INSTITUTO LELOIR
FUNDACIÓN

Hybridization of Chemoinformatic and Bioinformatic Information Networks for Graph Diffusion of Drug-Target Interactions

Bruno Kaufman[1], Dr. Ariel Chernomoretz[1,2]

1) Integrative Systems Biology Lab, Fundación Instituto Leloir
2) Department of Physics, FCEN, Universidad de Buenos Aires

Drug repurposing

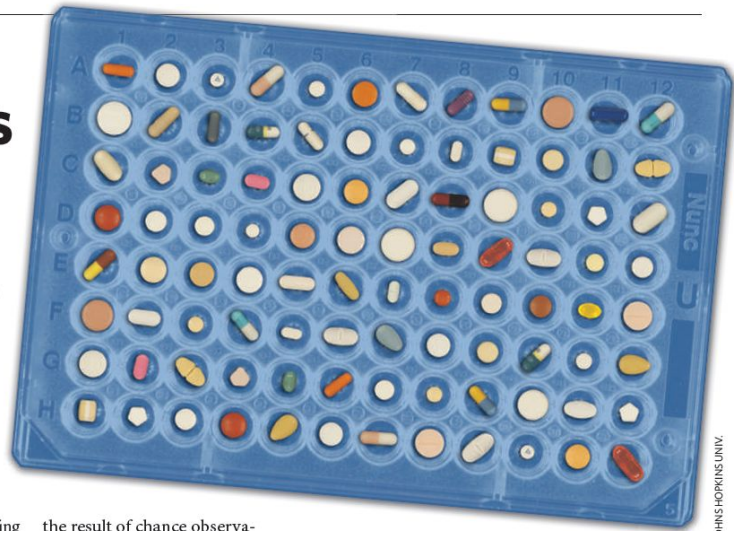
Novel approaches are needed when there is no clear financial incentive for biopharma companies

Vol 448|9 August 2007 nature

COMMENTARY

New uses for old drugs

It takes too long and costs too much to bring new drugs to market. So let's beef up efforts to screen existing drugs for new uses, argue **Curtis R. Chong** and **David J. Sullivan Jr.**



Fast, affordable drug development is a vision that contrasts sharply with the current state of drug discovery — which also neglects too many diseases of the poor. An analysis[1] of 68 approved drugs estimated that it takes an average of 15 years and US$800 million to bring a single drug to market. And despite a doubling

the result of chance observa-

"The most fruitful basis for the discovery of a new drug is to start with an old one"

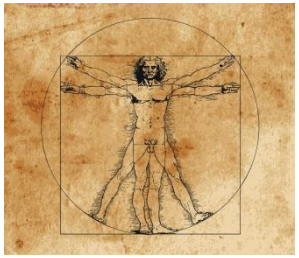
- Known pharmacokinetics
- Known safety profiles
- Often already approved by regulatory agencies for human use
- Already approved by regulation agencies for human use

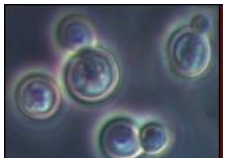
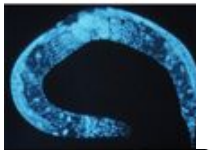
~40% of cost US$

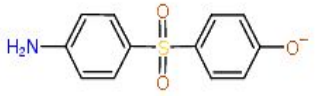
- Library screening: biochemical/pharmacological assays
- In-silico screening: chemoinformatics approaches / computational docking
- **Prioritization tasks: data-mining of biochemical available data**

Available data

Model organisms



Chemical compound

bioactive

Druggable target protein

PubChem

Compounds: 91,699,023 (2,283,528 tested)
Substances: 223,181,033 (3,576,050 tested)
BioAssays: 1,218,668 (91 RNAi)
BioActivities: 230,658,885
Protein Targets: 10,182
Gene Targets: 19,779

ChEMBL

Compounds: 1,928,903
BioActivities: 13,967,816
Protein Targets: 11,019
Publications: 62,501

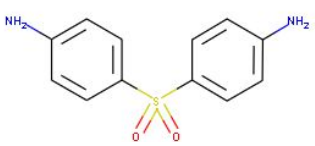
DrugBank
Drug & Drug Target Database

FDA-approved: 2000
Experimental: >6000
Protein Targets: 4333

Urán Landaburu, Lionel. *TDR Targets 6: driving drug discovery for human pathogens through intensive chemogenomic data integration.* Nucleic Acids Research (2020)
Agüero, Fernán. *Genomic-scale prioritization of drug targets: the TDR Targets database*. Nature Reviews Drug Discovery (2008)

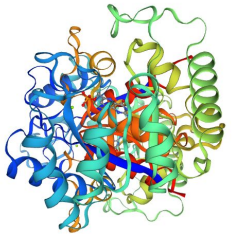
Our dataset

Known drug-protein interactions:
Curated ground truth from evidence of interactions.



Known drug
(dapsone)

bioactivity



Known target
(DHPS1, *mycobacterium leprae*)

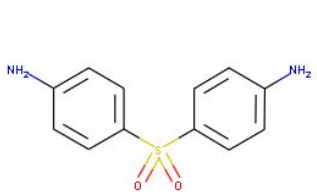
Molecules involved in known interactions: 1M out of 7M (14%)
Targets involved in known interactions: 5.7k out of 560k (1%)

Low information density

The problem: Inferring reliable predictions between potential drugs and useful targets.

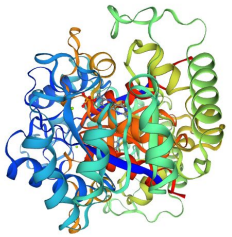
Our dataset

Known drug-protein interactions:
Curated ground truth from evidence of interactions.



Known drug
(dapsone)

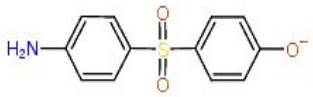
bioactivity



Known target
(DHPS1, *mycobacterium leprae*)

similarity

Molecular similarity measures:
Relating compounds to one another through structure, not function.



Potential drug

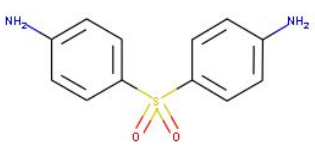
Molecules involved in known interactions: 1M out of 7M (14%)
Targets involved in known interactions: 5.7k out of 560k (1%)

Low information density

The problem: Inferring reliable predictions between potential drugs and useful targets.

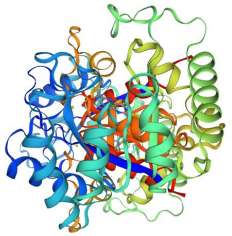
Our dataset

Known drug-protein interactions:
Curated ground truth from evidence of interactions.



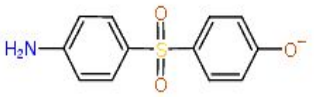
Known drug
(dapsone)

bioactivity



Known target
(DHPS1, *mycobacterium leprae*)

similarity



Potential drug

Molecular similarity measures:
Relating compounds to one another through structure, not function.

