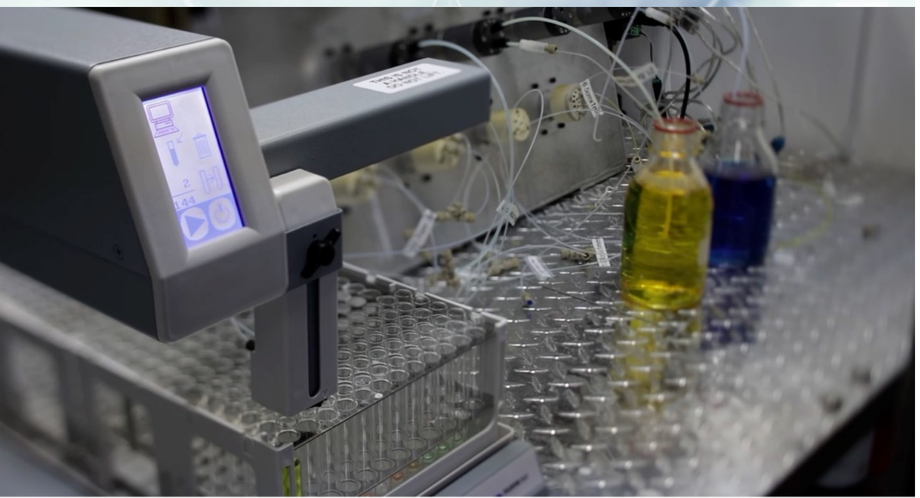


Accelerated discovery in chemistry: representation learning and recommendation systems

Dmitry Yu. Zubarev

IBM Research – Almaden

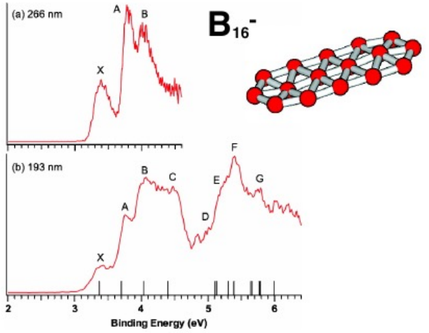
Brief scientific bio

Theoretical photoelectron spectroscopy of molecular clusters (Utah State University)

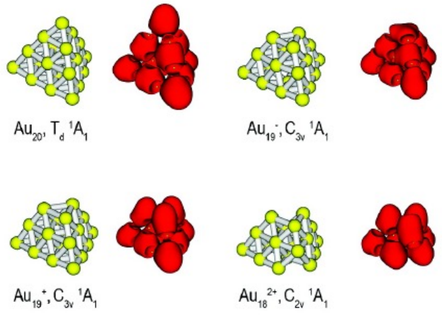
Quantum Monte-Carlo for fermionic systems (UC Berkeley)

Reconstruction of prebiological networks in origins of life (Harvard)

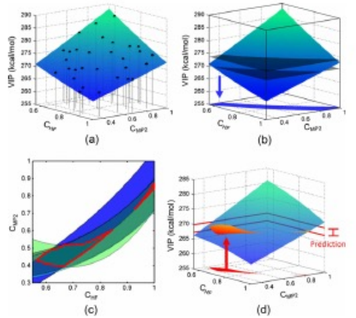
Photoelectron Spectroscopy in Molecular Beams



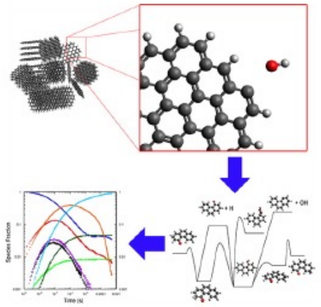
Adaptive Natural Density Partitioning

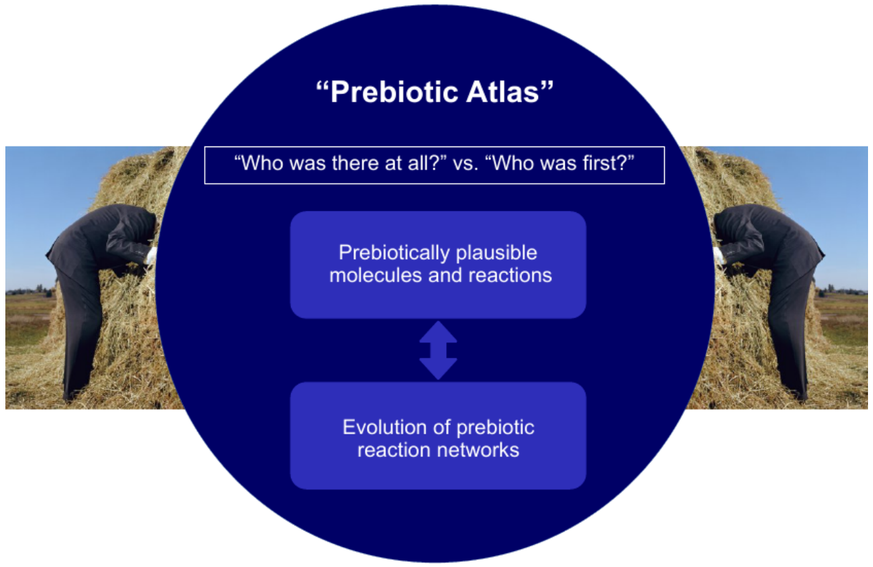


Quantum Chemistry with Error Bounds



Combustion Chemistry



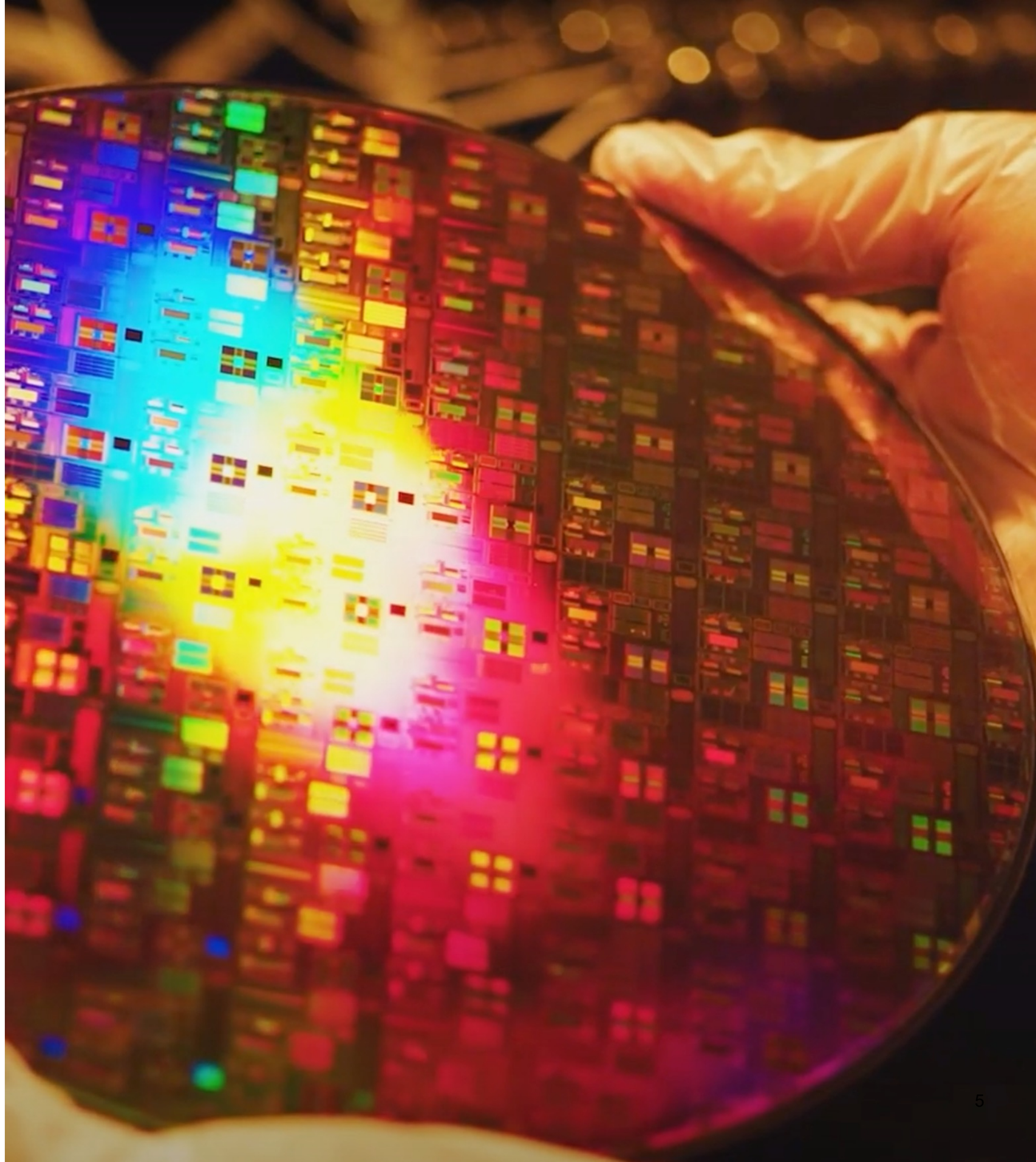


Outline

- Accelerated materials discovery
- Molecular representation
- Networks and network completion
- Representation learning and recommender systems
  - Example A: matrix factorization
    - antimicrobial resistance
  - Example B: semantic network embedding
    - ring-opening polymerization

