Fingerprint-based *in silico* models for the prediction of P-glycoprotein substrates and inhibitors

Poongavanam Vasanthanathan [a], Norbert Haider [b], and Gerhard F Ecker [a], *

[a] University of Vienna, Department of Medicinal Chemistry, Althanstrasse 14, 1090 Vienna, Austria

[b] University of Vienna, Department of Drug and Natural Product Synthesis, Althanstrasse 14, 1090 Vienna, Austria

Supplementary material

T1: List of physicochemical properties used for PCA

Properties	Description
apol	Sum of atomic polarizabilities
a_acc	Number of hydrogen bond acceptor atoms
a_acid	Number of acidic atoms
b_ar	Number of aromatic bonds
b_count	Number of bonds
b_double	Number of double bonds
b_heavy	Number of heavy-heavy bonds
LogP (o/w)	Log octanol/water partition coefficient
LogS	Log of solubility in water
MR	Molar refractivity
Radius	Smallest vertex eccentricity in graph
Reactive	Reactivity
Rings	Number of rings
TPSA	Topological polar surface area
Weight	Molecular weight
Zagreb	Zagreb index

T2: List of descriptors contributed to the P-gp Substrate and Inhibitor models

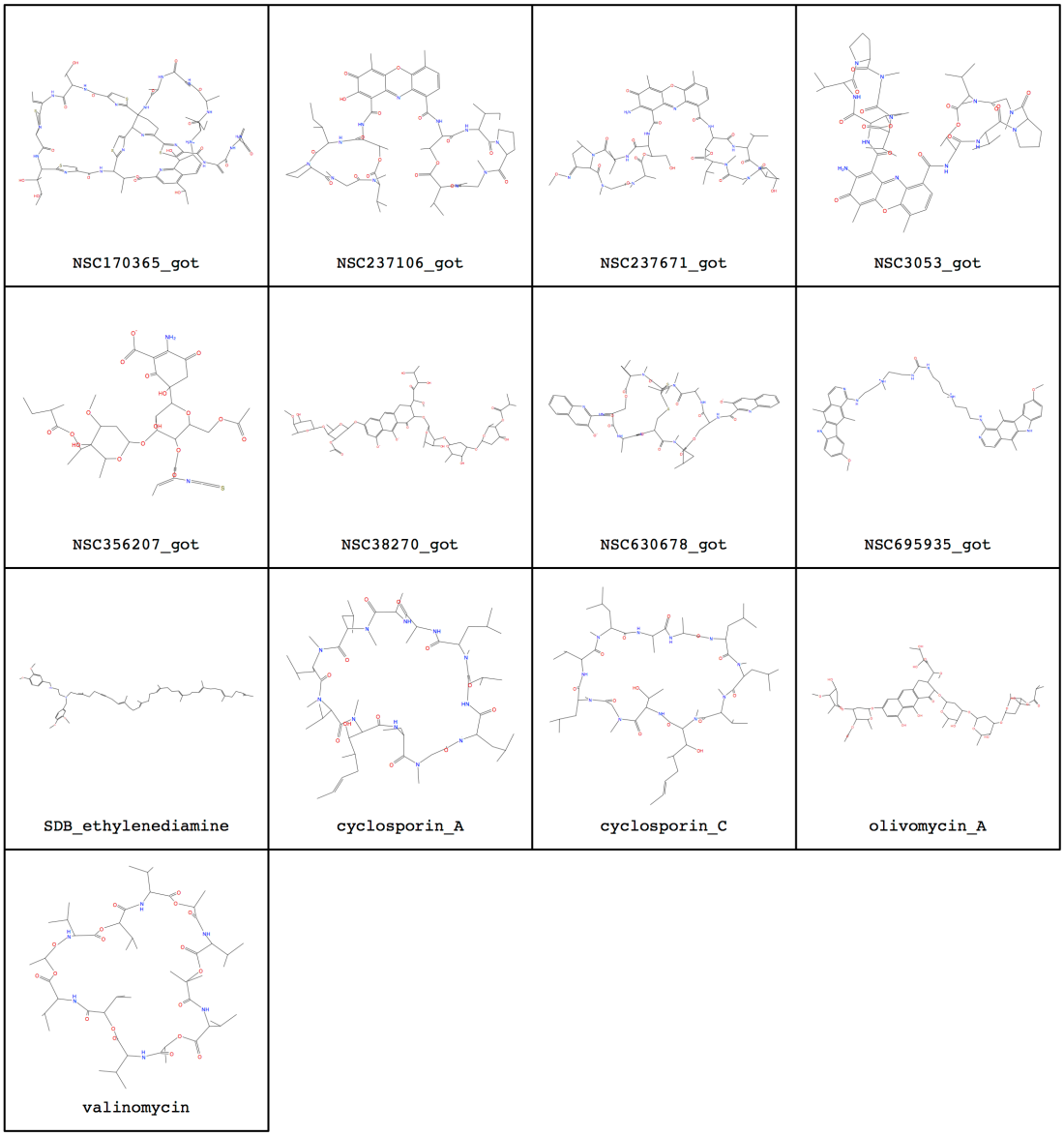
Models	Functional groups	Bin Number
Substrate/Non substrate	Cation	1
	Enol	25
	Hydroxy compounds	27
	Alcohol	28
	Primary Alcohol	29
	Secondary Alcohol	30
	1, 2-Diol	32
	Secondary Aliphatic Amine	52
	Secondary Aromatic Amine	54
	Tertiary Aliphatic Amine	56
	Lactam	84
	Carboxylic Acid Azides	86
	Nitrile	90
Inhibitor/Non-inhibitor	Carbony Compound	3
	Hydrazone	10
	Oxime Ether	14
	Acetal	19
	Hemiaminal	20
	Aminal	21
	Secondary Alcohol	30
	1, 2-Diol	32
	Phenol	34
	Alkyl Aryl ether	39
	Amine	47
	Primary Amine	48
	Secondary Aliphatic Amine	52
	Sec Alkyl Aryl Amine	53
	Tertiary Aliphatic Amine	56
	Carboxylic Acid	76
	Carboxylic Acid Ester	78
	Carboxylic Acid Secondary Amide	82
	Lactam	84
	Thiocarboxylic Acid Ester	101
	Iminohetarene	108
	Urea	133
	Thiourea	135
	Sulfonamide	164
	Aromatic compound	201
	Heterocyclic Compound	202

