WhichP450 – A multi-class categorical model to predict the major metabolising CYP450 isoform for a compound.

Journal Computer-Aided Molecular Design

Peter A. Hunt[*1], *Matthew D. Segall*[1], *Jonathan D. Tyzack*[2],

1 - Optibrium Ltd., F5-6 Blenheim House, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, UK
2 - The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridge, CB10 1SD, UK
peter@optibrium.com;

Supplementary Information

Expected Uniform Random Performance

When evaluating the expected success rates for uniform random chance, the number of isoforms associated with the metabolism of a compound and which top-*k* criterion is used need to be taken into account. Table S1 shows the probability for random success based upon these two factors.

Table S1. Probabilities of achieving a successful top-*k* result with a uniform random selection, depending on the number of observed isoforms for a compound.

	Top-1	*Top-2*	*Top-3*
Single isoform	1/7 = 0.14	1-(6/7*5/6) = 0.28	1-(6/7*5/6*4/5) = 0.42
2 isoforms	2/7 = 0.28	1-(5/7*4/6) = 0.52	1-(5/7*4/6*3/5) = 0.71
3 isoforms	3/7 = 0.42	1-(4/7*3/6) = 0.71	1-(4/7*3/6*2/5) = 0.88
4 isoforms	4/7 = 0.57	1-(3/7*2/6) = 0.85	1-(3/7*2/6*1/5) = 0.97
5 isoforms	5/7 = 0.71	1-(2/7*1/7) = 0.96	1.00

The probabilities in the top-2 and top-3 cases are calculated as the 1 – (probability of failing) and, as one might expect, if you have a list of 4 isoforms and are allowed 3 guesses then there is only a very small chance that one could get the prediction wrong. Hence, for isoform lists that are longer than 4 isoforms one is assured of success if given enough guesses. Also, as the test set composition changes, the distribution of compounds with either 1, 2, 3, 4 or 5 associated isoforms will vary. Therefore, the Expected Uniform Random results will vary for each test set and are calculated by multiplying the numbers of compounds with each length of isoform list by the probabilities in Table S1 to produce the expected number of compounds to be successfully predicted by random. The figures for each data set split are then averaged to produce the figures listed in Table 1 of the paper.

The Expected Uniform Random AUC on a ROC curve is 0.5 by definition (random is effectively the equal likelihood of picking a true positive over a false positive and hence is a diagonal line from the bottom left to top right corner of a ROC plot). One can demonstrate this within this data set by examining the case where a single major isoform is associated with a compound, the AUC values for

picking the correct isoform on the first, second, third, etc guess are 1, 5/6, 4/6, 3/6, 2/6, 1/6, 0 respectively. Summing the possible AUC values and multiplying by the equal probability of each outcome gives (21/6)*1/7 = 3/6 = 0.5. The AUC values can be calculated for every possible isoform selection scenario and number of major isoforms for a compound in a similar fashion and again the average AUC can be shown to be 0.5.

Performance Including Minor Isoforms

The inclusion of minor observed isoforms in the test set compounds produces the distributions of numbers of isoforms associated with each compound shown in Figure S1.

