bias in at least one tissue.

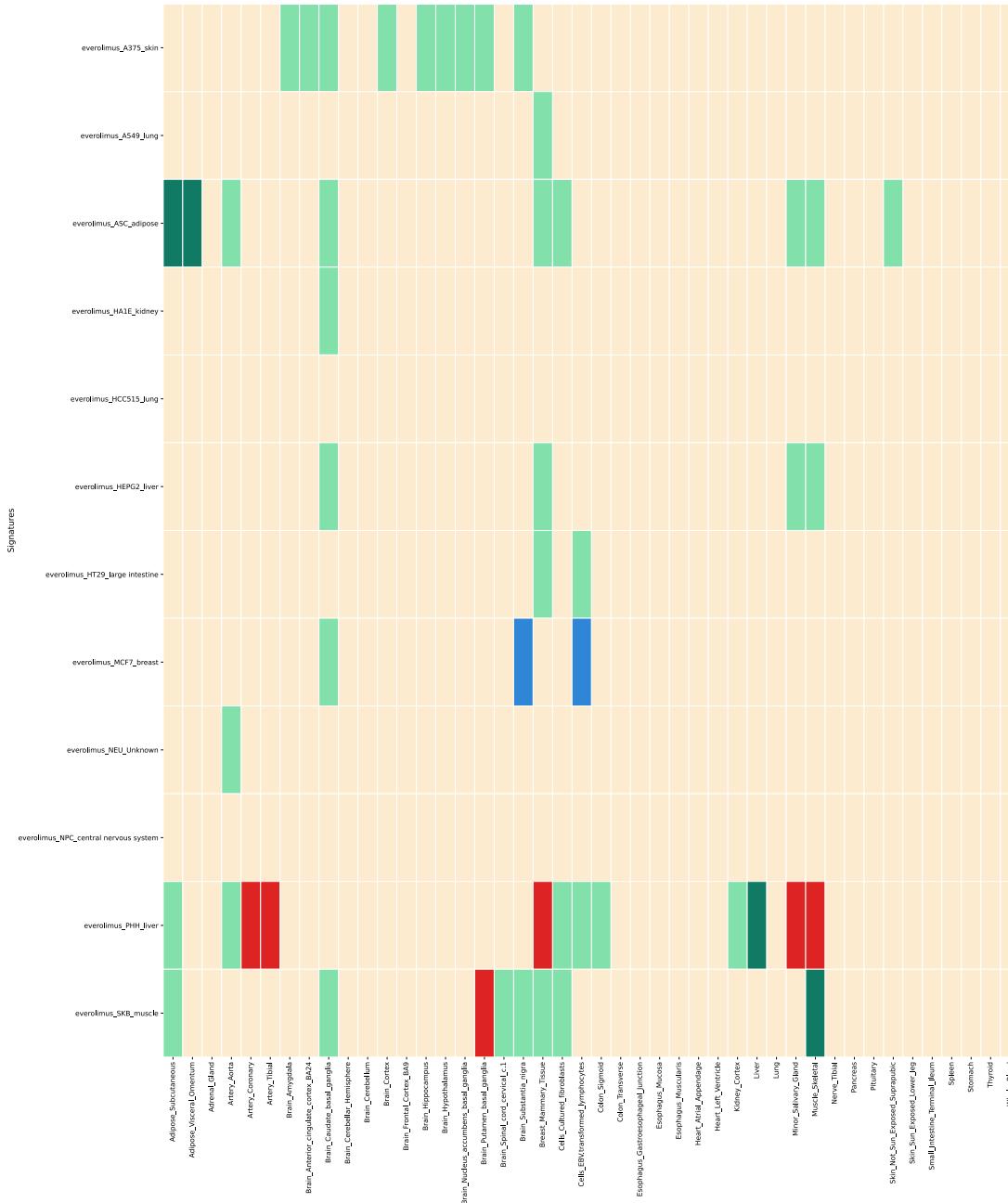
Everolimus: Drug to treat different types of malignant neoplasms.

- ❖ Light green: genes that change their expression when the drug is applied and have a sex-bias in certain tissues.
- ❖ Dark green: genes that also belong to the same type as the tissues.



