USI/SUPSI Istituto Dalle Molle di studi sull'intelligenza artificiale IDSIA

Reagent prediction with a transformer and its benefits for reaction product prediction

Mikhail Andronov, The Swiss AI Lab IDSIA

Advanced Machine Learning for Innovative Drug Discovery (AIDD) – Horizon 2020 Marie Skłodowska-Curie Innovative Training Network

Paper

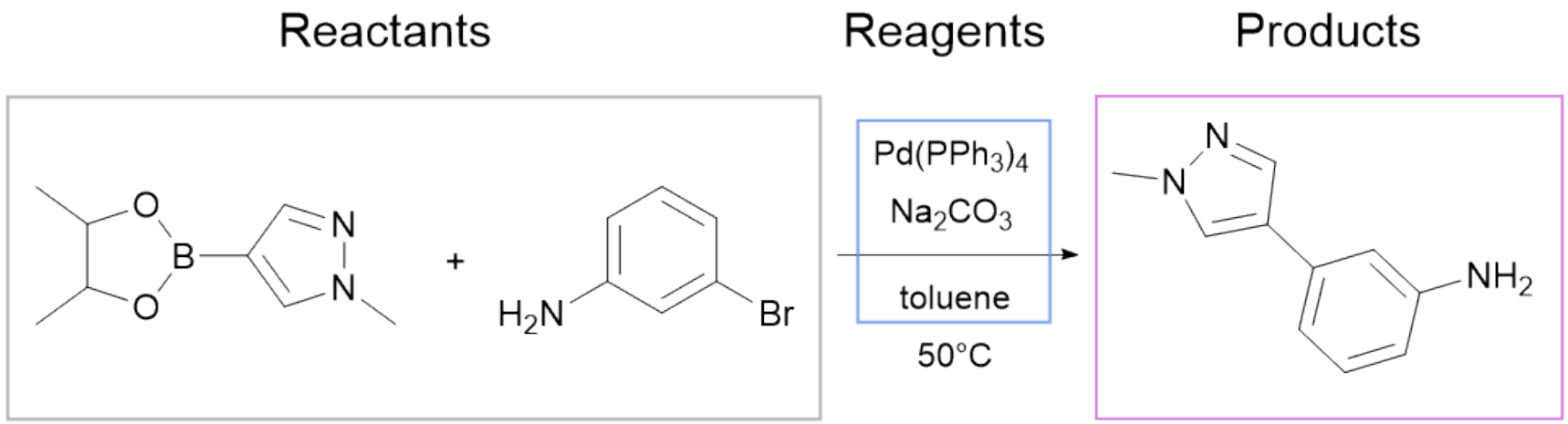
Cite this: DOI: 00.0000/xxxxxxxxxx

Reagent Prediction with a Molecular Transformer Improves Reaction Data Quality†

Mikhail Andronov,[a,e] Varvara Voinarovska,[b] Natalia Andronova,[c] Michael Wand,[a,d] Djork-Arné Clevert,[e] and Jürgen Schmidhuber[a,f]

Presented at 23rd EuroQSAR (Heidelberg, Germany)
Uploaded to *chemRxiv* in November
Under review to *Chemical Science*

Chemical reactions



A reaction type is defined by the reaction center and reagents.

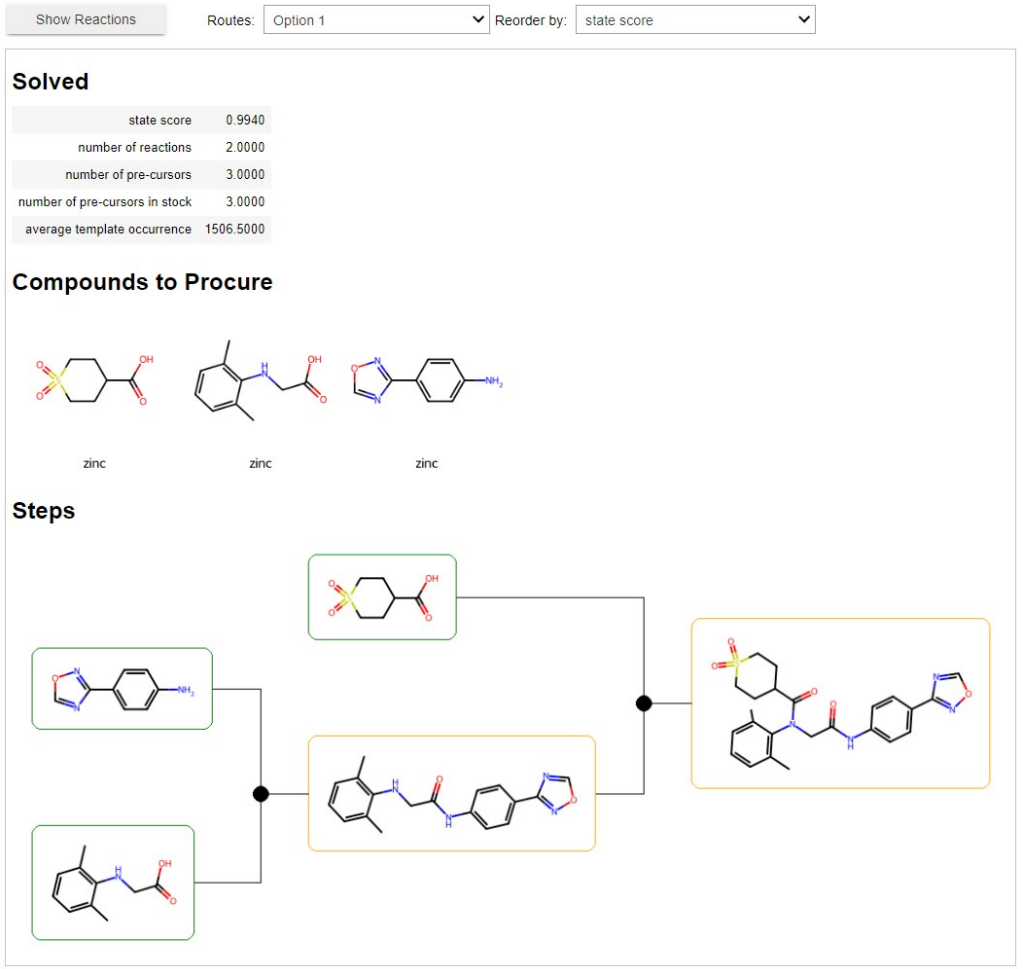
With different reagents, reactants can turn into different products.

