Toward minimalistic modeling of oral drug absorption

Tudor I. Oprea and Johan Gottfries

Medicinal Chemistry, AstraZeneca R&D Mölndal, S-43183 Mölndal, Sweden

Poor intestinal permeability of drugs constitutes a major bottleneck in the successful development of candidate drugs. Fast computational tools to help in designing compounds with increased probability of oral absorption are required, since both medicinal and combinatorial chemists are under pressure to consider increasing numbers of virtual and existing compounds. The QSAR paradigm for drug absorption is expressed as a function of molecular size, hydrogen-bonding capacity, and lipophilicity. A nonlinear PLS model that can be achieved with minimal computational efforts is described. The QSAR model correlates human intestinal absorption (%HIA) data, and apparent Caco-2 cell permeability data, to parameters calculated from molecular structures. Two properties were found to be relevant for absorption predictions, namely H-bonding capacity, and hydrophobic transferability. The parsimony principle was applied in several aspects: single conformers were used to compute molecular surface areas; the definitions of "polar" and "nonpolar" surfaces were done in a simplistic fashion; simple and fast 2D descriptors were used to estimate other properties; the 1 PLS component model was selected. These choices result in a minimalistic model for oral absorption. The use of both %HIA and Caco-2 permeability data was found to stabilize and improve the model. This QSAR model can serve as a simple, quantitative extension of the "rule of five" scheme (Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. Adv. Drug Deliv. Rev. 1997, ***23,*** *3–25), in a manner that can prove beneficial to the drug discovery process. © 2000 by Elsevier Science Inc.*

Keywords: Caco-2, drug absorption, hydrogen bonding, intestinal absorption, lipophilicity, mannitol, nonlinear model, permeability, PLS, polar surface area, QSAR, sulfasalazine

INTRODUCTION

Oral absorption of drugs is a major drive in the pharmaceutical industry. Because poor absorption characteristics constitute a bottleneck in the successful development of candidate drugs, understanding which properties need to be optimized in order to enhance oral absorption has become the subject of early-stage preclinical research, e.g., in medicinal and/or combinatorial chemistry settings. Medicinal and combinatorial chemistry facilities are under pressure to handle increasingly larger amounts of both virtual and existing compounds, while increasing their probability for good oral absorption and permeation. Fast computational tools that meet these requirements are currently under development.

One probability scheme in particular has proved to be extremely efficient in screening out compounds with potential problems.[1] Known as the "rule of five" probability scheme, it states that poor absorption or permeation is more probable when the molecular weight (MW) is over 500, when the calculated octanol/water partition coefficient (CLOGP) is over 5, when there are more than 5 hydrogen bond donors (HDO), and when the sum of nitrogen and oxygen atoms is above 10 (or hydrogen bond acceptors, HAC). Any pairwise combination of the following conditions: MW > 500, CLOGP > 5, HDO > 5, and HAC > 10, translates to poor permeability.[1] Despite its usefulness in assessing combinatorial and/or HTS (high-throughput screening) libraries, the "rule of five" does not estimate oral absorption in a quantitative manner. Furthermore, compounds that do not violate any of the "rule of five" criteria are not necessarily orally available. Therefore, there is a stringent need for quantitative models of oral absorption.[2–6]

The current QSAR paradigm for structure–permeability correlations, used to evaluate oral absorption, has been summarized[2] by Van de Waterbeemd et al. as follows:

$$\text{Oral absorption} = f(logD7.4,\ molecular\ size,\ H\text{-}bonding\ capacity) \quad (1)$$

where LogD7.4 is the distribution coefficient, namely the octanol/water partition coefficient at pH 7.4, molecular size is a measure related to mass, volumes, and surfaces of the molecule in question, and H-bonding capacity relates to the number and strength of the hydrogen bonds that can be donated and/or accepted by the model compound. All literature data surveyed thus far converge on the importance of hydrogen-bonding factors to oral absorption—be this calculated by HYBOT[7] or MolSurf,[8] or indirectly by estimating polar surface areas in

Color Plate for this article is on page 329.

Corresponding author: T.I. Oprea. Tel.: 46-31-776-10-00; fax: 46-31-776-37-00.

E-mail address: tudor.oprea@astrazeneca.com (T.I. Oprea)

Journal of Molecular Graphics and Modelling 17, 261–274, 1999
© 2000 by Elsevier Science Inc.
655 Avenue of the Americas, New York, NY 10010

1093-3263/99/$–see front matter
PII S1093-3263(99)00034-0

VolSurf,[9] SAVOL,[10] or in other molecular modeling software. Lipophilicity measures[11] can be estimated with Albert Leo's CLOGP program,[12] or the ACDLogP software.[13] The latter is part of a program that estimates LogD7.4 values, based on calculated LogP and pK_a values. Molecular size can be estimated by computing volumes and surfaces by a number of methods available in commercial molecular modeling software, e.g., by using SAVOL in SYBYL. Parameters computed with HYBOT,[7] MolSurf,[8] and VolSurf[14] have been successfully shown to correlate with bioavailability-related parameters.

This article describes a quantitative extension of the "rule of five" scheme, which can be achieved with minimal computational efforts. The descriptors used to correlate oral absorption were chosen on the basis of three criteria: (1) they are relatively fast and easy to calculate; (2) they have a direct physicochemical interpretation that is intuitive to the practising chemist; and (3) their use is intended for the combinatorial chemistry and/or HTS framework (i.e., efficient handling of hundreds of thousands of compounds). Furthermore, these descriptors are consistent with the current QSAR paradigm for oral absorption, as outlined above [see Eq. (1)]. Using data available from the literature, we propose an easy-to-interpret model that is stable for two independent measures of oral absorption, namely the percent human intestinal absorption (%HIA), and the apparent Caco-2 cell permeability coefficient (P_{app}).

MATERIALS AND METHODS

Compound selection

The data set consists of 85 compounds, for which Wessel and co-workers[4] gathered %HIA values from the literature. Sixteen of these 85 structures have Caco-2 cell permeability data, originating from Artursson and Karlsson,[15] that were used in the QSAR models proposed by Van de Waterbeemd et al.[2] and by Norinder et al.[3] In addition, Yazdanian et al.[16] reported Caco-2 cell permeability data for 29 of these 85 structures. Thus, three experimental measures were used to derive our QSAR model. %HIA was modeled with the logit transformation [see Eq. (2)], whereas $\log_{10}$ values were used for the P_{app} data from Artursson and Karlsson[15] (ALgPapp) and from Yazdanian et al.[16] (YLgPapp). Experimental data are summarized in Table 1.

$$\text{Logit}(\%\text{HIA}) = \log[(\%\text{HIA} + 0.1)/(100.1 - \%\text{HIA})] \quad (2)$$

Descriptor calculations

The initial list of QSAR descriptors was consistent with the current paradigm [see Eq. (1)]:

- *LogD7.4* from ACD Labs[17]: LogD7.4 takes into account the extent of ionization, as well as the partitioning of various microspecies (ionized and/or neutral) for each compound, in both aqueous and organic phases.[17] In addition to LogD7.4, CLOGP[12] and ACDLogP[13] were included, to estimate better the hydrophobicity-related aspects. While certain discrepancies exist between the ACDLogP and CLOGP methods regarding the LogP value for given compounds, the PLS modeling of hydrophobicity will be based on those data that coincide, and produce a PLS-weighted value for those situations in which large differences occur.
- *Molecular size*: Molecular size was evaluated by the following descriptors: molecular weight (MW), molecular volume (MolVol), molecular surface (TotArea), polar volume (PolVol), polar surface (PolArea), nonpolar volume (NPVol), and nonpolar surface (NPArea). Except for molecular weight, all other descriptors were calculated for each compound by SAVOL,[10] on a single conformer generated with CONCORD.[18] For the sake of simplicity, all noncarbon heavy atoms and their attached hydrogens were considered "polar," whereas all other hydrogens and the carbon atoms were considered "nonpolar." Note that "polar" in this definition does not overlap with "hydrophilic."
- *Hydrogen-bonding capacity*: Hydrogen-bonding capacity was estimated by HYBOT,[7] using four descriptors: HDOM, the maximum free energy H-bond donor factor (C_d); HDOS, the sum of C_d values; HACM, the maximum free energy H-bond acceptor factor (C_a); and HACS, the sum of C_a values. All C_d values were given a positive sign, as proposed in reference 2. In addition, we have used the simple count of H-bond donors (HDO) and H-bond acceptors (HAC), as implemented in SaSA, our in-house software.[19]
- *Additional parameters*: Additional parameters included simple counts of carbon (#C), nitrogen (#N), and oxygen (#O) atoms, the number of rotatable bonds (RTB), the number of rigid bonds (RGB), as well as a count of positive (N_POS) and negative (N_NEG) ionization centers, as implemented in SaSA.[19]

RTB is calculated as in Eq. (3), where N is the number of nonterminal freely rotatable bonds (e.g., single bonds observed in groups such as sulfonamides [N–S] or esters [C–O] are excluded); n_i is the number of single bonds in any ring i that has more than 5 single bonds; RGB$_i$ is the number of rigid bonds in ring i; ShB$_i$ is the number of common bonds between ring i and any other ring. RGB is the difference between the total number of bonds and the total number of rotatable bonds, which includes terminal single bonds besides RTB.

$$RTB = N + \sum_i (n_i - 4 - RGB_i - ShB_i) \quad (3)$$

- *Nonlinear parameters*: Nonlinear combinations for some of the preceding parameters were included, since previous QSAR studies[2, 4] indicated that nonlinear models may be better suited to explain oral absorption. Significant contributions from nonlinear terms were selected via PLS loadings, with the aim of capturing the manifest nonlinearity involved in oral absorption. These nonlinear combinations included DOSQ (squared HDOS); ACSQ (squared HACS); D7.4SQ (squared LogD7.4); PolASQ (squared PolArea); NPSQ (squared NPArea); !PolArea and !NPArea (PolArea and NPArea, normalized by TotArea); as well as cross-terms for HDOS and LogD7.4, as follows: DOS_PA (HDOS * PolArea); DOS_D7.4 (HDOS * LogD7.4); DOS_ACS (HDOS * HACS); D7.4_TA, D7.4_PA, D7.4_NP—which are the cross-terms of LogD7.4 with the total, polar, and nonpolar areas, respectively. Estimates of polar and nonpolar surface areas may show linear correlations with calculated LogP and HYBOT values. However, this problem is circumvented by the use of multivariate analysis methods. All descriptor values are listed in Table 2.

