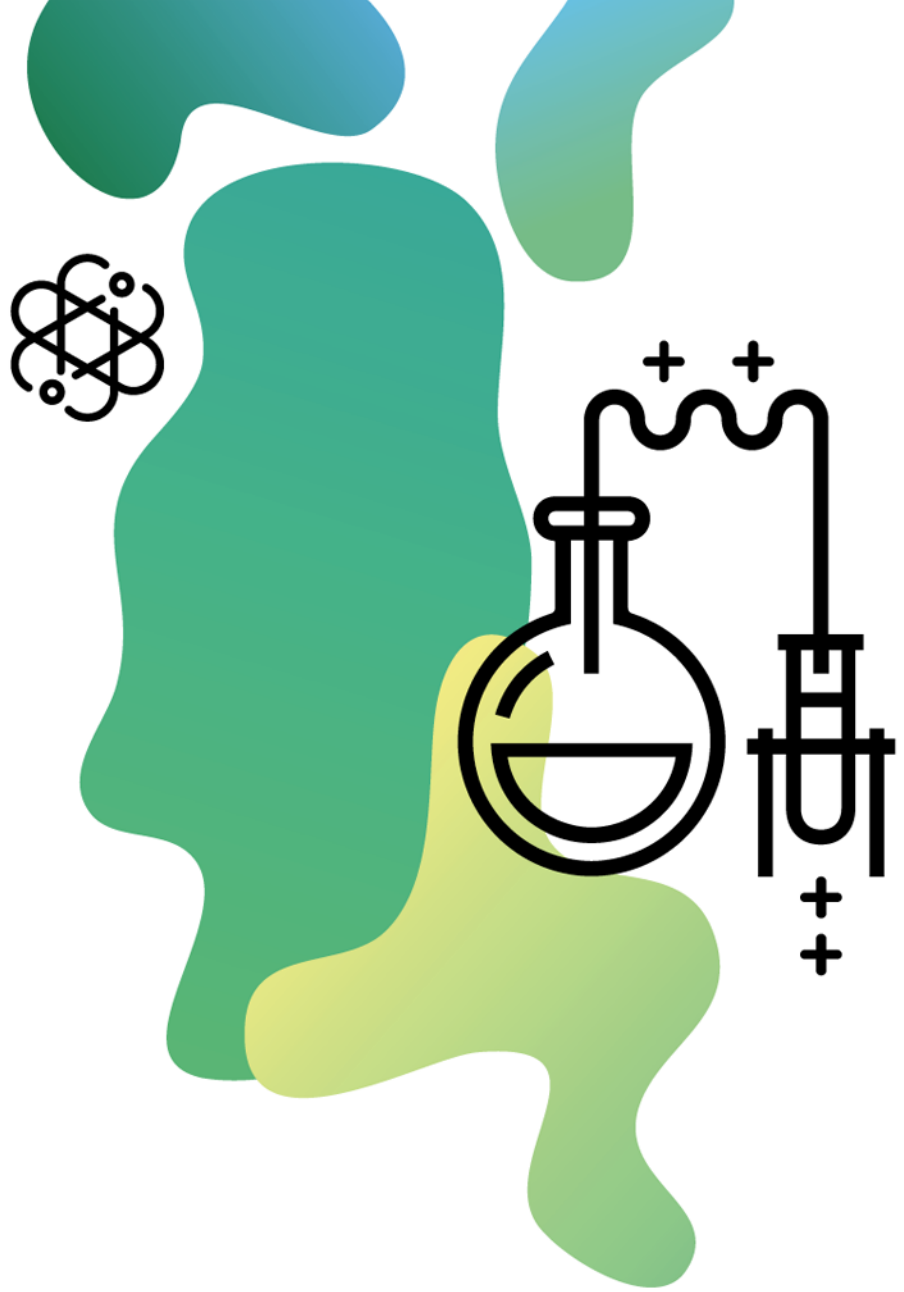
idorsia

Targeting of the disease related proteome by small molecules



ICCS
2018
Noordwijkerhout

Modest v. Korff

Outline

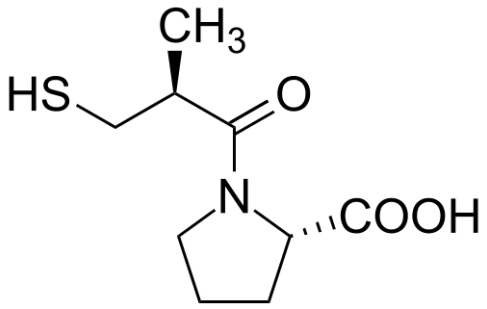
- Introduction
- Data sources
  - Diseases
  - Proteins
  - Compounds
- Data relations
  - Diseases – target proteins: relevance estimator:
  - Protein – protein: BLAST similarity
  - Compound –compound: Flexophore descriptor similarity
- Visualization: rubber bond map
- Summary & conclusions

What is pharmaceutical industry doing?

Deliver medicine

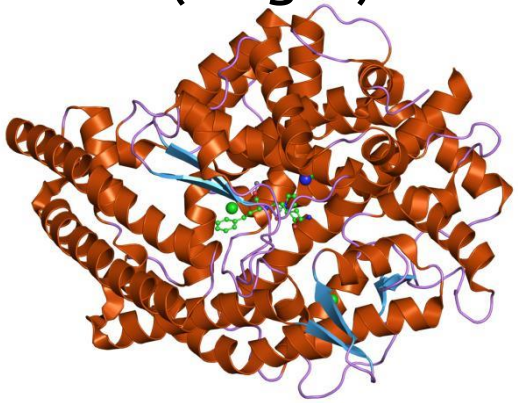
Switch back to normal condition (blood pressure)

Drug



Molecule
Captopril

Protein
(target)



Angiotensin converting enzyme

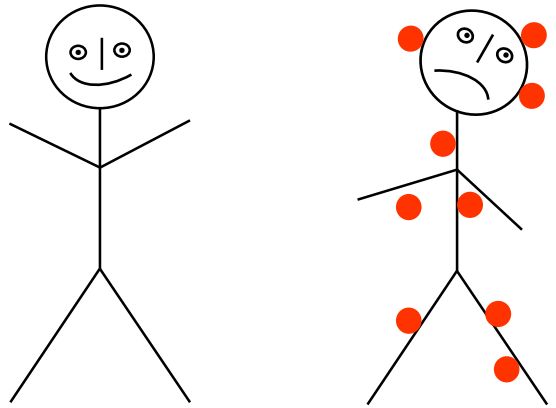
Disease

Hypertension

Diseases

A condition of the living animal or plant body or of one of its parts that impairs normal functioning and is typically manifested by distinguishing signs and symptoms.

- 4500 diseases indexed by MeSH
- Conditions to start drug discovery
  - Severe
  - No sufficient treatment available



MeSH: medical subject headings (Thesaurus)
Used for indexing MEDLINE
MEDLINE contains 27 million publication records from life sciences

Proteins in drug discovery

Drug targets are proteins (almost all of them)