Any part of a reaction can be predicted.

Why predict reagents?

1). **To help CASP**
Aizynthfinder generates routes without reagents.

2). **To address data flaws**
Many reactions in USPTO don’t always have well-specified reagents

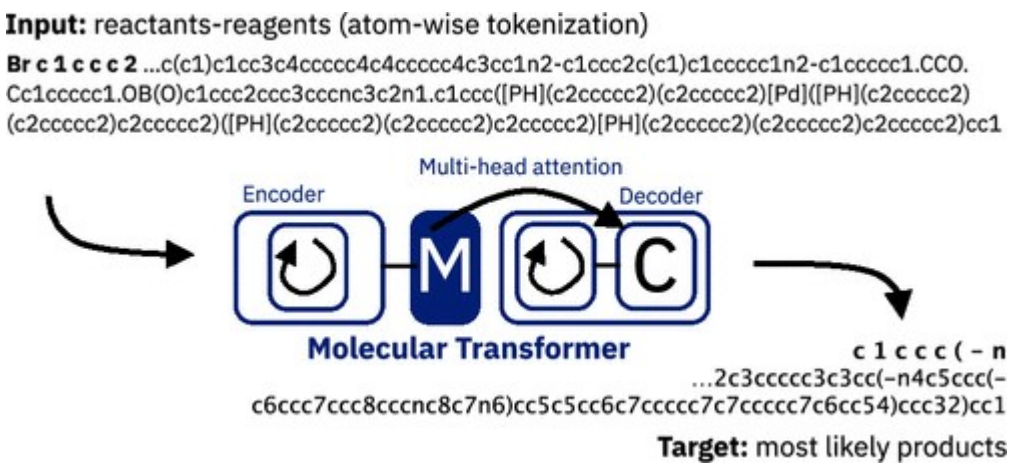


Literature: conditions prediction

Paper	Reaction types	Goal of predictions	Dataset	Model	Data format
Walker et al. 2019	Five name reactions	Solvent	Reaxys	SVM	Molecular fingerprints (OpenBabel)
Afonina et al. 2021	Hydrogenation reactions	Catalyst, temperature, pressure	Reaxys	MLP	Molecular fingerprints (ISIDA Fragmentor 2017)
Gao et al. 2018	Broad range of reactions	Restricted set of reagents, temperature	Reaxys	MLP	Molecular fingerprints (RDKit)
Maser et al. 2021	Four name reactions	Restricted set of reagents, temperature	Reaxys	GBM, GNN	Molecular graphs

Transformer

Today the standard base model for all kinds of NLP tasks.
Originally proposed for machine translation.



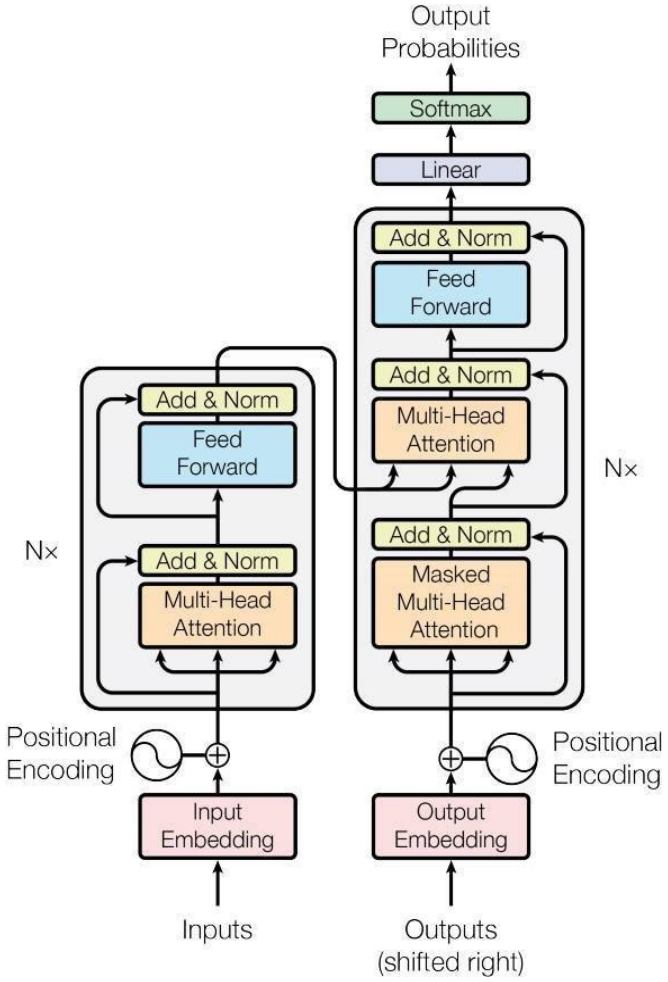
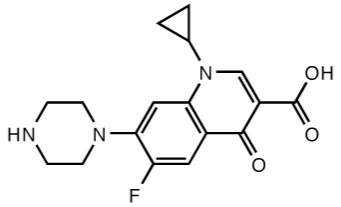


