INSTITUTE OF MOLECULAR AND TRANSLATIONAL MEDICINE

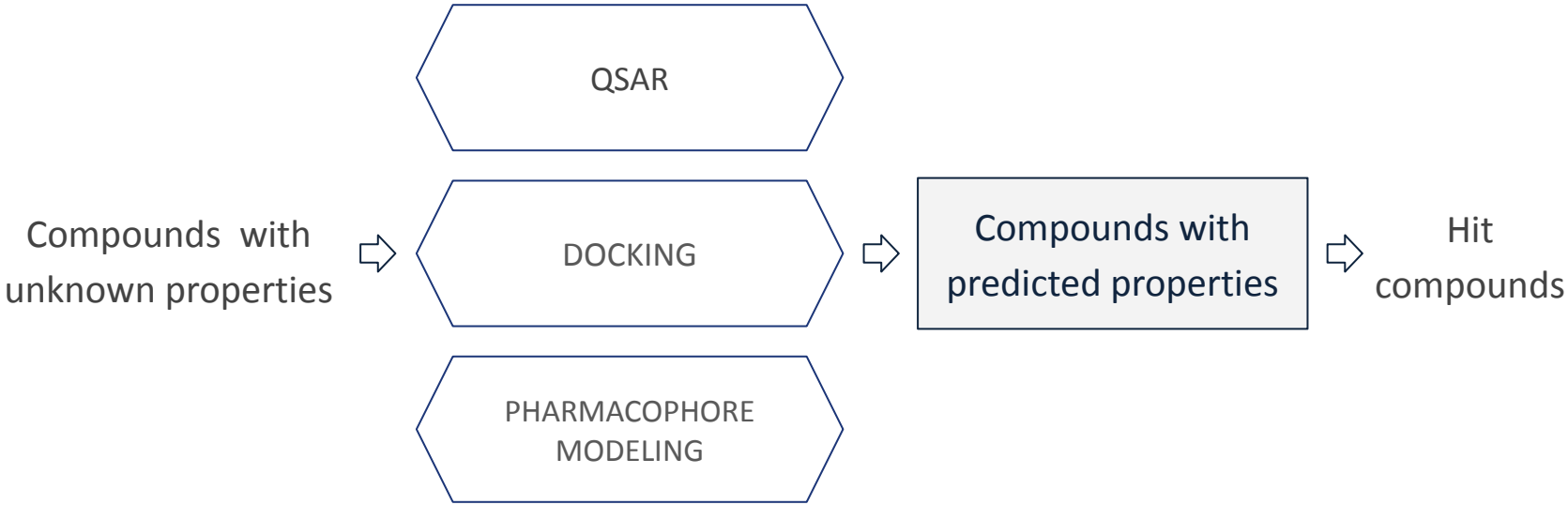
XXVIII Symposium on Bioinformatics and Computer-Aided Drug Discovery

De novo generation of synthetically feasible molecules

Minibaeva Guzel, Aleksandra Ivanová, Pavel Polishchuk

Institute of Molecular and Translational Medicine, Palacký University, Czech Republic

Virtual screening



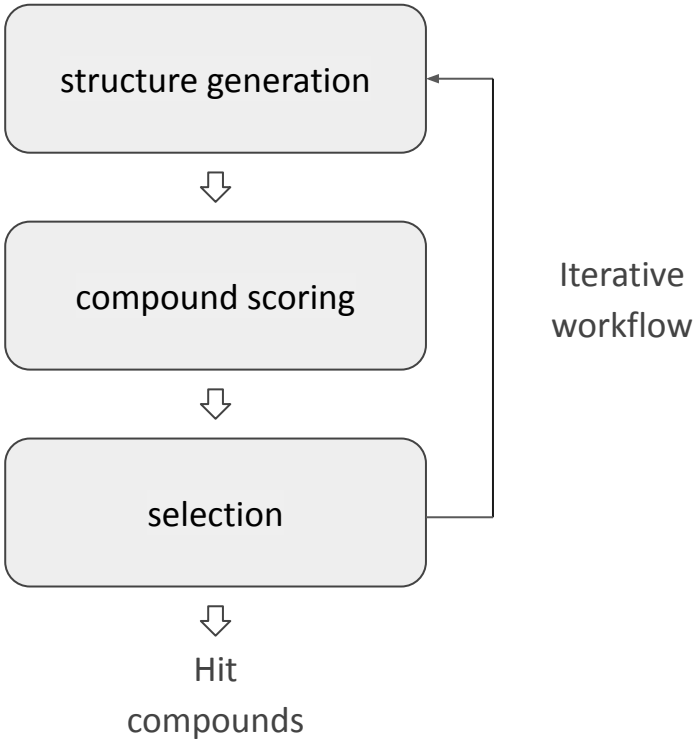
Polishchuk P., Madzhidov T., Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des 2013

De novo design

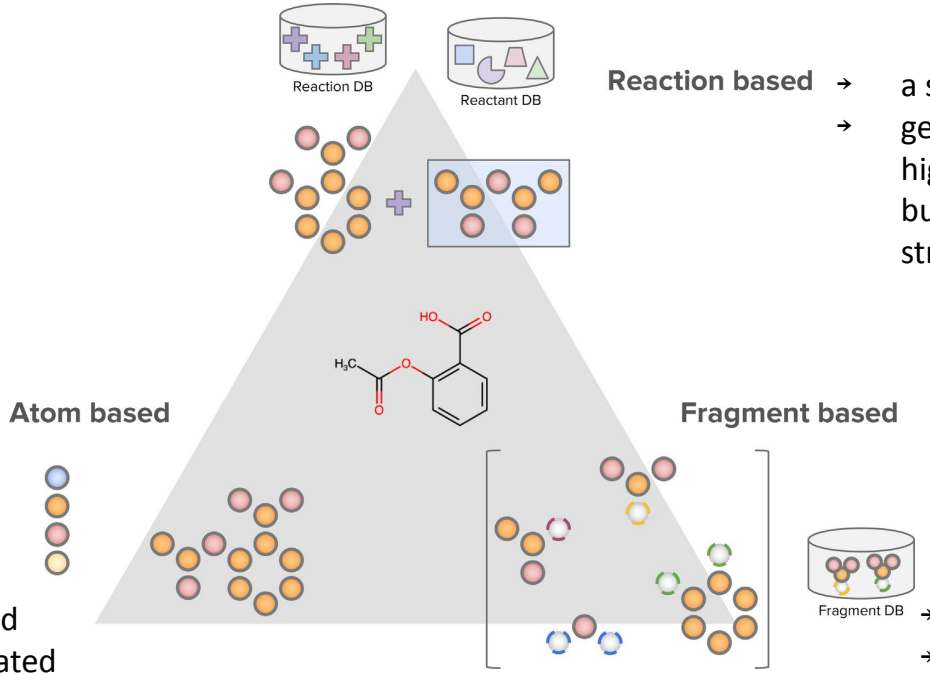
how to create/assembly new structures

how to estimate/predict a property of a compound

how to find compound with optimal properties



Atom/reaction/fragment-based approaches



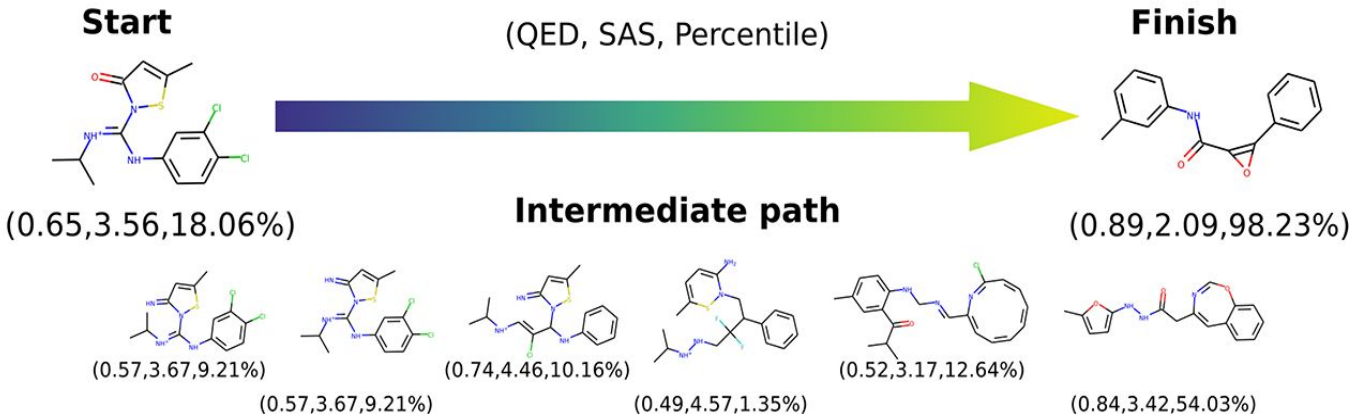
- a seed fragment is required
- generated structures with high chemical accessibility, but the variety of generated structures is very limited [4]

- a seed atom is required
- high number of generated structures with questionable chemical accessibility [3]

