



QSAR models for predicting enzymatic hydrolysis of new chemical entities in 'soft-drug' design

I. Massarelli ^{a,†}, M. Macchia ^b, F. Minutolo ^b, G. Prota ^b, A. M. Bianucci ^{b,*}

^a Department of Chemistry and Industrial Chemistry, University of Pisa, Via Risorgimento 35, 56126 Pisa, Italy

^b Department of Pharmaceutical Sciences, University of Pisa, Via Bonanno 6, 56126 Pisa, Italy

ARTICLE INFO

Article history:

Received 24 November 2008

Revised 4 April 2009

Accepted 9 April 2009

Available online 12 April 2009

Keywords:

Soft-drug

Enzymatic hydrolysis

QSAR

Classification methods

ABSTRACT

The work described here is aimed at developing QSAR models capable of predicting in vitro human plasma lability/stability. They were built based on a dataset comprising about 200 known compounds. 3D structures of the molecules were drawn, optimized and submitted to the calculation of molecular descriptors that enabled selecting different TR/TS set pairs, subsequently exploited to develop QSAR models. Several 'machine learning' algorithms were explored in order to obtain suitable classification models, which were then validated on the relevant TS sets. Moreover the predictive ability of the best performing models was assessed on a Prediction set (PS) comprising about 40 molecules, not strictly related, from a structural point of view, to the initial dataset, but (obviously) comprised within the validity domain of the QSAR models obtained. The study allowed selecting predictive models enabling the classification of New Chemical Entities with regard to hydrolysis rate, that may be exploited for soft-drug design.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

The terms soft-drug have been applied to compounds designed to exert their desired effect locally but which are inactivated in the circulation to reduce unwanted systemic effects. The ideal soft-drug would combine stability in the target tissue with very rapid inactivation in the blood.¹ It may be worth to point out here the difference between the soft-drug and pro-drug concepts, because confusion related to these two terms is still frequent. Pro-drugs are pharmacologically inactive compounds that result from transient chemical modifications of biologically active species. After administration, the pro-drug, by virtue of its improved characteristics, is more systemically and/or locally available than the parent drug. Before exerting its biological effect, however, the pro-drug must undergo chemical or biochemical conversion to the active form.² Soft-drug and pro-drug design share the importance of studying the lability/stability profile of the candidate compounds in the circulation. For this reason, one of the most used assay in both types of approaches is the measurement of the half-lives of potential soft-drugs or pro-drugs mainly in human blood^{3–7} or plasma.^{1,3–5,8–48} The great availability of in vitro hydrolysis data re-

ported in literature (in terms of half-lives, $t_{1/2}$, measured in human plasma) for a large variety of chemical compounds, suggested us to try to quantitatively describe such a relevant property with the aim of improve soft- or pro-drug designing.

Only a few works were afforded up to now by other research groups in this particular application.³ In particular Buchwald and Bodor³ carried out multi-linear regression studies on a dataset of about 80 non-congener carboxylic esters (for which human blood in vitro hydrolysis data were taken from literature) using hindrance related descriptors and a calculated log octanol–water partition coefficient ($Q_{log P}$). Shen and Thropsha⁴⁹ reported quantitative structure–property relationship (QSPR) models of metabolic turnover rate for more than 600 compounds in human S9 homogenate obtained from diverse chemicals proprietary to GlaxoSmithKline (GSK). The models were built with topological molecular descriptors such as molecular connectivity indices or atom pairs using the k -nearest neighbor-type methodologies. Finally, Jensen et al.⁵⁰ reported a QSAR study to estimate the in vitro metabolic stability of a data set of 130 calcitriol analogs in which the analogs were encoded with molecular structure descriptors computed with different commercial software. Partial Least Squares Regression (PLS) models were generated from the 130 analogs.

In the present paper a combination of binary (yes/no) classification methods is presented in order to obtain two groups of QSAR models for the prediction of the hydrolysis rates for New Chemical Entities (NCE).

Many molecular descriptors were computed at the E-Dragon⁵¹ server and used for model building, while the search of valid mod-

Abbreviations: PLS, partial least squares regression; QSAR, quantitative and structure–activity relationships; TR, training; TS, test; OECD, organization for economic co-operation and development.

* Corresponding author. Tel.: +39 0502219564; fax: +39 0502219605.

E-mail address: bianucci@dccl.unipi.it (A.M. Bianucci).

[†] Present address: UDR INSTM, Department of Pharmaceutical Sciences, University of Pisa, Via Bonanno 6, 56126 Pisa, Italy.

els was performed by applying several algorithms included in the Data Mining software WEKA.⁵² Rigorous validation analysis based on estimates of highly diagnostic statistical parameters and inspired to the guidelines emanated from OECD⁵³ (Organization for Economic Co-operation and Development) was applied for the seek of reliable models. The results are expected to significantly contribute at solving problems related to soft- and pro-drug design.

2. Theoretical calculations

2.1. Database collection and building

A dataset of 194 enantiomeric pure compounds (Table 1), for which hydrolysis studies in human plasma had already been performed, was collected from literature. The chemical diversity of the selected molecules is quite large as evidenced from their general structures (Chart 1). The half-live values, expressed in terms of $t_{1/2}$ (minutes) supplied the biological data (*target* property values) for QSAR analysis (Table 1). For computational needs, the target property values used in the QSAR analysis were converted into two different pairs of nominal classes (yes/no) on the basis of two different threshold values (15 min or 30 min). Table 1 also shows the class ascribed to each molecule in relation to the selected threshold.

The philosophy underlying the use of classification approaches is that there are many cases where high accuracy in predicting specific values of the *target* property does not present high priority with respect to the ability of roughly discriminating between molecules which possess or do not possess a given property (see, for example, different needs in *lead finding* and *lead optimization* tasks). In certain cases the lack of homogeneity in experimental protocols used for collecting biological data does not allow to develop well tuned QSAR models, while the same data may be successfully exploited for Classification tasks.

The structures of the molecules were built by using ISISDRAW program,⁵³ converted into 3D structures within the VEGA package⁵⁴ and were then optimized by means of the semi-empirical quantum mechanics method implemented in the program MOPAC,⁵⁵ where the AM1 Hamiltonian was used. The optimized structures were up-loaded on the E-Dragon server⁵¹ for the calculation of about 1600 molecular descriptors.

2.2. Rational TR/TS set splitting

A sphere-exclusion type algorithm,^{56–59} optimized 'in house',⁶⁰ was used to rationally split the initial dataset into training (TR) and test (TS) set pairs. Indeed one of the most challenging issues concerns the possibility that a TS set, needed for a first model validation step, properly represents the chemical space sampled by the TR set, used, in turn, for model building. Both sub-sets, obviously disjointed from each other, have to be rationally selected from any initially available dataset comprising known molecular structures and relevant biological property (*target* property) values.

The E-Dragon molecular descriptors,⁵¹ which were found to be shared by all the molecules in the dataset, were normalized and subsequently exploited for computing Euclidean distances between each pair of the molecules, in the multi-dimensional descriptor space. Descriptors were normalized according to the following formula:

$$X_{ij}^n = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}$$

where X_{ij} and X_{ij}^n are the non-normalized and normalized j -th ($j = 1, \dots, K$) descriptor values for compound i ($i = 1, \dots, N$), correspondingly, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for j th descriptor. Thus, for descriptors, $\min X_{ij}^n = 0$ and, $\max X_{ij}^n = 1$.

Table 1

The 194 molecules comprised in the starting dataset, with their IDs, bibliographic sources, $t_{1/2}$ values (min) measured in human plasma and class label (yes/no) attributed on the basis of the two selected thresholds (15 and 30 min)

