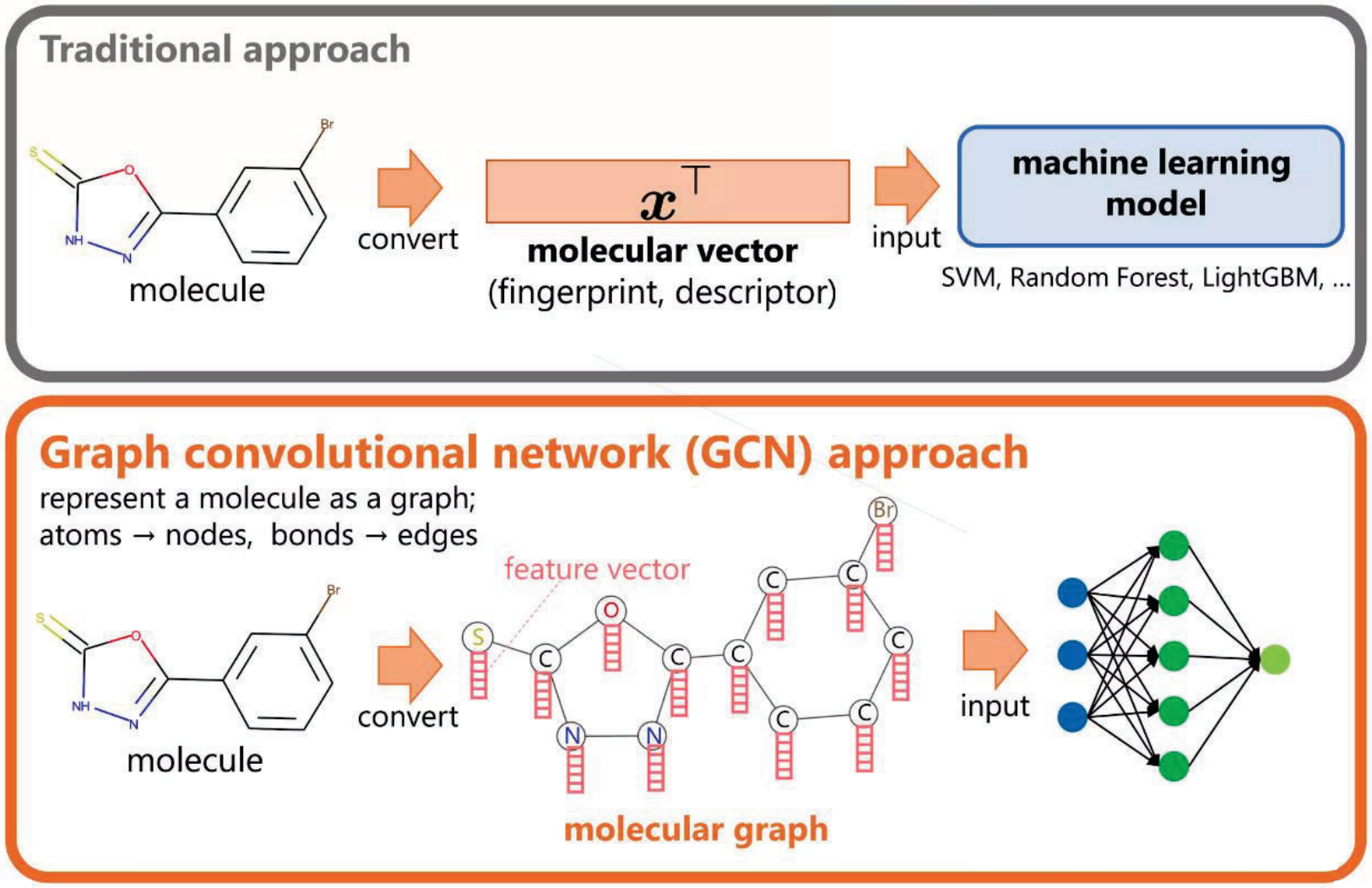
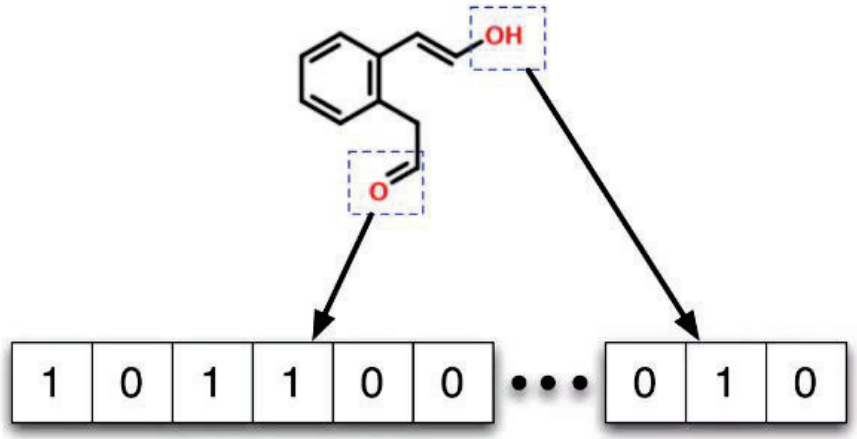
Two strategies for producing molecular representations



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

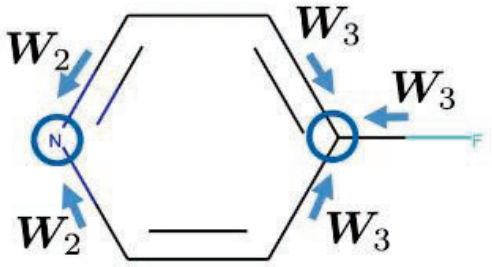
Input: molecule, radius R, hidden weights $H_1^1 \dots H_R^5$, output weights $W_1 \dots W_R$
Initialize: fingerprint vector $\mathbf{f} \leftarrow \mathbf{0}_S$
for each atom a in molecule
  $\mathbf{r}_a \leftarrow g(a)$ ▷ lookup atom features
for $L = 1$ to R ▷ for each layer
  for each atom a in 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.

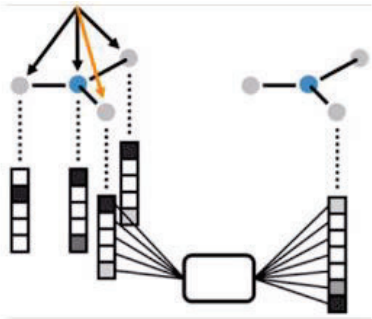
Duvenaud et al., NeurIPS 2015

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



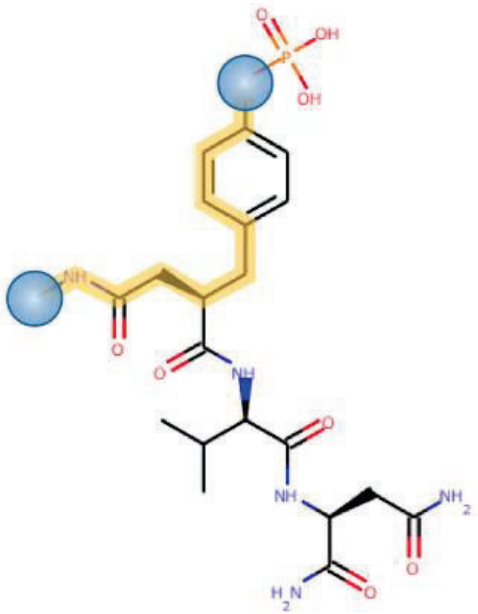
2) **Molecular graphs**
   - Update atom features by convolutional and pooling layers using adjacent atoms



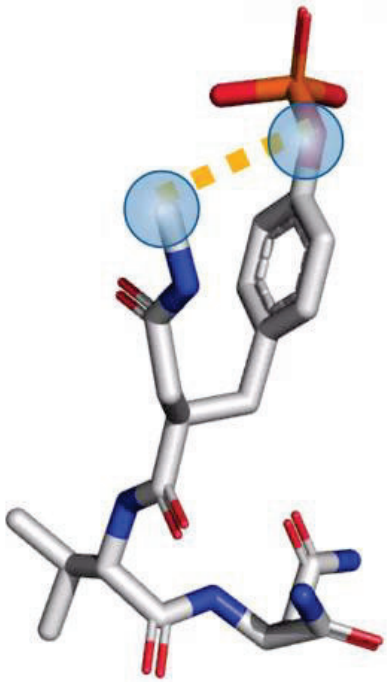
- **They did not consider property of edges (bonds)**
- **They did not consider atoms other than 1-neighbor**

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

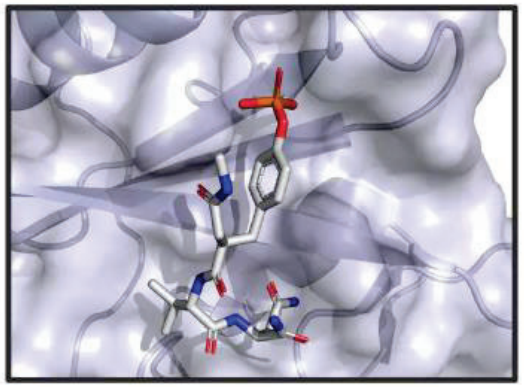
Graphs vs. 3D structures



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)

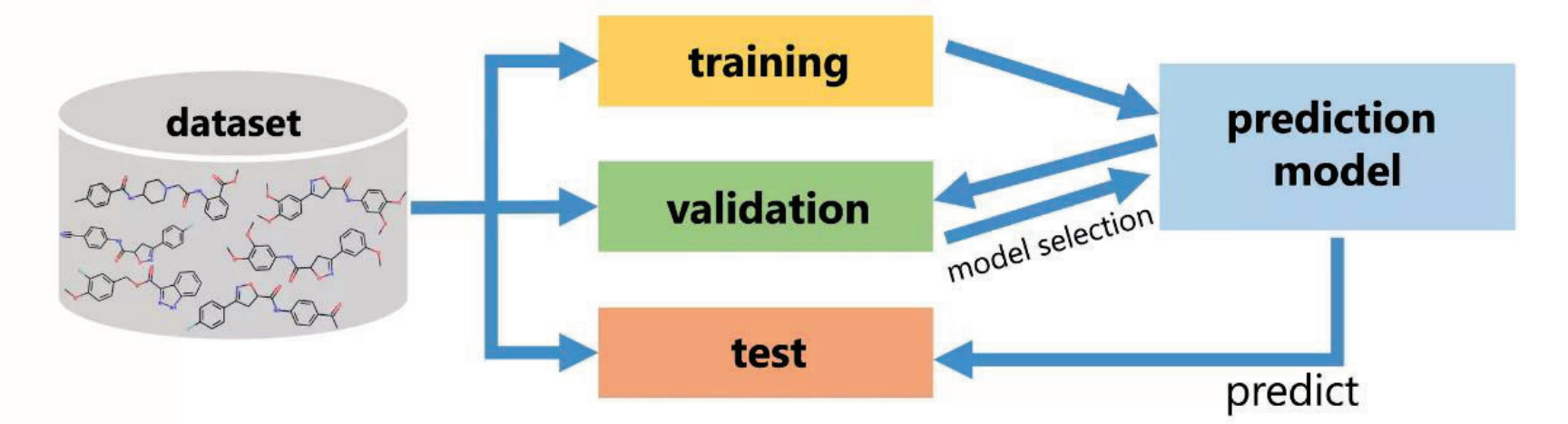
M: Metabolism
CYP2C9/2D6/3A4 Inhibition
CYP2C9/2D6/3A4 Substrate

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

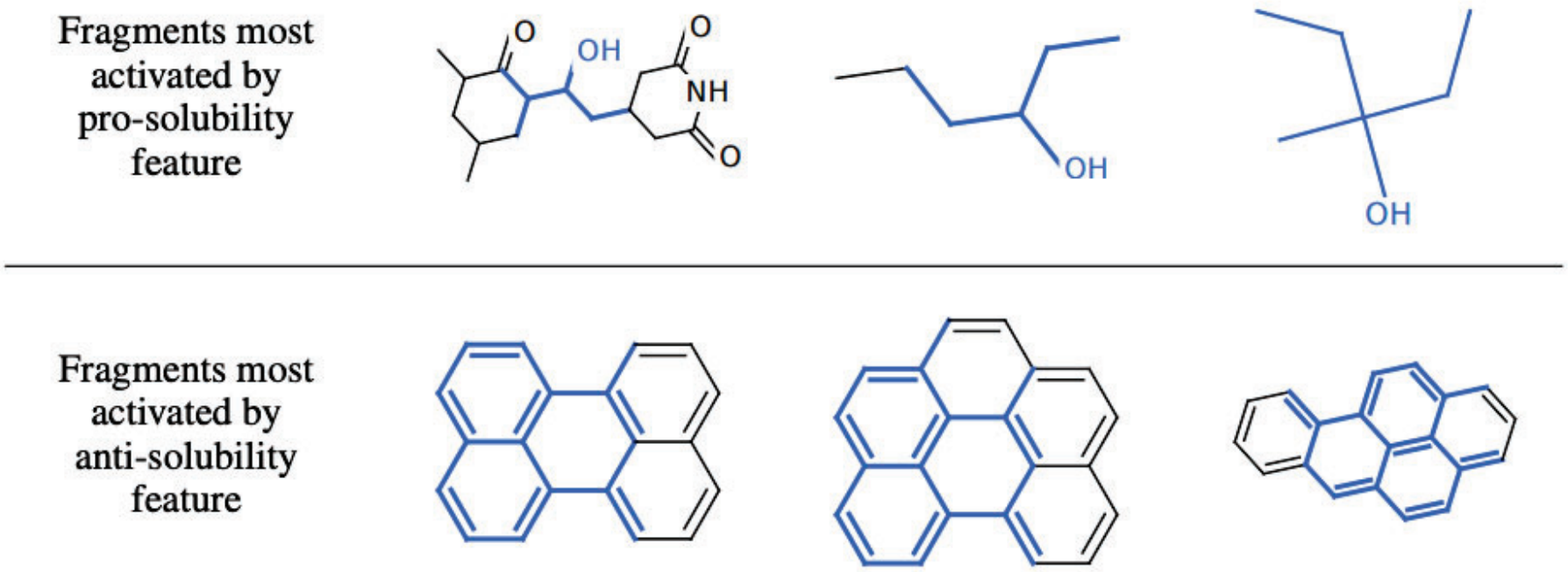


Figure 4: Examining fingerprints optimized for predicting solubility. Shown here are representative examples of molecular fragments (highlighted in blue) which most activate different features of the fingerprint. *Top row:* The feature most predictive of solubility. *Bottom row:* The feature most predictive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

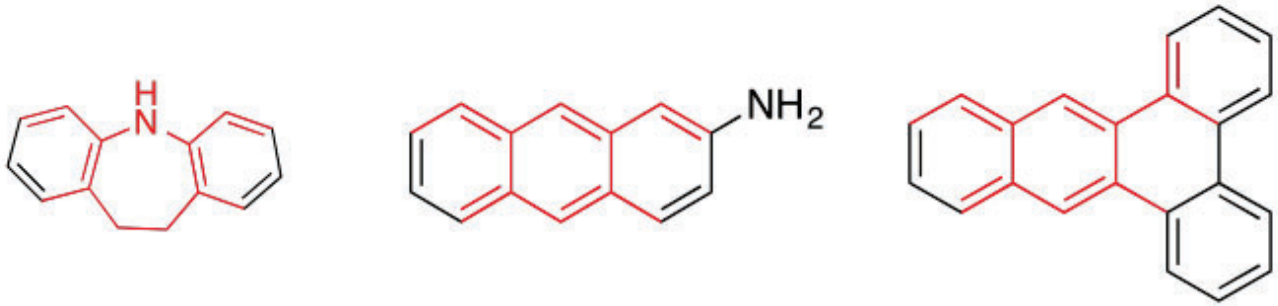


Figure 5: Visualizing fingerprints optimized for predicting toxicity. Shown here are representative samples of molecular fragments (highlighted in red) which most activate the feature most predictive of toxicity. *Top row:* the most predictive feature identifies groups containing a sulphur atom attached 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