"Guilt by association"
Similar drugs interact with a similar set of targets.

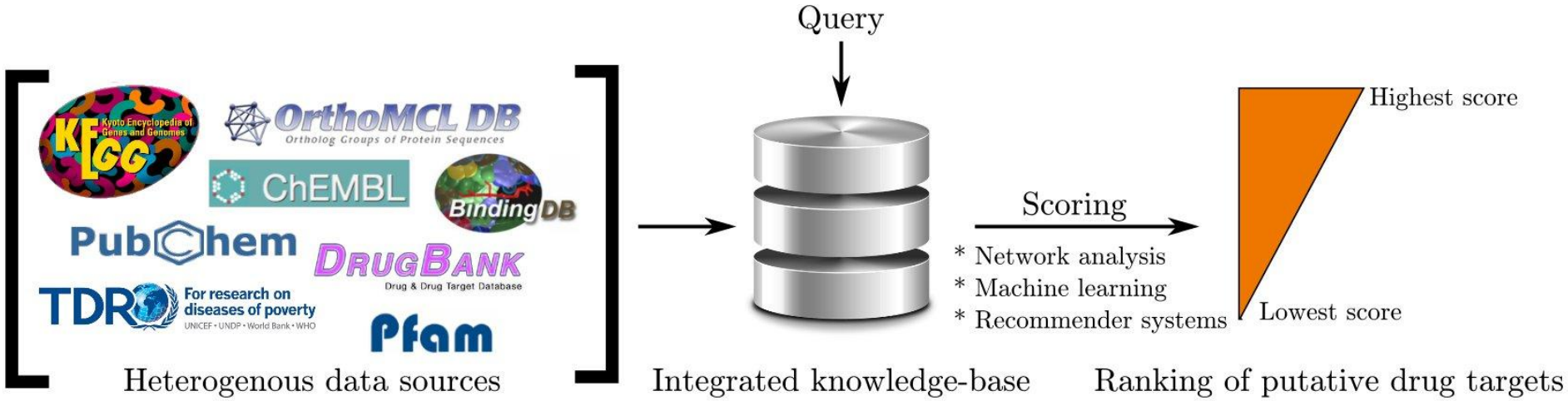
Molecules involved in known interactions: 1M out of 7M (14%)
Targets involved in known interactions: 5.7k out of 560k (1%)

Low information density

The solution: Aggregate heterogeneous similarity measures to infer new drug-target interactions.

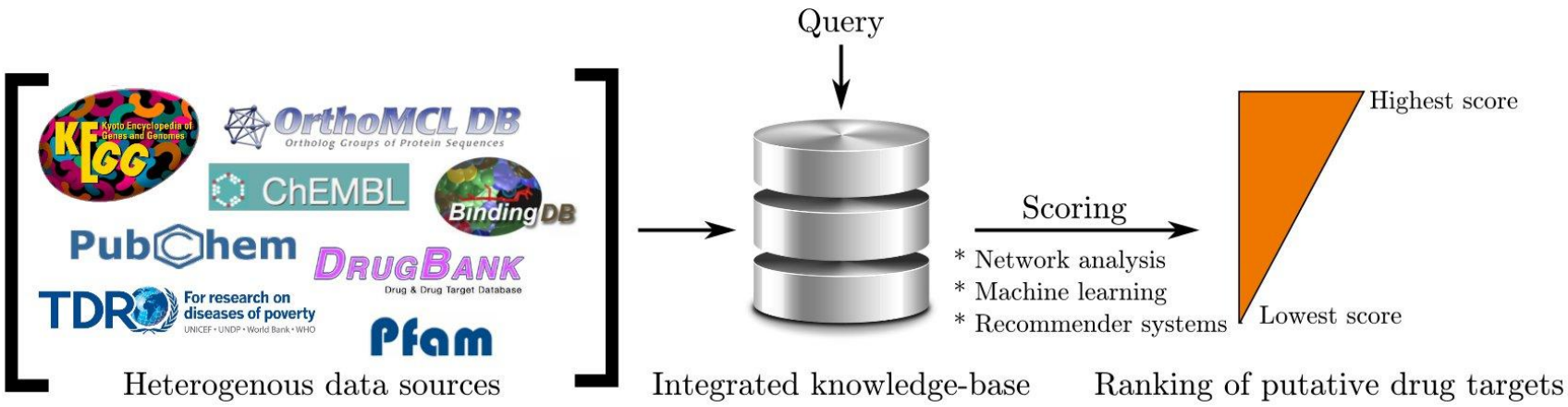
In silico Drug Target Prioritization

Prioritization strategies:

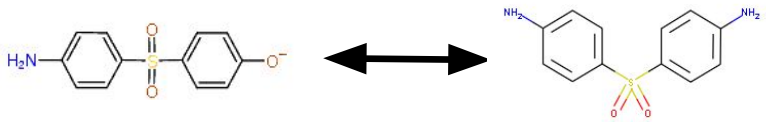


Our approach: Complex networks as a mathematical framework to **integrate data and perform prioritization tasks**.

Our procedure



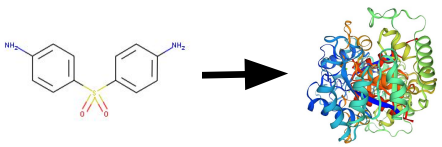
Multi-layered chemical space



Tanimoto layer: bitstring-based chemical similarities
Scaffold layer: structural scaffold similarities
Target layer: chemical similarities from shared targets