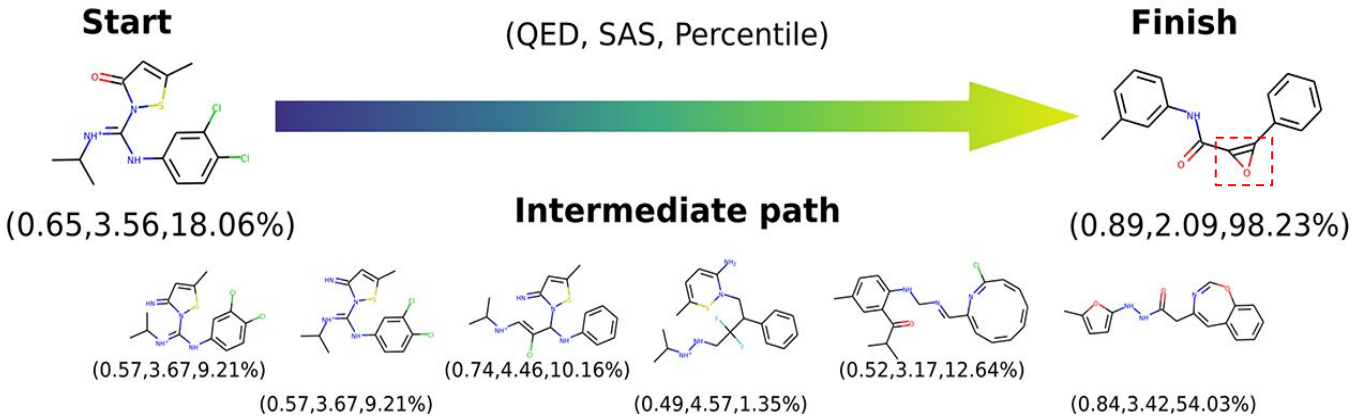
- a seed fragment is required
- the set of fragments directly determines the novelty and diversity of the generated compounds

[3] Hoksza, D. Molpher: A software framework for systematic chemical space exploration. Journal of Cheminformatics, 2014
[4] Batiste, L. Chemical Space Expansion of Bromodomain Ligands Guided by in Silico Virtual Couplings. ACS Central Science, 2018
[5] Meyers J. De novo molecular design and generative models. Drug Discovery Today, 2021

Deep learning models for structure generation



Gómez-Bombarelli, R. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science 2018



! structure filtering by chemical validity is necessary

A lot of methods to create new molecules, but we need to score a synthesizability

Gómez-Bombarelli, R. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science 2018

AutoGrow4: PARP inhibitors de novo design

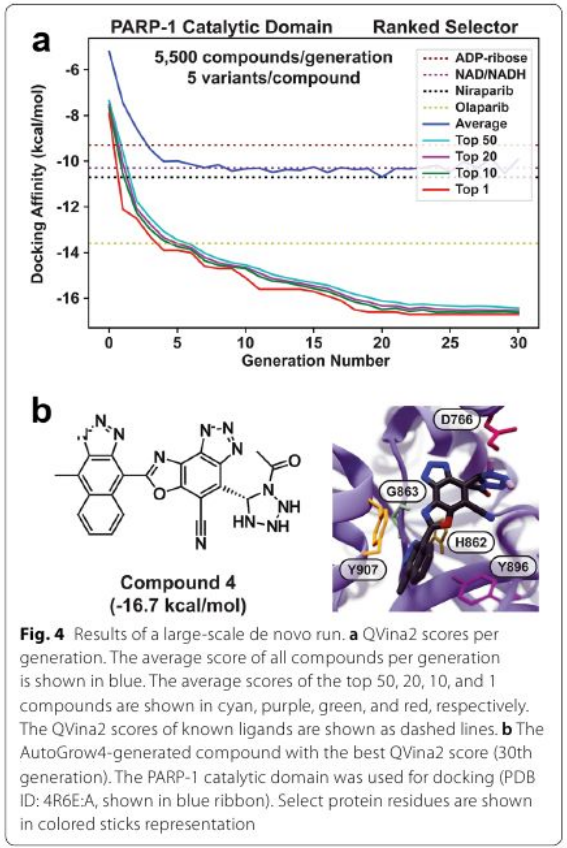
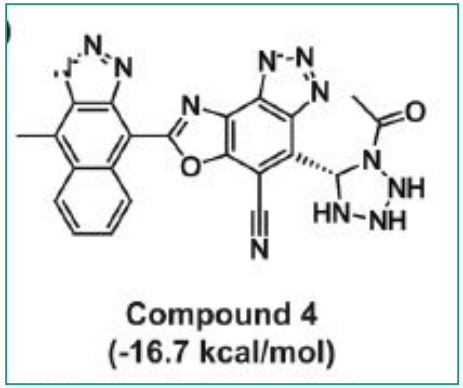
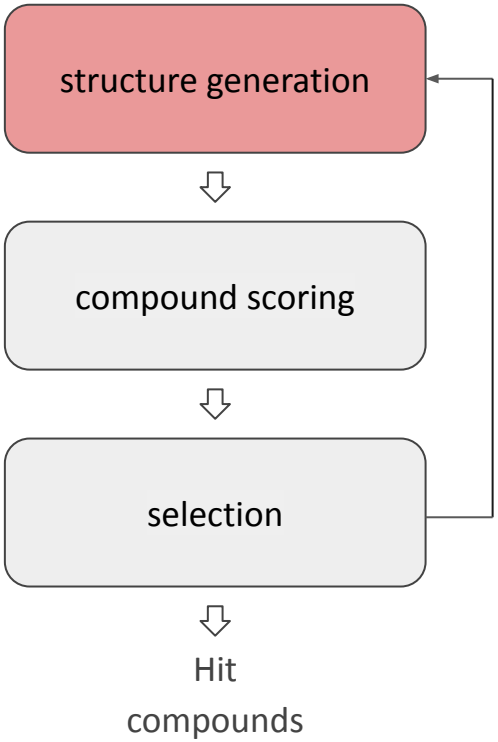


Fig. 4 Results of a large-scale de novo run. **a** QVina2 scores per generation. The average score of all compounds per generation is shown in blue. The average scores of the top 50, 20, 10, and 1 compounds are shown in cyan, purple, green, and red, respectively. The QVina2 scores of known ligands are shown as dashed lines. **b** The AutoGrow4-generated compound with the best QVina2 score (30th generation). The PARP-1 catalytic domain was used for docking (PDB ID: 4R6E:A, shown in blue ribbon). Select protein residues are shown in colored sticks representation



SA score: 4.61

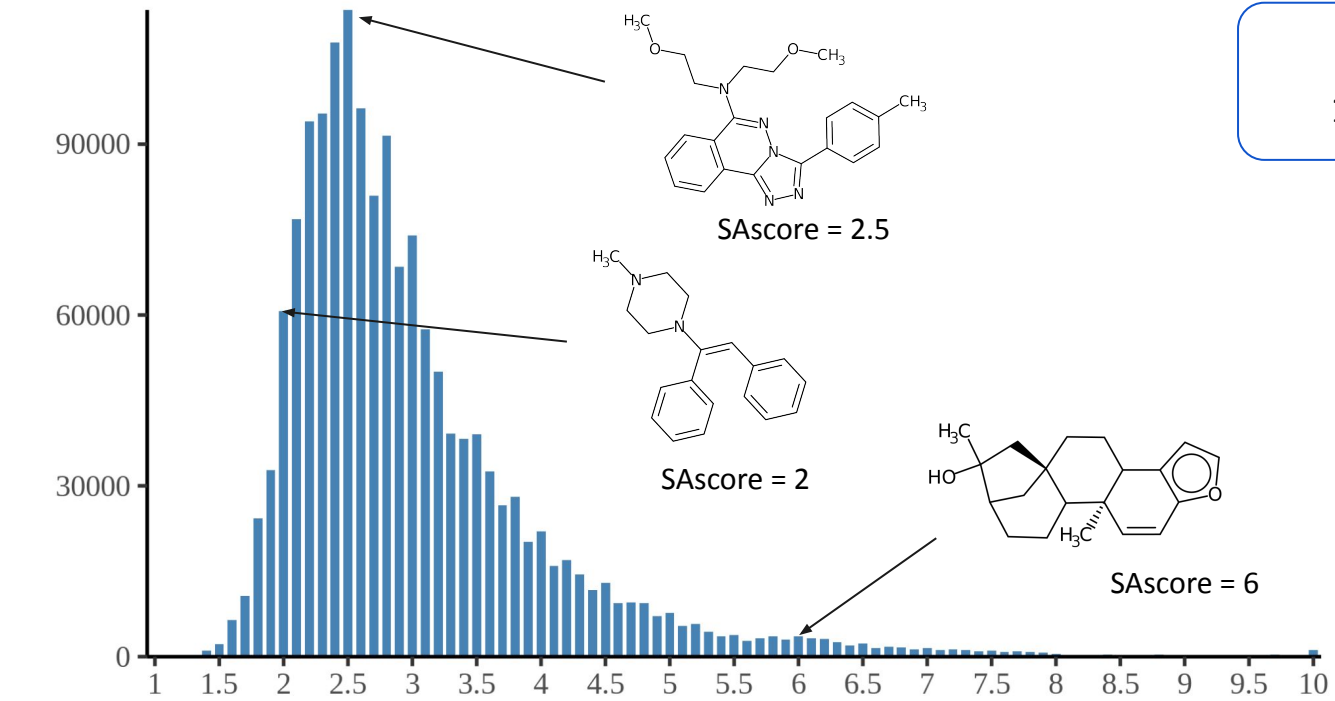


Spiegel, J. O.; Durrant, J. D., AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. Journal of Cheminformatics 2020

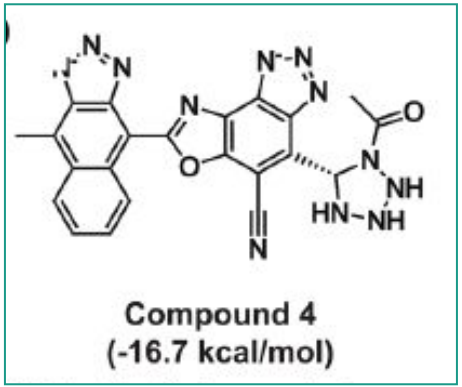
Aims

1. Develop de novo design approach which take into account synthetic accessibility of molecules (CReM)
2. Investigate the applicability of developed approach on a benchmarking study (CDK2)
3. Application to design inhibitors of SARS-CoV-2 main protease

Synthetic accessibility of molecules (SAscore)



