Vienna University, 12 October 2023

The first CACHE challenge: searching for hit molecules in ultra-large chemical databases

Pavel Polishchuk

Institute of Molecular and Translational Medicine
Faculty of Medicine and Dentistry
Palacky University
Czech Republic
pavlo.polishchuk@upol.cz
https://imtm.cz/chemoinformatics-and-drug-design

CACHE challenge

Competition among top chemoinformatics groups world-wide

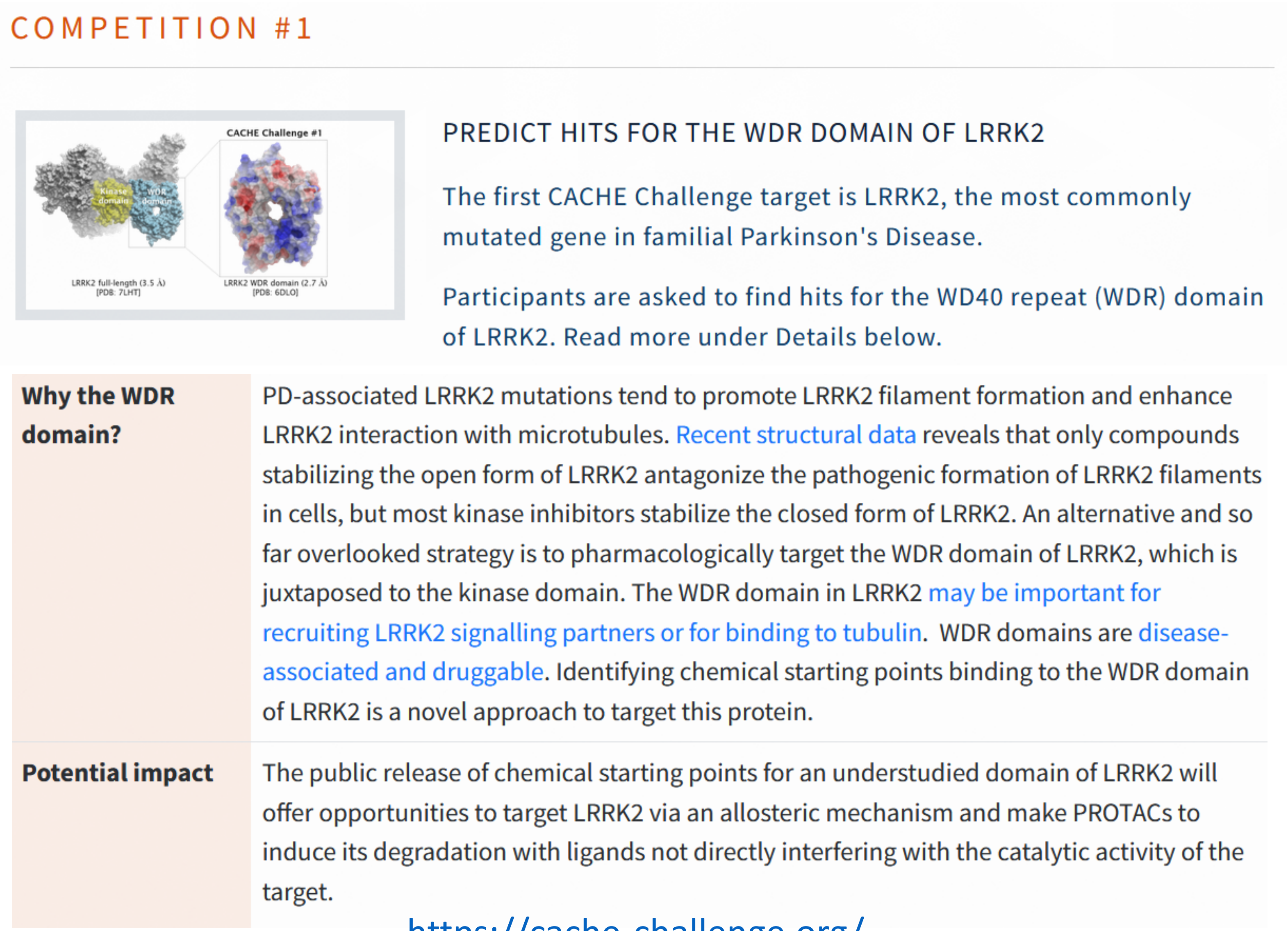
Benefits supposed by organizers:

1. Encourage development and improvement of computational tools
2. Create a platform for prospective validation and comparison of different modeling tools and pipelines
3. Identify hit compounds for challenging or emerging targets/diseases
4. Contribute to open science to accelerate researches in a chosen direction

Our motivation

1. Validate and improve our developed modeling tools in a competitive environment
2. Establish robust and reliable computational pipelines which can be further easily applied in other projects

The first CACHE challenge



CACHE Challenge #1
Kinase domain
WDR domain
LRRK2 full-length (3.5 Å) [PDB: 7LHT]
LRRK2 WDR domain (2.7 Å) [PDB: 6DLO]

COMPETITION #1

PREDICT HITS FOR THE WDR DOMAIN OF LRRK2

The first CACHE Challenge target is LRRK2, the most commonly mutated gene in familial Parkinson's Disease.

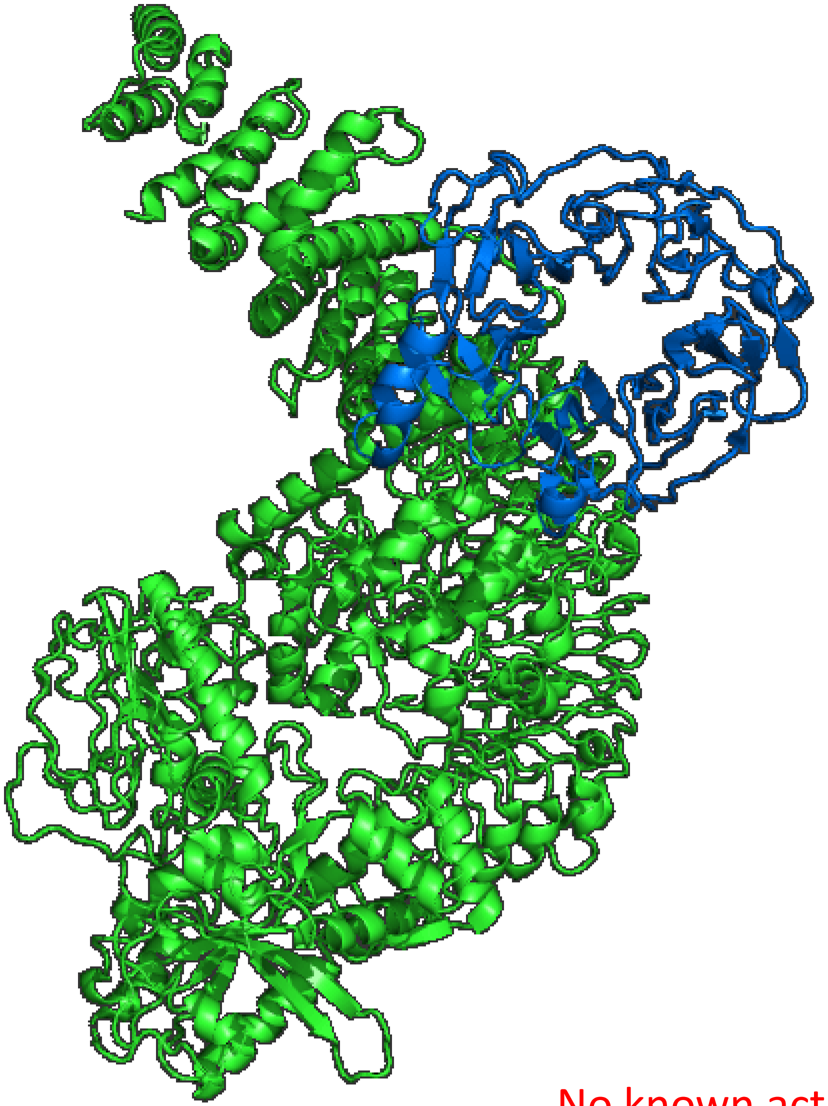
Participants are asked to find hits for the WD40 repeat (WDR) domain of LRRK2. Read more under Details below.

Why the WDR domain?	PD-associated LRRK2 mutations tend to promote LRRK2 filament formation and enhance LRRK2 interaction with microtubules. Recent structural data reveals that only compounds stabilizing the open form of LRRK2 antagonize the pathogenic formation of LRRK2 filaments in cells, but most kinase inhibitors stabilize the closed form of LRRK2. An alternative and so far overlooked strategy is to pharmacologically target the WDR domain of LRRK2, which is juxtaposed to the kinase domain. The WDR domain in LRRK2 may be important for recruiting LRRK2 signalling partners or for binding to tubulin. WDR domains are disease-associated and druggable. Identifying chemical starting points binding to the WDR domain of LRRK2 is a novel approach to target this protein.
Potential impact	The public release of chemical starting points for an understudied domain of LRRK2 will offer opportunities to target LRRK2 via an allosteric mechanism and make PROTACs to induce its degradation with ligands not directly interfering with the catalytic activity of the target.

https://cache-challenge.org/

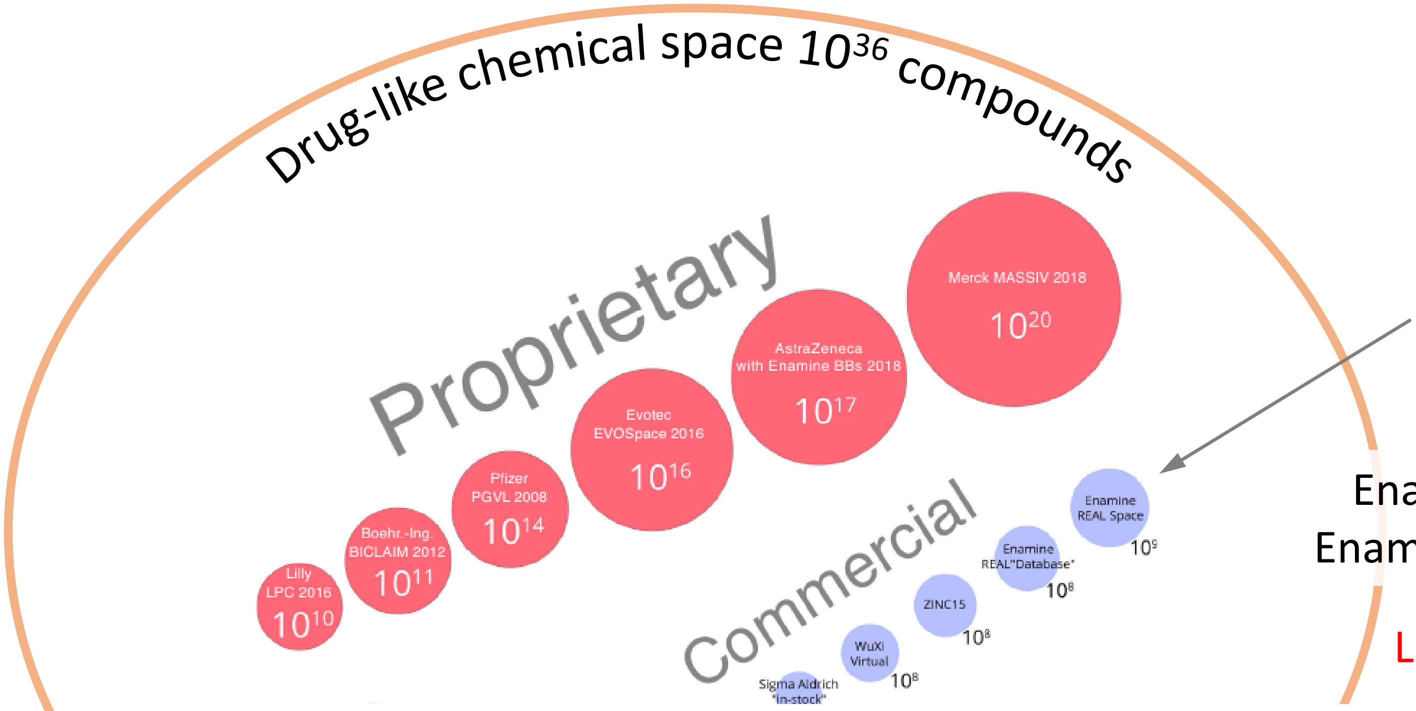
Ackloo, S. et al. *Nature Reviews Chemistry* **2022**, 6, 287-295.

