AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 10: AI for protein structure prediction, Drug discovery and therapeutic science, Structure- and sequence-based co-design, Small molecule design, Local and global optimization in molecular design, Conditional protein generation

HARVARD
MEDICAL SCHOOL

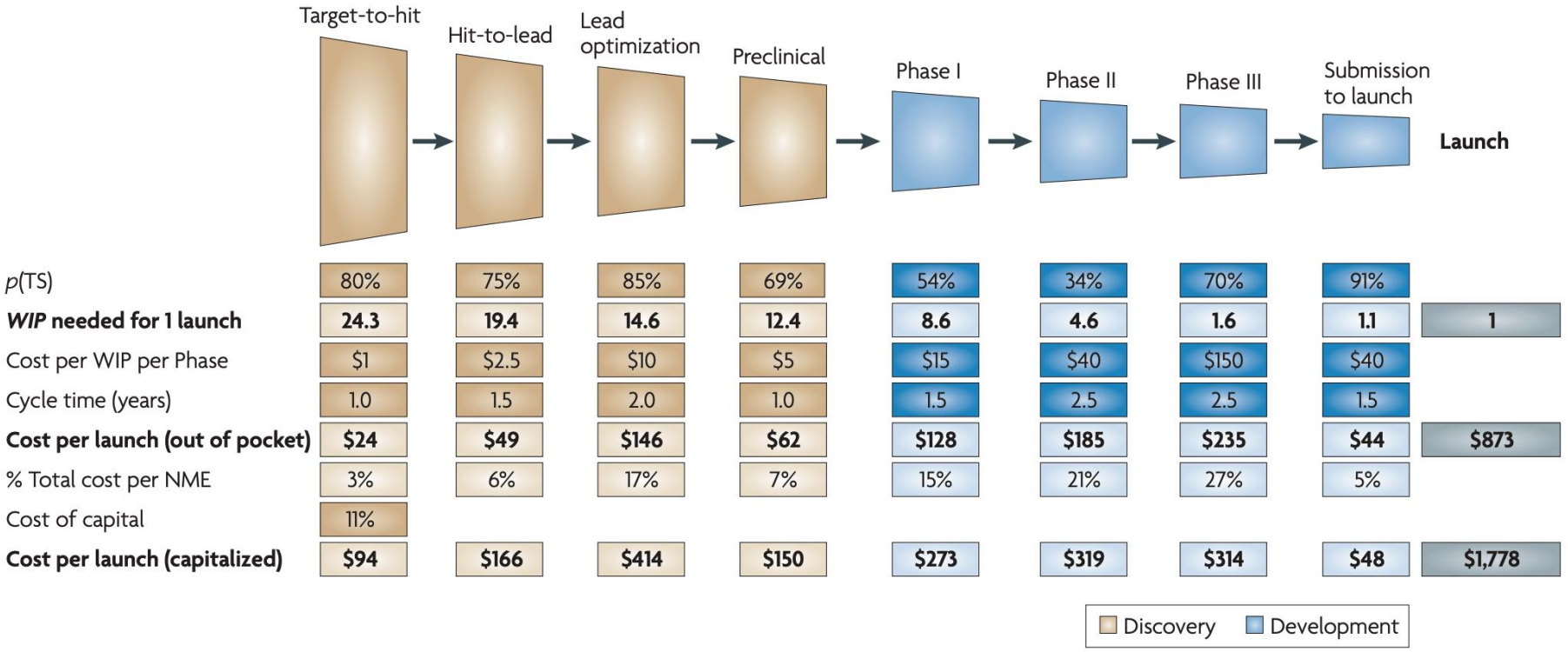
Kempner
INSTITUTE

For the Study of Natural
& Artificial Intelligence
at Harvard University

BROAD
INSTITUTE

Marinka Zitnik
marinka@hms.harvard.edu

Phases of drug discovery from initial stage (target-to-hit) to final stage (launch)



p(TS) – probability of successful transition from one stage to the next; NME – new molecular entity; WIP – work in process

Drug-like chemical space
10^{60} chemical compounds

Drugs available to humans
~10^4 ●

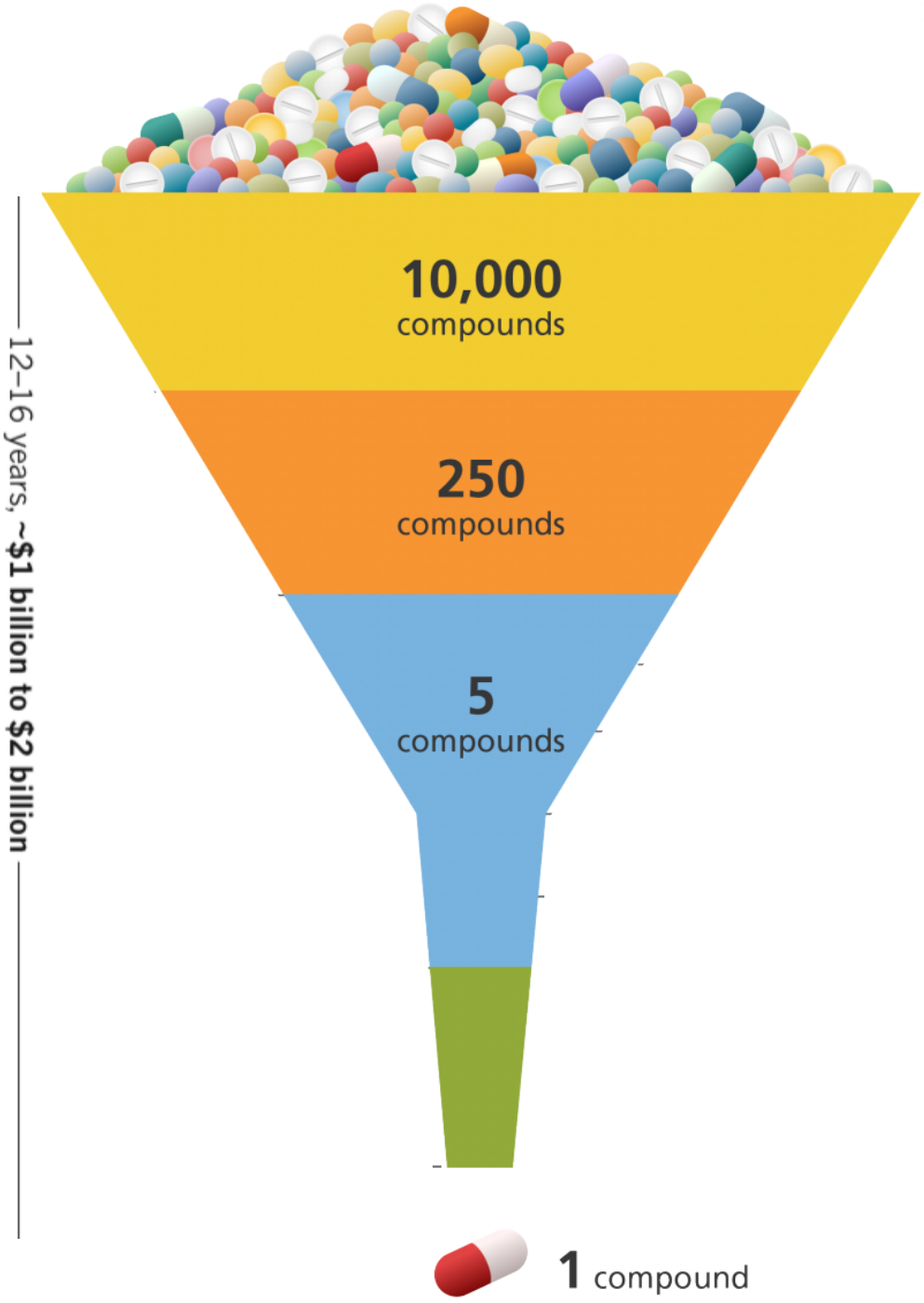
Step 1: Design and Discovery

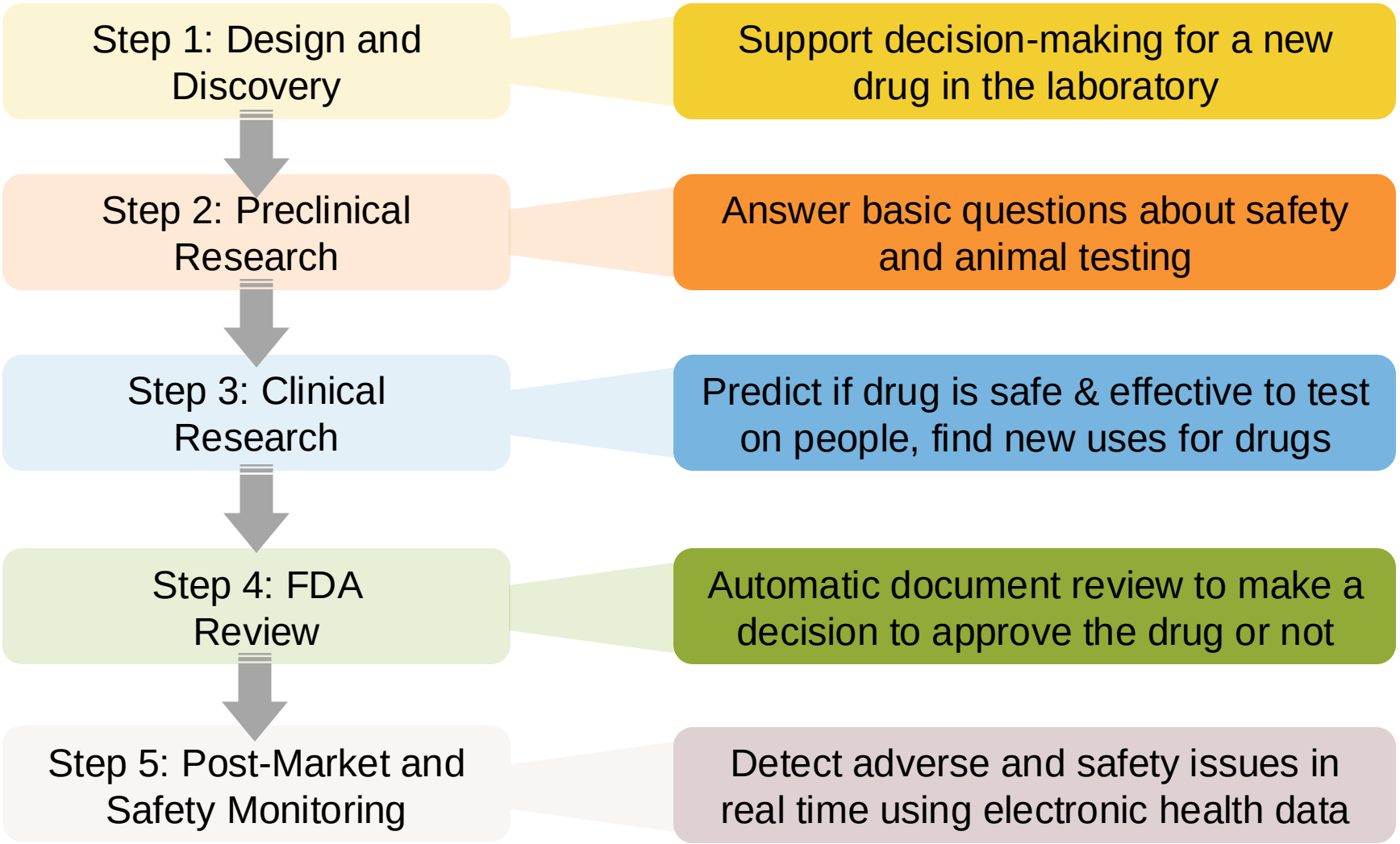
Step 2: Preclinical Research

Step 3: Clinical Research

Step 4: FDA Review

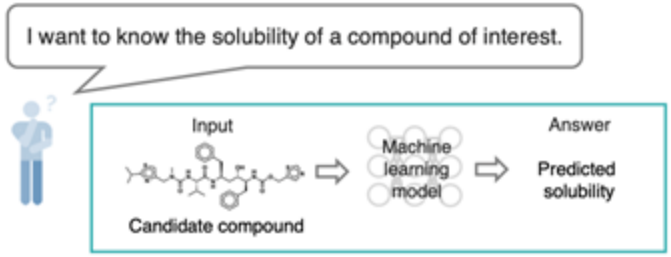
Step 5: Post-Market and Safety Monitoring



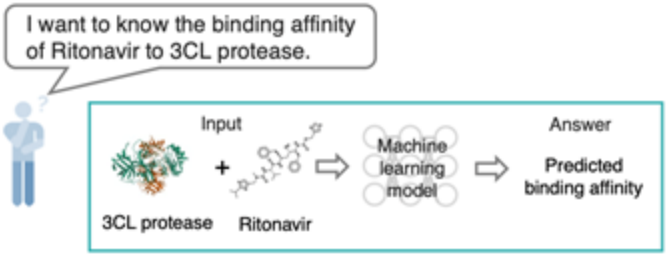

Step 1: Design and Discovery
Support decision-making for a new drug in the laboratory
Step 2: Preclinical Research
Answer basic questions about safety and animal testing
Step 3: Clinical Research
Predict if drug is safe & effective to test on people, find new uses for drugs
Step 4: FDA Review
Automatic document review to make a decision to approve the drug or not
Step 5: Post-Market and Safety Monitoring
Detect adverse and safety issues in real time using electronic health data

Outline for today's class

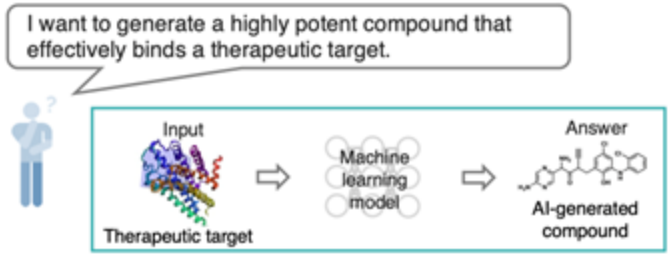
- Optimization & generation of small molecules



- Binding of drugs to therapeutic targets

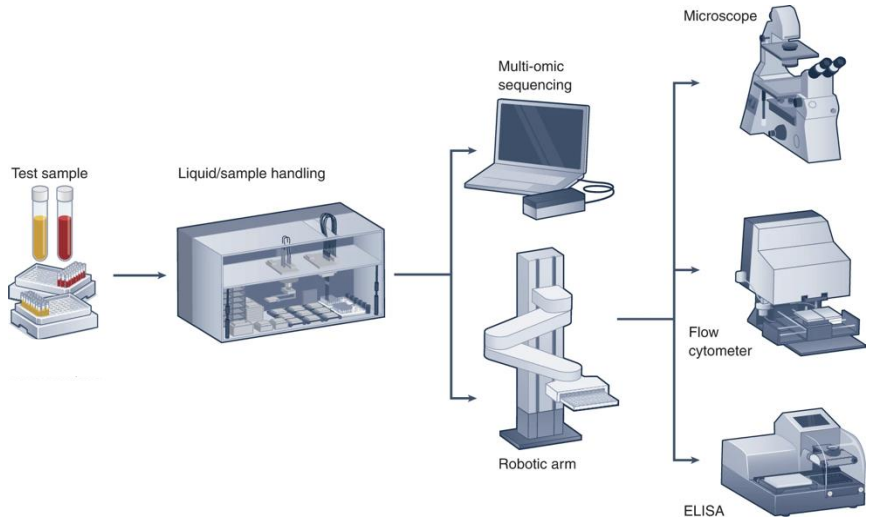


- High-throughput genetic & chemical perturbations



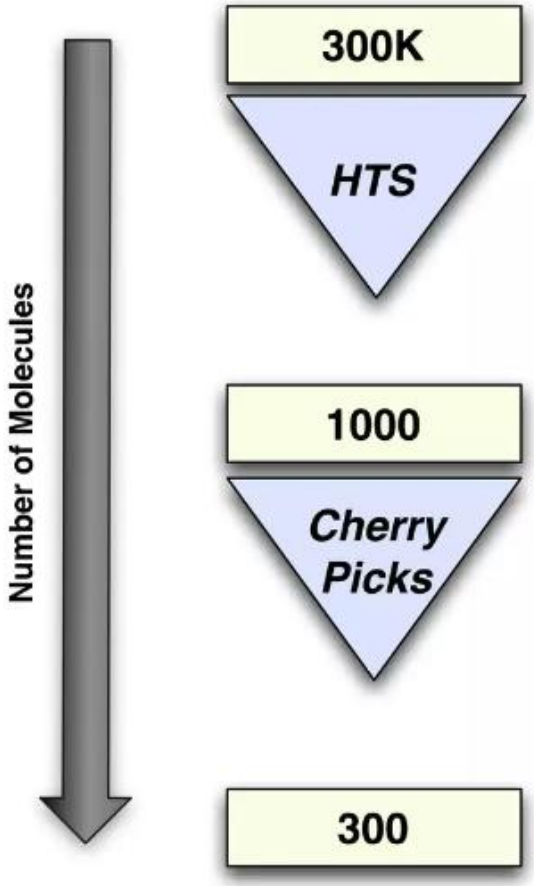
High throughput screening (HTS)

- Test thousands to hundreds of thousands of compounds in one or more assays
  - Biochemical, genetic, and pharmacological assays
- Integrate with robotics for self-driving lab
- **Goal:** Rapidly identify novel modulators of biological systems
  - Cellular basis of diseases
  - Therapeutic agents



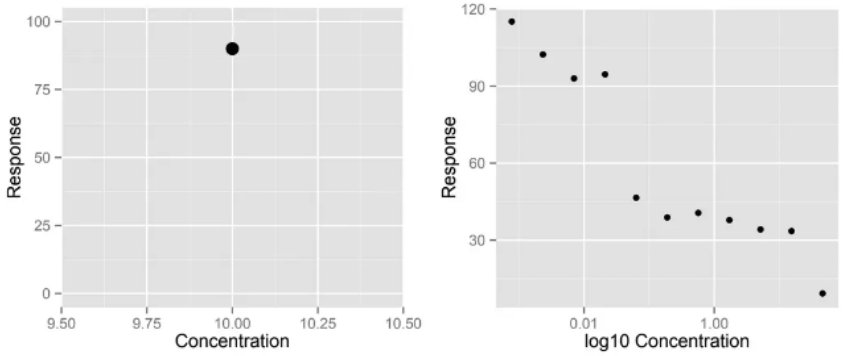
Goals of high throughput screening

- Rapidly screen large collections of compounds (chemical libraries)
- Efficiently identify active compounds
  - Test them in slower, accurate, expensive screens
- Use the data to learn what types of compounds tend to be active

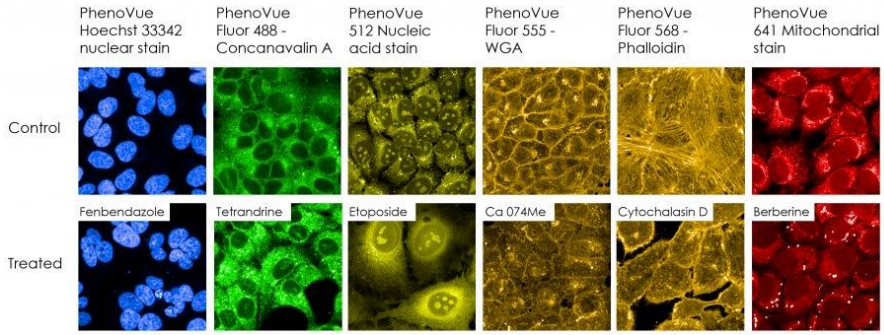


HTS data types

- Categorical: active/inactive or toxic/nontoxic
- Continuous: single-point or dose-response
- Multiple readouts:
  - Might read at different wavelengths or time points
  - More complex when dealing with images



Single-point vs. dose-response readouts



Cell painting for phenotypic drug discovery

HTS: Machine learning setup

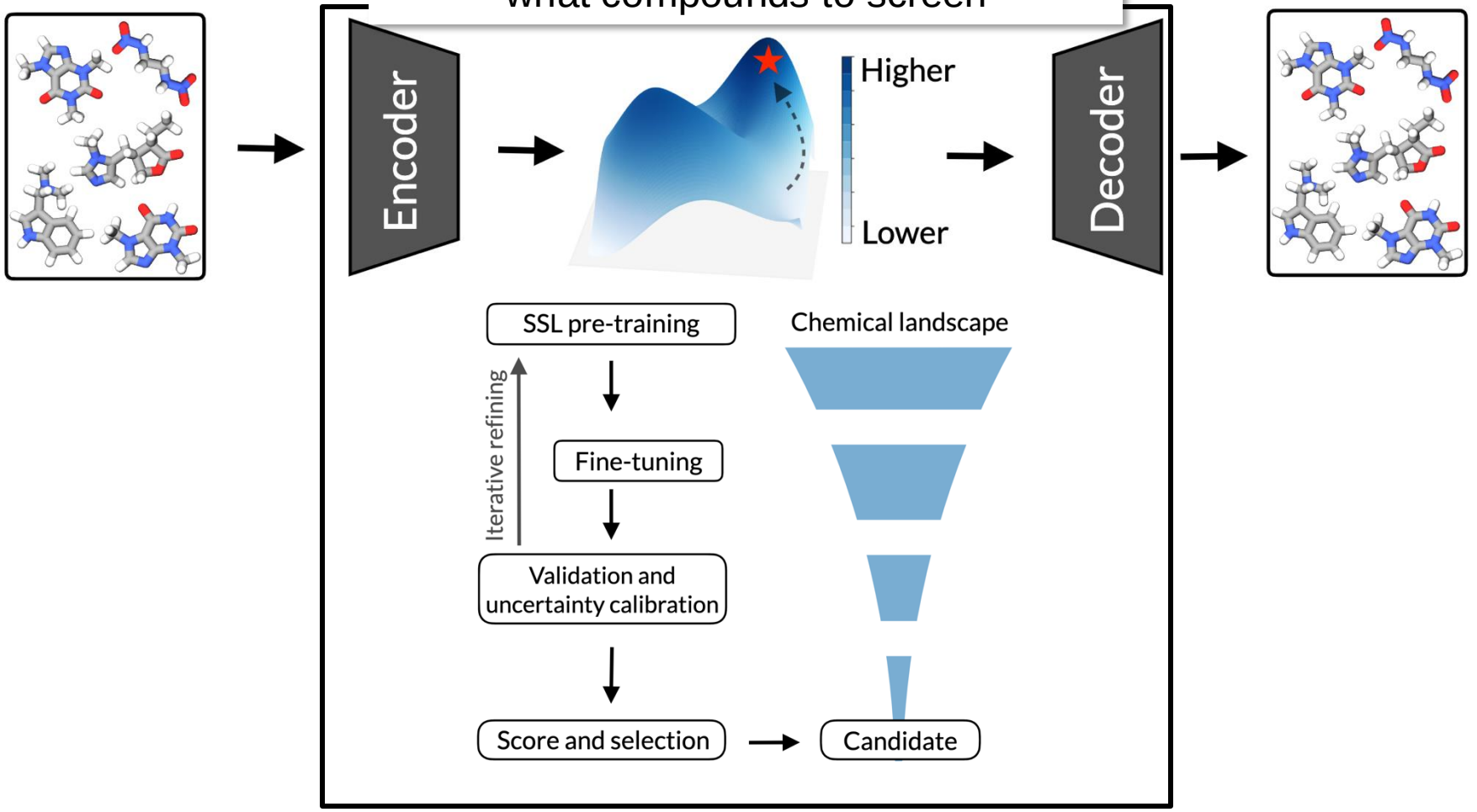
- HTS tests the activity of molecules:

$$Activity = f\left(Structure\right)$$

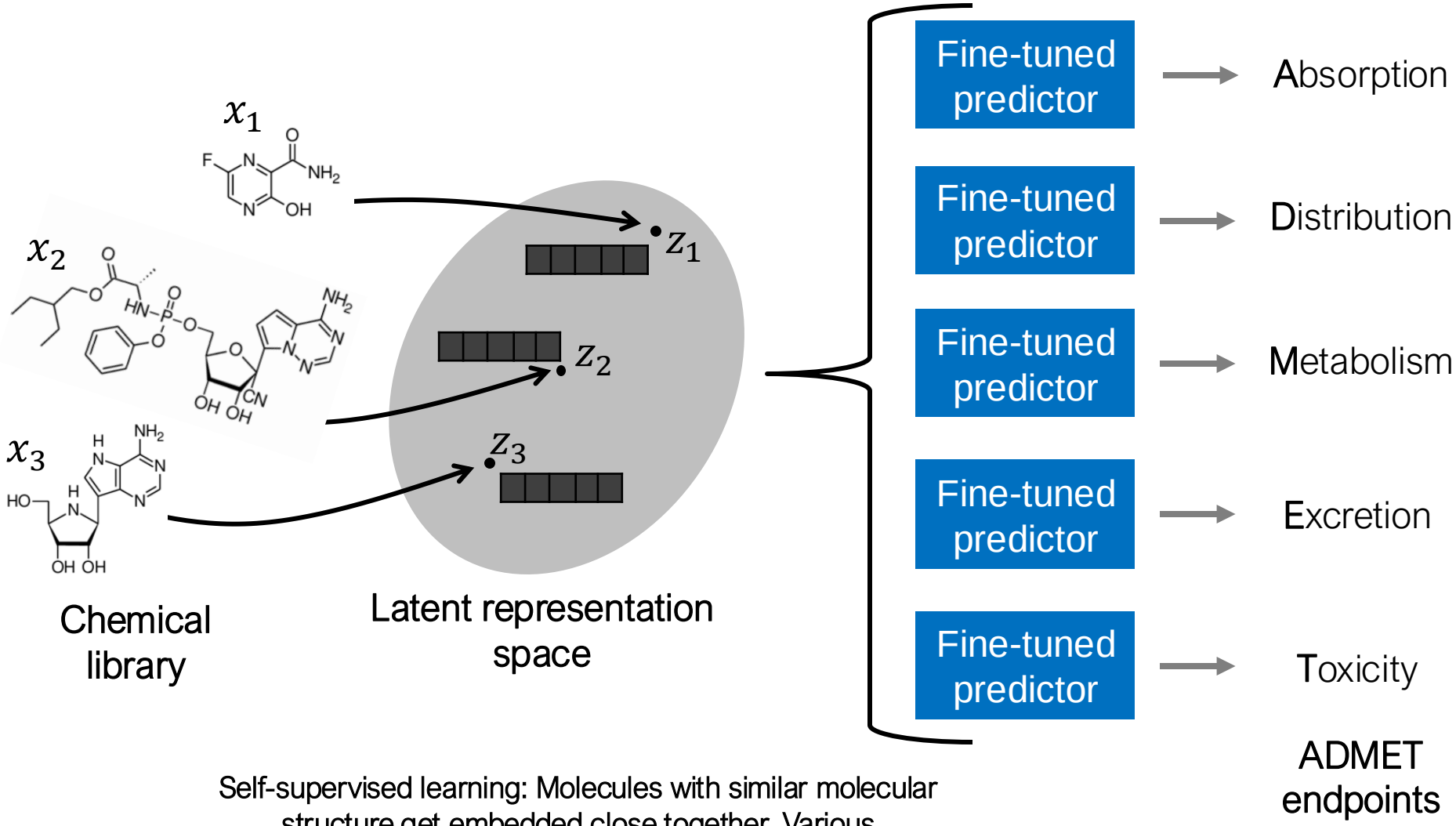
- We need to describe the molecular structure
  - Various discrete or real-valued descriptors
  - Surfaces (3D)
  - Binary fingerprints
  - Learned molecular embeddings

In-silico screening and optimization of molecular structure

Use computational models to suggest what compounds to screen



Molecular property prediction

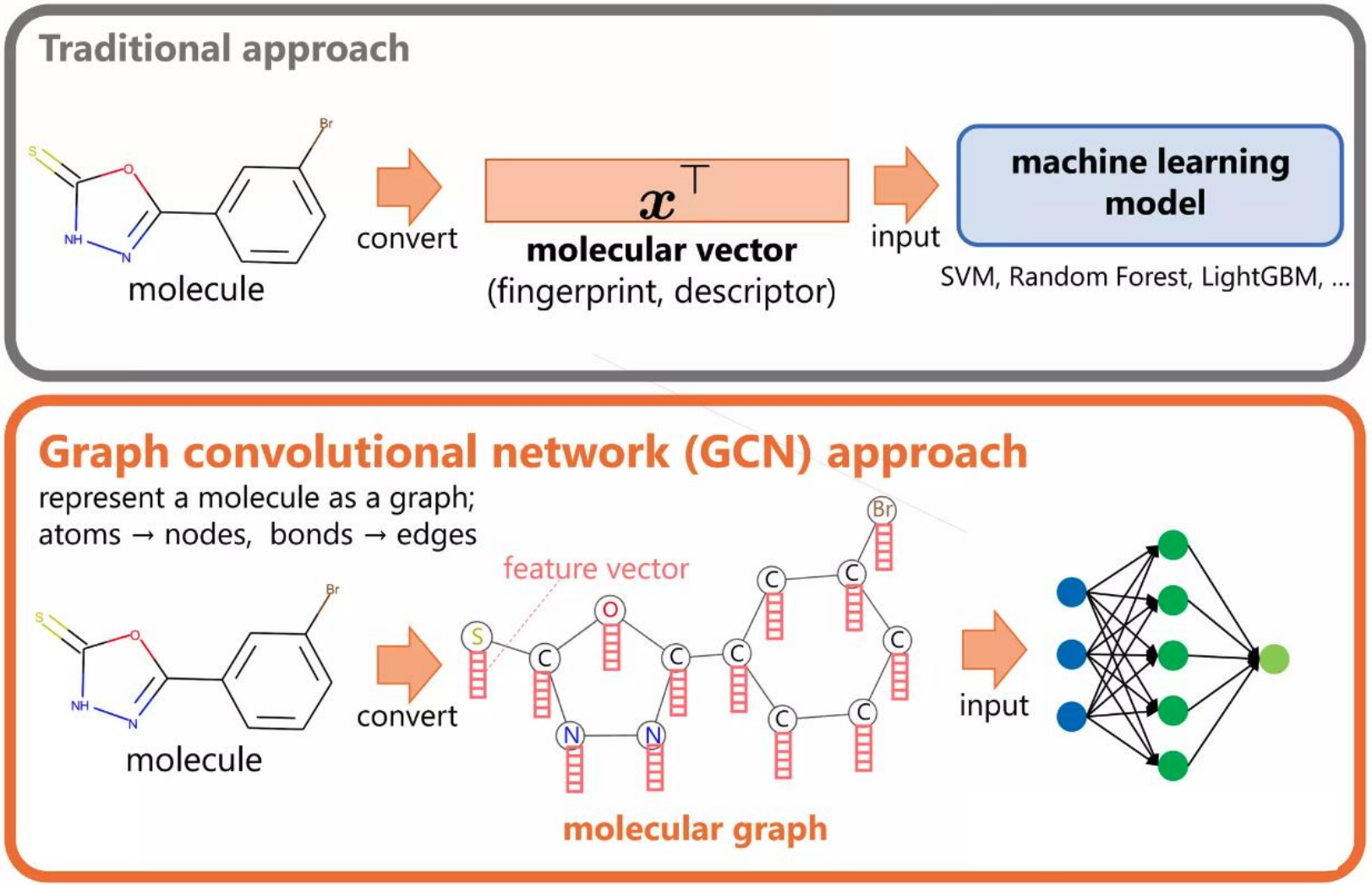


Self-supervised learning: Molecules with similar molecular structure get embedded close together. Various representations: Neural fingerprints, Attentive fingerprints, SMILES descriptors, Graphormer, Transformer-M, and others

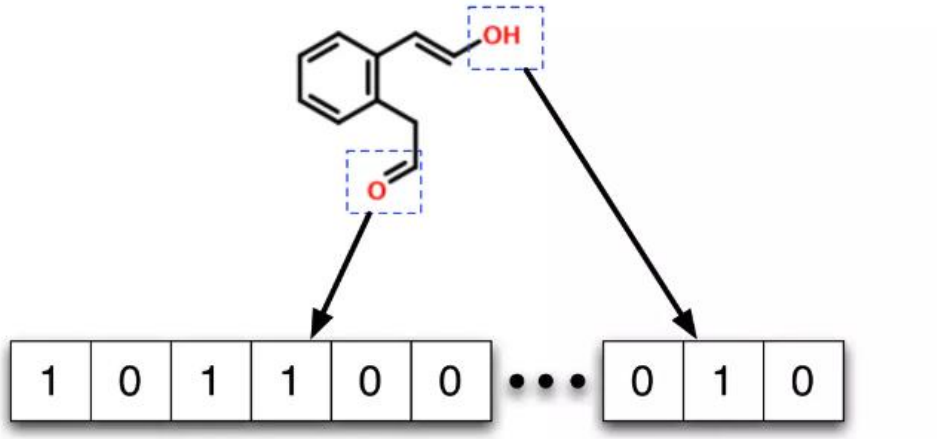
What can we use molecular representations for?

