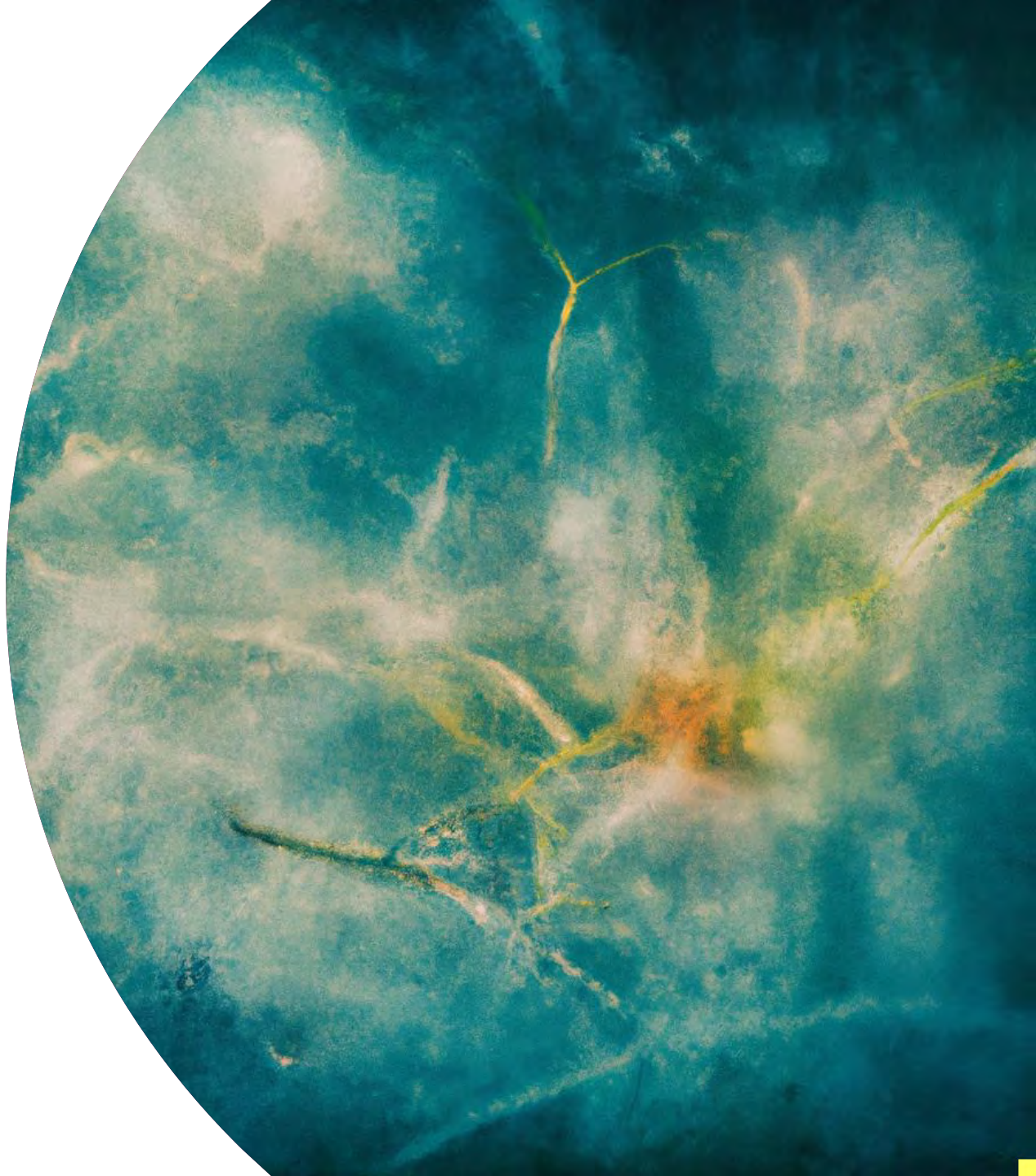
Paradigm shift in synthesis planning: how to exploit artificial intelligence

