QSAR study of pharmacological permeabilities

Mati Karelson*[a], Gunnar Karelson,[a,c] Tarmo Tamm,[b,c] Indrek Tulp,[b,c] Jaak Jänes,[b,c] Kaido Tämm,[b,c] Andre Lomaka,[a,c] Deniss Savchenko,[a,c] and Dimitar Dobchev[a,c]

[a] *Department of Chemistry, Tallinn University of Technology, Akadeemia tee 15, Tallinn 19086, Estonia*
[b] *Department of Chemistry, University of Tartu, Jakobi Street 2, Tartu 51014, Estonia*
[c] *MolCode, Ltd., Soola 8, Tartu 51013, Estonia*
Email: dimitar@molcode.com

Abstract
Quantitative Structure-Activity Relationship (QSAR) models are developed for three pharmacological permeabilities, i.e. two PAMPA apparent permeabilities ($logP_{app}$) at different pH values (pH 5.5 and pH 7.4) and Caco-2 cell monolayer apparent permeability ($logP_{app}$(Caco-2)). The compounds are represented by chemical descriptors calculated from their constitutional, geometrical and topological structure, and quantum mechanical wave function. The obtained linear (multilinear regression) and nonlinear (artificial neural network) models link the drug structures to their reported permeabilities. Each multilinear model was tested by leave-one-out and ABC methods whereas the neural networks were assessed using the test sets. All drug structures were investigated by conformational analysis in order to find the low energy conformers.

Keywords: QSAR, neural network, BMLR, PAMPA, Caco-2 cell monolayer assay, cell membrane permeability

Introduction

Within drug discovery, about 40%-50% of drug candidates fail during late-stage development or clinical trials due to their poor ADMET (absorption, distribution, metabolism, excretion and toxicity) properties.[1,2] During the past decade, however, several experimental techniques for aqueous solubility,[3] plasma protein binding[4-6] blood-brain barrier penetration and intestinal absorption[7-9] with corresponding *in silico* prediction methods[10-12] have been developed and integrated into the early phases of the drug discovery and evaluation to improve the efficiency and cost-effectiveness of pharmaceutical industry. Compared to the experimental approaches, the

computational methods have some important advantages: (i) there is no need to initially synthesize compounds for determining their ADMET properties and (ii) rapid processing of the property data using numerous different modeling tools. Consequently, carefully developed and rigorously validated *in silico* ADME prediction models applied for the experimental observations in early screening and evaluation of compounds allow accurately prediction of key information of new drug candidates and to remarkably enhance the productivity in drug discovery.

One of the most important ADME-Tox properties, the drug intestinal absorption, is a complicated process determined by certain physiological conditions (local pH, absorptive surface area), activities of enzymes/transporters/carriers of gastrointestinal tract and chemical properties (solubility, molecular size and stability) of a drug.[13] Although the computational (*in silico*) approach for the prediction of intestinal and oral absorption is very attractive, even satisfactory accuracy is difficult to achieve.[14,15]

Two of the most popular *in vitro* absorption/permeability models used today involve Caco-2 cell monolayer and Parallel Artificial Membrane Permeability Assays (PAMPA). Caco-2 cells are derived from human epithelial colon adenocarcinoma and retain many morphological and functional properties of the intestinal enterocytes.[16-18] Thus, Caco-2 cell monolayer assay provides information about the drug absorption potential at near physiological conditions; however, its use is often limited due to the long membrane growth cycle and high costs.[19-21] A less expensive high throughput alternative, the PAMPA, first introduced by Kansy *et al*.[22] has gained recent popularity. This easily automated *in vitro* drug absorption assay is based on the use of a filter-immobilized artificial lipid (phosphatidylcholine) membrane.[23] Several experimental conditions (different membrane lipid compositions or multiple pH measurements) have been proposed for the determination of artificial membrane permeability values.[24-27]

Numerous attempts have been made to explain and predict the absorption of drug candidates via physico-chemical properties, such as the experimental and calculated octanol-water partition coefficient, (logP),[28-30] the apparent distribution coefficient at pH=7.4 used for ionizable compounds, D (expressed as logD),[31] hydrogen bonding potential (ΔlogP)[32] and desolvation energy.[33,34] Such properties have been frequently correlated to intestinal absorption rate or cell membrane permeability. Furthermore, Palm *et al.* obtained excellent correlation between dynamic polar van der Waals' surface areas of homologous series of beta-adrenoreceptor antagonists and their absorption in both Caco-2 monolayer assay and rat intestinal segments.[35] These results indicate that the dynamic polar surface area could be a better theoretical descriptor for intestinal drug absorption than lipophilicity (logP, logD) or hydrogen bonding potential. Good correlation, however, is generally impaired when structural diversity is introduced. This problem can be diminished by combining several physico-chemical properties into one expression, with the aid of multiple linear regression analysis.[36, 37] In addition, Lipinski and coworkers at Pfizer Research Center, after analyzing the physico-chemical profiles of 2245 orally active drugs from the World Drug Index, proposed a set of general principles ("Rule of 5") that would help to distinguish the well-absorbed molecules from poorly-absorbed molecules. According to these principles, good absorption or permeability is more likely when:

1. Molecular weight is ≤ 500
2. LogP is ≤ 5
3. Number of H-bond donors (expressed as the sum of OHs and NHs) is ≤ 5
4. Number H-bond acceptors (expressed as the sum of Ns and Os) is ≤ 10

However, some therapeutic classes (antibiotics, antifungals, vitamins and cardiac glycosides) of compounds being substrates for naturally occurring transporters fall outside the rule of 5.[38]

Recently, Verma *et al.* have carried out comparative QSARs of PAMPA for profiling of drug absorption potential with respect to Caco-2 cells and human intestinal absorption.[39] They developed several models with various numbers of drugs (9-94) whose correlation ranged from 0.7-0.9 in terms of the coefficient of determination (R^2). The models were relying mainly on ClogP (lipophilicity) including also highly nonlinear terms.

Guangli and Yiyu developed QSAR models using the support vector machine (SVM) method and multilinear (MLR) approach based on 100 drugs for Caco-2 cell monolayer permeability.[40a] They obtained SVM models with (Spearmen R^2) $R^2 = 0.77$ for training set of 77 drugs and $R^2 = 0.72$ for test set of 23 drugs. Their MLR model consisting of 4 descriptors resulted in $R^2 = 0.54$ for the training set and 0.61 for the test set. All descriptors in the models relied on the partial charge distributions and hydrogen donor ability of the drugs.

Fujikawa et al. built QSAR models using permeability coefficients of both hydrophilic and hydrophobic chemicals determined by PAMPA. Combining the descriptors of lipophilicity (logP and logD) as well as hydrogen-accepting and hydrogen-donating ability, they obtained a bilinear QSAR model with $R^2_{cv}= 0.68$ for a set of 97 compounds. In addition, they compared the apparent permeability coefficients of PAMPA to Caco-2 cell monolayer assay and derived a QSAR model with $R^2_{cv}= 0.73$ for a set of 35 compounds.[40b]

The goal of the current study is to develop QSAR models based on three pharmacological permeabilities *i.e.* apparent permeability coefficients determined by PAMPA at two different pH levels and Caco-2 cell monolayer permeability coefficients, proceeding from the same, uniform initial set of molecular descriptors. The different models are compared in terms of descriptors involved and statistical parameters. Two approaches were taken in order to achieve this goal, namely, the development of multilinear mathematical equation and creation and training of artificial neural networks. Before the QSAR development all drugs were studied in terms of conformational search in order to find the optimum low energy structures.

