Hierarchical Generation of Molecular Graphs using Structural Motifs

Wengong Jin, Regina Barzilay, Tommi Jaakkola
MIT CSAIL

Drug Discovery via Generative Models

- Drug discovery: finding molecules with desired chemical properties
- The primary challenge: large search space

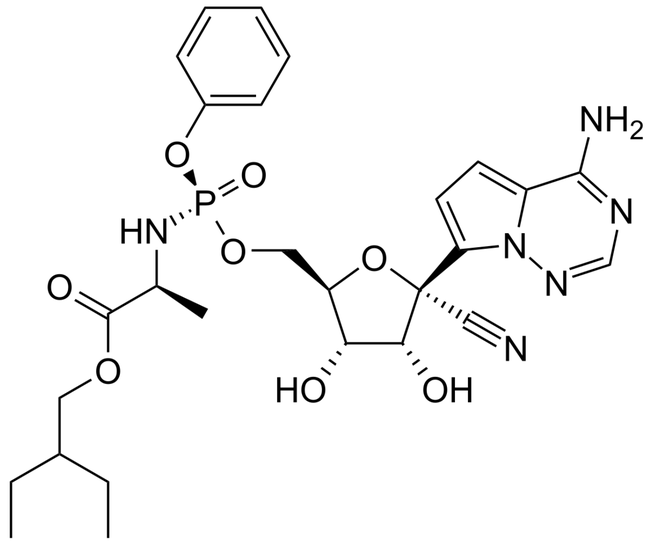
Criterion:
- Safe
- Cures COVID

Search →

10^{30}

Potential candidates

Find →



Remdesivir?

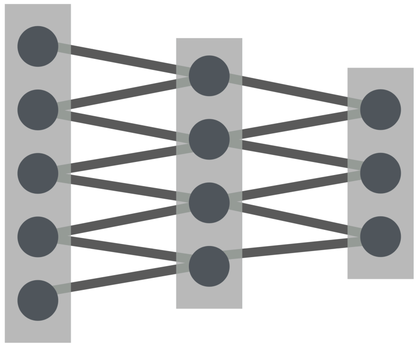
Drug Discovery via Generative Models

- Generative models can be used to efficiently search in the chemical space
- Given a specified criterion, the model generates a molecule with desired properties.

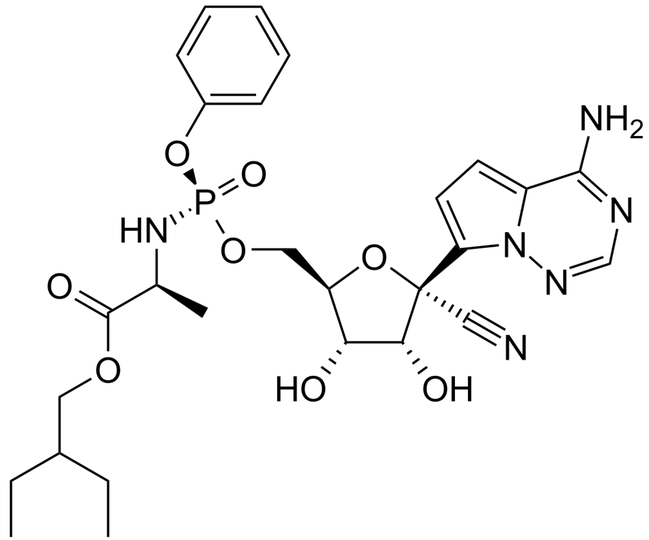
Criterion:

- Safe
- Cures COVID

Condition



Generate



Generative Model

Remdesivir

Molecular Graph Generation

- Consider connected graphs…
- Different type of graphs require different generation method.
- What kind of generation method is suitable for molecules?

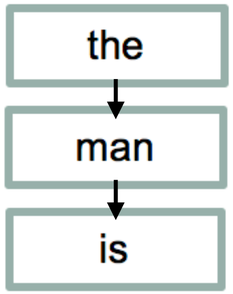
Line graph (text)

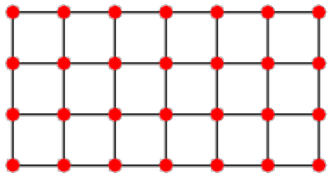
Grid graph (Images)

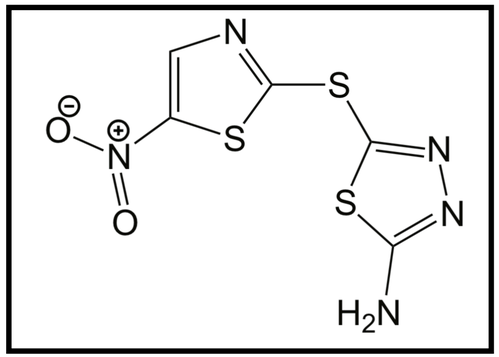
Low tree-width graph (molecule)

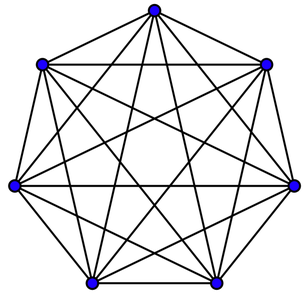
Fully connected graph

Complexity



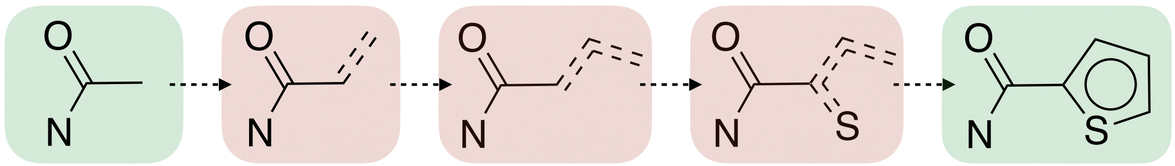






Previous Methods for Molecule Generation

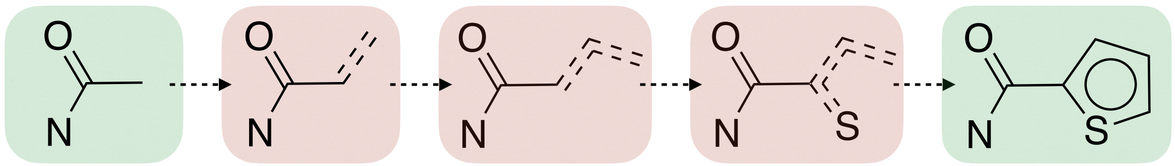
- **Atom based methods**: CG-VAE (Liu et al. 2018), DeepGMG (Li et al. 2018), GraphRNN (You et al. 2018), and more



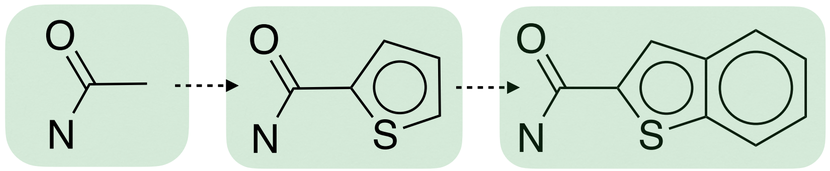
Atom based

Previous Methods for Molecule Generation

- **Atom based methods**: CG-VAE (Liu et al. 2018), DeepGMG (Li et al. 2018), GraphRNN (You et al. 2018), and more
- **Substructure based methods**: JT-VAE (Jin et al., 2018)
  - Incorporating inductive bias (i.e., low tree-width) into generation
  - Each time generate a cycle or edge



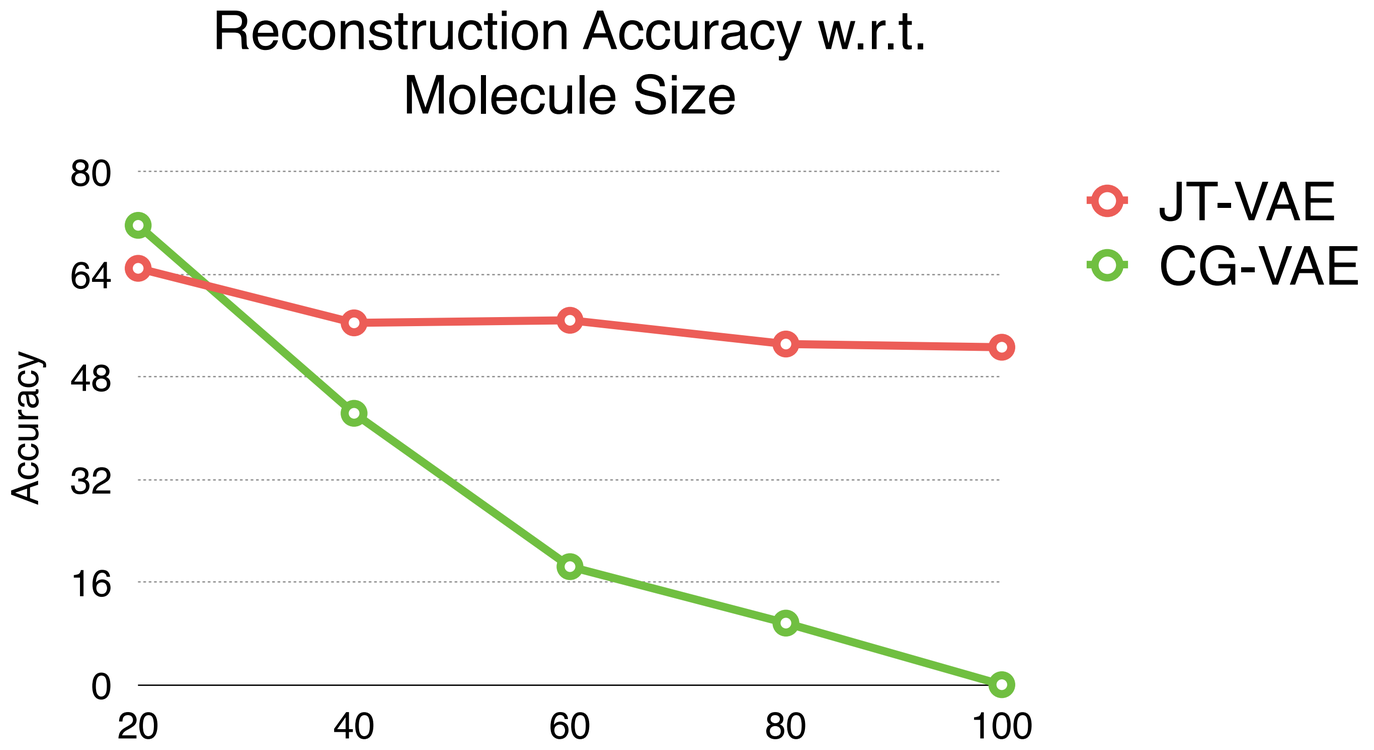
Atom based



Substructure based

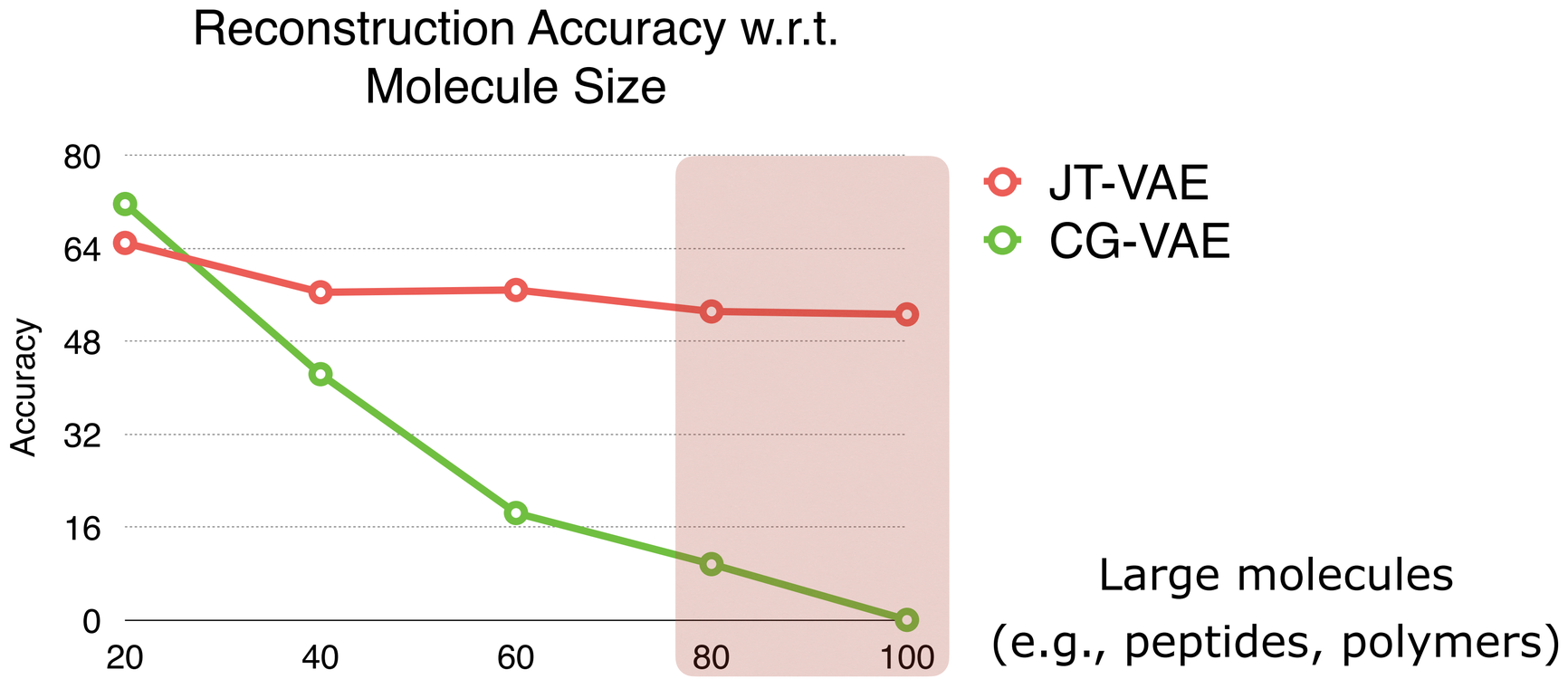
Previous methods: limitation

- **Atom based methods**: CG-VAE (Liu et al. 2018), DeepGMG (Li et al. 2018), GraphRNN (You et al. 2018), and more
- **Substructure based methods**: JT-VAE (Jin et al., 2018)



