European Journal of Medicinal Chemistry 46 (2011) 218–228

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Prediction of drug intestinal absorption by new linear and non-linear QSPR

Alan Talevi [a], Mohammad Goodarzi [b], Erlinda V. Ortiz [c], Pablo R. Duchowicz [b,*], Carolina L. Bellera [a], Guido Pesce [d], Eduardo A. Castro [b], Luis E. Bruno-Blanch [a]

[a] *Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP), 47 y 115, 1900 La Plata, Argentina*
[b] *Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, 1900 La Plata, Argentina*
[c] *Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, Av. Maximio Victoria 55, 4700 Catamarca, Argentina*
[d] *Department of Pharmacology, National Institute of Medicines (INAME), National Administration of Food, Medicines and Medical Technology (ANMAT), Buenos Aires, Argentina*

ARTICLE INFO

Article history:
Received 8 February 2010
Received in revised form
31 October 2010
Accepted 1 November 2010
Available online 9 November 2010

Keywords:
QSPR theory
Molecular descriptors
Replacement method
Drug intestinal absorption
Model's applicability domain
ADME properties

ABSTRACT

In order to minimize the high attrition rate that usually characterizes drug research and development projects, current medicinal chemists aim to characterize both pharmacological and ADME profiles at the beginning of drug R&D initiatives. Thus, the development of ADME High-Throughput Screening *in vitro* and *in silico* ADME models has become an important growing research area. Here we present new linear and non-linear predictive QSPR models to predict the human intestinal absorption rate, which are derived from a medium sized, balanced and diverse training set of organic compounds. The structure–property relationships so obtained involve only 4 molecular descriptors, and display an excellent ratio of number of cases to number of descriptors. Their adjustment of the training set data together with the performance achieved during the internal and external validation procedures are comparable to previously reported modeling efforts.

1. Introduction

The former paradigm of drug development has focused on the optimization of the molecule in order to gain potency and selectivity. As a consequence, drug development has been characterized by a high attrition rate, with about nine from ten drugs that have entered clinical trials which have not made it into the pharmaceutical market, mostly because of toxicity issues or the inability of the drug to reach its pharmacological target [1,2]. Moreover, modern Combinatorial Chemistry and High-Throughput Screening (HTS) technologies have tended to produce novel entities with poor ADME properties [3]. In order to reduce the rate of failure of drug development programs due to ADMET issues at late stages of the research (i.e. clinical trials), the modern paradigm of drug development has moved towards finding a balance between potency, bioavailability and safety from the very beginning of the project. This modern paradigm may be synthesized under the expression “to fail early is to fail cheap”, and is implemented by including at early stages parallel ADMET filters to discard chemical entities with unfavorable pharmacokinetic and toxicity profiles.

The oral route is generally considered the most convenient route of administration because of production costs, stability of the drug and ease of administration and transport. For this reason, some of the most important physicochemical properties currently assessed at the beginning of a novel drug project are aqueous solubility, intestinal permeability and oral bioavailability. A possible way for assessing ADMET-related physicochemical properties is to rely on *in vitro* assays. For instance, Caco-2 or MDCK cells and parallel artificial membrane permeation assays (PAMPA) [4], coupled with high-throughput liquid chromatography–mass spectrometry [5], and liquid chromatography using special stationary or mobile phases (e.g. the immobilized artificial membrane – IAM – technique) [6], have proved successful to simulate transport processes. Although there have been significant innovations in the area of high throughput *in vitro* ADME screening in the last few years [7,9], *in vitro* assays are laborious, expensive, time consuming and demand certain drug quantities, usually more than what is produced in a standard combinatorial library synthesis. Thus, *in vitro* ADME assays are still not entirely compatible with HTS technologies [8]: compounds are currently synthesized and pharmacologically screened much faster than the speed by which experimental ADME studies can be carried out, and ADME studies have become a bottleneck during modern drug discovery efforts. *In silico* models thus constitute an inexpensive and faster option to apply at early

[*] Corresponding author. Fax: +54 221 425 4642.
E-mail addresses: prduchowicz@yahoo.com.ar, pabloducho@gmail.com (P.R. Duchowicz).

0223-5234/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2010.11.005

stages of the screening process, and are completely compatible with HTS. These can be also applied to estimate ADME properties of a compound *prior* to synthesis, providing a reference frame to guide an 'informed synthesis' and improving the chances of landing in better chemical space [8–10]. When a large number of compounds have been synthesized, *in silico* ADME models can be build from experimental data of a representative subset of the compounds and then used to predict the attributes of the remainder [8–10]. *In silico* models of ADMET-related properties may rely on either theoretical or experimental descriptors to establish a quantitative relationship between the target property and the molecular structure of a set of compounds. Quantitative Structure Property–Activity Relationships (QSPR-QSAR) may be derived through a wide range of linear or non-linear models [11,12].

As part of our ongoing research to build QSAR models of ADME properties to assist drug development projects [13–16], we now develop both linear and non-linear models to estimate Human Intestinal Absorption rate (%HIA). Tables 1 and 2 include a summary of some of the most notable efforts to model Human Intestinal Absorption rate (%HIA), which have been reported in the last 10 years. Table 1 shows a summary of reported quantitative models, aimed to predict exact %HIA values; Table 2 presents a review of classificatory models, aimed to classify a given drug into one of two or more %HIA categories. Only the best model reported in each article, in terms of squared correlation coefficient (R^2) or root mean squared error (*RMSE*) in the training set in Table 1, or overall percentage of good classifications or *RMSE* in Table 2, is included in each of the tables. Note that many of the training and/or test sets of these models are clearly heavily biased towards highly permeable drugs [17–27,29–33], while some models present very low cases to descriptors ratio *N*/*D* (with the resultant risk of overfitting) or have been derived from scarce training sets that limit the applicability domain of the models. Moreover, some of them are based on experimental descriptors, jeopardizing their applicability in HTS campaigns (e.g. the model presented in Deconinck et al. [26] requires the determination of the retention time of each compound in three different HPLC systems). In some cases 3D descriptors are included in the models without a systematic conformational analysis of the structures of the training set [26]. Another limitation of these models is that they have been derived

Table 1
Summary of modeling efforts towards quantitative prediction of %HIA in the last ten years. 'In house or literature' indicates whether the %HIA data have been measured either by the authors or obtained from literature. Only results of the external validation have been summarized in the table, although in some cases the authors performed internal cross-validation procedures.

Authors	*N*	Data source	Descriptor types	Modeling technique	*N*/*d*	R^2 training set	*RMSE* training set	R^2 test set	*RMSE* test set
Wessel et al. [17]	76	Literature	1D–3D theoretical descriptors (functional group counts, topological, geometrical and electronic descriptors)	Combination of GA and ANN	12.7	–	9.4[b]	–	16.0[b]
Norinder et al. [18]	20	Literature	MolSurf theoretical physicochemical descriptors related to oral bioavailability	Principal Components Analysis plus PLS	6.7[c]	0.92	0.49[a]	–	–
Österberg and Norinder [19]	20	Literature	Number of H bond acceptor nitrogens and oxygens, number of H bond donors, log P	PLS	5.0	0.81	1.48[a]	–	–
Agatonovic-Kustrin et al. [20]	76	Literature	0D–3D theoretical descriptors (constitutional, topological, geometrical, quantum-chemical descriptors)	Combination of GA and ANN	5.1	–	0.59 [a]	0.802	0.42 [a]
Klopman et al. [21]	417	Literature	1D theoretical descriptors plus 6 basic parameters related to oral absorption	Combination of Group Contribution and CASE	11.3	0.79	12.3[b]	0.79	12.3[b]
Abraham et al. [22]	127	In house	Abraham's solvation parameters	MLR	25.4	0.80	0.29[a]	–	–
Niwa [23]	67	Literature	0D and 1D theoretical descriptors (constitutional descriptors and count of functional groups and atom types)	Neural networks (ANN)	9.6	–	6.5[b]	–	22.8[b]
Sun [24]	169	In house	Theoretical (Atom types)	PLS	56.3[c]	0.92	–	–	–
Yen et al. [25]	52	Literature	Experimental (chromatographic) descriptor (IAM chromatography) plus Molecular Modeling Pro topological, geometrical and physicochemical descriptors	Multiple Linear Regression (MLR)	5.8	0.68	–	–	–
Deconinck et al. [26]	67	Literature	Experimental (chromatographic) descriptors combined with Dragon and Hyperchem theoretical 1D–3D descriptors plus one of Abraham's solvation parameters.	Non-linear, MARS	9.6	0.93	–	–	–
Yan et al. [27]	380	Literature	Adriana Code and Cerius2 0D–2D theoretical descriptors (constitutional, functional group counts, topological and physicochemical descriptors)	Combination of Genetic Algorithms (GA), Partial Least Squares (PLS) and Support Vector Machine (SVM)	42.2	0.66	12.5[b]	0.77	16.0[b]
Reynolds et al. [28]	567	Literature	ADME Boxes and Algorithm Builder; log P experimental values were used when available	Non-Linear Least Squares (NLS)	43.6	0.93	9.5[b]	–	0.35[a],[d] 0.45[a],[e]
Guerra et al. [29]	37	Literature	Codes 2D (topological) theoretical descriptors	ANN	12.3	0.93	8.0[b]	–	–
Talevi et al. [this work]	120	Literature	0D–3D Dragon theoretical descriptors	Linear (MLR) and non-linear	30.0	0.80	0.18[a]	0.66	0.21[a]

N refers to the number of compounds in the training set. *N*/*d* refers to the ratio between the number of cases in the training set and the number of independent variables included in the best model.

