Electronic Supplementary Material (ESI) for Molecular BioSystems.

Effects of Protein Flexibility and Active Site Water Molecules on Prediction of Sites of Metabolism for Cytochrome P450 2C19 Substrates

Junhao Li, Jinya Cai, Haixia Su, Hanwen Du, Juan Zhang, Shihui Ding, Guixia Liu,

Yun Tang*, and Weihua Li*

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China

University of Science and Technology, Shanghai 200237, China

* Corresponding authors, Tel: +86-21-64250811; Fax: +86-21-64251033; E-mail:

whli@ecust.edu.cn (W. Li); ytang234@ecust.edu.cn (Y. Tang)

Contents:

Table S1 Statistical data for the 87 substrates (Dataset 1)

Atom name	Numbers[a]
C	1488 (78.3%)
N	164 (8.6%)
S	25 (1.3%)
Other[b]	224 (11.8%)

Bond type[c]	Numbers
ar	908 (44.4%)
am	44 (2.2%)
1	951 (46.5%)
2	137 (6.7%)
3	5 (0.2%)

SOM type[d]	Numbers
Aliphatic-hydroxylation	40 (20.4%)
Aromatic-hydroxylation	52 (26.5%)
N-dealkylation	58 (29.5%)
O-dealkylation	21 (10.7%)
N-oxidation	4 (2.0%)
S-oxidation	9 (4.6%)
Other	12 (6.1%)

Atom types[e]	Numbers
C.1	7
C.2	124
C.3	487
C.ar	868
C.cat	2
Cl	28
F	21
I	2
N.1	3
N.2	13
N.3	4
N.4	33
N.am	39
N.ar	18
N.pl3	54
O.2	89
O.3	75
O.co2	9
S.2	4
S.3	11
S.O	4
S.o2	6

[a.] Numbers of a specific properties, bracket is the percentage of this properties

[b.] Other atoms in this set are O, F, Cl, and I.

[c.] Tripo Sybyl bond types for all heavy atoms

[d.] Statistic data of experimental site of metabolism types

[e.] Tripo Sybyl atom types for all heavy atoms

Table S2 Statistical data for the 59 substrates (Dataset 2, validation set)

Atom name	Numbers[a]
C	924 (76.6%)
N	91 (7.5%)
S	12 (1.0%)
P	2 (0.2%)
Other[b]	178 (14.7%）

Bond type[c]	Numbers
ar	534 (22.4%)
am	18 (0.8%)
1	1766 (73.9%)
2	66 (2.8)
3	5 (0.2)

SOM type[d]	Numbers
Aliphatic-hydroxylation	29 (33.7%)
Aromatic-hydroxylation	9 (10.5%)
N-dealkylation	17 (19.8%)
O-dealkylation	24 (27.9%)
N-oxidation	1 (1.2%)
S-oxidation	4 (4.7%)
Other	2 (2.3%)

Atom types[e]	Numbers
Br	2
C.1	5
C.2	77
C.3	353
C.ar	489
Cl	25
F	19
N.1	5
N.2	5
N.3	10
N.4	18
N.am	17
N.ar	16
N.pl3	20
O.2	43
O.3	83
O.co2	6
P.3	2
S.3	7
S.o	2
S.o2	3

[a.] Numbers of a specific properties, bracket is the percentage of this properties

[b.] Other atoms in this set are O, F, Cl, and Br.

[c.] Tripo Sybyl bond types for all heavy atoms

[d.] Statistic data of experimental site of metabolism types

[e.] Tripo Sybyl atom types for all heavy atoms

Table S3 Comparison of the 12 Kclust outputs to the Crystal structure

	Volume ($Å^3$)	RMSD [a]
Snapshot 01	475	1.20
Snapshot 02	457	1.27
Snapshot 03	456	1.23
Snapshot 04	474	1.12
Snapshot 05	488	1.27
Snapshot 06	464	1.21
Snapshot 07	518	1.31
Snapshot 08	446	1.24
Snapshot 09	481	1.13
Snapshot 10	466	1.31
Snapshot 11	481	1.12
Snapshot 12	510	1.22
PDB_Chain A	444	NA

[a.] The RMSD value is calculated based on the superimposition of protein C_α atoms

Table S4 Detailed results for docking models alone (dataset 1)

		Simu_1[a]	Simu_2[b]	Simu_3[c]	AVG ± STD[d]
R[e]	Top-1 (pose) %	35.63	42.53	35.63	37.93 ± 3.25
	Top-2 (pose) %	54.02	54.02	55.17	54.41 ± 0.54
	Top-3 (pose) %	60.92	60.92	63.22	61.69 ± 1.08
RH[f]	Top-1 (pose) %	36.78	35.63	42.53	38.31 ± 3.02
	Top-2 (pose) %	51.72	54.02	49.43	51.72 ± 1.88
	Top-3 (pose) %	63.22	66.67	57.43	62.44 ± 3.81
S[g]	Top-1 (pose) %	35.63	34.48	37.93	36.02 ± 1.43
	Top-2 (pose) %	47.13	47.13	51.72	48.66 ± 2.17
	Top-3 (pose) %	63.22	59.77	62.07	61.69 ± 1.43
SH[h]	Top-1 (pose) %	41.38	36.78	41.38	39.85 ± 2.17
	Top-2 (pose) %	56.32	51.72	50.57	52.87 ± 2.48
	Top-3 (pose) %	60.92	66.67	59.77	62.45 ± 3.02
M[i]	Top-1 (pose) %	33.33	32.18	35.63	33.72 ± 1.43
	Top-2 (pose) %	42.53	40.23	42.53	41.76 ± 1.08
	Top-3 (pose) %	57.43	55.17	56.32	56.31 ± 0.92
MW[j]	Top-1 (pose) %	32.18	35.63	35.63	34.48 ± 1.63
	Top-2 (pose) %	43.68	43.68	42.53	43.30 ± 0.54
	Top-3 (pose) %	62.07	63.22	62.07	62.45 ± 0.54
PS[k]	Top-1 (pose) %	37.93	41.38	43.68	41.00 ± 2.36
	Top-2 (pose) %	56.32	58.62	62.07	59.00 ± 2.36
	Top-3 (pose) %	74.71	73.56	72.41	73.56 ± 0.94
MS[l]	Top-1 (pose) %	49.43	48.28	43.68	47.13 ± 2.48
	Top-2 (pose) %	68.97	65.52	62.07	65.52 ± 2.82
	Top-3 (pose) %	79.31	79.31	74.71	77.78 ± 2.17