LRRK2 and WDR domain



No known active molecules
No X-ray of protein-ligand complexes

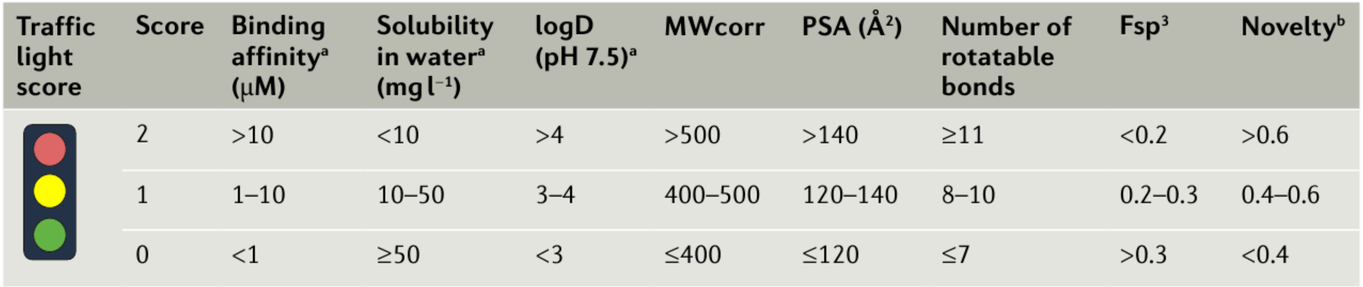
Chemical search space



Enamine Stock: 2.5M

Enamine Real Space: 16B

Large space to explore



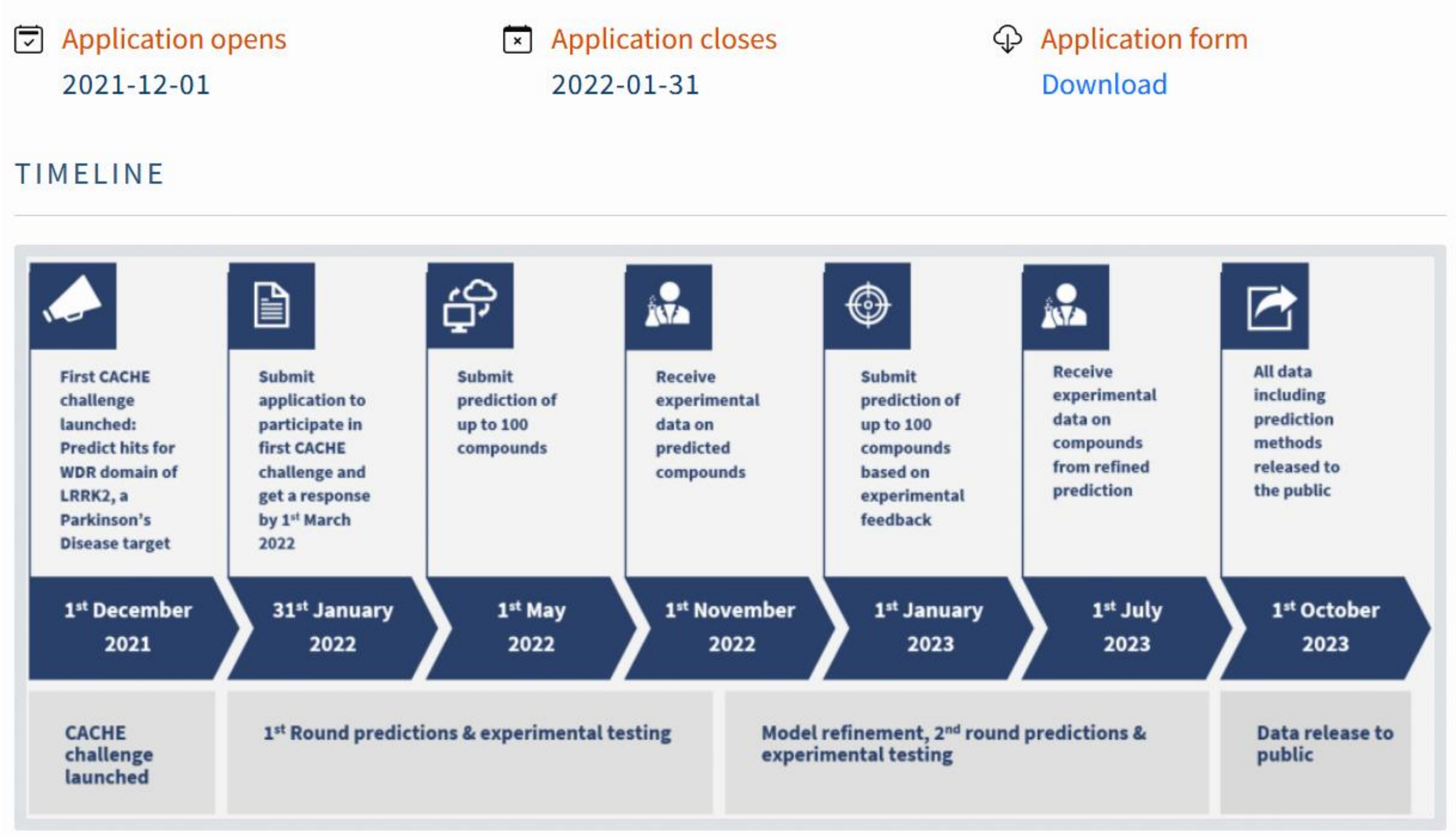
Traffic light score	Score	Binding affinity[a] (μM)	Solubility in water[a] (mg l⁻¹)	logD (pH 7.5)[a]	MWcorr	PSA (Å²)	Number of rotatable bonds	Fsp³	Novelty[b]
🔴	2	>10	<10	>4	>500	>140	≥11	<0.2	>0.6
🟡	1	1–10	10–50	3–4	400–500	120–140	8–10	0.2–0.3	0.4–0.6
🟢	0	<1	≥50	<3	≤400	≤120	≤7	>0.3	<0.4

Fsp³, fraction of sp³ hybridized carbon atoms, calculated based on Murcko scaffolds. [a]Measured experimentally. [b]Tanimoto distance relative to most similar structures binding that target, as calculated from RDKit. PSA, polar surface area.

Hoffmann, T.; Gastreich, M., The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* **2019**, 24, 1148-1156

Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A., Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design* **2013**, 27, 675-679

CACHE challenge pipeline



Application opens
2021-12-01

Application closes
2022-01-31

Application form
Download

TIMELINE

First CACHE challenge launched: Predict hits for WDR domain of LRRK2, a Parkinson's Disease target

1st December 2021

Submit application to participate in first CACHE challenge and get a response by 1st March 2022

31st January 2022

Submit prediction of up to 100 compounds

1st May 2022

Receive experimental data on predicted compounds

1st November 2022

Submit prediction of up to 100 compounds based on experimental feedback

1st January 2023

Receive experimental data on compounds from refined prediction

1st July 2023

All data including prediction methods released to the public

1st October 2023

CACHE challenge launched

1st Round predictions & experimental testing

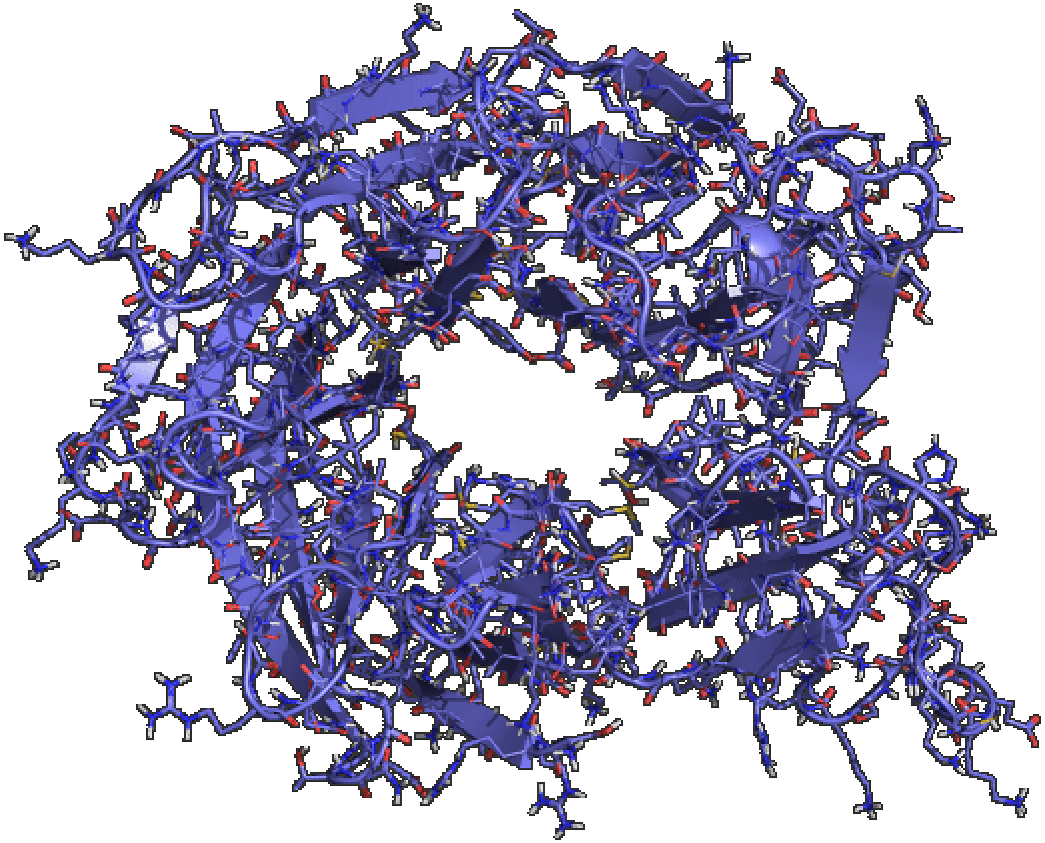
Model refinement, 2nd round predictions & experimental testing

Data release to public

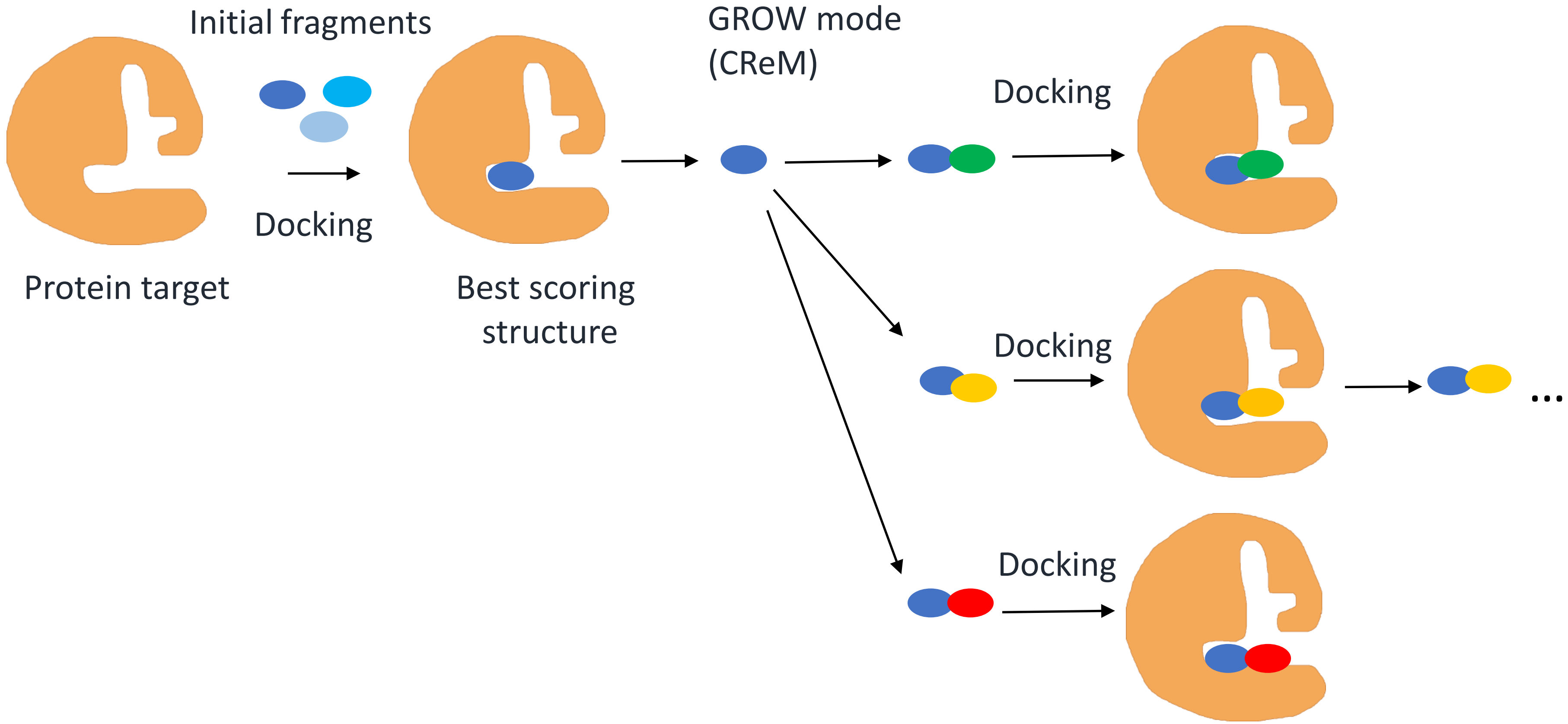
Round 1

WDR domain structure is **available**: 6DLO
Known ligand are **not available**

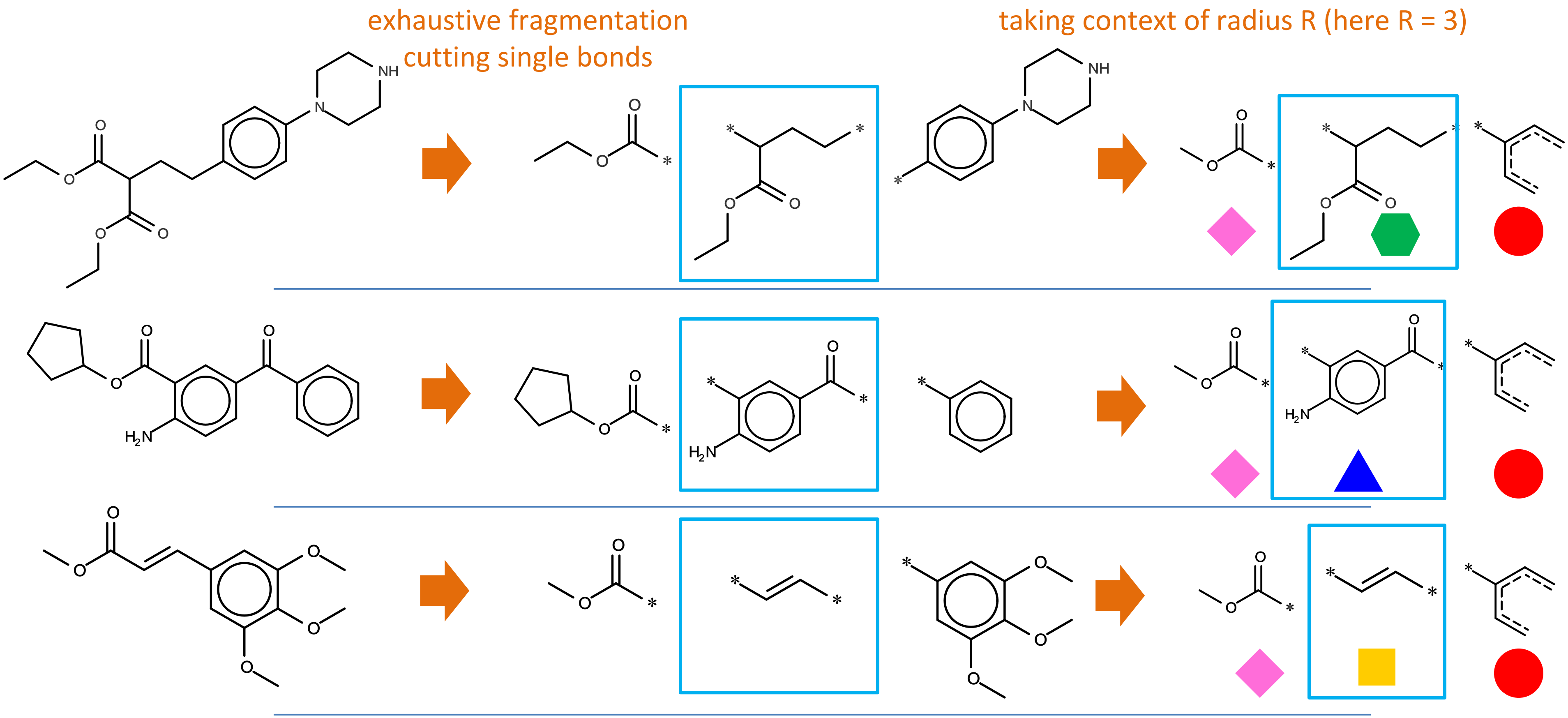
Only structure-based approaches are applicable: **molecular docking** and **dynamics**



