CReM: practical structure generation and optimization

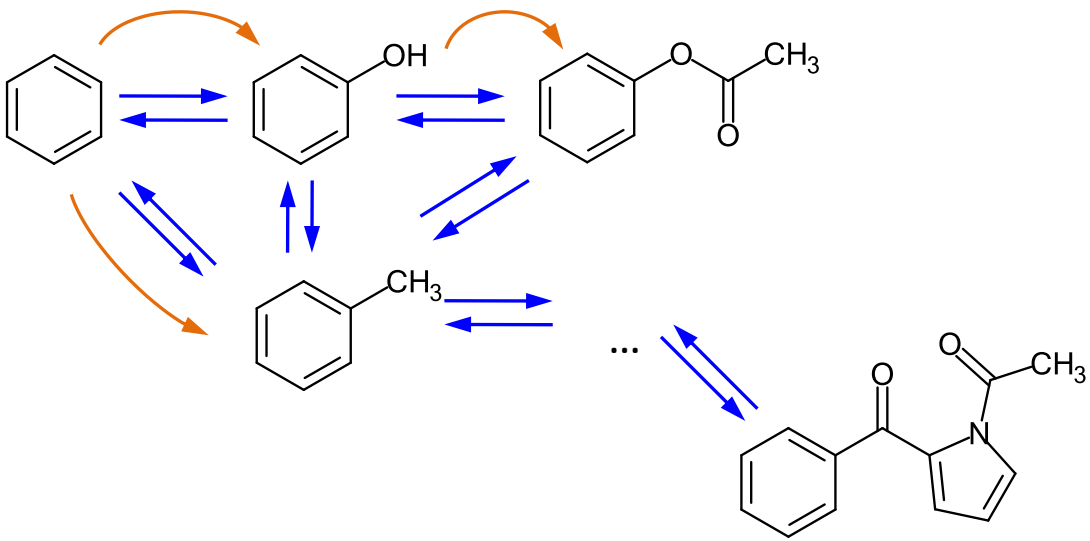
Pavel Polishchuk

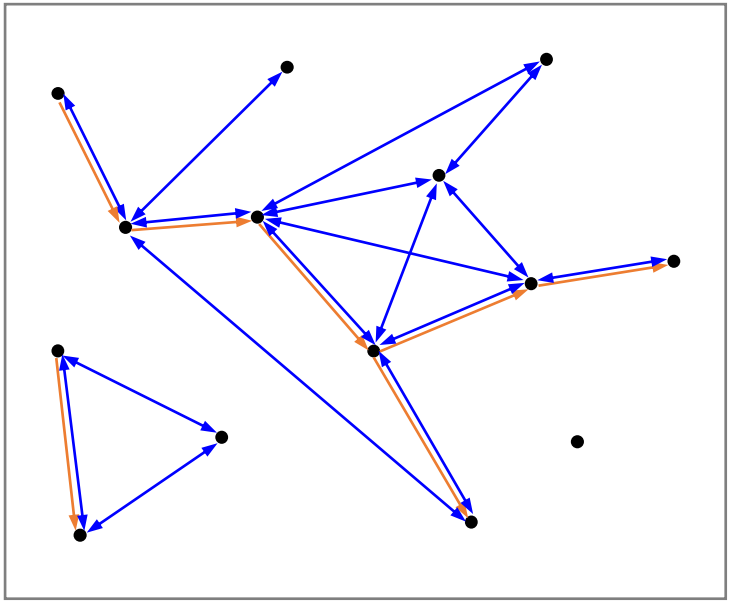
Institute of Molecular and Translational Medicine
Palacky University

pavlo.polishchuk@upol.cz

Reaction-based vs. fragment-based

	Reaction-based	Fragment-based
Prerequisites:	reaction rules set database of building blocks	database of fragments
Abilities & issues:	• molecules are more likely to be feasible • not all moves are allowed • usually only increase complexity • some molecules can be unreachable	• do not control synthetic feasibility • many moves are allowed • arbitrary direction of exploration • cover larger chemical space





Chemically reasonable mutations (CReM)

Polishchuk *J Cheminform* (2020) 12:28
https://doi.org/10.1186/s13321-020-00431-w

Journal of Cheminformatics

SOFTWARE Open Access

CReM: chemically reasonable mutations framework for structure generation

Pavel Polishchuk*

Abstract

Structure generators are widely used in de novo design studies and their performance substantially influences an outcome. Approaches based on the deep learning models and conventional atom-based approaches may result in invalid structures and fail to address their synthetic feasibility issues. On the other hand, conventional reaction-based approaches result in synthetically feasible compounds but novelty and diversity of generated compounds may be limited. Fragment-based approaches can provide both better novelty and diversity of generated compounds but the issue of synthetic complexity of generated structure was not explicitly addressed before. Here we developed a new framework of fragment-based structure generation that, by design, results in the chemically valid structures and provides flexible control over diversity, novelty, synthetic complexity and chemotypes of generated compounds. The framework was implemented as an open-source Python module and can be used to create custom workflows for the exploration of chemical space.

Keywords: De novo structure generation, De novo design, Matched molecular pairs

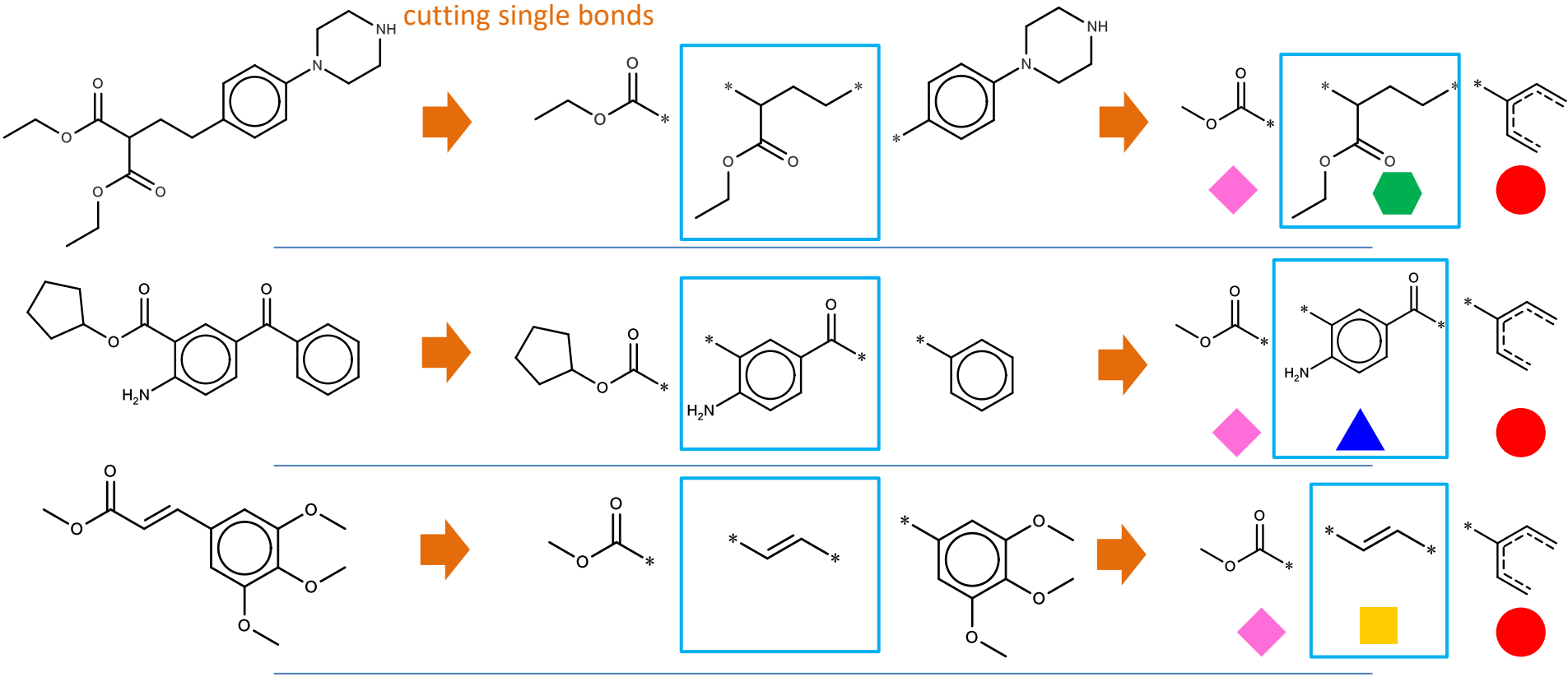
CReM is …

CReM is … not a Swiss army knife

Chemically reasonable mutations (CReM)

exhaustive fragmentation
cutting single bonds

taking context of radius R (here R = 3)



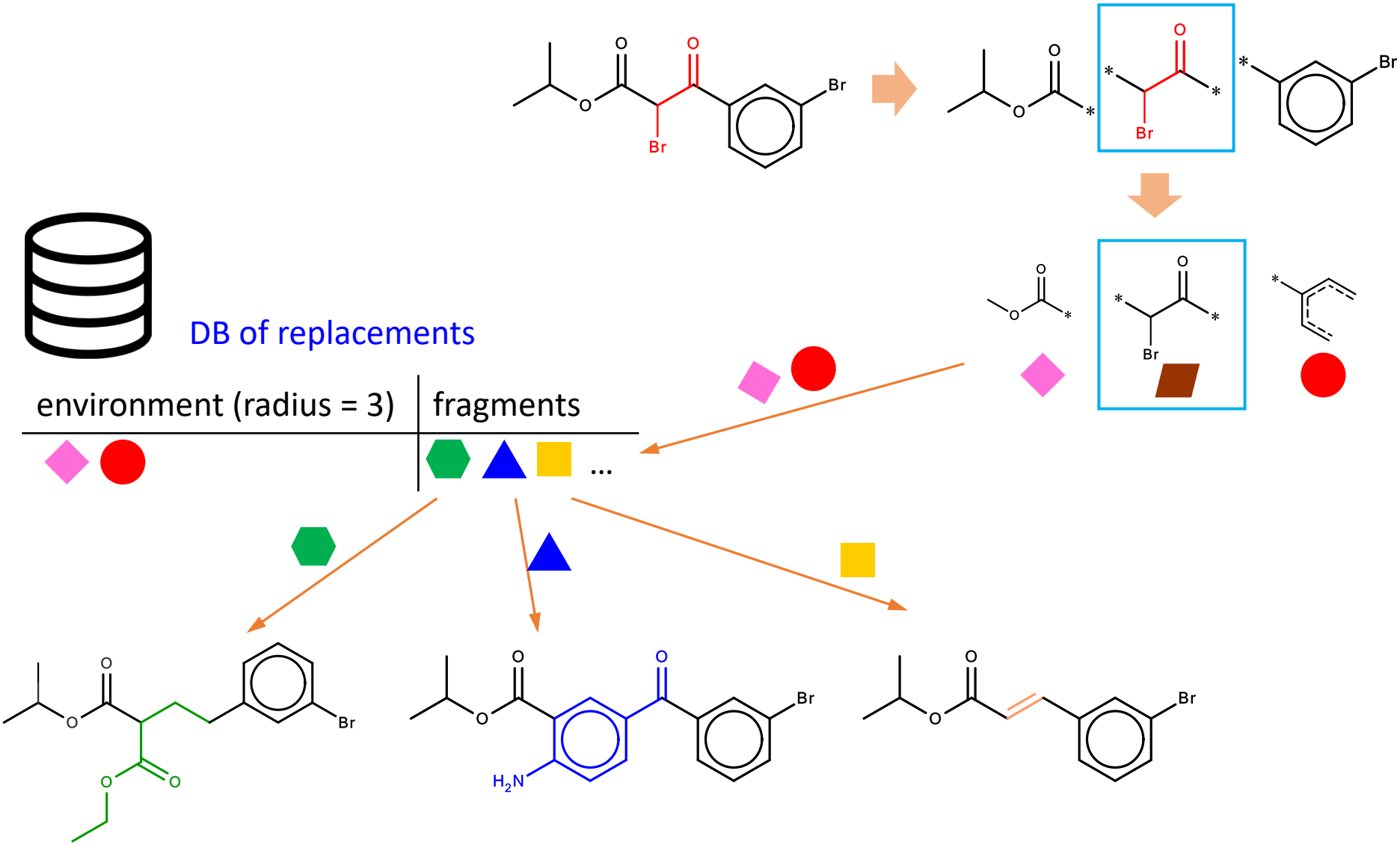
DB of replacements

environment (radius = 3)	fragments
◆ ●	⬢ ▲ ■ …
…	…

interchangeable fragments

Polishchuk, P., CReM: chemically reasonable mutations framework for structure generation. *J. Cheminf.* **2020**, 12 (1), 28.

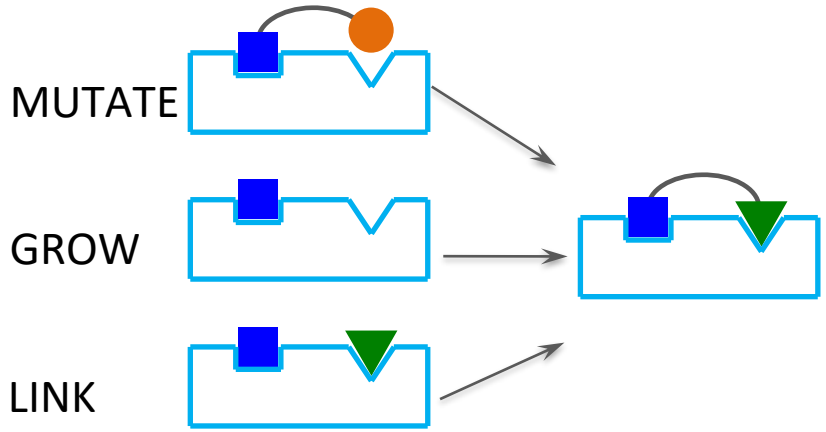
Chemically reasonable mutations (CReM)

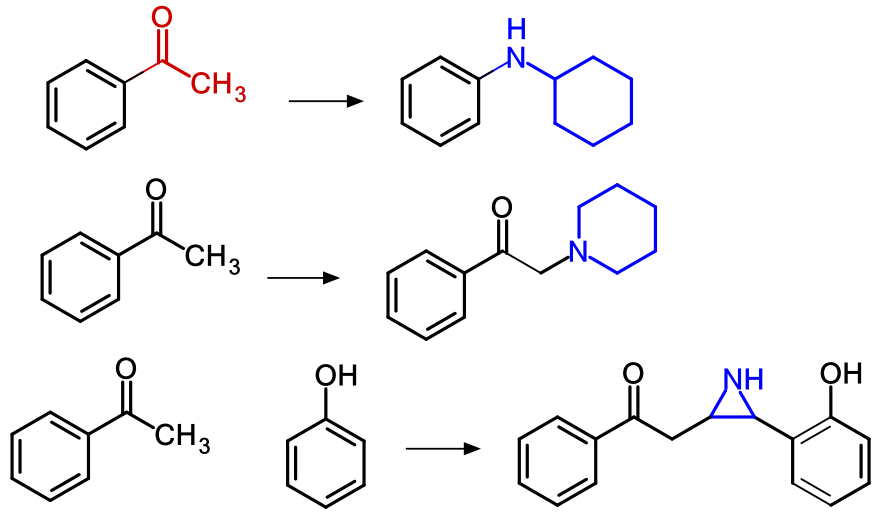


Generated structures are always chemically valid!

Polishchuk, P., CReM: chemically reasonable mutations framework for structure generation. *J. Cheminf.* **2020**, 12 (1), 28.

CReM features & applications

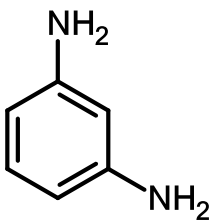




- use a **custom** (in-house) **fragment database** to generate more synthetically accessible compounds enriched with specific chemotypes
- choose larger **radiuses** to make replacements more conservative and resulting to more synthetically accessible compounds
- specify the **size of replaced and replacing fragments** to control granularity of steps in chemical space
- specify **atoms** to **protect** **or** **replace** to direct structural modifications
- specify the topological **distance** between attachment points in a linker