Internal ID	Original denomination	References	$t_{1/2}$ (min)	Threshold 15 min	Threshold 30 min
1	7b	38	0.12	Yes	Yes
2	3	39	0.19	Yes	Yes
3	3a	14	0.24	Yes	Yes
4	6	38	0.3	Yes	Yes
5	3j	14	0.41	Yes	Yes
6	3h	14	0.43	Yes	Yes
7	VI	41	0.8	Yes	Yes
8	5h	9	1	Yes	Yes
9	7c	38	1.38	Yes	Yes
10	3g	14	1.63	Yes	Yes
11	I	39	1.9	Yes	Yes
12	7f	38	1.93	Yes	Yes
13	VIII	41	2.3	Yes	Yes
14	3k	14	2.72	Yes	Yes
15	3a	40	3	Yes	Yes
16	3l	14	3.48	Yes	Yes
17	X	41	3.7	Yes	Yes
18	3g	24	3.77	Yes	Yes
19	3b	40	4	Yes	Yes
20	3e	40	4	Yes	Yes
21	3f	14	4.17	Yes	Yes
22	2a	24	4.17	Yes	Yes
23	IX	41	4.6	Yes	Yes
24	2b	29	5	Yes	Yes
25	3h	40	5	Yes	Yes
26	1p	30	5.1	Yes	Yes
27	SDN	5	5.4	Yes	Yes
28	II	27	5.5	Yes	Yes
29	6c	11	5.7	Yes	Yes
30	7a	38	5.78	Yes	Yes
31	7c	11	6	Yes	Yes
32	3f	40	6	Yes	Yes
33	D-Glu(4NO ₂ BnO)-Ala	22	6.8	Yes	Yes
34	5a	9	7	Yes	Yes
35	II	41	7	Yes	Yes
36	6	37	7.19	Yes	Yes
37	D-Asp(4NO ₂ BnO)-Ala	22	7.2	Yes	Yes
38	IV	41	7.5	Yes	Yes
39	3d	36	8	Yes	Yes
40	5	37	8.32	Yes	Yes
41	XI	41	8.5	Yes	Yes
42	3°	36	9	Yes	Yes
43	1d	34	9.02	Yes	Yes
44	D3	8	10	Yes	Yes
45	II	12	10	Yes	Yes
46	1q	30	10.1	Yes	Yes
47	3d	32	10.2	Yes	Yes
48	III	12	11.5	Yes	Yes
49	5b	9	12	Yes	Yes
50	1i	30	12	Yes	Yes
51	3c	40	12	Yes	Yes
52	5c	11	13.8	Yes	Yes
53	3b	36	13.9	Yes	Yes
54	5d	9	15	Yes	Yes
55	6	6	15	Yes	Yes
56	3l	40	15	Yes	Yes
57	8	6	16	No	Yes
58	3j	40	16	No	Yes
59	3e	24	16.6	No	Yes
60	3c	32	16.8	No	Yes
61	5c	9	17	No	Yes
62	3	6	17	No	Yes
63	3k	40	17	No	Yes
64	4a	4	18	No	Yes
65	3a	17	20	No	Yes
66	IV	27	21	No	Yes
67	3i	14	21.8	No	Yes
68	3g	40	24	No	Yes
69	6	1	24	No	Yes
70	Isobutyryl-ester	20	24.6	No	Yes
71	3d	40	25	No	Yes
72	V	41	25	No	Yes
73	3d	24	26	No	Yes

Table 1 (continued)

Internal ID	Original denomination	References	$t_{1/2}$ (min)	Threshold 15 min	Threshold 30 min
74	3h	24	26.9	No	Yes
75	III	27	27	No	Yes
76	4	39	27.3	No	Yes
77	4b	4	28	No	Yes
78	7b	11	29.5	No	Yes
79	3b	17	30	No	Yes
80	II	15	31.2	No	No
81	Isoval_Ester	20	32.4	No	No
82	III	41	33	No	No
83	3a	26	33.56	No	No
84	5i	9	34	No	No
85	3b	32	36.6	No	No
86	6a	11	38.6	No	No
87	NAP-deg	28	38.6	No	No
88	6b	11	38.8	No	No
89	Isoprop-carbonate	20	39	No	No
90	1h	30	40	No	No
91	7d	38	41.4	No	No
92	IV	12	42.5	No	No
93	4c	4	45	No	No
94	1n	30	46	No	No
95	3e	26	46.52	No	No
96	5g	9	47	No	No
97	3b	26	49.51	No	No
98	D11	25	50	No	No
99	VI	27	50	No	No
100	5b	11	51	No	No
101	3c	26	52.59	No	No
102	5a	11	53.9	No	No
103	9	6	54	No	No
104	7	37	55.75	No	No
105	3b	29	57	No	No
106	VII	41	57	No	No
107	3c	17	60	No	No
108	3d	17	60	No	No
109	L-Glu[1-(2-hydroxyethyl)thymine]-Sar	19	61.3	No	No
110	V	12	61.5	No	No
111	III	15	63	No	No
112	3c	36	63	No	No
113	7°	11	65	No	No
114	VII	27	70	No	No
115	I	12	71.5	No	No
116	1s	30	72	No	No
117	CA-DADLE	33	72	No	No
118	L-Glu[acyclovir]-Sar	19	72.7	No	No
119	Glu(Obzl)-Sar	13	73.2	No	No
120	V	15	73.2	No	No
121	1k	30	78	No	No
122	2	36	83	No	No
123	D2	8	88	No	No
124	3a	29	91	No	No
125	7	6	94	No	No
126	Butyl ester	42	96	No	No
127	D-Glu(CH ₂ BnO)-Ala	22	97	No	No
128	Ethyl ester	42	100	No	No
129	1	36	101.9	No	No
130	7g	38	102	No	No
131	4-Hydroxybutyl ester	42	105	No	No
132	VI	12	112	No	No
133	D-Glu(4FBnO)-Ala	22	120	No	No
134	1°	30	120	No	No
135	AOA-DADLE	33	124	No	No
136	7	39	124	No	No
137	3f	24	126	No	No
138	T1	25	130	No	No
139	2	31	132	No	No
140	3f	26	138	No	No
141	T2	25	150	No	No
142	3°	32	157.2	No	No
143	D-Glu(Obzl)-Ala	13	157.8	No	No
144	5	6	160	No	No
145	1l	30	162	No	No
146	D10	25	170	No	No
147	T5	25	170	No	No

Table 1 (continued)

Internal ID	Original denomination	References	$t_{1/2}$ (min)	Threshold 15 min	Threshold 30 min
148	Asp(Obzl)-Sar	13	180	No	No
149	D-Asp(Obzl)-Ala	48	211	No	No
150	1c	34	231	No	No
151	Isopropyl ester	42	238	No	No
152	3e	17	240	No	No
153	7e	38	244	No	No
154	D-Asp(4FBnO)-Ala	22	260	No	No
155	OMCA-DADLE	33	264	No	No
156	D1	8	273	No	No
157	1a	34	273	No	No
158	N-Pivaloyloxymethyl lidocaine	23	280	No	No
159	3b	24	301	No	No
160	8	11	312	No	No
161	NAP-Me	28	316	No	No
162	7	31	318	No	No
163	3i	40	330	No	No
164	III	35	334	No	No
165	Propantheline	6	370	No	No
166	1t	30	384	No	No
167	Valaciclovir	19	428	No	No
168	3a	24	434	No	No
169	3c	24	462	No	No
170	1b	34	484	No	No
171	3d	26	511.2	No	No
172	D9	25	530	No	No
173	PIV-Ester	20	562.2	No	No
174	T3	25	630	No	No
175	IV	35	730	No	No
176	3e	32	743.4	No	No
177	D2	25	770	No	No
178	D1	25	800	No	No
179	I	35	810	No	No
180	Mefenamic acid-guaiaicol ester	18	913.8	No	No
181	7j	38	954	No	No
182	8	31	960	No	No
183	V	35	1127	No	No
184	1m	30	1320	No	No
185	I	27	1560	No	No
186	D3	25	2000	No	No
187	9	31	2100	No	No
188	3	31	2280	No	No
189	Etbut-ester	20	2682	No	No
190	7i	38	2950	No	No
191	12	10	4440	No	No
192	D5	25	9100	No	No
193	DADLE	33	>24 h	No	No
194	tert-Butyl-carbonate	20	>24 h	No	No

The selection algorithm was applied to the dataset by using different similarity thresholds, in order to select different TR/TS set pairs which were subsequently exploited for building a number of QSAR models through the use of a variety of regression algorithms available within the Data Mining program, WEKA.⁵²

Different TR sets, thanks to the molecular diversity presented by compounds comprised in them, sample a particular chemical space. That leads to defining different Applicability Domains (AD) where each model is expected to possess high predictive power. Once a QSAR model passes the validation step on its own TS set, it may be exploited for predicting the *target* property for NCEs, provided that such new molecules are shown to be comprised in the relevant AD.