Data sets

In the attempt to predict pharmacological permeabilities, 3 different datasets were collected from the literature. The Caco-2 cell monolayer permeability data was obtained from Castillo-Garit *et al.*[41] This dataset was selected to cover a range of structural diversity, molecular weight, net charge and diverse absorption mechanisms of different compounds. Permeability is represented by the apparent permeability coefficient measured in cm/s on Caco-2 cell monolayer assay (P_{app}(Caco-2)). The dataset consists of 81 compounds.

The determination of artificial membrane permeability values requires two pH conditions (pH 7.4 and 5.5) to predict oral absorption. The apparent permeability coefficients of miscellaneous drugs determined by PAMPA at pH 5.5 (75 compounds) and pH 7.4 (62 compounds) were obtained from the literature.[39] Permeability data is represented by the apparent permeability coefficient measured in cm/s on PAMPA assay (P_{app}).

All experimental values for the three permeabilities along with the respective CAS numbers of the drugs are collected in Table 1. As can be seen from Table 1 the different data for most of the compounds are overlapped among the sets facilitating simultaneous comparison of the models in this study.

Table 1. Experimental (obsd) and predicted (calcd) permeation of used compounds

No.	CAS number	Compound	log P_{app} (pH 5.5)[x10^{-6}]			log P_{app} (pH 7.4) [x10^{-6}]			log P_{app} (Caco-2)		
			obsd	calcd MLR	calcd ANN	obsd	calcd MLR	calcd ANN	obsd	calcd MLR	calcd ANN
1	15722-48-2	Olsalazine							-6.960	-6.171	-6.138[a]
2	73384-59-5	Ceftriaxone							-6.880	-6.759	-6.734
3	23214-92-8	Doxorubicin	-0.520	-0.096	-0.525	-0.300	-0.276	-0.246	-6.800	-6.666	-6.589
4	58-94-6	Chlorothiazide	-0.700	-0.372	-0.758	0.110	-0.094	-0.094	-6.720	-6.59	-6.730
5	599-79-1	Sulfasalazine	-0.520	-0.042	-0.486				-6.710	-6.114	-6.087
6	54-31-9	Furosemide	-0.220	-0.315	-0.178	-0.220	-0.100	-0.239	-6.510	-6.039	-6.018
7	29122-68-7	Atenolol	-1.000	-0.386	-0.082[a]				-6.500	-5.619	-6.568
8	23031-25-6	Terbutaline							-6.380	-5.807	-5.849
9	28797-61-7	Pirenzepine							-6.360	-5.354	-6.276[a]
10	66357-35-5	Ranitidine							-6.310	-5.493	-6.479
11	127779-20-8	Saquinavir							-6.260	-5.785	-5.758
12	466-06-8	Proscillaridin							-6.200	-6.097	-6.068
13	15676-16-1	Sulpiride	-0.700	0.061	-0.848				-6.160	-6.167	-6.186
14	59277-89-3	Acyclovir							-6.150	-5.722	-5.949
15	126222-34-2	Remikiren							-6.130	-6.178	-6.203
16	26787-78-0	Amoxicillin							-6.100	-5.898	-5.912
17	58-93-5	Hydrochlorothiazide							-6.060	-6.034	-5.978
18	59865-13-3	Cyclosporine							-6.050	-6.428	-5.969[a]
19	6673-35-4	Practolol							-6.050	-5.425	-6.403
20	51-43-4	Epinephrine							-6.020	-5.214	-5.251
21	59-05-2	Methotrexate	-0.700	-1.320	-0.615				-5.920	-6.603	-6.583
22	51481-61-9	Cimetidine							-5.890	-5.049	-4.962

Table 1. Continued

23	37517-30-9	Acebutolol	−0.700	−0.153	−0.738	0.520	0.642	0.512	-5.830	-5.765	-5.821
24	57-50-1	Sucrose							-5.770	-5.909	-5.838
25	75847-73-3	Enalapril	0.530	0.386	0.318				-5.640	-5.204	-5.731
26	114-07-8	Erythromycin							-5.430	-5.849	-5.832
27	42200-33-9	Nadolol							-5.410	-5.679	-5.581
28	88495-63-0	Artesunate							-5.400	-5.462	-5.434[a]
29	66-22-8	Uracil							-5.370	-5.24	-5.367
30	57-13-6	Urea							-5.340	-5.22	-5.514
31	30516-87-1	Zidovudine	-0.220	-0.005	0.076	0.690	0.328	0.870[a]	-5.160	-4.953	-4.918
32	50-78-2	Acetylsalicylic acid	0.510	0.904	0.814	0.580	0.777	0.873	-5.060	-4.841	-4.771
33	36894-69-6	Labetalol				0.650	0.614	0.448	-5.030	-5.854	-5.846
34	51-61-6	Dopamine							-5.030	-5.06	-5.123
35	51-34-3	Scopolamine							-4.930	-5.116	-5.202
36	116644-53-2	Mibefradil							-4.870	-4.736	-4.710
37	26839-75-8	Timolol	0.230	0.255	0.211[a]	0.710	1.019	0.707	-4.850	-5.224	-5.307[a]
38	50-49-7	Imipramine	1.110	1.215	1.134	0.920	1.127	1.344	-4.850	-4.427	-4.362
39	79660-72-3	Fleroxacin							-4.810	-5.03	-5.064
40	69-72-7	Salicylic acid	1.330	0.878	1.152	0.520	0.677	0.674	-4.790	-4.85	-4.821
41	13523-86-9	Pindolol	1.120	0.699	0.858[a]	0.690	1.050	0.650	-4.780	-5.171	-5.158
42	39562-70-4	Nitrendipine							-4.770	-5.092	-5.020
43	50-28-2	Estradiol							-4.770	-4.556	-4.591
44	50-02-2	Dexamethasone	0.830	0.479	0.601[a]	0.910	0.808	0.847[a]	-4.750	-5.056	-5.015
45	54-11-5	Nicotine	1.170	1.473	1.233	1.330	1.237	1.194	-4.710	-4.33	-5.761[a]
46	71125-38-7	Meloxicam							-4.710	-5.726	-4.251
47	50-53-3	Chlorpromazine	1.070	1.427	1.535	0.600	0.877	0.654	-4.700	-4.518	-4.401
48	56-54-2	Quinidine	0.780	0.921	0.888	1.040	1.083	0.996	-4.690	-5.177	-4.724
49	53-86-1	Indomethacin	0.800	0.809	0.733	0.380	0.809	0.929	-4.690	-4.721	-5.128
50	56-75-7	Chloramphenicol	0.830	0.121	0.329	0.230	0.432	0.592[a]	-4.690	-4.834	-4.724
51	6452-71-7	Oxprenolol							-4.680	-5.081	-4.555
52	50-23-7	Hydrocortisone	0.490	0.550	0.494	0.530	0.477	0.468	-4.660	-4.834	-4.799
53	50-47-5	Desipramine	0.970	1.235	1.148	1.160	1.126	1.193	-4.640	-4.452	-4.279
54	72509-76-3	Felodipine							-4.640	-4.37	-4.437
55	13655-52-2	Alprenolol	0.150	0.527	0.473	1.180	1.048	1.026	-4.620	-4.663	-4.594
56	77-10-1	Phencyclidine							-4.610	-4.759	-4.723
57	37350-58-6	Metoprolol	0.080	0.184	0.045	0.540	0.803	0.836[a]	-4.590	-4.596	-4.613[a]
58	4205-90-7	Clonidine	1.300	1.280	1.169	1.150	0.776	0.895	-4.590	-4.315	-4.239
59	52-53-9	Verapamil	0.990	0.403	0.312[a]	0.870	0.822	0.979	-4.580	-4.692	-4.686
60	525-66-6	Propranolol	1.230	0.900	0.792	1.370	1.304	1.408	-4.580	-5.02	-4.938
61	57-41-0	Phenytoin	0.880	0.944	1.073	0.710	0.665	0.540	-4.570	-5.107	-5.104