Round 1: strategy 1 (de novo design)



Chemically reasonable mutations (CReM)



exhaustive fragmentation cutting single bonds

taking context of radius R (here R = 3)

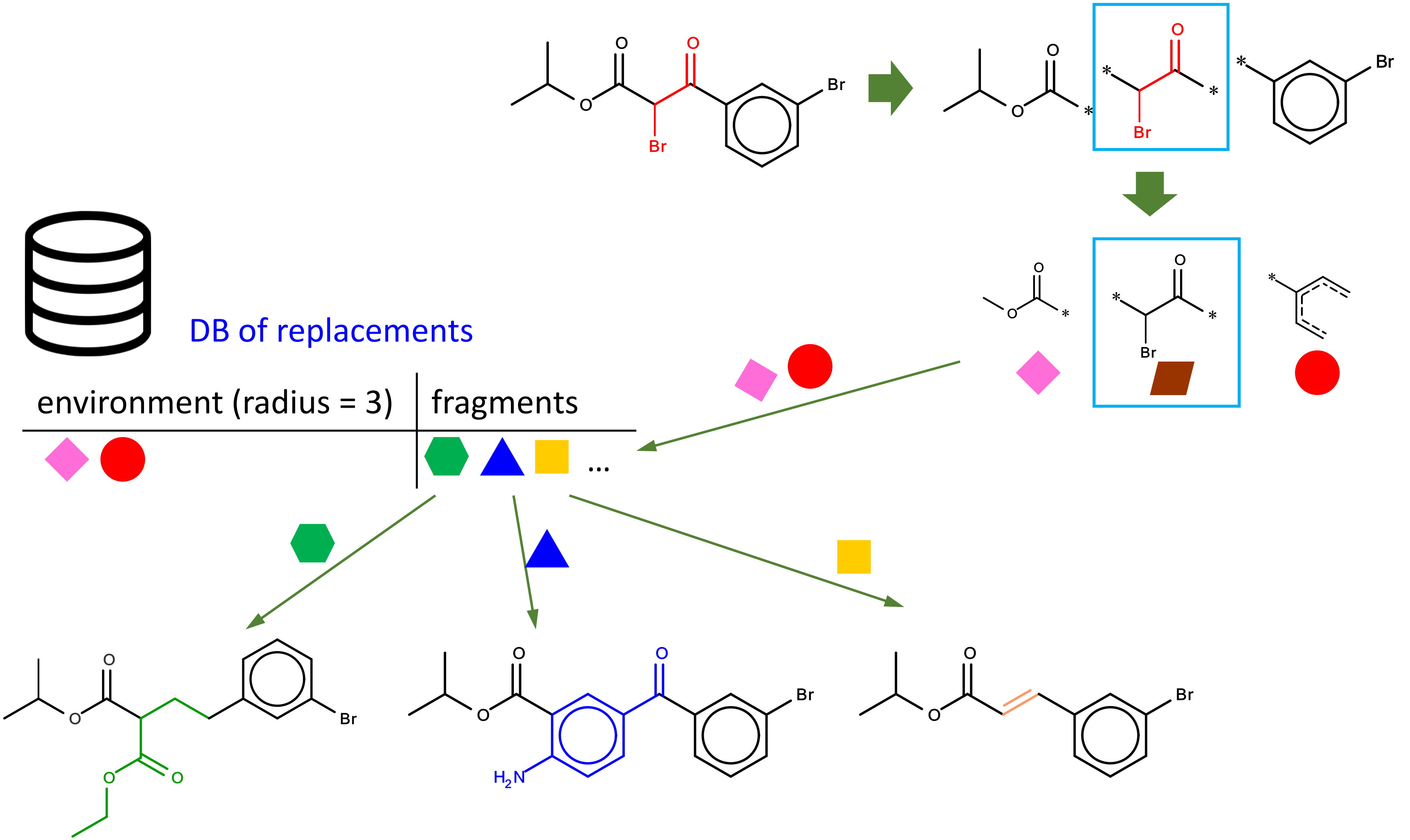
DB of replacements

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)



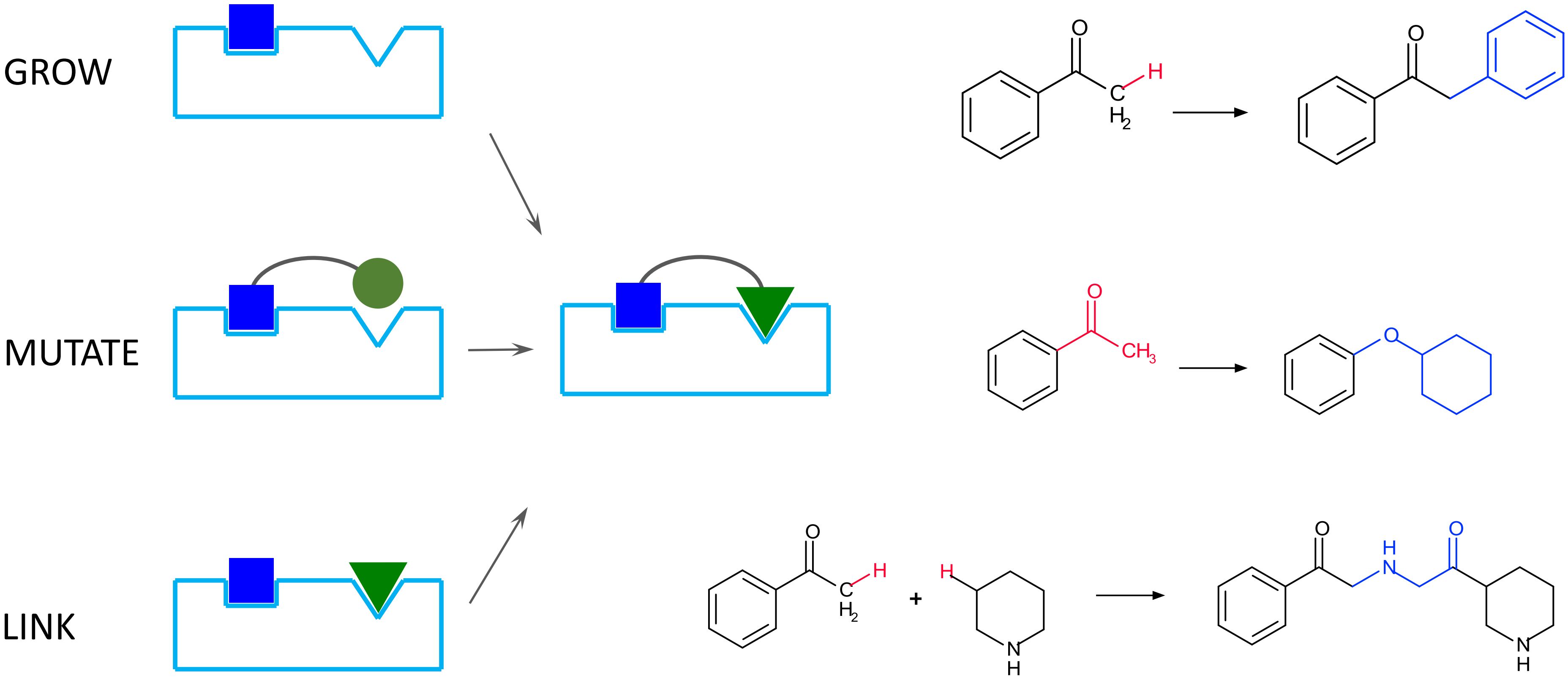
DB of replacements

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

Generated structures are always chemically valid!

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

Chemically reasonable mutations (CReM)



GROW

MUTATE

LINK

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

IMTM

Tweak synthetic accessibility 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

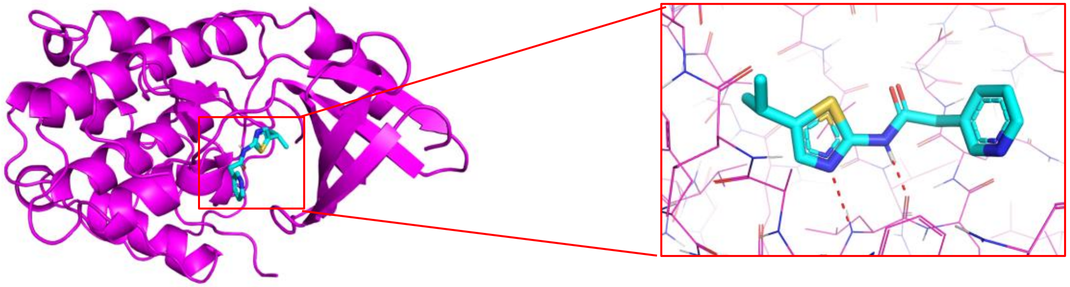
3

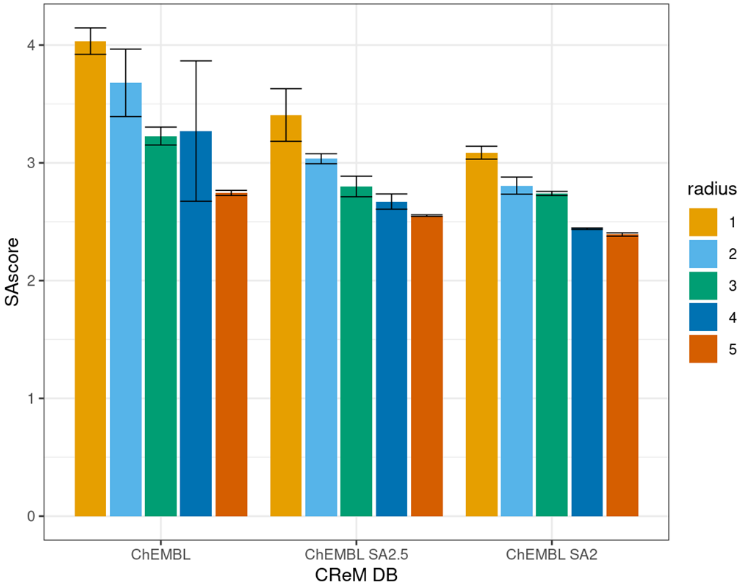
4 more conservative replacements

5

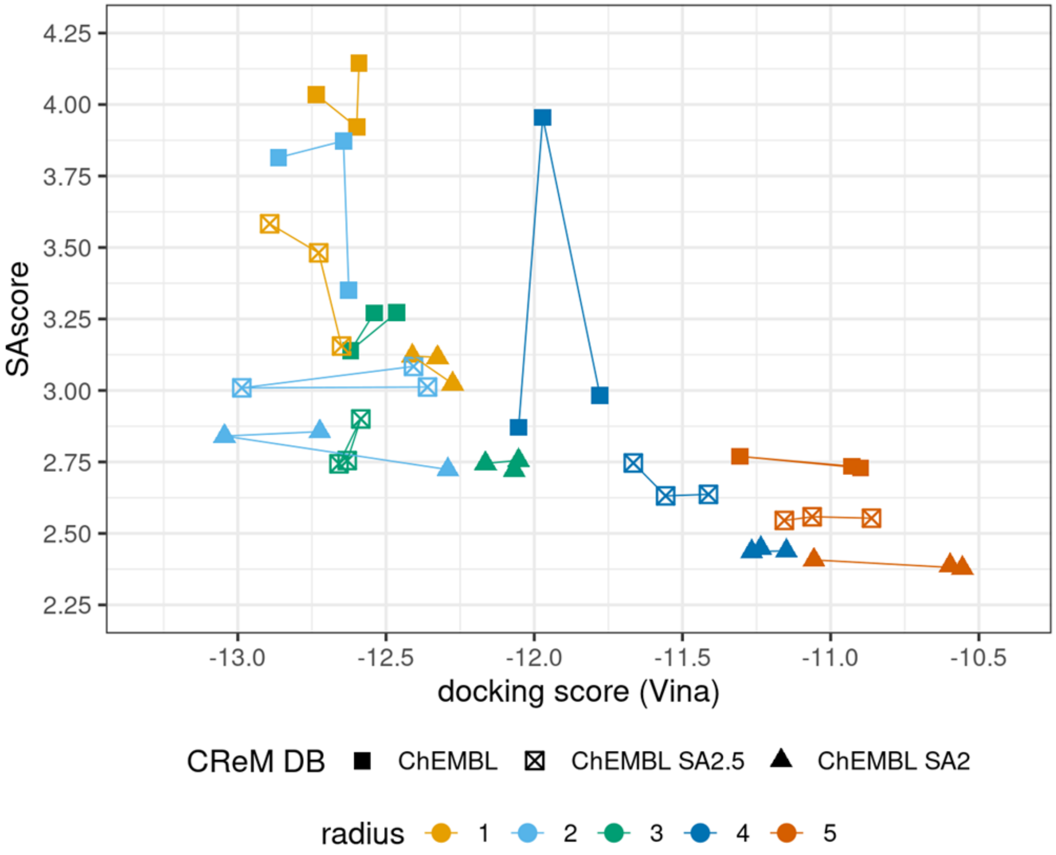
De novo design using docking (example)