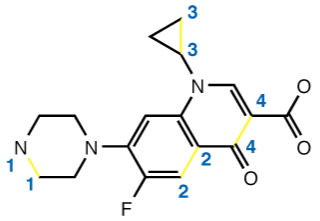
Fig. from Vaswani et al. 2017, Schwaller et al. 2019

SMILES

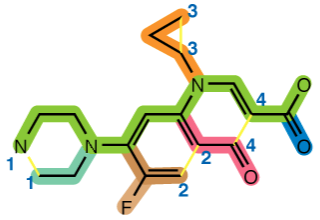
A



B



C



D

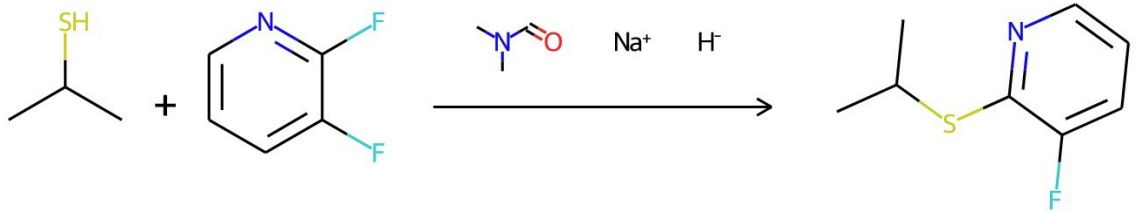
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

SMILES for cyprofloxacin

SMILES – a text notation of organic molecules designed for chemical information systems.
Reaction SMILES are to depict reactions.

The idea of SMILES is to build a spanning tree in the molecular graph.

CC(C)S.Fc1cccnc1F>CN(C)C=O.[Na+].[H-]>CC(C)Sc1ncccc1F



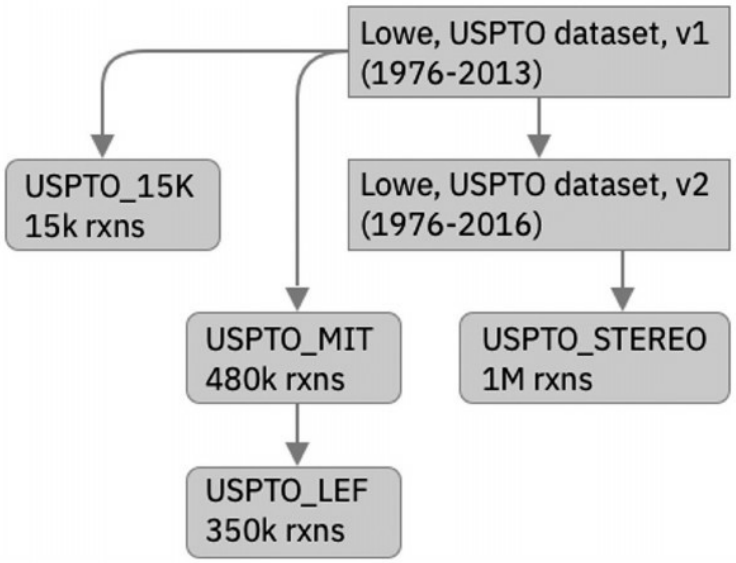
Reaction SMILES

Any part of a reaction SMILES can be predicted by masked language modeling

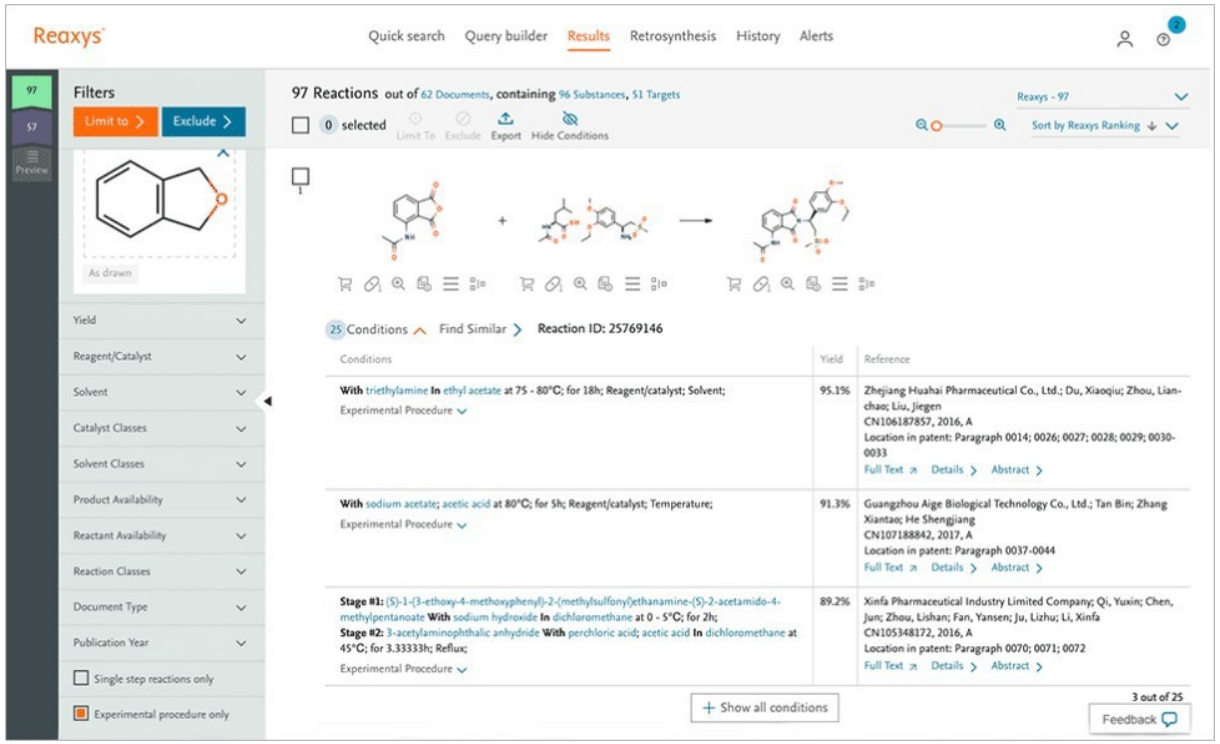
Chemical reaction data

Chemical reactions from US patents (USPTO dataset, 2012) – the only open chemical reaction dataset.