Samuel Genheden

MolecularAI, Discovery Sciences, AstraZeneca R&D, Gothenburg

2024-09-16



Acknowledgements


Annie Westerlund


Lakshidaa Saigiridharan


Mikhail Kabeshov


Thierry Kogej


Christos Kannas


Ola Engkvist


Varvara Voinarovska


Peter Hartog


Emma Rydholm

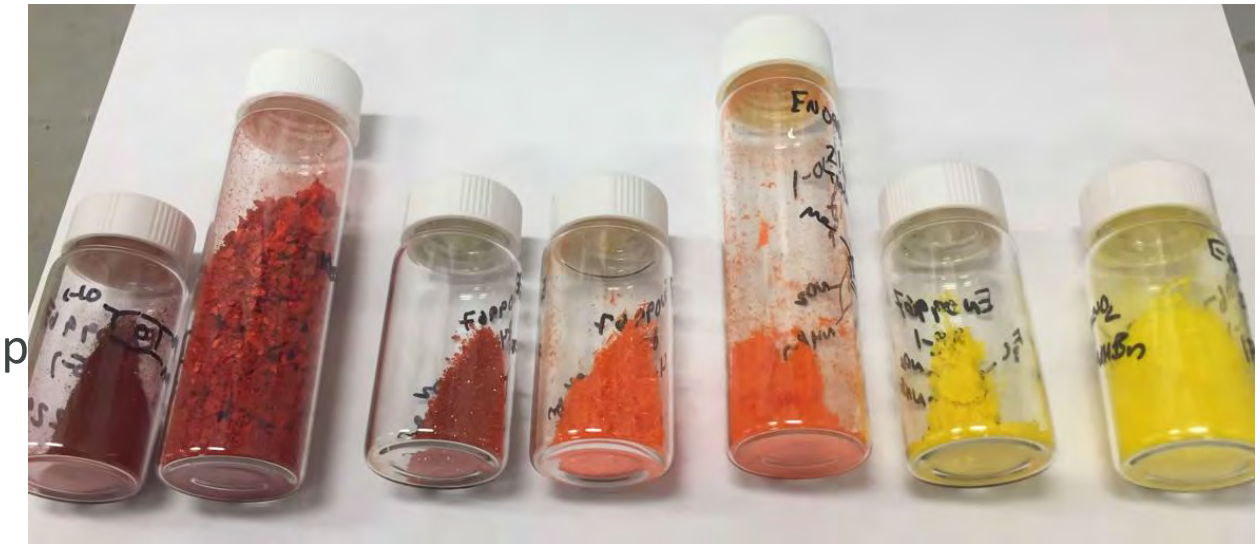

Supriya Kancharla


Siva Manohar


Rocío Mercado

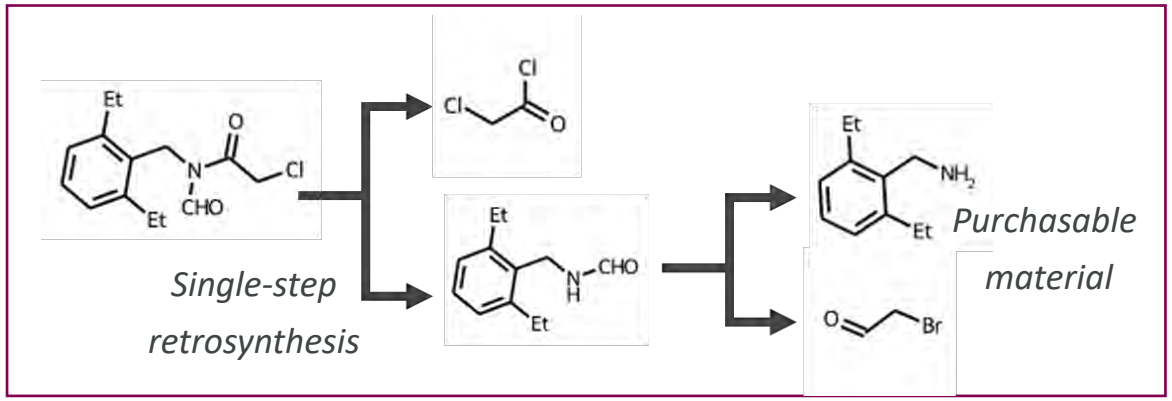
Synthetic Chemistry in Pharma

- Medicinal chemistry:
  - Invent a molecule to treat human disease
  - Design & synthesize new molecules: 1000s per p
  - Small scale: <0.005 g to 50 g
  - Diversity-oriented
  - Speed is key
- Process chemistry:
  - Make large quantities of that molecular for studies and clinic
  - Deliver large quantities of API for studies
  - 50 g to >1,000,000 g scale
  - Purity, sustainability and cost are key

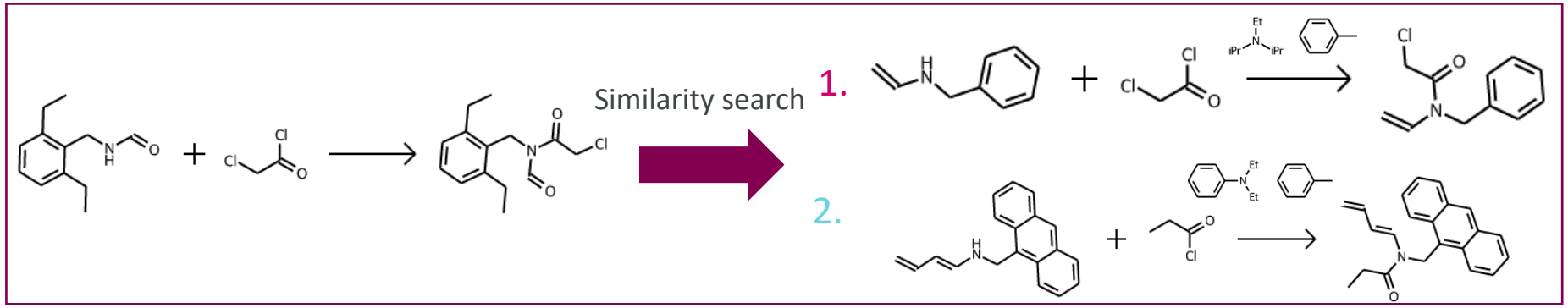


Computer-aided synthesis planning

- Retrosynthesis analysis
  - One-step retrosynthesis
  - Multi-step retrosynthesis or route prediction



- Database searches
  - Support for predictions
  - Experimental details



Software landscape

Free solutions

ASKCOS

Massachusetts Institute of Technology

AiZynthFinder

AstraZeneca

IBM RXN for Chemistry

+ Many GitHub repositories

Commercial solutions

Reaxys®

CAS

SYNTHIA™ Retrosynthesis Software

Molecule.one

spaya

Reaction data landscape

Proprietary sources

Reaxys

- > 30 M reactions
- 90% from research articles

NextMove Software

- > 10 M reactions
- US, European and Asian patents

ELN records

- Closed company resource
- AZ ELN is > 2 M reactions

CAS

- > 100 M reactions
- Rarely used for ML

Public-domain sources

US patents

- > 2M reactions
- Open-source extract from US patents
- Questionable quality
- Up to 2016

ORD

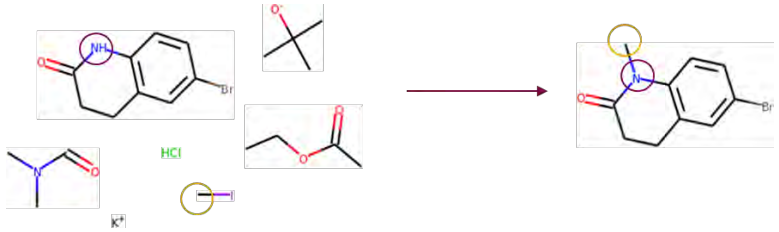
ORD

- > 2M reactions
- Initiative from academia with industry support
- Open data format
- Contains US patents and HTE datasets

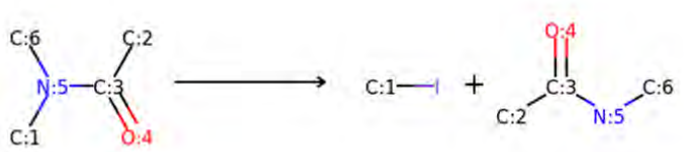
AiZynthFinder and template-based retrosynthesis

Reaction templates

1. Databases provide atom-mapped reactions



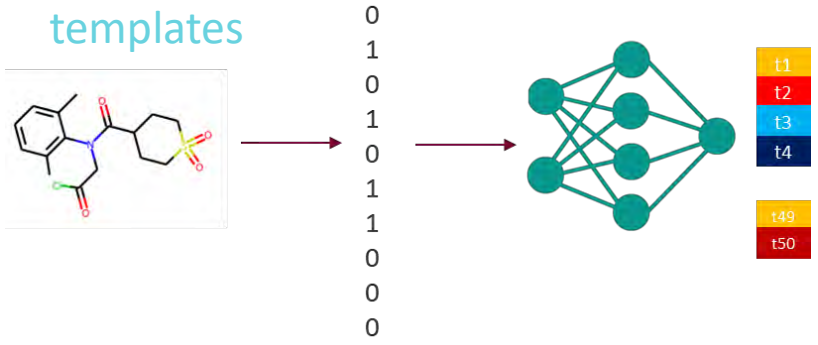
2. Extract chemical transformation rules: templates



