Supporting Information

Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters – An *in silico* Modeling Approach

Eleni Kotsampasakou and Gerhard F. Ecker*

University of Vienna, Department of Pharmaceutical Chemistry, Althanstrasse 14, 1090 Vienna, Austria

*e-mail: gerhard.f.ecker@univie.ac.at

List of Contents

Table S1: List of the 93 molecular 2D MOE descriptors and the 5 descriptors for BSEP, BCRP, P-gp, OATP1B1 and 1B3 inhibition prediction used for the cholestasis classification model for human data.

	MOE Descriptor	Description
1	apol	Sum of the atomic polarizabilities (including implicit hydrogens) with polarizabilities taken from [CRC 1994]

2	a_acc	Number of hydrogen bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
3	a_acid	Number of acidic atoms.
4	a_aro	Number of aromatic atoms.
5	a_count	Number of atoms (including implicit hydrogens). This is calculated as the sum of $(1 + h_i)$ over all non-trivial atoms *i*.
6	a_don	Number of hydrogen bond donor atoms (not counting basic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH).
7	a_donacc	Number of hydrogen bond donor and hydrogen bond acceptor atoms.
8	a_heavy	Number of heavy atoms $\#\{Z_i \mid Z_i > 1\}$.
9	a_hyd	Number of hydrophobic atoms.
10	a_IC	Atom information content (total). This is calculated to be a_ICM times *n*.
11	a_ICM	Atom information content (mean). This is the entropy of the element distribution in the molecule (including implicit hydrogens but not lone pair pseudo-atoms). Let n_i be the number of occurrences of atomic number *i* in the molecule. Let $p_i = n_i / n$ where *n* is the sum of the n_i. The value of a_ICM is the negative of the sum over all *i* of $p_i \log p_i$.
12	a_nBr	Number of bromine atoms: $\#\{Z_i \mid Z_i = 35\}$.
13	a_nC	Number of carbon atoms: $\#\{Z_i \mid Z_i = 6\}$.
14	a_nCl	Number of chlorine atoms: $\#\{Z_i \mid Z_i = 17\}$.
15	a_nF	Number of fluorine atoms: $\#\{Z_i \mid Z_i = 9\}$.
16	a_nH	Number of hydrogen atoms (including implicit hydrogens). This is calculated as the sum of h_i over all non-trivial atoms *i* plus the number of non-trivial hydrogen atoms.
17	A_nI	Number of iodine atoms: #{Zi | Zi = 53}
18	a_nN	Number of nitrogen atoms: $\#\{Z_i \mid Z_i = 7\}$.
19	a_nO	Number of oxygen atoms: $\#\{Z_i \mid Z_i = 8\}$.
20	a_nP	Number of phosphorus atoms: $\#\{Z_i \mid Z_i = 15\}$.
21	a_nS	Number of sulfur atoms: $\#\{Z_i \mid Z_i = 16\}$.
22	bpol	Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule (including implicit hydrogens) with polarizabilities taken from [CRC 1994].
23	b_1rotN	Number of rotatable single bonds. Conjugated single bonds are not included (e.g. ester and peptide bonds).
24	b_1rotR	Fraction of rotatable single bonds: b_1rotN divided by b_heavy.
25	b_ar	Number of aromatic bonds.
26	b_count	Number of bonds (including implicit hydrogens). This is calculated as the sum of $(d_i/2 + h_i)$ over all non-trivial atoms *i*.
27	b_double	Number of double bonds. Aromatic bonds are not considered to