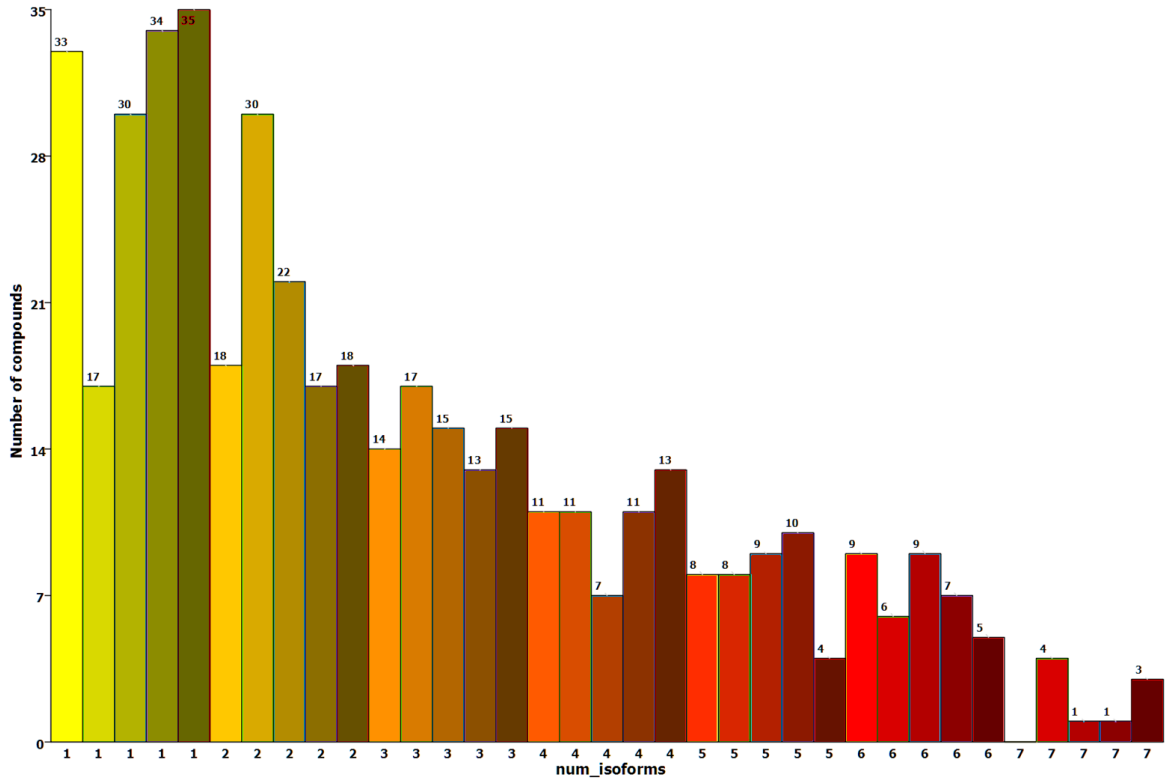

Figure S1. Distributions of the numbers of isoforms identified as major or minor metabolising enzymes for each compound in each test set split. The shading of the bars in each group indicates the 5 different test set splits.

Note the fact that the number of compounds with only 1 isoform has halved in these test sets compared to the test sets containing only major isoform information. The number of compounds with 3 or more isoforms has risen sharply and there are a small number of compounds associated with all 7 isoforms.

Figure S2 shows the distribution of observed isoforms when the minor isoforms are included for the test set compounds. Again, the CYPs are ordered by the frequency order as noted in Zanger *et al.* and one can see that the frequencies for the CYPs are now more in line with that seen by Zanger with the most frequent on the left to the least on the right.