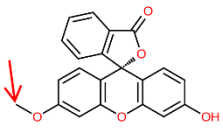
[a]. Simulation 1
[b]. Simulation 2
[c]. Simulation 3
[d]. Average accuracy of three parallel simulations and standard deviation
[e]. Results from rigid receptor
[f]. Results from rigid receptor with HOH601
[g]. Results from GOLD semi-flexible receptor
[h]. Results from GOLD semi-flexible receptor with HOH601
[i]. Results from MD snapshots
[j]. Results from MD snapshots with WAT5519
[k]. Results from PDB_SCs
[l]. Results from MD_SCs

Table S5 Detailed results for combined models (dataset 1)

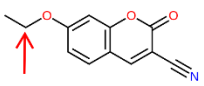
		Simu_1	Simu_2	Simu_3	AVG ± STD
R	Top-1 %	70.11	71.26	67.82	69.73 ± 1.43
	Top-2 %	78.16	72.41	74.71	75.10 ± 2.36
	Top-3 %	83.91	82.76	83.91	83.53 ± 0.54
RH	Top-1 %	64.37	63.22	66.67	64.75 ± 1.43
	Top-2%	75.86	73.56	72.41	73.95 ± 1.43
	Top-3 %	81.61	82.76	81.61	81.99 ± 0.54
S	Top-1 %	67.82	70.11	66.67	68.20 ± 1.43
	Top-2 %	78.16	77.01	72.41	75.86 ± 2.48
	Top-3 %	86.21	82.76	87.36	85.44 ± 1.95
SH	Top-1 %	63.22	67.82	68.97	66.67 ± 2.48
	Top-2 %	75.86	74.71	72.41	74.33 ± 1.43
	Top-3 %	81.61	82.76	88.51	84.29 ± 3.02
M	Top-1 %	75.86	75.86	70.11	73.95 ± 2.71
	Top-2 %	83.91	83.91	81.61	83.14 ± 1.08
	Top-3 %	86.21	86.21	88.51	86.97 ± 1.08
MW	Top-1 %	77.01	77.01	72.41	75.48 ± 2.17
	Top-2 %	83.91	85.06	82.76	83.91 ± 0.94
	Top-3 %	89.66	89.66	85.06	88.12 ± 2.17
PS	Top-1 %	78.16	77.01	78.16	77.78 ± 0.54
	Top-2 %	88.51	85.06	86.21	86.59 ± 1.44
	Top-3 %	90.80	89.66	90.80	90.42 ± 0.54
MS	Top-1 %	75.86	73.56	71.26	73.56 ± 1.88
	Top-2 %	86.21	86.21	83.91	85.44 ± 1.08
	Top-3 %	91.95	91.95	90.80	91.57 ± 0.54

Table S6 Detailed results for combined models (dataset 2)

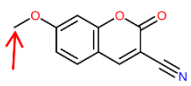
		Simu_1	Simu_2	Simu_3	AVG ± STD
Rigid	Top-1 %	50.85	54.24	55.93	53.67 ± 2.11
	Top-2 %	69.49	71.19	72.88	71.19 ± 1.38
	Top-3 %	83.05	84.75	84.75	84.18 ± 0.80
Semi-flex	Top-1 %	59.32	52.54	54.24	55.37 ± 2.88
	Top-2 %	74.58	69.49	71.19	71.75 ± 2.11
	Top-3 %	88.14	84.75	88.14	87.01 ± 1.60
MD_07	Top-1 %	59.32	61.02	59.32	59.89 ± 0.80
	Top-2 %	74.58	72.88	77.97	75.14 ± 2.11
	Top-3 %	84.75	79.66	83.05	82.49 ± 2.11
PDB_SC_02	Top-1 %	64.41	59.32	57.63	60.45 ± 2.88
	Top-2 %	74.58	77.97	72.88	75.14 ± 2.11
	Top-3 %	81.36	84.75	83.05	83.05 ± 1.38
MD_SC_34	Top-1 %	64.41	61.02	61.02	62.15 ± 1.60
	Top-2 %	79.66	79.66	77.97	79.10 ± 0.80
	Top-3 %	88.14	89.83	89.83	89.27 ± 0.80



