From data acquisition to ML model for predicting outcomes of chemical reactions

Molecule.one

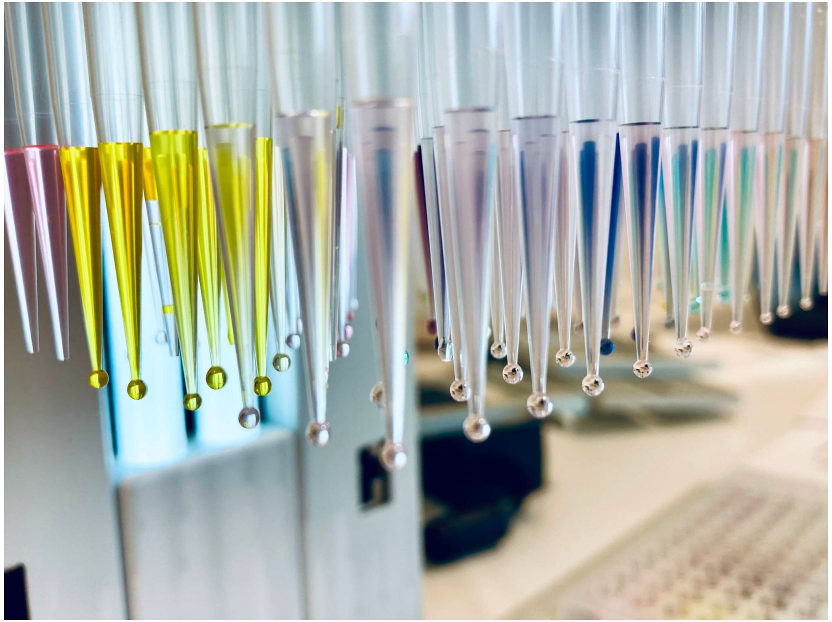
Michał Sadowski

Molecule.one: solving unpredictability of chemistry

Predicts chemical reactions with unprecedented accuracy for faster drug discovery

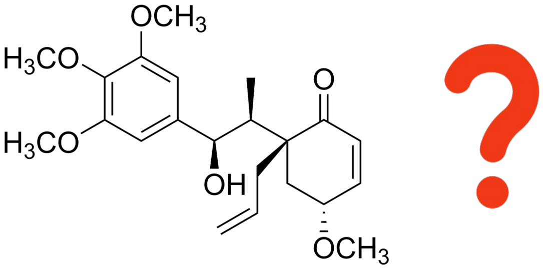
- We created the first deep learning based commercial software for synthesis planning (2019-2022)

- Currently focused training deep learning models on data from our own laboratory (2022)



Molecule.one

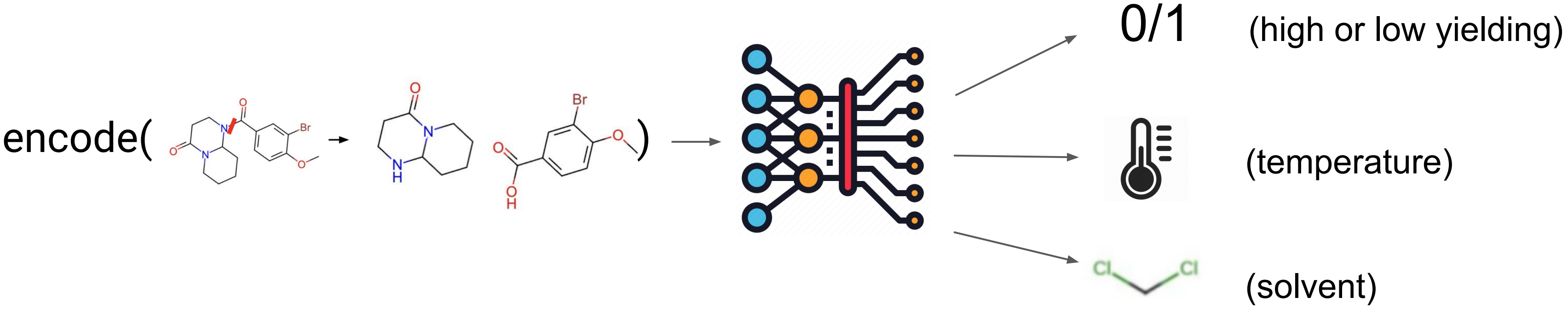
Bottlenecks in organic chemistry



Finding cheap synthesis plan is difficult

~50% of reactions fail

Deep learning for chemical reactions



encode( )

