Graph Representation Learning for Drug Discovery

Jian Tang

Mila-Quebec AI Institute

HEC Montreal

www.jian-tang.com

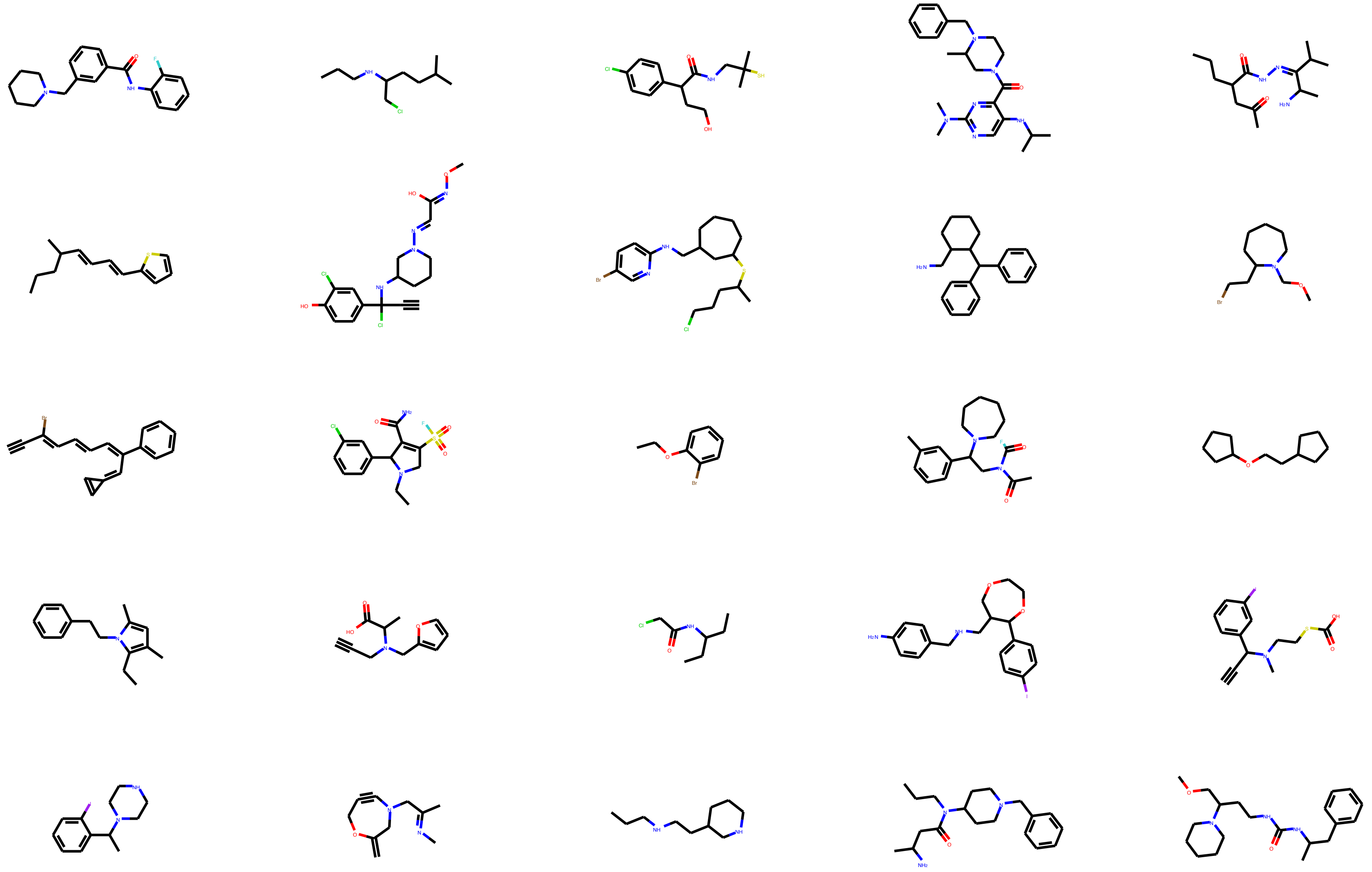
The Process of Drug Discovery

- A very long and costly process
  - On average takes more than 10 years and $2.5B to get a drug approved
- Big opportunities for AI to accelerate this process

Target → Lead Discovery 2 years → Lead Optimization 3 years → Preclinical Study 2 years → Clinical Trial

Lead Discovery	Lead Optimization	Preclinical Study	Clinical Trial
Screen millions of functional molecules; Found by serendipity: Penicillin	Modify the molecule to improve specific properties. e.g. toxicity, SA	In-vitro and in-vivo experiments; synthesis	Multiple Phases

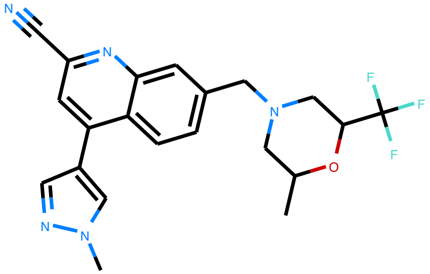
Molecules



Research Problems

Target → Lead Discovery 2 years → Lead Optimization 3 years → Preclinical Study 2 years → Clinical Trial