- **Search**
  - Given a potent active molecule, find similar ones (or dissimilar but also potent)
- **Prediction of various endpoints**
  - Given a set of active and inactive molecules, build a model to predict which members from a chemical library will be active
- **Clustering**
  - Given a set of molecules, do they cluster into structurally different groups?

Two strategies for producing molecular representations



Fingerprint representations



- Lots of types of fingerprints
- Keyed fingerprints indicate the presence or absence of a structural feature
- Length can vary from 166 to 4096 bits or more
- Fingerprints usually compared to each other using the Tanimoto metric

Towards neural fingerprints

Algorithm 1 Circular fingerprints

Input: molecule, radius R, fingerprint length S
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow [\mathbf{r}_a, \mathbf{r}_1, \dots, \mathbf{r}_N]$ ▷ concatenate
    $\mathbf{r}_a \leftarrow \text{hash}(\mathbf{v})$ ▷ hash function
    $i \leftarrow \text{mod}(r_a, S)$ ▷ convert to index
    $\mathbf{f}_i \leftarrow 1$ ▷ Write 1 at index
Return: binary vector $\mathbf{f}$

Algorithm 2 Neural graph fingerprints

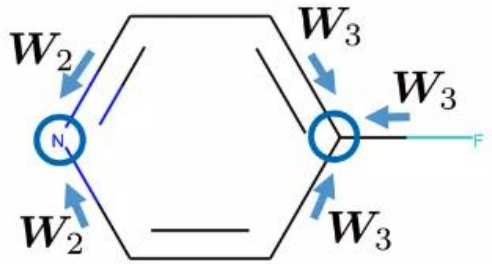
Input: molecule, radius R, hidden weights $H_1^1 \dots H_R^5$, output weights $W_1 \dots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in molecule
    $\mathbf{r}_1 \dots \mathbf{r}_N = \text{neighbors}(a)$
    $\mathbf{v} \leftarrow \mathbf{r}_a + \sum_{i=1}^{N} \mathbf{r}_i$ ▷ sum
    $\mathbf{r}_a \leftarrow \sigma(\mathbf{v} H_L^N)$ ▷ smooth function
    $\mathbf{i} \leftarrow \text{softmax}(\mathbf{r}_a W_L)$ ▷ sparsify
    $\mathbf{f} \leftarrow \mathbf{f} + \mathbf{i}$ ▷ add to fingerprint
Return: real-valued vector $\mathbf{f}$

Figure 2: Pseudocode of circular fingerprints (*left*) and neural graph fingerprints (*right*). Differences are highlighted in blue. Every non-differentiable operation is replaced with a differentiable analog.

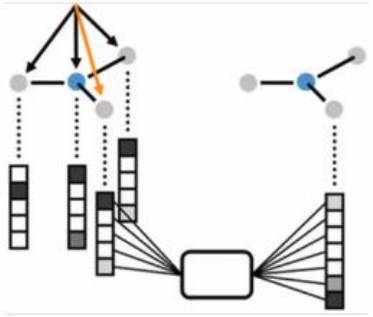
Duvenaud et al., NeurIPS 2015

Neural fingerprint representations

1) **Neural graph fingerprints**
   - Generate molecular fingerprints with a neural network
   - Update atom features using only adjacent atoms
   - Use different weights for node degrees



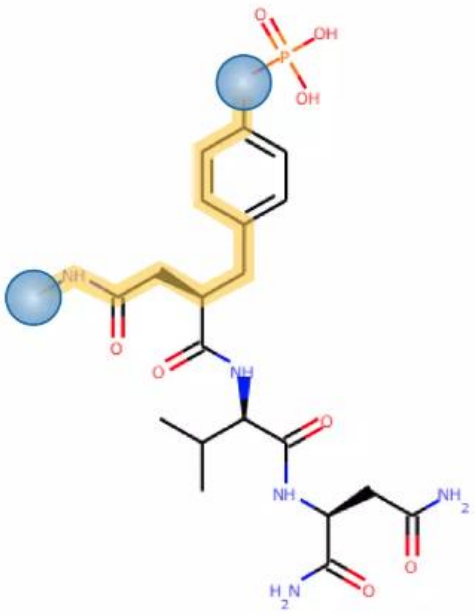
2) **Molecular graphs**
   - Update atom features by convolutional and pooling layers using adjacent atoms



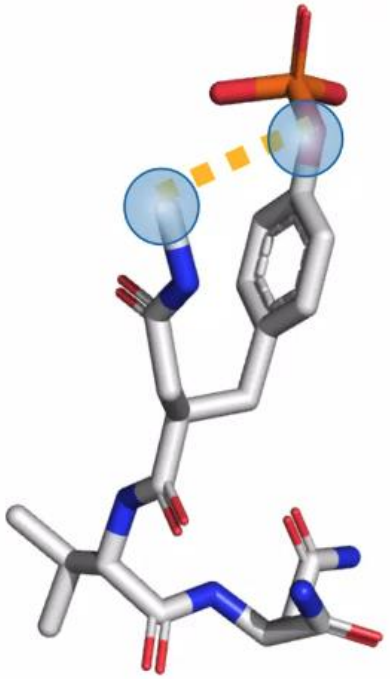
- **They did not consider property of edges (bonds)**
- **They did not consider atoms other than 1-neighbor**

Duvenaud et al., NeurIPS 2015; Altae Tran et al., ACS Central Science 2017

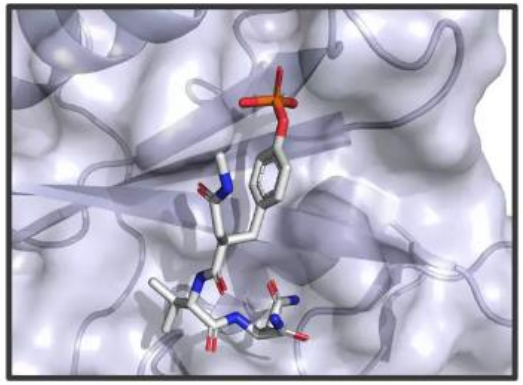
Graphs vs. 3D structures



Molecular Graph



3D Structure



The distance on the graph does not necessarily correlate with the Euclidean distance between atoms in the 3D structure

Need to consider modifying the definition of graph distance

Datasets

22 datasets with ADMET endpoints

A: Absorption
Caco2 (Cell Permeability)
HIA (Intestinal Absorption)
Pgp (P-glycoprotein)
Bioavailability
Lipophilicity
Solubility

D: Distribution
BBB (Blood-Brain Barrier)
PPBR (Plasma Protein Binding)
VDss (Volume of Distribution)

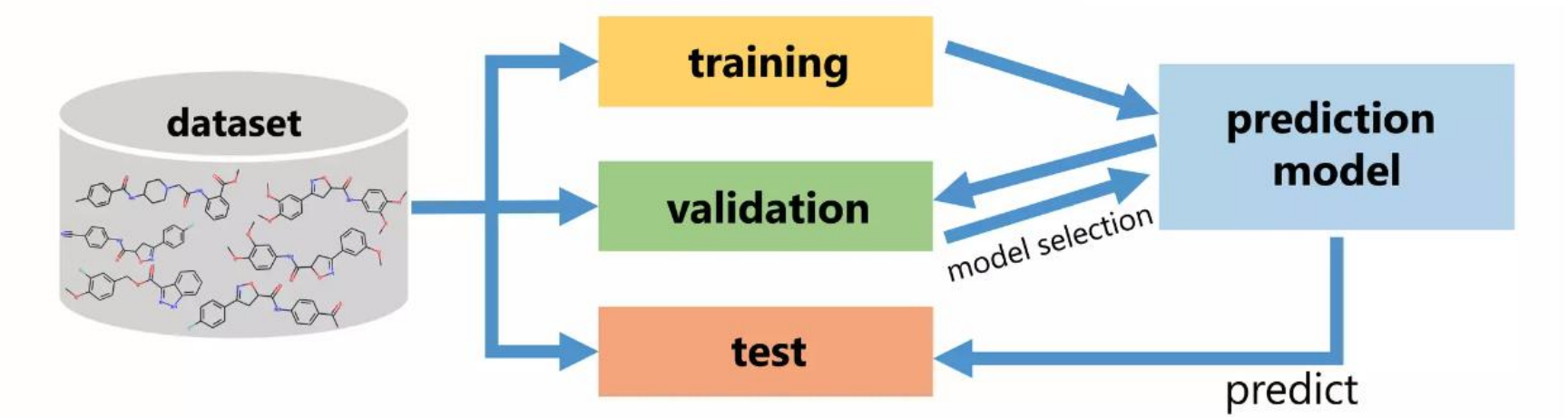
M: Metabolism
CYP2C9/2D6/3A4 Inhibition
CYP2C9/2D6/3A4 Substrate

E: Excretion
Half Life
Clearance (Hepatocyte)
Clearance (Microsome)

T: Toxicity
LD50 (Acute Toxicity)
hERG blocker
Ames Mutagenicity
Drug Induced Liver Injury

Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology,* 2022

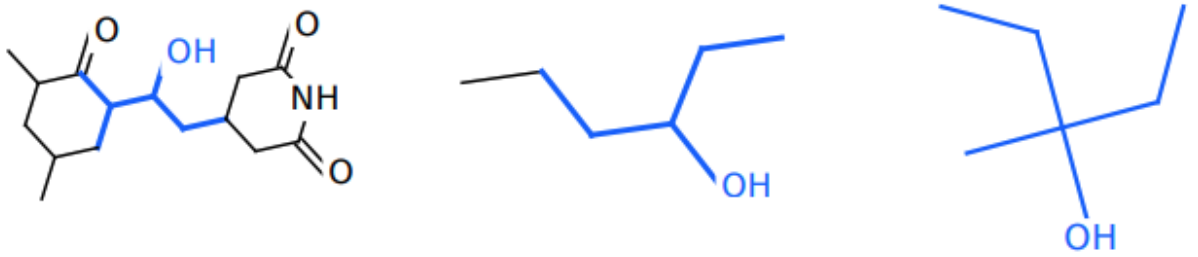
Experimental setup



- Demonstrate that fingerprints are interpretable
  - Show substructures which most activate individual features in a fingerprint vector
  - **Fingerprint features** can each only be activated by a single fragment of a single radius, except for accidental collisions
  - In contrast, **neural fingerprint features** can be activated by variations of the same structure, making them more interpretable, and allowing shorter feature vectors.

Results: Examining neural fingerprints

Fragments most activated by pro-solubility feature



Fragments most activated by anti-solubility feature

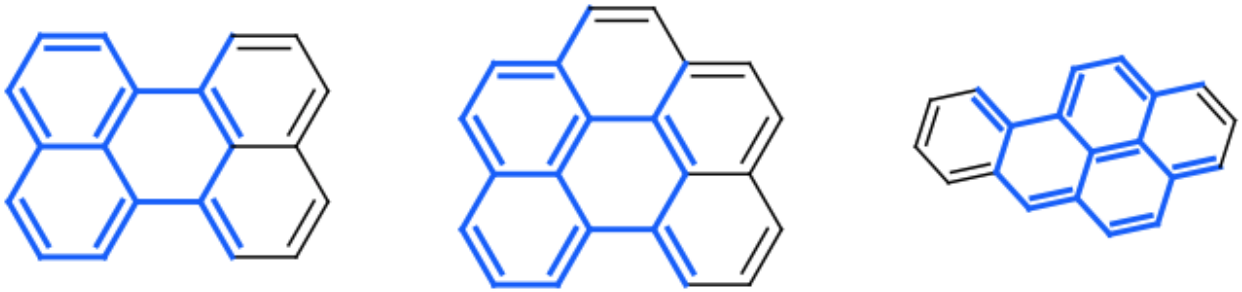
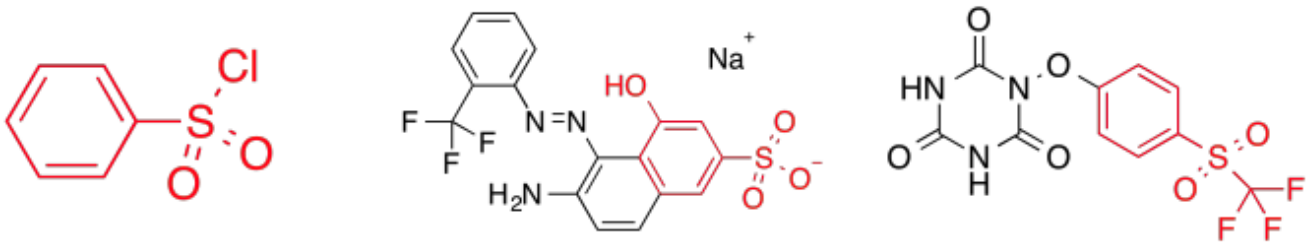


Figure 4: Examining fingerprints optimized for predicting solubility. Shown here are representative examples of molecular fragments (highlighted in blue) which most activate different features of the fingerprint. *Top row:* The feature most predictive of solubility. *Bottom row:* The feature most predictive of insolubility.

Duvenaud et al., NeurIPS 2015

Results: Examining neural fingerprints

Fragments most activated by toxicity feature on SR-MMP dataset



Fragments most activated by toxicity feature on NR-AHR dataset

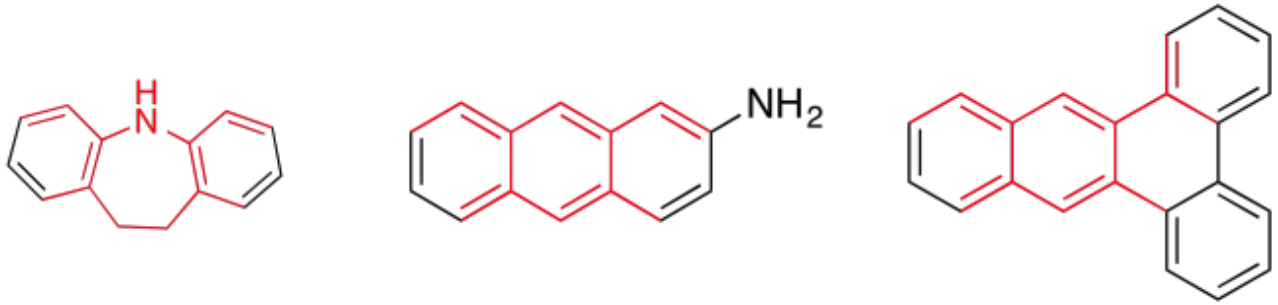


Figure 5: Visualizing fingerprints optimized for predicting toxicity. Shown here are representative samples of molecular fragments (highlighted in red) which most activate the feature most predictive of toxicity. *Top row:* the most predictive feature identifies groups containing a sulphur atom attached to an aromatic ring. *Bottom row:* the most predictive feature identifies fused aromatic rings, also known as polycyclic aromatic hydrocarbons, a well-known carcinogen.

Duvenaud et al., NeurIPS 2015

Results: Molecular property prediction

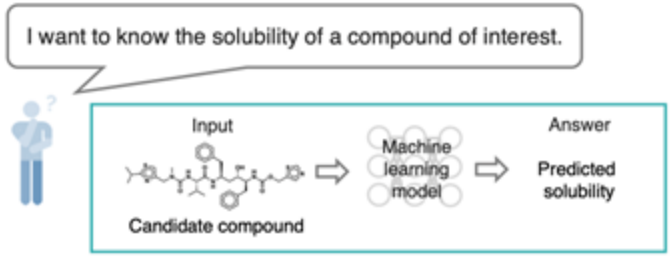
Raw Feature Type		Expert-Curated Methods		SMILES	Molecular Graph-Based Methods (state-of-the-Art in ML)				
Dataset	Metric	Morgan	RDKit2D	CNN	NeuralFP	GCN	AttentiveFP	AttrMasking	ContextPred
	# Params.	1477K	633K	227K	480K	192K	301K	2067K	2067K
TDC.Caco2 (↓)	MAE	0.908±0.060	**0.393±0.024**	0.446±0.036	0.530±0.102	0.599±0.104	0.401±0.032	0.546±0.052	0.502±0.036
TDC.HIA (↑)	AUROC	0.807±0.072	0.972±0.008	0.869±0.026	0.943±0.014	0.936±0.024	0.974±0.007	**0.978±0.006**	0.975±0.004
TDC.Pgp (↑)	AUROC	0.880±0.006	0.918±0.007	0.908±0.012	0.902±0.020	0.895±0.021	0.892±0.012	**0.929±0.006**	0.923±0.005
TDC.Bioav (↑)	AUROC	0.581±0.086	**0.672±0.021**	0.613±0.013	0.632±0.036	0.566±0.115	0.632±0.039	0.577±0.087	0.671±0.026
TDC.Lipo (↓)	MAE	0.701±0.009	0.574±0.017	0.743±0.020	0.563±0.023	0.541±0.011	0.572±0.007	0.547±0.024	**0.535±0.012**
TDC.AqSol (↓)	MAE	1.203±0.019	0.827±0.047	1.023±0.023	0.947±0.016	0.907±0.020	**0.776±0.008**	1.026±0.020	1.040±0.045
TDC.BBB (↑)	AUROC	0.823±0.015	0.889±0.016	0.781±0.030	0.836±0.009	0.842±0.016	0.855±0.011	0.892±0.012	**0.897±0.004**
TDC.PPBR (↓)	MAE	12.848±0.362	9.994±0.319	11.106±0.358	**9.292±0.384**	10.194±0.373	9.373±0.335	10.075±0.202	9.445±0.224
TDC.VD (↑)	Spearman	0.493±0.011	**0.561±0.025**	0.226±0.114	0.258±0.162	0.457±0.050	0.241±0.145	0.559±0.019	0.485±0.092
TDC.CYP2D6-I (↑)	AUPRC	0.587±0.011	0.616±0.007	0.544±0.053	0.627±0.009	0.616±0.020	0.646±0.014	0.721±0.009	**0.739±0.005**
TDC.CYP3A4-I (↑)	AUPRC	0.827±0.009	0.829±0.007	0.821±0.003	0.849±0.004	0.840±0.010	0.851±0.006	0.902±0.002	**0.904±0.002**
TDC.CYP2C9-I (↑)	AUPRC	0.715±0.004	0.742±0.006	0.713±0.006	0.739±0.010	0.735±0.004	0.749±0.004	0.829±0.003	**0.839±0.003**
TDC.CYP2D6-S (↑)	AUPRC	0.671±0.066	0.677±0.047	0.485±0.037	0.572±0.062	0.617±0.039	0.574±0.030	0.704±0.028	**0.736±0.024**
TDC.CYP3A4-S (↑)	AUROC	0.633±0.013	0.639±0.012	**0.662±0.031**	0.578±0.020	0.590±0.023	0.576±0.025	0.582±0.021	0.609±0.025
TDC.CYP2C9-S (↑)	AUPRC	0.380±0.015	0.360±0.040	0.367±0.059	0.359±0.059	0.344±0.051	0.375±0.032	0.381±0.045	**0.392±0.026**
TDC.Half_Life (↑)	Spearman	**0.329±0.083**	0.184±0.111	0.038±0.138	0.177±0.165	0.239±0.100	0.085±0.068	0.151±0.068	0.129±0.114
TDC.CL-Micro (↑)	Spearman	0.492±0.020	**0.586±0.014**	0.252±0.116	0.529±0.015	0.532±0.033	0.365±0.055	0.585±0.034	0.578±0.007
TDC.CL-Hepa (↑)	Spearman	0.272±0.068	0.382±0.007	0.235±0.021	0.401±0.037	0.366±0.063	0.289±0.022	0.413±0.028	**0.439±0.026**
TDC.hERG (↑)	AUROC	0.736±0.023	**0.841±0.020**	0.754±0.037	0.722±0.034	0.738±0.038	0.825±0.007	0.778±0.046	0.756±0.023
TDC.AMES (↑)	AUROC	0.794±0.008	0.823±0.011	0.776±0.015	0.823±0.006	0.818±0.010	0.814±0.008	**0.842±0.008**	0.837±0.009
TDC.DILI (↑)	AUROC	0.832±0.021	0.875±0.019	0.792±0.016	0.851±0.026	0.859±0.033	0.886±0.015	**0.919±0.008**	0.861±0.018
TDC.LD50 (↓)	MAE	0.649±0.019	0.678±0.003	0.675±0.011	0.667±0.020	0.649±0.026	0.678±0.012	**0.685±0.025**	0.669±0.030

- No single method performs the best across all scenarios
- Pre-training boost performance
- Pre-trained graph models yield strongest predictors overall

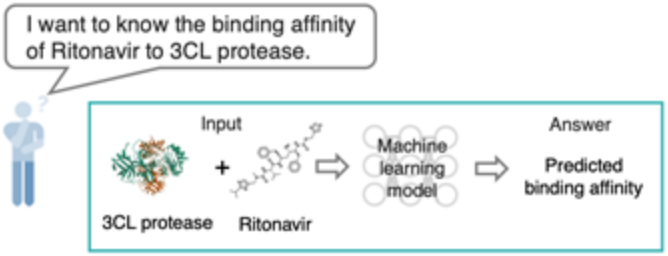
Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology,* 2022

Outline for today's class

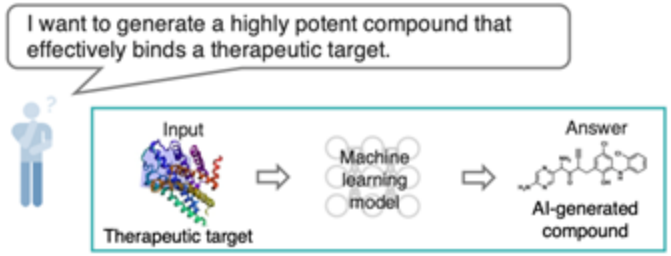
- Optimization & generation of small molecules



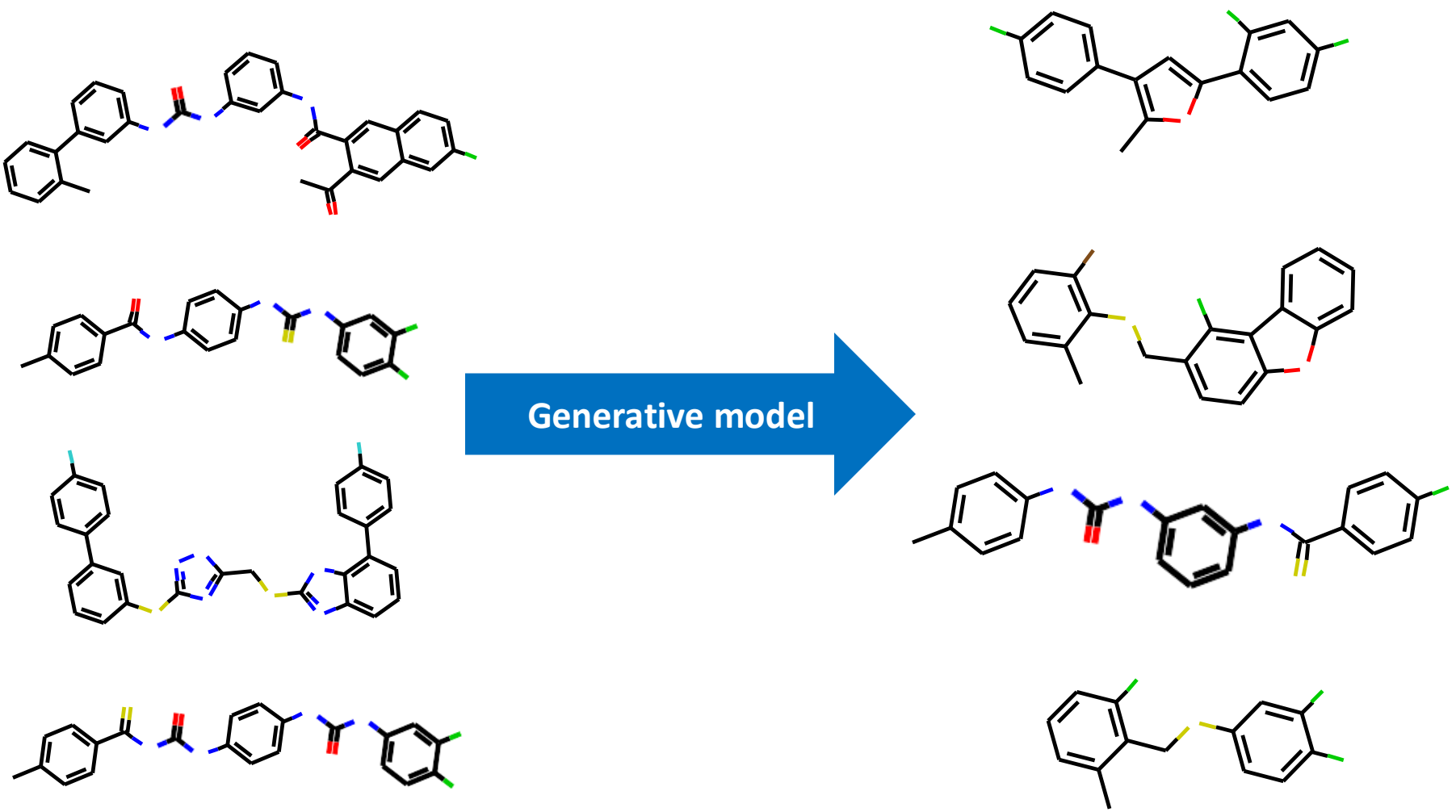
- Binding of drugs to therapeutic targets



- High-throughput genetic & chemical perturbations

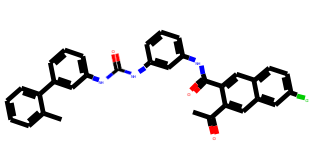


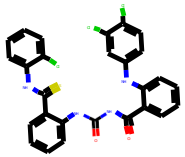
Molecular graph generation

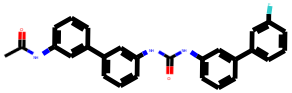


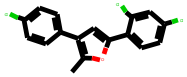
Details and description of other models at https://zitniklab.hms.harvard.edu/drugml