Table 1. Continued

62	81-81-2	Warfarin	1.020	0.698	0.986	1.090	0.885	0.908	-4.550	-4.271	-4.391
63	60-80-0	Antipyrine	1.300	0.984	1.200	1.120	1.370	1.391	-4.550	-4.855	-4.830
64	129618-40-2	Nevirapine							-4.520	-4.795	-4.874
65	5051-62-7	Guanabenz	0.200	0.128	0.112	1.240	1.053	1.132	-4.500	-5.028	-4.996
66	50-22-6	Corticosterone	1.590	0.554	1.477	1.340	0.653	1.186	-4.470	-4.628	-4.562[a]
67	36322-90-4	Piroxicam	0.920	0.472	0.460	0.910	0.545	0.683	-4.450	-5.222	-4.197
68	126-07-8	Griseofulvin	0.890	1.048	0.800	0.720	0.198	0.718	-4.440	-4.802	-4.742
69	58-15-1	Aminopyrine							-4.440	-4.046	-3.932
70	58-08-2	Caffeine	1.310	0.995	0.925	1.030	0.811	0.813[a]	-4.410	-4.835	-4.871
71	42399-41-7	Diltiazem	1.030	0.519	0.427	1.270	1.211	0.975	-4.380	-4.606	-4.580
72	57-83-0	Progesterone	-0.100	0.235	0.079[a]	0.600	0.509	0.365	-4.370	-4.643	-4.672
73	19216-56-9	Prazosin	0.400	0.343	0.131	1.130	0.877	0.654[a]	-4.360	-5.158	-4.137
74	58-55-9	Theophyline							-4.350	-4.702	-4.704
75	58-22-0	Testosterone							-4.340	-4.056	-4.123[a]
76	439-14-5	Diazepam							-4.320	-4.153	-4.171
77	15687-27-1	Ibuprofen	1.030	1.169	0.966	0.830	0.902	1.009	-4.280	-4.569	-4.291
78	137-58-6	Lidocaine							-4.210	-4.36	-4.295
79	22204-53-1	Naproxen	1.360	1.222	1.067[a]	1.030	0.969	1.092	-4.130	-4.77	-4.175
80	91-64-5	Coumarin	1.360	1.002	1.287	1.340	1.124	1.575	-4.110	-3.927	-3.923
81	67-56-1	Methyl alcohol							-3.880	-3.645	-4.514[a]
82	637-07-0	Clofibrate	-0.400	-0.351	-0.495[a]	-0.520	-0.289	-0.117[a]			
83	70458-96-7	Norfloxacine	-0.300	0.191	-0.146	-0.050	0.301	0.25[a]			
84	22916-47-8	Miconazole	-0.150	0.194	0.272						
85	33419-42-0	Etoposide	-0.150	0.143	0.045	-0.400	-0.314	-0.438			
86	60142-96-3	Gabapentin	0.080	0.446	0.098	0.080	0.397	0.092			
87	25614-03-3	Bromocriptine	0.110	0.608	0.431						
88	87-08-1	Penicillin V	0.200	0.634	0.463						
89	63590-64-7	Terazosine	0.230	-0.124	-0.096	0.940	0.852	0.923			
90	50-24-8	Prednisolone	0.340	0.480	0.403	0.760	0.854	0.753[a]			
91	103-90-2	Acetaminophen	0.360	0.623	0.807	0.540	0.785	0.734			
92	83-43-2	Methylprednisolone	0.410	0.507	0.399	0.770	0.892	0.765			
93	738-70-5	Trimethoprim	0.430	0.571	0.292	0.700	0.881	0.621			
94	3930-20-9	Sotalol	0.460	0.216	0.469[a]	0.040	-0.238	0.530			
95	43200-80-2	Zopiclone	0.510	0.645	0.398	0.950	0.767	0.526			
96	65277-42-1	Ketoconazole	0.520	0.083	0.036	0.080	0.082	0.314			
97	57-66-9	Probenecid	0.600	0.964	0.764	0.380	0.350	0.501			
98	62571-86-2	Captopril	0.640	0.496	0.323[a]	1.280	1.186	0.908			
99	28395-03-1	Bumetanide	0.660	0.247	0.243	-0.520	-0.189	-0.182			
100	78755-81-4	Flumazenil	0.680	1.023	0.780	0.780	0.858	0.827			

Table 1. Continued

101	74103-06-3	Ketorolac	0.710	0.553	0.980	0.150	0.732	0.676
102	81-07-2	Saccharin	0.850	1.154	1.354			
103	54910-89-3	Fluoxetine	0.870	1.065	0.630	1.150	1.332	1.371
104	148-82-3	Melphalan	1.010	0.418	1.155	0.760	0.914	0.525
105	15307-86-5	Diclofenac	1.030	1.035	1.152	1.100	0.918	1.155
106	298-46-4	Carbamazepine	1.080	1.024	1.004	1.050	1.097	1.051
107	22071-15-4	Ketoprofen	1.280	1.042	0.971	1.220	0.943	0.97[a]
108	5786-21-0	Clozapine	1.350	1.246	1.292	1.450	1.058	1.408
109	34841-39-9	Bupropion	1.680	1.058	0.961	1.150	1.056	1.189

[a] Data points used for test set for ANN

Methodology

1. Structure optimization and descriptor calculation

In the current study, we present QSAR models for $logP_{app}$(pH 5.5), $logP_{app}$(pH 7.4) and $logP_{app}$(Caco-2) involving theoretical descriptors, which were calculated solely from the molecular (drug) structure using CODESSA Pro [42] and QSARModel program.[43] These descriptors can be classified as: (i) constitutional, (ii) geometrical, (iii) topological, (iv) charge-related, (v) quantum chemical, and (vi) thermodynamic. [44a-44d] The total number of descriptors for each property ranged between 600 and 900 per compound.

The conformational search was also performed for all compounds using the MacroModel software package. For the calculations, MMFF94s - a static variant of Merck Molecular Force Field 94 (MMFF94)[46] was used. The energy minimization was carried out using the Polak-Ribiere Conjugate Gradient (PRCG) method with a gradient 0.05 kcal/Å as a stopping criterion was used. Furthermore, for the conformational search, Monte Carlo Multiple Minimum (MCMM) method was used where 100 steps per rotatable bond and up to 15000 maximum steps per compound were defined.[47] Conformers with the minimum potential energy were utilized as the MOPAC 6 [48] input structures. Within quantum-mechanical semi-empirical calculations the AM1 [49] parameterization was used with the gradient norm setting of 0.01 kcal/Å.

For reasons of comparison and completeness, optimization of the structures without prior conformational search was also performed. However, the models derived from these drug structures did not lead to significant QSAR equations and hence these results are not presented.

2. QSAR modeling

(a) Linear approach (BMLR)

The Best Multilinear Regression method[50] was used to find the best correlation models from selected non-collinear descriptors. The BMLR selects the best two-parameter regression equations, the best three-parameter regression equations, etc. on the basis of the highest R^2 and F

values in the step-wise regression procedure. Further, the selection of the final equation consists of comparison of R^2 and F augmentation of the statistical parameters of the best 2-, 3- and etc. descriptor equations. The selection of the optimum number of descriptors for the equation is performed by simple plots of R^2 vs number of descriptors i.e. if the difference in R^2 between n-1 and n- descriptor models drops below 0.2 then the procedure is stopped. The result obtained by BMLR is the "best" linear representation of the property in the given descriptors pool.

