Deep learning for drug discovery

Wengong Jin

Massachusetts Institute of Technology

Drug discovery is a time-consuming process

Average time/cost for designing one drug = 10 years + $2.6B

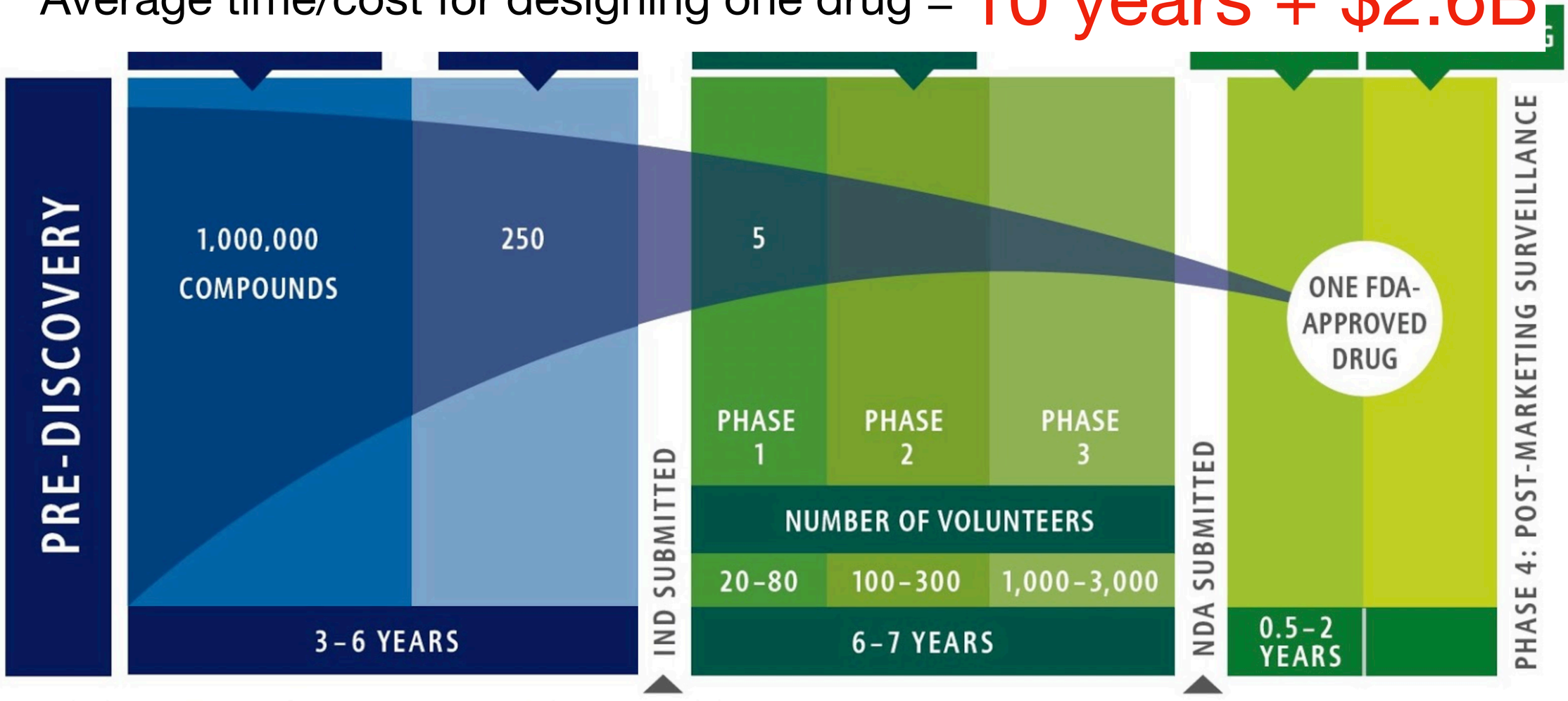
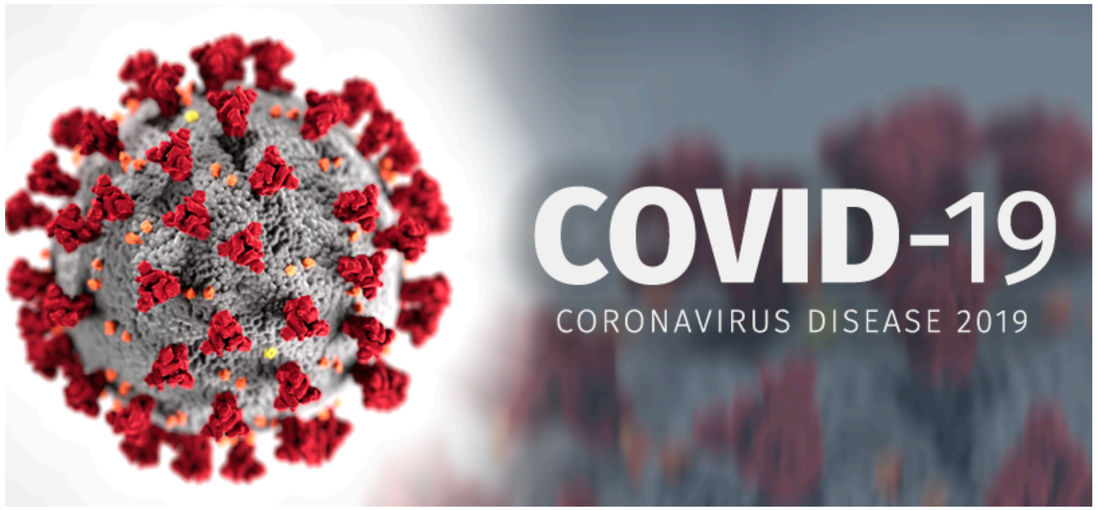


Figure source: Pharmaceutical Research and Manufacturers of America

Obviously, we can’t wait for 10 years…



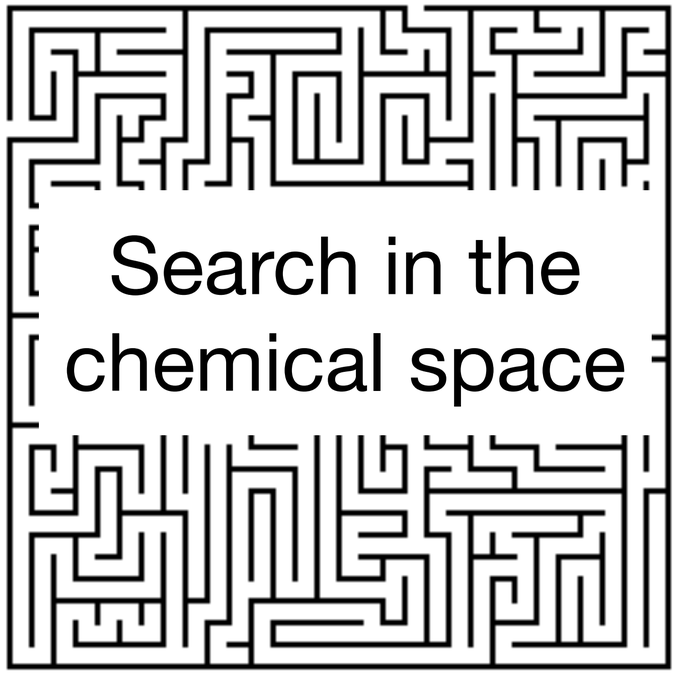
United States

Total cases	Recovered	Deaths
27.5M	-	481K
+99,565		+5,463

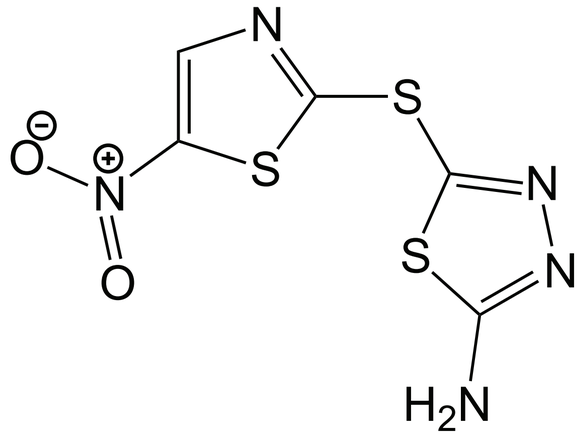
Worldwide

Total cases	Recovered	Deaths
108M	60.6M	2.38M

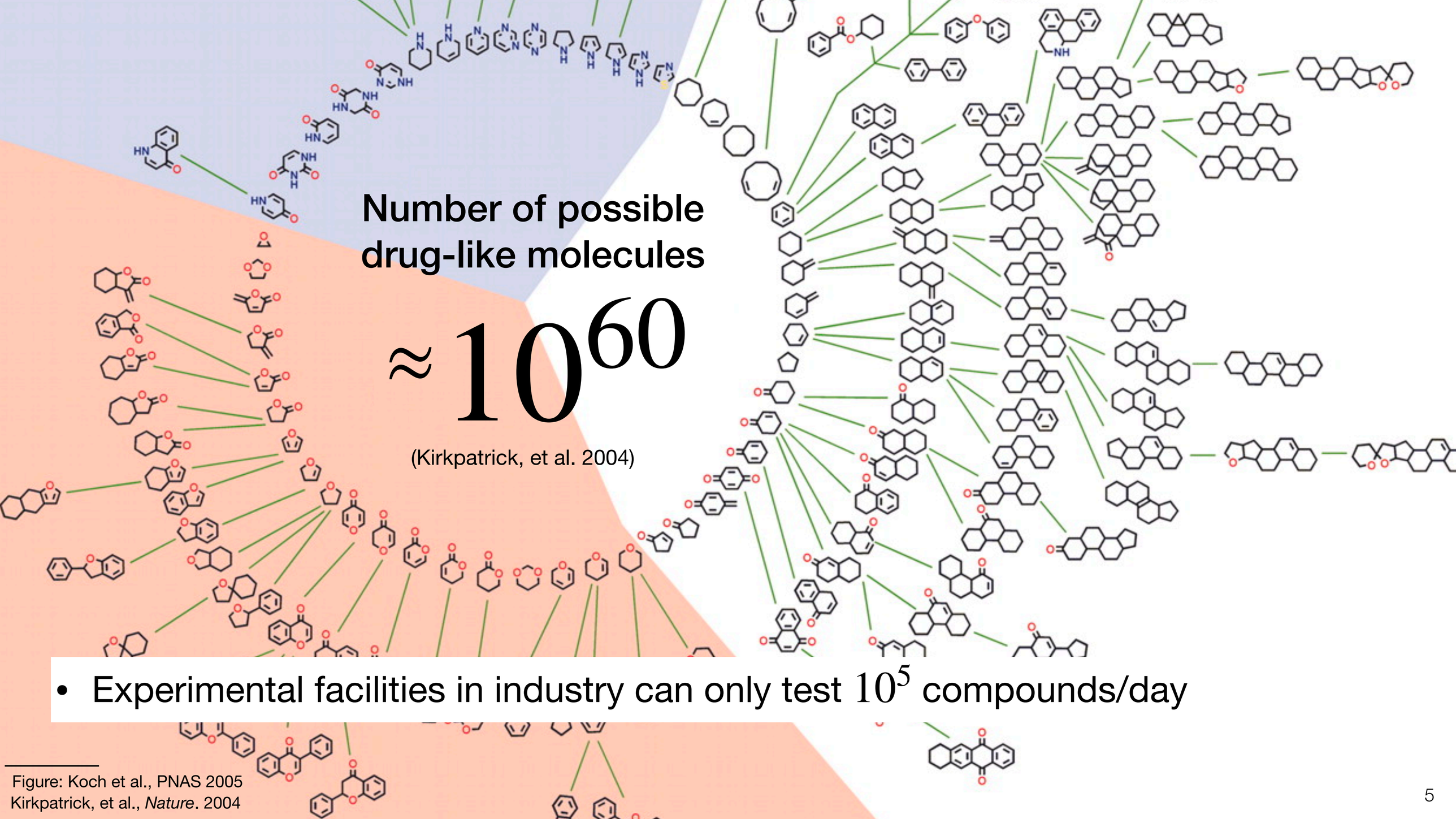
Drug discovery is a challenging search problem



A good drug (e.g., kills virus)



Data source: PhRMA.org



Number of possible drug-like molecules

$$\approx 10^{60}$$

(Kirkpatrick, et al. 2004)

- Experimental facilities in industry can only test 10^5 compounds/day

Figure: Koch et al., PNAS 2005
Kirkpatrick, et al., *Nature*. 2004

Automate drug discovery with computation

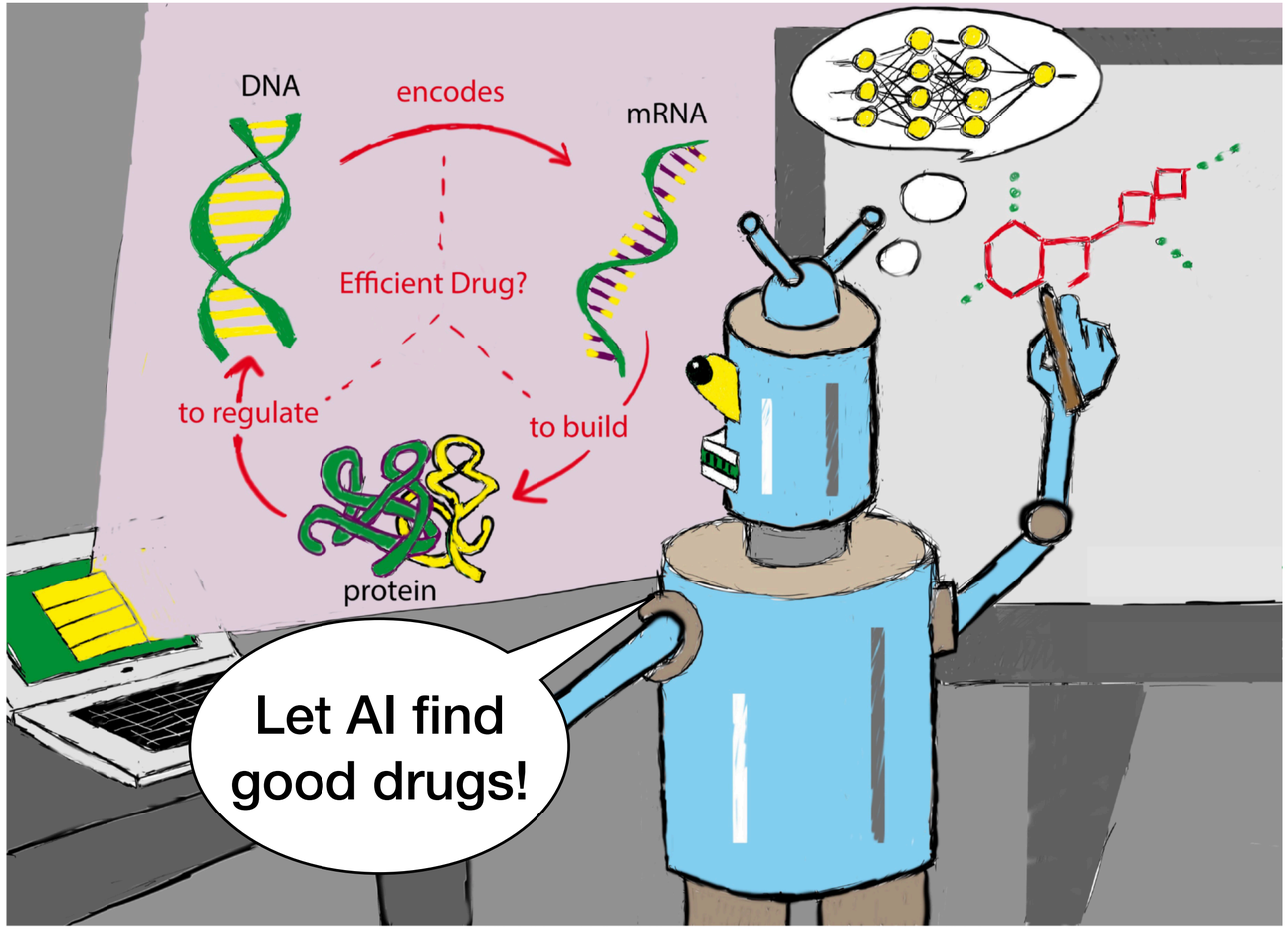
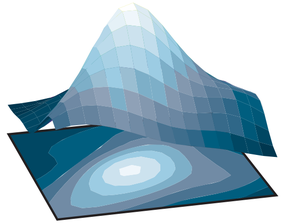


Figure source: Andrii Buvailo

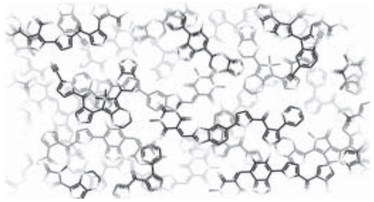
Computational drug discovery: three schemes

Functional space



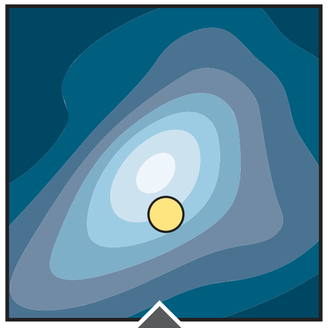
Desired properties (redox potential, solubility, toxicity)

Chemical space

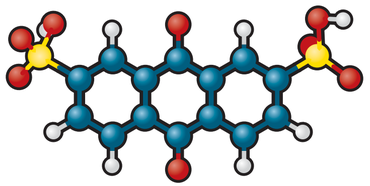


(Drug-like, photovoltaics, polymers, dyes)

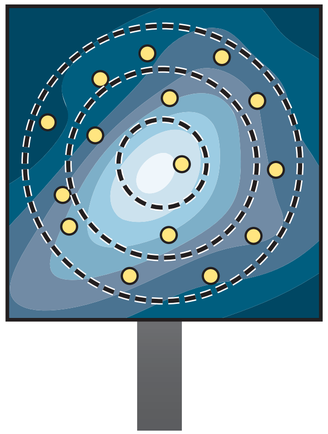
Simulation



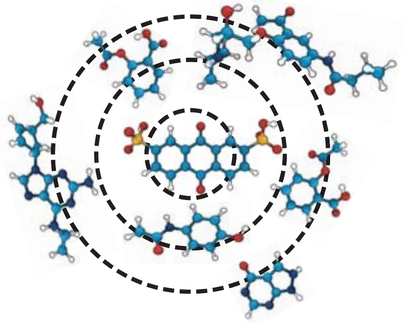
Experiment or simulation (Schrödinger equation)