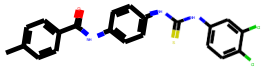
Molecular graph generation

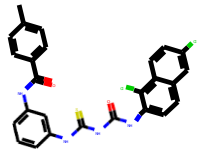

5.30

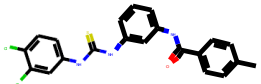

4.93

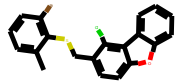

4.49

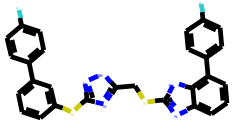

4.45

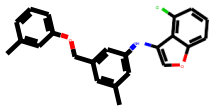

4.42

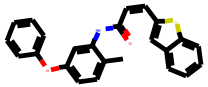

4.40

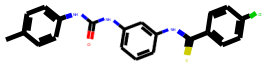

4.37

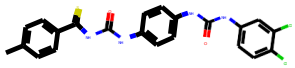

4.30

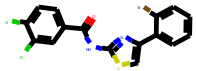

4.23

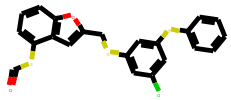

4.18

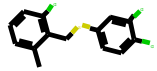

4.17


4.08


4.07

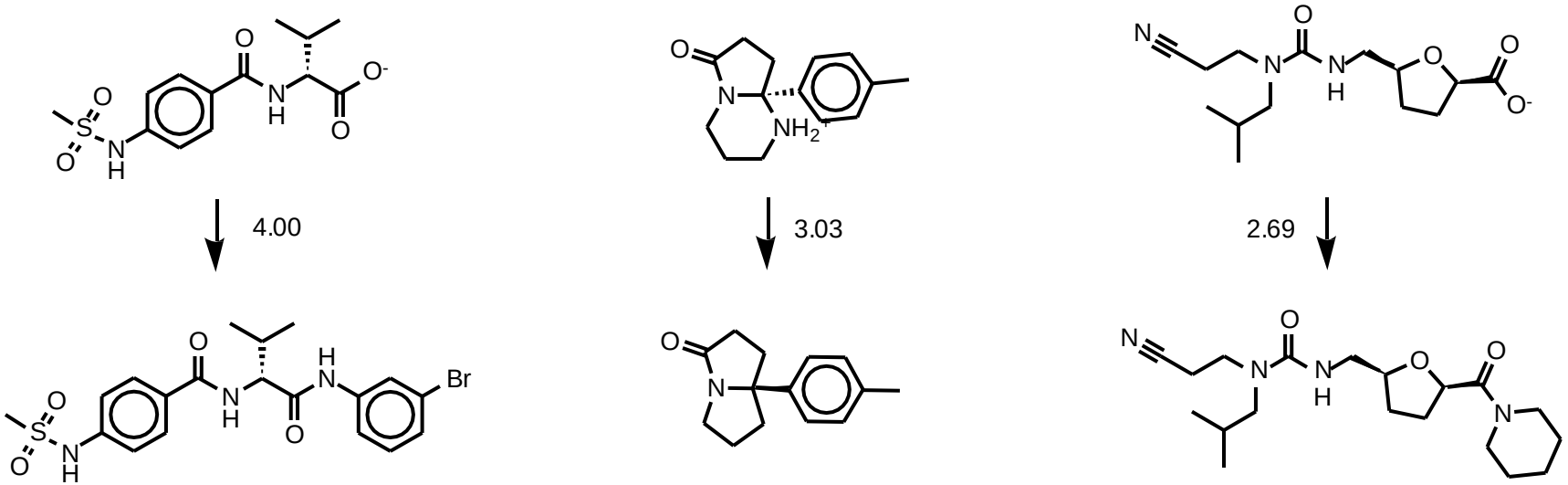

4.04


4.04


4.03

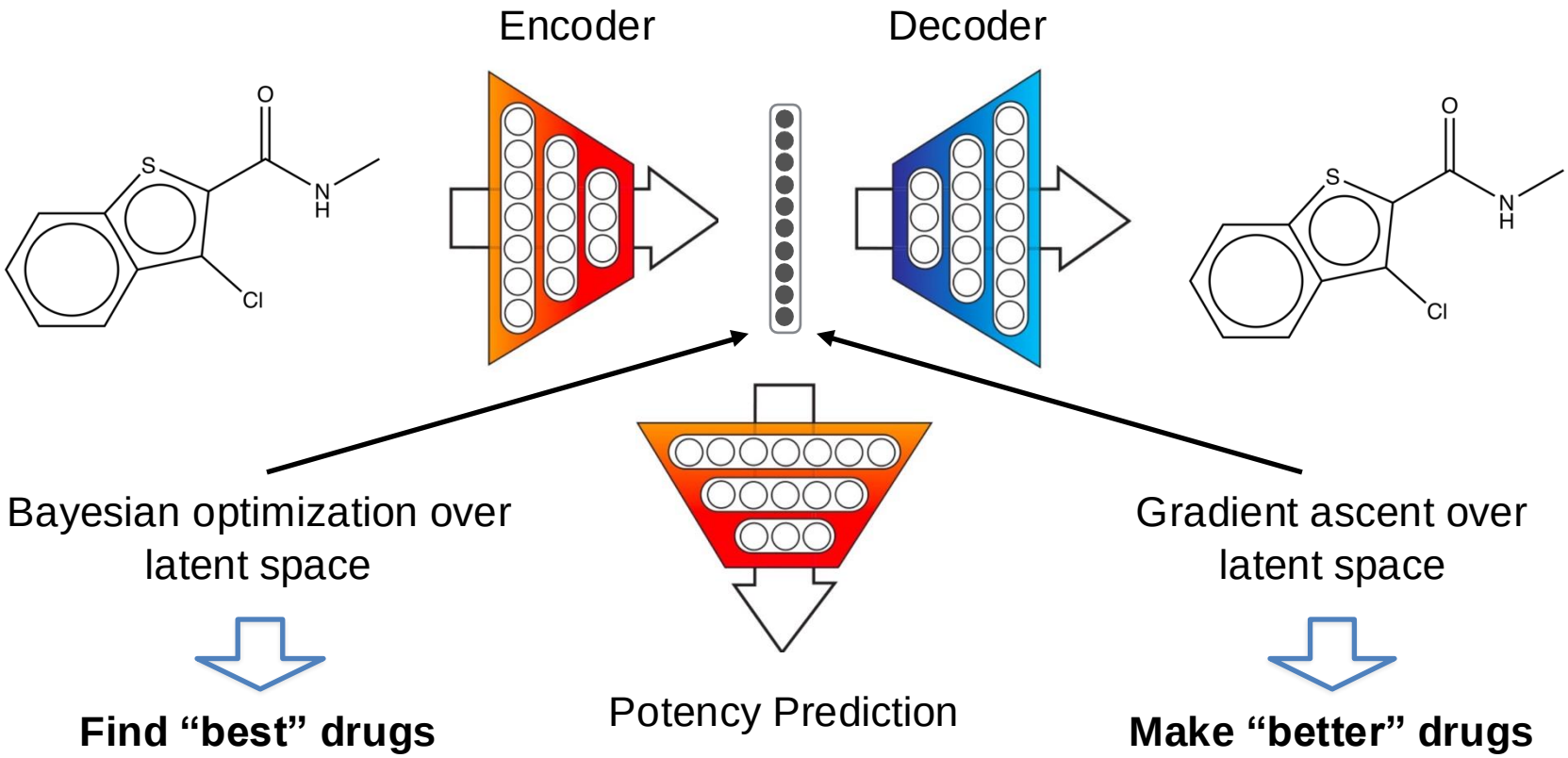
Generate molecules with high potency

Molecular graph generation



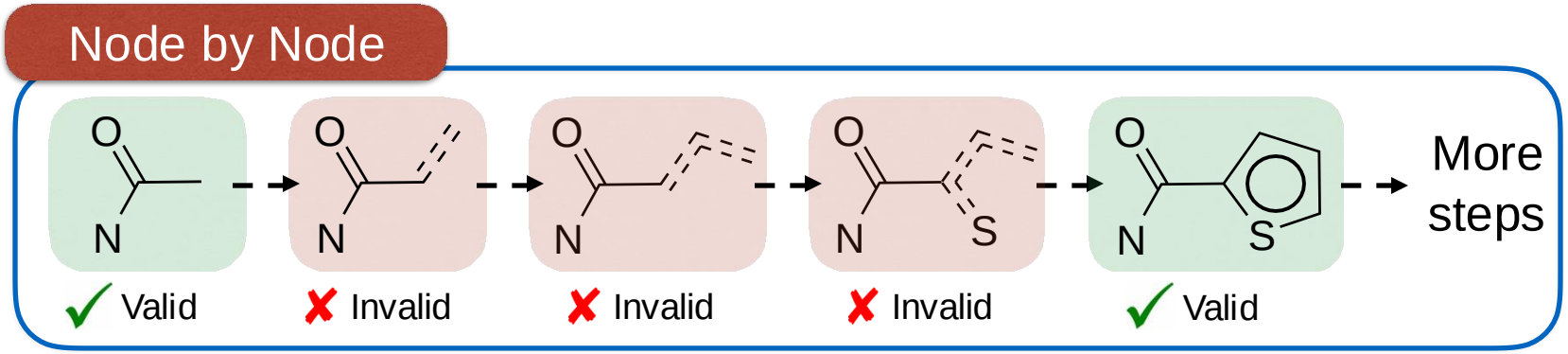
Modify molecules to increase potency

Molecular variational autoencoder



[1] Gomez-Bombarelli et al.,Automatic chemical design using a data-driven continuous representation of molecules, 2016

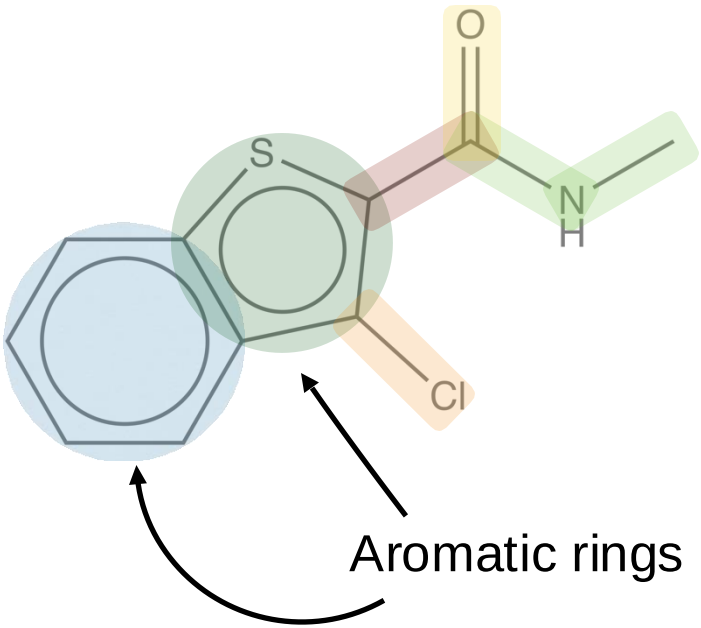
How to generate graphs?

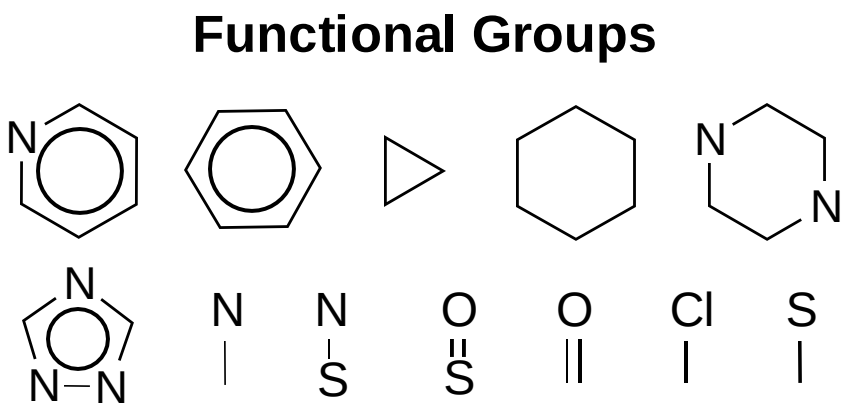


- Not every graphs is chemically valid
- Invalid intermediate states ⟶ hard to validate
- Very long intermediate steps ⟶ difficult to train (Li et al., 2018)

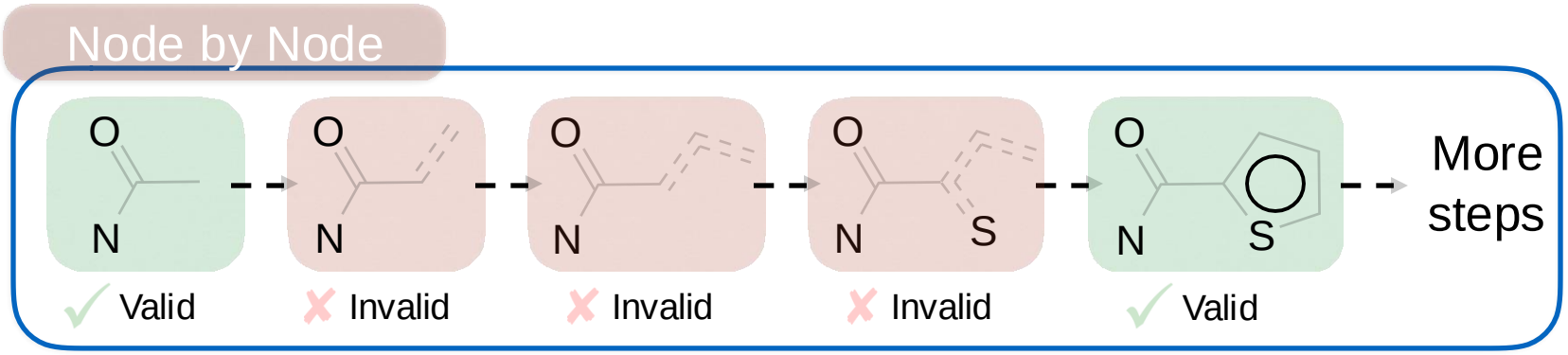
[2] Li et al., Learning Deep Generative Models of Graphs, 2018

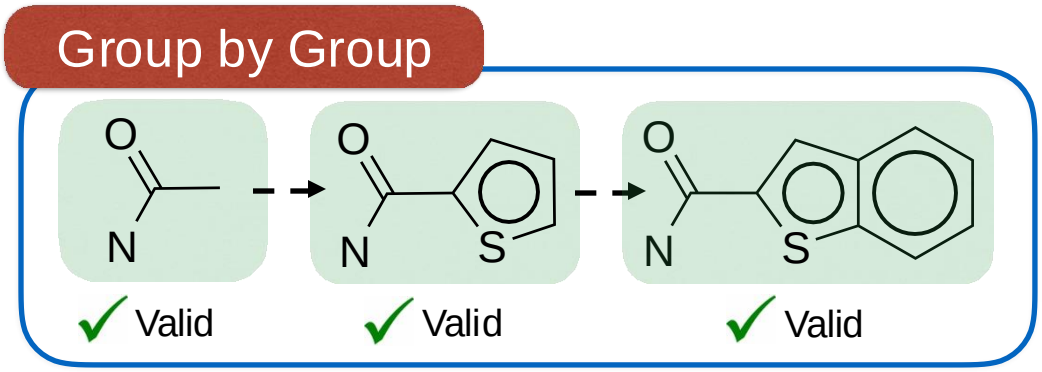
Functional groups





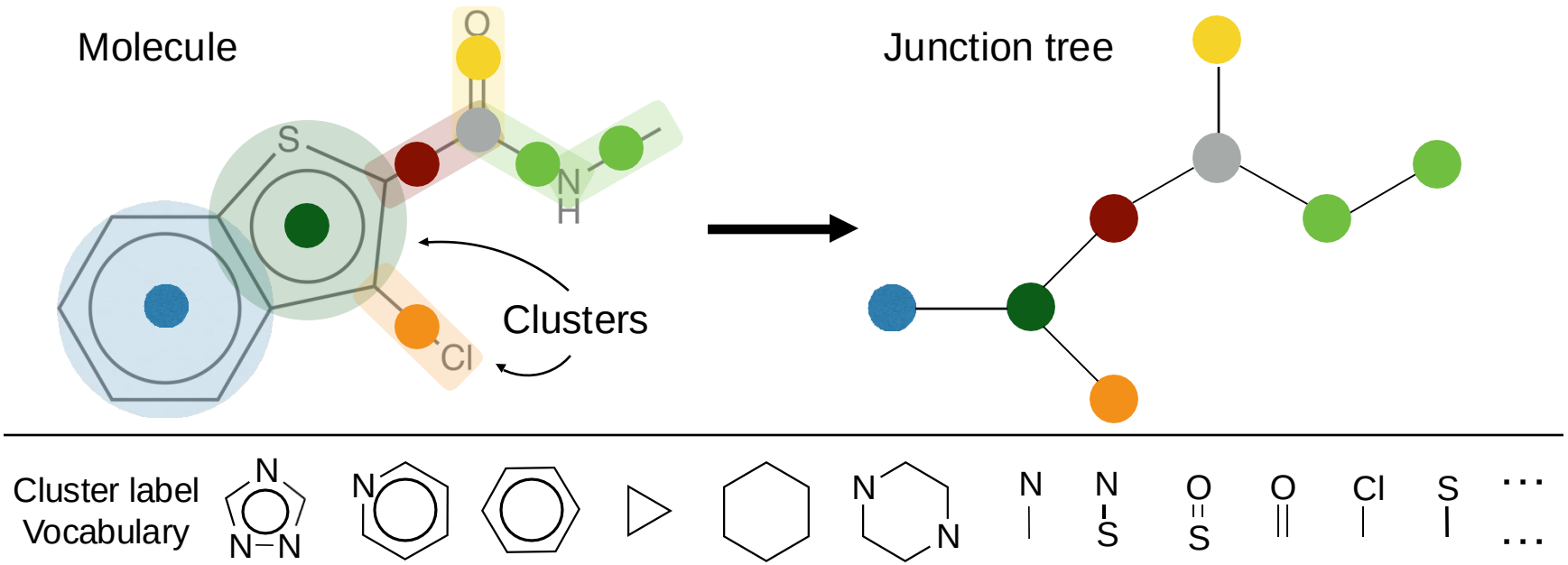
How to generate graphs?





- Shorter action sequence
- Easy to check validity

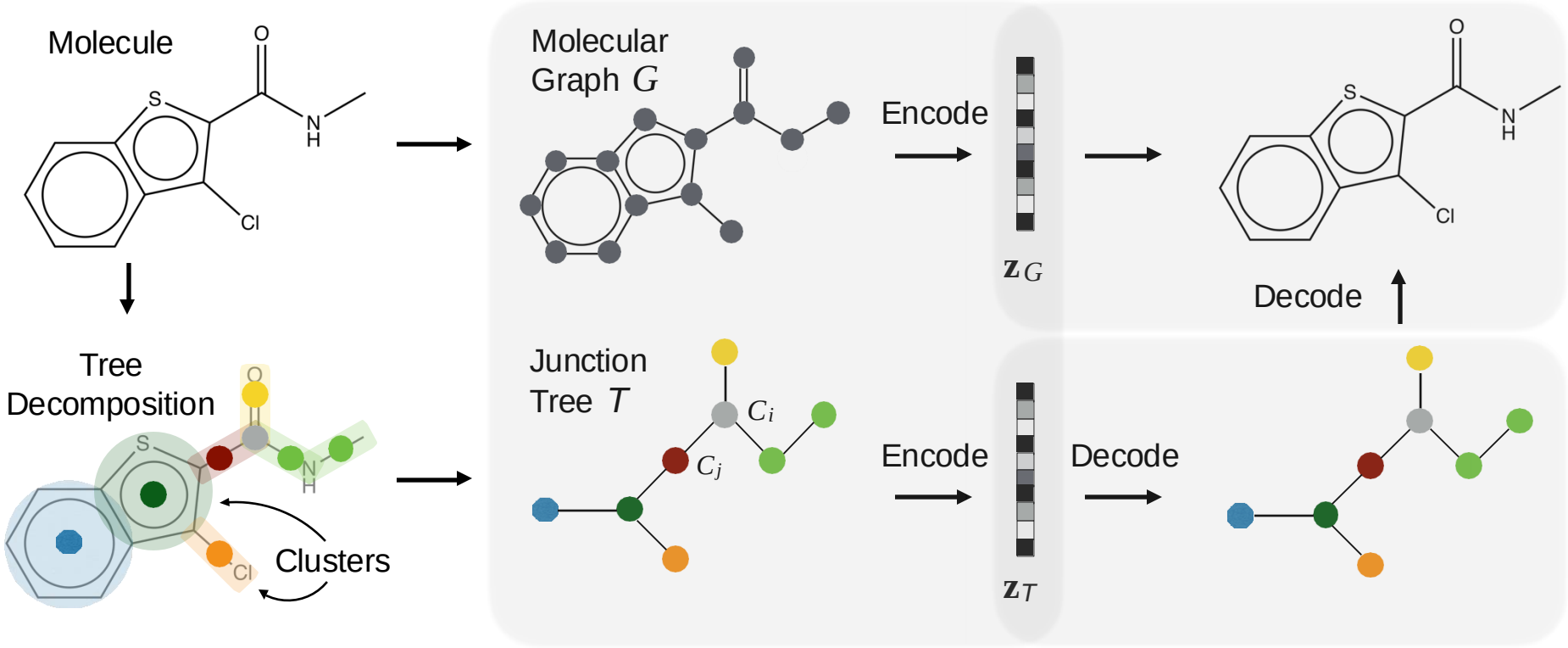
Tree decomposition



- Generate junction tree → Generate graph group by group
- Vocabulary size: less than 800 given 250K molecules

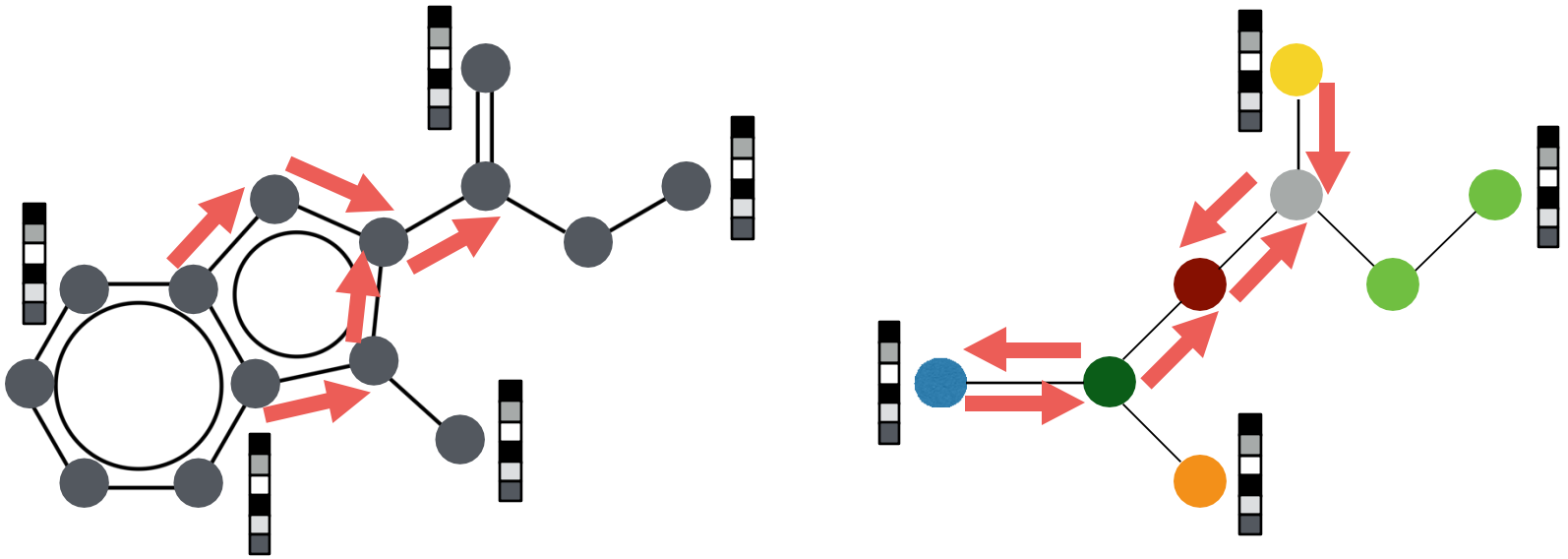
Jin et al., ICML 2018

Approach: Junction-tree variational autoencoder



Jin et al., ICML 2018

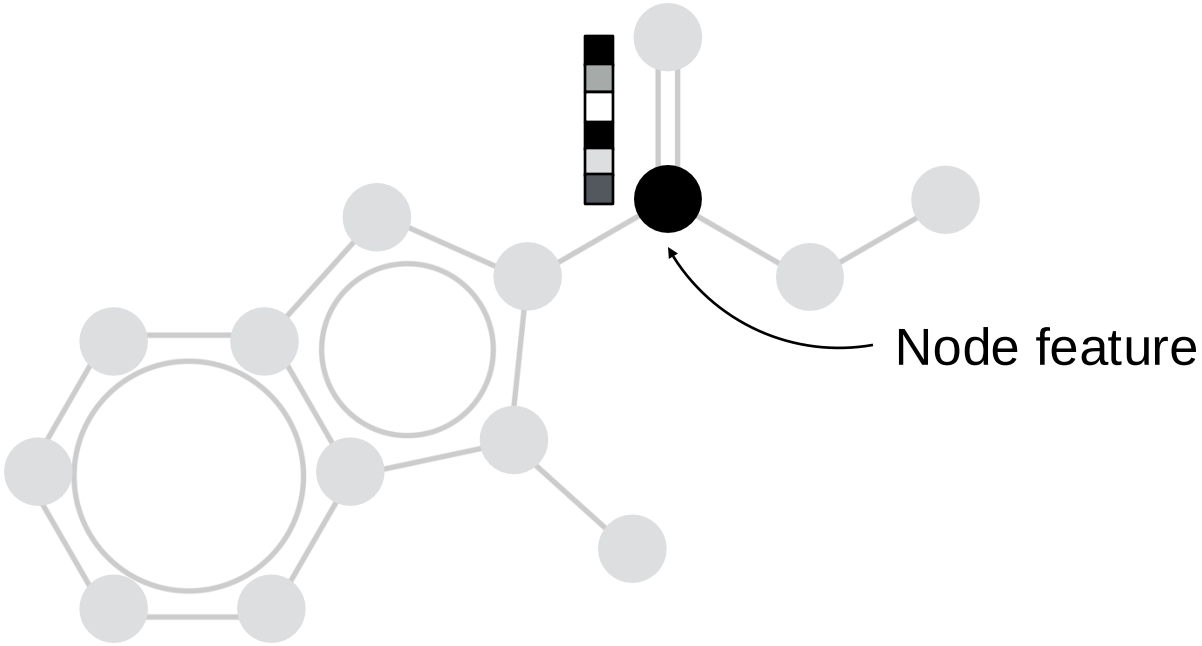
Graph and tree encoders



Neural Message Passing Network (MPN)

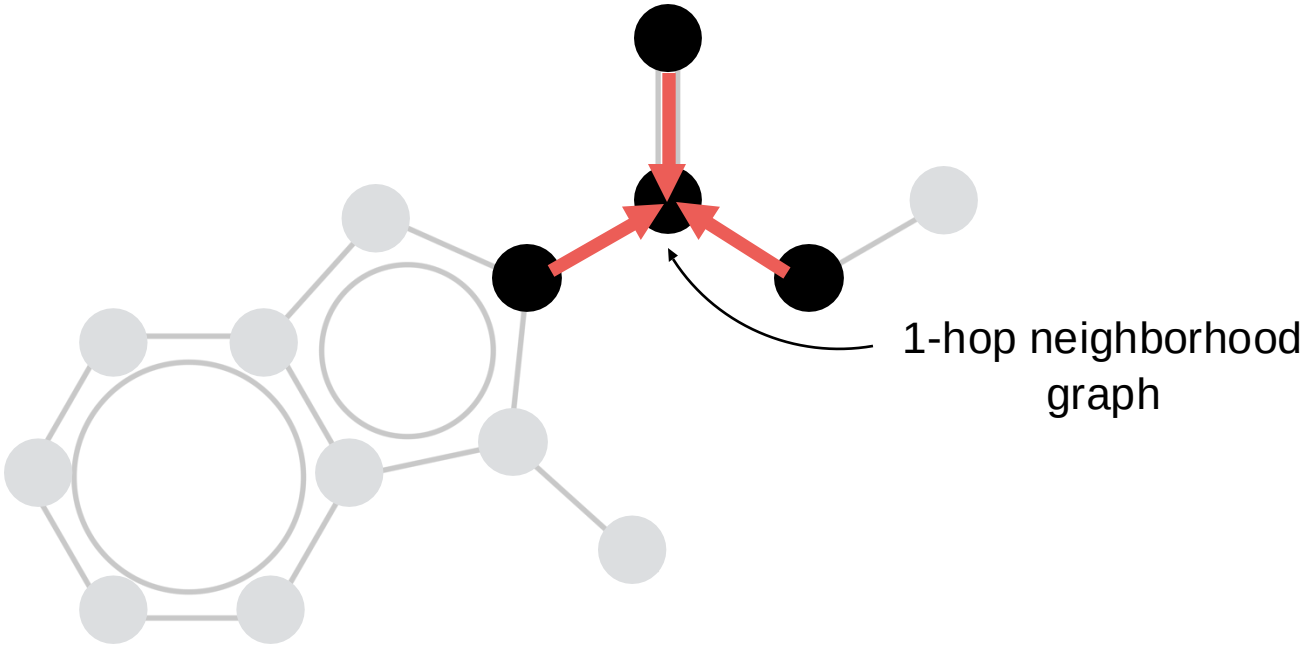
Jin et al., ICML 2018

Graph encoding



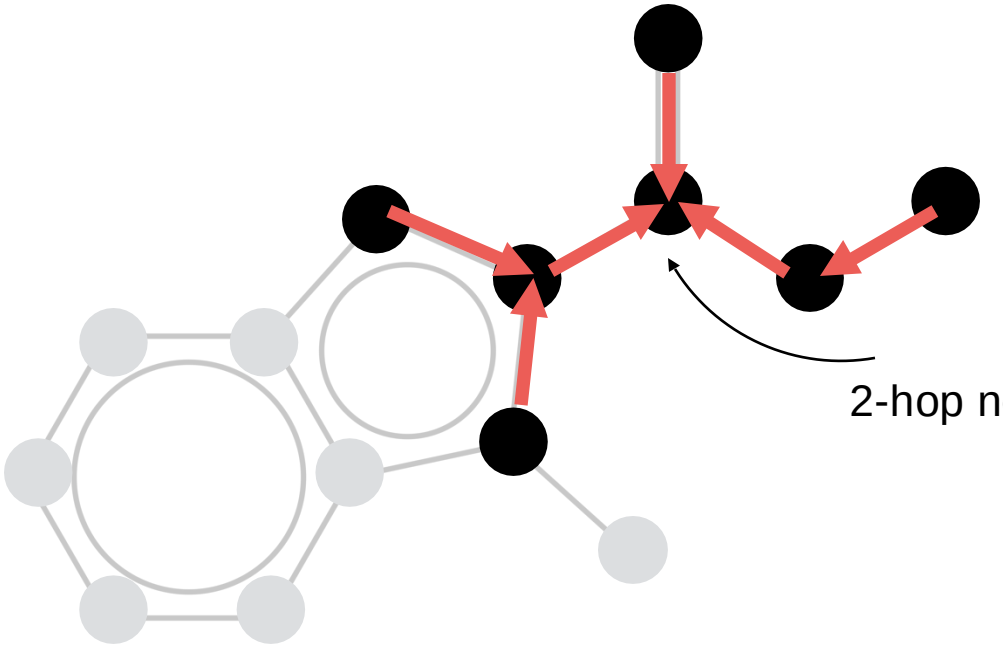
[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

Graph encoding



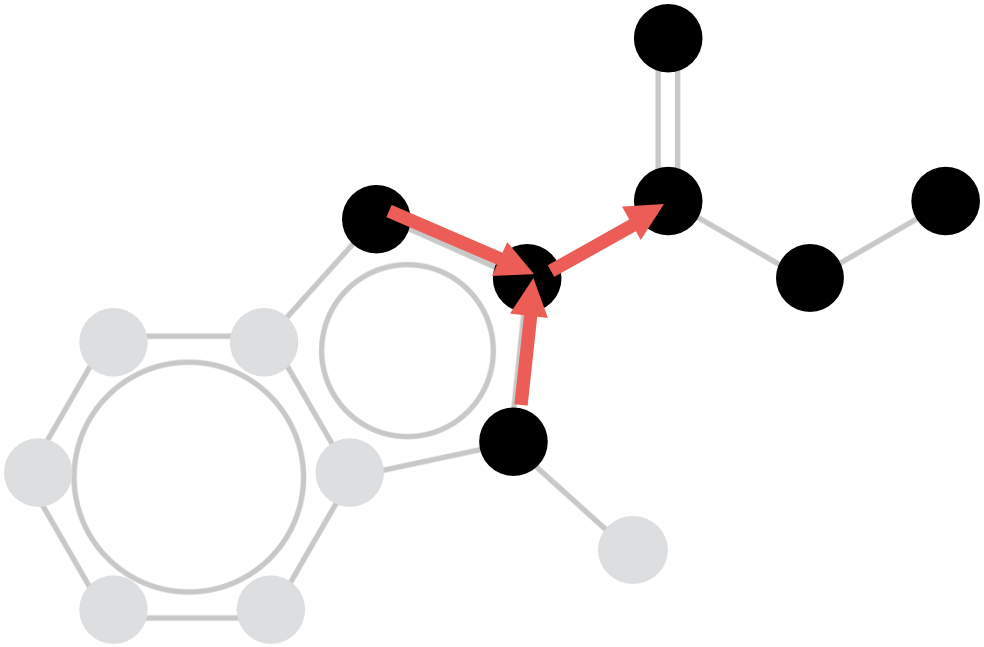
[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