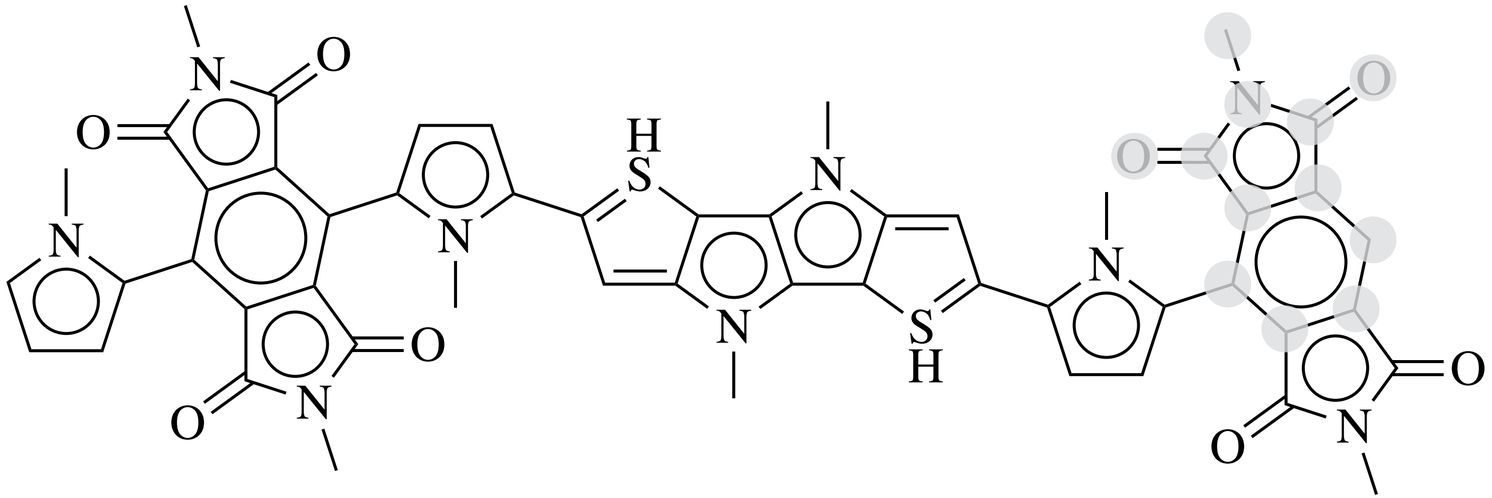
Previous methods: limitation

- **Atom based methods**: CG-VAE (Liu et al. 2018), DeepGMG (Li et al. 2018), GraphRNN (You et al. 2018), and more
- **Substructure based methods**: JT-VAE (Jin et al., 2018)



Failure in Generating Large Molecules

- **Atom based methods**: CG-VAE (Liu et al. 2018), DeepGMG (Li et al. 2018), GraphRNN (You et al. 2018), and more



CG-VAE

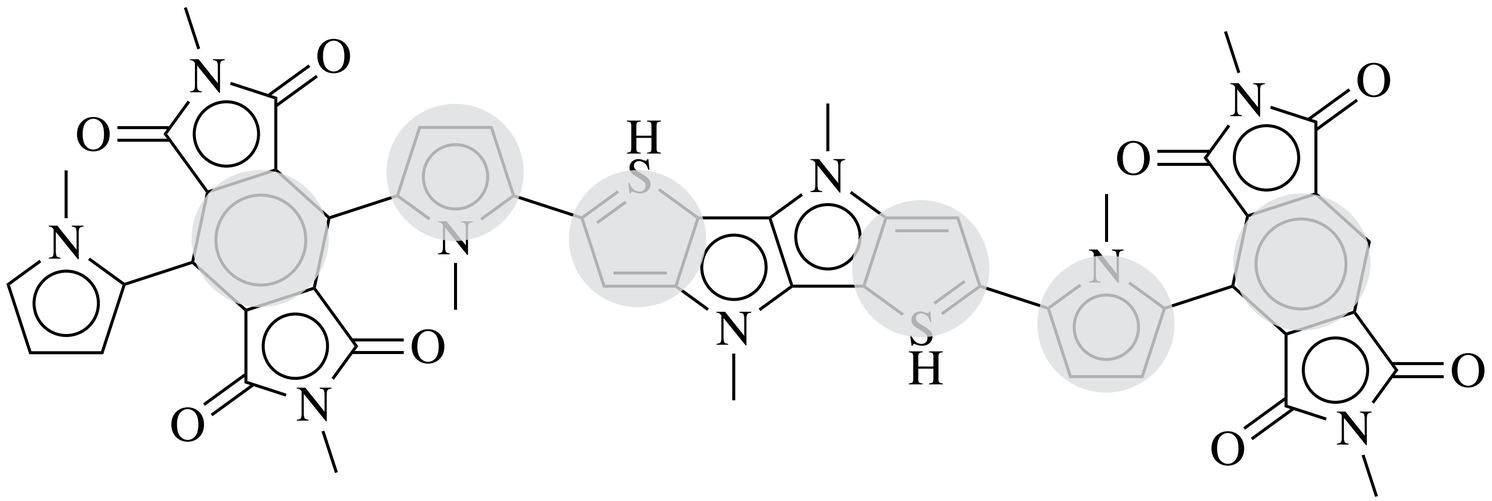
70 atom predictions

+ 70 bond predictions

- Many Generation Steps: Vanishing gradient + error accumulation

Failure in Generating Large Molecules

- **Atom based methods**: CG-VAE (Liu et al. 2018), DeepGMG (Li et al. 2018), GraphRNN (You et al. 2018), and more
- **Substructure based methods**: JT-VAE (Jin et al., 2018)



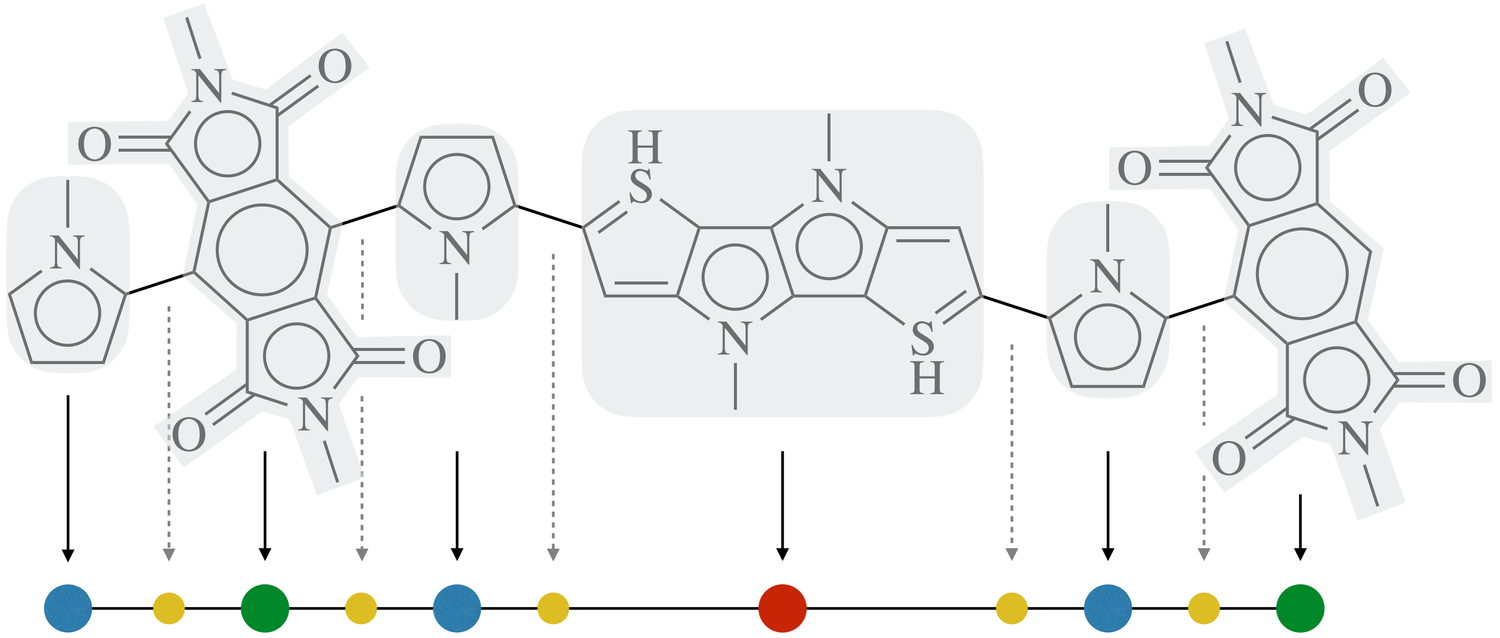
JT-VAE:

35 substructure (ring/bond) predictions

- JT-VAE decoder requires each substructure neighborhood to be assembled <u>in one go</u>, making it combinatorially challenging to handle large substructures.

Larger Building Blocks: Motifs

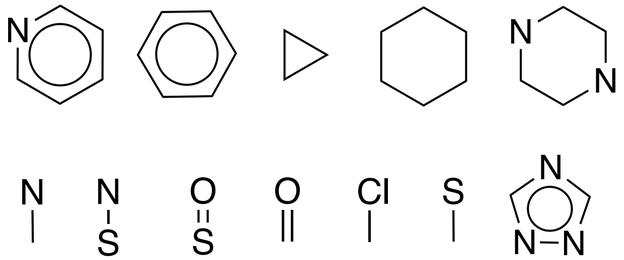
- JT-VAE only considered single rings and bonds as building blocks
- How about using larger building blocks — motifs with flexible structures, not restricted to rings and bonds?
- Large molecules such as polymers exhibit clear hierarchical structure, being built from repeated structural motifs.



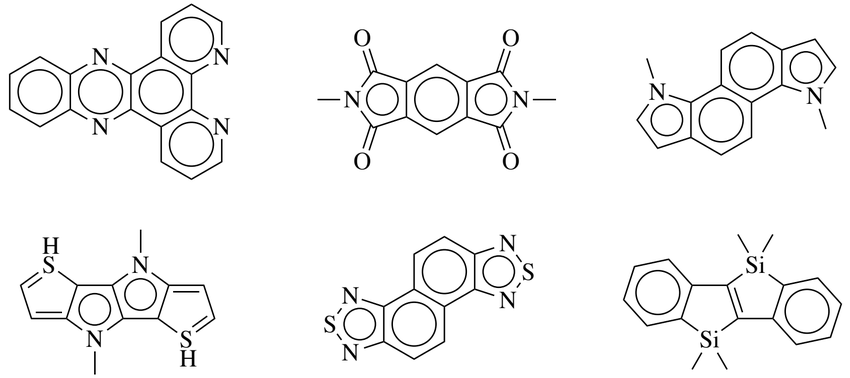
- Only 11 steps to generate this polymer structure.