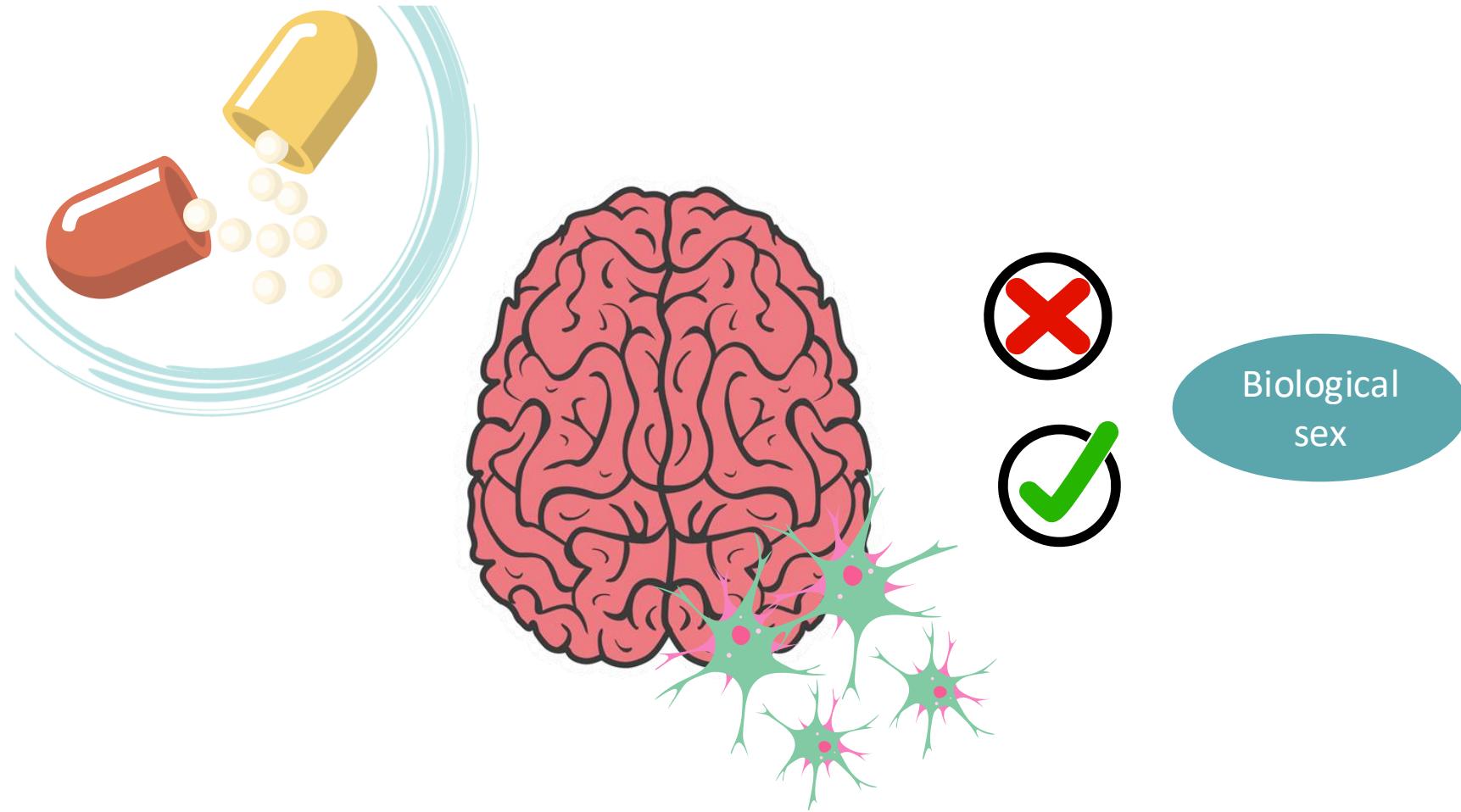
Sex-biased directions



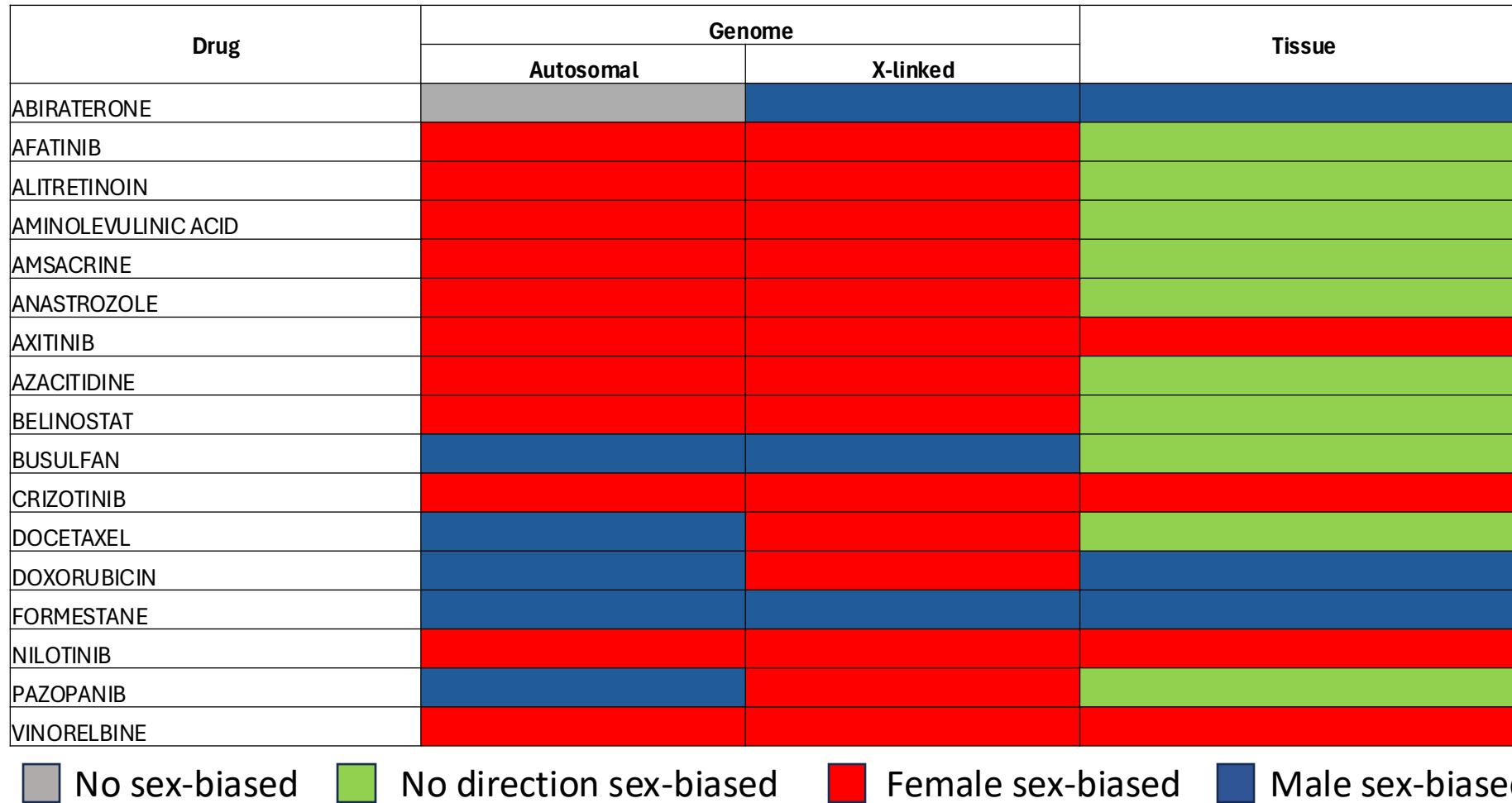


- █ No direction sex-biased
- █ Female sex-biased
- █ Male sex-biased

Methodology – Sex-biased cancer drugs



Results – Genome and Tissue





Phenotype: ADRs in cancer drugs by sex

Data integration - Cancer Drugs ADRs

