Supporting information

Authors:

Aleksandra Rybacka, Christina Rudén, Igor V. Tetko, Patrik L. Andersson

Manuscript title:

Identifying potential endocrine disruptors among industrial chemicals and their metabolites – development and evaluation of *in silico* tools

Content:

Description of the stepwise model development;

Figures:

Tables:

Table S2. Balanced accuracies of all QSAR models built for estrogen (A), androgen (B), and transthyretin (C) binding.

Table S3. Statistics of the estrogen, androgen, and transthyretin models applied to high and low production volume chemicals.

Table S4. Lists of descriptors and coefficients for PLS models of estrogen (A), androgen (B), and transthyretin (C) binding.

Table S5. Number of predicted binders to estrogen (E), androgen (A), and transthyretin (T) and combinations of these.

Table S6. Structural features comparisons (with the use of SetCompare utility) of predicted estrogen receptor binders among high and low production volume chemicals (E binders) with predicted non-binders whose metabolites are predicted binders (E metabolites).

Table S7. Structural features comparisons of predicted androgen receptor binders among high and low production volume chemicals (A binders) with predicted non-binders whose metabolites are predicted binders (A metabolites).

Table S8. Selected potential endocrine disruptors and reproductive toxicants among high and low production volume chemicals based on their presence on the lists of potential endocrine disruptors or classification as reproductive toxicant according to CLP regulation (No 1272/2008).

Files:

EAT data.xls Data used for training and validation of the EAT models

Section S1. Step 1 – Data collection

E data from the METI database (METI, 2014) were transformed from relative binding affinity (RBA) to IC_{50} by normalizing each chemical's RBA with the reference endogenous ligand 17β-estradiol (E2)(eq. 1).

$$IC_{50}\ of\ ligand = \frac{IC_{50}\ of\ E_2}{RBA} \times 100$$

The A data includes both 25 and 50% (IC_{25} and IC_{50}) inhibition of androgen-induced activity. Development of QSAR models using both responses have recently been successfully applied by Jensen et al., 2011 and Vinggaard et al., 2008. The inclusion of both IC_{25} and IC_{50} data enlarges the applicability domain of the A model by increasing the number of chemicals in the model but it also adds chemicals with weak A response and thus increases the sensitivity of the model.

All data can be found in a separate file (EAT data.xls).

Section S2. Step 2 – Chemical Variation Analysis

PCA is a well-established and powerful tool for pattern recognition in large datasets. It identifies a few uncorrelated linear combinations (principal components) in the original data. Each principal component captures as much variation as possible and is defined by a score and a loading vector. Scores show variance among objects, whereas loadings describe the contribution of individual descriptors to a given component. Descriptors on which PCA was built described mostly topological information about the molecules including connectivity, atom type, charge, hydrophobic and hydrophilic surface, weight, hydrogen bond donors/acceptors, structure flexibility and other. Detailed information on the descriptors was given previously (Rannar and Andersson, 2010; Rybacka et al., 2014); see also Table S1. The chemical structures were washed and optimized with a 94x Merck Molecular Force Field (MMFF) prior to chemical descriptor calculations. The washing procedure included a number of operations such as hydrogen correction, salt and solvent removal, and adjustment and enumeration of protonation states. Prior to PCA, the descriptors were autoscaled and if needed log-transformed (if they were not already logarithmic values). For log-transformation we used a function for variable transformation implemented in SIMCA where the base value (and sometimes also a multiplication) of the logarithm is adjusted automatically depending on the skewness of descriptor's probability distribution.

PCA based on every training set and the high- and low-production volume chemicals (H&LPVCs) clearly showed that the compounds used to develop the A model (the A set) represented a large portion of the chemical variation of the entire H&LPVC set (Figure S1B). The same pattern was seen for compounds with available data depicting binding potency to the estrogen receptor (the E set). These compounds were structurally diverse (similar to the A set) albeit with fewer examples of small polar compounds, as seen in the left lower corner of Figure S1A. In contrast, the compounds that bind to transthyretin (the T set) showed very little structural variation and clustered into two groups in the PCA score plot (Figure S1C). One cluster included halogenated aromatic compounds such as polyhalogenated biphenyls and diphenylethers (upper group in Figure S1C), whereas a second cluster

consisted of per- and poly-fluorinated compounds (lower group in Figure S1C). Since no compounds with MW > 960 g/mol were present in any of the three training sets, and H&LPVCs range up to 2104 g/mol in MW, no compounds were found on the right side of Figures S1A-C.

Section S3. Steps 3 and 4 – Model Development and metabolite simulation

Five different descriptor sets were used for QSAR modelling calculated using the Online Chemical Database (OCHEM (Sushko et al., 2011)). These included: electro-topological state indices (Hall and Kier, 1995; Kier and Hall, 1999); Estate that combined electronic and graph-topological information; ISIDA fragments (Varnek et al., 2008) that used 2D Lewis graph representations of the compounds; GSFragments (Skvortsova et al., 1999) that calculated the frequencies of certain special fragments; CDK descriptors (Steinbeck et al., 2003) that captured electronic, constitutional, topological and geometrical information; and Dragon descriptors (Todeschini and Consonni, 2000; version 6) that consisted of 20 different descriptor blocks. Additionally the octanol-water partition coefficient and water solubility were added as calculated by the ALOGPS program (Tetko and Tanchuk, 2002) implemented in the OCHEM software. More details about the descriptor sets can be found in the OCHEM user manual (http://docs.ochem.eu/display/MAN). The models built to aid the interpretation of the EAT models were based on the functional group counts including from 12 to 18 Dragon descriptors, depending on the model. The chemical structures from the EAT sets were standardized, neutralized, and optimized with Corina (Sadowski and Gasteiger, 1993) in OCHEM. Descriptors with a variance below 0.01 were excluded. If highly correlated (pair-wise Pearson's correlation coefficient R > 0.95), descriptors were used as groups in further modelling.

For finding covariance between calculated chemical descriptors (X-data) and E-, A- or T- data (Y-data), seven regression algorithms were studied. Linear correlations were searched using the following algorithms: multiple linear regression (MLR), Fast Stepwise (Stagewise) Multivariate Linear Regression (FSMLR), and Partial Least Squares (PLS) regression. In brief, FSMLR iteratively builds linear regression models and excludes highly correlated descriptors by means of so-called greedy algorithm (Cormen, 2001), and with PLS, X-data are transformed into a few latent variables aiming for the smallest squared difference between original and predicted objects. Apart from associative neural network (described in materials and methods section 2.3) three additional non-linear models were also used, including k-nearest neighbor (k-NN) regression, random forest exported from the

Weka 3 software (Waikato, 2014); WEKA-RF), and support vector machine using the open source library for support vector machines (LIBSVM). In brief, k-NN classifies a sample according to k-closest objects using voting/average values; WEKA-RF constructs a multitude of decision trees which outputs a class; and LIBSVM produces linear boundaries between object groups in a transformed space, where vector representation of the data is replaced with similarities to other data points (Chang and Lin, 2011). More details about the modelling algorithms can be found in the OCHEM user manual (http://docs.ochem.eu/display/MAN).

All models were validated internally and externally. Prior to modelling, every set was randomly split into training and test sets in 4:1 ratio. The seed random generator from Java was used. The seed number is provided for every model on OCHEM website hence models can be easily recalculated and reproduced. The training set was used in model development, whereas the test set was left for external validation. For internal validation, 5-fold cross-validation or stratified bagging was used. In 5-fold cross-validation, the original sample is randomly partitioned into five equal-sized subsamples—one subsample is retained as the validation data for testing the model whereas the remaining 4 subsamples are used as training data. This procedure is then repeated 5 times. Stratified bagging involves training multiple models based on randomly selected training sets which are created from the original set by sampling with replacement (Breiman, 1996). We used 64-times sampling for this study since the statistical parameters already reached plateau at this level.

For each model, we calculated the number of correctly predicted binders and non-binders (true positives (TPs) and negatives (TNs) respectively), binders incorrectly predicted to be non-binders (false negatives, FNs), and non-binders incorrectly predicted to be binders (false positives, FPs). These measures were used to calculate commonly used statistics by QSAR modellers for model evaluation (eq. 1-4).

(eq. 1) $$accuracy = \frac{TPs+TNs}{TPs+TNs+FPs+FNs}$$

(eq. 2) $$sensitivity = \frac{TPs}{TPs+FNs}$$

(eq. 3) $$specificity = \frac{TNs}{TNs+FPs}$$

(eq. 4) $$balanced\ accuracy = \frac{sensitivity+specificity}{2}$$

Section S4. Step 5 – Applicability domain method

The applicability domain consists of the chemical properties and features characterising the compounds that a model was trained with. If a query chemical is too structurally different from compounds in the training set then that chemical's prediction can be non-reliable and would therefore fall outside of the domain.