(b) Nonlinear approach (ANN)

Artificial neural networks (ANN)[51-53] have become an important modeling technique for QSAR and QSPR, and Artificial Neural Network modeling has been applied in numerous application areas of chemistry and pharmacy. [54-57] The mathematical adaptability of ANN commends them as a powerful tool for pattern classification and building predictive models. A particular advantage of ANNs is their inherent ability to incorporate nonlinear dependencies between the dependent and independent variables without using an explicit mathematical function.

A fully connected neural network with backpropagation of the error[58] was constructed and used in the building of the nonlinear models for all three properties. ANNs are composed of a number of single processing elements (PE) or units (nodes, neurons). The activation function used for the PEs was standard sigmoid function. PEs are connected with coefficients (weights) and are organized in a layered topology as follow: (i) the input layer – PEs (I_i in this work i=3-7) consisting of the molecular descriptors, (ii) the output layer (with PEs O_m – in this work m=1) O_i consisting of the calculated property ($logP_{app}$) and (iii) the hidden layers (with PEs H_h – in this work h=3-7) between them. The number of layers (3-in this work) and the number of units in each layer determines the functional complexity of the ANN as shown in Fig. 1. Each input layer PE corresponds to a single independent variable (molecular descriptor). Similarly, each output layer node corresponds to a different dependent variable (property under investigation).

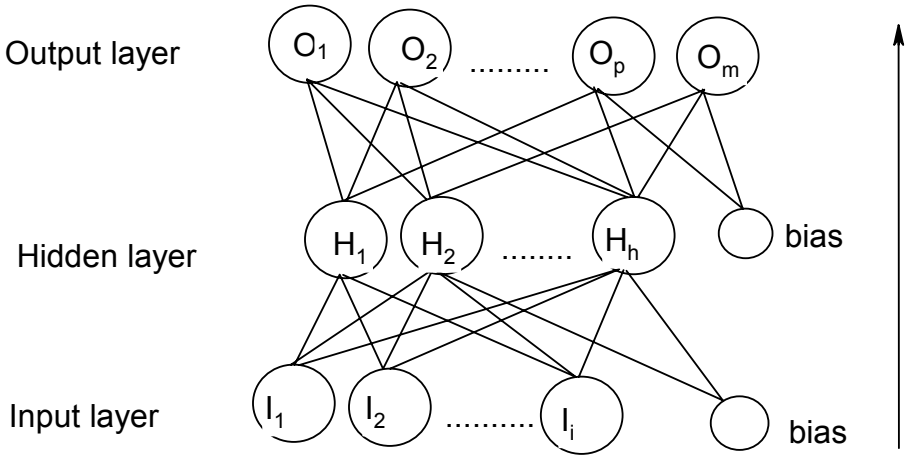

Figure 1. Three layer back propagation neural network.

In order to find the most important descriptors as inputs to the net, a sensitivity analysis was performed on a preselected descriptor space, based on the lowest root-mean squared error (RMS). This space was formed after applying the following criteria for reduction of the total descriptor space: i) all descriptors with variance less than 10^{-4} were excluded, ii) descriptors which did not correlate with the property more than R^2=0.2 were excluded also iii) by inspection of certain chemically irrelevant descriptors.

An important stage of the modeling is to define the proper architecture of the ANN models. Several ANN models with different architectures were built for each property. In the search for an optimal ANN architecture the lowest possible number of neurons was searched for, in order to follow the common principle of generality of the ANN prediction. In addition, we monitored the RMS for each different architecture (regarding the hidden units in the hidden layer) in order to select the one with the lowest RMS. The number of layers was chosen to be three-fold based on the common practice for the QSAR ANN modeling and by taking into account the number of data points so that to reduce the chance for overfitting during the training stage.

The ANN was trained on a training set selected from the total number of compounds where the weights were adjusted for each iteration by the delta rule. During the optimization, the RMS was monitored together with the RMS of a test set so that to avoid overfitting problems. An in-house program was used for the ANN calculations.

3. ABC Validation of the MLR models

To validate the multilinear models, the data was sorted in the ascending order according to the experimental value, and three subsets (A, B, C) were then formed: the 1st, 4th, 7th, etc. data points comprise the first subset (A), the 2nd, 5th, 8th, etc. comprise the second subset (B), and the 3rd, 6th, 9th, etc. comprise the third subset (C). The three training sets were prepared as the combinations of any two subsets (A and B), (A and C), and (B and C), respectively. The tested MLR model was then rebuilt for each of the training sets with the same descriptors but optimized regression coefficients, and used to predict the property values for the respective C, B and A subsets. The prediction was assessed based on the R^2 between the predicted and experimental property values.[55]

In addition to the ABC validation, the standard leave-one-out (LOO) cross-validation (R^2_{cv}) for all developed models was used.

Results and Discussion

1. MLR models for $logP_{app}$ (pH 5.5), $logP_{app}$(pH7.4) and $logP_{app}$(Caco-2)

The BMLR algorithm was used to generate several multilinear equations for each property with the number of descriptors between 2 and 7. The final equations were selected by taking into account two factors in order to obtain significant models: i) the number of the descriptors in the models should follow the basic "rule of thumb" *i.e.* not less than five data points per descriptor

and ii) relevance of the descriptors toward the nature of the phenomenon under investigation iii) increase in the difference of R^2 for the different models with different descriptor numbers (see Methodology). The resulted models for the three properties are summarized in Table 2 together with their statistical parameters and descriptors. A comparison of models 1 and 2 for the PAMPA permeabilities (see Table 2) shows that model 2 possesses better statistical characteristics. However, this model has been developed following a lower number of drugs.

Table 2. Statistical Parameters for BMLR Models of $logP_{app}$ (pH 5.5), $logP_{app}$(pH7.4) and logP(Caco-2)

No	BMLR equations	n	k	R^2	R^2_{cv}	F	s^2
1	$logP_{app}(5.5)$ = -16.98 - 0.078*D1 - 0.219*D2 + 282.2*D3 - 0.859*D4 + 19.90*D5 + 55.35*D6	70	6	0.670	0.589	20.86	0.150
2	$logP_{app}(7.4)$ = 6.21 - 0.009*D7 +0.004*D8 + 315.2*D9 -24.44*D10 + 0.356*D11 + 66.50*D12	62	6	0.759	0.674	28.90	0.065
3	$logP_{app}$(Caco-2) = -4.099 + 3755*D13 - 1.060*D14 -0.01418*D15 - 761.3*D16 - 0.0567*D17 + 0.0235*D18	81	6	0.724	0.680	32.39	0.196

NB: (number of data points (*n*), number of descriptors (*k*), squared correlation coefficient (R^2), cross validated squared correlation coefficient (R^2_{cv}, LOO), Fisher ratio (*F*), squared standard deviation (s^2).

In general, models 1 and 2 are of average quality although model 2 shows higher statistical performance according to R^2_{cv} (0.59 vs 0.67). Better MLR models (in terms of R^2 and F) could not be obtained for these data. One of the reasons is the structural diversity of the compounds in data sets used as well as the accuracy of the experimental data. The structural variability of the drugs poses difficulties to express PAMPA permeabilities (at both pH conditions) as simple linear function of descriptors. This fact was taken into account in one of the reports by Verma *et. al*.[39] by including an additional structural variable in their models that characterizes certain moieties of the drug. Their model for $logP_{app}$ (pH 5.5) shows a good correlation with R^2 = 0.728 (R^2_{cv} =0.69) developed for 60 drugs by using a nonlinear function of the hydrophobicity and two additional parameters. Our model (model 1) for this PAMPA permeability shows lower quality in terms of R^2 (=0.670). However, model 1 was developed on a larger data set compared to the Verma's equation (eq. 4 therein). In addition, a strong outlier appeared in model 1 which is significantly far from the prediction bands at the 95% level of confidence, as can be noted from Figure 2. Removing this outlier (Methotrexate – No. 21) leads to an improved correlation with R^2 = 0.701.