2.3. Statistical analysis for model validation

The evaluation of the goodness-of-fit for classification-based QSARs can be assessed in terms of their Cooper statistic.⁶¹ In

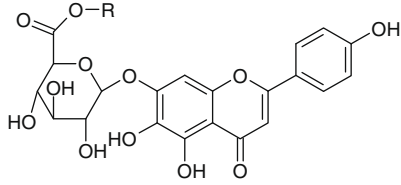
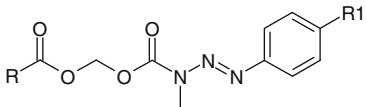
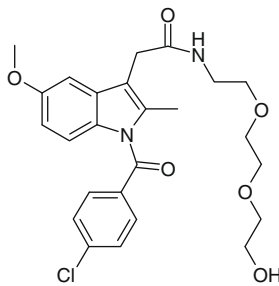
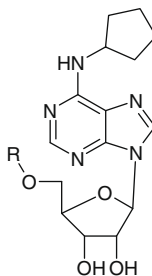
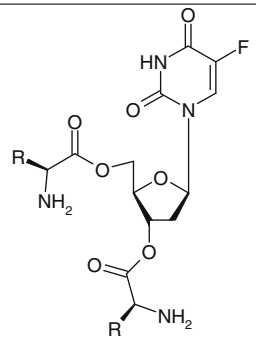
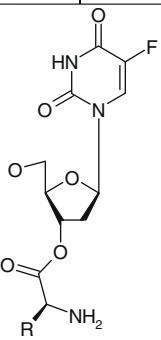
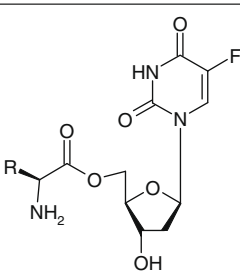
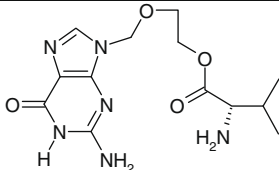
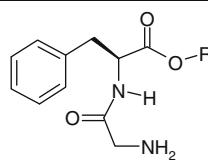
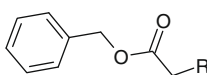
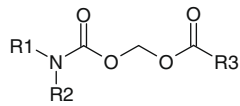
		
Ref. 8 (3 compounds)	Ref. 9 (7 compounds)	
		
Ref. 10	Ref. 4 (3 compounds)	
		
Ref. 11 (3 compounds)	Ref. 11 (3 compounds)	Ref. 11 (3 compounds)
		
Ref. 11	Ref. 12 (6 compounds)	
		
Ref. 13 (3 compounds)	Ref. 14 (8 compounds)	

Chart 1. Molecular scaffolds of compounds reported in Table 1 (R_i groups are not explicitly shown; see original references for details). Scaffolds highlight molecular diversity of the initial dataset.

practice, the results of the classification can be arranged in the so-called confusion or contingency matrix where the rows represent the reference classes, while the columns represent the predicted classes assigned through the classification model. The main diagonal represents the cases where the true class coin-

cides with the assigned class, that is, the number of objects correctly classified in each class, while the non-diagonal cells represent the misclassifications. Over-predictions fall in the upper right half matrix whereas under-predictions fall in the lower left half matrix.

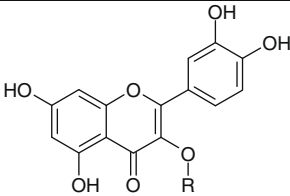
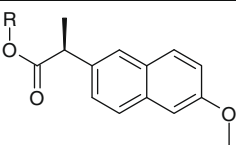
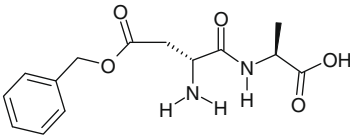
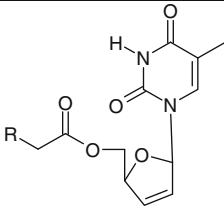
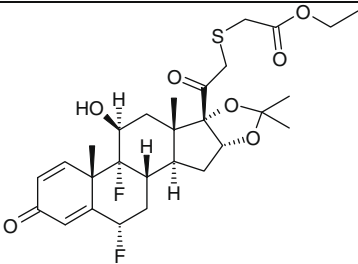
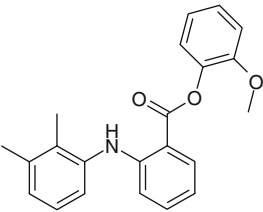
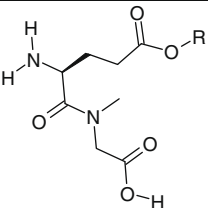
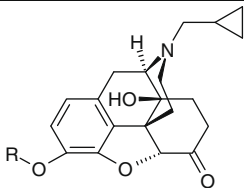
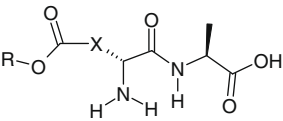
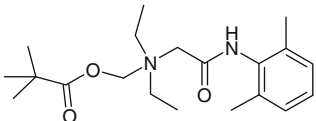
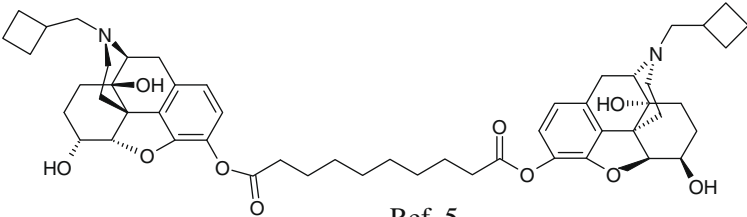
Ref. 5	
 <p>Ref. 15 (3 compounds)</p>	 <p>Ref. 28 (2 compounds)</p>
 <p>Ref. 48</p>	 <p>Ref. 17 (5 compounds)</p>
 <p>Ref. 1</p>	 <p>Ref. 18</p>
 <p>Ref. 19 (3 compounds)</p>	 <p>Ref. 20 (6 compounds)</p>
 <p>Ref. 22 (5 compounds)</p>	 <p>Ref. 23</p>
 <p>Ref. 5</p>	

Chart. 1 (continued)

When evaluating the results of a classification model, the reference status is generally considered the one where all of the objects are assigned to the class that is most represented. This reference

condition corresponds to the absence of a model, and is therefore called *No-model* condition. *Goodness-of-fit* values close to the ones of the *No-model* status give evidence of poor results from the clas-

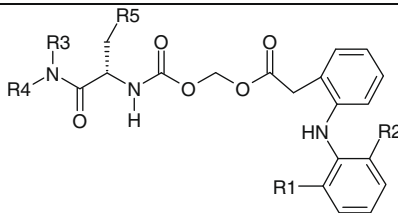
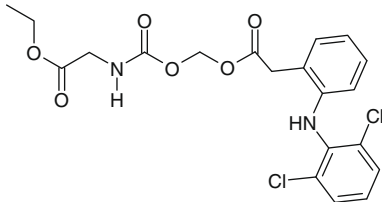
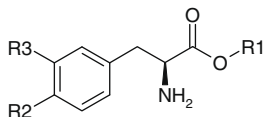
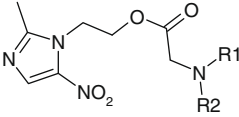
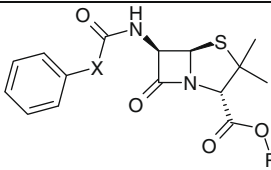
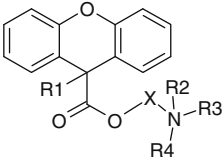
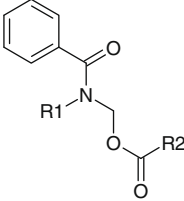
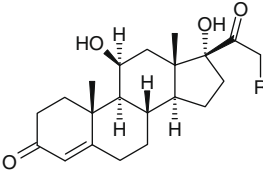
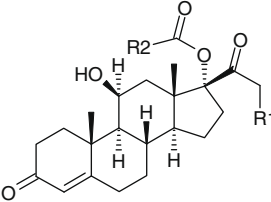
 <p>Ref. 24 (8 compounds)</p>	
 <p>Ref. 24</p>	 <p>Ref. 25 (11 compounds)</p>
 <p>Ref. 26 (6 compounds)</p>	 <p>Ref. 27 (6 compounds)</p>
 <p>Ref. 6 (7 compounds)</p>	 <p>Ref. 30 (11 compounds)</p>
 <p>Ref. 31 (3 compounds)</p>	 <p>Ref. 31 (2 compounds)</p>

Chart. 1 (continued)

sification model, as the *No-model* value is unique and independent from the classification method adopted.

Beyond the comparison with the *No-model*, the contingency matrix of a classification model offers other statistical parameters useful for estimating how well the model performs. Such parameters enable estimating the capability of the model to detect known active compounds (sensitivity), non-active compounds (specificity), and all chemicals in general (concordance or accuracy). Finally, other statistics were exploited, such as the kappa (κ) statistic.⁶² For an estimate of the *k* statistic values of a QSAR model the following criterion is usually adopted: 0.81–1.00: almost perfect,

0.61–0.80: substantially good, 0.41–0.60: moderately good, <0.41 poor model.