Graph encoding



[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

Graph encoding

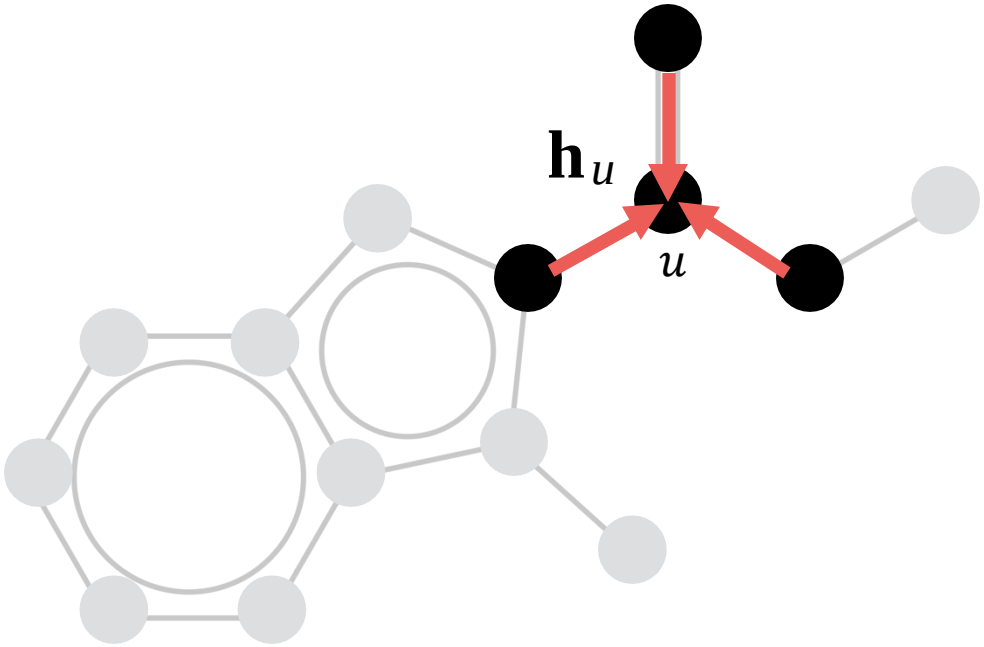


$$\boldsymbol{\nu}_{uv}^{(t)} = \tau(\mathbf{W}_1^g \mathbf{x}_u + \mathbf{W}_2^g \mathbf{x}_{uv} + \mathbf{W}_3^g \sum_{w \in N(u) \setminus v} \boldsymbol{\nu}_{wu}^{(t-1)})$$

Messages Node feature Edge feature

[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

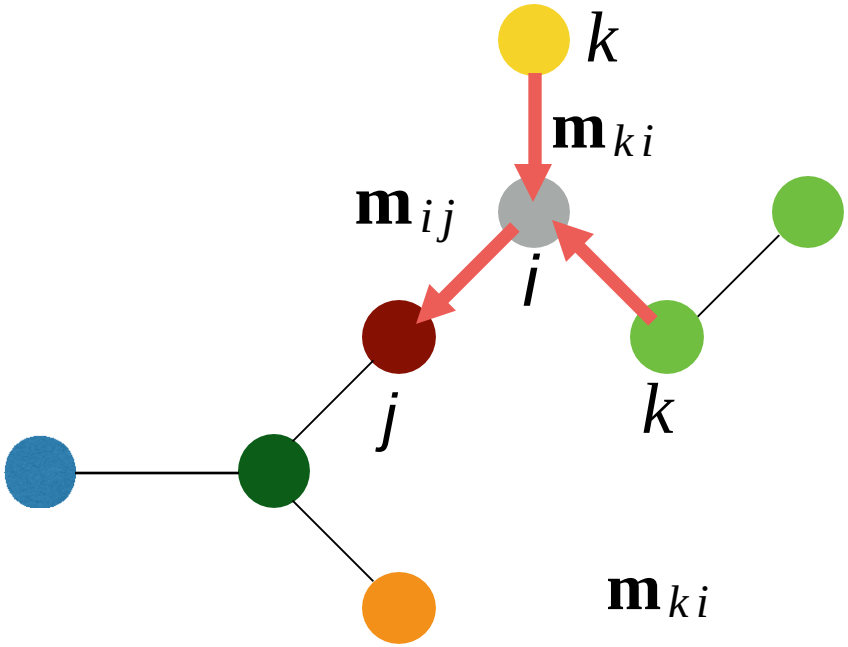
Graph encoding



$$\mathbf{h}_u = \tau(\mathbf{U}_1^g \mathbf{x}_u + \sum\nolimits_{v \in N(u)} \mathbf{U}_2^g \boldsymbol{\nu}_{vu}^{(T)})$$

[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

Tree encoding

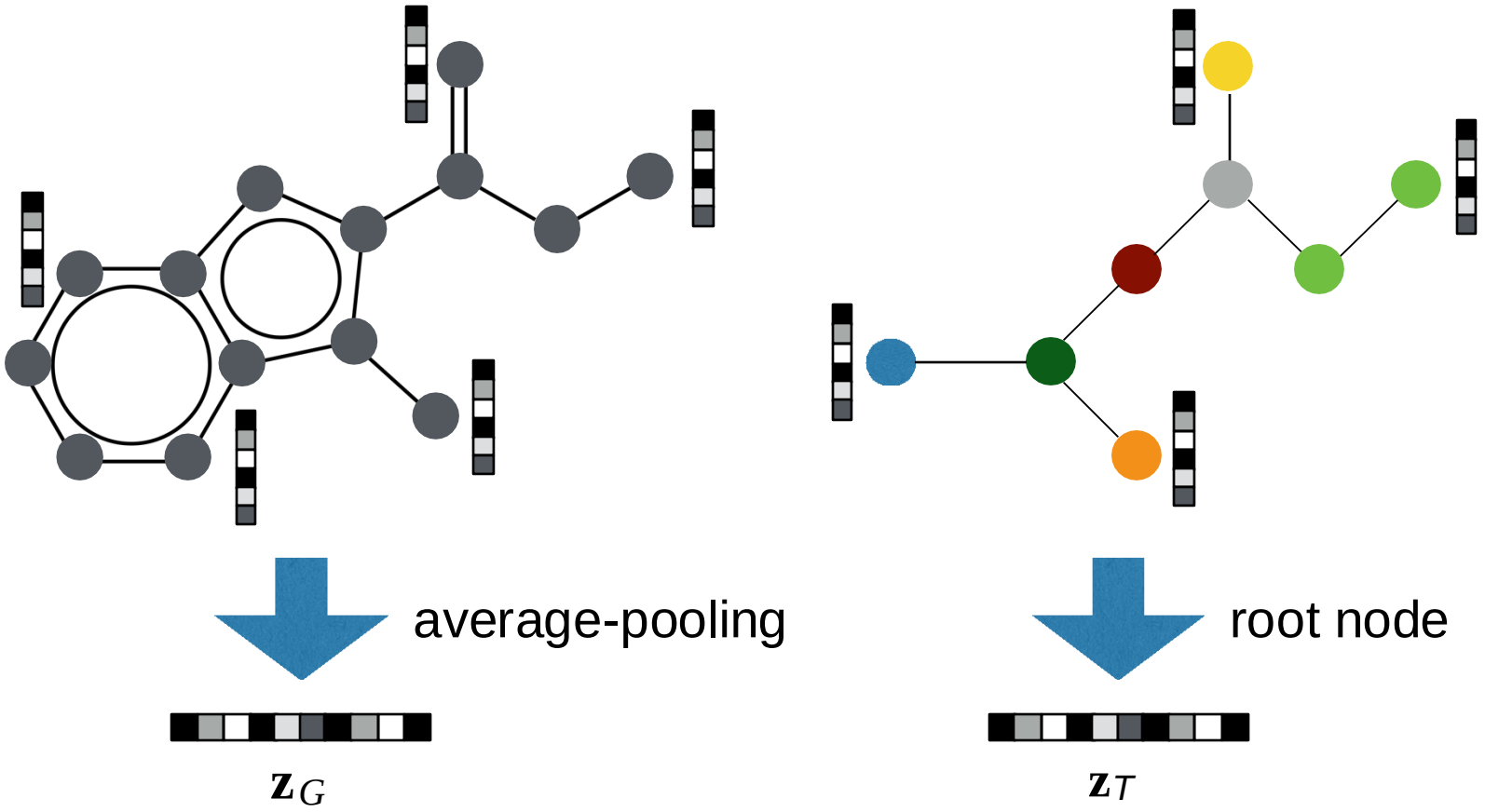


$$\mathbf{m}_{ij} = \text{GRU}(\mathbf{x}_i, \{\mathbf{m}_{ki}\}_{k \in N(i) \setminus j})$$

To capture long range interactions

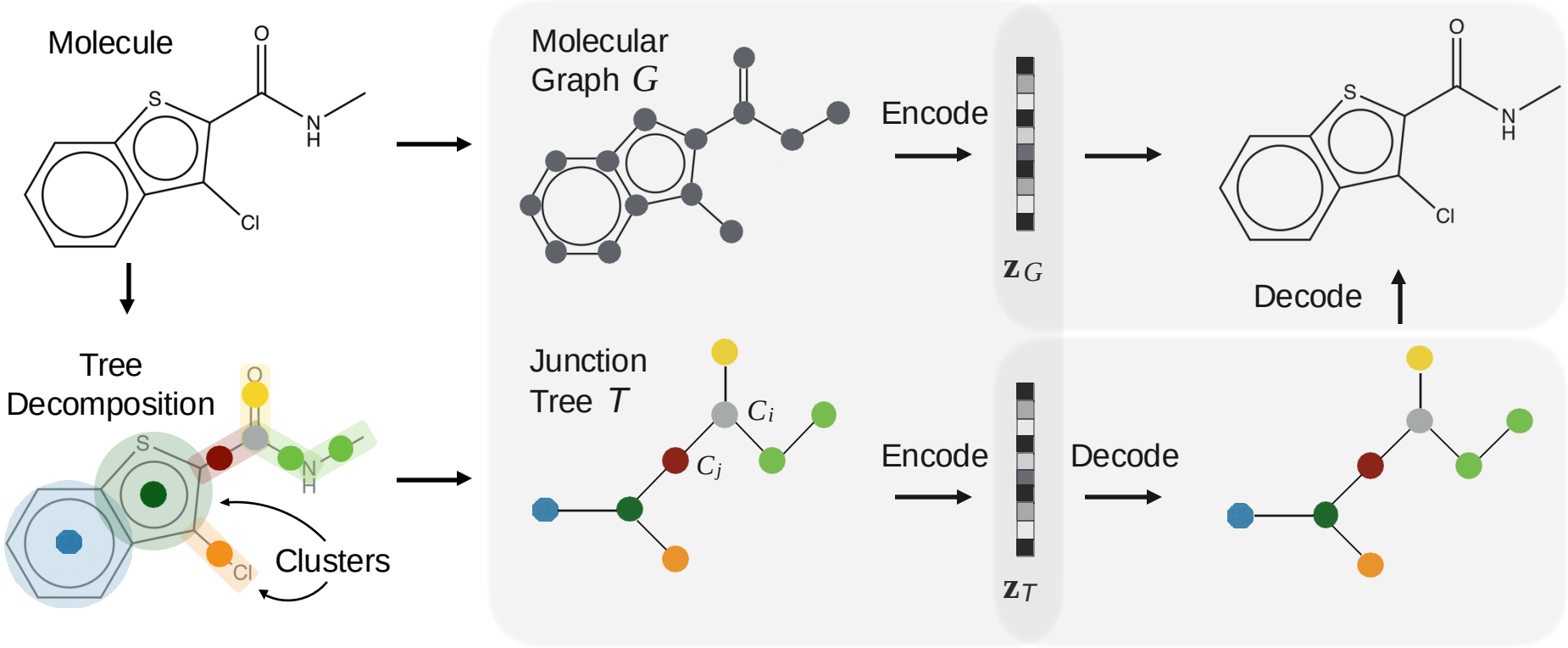
Jin et al., ICML 2018

Graph and tree encoders



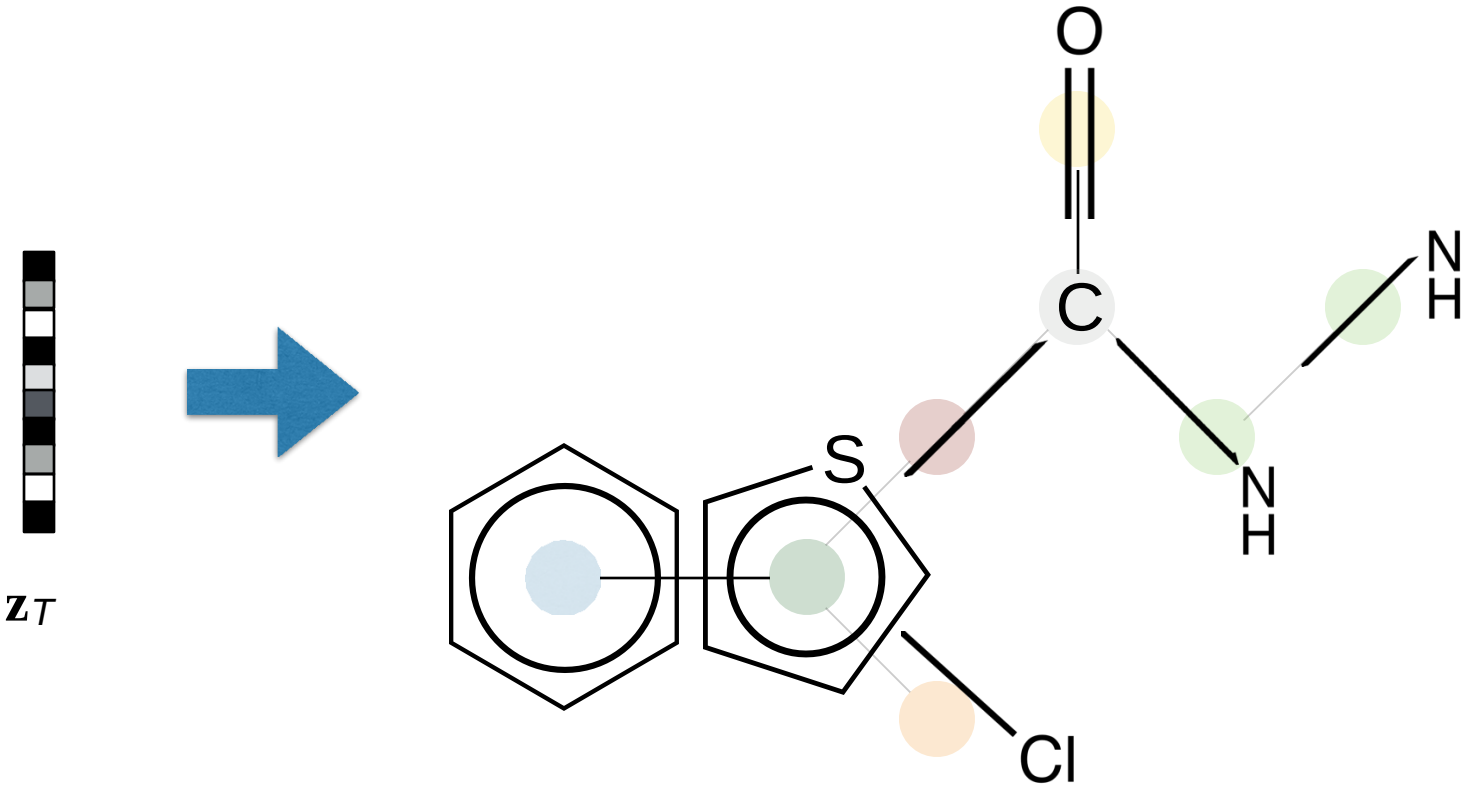
Jin et al., ICML 2018

Approach: Junction-tree variational autoencoder



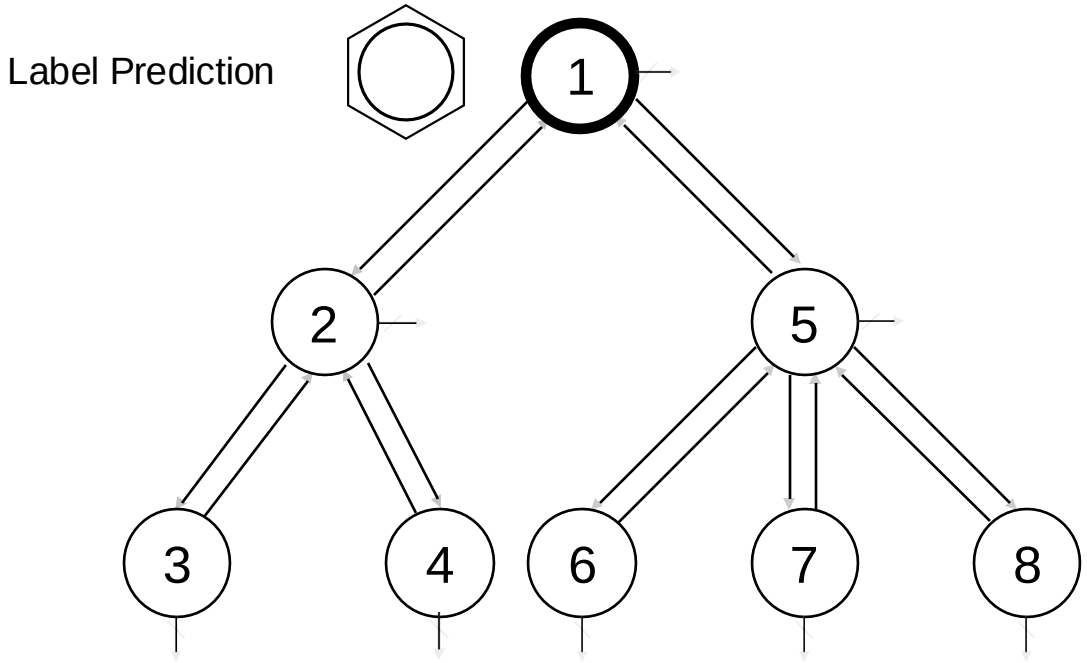
Jin et al., ICML 2018

Tree decoder



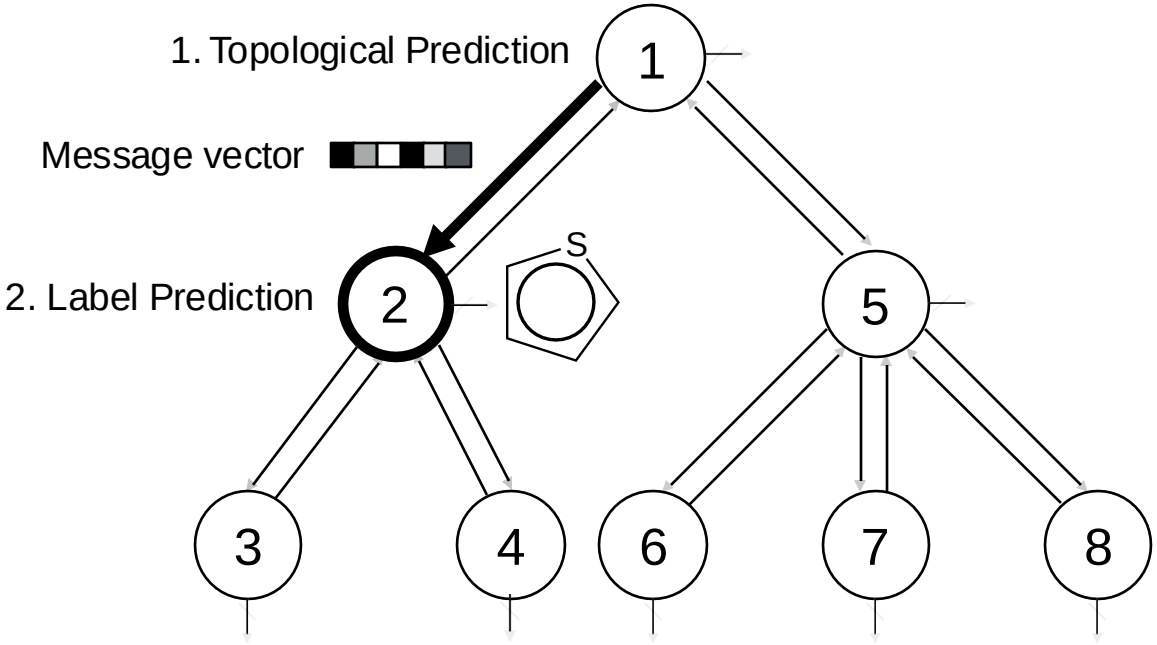
Jin et al., ICML 2018

Tree decoder



[4] Alvarez-Melis & Jaakkola, Tree-structured decoding with doubly-recurrent neural networks

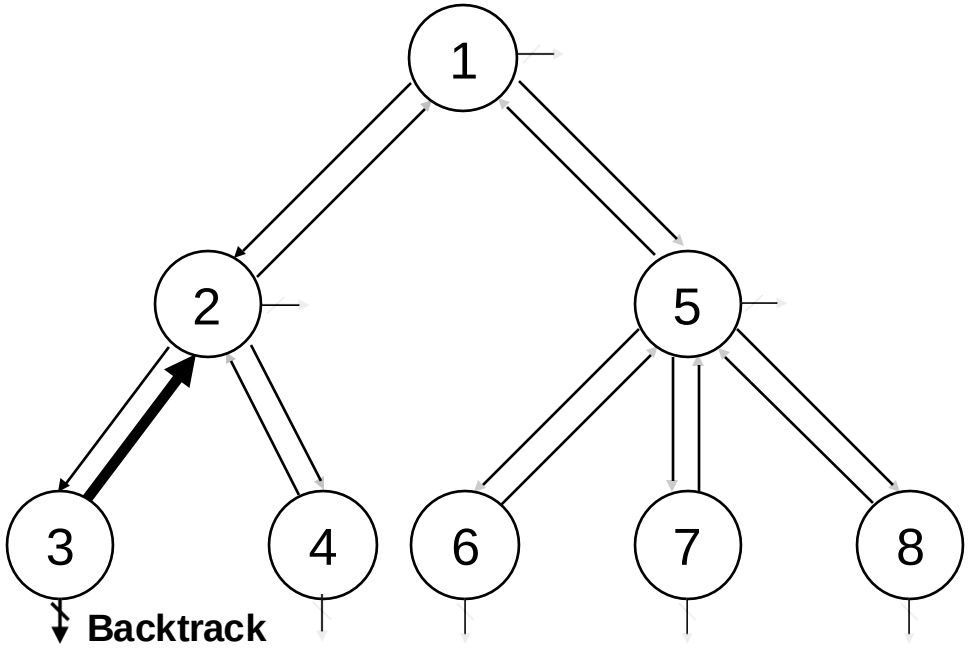
Tree decoder



Topological Prediction: Whether to expand a child or backtrack?

Label Prediction: What is the label of a node?

Tree decoder

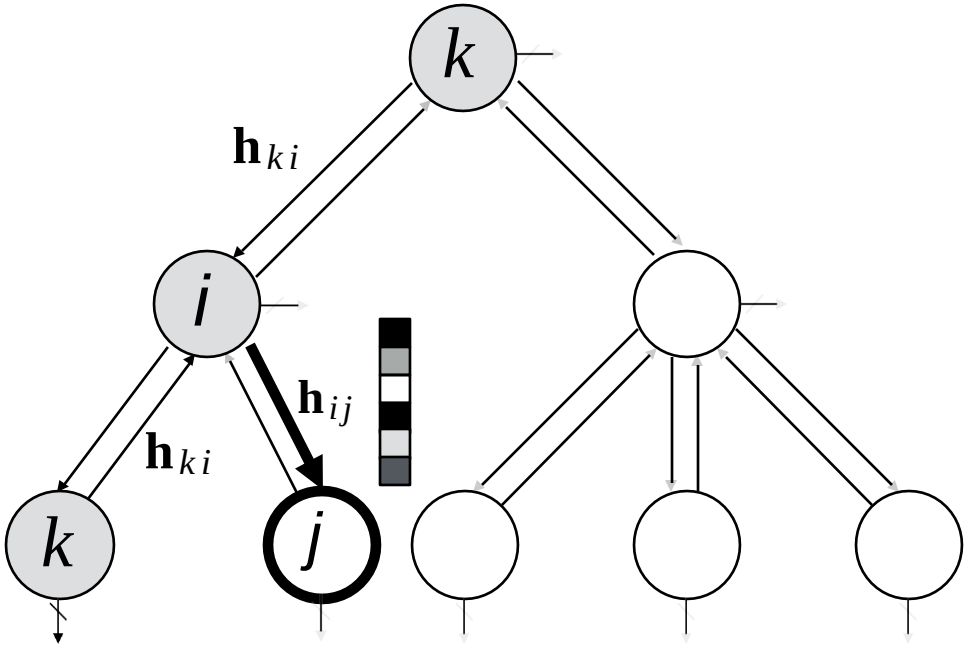


Topological Prediction

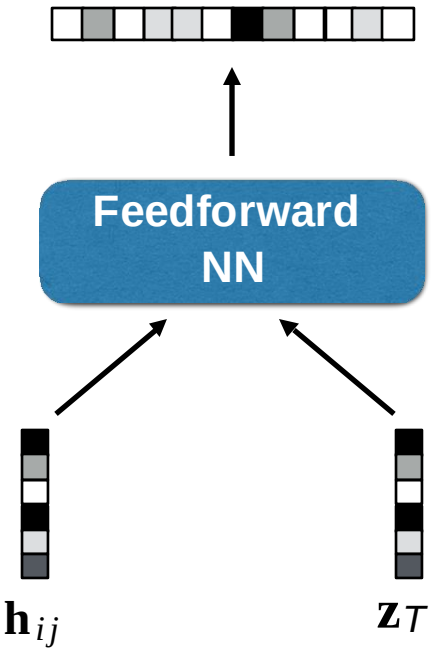
Topological Prediction: Whether to expand a node or backtrack?

Label Prediction: What is the label of a node?

Tree decoder



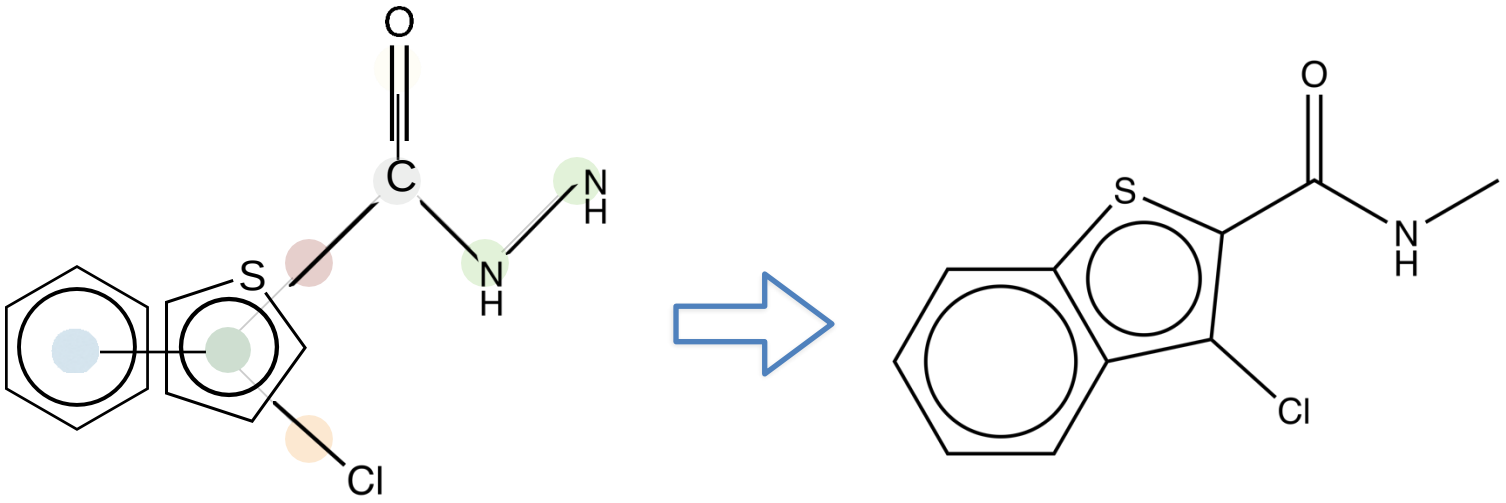
Label Prediction



$$\mathbf{h}_{ij} = \mathrm{GRU}(\mathbf{x}_i, \{\mathbf{h}_{ki}\}_{k2N_t(i)\setminus j})$$