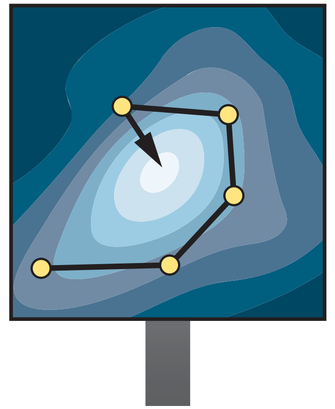
Virtual screening



High-throughput virtual screening (e.g., with 3 filtering stages)



De novo drug design



Optimization, evolutionary strategies, generative models (VAE, GAN, RL)

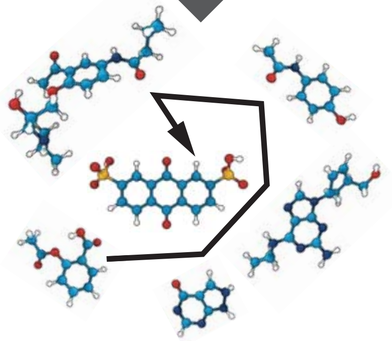


Figure source: Sanchez-Lengeling et al., Science 361, 360–365 (2018)

Simulation is often too slow

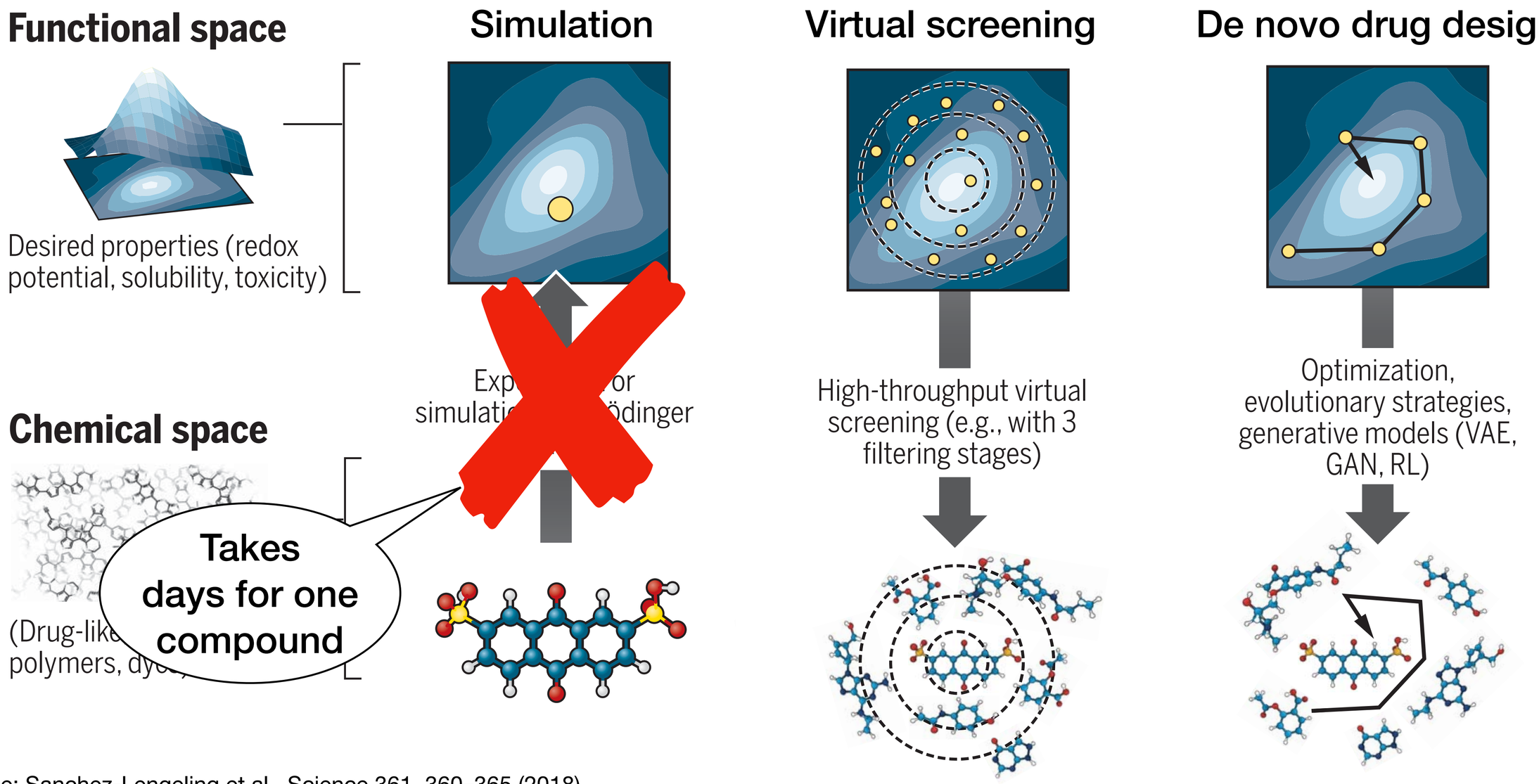
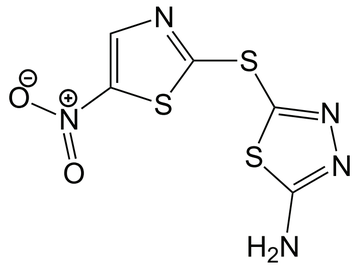


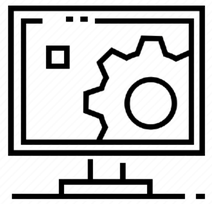
Figure source: Sanchez-Lengeling et al., Science 361, 360–365 (2018)

Virtual screening

- **Virtual screening:** assess whether a compound is a good drug using computation models (Walters et al., 1998; McGregor et al., 2007; …)



Compound



Virtual screening model

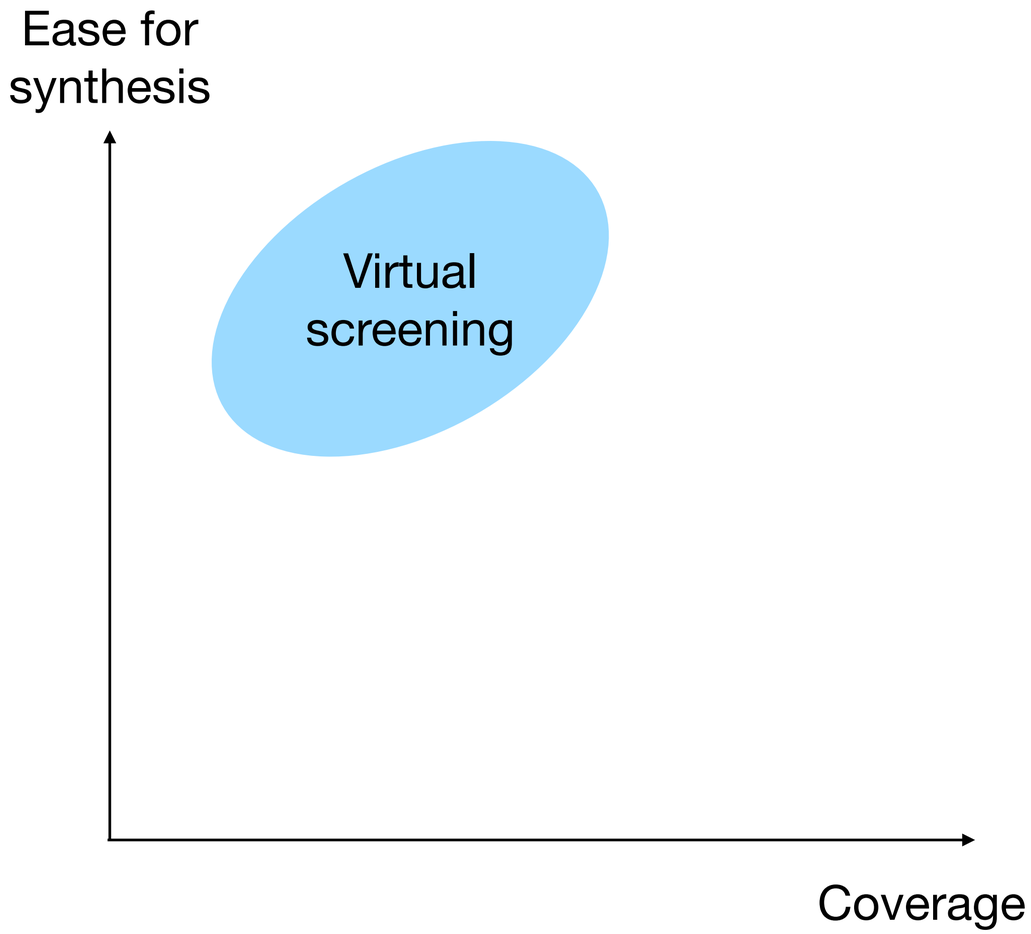
Prediction: good!

Experiments

- Virtual screening is much faster than experimental screening in web labs.
- It can test 10^8 compounds within a day, while experimental screening takes years
- It is also much cheaper than experimental screening

Virtual screening: inherent trade-off

- Virtual screening is restricted to commercially available compounds (e.g., ZINC library)
- Advantage: no need to synthesize any compounds (faster testing)
- Limitation 1: it loses coverage — at best, we can screen 10^9 compounds
- Limitation 2: traditional techniques are based on hand-crafted features

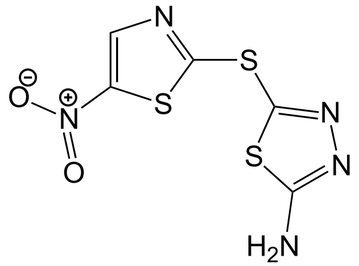


De novo drug design

- **De novo drug design**: directly generate a compound with desired properties (Moon et al., 1991; Clark et al., 1995; Schneider & Fechner, 2005; …)

Property criteria (potency, safety, …)

Drug design model

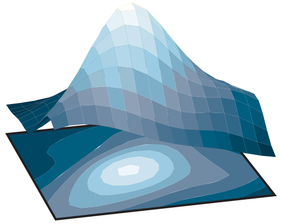


A good drug

Experiments

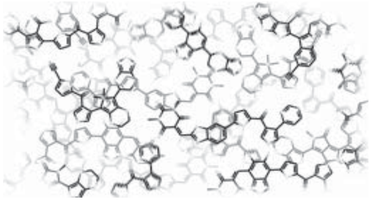
We need to solve an inverse problem

Functional space



Desired properties (redox potential, solubility, toxicity)

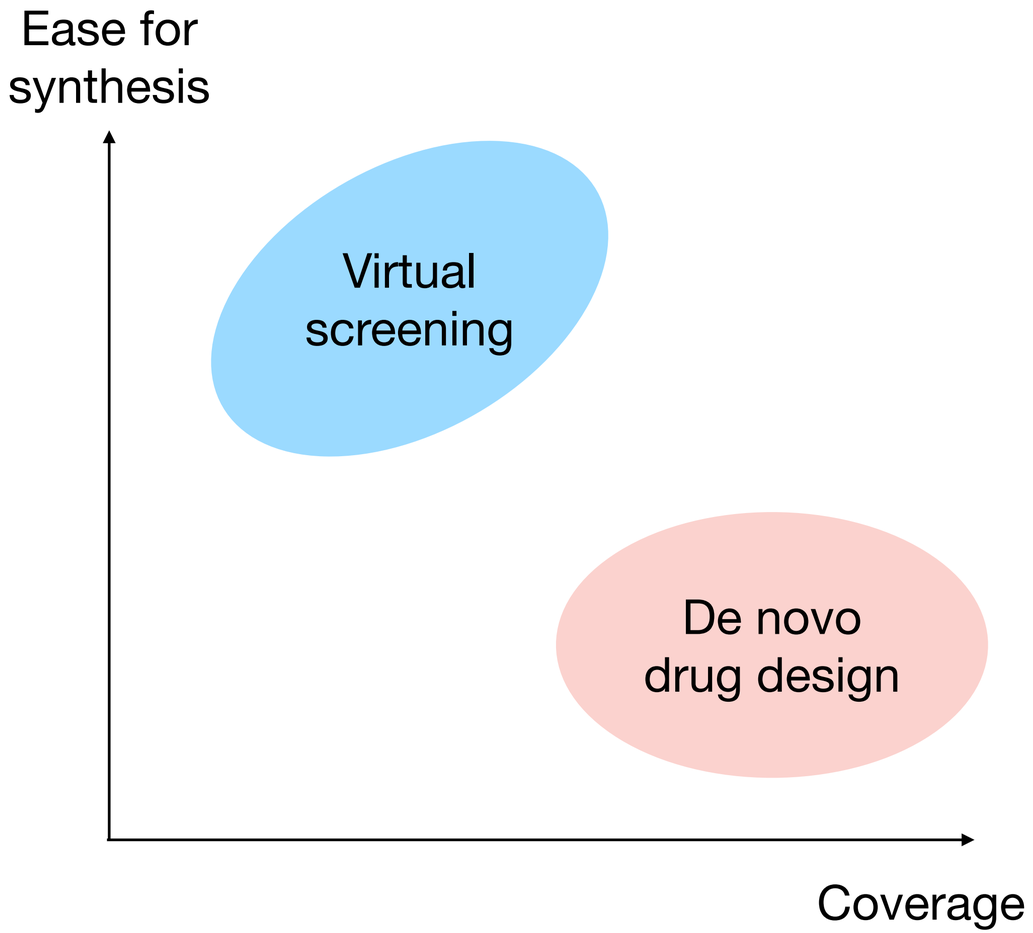
Chemical space



(Drug-like, photovoltaic polymers, dyes)

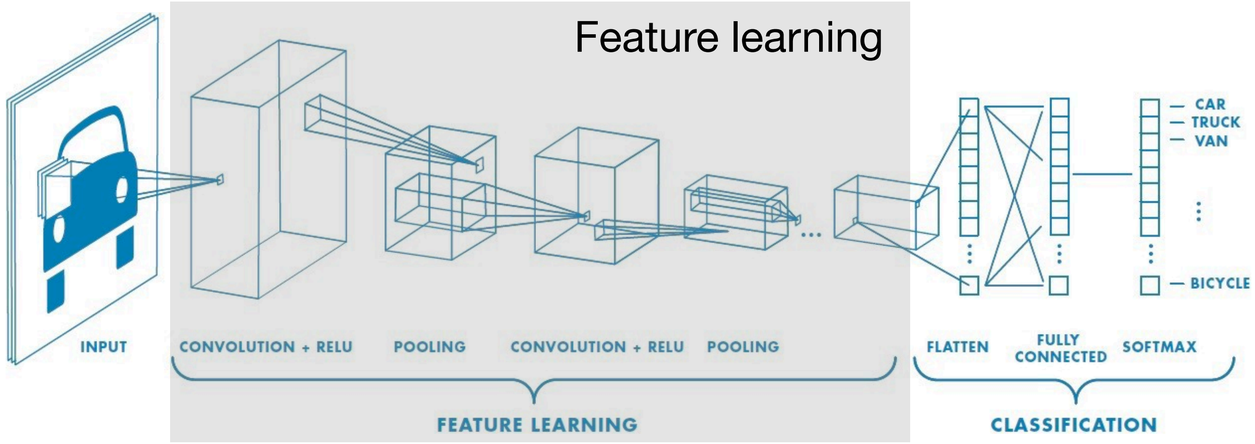
De novo drug design: inherent trade-off

- Virtual screening is restricted to commercially available compounds (e.g., ZINC library)
- Advantage: can explore the entire chemical space efficiently
- Limitation 1: we need to synthesize new compounds, which can be hard
- Limitation 2: traditional techniques explores the space based on hand-designed rules (e.g., genetic algorithms)



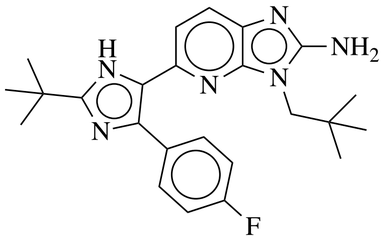
Deep learning: a promising direction

- Deep learning has achieved human-level accuracy in computer vision (He et al., 2016)



The key to success:
automatic feature learning

- **Virtual screening:** traditional methods are based on hand-crafted features



Use deep learning to learn features automatically

Prediction: good!