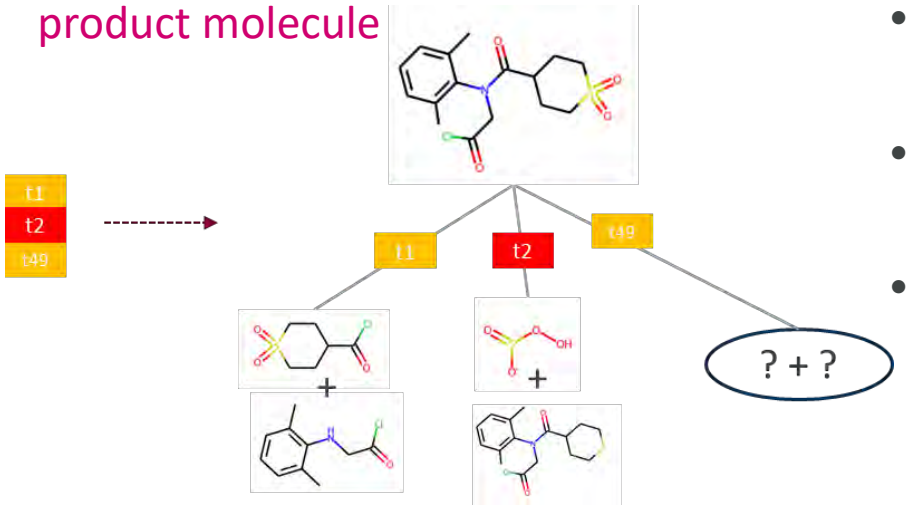
github.com/MolecularAI

One-step retrosynthesis

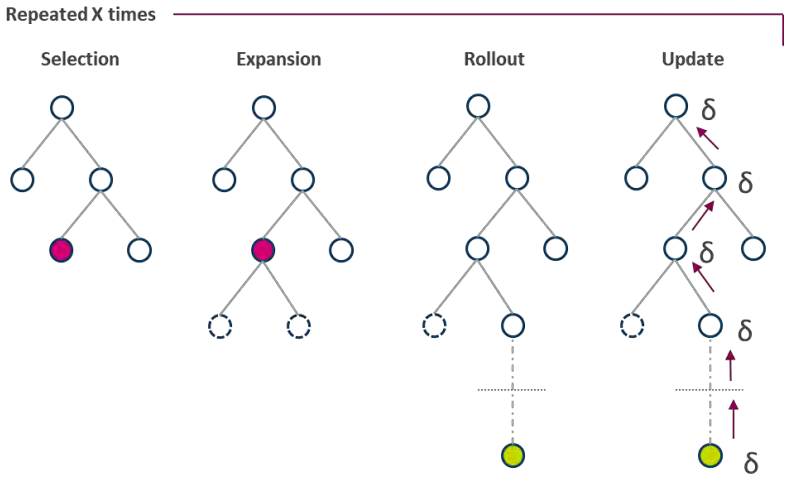
1. Train a neural network to rank templates



2. Apply top-ranked templates to product molecule



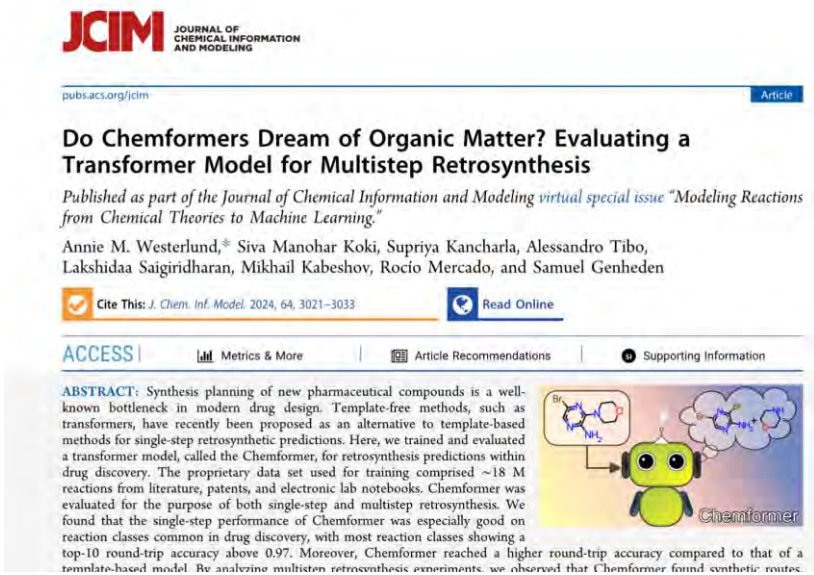
Monte-Carlo tree search



- Use a search algorithm to iteratively find a solution
- Uses one-step model at each iteration to expand the search tree
- Different scores to balance exploitation and exploration

#

Template-free model: Chemformer



JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

Do Chemformers Dream of Organic Matter? Evaluating a Transformer Model for Multistep Retrosynthesis

Published as part of the Journal of Chemical Information and Modeling virtual special issue "Modeling Reactions from Chemical Theories to Machine Learning."

Annie M. Westerlund,* Siva Manohar Koki, Supriya Kancharla, Alessandro Tibo, Lakshidaa Saigiridharan, Mikhail Kabeshov, Rocío Mercado, and Samuel Genheden

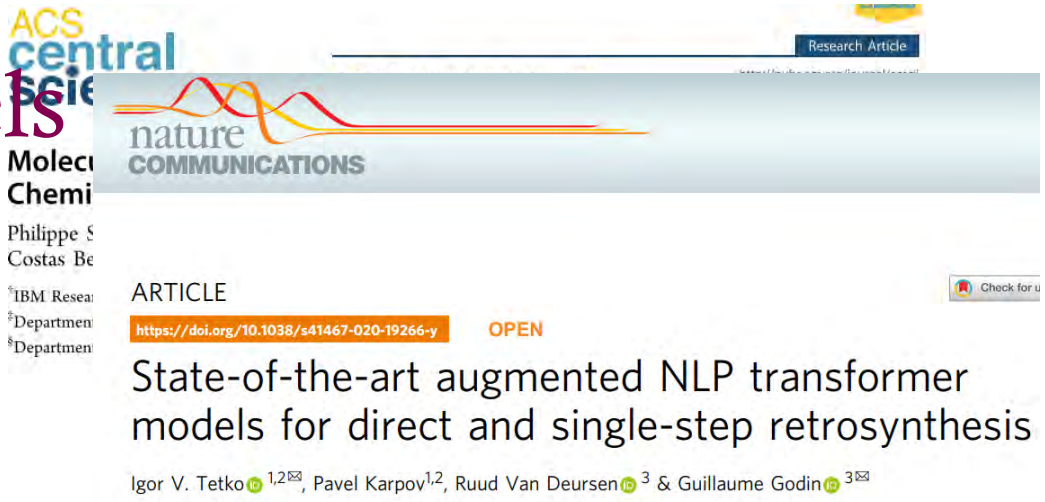
Cite This: J. Chem. Inf. Model. 2024, 64, 3021−3033 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Synthesis planning of new pharmaceutical compounds is a well-known bottleneck in modern drug design. Template-free methods, such as transformers, have recently been proposed as an alternative to template-based methods for single-step retrosynthetic predictions. Here, we trained and evaluated a transformer model, called the Chemformer, for retrosynthesis predictions within drug discovery. The proprietary data set used for training comprised ~18 M reactions from literature, patents, and electronic lab notebooks. Chemformer was evaluated for the purpose of both single-step and multistep retrosynthesis. We found that the single-step performance of Chemformer was especially good on reaction classes common in drug discovery, with most reaction classes showing a top-10 round-trip accuracy above 0.97. Moreover, Chemformer reached a higher round-trip accuracy compared to that of a template-based model. By analyzing multistep retrosynthesis experiments, we observed that Chemformer found synthetic routes