T3: Summary of obtained models from substrates and inhibitor classification

Models	No	Models	Dataset	TP	FN	TN	FP	Sen.	Spec	NPP	PPP	G-mean	AUC	F-Measure	BCR	MCC	Accuracy
Substrate	13	KNN	Training	105	37	102	38	0.74	0.73	0.73	0.73	0.73	0.73	0.74	0.73	0.47	0.73
			Test	75	26	41	60	0.74	0.41	0.61	0.56	0.55	0.57	0.64	0.57	0.16	0.57
			10-fold	188	55	167	74	0.77	0.69	0.75	0.72	0.73	0.73	0.74	0.73	0.47	0.73
		SVM	Training	91	51	104	36	0.64	0.74	0.67	0.72	0.69	0.69	0.68	0.69	0.39	0.69
			Test	67	34	57	44	0.66	0.56	0.63	0.60	0.61	0.61	0.63	0.61	0.23	0.61
			10-fold	152	91	159	82	0.63	0.66	0.64	0.65	0.64	0.64	0.64	0.64	0.29	0.64
		RF	Training	106	36	120	20	0.75	0.86	0.77	0.84	0.80	0.80	0.79	0.80	0.61	0.80
			Test	73	28	69	32	0.72	0.68	0.71	0.70	0.70	0.70	0.71	0.70	0.41	0.70
			10-fold	179	64	182	59	0.74	0.76	0.74	0.75	0.75	0.75	0.74	0.75	0.49	0.75
Inhibitors	26	KNN	Training	825	56	230	157	0.94	0.59	0.80	0.84	0.75	0.77	0.89	0.77	0.58	0.83
			Test	345	54	142	126	0.86	0.53	0.72	0.73	0.68	0.70	0.79	0.70	0.42	0.73
			10-fold	1153	127	378	277	0.90	0.58	0.75	0.81	0.72	0.74	0.85	0.74	0.51	0.79
		SVM	Training	800	81	190	197	0.91	0.49	0.70	0.80	0.67	0.70	0.85	0.70	0.45	0.78
			Test	345	54	129	139	0.86	0.48	0.70	0.71	0.65	0.67	0.78	0.67	0.38	0.71
			10-fold	1153	127	307	348	0.90	0.47	0.71	0.77	0.65	0.68	0.83	0.68	0.42	0.75
		RF	Training	845	36	289	98	0.96	0.75	0.89	0.90	0.85	0.85	0.93	0.85	0.74	0.89
			Test	334	65	168	100	0.84	0.63	0.72	0.77	0.72	0.73	0.80	0.73	0.48	0.75
			10-fold	1148	132	426	229	0.90	0.65	0.76	0.83	0.76	0.77	0.86	0.77	0.57	0.81

