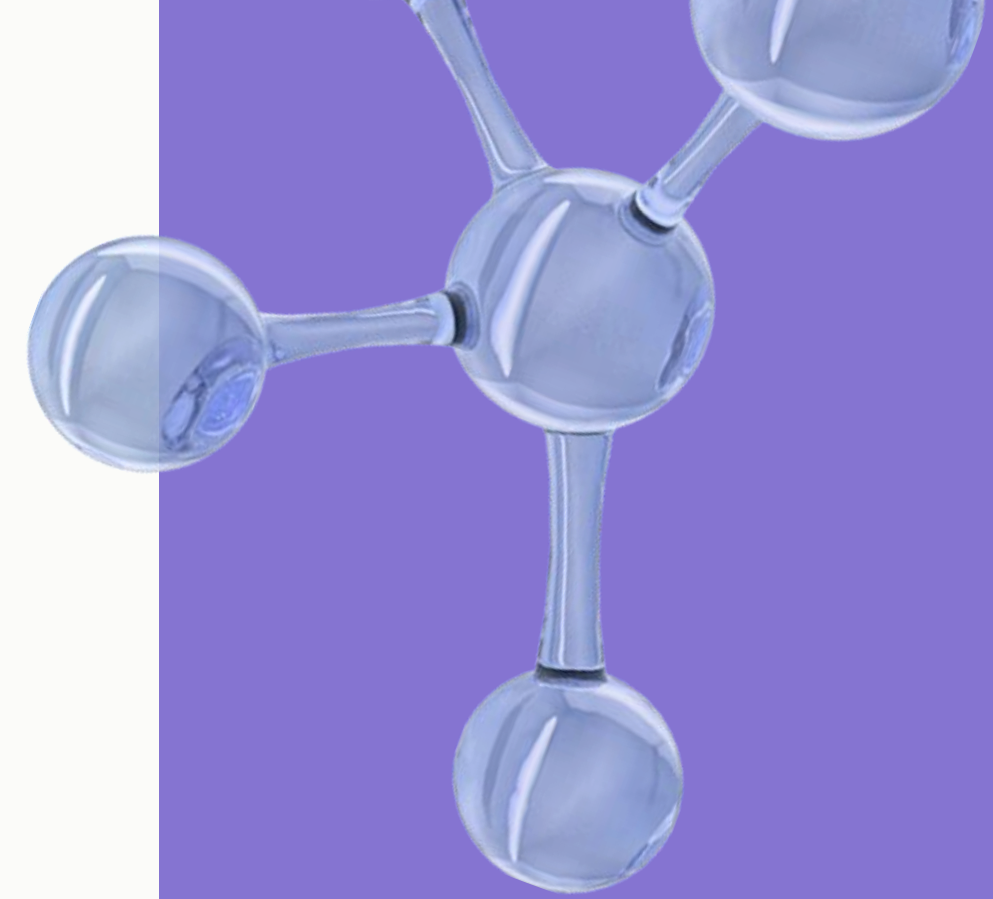# DATA ANALYTICS

FOR NETWORK BIOLOGY

Data and Methodologies in Network Biology

First steps to construct the biomedical knowledge graph:

- *bulding the dataframe and getting some insights*

47029	Symptom::D065906	Hyperlactatemia	Symptom
47030	Symptom::D066190	Allesthesia	Symptom
		0	

source	metaedge	target
Gene::9021	GpBP	Biological Process::GO:0071357

- *performing some preliminary analysis*

```
> Findings:
# the nodes data present 47031 different nodes
# symptoms' names (other than IDs) are usable as univocal references
# diseases' names (other than IDs) are usable as univocal references
# the edges data present 2250197 edges, some with equal source-target
      pair but different link type
```

- *constructing the MultiDiGraph*

Analyzing our graph

Computing some basic indicators:
- total number of entities for each unique kind;
- total number of each unique edge type;
- average in-degree and out-degree for each type of node.