Table 1. List of compounds and their respective biological activities[a]

Nr.crt.	Compound	%HIA	Logit(%HIA)	ALgPapp	YlgPapp
1	Acebutolol	89.5	0.927		−6.292
2	Acetaminophen	80	0.600		
3	Acetylsalicylic acid	100	3.000	−5.620	−5.041
4	Acrivastine	88	0.862		
5	Alprenolol	93	1.118	−4.393	−4.597
6	Amoxicillin	93.5	1.152		
7	Antipyrine	100	3.000		
8	Atenolol	50	0.000	−6.700	−6.276
9	Betaxolol	90	0.950		
10	Bromazepam	84	0.718		
11	Bumetanide	100	3.000		
12	Bupropion	87	0.823		
13	Caffeine	100	3.000		−4.511
14	Captopril	67	0.307		
15	Cefatrizine	76	0.499		
16	Cefuroxime	5	−1.271		
17	Cefuroximen axetil	36	−0.249		
18	Cephalexin	98	1.669		
19	Chloramphenicol	90	0.950		
20	Clorothiazide	13	−0.823		−4.701
21	Cimetidine	85	0.751		−5.863
22	Clonidine	95	1.271		−4.662
23	Corticosterone	100	3.000	−4.263	−4.674
24	Cromolyn	0.5	−2.220		
25	Desipramine	100	3.000		−4.613
26	Dexamethasone	100	3.000	−4.903	−4.914
27	Diazepam	100	3.000		−4.476
28	Doxorubicin	5	−1.271		
29	Enalaprilat	10	−0.950		
30	Etoposide	50	0.000		
31	Felodipine	100	3.000	−4.644	
32	Fluconazole	95	1.271		
33	Fluvastatin	100	3.000		
34	Furosemide	61	0.194		
35	Gabapentin	50	0.000		
36	Ganciclovir	3.8	−1.393		
37	Gentamycin	0	−3.000		
38	Guanabenz	75	0.476		
39	Hydrocortisone	91	1.00	−4.668	−4.854
40	Ibuprofen	1000	3.000		
41	Imipramine	95	1.271		
42	Ketoprofen	100	3.000		
43	Labetalol	95	1.271		
44	Lamotrigine	70	0.367		
45	Lisinopril	25	−0.476		
46	Loracarbef	100	3.000		
47	Lormetazepam	100	3.000		
48	Mannitol	15	−0.751	−6.745	−6.420
49	Methotrexate	100	3.000		
50	Methylprednisolone	82	0.657		
51	Metoprolol	95	1.271	−4.569	−4.625
52	Nadolol	34.5	−0.278		−5.411
53	Naproxen	99	1.955		
54	Norfloxacin	35	−0.268		

Continued

Table 1. Continued

Nr.crt.	Compound	%HIA	Logit(%HIA)	ALgPapp	YlgPapp
55	Olsalazine	2.3	−1.610	−6.959	
56	Ondansetron	100	3.000		
57	Oxprenolol	90	0.950		
58	Phenoxymethylpenicillin	45	−0.087		
59	Phenytoin	90	0.950		−4.574
60	Pindolol	90	0.950		−4.777
61	Practolol	100	3.000	−6.046	
62	Pravastatin	34	−0.287		
63	Prazosin	100	3.000		
64	Prednisolone	98.8	1.881		
65	Progesterone	91	1.000		−4.625
66	Propranolol	90	0.950	−4.378	−4.662
67	Propylthiouracil	75	0.476		
68	Quinidine	80	0.600		
69	Ranitidine	50	0.000		−6.310
70	Salicylic acid	100	3.000	−4.924	−4.658
71	Scopolamine	90	0.950		−4.928
72	Sorivudine	82	0.657		
73	Sotalol	95	1.271		
74	Sulfasalazine	10	−0.950	−6.886	−6.523
75	Sumatriptan	75	0.476		
76	Tenidap	90	0.950		
77	Terazosin	91	1.000		
78	Testosterone	100	3.000	−4.286	−4.604
79	Timolol	90	0.950		−4.893
80	Trimethoprim	97	1.496		
81	Trovafloxacin	88	0.862		
82	Valproic acid	100	3.000		
83	Warfarin	98	1.669	−4.417	−4.676
84	Zidovudine	100	3.000		−5.159
85	Zipracidone	60	0.176		

[a] Compiled from Wessel et al. (%HIA),[4] Norinder et al. (ALgPapp),[3] and Yazdanian et al. (YLgPapp).[16] A value of 10% HIA was used for sulfasalazine, consistent with reference 3. See text for details.

Statistics

All multivariate modeling of the HIA and related measurements were made by PLS[20,21] (projections to latent structures), using the Simca package.[22] The estimation of principal component significance was performed by the cross-validation (CV) procedure[23] and provided[22] as Q^2. Overfit was investigated by perturbation of the response values and the related loss of Q^2.

RESULTS AND DISCUSSION

Compound and variable selection

An initial set of 16 compounds was used in the model: acetylsalicylic acid, alprenolol, atenolol, corticosterone, dexamethasone, felodipine, hydrocortisone, mannitol, metoprolol, olsalazine, practolol, propranolol, salicylic acid, sulfasalazine, testosterone, and warfarin, for which both %HIA and Caco-2 cell permeability (ALgPapp) data were available. YLgPapp values were also available for these compounds, except for felodipine, olsalazine, and practolol—therefore the final PLS model was performed using three **Y** vectors, with 3 missing values on the YLgPapp column. Initial PLS models, using one **Y** variable at a time, indicated 1 dominating principal component (PC), with an occasional second significant PC (according to leave-two-out CV). Therefore, six PLS models were derived on the basis of these 16 compounds: the 1-PC model, and the optimal-PC model, according to leave-two-out CV.

These PLS models were analyzed in terms of the variable importance contribution (VIP) for each of the starting descriptors. This comparison was aimed at eliminating those descriptors that did not relate to oral absorption and/or bioavailability phenomena, and at focusing on descriptors that consistently contribute to the PLS models in all 3 responses. Only descriptors having a VIP value larger than 1 were considered during this analysis. The descriptors that occurred in at least 5 models, and had VIP values above the cutoff value, were automatically selected. Descriptors that had VIP $>$ 1.5 in at least one model (this being deemed as highly important) were also considered. RTB had VIP $>$ 1.65 in the PLS models for %HIA, and was thus included in the final list of descriptors. All other descriptors occurred in six models, except for HDO and #N, that had

Table 2. Values for each calculated descriptor.

