Ligand and structure-based classification models for Prediction of P-glycoprotein inhibitors

Freya Klepsch[§,‡,†], *Poongavanam Vasanthanathan*[§,†], *Gerhard F Ecker*[§,*]

[§]University of Vienna, Department of Medicinal Chemistry, Althanstraße 14, 1090 Vienna, Austria.

[‡]Current address: CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, 1090 Vienna, Austria

* Email: gerhard.f.ecker@univie.ac.at

[†] These authors contributed equally to this manuscript.

Supplementary material

SI-Table1: List of physicochemical properties used for PCA

Properties	Description	Properties	Description
logP(o/w)	Log of the partition coefficient (octanol/water)	**a_count**	Number of atoms
SMR	Molar refractivity (including implicit hydrogens)	**a_heavy**	Number of heavy atoms
TPSA	Total polar surface area	**a_hyd**	Number of hydrophobic atoms
weight	Molecular weight	**a_nC**	Number of carbon atoms
apol	Sum of the atomic polarizabilities	**a_nH**	Number of hydrogen atoms
bpol	Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms	**a_nO**	Number of oxygen atoms
density	Density	**b_ar**	Number of aromatic bonds
logs	Log of the solubility in water	**b_count**	Number of bonds
MR	Molar refractivity	**b_double**	Number of double bonds
pmi	Principal moment of inertia	**b_heavy**	Number of bonds between heavy atoms
pmiX	x component of pmi	**b_single**	Number of single bonds
pmiY	y component of pmi	**diameter**	Diameter
pmiZ	z component of pmi	**lip_acc**	Number of O and N atoms
vdw_area	Area of van der Waals surface	**lip_don**	Number of OH and NH atoms
vdw_vol	Van der Waals volume	**radius**	Radius
vol	Volume		

SI-Table 2: Results of training set prediction and 10 fold cross-validation

Validation	Descriptors	Models	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	G-Mean	MCC
Training Set	MOE	RF	840	355	5	1	1.00	0.99	1.00	0.99	0.99
		SVM	787	204	156	54	0.94	0.57	0.83	0.73	0.56
		KNN	841	360	0	0	1.00	1.00	1.00	1.00	1.00
		B-QSAR	751	270	90	90	0.89	0.75	0.85	0.82	0.64
	MACCS	RF	799	241	119	42	0.95	0.67	0.87	0.80	0.67
		SVM	785	142	218	56	0.93	0.39	0.77	0.61	0.40
		KNN	809	231	129	32	0.96	0.64	0.87	0.79	0.67
		B-QSAR	616	255	105	225	0.73	0.71	0.73	0.72	0.41
	SS-FP	RF	812	164	196	29	0.97	0.46	0.81	0.66	0.53
		SVM	807	141	219	34	0.96	0.39	0.79	0.61	0.46
		KNN	816	160	200	25	0.97	0.44	0.81	0.66	0.53
		B-QSAR	725	191	169	116	0.86	0.53	0.76	0.68	0.41
	Combined	RF	841	356	4	0	1.00	0.99	1.00	0.99	0.99
		SVM	788	223	137	53	0.94	0.62	0.84	0.76	0.61
		KNN	841	360	0	0	1.00	1.00	1.00	1.00	1.00
		B-QSAR	732	261	109	99	0.88	0.71	0.83	0.79	0.59

Training 10 fold CV	MOE	RF	780	245	115	61	0.93	0.68	0.85	0.79	0.64
		SVM	788	199	161	53	0.94	0.55	0.82	0.72	0.55
		KNN	746	258	102	95	0.89	0.72	0.84	0.80	0.61
		B-QSAR	747	266	94	94	0.89	0.74	0.84	0.81	0.63
	MACCS	RF	751	197	163	90	0.89	0.55	0.79	0.70	0.47
		SVM	782	137	223	59	0.93	0.38	0.77	0.59	0.38
		KNN	770	178	182	71	0.92	0.49	0.79	0.67	0.46
		B-QSAR	616	255	105	225	0.73	0.71	0.73	0.72	0.41
	SS-FP	RF	795	142	218	46	0.95	0.39	0.78	0.61	0.43
		SVM	805	140	220	36	0.96	0.39	0.79	0.61	0.45
		KNN	800	130	230	41	0.95	0.36	0.77	0.59	0.41
		B-QSAR	725	191	169	116	0.86	0.53	0.76	0.68	0.41
	Combined	RF	780	253	107	61	0.93	0.70	0.86	0.81	0.66
		SVM	786	216	144	55	0.93	0.60	0.83	0.75	0.59
		KNN	744	251	109	97	0.88	0.70	0.83	0.79	0.59
		B-QSAR	727	256	104	114	0.86	0.71	0.82	0.78	0.57