2.4. Machine learning

The selection of the best-performing algorithm together with an optimal set of molecular descriptors (among the about 1600 ones initially computed) was performed by using the WEKA package, version 3.5.6. WEKA is a JAVA software from the University of Waikato, New Zealand⁵² within the open source frame, issued under the GNU General Public License. The package provides a collection of

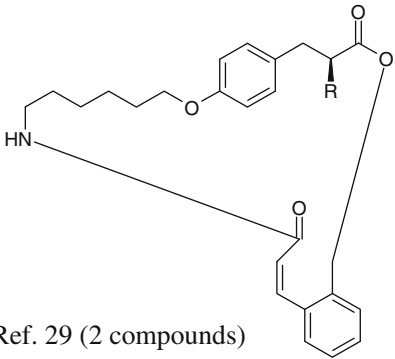
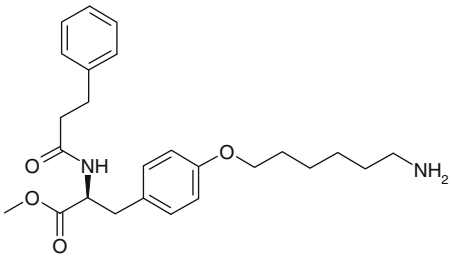
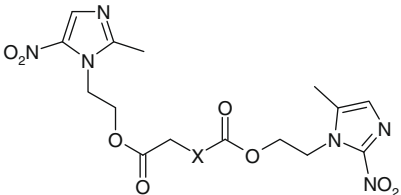
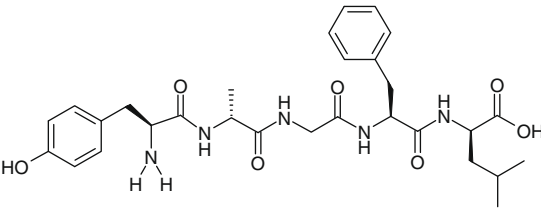
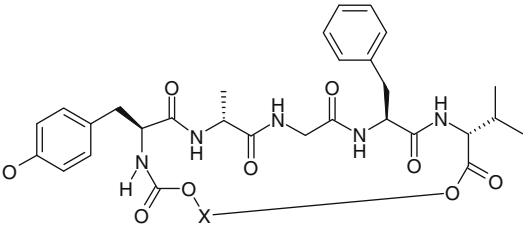
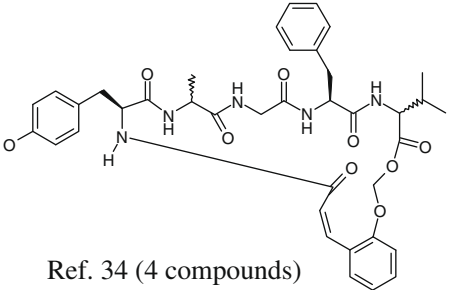
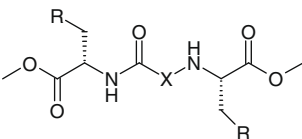
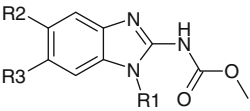
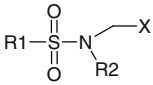
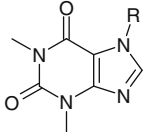
 <p>Ref. 29 (2 compounds)</p>	 <p>Ref. 29</p>
 <p>Ref. 32 (5 compounds)</p>	 <p>Ref. 33</p>
 <p>Ref. 33 (3 compounds)</p>	 <p>Ref. 34 (4 compounds)</p>
 <p>Ref. 36 (6 compounds)</p>	 <p>Ref. 37 (3 compounds)</p>
 <p>Ref. 38 (10 compounds)</p>	 <p>Ref. 39 (2 compounds)</p>

Chart. 1 (continued)

machine learning algorithms for data mining tasks. It contains tools for data pre-processing, classification, regression, clustering, association rules and visualization, and is well suited for developing new machine learning schemes.

In particular, with regard to algorithms utilized in building the models described in *Results*, Decision Tree-type algorithms, such as Random Tree, Random Forest and Best First decision-tree (BFTree) were the ones most exploited. Decision tree learning is a method commonly used in Data Mining. According to a technical definition, a Decision tree represents a disjunction of conjunctions of con-

straints on the attribute-values of instances, where the term 'instance' refers to the property of interest of each molecule and 'attribute' refer to the vector of molecular descriptors exploited for representing each molecular structure. See Ref. 52 for details.

Two additional Machine Learning methods gave interesting results during model development: Locally Weighted Learning (LWL) and Random committee. LWL uses an instance-based algorithm to assign instance weights which are then used by a specified Weighted-Instances-Handler. For more details see Ref. 63. Random committee is a method for building an ensemble of randomizable

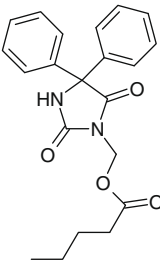
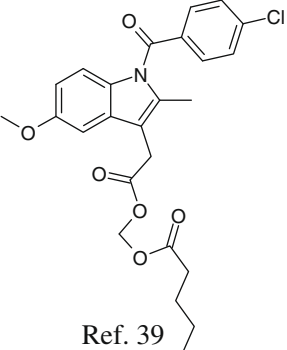
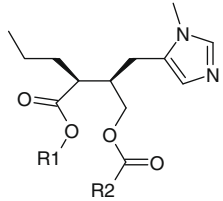
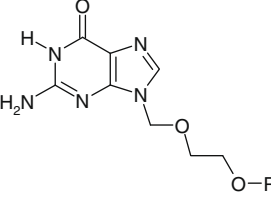
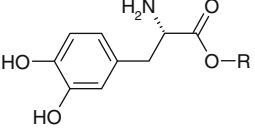
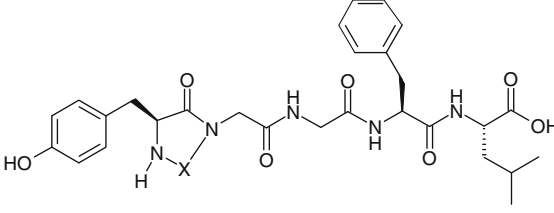
 <p>Ref. 39</p>	 <p>Ref. 39</p>
 <p>Ref. 40 (12 compounds)</p>	 <p>Ref. 41 (10 compounds)</p>
 <p>Ref. 42 (4 compounds)</p>	 <p>Ref. 35 (4 compounds)</p>

Chart. 1 (continued)

base classifiers. Each base classifiers is built by using a different random number seed (but based on the same dataset). The final prediction is a straight average of the predictions generated by the individual base classifiers.⁵²

3. Results

3.1. Calculation and selection of molecular descriptors

The molecules of the dataset were drawn by the ISISDRAW program,⁵³ converted in 3D structures within the VEGA package⁵⁴ and subjected to quantum-chemical and thermodynamic calculations. After that, E-Dragon molecular descriptors were calculated for all the molecules and about 1600 of them were found to be shared by all the molecules belonging to the whole dataset. In order to determine the proper combination of the molecular descriptors (usually referred as *attributes*), to be used in the search of good QSAR models, the CfsSubsetEval *attribute evaluator*⁶⁴ of WEKA was employed. It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class of compounds while having low inter-correlation are preferred.

Using a 10-fold cross-validation on the entire dataset, and performing 10 runs with different seed numbers to obtain averaged results, subsets of molecular descriptors were selected according to the averaged percentages of a ranking value. In experiments made in the *A case* ($t_{1/2}$ threshold: 15 min) 22 descriptors were se-

lected (Table 2), while for the *B case* ($t_{1/2}$ threshold: 30 min) 25 descriptors were selected (Table 3).

Interpretability of QSAR models plays a relevant role in the step of the drug design and depends on modeling technique and molecular descriptors involved. Often linear regression models are amenable to interpretation, but not very accurate. On the other hand some Machine Learning methods, such as Neural Networks behave as 'black boxes' but give more accurate results. Some other techniques (like Random Forest or others) are considered to lie in between.

In the case of models described here, the molecular properties represented by most molecular descriptors, identified as more significant during the selection, are not easy to be recognized, especially for the descriptors selected in the *A case*. Only in the list of descriptors selected in the *B case*, the molecular descriptors referred as nArCOOR [number of (aromatic) esters] and nArOH (number of aromatic hydroxyls) show a clear chemical meaning, being referred to well defined functional groups. In particular, the number of aromatic esters was identified as one of the most relevant descriptors in explaining hydrolysis rate as it can be reasonably expected. Obviously, the higher is the number of (aromatic) esters, the more easily the molecule undergoes hydrolysis, thus leading to a faster hydrolysis rate.