Consists of 1-2M reactions obtained by text mining, pretty noisy.



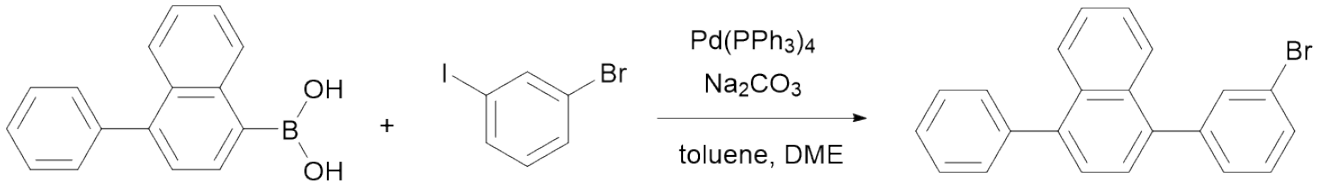
Reaxys – a proprietary expert-curated database from Elsevier, 56M reactions.



Pictures from P. Schwaller and T. Laino, Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions. January 1, 2019 , 61-79; https://www.elsevier.com/ data/assets/image/0020/1156070/Reaxys-Filter-Results.png

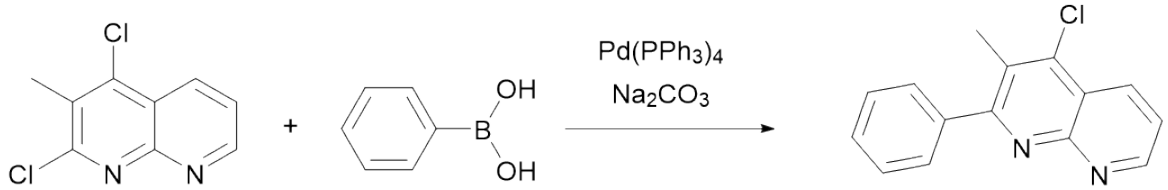
USPTO noise

US07985491B2, 2011



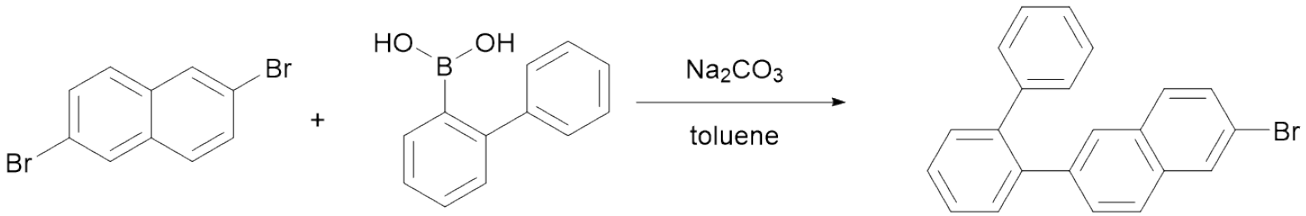
All reagents are specified

US08765940B2, 2014



No solvent is specified

US08853675B2, 2014



The Pd catalyst is missing

US09394290B2, 2016

All reagents are missing

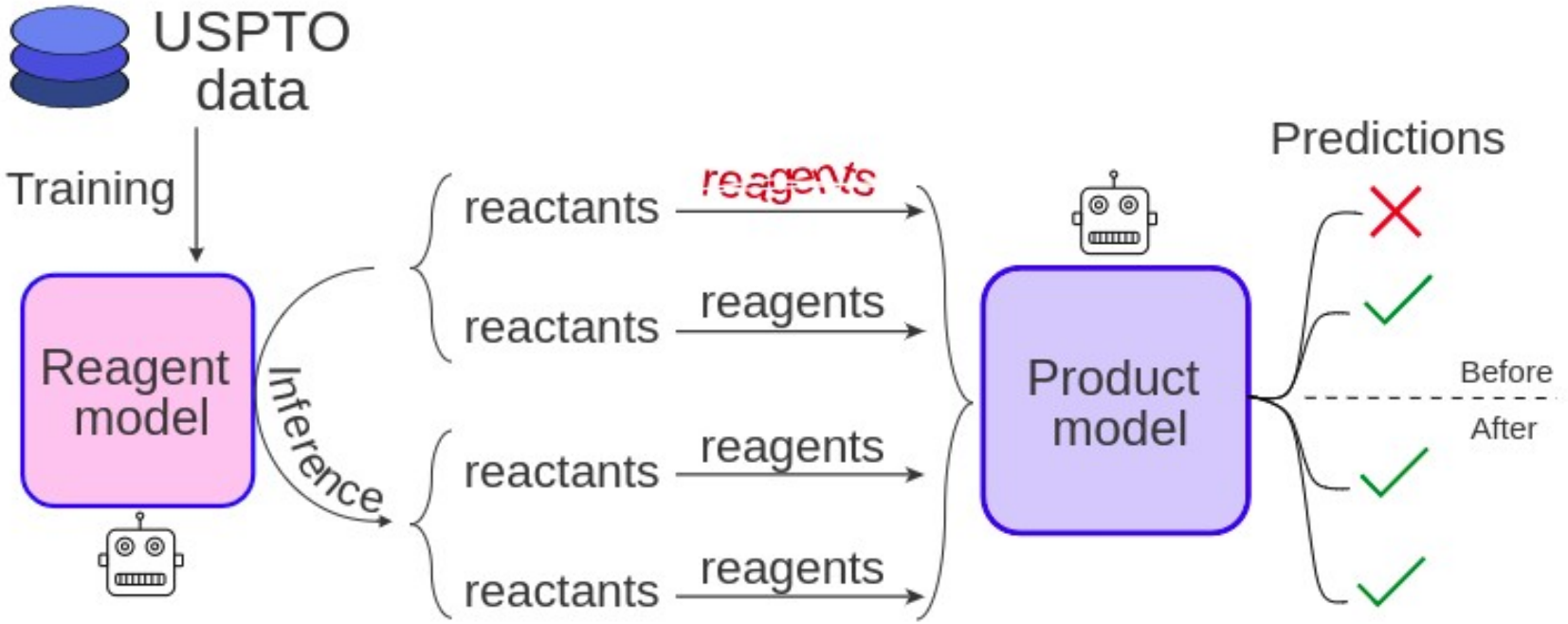
A catalyst, a base and a solvent are necessary.

