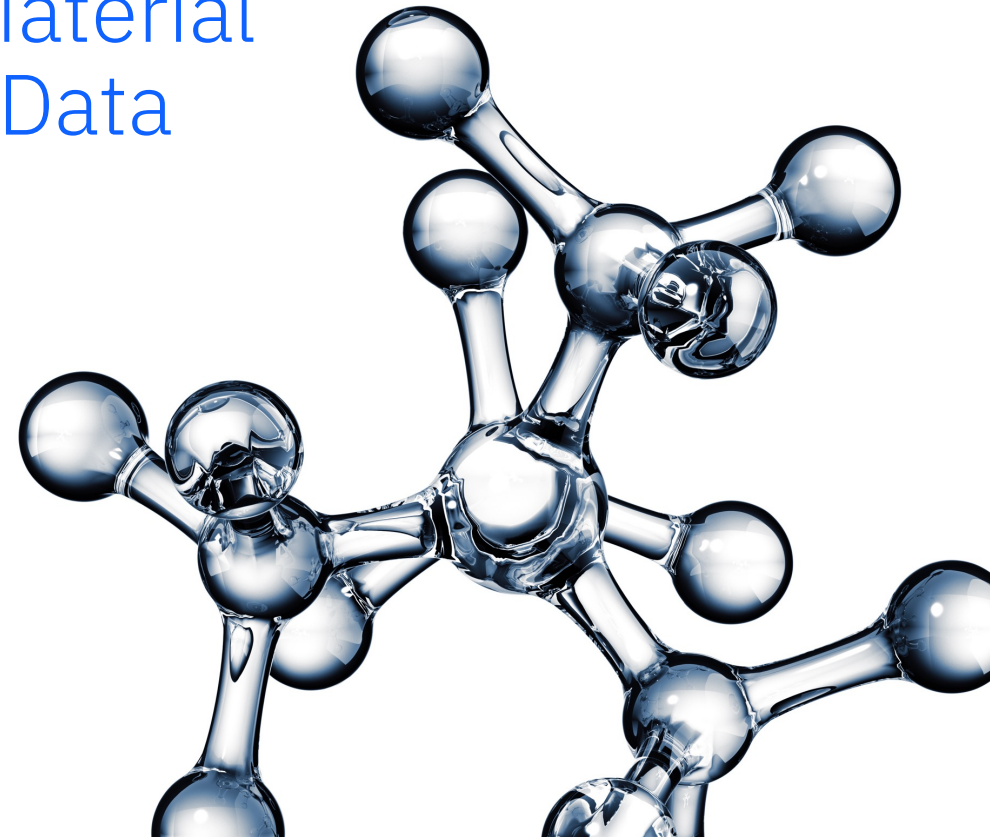
What's Next Seminar Series

Automating Drug and Material Discovery with Limited Data



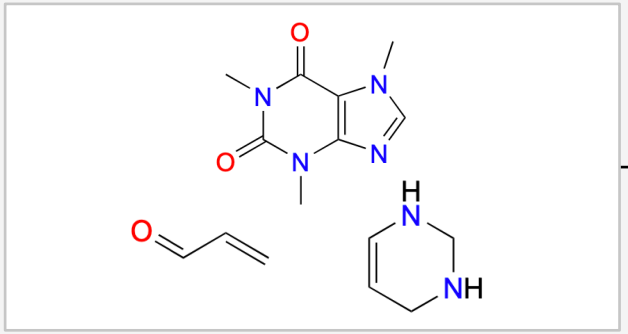
Jie Chen
MIT-IBM Watson AI Lab, IBM Research

IBM

Introduction

- Molecular generation is a key step in drug design and material discovery
- Deep learning-based generative models are quite effective, but data-hungry
- We propose a data-efficient generation method (E.g., requiring 10~100 training examples, as opposed to 81k in deep learning)
- In this method, we treat molecules as graphs and learn a grammar that generates them