He et al., “Deep residual learning for image recognition." *CVPR 2016*

Deep learning: a promising direction

- Deep generative models can generate realistic text and images with desired properties

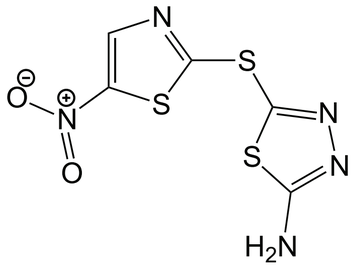
Ramesh et al., 2020

Generate an image of an armchair in the shape of avocado

- **De novo drug design**: generate a compound with desired properties

Property criteria (potency, safety, …)

Use deep generative models



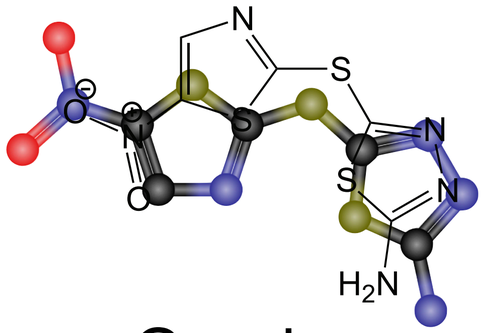
A good drug

Silver et al., “Mastering the game of Go with deep neural networks and tree search”, *Nature* (2016).
Ramesh et al., “DALL-E: creating images from text ”, *OpenAI blog*

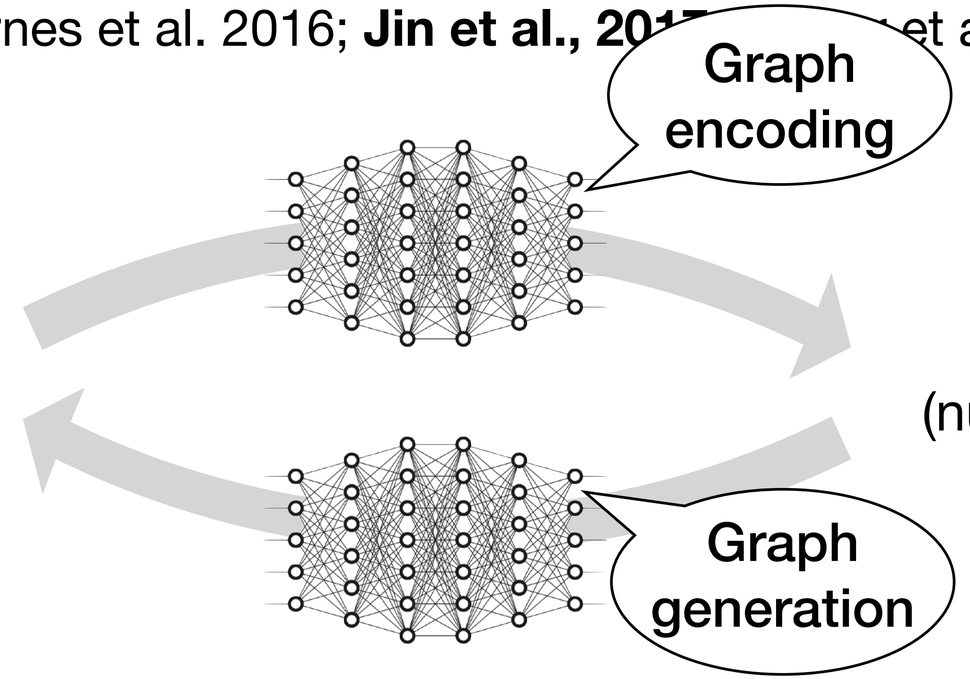
Main technique: graph neural networks

Virtual screening / molecular property prediction

(Duvenaud et al. 2015; Kearnes et al. 2016; **Jin et al., 201** et al., 2017; …)



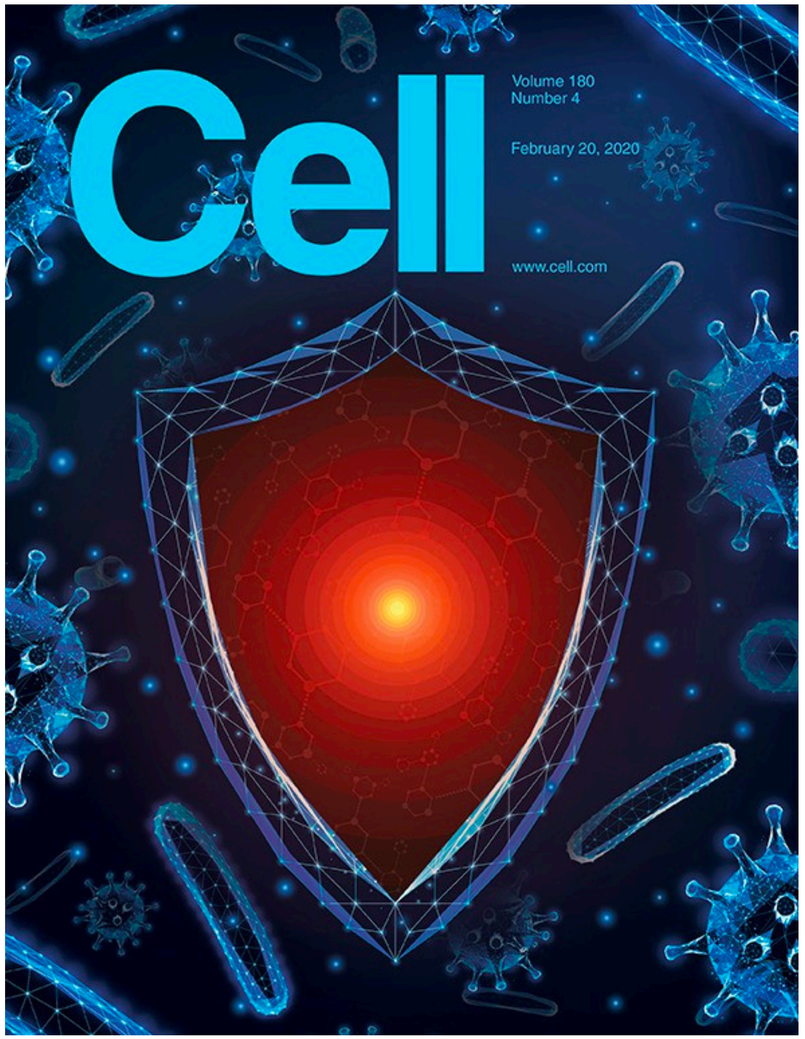
Graphs



De novo drug design

(Olivecrona et al., 2018; Gomez-bombarelli et al., 2018; **Jin et al., 2018**; Popova et al., 2018; …)

Example: discovery of new antibiotics



Powerful antibiotics discovered using AI

Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

nature

Powerful antibiotic discovered using machine learning for first time

Team at MIT says halicin kills some of the world's most dangerous strains

The Guardian

NEWS

Scientists discover powerful antibiotic using AI

BBC

Stokes, Yang, Swanson, Jin et al, *Cell* 2020

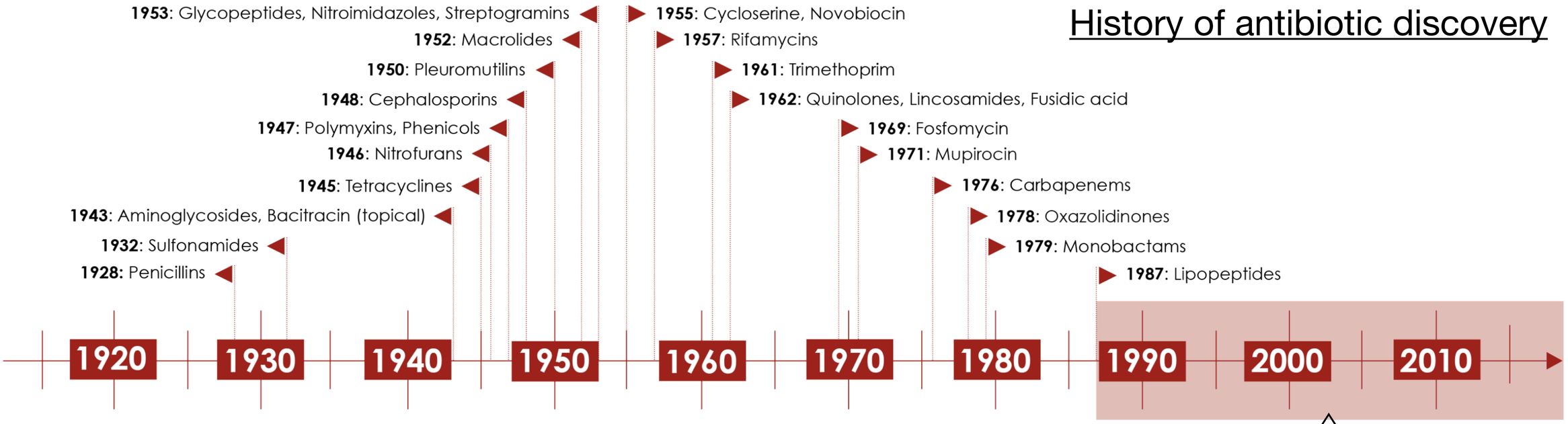
Outline of today's lecture

- **Part 1: graph neural networks for antibiotic discovery**
  [ICML'17, NeurIPS'17, JCIM'19, Cell'20]

- **Part 2: Incorporate biological knowledge into graph neural networks: application to COVID-19 drug combination discovery**
  [PNAS (In submission)]

- **Part 3: Generative models for de novo drug design**
  [ICML'18, ICLR'19, ICML'20a,b,c]

Part 1: antibiotic discovery



- After 1990s, we struggle to discover novel antibiotic classes (Silver et al., 2011; Brown et al., 2014; Shore & Coukell, 2016)
- We need novel antibiotic classes due to antibiotic resistance

Figure source: ReAct group

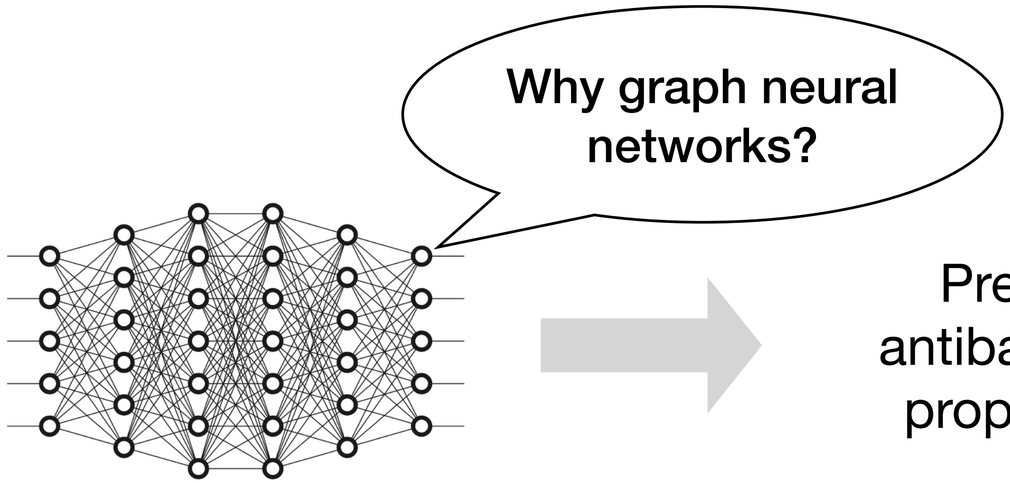
FDA = U.S. Food and Drug Administration

Virtual screening for antibiotic discovery

- Through collaboration with the Broad Institute, we collected 2560 molecules with measured growth inhibition against E. coli (BW25113)

Drug	Antibacterial
Nitrocefin	Yes
Reserpine	No
Penicillin	Yes
IQ-1S	No
.....	.....

Training data

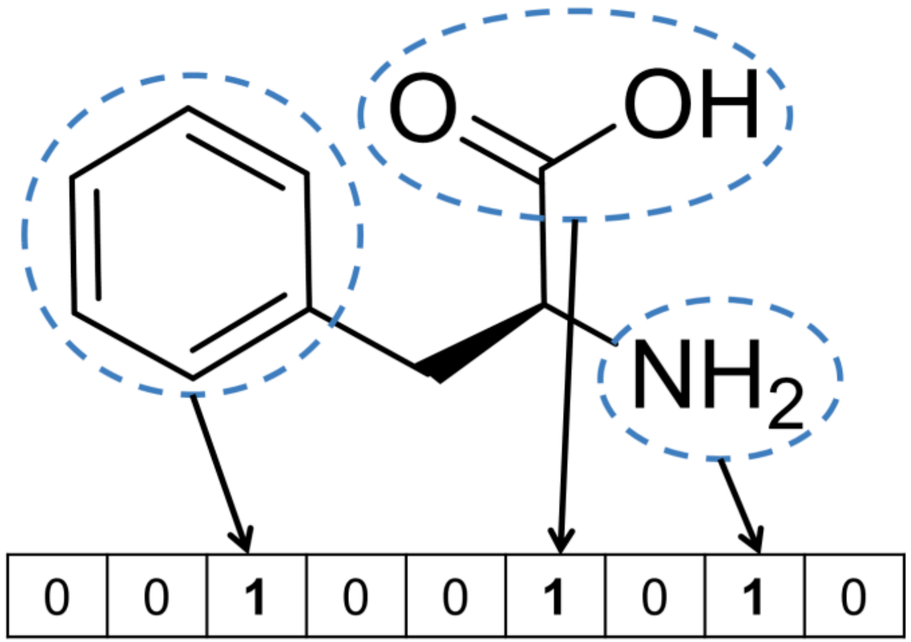


Graph neural network

Predict antibacterial properties

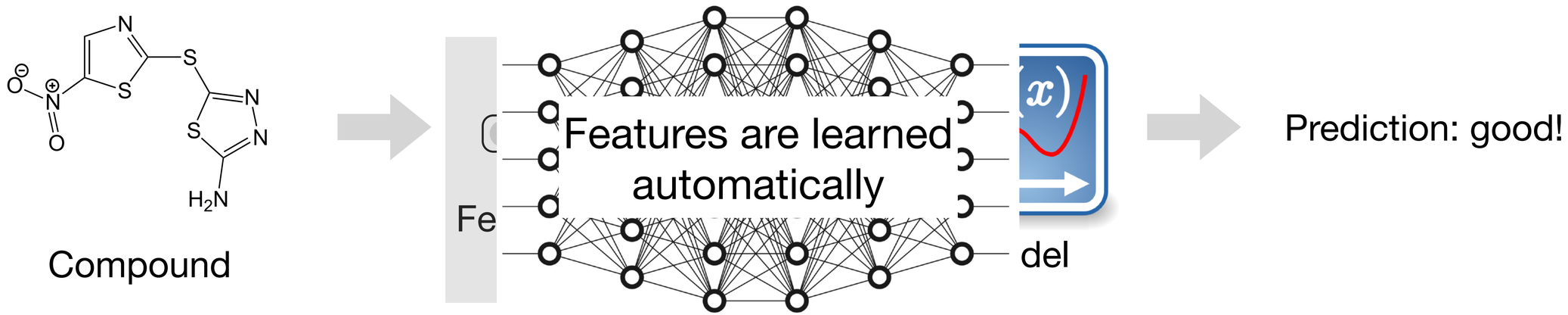
Traditional approach: hand-crafted features

- Traditional methods are based on fixed, hand-engineered molecular features.
- Molecular weight, number of heavy atoms
- More sophisticated features: Morgan fingerprint (Rogers & Hahn 2010)
- Exhaustive enumeration of all possible substructures, up to radius 3
- Result: high dimensional features (2048), different substructures merged by hash



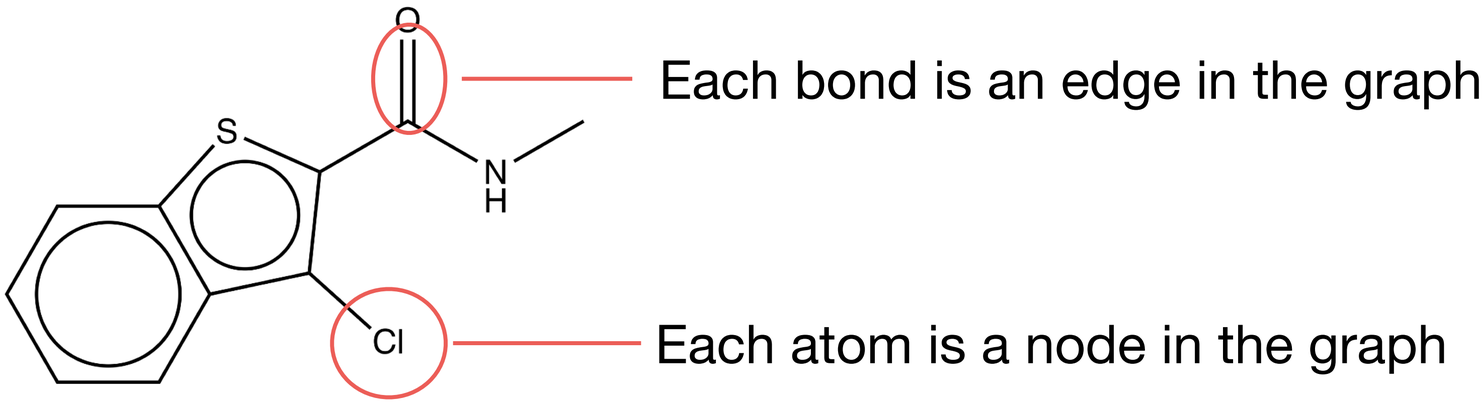
Problem of traditional features

- Traditional methods are based on fixed, hand-engineered molecular features.
  - Molecular weight, number of heavy atoms, etc.
- **Problem**: we don't know all the antibacterial patterns
  - So these hand-engineered features can **miss** some of the unknown patterns
- Graph neural networks automatically learn a feature representation from data

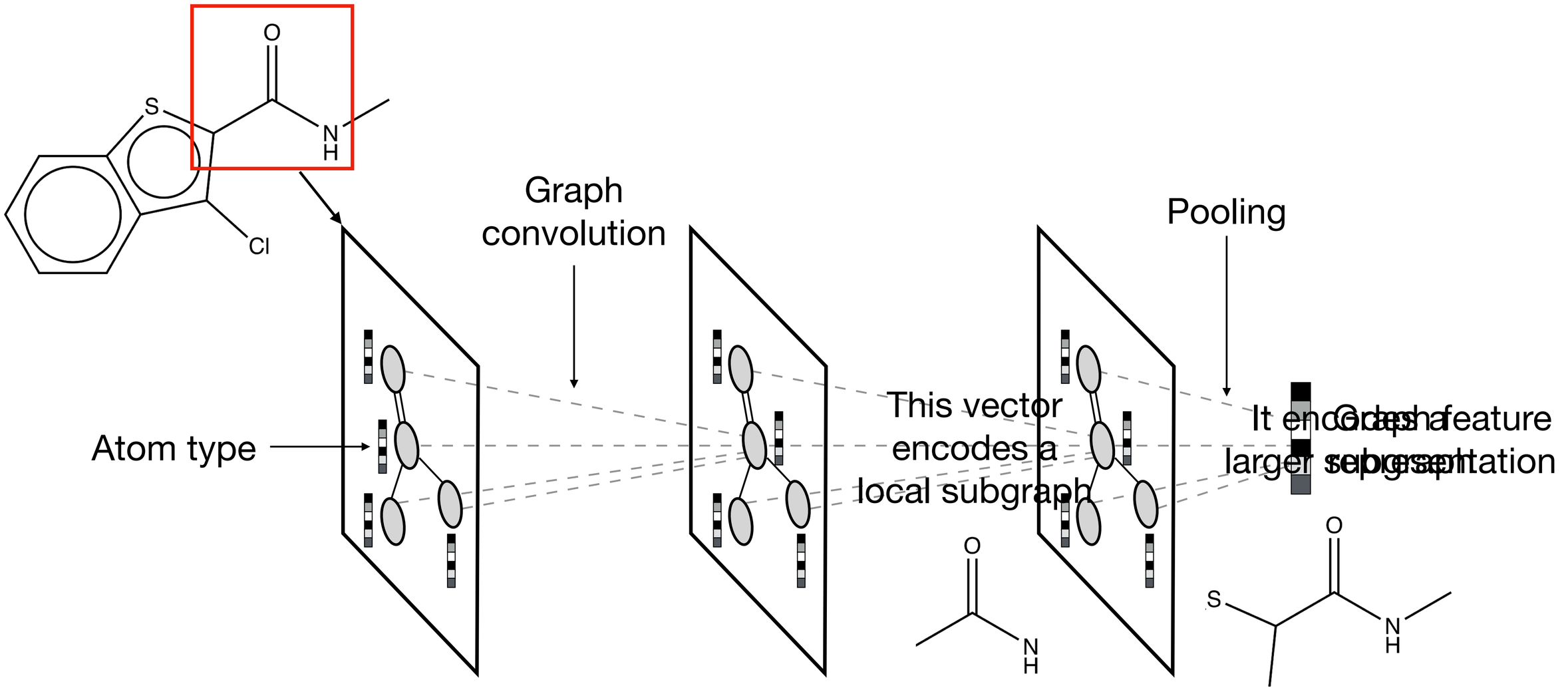


Graph neural network (GNN)