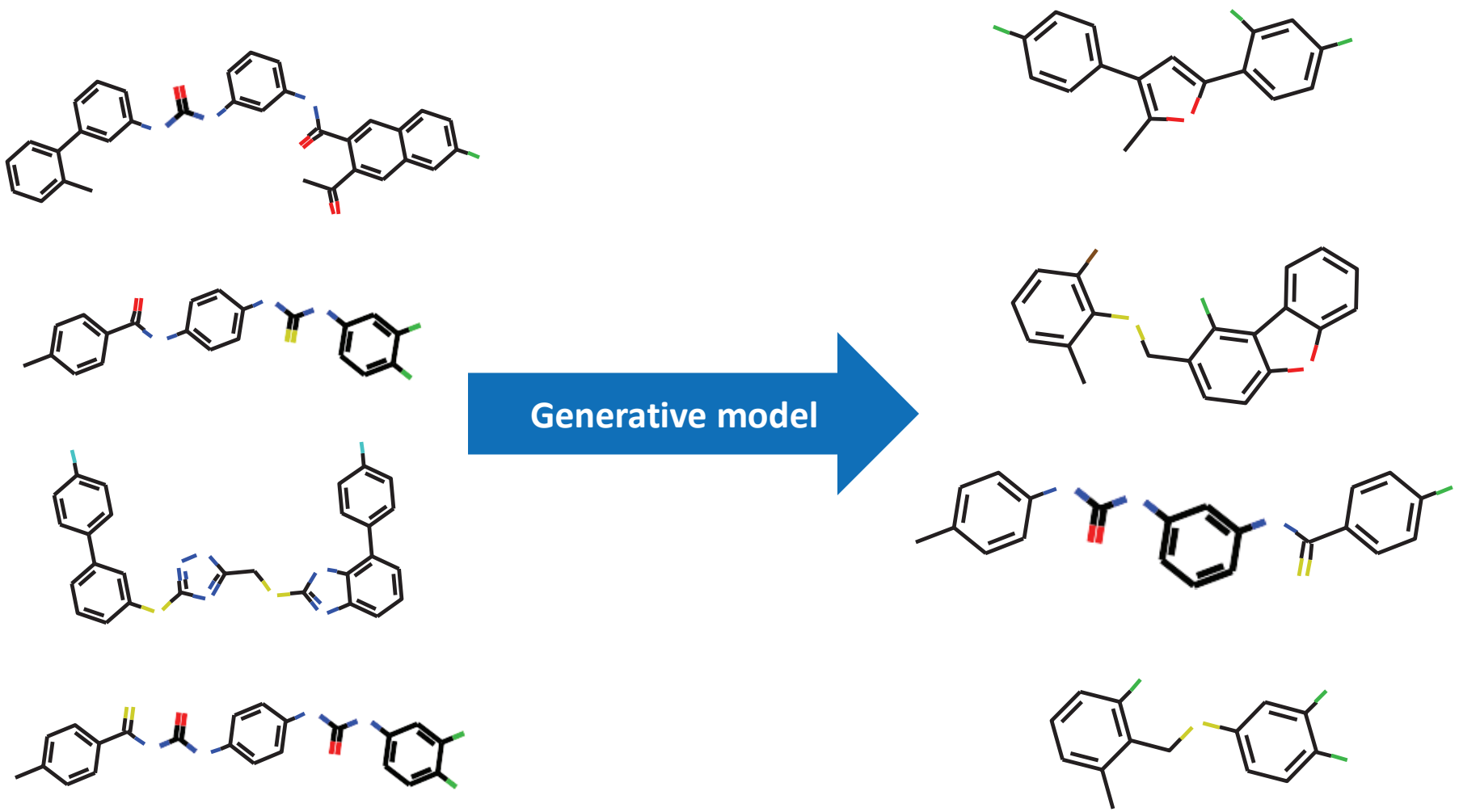
Results: Molecular property prediction

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

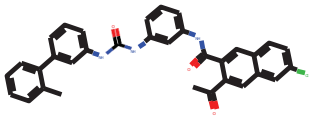
- No single method performs the best across all scenarios
- Pre-training boost performance
- Pre-trained graph models yield strongest predictors overall

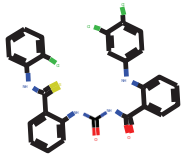
Therapeutics Data Commons: Machine Learning Datasets and Tasks for Therapeutics, NeurIPS, 2021
Artificial Intelligence Foundation for Therapeutic Science, *Nature Chemical Biology,* 2022

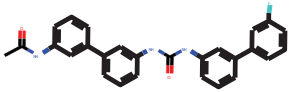
Molecular graph generation

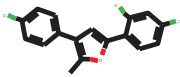


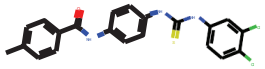
Molecular graph generation

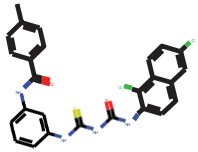

5.30

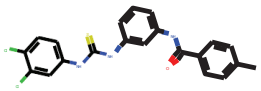

4.93

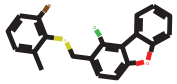

4.49

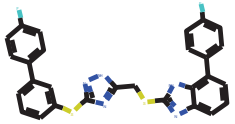

4.45

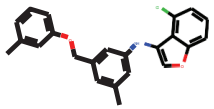

4.42

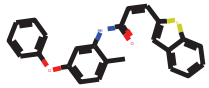

4.40

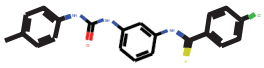

4.37

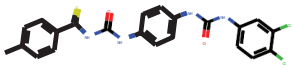

4.30

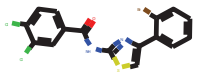

4.23

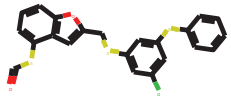

4.18

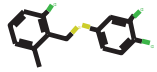

4.17


4.08


4.07

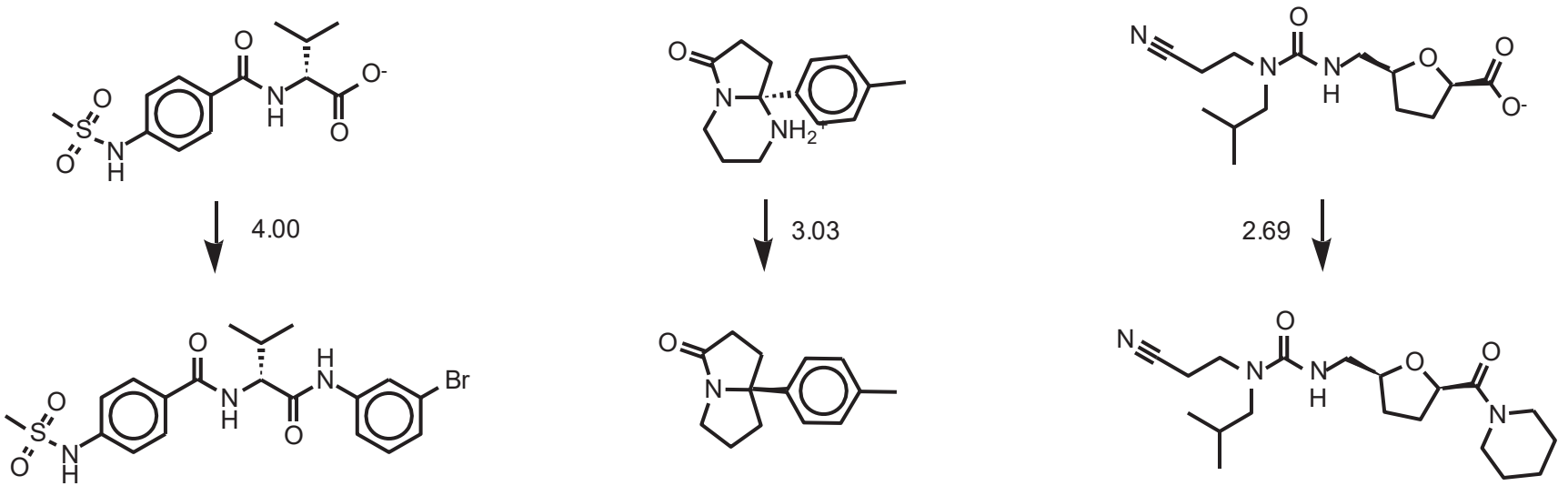

4.04


4.04


4.03

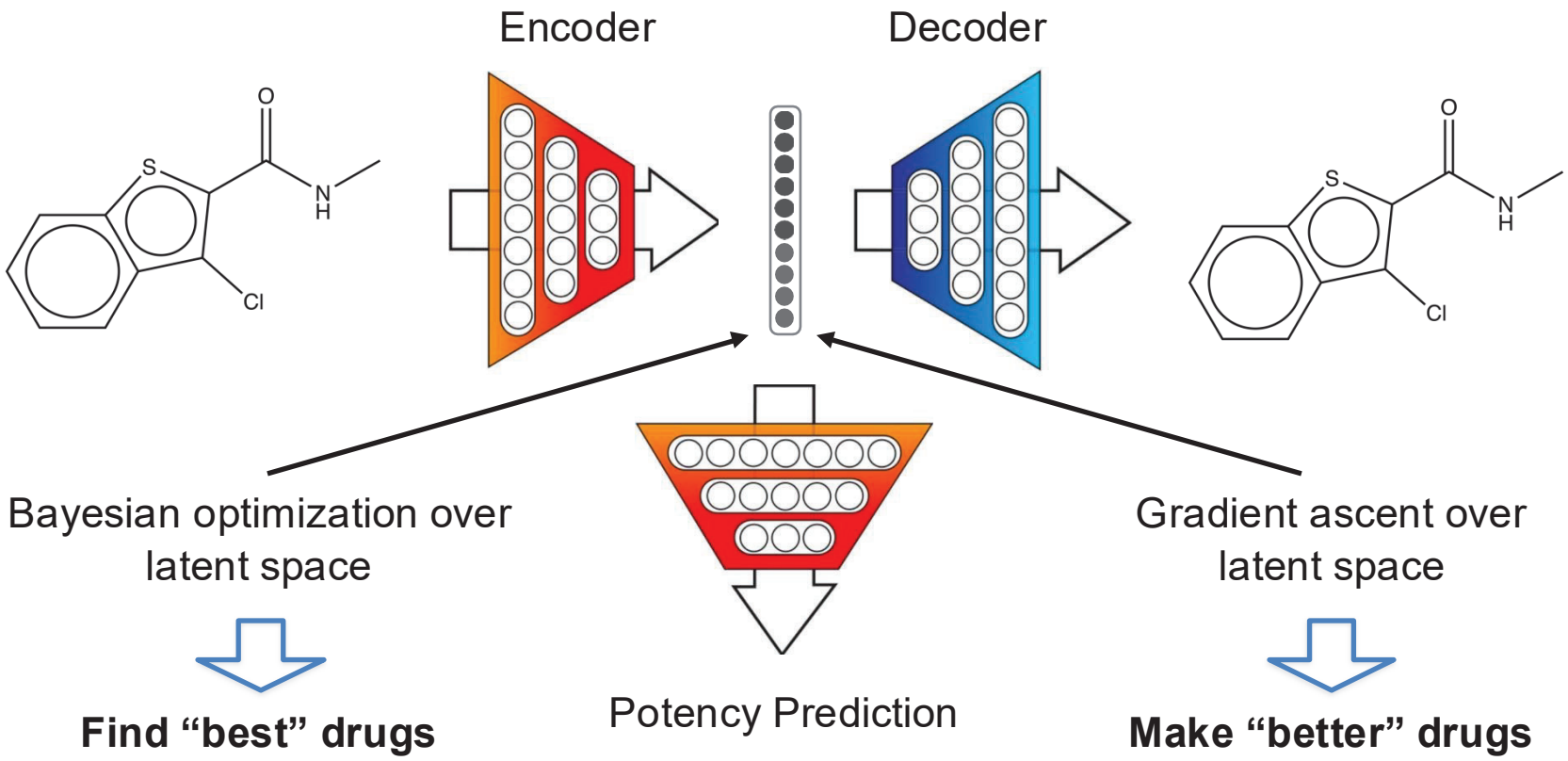
Generate molecules with high potency

Molecular graph generation



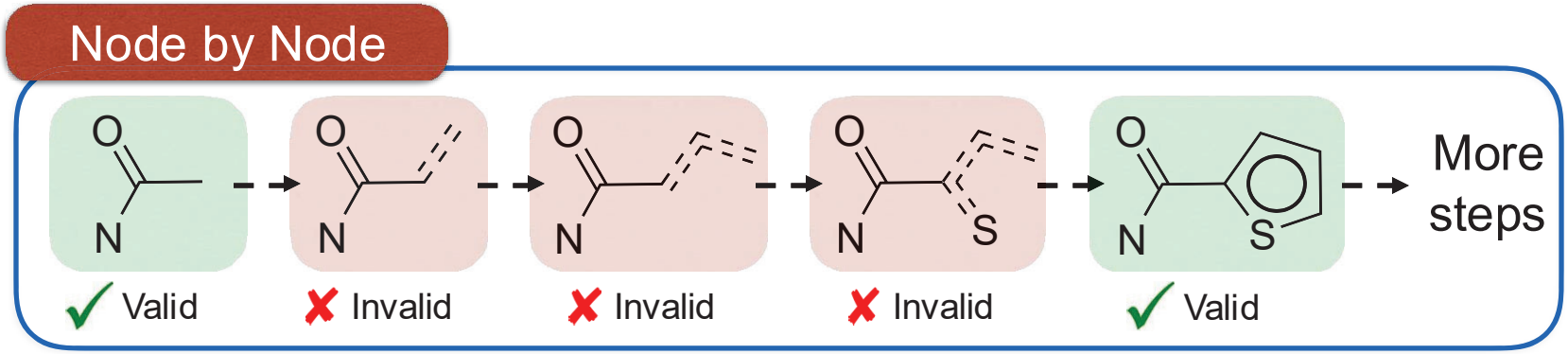
Modify molecules to increase potency

Molecular variational autoencoder



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

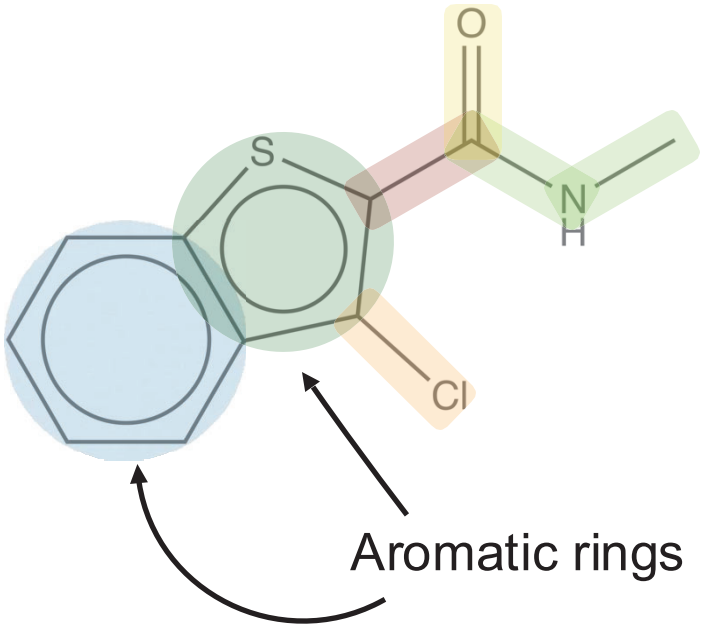
How to generate graphs?