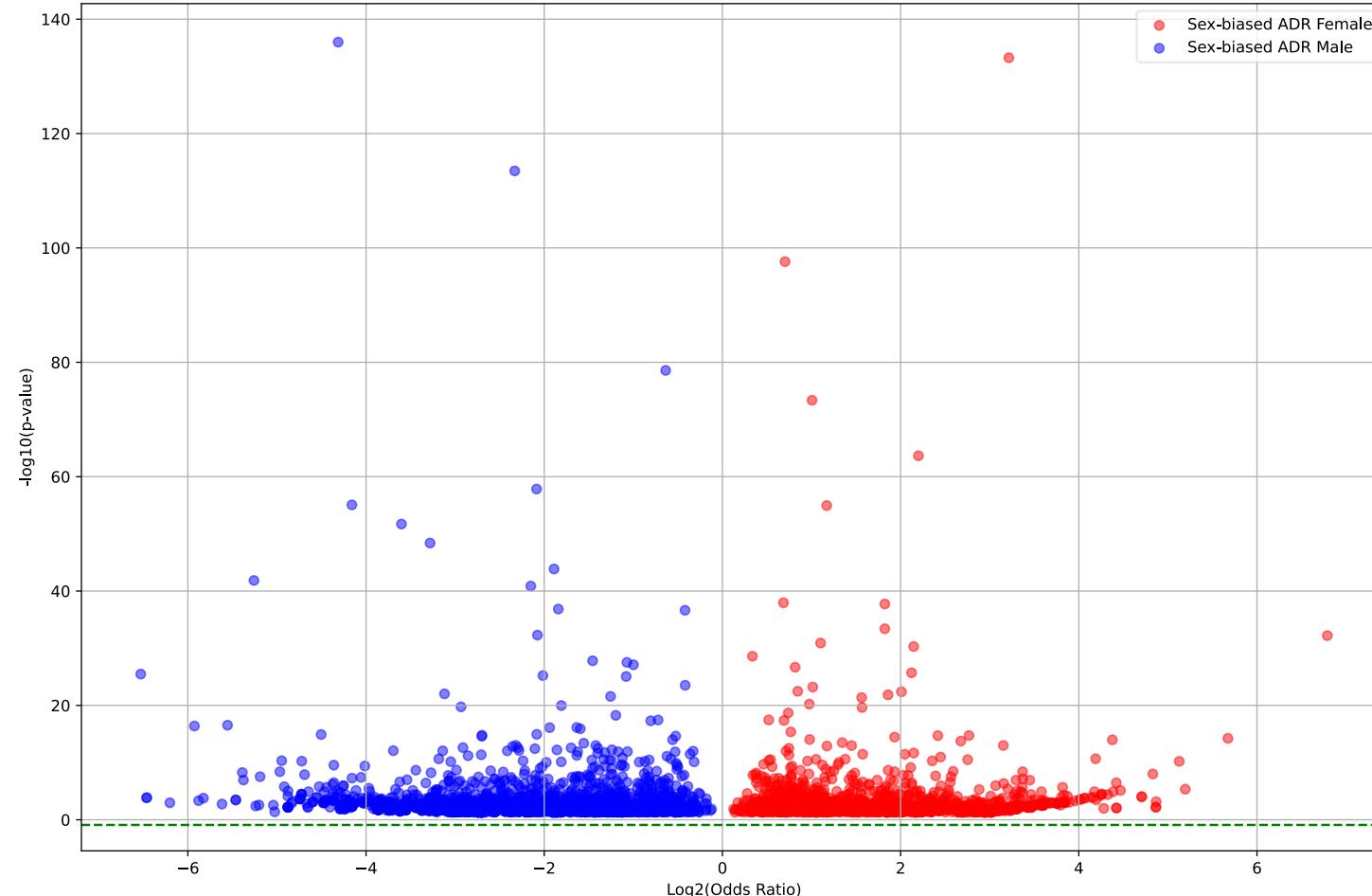
FDA Adverse Drug Reactions	Patient_id	Drugs	ADRs	Age	Sex	Date
	6161368	[10001488', '10002034', '10002855', '10002948', '10003210', '10003239', '10003246]	[10020772', '10021015', '10021024'], ['10021036', '10021519'], ['10022437', '10023232', '10025197',]	52	Male	2013-01-16
					
