MoLeR: Creating a path to more efficient drug design

Krzysztof Maziarz

Microsoft Research Cambridge

Generative Chemistry at MSR Cambridge 

Solving fundamental problems in Chemistry using Deep Learning

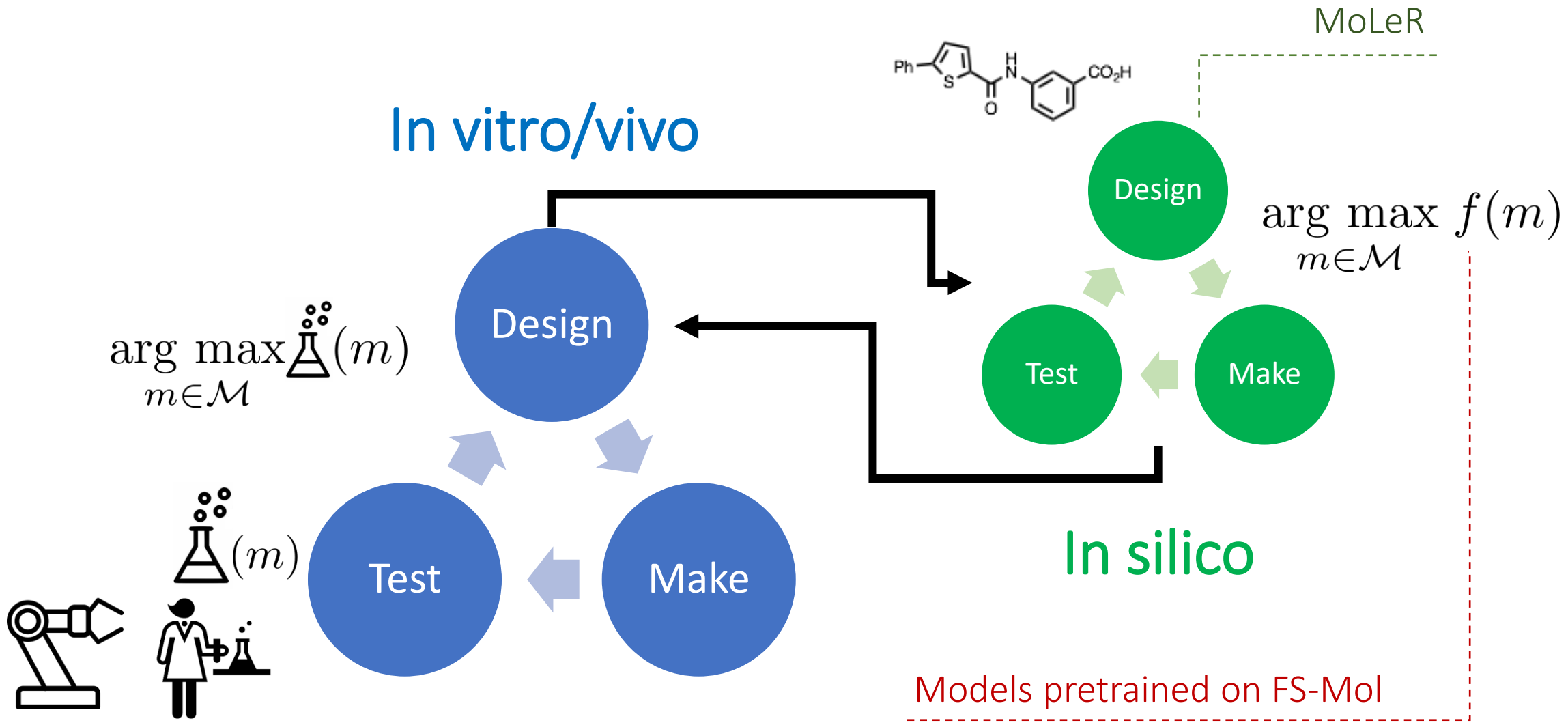
- molecular generation & optimization — MoLeR
- molecular property prediction — FS-Mol
- reaction prediction & retrosynthesis
- structure-based drug design

ongoing

[MoLeR] Learning to Extend Molecular Scaffolds with Structural Motifs

[FS-Mol] FS-Mol: A Few-Shot Learning Dataset of Molecules

Lead Optimization: Design-Make-Test



Molecule generation: yet another model?

Requirements:

- can be constrained by scaffold + specific attachment points 🔒
- not easily exploited during optimization ✕
- fast in practice 🚀

GNN-based, enforce constraints during sequential decoding

Assemble molecules from common fragments/motifs

Most GNN-based models unpractically slow, simple RNNs *much* faster

Molecule generation: yet another model?