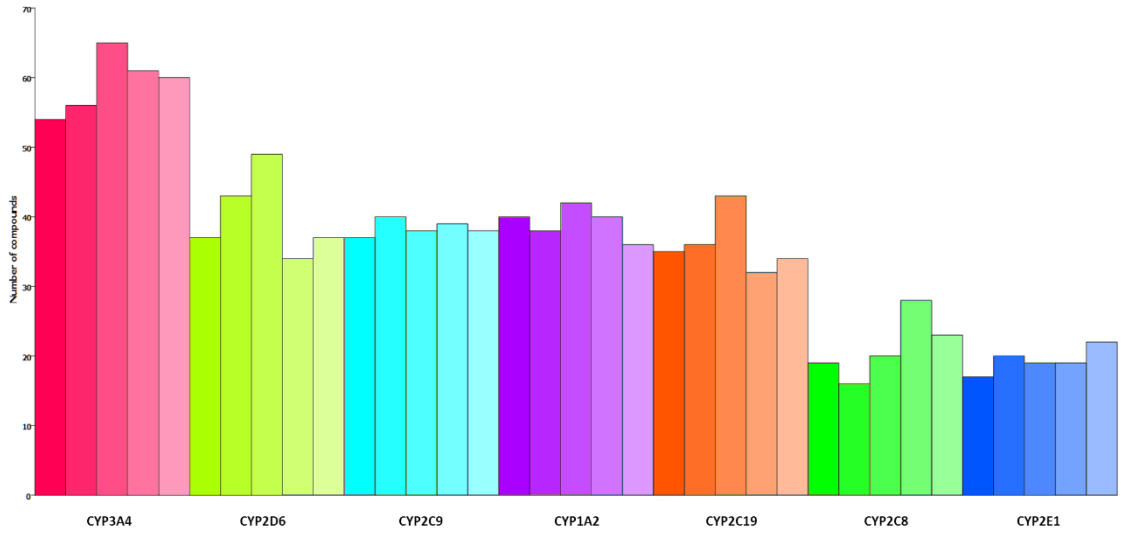

Figure S2. Distributions of the numbers of compounds metabolised by each isoform across the different test sets, when minor isoforms are included. The CYP450s are coloured consistently with Figure 2 and ordered by the frequency with which they are observed to metabolise marketed drugs, from the most common (left) to least common (right), as determined by Zanger *et al.*[2] The graduated colours indicate the 5 different set splits

As expected, the prediction performance increases to almost 90% for the top-1 criterion and well over 90% for the top-2 and top-3 criteria (as shown in Table S2), but the AUC values do not change significantly. This is due to the fact that missing an isoform by one place "costs" more in terms of AUC value when there are more isoforms listed, i.e. if a single isoform is misplaced by one place then the reduction in AUC value is 1/6, but if an incorrect isoform is placed ahead of isoforms from a list of 3 possible isoforms then the AUC value reduction is 1/4. So, as the prediction gets easier, the cost for making a wrong prediction gets greater.

Table S2. Top-*k* (*k*=1-3*)* and AUC results on the independent test sets for the models built and tested with each of the training/test set splits, now with minor isoforms included.

Set Split	Top-1 performance	Top-2 performance	Top-3 performance	Average AUC
1	81.7	93.5	96.8	0.85
2	94.6	97.8	98.9	0.81
3	82.8	93.5	97.8	0.84
4	92.5	98.9	98.9	0.85
5	90.3	98.9	98.9	0.84
Average	88.4	96.5	98.3	0.84

Table S3 shows comparisons between the models and the different random methods for the three top-*k* metrics and AUC. Although the chance of successful prediction has gone up with the inclusion of the minor isoforms, the models are still performing significantly better than any random method.

Table S3 – Summary of the statistics for the top-*k* and ROC AUC performances of the models and the four random methods. For each the average and standard deviation over the five data set splits is shown when taking the minor isoforms also into consideration.

Method	*Models*	*Expected Uniform Random*	*Uniform Random*	*Guided Random*	*Y-shuffled*	*Y-Scrambled*
Top-1 (%)	88.4 ± 5.2	34.7 ± 1.5	38.8 ± 4.4	45.6 ± 4.5	53.4 ± 4.2	53.1 ± 4.4
Top-2 (%)	96.5 ± 2.5	57.3 ± 2.2	59.6 ± 4.3	67.2 ± 4.1	76.3 ± 3.9	76.3 ± 3.9
Top-3 (%)	98.3 ± 0.9	67.3 ± 2.3	72.9 ± 3.9	80.0 ± 3.5	88.3 ± 3.5	87.8 ± 3.4
AUC	0.84 ± 0.02	0.5 ± 0.0	0.49 ± 0.007	0.58 ± 0.008	0.60 ± 0.004	0.60 ± 0.005

T-test evaluations using Excel

To assess the significance of the results for AUC and top-*k* performances, the "t-test: Two sample assuming unequal variances" tool in the Data Analysis tool pack within Excel was used on the following data.