2BTR
IC_{50} = 95 nM
docking score = -7.86





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



De novo design using docking (example)

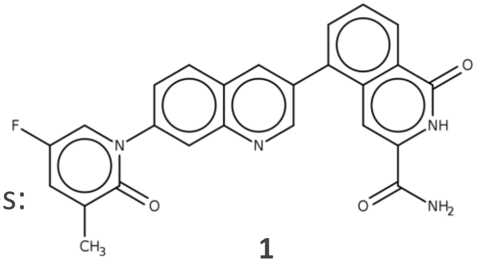
Constant conditions:

- hinge region binding
- ChEMBL SA2
- radius 2

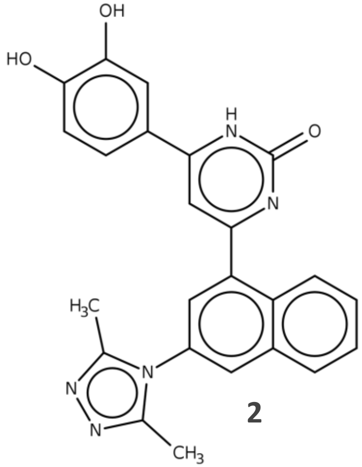
Variable conditions:

different CDK2 complexes:

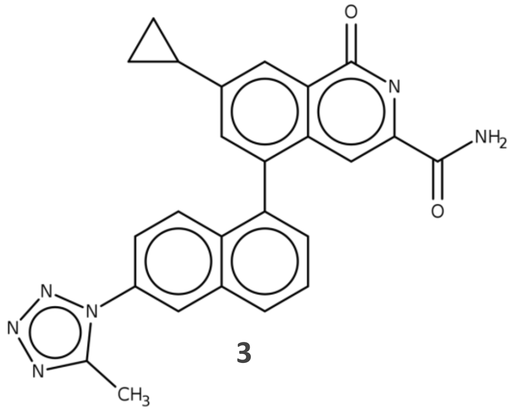
- 2BTR
- 2FVD
- 3RAL
- 6GUH



1

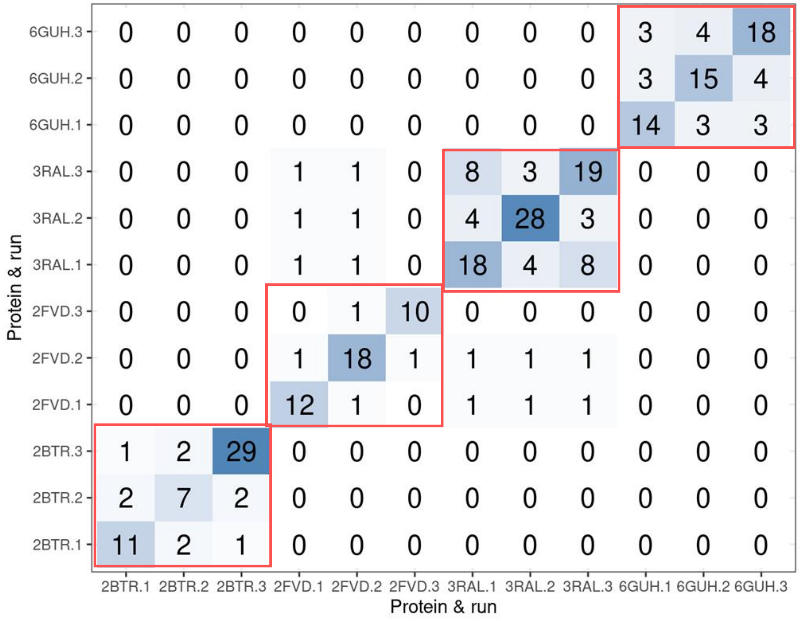


2



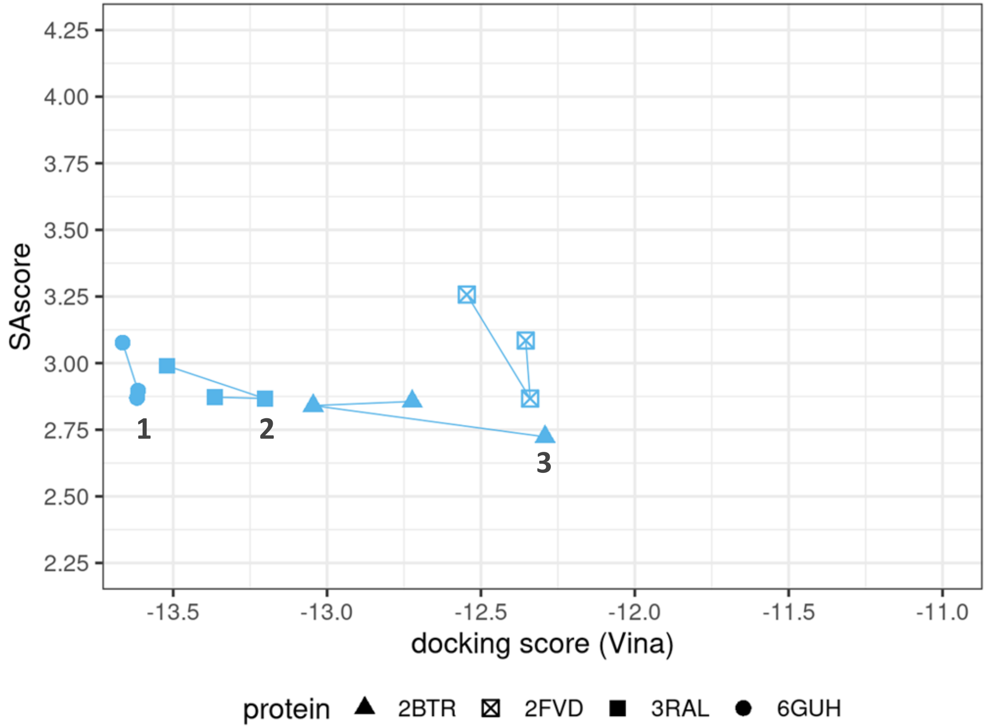
3

The number of distinct Murcko scaffolds in top 100 scored compounds in different runs and their intersection across runs

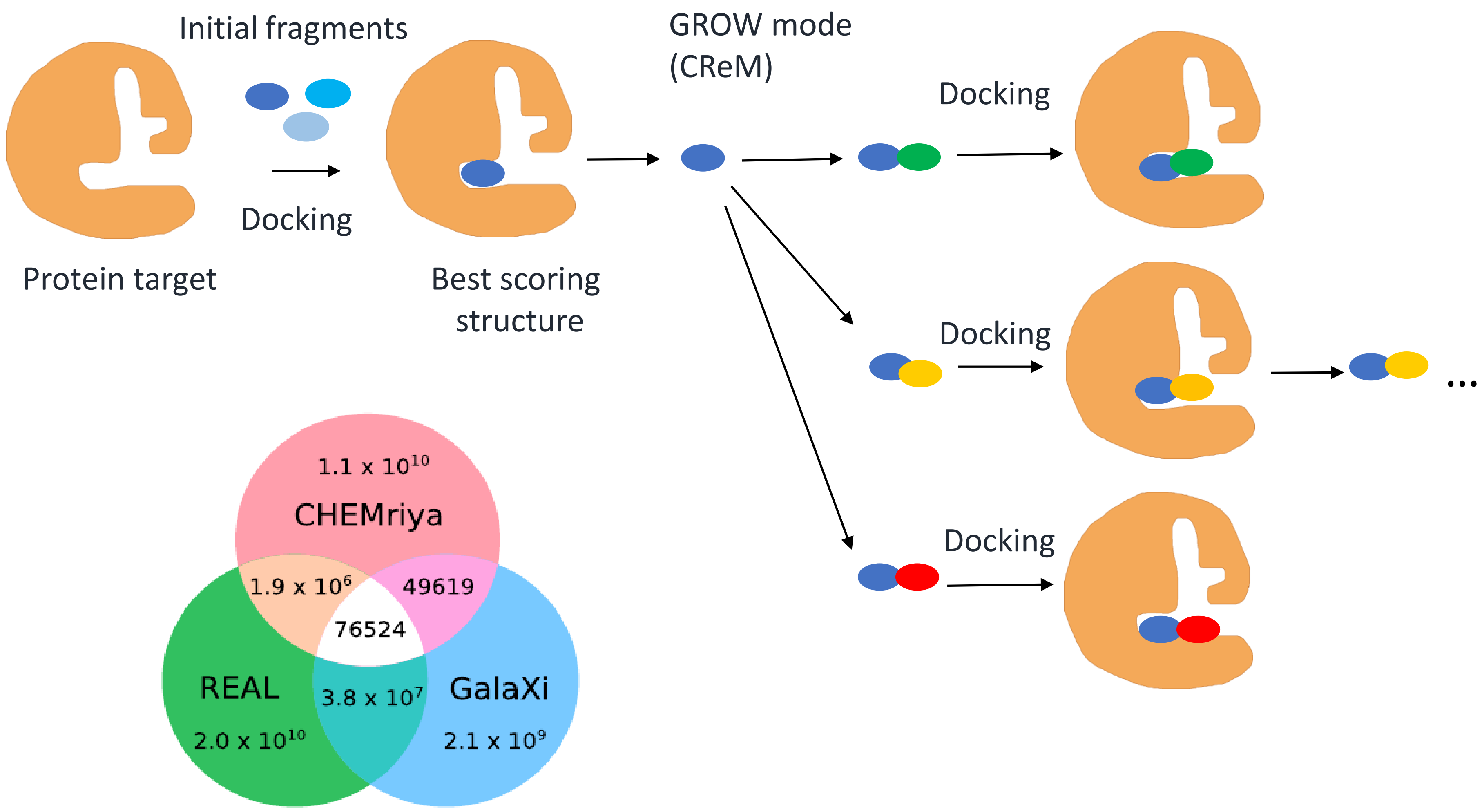


Protein & run	2BTR.1	2BTR.2	2BTR.3	2FVD.1	2FVD.2	2FVD.3	3RAL.1	3RAL.2	3RAL.3	6GUH.1	6GUH.2	6GUH.3
6GUH.3	0	0	0	0	0	0	0	0	0	3	4	18
6GUH.2	0	0	0	0	0	0	0	0	0	3	15	4
6GUH.1	0	0	0	0	0	0	0	0	0	14	3	3
3RAL.3	0	0	0	1	1	0	8	3	19	0	0	0
3RAL.2	0	0	0	1	1	0	4	28	3	0	0	0
3RAL.1	0	0	0	1	1	0	18	4	8	0	0	0
2FVD.3	0	0	0	0	1	10	0	0	0	0	0	0
2FVD.2	0	0	0	1	18	1	1	1	1	0	0	0
2FVD.1	0	0	0	12	1	0	1	1	1	0	0	0
2BTR.3	1	2	29	0	0	0	0	0	0	0	0	0
2BTR.2	2	7	2	0	0	0	0	0	0	0	0	0
2BTR.1	11	2	1	0	0	0	0	0	0	0	0	0

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

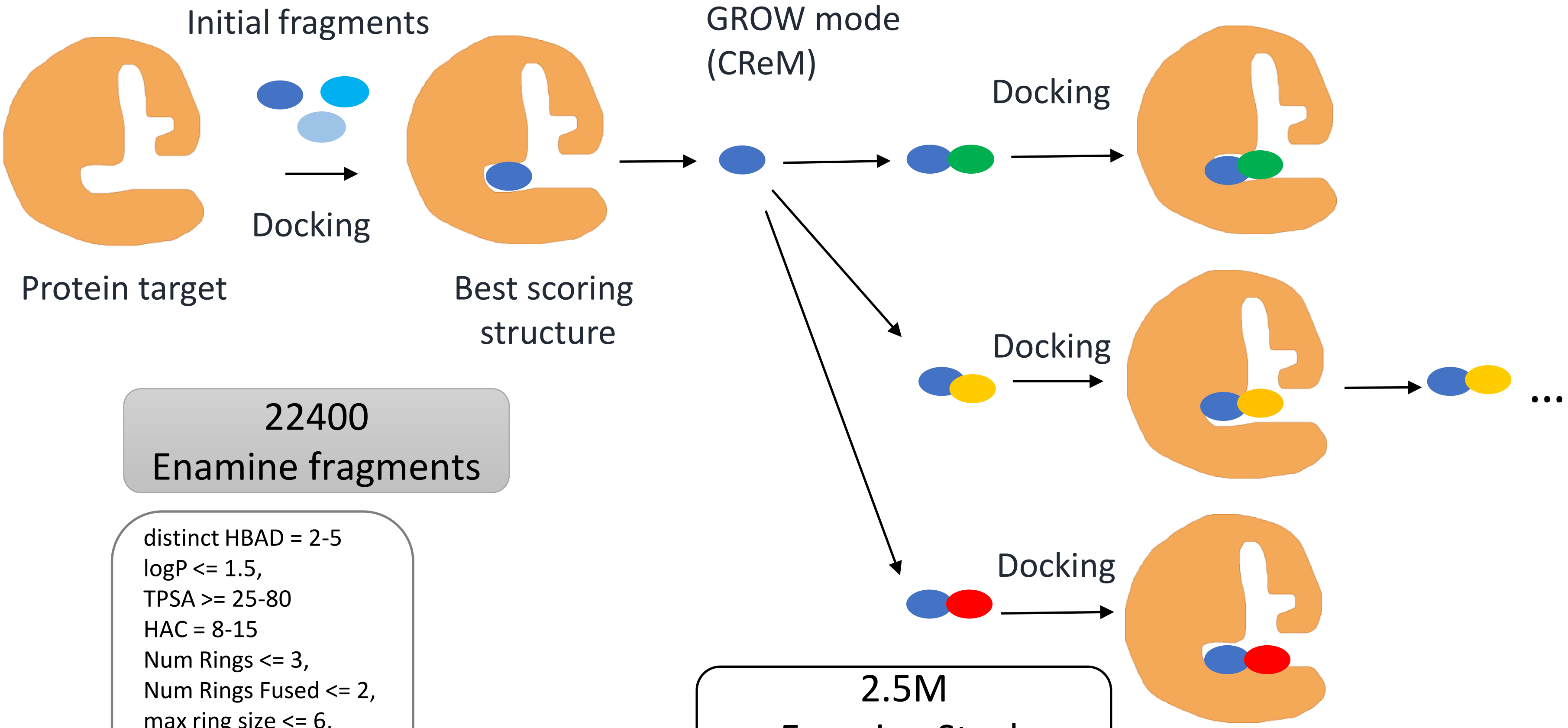


Round 1: strategy 1 (de novo design)



Bellmann, L.; Penner, P.; Gastreich, M.; Rarey, M., Comparison of Combinatorial Fragment Spaces and Its Application to Ultralarge Make-on-Demand Compound Catalogs. *J. Chem. Inf. Model.* **2022,** 62 (3), 553-566.