Compound	LogD7.4	clogP	ACDLogP	MW	MolVol	TotArea	PolVol	PolArea	NPVol	NPArea	HDOS	HDOM	HACS	HACM	HDO	HAC	RTB	RGB	#C	#N	#O	N_POS	N_NEG
Acebutolol	0.49	1.632	2.59	336.4	397.5	272.4	80	51.5	317.4	220.9	5.95	2.11	4.84	1.73	3	5	10	9	18	2	4	1	0
Acetaminophen	0.34	0.494	0.34	151.2	166.7	113.8	48.5	27.8	118.2	86	4.03	2.11	3.11	1.73	2	2	1	8	8	1	2	0	0
Acetylsalicylic acid	−2.51	1.023	1.19	180.2	181.7	126	56.5	31.9	125.3	94.1	2.11	2.11	3.8	1.73	0	3	2	9	9	0	4	0	1
Acrivastine	2.12	1.131	4.63	348.4	395.1	279.1	51.4	32.5	343.7	246.6	2.11	2.11	1.9	1.73	0	4	7	19	22	2	2	2	1
Alprenolol	2.14	3.558	3.18	262.4	319.5	218.9	52.5	35.5	267	183.4	3.62	1.92	0.4	0.4	2	4	8	8	15	2	2	1	0
Amoxicillin	−2.58	−1.936	0.61	365.4	365.1	257.6	150.4	90.7	214.6	166.9	8.94	2.11	8.06	1.73	4	6	4	18	16	3	5	1	1
Antipyrine	0.27	0.414	0.27	188.2	209.3	146.2	25.9	20.1	183.4	126.1	0	0	1.94	1.73	0	3	1	12	11	2	1	0	0
Atenolol	−2.06	−0.109	0.1	266.3	321.7	216.2	84.5	49.2	237.2	167	6.83	2.11	3.11	1.73	4	4	8	7	14	2	3	1	0
Betaxolol	0.49	2.169	2.66	307.4	390.1	262.3	54.5	34.1	335.6	228.3	4.03	2.11	1.17	1.17	2	4	11	9	18	1	3	1	0
Bromazepam	2.41	1.724	2.41	316.2	257.5	193.3	84.8	56	172.7	137.2	1.92	1.92	1.94	1.73	1	3	1	19	14	3	1	2	0
Bumetanide	−0.62	3.898	2.78	364.4	370.2	260	120.6	76.1	249.5	184	7.59	2.11	4.67	1.73	3	6	8	15	17	2	5	1	1
Bupropion	3.27	3.211	3.47	239.7	280.5	191.4	56.7	35.1	223.8	156.4	1.92	1.92	1.73	1.73	1	2	4	7	13	1	1	1	0
Caffeine	−0.08	−0.057	−0.08	194.2	193.3	135.2	55.4	40.7	137.8	94.5	0	0	3.88	1.73	0	6	0	12	8	4	2	0	0
Captopril	−1.98	1.016	1.51	217.3	236.6	160.8	83.1	50.9	153.5	110	2.11	2.11	3.84	1.73	0	3	4	7	9	1	3	0	1
Cefatrizine	−3.2	−3.378	0.11	462.5	425.5	309.8	208.9	133.9	216.6	175.9	10.86	2.11	8.06	1.73	5	9	7	24	18	6	5	1	1
Cefuroxime	−3.77	−0.174	0.29	424.4	387.6	275.9	180	113.7	207.6	162.3	6.83	2.11	7.89	1.73	3	8	7	21	16	4	8	0	1
Cefuroxime axetil	0.5	0.251	0.67	510.5	479.8	340.1	192	123.4	287.7	216.6	4.72	1.92	9.79	1.73	3	8	9	24	20	4	10	0	0
Cephalexin	−2.52	−1.901	0.65	347.4	345	244.9	131.4	80.1	213.5	164.8	6.83	2.11	6.89	1.73	3	5	4	19	16	3	4	1	1
Chloramphenicol	1.02	0.687	1.02	323.1	290	206	149.3	93.5	140.7	112.5	5.71	2.11	5.95	1.73	3	5	6	10	11	2	5	0	0
Clorothiazide	−0.68	−0.408	−0.03	295.7	223.5	164.3	150.3	101.7	70.2	62.6	5.48	1.92	5.43	1.33	3	5	1	15	7	3	4	1	0
Cimetidine	0.21	0.351	0.36	252.3	280.7	193.4	108.4	74.2	172.3	119.2	5.76	1.92	2.18	1.98	3	6	6	9	10	6	0	2	0
Clonidine	−1.07	1.367	1.49	230.1	213.2	152.3	95.5	61.4	117.7	90.9	3.84	1.92	0	0	2	3	1	12	9	3	0	1	0
Corticosterone	1.76	1.163	1.76	346.5	395.1	281.8	69.4	36.9	325.6	244.9	3.79	2.11	6.17	1.73	2	4	2	22	21	0	4	0	0
Cromolyn	−4.8	0.475	0.2	468.4	422.6	307.1	148.8	86.2	273.9	220.9	6.33	2.11	8.43	1.73	1	11	8	26	23	0	11	0	2
Desipramine	1.23	4.087	3.97	266.4	313	224	20.2	16.7	292.8	207.3	1.92	1.92	0	0	1	2	5	16	18	2	0	2	0
Dexamethasone	2.06	1.485	2.06	392.5	415.3	300	96.9	53.5	318.4	246.3	5.66	2.11	7.62	1.73	3	5	2	22	22	0	5	0	0
Diazepam	2.96	3.084	2.96	284.7	279.4	205.3	58.2	38.5	221.3	166.9	0	0	1.94	1.73	0	2	1	19	16	2	1	1	0
Doxorubicin	1.03	−1.451	2.79	543.5	525.6	382.6	187	108.5	338.6	274.2	12.68	2.11	12.9	1.73	7	12	8	27	27	1	11	1	0
Enalaprilat	−0.28	0.041	3.34	348.4	386.7	266.6	103.1	60.3	283.6	206.3	6.14	2.11	5.74	1.73	1	6	9	14	18	2	5	1	2
Etoposide	−4.47	−1.12	0.39	668.5	615.1	446.6	215.1	134.5	400	312.1	8.4	2.11	7.98	2.87	4	15	8	37	29	0	16	0	2
Felodipine	4.92	4.525	4.92	384.3	375.5	267.8	114.7	71.1	260.8	196.7	1.92	1.92	3.8	1.73	1	3	4	16	18	2	5	1	0
Fluconazole	0.3	−0.114	0.31	306.3	282	205.1	98.9	68.9	183.1	136.3	1.87	1.87	1.52	1.52	1	7	5	16	13	6	1	0	0
Fluvastatin	0.21	3.162	3.63	411.5	439.4	311.1	87.3	51.6	352.2	259.5	6.33	2.11	4.24	1.73	2	5	8	18	24	1	4	0	1
Furosemide	−0.99	1.139	2.92	330.7	287.5	207	142	91.7	145.5	115.4	7.59	2.11	4.67	1.73	3	6	5	14	12	2	5	1	1
Gabapentin	−1.31	−1.222	1.19	171.2	217.1	145.2	60.9	32.9	156.2	112.3	4.91	2.11	3.82	1.92	2	3	5	5	9	1	2	1	1
Ganciclovir	−4.25	−3.216	−2.15	255.2	254.3	175.2	133.3	81.7	121	93.4	8.08	1.92	4.81	1.73	5	9	5	11	9	5	4	1	0
Gentamycin	−8.4	−3.774	−1.89	477.6	550.04	381.28	200.7	122.4	326.7	249.1	18.33	2.11	7.19	1.52	11	12	13	12	21	5	7	5	0
Guanabenz	−0.45	2.965	2	231.1	211.6	152.3	122.4	78.3	89.2	74	6.77	2.05	2.96	2.46	4	4	2	9	8	4	0	2	0
Hydrocortisone	1.43	0.537	1.43	362.5	397.8	287.9	82.3	46.5	315.5	241.4	5.66	2.11	7.69	1.73	3	5	2	22	21	0	5	0	0
Ibuprofen	0.77	3.679	3.72	206.3	257.4	174.2	34.3	18.1	223.1	156.1	2.11	2.11	1.9	1.73	0	2	4	7	13	0	2	0	1
Imipramine	2.41	4.413	4.47	280.4	333.8	238.2	10.1	12.5	323.7	225.7	0	0	0	0	0	2	5	16	19	2	0	2	0
Ketoprofen	−0.31	2.761	2.81	254.3	274.8	193.6	49.4	25.9	225.3	167.7	2.11	2.11	3.63	1.73	0	3	4	14	16	0	3	0	1
Labetalol	0.95	2.5	2.87	328.4	374.1	261	89.6	52.2	284.5	208.7	8.94	2.11	4.28	1.73	5	4	8	13	19	2	3	1	0
Lamotrigine	−0.19	2.626	−0.19	256.1	226.6	164.7	140	87.7	86.5	77	5.6	1.4	2.22	1.11	4	5	1	12	9	5	0	2	0
Lisinopril	−0.15	−2.529	2.98	405.5	466.3	319.5	129.4	75	345	248	8.94	2.11	7.75	2.01	3	7	13	14	21	3	5	2	2
Loracarbef	−3.26	−0.472	−0.12	349.8	339	241.7	127	77.2	205.2	161.3	6.83	2.11	6.89	1.73	3	5	4	19	16	3	4	1	1
Lormetazepam	2.37	3.676	2.37	335.2	302.8	224.6	102.4	65.4	200.4	159.2	2.11	2.11	3.11	1.73	1	3	1	19	16	2	2	1	0

