# Integrative Topology Uncovers New Biology from Heterogeneous Omics Data

**Nataša Pržulj, PhD, MAE**

ICREA Research Professor
Barcelona Supercomputing Center

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

# Overview

**Medicine: complex world of inter-connected entities**

1. **Motivation**
2. **New Methods – Examples:** mine inter-connected data

   i. **<u>Single type of omics data</u>:**
   - **Molecular networks**
   - **Multi-scale organization**

   } → function, disease

   ii. **<u>Multiple layers of heterogeneous data:</u>**
   - **iCell**
   - **Patient-centered data integration → Precision medicine**
     - ✓ Stratification, biomarker discovery, drug repurposing
   - **Disease re-classification, GO reconstruction, Network alignment, …**
3. **Conclusions**

# Overview

**Medicine: complex world of inter-connected entities**

1. **Motivation**
2. **New Methods – Examples:** mine inter-connected data

   i. **<u>Single type of omics data</u>:**
   - **Molecular networks**
   - **Multi-scale organization**

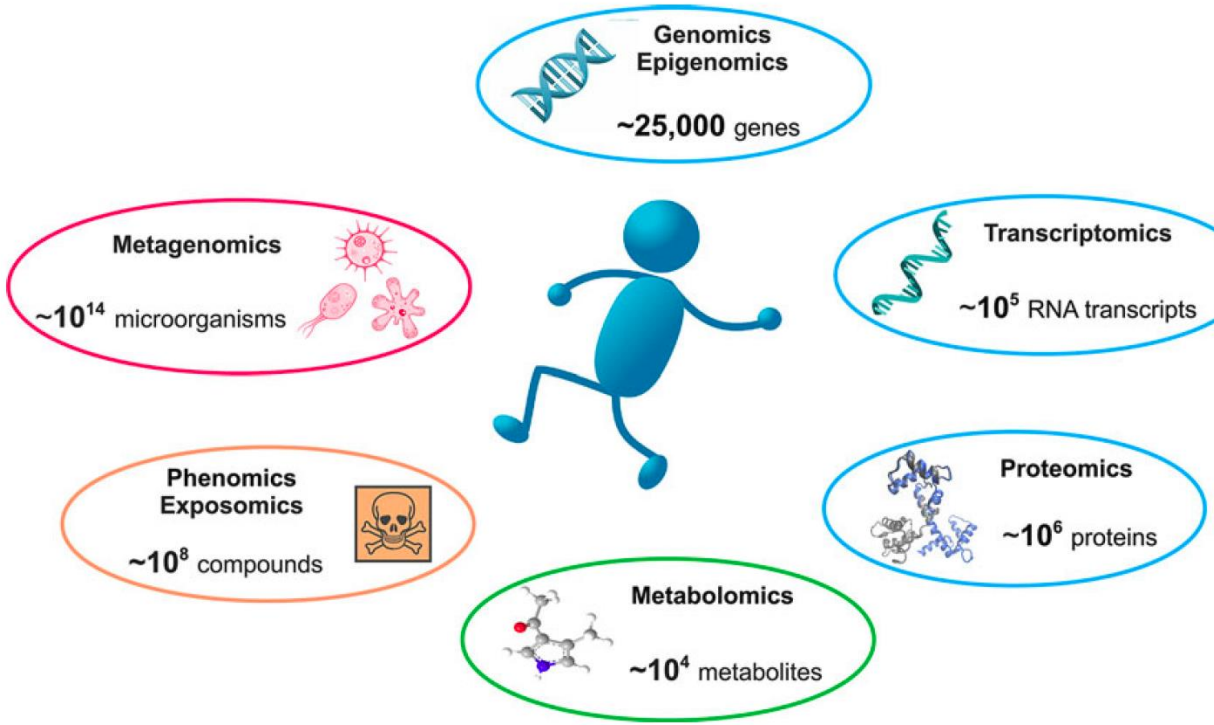
   } → function, disease

   ii. **<u>Multiple layers of heterogeneous data:</u>**
   - **iCell**
   - **Patient-centered data integration → Precision medicine**
     - ✓ Stratification, biomarker discovery, drug repurposing
   - **Disease re-classification, GO reconstruction, Network alignment, …**
3. **Conclusions**

# 1. Motivation
## Medicine: complex world of inter-connected entities

**Technological advances →**
**astounding harvest of various molecular and clinical data**

*Proteomics* 2016, *16*, 741–758



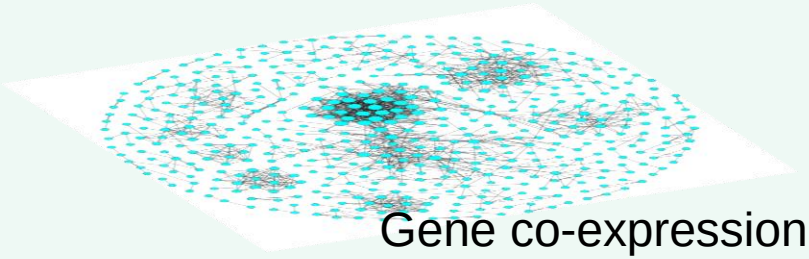

REVIEW

**Integrative methods for analyzing big data in precision medicine**

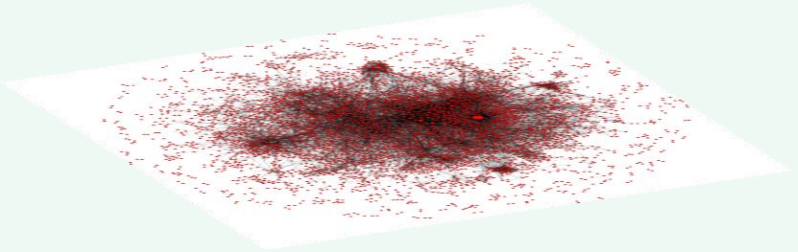
*Vladimir Gligorijević, Noël Malod-Dognin and Nataša Pržulj*

# 1. Motivation
## Medicine: complex world of inter-connected entities

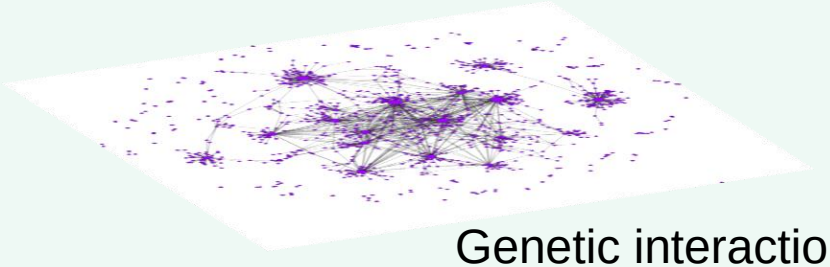
### Computational challenges



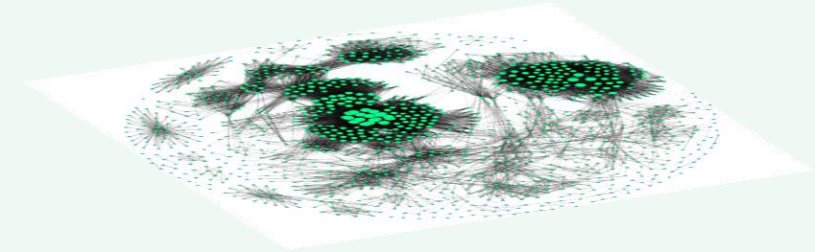
Gene co-expression



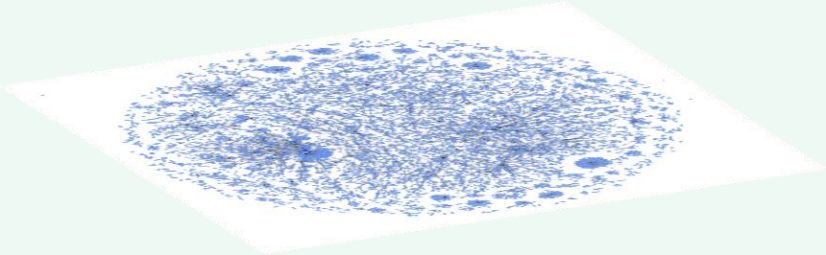
Protein-protein interactions (PPI)



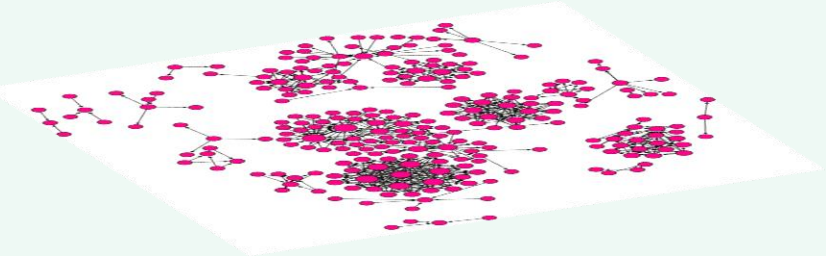
Genetic interactions



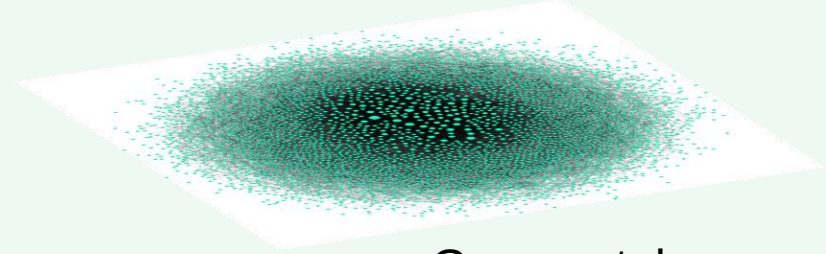
Metabolic interactions



Drug side-effects



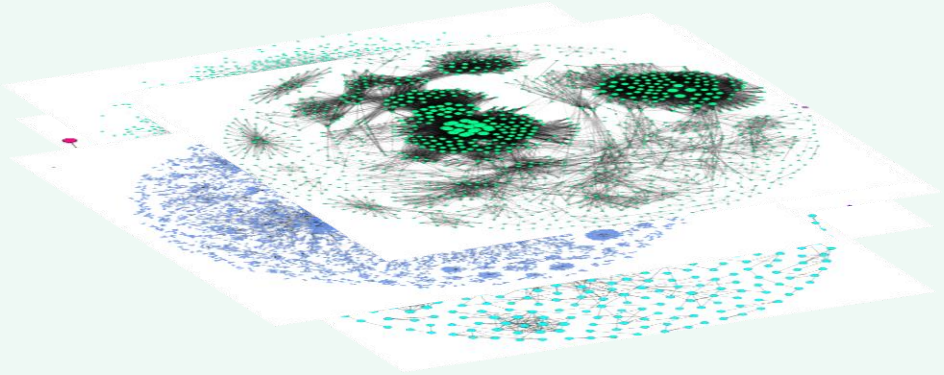
Disease ontology



Gene ontology

# 1. Motivation
## Medicine: complex world of inter-connected entities

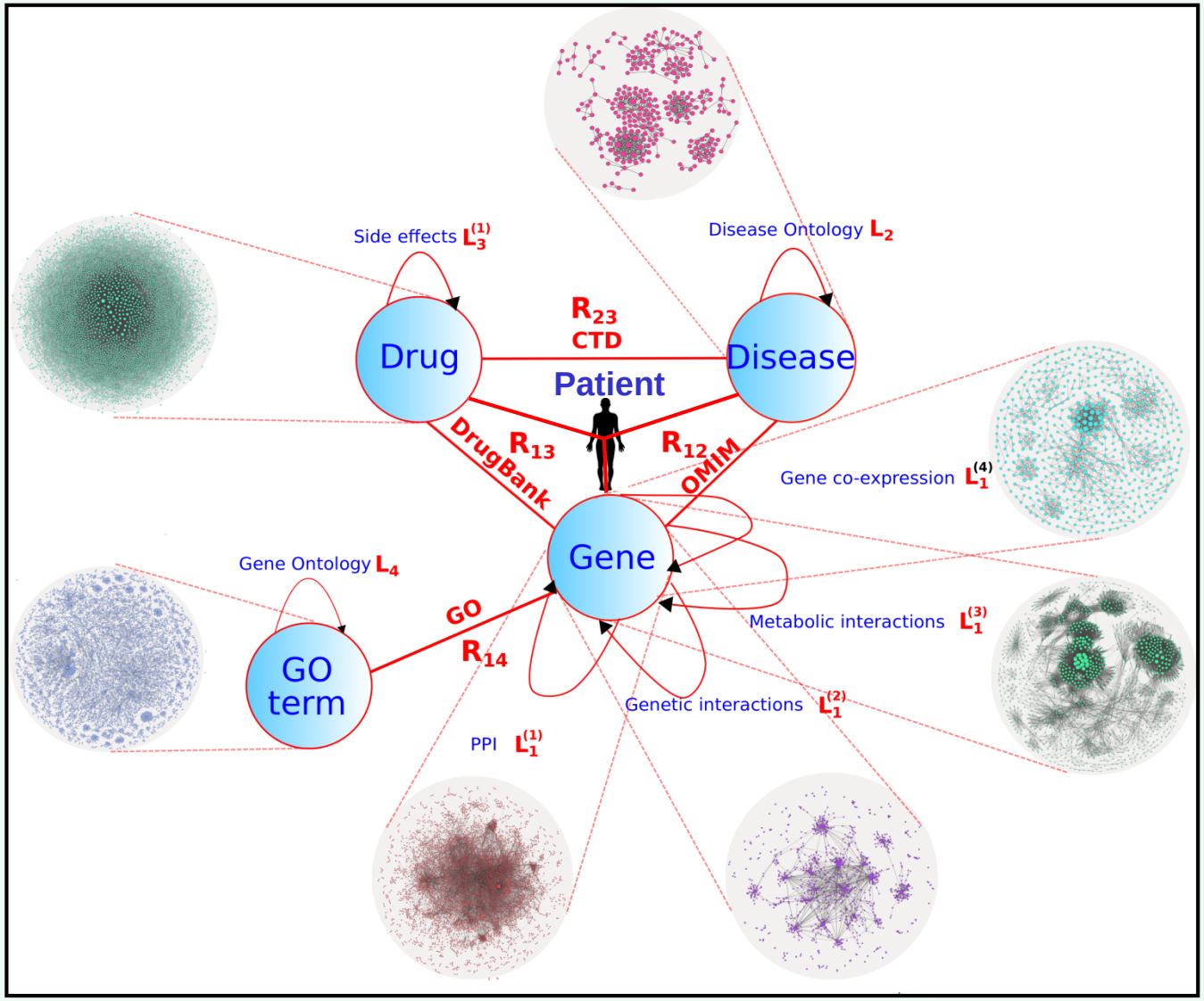
### Computational challenges



# 1. Motivation

## Medicine: complex world of inter-connected entities

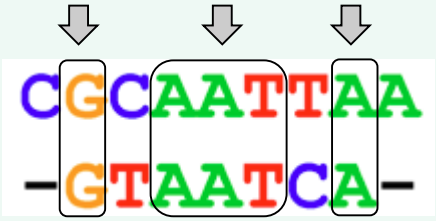
### Computational challenges



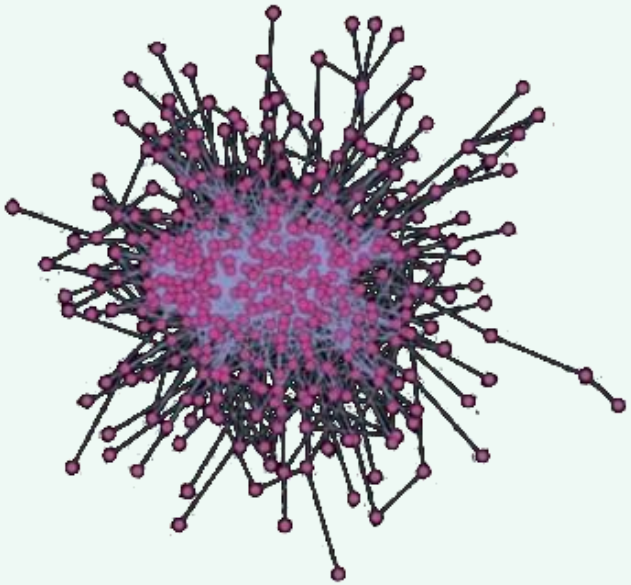
# 1. Motivation
## Medicine: complex world of inter-connected entities

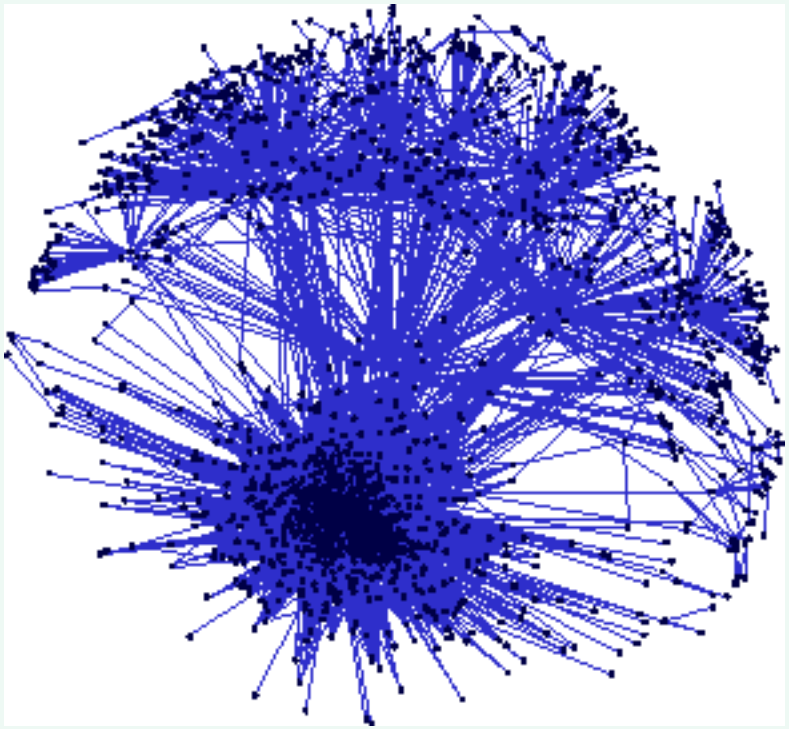
### Computational challenges

- Need new tools to mine complex data systems
- Why?
  - Analysing sequences: “computationally easy” → still lacking
  - Analysing interconnected heterogeneous data: “computationally hard”




- Sophisticated methods carefully tuned to extract new knowledge from particular data

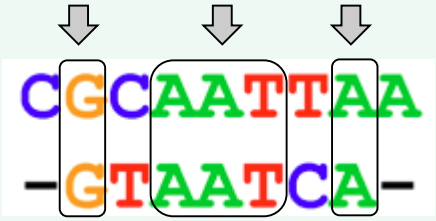




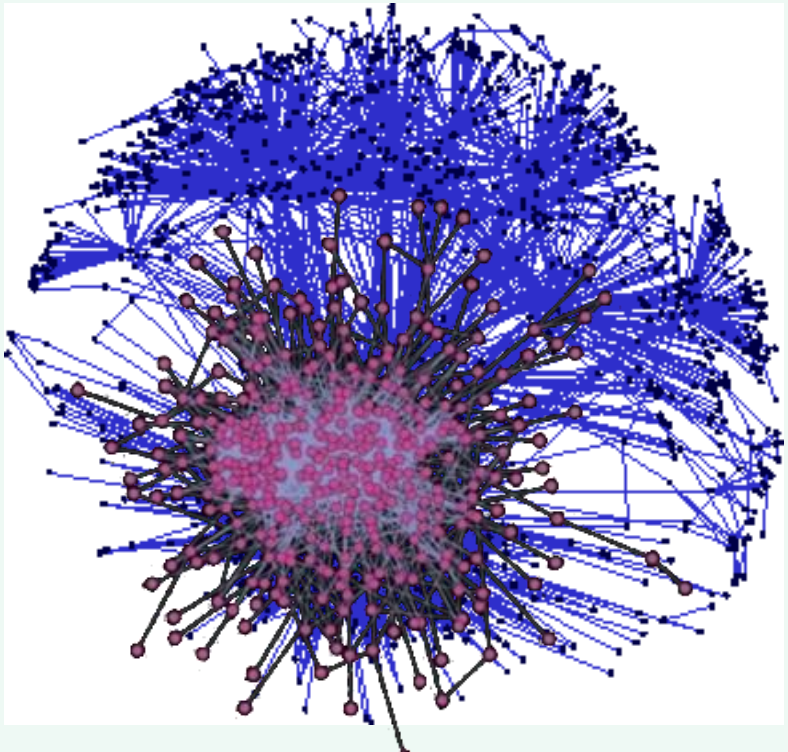
# 1. Motivation
## Medicine: complex world of inter-connected entities

### Computational challenges

- Need new tools to mine complex data systems
- Why?
  - Analysing sequences: “computationally easy” → still lacking
  - Analysing interconnected heterogeneous data: “computationally hard”




- Sophisticated methods carefully tuned to extract new knowledge from particular data



# Overview

**Medicine: complex world of inter-connected entities**

1. **Motivation**
2. **New Methods – Examples:** mine inter-connected data

   i. **Single type of omics data:**
   - **Molecular networks**
   - **Multi-scale organization**

   } → function, disease

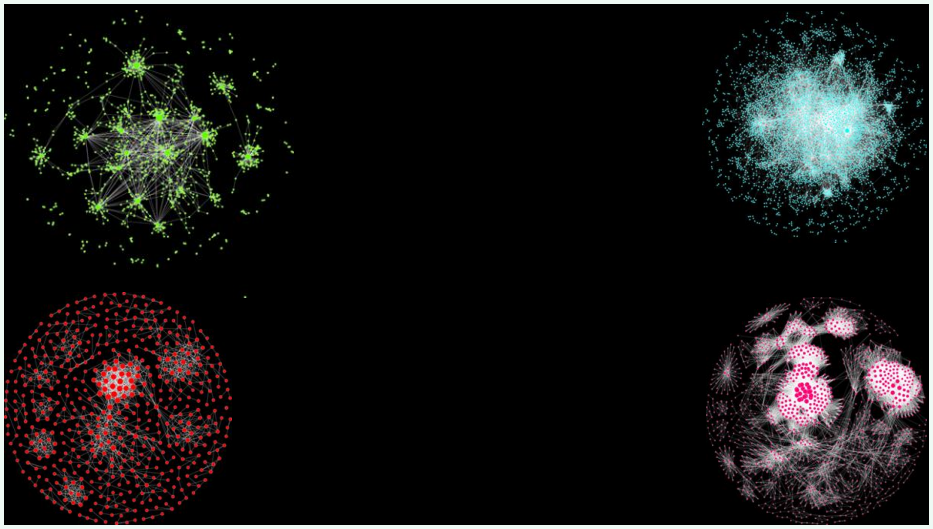
   ii. **Multiple layers of heterogeneous data:**
   - **iCell**
   - **Patient-centered data integration → Precision medicine**
     - ✓ Stratification, biomarker discovery, drug repurposing
   - **Disease re-classification, GO reconstruction, Network alignment, …**

3. **Conclusions**

# 2. Novel Methods

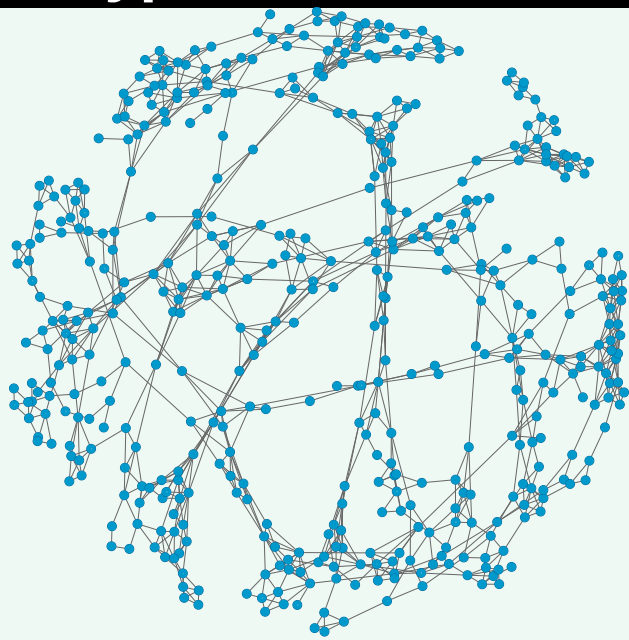
## Mine Inter-Connected Entities: One Network Type

i. Molecular Networks

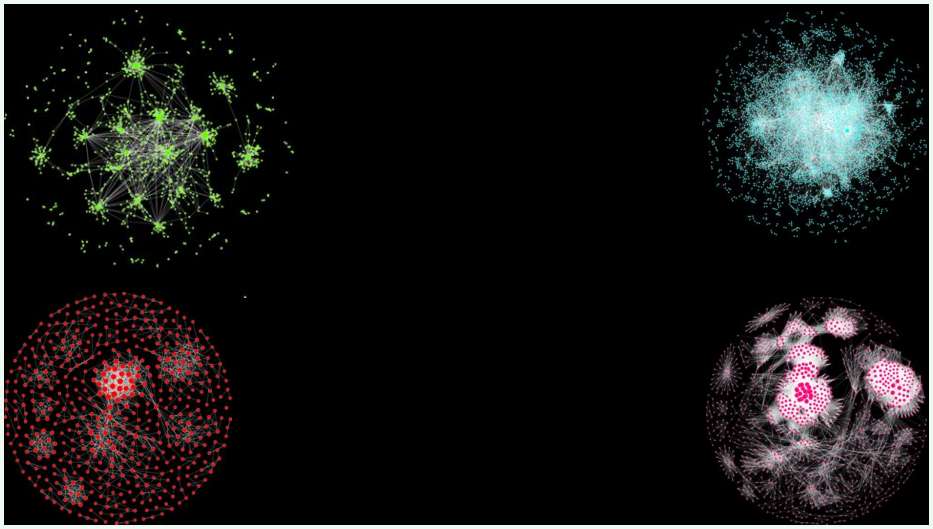


# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type



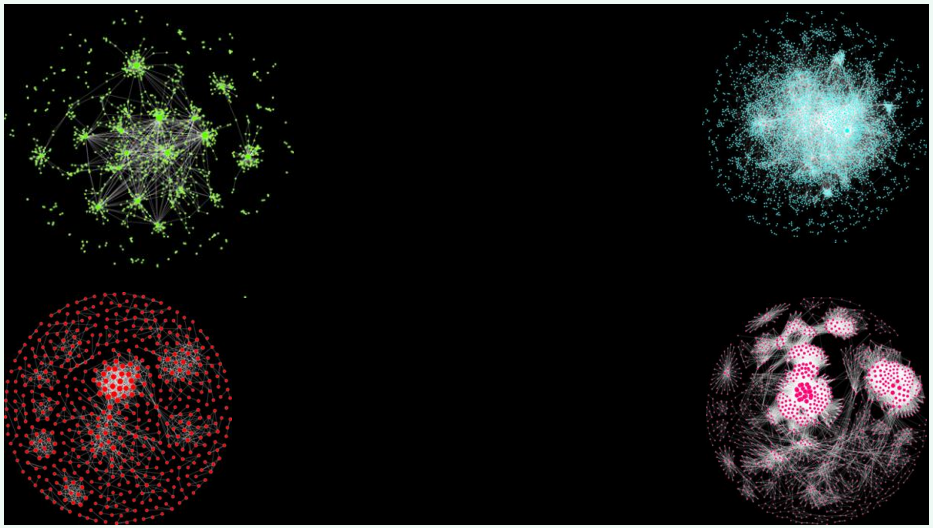
i. Molecular Networks

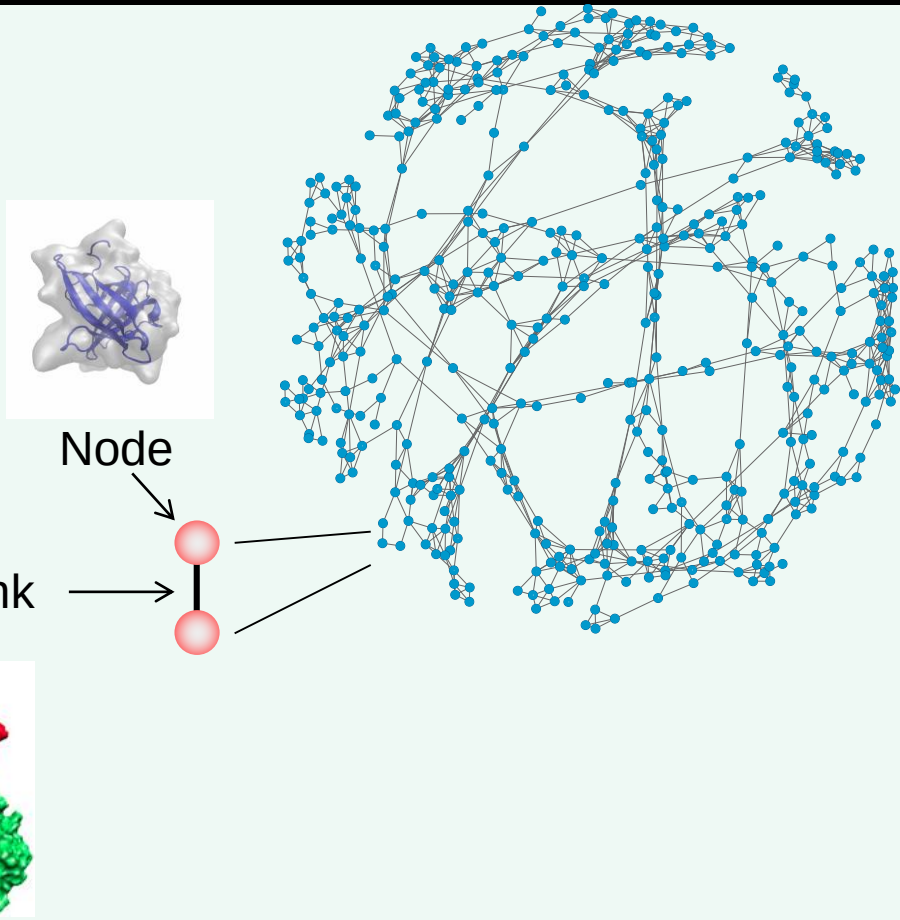


# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

### i. Molecular Networks

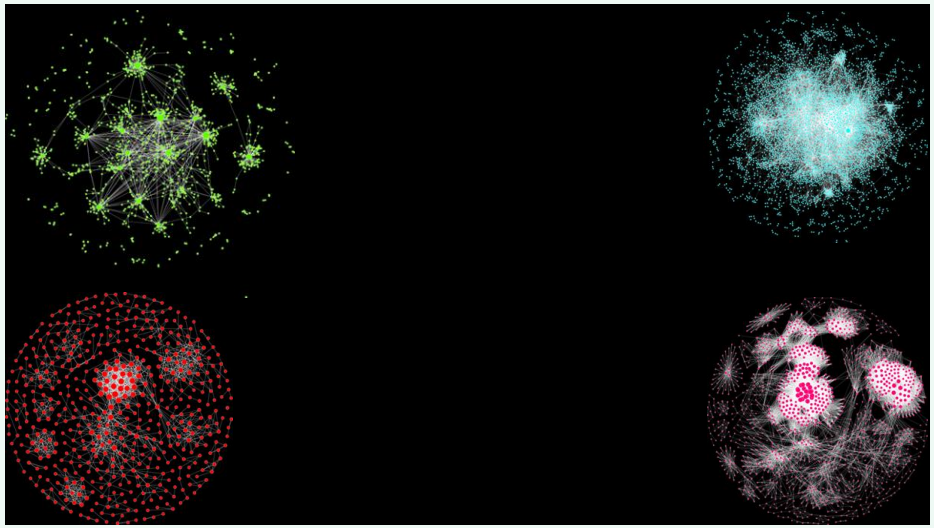


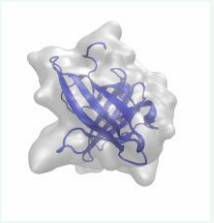


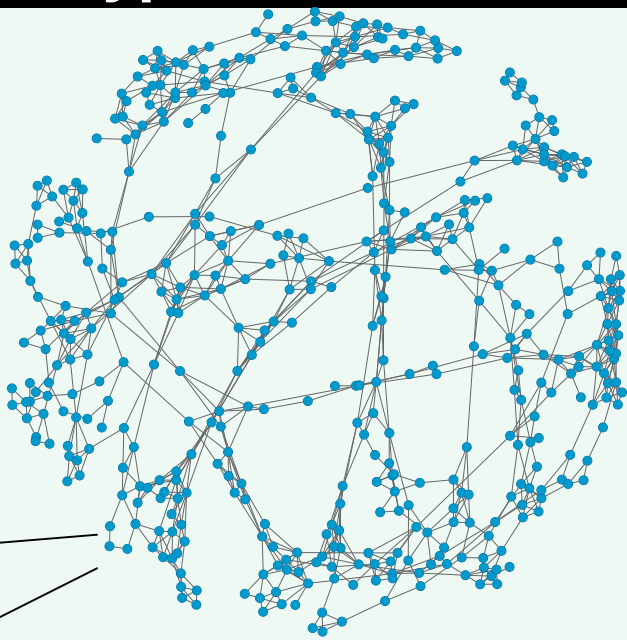
# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

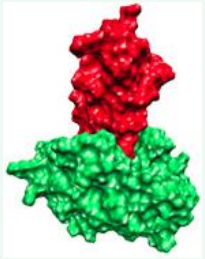
### i. Molecular Networks

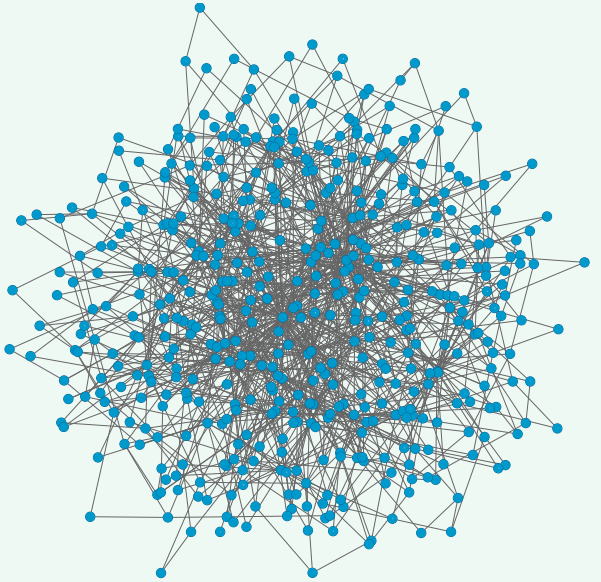








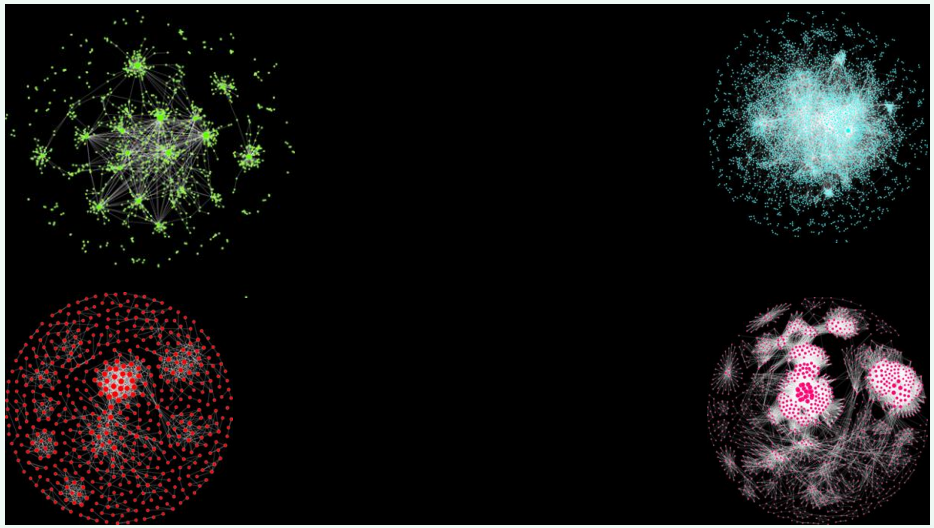


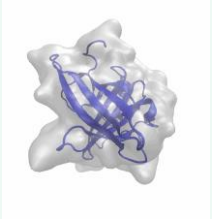


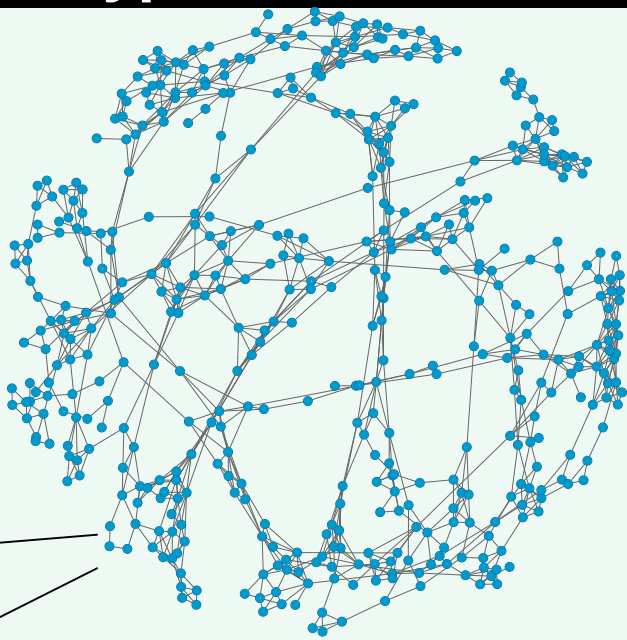
# 2. Novel Methods

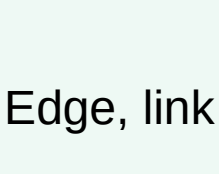
## Mine Inter-Connected Entities: One Network Type