Round 1: strategy 1 (de novo design)



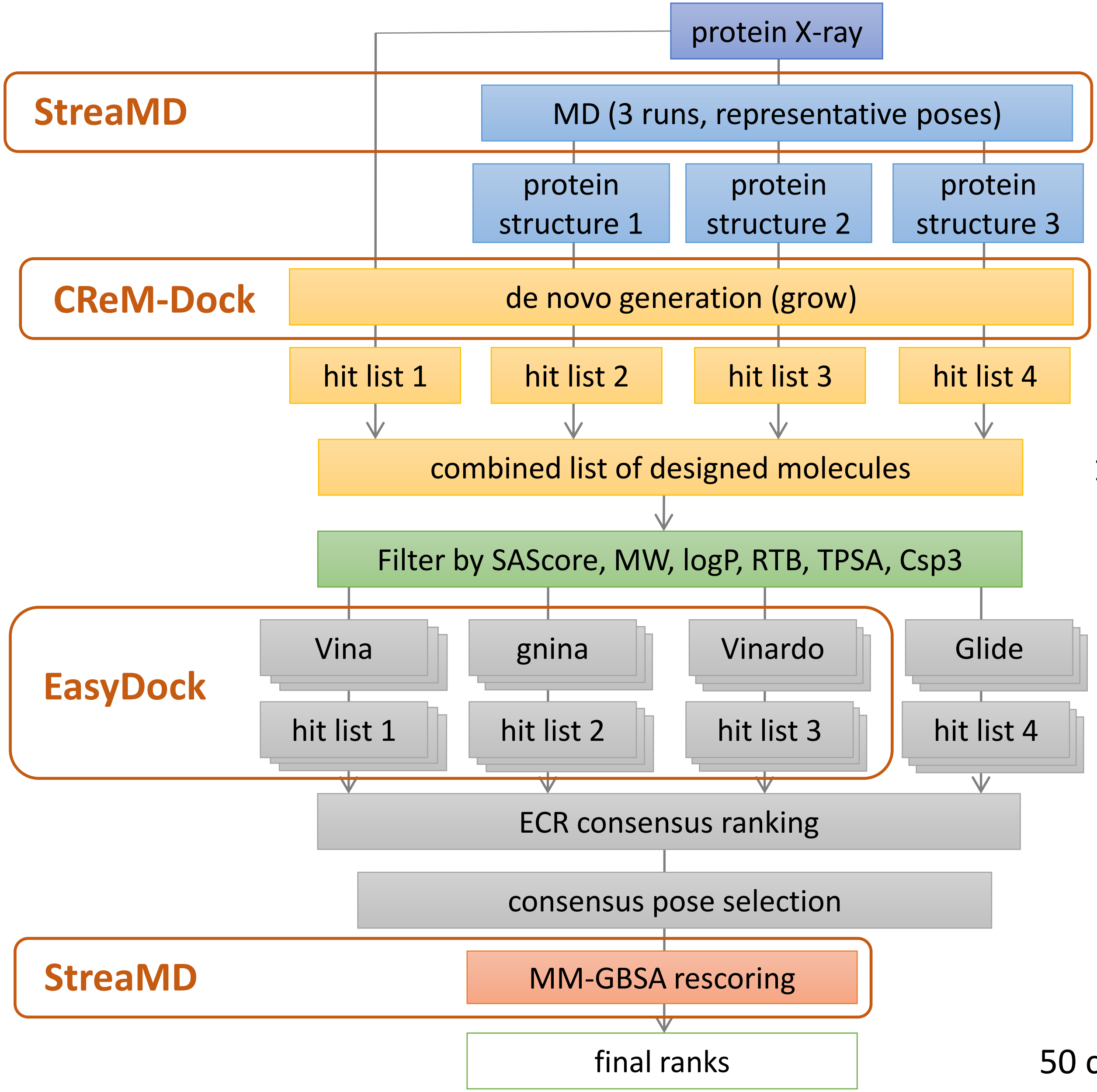
22400
Enamine fragments

distinct HBAD = 2-5
logP <= 1.5,
TPSA >= 25-80
HAC = 8-15
Num Rings <= 3,
Num Rings Fused <= 2,
max ring size <= 6,
nHal <= 1,
ChiralCenters <= 2,
FCsp3_BM >= 0.3

2.5M
Enamine Stock
(SA ≤ 2, SA ≤ 3)

860 000
fragments (1-10 atoms)

Round 1: strategy 1 (de novo design)



protein X-ray

StreaMD

MD (3 runs, representative poses)

protein structure 1

protein structure 2

protein structure 3

CReM-Dock

de novo generation (grow)

hit list 1

hit list 2

hit list 3

hit list 4

combined list of designed molecules

1M compounds

Filter by SAScore, MW, logP, RTB, TPSA, Csp3

267k compounds

EasyDock

Vina

gnina

Vinardo

Glide

3.5k compounds (1.3k scaffolds)

hit list 1

hit list 2

hit list 3

hit list 4

ECR consensus ranking

400 compounds

consensus pose selection

StreaMD

MM-GBSA rescoring

final ranks

50 compounds with distinct scaffolds

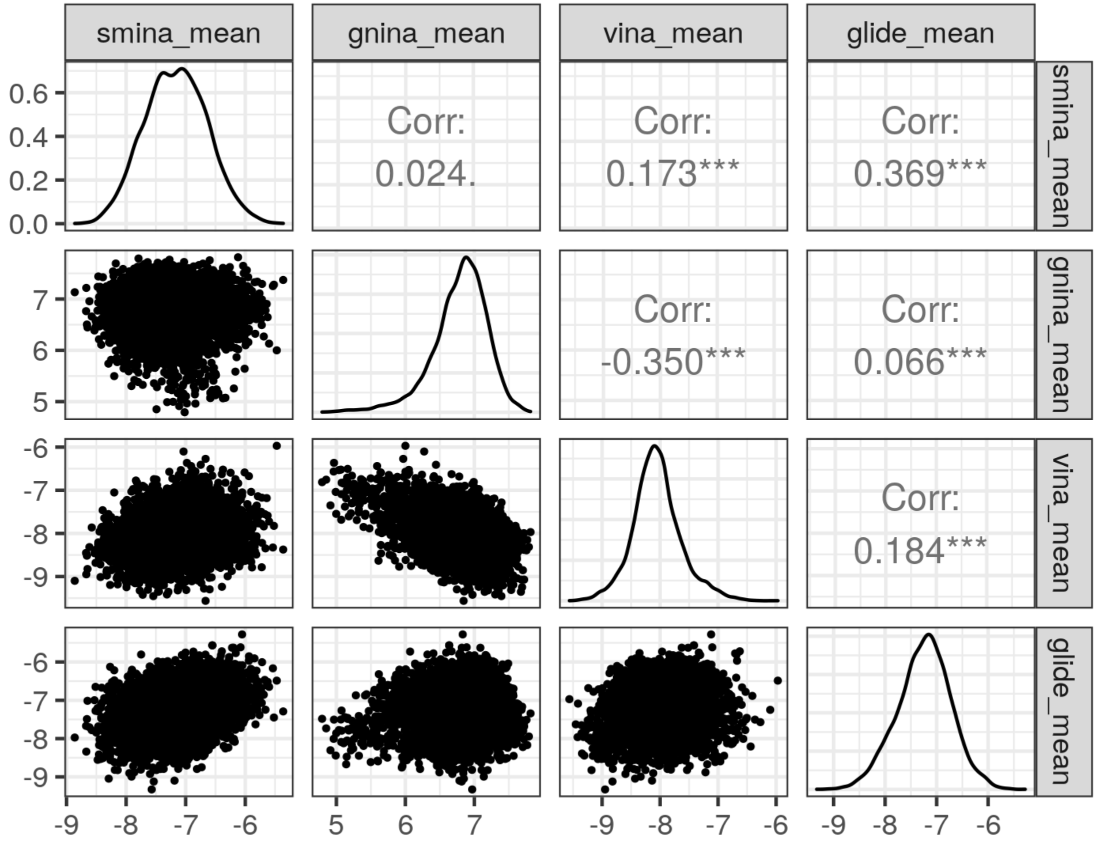
Round 1: strategy 1 (de novo design)

protein X-ray

StreaMD

CReM-Dock

h



s

ds

EasyDock

Vi

hit l

ds (1.3k scaffolds)

ls

consensus pose selection

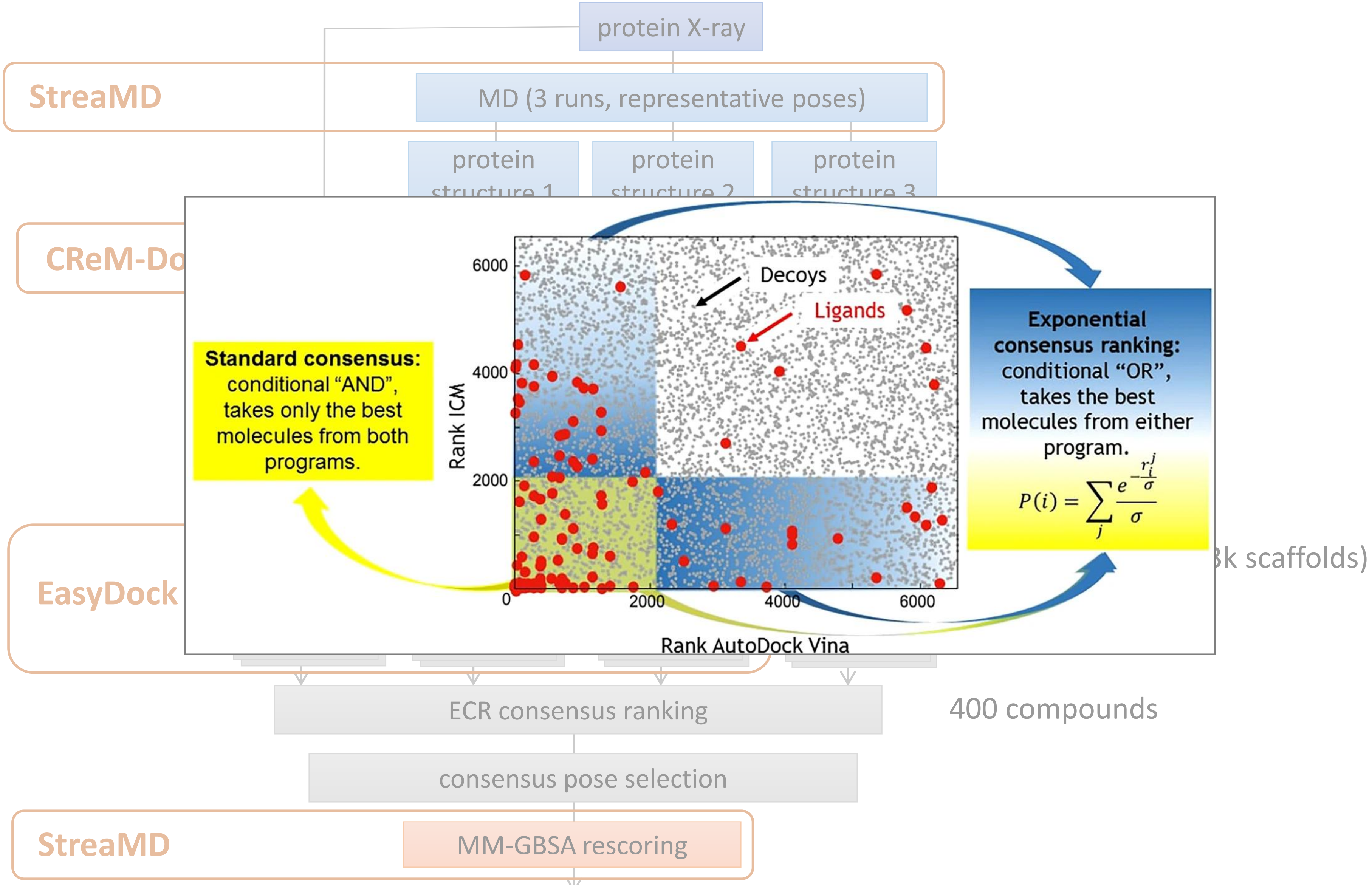
StreaMD

MM-GBSA rescoring

final ranks

50 compounds with distinct scaffolds

Round 1: strategy 1 (de novo design)



protein X-ray

StreaMD

MD (3 runs, representative poses)

protein structure 1

protein structure 2

protein structure 3

CReM-Do

EasyDock

k scaffolds)