[a] Expressed in log units (the dependent variable is some log transformation of an experimental variable linked to intestinal absorption, usually %HIA).

[b] Expressed in % units.

[c] The ratio is calculated considering the number of PLS latent variables as the number of independent variables, though these latent variables are combinations of a higher number of descriptors.

[d] *RMSE* reported when logarithm of absorption rate constant is taken as dependent variable.

[e] *RMSE* reported when logarithm of human permeability coefficients taken as dependent variable.

Table 2
Summary of modeling efforts towards classificatory models of %HIA in the last ten years.

Authors	*N*	Data source	Descriptor types	Modeling technique	*N*/*d*	% of good classifications in training set	% of good classifications in the test set
Niwa [23]	67	Literature	0D and 1D descriptors (constitutional descriptors and counts of functional groups and atom types)	Neural networks	9.6	100.0	80.0
Wolohan and Clark [30]	86	Literature	CoMFA and CoMSIA fields	Soft independent modeling of class analogies (SIMCA)	–	90.7	–
Cabrera Pérez et al. [31]	82	Literature	Topological	Linear Discriminant Analysis	16.4	89.0	92.4
Deconinck et al. [32]	141	Literature	Dragon and Hyperchem theoretical 1D–3D descriptors plus McGowan volumes.	Classification and regression trees	15.7	88.9	85.2
Hou et al. [33]	480	Literature	0D–2D descriptors related to oral bioavailability	Support vector machine	68.6	97.3	98.0
Yang et al. [34]	196	Literature	0D–2D descriptors	Combination of GA, conjugate gradient and Support vector machine	7.8	87.2	–

N refers to the number of compounds in the training set.

from absorption data gathered from literature instead of in house, standardized experiments; therefore, variability in the %HIA may reflect not only differences in the molecular structure of the compounds but differences in the experimental determination of the absorption rate as well.

In the present work we try to address some of the aforementioned issues, by means of balancing the distribution of low and high permeable compounds in a medium-size dataset in order to build new QSPR models of %HIA posing high *N*/*D* ratio. Each structure in the dataset is optimized through a systematic conformational analysis in order to obtain descriptor values from probable, low energy conformations.

2. Materials and methods

2.1. Dataset

The experimental %HIA permeability values of 160 organic compounds are extracted from the literature [20,26,32,33] and provided together with the chemical names are provided in Table 3. For modeling purposes, we convert %HIA into logarithm units, $\log_{10}(\%\text{HIA}+A)$. The value of the *A* parameter is selected in such a way to achieve the best predictions in the training set, and so we manually assign to *A* different numerical values until minimizing the root mean squared error of prediction (*RMSE*), leading to $A = 10$ for the linear models of Eqs. (6) and (7) (see below).

In order to design a balanced training set, we consider four categories of %HIA: category one: $\%\text{HIA} \leq 20\%$; category two: $20\% < \%\text{HIA} \leq 50\%$; category three: $50\% < \%\text{HIA} \leq 80\%$ and; category four: $\%\text{HIA} > 80\%$, and we try to obtain an even distribution of the 160 compounds among these categories. The 160 selected compounds are distributed as follows: category one, 46 compounds; category two, 26 compounds; category three, 41 compounds; category four, 46 compounds. This is not an entirely even distribution among the four categories but is not either highly biased towards highly permeable compounds. The complete molecular set is then split into a 90-compounds training set (train), 30-compounds validation set (val) and a 40-compounds test set (test) through systematic random sampling: the compounds which compose each category are alphabetically sorted and one each four compounds is removed to the test set (starting from a random compound in each category), obtaining a similar distribution of the modeled property in both the training and the test sets. It is to be noted that the training set covers the range of variation of the permeability values in the validation set. The %HIA experimental values are checked in Hazardous Substances Data Bank (HSDB) and Pubchem [35,36]; we discard compounds that according to different sources belong to different of the four categories considered.

Bibliographic search (taking into account Hou et al. [37] careful analysis on the transport mechanisms of their 648-compound permeability dataset, mainly) showed that around 10% of the compounds of our dataset have carrier-mediated transport (adefovir, foscarnet, k-sthrophantoside, ouabain, amiloride, fosfomycin, methyldopa, captopril, cefatrizine, enalapril maleate, amoxicilline and warfarin) or paracellular transport (Lucifer yellow, mannitol). At first, following Reynolds discussion (stating that some compounds that are substrates of ATP-dependent efflux carriers as digoxin have high experimental %HIA values, while other drugs suspected of active influx – such as tetracyclines, peptides and β-lactams – tend to present low %HIA values) we have chosen to keep this drugs in our dataset; however, since our results showed many of these compounds present high residuals, we have derived a second model without these potential outliers.

The structural diversity of the training set is assessed through the calculation of the average Tanimoto similarity (based on atom pairs) considering all possible pairs of compounds that can be derived from the training set. For this purpose, we use PowerMV software freely provided by the National Institute of Statistical Sciences [38]. According to the results, the average Tanimoto similarity for the dataset is 0.329 ($S = 0.415$) which confirms high structural diversity. The average similarity of the test set compounds to the training set compounds 0.333 ($S = 0.538$) is similar to the average similarity among the compounds of the dataset. A systematic conformational analysis of the 160 compounds that compose the dataset is performed with the Conformational Analysis tool from Hyperchem 6.03 package [39], using the Molecular Mechanics Force Field (MM^+) and a gradient norm limit of 0.1 kcal mol^{-1} (number of simultaneous variation range from1 to 8; acyclic torsion variation range from 60 to 180°; range for cyclic torsion variation from 30 to 120°). The geometries of three conformers of lowest energy are further refined by means of the Semiempirical Method PM3 (Parametric Method-3) using the Polak-Ribiere algorithm and a gradient norm limit of 0.05 kcal Å^{-1}, retaining the lowest energy conformation for calculating the molecular descriptors of each molecule.

2.2. Molecular descriptors

We compute 1497 molecular descriptors using the software Dragon [40], including descriptors of all types such as Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted

Table 3
Experimental and QSPR predicted log$_{10}$(%HIA + 10) for 160 heterogeneous organic compounds. The %HIA category of the compound is indicated by a number preceding the compound name.

