J Comput Aided Mol Des (2009) 23:309–319
DOI 10.1007/s10822-009-9261-8

QSAR application for the prediction of compound permeability with *in silico* descriptors in practical use

Kazuya Nakao · Masaaki Fujikawa · Ryo Shimizu · Miki Akamatsu

Received: 30 July 2008 / Accepted: 18 January 2009 / Published online: 25 February 2009

Abstract The parallel artificial membrane permeation assay (PAMPA) was developed as a model for the prediction of transcellular permeation in the process of drug absorption. In our previous report, it was revealed that PAMPA permeability is governed by log *P*, pK_a, and the hydrogen-bonding ability of compounds. In order to construct a new filtering method for selecting informative compounds from the whole combinatorial library, this study tried to predict PAMPA permeability with in silico descriptors. Log *P*, pK_a, and polar surface areas (PSA) as a hydrogen-bonding descriptor were calculated by commercially available or free-accessible web programs. Five-fold cross-validations and conventional regression analyses were examined with the training set for the entire 81 combinations with nine log *P*, three pK_a and three PSA descriptors. By comparison of statistical indices, four equations were selected and then the model with the best combination of in silico descriptors was determined based on the external validation. The PAMPA prediction equation obtained in this report could be applied for the prediction of both Caco-2 cell permeability and human intestinal absorption of mainly passively-transported drugs.

Keywords PAMPA · QSAR · Log *P* · pK_a · PSA · Caco-2

K. Nakao · R. Shimizu
Tanabe Seiyaku, Co. Ltd, Osaka 532-8505, Japan

M. Fujikawa · M. Akamatsu (✉)
Division of Environmental Science and Technology, Graduate School of Agriculture, Kyoto University, Kyoto 606-8502, Japan
e-mail: akamatsu@kais.kyoto-u.ac.jp

Introduction

In the early stage of drug discovery, a large number of diverse compounds are synthesized by the technique of combinatorial chemistry. However, this approach, of synthesizing large numbers of compounds designed from the combination of available building blocks and evaluating their activity in high throughput screening, is not always effective in finding new drugs. Filtering methods are needed for selecting only informative compounds from the entire compound library before their synthesis [1, 2].

In recent years, the evaluation of drug absorption in the early stage of drug discovery has been applied as one of the experimental filtering methods. Using various absorption models such as Caco-2 cells, MDCK cells, immobilized artificial membrane columns, and artificial membranes [3], compounds with good permeability can be selected as drug candidates.

The parallel artificial membrane permeation assay (PAMPA), developed by Kansy et al. [4] in 1998, is a model for the prediction of transcellular permeation in the process of drug absorption. The PAMPA is a high throughput system and has been nowadays automated by robotics. Many researchers have been improving the PAMPA system for the prediction of drug absorption, e.g. Bio-mimetic PAMPA [5–7] and Double-sink PAMPA [8], and challenging this system to predict passive blood-brain penetration [7] and human skin permeability [8]. When the permeation or absorption of compounds is evaluated by these systems, their data are collected and saved for the construction of prediction systems.

We have been interested in the PAMPA system as a permeation model through cell membranes, and the preceding study revealed that PAMPA permeability is governed by log *P* (*P*: 1-octanol/water partition coefficient), pK_a, and the

hydrogen-bonding ability of compounds [9, 10]. Log P could indicate that the permeation process of an undissociated compound form through artificial membranes is similar to the partitioning behavior between 1-octanol and water. The absolute value of the difference between the pK_a value and the experimental pH 7.3, |pK_a − pH|, is the descriptor that relates to the ratio between dissociated and undissociated molecules in an aqueous buffer solution. This descriptor could correct for the hydrophobicity of compounds existing in undissociated and various dissociated forms. The hydrogen-bonding ability of compounds was indicated by the polar surface area (PSA), which was calculated in a stable conformation (cPSA). The interaction between a compound and lecithin, a component of artificial membranes, differs from the interaction between the compound and 1-octanol because lecithin has more hydrogen-bonding sites than 1-octanol. The cPSA descriptor could correct for the effect of the difference in hydrogen-bonding ability between lecithin and 1-octanol on compound permeability.

Many commercially available programs can calculate the common physicochemical descriptors, log P, pK_a, and cPSA that are used for the prediction of PAMPA permeability. This study developed prediction models for the PAMPA permeability with various in silico descriptors to construct a new filtering method for selecting only useful compounds from compound libraries. By comparison of the statistical indices of each model, the optimal descriptor set suitable for the practical prediction of PAMPA permeability of compounds is discussed.

Method

The data set of 71 compounds included peptide related compounds, drugs, and other chemicals shown in Table 1. PAMPA permeability coefficients, log $P_{app\text{-}pampa}$, were taken from the preceding report [10]. As discussed previously, compounds with high hydrophobicity could encounter difficultly in passing through the unstirred water layer next to the membranes and also could be retained in the membranes [9, 10], so compounds with calculated log $P_{app\text{-}pampa}$ values greater than −4.5 were excluded from this analysis. The programs examined are shown in Table 2. Nine programs for the prediction of log P, three for pK_a, and three for PSA were tested. All PSA values, cPSA, TPSA, PSA/MNDO, and PSA/AM1 in Table 1, are shown as 1/100 squared angstrom unit. Each value of PSA/MNDO and PSA/AM1 was calculated in the geometry-optimized form of the molecule as reported in the preceding study in detail [10]. Values calculated by the programs are also listed in Table 1.

Seventy-one compounds were divided into two data sets; a training set with 60 compounds and a validated set with 11 compounds. In the preceding reports [9–11], log $P_{app\text{-}pampa}$ was predicted by three descriptors, measured log P (hereinafter the measured value of log P is indicated by log P_{oct}) from the database of MacLogP program version 4.0 [12], pK_a from the CQSAR database [13], and cPSA calculated by Sybyl MOLPROP [14]. The values of log P_{oct} ranged between −1.03 and 4.93, |pK_a − 7.3| between 0 and 4.32, and cPSA between 0.27 and 1.59. The 71 compounds were divided into two data sets so as both the training and the validated data could cover the whole range of each descriptor as evenly as possible [15].