…models tend to either be fast and simple, but with limited options for structural constraints, or give such options, but are slow and clunky ☹

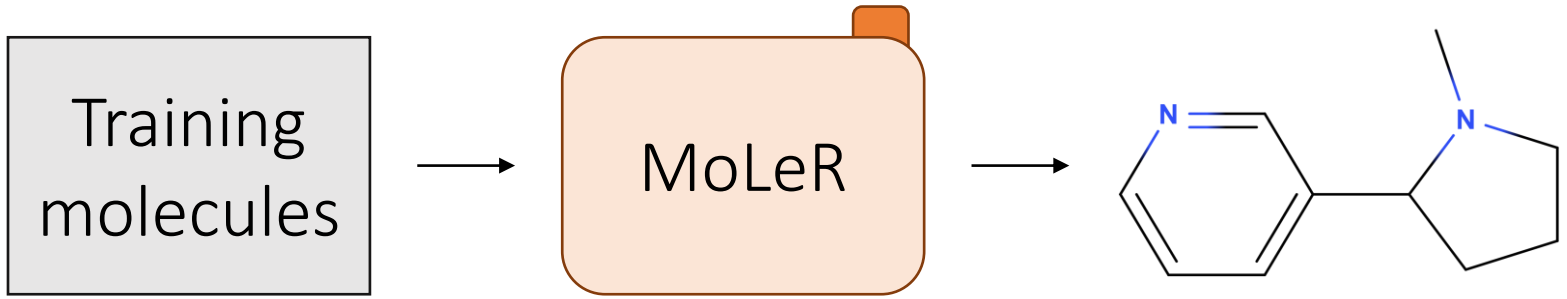
We made MoLeR to get something that

- supports scaffold/attachment constraints
- uses fragments (motifs)

…but is well-engineered, fast, and without unnecessary bells & whistles

MoLeR: Introduction

MoLeR is a generative model: given a training dataset, it can learn to produce more molecules from the same distribution.



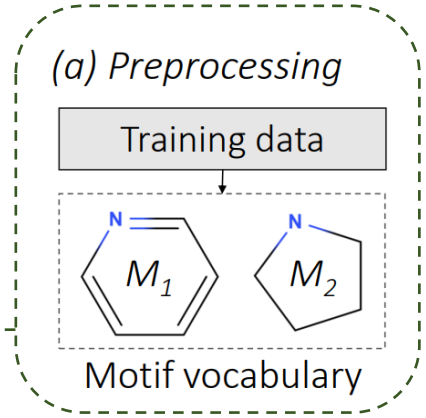
Scaffold information is only introduced during inference.

For optimization, we use a black-box optimization method MSO.

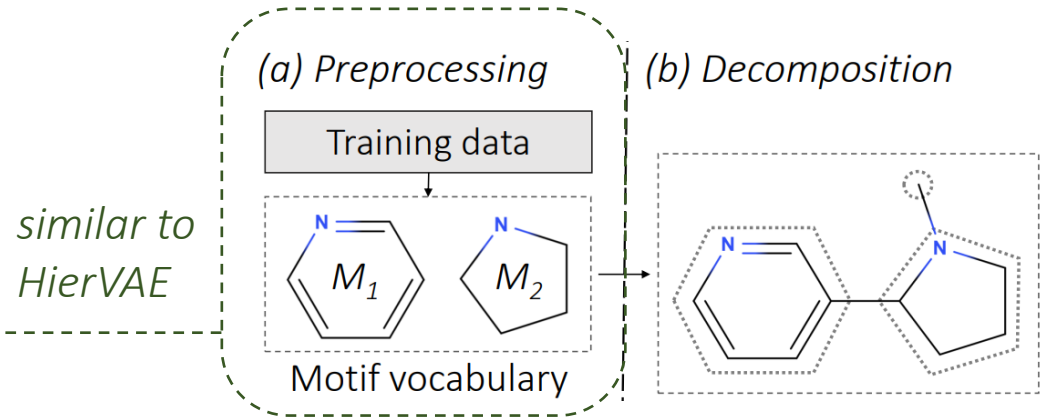
[MSO] Efficient multi-objective molecular optimization in a continuous latent space

