Journal of Enzyme Inhibition and Medicinal Chemistry, June 2004 Vol. 19 (3), pp. 237–248

Quantum Theoretic QSAR of Benzene Derivatives: Some Enzyme Inhibitors

CLAUDIU T. SUPURAN[a,*] and BRIAN W. CLARE[b,†]

[a]*Università degli Studi, Laboratorio di Chimica Inorganica e Bioinorganica, Via Gino Capponi 7, I-50121, Florence, Italy;* [b]*School of Biomedical and Chemical Science, The University of Western Australia, 35 Stirling Highway, Crawley W.A., Australia 6009*

(Received 18 January 2004)

Our previously developed approach to the development of QSAR equations for benzene derivatives, originally for phenylalkylamine hallucinogens, has been applied to four new systems: sulfonamide inhibitors of the enzymes carbonic anhydrase, thrombin, trypsin, and *Clostridium histolyticum* collagenase. The novel features involve the energies and nodal orientations of π-like orbitals, and an allowance for the symmetry of the benzene nucleus. The resulting equations give better fits, better predictivity and are more easily interpretable than those resulting from traditional QSAR methods.

Keywords: Enzyme inhibitor; Quantum QSAR; Orbital energy; Node orientation; Symmetry

INTRODUCTION

A recent publication introduced systematic procedures for obtaining QSARs for benzene derivatives using quantum chemical descriptors, and a method for resolving problems introduced by the symmetry of the benzene molecule.[1] A QSAR was presented for the phenylalkylamine hallucinogens in terms mainly of quantum theoretic descriptors, which included the energies of four near-frontier π-like orbitals and the orientation of the nodes in two of them. It is hypothesized that these orbitals interact with similar orbitals on the receptor, and this interaction is at a maximum when the energies of the interacting orbitals are similar, and when the nodes in the interacting orbitals nearly coincide. A computer program to calculate the orientation of these nodes is available from the authors.[2]

The present contribution is an extension of this work to some other receptors, namely carbonic anhydrase inhibitors,[3] inhibitors of the proteases thrombin and trypsin,[4] and *Clostridium histolyticum* collagenase.[5] We have previously developed QSARs for some of these ligands. All of these groups of ligands are substituted benzene derivatives. The determination of biological activities and the procedures for the quantum theoretic calculations have been described previously.[3,4,5] All molecular orbital calculations were at the AM1 level,[6] and solvation was allowed for using the COSMO procedure,[7] employing the MOPAC 93[8] molecular orbital package. Atomic charges were based on the calculated electrostatic potential.[9]

CALCULATIONS

The core of the statistical procedure has also been described in detail previously,[1] and involves "flipping" the drugs. Flipping consists of swapping the ortho (2,6) substituents with each other, and also the meta (3,5) substituents, as though the drug had rotated 180° around the 1,4 axis. Because in this rotation the angles between the nodes in the orbitals and the 1-position in the molecule would be transformed into their negatives, this change is also introduced. Regressions are carried out of the logarithm of biological activity on the physical descriptors with all combinations of drugs, flipped and unflipped, and that combination of flips producing the best fit is selected.

*Fax: +39-055-2757555. E-mail: claudiu.supuran@unifi.it
†Corresponding author. Tel.: +61-8-9337-7824. Fax: +61-8-9380-1005. E-mail: bwc@theochem.uwa.edu.au

ISSN 1475-6366 print/ISSN 1475-6374 online © 2004 Taylor & Francis Ltd
DOI: 10.1080/14756360410001689603

Because this involves a number of regressions that depends exponentially on the number of cases (drugs), it can be carried out rigorously only for small data sets. In this case it is done this way only for the CA inhibitors (27 cases), using the program FLIPALL. For the others an iterative approach is used, implemented in the program FLIPANNL. In the case of the CA inhibitors it is possible to compare the two approaches.

Having modified the data file by flipping the drugs when appropriate, the standard approaches of multiple regression are applied, using the "all possible subsets" algorithm of Furnival and Wilson[10] to obtain candidate models, and the Λ statistic to guard against colinearity.[11] Software implementing these procedures is available in the authors MARTHA package available on the WWW.[12]

RESULTS

Carbonic Anhydrase Inhibitors

The identities and activities of the inhibitors are presented in Table I, as previously described.[13,14] In our original reports on the QSAR of these compounds[15] we considered a wide range of calculated parameters, including beside the potential-based charges on the atoms of the sulfonamide moiety and the carbon bound to it, and HOMO and LUMO energies, polarizabilities, dipole moment vector, pKa, overall molecular dimensions and van der Waals volume, and electrophilic superdelocalizabilities of atoms. These were reduced in number in our subsequent report[3] because the number of compounds was reduced from 35 to 27 to homogenize the data set. For the purposes of this contribution, the number of descriptors was reduced still further, and in addition to the orbital energies and node angles, only the atomic charges, the local dipole index, an indicator variable I_p, (which is 1 if a styryl group is present in addition to the common benzene, and 0 otherwise) and the calculated lipophilicity were included. The atomic charges are known to be important from many studies. The local dipole index is the mean of the absolute difference in Mulliken charge between each bonded pair of atoms, that is, the magnitude of charge separation in the molecule, and we frequently find it to be important in QSAR studies.