Data Set Split	*Method*	*AUC*	*Top-1*	*Top-2*	*Top-3*
Split1	Model	0.86	71.00	82.80	90.30
Split2	Model	0.87	81.70	87.10	89.20
Split3	Model	0.86	68.80	83.90	93.50
Split4	Model	0.93	80.60	95.70	98.90
Split5	Model	0.91	79.60	92.50	94.60
Split1	Yscrambled Random	0.58	24.69	46.46	58.84
Split2	Yscrambled Random	0.63	30.67	54.56	70.03
Split3	Yscrambled Random	0.6	31.16	49.67	63.49
Split4	Yscrambled Random	0.61	29.81	50.29	65.56
Split5	Yscrambled Random	0.62	28.65	49.17	64.70
Split1	Shuffled Random	0.58	24.60	45.72	59.83
Split2	Shuffled Random	0.62	29.43	53.33	69.03
Split3	Shuffled Random	0.61	30.64	50.13	64.91
Split4	Shuffled Random	0.62	29.08	49.98	65.36
Split5	Shuffled Random	0.61	27.45	48.29	63.85
Split1	Uniform Random	0.5	18.19	34.50	49.53
Split2	Uniform Random	0.5	21.19	39.26	54.33
Split3	Uniform Random	0.5	20.13	37.44	52.58
Split4	Uniform Random	0.5	19.22	36.30	51.70
Split5	Uniform Random	0.5	18.94	35.58	50.98
Split1	Guided Random	0.58	22.63	42.14	58.42
Split2	Guided Random	0.6	27.40	49.08	66.04

Split3	Guided Random	0.58	25.69	45.47	61.69
Split4	Guided Random	0.59	25.26	45.67	62.36
Split5	Guided Random	0.59	24.42	44.69	61.39

The comparisons of the performances across the five splits in each of the top-*k* and AUC criteria were made for the real versus the random methodologies. The output from each of the t-tests is shown below.

AUC Models vs Uniform Random	*Variable 1*	*Variable 2*
Mean	0.5	0.886
Variance	0	0.00103
Observations	5	5
Hypothesized Mean Difference	0	
Df	4	
t Stat	-26.89389435	
P(T<=t) one-tail	5.68217E-06	
t Critical one-tail	2.131846786	
P(T<=t) two-tail	1.13643E-05	
t Critical two-tail	2.776445105	

AUC Models vs Guided Random	*Variable 1*	*Variable 2*
Mean	0.588	0.886
Variance	7E-05	0.00103
Observations	5	5
Hypothesized Mean Difference	0	
Df	5	
t Stat	-20.0911559	
P(T<=t) one-tail	2.82352E-06	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	5.64703E-06	
t Critical two-tail	2.570581836	

AUC Models vs Y-scrambled Random	*Variable 1*	*Variable 2*
Mean	0.608	0.886
Variance	0.00037	0.00103
Observations	5	5
Hypothesized Mean Difference	0	
Df	7	
t Stat	-16.61367767	
P(T<=t) one-tail	3.49653E-07	

t Critical one-tail	1.894578605
P(T<=t) two-tail	6.99306E-07
t Critical two-tail	2.364624252

AUC Models vs Y-shuffled	*Variable 1*	*Variable 2*
Mean	0.608	0.886
Variance	0.00027	0.00103
Observations	5	5
Hypothesized Mean Difference	0	
Df	6	
t Stat	-17.24082811	
P(T<=t) one-tail	1.21937E-06	
t Critical one-tail	1.943180281	
P(T<=t) two-tail	2.43875E-06	
t Critical two-tail	2.446911851	

Top-1 Models vs Uniform Random	*Variable 1*	*Variable 2*
Mean	19.53247312	76.34
Variance	1.338130651	35.718
Observations	5	5
Hypothesized Mean Difference	0	
Df	4	
t Stat	-20.86704009	
P(T<=t) one-tail	1.55832E-05	
t Critical one-tail	2.131846786	
P(T<=t) two-tail	3.11665E-05	
t Critical two-tail	2.776445105	