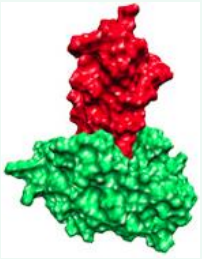
### i. Molecular Networks

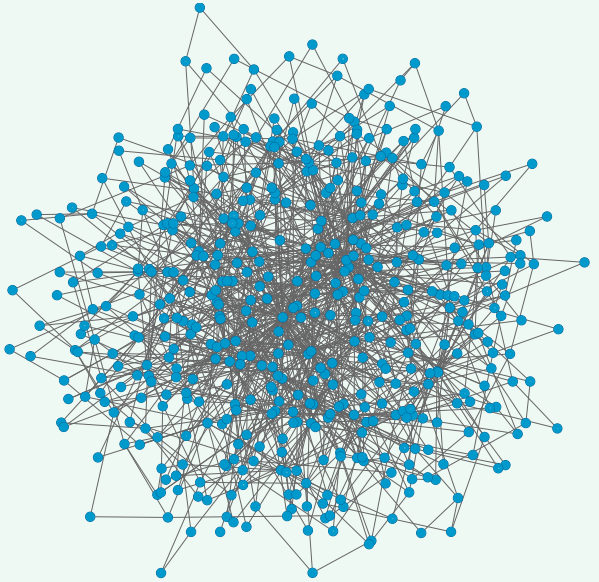








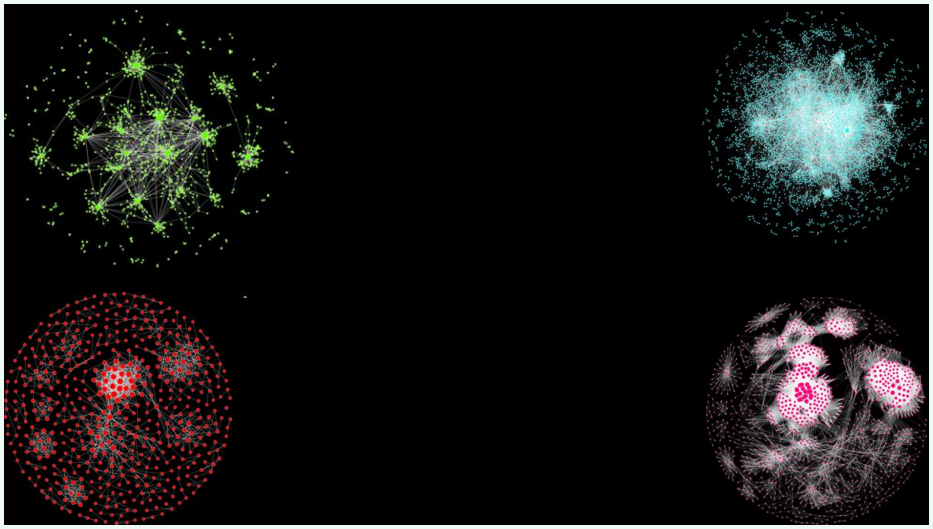


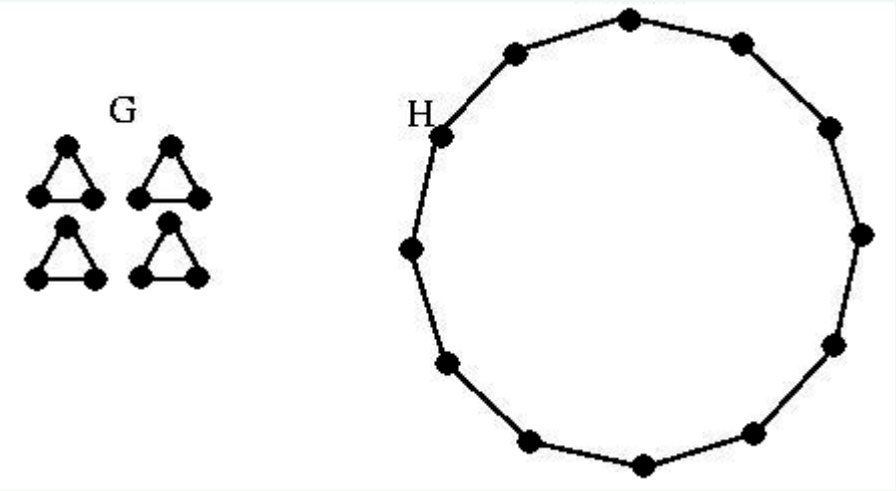
- The number of nodes
- The number of links
- Links of each node: *degree*
- Distribution of links (degrees)

# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

i. Molecular Networks





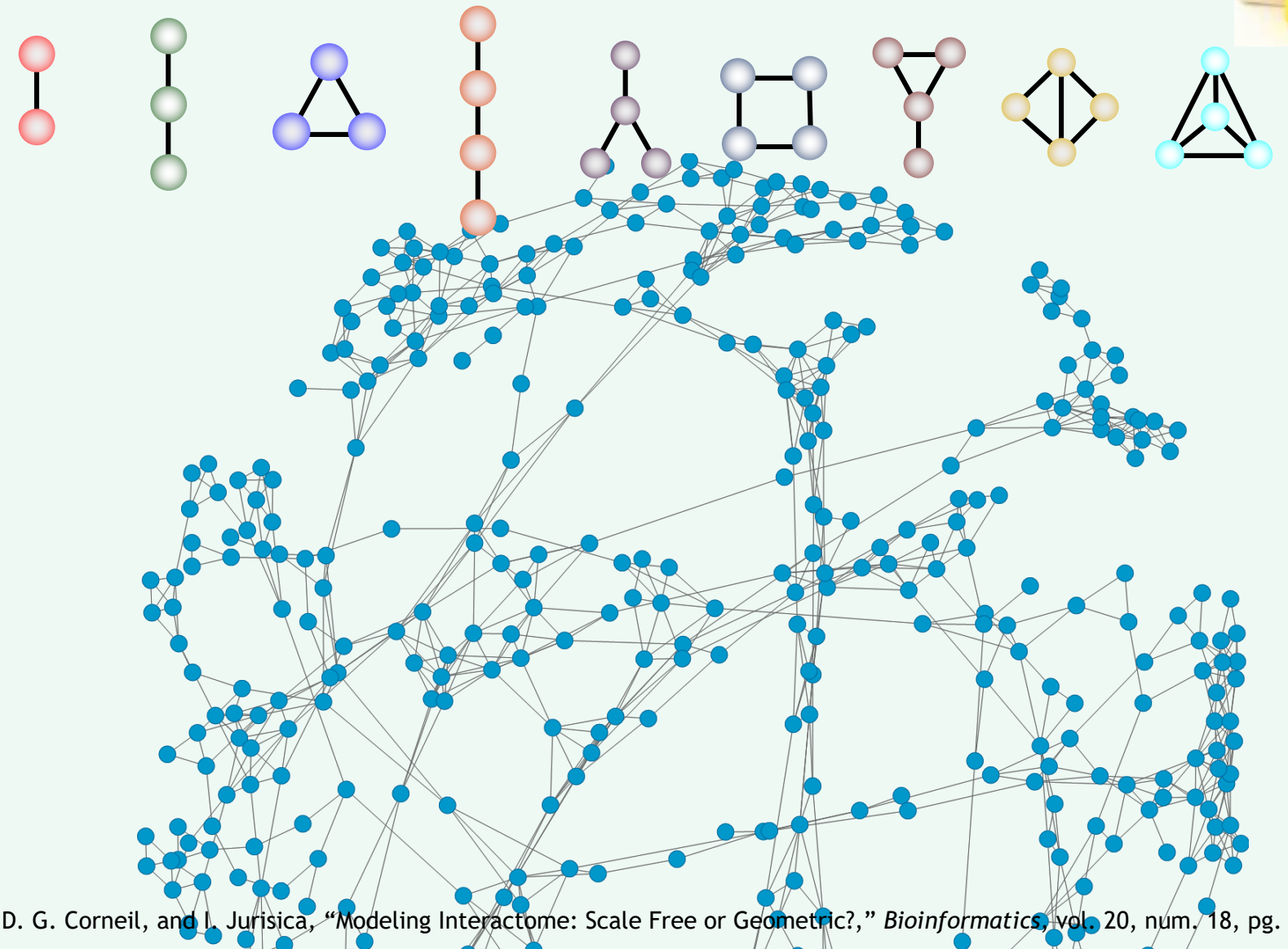

- The number of nodes
- The number of links
- Links of each node: *degree*
- Distribution of links (degrees)

# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

***Graphlets***

***"Legos of Networks"***

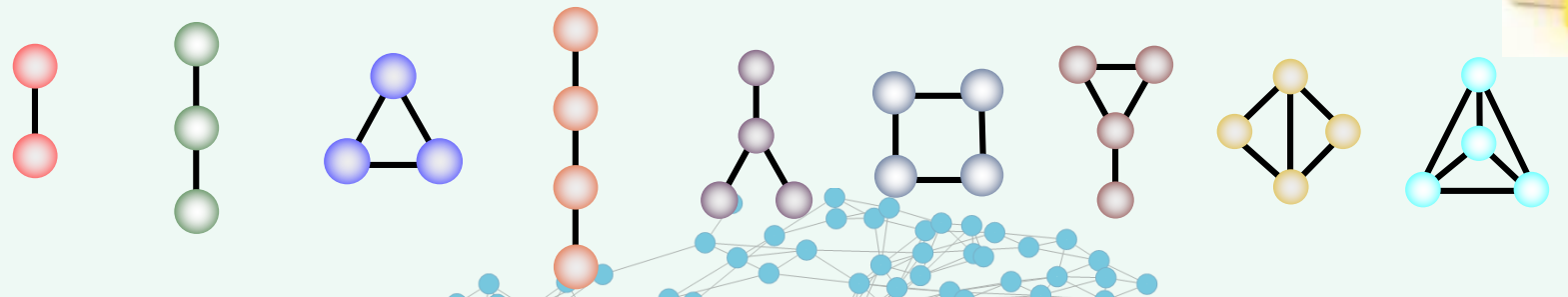


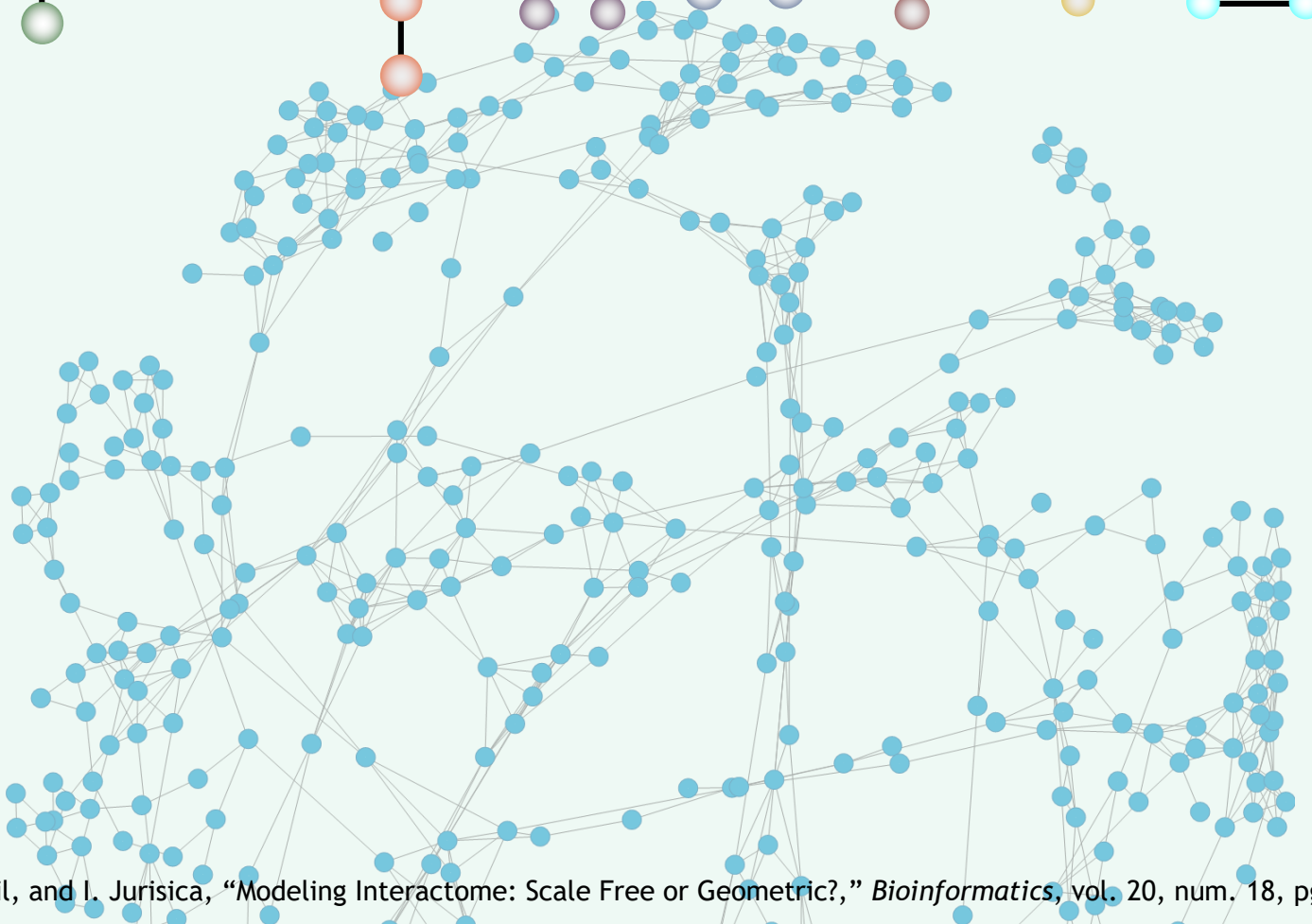
**N. Przulj**, D. G. Corneil, and I. Jurisica, "Modeling Interactome: Scale Free or Geometric?," *Bioinformatics*, vol. 20, num. 18, pg. 3508-3515, 2004.

# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

***Graphlets***

***"Legos of Networks"***





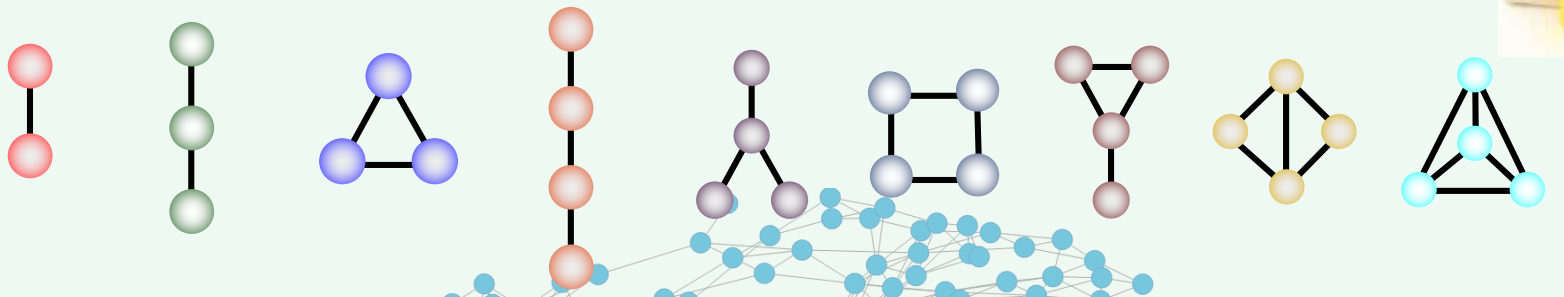
**N. Przulj**, D. G. Corneil, and I. Jurisica, "Modeling Interactome: Scale Free or Geometric?," *Bioinformatics*, vol. 20, num. 18, pg. 3508-3515, 2004.

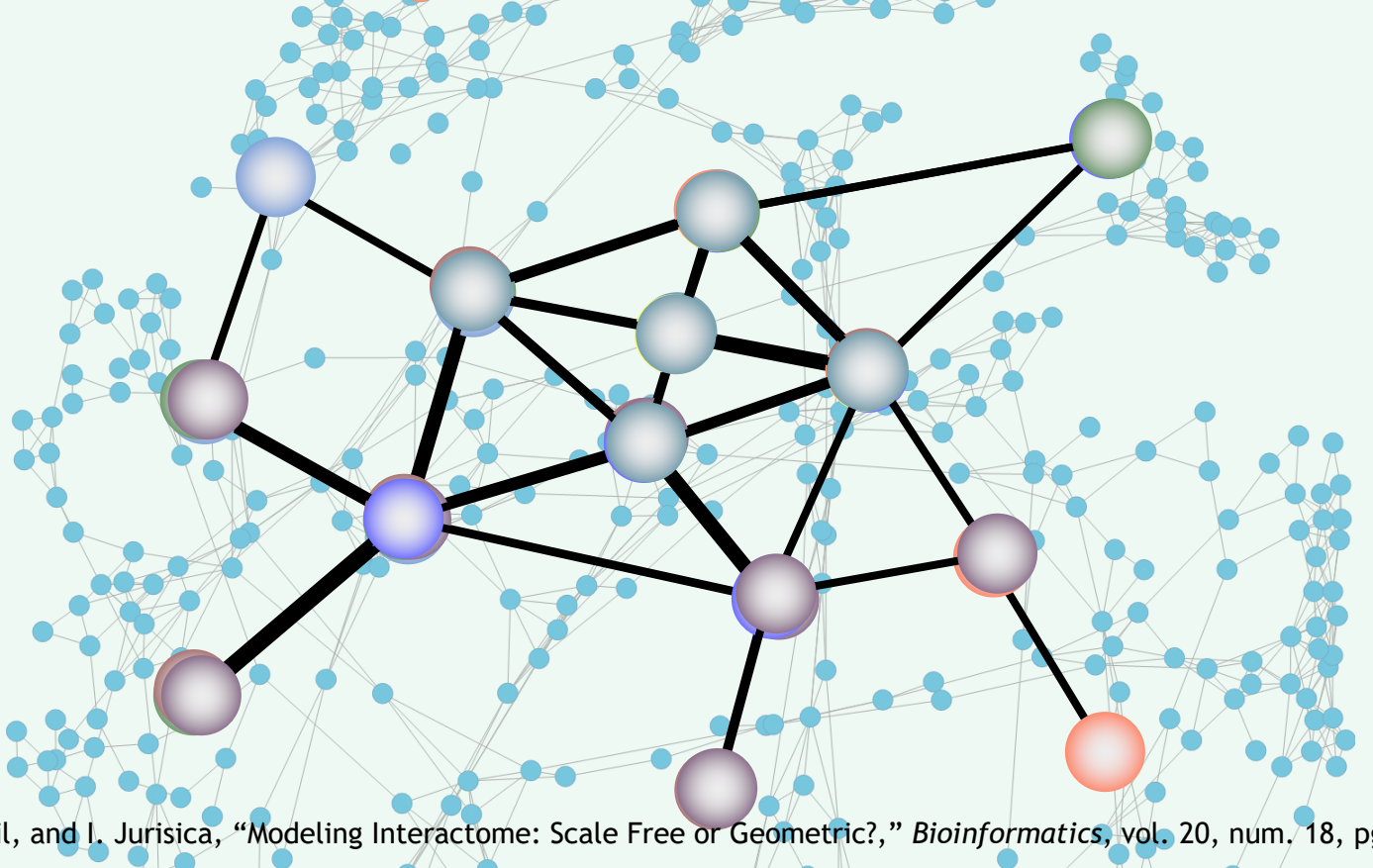
# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

***Graphlets***

***"Legos of Networks"***



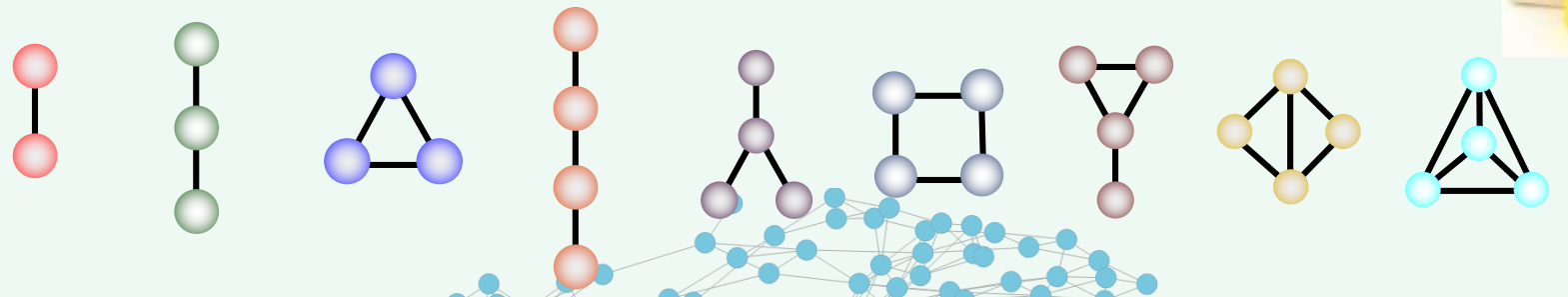


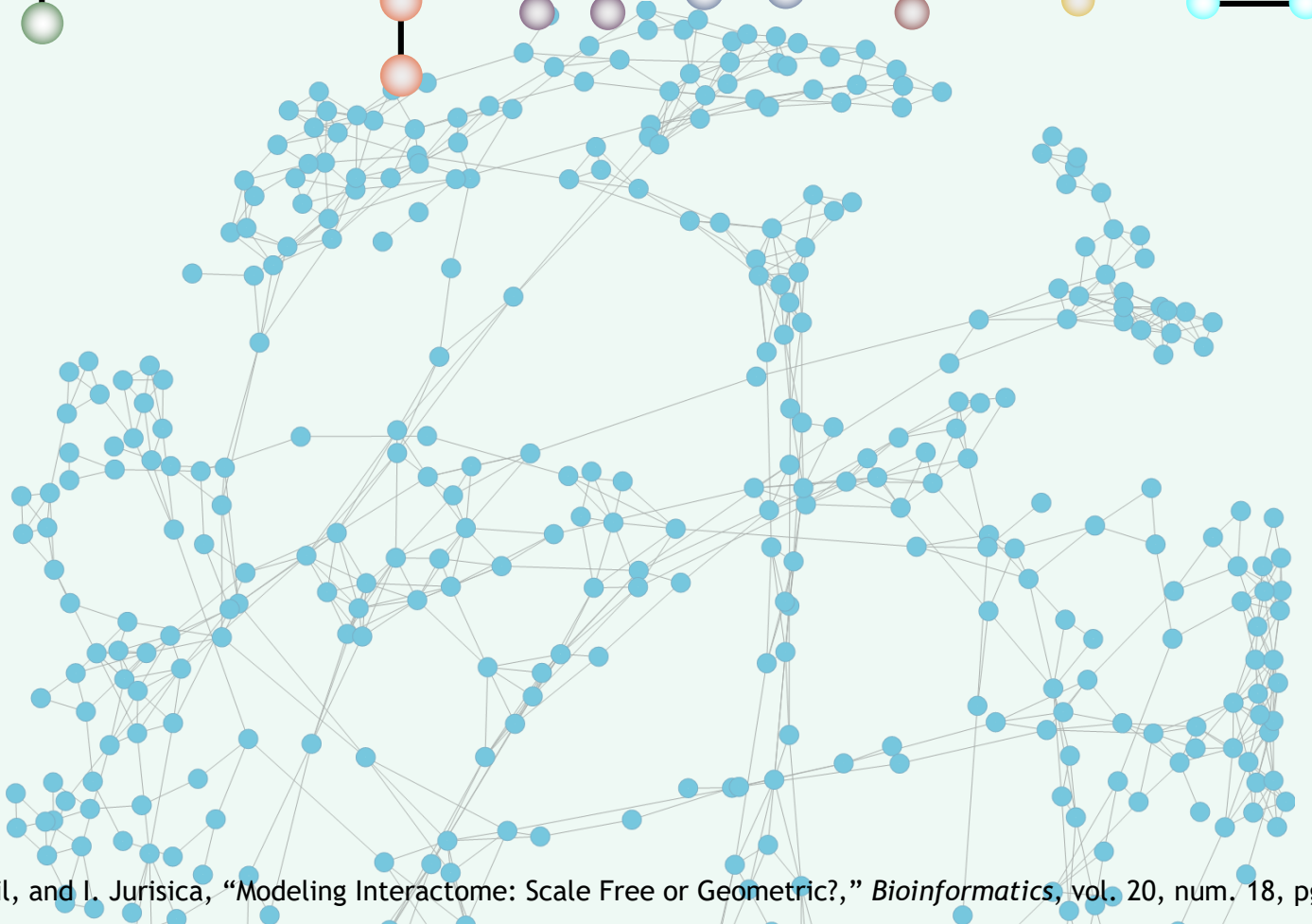
**N. Przulj**, D. G. Corneil, and I. Jurisica, "Modeling Interactome: Scale Free or Geometric?," *Bioinformatics*, vol. 20, num. 18, pg. 3508-3515, 2004.

# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

***Graphlets***

***"Legos of Networks"***





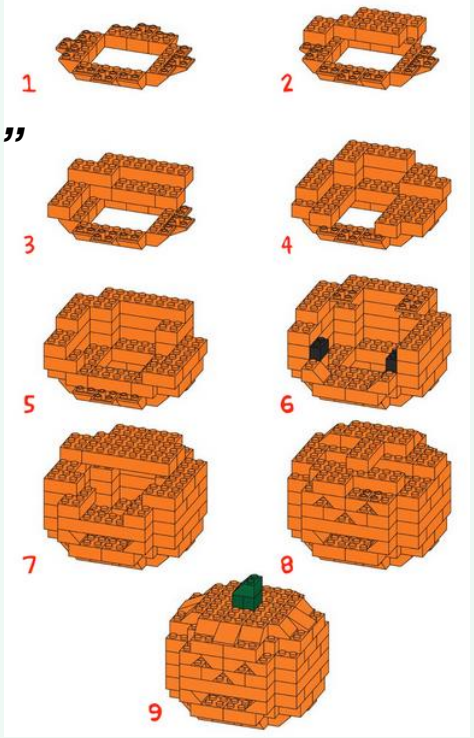
**N. Przulj**, D. G. Corneil, and I. Jurisica, "Modeling Interactome: Scale Free or Geometric?," *Bioinformatics*, vol. 20, num. 18, pg. 3508-3515, 2004.

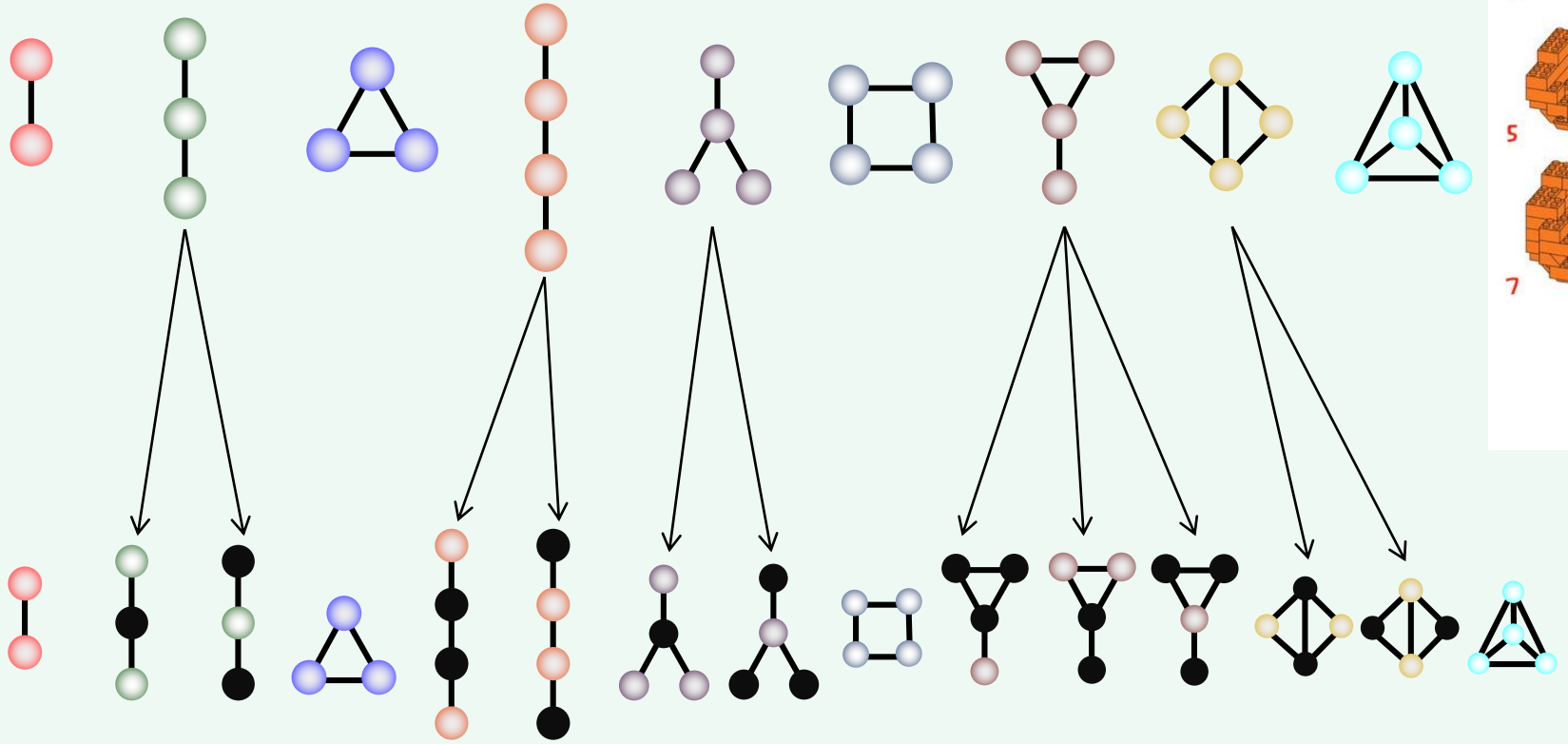
# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

**ERC StG**: 278212 (2012-2017): "Biological Network Topology Complements Genome as a Source of Biological Information"

***Graphlets***

***"Legos of Networks"***



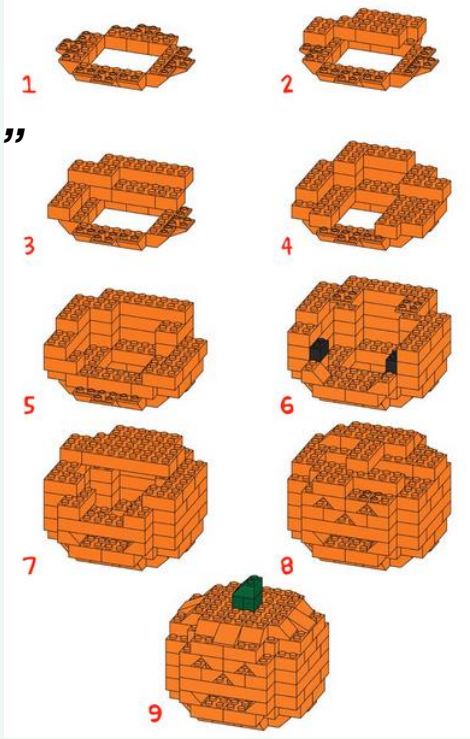
**N. Przulj**, "Biological Network Comparison Using Graphlet Degree Distribution," *Proceedings of the 2006 European Conference on Computational Biology, ECCB '06*, Eilat, Israel, January 21-24, 2007, **acceptance rate 18%.** *Bioinformatics*, volume 23, pages e177-e183, 2007

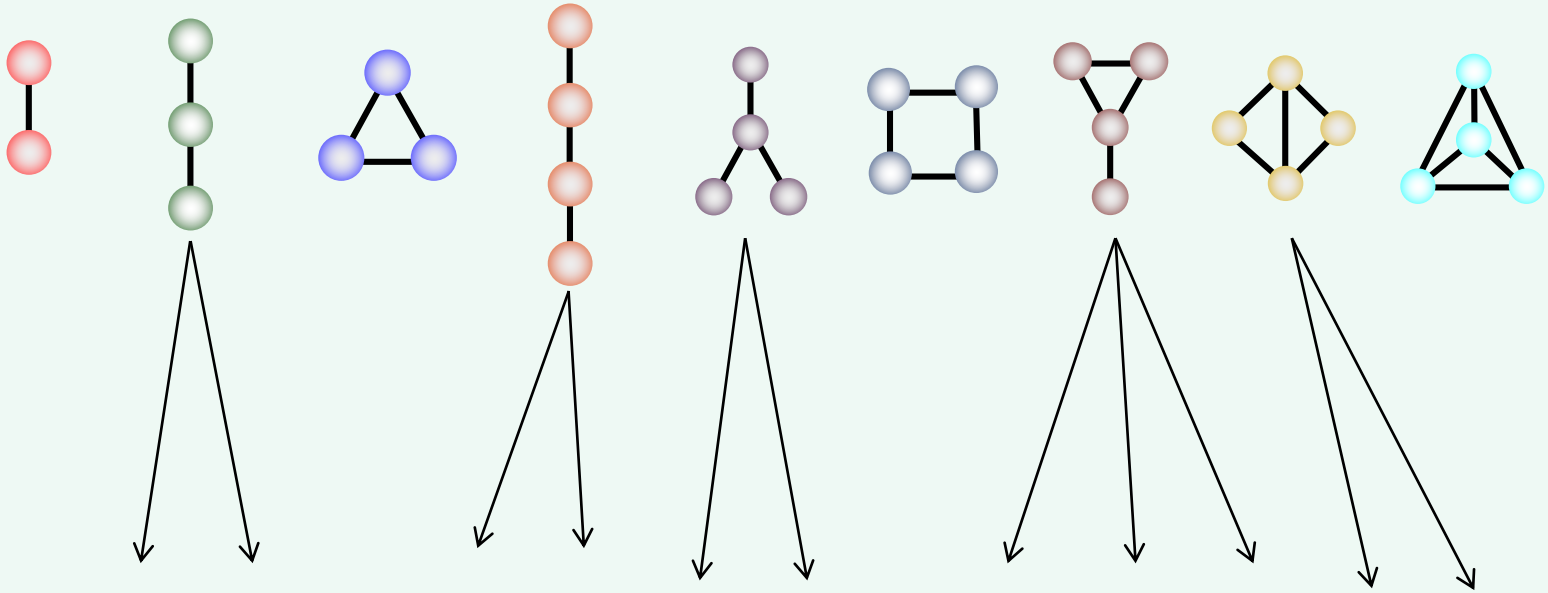
# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

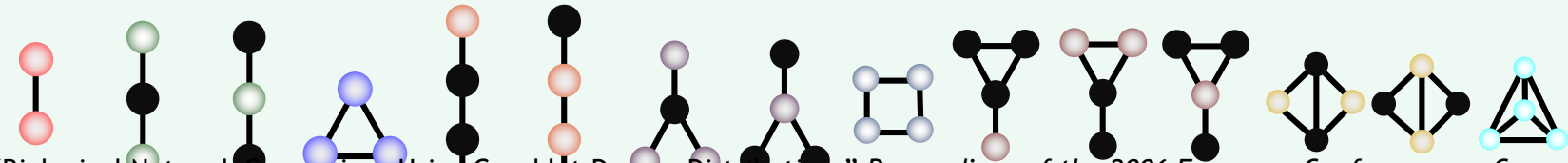
**ERC StG**: 278212 (2012-2017): "Biological Network Topology Complements Genome as a Source of Biological Information"