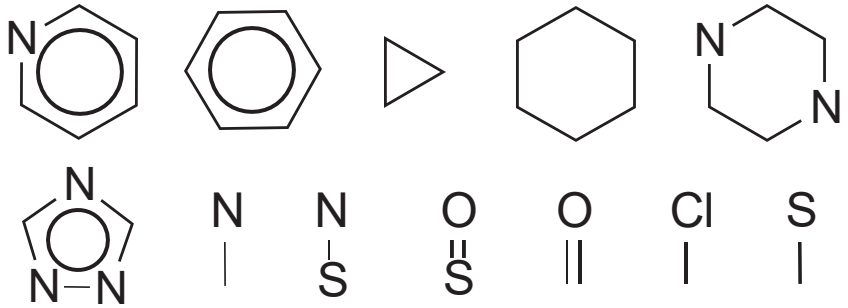
- Not every graphs is chemically valid
- Invalid intermediate states ⟶ hard to validate
- Very long intermediate steps ⟶ difficult to train (Li et al., 2018)

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

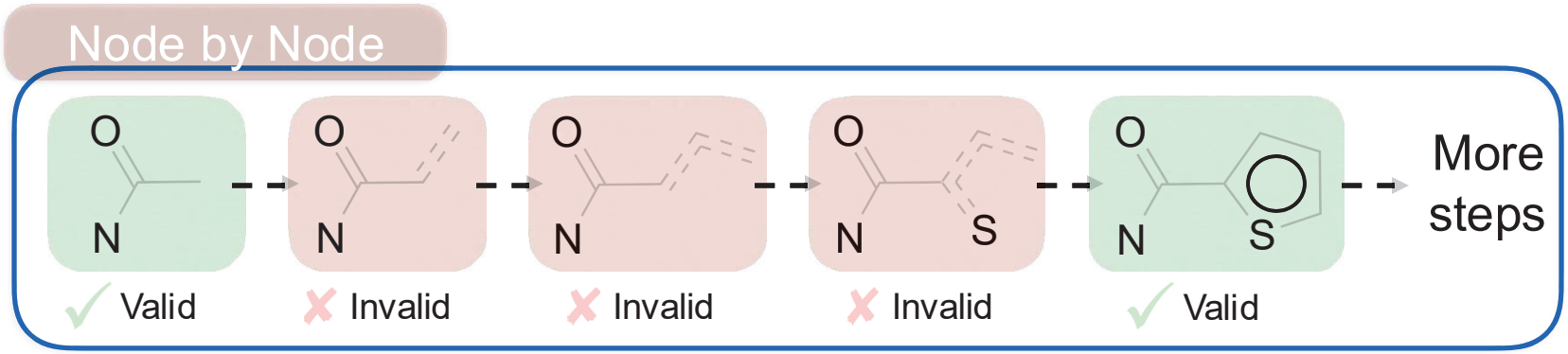
Functional groups

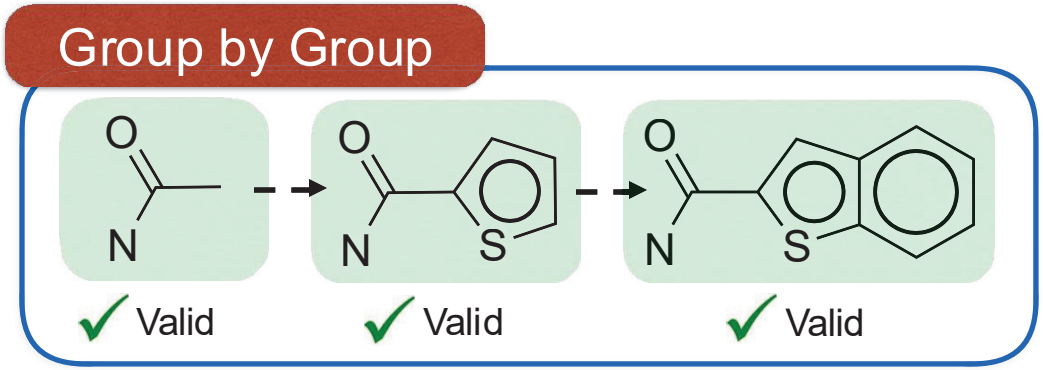


Functional Groups



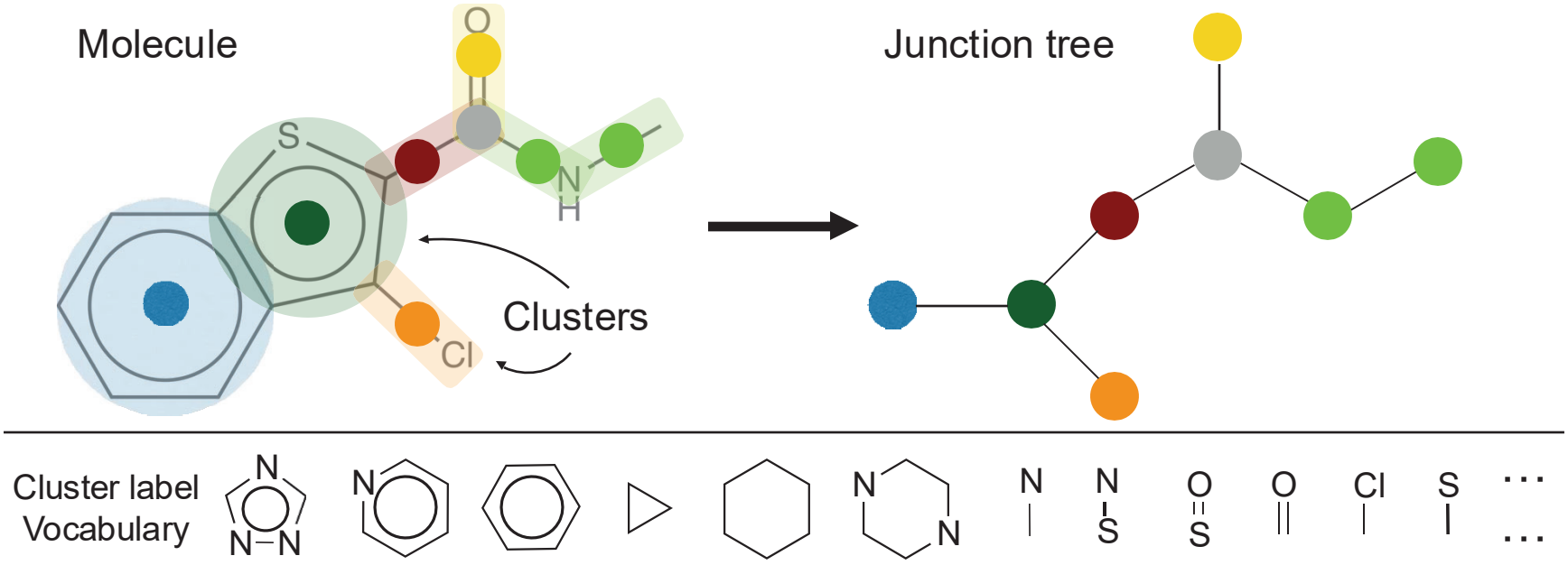
How to generate graphs?





- Shorter action sequence
- Easy to check validity

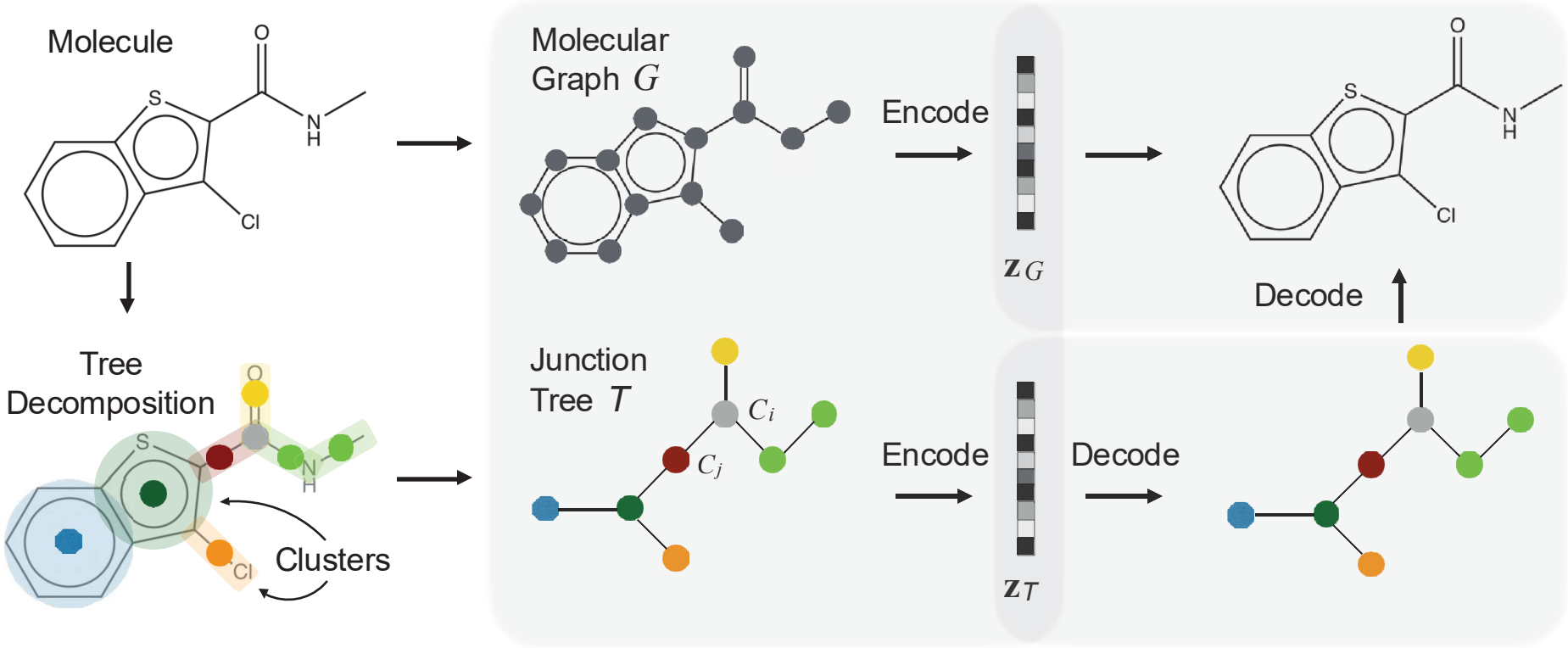
Tree decomposition



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

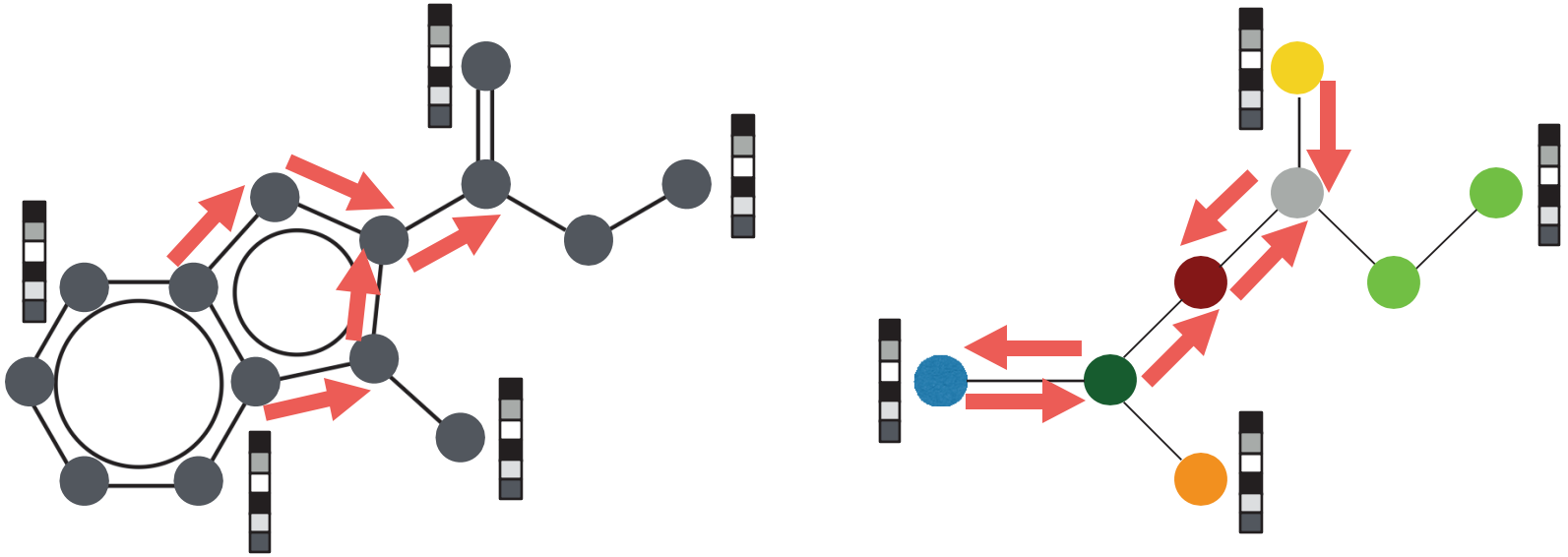
Jin et al., ICML 2018

Approach: Junction-tree variational autoencoder



Jin et al., ICML 2018

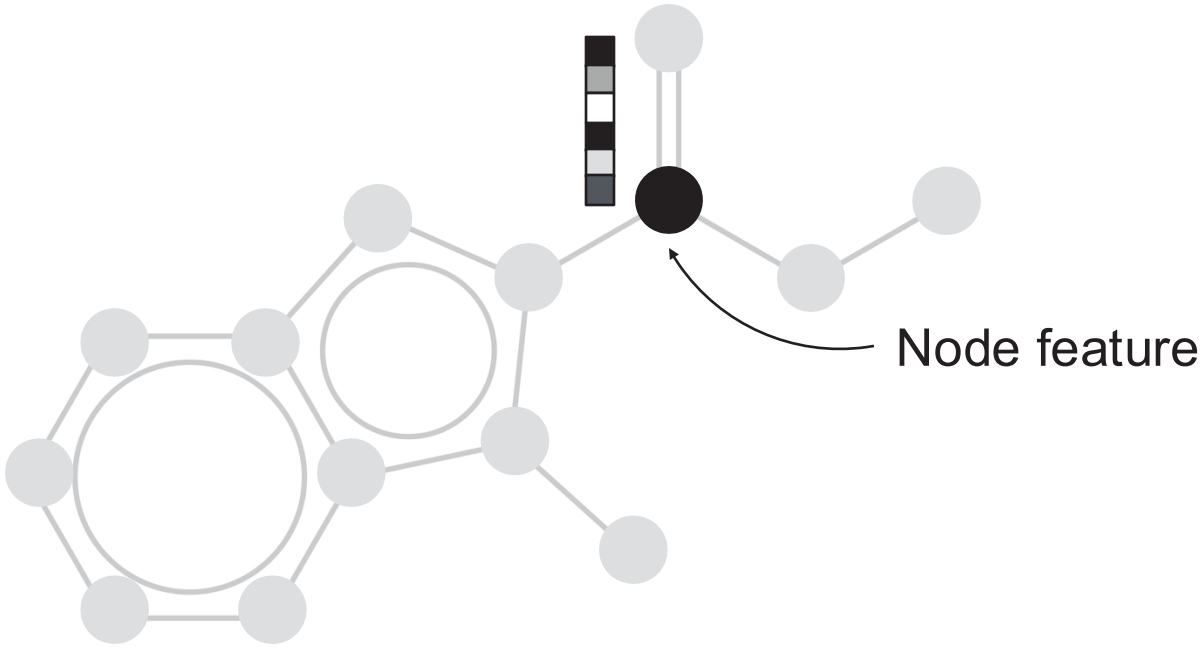
Graph and tree encoders



Neural Message Passing Network (MPN)

