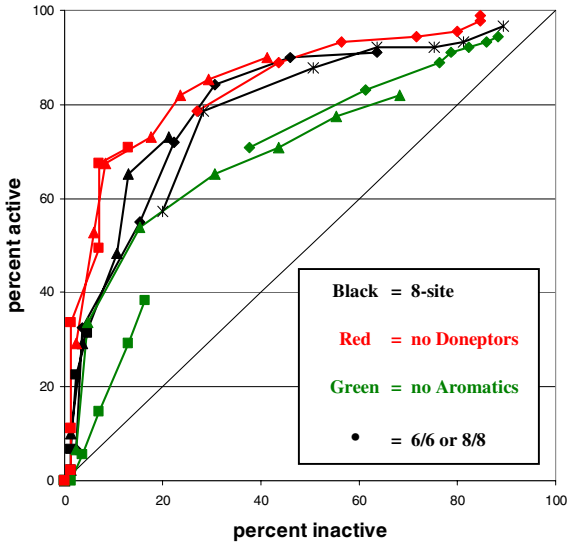
Supplemental Information

Simplification of Overlapping Pharmacophore Elements. The initial pharmacophore model of HIVp was created from a 1 ns trajectory using snapshots every 100 ps, aligned by Cα position. This resulted in an 8-site pharmacophore model in which two hydrogen-bonding elements were contained within the aromatic or aromatic/hydrophobic sites (Fig. S1A). This 8-site model was searched against the databases with moderate results. Requiring eight pharmacophore elements is a strict criterion, and we feared that it might be too specific. With radii two times the cluster RMSD only 2% of the known inhibitors were identified. Reducing the required elements to 6/8 enabled 72% of the known inhibitors to be identified with a false positive rate of 22% (Fig. S1B). Therefore, we investigated removing each of the overlapping elements in turn, resulting in two different, 6-site pharmacophore models. Removing the overlapping doneptor sites gives the most predictive model, which shows improvement over the 8-site model, predicting 73% known actives and only 18% false positives at the same radii size (Fig. S1B). However, removing the overlapping aromatic/hydrophobic sites (retaining the doneptor sites) leads to a significant loss in the predictive power of the model and a sharp increase in the number of false positives (up to 31% in Fig. S1B). Therefore, in the subsequent studies, any overlapping hydrogen-bonding sites were removed and aromatic/hydrophobic sites were retained.

A



B

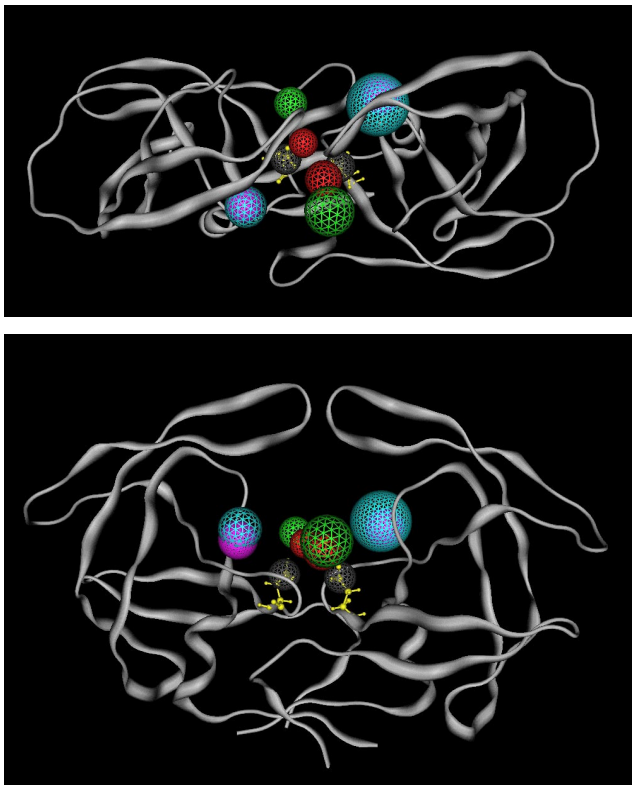


Fig. S1. **A)** ROC plots showing the performance of the 1ns, Cα 8-site model (black), the 6-site model removing the overlapping doneptor sites (red) and the 6-site model removing the overlapping aromatic/hydrophobic sites (dark green). **B)** Two views of the 8-site pharmacophore model (2xRMSD). The doneptor sites are magenta and almost entirely enveloped by the aromatic/hydrophobic sites.

The raw data for selected pharmacophore models. W is the factor by which the consensus cluster RMSD was multiplied. Data is reported as the % known active (number of known active hits / 89) and the % known inactive (number of known inactive hits / 85). Percentages are multiplied by 100 and rounded to the nearest whole number. The last column (% active / % inactive) is also calculated directly by the number of hits and any discrepancies with the first two columns are simply due to rounding the entries. Bold numbers in the first two columns note the best points on the ROC plots; the bold numbers in the third column note the models with the greatest bias against false positives (at the possible sacrifice of a good number of true positives).