The meaning of the remaining descriptors involved in the models, and listed in Tables 2 and 3, is briefly outlined below. They belong to 10 different groups: 2D autocorrelation descriptors, 3D Molecule Representation of Structure based on Electron diffraction (MoRSE) descriptors, constitutional descriptors, functional group

Table 2Molecular descriptors selected by the CfsSubsetEval attribute evaluator⁶⁴ of WEKA for the *A* case ($t_{1/2}$ threshold: 15 min)

ID	Molecular descriptor name	Descriptor type
AMW	Average molecular weight	Constitutional descriptors
Dp	D total accessibility index/weighted by atomic polarizabilities	WHIM descriptors
E2p	2nd component accessibility directional WHIM index/weighted by atomic polarizabilities	WHIM descriptors
GATS1e	Geary autocorrelation—lag 1/weighted by atomic Sanderson electronegativities	2D autocorrelations
GATS2v	Geary autocorrelation—lag 2/weighted by atomic van der Waals volumes	2D autocorrelations
GATS3m	Geary autocorrelation—lag 3/weighted by atomic masses	2D autocorrelations
GATS4p	Geary autocorrelation—lag 4/weighted by atomic polarizabilities	2D autocorrelations
HATSm	Leverage-weighted total index/weighted by atomic masses	GETAWAY descriptors
JGI1	Mean topological charge index of order1	Topological charge indices
Mor08v	3D-MorSE—signal 08/weighted by atomic van der Waals volumes	3D-MorSE descriptors
Mor16p	3D-MorSE—signal 16/weighted by atomic polarizabilities	3D-MorSE descriptors
Mor16v	3D-MorSE—signal 16/weighted by atomic van der Waals volumes	3D-MorSE descriptors
MSD	Mean square distance index (Balaban)	Topological descriptors
Mv	Mean atomic van der Waals volume (scaled on carbon atom)	Constitutional descriptors
PW2	Path/walk 2—Randic shape index	Topological descriptors
R4u	R autocorrelation of lag 4/unweighted	GETAWAY descriptors
R5u	R autocorrelation of lag 5/unweighted	GETAWAY descriptors
R5v+	R maximal autocorrelation of lag 5/weighted by atomic van der Waals volumes	GETAWAY descriptors
RDF030m	Radial distribution function—3.0/weighted by atomic masses	RDF descriptors
RDF035m	Radial distribution function—3.5/weighted by atomic masses	RDF descriptors
RDF130m	Radial distribution function—13.0/weighted by atomic masses	RDF descriptors
T(O...F)	Sum of topological distances between O...F	Topological descriptors

counts, geometrical descriptors, GETAWAY descriptors, Radial Distribution Function (RDF) descriptors, topological descriptors, topological charge indices and Weighted Holistic Invariant Molecular (WHIM) descriptors.

Some of them do not depend on molecular conformation while others are strictly related to the 3D arrangement of molecules. It implies that reliable values for the latter type of descriptors are obtained only if plausible conformations may be guessed with sufficient accuracy.

The '2D autocorrelations' type descriptors are computed from molecular graph as the sum of products of atom weights of the terminal atoms of all the paths for the considered path length (the so-called lag).

The 3D-MorSE descriptors provide 3D information from atomic coordinates by using the same transform as in electron diffraction (which uses them to prepare theoretical scattering curves).

Constitutional descriptors are simple one-D descriptors, independent from molecular connectivity and conformations. Example of these descriptors are atom and bond counts, molecular weight sum of atomic properties, etc.

Functional group counts are molecular descriptors based on the counting of chemical functional groups. The two descriptors previously mentioned [nArCOOR = number of (aromatic) esters and nArOH = number of aromatic hydroxyls] belong to such a category: in the most simple cases they may be directly related to specific chemical properties.

Geometrical descriptors consist of different kinds of conformationally dependent descriptors based on the molecular geometry.

According to the definition, GETAWAY (Geometry, Topology, and Atom-Weights Assembly) descriptors encode both geometrical information given by the Molecular Influence Matrix (MIM) (which in turn takes into account the relative position of atoms in a molec-

Table 3Molecular descriptors selected by the CfsSubsetEval attribute evaluator⁶⁴ of WEKA for the *B* case ($t_{1/2}$ threshold: 30 min)

ID	Molecular descriptor name	Descriptor type
Dp	D total accessibility index/weighted by atomic polarizabilities	WHIM descriptors
Dv	D total accessibility index/weighted by atomic van der Waals volumes	WHIM descriptors
GATS1e	Geary autocorrelation—lag 1/weighted by atomic Sanderson electronegativities	2D autocorrelations
GATS2v	Geary autocorrelation—lag 2/weighted by atomic van der Waals volumes	2D autocorrelations
GATS3m	Geary autocorrelation—lag 3/weighted by atomic masses	2D autocorrelations
GATS4p	Geary autocorrelation—lag 4/weighted by atomic polarizabilities	2D autocorrelations
HATS4u	Leverage-weighted autocorrelation of lag 4/unweighted	GETAWAY descriptors
HATS5e	Leverage-weighted autocorrelation of lag 5/weighted by atomic Sanderson electronegativities	GETAWAY descriptors
HATS5u	Leverage-weighted autocorrelation of lag 5/unweighted	GETAWAY descriptors
HATSm	Leverage-weighted total index/weighted by atomic masses	GETAWAY descriptors
JGI7	Mean topological charge index of order7	Topological charge indices
JGT	Global topological charge index	Topological charge indices
MATS1e	Moran autocorrelation—lag 1/weighted by atomic Sanderson electronegativities	2D autocorrelations
MATS3p	Moran autocorrelation—lag 3/weighted by atomic polarizabilities	2D autocorrelations
MATS8p	Moran autocorrelation—lag 8/weighted by atomic polarizabilities	2D autocorrelations
Mor08v	3D-MorSE—signal 08/weighted by atomic van der Waals volumes	3D-MorSE descriptors
Mor22v	3D-MorSE—signal 22/weighted by atomic van der Waals volumes	3D-MorSE descriptors
nArCOOR	Number of esters (aromatic)	Functional group counts
nArOH	Number of aromatic hydroxyls	Functional group counts
R5v+	R maximal autocorrelation of lag 5/weighted by atomic van der Waals volumes	GETAWAY descriptors
RDF030m	Radial Distribution Function—3.0/weighted by atomic masses	RDF descriptors
RDF035m	Radial Distribution Function—3.5/weighted by atomic masses	RDF descriptors
RDF095v	Radial Distribution Function—9.5/weighted by atomic van der Waals volumes	RDF descriptors
SPAN	span R	Geometrical descriptors
T(N...Cl)	Sum of topological distances between N...Cl	Topological descriptors

Table 4The selected QSAR models of the *A case*

Threshold	TR/TS	Algorithm	Statistics	Validation task			Prediction task
				TR	TR_CV_LOO	TS	PS
0.3	TR 164	Random tree (RT1)	Correctly classified instances	100%	68.90%	90%	75.68%
	TS 30		Kappa statistic	1	0.29	0.71	0.39
0.3	TR 164	Random tree (RT2)	Correctly classified instances	100%	78.05%	93.33%	78.38%
	TS 30		Kappa statistic	1	0.47	0.79	0.49
0.3	TR 164	LWL	Correctly classified instances	84.76%	73.17%	83.33%	67.57%
	TS 30		Kappa statistic	0.61	0.32	0.52	0.14
0.3	TR 164	Random forest (RF)	Correctly classified instances	99.39%	76.83%	93.33%	75.68%
	TS 30		Kappa statistic	0.99	0.44	0.76	0.39
0.3	TR 164	No model	Correctly classified instances	68.90%	68.90%	83.33%	64.86%
	TS 30		Kappa statistic	0	0	0	0

ular structure optimized in some way) and the topological information given by the molecular graph, weighted by chemical information encoded in selected atomic weights.

Radial Distribution Function descriptors are obtained by radial basis functions centered on different interatomic distances (from 0.5Å to 15.5Å).

Topological descriptors are obtained from molecular graph (usually H-depleted), that is, 2D-descriptors conformationally independent.

Topological charge indices describe charge transfer between pairs of atoms and therefore global charge transfer in a molecule.

Finally, WHIM descriptors are based on statistical indices calculated on the projections of atoms onto the three principal components obtained from weighted covariance matrices of the atomic coordinates. The aim is to capture 3D information regarding size, shape, symmetry and atom distributions with respect to invariant reference frames.

For more details about E-Dragon descriptors we remand to previously published studies.^{65–70}

3.2. Rational splitting of the dataset into training and test set (TR/TS)

The descriptors selected for each one of the *A* and *B* cases were exported and used as the input for the sphere-exclusion algorithm previously mentioned, that at first parsed and normalized them. Then, the similarities among each of the molecules of the dataset were calculated in terms of Euclidean distances in the multi-dimensional descriptor space. Among several similarity thresholds chosen during the TR/TS set splitting, the one corresponding to 0.3 value (in the *A case*), and the ones corresponding to 0.3 and 0.4 val-

ues (in the *B case*) turned out to lead to satisfactory QSAR models. When using such thresholds, the *A case* provided a pair of TR/TS sets respectively comprising 164 and 30 molecules. In the *B case* the use of the 0.3 threshold led to a TR set comprising 165 molecules and a TS set comprising 29 molecules, while the use of the 0.4 threshold gave a pair of TR/TS sets comprising 137 and 57 molecules, respectively.