$$\begin{matrix} A_t \\ A_s \\ A_{DT} \end{matrix}$$

Drug-Target associations



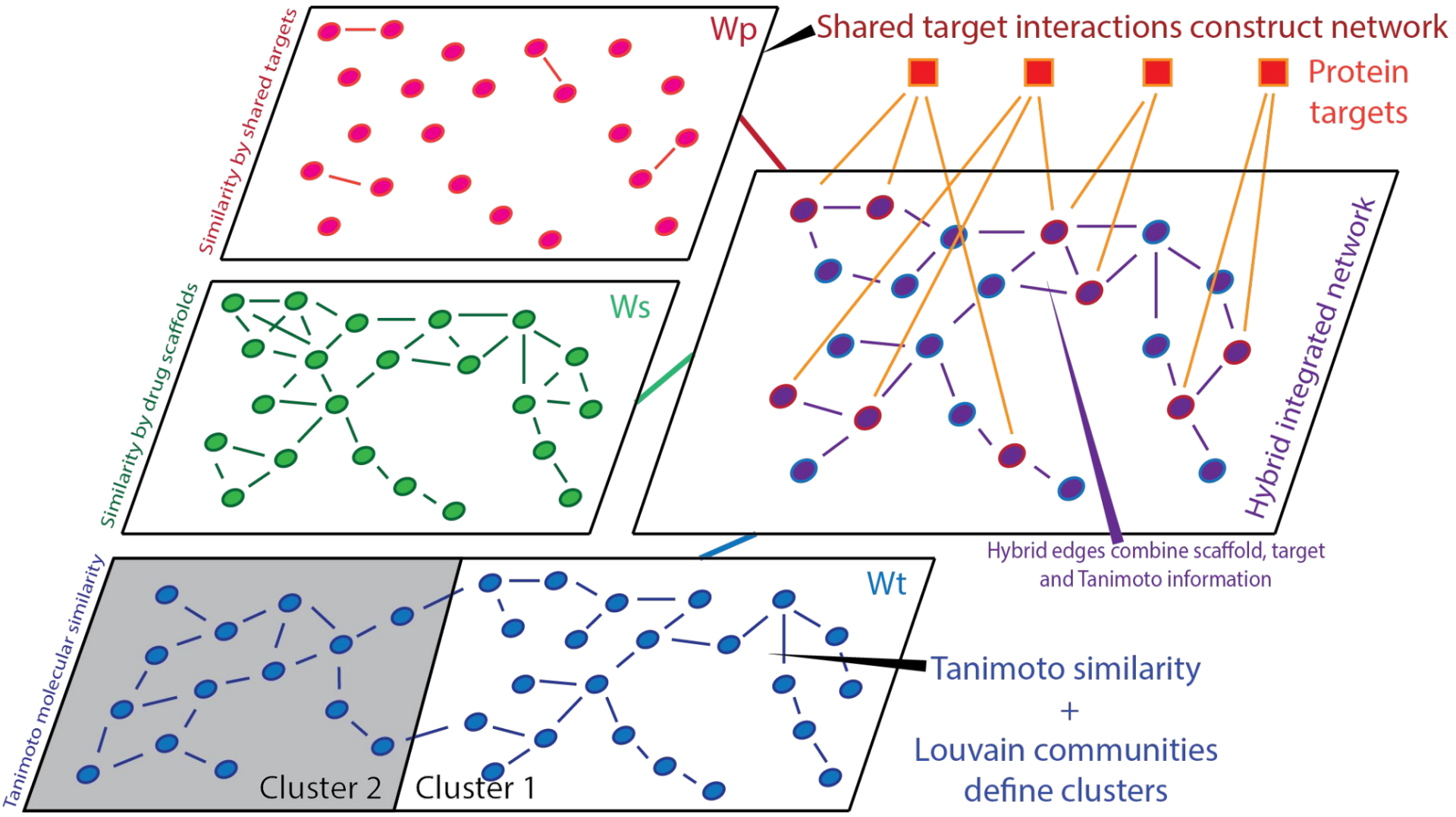
$$L = D - A$$

$$\frac{d\vec{y}}{dt} = \alpha(\hat{\vec{y}} - \vec{y}) - (1-\alpha)\overline{\overline{L}}\vec{y}$$

Network information within Laplace operator
Weighted sum combines three layers

Learn optimal network combination:
Diffuse from training set, rank node scores, compare to test set, optimize recovery score.

Integrating a Multilayer Network



Learning by communities (divide and conquer)

Clustering on chemical layer: Louvain communities identified on Tanimoto similarity network (well connected, no training/validation info).

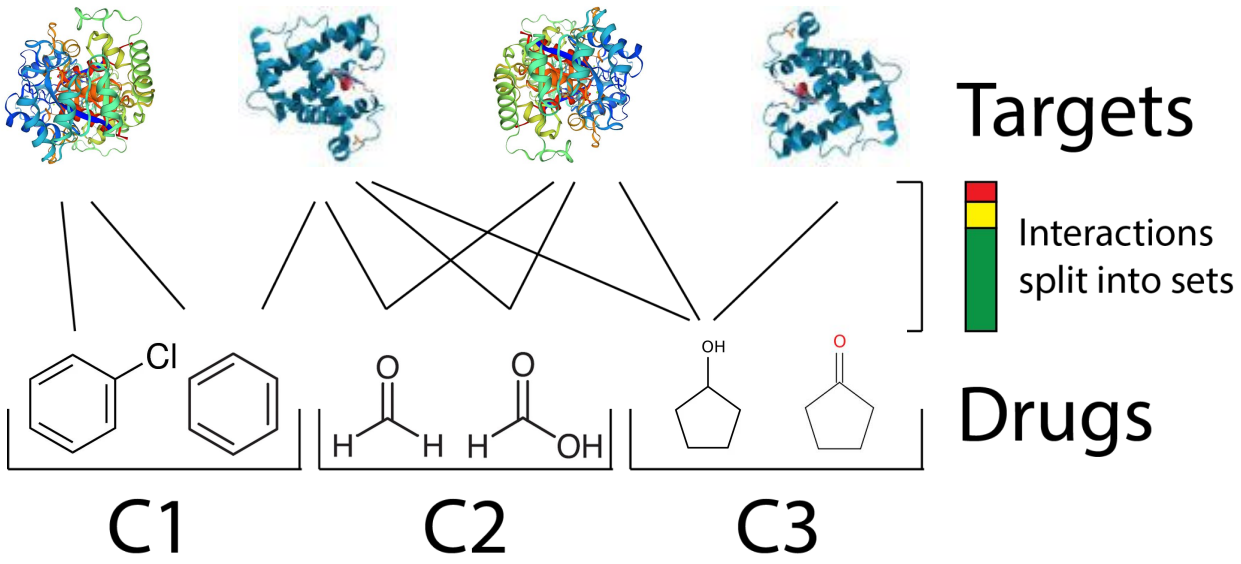
Community statistics:

- 6.1k out of 7.4k clusters contain at least 20 active drugs.
- 1.5M drugs represented within these clusters, 604k active drugs.

Split drug-target interactions:
Training (70%)
Test (20%)
Validation (10%)

For each cluster:

- Learn optimal network combination using training and test datasets.
- Perform network diffusion with optimal parameters to recover validation set.



Why not train on a per-target basis?