ID	Compound name	Exp.[a]	Eq. (5)	BRANN	LMANN	RBFNN	SVMR
1	1Acamprosate	1.322	1.636	1.534	1.529	1.534	1.584
2	1Acarbose^	1.079	0.778	1.109	1.060	1.109	1.077
3	1Adefovir	1.342	1.527	1.401	1.332	1.401	1.471
4	1Amygdalin^^	1.176	1.189	1.180	1.115	1.180	1.162
5	1Anphotericinb	1.176	1.210	1.199	1.152	1.199	1.288
6	1Arbekacin^	1.000	1.001	1.076	1.065	1.076	0.988
7	1Azlocillin	1.000	1.406	1.252	1.263	1.252	1.359
8	1Aztreonam^^	1.041	1.293	0.959	1.107	0.959	1.067
9	1Cefodizime	1.000	1.080	1.071	1.045	1.071	0.990
10	1Ceftriaxone	1.041	1.056	1.062	1.250	1.062	1.178
11	1Cefuroxime	1.176	1.335	1.167	1.214	1.167	1.318
12	1Cidofovir^^	1.114	1.463	1.272	1.229	1.272	1.366
13	1Cromolyn	1.021	1.211	1.198	1.036	1.198	1.084
14	1Doxorubicin	1.176	1.403	1.325	1.385	1.325	1.411
15	1Edetic acid	1.176	0.913	1.182	1.029	1.182	1.166
16	1Foscarnet^^	1.431	1.496	1.343	1.045	1.343	1.396
17	1Ganciclovir	1.134	1.451	1.444	1.420	1.444	1.479
18	1Gentamycin	1.000	1.238	1.195	1.140	1.195	1.194
19	1Imipenem^	1.176	1.428	1.255	1.276	1.255	1.439
20	1Iohexol^^	1.176	0.843	1.227	1.171	1.227	1.207
21	1Iotroxic acid	1.176	0.975	1.172	1.192	1.172	1.175
22	1Iothalamate	1.076	1.217	1.089	1.252	1.089	1.085
23	1Kanamycin	1.041	0.935	1.069	1.062	1.069	0.986
24	1k-sthrophantoside^^	1.415	1.184	1.187	1.099	1.187	1.223
25	1Lactulose	1.025	1.121	1.116	1.082	1.116	1.053
26	1Lucifer yellow	1.000	1.352	1.223	1.149	1.223	1.182
27	1Mannitol	1.415	1.424	1.423	1.371	1.423	1.459
28	1Meropenem^^	1.000	1.440	1.291	1.317	1.291	1.408
29	1Mezlocillin	1.000	1.344	1.090	1.133	1.090	1.109
30	1Mitoxanthrone^	1.176	1.445	1.398	1.487	1.398	1.461
31	1Moexiprildiacid	1.176	1.480	1.335	1.304	1.335	1.306
32	1Nedocromil^^	1.114	1.414	1.272	1.167	1.272	1.194
33	1Neomycin	1.041	0.772	1.058	1.059	1.058	1.032
34	1Netilmycin	1.000	1.268	1.218	1.171	1.218	1.238
35	1Olsalazine	1.090	1.375	1.243	1.137	1.243	1.254
36	1Ouabain^^	1.057	1.478	1.427	1.569	1.427	1.505
37	1Pamidronic acid	1.176	1.323	1.096	1.110	1.096	1.100
38	1Pentamidine	1.000	1.829	1.879	1.708	1.879	1.810
39	1Phthalylsulfathiazole	1.176	1.608	1.471	1.491	1.471	1.527
40	1Raffinose^^	1.013	0.881	1.088	1.061	1.088	1.017
41	1Risedronic acid	1.041	1.464	1.377	1.298	1.377	1.368
42	1Streptomycin^	1.041	1.309	1.164	1.070	1.164	1.078
43	1Streptosozin	1.000	1.260	1.142	1.128	1.142	1.171
44	1Succinylsulfathiazole^^	1.176	1.461	1.139	1.140	1.139	1.302
45	1Ticarcillin	1.176	1.309	1.164	1.070	1.164	1.078
46	1Tobramycin	1.000	1.041	1.078	1.069	1.078	0.994
47	2AAFC^^	1.623	1.681	1.800	1.835	1.800	1.821
48	2Amiloride	1.778	1.624	1.658	1.637	1.658	1.695
49	2Azithromycin^	1.663	1.541	1.536	1.684	1.536	1.578
50	2Benazepril	1.672	1.753	1.805	1.764	1.805	1.770
51	2Bromocriptine^^	1.580	1.614	1.510	1.417	1.510	1.555
52	2Cefpodoximeproxetyl	1.778	1.483	1.457	1.563	1.457	1.502
53	2Chlorothiazide^	1.528	1.495	1.425	1.335	1.425	1.396
54	2Cymarin	1.756	1.732	1.804	1.663	1.804	1.728
55	2Dihydroergotamine^^	1.653	1.869	1.881	1.695	1.881	1.822
56	2Famotidine	1.681	1.612	1.566	1.532	1.566	1.594
57	2Flucloxacillin^	1.699	1.703	1.732	1.839	1.732	1.714
58	2Fosfomycin	1.613	1.700	1.624	1.655	1.624	1.620
59	2Fosmidomycin^^	1.602	1.681	1.648	1.707	1.648	1.702
60	2Guanoxan	1.778	1.836	1.970	1.923	1.970	1.982
61	2Lincomycin^	1.574	1.529	1.591	1.691	1.591	1.599
62	2Lisinopril	1.544	1.451	1.340	1.274	1.340	1.325
63	2Lovastatin^^	1.608	2.075	1.911	1.812	1.911	1.889
64	2Metaproterenol	1.732	1.779	1.920	1.945	1.920	1.920
65	2Methyldopa^	1.708	1.691	1.780	1.829	1.780	1.832
66	2Nadolol	1.613	1.796	1.898	1.905	1.898	1.891
67	2Pafenolol^^	1.591	1.864	1.947	1.942	1.947	1.963
68	2Pravastatin	1.643	1.715	1.728	1.692	1.728	1.698
69	2Rimiterol^	1.763	1.789	1.935	1.941	1.935	1.944
70	2Sulpiride	1.732	1.710	1.755	1.799	1.755	1.774
71	2Trandolapril^^	1.778	1.748	1.798	1.755	1.798	1.764
72	2Zonavir	1.580	1.544	1.643	1.639	1.643	1.644

(*continued on next page*)

Table 3 (*continued*)

ID	Compound name	Exp.[a]	Eq. (5)	BRANN	LMANN	RBFNN	SVMR
73	3Almotriptan^	1.929	1.848	1.860	1.912	1.860	1.946
74	3Anagrelide	1.903	2.022	1.973	1.961	1.973	2.026
75	3Atenolol^^	1.785	1.829	1.962	1.944	1.962	1.978
76	3Benzbromarone	1.919	1.809	1.794	1.882	1.794	1.908
77	3Benserazide^	1.903	1.465	1.471	1.428	1.471	1.508
78	3Bromhexine	1.903	1.834	1.818	1.956	1.818	1.898
79	3Captopril^^	1.892	1.719	1.829	1.878	1.829	1.861
80	3Cefatrizine	1.934	1.162	1.609	1.239	1.609	1.191
81	3Cycloserine^	1.919	1.980	1.801	1.929	1.801	1.892
82	3Dipyridamole	1.833	1.576	1.625	1.774	1.625	1.622
83	3Eflornithine^^	1.813	1.714	1.773	1.845	1.773	1.830
84	3Enalapril maleate	1.881	1.670	1.720	1.815	1.720	1.727
85	3Ethambutol^	1.954	1.816	2.014	1.970	2.014	1.985
86	3Etodolac	1.903	1.795	1.898	1.947	1.898	1.879
87	3Famciclovir^^	1.903	1.670	1.813	1.913	1.813	1.786
88	3Fenoterol	1.845	1.800	1.887	1.838	1.887	1.851
89	3Furosemide^	1.851	1.551	1.299	1.267	1.299	1.464
90	3Guanabenz	1.954	2.057	2.006	1.911	2.006	1.941
91	3Hydrochlothiazide^^	1.875	1.509	1.453	1.366	1.453	1.423
92	3Isocarboxazid	1.903	1.818	1.971	1.930	1.971	1.976
93	3Ivermectin^	1.845	1.701	1.764	1.637	1.764	1.692
94	3Metformin	1.799	2.047	1.813	1.874	1.813	1.819
95	3Metolazone^^	1.863	1.902	1.888	1.922	1.888	1.994
96	3Mianserin	1.903	1.943	1.896	1.736	1.896	1.848
97	3Mibefradil^	1.898	2.083	1.913	1.817	1.913	1.900
98	3Moxisylyte	1.903	2.057	1.976	2.001	1.976	1.998
99	3Oxycodone^^	1.845	1.843	1.904	1.886	1.904	1.927
100	3Oxytetracycline	1.833	1.324	1.245	1.203	1.245	1.283
101	3Pimozide^	1.903	2.240	1.915	1.879	1.915	1.907
102	3Propylthiouracil	1.929	1.851	1.926	1.911	1.926	1.948
103	3Pyrbuterol^^	1.845	1.661	1.836	1.903	1.836	1.812
104	3Quetiapine	1.919	1.936	1.911	1.825	1.911	1.939
105	3Ramipril^	1.845	1.727	1.776	1.778	1.776	1.750
106	3Ranitidine	1.798	1.828	1.952	1.919	1.952	1.976
107	3Recainam^^	1.909	2.064	2.009	2.019	2.009	2.003
108	3Reproterol	1.845	1.521	1.561	1.701	1.561	1.580
109	3Terbutaline^	1.857	1.797	1.918	1.937	1.918	1.917
110	3Tolrestat	1.881	1.884	1.934	1.905	1.934	1.936
111	3Urapidil^^	1.945	1.761	1.888	1.926	1.888	1.870
112	3Valsartan	1.813	1.610	1.583	1.731	1.583	1.628
113	3Ziprasidone^	1.845	2.013	1.919	1.844	1.919	1.960
114	4Acebutolol	1.999	1.812	1.899	1.876	1.899	1.881
115	4Acetaminophen^^	1.978	2.008	1.998	1.891	1.998	1.922
116	4Almitrine	2.000	2.139	1.908	1.850	1.908	1.890
117	4Alprenolol^	2.016	1.961	1.955	1.938	1.955	1.976
118	4Aminopyrine	2.041	1.886	1.981	1.915	1.981	2.003
119	4Amoxicillin^^	2.016	1.432	1.274	1.293	1.274	1.700
120	4Antipyrine	2.041	1.952	2.011	1.910	2.011	1.998
121	4Aspirin	2.041	1.804	1.981	1.991	1.981	1.948
122	4Atropine	2.033	1.953	1.960	1.952	1.960	1.992
123	4Benzydamine^^	1.987	1.887	1.899	1.751	1.899	1.868
124	4Betaxolol	2.000	1.920	1.928	1.851	1.928	1.924
125	4Bupropion	1.987	2.100	1.975	2.022	1.975	1.993
126	4Caffeine^	2.041	1.704	1.832	1.869	1.832	1.852
127	4Chloramphenicol^^	2.000	1.852	1.951	1.923	1.951	1.990
128	4Clofibrate	1.987	2.078	1.986	2.020	1.986	2.002
129	4Codeine	2.021	1.962	1.933	1.915	1.933	1.990
130	4Diclofenac	2.041	2.062	2.051	2.104	2.051	2.068
131	4Disulfiram^^	2.029	1.823	1.883	1.877	1.883	1.910
132	4Felbamate	2.000	1.979	2.017	1.965	2.017	2.021
133	4Fluconazole^	2.027	1.644	1.775	1.898	1.775	1.746
134	4Hydrocortisone	2.004	1.801	1.867	1.742	1.867	1.816
135	4Ibuprofen^^	2.041	2.033	2.109	2.100	2.109	2.033
136	4Indomethacin	2.041	1.887	1.949	1.931	1.949	1.952
137	4Ketoprofen^	2.009	1.958	2.024	2.004	2.024	2.032
138	4Ketorolac	2.000	1.716	1.814	1.901	1.814	1.804
139	4Labetalol^^	2.021	1.902	1.929	1.886	1.929	1.940
140	4Lamivudine	1.987	1.625	1.749	1.766	1.749	1.753
141	4Lansoprazol	1.978	1.866	1.894	1.905	1.894	1.966
142	4Minoxidil^	2.033	1.849	1.981	1.924	1.981	1.991
143	4Moricizine^^	1.991	1.956	1.910	1.807	1.910	1.936
144	4Moxonidine	1.991	1.717	1.875	1.903	1.875	1.879
145	4Naproxen	2.037	1.957	2.058	2.068	2.058	2.021
146	4Nitrendipine	1.991	1.915	1.929	1.898	1.929	1.960
147	4Nordiazepam^^	2.037	2.125	2.000	2.043	2.000	1.991