Analyzing diseases-symptoms correlation:
- Analyzing the degrees of each disease-type node
- The disease-symptom sub-graph is extracted and the average number of symptoms related to a disease is computed.

	kind	total_nodes
0	Anatomy	402
1	Biological Process	11381
2	Cellular Component	1391
3	Compound	1552
4	Disease	137
5	Gene	20945
6	Molecular Function	2884
7	Pathway	1822

	metaedge	total_edges
0	AdG	102240
1	AeG	526407
2	AuG	97848
3	CbG	11571
4	CcSE	138944
5	CdG	21102
6	CpD	390
7	CrC	6486
8	CtD	755
9	CuG	18756
8		12623
9		7623
10		3602
		3357
		543
		7731
		61690
		147164
		559504
		73566
		97222
		84372
		265672
		1029

	kind	averageInDegree	averageOutDegree
0	Anatomy	8.960199	1807.201493
1	Biological Process	49.161234	0.000000
2	Cellular Component	52.887132	0.000000
3	Compound	4.842139	127.579897
4	Disease	12.321168	258.970803
5	Gene	61.132824	61.551206
6	Molecular Function	33.710818	0.000000
7	Pathway	46.307355	0.000000
8	Pharmacologic Class	0.000000	2.982609
9	Side Effect	24.231601	0.000000
10	Symptom	7.664384	0.000000

Disease-type nodes with higest out-degrees: _____

	id	name	out_degree
14784	Disease::DOID:1612	breast cancer	1115
14804	Disease::DOID:2531	hematologic cancer	1076
14808	Disease::DOID:2986	IgA glomerulonephritis	966
14793	Disease::DOID:1909	melanoma	931
14842	Disease::DOID:7148	rheumatoid arthritis	868

Disease-type nodes with higest in-degrees: _____

	id	name	in_degree
14739	Disease::DOID:10763	hypertension	73
14804	Disease::DOID:2531	hematologic cancer	53
14784	Disease::DOID:1612	breast cancer	44
14807	Disease::DOID:2841	asthma	41
14818	Disease::DOID:3393	coronary artery disease	40

```
Disease-symtpom sub-graph info:
Type of edges: ['DrD' 'DpS']

Average number of symptoms related
to each disease: 24.503649635036496
```

Jaccard similarity

Computing diseases' and symptoms' similarity

Similarity is computed at first with the Jaccard formula and then through our function (implemented for multiDiGraphs) for three couples of nodes of the same type:

1. *type-1 and type-2 diabetes mellitus (disease nodes);*
2. *eye pain and blindness (symptom nodes);*
3. *PTGS2 and JUN (gene nodes).*