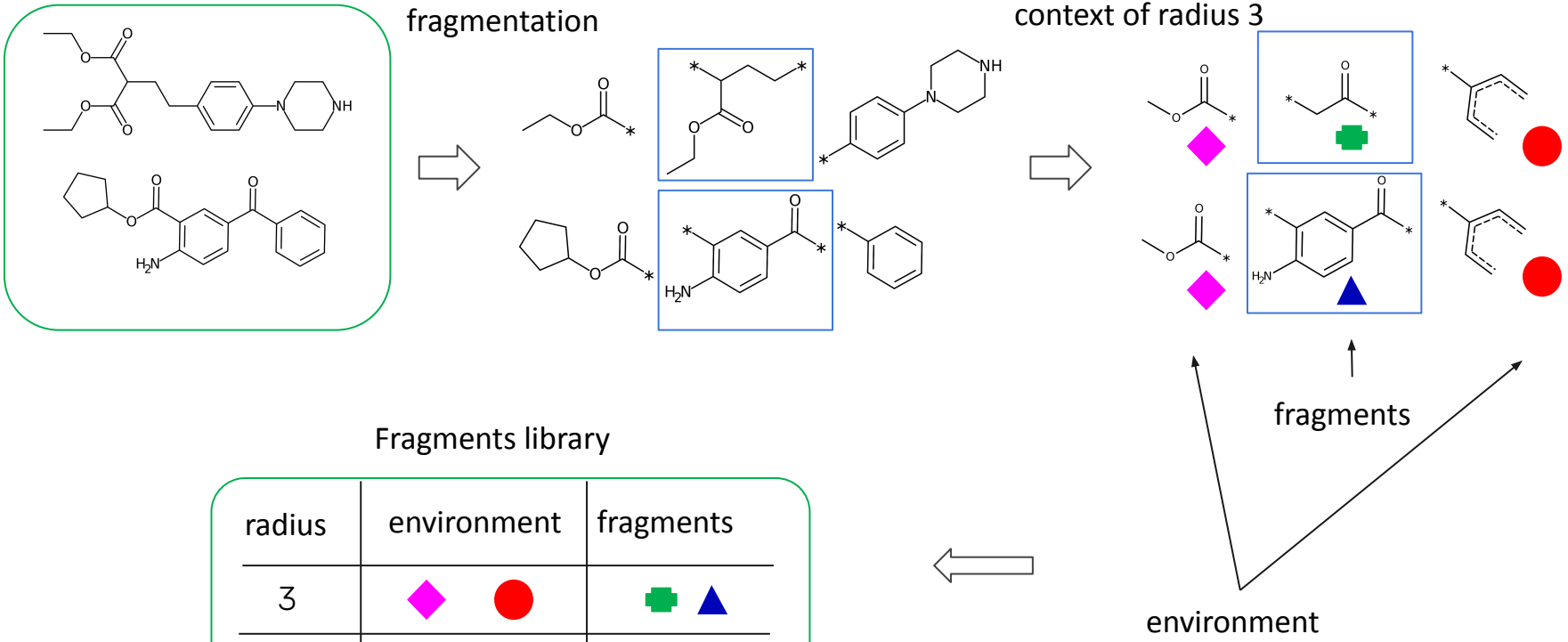
ChEMBL (version 22)
1 554 260 compounds



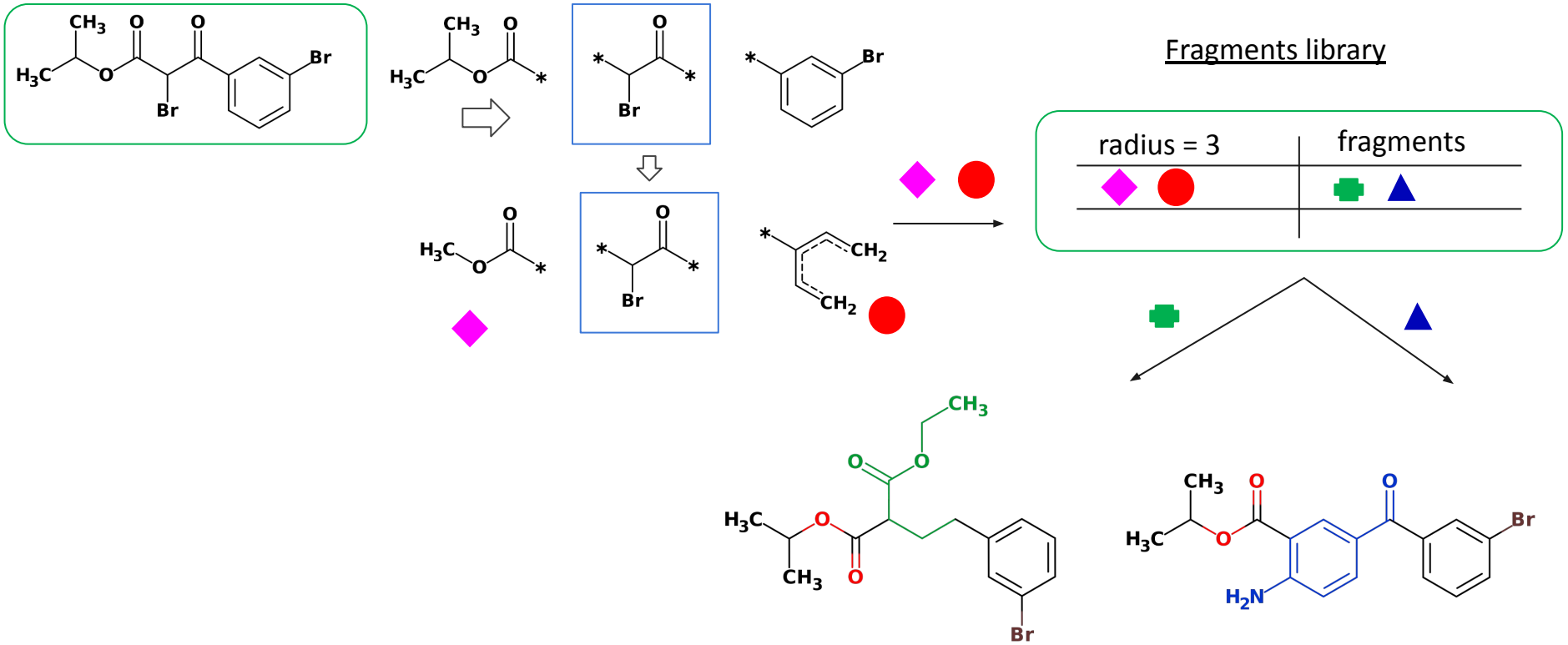
SA score: 4.61

Ertl, P. et al. Journal of Cheminformatics, 2009

Chemically reasonable mutations framework (CReM)



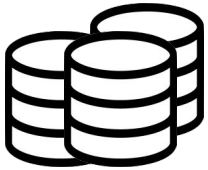
Polishchuk, P. Journal of Cheminformatics, 2020



Polishchuk, P. Journal of Cheminformatics, 2020

Control of synthetic feasibility within CReM

Content of fragmented library



all ChEMBL compounds (1 554 160)

compounds with SA score ≤ 2.5 (572 527)

compounds with SA score ≤ 2 (107 806)

Context radius

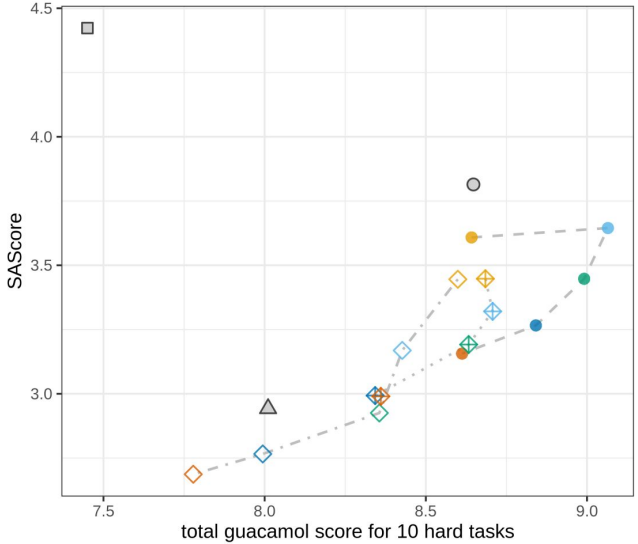
1 — less conservative replacements

2

3

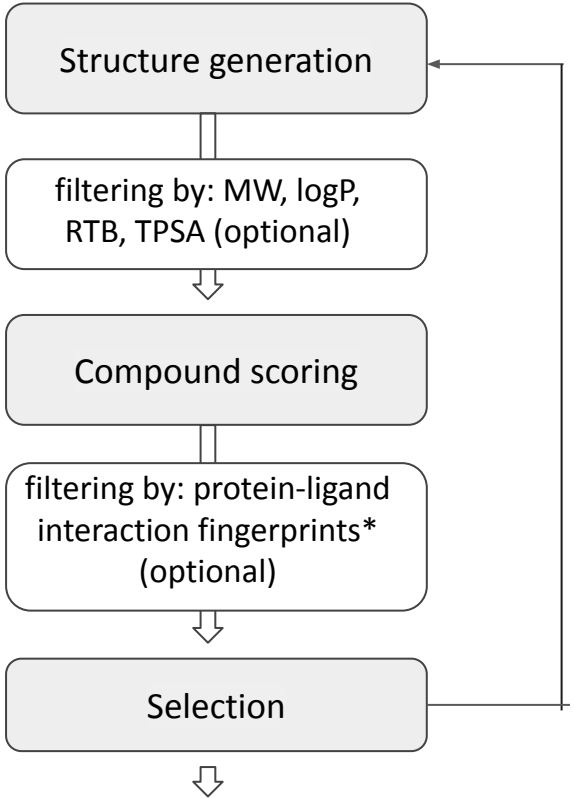
4 — more conservative replacements

5



Polishchuk, P. Control of Synthetic Feasibility of Compounds Generated with CReM. Journal of Chemical Information and Modeling, 2020

