Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 7738–7745
Bioorganic & Medicinal Chemistry

Hologram QSAR model for the prediction of human oral bioavailability

Tiago L. Moda,[a] Carlos A. Montanari[b] and Adriano D. Andricopulo[a,*]

[a]*Laboratório de Química Medicinal e Computacional, Centro de Biotecnologia Molecular Estrutural, Instituto de Física de São Carlos, Universidade de São Paulo, 13566-970 São Carlos, SP, Brazil*
[b]*Grupo de Química Medicinal de Produtos Naturais, Instituto de Química de São Carlos, Universidade de São Paulo, 13566-970 São Carlos, SP, Brazil*

Received 2 June 2007; revised 22 August 2007; accepted 28 August 2007
Available online 1 September 2007

Abstract—A drug intended for use in humans should have an ideal balance of pharmacokinetics and safety, as well as potency and selectivity. Unfavorable pharmacokinetics can negatively affect the clinical development of many otherwise promising drug candidates. A variety of in silico ADME (absorption, distribution, metabolism, and excretion) models are receiving increased attention due to a better appreciation that pharmacokinetic properties should be considered in early phases of the drug discovery process. Human oral bioavailability is an important pharmacokinetic property, which is directly related to the amount of drug available in the systemic circulation to exert pharmacological and therapeutic effects. In the present work, hologram quantitative structure–activity relationships (HQSAR) were performed on a training set of 250 structurally diverse molecules with known human oral bioavailability. The most significant HQSAR model ($q^2 = 0.70$, $r^2 = 0.93$) was obtained using atoms, bond, connection, and chirality as fragment distinction. The predictive ability of the model was evaluated by an external test set containing 52 molecules not included in the training set, and the predicted values were in good agreement with the experimental values. The HQSAR model should be useful for the design of new drug candidates having increased bioavailability as well as in the process of chemical library design, virtual screening, and high-throughput screening.

1. Introduction

The challenges facing the pharmaceutical industry are tremendous at every step of the drug discovery and development process. The high number of compounds emerging from combinatorial chemistry and high-throughput medicinal chemistry programs is increasing the demand for new compounds that need to be screened in a wide range of biological assays.[1] The classical strategy to screen thousands of compounds solely for potency and selectivity brought the pharmaceutical industry to face the reality of disproportionate attrition in advanced stages of clinical development, because many attractive compounds do not possess the required pharmacokinetic properties of a drug. New chemical entities (NCEs) expected to advance into clinical trials should have an ideal balance of pharmacodynamic and pharmacokinetic properties.[2] Problems with absorption, distribution, metabolism, and excretion (ADME) have been identified as a major cause of drug candidate failure in late stages of the pharmaceutical R&D process.[3,4]

The early evaluation of ADME properties in drug research has driven the need for large-scale screening methods. In vitro and in vivo ADME assays (e.g., Caco-2, PAMPA, MDCK, assessment of absorption, metabolism studies) are lengthy, complex, and relatively expensive in terms of resources, reagents, and detection techniques.[5] Computational methods have emerged during the past decade as a powerful strategy for the prediction of human pharmacokinetics.[6] In this regard, a variety of useful in silico ADME models has been developed with different levels of complexity for the screening of large data sets of compounds, creating tools that are faster, simpler, and more cost-effective than traditional experimental procedures.[6–9]

Quantitative structure–activity relationships (QSAR) employing both classical and modern technologies have

Keywords: ADME; Pharmacokinetics; In silico modeling; QSAR; Drug design.
* Corresponding author. Tel.: +55 16 3373 8095; fax: +55 16 3373 9881; e-mail: aandrico@if.sc.usp.br

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.08.060

proven useful in a large number of settings.[10,11] QSAR is a technology that generates descriptors based on molecular structures and uses computational algorithms to relate the key descriptors to the dependent property value of interest.[12] It is therefore possible to match these unique characteristics to pharmacokinetic properties, mainly due to the accumulation of suitable ADME data for model generation.[13,14] Human oral bioavailability, which is defined as the fraction of an administered dose of drug that reaches the systemic circulation, is a critical property to be considered during the early stages of discovery. Several reports in the literature indicate that there is a significant scientific and practical need for new tools for early prediction of oral bioavailability as well as other important pharmacokinetic properties.[15,16] In the present work, we have organized a data set of 302 structurally diverse molecules with known human oral bioavailability, and used the data to create predictive 2D QSAR models, employing the hologram QSAR (HQSAR) method.[17–19] The results of modeling this data set are reported herein.

2. Materials and methods

2.1. Computational approach

The QSAR analyses, calculations, and visualizations for HQSAR were performed using the SYBYL 7.2 package (Tripos Inc., St. Louis, USA) running on Red Hat Enterprise Linux workstations. A statistical cluster analysis was carried out with Tsar 3D version 3.3 (Accelrys, San Diego, USA) employing the complete linkage clustering method with no standardization as previously described.[18,19]

2.2. Data set

The data set of 302 structurally diverse molecules used in the QSAR analyses was collected from the literature.[20–22] For convenience, the list of compounds along with the corresponding human oral bioavailability data are shown in Table 3. The 3D structures of the molecules employed in this work were constructed using CONCORD and standard geometric parameters available in the SYBYL 7.2 molecular modeling package.

2.3. Hologram QSAR

Statistical HQSAR modeling was carried out as previously described using the standard parameters implemented in SYBYL 7.2.[17–19,23,24] Briefly, HQSAR requires only 2D structures and the property value as input. In this method, each molecule in the training set is broken down into several unique structural fragments, which are then arranged to form a molecular hologram, an extended form of fingerprint that encodes all possible molecular fragments (e.g., linear, branched, and overlapping) and maintains a count of the number of occurrences of each fragment. Incorporation of information about each fragment, and each of its constituent subfragments, implicitly encodes 3D structural information, such as hybridization and chirality.[17,19,24] With the transformation of the chemical representation of a molecule into its corresponding molecular hologram, this method requires no explicit 3D information (e.g., 3D structures, putative binding conformations, and molecular structural alignment).[17–19,25] HQSAR models can be affected by a number of parameters concerning hologram generation: hologram length, fragment size, and fragment distinction. Several combinations of fragment distinction were considered during the QSAR modeling runs. HQSAR analyses were performed by screening the 12 default series of hologram lengths. The influence of different fragment sizes was also investigated. The patterns of fragment counts from the training set molecules were then related to the experimental oral bioavailability data using full cross-validated r^2 (q^2) partial least squares (PLS) leave-one-out (LOO), and leave-many-out methods.[17–19,22,23] The predictive ability of the models was assessed by their q^2 values.