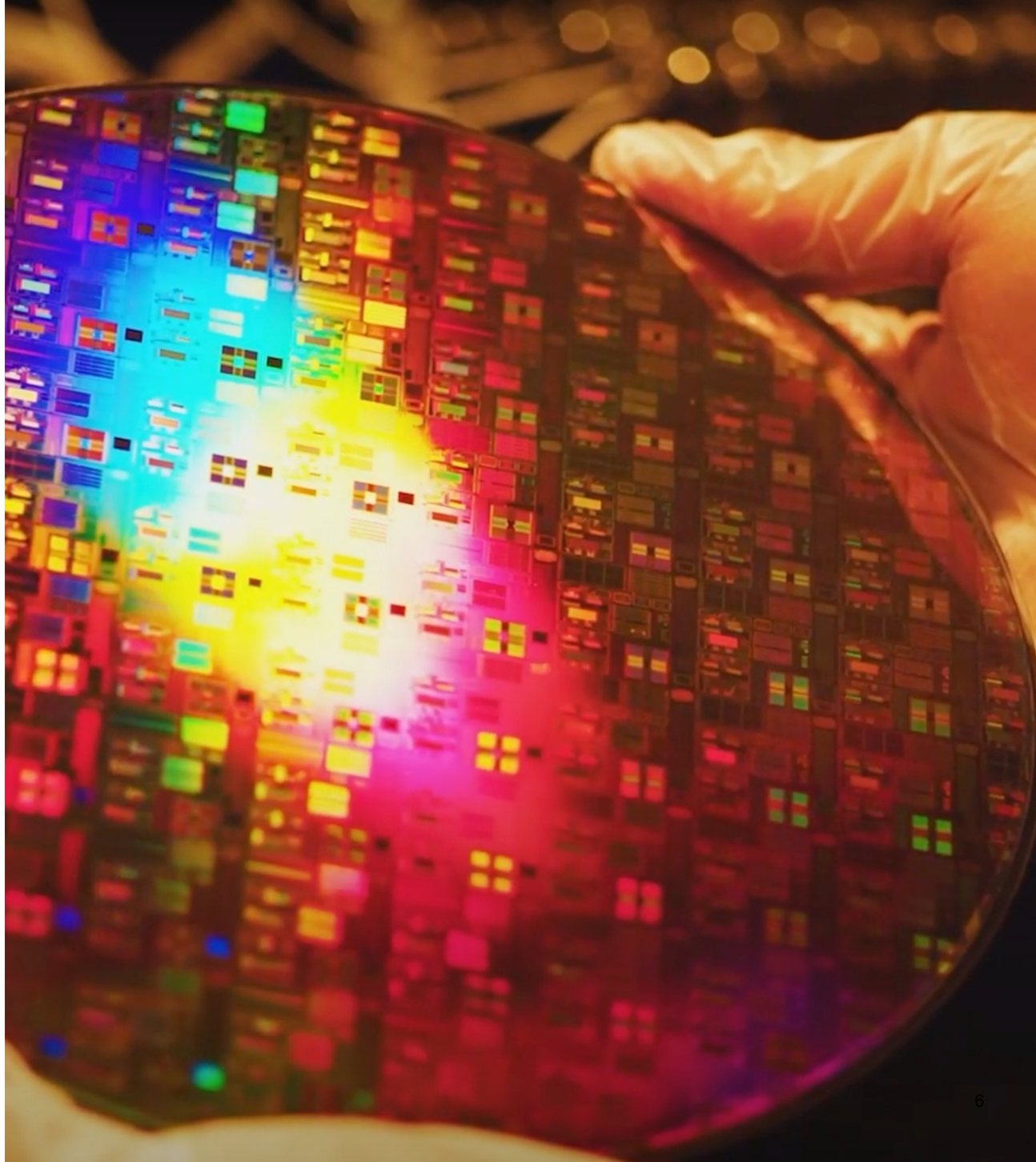
Accelerated materials discovery

10 years



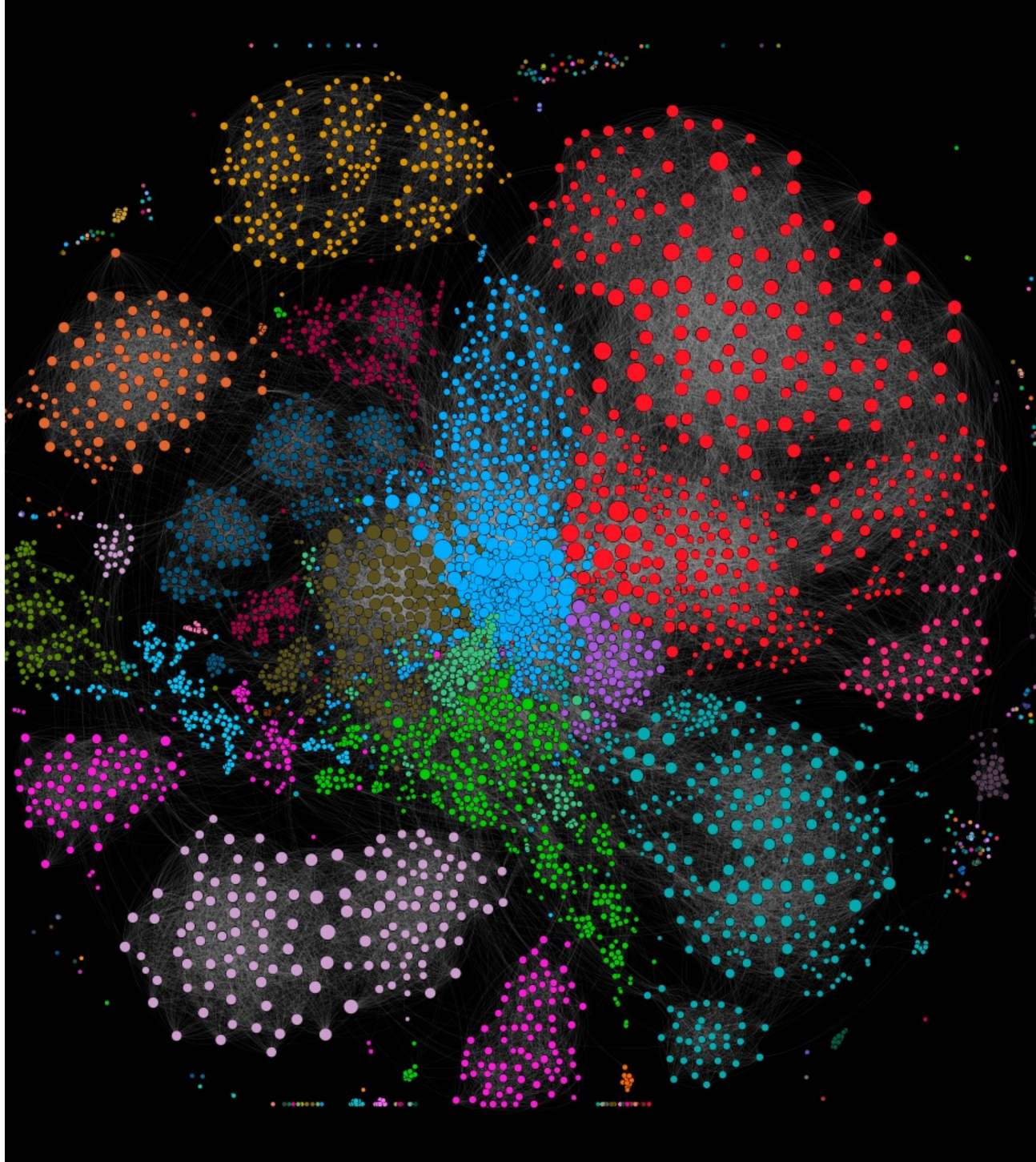
Time from identifying a discovery target to bringing material to market