1. Scaffold decoration
2. Enumeration of analog series
3. Hit expansion
4. Lead optimization
5. De novo design

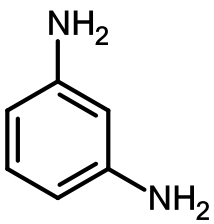
Radius of chemical context



GROW

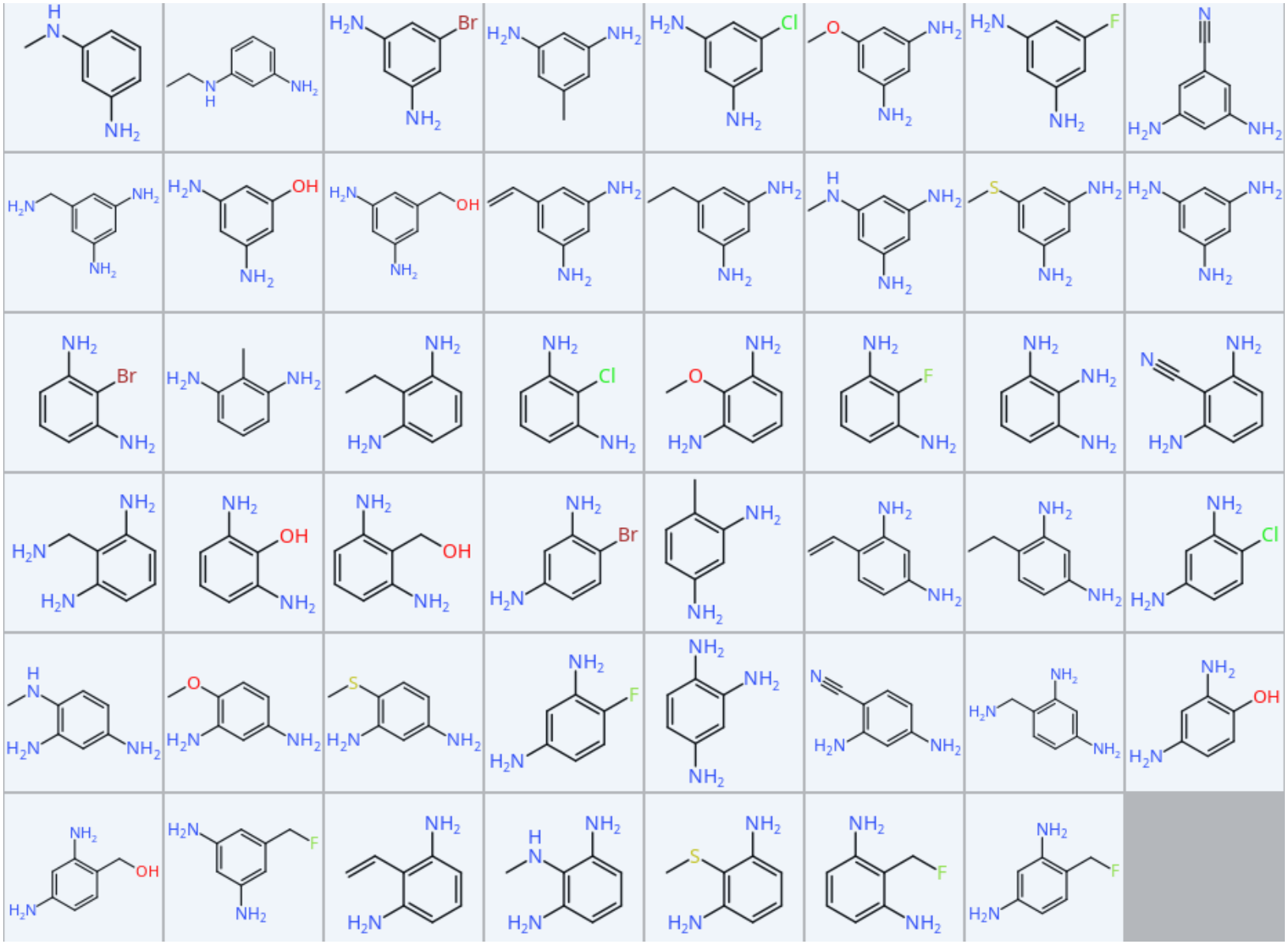
max_atoms=2

Radius of chemical context



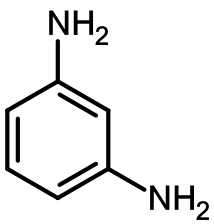
GROW

max_atoms=2



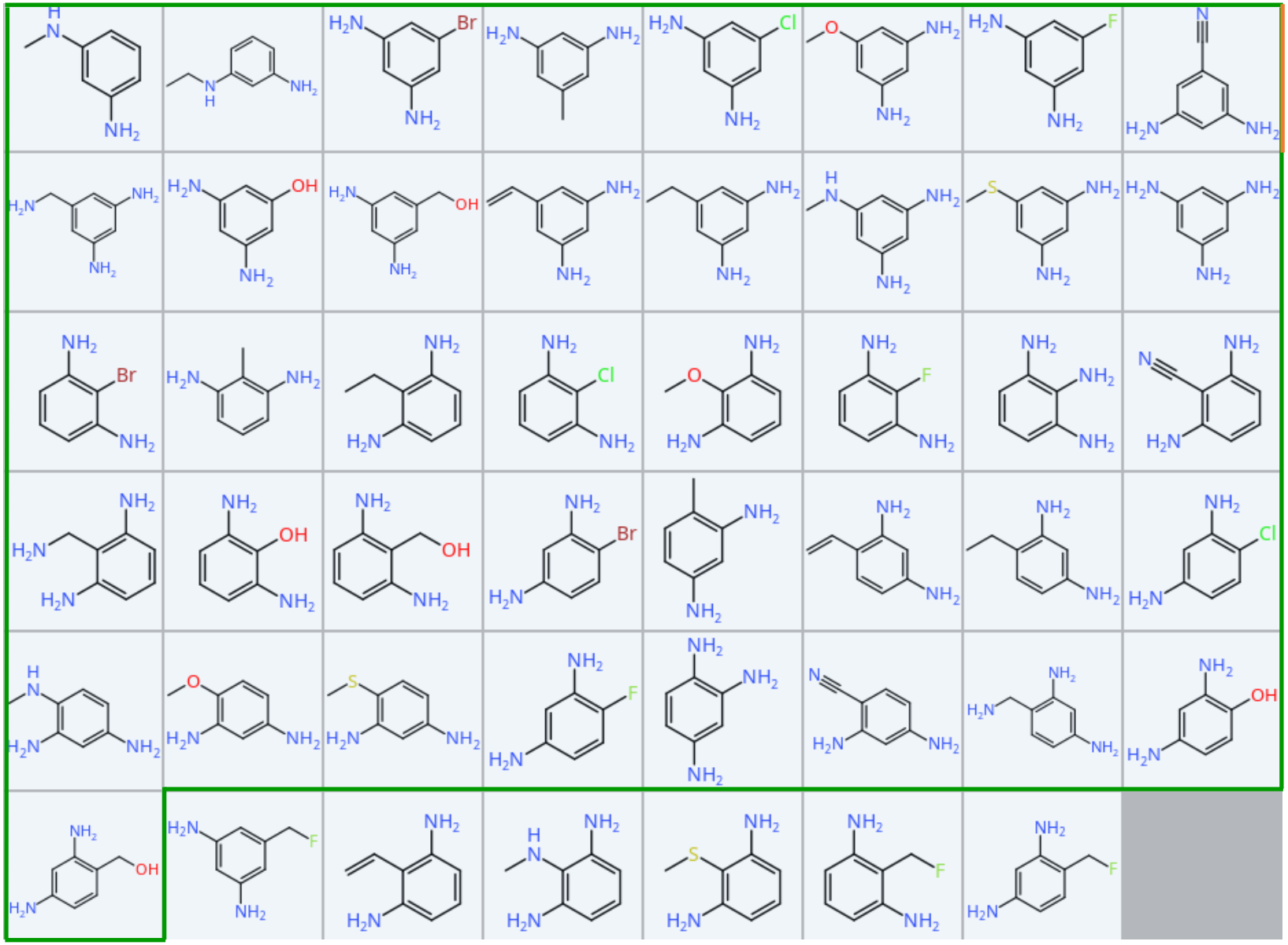
Radius 1

Radius of chemical context



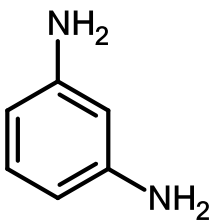
GROW

max_atoms=2



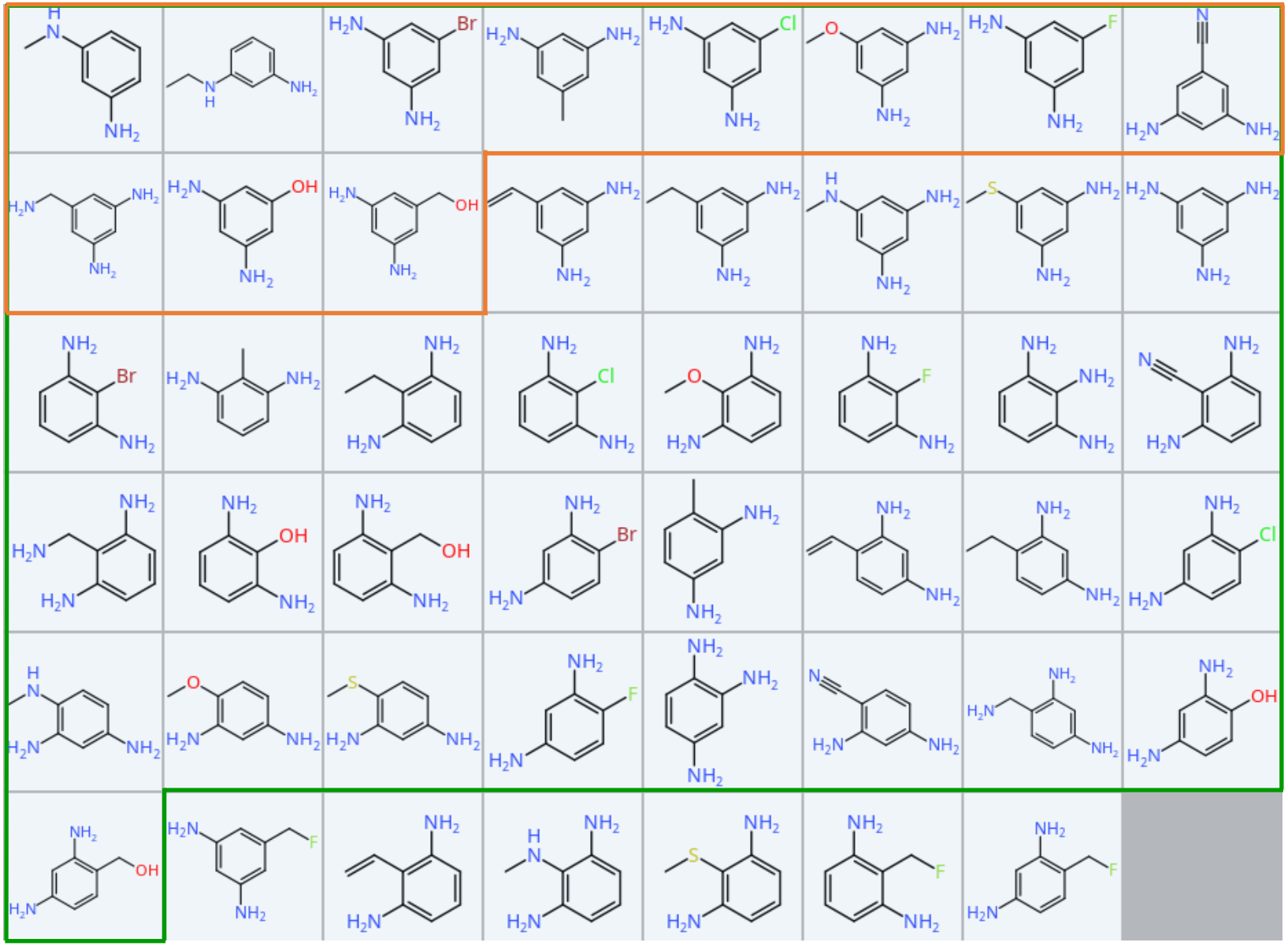
Radius 2

Radius of chemical context



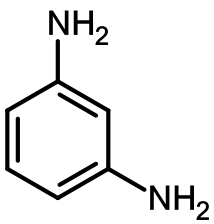
GROW

max_atoms=2



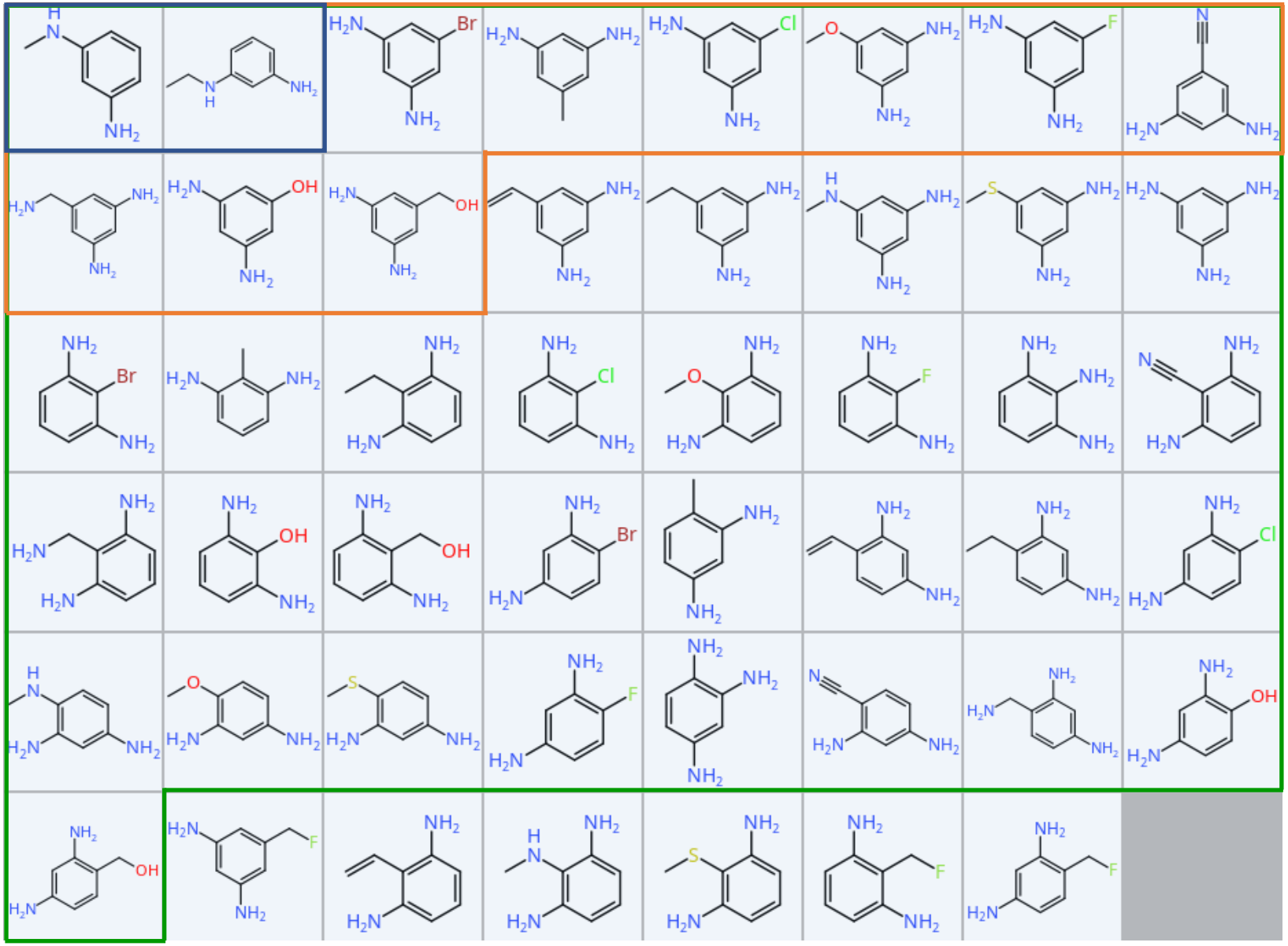
Radius 3

Radius of chemical context



GROW

max_atoms=2

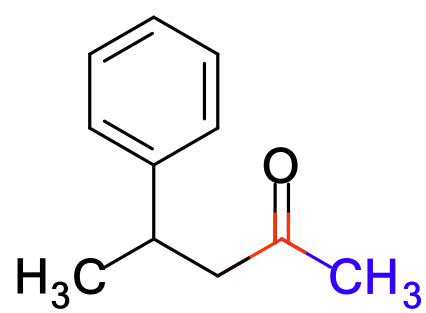


Radius 4

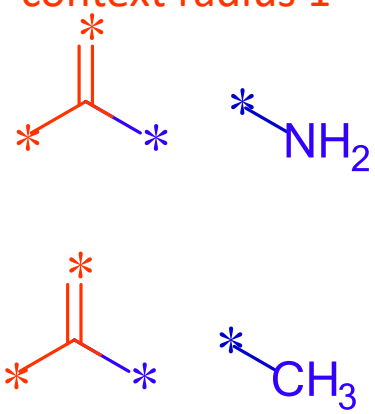
Radius of chemical context

Generated new chemotypes will have a size greater than a selected radius

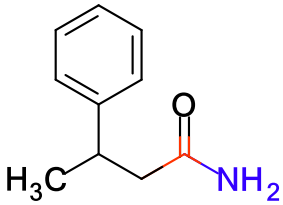
CReM DB - no amides

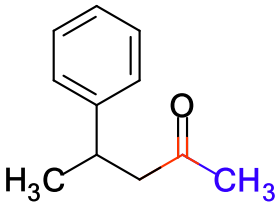


context radius 1



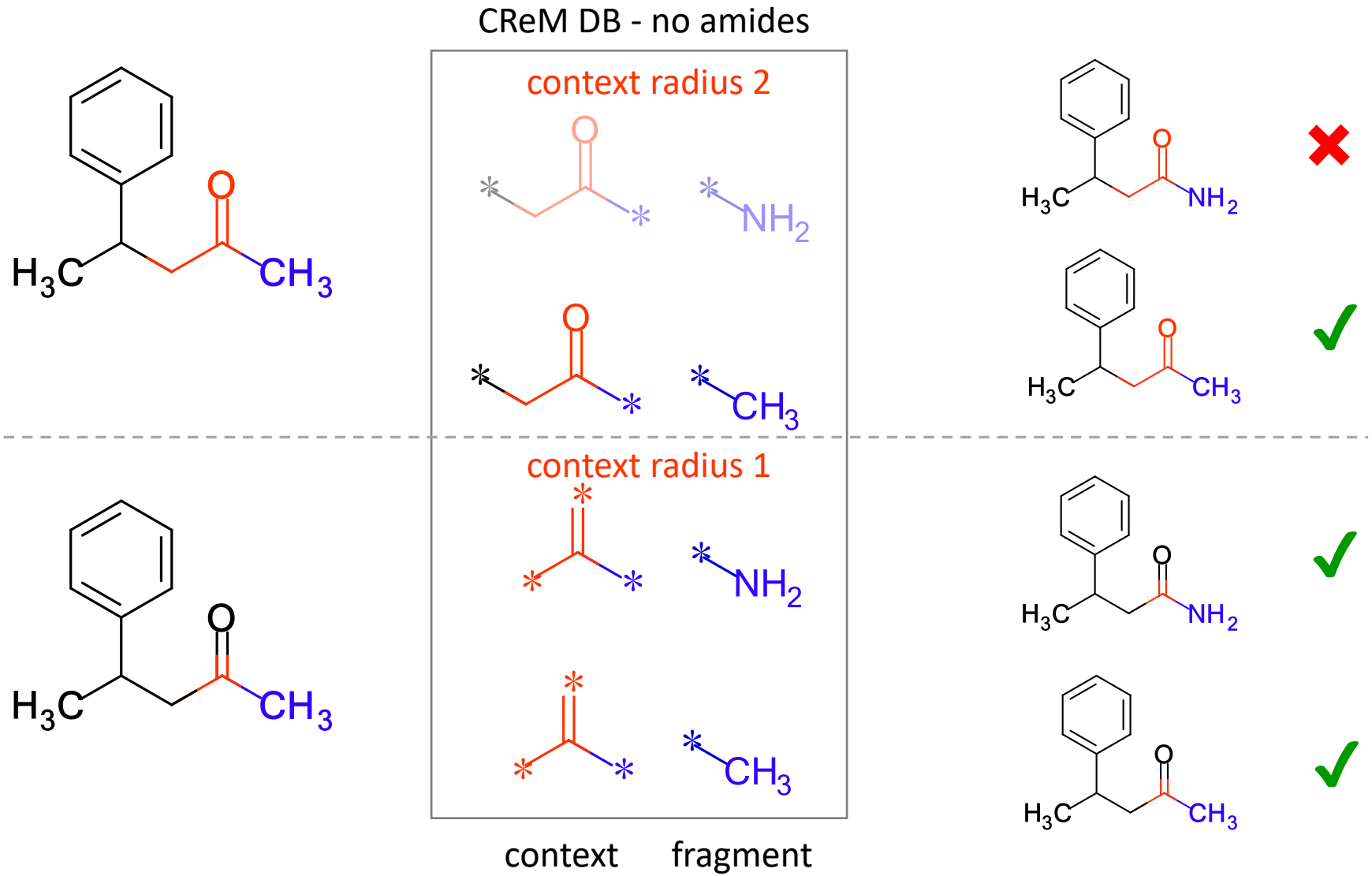
context fragment



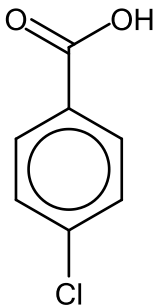


Radius of chemical context

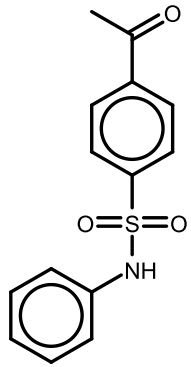
Generated new chemotypes will have a size greater than a selected radius



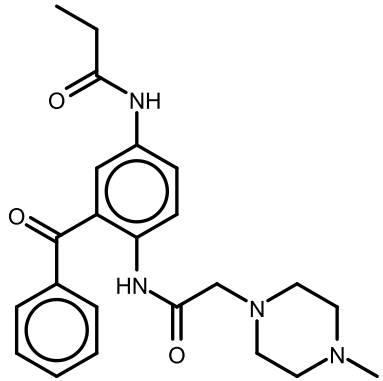
Synthetic accessibility of compounds



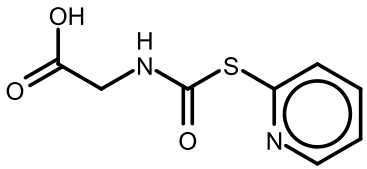
1.2
CHEMBL618



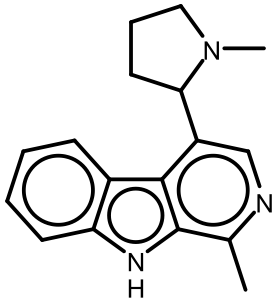
1.5
CHEMBL3310985



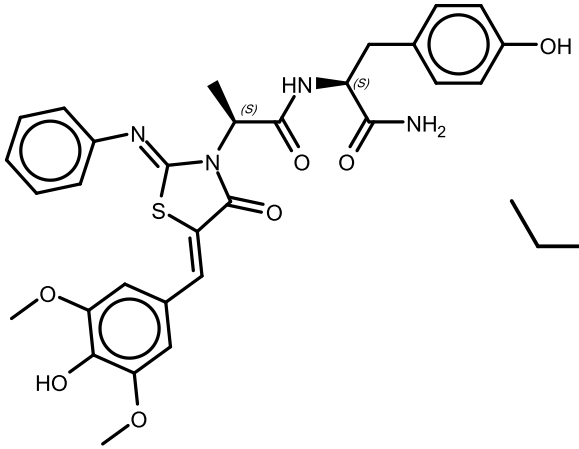
2.0
CHEMBL595820



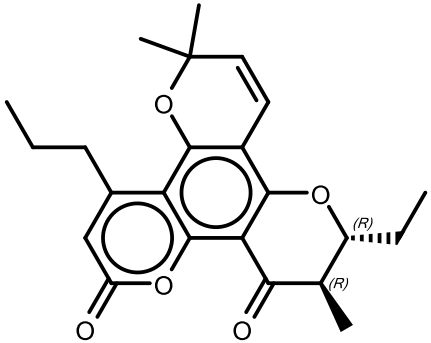
2.5
CHEMBL503660



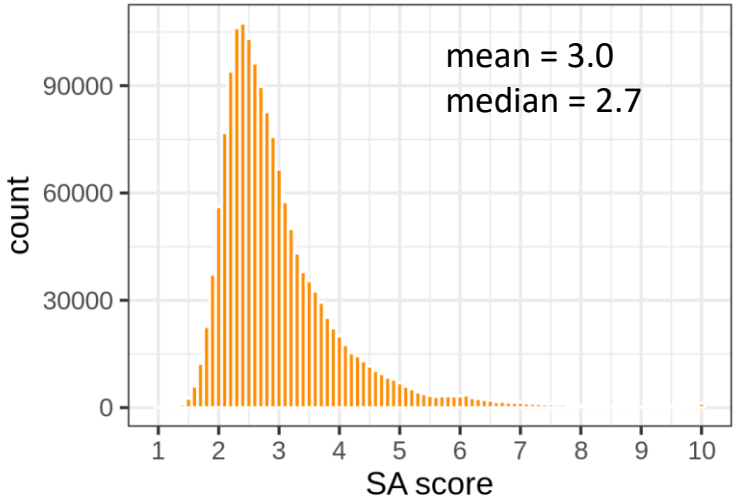
3.0
CHEMBL500286



3.5
CHEMBL582554

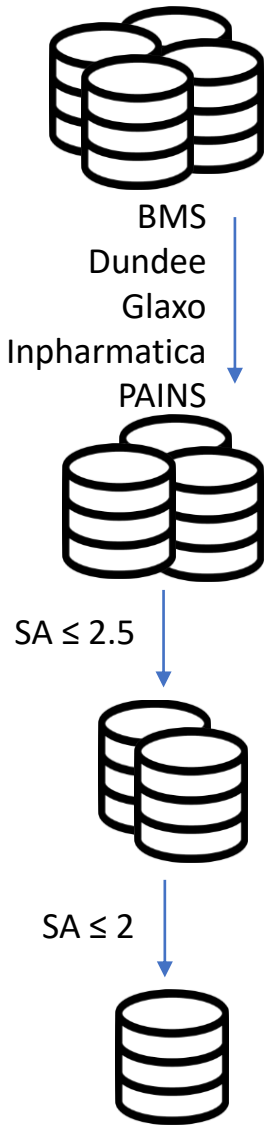


4.0
CHEMBL7633



Ertl, P.; Schuffenhauer, A., Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009**, 1, 8. (10.1186/1758-2946-1-8)