Table A 1 ns, ASP alignment

W	**% Known Active**			**% Known Inactive**			**% Active / % Inactive**		
	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6
1	0	33	83	0	2	25	–	14	3
1 1/3	3	57	89	1	6	44	3	10	2
1 2/3	17	69	92	1	13	62	14	5	1
2	40	83	96	4	18	66	11	5	1
2 1/3	62	87	96	6	28	79	11	3	1
2 2/3	67	88	99	12	41	84	6	2	1
3	73	89	99	15	49	87	5	2	1

Table B 1ns, Cα alignment

W	% Known Active			% Known Inactive			% Active / % Inactive		
	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6
1	0	29	79	0	2	27	–	12	3
1 1/3	2	53	89	1	6	44	2	9	2
1 2/3	11	67	93	1	8	56	10	8	2
2	34	73	94	1	18	72	29	4	1
2 1/3	49	82	96	7	24	80	7	3	1
2 2/3	67	85	98	7	29	85	10	3	1
3	71	90	99	13	41	85	5	2	1

Table C 2 ns, Cα alignment

W	% Known Active			% Known Inactive			% Active / % Inactive		
	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6
1	3	56	93	0	7	40	•	8	2
1 1/3	16	80	94	1	13	51	13	6	2
1 2/3	42	90	96	4	19	71	12	5	1
2	63	91	98	6	29	78	11	3	1
2 1/3	75	92	99	11	34	81	7	3	1
2 2/3	84	93	99	15	42	86	6	2	1
3	88	96	100	19	53	87	5	2	1

Table D 3 ns, Cα alignment

W	% Known Active			% Known Inactive			% Active / % Inactive		
	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6	6 of 6	5 of 6	4 of 6
1	21	79	94	1	13	59	18	6	2
1 1/3	51	90	98	5	21	73	11	4	1
1 2/3	67	92	99	8	34	82	8	3	1
2	85	94	99	11	42	86	8	2	1
2 1/3	90	97	100	16	52	86	5	2	1
2 2/3	92	100	100	21	61	88	4	2	1
3	93	100	100	28	66	88	3	2	1

Fig. S2. Known HIV-1p inhibitors taken from protein-ligand complexes present in the PDB (listed by PDB ID).