The assessment of applicability domain is crucial for assuring high reliability of obtained predictions. Many different approaches for AD assessment were studied (Sahigara et al., 2012; Sushko et al., 2010) yet still no method is ideal (Tetko et al., 2014). According to a number of recent studies accuracy of predictions practically depends neither on the used descriptors nor the used QSAR method but rather on the similarity of queried compounds to the training set molecules (Sheridan et al., 2004; Tetko et al., 2008; Sahigara et al., 2012). This implies also that some compounds tend to yield large errors in their predictions for every model due to their unique chemical and structural properties that may not be captured by descriptors used in the modelling. For this reason, in the AD assessment we have used a set of carefully chosen MOE descriptors that represent universal and interpretable properties related to e.g. connectivity, charge, and hydrophobicity (see Table S1 for detailed description of every used descriptor and section S2 for details). To highlight the most significant information and decrease correlation among descriptors we built PCA models for the E-, A-, and T-training set compounds. Every PCA model included only principal components with eigenvalues higher than 2. Compounds with chemical characteristics that differed significantly from the chemical variation presented by E-, A- or T-data received a high distance-to-the-model (DModX) value. New compounds (not present in the training set) can be predicted into the PCA (and hence have their DModX values calculated). A DModX value larger than around 2.5 usually defines compounds that fell outside of the domain (Wold et al., 2001). Herein we used a more cautious approach where chemicals with DModX > 2 were considered as outside of the applicability domain.

Section S5. Model interpretation with PLS

The EAT models presented in the main manuscript developed using associative neural network, were complemented with PLS models to ease interpretation of critical chemical factors for studied endpoints. The PLS models were built based on simple constitutional descriptors related to e.g., functional group counts. This procedure was recently applied by Vorberg and Tetko, 2014. These PLS models performed slightly worse than other models (4-9% lower balanced accuracy) but were valuable in increasing our understanding of the most critical chemical properties related to modelled response (Table S4). In brief, E- and T-binders are mostly lipophilic compounds with aromatic rings containing hydroxyl groups attached to an aromatic ring (nArOH). These compounds have frequently high numbers of N and O atoms that can function as H-bond donors (nHDon). Their features suggest that binders are easily polarizable which, together with information about molecular complexity and degree of substitution of the compounds, has proven to exert a significant impact on estrogenic activity (Liu et al., 2008). Additionally, the presence of the nHDon descriptor confirms observations by Papa et al. (2013) that information describing the frequency of C-O fragments at a topological distance of 7 – which identifies the length of the molecules and the presence of oxygen atoms – is relevant for T-binding (Papa et al., 2013). E-binders often contain the following functional groups: carbamates (nROCON), hydrazones (nC=N-N<) and sulphonamides (nSO2N); whereas T- and A-binders are, more frequently, aromatic ethers. Like E- and T-binders, A-binders are lipophilic and have high numbers of aromatic carbons, but they also show structural characteristics that are atypical for E- or T-binders. These characteristics include the presence of nitro-groups and aliphatic secondary and tertiary amides, as also observed by Jensen et al. (Jensen et al., 2011). The structural characteristics of typical non-binders were reflected by negative PLS variable coefficients; it is very unlikely that E-binders: are ethers; or contain nitro-groups on aromatic rings; or are secondary aliphatic amines and ketones. Large counts of secondary and quaternary carbons can characterize non-binders to transthyretin. Likewise, non-binders to the estrogen receptor have high numbers of

primary carbons in terminal positions. High numbers of secondary carbons were also frequently found among non-binders to the androgen receptor, along with aliphatic esters and furans.

Section S6. Applicability domain analysis of EAT models

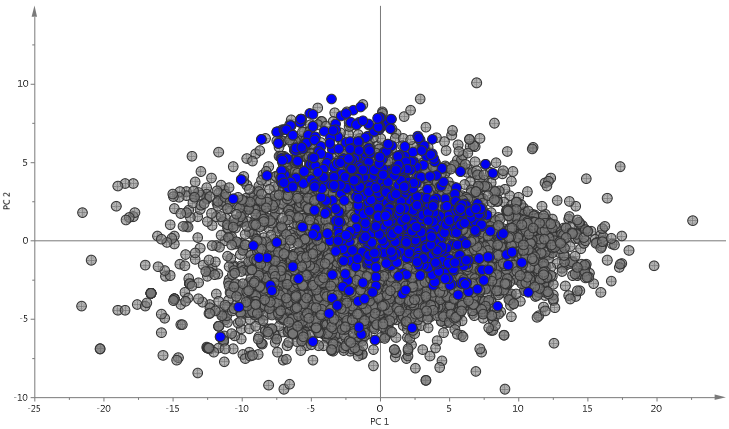
In the fifth step of the study we analysed chemical variation of compounds outside the applicability domain (AD) (step five in Figure 1). A group of 389 chemicals was found to be outside the AD for all three EAT models. These chemicals did not encompass the chemical variation of the H&LPVCs. PCA showed that compounds outside the AD could be separated into three groups (Figure S3A-B). The first group consisted of large and hydrophobic structures with many single bonds (the upper right corner of Figure S3A), such as pentaerythritoltetraoleate. The second group comprised chemicals with a high number of hydrogen acceptors, double bonds and rings, often with sulphuric fragments (the left bottom side of Figure S3A). The third group included a vast range of compounds but among them relatively small and volatile structures, such as acetylene or trifluoromethane, were found (very left side of the plot). Some compounds, such as phosphate-substituted amino acids (found in the middle of Figure S3A) were studied in higher dimensional PCs (3-8) but no grouping was found among them.

Section S7. Structure comparison

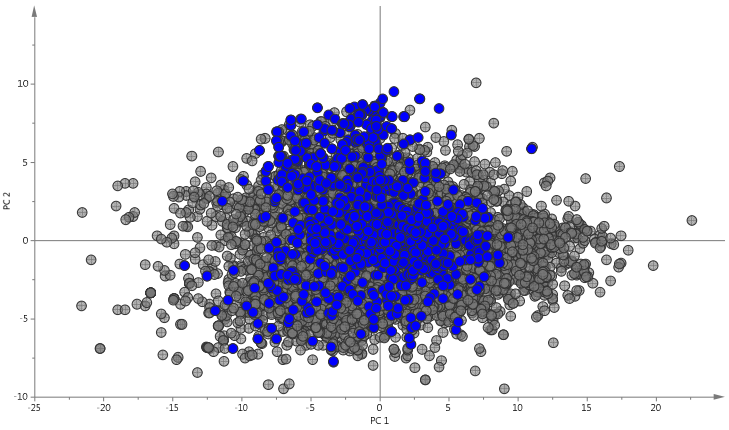
The analysis was done to investigate unique chemical properties of bioactivated E- and A-binders whereas the number of T-binders (50) was considered too few for the analysis. Among the bioactivated compounds, 18.2% were found to be nitro compounds and 16.9% aromatic nitro compounds; these portions were much greater than those of the active parent compounds (3.2 and 2.6%, respectively; Table S5). In addition, bioactivated H&LPVCs included more tertiary amines (29.9%, versus 16.5% for parent compounds) and nitriles (5.5% versus 1%). Several structural features were only present among the bioactivated compounds, such as nitro-haloarenes (6.3%) and aliphatic secondary and tertiary amines (2.7%). The descriptors, including nitro aromatics (nArNO2) and secondary amines (nRNHR), have high negative coefficients in the PLS model for E-binders (Table S4A) and it is not surprising that these chemical features are unlikely to occur among active parent compounds. It is, however, likely that bioactivation leads to the reduction of these structural features, e.g. when compounds undergo deamination (a typical reaction catalysed by cytochrome P450 in phase I metabolism) and thus become E-binders. The same analysis was done for compounds that become A-binders after metabolism (Table S6), however no typical structural features could be identified. Notably, the number of bioactivated compounds (251 chemicals) was only slightly higher than the number of deactivated compounds (187 chemicals). Fewer E- (60) and T- binders (2) were detoxified. T-binders and T-non-binders include a large portion of polyhalogenated aromatics (more information in the first paragraph of Section 3); considering that hydroxylation is a common biotransformation pathway of aromatics, and that the descriptor with the highest positive PLS-coefficient was the number of hydroxylated aromatics (nArOH) (Table S4C), it would not be surprising if bioactivation could lead to a significant increase in the number of T-binders (by 3-fold, as seen in this study). Such bioactivation was already proven to generate potential endocrine-disrupting metabolites for brominated aromatic flame retardants (Hamers et al., 2008).

Figure S1. PCA plots of training and test sets (blue) for estrogen (A), androgen (B), and transthyretin (C) models mapped with high and low production volume chemicals (grey). PC1 and PC2 explain 54% of the variance, and the model has in total 9 PCs (84% of the explained variance).

