



Article

# MetStabOn—Online Platform for Metabolic Stability Predictions

Sabina Podlewska \*  and Rafał Kafel

Institute of Pharmacology, Polish Academy of Sciences, Department of Medicinal Chemistry, Smętna Street 12, 31-343 Kraków, Poland; rafal.kafel@gmail.com

\* Correspondence: smusz@if-pan.krakow.pl; Tel.: +48-12-66-23-301

Received: 9 March 2018; Accepted: 28 March 2018; Published: 30 March 2018



**Abstract:** Metabolic stability is an important parameter to be optimized during the complex process of designing new active compounds. Tuning this parameter with the simultaneous maintenance of a desired compound's activity is not an easy task due to the extreme complexity of metabolic pathways in living organisms. In this study, the platform for *in silico* qualitative evaluation of metabolic stability, expressed as half-lifetime and clearance was developed. The platform is based on the application of machine learning methods and separate models for human, rat and mouse data were constructed. The compounds' evaluation is qualitative and two types of experiments can be performed—regression, which is when the compound is assigned to one of the metabolic stability classes (low, medium, high) on the basis of numerical value of the predicted half-lifetime, and classification, in which the molecule is directly assessed as low, medium or high stability. The results show that the models have good predictive power, with accuracy values over 0.7 for all cases, for Sequential Minimal Optimization (SMO), k-nearest neighbor (IBk) and Random Forest algorithms. Additionally, for each of the analyzed compounds, 10 of the most similar structures from the training set (in terms of Tanimoto metric similarity) are identified and made available for download as separate files for more detailed manual inspection. The predictive power of the models was confronted with the external dataset, containing metabolic stability assessment via the GUSAR software, leading to good consistency of results for SMOreg and Naïve Bayes (~0.8 on average). The tool is available online.

**Keywords:** metabolic stability; machine learning; ChEMBL database; regression; classification

## 1. Introduction

During the drug design process, attention is initially placed on obtaining the desired affinity with the appropriate receptors. However, failures of compounds at later stages of drug development are connected with other unfavorable physicochemical, pharmacokinetic, or toxic properties. The proper evaluation of these properties *in silico* is therefore just as important as the development of computational tools for accurate activity predictions [1–4].

A number of parameters can be set on the basis of which drug-like potential of the compounds is evaluated. One of the most popular groups of these properties are Lipinski's Rule of Five, which are one of the simplest and usually the first filters applied to disqualify compounds with an unfavorable physicochemical profile [5]. It is also important to provide proper compound solubility (both in water due to solubility in the fluids in the organism, and in non-polar solvents due to the provision of penetration of biological membranes, as well as the provision of proper equilibrium between the solubility in these two environments [6]), determine the ionization potential [7] and assure that the compound can penetrate the gut-blood and blood-brain barrier (in the case of drugs acting within the central nervous system) [8]. The permeability through the biological membranes is vital not only from

the point of view of determining the proper therapeutic dose but also because of the possible toxic effects. It is also important to analyze whether the compound will bind to the plasma proteins [9] as well as evaluate the half-life time or potential metabolic routes [3]. In terms of toxicity, predictions most often concern the possible interactions of the examined compound with other therapeutics and the possible undesirable modulations of other protein activities, such as hERG potassium channels [10] leading to the compound's cardio toxicity [11,12].

These abovementioned properties are connected with the characterization of compounds in terms of their Absorption, Distribution, Metabolism, Excretion, Toxicity (ADMET) properties [3,13–16]. Metabolism and metabolic stability are particularly considered in the study, as compounds need to have sufficient time to induce the desirable therapeutic effect. Additionally, although metabolites might possess desirable biological activity, transformations of biologically active substances can also lead to the formation of toxic products. Unfortunately, the *in silico* examination of metabolic stability is very difficult due to the extreme complexity of the metabolism process. However, although it is difficult to obtain a broad predictive model that can correctly evaluate compounds that cover a wide structural spectrum, studies on metabolic stability, as well as the construction of *in silico* tools that enable the computational evaluation of metabolic stability, are continuously carried out.

A number of approaches for the prediction of ADMET properties are already available. They are mostly ligand-based tools, and two classes of models are constructed—classification ones (mutagenic/non-mutagenic, stable/unstable, soluble/insoluble, etc. [17–19]) or regression tools [19–21]. The application of the tools of the latter type is connected with the formation of the QSAR-type [22–25] models, in which the quantitative impact of particular structural moieties on considered parameters is examined. Many comprehensive software packages for ADMET properties evaluation are available, such as ADMET Predictor [26], CASE ULTRA [27], DEREK [28], META-PC [29], METEOR [30,31], ONCOLOGIC [32], PASS [33], TOPKAT [34], and GUSAR [35]. Moreover, the initial characteristics of physicochemical and pharmacokinetic properties are offered in most packages of software for molecular modeling, such as QikProp in the Schrödinger Suite [36], Molecular Descriptors in the MOE [37], or ADMET and Predictive Toxicology from the BIOVIA Discovery Studio [34]. A number of individual ADMET properties can also be evaluated via various online servers, such as ALOGPS [38], Molinspiration [39] PreADMET [40], MetaPrint2D [41–43], MetaPred [44] or Pred-hERG [45]. A summary of some of the available tools is shown in Table 1.

**Table 1.** A summary of some of the available tools for ADMET properties predictions.

Package Name	Link	Availability	Description
ADMET Predictor	<a href="http://www.simulations-plus.com/">http://www.simulations-plus.com/</a>	commercial software	Comprehensive characteristic of physicochemical and ADMET properties of compounds, including cancerogenicity, mutagenicity, overall toxicity and possibility of interactions with 5 selected CYP isoforms
CASE ULTRA	<a href="http://www.multicase.com/case-ultra-models">http://www.multicase.com/case-ultra-models</a>	commercial software	A set of statistical and expert tools for evaluation of compounds toxicity
DEREK	<a href="http://www.lhasalimited.org">http://www.lhasalimited.org</a>	commercial software	Expert system for predicting toxicity of compounds, including cancerogenicity, mutagenicity, genotoxicity, teratogenicity, influence on fertility, irritating influence on skin or allergic effect
META-PC	<a href="http://www.multicase.com/meta-pc">http://www.multicase.com/meta-pc</a>	commercial software	Expert system for predicting products of compounds metabolism
METEOR	<a href="http://www.lhasalimited.org">http://www.lhasalimited.org</a>	commercial software	Expert system for predicting metabolic transformations
ONCOLOGIC	<a href="http://www2.epa.gov/tsca-screening-tools/oncologictm-computer-system-evaluate-carcinogenic-potential-chemicals">http://www2.epa.gov/tsca-screening-tools/oncologictm-computer-system-evaluate-carcinogenic-potential-chemicals</a>	free software	Predicting of cancerogenicity of compounds

Table 1. Cont.

Package Name	Link	Availability	Description
PASS	<a href="http://www.pharmaexpert.ru">http://www.pharmaexpert.ru</a>	commercial software (simplified version is freely available online)	Qualitative evaluation of above 3500 properties, including mechanisms of action, side and toxic effects, interaction with various enzymes and transport proteins, influence on genes expression
TOPKAT	<a href="http://accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html">http://accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html</a>	commercial software	Mutagenicity, cancerogenity, irritating action on skin, eyes, etc.
GUSAR	<a href="http://www.way2drug.com/gusar/index.html">http://www.way2drug.com/gusar/index.html</a>	commercial software	Evaluation of compounds toxicity and interaction with selected off-targets

In this study, we focus on one of the compound's parameters—metabolic stability. Usually, the tools for ADMET properties evaluation do not consider this property, as its predictions are very difficult due to the extreme complexity of the metabolic stability phenomenon and the great number of factors influencing this parameter. On the other hand, metabolic stability is a very important parameter, as the compound requires sufficient time to trigger the desired pharmacological response before its decomposition and moreover, the formed metabolites might not only be unable to provide the biological action of interest, but they can be toxic. For the evaluation of metabolic stability on the basis of experimental data, the most often used are half-lifetime ( $T_{1/2}$ ) data produced in assays using liver microsomes. There were already several machine learning-based approaches to construct general QSAR models for prediction of this parameter [46–49] and several other studies based on narrower chemical space of selected classes of compounds [50–54].