Table 3 (*continued*)

ID	Compound name	Exp.[a]	Eq. (5)	BRANN	LMANN	RBFNN	SVMR
148	4Oxyfedrine	1.978	1.924	1.920	1.814	1.920	1.909
149	4Propanolol^	2.037	1.984	1.953	1.948	1.953	1.993
150	4Rivastigmine	2.041	2.125	2.032	2.059	2.032	1.974
151	4Saccharin^^	1.991	1.761	1.676	1.809	1.676	1.755
152	4Sotalol	2.021	1.767	1.792	1.865	1.792	1.847
153	4Sultopride^	1.996	1.787	1.840	1.893	1.840	1.881
154	4Tenidap	1.996	1.811	1.892	1.887	1.892	1.905
155	4Timolol^^	2.021	1.577	1.677	1.740	1.677	1.668
156	4Tolbutamide	1.978	1.850	1.853	1.911	1.853	1.942
157	4Trapidil^	2.025	1.870	1.976	1.915	1.976	1.994
158	4Trimethoprim	2.029	1.789	1.916	1.933	1.916	1.926
159	4Warfarin^^	2.033	1.892	1.908	1.805	1.908	1.900
160	4Zalcitabine	1.978	1.681	1.857	1.880	1.857	1.849

^Validation set, ^^External test set.

[a] Exp.: experimental permeability.

Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), Molecular Walk Counts, BCUT descriptors, 2D-Autocorrelations, Aromaticity Indices, Randic Molecular Profiles, Radial Distribution Functions, Functional Groups, Atom-centred fragments, Empirical and Properties [41]. Furthermore, 4 molecular descriptors were derived taking into consideration the Lipinski's rule, based on combinations of the detour index *dd* from the Chemical Graph Theory (derived as the half sum of the elements of the Detour Matrix – DD) [42] together with molecular features such as the number of H donors (D), the number of H acceptors (A) and the number of heteroatoms (H) present in the molecular structure [13]. We also considered the square and cubic roots of these last descriptors. Finally, 5 quantum-chemical descriptors not provided by the program Dragon were added to the pool: molecular dipole moments, total energies, homo–lumo energies, and homo–lumo gap ($\Delta_{\mathrm{homo-lumo}}$).

2.3. Linear model search

In recent years researchers have focused an increasing attention on finding the most efficient tool for variable selection in QSPR/QSAR studies. There are a lot of feature selection methods to search the best structural descriptors from a pool of variables, and the Replacement Method (RM) [43,44], employed here, has been successfully used elsewhere [14,45,46]. In brief, the RM is an efficient optimization tool which generates multi-parametric linear regression QSPR models by searching the set $\mathbf{D}$ of D descriptors for an optimal subset $\mathbf{d}$ of $d << D$ ones with minimum model's standard deviation (S). The quality of the results achieved with this technique is quite close to that obtained by performing an exact (combinatorial) full search of molecular descriptors, although, of course, requires much less computational work. We used the computer Matlab 7.0 system for all our calculations [47].

2.4. Non-linear model search

In past years, non-linear models have proven to be fruitful in the study of permeability of compounds [17,26–29]. Reynolds et al. have addressed the necessity of non-linear models taking into consideration that at small values of %HIA, small variations in %HIA correspond to very large variations of the effective passive permeability coefficient that depends on paracellular and transcellular transport routes (Peff) (e.g. %HIA from 0 to 1% correspond to log Peff from −1 to −6). Therefore, linear models tend to overestimate small Peff values [28]. These observations are in good agreement with our present results (see discussion regarding this subject in the Results section).

ANNs are computational architectures constructed with the goal of mimicking biological neural networks. The ANN artificial counterpart reproduces a similar functionality to the biological one, i.e. calculates a weighted sum of input signals and compares it against the activation function (or threshold) [48]. Theoretically, ANN models are a kind of black box models, whose accuracy depends on the data presented to it during the training stage. The collection of well-distributed, enough number of observations, and accurately measured-simulated input data is the basic requirement to obtain an acceptable non-linear model. In this context, ANNs can be trained to recognize patterns and the so developed models allow generalizing their conclusions to be applicable on patterns not encountered previously. ANNs may be defined as structures comprised of densely interconnected adaptive simple processing elements or units that are capable of performing massively parallel computations for data processing and knowledge representation. Each of those elements, also called neurons, forms a weighted sum of its inputs, to which a constant term called bias is added. This sum is then passed through a transfer function: linear, sigmoid or hyperbolic tangent. Multilayer Perceptron ANNs (MLP-ANNs) are the best known and most widely used kind of ANN. Networks with interconnections that do not form any loops are called feedforward. Recurrent or non-feedforward networks in which there are one or more loops of interconnections are used for some kinds of applications [49,50].

The feedforward MLP-ANN model generally consists of three layers (input layer, hidden layer, and output layer). In this context, the MLP-ANN architecture will be described here as "MLP-ANN m–n–k", where m, n, and k represent the number of neuron in the input, hidden, and output layers. The training process determines the MLP-ANN weights by minimizing some error criterion, which is related to the square of the error function (difference between the observed and predicted outputs); once determined, the weights and the associated network architecture together constitute the model and are therefore stored. The MLP-ANN model with the error back-propagation algorithm is the most popular ANN model for prediction.

A number of advanced algorithms have been proposed so far in MLP-ANNs learning. The neural model adopted in this study is a fully-connected MLP-ANN model, coupled with different training methods of the Matlab Environment [51], such as Levenberg-Marquardt (LM) [52,53], Levenberg-Marquardt with Bayesian regularization (BR), and Radial Basis Function (RBF). Both LMANN and BRANN algorithms attempt to use second derivatives-related information to accelerate the learning optimization process. Radial basis functions are a special kind of artificial neural network proposed by Moody and Darken in the late 1980s [54,55] However, their roots are entrenched in much older pattern recognition techniques for example potential functions, clustering, functional approximation, spline interpolation and mixture models. RBF networks consist of three layers, input, hidden and output layers. The input layer consists of a non-linear n-dimensional vector. The hidden layer includes a number of RBF units and the bias. Each neuron on the hidden layer uses a radial basis function for operating on the input data as a non-linear transformation function. In this study Gaussian function is used.