SI-Table 3: List of MACCS fingerprints that contributed to the MACCS FP based models and their frequency of occurrences in inhibitor and non-inhibitor

MACCS key		Description	Inhibitors [%]	Non-Inhibitors [%]
8	OAA@1	a 4-membered heterocycle	0.19	2.82
17	CTC	two carbon atoms connected via a triple bond	0.37	1.50
50	C=C(C)	propene	93.03	74.44
54	QHAAQH	two hetero atoms linked via two atoms	5.48	21.24
69	QQH	two successive hetero atoms	1.12	9.77
75	A!N$A	piperidine nitrogen	70.17	44.55
76	C=C(A)A	disubstituted ethene	98.51	90.60
84	NH2	primary amine	3.62	17.48
86	CH2QCH2	hetero atom connected to two aliphatic carbon atoms	72.03	40.04
102	QO	hetero atom bound to oxygen	2.70	11.28
112	AA(A)(A)A	branched substructure of 5 atoms of any type	99.63	94.36
125	aromatic ring > 1	more than one aromatic ring	84.85	49.44
129	ACH2AACH2A	chain of 6 atoms with the second and the fourth being aliphatic carbon atoms	75.74	48.50
139	OH	hydroxyl group	27.97	50.75
145	6M ring > 1	more than one 6-membered ring	92.94	70.11

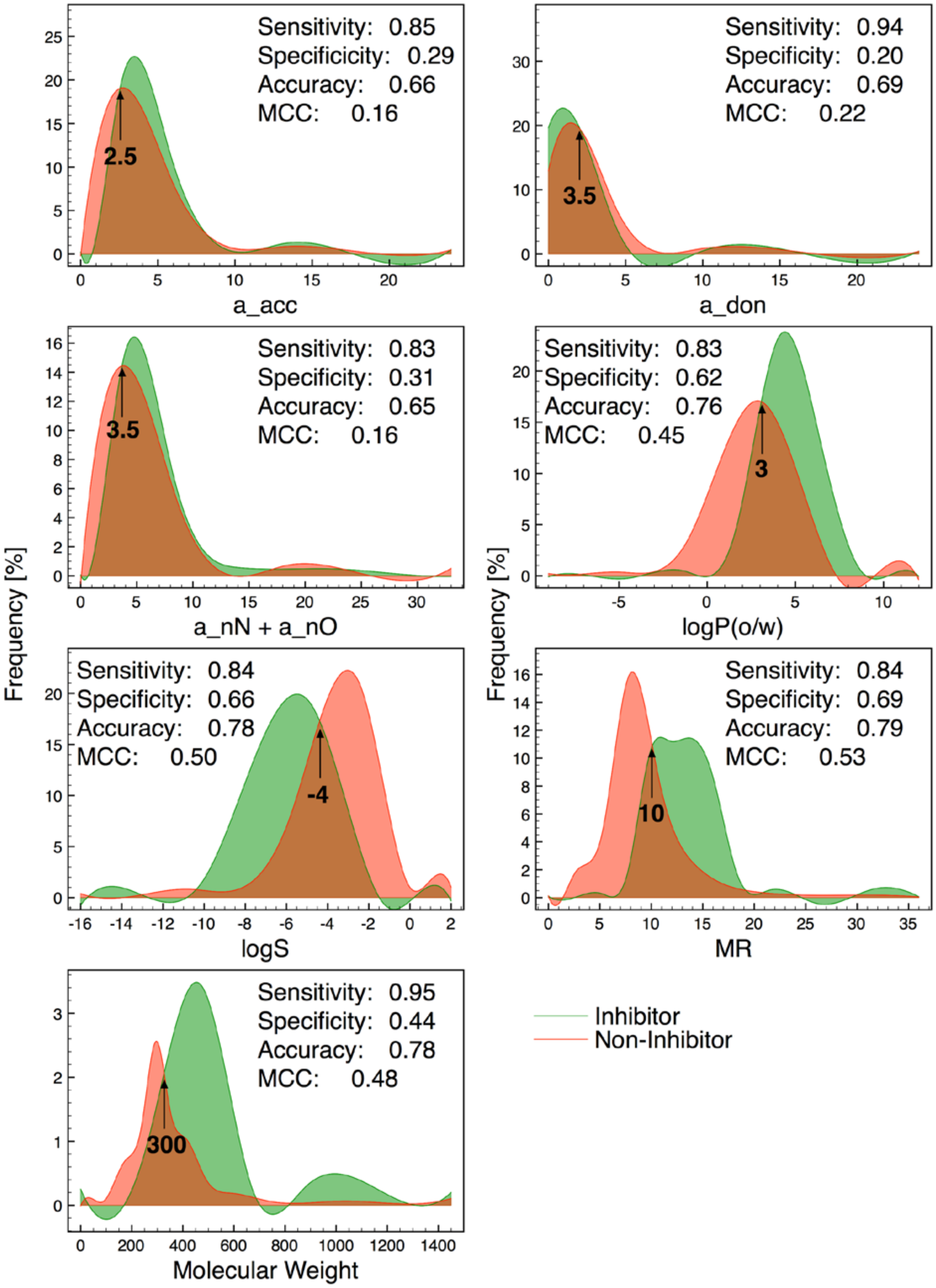
155 A!CH2!A	an aliphatic carbon atom connected with two atoms of any type	94.70	79.70
162 AROMATIC	at least one aromatic ring	94.70	81.58