		be double bonds.
28	b_heavy	Number of bonds between heavy atoms.
29	b_max1len	Maximum single bond chain length.
30	b_rotN	Number of rotatable bonds. A bond is rotatable if it has order 1, is not in a ring, and has at least two heavy neighbors.
31	b_rotR	Fraction of rotatable bonds: b_rotN divided by b_heavy.
32	b_single	Number of single bonds (including implicit hydrogens). Aromatic bonds are not considered to be single bonds.
33	b_triple	Number of triple bonds. Aromatic bonds are not considered to be triple bonds.
34	chiral_u	The number of unconstrained chiral centers.
35	density	Molecular mass density: Weight divided by vdw_vol (amu/Å^3).
36	diameter	Largest value in the distance matrix [Petitjean 1992]
37	lip_acc	The number of O and N atoms.
38	lip_don	The number of OH and NH atoms.
39	logP(o/w)	Log of the octanol/water partition coefficient (including implicit hydrogens). This property is calculated from a linear atom type model [LOGP 1998] with r^2 = 0.931, RMSE=0.393 on 1,827 molecules.
40	logS	Log of the aqueous solubility (mol/L). This property is calculated from an atom contribution linear atom type model [Hou 2004] with r^2 = 0.90, ~1,200 molecules.
41	mr	Molecular refractivity (including implicit hydrogens). This property is calculated from an 11 descriptor linear model [MREF 1998] with r^2 = 0.997, RMSE = 0.168 on 1,947 small molecules.
42	PC+	Total positive partial charge: the sum of the positive q_i. Q_PC+ is identical to PC+ which has been retained for compatibility.
43	PC-	Total negative partial charge: the sum of the negative q_i. Q_PC- is identical to PC- which has been retained for compatibility.
44 45	PEOE_PC+ Q_PC+	Total positive partial charge: the sum of the positive q_i.
46 47	PEOE_PC- Q_PC-	Total negative partial charge: the sum of the negative q_i.
48 49	PEOE_RPC+ Q_RPC+	Relative positive partial charge: the largest positive q_i divided by the sum of the positive q_i.
50 51	PEOE_RPC- Q_RPC-	Relative negative partial charge: the smallest negative q_i divided by the sum of the negative q_i.
52 53	PEOE_VSA_FHYD Q_VSA_FHYD	Fractional hydrophobic van der Waals surface area. This is the sum of the v_i such that $\|q_i\|$ is less than or equal to 0.2 divided by the total surface area. The v_i are calculated using a connection table approximation.
54 55	PEOE_VSA_FNEG Q_VSA_FNEG	Fractional negative van der Waals surface area. This is the sum of the v_i such that q_i is negative divided by the total surface area. The v_i are calculated using a connection table approximation.
56	PEOE_VSA_FPNEG	Fractional negative polar van der Waals surface area. This is the

57	Q_VSA_FPNEG	sum of the v_i such that q_i is less than -0.2 divided by the total surface area. The v_i are calculated using a connection table approximation.
58 59	PEOE_VSA_FPOL Q_VSA_FPOL	Fractional polar van der Waals surface area. This is the sum of the v_i such that $\|q_i\|$ is greater than 0.2 divided by the total surface area. The v_i are calculated using a connection table approximation.
60 61	PEOE_VSA_FPOS Q_VSA_FPOS	Fractional positive van der Waals surface area. This is the sum of the v_i such that q_i is non-negative divided by the total surface area. The v_i are calculated using a connection table approximation.
62 63	PEOE_VSA_FPPOS Q_VSA_FPPOS	Fractional positive polar van der Waals surface area. This is the sum of the v_i such that q_i is greater than 0.2 divided by the total surface area. The v_i are calculated using a connection table approximation.
64 65	PEOE_VSA_HYD Q_VSA_HYD	Total hydrophobic van der Waals surface area. This is the sum of the v_i such that $\|q_i\|$ is less than or equal to 0.2. The v_i are calculated using a connection table approximation.
66 67	PEOE_VSA_NEG Q_VSA_NEG	Total negative van der Waals surface area. This is the sum of the v_i such that q_i is negative. The v_i are calculated using a connection table approximation.
68 69	PEOE_VSA_PNEG Q_VSA_PNEG	Total negative polar van der Waals surface area. This is the sum of the v_i such that q_i is less than -0.2. The v_i are calculated using a connection table approximation.
70 71	PEOE_VSA_POL Q_VSA_POL	Total polar van der Waals surface area. This is the sum of the v_i such that $\|q_i\|$ is greater than 0.2. The v_i are calculated using a connection table approximation.
72 73	PEOE_VSA_POS Q_VSA_POS	Total positive van der Waals surface area. This is the sum of the v_i such that q_i is non-negative. The v_i are calculated using a connection table approximation.
74 75	PEOE_VSA_PPOS Q_VSA_PPOS	Total positive polar van der Waals surface area. This is the sum of the v_i such that q_i is greater than 0.2. The v_i are calculated using a connection table approximation.
76	radius	If r_i is the largest matrix entry in row *i* of the distance matrix *D*, then the radius is defined as the smallest of the r_i [Petitjean 1992].
77	reactive	Indicator of the presence of reactive groups. A non-zero value indicates that the molecule contains a reactive group. The table of reactive groups is based on the Oprea set [Oprea 2000] and includes metals, phospho-, N/O/S-N/O/S single bonds, thiols, acyl halides, Michael Acceptors, azides, esters, etc.
78	rings	The number of rings.
79	RPC+	Relative positive partial charge.
80	RPC-	Relative negative partial charge.
81	SlogP	Log of the octanol/water partition coefficient (including implicit hydrogens). This property is an atomic contribution model [Crippen 1999] that calculates logP from the given structure; i.e.