***Graphlets***

***"Legos of Networks"***



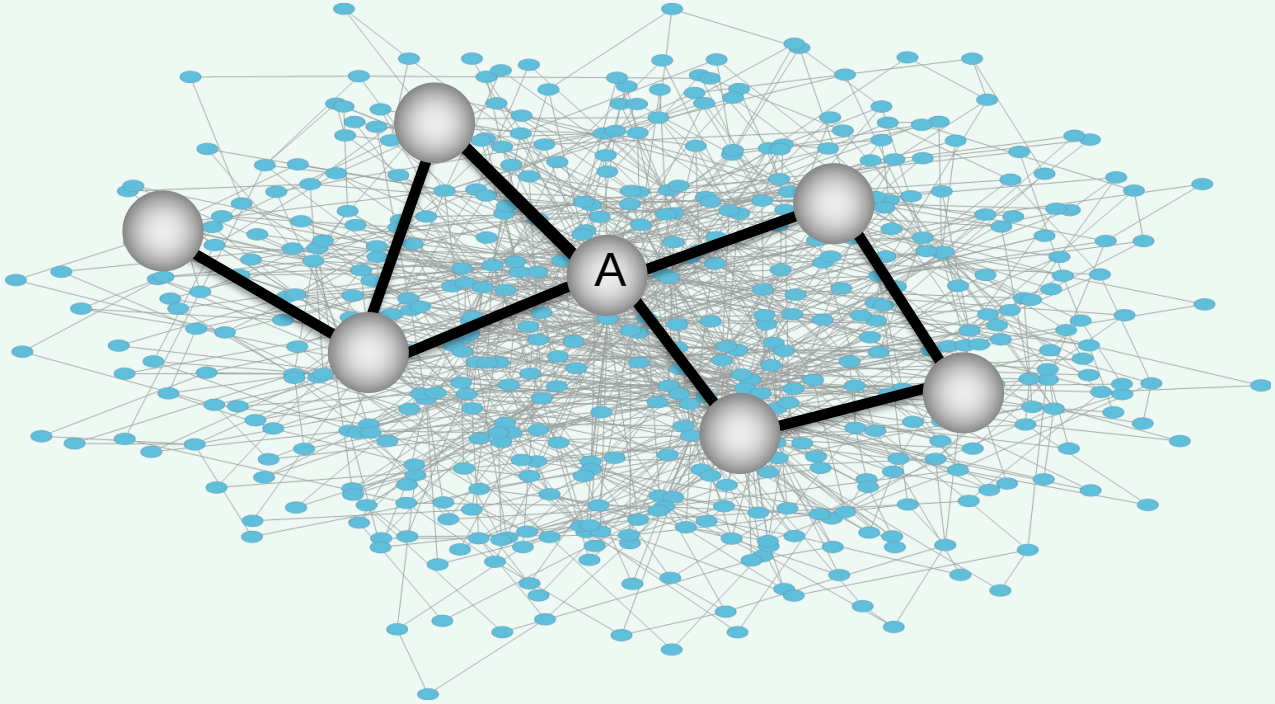





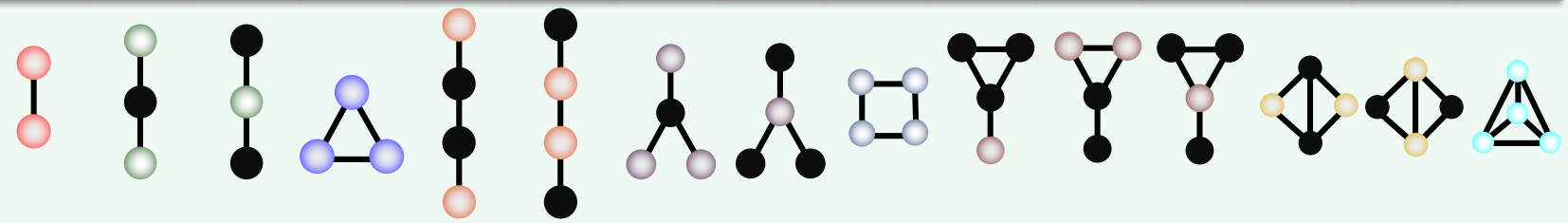
**N. Przulj**, "Biological Network Comparison Using Graphlet Degree Distribution," *Proceedings of the 2006 European Conference on Computational Biology, ECCB '06*, Eilat, Israel, January 21-24, 2007, **acceptance rate 18%.** *Bioinformatics*, volume 23, pages e177-e183, 2007

# 2. Novel Methods

## Mine the Medical World of Inter-Connected Entities

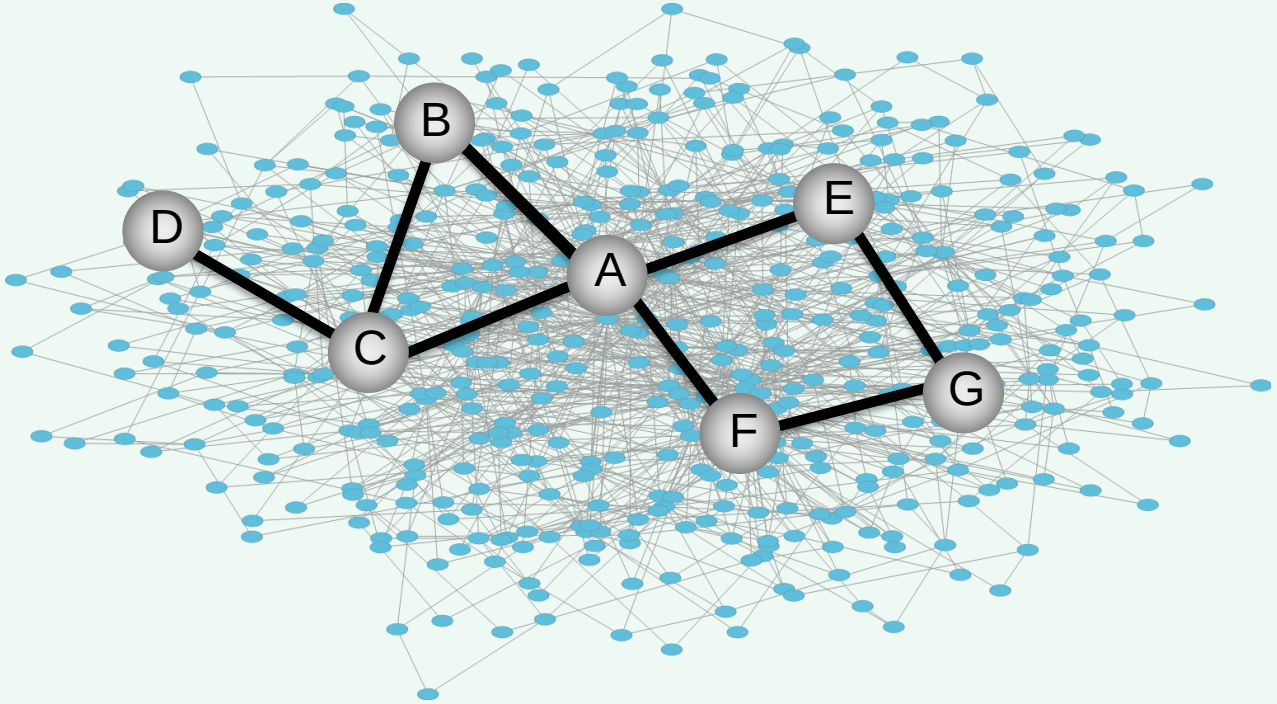


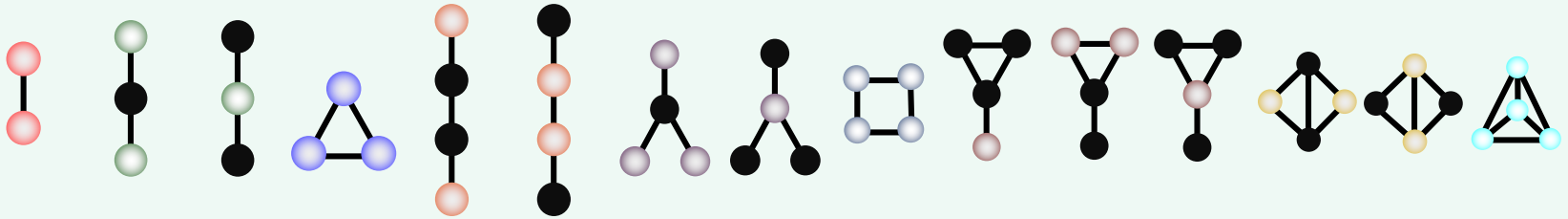

| Orbit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | 3 | 5 | 1 | 0 | 6 | 0 | 2 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |



# 2. Novel Methods

## Mine the Medical World of Inter-Connected Entities

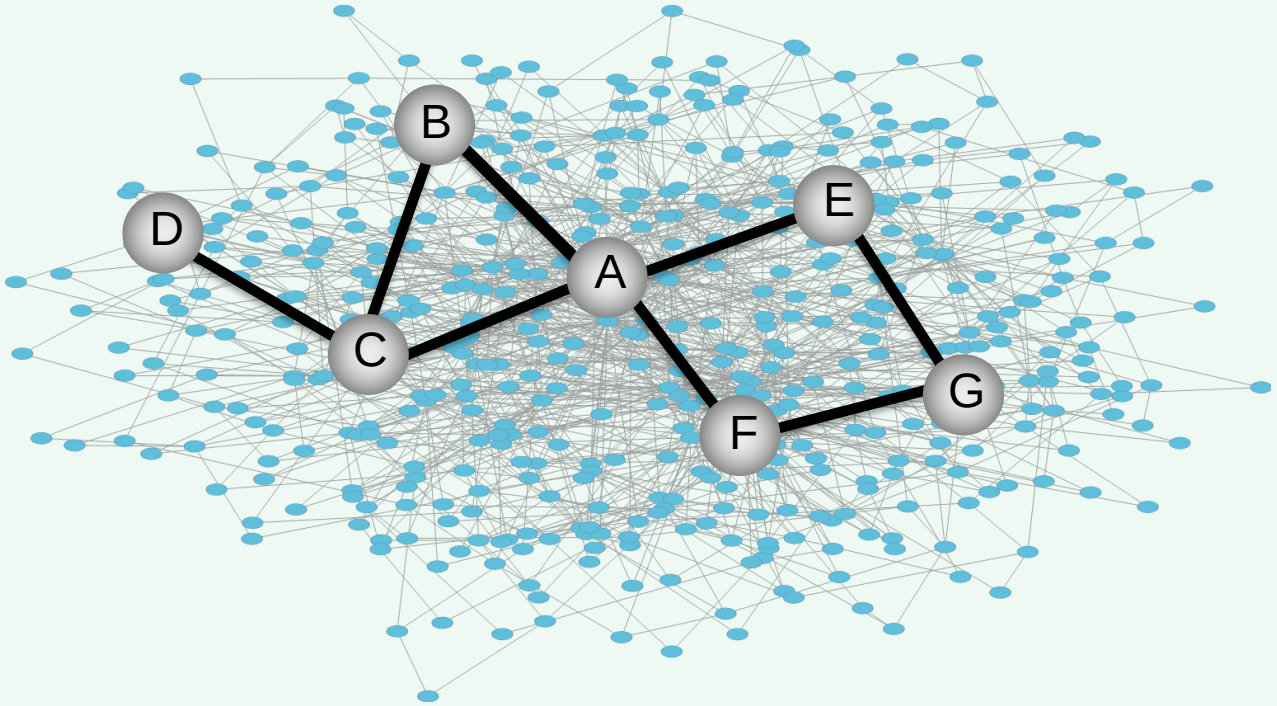


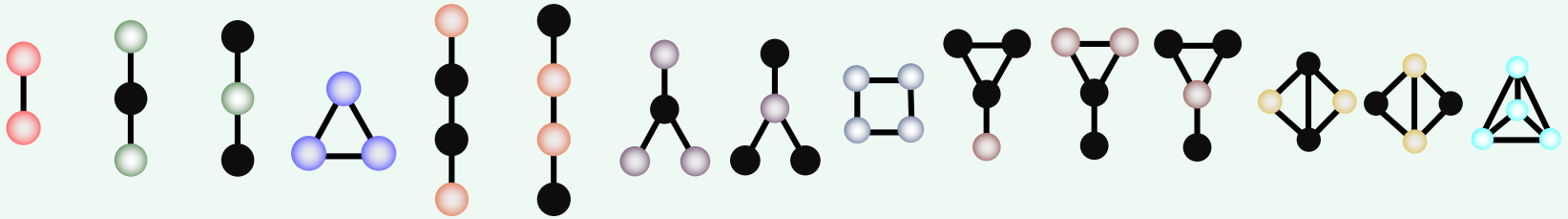



| Orbit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 4 | 3 | 5 | 1 | 0 | 6 | 0 | 2 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |

# 2. Novel Methods

## Mine the Medical World of Inter-Connected Entities

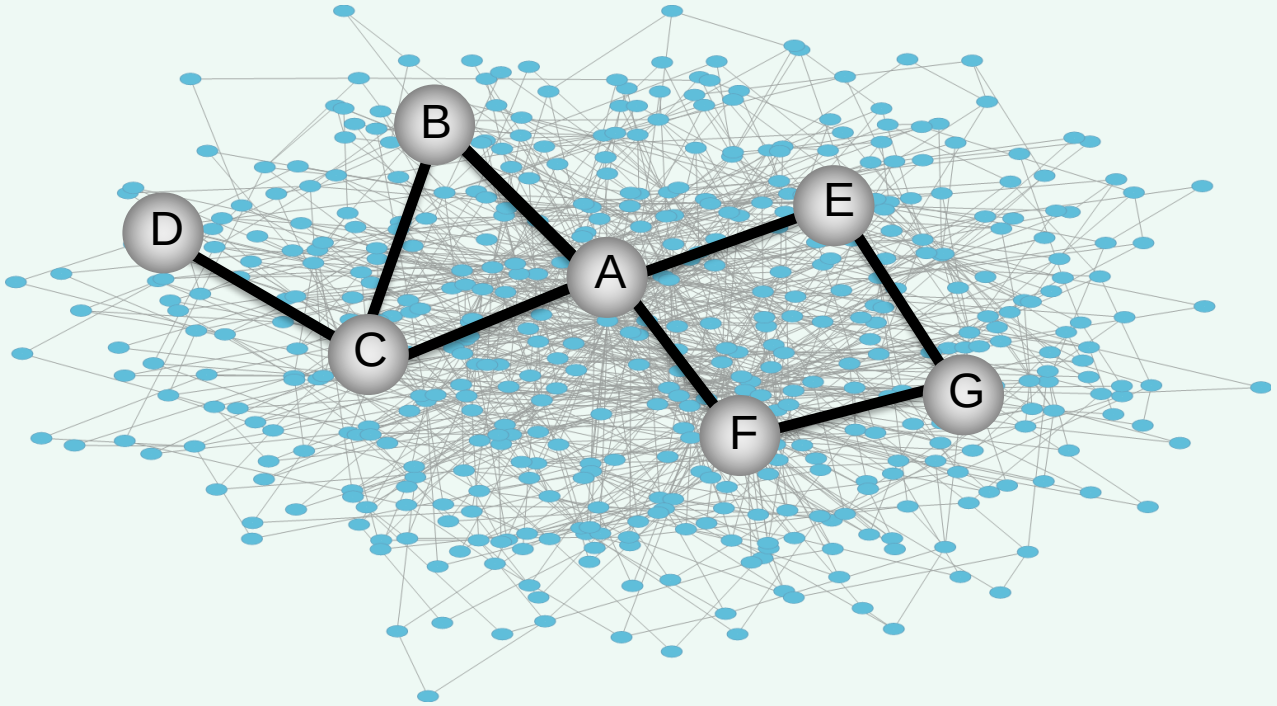


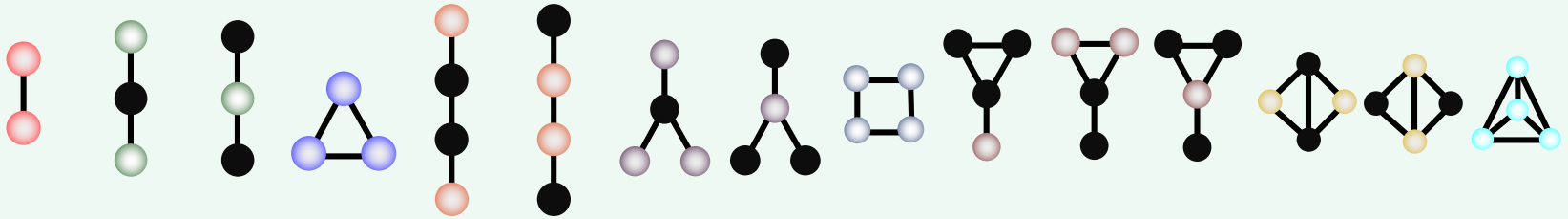



| Orbit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | 3 | 5 | 1 | 0 | 6 | 0 | 2 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| B | 2 | 3 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| C | 3 | 2 | 2 | 1 | 2 | 2 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| D | 1 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| E | 2 | 4 | 1 | 0 | 1 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| F | 2 | 4 | 1 | 0 | 1 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| G | 2 | 2 | 1 | 0 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

# 2. Novel Methods

## Mine the Medical World of Inter-Connected Entities

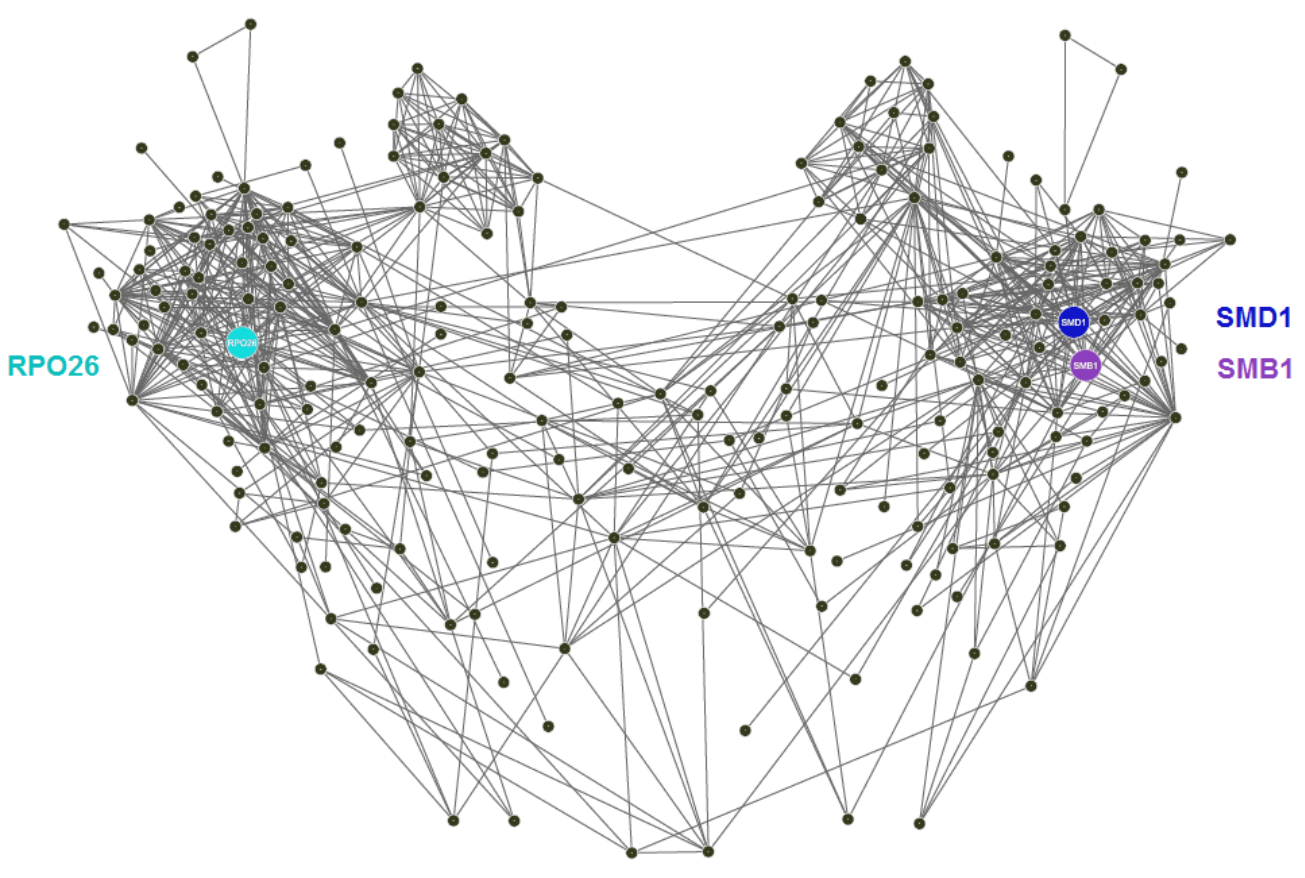


| Orbit | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 4 | 3 | 5 | 1 | 0 | 6 | 0 | 2 | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| B | 2 | 3 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| C | 3 | 2 | 2 | 1 | 2 | 2 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| D | 1 | 2 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| E | 2 | 4 | 1 | 0 | 1 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| F | 2 | 4 | 1 | 0 | 1 | 2 | 2 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| G | 2 | 2 | 1 | 0 | 4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

<u>90% similar wiring – significantly enriched</u>:

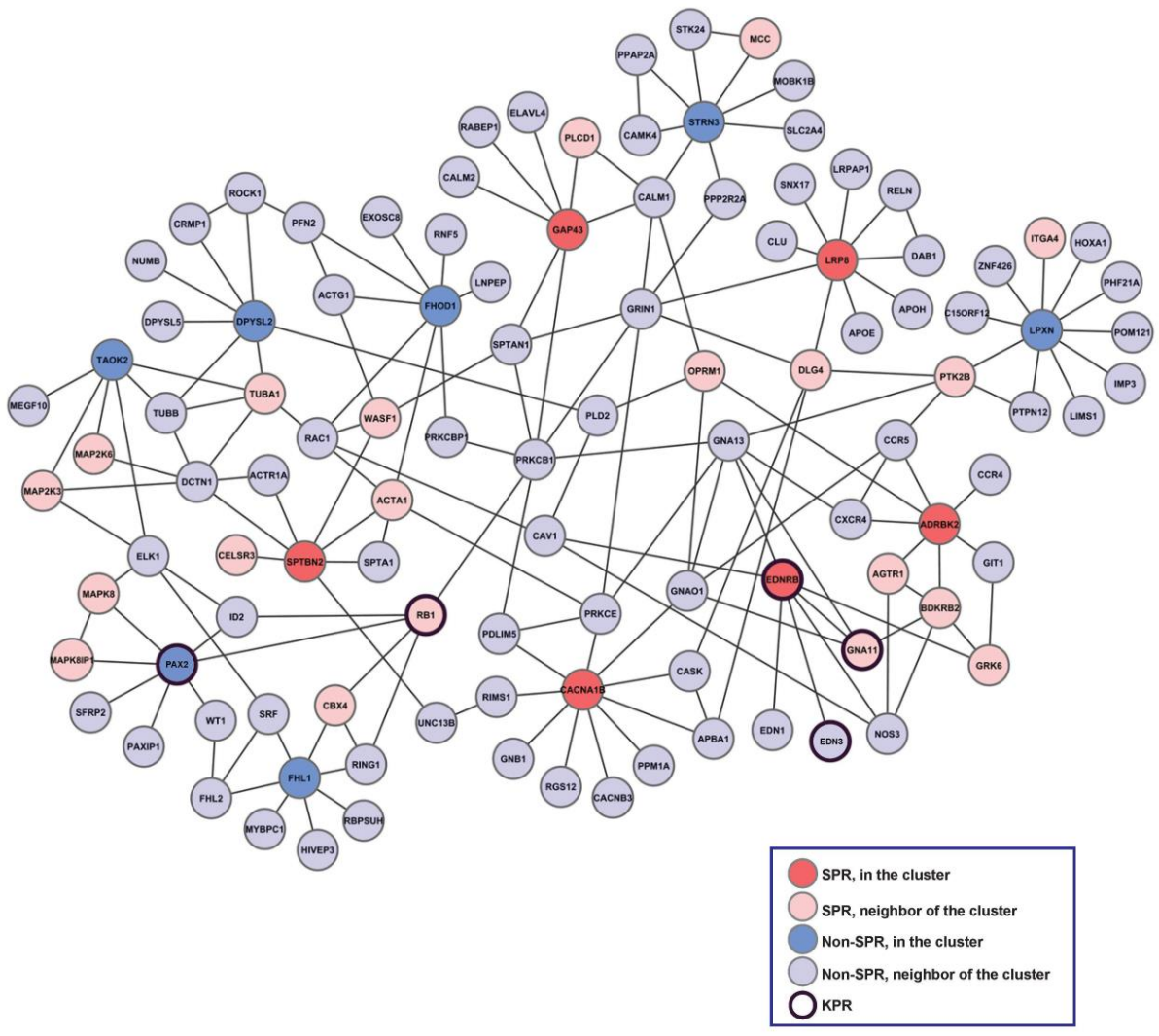



→ Biological function
→ Protein complexes
→ Sub-cellular localization
→ Tissue expression
→ Disease

T. Milenkovic and **N. Przulj**, "Uncovering Biological Network Function via Graphlet Degree Signatures", *Cancer Informatics*, vol. 4, pg. 257-273, 2008 (Highly accessed)

# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type




**Cancer research:**

→ New proteins for melanin production

→ Same cancer type: more similar wiring

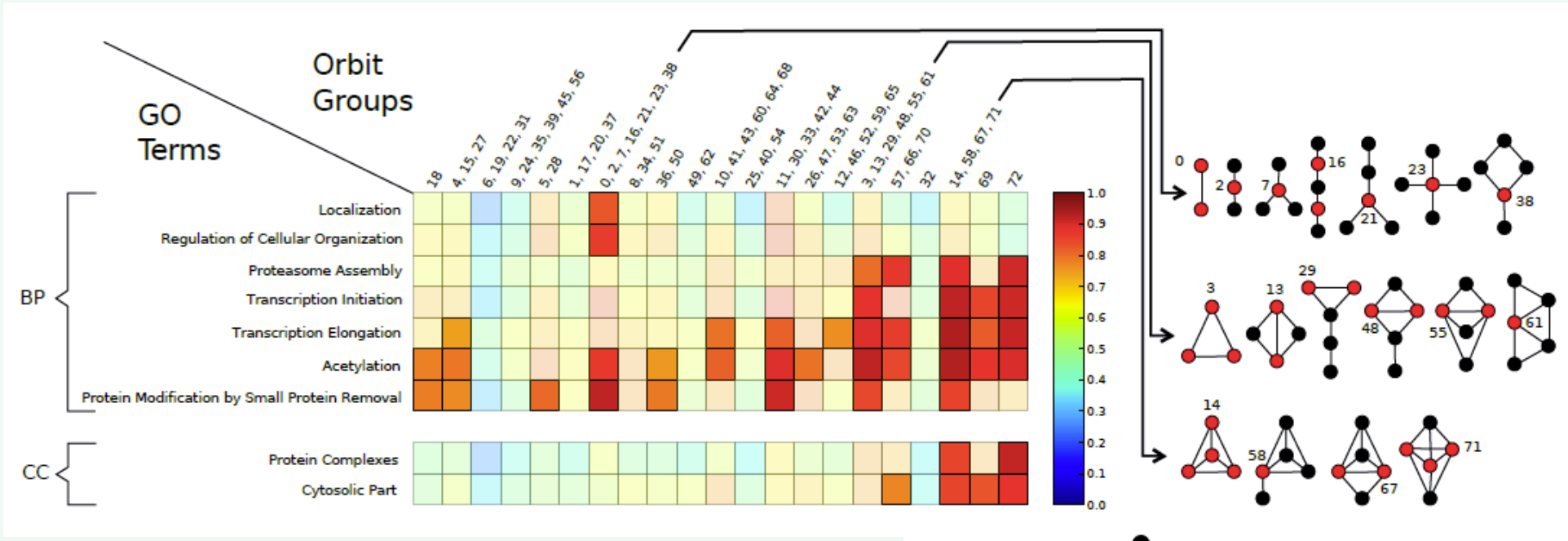
→ Far away in the network

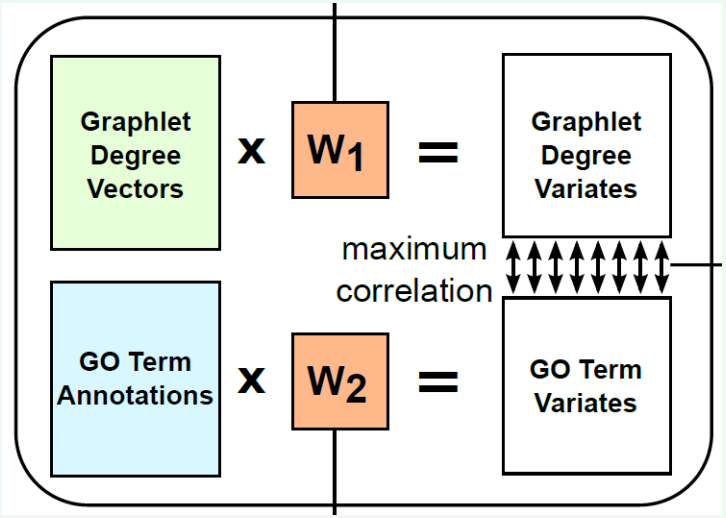

T. Milenković, V. Memisević, A. K. Ganesan, and **N. Pržulj**, *J. Roy. Soc. Interface*, 7(44):423-437, 2010

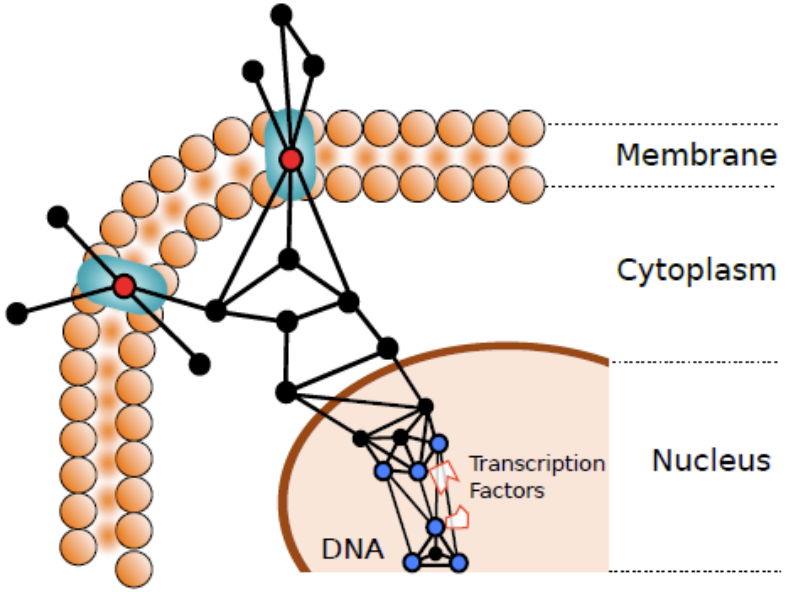
H. Ho, T. Milenković, V. Memisević, J. Aruri, **N. Pržulj**, and A. K. Ganesan, *BMC Systems Biology*, 4:84, 2010 (Highly accessed)

# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type



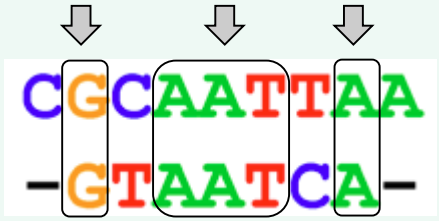



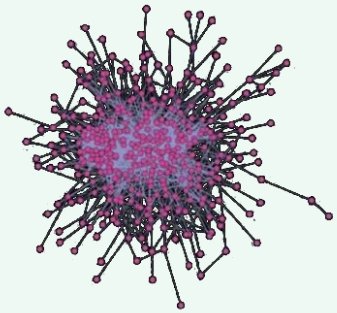

D. Davis, O. N. Yaveroglou, N. Malod-Dognin, A. Stojmirovic, **N. Przulj**, "Topology-Function Conservation in Protein-Protein Interaction Networks," *Bioinformatics* 31(10):1632-1639, 2015. IF=7.3

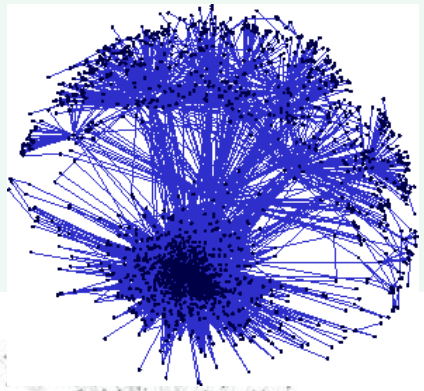
# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

### Network Alignment



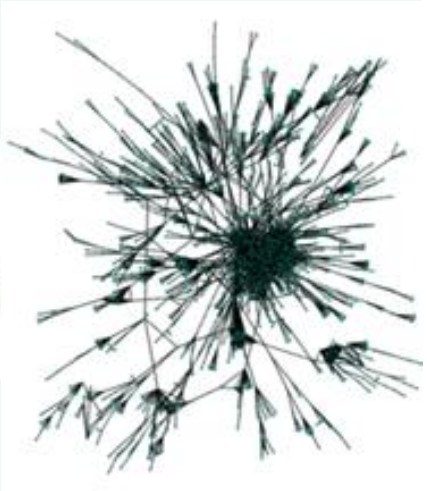





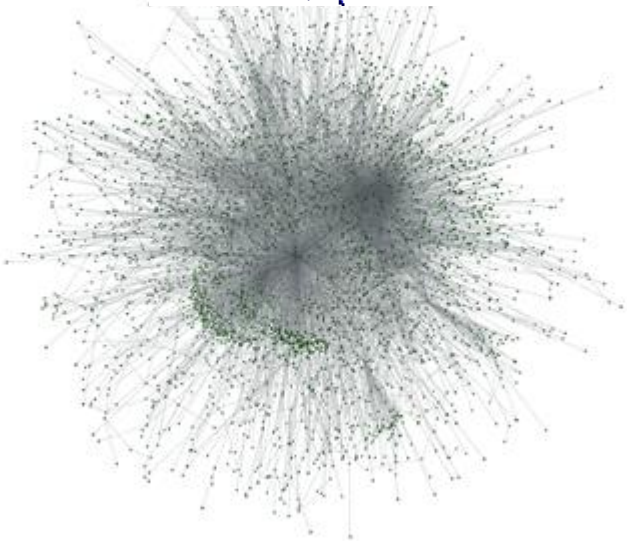
**Isorank:**
116 nodes
261 edges



**GRAAL:**
267 nodes
900 edges



**MI-GRAAL:**
1,858 nodes
3,467 edges



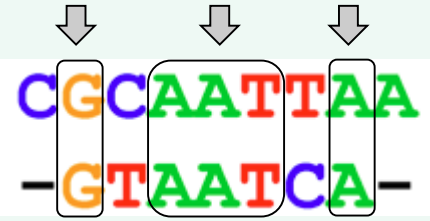
**L-GRAAL:**
5,726 nodes
16,084 edges
Yeast: 98% proteins
21% interactions

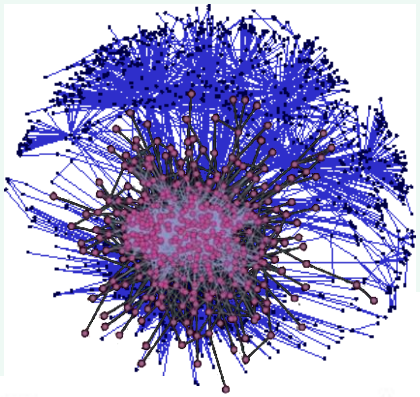

N. Malod-Dognin & N. Pržulj, L-GRAAL, *Bioinformatics*, doi: 10.1093/bioinformatics/btv130, 2015
N. Malod-Dognin & N. Pržulj, GR-ALIGN, *Bioinformatics*, doi:10.1093/bioinformatics/btu020, 2014
V. Memisevic & N. Pržulj, C-GRAAL, *Integrative Biology*, doi:10.1039/c2ib00140c, 2012
O. Kuchaiev & N. Pržulj, MI-GRAAL, *Bioinformatics*, 27(10): 1390-6, 2011
O. Kuchaiev, T. Milenkovic, V. Memisevic, W. Hayes, & N. Pržulj, *J. Royal Society Interface*, 7:1341-1354, 2010
T. Milenkovic, W.L. Wong, W. Hayes, & N. Pržulj, *Cancer Informatics*, 9:121-37, June 30, 2010 (Highly visible)

# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

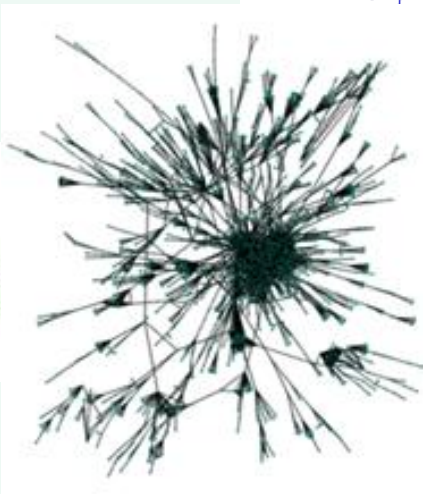
## Network Alignment



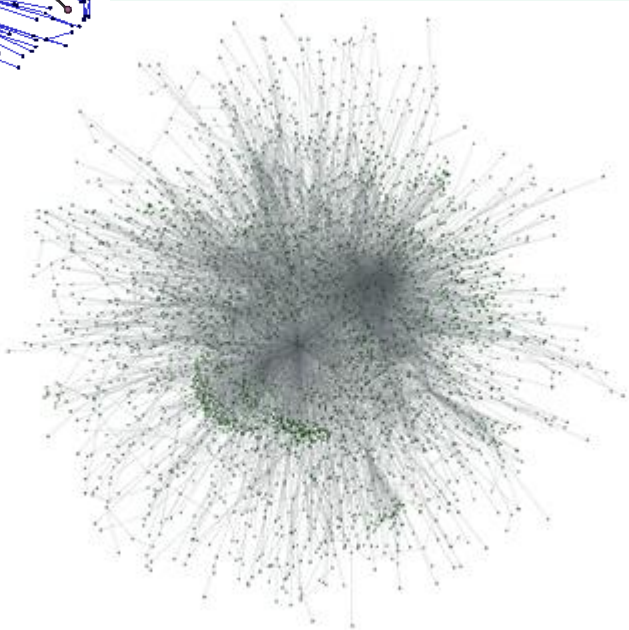



**Isorank:**
116 nodes
261 edges

**GRAAL:**
267 nodes
900 edges



**MI-GRAAL:**
1,858 nodes
3,467 edges



**L-GRAAL:**
5,726 nodes
16,084 edges
Yeast: 98% proteins
21% interactions

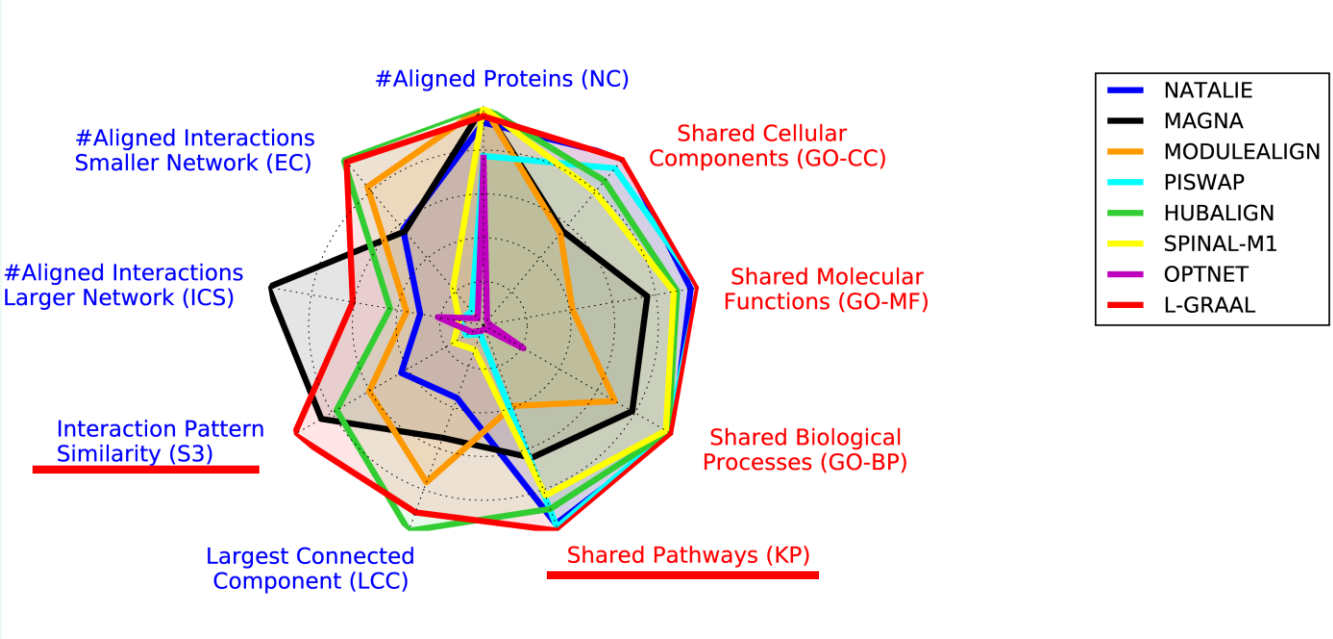

N. Malod-Dognin & N. Pržulj, L-GRAAL, *Bioinformatics*, doi: 10.1093/bioinformatics/btv130, 2015
N. Malod-Dognin & N. Pržulj, GR-ALIGN, *Bioinformatics*, doi:10.1093/bioinformatics/btu020, 2014
V. Memisevic & N. Pržulj, C-GRAAL, *Integrative Biology*, doi:10.1039/c2ib00140c, 2012
O. Kuchaiev & N. Pržulj, MI-GRAAL, *Bioinformatics*, 27(10): 1390-6, 2011
O. Kuchaiev, T. Milenkovic, V. Memisevic, W. Hayes, & N. Pržulj, *J. Royal Society Interface*, 7:1341-1354, 2010
T. Milenkovic, W.L. Wong, W. Hayes, & N. Pržulj, *Cancer Informatics*, 9:121-37, June 30, 2010 (Highly visible)

# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

### Alignment of PPI Networks – *Ulign*

- Many methods
- All heuristic
- No gold standard

- Questions:
  - Which aligner for which data?
  - Which scoring scheme for evaluation?
  - Coverage: biological and topological?
  - Contribution of topology vs sequence?