SI-Table 4: List of substructure fingerprints that contributed to the substructure FP based models and their frequency of occurrences in inhibitor and non-inhibitor

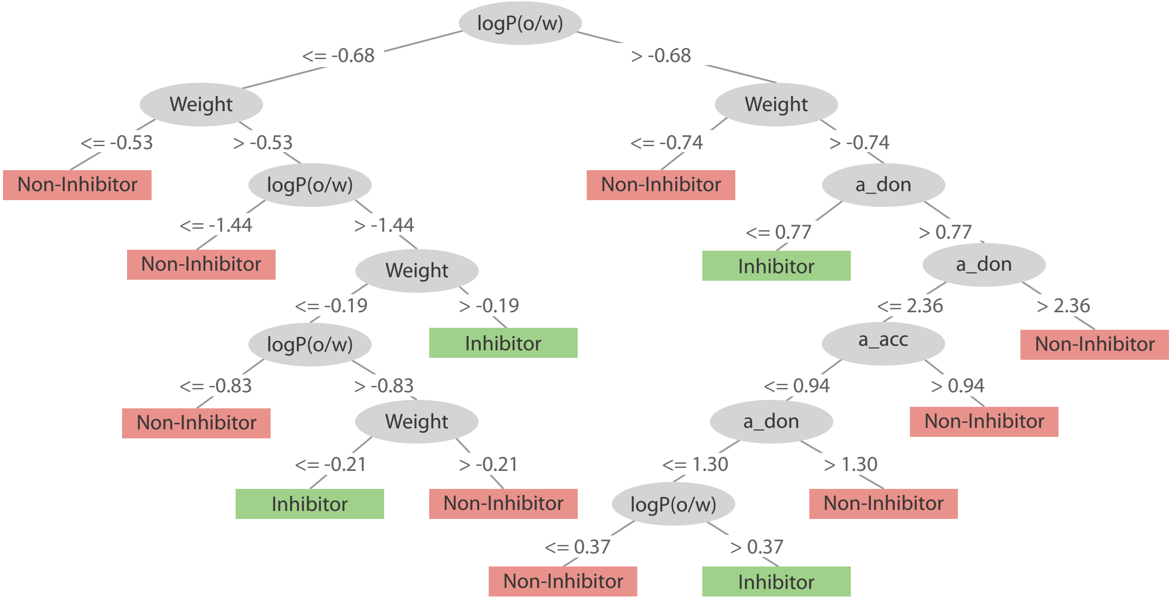
Substructure Fingerprint	Description	Inhibitor [%]	Non-Inhibitor [%]
SubFP2	secondary carbon	78.53	55.26
SubFP6	alkine	0.37	1.50
SubFP18	alkylarylether	48.42	24.44
SubFP23	amine	60.59	32.33
SubFP41	1,2-diol	0.56	5.83
SubFP84	carboxylic acid	0.93	13.91
SubFP90	carbothioic S ester	0.09	0.38
SubFP128	peptide C term	0.09	3.01
SubFP151	guanidine	0.09	1.88
SubFP169	phenol	6.13	15.23
SubFP170	1,2-diphenol	0.19	3.01
SubFP172	arylfluoride	7.53	2.26
SubFP214	sulfonic derivative	1.21	6.77
SubFP274	aromatic	94.70	81.58
SubFP287	conjugated double bond	96.84	87.41
SubFP302	rotatable bond	98.88	91.54