MoLeR: Deep dive

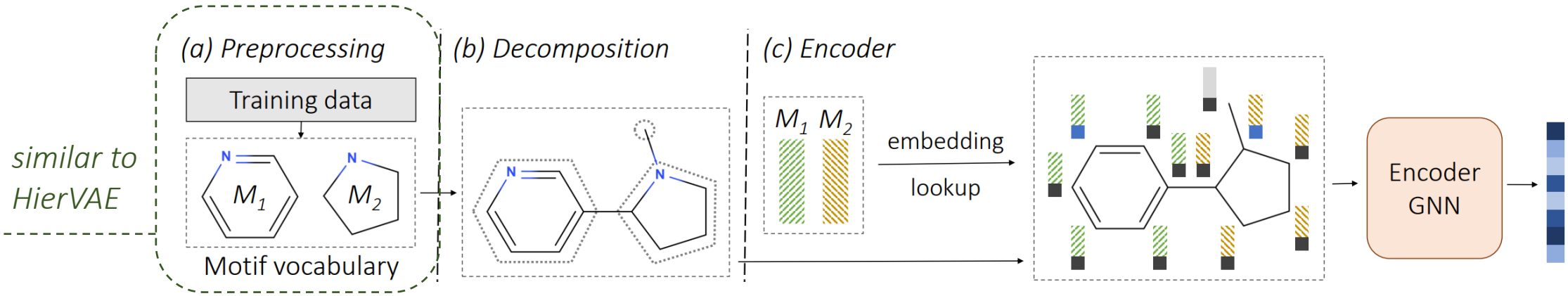
similar to HierVAE



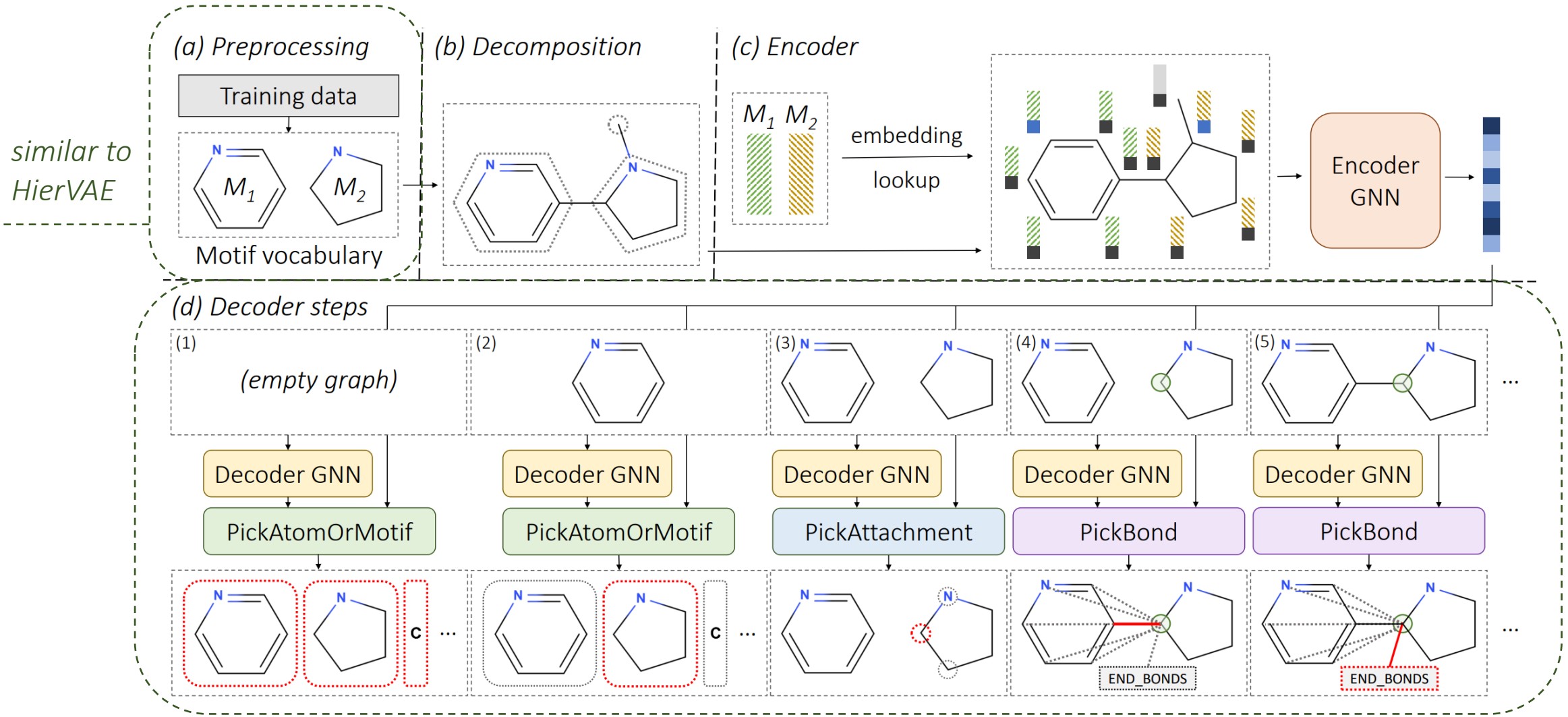
MoLeR: Deep dive



MoLeR: Deep dive



MoLeR: Deep dive



decoder steps independent given encoder output → can start from scaffold

Quantitative results: Guacamol

Method	GuacaMol	
	Score	Quality
Best of dataset	0.61	0.77
SMILES LSTM	0.87	0.77
SMILES GA	0.72	0.36
GRAPH MCTS	0.45	0.22
GRAPH GA	0.90	0.40
CDDD + MSO	0.90	0.58
MNCE-RL	0.92	0.54
MoLeR + MSO	0.82	0.75