Encodes the entire subtree of current state

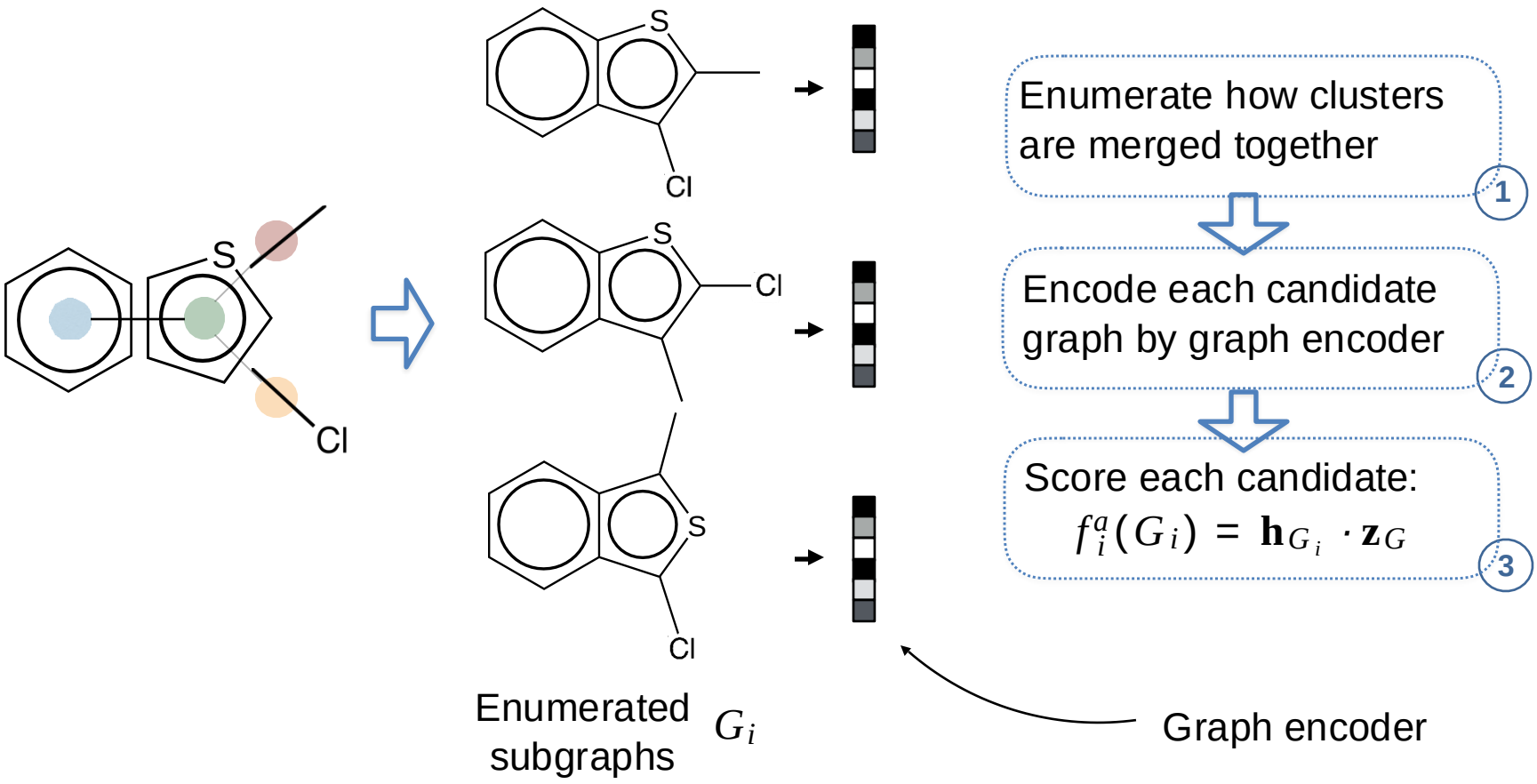
Graph decoder



Predicted Junction Tree

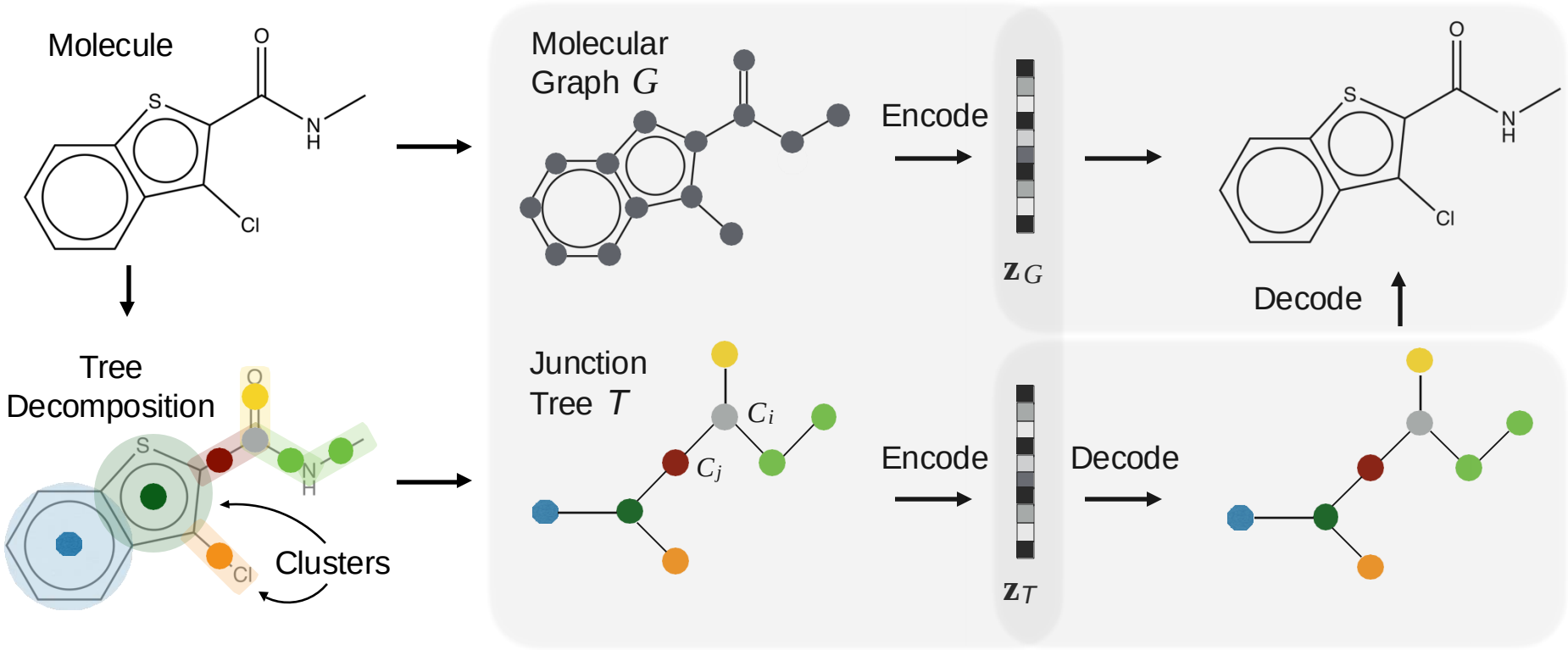
Molecular Graph

Graph decoder



$$\mathcal{L}_g(G) = \sum_i \left[f^a(G_i) - \log \sum_{G_i' \in \mathcal{G}_i} \exp(f^a(G_i')) \right] \quad (16)$$

Recap: Junction-tree variational autoencoder



Jin et al., ICML 2018

Experiments

- **Data**: 250K compounds from ZINC dataset
- **Molecule Generation**: How many molecules are valid when sampled from Gaussian prior?
- **Molecule Optimization**
  - **Global**: Find the best molecule in the entire latent space.
  - **Local**: Modify a molecule to increase its potency

Jin et al., ICML 2018

Baselines

SMILES string based:

1. Grammar VAE (GVAE) (Kusner et al., 2017);

2. Syntax-directed VAE (SD-VAE) (Dai et al., 2018)

Graph based:

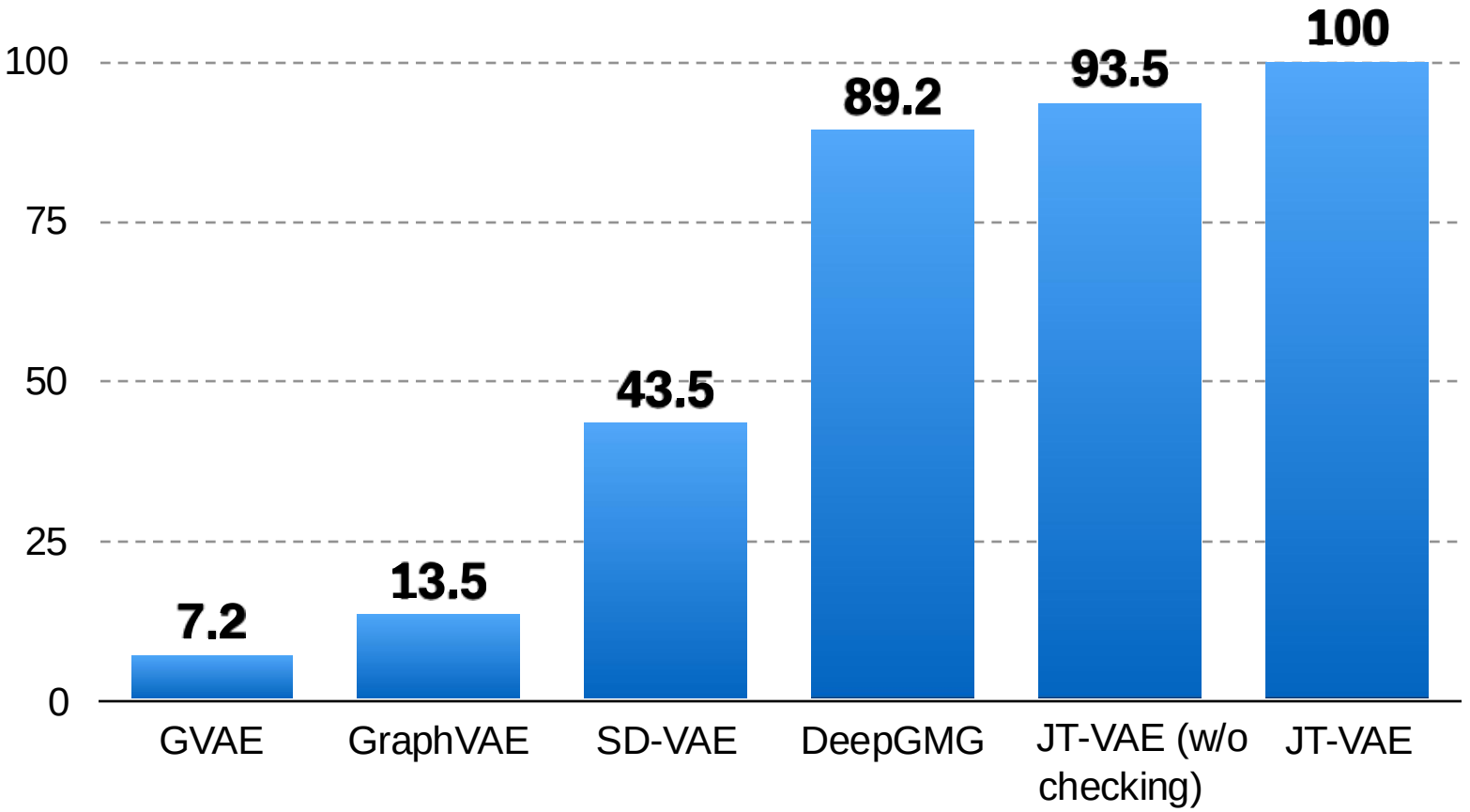
1. Graph VAE (Simonovsky & Komodakis, 2018)

2. DeepGMG (Li et al., 2018)

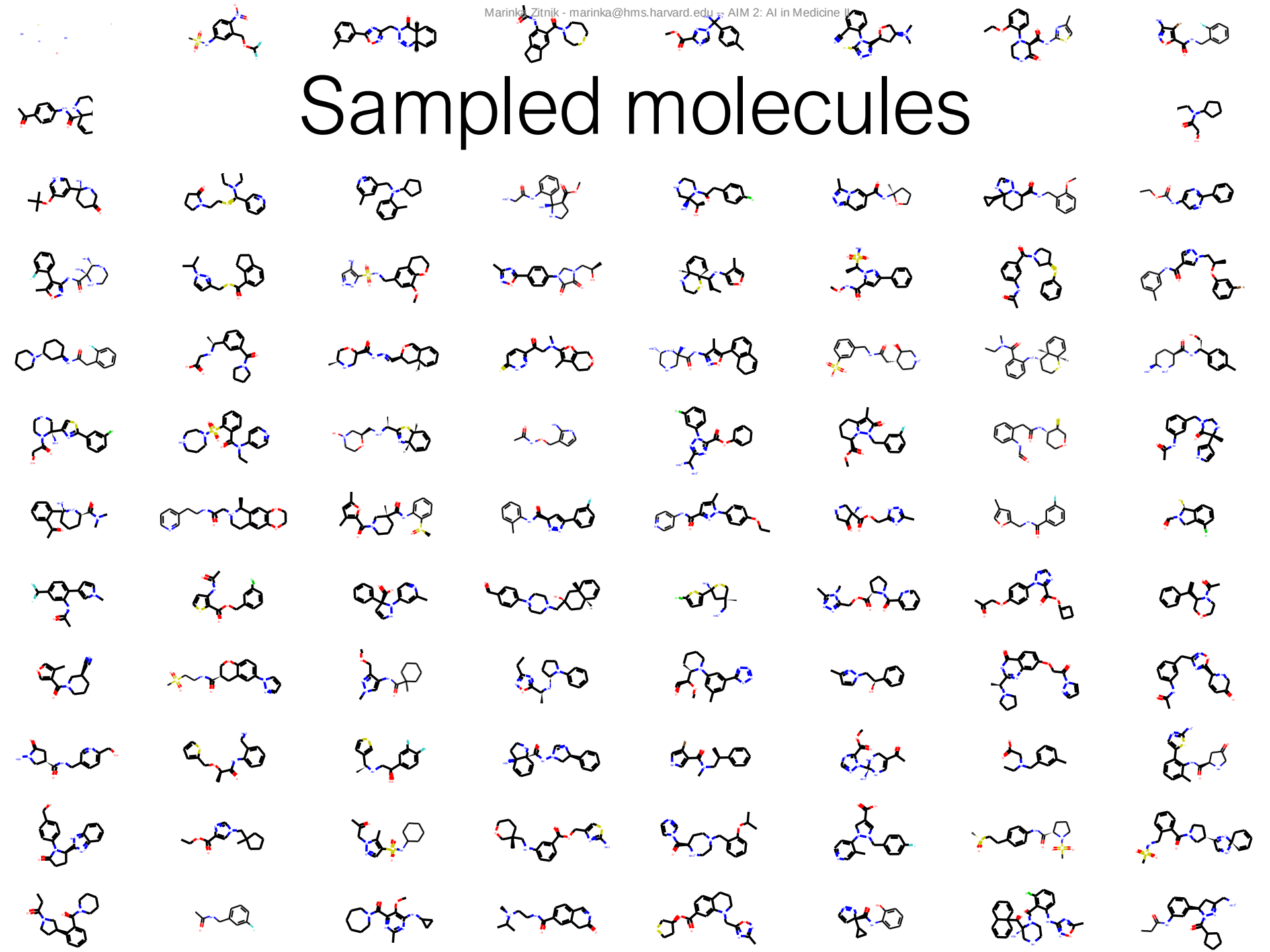
---

[2] Li et al., Learning Deep Generative Models of Graphs, 2018
5 Kusner et al., Grammar Variational Autoencoder, 2017
6 Dai et al., Syntax-directed Variational Autoencoder for structured data, 2018
7 Simonovsky & Komodakis, GraphVAE: Towards generation of small graphs using variational autoencoders

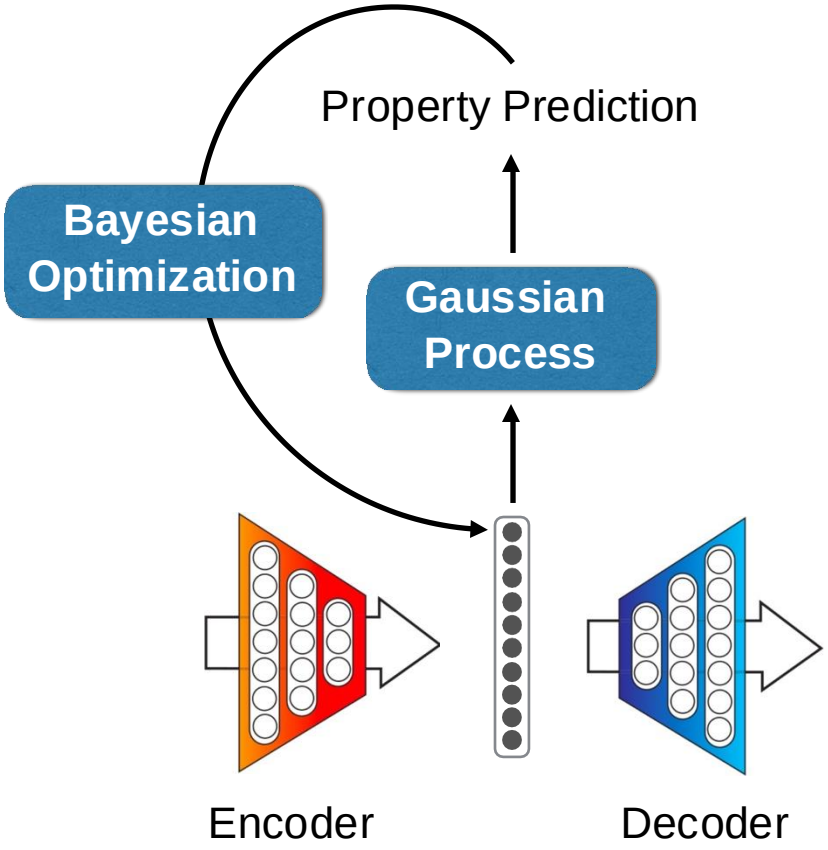
Molecule generation (Validity)

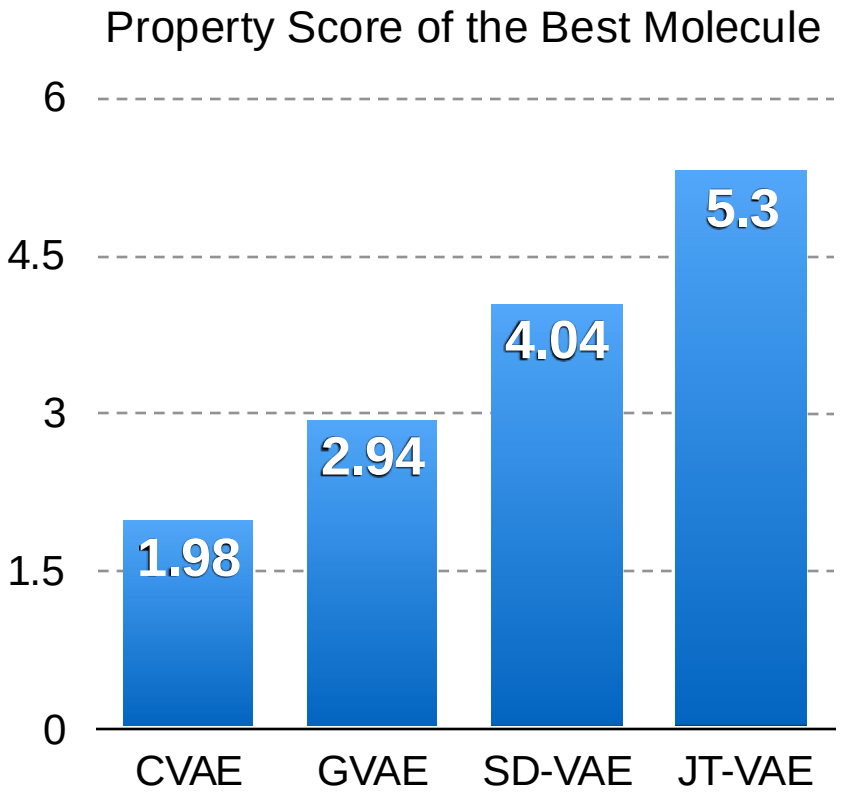


Sampled molecules



Molecule optimization (Global)

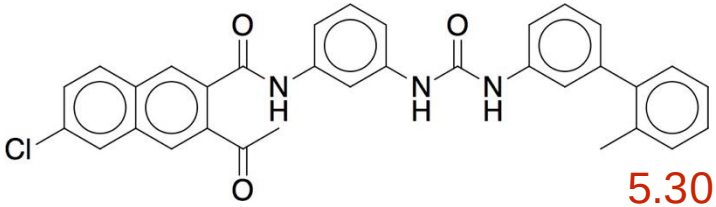




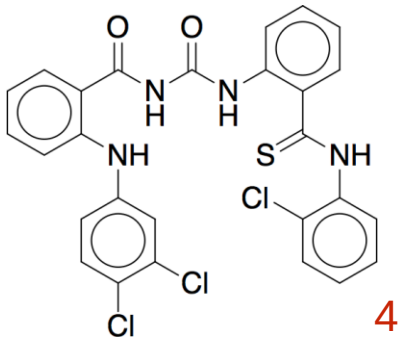
Property: Solubility + Ease of Synthesis

Jin et al., ICML 2018

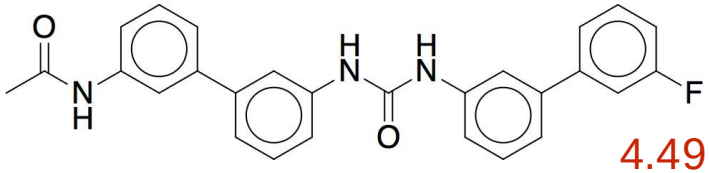
Molecule optimization (Global)



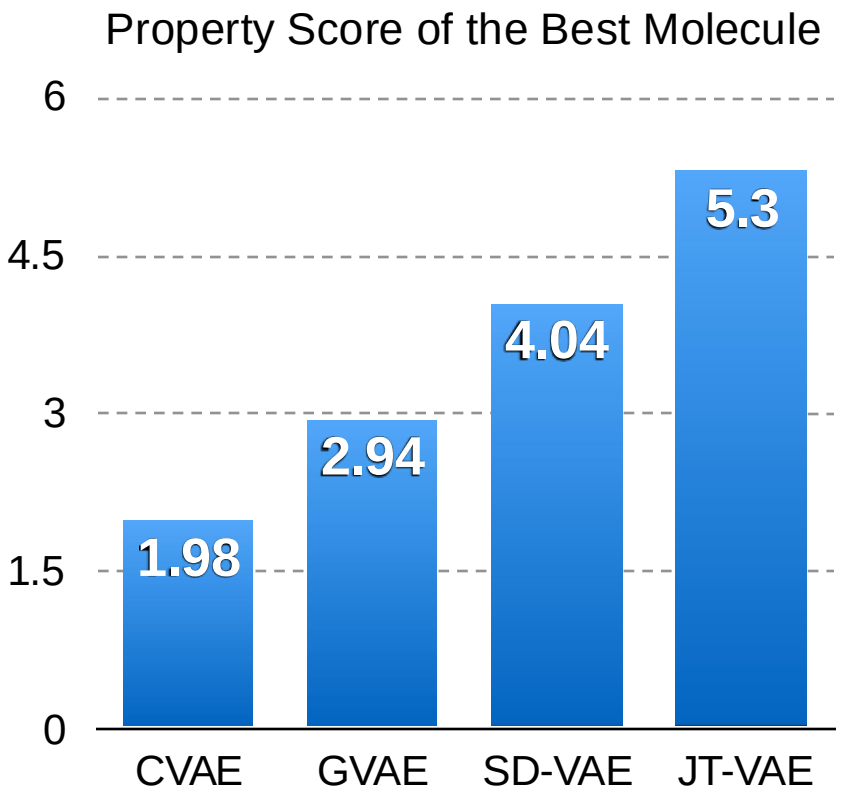
5.30



4.93



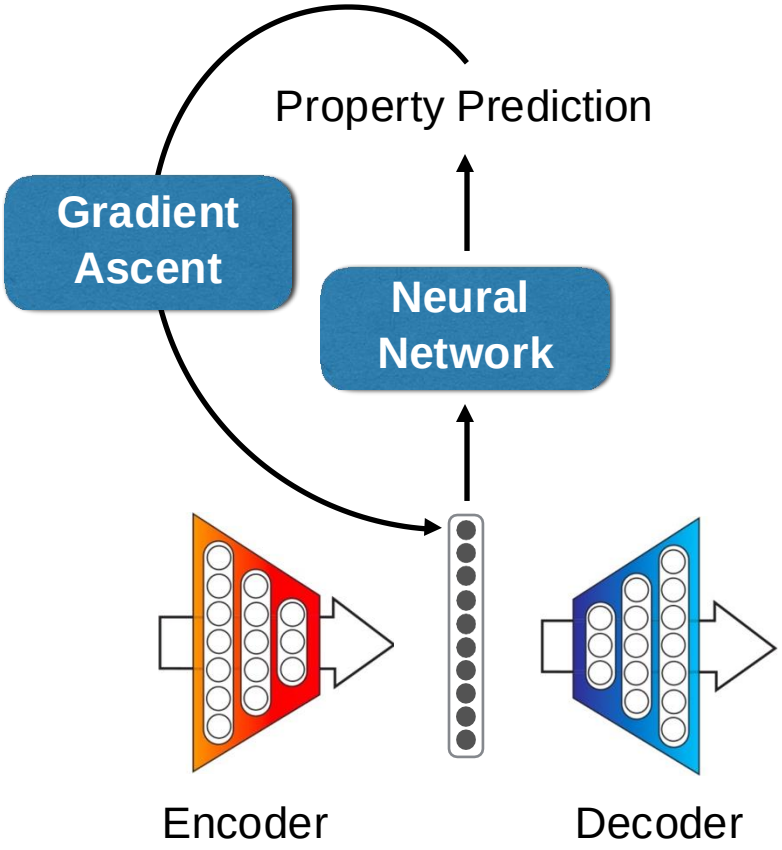
4.49

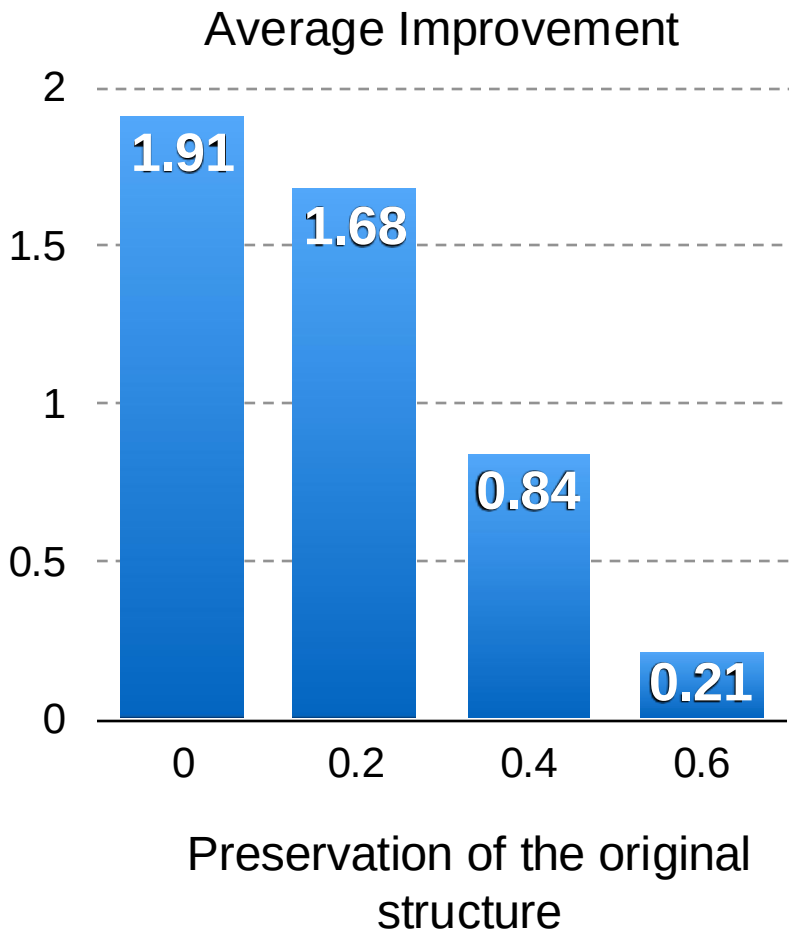


Property: Solubility + Ease of Synthesis

Jin et al., ICML 2018

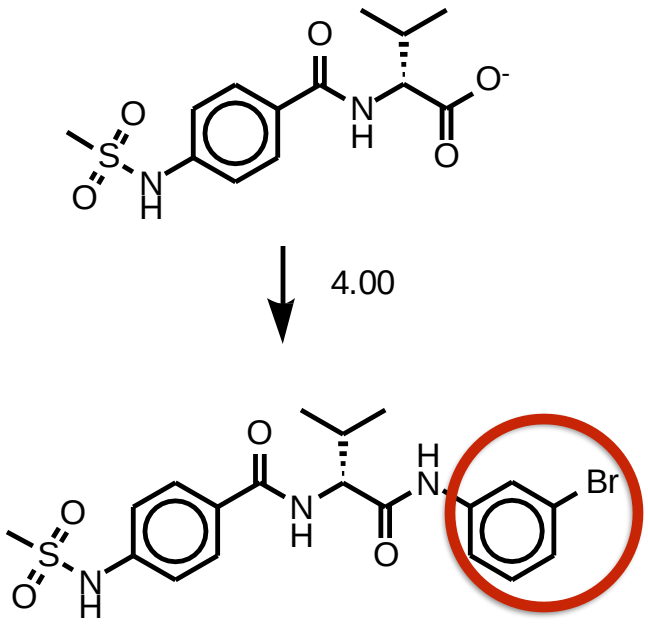
Molecule optimization (Local)



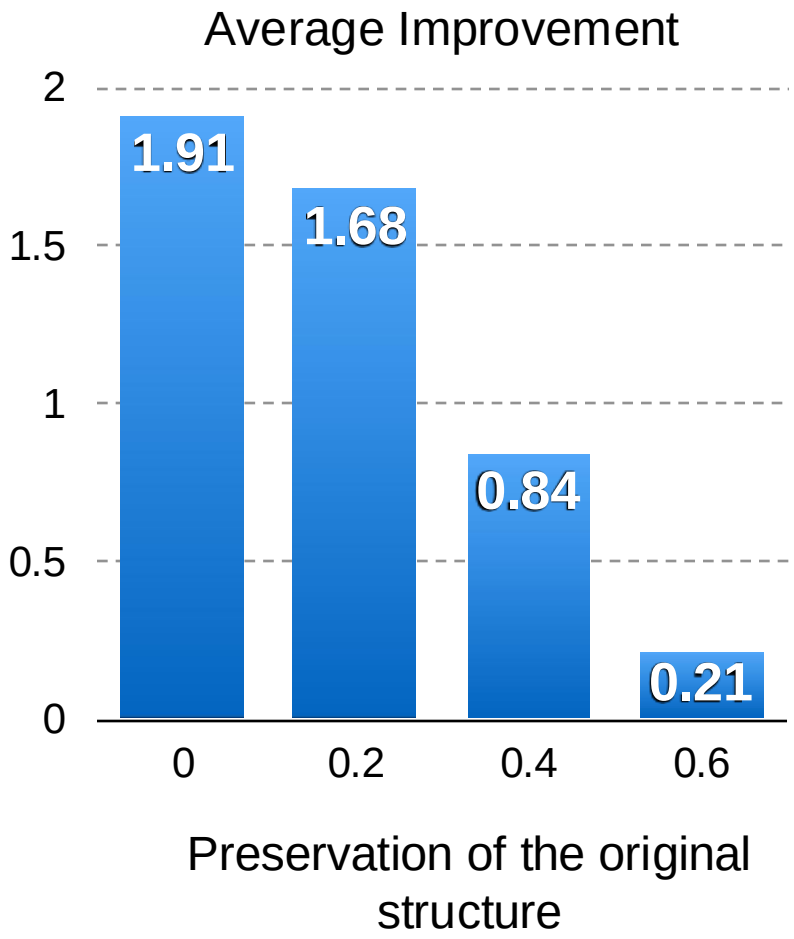


Jin et al., ICML 2018

Molecule optimization (Local)