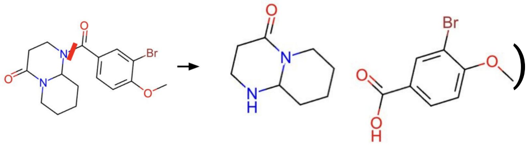
0/1 (high or low yielding)

(temperature)

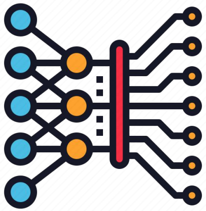
(solvent)

Deep learning for chemical reactions

encode(

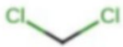


)



0/1 (high or low yielding)

(temperature)

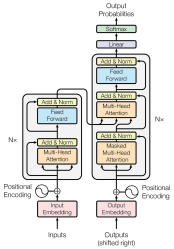


(solvent)

For example:

CC(=O)[OH].[OH]C>>CC(=O)OCC

SMILES notation

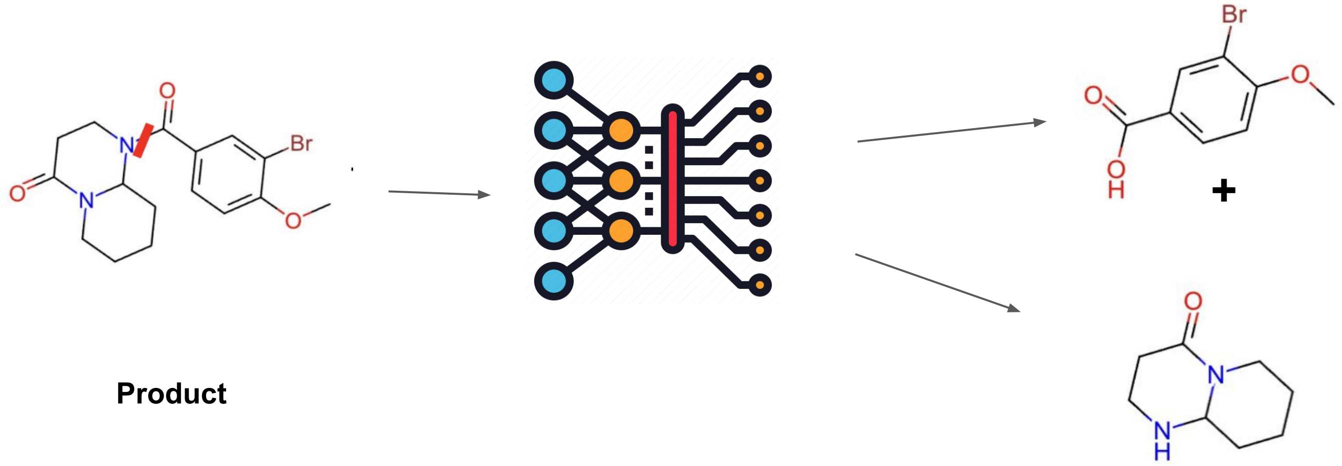


Transformer

0	4	2

Output tokens

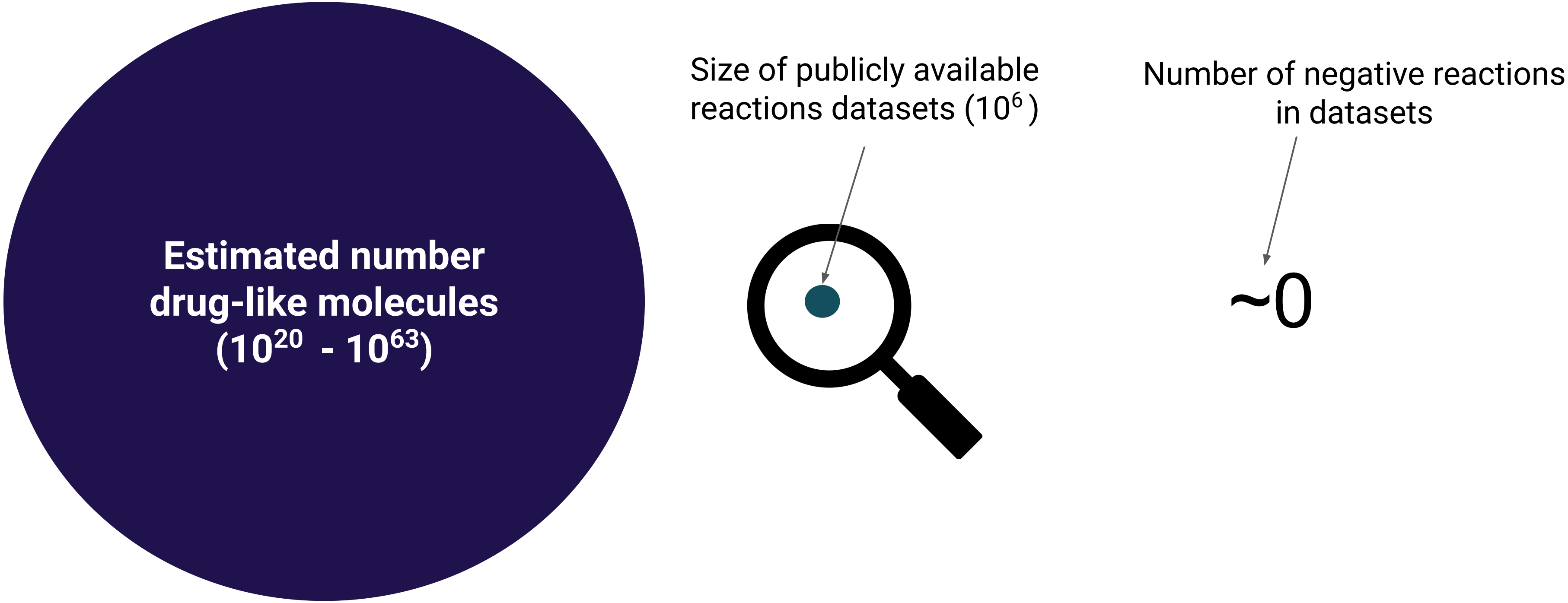
Finding synthesis plans with deep learning



Product

Substrates

Dataset sizes in chemistry are tiny



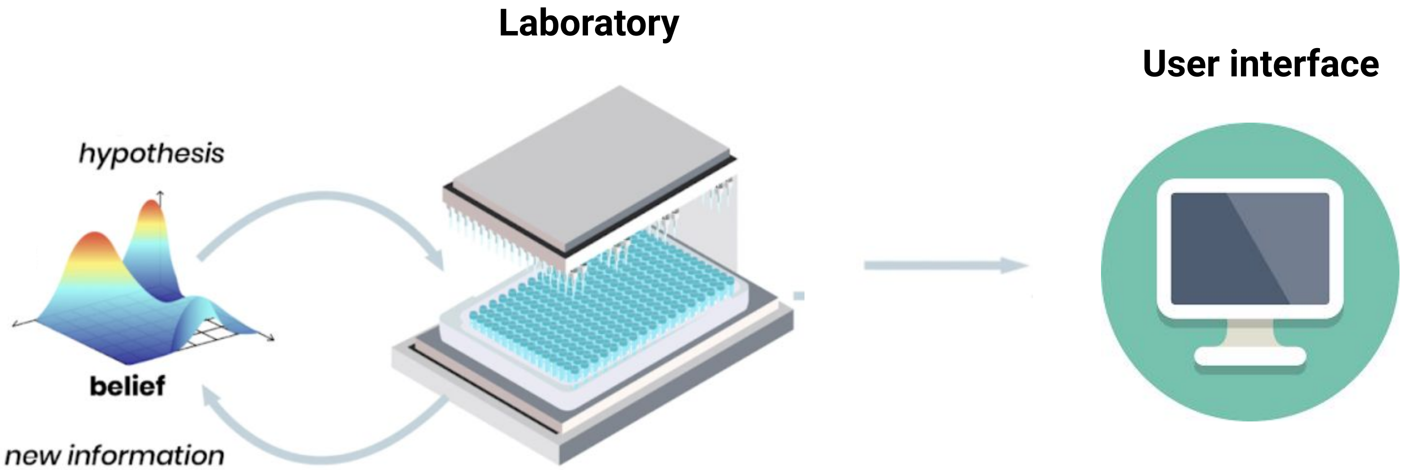
Estimated number drug-like molecules (10^{20} - 10^{63})

Size of publicly available reactions datasets (10^{6})