Top-1 Models vs Guided Random	*Variable 1*	*Variable 2*
Mean	25.07913978	76.34
Variance	3.057152041	35.718
Observations	5	5
Hypothesized Mean Difference	0	
Df	5	
t Stat	-18.40747153	
P(T<=t) one-tail	4.35179E-06	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	8.70357E-06	
t Critical two-tail	2.570581836	

Top-1 Models vs Y-scrambled	*Variable 1*	*Variable 2*
Mean	28.996	76.34
Variance	6.70228	35.718
Observations	5	5
Hypothesized Mean Difference	0	
Df	5	
t Stat	-16.25410963	
P(T<=t) one-tail	8.03535E-06	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	1.60707E-05	
t Critical two-tail	2.570581836	

Top-1 Models vs Y-Shuffled	*Variable 1*	*Variable 2*
Mean	28.24	76.34
Variance	5.43885	35.718
Observations	5	5
Hypothesized Mean Difference	0	
Df	5	
t Stat	-16.76521065	
P(T<=t) one-tail	6.89945E-06	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	1.37989E-05	
t Critical two-tail	2.570581836	

Top-2 Models vs Uniform Random	*Variable 1*	*Variable 2*
Mean	36.61548387	88.4
Variance	3.329625159	30.85
Observations	5	5
Hypothesized Mean Difference	0	
Df	5	
t Stat	-19.80620533	
P(T<=t) one-tail	3.03025E-06	
t Critical one-tail	2.015048373	
P(T<=t) two-tail	6.0605E-06	
t Critical two-tail	2.570581836	

Top-2 Models vs Guided Random	*Variable 1*	*Variable 2*

Mean	45.40989247	88.4
Variance	6.177827726	30.85
Observations	5	5
Hypothesized Mean Difference	0	
Df	6	
t Stat	-15.79753828	
P(T<=t) one-tail	2.04007E-06	
t Critical one-tail	1.943180281	
P(T<=t) two-tail	4.08014E-06	
t Critical two-tail	2.446911851	

Top-2 Models vs Y-scrambled	*Variable 1*	*Variable 2*
Mean	50.03	88.4
Variance	8.55065	30.85
Observations	5	5
Hypothesized Mean Difference	0	
Df	6	
t Stat	-13.66863379	
P(T<=t) one-tail	4.76292E-06	
t Critical one-tail	1.943180281	
P(T<=t) two-tail	9.52584E-06	
t Critical two-tail	2.446911851	

Top-2 Models vs Y-shuffled	*Variable 1*	*Variable 2*
Mean	49.49	88.4
Variance	7.76205	30.85
Observations	5	5
Hypothesized Mean Difference	0	
Df	6	
t Stat	-14.00183008	
P(T<=t) one-tail	4.13795E-06	
t Critical one-tail	1.943180281	
P(T<=t) two-tail	8.27591E-06	
t Critical two-tail	2.446911851	

Top-3 Models vs Uniform Random	*Variable 1*	*Variable 2*
Mean	51.82365591	93.3
Variance	3.217248237	14.725
Observations	5	5

Hypothesized Mean Difference	0
Df	6
t Stat	-21.89510541
P(T<=t) one-tail	2.9649E-07
t Critical one-tail	1.943180281
P(T<=t) two-tail	5.92979E-07
t Critical two-tail	2.446911851

Top-3 Models vs Guided Random	*Variable 1*	*Variable 2*
Mean	61.97978495	93.3
Variance	7.433124292	14.725
Observations	5	5
Hypothesized Mean Difference	0	
Df	7	
t Stat	-14.87795499	
P(T<=t) one-tail	7.42785E-07	
t Critical one-tail	1.894578605	
P(T<=t) two-tail	1.48557E-06	
t Critical two-tail	2.364624252	