Vladimir Vapnik proposed Support Vector Machine (SVM) [56] which the basic idea is to map the data **X** into a higher-dimensional feature space **F** via a non-linear mapping Φ and then to do linear regression in this space. Given training data $(\mathbf{x}_1,\mathbf{y}_1),\dots,(\mathbf{x}_N,\mathbf{y}_N)$, where vector $\mathbf{x_i}$ contains independent variables, vector $\mathbf{y_i}$ contains dependent variables and N is total number of data patterns, the Support Vector Machine Regression (SVMR) model can be obtained through solving the following optimization problem:

$$\min_{w,b} = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left(\xi_i+\xi_i^*\right) \tag{1}$$

subjected to:

$$w\Phi(x_i)+b-y_i \le \varepsilon+\xi_i,\ y_i-w\Phi(x_i)-b_i \le \varepsilon+\xi_i^* \quad \varepsilon,\xi_i,\xi_i^* \ge 0 \quad i=1,2,\dots,N$$

where Φ is included by a kernel function and it is a non-linear mapping and C is a regularized constant determining the trade off between the empirical risk and regularization term. However, we need to construct a Lagrange function and from the Karush-Kuhn-Tucker (KKT) conditions, in order to solve Eq. (1), it thus yields the following dual Lagrangian form [57]; Maximize:

$$\Phi\left(\alpha_i,\alpha_i^*\right) = -\varepsilon\sum_{i=1}^{N}\left(\alpha_i+\alpha_i^*\right)+\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)y_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)\left(\alpha_j-\alpha_j^*\right)K\left(x_i,x_j\right) \tag{2}$$

subjected to

$$\sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right) = 0,\ 0\le\alpha_i,\alpha_i^*\le C \quad i=1,\dots,N \tag{3}$$

$K(x_i,x_j)$ is the kernel function which represents the inner product between two vectors in the high dimensional (even infinite) feature space. In this study we used of $K(x_i,x_j)$ as Gaussian however after solving Lagrangian multipliers of Eq. (2), the following predictor, viz. SVR, can be got [58].

$$f(x,w) = \sum_{i=1}^{N}\left(\alpha_i-\alpha_i^*\right)K\left(x_i,x_j\right)+b \tag{4}$$

2.5. Applicability domain analysis

The applicability domain of the models was defined through the Extent of Extrapolation or leverage approach [59], which is based in computing the leverage h_i for each compound for which the QSPR model is used to predict the property under study. The leverage is defined as $h_i = x_i^T(X^TX)^{-1}x_i$, where x_i is the descriptor vector of the considered compound I and X is the model matrix derived from the training set descriptor values. The warning leverage is generally fixed at $3k/n$, k being the number of model parameters and n being the number of training set compounds.

3. Results and discussion

As a first step, we search for the best linear QSPR through exploring multi-parametric regression models on the Human Intestinal Absorption rate, with the purpose of identifying the representative structural features of organic compounds that lead to their permeability behavior. The RM variable subset selection approach enables to select the most relevant structural descriptors on the training molecular set, exploring a pool containing more than a thousand numerical variables computed with Dragon. Table 4 includes the best 1–6 molecular descriptors found by the RM, while a brief description for the meaning of each descriptor is supplied by Table 5. It can be appreciated from Table 4 that all these models obey the semiempirical "Rule of Thumb" [11], stating that at least five or six data points should be present for each descriptor, and the more rigorous events per independent variables ratio of 10, suggested by the studies from Peduzzi et al. [60,61]. Furthermore, it is clear that four descriptors are able to lead to the best values for the statistical parameters of the validation set, that is to say, a monitoring set that is partly considered during the model development, so we select the following relationship according to this criterion:

$$\begin{aligned}\log_{10}(\text{HIA}+10) = {} & 1.801(\pm0.1)-0.100(\pm0.02)\cdot BEHm1 \\ & +1.639(\pm0.4)\cdot RNCG \\ & -0.139(\pm0.04)\cdot nCOOH \\ & +0.118(\pm0.01)\cdot MLOGP\end{aligned} \tag{5}$$

$$N=90,\ d=4,\ R^2=0.659,\ S=0.24,\ F=41.3,\ p<10^{-4},\ R^2_{\text{max}}=0.084,\ R^2_{\text{loo}}=0.618,\ S_{\text{loo}}=0.25,\ S_{\text{Rand}}=0.39,\ N_{\text{val}}=30,\ R^2_{\text{val}}=0.654,\ S_{\text{val}}=0.22,\ N_{\text{test}}=40,\ R^2_{\text{test}}=0.551,\ S_{\text{test}}=0.26$$

In this equation, N is the number of compounds in the training set, d is the number of descriptors of the model, R is the correlation coefficient, S stands for the standard deviation of calibration of the model, p is the significance of the model, R_{max} is the maximum intercorrelation coefficient between descriptors participating in Eq. (5), and subindex *loo* stands for the Leave One Out Cross-Validation technique [62]. The S_{Rand} parameter represents the standard deviation according to the Y-Randomization technique [63] (500000 cases).

Table 3 provides the predicted $\log_{10}(\%\text{HIA}+10)$ values for all the studied compounds according to Eq. (5). Validation compounds are denoted with symbol ^ in this table. Fig. 1a includes a graphical representation of the predictions as function of the experimental permeability values and Fig. 1b plots the residuals as function of the predictions. In general, all the range of variation of the observations pertaining to the training and validation series are properly predicted, while exception of some compounds that display high

Table 4
The best linear QSPR obtained from a pool of 1497 calculated descriptors. The selected model appears in bold.

d	R^2	S	R^2_{val}	S_{val}	R^2_{test}	S_{test}	Descriptors
1	0.491	0.28	0.401	0.29	0.468	0.27	*nHAcc*
2	0.607	0.25	0.545	0.26	0.514	0.26	*nHAcc Mor20e*
3	0.616	0.25	0.613	0.23	0.487	0.28	*Ms H3p MLOGP*
4	**0.659**	**0.24**	**0.655**	**0.22**	**0.551**	**0.26**	***BEHm*1 *RNCG nCOOH MLOGP***
5	0.714	0.22	0.567	0.28	0.399	0.33	*Mor*11*m Mor20e Mor26p E3e nHAcc*
6	0.738	0.21	0.543	0.31	0.497	0.30	*D/Dr08 GATS2e Mor20u nCOOR nHAcc C*-040

Table 5
Notation for molecular descriptors involved in QSPR model.

Type	Dim[a]	Molecular descriptor	Description
3D-MoRSE	3D	*Mor20e*	3D-MoRSE-signal 20/weighted by atomic Sanderson electronegativities
		Mor26p	3D-MoRSE-signal 26/weighted by atomic polarizabilities
		Mor11m	3D-MoRSE-signal 11/weighted by atomic masses
		Mor20u	3D-MoRSE-signal 20/unweighted
GETAWAY	3D	*H3p*	H autocorrelation of lag 3/weighted by atomic polarizabilities
		$R5e^+$	R maximal autocorrelation of lag 5/weighted by atomic Sanderson electronegativities
Constitutional	0D	*Ms*	Mean electrotopological state
Atom-centred fragments	1D	*H*-046	H attached to C0(sp^3) no X attached to next C
Functional groups	1D	*nHAcc*	Number of acceptor atoms for hydrogen bonds (N, O, F)
		nCOOH	Number of carboxylic acids (aliphatic)
Properties	1D	*MLOGP*	Moriguchi octanol–water partition coefficient
BCUT	3D	*BEHm*1	Highest eigenvalue no. 1 of Burden matrix/weighted by atomic masses
Charge	3D	*RNCG*	Relative negative charge
WHIM	3D	*E3e*	3rd component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities
Topological	2D	*D*/*Dr*08	Distance/Detour ring index of order 8
2D-Autocorrelations	2D	*GATS2e*	Geary autocorrelation – lag 2/weighted by atomic Sanderson electronegativities
		*ATS*3*m*	Broto-Moreau autocorrelation of a topological structure-lag 3/weighted by atomic masses
Atom-centred fragments	1D	*C*-040	R–C(=X)/R–C#X/X– = C = X[b]

[a] Dim: dimensionality.

[b] R represents any group linked through carbon; X represents any electronegative atom (O, N, S, P, Se, halogens).

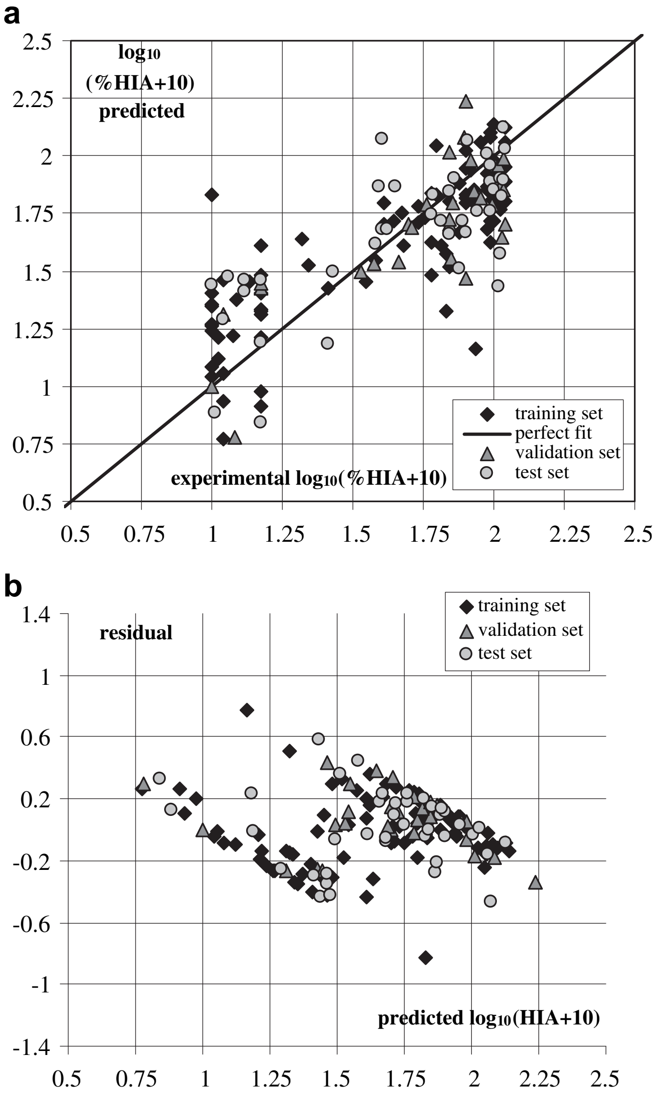

Fig. 1. a) Predicted (Eq. (5)) drug intestinal absorption as function of experimental values for the training, validation, and test sets. b) Residuals versus predicted permeabilities (Eq. (5)).