The effectiveness of the compounds as inhibitors of both CA I and CA II were considered separately, and because of the small numbers of compounds it was possible to use both the FLIPALL and FLIPANNL procedures. For CA I the Fishers F ratio was initially 2.28 and improved to 25.02 with FLIPANNL and 25.97 with FLIPALL. The flip statuses and flip significances are compared in Table II.

TABLE I Structures and activities of the CA inhibitors

R_1	R_2	No	K_I CA I ($\times 10^6$ M)	K_I CA II ($\times 10^8$ M)
Phenyl	H	**A1**	18	27
2-Hydroxyphenyl	H	**A2**	35	41
2-Nitrophenyl	H	**A3**	9	21
4-Chlorophenyl	H	**A4**	25	28
4-Hydroxyphenyl	H	**A5**	14	19
4-Methoxyphenyl	H	**A6**	13	19
4-Dimethylaminophenyl	H	**A7**	10	8
4-Nitrophenyl	H	**A8**	13	5
4-Cyanophenyl	H	**A9**	4	11
3-Methoxy-4-hydroxyphenyl	H	**A10**	5	8
3,4-Dimethoxyphenyl	H	**A11**	7	3
3-Methoxy-4-acetoxyphenyl	H	**A12**	3	10
2,3-Dihydroxy-5-formylphenyl	H	**A13**	4	2
2-Hydroxy-3-methoxy-5-formylphenyl	H	**A14**	5	3
3,4,5-Trimethoxyphenyl	H	**A15**	5	3
3-Methoxy-4-hydroxy-5-bromophenyl	H	**A16**	12	4
2-Pyridyl	H	**A17**	2	9
3-Pyridyl	H	**A18**	4	8
4-Pyridyl	H	**A19**	4	5
Phenyl	Styryl	**A20**	20.9	0.56
Phenyl	4-Methoxystyryl	**A21**	19	1.50
Phenyl	4-Dimethylaminostyryl	**A22**	16	1.69
Phenyl	3,4,5-Trimethoxystyryl	**A23**	10.7	2.35
4-Methoxyphenyl	3,4,5-Trimethoxystyryl	**A24**	12.5	1.27
4-Methoxyphenyl	3-Nitrostyryl	**A25**	6.3	0.65
4-Aminophenyl	3,4,5-Trimethoxystyryl	**A26**	10.6	0.85
4-Phenylphenyl	3,4,5-Trimethoxystyryl	**A27**	25.0	2.48

TABLE II Results of the two flip regression procedures FLIPANNL and FLIPALL on the CA I inhibitor data of Table I

		FLIPANNL		FLIPALL	
Compound	Symmetry	Status	Significance	Status	Significance
A1	*	−1	0.790	−1	0.681
A2		1	0.012	1	0.001
A3		−1	0.649	−1	0.218
A4	*	1	0.022	−1	0.019
A5	*	−1	0.885	1	0.845
A6	*	−1	0.950	−1	0.757
A7	*	1	0.960	1	0.892
A8	*	1	0.994	1	0.952
A9	*	1	0.951	−1	0.964
A10		1	0.514	1	0.121
A11		1	0.002	1	0.002
A12		1	0.013	1	0.006
A13		−1	0.005	1	0.559
A14		1	0.005	1	0.006
A15	*	1	0.086	1	0.104
A16		1	0.005	1	0.001
A17		−1	0.074	−1	0.002
A18		−1	0.064	1	0.087
A19	*	1	0.068	1	0.030
A20	*	−1	0.901	1	0.397
A21	*	−1	0.613	−1	0.280
A22	*	1	0.779	1	0.424
A23	*	1	0.065	1	0.134
A24	*	1	0.740	1	0.600
A25	*	−1	0.003	−1	0.010
A26	*	−1	0.087	1	0.016
A27	*	1	0.048	−1	0.022

An asterisk indicates those compounds for which the flipped and unflipped forms are chemically indistinguishable. A status of −1 indicates a compound which has flipped from the input form. And a significance less than 0.05 indicates a compound which has a preference for one orientation over the other. Only unsymmetrical (non asterisked) compounds are expected to have significant flips.