Here, we present a freely available online tool for metabolic stability predictions expressed as  $T_{1/2}$  or clearance. The tool uses ligand-based methodology and six machine learning methods to evaluate the compound stability, with separate models constructed for various species and assays performed on liver microsomes and plasma. Additionally, an analysis of available metabolic stability data was performed and the most similar compounds from the training set are provided for each of the submitted structures, enabling manual analysis of the results and comparisons. The outcome of the constructed tool was compared with the external National Cancer Institute (NCI) dataset containing a GUSAR-based evaluation of metabolic stability [35]. The tool and all obtained results are freely available at [http://skandal.if-pan.krakow.pl/met\\_stab\\_pred/](http://skandal.if-pan.krakow.pl/met_stab_pred/).

## 2. Materials and Methods

The data for the construction of the tool for metabolic stability predictions were collected from the ChEMBL database version 23 [55]. All records with the  $T_{1/2}$  and intrinsic clearance parameters reported were downloaded. Data preprocessing involved the following steps: selection of *in vitro* assays performed on liver microsomes or plasma, selection of records with standard unit “hr” for  $T_{1/2}$  and “ $\text{mL}\cdot\text{min}^{-1}\cdot\text{g}^{-1}$ ”, and division into separate sets referring to human, rat and mouse experiments. Due to the lack of sufficient number of entries for clearance data obtained via plasma-based assays, for plasma data, only records for  $T_{1/2}$  were gathered. The number of data points present in the respective datasets is gathered in Table 2.

Table 2. The number of compounds present in each dataset used for the predictive model's construction.

	Human		Rat		Mouse	
	Liver Microsomes	Plasma	Liver Microsomes	Plasma	Liver Microsomes	Plasma
$T_{1/2}$	2127	561	1308	277	808	62
clearance	2546	-	1244	-	266	-

The compounds were represented with the use of the 1- and 2-dimensional PaDEL-Descriptors [56] (1d2d descriptors) and Extended Fingerprint [57] (ExtFP) from the same software package. These forms

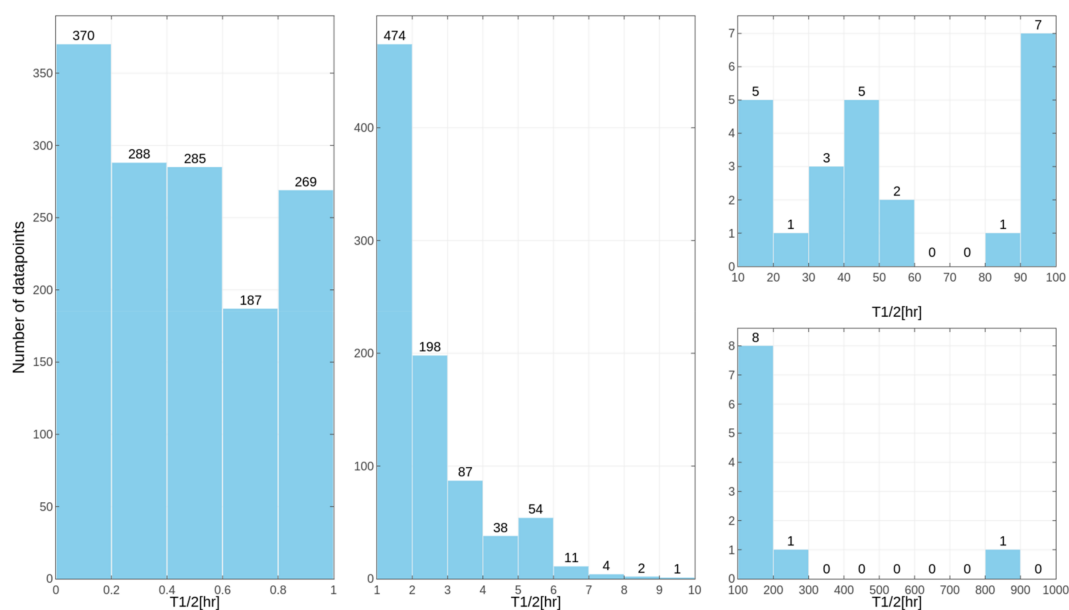
of representation were chosen after initial studies performed on a series of long-chain aryl piperazines, where several fingerprints implemented in PaDEL (Extended Fingerprint, MACCS Fingerprint [58], Klekota-Roth Fingerprint [59], Graph Fingerprint [57], PubChem Fingerprint [60], and Substructure Fingerprint [61]) and descriptor sets were tested. Three dimensional descriptors were not included due to the relatively high fraction of compounds for which they could not be calculated due to errors in their generation.

The constructed tool predicts the numerical value of metabolic stability with the predictive model based on the application of the two types of machine learning algorithms, regression and classification. From the first group of methods, one algorithm was used: SMOreg, which is a modification of the very popular and efficient algorithm Support Vector Machine (SVM) [62] into Sequential Minimal Optimization (SMO) [63] and adjusted for performing regression tasks. Additionally, five classification algorithms were used: SMO, Random Forest [64], Naïve Bayes [65], k-nearest neighbor (IBk) [66], and decision tree J48 [67]. However, in order to enable easier interpretability of the outcome of regression experiments, compounds are also divided into three classes according to metabolic stability values—low, medium, and high—and the results are colored accordingly. For each of the analyzed structures, the 10 most similar compounds from the training set (Tanimoto metric [68], topological fingerprint from RDKit package [69]) are found and provided in separate files for manual inspection (the particular chemical structure is provided only once, and the median half-lifetime value is given). Structures for analysis online can be submitted as a structure data file (sdf) or drawn using the MarvinJS [70] plugin. Regardless of the way the query is submitted, all the structures are shown in the results.