		the correct protonation state (washed structures). Results may vary from the logP(o/w) descriptor. The training set for SlogP was ~7000 structures.
82	SMR	Molecular refractivity (including implicit hydrogens). This property is an atomic contribution model [Crippen 1999] that assumes the correct protonation state (washed structures). The model was trained on ~7000 structures and results may vary from the mr descriptor.
83	TPSA	Polar surface area (Å^2) calculated using group contributions to approximate the polar surface area from connection table information only. The parameterization is that of Ertl *et al.* [Ertl 2000].
84	vdw_area	Area of van der Waals surface (Å^2) calculated using a connection table approximation.
85	vdw_vol	an der Waals volume (Å^3) calculated using a connection table approximation.
86	vsa_acc	Approximation to the sum of VDW surface areas (Å^2) of pure hydrogen bond acceptors (not counting acidic atoms and atoms that are both hydrogen bond donors and acceptors such as -OH).
87	vsa_acid	Approximation to the sum of VDW surface areas of acidic atoms (Å^2).
88	vsa_don	Approximation to the sum of VDW surface areas of pure hydrogen bond donors (not counting basic atoms and atoms that are both hydrogen bond donors and acceptors such as -OH) (Å^2).
89	vsa_hyd	Approximation to the sum of VDW surface areas of hydrophobic atoms (Å^2).
90	vsa_other	Approximation to the sum of VDW surface areas (Å^2) of atoms typed as "other".
91	vsa_pol	Approximation to the sum of VDW surface areas (Å^2) of polar atoms (atoms that are both hydrogen bond donors and acceptors), such as -OH.
92	Weight	Molecular weight (including implicit hydrogens) in atomic mass units with atomic weights taken from [CRC 1994].
93	zagreb	Zagreb index: the sum of d_i^2 over all heavy atoms i.
94	ABCB1 Inhib	P-gP inhibition prediction (float number score)
95	ABCG2 Inhib	BCRP inhibition prediction (float number score)
96	BSEP Inhib	BSEP inhibition prediction (float number score)
97	OATPB1_Inhib_Sum_binary	Sum of the binary scores of the 4 classification models for OATP1B1 inhibition (integer score between 0 and 4)
98	OATPB3_Inhib_Sum_binary	Sum of the binary scores of the 4 classification models for OATP1B3 inhibition (integer score between 0 and 4)

Table S2: Information for the transporter models for BSEP, BCRP, P-gp, OATP1B1 and 1B3 inhibition. The size of training set, the threshold of inhibition definition for the training set, the algorithm used and the AUC of the model are provided.

Transporter Inhibition Model	**Training set (compounds' number)**	**Threshold of inhibition for IC_{50} values**	**Algorithm**	**AUC**
BSEP	670	<10 µM: inhibitors >50 µM: noninhibitors 10 µM ≤IC_{50}≤50 µM: compounds removed	RF (10 trees) with feature selection	0.91 (10-fold CV)
BCRP	978	10 µM	Logistic regression	0.90 (10-fold CV)
P-gp	1180	10 µM	SVM	0.94 (10-fold CV)
OATP1B1 (ensemble of 4 models)	1708	10 µM	MetaCost8:1 +RF (10 trees) MetaCost8:1 + SMO (Puk kernel)	0.790-0.806 (10-fold CV)
OATP1B3 (ensemble of 4 models)	1725	10 µM	MetaCost13:1 +RF (10 trees) MetaCost13:1 + SMO (Puk kernel)	0.825-0.866 (10-fold CV)

Table S3: Number of reliable predictions for the cholestasis data for each model based on applicability domain defined according the Euclidean distances between training and test set. Since the model of BSEP inhibition was generated with confidential training data, we cannot report the exact number of reliable predictions.