					
Total of 1.048.576 registers						2004 -2022

Phenotype - Cancer Drugs ADRs

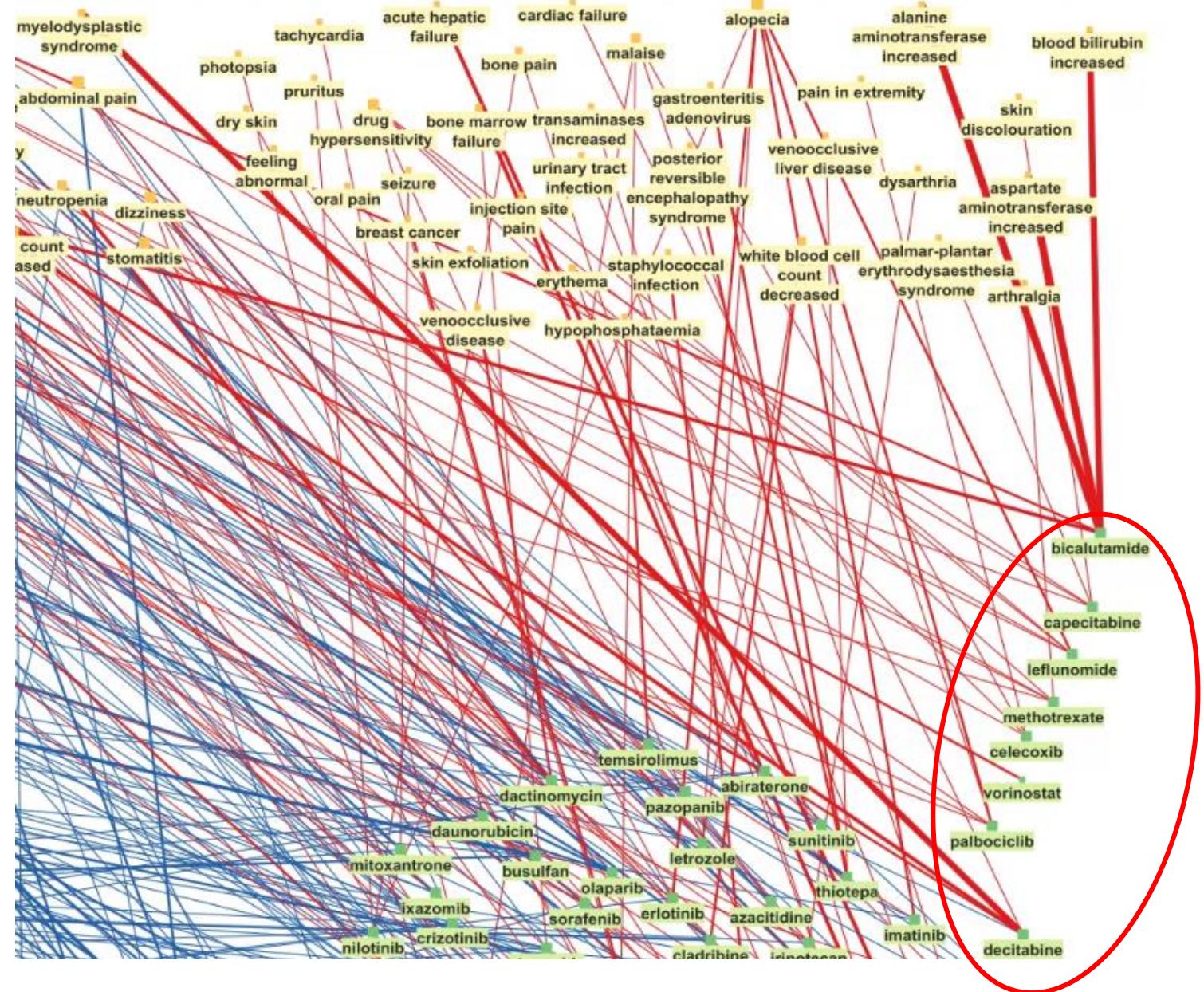


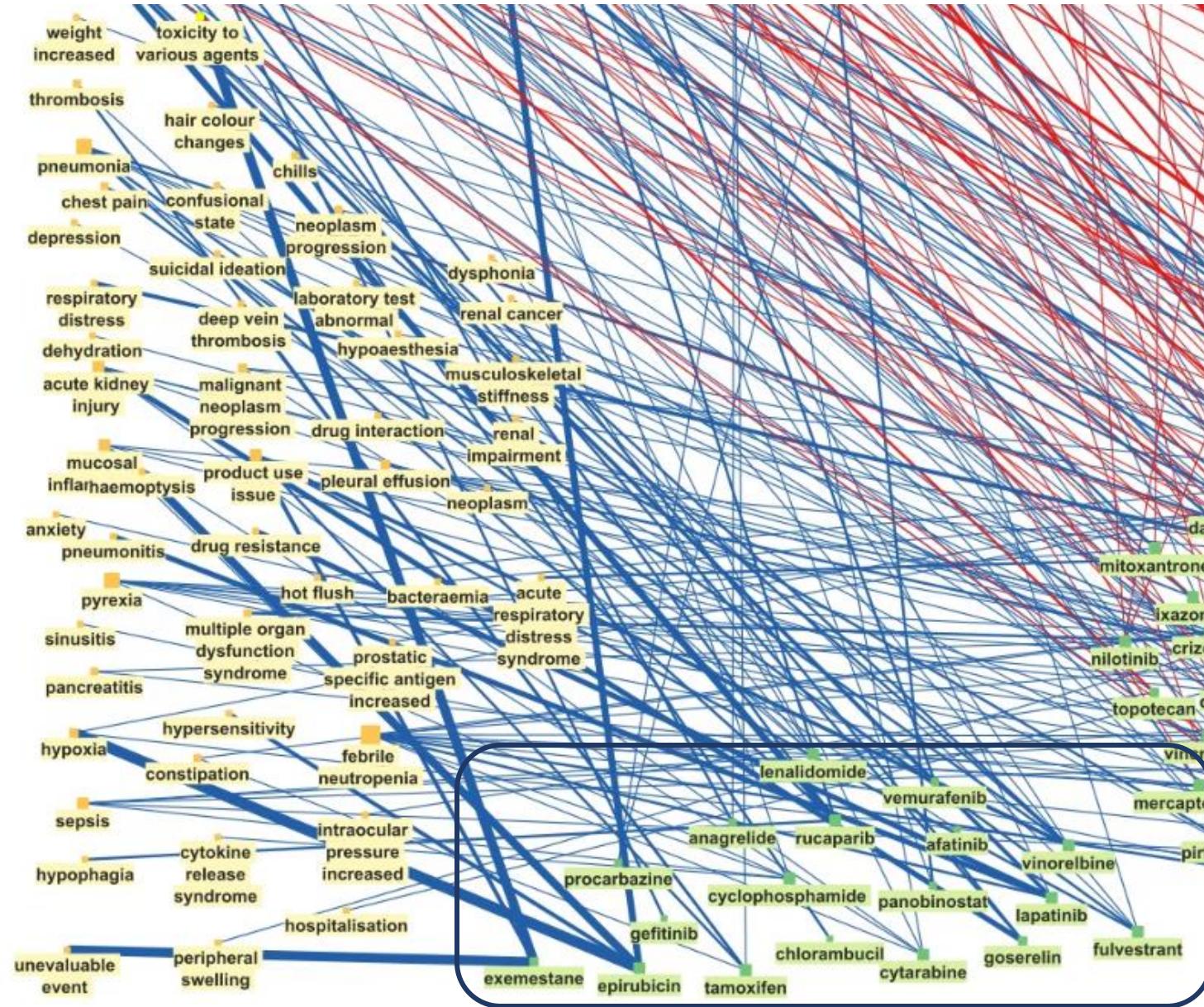
- Drug
- ADRs
- Number of ADRs in male
- Number of ADRs in female
- Sum ADRs by cancer drug
- **P-value**
- **Odds Ratio**
- Log2 Odds Ratio
- Log2 P-value
- **Sex-biased ADR**