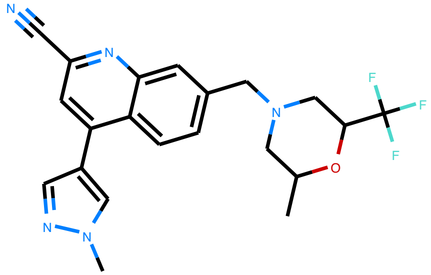
Property Prediction



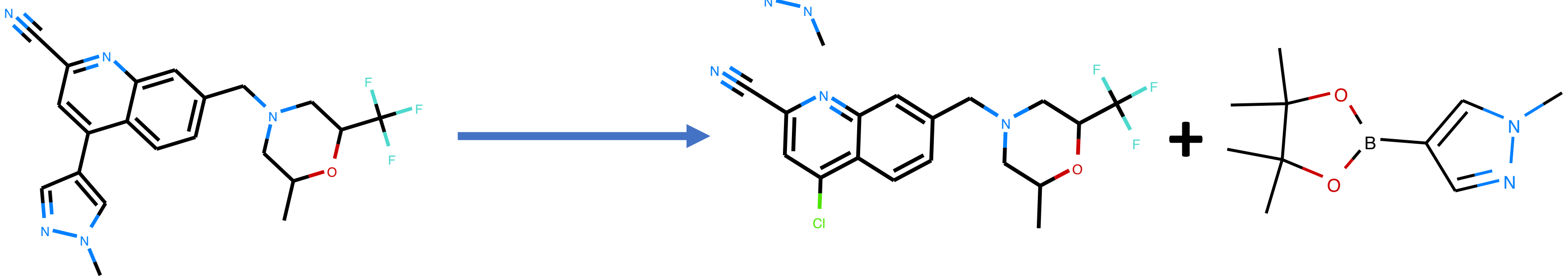
→ Property

Molecule Design and Optimization

Property →

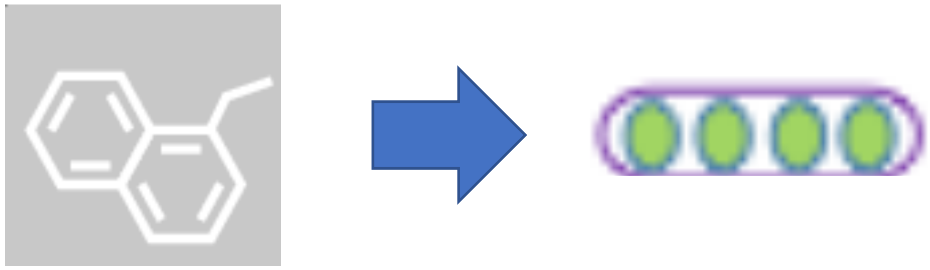


Retrosynthesis Prediction



Molecule Properties Prediction

- Predicting the properties of molecules or compounds is a fundamental problem in drug discovery
- Each molecule is represented as a graph
- The fundamental problem: how to represent **a whole molecule (graph)**



Graph Neural Networks

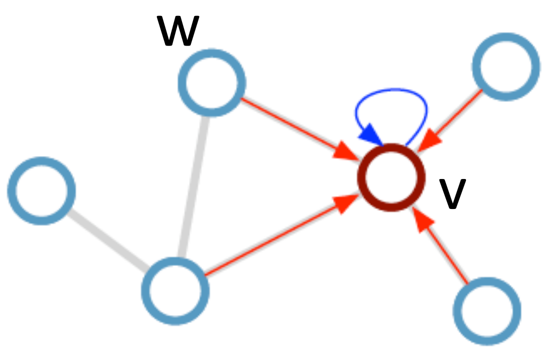
- Techniques for learning node/graph representations
  - Graph convolutional Networks (Kipf et al. 2016)
  - Graph attention networks (Veličković et al. 2017)
- Neural Message Passing (Gilmer et al. 2017)

MESSAGE PASSING: $M_k(h_v^k, h_w^k, e_{vw})$

AGGREGATE : $m_v^{k+1} = \mathrm{AGGREGATE}\{M_k\left(h_v^k, h_w^k, e_{vw}\right): w \in N(v)\}$

COMBINE : $h_v^{k+1} = \mathrm{COMBINE}(h_v^k, m_v^{k+1})$

READOUT: $g = \mathrm{READOUT}\{h_v^K: v \in G\}$



InfoGraph: Unsupervised and Semi-supervised Whole-Graph Representation Learning (Sun et al. ICLR'20)

- For supervised methods based on graph neural networks, a large number of labeled data are required for training
- In the domain of drug discovery, the number of labeled data are limited
  - A large amount of unlabeled data (molecules) are available
- This work: how to effectively learn whole graph representations in unsupervised or semi-supervised fashion

Fanyun Sun, Jordan Hoffman, Vikas Verma and Jian Tang. **InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization**. ICLR'20.

InfoGraph: Unsupervised Whole-Graph Representation Learning (Sun et al. ICLR'20)

- Maximizing the *mutual information* between the whole graph representation $H_\phi(G)$ and all the sub-structure representation h_ϕ^i .
  - Ensure the graph representation capture the predominant information among all the substructures
- K-layer graph neural networks:

$$h_v^{(k)} = \text{COMBINE}^{(k)}\left(h_v^{(k-1)}, \text{AGGREGATE}^{(k)}\left(\left\{\left(h_v^{(k-1)}, h_u^{(k-1)}, e_{uv}\right) : u \in \mathcal{N}(v)\right\}\right)\right)$$

- Summarize the local structure information at every node *i*:

$$h_\phi^i = \text{CONCAT}(\{h_i^{(k)}\}_{k=1}^K)$$

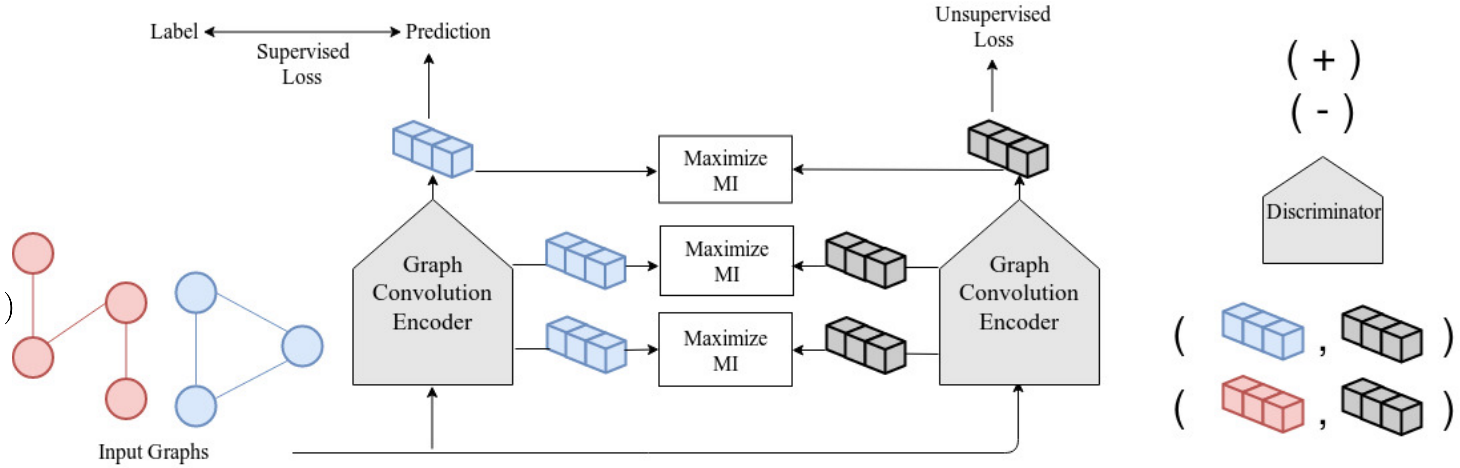
- Summarize the information of the whole graph:

$$H_\phi(G) = \text{READOUT}(\{h_\phi^i\}_{i=1}^N)$$

InfoGraph*: Semi-supervised Graph Representation Learning (Sun et al. ICLR'20)

