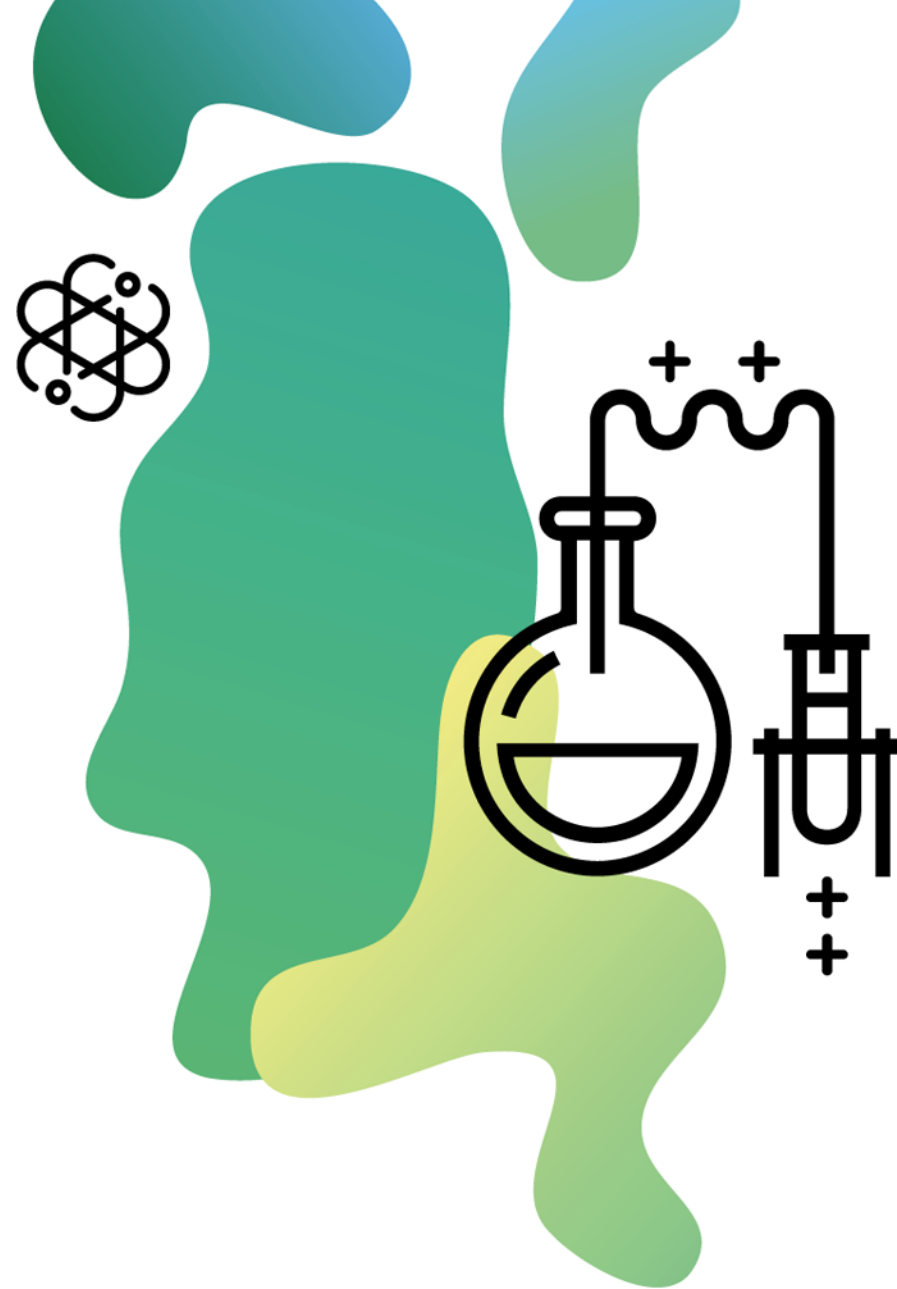




# Targeting of the disease related proteome by small molecules

Workshop on chemical information  
2018  
EPFL



# Outline

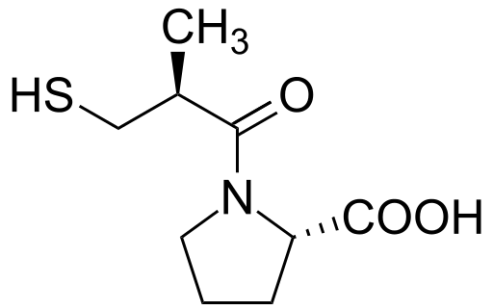
- Introduction
- Data sources
  - Diseases
  - Proteins
  - Compounds
- Data relations
  - Diseases – target proteins: relevance estimator:
  - Protein – protein: BLAST similarity
  - Compound –compound: Flexophore descriptor similarity
- Visualization: rubber bond map
- Summary & conclusions

# What is pharmaceutical industry doing?

Deliver medicine

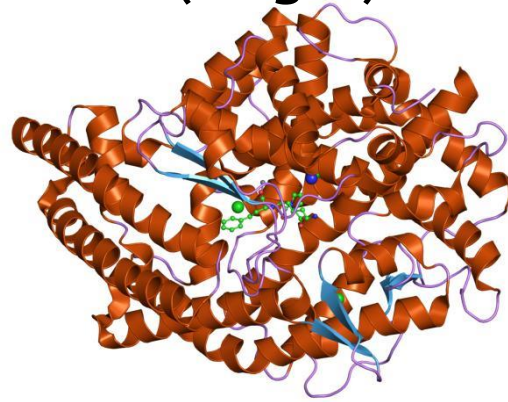
Switch back to normal condition (blood pressure)

Drug

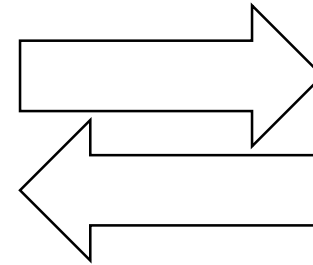


Molecule  
Captopril

Protein  
(target)



Angiotensin converting enzyme



Disease



Hypertension

# Diseases

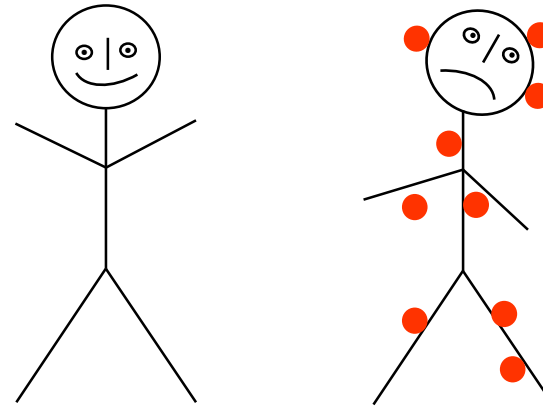
A condition of the living animal or plant body or of one of its parts that impairs normal functioning and is typically manifested by distinguishing signs and symptoms.