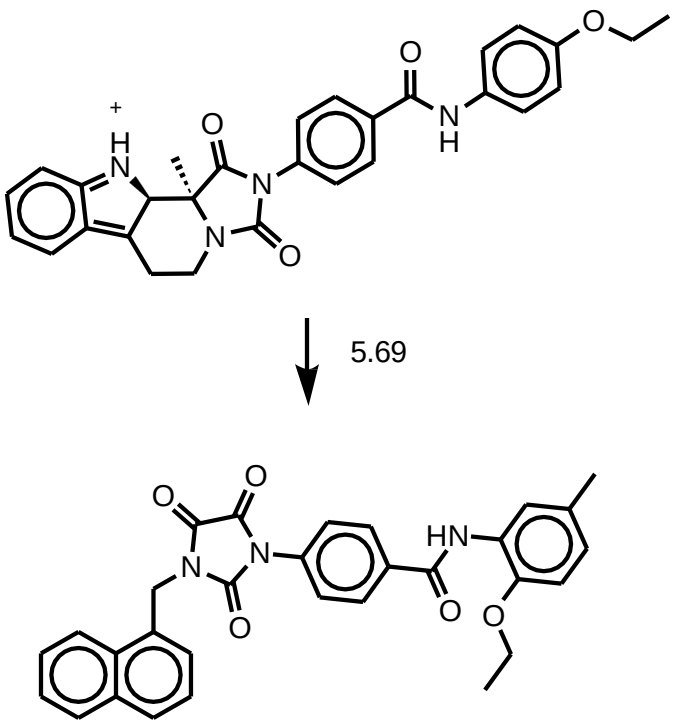


Preservation ≈ 0.6

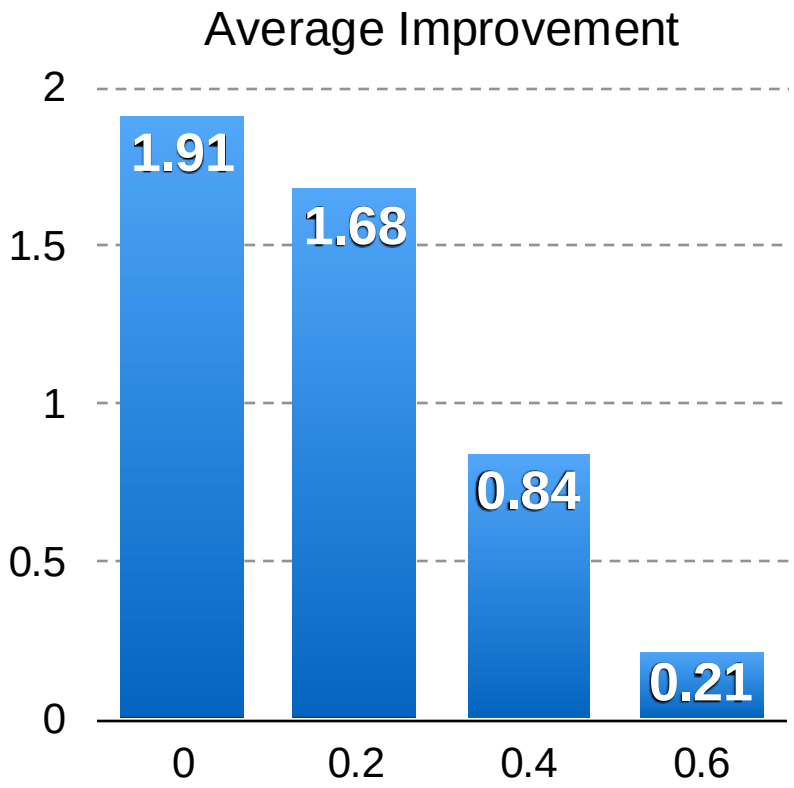


Jin et al., ICML 2018

Molecule optimization (Local)



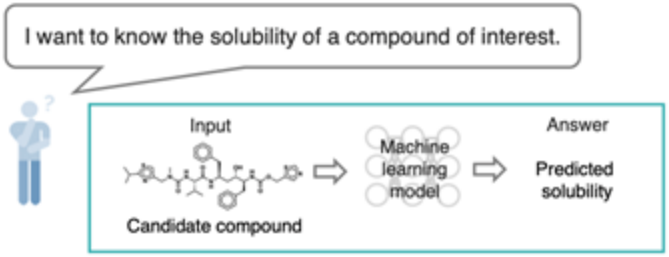
Preservation ≈ 0.4



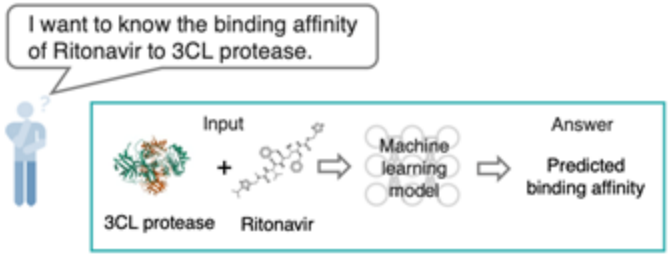
Jin et al., ICML 2018

Outline for today's class

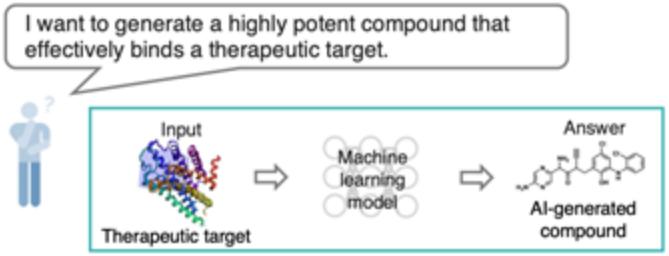
- Optimization & generation of small molecules



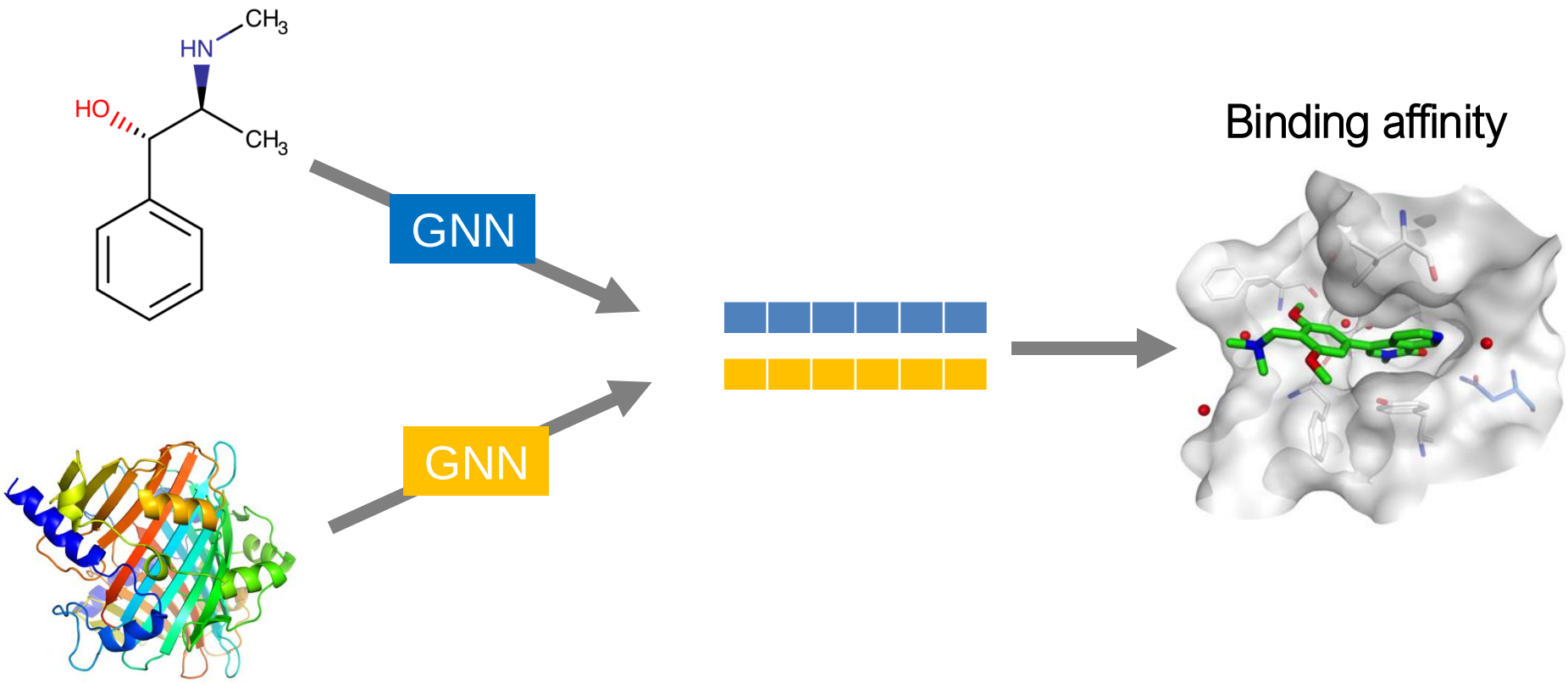
- Binding of drugs to therapeutic targets



- High-throughput genetic & chemical perturbations

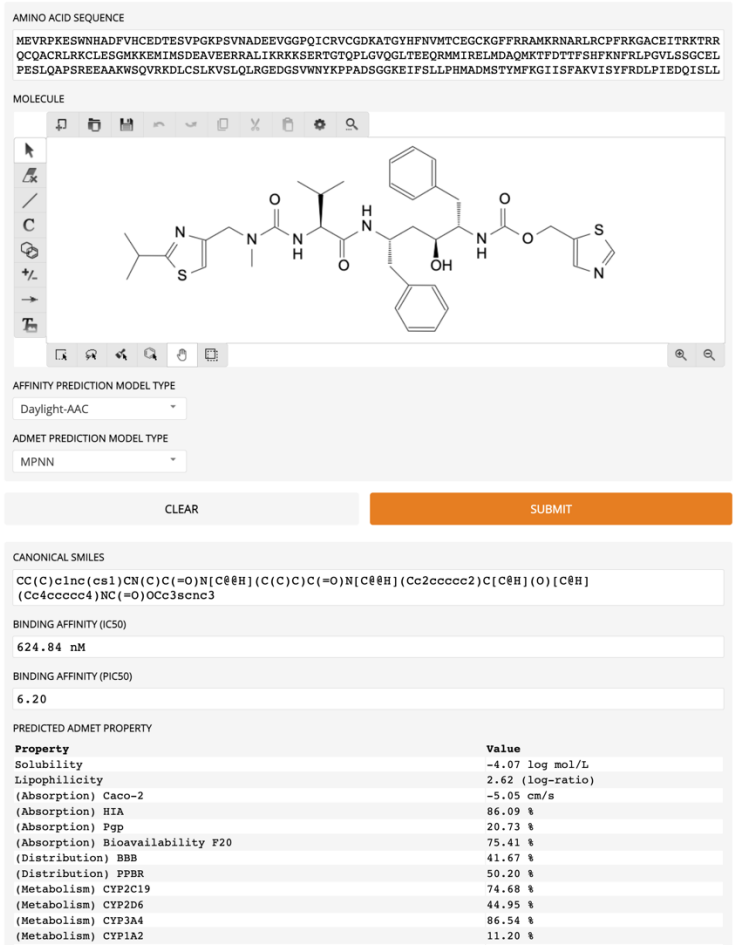


Geometric modeling of binding



Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology,* 2022

Results: Binding affinity prediction



- **ERM** is a standard strategy to minimize errors across all domains
- **MMD** minimizes maximum mean discrepancy across domains
- **CORAL** matches mean and covariance of features across domains
- **IRM** optimizes features using a cross-domain optimized linear classifier
- **GroupDRO** optimizes ERM and adjusts weights of domains with larger errors
- **Marginal transfer learning** augments features with marginal distributions
- **ANDMask** masks gradients that have inconsistent signs in the corresponding weights across domains

Modern data management
Human-AI collaboration

Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology,* 2022

Quick Check

https://forms.gle/72pBeaXADuhLnSmj9

AIM 2: Artificial Intelligence in Medicine II

Artificial Intelligence in Medicine II, Spring 2025

Lecture 10: AI for protein structure prediction, Drug discovery and therapeutic science, Structure- and sequence-based co-design, Biological foundation models

Course website and slides: https://zitniklab.hms.harvard.edu/AIM2

* Indicates required question

First and last name *

Your answer

Harvard email address *

Your answer

Describe two challenges that need to be addressed by generative models for molecular design. *

Your answer

What is the difference between traditional vs. neural fingerprint representations? *

Your answer

(1) What is the difference between local and global optimization in molecular design? (2) Give one use case for global optimization and one use case for local optimization. *

Your answer

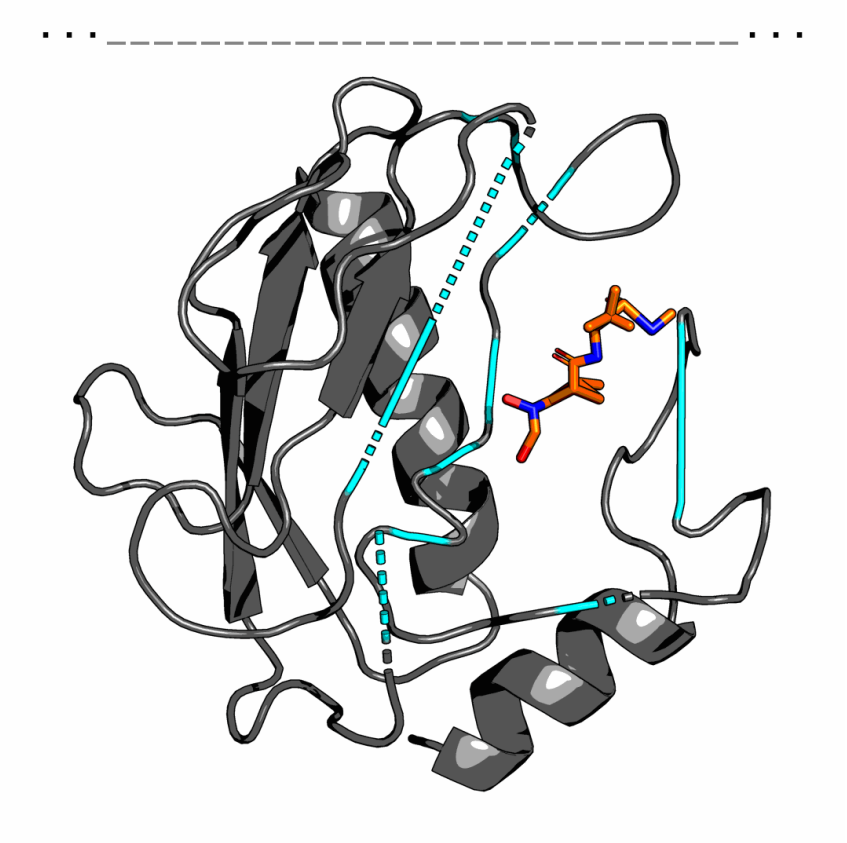
Submit

Clear form

Protein universe

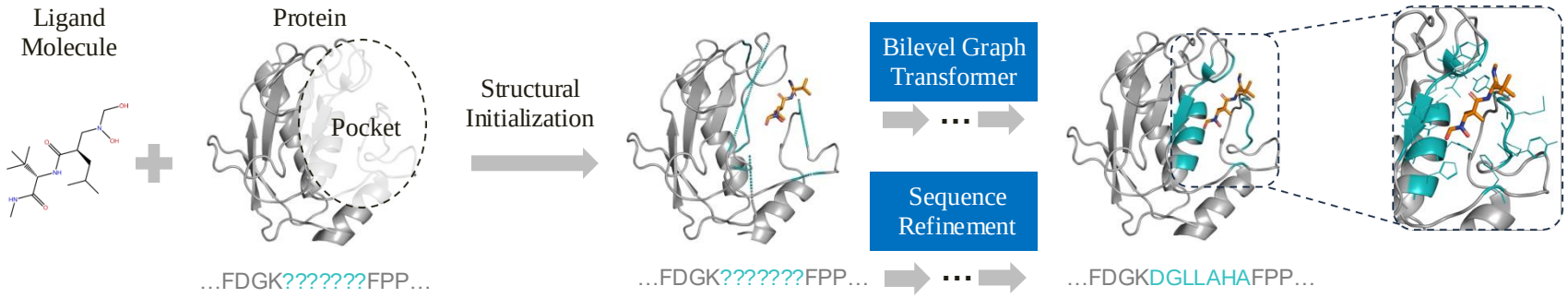
$\sim 20^{300}$

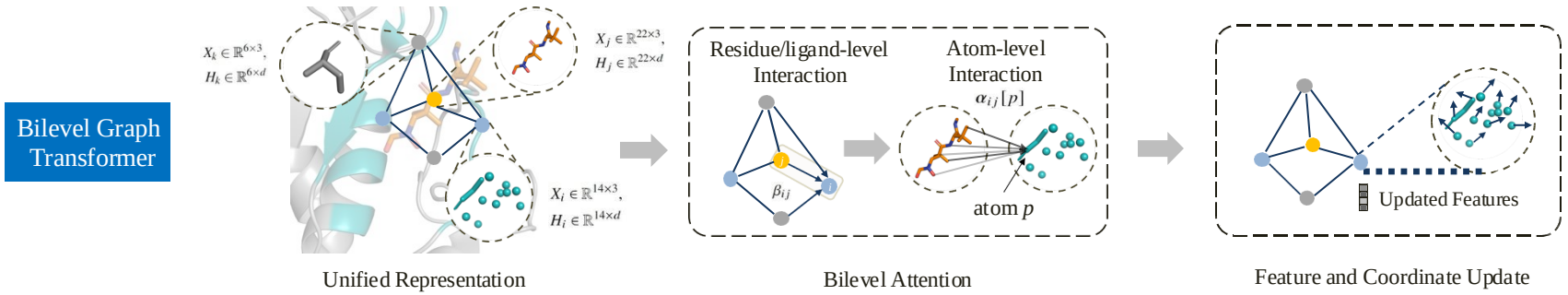
Sequence-structure molecular generation



Nature Methods'24; *Nature Chemical Biology*'22; *Nature Machine Intelligence*'24a, *Nature Machine Intelligence*'24b; *NeurIPS*'23; *ICML*'24; NeurIPS'23, NeurIPS'24

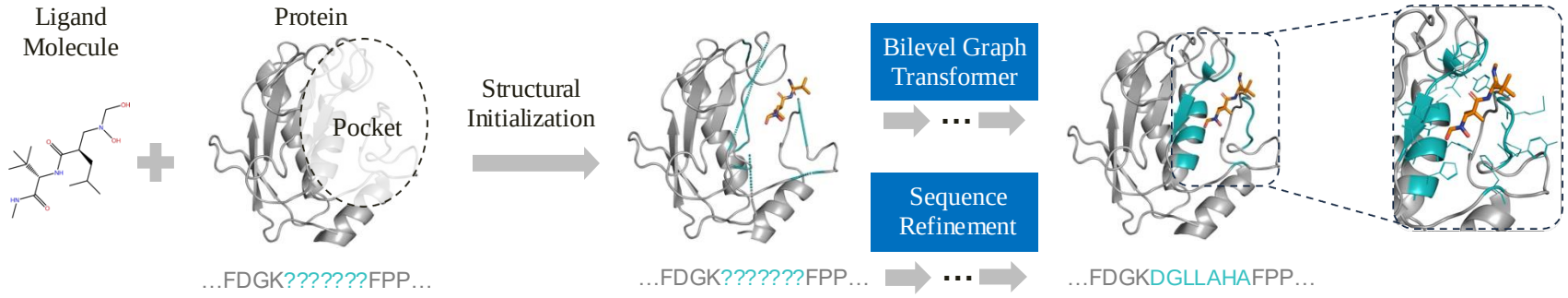
Iterative refinement of sequence and structure in the protein pocket to maximize binding affinity with small molecule ligands

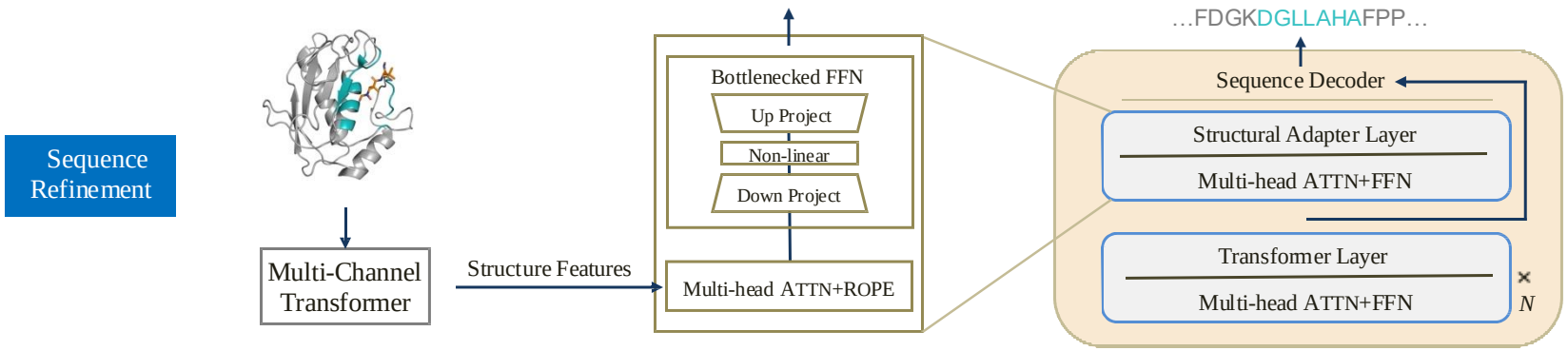




Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS'23*; Efficient Generation of Protein Pockets with PocketGen, *Nature Machine Intelligence'24*; Generalized Protein Pocket Generation with Prior-Informed Flow Matching, *NeurIPS'24*; Invariant Tokenization for Language Model Enabled Crystal Materials Generation, *NeurIPS'24;* Evaluating Generalizability of Artificial Intelligence Models for Molecular Datasets, *Nature Machine Intelligence'24*

Iterative refinement of sequence and structure in the protein pocket to maximize binding affinity with small molecule ligands

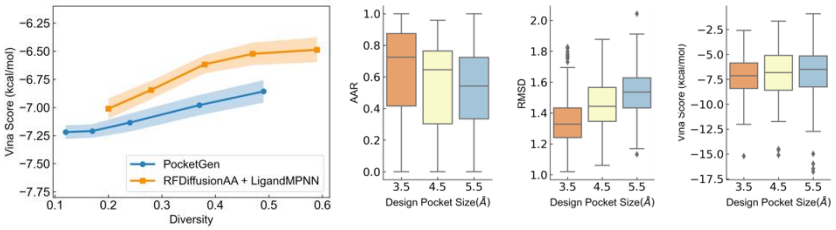




Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS'23*; Efficient Generation of Protein Pockets with PocketGen, *Nature Machine Intelligence'24*; Generalized Protein Pocket Generation with Prior-Informed Flow Matching, *NeurIPS'24*; Invariant Tokenization for Language Model Enabled Crystal Materials Generation, *NeurIPS'24;* Evaluating Generalizability of Artificial Intelligence Models for Molecular Datasets, *Nature Machine Intelligence'24*

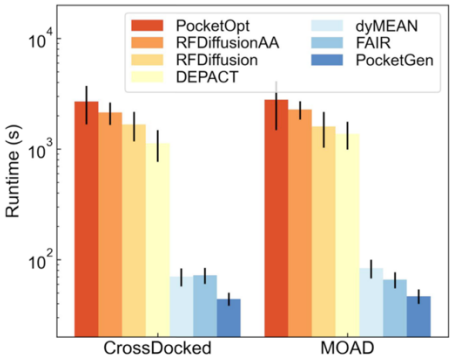
PocketGen generates protein pockets with high binding affinity and structural validity

	PocketOpt	DEPACT	dyMEAN	FAIR	RFDiffusion	RFDiffusionAA	PocketGen
Top-1 generated protein pocket							
Vina score (↓)	-9.216	-8.527	-8.540	-8.792	-9.037	-9.216	**-9.655**
Success Rate (↑)	0.92	0.75	0.76	0.80	0.89	0.93	**0.97**
pLDDT (↑)	-	82.1	83.3	83.2	84.5	86.3	**86.7**
scRMSD (↓)	-	0.705	0.703	0.680	0.676	0.654	**0.645**
scTM (↑)	-	0.901	0.906	0.899	0.924	0.931	**0.937**
Top-3 generated protein pockets							
Vina score (↓)	-8.878	-8.131	-8.196	-8.321	-8.876	-8.980	**-9.353**
pLDDT (↑)	-	81.9	82.8	83.1	84.6	**86.2**	**86.2**
scRMSD (↓)	-	0.706	0.724	0.685	0.679	**0.653**	0.657
scTM (↑)	-	0.896	0.892	0.897	0.929	0.930	**0.934**
Top-5 generated protein pockets							
Vina score (↓)	-8.702	-7.786	-7.974	-7.943	-8.510	-8.689	**-9.239**
pLDDT (↑)	-	82.2	82.9	83.3	84.3	85.7	**86.1**
scRMSD (↓)	-	0.717	0.725	0.690	0.680	0.656	**0.652**
scTM (↑)	-	0.892	0.903	0.886	0.926	0.929	**0.935**
Top-10 generated protein pockets							
Vina score (↓)	-8.556	-7.681	-7.690	-7.785	-8.352	-8.524	**-9.065**
pLDDT (↑)	-	81.5	82.7	83.0	84.2	85.3	**85.9**
scRMSD (↓)	-	0.710	0.734	0.705	0.684	**0.672**	0.678
scTM (↑)	-	0.895	0.896	0.884	0.924	0.929	**0.931**

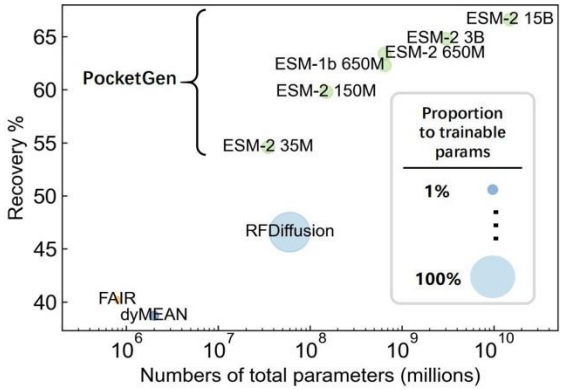


Improved structural validity, amino acid sequence recovery, and binding affinity with target ligands

Model	CrossDocked			Binding MOAD		
	AAR (↑)	Designability (↑)	Vina (↓)	AAR (↑)	Designability (↑)	Vina (↓)
Test set	-	0.77	-7.016	-	0.79	-8.076
DEPACT	31.52±3.26%	0.68±0.04	-6.632±0.18	35.30±2.19%	0.67±0.06	-7.571±0.15
dyMEAN	38.71±2.16%	0.71±0.03	-6.855±0.06	41.22±1.40%	0.70±0.03	-7.675±0.09
FAIR	40.16±1.17%	0.73±0.02	-7.015±0.12	43.68±0.92%	0.72±0.05	-7.930±0.15
RFDiffusion	46.57±2.07%	0.74±0.01	-6.936±0.07	45.31±2.73%	0.75±0.05	-7.942±0.14
RFDiffusionAA	50.85±1.85%	0.75±0.03	-7.012±0.09	49.09±2.49%	0.78±0.03	-8.020±0.11
PocketGen	**63.40±1.64%**	**0.77±0.02**	**-7.135±0.08**	**64.43±2.35%**	**0.80±0.04**	**-8.112±0.14**

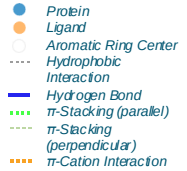


Better generation efficiency



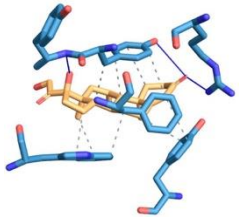
Performance wrt protein LM size

PocketGen can design protein pockets of antibodies, enzymes, and biosensors for target ligand molecules



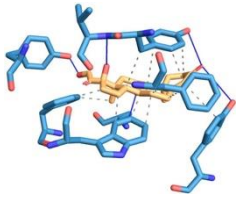
Cortisol (HCY)

…WFRYYDTMY…

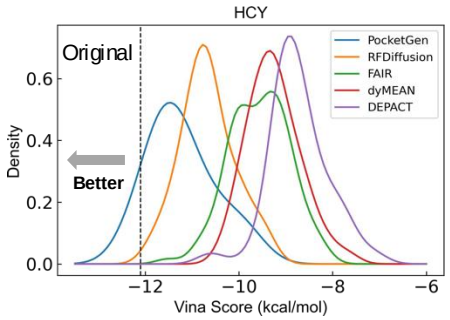


Original
(HP 12, HB 3)

…WFRYVYSFY…

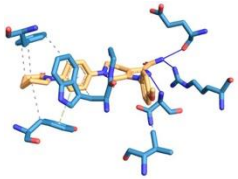


PocketGen
(HP 12, HB 5)



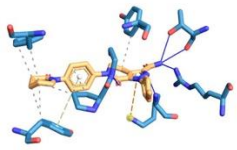
Apixaban (APX)

…YREFQVWGG…

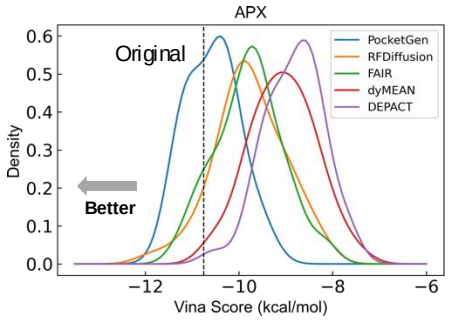


Original
(HP 7, HB 4, π 1)

…YRTFKVPGY…

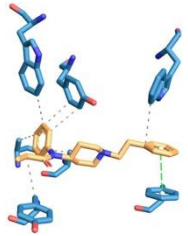


PocketGen
(HP 9, HB 5, π 2)



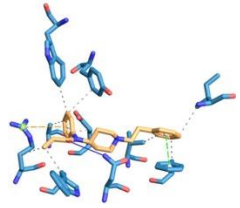
Fentanyl 7V7)

…FVWAWYWAY…

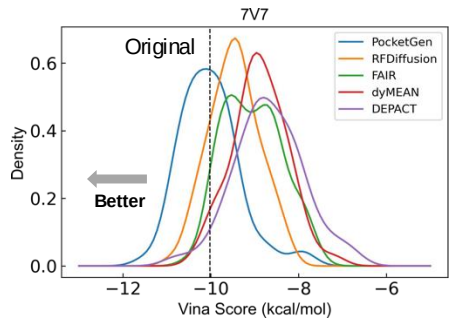


Original
(HP 7, HB 0, π 1)

…FVRAWYWEW…



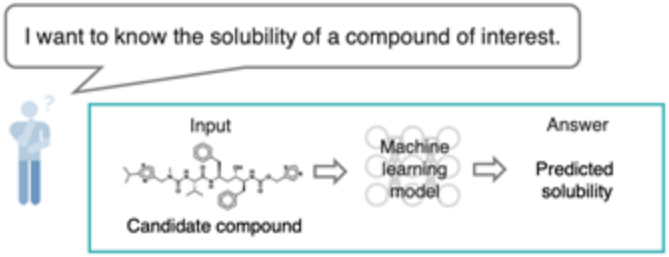
PocketGen
(HP 9, HB 1, π 2)



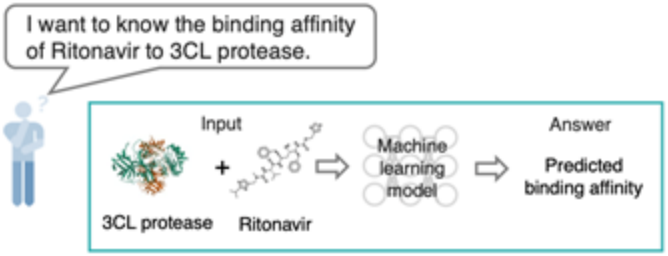
Full-Atom Protein Pocket Design via Iterative Refinement, *NeurIPS'23*; Efficient Generation of Protein Pockets with PocketGen, *Nature Machine Intelligence'24*;
Generalized Protein Pocket Generation with Prior-Informed Flow Matching, *NeurIPS'24;* AI Foundation for Therapeutic Science, *Nature Chemical Biology*

Outline for today's class

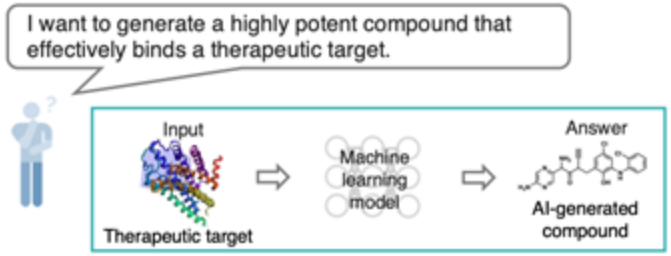
- Optimization & generation of small molecules



- Binding of drugs to therapeutic targets



- High-throughput genetic & chemical perturbations



Words and genes share a correspondence:
their **meanings** arise from their **context**.