- Two different encoders for the supervised and unsupervised tasks
- Maximize the *mutual information* of the representations learned by the two encoders at all levels (or layers)

$$
\begin{aligned}
\mathrm{L}_{\text{total}} = \sum_{\mathrm{i}=1}^{|\mathbb{G}^{\mathrm{L}}|} \mathrm{L}_{\text{supervised}}(\mathrm{y}_{\phi}(\mathrm{G}_{\mathrm{i}}), \mathrm{o}_{\mathrm{i}}) &+ \sum_{j=1}^{|\mathbb{G}^{L}|+|\mathbb{G}^{U}|} \mathrm{L}_{\text{unsupervised}}(h_{\varphi}(G_j); H_{\varphi}(G_j)) \\
&- \lambda \sum_{j=1}^{|\mathbb{G}^{L}|+|\mathbb{G}^{U}|} \frac{1}{|G_j|} \sum_{k=1}^{K} I(H_{\phi}^{k}(G_j); H_{\varphi}^{k}(G_j))
\end{aligned}
$$



Results on Graph Classification and Regression

Dataset	MUTAG	PTC-MR	RDT-B	RDT-M5K	IMDB-B	IMDB-M
(No. Graphs)	188	344	2000	4999	1000	1500
(No. classes)	2	2	2	5	2	3
(Avg. Graph Size)	17.93	14.29	429.63	508.52	19.77	13.00
Graph Kernels						
RW [14]	83.72 ± 1.50	57.85 ± 1.30	OMR	OMR	50.68 ± 0.26	34.65 ± 0.19
SP [3]	85.22 ± 2.43	58.24 ± 2.44	64.11 ± 0.14	39.55 ± 0.22	55.60 ± 0.22	37.99 ± 0.30
GK [55]	81.66 ± 2.11	57.26 ± 1.41	77.34 ± 0.18	41.01 ± 0.17	65.87 ± 0.98	43.89 ± 0.38
WL [54]	80.72 ± 3.00	57.97 ± 0.49	68.82 ± 0.41	46.06 ± 0.21	72.30 ± 3.44	46.95 ± 0.46
DGK [68]	87.44 ± 2.72	60.08 ± 2.55	78.04 ± 0.39	41.27 ± 0.18	66.96 ± 0.56	44.55 ± 0.52
MLG [28]	87.94 ± 1.61	$\mathbf{63.26 \pm 1.48}$	> 1 Day	> 1 Day	66.55 ± 0.25	41.17 ± 0.03
Other Unsupervised Methods						
node2vec [17]	72.63 ± 10.20	58.58 ± 8.00	-	-	-	-
sub2vec [1]	61.05 ± 15.80	59.99 ± 6.38	71.48 ± 0.41	36.68 ± 0.42	55.26 ± 1.54	36.67 ± 0.83
graph2vec [38]	83.15 ± 9.25	60.17 ± 6.86	75.78 ± 1.03	47.86 ± 0.26	71.1 ± 0.54	$\mathbf{50.44 \pm 0.87}$
InfoGraph	$\mathbf{89.01 \pm 1.13}$	61.65 ± 1.43	$\mathbf{82.50 \pm 1.42}$	$\mathbf{53.46 \pm 1.03}$	$\mathbf{73.03 \pm 0.87}$	49.69 ± 0.53

Table 1: Graph classification accuracy with unsupervised methods

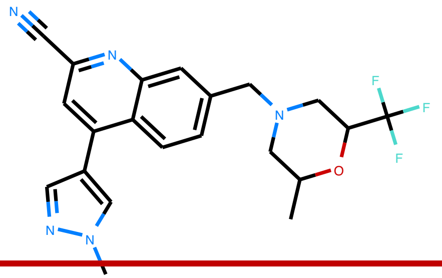
Target	Mu (0)	Alpha (1)	HOMO (2)	LUMO (3)	Gap (4)	R2 (5)	ZPVE(6)	U0 (7)	U (8)	H (9)	G(10)	Cv (11)
MAE	0.3201	0.5792	0.0060	0.0062	0.0091	10.0469	0.0007	0.3204	0.2934	0.2722	0.2948	0.2368
Semi-Supervised	Error Ratio											
Mean-Teachers	1.09	1.00	**0.99**	1.00	**0.97**	0.52	0.77	1.16	0.93	0.79	0.86	0.86
InfoGraph	1.02	0.97	1.02	**0.99**	1.01	0.71	0.96	0.85	0.93	0.93	0.99	1.00
InfoGraph*	**0.99**	**0.94**	**0.99**	**0.99**	0.98	**0.49**	**0.52**	**0.44**	**0.58**	**0.57**	**0.54**	**0.83**

Table 2: Results of semi-supervised experiments on QM9 data set.

Research Problems

Target → Lead Discovery 2 years → Lead Optimization 3 years → Preclinical Study 2 years → Clinical Trial