- 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

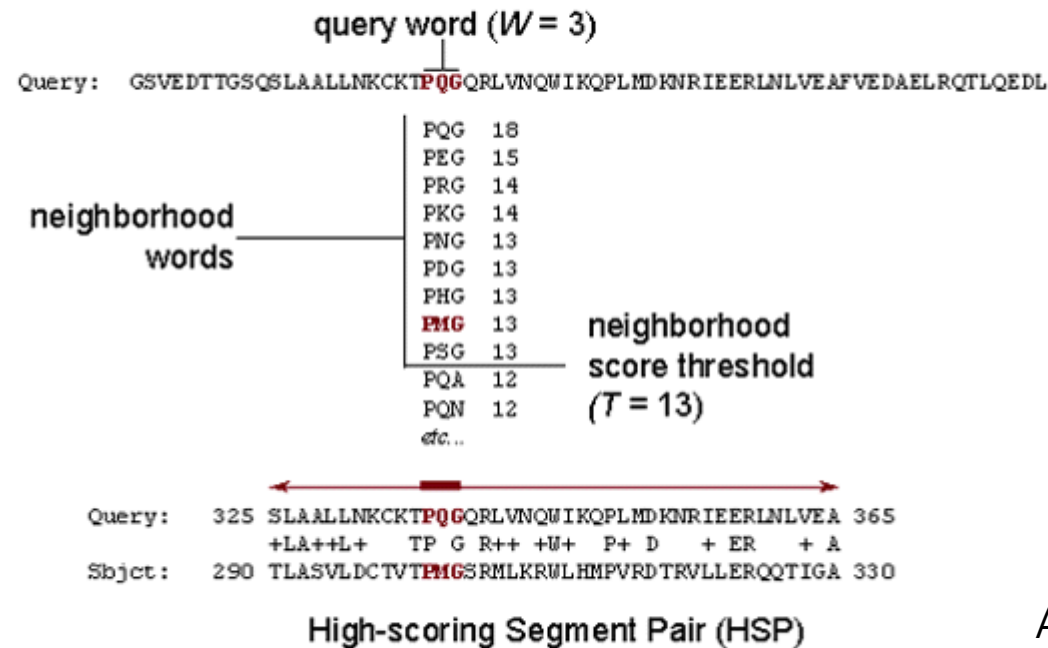
- Enzymes
  - Hydrolases (ACE, 1956)
  - Kinases (cancer therapy, emerging drug targets)
- Receptors
  - G protein coupled receptors (70 % of all drugs)
- Ion channels
  - Emerging drug targets
  - Anti targets (hERG)

# BLAST

Basic Local Alignment Search Tool

Protein similarity by amino acid sequence alignment

## The BLAST algorithm



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

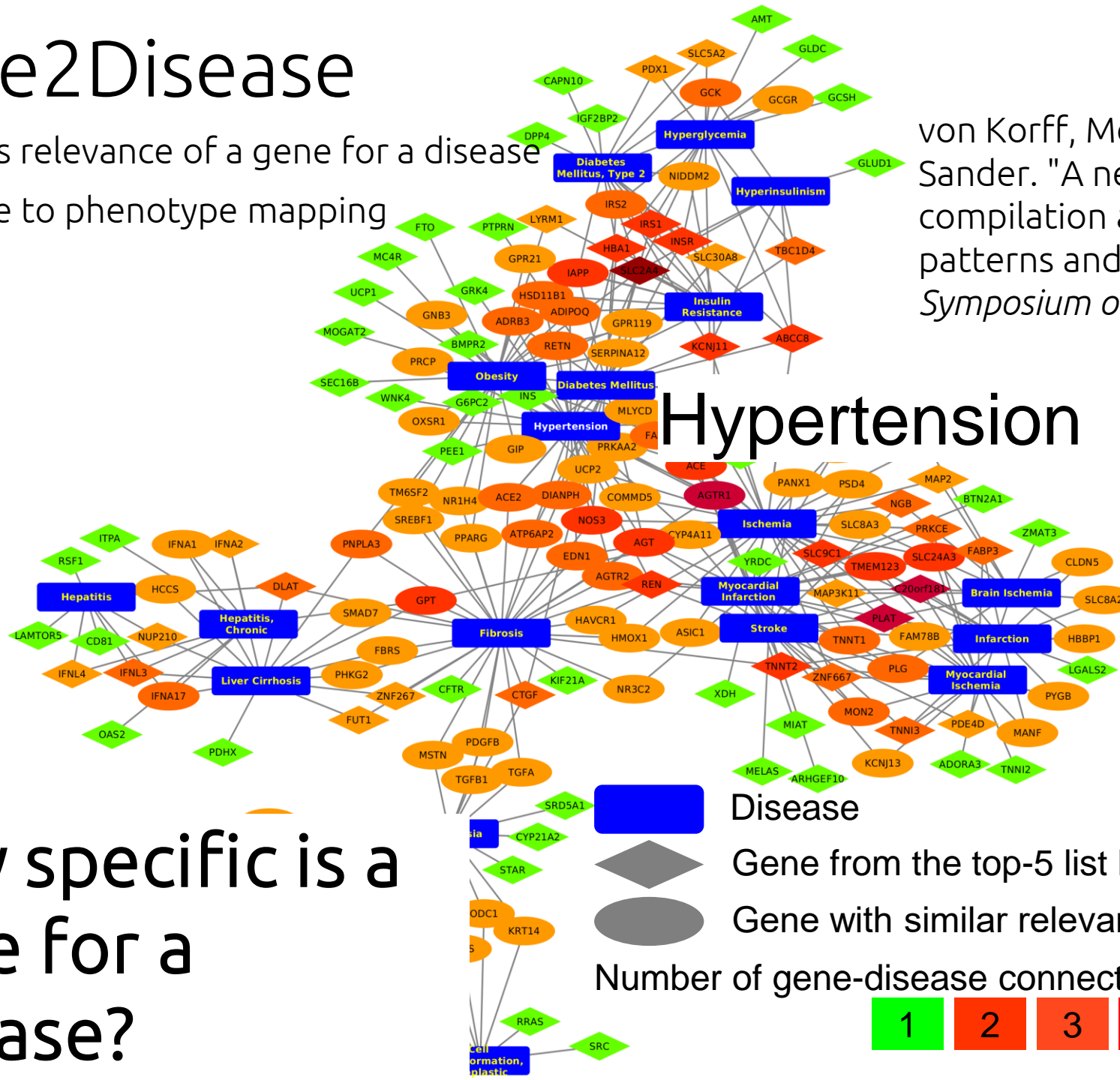
Sophisticated form of character comparison

# Gene2Disease

Estimates relevance of a gene for a disease

Genotype to phenotype mapping

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.



How specific is a gene for a disease?