Synthetic accessibility of compounds



ChEMBL22
(1.55 M)

CReM DB	n (fragmented molecules)	n (distinct fragments, 12 atoms)
all	818 174	988 585
SA2.5 (SA ≤ 2.5)	338 422	272 988
SA2 (SA ≤ 2)	67 970	55 498

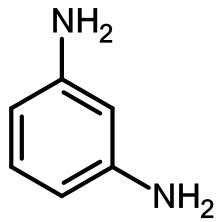
Estimated size of covered chemical space

CReM DB	radius	size
all	3	2.8×10^{17}
SA2.5	3	4.2×10^{16}
SA2	2	8.4×10^{16}
SA2	3	1.8×10^{15}
SA2	4	2.7×10^{13}

Pre-compiled
CReM databases (zenodo)

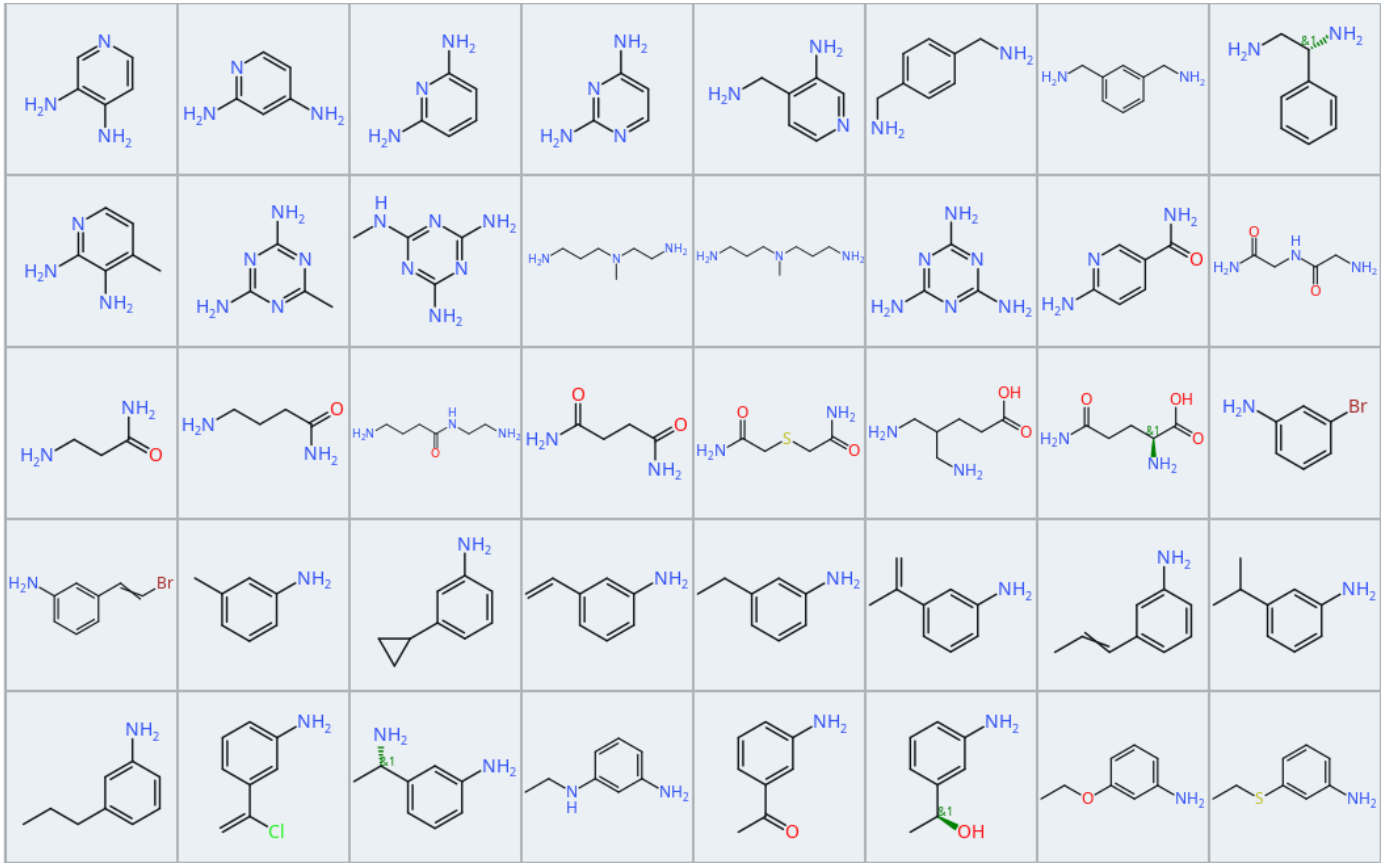
https://qsar4u.com/pages/crem.php

Synthetic accessibility of compounds

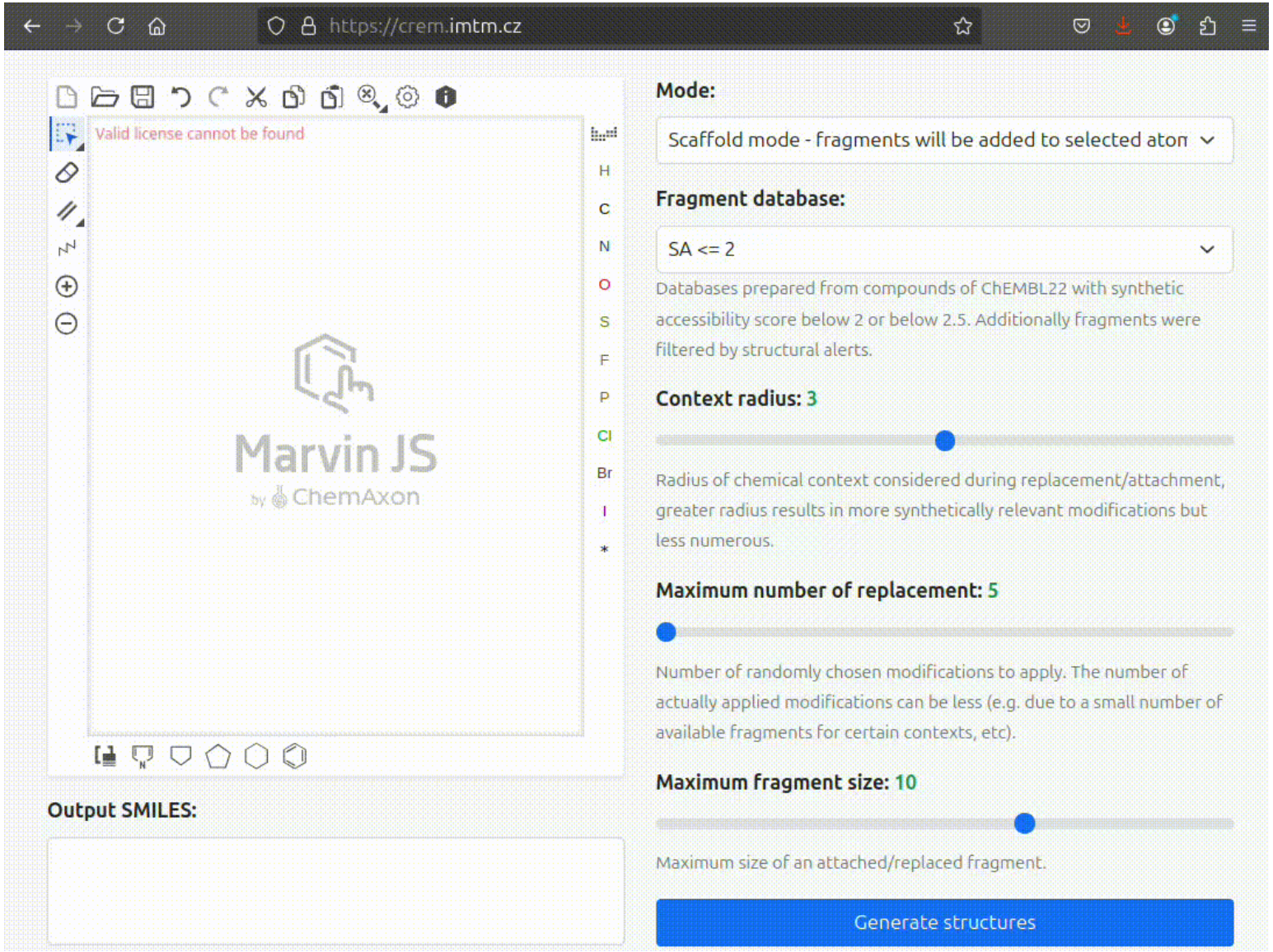


MUTATE

	radius 1	radius 2	radius 3	radius 4	radius 5
ChEMBL SA2.5	329	327	323	288	288
ChEMBL SA2	161	158	154	123	123