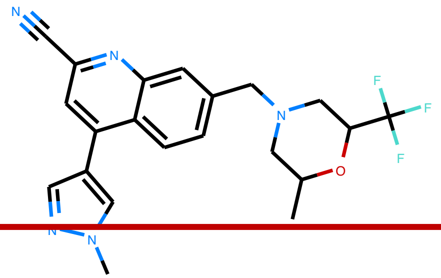
Property Prediction



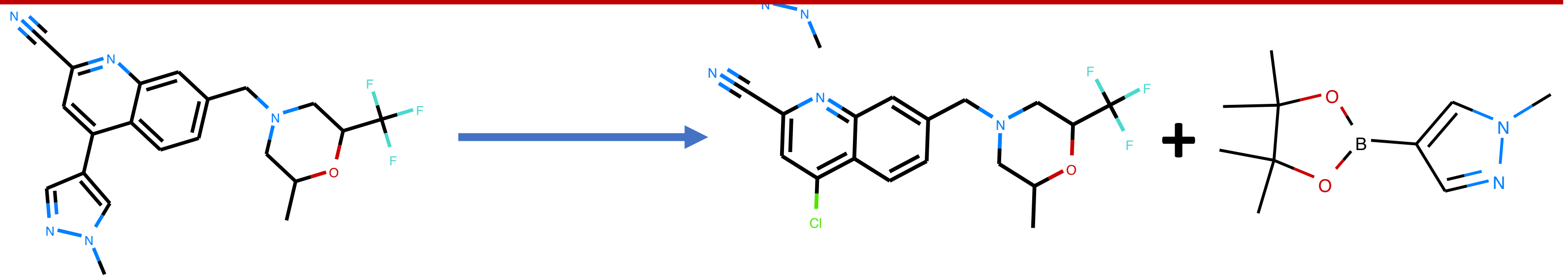
Property

Molecule Design and Optimization

Property



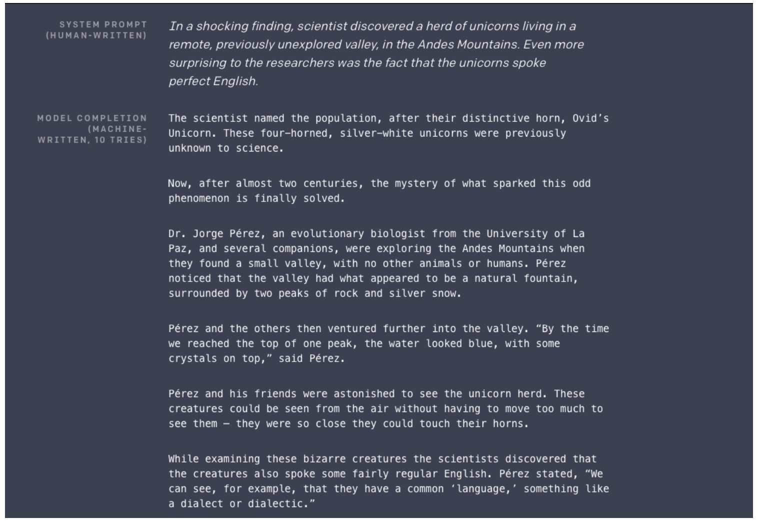
Retrosynthesis Prediction



Molecule Generation and Optimization

- Deep generative models for data generation

Image generation
(by StyleGAN, From Internet)



SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

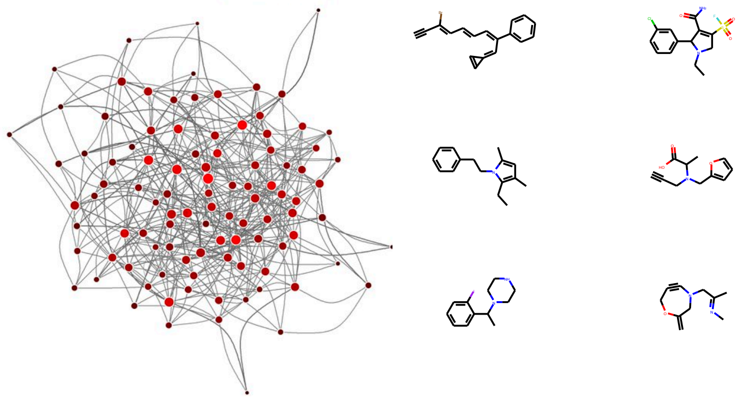
Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

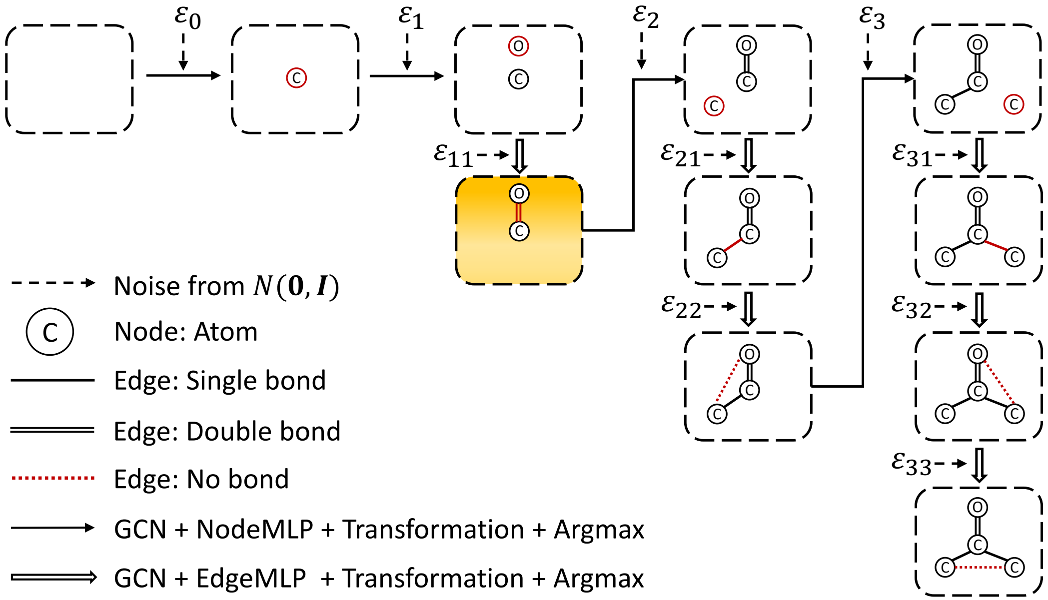
Text generated by by GPT-2,
Examples from Internet



Graphs?

GraphAF: an Autoregressive Flow for Molecular Graph Generation (Shi & Xu ICLR'20)

- Formulate graph generation as a sequential decision process
  - In each step, generate a new atom
  - Determine the bonds between the new atoms and existing atoms

