CReM-dock: de novo design of synthetically feasible compounds guided by molecular docking

Guzel Minibaeva[1], Haolin Du[2], Finlay Clark,[3] Julien Michel,[2] Pavel Polishchuk[1*]

[1] Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hněvotínská 1333/5, 779 00 Olomouc, Czech Republic

[2] EaStCHEM School of Chemistry, University of Edinburgh, Edinburgh, EH9 3FJ, United Kingdom

[3] School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom

pavlo.polishchuk@upol.cz

Supplementary materials

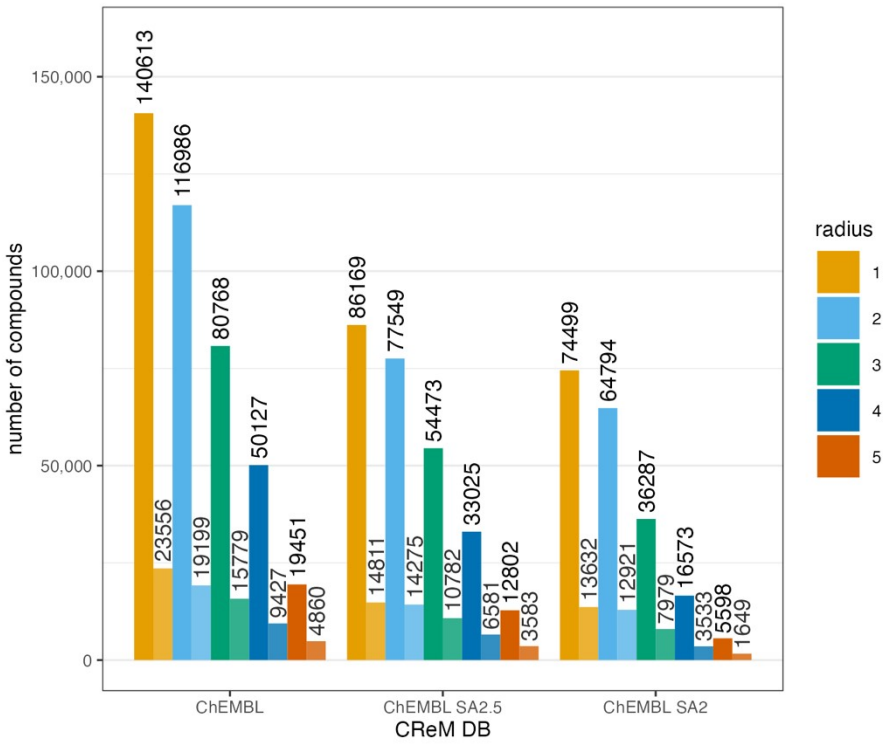


Figure S1. Average number of generated compounds for different generation settings (fragment CReM databases and context radiuses) across three runs. Darker color denotes the total number of generated compounds, lighter colors – the total number of generated compounds which satisfy the required ligand-protein interactions (H-bond donor and acceptor with Leu83).

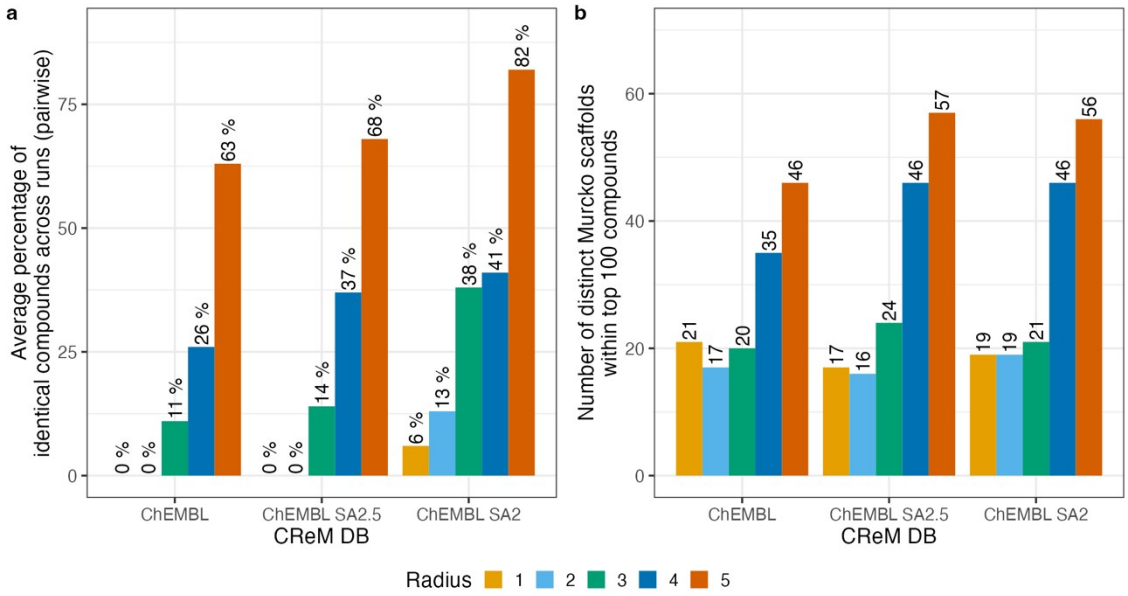


Figure S2. Statistics for top 100 compounds bound to the hinge region of CDK2. (a) Average percentage of identical structures among top 100 compounds between pairs of three independent runs for every set of generation settings. (b) The average number of distinct Murcko scaffolds among top 100 compounds across three independent runs.

