



Towards computationally efficient graph generative models for molecular search

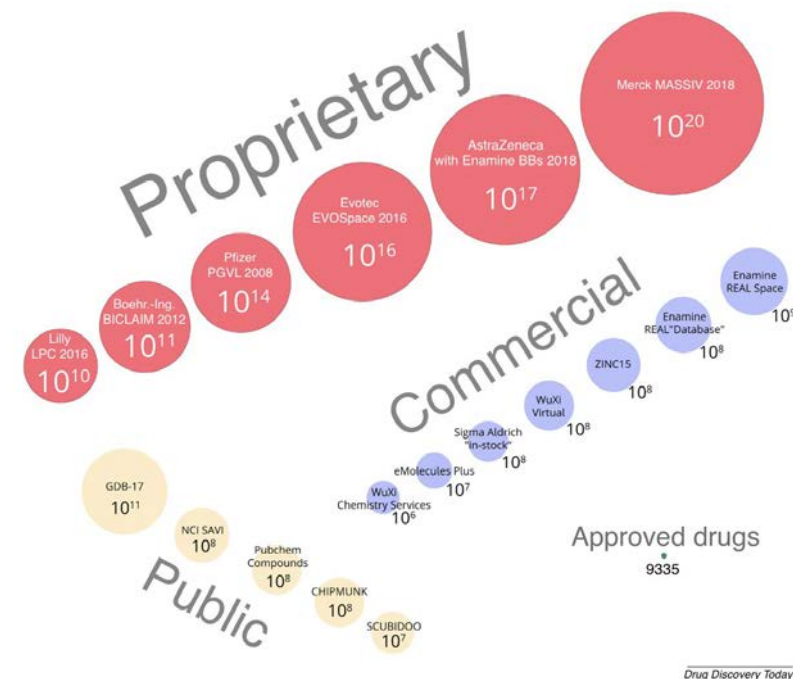


**Aryan Pedawi, Paweł Gniewek, Chaoyi Chang,
Brandon M. Anderson, Henry van den Bedem**
Atomwise Inc.

Motivation

Search in the non-enumerative regime

- Exponential growth in the size of make-on-demand combinatorial synthesis libraries
- Broad interest in search and search-adjacent problems applied to such libraries
 - Analogue/similarity search
 - Goal-directed search
 - Virtual screening
- Conventional enumerative approaches to these problems face significant limitations with the size of recent combinatorial synthesis libraries

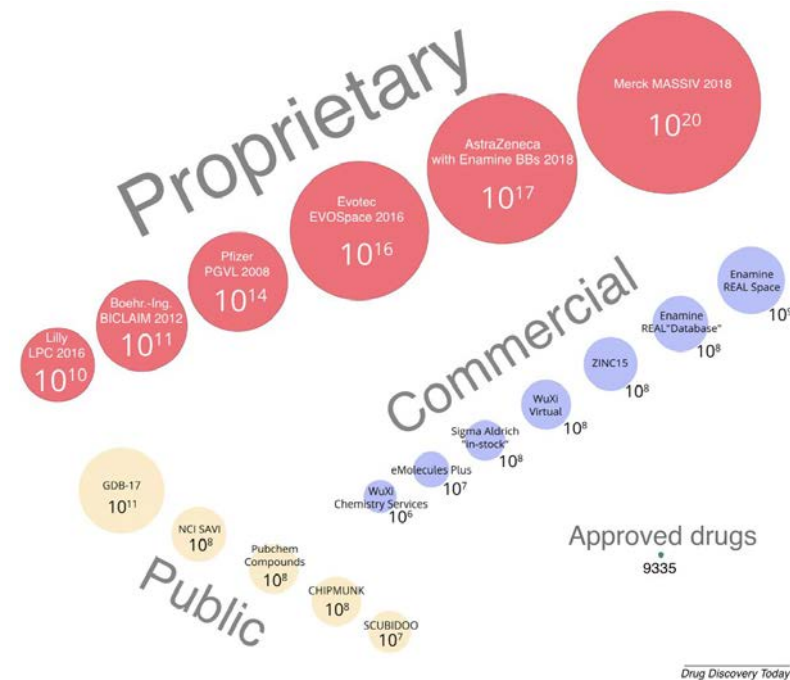


Hoffmann, Torsten, and Marcus Gastreich. "The next level in chemical space navigation: going far beyond enumerable compound libraries." *Drug Discovery Today* 24.5 (2019): 1148-1156.

Motivation

Search in the non-enumerative regime

- Exponential growth in the size of make-on-demand combinatorial synthesis libraries
- Broad interest in search and search-adjacent problems applied to such libraries
 - Analogue/similarity search
 - Goal-directed search
 - Virtual screening
- Conventional enumerative approaches to these problems face significant limitations with the size of recent combinatorial synthesis libraries



Hoffmann, Torsten, and Marcus Gastreich. "The next level in chemical space navigation: going far beyond enumerable compound libraries." *Drug Discovery Today* 24.5 (2019): 1148-1156.

In this work, we develop a new graph generative model specifically tailored for the navigation of ultra-large combinatorial synthesis libraries

Molecular generative models

Molecular generative models

- Lots of attention in developing and applying deep learning-based generative models to molecular datasets
- In goal-directed applications, such methods have shown promise in generating novel compounds with optimized properties
- Two dominant paradigms: string-based and graph-based generative models
 - String-based: SMILES-RNN, DeepSMILES, REINVENT, ...
 - Graph-based: CG-VAE, JT-VAE, RationaleRL, GraphINVENT, ...

Gómez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." *ACS Central Science* 4.2 (2018): 268-276.

O'Boyle, Noel, and Andrew Dalke. "DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures." (2018).

Blaschke, Thomas, et al. "REINVENT 2.0: an AI tool for de novo drug design." *Journal of Chemical Information and Modeling* 60.12 (2020): 5918-5922.

Liu, Qi, et al. "Constrained graph variational autoencoders for molecule design." *Advances in Neural Information Processing Systems* 31 (2018).

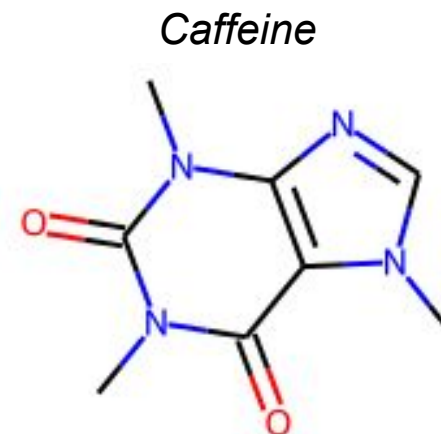
Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." *International Conference on Machine Learning*. PMLR, 2018.

Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Multi-objective molecule generation using interpretable substructures." *International Conference on Machine Learning*. PMLR, 2020.

Mercado, Rocío, et al. "Graph networks for molecular design." *Machine Learning: Science and Technology* 2.2 (2021): 025023.

Autoregression

- In both paradigms, the workhorse has been autoregression
 - Molecules are grown one token (string-based) or atom/bond (graph-based) at a time, until an “end” token is reached
- Autoregression has limitations
 - Canonicalization
 - Ensuring validity
 - Poor scaling when tasked with generating large molecules



1. Cn1c(=O)c2c(ncn2C)n(C)c1=O
2. n1cn(C)c2c(=O)n(c(n(C)c21)=O)C
3. Cn1c2ncn(C)c2c(n(C)c1=O)=O
4. c1n(C)c2c(n(c(=O)n(c2n1)C)C)=O
5. Cn1c(=O)c2c(ncn2C)n(c1=O)C
- ...

The in-library constraint

- Many early stage virtual screening pipelines seek to limit exploration to compounds that can be ordered from a make-on-demand catalog
- Existing generative models face significant challenges with satisfying this hard constraint
 - Projecting back to the library via analogue enumeration has limitations, especially with the increasingly larger size of such libraries

The in-library constraint

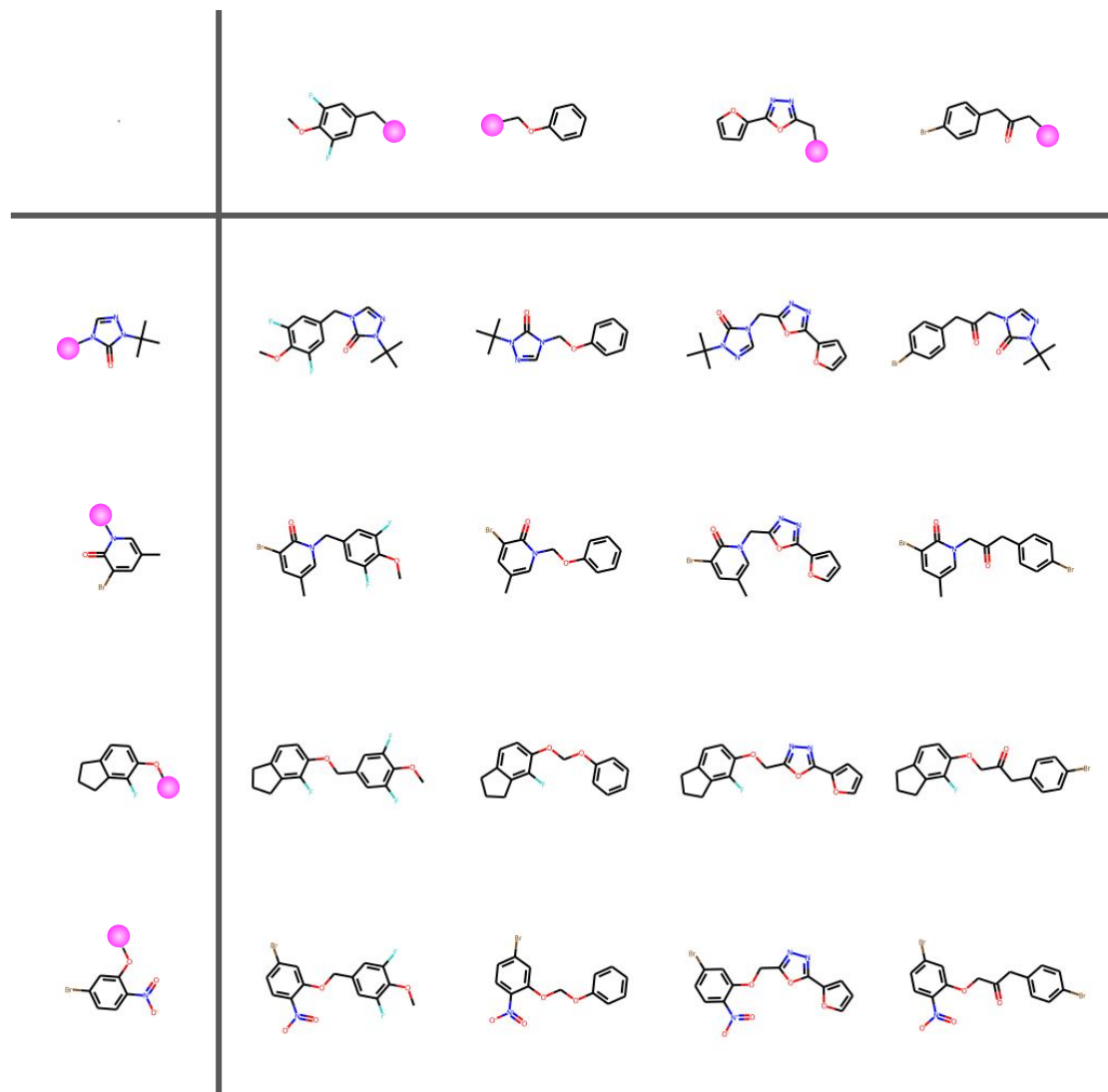
- Many early stage virtual screening pipelines seek to limit exploration to compounds that can be ordered from a make-on-demand catalog
- Existing generative models face significant challenges with satisfying this hard constraint
 - Projecting back to the library via analogue enumeration has limitations, especially with the increasingly larger size of such libraries