Good balance of score vs quality

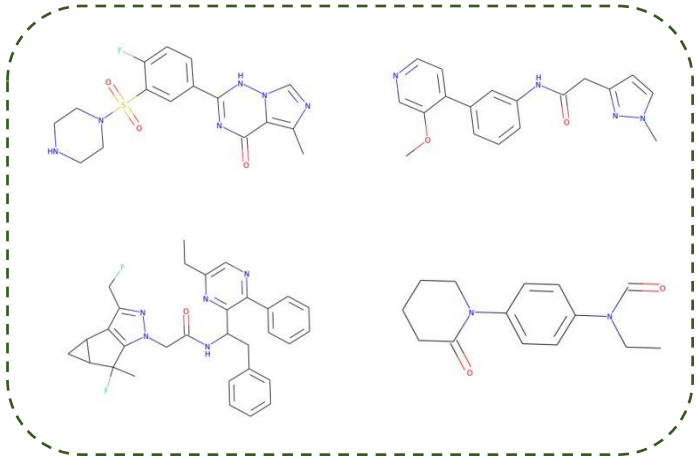
[GuacaMol] GuacaMol: Benchmarking Models for De Novo Molecular Design

Quantitative results: Guacamol

Method	GuacaMol		Scaffolds	
	Score	Quality	Score	Quality
Best of dataset	0.61	0.77	0.17	-
SMILES LSTM	0.87	0.77	0.45	-
SMILES GA	0.72	0.36	0.45	-
GRAPH MCTS	0.45	0.22	0.20	-
GRAPH GA	0.90	0.40	0.79	-
CDDD + MSO	0.90	0.58	0.92	0.59
MNCE-RL	0.92	0.54	0.95	0.47
MoLeR + MSO	0.82	0.75	0.93	0.63

Good balance of score vs quality

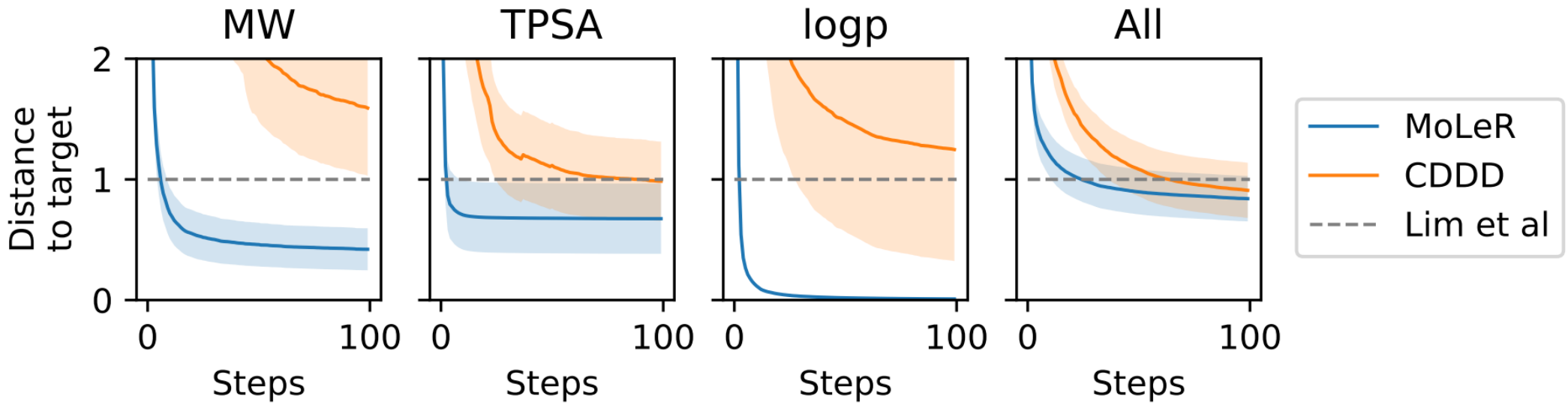
New benchmarks (complex scaffolds)