# ChEMBL database

Structures, biological activities, target protein identifiers

Version 23, quality 9 (highest)

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

## Most tested proteins

POLI (DNA polymerase iota ) 116'761

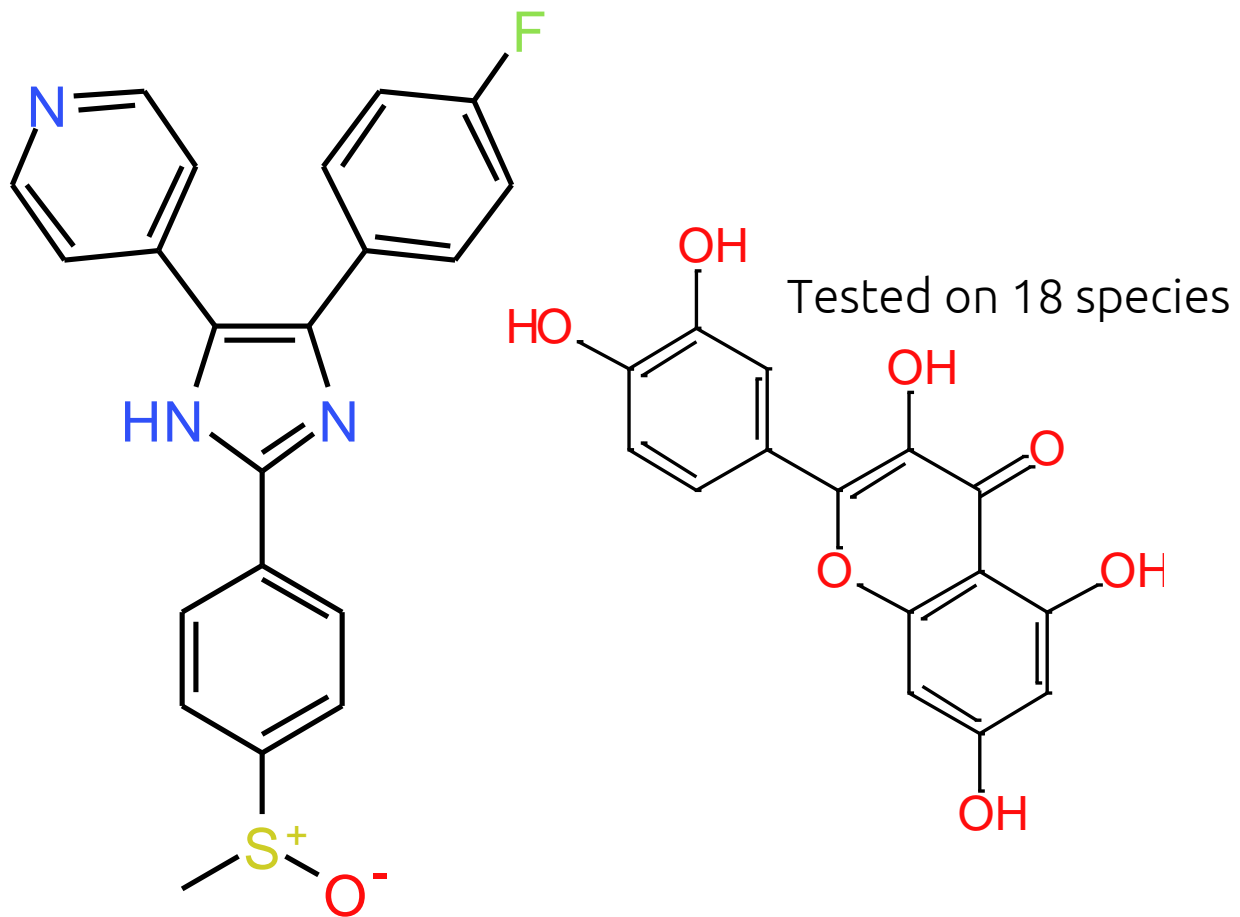
ATAD5 (ATPase family AAA domain-containing protein 5) 122'498

GMNN (Geminin, DNA replication inhibitor) 127'916

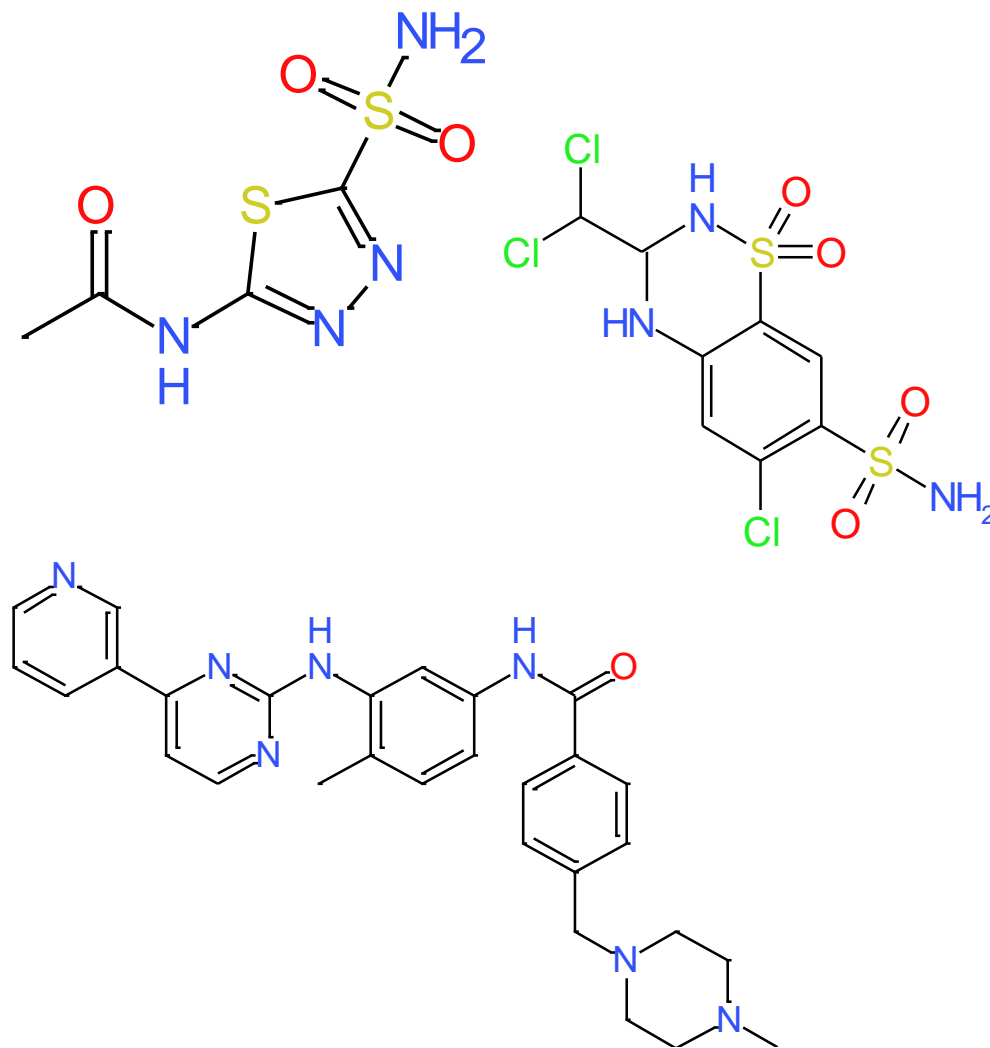
Gaulton, Anna, et al. "ChEMBL: a large-scale bioactivity database for drug discovery." *Nucleic acids research* 40.D1 (2011): D1100-D1107.