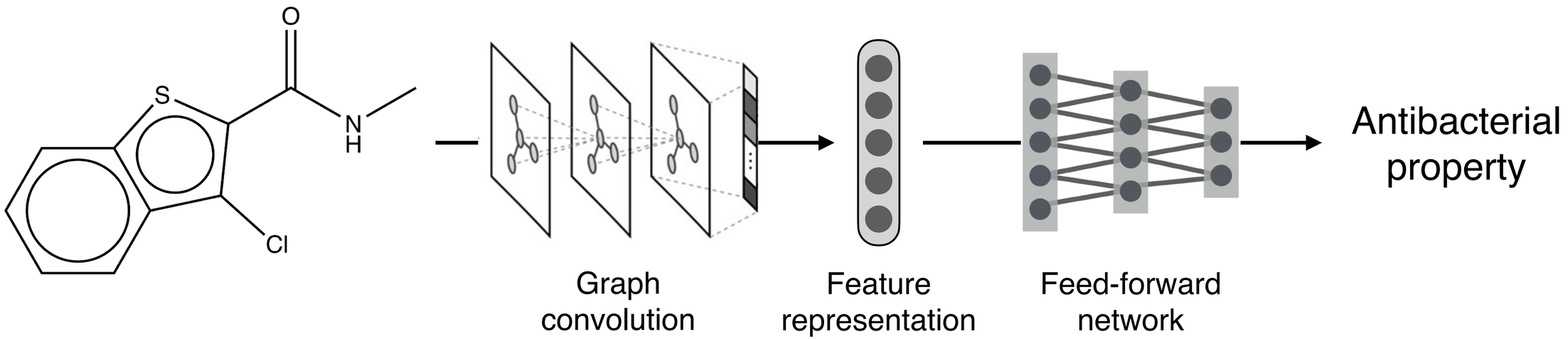
- Rich history of GNNs (Gori et al., 2005, Scarselli et al., 2009, Duvenaud et al. 2015, Kearnes et al. 2016, Jin et al., 2017, Gilmer et al., 2017, Zitnik et al., 2018, etc.)
- A molecule is represented as a graph

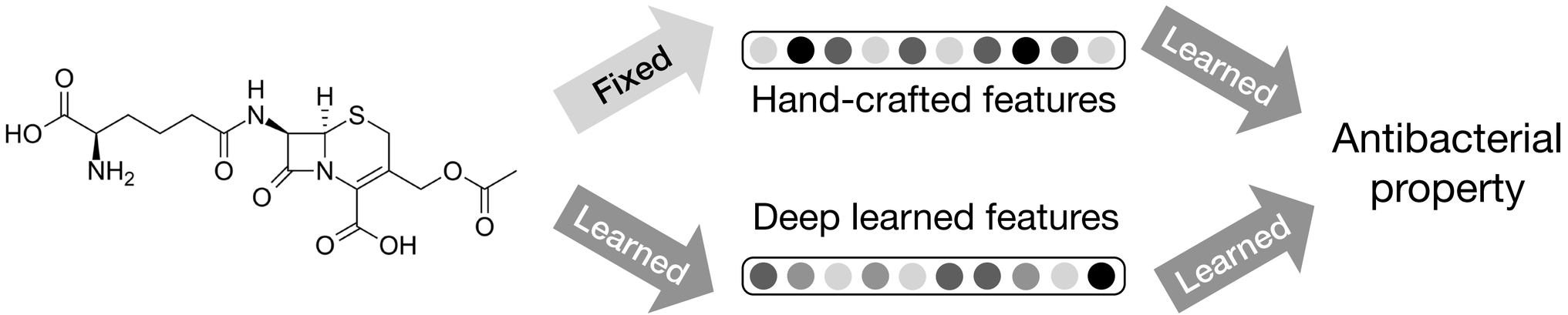


Graph neural network (GNN)



Graph neural network (GNN)

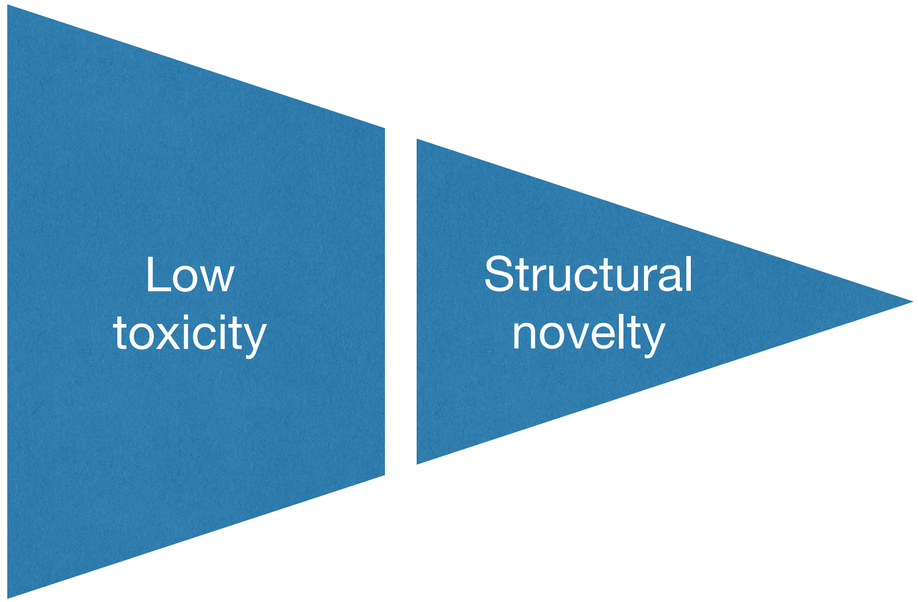


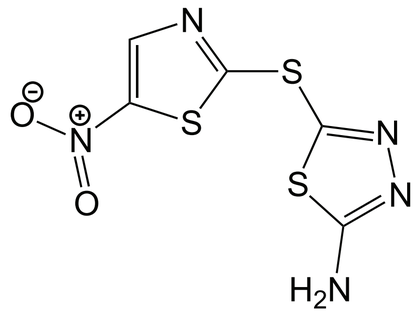


Use GNN for virtual screening

- We virtually screened 10^4 compounds in Broad drug repurposing hub
- We experimentally tested the top 99 compounds in the Broad Institute
- 51 of them are indeed antibacterial — hit rate = 51.5%

51 drugs

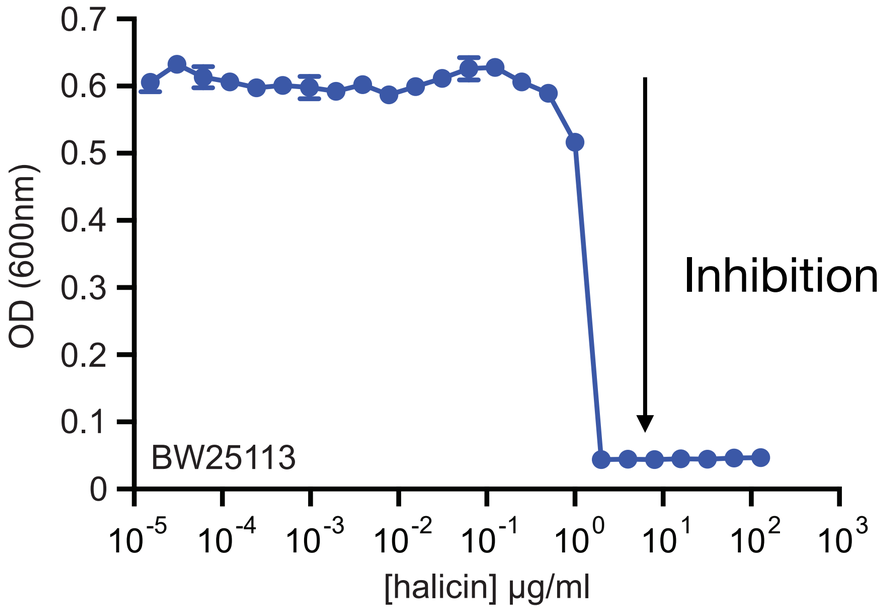


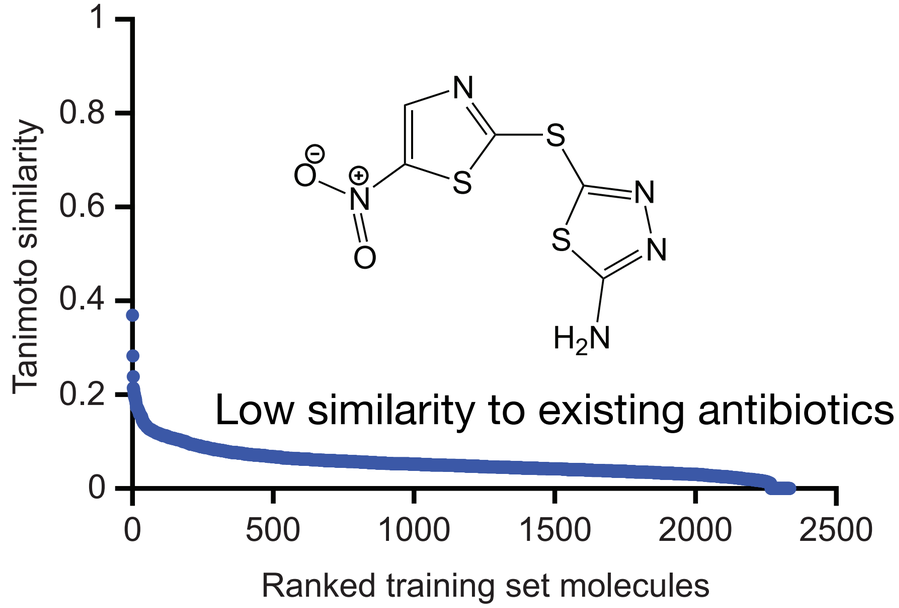


Compound SU3327
(renamed as Halicin)

Halicin is a novel and potent antibiotic

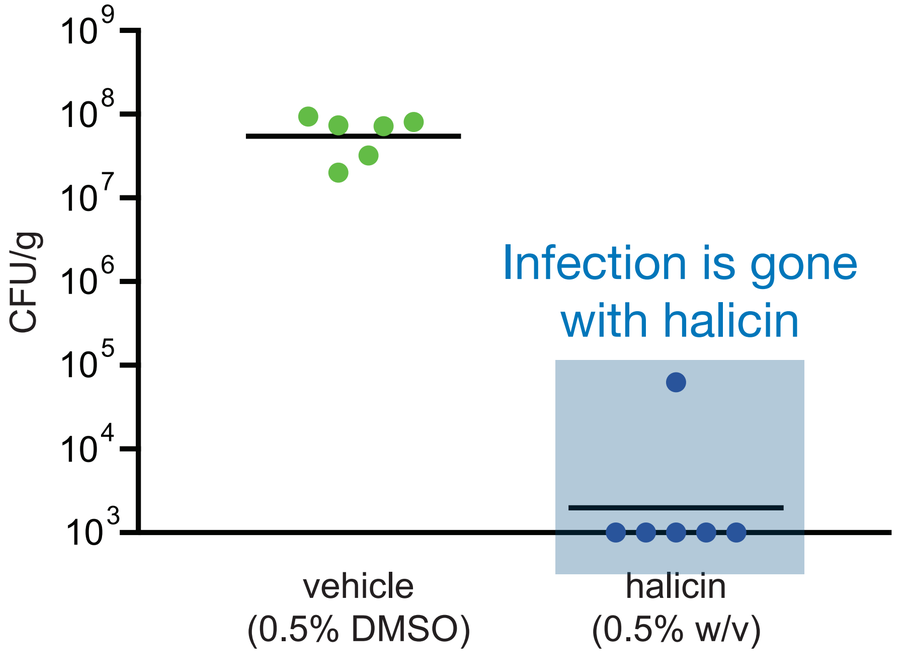
- Halicin shows potent growth inhibition against *E. coli in vitro*
- It is also structurally different from known antibiotics



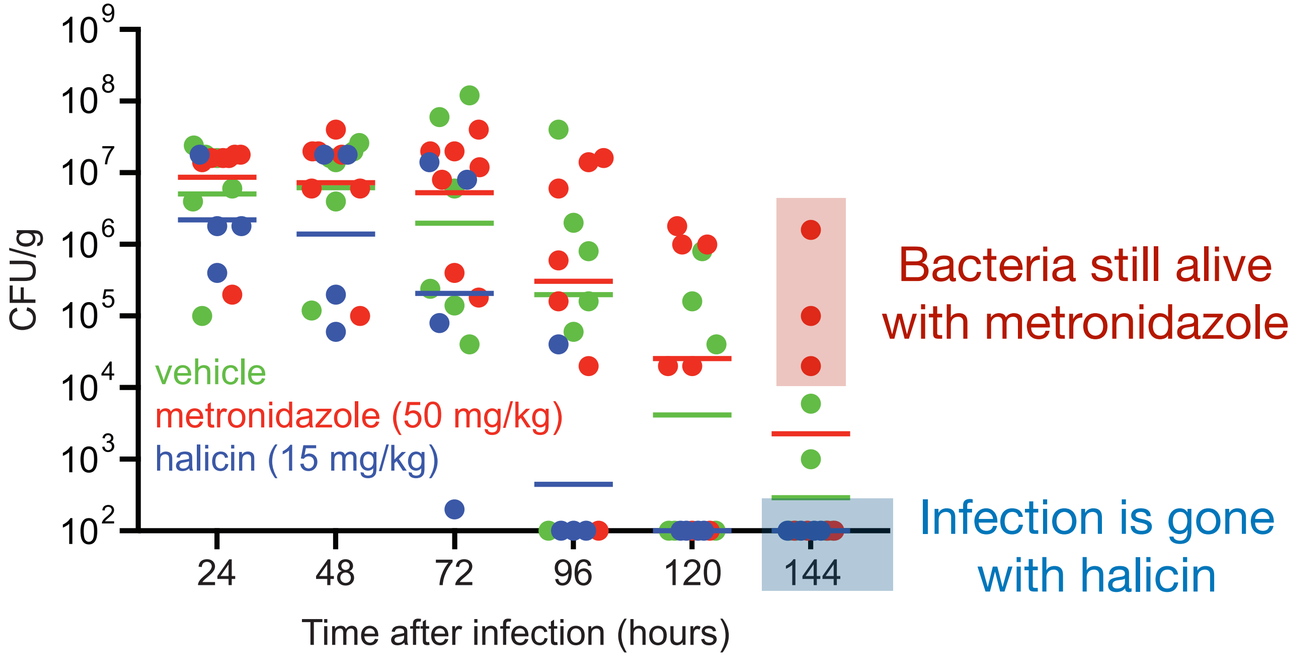


Halicin is potent to resistant bacteria in mice

Strong *in vivo* inhibition of
pan-resistant *A. baumannii*

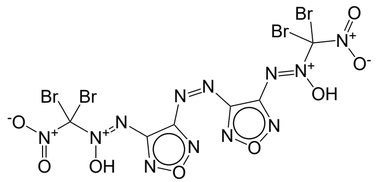


Strong *in vivo* inhibition against
resistant *C. difficile*



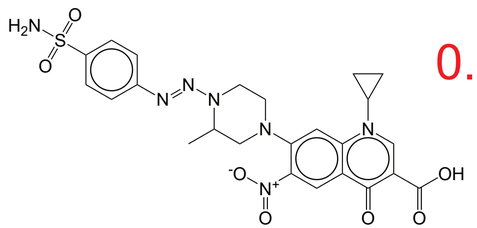
Large-scale virtual screening

- Applied the same model to screen 10^8 compounds in the ZINC library
- We identified **8** more compounds with inhibition against *E. coli* (**EC**), *S. aureus* (**SA**), *K. pneumoniae* (**KP**), *A. baumannii* (**AB**), or *P. aeruginosa* (**PA**) *in vitro*



EC	SA	KP	AB	PA
4	2	4	4	4

Minimum inhibitory concentration (μM)



EC	SA	KP	AB	PA
0.01	0.25	4	0.25	0.06

Minimum inhibitory concentration (μM)

Compare GNN with other models

- **Only GNN ranks Halicin among the top 100 compounds.**
- Given our budget, Halicin would not be discovered by other models

Model	Feature	Rank of Halicin
Graph neural network	**Learned**	**61**
Feed-forward neural network	RDKit features (fixed)	273
Feed-forward neural network	Morgan fingerprint (fixed)	1217
Random forest	Morgan fingerprint (fixed)	2640
Support vector machine	Morgan fingerprint (fixed)	771