3. Results and discussion

3.1. Data set characterization

The data set of 302 compounds employed in this work is structurally diverse, containing enzyme inhibitors, receptor agonists and antagonists as well as other biologically active agents of several important therapeutic classes, including antibiotics, analgesics, antivirals, anticancers, antibacterials, antifungals, antidepressants, antiepileptics, antihypertensives, anti-inflammatories, antiparasitics, anxiolytics, antipsychotics, antispasmodics, among others. An example of the high chemical and pharmacological diversity of the data set is shown in Figure 1. Human oral bioavailability is represented as the percentage of an administered dose of a drug that reaches the systemic circulation after oral administration. When two or more values of oral bioavailability for the same compound were available or a value had a range, the average value was employed. In cases where large differences (>20%) between two or more values were found, the compounds were not incorporated in the data set. Training set compounds containing one asymmetric (chiral) center, for which the corresponding oral bioavailability was determined for the racemate, were considered as the individual enantiomers and modeled accordingly.

The distribution of the human oral bioavailability data for the complete data set is presented in form of a histogram in Figure 2. Although weighted toward the high-bioavailability end of the spectrum, values of bioavailability (%) are acceptably distributed across the range of values. Basically, three distinct groups can be seen in Figure 2, indicated as 'low' (≤40%, 74 compounds), 'intermediate' (41–80%, 127 compounds) and 'high' (>80%, 101 compounds) bioavailability values. These compounds were grouped in such a way to provide reasonable ranges for comparison reasons.

The selection of appropriate training and test sets is of critical importance in the generation of robust QSAR models. A hierarchical cluster analysis was carried out

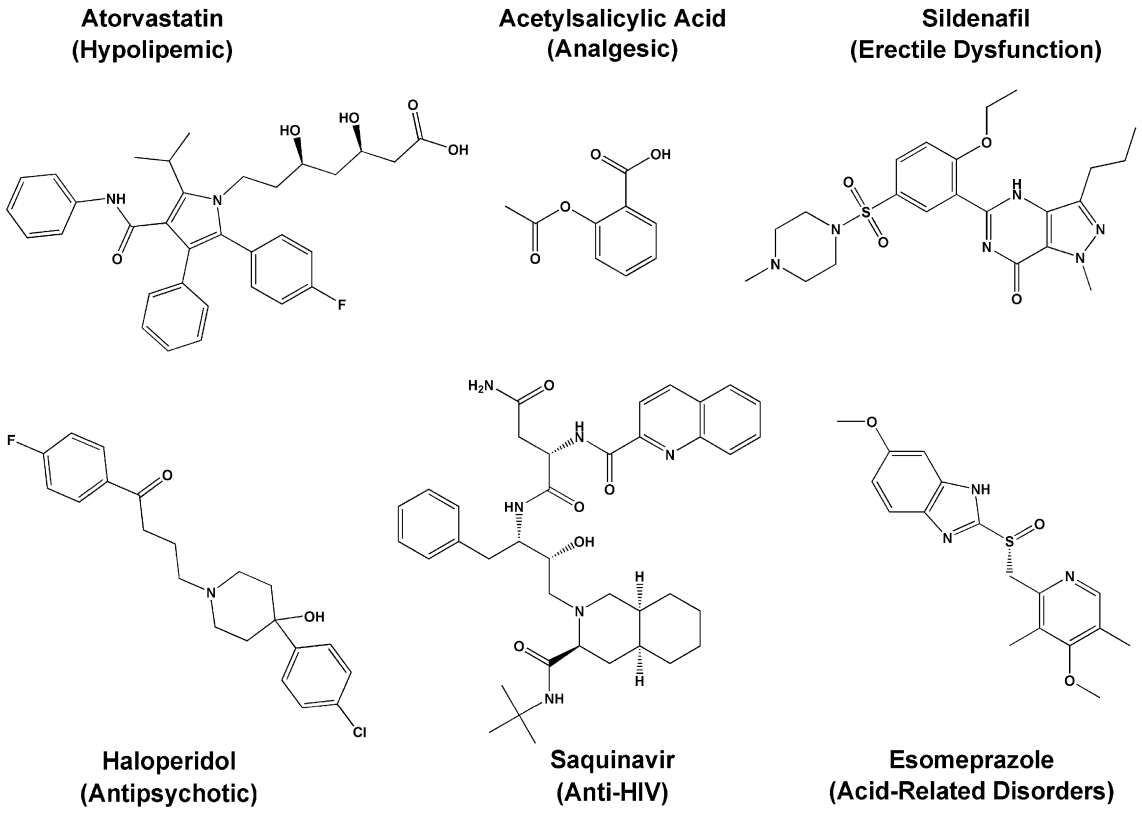

Figure 1. Chemical structure and therapeutic class of representative drugs of the data set.

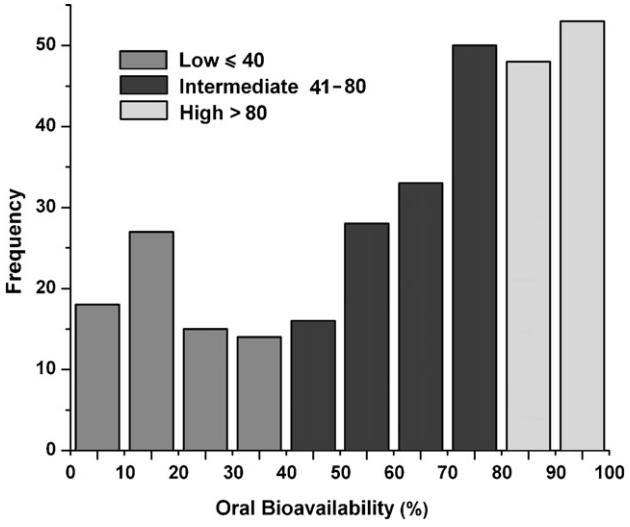

Figure 2. Histogram representation of the distribution of human oral bioavailability for the 302 data set compounds.