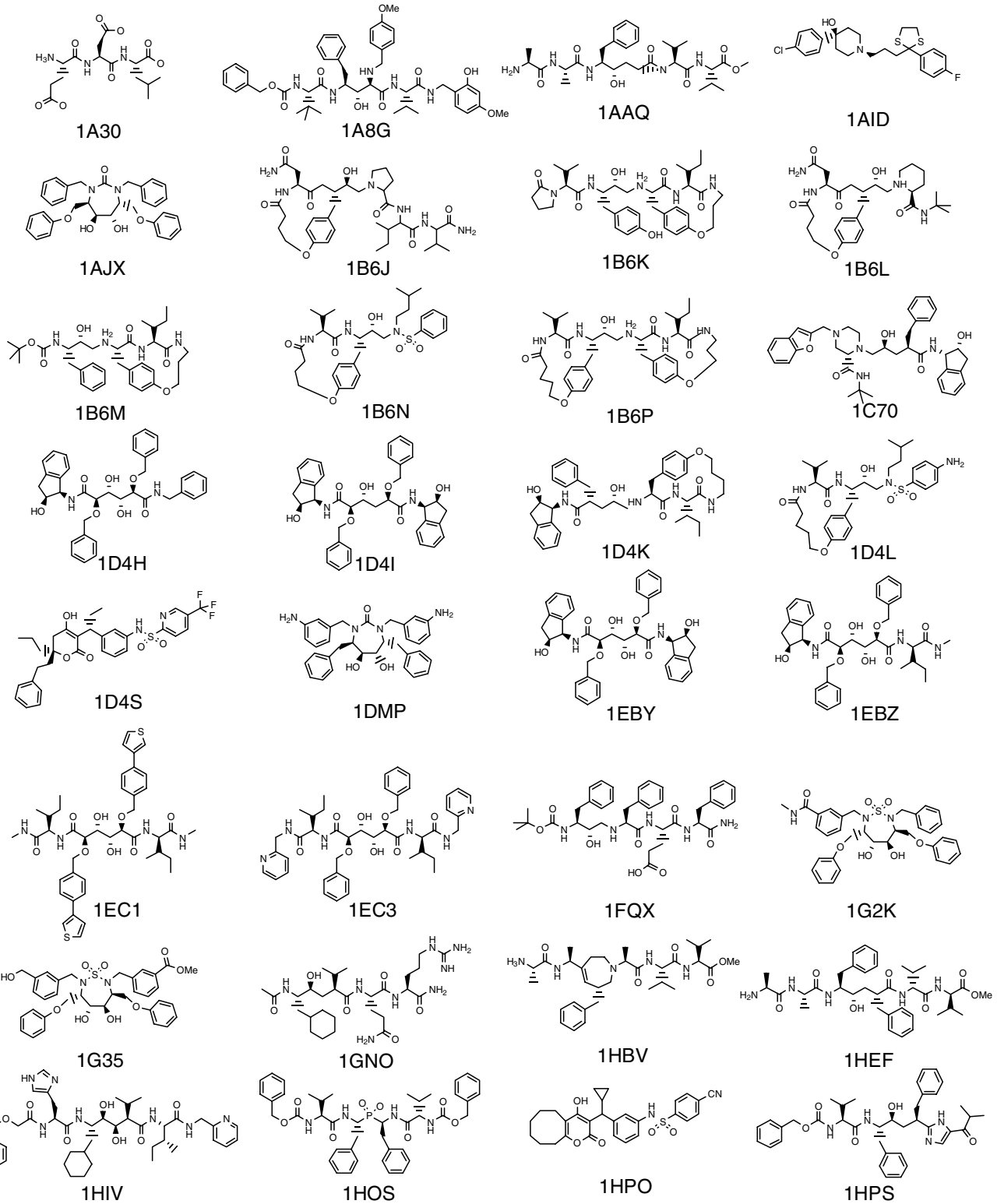


Fig. S2. cont. Known HIV-1p inhibitors taken from protein-ligand complexes present in the PDB (listed by PDB ID).