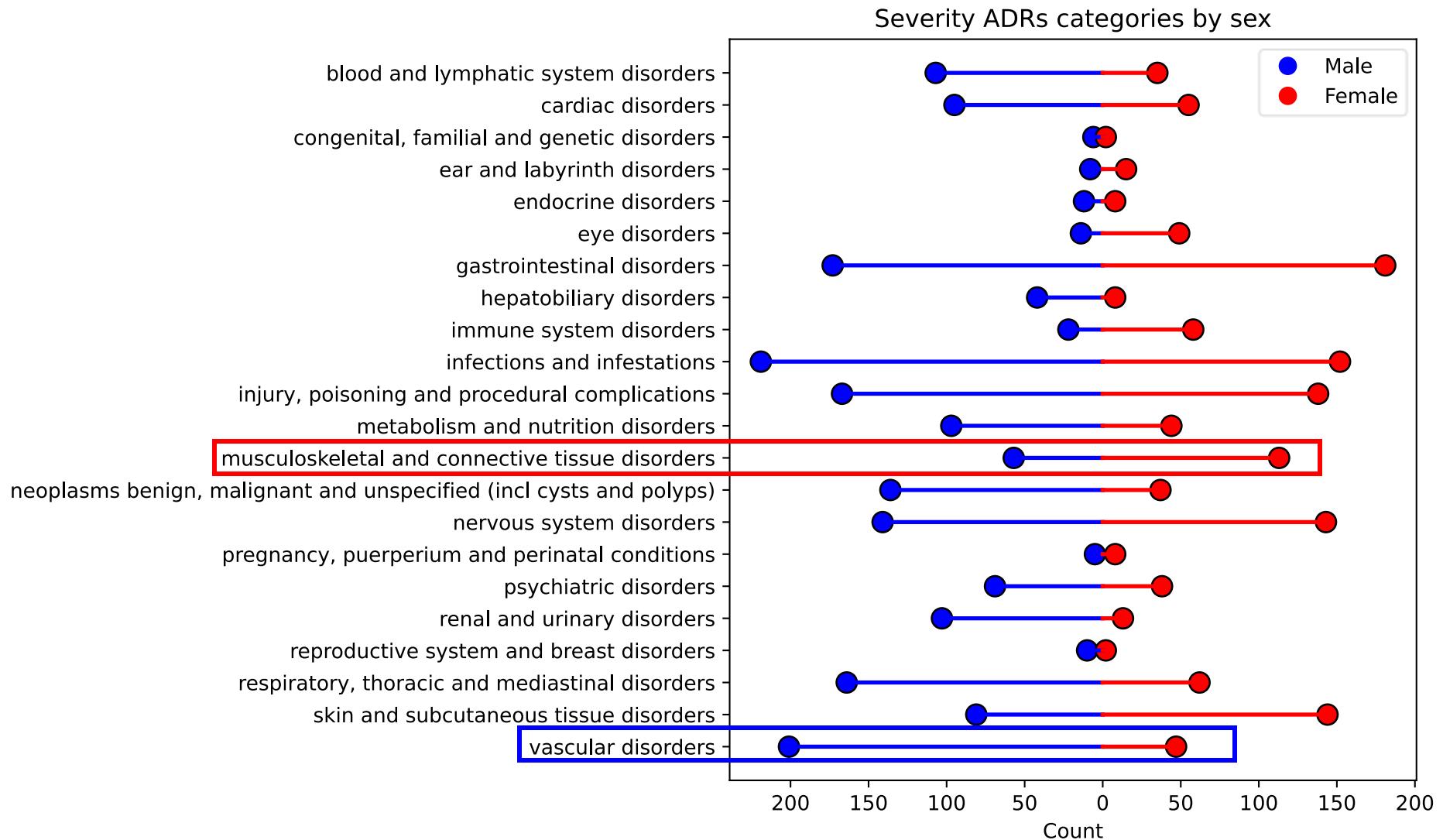
Sex-biased ADRs











Highlights

- Bias in Artificial Intelligence – Sex-bias in Biomedical domain.
- Drugs can induce changes in the gene expression of certain genes.
- X-linked and autosomal sex-biased genes present different enrichment in cancer drugs.
- The cancer drugs studied show evidence of being sex-biased drugs in particular tissues.
- Sex-specific ADRs are identified for the cancer drugs studied.

References

- Ayuso-Muñoz, A., et al. (2022). REDIRECTION: Generating drug repurposing hypotheses using link prediction with DISNET data. IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS 2022).
- Benson, T. (2010). Principles of Health Interoperability HL7 and SNOMED. Springer.
- Garcia del Valle, E.P. et al. (2021). A Meta-Path-Based Prediction Method for Disease Comorbidities. IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS 2021).
- Garcia del Valle, E.P. et al. (2021). DisMaNET: A network-based tool to cross map disease vocabularies. Computer Methods and Programs in Biomedicine.
- Goh, K.-I. et al. The human disease network. Proc. Natl. Acad. Sci. U. S. A. 104, 8685–8690 (2007).
- Hamilton, W.L., et al. (2018). Inductive Representation Learning on Large Graphs. arXiv:1706.02216

References

Harrer, S., Shah, P., Antony, B. & Hu, J. Artificial Intelligence for Clinical Trial Design. *Trends Pharmacol. Sci.* 40, 577–591 (2019)

Karagiannakos, S. (2021). Best Graph Neural Network architectures: GCN, GAT, MPNN and more. The AI Summer. URL: <https://theaisummer.com/gnn-architectures/>

Lagunes-García, G. et al. DISNET: a framework for extracting phenotypic disease information from public sources. *PeerJ* 8, e8580 (2020).