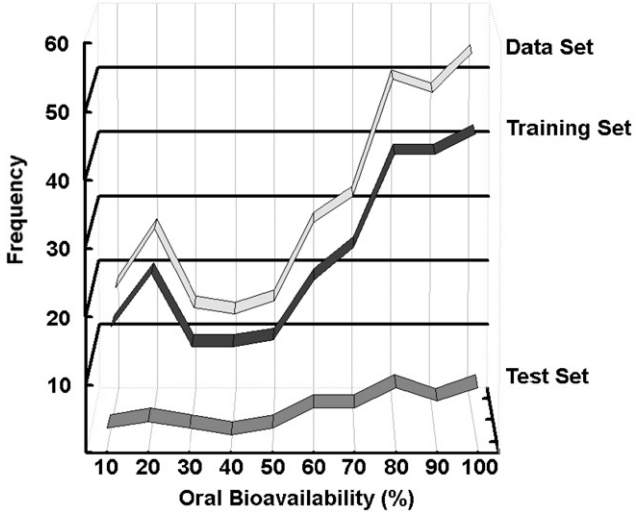

Figure 3. Data set, training set, and test set composition.

with the assistance of the Tsar 3D program, as depicted in Figure 3. In this study, the original data set of 302 compounds was arranged into training (**001**–**250**, Table 3) and test (**251**–**302**, Table 3) sets such that structurally diverse molecules possessing property values of a wide range were included in both sets. For example, the 302 data set drugs of different therapeutic classes were included in the final training and test sets, respectively, as follows (shown in percentages): analgesics (9.4 and 7.8); antibacterials (9.9 and 13.5); anticancers (5.6 and 2.0); antidepressants (4.2 and 1.5); antiemetics (3.3 and 2.0); antihypertensives (11.8 and 15.6); anti-inflammatories (6.1 and 7.8); antilipemics (2.0 and 5.9); contraceptives (1.9 and 1.9); immunosuppressives (1.9 and 1.9); and others (43.9 and 40.1). Thus, the data set is suitable for HQSAR model development in terms of structural diversity, distribution of drug classes and property values. The training set was then used to generate the HQSAR models, while the test set was hold out for the process of model external validation.

3.2. HQSAR modeling

HQSAR investigations require selecting values for parameters that specify hologram length, as well as the size and type of fragments that are to be encoded.[17,19] Holograms were created using the standard parameters implemented in SYBYL 7.2. The generation of molecular fragments was carried out using the following fragment distinctions: atoms (A), bonds (B), connections (C), hydrogen atoms (H), chirality (Ch), and donor and acceptor atoms (DA). Several combinations of these parameters were considered using the fragment size default (4–7), as follows: A/B, A/B/C, A/B/C/H, A/B/C/H/Ch, A/B/H, A/B/C/Ch, A/B/C/Ch/DA, A/B/DA, A/B/C/DA, A/B/H/DA, A/B/C/H/Ch/DA, A/B/C/H/DA, and A/B/Ch/DA. HQSAR analysis was performed over the 12 default series of hologram lengths of 53, 59, 61, 71, 83, 97, 151, 199, 257, 307, 353, and 401 bins, using the fragment size default (4–7). The statistical results from the PLS analyses for the 250 training set compounds using the several fragment distinction combinations are summarized in Table 1.

The best statistical results were obtained using either A/B/C ($q^2 = 0.52$, $r^2 = 0.89$) or A/B/C/Ch ($q^2 = 0.53$, $r^2 = 0.90$) as distinction information (models 2 or 6, respectively). The influence of different fragment sizes on the statistical parameters was further investigated for the HQSAR generated using A/B/C/Ch. Fragment size parameters control the minimum and maximum lengths of fragments to be included in the hologram fingerprint, therefore, a fundamental aspect is to be considered regarding the statistical generation of QSAR models.[17–19] These parameters can be adjusted to include larger or smaller fragments in the molecular holograms. As mentioned previously, the fragment size default (4–7), which has proved useful in a number of situations, was first used to derive the HQSAR models. The statistical results for the sequence of different fragment sizes evaluated (including, 2–5, 3–6, 4–7, 5–8, 6–9, 7–10, 8–11, and 9–12) are summarized in Table 2.

Table 1. Results of HQSAR analyses for various fragment distinctions on the key statistical parameters using the fragment size default (4–7)

Model	Fragment distinction	Statistical parameters			
		q^2	r^2	N	HL
1	A/B	0.43	0.60	5	53
2	A/B/C	0.52	0.89	8	353
3	A/B/C/H	0.50	0.81	8	257
4	A/B/C/H/Ch	0.48	0.79	7	257
5	A/B/H	0.46	0.75	7	401
6	A/B/C/Ch	0.53	0.90	8	353
7	A/B/C/Ch/DA	0.47	0.87	7	353
8	A/B/DA	0.41	0.73	8	353
9	A/B/C/DA	0.52	0.81	6	307
10	A/B/H/DA	0.38	0.70	7	401
11	A/B/C/H/Ch/DA	0.40	0.75	7	353
12	A/B/C/H/DA	0.44	0.76	6	353
13	A/B/Ch/DA	0.42	0.69	6	401

q^2, cross-validated correlation coefficient. r^2, noncross-validated correlation coefficient. SEE, noncross-validated standard error. N, optimal number of components. HL, hologram length.