For CA II the initial F ratio was 6.81 and both FLIPANNL and FLIPALL improved this to 38.86, giving identical solutions. Regressions were carried out in the output descriptors of the FLIP programs to reduce the number of variables from 15, which is more than 27 cases can support, and eliminate those that are statistically nonsignificant.

For CA I

$$\begin{aligned}\log C_I = {} & C_1Q_S + C_2Q_H + C_3 \log P + C_4E_{SH} \\ & + C_5E_H + C_6 \cos 2\Phi_H + C_7 \sin 2\Phi_H \\ & + C_8 \cos 4\Phi_L + C_9 \sin 4\Phi_L + C_{10} \\ N = 27, \quad & R^2 = 0.959, \quad Q^2 = 0.876, \\ F = 43.8, \quad & P = 5 \times 10^{-10}, \quad s = 0.08, \\ \Lambda = 4.01 & \end{aligned} \tag{1}$$

The regression coefficients and statistics for Equation (1) are presented in Table III.

Here N is the number of drugs in the regression, R^2 the squared multiple correlation coefficient, Q^2 the same based on the "leave one out" technique (i.e. on the predicted residuals), s the standard error of estimate and F the Fisher variance ratio. P is the probability based on F. In Table III α is the statistical significance of the variable in the presence of the other variables: a value of α greater than 0.05 indicates that the variable does not contribute significantly to the regression.[16] Such results are not presented, with one exception: if one of a sin/cos pair is significant, the other is also presented, even if not formally significant. This is because the two together define an optimum angle for the node, and should be considered as a single unit. To do otherwise would enforce a value of 0° or 90° on the angle, for which there is no justification. The statistic Λ, defined as:

$$\Lambda = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\lambda_i}$$

where n is the number of variables and the λ_i are the eigenvalues of the correlation matrix of independent variables,[11] measures the amount of colinearity in

TABLE III The statistics of Equation (1)

	1	2	3	4	5	6	7	8	9	10
C	−20.9	98.7	0.354	−0.736	0.221	−0.226	−0.042	0.021	−0.717	−23.7
σ	2.1	8.1	0.026	0.096	0.057	0.032	0.036	0.023	0.068	6.0
α	0.00000	0.00000	0.00000	0.00000	0.00114	0.00000	0.26406	0.38984	0.00000	.00012

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a t test.

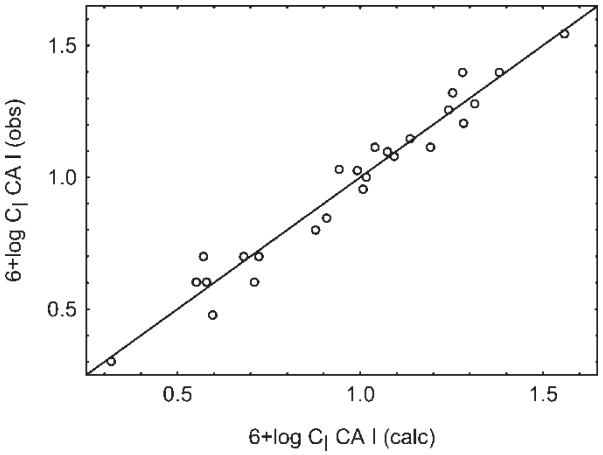


FIGURE 1. Plot of observed versus calculated (Equation 1) log inhibition constant for inhibitors of CA I.

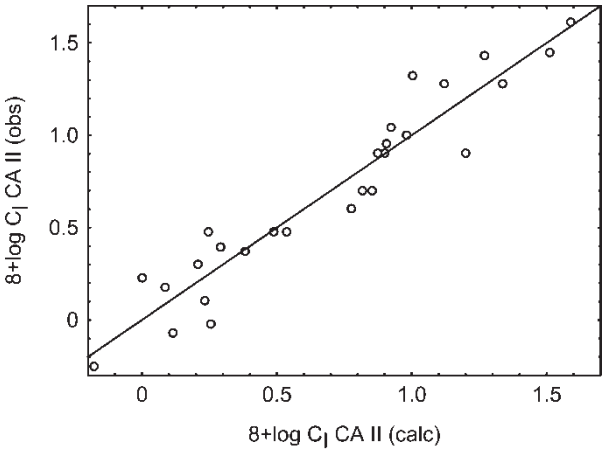


FIGURE 2. Plot of observed versus calculated (Equation 2) log inhibition constant for inhibitors of CA II.