Learned features are better than hand-designed features

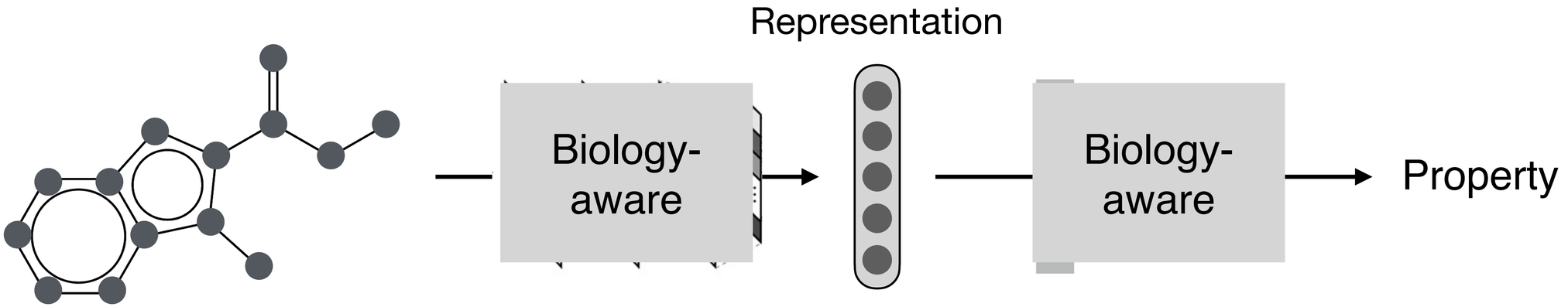
Part 2: infuse biological knowledge in GNNs

- **Part 1: graph neural networks for antibiotic discovery**
  [ICML'17, NeurIPS'17, JCIM'19, Cell'20]

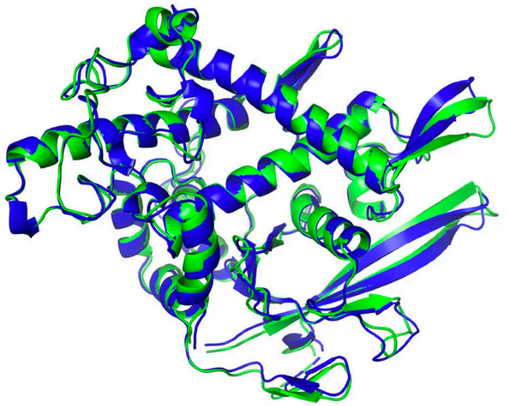
- **Part 2: Incorporate biological knowledge into graph neural networks: application to COVID-19 drug combination discovery**
  [PNAS (In submission)]

- **Part 3: Generative models for de novo drug design**
  [ICML'18, ICLR'19, ICML'20a,b,c]

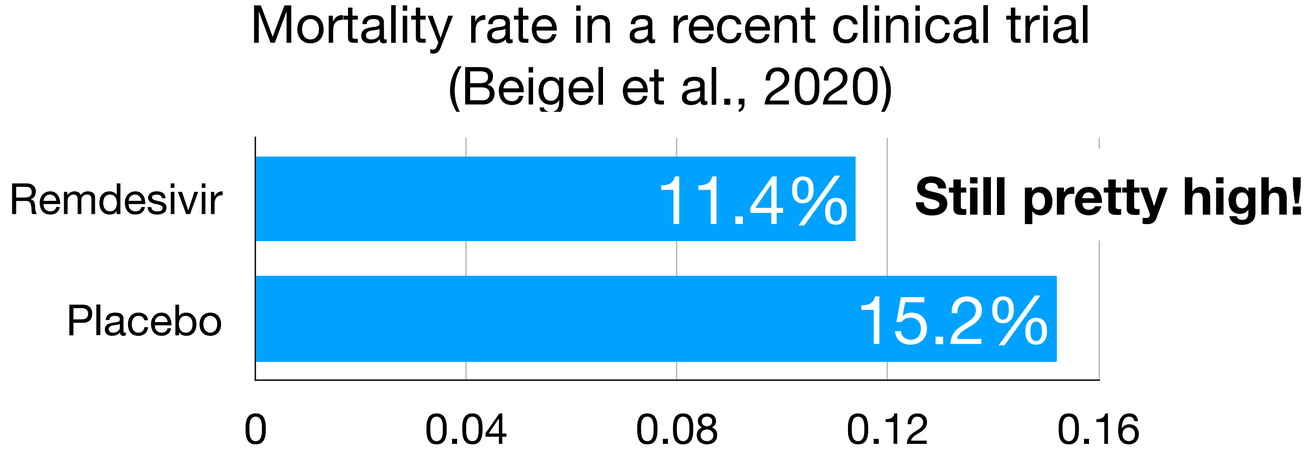
Motivation for biology-aware models



- Existing property prediction models only look at the chemical structure
- But properties may depend on additional biological information



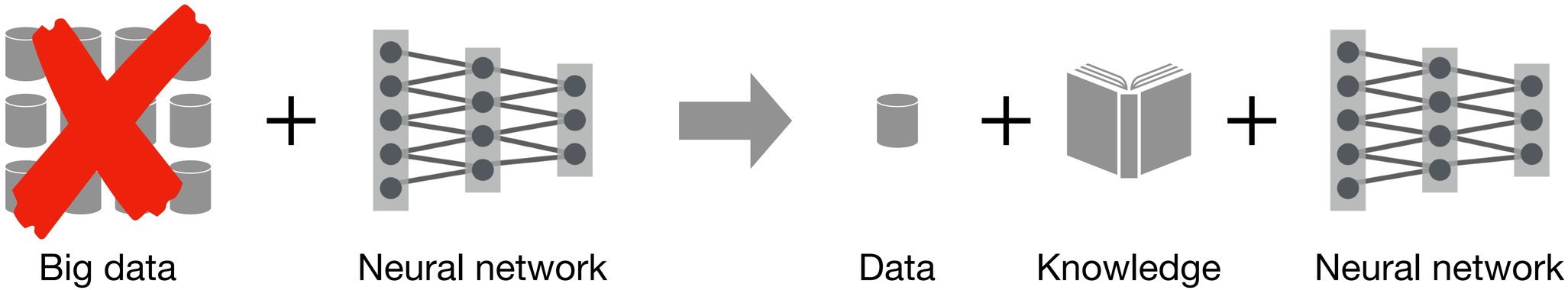
Case study: COVID-19 drug combinations



- Most HIV treatments are drug combinations
- effect(💊, 💊) >> effect(💊) + effect(💊)
- Can we find drug combinations for COVID?

Case study: COVID-19 drug combinations

- Two drugs are synergistic if effect(💊, 💊) >> effect(💊) + effect(💊)
- **Goal**: Train a model to predict whether a drug combination is synergistic
- **Challenge**: training data is limited (less than 200 drug combinations), but deep neural networks are very data hungry



Biological knowledge of viral replication

How can a drug block COVID-19 infection?

1. Block viral entry by inhibiting ACE2 or TMPRSS2
2. Inhibit viral proteases: 3CLpro, PLpro, RdRp
3. Inhibit host targets that interact with viral proteins (Gordon et al., 2020)

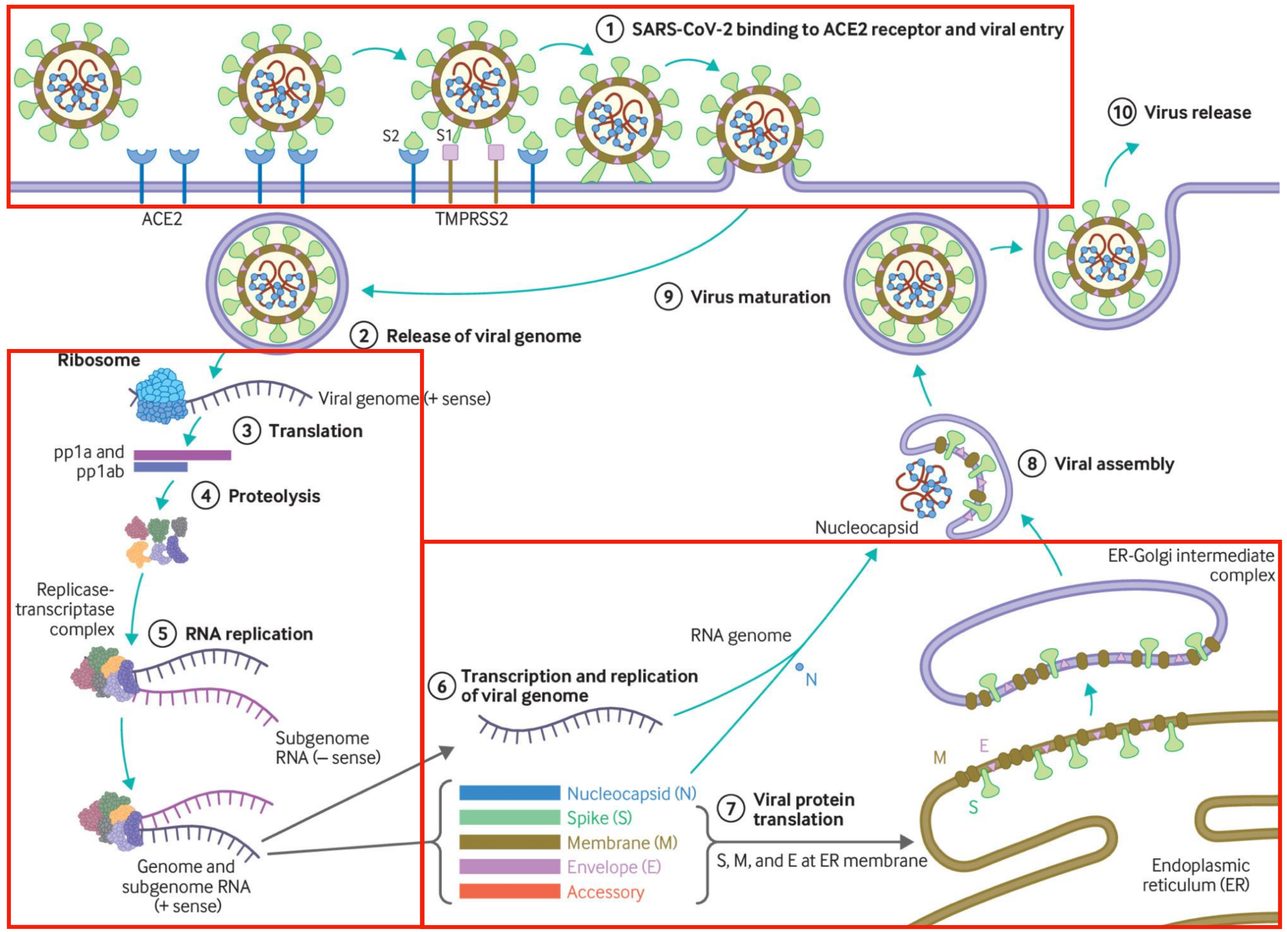
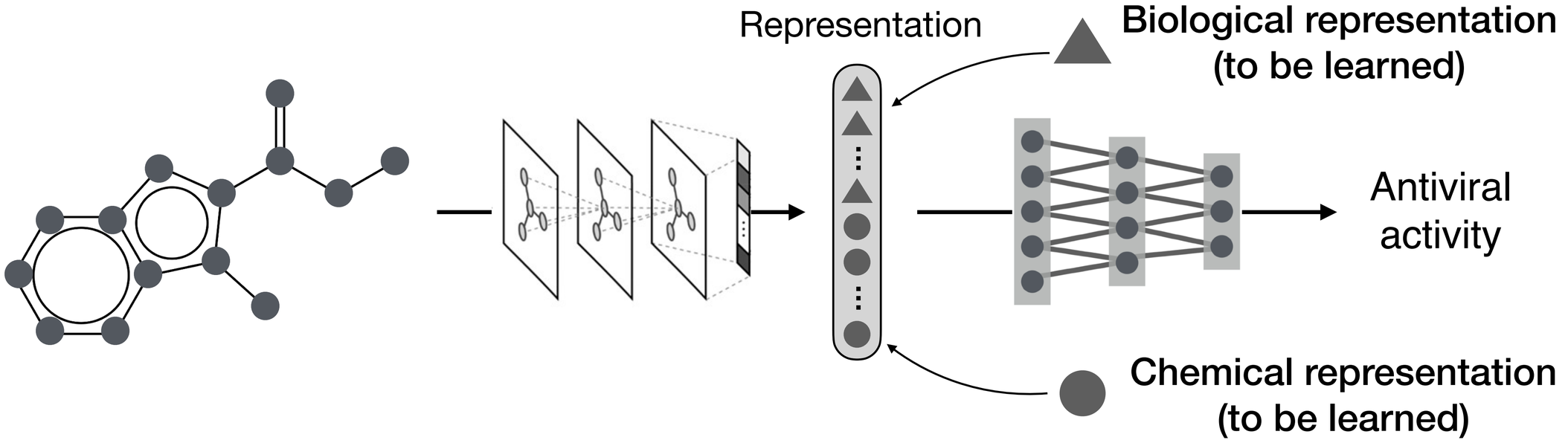


Figure source: Cevik et al., *BMJ* 2020

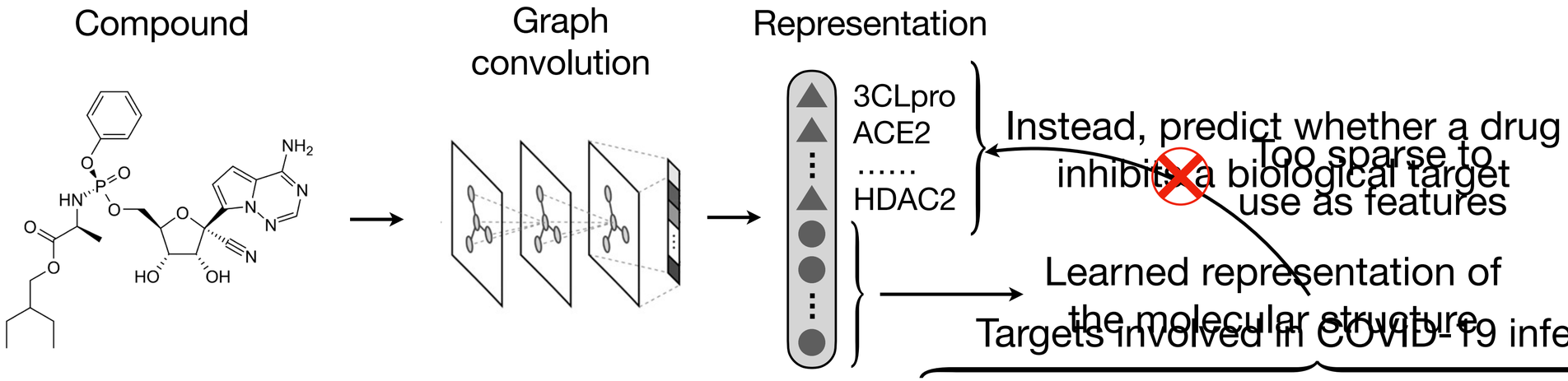
ComboNet incorporates biology & chemistry

- Synergy comes from inhibition of certain biological targets (e.g., proteins)
- Model biological interaction ⇒ additional data ⇒ better generalization



ComboNet learns drug-target interaction

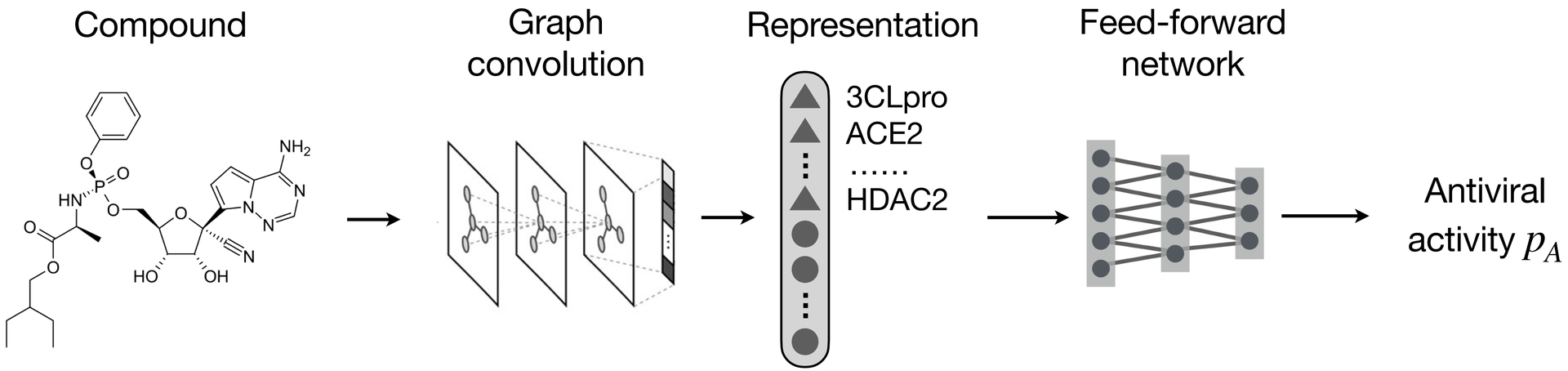
1. Predict drug-target interaction — whether drug A inhibits target B

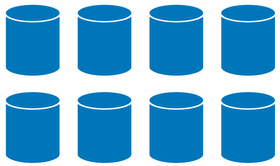


Drug-target interaction data
(ChEMBL and NCATS)