We utilize the structure of combinatorial synthesis libraries to develop a graph generative model that is guaranteed to remain in-library

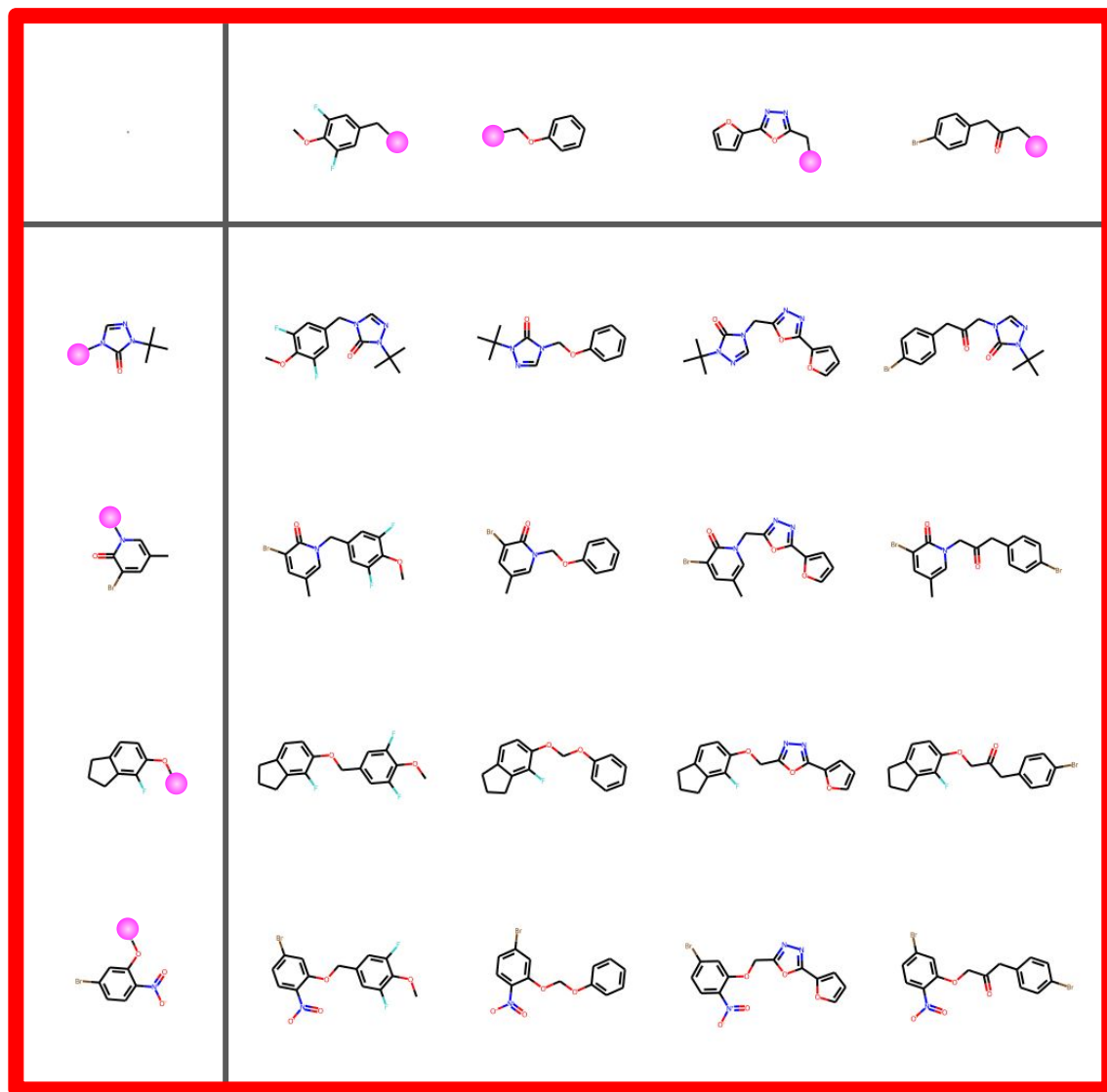
Our method utilizes minimal autoregression, improving scalability for in-library virtual discovery efforts

Combinatorial synthesis libraries

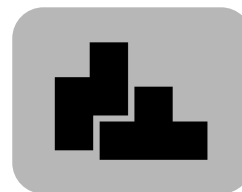
Synthesis tables



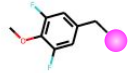
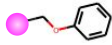
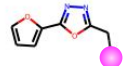
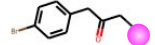
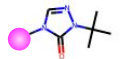
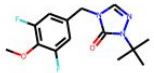
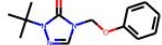
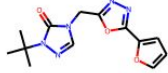
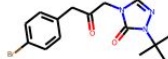

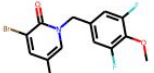
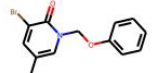
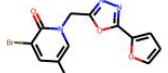
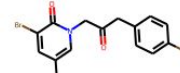
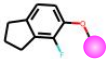
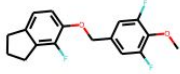
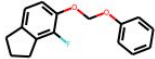
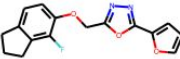
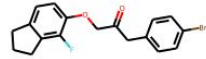
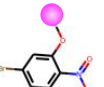
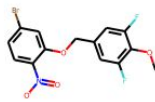
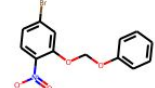

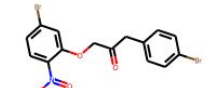
Synthesis tables



Reaction

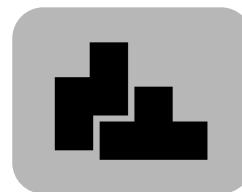


Synthesis tables

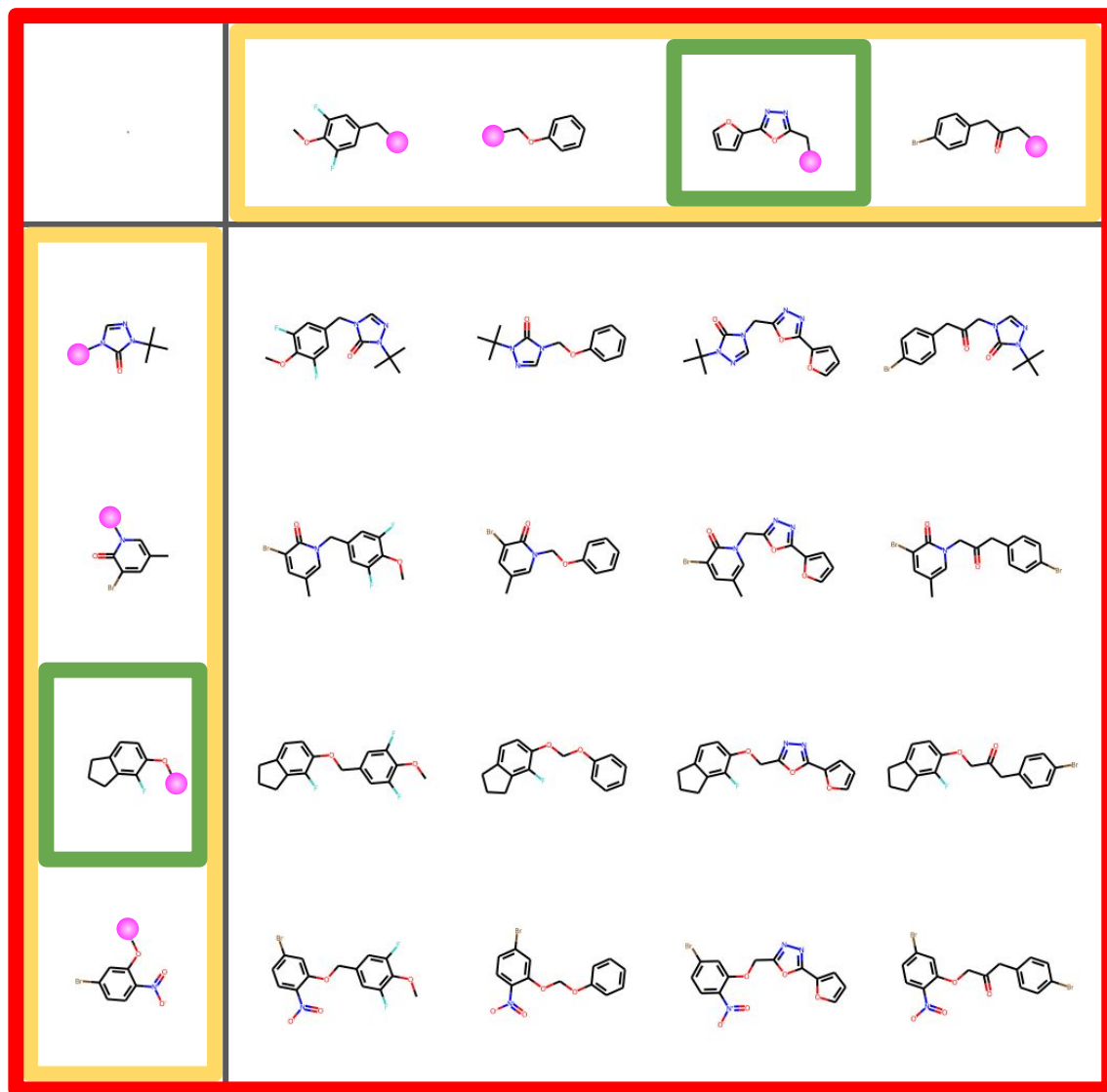
				
				
				
				
				

Reaction

R-Group



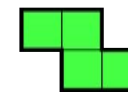
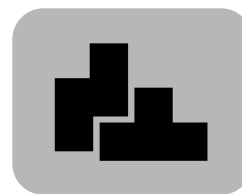
Synthesis tables