In order to select the best combination of descriptors for prediction, five-fold cross-validations and conventional regression analyses were examined with the training set for the entire 81 combinations with nine log P, three pK_a and three PSA descriptors. For five-fold cross-validation, the training data of 60 compounds was divided into five groups containing the data of 12 compounds each (indicated by "Group" in Table 1). The data of each of the five groups covered the whole range of values of log P_{oct}, |pK_a − 7.3|, and cPSA as evenly as possible. The first QSAR equation was formulated by regression analysis for the permeability of 48 compounds except for the compounds in Group 1. Then, log $P_{app\text{-}pampa}$ values were predicted with the compounds in Group 1 using the obtained equation. Next, the second QSAR equation for 48 compounds in Groups 1 and 3–5 was formulated similarly, and the log $P_{app\text{-}pampa}$ values of compounds in Group 2 were predicted by the equation. This step was repeated and five equations and five sets of the corresponding predicted values were obtained. Then using the whole data of the measured and predicted values of 60 compounds, the squared values of five-fold cross-validated correlation coefficient (r^2_{cv}) and the root mean squared errors ($RMSE_{cv}$) were calculated. Statistical indices were also calculated for the conventional regression analyses with 60 compounds.

Several equations having good predictability for PAMPA permeability coefficients were selected based on the r^2_{cv} and $RMSE_{cv}$ values from the five-fold cross-validations. Using these equations, log $P_{app\text{-}pampa}$ values of 11 compounds in a validated set were predicted and the correlation coefficient for prediction (r^2_{pred}) and $RMSE_{pred}$ were calculated with the measured and the predicted values. Finally the best prediction equation was obtained by comparing with the r^2_{pred} and $RMSE_{pred}$ values.

Results and Discussion

A QSAR Eq. 1 was formulated for the entire data of 60 compounds in the training set using the experimentally determined log P_{oct} and pK_a values and cPSA.

Table 1 Permeability coefficients and QSAR descriptors

No.	Compound	Group	log $P_{app\text{-}pampa}$	log P_{oct}	CLOGP	KOWWIN	ACD/LogP	miLogP	AB/LogP	AlogP98	XLOGP
Training set											
1	Boc–Trp	2	−5.74	2.65	2.80	2.76	2.89	2.66	1.82	2.76	2.61
2	Cbz–Trp	3	−5.55	3.20	3.25	3.11	3.50	3.07	2.71	3.42	3.01
3	Fmoc–Trp	4	−4.89	4.93	5.08	4.71	5.33	4.83	4.20	4.78	4.58
4	Ac–Trp–NH_2	4	−5.72	0.42	−0.19	0.20	0.26	0.69	0.27	0.56	0.41
5	Gly–Trp–NH_2	5	−6.38	−0.48	−1.76	−0.87	−0.18	−0.04	−1.76	−0.40	−0.70
6	Phe–Trp–NH_2	5	−5.22	1.38	1.01	1.25	1.96	1.32	0.16	1.64	1.27
7	Trp–Ala–Val–NH_2	2	−6.62	0.40	0.26	0.04	0.48	0.08	−1.03	0.47	0.18
8	Ac–Trp–Val–NH_2	1	−6.18	0.73	0.48	0.69	0.63	0.91	0.67	0.95	0.76
9	Ac–D–Trp–Val–NH_2	3	−6.14	0.65	0.48	0.69	0.63	0.91	0.67	0.95	0.76
10	Ac–Tyr–Leu–NH_2	4	−6.54	0.32	0.35	0.64	0.25	0.81	0.45	0.80	1.06
11	Cyclo(–Trp–Tyr)	2	−6.17	1.11	1.47	1.00	1.00	1.64	1.97	2.25	1.95
12	Cyclo(–Trp–Trp)	2	−5.04	2.04	2.13	1.55	1.66	2.27	2.67	2.79	2.43
13	Tryptophol	4	−4.72	1.54	1.32	1.63	1.28	1.63	1.71	1.84	1.44
14	Tryptamine	3	−5.27	1.35	1.42	1.27	1.28	1.07	1.49	1.55	1.37
15	Indole-3-carboxylic acid	4	−6.10	1.99	2.13	1.94	1.99	1.66	2.51	1.75	1.71
16	Indole-3-acetic acid	4	−6.40	1.41	1.40	1.49	1.43	1.51	1.63	1.79	1.52
17	Acebutolol	1	−6.44	1.71	1.70	1.19	1.95	2.25	1.77	1.62	1.68
18	Acetaminophen	2	−6.04	0.51	0.49	0.27	0.34	0.68	0.23	0.71	0.45
19	Alprenolol	1	−4.94	2.89	2.65	2.81	2.88	2.58	2.94	2.64	2.84
20	Antipyrine	1	−5.54	0.23	0.20	0.59	0.27	1.40	0.54	1.61	1.03
21	Caffeine	3	−5.41	−0.07	−0.06	0.16	−0.13	0.06	−0.45	0.42	−0.50
22	Chloramphenicol	3	−5.41	1.14	1.28	0.92	1.02	0.73	0.82	1.03	0.69
23	Clonidine	5	−4.98	1.43	1.41	1.85	1.41	2.61	2.07	2.35	2.68
24	Corticosterone	1	−4.77	1.94	2.32	1.99	1.76	1.88	2.22	2.02	1.67
25	Dexamethasone	2	−5.37	2.01	1.75	1.72	1.87	2.06	1.80	1.71	1.14
26	Diltiazem	2	−4.72	2.80	3.65	2.79	3.63	3.34	2.83	3.09	2.84
27	Furosemide	2	−6.47	2.03	1.87	2.32	3.00	1.77	2.27	1.40	1.41
28	Hydrochlorothiazide	3	−6.69	−0.07	−0.40	−0.10	−0.07	−0.06	−0.38	−0.28	−0.47
29	Hydrocortisone	3	−5.45	1.61	1.70	1.62	1.43	1.62	1.60	1.28	0.52
30	Ibuprofen	2	−4.67	3.50	3.68	3.79	3.72	3.46	3.44	3.61	3.64
31	Ketoprofen	3	−5.55	3.12	2.76	3.00	2.81	3.62	2.54	3.36	3.22
32	Metoprolol	5	−5.10	1.88	1.35	1.69	1.79	1.97	1.72	1.76	1.63
33	Naproxen	4	−5.30	3.34	2.82	3.10	3.00	3.38	3.01	2.85	2.84
34	Norfloxacin	5	−6.71	−1.03	−0.99	−0.31	1.48	−0.69	−0.88	1.29	1.05
35	Oxprenolol	1	−4.83	2.10	2.09	1.83	2.29	2.07	2.26	2.23	2.02