ECR consensus ranking

400 compounds

consensus pose selection

StreaMD

MM-GBSA rescoring

final ranks

50 compounds with distinct scaffolds

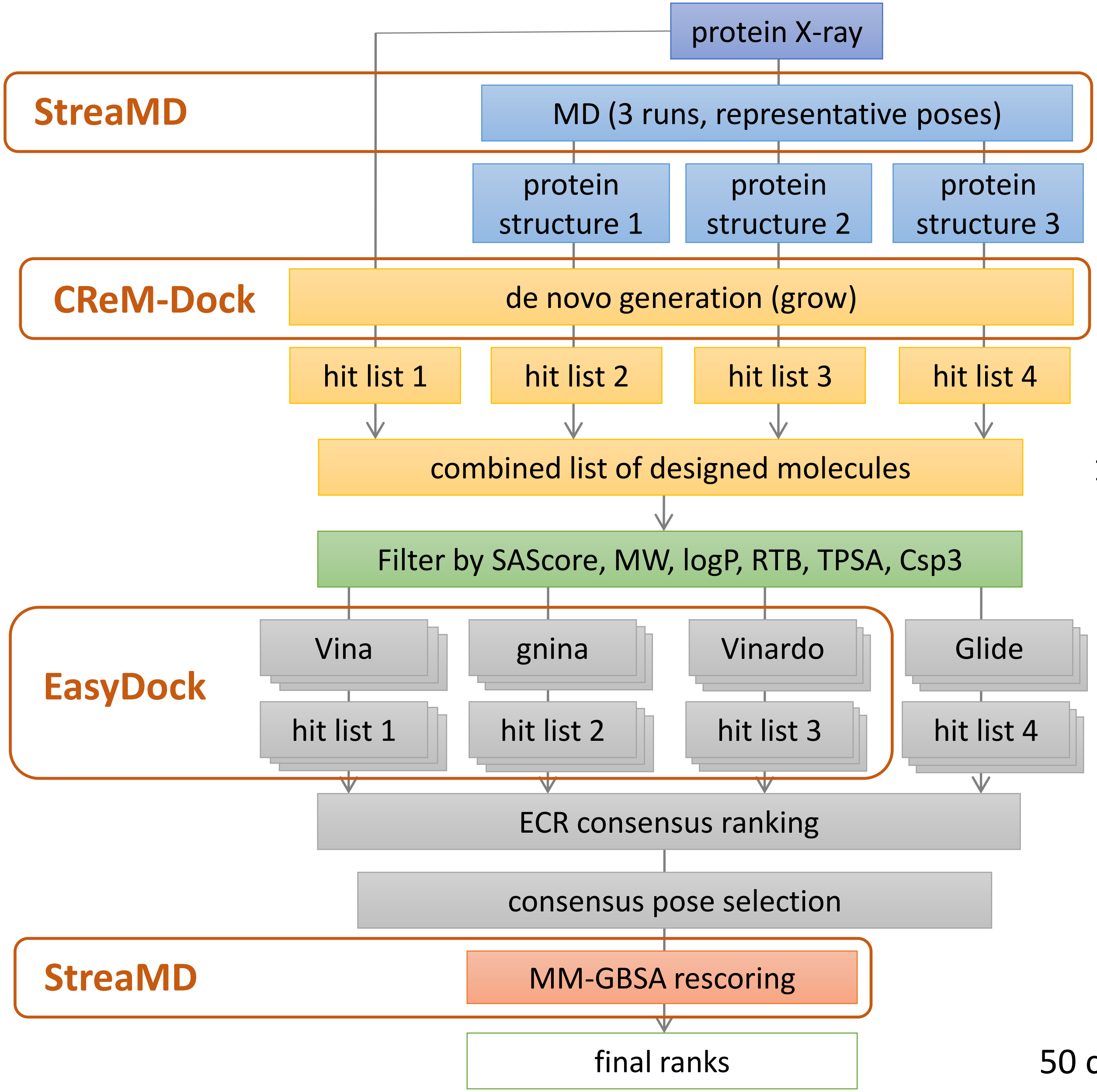
Standard consensus: conditional "AND", takes only the best molecules from both programs.

Exponential consensus ranking: conditional "OR", takes the best molecules from either program.

$$P(i) = \sum_j \frac{e^{-\frac{r_i^j}{\sigma}}}{\sigma}$$

Palacio-Rodríguez, K.; Lans, I.; Cavasotto, C. N.; Cossio, P. Exponential consensus ranking improves the outcome in docking and receptor ensemble docking. *Scientific Reports* **2019**, 9 (1), 5142.

Round 1: strategy 1 (de novo design)



protein X-ray

StreaMD

MD (3 runs, representative poses)

protein structure 1

protein structure 2

protein structure 3

CReM-Dock

de novo generation (grow)

hit list 1

hit list 2

hit list 3

hit list 4

combined list of designed molecules

Filter by SAScore, MW, logP, RTB, TPSA, Csp3

EasyDock

Vina

gnina

Vinardo

Glide

hit list 1

hit list 2

hit list 3

hit list 4

ECR consensus ranking

consensus pose selection

StreaMD

MM-GBSA rescoring

final ranks

1M compounds

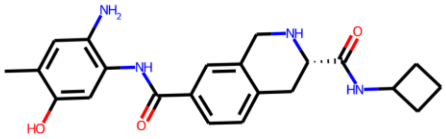
267k compounds

3.5k compounds (1.3k scaffolds)

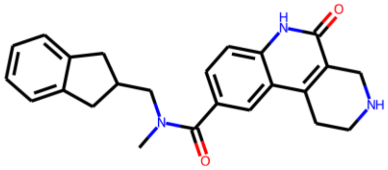
400 compounds

50 compounds with distinct scaffolds

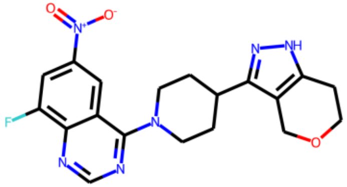
Round 1: strategy 1 (de novo design)



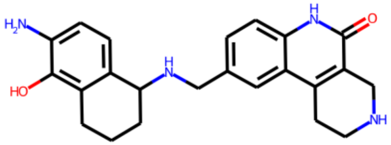
CREM0402551



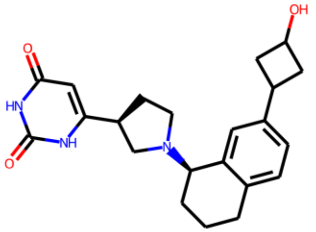
CREM0978670



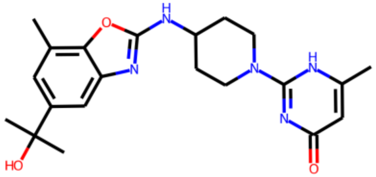
CREM1515848



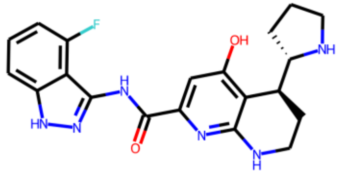
CREM1480106



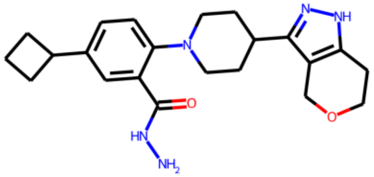
CREM1777121



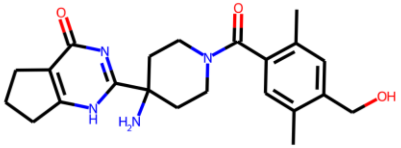
CREM0329741



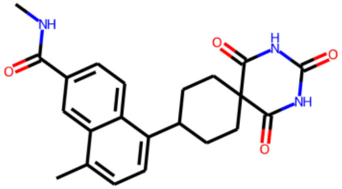
CREM1661038



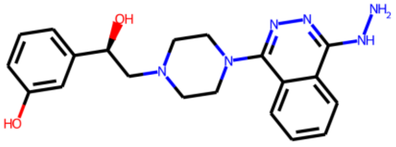
CREM1506273



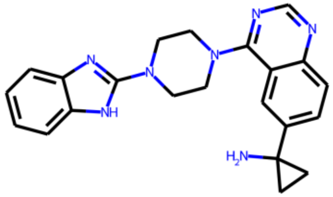
CREM0340409



CREM1089720



CREM1507777



CREM1468894

50 de novo compounds

SA score < 3

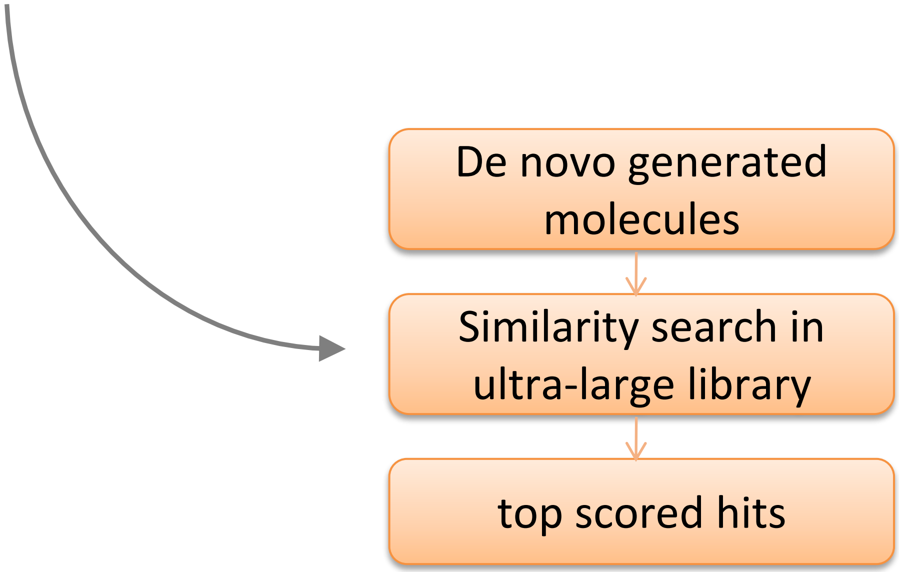
11 reconstructed retrosynthetic pathways with AiZynthFinder (2-5 steps)

Round 1: strategy 2 (similarity search)

Enamine Real Space: 16B

Docking of a whole ultra-large library (>10 B compounds) is extremely expensive

(if one docking takes 1 sec, it will take 317 years on a single core)

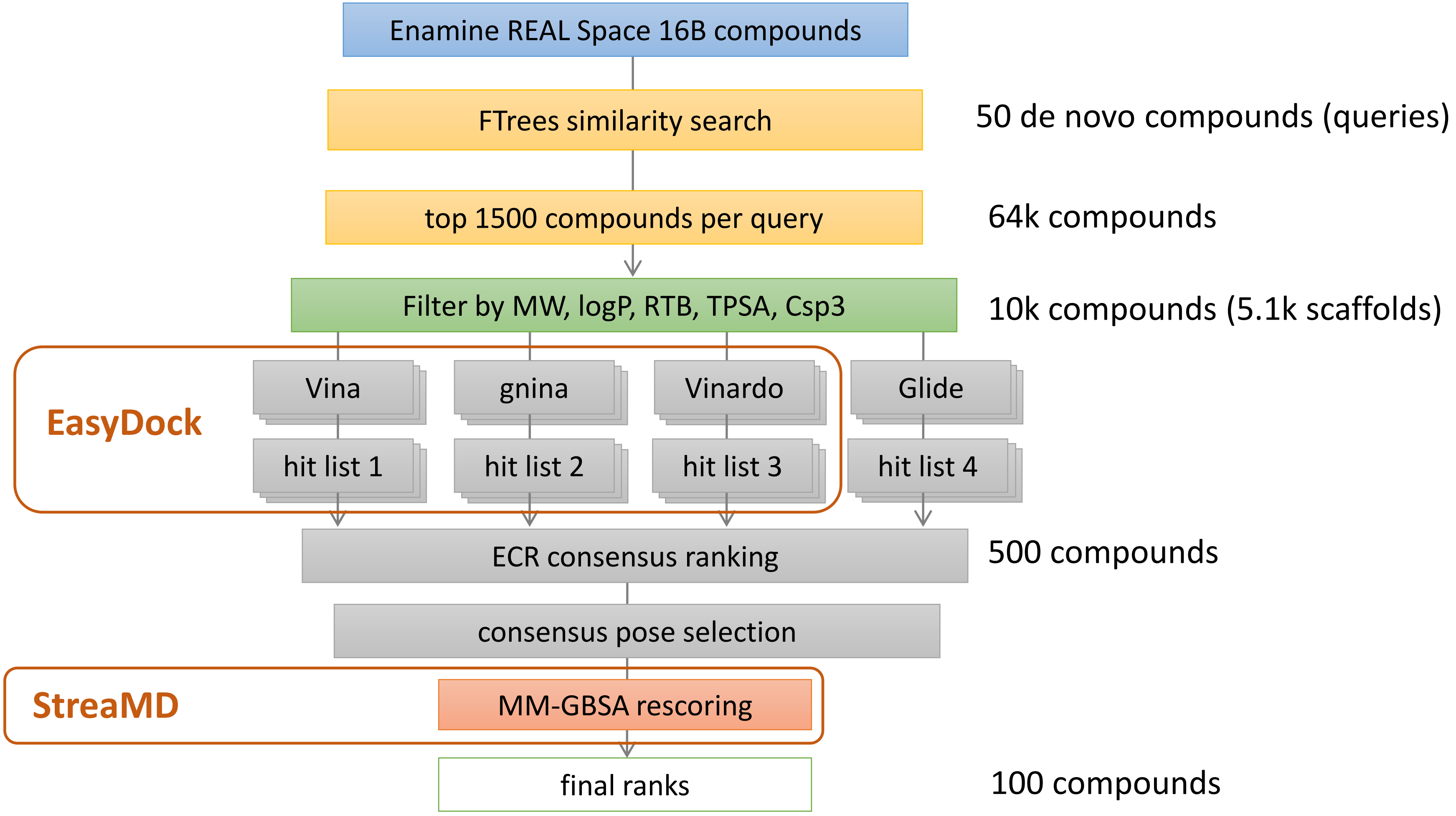


De novo generated molecules

Similarity search in ultra-large library

top scored hits

Round 1: strategy 2 (similarity search)



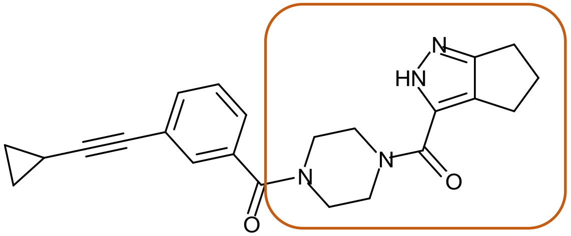
Round 1: experimental results

50 de novo + 100 similar compounds

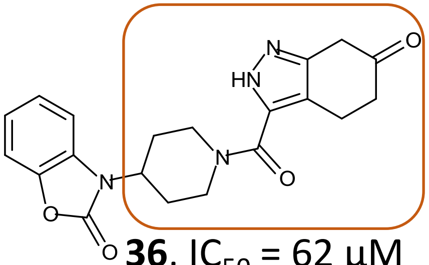
91 compounds were selected (within the budget 9000$)