[GuacaMol] GuacaMol: Benchmarking Models for De Novo Molecular Design

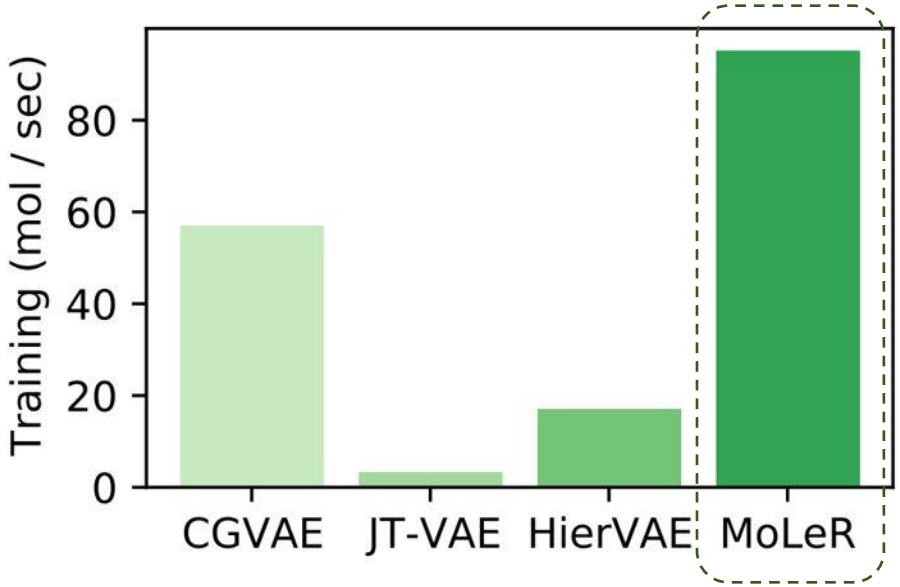
Quantitative results: more scaffold-based tasks

MoLeR more robust than CDDD with a soft scaffold constraint in scoring.

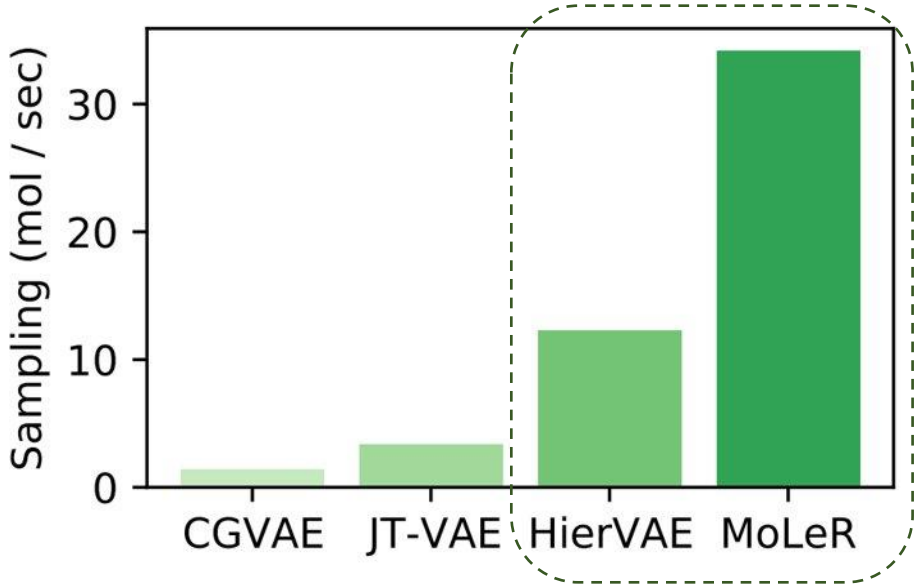


[Lim et al] Scaffold-based molecular design using graph generative model

Quantitative results: speed



Training not autoregressive → much faster



sampling 2.7x faster than HierVAE

[CGVAE] Constrained Graph Variational Autoencoders for Molecule Design
[JT-VAE] Junction Tree Variational Autoencoder for Molecular Graph Generation
[HierVAE] Hierarchical Generation of Molecular Graphs using Structural Motifs

MoLeR: Further reading

LEARNING TO EXTEND MOLECULAR SCAFFOLDS WITH STRUCTURAL MOTIFS

Krzysztof Maziarz* Microsoft Research United Kingdom

Henry Jackson-Flux Microsoft Research United Kingdom

Pashmina Cameron Microsoft Research United Kingdom

Finton Sirockin Novartis Switzerland

Nadine Schneider Novartis Switzerland

Nikolaus Stiefl Novartis Switzerland

Marwin Segler Microsoft Research United Kingdom

Marc Brockschmidt Microsoft Research United Kingdom

(ICLR ‘22 paper)

Microsoft Research Blog

MoLeR: Creating a path to more efficient drug design

Published April 27, 2022

By Krzysztof Maziarz, Senior Applied Researcher; Marc Brockschmidt, Senior Principal Researcher; Marwin Segler, Principal Researcher

(MSR blogpost)

MoLeR is open-source!