Continued

Table 2. Continued

Compound	LogD7.4	clogP	ACDLogP	MW	MolVol	TotArea	PolVol	PolArea	NPVol	NPArea	HDOS	HDOM	HACS	HACM	HDO	HAC	RTB	RGB	#C	#N	#O	N_POS	N_NEG
Mannitol	−4.96	−4.67	−4.96	182.2	210.8	133.9	118.8	63.5	92	70.3	11.8	2.11	7.76	1.54	6	6	5	0	6	0	6	0	0
Methotrexate	−3.4	−0.865	−0.09	454.4	453.8	322.2	200.4	122.4	253.5	199.8	11.74	2.11	7.96	1.73	5	12	9	21	20	8	5	3	2
Methylprednisolone	2.18	1.642	2.18	374.5	410.9	296.3	83.4	46.4	327.5	250	5.66	2.11	7.69	1.73	3	5	2	22	22	0	5	0	0
Metoprolol	−0.41	1.196	1.76	267.4	338.1	226.2	54.7	34.1	283.4	192.2	4.03	2.11	1.17	1.17	2	4	9	6	15	1	3	1	0
Nadolol	−0.84	0.229	1.29	309.4	369.1	252.6	76.4	45.2	284.8	203.9	8.25	2.11	3.51	1.17	4	5	7	10	17	1	4	1	0
Naproxen	0.29	2.816	3	230.3	249.6	175.5	43.8	24.4	205.8	151.1	2.11	2.11	1.9	1.73	0	3	3	12	14	0	3	0	1
Norfloxacin	−1.03	1.732	1.49	319.3	322.2	229.8	89.5	56.4	232.7	173.3	4.03	2.11	3.63	1.73	1	6	5	17	16	3	3	3	1
Olsalazine	−1.06	4.501	3.94	302.2	272.8	199.3	119.2	71.7	153.6	172.6	8.44	2.11	6.14	1.73	2	8	4	15	14	2	6	2	2
Ondansetron	2.12	2.643	2.49	293.4	314.1	227.4	37.2	28.4	276.9	199	0	0	1.73	1.73	0	4	3	20	18	3	1	0	0
Oxprenolol	0.16	1.692	2.29	265.4	329.1	221	54.4	33.9	274.7	187.1	4.03	2.11	1.17	1.17	2	4	9	7	15	1	3	1	0
Phenoxymethylpenicillin	−2.14	1.879	1.88	350.4	349.8	247.1	114.4	71.8	235.4	175.3	4.03	2.11	5.78	1.73	1	5	5	18	16	2	5	0	1
Phenytoin	1.95	2.09	2.28	252.3	252.3	184.3	67.1	43.4	192.7	148.7	4.03	2.11	1.73	1.73	2	3	2	18	15	2	2	1	0
Pindolol	−0.22	1.671	1.97	248.3	293.5	201.6	60.2	37.8	233.4	163.8	5.95	2.11	1.17	1.17	3	4	6	10	14	2	2	1	0
Practolol	−1.39	0.755	0.76	266.3	320.7	216.1	73.4	44.9	247.2	171.2	5.95	2.11	3.11	1.73	3	4	7	8	14	2	3	1	0
Pravastatin	−2.05	0.544	1.35	424.5	492.2	339.3	111.4	62.7	380.8	276.6	8.44	2.11	7.31	1.73	3	6	10	14	23	0	7	0	1
Prazosin	−6.05	2.21	−2.06	383.4	384.3	278.4	100	70	284.3	208.5	2.8	1.4	3.05	1.73	2	8	6	22	19	5	4	2	0
Prednisolone	1.69	1.123	1.69	360.5	389	282.6	82.4	46.5	306.6	236.1	5.66	2.11	7.69	1.73	3	5	2	22	21	0	5	0	0
Progesterone	4.04	3.775	4.04	314.5	375	268	30.2	15.7	344.7	252.3	0	0	3.46	1.73	0	2	1	22	21	0	2	0	0
Propranolol	0.96	2.753	3.1	259.4	308.2	213.6	44.5	27.5	263.8	186	4.03	2.11	1.17	1.17	2	3	6	11	16	1	2	1	0
Propylthiouracil	1.09	1.37	1.37	170.2	182.3	125	72.9	46	109.4	78.9	3.84	1.92	2.77	1.73	2	3	2	8	7	2	1	0	0
Quinidine	1.82	2.931	3.44	324.4	358.6	259.2	46.7	31.4	311.9	227.8	2.11	2.11	1.17	1.17	1	4	4	21	20	2	2	2	0
Ranitidine	0.54	1.327	1.28	314.4	345.4	237.2	99.5	71.4	245.9	165.8	3.84	1.92	1.3	0.65	2	6	10	8	13	4	3	3	0
Salicylic acid	−1.86	2.187	2.06	138.1	136.4	95.7	48.2	27.5	88.2	68.3	4.22	2.11	3.07	1.73	1	3	1	7	7	0	3	0	1
Scopolamine	0.62	−0.195	1.34	303.4	325.6	231.7	56.4	36.1	269.2	195.6	1.68	1.68	3.44	1.73	1	4	4	19	17	1	4	1	0
Sorivudine	−1.11	−1.788	−0.74	349.1	277.9	201.2	138.9	88.9	139	112.3	7.82	2.11	7.76	1.73	4	8	4	13	11	2	6	0	0
Sotalol	−1.86	0.226	0.32	272.4	305.9	207.7	89.2	59.1	216.6	148.6	5.95	2.11	4.77	1.6	3	4	5	9	12	2	3	1	0
Sulfasalazine	−0.78	3.831	3.18	398.4	362.6	266.1	135.5	90.1	227	176	6.14	2.11	5.77	1.73	2	8	5	23	18	4	5	3	1
Sumatriptan	−1.26	0.583	0.79	295.4	323.1	224.4	74.1	53.8	285	170.6	3.84	1.92	2.77	1.23	2	4	5	13	14	3	2	1	0
Tenidap	−0.58	2.29	2.29	320.8	276.8	204.6	122.5	76.9	154.4	127.7	4.91	2.11	5.05	1.73	3	3	1	19	14	2	3	0	0
Terazosin	−4.64	2.342	−0.64	387.4	404	289.5	100.8	70.4	303.2	219.1	2.8	1.4	3.05	1.73	2	8	7	21	19	5	4	2	0
Testosterone	3.47	3.219	3.48	288.4	342.4	245.3	45	18.4	307.4	226.8	2.11	2.11	2.9	1.73	1	2	0	21	19	0	2	0	0
Timolol	−2.01	1.609	−0.15	316.4	353.2	242.6	105.7	72	247.5	170.6	4.03	2.11	1.17	1.17	2	7	9	9	13	4	3	2	0
Trimethoprim	0.52	0.334	0.79	290.3	312.8	219.1	102	64.7	210.9	154.3	5.6	1.4	2.22	1.11	4	7	5	12	14	3	2	2	0
Trovafloxacin	−0.97	1.708	1.57	416.4	369.7	272.2	136.4	82.2	233.4	189.9	4.91	2.11	4.74	1.73	2	7	3	26	20	4	3	4	1
Valproic acid	0.15	2.72	2.72	144.2	200.6	128.4	34.3	18.1	166.3	110.4	2.11	2.11	1.9	1.73	0	2	5	1	8	0	2	0	1
Warfarin	0.59	2.785	3.47	308.3	314.6	228.3	56.5	31.9	258.1	196.3	2.11	2.11	4.8	1.73	1	4	4	19	19	0	4	0	0
Zidovudine	−0.59	−0.327	−0.58	267.2	264.3	184.8	117.7	74.4	146.6	110.4	3.6	1.92	6.22	1.73	2	8	4	14	10	5	4	1	0
Zipracidone	3.36	4.421	4.26	412.9	402.1	295.7	102.4	70.6	299.8	225.2	1.92	1.92	1.94	1.73	1	4	6	25	21	4	1	2	0

[a] See text for details.

Table 3. Values for the nonlinear descriptors included in the training set[a]

Compound	!PolArea	!NPArea	D7.4_TA	D7.4_PA	D7.4_NP	PolASQ	DOSQ	DOS_ACS	DOS_PA	DOS_D7.4
Acetylsalicylic acid	0.253	0.747	−316.260	−80.069	−236.191	1017.610	4.452	8.018	67.309	−5.296
Alprenolol	0.162	0.838	468.446	75.970	392.476	1260.250	13.104	1.448	128.510	7.747
Atenolol	0.228	0.772	−445.372	−101.352	−344.020	2420.640	46.649	21.241	336.036	−14.070
Corticosterone	0.131	0.869	495.968	64.944	431.024	1361.610	14.364	23.384	139.851	6.670
Dexamethasone	0.178	0.821	618.000	110.210	507.378	2862.250	32.036	43.129	302.810	11.660
Felodipine	0.265	0.735	1317.576	349.812	967.764	5055.210	3.686	7.296	136.512	9.446
Hydrocortisone	0.162	0.838	411.697	66.495	345.202	2162.250	32.036	43.525	263.190	8.094
Mannitol	0.474	0.525	−664.144	−314.960	−348.688	4032.250	139.240	91.568	749.300	−58.528
Metoprolol	0.151	0.850	−92.742	−13.981	−78.802	1162.810	16.241	4.715	137.423	−1.652
Olsalazine	0.360	0.866	−211.258	−76.002	−182.956	5140.890	71.234	51.822	605.148	−8.946
Practolol	0.208	0.792	−300.379	−62.411	−237.968	2016.010	35.403	18.505	267.155	−8.271
Propranolol	0.129	0.871	205.056	26.400	178.560	756.250	16.241	4.715	110.825	3.869
Salicylic acid	0.287	0.714	−178.002	−51.150	−127.038	756.250	17.808	12.955	116.050	−7.849
Sulfasalazine	0.339	0.661	−207.558	−70.278	−137.280	8118.010	37.700	35.428	553.214	−4.789
Testosterone	0.075	0.925	851.191	63.848	786.996	338.560	4.452	6.119	38.824	7.322
Warfarin	0.140	0.860	134.697	18.821	115.817	1017.610	4.452	10.128	67.309	1.245

[a] See text for details.

VIP > 1.0 in five models only. Descriptors selected after the VIP analysis were as follows: LogD7.4, ACDLogP, PolArea, NPArea, HDOS, HDO, HAC, RTB, and #N. Several cross-terms, listed in Table 3, were also above the VIP cutoff in at least 5 models. Two descriptors (TotArea and HACS) that had VIP values below 1.0 were present in combination with other descriptors, as cross-terms. These were included in the final selection for computational reasons, and to ensure a direct interpretability of the PLS models.

Analysis of the final model: Is there a sigmoidal relationship?

The final PLS model was based on 1 PC with R^2 between 0.5 and 0.8 and Q^2 between 0.4 and 0.8 (Figure 1). The PLS inner relation, i.e., the t_1 versus u_1 correlation, revealed a sigmoidal shape (see Figure 2). This was discussed by Van de Waterbeemd et al.[2] and Norinder et al.,[3] who plotted LogD7.4 versus ALgPapp,[2] and ALgPapp versus percent oral absorption in humans,[3] respectively. While the sigmoidal relationship appeared to exist when plotting LogD7.4 versus ALgPapp in our training set (16 compounds), this was not observed when plotting LogD7.4 versus %HIA (85 compounds), LogD7.4 versus YlgPapp (29 compounds), ALgPapp versus %HIA (16 compounds), or even YLgPapp versus %HIA (29 compounds) (plots not shown). Although it could be argued that LogD7.4 is a calculated parameter subject to computational errors, the sigmoidal relationship was not present even when plotting %HIA against two other experimental values, namely ALgPapp and YLgPapp. However, a sigmoidal relationship may be expected when plotting %HIA against a nonrestricted, continuous variable such as Caco-2 cell permeability, because HIA values are restricted to the 0–100% interval.

The sigmoidal curve in Figure 2 might be an artifact, since two of the datapoints located at the lower end of the sigmoid, sulfasalazine and mannitol, are in question. Sulfasalazine, plotted with an oral absorption value between 10 and 15% in Figure 4 of reference 3, has a 65% HIA value according to

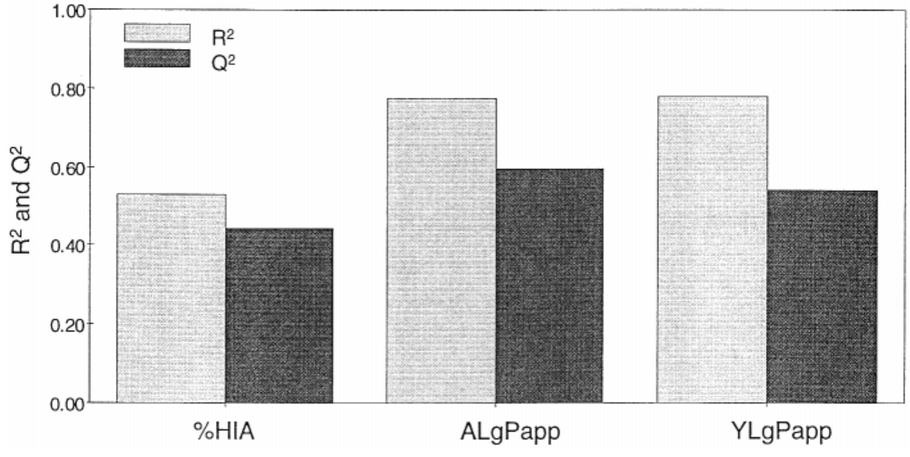

Figure 1. Regression parameters for the 16-compounds, 1-component model. Note that YLgPapp has 18.75% missing values.