ComboNet learns antiviral activity

2. Single-agent antiviral activity prediction



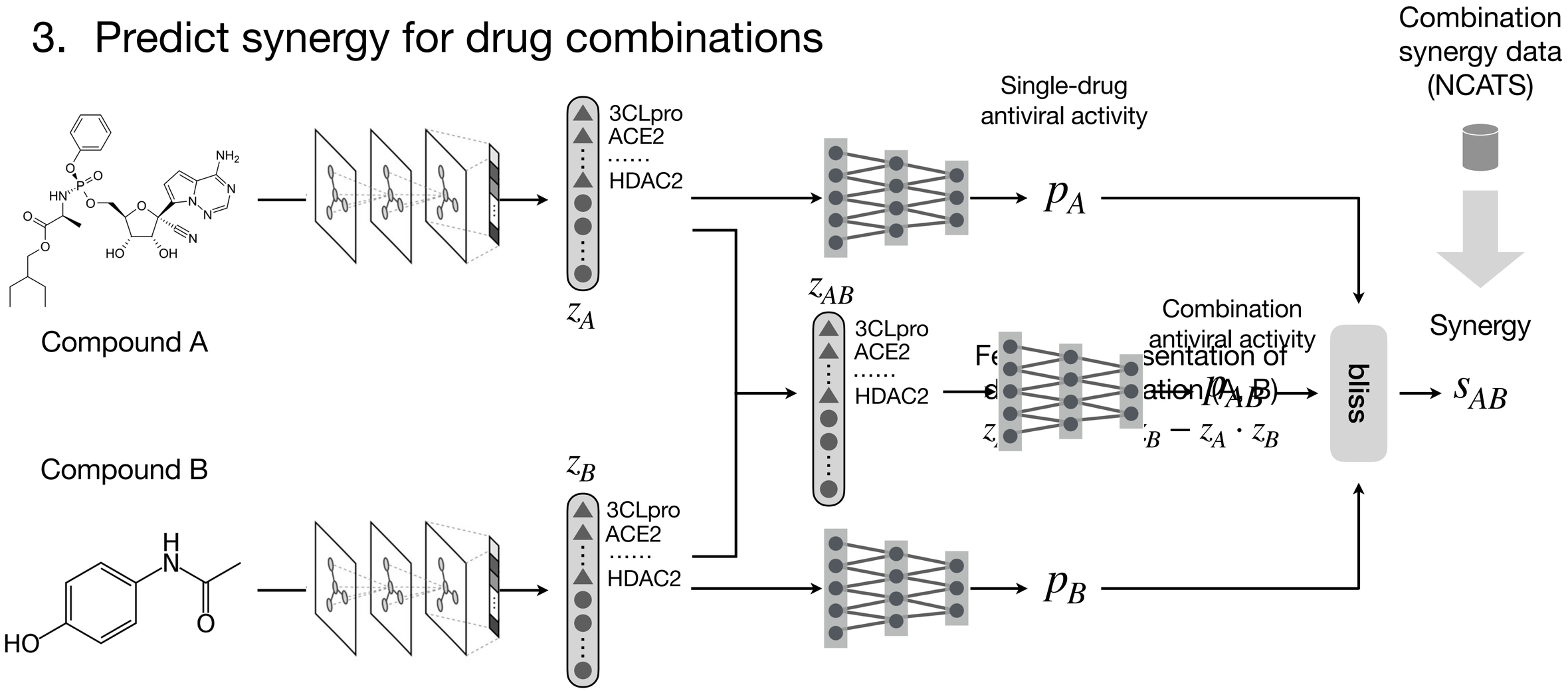


Single-drug antiviral activity data (NCATS)

Drug	Reserpine	Remdesivir	Penicillin	Halicin
Antiviral?	Yes	Yes	No	No

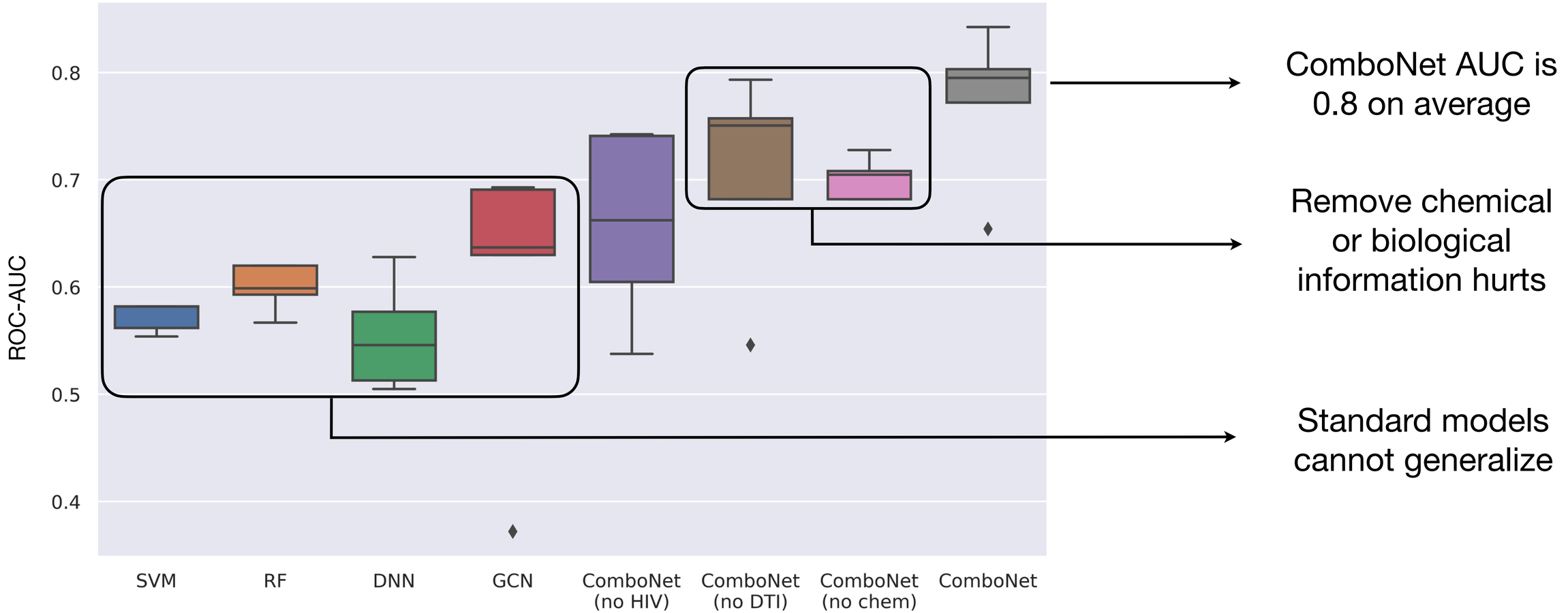
ComboNet learns antiviral synergy

3. Predict synergy for drug combinations



ComboNet performance

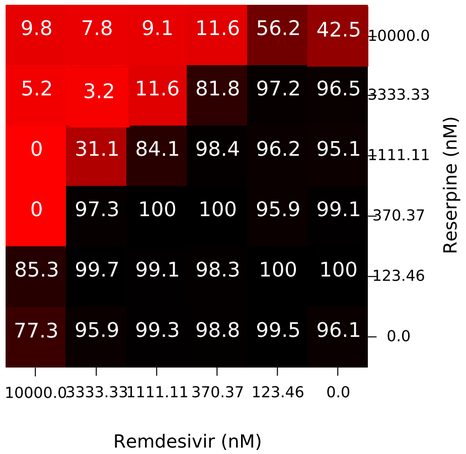
- Training set (88 drug combinations); Test set (71 drug combinations)



Discover new drug combinations

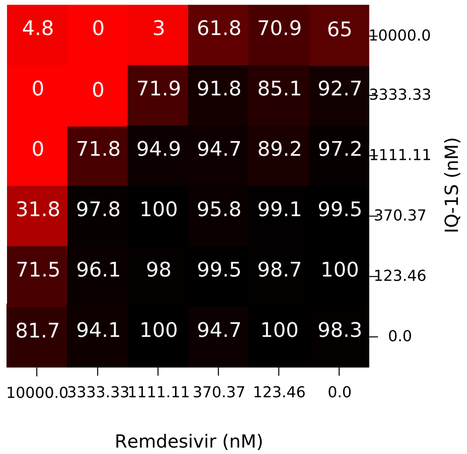
- Collaboration with National Center for Advancing Translational Science (NCATS)
- We experimentally tested top drug combinations in NCATS Vero E6 cell assays
- Further studying these combinations in human cell lines

Remdesivir + reserpine



Drug	Virus alive (%)
Remdesivir	77.3%
Reserpine	42.5%
Combination	**3.2%**

Remdesivir + IQ-1S



Drug	Virus alive (%)
Remdesivir	81.7%
IQ-1S	65%
Combination	**0%**

Part 3: de novo drug design

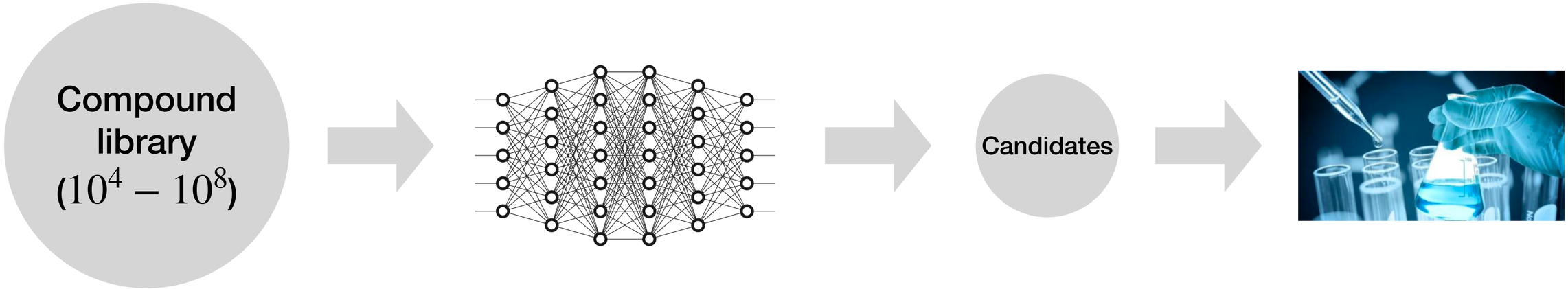
- **Part 1: graph neural networks for antibiotic discovery**
  [ICML'17, NeurIPS'17, JCIM'19, Cell'20]

- **Part 2: Incorporate biological knowledge into graph neural networks: application to COVID-19 drug combination discovery**
  [PNAS (In submission)]

- **Part 3: Generative models for de novo drug design**
  [ICML'18, ICLR'19, ICML'20a,b,c]

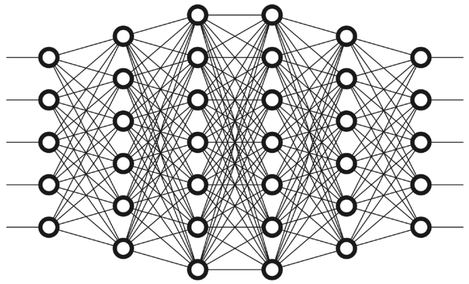
Motivation for de novo drug design

- Deep learning can discover new antibiotics and COVID-19 drugs
- Simple approach: train a GNN to rank all the compounds in our library
  - Reason: maximize the speed of experimental validation
- **Problem**: number of drug like molecules = 10^{60}. We can't rank all of them.



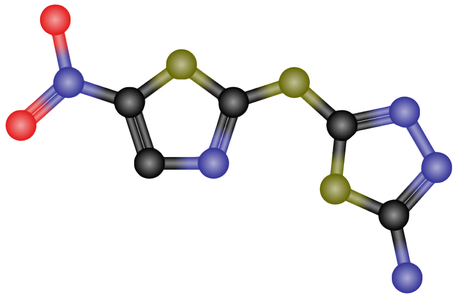
Graph generation for de novo drug design

- Learn a distribution whose mass is concentrated around “good” molecules
- Let’s train a generative model to directly generate “good” molecules
- It can efficiently explore the entire chemical space (10^{60} molecules)



Generative model

Generate

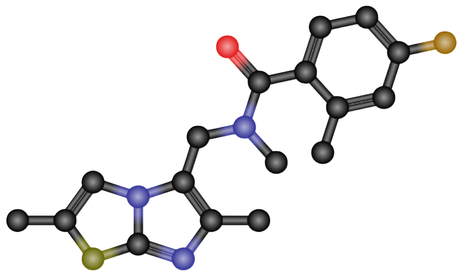


A good molecule

How to generate molecular graphs?

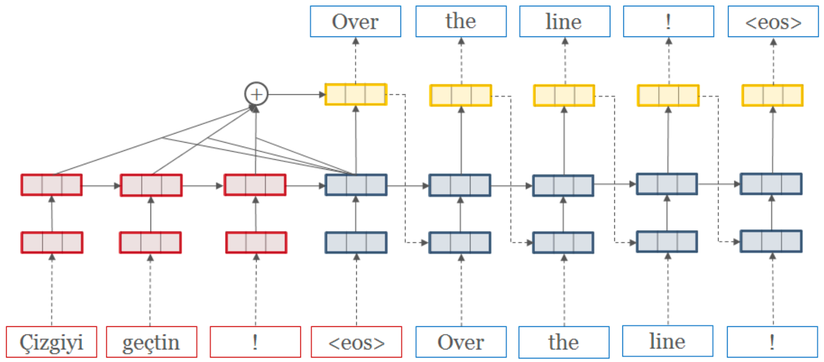
Previous solution 1: sequence-based methods

- Prior work used recurrent neural networks to generate molecular graphs
  (Olivecrona et al., 2018; Gomez-bombarelli et al., 2018; Popova et al., 2018; …)
- Convert a molecule into a SMILES string (a domain specific language)
  (Weininger, 1988)



Convert it into a SMILES string

Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1



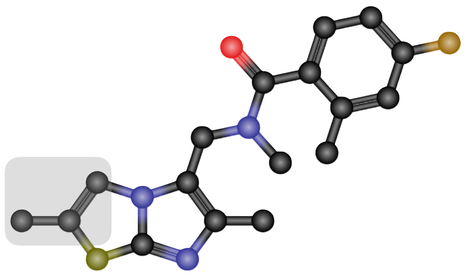
Recurrent neural networks (RNNs)

Weininger, D. SMILES, a chemical language and information system. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Problems of sequence-based approach

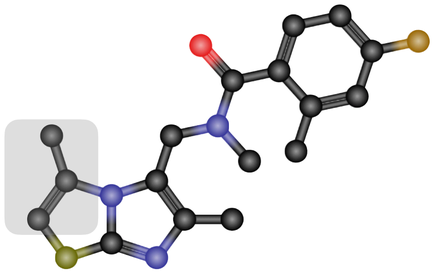
- Prior work used sequence-based generative models for molecular graphs
  (Olivecrona et al., 2018; Gomez-bombarelli et al., 2018; Popova et al., 2018; …)
- But this string representation is quite brittle…

Two almost identical graphs



→ Cc1cn2c(CN(C)C(=O)c3ccc(F)cc3C)c(C)nc2s1

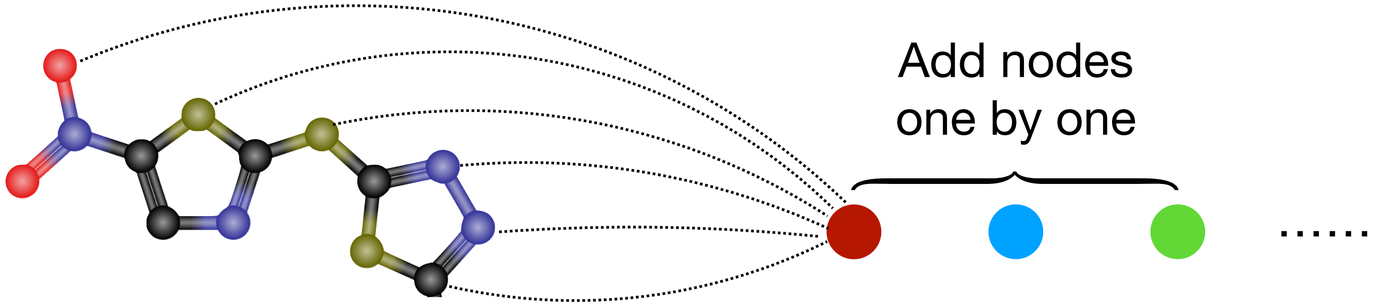
Quite different strings



→ Cc1cc(F)ccc1C(=O)N(C)Cc1c(C)nc2scc(C)n12

Previous solution: node-by-node generation

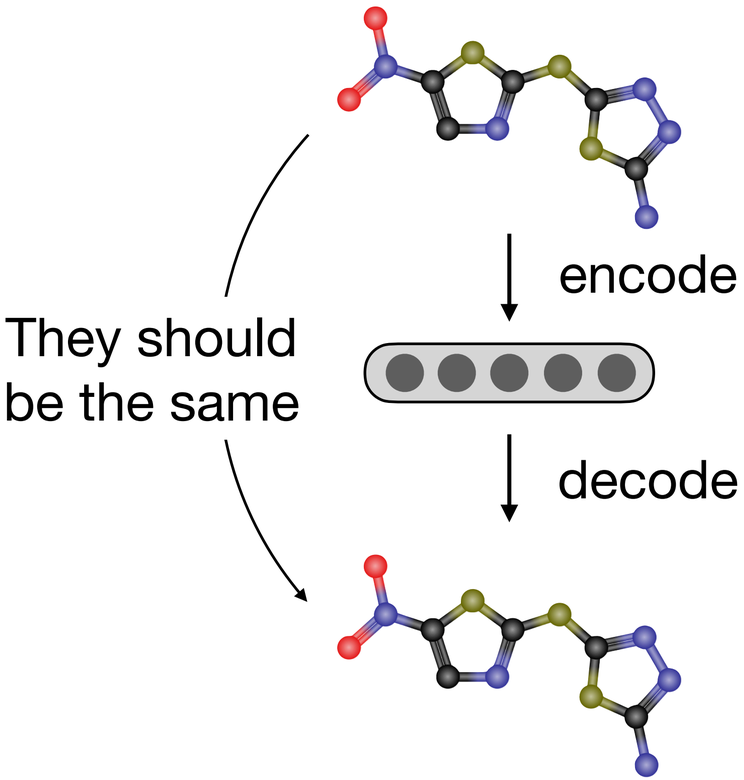
- A straightforward approach: generate a graph node-by-node (Liu et al., 2018)

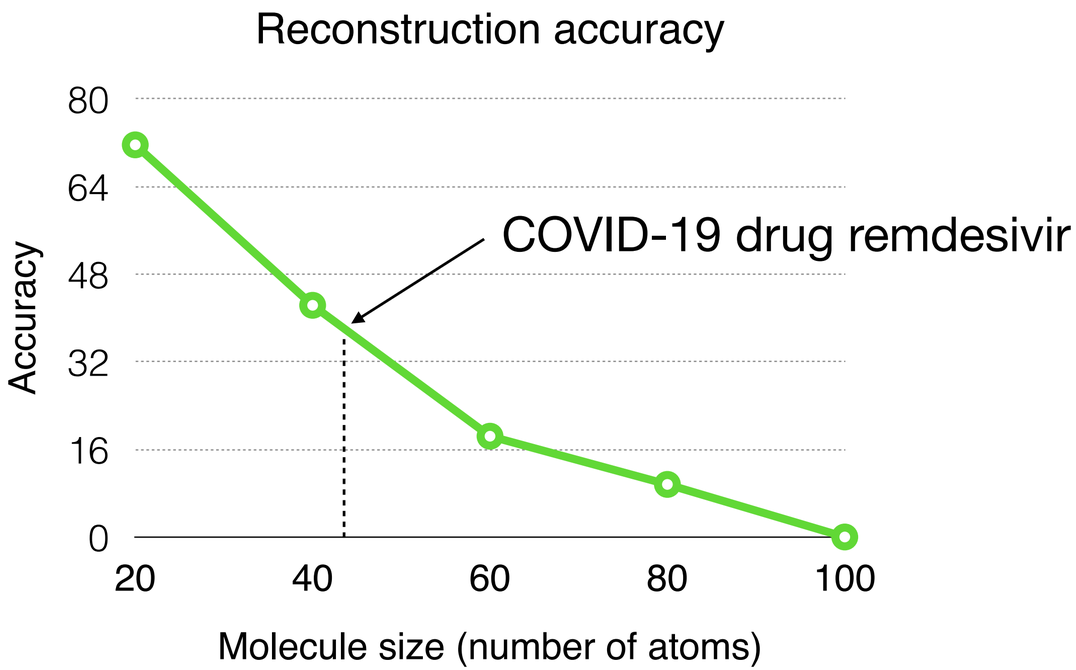


- Molecules are typically sparse: N nodes, $O(N)$ edges
- However, it needs to make $O(N)$ edge predictions in each step
- In total: $O(N^2)$ edge predictions

Failure of node-by-node generation

- Node-by-node generation via a variational auto encoder (VAE) (Liu et al., 2018)
- Diagnostic test: can the decoder reconstruct an input molecule?