molecule-generation 0.2.0

```
pip install molecule-generation
```

MoLeR: A Model for Molecule Generation

101 stars
9 watching
18 forks

CI passing | license MIT | pypi v0.2.0 | code style black

This repository contains training and inference code for the MoLeR model introduced in Learning to Extend Molecular Scaffolds with Structural Motifs. We also include our implementation of CGVAE, but it currently lacks integration with the high-level model interface, and is provided mostly for reference.

github.com/microsoft/molecule-generation

A MoLeR checkpoint trained using the default hyperparameters is available here. This file needs to be saved in a fresh folder `MODEL_DIR` (e.g., `/tmp/MoLeR_checkpoint`) and be renamed to have the `.pkl` ending (e.g., to `GNN_Edge_MLP_MoLeR__2022-02-24_07-16-23_best.pkl`). Then you can sample 10 molecules by running

```
molecule_generation sample MODEL_DIR 10
```

Application at Novartis: projects

MoLeR is part of a Generative Chemistry system built at Novartis.

	# design cycles	# selected for synthesis	# synthesized	# with desired profile	Status
Project A	3	6	3	3	ongoing
Project B	3	5	1	0	on hold
Project C	3	6	2	2	ongoing

anonymized

in-silico property predictions often carry over to real life

This is just a pilot, many more projects are coming.

Application at Novartis: Chemist feedback 

"The nice thing from GenChem is the exhaustive coverage of chemical space around our active (…) compounds, keeping within suitable property space"

“Many molecules look really synthesizable”

"Example of designs that we would not have considered without GenChem“

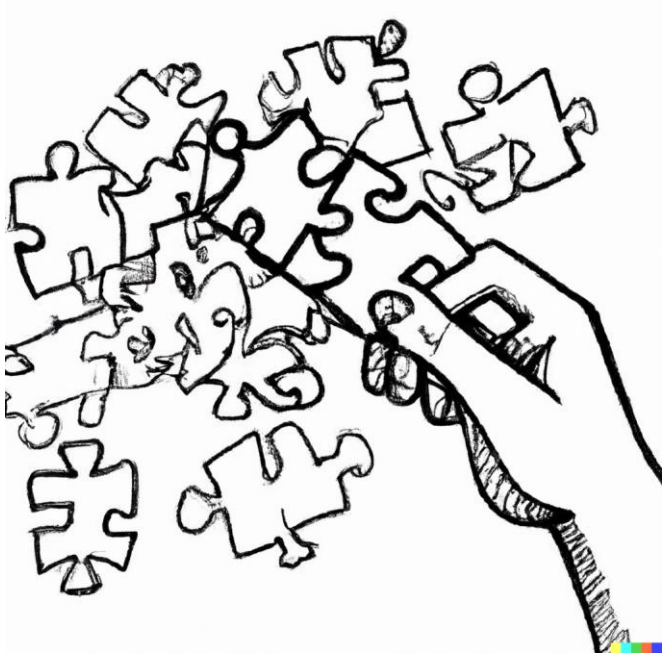
"The team had considered hydroxy-XXX before and didn't move on it. GenChem suggested it again and we made it this time around, it is nicely active, one of the most potent compounds"

ML inspiring chemists, not replacing them

Application at Novartis: Learnings

Another piece of feedback: *“The molecules look great, but maybe too similar to the ones we designed.”*
→ Our interpretation: the platform passed the “Turing test”

- Good software engineering is critical
- Everything takes longer than expected ☺
- Connecting many different ML models/algorithms into one pipeline is more complex than it sounds