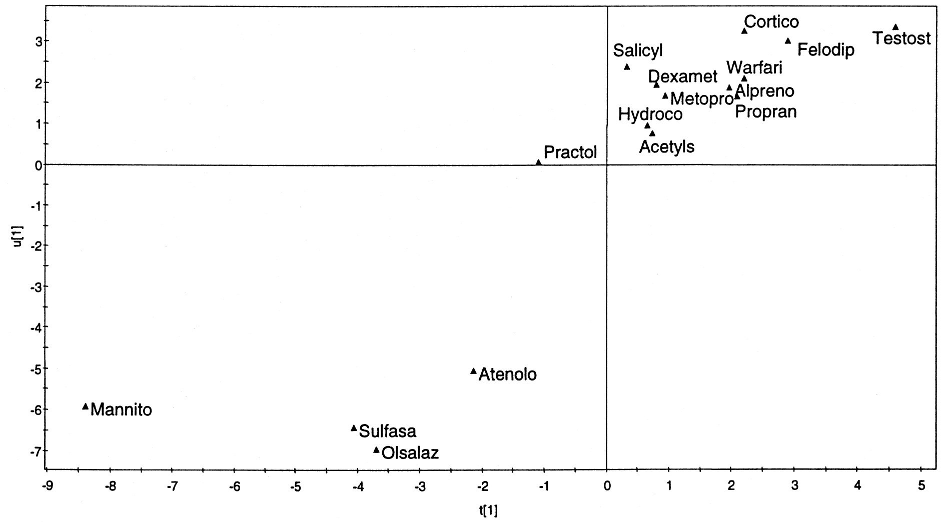

Figure 2. The t_1 *vs* u_1 *plot showing the PLS inner relation, for the 16-compounds, 1-component model.*

reference 4. Sulfasalazine is biotransformed by bacterial azoreductases, via azo bond reduction, in the cecum and colon[24]—see also Figure 3. In fact, sulfasalazine is considered a poorly absorbed sulfonamide[25] since only a small fraction of orally ingested sulfasalazine is absorbed from the small intestine,[24] then excreted unchanged in the urine. Sulfapyridine, the biotransformation product of sulfasalazine, is then absorbed from the colon. In fact, the 65% HIA quoted in reference 4 for sulfasalazine relates to the systemic availability of enteric-coated tablets of sulfapyridine, whereas the oral absorption of sulfapyridine itself is 92.7%, according to reference 24. To summarize, sulfasalazine is a poorly absorbed drug, and the percent oral absorption value used in reference 3 is closer to reality. A 10% value for HIA was used for sulfasalazine in this study (see Table 1). Furthermore, olsalazine is a salicylate from the same class as sulfasalazine, contains the same azo bond, and is also metabolized in the colon.

On the other hand, mannitol is generally accepted in the literature as a marker for paracellular transport,[26] because it is small and hydrophilic, which means that it cannot cross lipid cell membranes, but is likely to permeate across the intercellular gaps. Since three of four points in the lower part of the sigmoid are in doubt (see Figure 2), the significance of this sigmoidal relationship requires further studies, particularly by adding compounds with medium-to-poor oral absorption and permeability properties to the QSAR model. In fact, the existence of this sigmoid was also questioned by Winiwarter et al.,[27] who used PLS models to explain Caco-2 P_{eff} data. As an example, we suggest that Caco-2 data need to be measured for compounds like prazosin and terazosin, which have %HIA values above 90% in reference 4 and Table 1 (only 68% for prazosin according to Benet et al.[28]), but were predicted to have low permeability properties (e.g., HIA $< 5\%$ and ALgPapp < -6 for both compounds). Measured Caco-2 cell permeability was not available in the literature for these two compounds, at the time of this writing. Our model fails to explain their excellent absorption properties, despite the fact that their computed nonpolar surface area is over 70%.

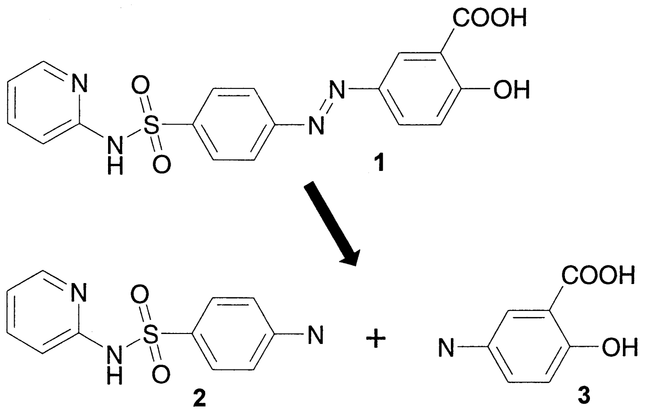

Figure 3. Sulfasalazine (1) is biotransformed into sulfapyridine (2) and 5-aminosalicylic acid (3), by intestinal bacteria. See text for details.

Other compounds that are outliers in computational models of oral absorption have recently been discussed by Clark.[29] These include methotrexate, zidovudine, amoxicillin, cefuroxime axetil, lisinopril, and possibly etoposide. Most of these compounds are absorbed by processes other than passive diffusion. The sigmoidal relationship between %HIA and polar surface area is questioned on the basis of experimental data availability, particularly with regard to the mode of absorption, which tends to add noise to the model.[29]

High regression coefficients, as seen in the present study, may be due to overfit. This was investigated by redoing the regression with scrambled **Y** data. The procedure indicated that the model lost its explained variance, as measured by Q^2, with the grade of correlation between the true response vector and the perturbated vector (Figure 4a–c). The column chart in Figure 5 shows the loadings (w^*c) for all descriptors selected in the final model. This model was used to predict the remaining 69 compounds, according to the availability of their respective responses—see Figure 6a and b.

Analysis of the final model: Property correlations to absorption and permeability

The loadings (w^*c) plot shows the correlation coefficient of each variable to the principal component(s) in a PLS model.

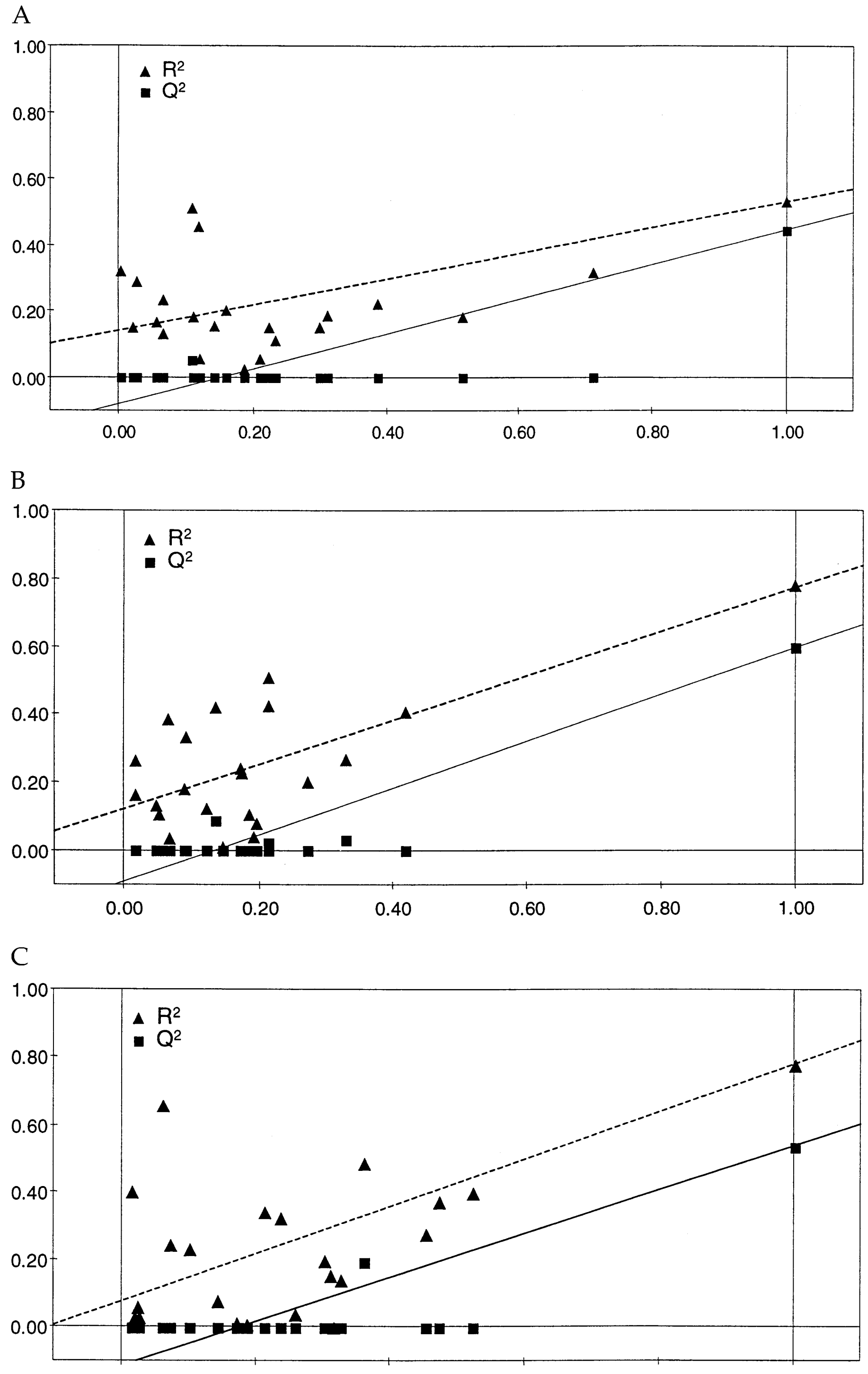

Figure 4. Regression parameters for the 16-compounds, 1-component model for %HIA (a), ALgPapp (b), and YLgPapp (c). R^2 *and* Q^2 *values are plotted against the regression coefficient of the* **Y** *vector itself, that shows the degree of scrambling for each model, respectively. Note that, while there appear to exist good models according to the fraction of explained variance (*R^2*—marked by triangles), these models are not robust according to internal cross-validation (*Q^2*—marked by squares). This indicates that those relatively high* R^2 *models are due to chance correlation, while the final unscrambled PLS model is significant.*

This provides an overview of the correlation pattern extracted by the model. For the present model, the loadings were clustered in the following manner: (1) variables related to H-bonding capacity (#N, HDO, HAC, HDOS, HACS, PolArea), which revealed inverse correlations to the **Y** block (i.e., %HIA, ALgPapp, and YLgPapp); (2) variables related to hydrophobic transferability (ACDLogP, LogD7.4, NPArea), which revealed direct correlations to the **Y** block, and (3) variables related to size (TotArea) and entropy (RTB), which had minor contributions. These loadings are plotted in Figure 5. In addition to the linear terms of the model, we found two significant quadratic terms (i.e., DOSQ and PASQ), both with negative signs indicating a maximum value, or a plateau, for optimal absorption. Six significant cross-terms indicating complex nonlinear relationships within the property space were included, as shown above.