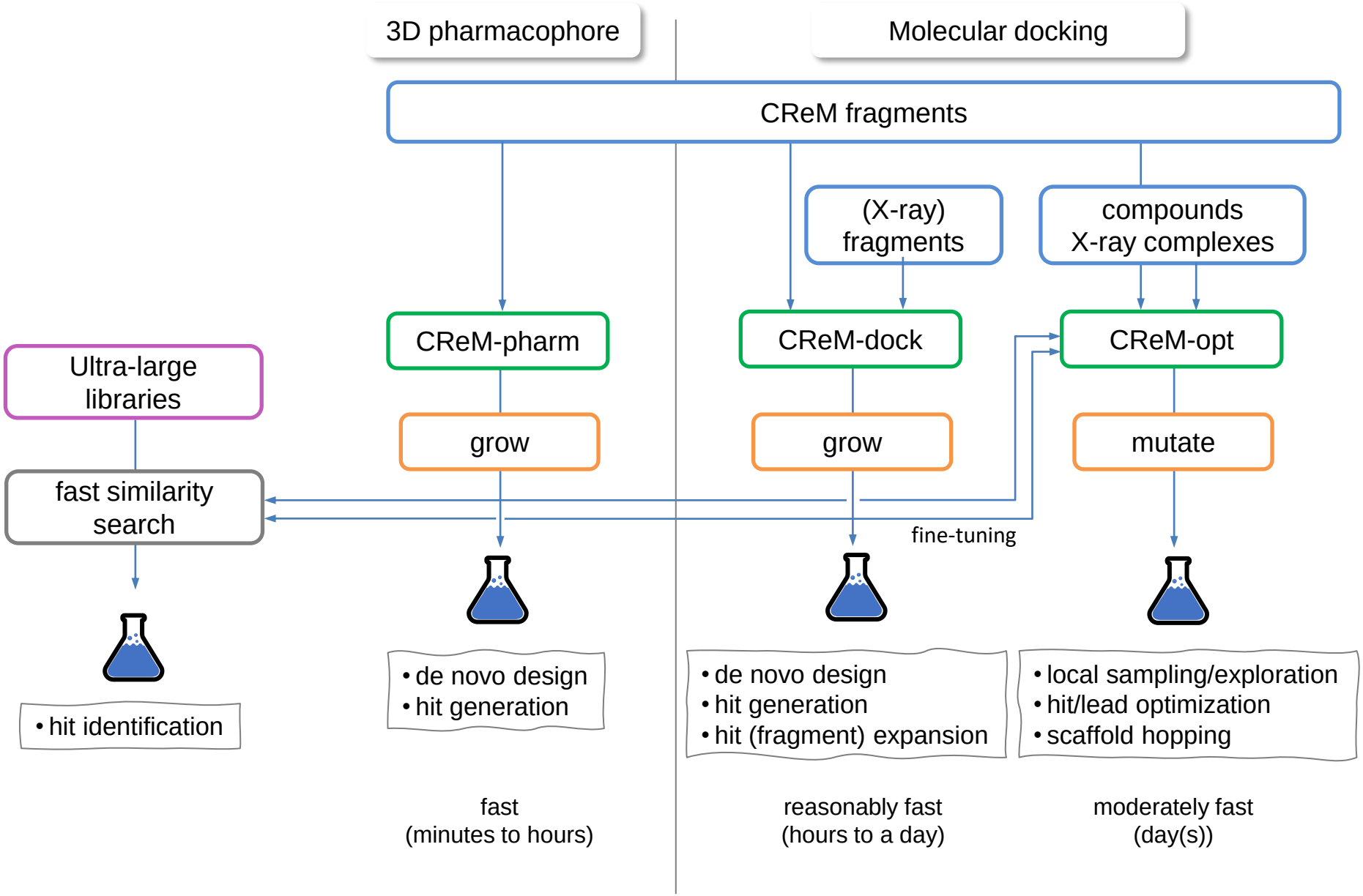


CReM online



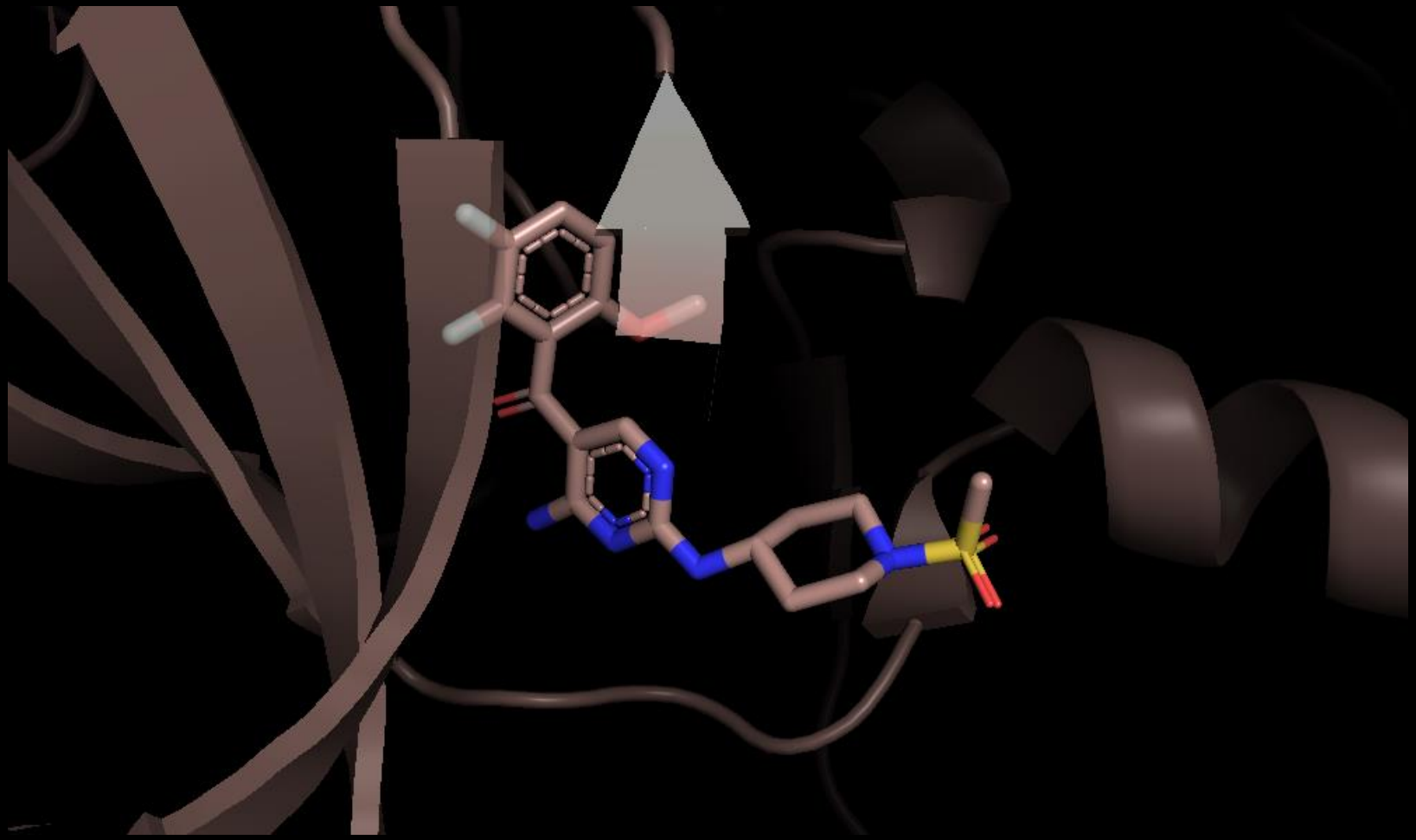
crem.imtm.cz

CReM-based applications

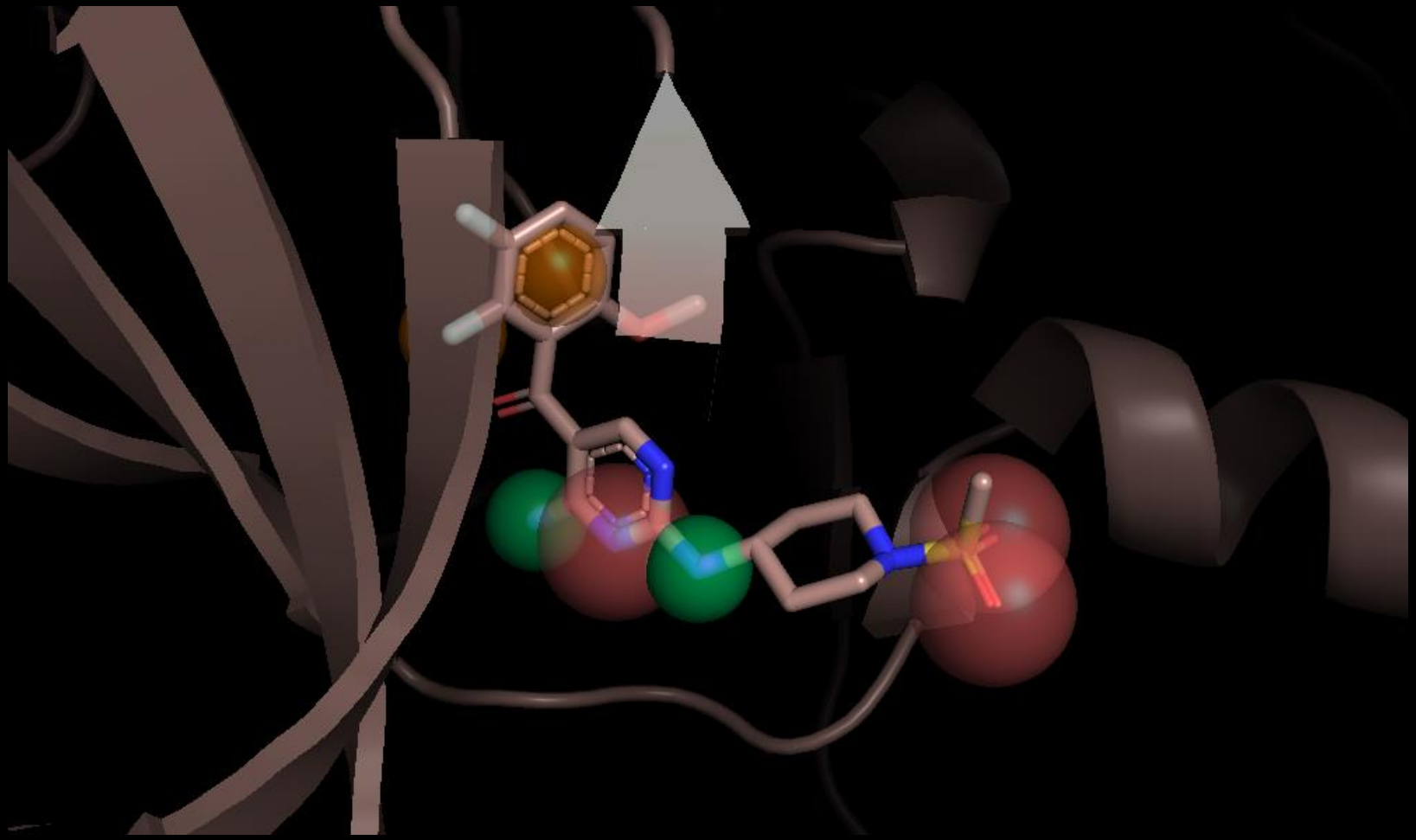


CReM-pharm

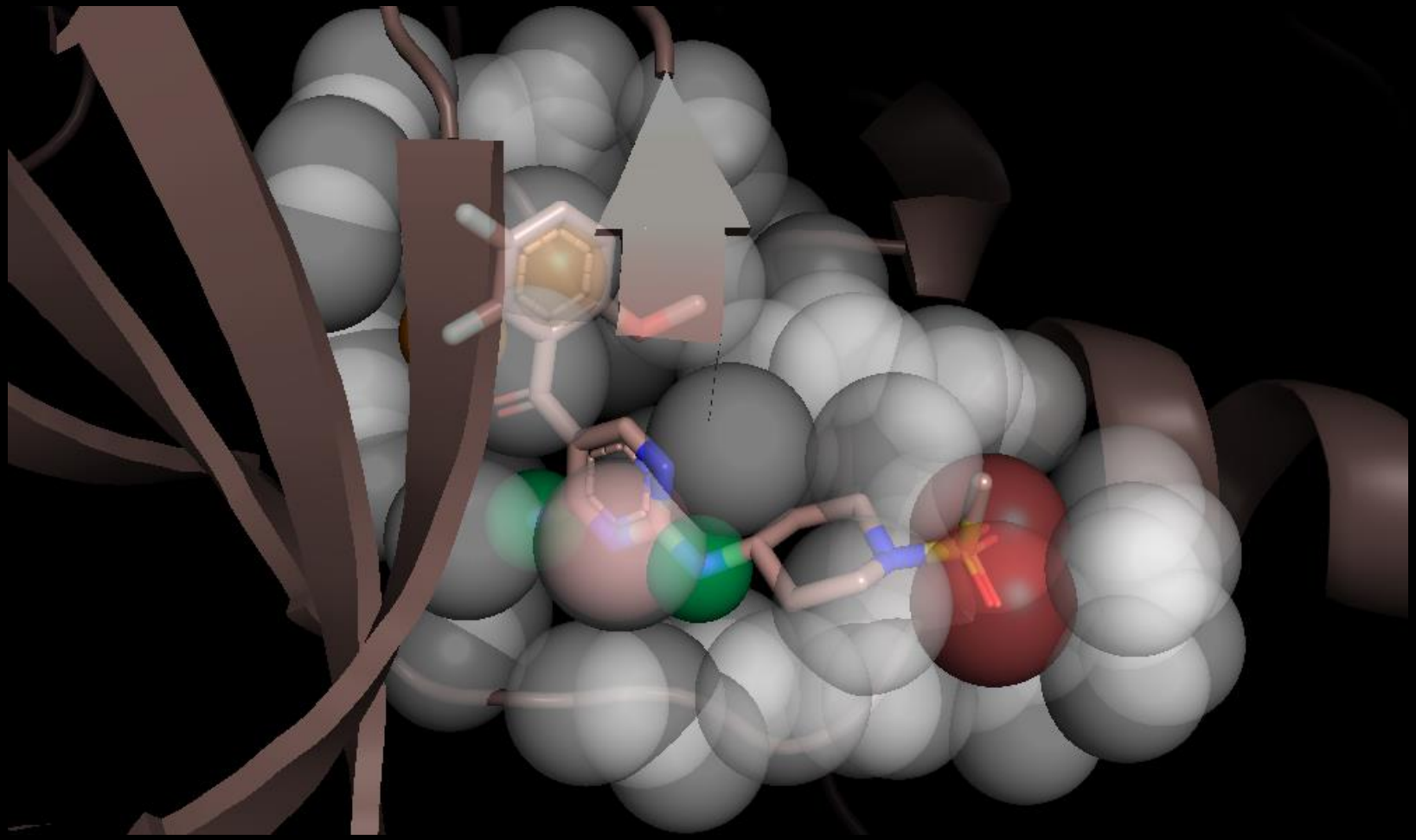
CReM-pharm: generation example



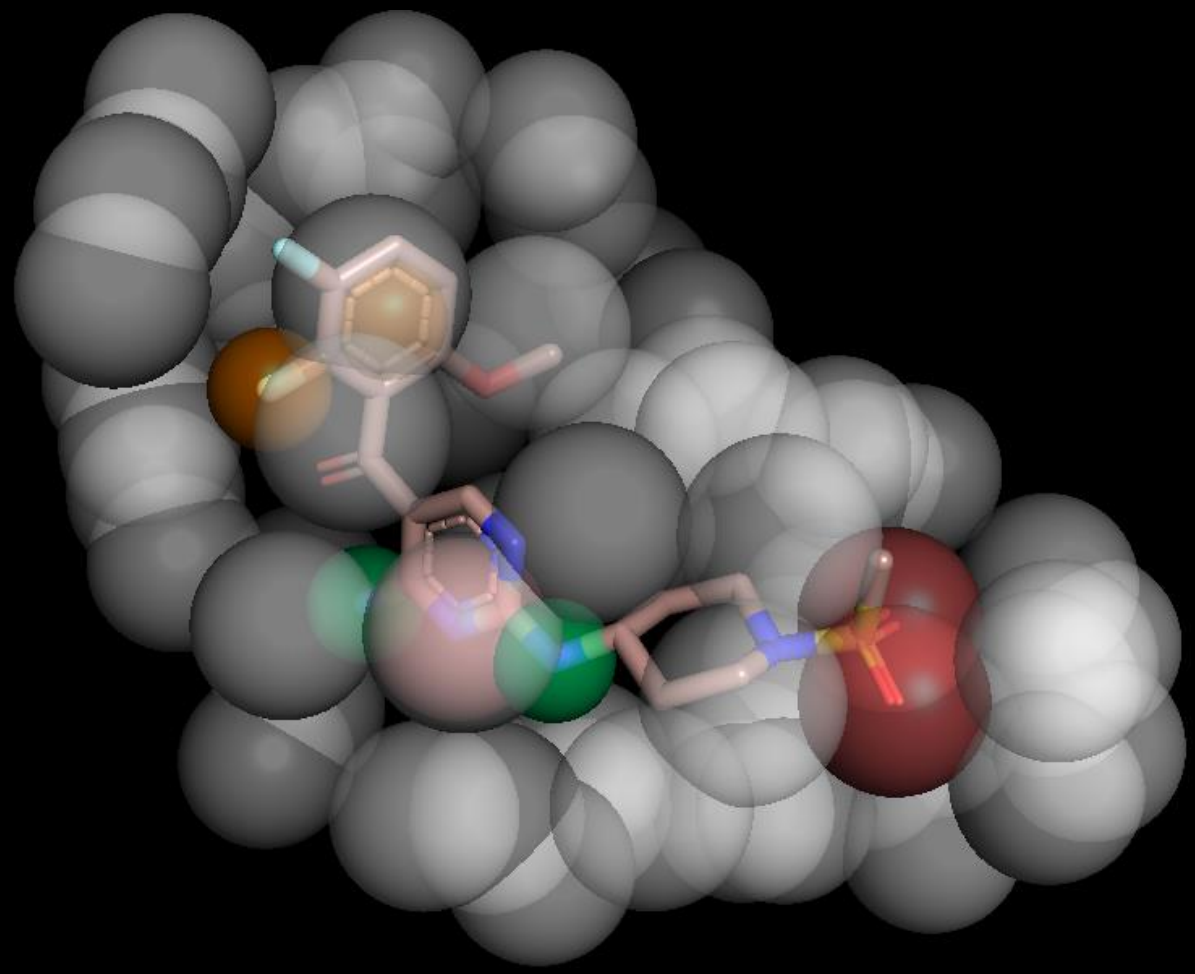
CReM-pharm: generation example



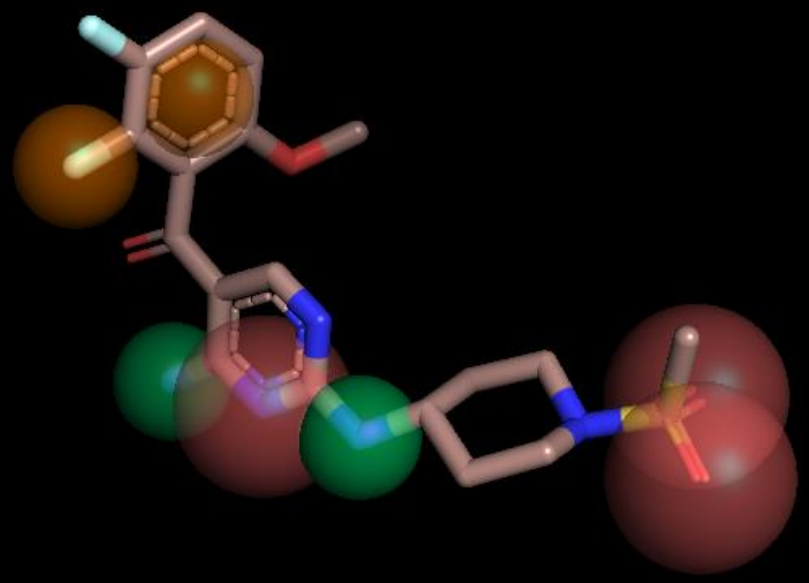
CReM-pharm: generation example



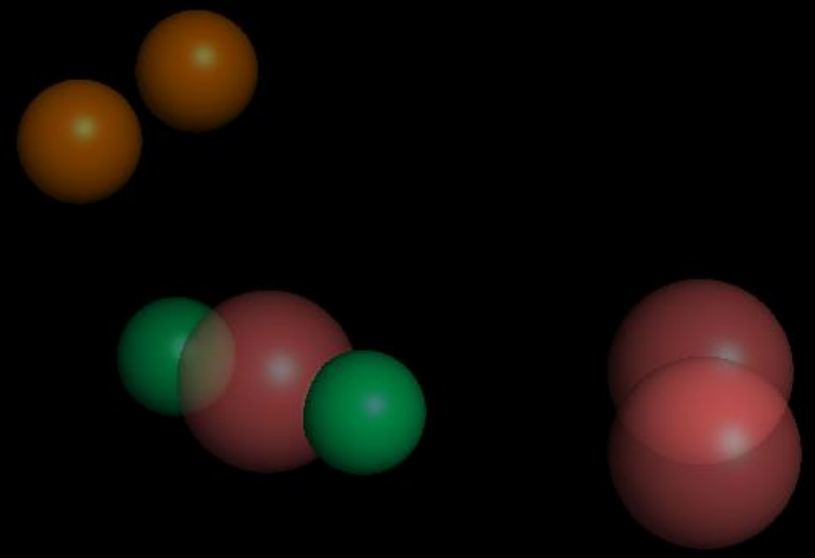
CReM-pharm: generation example



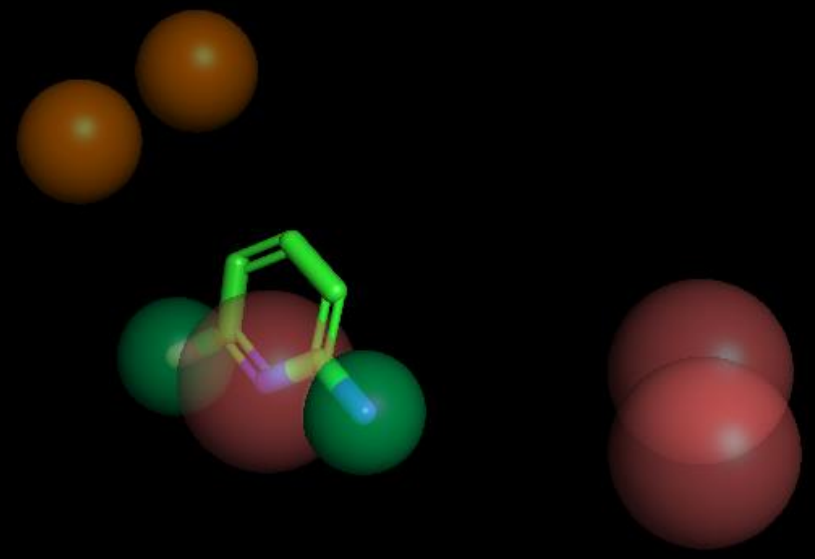
CReM-pharm: generation example



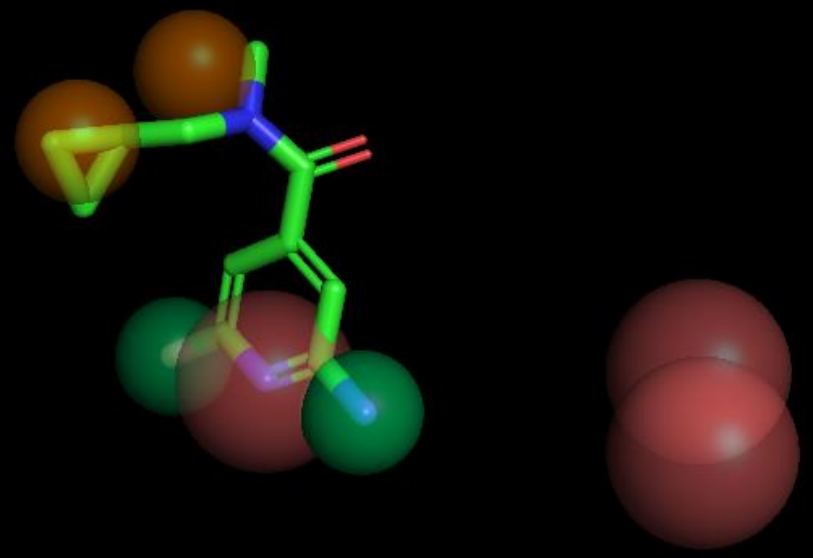
CReM-pharm: generation example



CReM-pharm: generation example

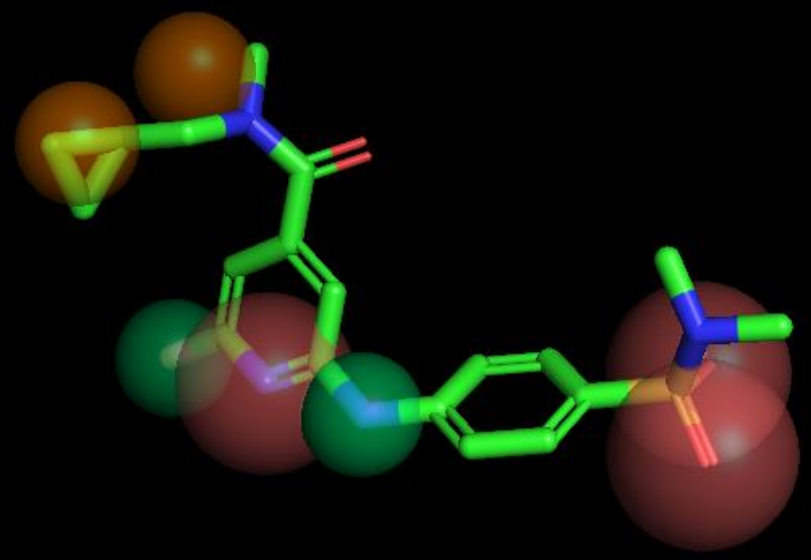


CReM-pharm: generation example

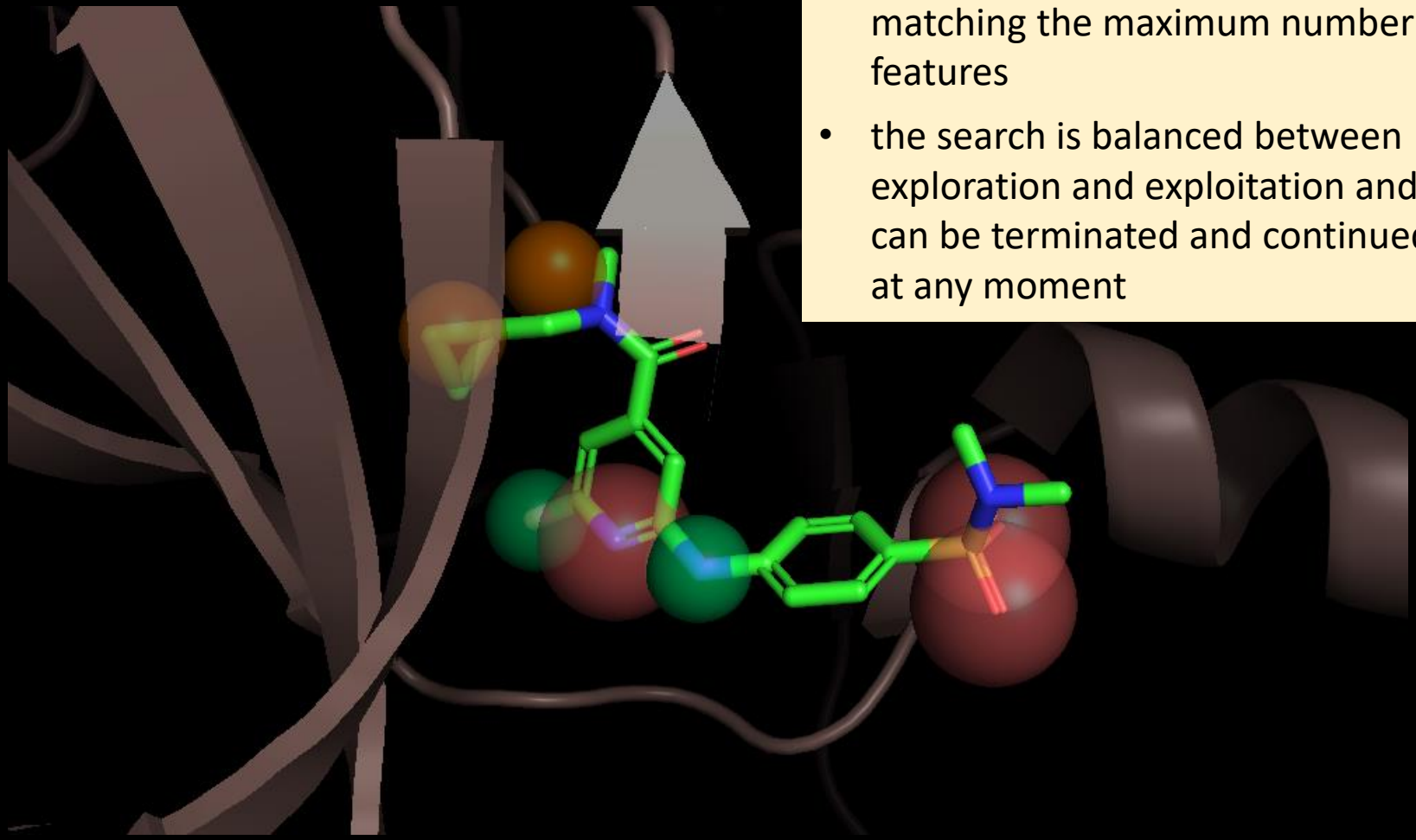


Seidel, T.; Permann, C.; Wieder, O.; Kohlbacher, S. M.; Langer, T. High-Quality Conformer Generation with CONFORGE: Algorithm and Performance Assessment. *J. Chem. Inf. Model.* **2023**, 63 (17), 5549-5570.

