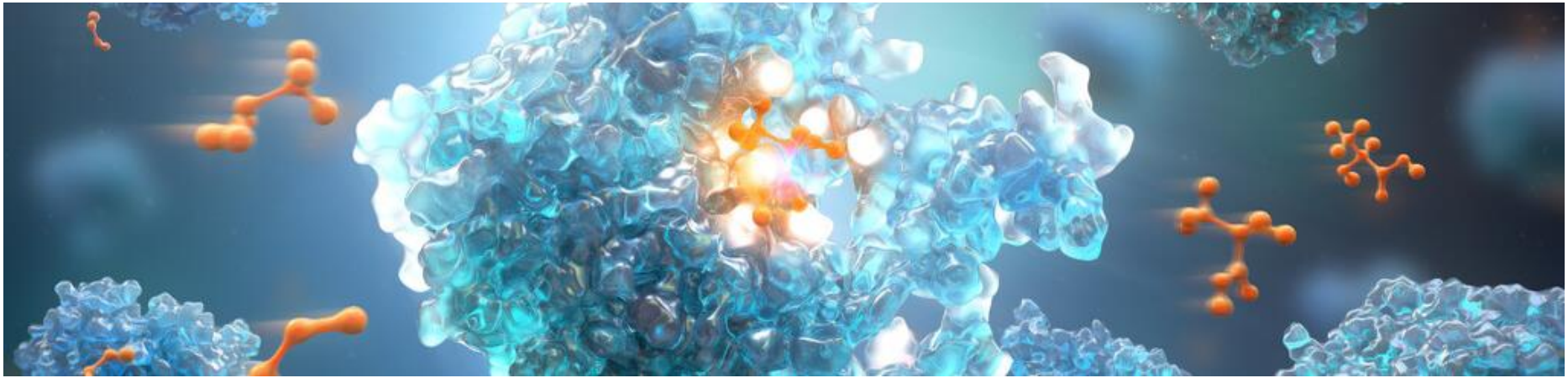
RDKit derived reaction labels for improved retrosynthetic route finding

Esben Jannik Bjerrum, Principal Scientist

RDKit UGM 2020

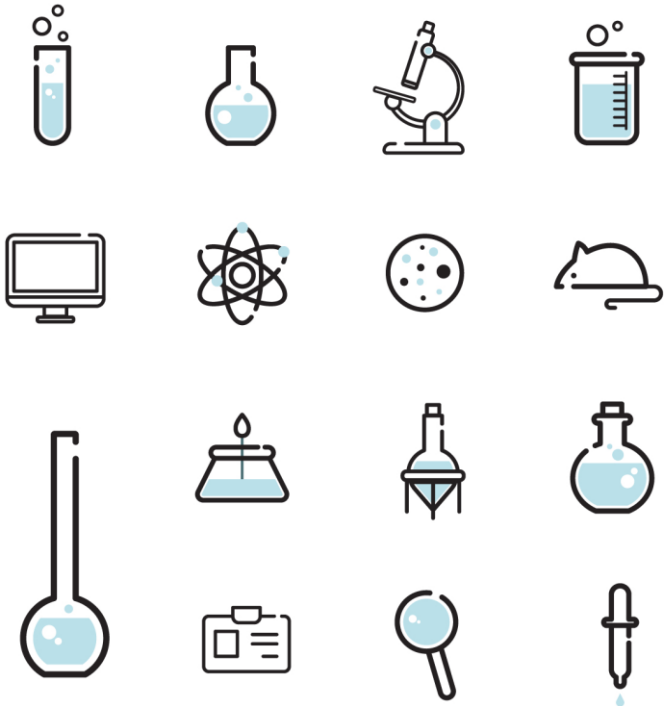
2020 may 29



Agenda

- Speeding up design-make-test-analyze cycles with machine learning
- Reaction prediction
- Retrosynthesis planning with MCTS tree search

- Policy model failures
- Artificial Labels and applicability filtering
- Slow combinations of template applications

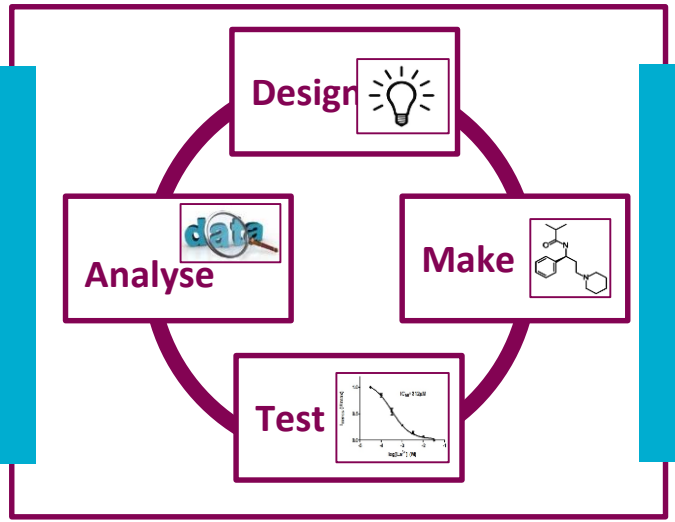


Design-Make-Test-Analyse cycles in Drug Discovery (DMTA)

Drug target

Chemical starting point ("Hit") found through HTS, DEL, fragment screening or knowledge

- Weakly active
- Target unselective
- Toxicity risk
- Low metabolic stability



Candidate drug

- Highly potent
- Effective in *in vivo* models
- Metabolically stable
- No toxicity issues

~3 years

Multiple of DMTA cycles

4-6 weeks per cycle

Hand-overs between multiple labs

The challenge: *Find ways to speed up and improve the process using AI*

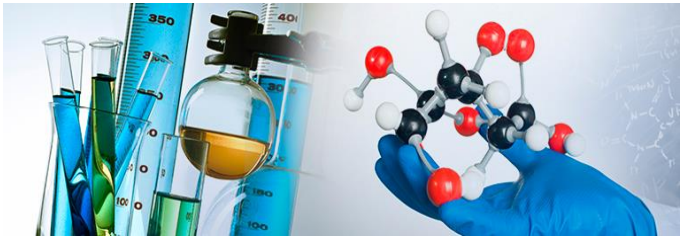
Drug Design

Molecular AI group provides tools for the projects:

What to make next?