Our team

Marwin Segler

Sarah Lewis

Megan Stanley

Nikolaus Stiefl

Nadine Schneider

Finton Sirockin

Raquel Rodriguez Perez

Jose Jimenez Luna

Benedict Irwin

Krzysztof Maziarz

Hubert Misztela

Jessica Lanini

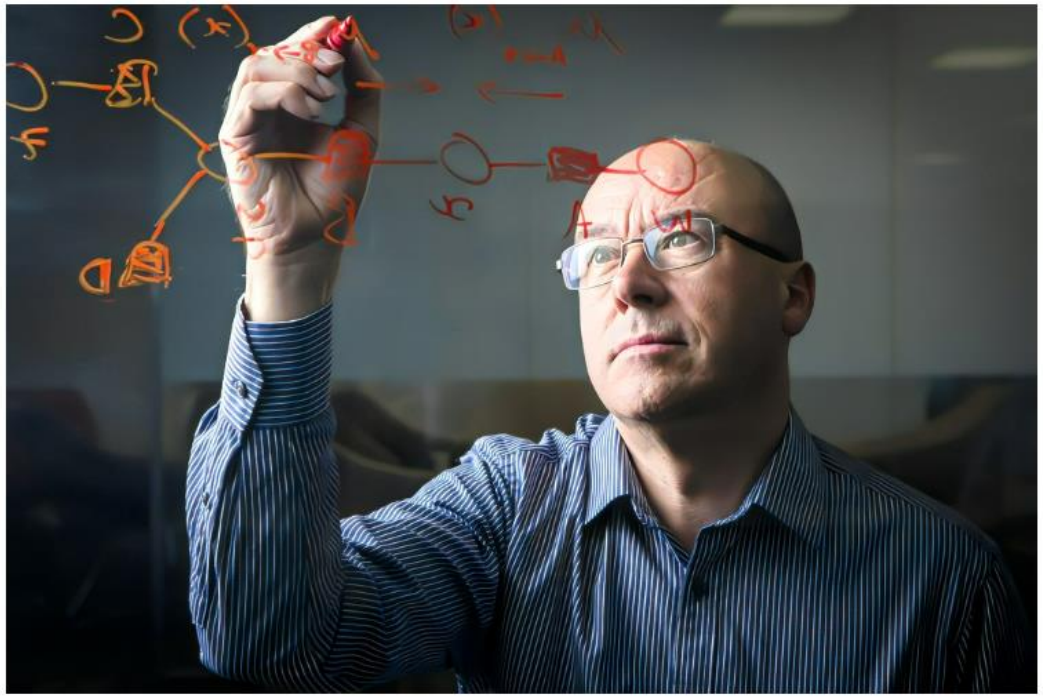
Qurrat Ul Ain

Ohhyeon Choung

Microsoft Research

Novartis

AI4Science at Microsoft



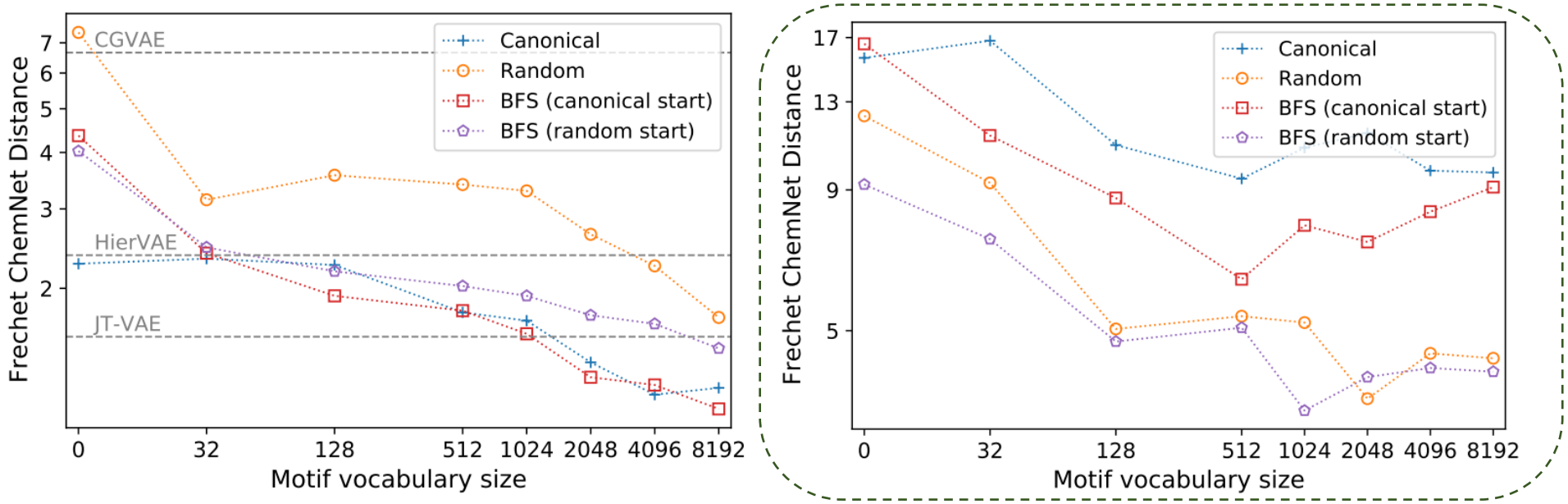
"Over the coming decade, deep learning looks set to have a transformational impact on the natural sciences. The consequences are potentially far-reaching and could dramatically improve our ability to model and predict natural phenomena over widely varying scales of space and time. Our AI4Science team encompasses world experts in machine learning, quantum physics, computational chemistry, molecular biology, fluid dynamics, software engineering, and other disciplines, who are working together to tackle some of the most pressing challenges in this field."

— Professor Chris Bishop, Technical Fellow, and Director, AI4Science

We're hiring!