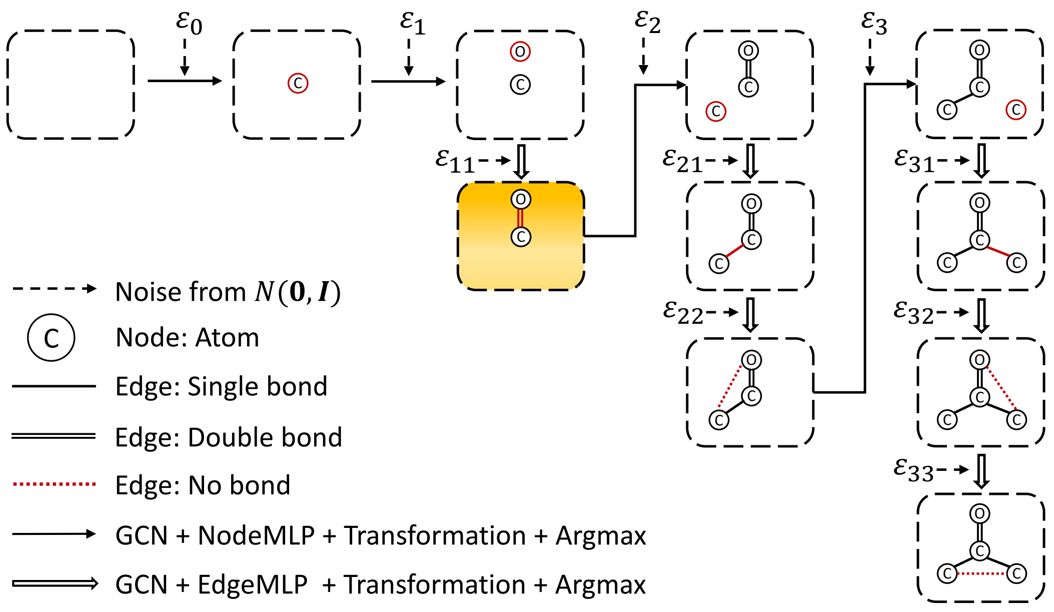


(a) Sampling Framework

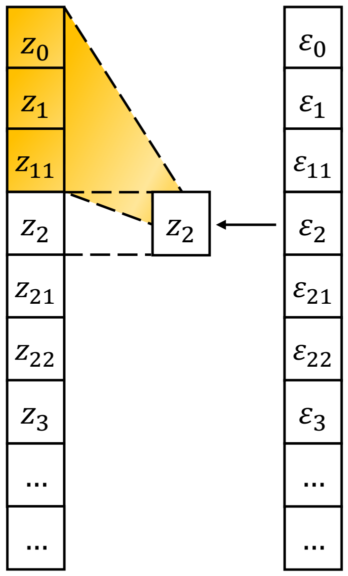
Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang.
"GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation." ICLR'20.

GraphAF: an Autoregressive Flow for Molecular Graph Generation

- Traverse a graph through BFS-order
  - Transform each graph into a sequence of nodes and edges
- Defines an invertible mapping from a base distribution (Gaussian distribution) to the observations (graph nodes and edge sequences)



(a) Sampling Framework

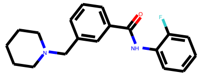
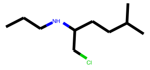
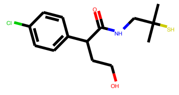
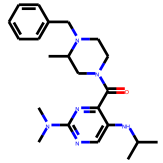
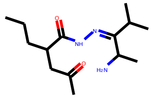
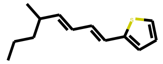
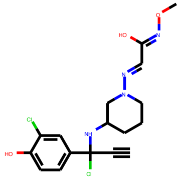
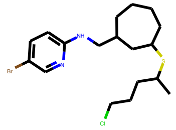
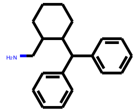
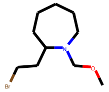
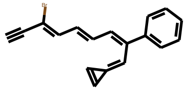
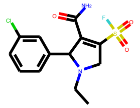
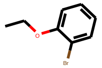
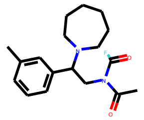
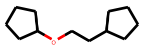
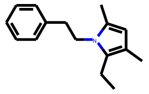
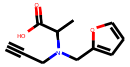
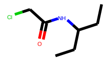
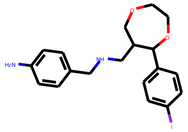
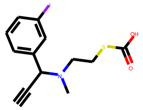
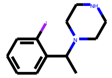
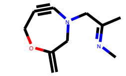
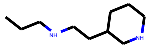
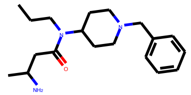
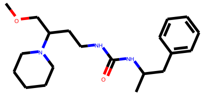
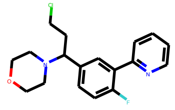
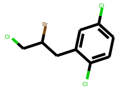
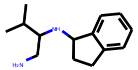
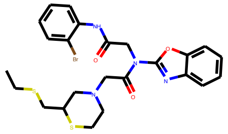
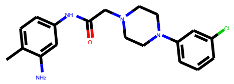


(b) Autoregressive Flow

Molecule Generation

- Training Data: ZINC250K
  - 250K drug-like molecules with a maximum atom number of 38
  - 9 atom types and 3 edge types

Method	Validity	Validity w/o check	Uniqueness	Novelty	Reconstruction
JT-VAE	100%	—	100%‡	100%‡	76.7%
GCPN	100%	20%†	99.97%‡	100%‡	—
MRNN	100%	65%	99.89%	100%	—
GraphNVP	42.60%	—	94.80%	100%	100%
GraphAF	100%	68%	99.10%	100%	100%

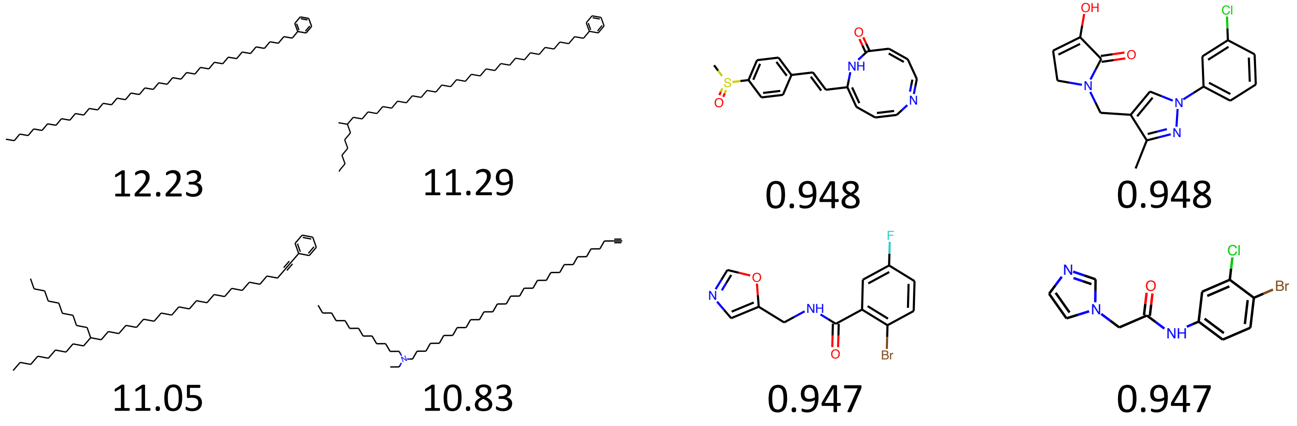
Goal-Directed Molecule Generation with Reinforcement Learning