De novo design

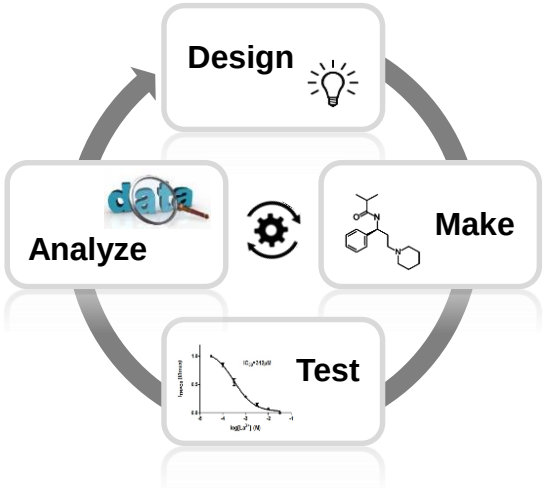
How to make it?

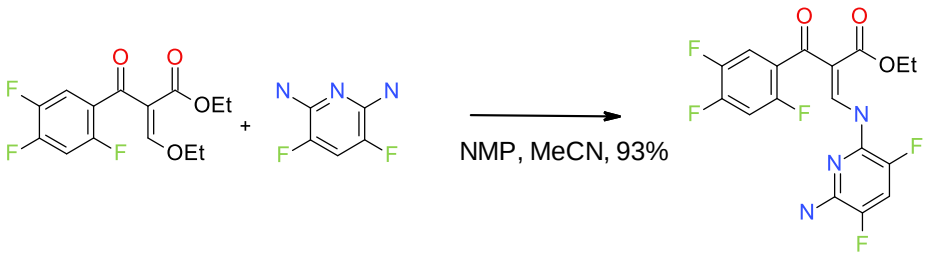


Retrosynthesis

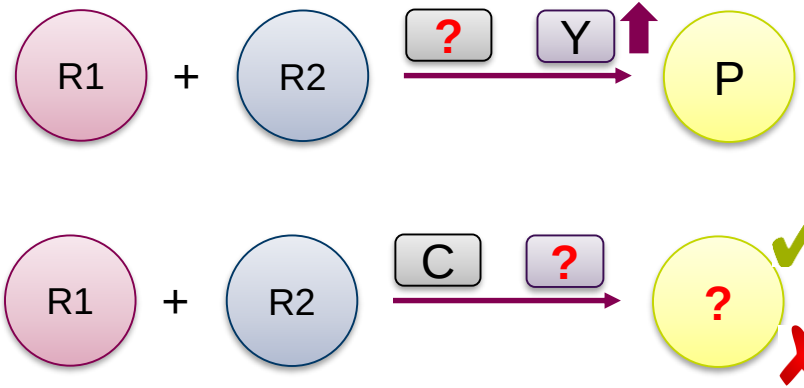
RDKit UGM 2019: SMILES, RNNs and RDKit, - To the molecular universe and beyond