Leskovec, J., et al. CS224W: Machine Learning with Graphs. Stanford University. URL: <http://web.stanford.edu/class/cs224w/>

Low, Z. Y., Farouk, I. A. & Lal, S. K. Drug Repositioning: New Approaches and Future Prospects for Life-Debilitating Diseases and the COVID-19 Pandemic Outbreak. *Viruses* 12, 1058 (2020)

Otero-Carrasco, B. et al. (2022). Repositioning Drugs for Rare Diseases Based on Biological Features and Computational Approaches. *Healthcare*.

References

Parisi D, Adasme MF, Sveshnikova A, Bolz SN, Moreau Y, Schroeder M. Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs. *Comput Struct Biotechnol J* 2020;18:1043-55

Prieto Santamaría, L., Díaz Uzquiano, M., Ugarte Carro, E., Ortiz-Roldán, N., Pérez Gallardo, Y., Rodríguez-González, A. Integrating heterogeneous data to facilitate COVID-19 drug repurposing. *Drug Discovery Today*. 2021.

Prieto Santamaría, L., Ugarte Carro, E., Díaz Uzquiano, M., Menasalvas Ruiz, E., Pérez Gallardo, Y., Rodríguez-González, A. A data-driven methodology to-wards evaluating the potential of drug repurposing hypotheses. *Computational and Structural Biotechnology Journal*. 2021.

Prieto-Santamaría, L., et al. (2021). Classifying diseases by using biological features to identify potential nosological models. *Scientific Reports*.

References

Rodríguez-González, A., Alor-Hernández, G. (2012). An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies. *Computers in Biology and Medicine*.

Rodríguez-González, A., Costumero, R., Martinez-Romero, M., Wilkinson, M.D., Menasalvas-Ruiz, E. Extracting Diagnostic Knowledge from MedLine Plus: a Comparison between MetaMap and cTAKES Approaches. *Current Bioinformatics*. 2016.

Rodríguez-González, A., Martinez-Romero, M., Costumero, R., Wilkinson, M.D., Menasalvas-Ruiz, E., Diagnostic knowledge extraction from MedLine Plus: an application for infectious diseases. *9th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB 2015)*. 2015.

Schmidt, H. H. H. *The End of Medicine as We Know it -- and why Your Health Has a Future*. (Springer Nature, 2022).

Upton, F., DeGette, D. (2015). Can we find cures for 7,000 diseases?. CNN.com. URL: <http://edition.cnn.com/2015/01/13/opinion/upton-degette-cure-diseases/>

References

Zhang, X-M. et al. (2021). Graph Neural Networks and Their Current Applications in Bioinformatics. *Frontiers in Genetics*.

Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms-disease network. *Nat. Commun.* 5, 4212 (2014).

References

- T. T. Ashburn and K. B. Thor, “Drug repositioning: identifying and developing new uses for existing drugs,” *Nature Reviews Drug Discovery*, vol. 3, no. 8, pp. 673–683, Aug. 2004, doi: 10.1038/nrd1468.
- R. Artiñano-Muñoz, L. Prieto-Santamaría, A. Pérez-Pérez, and A. Rodríguez-González, “DRAGON: Drug Repurposing via Graph Neural Networks with Drug and Protein Embeddings as Features,” in 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), Jun. 2024.
- A. Ayuso-Muñoz et al., “REDIRECTION: Generating drug repurposing hypotheses using link prediction with DISNET data,” in 2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS), Jul. 2022, pp. 7–12. doi: 10.1109/CBMS55023.2022.00009.
- A. Ayuso-Muñoz, L. Prieto-Santamaría, E. U. Carro, E. Serrano, and A. Rodríguez-González, “Uncovering hidden therapeutic indications through drug repurposing with graph neural networks and heterogeneous data,” *Artificial Intelligence in Medicine*, p. 102687, Oct. 2023, doi: 10.1016/j.artmed.2023.102687.
- A. Ayuso-Muñoz, L. Prieto-Santamaría, A. Álvarez-Pérez, B. Otero-Carrasco, E. Serrano, and A. Rodríguez-González, “Enhancing Drug Repurposing on Graphs by Integrating Drug Molecular Structure as Feature,” in 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), Jun. 2023, pp. 192–197. doi: 10.1109/CBMS58004.2023.00215.
- A.-L. Barabási, “Network Medicine — From Obesity to the ‘Diseasome,’” *New England Journal of Medicine*, vol. 357, no. 4, pp. 404–407, Jul. 2007, doi: 10.1056/NEJMMe078114.