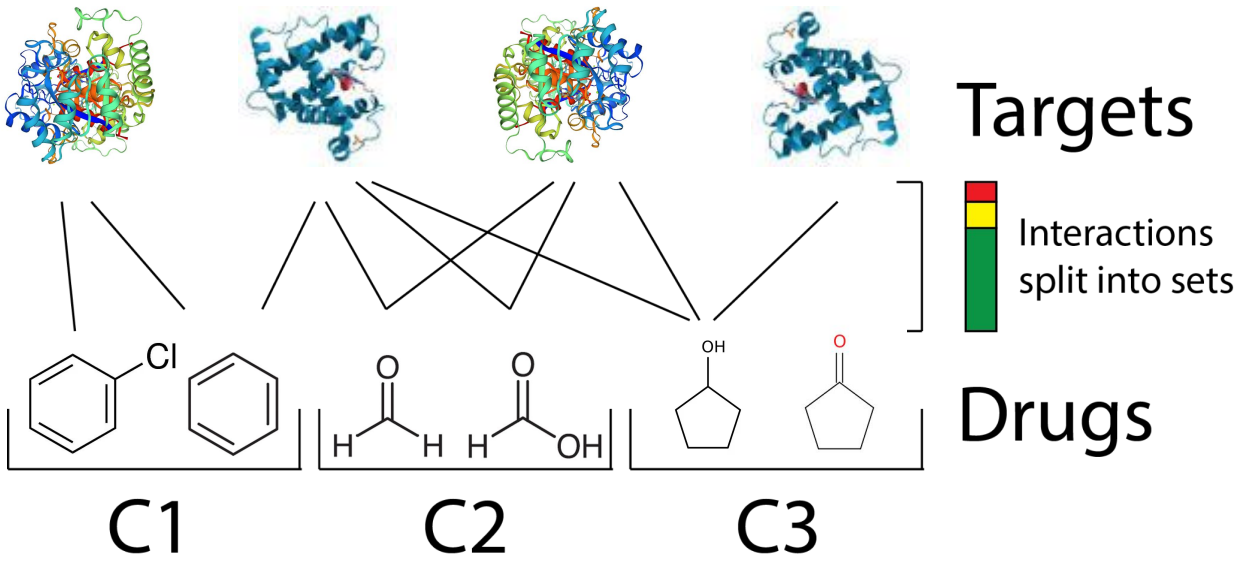
Previous method: Individual learning. Per-target training severely underperformed existing methods due to low information density. **Usable targets < 500.**

New method benefit: Local learning based on topology. Per-cluster knowledge pool counteracts this disadvantage, yielding better training. **Usable targets > 3000.**

Split drug-target interactions:
Training (70%)
Test (20%)
Validation (10%)

For each cluster:

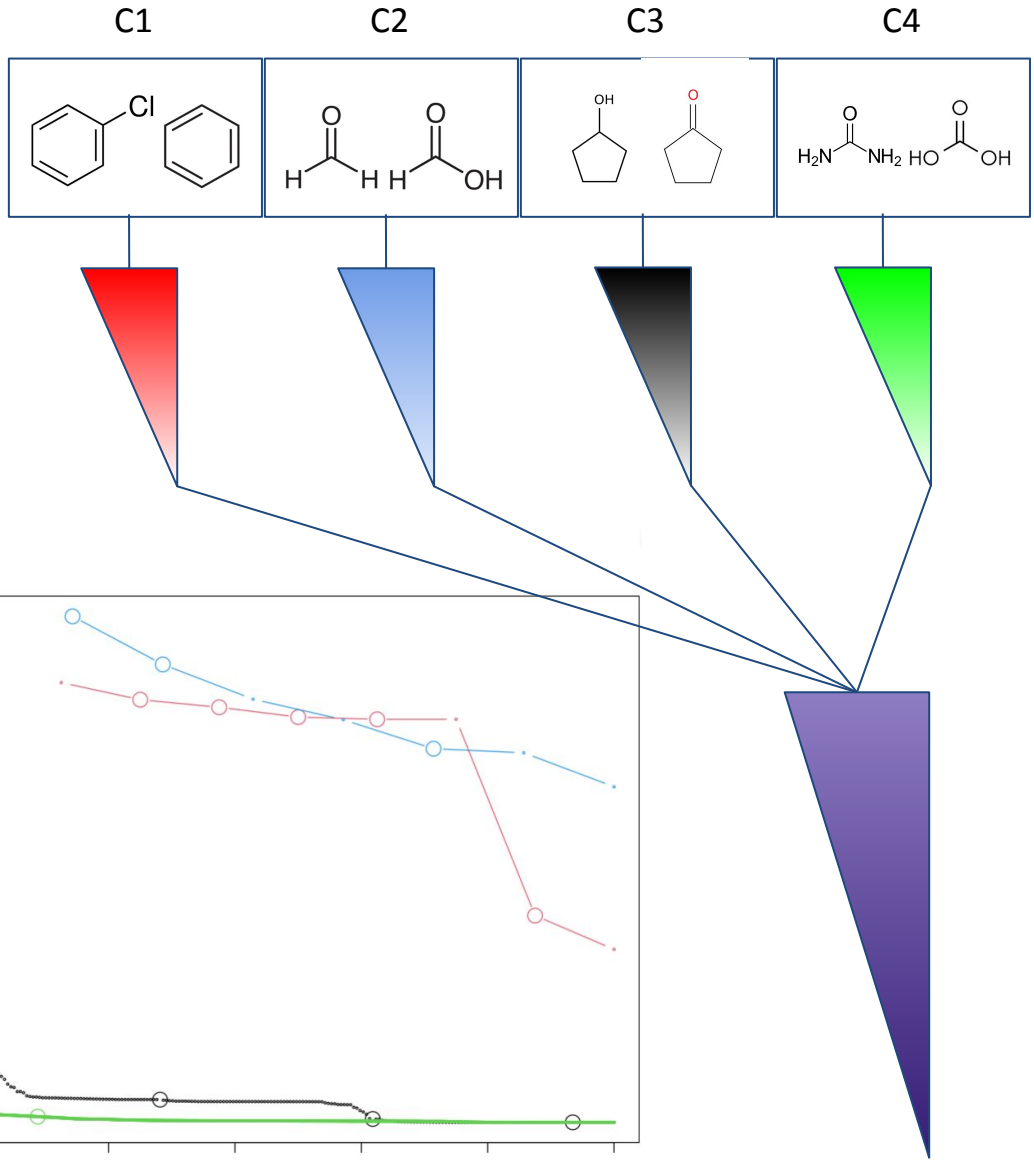
- Learn optimal network combination using training and test datasets.
- Perform network diffusion with optimal parameters to recover validation set.

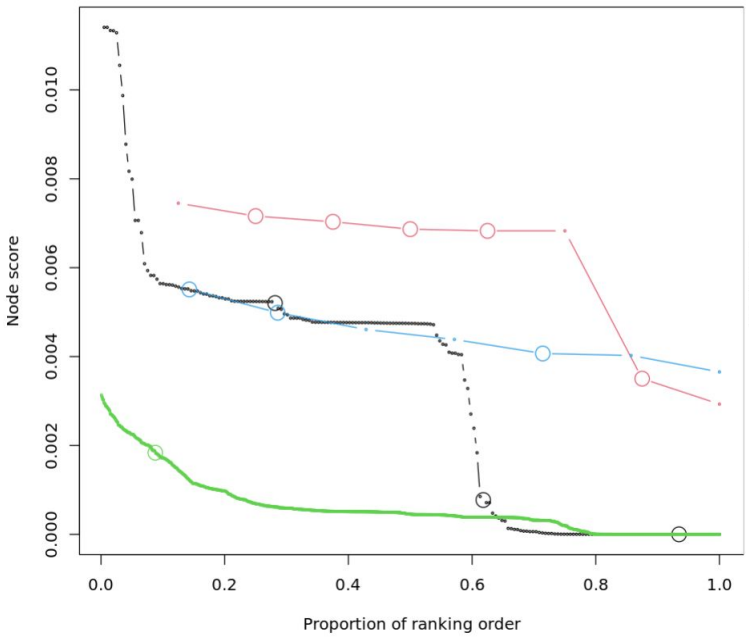


Ranking consolidation

Individual clusters have independent rankings. Drug scores are normalized over their clusters, then integrated.