Our jaccard similarity VS normal Jaccard formula

```
Jaccard similarity 1: 0.13267326732673268
Jaccard similarity 2: 0.35
Jaccard similarity 3: 0.10552763819095477
```
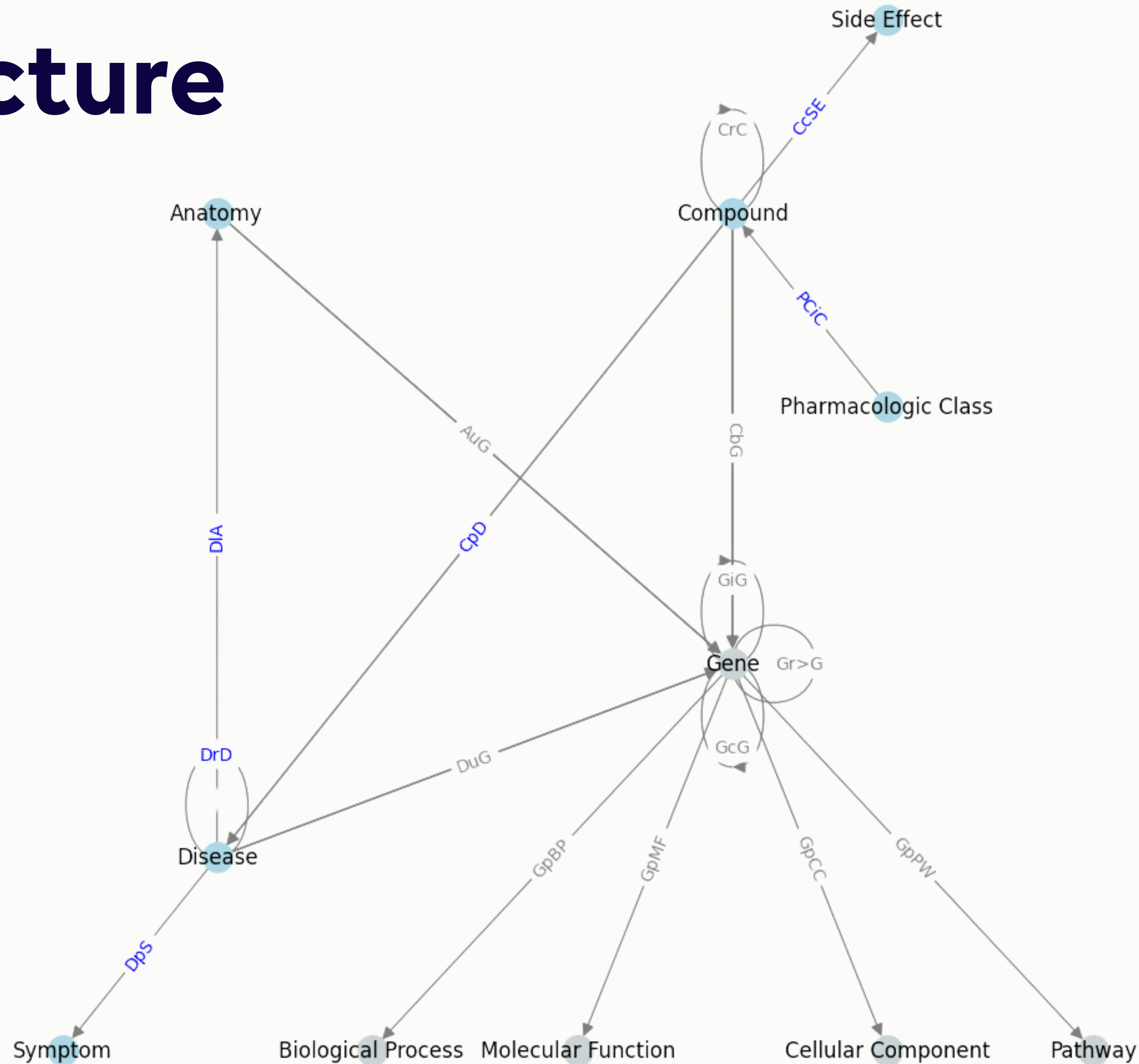
*Using the **multiDiGraph-based similarity formula** on two gene-type nodes, a different value is obtained (presence of multi-directional and multiple edges)*

```
Using the similarity function for multiDiGraphs:
Adapted Jaccard similarity 1: 0.13267326732673268
Adapted Jaccard similarity 2: 0.35
Adapted Jaccard similarity 3: 0.09482758620689655
```
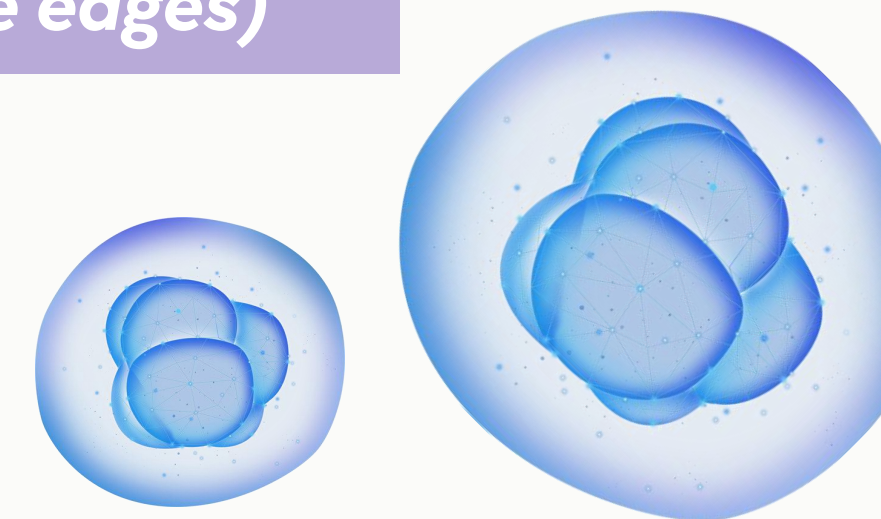
Graph structure

Hypoteses

- Disease-to-symptom edges mean that the particular disease necessarily implies the appeareance of all the connected symptoms (while not viceversa), thus the main symptoms are considered to be the ones triggered by a lower number of different diseases

- Compound-to-disease edges mean that the particular compound perfectly treats (at the same exact extent as all the other compounds connected to that disease node) the specific illness

- Compound-to-side-effect edges mean that the particular compound eventually casues the appearance of that specific side effect with an undefined constant probability, thus (being all side effects comparable by severeness) the best compound for a disease is determinable as the connected one with fewer possible side effects

- Disease-to-anatomy edges mean that the particular disease could affect the specific anatomy, thus the main anatomies are considered to be the ones affectable by a lower number of different diseases

- Disease-to-diseases edges connect pairs of diseases that could probably appear together in a patient, thus (if no univocal diagnosis is obtainable based on observed symptoms olny) eventually multiple possible diagnosis are ordered based on their eventual connections with already diagnosed diseases

patientRand function

Takes as input G, s, irrs, d where:
- s is the number of relevant symptoms to randomize
- irrs the number of irrelevant symptoms to randomize
- d is the number of already known diseases to randomize