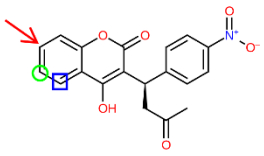
3-O-methylfluorescein



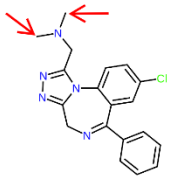
3-cyano-7-ethoxycoumarin



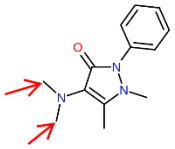
3-cyano-7-methoxycoumarin



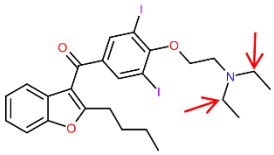
acenocoumarol



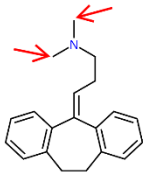
adinazolam



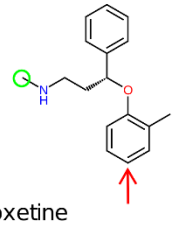
aminopyrine



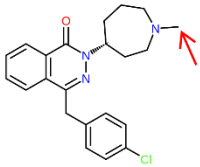
amiodarone



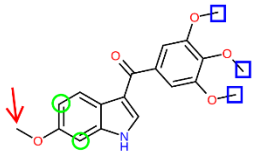
amitriptyline



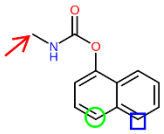
atomoxetine



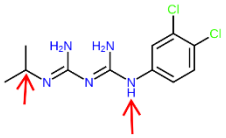
azelastine



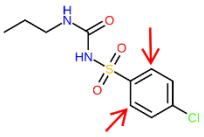
bpr0l075



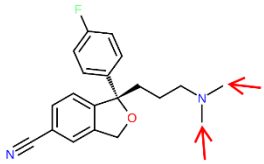
carbaryl



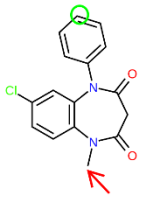
chlorproguanil



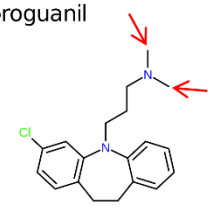
chlorpropamide



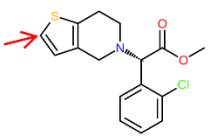
citalopram



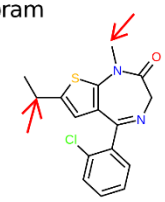
clobazam



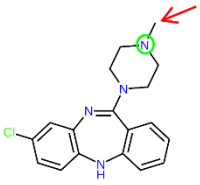
clomipramine



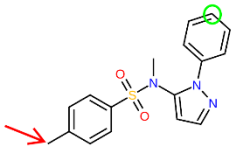
clopidogrel



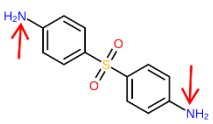
clotiazepam



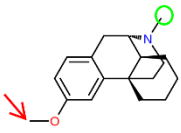
clozapine



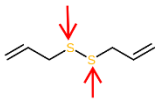
compound-4



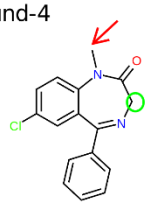
dapsone



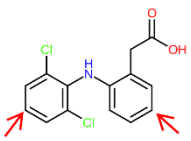
dextromethorphan



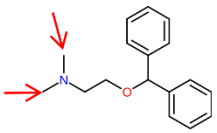
diallyl-disulfide



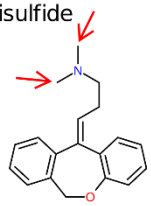
diazepam



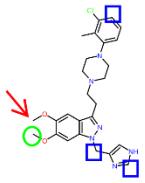
diclofenac



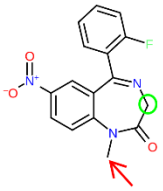
diphenhydramine



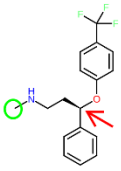
doxepin



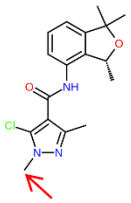
dy-9760e



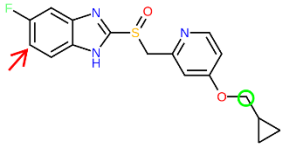
flunitrazepam



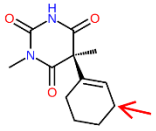
fluoxetine



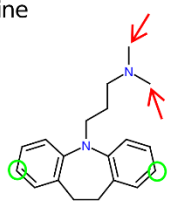
furametpyr



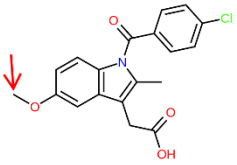
h259-31



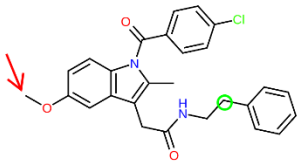
hexobarbital



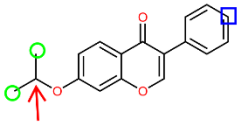
imipramine



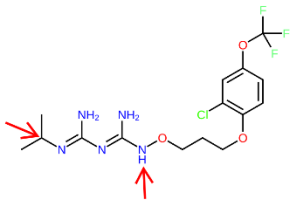
indomethacin



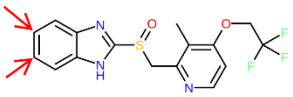
indomethacin-phenethylamide



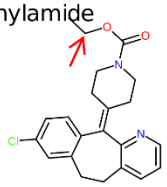
ipriflavone



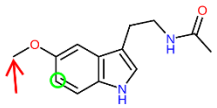
jpc-2056



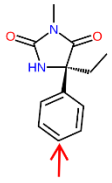
lansoprazole



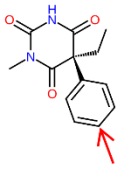
loratadine



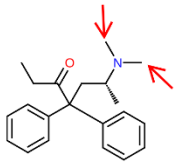
melatonin



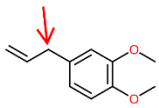
mephenytoin



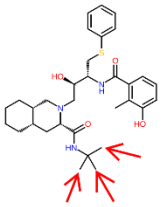