From Design to Compound: Make step





Different Objectives for Synthetis Prediction



Condition Prediction

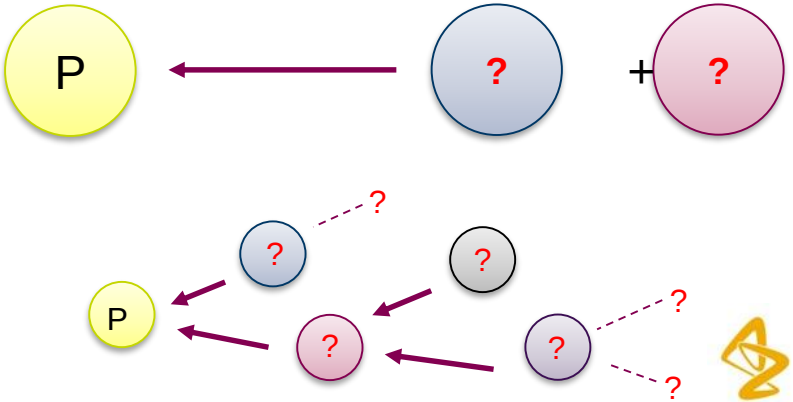
Reaction Feasibility

Forward

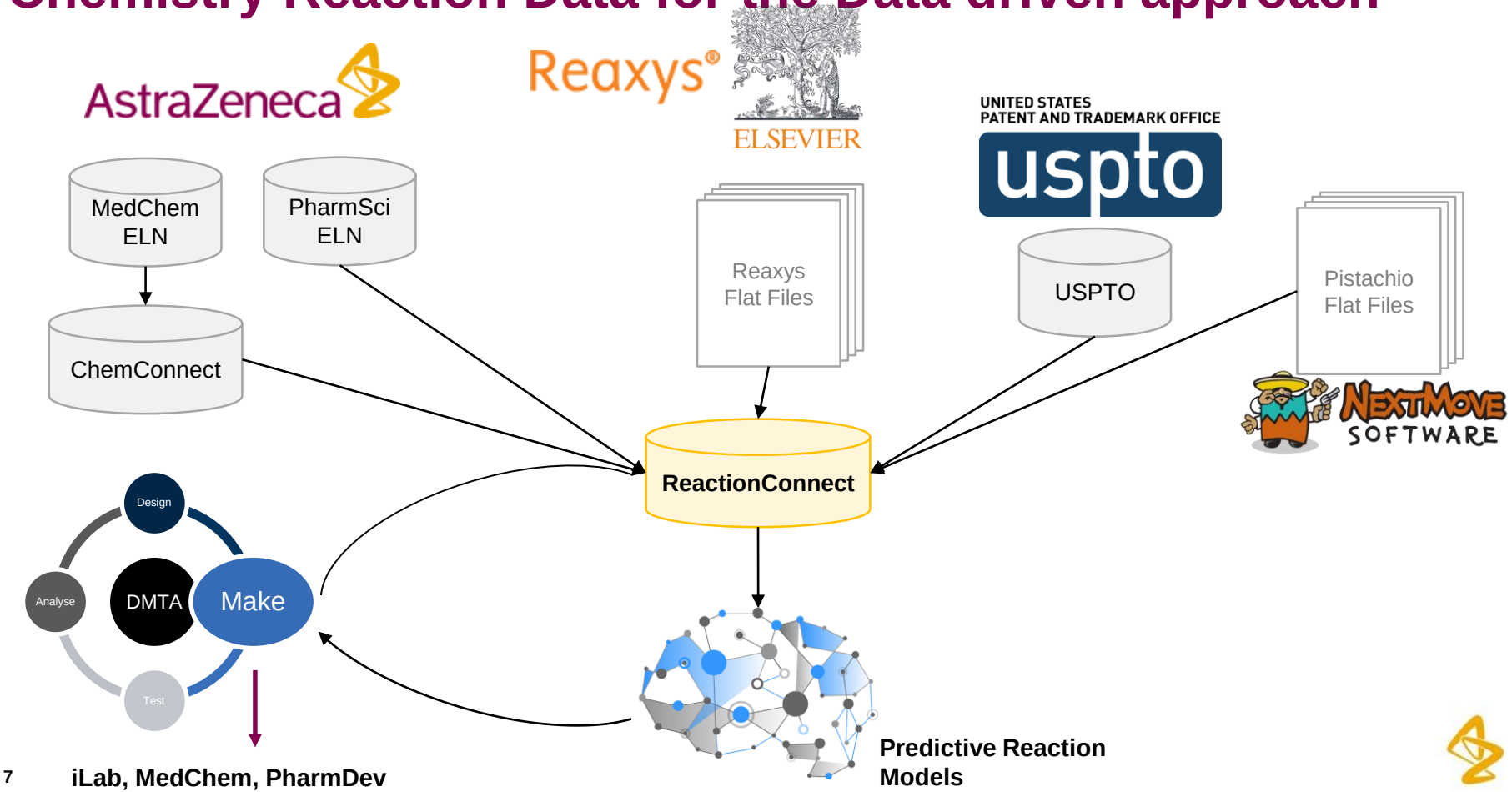
Backward

Retro-synthesis 1-step

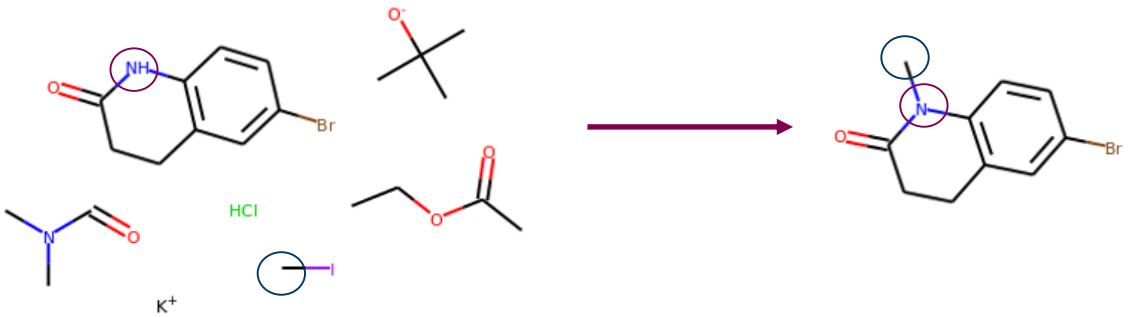
Retro synthetic planning



Chemistry Reaction Data for the Data driven approach

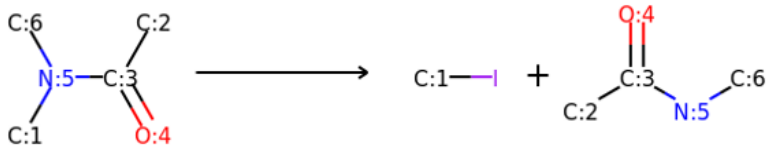


Template Extraction – here from the USPTO dataset



RDChiral for template extraction and application https://github.com/connorcoley/rdchiral

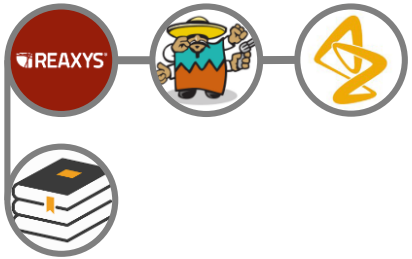
The extracted template (R1)

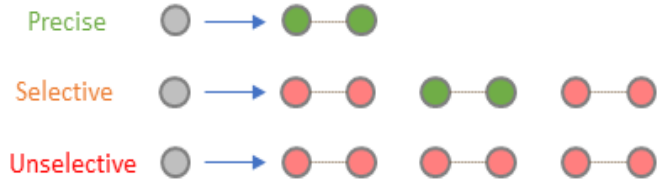


Dataset	Size	Templates Extracted
Pistachio (incl. PGs)	6,839,427	308,951
USPTO 1976-2016	3,748,191	252,877
Reaxys	6,540,786	361 603
All Data	**17 523 783**	**675 530**

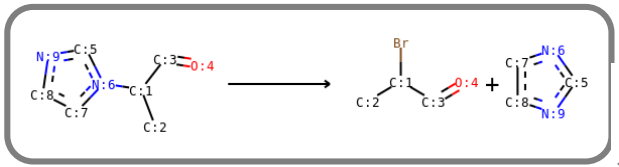
'([C:2]-[C:3](=[O;D1;H0:4])-[N;H0;D3;+0:5](-[**CH3**;D1;+0:1])-[**c:6**])>>(I-[**CH3**;D1;+0:1]).([C:2]-[C:3](=[O;D1;H0:4])-[NH;D2;+0:5]-[**c:6**])'