Protein as a switch, changing physiological condition

- Enzymes
  - Hydrolases (ACE, 1956)
  - Kinases (cancer therapy, ermerging drug targets)
- Receptors
  - G protein coupled receptors (70 % of all drugs)
- Ion channels
  - Emerging drug targets
  - Anti targets (hERG)

idorsia

UniProt

Protein database

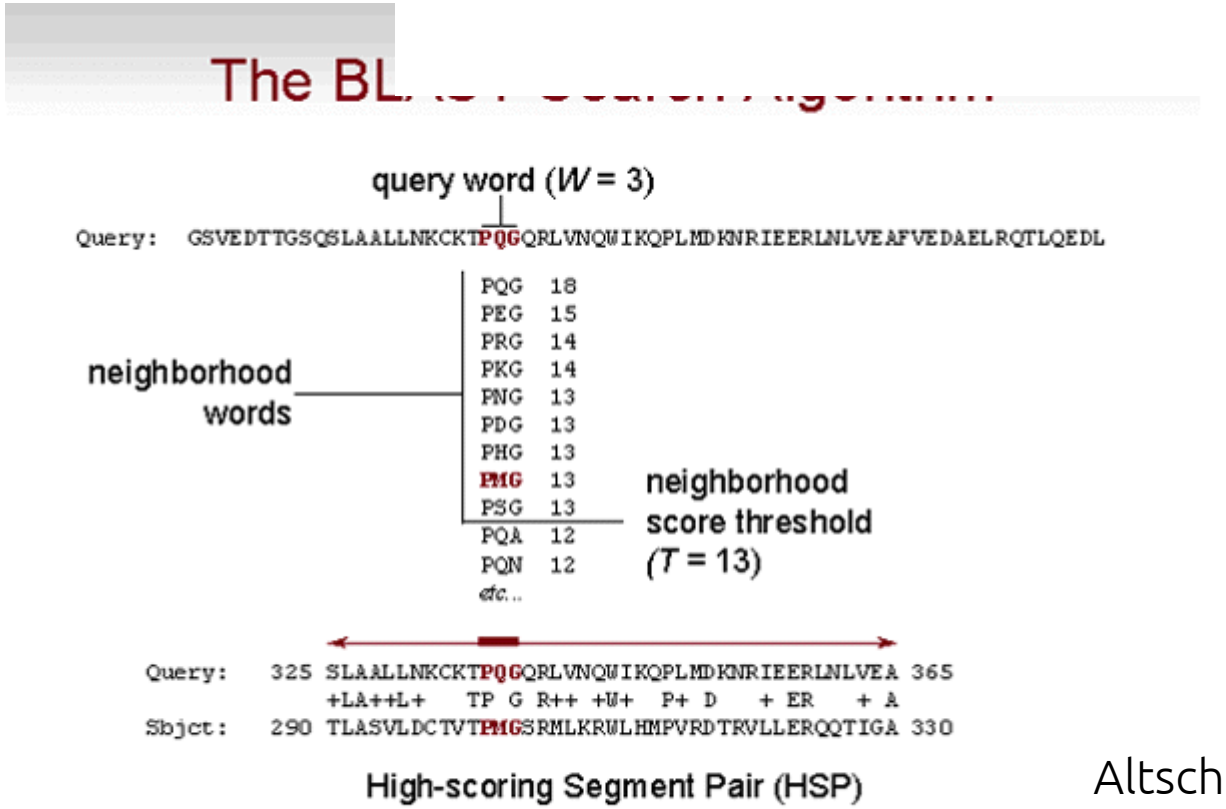
- Nucleotide sequences
- Gene- and protein names
- Amino acid sequences
- Annotation data
- Protein sequence comparison (BLAST)

UniProt Consortium. "UniProt: the universal protein knowledgebase." *Nucleic acids research* 45.D1 (2016): D158-D169.

BLAST

Basic Local Alignment Search Tool

Protein similarity by amino acid sequence alignment



Sophisticated form of character comparison

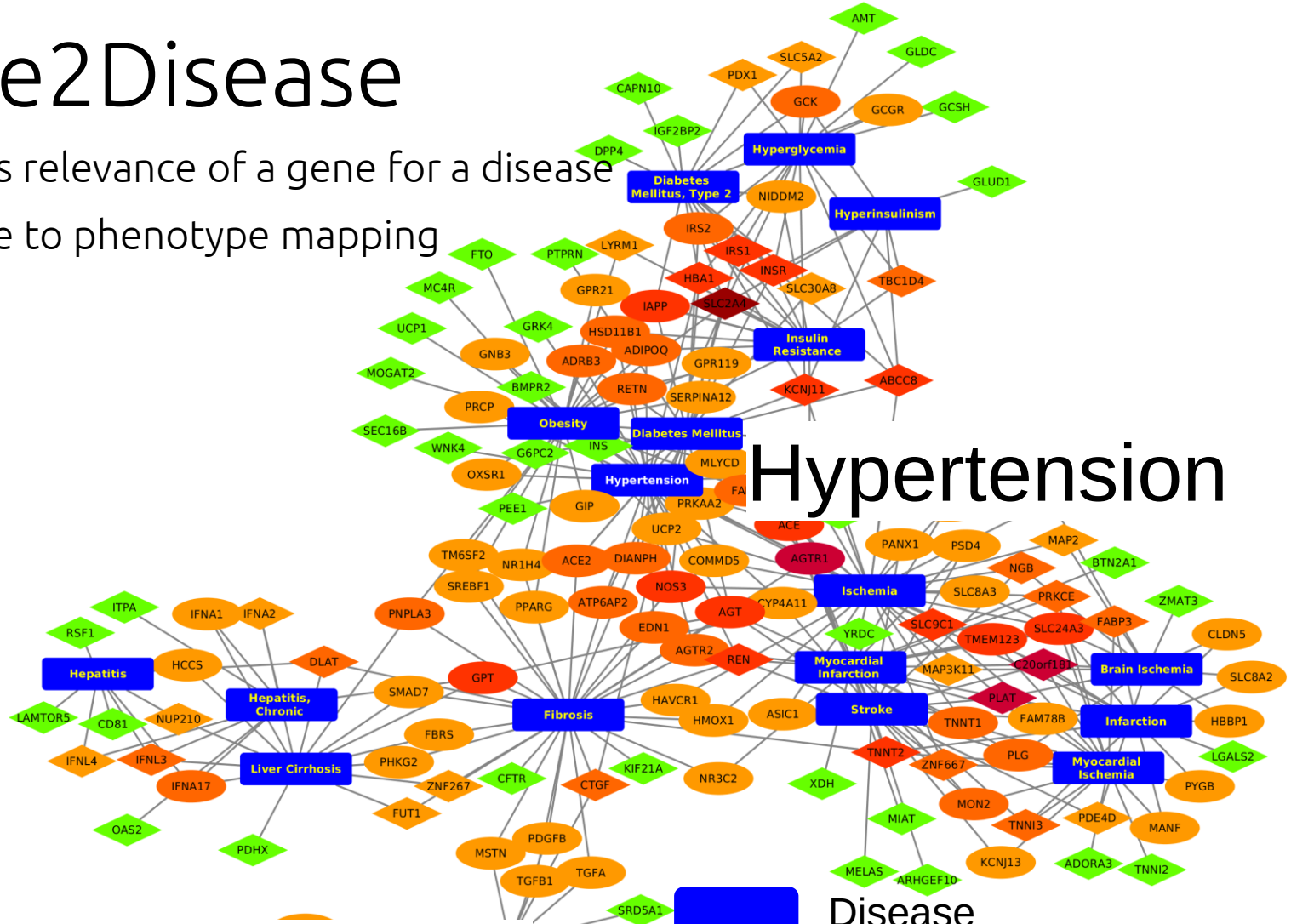
Altschul, Stephen F., et al. "Basic local alignment search tool." *Journal of molecular biology* 215.3 (1990): 403-410.

idorsia

Gene2Disease

Estimates relevance of a gene for a disease

Genotype to phenotype mapping



How specific is a gene for a disease?