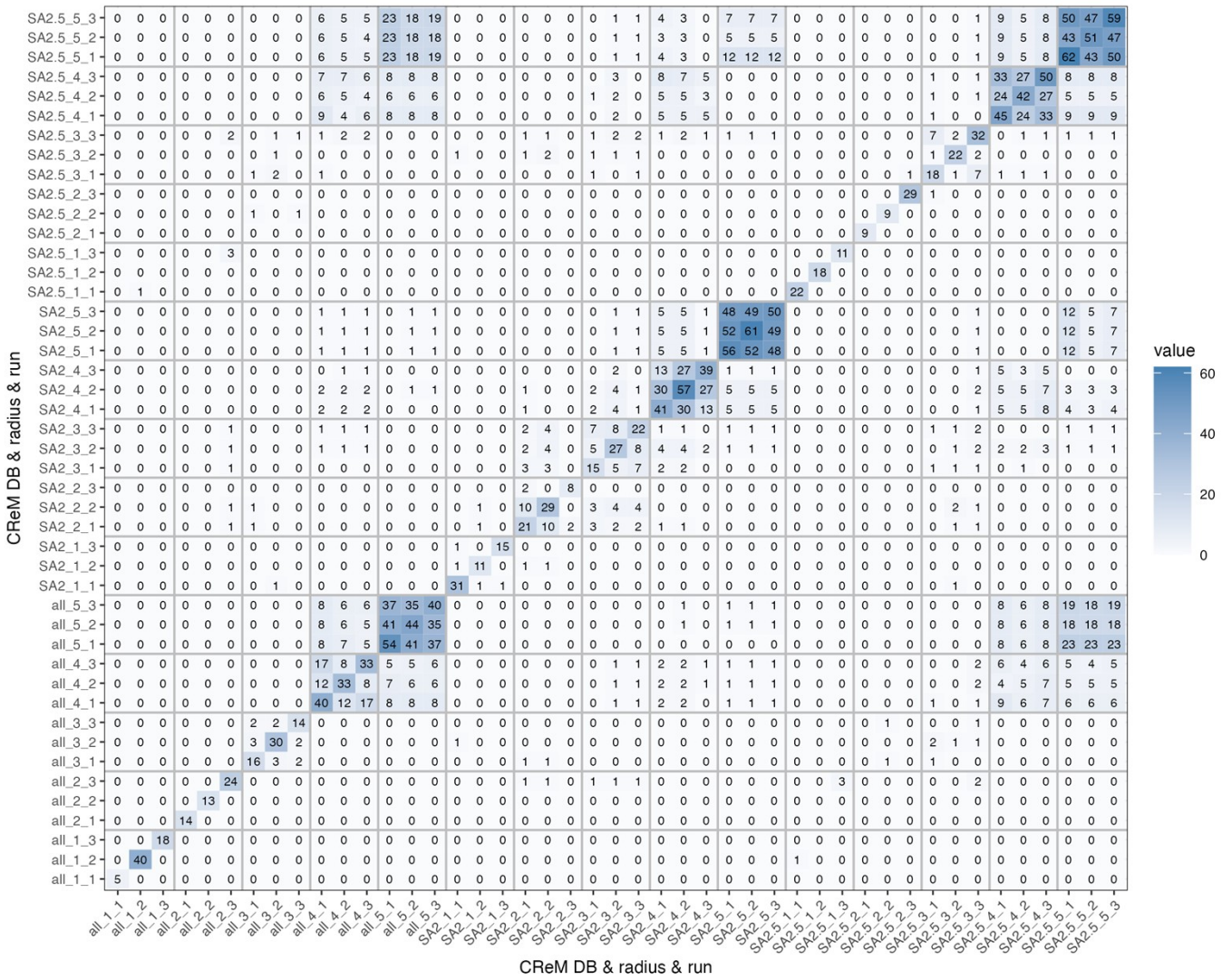


Figure S3. The number of distinct scaffolds for individual runs is in the diagonal cells. Other cells contain the number of common scaffolds between each pair of runs.