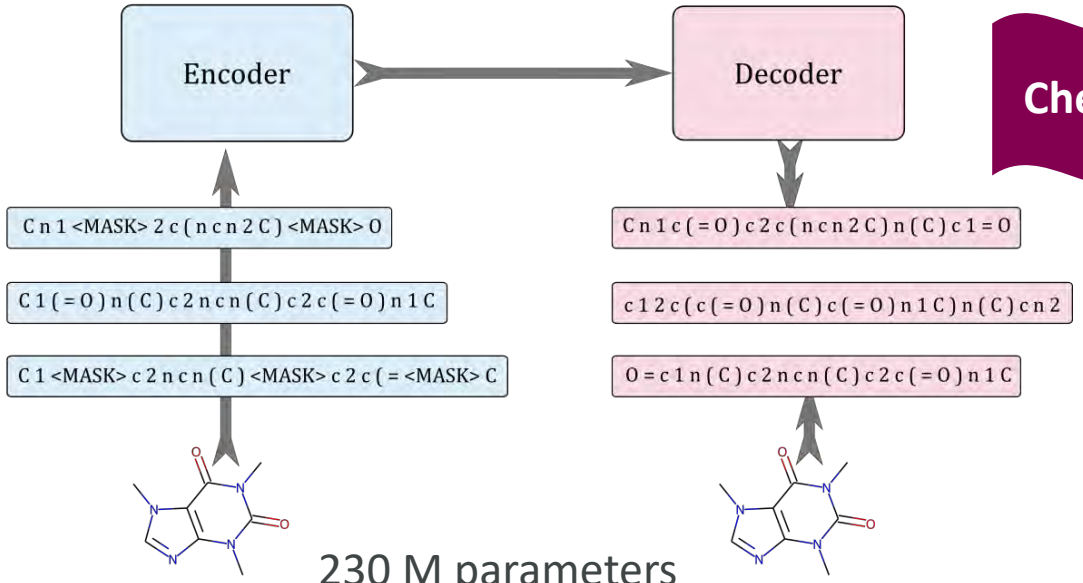
Westerlund et al. J. Chem. Inf. Model. 2024, 64, 3021−3033

Template-free / Generative models



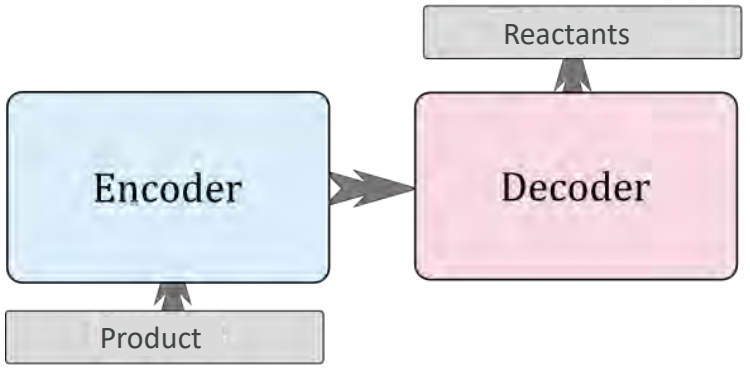
- Treat retrosynthesis as a language problem
- Translating from product to reactants
- Model will learn chemical rules *and* the prioritization of those rules

1. Pretrain a *transformer* model (BART) to learn SMILES



230 M parameters

2. Fine-tune a model to perform retrosynthesis



On the promises of template-free models

- Template-free retrosynthesis models do not need
  - Reactions with atom-mapping
  - Extracted templates
- Better scaling to larger chemical spaces
- Possibility to extrapolate outside known space

Data sources and trained models

Reaxys®

- > 30 M reactions
- 90% from research articles

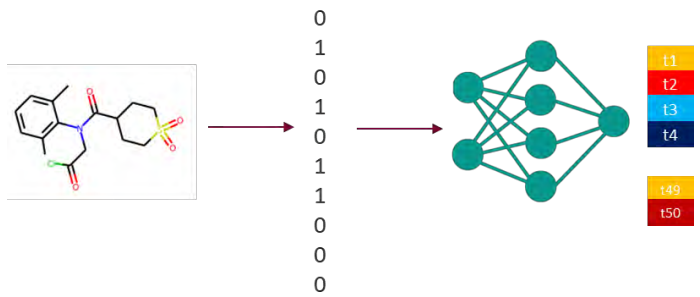
NextMove Software

- > 10 M reactions
- US, European and Asian patents

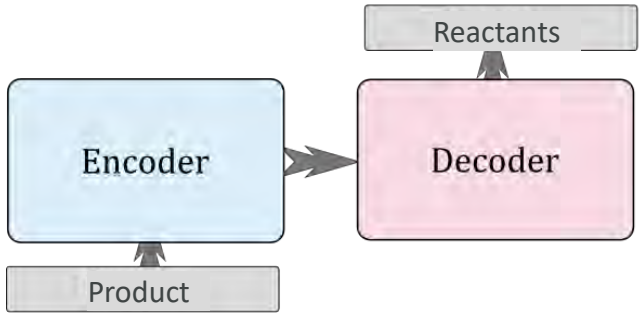
ELN records

- AZ ELN is > 2 M reactions

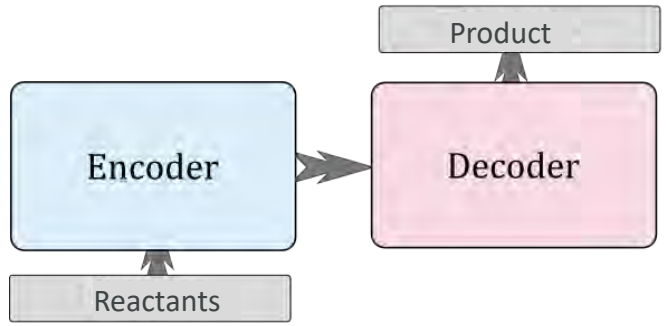
18M unique reactions



Template-based model(s)



Chemformer retrosynthesis



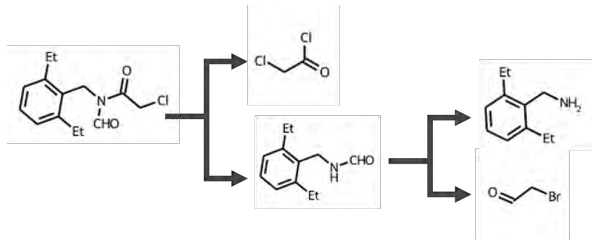
Chemformer forward

(for validation)