Template Extraction

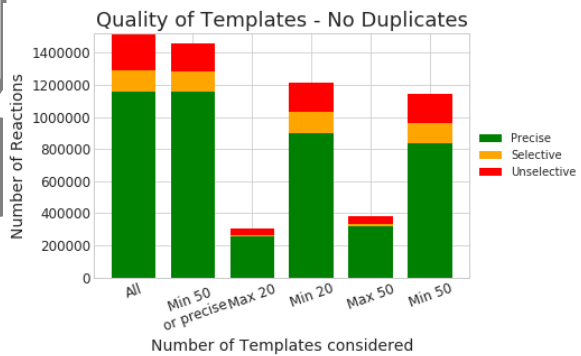


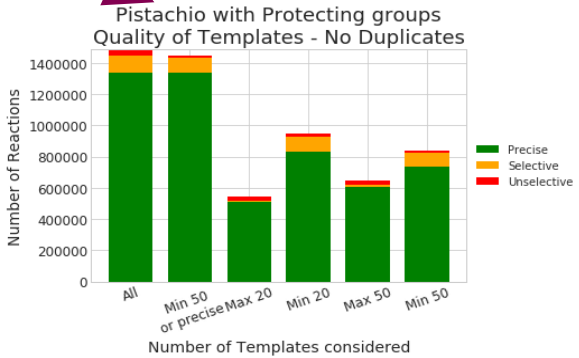


Explicit Handling of Protection Groups increase template quality



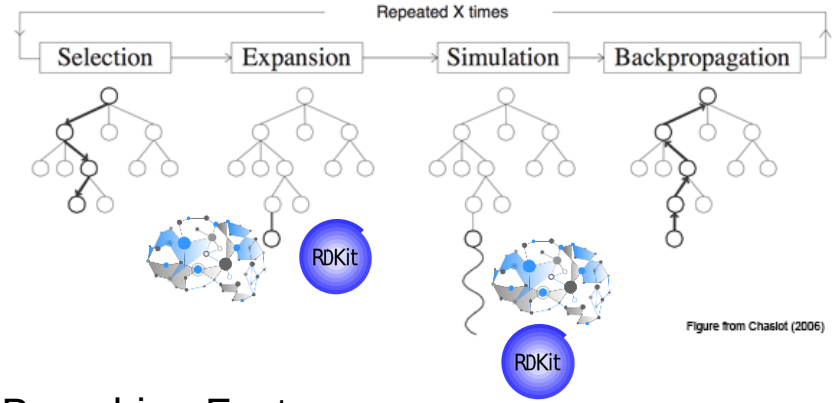
Retro Reaction Templates Extracted

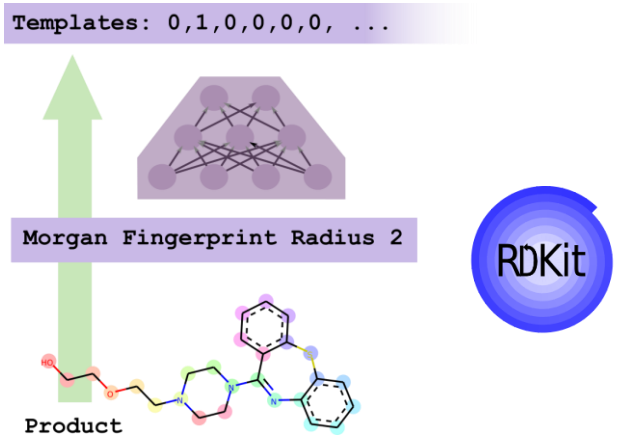




Searching the tree of possible reaction routes

Monte Carlo Tree Search





Branching Factors

	Chess	Go	Retrosynthesis
Search Breadth	~35	~250	> 500,000
Search Depth	~80	~150	~12