Abbreviations: No: Number of descriptors used for models, KNN: Kappa nearest neighbor, SVM: Support vector machine, RF: Random forest, TP: True positive, TN: True negative, FP: False positive, FN: False negative, Sen.: sensitivity, Spec.: Specificity, NPP: Negative predictive power, PPP: Positive predictive power, AUC: Area under curve, BCR: Balanced classification rate, MCC: Matthews correlation coefficient, 10-fold: 10-fold cross validation of whole dataset.

F1: Structures of Outliers, First eight compounds (NSC170365, NSC237106, NSC237671, NSC3053, NSC356207, NSC356207, NSC38270, NSC695935) were present in Szakács et al., dataset and last five compounds (SDB-ethylenediamine, Cyclosporine-A, Cyclosporin-C, Olivomycin-A, Valinomycin) were present in Literature dataset.



F2: Distribution of LogP (O/W) of substrate and non-substrate.

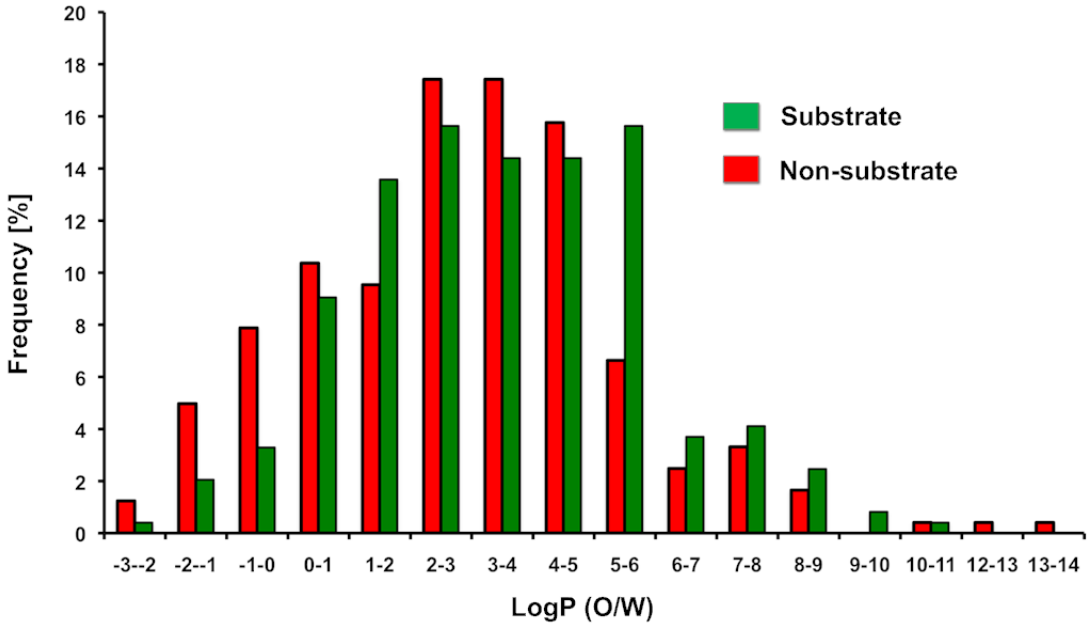


Fig.3: Applicability domain experiment shown in PCA plot