- Fine tune the generation policy with reinforcement learning to optimize the properties of generated molecules
- **State**: current subgraph G_i
- **Action**: generating a new atom (i.e. $\mathrm{p}(X_i|G_i)$) or a new edge ($\mathrm{p}(A_{ij}|G_i, X_i, A_{i,1:j-1})$).
- **Reward Design**: the properties of molecules (final reward) and chemical validity (intermediate and final reward)

Molecule Optimization

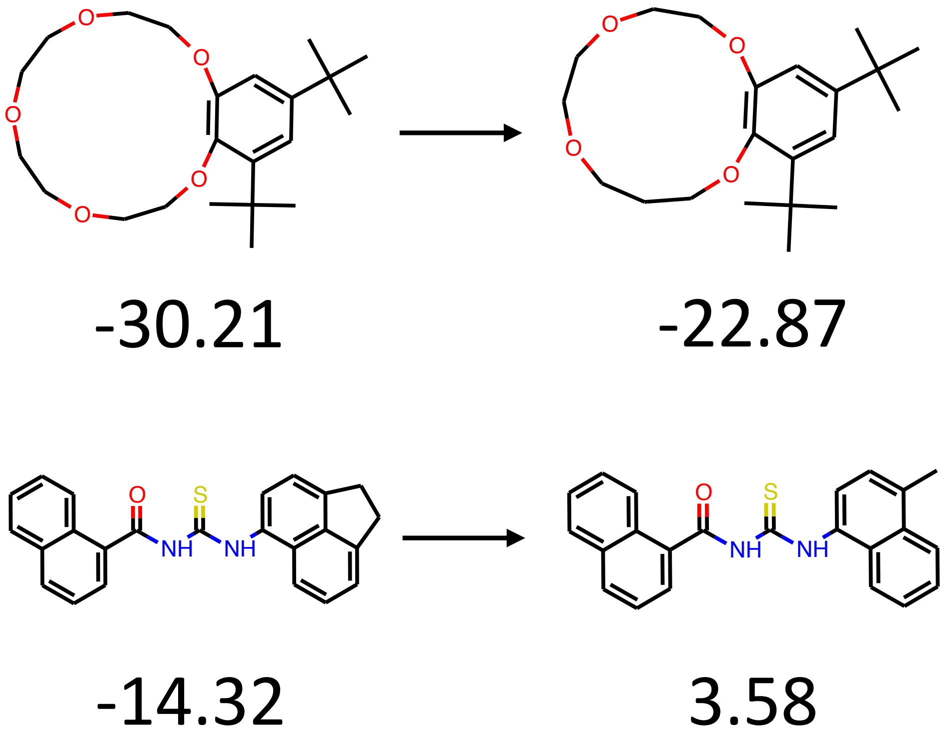
- Properties
  - Penalized logP
  - QED (druglikeness)

Method	Penalized logP 1st	2nd	3rd	Validity	QED 1st	2nd	3rd	Validity
ZINC (Dataset)	4.52	4.30	4.23	100.0%	0.948	0.948	0.948	100.0%
JT-VAE (Jin et al., 2018)	5.30	4.93	4.49	100.0%	0.925	0.911	0.910	100.0%
GCPN (You et al., 2018a)	7.98	7.85	7.80	100.0%	**0.948**	0.947	0.946	100.0%
MRNN[1] (Popova et al., 2019)	8.63	6.08	4.73	100.0%	0.844	0.796	0.736	100.0%
GraphAF	**12.23**	**11.29**	**11.05**	100.0%	**0.948**	**0.948**	**0.947**	100.0%



(a) Penalized logP optimization (b) QED optimization

Constrained Optimization

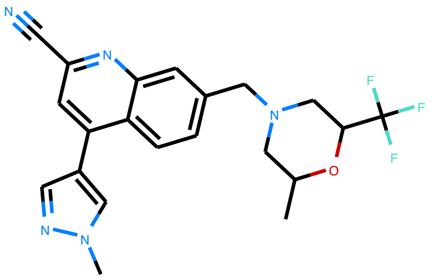


(c) Constrained optimization

Research Problems

Target → Lead Discovery 2 years → Lead Optimization 3 years → Preclinical Study 2 years → Clinical Trial