De novo design

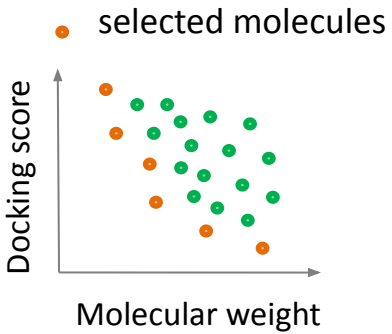


Hit compounds with high docking score, favorable physicochemical properties and specific protein-ligand contacts

CReM
n replacements = 2000
fragment size = 1-10

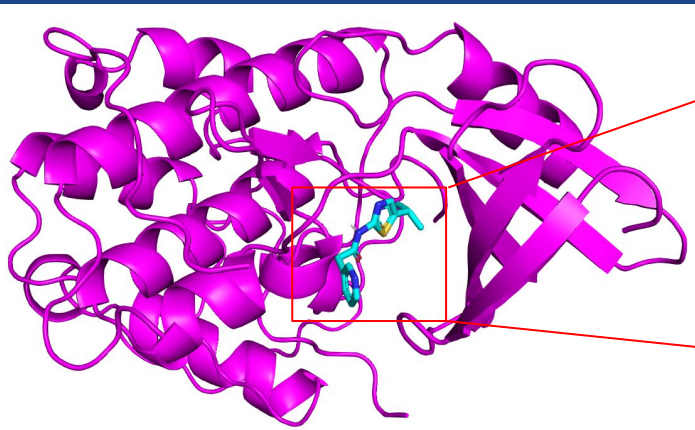
docking score (Vina)
docking score (Vina) + QED

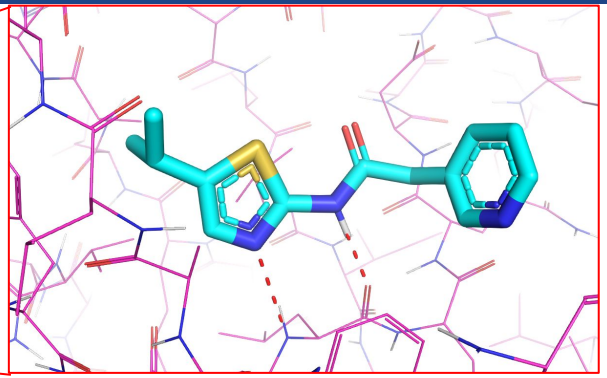
Selection:
- greedy
- clustering (k-means)
- Pareto (docking score vs. MW)



*github.com/chemosim-lab/ProLIF.git

De novo design using docking (CDK2 inhibitors)

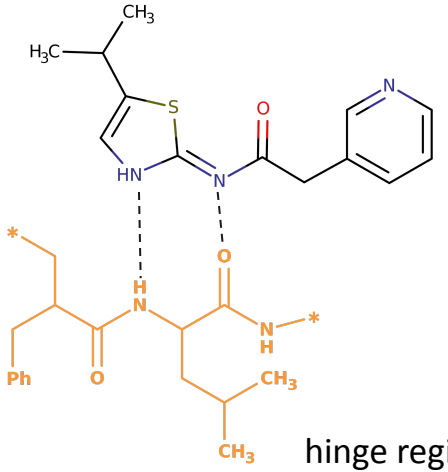




2BTR

IC_{50} = 95 nM

docking score = -7.86



hinge region

De novo design using docking (CDK2 inhibitors)

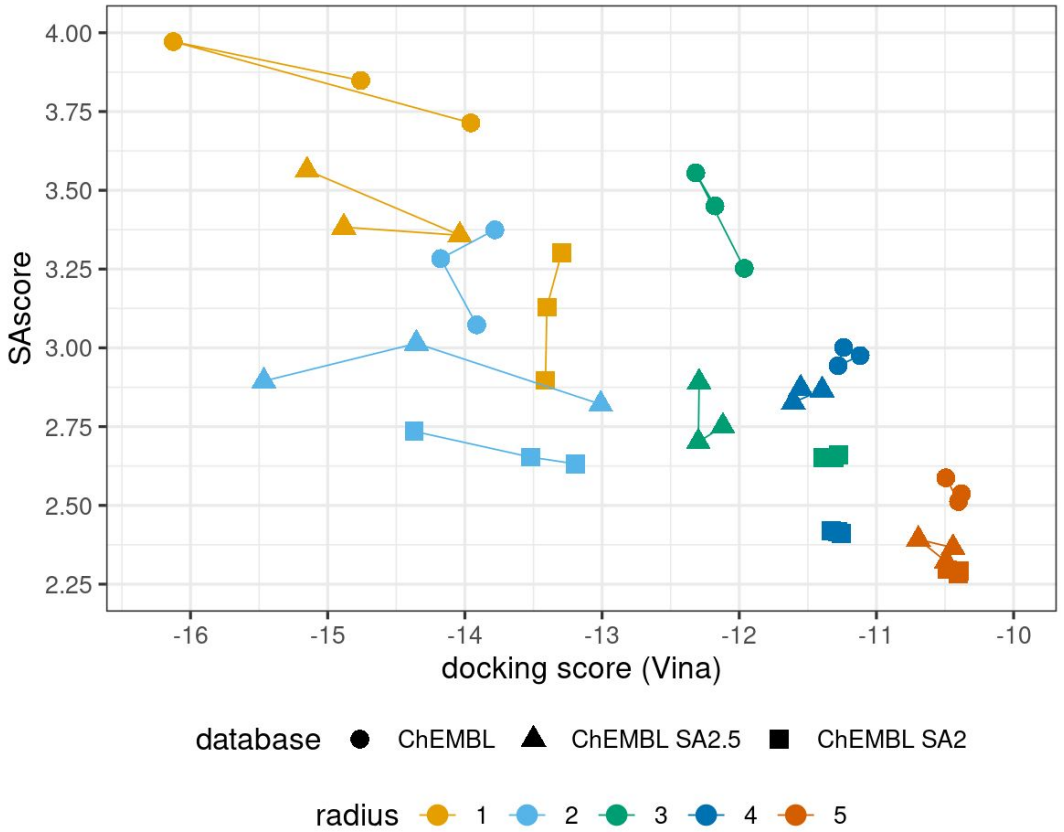
Constant conditions:

- MW ≤ 450
- RTB ≤ 5
- logP ≤ 4
- TPSA ≤ 120
- hinge region binding
- selection algorithm: clustering
  - nclusters = 25
  - nmols per cluster = 2

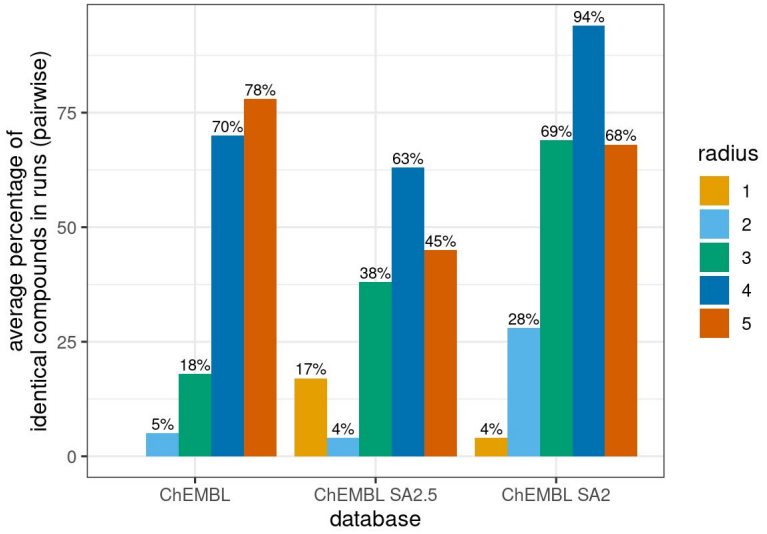
Variable conditions:

CReM fragment bases:
ChEMBL, ChEMBL SA2.5, ChEMBL SA2
radius: 1, 2, 3, 4, 5

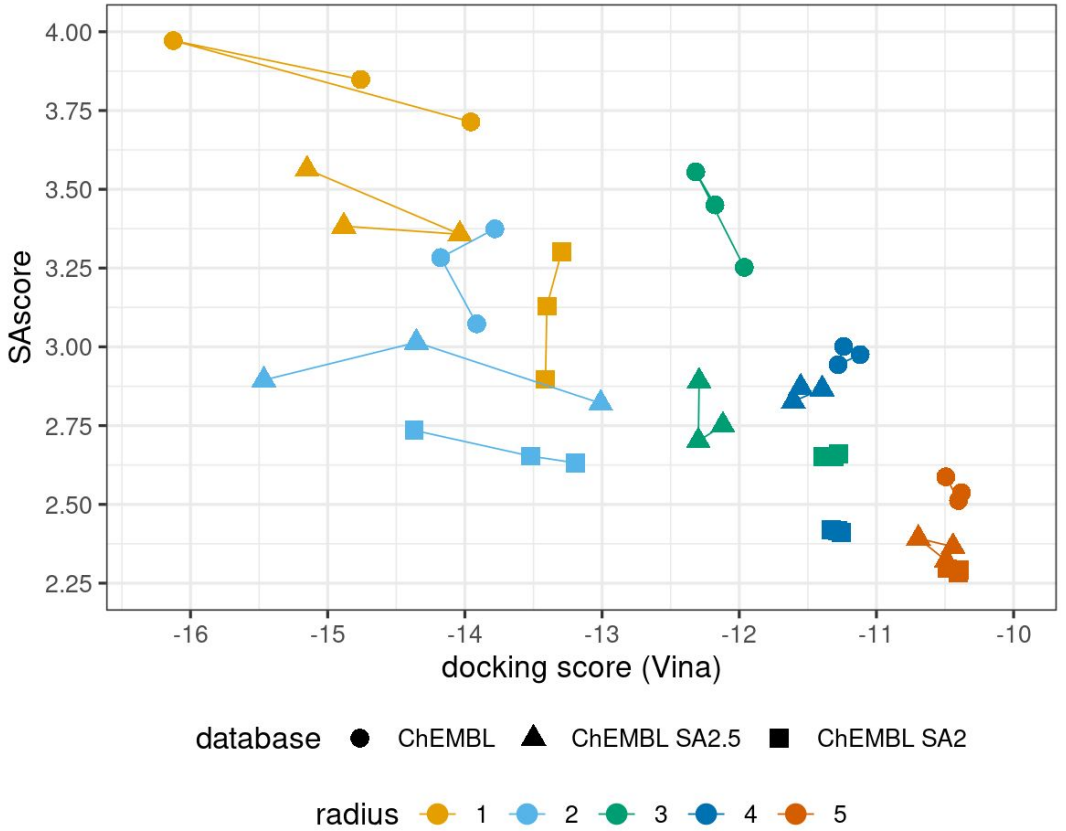
Average docking and SA scores for top 100 molecules from each run