residuals in Fig. 1b are **38** (Pentamidine, residual = 0.83) and **80** (Cefatrizine, residual = 0.77). Both outliers have residuals exceeding the value 3.*S* logarithm units. However, as the purpose of present work is to derive a general model having applicability to any drug without restrictions, we decide not removing these molecules from the training set. It is also evident, however, from Fig. 1, that the linear model has a tendency to overestimate the % HIA of compounds with low intestinal absorption (see the number of points that concentrate over the regression line at low %HIA experimental values, in Fig. 1a). This is in good agreement with Reynolds' statement regarding how linear models tend to overestimate low %HIA, and how non-linear models can overcome this issue [28]. Clustering of the points in both scatter plots is also evident, suggesting two independent linear models might be developed for compounds with low and high %HIA. However, in that case, it will be difficult to systematically define *a priori* – when the hypothetic new models were applied for predicting the property – which of the predicted compounds correspond to low %HIA and which to high %HIA. An *a priori* classification scheme might be developed in the future (e.g. a discriminant function) to define whether a compound has high probability of having low or high permeability, so that two independent linear models (one for drugs with high %HIA and the other for drugs with low %HIA) are applied.

We verify that the linear QSPR established does not function only correlatively but would also function similarly well for the prediction of new permeability data not contemplated during the training stage of the model. The approval of the internal validation process of Eq. (5) is evidenced by the stability of the equation upon the inclusion/exclusion of compounds from the training set, measured via the commonly employed *loo* parameter (R_{loo} and S_{loo}). Furthermore, the external validation of Eq. (5) involves the prediction of 40 fresh test set structures, achieving a statistical quality (R_{test}, S_{test}) that is in line with that found for the training set. The predictions for these test set compounds (denoted with ^^ in Table 3) are able to provide insight on the variation of the experimental property for such "unknown structures", demonstrating the correct functioning of the QSPR. Finally, as a further step to assess the robustness of present equation, we applied the *y*-randomization procedure, demonstrating that the calibration does not result from happenstance and therefore, correspond to a valid structure–permeability relationship.

In QSPR-QSAR studies most molecular descriptors do not have a specific and direct (transparent) physical meaning, being defined as mathematical (abstract) quantities. They only represent a numerical feature (attribute) characterizing the chemical structure and obtained solely from it. Molecular descriptors are not numerical codes, in the sense that one cannot reconstruct the entire molecular structure from the single knowledge of the values of the descriptor. The main advantage of these numerical attributes is that they are able to combine in some optimal way in the linear regression, for predicting in the best possible way the property/activity being studied. The subset of descriptors found by the RM does not incorporate redundant structural information, as evidenced by the highest value of intercorrelation coefficient between descriptors of $R^2_{\max} = 0.084$, thus avoiding collineality among the four numerical variables, which may yield to highly unstable models. The four descriptors of Eq. (5) are: two three-dimensionals: BCUT: *BEHm*1, the highest eigenvalue no. 1 of Burden matrix/weighted by atomic masses; Charge: RNCG, the relative negative charge; and two one-dimensionals: Functional Groups: *n*COOH, number of carboxylic acids (aliphatic); Properties: MLOGP, Moriguchi octanol–water partition coefficient. The specific definition of these variables can be found elsewhere [40]. The numerical values of descriptors appearing in Eq. (5) are provided in Table 1S of Supplementary Material section.

The Burden matrix (**B**) corresponds to a modified Adjacency matrix (**A**) with the main diagonal elements being weighted with a given atomic property. The BCUT type of descriptors are obtained as positive and negative eigenvalues of **B**, in present case *BEHm*1 is the highest eigenvalue no. 1, and the atomic masses are employed as weights in **A**. *BEHm*1 does not have a direct/transparent interpretation as it is a purely mathematical quantity derived from the molecular graph. Charge-based descriptors are reliable only when quantum molecular methods are used for estimating charges, such as PM3 Semiempirical Method employed here. RNCG is the relative negative charge (ratio between maximum negative atomic charge and the molecular negative charge) [40].

Now, we investigate the role of the compounds that have carrier-mediated transport (adefovir, foscarnet, k-sthrophantoside, ouabain, amiloride, fosfomycin, methyldopa, captopril, cefatrizine, enalapril maleate, amoxicilline and warfarin) and paracellular transport (Lucifer yellow, mannitol) by removing them from the analysis, which leads to the following QSAR:

$$\begin{aligned}\log_{10}(\mathrm{HIA}+10) = {} & 1.823(\pm 0.05) - 0.00324(\pm 0.0008)\cdot \\ & \mathit{ATS3m} - 0.172(\pm 0.04)\cdot \mathit{nCOOH} \\ & + 0.111(\pm 0.01)\cdot \mathit{MLOGP} \qquad (6)\end{aligned}$$

$$\begin{aligned}& N = 83,\ d = 3,\ R^2 = 0.663,\ S = 0.24,\ F = 51.6,\ p \\ & < 10^{-4},\ R^2_{\max} = 0.071,\ R^2_{loo} = 0.615,\ S_{loo} = 0.25,\ S_{\mathrm{Rand}} \\ & = 0.36,\ N_{\mathrm{val}} = 29,\ R^2_{\mathrm{val}} = 0.605,\ S_{\mathrm{val}} = 0.24,\ N_{\mathrm{test}} \\ & = 34,\ R^2_{\mathrm{test}} = 0.643,\ S_{\mathrm{test}} = 0.22\end{aligned}$$

As can be appreciated, the statistical comparison between Eqs. (5) and (6) achieves close results, for the training, validation and test set, although Eq. (6) involves one less descriptor. Here, *ATS3m* is a structural variable introduced by Broto-Moreau [40] which corresponds to a bi-dimensional autocorrelation between atom pairs in the molecule, with the main purpose of capturing in the molecular descriptor the degree of interaction between such pairs. These indices are readily calculated from the graph, i.e.: by summing products of terms that include the atomic masses for the terminal atoms in all the paths of length (lag) 3. *ATS3m* has physical meaning, as it measures the degree of complexity of the chemical structure in terms of its degree of ramification and size. It is not possible to improve the quality of Eq. (6) by adding an additional variable, as its predictive performance deteriorates. We consider Eq. (5) to be more general and thus this is the QSAR we adopt as linear model.

As it is evident, the statistical quality of Eq. (5) should be improved further. In order to find out the non-linearity behavior and predictive ability of non-linear methods, we perform several neural networks. The network used in this study is constructed based on the four selected descriptors and consisted of three layers: input layer, hidden layer and output layer. Firstly, we train Back-Propagation models by resorting to two popular and powerful activity functions, namely, Levenberg-Marquardt (LM) and Bayesian Regulation (BR). The input values are auto-scaled and the initial weights are selected randomly between −0.3 and 0.3, and after that the parameters of the nodes in the hidden layer, weights and biases learning rates and momentum values are optimized based on minimum root mean square error (*RMSE*). Note that training of the network is stopped when overtraining begins by using the validation set [62,63].

On the other hand, with the purpose of exploring the predictive ability of another type of ANN model, we establish a Radial Basis Function Neural Network (RBFNN) by using the same split dataset. It should be noted that the predictive power of the RBFNN is also affected by various parameters such as the number of hidden neurons of radial basis functions, the center and width of each radial basis function, and the connection weight between both hidden layer and output units. The Forward Subset Selection method, which does not need to fix the number of hidden layer units, is used to determine the centers of RBFNN. The parameters are optimized based on minimum *RMSE* based on leave one out cross-validation [64,65], and the adjustment of the connection weight between the hidden and output layers is performed using a least squares solution [66].

A completely different approach we apply on this dataset is Support Vector Machine as a regression (SVMR) and based on a non-linear method. In fact, SVMR prediction ability is affected by kernel type (γ), capacity parameter (C) which is controlling the trade off, and the ε-insensitive loss function [67]. We first normalize the input values and then use 5-fold cross-validation for optimizing the parameters, selecting their optimal values based on maximum accuracy of the model and minimum *RMSE*. The kernel function used is the Gaussian radial basis function (RBF) kernel:

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2}\right) \qquad (7)$$

Table 6
Statistical parameters obtained for the various non-linear models.