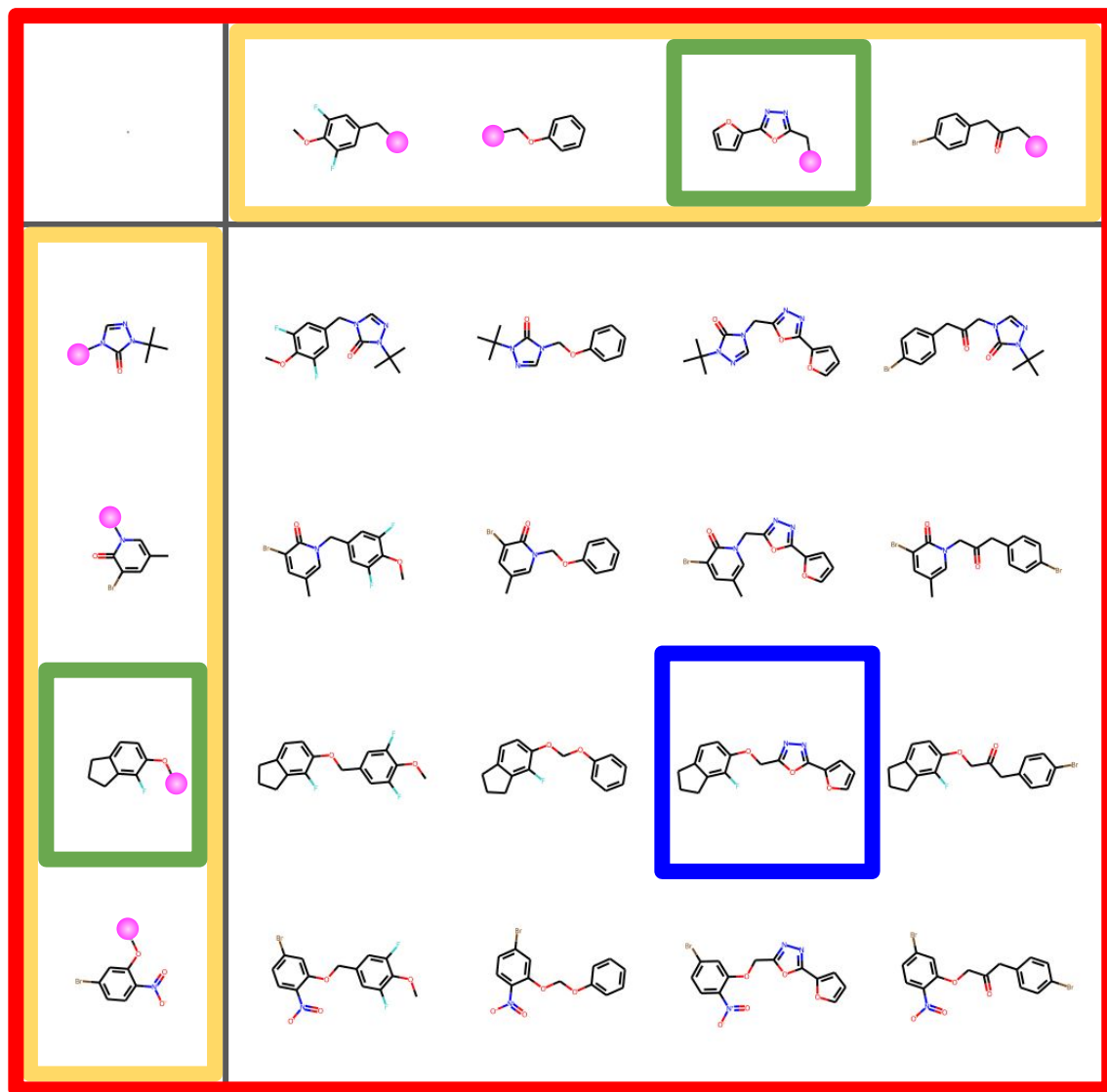
Reaction

R-Group

Synthon



Synthesis tables

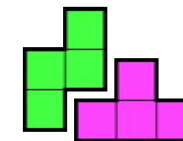
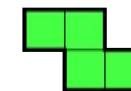
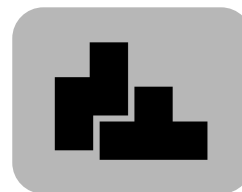


Reaction

R-Group

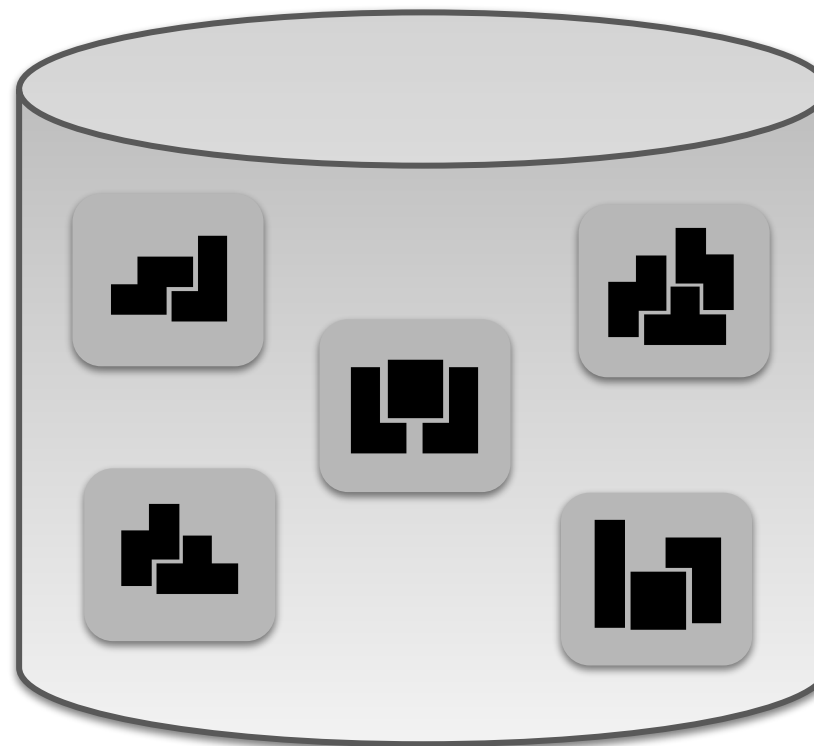
Synthon

Product



Combinatorial synthesis libraries

- Combinatorial synthesis libraries (CSLs) are comprised of a potentially large number of such k -component synthesis tables
- April 2020 Enamine REAL space
 - 340K synthons
 - 1300 reactions
 - **16B products**



Architecture

Combinatorial Synthesis Library Variational Auto-Encoder

A new graph generative model to enable navigation of combinatorial synthesis libraries

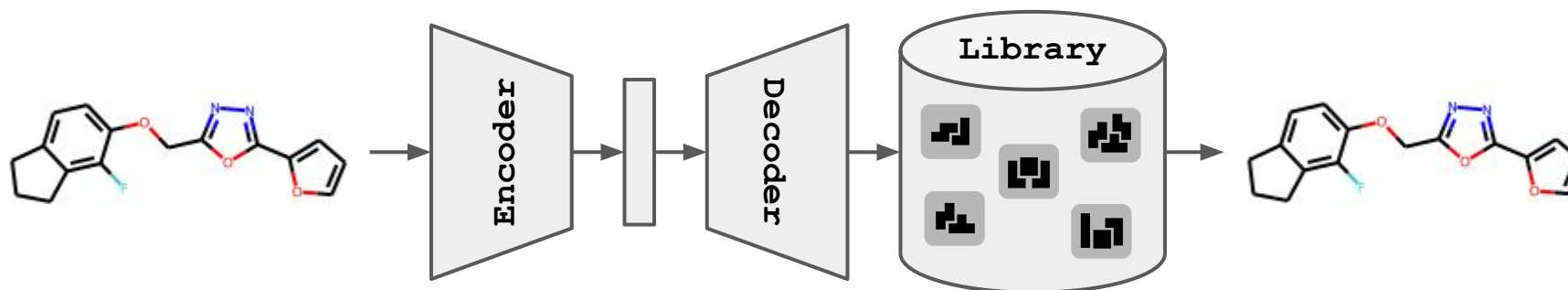
- We propose a new generative model, called the Combinatorial Synthesis Library Variational Auto-Encoder (CSLVAE), which is tailored to CSLs
 - Satisfies the hard constraint of restricting generated compounds to a provided CSL
 - Decoder design improves on computational complexity compared to existing molecular generative models and analog enumeration approaches

Manuscript currently under review

CSLVAE

Overview of architecture

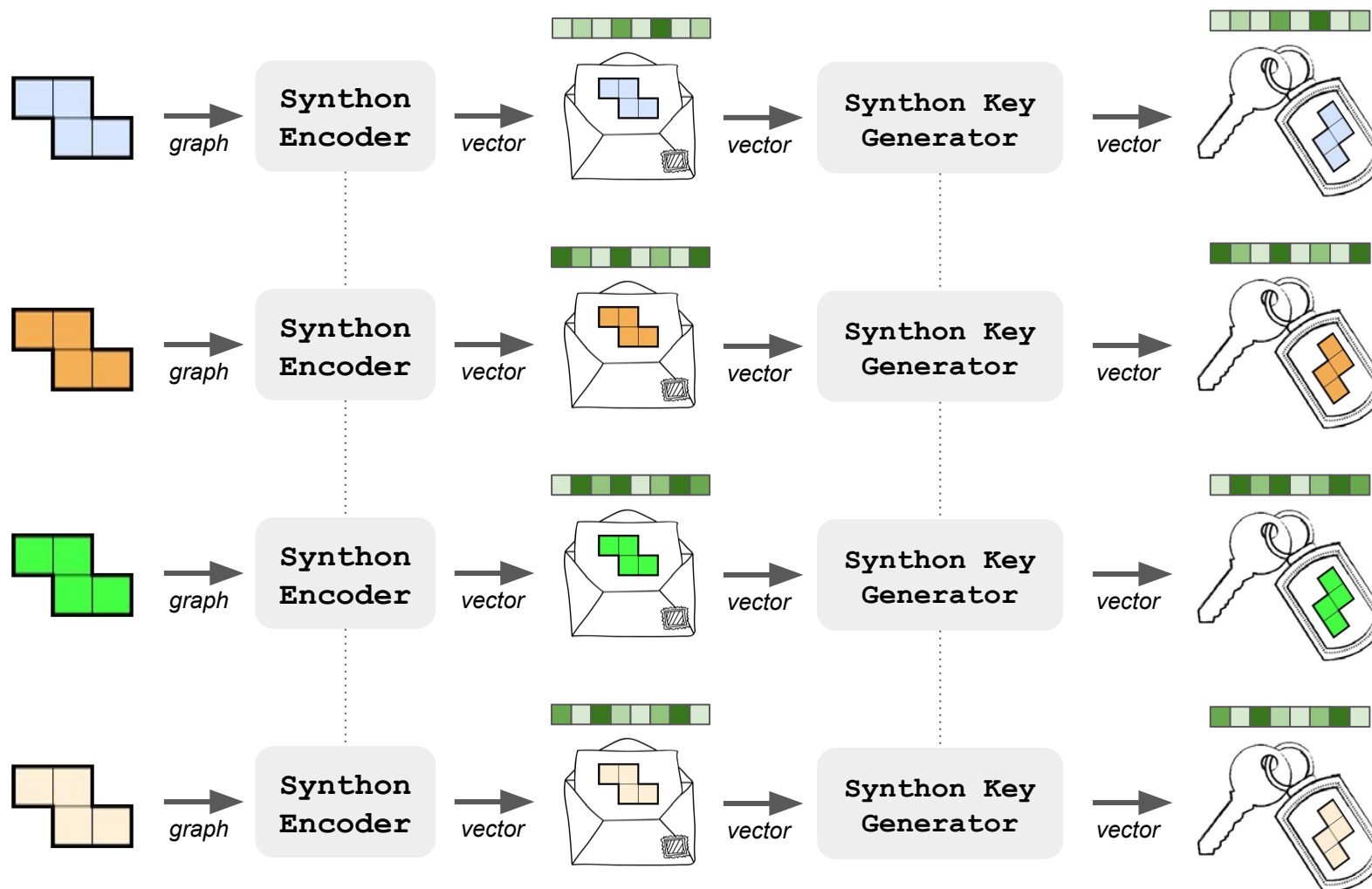
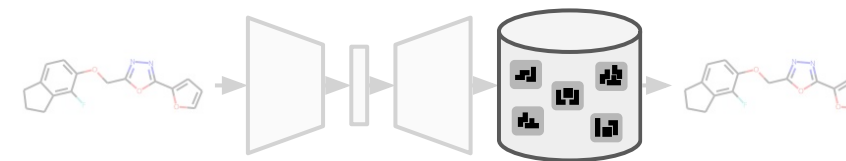
- CSLVAE uses an autoencoder as an engine for database retrieval, fusing deep generative models with neural databases