# ChEMBL database compounds



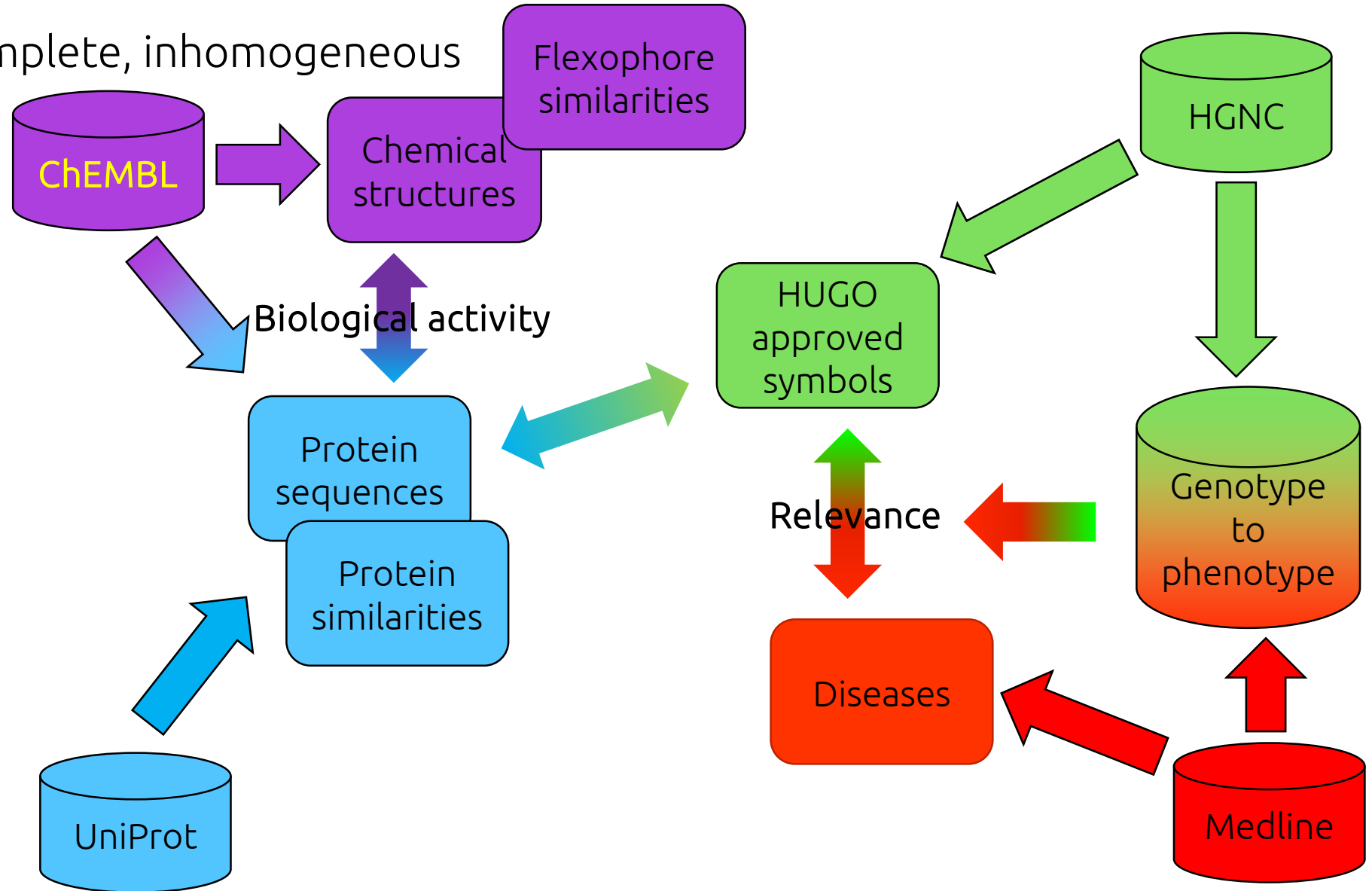
Tested on >500 targets



Active on >100 targets

# Data relations

Many, incomplete, inhomogeneous



# Map desired



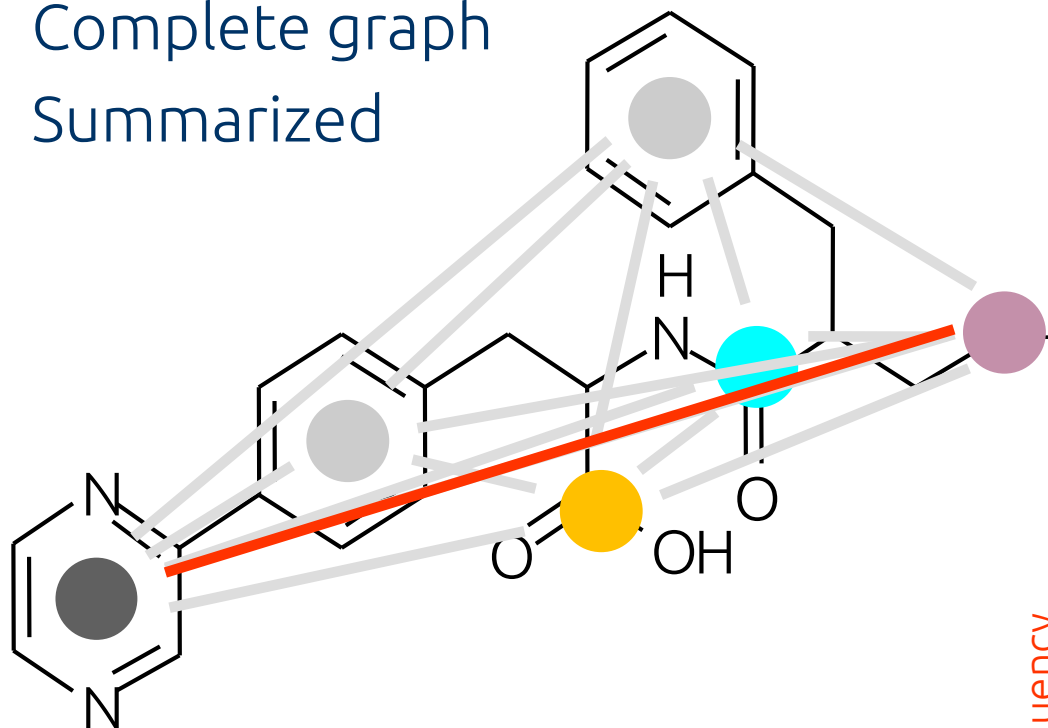
Martin Waldseemüller's World Map of 1507, the first map to use the name "America" to label the New World



# Compound – compound similarity

Pharmacophore descriptor: Flexophore

Complete graph  
Summarized



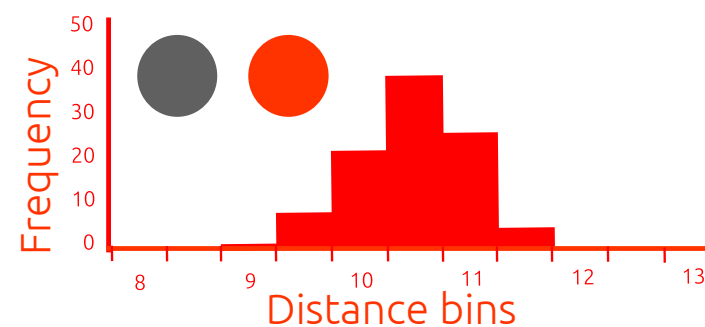
Nodes:

MM2 atom types



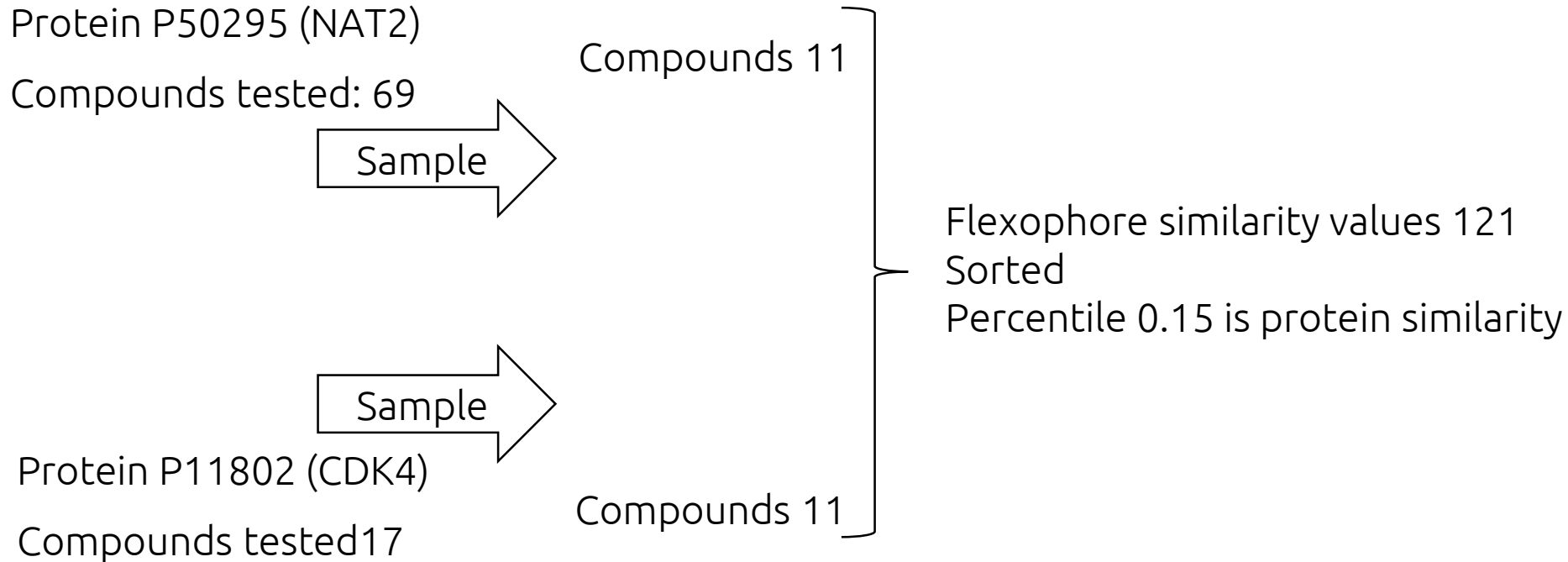
Edges:

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.

# 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

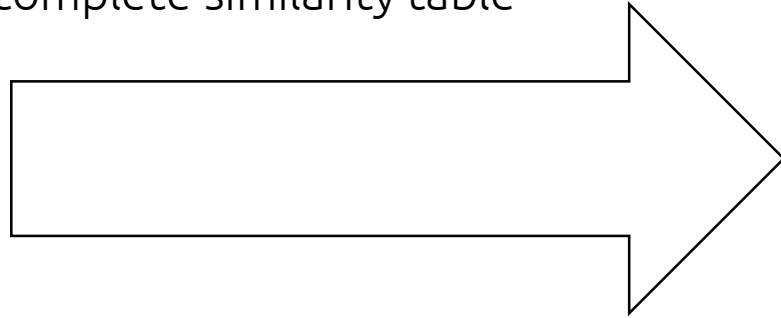
- ChEMBL structure records 890'000
- Activity values 4'600'00
- Protein identifiers 4'200
- Similarity values 8'700'000 from Flexophore