10 years



10^{60}

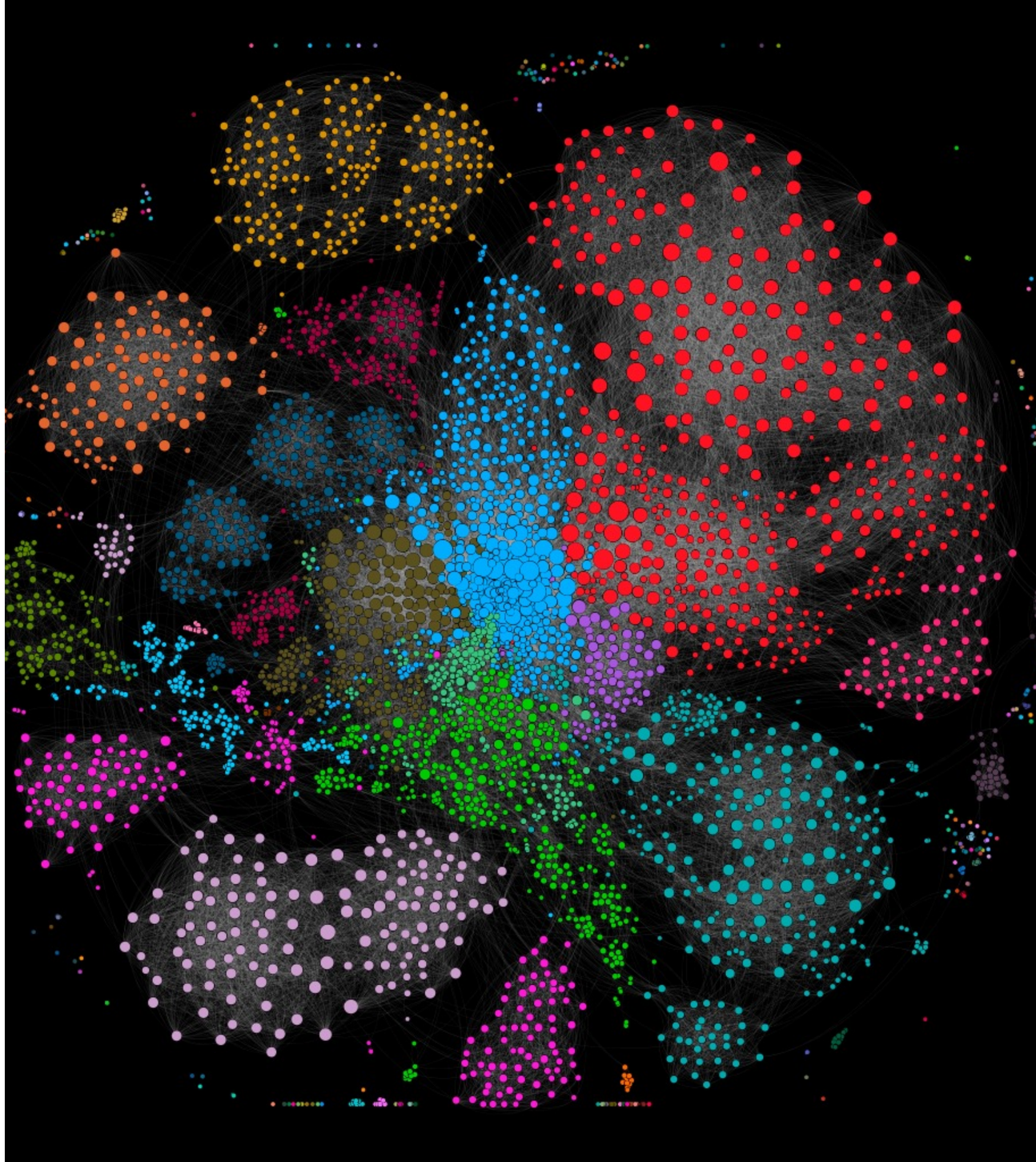
10^{8}



Number of molecules that might be accessible vs the content of PubChem

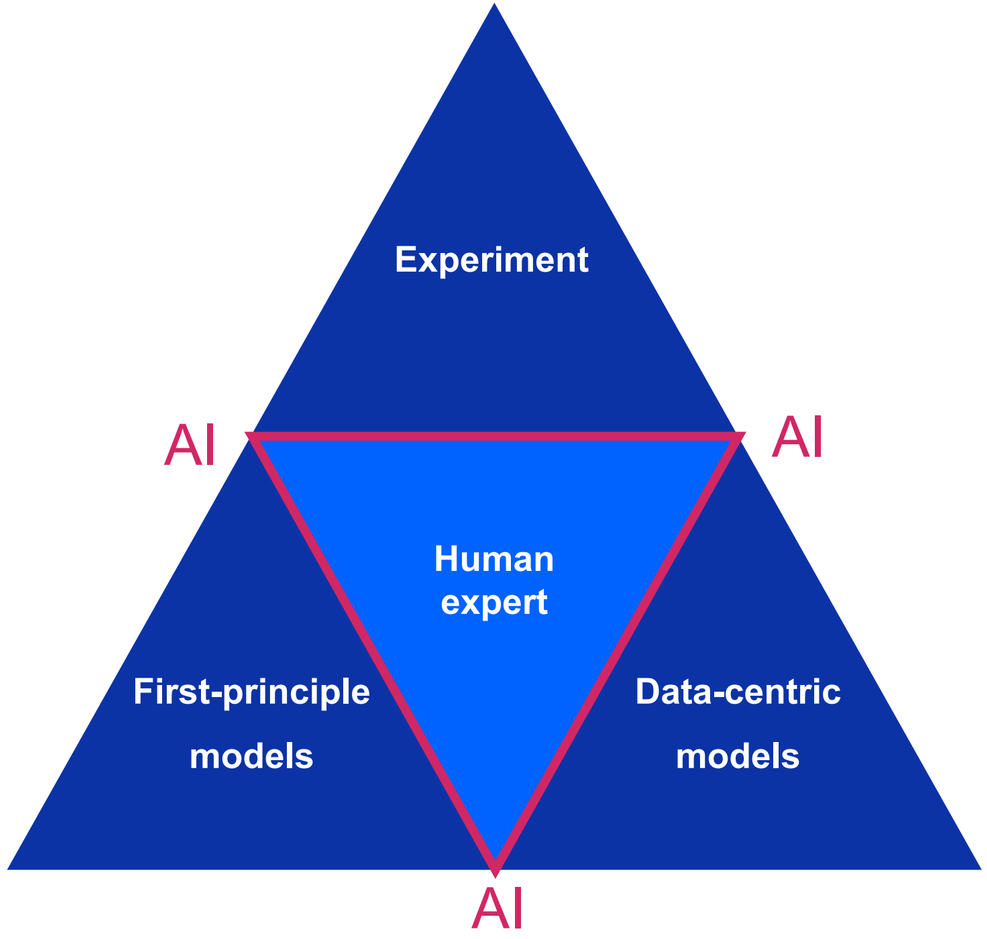
10^{60}

10^{8}



Accelerated materials discovery

- Constraints of materials research:
  - Data is scarce, sparse, and non-existent
  - Data acquisition has tangible costs (regardless if computational or experimental)
  - Deliverables are not flexible, and success is never guaranteed
- Goals:
  - end-to-end discovery of **molecular materials**
  - identify and resolve bottlenecks between components of the scientific process
  - focus on the evolution of human-AI interactions in scientific discovery



Molecular representation

Molecular formula

C_7H_6O

benzaldehyde tropone quinomethane

- Informs mass conservation
- Lacks uniqueness

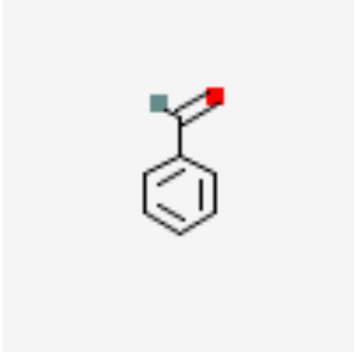
- …the chemical nature of a compound molecule depends on the nature and quantity of its elementary constituents and **its chemical structure.**

Structural theory

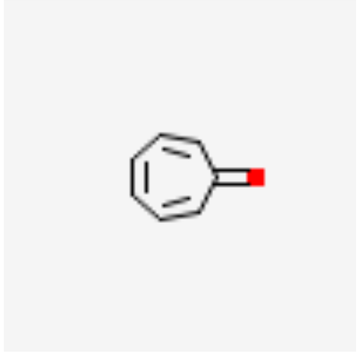
Lewis structure

C_7H_6O

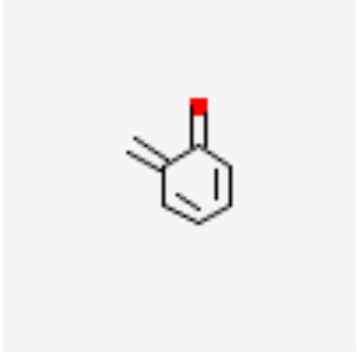
benzaldehyde



tropone



quinomethane

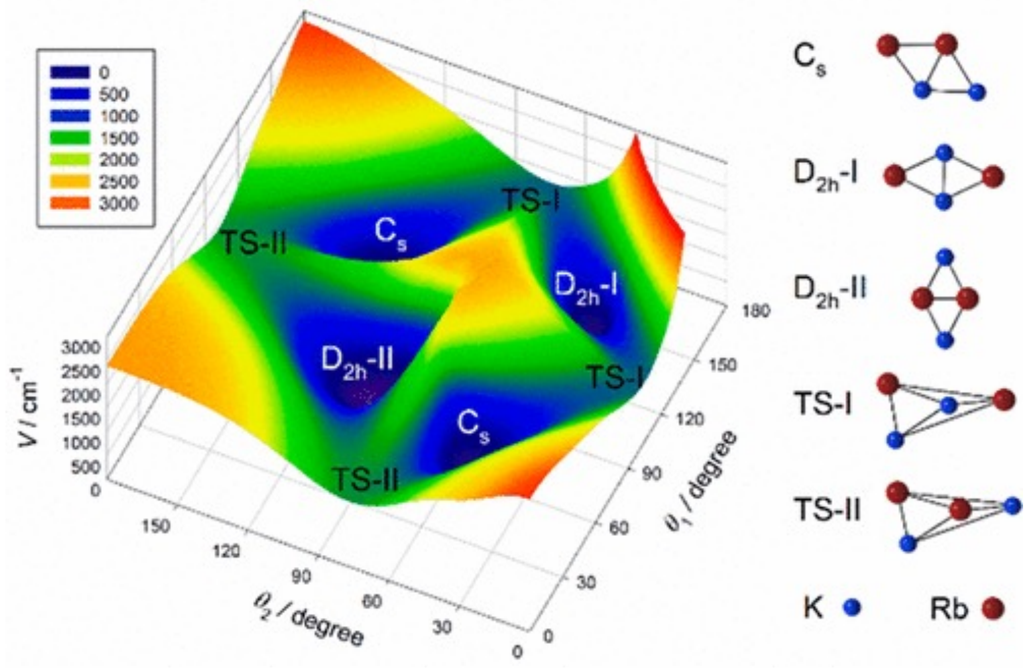


- Ad-hoc, predates quantum mechanics
- Electron counting, formal charge, octet rule

- De-facto language of organic chemistry
- Tons of heuristics that enable interpretation of empirical results, hypotheses generation and falsification

Quantum chemistry

Born-Oppenheimer approximation, potential energy surface



- Rigorous
- Excruciatingly detailed and verbose
- Costly

- Informs energetics and mechanistic analysis directly comparable with empirical evidence

Cheminformatic identifiers

- Digital representation
  - storage
  - search
  - analysis
- Convenient level of abstraction

2.1.1 IUPAC Name

benzaldehyde

Computed by Lexichem TK 2.7.0 (PubChem release 2021.05.07)

PubChem

International Chemical Identifier

2.1.2 InChI

InChI=1S/C7H6O/c8-6-7-4-2-1-3-5-7/h1-6H

Computed by InChI 1.0.6 (PubChem release 2021.05.07)

PubChem

2.1.3 InChI Key

HUMNYLRZRPPJDN-UHFFFAOYSA-N

Computed by InChI 1.0.6 (PubChem release 2021.05.07)

PubChem

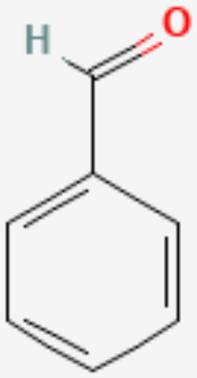
Simplified Molecular-Input Line-Entry System

2.1.4 Canonical SMILES

C1=CC=C(C=C1)C=O

Computed by OEChem 2.3.0 (PubChem release 2021.05.07)

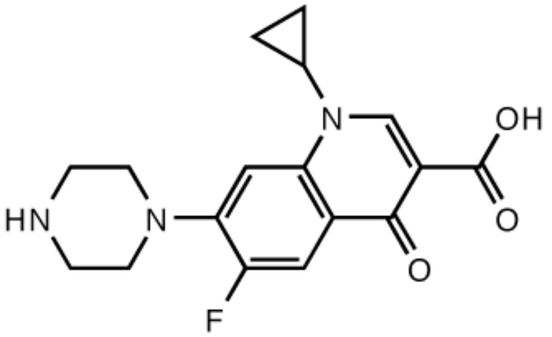
PubChem



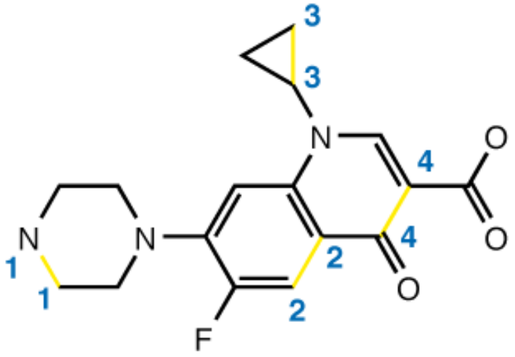
Cheminformatic identifiers

- Organic molecules are typically described as graphs
  - Atoms are nodes
  - Bonds are edges
  - Nodes and edges carry attributes
  - Graph layout follows chemical conventions
- String representation (SMILES, SELFIE) encode graphs

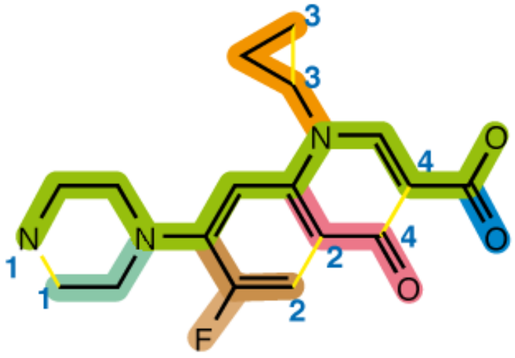
A



B



C

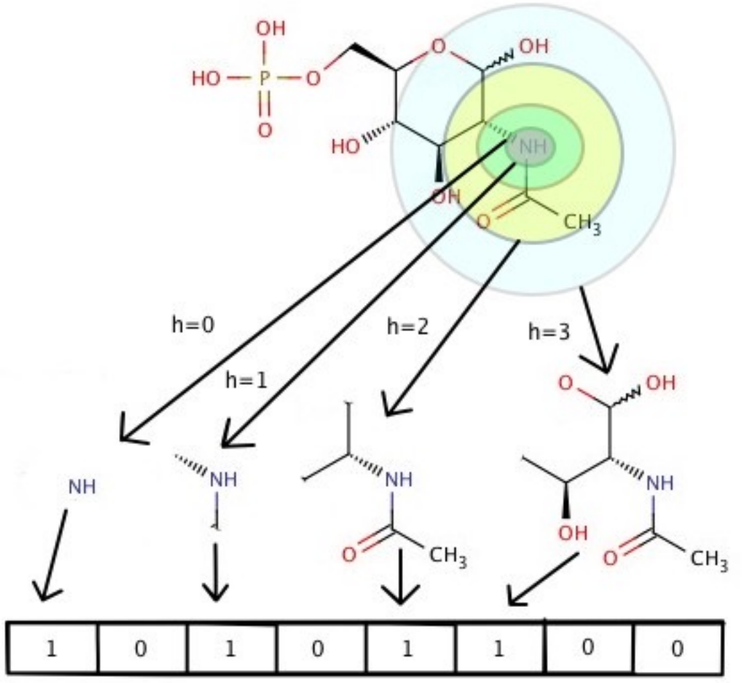


D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Molecular features

- Functional group
  - a transferable moiety associated with specific properties and activity
- Functional groups are sub-graphs in molecular graphs
- Chemically motivated featurization
  - enumeration of functional groups or fragments
  - enumeration of subgraphs