- Three components:
 1. Library encoder
 2. Molecular encoder
 3. Molecular decoder

Library encoder

Encoding the synthon representations and keys



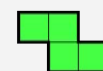
Reaction



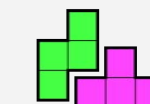
R-Group



Synthon

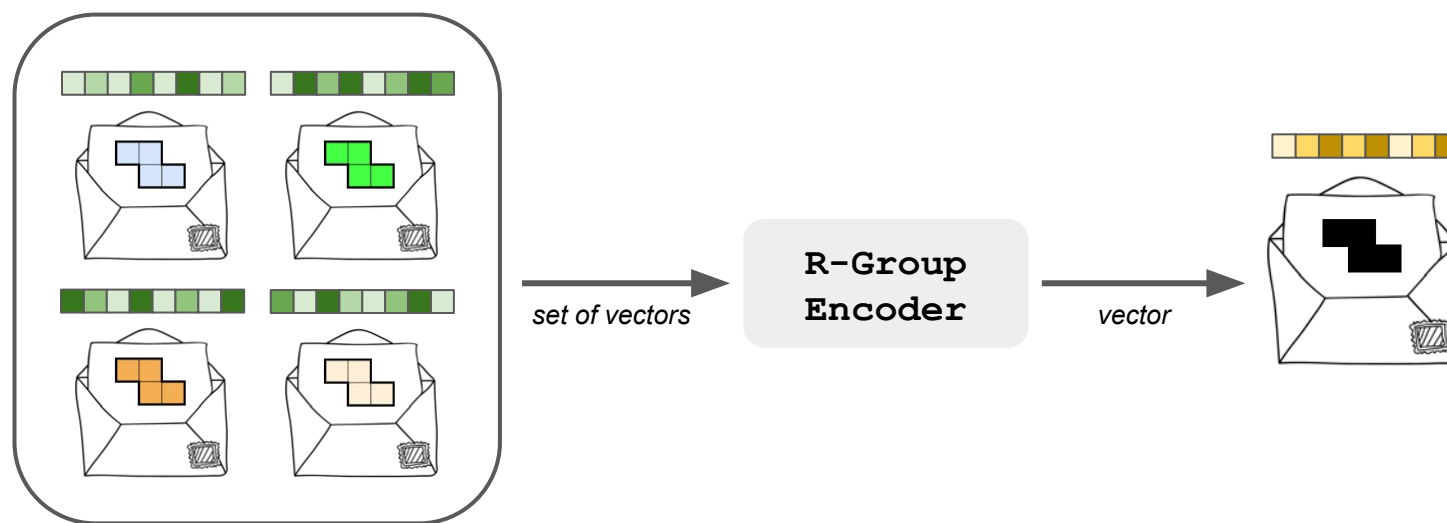
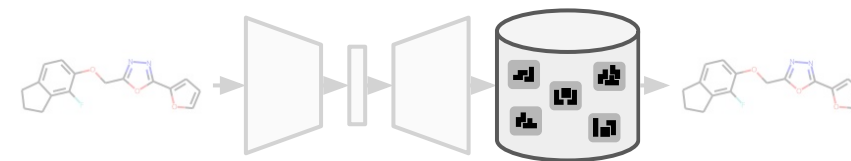


Product



Library encoder

Encoding the R-group representations



Reaction



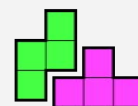
R-Group



Synthon

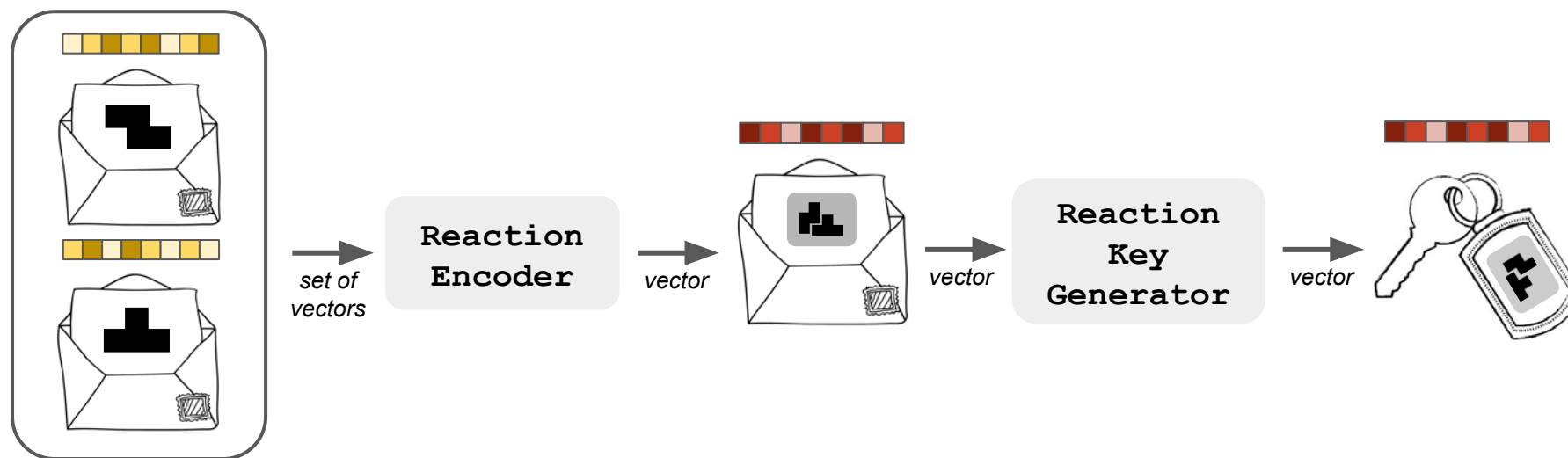
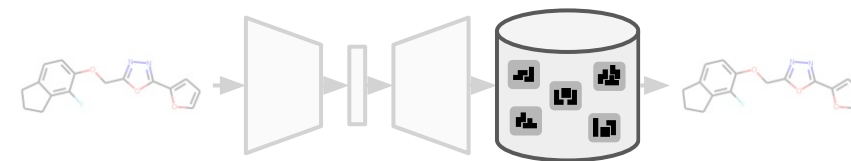


Product



Library encoder

Encoding the reaction representations and keys



Reaction



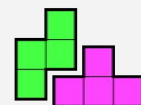
R-Group



Synthon

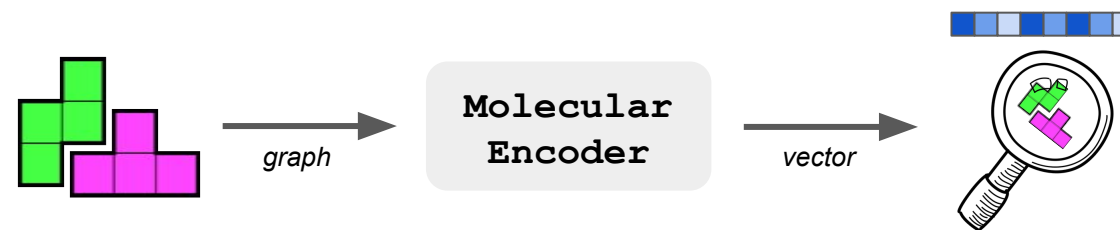
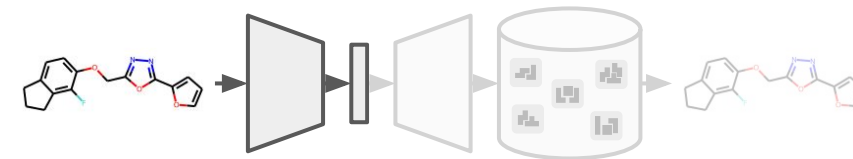


Product



Molecular encoder

Encoding the query molecule



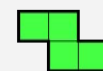
Reaction



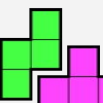
R-Group



Synthon

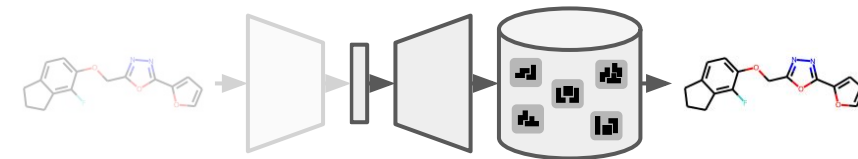
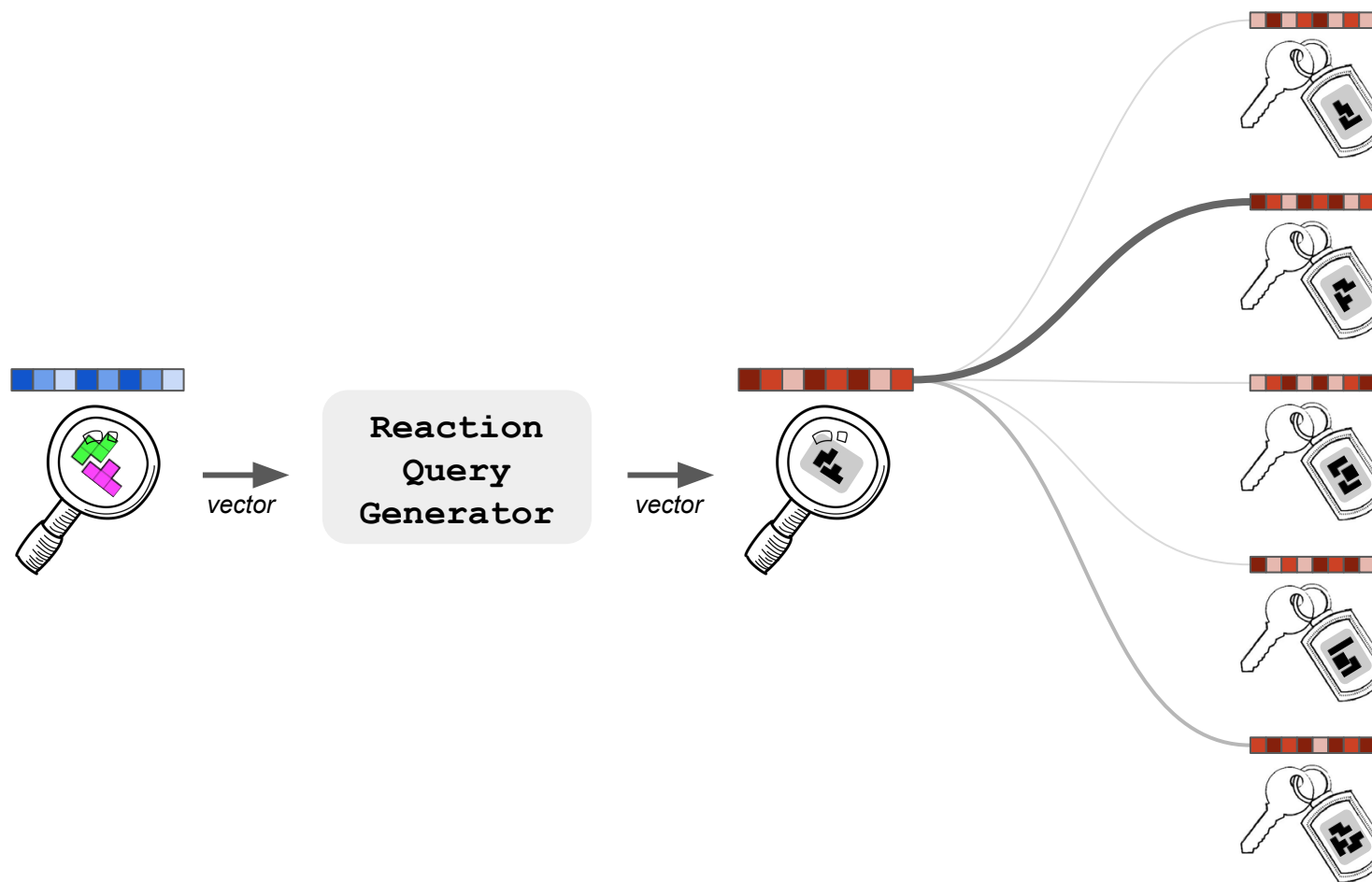