Table 1 continued

No.	Compound	Group	log $P_{app\text{-}pampa}$	log P_{oct}	CLOGP	KOWWIN	ACD/LogP	miLogP	AB/LogP	AlogP98	XLOGP
36	Practolol	2	−5.98	0.79	0.75	0.53	0.60	1.03	0.81	0.75	1.02
37	Prednisolone	4	−5.47	1.62	1.38	1.40	1.49	1.60	1.56	1.26	1.02
38	Propranolol	5	−4.58	2.98	2.75	2.60	3.10	2.97	3.04	2.54	3.03
39	Ranitidine	2	−6.05	0.27	0.63	0.29	1.23	0.33	0.13	1.04	1.28
40	Salicylic acid	1	−5.92	2.26	2.19	2.24	2.06	1.87	2.04	1.22	2.43
41	Theophylline	1	−5.31	−0.02	−0.06	−0.39	−0.17	0.00	0.12	0.22	−0.80
42	Trimethoprim	4	−5.50	0.91	0.88	0.73	0.79	1.00	0.96	1.54	0.65
43	Aminopyrine	5	−4.76	1.00	0.57	0.60	0.76	1.43	1.11	1.38	0.83
44	Piroxicam	3	−4.96	1.98	1.89	2.58	1.71	2.06	2.39	0.50	2.65
45	Pirenzepine	4	−6.05	0.10	0.17	1.68	−0.08	1.34	0.64	1.04	0.58
46	2-Nitrophenol	5	−4.24	1.79	1.85	1.91	1.71	1.63	1.99	1.48	2.15
47	3-Nitroaniline	2	−4.67	1.37	1.26	1.47	1.37	0.95	1.44	0.98	1.10
48	3-Nitrophenol	5	−4.77	2.00	1.85	1.91	1.93	1.39	1.85	1.48	1.51
49	4-Cyanophenol	3	−4.77	1.60	1.60	1.61	1.60	1.21	1.45	1.47	1.34
50	4-Nitroaniline	3	−4.68	1.39	1.26	1.47	1.39	0.97	1.19	0.98	1.10
51	Aniline	1	−4.12	0.90	0.92	1.08	0.94	1.01	1.14	1.08	1.21
52	Diethyl phthalate	4	−4.82	2.47	2.62	2.65	2.70	2.31	2.55	2.24	2.74
53	Dimethyl phthalate	4	−4.36	1.56	1.56	1.66	1.64	1.55	1.57	1.54	1.89
54	Hydroquinone	5	−5.14	0.59	0.81	1.03	0.64	0.98	0.69	1.35	0.79
55	Phenol	1	−4.33	1.47	1.48	1.51	1.48	1.46	1.59	1.59	1.62
56	Atrazine	3	−4.53	2.61	2.50	2.82	2.63	2.55	2.52	2.54	1.66
57	DMTP	5	−4.97	2.50	2.77	1.58	2.03	1.79	0.81	2.73	1.21
58	Imidacloprid	1	−5.34	0.59	0.67	−0.41	−0.43	0.84	1.20	1.05	2.17
59	Chromafenozide	1	−4.78	2.70	4.80	4.40	4.51	4.09	5.53	4.98	5.65
60	RH-5849	5	−4.78	2.45	2.48	2.49	2.33	2.34	3.41	3.01	4.22
Validated set											
61	Trp–NH_2		−5.56	0.30	−0.15	−0.05	0.11	0.42	−1.09	0.54	0.13
62	Ac–Tyr–Phe–NH_2		−6.92	0.54	0.31	0.95	0.71	0.96	0.82	1.13	1.07
63	Cyclo(–Trp–Gly)		−6.02	0.07	−0.32	−0.64	−0.45	0.83	0.61	0.45	0.39
64	Indole-3-acetamide		−5.21	0.75	0.44	0.61	0.37	1.00	0.76	1.16	0.79
65	Indole-3-propionic acid		−5.68	1.75	1.89	2.35	1.76	2.03	1.79	2.24	1.69
66	Coumarin		−4.55	1.39	1.41	1.51	1.39	2.01	1.66	1.90	1.48
67	Labetalol		−5.18	3.09	2.50	2.41	2.31	2.85	2.24	2.36	2.52
68	Nadolol		−6.15	0.71	0.38	1.17	1.29	1.15	0.54	1.15	1.17
69	Pindolol		−5.31	1.75	1.67	1.48	1.97	1.98	2.15	1.93	1.92
70	Phenytoin		−4.41	2.26	2.08	2.16	2.52	2.18	1.96	2.30	2.22
71	2-Nitroaniline		−4.54	1.85	1.80	2.02	1.83	1.33	1.34	0.98	1.74