CReM-pharm: generation example

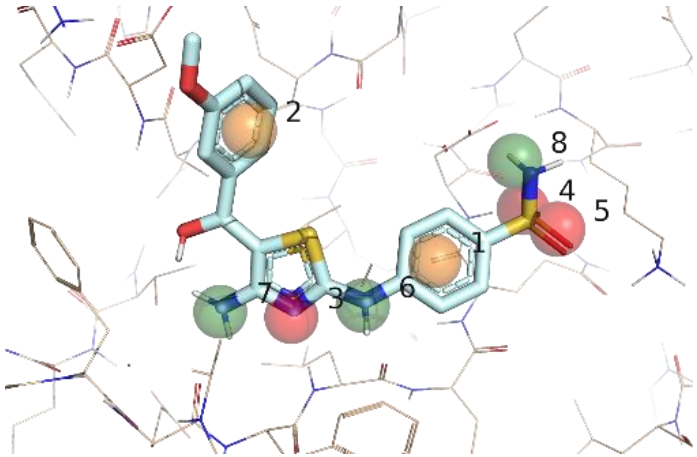


CReM-pharm: generation example



- design minimum-sized molecules matching the maximum number of features
- the search is balanced between exploration and exploitation and can be terminated and continued at any moment

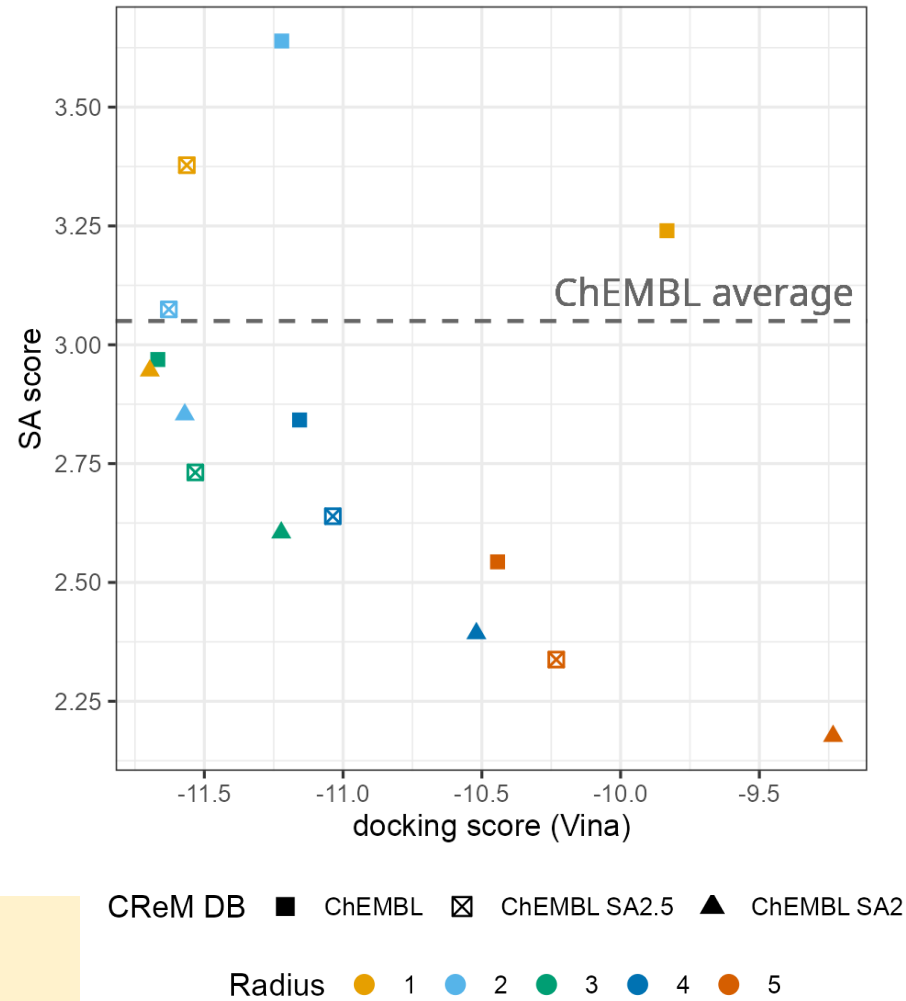
CReM-pharm: CDK2 example



3RAL

Settings:

- MW ≤ 450, logP ≤ 4, TPSA ≤ 120, RTB ≤ 7
- maximum number of replacements: all
- top 100 compounds by docking score

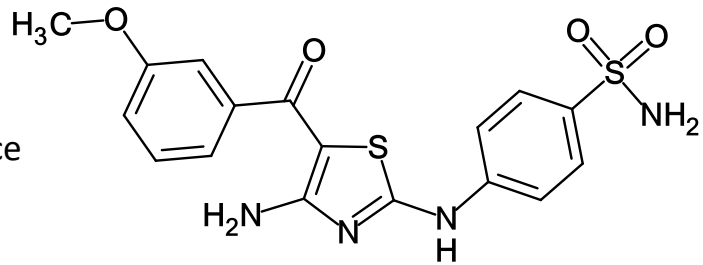


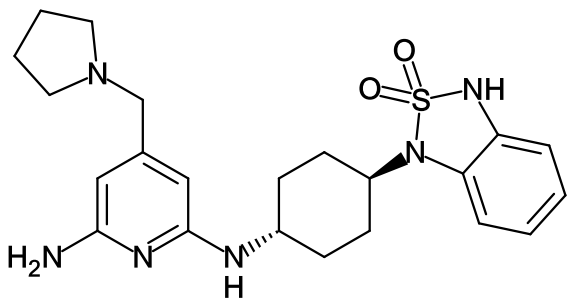
- a clear trade-off between SA and docking scores
- SA scores are predictably changed with changing of a radius and a fragment database

CReM-pharm example

CDK2 (3RAL)

reference ligand

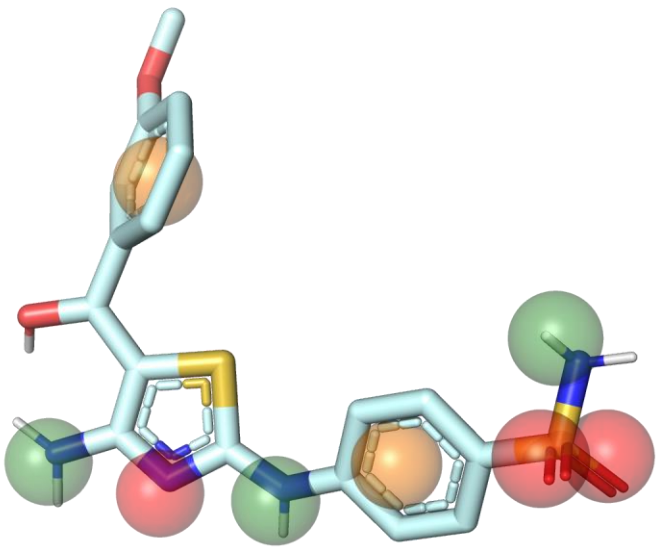


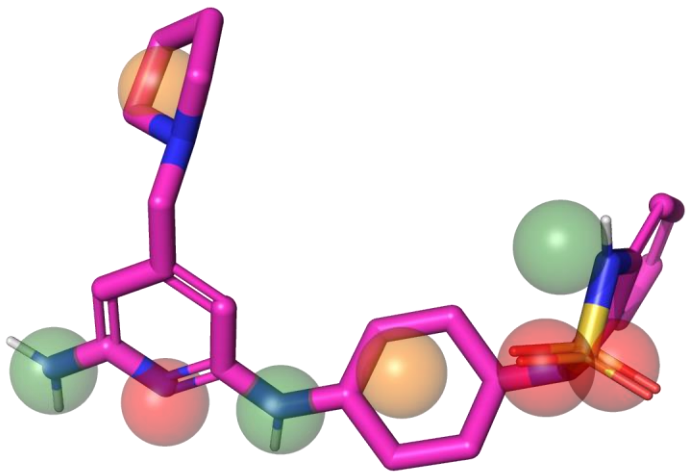


designed ligand

-11.4 / 3.0

docking score / SA score

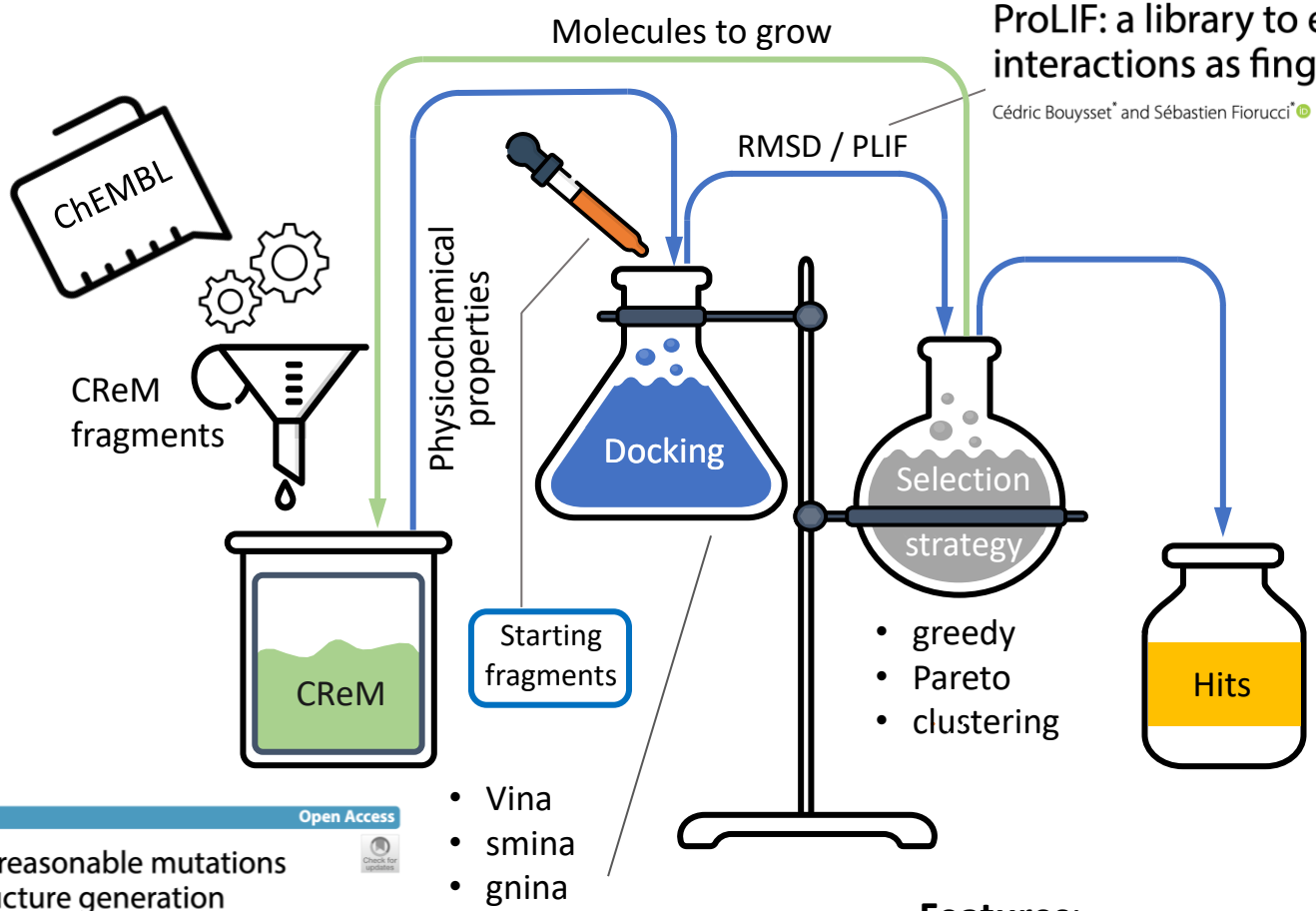




- designed compounds have high docking scores and fit to protein pockets
- SA scores are not very sensitive to complexity of pharmacophore models

CReM-dock

CReM-dock



SOFTWARE

ProLIF: a library to encode molecular interactions as fingerprints

Cédric Bouysset* and Sébastien Fiorucci*

SOFTWARE Open Access

CReM: chemically reasonable mutations framework for structure generation

Pavel Polishchuk*

SOFTWARE Open Access

EasyDock: customizable and scalable docking tool

Guzel Minibaeva[1], Aleksandra Ivanova[1] and Pavel Polishchuk[1*]

Features:

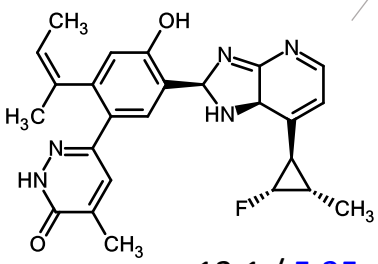
- control physicochemical properties
- control protein-ligand interactions
- keep the initial pose
- support different docking tools via EasyDock

Minibaeva and Polishchuk, CReM-dock: de novo design of synthetically feasible compounds guided by molecular docking, https://doi.org/10.26434/chemrxiv-2024-fpzqb-v2

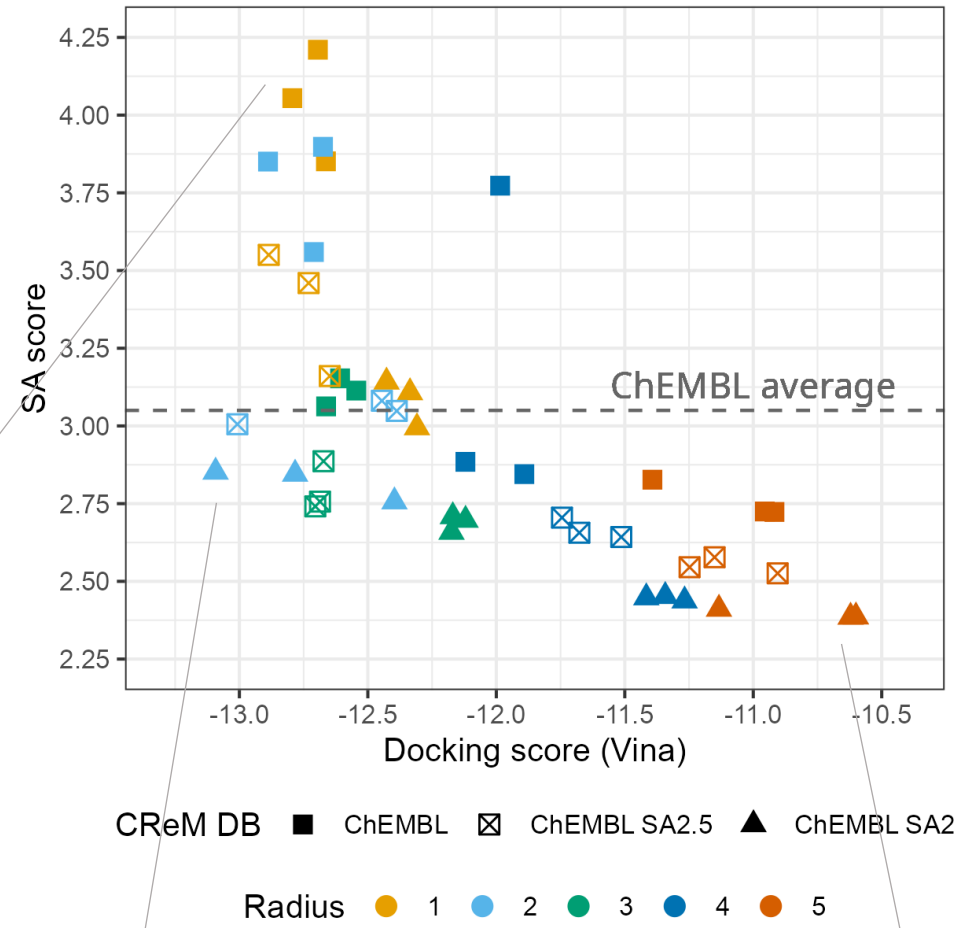
CReM-dock: CDK2 example

Settings:

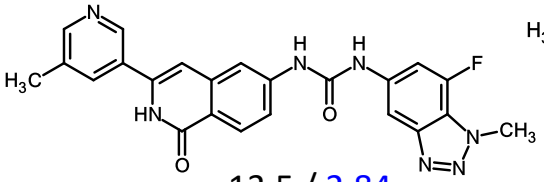
- CDK2 (2BTR)
- MW ≤ 450, logP ≤ 4, TPSA ≤ 120, RTB ≤ 7
- PLIF – hinge region interaction
- maximum number of replacements: 2000
- selection strategy: clustering (25 clusters, top 2 mols)
- 3 independent runs
- top 100 compounds by docking score



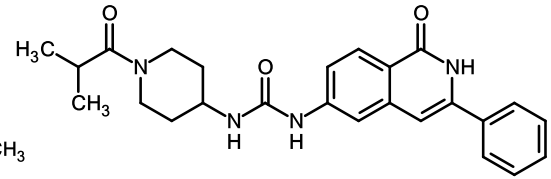
-13.1 / 5.25



- a clear trade-off between SA and docking scores
- SA scores are predictably changed with changing of a radius and a fragment database