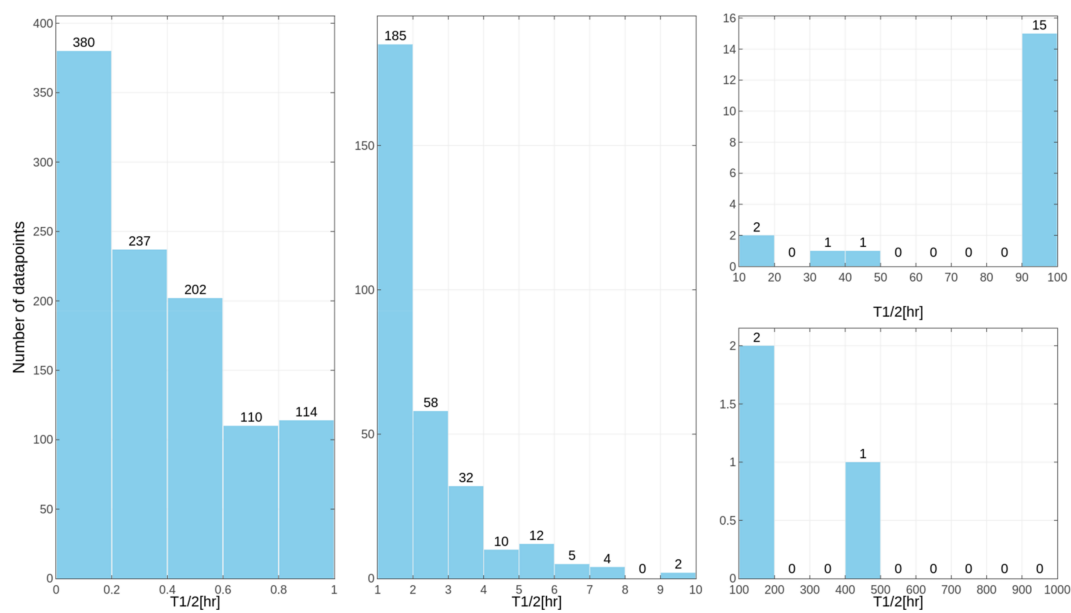
### 3. Results and Discussion

#### 3.1. The Importance of Separate Models for Different Species

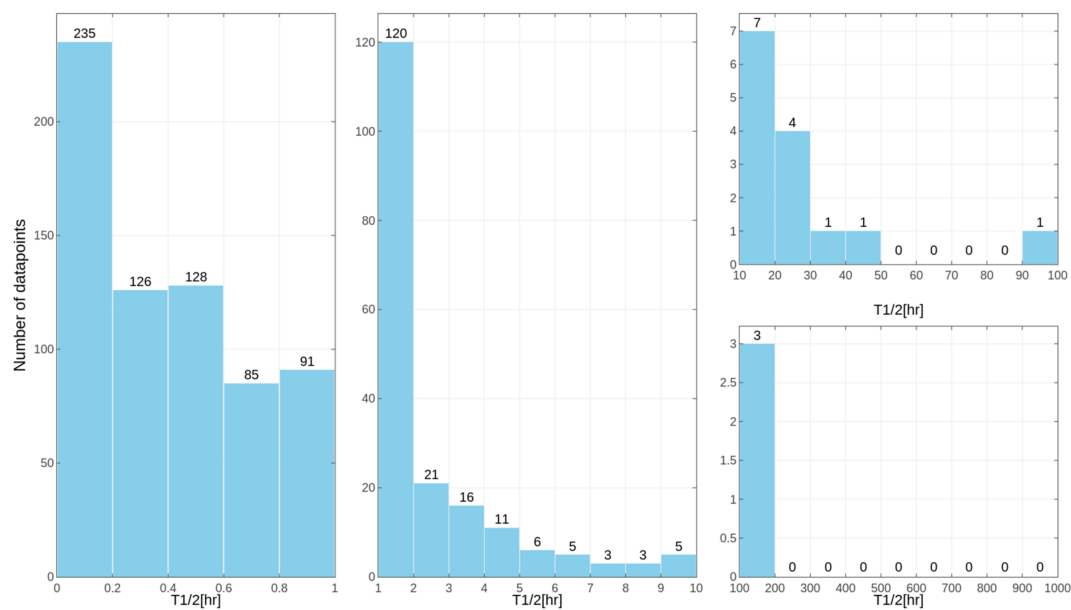
Separate models were constructed for different species. This approach was applied due to the relatively high differences in results of the *in vitro* tests for some of the compounds, despite the similar overall distribution of data points for human, rat and mouse data with the majority of very unstable compounds (Figures 1–3 presenting the distribution of  $T_{1/2}$ ). Some examples of the abovementioned problems are shown in Figure 4.



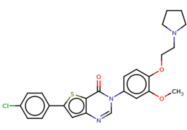
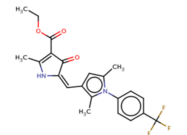
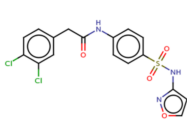
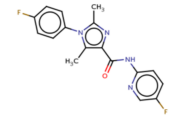
**Figure 1.** Distribution of compound half-lifetimes in the constructed datasets referring to experiments performed on human samples. For better visualization, the dataset was divided into several parts.



**Figure 2.** Distribution of compound half-lifetimes in the constructed datasets referring to experiments performed on rat samples. For better visualization, the dataset was divided into several parts.



**Figure 3.** Distribution of compound half-lifetimes in the constructed datasets referring to experiments performed on mouse samples. For better visualization, the dataset was divided into several parts.

CHEMBL ID	Compound structure	T <sub>1/2</sub> [hr]		
		human	rat	mouse
CHEMBL214957		10.05	3.267	9.567
CHEMBL2335990		7.25	0.9333	0.5483
CHEMBL3108858		5.5	0.5	3.85
CHEMBL2346736		5	1.833	0.5333

**Figure 4.** Examples of differences in results of metabolic stability tests for human, rat and mouse models.

For all the compounds presented in the Figure 4, there were substantial differences in the results of metabolic stability examinations based on different species. For example, for compound CHEMBL214957, the human and mouse-based data were quite consistent (with stability experiments outcome being 10.05 and 9.57 hr, respectively); however, for experiments, in which rat liver microsomes were used, the obtained T<sub>1/2</sub> value was equal to 3.27. For compound CHEMBL2335990, more consistency was observed for rat and mouse-based experiments with the stability values of 0.93 and 0.55 hr, respectively, whereas in experiments using human microsomes, the obtained T<sub>1/2</sub> value was equal to 7.25 hr. For the other two compounds presented in Figure 4, that is CHEMBL3108858 and CHEMBL2346736, no similarities between any two experiments were observed, and the T<sub>1/2</sub> values varied from 0.5 to 5.5 hr.

Therefore, the data referring to human, rat and mouse-based experiments were not mixed, and separate models were constructed for each of these experimental conditions.

In order to provide a more general picture of variations in experimental results of metabolic stability for different models (human, rat, mouse), all compounds for which the half-lifetimes were provided for all the models were identified and standard deviations ( $\sigma$ ) given by the following equation were calculated:

$$\sigma = \sqrt{\frac{\sum_i^3 (x_i - \mu)^2}{3}} \quad (1)$$

where  $x_i$  is half-lifetime value measured for a particular model;  $\mu$ —mean value of all three measures (human, mouse, rat) provided for a particular compound

The obtained values of standard deviations were presented in Figure 5 (the ChEMBL identifiers of compounds with metabolic stability data for human, rat, and mouse models with standard deviation values are provided in the Supplementary Materials).

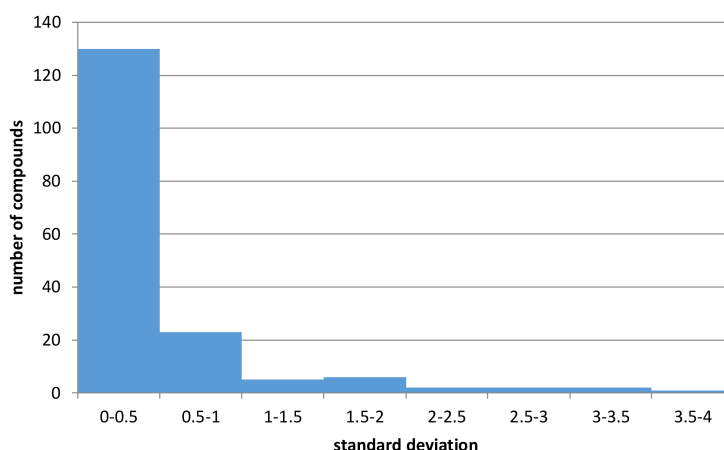


Figure 5. Standard deviation values of half-lifetimes between human, rat and mouse data.

The analysis of the histogram in Figure 5 indicates and confirms the relatively high variation of data. Out of 249 structures containing data referring to all three experimental conditions (human, rat, mouse) to which Figure 5 refers, for over 16% of them, standard deviation of half-lifetimes values was higher than 0.5, leading to swap of metabolic classes. Moreover, for 172 structures, it was above 0. Additionally, although for datasets containing human-based data, the percentage of datapoints with non-zero standard deviation was equal to 8% for mouse-based datasets, which contained much less records, the percentage of non-zero data was equal to 20%. All these results support the construction of separate predictive models for human, rat, and mouse-based data.