Chemformer vs Template-based: Multi-step performance

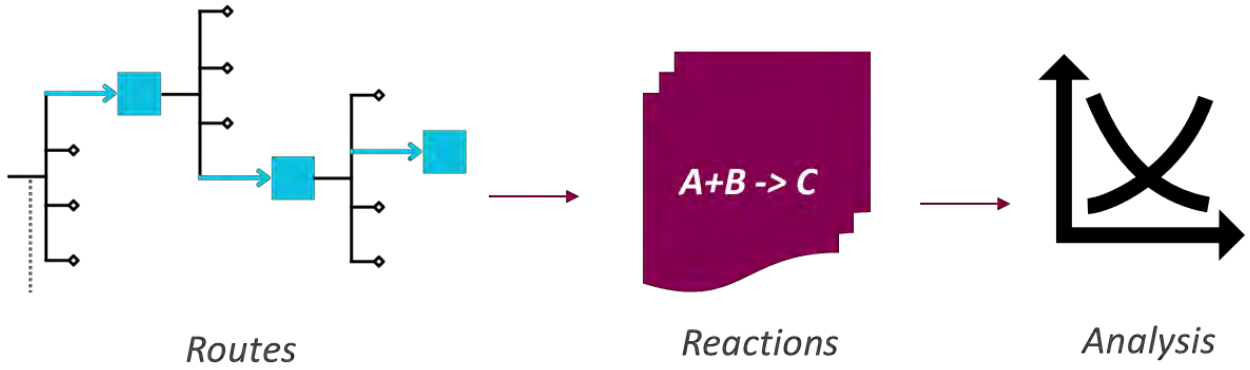
- Evaluate route prediction on 5K compounds designed by AZ chemists
- Stock is AstraZeneca internal building block collection (5M)
- Run 100 iteration of retrosynthesis
- Implement novel algorithms for speeding-up Chemformer performance

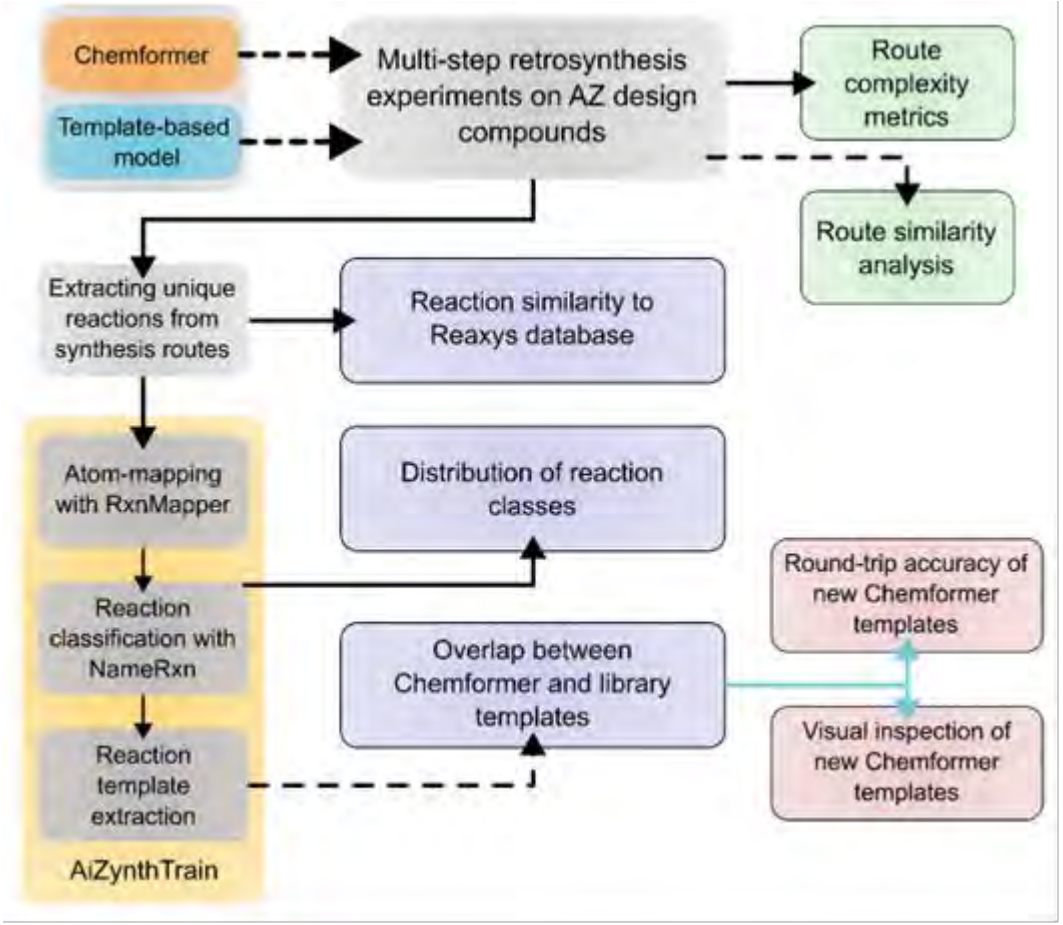
Can we find a route to purchasable material?



Solvability case	Percentage
Solved by both	71.6
Only solved by chemformer	23.4
Only solved by template-based	0.74
Solved by none	4.2
Solved by any	95.8

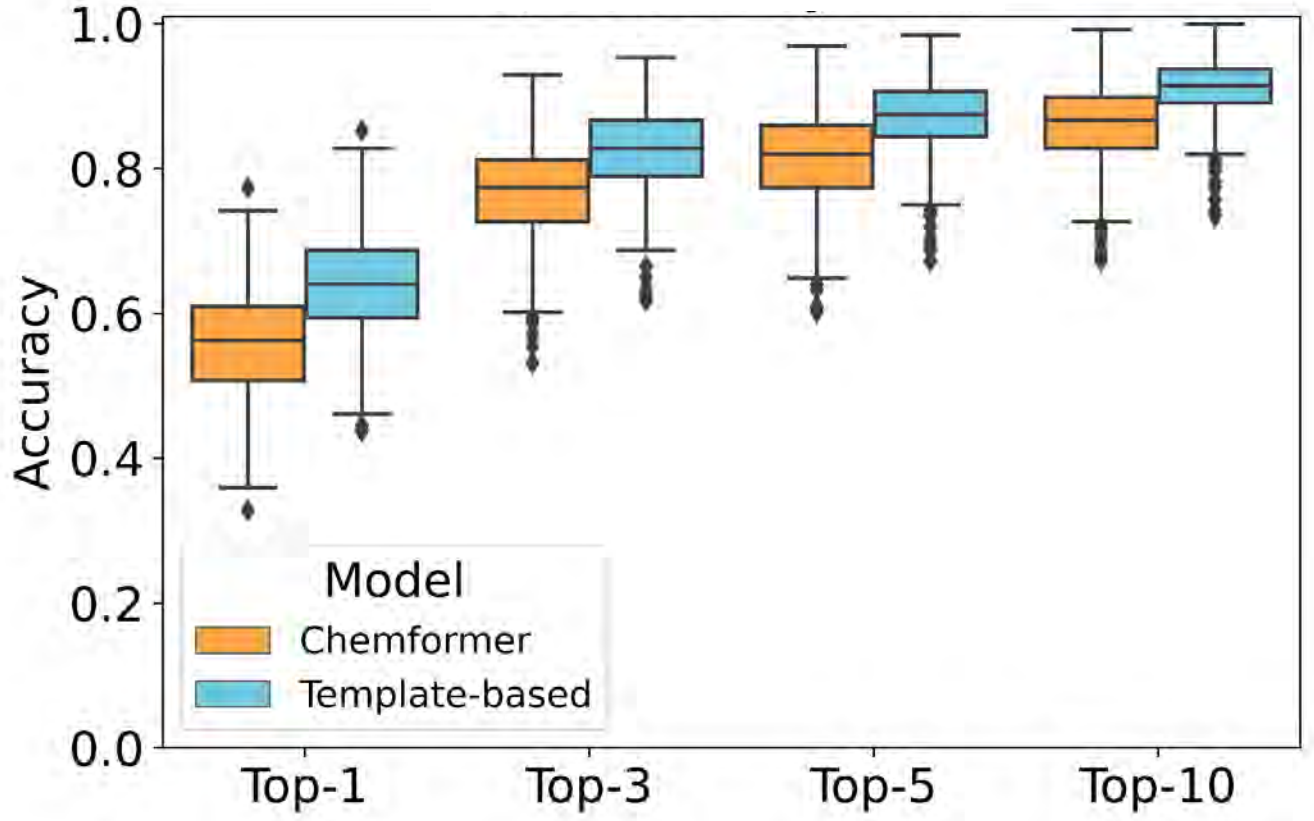
Do Chemformers dream of organic matter?