NLP Analogy

- Atom-based generation == character-based generation
- Substructure-based generation == word-based generation
- Motif-based generation == phrase-based generation



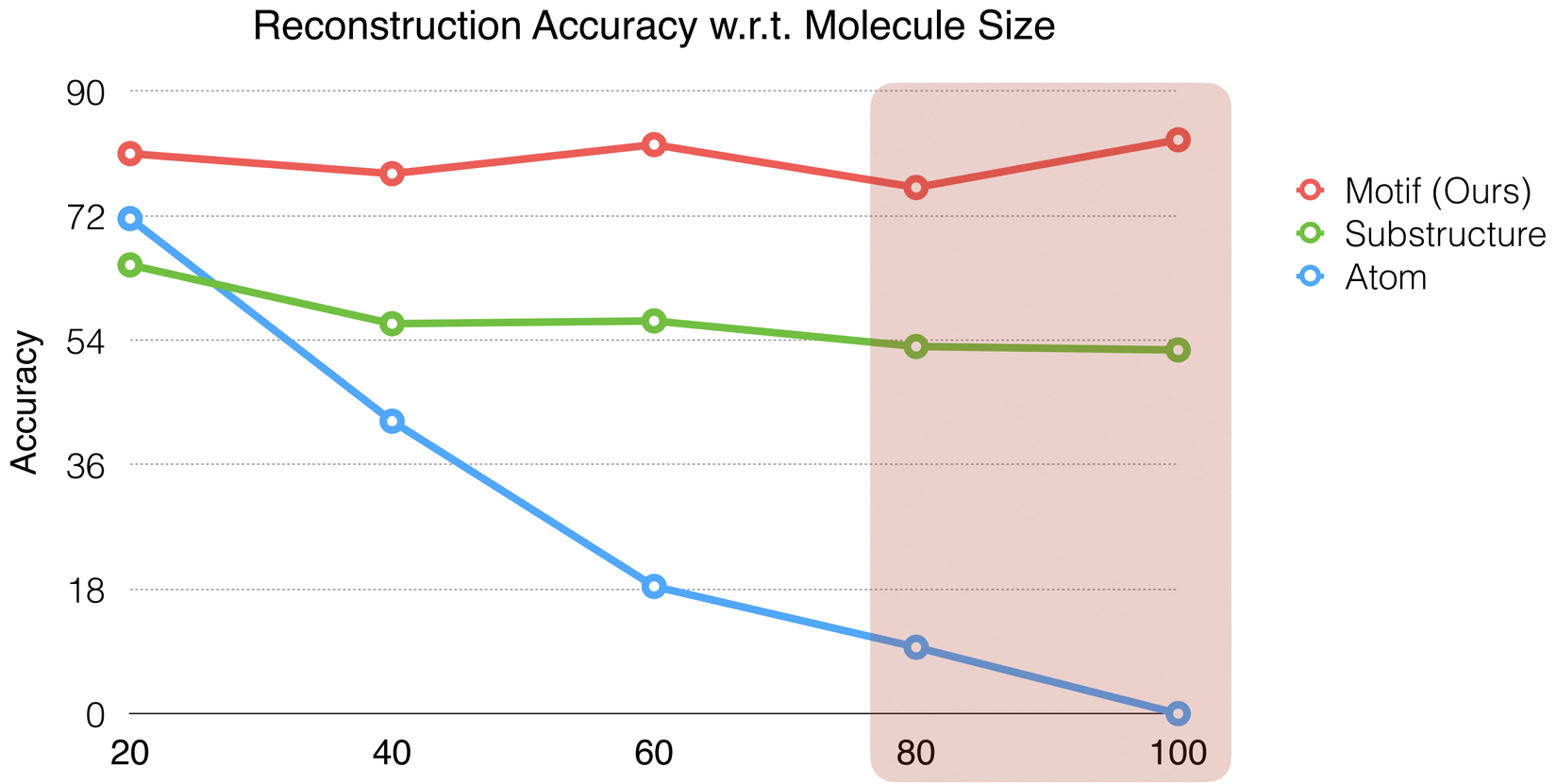
- Substructures
- (ring and bond only)
- Word-based generation



- Motifs
- (structures can be flexible)
- Phrase-based generation

Our New Architecture: HierVAE

- **Generates molecules motif by motif**
  - Faster and more efficient
  - Much higher reconstruction accuracy for large molecules



Our New Architecture: HierVAE

- **Motif extraction from data**
  - Motif extraction is based on heuristics
  - Later I will discuss how motifs can be learned (based on given properties).

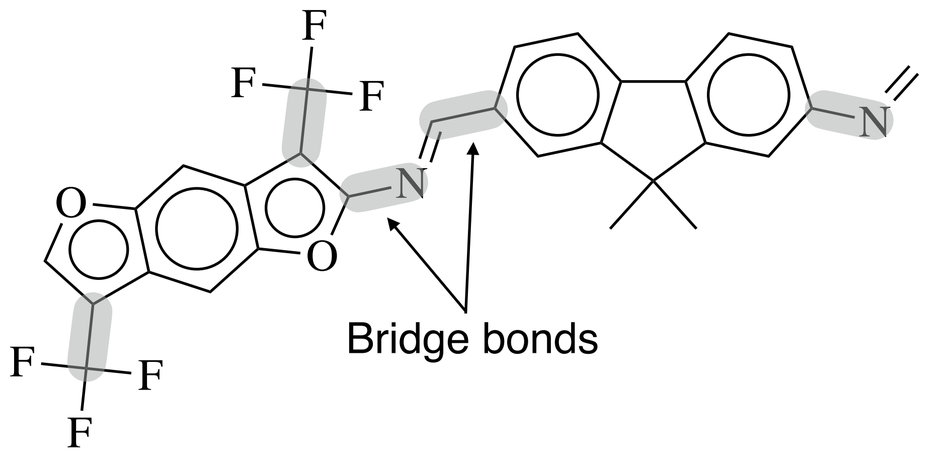
- **Hierarchical Graph Encoder**
  - Representing molecules at both motif and atom level.
  - Designed to match the decoding process

- **Hierarchical Graph Decoder**
  - Each generation step needs to resolve:
    1. What's the next motif?
    2. How it should be attached to current graph?

Motif Extraction Algorithm

- A molecule is decomposed into disconnected motifs as follows:

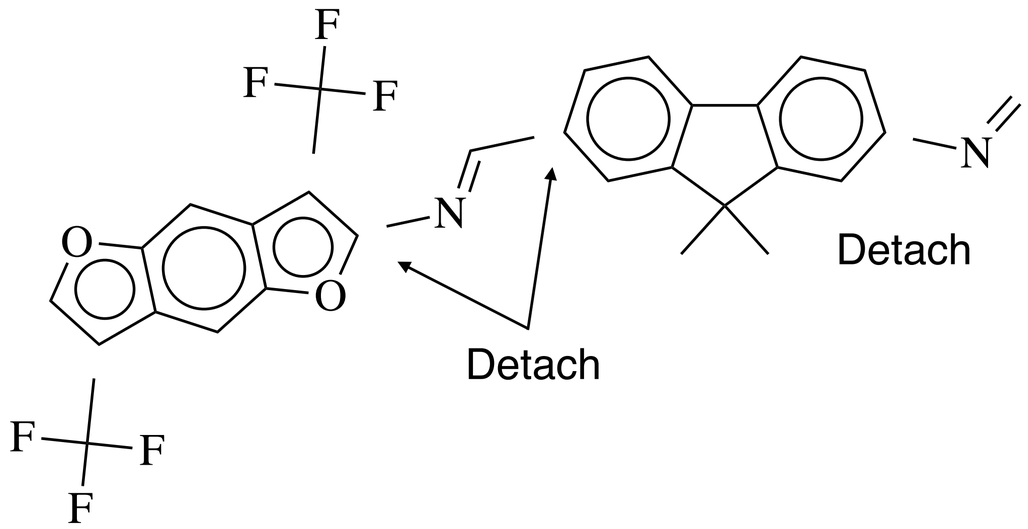
1. Find all the bridge bonds (u, v) such that either u or v is part of a ring.



Motif Extraction Algorithm

- A molecule is decomposed into disconnected motifs as follows:

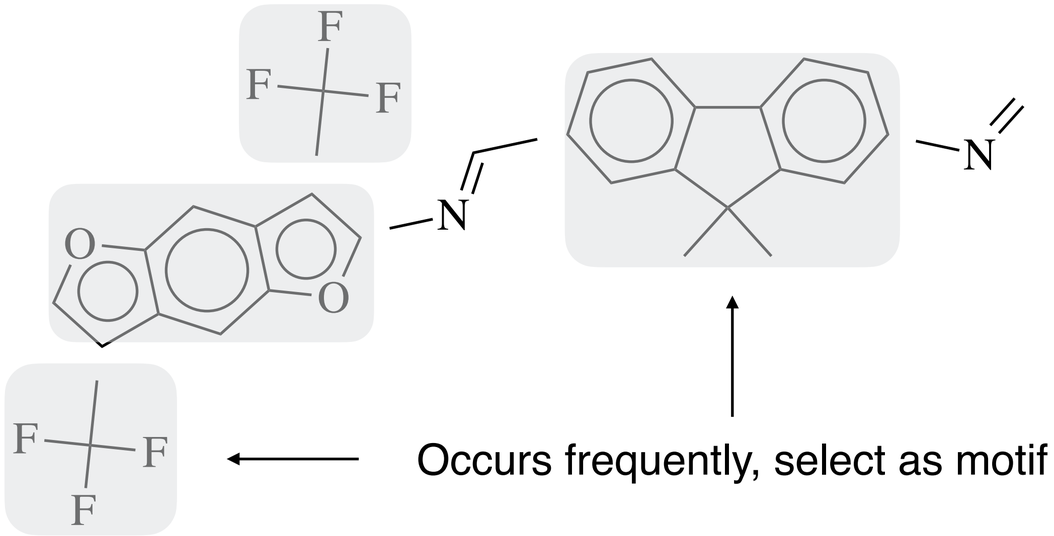
1. Find all the bridge bonds (u, v) such that either u or v is part of a ring.
2. Detach all bridge bonds from its neighbors.



Motif Extraction Algorithm

- A molecule is decomposed into disconnected components as follows:

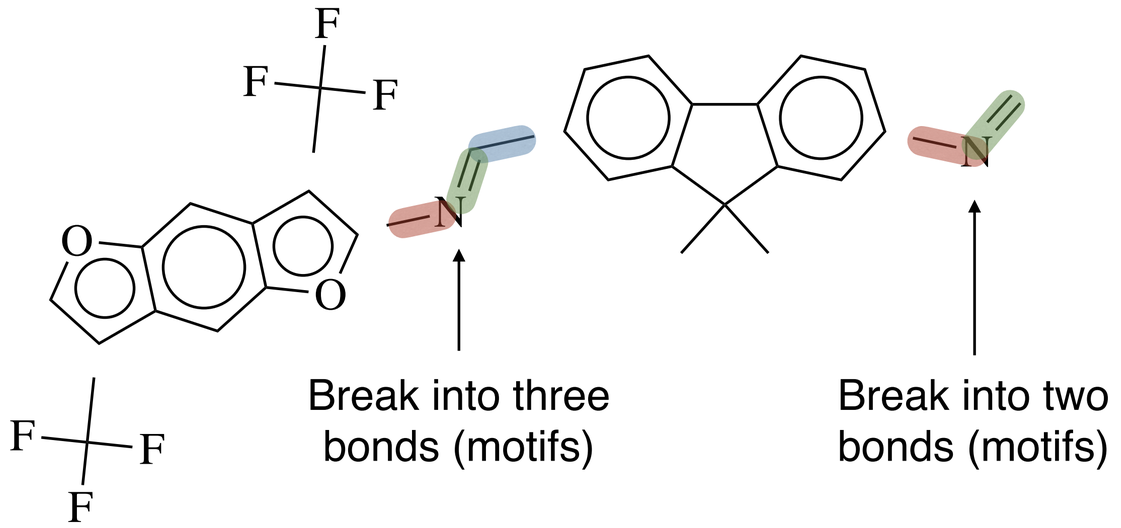
1. Find all the bridge bonds (u, v) such that either u or v is part of a ring.
2. Detach all bridge bonds from its neighbors.
3. Select all components as motifs if it occurs frequently in the training set.



Motif Extraction Algorithm

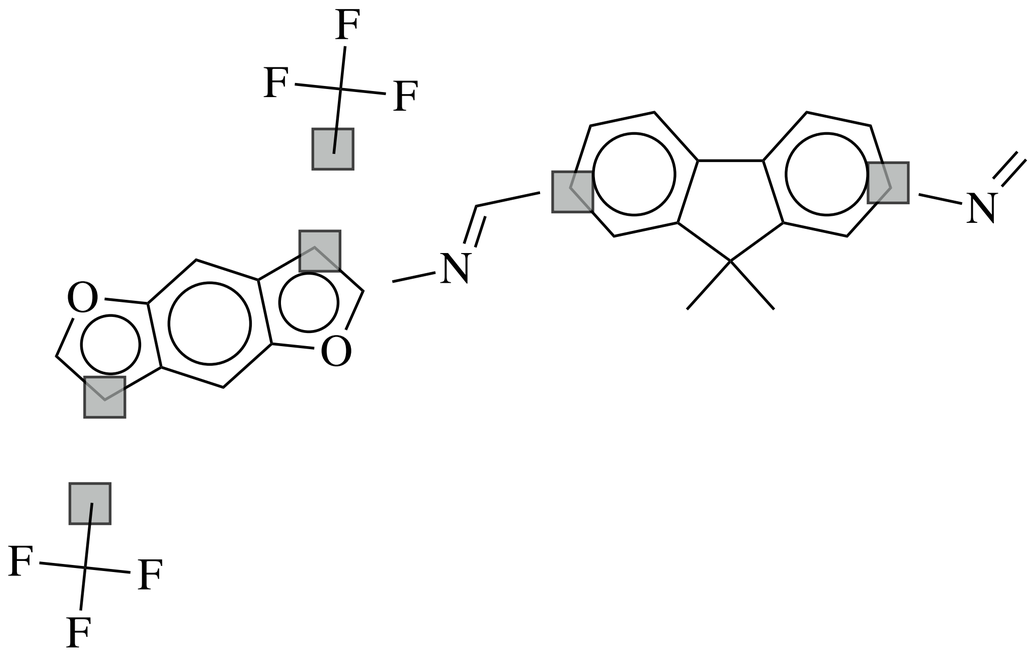
- A molecule is decomposed into disconnected components as follows:

1. Find all the bridge bonds (u, v) such that either u or v is part of a ring.
2. Detach all bridge bonds from its neighbors.
3. Select all components as motifs if it occurs frequently in the training set.
4. If a component is not selected, further decompose it into basic rings and bonds.