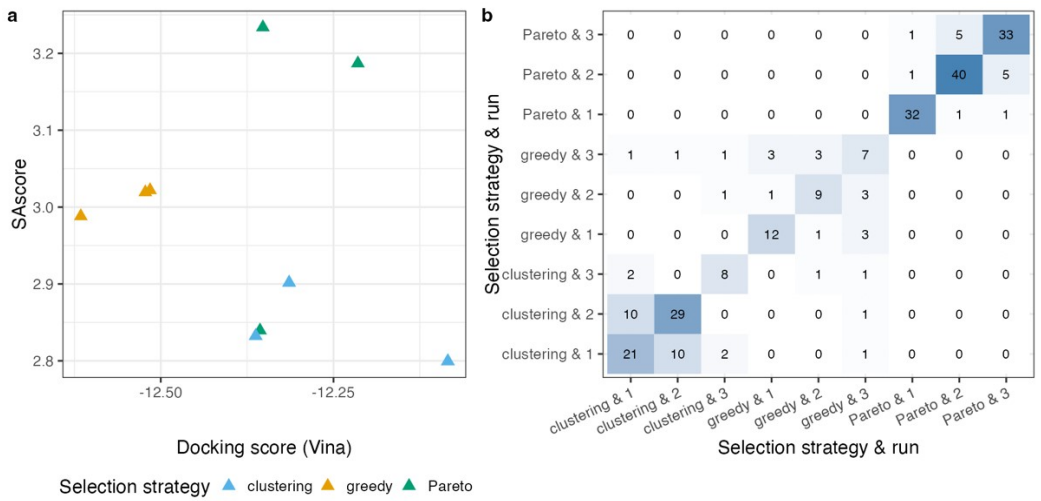


Figure S4. Statistics for top 100 molecules bound to the hinge region from three independent runs using different selection strategy (greedy, clustering and Pareto) and cluster settings (all runs used ChEMBL SA2 fragment database and radius 2). (a) Average docking and SA scores for top 100 molecules. (b) The number of distinct Murcko scaffolds among top 100 compounds for different selection strategies.

Figure S5 UMAP (the number of neighbors is 10) of top 100 compounds generated by CReM-dock and REINVENT for different targets in comparison with 50000 randomly selected ChEMBL compounds representing a reference space.