Scaled

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

# 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.

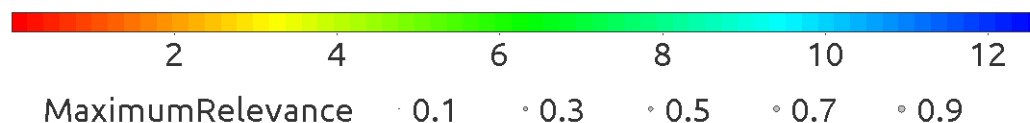


Protein  
map with  
rubber  
bond  
scaling

The map

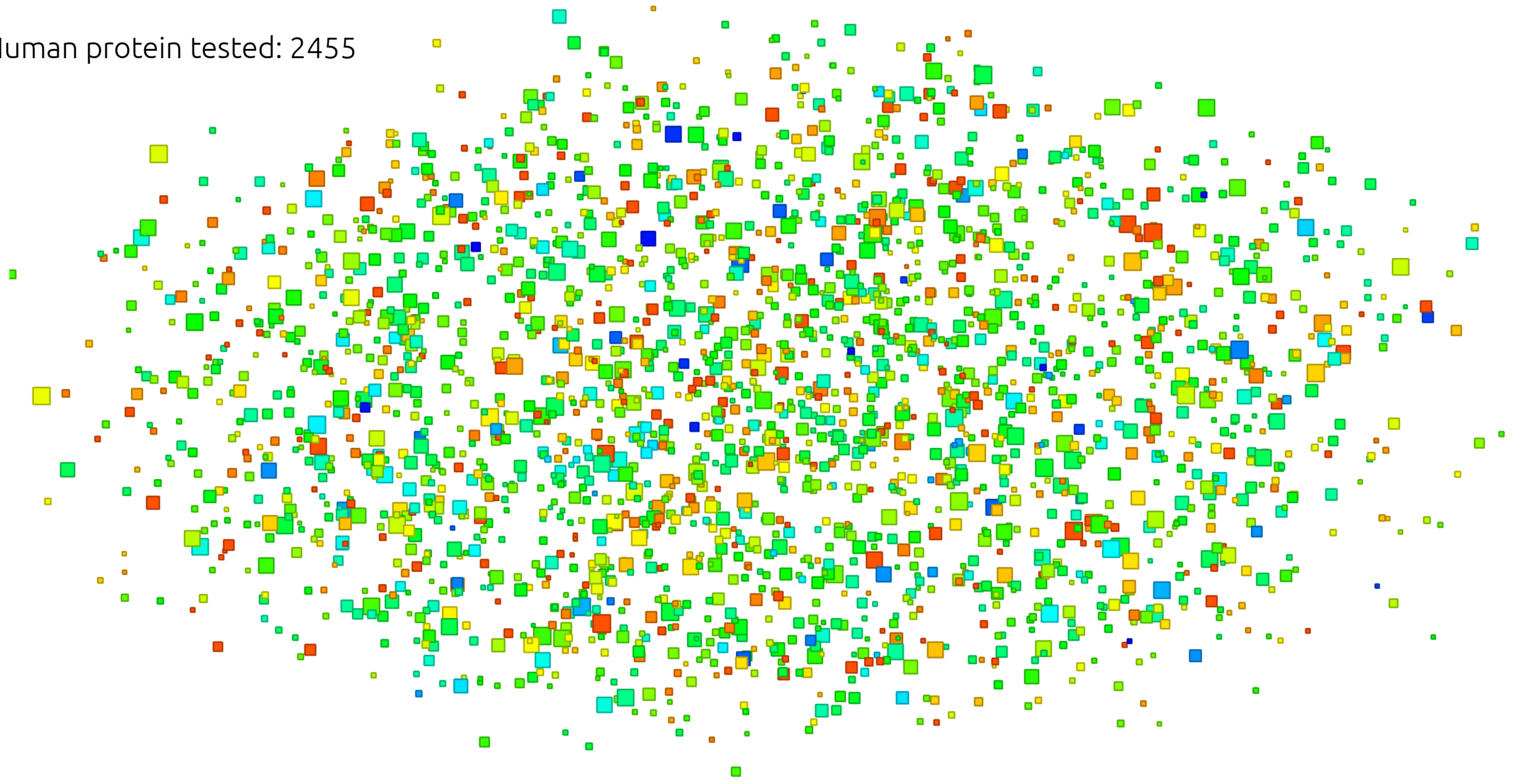
Proteins 200 k

CompoundsOnTargetLog

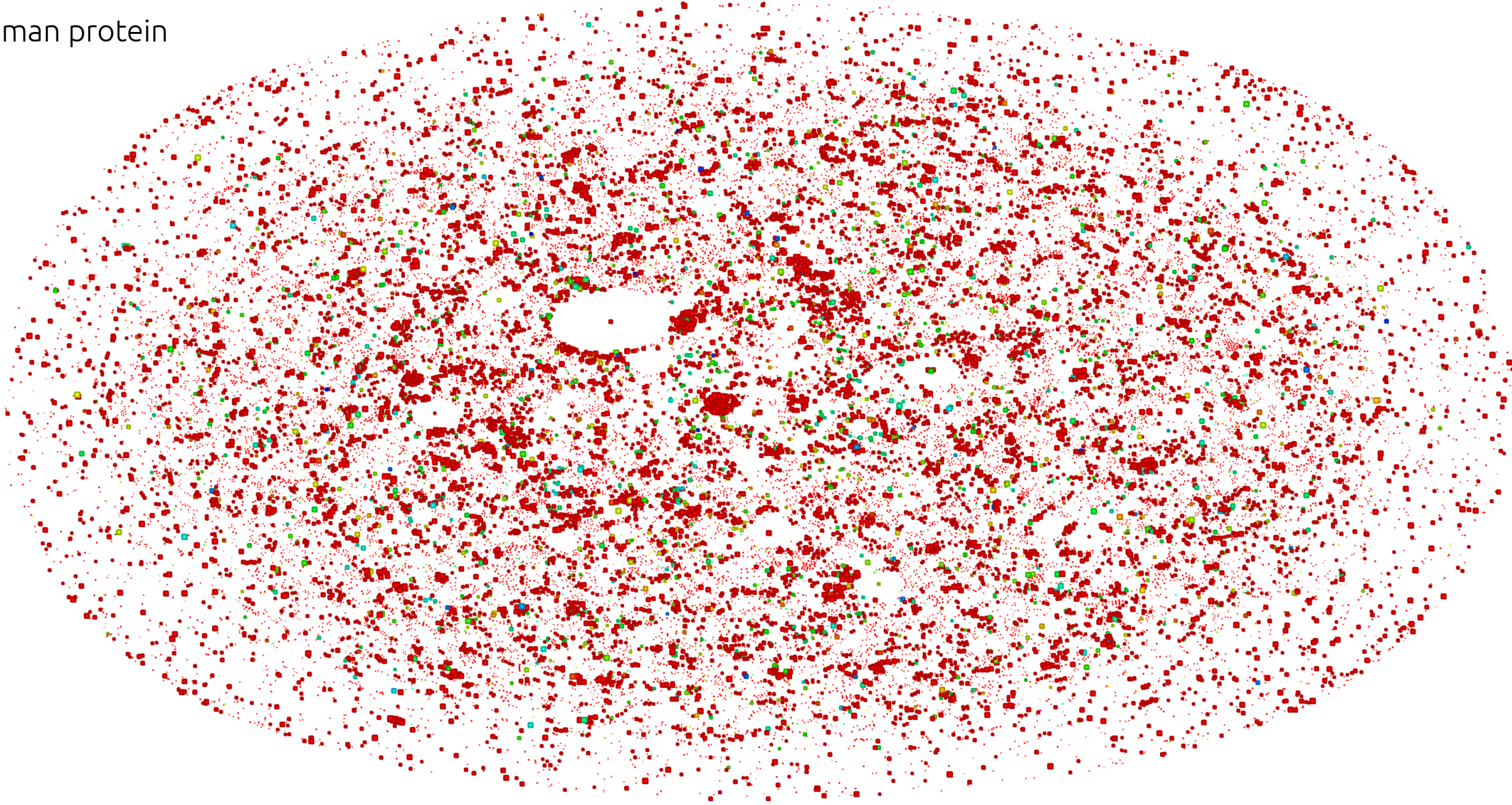