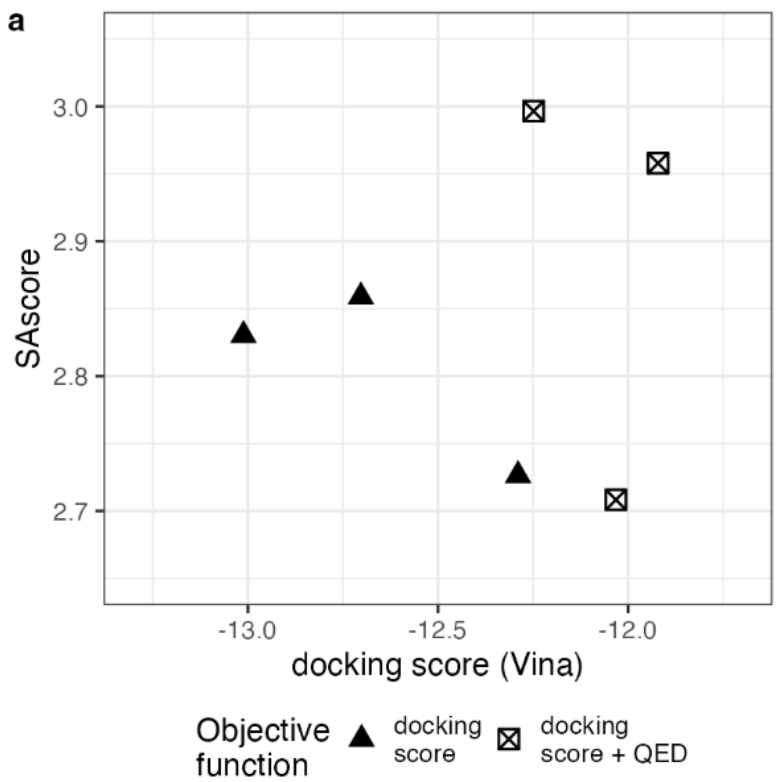
-13.5 / 2.84

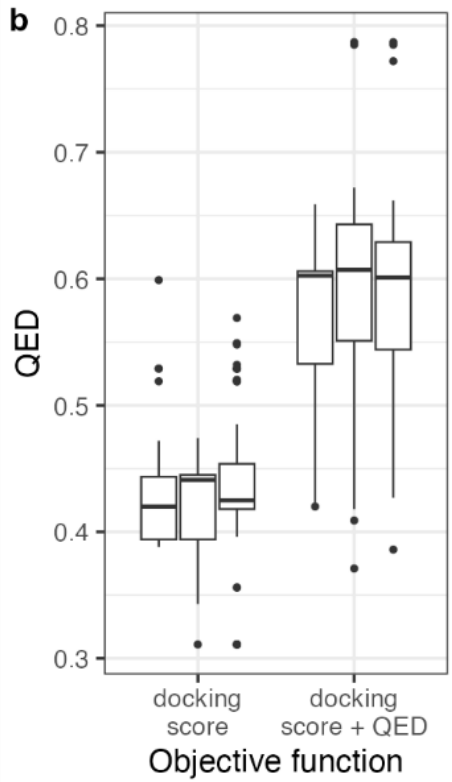


-11.8 / 2.34

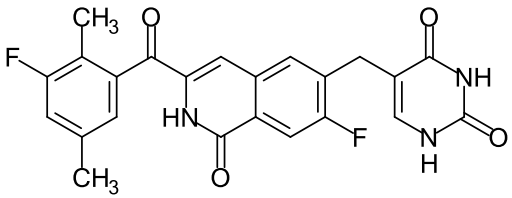
docking score / SA score

CReM-dock: augmentation with QED

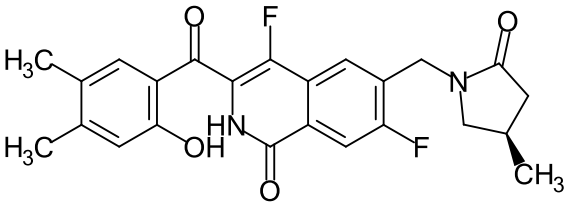




Docking + QED

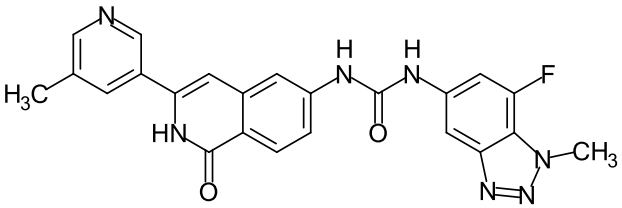


-12.8 / 2.93 / 0.43

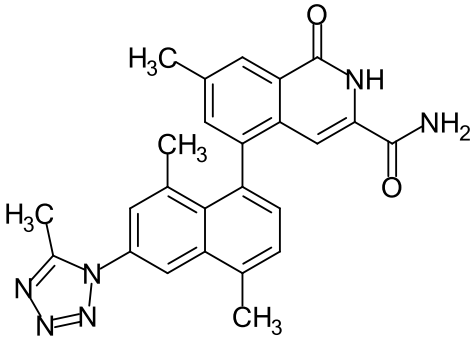


-12.3 / 3.41 / 0.63

Docking



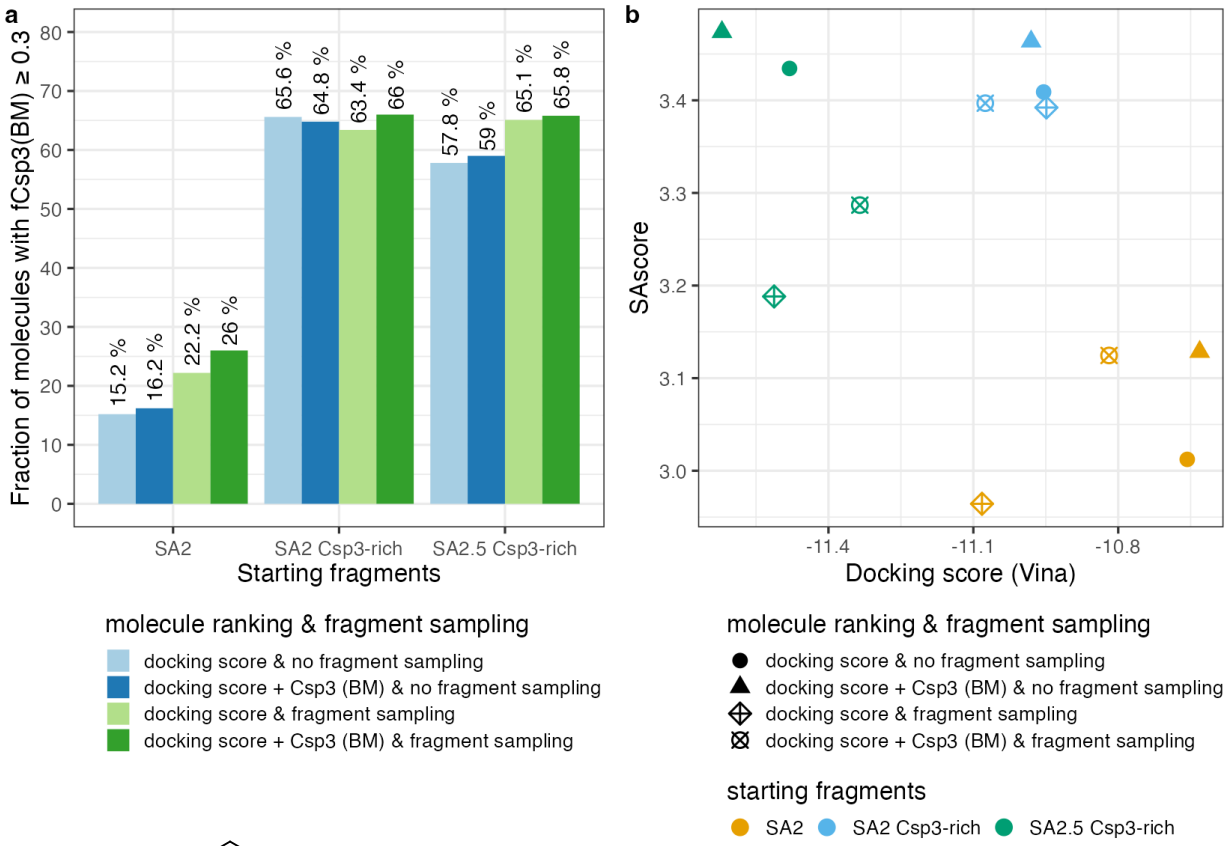
-13.5 / 2.84 / 0.39

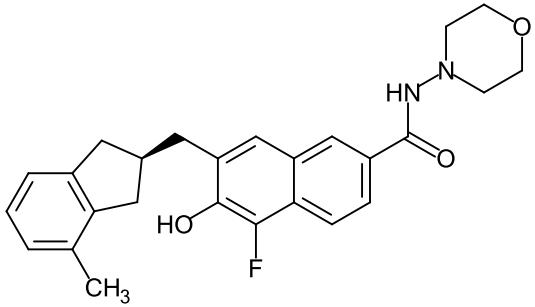


-13.8 / 2.92 / 0.45

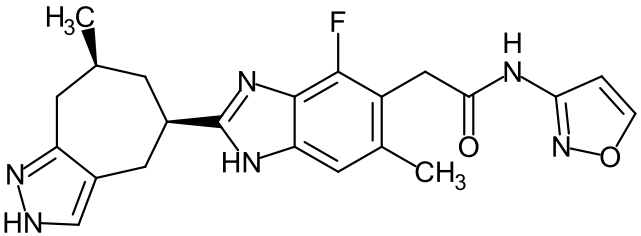
docking score / SA score / QED

CReM-dock: Csp^3-biasing

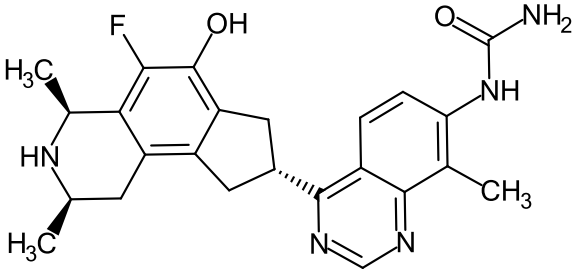




-11.7 / 3.32



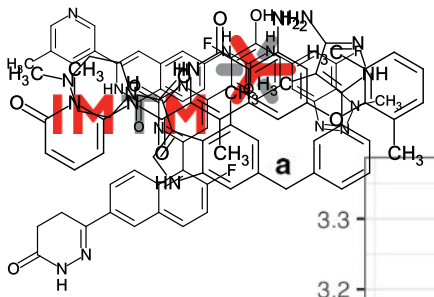
-11.8 / 4.42

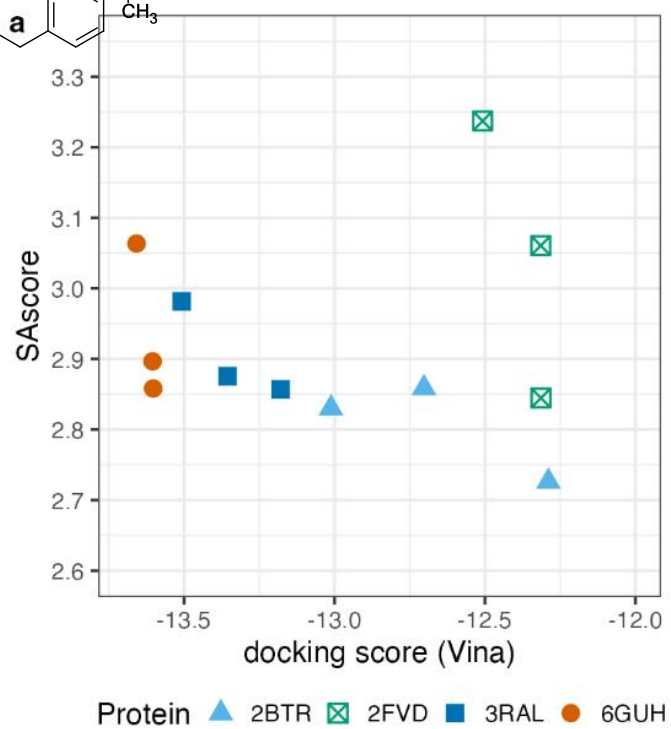


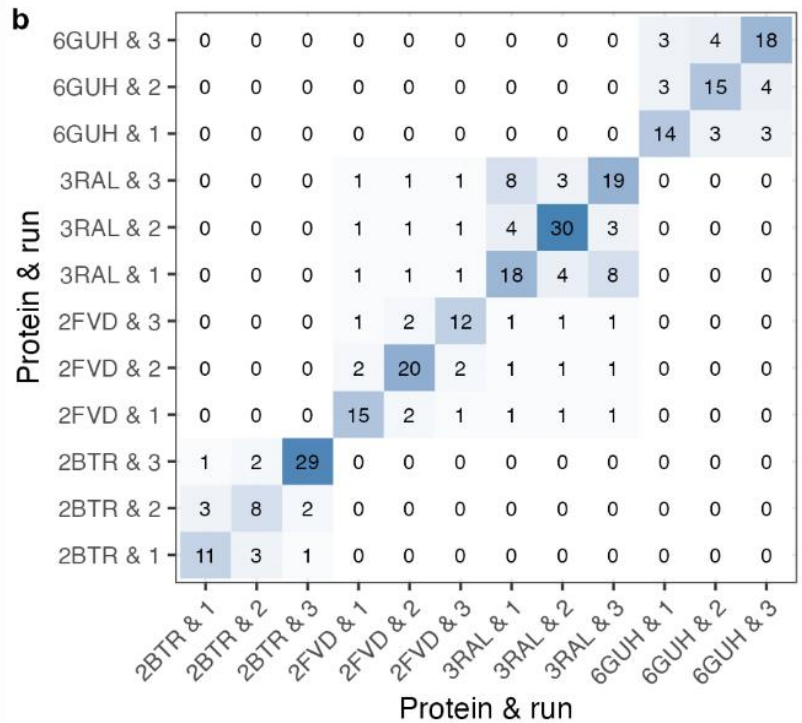
-12.3 / 4.39

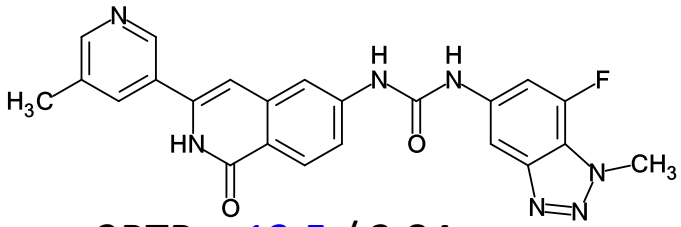
docking score / SA score

CReM-dock: different CDK2 complexes

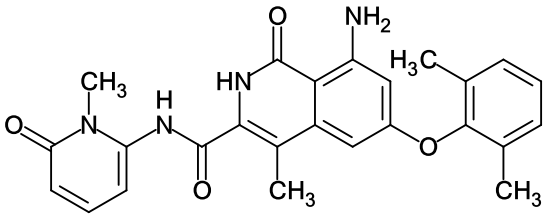




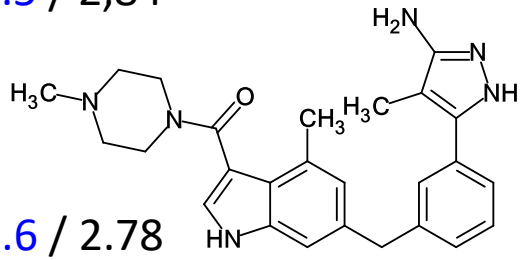




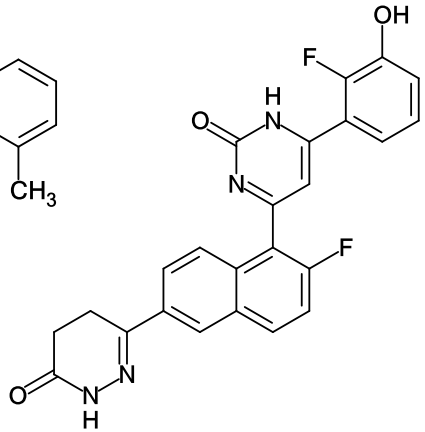
2BTR -13.5 / 2,84



3RAL -12.7 / 2.86



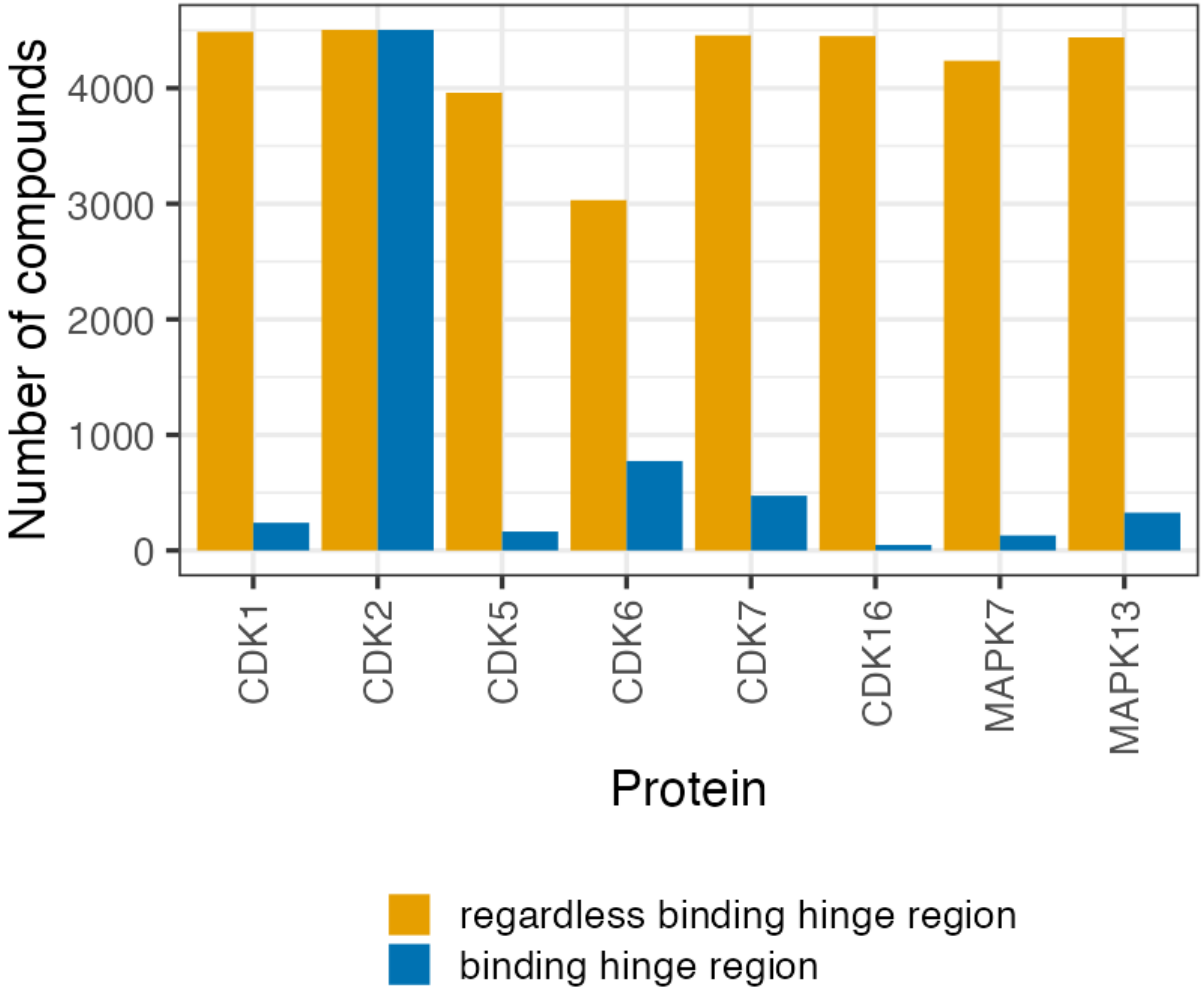
2FVD -12.6 / 2.78



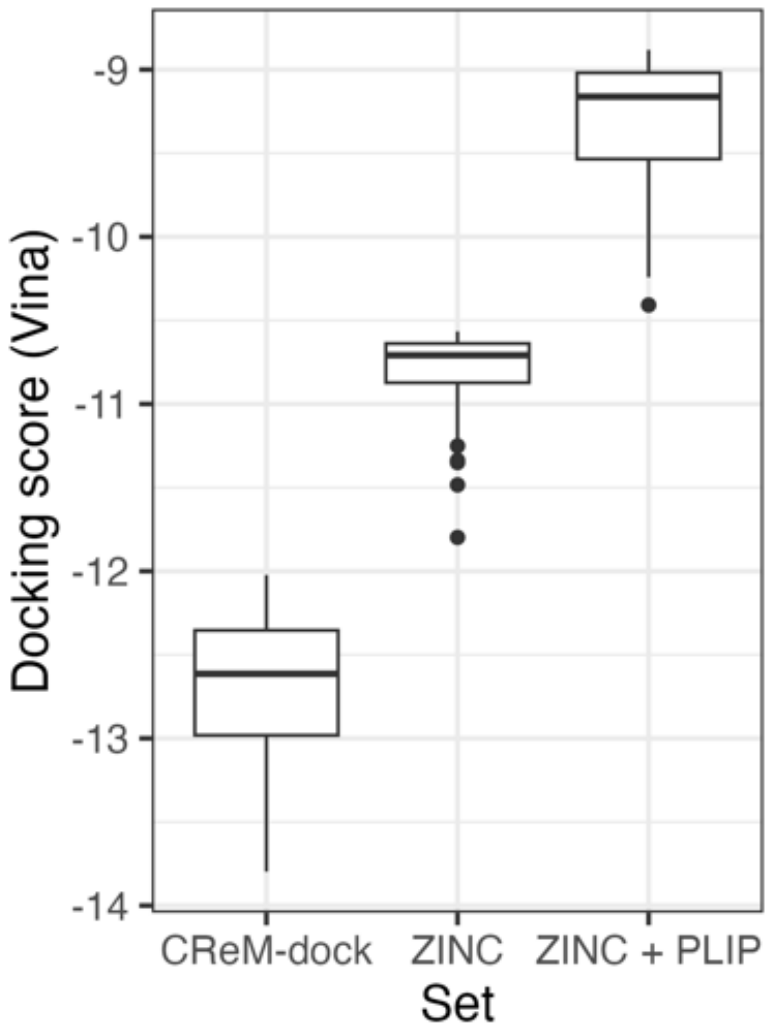
6GUH -14.3 / 3.06

docking score / SA score

CReM-dock: selectivity of designed CDK2 inhibitors

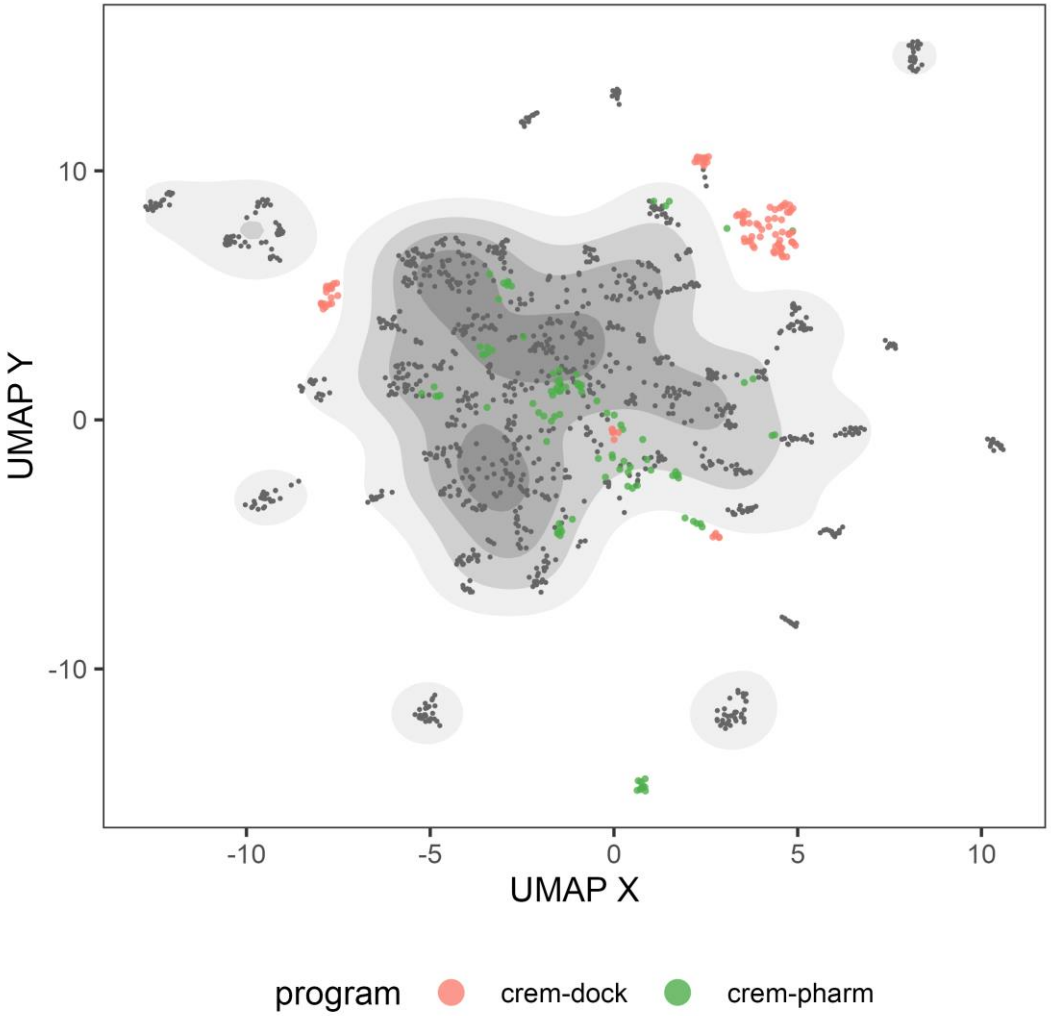


CReM-dock vs virtual screening



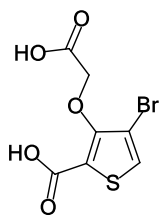
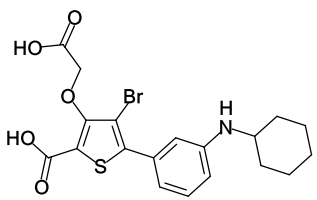
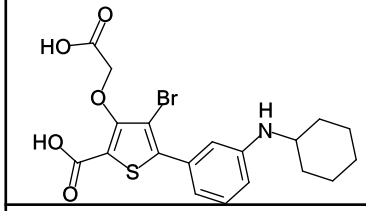
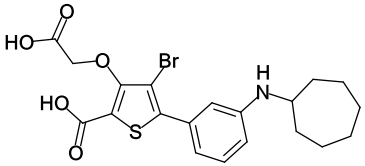
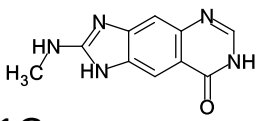
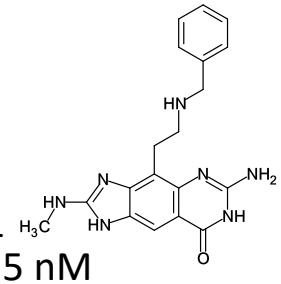
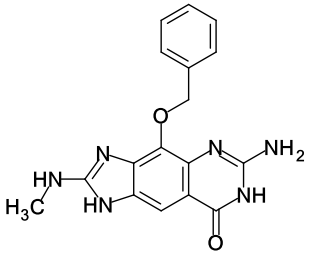
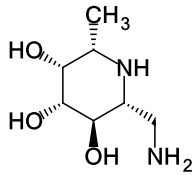
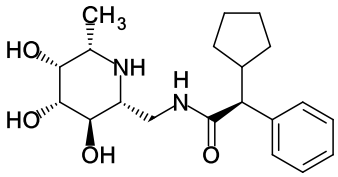
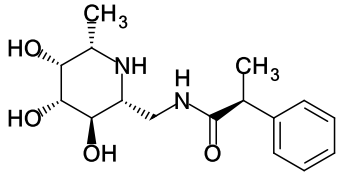
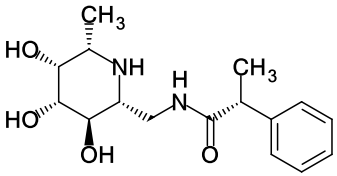
Chemical space overlap CReM-dock and CReM-pharm