A



B



C

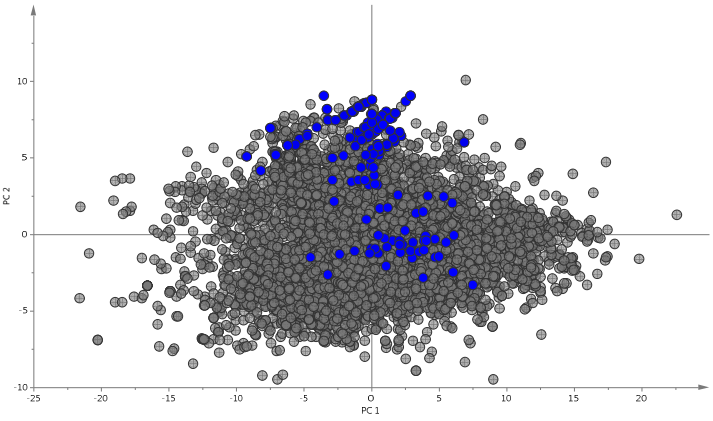
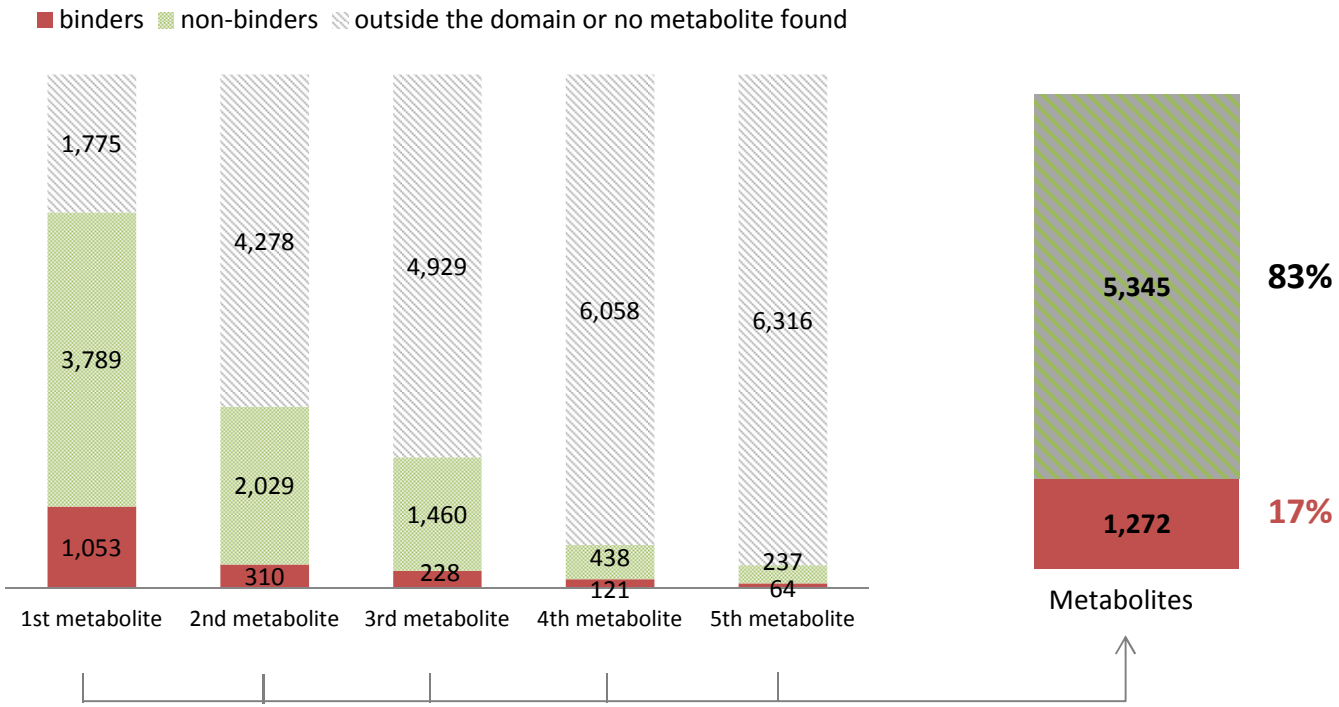
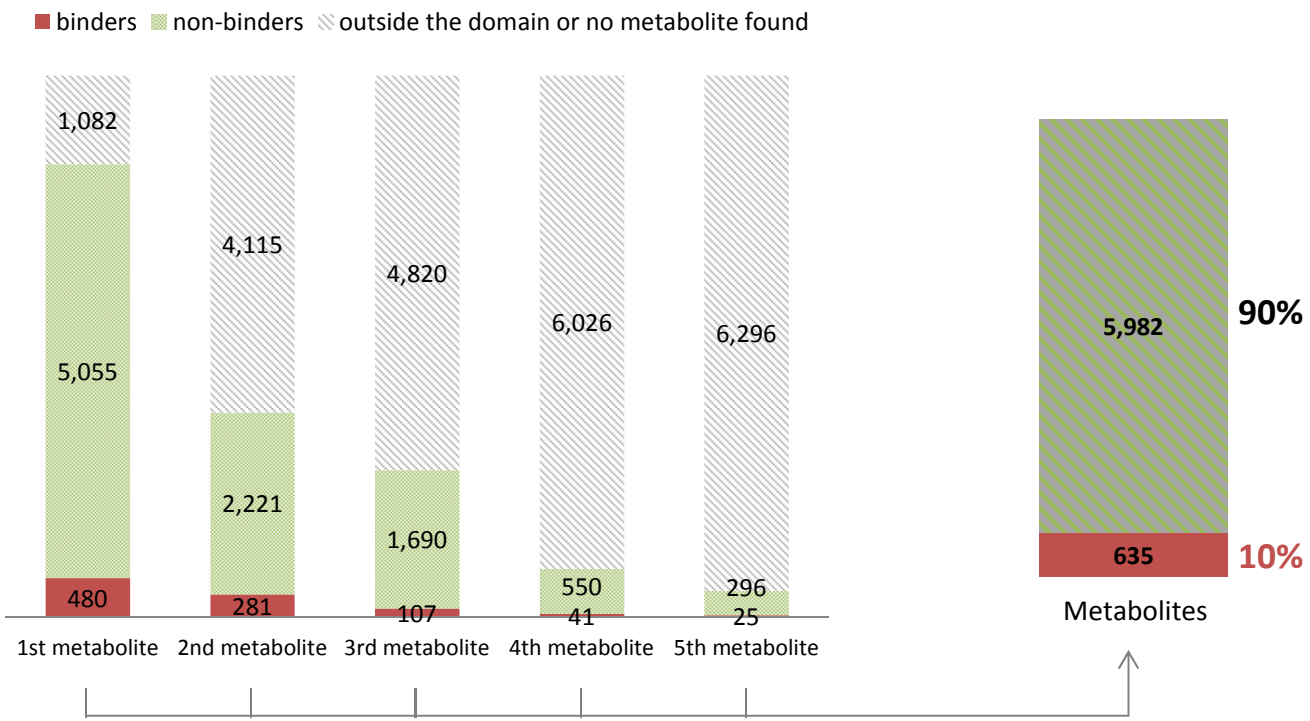


Figure S2. Ratios of predicted binders, non-binders, and compounds outside the applicability domains of the estrogen (A), androgen (B), and transthyretin (C) models for the 5 most likely (according to MetaSite) formed metabolites of high and low production volume chemicals (6,617 chemicals). To the right of all the graphs, the final numbers of compounds that were binders (red) and non-binders/outside the domain (green dots/grey stripes) are given.

A



B


binders
non-binders
outside the domain or no metabolite found
1,082
5,055
480
4,115
2,221
281
4,820
1,690
107
6,026
550
41
6,296
296
25
1st metabolite
2nd metabolite
3rd metabolite
4th metabolite
5th metabolite
5,982
635
90%
10%
Metabolites