82 compounds were synthesized

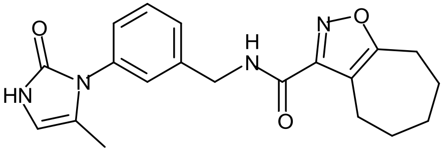
8 compounds demonstrated activity (K_d = 25-117 μM by SPR)



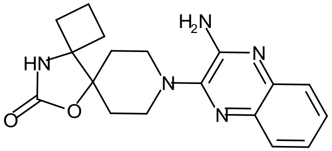
1, IC_{50} = 61 μM



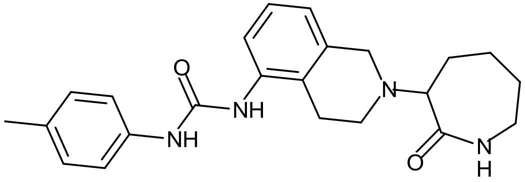
36, IC_{50} = 62 μM



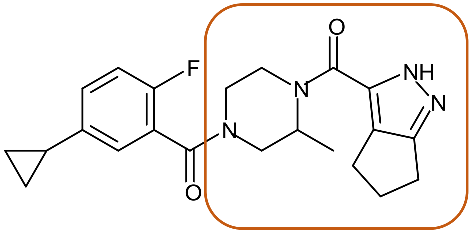
59, IC_{50} = 32 μM



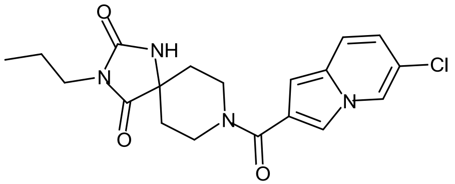
62, IC_{50} = 25 μM



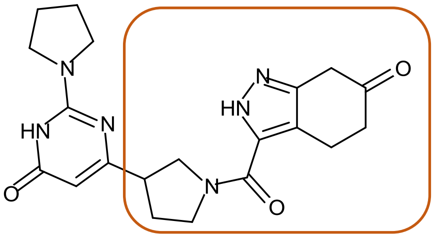
65, IC_{50} = 56 μM



69, IC_{50} = 117 μM

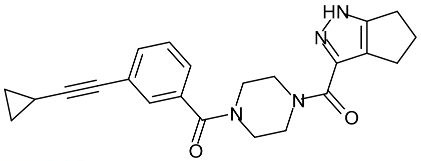


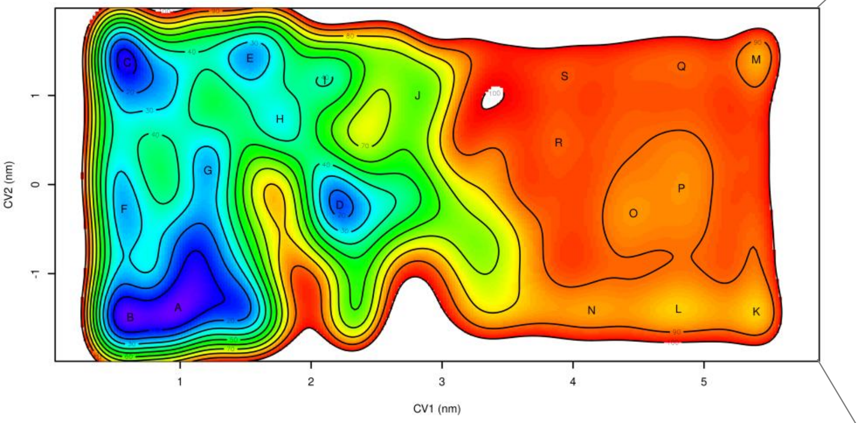
73, IC_{50} = 31 μM

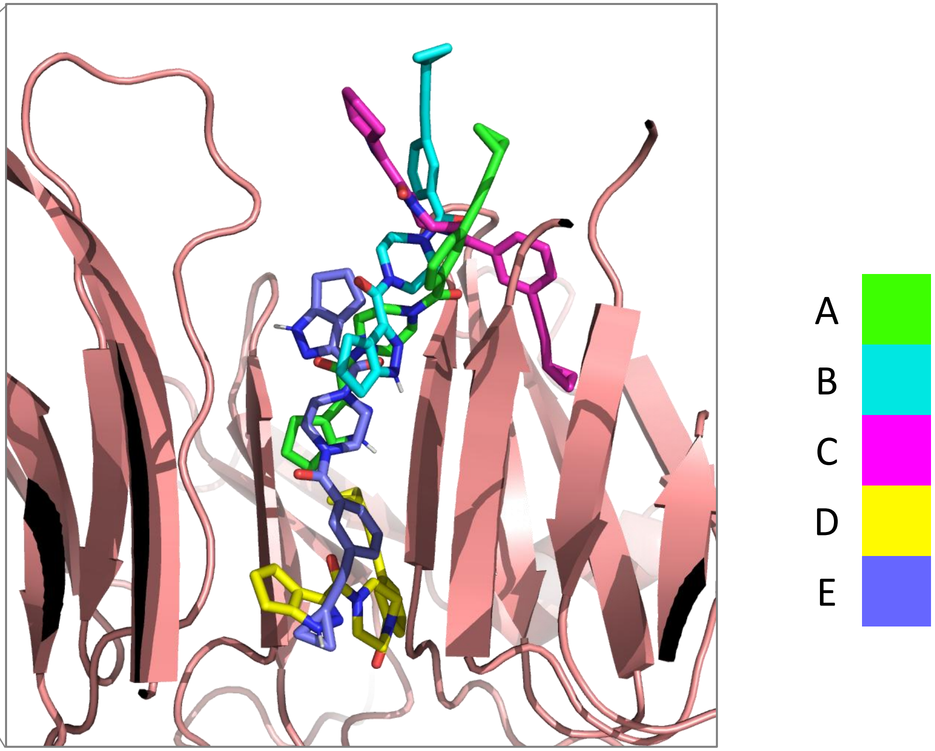


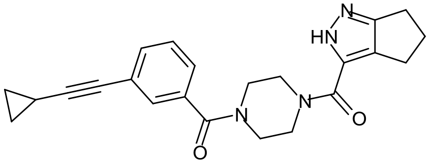
76, IC_{50} = 74 μM

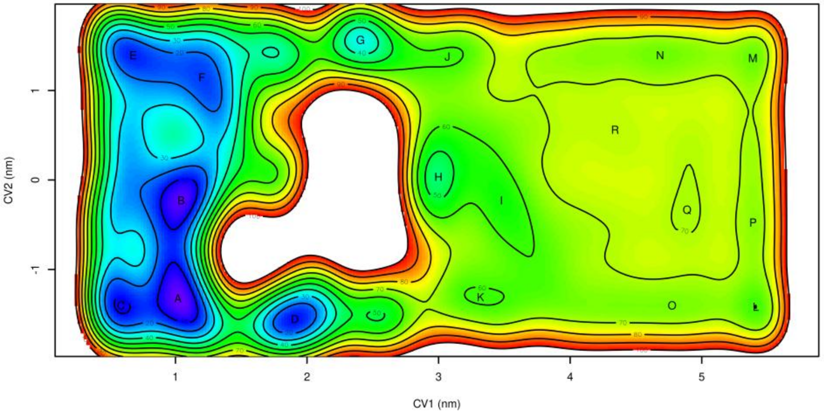
Round 2: hit optimization (metadynamics)

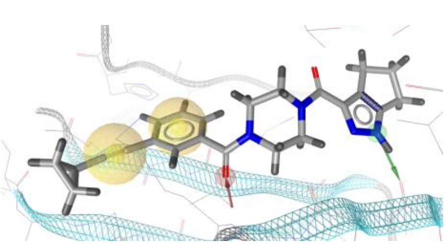


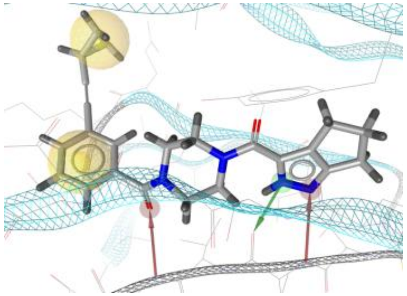






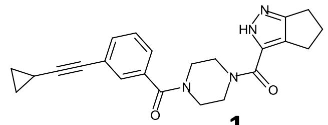


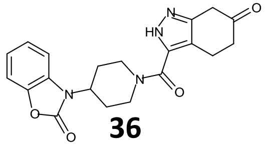


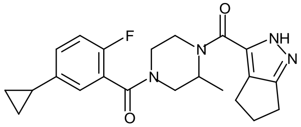


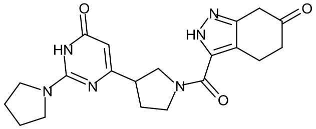
Round 2: hit optimization (compound pool 1)

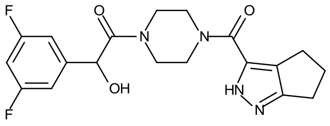
actives

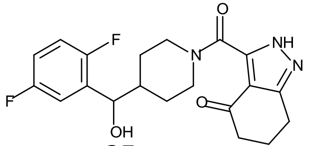

1


36

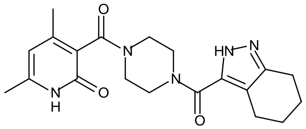

69

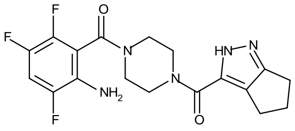

76

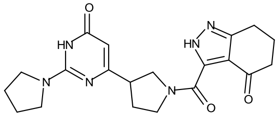

24

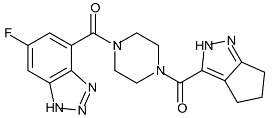

35

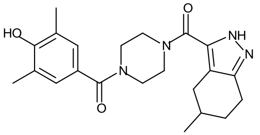
inactives


7

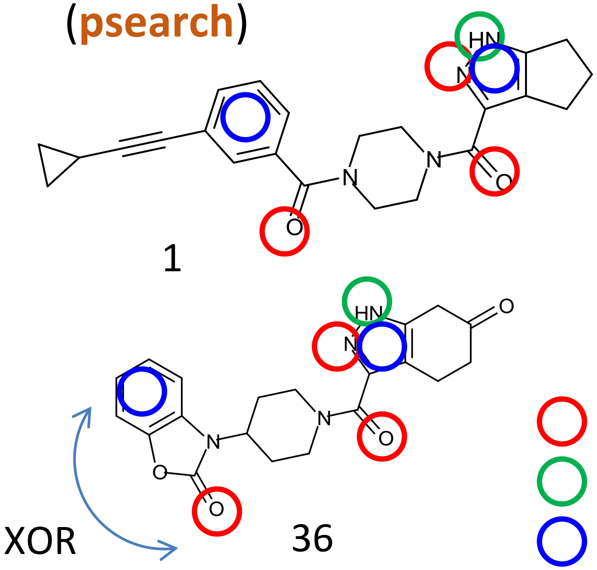

77


79


82


87

3D ligand-based pharmacophores (**psearch**)



1

36

precision: 0.43-0.5
recall: 0.75
EF: 7.2-8.4

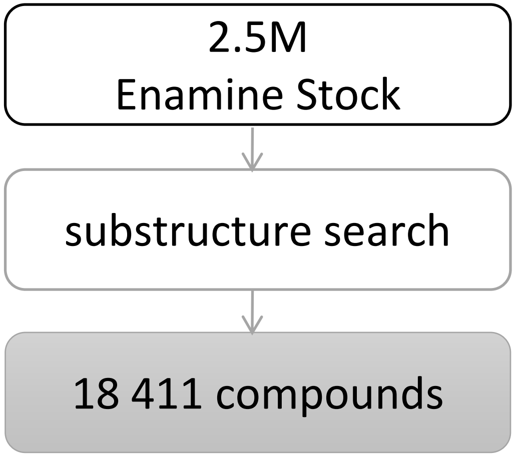
○ H-bond acceptor
○ H-bond donor
○ aromatic/hydrophobic

2.5M
Enamine Stock

the most restrictive pharmacophore model

155 compounds

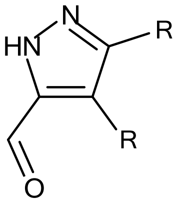
Round 2: hit optimization (compound pool 2)

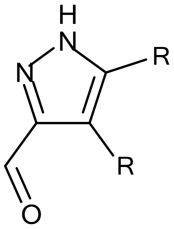


2.5M
Enamine Stock

substructure search

18 411 compounds





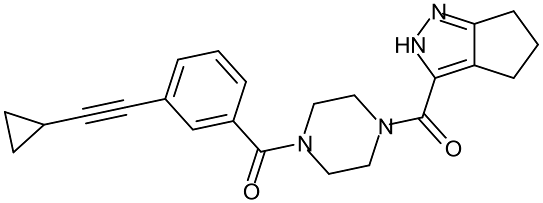
chemicalite-scripts

Round 2: hit optimization (compound pool 3)