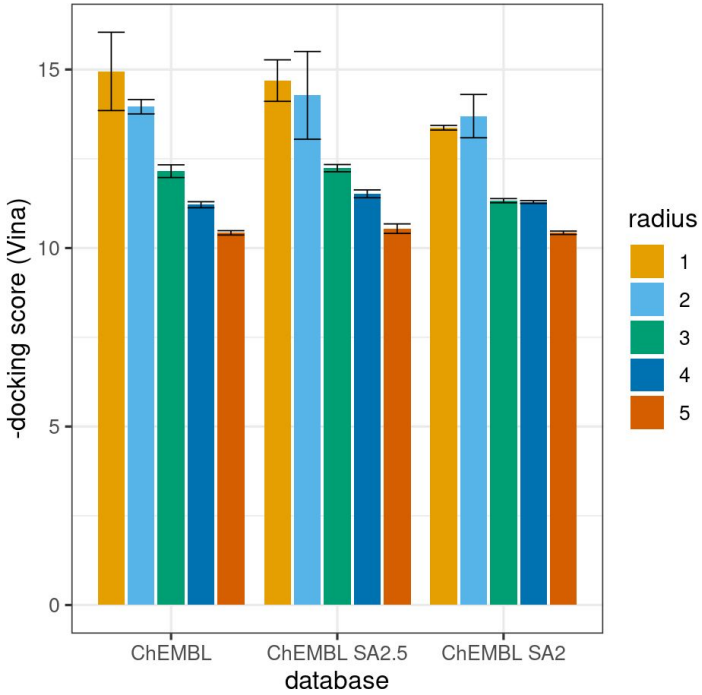
De novo design using docking (CDK2 inhibitors)



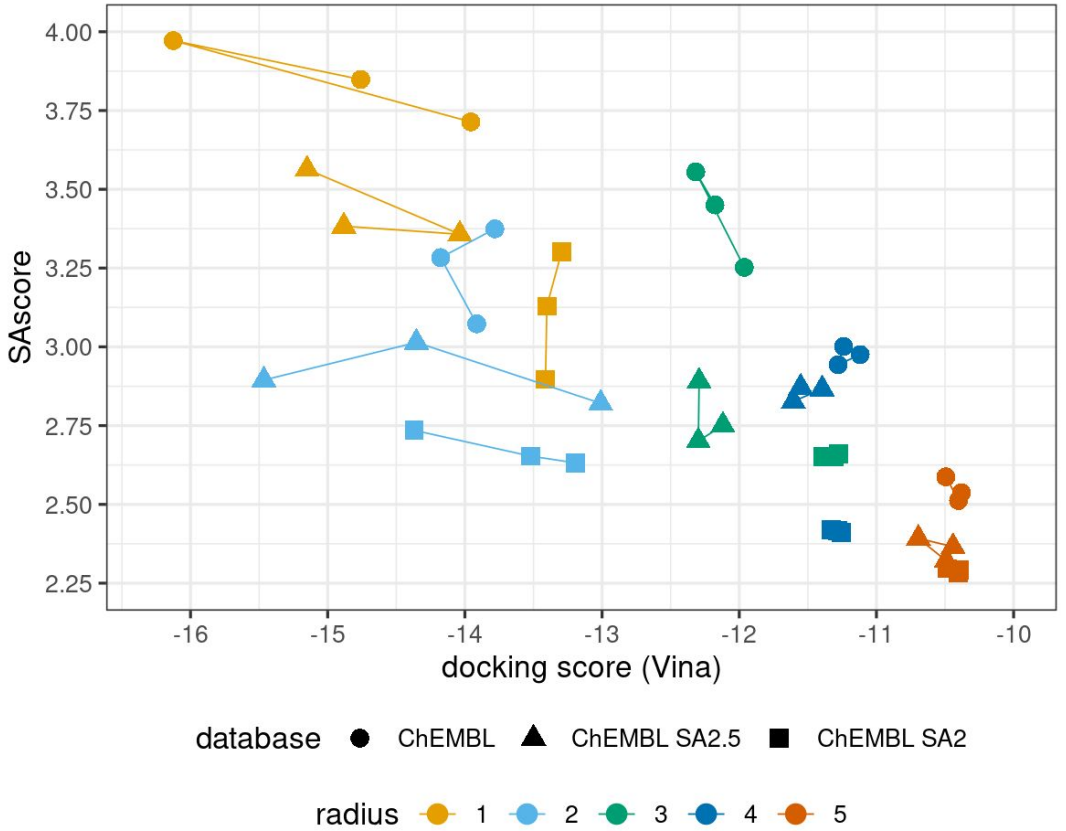
Average docking and SA scores for top 100 molecules from each run



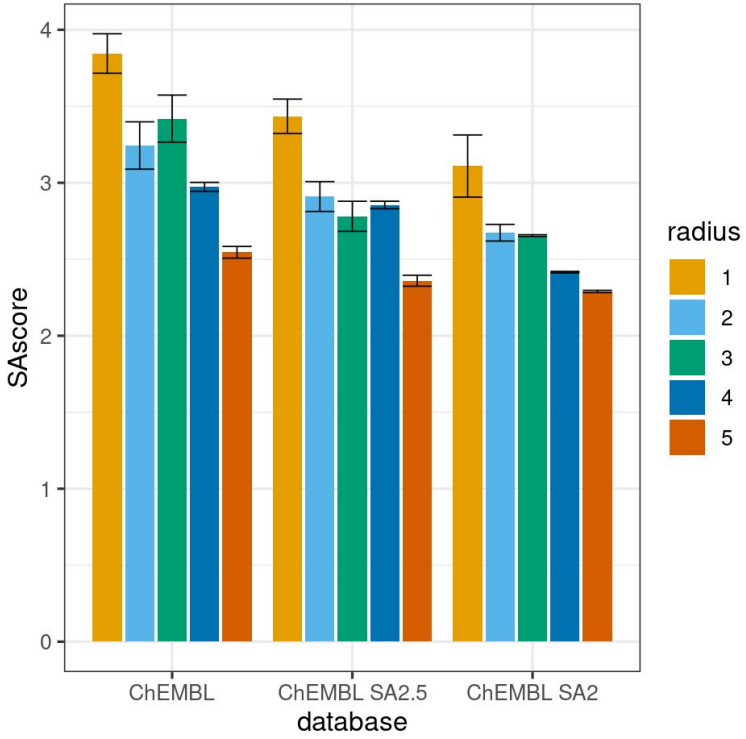
De novo design using docking (CDK2 inhibitors)



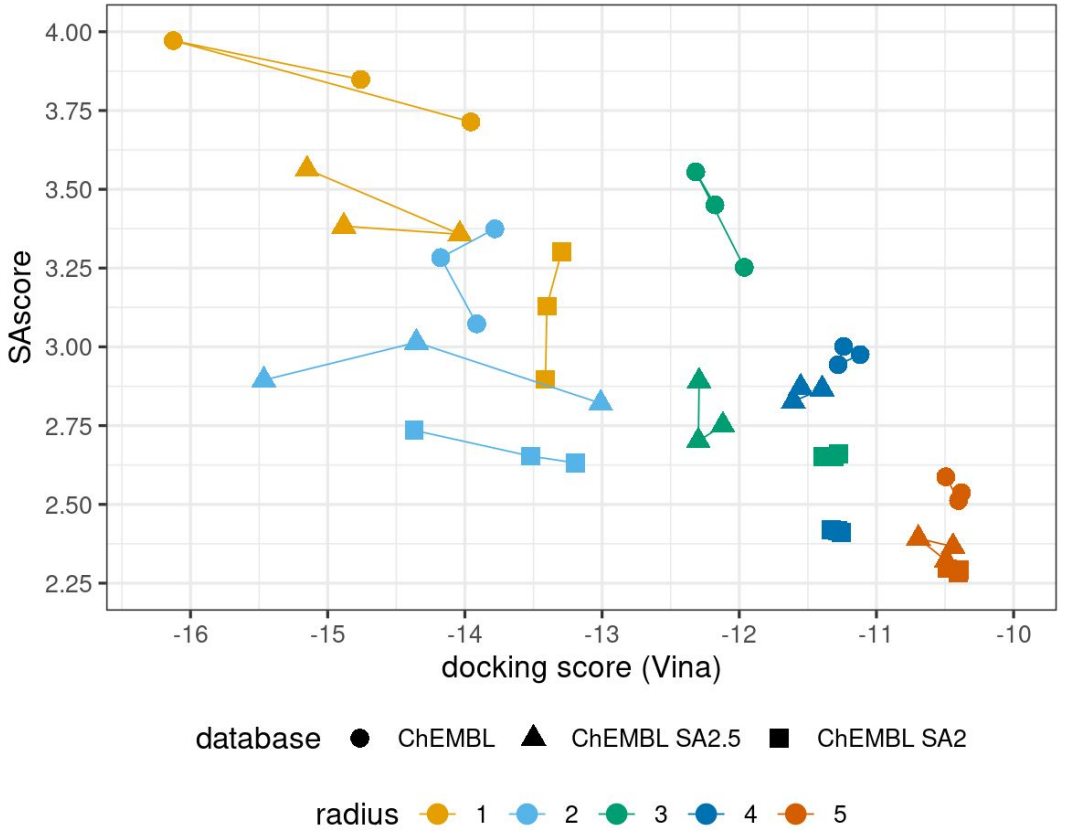
Average docking and SA scores for top 100 molecules from each run