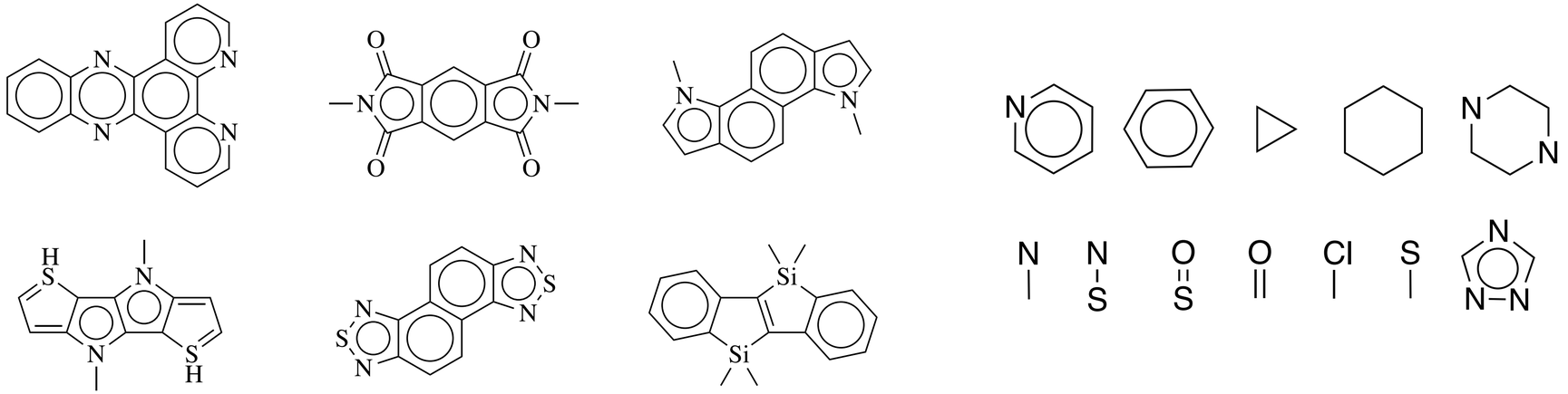
Mark attaching points

- Motif decomposition loses atom-level connectivity information
- For ease of reconstruction, we propose to mark attaching points in each motif.

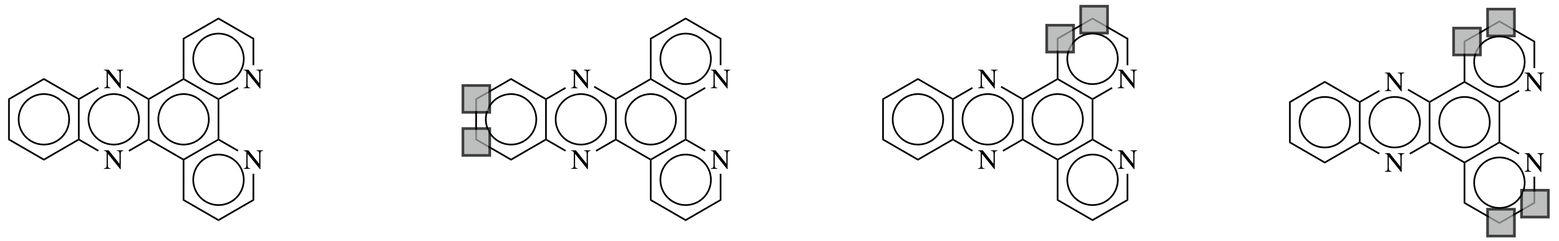


Motif Vocabulary

- We can construct a motif vocabulary given a training set (usually <500)

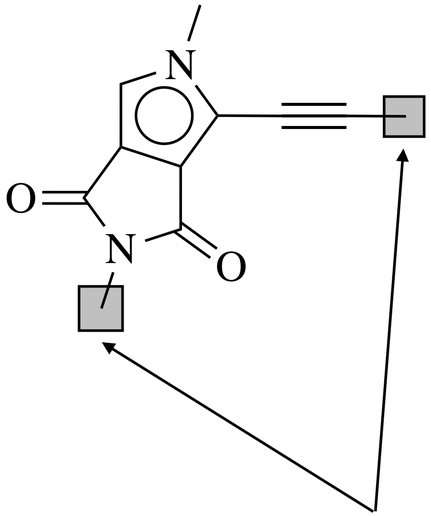


- Each motif also has a vocabulary of possible attaching point configurations.
  - Usually less than 10 because motifs have regular attachment patterns.
  - The attachment vocabulary covers >97% of the molecules in test set.



Generation Process

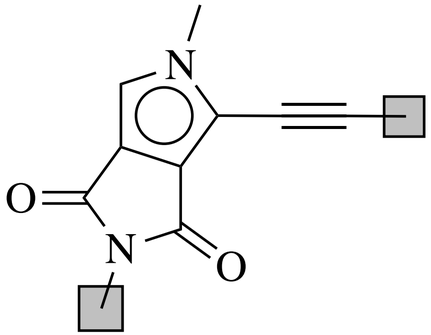
Current state



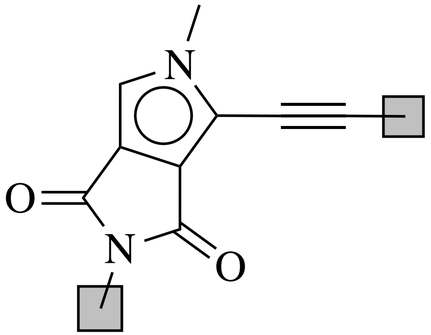
During generation, we maintain all possible positions to which new motifs will be attached