SI-Table 5: Classification performance of models based on physicochemical properties, performed on the Dolghih et al. Dataset

Model	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy	G-Mean	MCC
logP	8	101	10	16	0.33	0.91	0.81	0.55	0.27
MW	23	86	25	1	0.96	0.77	0.81	0.86	0.59
logS	10	99	12	14	0.42	0.89	0.81	0.61	0.32
MR	18	100	11	6	0.75	0.90	0.87	0.82	0.61

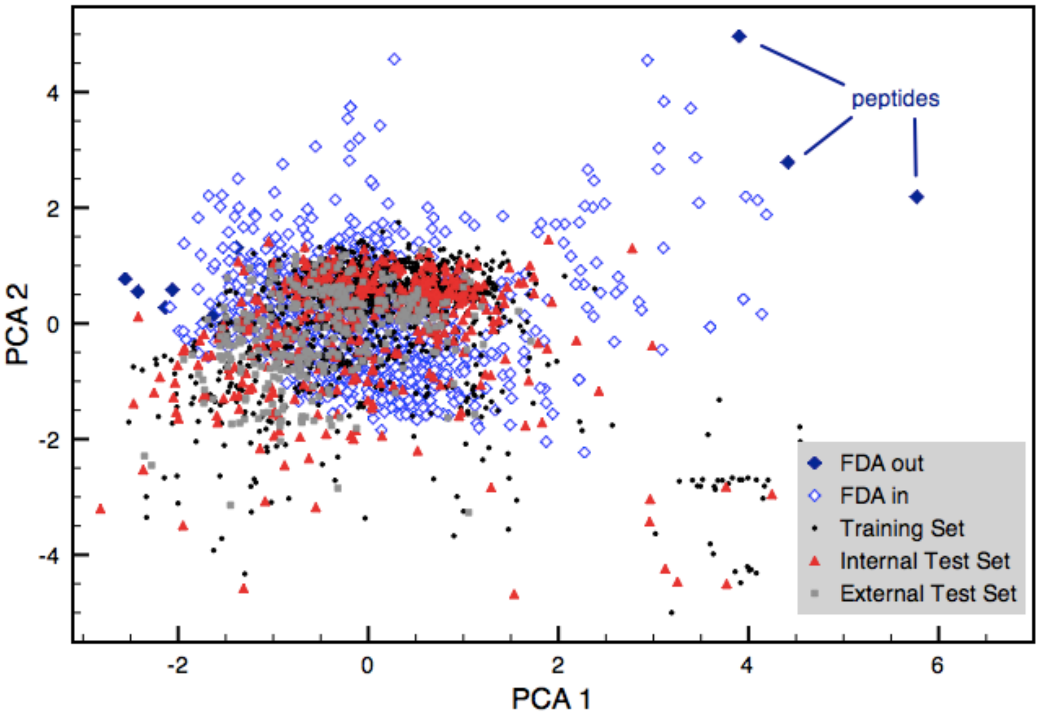


SI-Figure 1: Distribution plots from physicochemical properties



SI-Figure 2: Decision tree generated by using "Rule-of-Five" descriptors.

Note: Inhibitor denoted as 1 and non-inhibitor denoted as 0.



SI-Figure 3: Applicability domain experiment using "ED approach" Compounds shown as follows: Training compounds: Gray dots, FDA compounds: red square, Test compounds: Black cross.