C

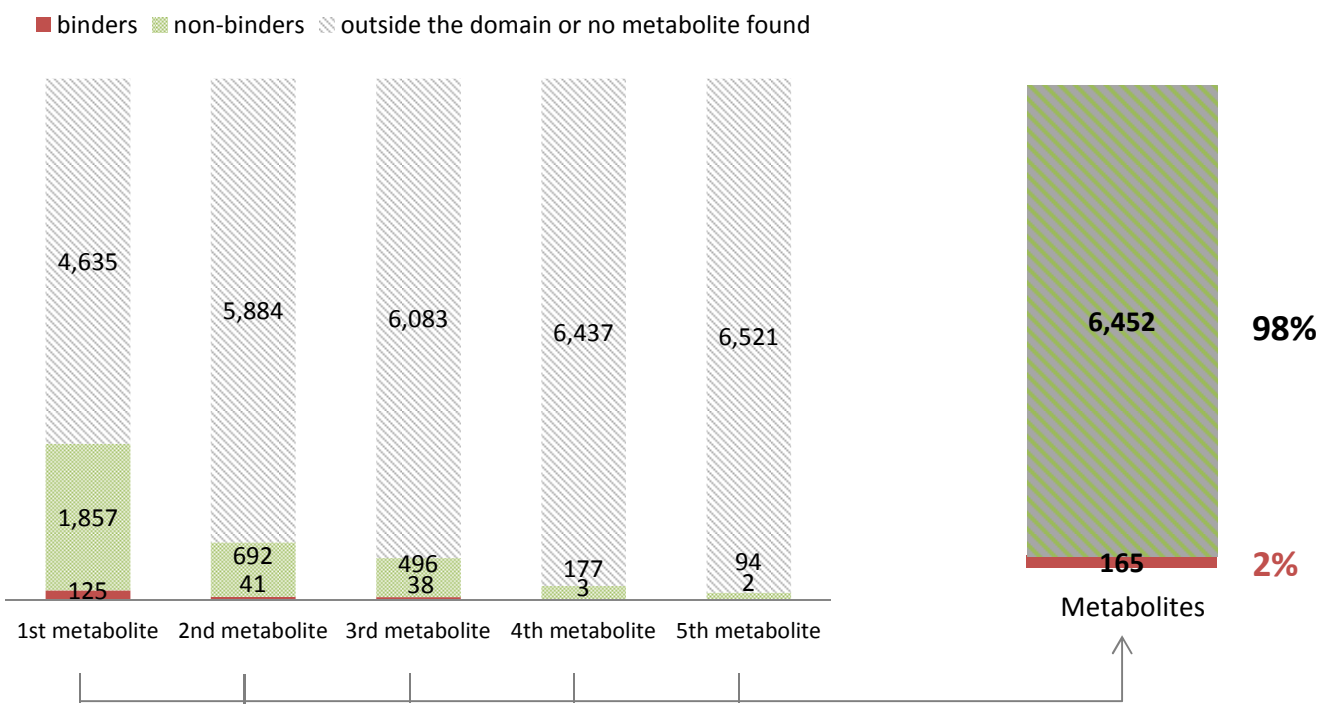
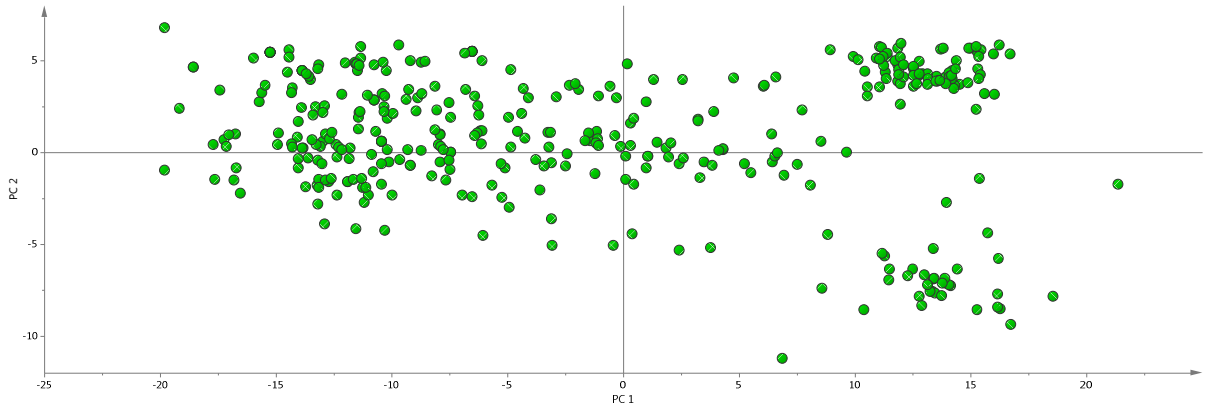

binders
non-binders
outside the domain or no metabolite found
4,635
1,857
125
5,884
692
41
6,083
496
38
6,437
177
3
6,521
94
2
1st metabolite
2nd metabolite
3rd metabolite
4th metabolite
5th metabolite
6,452
165
Metabolites
98%
2%

Figure S3. PCA score plot (A) and loading plot (B) of the 389 compounds identified as out of domain (OAD) by all three EAT models based on simple MOE descriptors. First two PCs explain 54% of variation (PC1 – 42% and PC2 – 12 %).

A



B

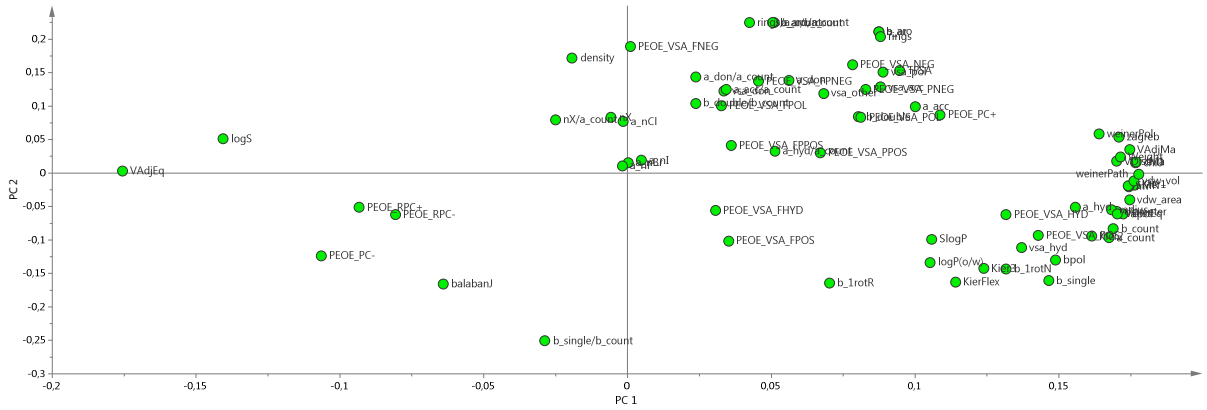


Table S1. The list of MOE descriptors used in PCA analysis

Descriptor	Description
VDistEq	If m is the sum of the distance matrix entries then VdistEq is defined to be the sum of $\log_2 m$ - $p_i \log_2 p_i$ / m where p_i is the number of distance matrix entries equal to i.
VDistMa	If m is the sum of the distance matrix entries then VDistMa is defined to be the sum of $\log_2 m$ - $D_{ij} \log_2 D_{ij}$ / m over all i and j.
b_1rotR	Fraction of rotatable single bonds: b_1rotN divided by b_heavy.
Weight	Molecular weight (including implicit hydrogens) with atomic weights taken from [CRC 1994].
chi0	Atomic connectivity index (order 0) from [Hall 1991] and [Hall 1977]. This is calculated as the sum of 1/sqrt(d_i) over all heavy atoms i with $d_i > 0$.
chi1	Atomic connectivity index (order 1) from [Hall 1991] and [Hall 1977]. This is calculated as the sum of 1/sqrt($d_i d_j$) over all bonds between heavy atoms i and j where $i < j$.
VAdjEq	Vertex adjacency information (equality): -(1-f)$\log_2$(1-f) - $f \log_2 f$ where $f = (n^2 - m) / n^2$, n is the number of heavy atoms and m is the number of heavy-heavy bonds. If f is not in the open interval (0,1), then 0 is returned.
VAdjMa	Vertex adjacency information (magnitude): 1 + $\log_2 m$ where m is the number of heavy-heavy bonds. If m is zero, then zero is returned.
balabanJ	Balaban's connectivity topological index [Balaban 1982].
PEOE_PC+	Total positive partial charge: the sum of the positive qi. Q_PC+ is identical to PC+ which has been retained for compatibility.
PEOE_PC-	Total negative partial charge: the sum of the negative q_i. Q_PC- is identical to PC- which has been retained for compatibility.
PEOE_RPC+	Relative positive partial charge: the largest positive q_i divided by the sum of the positive q_i. Q_RPC+ is identical to RPC+ which has been retained for compatibility.
PEOE_RPC-	Relative negative partial charge: the smallest negative q_i divided by the sum of the negative q_i. Q_RPC- is identical to RPC- which has been retained for compatibility.
PEOE_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.
PEOE_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.
PEOE_VSA_FPNEG	Fractional negative polar van der Waals surface area. This is the 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.
PEOE_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.
PEOE_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.

PEOE_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.
PEOE_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.
PEOE_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.
PEOE_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.
PEOE_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.
PEOE_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.
PEOE_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.
Kier1	First kappa shape index: $(n\text{-}1)^2 / m^2$ [Hall 1991].
Kier2	Second kappa shape index: $(n\text{-}1)^2 / m^2$ [Hall 1991].
Kier3	Third kappa shape index: $(n\text{-}1)\ (n\text{-}3)^2 / p_3^2$ for odd n, and $(n\text{-}3)\ (n\text{-}2)^2 / p_3^2$ for even n [Hall 1991].
KierFlex	Kier molecular flexibility index: (KierA1) (KierA2) / n [Hall 1991].
logS	Log of the aqueous solubility This property is calculated from an atom contribution linear atom type model [Hou 2004] with r^2 = 0.90, ~1,200 molecules.
apol	Sum of the atomic polarizabilities (including implicit hydrogens) with polarizabilities taken from [CRC 1994].
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].
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.
vsa_acc	Approximation to the sum of VDW surface areas of pure hydrogen bond acceptors (not counting acidic atoms and atoms that are both hydrogen bond donors and acceptors such as -OH).
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).
vsa_hyd	Approximation to the sum of VDW surface areas of hydrophobic atoms.
vsa_other	Approximation to the sum of VDW surface areas of atoms typed as "other".
vsa_pol	Approximation to the sum of VDW surface areas of polar atoms (atoms that are both hydrogen bond donors and acceptors), such as -OH.