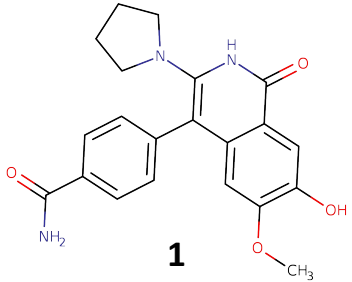
De novo design using docking (CDK2 inhibitors)



Average docking and SA scores for top 100 molecules from each run

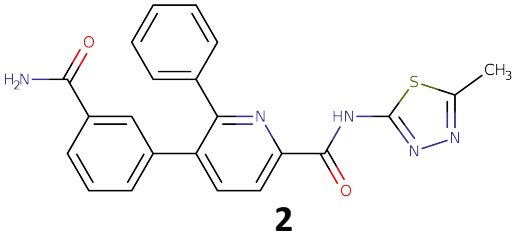


De novo design using docking (CDK2 inhibitors)



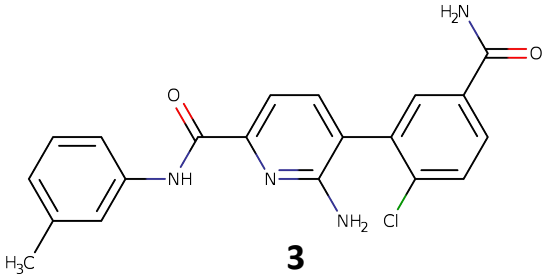
1

docking score = -16.5, SAscore 2.49



2

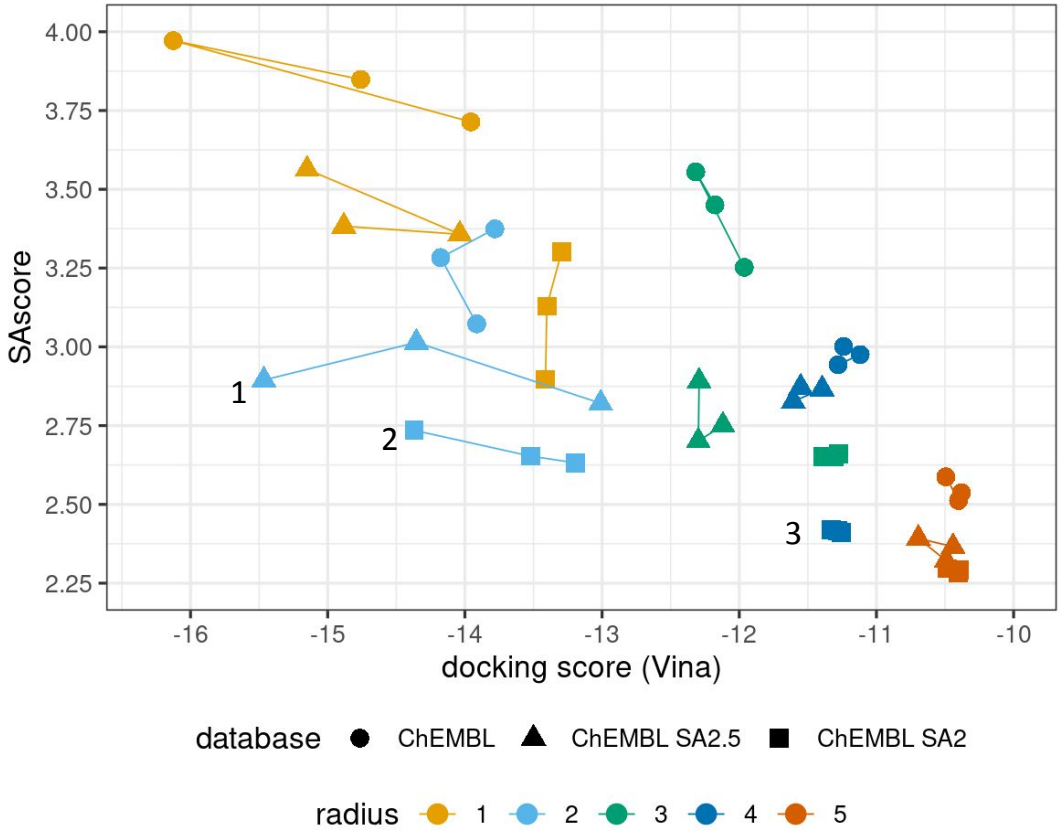
docking score = -14.87, SAscore 2.25



3

docking score = -11.71, SAscore 2.16

Average docking and SA scores for top 100 molecules from each run



De novo design using docking (CDK2 inhibitors)

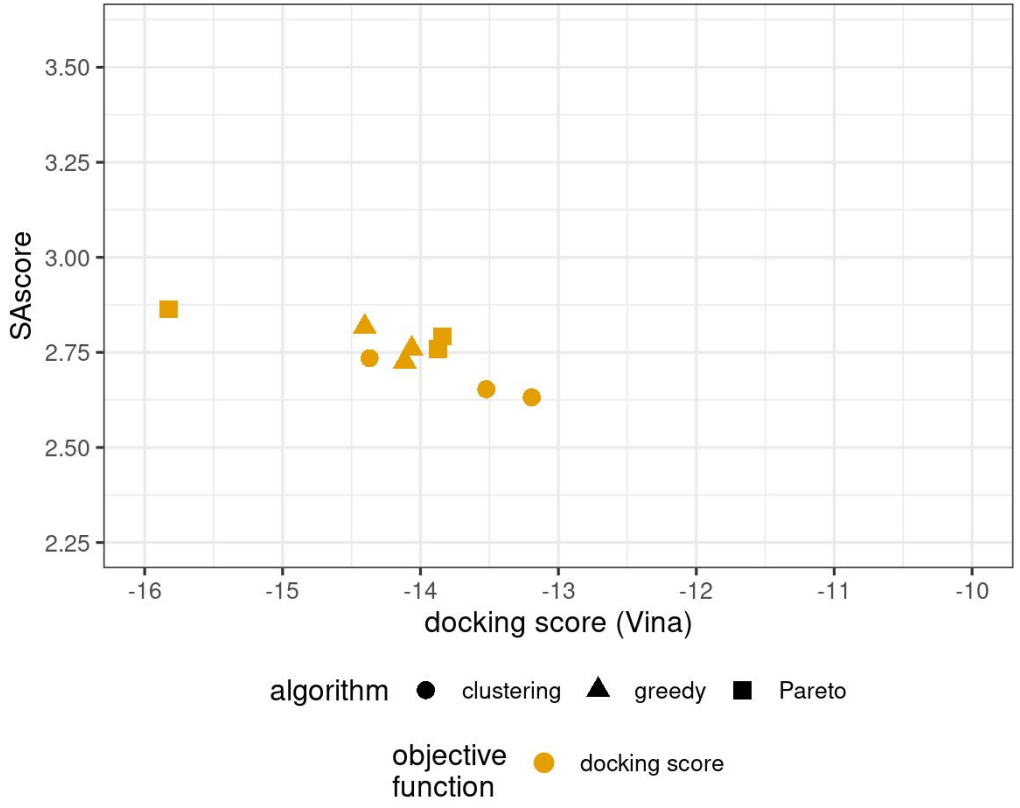
Variable conditions:

selection algorithms:

- ➔ clustering
- ➔ greedy
- ➔ using Pareto front

Constant conditions:

- MW ≤ 450
- RTB ≤ 5
- logP ≤ 4
- TPSA ≤ 120
- hinge region binding
- ChEMBL SA2
- radius 2



De novo design using docking (CDK2 inhibitors)

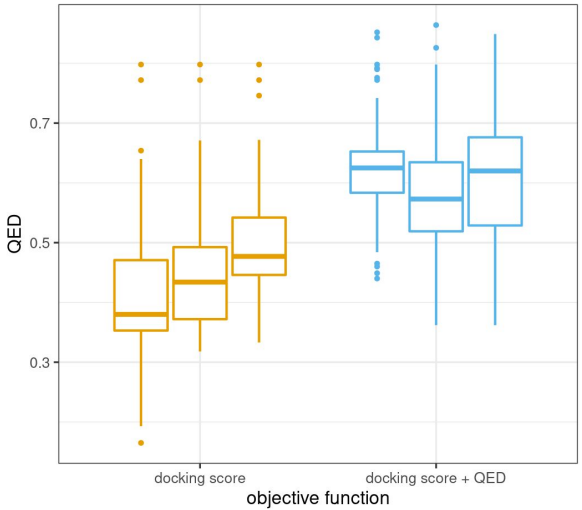
Variable conditions:

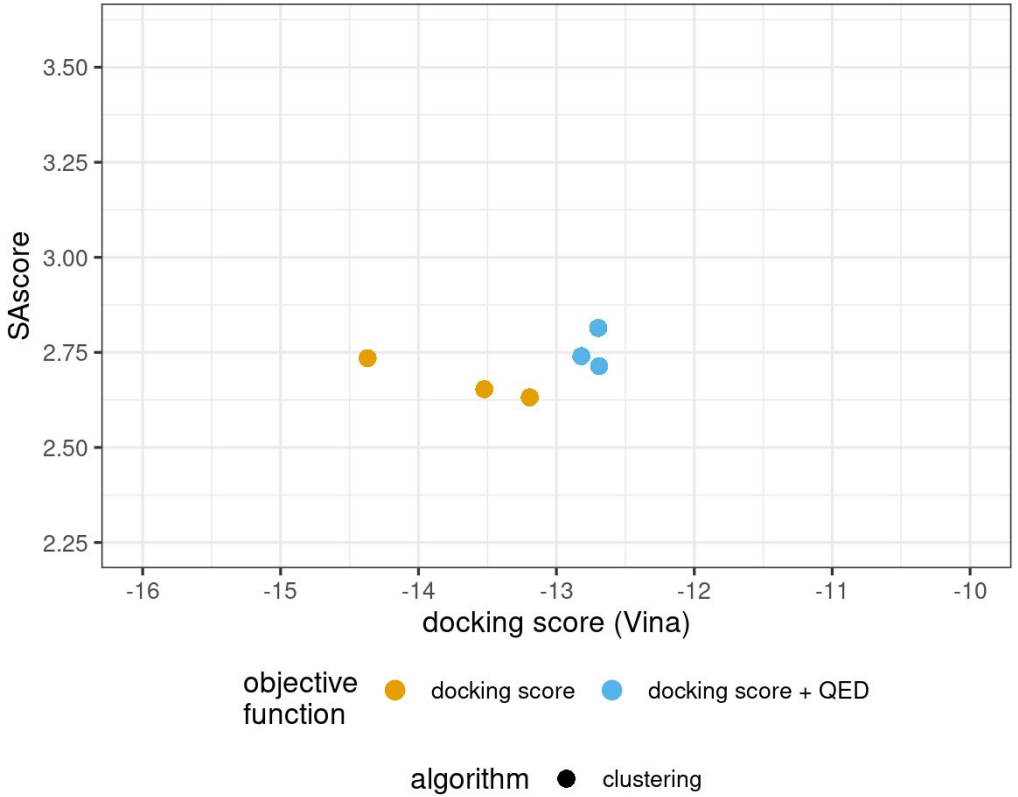
objective functions:

- ➔ docking score
- ➔ docking score + QED

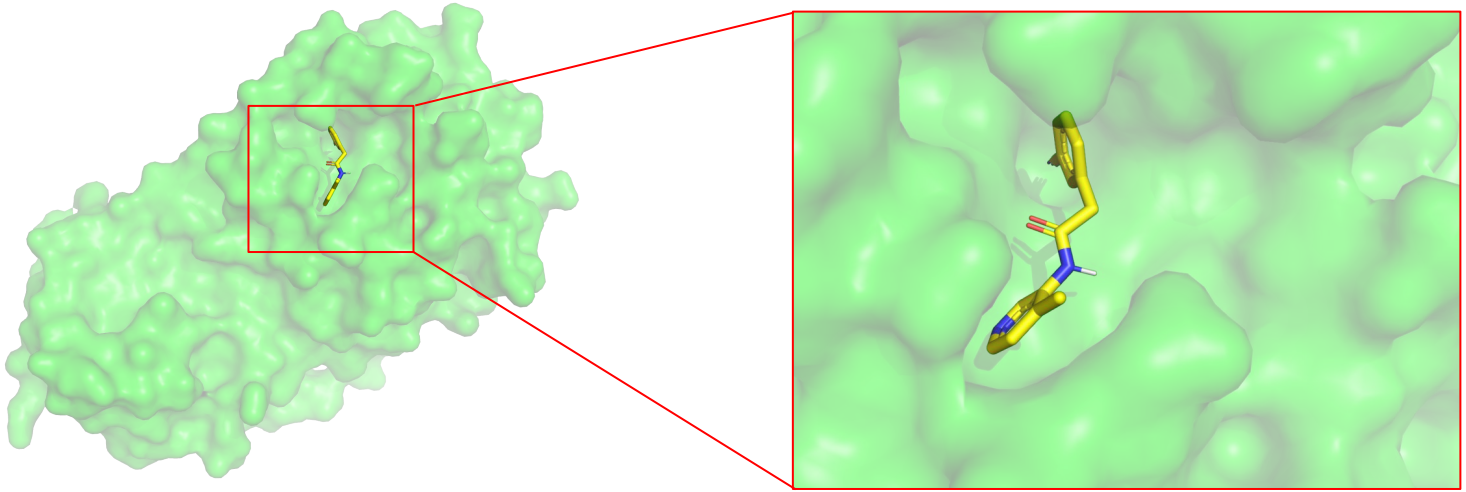
Constant conditions:

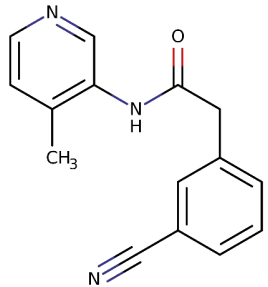
- MW ≤ 450
- RTB ≤ 5
- logP ≤ 4
- TPSA ≤ 120
- hinge region binding
- ChEMBL SA2
- radius 2
- clustering





Hit expansion: inhibitors of main protease SARS-CoV-2



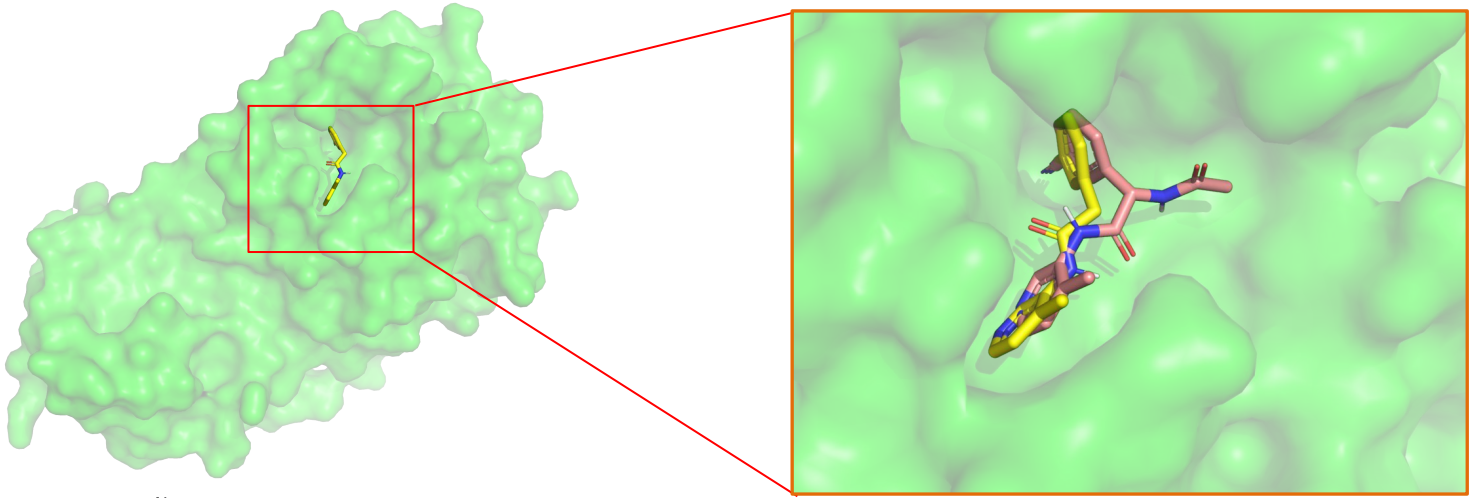


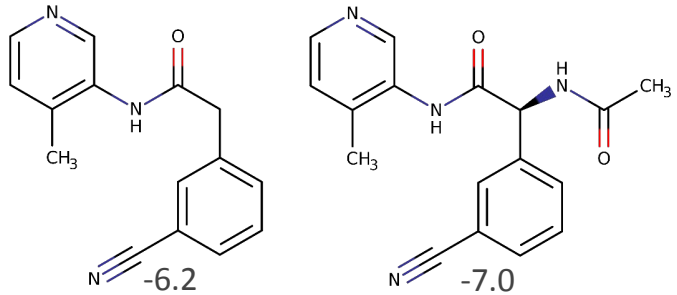
-6.2

docking score (Autodock Vina)

* in collaboration LifeChemicals company (Ukraine) and UOCHB (Prague)

Hit expansion: inhibitors of main protease SARS-CoV-2

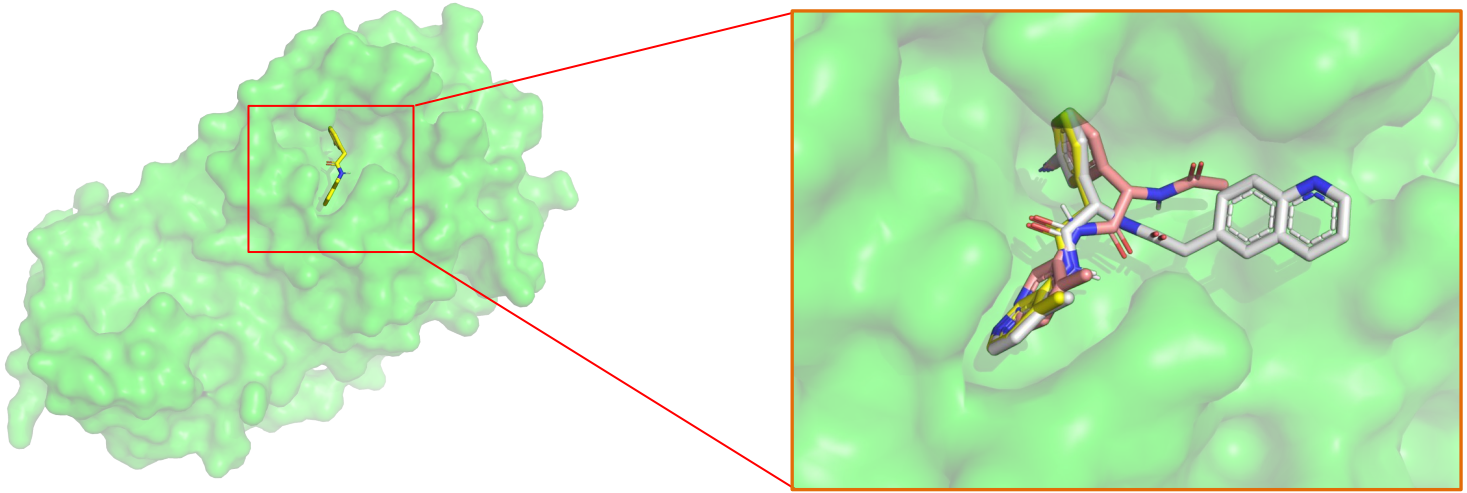


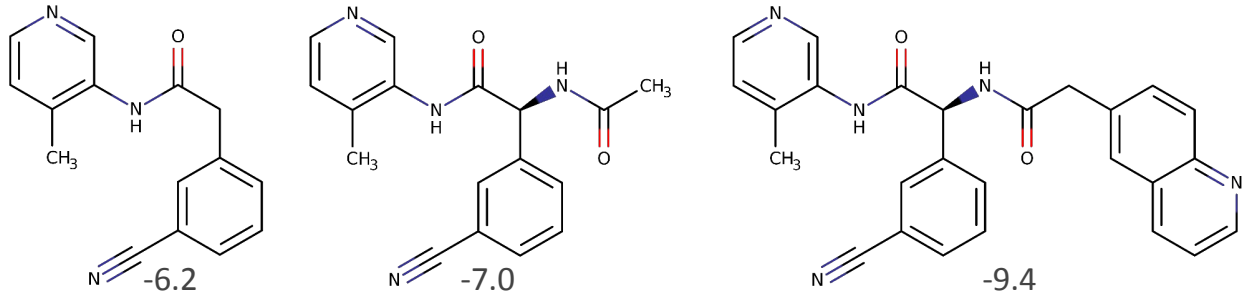


docking score (Autodock Vina)