References

- K.-I. Goh and I.-G. Choi, “Exploring the human diseasesome: the human disease network,” *Brief Funct Genomics*, vol. 11, no. 6, pp. 533–542, Nov. 2012, doi: 10.1093/bfgp/els032.
- K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, “The human disease network,” *PNAS*, vol. 104, no. 21, pp. 8685–8690, May 2007, doi: 10.1073/pnas.0701361104.
- J. A. Greene and J. Loscalzo, “Putting the Patient Back Together - Social Medicine, Network Medicine, and the Limits of Reductionism,” *N. Engl. J. Med.*, vol. 377, no. 25, pp. 2493–2499, 21 2017, doi: 10.1056/NEJMms1706744.
- E. Guney, J. Menche, M. Vidal, and A.-L. Barabási, “Network-based in silico drug efficacy screening,” *Nat Commun*, vol. 7, no. 1, Art. no. 1, Feb. 2016, doi: 10.1038/ncomms10331.
- W. Hamilton, Graph Representation Learning. in *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Cham: Springer International Publishing, 2020. doi: 10.1007/978-3-031-01588-5.
- C. A. Hidalgo, N. Blumm, A.-L. Barabási, and N. A. Christakis, “A Dynamic Network Approach for the Study of Human Phenotypes,” *PLoS Comput Biol*, vol. 5, no. 4, Apr. 2009, doi: 10.1371/journal.pcbi.1000353.
- D.-S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. N. Oltvai, and A.-L. Barabási, “The implications of human metabolic network topology for disease comorbidity,” *Proc Natl Acad Sci U S A*, vol. 105, no. 29, pp. 9880–9885, Jul. 2008, doi: 10.1073/pnas.0802208105.

References

- J. Leskovec, 2021. Stanford CS224W: Graph Neural Networks. URL: <https://web.stanford.edu/class/cs224w/> (accessed 6.7.22).
- M. Marín Tercero, L. Prieto-Santamaría, and A. Rodríguez-González, “Exploring Drug Repurposing Opportunities for Schizophrenia: A Network Medicine Approach,” in 2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS), Jun. 2024.
- J. Menche et al., “Uncovering disease-disease relationships through the incomplete interactome,” *Science*, vol. 347, no. 6224, p. 1257601, Feb. 2015, doi: 10.1126/science.1257601.
- D. Parisi, M. F. Adasme, A. Sveshnikova, S. N. Bolz, Y. Moreau, and M. Schroeder, “Drug repositioning or target repositioning: A structural perspective of drug-target-indication relationship for available repurposed drugs,” *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1043–1055, Jan. 2020, doi: 10.1016/j.csbj.2020.04.004.
- L. Prieto Santamaría, E. Ugarte Carro, M. Díaz Uzquiano, E. Menasalvas Ruiz, Y. Pérez Gallardo, and A. Rodríguez-González, “A data-driven methodology towards evaluating the potential of drug repurposing hypotheses,” *Comput Struct Biotechnol J*, vol. 19, pp. 4559–4573, Aug. 2021, doi: 10.1016/j.csbj.2021.08.003.
- L. Prieto Santamaría, M. Díaz Uzquiano, E. Ugarte Carro, N. Ortiz-Roldán, Y. Pérez Gallardo, and A. Rodríguez-González, “Integrating heterogeneous data to facilitate COVID-19 drug repurposing,” *Drug Discovery Today*, vol. 27, no. 2, pp. 558–566, Feb. 2022, doi: 10.1016/j.drudis.2021.10.002.

References

- C. Ranjan, S. Ebrahimi, and K. Paynabar, “Sequence graph transform (SGT): a feature embedding function for sequence data mining,” *Data Min. Knowl. Discov.*, vol. 36, no. 2, pp. 668–708, Mar. 2022, doi: 10.1007/s10618-021-00813-0.
- repoDB: Drug Repositioning Database [WWW Document], n.d. URL <https://unmtid-shinyapps.net/shiny/repodb/> (accessed 6.3.22).
- X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, “Human symptoms–disease network,” *Nature Communications*, vol. 5, p. 4212, Jun. 2014, doi: 10.1038/ncomms5212.

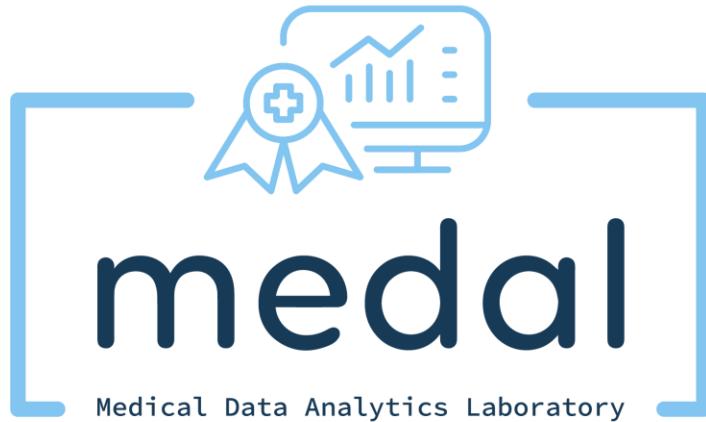


POLITÉCNICA

"Ingeniamos el futuro"



center for
biomedical
technology



<https://medal.ctb.upm.es/>
 @MEDAL_CTB

Disease understanding: Dealing with complex and unstructured big data in biomedical domain