By plotting the response surface for ALgPapp as a function of PolArea and HDOS, one can notice that the combination of

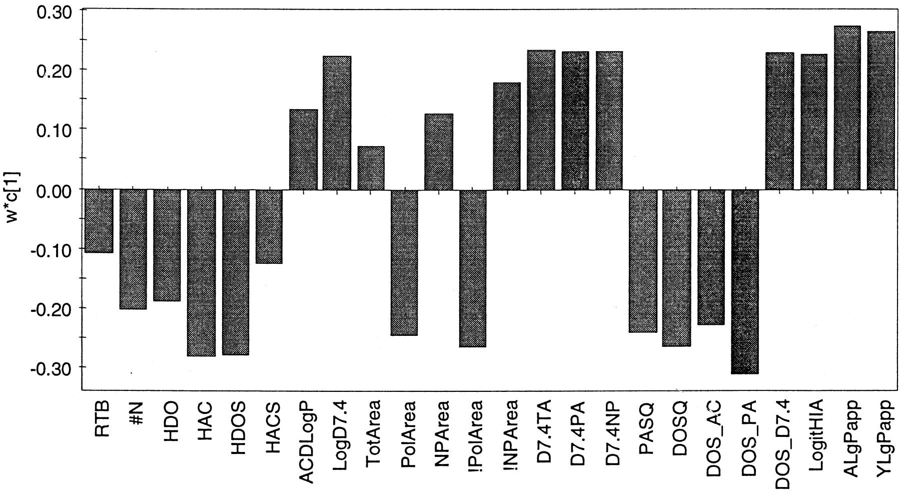

Figure 5. Loadings plot for the 16-compounds, 1-PC model. The variable-to-PLS component correlation coefficients are shown for all descriptors. Higher absolute values suggest larger contribution to the model. Positive values show direct correlations, whereas negative values imply inverse correlations with the ***Y*** *vectors.*

many H-bond donors and a large polar surface area yields low permeability values, while any other combination results in reasonable permeability—see Color Plate 1a. ALgPapp was also plotted as a function of PolArea and LogD7.4; see Color Plate 1b. It can be observed that, for molecules with large polar surface area, permeability increases with LogD7.4, while for molecules with small polar surface area, the distribution coefficient (LogD7.4) appears to have little effect on intestinal permeability (ALgPapp). This observation can be traced to small water-soluble, lipid-insoluble molecules such as mannitol or urea, which have low LogD7.4 values, as well as small polar surface areas. Such molecules are likely to exhibit good intestinal passage due to other mechanisms, besides passive transmembrane diffusion, e.g., solvent drag through nonrestrictive aqueous pores induced by hyperosmolality,[30] or perhaps via paracellular diffusion.[15, 31] This response surface plot is in agreement with a recent proposal by Clark: Compounds that have (calculated) polar surface area > 140 Å^2 are likely to have more problems with oral absorption, compounds with values $<$ 61 Å^2 are likely to have "good" oral absorption properties, whereas compounds with polar surface area between 61 and 140 Å^2 are deemed "OK."[29] In this study, the same procedure as described by Clark, namely a SAVOL calculation on a CONCORD-based structure, was used. Therefore, a good agreement between the estimated polar surface areas was found ($R^2 = 0.774$, $n = 85$). However, because of divergent definitions of the "polar" atoms and in the atomic radii, an offset difference between the two estimates is observed.

In the final model, LogD7.4 is also combined in cross-terms with NPArea, TotArea, and HDOS. Response surface areas revealed that the combination of low LogD7.4 with many H-bond donors, or with a large total area, or with a large nonpolar surface area, is correlated to poor intestinal permeability. High LogD7.4 values correlate to significantly improved permeability, even in the presence of, e.g., many H-bond donors—at least within the limits of this data set. Furthermore, an increase in the number and acidity (C_d) of H-bond donors has a higher influence on intestinal permeabilty, as opposed to an increase in the number and basicity (C_a) of H-bond acceptors, suggesting that donors have a somewhat higher leverage on oral absorption.

The selected parameters were computed extremely fast, using single conformers for molecular surface areas. This should be compared with previously published models that used *ab initio* quantum mechanical calculations—e.g., in MolSurf,[3] extended analyses of surfaces and volumes extracted from GRID-based calculations—e.g., in VolSurf,[14] or even on the basis of multiple conformational analyses of polar surface areas.[5] The present model, while comparable to others already present in the literature,[2] is a simplification of this work, since it does not require descriptors that estimate the water-accessible volume and/or surfaces. It is possible to obtain higher component models for Caco-2 cell permeability, as reported by Van de Waterbeemd et al.,[2] but this was not possible for %HIA. Therefore, we applied the *parsimony principle* (less is better), and restricted ourselves to the 1-PC model to explain all 3 responses. In agreement with Winiwarter et al.,[27] we did not observe any correlation between molecular weight and the three measures of oral absorption and permeability included in the present study.

Our selection of descriptors is aimed at direct model interpretability. The PLS algorithm provides the possibility to investigate the contribution of each descriptor in oral absorption models. By combining PLS with simple and intuitive descriptors (e.g., LogD7.4, polar surface area, H-bond capacity), we tried to improve model utility. This increases the possibility that the medicinal chemist suggests novel structures with higher probability of having improved oral absorption. Our model is consistent with previously published QSARs (references 1, 2, 3, 5, 14, 16, 27, and 29), with respect to the physicochemical properties and model interpretation. Other parameters, such as the cubic root of the gravity index (GRAV), and the molecular surface area projected on the *YZ* plane (SHDW), have been used by Wessel et al.[4] These were combined with the number of single bonds and several charge- and surface-related parameters pertaining to donatable hydrogen atoms. Given that atomic weight does not change during normal experimental conditions, GRAV measures molecular size, since it estimates atomic mass distribution through bond space. Therefore, this model[4] is also consistent with the QSAR paradigm outlined in Eq. (1).

External predictivity

The QSAR model outlined above did not explain the permeability properties of every single compound used in the test set.

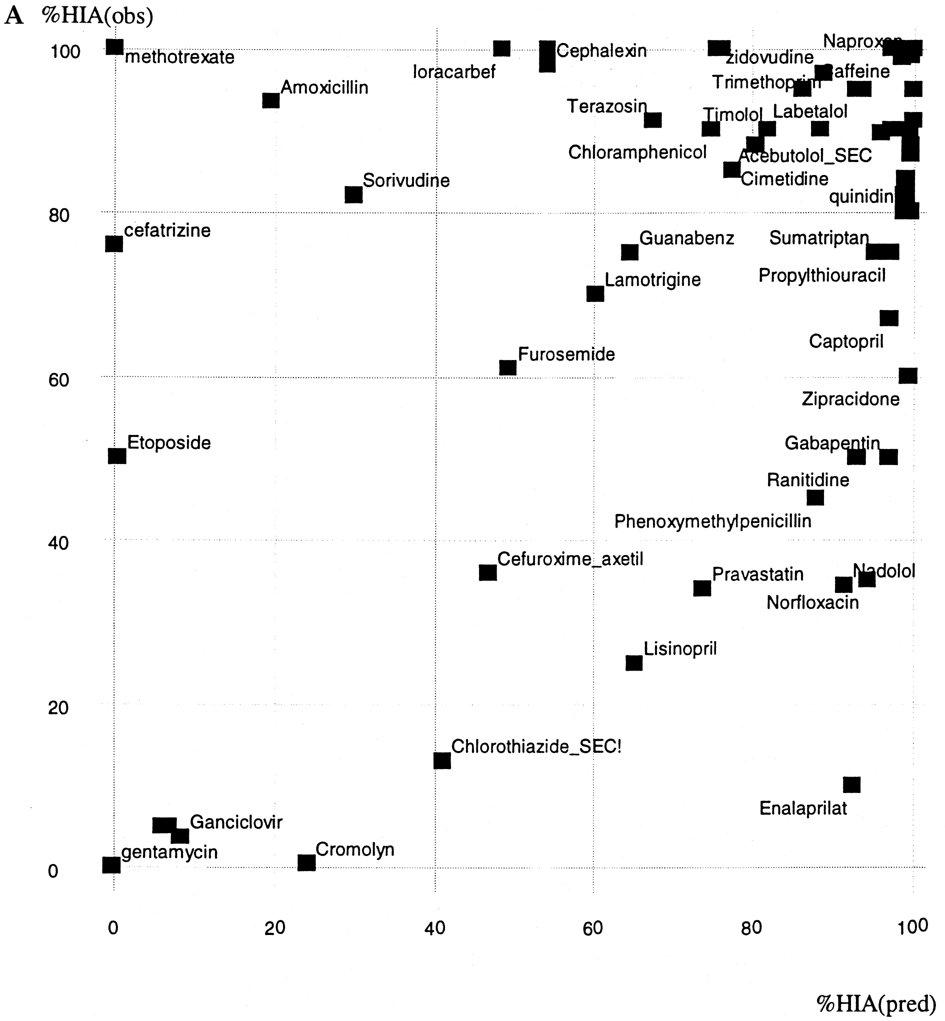

Figure 6. Predicted versus observed %HIA for 69 compounds (a), and predicted versus observed YLgPapp for 16 compounds (b—next page), using on the 16-compounds, 1-component model.