Takeaway

- Structural theory is the primary driver of feature engineering in cheminformatics
- Features describe parts of a molecule – molecular moieties – and some notion of their mutual positions
- Limiting cases
  - features are engineered using case-specific chemical knowledge (cf., group contribution method for a class of compounds)
  - features are constructed exhaustively and are agnostic to the specifics of the problem

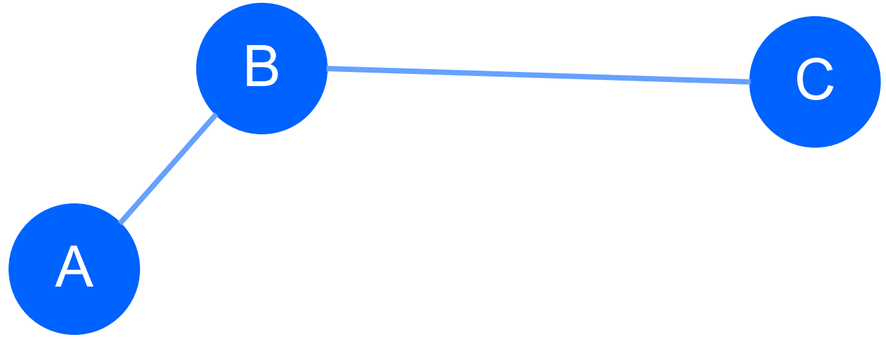
Networks and network completion

Networks

- Graph – a set of objects and relations
  - Objects – vertices (V)
  - Relations – groups of vertices, edges (E)
- Network – graph with attributes
- Network representation
  - Sets
  - Table (adjacency matrix)
  - Visualization

$G = (V, E)$
$V = (A, B, C)$
$E = (\{A, B\}, \{B, C\})$

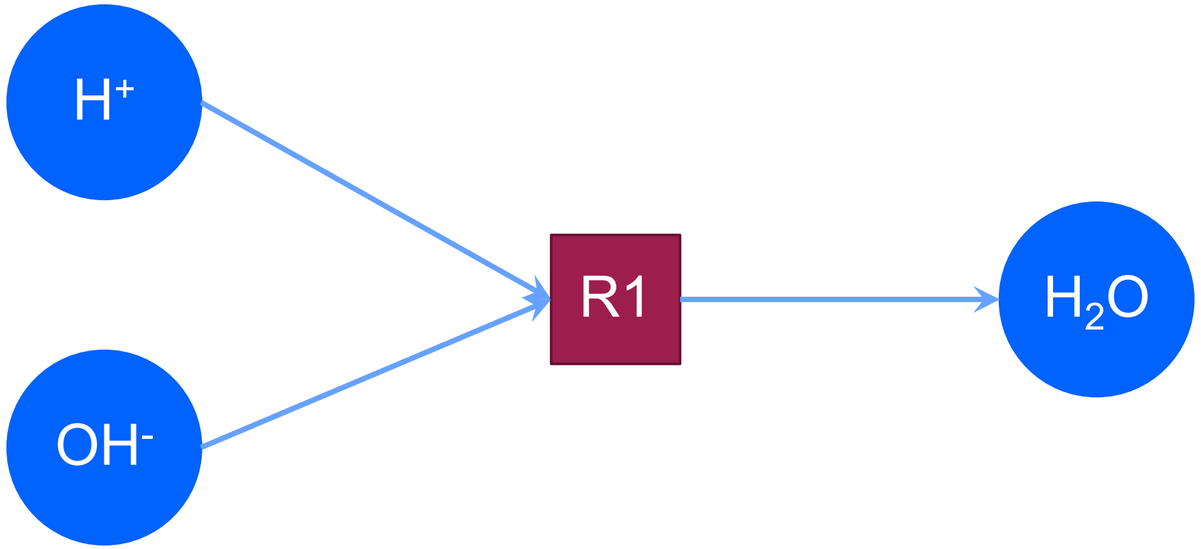
	A	B	C
A		**1**	
B	**1**		**1**
C		**1**	



Networks

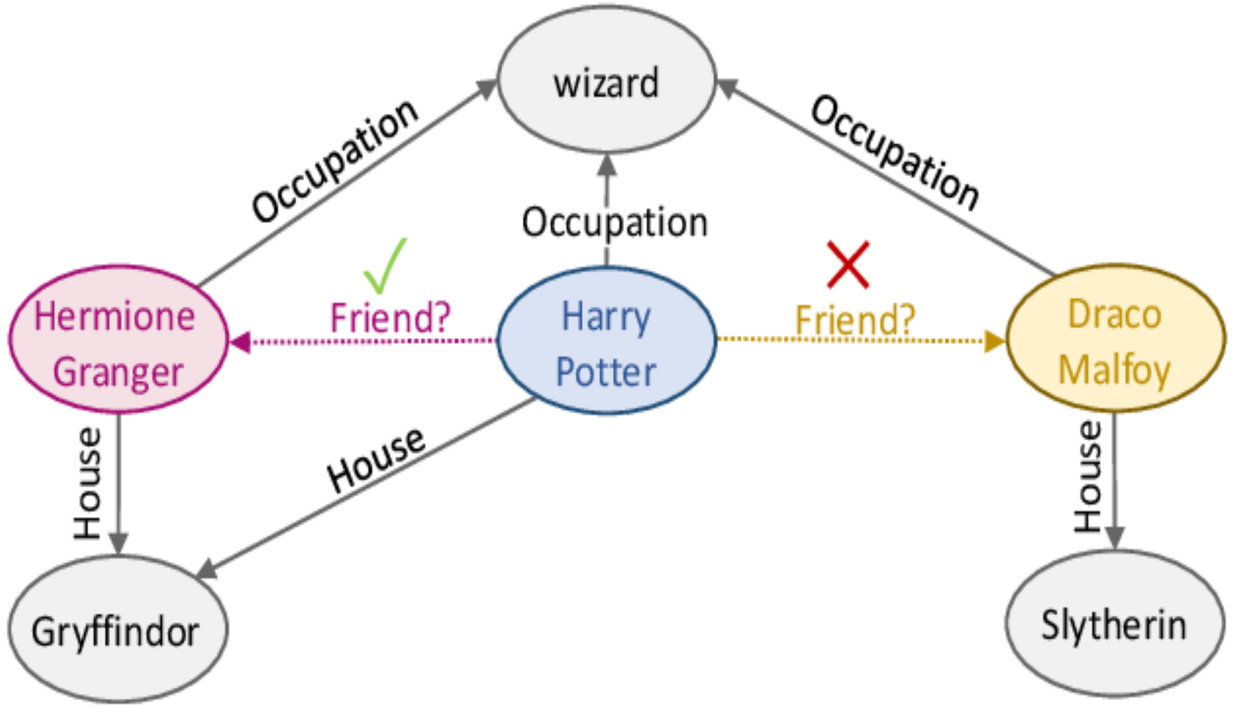
- Bi-partite networks
  - Objects (vertices) belong to two distinct classes (partitions)
  - Edges connect nodes in different partitions only
  - Some examples
    - product ratings
    - chemical reactions

User \ Item	*1984* by George Orwell	*Brave New World* by Aldous Huxley	*On the Road* by Jack Kerouac	*Of Mice and Men* by John Steinbeck
User X	✓	✓	?	✓
User Y	✓	?	✓	✓



Network completion

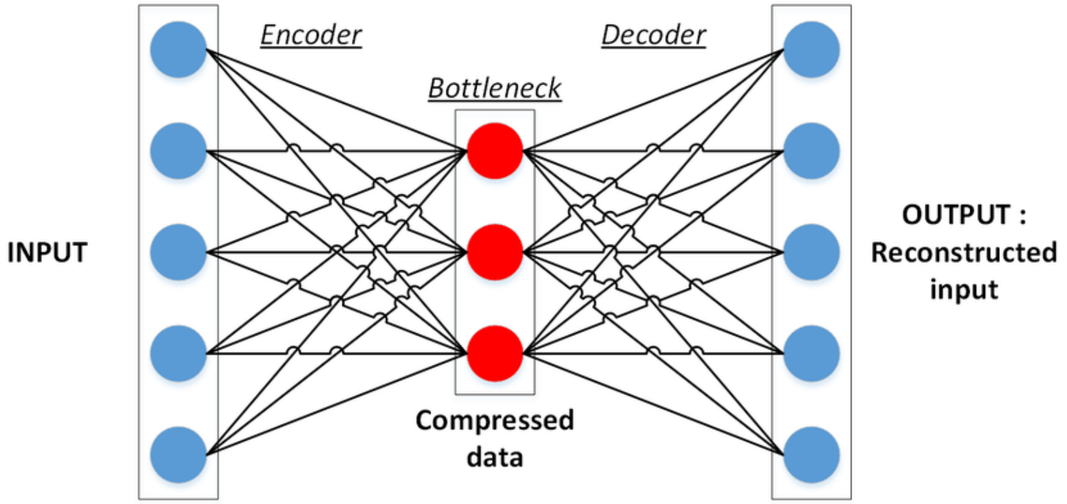
- Scientific data and knowledge is incomplete
- Networks representing available data and knowledge are incomplete, too!
  - missing nodes
  - missing links
- Network completion – inference of the missing parts of the networks
  - **Amounts to “recommendation”**
- Do we know if the “missing” link is a statement of absence of evidence or evidence of absence?
- Do we know certainty of the existing links?