Paper idea

Reagent and product models are transformers.

We can use a reagent model to improve product prediction models.

Model-agnostic in principle

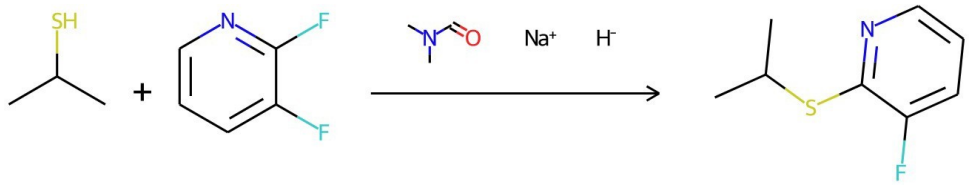


Training set

Training on full USPTO without USPTO MIT test. Final size ~ 1M reactions

Preprocessing:

1). Delete atom mapping.
2). Mix up precursors and extract reagents with RDKit.
3). Remove reagents which are too rare.
4). Augment data.
5). Sort reagents by roles (catalyst, solvent, etc.) using heuristics.



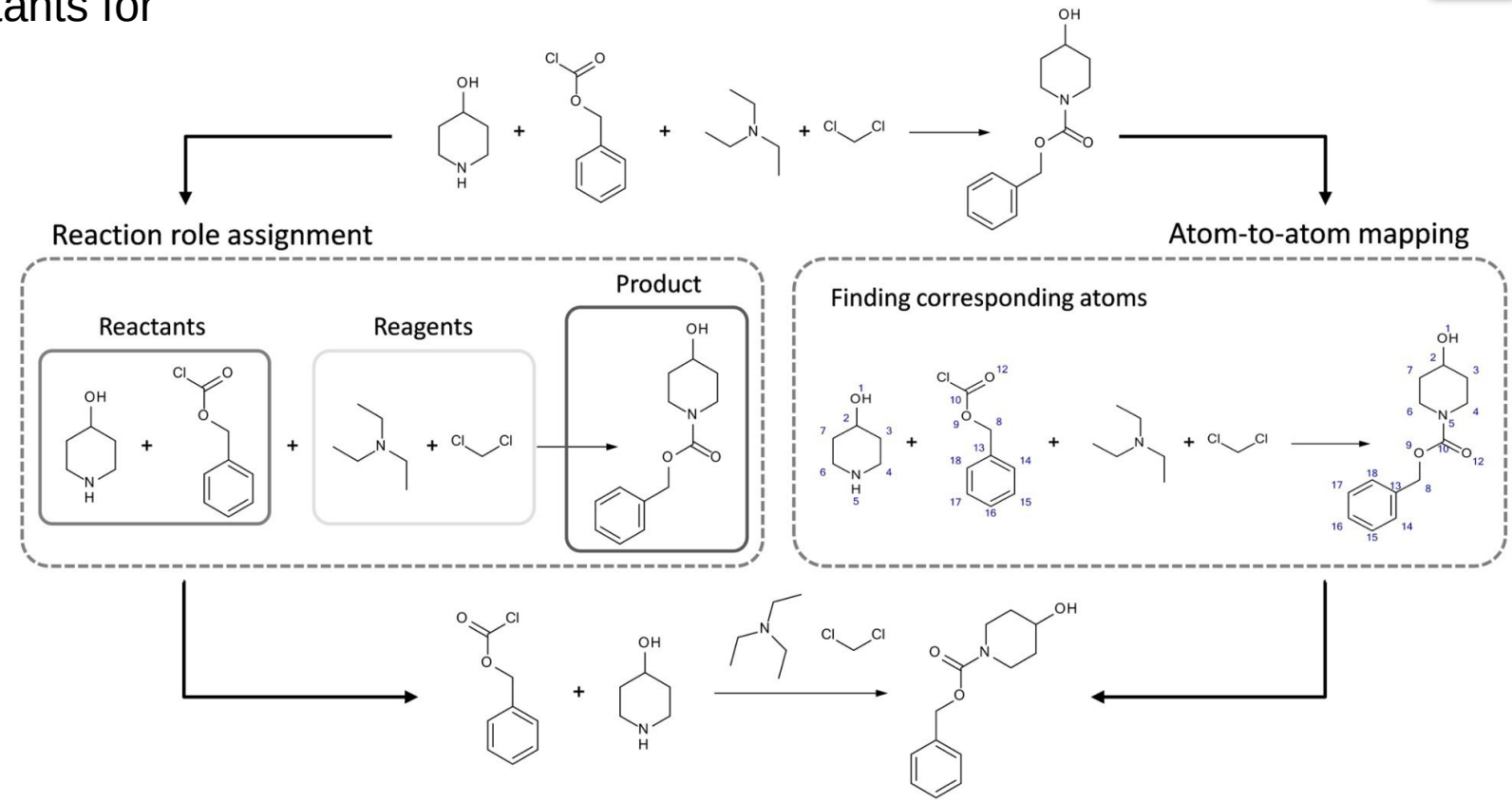
`CC(C)S.Fc1cccnc1F>>CC(C)Sc1ncccc1F` `[H-].[Na+].CN(C)C=O`