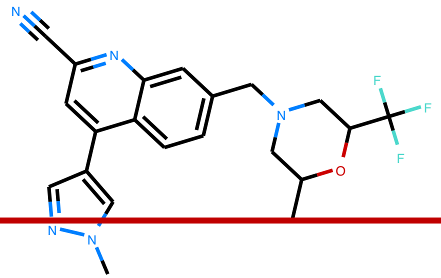
Property Prediction



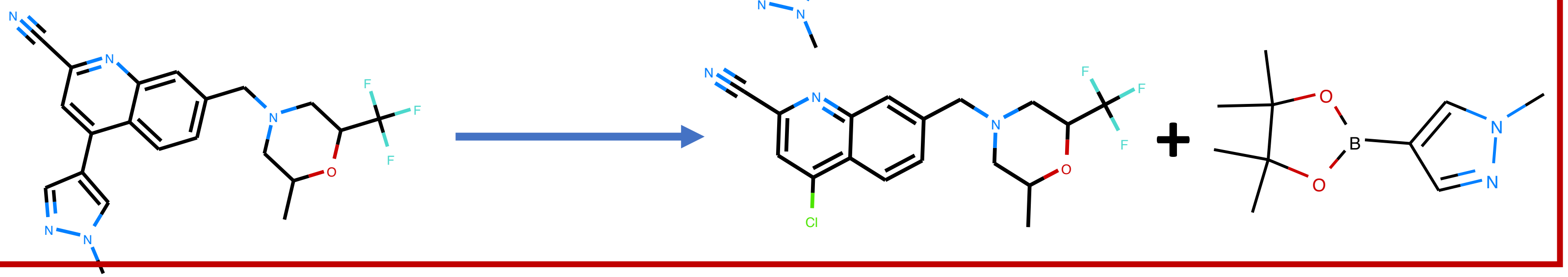
Property

Molecule Design and Optimization

Property

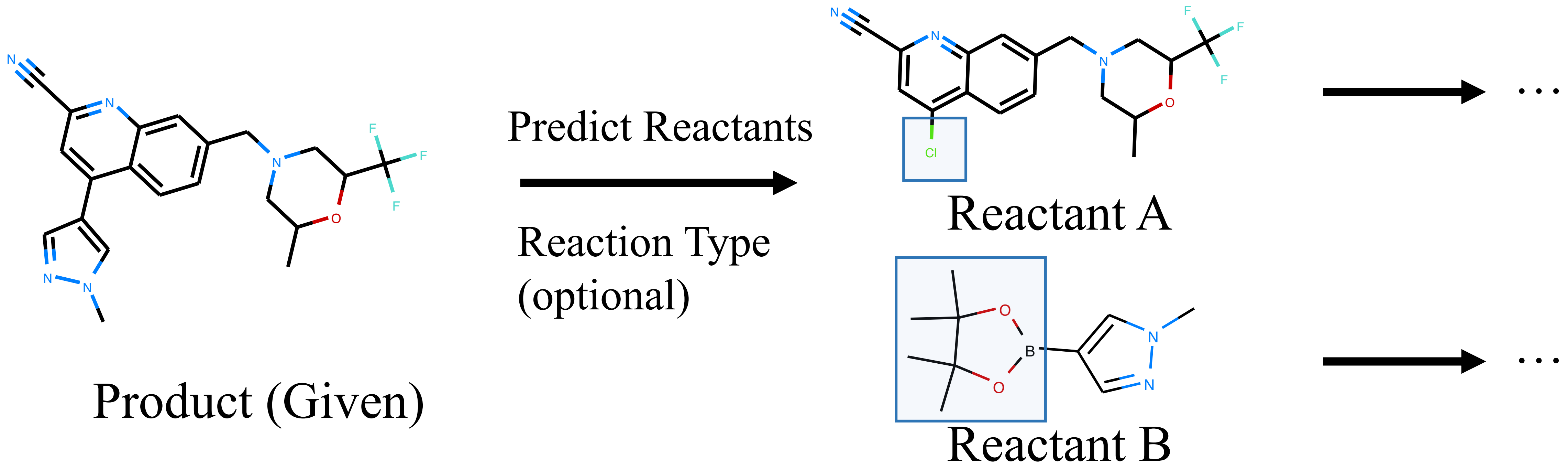


Retrosynthesis Prediction



Retrosynthesis Prediction

- Once a molecular structure is designed, how to synthesize it?
- Retrosynthesis planning/prediction
  - Identify a set of reactants to synthesize a target molecule

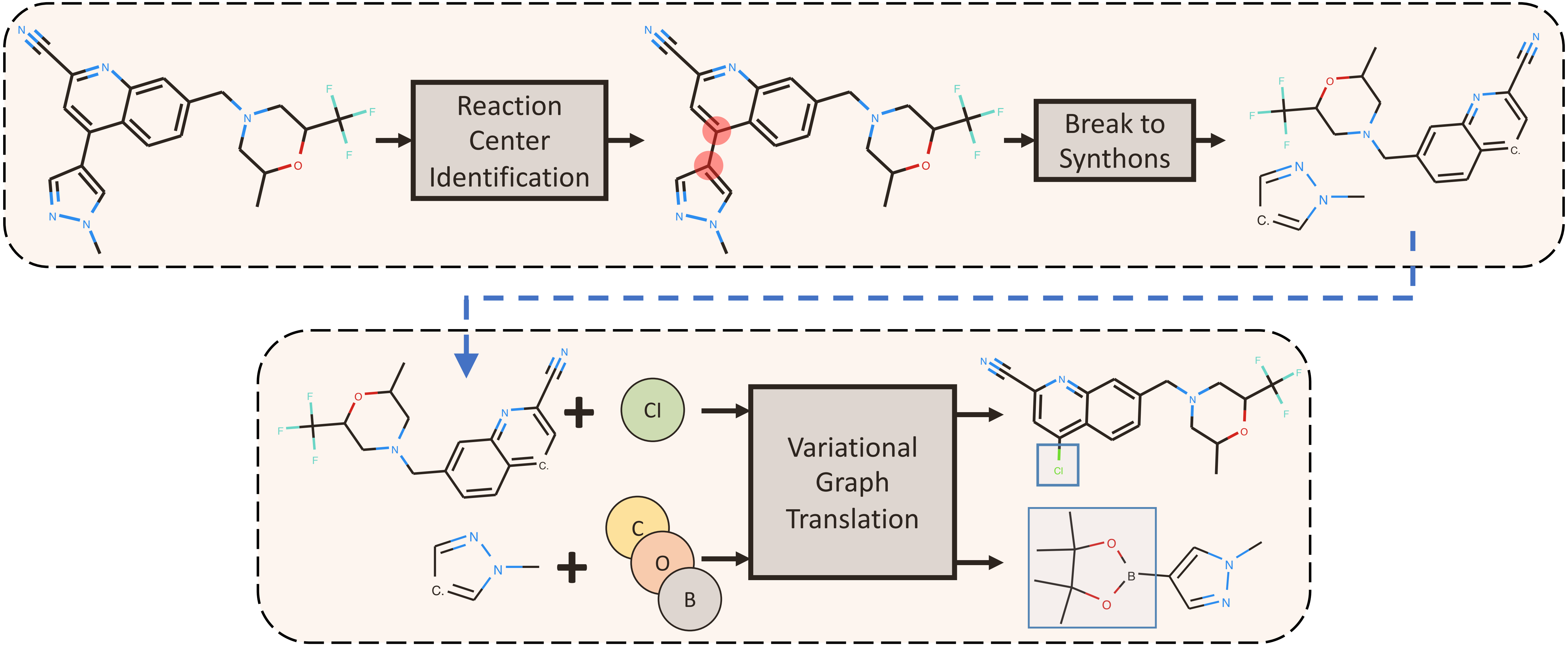


A Graph to Graphs Framework for Retrosynthesis Prediction (Shi et al. 2020)

- Each molecule is represented as a molecular graph
- Formulate the problem as a graph (product molecule) to a set of graphs (reactants)
- The whole framework are divided into two stages
  - Reaction center identification
  - Graph Translation

Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang and Jian Tang. **A Graph to Graphs Framework for Retrosynthesis Prediction**. ICML, 2020.