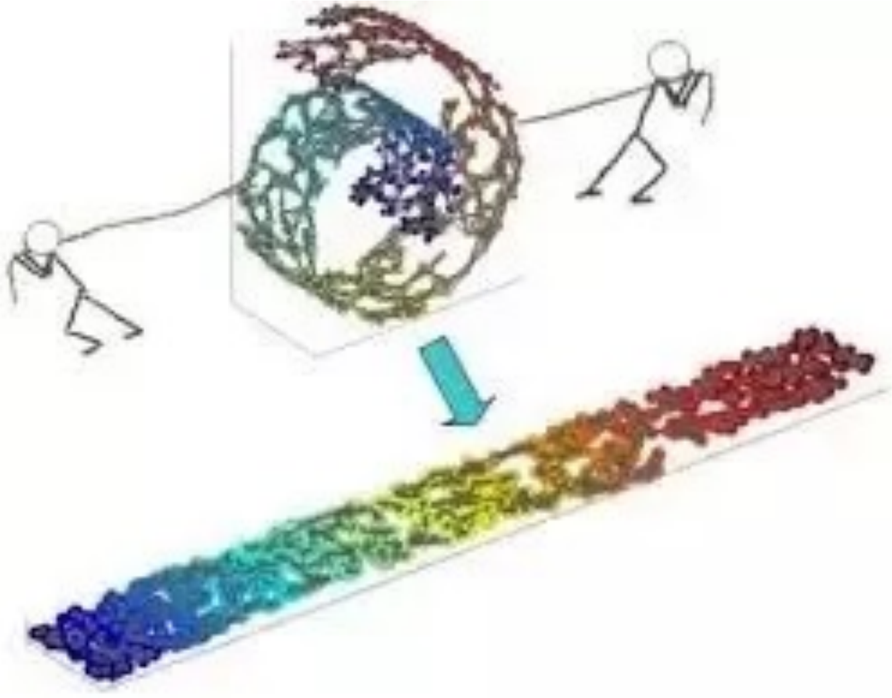


Representation learning and recommender systems

Representation learning

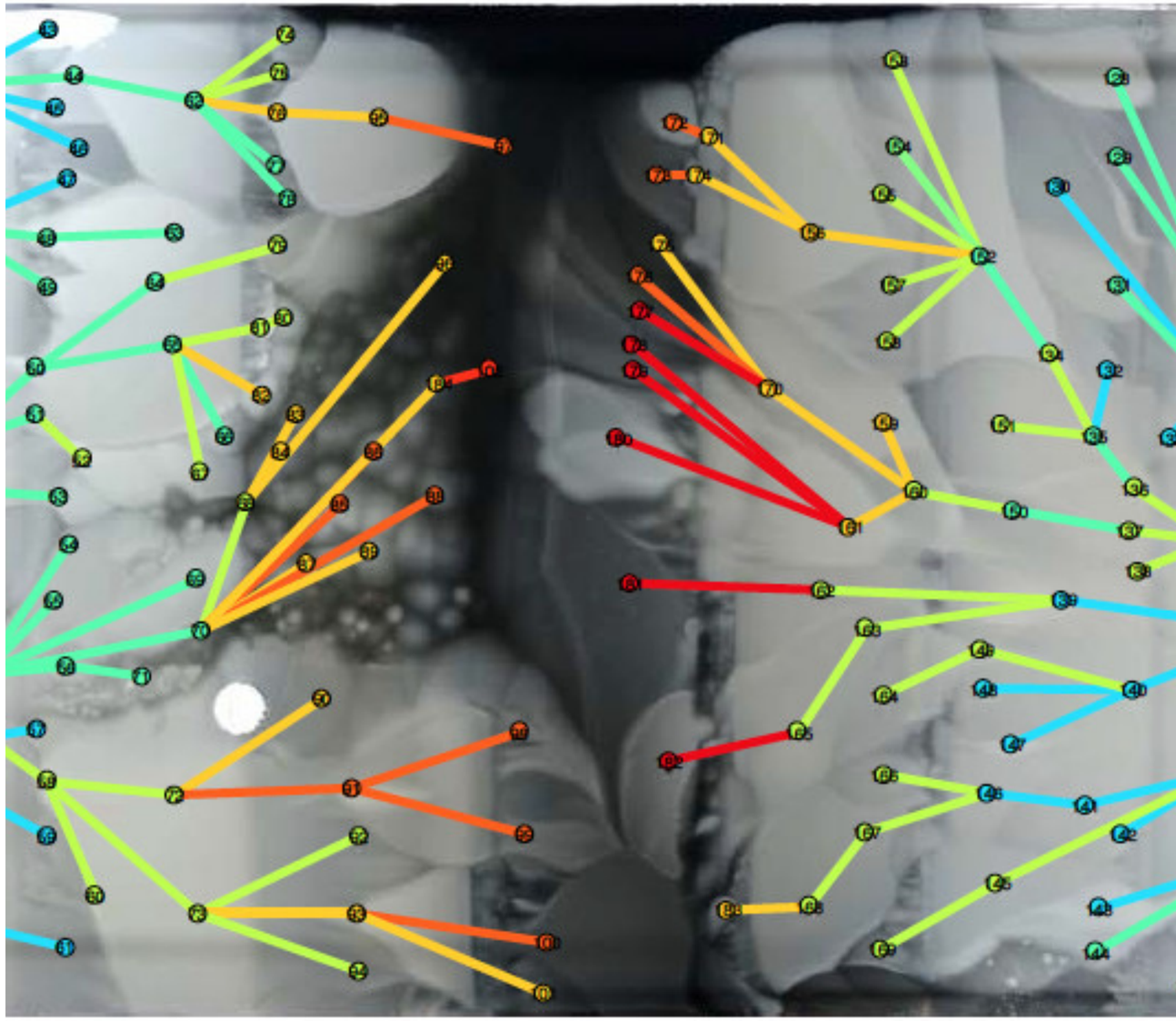
- Features are not available upfront, but learned – they are an **artifact** of the statistical learning method
- Simple ML examples where features are learned from **structural representation**
  - auto-encoders
  - manifold embedding derived from similarity or distance matrices
- Our task: learn features from the **observed relations**
  - Given a network, learn representation
  - Knowing representation, predict missing links





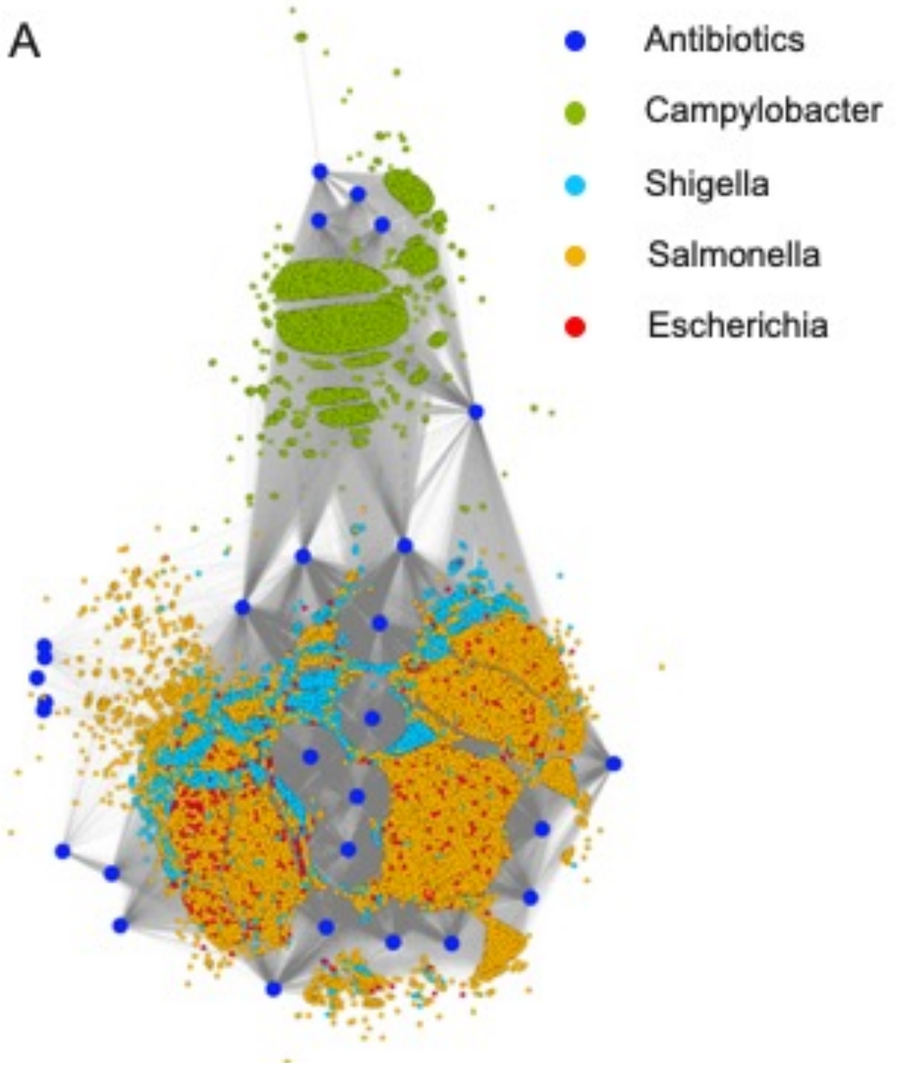
Deaths attributable to antimicrobial resistance (2014 estimates)

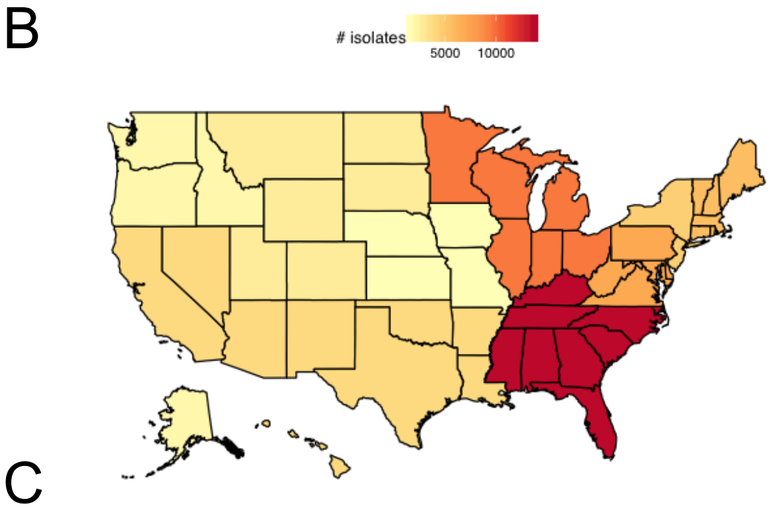
10M in 2050

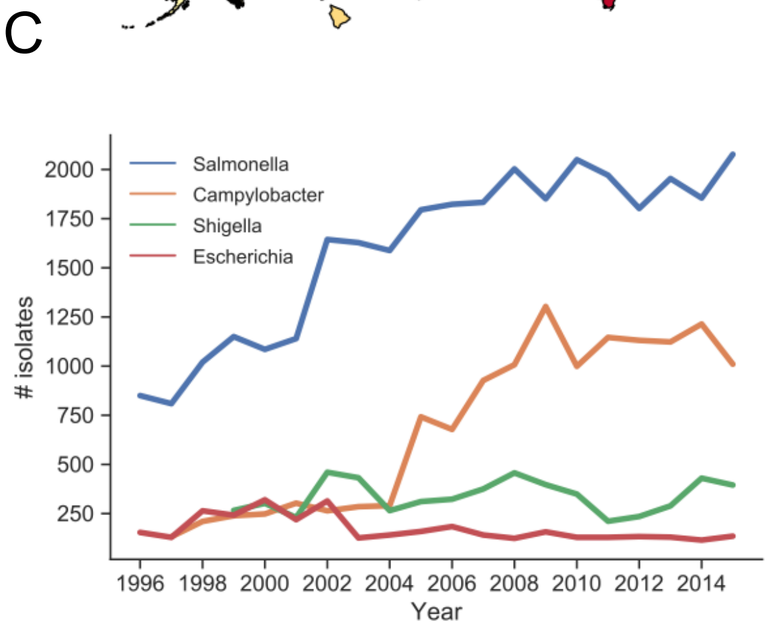


Example A: antimicrobial resistance

- Human isolate data from 1996 to 2015 (NARMS CDC)
- Metadata: classification, antibiotic susceptibility, site of isolation, year of specimen collection, region, and age
- 18 antibiotics







Matrix factorization with antibiogram

	Antibiotic 1	Antibiotic 2	Antibiotic 3
Isolate 1	**0**	**0**	**1**
Isolate 2	**1**	**1**	**1**
Isolate 3	**-1**	**0**	**0**
Isolate 4	**0**	**1**	**0**