source target

Reaction role assignment

An RDKit procedure to separate reactants for reagents (Schneider et al. 2016).

Atom mapping not needed.



Pics from Schneider et al., J. Chem. Inf. Model. 2016, 56, 12, 2336–2346

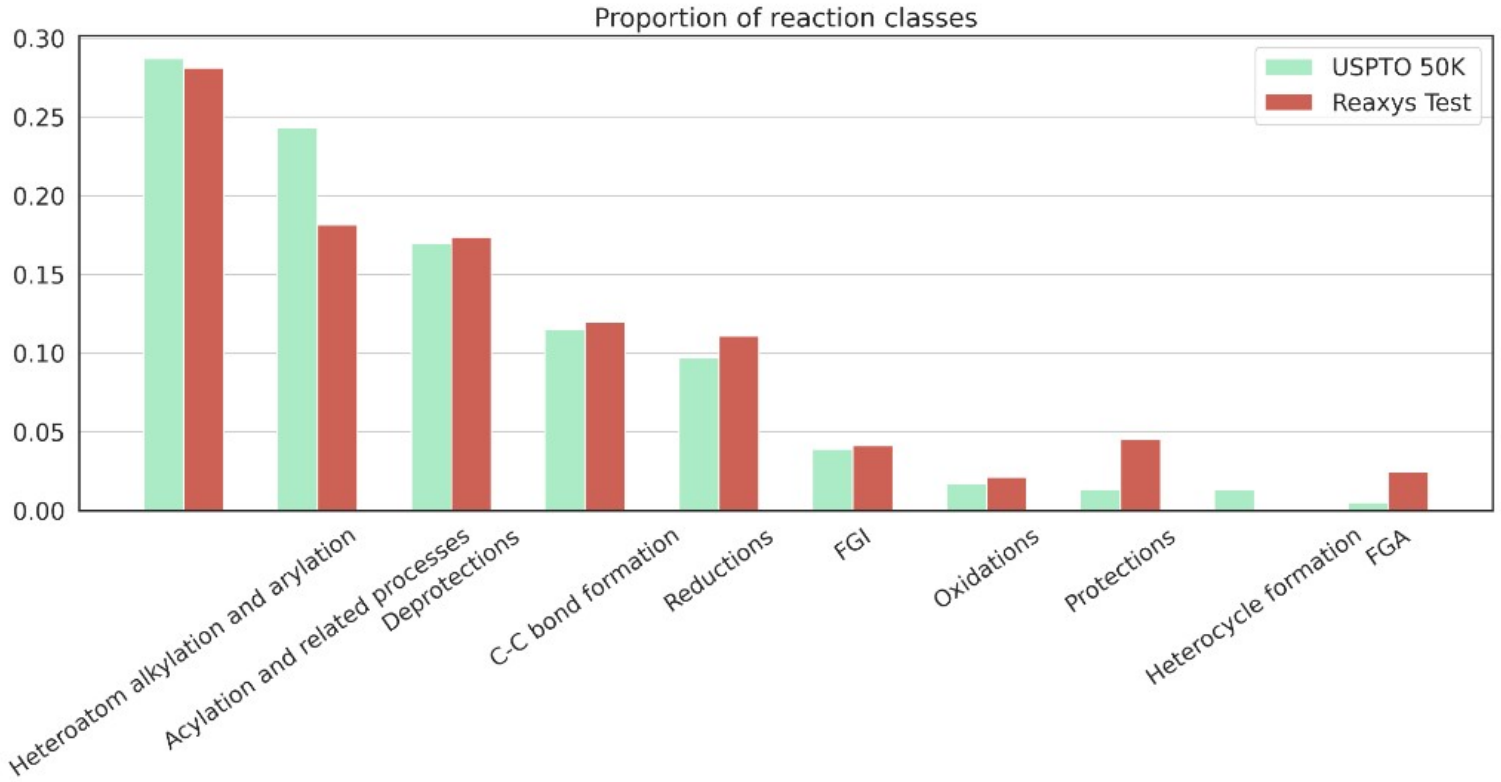
Test set

We used a subset of Reaxys for testing purposes.

Size: 96972 reactions.

