b-it LSi
Life
Science
Informatics

LIMES
Life & Medical Sciences Institute

Chemical Space Networks and SAR Visualization

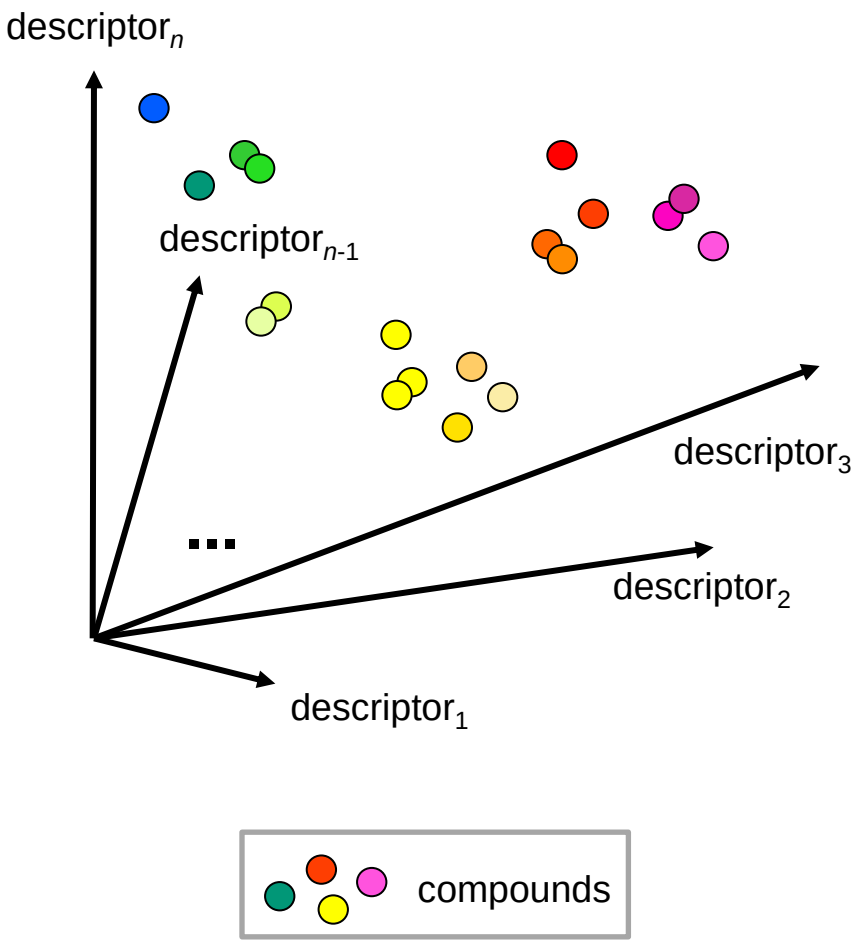
Martin Vogt, Jürgen Bajorath
Life Science Informatics, B-IT
University of Bonn

January 11, 2017

universitätbonn

Chemical Space Representations

- **Coordinate-based** chemical space design

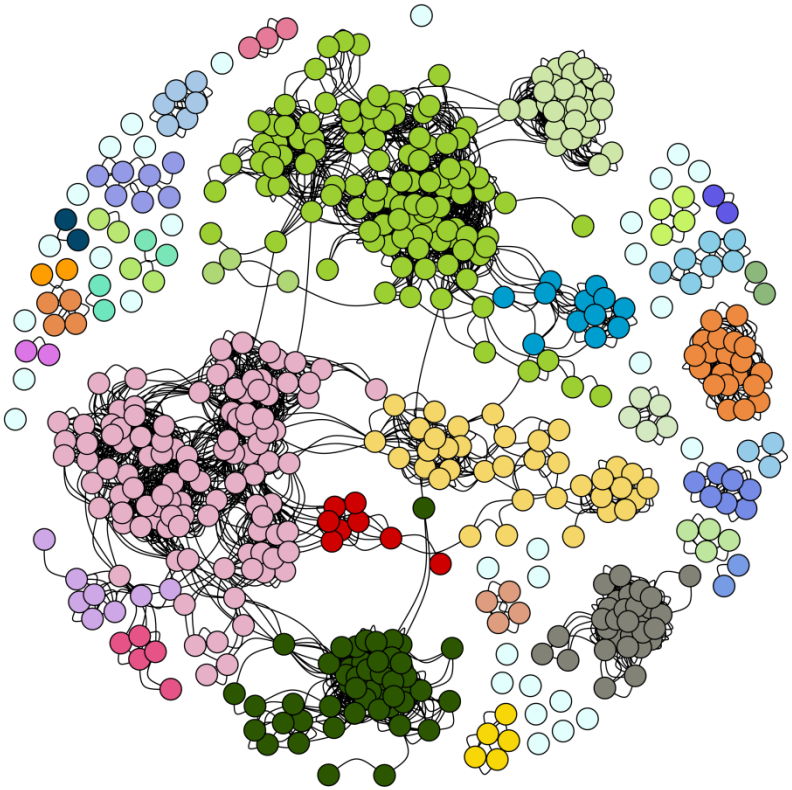


Chemical Space Representations

- **Coordinate-free** representations

Chemical Space Representations

- **Coordinate-free** representations
- **Molecular networks**

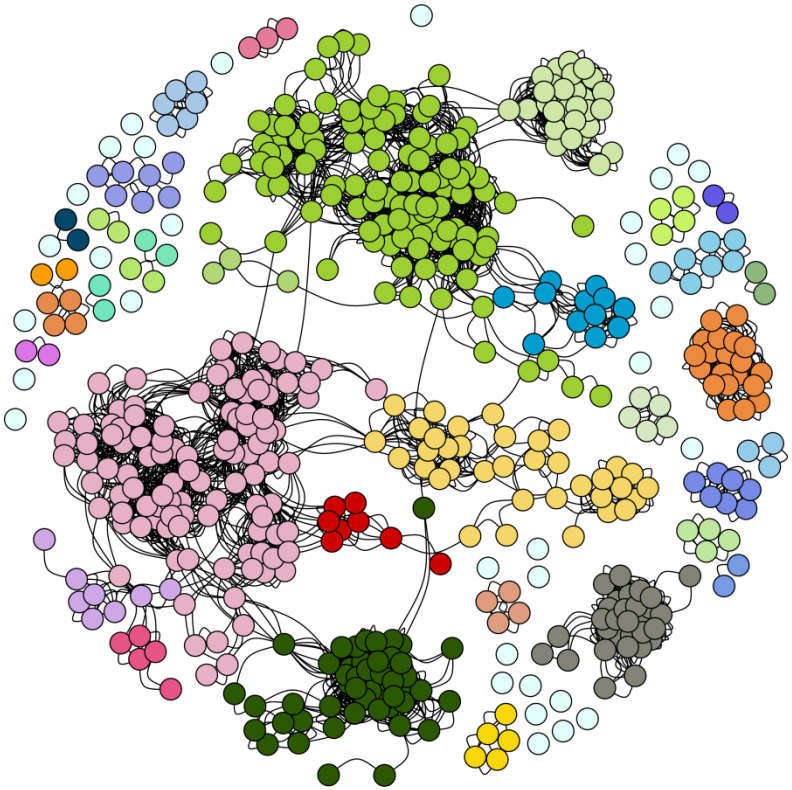


Dopamine D4 receptor ligands

Chemical Space Representations

- **Coordinate-free** representations

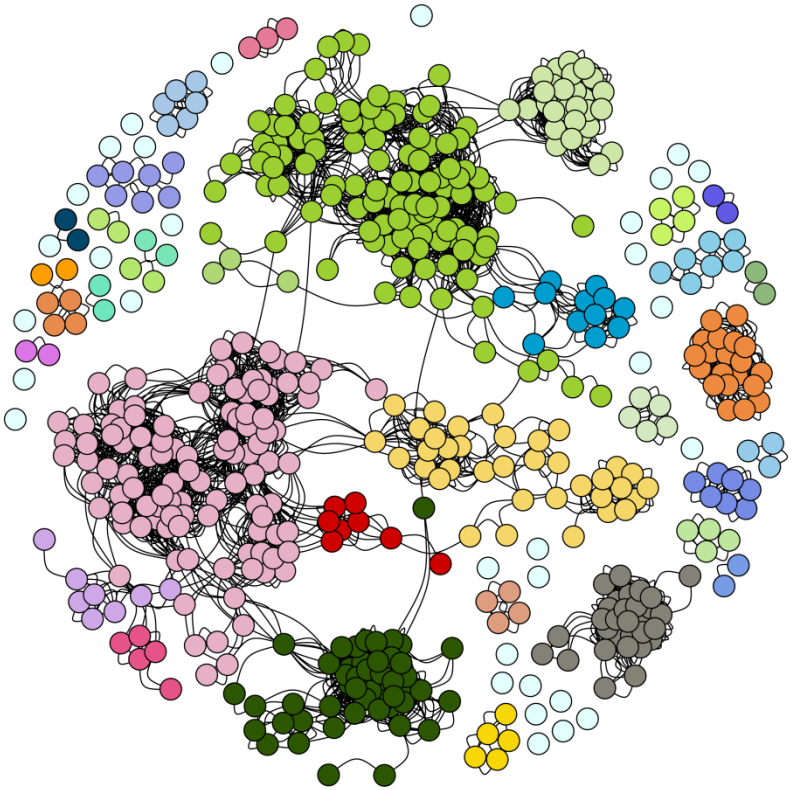
- **'Chemical space networks'**
  - similarity-based compound networks
  - **nodes:** compounds
  - **edges:** pairwise similarity relationships



  - exploring **biologically relevant chemical space**

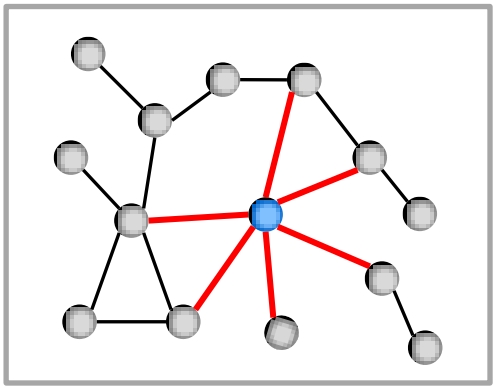
Chemical Space Networks (CSNs)

- **CSNs:** immediate graphical access and interpretability
- **Quantitative analysis**
  - statistical concepts from **network science**
  - network properties

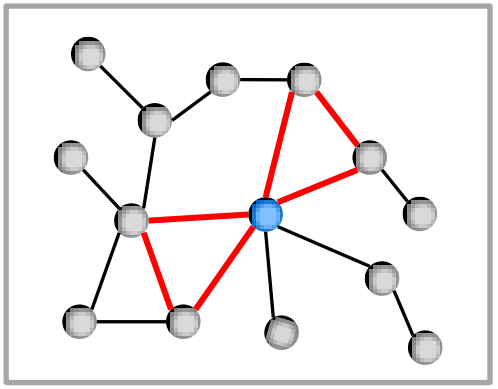


Network Properties

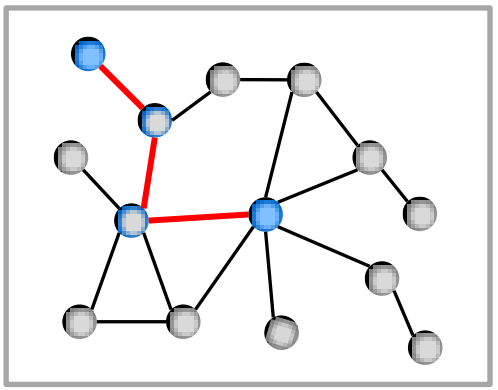
Node degree



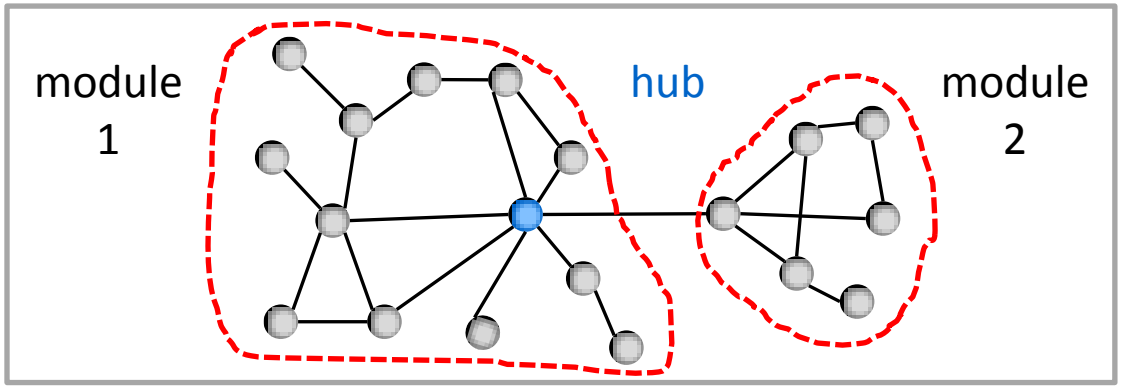
Clustering coefficient



Path length



Community structure



Network density

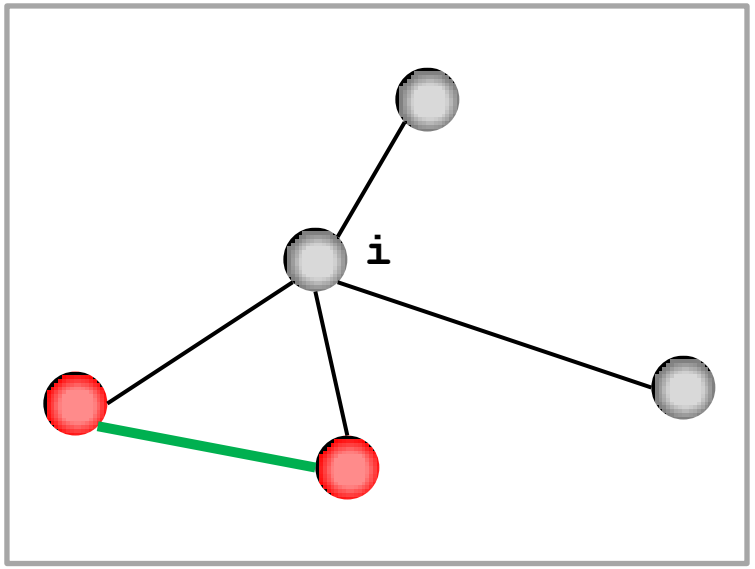
Degree assortativity

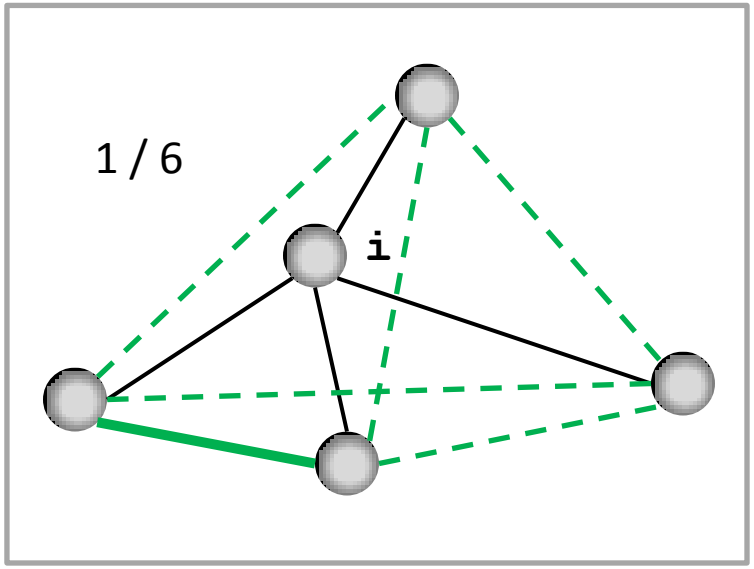
Modularity

Clustering Coefficient

- **Clustering coefficient:** degree to which neighbors of a given node are connected to each other