Dataset is imbalanced

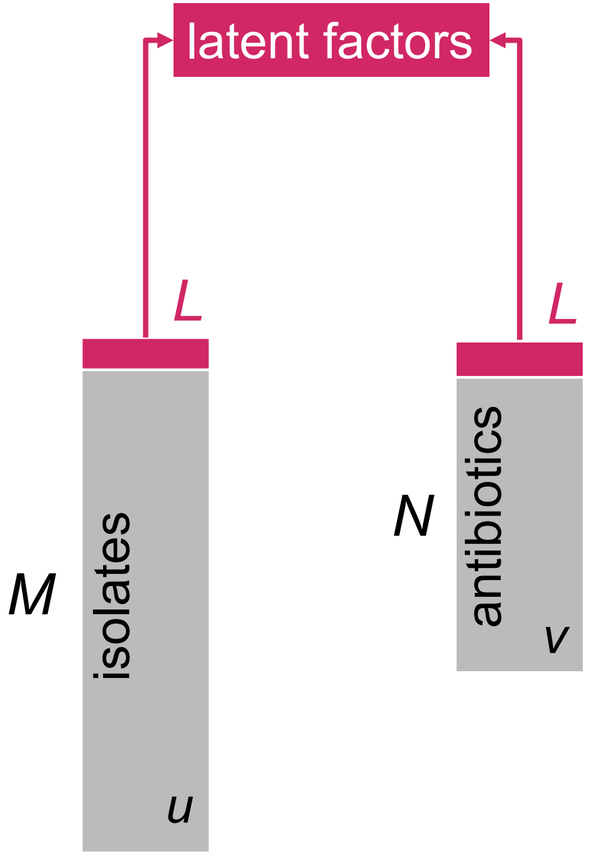
Susceptible (1): 92%

Resistant (-1): 8%

Critical to predict correctly **resistant isolates**

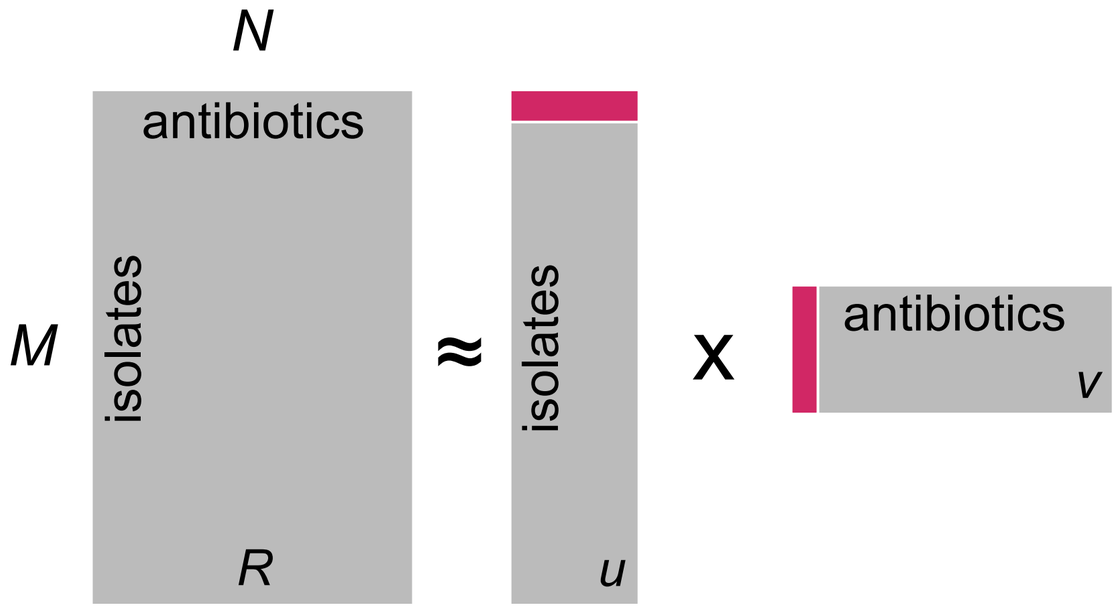
Matrix factorization with antibiogram

We assume that isolates *u* and antibiotics *v* are represented as points in the same latent space of some dimensionality (number of latent factors, *L*)



Matrix factorization with antibiogram

We approximate antibiogram as a product of the latent vectors

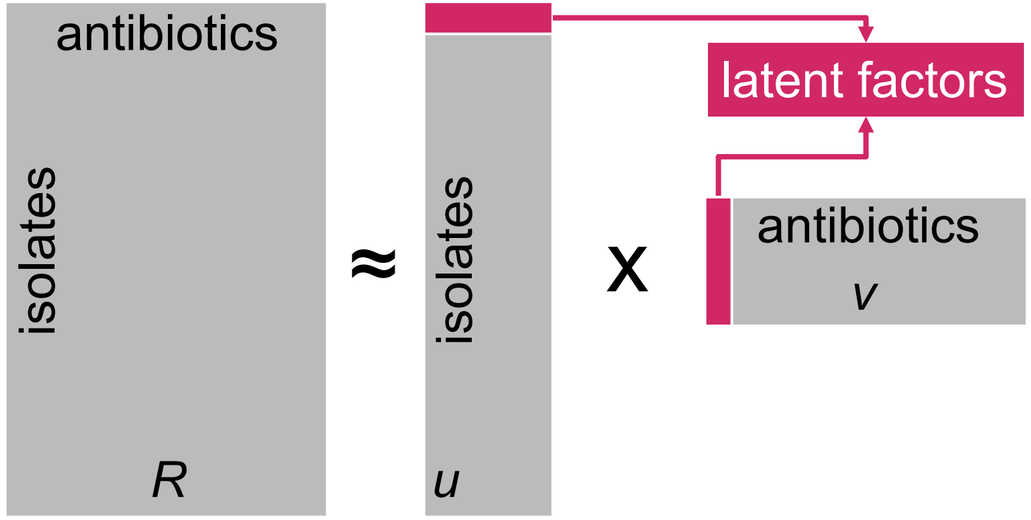


Matrix factorization with antibiogram

Finally, we use factorization algorithm (defined by constraints and loss functions) to find optimal decomposition *u* and *v*

Various ad hoc decisions might be helpful (e.g., separate recommenders for susceptible and resistant classes)

Reconstructed antibiogram contains new links!



$$R_{ij} = \begin{cases} 1, & \text{if isolate } i \text{ is susceptible against antibiotic } j \\ -1, & \text{if isolate } i \text{ is resistant against antibiotic } j \\ 0, & \text{unknown} \end{cases}$$

$$J = \sum_{i,j \in R_K} (R_{ij} - u_i^T v_j)^2 + \lambda (\|u_i\|^2 + \|v_j\|^2)$$

objective function for Alternating Least Squares

Matrix factorization with antibiogram

	Antibiotic 1	Antibiotic 2	Antibiotic 3
Isolate 1	0	0	1
Isolate 2	1	1	1
Isolate 3	-1	0	0
Isolate 4	0	1	0

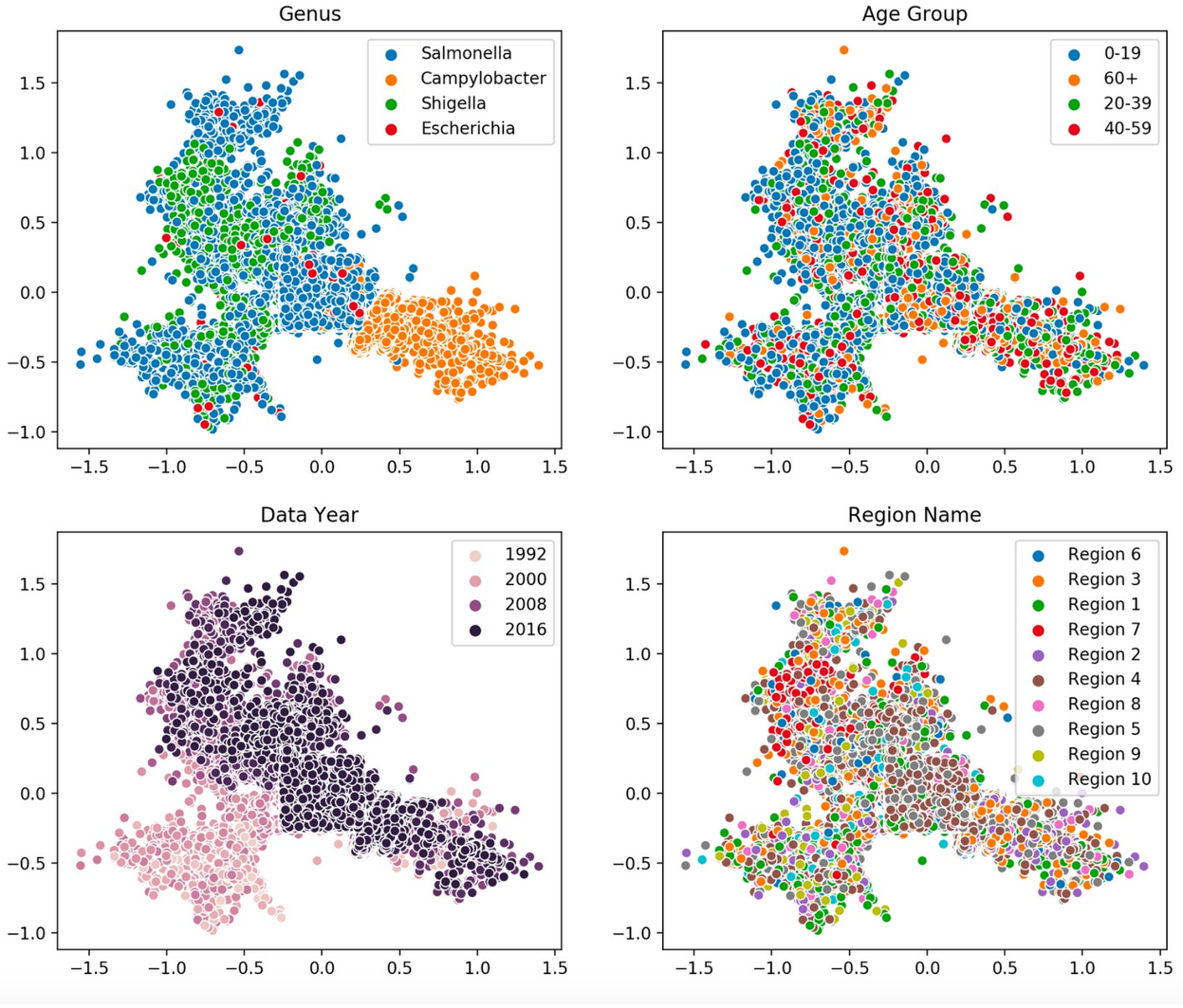
original antibiogram

	Antibiotic 1	Antibiotic 2	Antibiotic 3
Isolate 1	0.34	0.77	0.75
Isolate 2	0.93	0.99	0.91
Isolate 3	-0.7	0.68	0.02
Isolate 4	0	0.87	0.1

reconstructed antibiogram

Latent space

- Example of a 2D latent space of isolates (matrix factorization with 2 factors)
- Learned representation of the resistant class shows clustering wrt. Genus and Year, but not with Age Group and Region
  - makes sense from the evolutionary point of view
  - further interpretation of factors might benefit from genomic data, mechanism of antibiotic action, etc.