CDK2 (2BTR)



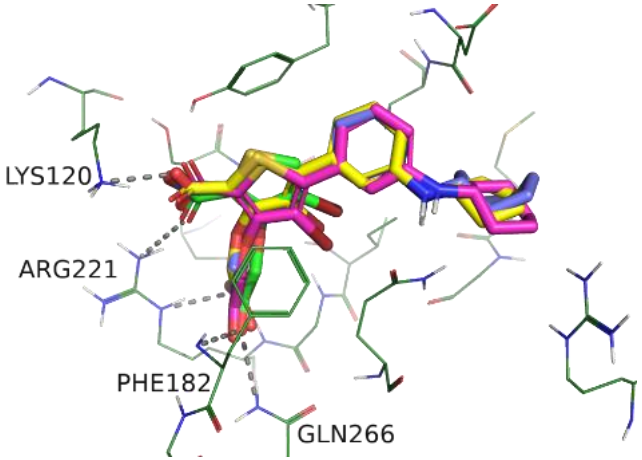
- CReM-dock and CReM-pharm structures generated for the same protein structure do not overlap much. Therefore, it can be suggested to use both approaches to get a greater number of diverse solutions

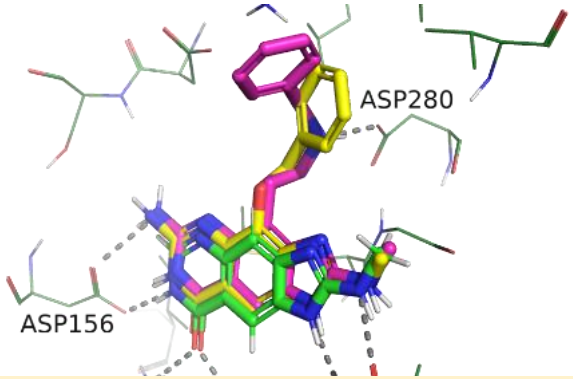
CReM-dock: fragment expansion

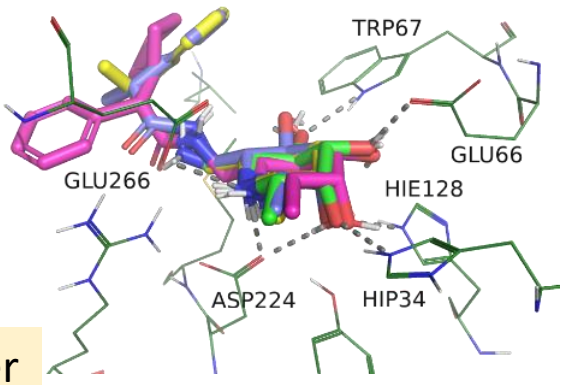
Starting ligand	Target ligand	Similarity of starting and target	Generated molecules most similar to the target one	Similarity to the target ligand	RMSD to the starting ligand, Å
 2HB1 K_i = 160 µM	 2QBS K_i = 210 nM	0.36		1	1.25
				1	1.52
 3S1G K_i = 6500 nM	 3GC4 K_i = 25 nM	0.32		0.63	0.06
 2ZWZ K_i = 16.3 nM	 2ZX9 K_i = 0.054 nM	0.32		0.69	0.86
				0.69	1.03

CReM-dock: fragment expansion

starting ligand
target ligand

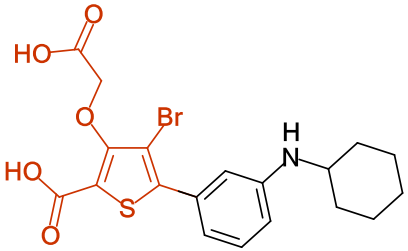


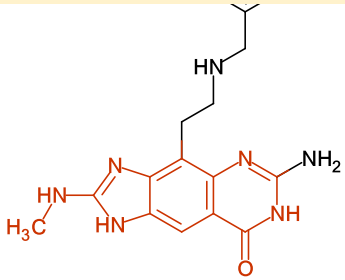


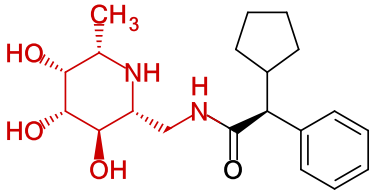


Fragments may grow in a proper direction which was previously explored as active

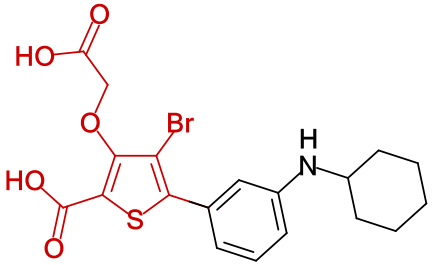
target ligand

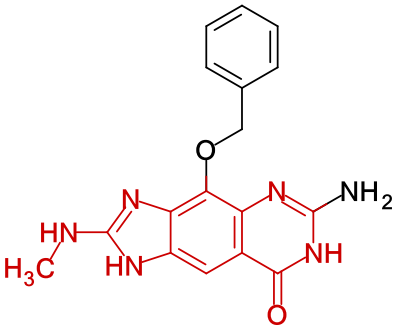


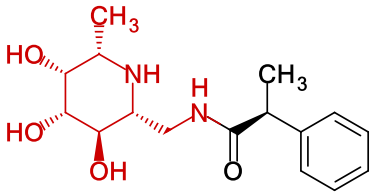




most similar ligand

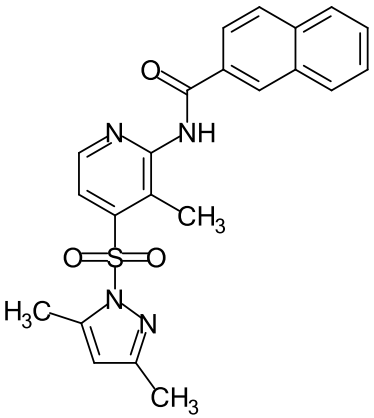




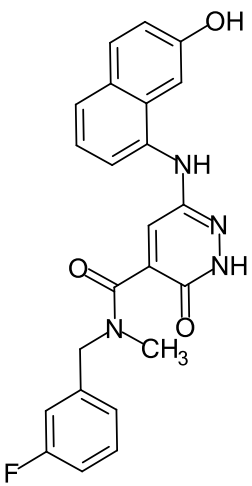


CReM-opt: local chemical space exploration

CReM-pharm



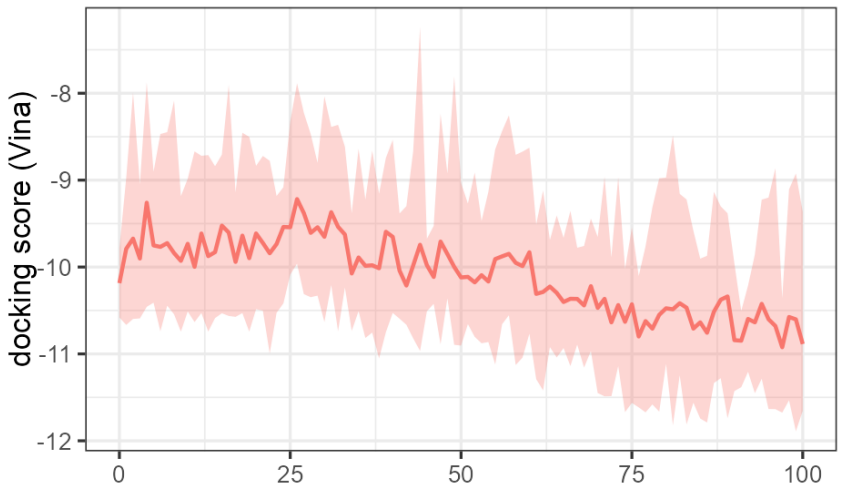
-10.4 / 2.56

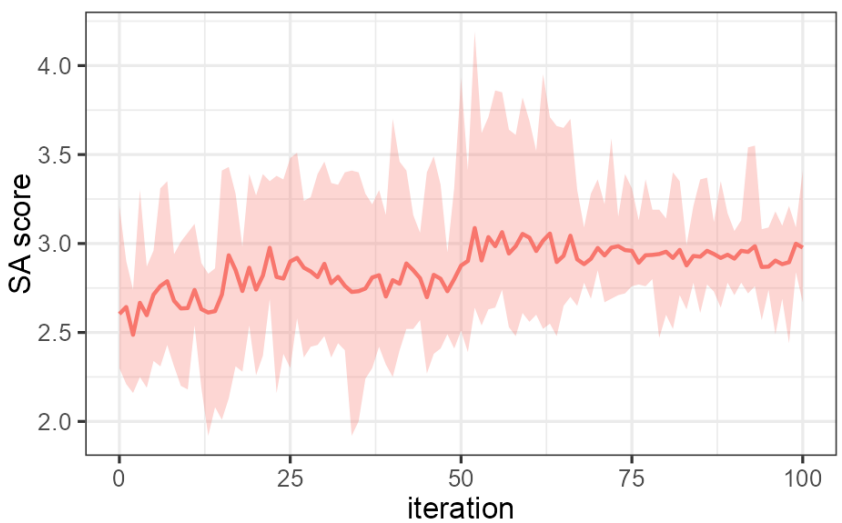


-10.2 / 2.46

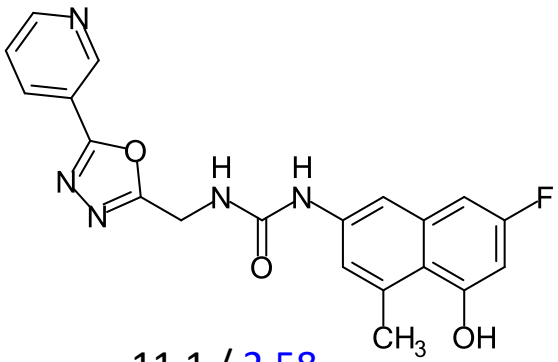
docking score / SA score

ChEMBL SA2.5 fragment database, radius 3

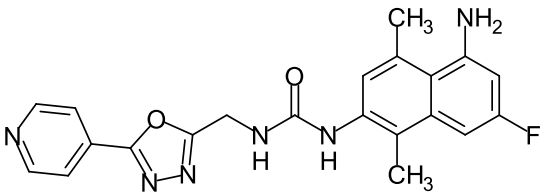




CReM-opt

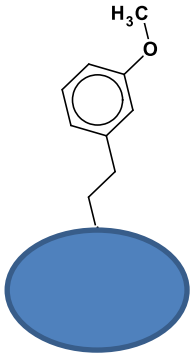


-11.1 / 2.58



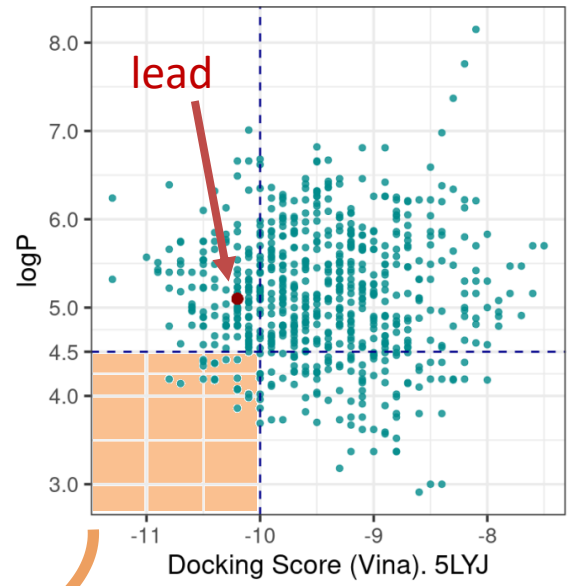
-11.0 / 2.69

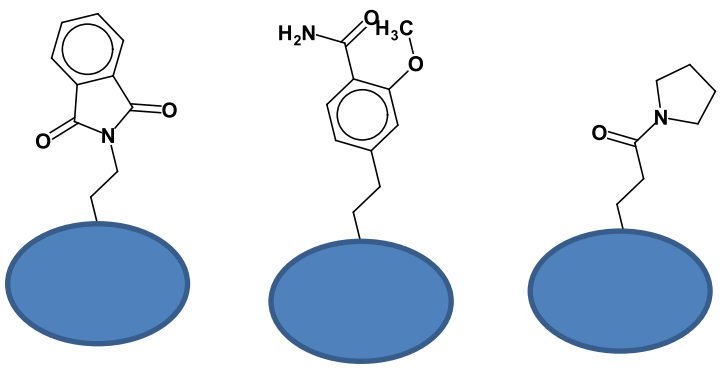
Optimization of tubulin inhibitors



docking score: -10.2

Cell line	IC_{50}, µM
A549	0.033
CCRF-CEM	0.058
CEM-DNR	0.097
HCT116	0.029
HCT116p53-	0.029
K562	0.029
K562-TAX	0.087
U2OS	0.038
BJ	>50

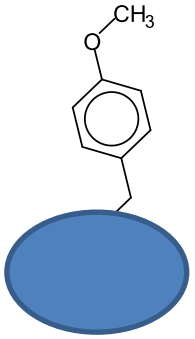




-10.7 -10.4 -10.0

docking score

Cell line	IC_{50}, µM
A549	8.84
CCRF-CEM	6.46
CEM-DNR	-
HCT116	9.18
HCT116p53-	9.29
K562	2.65
K562-TAX	-
U2OS	6.44
BJ	> 50



Cell line	IC_{50}, µM
A549	0.034
CCRF-CEM	0.018
CEM-DNR	0.029
HCT116	0.017
HCT116p53-	0.021
K562	0.013
K562-TAX	0.030
U2OS	0.018
BJ	>50

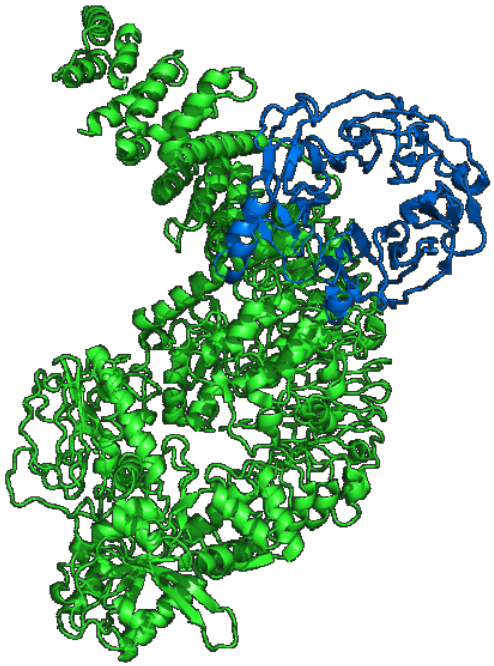
* in collaboration with Miroslav Soural group (UPOL)

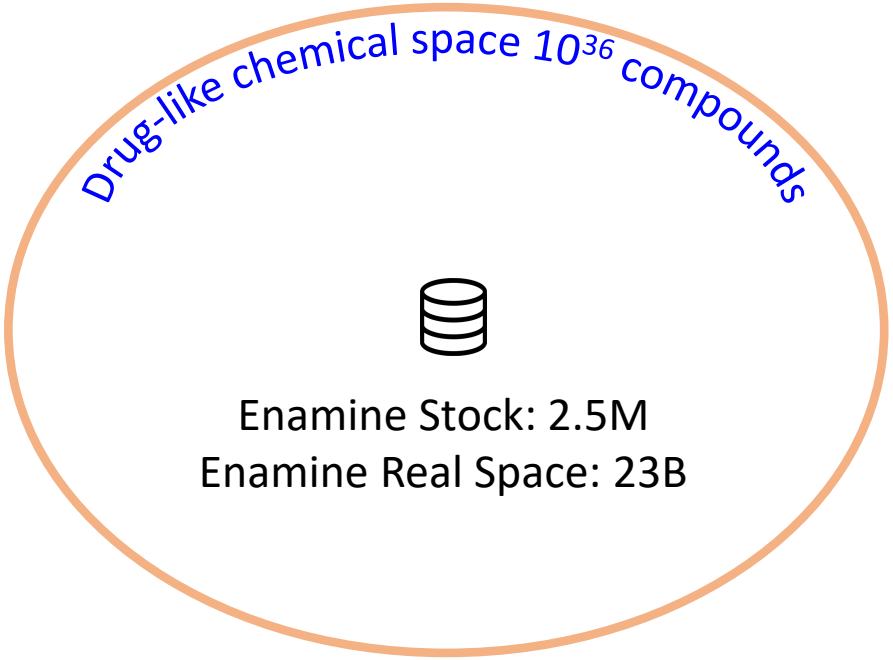
Searching for hits in ultra-large libraries guided by de novo design