Number of negative reactions in datasets

~0

Solution: Closing the loop in organic chemistry



Laboratory

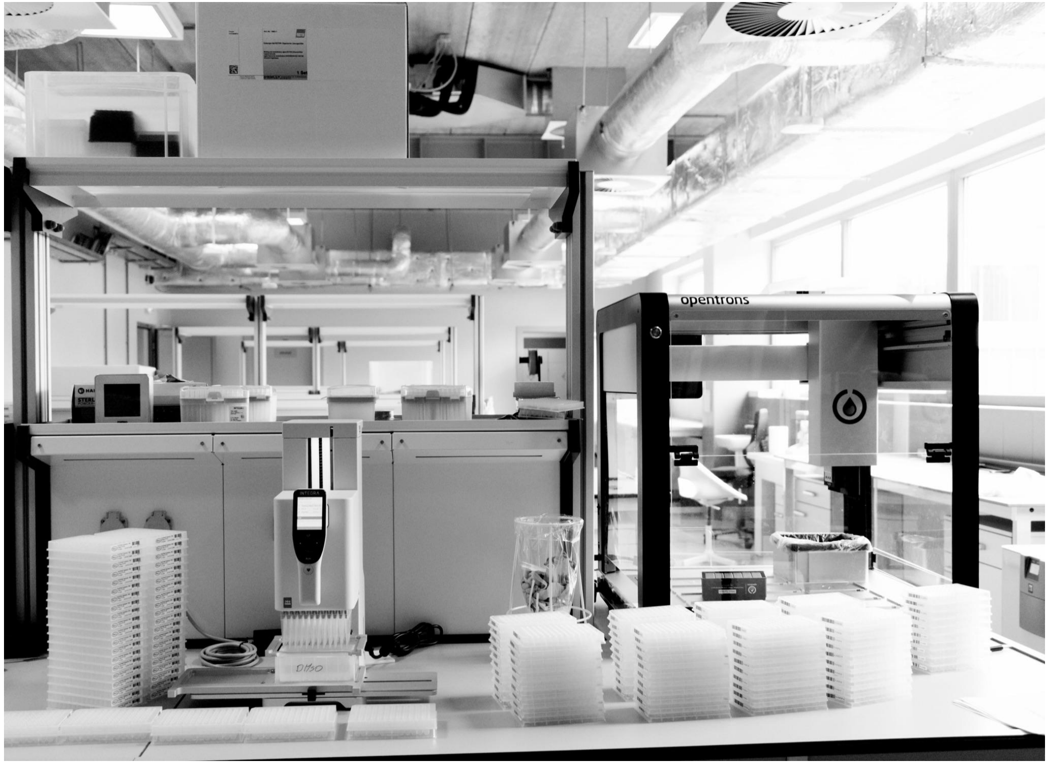
User interface

hypothesis

belief

new information

HT laboratory @ Molecule.one



- We set up our own laboratory
- Designed to perform diverse reactions based on automatic recommendations

HT laboratory @ Molecule.one

Link to the video

- It would not be possible without automation
- We are capable of performing a few thousand reactions per week
- It is more than a typical chemist performs in a year

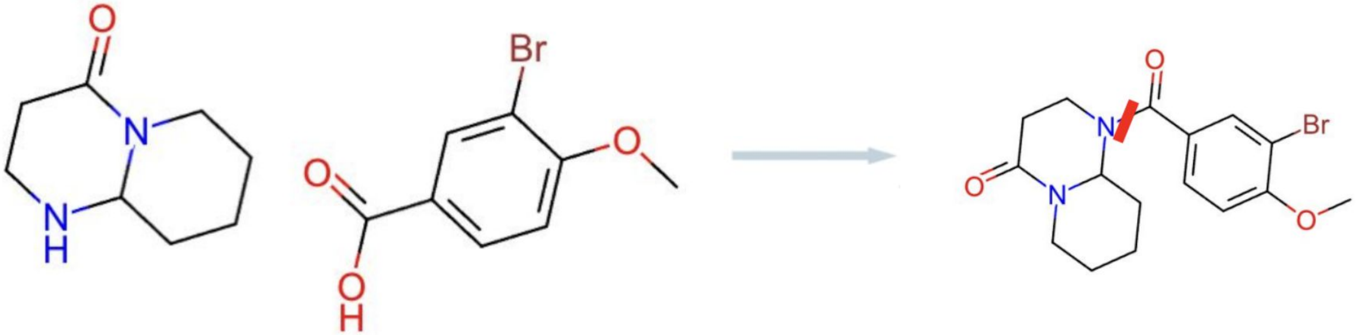
What reactions should we focus on?