mephobarbital



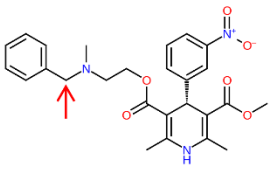
methadone



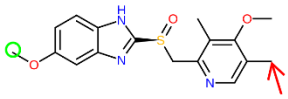
methyleugenol



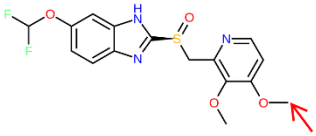
nelfinavir



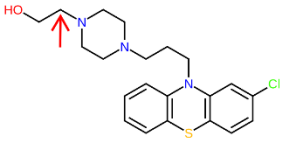
nicardipine



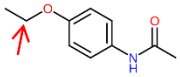
omeprazole



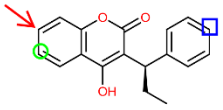
pantoprazole



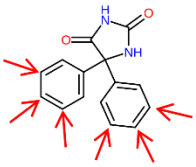
perphenazine



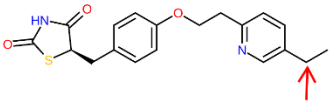
phenacetin



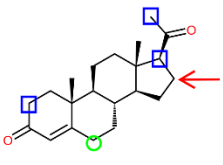
phenprocoumon



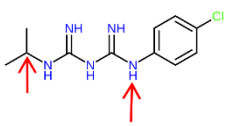
phenytoin



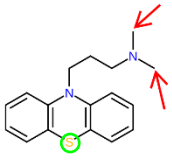
pioglitazone



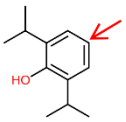
progesterone



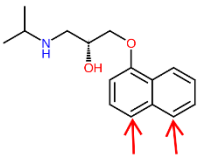
proguanil



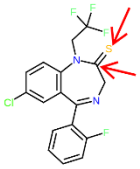
promazine



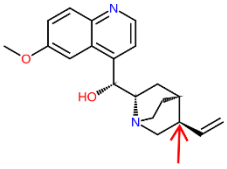
propofol



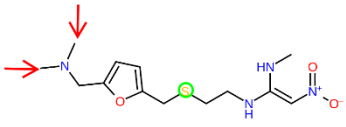
propranolol



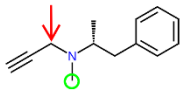
quazepam



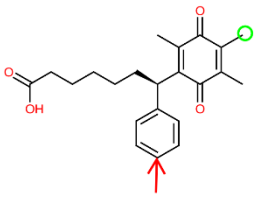
quinine



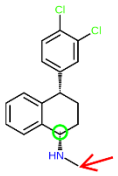
ranitidine



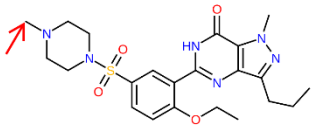
selegiline



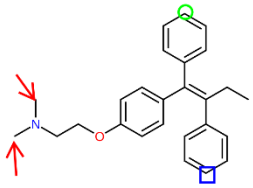
seratrodast



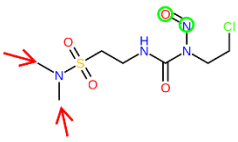
sertraline



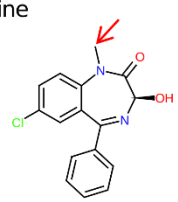
sildenafil



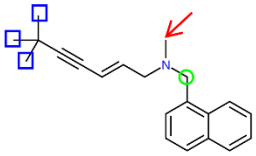
tamoxifen



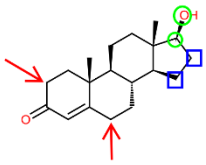
tauromustine



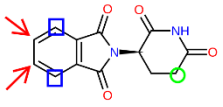
temazepam



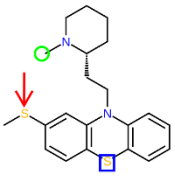
terbinafine



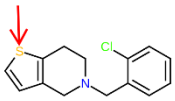
testosterone



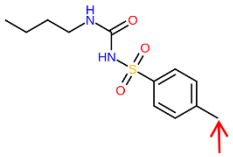
thalidomide



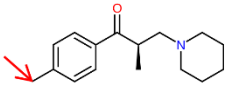
thioridazine



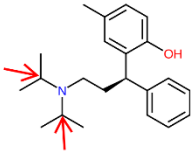
ticlopidine



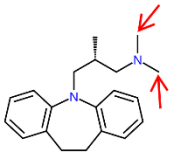
tolbutamide



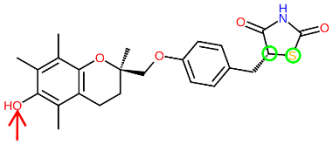
tolperisone



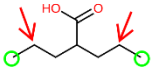
tolterodine



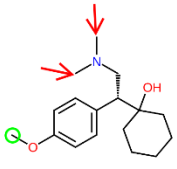
trimipramine



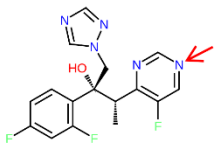
troglitazone



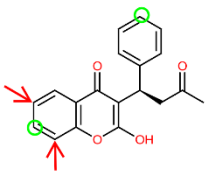
valproic-acid



venlafaxine



voriconazole



warfarin-r

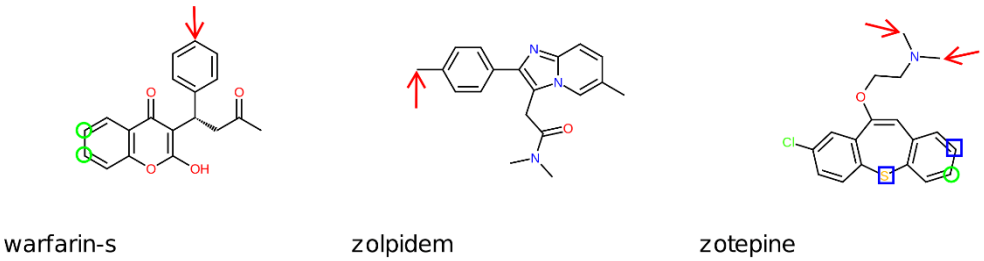
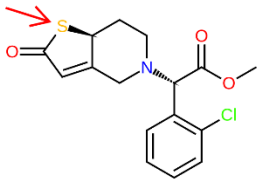
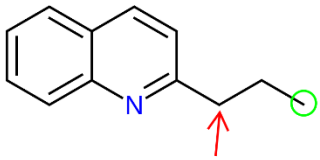


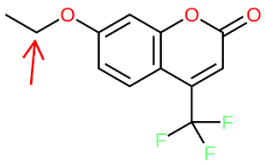
Fig. S1 2D structure of dataset 1 (Red arrays represent the primary SOMs, while the green circles and blue squares represent the secondary and tertiary SOMs, respectively.)