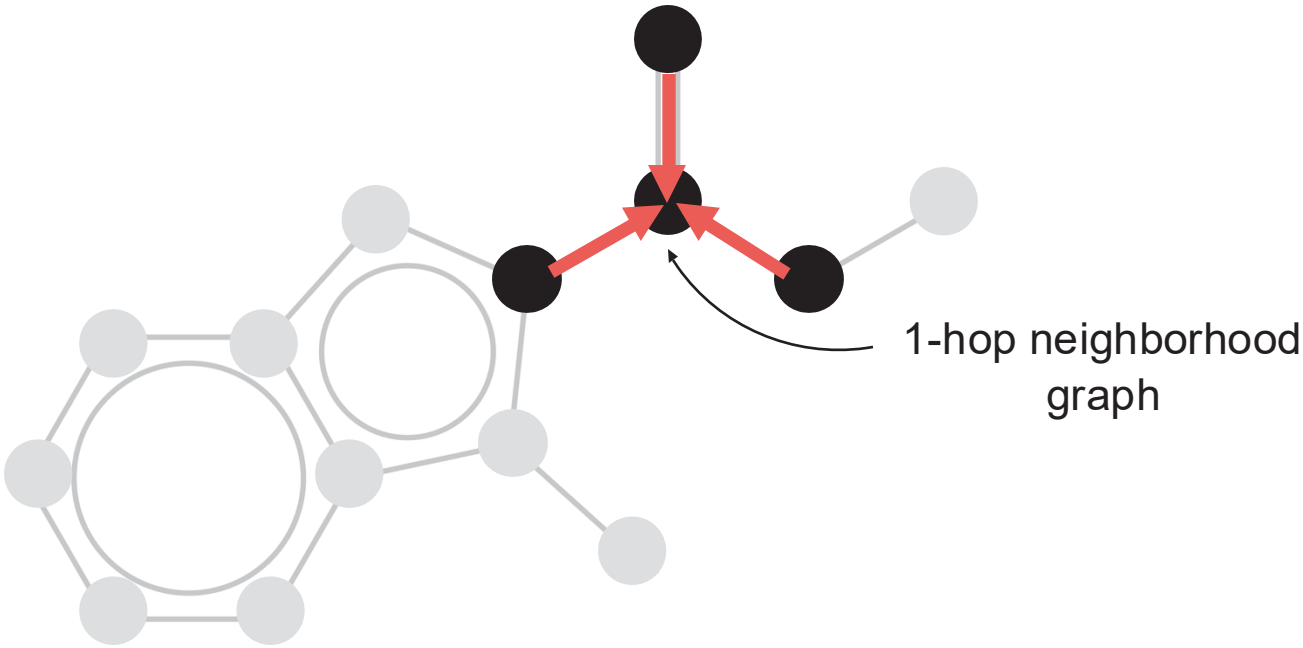
Jin et al., ICML 2018

Graph encoding



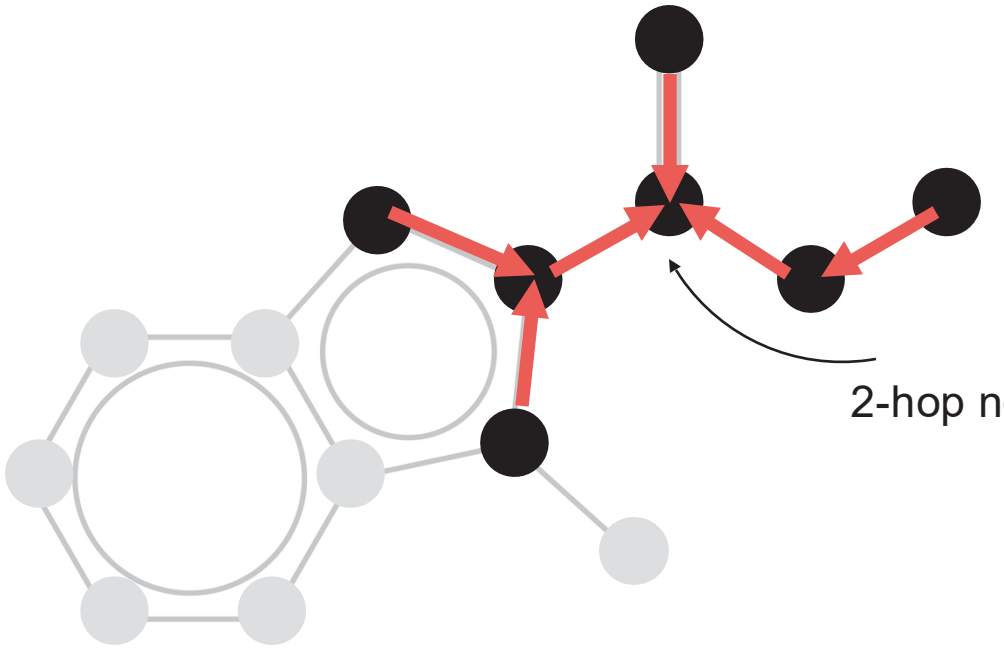
[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

Graph encoding



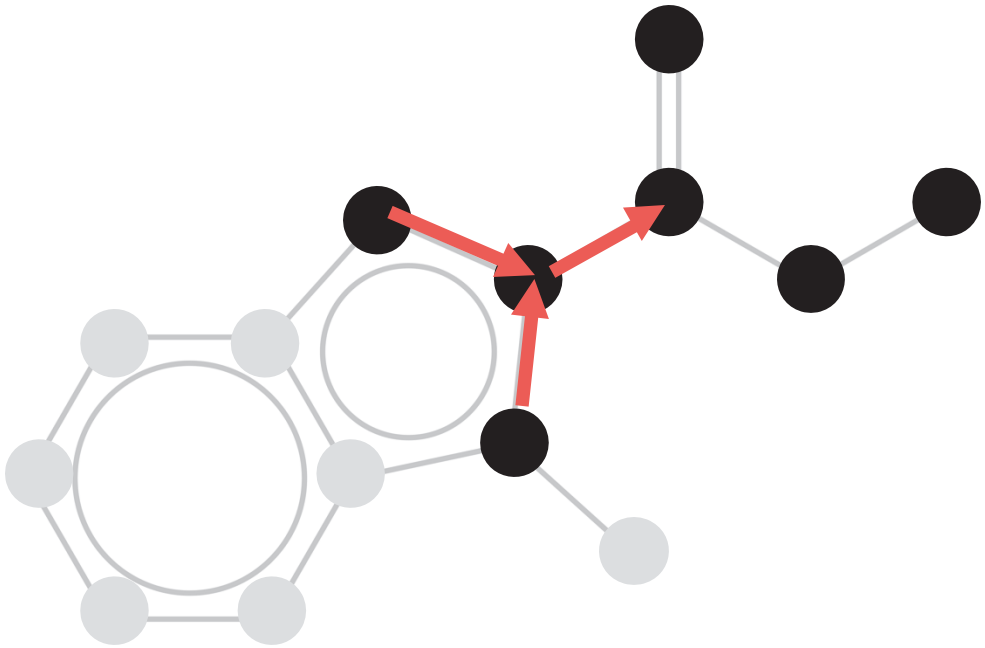
[3] Dai et al., Discriminative embeddings of latent variable models for structured data, 2016

Graph encoding



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

Graph encoding

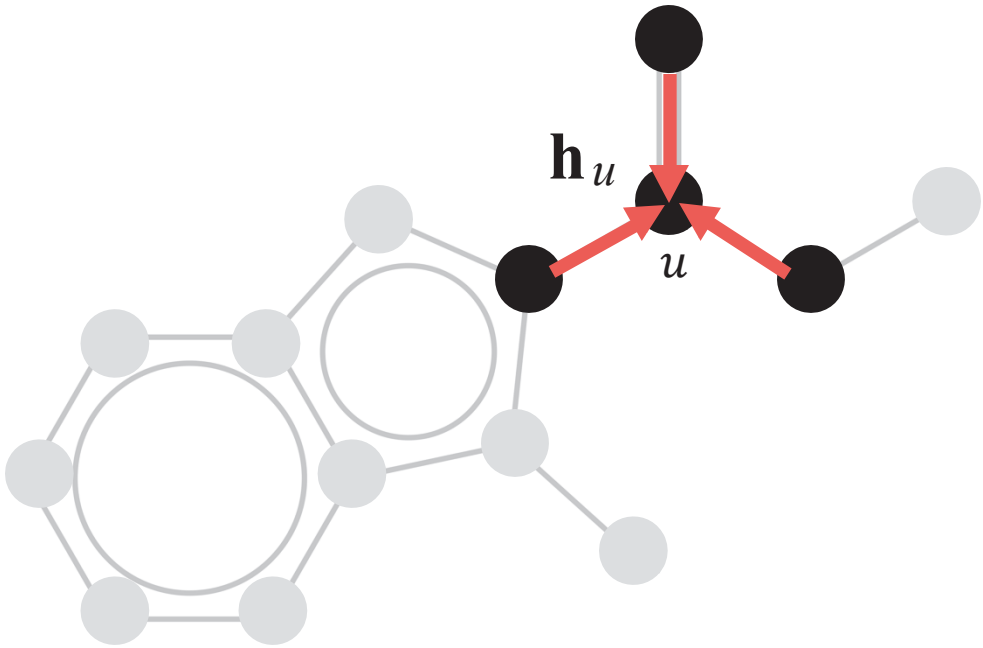


$$\boldsymbol{\nu}_{uv}^{(t)} = \tau(\mathbf{W}_1^g \mathbf{x}_u + \mathbf{W}_2^g \mathbf{x}_{uv} + \mathbf{W}_3^g \sum_{w \in N(u) \setminus v} \boldsymbol{\nu}_{wu}^{(t-1)})$$

Messages Node feature Edge feature

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

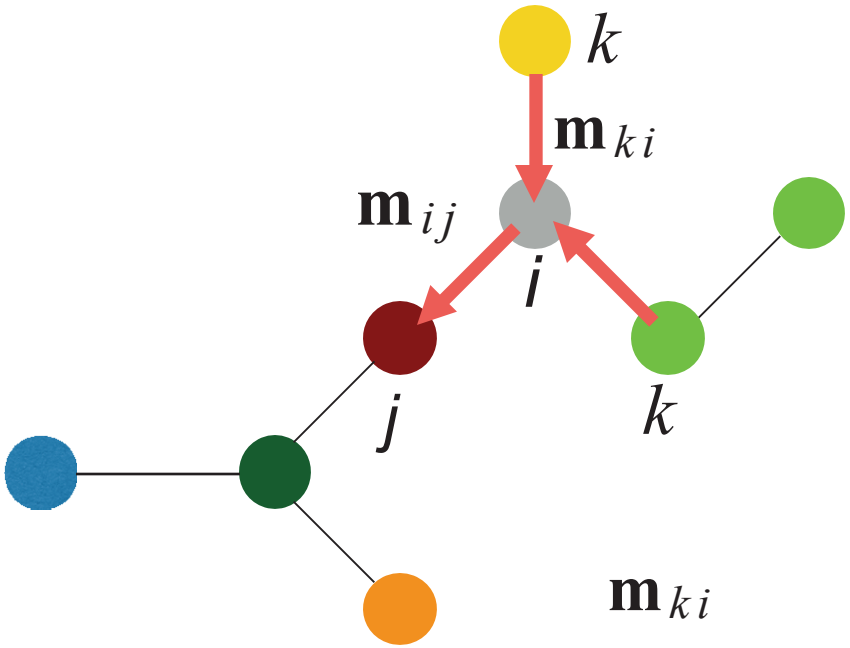
Graph encoding



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

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

Tree encoding

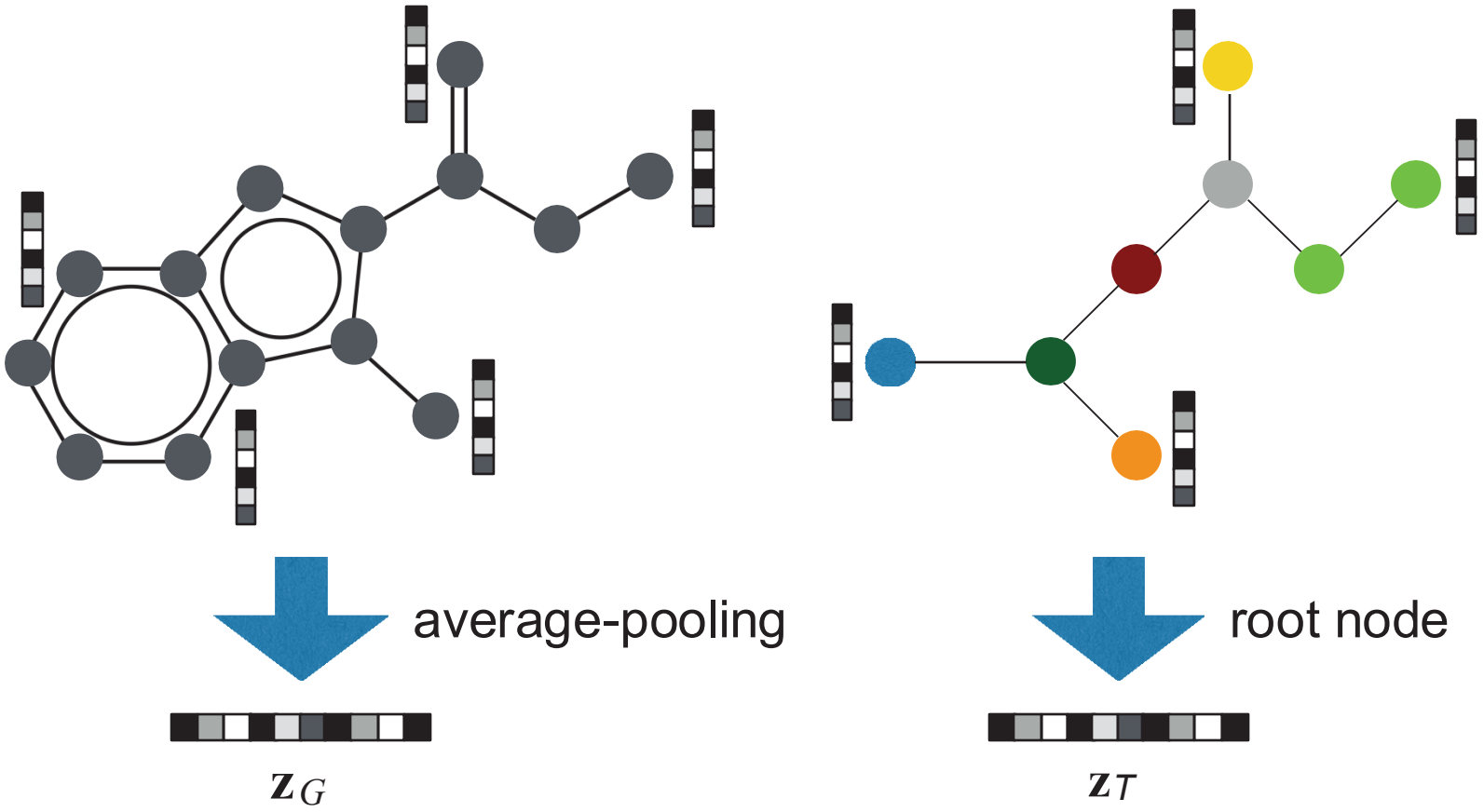


$$\mathbf{m}_{ij} = \mathrm{GRU}(\mathbf{x}_i, \{\mathbf{m}_{ki}\}_{k \in N(i) \setminus j})$$