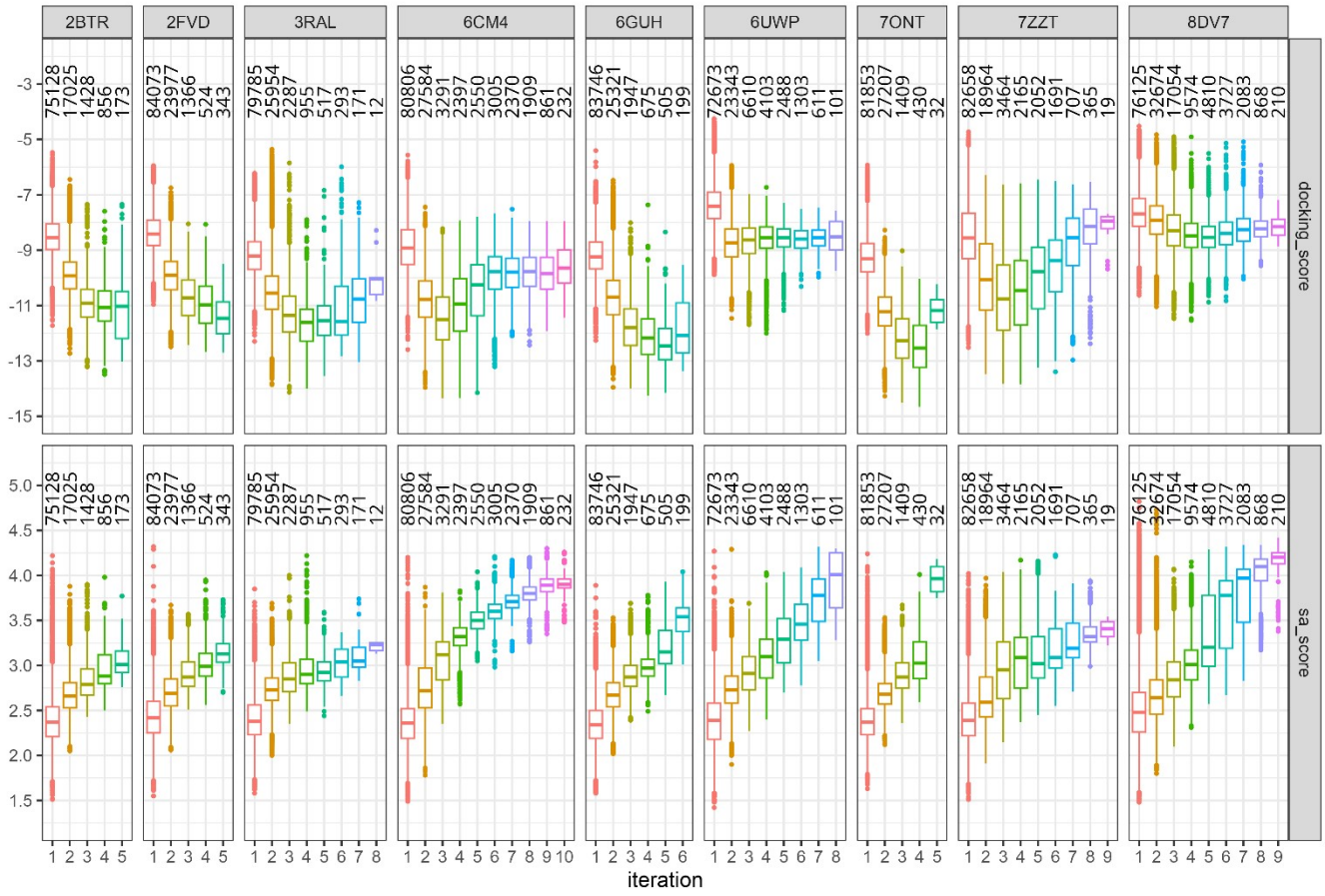


Figure S6 Distribution of docking and SA scores of all generated compounds on each iteration. The settings used for generation are the following: fragment database = ChEMBL SA2, starting fragments = ChEMBL SA2, radius = 2, selection strategy = clustering with top 2 picked compounds from each of 25 clusters, molecular mass ≤ 450 Da, the number of rotatable bonds ≤ 5, lipophilicity ≤ 4, topological polar surface area ≤ 120Å^2.

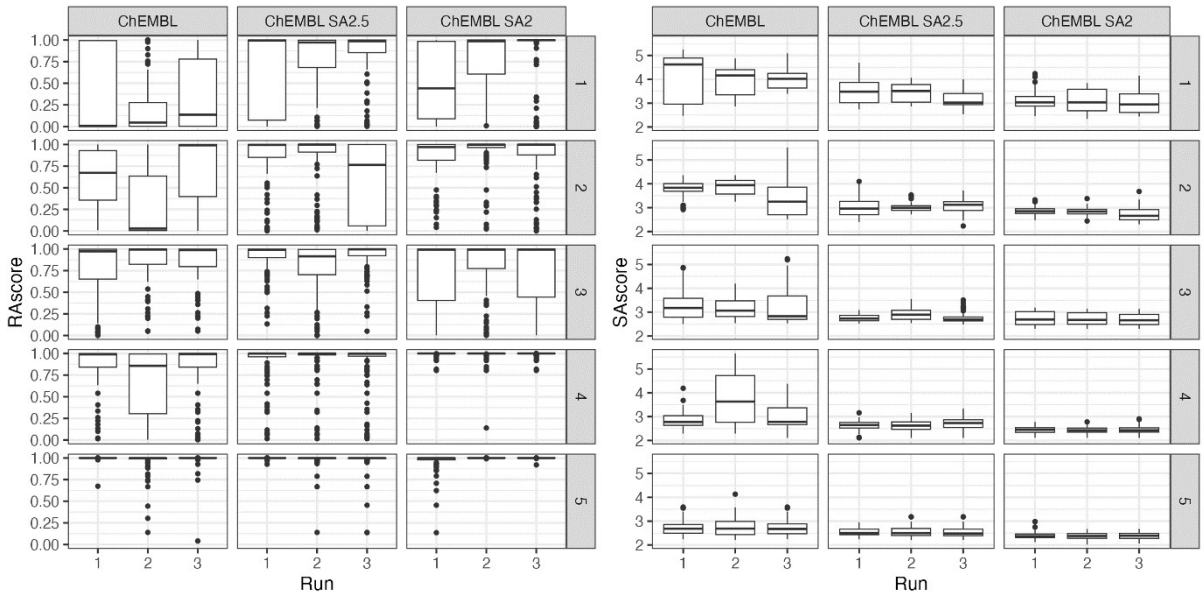


Figure S7 Distribution of RA and SA scores for top 100 structures generated by CReM-dock for 2BTR structure (CDK2) bound to the hinge region of the protein.