Table 1 continued

No.	Compound	ALOGPs	IA_logP	$\lvert pK_a - pH \rvert$	|ACD/pKa − pH|	|AB/pKa − pH|	|Marvin_pKa − pH|	cPSA	TPSA	PSA/MNDO	PSA/AM1
Training set											
1	Boc–Trp	2.79	1.69	3.60	3.30	5.7	2.97	0.995	0.914	0.529	0.635
2	Cbz–Trp	2.74	1.51	3.60	3.32	5.7	3.14	1.076	0.914	0.535	0.664
3	Fmoc–Trp	4.20	3.48	3.60	3.41	5.7	3.18	1.097	0.914	0.534	0.714
4	Ac–Trp–NH_2	0.77	0.20	0	0	0	0	0.907	0.880	0.544	0.747
5	Gly–Trp–NH_2	−0.25	−0.93	0	0	0.9	0	1.363	1.140	0.617	0.767
6	Phe–Trp–NH_2	1.36	1.74	0	0	0.2	0	1.345	1.140	0.626	0.839
7	Trp–Ala–Val–NH_2	0.46	1.85	0	0	0.2	0	1.593	1.431	0.670	0.855
8	Ac–Trp–Val–NH_2	0.57	0.83	0	0	0	0	1.320	1.171	0.649	0.832
9	Ac–D–Trp–Val–NH_2	0.57	0.83	0	0	0	0	1.184	1.171	0.578	0.774
10	Ac–Tyr–Leu–NH_2	0.31	−0.65	0	0	2.6	0	1.183	1.215	0.588	0.677
11	Cyclo(–Trp–Tyr)	1.98	1.05	0	0	0	0	1.201	0.942	0.520	0.610
12	Cyclo(–Trp–Trp)	2.23	1.09	0	0	0	0	1.119	0.898	0.455	0.609
13	Tryptophol	1.82	1.51	0	0	0	0	0.491	0.360	0.256	0.410
14	Tryptamine	1.21	1.53	2.90	3.27	1.5	2.43	0.551	0.418	0.163	0.274
15	Indole-3-carboxylic acid	1.79	2.26	3.10	3.24	2.2	3.78	0.689	0.531	0.303	0.326
16	Indole-3-acetic acid	1.87	1.54	3.00	2.81	2.7	2.64	0.641	0.531	0.319	0.438
17	Acebutolol	1.43	1.44	2.11	1.80	2.3	2.27	1.006	0.877	0.435	0.493
18	Acetaminophen	0.51	0.65	0	0	0	0	0.624	0.493	0.323	0.371
19	Alprenolol	2.59	2.81	2.30	1.86	2.3	2.37	0.493	0.415	0.299	0.320
20	Antipyrine	1.01	1.21	0	0	0	0	0.279	0.269	0.173	0.218
21	Caffeine	−0.23	−0.06	0	0	0	0	0.695	0.618	0.369	0.401
22	Chloramphenicol	1.15	0.65	0	0	0	0	1.181	1.127	0.697	0.715
23	Clonidine	2.56	2.08	0.75	0.80	0.6	0.86	0.468	0.364	0.730	0.845
24	Corticosterone	2.09	1.94	0	0	0	0	0.777	0.746	0.487	0.504
25	Dexamethasone	1.93	1.82	0	0	0	0	1.007	0.948	0.775	0.818
26	Diltiazem	3.09	3.08	0.76	1.64	0.9	0.88	0.783	0.591	0.434	0.515
27	Furosemide	2.71	1.69	3.96	4.26	3.8	3.05	1.443	1.226	0.711	0.650
28	Hydrochlorothiazide	−0.15	−0.23	0	0	0	0	1.525	1.184	1.162	1.053
29	Hydrocortisone	1.79	1.67	0	0	0	0	0.951	0.948	0.591	0.596
30	Ibuprofen	3.50	3.54	2.92	2.89	3.0	2.45	0.430	0.373	0.220	0.239
31	Ketoprofen	3.30	3.07	3.01	3.07	3.0	3.52	0.615	0.544	0.330	0.332
32	Metoprolol	1.80	1.91	2.45	1.87	2.3	2.37	0.622	0.507	0.372	0.392
33	Naproxen	3.29	3.11	3.29	2.46	3.0	3.11	0.532	0.465	0.289	0.307
34	Norfloxacin	−0.05	−0.34	2.08	1.34	2.7	2.91	0.808	0.746	0.716	0.670
35	Oxprenolol	2.44	2.07	2.30	1.87	2.3	2.37	0.624	0.507	0.376	0.384
36	Practolol	0.53	0.84	2.10	1.86	2.3	2.37	0.858	0.706	0.490	0.546
37	Prednisolone	1.66	1.27	0	0	0	0	0.982	0.948	0.611	0.603
38	Propranolol	3.03	2.96	2.30	1.84	2.3	2.37	0.483	0.415	0.251	0.260

Table 1 continued

No.	Compound	ALOGPs	IA_logP	$\lvert pK_a - pH \rvert$	|ACD/pKa − pH|	|AB/pKa − pH|	|Marvin_pKa − pH|	cPSA	TPSA	PSA/MNDO	PSA/AM1
39	Ranitidine	0.79	0.25	0.88	1.10	1.2	0.78	1.095	0.836	0.488	0.546
40	Salicylic acid	1.96	1.93	4.32	4.29	4.3	4.51	0.730	0.575	0.330	0.345
41	Theophylline	−0.25	−0.20	0	0	0	0	0.746	0.727	0.612	0.670
42	Trimethoprim	1.26	1.21	0	0	0	0	1.219	1.055	0.597	0.697
43	Aminopyrine	1.13	1.27	0	0	0	0	0.266	0.302	0.218	0.267
44	Piroxicam	2.20	2.44	2.23	0	2.1	5.51	1.045	0.996	0.438	0.437
45	Pirenzepine	1.26	0.71	0.70	0	0.8	0.29	0.765	0.688	0.438	0.496
46	2-Nitrophenol	1.91	1.78	0.08	0.16	0	0.87	0.699	0.661	0.261	0.250
47	3-Nitroaniline	1.53	1.35	0	0	0	0	0.821	0.718	0.422	0.458
48	3-Nitrophenol	1.92	1.82	0	0	0	0	0.743	0.661	0.419	0.449
49	4-Cyanophenol	1.68	1.50	0	0	0	0	0.524	0.440	0.241	0.254
50	4-Nitroaniline	1.50	1.32	0	0	0	0	0.823	0.718	0.432	0.483
51	Aniline	0.89	1.07	0	0	0	0	0.335	0.260	0.097	0.186
52	Diethyl phthalate	2.60	2.63	0	0	0	0	0.557	0.526	0.338	0.367
53	Dimethyl phthalate	1.96	1.80	0	0	0	0	0.563	0.526	0.286	0.301
54	Hydroquinone	0.71	0.65	0	0	0	0	0.504	0.405	0.254	0.276
55	Phenol	1.39	1.39	0	0	0	0	0.280	0.202	0.124	0.133
56	Atrazine	2.70	2.80	0	0	0	0	0.842	0.627	0.320	0.364
57	DMTP	0.54	2.15	0	0	0	0	1.331	0.626	0.875	1.164
58	Imidacloprid	0.62	1.51	0	0	0	0	0.938	0.863	0.882	0.901
59	Chromafenozide	4.16	4.33	0	0	0	0	0.679	0.586	0.353	0.448
60	RH-5849	2.40	2.57	0	0	0	0	0.561	0.494	0.319	0.350
Validated set											
61	Trp–NH_2	0.35	−0.33	0.20	0.88	0.2	0	0.982	0.849	0.389	0.543
62	Ac–Tyr–Phe–NH_2	0.61	0.35	0	0	2.6	0	1.381	1.215	0.691	0.802
63	Cyclo(–Trp–Gly)	0.69	−0.34	0	0	0	0	0.961	0.740	0.425	0.518
64	Indole-3-acetamide	1.17	0.87	0	0	0	0	0.693	0.589	0.351	0.537
65	Indole-3-propionic acid	2.04	1.64	2.60	2.53	2.7	2.50	0.656	0.531	0.321	0.396
66	Coumarin	1.66	2.21	0	0	0	0	0.330	0.302	0.173	0.192
67	Labetalol	1.73	0.77	2.10	1.90	1.3	0	1.085	0.956	0.498	0.548
68	Nadolol	1.23	0.98	2.37	1.87	2.1	2.46	0.957	0.820	0.496	0.505
69	Pindolol	2.17	1.69	2.24	1.90	2.3	2.37	0.670	0.573	0.307	0.469
70	Phenytoin	2.26	2.58	0	0	0	0	0.708	0.582	0.311	0.338
71	2-Nitroaniline	1.43	1.35	0	0	0	0	0.776	0.718	0.325	0.339