TABLE IV The statistics of Equation (2)

	1	2	3	4	5	6	7	8	9
C	62.2	−0.160	−0.482	0.990	−0.263	0.222	0.066	0.270	−46.7
σ	9.4	0.033	0.108	0.355	0.059	0.080	0.061	0.056	8.1
α	0.00000	0.00016	0.00003	0.01217	0.00033	0.01261	0.29124	0.00015	0.00000

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a t test.

the equation. Values of Λ exceeding 5.0 indicate unacceptable colinearity, while for an orthogonal correlation matrix $\Lambda = 1$. F is the Fisher variance ratio, and P the probability (the statistical significance) based on this. s is the standard error of estimate. Figure 1 shows a plot of the calculated against observed $\log C_I$ for Equation (1). Our previous R^2 for this set of drugs using classical descriptors

TABLE V Sulfonylguanidines **B1**–**B24** prepared in the present study, with their inhibition data against human thrombin and human trypsin

$$RSO_2N=C(NH_2)_2$$
B1–B24

R	Compound	K_I (nM)[a] Thrombin	Trypsin
p-F-C_6H_4-	**B1**	240	1090
p-Cl-C_6H_4-	**B2**	225	1170
p-Br-C_6H_4-	**B3**	220	1230
p-CH_3-C_6H_4-	**B4**	290	1810
p-O_2N-C_6H_4-	**B5**	180	975
m-O_2N-C_6H_4-	**B6**	190	1100
o-O_2N-C_6H_4-	**B7**	320	1850
3-Cl-4-O_2N-C_6H_3-	**B8**	160	990
p-AcNH-C_6H_4-	**B9**	195	1070
p-H_2N-C_6H_4-	**B10**	95	1350
m-H_2N-C_6H_4-	**B11**	107	1145
C_6F_5-	**B12**	146	1350
o-HOOC-C_6H_4-	**B13**	240	1445
m-HOOC-C_6H_4-	**B14**	121	1500
p-HOOC-C_6H_4-	**B15**	104	1235
o-HOOC-C_6Br_4-	**B16**	225	1250
p-CH_3O-C_6H_4-	**B17**	240	1320
2,4,6-$(CH_3)_3$-C_6H_2-	**B18**	255	1450
4-CH_3O-3-H_2N-C_6H_3-	**B19**	103	1080
2-HO-3,5-Cl_2-C_6H_2-	**B20**	152	1420
4-Me_2N-C_6H_4-N=N-C_6H_4-	**B21**	134	1245
5-Dimethylamino-1-naphthyl-	**B22**	120	1150
1-Naphthyl	**B23**	136	1230
2-Naphthyl	**B24**	132	1300

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ± 5–10% of the shown values.

TABLE VI Derivatives **B25–B32** obtained from sulfaguanidine **B10** as lead, with their inhibition data against human thrombin and human trypsin

$$p-\mathrm{RNH-C_6H_4-SO_2N=C(NH_2)_2}$$
B25–B32

R	Compound	K_I (nM)[a] Thrombin	Trypsin
Cbz-*D*-Phe	**B25**	54	1285
ts-*D*-Phe	**B26**	43	1250
ts-*L*-Pro	**B27**	48	1445
ts-*D*-PhePro	**B28**	12	1315
Cbz-*D*-PhePro	**B29**	13	1360
ts-GlyHis	**B30**	18	1455
ts-β-AlaHis	**B31**	15	1350
ts-*L*-ProGly	**B32**	21	1400

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ± 5–10% of the shown values. *Cbz = $PhCH_2OCO$; ts = p-$MeC_6H_4SO_2NHCO$-; these groups acylate the amino-terminal H_2N moiety. When configuration is not specified, *L*-amino acid moieties were employed. The usual polypeptide formalism is used: the amino-terminal residue is written first (and it is always protected either by the Cbz or the ts moieties), whereas the carboxyterminal residue is acylating the sulfaguanidine N-4 amino group.

and without flipping was 0.89, and the use of nodal angles did not improve it.[3]

For CA II

$$\log C_I = C_1Q_H + C_2\log P + C_3E_{SH} + C_4E_{SL} + C_5\cos 2\Phi_H + C_6\sin 2\Phi_H + C_7\cos 4\Phi_L + C_8\sin 4\Phi_L + C_9 \quad (2)$$

$$N = 27,\ R^2 = 0.911,\ Q^2 = 0.780,\ F = 22.9,\ P = 6\times10^{-8},\ s = 0.18,\ \Lambda = 1.68$$

Figure 2 shows a plot of the calculated against observed $\log C_I$ for Equation (2), and the regression coefficients and statistics for Equation (2) are given in Table IV.