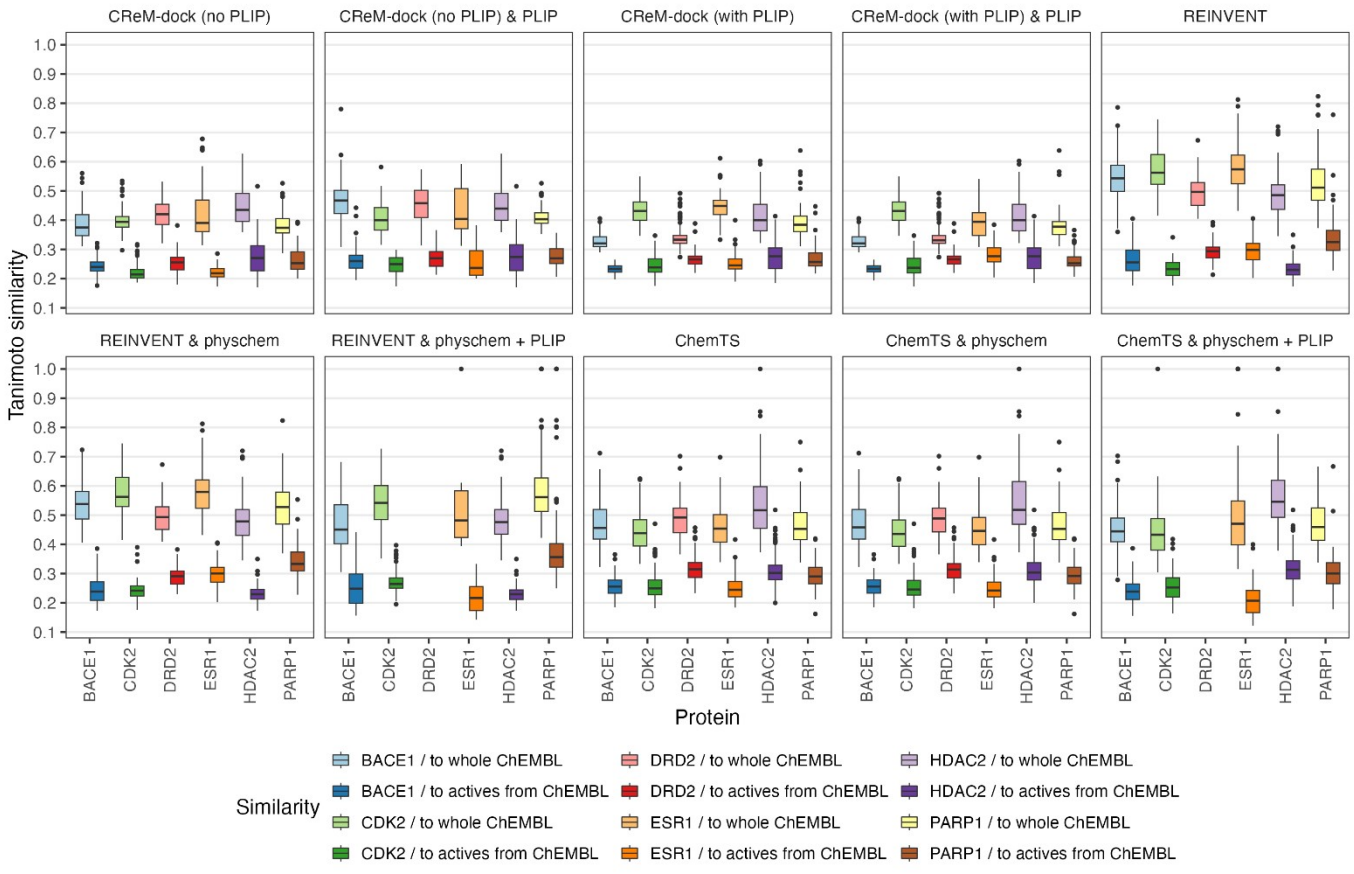


Figure S8 Novelty of top 100 structures by docking scores generated by CReM-dock, REINVENT and ChemTS (the lower the better).

Table S1. The number of generated compounds using different selection strategies and the number of compounds bind to the hinge region.

	selection_strategy	run	n	n_plif	%_plif
1	Pareto	1	121568	23106	0.190
2	Pareto	2	108841	23149	0.213
3	Pareto	3	148579	37712	0.254
4	clustering	1	89530	15002	0.168
5	clustering	2	87341	13241	0.152
6	clustering	3	89032	15145	0.170
7	greedy	1	90098	18294	0.203
8	greedy	2	90073	17680	0.196
9	greedy	3	91573	18384	0.201

Table S2. Proteins chosen for the studies, the associated important protein-ligand interaction patterns, the total number of active ligands extracted from ChEMBL33 and the median docking score of these actives.

Kinase	PDB code	Important protein-ligand interaction patterns (PLIP)	Number of known actives from ChEMBL33 (pIC_{50}/pK_i ≥ 6) and MW ≤ 500	Median docking score of actives
CDK1	6GU2	Glu84 (HB donor); Leu86 (HB acceptor); Leu86 (HB donor)	742	-8.67
CDK2	2BTR	Glu81(HB donor); Leu83 (HB acceptor); Leu83 (HB donor)	1001	-8.9
	2FVD	Glu81(HB donor); Leu83 (HB acceptor); Leu83 (HB donor)		
	3RAL	Glu81(HB donor); Leu83 (HB acceptor); Leu83 (HB donor)		
	6GUH	Leu85 (HB acceptor); Leu85 (HB donor)		
CDK5	4AU8	Cys83 (HB acceptor); Cys83 (HB donor); Glu81 (HB donor)	304	-9.06
CDK6	6OQO	Val101 (HB acceptor); Val101 (HB donor); Glu99 (HB donor)	301	-10.16
CDK7	8P4Z	Met94 (HB acceptor); Met94 (HB donor); Asp92(HB donor)	222	-8.54
CDK16	5G6V	Leu243 (HB acceptor); Leu243 (HB donor); Glu241 (HB donor)	23	-8.77
MAPK7	5BYZ	Met140 (HB acceptor); Met140 (HB donor); Asp138 (HB donor)	73	-8.8
MAPK13	5EKO	Met110 (HB acceptor); Met110 (HB donor); Pro108 (HB donor)	80	-8.01
BACE1	6UWP	Asp228 (HB donor); Asp32 (HB donor)	4067	-9.11
DRD2	6CM4	Asp114 (cationic)	5109	-9.98
ESR1	8DV7	Glu353 (HB donor); Arg394 (HB acceptor); His524 (HB acceptor)	1567	-9.29
HDAC2	7ZZT	Zn601 (metal acceptor)	950	-8.89
PARP1	7ONT	Gly863 (HB donor); Gly863 (HB acceptor)	2758	-10.32