3.3. Model building and validation

The use of several classification methods was tried, by exploiting each one of the TR sets previously identified, in order to obtain QSAR models endowed with good predictive ability. All the methods employed, ranging from algorithms based on decision trees, to support vector machines, neural networks, and so on, are available within the WEKA package.⁵² Among the several algorithms attempted only the ones listed in [Tables 4 and 5](#) enabled the construction of QSAR models which turned out to pass the validation step. In the first group of models (*A case*), Random tree (models RT1 and RT2 of [Table 4](#)), Locally Weighted Learning (LWL) and Random Forest (RF) turned out to be the algorithms that gave better results. The size of trees in the RT1 and RT2 models was 59 and 133, respectively. RF model consisted of 8 trees, each constructed considering five random features. The graphs of the obtained trees, where possible, are reported in [Supplementary data](#).

In the second group of models (*B case*) Random tree (RT in [Table 5](#)), Random committee (RC), Best First decision-tree (BFT) and Random Forest (RF) were the algorithms that gave better results. The size of the trees in RT, RC and BFT models was 99 and 93, 23, respectively. RF model consisted in nine trees, each constructed

Table 5The selected QSAR models of the *B case*

Threshold	TR/TS	Algorithm	Statistics	Validation task			Prediction task
				TR	TR_CV_LOO	TS	PS
0.4	TR 137	Random tree (RT)	Correctly classified instances	100%	56.20%	61.4%	67.57%
	TS 57		Kappa statistic	1	0.29	0.71	0.39
0.4	TR 137	No model	Correctly classified instances	55.47%	0.55%	68.42%	56.76%
	TS 57		Kappa statistic	0	0	0	0
0.3	TR 165	Random committee (RC)	Correctly classified instances	100%	65.45%	86.21%	67.57%
	TS 29		Kappa statistic	1	0.31	0.68	0.30
0.3	TR 165	BFT tree (BFTT)	Correctly classified instances	87.27%	62.42%	96.55%	64.86%
	TS 29		Kappa statistic	0.74	0.23	0.91	0.22
0.3	TR 165	Random forest (RF)	Correctly classified instances	99.39%	69.70%	89.66%	62.16%
	TS 29		Kappa statistic	0.99	0.38	0.75	0.15
0.3	TR 165	No model	Correctly classified instances	56.36%	56.36%	75.86%	56.76%
	TS 29		Kappa statistic	0	0	0	0

considering five random features. The graphs of the obtained trees, where possible, are also reported in [Supplementary data](#).

In this work we report two groups of QSAR models (A and B cases) showing good performances. Model performances were evaluated in terms of Cooper and *k* statistics (see Section 2.3) computed on: (i) TR set; (ii) TR set through Leave-One-Out (LOO) cross-validation; (iii) TS set. The *Correctly classified Instances* (referred as *accuracy*), and the *k* statistic values are reported in [Tables 4 and 5](#) for each model. Moreover, in [Tables 4 and 5](#) the statistics referred to the *No-model* condition (see Section 2.3) are reported for comparison.

Among hundreds of attempts, four models capable to discern NCEs with $t_{1/2} > \text{or} \leq$ than 15 min (A case, reported in [Table 4](#)) and 4 models capable to discern NCEs with $t_{1/2} > \text{or} \leq$ than 30 min (B case, reported in [Table 5](#)) were selected. By combining results coming from the two groups of models, information needed to correctly classify the hydrolysis rate of NCEs as fast (≤ 15 min), medium ($15 \text{ min} < t_{1/2} \leq 30 \text{ min}$) or slow ($> 30 \text{ min}$) is supplied. Classification of NCE on the basis of the above models is expected to be very useful in designing new molecules of therapeutic interest. Application of particular interest may be made in the field of soft-drug design, since the administered drug is required to exert its pharmacological effects on a certain local target and to undergo to a rapid hydrolysis so that systemic effects could be avoided or, at least, limited.

It can be observed, from [Tables 4 and 5](#), that different models were obtained by applying the same algorithm. Different models are generated by the same algorithm since different training parameters may be used during the model construction. For example, the difference between the RT1, RT2 ([Table 4](#)) and RT ([Table 5](#)) models are related to a parameter indicating the number of attributes randomly chosen when building them, which was set to 4, 10 and 1, respectively.

Furthermore the difference between RF models obtained in [Tables 4 and 5](#) consist of different values of the number of trees to be generated during the calculation (8 and 9, respectively) and in different values of the maximum depth of the trees (10 and 14, respectively).

3.4. Applicability domain

The predictive power of a QSAR model, which had already passed the validation step, depends upon the Applicability Domain (AD) where predictions are being carried out. Only for molecules belonging to such a domain, once submitted to the prediction task, the model is expected to return reliable values. Several methods to define the Applicability Domain of a QSAR model are reported in the literature.⁷¹ One of the simplest criteria was considered here, in which the chemical space of model validity is approximately defined by the range (max and min) of the values taken by each one

Table 6

Molecules initially selected for being comprised in PS, with their IDs, bibliographic sources, $t_{1/2}$ values (min) measured in human plasma and class label (yes/no), attributed on the basis of the two selected thresholds (15 and 30 min)

Internal id	Original denomination	References	$t_{1/2}$ (min)	Threshold 15 min	Threshold 30 min
195	Ic_R	46	1	Yes	Yes
196	Ila_R	46	1.1	Yes	Yes
197	Ic_S	46	1.4	Yes	Yes
198	Ib_R	46	1.5	Yes	Yes
199	Ilb_R	46	1.5	Yes	Yes
200	Ilb_S	46	2.6	Yes	Yes
201	13a	47	4	Yes	Yes
202	13b	47	4	Yes	Yes
203	23b	47	5	Yes	Yes
204	If_R	46	5.2	Yes	Yes
205	Ib_S	46	7	Yes	Yes
206	VII	41	9	Yes	Yes
207	IV	41	10	Yes	Yes
208	3	44	10.3	Yes	Yes
209	Ila_S	46	11	Yes	Yes
210	V	41	17	No	Yes
211	Ilc_R	46	18	No	Yes
212	Ie_R	46	19	No	Yes
213	III	41	36	No	No
214	2	45	37.2	No	No
215	Ild_R	46	38	No	No
216	1	44	52.8	No	No
217	3	45	56.4	No	No
218	4	44	81	No	No
219	2	44	97.8	No	No
220	If_S	46	135	No	No
221	VI	41	240	No	No
222	Ild_S	46	355	No	No
223	Ilc_S	46	445	No	No
224	10	45	498	No	No
225	Ie_S	46	960	No	No
226	5	45	1062	No	No
227	11	45	3390	No	No
228	6	45	3744	No	No
229	Ia_R	46	5580	No	No
230	Ia_S	46	15,900	No	No
231	4	45	>18,000	No	No
232	7	45	>18,000	No	No
233	8	45	>18,000	No	No

The two molecules, that were found to fall out of the AD of the selected QSAR models, are in bold.

of the descriptors involved in the equations which define the model itself.

3.5. Prediction set

In view of successfully applying to widely diverse molecular libraries any QSAR model, which had already passed preliminary validation steps, some Authors suggest to exploit, as a supplemental *test set*, a further set of known data, referring to molecules taken from a different database, with respect to the initial one (i.e., the one exploited for model development). It is usually referred as *Prediction set* (PS). Obviously, such PS must contain compounds that fall into the AD defined by the model to be further validated. In this perspective an additional pool of 39 molecules (Table 6), the structure of which are not strictly related to the ones belonging to the initial dataset, was collected from the literature. Data referring to half-live values in human plasma were available for these molecules. The relevant molecular scaffolds are reported in Chart 2.

The above molecules were checked out in order to verify if they belonged to the model ADs. According to the adopted criterion, the values of descriptors involved in each model were computed for all the available 39 molecules, in order to check if they fell in the ranges of values (max and min) covered by the corresponding descriptors for molecules comprised in each TR set. Only two molecules (internal ID: 201 and 202) showed at least one descriptor value outside the ranges defining the ADs; they were leaved out from the PS. This can be considered a proof that the large and chemically diverse dataset, exploited for model building, is capable of covering a wide chemical space, which makes the ADs of the obtained models wide as well.

The remaining 37 molecules (PS) were then submitted to hydrolysis rate prediction by using all the selected models for the *A* and *B* cases. Results of the predictions task for each group of four models are reported in Tables 4 and 5; they show that

the proposed models give slightly better (*A* case) or similar (*B* case) performances in comparison to the TS set. It supports the expectation that the models proposed here may be successfully exploited for practical applications in the field of soft-drugs design.