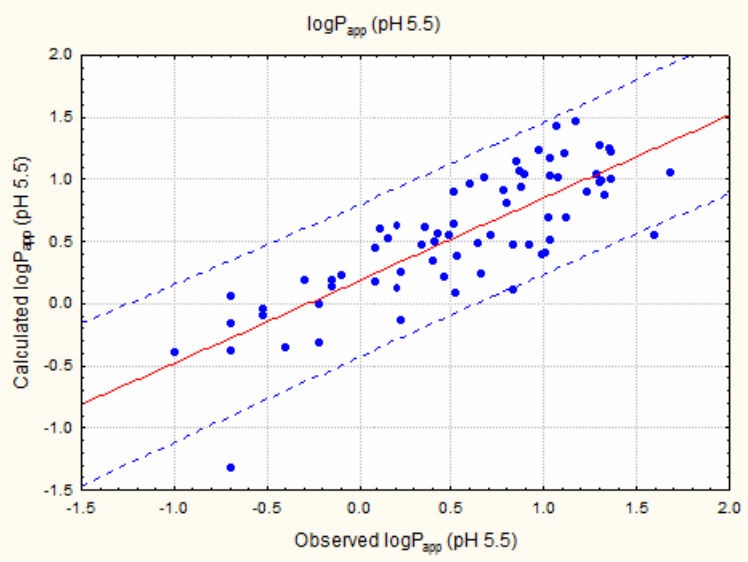

Figure 2. Experimental and predicted $logP_{app}$(pH 5.5) values based on equation 1 in Table 2.

The correlation between the experimental and predicted $logP_{app}$(pH 7.4) values based on equation 2 in Table 2 are presented in Figure 3. A general comparison of this model to the model for the same permeability of Verma's model (eq.5 therein), indicates that the QSAR model 2 has better statistics in terms of R^2 (0.759 vs0.743) and s^2 (0.065 vs 0.096). Moreover, model 2 is based on a larger data set (62 vs 55).

The model 3 (see Table 2 and Figure 4) for the Caco-2 cell monolayer set can be considered the best of the three, at R^2 = 0.724, (R^2_{cv} = 0.680), 81 structures, notably for the highest R^2_{cv} value. Direct comparison with other published results is difficult, as the datasets are not directly comparable. However, statistically the model is at least comparable (if not better) than other linear models for sets of comparable size and diversity [ref 16, and therein]; nonlinear models have shown better results.

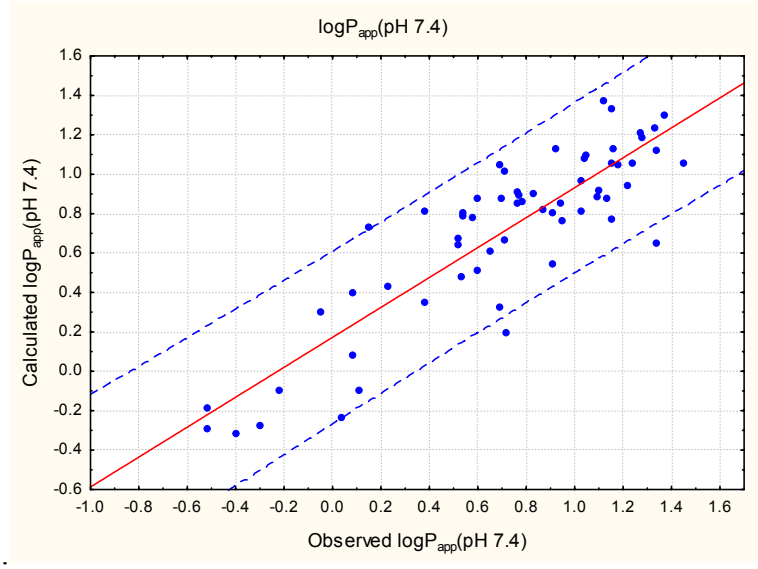

Figure 3. Experimental and predicted $logP_{app}$(pH 7.4) values based on equation 2 in Table 2.

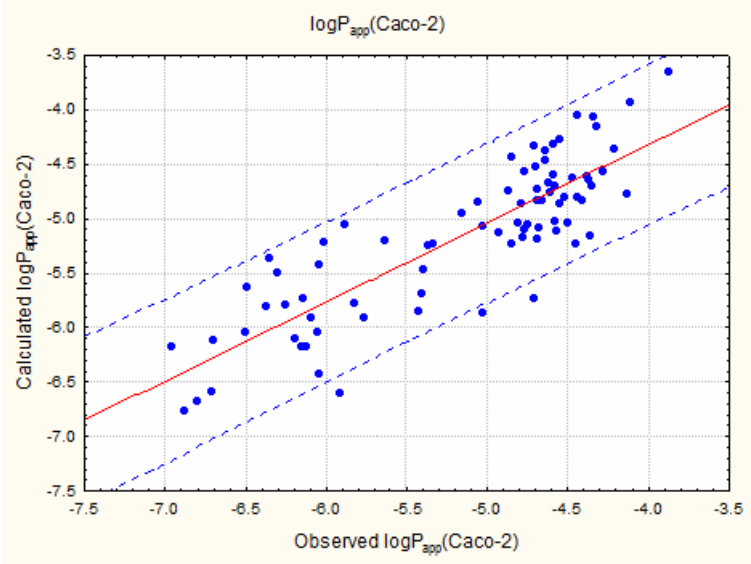

Figure 4. Experimental and predicted $logP_{app}$(Caco-2) values based on equation 3 in Table 2.

The descriptors involved in the QSAR models in Table 2 are collected in Table 3. From the Table, it can be noted that the descriptors for both PAMPA permeabilities are quite similar; the Caco-2 cell monolayer assay model descriptors form a slightly different subset. The precise formulation of all descriptors can be found in ref.[44a-44d] According to the t-statistics of the descriptors for models 1 and 2, the most significant variables are D2 and D7, respectively. D2 is an indicator for the molecular geometrical shape and also it encodes the branchiness of the drug

whereas D7 is related to the hydrogen acceptor ability of the drug and the respective partial charge on the hydrogen surface area. Increasing the values of these variables leads to decreased $logP_{app}$ values for both models 1 and 2. Hydrogen bonding is also of major importance in the Caco-2 monolayer assay model, as indicated by most of the descriptors, and the most significant (according to t-statistics) descriptor D16, in particular.

Fujikawa *et al*. have developed a QSAR model for the PAMPA data in which the hydrogen-accepting ability (SA_{HA}), and hydrogen-donating ability (SA_{HD}) descriptors were proven to be very important.[40b] Similar descriptors also appear in our models of Table 2 *i.e.* D1 - *HA dependent HDCA-2 (AM1) (all)*, D7 - *HACA H-acceptors charged surface area (AM1)*. Moreover, these hydrogen bonding ability descriptors, D1 and D7, follow the same sign (minus) in the equations as the SA_{HA} and SA_{HD} descriptors in Fujikawa's model (eq. 1 therein).