Restructuring the ranking to reflect true recovery.





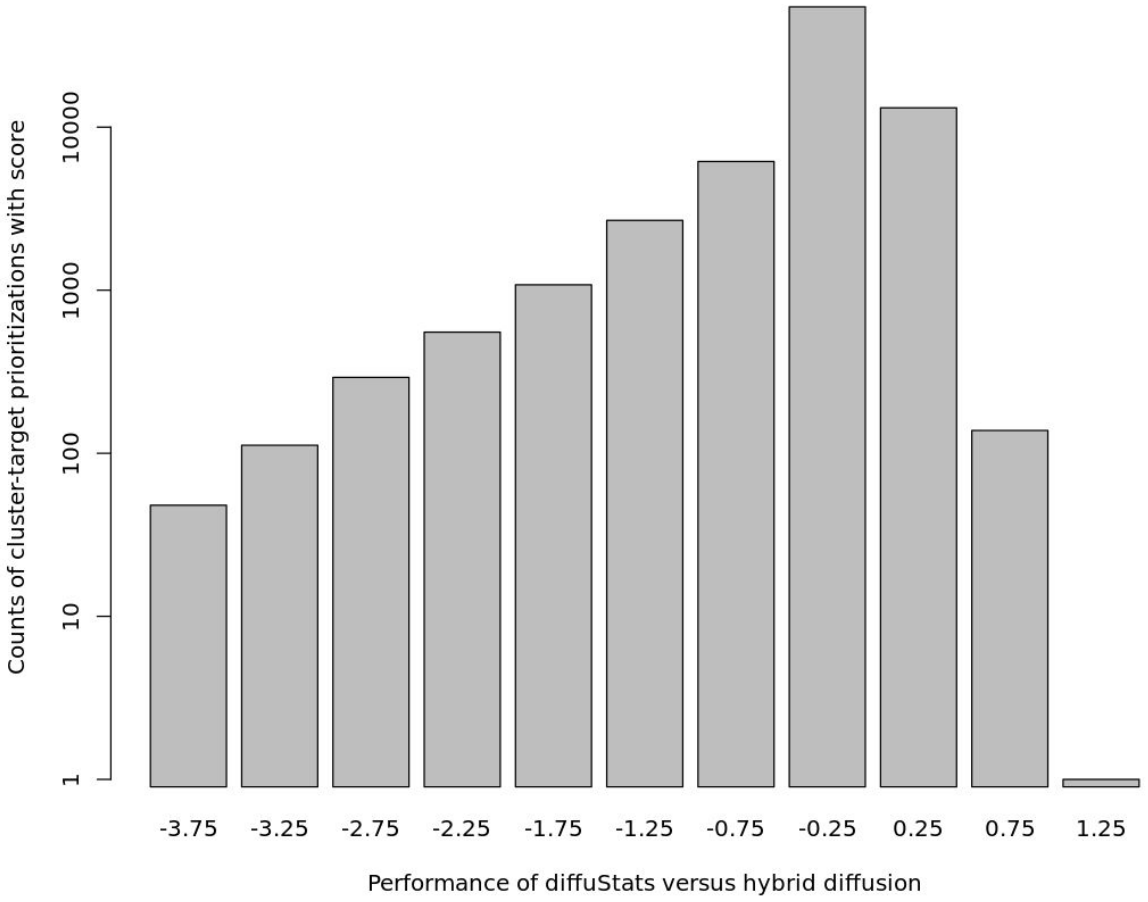
Comparison to existing implementations

diffuStats applies statistical considerations on result of network diffusion.
Ranges and qualities of scores make **consolidation unreliable.**

Performs better in **only 10k** cluster-by-cluster prioritizations, compared to **40k** for our method.

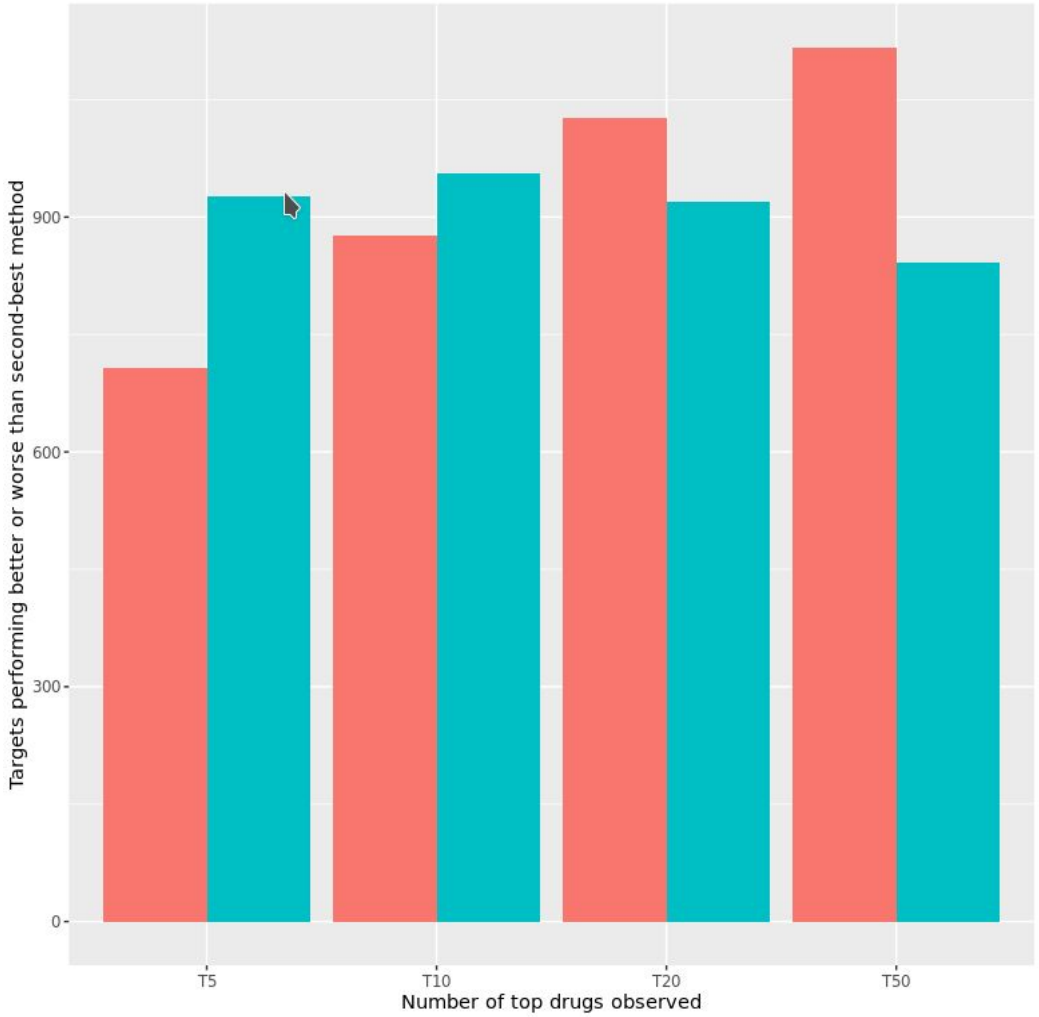
Underperforms our method even without consolidating.