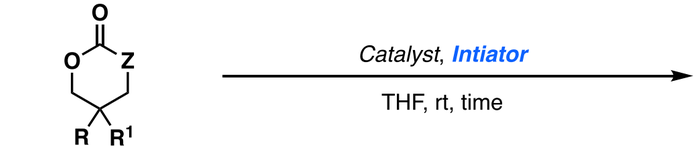
Number of experiments with 104 unique catalyst/co-catalyst and 289 monomer/initiator pairs

~30K

Example B: ring-opening polymerization experiments

- Target: monomer conversion and dispersity
- Predict combination of monomer and catalyst
- Model of ”outcome - experimental parameters” relations to achieve “embarrassingly actionable” hypotheses

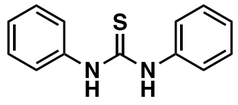
ROP of Cyclic Carbonates and Lactones



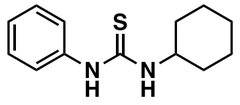
Cyclic Monomers
Z = C, O

Polyesters or Polycarbonates

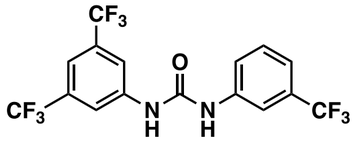
Catalysts



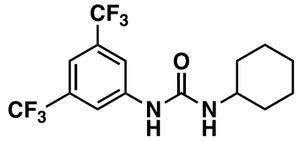
TU-(Ph)$_2$



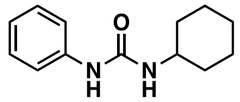
TU-Ph-Cy



U-3-CF$_3$

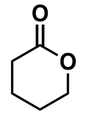


U-2-CF$_3$-Cy

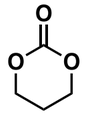


U-Ph-Cy

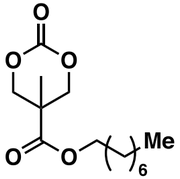
Monomers



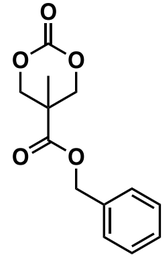
VL



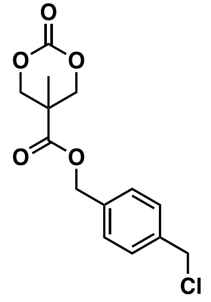
TMC



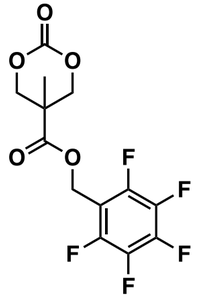
TMC-Octyl



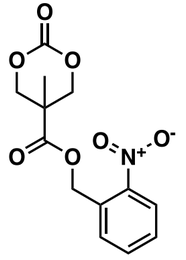
TMC-Bn



TMC-BnCl



TMC-BnF$_5$

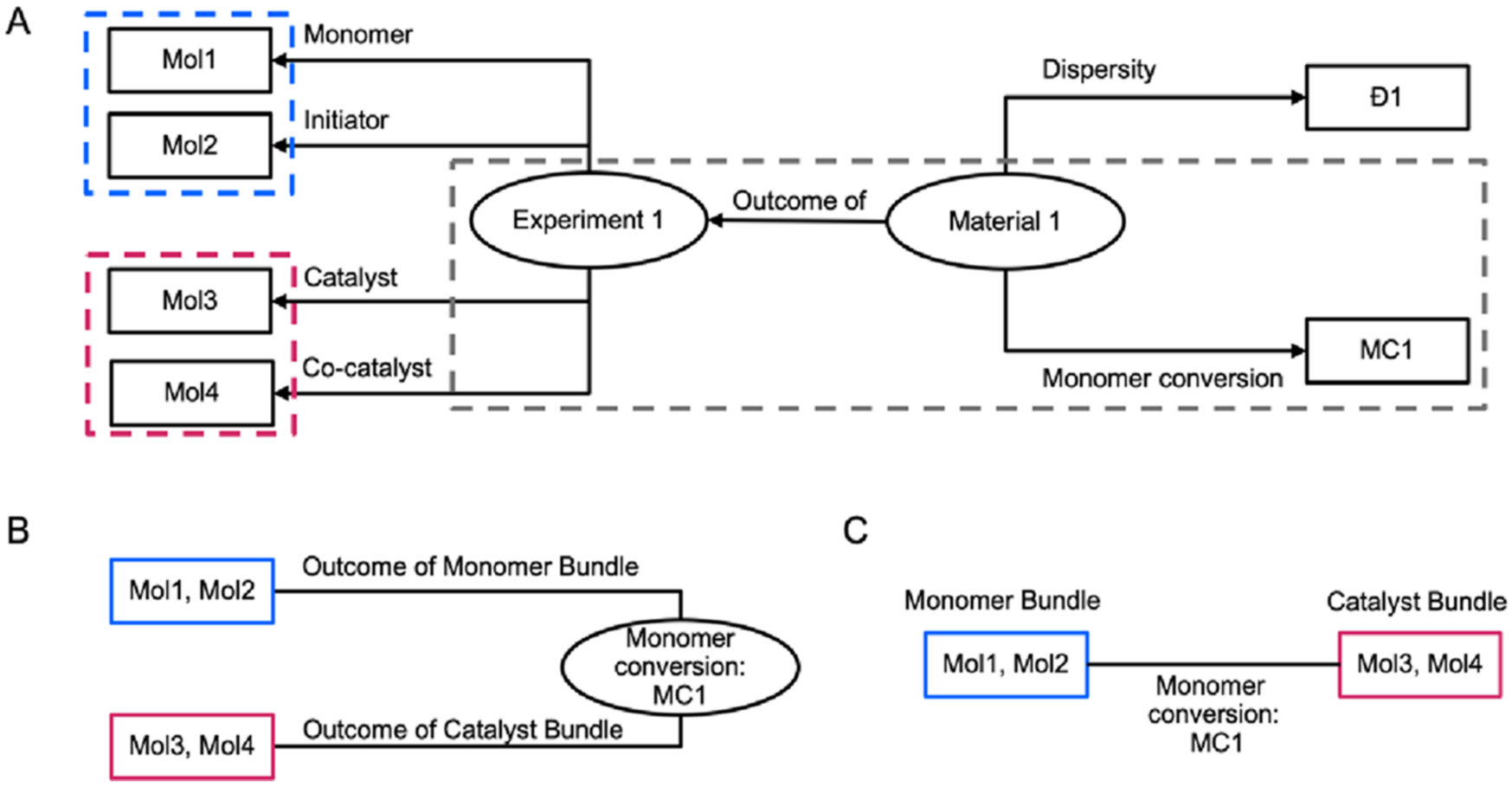


TMC-NitroBn

Experiment knowledge graph

We start with the network (A) that captures experiment specification and collapse some of the relations.

Bi-partite network (B) is further simplified to construct a simple network (C) where monomer/initiator and catalyst/co-catalyst pairs are nodes, and experiments are links

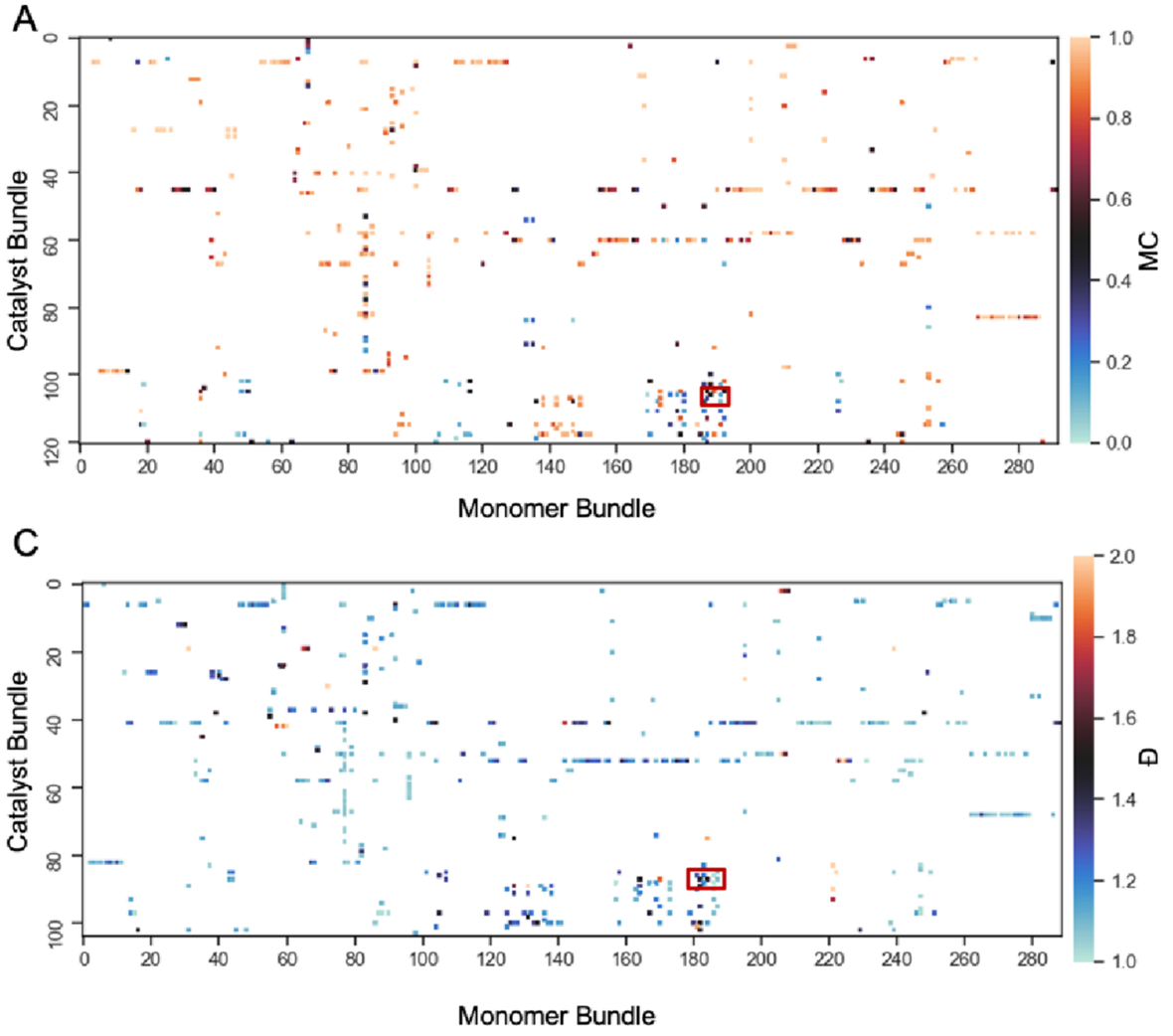


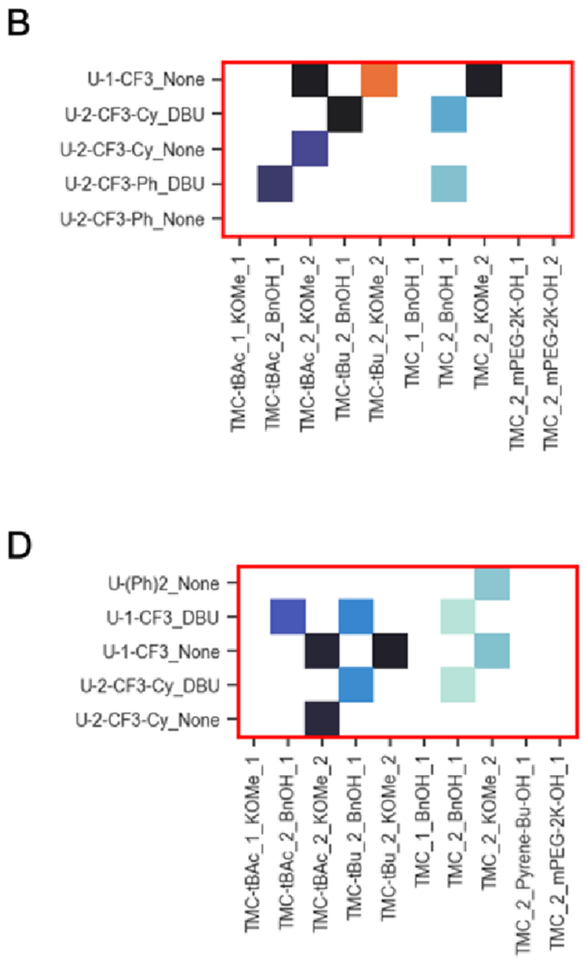
Setting up network completion

Two partitions: “catalyst” and “monomer”