Still it is impossible to generate enough data for model to predict every reaction

We’ve chosen type of reactions that are:

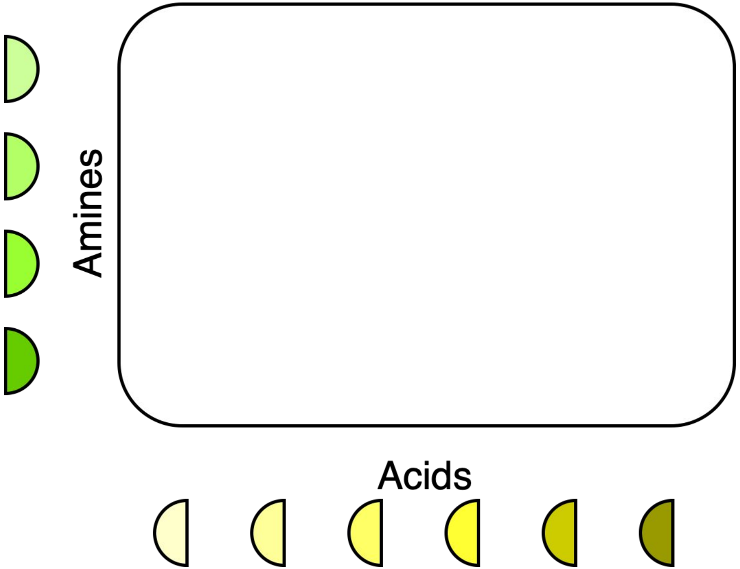
- Within the most commonly used reactions in MedChem
- Can be performed in a high-throughput laboratory
- Difficult to predict outcomes of reaction