Table 2. HQSAR analysis for the influence of various fragment sizes on the key statistical parameters using the fragment distinction A/B/C/Ch

Fragment size	Statistical parameters				
	q^2	r^2	SEE	N	HL
2–5	0.42	0.70	16.11	6	307
3–6	0.47	0.81	12.81	6	353
4–7	0.53	0.90	9.14	8	353
5–8	0.42	0.82	12.37	6	353
6–9	0.47	0.81	12.60	7	151
7–10	0.70	0.93	7.60	8	199
8–11	0.44	0.83	11.95	7	199
9–12	0.35	0.64	17.47	4	199

q^2, cross-validated correlation coefficient. r^2, noncross-validated correlation coefficient. SEE, noncross-validated standard error. N, optimal number of components. HL, hologram length.

The variation of fragment size led to the generation of a substantially better HQSAR model compared to that derived for the fragment size default (4–7). As it can be seen, the best statistical results ($q^2 = 0.70$, $r^2 = 0.93$) among all models were obtained with the fragment size 7–10. A similar procedure was applied for the evaluation of the fragment distinction A/B/C of model 2 (results not shown), but the results were not any better than those described for the HQSAR model 6. The stability of the best model (fragment distinction A/B/C/Ch, fragment size 7–10) was tested using 10 different data set splits, varying the relative distribution of the groups low—med—high. Values of q^2 between 0.66 and 0.70 were obtained indicating the high stability of the final model (results not shown).

In terms of validation of a QSAR model, a measure of internal consistency is available in the form of q^2 (Tables 1 and 2). However, the most important test of a QSAR model is its ability to predict the property value for new compounds. As the structure encoded in a 2D fingerprint is directly related to the human oral bioavailability of the training set molecules, HQSAR models should be able to predict the bioavailability of new structurally diverse compounds from its fingerprint. Thus, the predictive power of the best HQSAR model derived using the 250 training set molecules (fragment distinction: A/B/C/Ch; fragment size: 7–10, Table 2) was assessed by predicting bioavailability values for 52 test set molecules, which were not included in the training set for model generation. What is needed is a measure of how close the actual data is to the predicted data of an external data set. The results obtained are presented in Table 3, while the graphic representation for the experimental versus predicted oral bioavailability values of both training (model generation) and test (external evaluation) sets are displayed in Figure 4.

The results show that the test set compounds, which represent the different structural features incorporated within the training set, are reasonably well predicted. The good agreement between experimental and predicted bioavailability values indicates the robustness of the HQSAR model. The predictive power ($r^2_{\text{pred}} = 0.85$) of the model generated is remarkable given the inherent complexity of the target property and the large chemical universe of

Table 3. Experimental, predicted, and residual values of human oral bioavailability for both training and test set compounds for the final HQSAR model