Top-3 Models vs Y-scrambled	*Variable 1*	*Variable 2*
Mean	64.524	93.3
Variance	16.19933	14.725
Observations	5	5
Hypothesized Mean Difference	0	
Df	8	
t Stat	-11.57085037	
P(T<=t) one-tail	1.41412E-06	
t Critical one-tail	1.859548038	
P(T<=t) two-tail	2.82823E-06	
t Critical two-tail	2.306004135	

Top-3 Models vs Y-shuffled	*Variable 1*	*Variable 2*
Mean	64.596	93.3
Variance	10.90348	14.725
Observations	5	5
Hypothesized Mean Difference	0	
Df	8	

t Stat	-12.67844523
P(T<=t) one-tail	7.04193E-07
t Critical one-tail	1.859548038
P(T<=t) two-tail	1.40839E-06
t Critical two-tail	2.306004135

Main data set

The data set (compounds and associated isoforms) has been published previously for the reference

"Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations" - Jonathan D. Tyzack, Peter A. Hunt, and Matthew D. Segall*

J. Chem. Inf. Model., 2016, 56 (11), pp 2180–2193 - DOI: 10.1021/acs.jcim.6b00233

and is available in the supplementary information following the link below

https://pubs.acs.org/doi/suppl/10.1021/acs.jcim.6b00233

29 Compound further external test set

smiles	name	major	minor
CCC1=CC=C(CCOC2=CC=C(CC3SC(=O)NC3=O)C=C2)N=C1	pioglitazone	CYP2C8	CYP3A4
N#CCC(C1CCCC1)[N]2C=C(C=N2)C3=C4C=C[NH]C4=NC=N3	ruxolitinib	CYP3A4	
C1CCN2C[C@@H]3C[C@@H](CN4CCCC[C@H]34)[C@@H]2C1	sparteine	CYP2D6	
N[S](=O)(=O)NCCNC1=NON=C1C(=N)NC2=CC=C(F)C(=C2)Br	epacadostat_M11	CYP3A4	CYP2C19; CYP1A2
FC(F)OC(Cl)C(F)(F)F	isoflurane	CYP2E1	
OC(CCCN1CCC(O)(CC1)C2=CC=C(Br)C=C2)C3=CC=C(F)C=C3	dihydrobromperidol	CYP3A4	
OC1(CCN(CCCC(=O)C2=CC=C(F)C=C2)CC1)C3=CC=C(Br)C=C3	bromperidol	CYP3A4	
C[S](=O)(=O)CCNCC1=CC=C(O1)C2=CC=C3N=CN=C(NC4=CC=C(OCC5=CC(=CC=C5)F)C(=C4)Cl)C3=C2	lapatinib	CYP3A4	CYP3A5
C1=CC=C2C(=C1)C=CC3=CC4=C5C=CC=CC5=CC=C4C=C23	dibenzo_ah_anthracene	CYP1A2; CYP2C9	CYP2B6
CNC(=O)C1=C(F)C=C(C=C1)N2C(=S)N(C(=O)C2(C)C)C3=CC=C(C#N)C(=C3)C(F)(F)F	enzalutamide	CYP2C8	CYP3A4;CYP3A5
[H][C@@]12OC3=C(OC)C=CC4=C3[C@@]11CCN(C)[C@]([H])(C4)[C@]1([H])CC[C@@H]2O	dihydrocodeine	CYP3A4	CYP2D6
CC1(C)CCC(=C(C1)C2=CC=C(Cl)C=C2)CN3CCN(CC3)C4=CC(=C(C=C4)C(=O)N[S](=O)(=O)C5=CC(=C(NCC6CCOCC6)C=C5)[N](=O)=O)OC7=CN=C8[NH]C=CC8=C7	venetoclax	CYP3A4	
CC1(C)CC(O)C(=C(C1)C2=CC=C(Cl)C=C2)CN3CCN(CC3)C4=CC(=C(C=C4)C(=O)N[S](=O)(=O)C5=CC(=C(NCC6CCOCC6)C=C5)[N](=O)=O)OC7=CN=C8[NH]C=CC8=C7	venetoclax_M5	CYP3A4	