Product



Molecular decoder

Decoding the reaction type



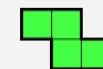
Reaction



R-Group



Synthon

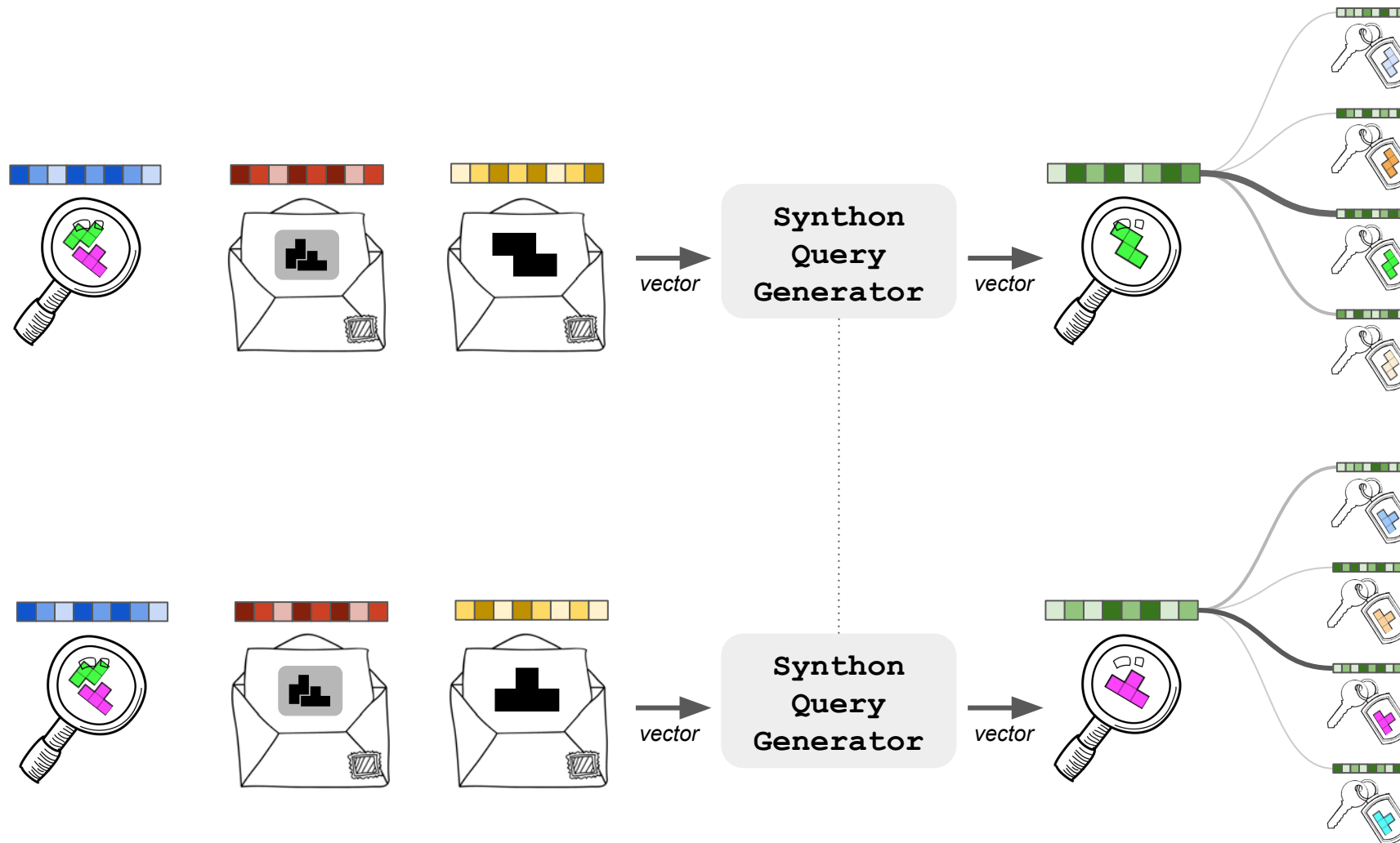
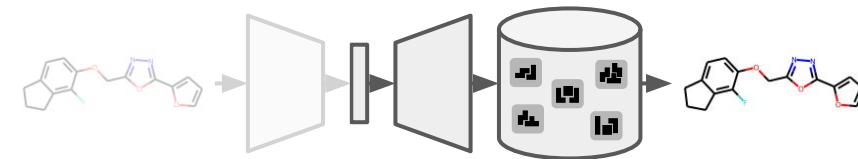


Product



Molecular decoder


Decoding one synthon per R-group given reaction type



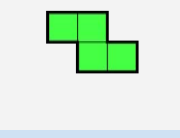
Reaction



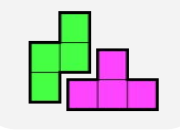
R-Group



Synthon



Product



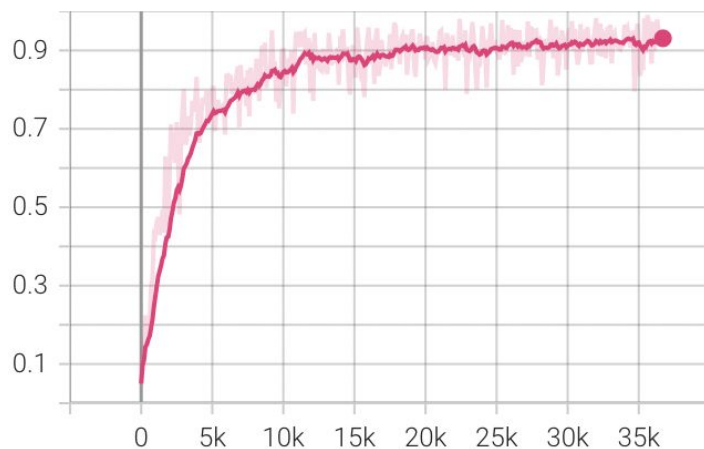
Training and inference details

- Training
 - The library encoder is shown mini-batches of the full library during training
 - Each library mini-batch covers a chemical space on the order of a few million compounds
 - Teacher forcing
- Inference
 - Given a CSL, we pre-compute all of the synthon/R-group/reaction representations and keys and cache them to the PyTorch module as buffers
 - The forward call takes a batch of molecular queries as input and retrieves a corresponding batch of compounds from the CSL

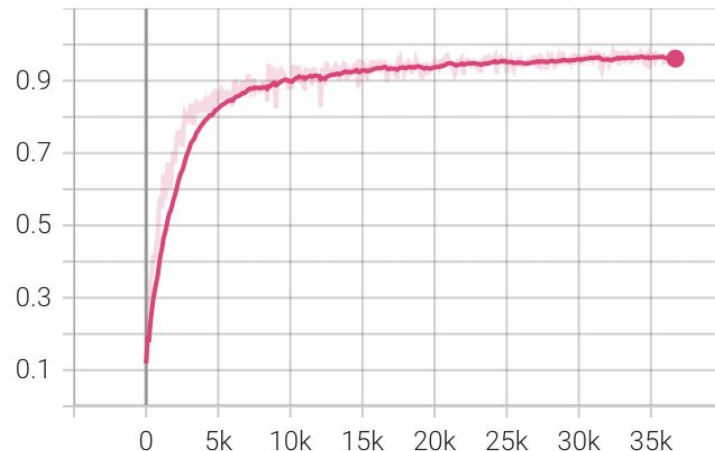
Results

Performance on Enamine REAL

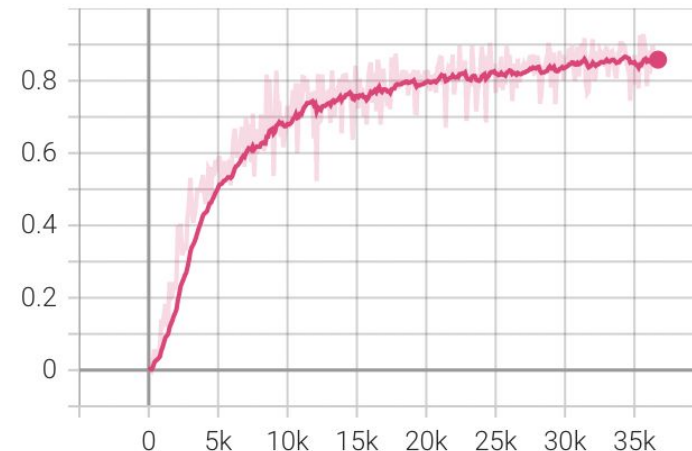
Reaction accuracy
tag: Reaction accuracy