Outputs:
- patient: contains randomized lists of relevant and irrelevant symptoms and known diseases.
- possibleDiseases: all potential diagnoses based on relevant symptoms

Key Points:
- ensures relevant symptoms are all related to at least one disease.
- ensures irrelevant symptoms do not relate to any common diseases associated with relevant symptoms

chatDOC function

Function to analyze a hypothetical patient situation starting from suffered symptoms and eventually already diagnosed diseases, assuming that:

- *The patient comes to chatDOC recognizing of one and only one main symptom*
- *The patient eventually also report (in order of importance to him/her) other secondary symptoms and already diagnosed diseases (which may be uncorrelated with the main one)*
- *The patient has one and only one disease to be diagnosed, if no symptoms are recognized not correlated to an already diagnosed disease, that diagnosis will be just confirmed.*

OUTPUT:

A univocal diagnosis of the responsible disease, indicating relevant revealing symptoms and known diseases and the diagnosed disease's additional verifiable symptoms, affected anatomies and (eventually) associated compounds for treatment
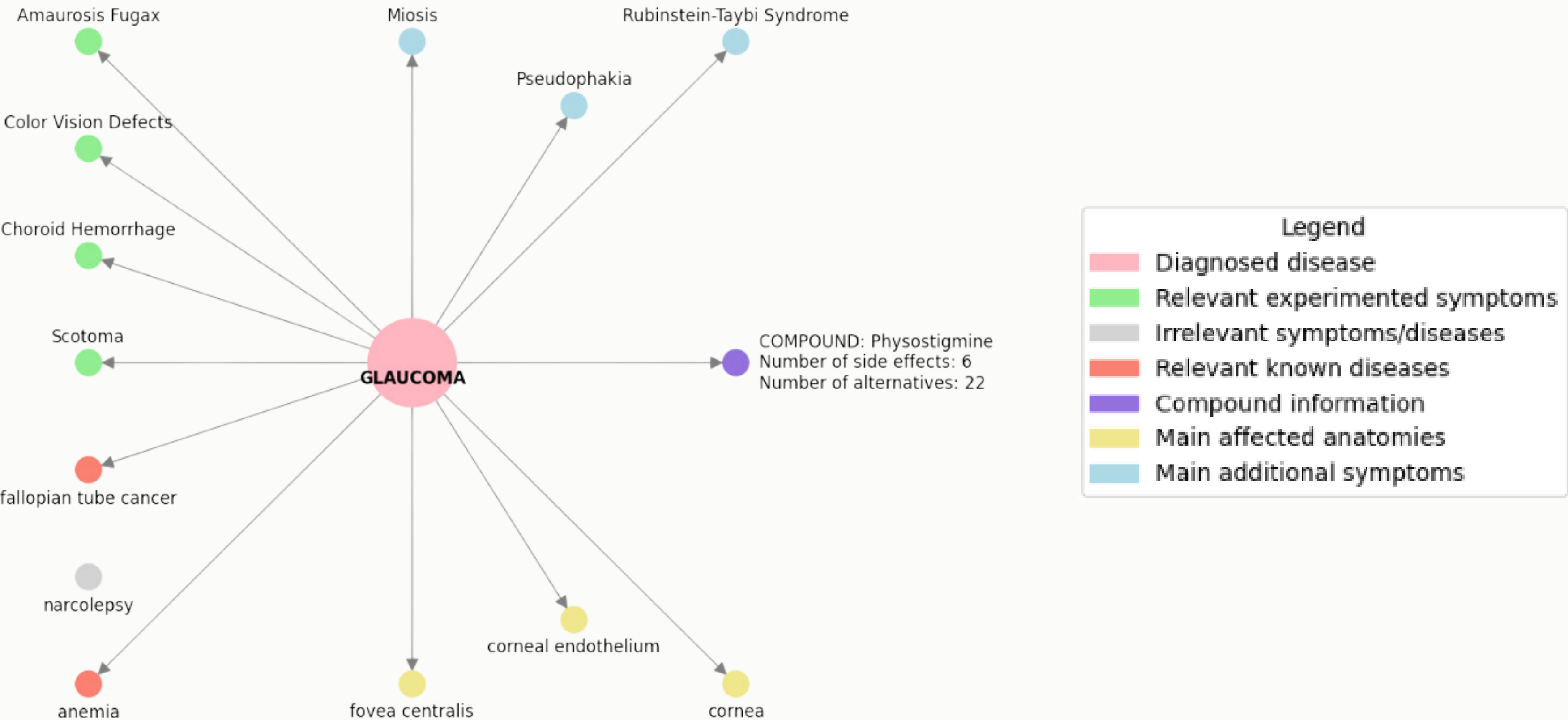
OR