4. Conclusion

The development of QSAR models supplies a very helpful tool in drug discovery and in the field of ADME/Tox predictions. Understanding the relationships between the structural features of a series of compounds and their biological activity/toxicity/metabolism allows optimizing features responsible for the wanted/unwanted properties. This presents significant relevance, since QSAR models allow to quantitatively predict the biological property or the toxicity profile of newly designed compounds before their synthesis. It allows removing undesirable molecules at early stages of their development, thus preventing waste of resources.

In this paper, two groups of QSAR models were developed on a dataset of about 200 chemically diverse molecules reported in the literature with their *in vitro* hydrolysis rates in human plasma. Molecular descriptors from E-Dragon server⁵¹ were computed for the molecules and filtered according to their significance through a subroutine available in the data mining program, WEKA.⁵² The half-live values, expressed in terms of $t_{1/2}$ (in minutes) supplied the biological data for QSAR analysis. They were converted, for computational needs, into two nominal classes (yes/no) according to two different threshold values utilized in predictive model building. Then, particular care was put on the rational splitting of the dataset into TR and TS sets before the development of the models, as optimal sampling of the available known molecular structures and relevant biological properties is required for obtaining highly reliable models. An algorithm based on the sphere-exclusion theory was used to produce TR/TS sets optimally selected from the whole available data set of known molecules.

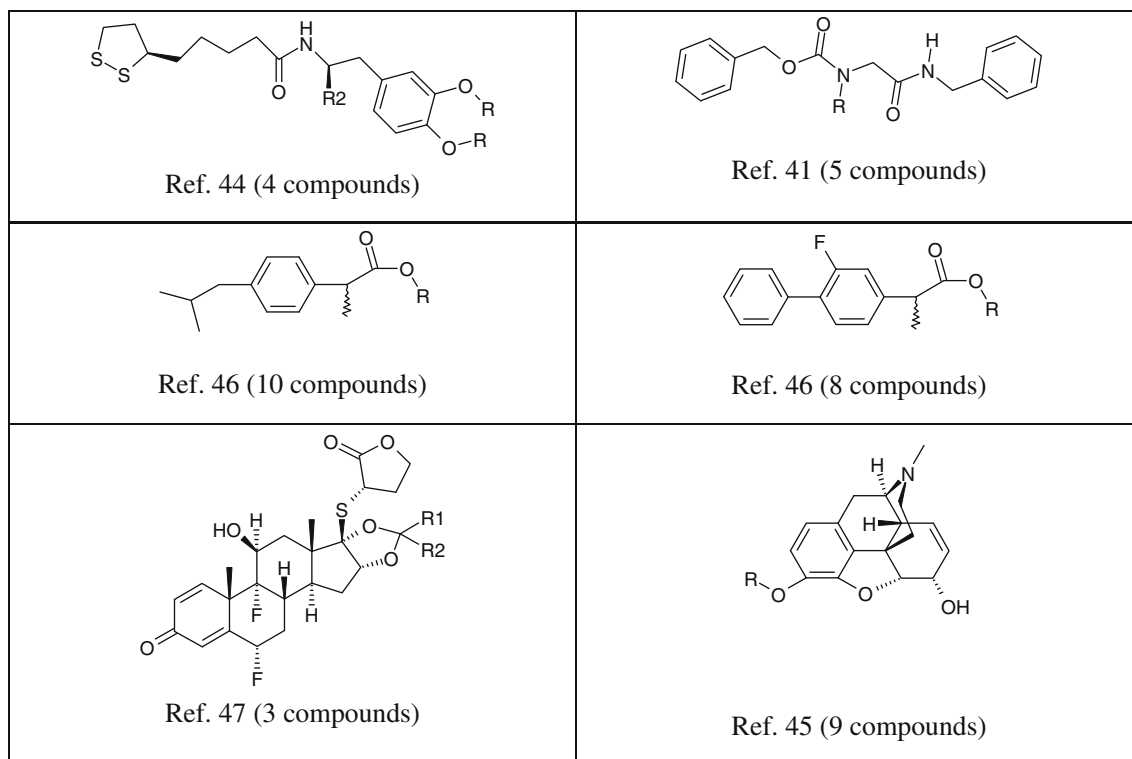


Chart 2. Molecular scaffolds of compounds reported in Table 6 (R_i groups are not explicitly shown; see original references for details). Scaffolds highlight molecular diversity of the PS.

Several hundreds of models were obtained and analyzed, starting from the different TR sets previously selected, by using a number of algorithms available in the WEKA package for classification purposes. Among them, 4 models capable to discern NCEs with $t_{1/2}$ > or \leq than 15 min (A case), and 4 models capable to discern NCEs with $t_{1/2}$ > or \leq than 30 min (B Case) were shown to possess high predictive power. The information supplied by a combination of the results obtained from the two type of models turned out to be rich enough to enable classifying the hydrolysis rate of NCEs as fast (≤ 15 min), medium ($15 \text{ min} < t_{1/2} \leq 30 \text{ min}$) or slow ($> 30 \text{ min}$).

We can conclude that the two groups of models proposed here are well suited for enabling predictions of hydrolytic rates of NCEs at the early stage of their development. That is expected to positively contribute at the design of soft-drugs, since the administered drug is required to exercise its pharmacological effects on a certain local target and then undergo to rapid hydrolysis, so that systemic effects could be avoided or, at least, limited. The unique condition to be observed to make reliable predictions on NCEs is that they must belong to the applicability domain defined by the TR set where the predictive model was trained. Only the 5.4% of the molecules belonging to the PS tested in this work, turned out do not be comprised in it. It suggests that the initially selected dataset, comprising many chemically diverse molecules, is capable of covering a wide chemical space (rich sampling). It makes the relevant ADs, where predictions have to be taken as valid, wide as well.

As a very final conclusion it has to be pointed out that, among the plethora of approaches recently developed for QSAR applications, and available at free web sites, it is always possible to identify computational methods suitable for treating, with highly satisfying results, any QSAR problem of interest for Medicinal Chemists, provided that protocols enabling the construction of QSAR models are carefully applied with particular regard to the definition of the chemical space of validity of the models under development.

Acknowledgments

The Authors wish to thank Fondazione Salvatore Maugeri, Pavia (Italy) for scientific and financial support.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.bmc.2009.04.014](https://doi.org/10.1016/j.bmc.2009.04.014).