Disease

Gene from the top-5 list by relevance

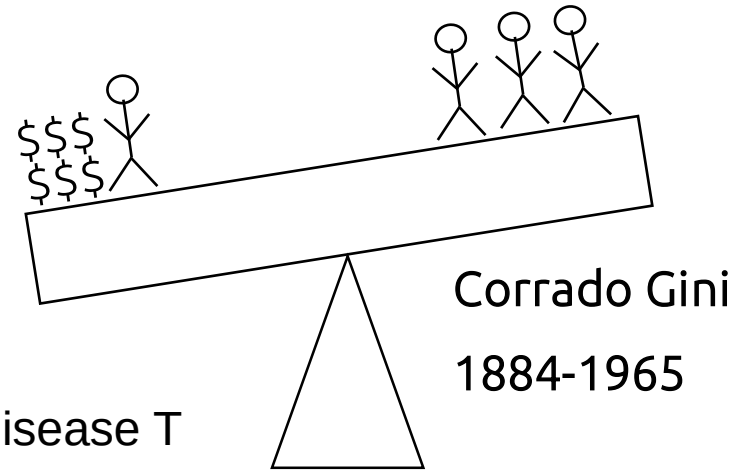
Gene with similar relevance for ≥2 diseases

Number of gene-disease connections:

1 2 3 >3

idorsia

Gini, ranks and the relevance estimator



Corrado Gini

1884-1965

For a disease T

$$G_T = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n\sum_{i=1}^{n}x_i}$$

$$f_{A,T} = \frac{\text{\# publications gene } A_T}{\text{\# publications for all genes}_T}$$

$$re_{A,T} = G_T\, f_{A,T}\, R_{rel,A,T}$$

$R_{rel,A,T}$= relative rank

PubCount	Rank	Relative Rank
45	1	0.83
20	2	0.67
10	3	0.5
10	3	0.5
6	4	0.33
5	5	0.17
1	6	0

von Korff, Modest, Tobias Fink, and Thomas Sander. "A new relevance estimator for the compilation and visualization of disease patterns and potential drug targets" *Pacific Symposium on Biocomputing 2017*. 2017.

ChEMBL database

Structures, biological activities, target protein identifiers

Version 23, quality 9 (higest)

- Molecule structures 900’000 unique
- Biological activity values 4 million
- Protein accession identifiers 4’000

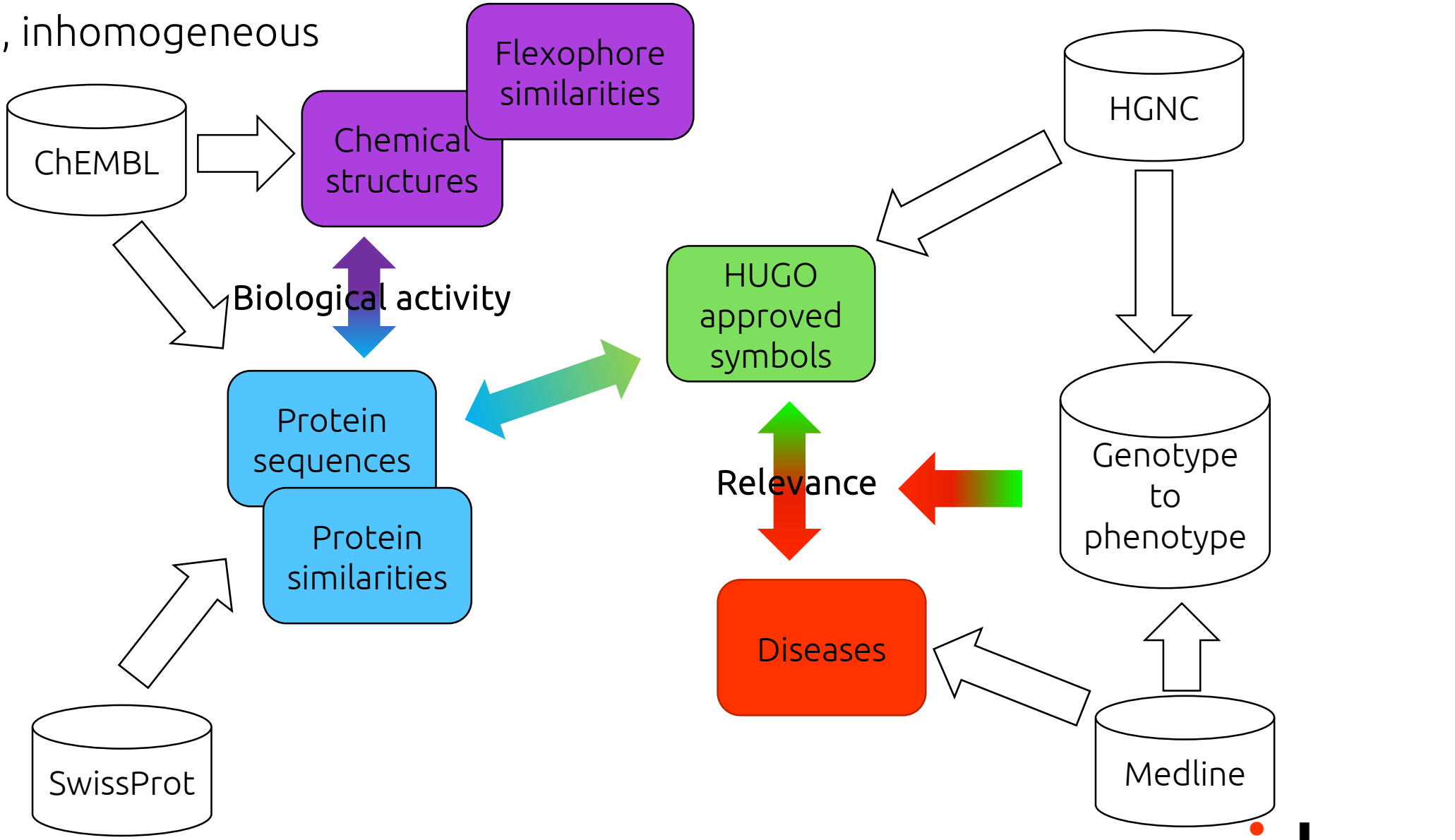
Most tested proteins
POLI (DNA polymerase iota) 116’761
ATAD5 (ATPase family AAA domain-containing protein 5) 122’498
GMNN (Geminin, DNA replication inhibitor)127’916

Gaulton, Anna, et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic acids research* 40.D1 (2011): D1100-D1107.

idorsia

Data relations

Incomplete, inhomogeneous



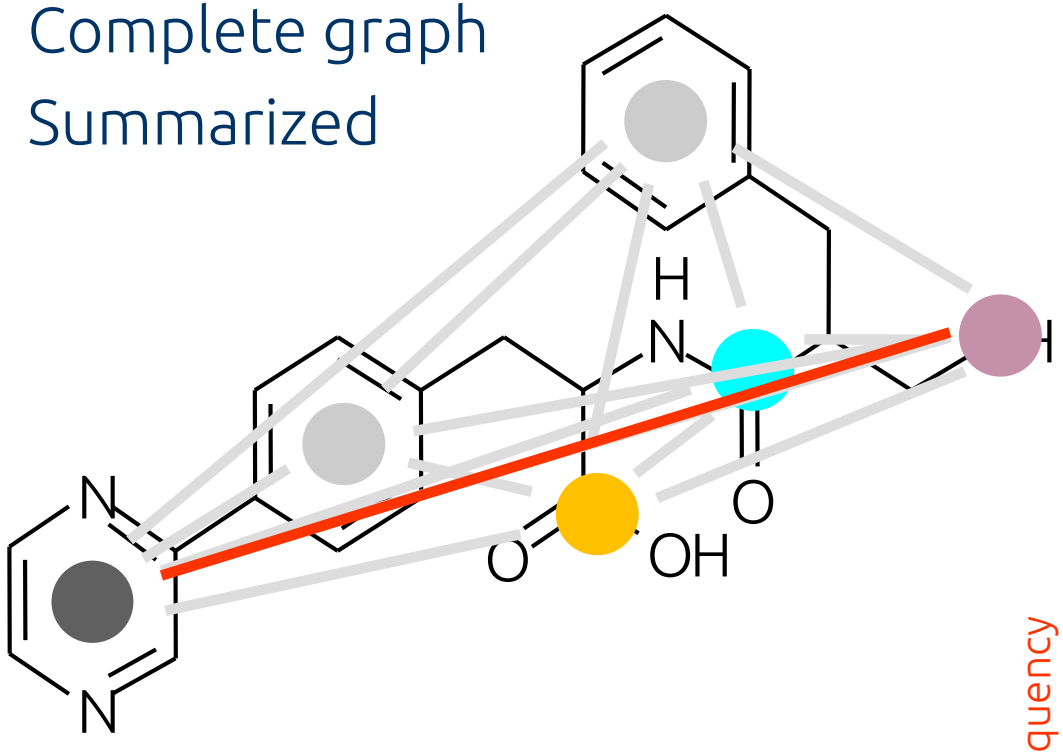
idorsia

Compound – compound similarity

Pharmacophore descriptor: Flexophore

Complete graph

Summarized

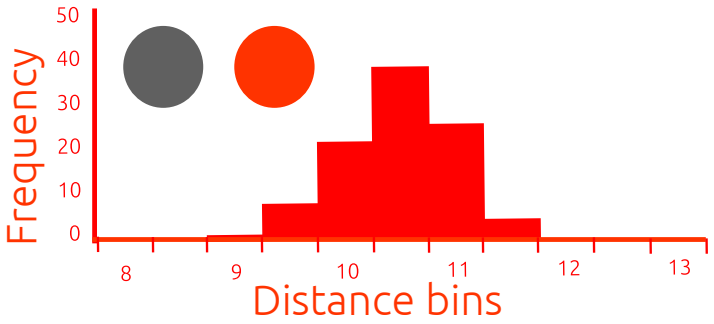


Nodes:

MM2 atom types

Edges:

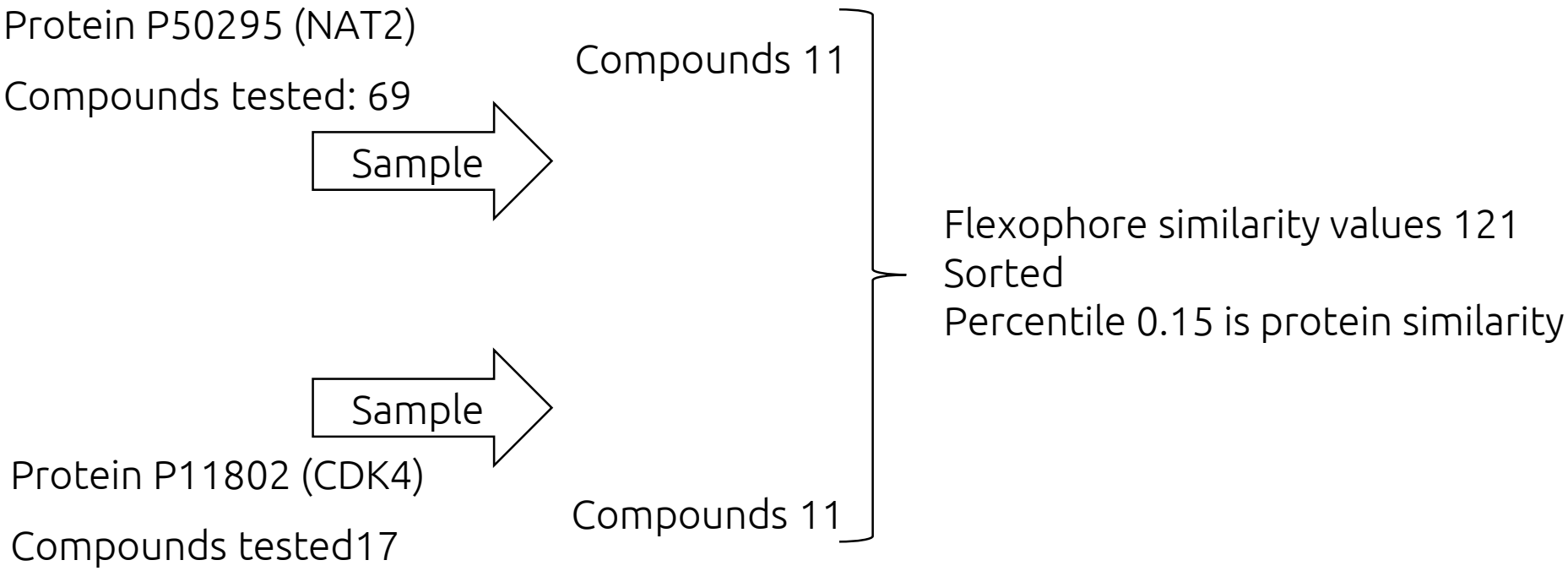
Distance histograms



von Korff, Modest, Joel Freyss, and Thomas Sander. "Flexophore, a new versatile 3D pharmacophore descriptor that considers molecular flexibility." *Journal of chemical information and modeling* 48.4 (2008): 797-810.

idorsia

Protein similarity by ligand similarity



Mixed model

Protein similarity

- HUGO approved symbols 42’000
- Protein identifiers 200’000
- Similarity values 7’200’000 from BLAST

Ligand similarity

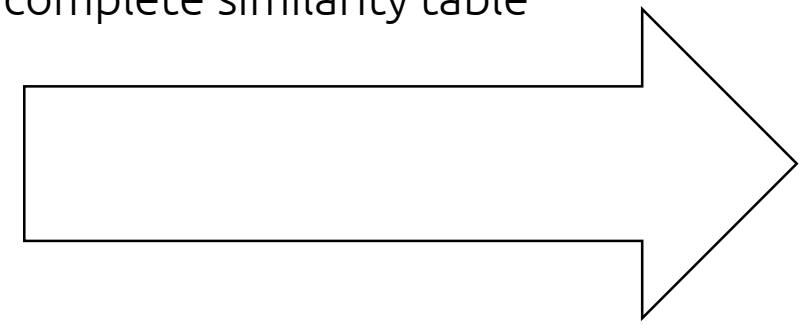
- ChEMBL structure records 890’000
- Activity values 4’600’00
- Protein identifiers 4’200
- Similarity values 8’700’000 from Flexophore

Scaled
Equal identifier pairs replaced by the more similar one
Table 15’000’000 similarity values

idorsia

How to make a map?

- Side conditions
  - Many objects (200’000)
  - Incomplete similarity table



2D Rubber Bond Scaling

Force field like arrangement in 2D space

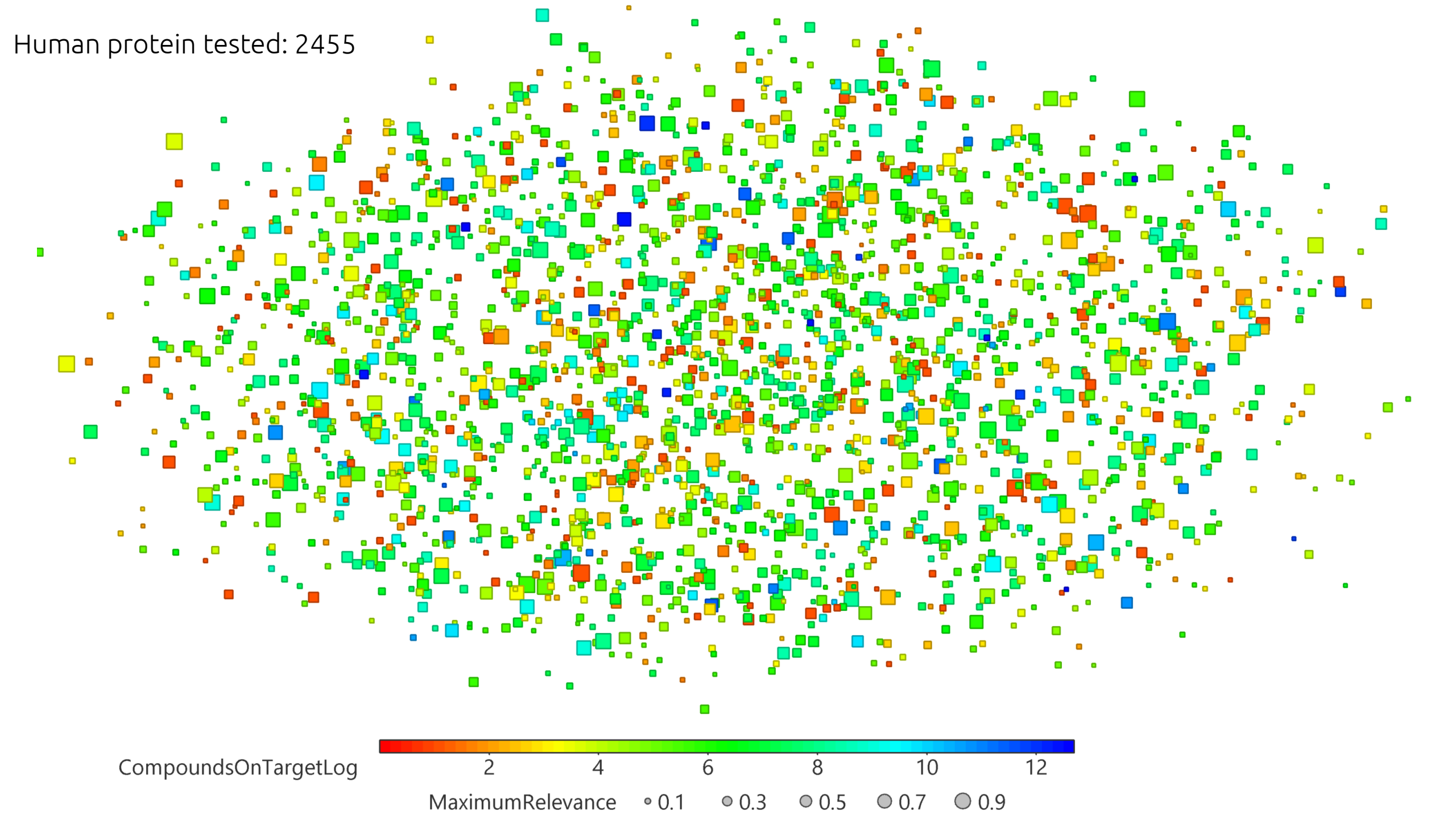
Sander, Thomas, et al. "DataWarrior: an open-source program for chemistry aware data visualization and analysis." *Journal of chemical information and modeling* 55.2 (2015): 460-473.