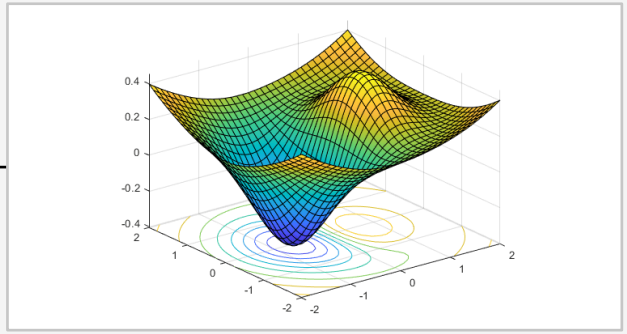
Graph-based design for drugs/materials

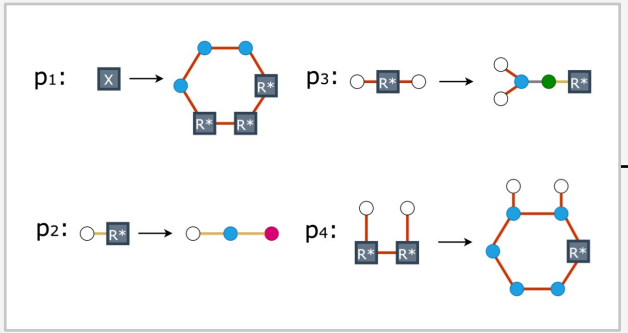


New, better data

Data as graph

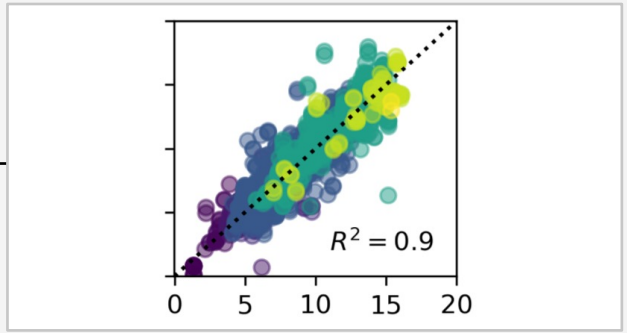
Property optimization



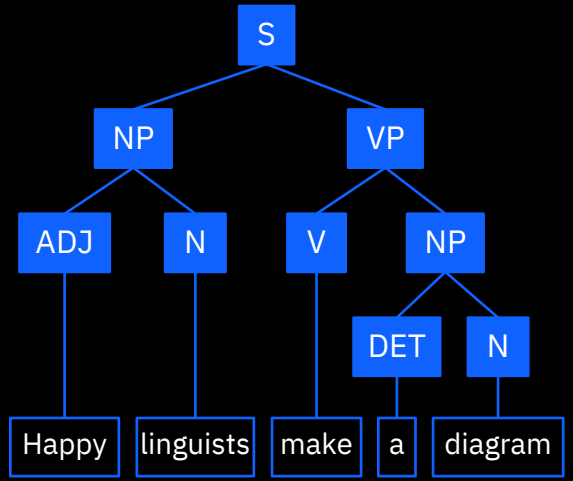


Graph grammar

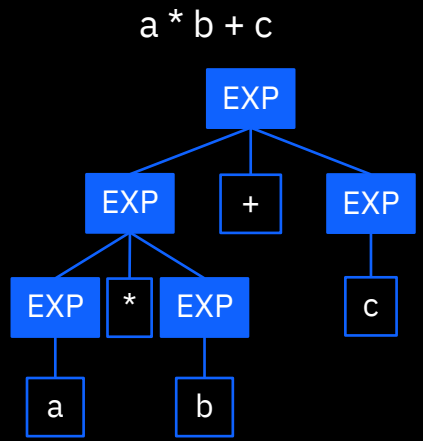