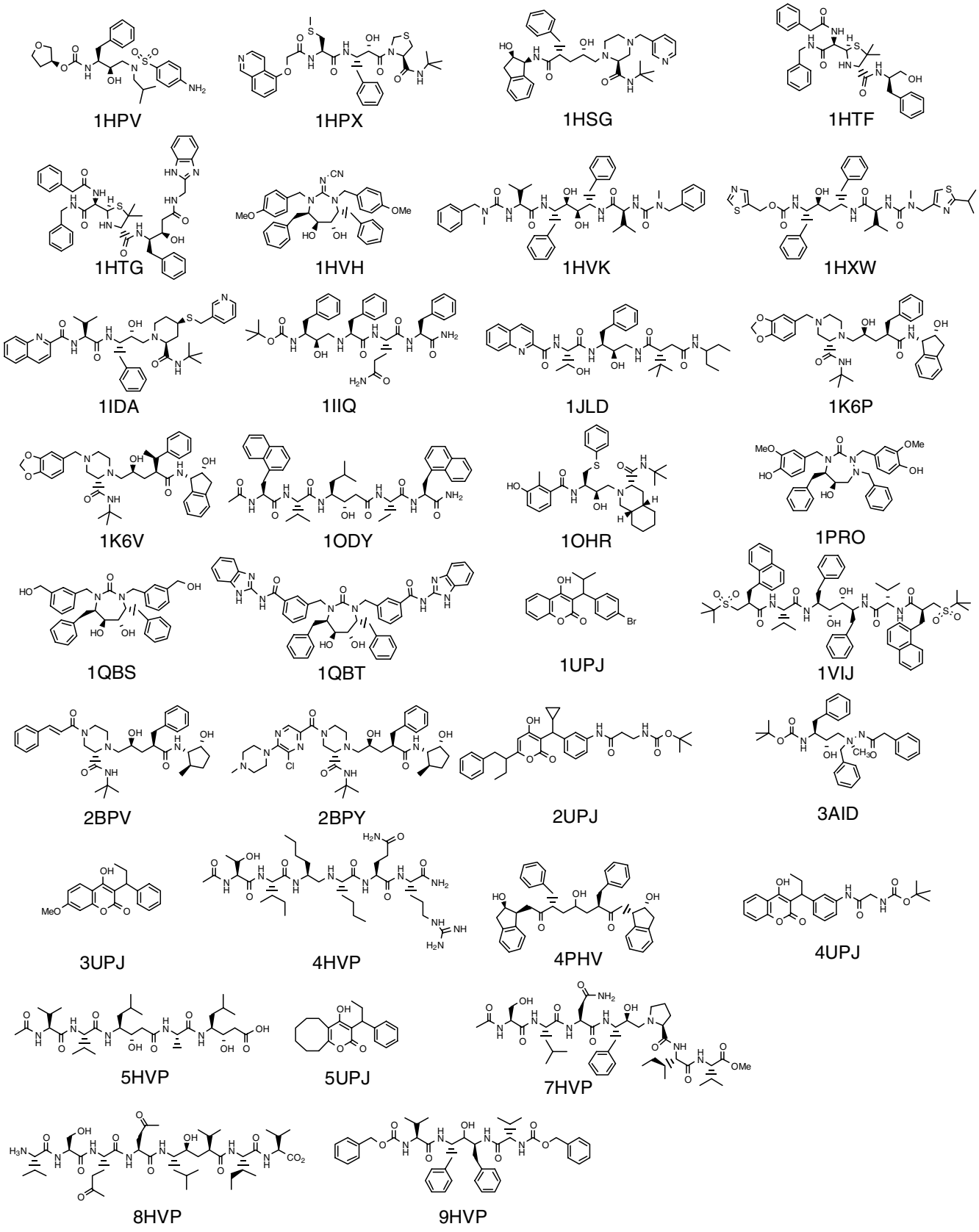


Fig. S2. cont. Known HIV-1p inhibitors taken from the literature (listed by ligand name).

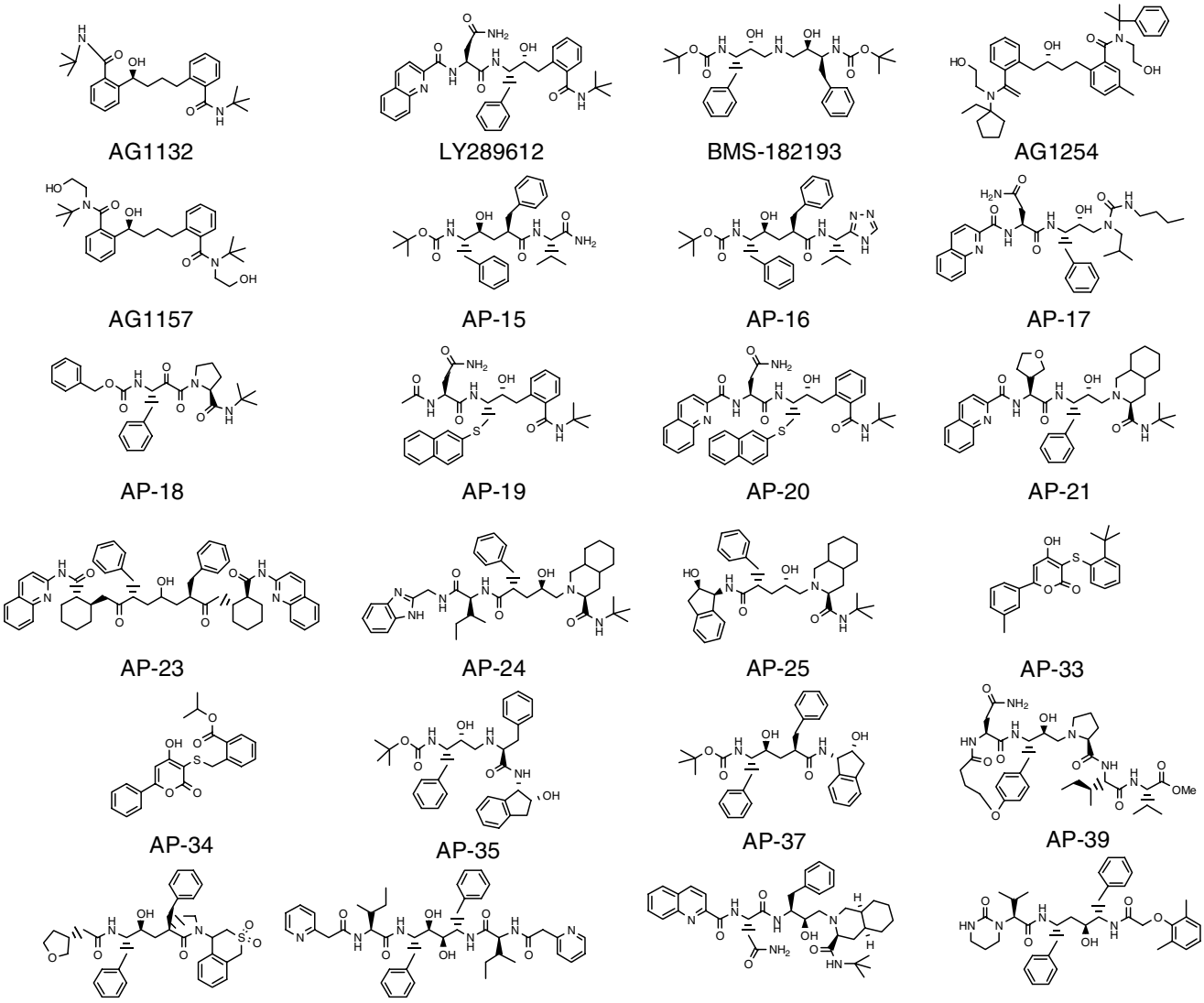


Fig. S3. Drug-like, known inactive compounds taken from the CMC database (listed by CMC code).