Table 2 *In silico* prediction programs

	Program	Method	Web Site	Reference
log *P*	CLOGP	Fragment-based	www.biobyte.com	[17]
	KOWWIN	Fragment-based	www.vcclab.org, www.syrres.com	[18]
	ACD/LogP	Fragment-based	www.acdlabs.com	[19]
	miLogP	Fragment-based	www.molinspiration.com	[20]
	AB/LogP	Fragment-based	www.ap-algorithms.com,www.vcclab.org	[21]
	AlogP98	Atom-based	www.accelrys.com	[22]
	XLOGP	Atom-based	http://physbio.mssm.edu/docs/molmod /xlogp/ [a]	[23]
	ALOGPs	Descriptor-based	www.vcclab.org	[24, 25]
	IA_logP	Descriptor-based	www.logp.com [a]	[26]
pK_a	ACD/pKa	Hammett-σ-based	www.acdlabs.com	[27]
	AB/pKa	Hammett-σ-based	www.ap-algorithms.com	[28]
	Marvin_pKa	Charge-based	www.chemaxon.com	[29]
PSA	TPSA	Fragment-based	www.molinspiration.com	[30]
	PSA/MNDO	Electrostatic potential	www.wavefun.com	[31]
	PSA/AM1	Electrostatic potential	www.wavefun.com	[31]

[a] On-line version of the algorithms was able to be accessed at this web site in 2006

$$\log P_{\text{app-pampa}} = 0.430(0.099)\log P_{\text{oct}} - 0.292(0.078)\,|\text{p}K_a - \text{pH}| - 1.059(0.290)\,\text{cPSA} - 4.862(0.311) \quad (1)$$

$n = 60$, SD = 0.359, $r^2 = 0.745$, $F_{3,56} = 54.591$, $r^2_{cv} = 0.712$, $\text{RMSE}_{cv} = 0.368$,

where n is the number of compounds, SD is the standard deviation, r is the correlation coefficient, F is the ratio between regression and residual variances, and the figures in parentheses are the 95% confidence intervals. The values of the statistical indices, the r^2 and SD for the conventional regression analysis as well as r^2_{cv} and RMSE_{cv}, for five-fold cross-validation of every 81 combination with the corresponding values of Eq. 1 are shown in Table 3.

Since five-fold cross-validation is a useful method for evaluating predictability of equations within the range of training set, the following four Eqs. 2–5 with $r^2_{cv} \geq 0.60$ and $\text{RMSE}_{cv} \leq 0.45$ were firstly selected as the good prediction equations for $\log P_{\text{app-pampa}}$ values. These equations also had high statistical quality in terms of conventional r^2 and SD.

$$\log P_{\text{app-pampa}} = 0.325(0.092)\,\text{CLOGP} - 0.275(0.085)\,|\text{ACD/pKa} - \text{pH}| - 1.202(0.362)\,\text{TPSA} - 4.759(0.333) \quad (2)$$

$n = 60$, SD = 0.389, $r^2 = 0.699$, $F_{3,56} = 43.428$, $r^2_{cv} = 0.653$, $\text{RMSE}_{cv} = 0.404$

$$\log P_{\text{app-pampa}} = 0.348(0.107)\,\text{KOWWIN} - 0.278(0.090)\,|\text{ACD/pKa} - \text{pH}| - 1.174(0.378)\,\text{TPSA} - 4.818(0.359) \quad (3)$$

$n = 60$, SD = 0.404, $r^2 = 0.677$, $F_{3,56} = 39.199$, $r^2_{cv} = 0.620$, $\text{RMSE}_{cv} = 0.423$

Table 3 Correlation coefficient (r^2), standard deviation (SD), cross-validated r^2 (r^2_{cv}) and $RMSE_{cv}$ of equations with every combination of in silico descriptors