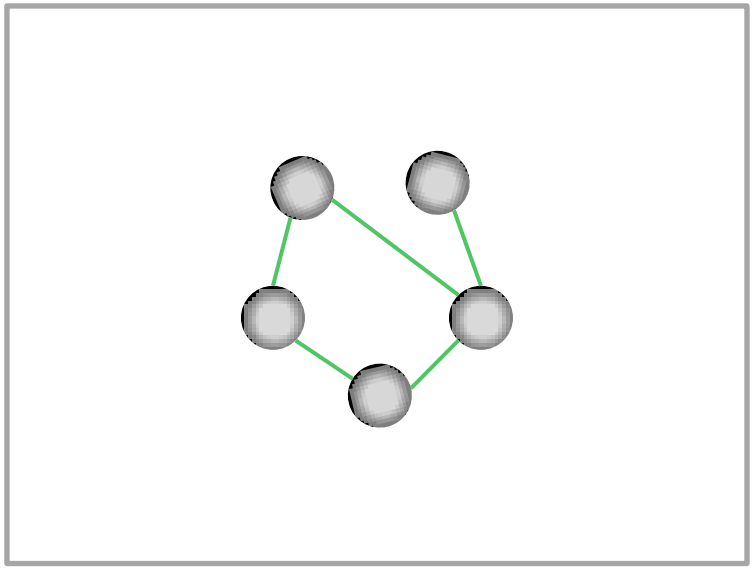
- Clustering coefficient of a network: average of all node coefficients

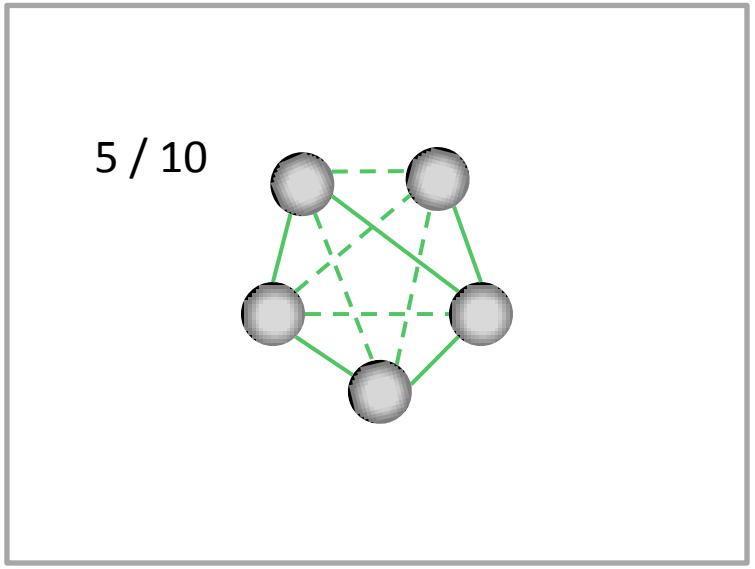




Network Density

- **Network edge density** defined as the:
  (Number of observed edges) / (Number of possible edges)

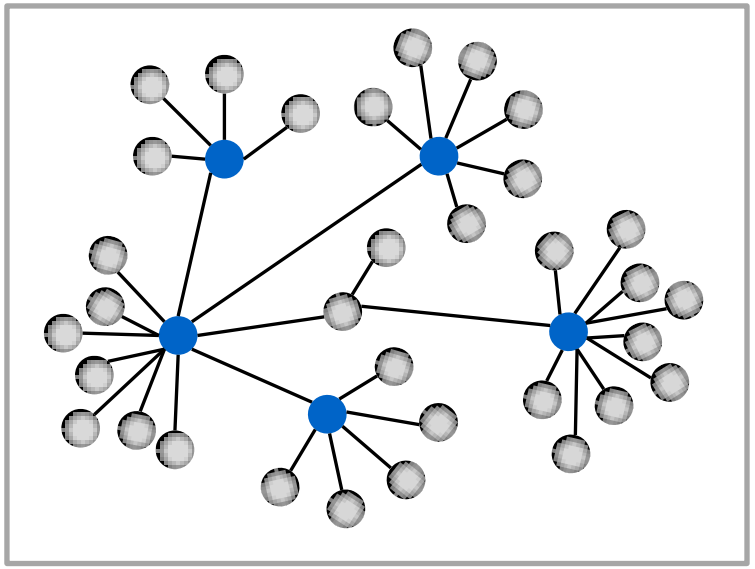




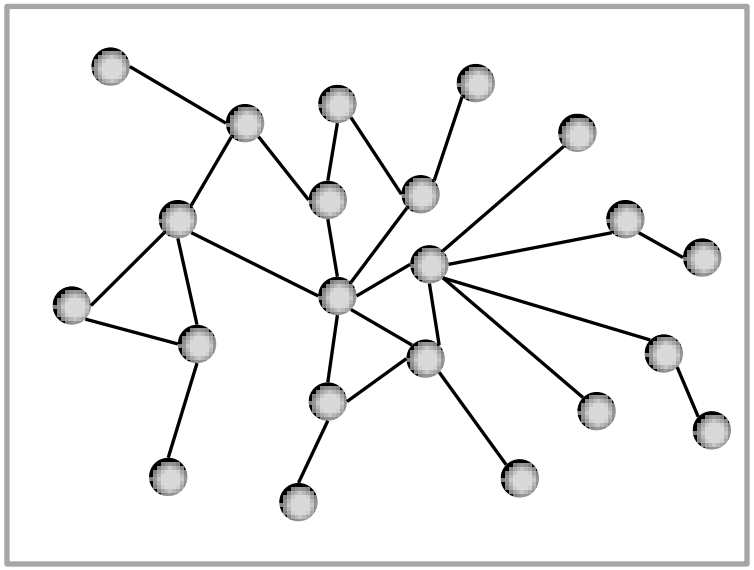
Degree Assortativity

- **Assortativity:** defined as the correlation coefficient between the degree of pairs of connected nodes
- Hubs lead to disassortative networks

disassortative



assortative



Assortativity and Homophily

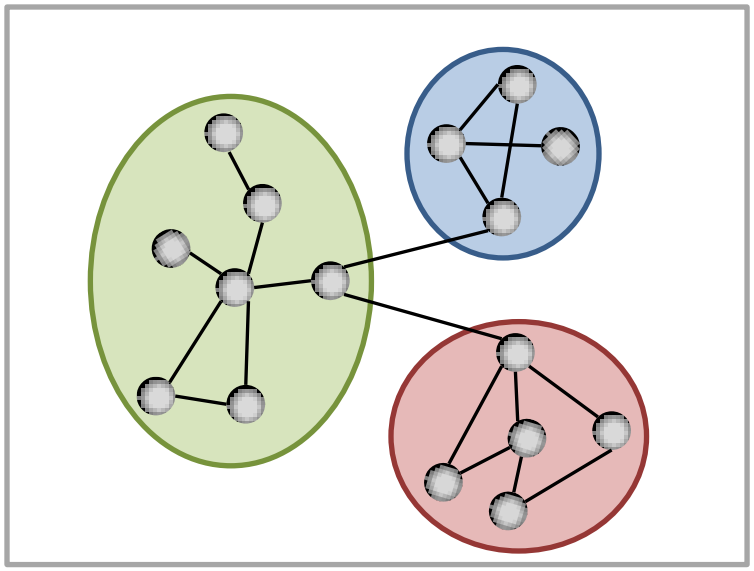
- **Assortativity:** defined as the correlation coefficient between the degree of pairs of connected nodes

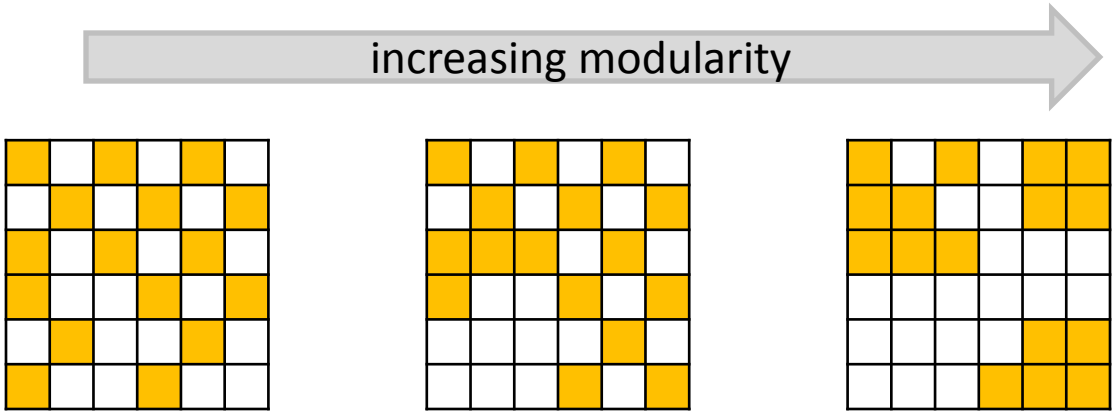
- ***Homophily principle*** from **network science**: nodes with similar ***latent characteristics*** are more likely connected than others (social networks)
  - *latent characteristic* of CSNs: **compound activity**
  - activity annotation through node coloring: **SAR visualization**

- **High assortativity** is a consequence of **homophily**

Modularity

- Modularity measures global separation of nodes into **communities** (**clusters**)
- ‘Small world‘ character
- Compound communities in CSNs ➔ **SAR analysis**





CSNs of Different Design

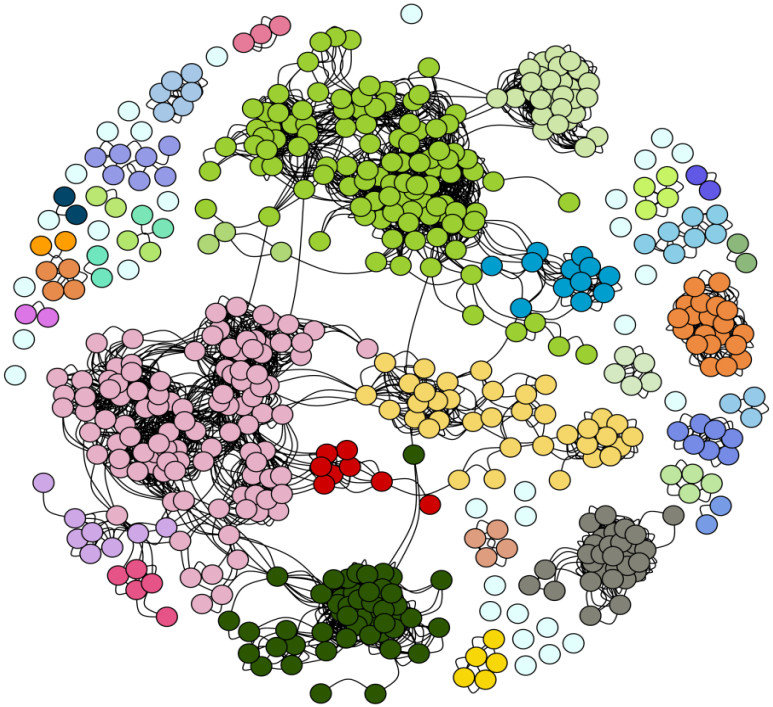
Similarity as a design variable

- **Threshold CSN (THR-CSN)**
  - continuous similarity metric (Tanimoto coefficient, Tc)
- **Matched Molecular Pair CSN (MMP-CSN)**
  - substructure-based similarity criterion
- **Tversky-CSN (TV-CSN)**
  - asymmetric similarity relationships