- Map biologically and topologically ***different*** network regions
- Each covers only about 50% of the proteins of the larger network

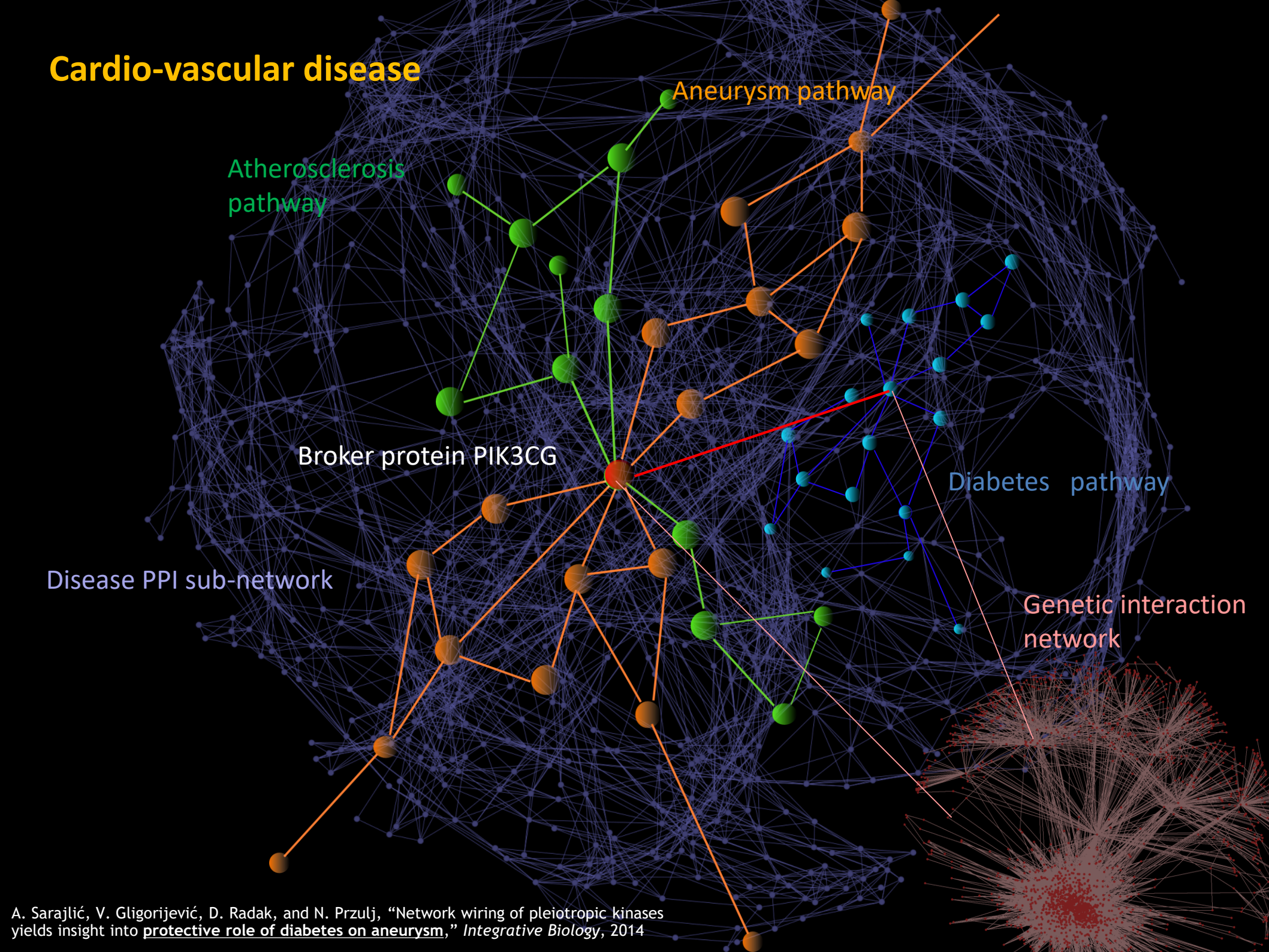
- **Together** – map **entire** networks → ***Ulign***
  - Biologically coherent

- The most topologically coherent – using topology only
- The most biologically coherent – using sequence only

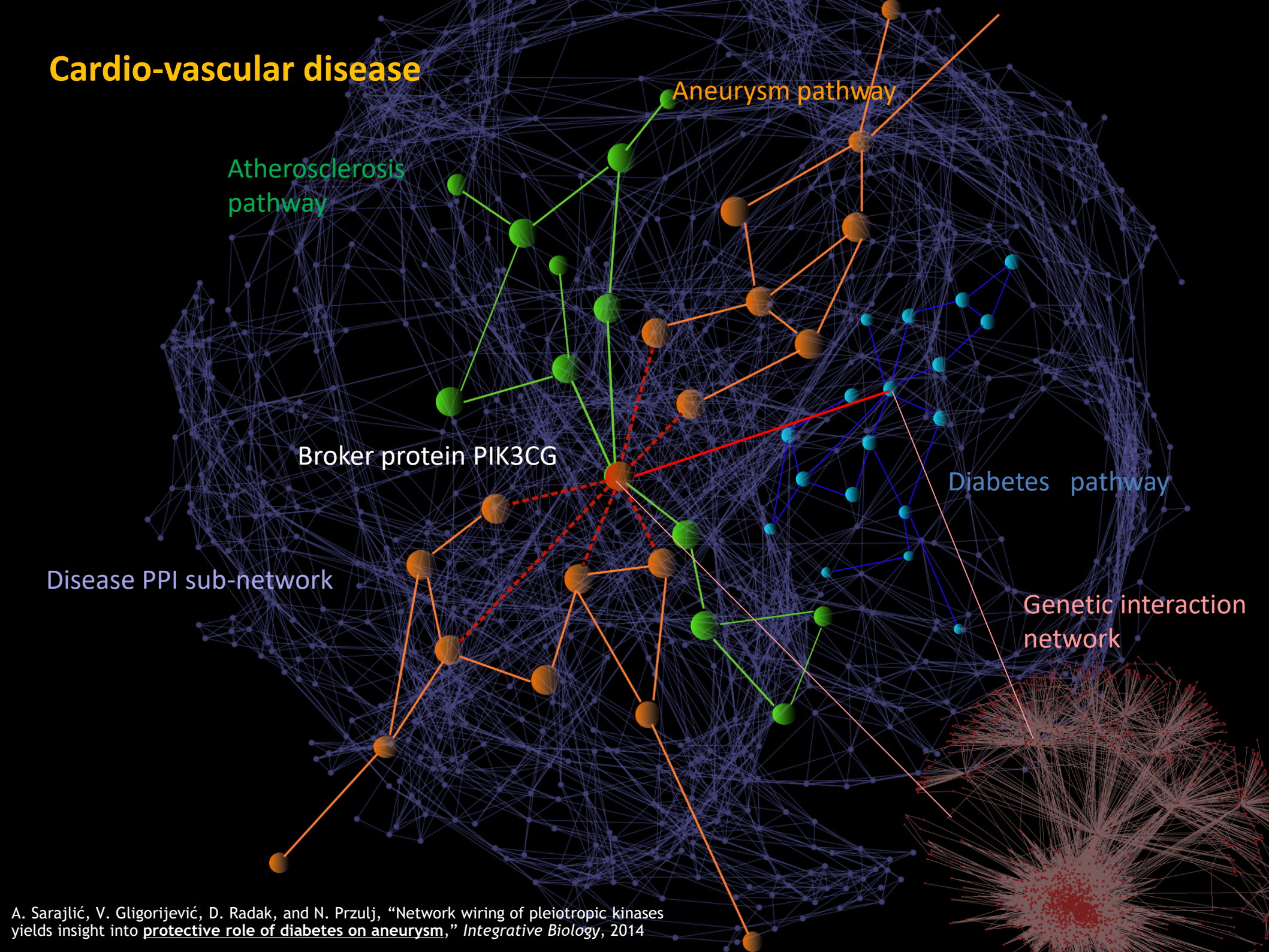
Why?
- Existing annotations ill-suited?
- Methodological limitations?

→ **Combine** topology and sequence information

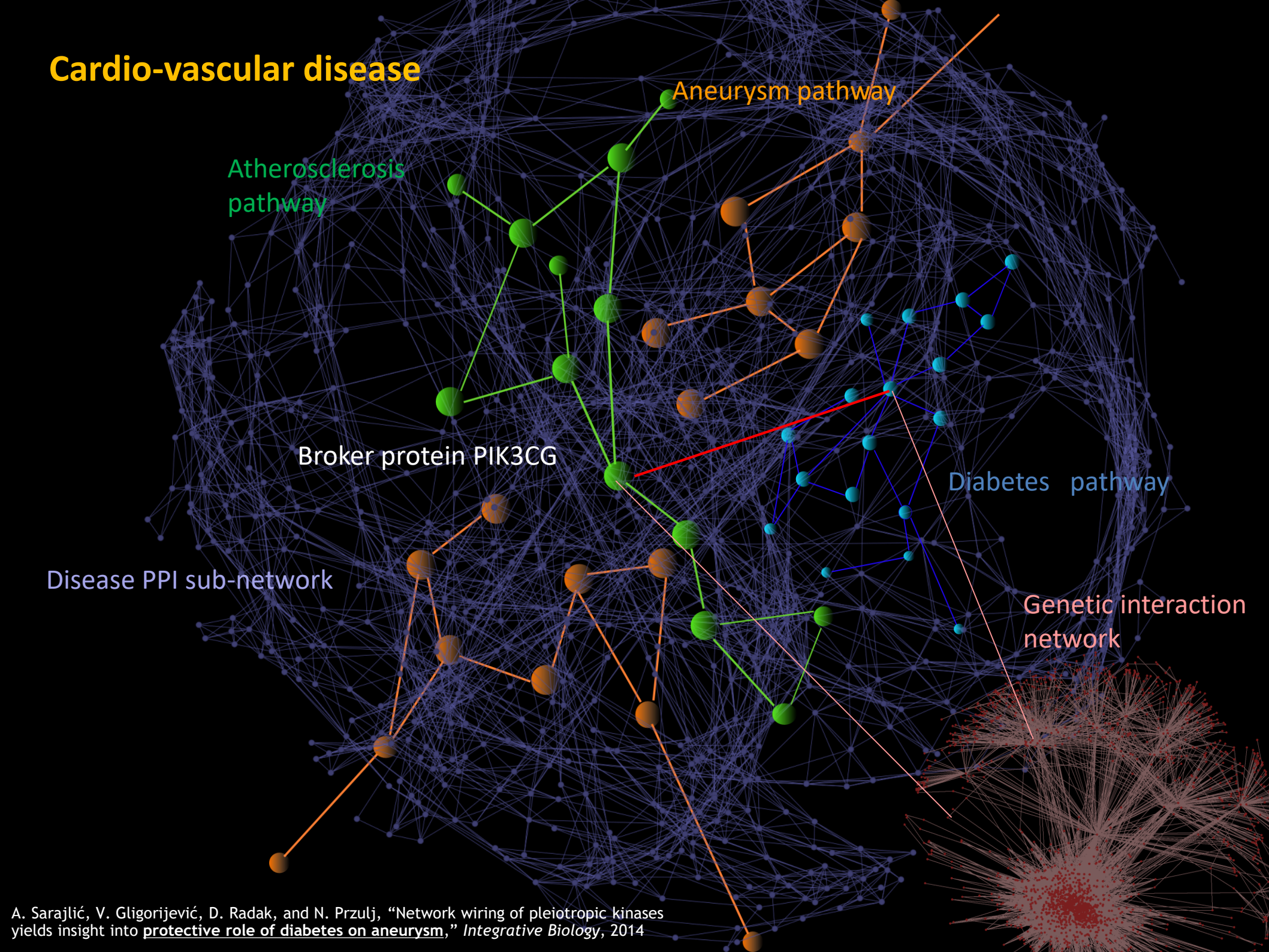
N. Malod-Dognin, K. Ban and **N. Przulj,** Unified Alignment of Protein-Protein Interaction Networks, *Scientific Reports - Nature*, 7:953, 2017




A. Sarajlić, V. Gligorijević, D. Radak, and N. Przulj, "Network wiring of pleiotropic kinases yields insight into **protective role of diabetes on aneurysm**," *Integrative Biology*, 2014




A. Sarajlić, V. Gligorijević, D. Radak, and N. Przulj, "Network wiring of pleiotropic kinases yields insight into **protective role of diabetes on aneurysm**," *Integrative Biology*, 2014

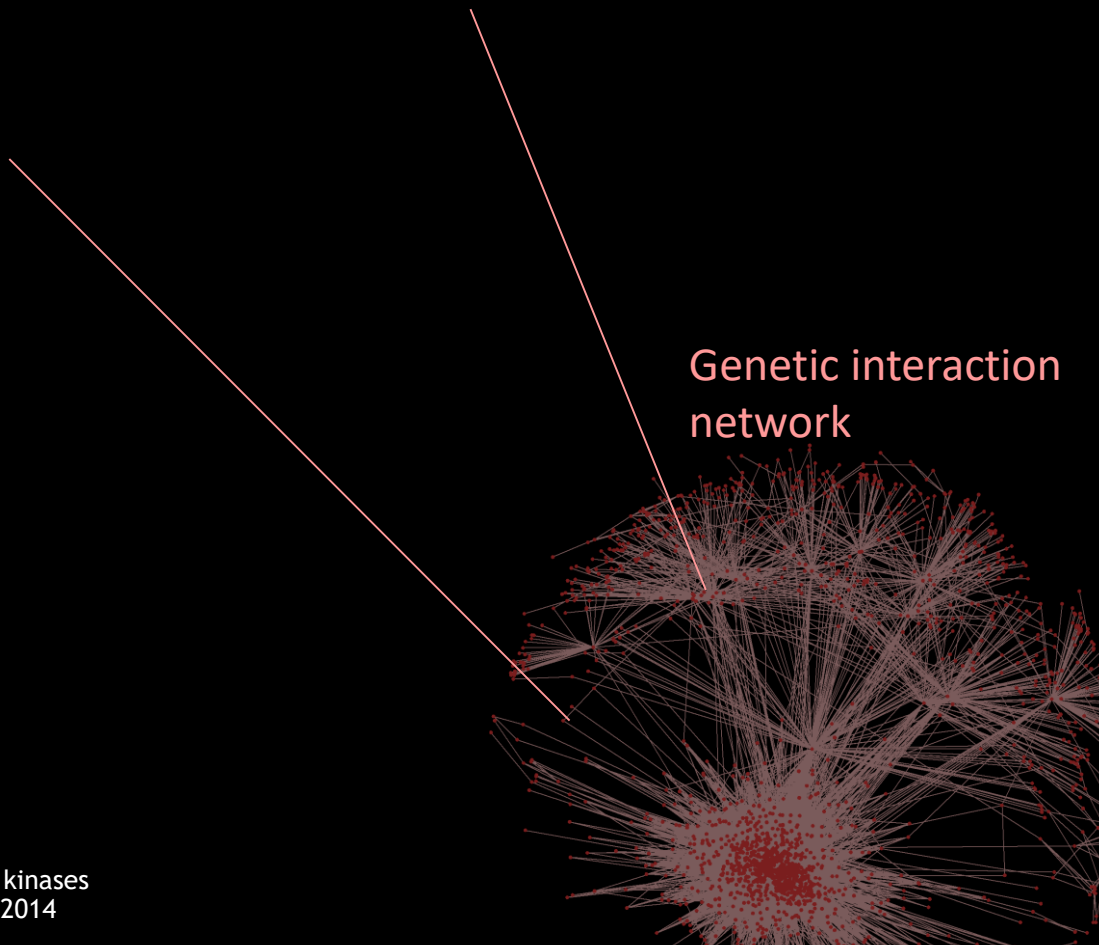



A. Sarajlić, V. Gligorijević, D. Radak, and N. Przulj, "Network wiring of pleiotropic kinases yields insight into **protective role of diabetes on aneurysm**," *Integrative Biology*, 2014

# Cardio-vascular disease

Disease PPI sub-network

Genetic interaction network



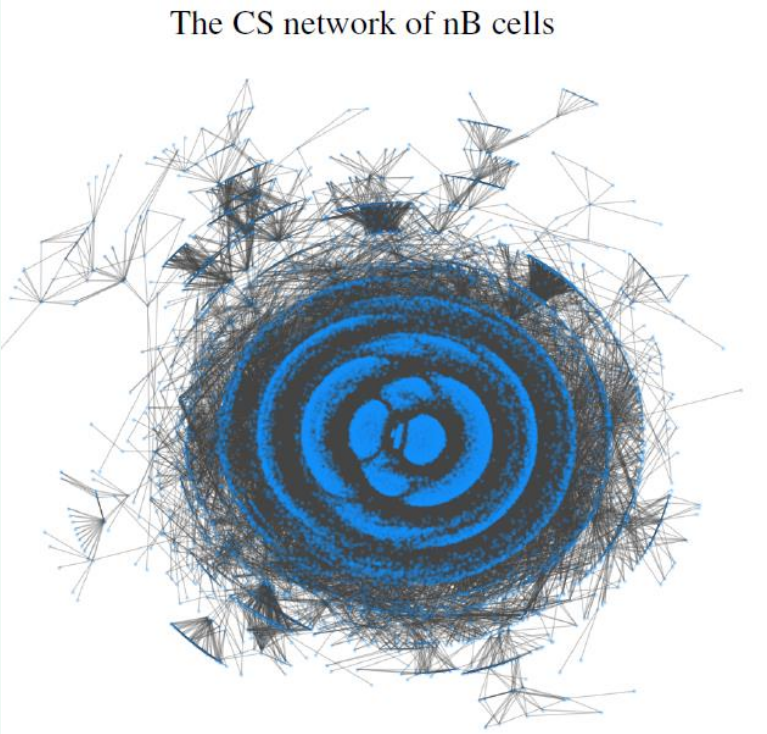
A. Sarajlić, V. Gligorijević, D. Radak, and N. Przulj, “Network wiring of pleiotropic kinases yields insight into **protective role of diabetes on aneurysm**,” *Integrative Biology*, 2014

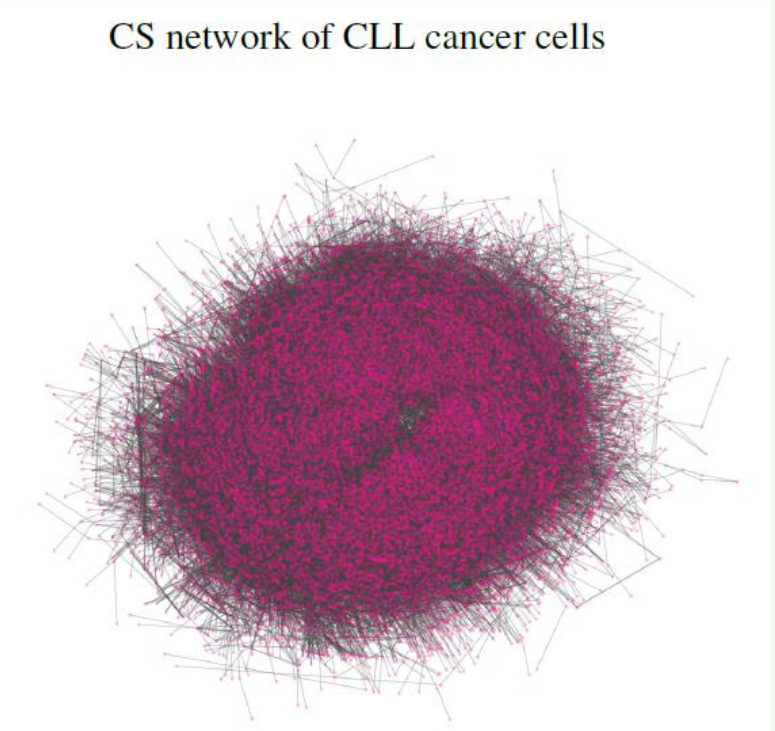
# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

**CLL Leukemia:**

Chromatin structure network: Hi-C data on CLL cells and nB control cells:

→ disrupted modular organization and functional coherence in CLL




Noel Malod-Dognin, Vera Pancaldi, Alfonso Valencia and **Natasa Przlj**: "Revealing structural markers of leukemia genes in chromatin structure networks," submitted

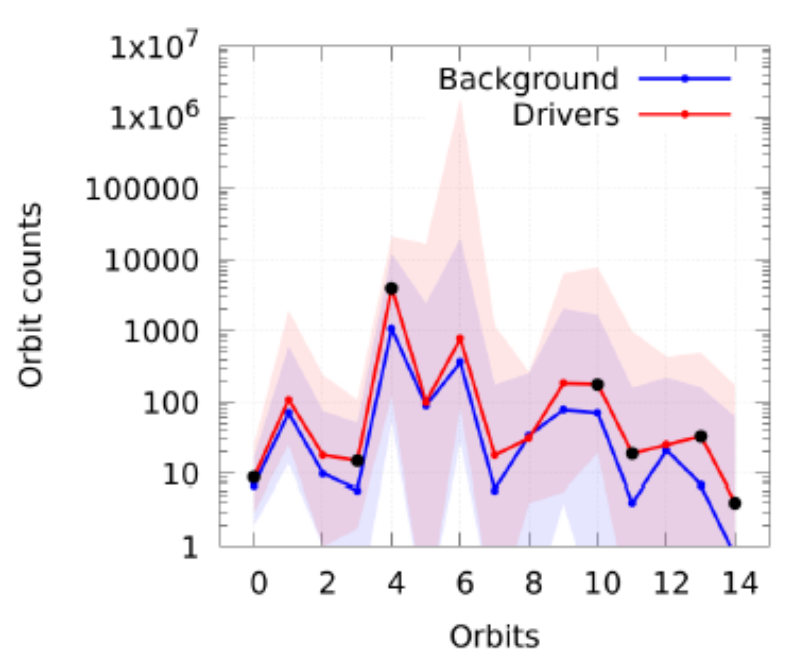
# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

**CLL Leukemia:**

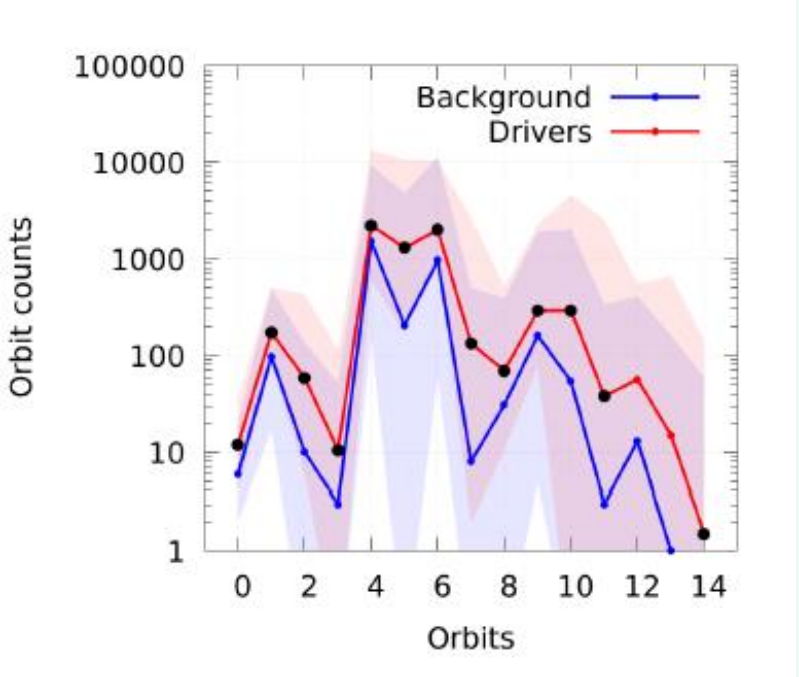
Chromatin structure network: Hi-C data on CLL cells and nB control cells:

→ BUT: the structural difference exists even before CLL!

nB control cells




CLL cells




Noel Malod-Dognin, Vera Pancaldi, Alfonso Valencia and **Natasa Przlj**: "Revealing structural markers of leukemia genes in chromatin structure networks," submitted

# 2. Novel Methods

## Mine Inter-Connected Entities: One Network Type

**Rare thrombophilia:**

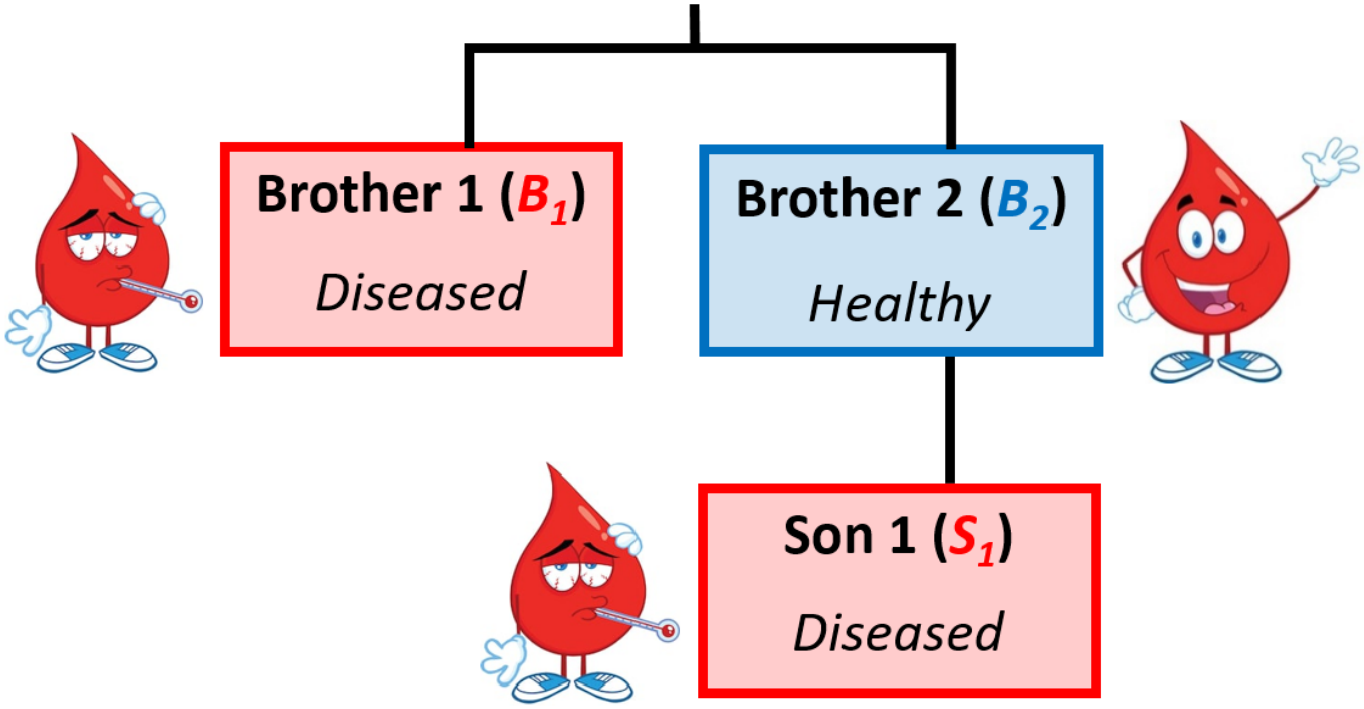



**Figure 1. Observed phenotypes.**

# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

- ✓ The best performing
- ✓ Robust
- ✓ …

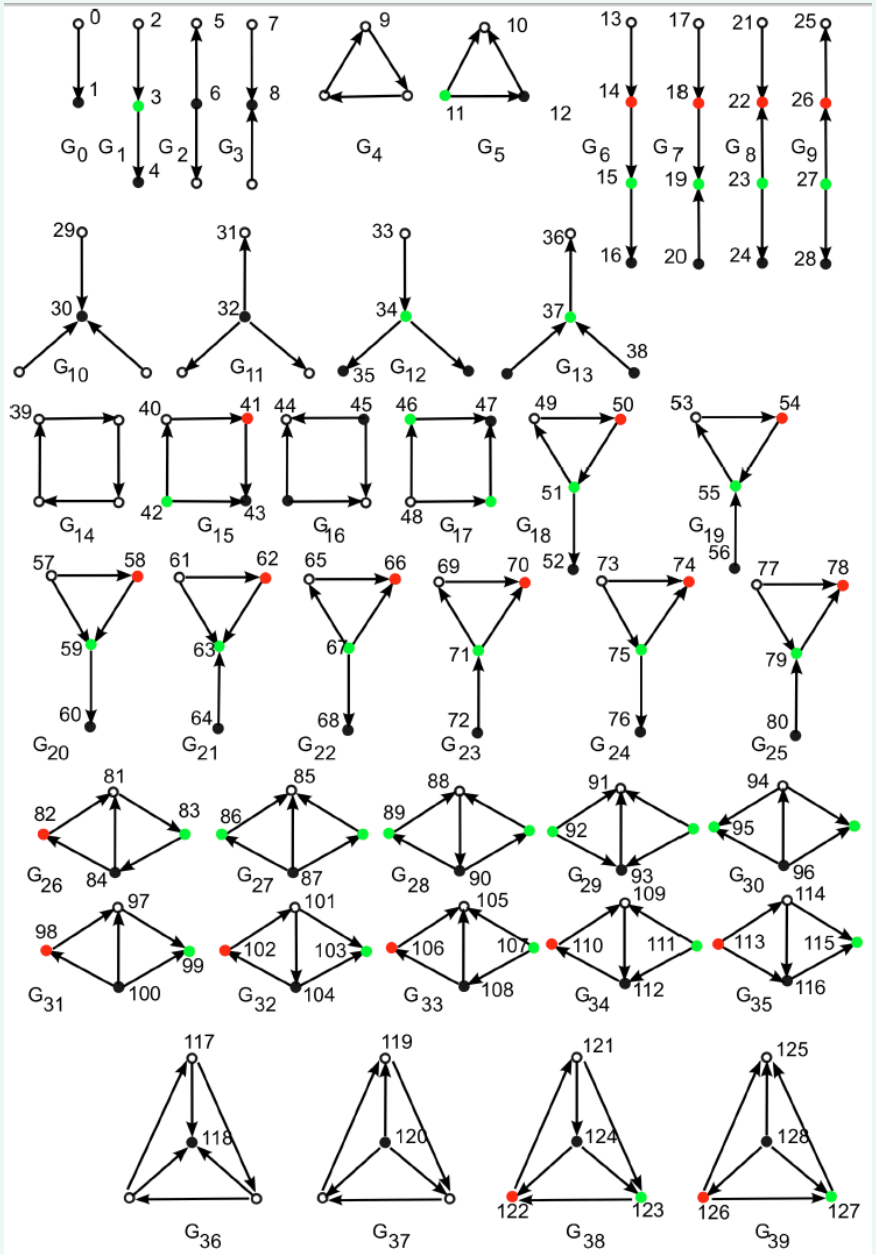
- ❑ PPI networks are ***geometric***

**N. Przulj**, D. G. Corneil, and I. Jurisica, "Modeling Interactome: Scale Free or Geometric?," *Bioinformatics*, vol. 20, num. 18, pg. 3508-3515, 2004.

**N. Przulj**, "Biological Network Comparison Using Graphlet Degree Distribution," Proceedings of the 2006 European Conference on Computational Biology, ECCB '06, Eilat, Israel, January 21-24, 2007, **acceptance rate 18%.** *Bioinformatics*, volume 23, pages e177-e183, 2007

…

- ➢ **Directed Networks**
- ➢ **Track dynamics**

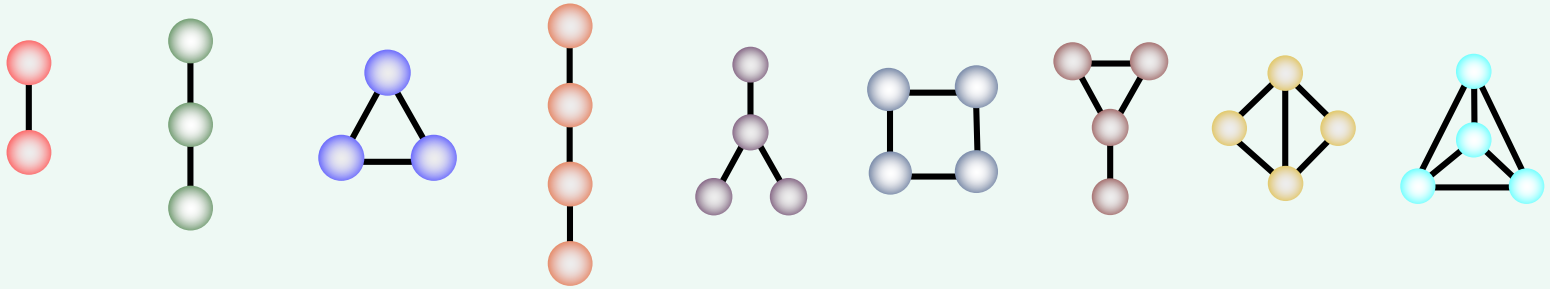


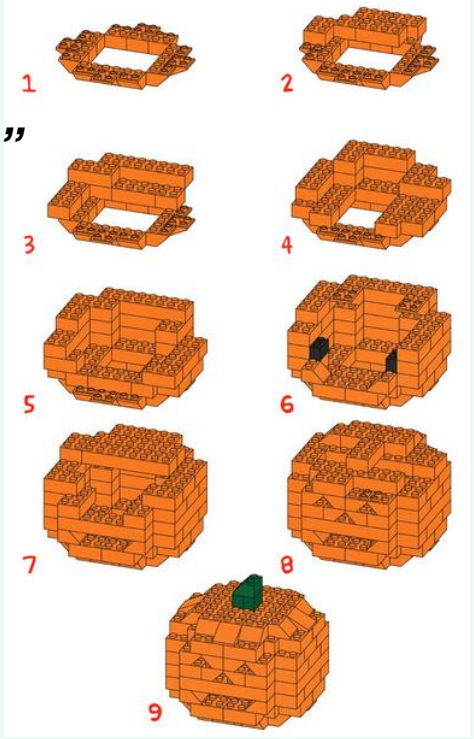

A. Sarajlic, N. Malod-Dognin, O. N. Yaveroglou, and **N. Przulj,** "Graphlet-based Characterization of Directed Networks," *Scientific Reports - Nature*, 6:35098, 2016

# 2. Novel Methods
## Mine Inter-Connected Entities: One Network Type

***Graphlets***

***"Legos of Networks"***






https://scholar.google.ca/scholar?hl=en&q=graphlets&btnG=&as_sdt= 90% Search

Web Images More... Sign in

Google graphlets

Scholar About 1,490 results (0.05 sec) My Citations

Articles
Case law
My library

Any time
Since 2017
Since 2016
Since 2013
Custom range...