To capture long range interactions

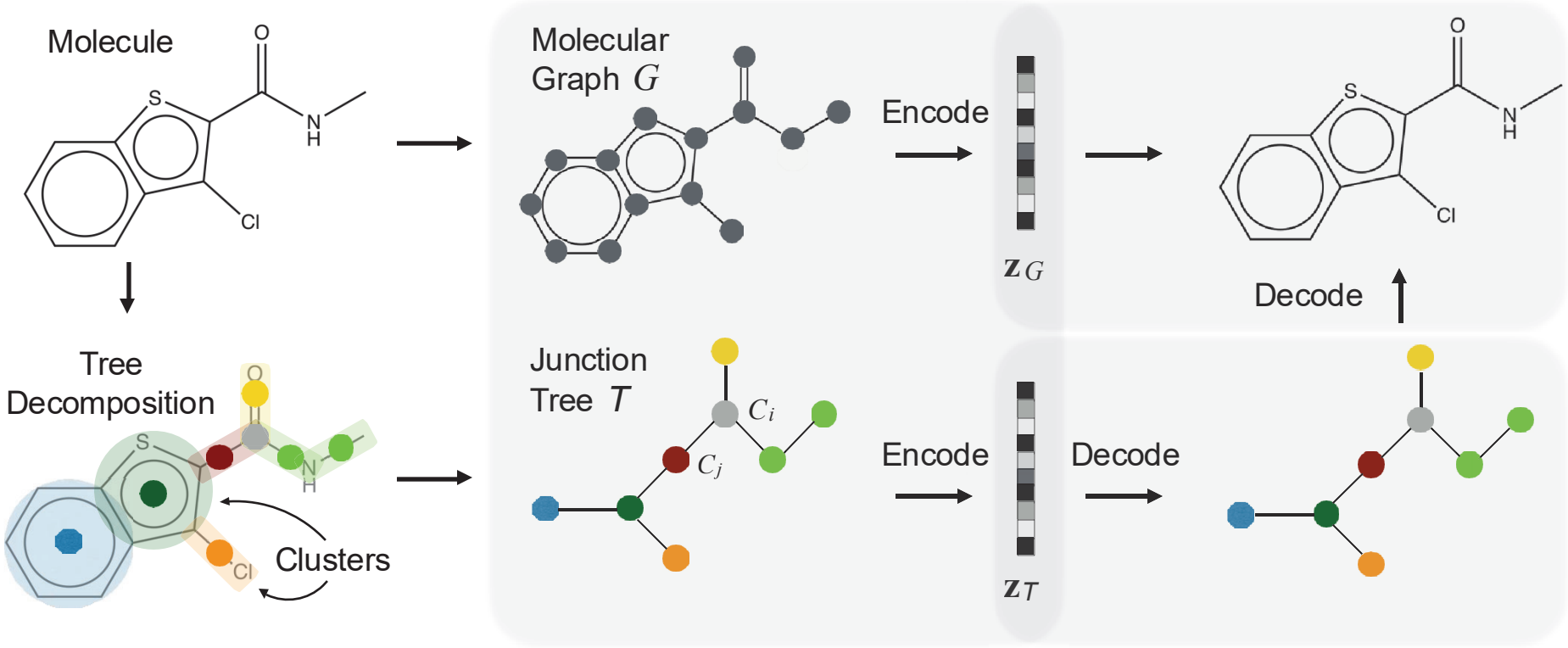
Jin et al., ICML 2018

Graph and tree encoders



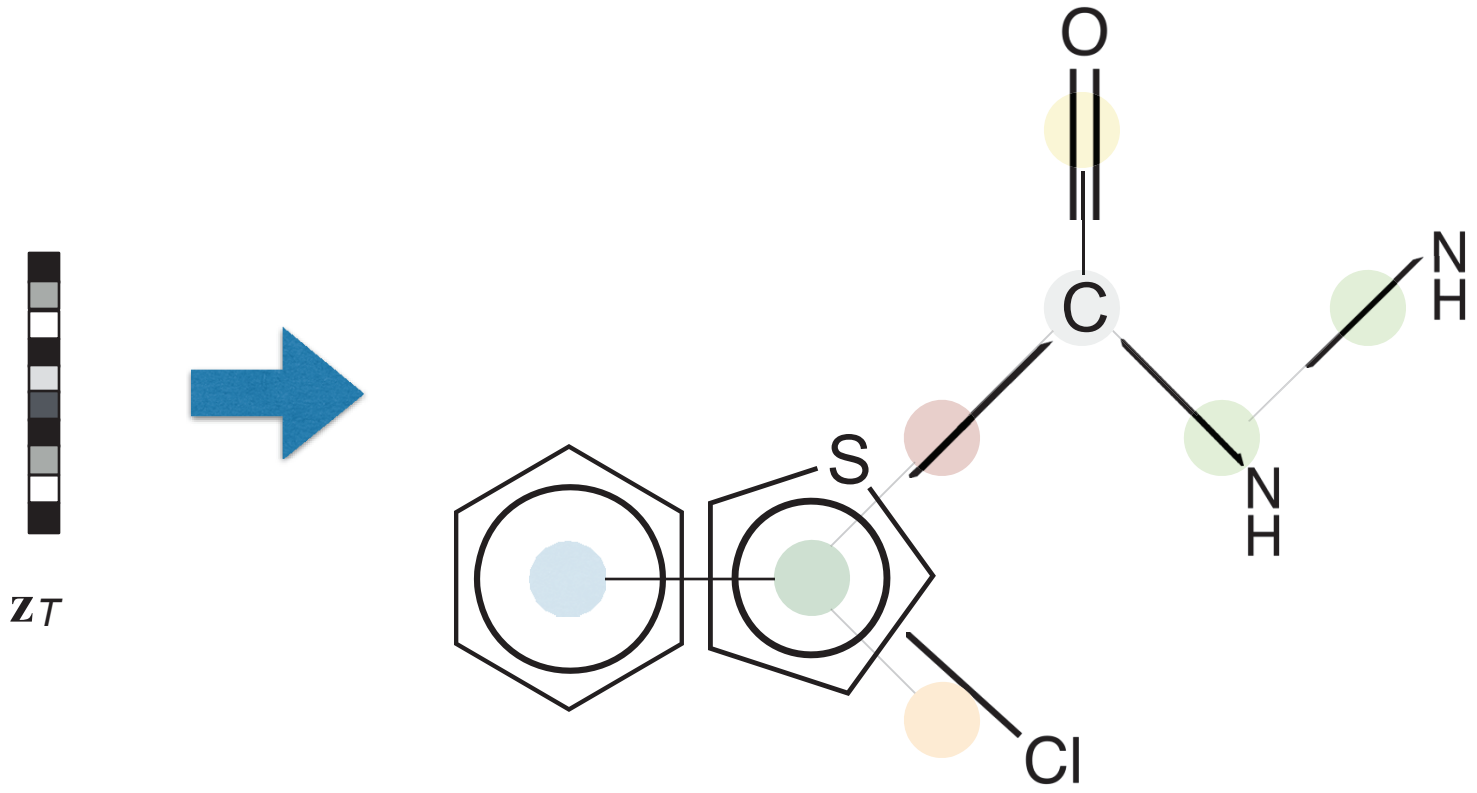
Jin et al., ICML 2018

Approach: Junction-tree variational autoencoder



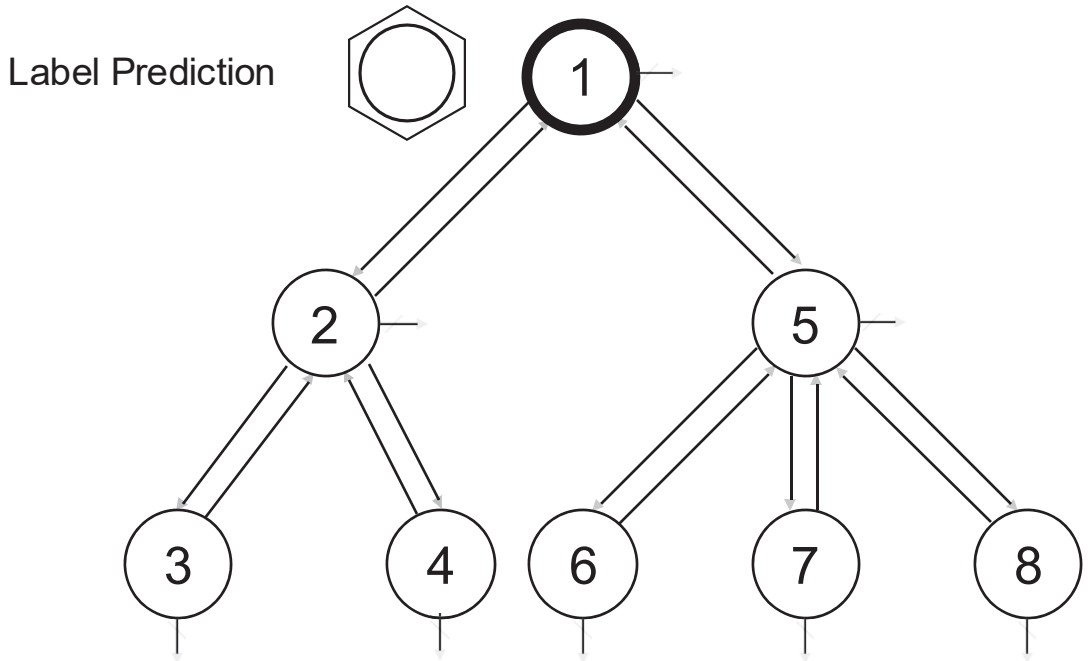
Jin et al., ICML 2018

Tree decoder



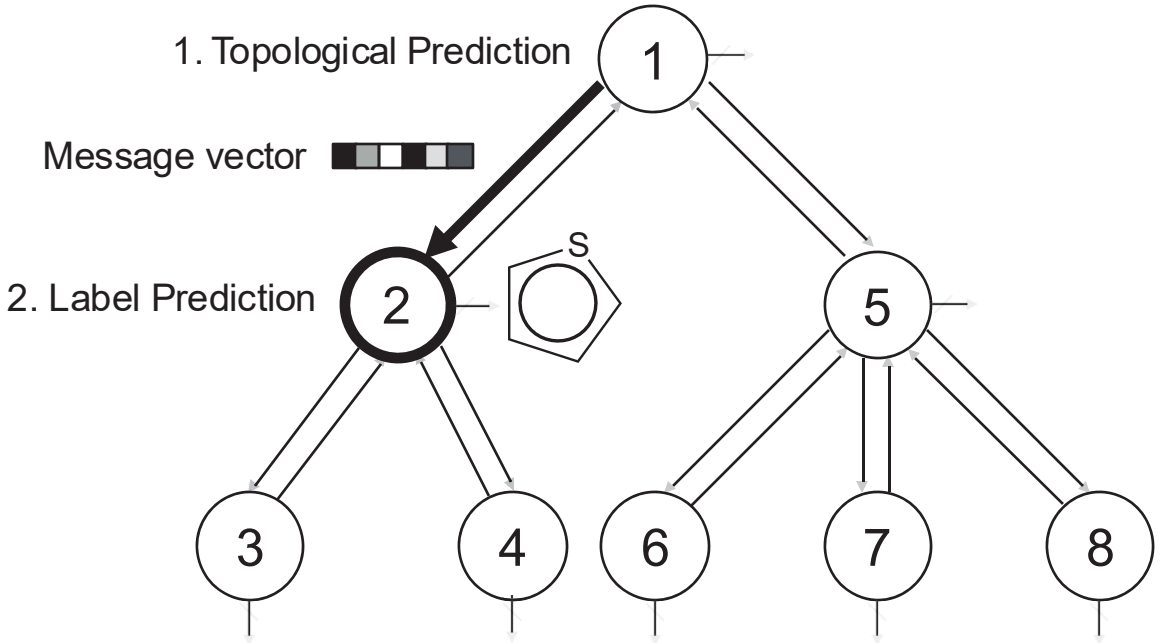
Jin et al., ICML 2018

Tree decoder



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

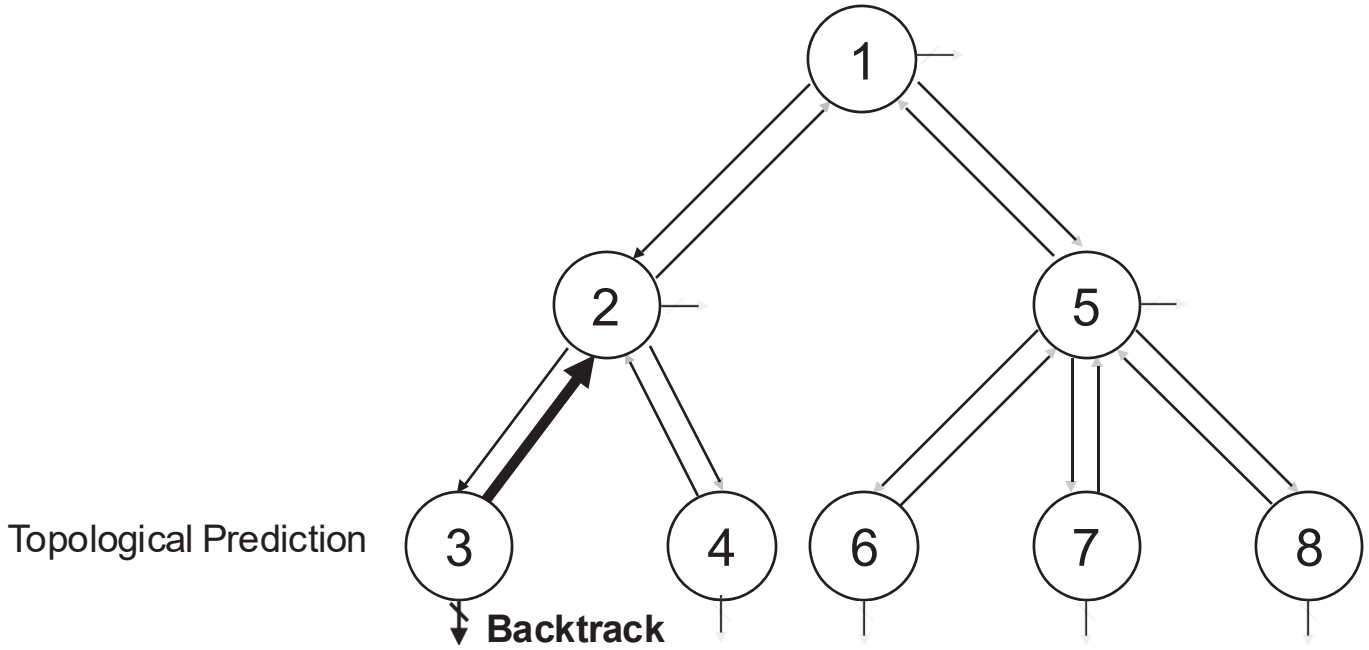
Tree decoder



Topological Prediction: Whether to expand a child or backtrack?