A scheme of all the predictive approaches that were used in the study is presented in Figure 6. Taking into account the combination of all datasets used, prediction algorithms and compounds representation, the total number of predictive models provided as a result of the study is equal to 108; the online version of the tool for metabolic stability predictions includes 36 models referring to experimental data produced on liver microsomes and reported as  $T_{1/2}$ . Remaining models are available at [http://skandal.if-pan.krakow.pl/met\\_stab\\_pred/](http://skandal.if-pan.krakow.pl/met_stab_pred/) and can directly be used within the WEKA software [71].

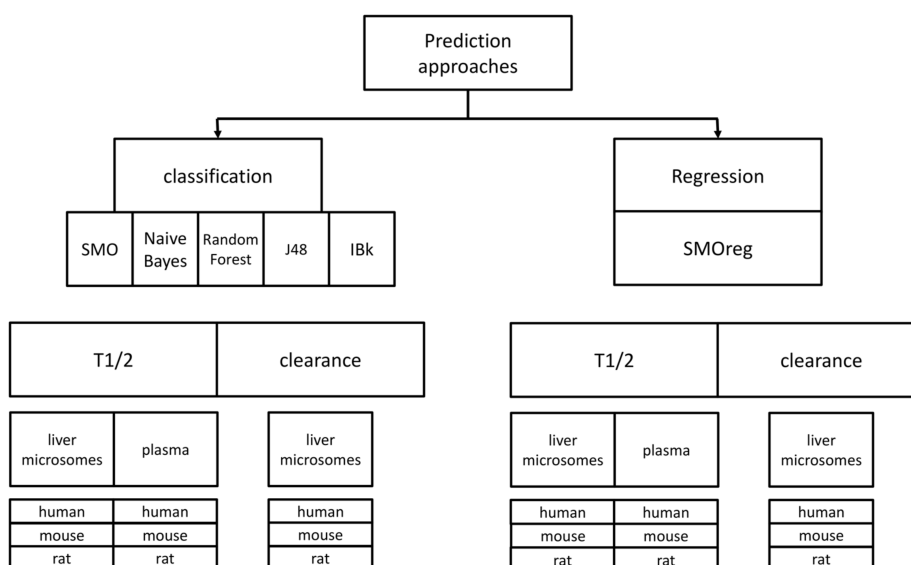


Figure 6. Scheme of the prediction approaches covered in the study.

### 3.2. Evaluation of Models in 10-Fold Cross-Validation

Ten-fold cross-validation (CV) studies were performed for optimization and evaluation of the constructed predictive models. WEKA implementation [71] was used for all the algorithms used in the study.

The division into metabolic stability classes was applied, and evaluation parameters were subsequently calculated: overall accuracy, AUROC [72] in case of classification models, and recall and precision for each class.

The cutoffs for metabolic stability class division were as follows ( $T_{1/2}$  expressed in hours):

- $\leq 0.6$ —low
- $(0.6-2.32)>$ —medium
- $> 2.32$ —high

The number of compounds belonging to each stability class are presented in Table 3.

**Table 3.** Statistics of number of compounds belonging to each class.

Class/Number of Compounds	Human	Rat	Mouse
Low	928 (44%)	814 (62%)	486 (60%)
Medium	937 (44%)	382 (29%)	252 (31%)
High	262 (12%)	112 (9%)	70 (9%)
Total	2127	1308	808

The class distribution was not uniform and in all cases the highest number of compounds belonged to the group of low metabolic stability. The highest variation was observed for rat and mouse models, with 62% and 60% of compounds exhibiting low values of half-lifetimes, respectively. For both of these cases, compounds of high stability constituted 9% of the whole dataset, and the fraction of medium stability compounds in the whole dataset was equal to 29% and 31% for rat and mouse models, respectively. For human models, the compounds were more uniformly distributed among classes with 44% of compounds belonging to low and medium stability class, and the remaining 12% of datapoints referred to compounds characterized by high metabolic stability.

The values of evaluating parameters obtained in 10-fold CV studies are presented in Table 4 for  $T_{1/2}$  liver microsomes data, all results for the remaining datasets are present in the Supplementary Materials.

The parameters values above 0.7 are presented in bold. In order to facilitate the results interpretation, the data were also presented in the respective figures (Figure 7 for  $T_{1/2}$  liver microsomes; remaining data in Supplementary Materials).

**Table 4.** Evaluation parameters obtained in 10-fold CV for data ( $T_{1/2}$ ) produced on liver microsomes. Values above 0.7 are depicted in bold.

		1d2d Descriptors						ExtFP						
		Class	SMOreg	SMO	IBk	Naïve Bayes	Random Forest	J48	SMOreg	SMO	IBk	Naïve Bayes	Random Forest	J48
human	Recall	Low	0.176	<b>0.792</b>	<b>0.773</b>	0.344	<b>0.778</b>	0.696	0.650	<b>0.768</b>	<b>0.761</b>	0.626	<b>0.775</b>	<b>0.745</b>
		Medium	<b>0.872</b>	<b>0.733</b>	<b>0.713</b>	<b>0.701</b>	<b>0.753</b>	0.679	<b>0.834</b>	<b>0.743</b>	0.697	0.508	<b>0.740</b>	0.673
		High	0.561	0.573	0.557	0.480	0.447	0.463	0.389	0.515	0.592	0.603	0.527	0.500
	Precision	Low	<b>0.873</b>	<b>0.781</b>	<b>0.759</b>	0.688	<b>0.766</b>	0.695	<b>0.841</b>	<b>0.778</b>	<b>0.749</b>	0.693	<b>0.777</b>	<b>0.745</b>
		Medium	0.477	<b>0.725</b>	<b>0.709</b>	0.520	0.692	0.649	0.618	0.695	<b>0.700</b>	0.631	<b>0.701</b>	0.676
		High	0.573	0.632	0.609	0.310	<b>0.710</b>	0.562	0.662	0.643	0.615	0.295	0.648	0.491
Overall accuracy			0.524	<b>0.739</b>	<b>0.720</b>	0.517	<b>0.726</b>	0.660	0.698	<b>0.725</b>	<b>0.711</b>	0.571	<b>0.728</b>	0.682
AUROC				<b>0.836</b>	<b>0.800</b>	<b>0.708</b>	<b>0.886</b>	<b>0.7333</b>		<b>0.821</b>	<b>0.792</b>	<b>0.757</b>	<b>0.881</b>	<b>0.781</b>



Table 4. Cont.