CC1(CO)CCC(=C(C1)C2=CC=C(Cl)C=C2)CN3CCN(CC3)C4=CC(=C(C=C4)C(=O)N[S](=O)(=O)C5=CC(=C(NCC6CCOCC6)C=C5)[N](=O)=O)OC7=CN=C8[NH]C=CC8=C7	venetoclax_M2	CYP3A4	
CC(C)(C)C1=NC(=C(S1)C2=CC=NC(=N2)N)C3=CC=CC(=C3F)N[S](=O)(=O)C4=C(F)C=CC=C4F	dabrafenib	CYP2C8; CYP3A4	
CC(C)(CO)C1=NC(=C(S1)C2=CC=NC(=N2)N)C3=CC=CC(=C3F)N[S](=O)(=O)C4=C(F)C=CC=C4F	Hydroxy_dabrafenib	CYP3A4	
CCC1=C(C)C=C(C(=O)NC2(CCCCC2)C(O)=O)[C](=O)[N]1CC3=CC=C(F)C=C3	S-777469	CYP2C9	
CCC1=C(CO)C=C(C(=O)NC2(CCCCC2)C(O)=O)[C](=O)[N]1CC3=CC=C(F)C=C3	S-777469_5HM	CYP2C9	
O=C1NC(=O)[C@H]([C@@H]1C2=C[NH]C3=CC=CC=C23)C4=C[N]5CCCC6=CC=CC4=C56	tivatinib	CYP2C19;CYP3A4;CYP3A5	
CO[C@@H]1[C@@H](C[C@H]2O[C@]1(C)[N]3C4=CC=CC=C4C5=C3C6=C(C7=C(C=CC=C7)[N]26)C8=C5CNC8=O)N(C)C(=O)C9=CC=CC=C9	midostaurin	CYP3A4	
C[C@]12CC[C@H](O)CC1CCC3C2CC[C@]4(C)[C@H](CC[C@]34O)C5=CO[C](=O)C=C5	bufalin	CYP3A4	
CCC1=C(C=C2C(=C1)C(=O)C3=C([NH]C4=CC(=CC=C34)C#N)C2(C)C)N5CCC(CC5)N6CCOCC6	alectinib	CYP3A4	
CC1CCN(CCN1C(=O)C2=CC(=CC=C2[N]3N=CC=N3)C)C4=NC5=CC(=CC=C5O4)Cl	suvorexant	CYP2C19;CYP3A4	
O=[C]1C=CC2=C3[N]1C[C@@H](CN4CCC(CC4)NCC5=NC=C6OCCCC6=C5)[N]3[C](=O)C=N2	GSK2140944	CYP3A4	
CNCC1=C[N](C(=C1)C2=CC=CC=C2F)[S](=O)(=O)C3=CN=CC=C3	Vonoprazan_TAK438	CYP3A4	CYP2B6;CYP2D6;CYP2C19
CCOC1=CC=C(C=C1)[N]2[C](=O)C3=CC=CN=C3N=C2[C@@H](C)N(CC4=CC=CN=C4)C(=O)CC5=CC=C(OC(F)(F)F)C=C5	AMG487	CYP3A4	
C[C@@H](N(CC1=CC=CN=C1)C(=O)CC2=CC=C(OC(F)(F)F)C=C2)C3=NC4=NC=CC=C4[C](=O)[N]3C5=CC=C(O)C=C5	AMG487_M2	CYP3A4	
CN1C[C@@]2(C=C)[C@@H]3C[C@H]4OC[C@@H]3[C@@H]1C[C@]25C4=NC6=CC=CC=C56	koumine	CYP3A4;CYP3A5	
OC1=CC=C(Cl)C=C1C(=O)NC2=C(Cl)C=C(C=C2)[N](=O)=O	Niclosamide	CYP1A2	