Sort by relevance
Sort by date

include patents
include citations

**Discovering discriminative graphlets for aerial image categories recognition**
L Zhang, Y Han, Y Yang, M Song... - IEEE Transactions on ..., 2013 - ieeexplore.ieee.org
Abstract: Recognizing aerial image categories is useful for scene annotation and surveillance. Local features have been demonstrated to be robust to image transformations, including occlusions and clutters. However, the geometric property of an aerial image (ie,
Cited by 111 Related articles All 10 versions Cite Save

**Probabilistic graphlet transfer for photo cropping**
L Zhang, M Song, Q Zhao, X Liu, J Bu... - IEEE Transactions on ..., 2013 - ieeexplore.ieee.org
Abstract: As one of the most basic photo manipulation processes, photo cropping is widely used in the printing, graphic design, and photography industries. In this paper, we introduce graphlets (ie, small connected subgraphs) to represent a photo's aesthetic features, and
Cited by 95 Related articles All 16 versions Cite Save

**Model selection for social networks using graphlets**
J Janssen, M Hurshman, N Kalyaniwalla - Internet Mathematics, 2012 - Taylor & Francis
Several network models have been proposed to explain the link structure observed in online social networks. This paper addresses the problem of choosing the model that best fits a given real-world network. We implement a model-selection method based on unsupervised
Cited by 25 Related articles All 10 versions Cite Save

**Integrating local features into discriminative graphlets for scene classification**
L Zhang, W Bian, M Song, D Tao, X Liu - International Conference on ..., 2011 - Springer
Abstract Scene classification plays an important role in multimedia information retrieval. Since local features are robust to image transformation, they have been used extensively for scene classification. However, it is difficult to encode the spatial relations of local features in
Cited by 9 Related articles All 4 versions Cite Save

**Ego-centric graphlets for personality and affective states recognition**
S Teso, J Staiano, B Lepri, A Passerini... - Social Computing ( ..., 2013 - ieeexplore.ieee.org
Abstract: Do we tend to perceive ourselves more creative when surrounded by creative people? Or rather the opposite holds? Such information is very valuable to understand how to optimize work processes and boost people's productivity along with their happiness and
Cited by 5 Related articles All 14 versions Cite Save

**From quasirandom graphs to graph limits and graphlets**
F Chung - Advances in Applied Mathematics, 2014 - Elsevier
Abstract We generalize the notion of quasirandomness which concerns a class of equivalent properties that random graphs satisfy. We show that the convergence of a graph sequence under the spectral distance is equivalent to the convergence using the (normalized) cut
Cited by 4 Related articles All 10 versions Cite Save

# Overview

**Medicine: complex world of inter-connected entities**

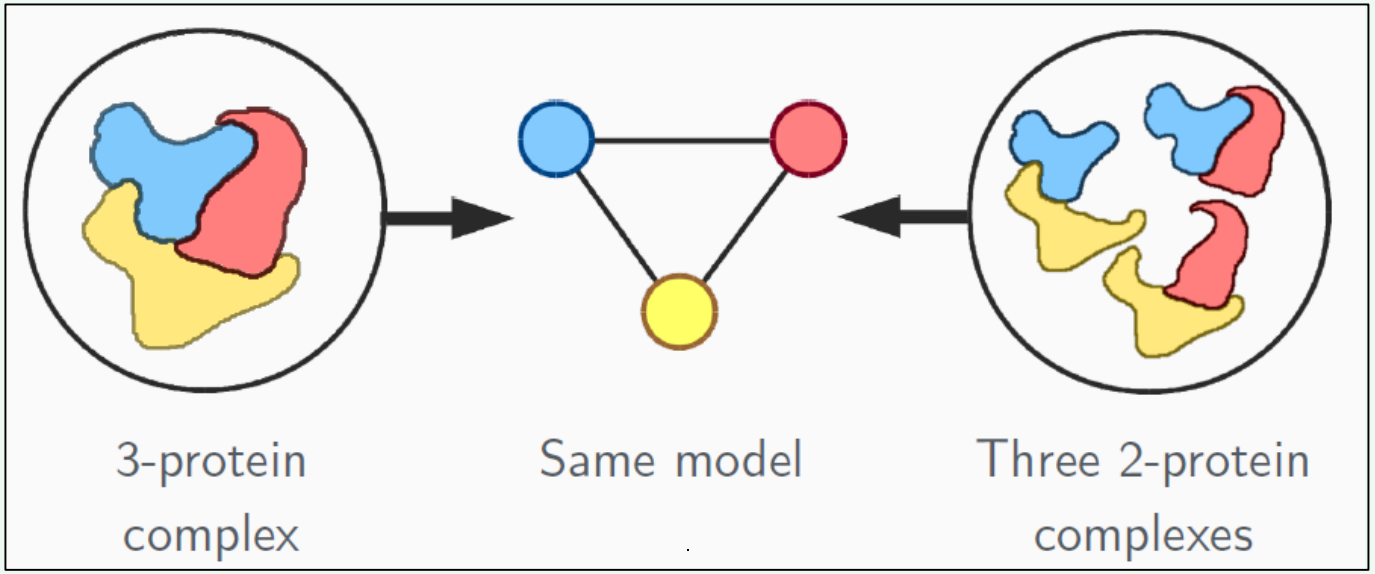
1. **Motivation**
2. **New Methods – Examples:** mine inter-connected data

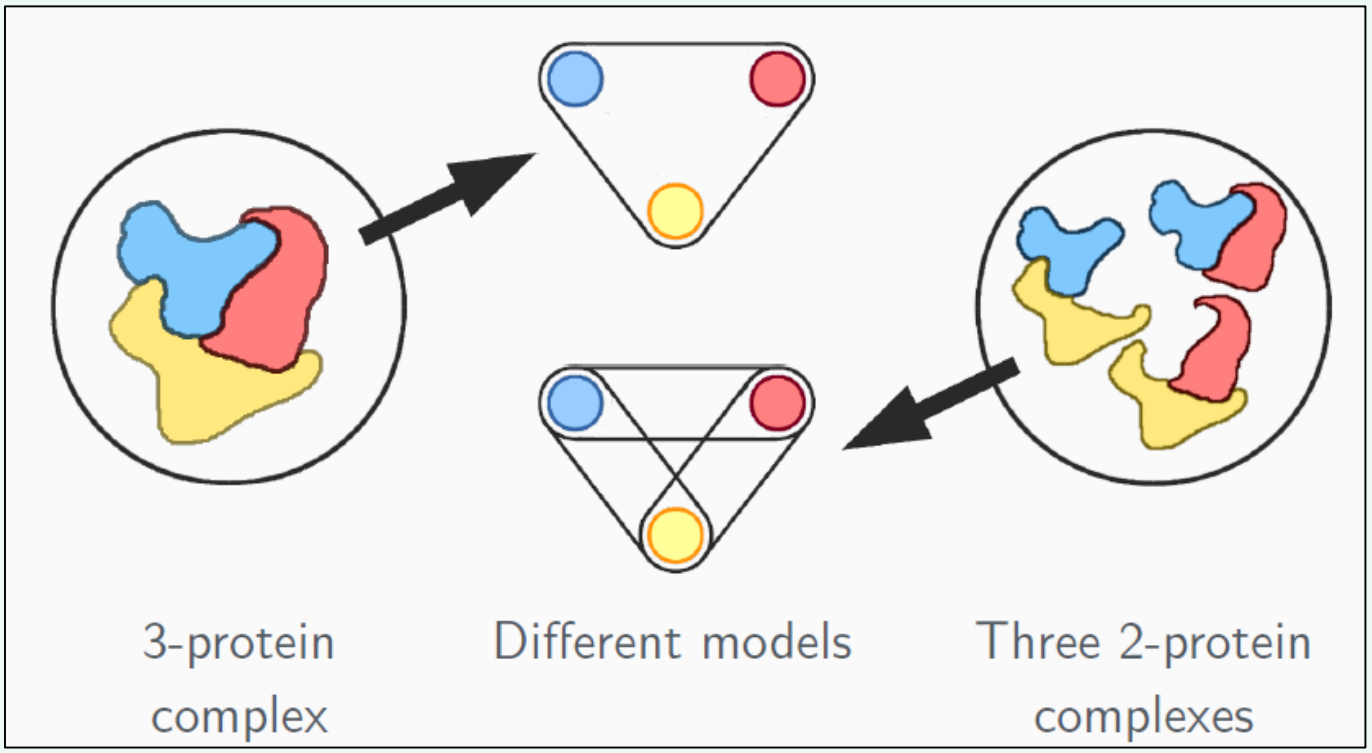
   i. **Single type of omics data:**
   - **Molecular networks**
   - **Multi-scale organization**

   } → function, disease

   ii. **Multiple layers of heterogeneous data:**
   - **iCell**
   - **Patient-centered data integration → Precision medicine**
     - ✓ Stratification, biomarker discovery, drug repurposing
   - **Disease re-classification, GO reconstruction, Network alignment, …**
3. **Conclusions**

# 2. Novel Methods
## Multi-scale organization



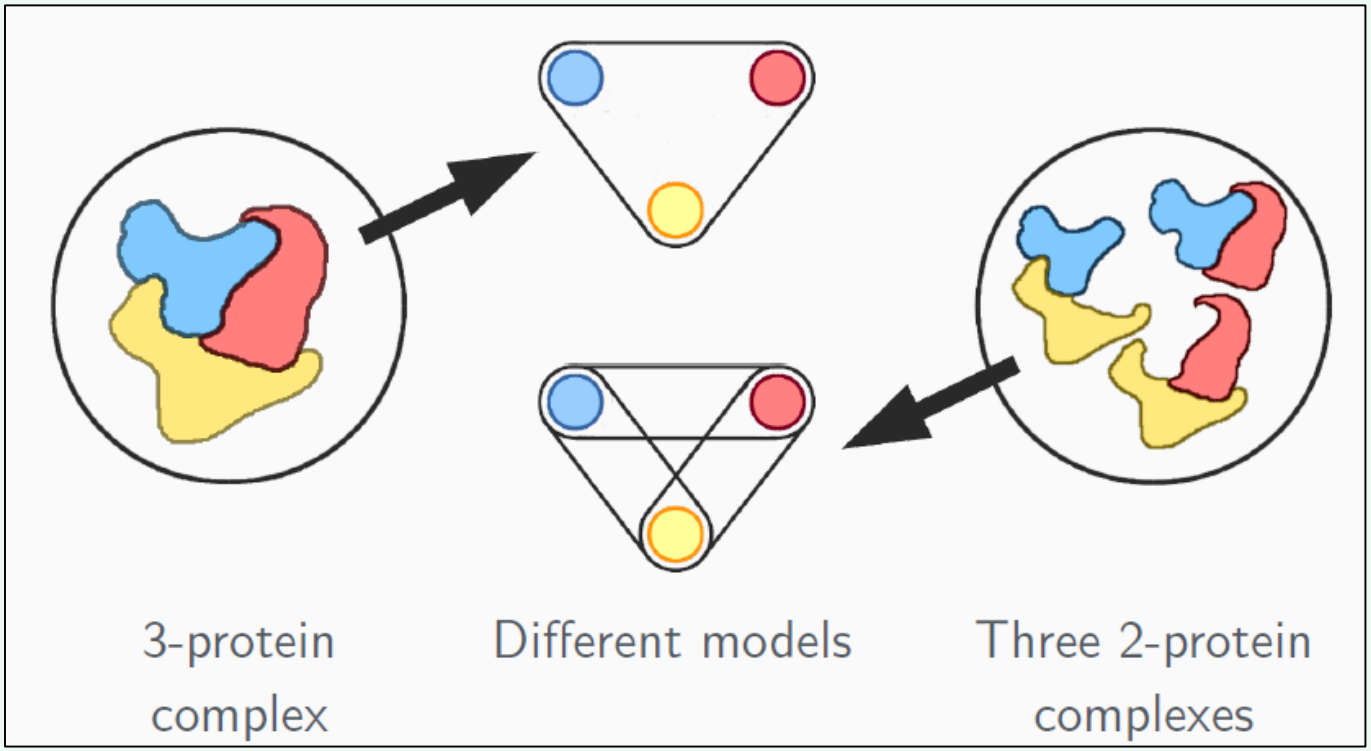
# 2. Novel Methods
## Multi-scale organization

**Hypergraphs**: extension of graphs (C. Berge, 1989)

- Hyperedges: sets of many nodes

**Hypergraphs in systems biology:**

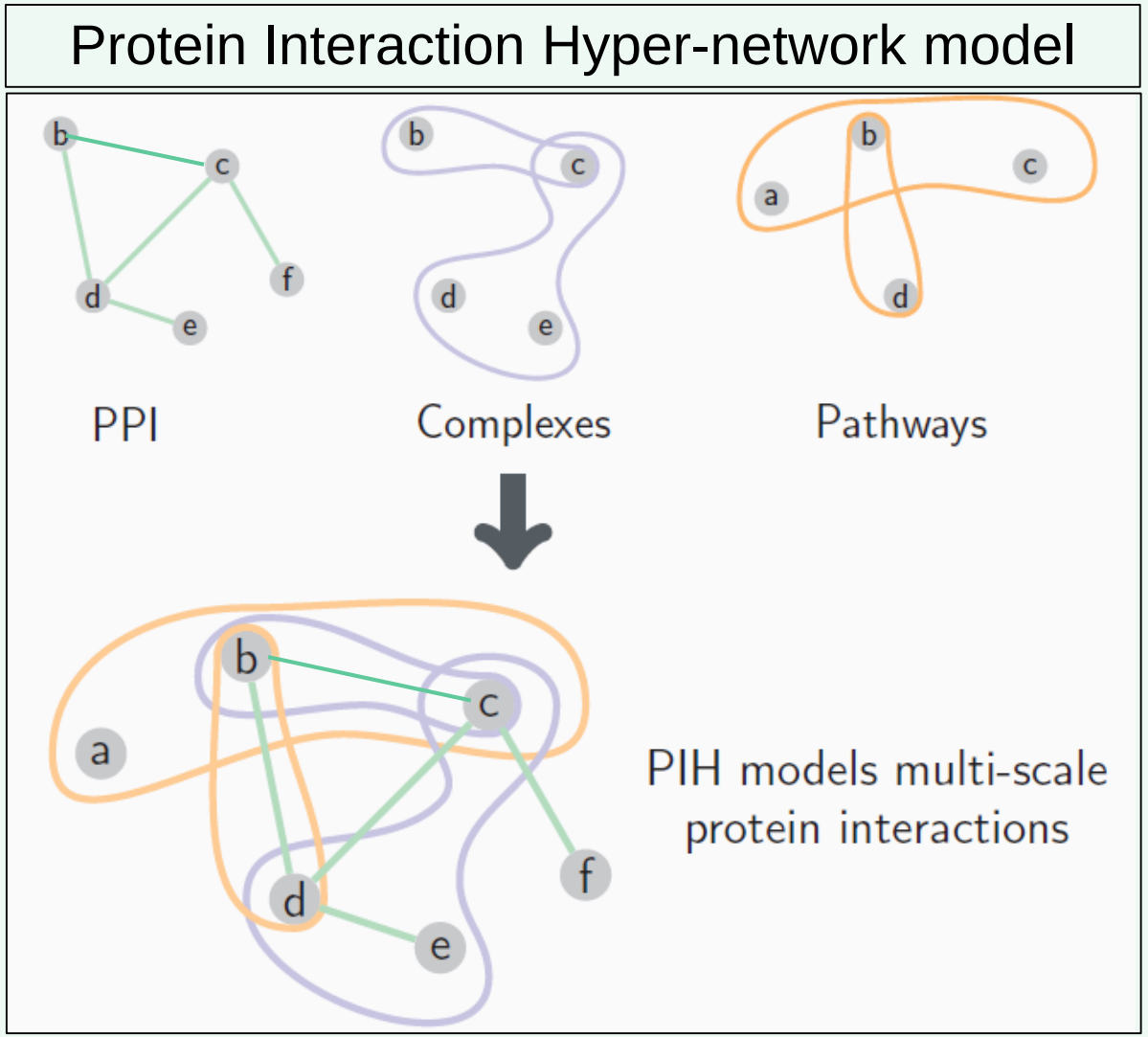
- ***Centrality*** and ***clustering*** [E. Estrada *et al.*, 2006]
- ***Degree distribution*** [M. Latapey *et al*., 2008]
- General modeling [S. Klamt *et al.*, . PLoS Comput Biol 5(5), 2009]



# 2. Novel Methods
## Multi-scale organization

***Hypergraphlets***

- Thomas Gaudelet *et al.*, ECCB 2018



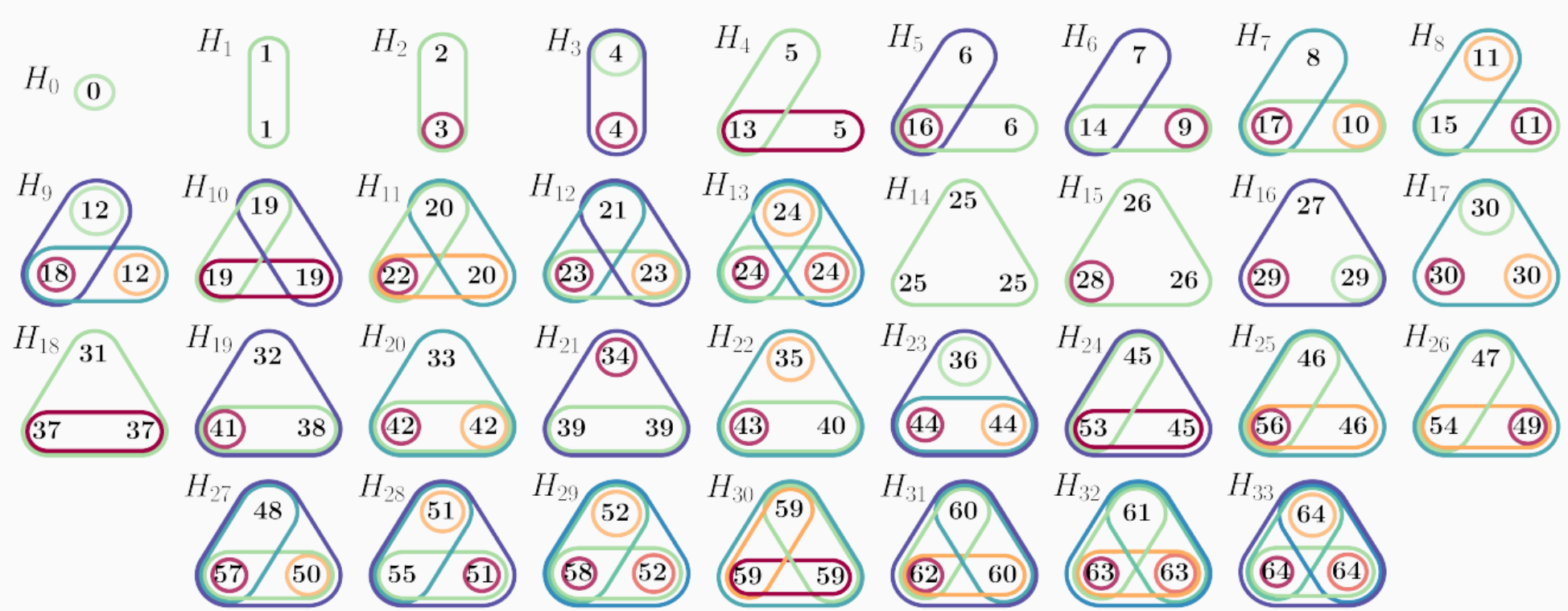

T. Gaudelet, N. Malod-Dognin and **N. Przulj**, “Higher order molecular organisation as a source of biological function,” *Bioinformatics*, ECCB’18

# 2. Novel Methods
## Multi-scale organization

***Hypergraphlets***

- Thomas Gaudelet *et al.*, ECCB 2018

***Hypergraphlet*** orbits:



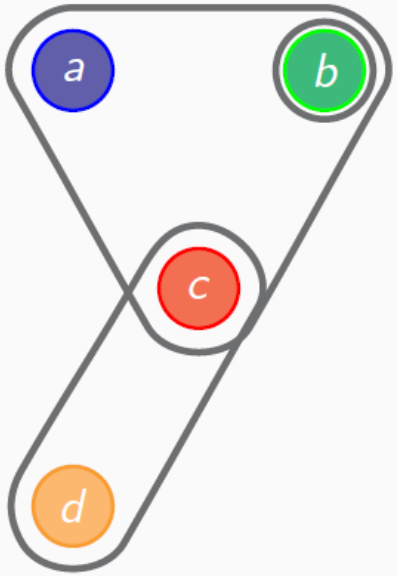

We consider up to 4-node hypergraphlets, having *6,369* orbits

T. Gaudelet, N. Malod-Dognin and **N. Przulj**, "Higher order molecular organisation as a source of biological function," *Bioinformatics*, ECCB'18

# 2. Novel Methods
## Multi-scale organization

### *Hypergraphlets*
➢ Thomas Gaudelet *et al.*, ECCB 2018

- The $i^{th}$ hypergraphlet degree of a vertex corresponds to the number of times the vertex belongs to orbit $i$
- The 6,369-dimensional feature vector obtained for each vertex is termed the *Hypergraphlet Degree Vector (HDV)*




| $C_0$ | $C_1$ | $C_2$ | ... | $C_5$ | ... | $C_7$ | ... |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 2 | ... | 1 | ... | 0 | ... |
| 2 | 0 | 0 | ... | 0 | ... | 0 | ... |
| 2 | 0 | 0 | ... | 0 | ... | 0 | ... |
| 1 | 0 | 1 | ... | 1 | ... | 1 | ... |

Hypergraphlet Degree Vectors (HDVs)
(number of columns = number of orbits)

T. Gaudelet, N. Malod-Dognin and **N. Przulj**, "Higher order molecular organisation as a source of biological function," *Bioinformatics*, ECCB'18

# 2. Novel Methods
## Multi-scale organization

***Hypergraphlets***

➢ Thomas Gaudelet *et al.*, ECCB 2018

**Identifying links between vertices' topology and functions:**

- Canonical Correlation Analysis (CCA)
  - Finds linear connections between HDVs and biological annotations

**Measuring if proteins with similar wirings have similar biological functions:**

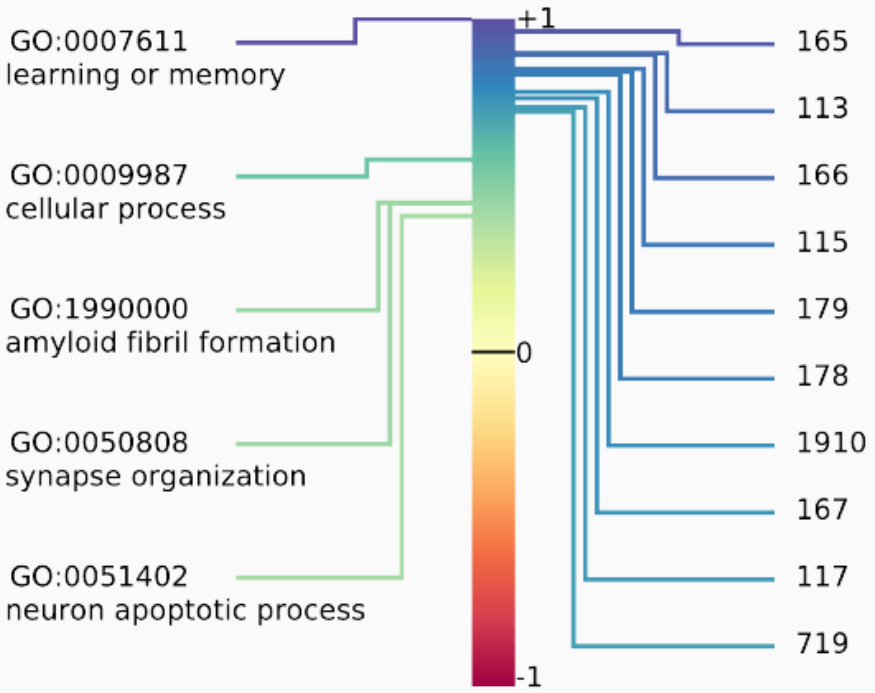
- HDV-based clustering & enrichment analysis
  - k-means clustering algorithm
  - enrichment measured with respect to biological annotations

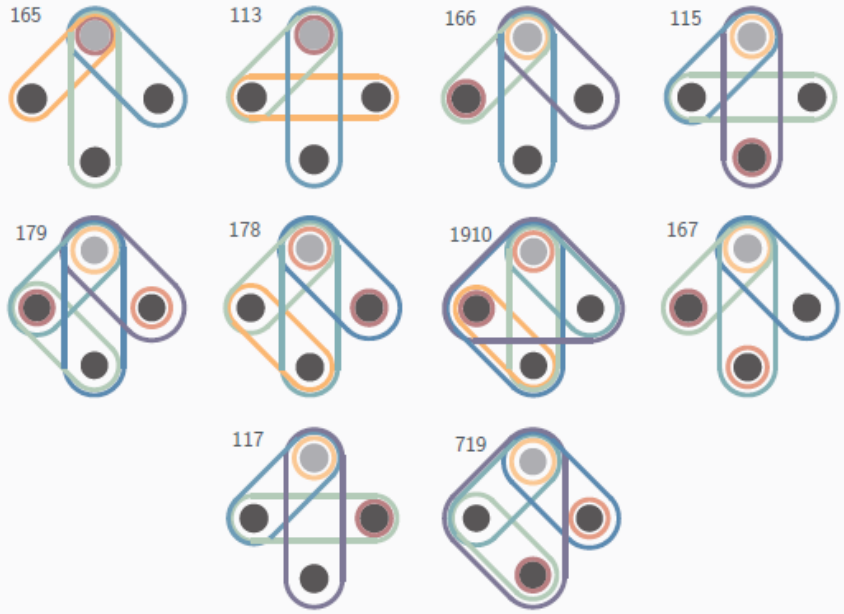
T. Gaudelet, N. Malod-Dognin and **N. Przulj**, "Higher order molecular organisation as a source of biological function," *Bioinformatics*, ECCB'18

# 2. Novel Methods
## Multi-scale organization

***Hypergraphlets***

➢ Thomas Gaudelet *et al.*, ECCB 2018



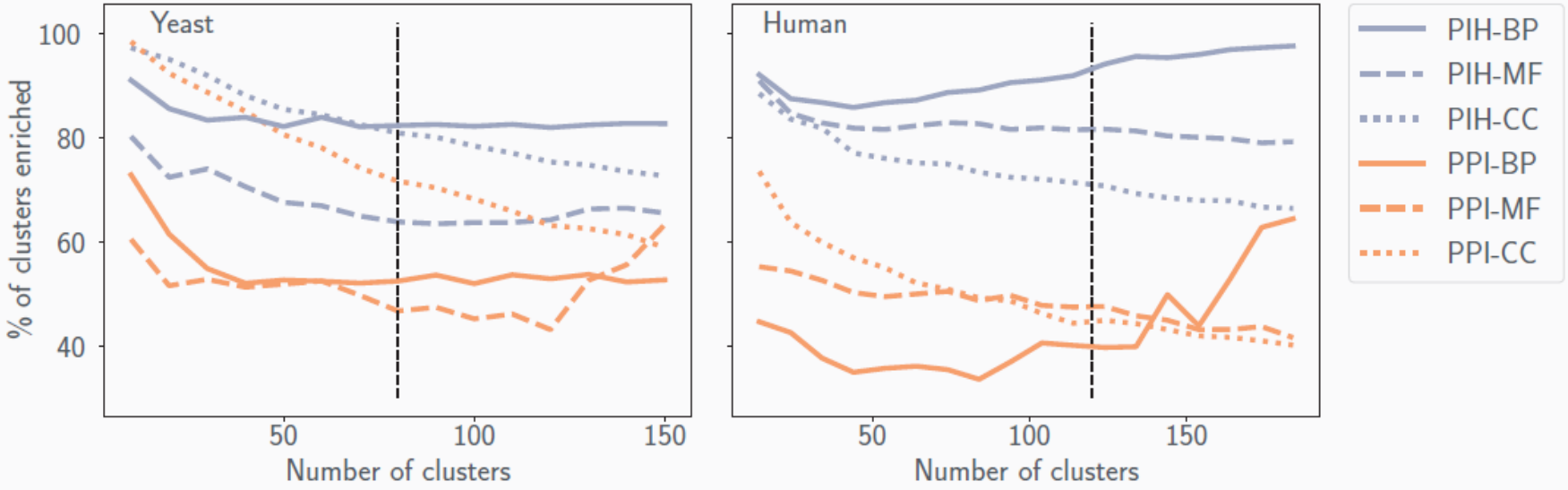

- Strong links between specific orbits and GO–BP annotations
- Functions identified are related

T. Gaudelet, N. Malod-Dognin and **N. Przulj**, “Higher order molecular organisation as a source of biological function,” *Bioinformatics*, ECCB’18

# 2. Novel Methods
## Multi-scale organization

### *Hypergraphlets*

➢ Thomas Gaudelet *et al.*, ECCB 2018




| | Biological Process | | Molecular Function | | Cellular Component | |
|---|---|---|---|---|---|---|
| | **PIH** | PPI | **PIH** | PPI | **PIH** | PPI |
| Yeast | **91.1%** (79) | 68.75% (80) | **71.8%** (79) | 52.5% (80) | **88.6%** (79) | 72.2% (79) |
| Human | **100.0%** (105) | 41.7% (120) | **95.0%** (98) | 55.8% (120) | **76.1%** (105) | 51.7% (120) |

- Proteins with similar wiring have related biological functions
- PIH representations lead to higher enrichments values

T. Gaudelet, N. Malod-Dognin and **N. Przulj**, "Higher order molecular organisation as a source of biological function," *Bioinformatics*, ECCB'18

# 2. Novel Methods
## Multi-scale organization

***Hypergraphlets***

- Thomas Gaudelet *et al.*, ECCB 2018

We associate to each uncharacterised protein the most significantly enriched GO annotations in its cluster. We validate our predictions through litterature curation.

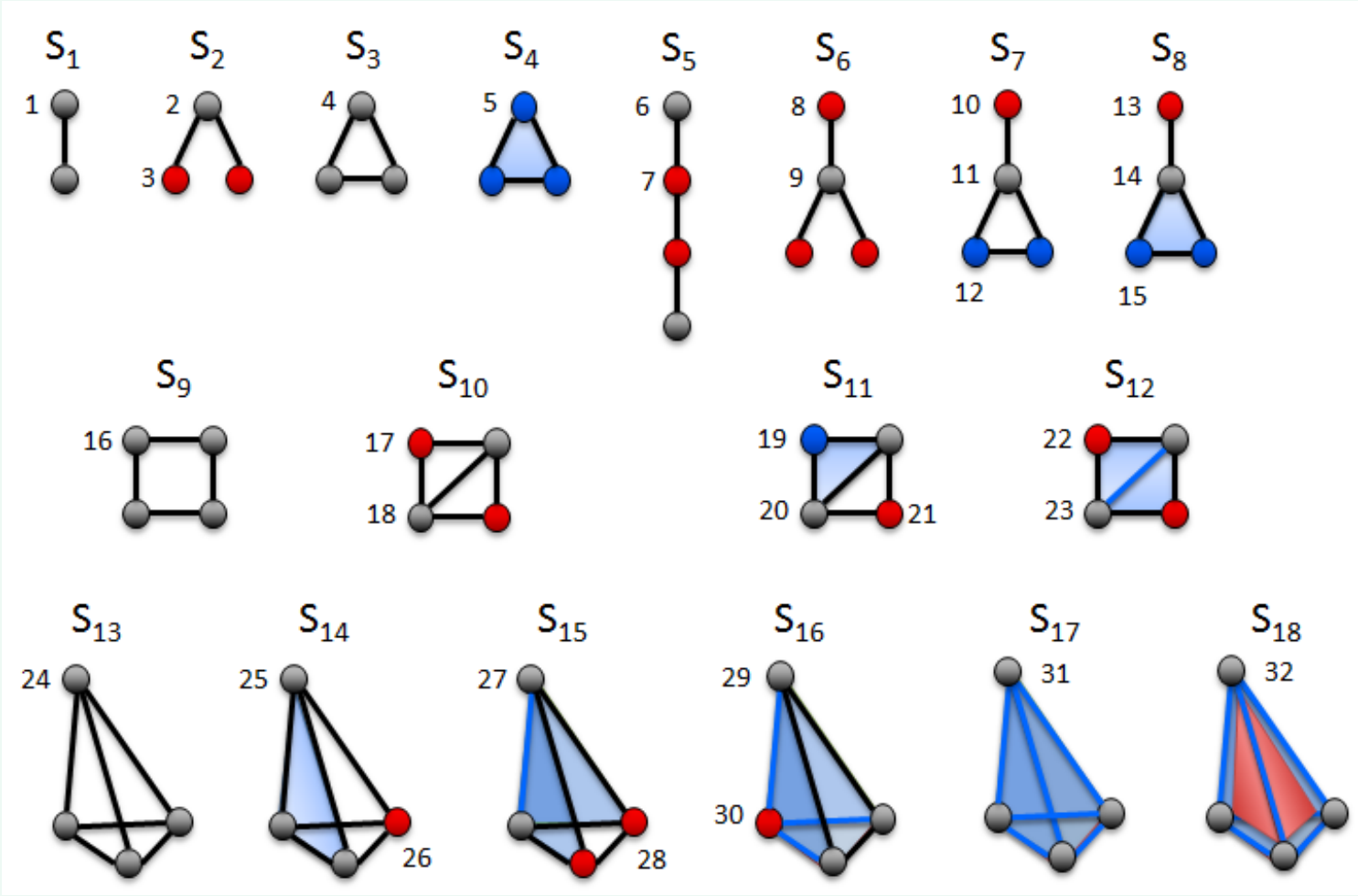
| GO ID | GO function | Proteins Symbol |
|---|---|---|
| **GO:0006334** | **nucleosome assembly** | **HIST1H2AJ** |
| GO:0001580 | detection of chemical stimulus involved in sensory perception of bitter taste | LOC107987462; LOC107987425; LOC102725035 |
| **GO:0006364** | **rRNA processing** | **LOC101929876** |
| GO:0035987 | endodermal cell differentiation | MIR711 |
| GO:0051292 | nuclear pore complex assembly | MIR4260 |
| **GO:0016579** | **protein deubiquitination** | **MIR6764** |
| **GO:0030199** | **collagen fibril organization** | **MIR3606** |
| **GO:0030216** | **keratinocyte differentiation** | **KRTAP4-7** |
| GO:0006997 | nucleus organization | LOC101060521; MIR1181 |
| **GO:0052695** | **cellular glucuronidation** | **UGT2A2**; LOC102724788; GUCY2EP |

T. Gaudelet, N. Malod-Dognin and **N. Przulj**, “Higher order molecular organisation as a source of biological function,” *Bioinformatics*, ECCB’18

# 2. Novel Methods
## Multi-scale organization

***Simplets***

- Noel Malod-Dognin & N. Przulj, *Bioinformatics*, 2019




N. Malod-Dognin and **N. Przulj**, "Functional geometry of protein interactomes," *Bioinformatics*, btz146, 01 March 2019

# 2. Novel Methods
## Multi-scale organization

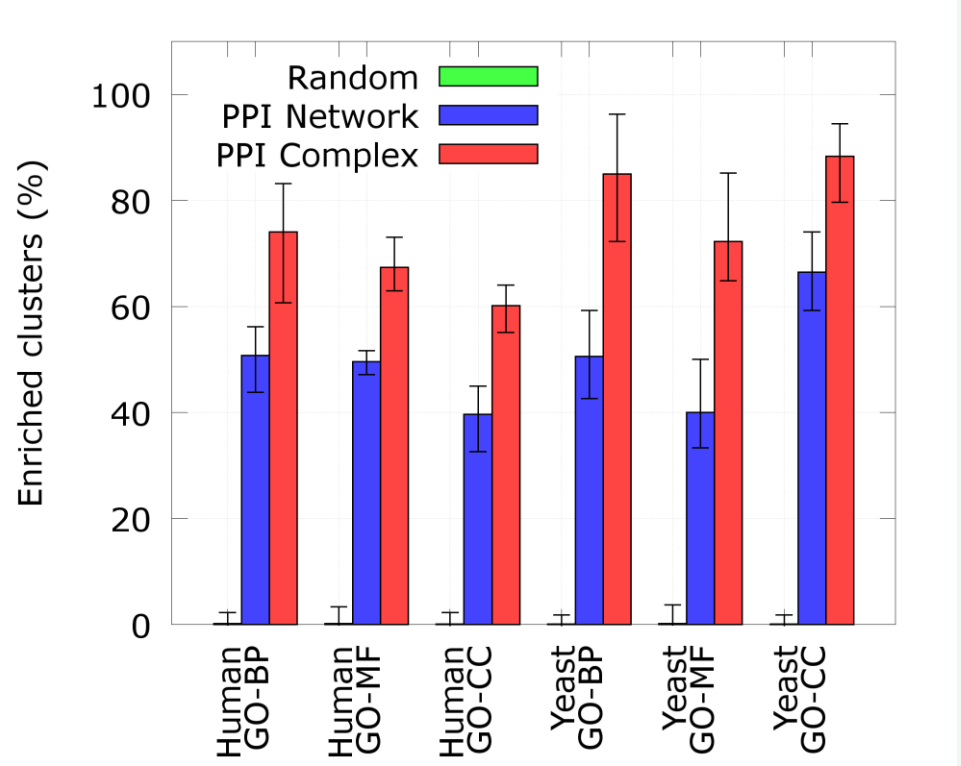
### *Simplets*

➢ Noel Malod-Dognin & N. Przulj, *Bioinformatics*, 2019

We compare two models:

- **PPI network**: 1-dimensional simplex
- **PPI complex**:
  - PPI network
  - additionally connect by simplices all the proteins in a common complex
- **Human** and **yeast**
- Cluster proteins in **PPI** and in "**PPI Complex"** according to the similarity of their wiring patterns (based on simplets)

➢ **RESULT:**
**Clusters of genes in "PPI Complex" better functional enrichments**




N. Malod-Dognin and **N. Przulj**, "Functional geometry of protein interactomes," *Bioinformatics*, btz146, 01 March 2019

# Overview

**Medicine: complex world of inter-connected entities**

1. **Motivation**
2. **New Methods – Examples:** mine inter-connected data

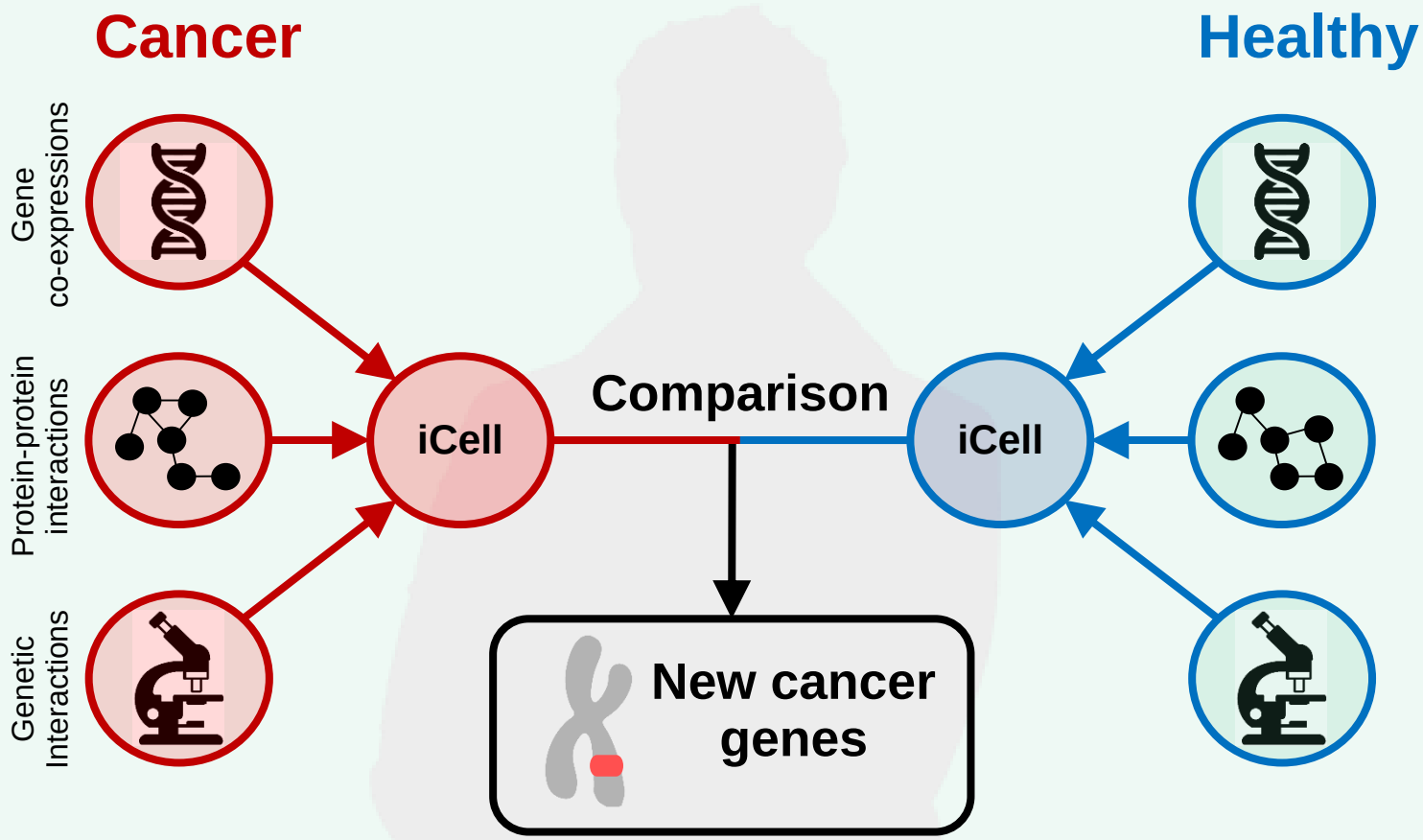
   i. **Single type of omics data:**
   - **Molecular networks**
   - **Multi-scale organization**

   } → function, disease

   ii. **Multiple layers of heterogeneous data:**
   - **iCell**
   - **Patient-centered data integration → Precision medicine**
     - ✓ Stratification, biomarker discovery, drug repurposing
   - **Disease re-classification, GO reconstruction, Network alignment, …**
3. **Conclusions**

# 2. Novel Methods

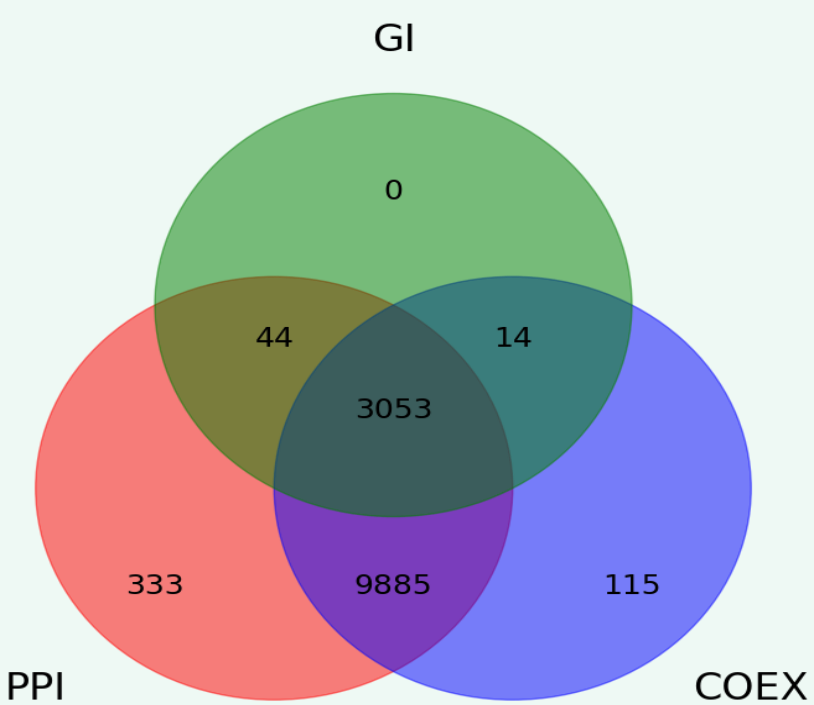
## iCell: Tissue-specific integration of heterogeneous omics data



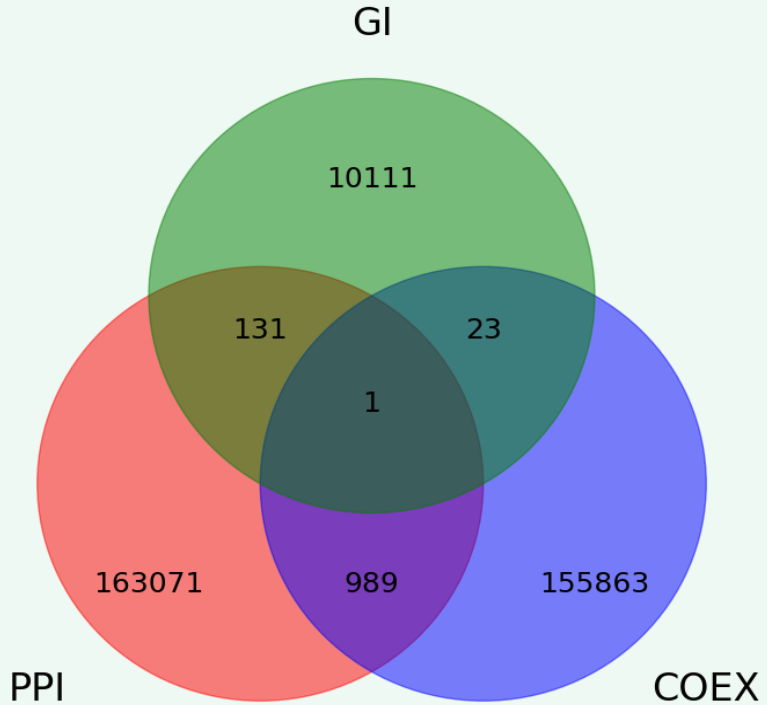

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Different data types **complement** each other




Genes




Interactions

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Different data types **complement** each other

→ **Most data integration methods fail**

- ❑ Similarity network fusion (SNF): returns an empty integrated net.
- ❑ Natural Gradient Weighted Simultaneous Symmetric NMTF:
  - diverges after ≈100 iterations;
  - if stopped before clusters not functionally consistent
- ❑ GraphFuse (tensor factorization): memory issues, can't process out data
- ❑ Spectral clustering on multi-layer graphs (SC-ML):
  - Doesn't converge, doesn't produce clusters
- ❑ Markov CLustering (MCL):
  - Very large number of very small clusters
  - Many nodes (genes) left isolated

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Different data types **complement** each other

→ **Create tissue specific networks**

1. **PPI**s: from IID
2. **COEX:** from COEXPRESdb
3. **GI**s: from BioGRID and SynLethDB
4. **Tissue-specific gene expression:** Human Protein Atlas (HPA)
   - Not microarray, or RNA-seq
   - It's antibody staining → low throughput, low number of samples
   - Doesn't represent all patients, just a small number of them

Only consider genes with:
- expression available in HPA and
- >1 PPI in IID

**Create tissue-specific networks**:
- **nodes** are genes **expressed in the tissue (from HPA)**
- **linked** by interactions

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Different data types **complement** each other

→ **Create tissue specific networks**

**Cancers**:
- Breast
- Prostate
- Lung
- Colorectal

**Controls**:
- Breast glandular cells
- Prostate glandular cells
- Lung pneumocytes
- Colorectal glandular cells

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Different data types **complement** each other

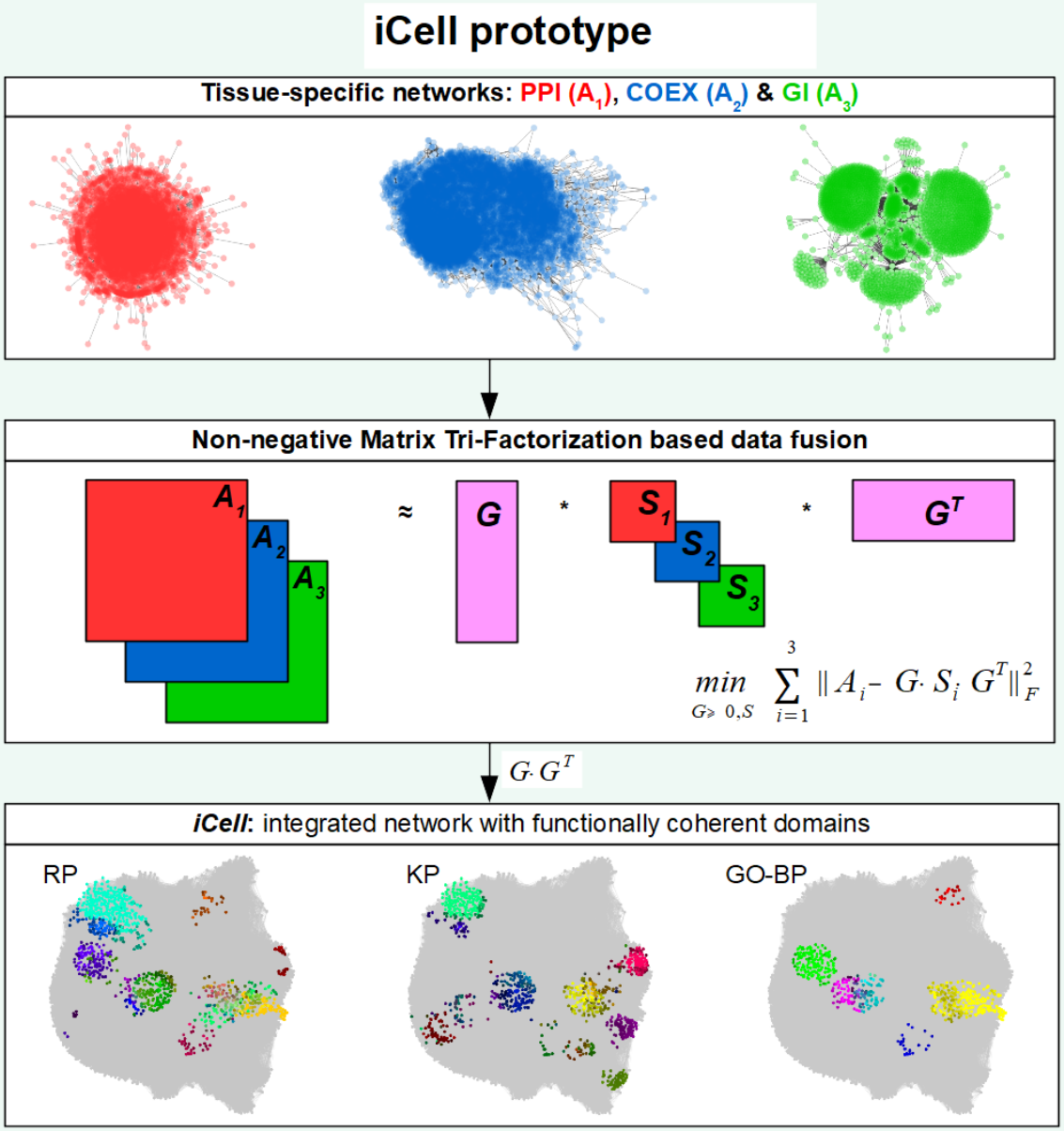
→ **Create tissue specific**

**Cancers**:
- Breast
- Prostate
- Lung
- Colorectal

| | Network statistics | | | | | |
|---|---|---|---|---|---|---|
| | PPI | | COEX | | GI | |
| Tissue | #Node | #Edge | #Node | #Edge | #Node | #Edge |
| Breast control | 9,188 | 106,198 | 9,233 | 84,930 | 2,269 | 4,998 |
| Prostate control | 8,963 | 97,699 | 9,051 | 81,649 | 2,189 | 5,543 |
| Lung control | 6,753 | 63,087 | 7,022 | 50,184 | 1,658 | 3,204 |
| Colon control | 10,257 | 120,851 | 10,263 | 103,106 | 2,487 | 6,766 |
| Breast cancer * | 8,260 | 93,416 | 8,378 | 74,147 | 2,027 | 4,679 |
| Carcinoid | 8,064 | 85,693 | 8,242 | 69,852 | 1,981 | 4,603 |
| Cervical cancer | 7,137 | 77,122 | 7,303 | 58,874 | 1,790 | 3,984 |
| Colorectal cancer * | 8,760 | 100,196 | 8,844 | 80,902 | 2,206 | 5,981 |
| Endometrial cancer | 7,632 | 82,061 | 7,788 | 64,467 | 1,825 | 4,210 |
| Glioma | 6,467 | 68,374 | 6,672 | 48,599 | 1,464 | 2,826 |
| Head and neck cancer | 8,446 | 97,078 | 8,554 | 75,823 | 2,154 | 5,440 |
| Liver cancer | 7,632 | 75,625 | 7,833 | 63,646 | 1,843 | 4,253 |
| Lung cancer * | 6,839 | 70,437 | 6,980 | 53,857 | 1,738 | 3,724 |
| Lymphoma | 5,373 | 53,498 | 5,599 | 38,831 | 1,363 | 2,693 |
| Melanoma | 7,672 | 83,714 | 7,818 | 65,731 | 1,884 | 3,856 |
| Ovarian cancer | 7,915 | 86,299 | 8,065 | 69,074 | 1,937 | 4,326 |
| Pancreatic cancer | 8,187 | 89,535 | 8,300 | 71,938 | 1,976 | 4,947 |
| Prostate cancer * | 7,675 | 79,969 | 7,851 | 64,625 | 1,890 | 5,122 |
| Renal cancer | 5,983 | 52,481 | 6,237 | 41,114 | 1,459 | 2,982 |
| Skin cancer | 6,549 | 70,117 | 6,719 | 51,275 | 1,683 | 3,586 |
| Stomach cancer | 7,409 | 79,488 | 7,575 | 62,078 | 1,866 | 4,585 |
| Testis cancer | 7,127 | 78,498 | 7,269 | 58,912 | 1,760 | 3,793 |
| Thyroid cancer | 9,213 | 104,463 | 9,301 | 86,323 | 2,256 | 5,951 |
| Urothelial cancer | 7,733 | 86,519 | 7,852 | 66,547 | 1,952 | 4,080 |

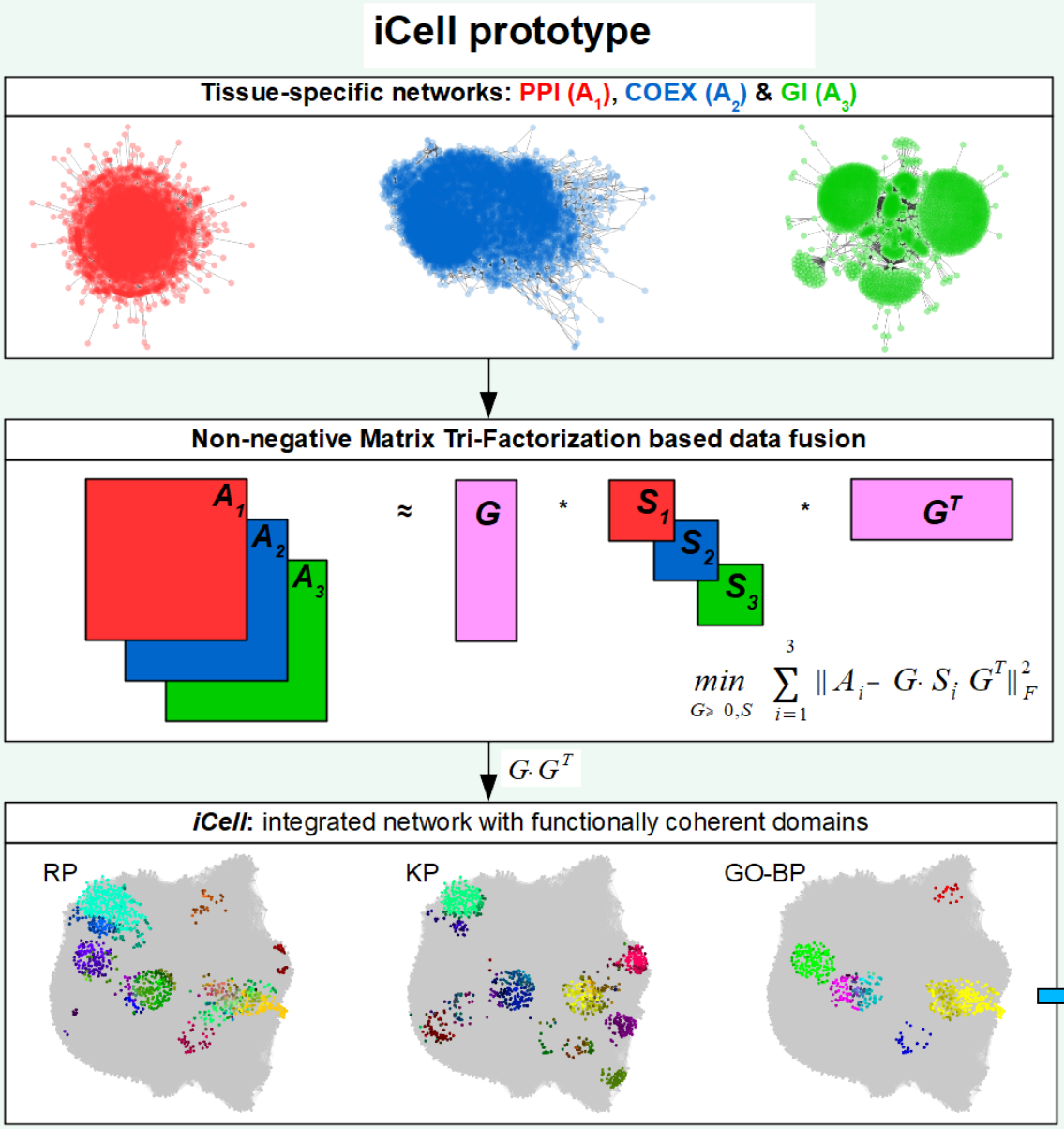
Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data




Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

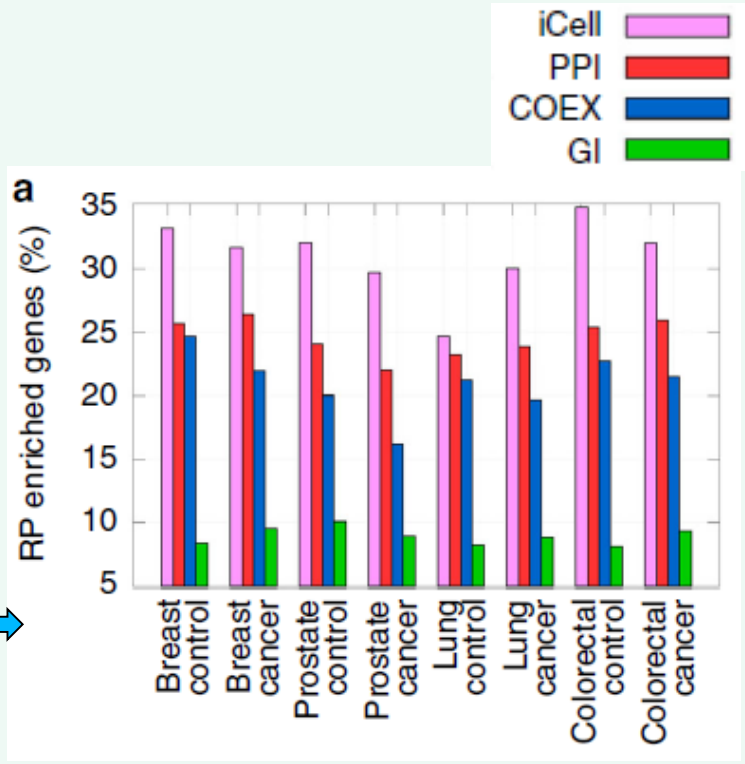
# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data




**iCells capture:**

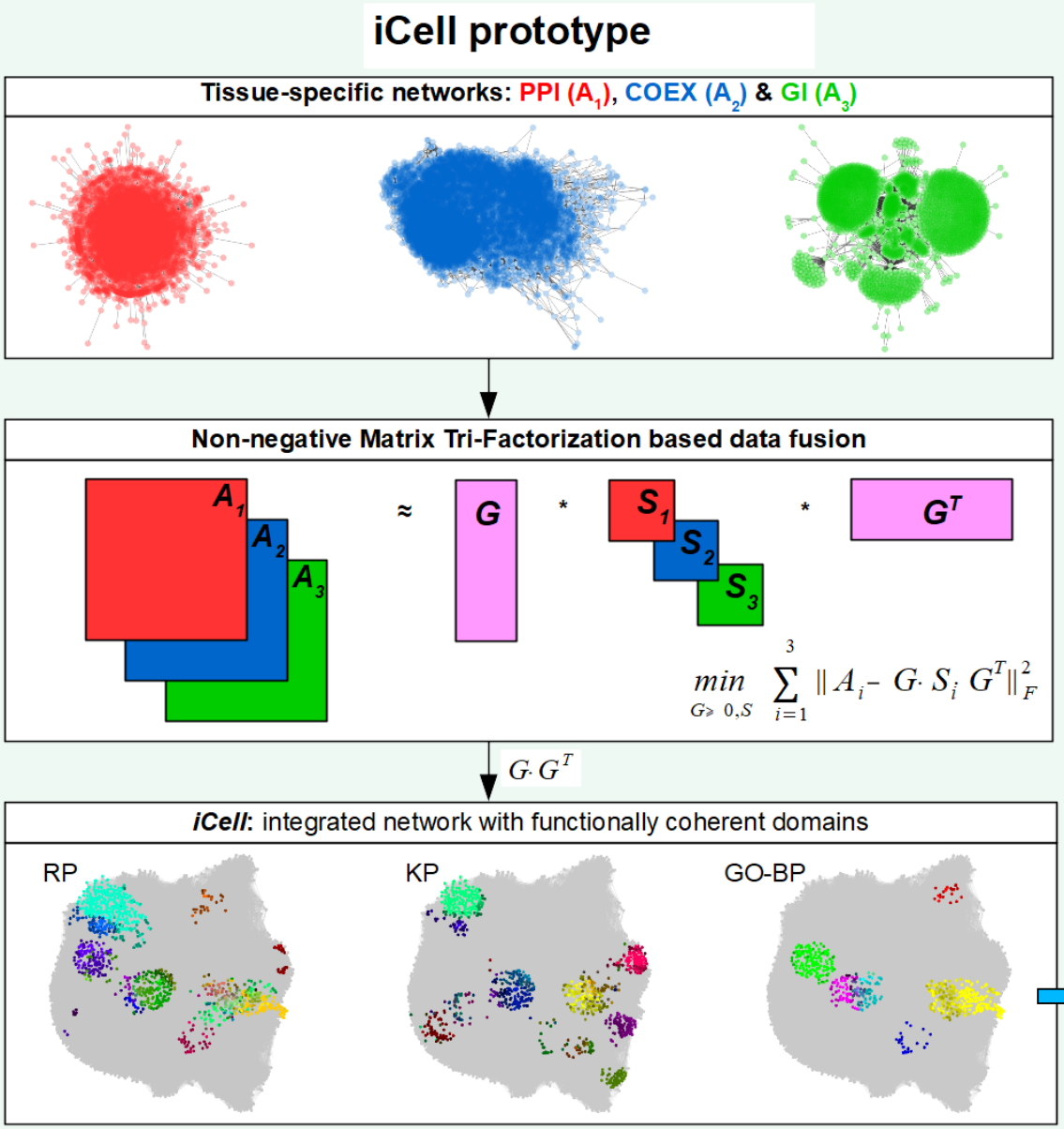
- **additional functional** info
- **emerging** from the **fusion**
- despite ≈ no network overlap

Clusters: k-means equivalent

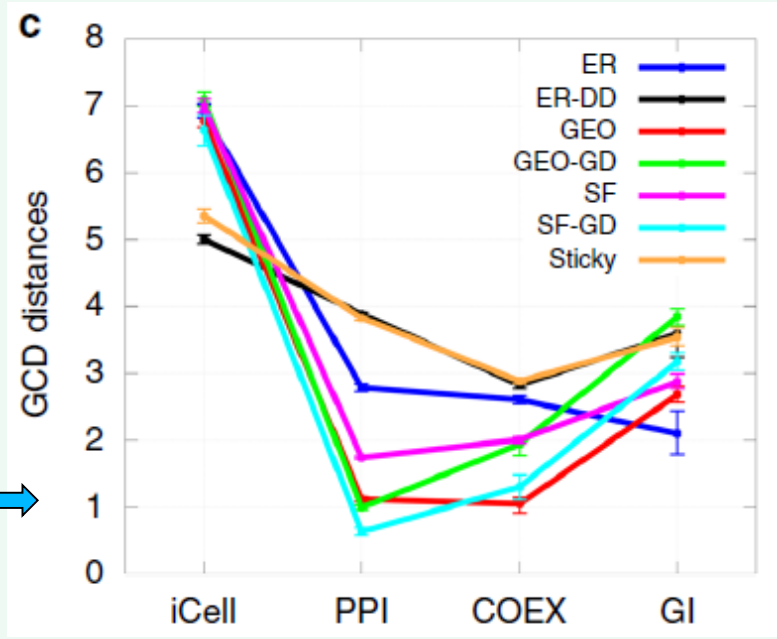



Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

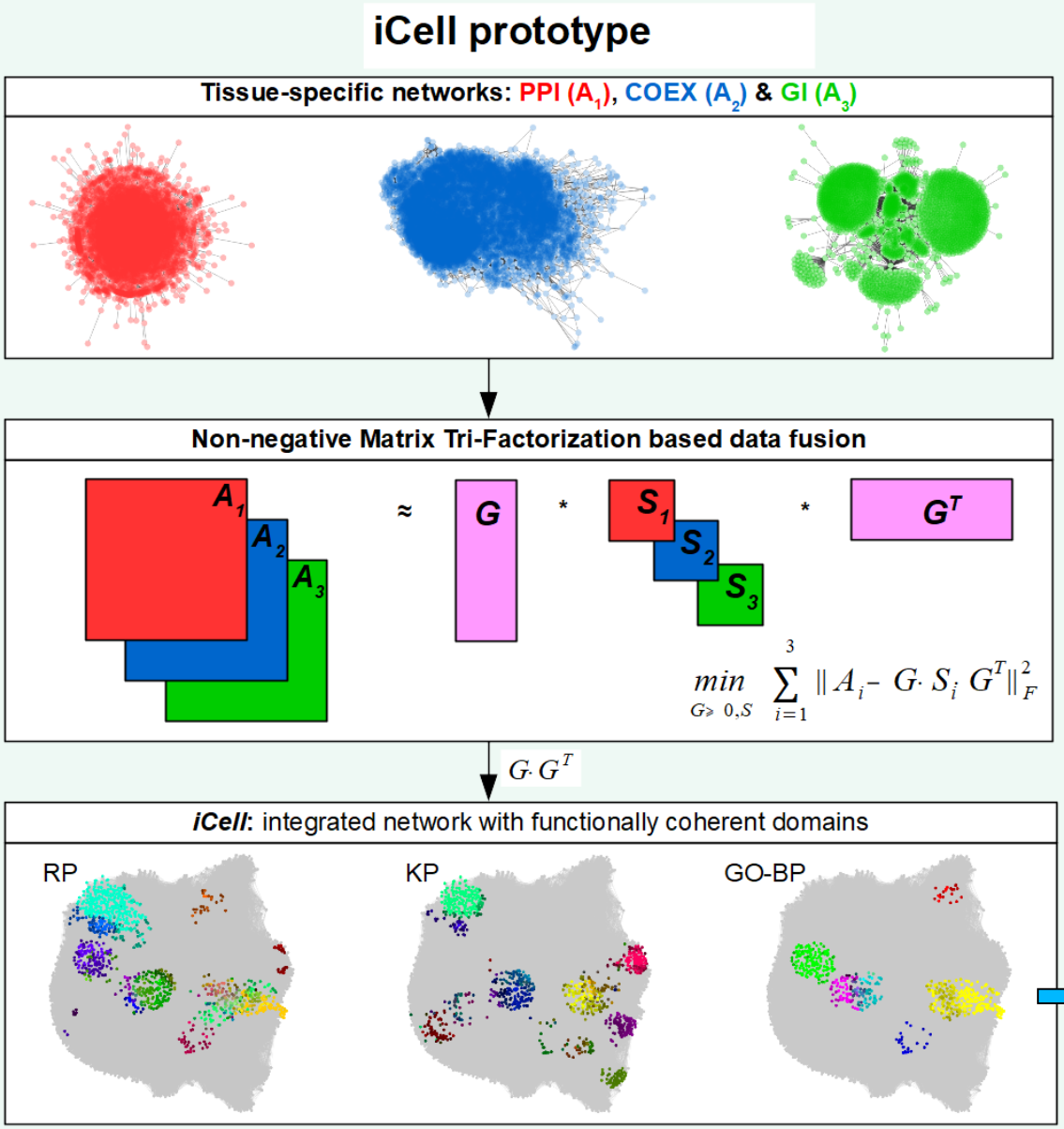


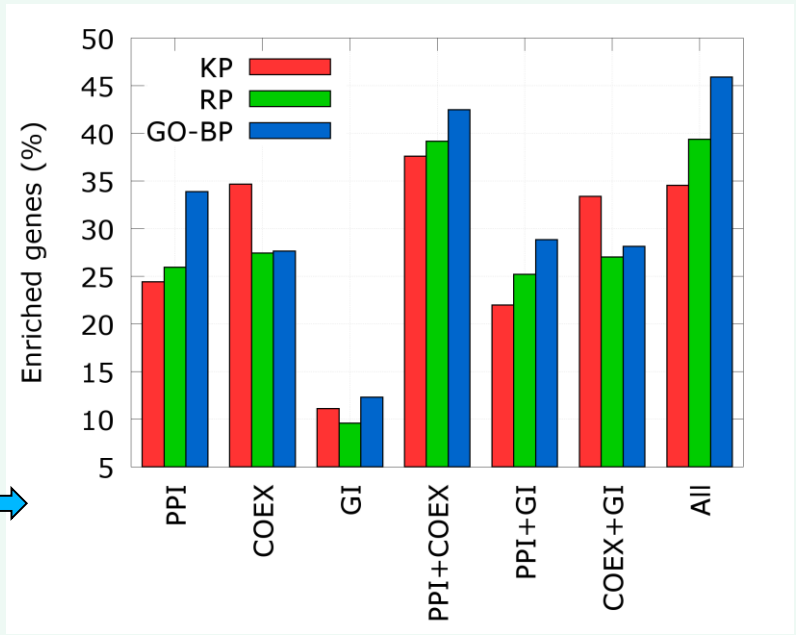

iCells are not random




Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data




Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

→ Create tissue specific networks (PPI, GI, COEX) and their iCells

**Cancers**:
- Breast
- Prostate
- Lung
- Colorectal

**Controls**:
- Breast glandular cells
- Prostate glandular cells
- Lung pneumocytes
- Colorectal glandular cells

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

→ Create tissue specific networks (PPI, GI, COEX) and their iCells

**Cancers**:
- Breast
- Prostate
- Lung
- Colorectal

**Controls**:
- Breast glandular cells
- Prostate glandular cells
- Lung pneumocytes
- Colorectal glandular cells

In Human Protein Atlas: a gene either expressed or not in a tissue

Find between cancer and control tissue, genes:
- **Cancer-silenced**: expressed in control, but not in cancer
- **Cancer-activated**: expressed in cancer, but not in control
- **Always-silenced**: not expressed in either cancer or control
- **Always-expressed**: expressed in both (maybe at different levels)

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

→ Create tissue specific networks (PPI, GI, COEX) and their iCells

**Cancers**:
- Breast
- Prostate
- Lung
- Colorectal

**Controls**:
- Breast glandular cells
- Prostate glandular cells
- Lung pneumocytes
- Colorectal glandular cells

In Human Protein Atlas: a gene either expressed or not in a tissue

Find between cancer and control tissue, genes:
- **Cancer-silenced**: expressed in control, but not in cancer
- **Cancer-activated**: expressed in cancer, but not in control
- **Always-silenced**: not expressed in either cancer or control
- **Always-expressed**: expressed in both (maybe at different levels)
  - ➢ Enriched in drivers

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

→ Create tissue specific networks (PPI, GI, COEX) and their iCells

**Cancers**:
- Breast
- Prostate
- Lung
- Colorectal

**Controls**:
- Breast glandular cells
- Prostate glandular cells
- Lung pneumocytes
- Colorectal glandular cells

**Rewiring** of **always-expressed** genes ↔ involvement in cancer?

- In **cancer** and **control: *iCells***, ***PPI***, ***GI***, ***COEX***
  → Most GDV-rewired
  → Check enrichment in drivers?