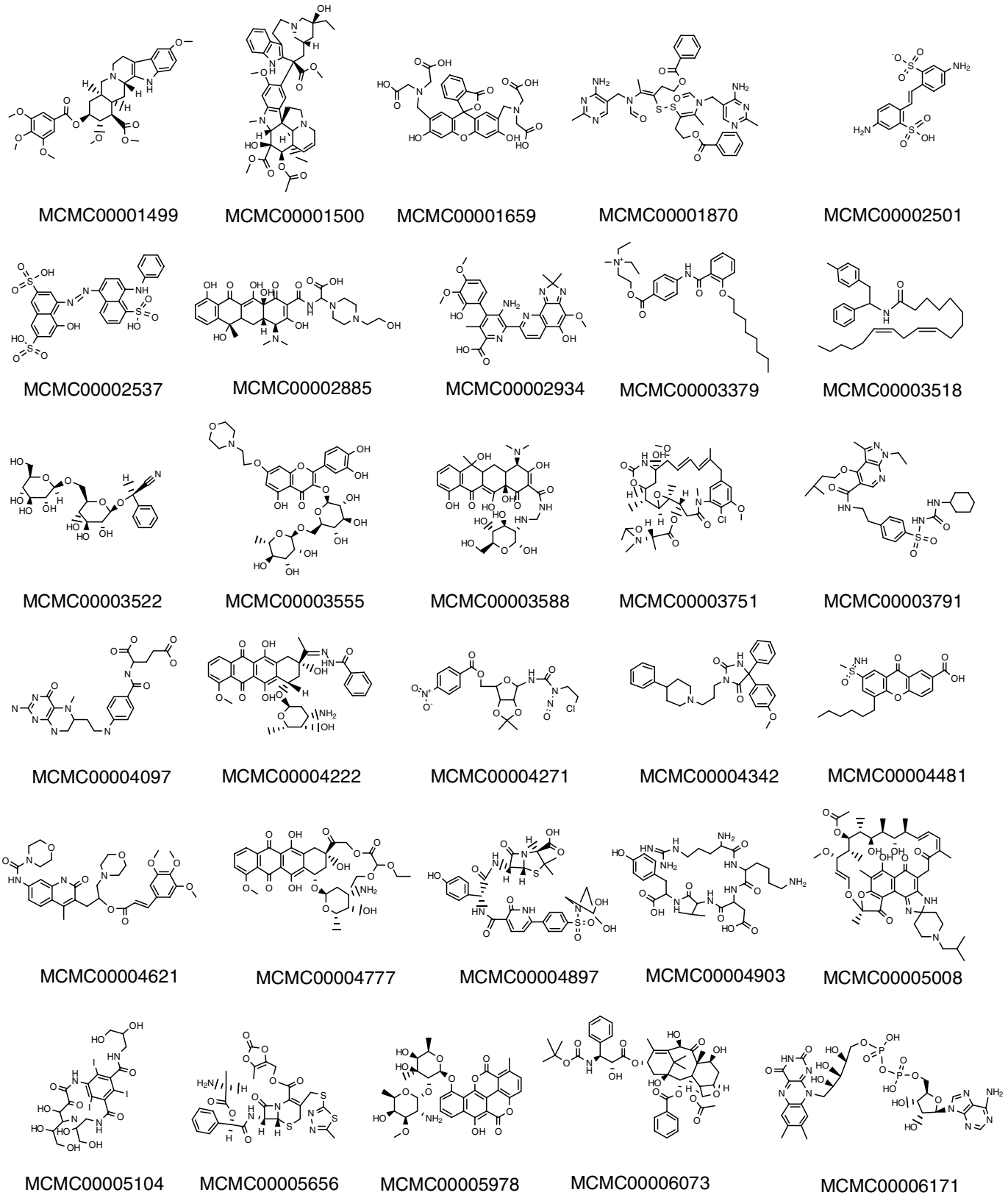
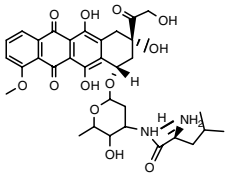
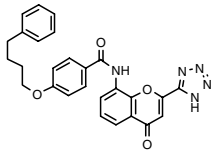


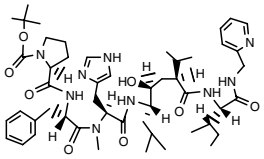
Fig. S3. cont. Drug-like, known inactive compounds taken from the CMC database (listed by CMC code).