On the contrary, one-third of the Caco-2 cell permeability data from Yazdanian et al.[16] were not explained within reasonable limits (see Figure 6b), while one-third of the %HIA values were predicted with more than 20% HIA units error (see Figure 6a). This was far from discouraging, since 11 of the 16 compounds (68.7%) in this external set for YLgPapp were predicted within 0.6 log unit error, and 46 of the 69 compounds (66.7%) from the external test set for %HIA were predicted within 23% HIA unit error. The measurement of Caco-2 cell permeability is also prone to errors: Alprenolol, for instance, has a LogPapp value of −4.393 in reference 15, −4.597 in reference 16, and −3.616 in reference 5, thus spanning 1 log unit in three different experiments. Although we believe −3.616 to be the correct value,[32] only the first two values were used in this report. A comparison between ALgPapp and YLgPapp values in Table 1 reveals further variations, indicating differences between the two experimental techniques.

The root mean square error of prediction (RMESP) for %HIA is 32.5% when considering the entire test set (69 compounds; Figure 6a), and only slightly improved (28.6%) when removing the 6 outliers discussed by Clark.[29] The RMESP for YLgPapp is 0.64 log unit (16 compounds; Figure 6b). The overall degree of accuracy of these predictions is rather poor, but the complexity of this problem is further outlined if one compares the "Goodman & Gilman" compilation table for percent oral availability (%ORAL) data,[28] versus the %HIA data from Wessel et al.[4] (both are for humans): $R^2 = 0.42$ ($n = 58$) for the %ORAL versus %HIA plot (Figure 7), and $R^2 = 0.33$ ($n = 63$) for the %HIA predicted versus %HIA observed, respectively (Figure 6a).

The fact that 31% (18 of 58) of the compounds plotted in Figure 7 were outside the 20% unit error indicates that drug permeability, be it intestinal absorption or oral avilability, is an extremely variable biological parameter, and that the errors of the current model are within the same order of magnitude as those from the available experimental data. The variability is, at least in part, due to the fact that the fraction of the administered drug that is absorbed intact depends on other factors besides epithelial permeability. The area of the absorbing surface and the residence time of the drug at the absorbing surface have a direct influence on the fraction of the drug dose that is absorbed (%F), while several other factors (e.g., complexation with bile salts, metabolic reactions in the intestine, etc.) may decrease %F.[31] The %ORAL estimates include specific regions

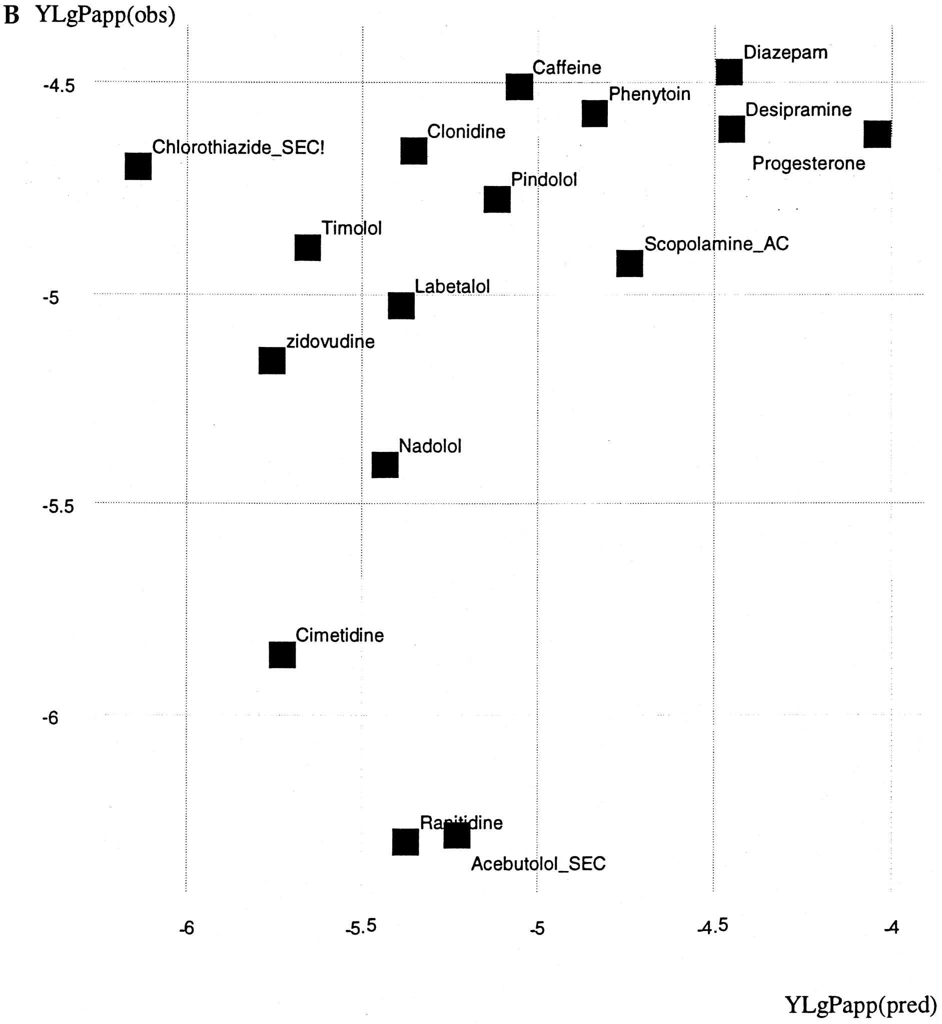

Figure 6. Continued.

of the gastrointestinal tract that may absorb (or not) drugs at different rates, further substantiating the complexity of the problem. On the other hand, there are drugs with high intestinal permeability value that undergo extensive hepatic extraction, e.g., felodipine, imipramine, propranolol, etc., that are typically situated in the upper left corner in Figure 7. Thus, %ORAL becomes a *global* parameter that incorporates other phenomena besides intestinal absorption (%HIA).

The use of the parsimony principle minimizes the risk of overfitting in order to improve regression coefficients—an avenue that may be tempting to pursue in such an area as difficult to model as oral absorption. A model should not fit data with higher precision than that of the measured biological data. The comparison of measured absorption data from two different sources, plotted in Figure 7, confirmed that the present model extracted the optimal amount of signal that could be obtained from this dataset. This was verified in a number of ways (see Figures 1, 2, and 4–7, and Color Plate 1), and is indicative of the fact that a minimalistic model was achieved. Attempts to force higher explained variance are equivalent to overfitting. In light of this comparison (Figure 7), one must ask the question how the neural networks based model that has 9.4% HIA unit error[4] should be interpreted.

CONCLUSIONS

The greatest concern regarding this data set is interexperimental variability, and the lack of reliable data for the examined compounds. Over 30% of the compounds plotted in Figure 7 are outside the 20% error. It is reasonable to assume that such variability has a direct effect on model computation, its reliability and, indeed, its interpretability. On the other hand, the two different sources for Caco-2 cell permeability used in this study appear to be more consistent. Therefore, we propose that oral absorption and Caco-2 cell permeability data need to be modeled simultaneously, as shown in the present article. Such data are complementary, since other factors besides Caco-2 (epithelial) permeability play an important role during oral drug absorption, i.e., the area of the absorbing surface and the residence time of the drug at that surface. Furthermore, Caco-2 cell permeability is not bordered by a fixed interval (i.e., between 0 and 100%, like HIA), and is likely to attenuate the ambiguity caused by other mechanisms of absorption, ambiguity that sometimes originates from in vivo studies in humans.

In the present model, several properties were found to be relevant: variables related to H-bonding capacity, that reveal inverse correlations to permeability, and variables related to