Table S3. Important protein-ligand contacts observed in 2HB1, 2ZWZ, 3S1G complexes and encoded as required in the fragment expansion study.

PDB	PLIP
2HB1	Lys120 (HB acceptor), Arg221 (anionic), Gln266 (HB acceptor), Phe182 (HB acceptor)
3S1G	Ala232 (HB donor), Asp156 (HB donor), Gln203 (HB acceptor), Gly230 (HB acceptor), Leu231 (HB donor)
2ZWZ	Asp224 (HB donor), Glu266 (HB donor), Glu66 (HB donor), Hie128 (HB acceptor), Hip34 (HB acceptor), Trp67 (HB acceptor)

Table S4. The total number of structures generated by CReM-dock (no PLIP) on each iteration and the number of those structures satisfying given PLIP patterns.

protein	iteration	n	n (PLIP > 0.6)	fraction of structures with PLIP > 0.6, %
BACE1	0	23840	86	0.36
BACE1	1	44314	293	0.66
BACE1	2	23806	38	0.16
BACE1	3	1577	1	0.06
BACE1	4	574	0	0
BACE1	5	23	0	0
CDK2	0	23840	1542	6.47
CDK2	1	43851	2236	5.1
CDK2	2	22089	1103	4.99
CDK2	3	1332	29	2.18
CDK2	4	408	1	0.25
CDK2	5	346	0	0
DRD2	0	23840	1136	4.77
DRD2	1	44593	2631	5.9
DRD2	2	21957	638	2.91
DRD2	3	2187	46	2.1
DRD2	4	1342	19	1.42
DRD2	5	869	26	2.99
DRD2	6	1031	57	5.53
DRD2	7	990	54	5.45
DRD2	8	518	9	1.74
DRD2	9	15	0	0
ESR1	0	23840	158	0.66
ESR1	1	43912	12	0.03
ESR1	2	20233	9	0.04

protein	iteration	n	n (PLIP > 0.6)	fraction of structures with PLIP > 0.6, %
ESR1	3	1936	1	0.05
ESR1	4	1335	2	0.15
ESR1	5	1529	1	0.07
ESR1	6	1484	3	0.2
ESR1	7	1310	0	0
ESR1	8	809	0	0
HDAC2	0	23840	7613	31.93
HDAC2	1	47242	22474	47.57
HDAC2	2	11160	5835	52.28
HDAC2	3	1911	1231	64.42
HDAC2	4	2707	1575	58.18
HDAC2	5	2458	1416	57.61
HDAC2	6	1937	1235	63.76
HDAC2	7	1675	986	58.87
HDAC2	8	1424	690	48.46
HDAC2	9	782	316	40.41
HDAC2	10	322	107	33.23
PARP1	0	23840	1004	4.21
PARP1	1	44193	5974	13.52
PARP1	2	21715	1562	7.19
PARP1	3	1092	75	6.87
PARP1	4	561	37	6.6
PARP1	5	28	0	0

Table S5. Top scoring structures generated by CReM-dock, REINVENT4 and ChemTS. The numbers below structures are SA and docking scores, respectively.