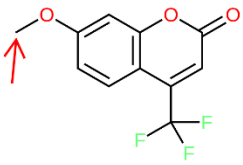
2-oxo-clopidogrel



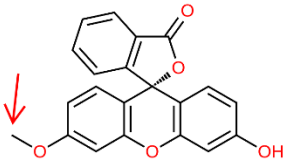
2n-propylquinoline



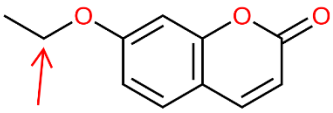
3-ethoxy-4-trifluoromethylcoumarin



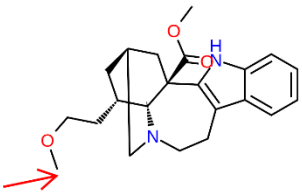
3-methoxy-4-trifluoromethylcoumarin



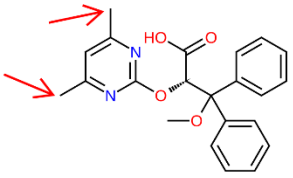
3-o-methylfluorescein



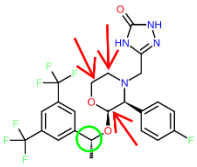
7-ethoxycoumarin



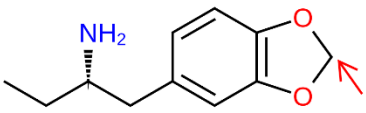
18-methoxycoronaridine



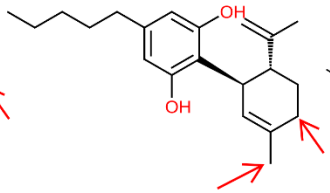
ambrisentan



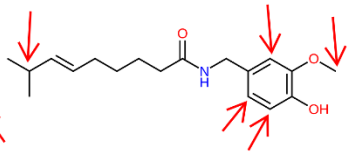
aprepitant



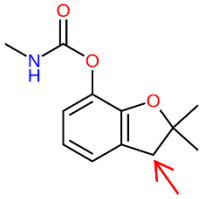
bdb



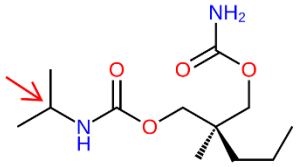
cannabidiol



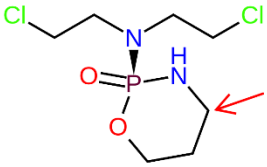
capsaicin



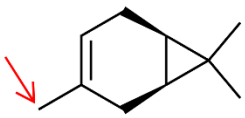
carbofuran



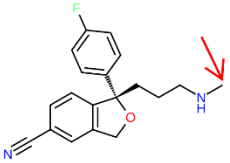
carisoprodol



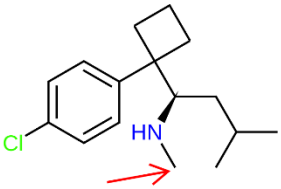
cyclophosphamide



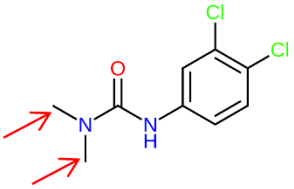
delta3-carene



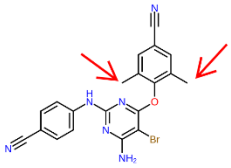
desmethyl-citalopram



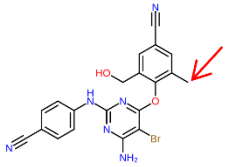
desmethylsibutramine



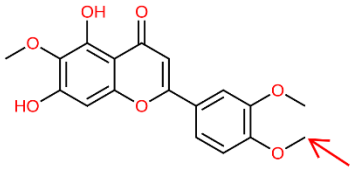
diuron



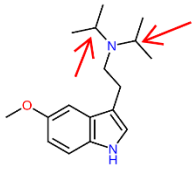
etravirine



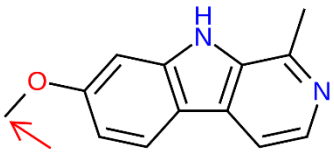
etravirine-mono-oh



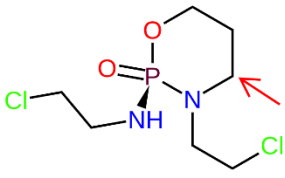
eupatilin



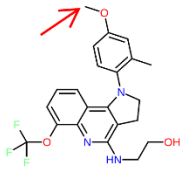
foxy



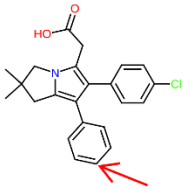
harmine



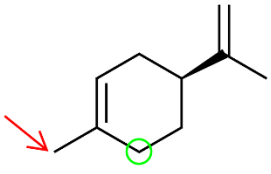
ifosfamide



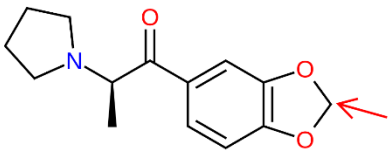
kr-60436



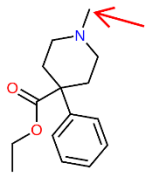
licofelone



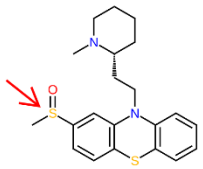
limonene