Neural Network selects and prioritizes
=> More Manageable Problem

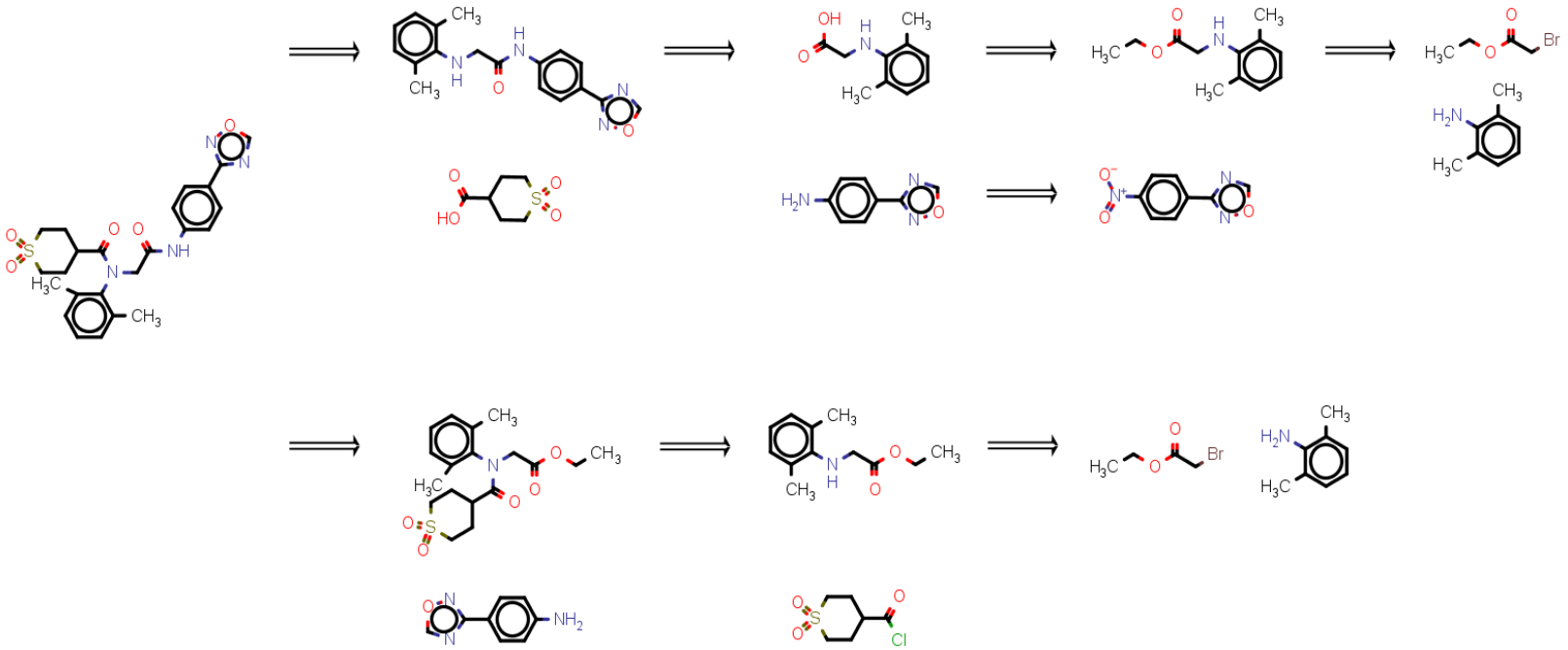
Alpha Go architecture

Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978 .

Results in Seconds to Minutes

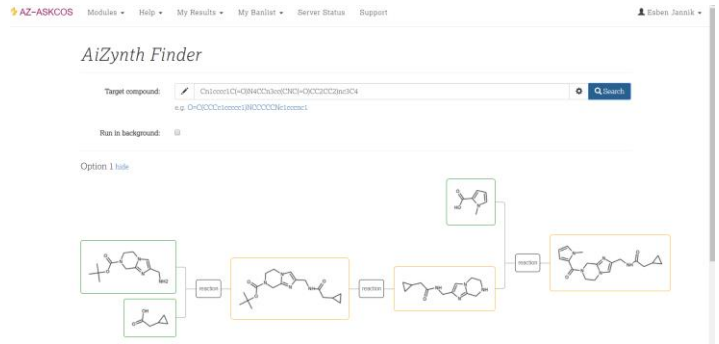
Model: USPTO

Time taken: 3.26 s

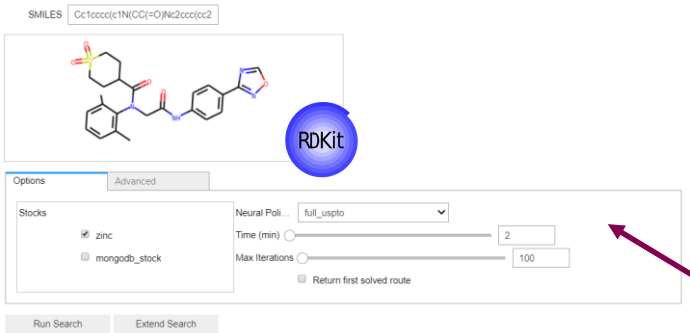


Making the tool available

Web-GUI based on MIT MLDPS consortium tools

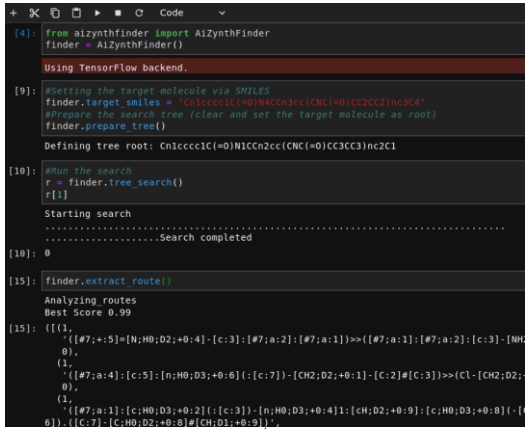


Jupyter based GUI



The Value: Chemists can quickly get suggested routes/ideas to purchasable compounds. Cheminformaticians can filter datasets into “synthesizable/not-synthasizable”

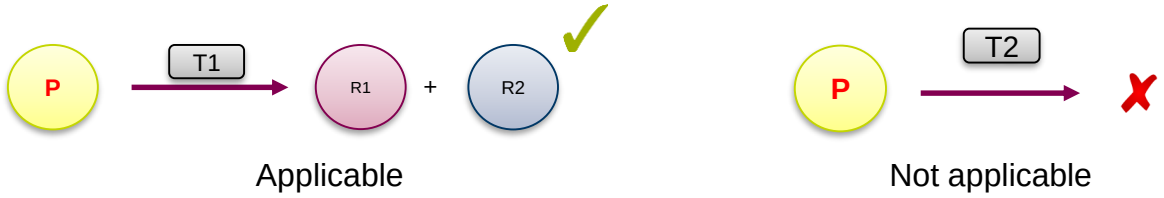
Scripting access via Python Objects



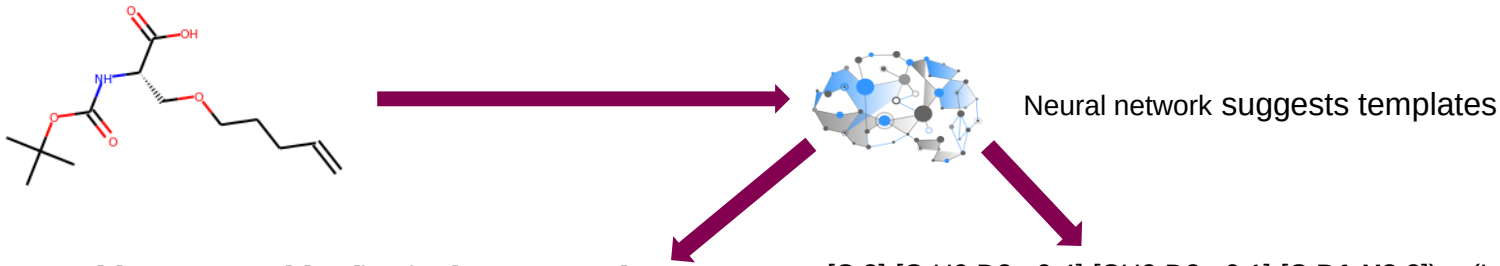
Open Sourced: https://github.com/MolecularAI/aizynthfinder

Genheden et al. AiZynthFinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. **2020**. https://doi.org/10.26434/chemrxiv.12465371.v1 .