No	Log *P*	pK_a	PSA	SD	r^2	$RMSE_{cv}$	r^2_{cv}
1	Log Poct	pKa	cPSA	0.359	0.745	0.368	0.712
2	CLOGP	ACD/pKa	TPSA	0.389	0.699	0.404	0.653
3	CLOGP	ACD/pKa	PSA/MNDO	0.434	0.627	0.453	0.564
4	CLOGP	ACD/pKa	PSA/AM1	0.430	0.633	0.458	0.555
5	CLOGP	AB/pKa	TPSA	0.422	0.647	0.438	0.592
6	CLOGP	AB/pKa	PSA/MNDO	0.445	0.607	0.465	0.541
7	CLOGP	AB/pKa	PSA/AM1	0.446	0.606	0.471	0.528
8	CLOGP	Marvin_pKa	TPSA	0.442	0.613	0.472	0.527
9	CLOGP	Marvin_pKa	PSA/MNDO	0.476	0.550	0.505	0.459
10	CLOGP	Marvin_pKa	PSA/AM1	0.469	0.564	0.507	0.453
11	KOWWIN	ACD/pKa	TPSA	0.404	0.677	0.423	0.620
12	KOWWIN	ACD/pKa	PSA/MNDO	0.458	0.583	0.488	0.494
13	KOWWIN	ACD/pKa	PSA/AM1	0.461	0.580	0.497	0.475
14	KOWWIN	AB/pKa	TPSA	0.422	0.648	0.445	0.579
15	KOWWIN	AB/pKa	PSA/MNDO	0.454	0.592	0.482	0.507
16	KOWWIN	AB/pKa	PSA/AM1	0.457	0.586	0.491	0.487
17	KOWWIN	Marvin_pKa	TPSA	0.442	0.612	0.473	0.526
18	KOWWIN	Marvin_pKa	PSA/MNDO	0.486	0.532	0.518	0.431
19	KOWWIN	Marvin_pKa	PSA/AM1	0.482	0.539	0.523	0.419
20	ACD/LogP	ACD/pKa	TPSA	0.403	0.678	0.427	0.613
21	ACD/LogP	ACD/pKa	PSA/MNDO	0.458	0.584	0.492	0.486
22	ACD/LogP	ACD/pKa	PSA/AM1	0.459	0.582	0.497	0.476
23	ACD/LogP	AB/pKa	TPSA	0.433	0.628	0.456	0.558
24	ACD/LogP	AB/pKa	PSA/MNDO	0.462	0.577	0.491	0.488
25	ACD/LogP	AB/pKa	PSA/AM1	0.466	0.569	0.499	0.471
26	ACD/LogP	Marvin_pKa	TPSA	0.461	0.579	0.505	0.458
27	ACD/LogP	Marvin_pKa	PSA/MNDO	0.502	0.500	0.543	0.375
28	ACD/LogP	Marvin_pKa	PSA/AM1	0.498	0.508	0.546	0.366
29	miLogP	ACD/pKa	TPSA	0.437	0.621	0.450	0.569
30	miLogP	ACD/pKa	PSA/MNDO	0.475	0.553	0.496	0.478
31	miLogP	ACD/pKa	PSA/AM1	0.471	0.561	0.500	0.470
32	miLogP	AB/pKa	TPSA	0.454	0.592	0.468	0.534
33	miLogP	AB/pKa	PSA/MNDO	0.473	0.557	0.491	0.487
34	miLogP	AB/pKa	PSA/AM1	0.471	0.560	0.496	0.477
35	miLogP	Marvin_pKa	TPSA	0.475	0.553	0.498	0.474
36	miLogP	Marvin_pKa	PSA/MNDO	0.505	0.494	0.530	0.404
37	miLogP	Marvin_pKa	PSA/AM1	0.497	0.511	0.531	0.401
38	AB/LogP	ACD/pKa	TPSA	0.427	0.639	0.445	0.580
39	AB/LogP	ACD/pKa	PSA/MNDO	0.465	0.572	0.488	0.495
40	AB/LogP	ACD/pKa	PSA/AM1	0.472	0.558	0.509	0.450
41	AB/LogP	AB/pKa	TPSA	0.455	0.591	0.471	0.528
42	AB/LogP	AB/pKa	PSA/MNDO	0.477	0.548	0.498	0.474
43	AB/LogP	AB/pKa	PSA/AM1	0.486	0.532	0.519	0.429
44	AB/LogP	Marvin_pKa	TPSA	0.458	0.585	0.486	0.498
45	AB/LogP	Marvin_pKa	PSA/MNDO	0.488	0.527	0.517	0.432
46	AB/LogP	Marvin_pKa	PSA/AM1	0.491	0.522	0.534	0.394
47	AlogP98	ACD/pKa	TPSA	0.471	0.561	0.486	0.498
48	AlogP98	ACD/pKa	PSA/MNDO	0.494	0.516	0.516	0.435

Table 3 continued

No	Log *P*	pK_a	PSA	SD	r^2	$RMSE_{cv}$	r^2_{cv}
49	AlogP98	ACD/pKa	PSA/AM1	0.485	0.534	0.512	0.444
50	AlogP98	AB/pKa	TPSA	0.487	0.530	0.503	0.463
51	AlogP98	AB/pKa	PSA/MNDO	0.497	0.511	0.517	0.433
52	AlogP98	AB/pKa	PSA/AM1	0.491	0.523	0.515	0.437
53	AlogP98	Marvin_pKa	TPSA	0.508	0.488	0.537	0.388
54	AlogP98	Marvin_pKa	PSA/MNDO	0.530	0.443	0.560	0.333
55	AlogP98	Marvin_pKa	PSA/AM1	0.520	0.464	0.558	0.338
56	XLOGP	ACD/pKa	TPSA	0.461	0.578	0.481	0.509
57	XLOGP	ACD/pKa	PSA/MNDO	0.497	0.510	0.516	0.435
58	XLOGP	ACD/pKa	PSA/AM1	0.499	0.505	0.527	0.410
59	XLOGP	AB/pKa	TPSA	0.473	0.556	0.502	0.465
60	XLOGP	AB/pKa	PSA/MNDO	0.493	0.519	0.519	0.428
61	XLOGP	AB/pKa	PSA/AM1	0.496	0.512	0.532	0.399
62	XLOGP	Marvin_pKa	TPSA	0.483	0.537	0.511	0.445
63	XLOGP	Marvin_pKa	PSA/MNDO	0.512	0.481	0.535	0.393
64	XLOGP	Marvin_pKa	PSA/AM1	0.508	0.488	0.543	0.375
65	ALOGPs	ACD/pKa	TPSA	0.412	0.663	0.427	0.613
66	ALOGPs	ACD/pKa	PSA/MNDO	0.467	0.568	0.489	0.492
67	ALOGPs	ACD/pKa	PSA/AM1	0.473	0.557	0.506	0.456
68	ALOGPs	AB/pKa	TPSA	0.442	0.613	0.460	0.552
69	ALOGPs	AB/pKa	PSA/MNDO	0.475	0.553	0.496	0.478
70	ALOGPs	AB/pKa	PSA/AM1	0.482	0.539	0.513	0.441
71	ALOGPs	Marvin_pKa	TPSA	0.456	0.588	0.488	0.494
72	ALOGPs	Marvin_pKa	PSA/MNDO	0.499	0.506	0.529	0.406
73	ALOGPs	Marvin_pKa	PSA/AM1	0.500	0.505	0.541	0.379
74	IA_logP	ACD/pKa	TPSA	0.426	0.640	0.440	0.588
75	IA_logP	ACD/pKa	PSA/MNDO	0.457	0.586	0.474	0.523
76	IA_logP	ACD/pKa	PSA/AM1	0.458	0.584	0.484	0.503
77	IA_logP	AB/pKa	TPSA	0.462	0.577	0.476	0.520
78	IA_logP	AB/pKa	PSA/MNDO	0.481	0.542	0.494	0.482
79	IA_logP	AB/pKa	PSA/AM1	0.484	0.535	0.506	0.456
80	IA_logP	Marvin_pKa	TPSA	0.451	0.597	0.482	0.506
81	IA_logP	Marvin_pKa	PSA/MNDO	0.475	0.553	0.500	0.470
82	IA_logP	Marvin_pKa	PSA/AM1	0.471	0.561	0.504	0.460

$$\log P_{\text{app-pampa}} = 0.350(0.107)\,\text{ACD/LogP} - 0.314(0.095)\,|\text{ACD/pKa} - \text{pH}| - 1.311(0.369)\,\text{TPSA} - 4.714(0.342) \quad (4)$$