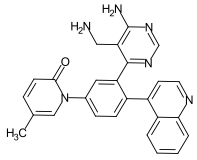
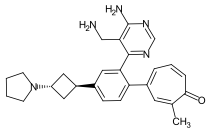
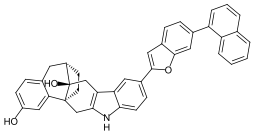
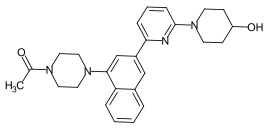
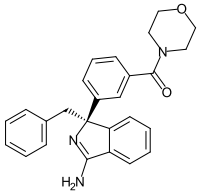
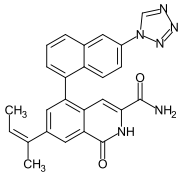
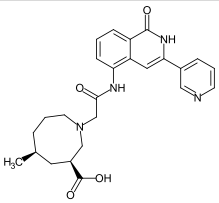
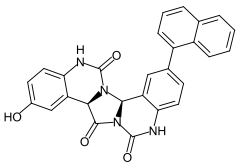
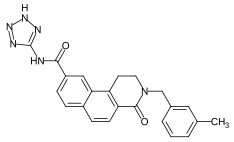
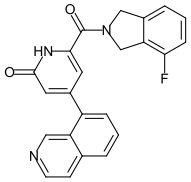
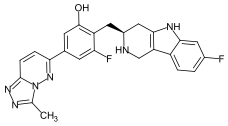
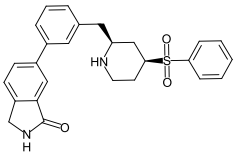
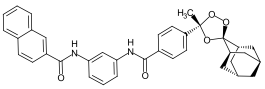
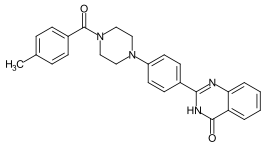
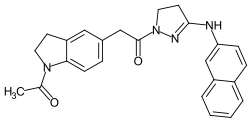
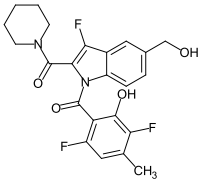
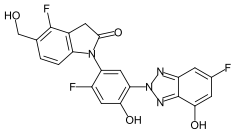
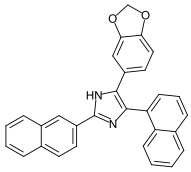
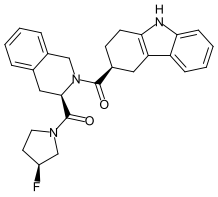
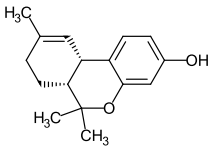
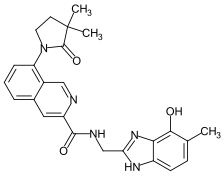
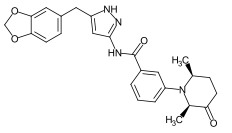
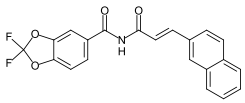
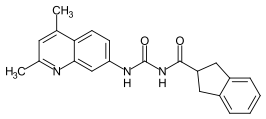
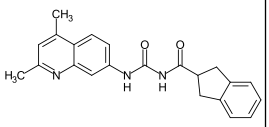
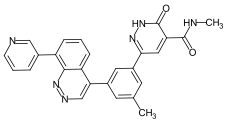
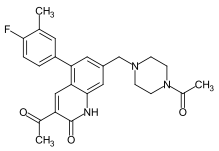
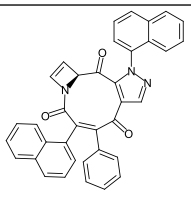
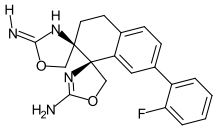
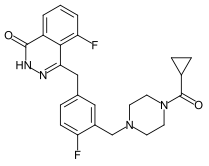
	CReM-dock (Docking)	CReM-dock (Docking + QED)	REINVENT4	REINVENT4 (physicochemical filters)	REINVENT4 (physicochemical filters + PLIP)
BACE1	 2.73; -11.0	 2.88; -10.9	 4.71; -13.8	 2.40; -12.1	 3.03; -10.0
CDK2	 2.95; -13.7	 3.22; -12.2	 3.63; -13.6	 2.52; -12.7	 2.59; -11.6
DRD2	 3.60; -14.4	 3.39; -14.0	 4.82; -15.8	 2.11; -14.0	 2.66; -13.6
ESR1	 2.99; -11.5	 3.27; -11.6	 2.35; -13.4	 3.62; -12.5	 3.47; -9.00
HDAC2	 2.94; -13.8	 3.53; -13.6	 2.41; -13.7	 2.3; -13.7	 2.3; -13.7
PARP1	 2.75; -14.6	 2.48; -13.9	 3.90; -16.4	 5.15; -13.7	 2.55; -12.8

Table S5 (continue)