		1d2d Descriptors						ExtFP						
	Class	SMOreg	SMO	IBk	Naïve Bayes	Random Forest	J48	SMOreg	SMO	IBk	Naïve Bayes	Random Forest	J48	
rat	Recall	Low	0.467	<b>0.903</b>	<b>0.877</b>	<b>0.768</b>	<b>0.935</b>	<b>0.821</b>	0.565	<b>0.904</b>	<b>0.870</b>	<b>0.713</b>	<b>0.903</b>	<b>0.827</b>
		Medium	<b>0.752</b>	0.617	0.644	0.114	0.561	0.542	0.427	0.605	0.607	0.573	0.586	0.576
		High	0.514	0.400	0.476	0.648	0.228	0.390	0.598	0.384	0.429	0.553	0.348	0.429
	Precision	Low	<b>0.896</b>	<b>0.838</b>	<b>0.848</b>	<b>0.748</b>	<b>0.799</b>	<b>0.804</b>	<b>0.759</b>	<b>0.837</b>	<b>0.834</b>	<b>0.841</b>	<b>0.828</b>	<b>0.809</b>
		Medium	0.379	0.680	0.663	0.512	0.694	0.570	0.354	0.672	0.637	0.489	0.655	0.598
		High	0.422	0.545	0.568	0.181	0.615	0.387	0.276	0.506	0.505	0.365	0.500	0.444
	Overall accuracy		0.553	<b>0.777</b>	<b>0.775</b>	0.566	<b>0.767</b>	<b>0.704</b>	0.528	<b>0.771</b>	<b>0.754</b>	0.657	<b>0.762</b>	<b>0.718</b>
	AUROC			<b>0.819</b>	<b>0.813</b>	0.698	<b>0.912</b>	<b>0.773</b>		<b>0.817</b>	<b>0.870</b>	<b>0.821</b>	<b>0.906</b>	<b>0.774</b>
	mouse	Recall	Low	0.570	<b>0.881</b>	<b>0.834</b>	0.650	<b>0.896</b>	<b>0.772</b>	<b>0.706</b>	<b>0.825</b>	<b>0.860</b>	<b>0.718</b>	<b>0.872</b>
Medium			<b>0.796</b>	0.601	0.622	0.248	0.521	0.584	0.723	0.667	0.615	0.615	0.579	0.591
High			0.500	0.329	0.486	0.728	0.357	0.271	0.529	0.557	0.343	0.500	0.457	0.386
Precision		Low	<b>0.888</b>	<b>0.781</b>	<b>0.808</b>	<b>0.738</b>	<b>0.753</b>	<b>0.780</b>	<b>0.848</b>	<b>0.815</b>	<b>0.787</b>	<b>0.841</b>	<b>0.782</b>	<b>0.766</b>
		Medium	0.448	0.678	0.643	0.546	0.685	0.565	0.525	0.648	0.654	0.525	0.679	0.575
		High	0.614	0.622	0.540	0.199	0.658	0.284	0.638	0.684	0.600	0.357	0.627	0.509
Overall accuracy			0.632	<b>0.743</b>	<b>0.736</b>	0.533	<b>0.730</b>	0.667	0.696	<b>0.751</b>	<b>0.737</b>	0.665	<b>0.743</b>	0.686
AUROC				<b>0.753</b>	<b>0.781</b>	0.673	<b>0.872</b>	<b>0.729</b>		<b>0.776</b>	<b>0.846</b>	<b>0.809</b>	<b>0.848</b>	<b>0.742</b>

In general, the values of evaluating parameters are high and show that the constructed models are capable of making a valid evaluation of metabolic stability expressed as  $T_{1/2}$ .

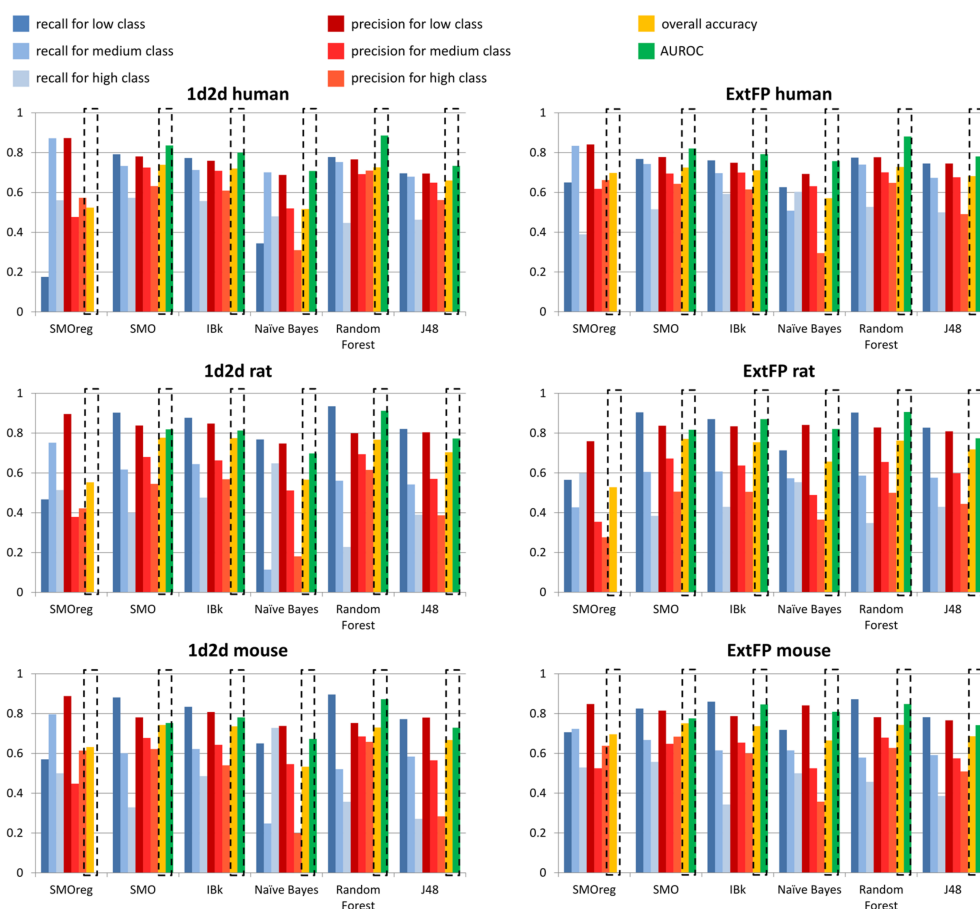


Figure 7. Visualization of evaluating parameters values obtained in 10-fold CV studies ( $T_{1/2}$  liver microsomes data).

The general observation is that 1d2d descriptors provided higher values of evaluating parameters than ExtFP. It is also visible that there are high variations for recall and precision values depending on the metabolic stability class considered. However, for the great majority of cases, low stability class was the one that led to the highest recall and precision values. However, the highest number of cases belonged to the low stability class in all human, mouse and rat datasets. Relatively, the highest recall values were obtained for rat models, for Random Forest and SMO models (0.935 and 0.903, respectively) with compounds represented by 1d2d descriptors; for ExtFP representation, the recall values were on similar level for these methods: 0.903, and 0.904, respectively. On the other hand, it was IBk (for 1d2d descriptors) and Naïve Bayes (for ExtFP) that provided the highest precision values: 0.848 and 0.841, respectively.

Taking into account the overall accuracy of predictions, the SMO, IBk and Random Forest methods were the only ones that consistently provided overall accuracy over 0.7. For human models, it was 0.739 for SMO and 0.728 for Random Forest (for 1d2d descriptors and ExtFP representations, respectively) that were the highest overall accuracies; however, for mouse and rat data, it was SMO that provided the best performance of predictive models for both compounds representations, with values varying from 0.743 to 0.777. The other parameter that provided information on the general performance of the model was AUROC, which in general adopted higher values than overall accuracy, reaching values close to 0.9: 0.886 and 0.881 for Random Forest models constructed on human data, 0.912 and 0.906 for Random Forest models that used mouse data, and 0.872 and 0.848 for Random Forest model built on rat data (for 1d2d descriptors and ExtFP representations, respectively).

### 3.3. Comparison of the Constructed Tool Outcome with the Predictions on the External Dataset

The outcome of the constructed tool was compared with the external NCI dataset containing GUSAR-based evaluation of metabolic stability [35]. The predictions provided there were only binary (stable/unstable). After the removal of compounds with errors, and those containing heavy atoms, such as Pb, Ag, Se, Te (leading to errors in descriptors calculations), the predictions with models produced on  $T_{1/2}$  liver microsomes data were carried out (as the dataset originally contained such data). The results obtained via these two approaches were compared by accuracy indicating the fraction of the same predictions (Table 5).

**Table 5.** Accuracies of predictions on external test set ( $T_{1/2}$  human data on liver microsomes). Values above 0.7 are depicted in bold.