Table 3. Descriptors[#] involved in BMLR models in Table 2

Desc ID	Descriptor name
D1	HA dependent HDCA-2 (AM1) (all)
D2	Kier shape index (order 3)
D3	Minimum 1-electron reactivity index (AM1) for Cl atoms
D4	Minimum (>0.1) bond order (AM1) for C atoms
D5	Average bond order (AM1) for H atoms
D6	Minimum 1-electron reactivity index (AM1) for O atoms
D7	HACA H-acceptors charged surface area (AM1)
D8	WNSA2 Weighted PNSA (PNSA2*TMSA/1000) (Zefirov)
D9	Maximum 1-electron reactivity index (AM1) for Cl atoms
D10	Maximum electrophilic reactivity index (AM1)
D11	Highest total interaction (AM1)
D12	Average 1-electron reactivity index (AM1) for O atoms
D13	Average 1-electron reactivity index (AM1) for H atoms
D14	Difference (Pos - Neg) in Charged Partial Surface Area (AM1)
D15	HBCA H-bonding charged surface area (AM1)
D16	Minimum 1-electron reactivity index (AM1) for H atoms
D17	HACA-1 (Zefirov)
D18	RNCS Relative negative charged SA (SAMNEG*RNCG) (Zefirov)

#-Descriptor values included in Supplementary Information (SI1)

The following group of descriptors is molecular orbital derived: D3, D9, D10, D11, D6 and D12. In fact, these descriptors indicate the importance of the presence of certain atoms for the PAMPA permeability, namely, O and Cl atoms. The reactivity indices are explicitly related to the HOMO and LUMO energies and thus to the electrophilicity and nucleophilicity of the drug as was also indicated by Fujikawa and Verma. Notably, in the case of PAMPA permeabilities at the

different pH conditions, the minimums of electronic descriptors are involved in the model for $logP_{app}$(pH 5.5) whereas in the case of $logP_{app}$(pH 7.4) the maximum values of electronic descriptors appear in the best model.

The last group of descriptors is related to the stability of the chemical bond and surface area of the drug molecule: the descriptors D4, D5 and D8. The D8 – *WNSA2* is defined as the total sum of partial areas of the drug which possess negative partial charges times the total solvational area of the drug divided by 1000. This is the second descriptor by significance according to the t-test in model 2.

The descriptors for Caco-2 monolayer assay model (D13-D18) reflect the same features of the drug molecules as those in PAMPA models. The most significant descriptors in the equation 3 (see Tables 2 and 3) are D15 and D14 indicating that the hydrogen-bonding charged areas and the excess of the total charged surface area are the main factors for this model. The remaining descriptors contribute additionally to the charged areas (D18) and particularly areas with hydrogen donor/acceptor abilities (D17) and electrophilicity and nucleophilicity (D16, D13).

All descriptors in Table 3 are related to the charge, hydrogen acceptor and donor potentials as well as polar molecular surface area, electrophilicity and nucleophilicity of the drugs. These features have been found to play a crucial role in the kinetics and dynamics of intestinal permeation.[28-34,38] Caco-2 cell monolayer mimicking the morphology of the gastrointestinal tract allows study of both passive (transcellular and paracellular transport) and active (carrier-mediated influx and efflux) absorption mechanisms of drugs, while PAMPA provides the investigation of diffusion via passive transcellular (through the cell membrane of enterocytes) pathway. Since the absorption of many drugs across the intestinal epithelium may occur mainly by passive diffusion, a linear correlation between the apparent permeability coefficients of these two analytical methods is likely to exist in a certain degree.[40b]

1. Validation of the MLR models

One of the main goals of this study was also to develop MLR models as general as possible in order to cover a larger chemical space of applicability. Therefore, for the sake of generality, we did not use the division of data into training and external test sets.

Instead of direct external validation, a type of leave-many-out validation, *i.e.* the ABC validation (see Methodology) was employed to mimic external validation. The efficiency of QSAR models to predict the logarithmic scales of the PAMPA and Caco-2 cell monolayer permeabilities was assessed by the squared correlation coefficients $R^2_{(Pred)}$ between experimental and predicted data for test sets (A, B or C). The results from the ABC validation are presented in Table 4. The overall assessment of the predictions for both PAMPA models is of average quality as can be noted by the differences between the average $R^2_{(Fit)}$ and $R^2_{(Pred)}$. The reason for this difference is that the variability in the drugs cannot be easily accounted with linear function from the descriptors in the models. As demonstrated later, accounting for the nonlinearities improves significantly the predictions. The ABC validation of the linear model of the Caco-2 cell

monolayer set (although uneven between subsets) showed higher overall stability as compared to the PAMPA sets.

Table 4. Statistical characteristics for ABC validation of the multilinear model in Table 2

TRAINING SET	N	$R^2_{(Fit)}$	$R^2_{cv\ (Fit)}$	TEST SET	N	$R^2_{(Pred)}$
BMLR model for $logP_{app}$ (5.5)						
A+B	47	0.709	0.591	C	23	0.577
B+C	47	0.701	0.589	A	23	0.519
A+C	46	0.653	0.539	B	24	0.556
Average		0.688	0.573			0.551
BMLR model for $logP_{app}$ (7.4)						
A+B	42	0.814	0.694	C	20	0.444
B+C	42	0.802	0.729	A	20	0.630
A+C	43	0.752	0.601	B	19	0.578
Average		0.789	0.675			0.551
BMLR model for $logP_{app}$(Caco-2)						
A+B	54	0.705	0.609	C	27	0.781
B+C	54	0.733	0.671	A	27	0.722
A+C	54	0.761	0.694	B	27	0.650
Average		0.733	0.658			0.718

N – number of compounds used in the validation models
$R^2_{(Fit)}$ – squared correlation coefficient of the multilinear equation for the binary sets used as training sets (A+B, A+C or B+C)
$R^2_{cv(Fit)}$ – cross-validated squared correlation coefficient of the multilinear equations for the binary sets (leave-one-out approach)
$R^2_{(Pred)}$ – squared correlation coefficient for the regression between the predicted values (from the models of the binary sets) for the test sets and the respective experimental values

1. Neural network models

Before the neural network treatment was started, the experimental logarithmic values and descriptor values were both normalized to a range 0-0.9 for internal consistency. Then the significant descriptors were selected by reducing the initial descriptor pool as described in the Methodology part. For each property, the available experimental data were divided into training and test set. To preserve the generality of the network models, all test sets consisted of not more than ten data points. Further, sensitivity analyses were performed on the reduced descriptor space by constructing 1-1-1 neural networks and then the descriptors that produced the lowest RMS

error were selected. Several neural network models with different architecture were investigated for each property.

The best ANN model found for $logP_{app}(5.5)$ was 5-4-1, *i.e.* having five neurons (descriptors) in the input layer, four neurons in the hidden layer and one neuron in the output layer. The input neurons consisted of the following descriptors*: HA dependent HDCA-2/SQRT(TMSA) (AM1), Minimum 1-electron reactivity index (AM1) for Cl atoms, Kier shape index (order 3), Minimum (>0.1) bond order (AM1) for C atoms* and *Maximum net atomic charge (AM1) for Cl atoms*. The training of the network finished at approximate 350 iterations with RMS of the training set 0.266 and for the test set 0.406. The experimental and predicted values for $logP_{app}(5.5)$ for both sets are collected in Table 1 and shown in Figure 5.

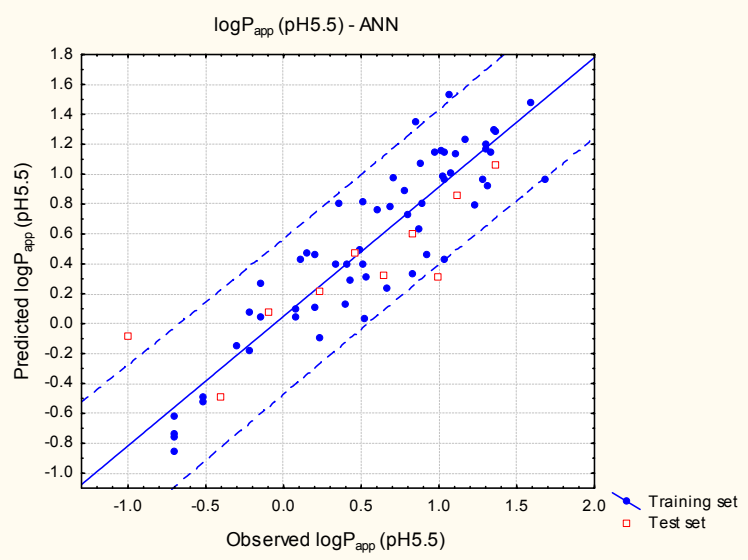

Figure 5. Experimental and predicted $logP_{app}$(pH 5.5) values based on 5-4-1 ANN model (Training set 61 and test set 10 data points).