When no univocal diagnosis seems possible, a list of possible diseases correlated with the suffered symptoms and known diseases
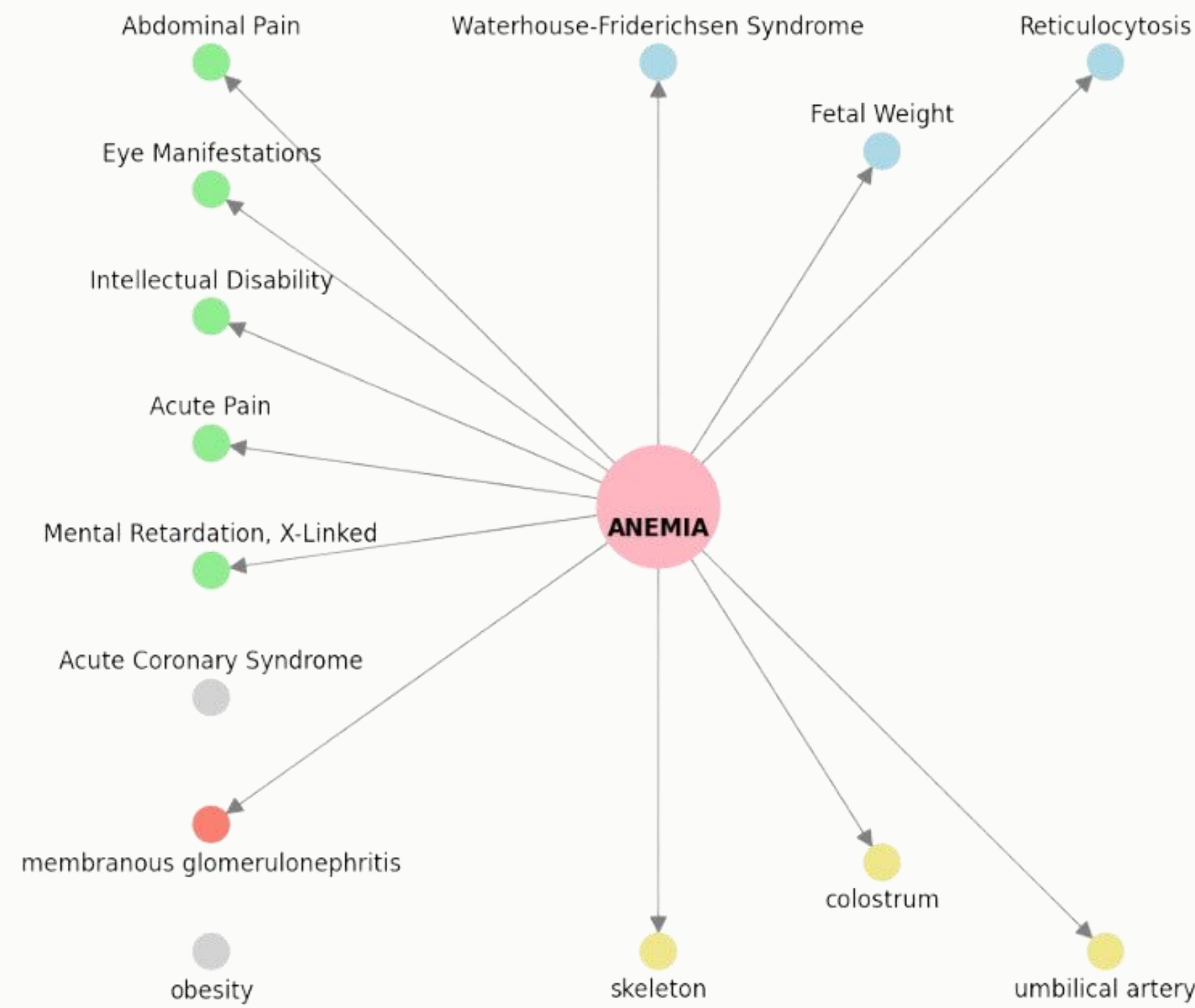
Disease univocally diagnosed

chatDOC's diagnostic analysis

Amaurosis Fugax
Miosis
Rubinstein-Taybi Syndrome
Pseudophakia
Color Vision Defects
Choroid Hemorrhage
Scotoma