Reagent SMILES determined by *PubChemPy.*
Reaction types determined by *NameRXN.*

Design goal: similarity to USPTO 50K in terms of types distribution



Discussion: overall performance

Exact match accuracy
Predicted only the molecules in the ground truth and all of them. A.C.B. ~ A.B.C
Partial match accuracy
Some of the molecules are predicted correctly. A.B ~ A.C.D.
Recall
#(correctly predicted) / #(molecules in target)

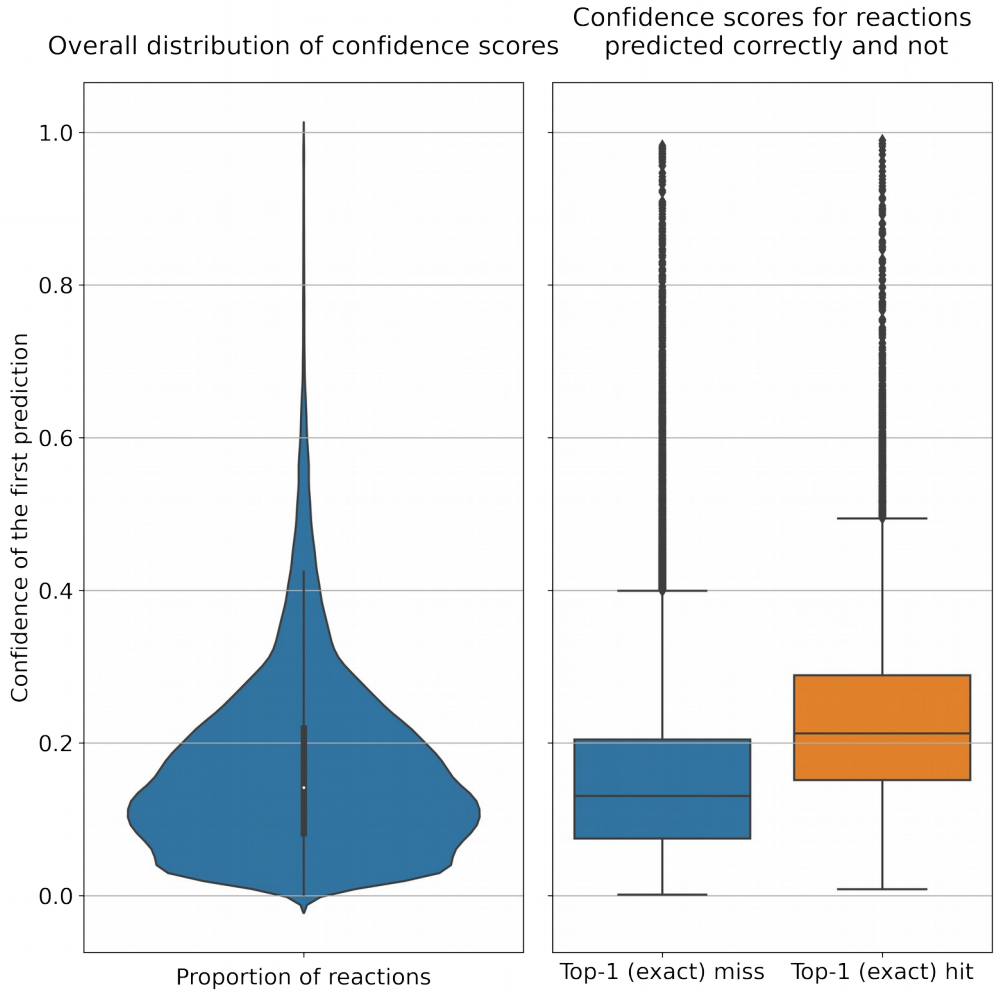
Metric	Top-1	Top-2	Top-3	Top-4	Top-5
Exact match accuracy	17.0	24.7	29.2	31.8	33.5
Full recall	19.2	28.4	31.5	39.3	42.8
Partial match accuracy	70.9	80.5	89.4	87.3	88.9

Model confidence

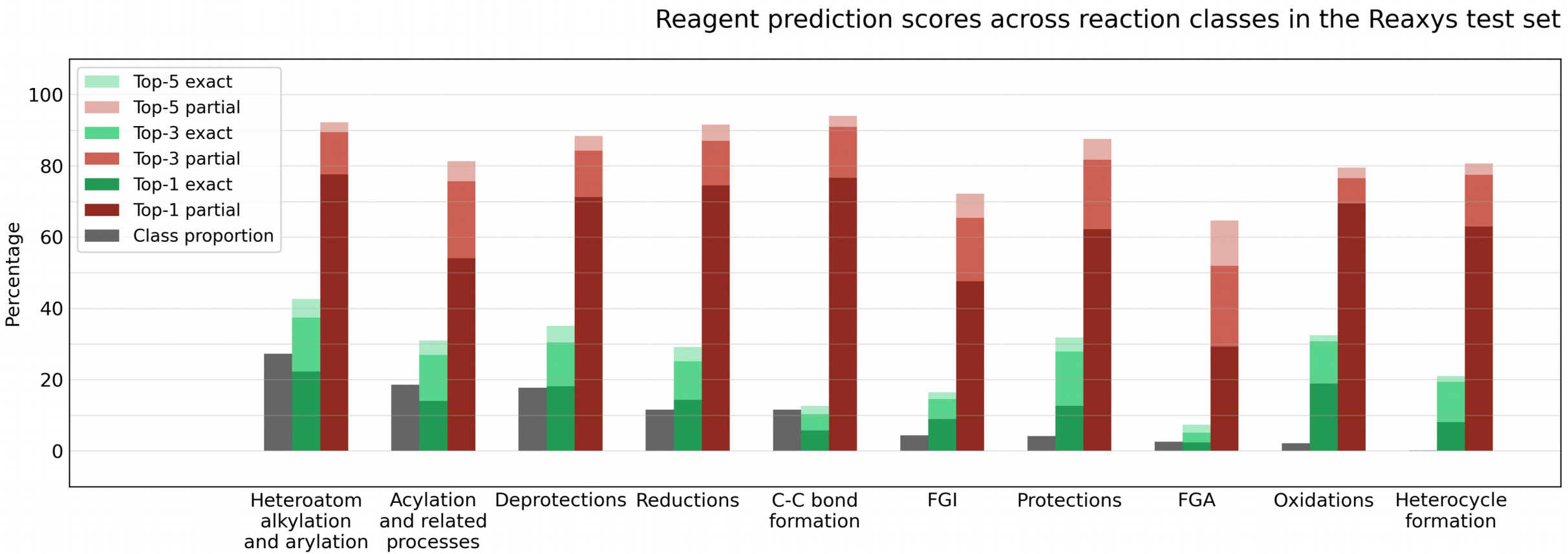
Confidence: product of the probabilities of all tokens in the generated sequence.