Are the suggested reactions feasible?

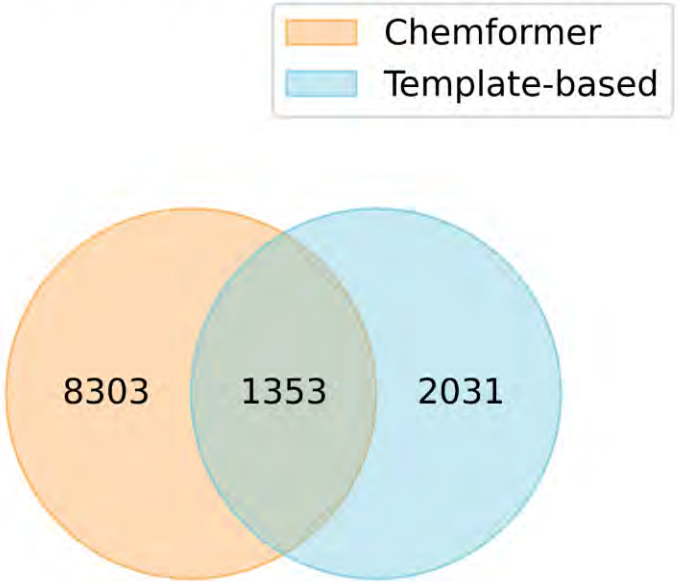
Round-trip accuracy calculated with forward Chemformer, showing how feasible the predicted reactions are



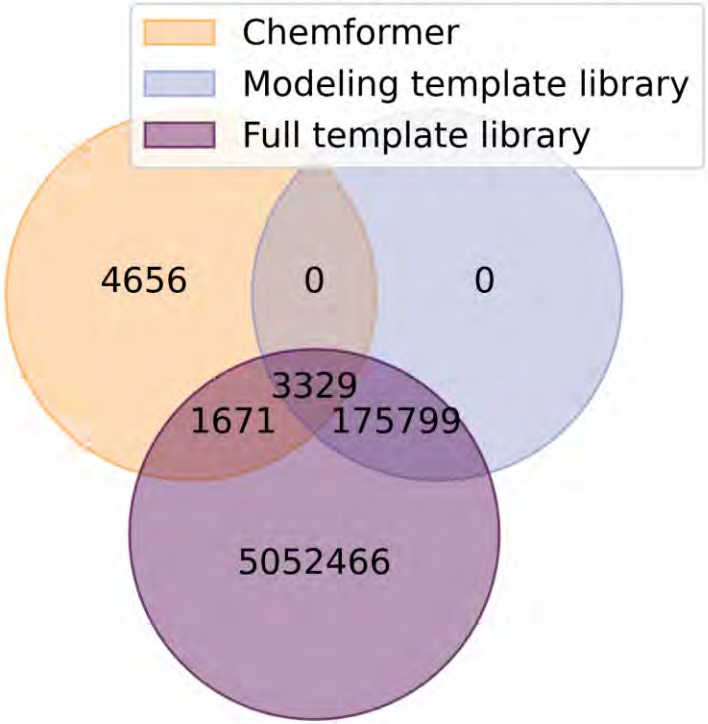
Are the models exploring different chemical space?

Extract templates from predicted reaction and look at space of unique templates

a) Chemformer vs. template-based



b) Chemformer vs. library



What is ahead of us?

- Chemformer appears to be a viable option to template-based retrosynthesis
- Hallucination occurs, but is perhaps acceptable
  - Hard to evaluate plausibility on a large scale
- Two major issues
  - Computational speed (engineering)
  - Interpretability and experimental support (science)
- Possible solution: a two-staged approach
  1. Predict with template-based retrosynthesis
  2. If fails to find synthetic route, try again with Chemformer

#

Constrained retrosynthesis

Constrained synthesis planning with disconnection-aware transformer and multi-objective search

24 May 2024, Version 1

Working Paper

Annie M. Westerlund, Lakshidaa Saigiridharan, Samuel Genheden

Show author details

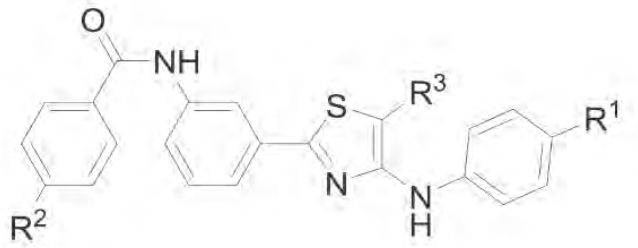
This content is a preprint and has not undergone peer review at the time of posting.

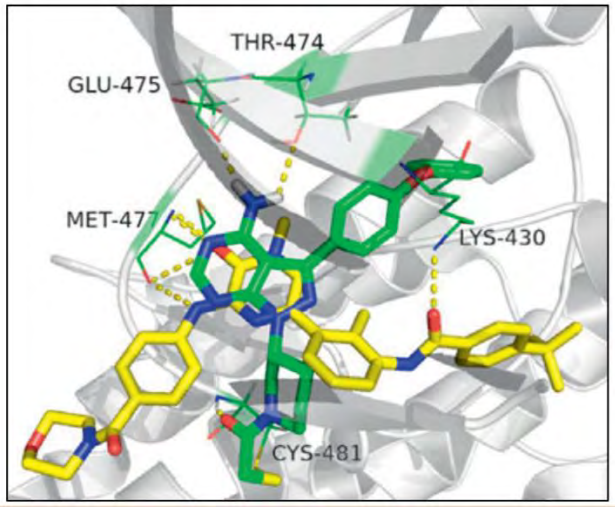
Download | Cite | Comment

Abstract

Designing synthesis routes with shared intermediates for a set of target compounds is a common task in drug
tep retrosynthesis tools such as AiZynthFinder are frequently used by chemists to generate
Although these tools can find solved routes for a majority of target compounds, they may not
vhich comply with specific bond constraints. Such bond constraints could be defined in the
synthesis routes with common intermediates for the set of compounds. Here, we present a novel
proach which aims to generate routes in the feasible region defined by these constraints. The
ivided into bonds to break and bonds to freeze. First, we introduce a filter in the search which

Synthesis of compound series



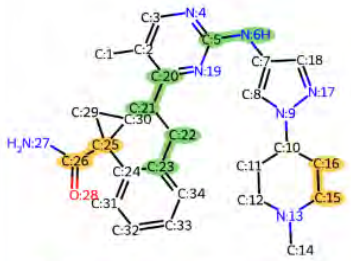


- In drug discovery you often explore series of related compounds
- Investigate structure-activity relationship for one or more vectors
- Advantageous to synthesis these compounds in similar fashion
- This is not supported by available retrosynthesis tools

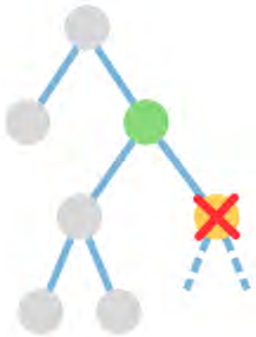
Guo et al. Eur. J. Med. Chem. 2019, 178, 767–781.
https://doi.org/10.1016/j.ejmech.2019.06.035.