Universal CSN Implementation

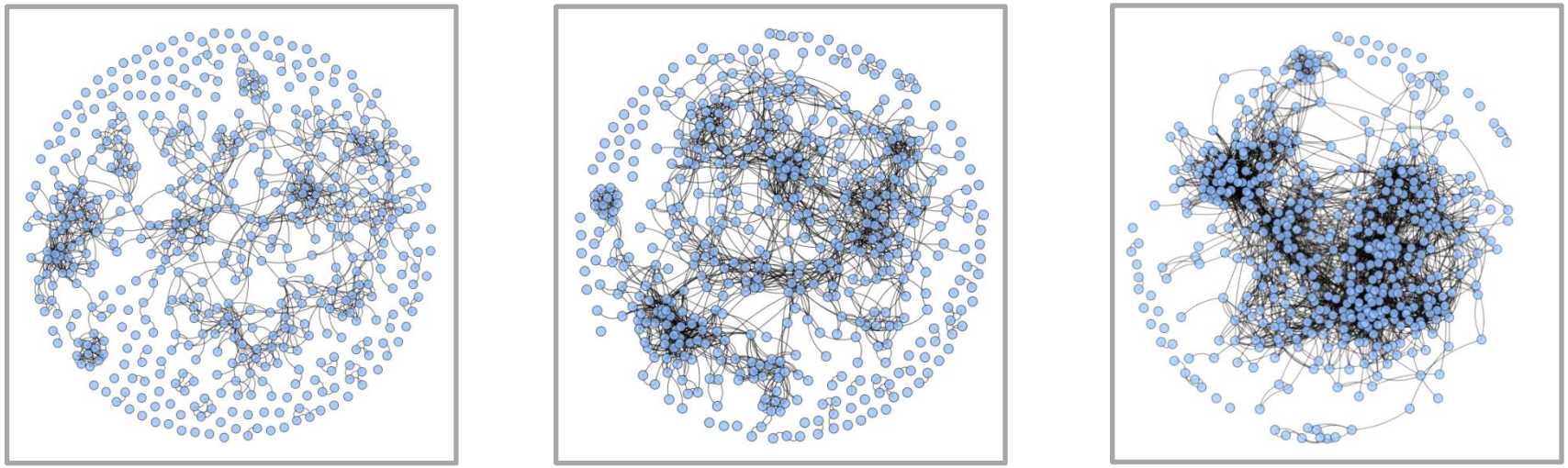
- Java universial network/graph framework (JUNG)

- Fruchterman-Reingold layout algorithm



THR-CSNs

- A Tc matrix can be transformed into many different CSNs
- Each CSN is associated with a specific similarity threshold

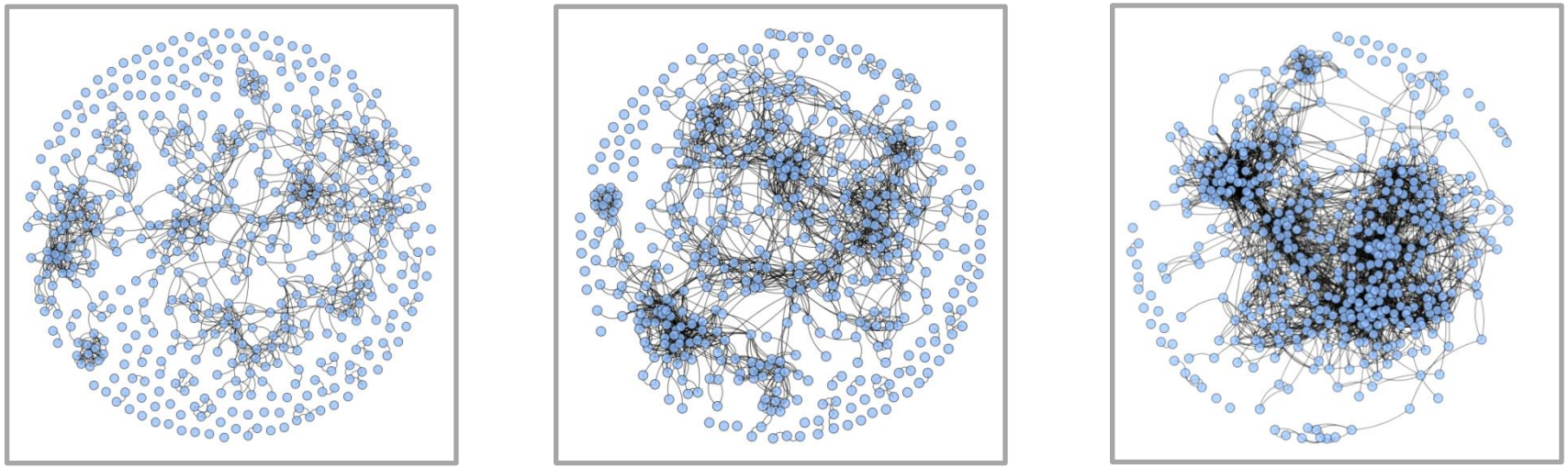


Increasing Tc threshold

THR-CSNs

- Threshold values and edge density are inversely related
- **Network properties strongly depend on edge density**

Increasing density



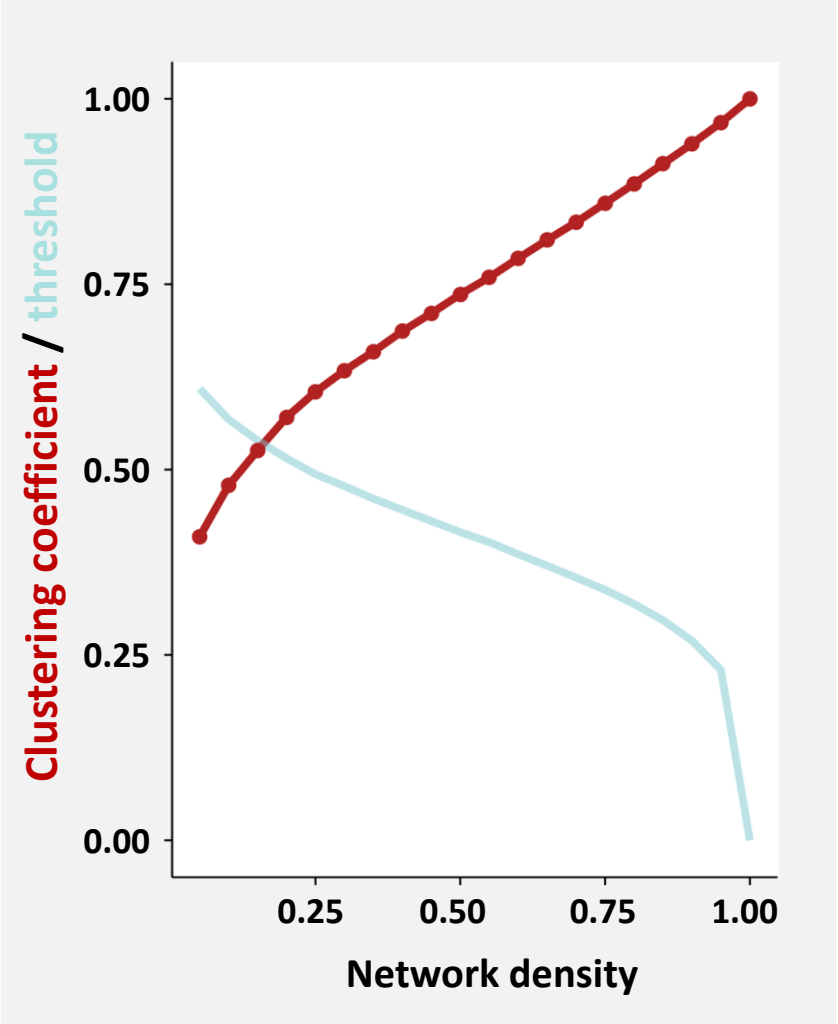
Increasing Tc threshold

Density Dependence of Network Properties

Increasing network density:

- Increase of
  - clustering coefficient

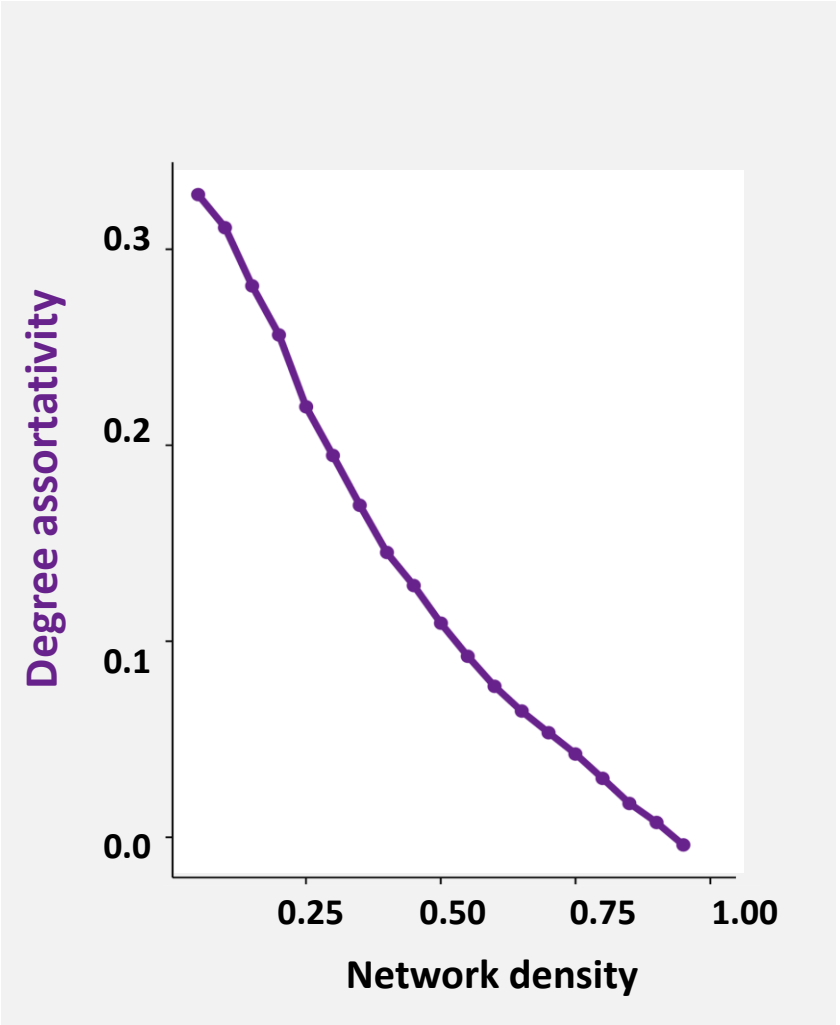
THR-CSN of 1000 randomly selected ZINC compounds (MACCS Tc)



Density Dependence of Network Properties

Increasing network density:

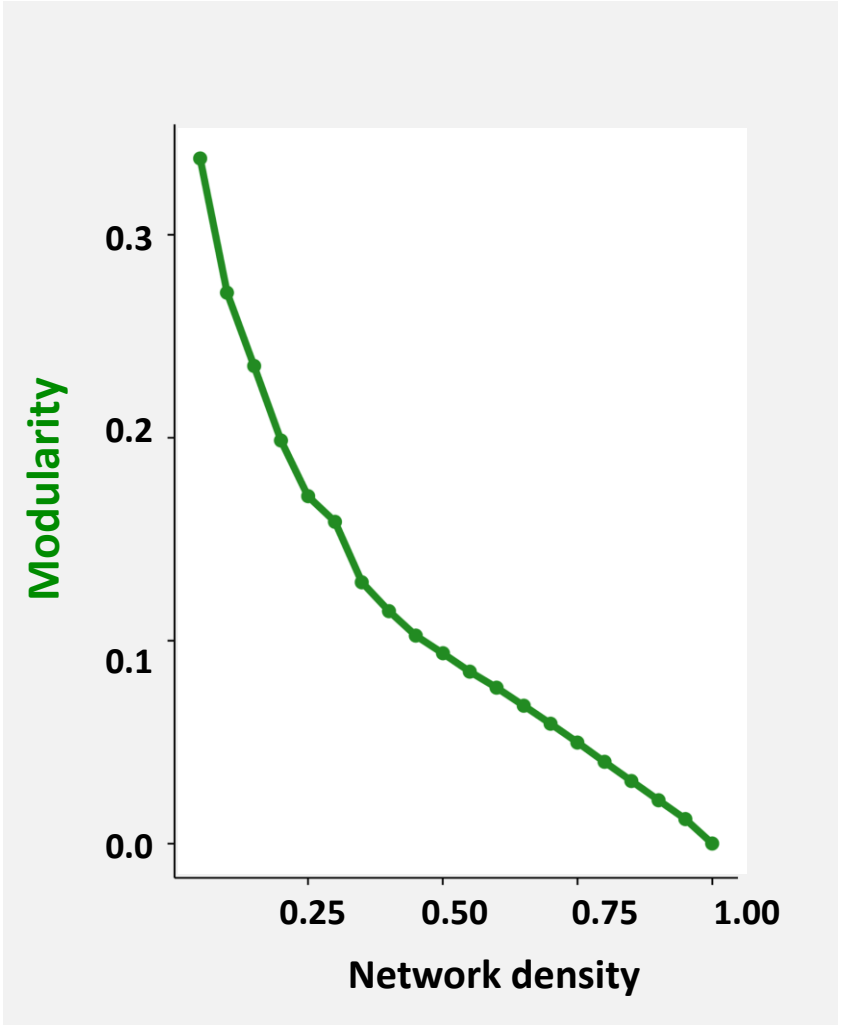
- Increase of
  - clustering coefficient
- Decrease of
  - degree assortativity



Density Dependence of Network Properties

Increasing network density:

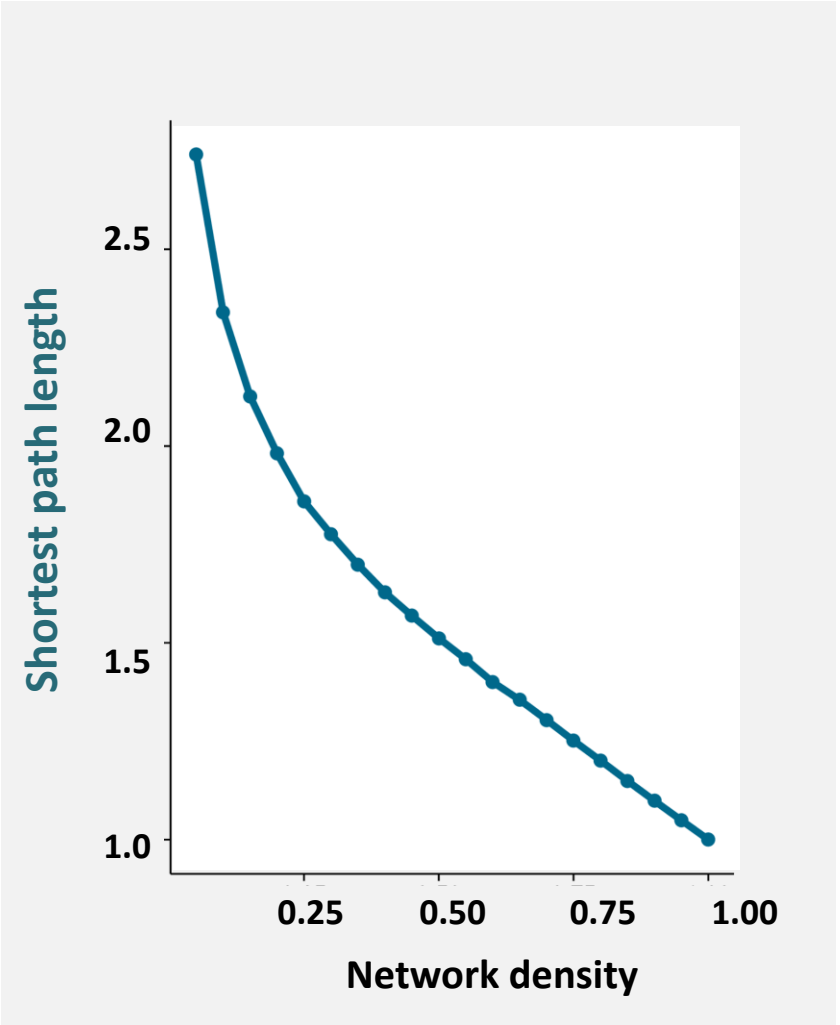
- Increase of
  - clustering coefficient
- Decrease of
  - degree assortativity
  - modularity



Density Dependence of Network Properties

Increasing network density:

- Increase of
  - clustering coefficient
- Decrease of
  - degree assortativity
  - modularity
  - shortest path length



Comparison of THR-CSNs

- THR-CSNs typically display **high modularity** and **assortativity** at **low network density**
- High modularity and assortativity characterize THR-CSNs with clear **compound community structures**
- THR-CSNs are difficult to compare at constant Tc values, due to **compound class-dependent similarity values**
- THR-CSNs are best compared at **constant low density**, e.g. 2.5%

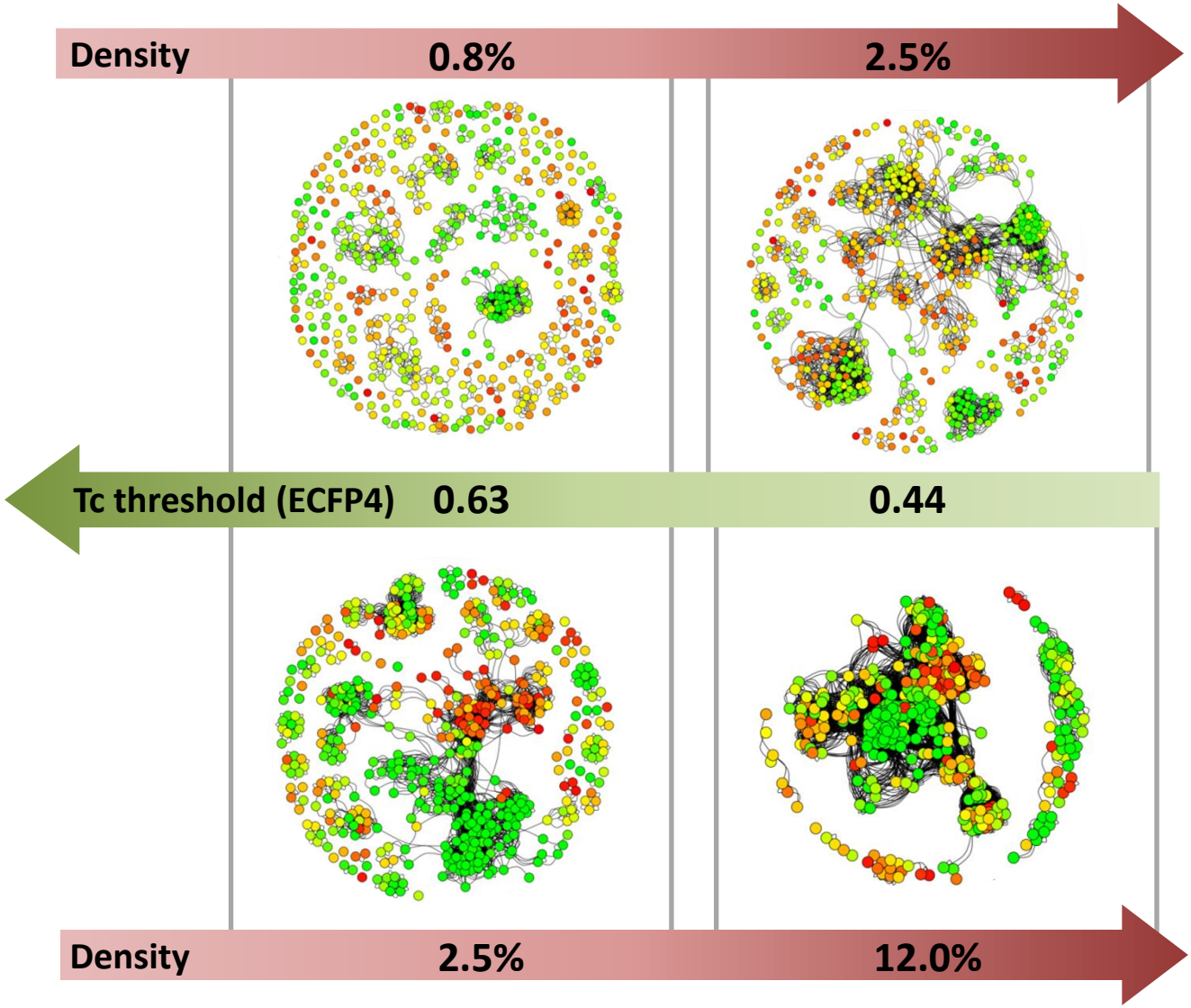
Comparison of THR-CSNs

Serotonin 7 (5-HT7) receptor ligands (**structurally diverse**)

THR-CSNs

Potency
- low
- medium
- high

Delta-type opioid receptor ligands (**homogeneous**)



THR-CSNs of Bioactive Compounds

- THR-CSNs
  - for **data sets of varying diversity** (120 sets of 1000 ZINC cpds)
  - for **bioactive compounds** (21 ChEMBL data sets, 522-973 cpds)
  - compared at constant edge density (2.5%)

Constant density: 0.025

Data sets:

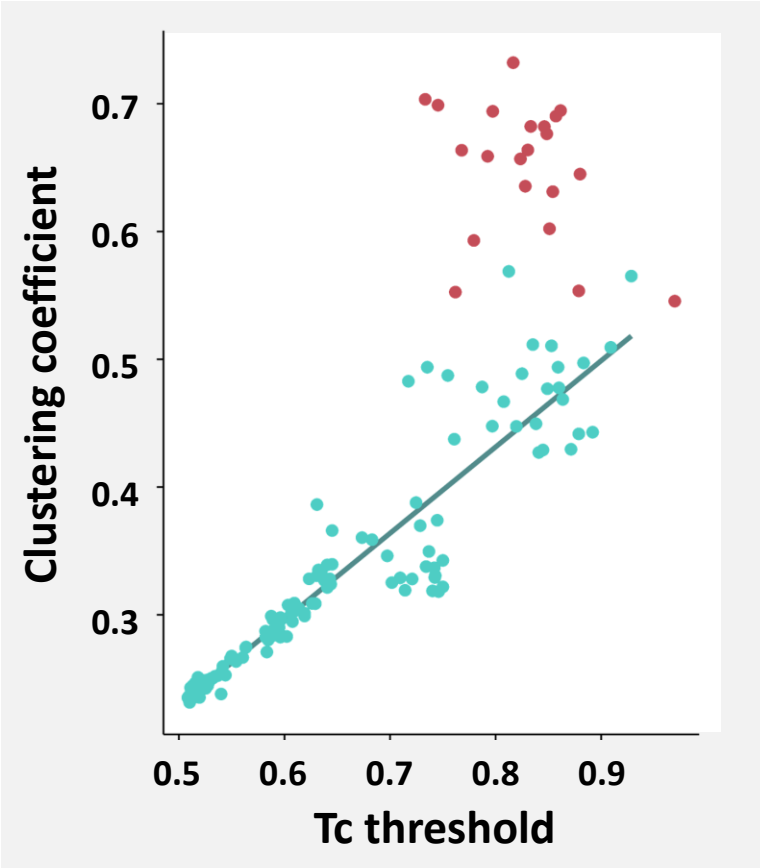
- ZINC data sets (120)
  - diverse (64)
  - random (10)
  - similar (46)
- ChEMBL data sets (21)

THR-CSNs of Bioactive Compounds

- At constant network density, bioactive compound CSNs have **larger clustering coefficients**

Constant density: 0.025

Data sets:
- ZINC data sets (120)
- ChEMBL data sets (21)

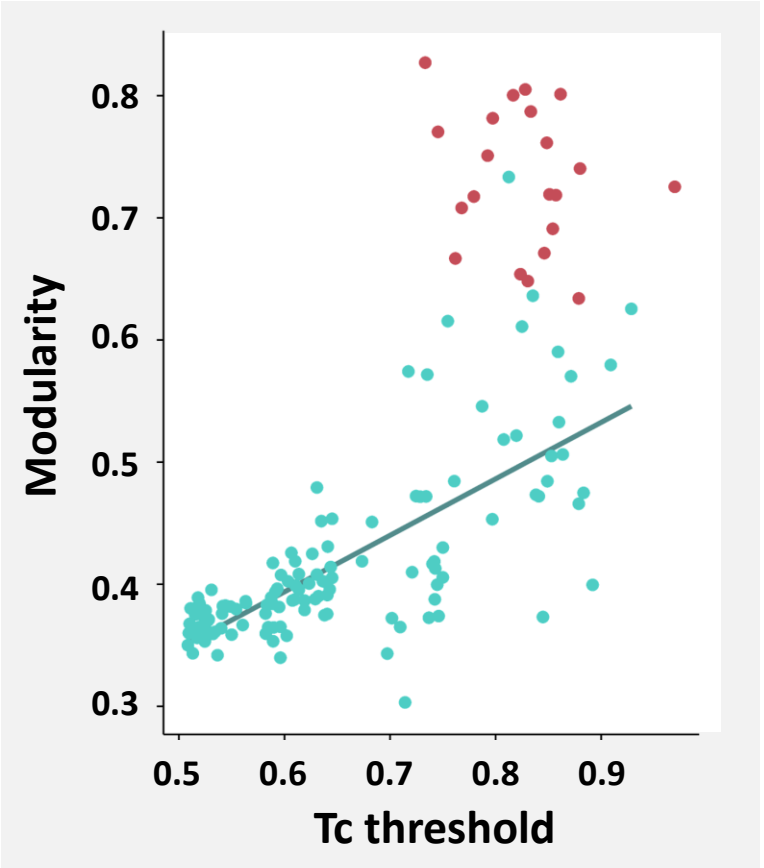


THR-CSNs of Bioactive Compounds

- At constant network density, bioactive compound CSNs have **higher modularity**

Constant density: 0.025

Data sets:
- ZINC data sets (120)
- ChEMBL data sets (21)

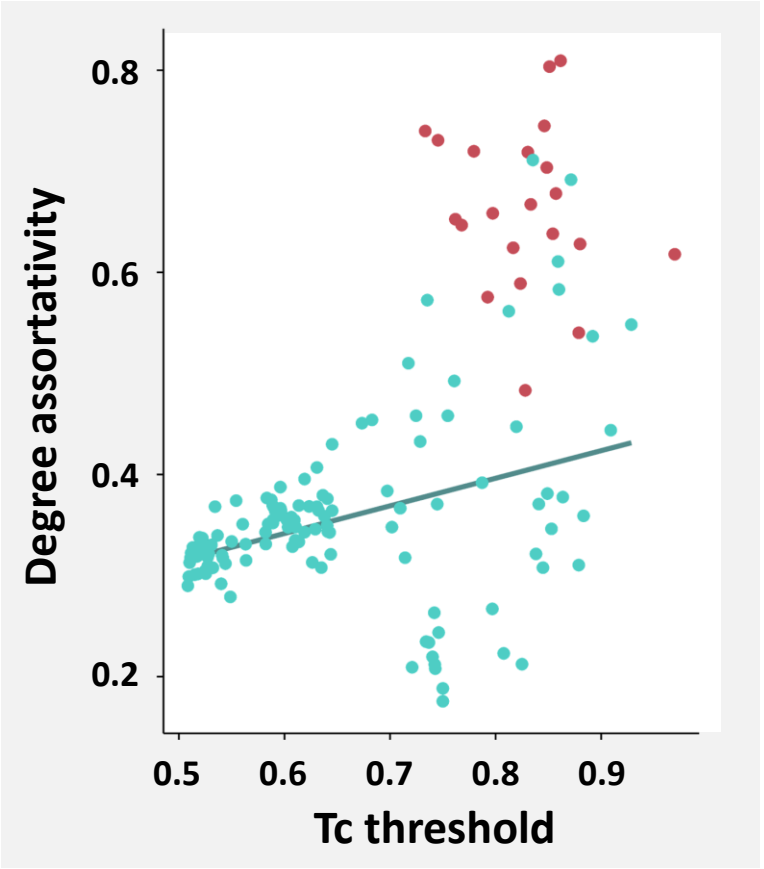


THR-CSNs of Bioactive Compounds

- At constant network density, bioactive compound CSNs have **higher assortativity**