Synthon accuracy
tag: Synthon accuracy



Full accuracy
tag: Full accuracy



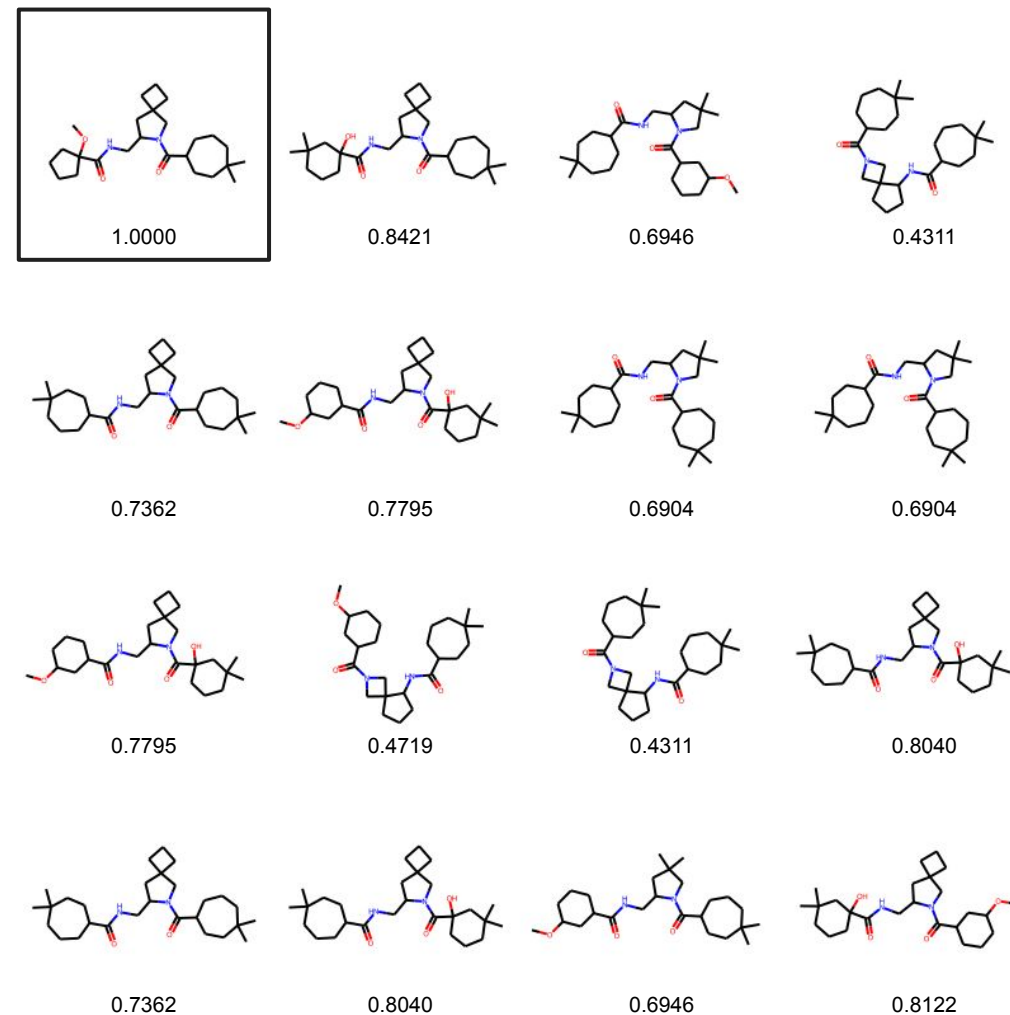
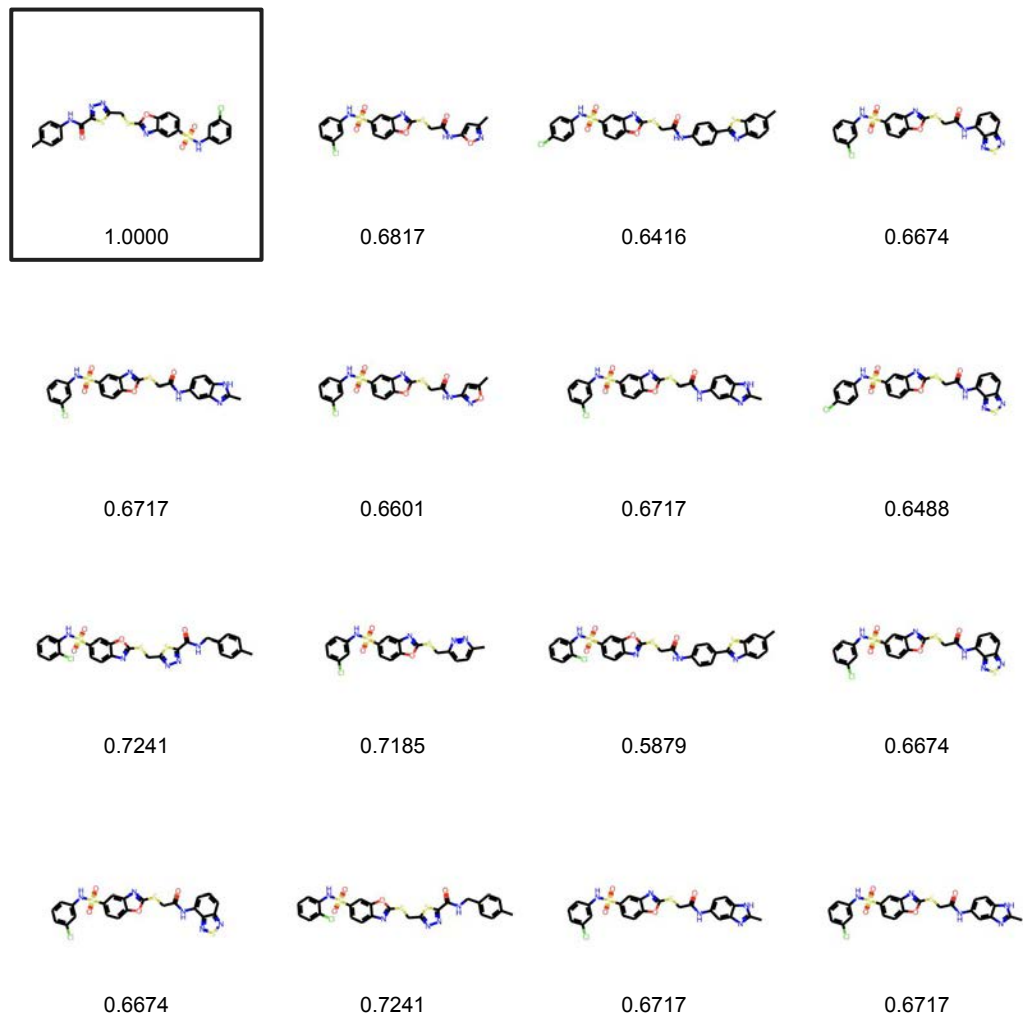
	JT-VAE	RationaleRL	CSLVAE (ours)
# Parameters	4.7M	3.4M	380K
Validity	100.0%	100.0%	100.0%
Uniqueness	80.1%	96.3%	98.8%
Average likelihood	18.7%	62.3%	72.4%
In-library proportion	2.9%	50.9%	100.0%

Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Junction tree variational autoencoder for molecular graph generation." *International Conference on Machine Learning*. PMLR, 2018.

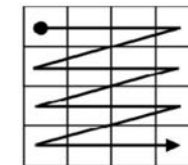
Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. "Multi-objective molecule generation using interpretable substructures." *International Conference on Machine Learning*. PMLR, 2020.

Decoding analogues

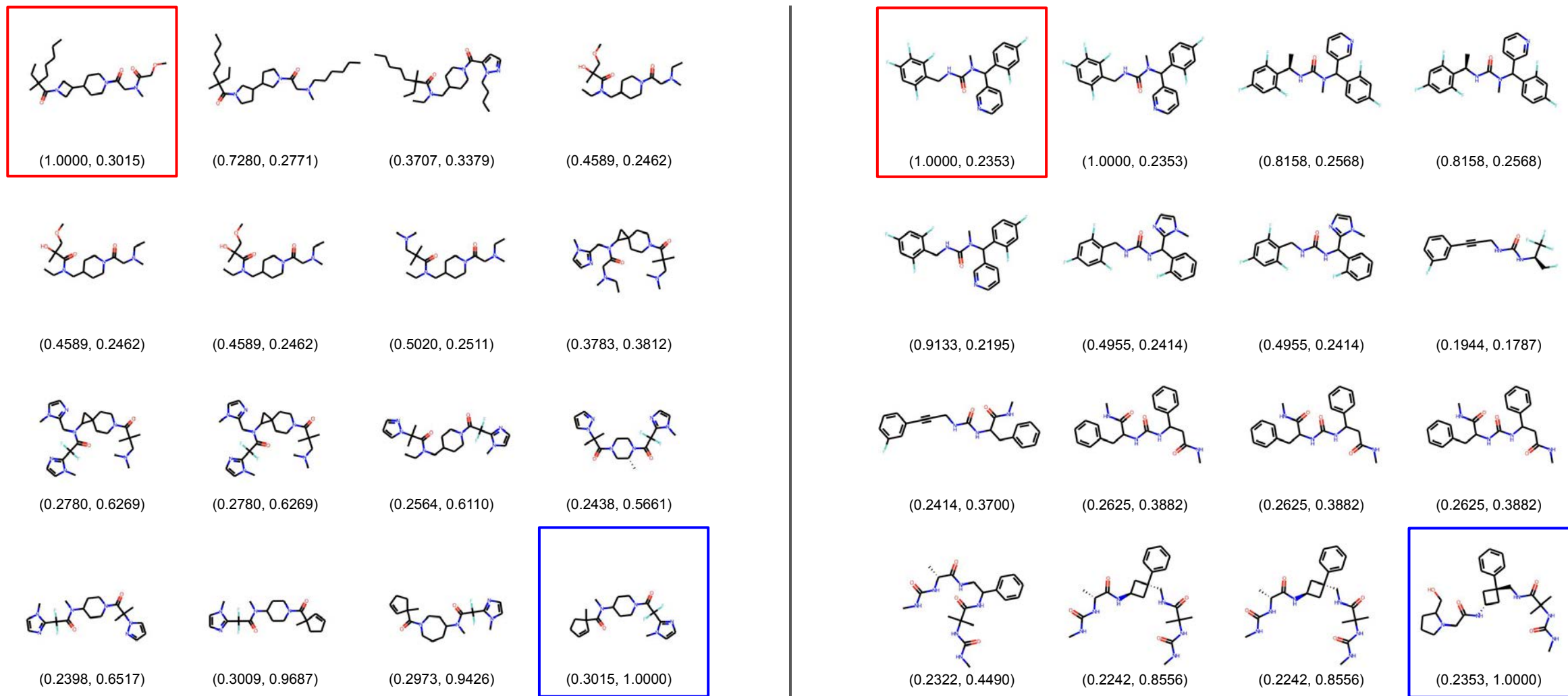
Stochastically decoding a molecular query into REAL