Constrained synthesis planning

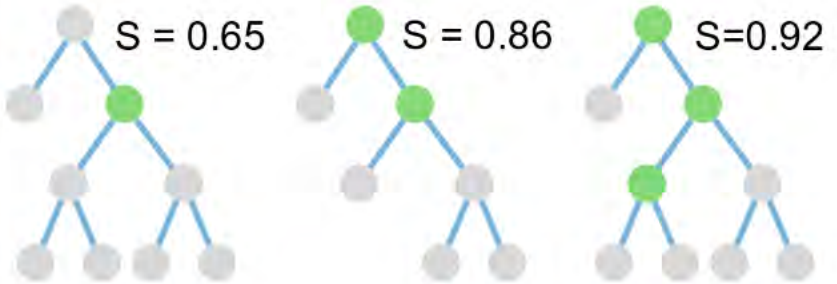
User define constraints



Frozen bonds violations rejected

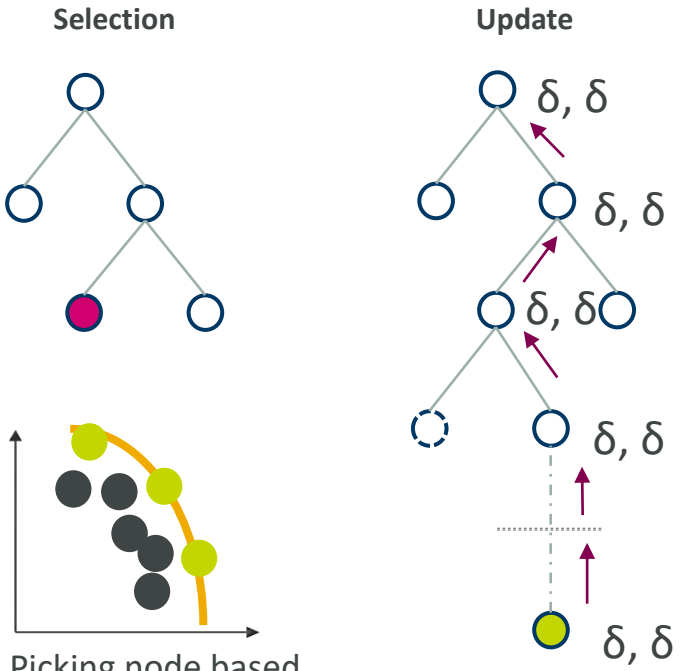


Score routes based on when bonds are broken

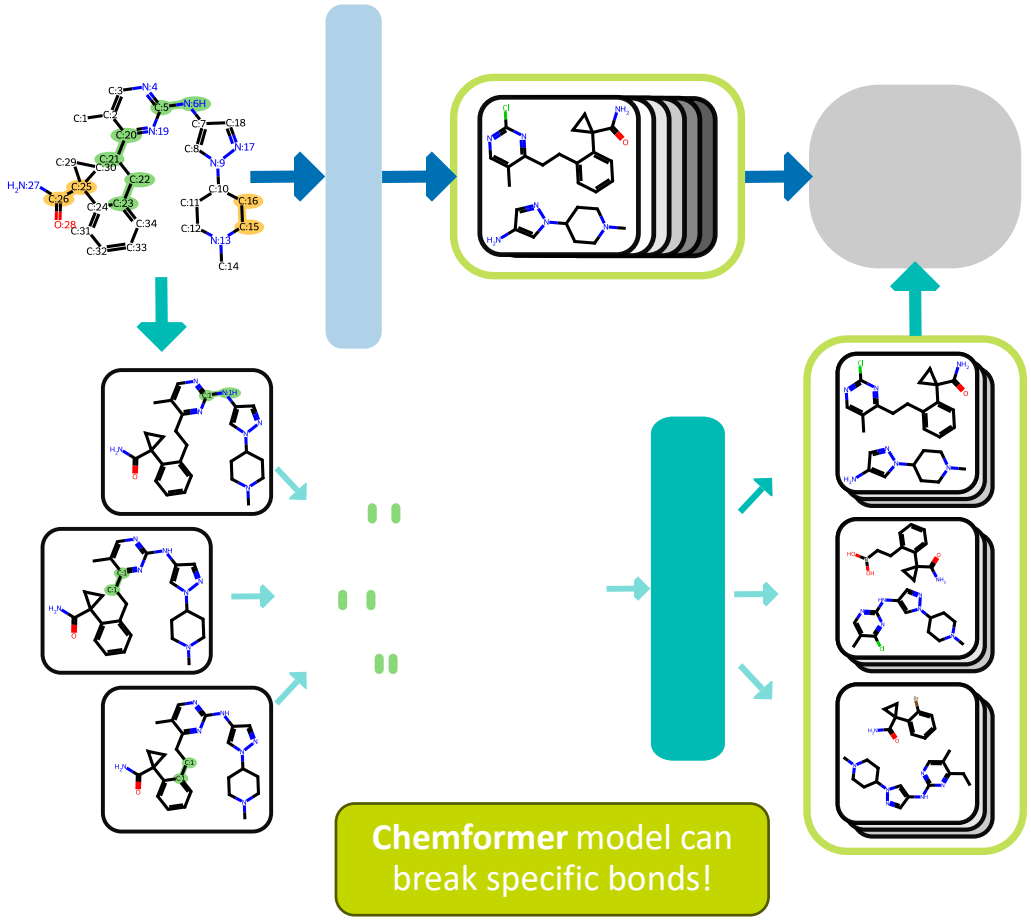


Novel techniques for enhancing bond-breaking

Selecting and scoring routes with multiple objectives

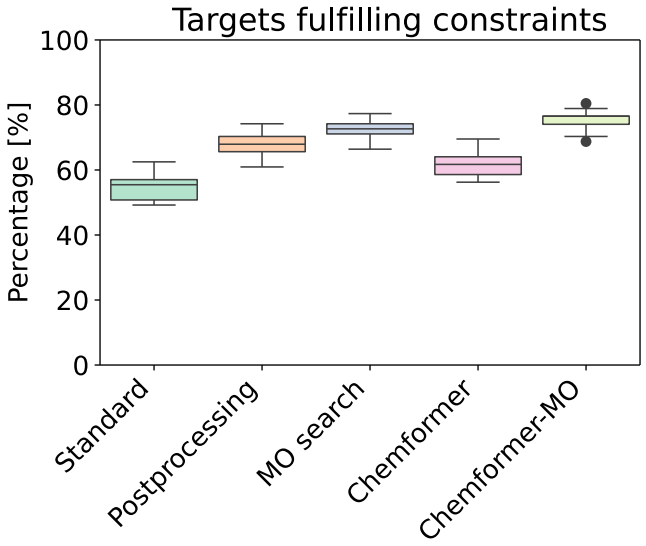


Enriching templated-based model with direct bond breaking

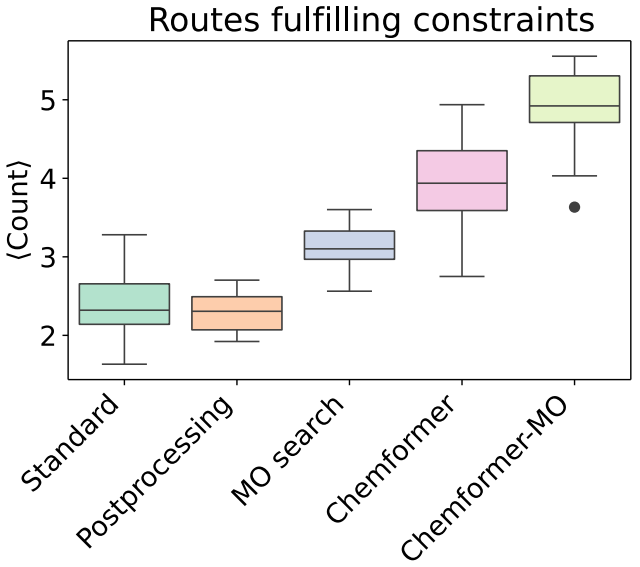


Benchmarking on synthetic data

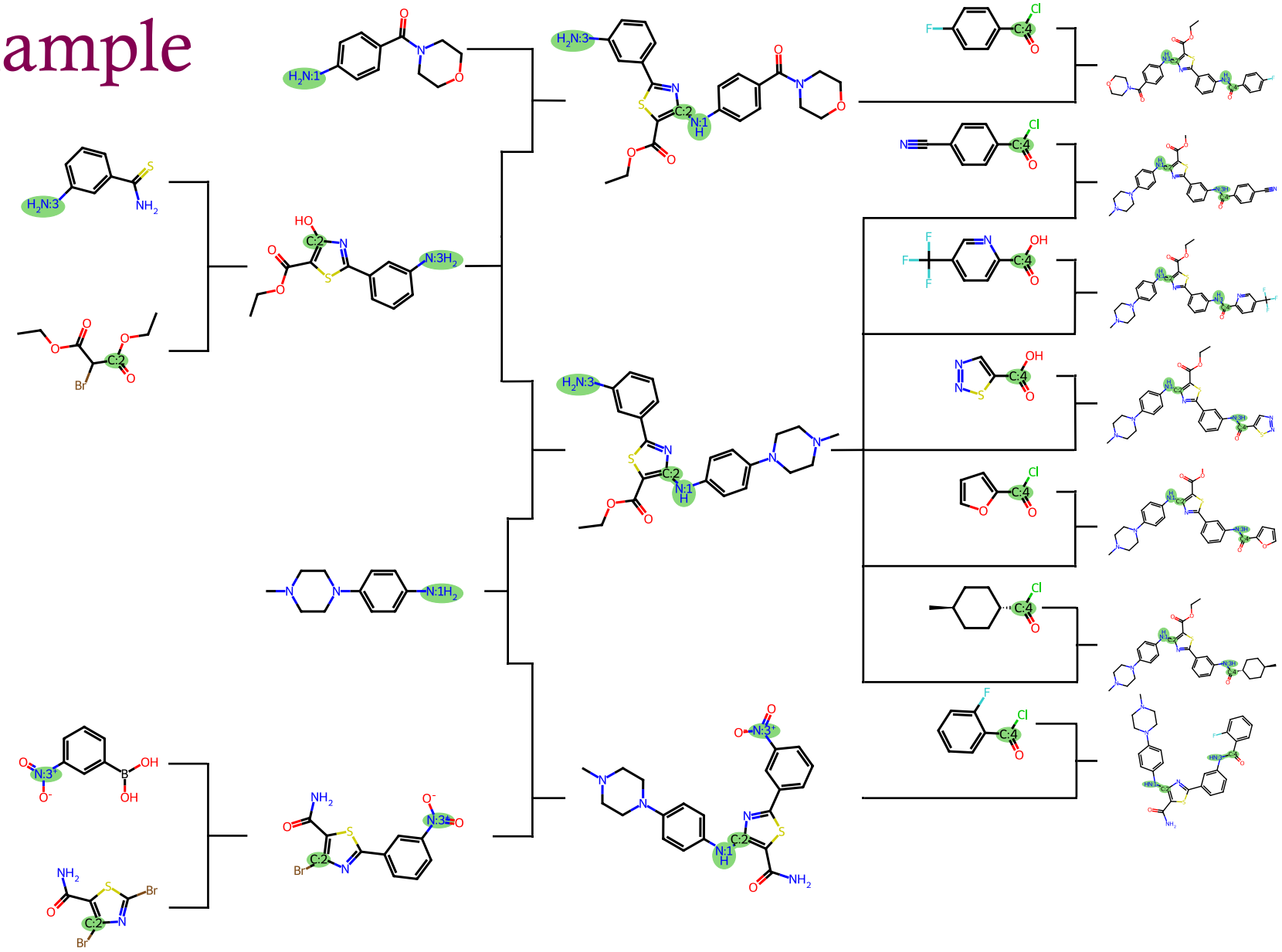
- Extracted synthesis routes from patents and J Med Chem
- Constructing bonds to break or freeze
- Compare five approaches to treating bonds to break



- We can increase the score for bond breaking by steering the search with multi-objective search
- We can increase the number of routes satisfying the constrained with a Chemformer model



Real-world example



Future outlook

Challenge 1: route scoring and comparison

- AiZynthFinder has been used for 4 years in production and have predicted synthetic routes for hundred of thousands of compounds
- … but we don’t know if the various improvements made have had a significant effect
- Comparing routes are difficult and mean different things to different people
- What is the best reference set to use for benchmarking?
- Should one compare routes, or should one just score new predictions and monitor an improvement in the score over time?

Challenge 2: human-like routes

- The end-goal is that we predict human-like routes with a high confidence of success that can directly taken to the lab
- Currently, we use retrosynthesis as an idea generation
- Currently, route predictions overuse protection/deprotection steps in an unbalanced way, order of steps are non-optimal, steps does not make sense from a forward synthesis perspective, etc.
- To bridge this gap, we need novel algorithms, scores, models etc

Acknowledgements


Annie Westerlund


Lakshidaa Saigiridharan


Mikhail Kabeshov


Thierry Kogej


Christos Kannas


Ola Engkvist


Varvara Voinarovska


Peter Hartog


Emma Rydholm


Supriya Kancharla


Siva Manohar


Rocío Mercado

Thanks!