No.	Compound	Exp[a]	Pred[b]	Res[c]	No.	Compound	Exp[a]	Pred[b]	Res[c]
Training set									
001	Abacavir[22]	86.00	79.66	6.34	**002**	Acetaminophen[22]	88.00	75.20	12.80
003	Acetylsalicylic acid[22]	68.00	74.15	−6.15	**004**	Albuterol[22] (S)	71.00	70.88	0.12
005	Almotriptan[21]	70.00	69.11	0.89	**006**	Alosetron[21]	55.00	64.95	−9.95
007	Anastrozole[22]	84.00	86.12	−2.12	**008**	Aprepitant[21]	62.50	74.29	−11.79
009	Atomoxetine[21]	63.00	77.19	−14.19	**010**	Atorvastatin[21]	14.00	19.60	−5.60
011	Atovaquone[21]	23.00	17.38	5.62	**012**	Bepridil[23] (R)	60.00	63.56	−3.56
013	Bosentan[21]	50.00	47.62	2.38	**014**	Bromocriptine[22]	4.50	−2.28	6.78
015	Budesonide[23]	11.00	21.23	−10.23	**016**	Bufuralol[23] (R)	46.00	61.48	−15.48
017	Bufuralol[23] (S)	46.00	58.41	−12.41	**018**	Bumetanide[22]	81.00	72.65	8.35
019	Bupropion[23] (R)	70.00	76.94	−6.94	**020**	Bupropion[23] (S)	70.00	64.66	5.34
021	Busulfan[21]	80.00	71.51	8.49	**022**	Calcitriol[22]	61.00	61.44	−0.44
023	Candesartan[21] (R)	15.00	2.45	12.55	**024**	Candesartan[21] (S)	15.00	9.60	5.40
025	Carbamazepine[22]	70.00	69.62	0.38	**026**	Cefaclor[23]	90.00	88.57	1.43
027	Cefazolin[22]	90.00	81.51	8.49	**028**	Cephalexin[22]	90.00	96.70	−6.70
029	Chlorambucil[22]	87.00	78.84	8.16	**030**	Chloramphenicol[23]	69.00	79.49	−10.49
031	Chloramphenicol palmitate[23]	80.00	75.36	4.64	**032**	Chloroquine[22] (S)	80.00	79.12	0.88
033	Chlorpromazine[22]	32.00	24.71	7.29	**034**	Chlortetracycline[23]	27.50	19.59	7.91
035	Chlorthalidone[22] (R)	64.00	62.22	1.78	**036**	Cicloprolol[23] (R)	100.00	92.67	7.33
037	Cicloprolol[23] (S)	100.00	95.18	4.82	**038**	Cimetropium bromide[23]	2.00	22.22	−20.22
039	Ciprofloxacin[21]	70.00	83.34	−13.34	**040**	Citalopram[22] (R)	80.00	87.92	−7.92
041	Citalopram[22] (S)	80.00	68.42	11.58	**042**	Clonidine[22]	60.00	74.50	−14.50
043	Clorazepate[22] (R)	91.00	91.27	−0.27	**044**	Clorazepate[22] (S)	91.00	92.96	−1.96
045	Chloroquine[22] (R)	80.00	77.87	2.13	**046**	Clozapine[22]	55.00	65.38	−10.38
047	Cyclophosphamide[22]	74.00	72.82	1.18	**048**	Dapsone[23]	93.00	90.53	2.47
049	Delavirdine[21]	96.00	92.66	3.34	**050**	Diazepam[22]	100.00	87.63	12.37
051	Diflunisal[23]	100.00	90.31	9.69	**052**	Digitoxin[23]	88.50	86.82	1.68
053	Digoxin[21]	95.00	90.23	4.77	**054**	Dihydroergosine[23]	10.00	6.35	3.65
055	Diltiazem[21]	40.00	47.61	−7.61	**056**	Diphenhydramine[22]	72.00	81.68	−9.68
057	Disopiramide[22] (R)	83.00	70.21	12.79	**058**	Disopiramide[22] (S)	83.00	95.43	−12.43
059	Dixyrazine[23] (S)	10.00	5.69	4.31	**060**	Domperidone[23]	14.00	14.34	−0.34
061	Doxapram[23] (R)	61.00	63.03	−2.03	**062**	Doxapram[23] (S)	61.00	45.34	15.66
063	Doxazosin[23]	65.00	62.11	2.89	**064**	Doxycycline[22]	93.00	90.60	2.40
065	Drospirenone[21]	76.00	76.08	−0.08	**066**	Dutasteride[21]	60.00	73.93	−13.93
067	Eletriptan[21]	62.50	62.38	0.12	**068**	Emtricitabine[21]	93.00	82.78	10.22
069	Enalapril[22]	41.00	45.59	−4.59	**070**	Endralazine[23]	75.00	78.38	−3.38
071	Eproxindine[23]	70.00	66.72	3.28	**072**	Erythromycin[22]	35.00	40.74	−5.74
073	Estradiol valerate[23]	3.00	13.19	−10.19	**074**	Ethambutol[22]	77.00	82.66	−5.66
075	Ethinyl estradiol[21]	40.00	38.26	1.74	**076**	Esomeprazole[21]	90.00	96.50	−6.50
077	Famciclovir[21]	77.00	70.43	6.57	**078**	Felodipine[22] (R)	15.00	6.13	8.87
079	Felodipine[22] (S)	15.00	4.82	10.18	**080**	Fenflumizole[23]	50.00	49.62	0.38
081	Fenfluramine[23] (R)	89.00	83.38	5.62	**082**	Fenoprofen[23] (R)	80.00	89.82	−9.82
083	Flecainide[23] (R)	95.00	82.15	12.85	**084**	Flecainide[23] (S)	95.00	83.34	11.66
085	Flucloxacillin[23]	49.00	38.61	10.39	**086**	Flucytosine[21]	83.50	81.52	1.98
087	Flunisolide[23]	20.00	23.63	−3.63	**088**	Fluocortolone[23]	83.50	77.39	6.11
089	Flurbiprofen[22] (R)	92.00	85.47	6.53	**090**	Fluticasone[21]	1.00	0.06	0.94
091	Gatifloxacin[22] (R)	96.00	98.47	−2.47	**092**	Gefitinib[21]	60.00	61.81	−1.81
093	Gemifloxacin[21] (S)	71.00	72.86	−1.86	**094**	Glipizide[22]	95.00	92.13	2.87
095	Glyburide[22]	95.00	96.27	−1.27	**096**	Granisetron[22]	60.00	69.74	−9.74
097	Haloperidol[22]	60.00	59.64	0.36	**098**	Hydrochlorothiazide[22]	71.00	84.86	−13.86
099	Hydromorphone[21]	24.00	21.00	3.00	**100**	Ibuprofen[22] (S)	80.00	72.21	7.79
101	Idarubicin[22]	28.00	34.18	−6.18	**102**	Imipramine[22]	42.00	53.33	−11.33
103	Irbesartan[21]	70.00	77.05	−7.05	**104**	Isradipine[22] (R)	19.50	14.26	5.24
105	Itraconazole[21]	55.00	54.72	0.28	**106**	Ketoprofen[23] (S)	85.00	89.63	−4.63
107	Lamivudine[22]	86.00	86.69	−0.69	**108**	Lansoprazole[22] (R)	80.00	83.57	−3.57
109	Lansoprazole[22] (S)	80.00	83.57	−3.57	**110**	Letrozole[22]	99.90	98.34	1.56
111	LAAM[22]	47.00	50.76	−3.76	**112**	Levonorgestrel[22]	87.00	93.83	−6.83
113	Linezolid[21]	100.00	95.28	4.72	**114**	Lormetazepam[23] (R)	75.00	85.19	−10.19
115	Losartan[21]	33.00	42.71	−9.71	**116**	Lovastatin[22]	5.00	15.83	−10.83
117	Melphalan[22] (R)	71.00	72.46	−1.46	**118**	Melphalan[22] (S)	71.00	64.79	6.21
119	Mepindolol[23] (R)	82.00	82.63	−0.63	**120**	Mepindolol[23] (S)	82.00	84.74	−2.74
121	Metergoline[23]	23.00	23.30	−0.30	**122**	Methacycline[23]	58.00	60.17	−2.17
123	Methadone[22] (R)	92.00	84.31	7.69	**124**	Methadone[22] (S)	92.00	88.91	3.09
125	Methimazole[23]	93.00	79.60	13.40	**126**	Methotrexate[22] (R)	70.00	64.41	5.59

Table 3 (*continued*)