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.
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.
TPSA	Polar surface area 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].
density	Molecular mass density: Weight divided by vdw_vol.
vdw_area	Area of van der Waals surface calculated using a connection table approximation.
vdw_vol	van der Waals volume calculated using a connection table approximation.
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.
diameter	Largest value in the distance matrix [Petitjean 1992].
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].
wiener Path	Wiener path number: half the sum of all the distance matrix entries as defined in [Balaban 1979] and [Wiener 1947].
wiener Pol	Wiener polarity number: half the sum of all the distance matrix entries with a value of 3 as defined in [Balaban 1979].
a_aro	Number of aromatic atoms.
b_1rotN	Number of rotatable single bonds. Conjugated single bonds are not included (e.g., ester and peptide bonds).
b_ar	Number of aromatic bonds.
b_double	Number of double bonds. Aromatic bonds are not considered to be double bonds.
rings	The number of rings.
zagreb	Zagreb index: the sum of d_i^2 over all heavy atoms i.
b_double/b_count	Number of double bonds. / Number of bonds (including implicit hydrogens). This is calculated as the sum of (di/2 + hi) over all non-trivial atoms i.
b_ar/b_count	Number of aromatic bonds / Number of bonds
b_single/b_count	Number of single bonds / Number of bonds
a_aro/a_count	Number of aromatic atoms / Number of atoms
a_don/a_count	Number of hydrogen bond donor atoms / Number of atoms
a_acc/a_count	Number of hydrogen bond acceptor atoms (not counting acidic atoms but counting atoms that are both hydrogen bond donors and acceptors such as -OH) / Number of atoms
a_hyd / a_count	Number of hydrophobic atoms / Number of atoms
nX	Number of halogen atoms

nX/a_count	Number of halogen atoms / Number of atoms
rings/a_count	The number of rings / Number of atoms

Table S2. Balanced accuracies of all QSAR models built for estrogen (A), androgen (B), and transthyretin (C) binding. The abbreviations in the first row stand for the following machine learning methods: Partial Least Squares (PLS), Associative Neural Networks (ASNN), k-Nearest Neighbour (k-NN), Supporting Vector Machine using the LibSVM, Fast Stepwise (Stagewise) Multivariate Linear Regression (FSMLR), Multilinear Regression Analysis (MLRA), and Random Forest method implemented in Weka software (WEKA-RF)[1]. For every machine learning method an internal validation procedures were used: cross-validation (CV) or bagging procedure (bag). Statistics of the models chosen for further modelling are bolded.

A

Descriptors/machine learning method		ASNN		k-NN		LibSVM		FSMLR		MLRA		PLS		WEKA-RF	
		CV	bag	CV	bag	CV	bag	CV	bag	CV	bag	CV	bag	CV	bag
CDK[2]	Training set	84%	86%	82%	81%	84%	84%	77%	73%	82%	84%	79%	83%	80%	83%
	Test set	88%	92%	88%	87%	82%	90%	88%	87%	83%	87%	84%	87%	83%	89%
Dragon6 (blocks: 1-29)	Training set	86%	**87%**	82%	81%	85%	87%	83%	85%	79%	85%	82%	84%	82%	83%
	Test set	89%	**88%**	86%	88%	92%	92%	85%	75%	80%	84%	84%	86%	85%	90%
ALogPS, OEstate	Training set	84%	85%	79%	81%	83%	84%	80%	83%	82%	82%	81%	83%	84%	84%
	Test set	84%	85%	83%	86%	87%	85%	81%	86%	84%	82%	83%	87%	90%	90%
Fragmentor (Length 2 – 4)	Training set	84%	85%	75%	76%	83%	84%	80%	83%	79%	81%	81%	82%	81%	83%
	Test set	85%	83%	79%	83%	84%	89%	80%	85%	81%	81%	83%	82%	87%	89%
GSFrag	Training set	83%	84%	77%	78%	77%	82%	77%	79%	78%	79%	77%	78%	80%	81%
	Test set	87%	88%	81%	83%	82%	89%	79%	81%	81%	80%	83%	80%	79%	85%

B

Descriptors/machine learning method		ASNN		k-NN		LibSVM		FSMLR		MLRA		PLS		WEKA-RF	
		CV	bag	CV	bag	CV	bag	CV	bag	CV	bag	CV	bag	CV	bag
CDK[2]	Training set	77%	76%	73%	74%	74%	77%	65%	66%	70%	73%	64%	67%	70%	72%
	Test set	78%	**77%**	80%	78%	81%	80%	65%	71%	75%	73%	71%	74%	75%	78%
Dragon6 (blocks: 1-29)	Training set	77%	**77%**	75%	73%	74%	76%	64%	72%	69%	72%	64%	76%	71%	72%
	Test set	79%	75%	77%	74%	79%	81%	71%	76%	62%	75%	61%	77%	72%	77%
ALogPS, OEstate	Training set	74%	76%	73%	72%	75%	76%	64%	70%	69%	72%	67%	72%	73%	73%
	Test set	79%	78%	76%	79%	74%	77%	65%	76%	74%	73%	68%	71%	78%	78%
Fragmentor (Length 2 – 4)	Training set	75%	75%	70%	68%	75%	78%	68%	71%	70%	72%	64%	72%	73%	73%
	Test set	79%	78%	74%	76%	78%	74%	64%	72%	69%	69%	72%	73%	78%	76%
GSFrag	Training set	74%	75%	71%	69%	74%	74%	62%	66%	71%	72%	63%	66%	71%	72%
	Test set	80%	81%	75%	77%	78%	79%	70%	71%	74%	76%	64%	69%	74%	76%

C

Descriptors/machine learning method		ASNN		k-NN		LibSVM		FSMLR		MLRA		PLS		WEKA-RF	
		CV	bag	CV	bag	CV	bag	CV	bag	CV	bag	CV	bag	CV	bag
CDK[2]	Training set	82%	86%	81%	83%	81%	87%	83%	84%	80%	82%	80%	84%	86%	88%
	Test set	90%	86%	85%	84%	82%	87%	79%	76%	76%	77%	82%	79%	78%	85%
Dragon6 (blocks: 1-29)	Training set	88%	**89%**	82%	88%	87%	89%	80%	87%	78%	65%	85%	88%	85%	88%
	Test set	89%	**89%**	86%	82%	87%	89%	79%	86%	84%	76%	80%	89%	85%	83%
ALogPS, OEstate	Training set	85%	84%	79%	80%	81%	85%	79%	81%	77%	74%	79%	79%	88%	88%
	Test set	82%	79%	79%	79%	81%	80%	65%	69%	70%	73%	69%	69%	64%	85%
Fragmentor (Length 2 – 4)	Training set	82%	81%	84%	79%	83%	81%	81%	80%	79%	77%	81%	79%	87%	85%
	Test set	76%	80%	68%	82%	79%	82%	66%	67%	62%	64%	80%	72%	82%	79%
GSFrag	Training set	84%	88%	75%	78%	81%	89%	78%	89%	79%	78%	80%	84%	83%	87%
	Test set	82%	84%	79%	79%	78%	91%	66%	69%	69%	77%	71%	76%	78%	85%

[1] more details at www.cs.waikato.ac.nz/ml/weka/

[2] constitutional, topological, geometrical, electronic and hybrid descriptors

Table S3. Statistics of selected estrogen, androgen, and transthyretin Associative Neural Networks models.

Predicted endpoint		Number of compounds	Balanced accuracy	Sensitivity	Specificity
Estrogen receptor binding	Training set	743	87%	86%	87%
	Test set	186	91%	94%	88%
Androgen receptor binding	Training set	744	77%	68%	86%
	Test set	186	81%	74%	88%
Transthyretin binding	Training set	162	89%	86%	92%
	Test set	41	89%	83%	94%

Table S4. Lists of descriptors and coefficients for PLS models of estrogen (A), androgen (B), and transthyretin (C) binding.