Gene perturbation measurements across diverse cell contexts induce **semantics for genes**

(under the right approach)

“apple” is a **polysemic** word…

… whose **particular meaning** is resolved via **sentence context**.

Google

grow an apple

grow an apple **tree**

grow an apple **tree from seed**

grow an apple **tree in a pot**

grow an apple **tree indoors**

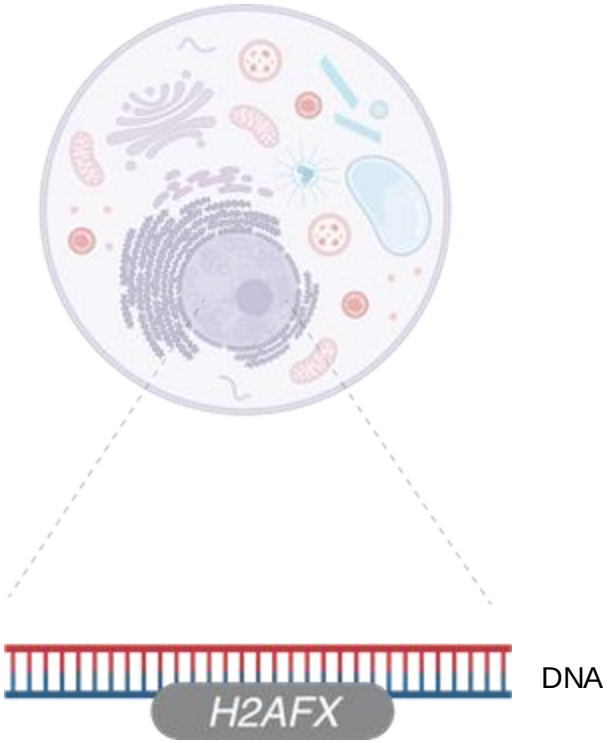
buy an apple

buy an apple **watch**

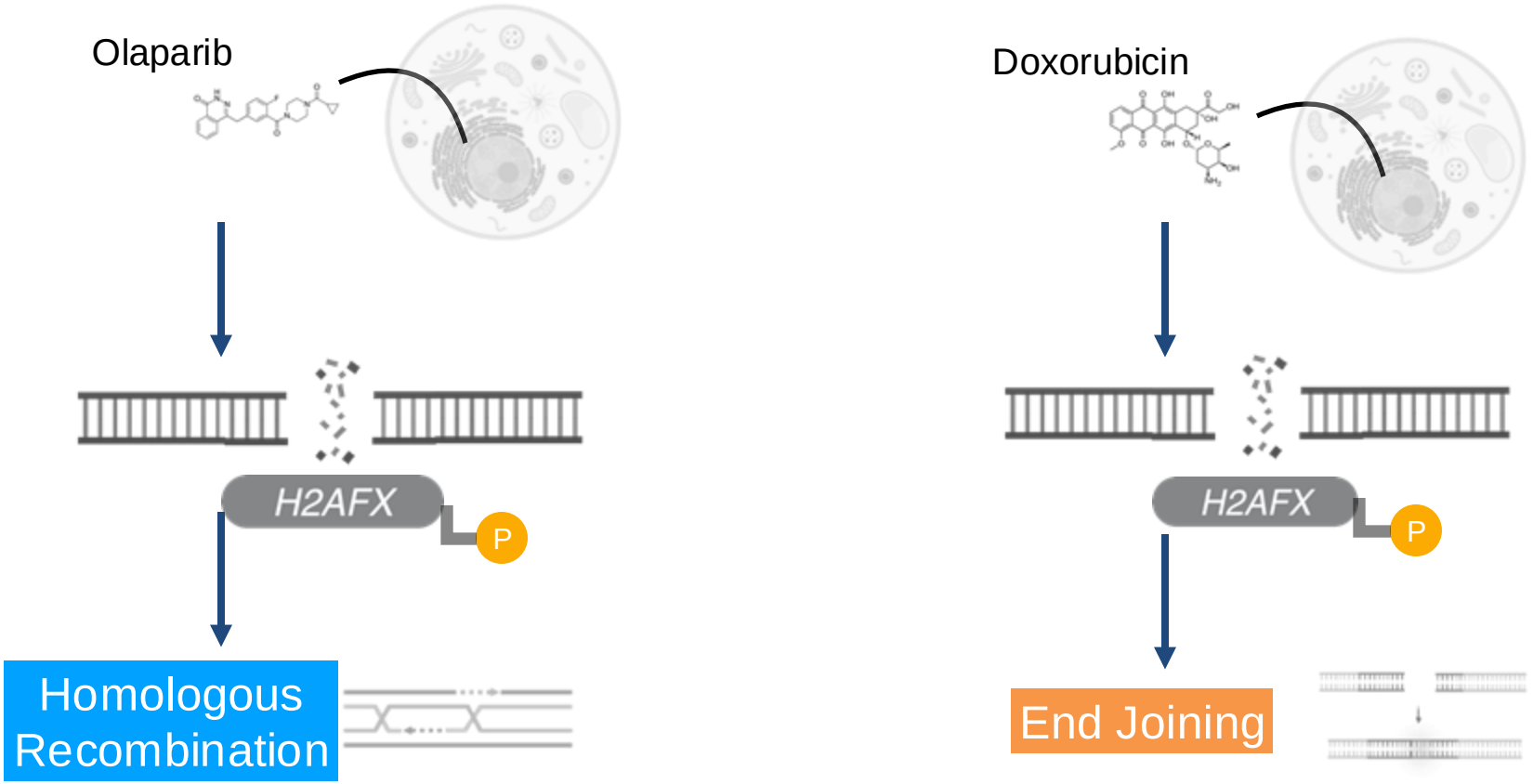
buy an apple **gift card**

buy an apple **tv**

H2AFX is a **pleiotropic** gene…



… whose **particular function** is resolved via **cell context**.



While unsupervised learning of word polysemy is **common**…

Data: corpus of sentence contexts

Approach: word embeddings w/ linear semantics

king - man + woman ≈ queen

unsupervised learning of gene pleiotropy is **unsolved**

Data: **?**

Approach: **?**

geneA - func1 + func2 ≈ geneB

Our goal for today

Unsupervised learning of gene pleiotropy with applications to therapeutic science

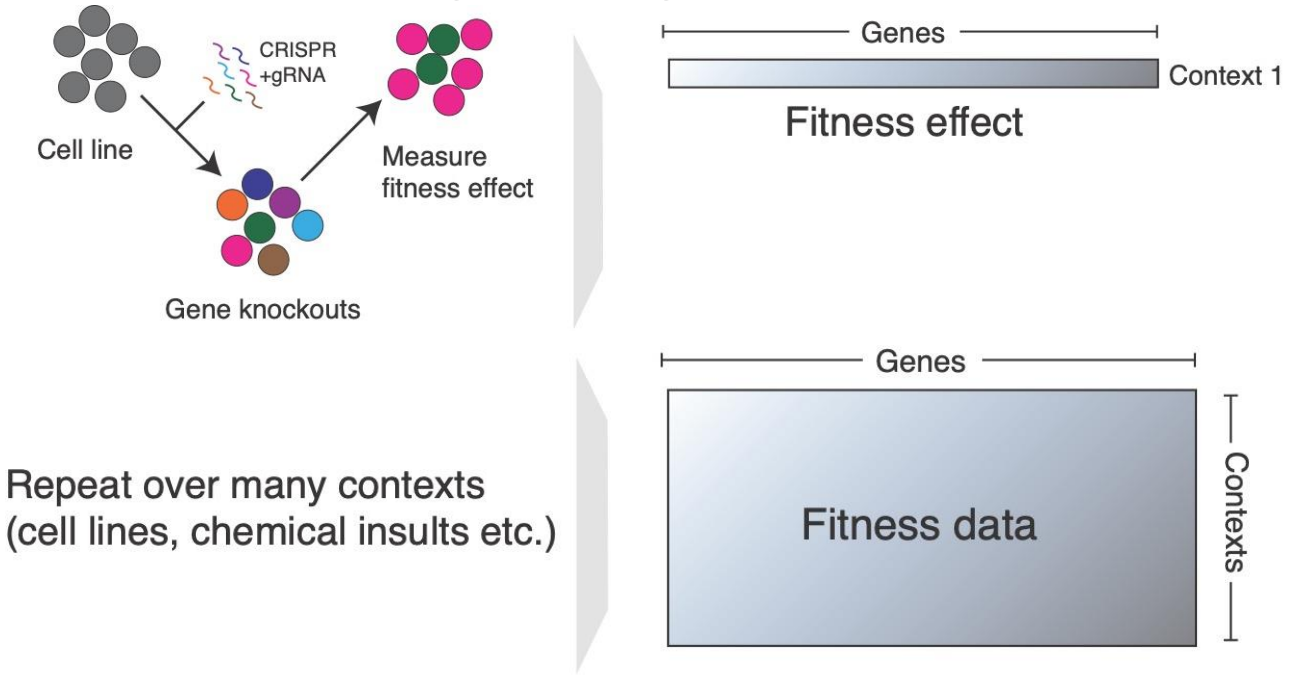
Data: ?

Approach: ?

geneA - func1 + func2 ≈ geneB

Data

Use **gene perturbation effect measurements** for inferring biological functions

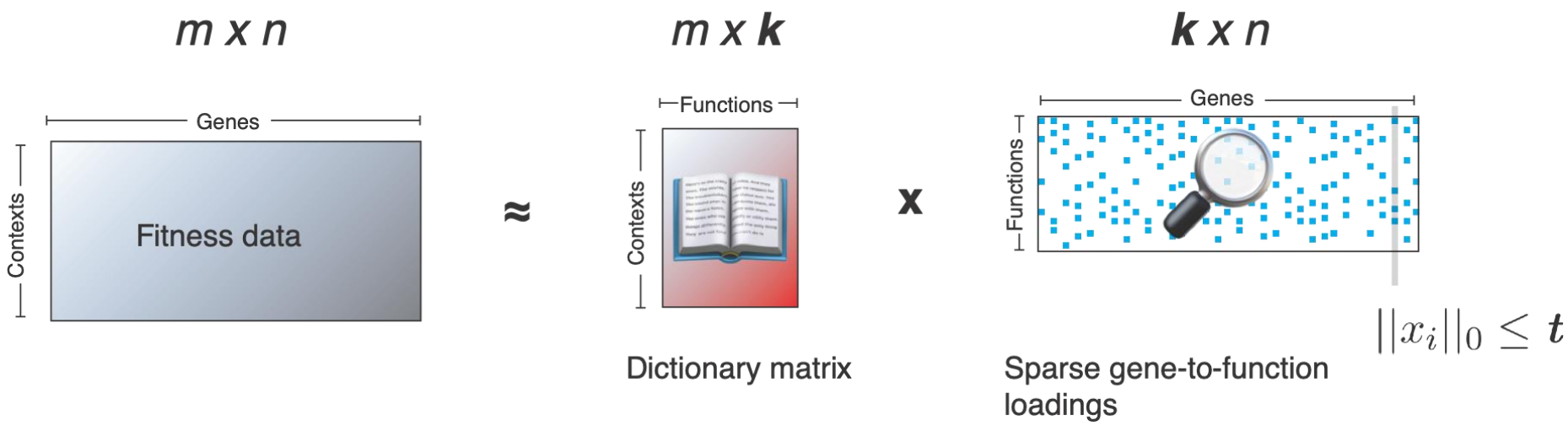


Why perturbation datasets? Alternative data types:

- **Transcriptomics:** gene co-expression is necessary but not sufficient for co-function
- **Protein-protein interactions:** direct interactions are not necessary for co-function

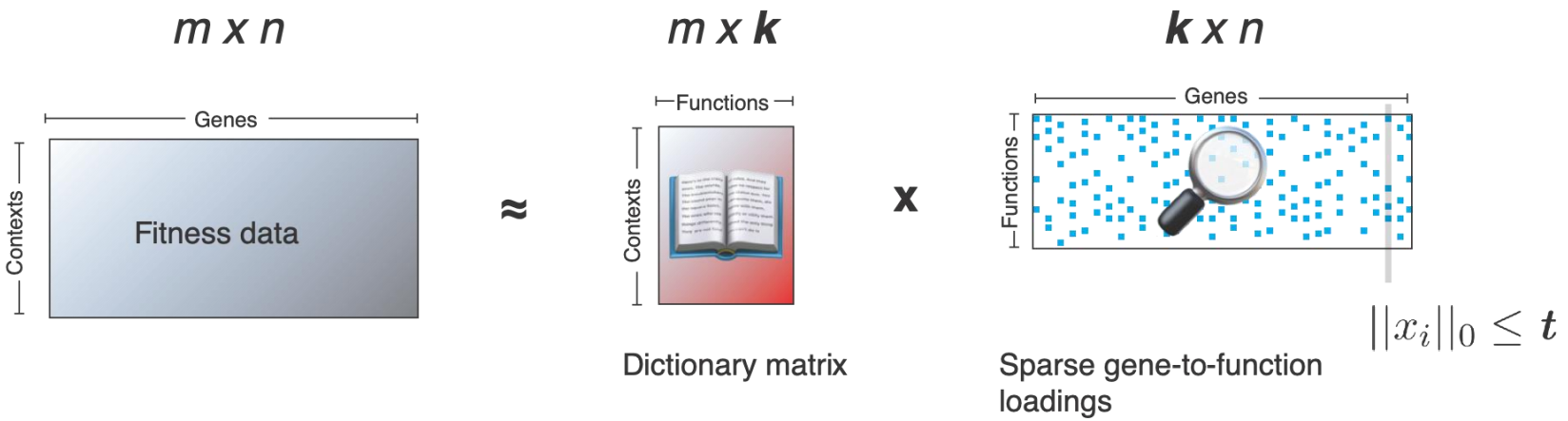
Approach: Webster

- **Low-dimensional vector embeddings** that satisfy three criteria:
  - Sparse
  - Latents are biologically meaningful
  - Account for redundancy between cell contexts



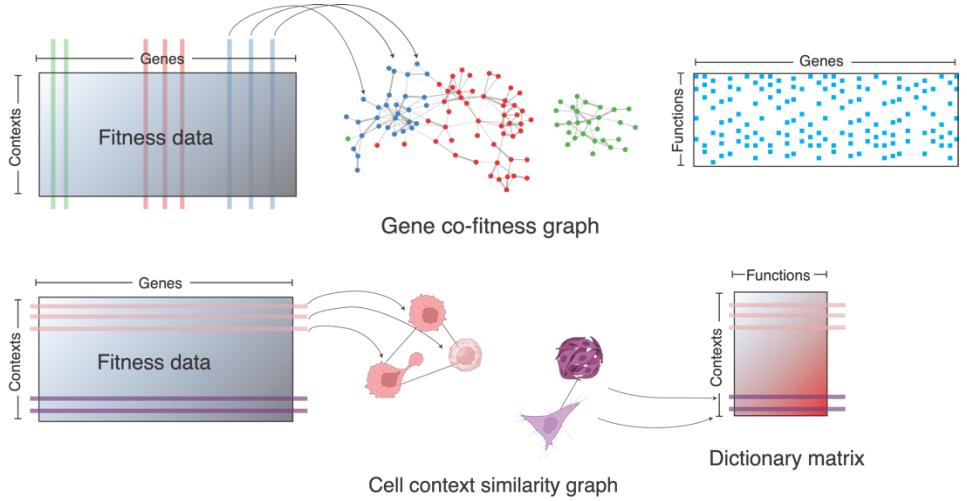
Approach: Webster

Webster learns a dictionary matrix that **sparsely** approximates gene effects…



1 … while preserving interpretable relationships between genes

2 … and accounting for redundancies between cell contexts

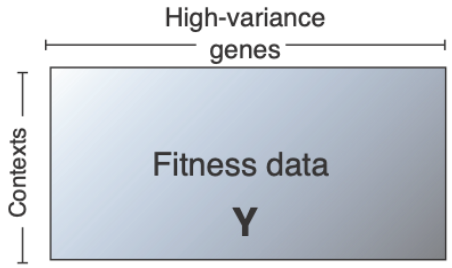


Overview of Webster

Preprocessing

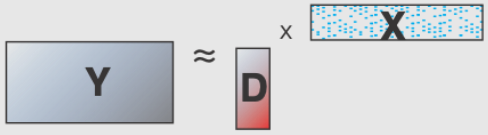
Raw fitness data

Standardize cell lines
Center gene effects
Filter genes by variance



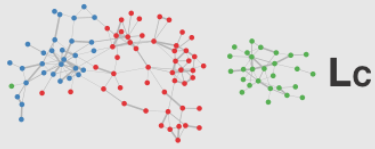
Graph-regularized dictionary learning
Objectives

Reduce dimensionality



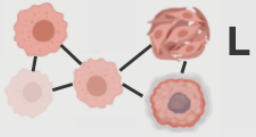
$$\|Y - DX\|_F^2$$

Preserve gene similarity



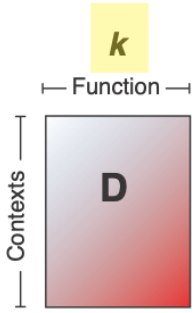
$$Tr(XL_cX^T)$$

Preserve cell context similarity

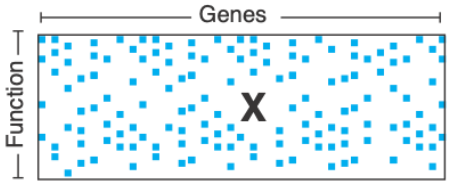


$$Tr(D^T LD)$$

Output



Dictionary matrix

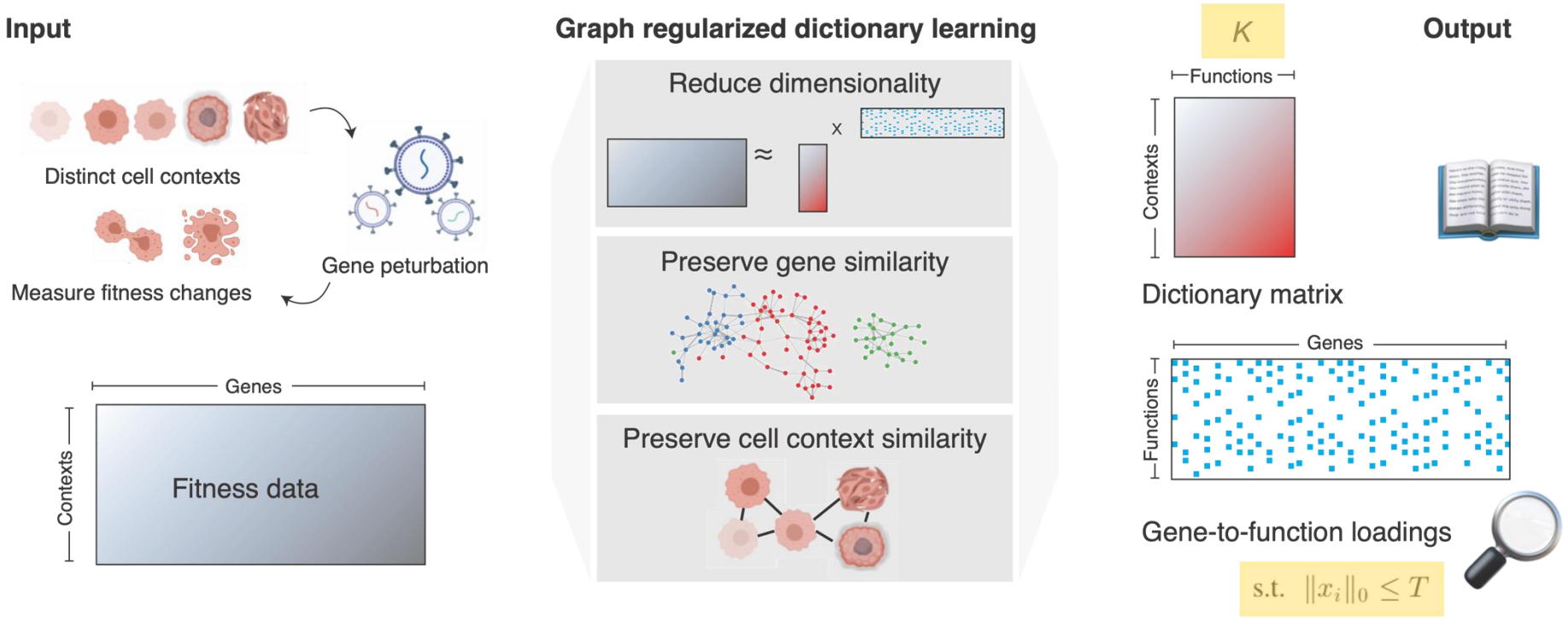


Gene-to-function loadings

$$||x_i||_0 \leq t$$

= key hyperparameters

Its key parameters are dictionary size (K) and sparsity on loadings (T)

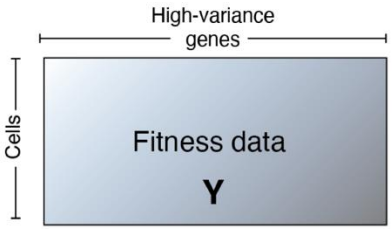


Model optimization

Preprocessing

Raw fitness data

Standardize cell lines
Filter genes by variance

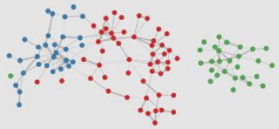


Graph-regularized dictionary learning
Objectives

Reduce dimensionality

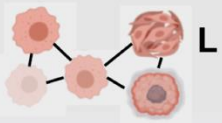
$$\|Y - DX\|_F^2$$

Preserve gene dependency manifold



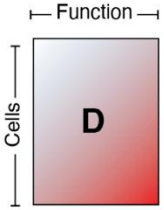
$$Tr(XL_cX^T)$$

Preserve cell state manifold

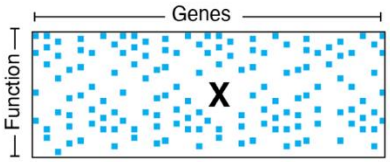


$$Tr(D^T LD)$$

Output



Dictionary matrix



Gene-to-funciton loadings

$$\text{s.t.} \quad \|x_i\|_0 \leq T$$

Parameters

$$\arg\min_{D,X} \|Y - DX\|_F^2 + \alpha Tr(D^T LD) + \beta Tr(XL_cX^T) \quad \text{s.t.} \quad \|x_i\|_0 \leq T \quad \forall i.$$

k	=latent dimension size
L	=cell Laplacian (num neighbors, metric)
Lc	=gene Laplacian (num neighbors, metric)
α	=weight of cell Laplacian
β	=weight of gene Laplacian
T	=sparsity

Applications to three screens of gene perturbation effects

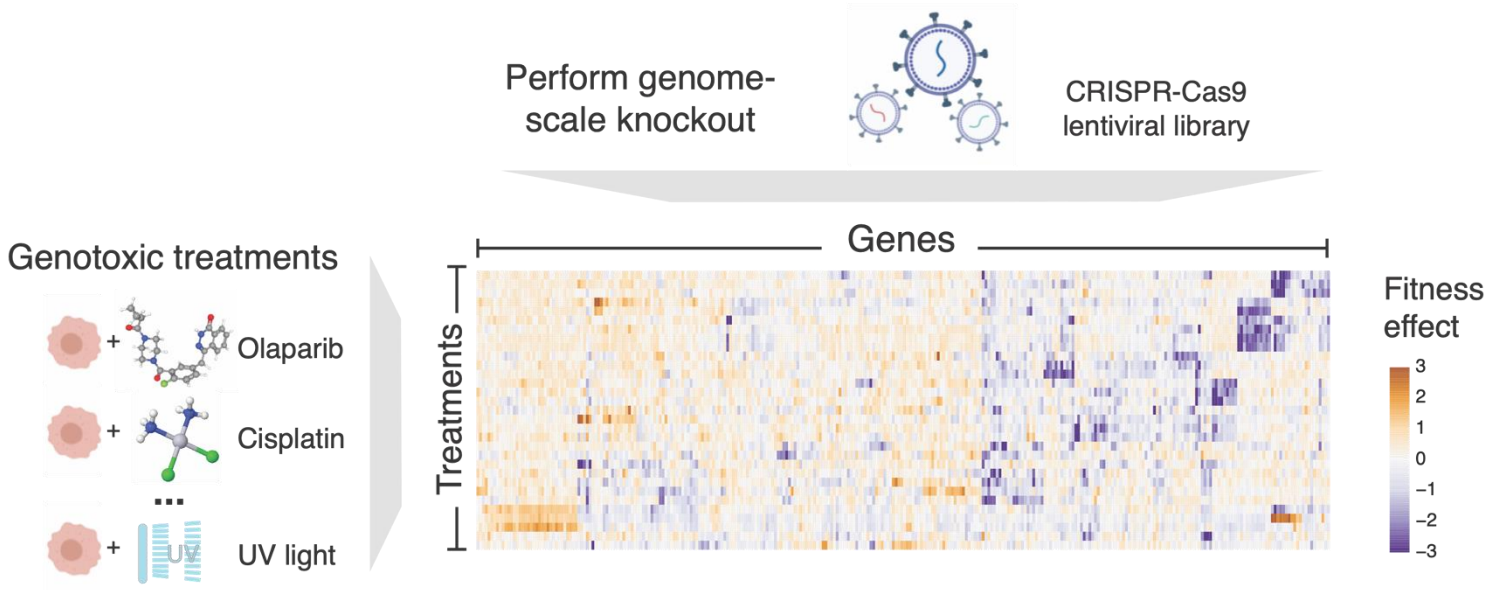
1) Genotoxic screens

2) Cancer fitness screens

3) Compound sensitivity screens

Part 1: Genotoxic screens

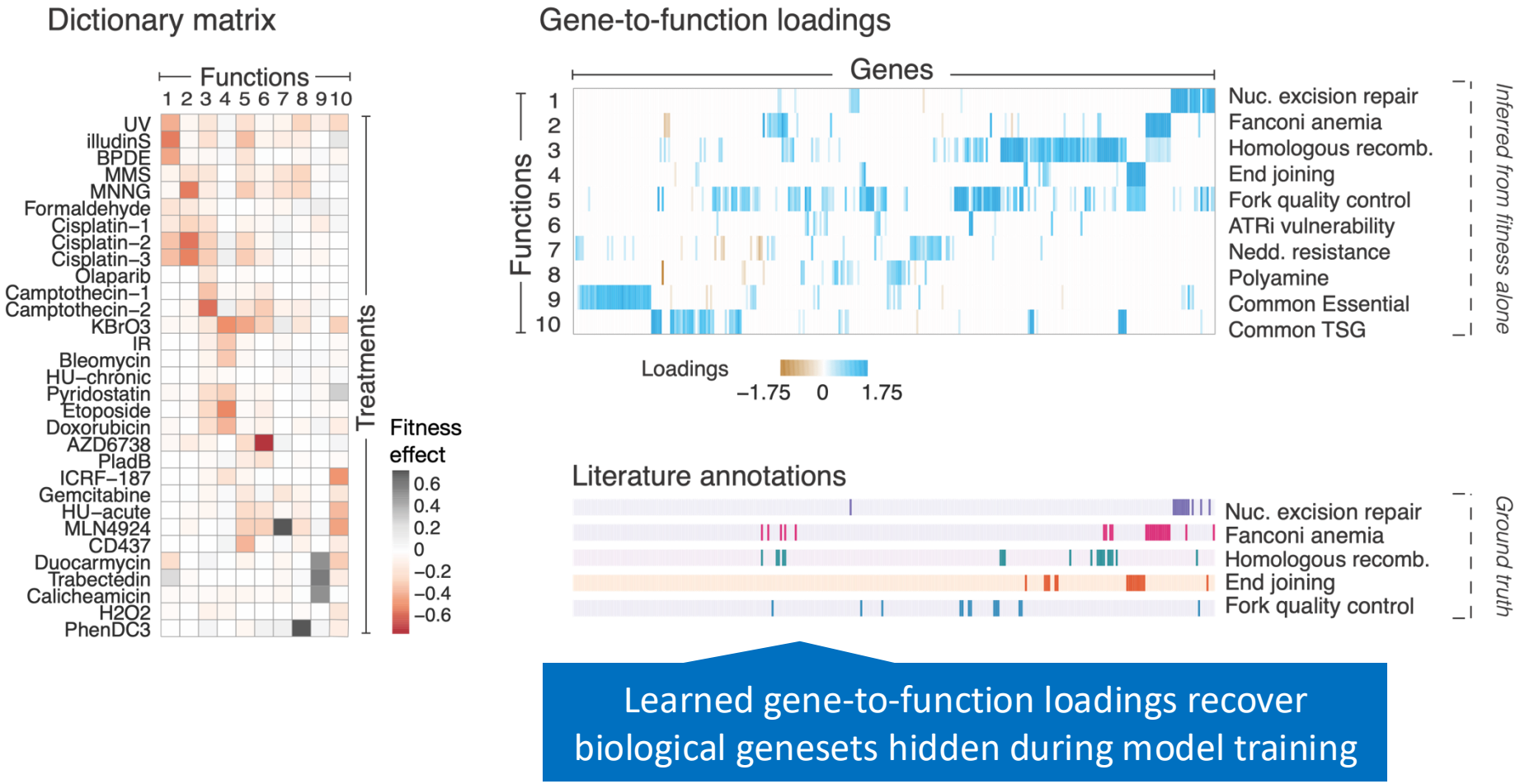
Olivieri et al. 2020: fitness effect of gene knockout in presence of genotoxins