		SMOreg	SMO	IBk	Naïve Bayes	Random Forest	J48
Medium class predictions removed	1d2d descriptors	<b>0.89</b>	0.58	0.22	<b>0.81</b>	0.61	0.51
	ExtFP	0.39	0.27	0.44	<b>0.72</b>	0.23	0.38
Medium class predictions shifted to high class	1d2d descriptors	<b>0.78</b>	0.66	0.22	<b>0.77</b>	<b>0.71</b>	0.58
	ExtFP	<b>0.77</b>	0.64	0.61	<b>0.72</b>	0.54	0.53

The comparison of the predictions obtained via the constructed tool with the GUSAR predictions indicate great dependence of the results on the machine learning algorithm applied. The most consistent predictions with the output of the GUSAR software were provided by the SMOreg and Naïve Bayes algorithms. As the GUSAR predictions were binary (the compounds were evaluated only as stable or unstable) and our tool evaluates compounds as low, medium or high stability, two approaches of dealing with records assigned to medium stability class were applied: such entries were removed before the accuracy calculation, or the medium class assignments were shifted to the high stability class, as being more populated. With the removal of records assigned to medium stability class, the most consistent predictions with the GUSAR software outcome were obtained with the use of the SMOreg algorithm with 1d2d descriptors used for compounds representation (accuracy of 0.89). Surprisingly, the application

of SMOreg with ExtFP for compounds representation led to high inconsistency of the constructed tool outcome with the GUSAR-based evaluation with the accuracy of 0.39. The GUSAR software output was also in line with the Naïve Bayes predictions, for both 1d2d descriptors and ExtFP compounds representation, with accuracies of 0.81 and 0.72, respectively. In general, for sets with medium class predictions removed, the application of 1d2d descriptors for compounds representation led to more consistent results with the GUSAR software than ExtFP (for all algorithms but IBk, the accuracies were much higher for the former compounds representation, by from 0.13 for J48, through 0.38 for Random Forest, to 0.50 for SMOreg). The advantage of the 1d2d descriptors representation over ExtFP was not visible for the situation, when the medium class predictions were manually shifted to the group of records referring to high stability compounds. For the most consistent with the GUSAR predictions, SMOreg, the difference in accuracy between these two forms of representations was only 0.01 (0.78 vs. 0.77). For Naïve Bayes, which also produced results consistent with the GUSAR software, the difference was equal to 0.05 (accuracy of 0.77 vs. 0.72 for 1d2d descriptors and ExtFP, respectively). Random Forest predictions were in line with the output of the GUSAR software only for the 1d2d descriptors (0.71 accuracy for this representation vs. 0.54 for ExtFP).

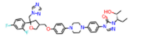
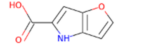
### 3.4. Example Results

A screenshot from the example output of the tool is presented in Figure 8. The summary describes the compound representation used, the predictive model applied, the number of input compounds and the number of compounds assigned to a particular metabolic stability class. Detailed results are gathered in a table with the structure and simplified molecular-input line-entry system (SMILES) of a compound, the predicted value of half-life and the metabolic stability class to which a compound was assigned. Additionally, in order to perform a more detailed analysis, the 10 most similar compounds from the training set (in terms of Tanimoto metric-based similarity) can be downloaded for each of the analyzed structures (due to the computational resources limitations, this option is available for a maximum number of 100 structures submitted at one time). As the training sets contained all available data, in order to prevent situations in which the number of structures for manual inspection is very restricted, the particular chemical structure was listed only once, and in case of multiple metabolic stability entries available, median values of half-lifetimes were provided. The abovementioned data can be downloaded separately for each of the analyzed compounds or as one zipped directory.

**Summary**

Representation: 1d2d  
Model: SMO\_human

Number of all compounds: 4  
Number of compounds with high metabolic stability: 1  
Number of compounds with medium metabolic stability: 2  
Number of compounds with low metabolic stability: 1

Inst	Structure	Smiles	Predicted half-life [hr]	Metabolic stability	Similar cmds
1		<chem>CCC(O)C)n1c(=O)n(-c2ccc(N3CCN(c4ccc(OCC5COC(Cn6cncn6)(c6c(F)cc(F)cc6)C5)cc4)CC3)cc2)cn1</chem>	1.1	low	<a href="#">Download</a>
2		<chem>O=C(O)c1[nH]c2ccoc2c1</chem>	3.3	high	<a href="#">Download</a>

**Figure 8.** Screenshot from example results, containing summary of predictions and predicted metabolic stabilities with coloring corresponding to metabolic stability results.

#### 4. Materials and Methods

The models' parameters were optimized with the set of values gathered in Table 6. For each model, the value that provided the lowest overall accuracy was selected. The conditions selected for each model are shown in Table 7.

The overall accuracy, recall and precision were calculated using the following equations:

$$\text{overall accuracy} = \frac{\text{number of correct predictions}}{\text{number of all predictions}} \quad (2)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

where TP is the number of instances correctly assigned to a particular class, FN is the number of instances belonging to particular class incorrectly assigned to another one, and FP is the number of instances incorrectly assigned to particular class.

**Table 6.** Optimization conditions for SMOreg, SMO, Random Forest, and IBk.

Method	Parameter	Tested Values
SMOreg/SMO	C	0.01, 0.1, 1, 10, 100, 1000
	Gamma	0.001, 0.01, 0.1, 1, 10
	Operations on data	Normalization standardization
Random Forest	Number of trees	10, 100, 1000
IBk	Number of nearest neighbors	1, 2, 3, 4, 5

**Table 7.** Conditions selected for each model.

Compounds Representation	Method	Parameter	Human	Rat	Mouse
1d2d descriptors	SMOreg	C	0.01	0.1	0.1
		gamma	0.1	0.1	0.1
		Operations on data	normalization	normalization	normalization
ExtFP	SMO	C	0.1	1	1
		gamma	0.001	0.001	0.001
		Operations on data	standardization	standardization	standardization
1d2d descriptors	SMO	C	100	100	100
		gamma	0.01	0.1	0.1
		Operations on data	normalization	normalization	normalization
ExtFP	SMO	C	10	10	10
		gamma	0.01	0.01	0.001
		Operations on data	normalization	normalization	normalization
1d2d descriptors	Random Forest	Number of trees	1000	1000	1000
ExtFP			1000	100	100
1d2d descriptors	IBk	Number of nearest neighbors	1	1	1
ExtFP			1	5	1

#### 5. Conclusions

In summary, a tool for the qualitative evaluation of metabolic stability expressed as half-lifetime was constructed. It uses regression and classification tools to provide the assignment of a compound to a particular stability class (low, medium, high), 1d2d descriptors and ExtFP for compound representation,

and the SMOreg, Random Forest, SMO, IBk, Naïve Bayes and J48 machine learning algorithms for making predictions. The tool is freely available online and allows for the submission of structures via sdf files or through drawing. Separate predictive models were constructed for human, rat and mouse data, and for data obtained in experiments using liver microsomes and plasma, as well as for data with metabolic stability expressed as  $T_{1/2}$  and clearance. A detailed retrospective analysis and the application of the constructed model to the external dataset proved the usefulness of the developed tool. The tool can be very useful in designing new potential drugs, and in enabling a fast initial evaluation of a compound's metabolic stability.

**Supplementary Materials:** Supplementary materials can be found at <http://www.mdpi.com/1422-0067/19/4/1040/s1>.

**Acknowledgments:** The study was supported by the National Science Centre, Poland within the HARMONIA 7 grant 2015/18/M/NZ7/00377.