Property prediction



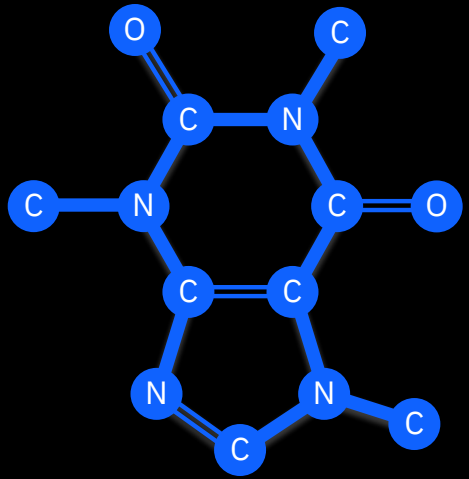
Grammars



English

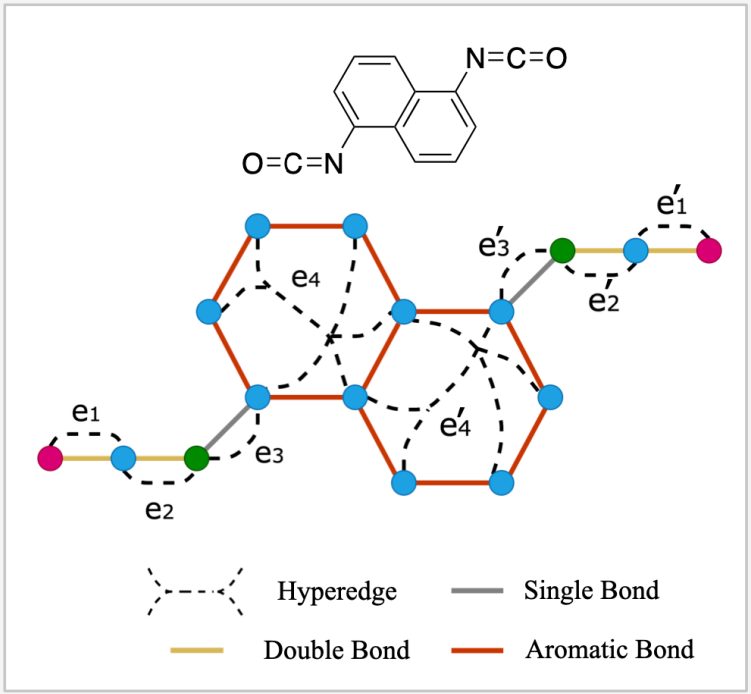


Programs

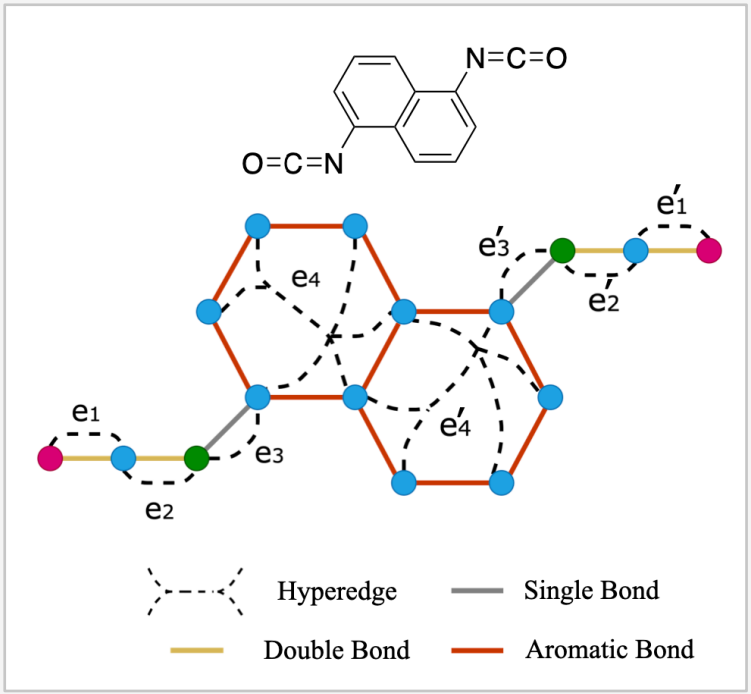


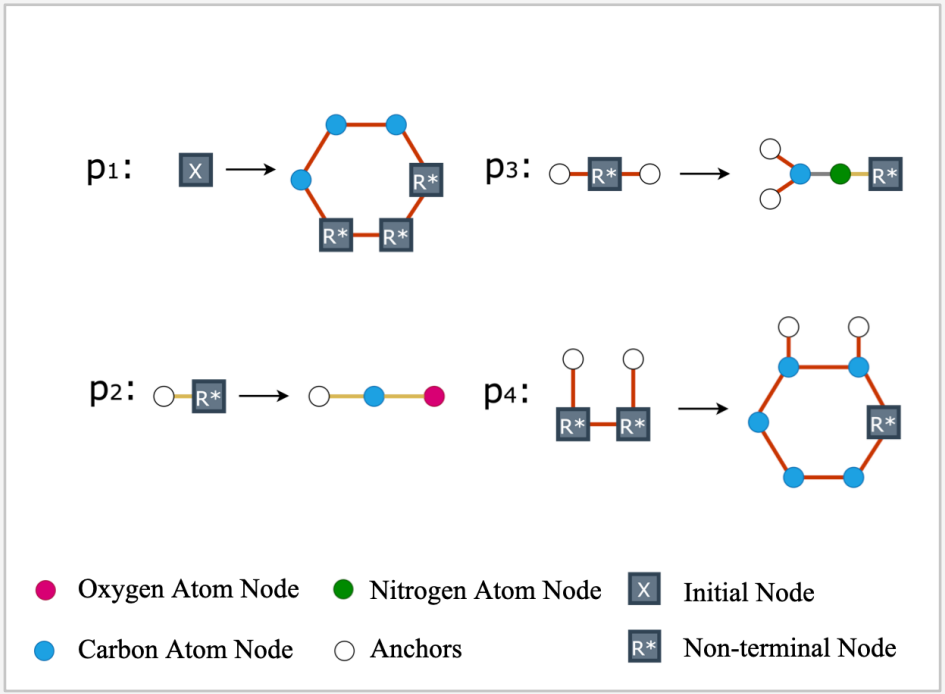
Graphs also have a grammar

Molecular graphs

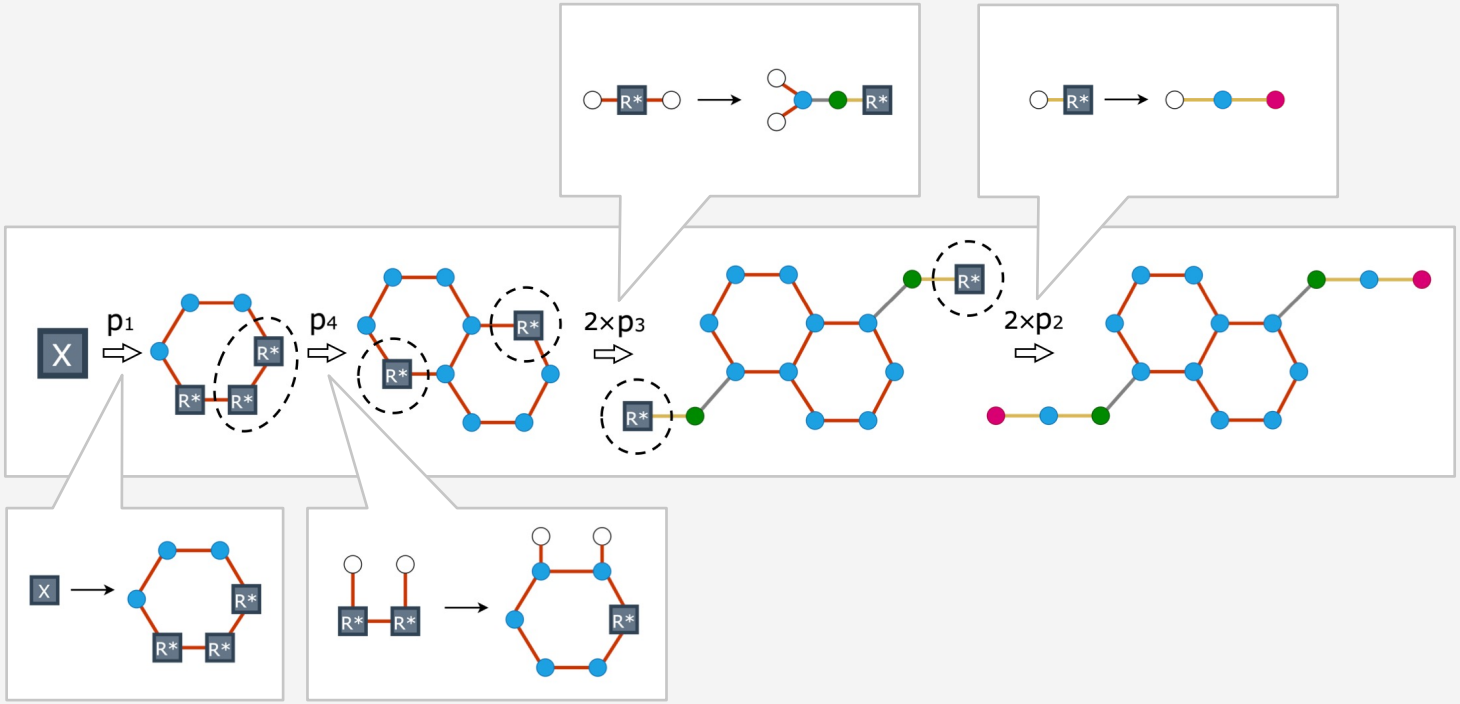