Label Prediction: What is the label of a node?

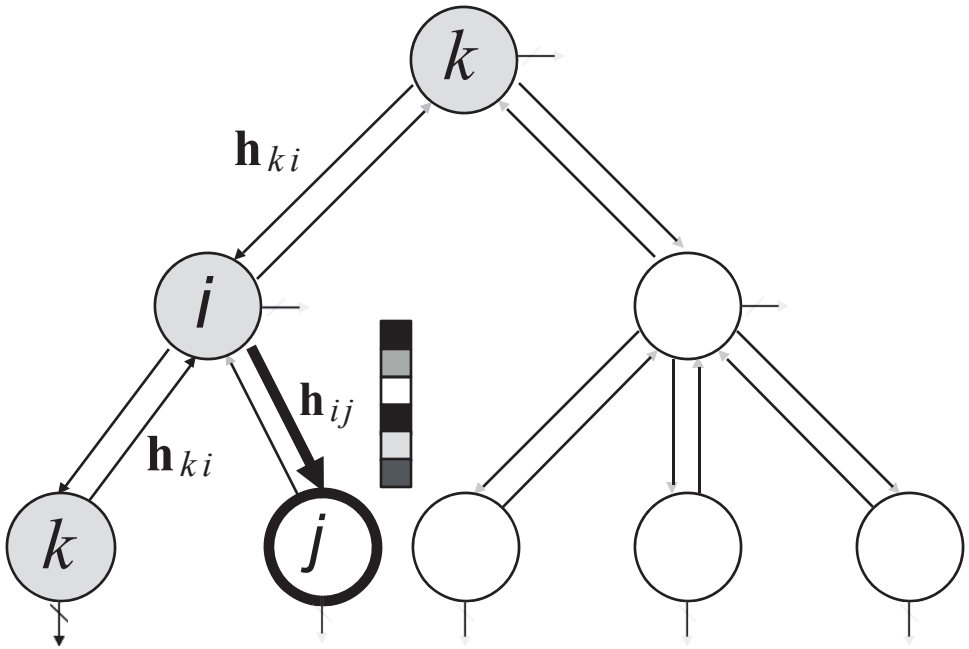
Tree decoder

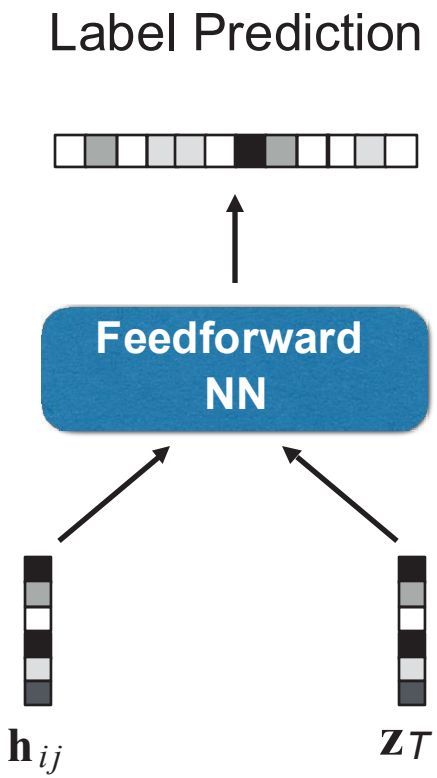


Topological Prediction: Whether to expand a node or backtrack?

Label Prediction: What is the label of a node?

Tree decoder

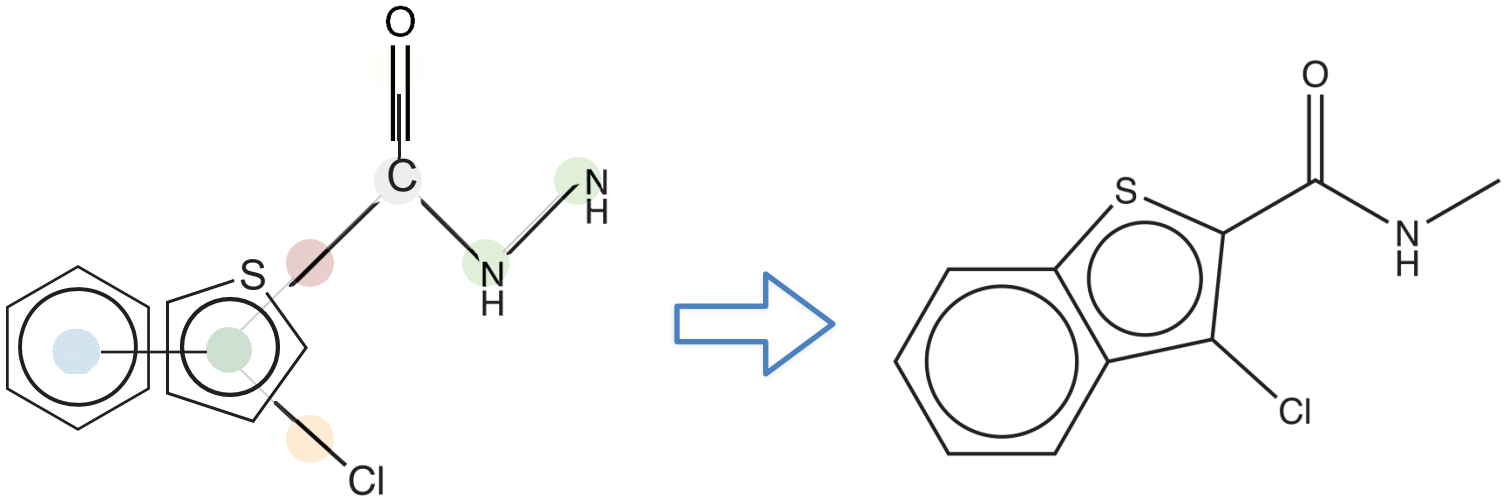




$$\mathbf{h}_{ij} = \mathrm{GRU}(\mathbf{x}_i, \{\mathbf{h}_{ki}\}_{k \in N_t(i) \setminus j})$$

Encodes the entire subtree of current state

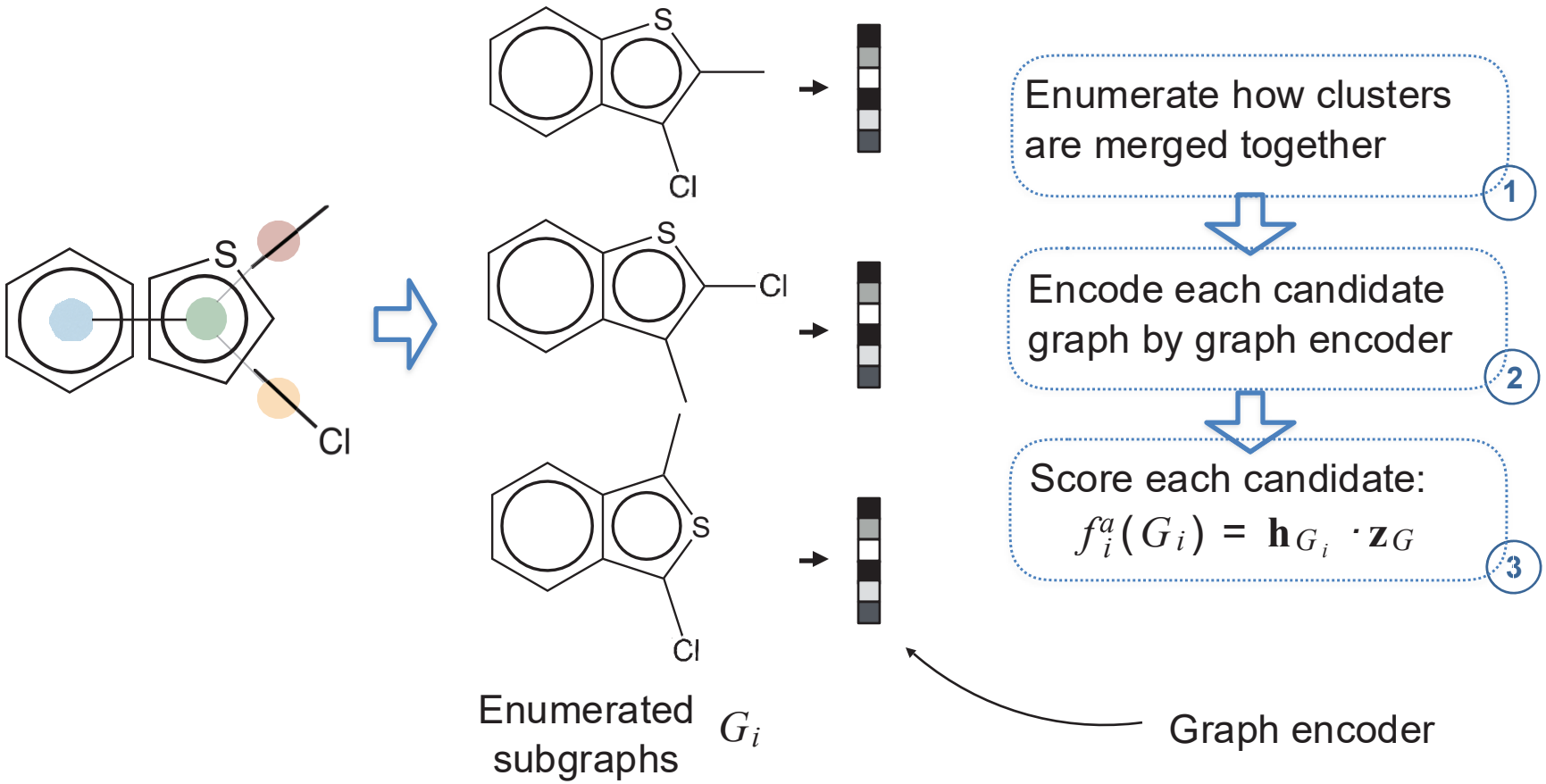
Graph decoder



Predicted Junction Tree

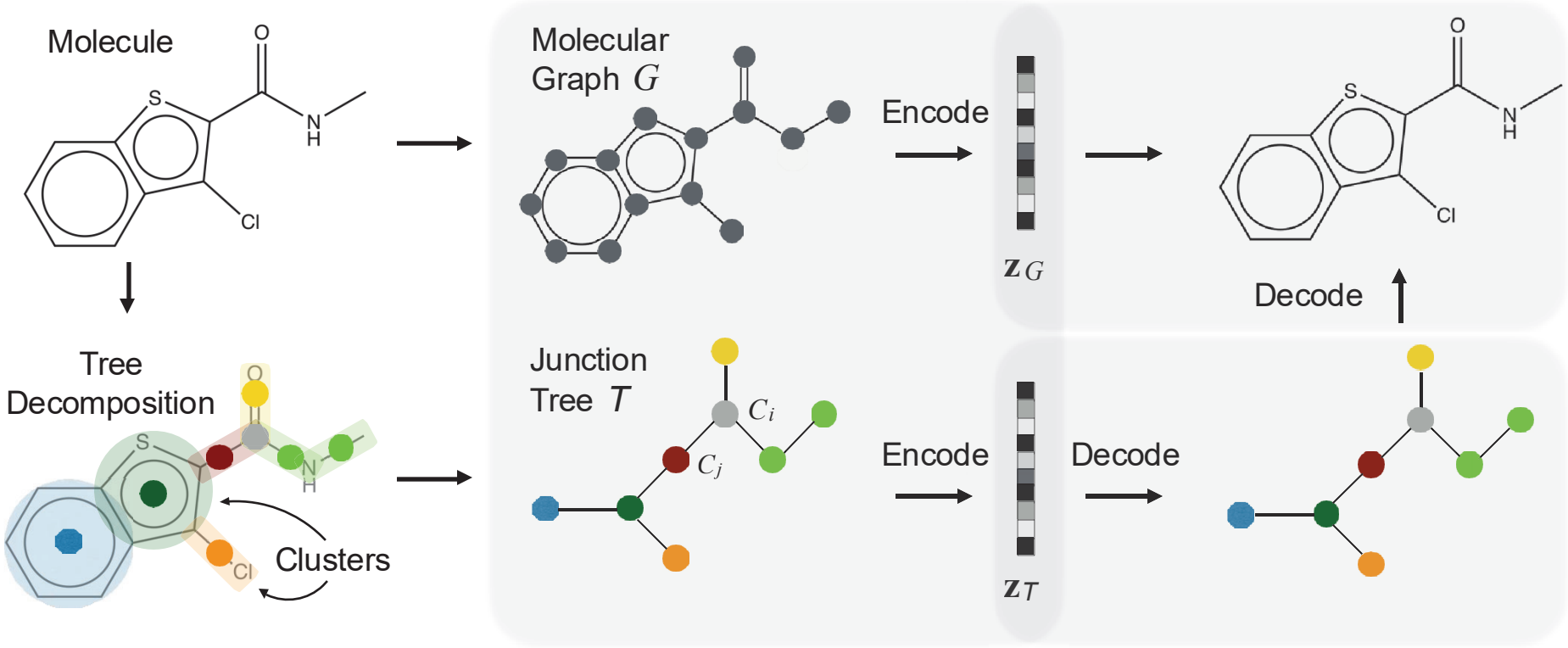
Molecular Graph

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:

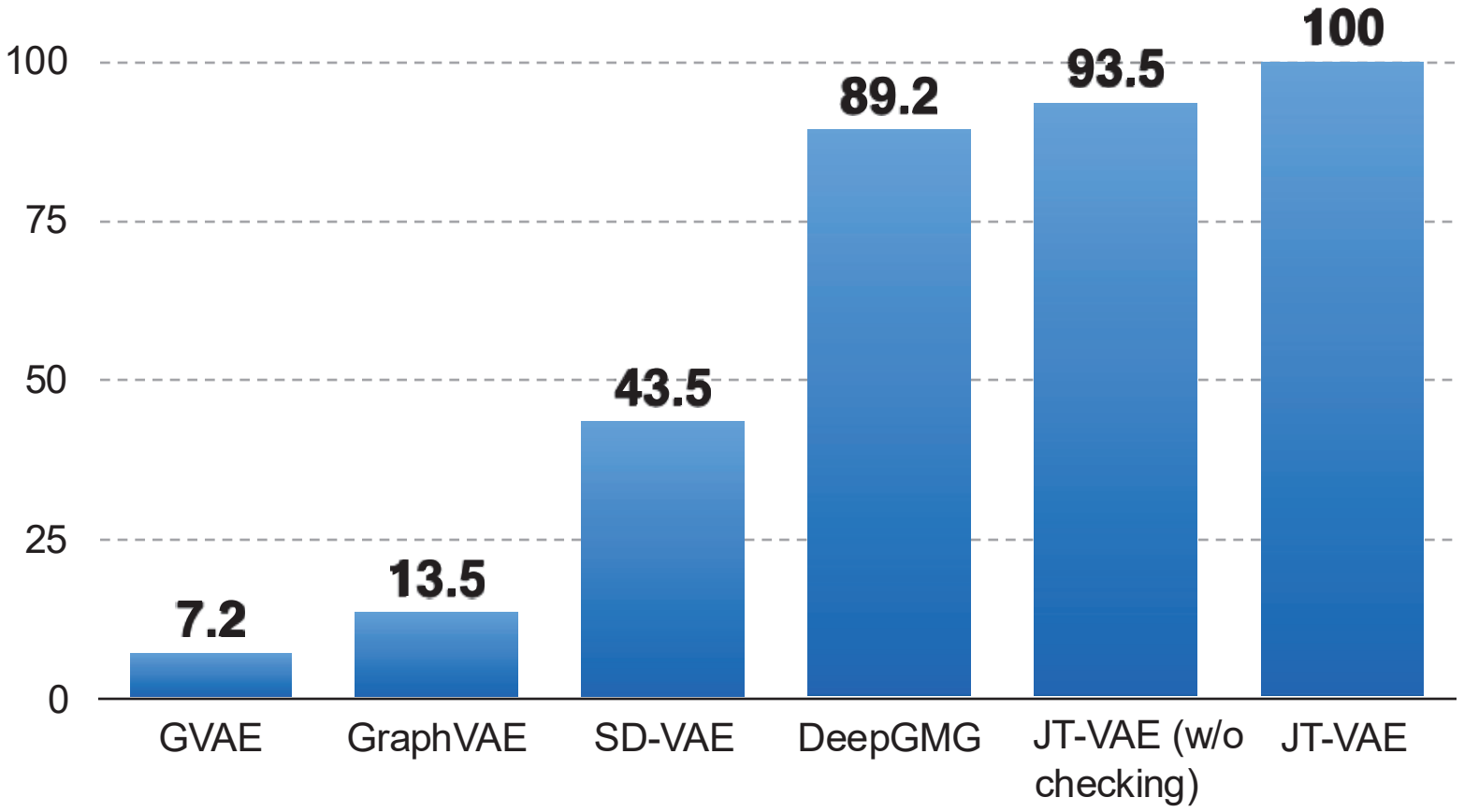
1. Graph VAE (Simonovsky & Komodakis, 2018)