First experiments - Amide couplings

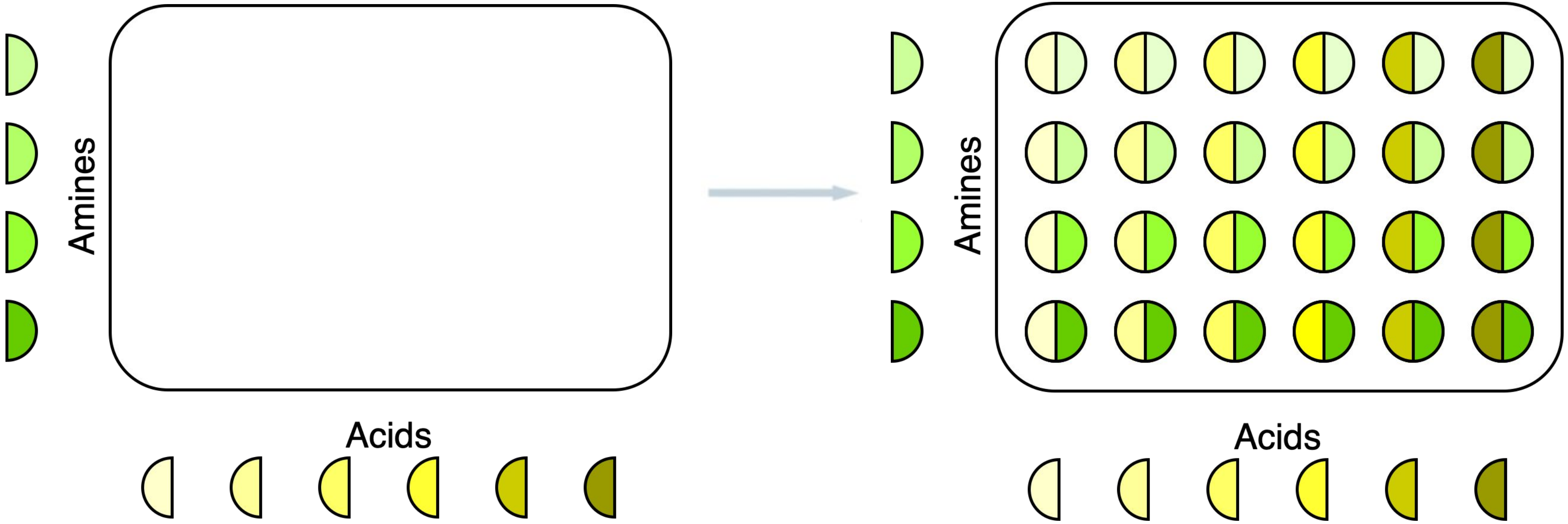


amine **carboxylic acid** **amide**

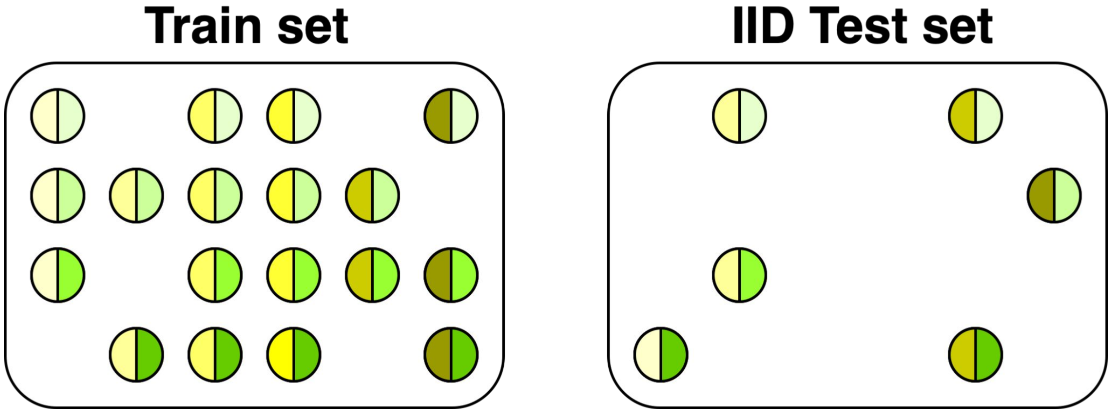
Grid dataset design



Grid dataset design



Model evaluation