idorsia

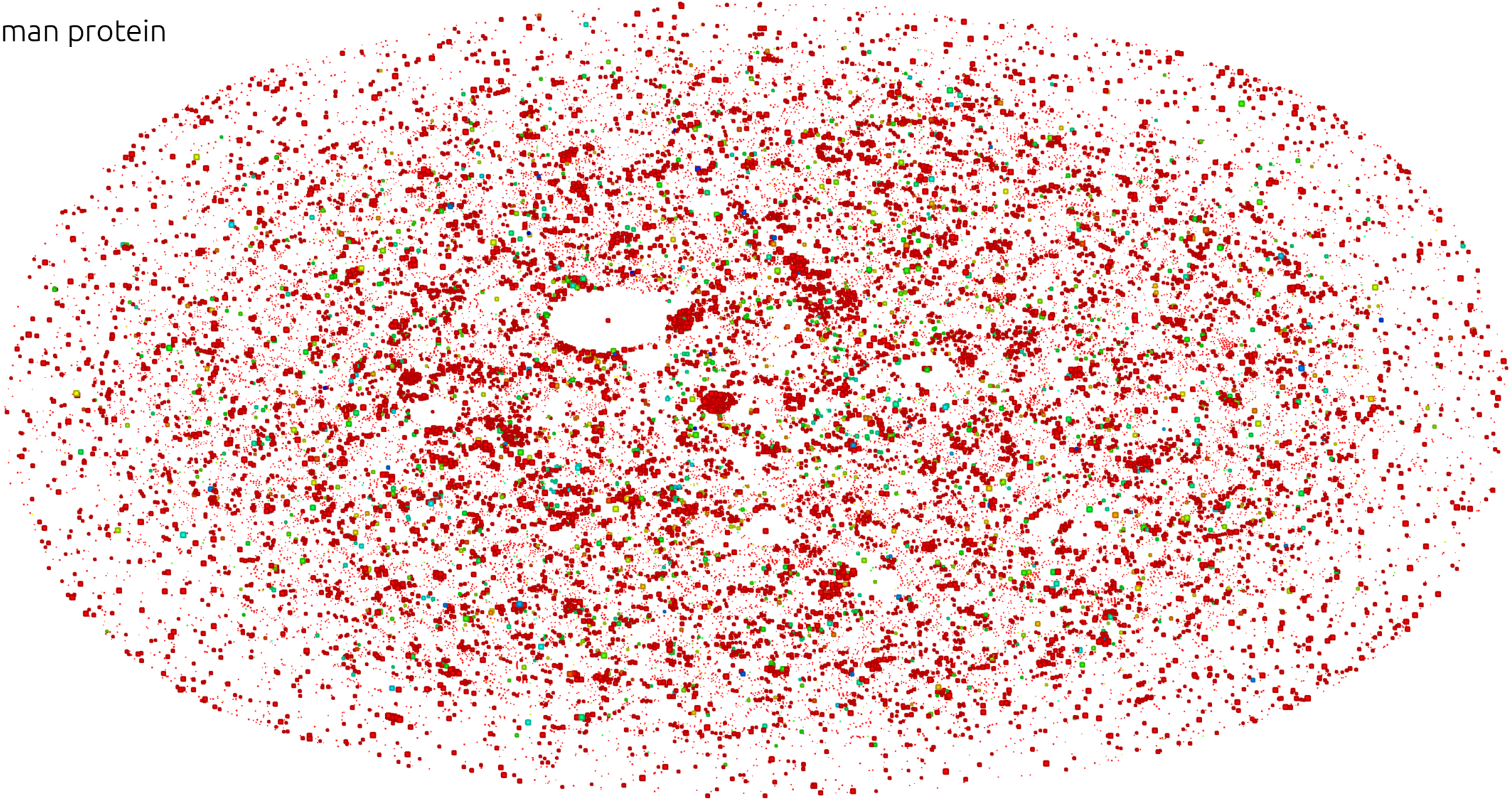
Protein map with rubber bond scaling

Proteins 200 k

idorsia

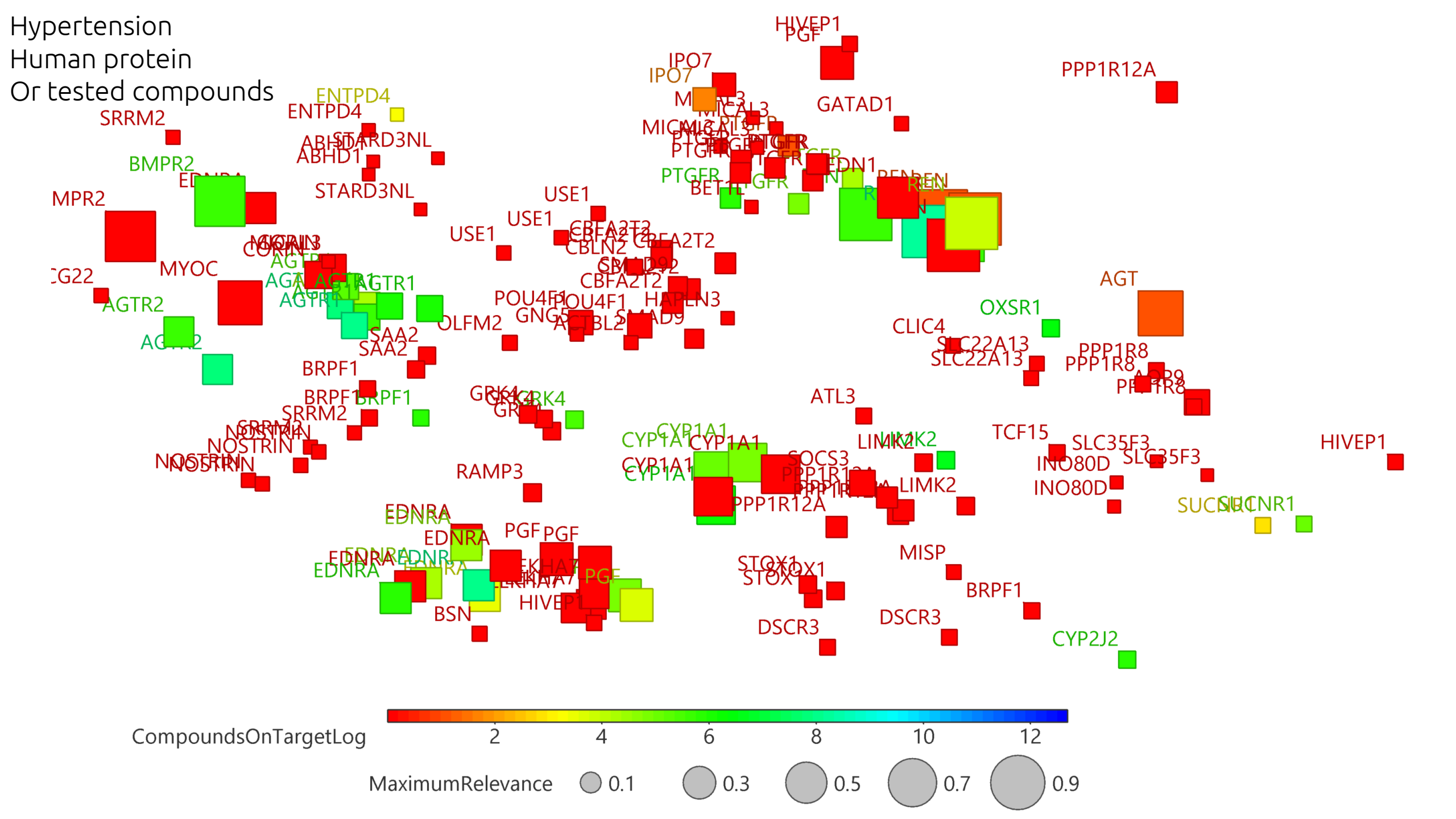

Human protein tested: 2455
CompoundsOnTargetLog
2
4
6
8
10
12
MaximumRelevance
0.1
0.3
0.5
0.7
0.9