2. DeepGMG (Li et al., 2018)

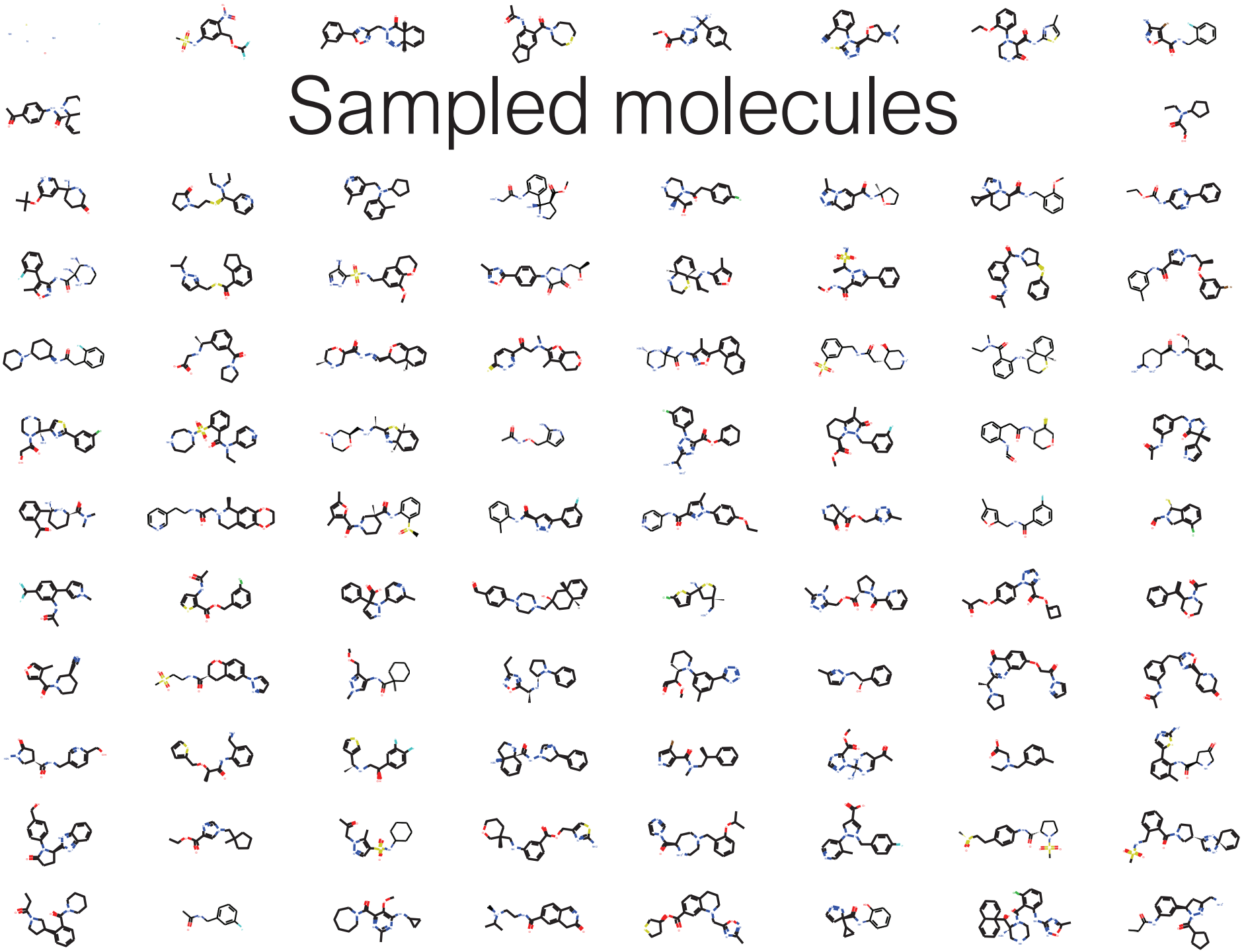
---

[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

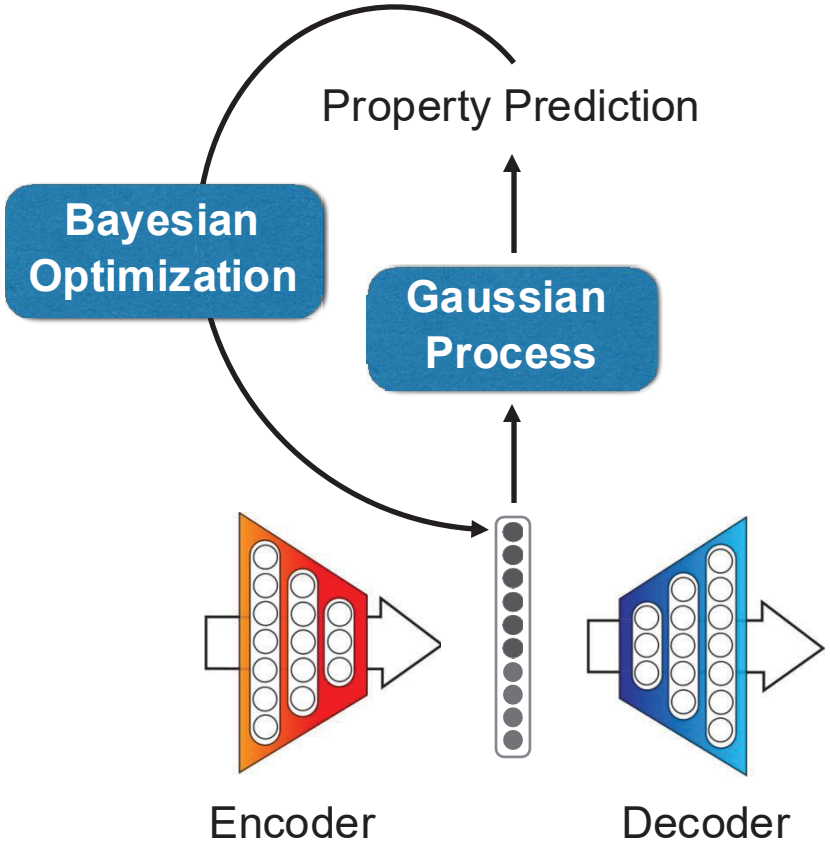
Molecule generation (Validity)

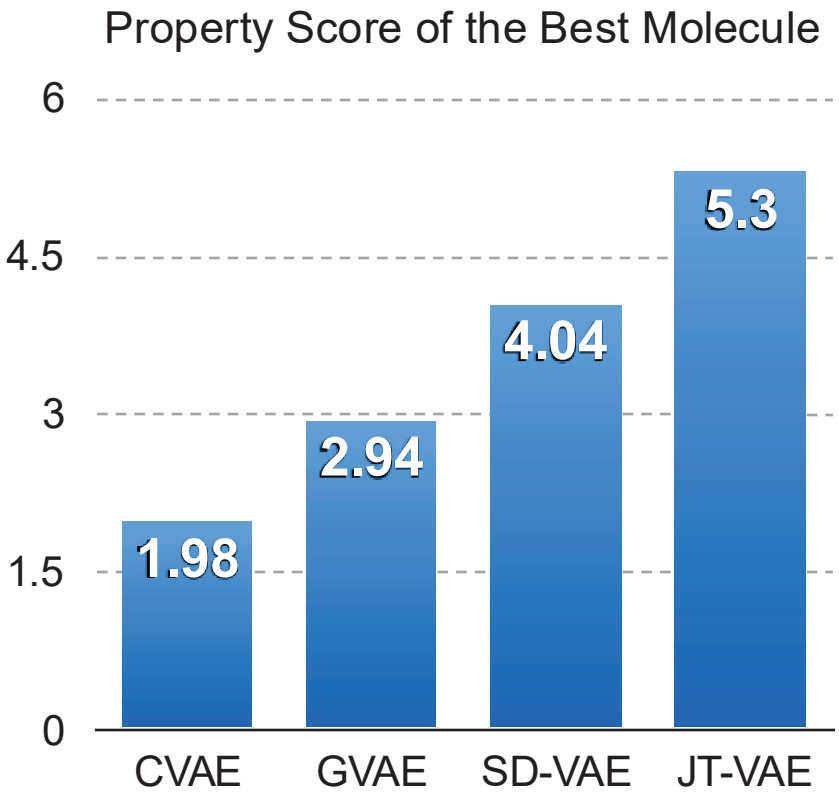


Sampled molecules



Molecule optimization (Global)

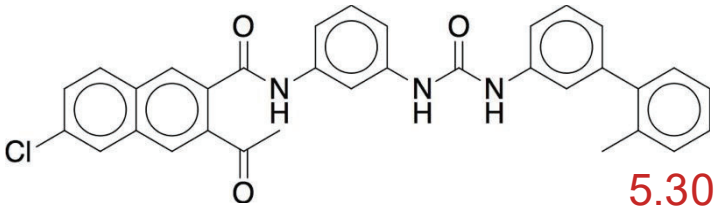




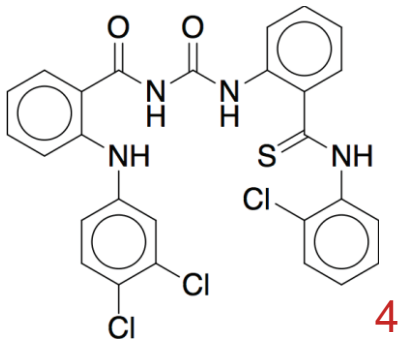
Property: Solubility + Ease of Synthesis

Jin et al., ICML 2018

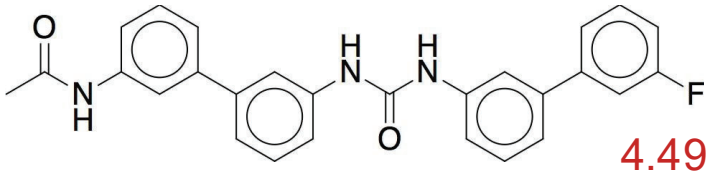
Molecule optimization (Global)