Molecular graphs and graph grammar

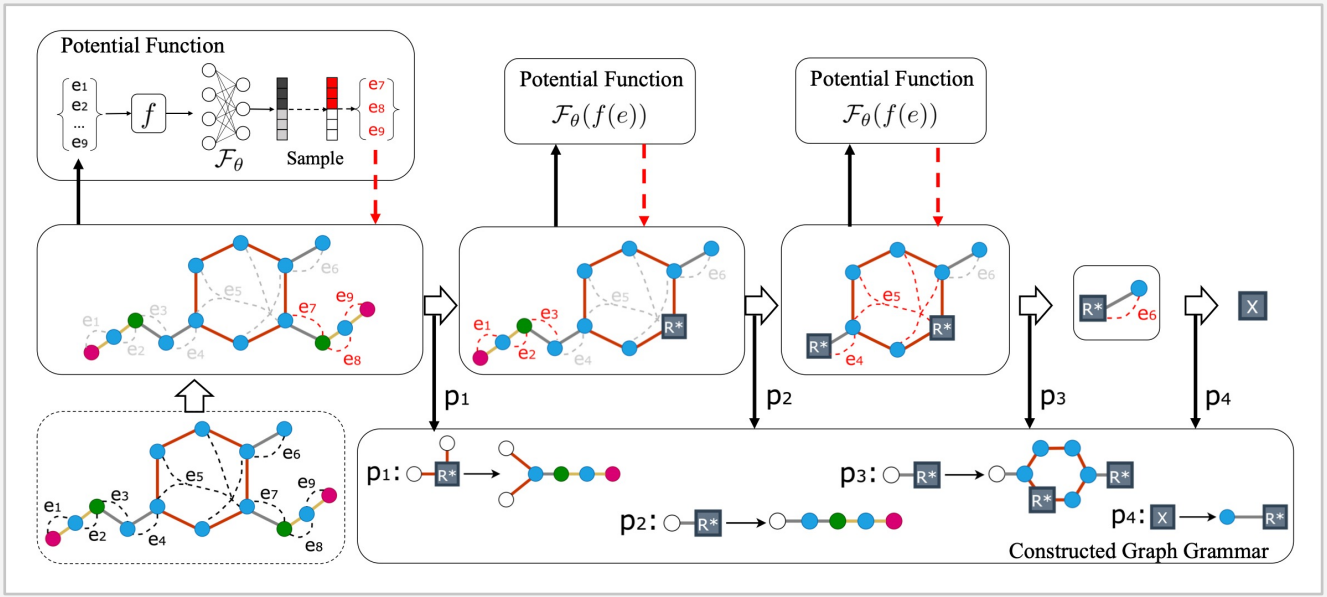




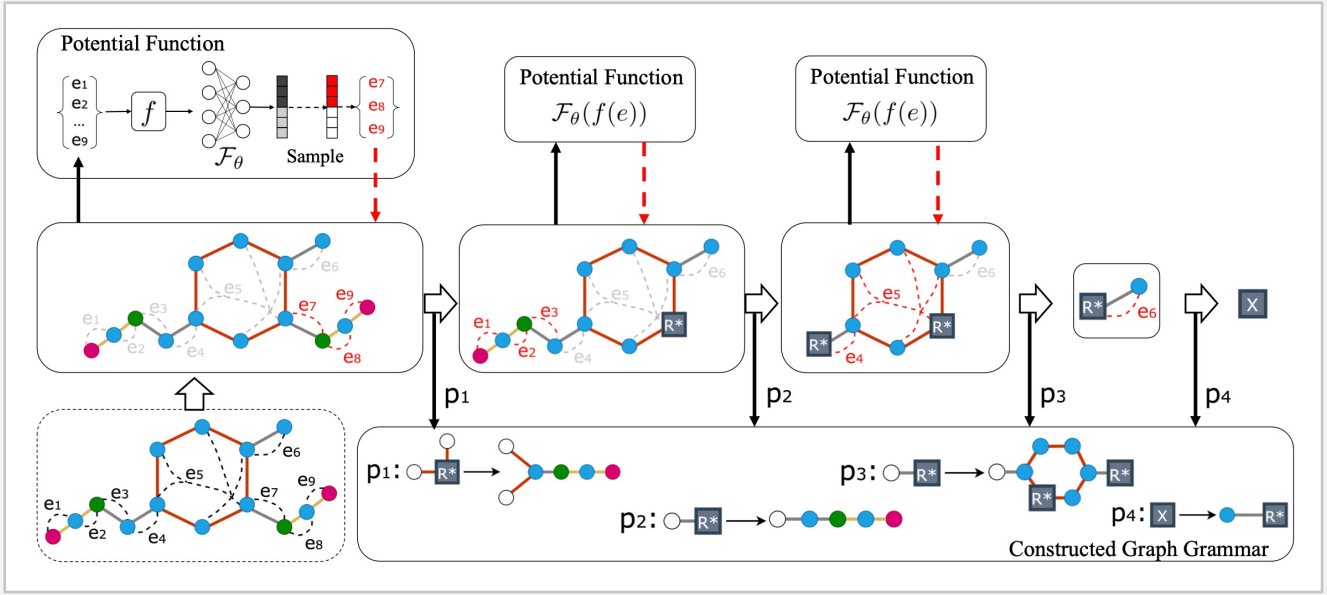
Graph generation using grammar



How can we get a grammar?



How can we get an optimal grammar?



There are many ways to remove a component from a graph and make a production rule.