A

Coefficient	Descriptor	Description	Coefficient	Descriptor	Description
+0.186	nArOH	Number of aromatic hydroxyls	-0.0954	ALogPS_logS	Water solubility
+0.11	ALogPS_logP	Octanol/water coefficient	-0.0912	nRCO	Number of ketones (aliphatic)
+0.0992	nCrs	Number of ring secondary C(sp3)	-0.0671	nCp	Number of terminal primary C(sp3)
+0.0666	nR#CH/X	Number of terminal C(sp)	-0.0635	nArX	Number of X on aromatic ring
+0.0656	nROCON	Number of (thio-) carbamates (aliphatic)	-0.0599	nArOR	Number of ethers (aromatic)
+0.0617	nHDon	Number of donor atoms for H-bonds (N and O)	-0.0317	nArNO2	Number of nitro groups (aromatic)
+0.0303	nC=N-N<	Number of hydrazones	-0.0304	nRNHR	Number of secondary amines (aliphatic)
+0.0276	nCb-	Number of substituted benzene C(sp2)	-0.0178	nArCHO	Number of aldehydes (aromatic)
+0.024	nSO2N	Number of sulfonamides (thio-/dithio-)	-0.0176	nR=CX2	Number of nR=CX2

B

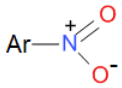
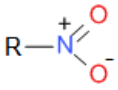
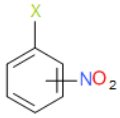
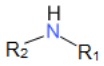
Coefficient	Descriptor	Description	Coefficient	Descriptor	Description
+0.0559	nRCONHR	Number of secondary amides (aliphatic)	-0.0761	ALogPS_logS	Water solubility
+0.0505	ALogPS_logP	Octanol/water coefficient	-0.0393	nCs	Number of total secondary C(sp3)
+0.0481	nCrq	Number of ring quaternary C(sp3)	-0.0355	nR=CHX	Number of R=CHX

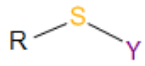
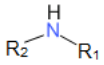
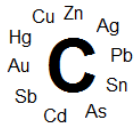
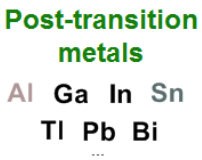
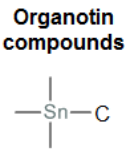
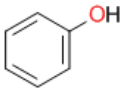
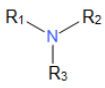
+0.0428	nR=Cs	Number of aliphatic secondary C(sp2)	-0.0334	nRCOOR	Number of esters (aliphatic)
+0.0398	nCb-	Number of substituted C(sp2) in benzenes	-0.0324	nR=CRX	Number of R=CRX
+0.037	nArNO2	Number of nitro groups (aromatic)	-0.0305	nFuranes	Number of furanes
+0.0344	nCar	Number of aromatic C(sp2)			
+0.0322	nCHRX2	Number of CHRX2			
+0.0318	nRCONR2	Number of tertiaryamides			

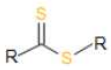
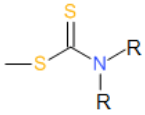
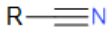
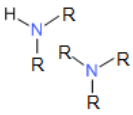
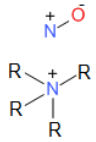
C

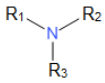
Coefficient	**Descriptor**	**Description**	**Coefficient**	**Descriptor**	**Description**
+0.18	nArOH	Number of aromatic hydroxyls	-0.113	nCq	Number of total quaternary C(sp3)
+0.132	nHDon	Number of donor atoms for H-bonds (N and O)	-0.0781	nCp	Number of terminal primary C(sp3)
+0.0882	nCb-	Number of substituted benzenes C(sp2)	-0.0655	nCs	Number of total secondary C(sp3)
+0.0783	nCRX3	Number of CRX3	-0.0593	nCbH	Number of unsubstituted benzenes
+0.066	nCH2RX	Number of CH2RX			
+0.0571	ALogPS_logP	Water/octanol coefficient			
+0.0548	nArOR	Number of ethers (aromatic)			
+0.0514	nCrq	Number of ring quaternary C(sp3)			

Table S5. Structural features comparisons (with the use of SetCompare utility in OCHEM software) of predicted estrogen receptor binders among high and low production volume chemicals (E binders) with predicted non-binders whose metabolites are predicted binders (E metabolites). The estrogen binding was predicted by the ASNN model.

Descriptor	E-binders (620 compounds)	E parent-metabolites (804 compounds)	Enrichment factor	p-Value
Aromatic nitro 	16 (2.6%)	136 (16.9%)	6.6	-1.09E-20
	20 (3.2%)	146 (18.2%)	5.6	-1.65E-20
Alcohols (R – OH)	262 (42.3%)	176 (21.9%)	1.9	1.3E-16
Alcohols or phenols	284 (45.8%)	201 (25.0%)	1.8	1.76E-16
Nitro-haloarenes 	0	51 (6.3%)	Inf.	-1.07E-13
Secondary aromatic amines 	86 (13.9%)	28 (3.5%)	4.0	5.78E-13
Contains metals 	46 (7.4%)	5 (0.6%)	11.9	1.57E-12

	42 (6.8%)	4 (0.5%)	13.6	6.04E-12
				
Secondary amines	131 (21.1%)	71 (8.8%)	2.4	3.93E-11
				
Organo metallic compounds	28 (4.5%)	0	Inf.	5.44E-11
				
	28 (4.5%)	0	Inf.	5.44E-11
	26 (4.2%)	0	Inf.	3.02E-10
Phenols	165 (26.6%)	111 (13.8%)	1.9	1.16E-9
				
Tertiary amine	102 (16.5%)	240 (29.9%)	1.8	-2.0E-9
				

Ortho- or paraalkylphenols	66 (10.6%)	24 (3.0%)	3.6	3.27E-9
tioesthers	26 (4.2%)	2 (0.2%)	16.9	3.98E-8
				
dithiocarbamates	23 (3.7%)	1 (0.1%)	29.8	5.56E-8
				
Nitriles	6 (1%)	44 (5.5%)	5.7	-1.08E-6
				
Aliphatic secondary and teriary amines	0	22 (2.7%)	Inf.	-3.04E-6
				
Quaternary salts (including N-oxides)	13 (2.1%)	58 (7.2%)	2.4	-4.01E-6
				
Teriary mixed amines (aryl alkyl)	38 (6.1%)	107 (13.3%)	2.2	-4.25E-6



Azo-type (general)	50 (8.1%)	128 (15.9%)	2.0	-4.41E-6

Aromatic azo	47 (7.6%)	119 (14.8%)	2.0	-1.33E-5

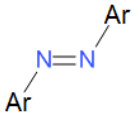
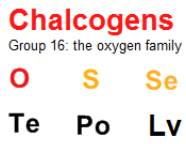
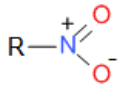
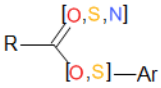
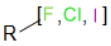
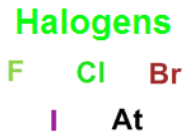
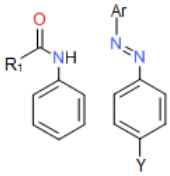
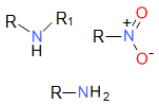
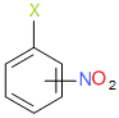


Table S6. Structural features comparisons (with the use of SetCompare utility in OCHEM software) of predicted androgen receptor binders among high and low production volume chemicals (A binders) with predicted non-binders whose metabolites are predicted binders (A metabolites). The androgen binding was predicted by the ASNN model.

Descriptor	A-binders (610 compounds)	A-metabolites (251 compounds)	Enrichment factor	p-Value
	545 (89.3%)	133 (53.0%)	1.7	5.82E-30
Aromatic nitro 	104 (17.0%)	2 (0.8%)	21.4	1.01E-14
	105 (17.2%)	2 (0.8%)	21.6	6.89E-15
Halogen derivatives (alkyl or aryl) 	207 (33.9%)	37 (14.7%)	2.3	2.89E-9
Aryl halide 	178 (29.2%)	29 (11.6%)	2.5	6.67E-9
	149 (24.4%)	23 (9.2%)	2.7	6.52E-8

	82 (13.4%)	6 (2.4%)	5.6	6.54E-8
	158 (25.9%)	26 (10.4%)	2.5	9.01E-8
	209 (34.3%)	43 (17.1%)	2.0	1.74E-7
Aromatic amines precursors 	78 (12.8%)	6 (2.4%)	5.3	2.26E-7
Aromatic N groups 	141 (23.1%)	23 (9.2%)	2.5	5.05E-7
Nitro-haloarenes 	39 (6.4%)	0	Inf.	1.01E-6
	179 (29.3%)	36 (14.3%)	2.0	1.31E-6

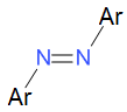
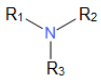
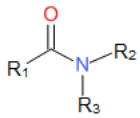
	51 (8.4%)	2 (0.8%)	10.5	1.76E-6
	173 (28.4%)	35 (13.9%)	2.0	2.52E-6
Alcohols or phenols	186 (30.5%)	41 (16.3%)	1.9	7.64E-6
alcohols	191 (31.3%)	44 (17.5%)	1.8	1.67E-5
Carboxylic acid amines	76 (12.5%)	9 (3.6%)	3.5	1.68E-5
				
H-N-C=O	60 (9.8%)	7 (2.8%)	3.5	1.38E-4

Table S7. Selected potential endocrine disruptors and reproductive toxicants among high and low production volume chemicals based on their presence on the lists of potential endocrine disruptors or classification as reproductive toxicant according to CLP regulation (No 1272/2008). Results from EAT models are given in columns 'E' (estrogen binding model), 'T' (transthyretin binding model), and 'A' (androgen antagonistic model). Compounds predicted as binders are marked with 'yes', non-binders with 'no' and compounds outside the domain of a model with 'OAD' (in 'E', 'A' and 'T' columns). References on phase I metabolism data are provided ('reference' column).