Model	**Number of Reliable Predictions for Cholestasis Training set 578 compounds**	**Number of Reliable Predictions for Cholestasis Test Set 1347 compounds**	**Number of Reliable Predictions for Cholestasis Merged Training Set 1904 compounds**
BSEP inhibition	confidential	confidential	confidential
BCRP inhibition	562/578 (97.2%)	1290/1347 (95.8%)	1831/1904 (93.9%)
P-gp inhibition	557/578 (95.8%)	1254/1347 (93.1%)	1788/1904 (96.2%)
OATP1B1 inhibition	574/578 (99.3%)	1342/1347 (99.5%)	1895/1904 (99.5%)
OATP1B3 inhibition	574/578 (99.3%)	1342/1347 (99.5%)	1895/1904 (99.5%)
Cholestasis	-	1331/1347 (98.8%)	-

Table S4. Performance of the model trained on the merged data for cholestasis (1904 compounds) and respective p-values. The performance is obtained from 50 iterations for 10-fold cross validation using 93 2D MOE descriptors, with or without transporters predictions and it is provided for accuracy, sensitivity, specificity, AUC and precision. The p-values were obtained by performing a two-sample paired t-test. The model was generated using MetaCost with a cost matrix of [0.0, 1.0; 5.0, 0.0] and SVM as a base classifier using Polynomial kernel (exp=2).

	Accuracy	**Sensitivity**	**Specificity**	**AUC**	**Precision**
93 2D MOE descriptors +Transporters predictions	0.690 ±0.007	0.595 ±0.013	0.711 ±0.008	0.726 ±0.005	0.321 ±0.008
93 2D MOE descriptors	0.670 ±0.007	0.574 ±0.014	0.692 ±0.008	0.690 ±0.006	0.299 ±0.007
p-value	$<2.2*10^{-16}$	$7.949*10^{-12}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$

Table S5. p-values from the respective two-sample t-tests comparing several model-pairs, bases on the usage of transporters predictions as descriptors (dscrs). The compared statistics metrics are accuracy, sensitivity, specificity, MCC, AUC, precision and weighted average precision. The outcome for each comparison is also provided in the conclusions column.

Comparisons	**Accuracy**	**Sensitivity**	**Specificity**	**MCC**	**AUC**	**Precision**	**Weighted Precision**	**Conclusions**
p-values: **i)comparison 93 2D dscrs + transp vs 93 2D dsrs**	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$9.977*10^{-4}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	For all statistics metrics, using 93 2D dscrs + transporters performs better
p-values: **ii)comparison 93 2D dscrs + BSEP vs 93 2D dsrs**	$5.786*10^{-7}$	0.01742	$5.8*10^{-11}$	0.1766	0.537	0.001289	0.3734	In terms of MCC, AUC and weighted precision, the two models perform equally. For the rest of the statistics metrics, including BSEP to the 93 2D dscrs yields better performance.
p-values: **iii)comparison 93 2D dscrs + transp vs 93 2D dscrs + BSEP**	$9.151*10^{-10}$	$<2.2*10^{-16}$	$6.117*10^{-4}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	$<2.2*10^{-16}$	For all statistics metrics using 93 2D dscrs + transporters performs better than using only BSEP.
p-values: **iv)comparison 93 2D dscrs + transp vs 93 2D dscrs + transporters without BSEP**	$9.387*10^{-4}$	0.07589	$2.021*10^{-6}$	0.3016	$1.411*10^{-6}$	$6.253*10^{-3}$	0.01355	For accuracy, specificity, AUC, precision and weighted precision the performance of the model is better when all transporters are used. For sensitivity and MCC the two models perform equally.

When p-values are too small to calculate, the < sign is introduced.