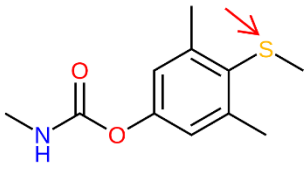
mdppp



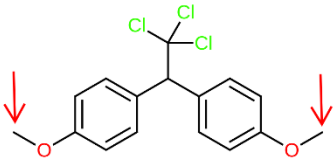
meperidine



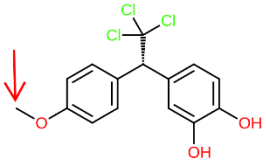
mesoridazine



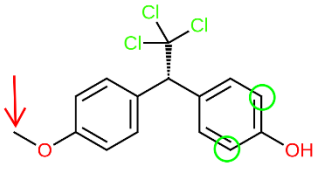
methiocarb



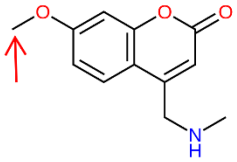
methoxychlor



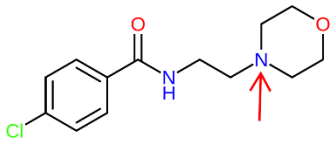
methoxychlor-catechol



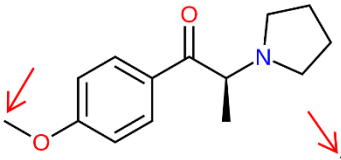
methoxychlor-mono-oh



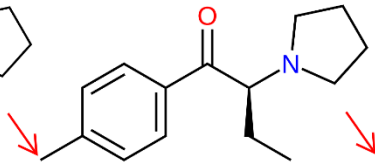
mmamc



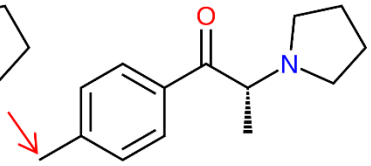
moclobemide



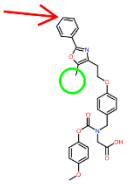
moppp



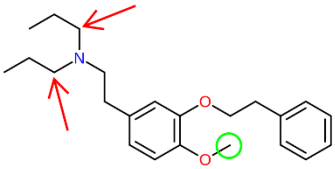
mpbp



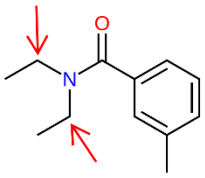
mppp



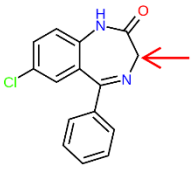
muraglitazar



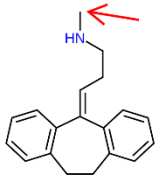
ne-100



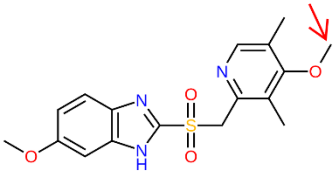
nn-dimethyl-m-toluamide



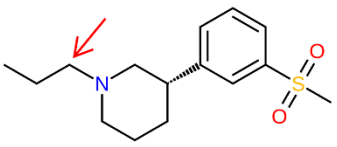
nordiazepam



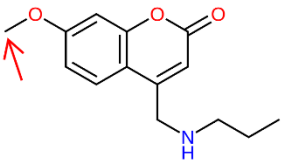
nortriptyline



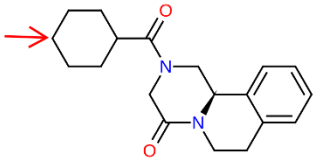
omeprazole-sulfone



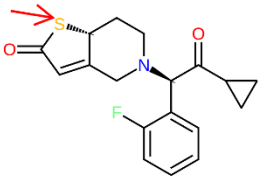
osu-6162



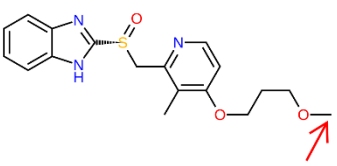
pmamc



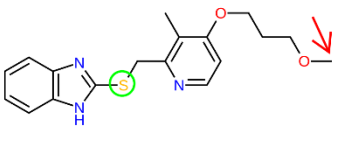
praziquantel



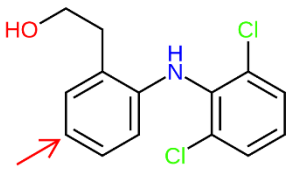
r-95913



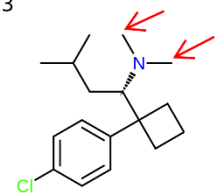
rabeprazole



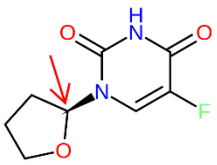
rabeprazole-thioether



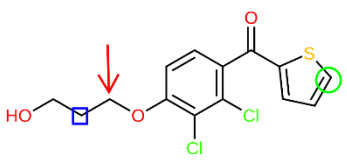
reduced-diclofenac



sibutramine



tegafur



tienilic-acid

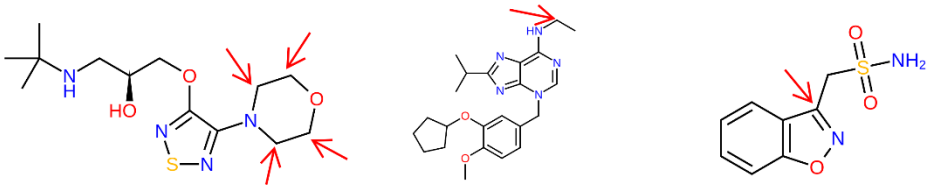


Fig. S2 2D structure of dataset 2 (Red arrays represent the primary SOMs, while the green circles and blue squares represent the secondary and tertiary SOMs, respectively.)