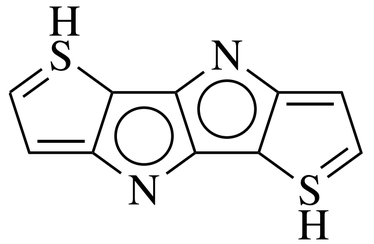
Generation Process

Current state

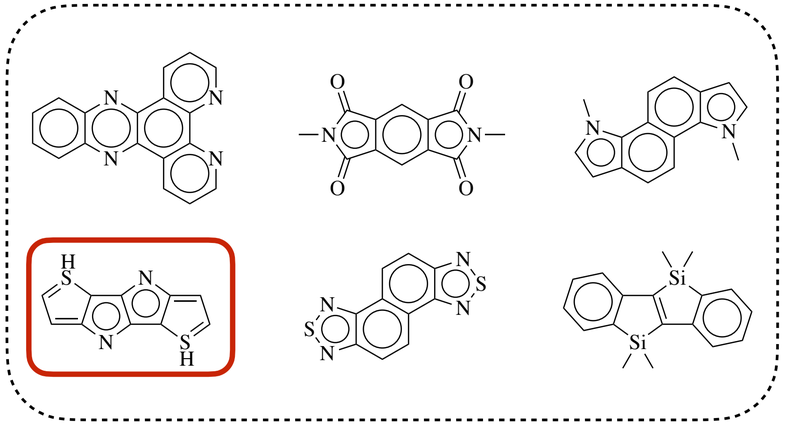


Step 1: Motif Prediction



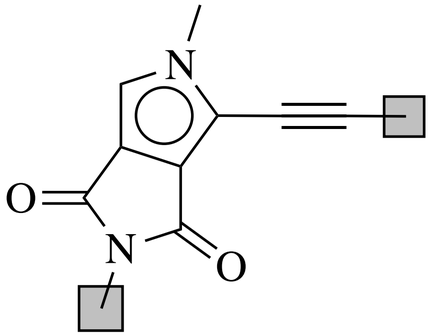


Motif Vocabulary

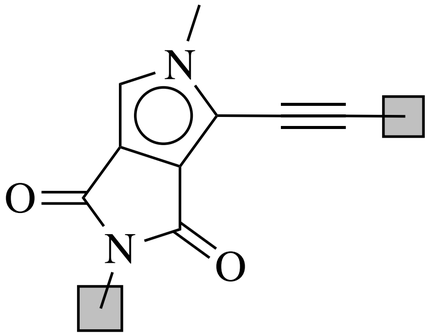


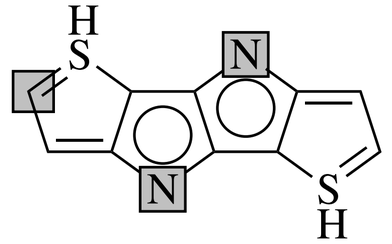
Generation Process

Current state

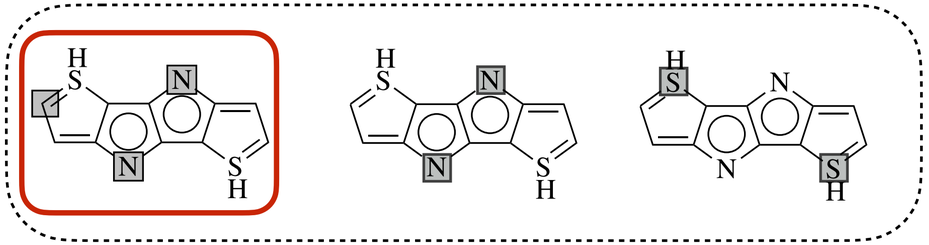


Step 2: Attachment Prediction



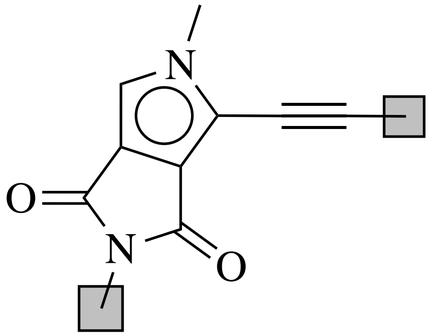


Attachment Vocabulary

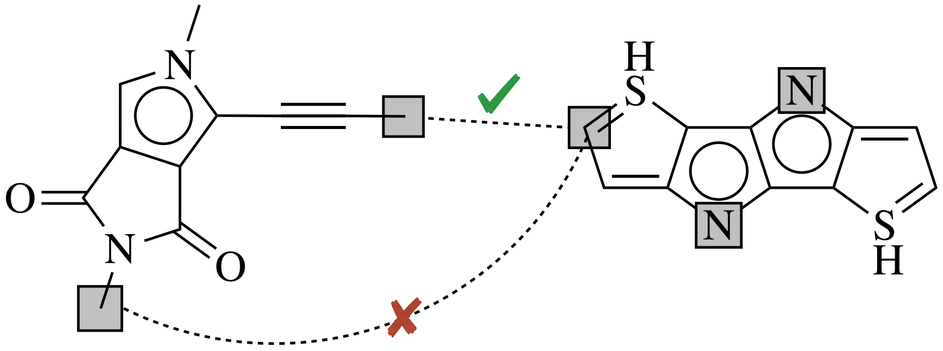


Generation Process

Current state

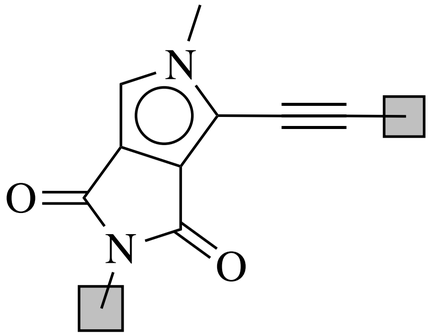


Step 3: Graph Prediction

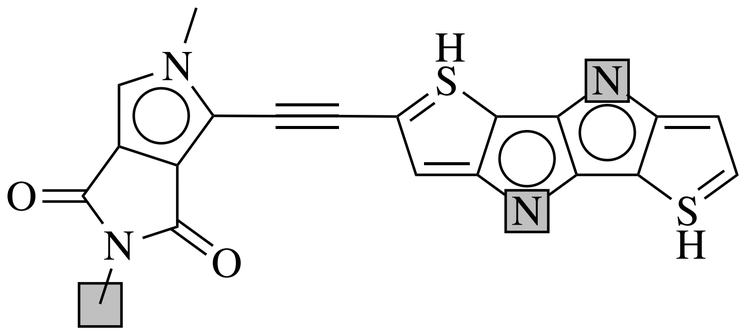


Generation Process

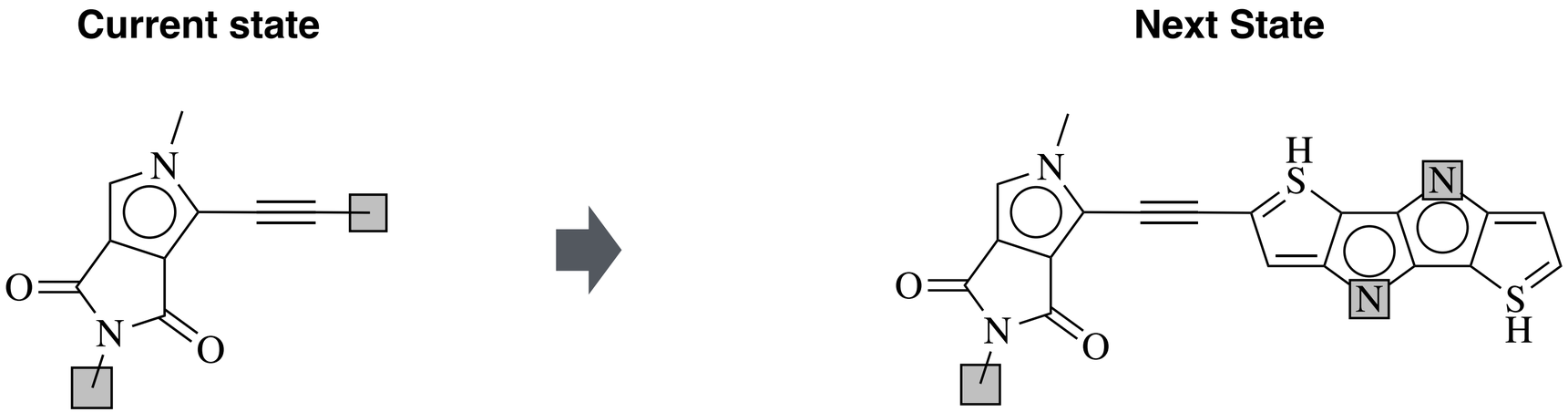
Current state



Next State



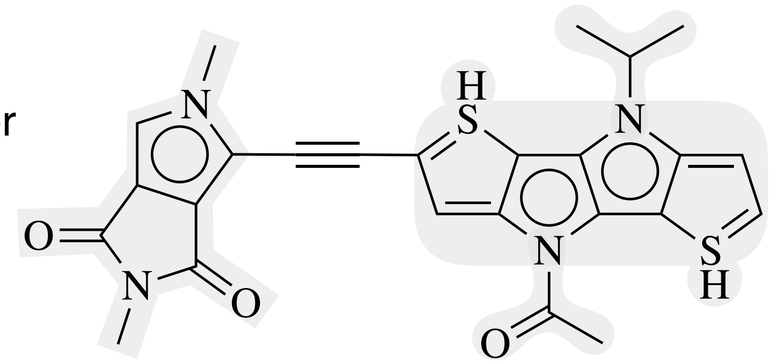
Generation Process



- JT-VAE assembles each neighborhood (multiple motifs) in one go.
- HierVAE decomposes the assembly process into multiple “baby steps”
  - First predict attaching points, then matching atoms.
  - Assembles one motif at a time, not the entire neighborhood.

Hierarchical Graph Encoder (bottom up)

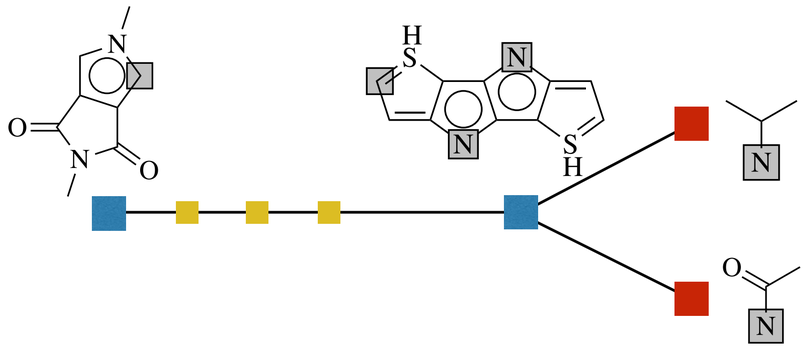
Atom Layer



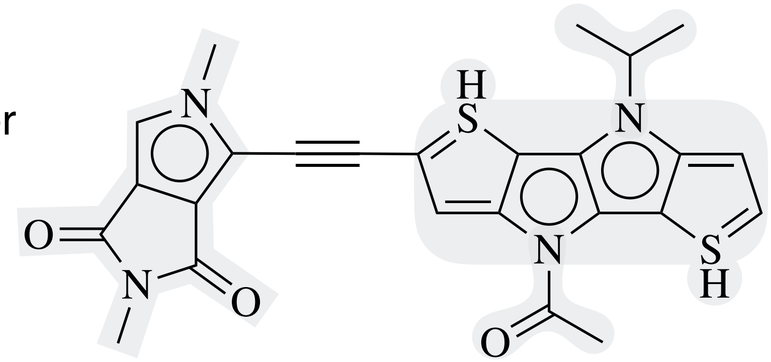
- Atom layer serves graph prediction (step 3)

Hierarchical Graph Encoder (bottom up)

Attachment Layer



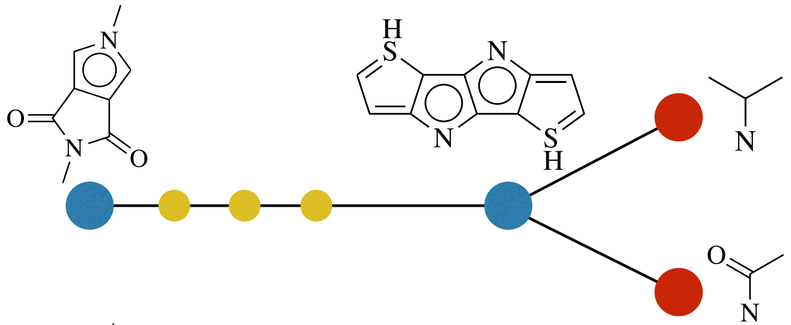
Atom Layer



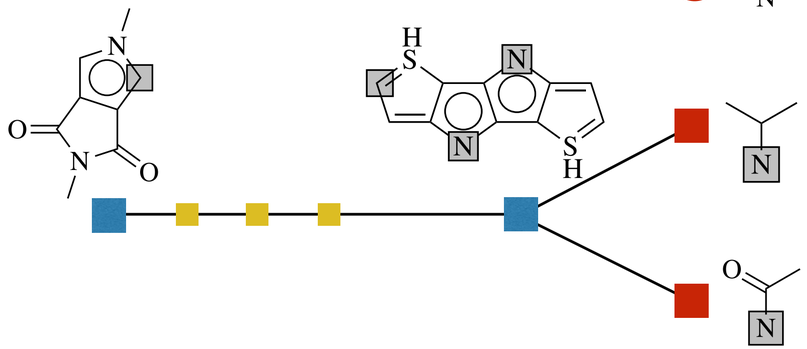
- Attachment layer serves attachment prediction (step 2)
- Atom layer serves graph prediction (step 3)

Hierarchical Graph Encoder (bottom up)

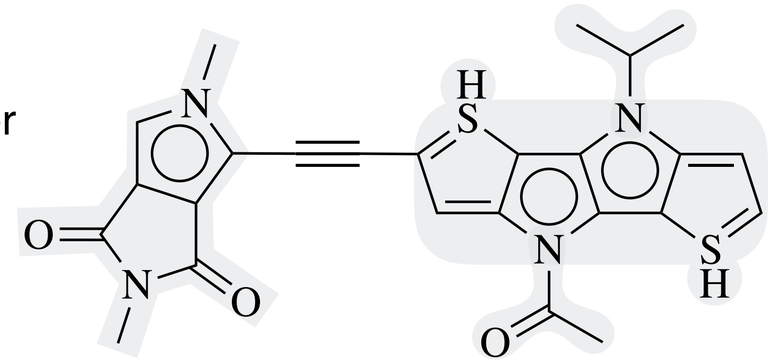
Motif Layer



Attachment Layer

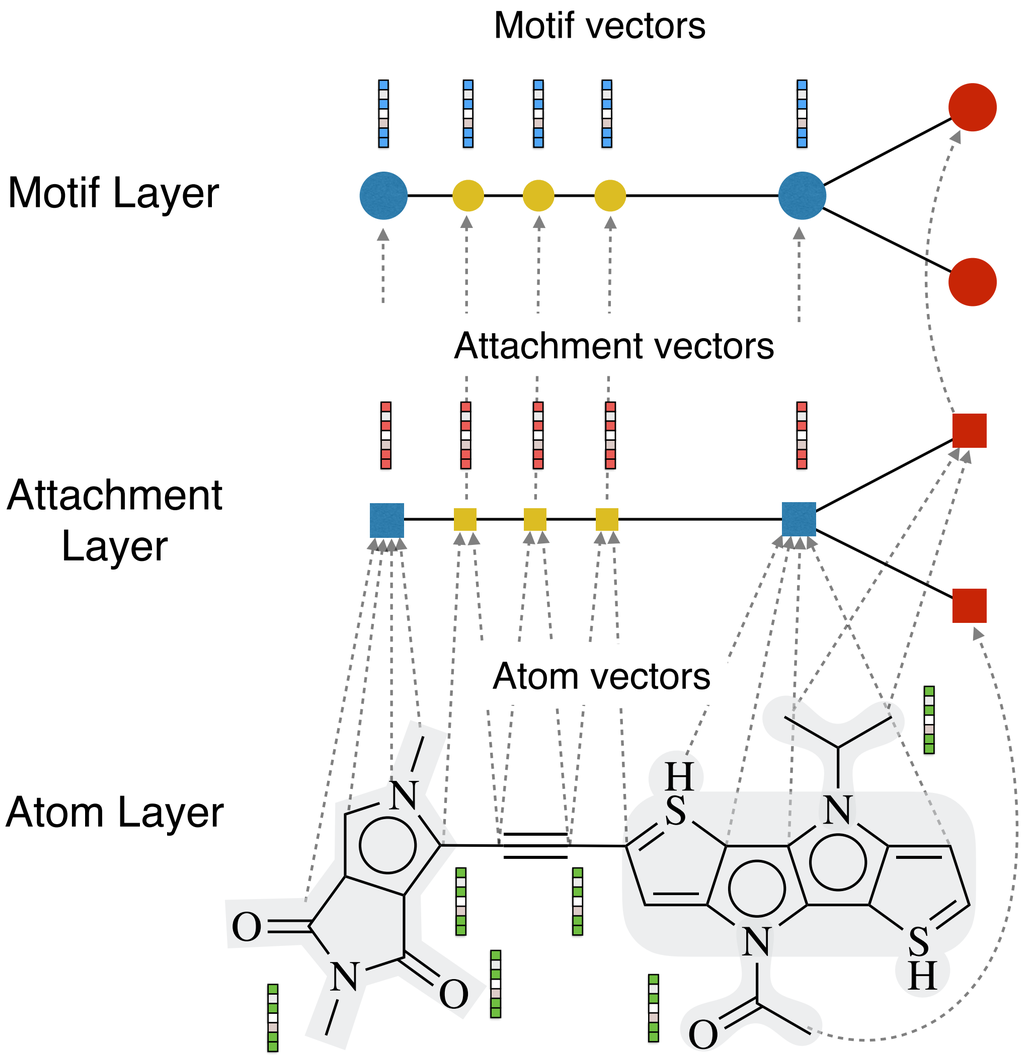


Atom Layer



- Motif layer designed for motif prediction (step 1)
- Attachment layer is designed for attachment prediction (step 2)
- Atom layer is designed for graph prediction (step 3)

Hierarchical Graph Encoder (bottom up)