GLAUCOMA

COMPOUND: Physostigmine
Number of side effects: 6
Number of alternatives: 22

fallopian tube cancer
narcolepsy
corneal endothelium
anemia
fovea centralis
cornea

Legend
Diagnosed disease
Relevant experimented symptoms
Irrelevant symptoms/diseases
Relevant known diseases
Compound information
Main affected anatomies
Main additional symptoms

Disease univocally diagnosed, compound info not available
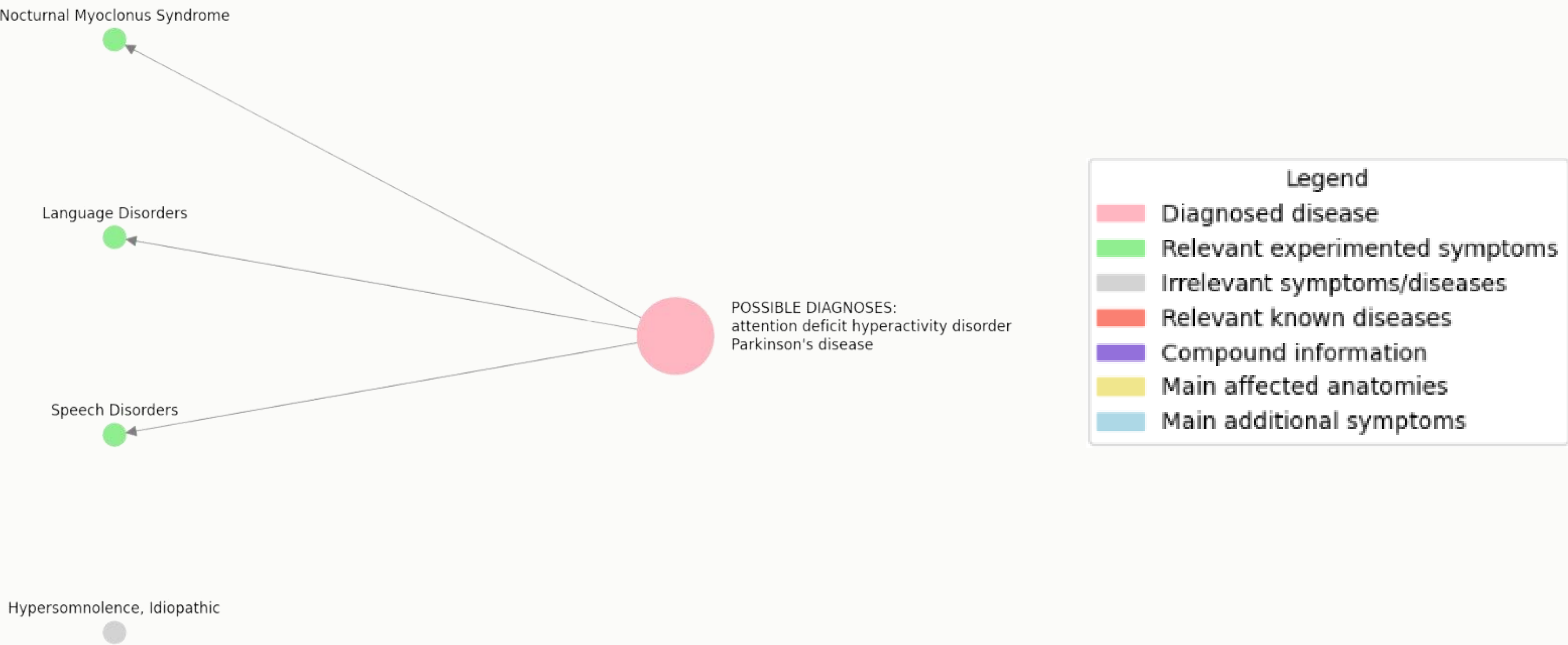