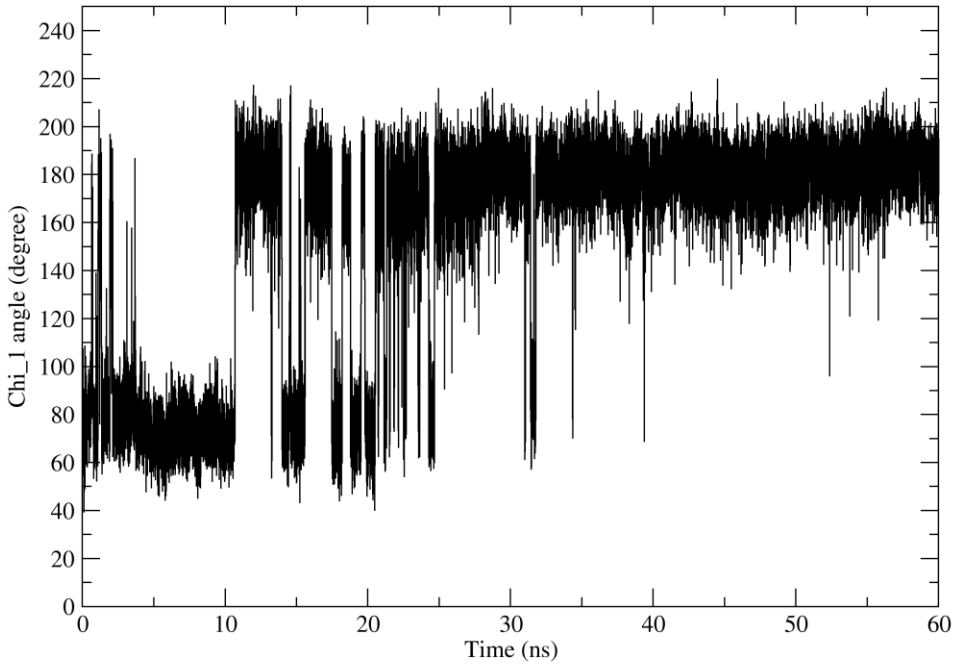


Fig. S3 The χ_1 angle of Phe476 during MD simulation

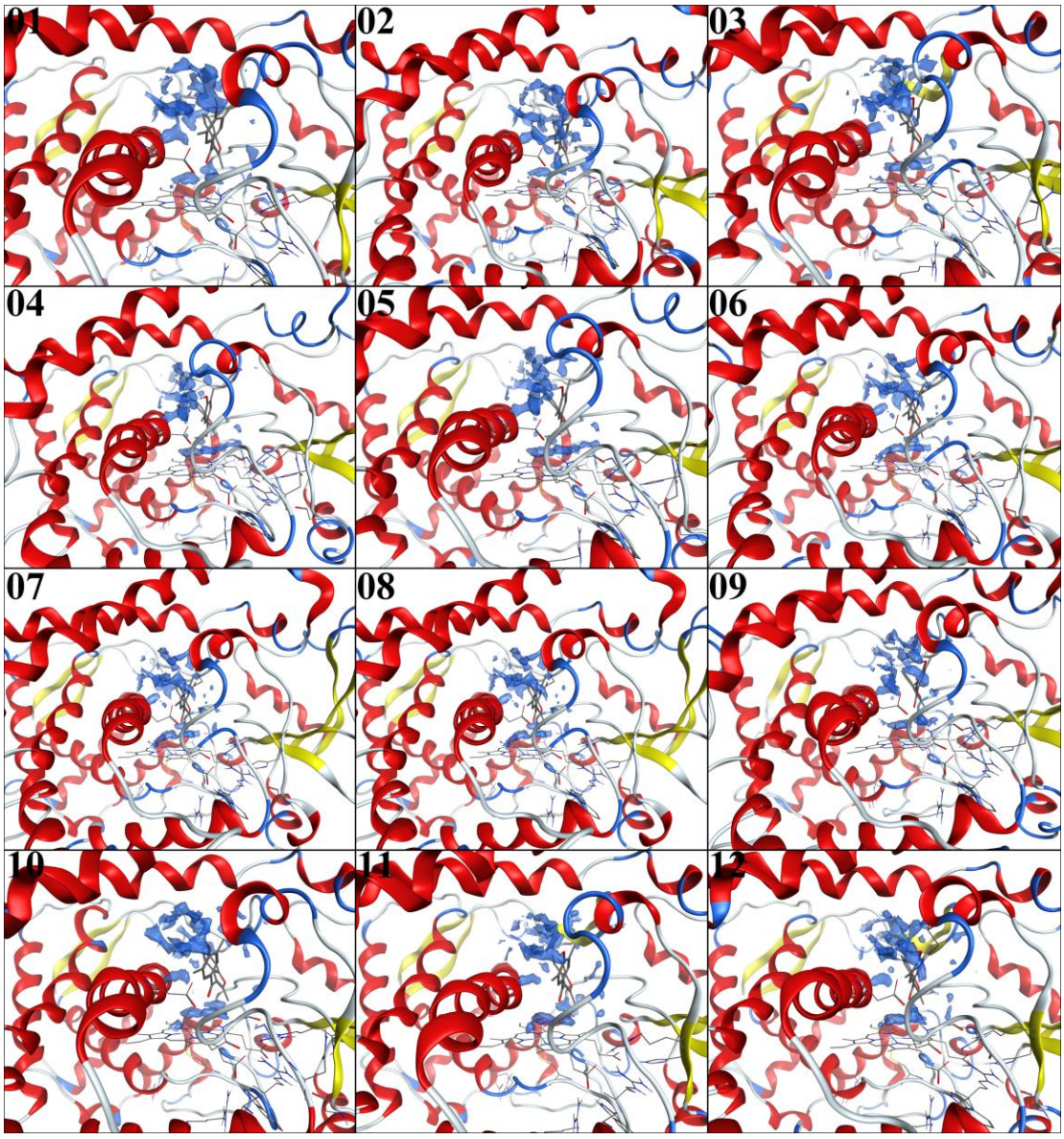


Fig. S4 The energy stable water distribution calculated by MOE 3D-RISM algorithm (with 10 Å of ligand 0XV, salt concentration of 100 mM. From Fig. S4, the area near the middle of helix I is one of the conserved hydration sites in the 12 clustered snapshots, indicating that it is reasonable to identify the hydration site by use of the dynamic WAT5519.)

Interpretation of the supporting animation movie

This movie was generated using snapshots extracted from 20-22ns (10ps per snapshot). Cl - and irrelevant water molecules are removed. Substrates Recognized Site 1 to 6 are coloring in magnetic, green, violet, yellow, cyan, and red, respectively. Water molecules are represented in red sphere, expect for WAT5519, which is represented in cyan sphere. The PDB ligand is shown in green sticks, and heme and its ligated residue Cys435 are represented in red sticks.

Additional interpretation of the criterion for docking alone

In this work, we prepared different kinds of receptors for docking. Each docking run produced 30 outputs and the top 3 clustered outputs were considered. This resulted in a total of 125*3 (85*3 for dataset 2) outputs that need to be examined for a specific receptor. Therefore, a combined protocol of automated scripts and visual inspection was adopted for determining the sites of metabolism. Automated scripts were implemented to exclude those outputs that do not meet the criterion of the 6 Å rule. Then, the protein structures with good performance were further selected for visual inspection by considering the SOM orientation. For example, in simulation 1 using dataset 1, the accuracy of the selected MD_SC_34 with the criterion of 6Å rule is 69.0%, 79.3%, and 83.9% at the top-1, top-2, and top-3 pose, respectively. By considering the site orientation with visual inspection, the prediction accuracy decreased to 49.4%, 69.0%, and 79.3% at the top-1, top-2, and top-3 pose, respectively (refer to “Simu_1” in Table S4).