CAS number	name	E	A	T	E[1]	A[1]	T[1]	reference	reference agreement[2]
1836-75-5	Nitrofen	no	yes	no	yes	yes	yes	(Brown and Manson, 1986)	yes
108-73-6	Phloroglucinol	no	no	no	yes	OAD	yes	(Monge et al., 1984)	no
1675-54-3	BADGE	no	no	no	yes	yes	OAD	(Bingham et al., 2001)	no
92-52-4	Biphenyl	OAD	no	OAD	yes	yes	OAD	(Meyer, 1977)	yes
569-64-2	Machite green	no	no	OAD	yes	yes	OAD	(Culp et al., 1999)	yes
1091-93-6	3-methoxyestra-2,5(10)-dien-17beta-ol	no	no	OAD	yes	yes	OAD	no data found	-
60628-96-8	bifonazole	no	no	no	yes	yes	OAD	no data found	-
439-14-5	diazepam	no	yes	no	yes	yes	OAD	(Umezawa et al., 2008)	yes
50-48-6	amitriptyline	no	yes	OAD	yes	yes	OAD	(Olesen and Linnet, 1997)	yes
7681-93-8	natamycin	no	yes	OAD	yes	yes	OAD	(EFSA, 2009)	no
117-81-7	Bis(2-ethylhexyl) phthalate	no	no	no	yes	OAD	OAD	(JRC, 2008)	no
121-75-5	malathion	no	no	OAD	yes	OAD	OAD	(Buratti et al., 2005)	yes
15087-24-8	Benzylidene camphor	no	no	no	yes	OAD	OAD	(SCCS, 2013)	no
19044-88-3	Oryzalin	no	no	OAD	yes	OAD	OAD	(U.S.EPA, 1994)	yes

101-20-2	Triclocarban	no	no	OAD	yes	OAD	OAD	(Scheb et al., 2011)	yes
57-68-1	Sulfadimidine	no	no	OAD	yes	OAD	OAD	(Paulson et al., 1987)	yes
81-11-8	4,4'-diaminostilbene-2,2'-disulphonic acid	no	no	OAD	yes	OAD	OAD	no data found	-
88-30-2	Alpha,alpha,alpha-trifluoro-4-nitro-m-cresol	no	no	no	yes	OAD	OAD	(Lech, 1971; Lech and Costrini, 1972)	no
91-53-2	Ethoxyquin	no	no	OAD	yes	OAD	OAD	(Burka et al., 1996; Skaare and Solheim, 1979)	yes
1689-99-2	2,6-dibromo-4-cyanophenyl octanoate	no	no	no	yes	OAD	OAD	(Roberts et al., 1998)	no
3861-47-0	4-cyano-2,6-diiodophenyl octanoate	no	OAD	no	yes	OAD	OAD	(Roberts et al., 1998)[3]	[3]
525-66-6	propranolol	no	no	OAD	yes	OAD	OAD	(Masubuchi et al., 1994)	yes
66357-35-5	ranitidine	no	no	OAD	yes	OAD	OAD	(Cross et al., 1995)	yes
35554-44-0	imazalin	no	yes	no	OAD	yes	yes	(Mannens et al., 1993 (Unpublished work))[4]	[4]
66246-88-6	penconazole	no	yes	no	OAD	yes	yes	(EFSA, 2008)	[5]
94-82-6	4-(2,4-dichlorophenoxy)butyric acid	no	no	no	OAD	yes	yes	(EC, 2002)	yes
59-50-7	Chlorocresol	no	yes	no	OAD	OAD	yes	[6]	no
118-74-1	Hexachlorobenzene	no	no	no	OAD	OAD	yes	(To-Figuer	yes

								as et al., 1997)	
87-65-0	2,6-dichlorophenol	no	no	no	OAD	OAD	yes	(Ministry of Economy)	no
1570-64-5	4-chloro-o-cresol	no	yes	no	OAD	OAD	yes	no data found	-
1689-84-5	3,5-dibromo-4-hydroxybenzonitrile	no	no	no	OAD	OAD	yes	(Roberts et al., 1998)	no
32534-81-9	Diphenylether, pentabromoderivative (BDE-99)	OAD	OAD	no	OAD	OAD	yes	(Erratico et al., 2011; Stapleton et al., 2009)	yes
106-48-9	4-chlorophenol	no	no	no	OAD	OAD	yes	(Ahlborg et al., 1980; Call et al., 1980)	no
94-74-6	(4-chloro-2-methylphenoxy)acetic acid	no	no	no	OAD	OAD	yes	(Lappin et al., 2002)	no
95-50-1	1,2-dichlorobenzene	no	no	no	OAD	OAD	yes	(Hissink et al., 1996)	yes
99-93-4	4'-hydroxyacetophenone	no	no	no	OAD	OAD	yes	no data found	-
32536-52-0	Diphenylether, octabromoderivative	OAD	OAD	no	OAD	OAD	yes	no data found	-
5216-25-1	Alpha,alpha,alpha,4-tetrachlorotoluene	no	no	no	OAD	OAD	yes	(U.S.EPA, 2013)	no

[1]results for metabolites
[2]data agrees between experimental observations, see indicated reference, and MetaSite
[3]authors speculates on metabolism
[4]three major metabolites were identified but the compounds were metabolized into at least 25 metabolites
[5]non-binding metabolite was correctly simulated
[6]data taken from registration dossier available at echa.europa.eu with the help of echemportal.org

References

Ahlborg UG, Thunberg TM, Spencer HC. Chlorinated Phenols: Occurrence, Toxicity, Metabolism, And Environmental Impact. Cri. Rev. Toxicol. 1980; 7: 1-35.

Bingham E, Cohrssen B, Powell CH. Patty's Toxicology. Vol 1-9. New York, N.Y., 2001.

Breiman L. Bagging predictors. Machine Learning 1996; 24: 123-140.

Brown T, Manson J. Further characterization of the distribution and metabolism of nitrofen in the pregnant rat. Teratology 1986; 34: 129-39.

Buratti FM, D'Aniello A, Volpe MT, Meneguz A, Testai E. Malathion bioactivation in the human liver: the contribution of different cytochrome P450 isoforms. Drug Metab. Dispos. 2005; 33: 295-302.

Burka L, Sanders J, Matthews H. Comparative metabolism and disposition of ethoxyquin in rat and mouse. II. Metabolism. Xenobiotica 1996; 26: 597-611.

Call D, Brooke L, Lu P. Uptake, elimination, and metabolism of three phenols by fathead minnows. Arch. Environ. Con. Tox. 1980; 9: 699-714.

Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2011; 2: 1-27.

Cormen, T.H.L., Charles E.; Rivest, Ronald L.; Stein, Clifford, 2001. Introduction to algorithms. the MIT Press; 2nd edition.

Cross D, Bell J, Wilson K. Kinetics of ranitidine metabolism in dog and rat isolated hepatocytes. Xenobiotica 1995; 25: 367-75.

Culp SJ, Blankenship LR, Kusewitt DF, Doerge DR, Mulligan LT, Beland FA. Toxicity and metabolism of malachite green and leucomalachite green during short-term feeding to Fischer 344 rats and B6C3F1 mice. Chem-Biol. Interact. 1999; 122: 153-170.

European Commission, Health & Consumer Protection Directorate. Review report for the active substance 2,4-DB. Available at http://ec.europa.eu/food/fs/sfp/ph_ps/pro/eva/existing/list1_2-4-db_en.pdf. [Last accessed July 03 2015]

European Food Safety Authority. Conclusion on the peer review of pencazole. Available at http://www.efsa.europa.eu/en/efsajournal/doc/175r.pdf. [Last accessed July 03 2015]

EFSA Panel on Food Additives and Nutrient Sources added to Food. Scientific Opinion on the use of natamycin (E 235) as a food additive. 2009. 12. Available at www.efsa.europa.eu. [Last accessed July 03 2015]

Erratico CA, Moffatt SC, Bandiera SM. Comparative Oxidative Metabolism of BDE-47 and BDE-99 by Rat Hepatic Microsomes. Toxicol. Sci. 2011; 123: 37-47.

Hall LH, Kier LB. Electrotopological state indices for atom types - a novel combination of electronic, topological, and valence state information. J. Chem. Inf. Comp. Sci. 1995; 35: 1039-1045.

Hamers T, Kamstra JH, Sonneveld E, Murk AJ, Visser TJ, Van Velzen MJM, et al. Biotransformation of brominated flame retardants into potentially endocrine-disrupting metabolites, with special attention to 2,2,4,4-tetrabromodiphenyl ether (BDE-47). Mol. Nutr. Food Res. 2008; 52: 284-298.

Hissink AM, Oudshoorn MJ, Van Ommen B, Haenen GRMM, Van Bladeren PJ. Differences in Cytochrome P450-Mediated Biotransformation of 1,2-Dichlorobenzene by Rat and Man: Implications for Human Risk Assessment. Chem. Res. Toxicol. 1996; 9: 1249-1256.

Jensen GE, Nikolov NG, Wedebye EB, Ringsted T, Niemelä JR. QSAR models for anti-androgenic effect – a preliminary study. SAR QSAR Environ. Res. 2011; 22: 35-49.