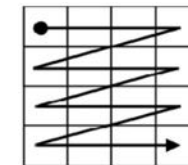
Interpolating compounds



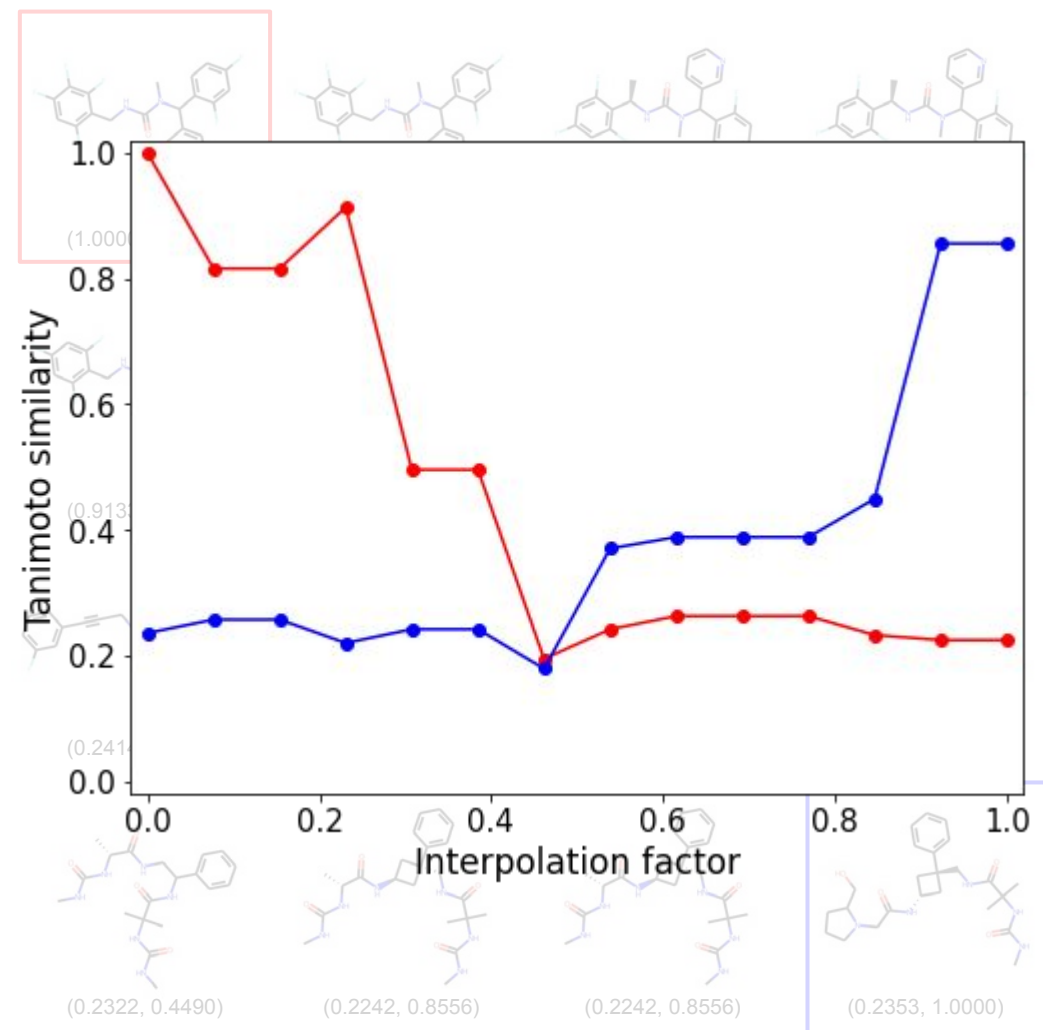
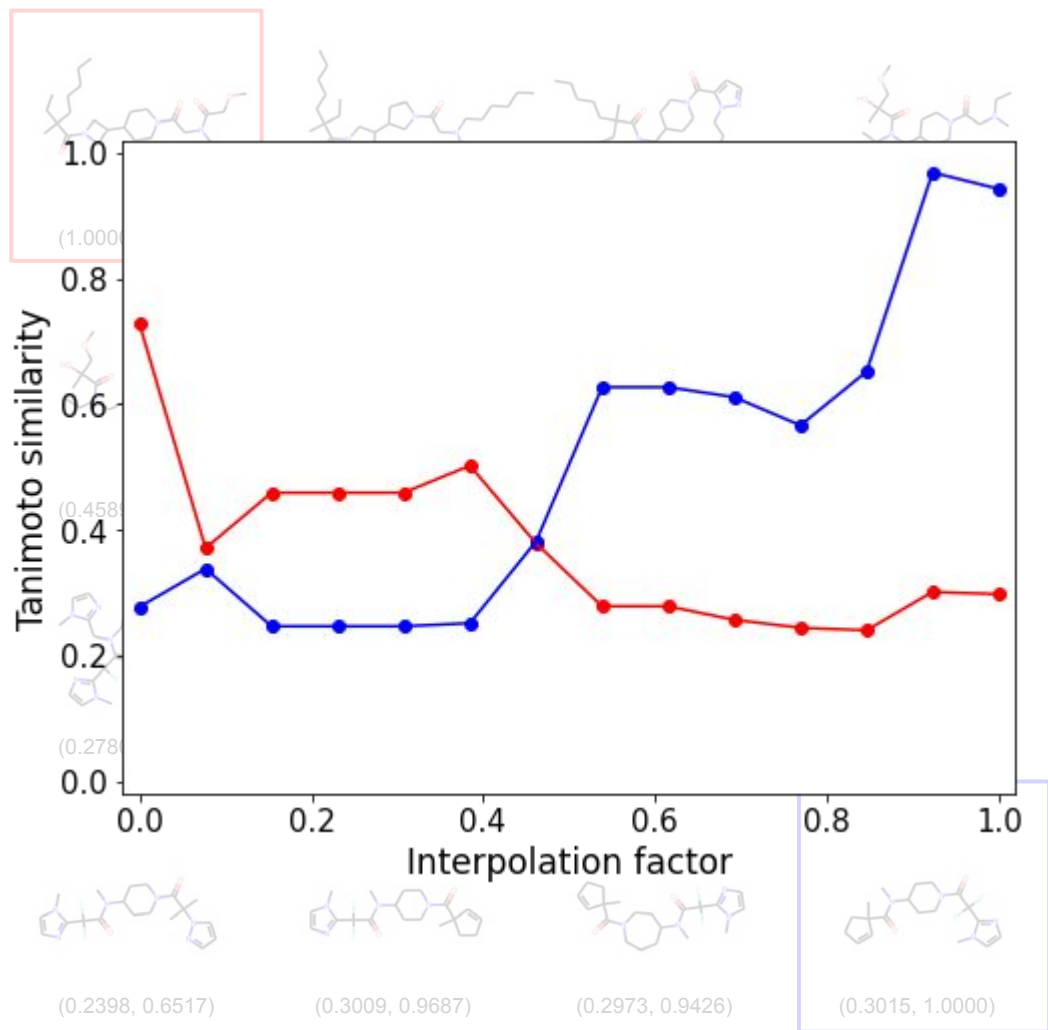
Decoding interpolated molecular queries into REAL



Interpolating compounds



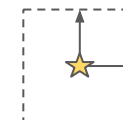
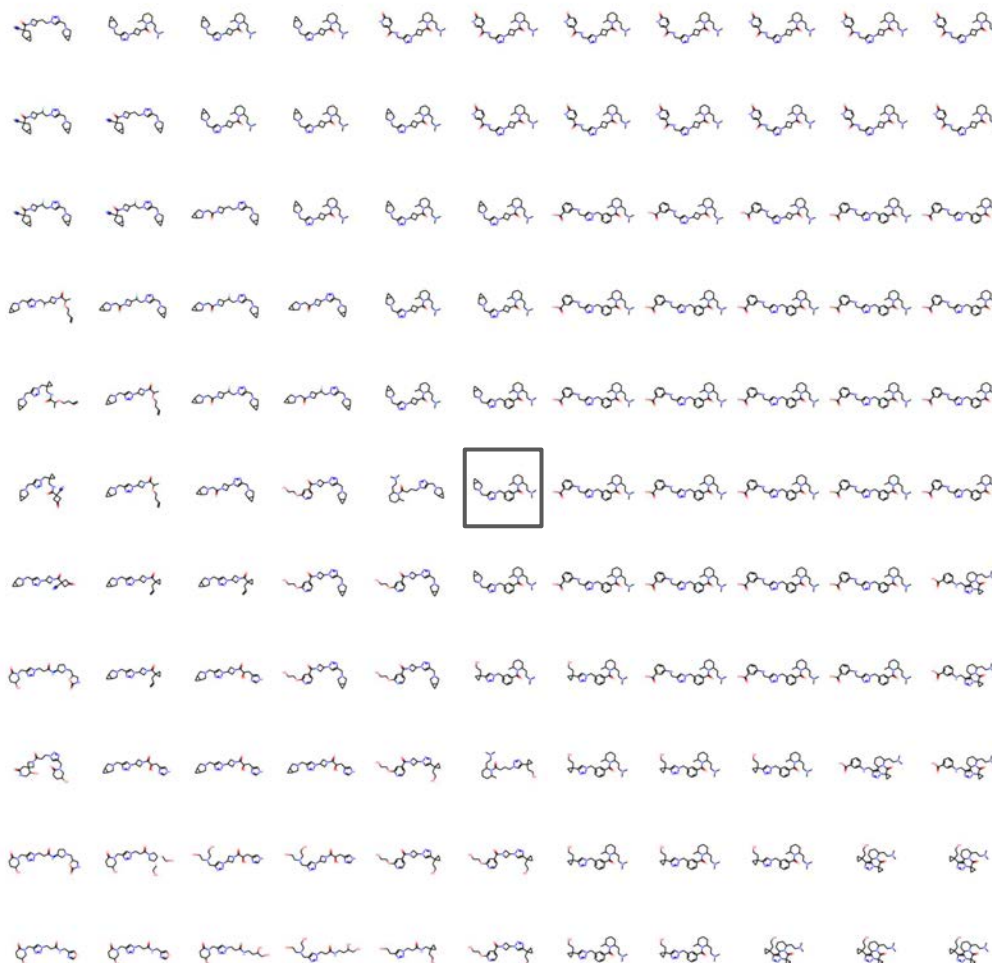
Decoding interpolated molecular queries into REAL



Visualizing local neighborhoods

Decoding molecular queries on a random 2D cross-section into REAL

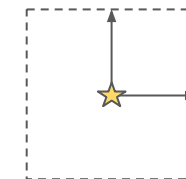
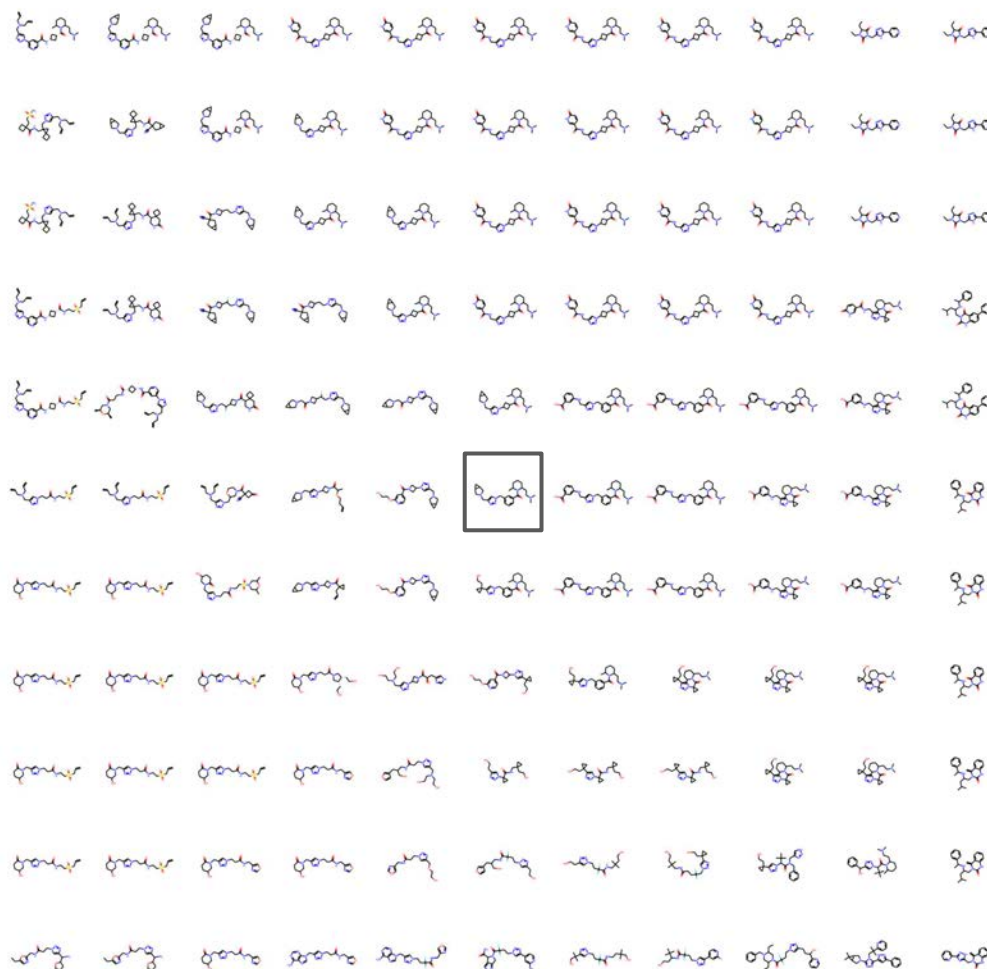
1. Sample a random molecular query from the prior
 2. Sample two random directions
 3. Form a 2D hyperplane from these three points
 4. Select molecular queries evenly on this 2D cross-section and decode into REAL
- Qualitative exercise to visualize structure of the learned latent space