Two recommendation problems: monomer conversion and dispersion

Sparse coverage of the experiment design space – most experiments are new





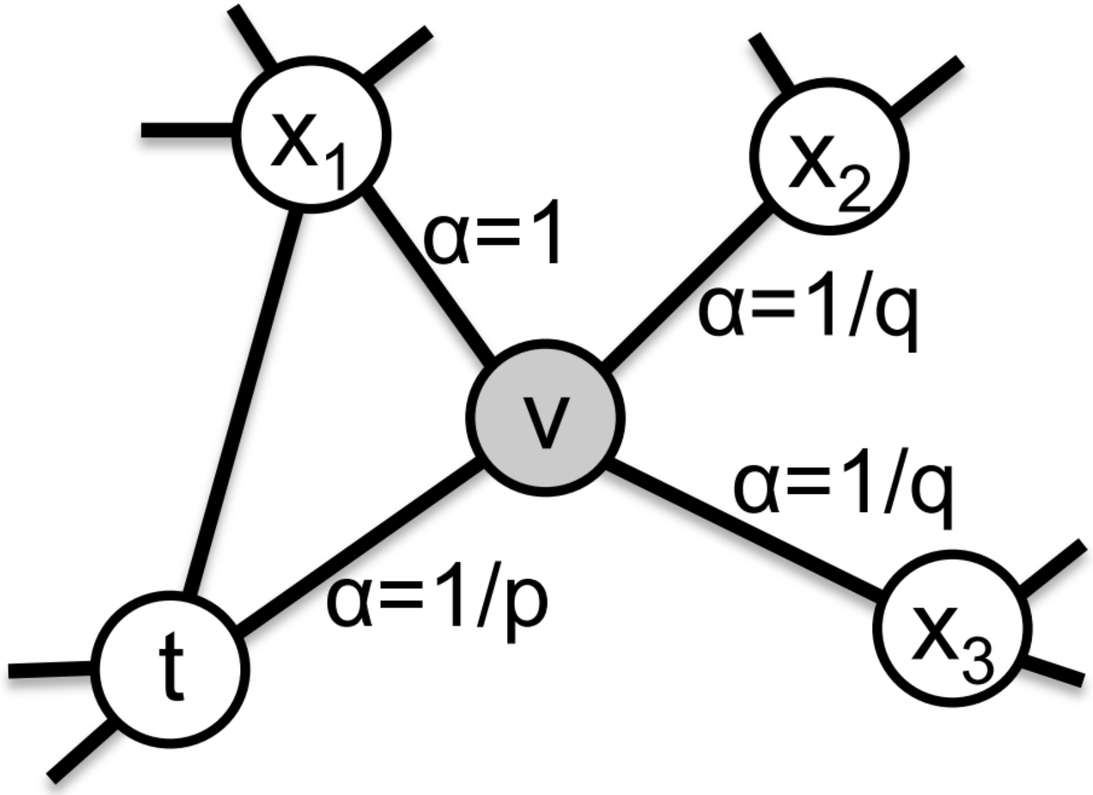
Semantic network embedding
node2vec

t – previous node

v – current node

x_i – possible next node

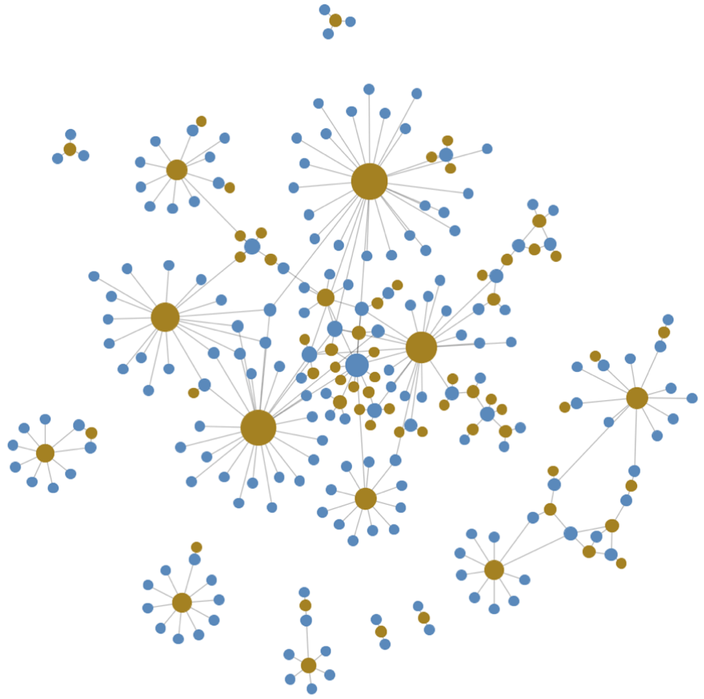
d_{tx} - shortest path
distance from *t* to *x*



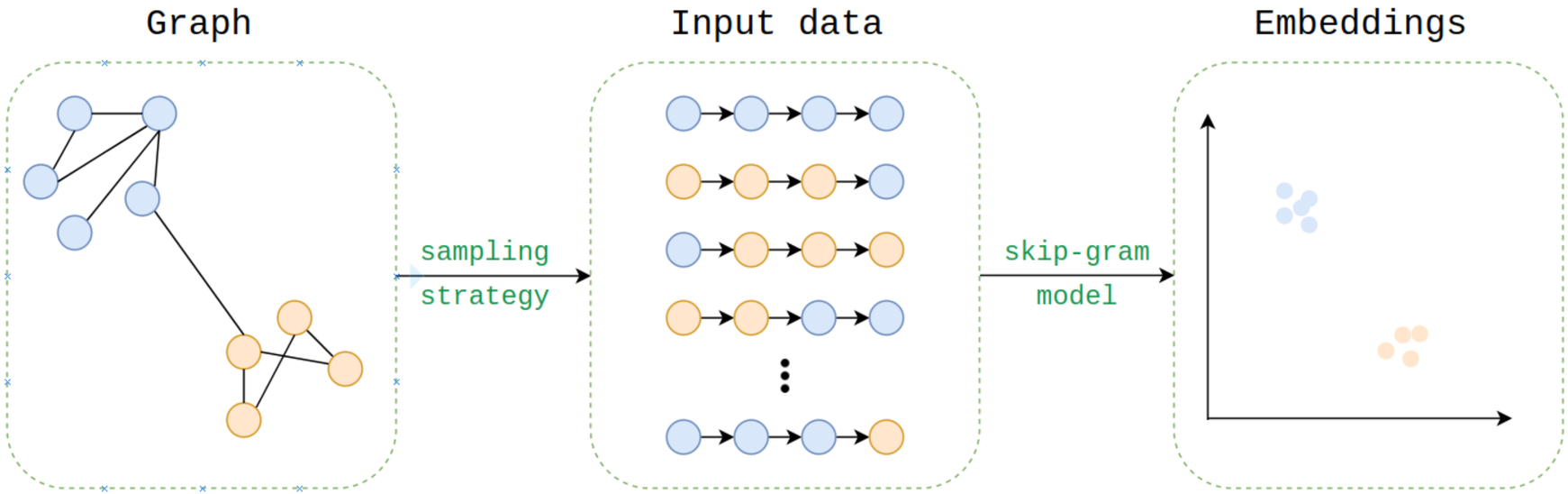
Search bias (breadth-first/depth-first tuning):

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

Semantic network embedding
node2vec



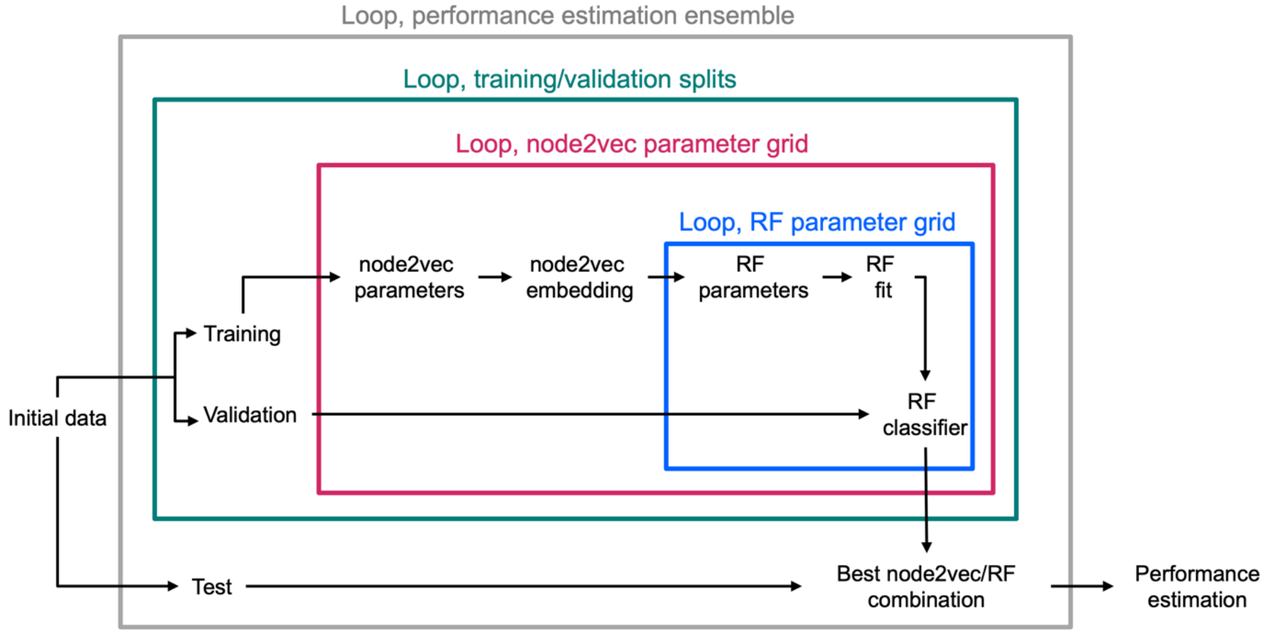
Network of historical experiments with target monomer conversion



node2vec process

Semantic network embedding *node2vec*

- *node2vec* produces embeddings of the nodes
- embeddings of the links are obtained from the node embeddings (e.g., element-wise product)
- embeddings of known links are used to train a classifier to infer probabilities of the new links



Takeaway

- It is straightforward to construct similarity networks (known features and defined distances)
- Networks defined by relations (especially those that we want to model) have special utility – they lend themselves to **learning features relevant to the problem**
- Representation learning on networks naturally extends to network completion – **prediction of missing links** (relations)
- Variety of algorithms is available from linear (matrix factorization) to non-linear (deep learning, node2vec)