Human protein tested: 2455



Human protein



CompoundsOnTargetLog



2

4

6

8

10

12

MaximumRelevance

0.1

0.3

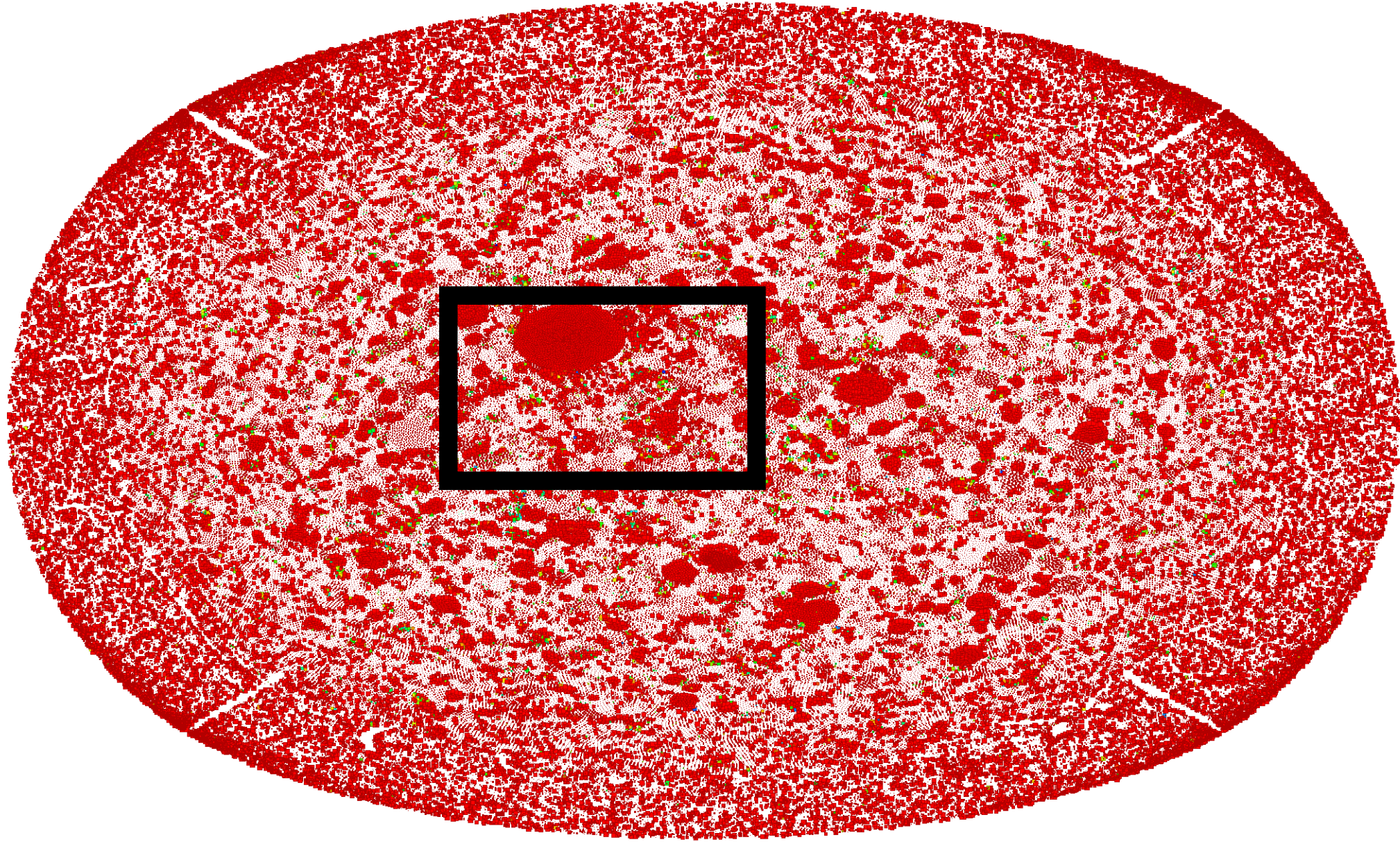
0.5

0.7

0.9









Mitochondrially  
encoded  
cytochrome  
C oxidase

Mitochondrially  
encoded  
cytochrome B

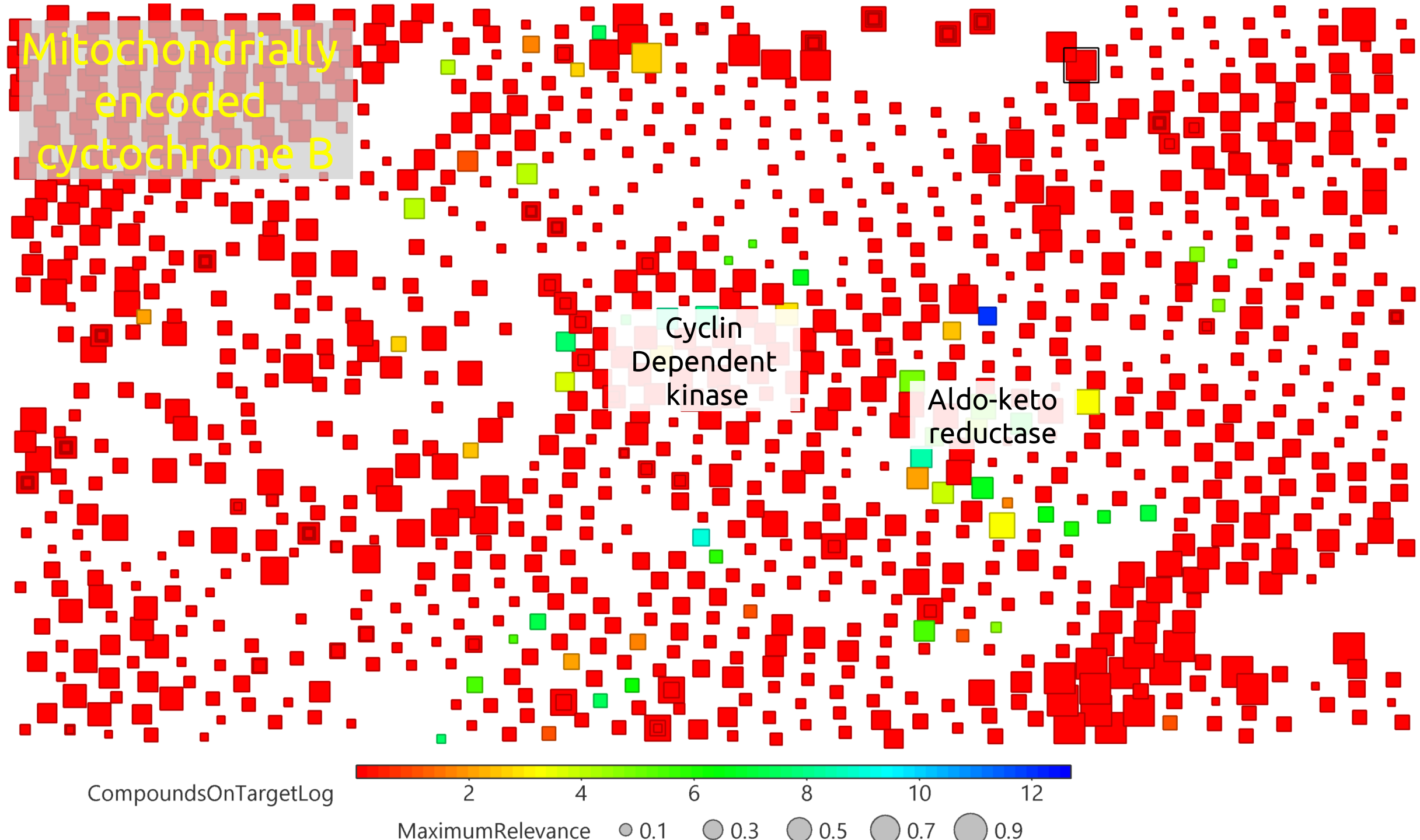
Fibroblast  