Parameter	Set	BRANN	LMANN	RBFNN	SVMR
RMSE	Training	0.178	0.178	0.216	0.182
	Validation	0.166	0.178	0.185	–
	Test	0.208	0.225	0.235	0.189
RSEP (%)	Training	10.602	10.619	12.848	10.673
	Validation	9.334	10.053	10.407	–
	Test	12.168	13.172	13.753	11.037
MAE (%)	Training	3.679	3.741	4.279	3.187
	Validation	6.216	6.373	6.978	–
	Test	6.069	6.313	6.900	5.882
F-test	Training	348.694	342.887	208.724	396.899
	Validation	85.743	70.134	63.157	–
	Test	74.212	62.577	48.518	95.257
t-test	Training	18.673	18.517	14.447	19.922
	Validation	9.260	8.375	7.947	–
	Test	8.615	7.911	6.965	9.760
R^2	Training	0.639	0.796	0.702	0.771
	Validation	0.754	0.716	0.692	–
	Test	0.661	0.622	0.561	0.716
PRESS	Training	2.852	2.862	4.189	3.966
	Validation	0.823	0.954	1.023	–
	Test	1.734	2.032	2.215	1.427
SST	Training	13.921	13.921	13.921	17.289
	Validation	3.083	3.083	3.083	–
	Test	4.982	4.982	4.982	4.982
Ratio (*PRESS*/*SST*)	Training	0.205	0.206	0.301	0.229
	Validation	0.267	0.310	0.332	–
	Test	0.348	0.408	0.445	0.286

where $\mathbf{x_i}$ and $\mathbf{x_j}$ are input space vectors and σ^2 denotes the width of the Gaussian kernel.

The statistical parameters for all the calculated non-linear models are listed in Table 6 and, as can be appreciated, present results suggest that all these methods perform similarly well, with the BRANN technique performing the best for the training and validation sets. This conclusion is demonstrated by comparison of the *RMSE*, the relative standard error of prediction (*RSEP*), mean absolute error (*MAE*), Fischer test (*F*), *t*-test, predicted residual sum of squares in **Y** (*PRESS*), and sum of squares total (*SST*). Fig. 2a includes a graphical representation of the BRANN predictions as function of the experimental permeability values and Fig. 2b plots the residuals as function of the BRANN predictions. We have checked that the best linear and non-linear models accomplish the following validation criteria suggested by Golbraikh and coworkers to assure predictive capability [68,69]:

- R^2_{loo} above 0.6.
- R^2 above 0.6
- $R^2 - R^2_0/R^2 < 0.1$ and $0.85 \le k \le 1.15$
- $R^2 - R'^2_0/R^2 < 0.1$ and $0.85 \le k' \le 1.15$
- Probability of chance correlation very low, with none of the randomized models satisfying the previous four conditions.

where R^2_0 is the determination coefficient of the regression line of the predicted versus the observed values of the studied property when the intercept is forced to assume zero value, k is the slope of that regression line (predicted versus observed), R'^2_0 is the determination coefficient of the regression line of the observed versus predicted values of the property (with intercept zero) and k' is the slope of that regression line (observed versus predicted).

The applicability domain analysis indicates that 39 out of 40 compounds included in the test set belong to the applicability domain of the models; iohexol is the only compound whose leverage exceeds the warning leverage. The leverage for each compound and the warning leverage are presented in Table 7.

Table 7
Leverages of the 40 compounds used for external validation purposes; only one compound (in italic) lies outside the applicability domain.

Compound ID	h_i	Compound ID	h_i
4	0.05	**83**	0.07
8	0.03	**87**	0.01
12	0.04	**91**	0.02
16	0.11	**95**	0.02
20	***0.30***	**99**	0.02
24	0.06	**103**	0.01
28	0.02	**107**	0.03
32	0.07	**111**	0.01
36	0.03	**115**	0.06
40	0.11	**119**	0.02
44	0.02	**123**	0.03
47	0.03	**127**	0.02
51	0.12	**131**	0.02
55	0.04	**135**	0.06
59	0.04	**139**	0.02
63	0.05	**143**	0.03
67	0.01	**147**	0.04
71	0.03	**151**	0.03
75	0.02	**155**	0.02
79	0.04	**159**	0.02

Warning leverage h* = 0.13.

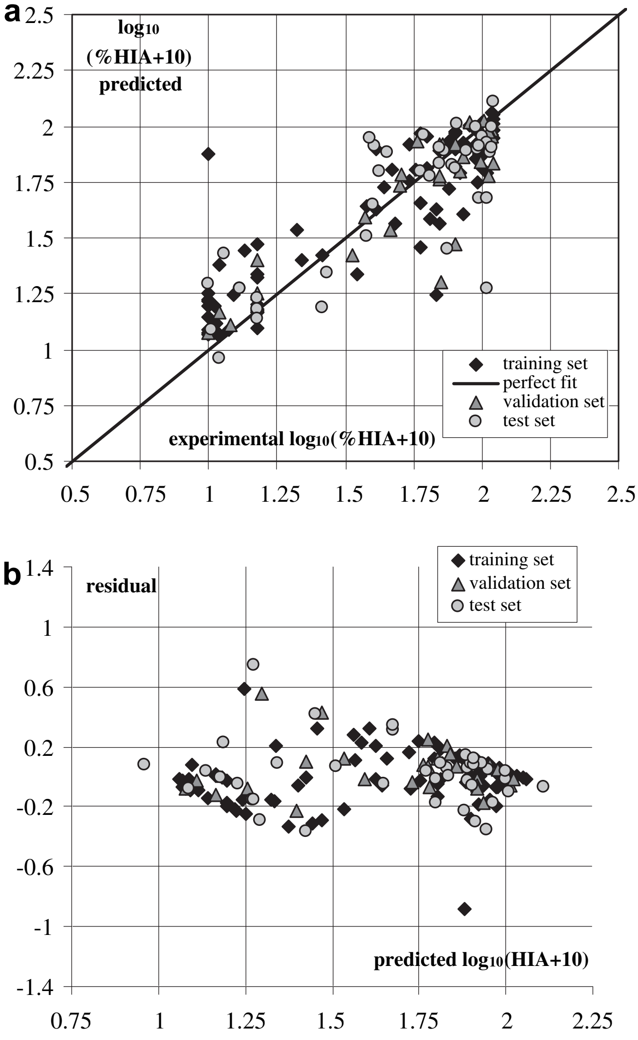

Fig. 2. a) BRANN predicted drug intestinal absorption as function of experimental values for the training, validation, and test sets. b) Residuals versus predicted permeabilities (BRANN).

4. Conclusions

Quantitative structure–permeability relationships are established in this study, by means of linear and non-linear methodologies. The novelty of this work relies on the use of a balanced distribution of low and high permeable organic compounds leading to a medium-size dataset, where the models established pose a high number of observations to descriptors ratio. The statistical results achieved in present analysis for this heterogeneous molecular set are satisfactory in the sense that these can be compared to previously reported ones of the literature. A main characteristic of the models established is that we do not discard compounds from the training set and include all of them during the QSPR analysis. As previously reported by many authors, the non-linear approaches outperform the linear approach in explaining the data variance (e.g. linear models tend to overestimate the %HIA for low permeable drugs; this bias is not so marked in the non-linear models). Clustering of the dataset in two evident groups is observed; future efforts should concentrate in generating an *a priori* classificatory scheme so that two independent models (one for low permeable compounds and the other for high permeable compounds) could be applied. As future tasks of our drug research and development program, we pretend to continue analyzing different ADME/Tox properties of drugs using different classes of computed theoretical structural descriptors combined with linear/non-linear methodologies.

Acknowledgements

We would like to thank to the National Council of Scientific and Technical Research of Argentina (Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET) and to University of La Plata for supporting this work. We are thankful to the reviewers for their thoughtful advices.

Appendix. Supplementary material

Supplementary data related to this article can be found online at doi:10.1016/j.ejmech.2010.11.005.