No.	Compound	Exp[a]	Pred[b]	Res[c]	No.	Compound	Exp[a]	Pred[b]	Res[c]
127	Methylphenobarbital[23] (R)	73.00	74.12	−1.12	**128**	Methylphenobarbital[23] (S)	73.00	75.48	−2.48
129	Methylprednisolone[22]	82.00	82.15	−0.15	**130**	Methysergide[23]	13.00	8.24	4.76
131	Metoclopramide[22]	76.00	84.05	−8.05	**132**	Metopimazine[23]	19.00	20.14	−1.14
133	Mexiletine[22] (R)	87.00	79.35	7.65	**134**	Mianserin[23] (S)	20.00	28.41	−8.41
135	Midalcipran[23]	84.00	79.19	4.81	**136**	Midodrine[21] (R)	93.00	81.10	11.90
137	Midodrine[21] (S)	93.00	80.82	12.18	**138**	Milrinone[23]	92.00	80.02	11.98
139	Minocycline[22]	97.50	99.46	−1.96	**140**	Moexipril[21]	13.00	10.74	2.26
141	Montelukast[21]	73.00	62.17	10.83	**142**	Morphine[21]	40.00	47.75	−7.75
143	Moxifloxacin[21]	90.00	93.29	−3.29	**144**	Mycophenolate[21]	72.00	74.82	−2.82
145	Nadolol[23] (R)	34.00	45.13	−11.13	**146**	Nadolol[23] (S)	34.00	47.68	−13.68
147	Nalbuphine[22]	11.00	4.75	6.25	**148**	Naloxone[22]	2.00	3.53	−1.53
149	Naltrexone[23]	20.00	24.33	−4.33	**150**	Naratriptan[22]	67.50	53.78	13.72
151	Nevirapine[21]	93.00	86.92	6.08	**152**	Nifedipine[22]	50.00	58.17	−8.17
153	Nimodipine[21] (S)	13.00	26.35	−13.35	**154**	Nitrendipine[23] (S)	16.00	30.91	−14.91
155	Nitrofurantoin[22]	87.00	81.34	5.66	**156**	Nizatidine[21]	70.00	77.26	−7.26
157	Nomifensine[23] (S)	27.00	33.66	−6.66	**158**	Norfenfluramine[23] (R)	85.00	84.41	0.59
159	Norfenfluramine[23] (S)	85.00	84.02	0.98	**160**	Norzimelidine[23]	66.00	77.27	−11.27
161	Ofloxacin[22] (R)	97.50	95.15	2.35	**162**	Ofloxacin[22] (S)	97.50	96.90	0.60
163	Olmesartan[21]	26.00	27.64	−1.64	**164**	Ondansetron[22] (R)	62.00	63.33	−1.33
165	Ondansetron[22] (S)	62.00	64.69	−2.69	**166**	Oseltamivir[21]	75.00	83.14	−8.14
167	Oxytetracycline[23]	58.00	60.40	−2.40	**168**	Pantoprazole[21] (R)	77.00	74.51	2.49
169	Pantoprazole[21] (S)	77.00	74.51	2.49	**170**	Penicillin G[23]	22.50	31.35	−8.85
171	Pentazocine[23]	18.00	23.28	−5.28	**172**	Phencyclidine[23]	72.00	61.14	10.86
173	Phenobarbital[23]	96.00	92.26	3.74	**174**	Phenylethylmalonamide[23]	91.00	91.64	−0.64
175	Phenytoin[22]	90.00	92.95	−2.95	**176**	Physostigmine[23]	6.00	15.33	−9.33
177	Pimozide[22]	50.00	39.30	10.70	**178**	Pinacidil[23] (R)	57.00	71.58	−14.58
179	Pinacidil[23] (S)	57.00	69.34	−12.34	**180**	Pipotiazine[23]	26.00	23.90	2.10
181	Piroxicam[23]	100.00	96.44	3.56	**182**	Pravastatin[21]	17.00	1.88	15.12
183	Prednisolone phosphate[22]	82.00	83.75	−1.75	**184**	Prednisone[22]	80.00	66.84	13.16
185	Primaquine[23] (R)	96.00	84.76	11.24	**186**	Primaquine[23] (S)	96.00	88.22	7.78
187	Primidone[23]	92.00	96.90	−4.90	**188**	Procyclidine[23] (R)	75.00	77.83	−2.83
189	Procyclidine[23](S)	75.00	75.21	−0.21	**190**	Promethazine[23] (S)	25.00	30.67	−5.67
191	Propiomazine[23] (R)	33.00	36.15	−3.15	**192**	Propylthiouracil[23]	78.00	75.20	2.80
193	Proscillaridin[23]	7.00	6.59	0.41	**194**	Protriptyline[23]	85.00	84.93	0.07
195	Proxyphylline[23] (R)	100.00	92.14	7.86	**196**	Proxyphylline[23] (S)	100.00	94.69	5.31
197	Quinidine[22]	75.00	63.57	11.43	**198**	Raloxifene[21]	2.00	−5.10	7.10
199	Repaglinide[21]	56.00	57.45	−1.45	**200**	Ribavirin[21]	64.00	65.56	−1.56
201	Rifampin[21]	88.80	87.38	1.42	**202**	Rizatriptan[21]	45.00	57.70	−12.70
203	Rofecoxib[21]	93.00	94.59	−1.59	**204**	Saccharin[23]	84.00	82.37	1.63
205	Saquinavir[21]	4.00	−0.83	4.83	**206**	Sildenafil[21]	40.00	34.79	5.21
207	Sirolimus[21]	14.00	6.18	7.82	**208**	Sobrerol[23] (trans)	72.00	69.4	2.60
209	Sotalol[23] (R)	60.00	62.28	−2.28	**210**	Sotalol[23] (S)	60.00	58.41	1.59
211	Stavudine[22]	82.00	79.30	2.70	**212**	Suprofen[23] (R)	92.00	80.85	11.15
213	Suprofen[23] (S)	92.00	78.06	13.94	**214**	Tacrolimus[21]	17.00	15.47	1.53
215	Telenzepine[23]	54.00	51.91	2.09	**216**	Telithromycin[21]	57.00	48.81	8.19
217	Testosterone[23]	7.00	0.67	6.33	**218**	Theophylline[21]	98.80	86.12	12.68
219	Tiagabine[21]	90.00	88.68	1.32	**220**	Tiapamil[23]	22.00	23.94	−1.94
221	Tocainide[22] (R)	89.00	79.35	9.65	**222**	Tocainide[22](S)	89.00	79.16	9.84
223	Tolbutamide[22]	85.00	85.90	−0.90	**224**	Toliprolol[23] (S)	90.00	80.19	9.81
225	Topiramate[22]	70.00	71.56	−1.56	**226**	Topotecan[22]	32.00	32.55	−0.55
227	Torasemide[23]	91.00	89.60	1.40	**228**	Tramadol[21]	75.00	70.63	4.37
229	Trandolapril[21]	10.00	7.92	2.08	**230**	Triamterene[22]	51.00	59.44	−8.44
231	Triazolam[22]	44.00	44.55	−0.55	**232**	Trimethoprim[22]	63.00	70.08	−7.08
233	Trospium[21]	9.60	6.38	3.22	**234**	Valganciclovir[21] (R)	59.40	75.30	−15.90
235	Valganciclovir[21] (S)	59.40	71.01	−11.61	**236**	Valproic acid[21]	90.00	78.64	11.36
237	Venalafaxine[22] (S)	45.00	42.54	2.46	**238**	Verapamil[21] (S)	13.50	16.39	−2.89
239	Verdenafil[21]	15.00	15.07	−0.07	**240**	Viloxazine[23] (R)	85.00	83.58	1.42
241	Viloxazine[23] (S)	85.00	82.07	2.93	**242**	Warfarin[22] (R)	93.00	93.35	−0.35
243	Warfarin[22] (S)	93.00	90.07	2.93	**244**	Zaleplon[21]	30.00	38.45	−8.45
245	Zidovudine[22]	64.00	75.48	−11.48	**246**	Ziprasidone[21]	60.00	74.32	−14.32
247	Zolmitriptan[21]	40.00	54.00	−14.00	**248**	Zolpiclone[23] (R)	80.00	79.36	0.64
249	Zolpiclone[23] (S)	80.00	90.18	−10.18	**250**	Zolpidem[22]	72.00	73.19	−1.19
Test set									
251	Amiodarone[22]	50.00	55.77	−5.77	**252**	Bepridil[23] (S)	60.00	49.36	10.64
253	Caffeine[23]	100.00	94.56	5.44	**254**	Chlorthalidone[22] (S)	64.00	48.81	15.19