growth  
factor  
receptor

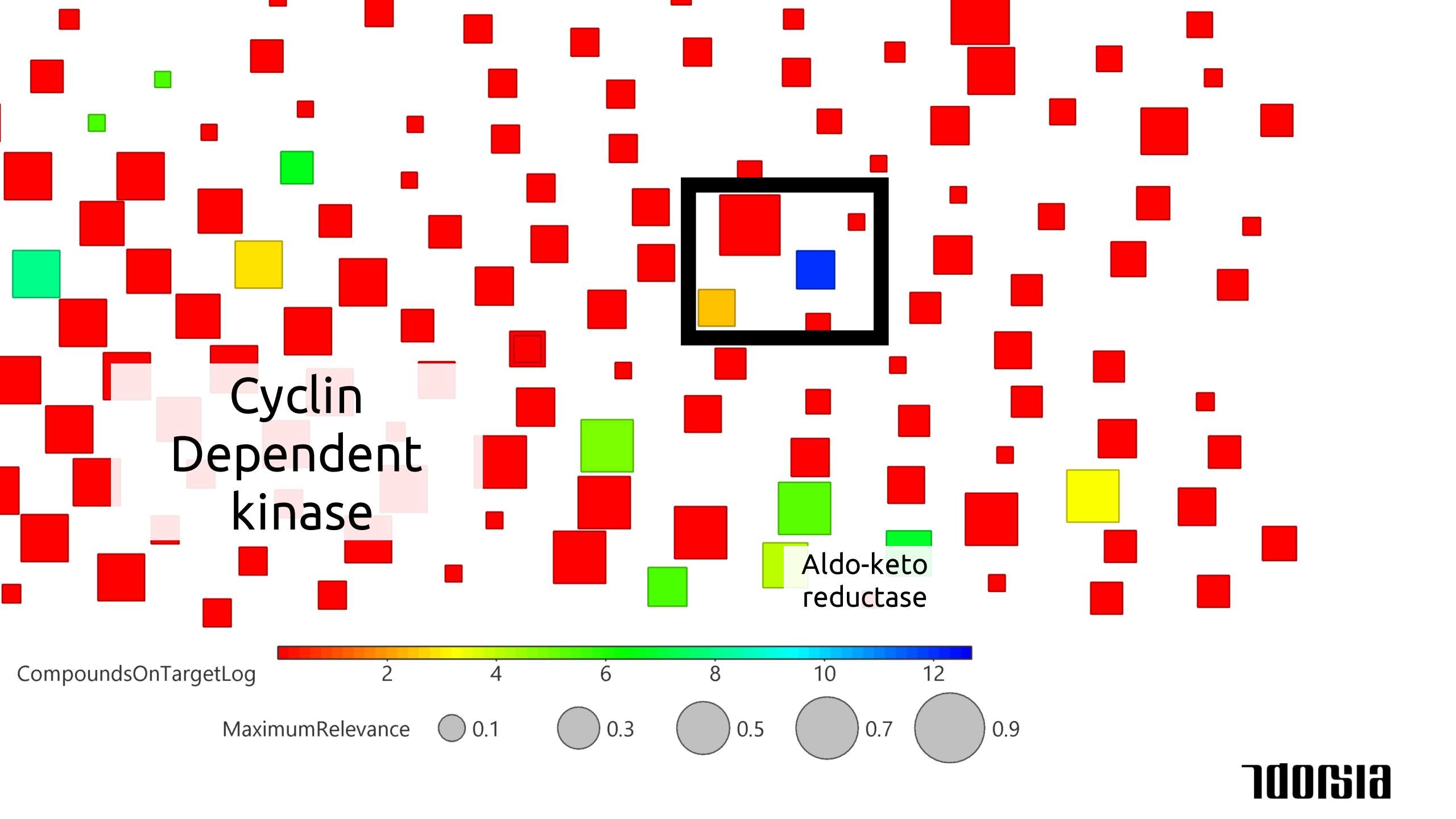
Cyclin  
Dependent  
kinase

CompoundsOnTargetLog

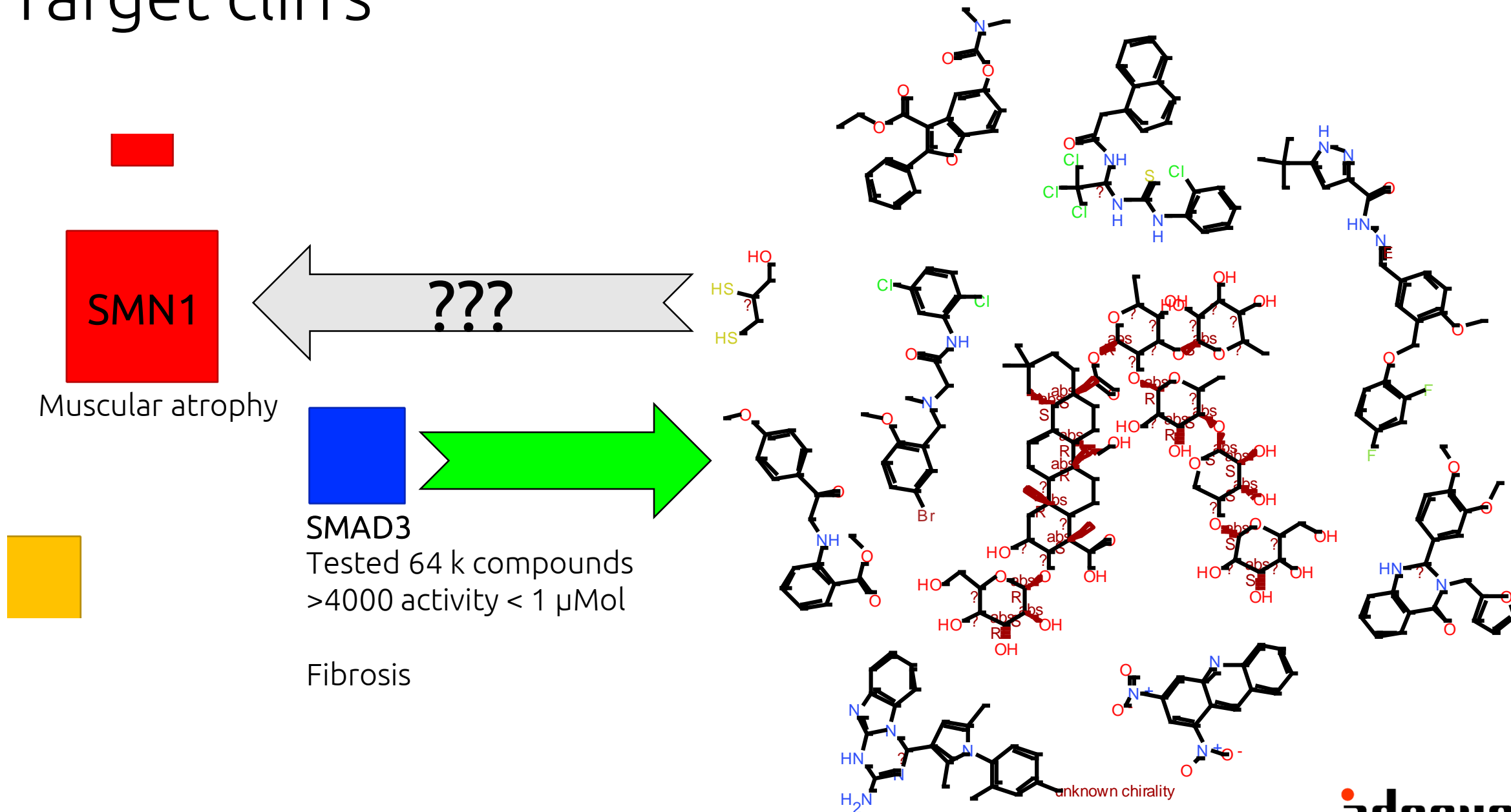


MaximumRelevance    • 0.1    • 0.3    • 0.5    • 0.7    • 0.9





# Target cliffs





# 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

idorsia

Thank you