Second-best method to compare to will be the **pure Tanimoto network.**



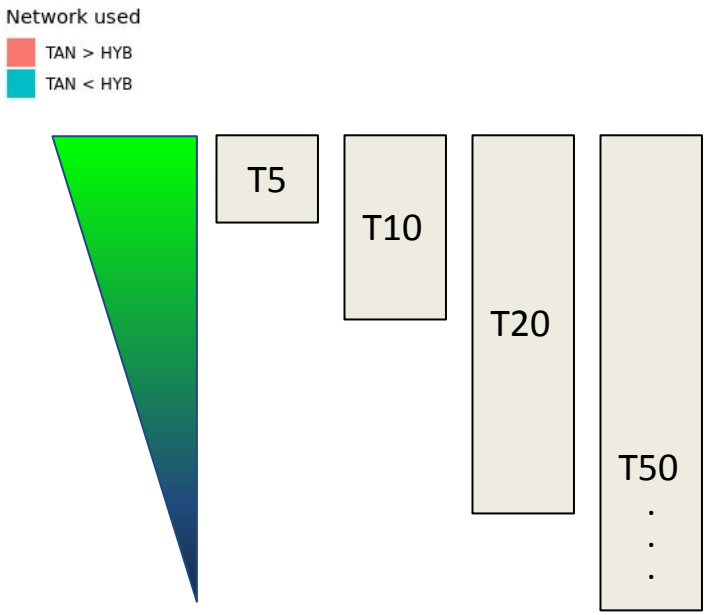
Performance improvement

Second-best method: Non-hybridized, pure Tanimoto drug layer.



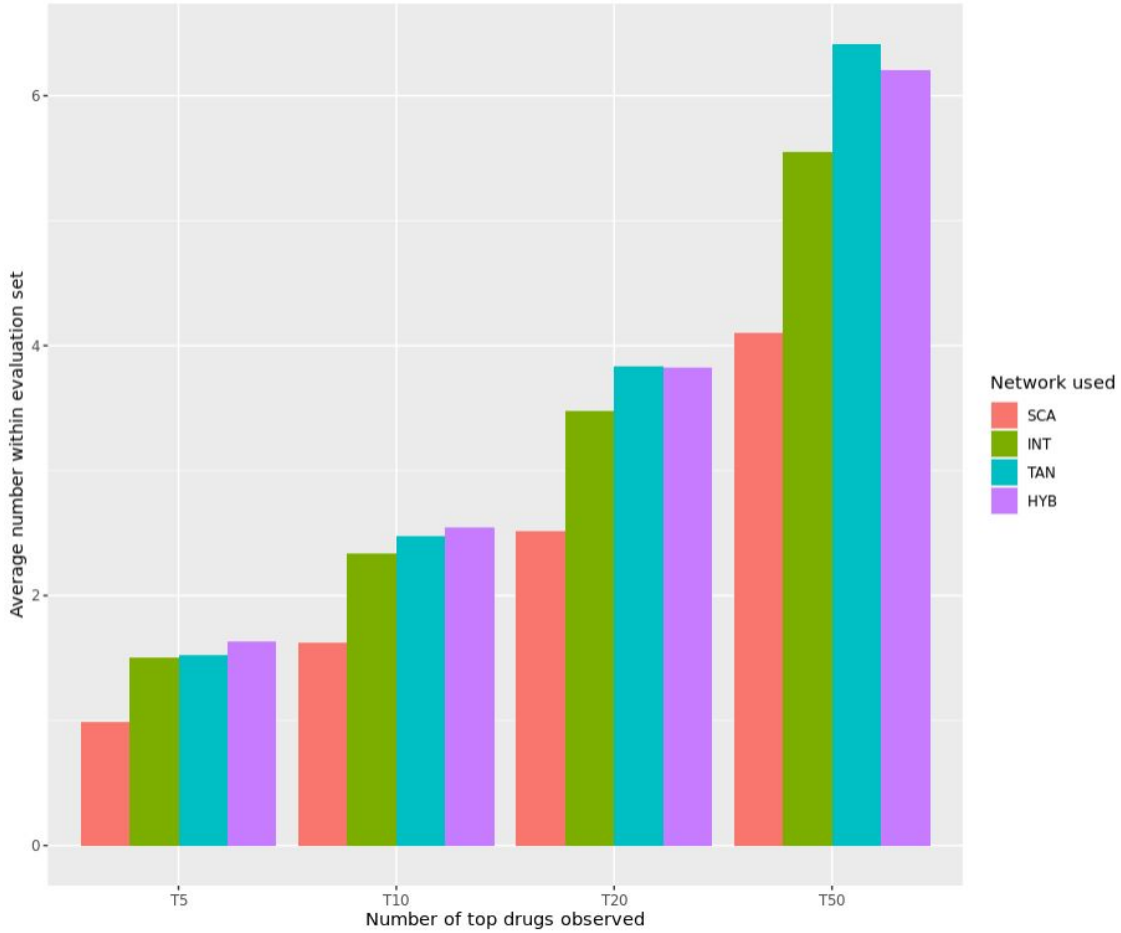
Ranking-oriented improvement: The hybrid approach prioritizes the top-5 score over others.

Concise recommendations for target-specific queries.



Performance improvement

The hybrid network outperforms individual network layers for start of ranking



Consistent results: Average recall shows trend parallel to that previously shown.

Cost function gives priority to first elements in ranking. Training the network parameters with an independent measure (not T5-T50) produced favorable results for T5-T10.

Summary

Divide and conquer: Previous methods (targetwise training) suffered from the curse of low information density: **small training sets, even smaller validation sets.** Dividing by cluster **aggregates information by similar drugs**, providing a **strong basis for training and validating.**

Heterogeneous knowledge pool: Tanimoto, scaffold and drug-target measures constitute **qualitatively different meanings**. Yet, they may be combined through training to **provide higher recovery scores**.

Prioritizing topmost ranking order: Our training acquires the network combination needed to **improve top 5-10 scores**, delivering only the most reliable predictions for queries.