Constant density: 0.025

Data sets:
- ZINC data sets (120)
- ChEMBL data sets (21)



THR-CSNs of Bioactive Compounds

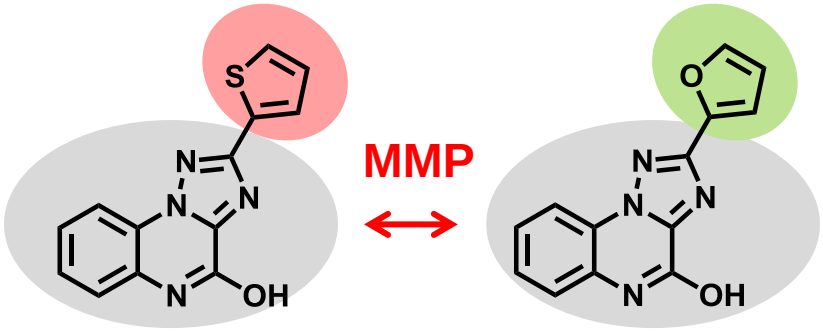
- Characteristics at low network density
  - large clustering coefficients
  - high assortativity
  - high modularity
  - extensive community structures

  - **homophily principle as a major determinant of THR-CSN topology** when charting biologically relevant chemical space (similar to social networks)
  - shared activity as a *latent characteristic*

Substructure-Based Similarity

- Alternative CSN representation designed by applying the matched molecular pair (**MMP**) formalism

- Formation of MMPs as a similarity criterion: **MMP-CSN**



MMP-CSNs vs. THR-CSNs

- **THR-CSN**
  - Tanimoto similarity
  - varying similarity threshold / varying density

- **MMP-CSN**
  - substructure-based similarity
  - constant density

- **CSN comparison**
  - MMP-CSN, determine edge density
  - THR-CSN, adjust threshold to match MMP-CSN density

Comparison of MMP- and THR-CSNs

- 154 activity classes (ChEMBL)
- Network property analysis / key findings
  - comparably high assortativity and modularity
  - surprisingly similar global topologies
  - community structures / small world character

Exemplary Comparison

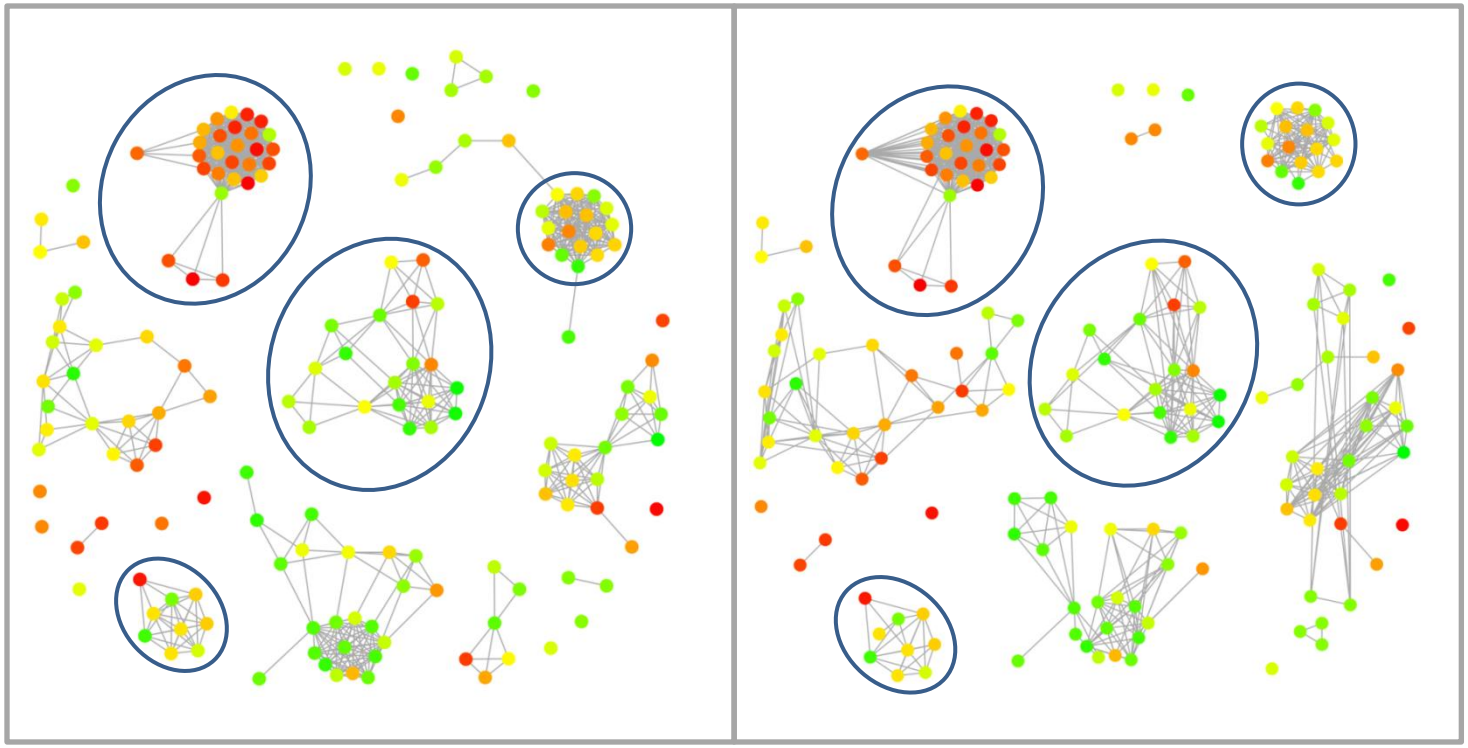
Assortativity	0.92	0.92
Cluster coefficient	0.71	0.69
Modularity	0.77	0.78
	MMP-CSN, Density: 5.2%	**THR-CSN**, Tc: 0.58, Density: 5.2%

Orexin receptor 2 ligands

pK_i

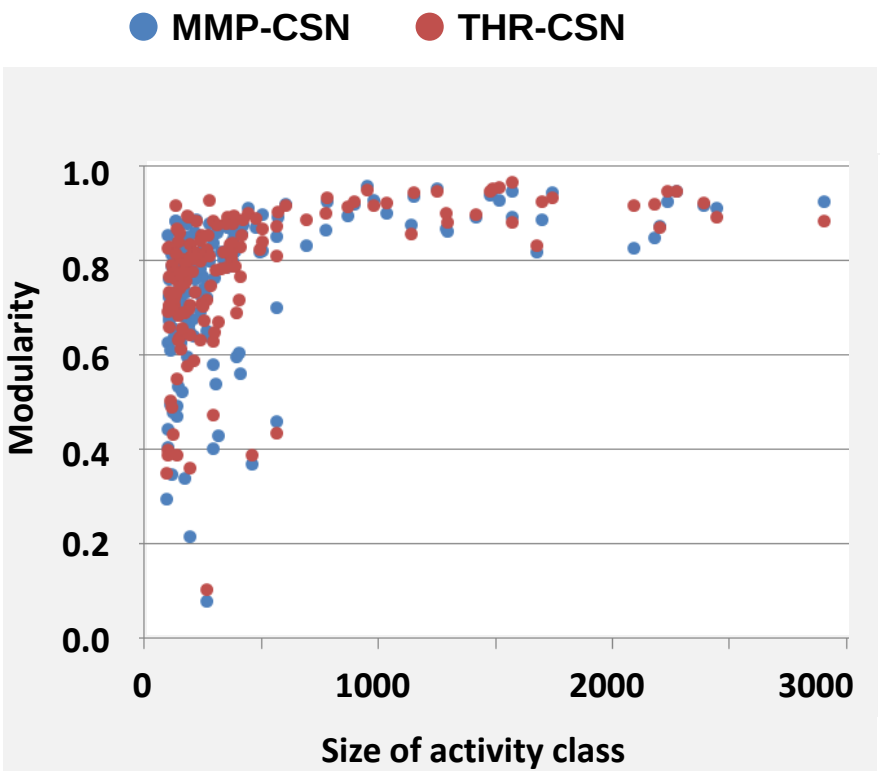
5.7

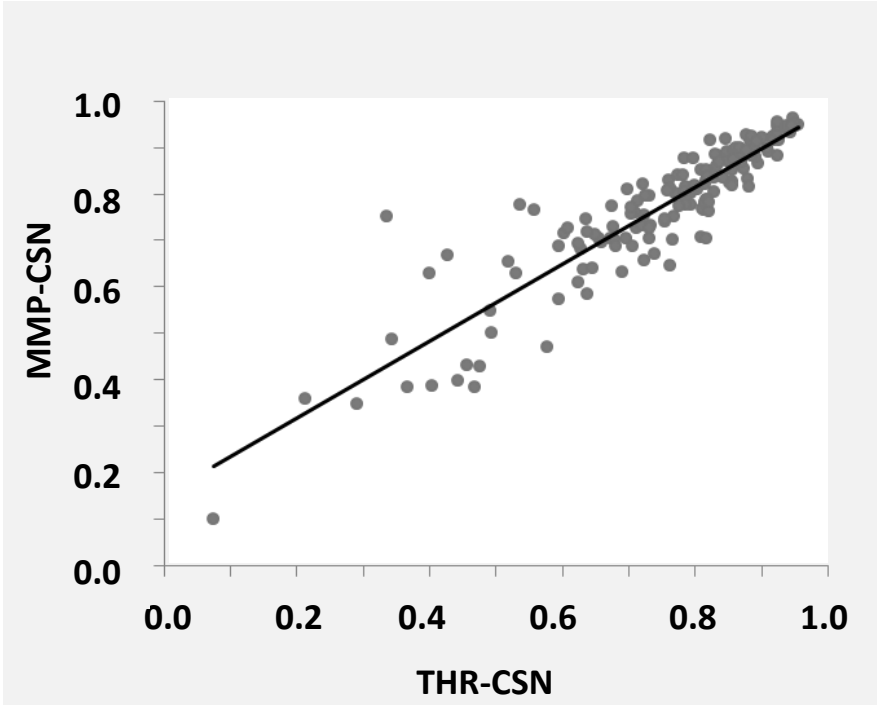
10.2



Comparison of MMP- and THR-CSNs

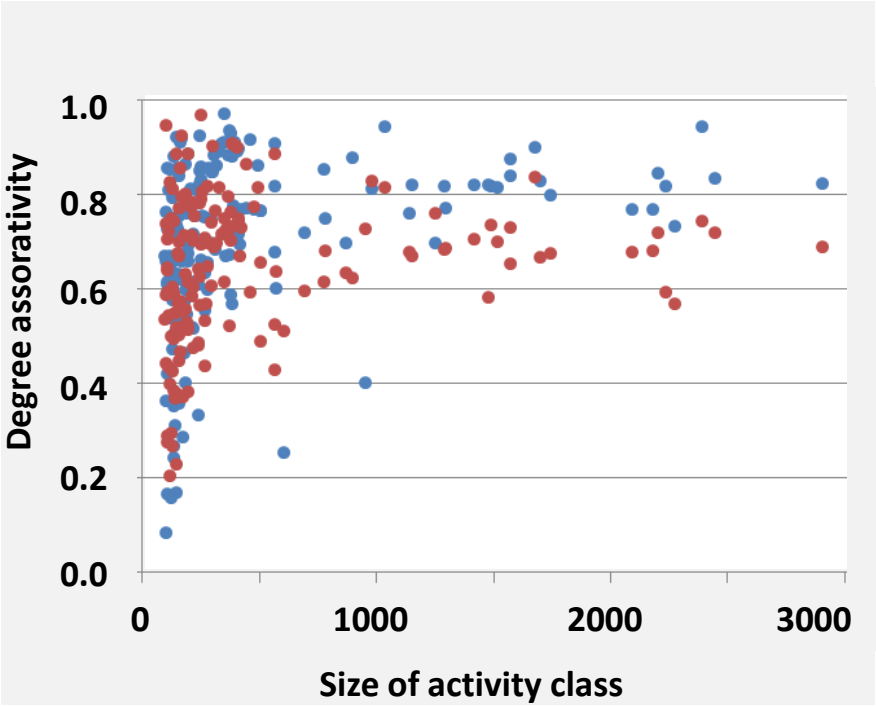
- **Modularity** (and **clustering coefficient**)
  - large values / **high correlation**

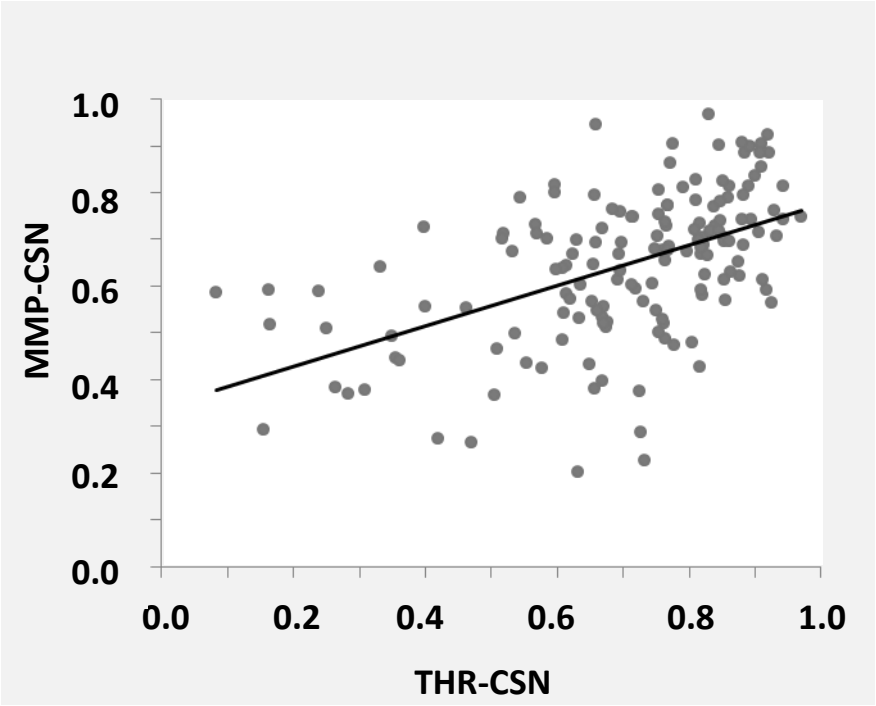




Comparison of MMP- and THR-CSNs

- **Assortativity**
  - large values / **low correlation**





Comparison of MMP- and THR-CSNs

- Network property analysis / key findings

  - **assortativity** as the major distinguishing feature

  - despite similar global topologies similarity relationships in compound communities systematically differ