chatDOC's diagnostic analysis
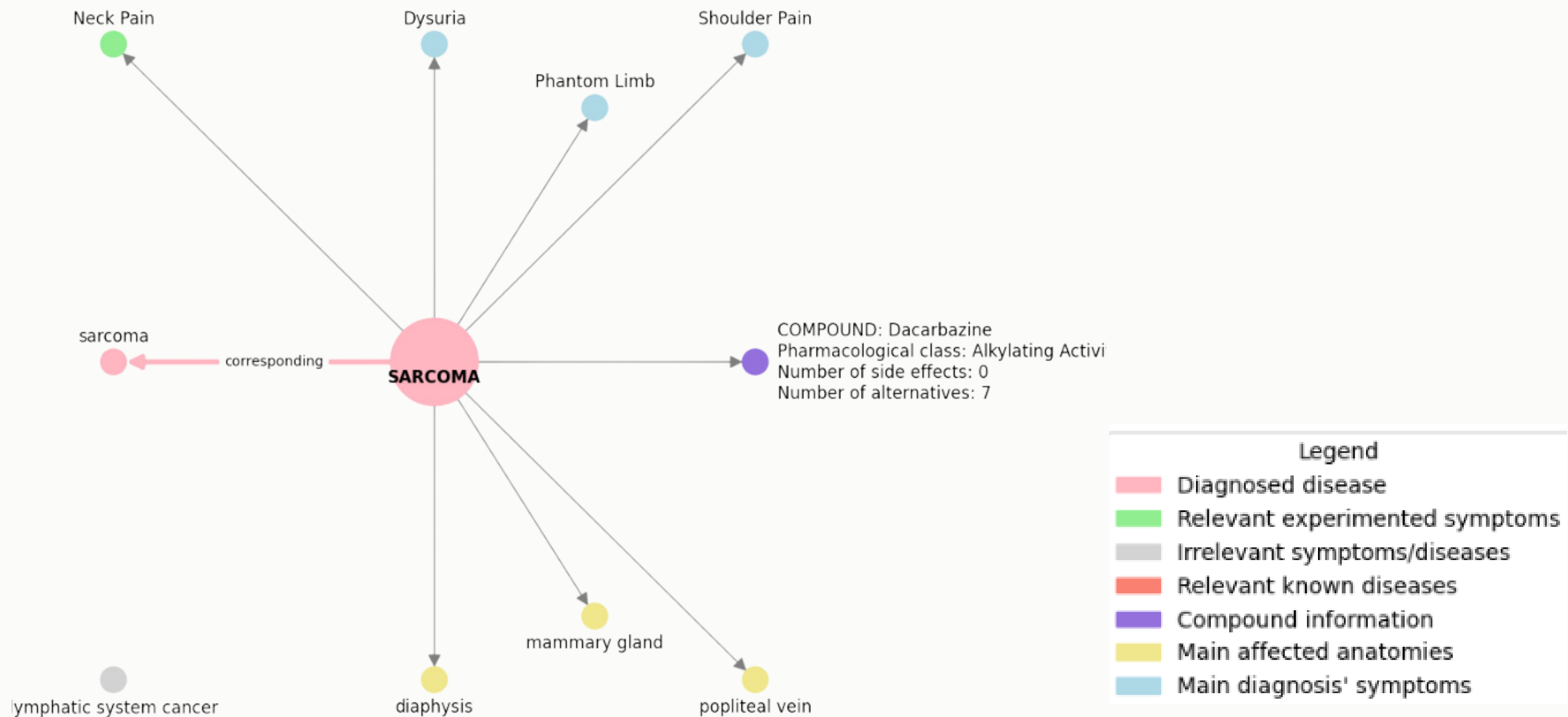
Impossible univocal diagnosis multiple diseases possible

Diagnosis matches known disease

chatDOC's diagnostic analysis

Neck Pain

Dysuria

Shoulder Pain

Phantom Limb

sarcoma

← corresponding ← SARCOMA

COMPOUND: Dacarbazine
Pharmacological class: Alkylating Activi
Number of side effects: 0
Number of alternatives: 7

lymphatic system cancer

diaphysis

mammary gland

popliteal vein

Legend
- Diagnosed disease
- Relevant experimented symptoms
- Irrelevant symptoms/diseases
- Relevant known diseases
- Compound information
- Main affected anatomies
- Main diagnosis' symptoms