- Run motif layer message passing network

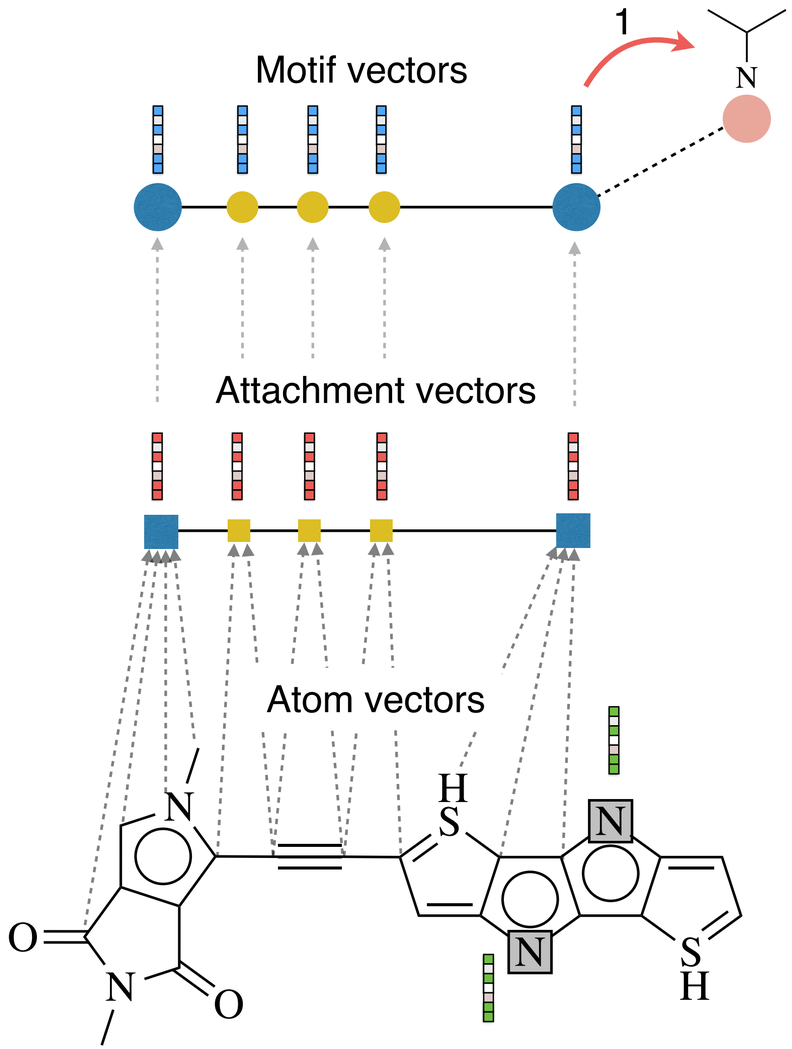
Propagate messages to corresponding nodes

- Run attachment layer message passing network

Propagate messages to corresponding nodes

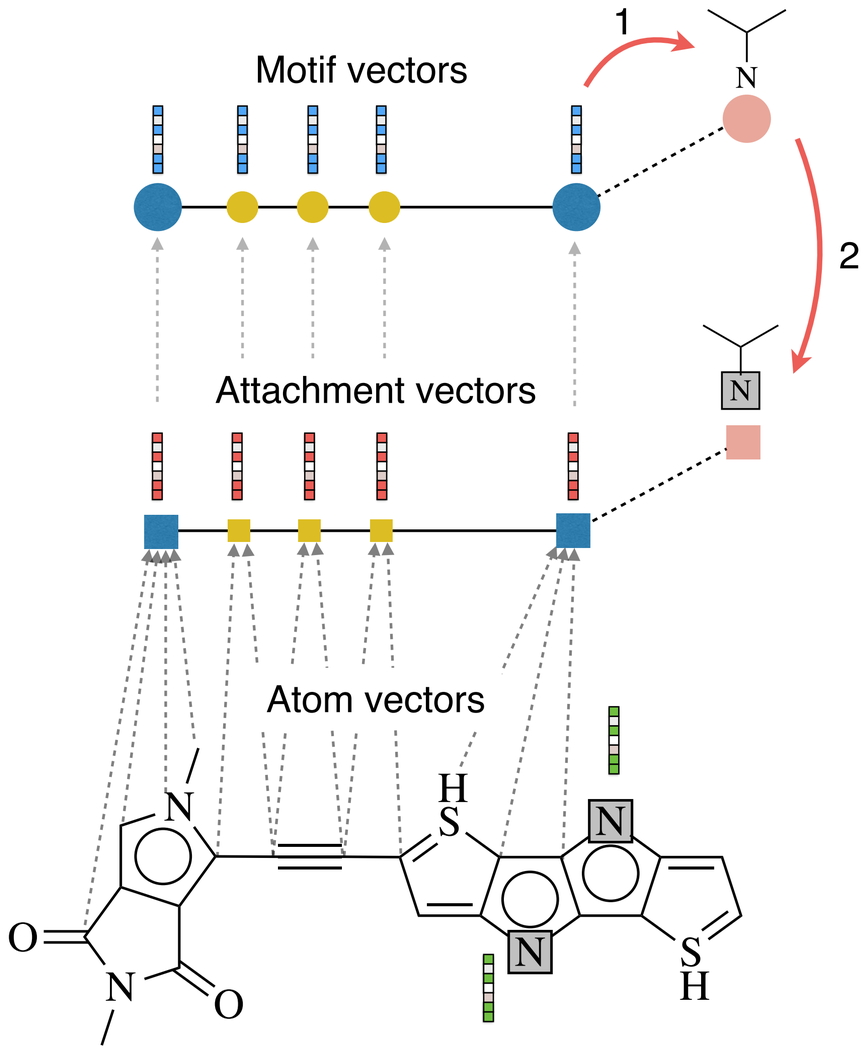
- Run atom layer message passing network

Hierarchical Graph Decoder (top down)



- Motif Prediction
  - Classification: predict the right motif in the vocabulary

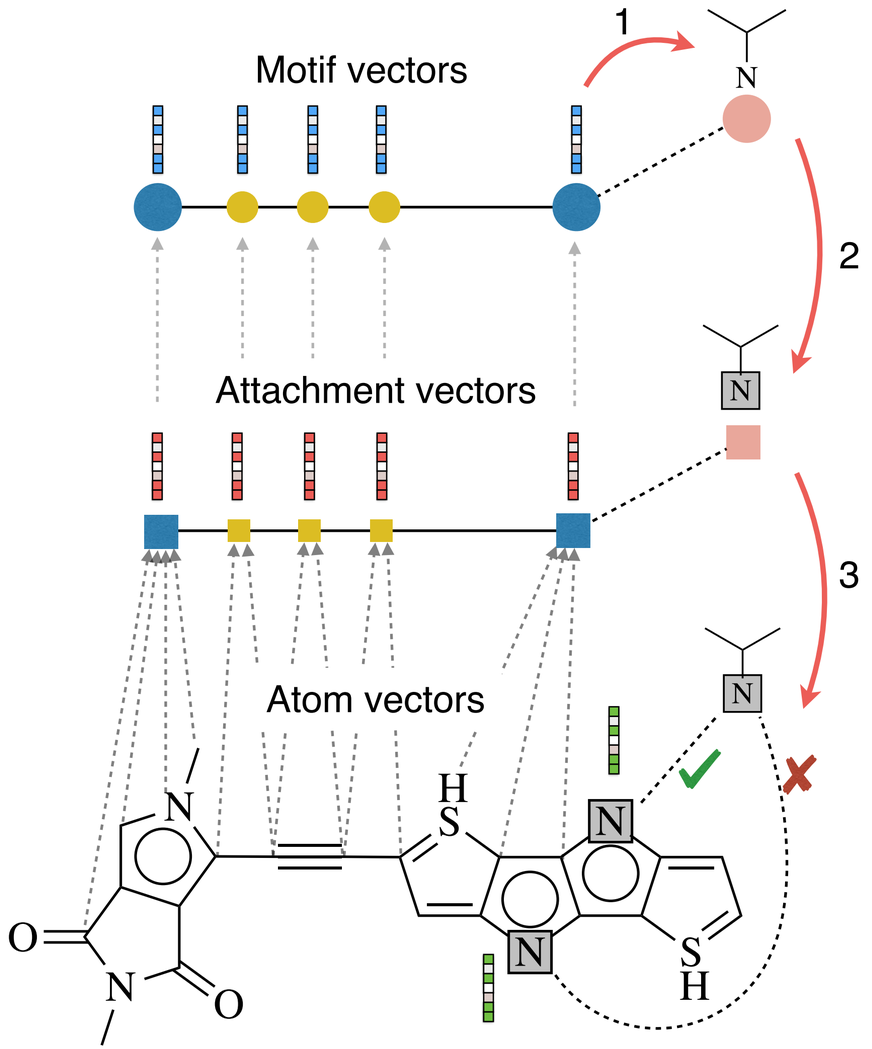
Hierarchical Graph Decoder (top down)



- Motif Prediction
  - Classification: predict the right motif in the vocabulary

- Attachment Prediction
  - Classification: predict the right attachment in the vocabulary

Hierarchical Graph Decoder (top down)



- Motif Prediction
  - Classification: predict the right motif in the vocabulary
- Attachment Prediction
  - Classification: predict the right attachment in the vocabulary
- Graph Prediction:
  - Classification: predict the corresponding matching atoms

Experiment 1: Polymer Generation

Dataset [1]: 86K polymers (76K training, 5K validation, 5K testing)

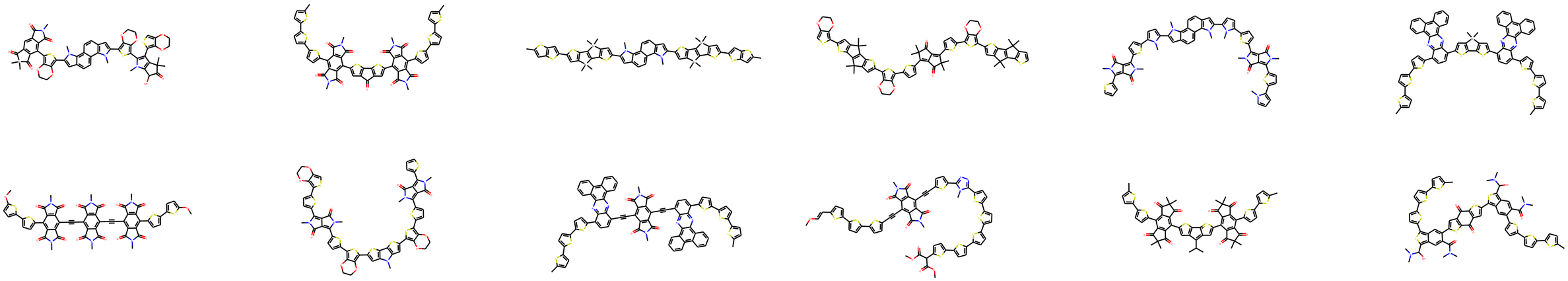
Evaluation Metrics: Sample 5000 molecules from models

- Reconstruction accuracy
- Validity
- Uniqueness
- Diversity
- Property statistics: Frechet distance between property distributions of molecules in the generated set and test set (logP, QED, SA, molecular weight).
- Structural statistics:
  - Nearest neighbor similarity (SNN)
  - Fragment similarity (Frag)
  - Scaffold similarity (Scaf)

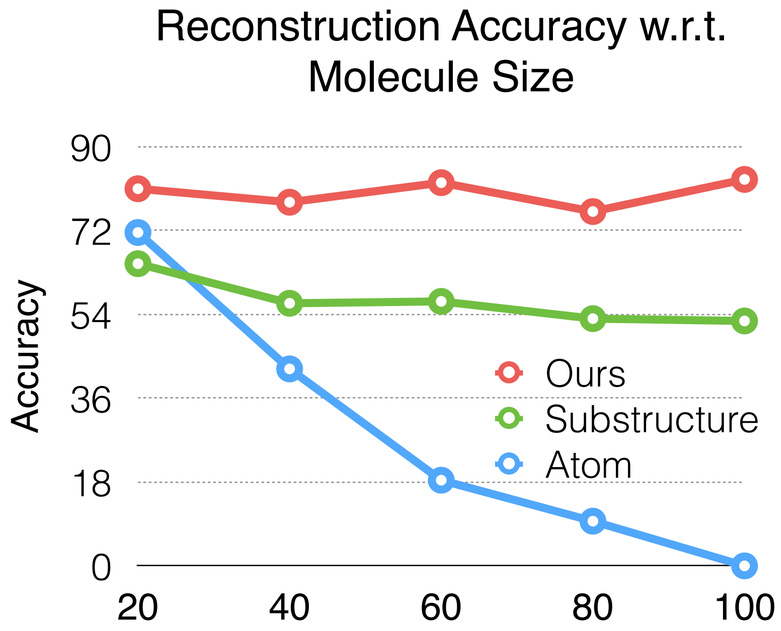
[1] St. John et al., “Message-passing neural networks for high-throughput polymer screening.” The Journal of chemical physics, 150 (23):234111, 2019