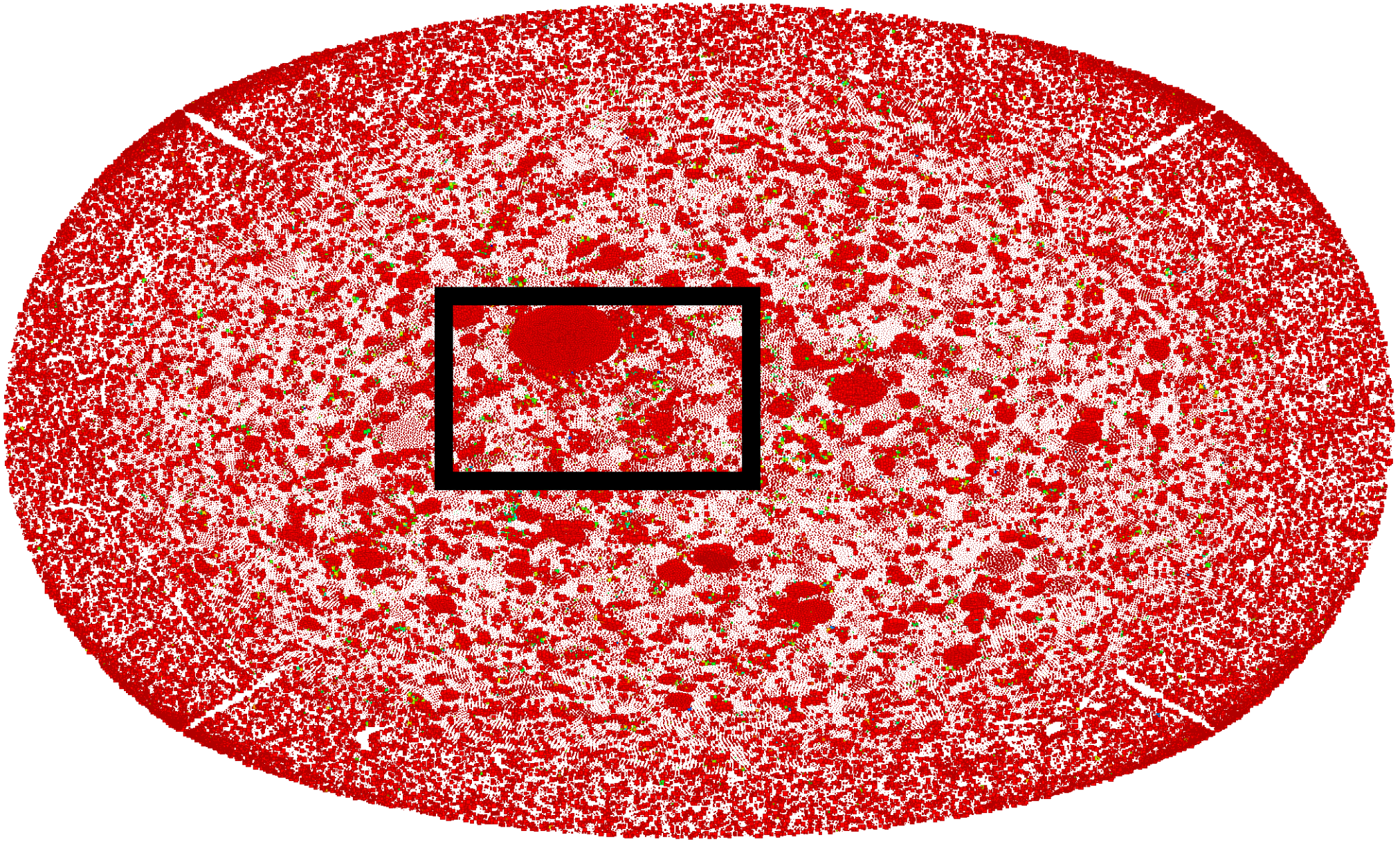
Human protein

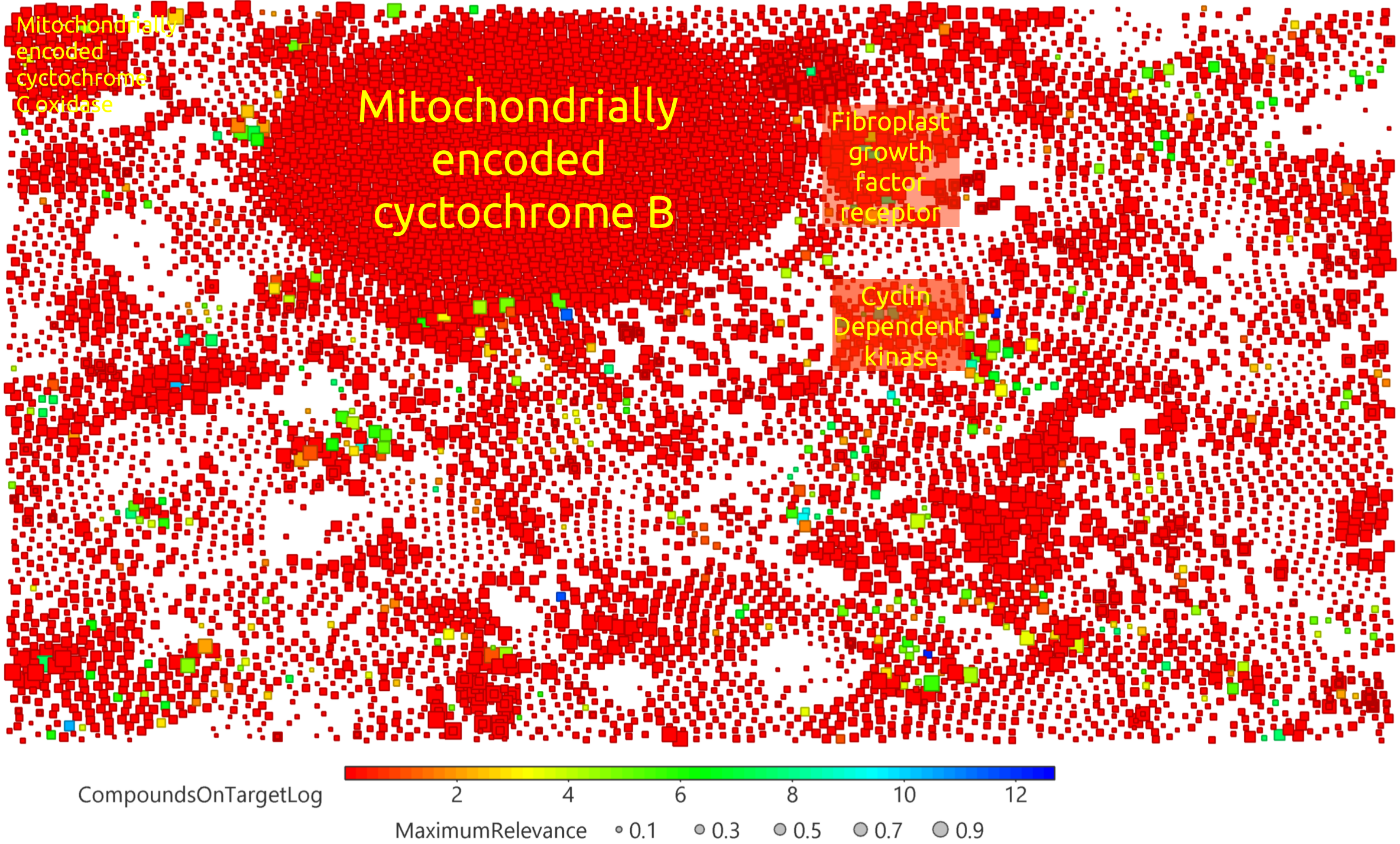


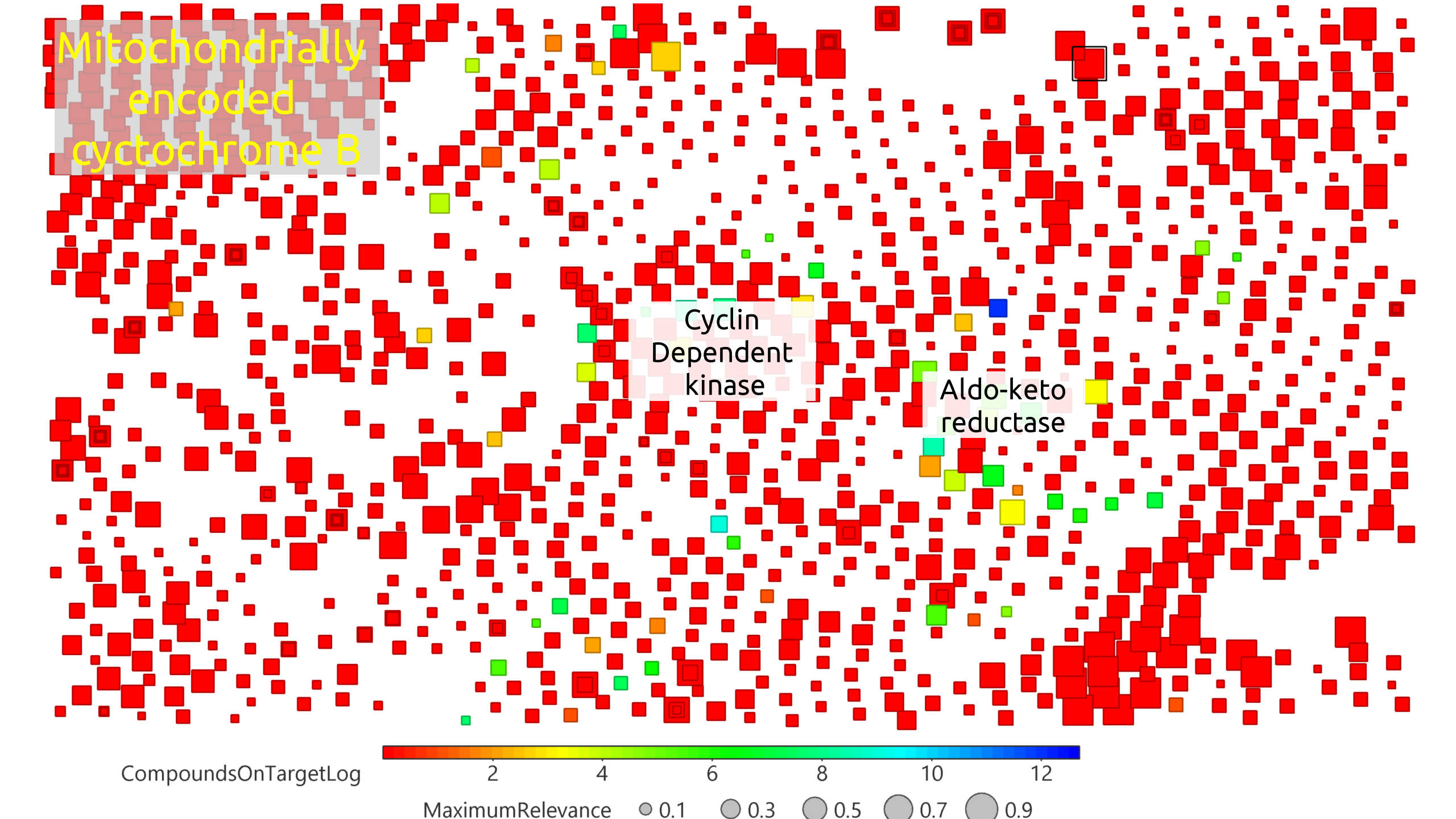
CompoundsOnTargetLog 2 4 6 8 10 12

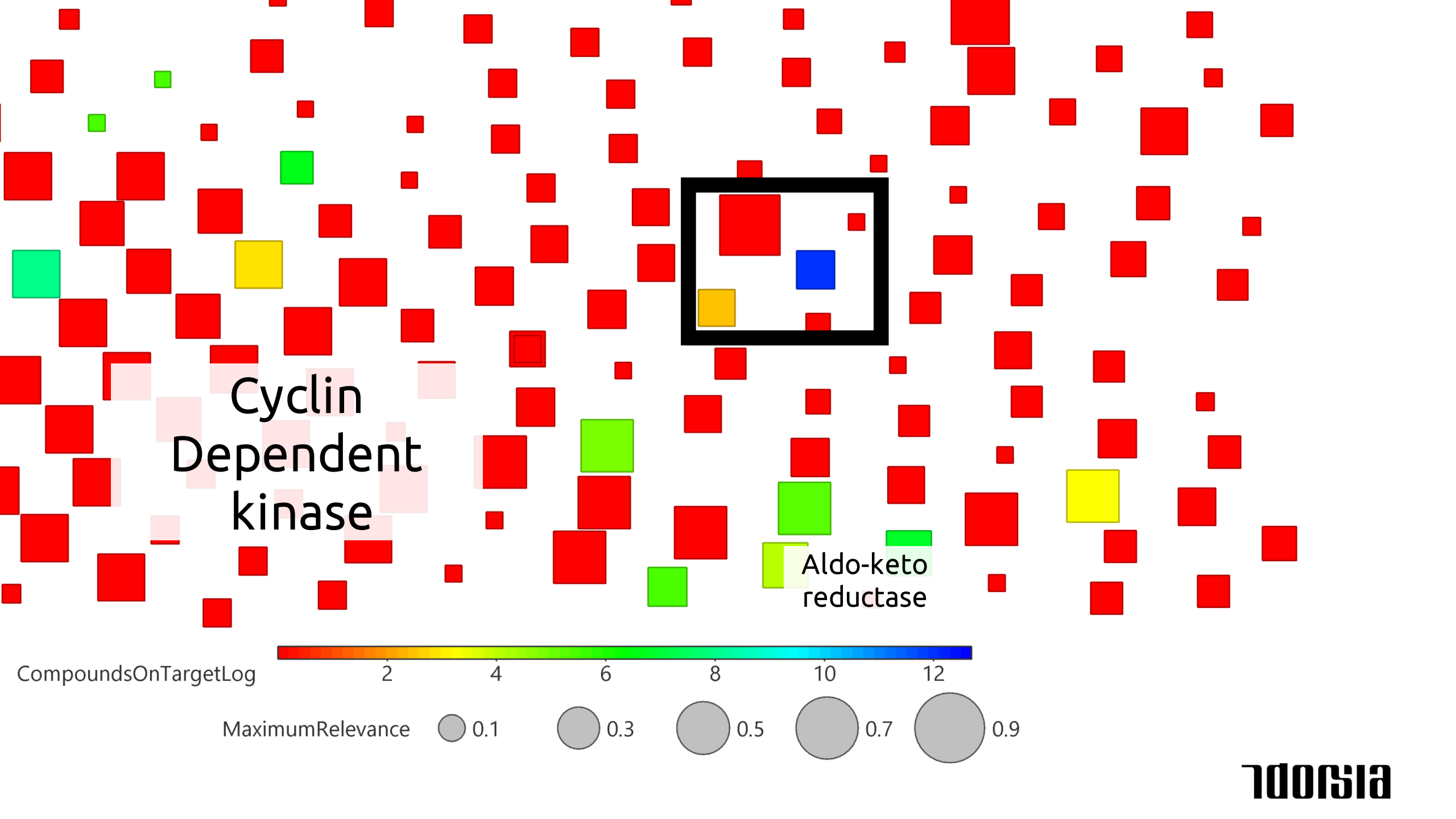
MaximumRelevance 0.1 0.3 0.5 0.7 0.9


Hypertension
Human protein
Or tested compounds
HIVEP1
PGF
IPO7
PPP1R12A
GATAD1
ENTPD4
SRRM2
BMPR2
MPR2
MYOC
AGTR2
AGTR1
USE1
CBFA2T2
CBLN2
POU4F1
HAPLN3
OLFM2
SAA2
BRPF1
GRK4
RAMP3
EDNRA
PGF
HIVEP1
BSN
ATL3
CYP1A1
SOCS3
PPP1R12A
LIMK2
OXSR1
CLIC4
SLC22A13
PPP1R8
AGT
TCF15
SLC35F3
INO80D
SUCNR1
MISP
STOX1
BRPF1
DSCR3
CYP2J2
CompoundsOnTargetLog
2