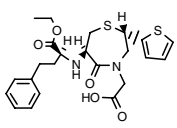
MCMC00006250



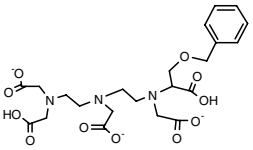
MCMC00006316



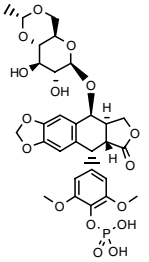
MCMC00006317



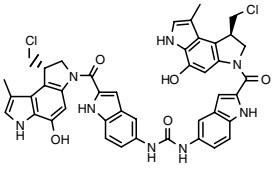
MCMC00006365



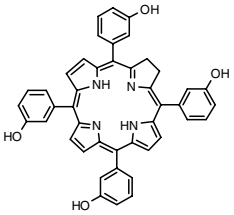
MCMC00006378



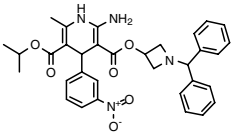
MCMC00006393



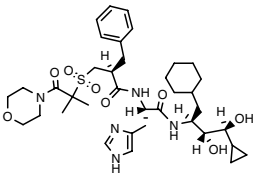
MCMC00006459



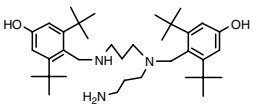
MCMC00006792



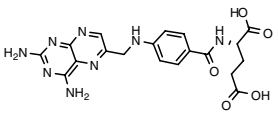
MCMC00006794



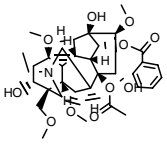
MCMC00006832



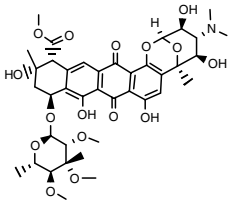
MCMC00006897



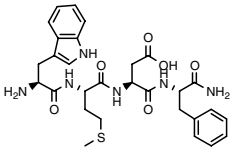
MCMC00006938



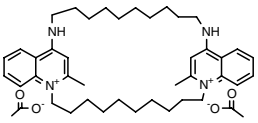
MCMC00007036



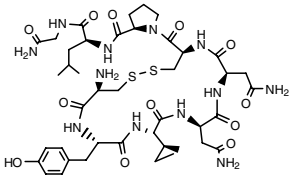
MCMC00007096



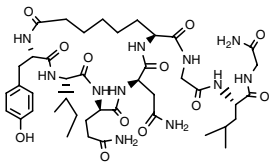
MCMC00007277



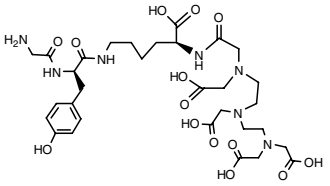
MCMC00007335



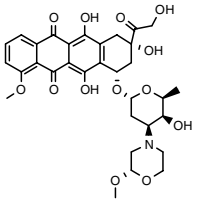
MCMC00007338



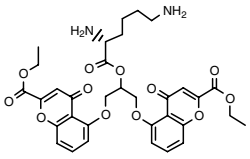
MCMC00007395



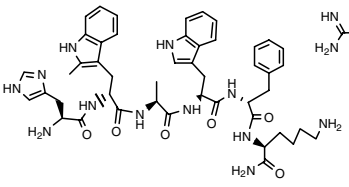
MCMC00007407



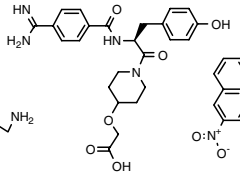
MCMC00007459

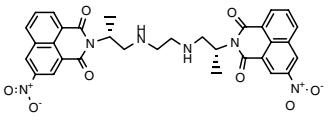


MCMC00007462

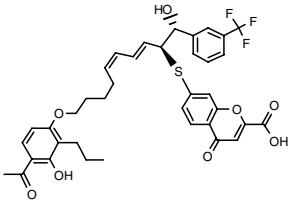


MCMC00007527

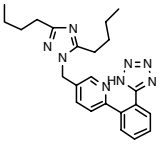




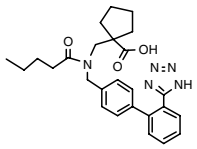
MCMC00007536 MCMC00007538



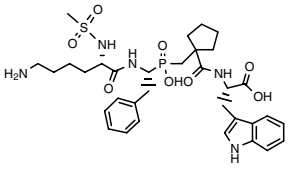
MCMC00007557



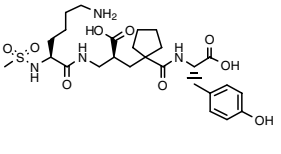
MCMC00007607



MCMC00007610



MCMC00007675



MCMC00007731

Fig. S3. . cont. Drug-like, known inactive compounds taken from the CMC database (listed by CMC code).