$n = 60$, SD = 0.403, $r^2 = 0.678$, $F_{3,56} = 39.332$, $r^2_{cv} = 0.613$, $RMSE_{cv} = 0.427$

$$\log P_{\text{app-pampa}} = 0.363(0.117)\ \text{ALOGPs} - 0.274(0.092)\,||\text{ACD/pKa} - \text{pH}|| - 1.207(0.384)\,\text{TPSA} - 4.860(0.379) \quad (5)$$

$n = 60$, SD = 0.412, $r^2 = 0.663$, $F_{3,56} = 36.722$, $r^2_{cv} = 0.613$, $RMSE_{cv} = 0.427$

In these equations, CLOGP, KOWWIN, ACD/LogP, and ALOGPs as log *P*, ACD/pKa as pK_a, and TPSA as PSA are used. Eq. 2 by CLOGP was slightly better than the other Eqs. 3–5, but was not as good as Eq. 1 which was derived based on experimentally determined log P_{oct} and pK_a.

Regarding log *P* calculation, there are three types of calculation methods of log *P*, fragment-based, atom-based, and molecular descriptor-based approaches. The three results (2–4) for the prediction of log $P_{\text{app-pampa}}$ were obtained by the fragment-based approach, which is considered to be the most reliable in general [16]. In the fragment-based approach, the log *P* value is calculated by addition of not only the hydrophobicity of fragments but also the physicochemical interaction terms between

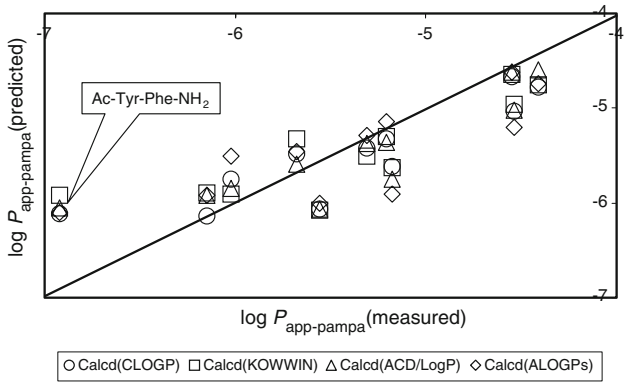

Fig. 1 Correlation between measured and predicted log $P_{\text{app-pampa}}$ values

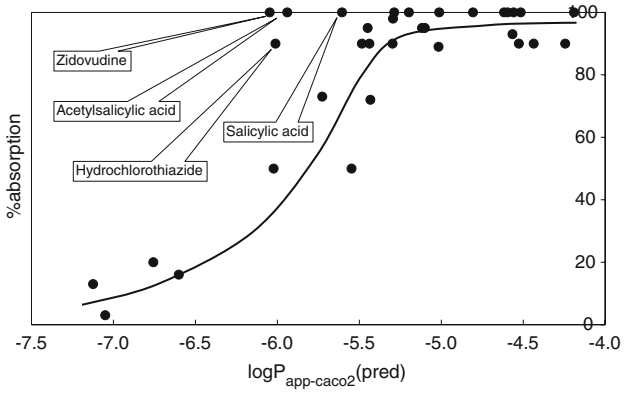

Fig. 2 Correlations between predicted Caco-2 permeation and percent oral absorption

fragments. Consequently, the fragment-based approach showed better results for predicting log $P_{\text{app-pampa}}$ than simple atom-based or molecular descriptor-based approaches for calculating log P, which is the only summation of atomic hydrophobicity or molecular descriptors, respectively.

ACD/pKa is one of the Hammett-σ-based approaches. In this approach, pK_a is calculated by combining the resonance and induced effects of functional groups on the ionization center. Both effects were indicated by empirical Hammett σ values. In the partial-charge-based pK_a estimations, each microscopic pK_a in all ionized forms is calculated by regression equations with the partial charge and then all the microscopic pK_a values are combined into the macroscopic pK_a. The pK_a calculation method using Hammett σ seemed to be more suitable for predicting log $P_{\text{app-pampa}}$ than the partial charge method.

Regarding PSA calculation, TPSA (topological PSA), which calculates PSA by adding the surface area of each polar atom type, and PSA weighted by the electrostatic potential (PSA/MNDO, PSA/AM1) were examined. The interaction between a ligand and lecithin membranes differs from that between the ligand and 1-octanol. In case of predicting log $P_{\text{app-pampa}}$, this difference seemed to be estimated by TPSA better than the others using the electrostatic potential.

The log $P_{\text{app-pampa}}$ values of 11 compounds for validation were calculated by the Eqs. 2–5, and the correlations of the observed and predicted log $P_{\text{app-pampa}}$ values of 11 compounds are shown in Fig. 1. In all calculations, the compounds with high log $P_{\text{app-pampa}}$ values were slightly underestimated, and compounds with low log $P_{\text{app-pampa}}$ values were overestimated. Particularly Ac–Tyr–Phe–NH_2, which had the lowest log $P_{\text{app-pampa}}$ value, was overestimated. The log $P_{\text{app-pampa}}$ value of Ac–Tyr–Phe–NH_2, −6.92, is below the range of the training data. Prediction systems cannot usually work well in extrapolative regions. When using the log $P_{\text{app-pampa}}$ prediction system, users should be careful with predicted values in extrapolative regions. The r^2_{pred} and $\text{RMSE}_{\text{pred}}$ were 0.724 and 0.385 in Eq. 2, 0.656 and 0.430 in Eq. 3, 0.710 and 0.394 in Eq. 4, and 0.606 and 0.460 in Eq. 5. Based on the r^2_{pred} and $\text{RMSE}_{\text{pred}}$ values, the Eq. 2 was selected as the best model for prediction of PAMPA permeability.