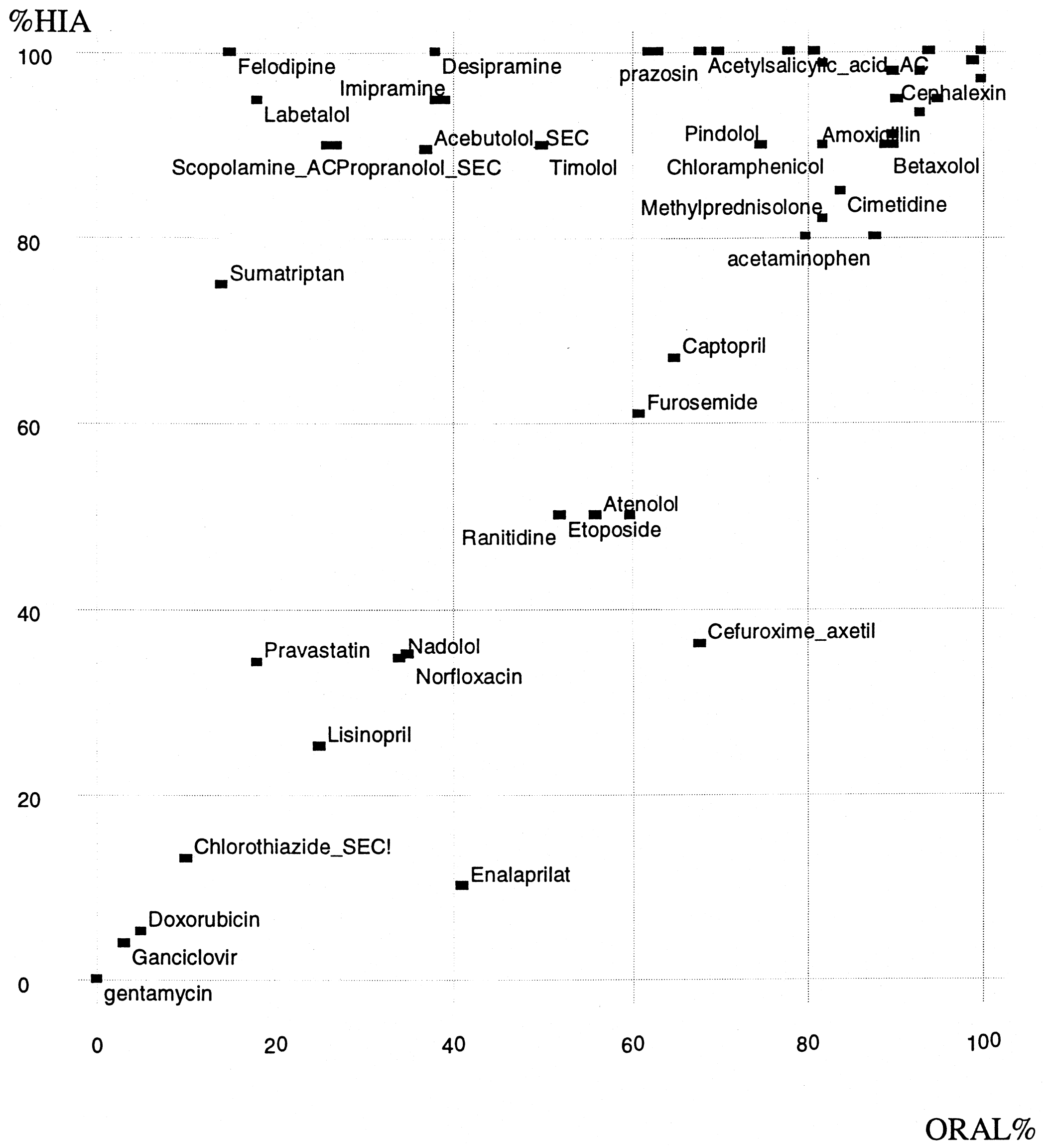

Figure 7. Percentage oral availability (%ORAL) versus percent intestinal absorption (%HIA) data in humans for 56 compounds, taken from references 28 and 4, respectively.

hydrophobic transferability, that reveal direct correlations to permeability. These findings are consistent with the QSAR paradigm outlined in Eq. (1), if one includes the minor effect of molecular size, reflected by the total molecular surface area. The model is nonlinear, as discussed above. The relationships between these properties and absorption estimates need not be sigmoidal unless the percentage scale is used for, e.g., HIA.

The parsimony principle was applied in several aspects, as follows: (1) single conformations were used to compute molecular surface areas; (2) the definitions of "polar" and "nonpolar" surfaces were made in a simplistic fashion; (3) simple and fast 2D descriptors were used to estimate other properties; and (4) the 1-PC model was selected from the projection analysis. Taken together, these aspects indicate an approach toward what we call a "minimalistic" model for oral absorption. We suggest that this model may serve as an extension of the "rule of five" test, for example, in a combinatorial library synthesis planning mode, by helping the chemist to focus on compounds with increased probability of oral absorption. With the caveat that its error rate is 30%, the minimalistic model can be used in a manner that might improve the drug discovery process.

ACKNOWLEDGMENTS

We thank Dr. Anna-Lena Ungell (AstraZeneca R&D Mölndal, Sweden), Dr. Ulf Norinder (AstraZeneca R&D Södertälje, Sweden), and Prof. Peter Langguth (University of Mainz, Germany) for fruitful discussions.

REFERENCES

1 Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug

discovery and development settings. *Adv. Drug Deliv. Rev.* 1997, **23,** 3–25

2 Van de Waterbeemd, H., Camenisch, G., Folkers, G., and Raevsky, O.A. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant. Struct.-Act. Relat.* 1996, **15,** 480–490

3 Norinder, U., Österberg, T., and Artursson, P. Theroetical calculations nad prediciton of Caco-2 cell permeability using MolSurf parametrization and PLS statistics. *Pharm. Res.* 1997, **14,** 1786–1791

4 Wessel, M.D., Jurs, P.C., Tolan, J.W., and Muskal, S.M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* 1998, **38,** 726–735

5 Palm, K., Luthman, K., Ungell, A.L., Strandlund, G., Beigi, F., Lundahl, P., and Artursson, P. Evaluation of dynamic polar molecular surface area as predictor of drug absorption: Comparison with other computational and experimental predictors. *J. Med. Chem.* 1998, **41,** 5382–5392

6 Krarup, L.H., Christensen, I.T., Hovgaard, L., and Frokjaer, S. Predicting drug absorption from molecular surface properties based on molecular dynamics simulations. *Pharm. Res.* 1998, **15,** 972–978

7 Raevsky, O.A., Grigor'ev, V.Y., Kireev, D., and Zefirov, N.S. Complete thermodynamic description of H-bonding in the framework of multiplicative approach. *Quant. Struct.-Act. Relat.* 1992, **11,** 49–64; HYBOT is available from pION, Inc., Cambridge, Massachussets, http://www.pion-inc.com

8 MolSurf is available from Qemist AB, Karlskoga, Sweden, http://members.xoom.com/Qemist/

9 VolSurf is available from MIA srl, Perugia, Italy, http://www.miasrl.com/volsurf.htm

10 SAVOL is integrated in the SYBYL suite, available from Tripos, Inc., St. Louis, Missouri, http://www.tripos.com

11 Hansch, C., and Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. American Chemical Society, Washington, D.C., 1995, pp. 125–168

12 CLOGP is available from Biobyte, Inc., Claremont, California, http://clogp.pomona.edu/

13 ACDLogP is available from ACD Labs, Toronto, Canada, http://www.acdlabs.com/

14 Guba, W., and Cruciani, G. Molecular field-derived descriptors for the multi-variate modeling of pharmacokinetic data. In: Molecular Modeling and Prediction of Bioactivity (Gundertofke, K. and Jørgensen, F.S., eds.), Kluwer Academic/Plenum Publishers, New York, 2000, pp. 89–94

15 Artursson, P., and Karlsson, J. Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells. *Biochem. Biophys. Res. Commun.* 1991, **175,** 880–885

16 Yazdanian, M., Glynn, S.L., Wright, J.L., and Hawi, A. Correlating partitioning and Caco-2 cell permeability of structurally diverse small molecular weight compounds. *Pharm. Res.* 1998, **15,** 1490–1494

17 See the ACD Labs webpage http://www.acdlabs.com/products/phys−chem−lab/logd/ for details

18 CONCORD 4.0.2 is available from Tripos, Inc., http://www.tripos.com

19 Shcherbukhin, V.V., and Olsson, T. SaSA (Synthesis and Structure Administration) program, personal communication.

20 Wold, S., Ruhe, A., Wold, H., and Dunn, W.J., III. The collinearity problem in linear regression. The partial least squares approach to generalised inverses. *J. Sci. Stat. Comput.* 1984, **5,** 735–743

21 Höskuldsson, A. PLS regression methods. *J. Chemom.* 1988, **2,** 211–228

22 Simca-P7 is available from Umetri AB, Umeå, Sweden, http://www.umetri.se/

23 Wold, S. Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* 1978, **20,** 397–405

24 Pieniaszek, H.J., Jr., Resetarits, D.E., Wilferth, W.W., Blumenthal, H.P., and Bates, T.R. Relative systemic availability of sulfapyridine from commercial enteric-coated and uncoated sulfasalazine tablets. *J. Clin. Pharmacol.* 1979, **19,** 39–45

25 Mandell, G.L., and Petri, W.A., Jr. Antimicrobial agents: Sulfonamides, trimethoprim-sulfamethoxazole, quinolones, and agents for urinary tract infections. In: *Goodman & Gilman's Pharmacological Basis of Therapeutics* (Hardman, J.G., Limbird, L.E., Molinoff, P.B., Ruddon, R.W. and Goodman Gillman, A., eds.), 9th Ed. McGraw-Hill, New York, 1996, pp. 1057–1072

26 Barthe, L., Woodley, J.F., Kenworthy, S., and Houin, G. An improved everted gut sac as a simple and accurate technique to measure paracellular transport across the small intestine. *Eur. J. Drug Metab. Pharmacokinet.* 1998, **23,** 313–323

27 Winiwarter, S., Bonham, N.M., Ax, F., Hallberg, A., Lennernäs, H., and Karlén, A. Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach. *J. Med. Chem.* 1998, **41,** 4939–4949

28 Benet, L.Z., Öie, S., and Schwartz, J.B. Design and optimization of dosage regimens; pharmacokinetic data. In: *Goodman & Gilman's Pharmacological Basis of Therapeutics* (Hardman, J.G., Limbird, L.E., Molinoff, P.B., Ruddon, R.W. and Goodman Gilman, A., eds.), 9th Ed. McGraw-Hill, New York, 1996, pp. 1707–1792

29 Clark, D.E. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. *J. Pharm. Sci.* 1999, **88,** 807–814

30 Bijlsma, P.B., Peeters, R.A., Groot, J.A., Dekker, P.R., Taminiau, J.A., and Van der Meer, R. Differential in vivo and in vitro intestinal permeability to lactulose and mannitol in animals and humans: A hypothesis. *Gastroenterology* 1995, **108,** 687–696

31 Langguth, P., and Östgren, M. Intestinal permeability and absorption of peptide drugs. In: *Clinical Pharmacology*, Vol. 15; Gramatte, T., and Weiss, M. (eds.). *Aspekte der intestinalen Absorption und der Modellentwicklung in Pharmakokinetik und Pharmacokynamik.* Zuckschwerdt Verlag, Munich, Germany, 1998

32 Ungell, Anna-Lena, personal communication