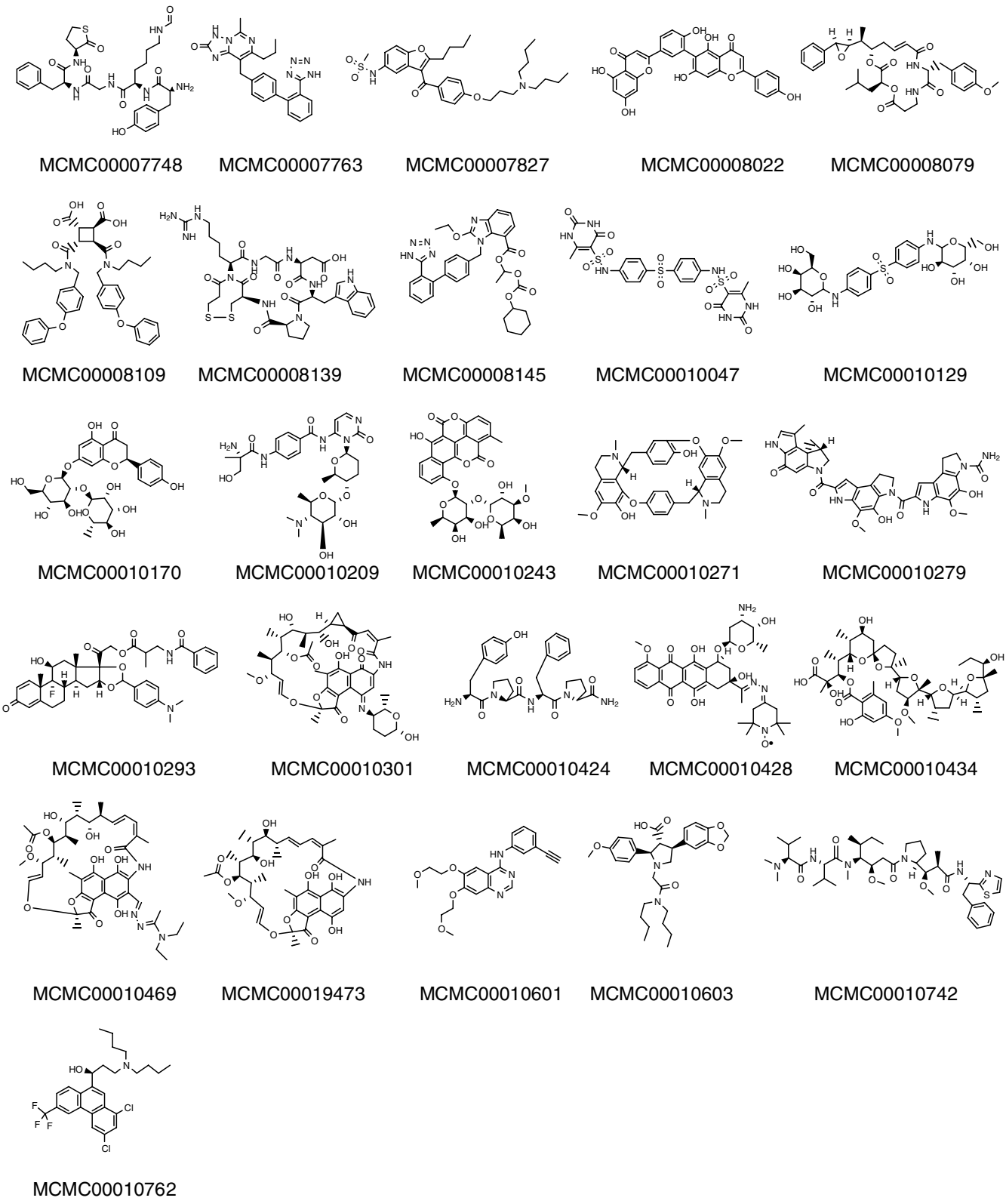
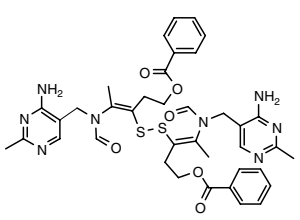
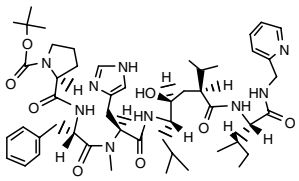


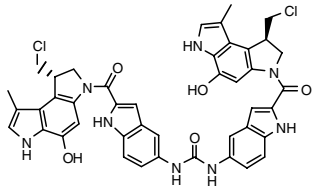
Fig. S4. False positives for the 2ns, 5/6 12/3xRMSD pharmacophore model.