(*continued on next page*)

Table 3 (*continued*)

No.	Compound	Exp[a]	Pred[b]	Res[c]	No.	Compound	Exp[a]	Pred[b]	Res[c]
255	Cimetidine[22]	60.00	77.71	−17.71	**256**	Clarithromycin[21]	50.00	44.82	5.18
257	Clavulanate[22]	75.00	63.22	11.78	**258**	Clofibrate[22]	95.00	79.76	15.24
259	Clonazepam[21]	90.00	79.95	10.05	**260**	Cloxacillin[23]	37.00	35.60	1.40
261	Codeine[23]	55.00	45.49	9.51	**262**	Dixyrazine[23] (R)	10.00	12.33	−2.33
263	Dronabinol[21]	15.00	5.68	9.32	**264**	Escitalopram[21] (S)	80.00	68.42	11.58
265	Fenfluramine[23] (S)	89.00	81.39	7.61	**266**	Fenoprofen[23] (S)	80.00	88.32	−8.32
267	Fenoximone[23]	53.00	63.32	−10.32	**268**	Flecainide[22] (R)	70.00	82.15	−12.15
269	Fluphenazine[22]	2.70	19.42	−16.72	**270**	Flurbiprofen[22] (S)	92.00	78.71	13.29
271	Fluvastatin[21]	24.00	30.79	−6.79	**272**	Gabapentin[21]	60.00	73.38	−13.38
273	Gatifloxacin[22] (S)	96.00	104.69	−8.69	**274**	Gemifloxacin[21]	71.00	66.47	4.53
275	Glimepiride[22]	100.00	107.33	−7.33	**276**	Ibuprofen[22] (R)	80.00	70.43	9.57
277	Isradipine[22] (S)	19.50	10.74	8.76	**278**	Ketoprofen[23] (R)	85.00	96.79	−11.79
279	Levofloxacin[22]	99.00	96.89	2.11	**280**	Lorazepam[22] (R)	93.00	83.31	9.69
281	Lorazepam[22] (S)	93.00	78.53	14.47	**282**	Lormetazepam[23] (S)	75.00	81.87	−6.87
283	Methotrexate[22] (S)	70.00	59.45	10.55	**284**	Mexiletine[22] (S)	87.00	76.76	10.24
285	Midazolam[23]	40.50	54.81	−14.31	**286**	Nateglinide[21]	73.00	60.01	12.99
287	Nimodipine[21]	13.00	28.34	−15.34	**288**	Nitrazepam[23]	78.00	87.02	−9.02
289	Nitrendipine[23] (R)	16.00	31.80	−15.80	**290**	Nomifensine[23] (R)	27.00	43.55	−16.55
291	Norethindrone[23]	64.00	77.49	−13.49	**292**	Pirazolac[23]	93.50	101.81	−8.31
293	Pramipexole[21]	90.00	81.72	8.28	**294**	Prazosin[22]	68.00	74.90	−6.90
295	Procainamide[22]	83.00	70.91	12.09	**296**	Risperidone[21]	70.00	80.11	−10.11
297	Ropinirole[21]	55.00	68.84	−13.84	**298**	Simvastatin[22]	5.00	4.90	0.10
299	Spironolactone[23]	25.00	18.91	6.09	**300**	Tetracycline[22]	77.00	94.25	−17.25
301	Toliprolol[23] (R)	90.00	77.18	12.82	**302**	Venlafaxine[22] (R)	45.00	47.45	−2.45

[a] Experimental.
[b] Predicted.
[c] Residual, the difference between experimental and predicted values.

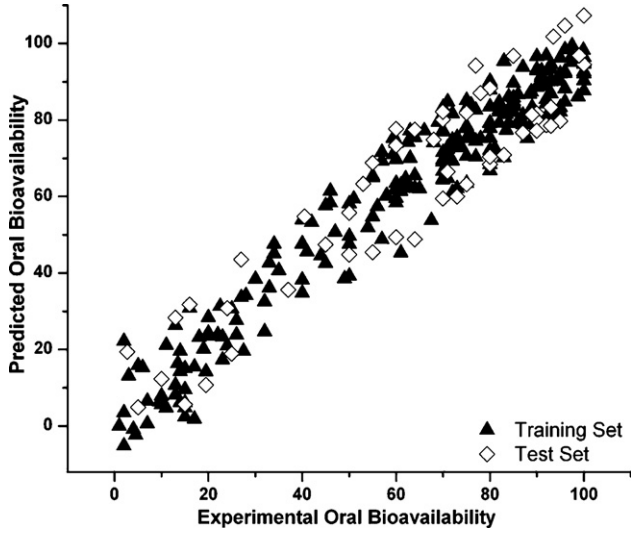

Figure 4. Plot of predicted versus experimental human oral bioavailability for training (250 molecules, model generation) and test sets (52 molecules, external validation).