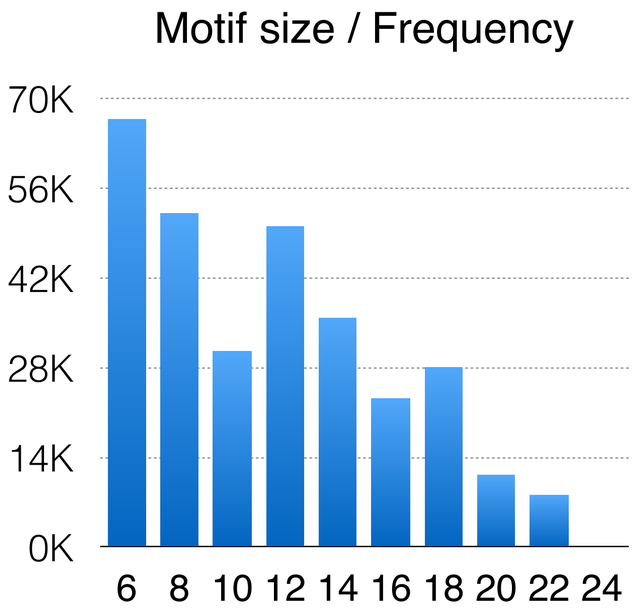
Experiment 1: Polymer Generation

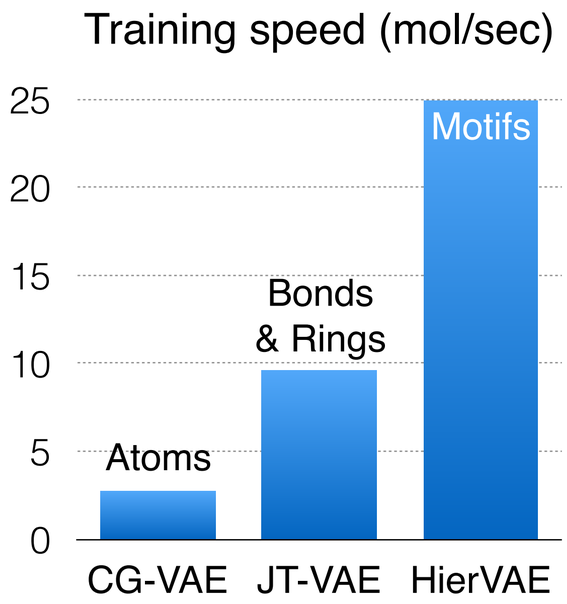
Method	Reconstruction / Sample Quality (↑)				Property Statistics (↓)				Structural Statistics (↑)		
	Recon.	Valid	Unique	Div.	logP	SA	QED	MW	SNN	Frag.	Scaf.
Real data	-	100%	100%	0.823	0.094	6.7e-5	1.7e-5	82.3	0.706	0.995	0.462
SMILES	21.5%	93.1%	**97.3%**	0.821	1.471	0.011	**5.4e-4**	4963	0.704	0.981	0.385
CG-VAE	42.4%	**100%**	96.2%	**0.879**	3.958	2.600	0.0030	3944	0.204	0.372	0.001
JT-VAE	58.5%	**100%**	94.1%	0.864	2.645	0.157	0.0075	10867	0.522	0.925	0.297
HierVAE	**79.9%**	**100%**	97.0%	0.817	**0.525**	**0.007**	5.7e-4	**1928**	**0.708**	**0.984**	**0.390**



Experiment 1: Polymer Generation







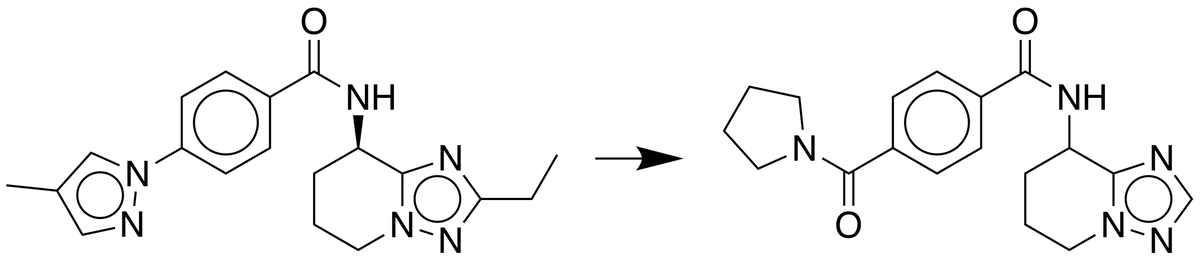
Experiment 2: Lead optimization

- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications (first introduced in Jin et al., 2019)

Learning Multimodal Graph-to-Graph Translation for Molecular Optimization, W. Jin, R. Barzilay, T. Jaakkola, ICLR 2019

Experiment 2: Lead optimization

- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications (first introduced in Jin et al., 2019)



Source Molecule (QED=0.784)

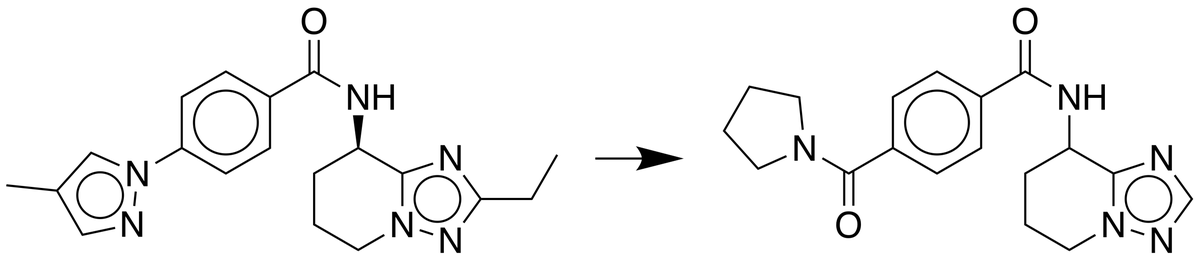
QED=0.924

- Similar but …
- Better drug-likeness

Learning Multimodal Graph-to-Graph Translation for Molecular Optimization, W. Jin, R. Barzilay, T. Jaakkola, ICLR 2019

Experiment 2: Lead optimization

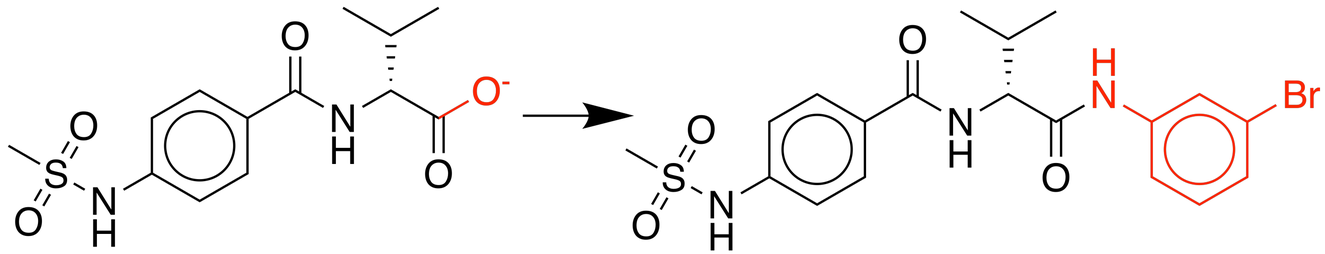
- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications (first introduced in Jin et al., 2019)



Source Molecule (QED=0.784)

QED=0.924

- Similar but …
- Better drug-likeness

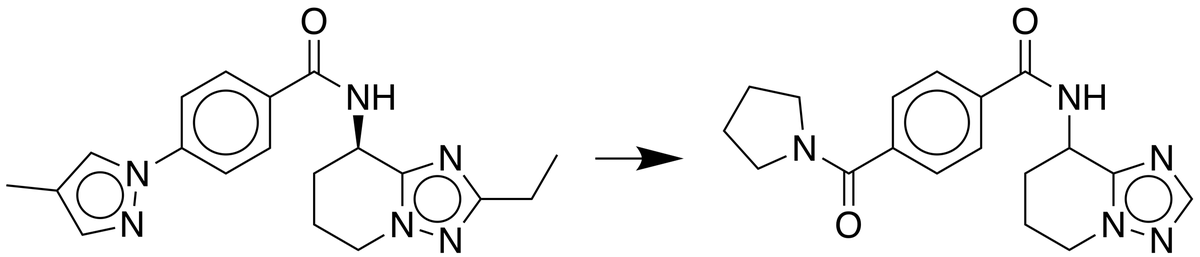


- Similar but …
- Better solubility

Learning Multimodal Graph-to-Graph Translation for Molecular Optimization, W. Jin, R. Barzilay, T. Jaakkola, ICLR 2019

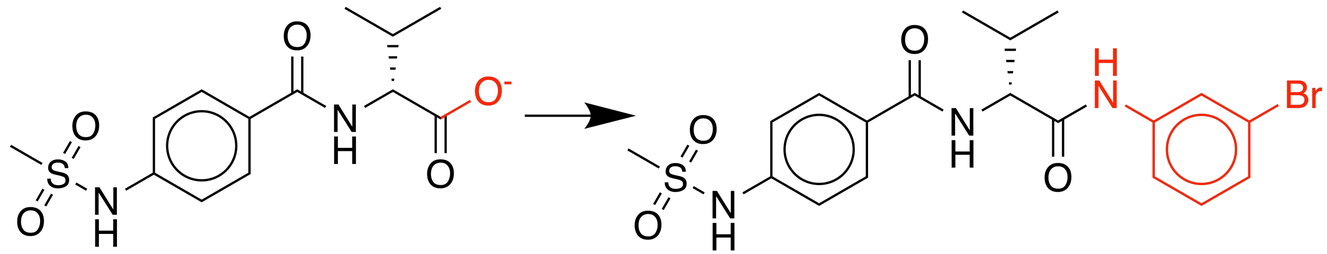
Experiment 2: Lead optimization

- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications (first introduced in Jin et al., 2019)



Source Molecule (QED=0.784) QED=0.924

- Similar but …
- Better drug-likeness



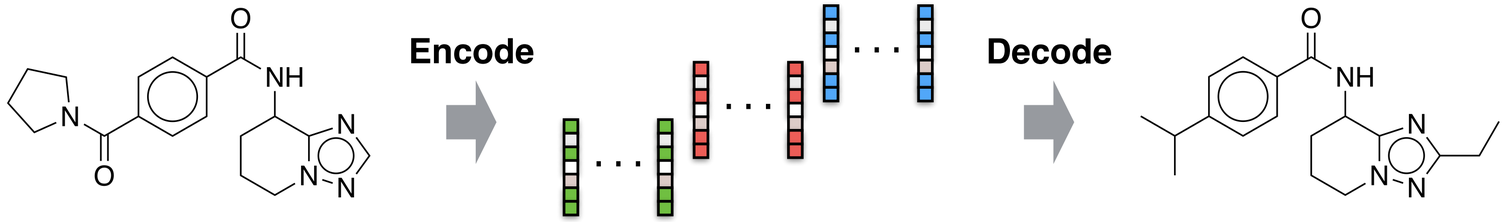
- Similar but …
- Better solubility

- <u>Need to learn a molecule-to-molecule mapping (i.e., graph-to-graph)</u>

Learning Multimodal Graph-to-Graph Translation for Molecular Optimization, W. Jin, R. Barzilay, T. Jaakkola, ICLR 2019

Lead optimization as Graph Translation

- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications (first introduced in Jin et al., 2019)

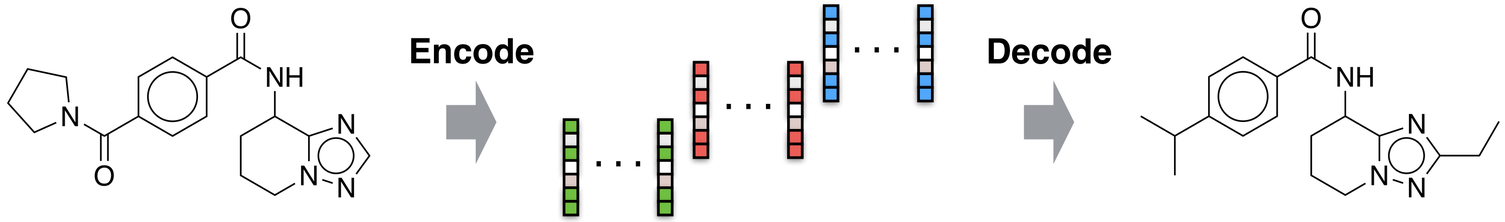
Source X 

Y Target

Learning Multimodal Graph-to-Graph Translation for Molecular Optimization, W. Jin, R. Barzilay, T. Jaakkola, ICLR 2019

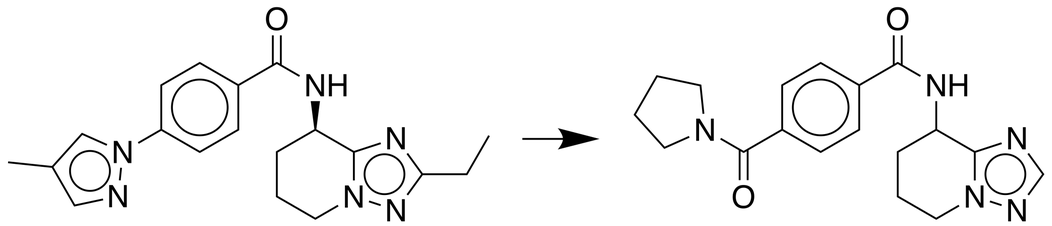
Lead optimization as Graph Translation

- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications (first introduced in Jin et al., 2019)

Source X 

Y Target

- The training set consists of (source, target) molecular pairs, e.g.,

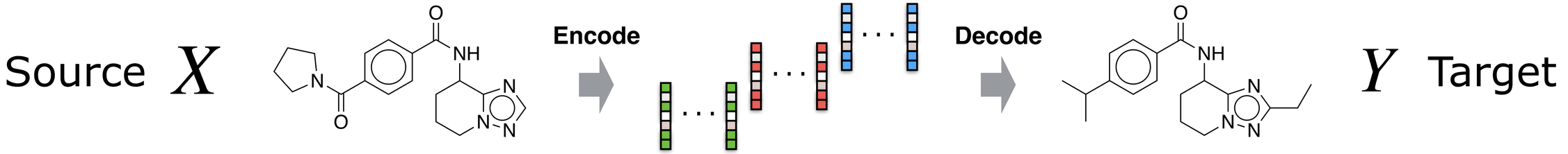
Source  Target

... ...