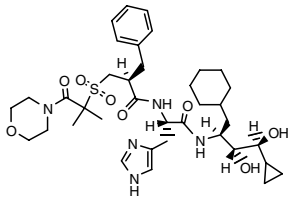
MCMC00001870



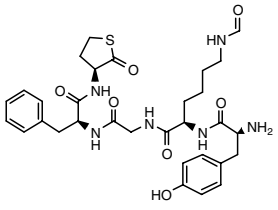
MCMC00006317
renin inhibitor



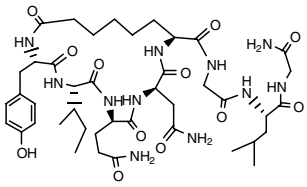
MCMC00006459



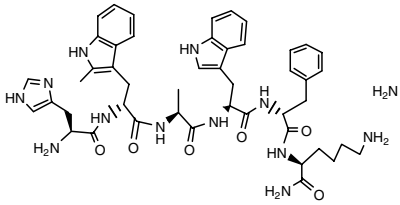
MCMC00006832
renin inhibitor



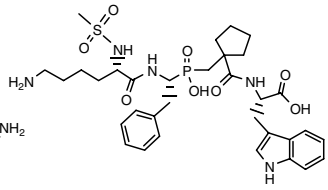
MCMC00007748
peptide analog



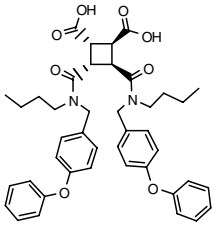
MCMC00007395
peptide analog



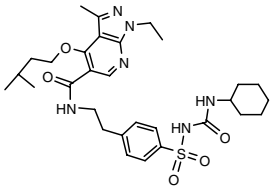
MCMC00007527
peptide analog



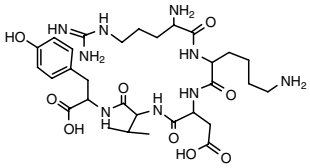
MCMC00007675
peptide TS mimic



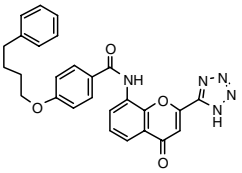
MCMC00008109



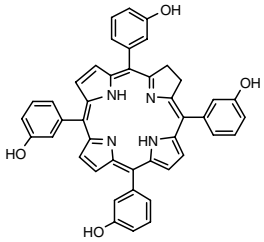
MCMC00003791



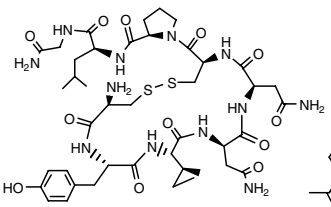
MCMC00004903
peptide analog



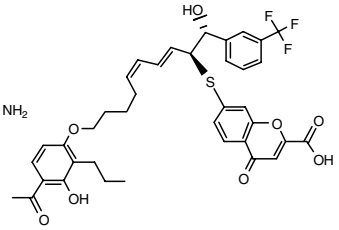
MCMC00006316



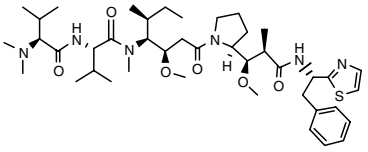
MCMC00006792



MCMC00007338
peptide analog

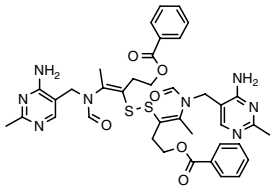


MCMC00007557

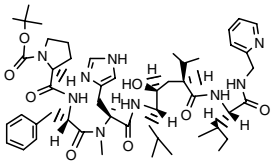


MCMC00010742
peptide analog

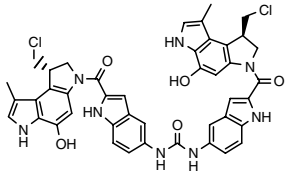
Fig. S5. False Positives for the 3ns model, 6/6, 2xRMSD pharmacophore model.



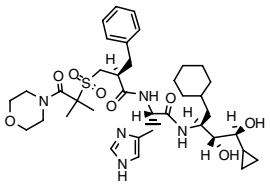
MCMC00001870



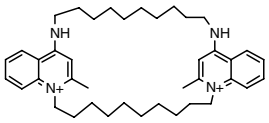
MCMC00006317
renin inhibitor



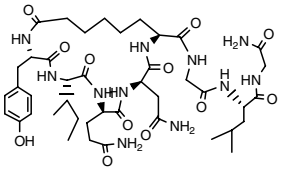
MCMC00006459



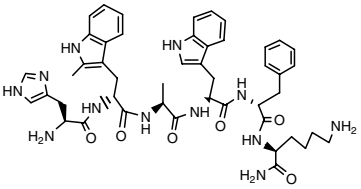
MCMC00006832
renin inhibitor



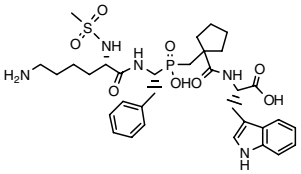
MCMC00007335



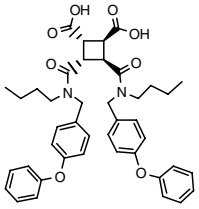
MCMC00007395
peptide analog



MCMC00007527
peptide analog



MCMC00007675
peptide TS mimic



MCMC00008109