Commision JE. Institute for Health and Consumer Protection. Bid (2-ethylhexyl) phthalate (DEHP) - Summary Risk Assessment Report (EUR 23384 EN/2). 2008. Available at available online at: echa.europa.eu. [Last accessed July 03 2015]

Kier LB, Hall LH. Molecular Structure Description: The Electropological State: Academic Press: London, 1999.
Lappin GJ, Hardwick TD, Stow R, Pigott GH, Ravenzwaay Bv. Absorption, metabolism and excretion of 4-chloro-2-methylphenoxyacetic acid (MCPA) in rat and dog. Xenobiotica 2002; 32: 153-163.
Lech JJ. Metabolism of 3-trifluoromethyl-4-nitrophenol in the rat. Toxicol. Appl. Pharm. 1971; 20: 216-226.
Lech JJ, Costrini NV. In vitro and in vivo metabolism of 3-trifluoromethyl-4-nitrophenol (TFM) in rainbow trout. Comp. Gen. Pharmacol. 1972; 3: 160-166.
Liu H, Papa E, Gramatica P. Evaluation and QSAR modeling on multiple endpoints of estrogen activity based on different bioassays. Chemosphere 2008; 70: 1889-1897.
Liu RF, Liu J, Tawa G, Wallqvist A. 2D SMARTCyp Reactivity-Based Site of Metabolism Prediction for Major Drug-Metabolizing Cytochrome P450 Enzymes. J. Chem. Inf. Model. 2012; 52: 1698-1712.
Mannens G, Van Leemput L, Heykants J. General metabolism of imazalil, 1993 (Unpublished work).
Masubuchi Y, Hosokawa S, Horie T, Suzuki T, Ohmori S, Kitada M, et al. Cytochrome P450 isozymes involved in propranolol metabolism in human liver microsomes. The role of CYP2D6 as ring-hydroxylase and CYP1A2 as N-desisopropylase. Drug Metab. Dispos. 1994; 22: 909-915.
METI, 2014. Receptor Binding Assay database. Ministry of Economy, Trade and Industry (METI).
Meyer T. The metabolism of biphenyl. IV. Phenolic metabolites in the guinea pig and the rabbit. Acta Pharmacol. Tox. 1977; 40: 193-200.
Hazard assessment of 2,4-dichlorophenol.Available at http://www.meti.go.jp/. [Last accessed July 03 2014]
Monge P, Solheim E, Scheline RR. Dihydrochalcone metabolism in the rat: Phloretin. Xenobiotica 1984; 14: 917-924.
Olesen OV, Linnet K. Metabolism of the tricyclic antidepressant amitriptyline by cDNA-expressed human cytochrome P450 enzymes. Pharmacology 1997; 55: 235-43.
Papa E, Kovarich S, Gramatica P. QSAR prediction of the competitive interaction of emerging halogenated pollutants with human transthyretin£. SAR QSAR Environ. Res. 2013; 24: 333-349.
Paulson GD, Feil VJ, Macgregor JT. Formation of a diazonium cation intermediate in the metabolism of sulphamethazine to desaminosulphamethazine in the rat. Xenobiotica 1987; 17: 697-707.
Rannar S, Andersson PL. A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment. J. Chem. Inf. Model. 2010; 50: 30-36.
Roberts TR, Hutson D, Lee P. Metabolic Pathways of Agrochemicals, Part 1: Herbicides and Plant Growth Regulators. Cambridge: GBR: Royal Society of Chemistry, 1998.
Rybacka A, Rudén C, Andersson PL. On the Use of In Silico Tools for Prioritising Toxicity Testing of the Low-Volume Industrial Chemicals in REACH. Basic Clin. Pharmacol. 2014: 115(1):77-87.
Sadowski J, Gasteiger J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem. Rev. 1993; 93: 2567-2581.
Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 2012; 17: 4791-4810.
SCCS. Scientific Committee on Consumer Safety. European Commission. OPINION ON 3-Benzylidene camphor. 2013. Available at http://ec.europa.eu/health/scientific_committees/consumer_safety/docs/sccs_o_134.pdf. [Last accessed July 03 2015]
Schebb NH, Flores I, Kurobe T, Franze B, Ranganathan A, Hammock BD, et al. Bioconcentration, metabolism and excretion of triclocarban in larval Qurt medaka (Oryzias latipes). Aquat. Toxicol. 2011; 105: 448-454.
Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. J. Chem. Inf. Comp. Sci. 2004; 44: 1912-1928.

Shin YG, Le H, Khojasteh C, Hop C. Comparison of Metabolic Soft Spot Predictions of CYP3A4, CYP2C9 and CYP2D6 Substrates Using MetaSite and StarDrop. Comb. Chem. High T. Scr. 2011; 14: 811-823.

Skaare J, Solheim E. Studies on the metabolism of the antioxidant ethoxyquin, 6-ethoxy-2,2,4-trimethyl-1,2-dihydroquinoline in the rat. Xenobiotica 1979; 9: 649-57.

Skvortsova MI, Baskin II, Skvortsov LA, Palyulin VA, Zefirov NS, Stankevich IV. Chemical graphs and their basis invariants. J. Mol. Struc-THEOCHEM 1999; 466: 211-217.

Stapleton HM, Kelly SM, Pei R, Letcher RJ, Gunsch C. Metabolism of polybrominated diphenyl ethers (PBDEs) by human hepatocytes in vitro. Environ. Health Persp. 2009; 117: 197-202.

Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. J. Chem. Inf. Comp. Sci. 2003; 43: 493-500.

Sushko I, Novotarskyi S, Korner R, Pandey AK, Cherkasov A, Lo JZ, et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. J. Chem. Inf. Model. 2010; 50: 2094-2111.

Sushko I, Novotarskyi S, Körner R, Pandey A, Rupp M, Teetz W, et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. J. Comput. Aid. Mol. Des. 2011; 25: 533-554.

T'Jollyn H, Boussery K, Mortishire-Smith RJ, Coe K, De Boeck B, Van Bocxlaer JF, et al. Evaluation of Three State-of-the-Art Metabolite Prediction Software Packages (Meteor, MetaSite, and StarDrop) through Independent and Synergistic Use. Drug Metab. Dispos. 2011; 39: 2066-2075.

Tetko IV, Sushko I, Pandey AK, Zhu H, Tropsha A, Papa E, et al. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. J. Chem. Inf. Model. 2008; 48: 1733-1746.

Tetko IV, Sushko Y, Novotarskyi S, Patiny L, Kondratov I, Petrenko AE, et al. How Accurately Can We Predict the Melting Points of Drug-like Compounds? J. Chem. Inf. Model. 2014; 54: 3320-3329.

Tetko IV, Tanchuk VY. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. J. Chem. Inf. Comp. Sci. 2002; 42: 1136-1145.

To-Figueras J, Sala M, Otero R, Barrot C, Santiago-Silva M, Rodamilans M, et al. Metabolism of hexachlorobenzene in humans: association between serum levels and urinary metabolites in a highly exposed population. Environ. Health Persp. 1997; 105: 78-83.

Todeschini R, Consonni V. Handbook of Molecular Descriptors: WILEY-VCH: Weinheim, 2000.

Trunzer M, Faller B, Zimmerlin A. Metabolic Soft Spot Identification and Compound Optimization in Early Discovery Phases Using MetaSite and LC-MS/MS Validation. J. Med. Chem. 2009; 52: 329-335.

U.S.EPA. Reregistration Eligibility Decision (RED) Oryzalin. Available at http://www.epa.gov/pesticides/reregistration/REDs/0186.pdf. [Last accessed July 03 2015]

U.S.EPA. Research and Development: Health and Environmental Effects Document for Selected Chlorinated Toluenes. BiblioGov, U. S. Environmental Protection Agency, 2013.

Umezawa H, Lee XP, Arima Y, Hasegawa C, Marumo A, Kumuzawa T, et al. Determination of diazepam and its metabolites in human urine by liquid chromatography/tandem mass spectrometry using a hydrophilic polymer column. Rapid commun. mass sp. : RCM. 2008; 22: 2333-41.

The University of Waikato. Weka 3: Data Mining Software in Java. Available at http://www.cs.waikato.ac.nz/ml/weka/. [Last accessed July 03 2014]

Varnek A, Fourches D, Horvath D, Klimchuk O, Gaudin C, Vayer P, et al. ISIDA - Platform for virtual screening based on fragment and pharmacophoric descriptors. Curr. Comput-Aid. Drug 2008; 4: 191-198.

Vinggaard AM, Niemelä J, Wedebye EB, Jensen GE. Screening of 397 Chemicals and Development of a Quantitative Structure−Activity Relationship Model for Androgen Receptor Antagonism. Chem. Res. Toxicol. 2008; 21: 813-823.

Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. 2001; 58: 109-130.

Vorberg S, Tetko IV. Modeling the Biodegradability of Chemical Compounds Using the Online CHEmical Modeling Environment (OCHEM). Mol. Inform. 2014; 33: 73-85.

Zhou DS, Afzelius L, Grimm SW, Andersson TB, Zauhar RJ, Zamora I. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. Drug Metab. Dispos. 2006; 34: 976-983.