Homophily principle influences THR- and MMP-CSNs in different ways

CSNs with Asymmetric Similarity

- Asymmetric similarity measures
  - assign different weights to features of *A* and *B*
  - comparison of *A to B* and *B to A* yields different values
  - directed similarity relationships

Tversky Index(Tv)

- **Tanimoto coefficient** (Tc)

$$Tc\ (A,B) = \frac{c}{a+b-c}$$

a: features of A
b: features of B
c: features of A and B

- **Tversky index** (Tv)

$$Tv\ (A,B,\alpha,\beta) = \frac{c}{\alpha(a-c)+\beta(b-c)+c}, \alpha,\beta \geq 0$$

- α and β are weighting factors for the distinguishing features of A and B, respectively

Normalization of Tv

- Tv can be normalized to enable single-parameter variation

$\alpha + \beta = k$ for an arbitary value $k > 0$

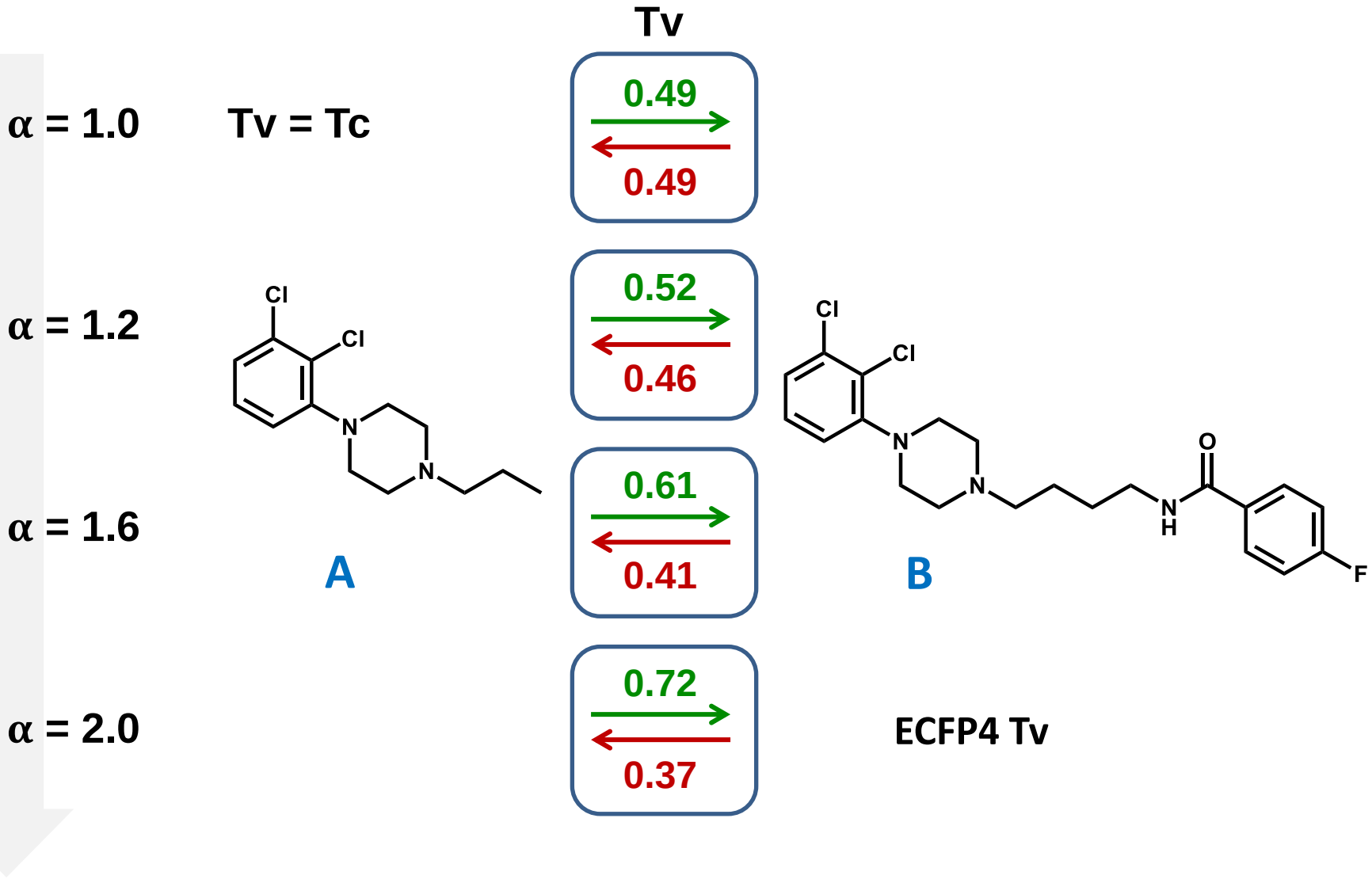
- Tv can be expressed using the single parameter α

$$\mathrm{Tv}_k(\mathrm{A},\mathrm{B},\alpha) = \mathrm{Tv}(\mathrm{A},\mathrm{B},\alpha,k-\alpha)$$