# chatDOC interaction

```
Hi! How can I help you today? Which is your main symptom?        eye
There are 3 symptoms related to your description:
1 - Eye Pain
2 - Eye Hemorrhage
3 - Eye Manifestations
Please select the number of the main symptom you are actually experimenting:    4
Invalid slection, please input the number corresponding to the symptom of interest as from the above list!
Please select the number of the main symptom you are actually experimenting:    1
Your main symptom is then: Eye Pain !
Do you already recognize specific other symptoms?        [Y/N]   yes
Please specify all already recognized secondary symptoms ordered by relevance and separated by commas:  head, Symptom::D058447, diabete
There are 2 symptoms related to your description 'head':
1 - Coma, Post-Head Injury
2 - Headache
Please select the number of the secondary symptom you are actually experimenting:        2
The following inputs are not recognizable as symptoms: diabete !
Have you already been diagnosed with other diseases?    [Y/N]   dont know
Input not recognized, please just answer yes or no to the following question!
Have you already been diagnosed with other diseases?    [Y/N]   y
Please specify all already diagnosed diseases ordered by relevance and separated by commas:     diabete, autistic
There are 3 diseases related to your description 'diabete':
1 - gestational diabetes
2 - type 1 diabetes mellitus
3 - type 2 diabetes mellitus
Please select the number of the disease you have actually beign diagnosed with: 2
The relevant symptoms detected are: Eye Pain; Headache !
The relevant diseases detected are: type 1 diabetes mellitus; autistic disorder !
You may have this disease: germ cell cancer !
Therefore, you may also recognize the following main symptoms: Ophthalmoplegic Migraine; Necrolytic Migratory Erythema; Gerstmann Syndrome !
The main affected anatomies could be: macula; geniculate ganglion; cochlear nucleus !
There are 7 compounds that may treat the disease!
The suggested compound is: Vinblastine (with 59 side effects)!
The relevant pharmacologic class is: Vinca Alkaloids !
```

If no patient profile is given as input, the function activates an I/O-mode to progressively acquire data chatting with the user.

Demonstration >
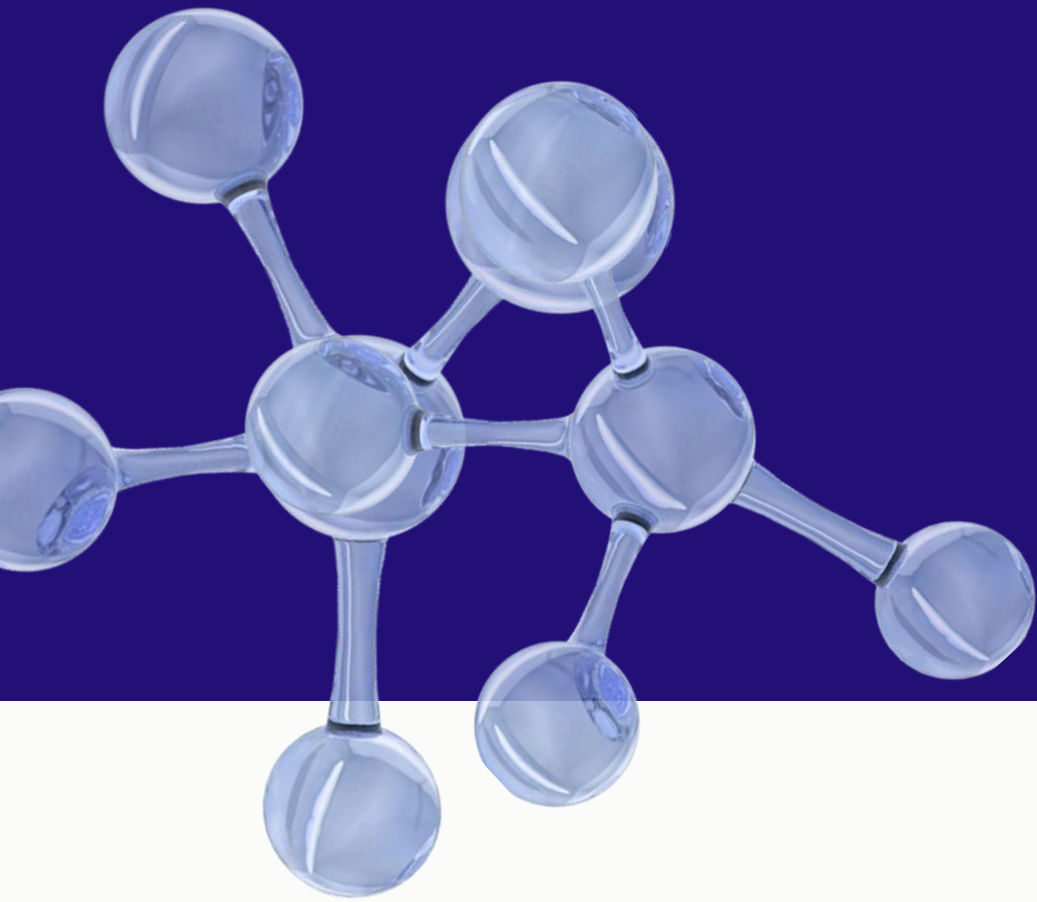
Thank you for your attention!

De Marinis Federico

Di Mario Eleonora   GROUP 1

Spadaccia Matteo