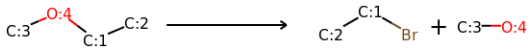
An Issue: Policy suggested templates don't necessarily work



Compound = Chem.MolFromSmiles('C=CCCCOC[C@H](NC(=O)OC(C)(C)C)C(=O)O')



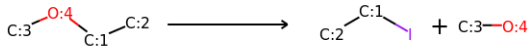
[C:3]-[O;H0;D2;+0:4]-[CH2;D2;+0:1]-[C:2])>>(Br-[CH2;D2;+0:1]-[C:2]).([C:3]-[OH;D1;+0:4]



```
top_reaction = AllChem.ReactionFromSmarts(template_0.retro_template)
outcome = top_reaction.RunReactants([compound])
outcome
```

```
((<rdkit.Chem.rdchem.Mol at 0x7f59342ef920>,
  <rdkit.Chem.rdchem.Mol at 0x7f59b034bd40>),
 (<rdkit.Chem.rdchem.Mol at 0x7f58b9668030>,
  <rdkit.Chem.rdchem.Mol at 0x7f58b96687c0>))
```

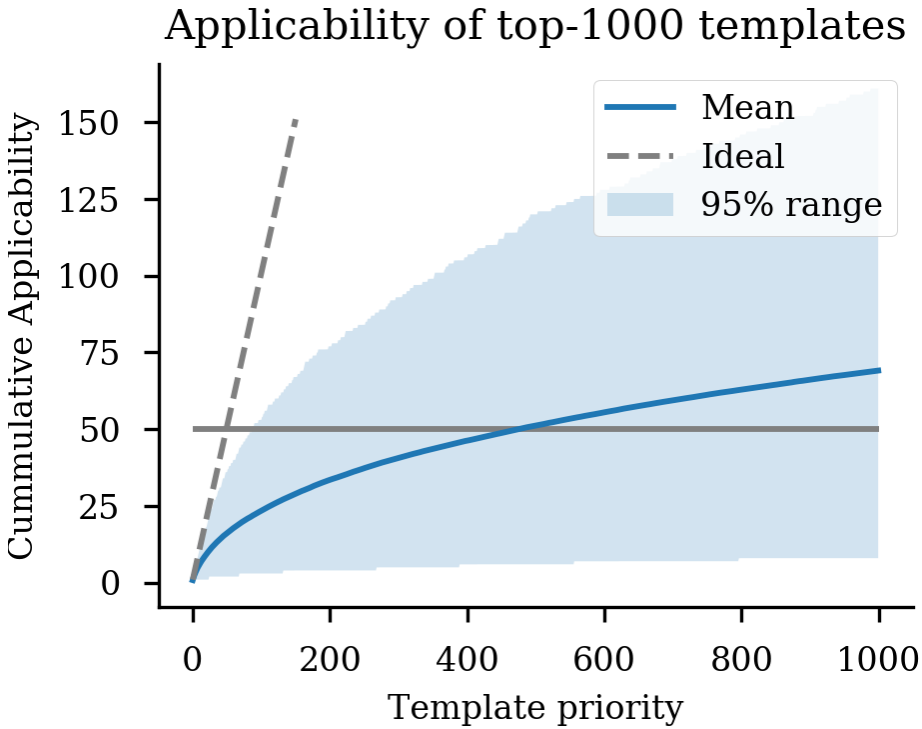
[C:3]-[O;H0;D2;+0:4]-[CH2;D2;+0:1]-[C;**D1;H3**:2])>>(I-[CH2;D2;+0:1]-[C;D1;H3:2]).([C:3]-[OH;D1;+0:4]



```
top3_reaction = AllChem.ReactionFromSmarts(template_3.retro_template)
top3_reaction.RunReactants([compound])
```

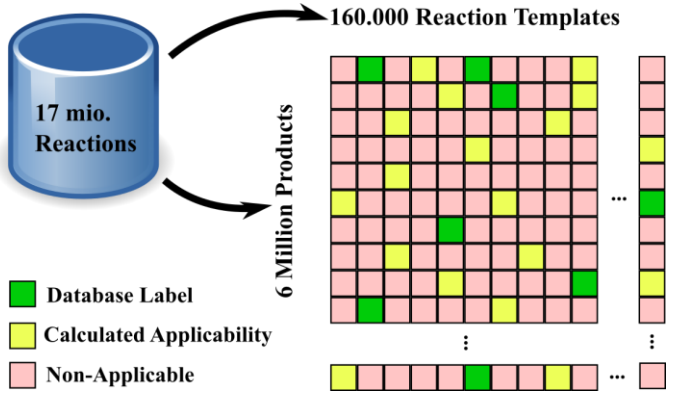
```
()
```

Cummulative Applicability

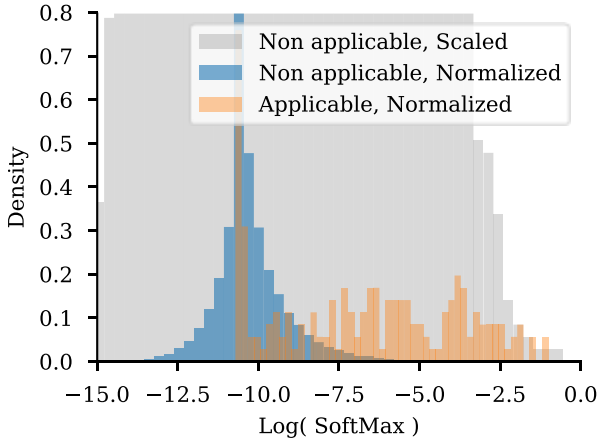


Sometimes we need to try a lot to get 50 working templates for the tree-search!