$$= \frac{c}{a(a-c)+(k-a)(b-c)+c}, a \hat{\mathrm{I}} \ [0,k]$$

$k = 2$: Tv becomes Tc if equal weights of 1 are put on A and B

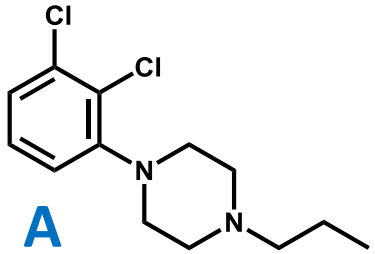
Asymmetry Through α Variation



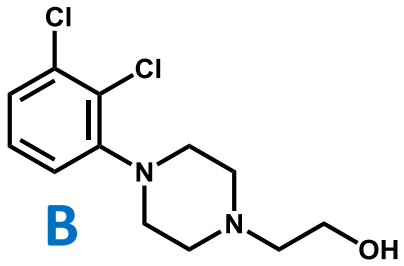
TV-CSN: Asymmetric Threshold CSN

α = 2.0 Tv threshold: 0.60 **Edge construction rules**

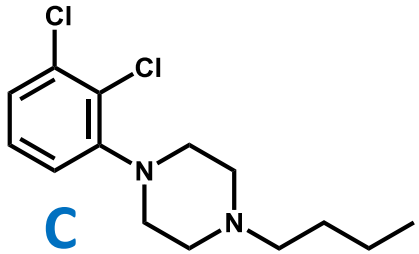
Tv



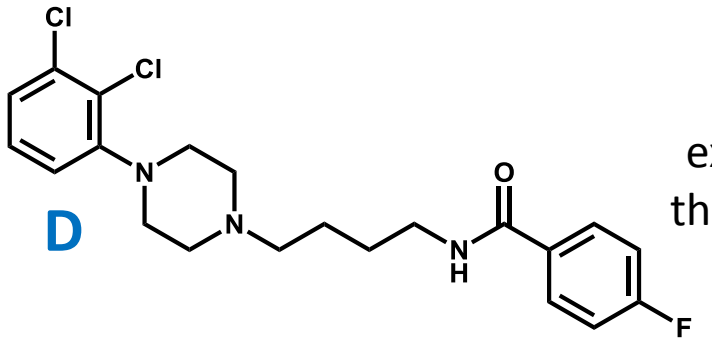
A



B



C



D

0.72 → / ← 0.72

0.88 → / ← 0.78

0.72 → / ← 0.37

equal values

both exceed threshold → 0.88

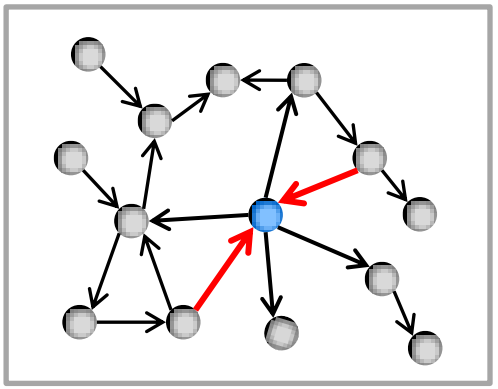
one exceeds threshold → 0.72

TV-CSN Asymmetry

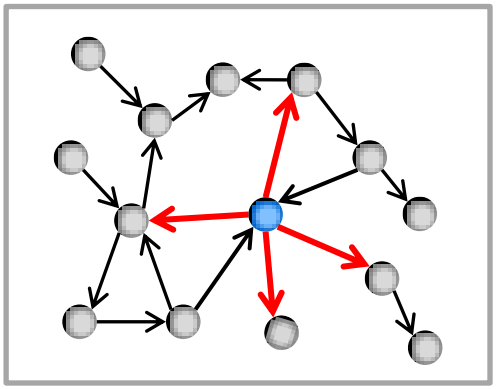
Edges are **directed**

Nodes have **in-** and **out-degrees**

Node **in**-degree: 2



Node **out**-degree: 4

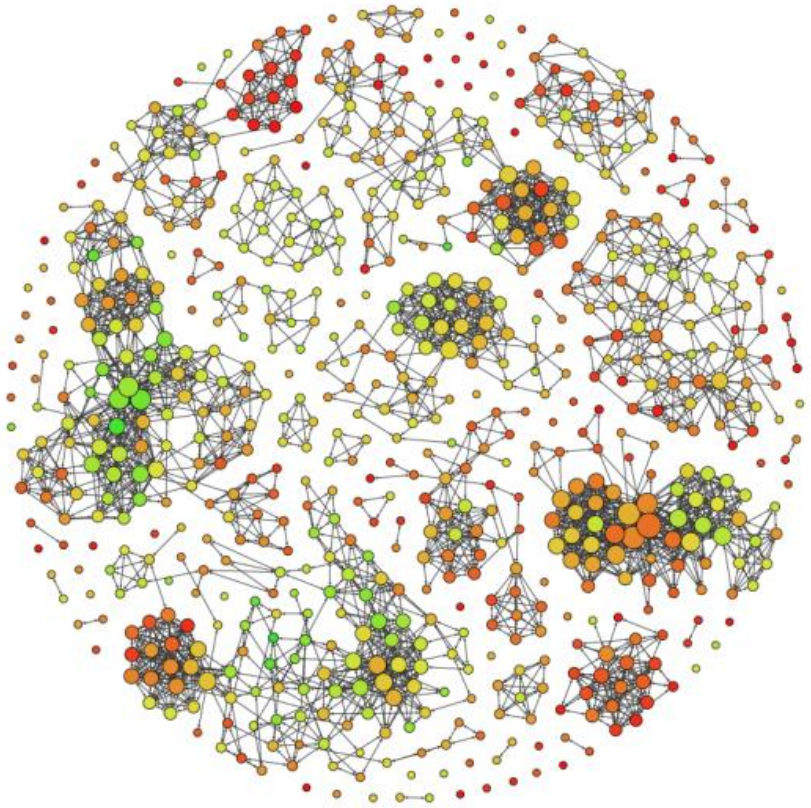


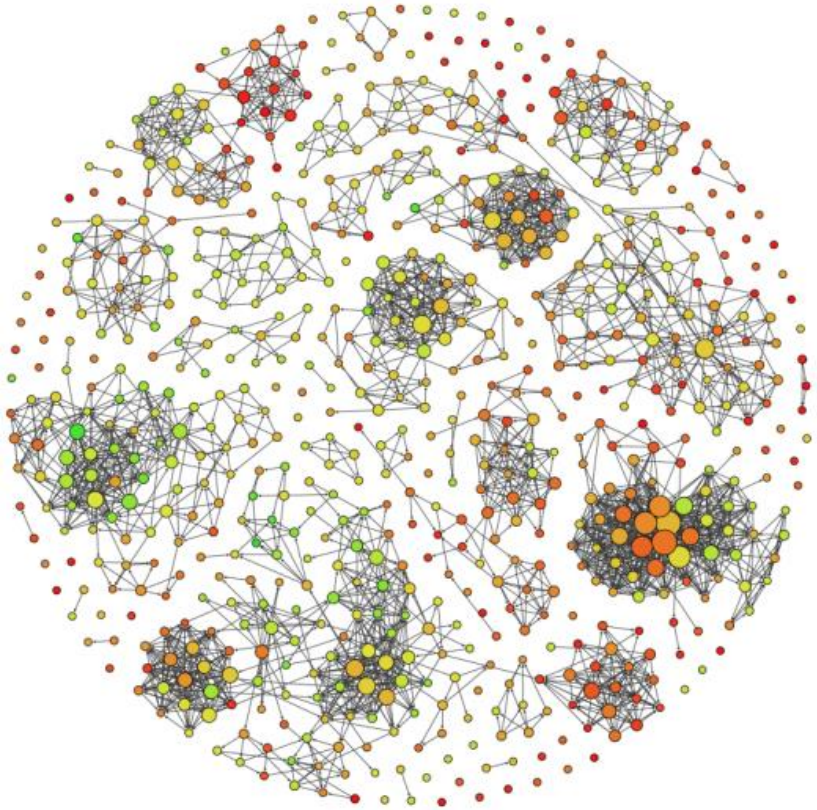
THR-CSN vs. TV-CSNs

Transforming a THR-CSN into a TV-CSN

Histamine H4 receptor ligands 5.0 10.4 pK_i

α = **1.0** (THR-CSN) **1.2**



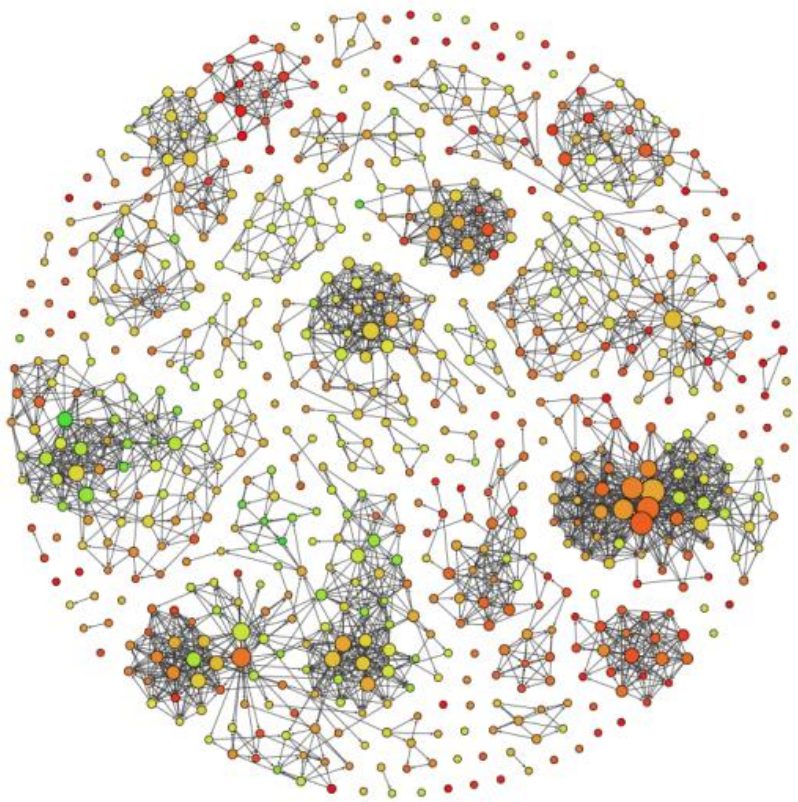


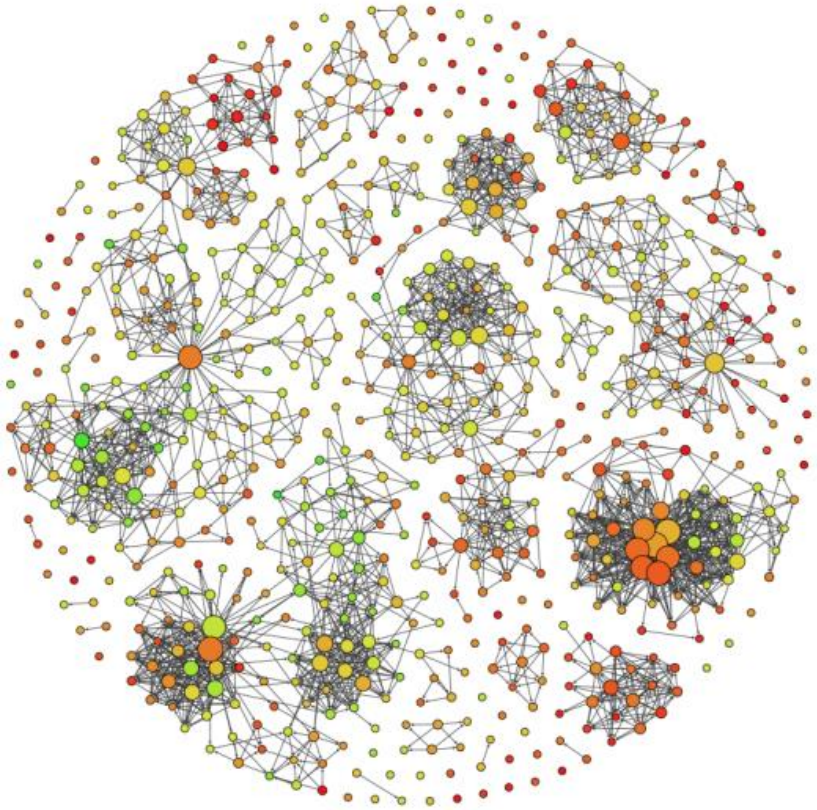
THR-CSN vs. TV-CSNs

Increasing asymmetry of similarity

Histamine H4 receptor ligands 5.0 10.4 pK_i

α = **1.6** **2.0**

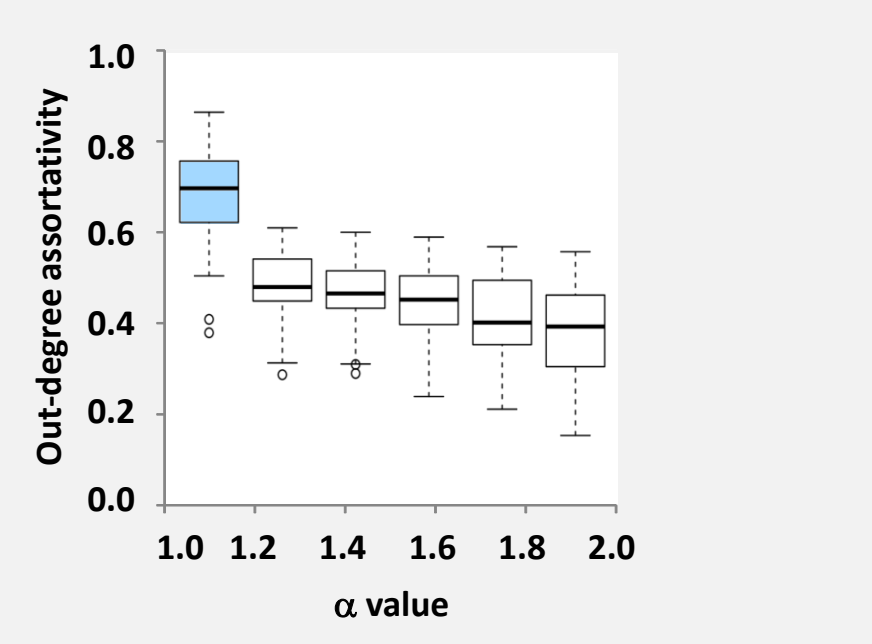




Network Properties: THR- vs. TV-CSNs

- **Major difference** - decrease in out-degree assortativity with increasing asymmetry

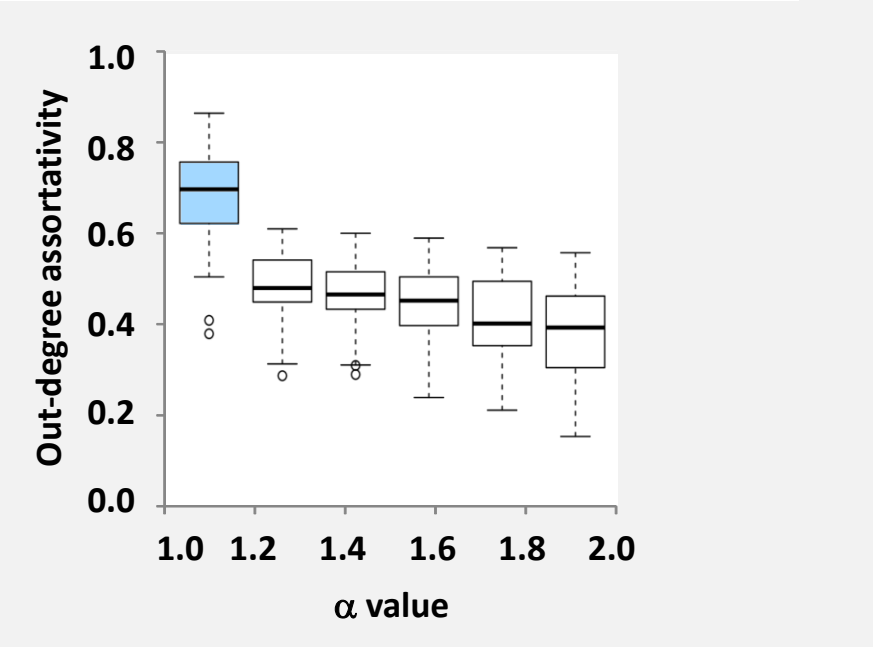
THR- vs. TV-CSNs with varying α settings for 36 activity classes (ChEMBL)