In our previous study [10], it was revealed that Caco-2 cell permeability and PAMPA permeability were well correlated with each other. With 35 compounds (7 peptide-related compounds, 20 commercial drugs, and 8 chemicals), the following Eq. 6 was obtained.

$$\log P_{\text{app-caco-2}} = 1.03(0.21)\ \log P_{\text{app-pampa}} + 0.49(1.09) \quad (6)$$

$n = 35$, SD $= 0.35$, $r^2 = 0.76$

In addition, human oral absorption was reported to have a sigmoidal relationship with Caco-2 cell permeability [32]. In order to examine the ability of in silico descriptors for predicting human oral absorption, Caco-2 cell permeability was calculated by Eqs. 2 and 6 using calculated log P, pK_a, and PSA and the obtained Caco-2 permeability values were plotted against percent oral absorption values which were reported by Yazdanian et al. [33]. The result is shown in Fig. 2. Almost all compounds were plotted around the sigmoidal curve as expected. Even if the compounds interact with various active transporters, their main absorption route should be passive permeation. The percent oral absorption of only four drugs, salicylic acid, acetylsalicylic acid, hydrochlorothiazide, and zidovudine, was underestimated. This might be because these compounds could be mainly absorbed through other routes. It was reported that salicylic acid, one of metabolites of acetylsalicylic acid, was actively transported by MCT-1 [34]. Zidovudine and hydrochlorothiazide could be also transported by transporter proteins in the gastrointestinal tract, but there is no clear evidence for this at present. Even if

such a limitation exists, it was indicated that in silico descriptors are useful for the prediction of not only Caco-2 cell permeability but also human oral absorption.

Conclusion

This study showed that the combination of CLOGP, ACD/pKa, and TPSA could accurately predict the permeability of compounds. Using this in silico permeability prediction system, it is possible to estimate the permeability of compounds before they are synthesized. As a result, compounds with unfavorable physicochemical properties as drug candidates can be effectively removed. However, it should be emphasized that the measured values were always better than the predicted values; therefore, continued efforts are required to obtain experimental data as well as to predict virtual data. The physicochemical properties of at least a small number of compounds should be measured and then compared with the predicted values, and the latter should be corrected if necessary. The prediction procedure reported here along with real measured data should allow the more effective selection of lead compounds of drug candidates.

Acknowledgments The authors are grateful to Dr. Hiroshi Ohmizu in Tanabe Seiyaku, Co., Ltd for his stimulating discussions and continuing encouragement. The authors would like to thank Mr. Masaaki Asao in Tanabe Seiyaku, Co., Ltd for his skillful collaboration of statistical analyses. The authors also appreciate to the reviewer for his/her helpful advices and fruitful suggestions.

References

1. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997) Adv Drug Deliv Rev 23:3. doi:10.1016/S0169-409X(96)00423-1
2. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD (2002) J Med Chem 45:2615. doi:10.1021/jm020017n
3. Lin J, Sahakian DC, de Morais SMF, Xu JJ, Polzer RJ, Winter SM (2003) Curr Top Med Chem 3:1125. doi:10.2174/1568026033452096
4. Kansy M, Senner F, Gubernator K (1998) J Med Chem 41:1007. doi:10.1021/jm970530e
5. Sugano K, Hamada H, Machida M, Ushio H (2001) J Biomol Screen 6:189. doi:10.1177/108705710100600309
6. Sugano K, Takata N, Machida M, Saitoh K, Terada K (2002) Int J Pharm 241:241. doi:10.1016/S0378-5173(02)00240-5
7. Sugano K, Nabuchi Y, Machida M, Aso Y (2003) Int J Pharm 257:245. doi:10.1016/S0378-5173(03)00161-3
8. Avdeef A (2003) Absorption and drug development. Wiley-Interscience, New Jersey
9. Fujikawa M, Ano R, Nakao K, Shimizu R, Akamatsu M (2005) Bioorg Med Chem 13:4721. doi:10.1016/j.bmc.2005.04.076
10. Fujikawa M, Nakao K, Shimizu R, Akamatsu M (2007) Bioorg Med Chem 15:3756. doi:10.1016/j.bmc.2007.03.040
11. Ano R, Kimura Y, Shima M, Matsuno R, Ueno T, Akamatsu M (2004) Bioorg Med Chem 12:257. doi:10.1016/j.bmc.2003.10.002
12. MacLogP 4.0; Biobyte Corp., Claremont
13. CQSAR 4.31; Biobyte Corp., Claremont
14. SYBYL Molecular Modeling Software; Tripos Associates Inc., St Louis
15. Leonard JT, Roy K (2006) QSAR Comb Sci 25:235. doi:10.1002/qsar.200510161
16. Mannhold R, Dross K (1996) Quant Struct Act Relat 15:403. doi:10.1002/qsar.19960150506
17. Leo AJ, Hoekman D (2000) Perspect Drug Discov Des 18:19 (17. CLOGP ver.4)
18. Meylan WM, Howard PH (1995) J Pharm Sci 84:83
19. Petrauskas AA, Kolovanov EA (2000) Perspect Drug Discov Des 19:99 (ACD/LogP ver. 10)
20. miLogP ver. 2.2.; Molinspiration Cheminformatics, Slovak Republic
21. Japertas P, Didziapetris R, Petrauskas A (2002) Quant Struct Act Relat 21:23
22. Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) J Phys Chem A 102:3762
23. Wang R, Fu Y, Lai L (1997) J Chem Inf Comput Sci 37:615 (XLOGP ver.1)
24. Tetko IV, Tanchuk VY (2002) J Chem Inf Comput Sci 42:1136 (ALOGPs ver. 2.1)
25. Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, Palyulin VA, Radchenko EV, Zefirov NS, Makarenko AS, Tanchuk VY, Prokopenko VV (2005) J Comput Aided Mol Des 19:453. doi:10.1007/s10822-005-8694-y
26. Tetko IV (2005) Drug Discov Today 10:1497
27. Meloun M, Bordovska S (2007) Anal Bioanal Chem 389:1267 (ACD/pKa ver.9.04)
28. AB/pKa in Virtual Computational Chemistry Laboratory (VCCLAB, http://www.vcclab.org, 2005)
29. Csizmadia F, Szegezdi J, Darvas F (1993) In: Wermuth CG (ed) Trends in QSAR and Molecular Modeling 92. ESCOM, Leiden (Marvin_pKa ver.4.1), p 507
30. Ertl P, Rohde B, Selzer P (2000) J Med Chem 43:3714. doi:10.1021/jm000942e
31. Hehre WJ (2003) A guide to molecular mechanics and quantum chemical calculations. Wavefunction Inc., Irvine, p 72
32. Stenberg P, Norinder U, Luthman K, Artursson P (2001) J Med Chem 44:1927. doi:10.1021/jm001101a
33. Yazdanian M, Glynn SL, Wright JL, Hawi A (1998) Phram Res 15:1490. doi:10.1023/A:1011930411574
34. Morita N, Kusuhara H, Sekine T, Endou S, Sugiyama Y (2001) J Pharmacol Exp Ther 298:1179