Additional predictions based on MD simulation with apo form of CYP2C19

In the preparation stage of this manuscript, we had performed MD simulations in both apo and holo forms of CYP2C19. In the 30 ns simulation of the apo form, we observed that the side chains of certain active site residues shrank into the active site (especially in helices of F and I). The representative snapshots from 4 major clusters were superimposed, as shown in Fig. S5. The active site volume was calculated with POVME 2.0 and presented in Table S7. The substrates in dataset 1 were docked into the active site of the 4 representative snapshots. The results were summarized in Tables S8 and S9. Compared to those from the holo form (Table 1), slight improvements were observed in docking models. In contrast, there was a slight decrease in combined models when using snapshots from apo form. However, predictions from MD apo form could not be superior to those from tCONCOORD sampled structures, indicating that the degrees of receptor flexibility using apo form MD is still insufficient for docking models.

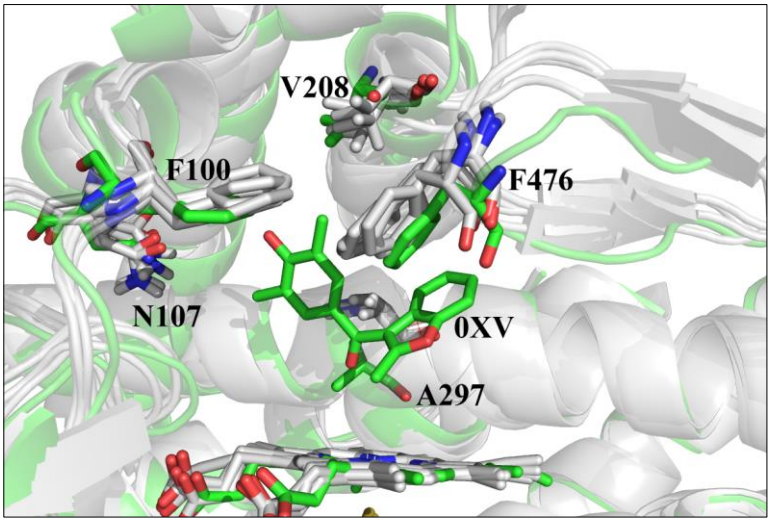


Fig. S5 Superposition of 4 representative snapshots from apo MD simulation to the crystal structure. The MD snapshots were represented in gray cartoons and the crystal structure was in green cartoon. The crystallographic ligand 0XV was shown in green sticks.

Table S7 Active site volume and C_α RMSD when compared to crystal structure

	Volume[a] (Å^3)	RMSD
SN1[b]	300	1.23
SN2	243	1.32
SN3	200	1.16
SN4	294	1.30
PDB_Chain A	444	NA

[a:] The active site volume for the four represented snapshots from aop-MD were calculated by POVME 2.0.

[b]: Snapshot ID

Table S8 Predictions using apo MD snapshots

	Docking models				Docking combined with SMARTCYP			
	SN1	SN2	SN3	SN4	SN1	SN2	SN3	SN4
Top-1 pose %	36.78	35.63	14.94	31.03	71.26	68.97	19.54	67.82
Top-2 pose %	50.57	44.83	16.09	47.13	82.76	77.01	22.99	77.01
Top-3 pose %	57.47	54.02	19.54	54.02	89.66	82.76	25.29	83.91

We also analyzed the water behavior in the active site of the apo form during MD simulation. We found that only WAT469 was conserved in the active site of these snapshots (Fig. S6). Thus this water molecule was kept in the docking of the substrates in dataset 1. The results were presented in Table S9. The prediction accuracies were not as good as those based on the ligand-bound complex. In addition, the effect of WAT469 on predictions appeared to be receptor- and methodology-dependent. This conclusion is consistent with the one in the manuscript.

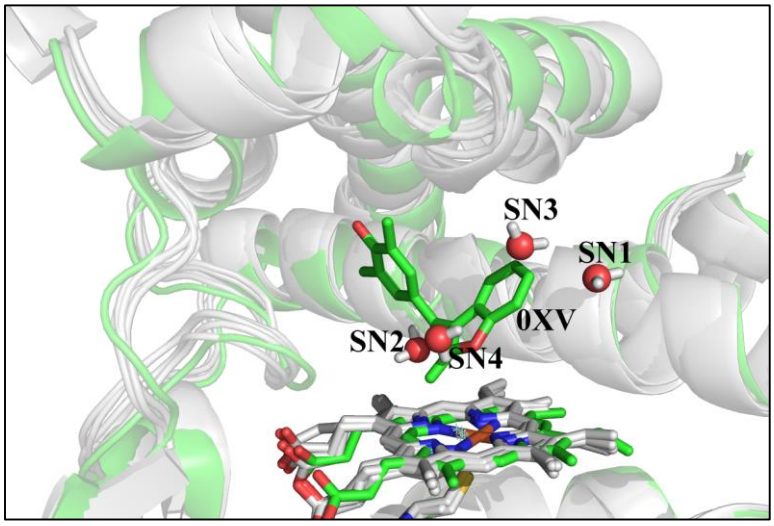


Fig. S6 The distribution of the conserved WAT469 in the four MD snapshots: SNs 1-4. These snapshots (represented in gray cartoon) were superposed to crystal structure (colored in green cartoon, with the ligand showed in green sticks).

Table S9 Predictions using apo MD snapshots with WAT469 in active site

	Docking models				Docking combined with SMARTCYP			
	SN1	SN2	SN3	SN4	SN1	SN2	SN3	SN4
Top-1 pose %	33.33	35.63	12.64	35.63	68.97	71.26	19.54	68.97
Top-2 pose %	43.68	41.38	14.94	47.13	82.76	82.76	22.99	78.16
Top-3 pose %	57.47	48.28	16.09	58.62	89.66	83.91	25.29	87.36