**Author Contributions:** All authors designed and performed experiments, analyzed data and wrote the paper.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

## Abbreviations

1d2d	1-and 2-dimensional descriptors
ExtFP	Extended Fingerprint
SVM	Support Vector Machine
SMO	Sequential Minimal Optimization
CV	Cross-validation

## References

1. Rankovic, Z. CNS Drug Design: Balancing Physicochemical Properties for Optimal Brain Exposure. *J. Med. Chem.* **2015**, *58*, 2584–2608. [[CrossRef](#)] [[PubMed](#)]
2. Kerns, E.H.; Di, L. *Drug-Like Properties: Concepts, Structure Design and Methods from ADME to Toxicity Optimization*; Elsevier: New York, NY, USA, 2008; ISBN 978-0-12-369520-8.
3. Masimirembwa, C.M.; Bredberg, U.; Andersson, T.B. Metabolic stability for drug discovery and development: Pharmacokinetic and biochemical challenges. *Clin. Pharmacokinet.* **2003**, *42*, 515–528. [[CrossRef](#)] [[PubMed](#)]
4. Thompson, T.N. Optimization of metabolic stability as a goal of modern drug design. *Med. Res. Rev.* **2001**, *21*, 412–449. [[CrossRef](#)] [[PubMed](#)]
5. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3–25. [[CrossRef](#)]
6. Di, L.; Fish, P.V.; Mano, T. Bridging solubility between drug discovery and development. *Drug Discov. Today* **2012**, *17*, 486–495. [[CrossRef](#)] [[PubMed](#)]
7. Prankerd, R.J. Critical Compilation of pKa Values for Pharmaceutical Substance. In *Profiles of Drug Substances Excipients and Related Methodology*, 1st ed.; Academic Press: Cambridge, MA, USA, 2008; Volume 33, pp. 1–33, ISBN 978-0-12-260833-9.
8. Sugano, K.; Kansy, M.; Artursson, P.; Avdeef, A.; Bendels, S.; Di, L.; Ecker, G.F.; Faller, B.; Fischer, H.; Gerebtzoff, G.; et al. Coexistence of passive and carrier-mediated processes in drug transport. *Nat. Rev. Drug Discov.* **2010**, *9*, 597–614. [[CrossRef](#)] [[PubMed](#)]
9. Trainor, G.L. The importance of plasma protein binding in drug discovery. *Expert Opin. Drug Discov.* **2007**, *2*, 51–64. [[CrossRef](#)] [[PubMed](#)]
10. Aronov, A.M. Predictive in silico modeling for hERG channel blockers. *Drug Discov. Today* **2005**, *10*, 149–155. [[CrossRef](#)]
11. Van de Waterbeemd, H.; Gifford, E. ADMET in silico modelling: Towards prediction paradise? *Nat. Rev. Drug Discov.* **2003**, *2*, 192–204. [[CrossRef](#)] [[PubMed](#)]

12. Hughes, J.P.; Rees, S.; Kalindjian, S.B.; Philpott, K.L. Principles of early drug discovery. *Br. J. Pharmacol.* **2011**, *162*, 1239–1249. [[CrossRef](#)] [[PubMed](#)]
13. Patil, P.S. Drug Discovery and ADMET process: A Review. *Int. J. Adv. Res. Biol. Sci.* **2016**, *3*, 181–192.
14. Wang, J. Comprehensive assessment of ADMET risks in drug discovery. *Curr. Pharm. Des.* **2009**, *15*, 2195–2219. [[CrossRef](#)] [[PubMed](#)]
15. Wang, Y.; Xing, J.; Xu, Y.; Zhou, N.; Peng, J.; Xiong, Z.; Liu, X.; Luo, X.; Luo, C.; Chen, K.; et al. In silico ADME/T modelling for rational drug design. *Q. Rev. Biophys.* **2015**, *48*, 488–515. [[CrossRef](#)] [[PubMed](#)]
16. Li, D.; Chen, L.; Li, Y.; Tian, S.; Sun, H.; Hou, T. ADMET Evaluation in Drug Discovery. 13. Development of in silico Prediction Models for P-Glycoprotein Substrates. *Mol. Pharm.* **2014**, *11*, 716–726. [[CrossRef](#)] [[PubMed](#)]
17. Wang, N.N.; Dong, J.; Deng, Y.H.; Zhu, M.F.; Wen, M.; Yao, Z.J.; Lu, A.P.; Wang, J.B.; Cao, D.S. ADME Properties Evaluation in Drug Discovery: Prediction of Caco-2 Cell Permeability Using a Combination of NSGA-II and Boosting. *J. Chem. Inf. Model.* **2016**, *56*, 129–138. [[CrossRef](#)] [[PubMed](#)]
18. Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharm.* **2012**, *9*, 996–1010. [[CrossRef](#)] [[PubMed](#)]
19. Pires, D.E.V.; Blundell, T.L.; Ascher, D.B. pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures. *J. Med. Chem.* **2015**, *58*, 4066–4072. [[CrossRef](#)] [[PubMed](#)]
20. Yadav, D.K.; Khan, F. QSAR, docking and ADMET studies of camptothecin derivatives as inhibitors of DNA topoisomerase-I. *J. Chemom.* **2013**, *27*, 21–33. [[CrossRef](#)]
21. Cheng, F.; Li, W.; Zhou, Y.; Shen, J.; Wu, Z.; Liu, G.; Lee, P.W.; Tang, Y. admetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* **2012**, *52*, 3099–3105. [[CrossRef](#)] [[PubMed](#)]
22. Wang, T.; Wu, M.B.; Lin, J.P.; Yang, L.R. Quantitative structure-activity relationship: Promising advances in drug discovery platforms. *Expert Opin. Drug Discov.* **2015**, *10*, 1283–1300. [[CrossRef](#)] [[PubMed](#)]
23. Danishuddin, K.A.U. Descriptors and their selection methods in QSAR analysis: Paradigm for drug design. *Drug Discov. Today* **2016**, *21*, 1291–1302. [[CrossRef](#)] [[PubMed](#)]
24. Nikolic, K.; Mavridis, L.; Djikic, T.; Vucicevic, J.; Agbaba, D.; Yelekci, K.; Mitchell, J.B. Drug Design for CNS Diseases: Polypharmacological Profiling of Compounds Using Cheminformatic, 3D-QSAR and Virtual Screening Methodologies. *Front. Neurosci.* **2016**, *10*, 265. [[CrossRef](#)] [[PubMed](#)]
25. Wang, T.; Yuan, X.S.; Wu, M.B.; Lin, J.P.; Yang, L. The advancement of multidimensional QSAR for novel drug discovery—Where are we headed? *Expert Opin. Drug Discov.* **2017**, *12*, 769–784. [[CrossRef](#)] [[PubMed](#)]
26. ADMET Predictor. Available online: <http://www.simulations-plus.com/software/admet-property-prediction-qsar/> (accessed on 19 December 2017).
27. Chakravarti, S.K.; Saiakhov, R.D.; Klopman, G. Optimizing predictive performance of CASE Ultra expert system models using the applicability domains of individual toxicity alerts. *J. Chem. Inf. Model.* **2012**, *52*, 2609–2618. [[CrossRef](#)] [[PubMed](#)]
28. Derek Nexus. Available online: <https://www.lhasalimited.org/products/derek-nexus.htm> (accessed on 19 December 2017).
29. Meta-PC. Available online: <http://www.multicase.com/meta-pc> (accessed on 19 December 2017).
30. Marchant, C.A.; Briggs, K.A.; Long, A. In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic. *Toxicol. Mech. Methods* **2008**, *18*, 177–187. [[CrossRef](#)] [[PubMed](#)]
31. Judson, P.N.; Long, A.; Murray, E.; Patel, M. Assessing Confidence in Predictions Using Veracity and Utility—A Case Study on the Prediction of Mammalian Metabolism by Meteor Nexus. *Mol. Inform.* **2015**, *34*, 284–291. [[CrossRef](#)] [[PubMed](#)]
32. Oncologic<sup>TM</sup>. Available online: <https://www.epa.gov/tsca-screening-tools/oncologictm-computer-system-evaluate-carcinogenic-potential-chemicals> (accessed on 19 December 2017).
33. PASS. Available online: <https://www.ncss.com/software/pass/> (accessed on 19 December 2017).
34. Discovery Studio. Available online: <http://accelrys.com/products/collaborative-science/biovia-discovery-studio/qsar-admet-and-predictive-toxicology.html> (accessed on 19 December 2017).
35. Lagunin, A.; Zakharov, A.; Filimonov, D.; Poroikov, V. QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction. *Mol. Inform.* **2011**, *30*, 241–250. [[CrossRef](#)] [[PubMed](#)]