The estimation of these predictions in terms of the coefficient of determination is R^2_{train} =0.820 and R^2_{test} = 0.740.

The next ANN model developed for the $logP_{app}(7.4)$ had also the architecture 5-4-1 while the input neurons consisted of the following descriptors*: FPSA3 Fractional PPSA (PPSA-3/TMSA) (AM1), WNSA3 Weighted PNSA (PNSA3*TMSA/1000) (Zefirov), Maximum 1-electron reactivity index (AM1) for Cl atoms, Maximum electrophilic reactivity index (AM1) for C atoms, Average 1-electron reactivity index (AM1) for O atoms*. This model was trained up to 411 iterations and the results from the predictions for both, training and test sets are shown in Figure 6 and Table 1. The statistical parameters for this models resulted in

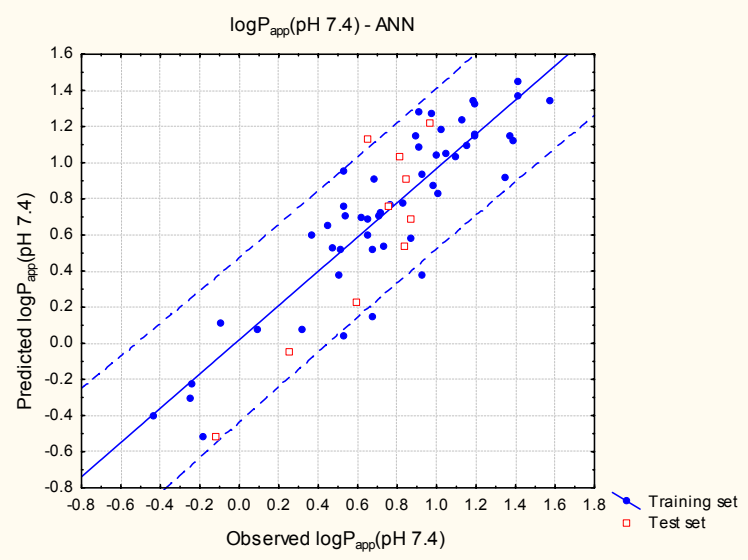

Figure 6. Experimental and predicted logP$_{app}$(pH 7.4) values based on 5-4-1 ANN model (Training set 52 and test set 10 data points)

R^2_{train} =0.801 and R^2_{test} = 0.791 and the achieved RMS error 0.217 and 0.290 for the training and test sets, respectively.

The final model developed was for the Caco-2 cell monolayer permeability logP$_{app}$(Caco-2) with topology 5-3-1. The following descriptors were used as inputs to the network: *HBCA H-bonding charged surface area (AM1), HACA-1 (Zefirov), Minimum 1-electron reactivity index (AM1) for H atoms, Difference (Pos - Neg) in Charged Partial Surface Area (AM1)* and *Average 1-electron reactivity index (AM1) for H atoms*. The experimental and predicted values for logP$_{app}$(Caco-2) from this ANN model are collected in Table 1 and graphically presented in Figure 7. The RMS error for the training and test sets are 0.333 and 0.496, respectively. The squared correlation coefficients for both sets, shown in Figure 7 are 0.823 (training) and 0.746 (test).

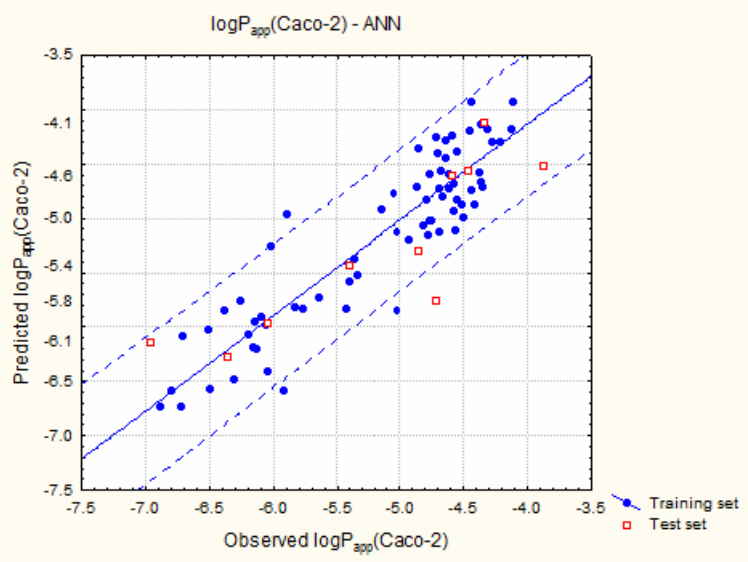

Figure 7. Experimental and predicted logP$_{app}$(Caco-2) values based on 5-3-1 ANN model (Training set 71 and test set 10 data points).

It was not surprising that most of the descriptors in the ANN models are quite similar to the descriptors involved in the MLR models. In general, the descriptors in ANN models reflect the charged surface areas, hydrogen donor/acceptor and lipophilic abilities of the drugs as in MLR models.

Conclusions

Our present attempt to correlate three permeabilities namely logP$_{app}$ (pH 5.5), logP$_{app}$ (pH 7.4) and logP$_{app}$(Caco-2) with theoretically calculated molecular descriptors has led to relatively successful QSAR models that relate these complex pharmacological and medicinal properties to structural characteristics of the drugs. Notably, all descriptors appearing in the multiparameter regression equations and the ANN model have been derived from theoretical molecular calculations. The current computational power available for chemical research allows such calculations for large data sets in realistic time. Thus, in principle, the QSAR models developed in our present work can be used for the prediction and screening of the above permeabilities. The descriptors appearing in these models can be related to the essential electrostatic, hydrogen acceptor/donor and lipophilic interactions between the drug and the cell membrane.

The results obtained for this work indicate that the regression and ANN models exhibit reasonable prediction capabilities. Though the linear model was developed mainly for the purpose of structure-activity interpretation, the ANN model was primarily developed for predictions and classification. It is worth noting that the drug structures obtained by conformational search resulted in better models for both BMLR and ANN.

In summary, this work should be able to provide prediction and screening of analogous drugs for their permeabilities.