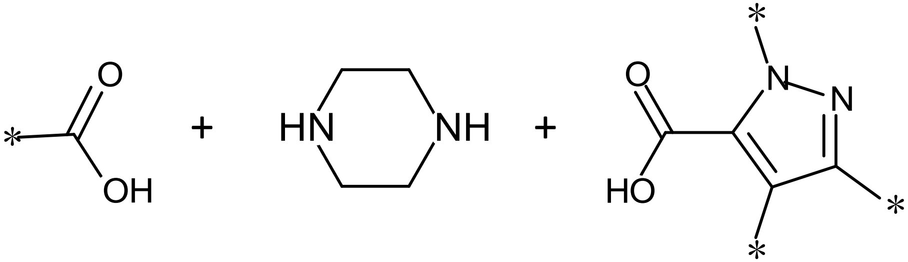
Enamine fragments

substructure search

18 845 building blocks



1, IC_{50} = 61 µM



Enamine fragments

substructure search

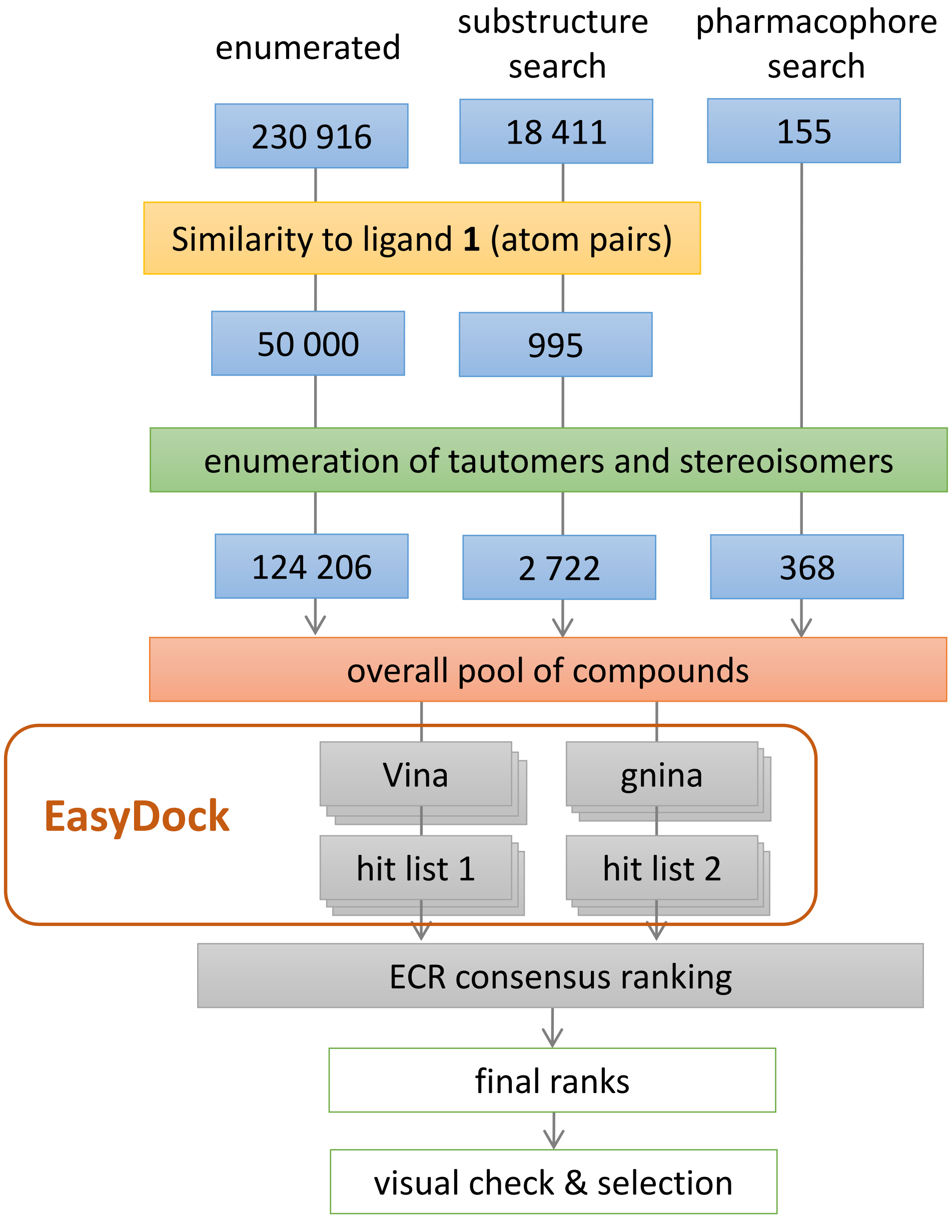
474 building blocks

2 943 486 enumerated molecules

Filter by MW, logP, TPSA, RTB, Csp3

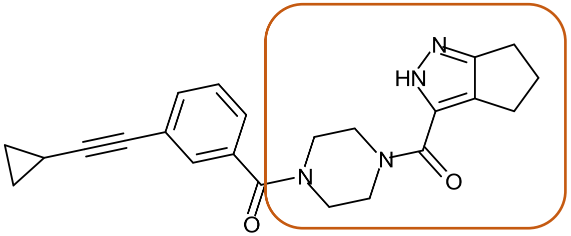
230 916 compounds

Round 2: hit optimization (screening pipeline)

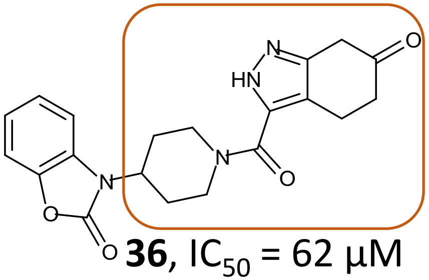


Round 2: hit optimization (experimental results)

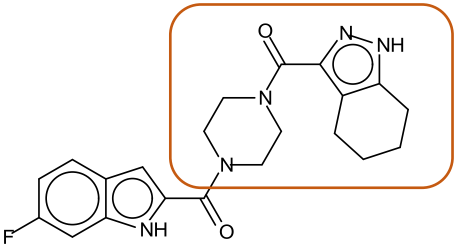
38 compounds were selected (within the budget 4500$)
35 compounds were synthesized
4 compounds demonstrated dose-response effect in SPR
1 scaffold had confirmed selectivity



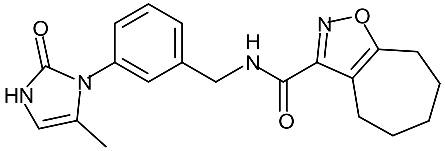
1, IC_{50} = 61 µM



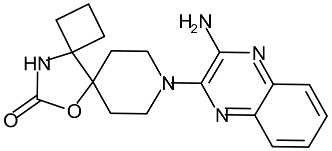
36, IC_{50} = 62 µM



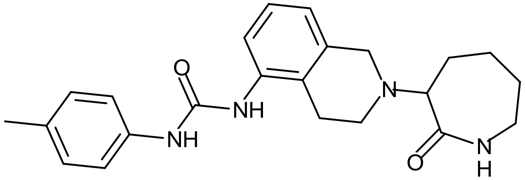
HO-15, IC_{50} = 71 µM



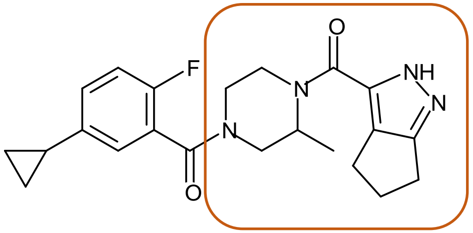
59, IC_{50} = 32 µM



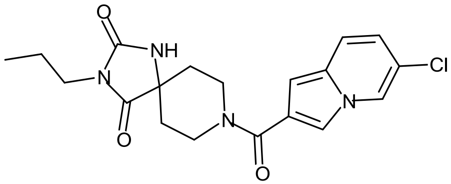
62, IC_{50} = 25 µM



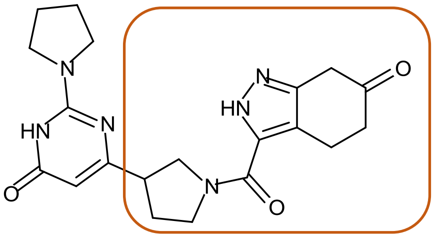
65, IC_{50} = 56 µM



69, IC_{50} = 117 µM



73, IC_{50} = 31 µM



76, IC_{50} = 74 µM

IMTM

Overall statistics of all groups

Round1 compounds	Round1 hits	Round2 compounds	Round2 SPR hits	Selective scaffolds confirmed in orthogonal methods	Promising chemical series
72	4	23	3	2	1
84	2	33	10	2	1
84	10	44	9	1	1
82	8	35	4	1	1
59	7	37	11	1	1
94	5	32	8	1	
92	4	39	6	1	
113	3	49	6	1	1
37	2	47	7	1	1
101	1	38	5	1	
98	3	46	4	0-2	
99	11	47	3	0	
100	4	49	3	0	
100	2	41	8	0	
105	2	25	1	0	
65	2	44	4	0	
91	2	36	4	0	
101	1	49	4	0	
79	0	0	0	0	
95	0	0	0	0	
71	0	0	0	0	
83	0	0	0	0	
50	0	0	0	0	

Conclusions

1. You should always have plan B, C, D...
2. Unbiased *in silico* hit selection works (hit rate at Round 1 was almost 10%)
3. The proposed strategy to search for hits in ultra-large libraries using similarity search guided by de novo designed compounds works
4. The designed multi-step virtual screening pipeline which includes docking to multiple apo-protein structures, consensus scoring and re-scoring using MM-GBSA approach also works
5. At the Round 2 we used a simplified screening strategy, however, still found a confirmed hit which belongs to the interesting chemical series according to evaluation of the organizer committee.
6. This project accelerated the development of new tools for automated docking (EasyDock) and molecular dynamics (StreaMD) which run on supercomputers. It allowed validate our de novo generation approach (CReM-Dock) and 3D ligand-based pharmacophore modeling tool (psearch) and FTrees tool for similarity search in large databases provided by BioSolvIT company.

EasyDock

Features:

1. User-friendly CLI application: input SMILES - output SQLite database (no issues with PDB/PDBQT conversion)
2. Support of Vina, Smina and Gnina, but can be easily extended to other programs
3. Support of docking of boron-containing compounds
4. Almost linear scalability over a cluster using Dask library

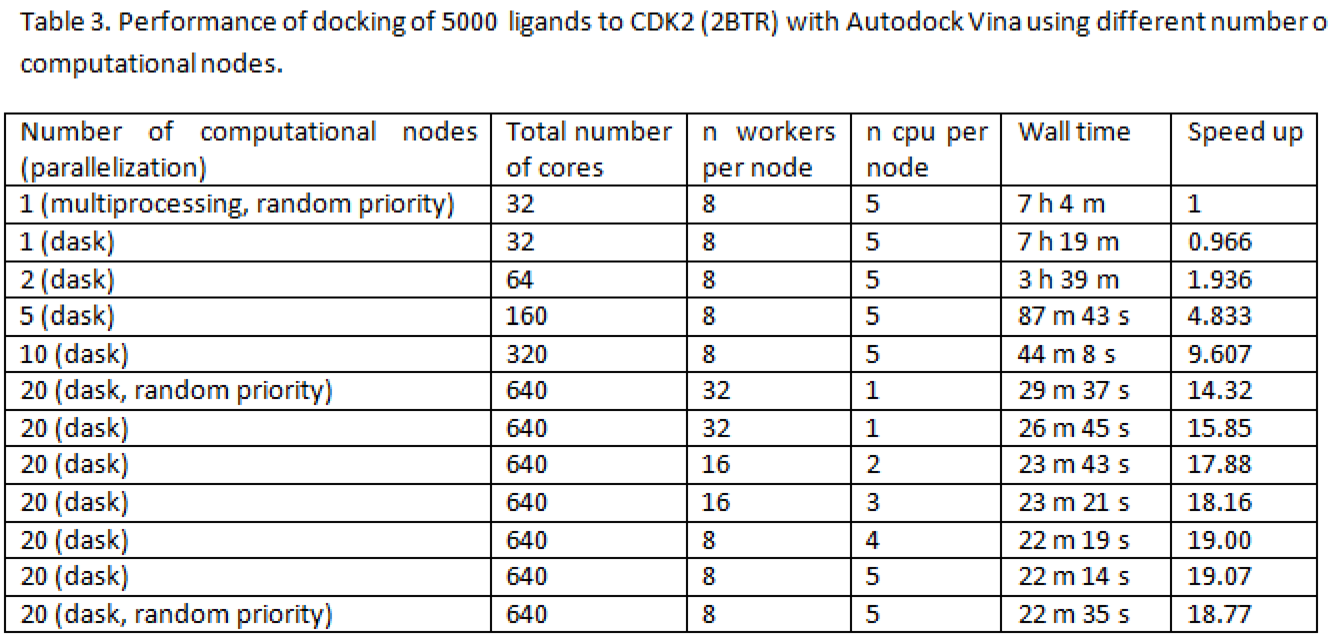


Table 3. Performance of docking of 5000 ligands to CDK2 (2BTR) with Autodock Vina using different number of computational nodes.

Number of computational nodes (parallelization)	Total number of cores	n workers per node	n cpu per node	Wall time	Speed up
1 (multiprocessing, random priority)	32	8	5	7 h 4 m	1
1 (dask)	32	8	5	7 h 19 m	0.966
2 (dask)	64	8	5	3 h 39 m	1.936
5 (dask)	160	8	5	87 m 43 s	4.833
10 (dask)	320	8	5	44 m 8 s	9.607
20 (dask, random priority)	640	32	1	29 m 37 s	14.32
20 (dask)	640	32	1	26 m 45 s	15.85
20 (dask)	640	16	2	23 m 43 s	17.88
20 (dask)	640	16	3	23 m 21 s	18.16
20 (dask)	640	8	4	22 m 19 s	19.00
20 (dask)	640	8	5	22 m 14 s	19.07
20 (dask, random priority)	640	8	5	22 m 35 s	18.77

https://github.com/ci-lab-cz/easydock

StreaMD

Features:

1. Molecular dynamic simulation for different systems:
   a) protein in water;
   b) protein - ligand;
   c) protein - cofactor (multiple);
   d) protein - ligand - cofactor (multiple);
2. Simulations of boron-containing molecules using Gaussian
3. Distributed computing using Dask library
4. Ability to extend time of MD simulations
5. Easy to continue an interrupted simulations by simply invoking the same command
6. Integrated support of end-state free energy calculations (gmx_MMPBSA) and protein-ligand interaction analysis (ProLIF)

https://github.com/ci-lab-cz/md-scripts

IMTM

Software

De novo design

CReM - Python module for structure generation https://github.com/DrrDom/crem

CReM-Dock – automated de novo generation guided by docking (not publicly available)

3D pharmacophore modeling

psearch – automated 3D ligand-based modeling and screening https://github.com/meddwl/psearch

Automated pipelines

easydock – Python module to run automatic molecular docking using vina, smina and gnina across multiple servers (cluster) https://github.com/ci-lab-cz/easydock

StreaMD – automated pipeline for high-throughput MD simulations https://github.com/ci-lab-cz/md-scripts

Auxiliary RDKit repositories

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

chemicalite-scripts - scripts to create local databases for similarity and substructure search using RDKit and Chemicalite https://github.com/DrrDom/chemicalite-scripts

Third-party software

FTrees – similarity search in Enamine REAL Space (BioSolveIT)

GROMACS – molecular dynamic simulations

R/RStudio – programming language and IDE for data analysis