We need to leverage inductive bias

Sequence (text)

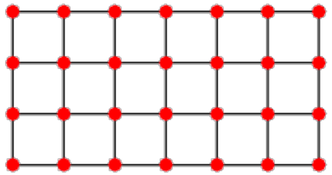
Grid graph (images)

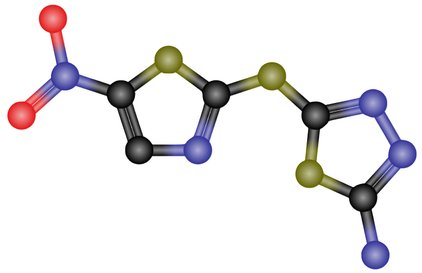
Molecular graphs (low treewidth)

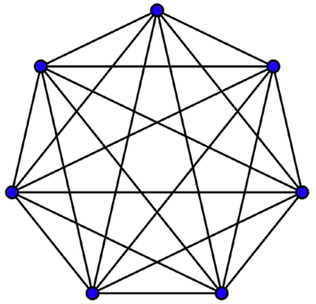
Inductive bias?

Dense graphs

Complexity



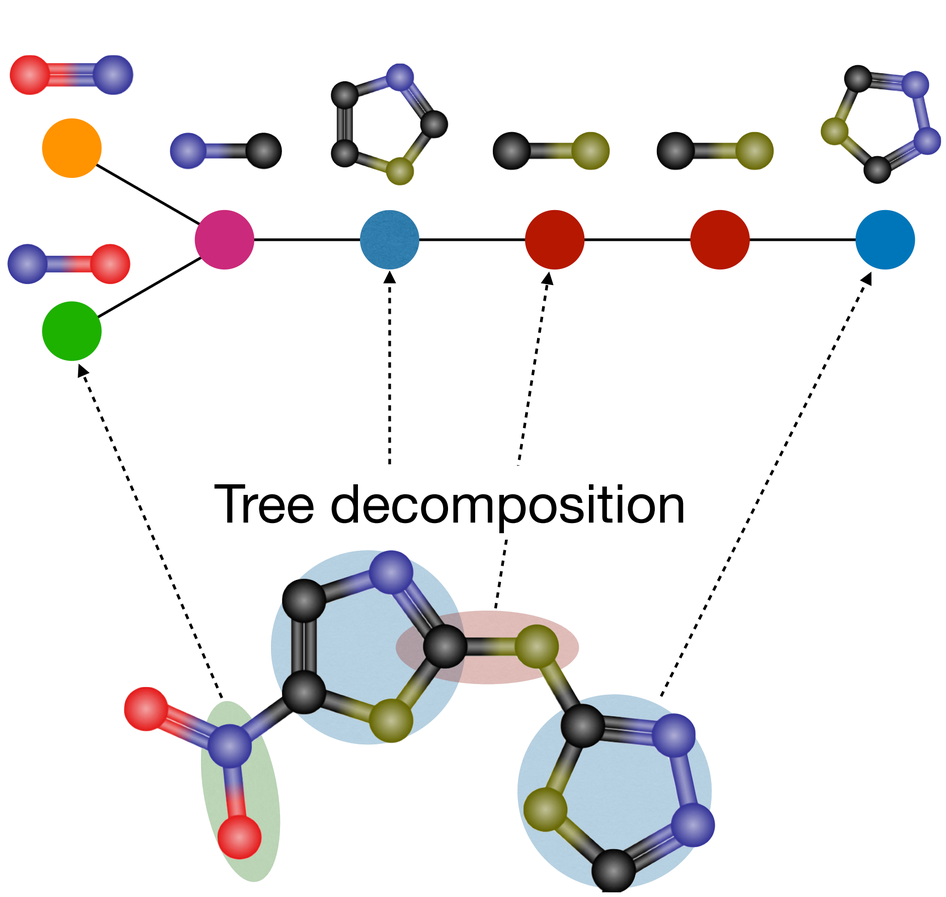




What's up?

All models leverage the inductive bias of the structure.

Junction tree variational autoencoder



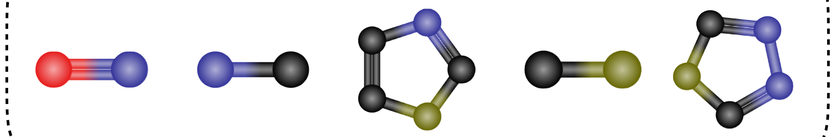
Motif

Motifs are small due to low treewidth

Junction tree

Molecular graph

Motif vocabulary

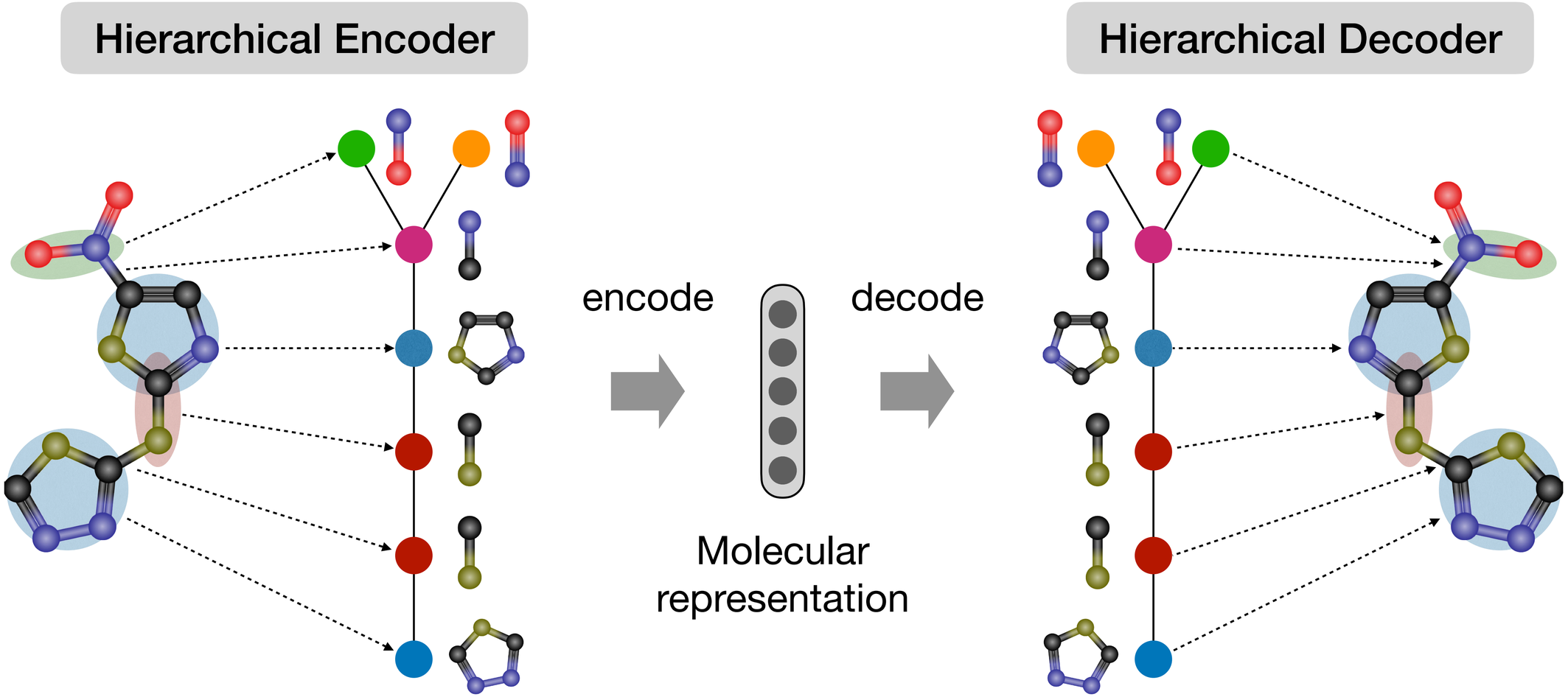


250K graphs ⇒ 638 motifs

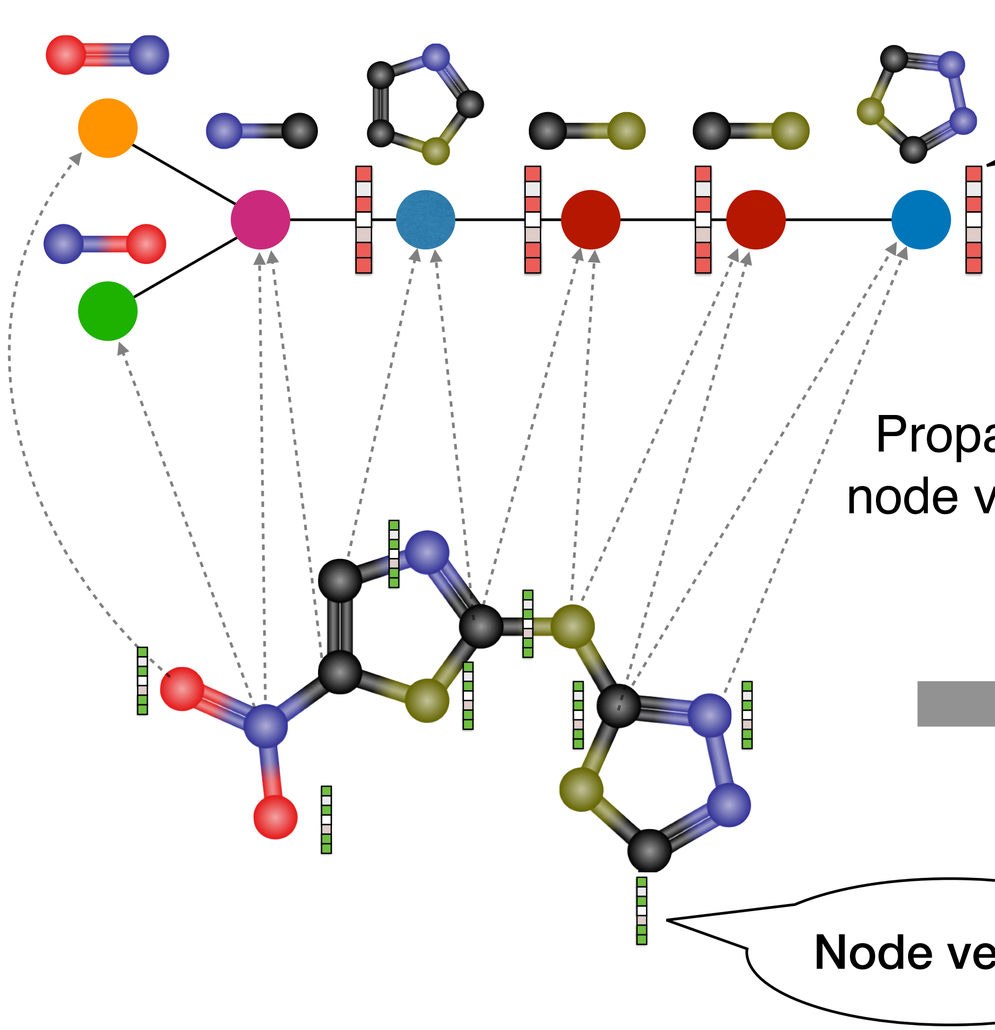
99.9% coverage (new graphs)

Inspired by the junction tree algorithm in graphical models.