Supplementary Information

Descriptor values in Table 3 available in SI1

References

1. van de Waterbeemd, H.; Gifford, E. *Nat. Rev. Drug Discov.* **2003**, *2*, 192.
2. Saunders, K. C. *Drug Discov. Today* **2004**, *1*, 373.
3. Lipinski, C. A. *J. Pharmacol. Toxicol. Meth.* **2000**, *44*, 235.
4. Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. *J. Med. Chem.* **2001**, *44*, 1927.
5. Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. *J. Pharmacol. Sci.* **2001**, *90*, 749.
6. Zsila, F.; Iwao, Y. *Biochim. Biophys. Acta-Gen. Subj.* **2007**, *1770*, 797.
7. Cucullo, L.; Hossain, M.; Rapp, E.; Manders, T.; Marchi, N.; Janigro, D. *Epilepsia*, **2007**, *48*, 505.
8. Garberg, P.; Ball, M.; Borg, N.; Cecchelli, R.; Fenart, L.; Hurst, R. D.; Lindmark, T.; Mabondzo, A.; Nilsson, J. E.; Raub, T. J.; Stanimirovic, D.; Terasaki, T.; Öberg, J.-O.; Österberg, T. *Toxicol. Vitro* **2005**, *19*, 299.
9. Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. *Pharmaceut. Res.* **1998**, *15*, 1490.
10. Wichmann, K.; Diedenhofen, M.; Klamt, A. *J. Chem. Inf. Model.* **2007**, *47,* 228.
11. Gunturi, S. B.; Narayanan, R. *QSAR Comb. Sci.* **2007**, *26*, 653.
12. Hall, L. M.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.* **2003**, *43,* 2120.
13. Martinez, M. N.; Amidon, G. L. *J. Clin. Pharmacol.* **2002**, *42***,** 620.
14. Ekins, S.; Waller, C. L.; Swaan, P. W.; Cruciani, G.; Wrighton, S. A.; Wikel, J. H. *J. Pharmacol. Toxicol. Meth.* **2000**, *44*, 251.
15. Norris, D. A.; Leesman, G. D.; Sinko, P. J.; Grass, G. M. *J. Contr. Release* **2000**, *65*, 55.
16. Bohets, H.; Annaert, P.; Mannens, G.; van Beijsterveldt, L.; Anciaux, K.; Verboven, P.; Meuldermans, W.; Lavrijsen, K. *Curr. Top. Med. Chem.* **2001**, *1*, 367.
17. Hidalgo, I. J.; Raub, T. J.; Borchardt, R. T. *Gastroenterology* **1989**, *96*, 736.
18. Hilgers, A. R.; Conradi, R. A.; Burton, P. S. *Pharm. Res.* **1990**, *7*, 902.
19. D'Souza, V. M.; Shertzer, H. G.; Menon, A. G.; Pauletti, G. M. *AAPS PharmSci* **2003**, *5*, article 24.
20. Bravo, S. A.; Nielsen, C. U.; Amstrup, J.; Frokjaer S.; Brodin, B. *Eur. J. Pharm. Sci.* **2004**, *21*, 77.

21. Anderle, P.; Niederer, E.; Rubas, W.; Hilgendorf, C.; Spahn-Langguth, H.; Wunderli-allenspach, H.; Merkle, H. P.; Langguth, P. *J. Pharm. Sci.* **1998**, *87*, 757.
22. Kansy, M.; Senner, F.; Gubernator, K. *J. Med. Chem.* **1998**, *41*, 1007.
23. Thompson, M.; Krull, U. J., Worsfold, P. J. *Anal. Chim. Acta* **1980**, *117*, 133.
24. Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. *J. Med. Chem.* **2002**, *45*, 2615.
25. Kerns, E. H.; Di, L.; Petuskey, S.; Farris, M.; Ley, R.; Jupp, P. *J. Pharm. Sci.* **2004**, *93*, 1440.
26. Wohnsland, F.; Faller, B. *J. Med. Chem.* **2001**, *44*, 923.
27. Avdeef, A.; Testa, B. *Cell. Mol. Life Sci.* **2002**, *59*, 1681.
28. Leo, A.; Hansch, C.; Elkins, D. *Chem. Rev.* **1971**, *71*, 525.
29. Hansch, C.; Dunn, W. J. *J. Pharm. Sci.* **1972**, *61*, 1.
30. Wils, P.; Warnery, A.; Phung-Ba, V.; Legrain S.; Scherman, D. *J. Pharmacol. Exp. Ther.* **1994**, *269*, 654.
31. Buur, A.; Trier, L.; Magnusson, C.; Artursson, P. *Int. J. Pharm.* **1996**, *129*, 223.
32. Raevsky, O. A.; Schaper, K.-J. *Eur. J. Med. Chem.* **1998**, *33*, 799.
33. Burton, P. S.; Conradi, R. A.; Hilgers, A. R.; Ho, N. F. H.; Maggiora, L. L. *J. Contr. Release* **1992**, *19*, 87.
34. Wright, L. L.; Painter, G. R. *Mol. Pharmacol.* **1992**, *41*, 957.
35. Palm, K.; Luthman, K.; Ungell, A.-L.; Strandlund, G.; Artursson, P. *Pharm. Sci.* **1996**, *85*, 32.
36. van de Waterbeemd, H.; Kansy, M. *Chimia* **1992**, *46*, 299.
37. Potts, R. O.; Guy, R. H. *Pharm. Res.* **1993**, *10*, 635.
38. Lipinski, C. A.; Lomardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **1997**, *23*, 3.
39. Verma, R. P.; Hansch, C.; Selassie, C. D. *J. Comput. Aided Mol. Des.* **2007,** *21*, 3.
40. (a) Guangli, M.; Yiyu, C. *J. Pharm. Pharmaceut. Sci.* **2006**, *9*, 210. (b) Fujikawa, M.; Nakao, K.; Shimizu, R.; Akamatsu, M. *Bioorg. Med. Chem*. **2007**, *15*, 3756.
41. Castillo-Garit, J. A.; Marrero-Ponce, Y.; Torrens, F.; García-Domenech, R. *J. Pharm. Sci.* **2008**, *97*, 1946.
42. www.codessa-pro.com
43. www.molcode.com
44. (a) Karelson, M. *Molecular Descriptors in QSAR/QSPR*. Wiley & Sons: New York, 2000. (b) Cronin, M.; Hewitt, M. *Compreh. Med. Chem.* **2006**, *5*, 725. (c) Bagchi, M.; Mills, D.; Basak, S. *J. Mol. Mod*. **2007**, *13*, 111. (d) Duchowicz, P.; Castro, E.; Toropov, A.; Benfenati, E. *Topics Heterocycl. Chem.* **2006**, *3*, 1. (e) Trinajstic, N.; Mladen, M.; Raic-Malic, S.; Raos, N. *Croat. Chim. Acta* **2007**, *80*, 1. (c) Todeschini, R; Consonni, V. *Molecular Descriptors in Chemometrics*, Wiley-VCH Weinheim, Ed. Manhold, R; Kubiniy, H. 2009. (d) Devillers, J.; Balaban, A. *Topological indices and related descriptors in QSAR and QSPR*, Gordon and Breach Sci. Publisher, 1999.
45. Schrödinger, MacroModel, version 9.5, Schrödinger, LLC, New York, NY, 2007.
46. Halgren, T. A. *J. Comput. Chem*. **1999**, *20*, 720.
47. MacroModel, version 9.5 User Manual, Schrödinger, LLC, New York, NY, 2007.

48. Stewart, J. J. P. MOPAC Program Package 6.0. QCPE No. 455, 1990.
49. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc*. **1985**, *107*, 3902.
50. Katritzky, A. R.; Lobanov, V.; Karelson, M. *J. Phys. Chem.* **1996**, *100*, 10400.
51. Goll, S.; Jurs, P. *J. Chem. Inf. Comput. Sci.* **1999,** *39*, 1081.
52. Tetteh, J.; Suzuki, T.; Metcalfe, E.; Howells, S. *J. Chem. Inf. Comput.Sci*. **1999**, *39*, 491.
53. Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: an Introduction*; VCH-Verlag: Weinheim, 1993; pp 213-228.
54. Burns, J. A.; Whitesides, G. *Chem. Rev.* **1993**, *93*, 2583.
55. Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2005**, *13*, 6598.
56. Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. *J. Med. Chem.* **2006**, *49*, 3305.
57. Karelson, M.; Dobchev, D. A.; Kulshyn, O. V.; Katritzky, A. *J. Chem. Inf. Mod.*, **2006**, *46*, 1897.
58. Haykin, S. *Neural Networks. A Comprehensive Foundation*; Pearson, 1999; pp 156-256.
59. Galbershtam, N.; Baskin, I.; Paljulin, V.; Zefirov, N. *Uspehi Himii*, **2003**, *72*, 706.