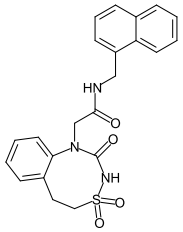
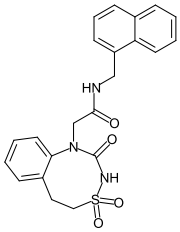
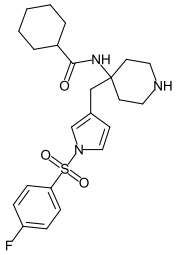
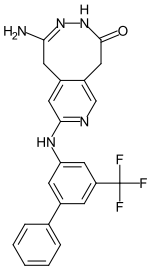
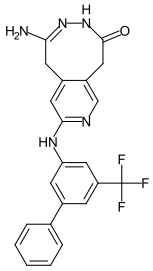
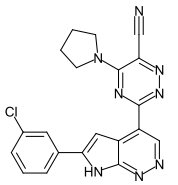
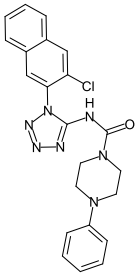
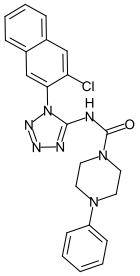
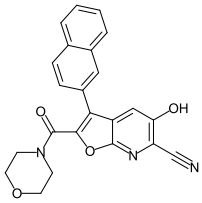
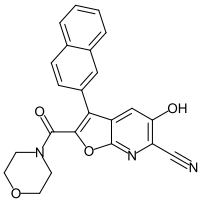
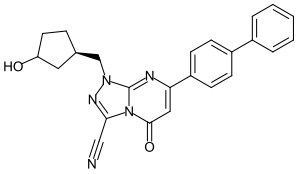
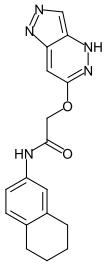
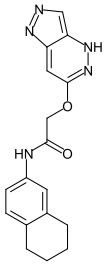
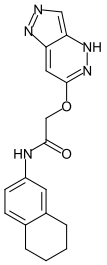
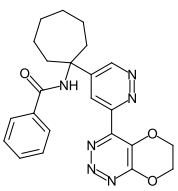
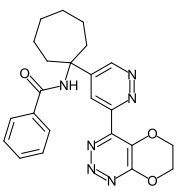
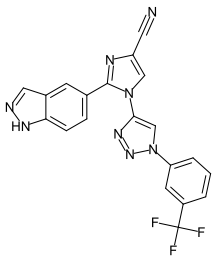
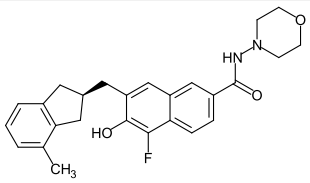
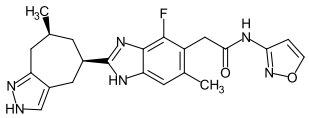
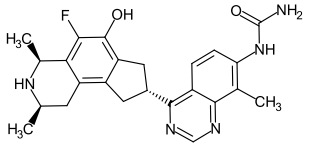
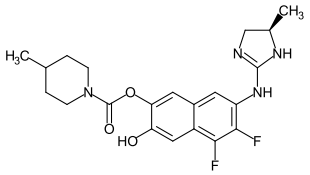
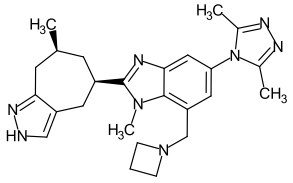
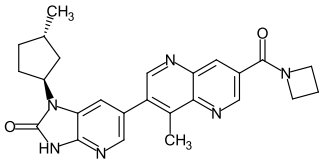
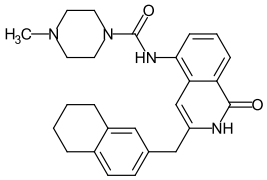
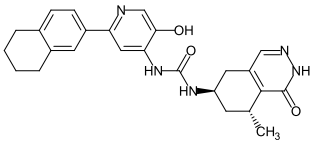
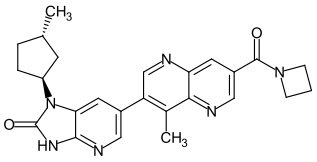
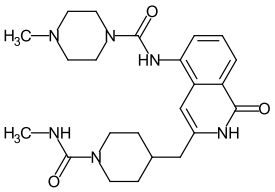
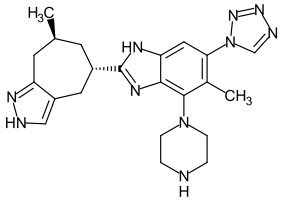
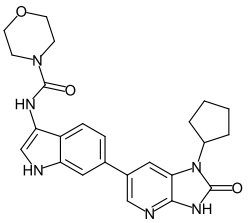
	ChemTS	ChemTS (physicochemical filters)	ChemTS (physicochemical filters + PLIP)
BACE1	 2.56; -11.2	 2.56; -11.2	 2.91; -9.31
CDK2	 2.88; 11.9	 2.88; 11.9	 3.00; -10.5
DRD2	 2.41; -13.4	 2.41; -13.4	-
ESR1	 2.80; -11.2	 2.80; -11.2	 3.49; -10.7
HDAC2	 2.70; -12.1	 2.70; -12.1	 2.70; -12.1
PARP1	 3.18; -13.2	 3.18; -13.2	 3.10; -12.0

Table S6. Top scored generated compounds across three independent runs for each combination of settings. Compounds bind to the hinge region and have the fraction of sp^3 carbon atoms in scaffolds equal or greater than 0.3.

	Starting fragments		
	SA2	SA2 Csp^3-rich	SA2.5 Csp^3-rich
docking score & no fragment sampling	 3.32; -11.7	 4.22; -11.8	 4.39; -12.3
docking score & Csp^3 fragment sampling	 3.56; -11.8	 4.19; -11.7	 3.62; -12.5
docking score + Csp^3 (BM) & no fragment sampling	 2.51; -11.6	 3.78; -12.3	 3.62; -12.5
docking score + Csp^3 (BM) & Csp^3 fragment sampling	 2.65; -11.6	 4.35; -12.0	 2.78; -12.3