36. Schrödinger Release 2017-2: QikProp; Schrödinger, LLC: New York, NY, USA, 2017.
37. Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Mod.* **2000**, *18*, 464–477. [[CrossRef](#)]
38. Tetko, I.V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V.A.; Radchenko, E.V.; Zefirov, N.S.; Makarenko, A.S.; et al. Virtual computational chemistry laboratory—Design and description. *J. Comput. Aided Mol. Des.* **2002**, *19*, 453–463. [[CrossRef](#)] [[PubMed](#)]
39. Molinspiration. Available online: [www.molinspiration.com](http://www.molinspiration.com) (accessed on 19 December 2017).
40. Lee, S.K.; Lee, I.H.; Kim, H.J.; Chang, G.S.; Chung, J.E.; No, K.T. The PreADME Approach: Web-based program for rapid prediction of physico-chemical, drug absorption and drug-like properties. In *EuroQSAR 2002 Designing Drugs and Crop Protectants: Processes, Problems and Solutions*; Blackwell Publishing: Malden, MA, USA, 2003; pp. 418–420. ISBN 1405125160.
41. Boyer, S.; Zamora, I. New methods in predictive metabolism. *J. Comput. Aided Mol. Des.* **2002**, *16*, 403–413. [[CrossRef](#)] [[PubMed](#)]
42. Bender, A.; Mussa, H.Y.; Glen, R.C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718. [[CrossRef](#)] [[PubMed](#)]
43. Boyer, S.; Arnby, C.H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R.C. Reaction Site Mapping of Xenobiotic Biotransformations. *J. Chem. Inf. Model.* **2007**, *47*, 583–590. [[CrossRef](#)] [[PubMed](#)]
44. Mishra, N.K.; Agarwal, S.; Raghava, G.P. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol.* **2010**, *10*, 8. [[CrossRef](#)] [[PubMed](#)]
45. Braga, R.C.; Alves, V.M.; Silva, M.F.; Muratov, E.; Fourches, D.; Lião, L.M.; Tropsha, A.; Andrade, C.H. Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Mol. Inform.* **2015**, *34*, 698–701. [[CrossRef](#)] [[PubMed](#)]
46. Lee, P.H.; Cucurull-Sanchez, L.; Lu, J.; Du, Y.J. Development of in silico models for human liver microsomal stability. *J. Comput. Aided Mol. Des.* **2007**, *21*, 665–673. [[CrossRef](#)] [[PubMed](#)]
47. Sakiyama, Y.; Yuki, H.; Moriya, T.; Hattori, K.; Suzuki, M.; Shimada, K.; Honma, T. Predicting human liver microsomal stability with machine learning techniques. *J. Mol. Graph. Model.* **2008**, *26*, 907–915. [[CrossRef](#)] [[PubMed](#)]
48. Schwaighofer, A.; Schroeter, T.; Mika, S.; Hansen, K.; ter Laak, A.; Lienau, P.; Reichel, A.; Heinrich, N.; Müller, K.-R. A probabilistic approach to classifying metabolic stability. *J. Chem. Inf. Model.* **2008**, *48*, 785–796. [[CrossRef](#)] [[PubMed](#)]
49. Hu, Y.; Unwalla, R.; Denny, R.A.; Bikker, J.; Di, L.; Humblet, C. Development of QSAR models for microsomal stability: Identification of good and bad structural features for rat, human and mouse microsomal stability. *J. Comput. Aided Mol. Des.* **2010**, *24*, 23–35. [[CrossRef](#)] [[PubMed](#)]
50. Bursi, R.; de Gooyer, M.E.; Grootenhuis, A.; Jacobs, P.L.; van der Louw, J.; Leysen, D. (Q) SAR Study on the Metabolic Stability of Steroidal Androgens. *J. Mol. Graph. Model.* **2001**, *19*, 552–556. [[CrossRef](#)]
51. Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V.; Tropsha, A. Development and validation of k-nearest-neighbor QSPR models of metabolic stability of drug candidates. *J. Med. Chem.* **2003**, *46*, 3013–3020. [[CrossRef](#)] [[PubMed](#)]
52. Jensen, B.F.; Sørensen, M.D.; Kissmeyer, A.M.; Björkling, F.; Sonne, K.; Engelsen, S.B.; Nørgaard, L. Prediction of in vitro metabolic stability of calcitriol analogs by QSAR. *J. Comput. Aided Mol. Des.* **2003**, *17*, 849–859. [[CrossRef](#)] [[PubMed](#)]
53. Gombar, V.K.; Alberts, J.J.; Cassidy, K.C.; Mattioni, B.E.; Mohutsky, M.A. In silico metabolism studies in drug discovery: Prediction of metabolic stability. *J. Comput. Aided Drug Des.* **2006**, *2*, 177–188. [[CrossRef](#)]
54. Ulenberg, S.; Belka, M.; Król, M.; Herold, F.; Hewelt-Belka, W.; Kot-Wasik, A.; Bączek, T. Prediction of Overall In Vitro Microsomal Stability of Drug Candidates Based on Molecular Modeling and Support Vector Machines. Case Study of Novel Arylpiperazines Derivatives. *PLoS ONE* **2015**, *10*, e0122772. [[CrossRef](#)] [[PubMed](#)]
55. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2011**, *40*, D1100–D1107. [[CrossRef](#)] [[PubMed](#)]
56. Yap, C.W.E.I. Software news and update PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2010**, *32*, 1466–1474. [[CrossRef](#)] [[PubMed](#)]

57. 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. Comput. Sci.* **2003**, *43*, 493–500. [[CrossRef](#)] [[PubMed](#)]
58. Ewing, T.; Baber, J.C.; Feher, M. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 2423–2431. [[CrossRef](#)] [[PubMed](#)]
59. Klekota, J.; Roth, F.P. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008**, *24*, 2518–2525. [[CrossRef](#)] [[PubMed](#)]
60. National Center for Biotechnology Information. All Resources. Downloads. FTP: Pubchem. Available online: [https://astro.temple.edu/~tua87106/list\\_fingerprints.pdf](https://astro.temple.edu/~tua87106/list_fingerprints.pdf) (accessed on 19 December 2017).
61. Laggner, C. SMARTS Patterns for Functional Group Classification. Available online: <http://semanticchemistry.googlecode.com/svn-history/r41/wiki/InteLigand.wiki2009> (accessed on 19 December 2017).
62. Cortes, C.; Vapnik, V. Support-vector network. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
63. Shevade, S.K.; Keerthi, S.S.; Murthy, K.R.K. Improvements to the SMO algorithm for SVM regression. *IEEE Trans. Neural Netw.* **2000**, *11*, 1188–1193. [[CrossRef](#)] [[PubMed](#)]
64. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
65. Hand, D.J.; Yu, K. Idiot's Bayes—Not so stupid after all? *Int. Stat. Rev.* **2001**, *69*, 385–399. [[CrossRef](#)]
66. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185. [[CrossRef](#)]
67. Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques. *Informatica* **2007**, *31*, 249–268.
68. Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **2015**, *7*, 20. [[CrossRef](#)] [[PubMed](#)]
69. RDKit: Open-Source Cheminformatics. Available online: <http://www.rdkit.org> (accessed on 19 December 2017).
70. Marvin Was Used for Drawing, Displaying and Characterizing Chemical Structures, Substructures and Reactions, Marvin 17.17.0, 2017, ChemAxon. Available online: <http://www.chemaxon.com> (accessed on 19 December 2017).
71. Hall, M. The WEKA data mining software: An update. *SIGKDD Explor.* **2009**, *11*, 10–18. [[CrossRef](#)]
72. Till, D.J.; Hand, R.J. A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems. *Mach. Learn.* **2012**, *45*, 171–186. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).