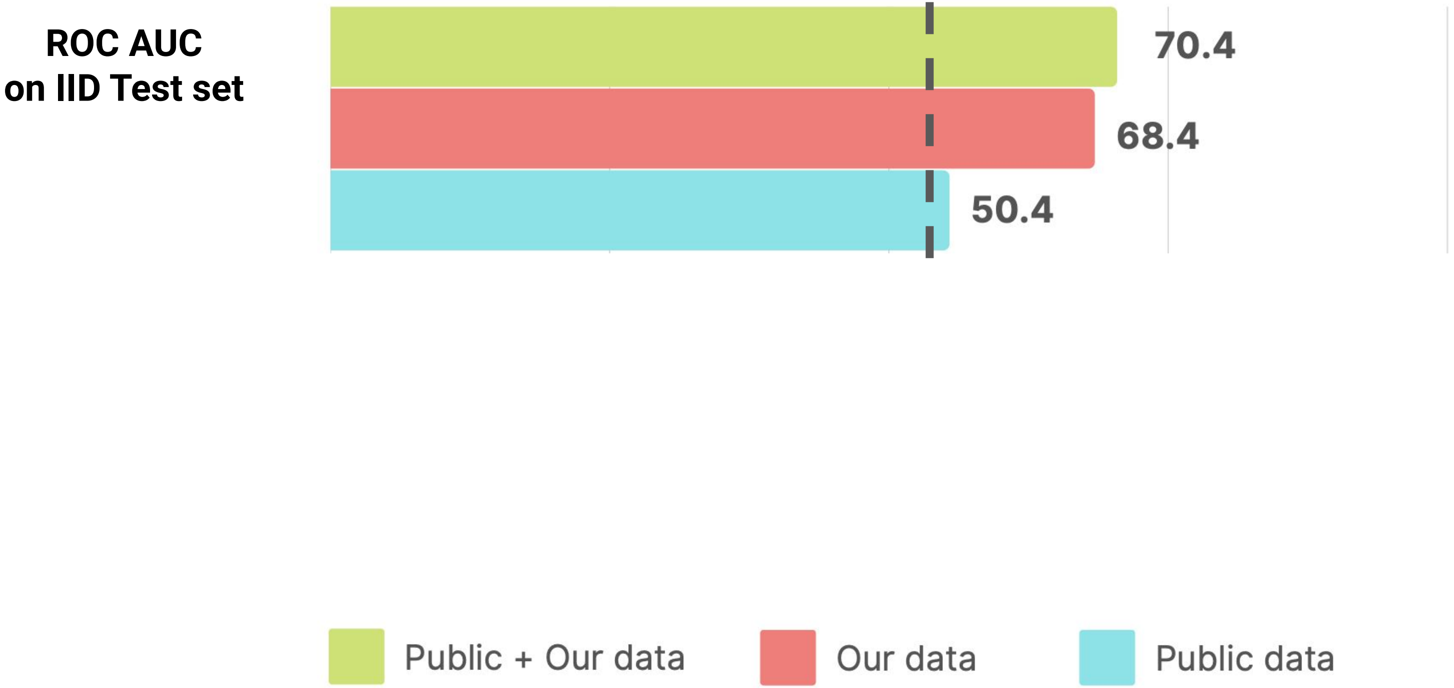


Train set

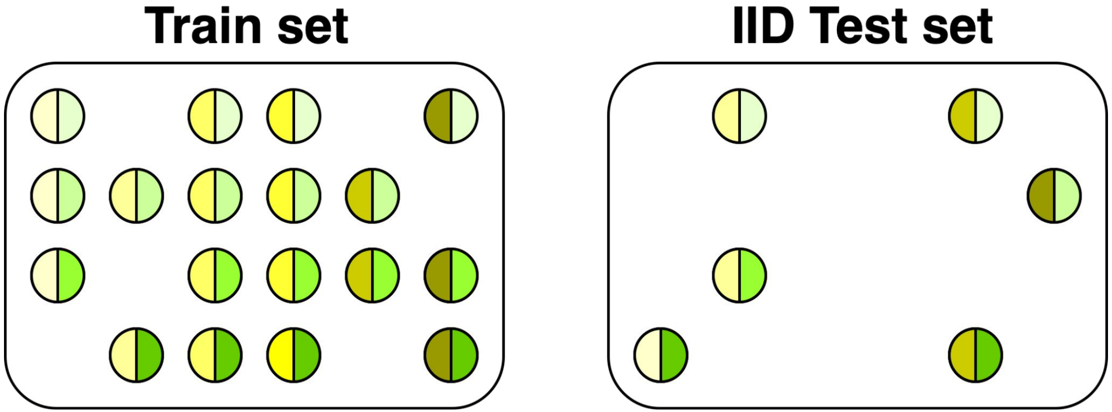
IID Test set

Evaluation on IID Test set

Using lab data enables achieving significantly better results than using public data