Network Properties: THR- vs. TV-CSNs

- Successive formation of nodes with uneven degrees

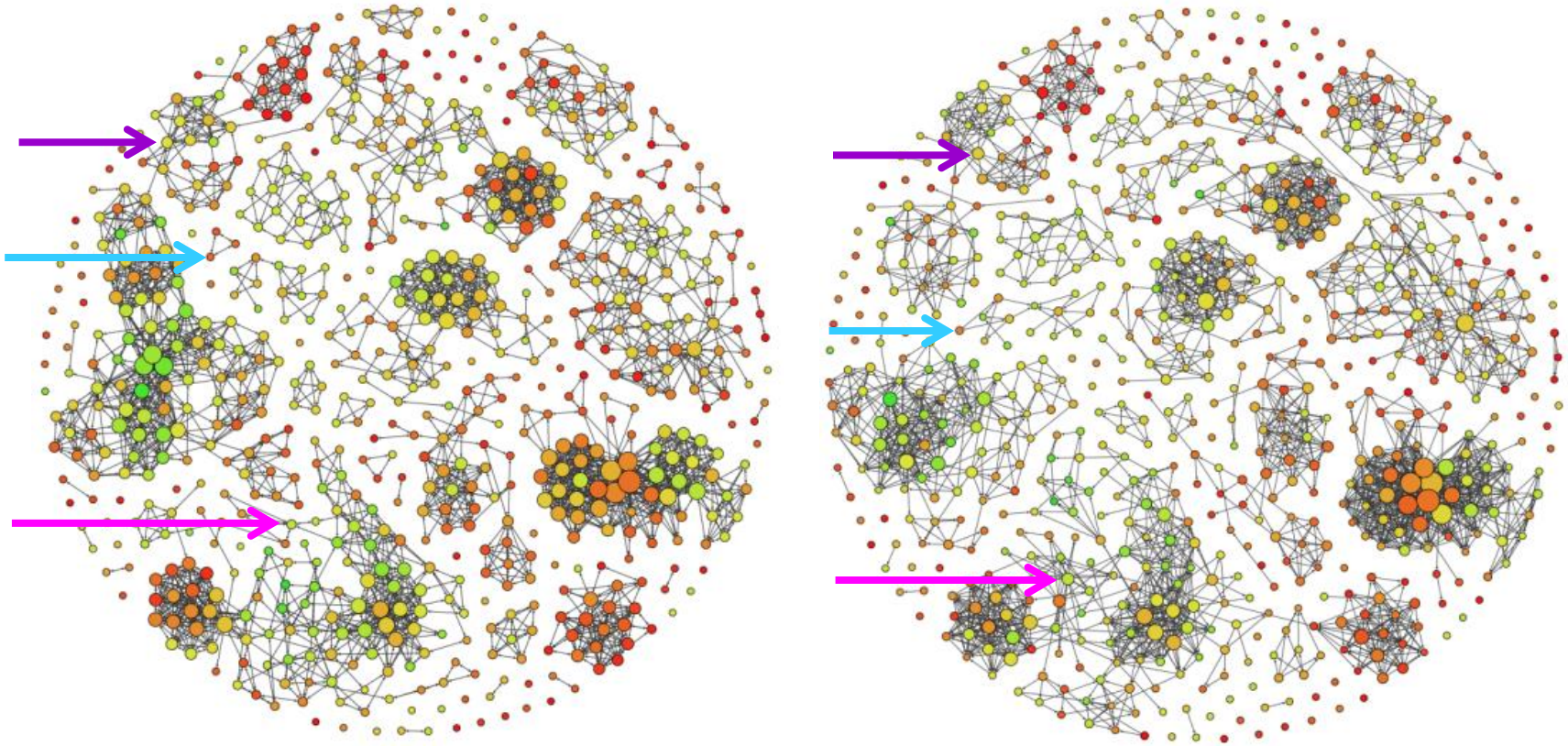
- Emergence of **hubs** in TV-CSNs



Emergence of Hubs in TV-CSNs

Nodes are **scaled in size** according to their **out-degrees**

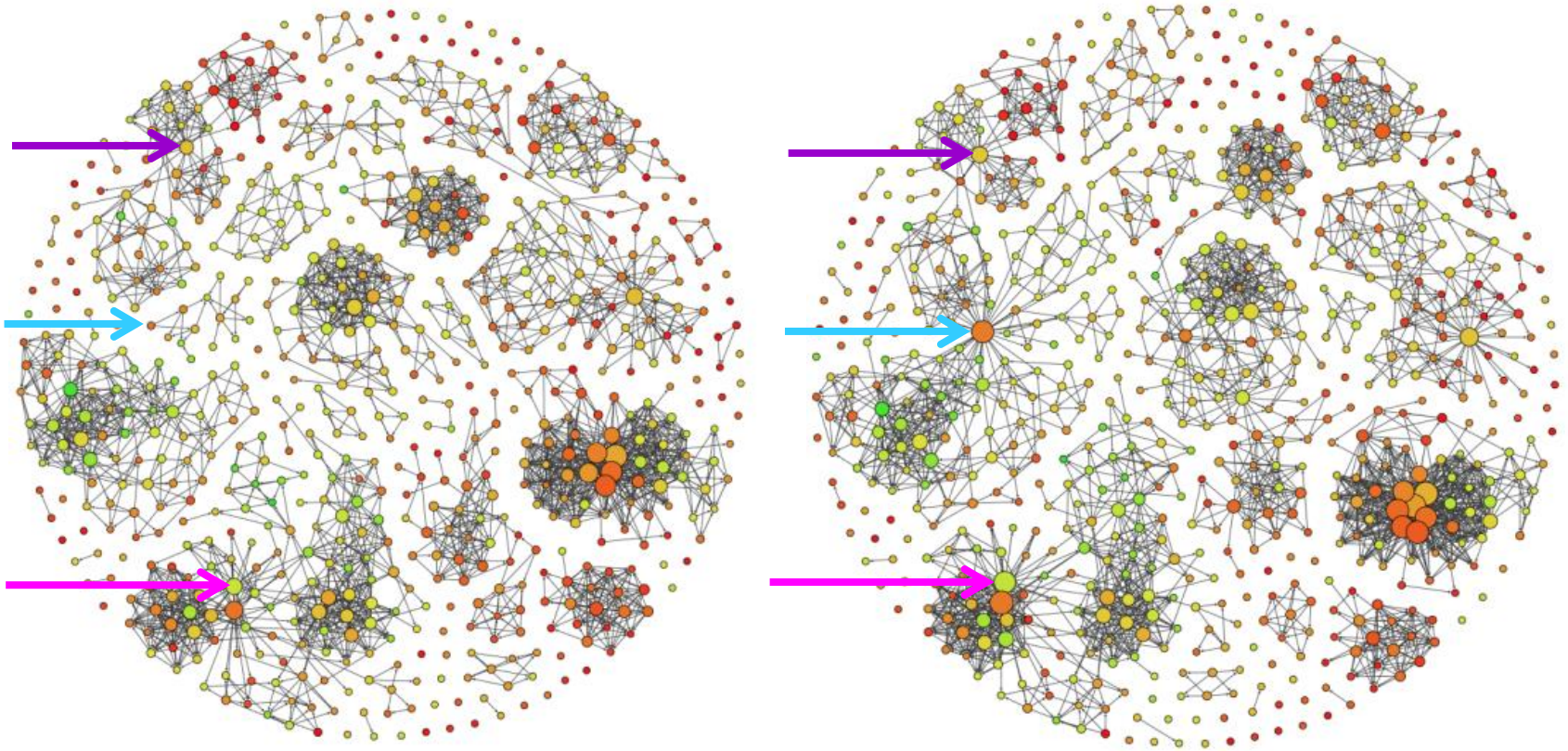
α = **1.0 (THR-CSN)** → **1.2**



Emergence of Hubs

Histamine H4 receptor ligands 5.0 10.4 pK_i

α = 1.6 2.0



Asymmetry and Scale-Free Nature

- **Hubs** often indicate **scale-free network character**
- In scale-free networks, the degree distribution follows a power law:

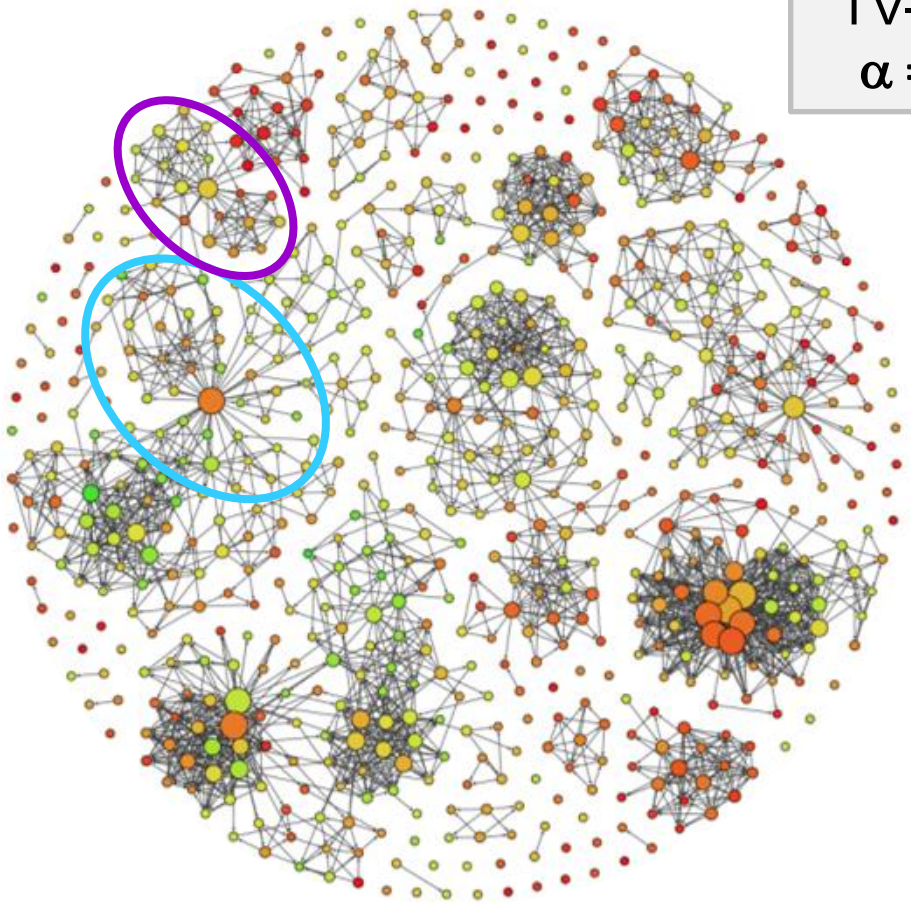
$$p(k) \propto k^{-\gamma}$$

  - γ: constant with values of $2 \leq \gamma \leq 3$ for scale-free networks

Number of TV-CSNs of 36 activity classes fitting a power law with $2 \leq \gamma \leq 3$

	α value					
	1.0	**1.2**	**1.4**	**1.6**	**1.8**	**2.0**
TV-CSNs	2	9	9	11	16	17

Hubs as Focal Points of SAR Analysis



TV-CSN
α = 2.0

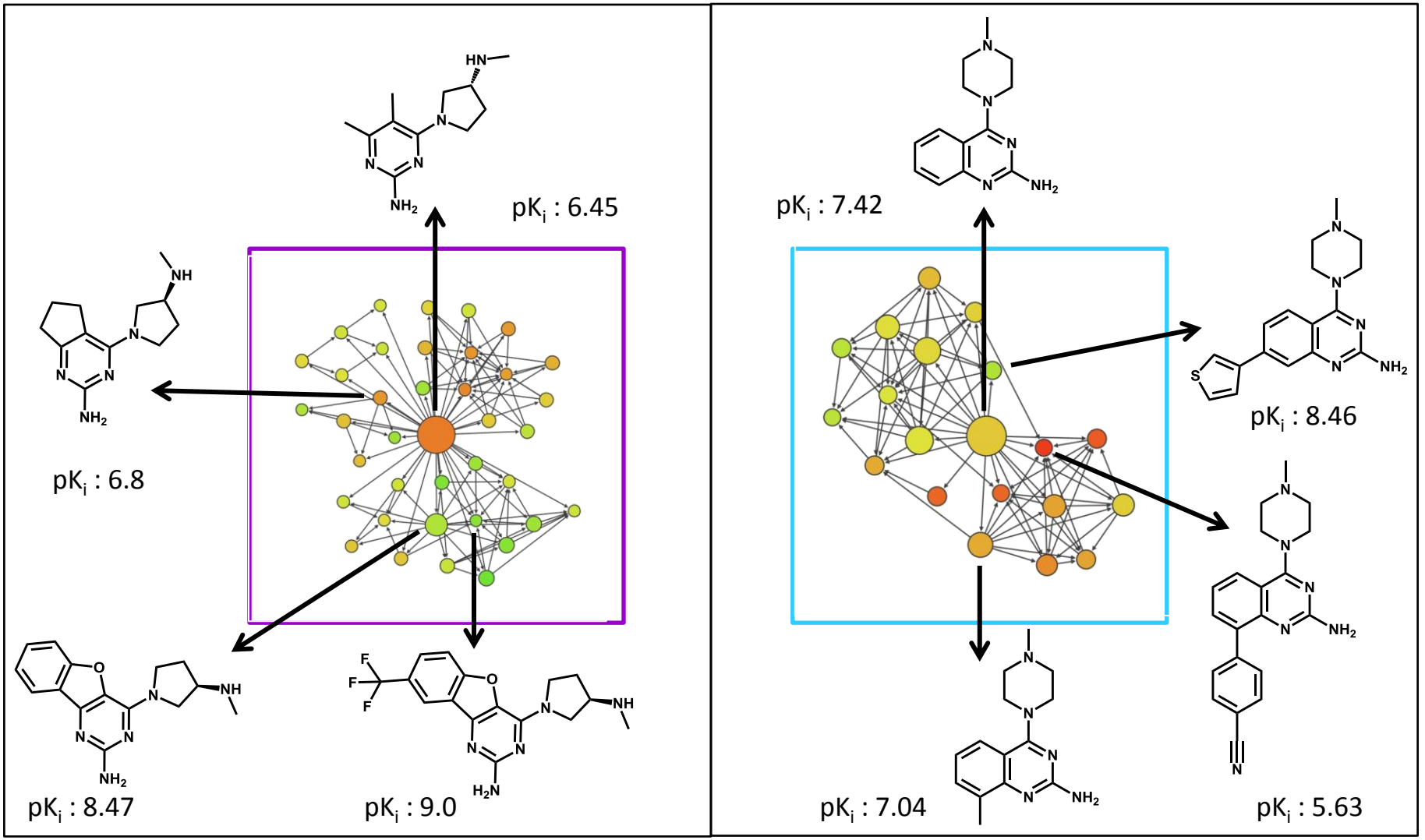
Histamine H4 receptor ligands

5.0 10.4

pK_i

Nodes are **scaled in size** according to their **out-degrees**

Hubs as Focal Points of SAR Analysis



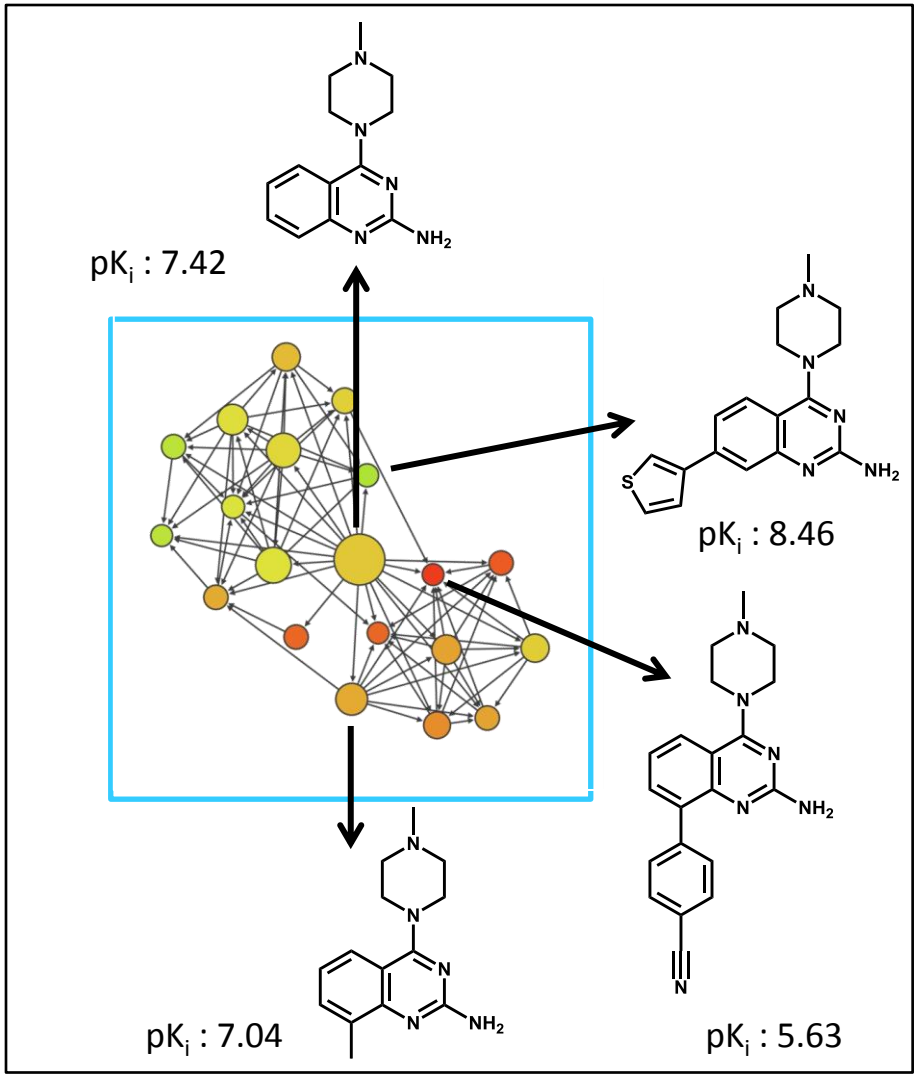
Hubs as Focal Points of SAR Analysis

From **hubs**

(i) pathways with compounds of increasing size can be traced

(ii) potency progression can be monitored

Lead optimization scenario



Conclusions

- Chemical space networks
  - paradigm for coordinate-free chemical space representation
  - characterization using statistical concepts from network science
  - designed for analyzing active compounds and SARs

- THR-CSNs
  - homophily principle as a major determinant of CSN topology
  - CSNs are best studied and compared at constant low edge density
  - THR-CSNs of random and bioactive compound samples are distinct

Conclusions

- MMP-CSNs
  - substructure-based similarity relationships
  - THR- and MMP-CSNs have similar topologies and small world character
  - homophily principle affects THR- and MMP-CSNs in different ways

- TV-CSNs
  - asymmetric similarity relationships
  - emergence of hubs and scale-free character
  - pathways of compounds of increasing size/complexity centered on hubs

b-it LSi
Life
Science
Informatics

LIMES
Life & Medical Sciences Institute

Acknowledgment

Magdalena Zwierzyna
Bijun Zhang
Mengjun Wu
Dagmar Stumpfe
Gerald Maggiora

universitätbonn