Details: hierarchical encoder & decoder



Hierarchical graph encoder

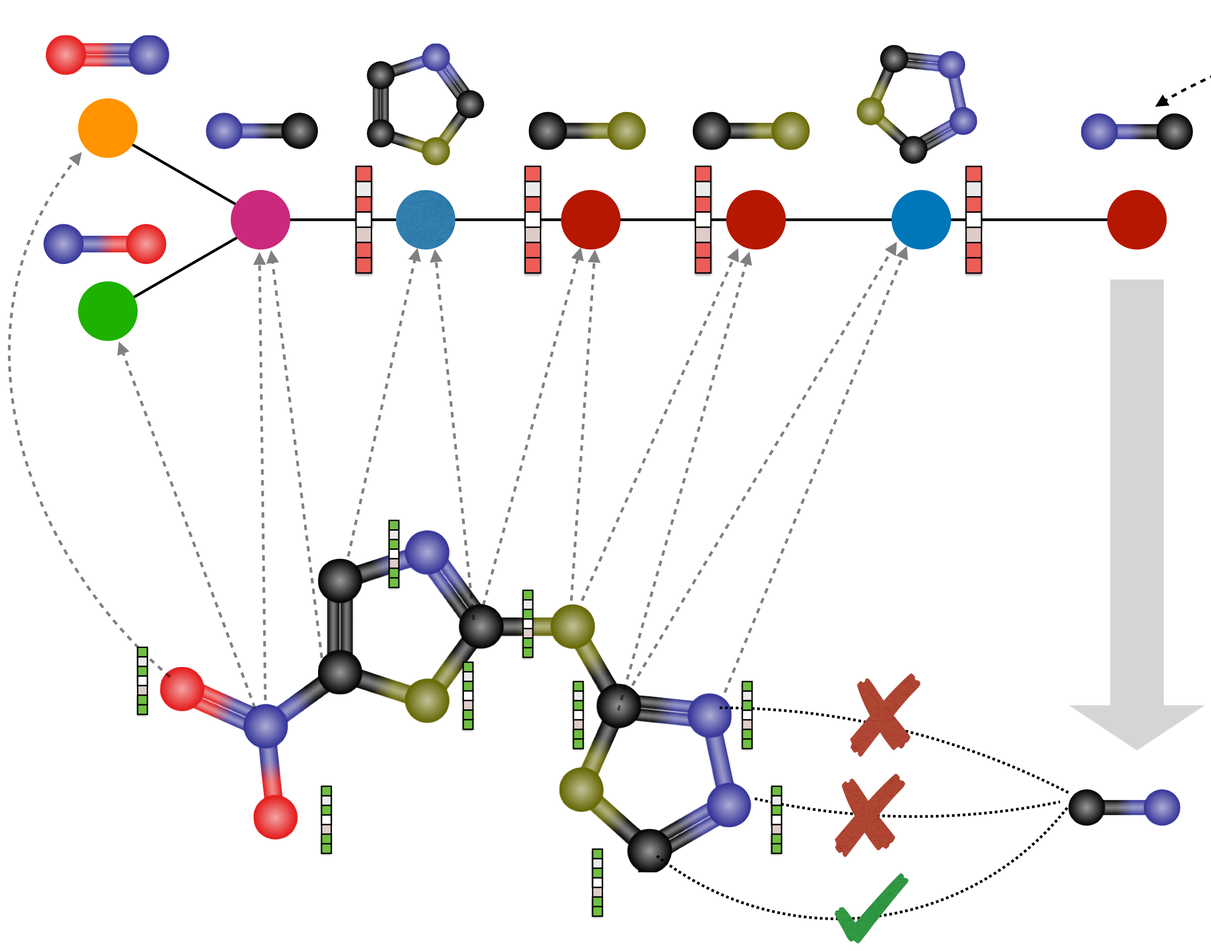


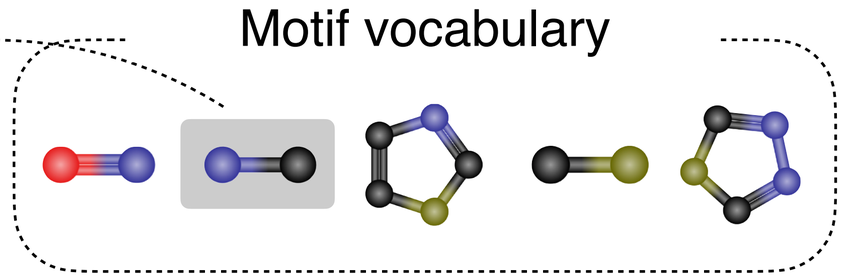
Propagate
node vectors

Run graph convolution in the
junction tree

Run graph convolution in the
molecular graph

Hierarchical graph decoder

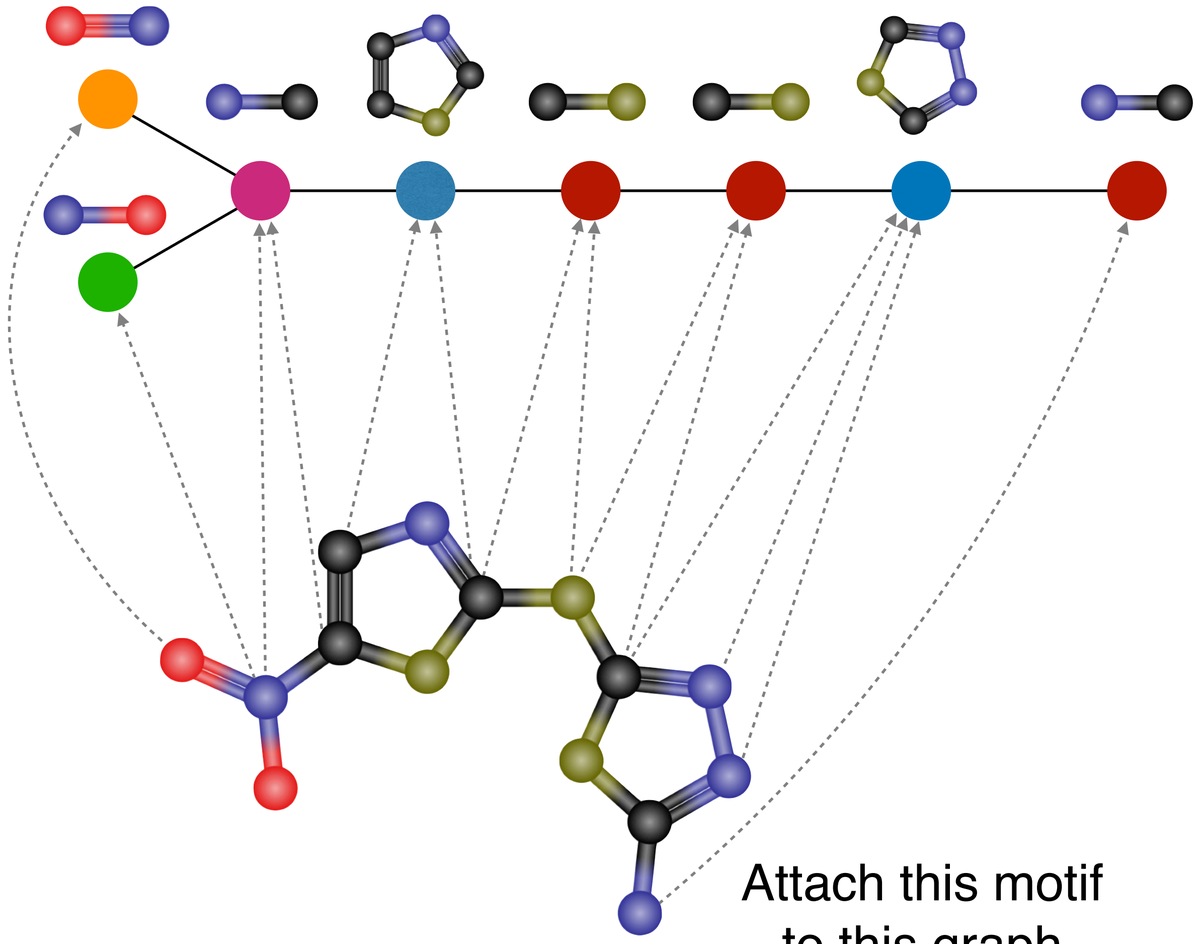




Step 1: predict the next motif

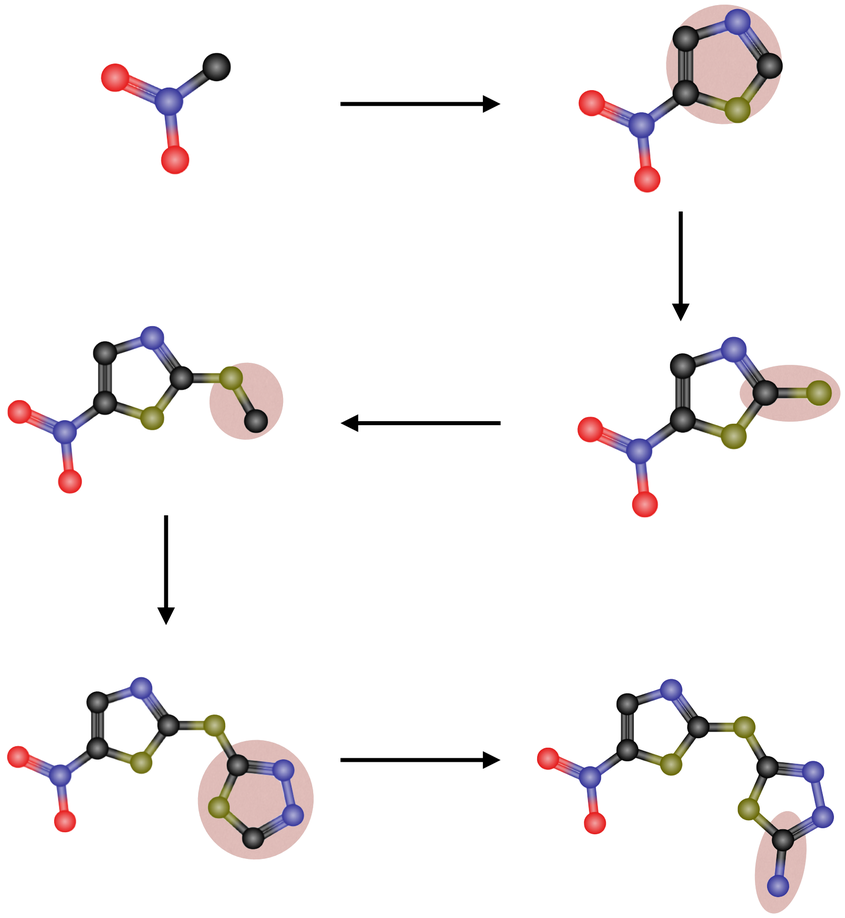
Step 2: predict how to attach this motif to the current graph

Hierarchical graph decoder



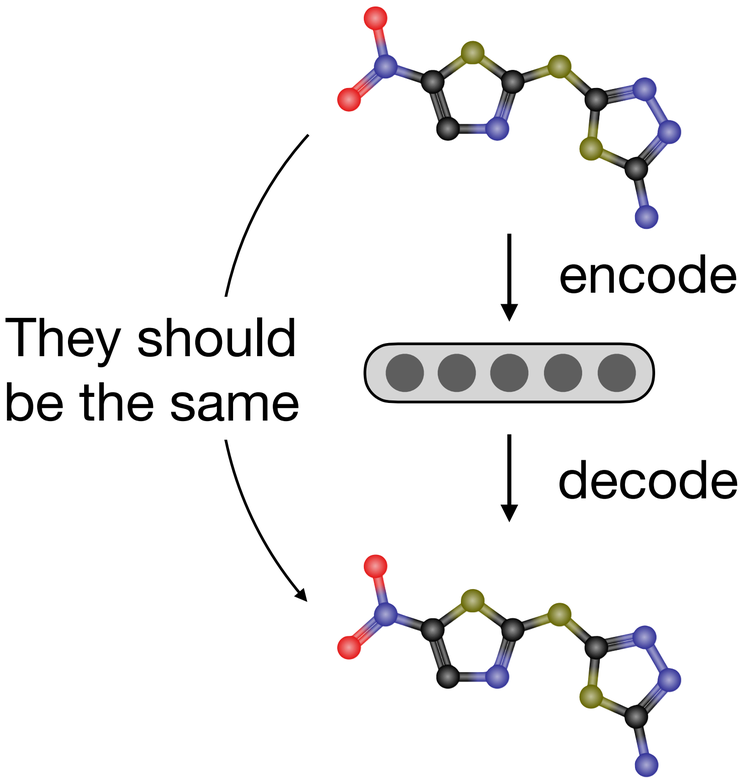
Attach this motif
to this graph

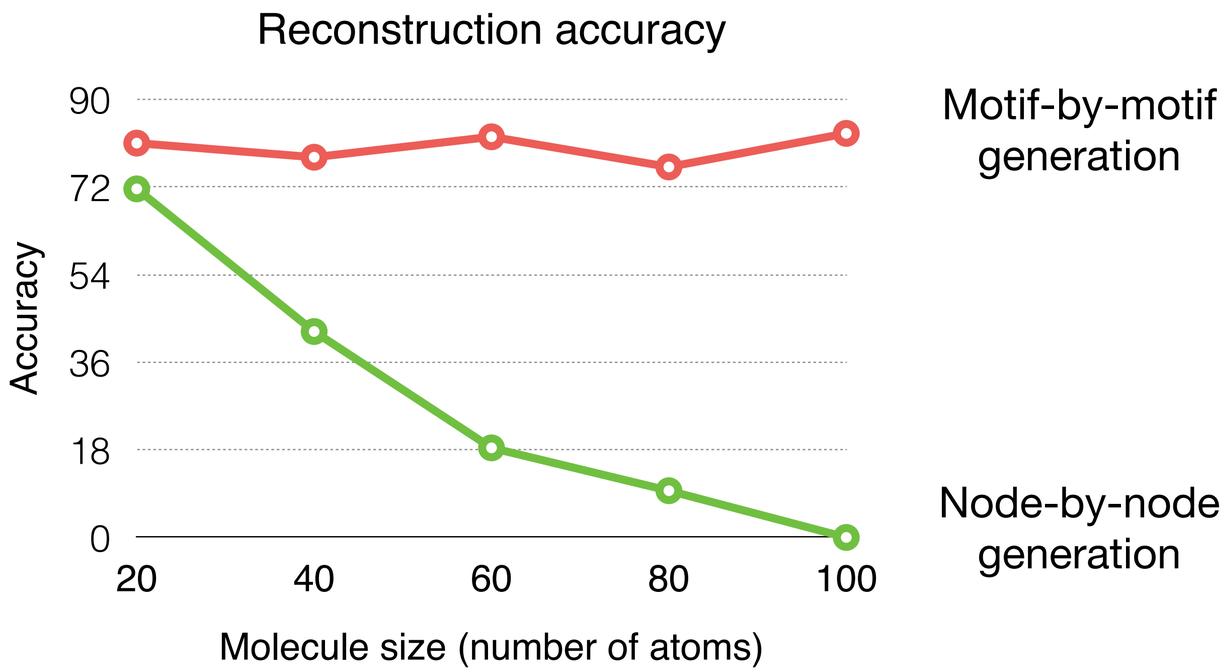
Motif-by-motif generation



Motif-by-motif versus node-by-node

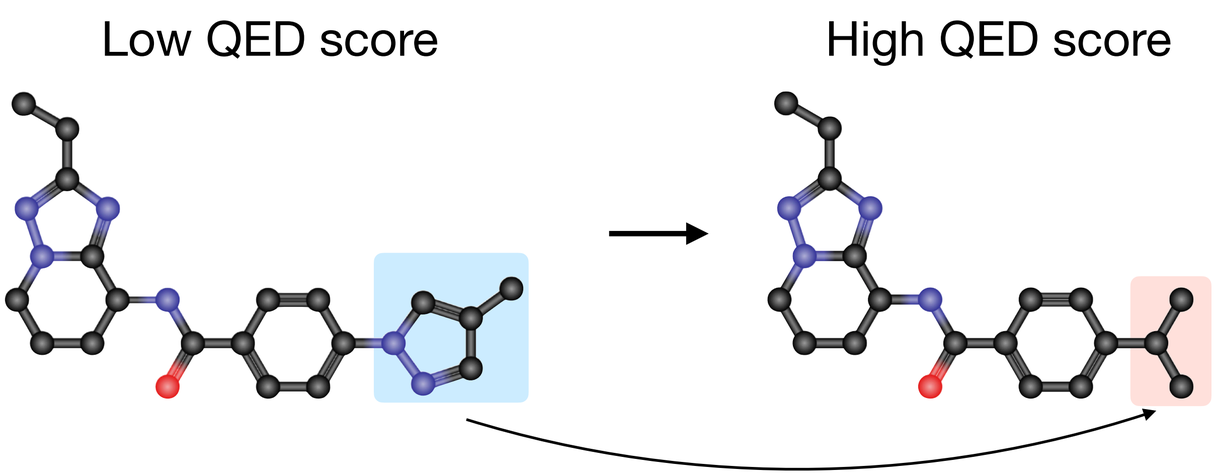
- Training objective: minimize reconstruction loss
- Motif-by-motif generation is able to reconstruct large molecules!

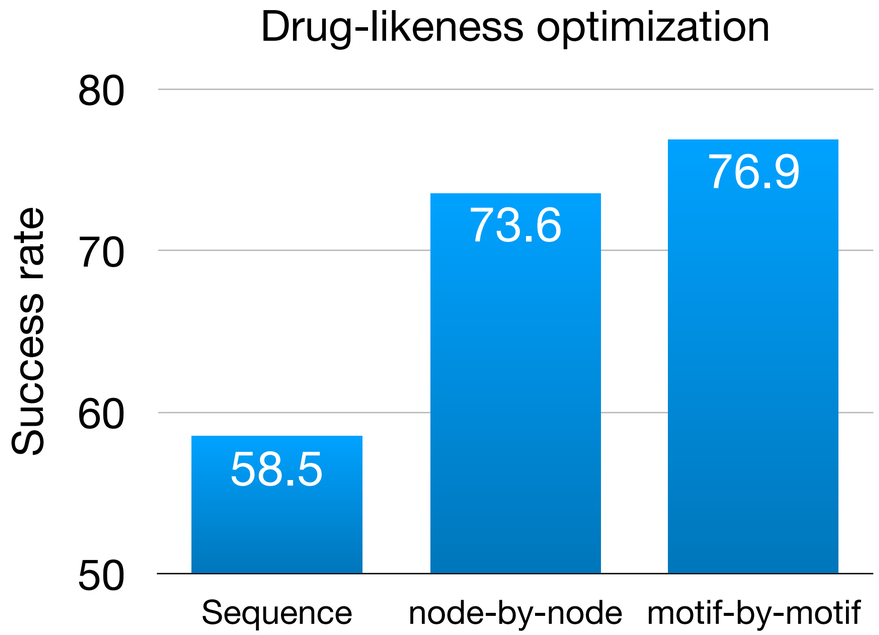




Results: molecular optimization

- Task: learn to modify a non-drug-like molecule into a drug-like molecule
- Drug-likeness is measured by QED scores (Bickerton et al., 2012)





Part 3: de novo drug design

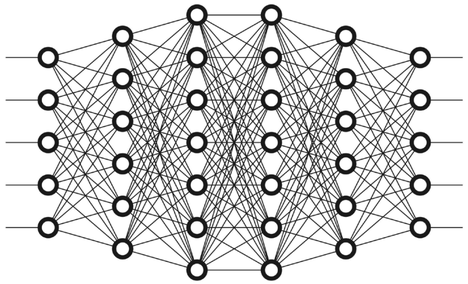
- **Part 1: graph neural networks for antibiotic discovery**
  [ICML'17, NeurIPS'17, JCIM'19, Cell'20]

- **Part 2: Incorporate biological knowledge into graph neural networks: application to COVID-19 drug combination discovery**
  [PNAS (In submission)]

- **Part 3: Generative models for de novo drug design**
  [ICML'18, ICLR'19, ICML'20a,b,c]

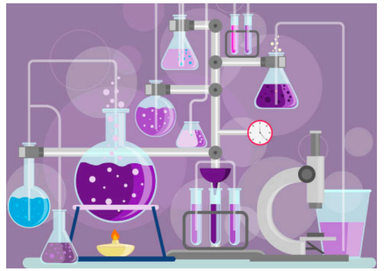
Deep learning for molecular sciences

Deep learning



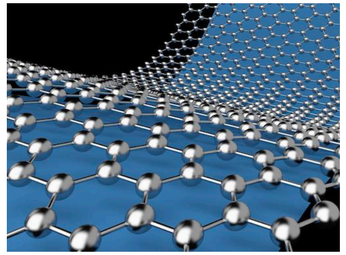
Drug discovery
(e.g., de novo drug design)

- Dahl et al., 2015;
- Stokes et al., 2020;
- Jin et al., 2018; ……



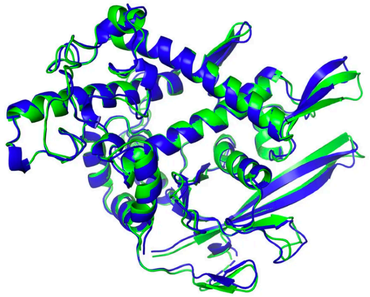
Chemistry
(e.g., reaction prediction)

- Duvenaud et al., 2015;
- Coley et al., 2019;
- Jin et al., 2017; ……



Material Science
(e.g., polymer design)

- Gomez-bombarelli et al., 2018
- Xie et al., 2019
- Jin et al., 2020; ……

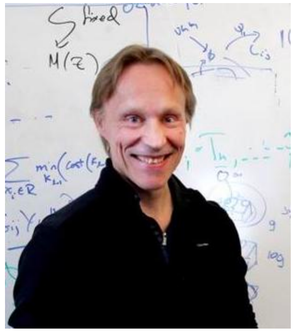


Biology
(e.g., protein folding)

- Rao et al., 2019;
- Senior et al., 2020;
- Jin et al., 2020; ……

Thanks to my collaborators


Regina Barzilay


Tommi Jaakkola


Klavs Jensen


William Green


Phillip A. Sharp


James Collins


Caroline Uhler


Rafael Gomez-Bombarelli


Connor Coley


Camille Bilodeau


Peter Sorger


Rachel Wu


Jonathan Stokes


Kyle Swanson

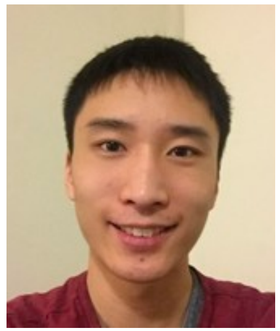

David Alvarez-Melis


Guang-He Lee


Allison Tam


Nienke Moret


Anne Fischer


Kevin Yang


Tao Lei

Thanks to my collaborators