Visualizing local neighborhoods

Decoding molecular queries on a random 2D cross-section into REAL

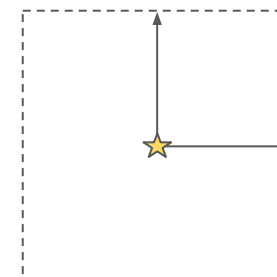
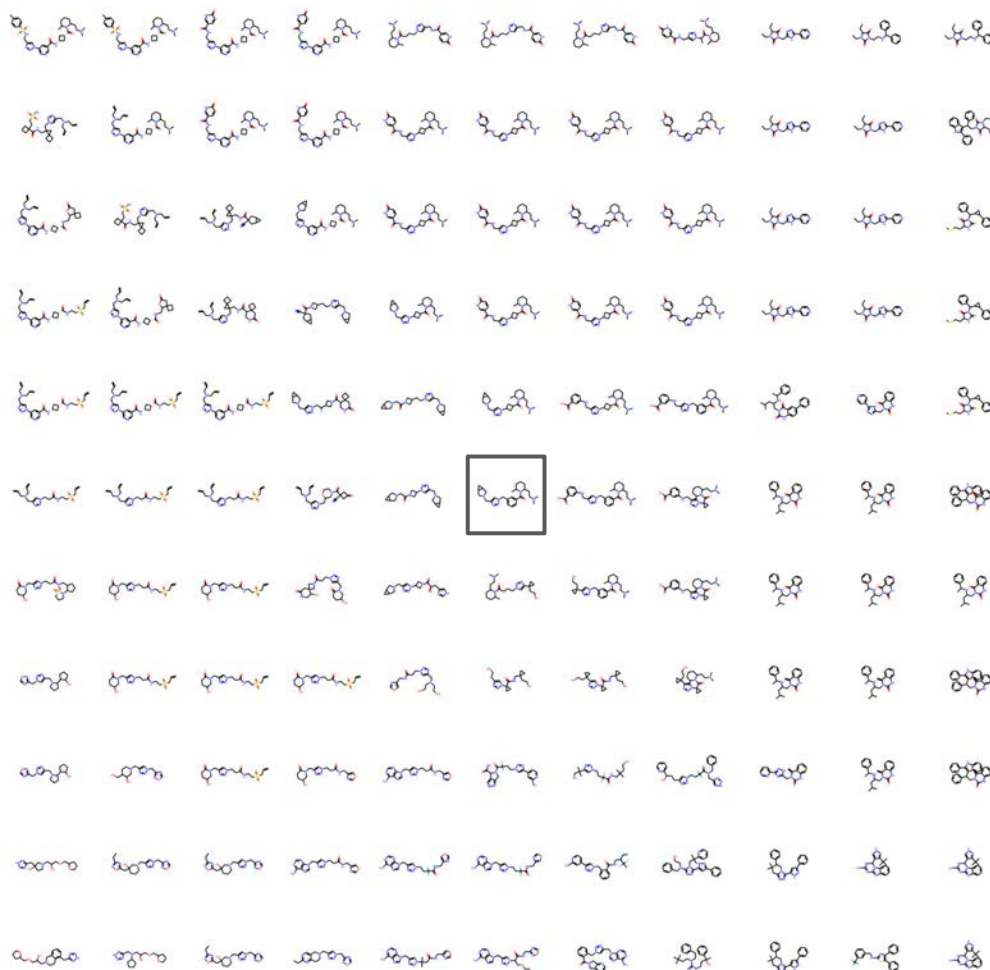
1. Sample a random molecular query from the prior
 2. Sample two random directions
 3. Form a 2D hyperplane from these three points
 4. Select molecular queries evenly on this 2D cross-section and decode into REAL
- Qualitative exercise to visualize structure of the learned latent space



Visualizing local neighborhoods

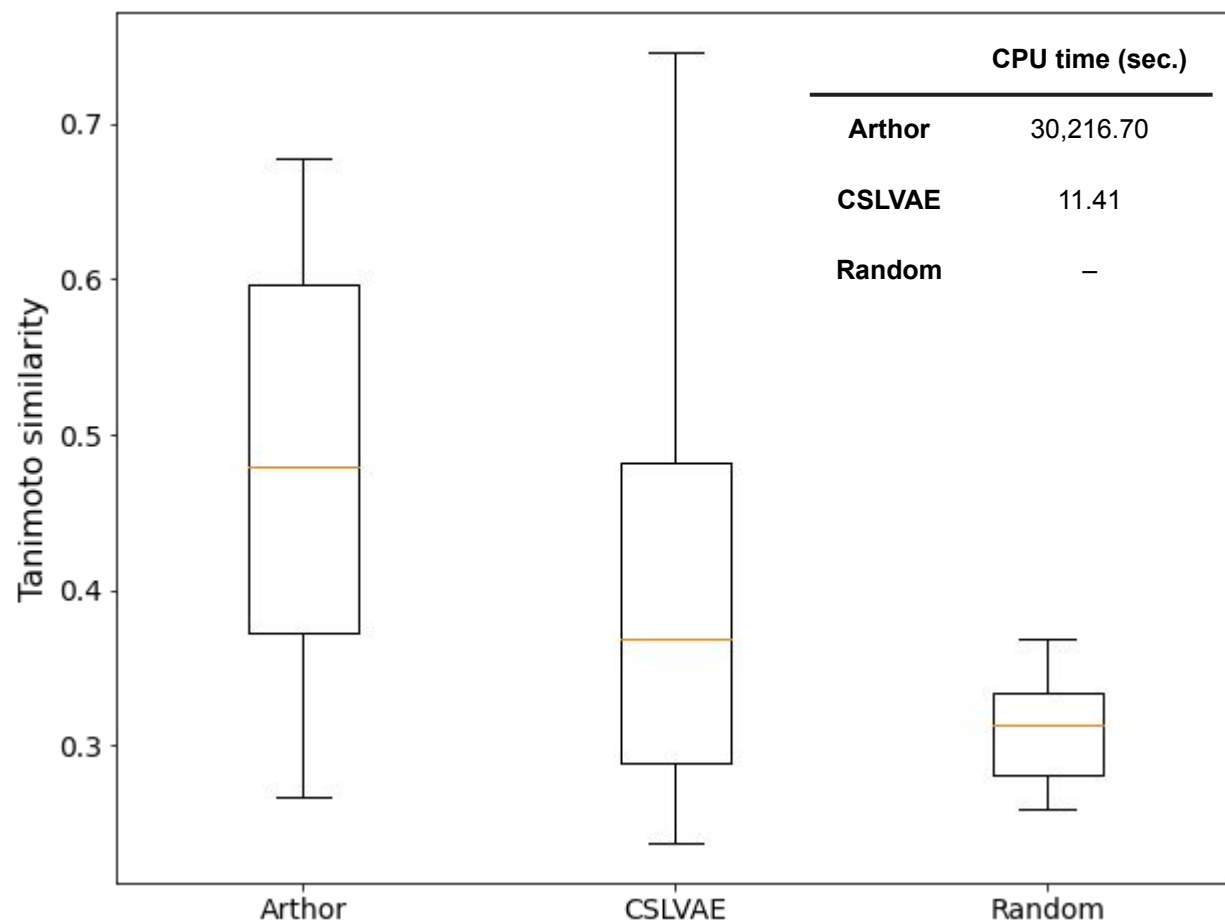
Decoding molecular queries on a random 2D cross-section into REAL

1. Sample a random molecular query from the prior
 2. Sample two random directions
 3. Form a 2D hyperplane from these three points
 4. Select molecular queries evenly on this 2D cross-section and decode into REAL
- Qualitative exercise to visualize structure of the learned latent space



An exercise in CSLVAE's analoging potential

Autoencoding with CSLVAE as an out-of-the-box analoging procedure



- Selected 24 of the [FDA Novel Drug Approvals for 2021](#)
- For each compound, we analog into REAL via autoencoding with CSLVAE, and compare to a naive baseline and Arthor (from NextMove) as a source of ground-truth
- For each method, we select 100 compounds and re-compute ECFP4 Tanimoto similarity in RDKit, selecting the top-1 analogue for each
- Boxplot shows distribution of Tanimoto similarities for top-1 analogues

Wrap-up

Properties of CSLVAE

- Only one step of autoregression in the decoder, irrespective of graph size
- CSLVAE-style retrieval has provably logarithmic computational complexity
- The molecular encoder can take any (valid) molecular graph as input, but the molecular decoder is guaranteed to stay in the library
- CSLVAE is inductive: the CSL is itself an input to the model
- Molecular encoder is permutation invariant, library encoder is permutation equivariant, molecular decoder does not suffer from canonicalization ambiguity

Conclusion

- We develop a new graph generative model, CSLVAE, to enable the navigation of ultra-large combinatorial synthesis libraries
- Our method has favorable scaling properties for non-enumerative libraries
- CSLVAE learns a latent space with “smooth” transitions in chemical space
- We demonstrate CSLVAE’s capabilities for out-of-the-box analogue enumeration as a proof-of-concept
- Future work to utilize CSLVAE in a virtual screening context forthcoming

An efficient graph generative model for navigating ultra-large combinatorial synthesis libraries

Aryan Pedawi
Atomwise Inc.
aryan@atomwise.com

Pawel Gniewek
Atomwise Inc.
pawel@atomwise.com

Chaoyi Chang
Atomwise Inc.
cchang373@atomwise.com

Brandon Anderson^{*†}
Atomic AI
branderson@gmail.com

Henry van den Bedem[†]
Atomwise Inc.
vdbedem@atomwise.com

Abstract

Virtual, make-on-demand chemical libraries have transformed early-stage drug discovery by unlocking vast, synthetically accessible regions of chemical space. Recent years have witnessed rapid growth in these libraries from millions to trillions of compounds, hiding undiscovered, potent hits for a variety of therapeutic targets. However, they are quickly approaching a size beyond that which permits explicit enumeration, presenting new challenges for virtual screening. To overcome these challenges, we propose the **Combinatorial Synthesis Library Variational Auto-Encoder (CSLVAE)**. The proposed generative model represents such libraries as a differentiable, hierarchically-organized database. Given a compound from the library, the molecular encoder constructs a query for retrieval, which is utilized by the molecular decoder to reconstruct the compound by first decoding its chemical reaction and subsequently decoding its reactants. Our design minimizes autoregression in the decoder, facilitating the generation of large, valid molecular graphs. Our method performs fast and parallel batch inference for ultra-large synthesis libraries, enabling a number of important applications in early-stage drug discovery. Compounds proposed by our method are guaranteed to be in the library, and thus synthetically and cost-effectively accessible. Importantly, CSLVAE can encode out-of-library compounds and search for in-library analogues. In experiments, we demonstrate the capabilities of the proposed method in the navigation of massive combinatorial synthesis libraries.



Aryan Pedawi

Atomwise Inc.

aryan@atomwise.com

Thank you!

Co-authors

Paweł Gniewek, Chaoyi Chang, Brandon M. Anderson, Henry van den Bedem

Support

Many wonderful colleagues at Atomwise,
and all of you for listening!