RDKit derived artificial Labels for filter training



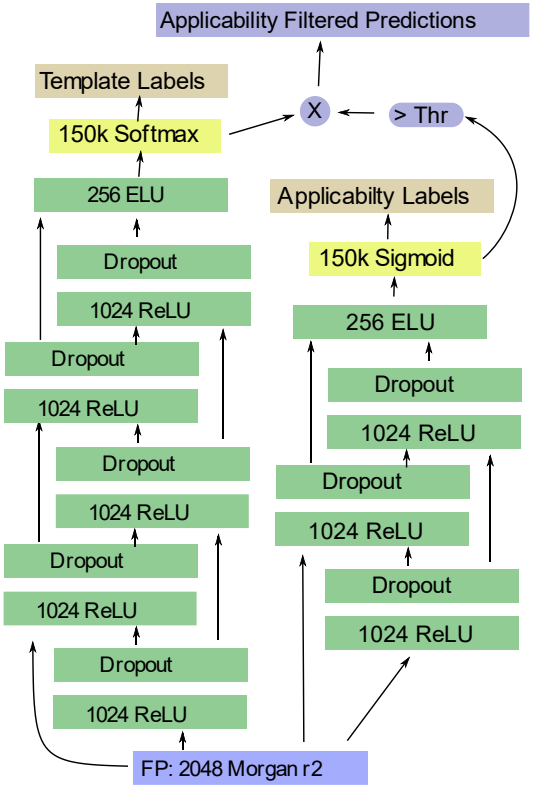
960 000 000 000 REACTION MATCHES

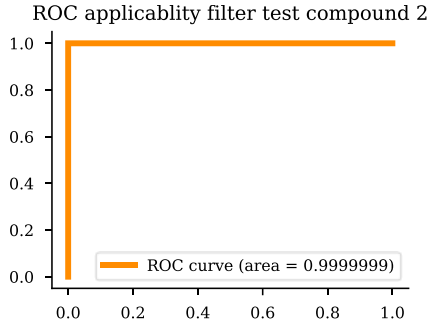


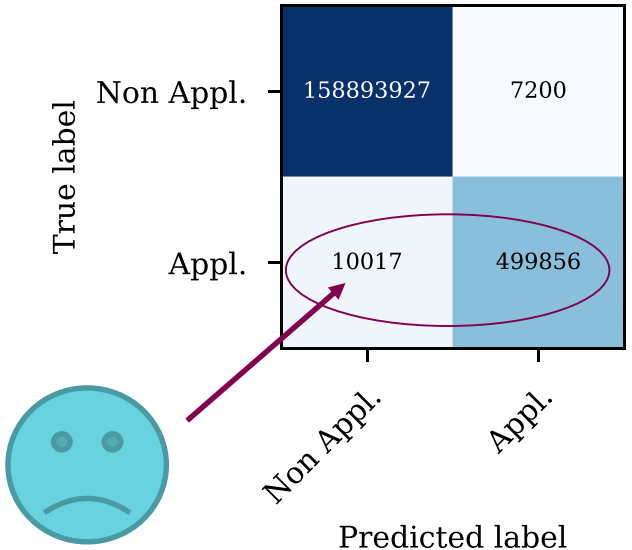
♥Scipy Sparse♥

Orders of magnitude more non-applicable than applicable

Filtering via second neural network







Recall must be high!

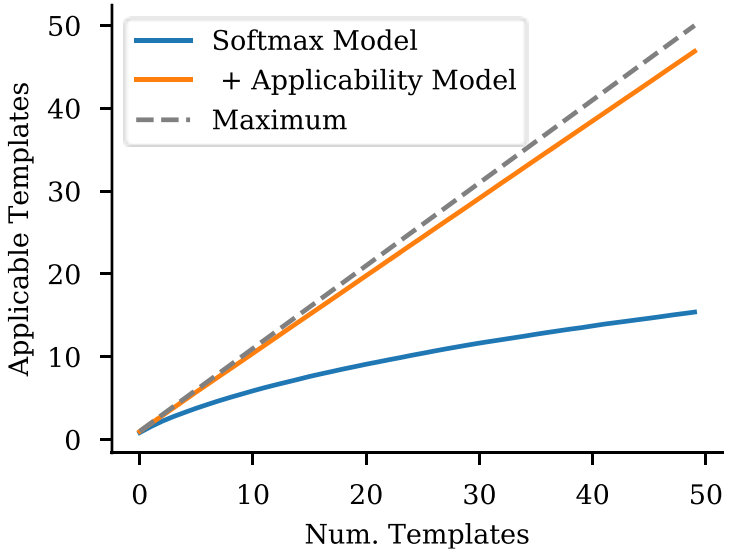
Recall = TP/P

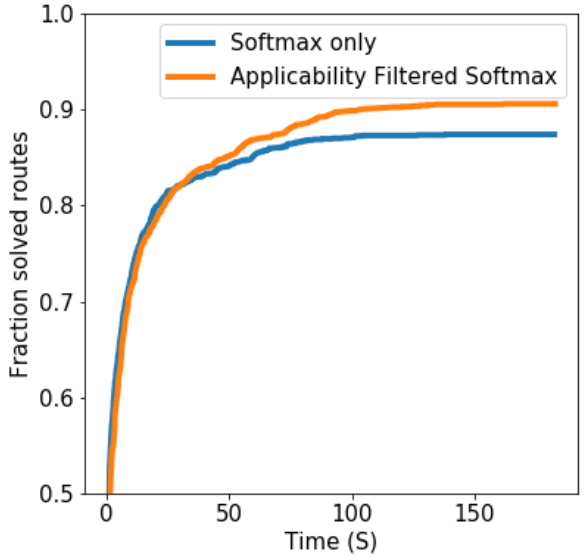
Why does it work so suspiciously well?