We parameterize the selection of edges and learn the model parameters by optimizing metrics of interest:

- molecule diversity
- synthesizability

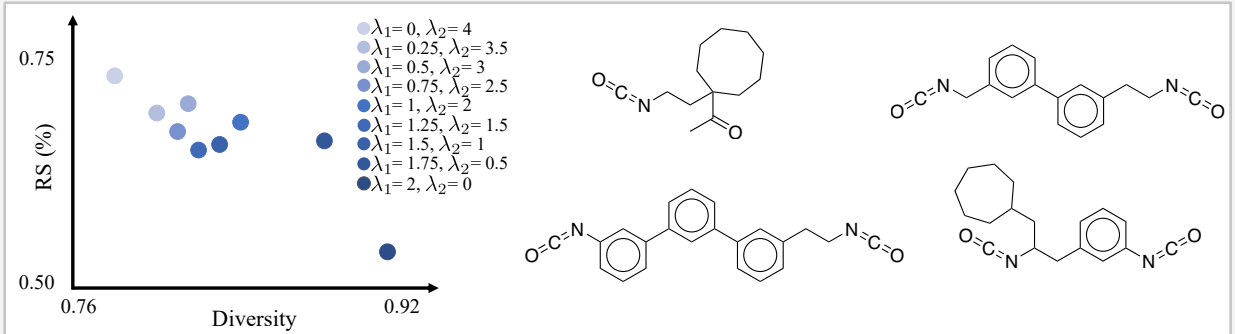
Experimental results compared with state-of-the-art methods

Method	Valid	Unique	Div.	Chamfer	RS	Memb.
Train data	100%	100%	0.61	0.00	100%	100%
GraphNVP	0.16%	-	-	-	0.00%	0.00%
JT-VAE	100%	5.8%	0.72	0.85	5.50%	66.5%
NierVAE	100%	99.6%	0.83	0.76	1.85%	0.05%
MHG	100%	75.9%	0.88	0.83	2.97%	12.1%
STONED	100%	100%	0.85	0.86	5.63%	79.8%
DEG	100%	100%	0.86	0.87	27.2%	96.3%
		Sample quality			Synthesizability	Class