References

[1] A.M. Davis, R. Riley, Curr. Opin. Chem. 8 (2004) 378–386.
[2] D. Schuster, C. Laggner, T. Langer, Curr. Pharm. Des. 11 (2005) 3545–3559.
[3] N. Singh, R. Guha, M.A. Giulianotti, C. Pinila, R.A. Houghten, J.L. Medina-Franco, J. Chem. Inf. Model. 49 (2009) 1010–1024.
[4] K.C. Sounders, Drug Discov. Today Technol. 1 (2004) 373–380.
[5] B.L. Ackermann, M.J. Berna, A.T. Murphy, Curr. Top. Med. Chem. 2 (2002) 53–66.
[6] J. Kotecha, S. Shah, I. Rathod, G. Subbaiah, Int. J. Pharm. 333 (2007) 127–135.
[7] T.J. Carlson, M.B. Fisher, Comb. Chem. High Throughput Screen. 11 (2008) 258–264.
[8] C.E.C.A. Hop, M.J. Cole, R.E. Davidson, D.B. Duignan, J. Federico, J.S. Janiszewski, K. Jenkins, S. Krueger, R. Lebowitz, T.E. Liston, W. Mitchell, M. Snyder, S.J. Steyn, J.R. Soglia, C. Taylor, M.D. Troutman, J. Umland, M. West, K.M. Whalen, V. Zelesky, S.X. Zhao, Curr. Drug Metab. 9 (2008) 847–853.
[9] A. Lahoz, L. Gombau, M.T. Donato, J.V. Castell, M.J. Gómez-Lechón, Mini rev. Med. Chem. 6 (2006) 1053–1062.
[10] C.L. Stoner, M. Troutman, H. Gao, K. Johnson, C. Stankovic, J. Brodfuehrer, E. Gifford, M. Chang, Lett. Drug Des. Discov. 3 (2006) 575–581.
[11] C. Hansch, A. Leo, Exploring QSAR. Fundamentals and Applications in Chemistry and Biology. American Chemical Society, Washington, D.C, 1995.
[12] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. 24 (1995) 279–287.
[13] P.R. Duchowicz, A. Talevi, C. Bellera, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. 15 (2007) 3711–3719.
[14] P.R. Duchowicz, A. Talevi, L.E. Bruno-Blanch, E.A. Castro, Bioorg. Med. Chem. 16 (2008) 7944–7955.
[15] P.R. Duchowicz, E.A. Castro, Int. J. Mol. Sci. 10 (2009) 2558–2577.
[16] A. Talevi, E.A. Castro, L.E. Bruno-Blanch, J. Arg. Chem. Soc. 44 (2006) 129–141.
[17] M.D. Wessel, P.C. Jurs, J.W. Tolan, S.M. Muskal, J. Chem. Inf. Comput. Sci. 38 (1998) 726–735.
[18] U. Norinder, T. Osterberg, P. Artursson, Eur. J. Pharm. Sci. 8 (1999) 49–56.
[19] T. Österberg, U. Norinder, J. Chem. Inf. Comput. Sci. 40 (2000) 1408–1411.
[20] S. Agatonovic-Kustrin, R. Beresford, A. Pauzi, M. Yusof, J. Pharmaceut. Biomed. Anal. 25 (2001) 227–237.
[21] G. Klopman, L.R. Stefan, R.D. Saiakhov, Eur. J. Pharm. Sci. 17 (2002) 253–263.
[22] M.H. Abraham, Y.H. Zhao, J. Le, A. Hersey, C.N. Luscombe, D.P. Reynolds, G. Beck, B. Sherborne, I. Cooper, Eur. J. Med. Chem. 37 (2002) 595–605.
[23] T. Niwa, J. Chem. Inf. Comput. Sci. 43 (2003) 113–184.
[24] H. Sun, J. Chem. Inf. Comput. Sci. 44 (2004) 748–757.
[25] T.E. Yen, S. Agatonovic-Kustrin, A.M. Evans, R.L. Nation, J. Ryand, J. Pharm. Biomed. Anal. 38 (2005) 472–478.
[26] E. Deconinck, H. Ates, N. Callebaut, Y. Van Gyseghemb, Y. Vander Heyden, J. Cromatograph. A 1138 (2007) 190–202.
[27] A. Yan, Z. Wang, Z. Cai, Int. J. Mol. Sci. 9 (2008) 1961–1976.
[28] D.P. Reynolds, K. Lanevskij, P. Japertas, R. Didziapetris, A. Petrauskas, J. Pharm. Sci. 98 (2009) 4039–4054.
[29] A. Guerra, N.E. Campillo, J.A. Páez, Eur. J. Med. Chem. 46 (2010) 930–940.
[30] P.R.N. Wolohan, R.D. Clark, J. Comput. Aided Mol. Des 17 (2003) 65–76.
[31] M.A. Cabrera Pérez, M. Bermejo Sanz, L. Ramos Torres, R. Grau Ávalos, M. Pérez González, H. González Díaz, Eur. J. Med. Chem. 39 (2004) 905–916.
[32] E. Deconinck, T. Hancock, D. Cooman, D.L. Massart, Y. Vander Heyden, J. Pharmaceut. Biomed. Anal. 39 (2005) 91–103.
[33] T. Hou, J. Wang, Y. Li, J. Chem. Inf. Model. 47 (2007) 2408–2415.
[34] S. Yang, Q. Huang, L. Li, C. Ma, H. Zhang, R. Bai, Q. Teng, M. Xiang, Y. Wei, Artif. Intell. Med. 46 (2009) 155–163.
[35] Division of Specialized Information Services, National Library of Medicine, National Institute of Health, US Department of Health and Human Services. Hazardous Substances Data Bank. http://toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?HSDB.
[36] National Center for Biotechnology Information, National Library of Medicine, National Institute of Health, US Department of Health and Human Services. The Pubchem Project. http://pubchem.ncbi.nlm.nih.gov/.
[37] T. Hou, J. Wang, W. Zhang, X. Xu, J. Chem. Inf. Model. 47 (2007) 208–218.
[38] K. Liu, J. Feng, S. Stanley Young, J. Chem. Inf. Model. 45 (2005) 515–522.
[39] Hyperchem 7. Hypercube, Inc., Gainesville, 2007. http://www.hyper.com.
[40] Dragon Milano Chemometrics and QSAR Research Group, http://michem.disat.unimib.it/chm
[41] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics. Wiley-VCH, Weinheim, 2009.
[42] F. Harary, Graph Theory. Addison-Wesley, 1969.
[43] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, Chem. Phys. Lett. 412 (2005) 376–380.
[44] P.R. Duchowicz, E.A. Castro, F.M. Fernández, MATCH Commun. Math. Comput. Chem. 55 (2006) 179–192.
[45] P.R. Duchowicz, M.G. Vitale, E.A. Castro, M. Fernandez, J. Caballero, Bioorg. Med. Chem. 15 (2007) 2680–2689.
[46] M. Goodarzi, P.R. Duchowicz, C.H. Wu, F.M. Fernández, E.A. Castro, J. Chem. Inf. Model. 49 (2009) 1475–1485.
[47] Matlab 7.0, The MathWorks Inc.
[48] S. Haykin, Neural Networks: A Comprehensive Foundation. Macmillan College Publishing Company, New York, 1994.
[49] Y.H. Hu, J.-N. Hwang, Handbook of Neural Network Signal Processing. CRC Press LLC, Boca Raton, Florida, 2002.
[50] R.J. Schalkoff, Artificial Neural Networks. McGraw-Hill, New York, 1997.
[51] H. Demuth, M. Beale, Neural Network Toolbox for Use with Matlab. The Mathworks, Inc., Natick, 1998.
[52] K. Levenberg, Quart. Appl. Math. 2 (1944) 164–168.
[53] D. Marquardt, J. Soc. Ind. Appl. Math. 11 (1963) 431–441.
[54] T.J. Moody, C.J. Darken, Neural Comput. 1 (1989) 151–160.
[55] C. Darken, J. Moody, IEEE INNS International Joint Conference on Neural Networks, IEEE Press, New York (1990) 233–238.
[56] V.N. Vapnik, Statistical Learning Theory. Wiley, New York, 1998.
[57] V.N. Vapnik, The Nature of Statistical Learning Theory. Springer-Verlag, New York, USA, 2000.
[58] Y.P. Zhao, J.G. Sun, Expert Syst. Appl. (2010). doi:10.1016/j.eswa.2009.12.082.
[59] A. Tropsha, P. Gramatica, V. Gombar, QSAR Comb. Sci. 22 (2003) 69–77.
[60] P.N. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, J. Clin. Epidemiol. 49 (1996) 1373.
[61] P.N. Peduzzi, J. Concato, A.R. Feinstein, T.R. Holford, J. Clin. Epidemiol. 48 (1995) 1503–1510.
[62] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579–586.
[63] S. Wold, L. Eriksson, Statistical validation of QSAR results. in: H. Van de Waterbeemd (Ed.), Chemometrics Methods in Molecular Design. VCH, Weinheim, 1995, pp. 309–318.
[64] M. Goodarzi, M.P. Freitas, R. Jensen, J. Chem. Inf. Model. 49 (2009) 824–832.
[65] M.J.L. Orr, Introduction to Radial Basis Function Networks. Centre for Cognitive Science, Edinburgh University, Edinburgh, 1996.
[66] M.J.L. Orr, MATLAB Routines for Subset Selection and Ridge Regression in Linear Neural Networks. Centre for Cognitive Science, Edinburgh University, Edinburgh, 1996.
[67] X.J. Yao, A. Panaye, J.P. Doucet, H.F. Chen, R.S. Zhang, B.T. Fan, M.C. Liu, Z.D. Hu, Anal. Chim. Acta 535 (2005) 259–273.
[68] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–279.
[69] A. Golbraikh, M. Shen, Z. Xiao, Y.D. Xiao, K.H. Lee, A. Tropsha, J. Comput. Aided Mol. Des. 17 (2003) 241–253.