References and notes

- Biggadike, K.; Angell, R. M.; Burgess, C. M.; Farrell, R. M.; Hancock, A. P.; Harker, A. J.; Irving, W. R.; Ioannou, C.; Procopiou, P. A.; Shaw, R. E.; Solanke, Y. E.; Singh, O. M. P.; Snowden, M. A.; Stubbs, R. J.; Walton, S.; Weston, H. E. *J. Med. Chem.* **2000**, *43*, 19.
- Bodor, N.; Buchwald, P. *Med. Res. Rev.* **2000**, *20*, 58.
- Buchwald, P.; Bodor, N. *J. Med. Chem.* **1999**, *42*, 5160.
- Dalpiaz, A.; Scatturin, A.; Menegatti, E.; Bortolotti, F.; Pavan, B.; Biondi, C.; Durini, E.; Manfredini, S. *Pharm. Res.* **2001**, *18*, 531.
- Pao, L.; Hsiong, C.; Hu, O. Y.; Wang, J.; Ho, S. *Drug Metab. Dispos.* **2005**, *33*, 395.
- Brouillette, G.; Kawamura, M.; Kumar, G. N.; Bodor, N. *J. Pharm. Sci.* **1996**, *85*, 619.
- Prueksaritanont, T.; Gorham, L. M.; Breslin, M. J.; Hutchinson, J. H.; Hartman, G. D.; Vyas, K. P.; Baillie, T. A. *Drug Metab. Dispos.* **1997**, *25*, 978.
- Cao, F.; Guo, J.; Ping, Q.; Liao, Z. *Eur. J. Pharm. Sci.* **2006**, *29*, 385.
- Carvalho, E.; Francisco, A. P.; Iley, J.; Rosa, E. *Bioorg. Med. Chem.* **2000**, *8*, 1719.
- Chandrasekaran, S.; Al-Ghananeem, A. M.; Riggs, R. M.; Crooks, P. A. *Bioorg. Med. Chem.* **2006**, *16*, 1874.
- Landowski, C. P.; Vig, B. S.; Song, X.; Amidon, G. L. *Mol. Cancer Ther.* **2005**, *4*, 659.
- Larsen, S. W.; Ankersen, M.; Larsen, C. *Eur. J. Pharm. Sci.* **2004**, *22*, 399.
- Lepist, E.; Kusk, T.; Larsen, D. H.; Andersen, D.; Frokjaer, S.; Taub, M. E.; Veski, P.; Lennernas, H.; Friedrichsen, G.; Steffansen, B. *Eur. J. Pharm. Sci.* **2000**, *11*, 43.
- Mendes, E.; Furtado, T.; Neres, J.; Iley, J.; Jarvinen, T.; Rautio, J.; Moreira, R. *Bioorg. Med. Chem.* **2002**, *10*, 809.
- Montenegro, L.; Carbone, C.; Maniscalco, C.; Lambusta, D.; Nicolosi, G.; Ventura, C. A.; Puglisi, G. *Int. J. Pharm.* **2007**, *336*, 257.
- Sriram, D.; Srichakravarthy, N.; Bal, T. R.; Yogeeswari, P. *Biomed. Pharmacother.* **2005**, *59*, 452.
- Sriram, D.; Yogeeswari, P.; Srichakravarthy, N.; Bal, T. R. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1085.
- Tantishaiyakul, V.; Wiwattanawongsa, K.; Pinsuan, S.; Kasiwong, S.; Phadoongsombut, N.; Kaewnopparat, S.; Kaewnopparat, N.; Rojanasakul, Y. *Pharm. Res.* **2002**, *19*, 1013.
- Thomsen, A. E.; Friedrichsen, G. M.; Sorensen, A. H.; Andersen, R.; Nielsen, C. U.; Brodin, B.; Begtrup, M.; Frokjaer, S.; Steffansen, B. *J. Control. Release* **2003**, *86*, 279.
- Vaddi, H. K.; Hamad, M. O.; Chen, J.; Banks, S. L.; Crooks, P. A.; Stinchcomb, A. L. *Pharm. Res.* **2005**, *22*, 758.
- Yang, Y. W.; Lee, J. S.; Kim, I.; Jung, Y. J.; Kim, Y. M. *Eur. J. Pharm. Biopharm.* **2007**, *66*, 260.
- Nielsen, C. U.; Andersen, R.; Brodin, B.; Frokjaer, S.; Steffansen, B. *J. Control. Release* **2001**, *73*, 21.
- Nielsen, A. B.; Buur, A.; Larsen, C. *Eur. J. Pharm. Sci.* **2005**, *24*, 433.
- Ribeiro, L.; Silva, N.; Iley, J.; Rautio, J.; Jarvinen, T.; Mota-Filipe, H.; Moreira, R.; Mendes, E. *Arch. Pharm. Chem. Life Sci.* **2007**, *340*, 32.
- Brunner-guenat, M.; Carrupt, P.; Lisa, G.; Testa, B.; Rose, S.; Thomas, K.; Jenner, P.; Ventura, P. *J. Pharm. Pharmacol.* **1995**, *47*, 861.
- Mahfouz, N. M.; Hassan, M. A. *J. Pharm. Pharmacol.* **2001**, *53*, 841.
- Nielsen, N. M.; Bundgaard, H. *J. Pharm. Pharmacol.* **1988**, *40*, 506.
- Najlah, M.; Freeman, S.; Attwood, D.; D'Emanuele, A. *Int. J. Pharm.* **2006**, *308*, 175.
- Camenish, G. P.; Wang, W.; Wang, B.; Borchardt, R. T. *Pharm. Res.* **1998**, *15*, 1174.
- Iley, J.; Moreira, R.; Calheiros, T.; Mendes, E. *Pharm. Res.* **1997**, *14*, 1634.
- Little, R. J.; Bodor, N.; Loftsson, T. *Pharm. Res.* **1999**, *16*, 961.
- Mahfouz, N. M.; Aboul-Fadl, T.; Diab, A. K. *Eur. J. Med. Chem.* **1998**, *33*, 675.
- Yang, J. Z.; Chen, W.; Borchardt, R. T. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 840.
- Liederer, B. M.; Borchardt, R. T. *J. Pharm. Sci.* **2005**, *94*, 2198.
- Bak, A.; Fich, M.; Larsen, B. D.; Frokjaer, S.; Friis, G. *Eur. J. Pharm. Sci.* **1999**, *7*, 317.
- Di Stefano, A.; Mosciatti, B.; Cingolani, G. M.; Giorgioni, G.; Ricciuti, M.; Cacciatore, I.; Sozio, P.; Claudi, F. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1085.
- Hernandez-Luis, F.; Hernandez-Campos, A.; Yopez-Mulia, L.; Cedillo, R.; Castillo, R. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 1359.
- Lopes, F.; Moreira, R.; Iley, J. *Bioorg. Med. Chem.* **2000**, *8*, 707.
- Redden, P. R.; Melanson, R. L.; Douglas, J. E.; Dick, A. J. *Int. J. Pharm.* **1999**, *180*, 151.
- Bundgaard, H.; Falch, E.; Larsen, C.; Mosher, G. L.; Mikkelsen, T. J. *J. Pharm. Sci.* **1986**, *75*, 775.
- Bundgaard, H.; Jensen, E.; Falch, E. *Pharm. Res.* **1991**, *8*, 1087.
- Fix, J. A.; Alexander, J.; Cortese, M.; Engle, K.; Leppert, P.; Repta, A. *J. Pharm. Res.* **1989**, *6*, 501.
- Carvalho, E.; Iley, J.; Perry, M.; Rosa, E. *Pharm. Res.* **1998**, *15*, 931.
- Di Stefano, A.; Sozio, P.; Cocco, A.; Iannitelli, A.; Santucci, E.; Costa, M.; Pecci, L.; Nasuti, C.; Cantalamessa, F.; Pinnen, F. *J. Med. Chem.* **2006**, *49*, 1486.
- Mignat, C.; Heber, D.; Schlicht, H.; Ziegler, A. *J. Pharm. Sci.* **1996**, *85*, 690.
- Mork, N.; Bundgaard, H. *Pharm. Res.* **1992**, *9*, 492.
- Procopiou, P. A.; Biggadike, K.; English, A. F.; Farrell, R. M.; Hagger, G. N.; Hancock, A. P.; Haase, M. V.; Irving, W. R.; Sareen, M.; Snowden, M. A.; Solanke, Y. E.; Tralau-Stewart, C. J.; Walton, S. E.; Wood, J. A. *J. Med. Chem.* **2001**, *44*, 602.
- Nielsen, C. U.; Andersen, R.; Brodin, B.; Frokjaer, S.; Taub, M. E.; Steffansen, B. *J. Control. Release* **2001**, *76*, 129.
- Shen, M.; Xiao, Y.; Golbraikh, A.; Gombar, V.; Thropsha, A. *J. Med. Chem.* **2003**, *46*, 3013.
- Jensen, B. F.; Sorensen, M. D.; Kissmeyer, A.; Bjorkling, F.; Sonne, K.; Engelsen, S. B.; Norgaard, L. *J. Comput. Aided Des.* **2003**, *17*, 849.
- 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.; Tanchuk, V. Y.; Prokopenko, V. V. *J. Comput. Aided Mol. Des.* **2005**, *19*, 453.
- Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools and techniques*; Morgan Kaufmann: San Francisco, 2005.
- <http://www.mdli.com/>.
- Pedretti, A.; Villa, L.; Vistoli, G. *J. Comput. Aided Mol. Des.* **2004**, *18*, 167.
- Dewar, M. J. S.; Zebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902.
- Hudson, B. D.; Hyde, R. M.; Rahr, E.; Wood, J. *Quant. Struct.-Act. Relat.* **1996**, *15*, 285.
- Snarey, M.; Terrett, N. K.; Willett, P.; Wilton, D. J. *J. Mol. Graphics Modell.* **1997**, *15*, 373.
- Nilakantan, R.; Bauman, N.; Haraki, K. S. *J. Comput. Aided Mol. Des.* **1997**, *11*, 447.
- Clark, R. D. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1181.
- Coi, A.; Fiamingo, F. L.; Livi, O.; Calderone, V.; Martelli, A.; Massarelli, I.; Bianucci, A. M. *Bioorg. Med. Chem.* **2009**, *17*, 319.
- Guidance document on the validation of (quantitative) structure-activity relationship [(Q)SAR] models. OECD 2007.
- Kraemer, H. C. In *Kappa Coefficient' in Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons: New York, 1982.

63. Atkeson, C. G.; Moore, A. W.; Schaal, S. *Art. Intel. Rev.* **1997**, *11*, 11.
64. Hall, M. A. *Correlation-based Feature Subset Selection for Machine Learning*; Hamilton: New Zealand, 1998.
65. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
66. Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.
67. Balaban, A. T. *From Chemical Topology to 3D Molecular Geometry*; Plenum Press: New York, 1997.
68. Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D QSAR in Drug Design*; Kluwer/ESCOM: Dordrecht, 1998.
69. Consonni, V.; Todeschini, R.; Hottje, H. D.; Sippl, W. *Rational Approaches to Drug Design*; Prous Science: Barcelona, 2001.
70. Randic, M. *Acta Chim. Slov.* **1998**, *45*, 239.
71. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; van de Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. *ATLA* **2005**, *33*, 155.