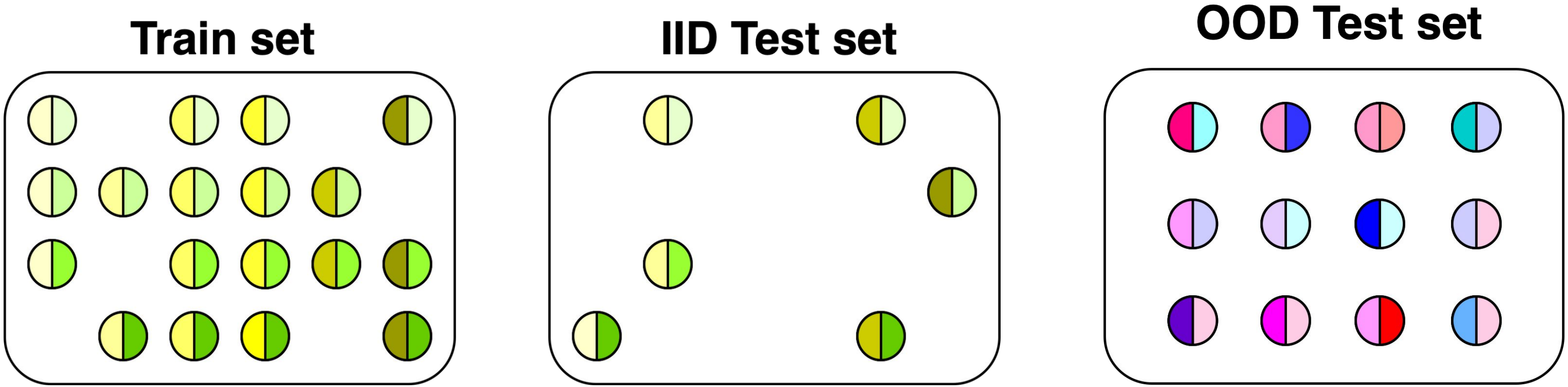
Model evaluation



Train set

IID Test set

Model evaluation - out of distribution

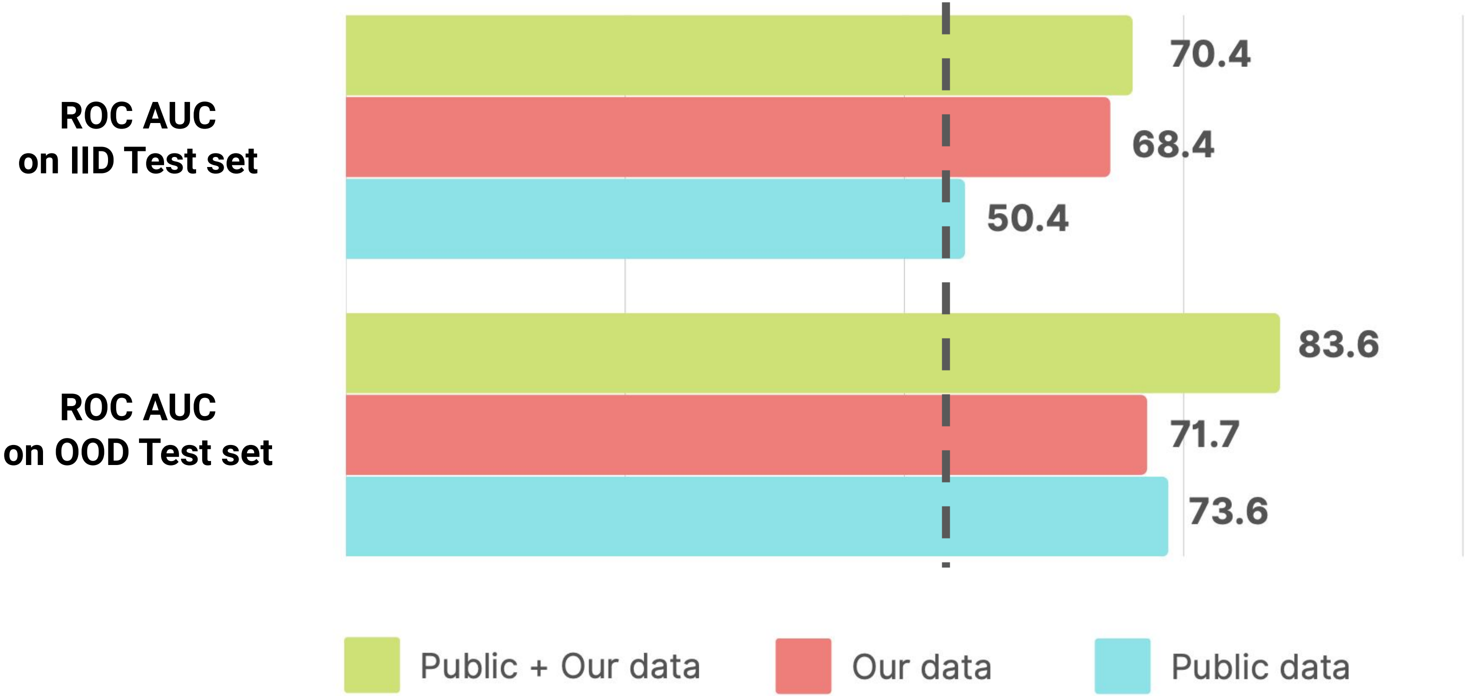


Train set

IID Test set

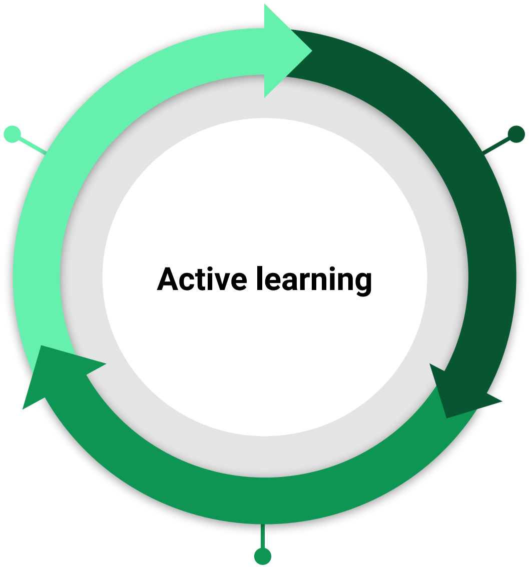
OOD Test set

Evaluation on IID and OOD Test set



Open challenges

- How to estimate the model uncertainty?
- How to preserve the variety of performed reactions?
- How to better predict reaction conditions?
- How to implement data pipelines and ensure data correctness?



Add performed reactions to the dataset

Active learning

Reactions selection
Choose reactions that are difficult to predict

Perform reactions in the lab

Thank you

- High failure rate of organic reactions is a key bottleneck in drug discovery
- Molecule.one solution: a closed loop with in-house wet lab focused on high throughput (performs in 1 week more reactions that a chemist usually performs in a year)
- First experiments show strong improvement of deep learning models; public data alone doesn't allow accurate predictions
- Many challenges to solve, especially related to active learning/continual learning!

Molecule.one

Thank you

- High failure rate of organic reactions is a key bottleneck in drug discovery
- Molecule.one solution: a closed loop with in-house wet lab focused on high throughput (performs in 1 week more reactions that a chemist usually performs in a year)
- First experiments show strong improvement of deep learning models; public data alone doesn't allow accurate predictions
- Many challenges to solve, especially related to active learning/continual learning!

Molecule.one

We are hiring!