* in collaboration LifeChemicals company (Ukraine) and UOCHB (Prague)

Hit expansion: inhibitors of main protease SARS-CoV-2

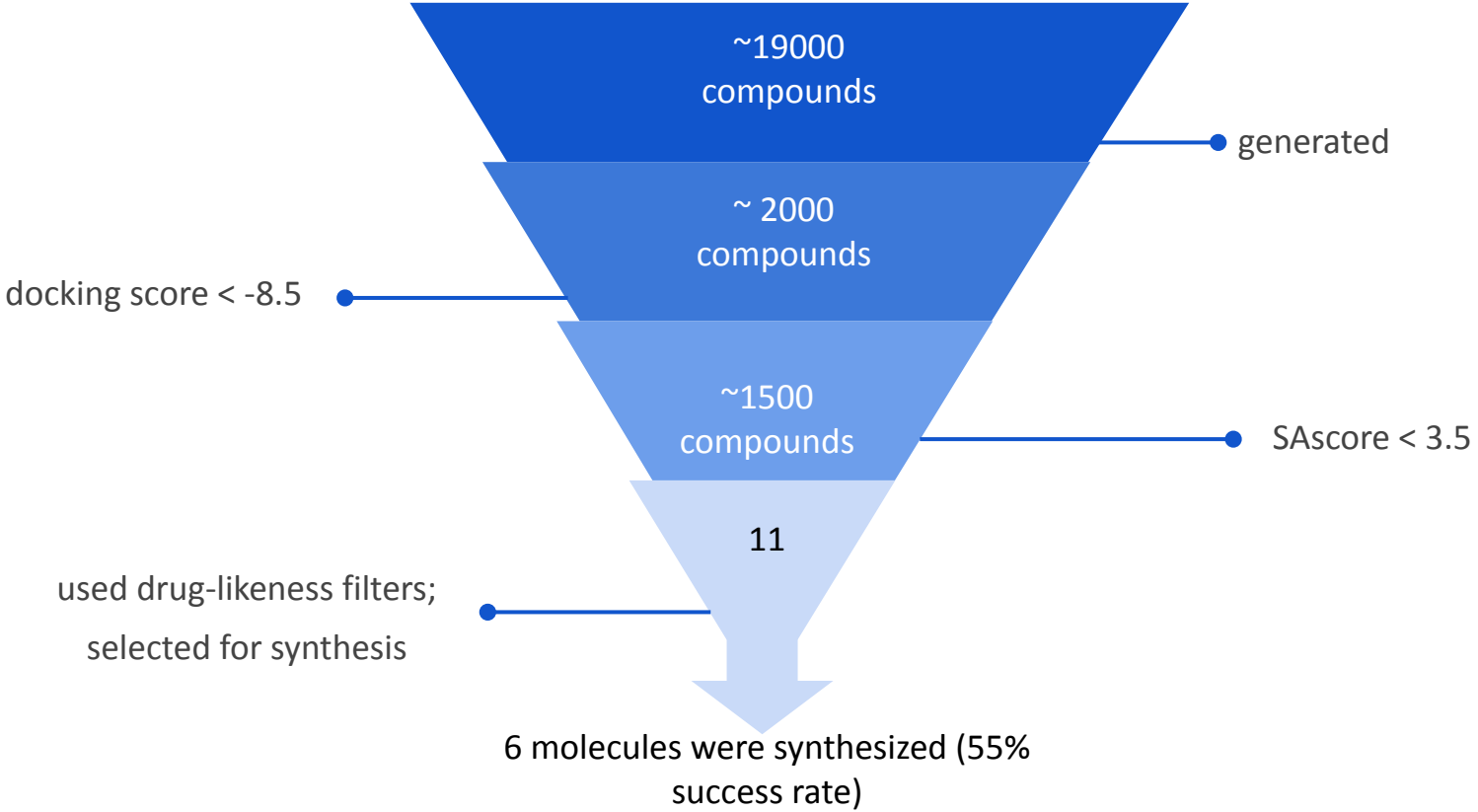


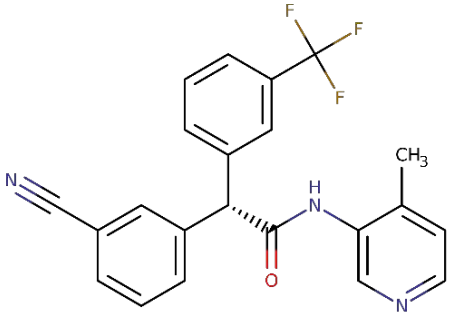


docking score (Autodock Vina)

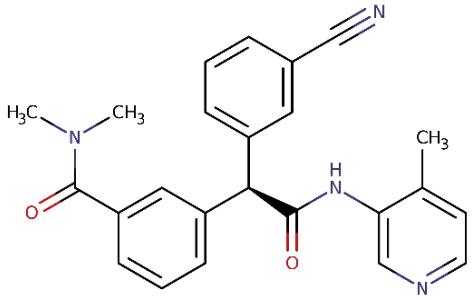
* in collaboration LifeChemicals company (Ukraine) and UOCHB (Prague)

Results

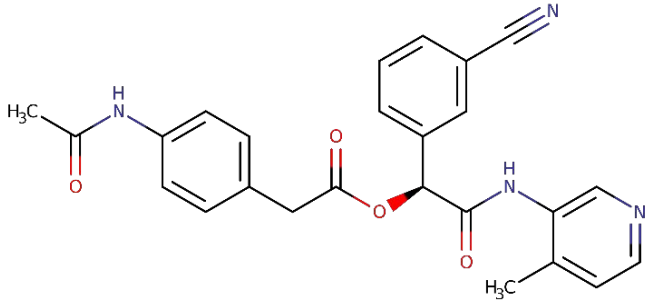

~19000 compounds
generated
~ 2000 compounds
docking score < -8.5
~1500 compounds
SAscore < 3.5
11
used drug-likeness filters;
selected for synthesis
6 molecules were synthesized (55% success rate)



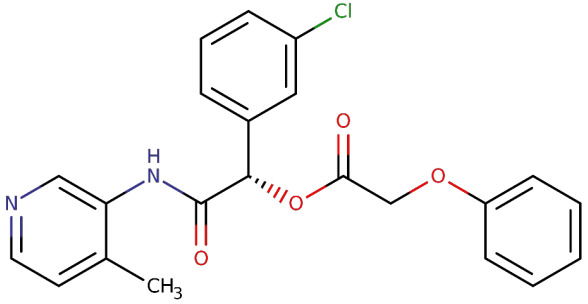
docking score = -9.4, SAscore 2.84



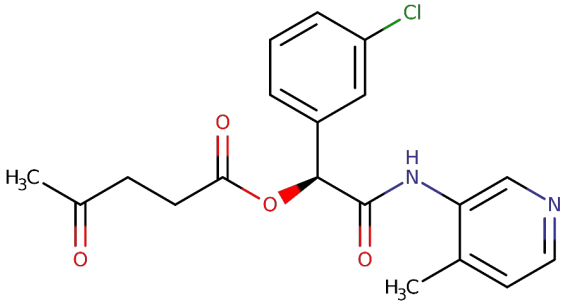
docking score = -9.8, SAscore 2.82



docking score = -9.8, SAscore 3.18



docking score = -9.7, SAscore 2.58



docking score = -9.5, SAscore 2.73

* in collaboration LifeChemicals company (Ukraine) and UOCHB (Prague)

Summary

1. The developed tool was able to autonomously generate synthetically accessible molecules.
2. Choosing more restricted fragment databases and greater context radius one may improve synthetic accessibility of generated molecules.
3. Docking score depends stronger on a chosen radius rather than on a fragment database.
4. Using greedy selection results in highly reproducible runs but with lower diversity of generated molecules, whereas selection based on clustering and Pareto approaches gives more diverse and variable output.
5. Objective function can be adjusted with additional parameters, for example drug-likeness to bias generation towards a more favorable region of chemical space.
6. The designed molecules demonstrated moderate real synthetic feasibility in the task of searching of inhibitors of SARS-CoV-2 main protease.

Future directions

1. Compare with state-of-the-art tools: OpenGrowth, AutoGrow4
2. Study success rates of chemical syntheses based on custom fragment databases (in collaboration with LIfeChemicals).
3. Application of the developed tool to ongoing medicinal chemistry projects on hit identification and lead optimization, e.g. CACHE challenge, internal projects.

Acknowledgments

UOCHB:
Tatana Majerova

LifeChemicals:
Vladimir Fetyukhin

This work was funded by the INTER-EXCELLENCE LTARF18013 project (MEYS), the European Regional Development Fund - Project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868) and ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131)

Thank you for attention

guzel.minibaeva01@upol.cz