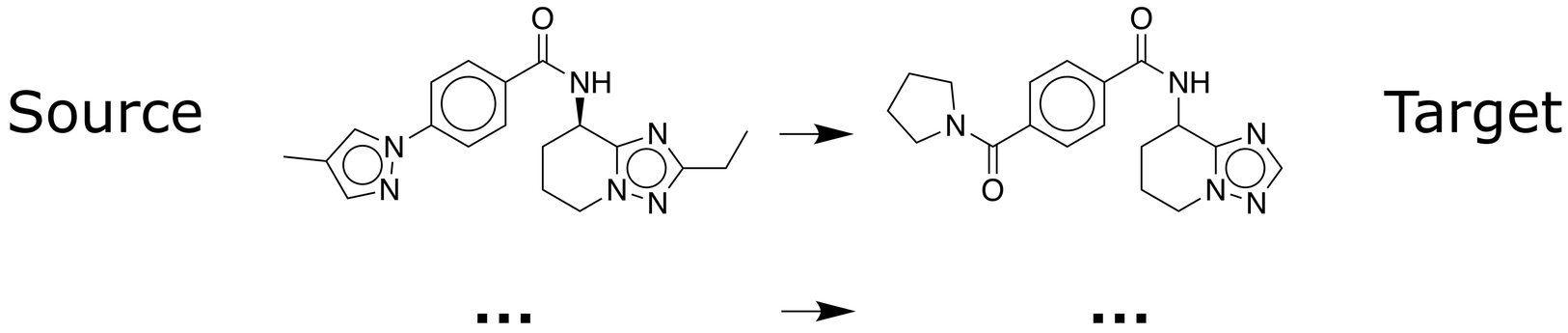
Learning Multimodal Graph-to-Graph Translation for Molecular Optimization, W. Jin, R. Barzilay, T. Jaakkola, ICLR 2019

Lead optimization as Graph Translation

- **Goal:** We aim to transform given molecules into molecules that satisfy given design specifications



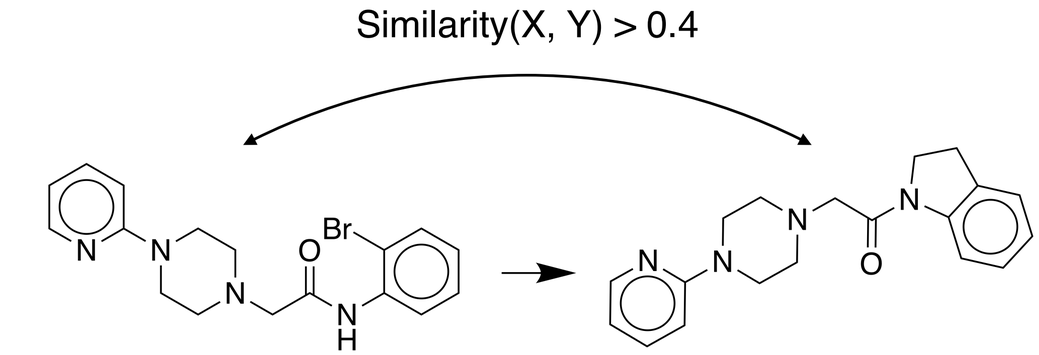
- The training set consists of (source, target) molecular pairs, e.g.,

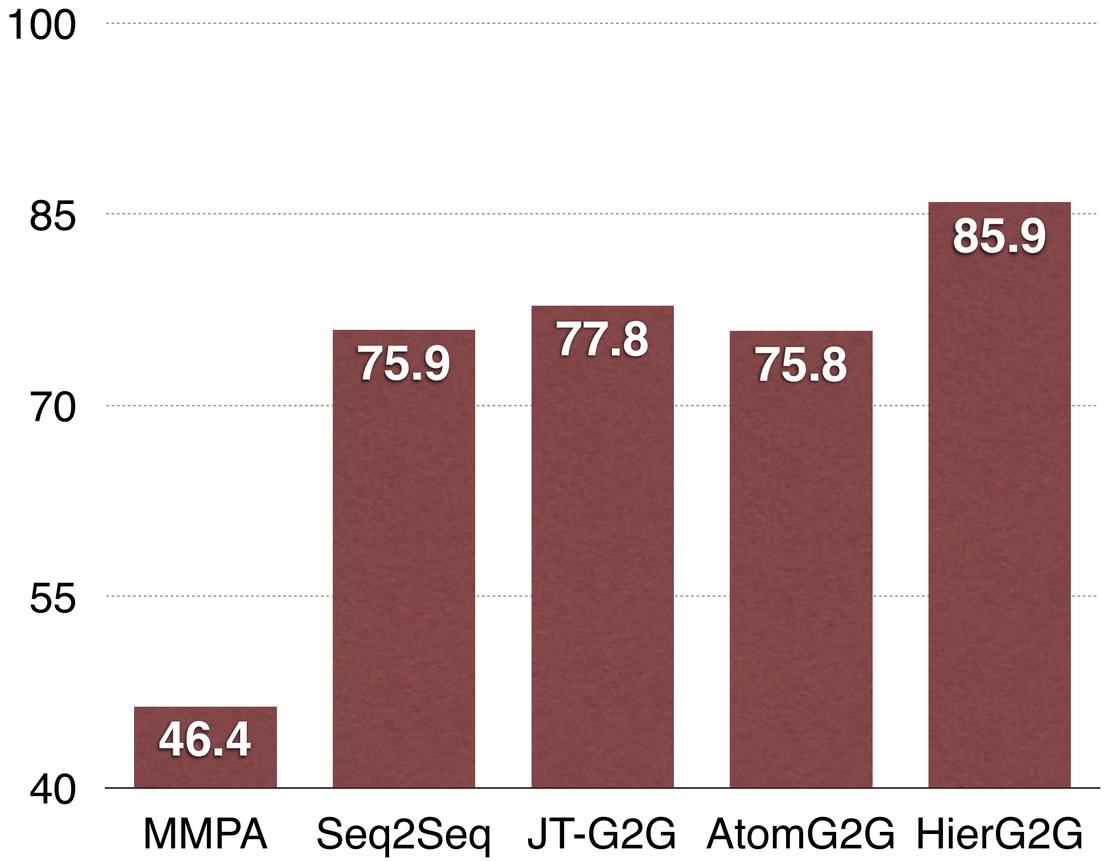


- Easy to modify HierVAE into a translation model (just add attention layers)

DRD2 Optimization

- Single property optimization: DRD2 success % (from inactive to active)



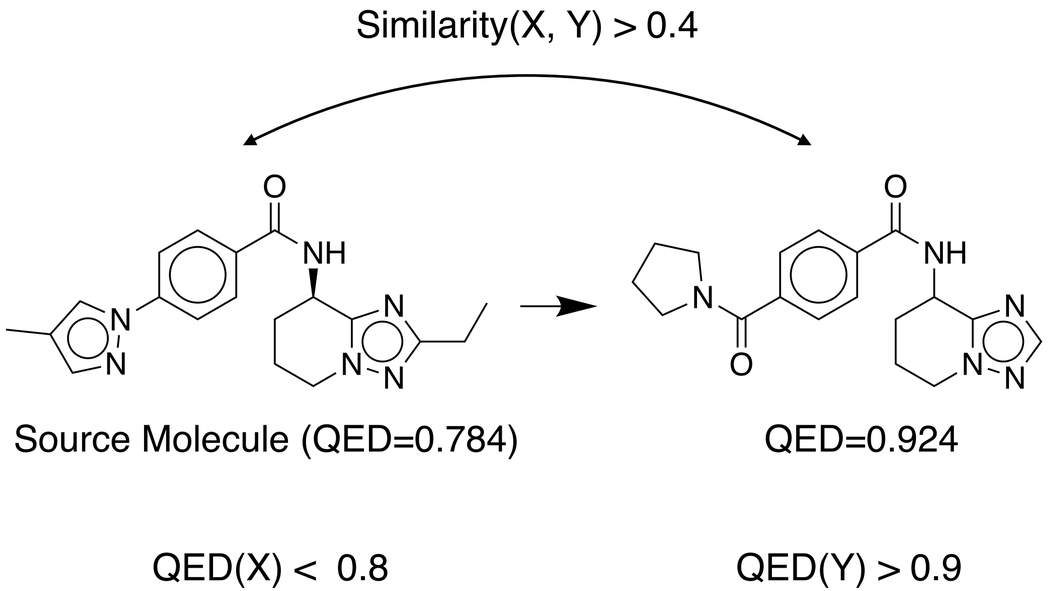


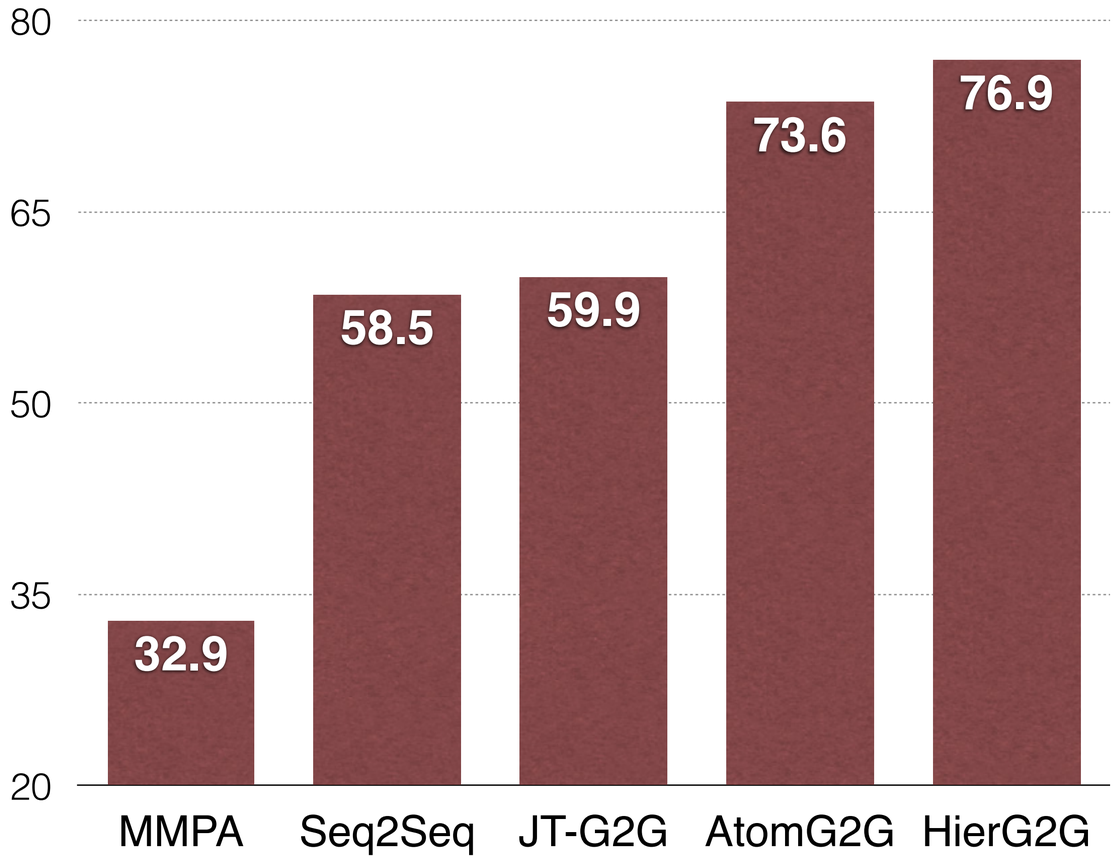
- We use a property predictor [1] to evaluate DRD2 activity of generated compounds

[1] Olivecrona et al., Molecular de-novo design through deep reinforcement learning, J. Chem. Inf. Model. 2017

QED Optimization

- Single property optimization: drug-likeness (QED) success %





- QED is computed by RDKit

Summary

- Molecular graph generation is an important problem for ML and drug discovery
- In this paper, we proposed HierVAE to generate molecules motif by motif.
- HierVAE works better than previous methods, both in large molecules (polymers) as well as small molecules (graph translation).
- Since motifs structures are flexible, how should we construct a good motif vocabulary?
  - Jin et al., Multi-objective molecule generation using interpretable substructures. ICML 2020
  - Use interpretability techniques to construct a motif vocabulary relevant for downstream task (poster ID 2748)