CACHE challenge #1: LRRK2 and WDR domain

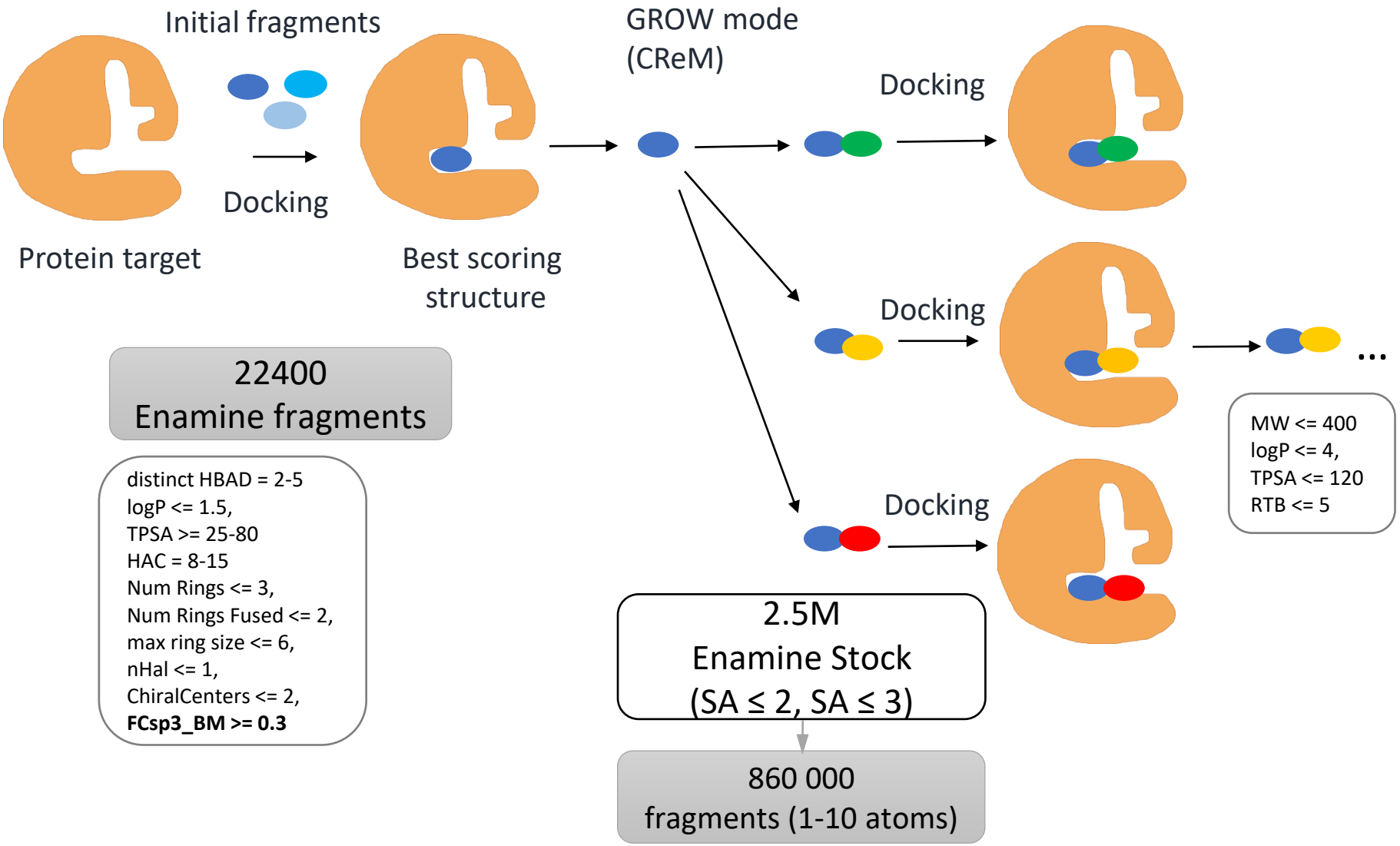
No X-ray of protein-ligand complexes:
- unknown binding site
- unknown conformation of a protein in a bound state

No known active molecules:
- large chemical space to explore

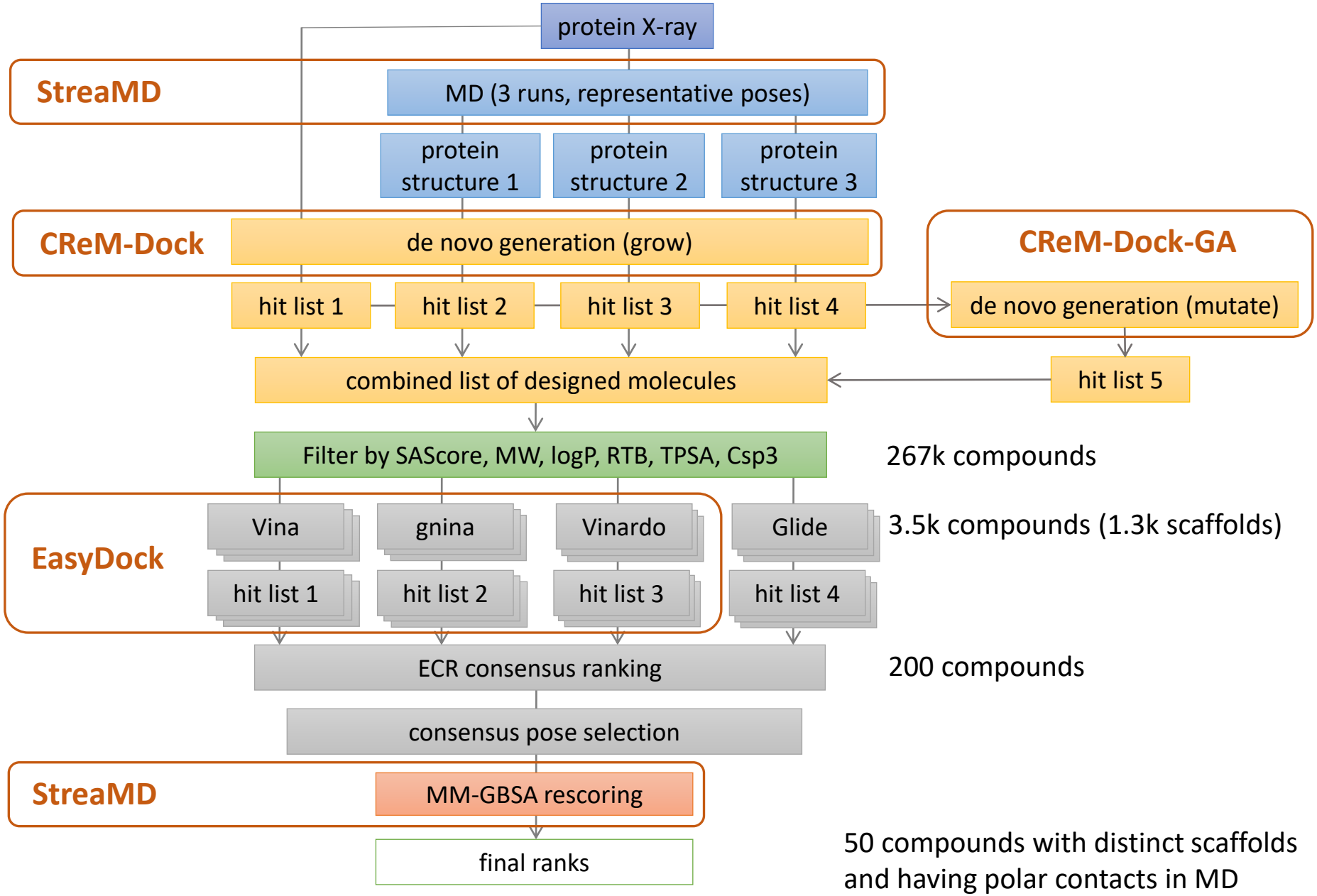




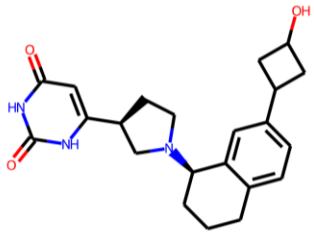
Round 1: strategy 1 (de novo design be CReM-dock)



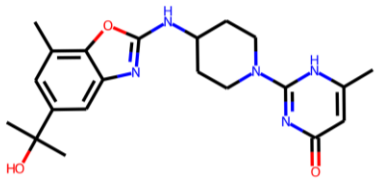
Round 1: strategy 1 (de novo design pipeline)



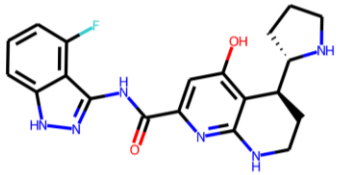
Round 1: strategy 1 (de novo design pipeline)



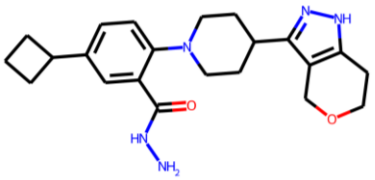
CREM1777121



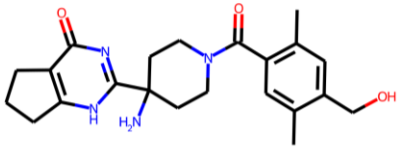
CREM0329741



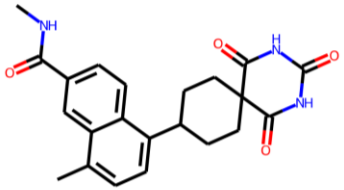
CREM1661038



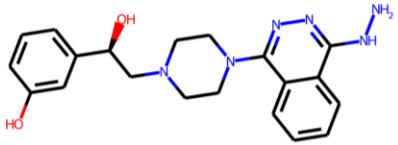
CREM1506273



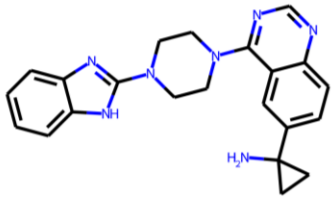
CREM0340409



CREM1089720

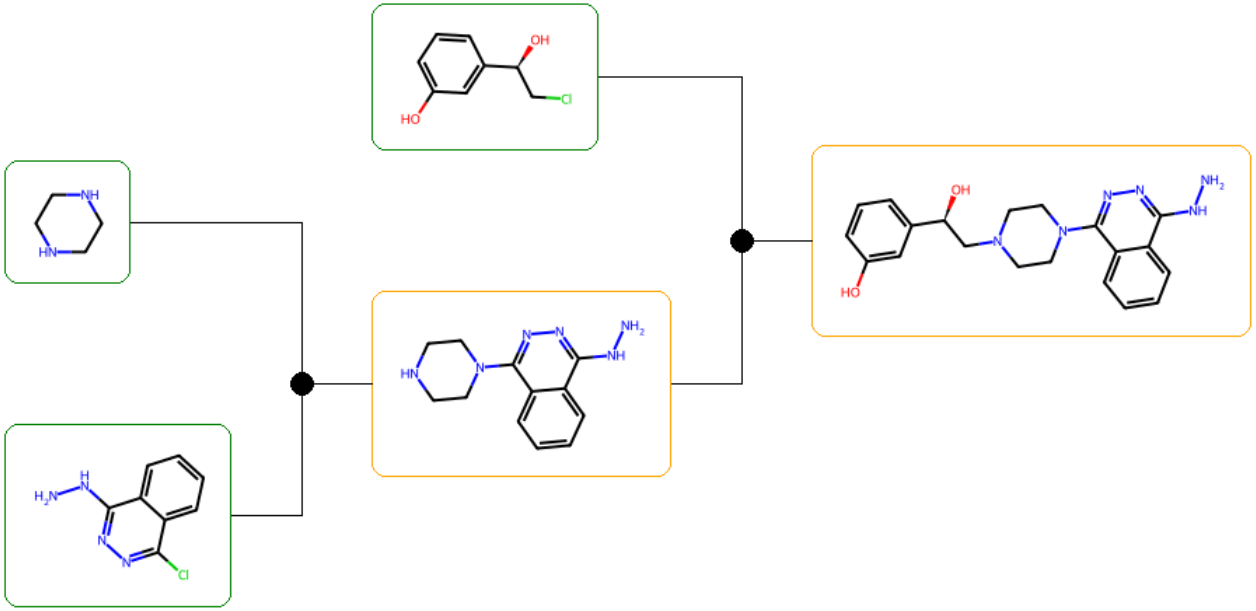


CREM1507777

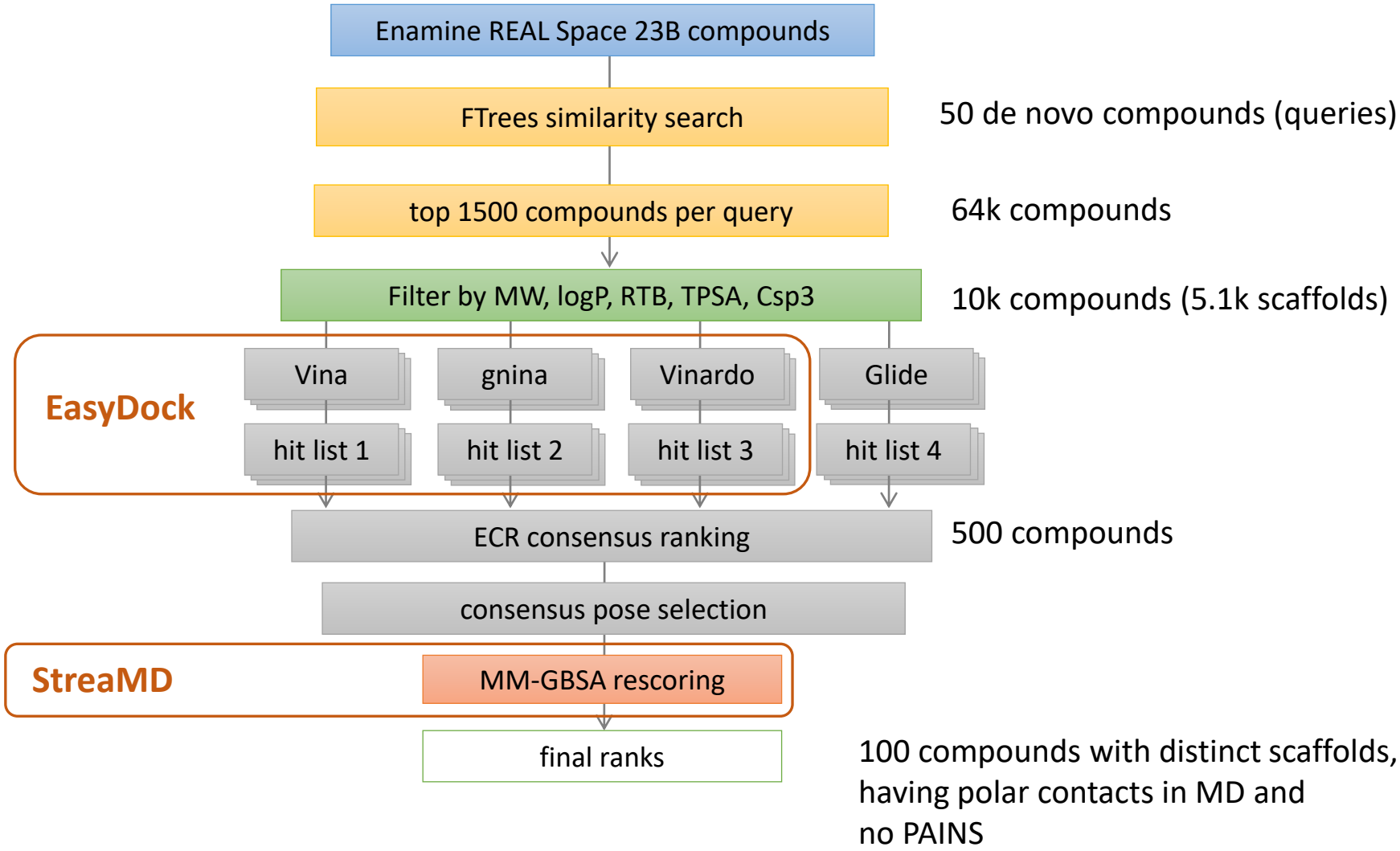


CREM1468894

- 50 de novo compounds
- SA score < 3
- 11 reconstructed retrosynthetic pathways with AiZynthFinder (2-5 steps)

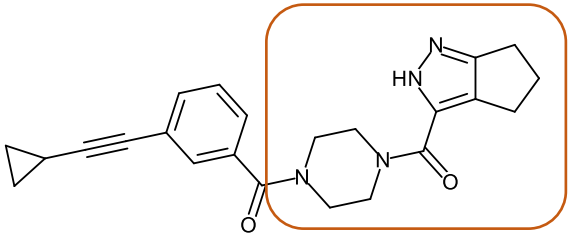


Round 1: strategy 1 (similarity search)

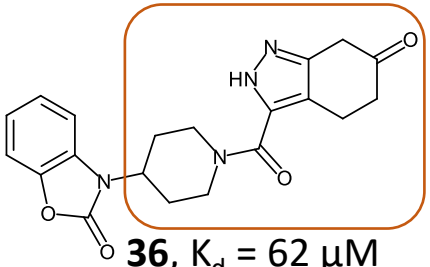


Rounds 1 & 2: results

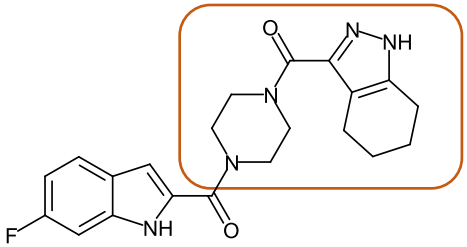
- 50 de novo + 100 similar compounds
- 91 compounds were selected (within the budget 9000$)
- 82 compounds were synthesized
- 8 compounds demonstrated activity (K_d = 25-117 µM by SPR)



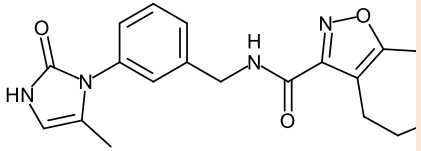
1, K_d = 61 µM



36, K_d = 62 µM

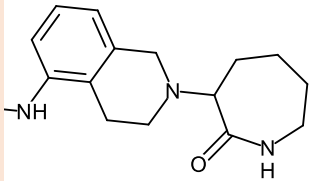


HO-15, K_d = 71 µM

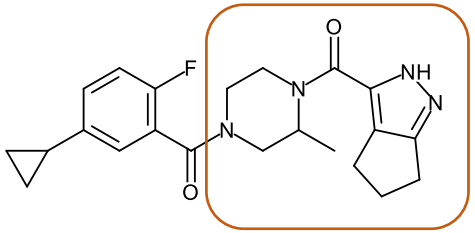


59, K_d = 32 µM

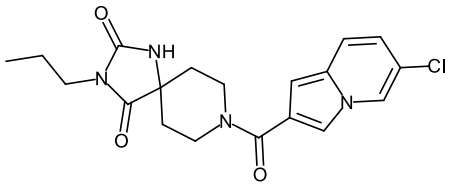
- no human decision and compound selection across the whole pipeline
- 1.27 million docking events and 700 short MD simulations were made



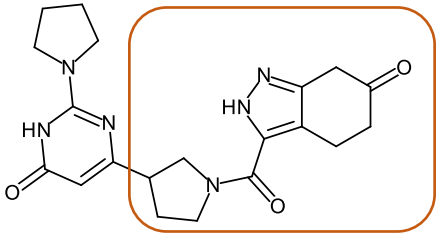
K_d = 56 µM



69, K_d = 117 µM



73, K_d = 31 µM



76, K_d = 74 µM

Summary

CReM is highly flexible and can be combined with relevant modeling tools to address different tasks

- scaffold decoration
- fragment expansion
- hit/lead generation/optimization
- de novo design

Synthetic accessibility of generated compounds depends on CReM settings rather than on a computational approach or model/protein complexity

```
$ pypistats overall crem
```

category	percent	downloads
with_mirrors	100.00%	31,076
without_mirrors	66.99%	20,819
Total		31,076

```
Date range: 2025-03-14 - 2025-09-10
```

The work was supported by the Ministry of Education, Youth and Sports of the Czech Republic through INTER_EXCELLENCE II grant LUAUS23262, the e-INFRA CZ (ID:90254), projects ELIXIR-CZ (LM2023055) and CZ-OPENSCREEN (LM2023052)

Links

De novo design / optimization

CReM - Python module for structure generation	https://github.com/DrrDom/crem
CReM-Dock – de novo generation guided by docking	https://github.com/ci-lab-cz/crem-dock
CReM-opt – structure optimization guided by docking	(not publicly available yet)
CReM-pharm – de novo generation guided by 3D pharmacophores	https://github.com/ci-lab-cz/crem-pharm

Automated pipelines

EasyDock – fully automated distributed molecular docking pipeline	https://github.com/ci-lab-cz/easydock
StreaMD – automated pipeline for high-throughput MD simulations	https://github.com/ci-lab-cz/streamd

Auxiliary RDKit repositories

rdkit-scripts - various RDKit scripts	https://github.com/DrrDom/rdkit-scripts

References

(1) Polishchuk, P. CReM: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics* **2020**, 12 (1), 28.

(2) Polishchuk, P. Control of Synthetic Feasibility of Compounds Generated with CReM. *Journal of Chemical Information and Modeling* **2020**, 60 (12), 6074-6080.

(3) Minibaeva, G.; Polishchuk, P. CReM-dock: de novo design of synthetically feasible compounds guided by molecular docking. *ChemRxiv* ***2024.***

(4) Minibaeva, G.; Ivanova, A.; Polishchuk, P. EasyDock: customizable and scalable docking tool. *Journal of Cheminformatics* **2023**, 15 (1), 102.

(5) Ivanova, A.; Mokshyna, O.; Polishchuk, P. StreaMD: the toolkit for high-throughput molecular dynamics simulations. *Journal of Cheminformatics* **2024**, 16 (1), 123.

The group of chemoinformatics and drug design

Aleksandra Ivanova

Guzel Minibaeva

Alina Kutlushina

Dinesh Kumar Sriramulu Ph.D.