Acknowledgements

Integrative Systems Biology, Fundación Instituto Leloir
Lab director: Dr. Ariel Chernomoretz

Postdoctoral researchers:
Dr. Maximiliano Beckel

Ph. D. candidates:
Ariel Berardino

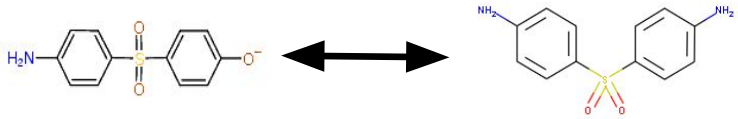
BSc. candidates:
Ingrid Heuer
Romina D'Alessandro
Bautista Buyatti

universidad de buenos aires - exactas
departamento de física

INSTITUTO LELOIR
FUNDACIÓN

CONICET

Appendix A: Tanimoto similarity

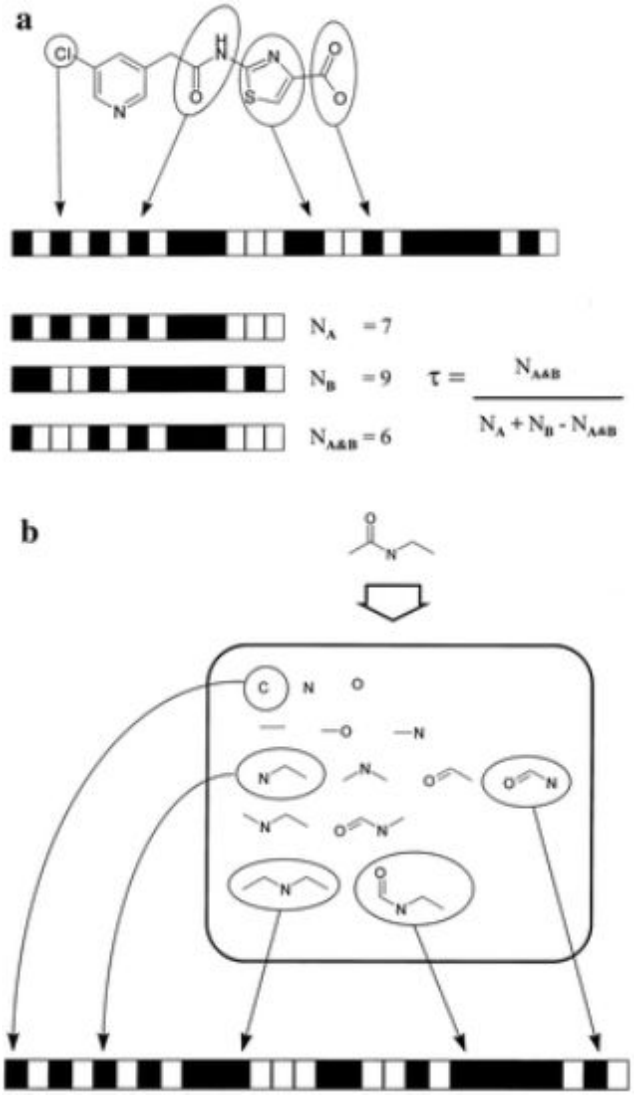


Bitstring: Array indicating whether specific components or qualities exist within a molecule.

Jaccard coefficient: The significance of sharing components of the bitstring is evaluated using their intersection and union:

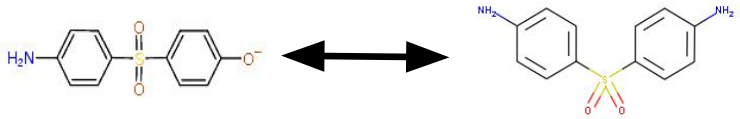
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Network construction: Weighted, using scores > 0.8

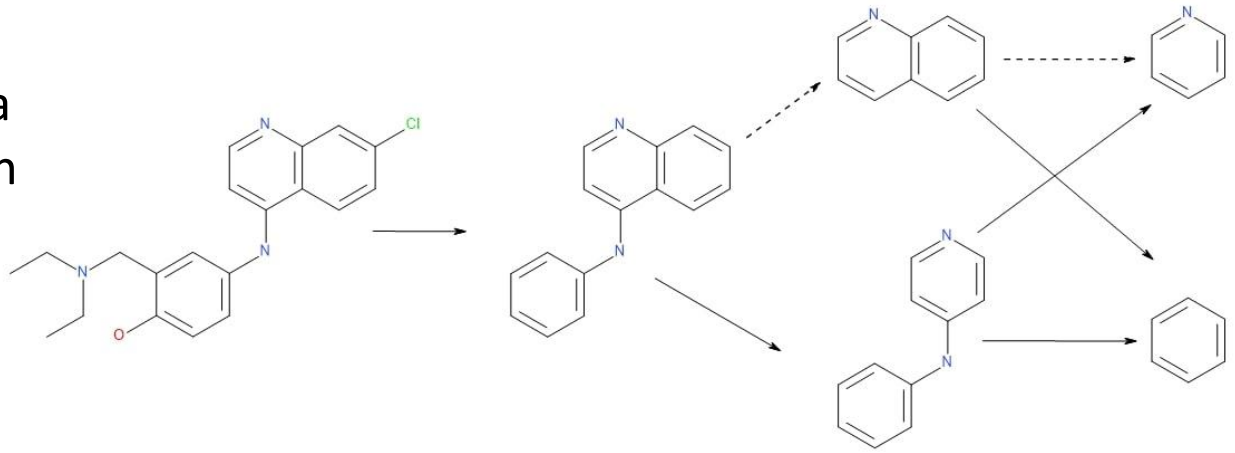


Flower, Darren R., *On the Properties of Bit String-Based Measures of Chemical Similarity*. (1998).

Appendix B: Scaffold similarity



Bemis-Murcko:
Removing “layers” of a molecule, starting with side structures.

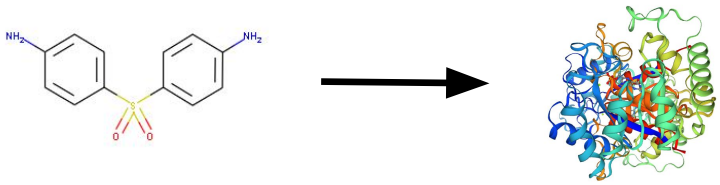


Levels: Repeating the process will yield smaller scaffolds that more dissimilar molecules will share.

Network construction: If molecules meet at a layer 3 scaffold or less, they are assigned a similarity score inversely proportional to the layer depth.