Webster approximates the input data matrix…

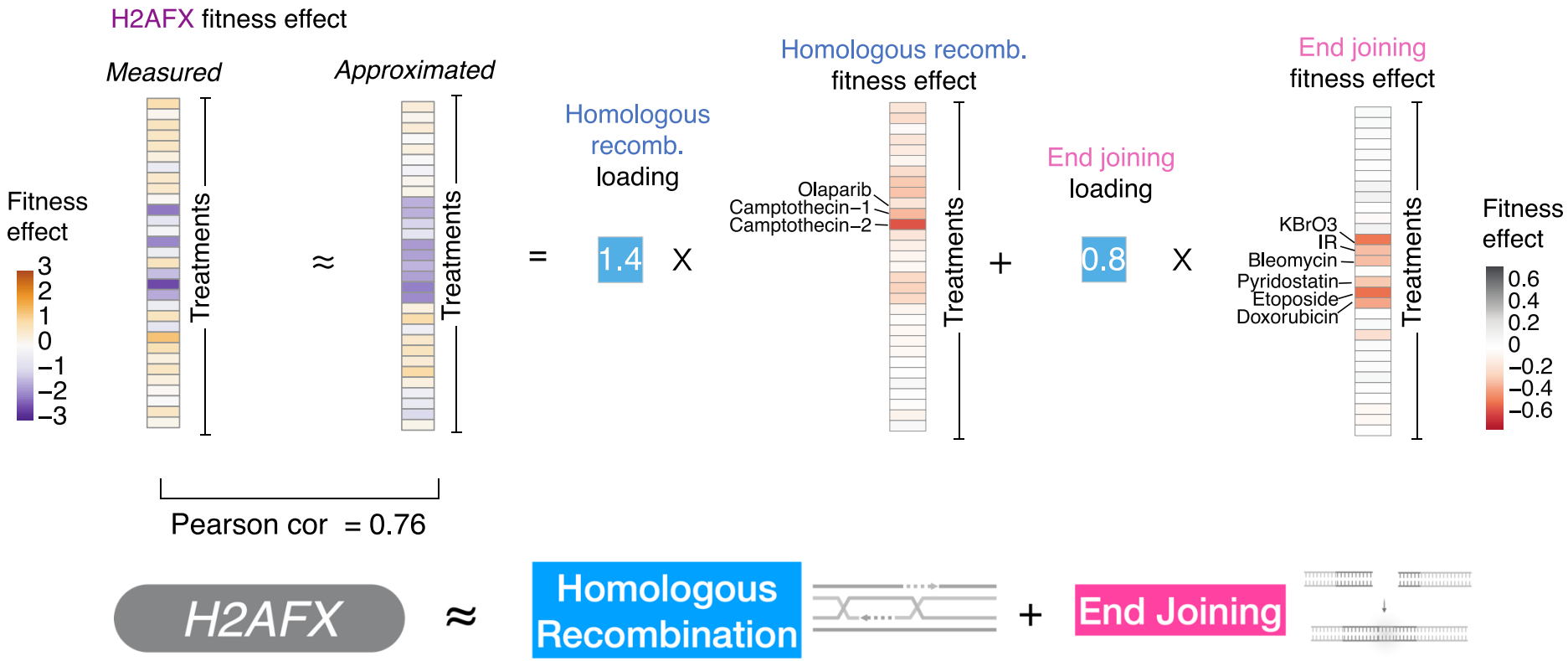
k=10
t=2

… as a product between a dictionary matrix and a loadings matrix



Learned gene-to-function loadings recover biological genesets hidden during model training

Latents inferred by the model recapitulate pleiotropy *without prior knowledge*

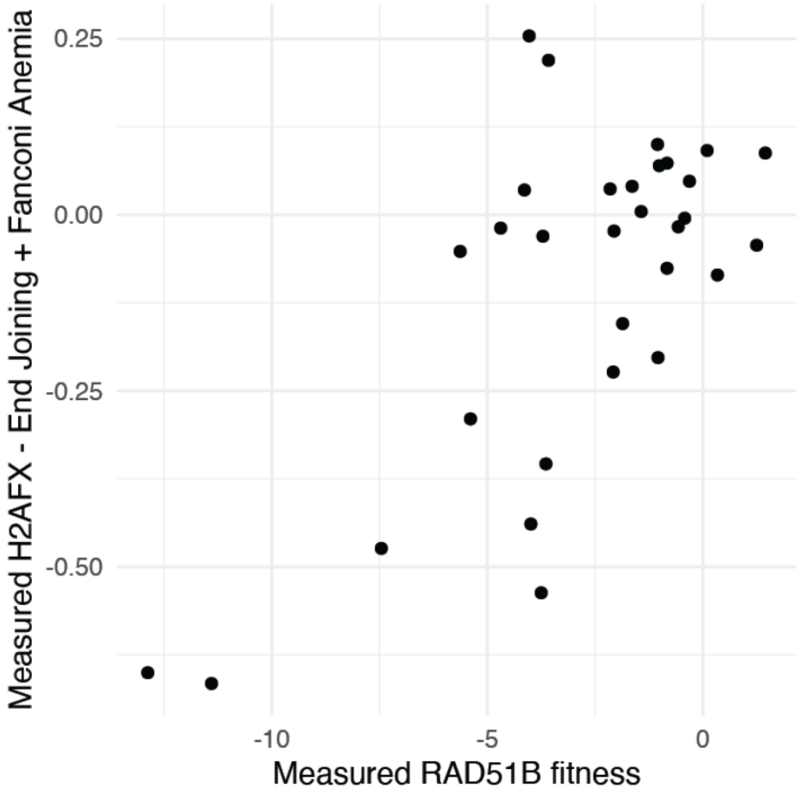


(hidden during model training!)

Latents are biologically meaningful

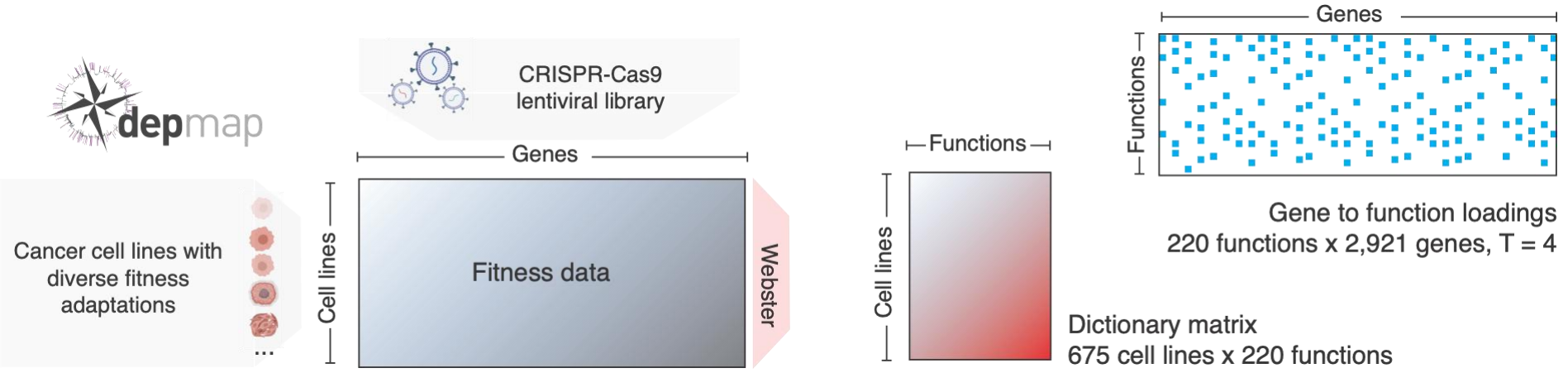
geneA - func1 + func2 ≈ geneB

H2AFX - End Joining + Fanconi Anemia ≈ RAD51B



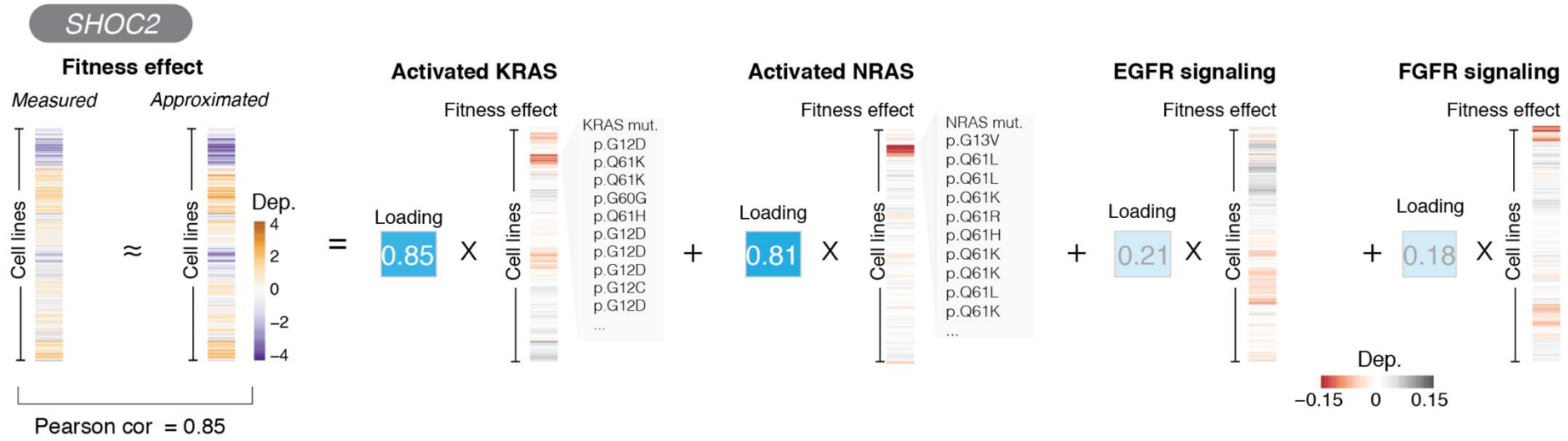
= cell context (treatment)

Part 2: Cancer fitness screens

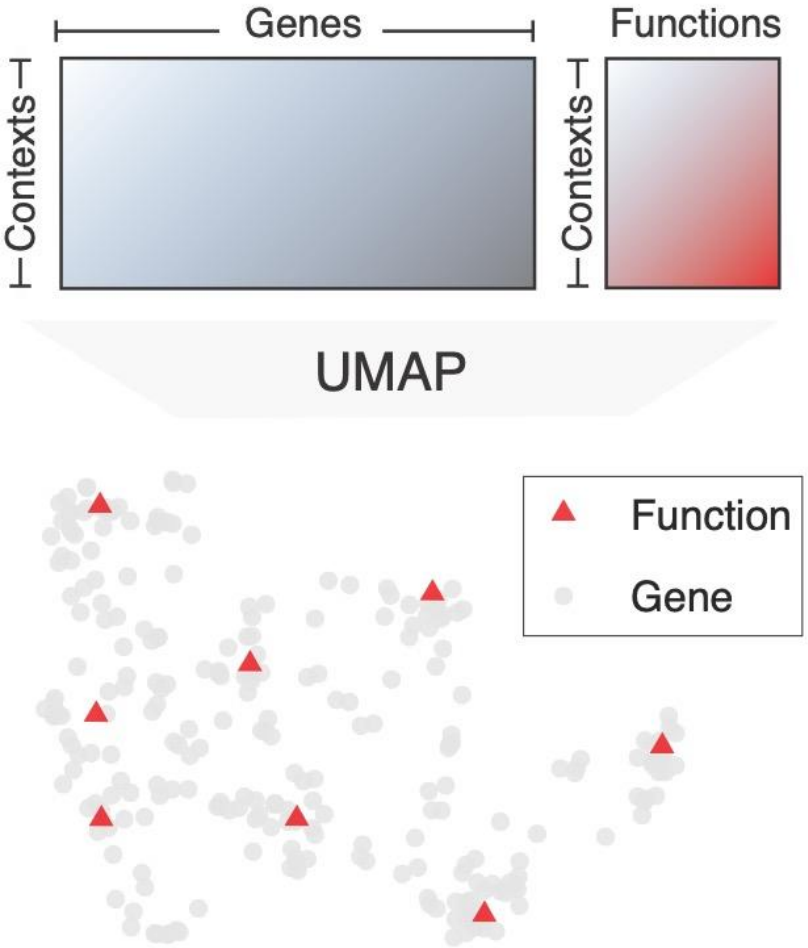


Pleiotropic genes obey linear semantics in the latent space

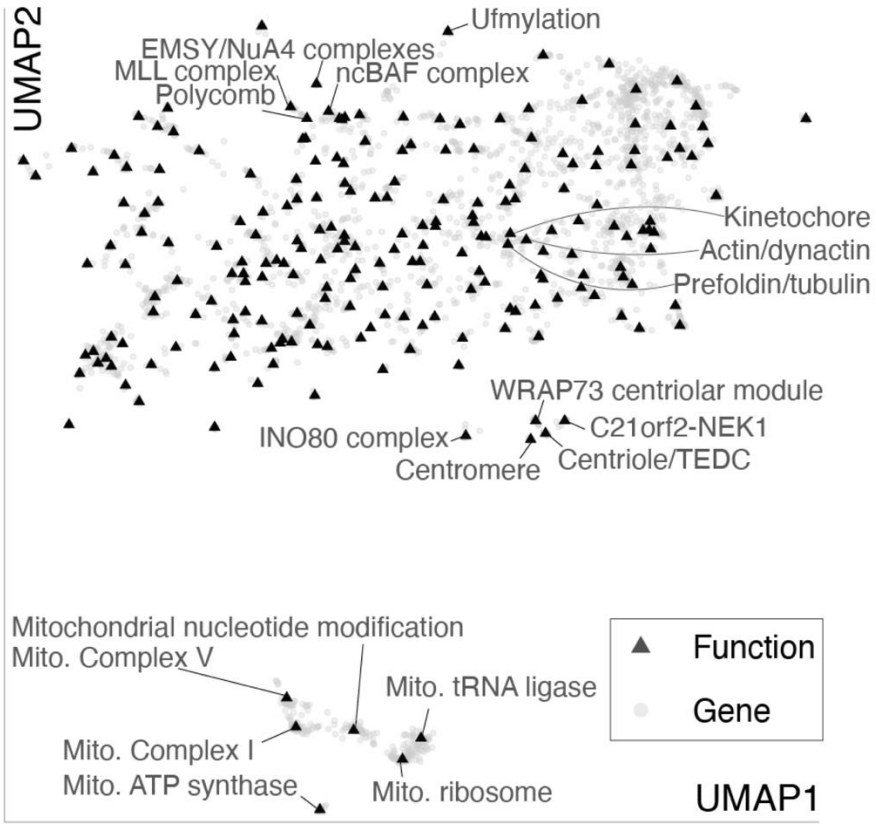
SHOC2 ≈ Activated KRAS + Activated NRAS + EGFR Signaling + FGFR Signaling



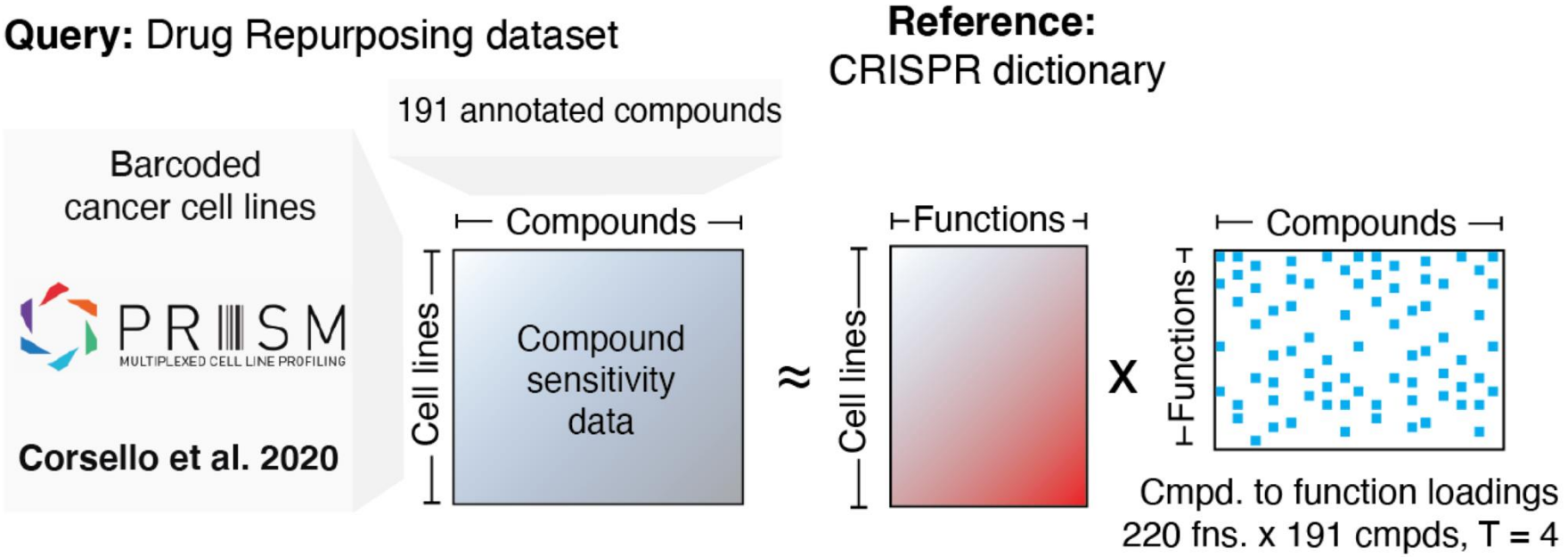
Joint embedding space of genes and functions



It captures interpretable processes in cancer

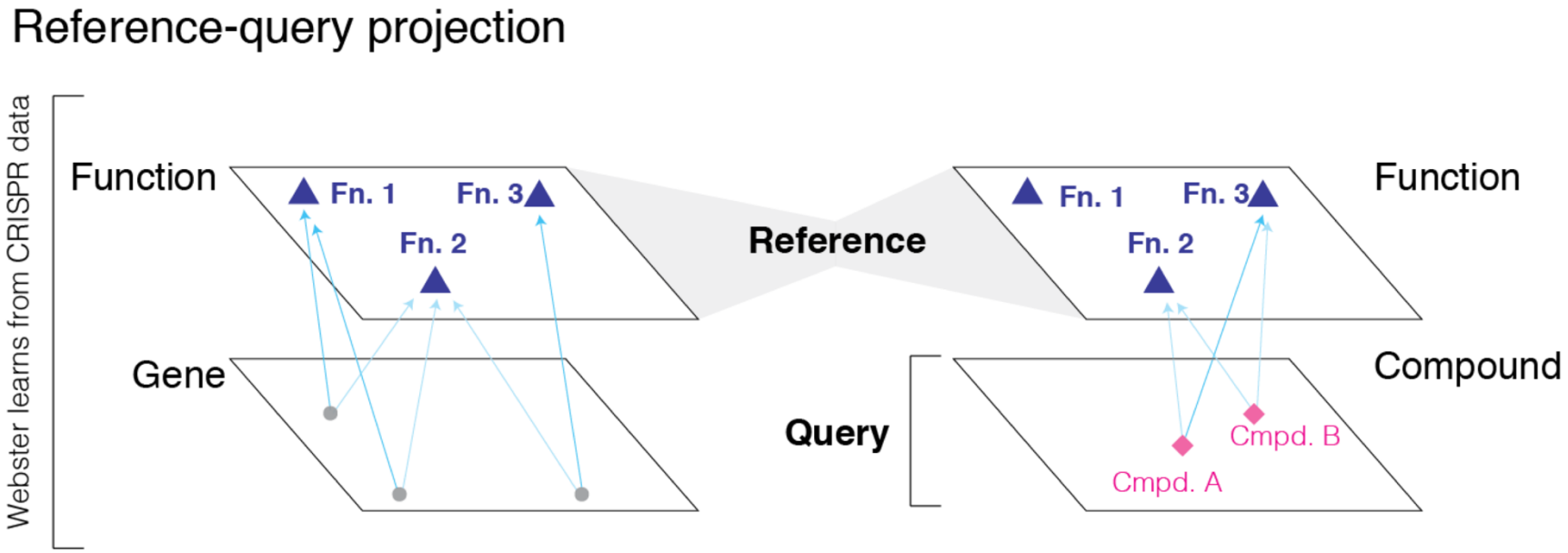


Part 3: Compound sensitivity screens



Modeling compound sensitivity profiles as mixtures of functions learned from CRISPR

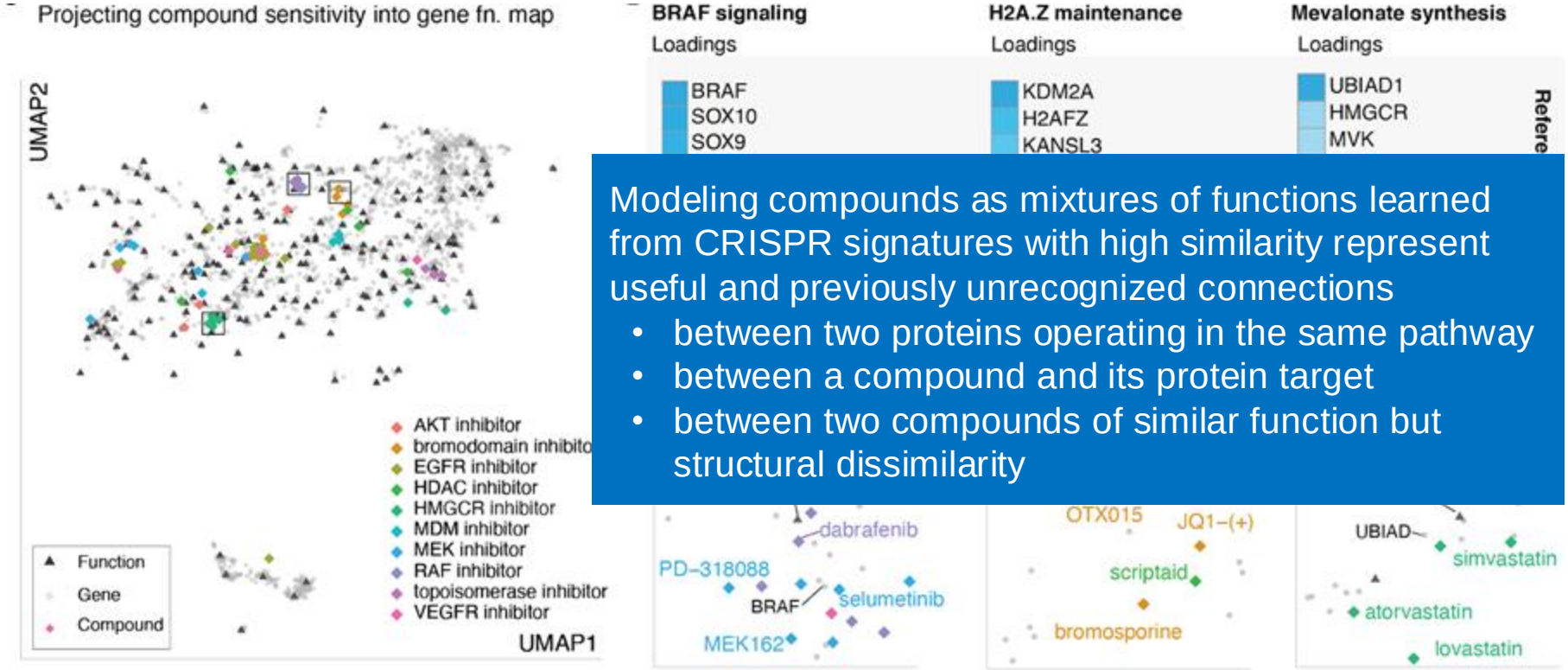
Modeling compounds as mixtures of latent functions



- Modeling compounds as mixtures of functions learned from CRISPR signatures with high similarity represent useful and previously unrecognized connections
  - between two proteins operating in the same pathway
  - between a small-molecule and its protein target
  - between two small-molecules of similar function but structural dissimilarity
- Such a catalog of connections can serve as a functional look-up table of compounds to predict sensitivity and genotoxic profiles and to inform therapeutic use

Compounds' mechanisms of action

Compounds are embedded nearby gene functions, reflecting their mechanism of action



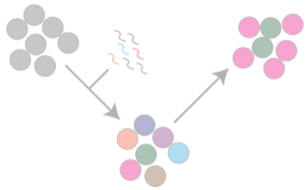
Modeling compounds as mixtures of functions learned from CRISPR signatures with high similarity represent useful and previously unrecognized connections

- between two proteins operating in the same pathway
- between a compound and its protein target
- between two compounds of similar function but structural dissimilarity

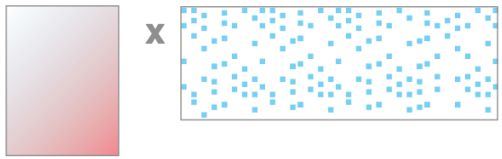
Key takeaways

- Analogously to word semantics, genes can be modeled as **distributions over latent bio functions**
  - **Sparse learning** is an effective strategy for learning bio functions from high-dimensional chemical and genetic perturbations
  - New perturbations can be **projected** into learned space

Data: high-dimensional gene perturbation measurements

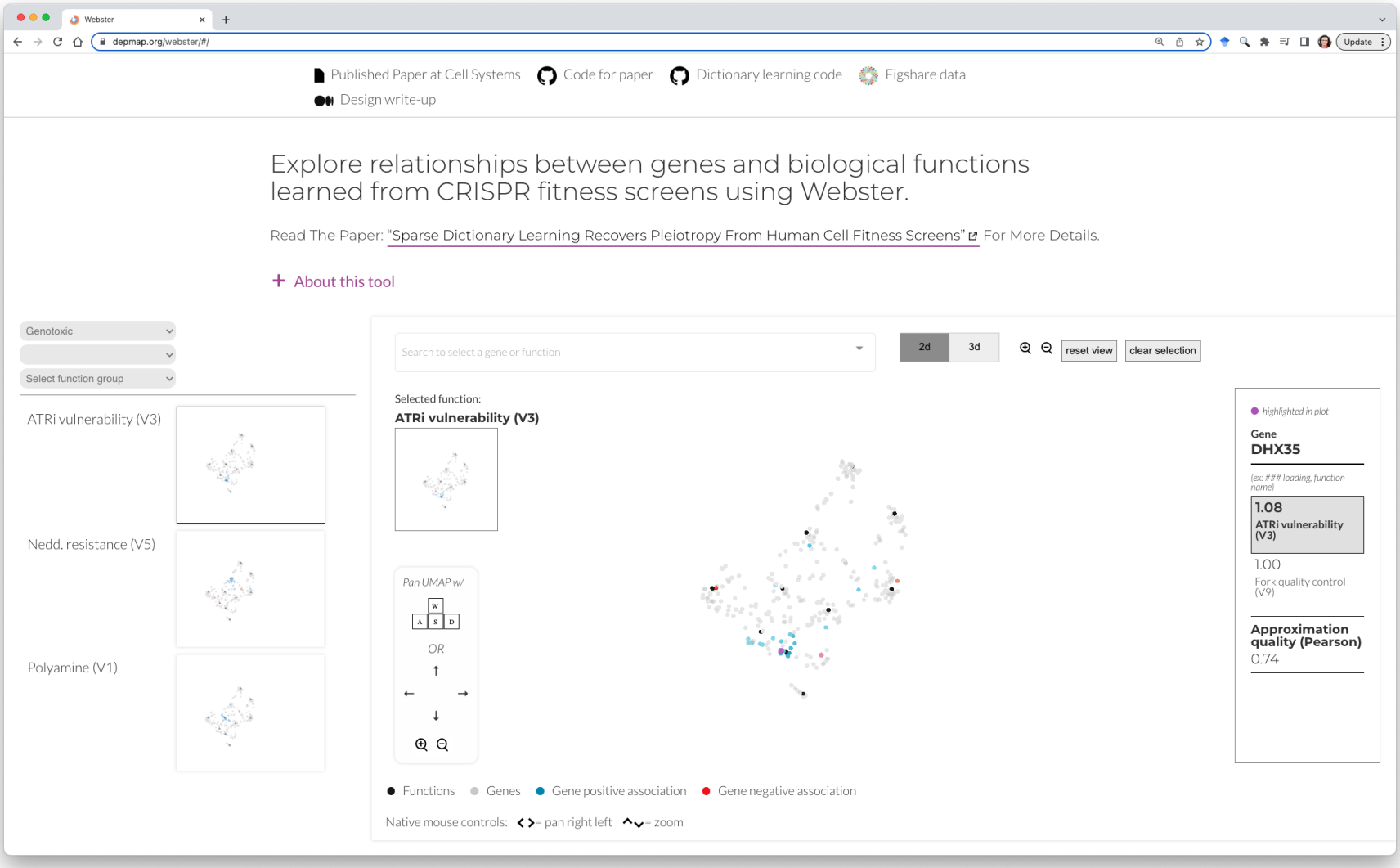


Approach: sparse approximation embeddings



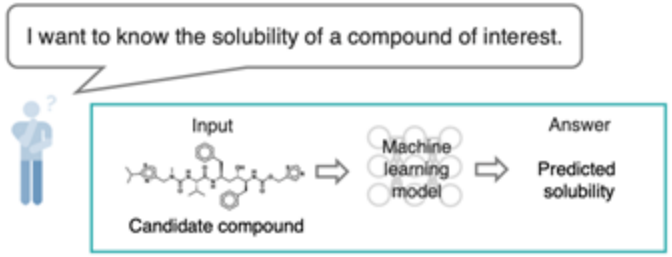
geneA - func1 + func2 ≈ geneB

https://depmap.org/webster

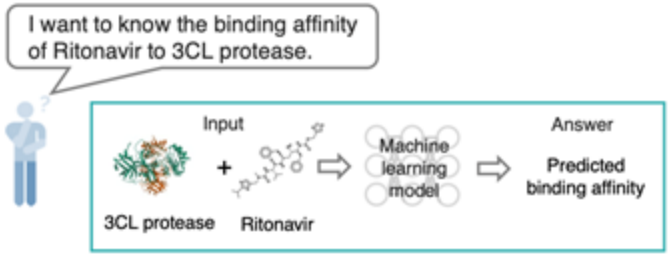


Outline for today's class

- Optimization & generation of small molecules



- Binding of drugs to therapeutic targets



- High-throughput genetic & chemical perturbations