- Noise-free artificial data
- Morgan fingerprints contain the relevant information
- Easy to rule out negatives (e.g. atomtype not found in template => non-applicable)

However, we do take a “slow” serial process on 150.000 templates (seconds) and turn it into a fast parallel process on the GPU (milliseconds)

Improved filtering of templates gives more solved routes

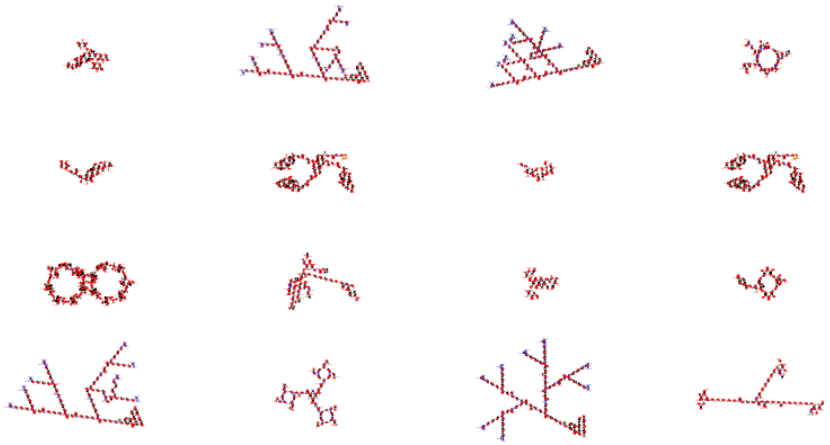




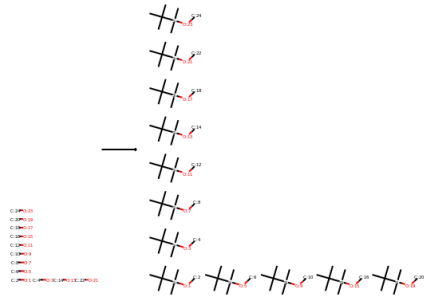
Bjerrum, Esben Jannik; Thakkar, Amol; Engkvist, Ola (2020): Artificial Applicability Labels for Improving Policies in Retrosynthesis Prediction. ChemRxiv. Preprint. https://doi.org/10.26434/chemrxiv.12249458.v1

Long application times for some template-compound combinations

Examples of Filtered compounds



Example of slow templates



In-silico testing of the templates revealed some unusual templates

Conclusions

- Data driven retro-synthetic algorithms are performant
- Specialized neural networks can provide alternatives in single step predictions
- Policy networks gets many template suggestions wrong
- Calculation of artificial labels for training pre-filter networks can improve route search performance

Acknowledgements

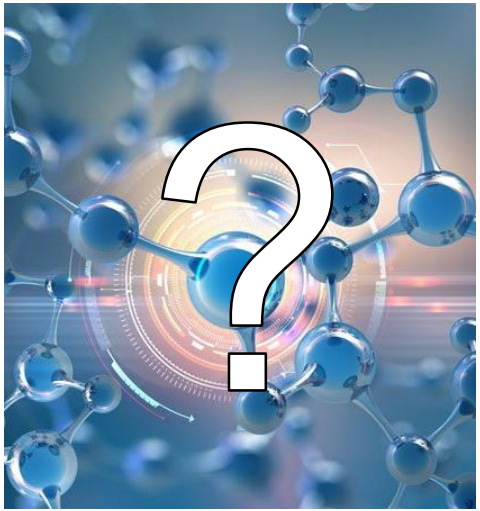
Molecular AI group:
Ola Engkvist, Associate Director, Molecular AI
Jiazhen He, post.doc. Molecular AI
Amol Thakkar, Ph.D student, BIGCHEM
Dean Sumner, Graduate Scientist, Graduate Programme
Veronika Chadimova, Graduate Scientist, Graduate Programme
Samuel Genheden, Data Scientist/Software Engineer
Atanas Patronov, Associate Principal Scientist
Isabella Feierberg, Associate Principal Scientist
Thierry Kogej, Associate Principal Scientist
Preeti Lyer, Machine Learning and Cheminformatics Experts
Christian Margreitter, Data Scientist
Papadopoulos, Kostas, Associate Principal Scientist
Lewis Mervin, Machine Learning and Cheminformatics Expert
Christos Kannas Machine Learning/Cheminformatics Expert
Alexey Voronov, Data Scientist/Software Engineer

RDKit community

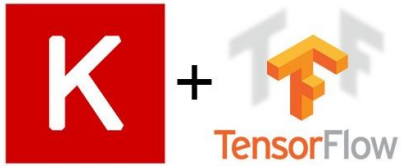
Panagiotis-Christos Kotsias, Graduate Scientist, Graduate Programme
Josep Arus Pous, Ph.D student, BIGCHEM
Rocio Mercado, Post.doc
Tomas Bastys, Post.doc
Simon Johansson, Ph.D Student WASP
Hampus Gummesson Svensson, Ph.D Student WASP
Sebastian Nilsson, Master Student
Tobias Rastemo, Master Student
Emil Sandström, Master Student
Jonathan Sundkvist, Master Student
Huifang You, Master Student
Carl Blomgren, Master Student

Collaborators:
Prof. Dr. Jean-Louis Reymond · Dept. of Chemistry & Biochemistry University of Berne
Christian Tyrchan, Team Leader - Computational Chemistry
Boris Sattarov, Informatics Programmer, Science Data Software LLC
Hongming Chen, Professor, Centre of Chemistry and Chemical Biology, Guangzhou, China
Nidhal Selmi, Research Outsourcing Specialist, Hit Discovery
Peter Varkonyi, Senior Research Scientist | Computational Chemistry

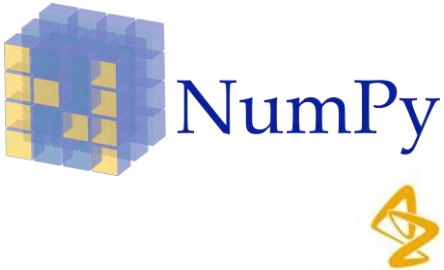
Questions



Toolkits – Source code - Links



ReInvent: https://github.com/MolecularAI/Reinvent
Molvecgen: https://github.com/Ebjerrum/molvecgen
Deep Drug Coder:https://github.com/pcko1/Deep-Drug-Coder
AiZynthFinder: https://github.com/MolecularAI/aizynthfinder
Blogposts: www.cheminformania.com



Confidentiality Notice