The G2Gs Framework (Shi et al. 2020)

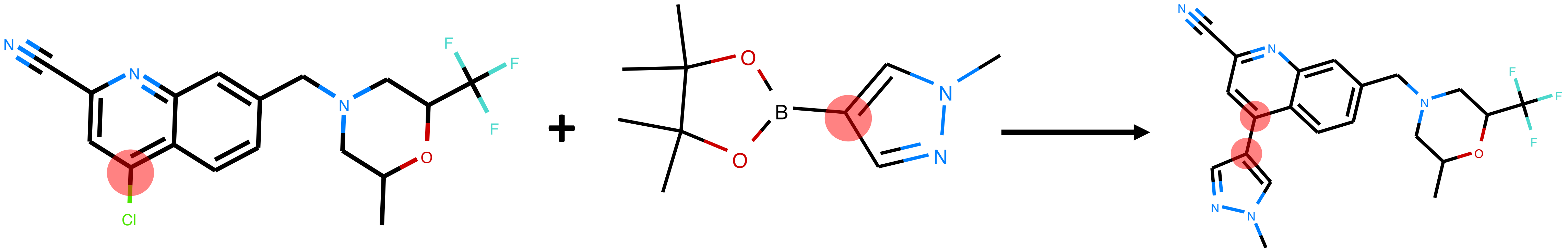


Shi et al., 2020, A Graph to Graphs Framework for Retrosynthesis Prediction

Reaction Center Prediction

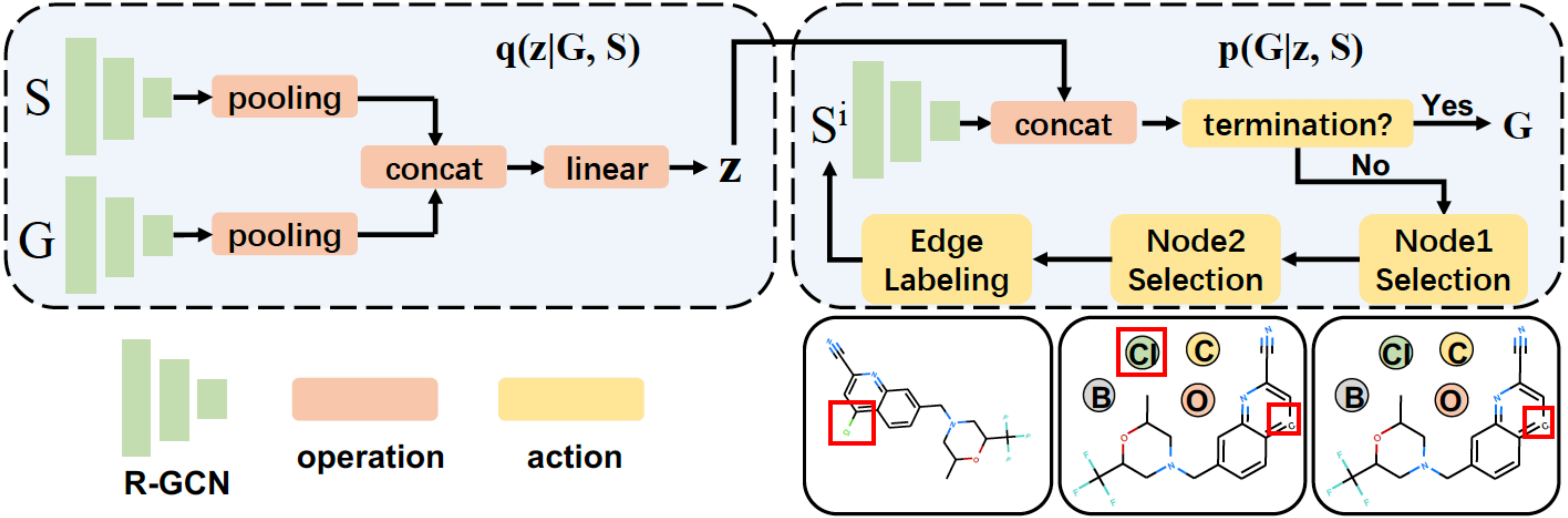
An atom pair *(i, j)* is a reaction center if:

- There is a bond between atom *i* and atom *j* in product
- There is no bond between atom *i* and atom *j* in reactants



Graph Translation

- Translate the incomplete synthon to the final reactant
- A variational graph to graph framework
  - A latent variable z is introduced to capture the uncertainty during translation



Experiments

- Experiment Setup
  - Benchmark data set USPTO-50K, containing 50k atom-mapped reactions
  - Evaluation metrics: top-k exact match (based on canonical SMILES) accuracy

Table 1. Top-k exact match accuracy when reaction class is given. Results of all baselines are directly taken from (Dai et al., 2019).

Methods	Top-k accuracy %			
	1	3	5	10
Template-free				
Seq2seq	37.4	52.4	57.0	61.7
G2Gs	**61.0**	**81.3**	**86.0**	**88.7**
Template-based				
Retrosim	52.9	73.8	81.2	88.1
Neuralsym	55.3	76.0	81.4	85.1
GLN	**64.2**	**79.1**	**85.2**	**90.0**

Table 2. Top-k exact match accuracy when reaction class is unknown. Results of all baselines are taken from (Dai et al., 2019).

Methods	Top-k accuracy %			
	1	3	5	10
Template-free				
Transformer	37.9	57.3	62.7	/
G2Gs	**48.9**	**67.6**	**72.5**	**75.5**
Template-based				
Retrosim	37.3	54.7	63.3	74.1
Neuralsym	44.4	65.3	72.4	78.9
GLN	**52.5**	**69.0**	**75.6**	**83.7**

Conclusion

- Drug discovery is slow and expensive
  - Great potential for AI in accelerating the process
- Great representation learning for drug discovery
  - Properties prediction
  - De novo molecule design and optimization
  - Retrosynthesis
- Next Step: Drug Discovery with Limited Labeled Data
  - Self-supervised Learning
  - Multi-task/Transfer Learning, Few-shot Learning

Thanks!

- **Current Students**
    - Meng Qu
    - Zhaocheng Zhu
    - Andreea Deac
    - Louis-Pascal Xhonneux
    - Shengchao Liu
    - Chence Shi
    - Minkai Xu
- **Collaborators and previous students**:, Yoshua Bengio, Pietro Liò, Fanyun Sun, Hongyu Guo, Jordan Hoffmann, Vikas Verma,….

NSERC CRSNG

CIFAR

Microsoft Research

amazon

IVADO
INSTITUT DE VALORISATION DES DONNÉES

Canada
NRC·CNRC

Tencent AI Lab

SAMSUNG