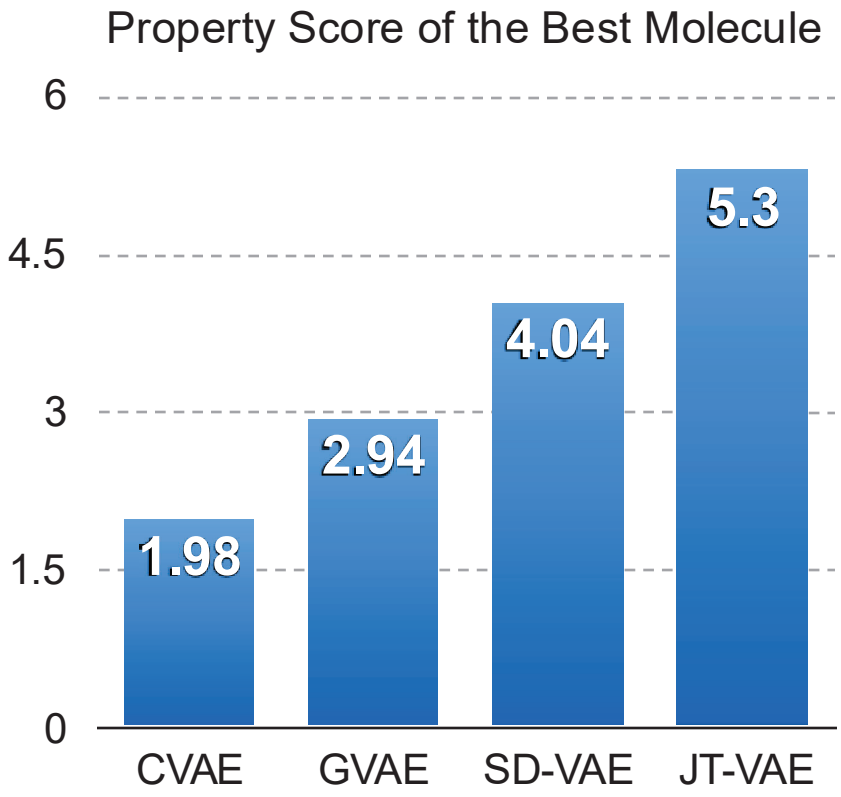
5.30



4.93



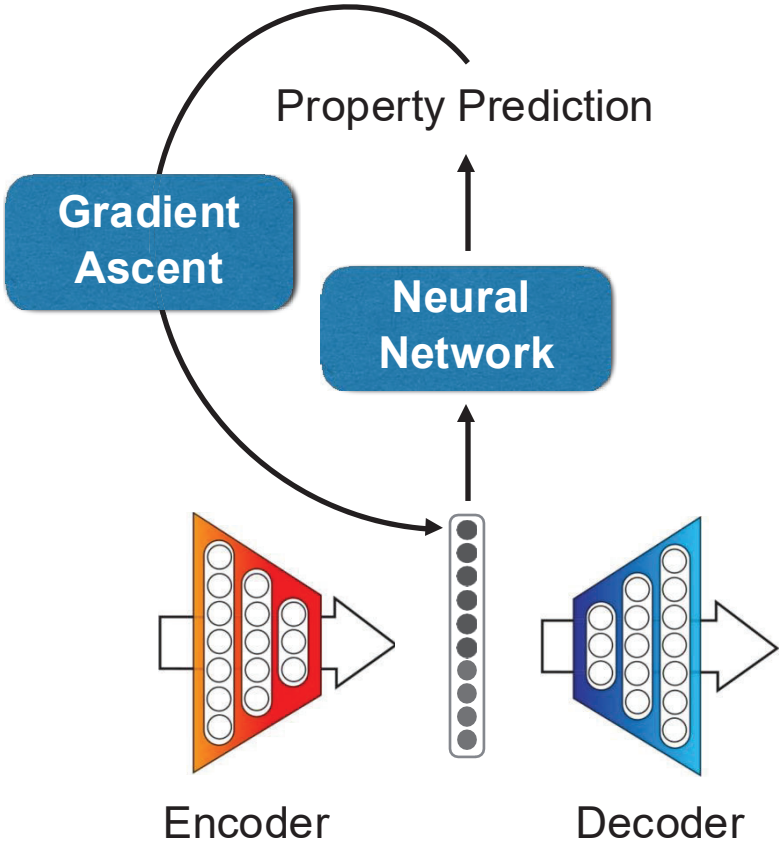
4.49

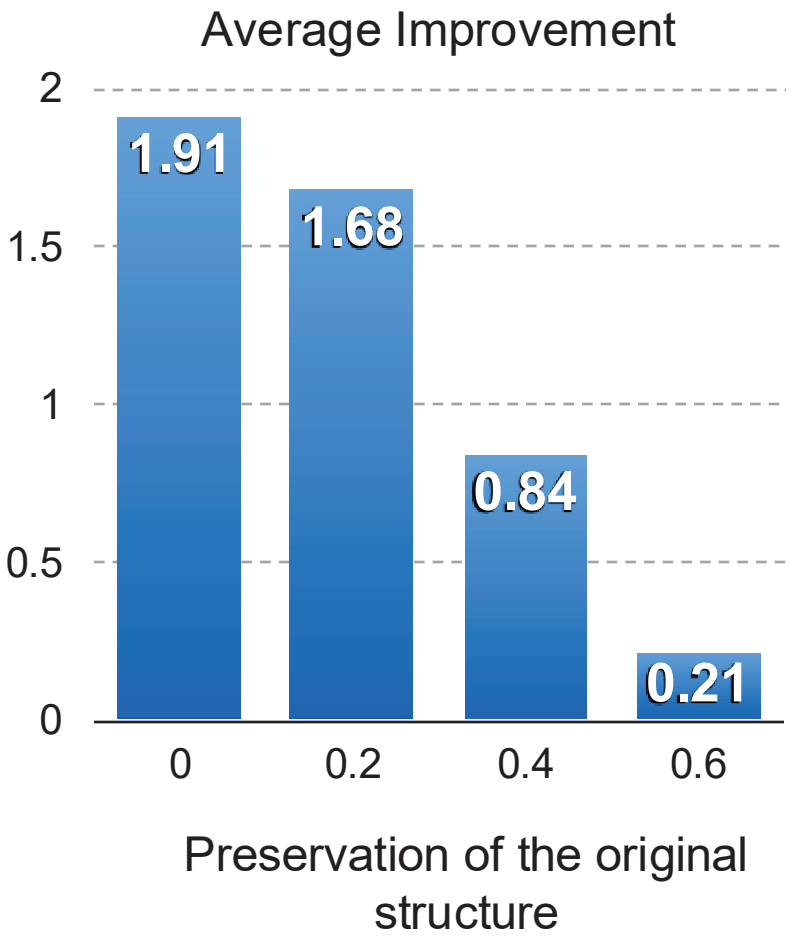


Property: Solubility + Ease of Synthesis

Jin et al., ICML 2018

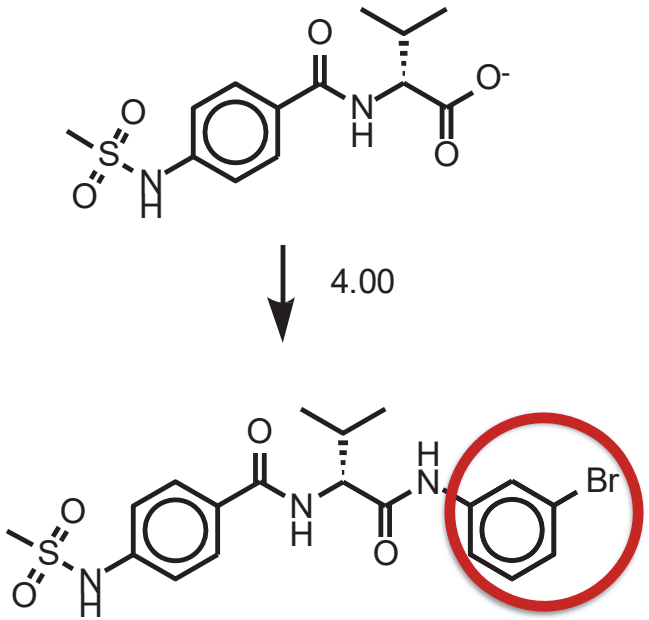
Molecule optimization (Local)



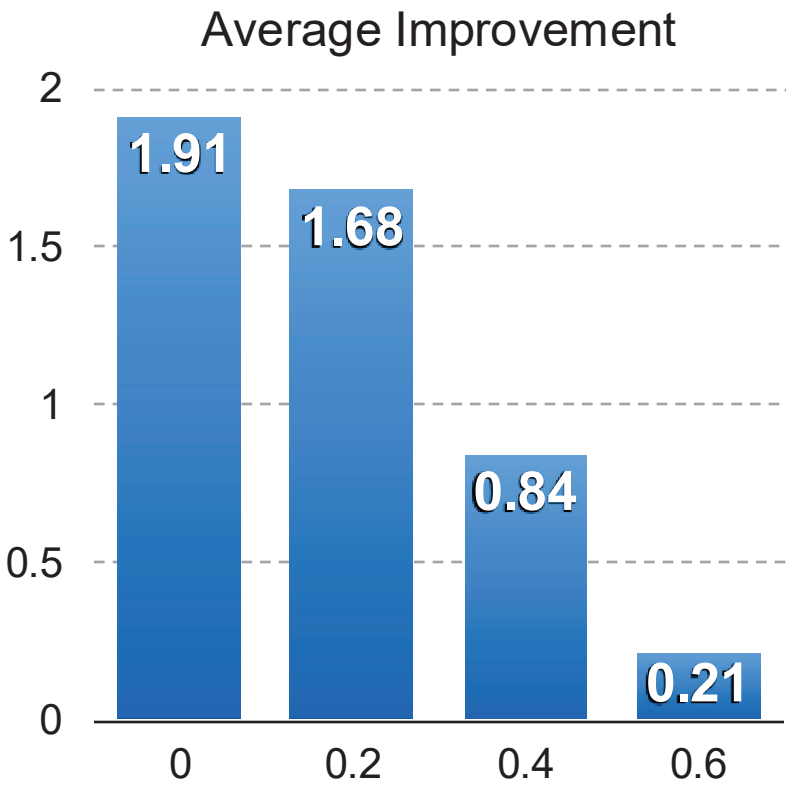


Jin et al., ICML 2018

Molecule optimization (Local)

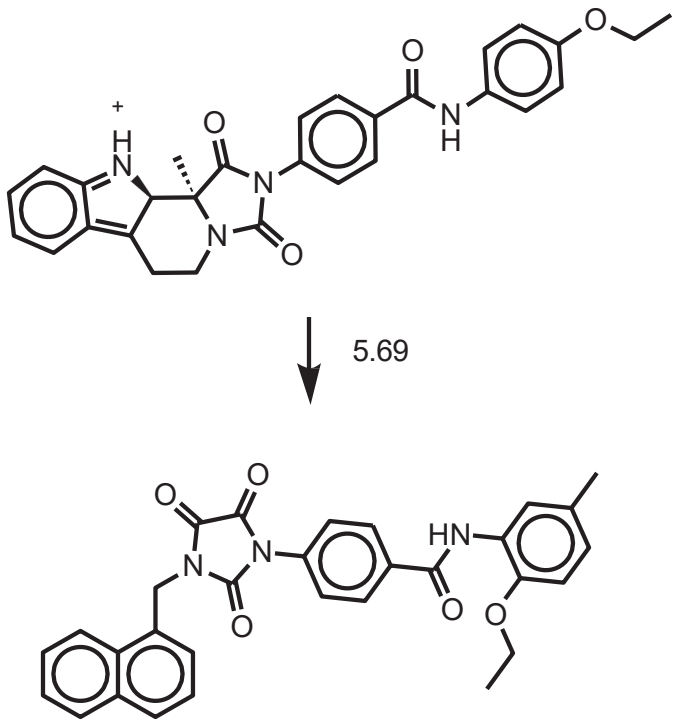


Preservation ≈ 0.6

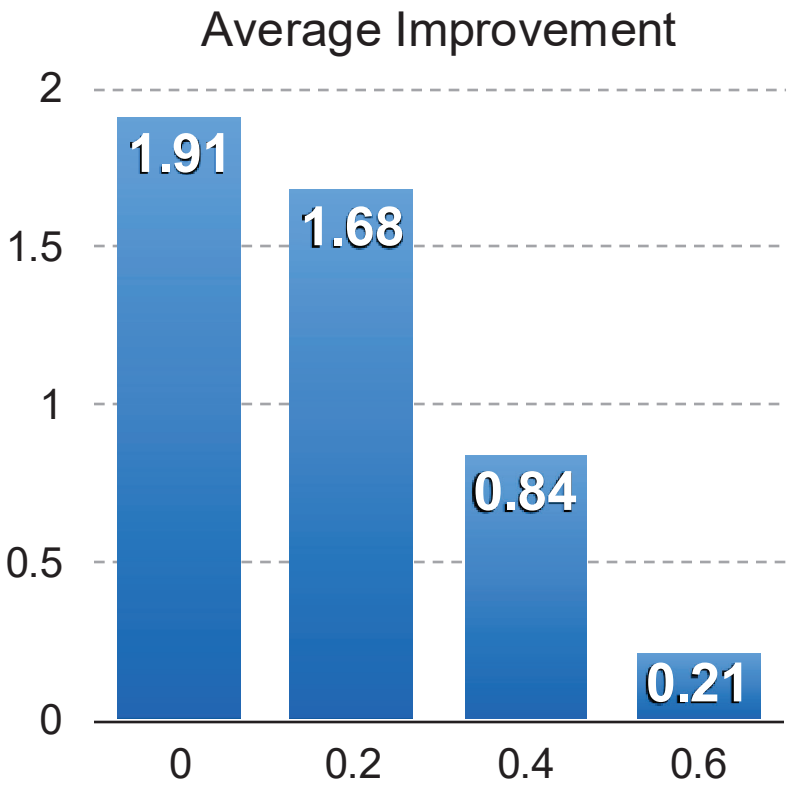


Jin et al., ICML 2018

Molecule optimization (Local)

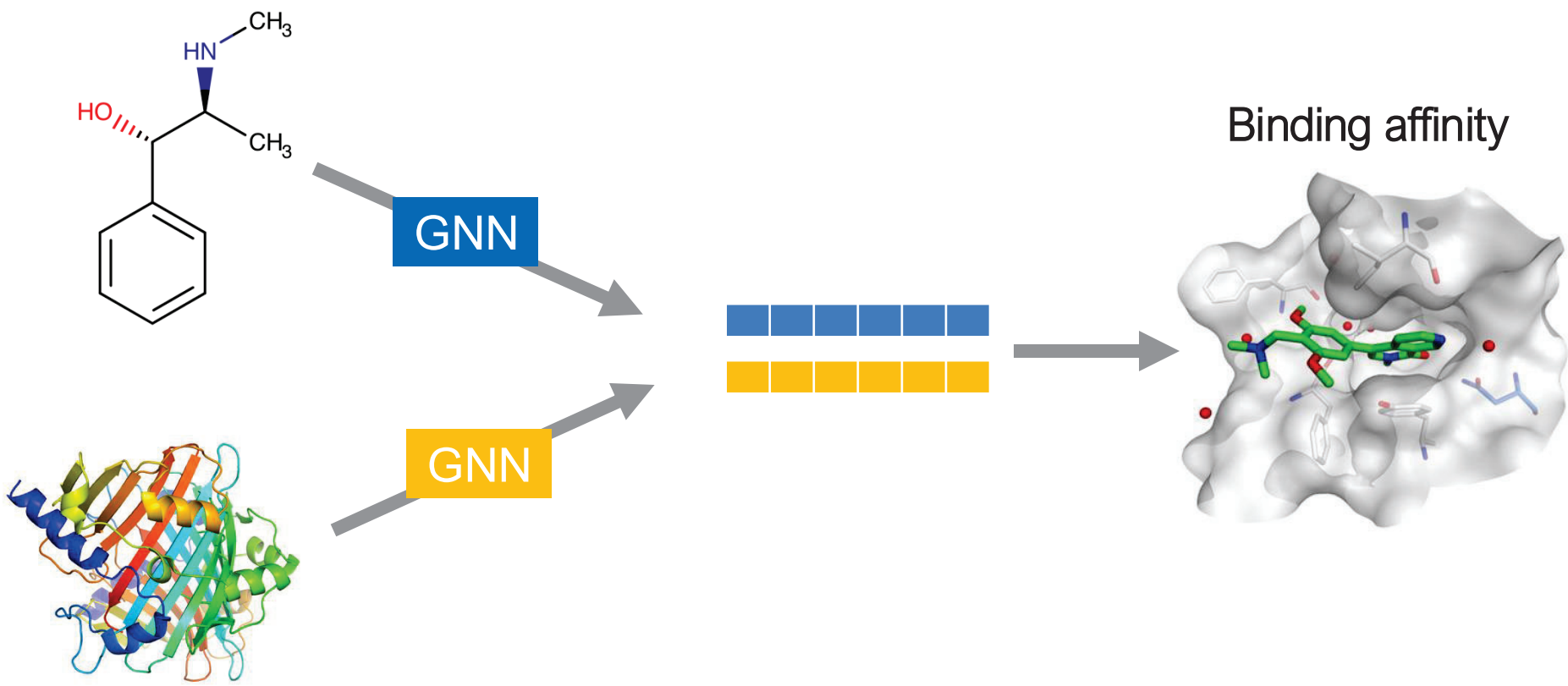


Preservation ≈ 0.4



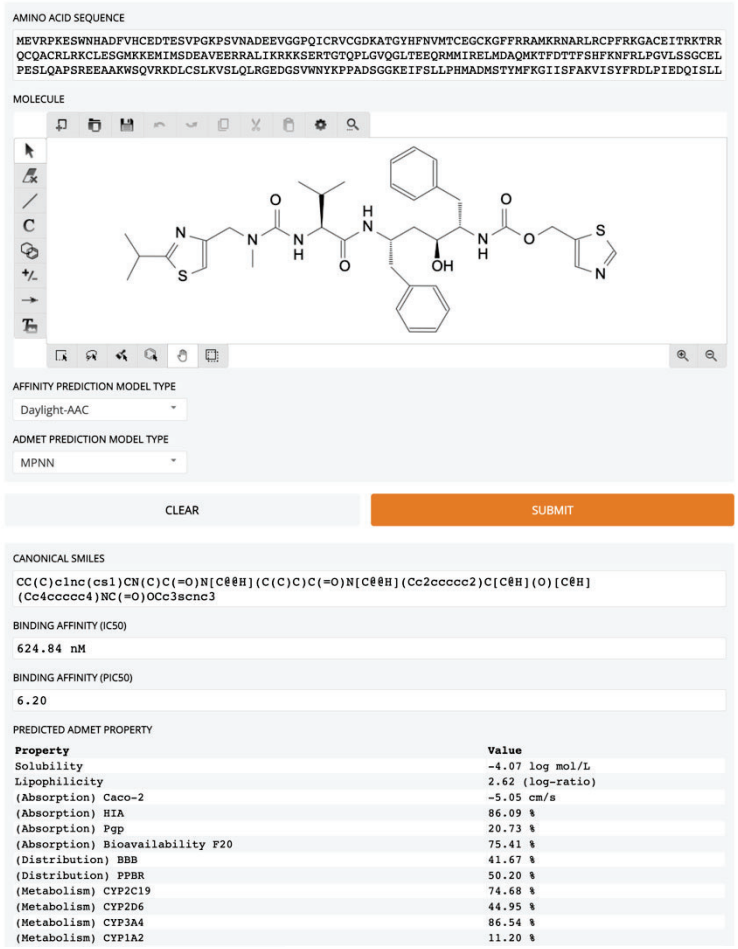
Jin et al., ICML 2018

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