4
6
8
10
12
MaximumRelevance
0.1
0.3
0.5
0.7
0.9


CompoundsOnTargetLog
2
4
6
8
10
12
MaximumRelevance
0.1
0.3
0.5
0.7
0.9


Mitochondrially encoded cyctochrome C oxidase
Mitochondrially encoded cyctochrome B
Fibroplast growth factor receptor
Cyclin Dependent kinase
CompoundsOnTargetLog
2
4
6
8
10
12
MaximumRelevance
0.1
0.3
0.5
0.7
0.9


Mitochondrially encoded cyctochrome B
Cyclin Dependent kinase
Aldo-keto reductase
CompoundsOnTargetLog
2
4
6
8
10
12
MaximumRelevance
0.1
0.3
0.5
0.7
0.9


Cyclin
Dependent
kinase
Aldo-keto
reductase
CompoundsOnTargetLog
2
4
6
8
10
12
MaximumRelevance
0.1
0.3
0.5
0.7
0.9

Target cliffs

SMN1

???

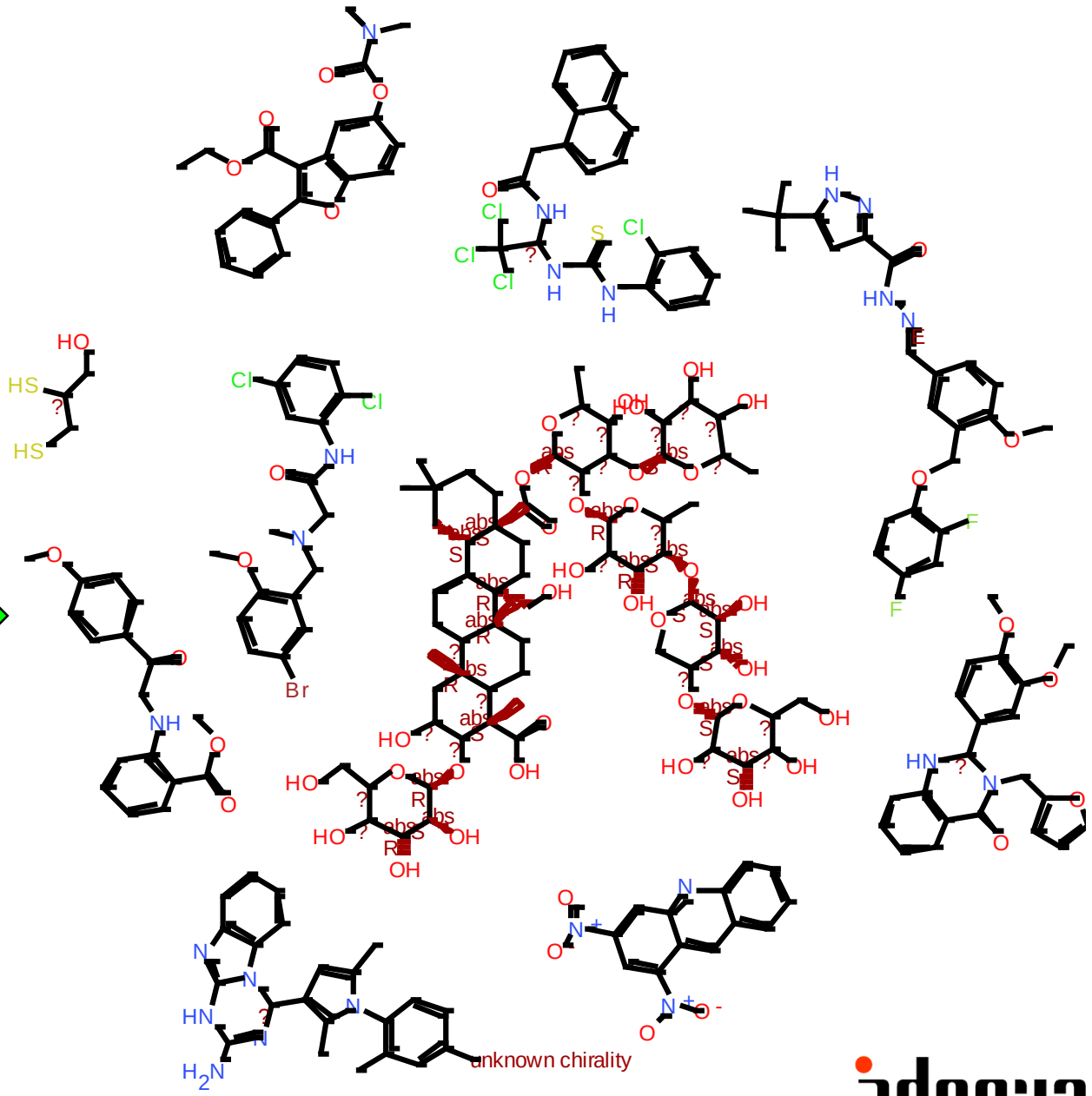
Muscular atrophy

SMAD3

Tested 64 k compounds

>4000 activity < 1 µMol

Fibrosis



idorsia

Summary & conclusions

- Around 4000 target proteins are covered by ChEMBL bioactivity
- Almost 200’000 additional proteins were analyzed
- 15 million similarity relations were derived from compound- and protein- similarity
- DataWarrior rubber bond scaling mapped all proteins
- Visualization largest part of the known genome together with tested compounds
- Target cliffs are valuable starting points for drug discovery

Thank you