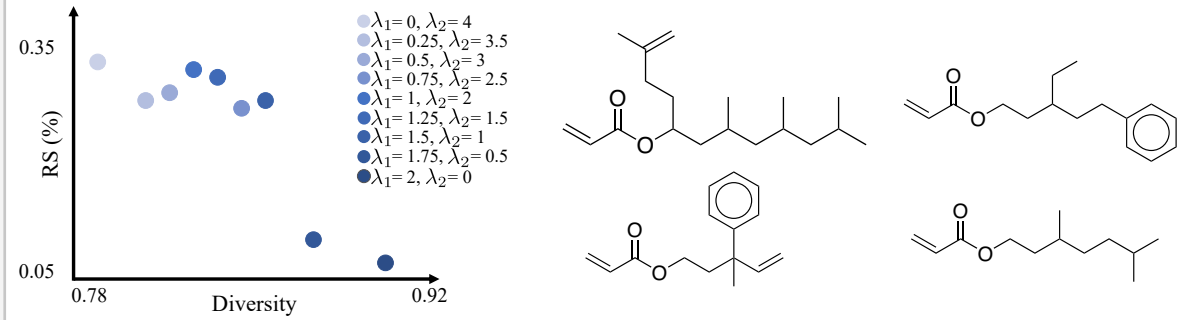
These are results for isocyanates (11 training examples)

Results for acrylates and chain extenders (not shown) conclude similar findings

Examples of generated molecules demonstrating a variety of structures

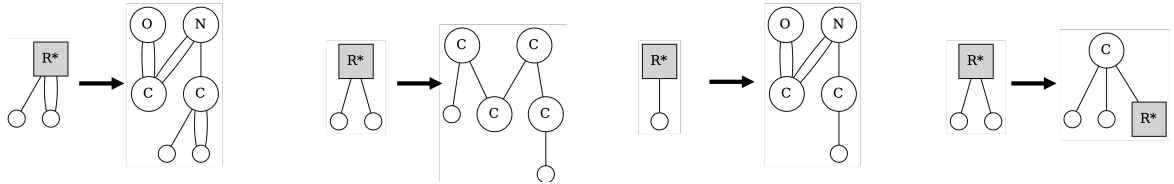


(a) Analysis on Isocyanate dataset

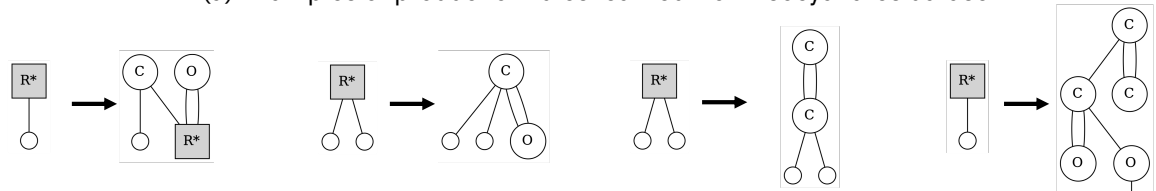


(b) Analysis on Acrylate dataset

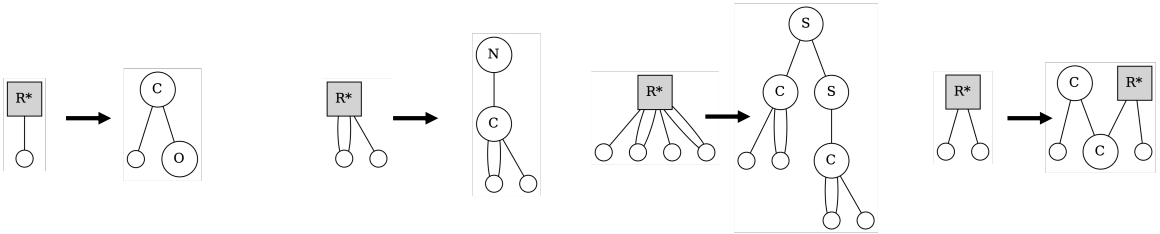
Examples of generated rules



(a) Examples of production rules learned from Isocyanates dataset



(b) Examples of production rules learned from Acrylates dataset



(b) Examples of production rules learned from Chain Extenders dataset

Summary

- We have presented a molecular generation method by using limited training data
- In this method, we treat molecules as graphs and learn a grammar that generates them
- The automated generation will significantly speed up the pipeline of identifying better molecules, used in drugs and materials

- Paper: Guo et al. Data-Efficient Graph Grammar Learning for Molecular Generation. ICLR, 2022