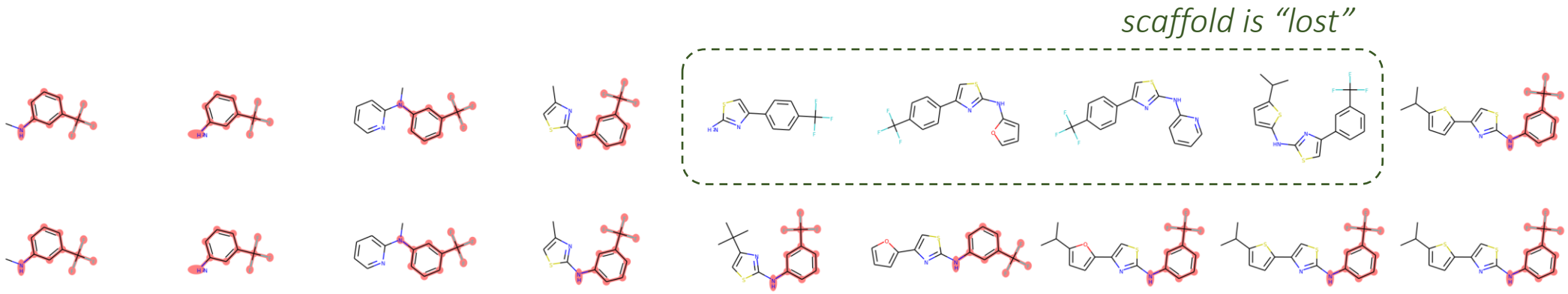
Backup slides

Quantitative results: generation order



randomized generation order crucial for working with scaffolds

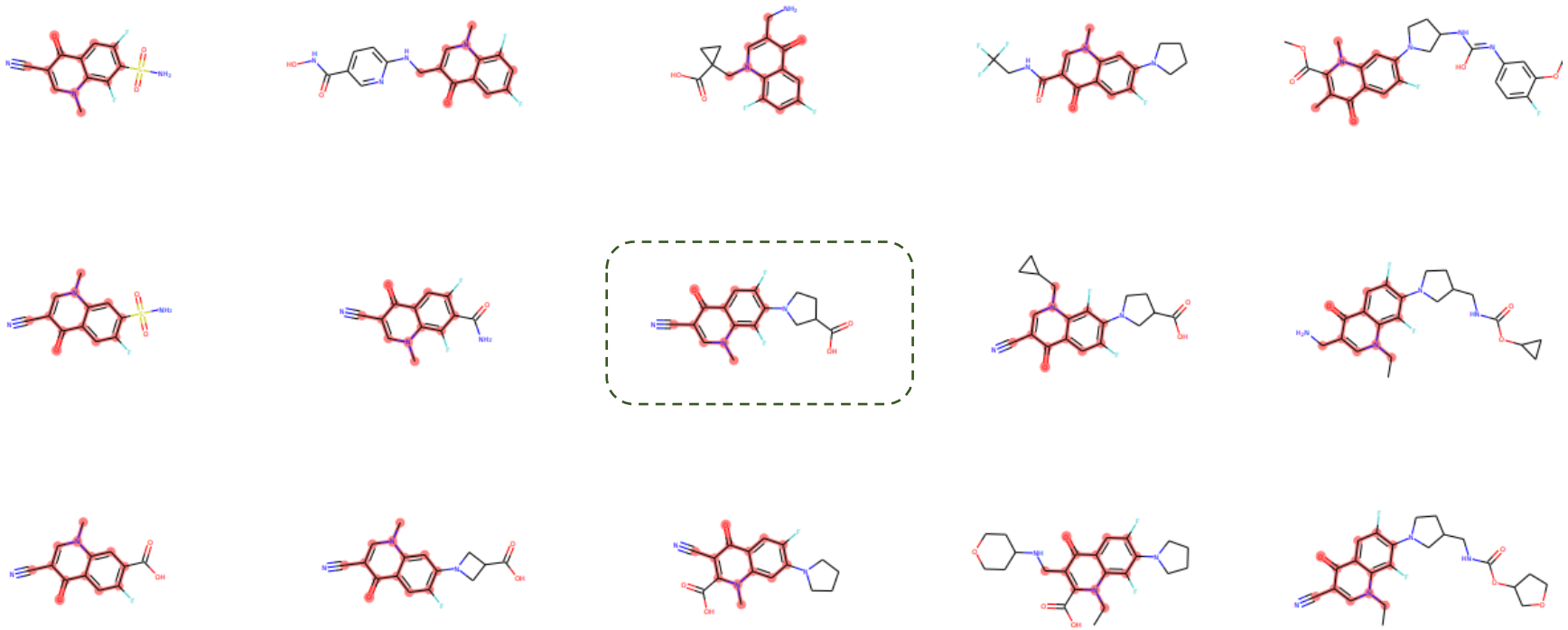
Qualitative results: interpolation



Interpolation without (top) and with (bottom) the scaffold constraint.

Qualitative results: latent neighborhood

We can visualize the scaffold-constrained neighborhood of a given mol.



Correlates with size