structural types from which the model was derived. This model can be employed to assist the processes of chemical library design, virtual screening, and high-throughput screening. Compound libraries usually possess broad chemical diversity, therefore, in silico ADME models that are needed to screen these libraries should inevitably be global models designed to cover a broad scope of the chemical space. This was achieved by training the model with compounds from numerous chemical classes. The level of predictability is also adequate for the goal at this stage, given the chemical diversity and number of compounds involved. However, the model possesses some limitations that should be considered when interpreting the results. Considering that the majority of the data set members are drugs, those compounds which do not follow Lipinski's Rule of 5[26] would not have significant representation in the data set. Poorly soluble and not orally bioavailable drugs are not included in the data set, therefore, the model is not suitable in this case. In addition, in silico ADME approaches face considerable challenges in drug design. For instance, human oral bioavailability is a difficult property to measure because of the complexity of the underlying biological processes. These include variability of the experimental data, determination of the standard error, and validation of the massive amount of data, among others. In general, in silico ADME models require constant refinement and updating to ensure their applicability to developing medicinal chemistry in a way to continuously improve their quality and predictive power.

4. Conclusions

Drug discovery programs generally focus on the development of orally administered drugs for reasons of convenience, acquiescence, and market perspectives. The use of computational models in the prediction of pharmacokinetic properties of compounds is growing rapidly in drug discovery due to the benefits they provide in throughput and early application in the design of new drug candidates. The use of in silico ADME models has progressed with significant improvements in predictability and simplification, among other relevant aspects with the advance of new technologies. The HQSAR

model described herein shows both good internal and external consistency, and should be useful for the design of new drug candidates having improved oral bioavailability. Although such a model cannot completely replace in vitro and in vivo testing, it is a useful tool for rapidly screening compounds for their probable behavior, allowing early elimination of unfavorable candidates.

Acknowledgments

We gratefully acknowledge financial support from FAPESP (The State of São Paulo Research Foundation), CNPq (The National Council for Scientific and Technological Development) and CAPES (The Brazilian Coordination for the Improvement of Higher Education Personnel), Brazil.

References and notes

1. O'Brien, S. E.; de Groot, M. J. *J. Med. Chem.* **2005**, *48*, 1287.
2. Mager, D. E. *Adv. Drug Deliv. Rev.* **2006**, *58*, 1326.
3. van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discovery* **2003**, *2*, 192.
4. Lombardo, F.; Gifford, E.; Shalaeva, M. *Mini-Rev. Med. Chem.* **2003**, *3*, 861.
5. Saunders, K. C. *Drug Discovery Today: Technol.* **2004**, *1*, 373.
6. Ekins, S.; Nikolsky, Y.; Nikolskaya, T. *Trends Pharmacol. Sci.* **2005**, *26*, 202.
7. Weiner, D. *Drug Discovery Dev.* **2006**, *12*, 54.
8. Hou, T.; Wang, J.; Zhang, W.; Xu, X. *J. Chem. Inf. Model.* **2007**, *47*, 208.
9. Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. *J. Med. Chem.* **2007**, *50*, 501.
10. Leitão, A.; Andricopulo, A. D.; Oliva, G.; Pupo, M. T.; de Marchi, A. A.; Vieira, P. C.; da Silva, M. F. G. F.; Ferreira, V. F.; de Souza, M. C. B. V.; Sa, M. M.; Moraes, V. R. S.; Montanari, C. A. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2199.
11. Yoshida, F.; Topliss, J. G. *J. Med. Chem.* **2000**, *43*, 2575.
12. Hansch, C.; Kurup, A.; Garg, R.; Gao, H. *Chem. Rev.* **2001**, *101*, 619.
13. Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. *Bioorg. Med. Chem.* **2004**, *12*, 3391.
14. Jónsdóttir, S. Ó.; Jorgensen, F. S.; Brunak, S. *Bioinformatics* **2005**, *21*, 2145.
15. Beresford, A. P.; Selick, H. E.; Tarbit, M. H. *Drug Discovery Today* **2002**, *7*, 109.
16. Tetko, I. V.; Bruneau, P.; Mewes, H. W.; Rohrer, D. C.; Poda, G. I. *Drug Discovery Today* **2006**, *11*, 700.
17. Castilho, M. S.; Postigo, M. P.; de Paula, C. B. V.; Montanari, C. A.; Oliva, G.; Andricopulo, A. D. *Bioorg. Med. Chem.* **2006**, *14*, 516.
18. Salum, L. B.; Polikarpov, I.; Andricopulo, A. D. *J. Mol. Graphics Modell.* **2007**, *25*, 434.
19. Honorio, K. M.; Garratt, R. C.; Andricopulo, A. D. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3119.
20. Physicians' Desk Reference (PDR), 59th ed.; Thomson, 2005.
21. Hardman, J. G.; Limbird, L. E.; Gilman, A. G.. *Pharmacokinetic Data*, 10th ed. In *Goodman s The Pharmacological Basis of Therapeutics*; McGraw-Hill: United States of America, 2001, p. 1924.
22. Sietsema, W. K. *Int. J. Clin. Pharmacol. Ther. Toxicol.* **1989**, *27*, 179.
23. Honorio, K. M.; Garratt, R. C.; Polikarpov, I.; Andricopulo, A. D. *Lett. Drug Des. Discovery* **2006**, *3*, 261.
24. HQSAR™ Manual, SYBYL 7.2, Tripos Inc., St. Louis, MO, 2003.
25. Castilho, M. S.; Guido, R. V. C.; Andricopulo, A. D. *Lett. Drug Des. Discovery* **2007**, *4*, 106.
26. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeny, P. J.; Feeney, P. J. *Adv. Drug Delivery Rev.* **1997**, *23*, 3.