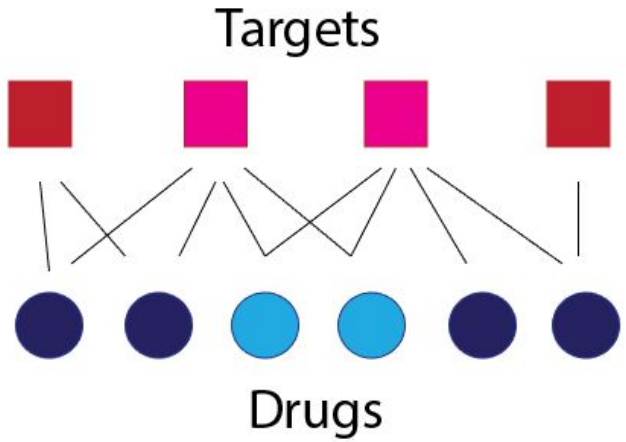
Bemis, Guy W., and Murcko, Mark A., *The properties of known drugs. 1. Molecular frameworks.* (1996).

Appendix C: Drug-target proximity



Interaction: A drug interferes with the protein's function, disrupting or altering its ordinary purpose.

Weighted proximity: Two drugs with a number of targets in common are assigned a link value by Zhou's bipartite compression. This weighs proximity by the **promiscuity (degree)** of both nodes.



$$s_{ij} = \frac{\sum_{\alpha=1}^{o} a_{\alpha i} a_{\alpha j}}{\sqrt{k_i k_j}}$$

PubChem

92M compounds, 2M tested.
230M drug-target interactions
10k protein targets, 20k gene targets

ChEMBL

2M compounds
14M drug-target interactions
11k protein targets

DrugBank
Drug & Drug Target Database

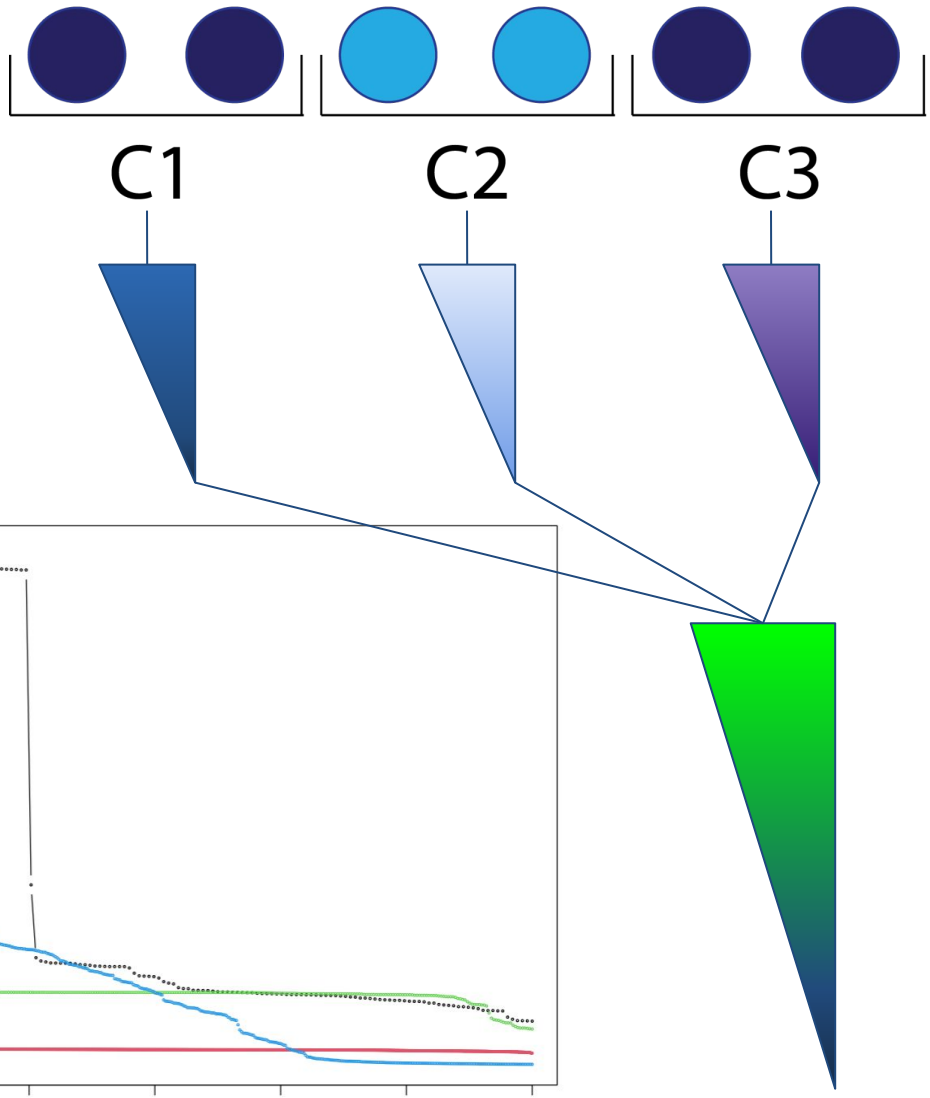
2k FDA-approved
>6k experimental
4k protein targets

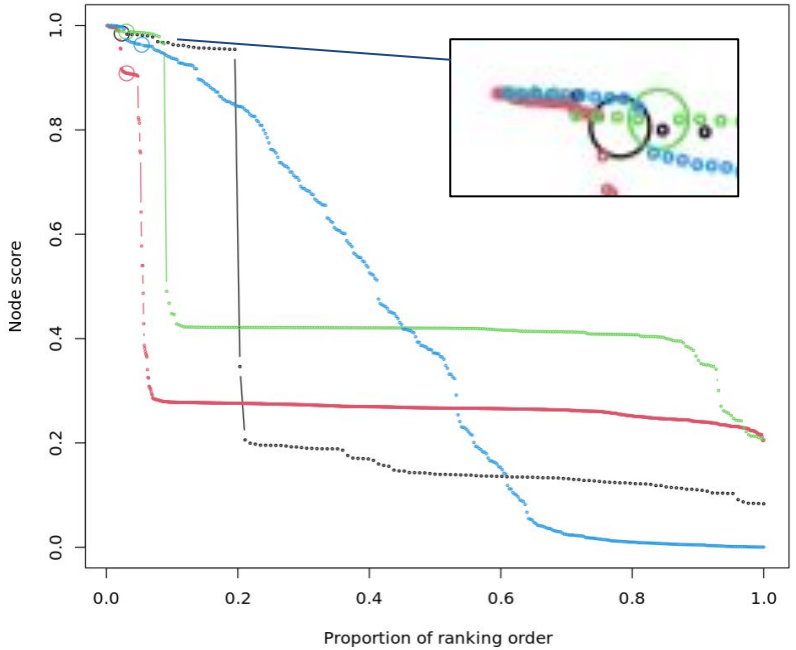
Zhou, Tao et al, *Solving the apparent diversity-accuracy dilemma of recommender systems.* (2010).

Appendix D:
Consolidation attempts on old methods

Known libraries adjust values through statistical criteria. Rankings given by their methods produce erratic results upon normalizing.

Existing methods don't conform to a "divide and conquer" cluster approach.





Appendix E:
Fitting consistency through different validations

Variance in parameters shows robustness. Results are replicable and minimally affected by random selection of datasets. **Median CV: 1%.**