→ **Only in iCells – rewiring indicative of oncogenicity**
- Top 500 most-rewired: enriched in drivers and cancer-related pathways

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Take top 20 of most-rewired in ***cancer iCells*** for the four cancers:

- 63 unique genes
- Almost all **validated**: literature, survival, knockdown in cancer changes cell viability

Note: only 17 of 63 (27%) are *differentially expressed* in cancer

**Top 20 most-rewired genes in breast cancer**

| Gene | Literature support | Patient survival curve diff. (p-val) | Cell viability change (p-val) |
|---|---|---|---|
| XKR3 | PMID: 19592507 | 4.57E-01 | 4.04E-02 |
| TOPAZ1 | PMID: 23478628 | | 4.04E-02 |
| HLA-DQA2 | PMID: 27539887 | 4.06E-03 | |
| ECT2L | intOgen | 2.88E-02 | 5.00E-01 |
| CD300LD | | | 4.04E-02 |
| GDF6 | PMID: 17616940 | 1.13E-01 | 4.04E-02 |
| PNMA6A | | 2.14E-02 | 4.04E-02 |
| MAGEB16 | PMID: 11454705 | | 4.04E-02 |
| ERICH6B | PMID: 26828653 | 6.77E-03 | 4.04E-02 |
| NAE1 | PMID: 22874562 | 3.22E-02 | 4.04E-02 |
| NTRK1 | intOgen | 5.89E-03 | 4.04E-02 |
| CCNB1 | PMID: 27903976 | 4.12E-02 | 4.04E-02 |
| MRPL3 | | 1.75E-02 | 4.04E-02 |
| PSMC3 | | 2.01E-02 | 4.04E-02 |
| MRPL50 | | 6.17E-02 | 4.04E-02 |
| CD300LG | | 2.38E-02 | 4.04E-02 |
| C9orf163 | | | 4.04E-02 |
| MRPL4 | | 3.33E-01 | 4.04E-02 |
| COPS5 | intOgen | 1.90E-03 | 5.00E-01 |
| MRPL42 | | 9.32E-02 | 1.91E-01 |

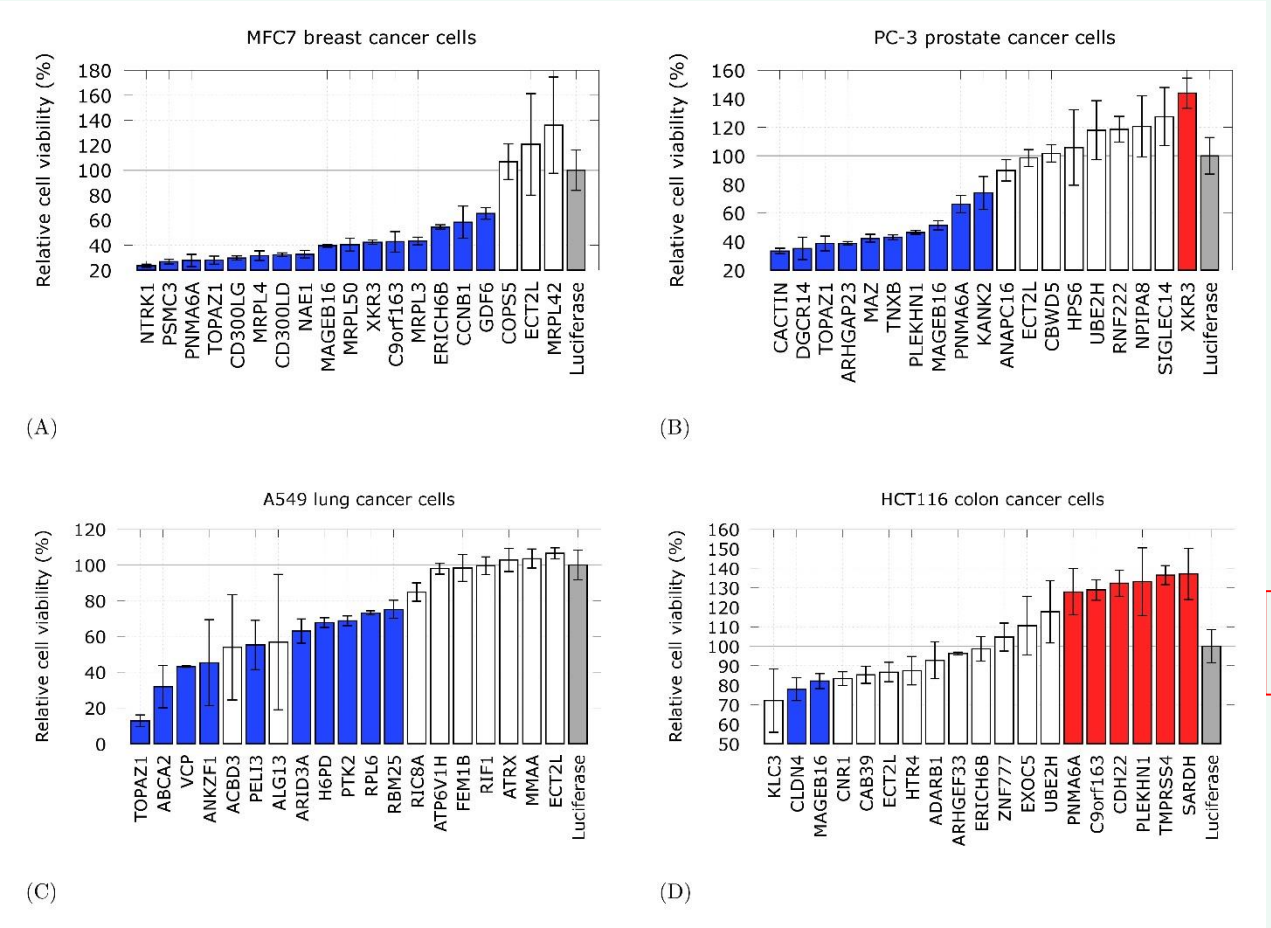
# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Take top 20 of most-rewired in ***cancer iCells*** for the four cancers:

- 63 unique genes
- Almost all **validated**: literature, survival, knockdown in cancer changes cell viability

Note: only 17 of 63 (27%) are *differentially expressed* in cancer

Significantly reduced cell viability in cancer



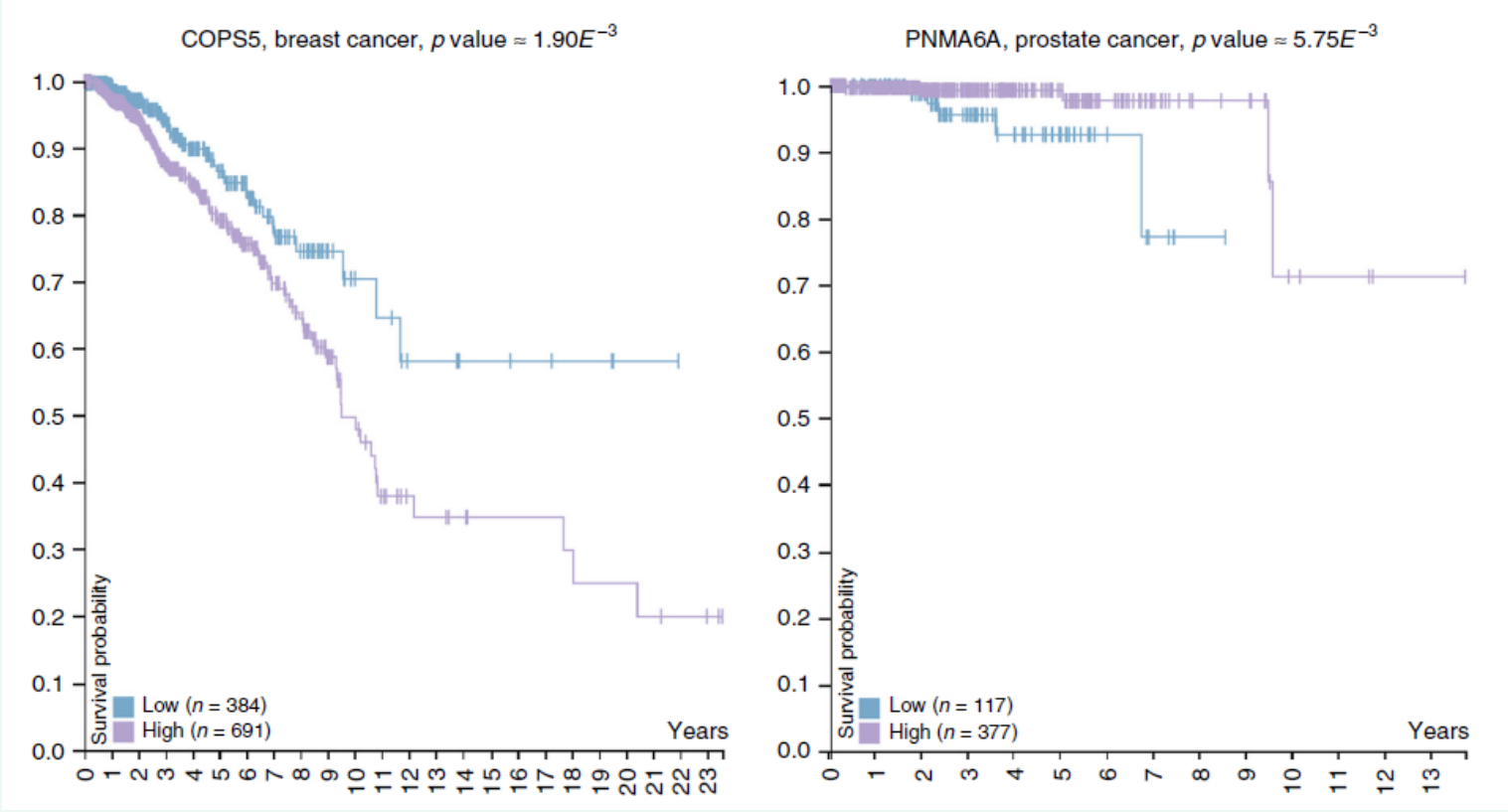

Significantly increased cell viability in cancer

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

Take top 20 of most-rewired in ***cancer iCells*** for the four cancers:

- 63 unique genes
- Almost all **validated**: literature, survival, knockdown in cancer changes cell viability

Note: only 17 of 63 (27%) are *differentially expressed* in cancer




Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

**Pan-cancer**: 16 more cancers → 20 in total

carcinoid, cervical, endometrial, glioma, head and neck, liver, lymphoma, melanoma, ovarian, pancreatic, renal, skin, stomach, testis, thyroid, and urothelial cancer

Make their ***cancer iCells***

Are similarly wired genes in different cancer iCells cancer-related?

- 3,077 genes expressed in all 20 cancer types
  → “pan-cancer expressed”
- Find the most wiring-similar out of the 3,077 genes common to cancer iCells
- Top 500: significantly enriched in drivers

# 2. Novel Methods

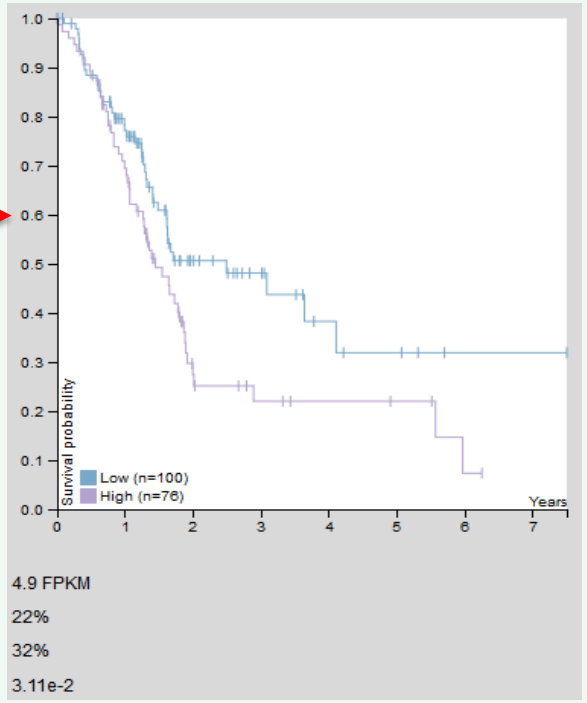
## iCell: Tissue-specific integration of heterogeneous omics data

- **NUDT8**: mitochondrial Nudix Hydrolase
  - ➢ Different survivals of patients of 8 cancer types:

| Rank | Gene | Evidence |
|---|---|---|
| 1 | NUDT8 | |
| 2 | HLA-DQA2 | PMID: 27539887 |
| 3 | ECT2L | intOgen |
| 4 | CUL5 | PMID: 24760825 |
| 5 | ENO1 | PMID: 26734996 |
| 6 | CCDC8 | PMID: 26052355 |
| 7 | CUL2 | PMID: 20078552 |
| 8 | VCP | PMID: 18798739 |
| 9 | TARDBP | PMID: 22146597 |
| 10 | NPM1 | PMID: 26894557 |
| 11 | SHMT2 | PMID: 27666119 |
| 12 | HNRNPU | PMID: 20010808 |
| 13 | FUS | PMID: 21169411 |
| 14 | SRRM2 | PMID: 26135620 |
| 15 | COPS5 (CSN5) | intOgen |
| 16 | DHX9 | PMID: 26973242 |
| 17 | GRB2 | PMID: 25031732 |
| 18 | ILF3 | PMID: 22842455 |
| 19 | OTUB1 | PMID: 25431208 |
| 20 | EEF1A1 (CCS-3) | PMID: 16828757 |

**From Human Protein Atlas**, significant patient stratifications for 8 cancer types:

- ❖ **Lung** (4.12E-2),
- ❖ **Liver** (2.69E-2),
- ❖ **Pancreatic** (3.11E-2)
- ❖ **Head and neck** (2.36E-3),
- ❖ **Stomach** (4.70E-2),
- ❖ **Renal** (3.70E-4),
- ❖ **Cervical** (7.26E-3)
- ❖ **Ovarian** (2.69E-2)



# 2. Novel Methods
## iCell: Tissue-specific integration of heterogeneous omics data

## iCell Conclusions:

- **iCell concept:**
  - ✓ Integrates tissue specific heterogeneous molecular networks
- **New data integration and analytics framework**
- **Cancer-specific and pan-cancer studies**
  - ✓ ***New cancer-related genes***
  - ✓ ***Differential expression limited – need to go beyond***
- **Generic:**
  - ✓ Can accommodate single-cell omics data
  - ✓ Study structure, heterogeneity & dynamics of tumour function & progression
  - ✓ Can include additional omics data of interest, e.g. epigenomic
- **Enables integrative omics analyses of all cells**
- **Other dieses, cell differentiation and specialization, ageing …**

Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications*** 10:805, 2019

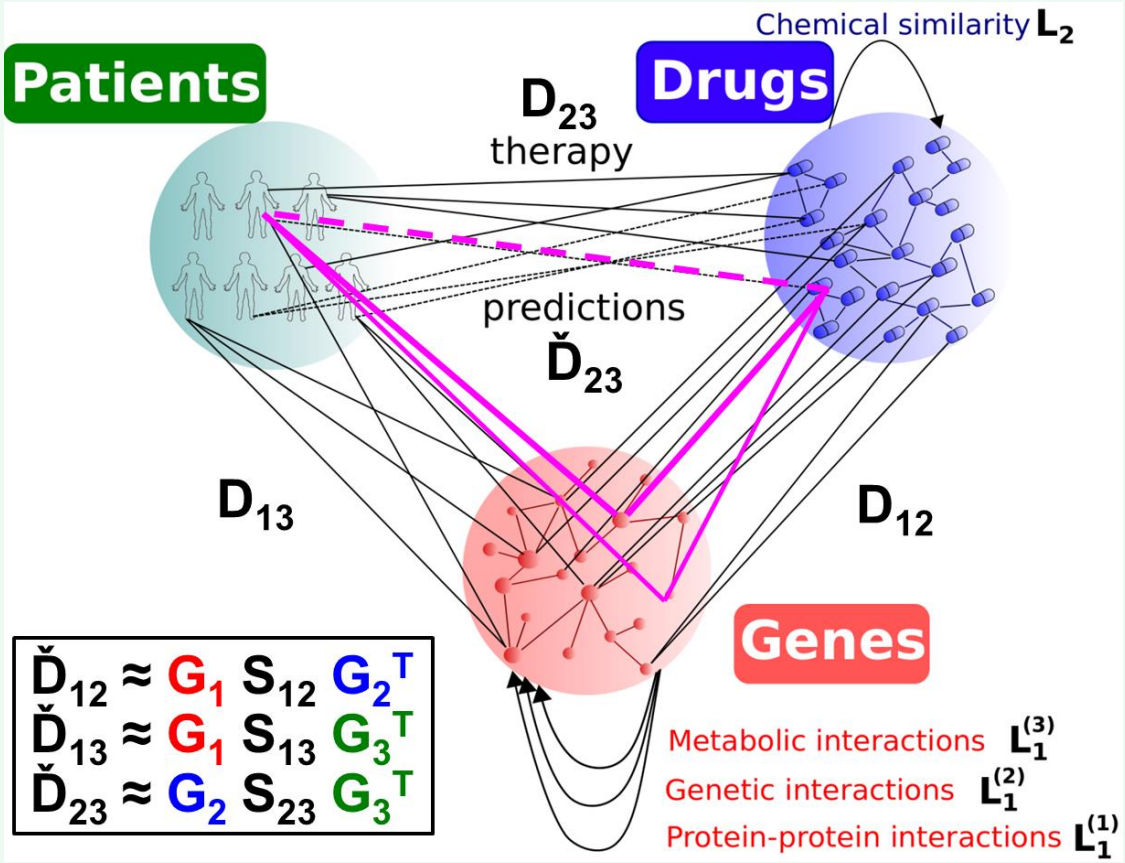
# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Multi-disciplinary, data-fusion methodology

**Motivation:**

- Captures all systems-level
- Captures how data relate
- Mechanistic explanations

**PNMTF → Personalize Treatment**




$$\min\{\sum_{1\le i\le j\le p} [\, ||W_{ij} \circ (D_{ij} - G_i S_{ij} G_j^T)||^2 + \alpha\, ||S_{ij}||^2 + \alpha_i\, tr(G_i^T L_i G_i) + \alpha_j\, tr(G_j^T L_j G_j)] : G_i, S_{ij} \ge 0\}$$

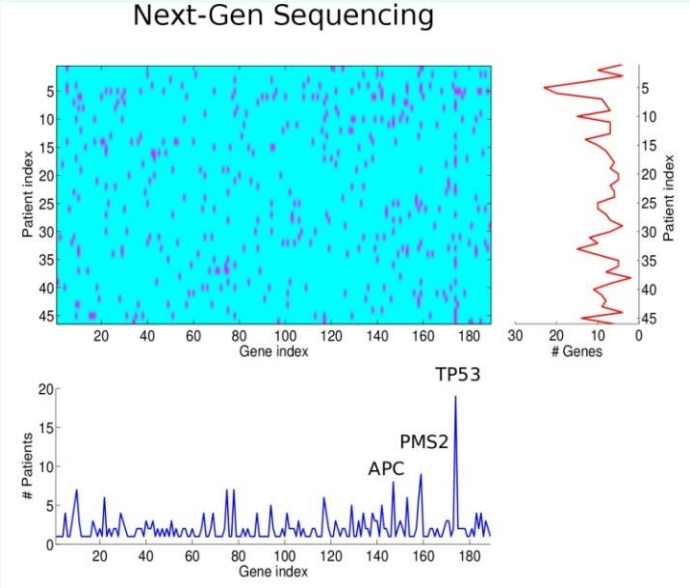
$\alpha\, ||S_{ij}||^2$ maintain sparsity of $S_{ij}$ , $\alpha_i\, tr(G_i^T L_i G_i)$ and $\alpha_j\, tr(G_j^T L_j G_j)$ adding prior knowledge (penalties), $G_i$, $S_{ij} \ge 0$ is needed for cluster interpretation

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

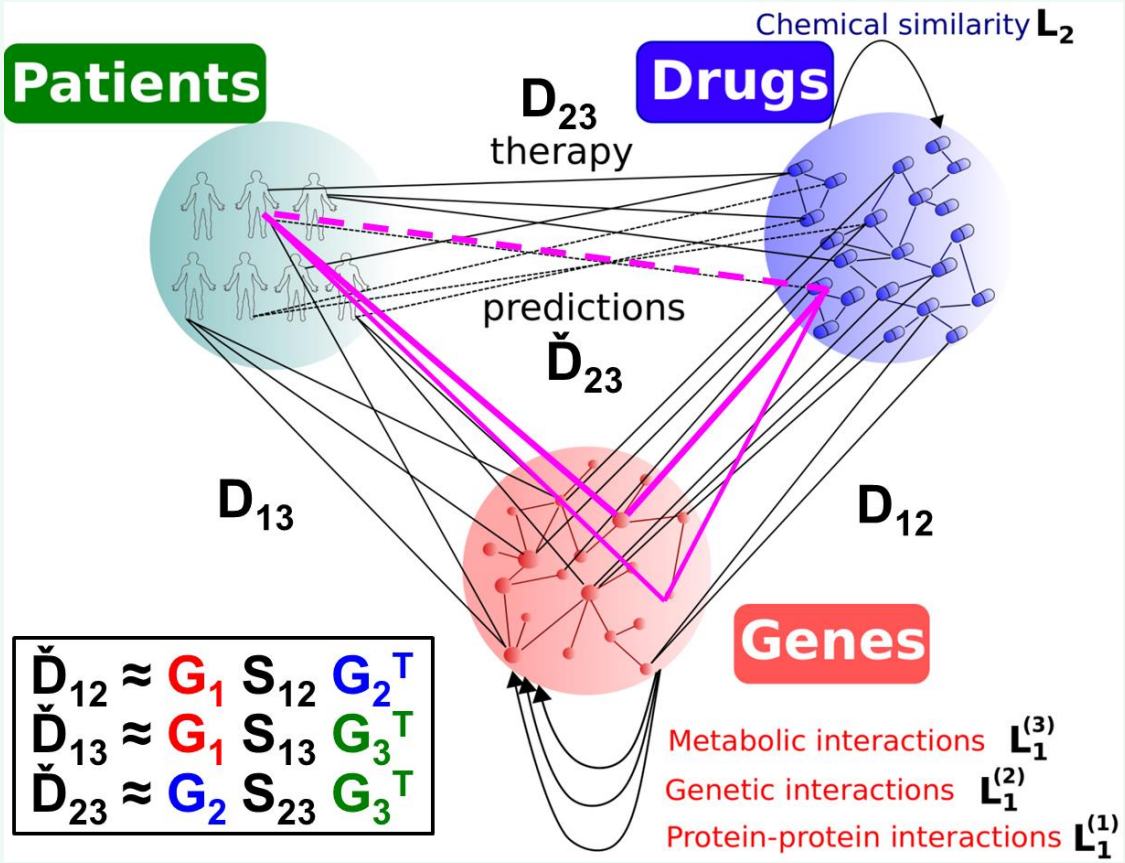
### Multi-disciplinary, data-fusion methodology

**Motivation:**

- Captures all systems-level
- Captures how data relate
- Mechanistic explanations




**PNMTF → Personalize Treatment**




$$\min\{\sum_{1\le i\le j\le p} [\, \|W_{ij} \circ (D_{ij} - G_i S_{ij} G_j^T)\|^2 + \alpha \|S_{ij}\|^2 + \alpha_i \, tr(G_i^T L_i G_i) + \alpha_j \, tr(G_j^T L_j G_j)] : G_i, S_{ij} \ge 0\}$$

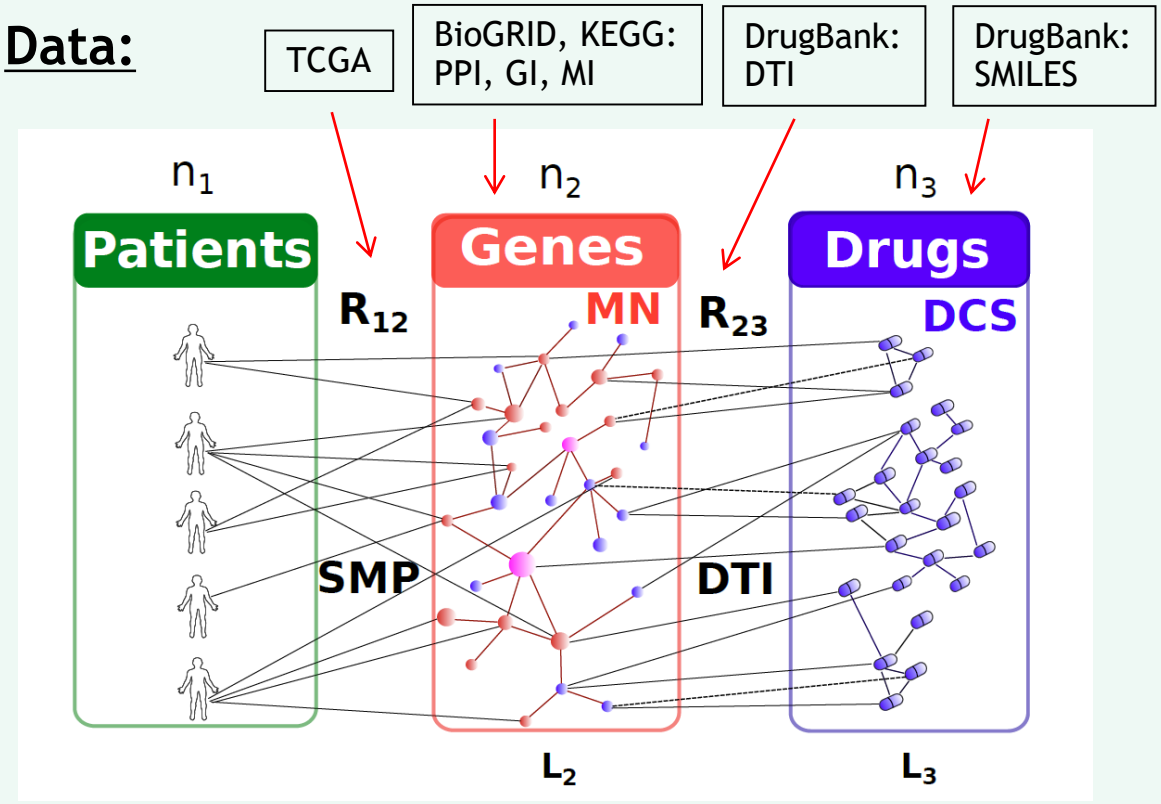
$\alpha \|S_{ij}\|^2$ maintain sparsity of $S_{ij}$ , $\alpha_i \, tr(G_i^T L_i G_i)$ and $\alpha_j \, tr(G_j^T L_j G_j)$ adding prior knowledge (penalties), $G_i$, $S_{ij} \ge 0$ is needed for cluster interpretation

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

## Patient-Specific Data Fusion → Personalized Treatment

**Co-clustering: patients, genes and drugs**

**Data:**




**353 serous ovarian cancer patients from TCGA:**

1. Patient stratification
2. Driver gene prediction
3. Drug repurposing

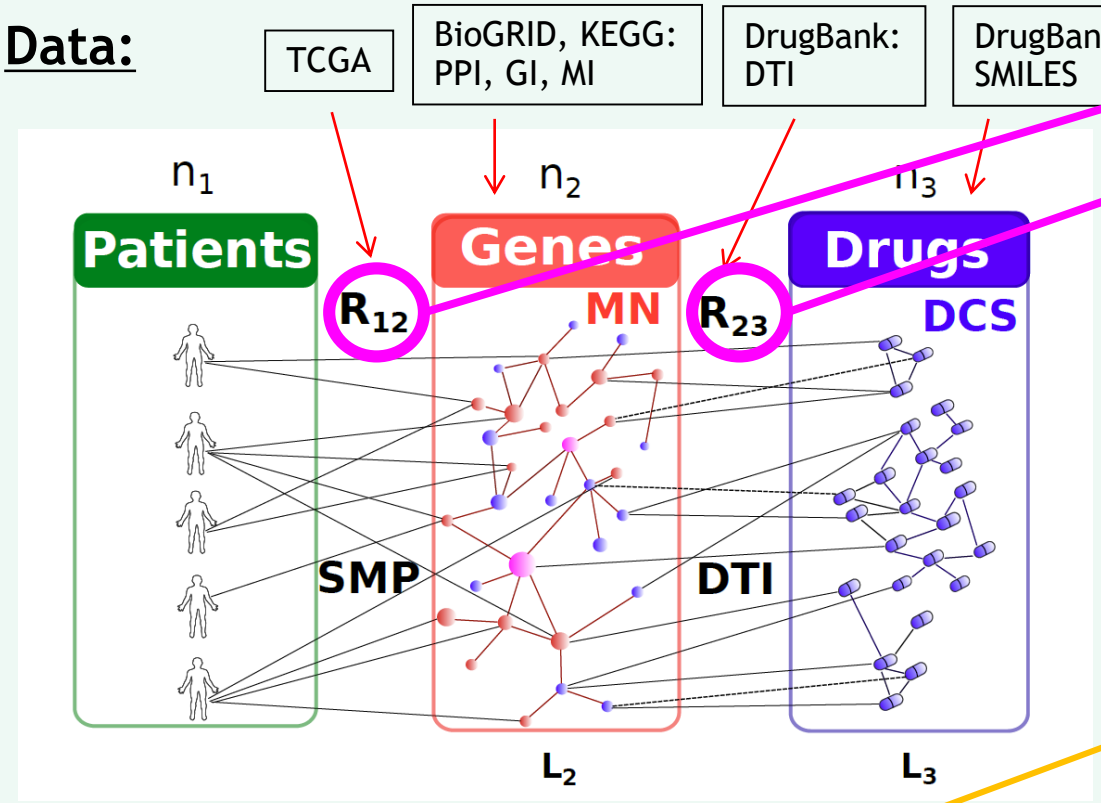
V. Gligorijevic, N. Malod-Dognin and **N. Przulj**, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Patient-Specific Data Fusion → Personalized Treatment

**Co-clustering: patients, genes and drugs**

**Data:**




$$R_{12} \approx G_1 H_{12} G_2^T$$
$$R_{23} \approx G_2 H_{23} G_3^T$$

$k_1 \ll n_1$ – patient clusters
$k_2 \ll n_2$ – gene clusters
$k_3 \ll n_3$ – drug clusters
$G_1$, $G_2$ and $G_3$ are cluster indicator matrices

**Ovarian cancer patients:**

1. Patient stratification → $\hat{C}_1$
2. Driver gene prediction → $\hat{C}_2$
3. Drug repurposing → $\hat{R}_{23}$

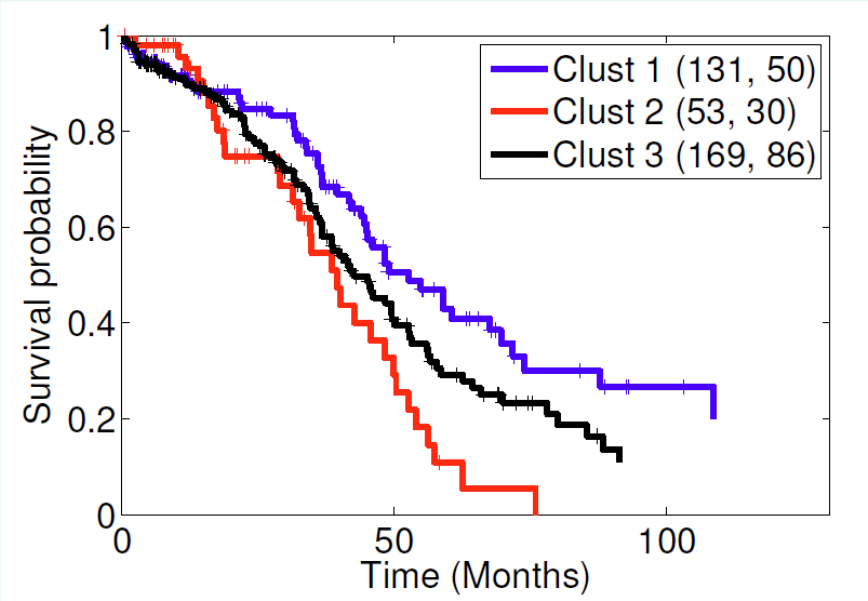
$$\min_{G_i \geq 0, 1 \leq i \leq 3} J = \min_{G_i \geq 0, 1 \leq i \leq 3} \left[ \| R_{12} - G_1 H_{12} G_2^T \|_F^2 + \| R_{23} - G_2 H_{23} G_3^T \|_F^2 + tr(G_2^T L_2 G_2) + tr(G_3^T L_3 G_3) \right]$$

V. Gligorijevic, N. Malod-Dognin and **N. Przulj**, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

## Patient-Specific Data Fusion → Personalized Treatment

**Some results:**




Kaplan-Meier survival curves for 3 patient groups found by GNMTF (log-rank p-val = $5.3 \times 10^{-3}$)

$\mathbf{R}_{12} \approx \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T$
$\mathbf{R}_{23} \approx \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T$

$k_1 \ll n_1$ - patient clusters
$k_2 \ll n_2$ - gene clusters
$k_3 \ll n_3$ - drug clusters
$\mathbf{G}_1$, $\mathbf{G}_2$ and $\mathbf{G}_3$ are cluster indicator matrices

Ovarian cancer patients:

1. Patient stratification → $\hat{\mathbf{C}}_1$
2. Driver gene prediction → $\hat{\mathbf{C}}_2$
3. Drug repurposing → $\hat{\mathbf{R}}_{23}$

$$\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \Big[ \| \mathbf{R}_{12} - \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \|_F^2 + \| \mathbf{R}_{23} - \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \|_F^2 + tr(\mathbf{G}_2^T \mathbf{L}_2 \mathbf{G}_2) + tr(\mathbf{G}_3^T \mathbf{L}_3 \mathbf{G}_3) \Big]$$

V. Gligorijevic, N. Malod-Dognin and **N. Przulj**, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

## Patient-Specific Data Fusion → Personalized Treatment

**Some results:** ~40% of our 809 predicted driver genes in CCGD, Census, or IntOGen

| New driver | Known drivers | Score | DB |
|---|---|---|---|
| ADAM32 | BMPR2 | 1.000 | – |
| REG1P | CLASP2 | 1.000 | – |
| PCDHA2 | CHD4 | 1.000 | – |
| NCR1 | BMPR2 | 1.000 | – |
| USPL1 | CLASP2 | 1.000 | – |
| GDPD3 | DDX5 | 1.000 | – |
| LECT1 | CLASP2 | 1.000 | CCGD |
| IL25 | CDK12, CCAR1 | 0.975 | – |
| BAK1 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| MOGAT2 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| CHAF1A | ATRX, TFDP1, NDRG1 | 0.967 | CCGD |
| PITX2 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| SIN3B | ATRX, TFDP1, NDRG1 | 0.967 | – |
| RPL30 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| GRWD1 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| SNAI1 | ATRX, TFDP1, NDRG1 | 0.967 | CCGD |
| RBMXP4 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| CPNE7 | ATRX, TFDP1, NDRG1 | 0.967 | – |
| HIPK3 | ATRX, TFDP1, NDRG1 | 0.967 | CCGD |
| EPOR | ATRX, TFDP1, NDRG1 | 0.967 | CCGD |

↔ TGFs, cell proliferation & progression
↔ proliferation, migration, anti-apoptosis; prognosis markers

Ovarian cancer patients:

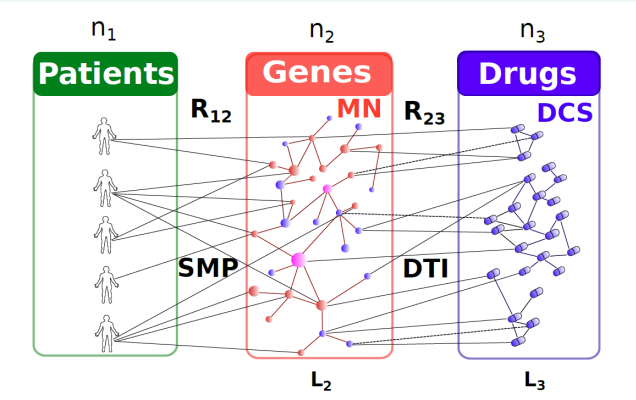
1. Patient stratification → $\hat{\mathbf{C}}_1$
2. **Driver gene prediction** → $\hat{\mathbf{C}}_2$
3. Drug repurposing → $\hat{\mathbf{R}}_{23}$

$$\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \Big[ \| \mathbf{R}_{12} - \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \|_F^2 + \| \mathbf{R}_{23} - \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \|_F^2 + tr(\mathbf{G}_2^T \mathbf{L}_2 \mathbf{G}_2) + tr(\mathbf{G}_3^T \mathbf{L}_3 \mathbf{G}_3) \Big]$$

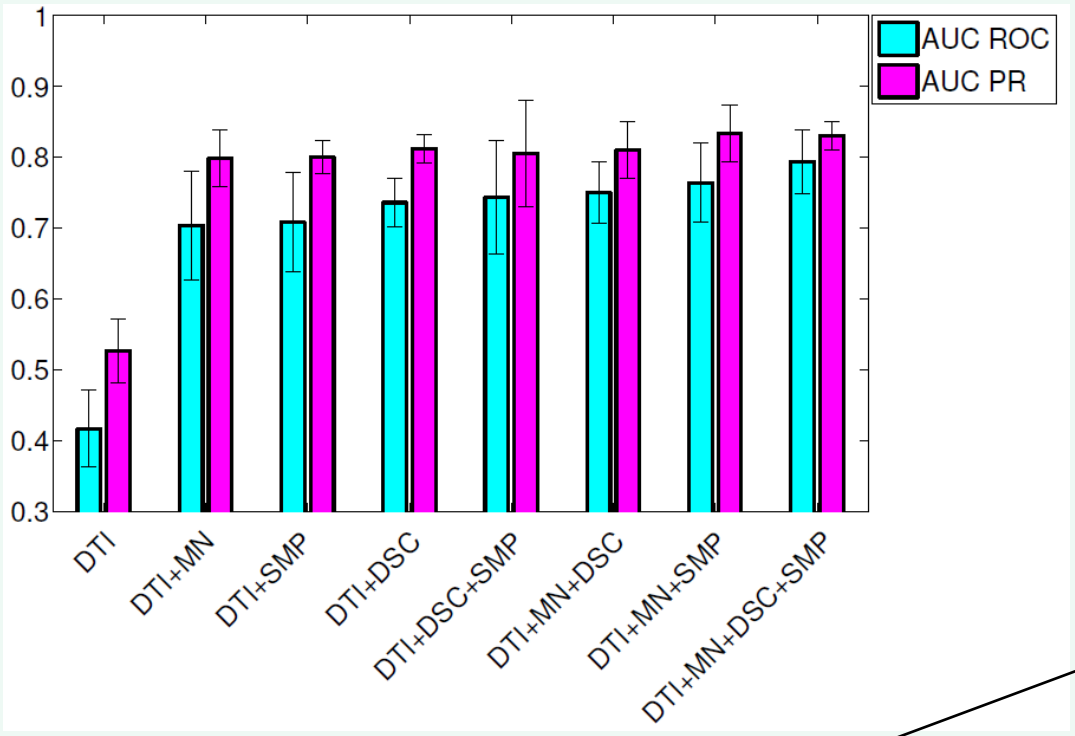
V. Gligorijevic, N. Malod-Dognin and **N. Przulj**, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

## Patient-Specific Data Fusion → Personalized Treatment




**Some results:** 5-fold cross validation, average AUC: ROC and PR




Ovarian cancer patients:

1. Patient stratification → $\hat{\mathbf{C}}_1$
2. Driver gene prediction → $\hat{\mathbf{C}}_2$
3. Drug repurposing → $\hat{\mathbf{R}}_{23}$

$$\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \Big[ \| \mathbf{R}_{12} - \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \|_F^2 + \| \mathbf{R}_{23} - \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \|_F^2 + tr(\mathbf{G}_2^T \mathbf{L}_2 \mathbf{G}_2) + tr(\mathbf{G}_3^T \mathbf{L}_3 \mathbf{G}_3) \Big]$$

V. Gligorijevic, N. Malod-Dognin and **N. Przulj**, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

## Patient-Specific Data Fusion → Personalized Treatment

**Some results:** 37% of our ~225K predicted DTIs confirmed in MATADOR or CTD

| Gene | Drug | Score | Clusters | DB |
|---|---|---|---|---|
| KIT | ATP | 0.873 | 1, 2, 3 | – |
| GABRQ | Adinazolam | 0.808 | 1 | M |
| GABRQ | Fludiazepam | 0.808 | 1 | M |
| GABRQ | Cinolazepam | 0.809 | 1 | M |
| GABRQ | Clotiazepam | 0.809 | 1 | M |
| HTR2A | Dopamine | 0.809 | 1, 3 | C, M |
| GRIN3A | Pethidine | 0.801 | 1, 2 | – |
| CACNA2D1 | Verapamil | 0.761 | 1, 3 | M |
| PDGFRB | ATP | 0.724 | 1, 2 | – |
| KDR | ATP | 0.724 | 1, 3 | C |
| HTR1A | Mirtazapine | 0.720 | 1, 2 | C ,M |
| GABRA6 | Adinazolam | 0.688 | 1 | M |
| GABRA6 | Fludiazepam | 0.688 | 1 | M |
| GABRA6 | Cinolazepam | 0.688 | 1 | M |
| GABRA6 | Clotiazepam | 0.688 | 1 | M |
| GABRA4 | Adinazolam | 0.687 | 1, 3 | M |
| GABRA4 | Fludiazepam | 0.687 | 1, 3 | M |
| GABRA4 | Cinolazepam | 0.687 | 1, 3 | M |
| GABRA4 | Clotiazepam | 0.687 | 1, 3 | M |
| CACNA1D | Magnesium Sulfate | 0.676 | 1, 2, 3 | M |

Ovarian cancer patients:

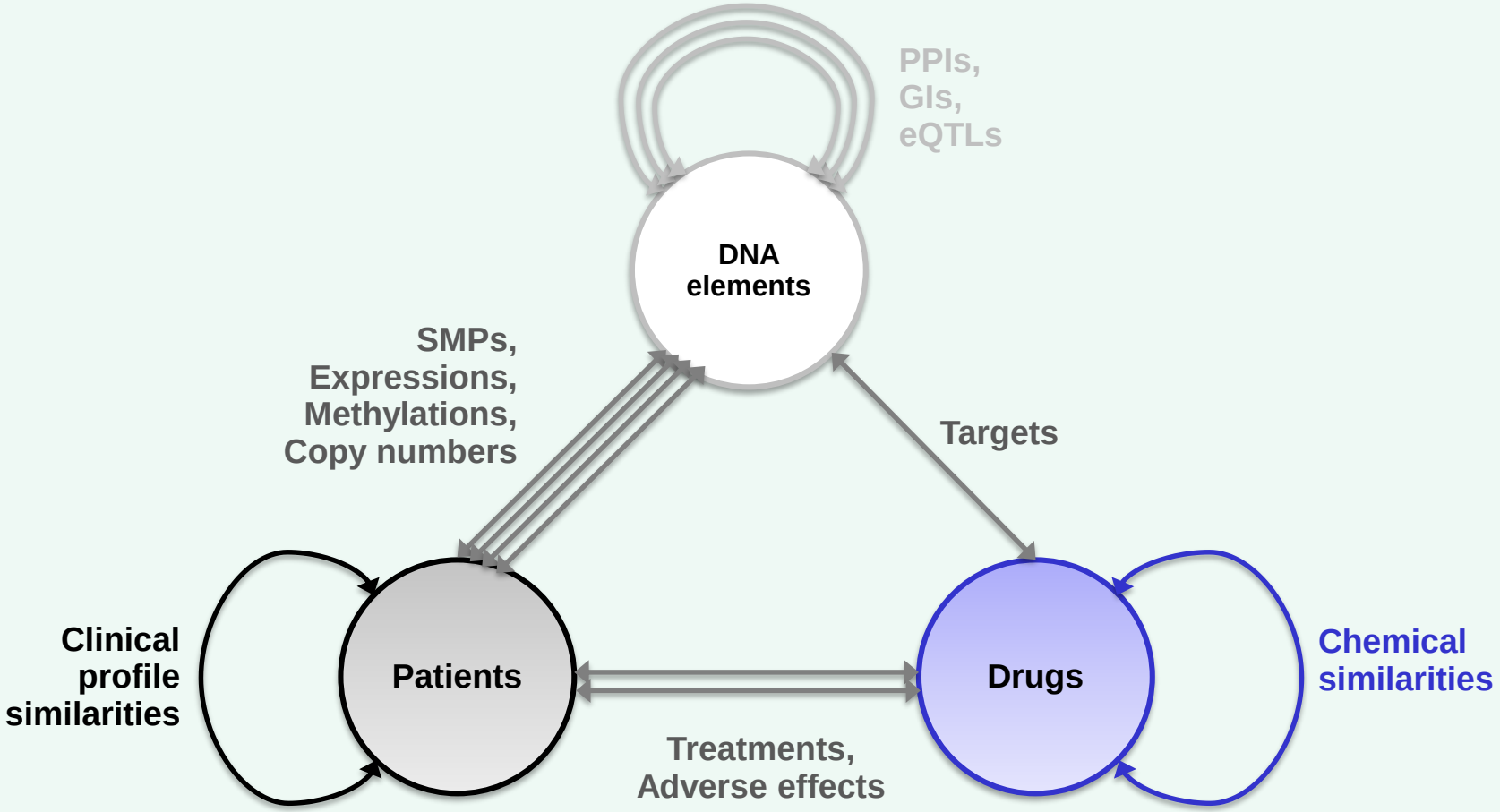
1. Patient stratification → $\hat{\mathbf{C}}_1$
2. Driver gene prediction → $\hat{\mathbf{C}}_2$
3. **Drug repurposing** → $\hat{\mathbf{R}}_{23}$

$$\min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} J = \min_{\mathbf{G}_i \geq 0, 1 \leq i \leq 3} \Big[ \| \mathbf{R}_{12} - \mathbf{G}_1 \mathbf{H}_{12} \mathbf{G}_2^T \|_F^2 + \| \mathbf{R}_{23} - \mathbf{G}_2 \mathbf{H}_{23} \mathbf{G}_3^T \|_F^2 + tr(\mathbf{G}_2^T \mathbf{L}_2 \mathbf{G}_2) + tr(\mathbf{G}_3^T \mathbf{L}_3 \mathbf{G}_3) \Big]$$

V. Gligorijevic, N. Malod-Dognin and **N. Przulj**, Patient-specific data fusion for cancer stratification and personalized treatment, *PSB*, 2016

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Patient-Specific Data Fusion → Personalized Treatment



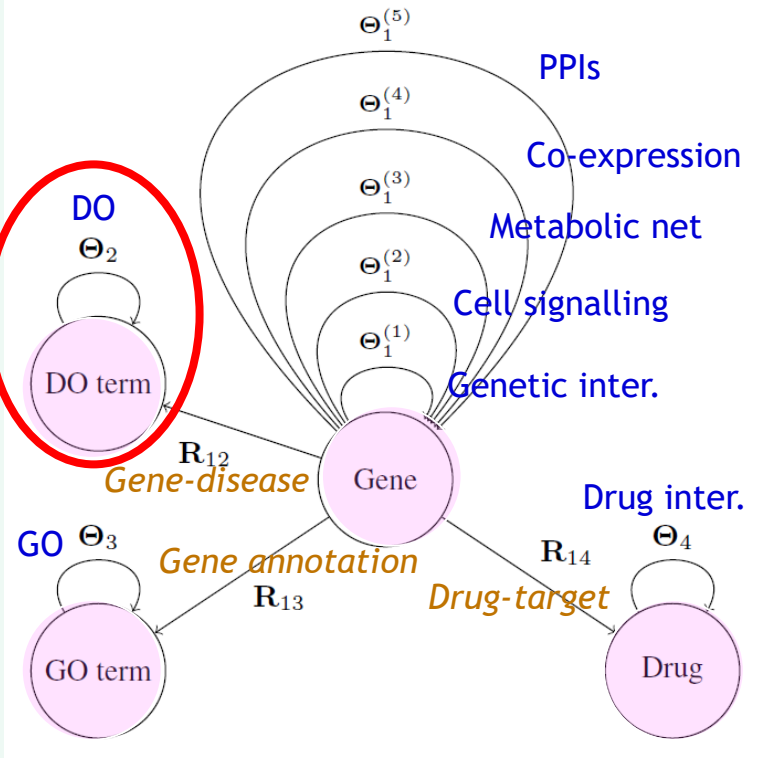

**Obstacles:**

1. **Different NP-hard continuous optimization problem:**
   - propose objective function,
   - optimization solver — prove convergence and correctness
2. **Optimization is slow → HPC**

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Disease Classification from Systems-Level Molecular Data

Method




4 Objects: Genes, GO terms, DO terms, Drugs
Constraints: $\Theta_i$ *(network topology, ontology relations)*
Relation matrices: $R_{ij}$

Some Results:

→ 14 disease-disease associations currently not present in DO:

- evidence for their relationships through *comorbidity* data and *literature curation*

→ GI the most important predictor of a link between diseases, despite small

→ Omission of any one of the included data sources reduces prediction quality

- Importance of systems-level data fusion

→ **DO ∩ disease class → 80% DO from *only* network data**

M. Zitnik, V. Janjic, C. Larminie , B. Zupan, and **N. Przulj,** Discovering disease-disease associations by fusing systems-level molecular data, *Scientific Reports - Nature*, 3:3202, 2013

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Disease Classification from Systems-Level Molecular Data

- Co-clustering GO terms, DO terms, Genes and Drugs under pairwise constraints:

$$\text{Input:}\quad \mathbf{R}=\begin{bmatrix} 0 & \mathbf{R_{12}} & \mathbf{R_{13}} & \mathbf{R_{14}} \\ \mathbf{R_{12}^T} & 0 & 0 & 0 \\ \mathbf{R_{13}^T} & 0 & 0 & 0 \\ \mathbf{R_{14}^T} & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{\Theta}=\begin{bmatrix} \mathbf{\Theta_1^{(t)}} & 0 & 0 & 0 \\ 0 & \mathbf{\Theta_2} & 0 & 0 \\ 0 & 0 & \mathbf{\Theta_3} & 0 \\ 0 & 0 & 0 & \mathbf{\Theta_4} \end{bmatrix}$$

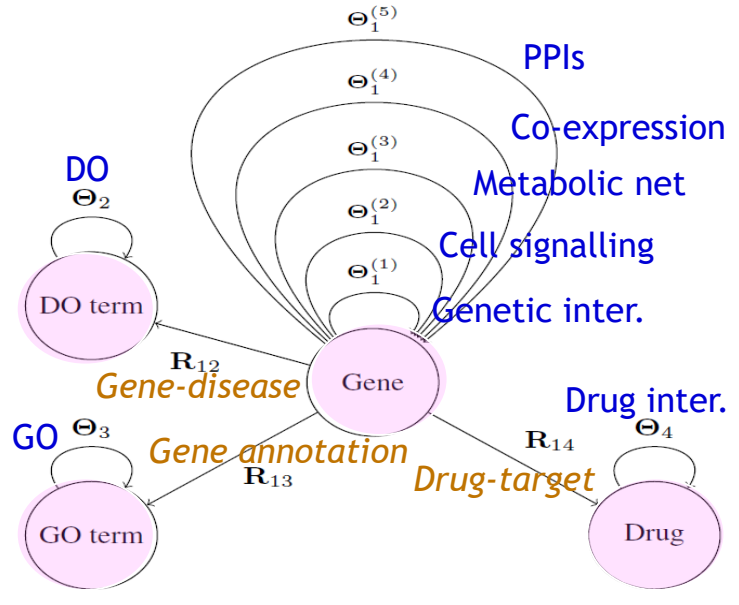
$$\text{Output:}\quad \mathbf{S}=\begin{bmatrix} 0 & \mathbf{S_{12}} & \mathbf{S_{13}} & \mathbf{S_{14}} \\ \mathbf{S_{21}} & 0 & 0 & 0 \\ \mathbf{S_{31}} & 0 & 0 & 0 \\ \mathbf{S_{41}} & 0 & 0 & 0 \end{bmatrix} \quad \mathbf{G}=\begin{bmatrix} \mathbf{G_1} & 0 & 0 & 0 \\ 0 & \mathbf{G_2} & 0 & 0 \\ 0 & 0 & \mathbf{G_3} & 0 \\ 0 & 0 & 0 & \mathbf{G_4} \end{bmatrix}$$

- Minimizing Frobenious distance between $R_{ij}$ and $G_i S_{ij} G_j^T$, for all relation matrices:
  - *i = {Genes}, j = {DO terms, GO terms, Drugs}*
  - $G_i$ is a cluster indicator matrix for data type *i* (genes, DO terms, GO terms and Drugs)

with additional penalty terms:

$$\min_{\mathbf{G}\geq 0} J = \min_{\mathbf{G}\geq 0}\left[\parallel \mathbf{R}-\mathbf{GSG}^T \parallel_F^2 + \left(\sum_{t=1}^{5} tr(\mathbf{G}^T\mathbf{\Theta^{(t)}}\mathbf{G})\right)\right]$$

- Interested in $G_2$ (DO terms)
  - used for cluster assignment and inferring new disease associations from clusters

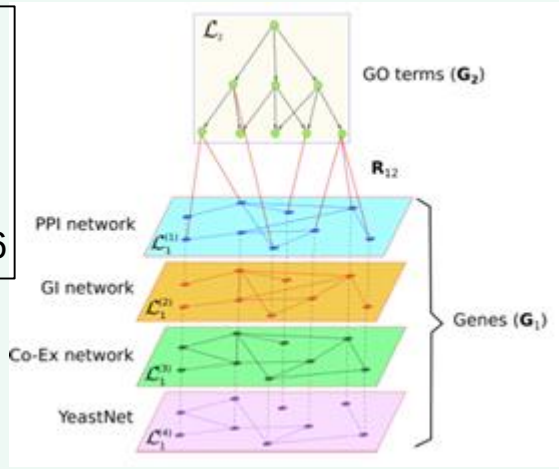



M. Zitnik, V. Janjic, C. Larminie , B. Zupan, and **N. Przulj,** Discovering disease-disease associations by fusing systems-level molecular data, *Scientific Reports - Nature*, 3:3202, 2013

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Gene Ontology from Systems-Level Molecular Data

- Outperform Dutkowski *et al*. [2013]
- 96% of GO reconstructed!
- Correct assignment of GO terms to genes (3-fold cross-validation, AUC=0.874±0.002)
- Graphlets improve results
- **Validated biologically** by Bonne's yeast Genetic Interaction profile data, *Science*, 2016




→ Optimization problem which minimizes $\| R_{12} - G_1 S_{12} G_2^T \|^2_F$
under the guidance of *pairwise constraints*
(***connectivity*** *and* ***GDV similarity***) between genes in networks:

$$\min_{G_1 \geq 0, G_2 \geq 0} J = \min_{G_1 \geq 0, G_2 \geq 0} \left[ \| R_{12} - G_1 S_{12} G_2^T \|_F^2 + \left( \sum_{i=1}^{4} tr(G_1^T L_1^{(i)} G_1) \right) + \left( \sum_{i=1}^{4} tr(G_1^T \Lambda_1^{(i)} G_1) \right) + tr(G_2^T L_2 G_2) \right]$$

→ using topology of molecular networks as constraints (penalty terms) in this optimization problem:

→ $L_1^{(i)}$ is Laplacian of **adjacency matrix** of a molecular network *i=1,2,3,4:*
$L_1^{(i)} = D^i - A^i$, $D^i$ is diagonal matrix of degrees (row summation of $A^i$), $A^i$ is adjacency matrix

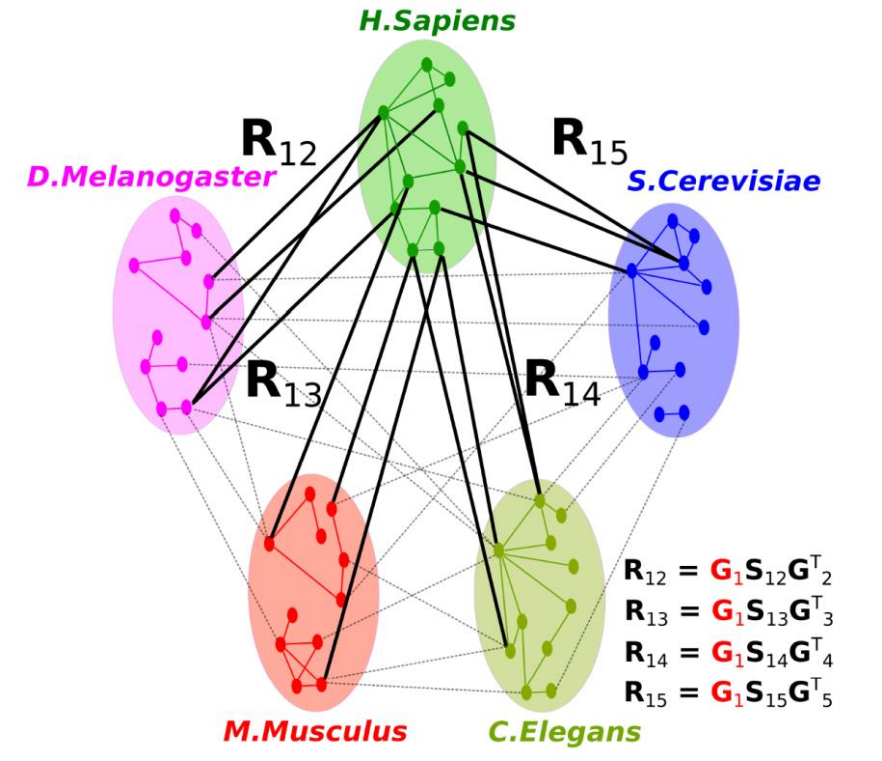
→ $\Lambda_1^{(i)}$ are Laplacians of **GDV similarity matrices** over all genes for each molecular network *i*:
$\Lambda_1^{(i)} = D^i - \sigma^{(i)}$, $D^i$ is diagonal matrix of row summation of $\sigma^{(i)}$, $\sigma^{(i)}$ is binary GDV similarity matrix (containing only significantly similar gene/protein pairs)

→ $L_2$ is Laplacian of **Gene Ontology graph**

V. Gligorijevic, V. Janjic and **N. Przulj**, Integration of molecular network data reconstructs GO, *Bioinformatics, Vol. 30 ECCB 2014, i594-i600 (14% acceptance rate)*, 2014

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Multiple Network Alignment: *Fuse*




**Algorithm 1** Approximate maximum weight $k$-partite matching.

**Input** $G = (\bigcup_{i=1}^{k} V_i, E, W)$
**for** $i = \{2, \ldots, k\}$ **do**
  Find maximum weight bipartite matching $F_{1,i}$ of $G[V_1, V_i]$
  Construct $G_{1i}$, the merge of $V_1$ and $V_i$ from $G$ along $F_{1,i}$
  Set $G = G_{1i}$, and relabel $V_{1i}$ as $V_1$
$C = \{\emptyset\}$
**for** each merged node $u$ in $V_1$ **do**
  Cluster $C_u$ is the set of nodes that are merged into $u$
  Add $C_u$ to $C$
**Output** $C$

We use a block-based representation of relation ($\mathbf{R}$) and Laplacian ($\mathbf{L}$) matrices and matrix factors ($\mathbf{S}$ and $\mathbf{G}$) for our 5 PPI networks as follows:

$$\mathbf{R} = \begin{bmatrix} 0 & \mathbf{R}_{12} & \ldots & \mathbf{R}_{15} \\ \mathbf{R}_{12}^T & 0 & \ldots & \mathbf{R}_{25} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{15}^T & \mathbf{R}_{25}^T & \ldots & 0 \end{bmatrix}, \quad \mathbf{L} = \begin{bmatrix} \mathbf{L}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{L}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{L}_5 \end{bmatrix};$$

$$\mathbf{S} = \begin{bmatrix} 0 & \mathbf{S}_{12} & \ldots & \mathbf{S}_{15} \\ \mathbf{S}_{12}^T & 0 & \ldots & \mathbf{S}_{25} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_{15}^T & \mathbf{S}_{25}^T & \ldots & 0 \end{bmatrix}, \quad \mathbf{G} = \begin{bmatrix} \mathbf{G}_1 & 0 & \ldots & 0 \\ 0 & \mathbf{G}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{G}_5 \end{bmatrix}$$

To simultaneously factorize all relation matrices, $\mathbf{R}_{ij} \approx \mathbf{G}_i \mathbf{S}_{ij} \mathbf{G}_j^T$, $0 \leq i, j \leq 5$, under the constraints of PPI networks, we minimize the following objective function:

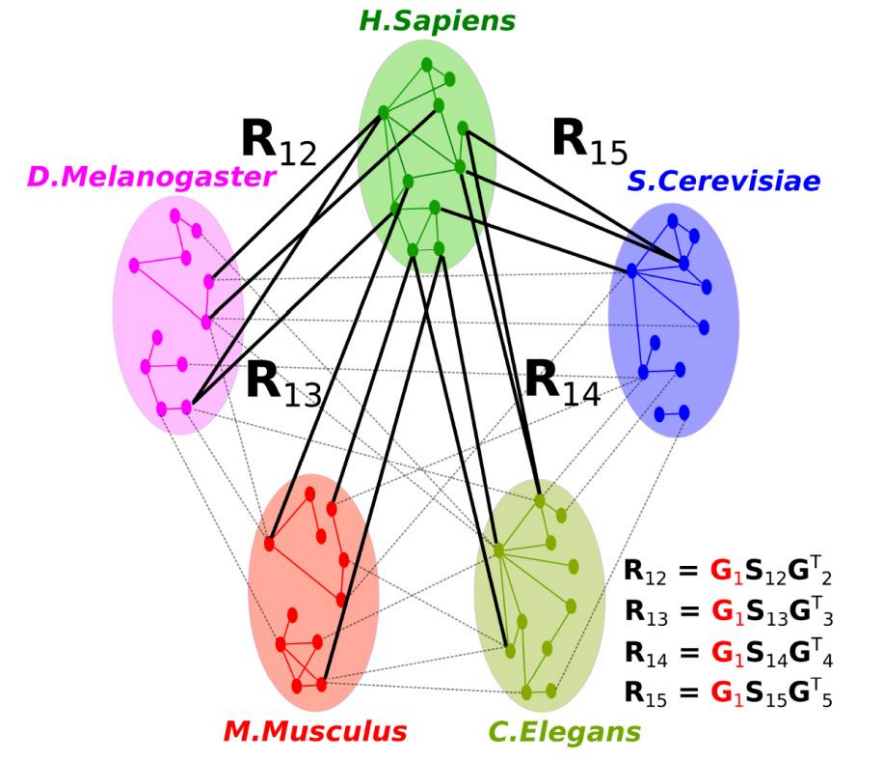
$$\min_{\mathbf{G} \geq 0} J = \left[ \| \mathbf{R} - \mathbf{G}\mathbf{S}\mathbf{G}^T \|_F^2 + \gamma Tr(\mathbf{G}^T \mathbf{L} \mathbf{G}) \right] \quad (2)$$

where $Tr$ denotes the trace of a matrix and $\gamma$ is a regularization parameter which balances the influence of network topologies in reconstruction of the relation matrix. The second term of equation 2 is the penalization term.

V. Gligorijevic, N. Malod-Dognin and **N. Przulj,** Fuse: Multiple network alignment via data fusion, *Bioinformatics*, 32(8):1195-203, 2016. IF=7.3

# 2. Novel Methods
## Mine the Medical World of Inter-Connected Entities

### Multiple Network Alignment: *Fuse*



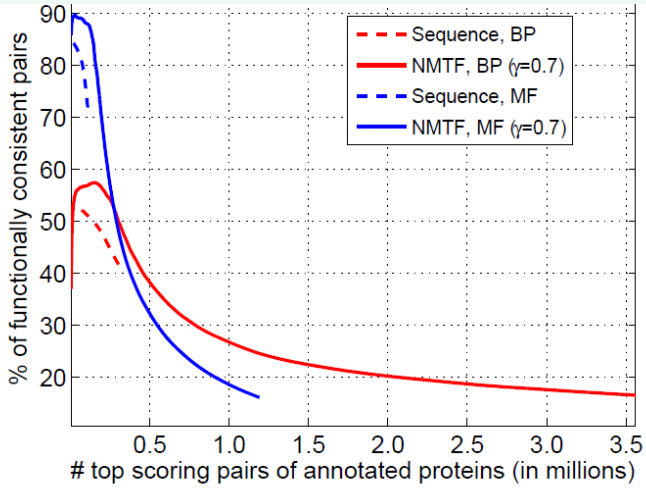




**Fig. 2. Functional consistency of NMTF associations.** For both NMTF associations and sequence similarity of protein pairs, we plot the cumulative number of protein pairs with both proteins annotated ($x$-axis) against the percentages of them sharing GO terms ($y$-axis). Biological process (BP) and molecular function (MF) annotations are considered separately.

V. Gligorijevic, N. Malod-Dognin and **N. Przulj,** Fuse: Multiple network alignment via data fusion, *Bioinformatics*, 32(8):1195-203, 2016. IF=7.3

# Overview

**Medicine: complex world of inter-connected entities**

1. **Motivation**
2. **New Methods – Examples:** mine inter-connected data

   i. **<u>Single type of omics data</u>:**
   - **Molecular networks**
   - **Multi-scale organization**

   } → function, disease

   ii. **<u>Multiple layers of heterogeneous data:</u>**
   - **iCell**
   - **Patient-centered data integration → Precision medicine**
     - ✓ Stratification, biomarker discovery, drug repurposing
   - **Disease re-classification, GO reconstruction, Network alignment, …**
3. **Conclusions**

# 3. Conclusions

**Biomedical Data: complex system of heterogeneous interacting entities**

- Large
- Heterogeneous
- Highly dimensional
- Growing Complexity
- Noisy
- Dynamic
- Different time and space scales

# 3. Conclusions

**<u>Biomedical Data</u>: complex system of heterogeneous interacting entities**

- Large
- Heterogeneous
- Highly dimensional
- Growing Complexity
- Noisy
- Dynamic
- Different time and space scales

- Each type: ***limited***, but ***complementary*** information
- **Seek principled, joint organization and mining within the same framework**

# 3. Conclusions

**<u>Biomedical Data</u>: complex system of heterogeneous interacting entities**

- Large
- Heterogeneous
- Highly dimensional
- Growing Complexity
- Noisy
- Dynamic
- Different time and space scales

- Each type: ***limited***, but ***complementary*** information
- **Seek principled, joint organization and mining within the same framework**

€2M **ERC Consolidator Grant** for 2018-2023
Title: “Integrated Connectedness for a New Representation of Biology”

- Post-Doc positions
- PhD student positions

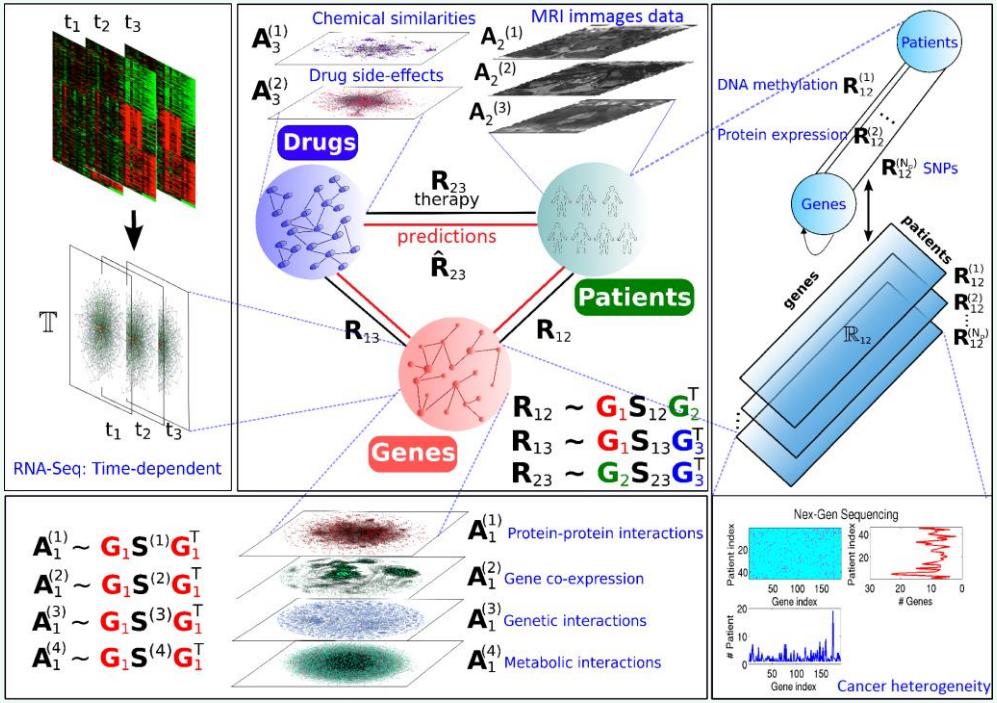
**<u>JnJ</u>**:
- Post-Doc position

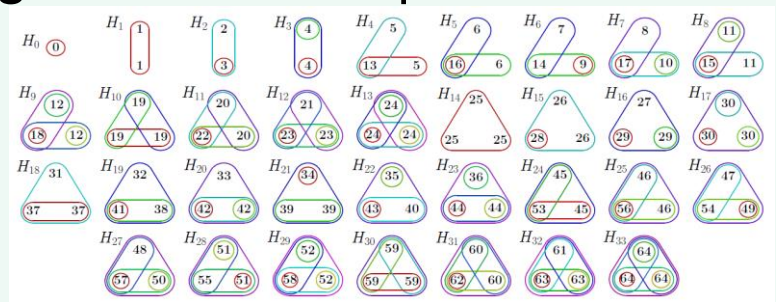
# 3. Conclusions

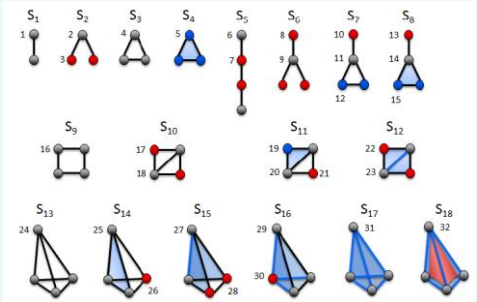
## Holistically Mine All Available Data

### Methodologies

- **Mathematical formalisms**
  - Capture **multi-scale organization**
  - **Dynamics, stochasticity** of the data,…
  - E.g., multiplex networks, hypergraphs, simplicial complexes …
- **Algorithms** to compute and **extract information** from those formalisms






T. Gaudelet, N. Malod-Dognin and **N. Przulj**, "Higher order molecular organisation as a source of biological function," ***Bioinformatics***, ECCB'18

N. Malod-Dognin and **N. Przulj**, "Functional geometry of protein-protein interaction networks," ***Bioinformatics***, 2019

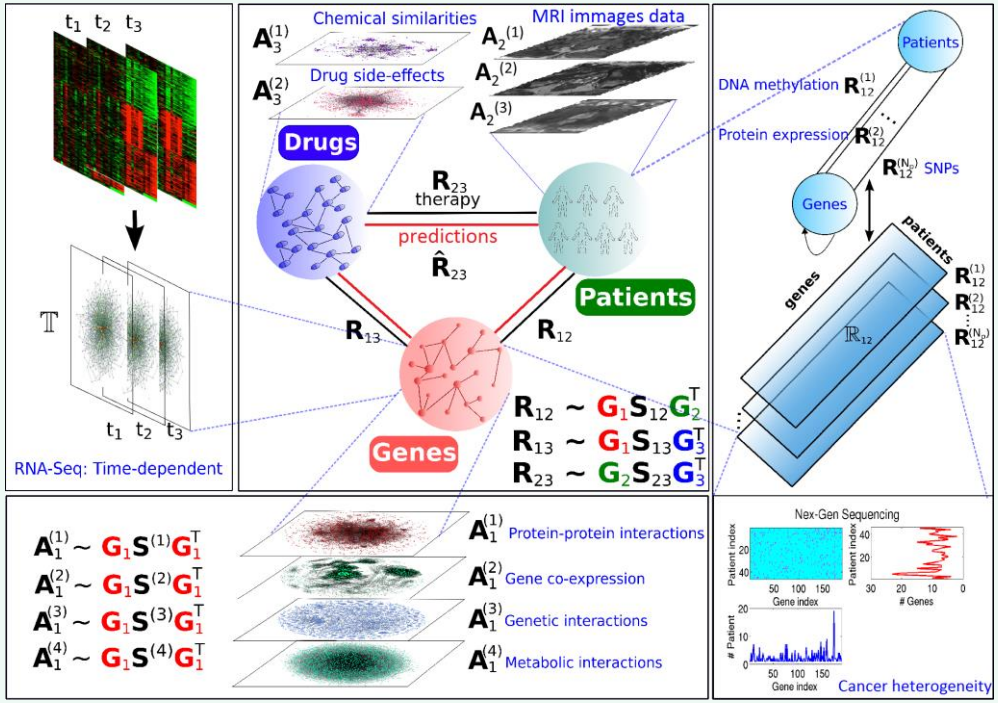
Noël Malod-Dognin, Julia Petschnigg, Sam F. L. Windels, Janez Povh, Harry Hemmingway, Robin Ketteler and **Nataša Pržulj**, "iCell: integrated cells uncover new cancer genes," ***Nature Communications***, 2019

# 3. Conclusions

## Holistically Mine All Available Data

### Methodologies

- **Mathematical formalisms**
  - Capture **multi-scale organization**
  - **Dynamics, stochasticity** of the data,…
  - E.g., multiplex networks, hypergraphs, simplicial complexes …
- **Algorithms** to compute and **extract information** from those formalisms



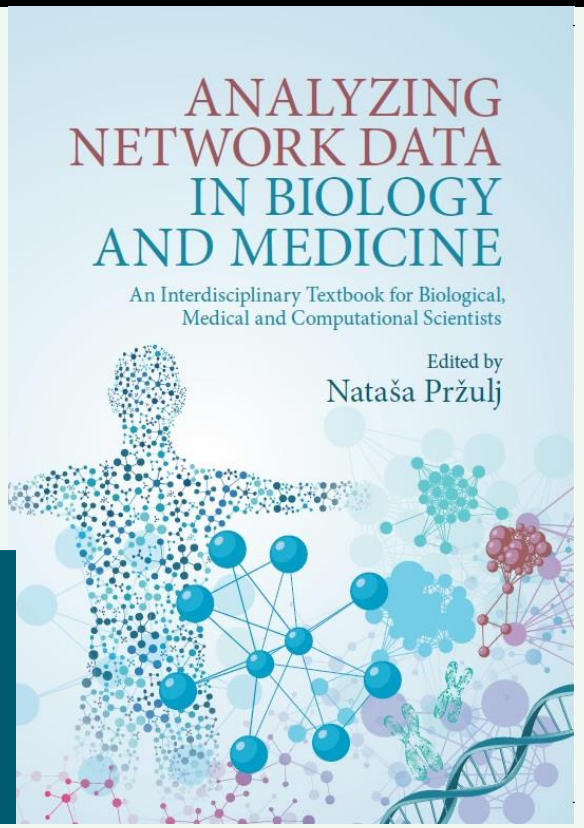

**Computational issues** remain to be addressed, arising from intractability:

- large sizes, complexity, heterogeneity, noisiness, and
- different time and space scales of the data

**"Embedded" data scientists**: problem-specific heuristic methods, HPC

# 3. Conclusions

## Holistically Mine All Available Data




### Methodologies

- **Mathematical formalisms**
  - Capture **multi-scale organization**
  - **Dynamics, stochasticity** of the data,…
  - E.g., multiplex networks, hypergraphs, simplicial complexes …
- **Algorithms** to compute and **extract information** from those formalisms

**Computational issues** remain to be addressed, arising from intractability:

- large sizes, complexity, heterogeneity, noisiness, and
- different time and space scales of the data

**"Embedded" data scientists**: problem-specific heuristic methods, HPC

# Acknowledgements

ICREA

## Funding:


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

erc European Research Council
Funding for top researchers from anywhere in the world

NSF

NSERC CRSNG

Ontario

gsk

IBM

Google

J&J

Farr The Farr Institute of Health Informatics Research

Pp the PROSTATE project

arrs SLOVENIAN RESEARCH AGENCY

Република Србија
МИНИСТАРСТВО ПРОСВЕТЕ НАУКЕ И ТЕХНОЛОШКОГ РАЗВОЈА


## Group members (present and past):

1. Dr. Noel Malod-Dogning
2. Dr. Julia Petschnigg
3. Dr. Chhedi Gupta
4. Dr. Remi Momo
5. Sam Windels
6. Thomas Gaudelet
7. Dr. Omer Yaveroglu
8. Prof. Tijana Milenković
9. Dr. Oleksii Kuchaiev
10. Dr. Vesna Memišević
11. Vladimir Gligorijevic












bcs The Chartered Institute for IT

ACADEMIA EUROPAEA } The Academy of Europe

# Thank you

**Comments and Questions**