The reagent model is much less confident than a product model.

For the latter, it is close to 1 almost all of the time.



Performance across reaction types

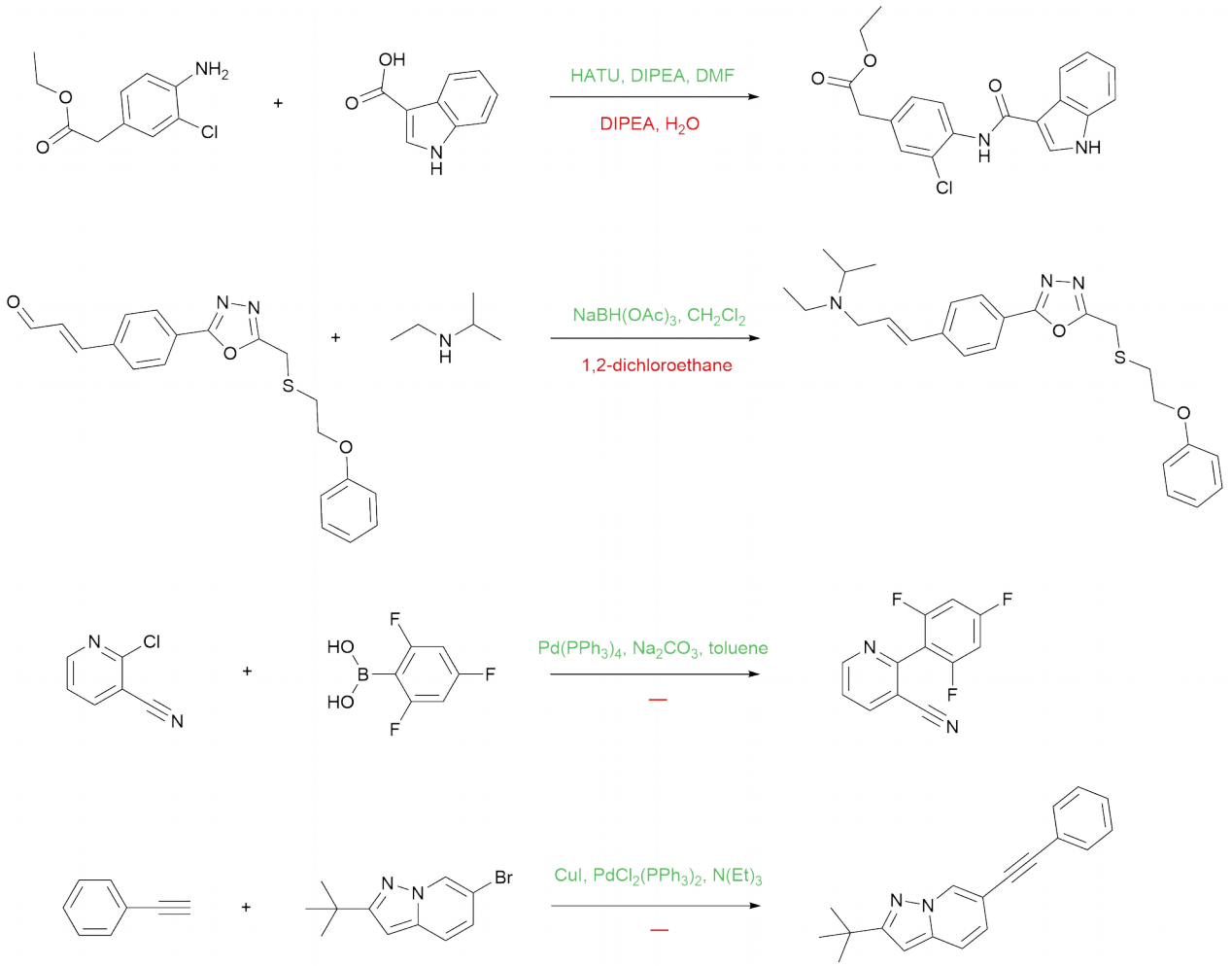


Reagent improvement

Strategy
Replace if more molecules were predicted than there was reported.

Reagents changed in ~25% of reactions

Restored catalysts, reducing agents, etc.



Product prediction

The new model performs better than the old model on both Reaxys and USPTO in both separated and mixed settings.

	Reaxys	USPTO MIT
MT, no reagents	77.3	84.0
MT base, mixed	82.0	87.7
MT new, mixed	**83.0**	**88.3**
MT base, separated	84.3	89.2
MT new, separated	**84.6**	**89.6**

MT base: trained on basic USPTO.
MT new: trained on USPTO with reconstructed reagents.

Statistical significance: McNemar's test

	F_1 incorrect	F_1 correct
F_2 incorrect	A	B
F_2 correct	C	D

$$x = \frac{(|B - C| - 1)^2}{B + C}$$

Chi-squared distribution (1 degree of freedom).
Null hypothesis: difference is accidental

Conclusion

> Transformer can be succesfully used to suggest reagents for organic reactions.

> We used the strategy to train a model on USPTO and test it on Reaxys.

> We used a reagent model to improve a product model in a *self-supervised* and *model-agnostic* fashion.

> We beat the score of the Molecular Transformer on USPTO MIT.

Thank you for your attention!