Thus in addition to the orbital parameters, only the charges on the sulfonamide S and H, and the lipophilicity are significant. The QSAR for CA I is substantially better than that for CA II, but it does have more colinearity. Our previous R^2 was 0.68 without the angle descriptors, and 0.70 with them but not using the flip procedure.[3]

Thrombin Inhibitors

In our previous study of trypsin and thrombin inhibitors[4] we again considered a wide range of descriptors. For the purposes of the present contribution these will be restricted to, in addition to the FOPA variables, the lipophilicity, the diagonal components of the polarizability tensor, the local dipole index, the solvation energy calculated as the difference between ΔH_f obtained by the COSMO procedure and that calculated for vacuum, the charges on the atoms of the sulfonamide group, and indicator variables I_A and I_S. For sulfonylguanidines both I_A and I_S were zero, while for sulfonylaminoguanidines I_A was 1 and for isoureas I_S was 1. This is a Free-Wilson approach to distinguishing between the activities of drugs belonging to each of the three pharmacophoric

TABLE VII Sulfonyl-*O*-methyl-isoureas **C1–C24** prepared in the present study, with their inhibition data against human thrombin and human trypsin

$$\mathrm{RSO_2N=C(NH_2)OMe}$$
C1–C24

R	Compound	K_I (nM)[a] Thrombin	Trypsin
p-F-C_6H_4-	**C1**	280	1200
p-Cl-C_6H_4-	**C2**	266	1325
p-Br-C_6H_4-	**C3**	265	1370
p-CH_3-C_6H_4-	**C4**	328	2150
p-O_2N-C_6H_4-	**C5**	189	1235
m-O_2N-C_6H_4-	**C6**	213	1375
o-O_2N-C_6H_4-	**C7**	355	2340
3-Cl-4-O_2N-C_6H_3-	**C8**	177	1200
p-AcNH-C_6H_4-	**C9**	202	1345
p-H_2N-C_6H_4-	**C10**	106	1580
m-H_2N-C_6H_4-	**C11**	99	1370
C_6F_5-	**C12**	153	1425
o-HOOC-C_6H_4-	**C13**	325	1550
m-HOOC-C_6H_4-	**C14**	197	1595
p-HOOC-C_6H_4-	**C15**	133	1340
o-HOOC-C_6Br_4-	**C16**	239	1300
p-CH_3O-C_6H_4-	**C17**	276	1460
2,4,6-$(CH_3)_3$-C_6H_2-	**C18**	348	1785
4-CH_3O-3-H_2N-C_6H_3-	**C19**	96	1235
2-HO-3,5-Cl_2-C_6H_2-	**C20**	170	1550
4-Me_2N-C_6H_4-N=N-C_6H_4-	**C21**	169	1350
5-Dimethylamino-1-naphthyl-	**C22**	138	1375
1-Naphthyl	**C23**	154	1425
2-Naphthyl	**C24**	147	1445

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ± 5–10% of the shown values.

TABLE VIII Derivatives obtained from sulfanilyl-*O*-methyl-isourea **C10** as lead and their inhibition data against human thrombin and human trypsin

$$p-\mathrm{RNH}-\mathrm{C_6H_4}-\underset{\mathbf{C25-C32}}{\mathrm{SO_2N=C(NH_2)OMe}}$$

R	Compound	K_I (nM)[a] Thrombin	Trypsin
Cbz-*D*-Phe	**C25**	62	1425
ts-*D*-Phe	**C26**	60	1320
ts-*L*-Pro	**C27**	63	1500
ts-*D*-PhePro	**C28**	18	1420
Cbz-*D*-PhePro	**C29**	16	1435
ts-GlyHis	**C30**	21	1550
ts-β-AlaHis	**C31**	19	1420
ts-*L*-ProGly	**C32**	27	1540

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ± 5–10% of the shown values. *Cbz = $PhCH_2OCO$; ts = p-$MeC_6H_4SO_2$-NHCO-; these groups acylate the amino-terminal H_2N moiety. When configuration is not specified, it means that *L*-amino acid moieties were employed. The usual polypeptide formalism is used: the amino-terminal residue is written first (and it is always protected either by the Cbz or the ts moieties), whereas the carboxyterminal residue is acylating the sulfanilyl-*O*-methyl-isourea N-4 amino group.

classes. For the inhibitors in Tables V–X.

$$\begin{aligned}\log K_I = {} & C_1 \log P + C_2 I_S + C_3 \Pi_{xx} + C_4 \Pi_{yy} \\ & + C_5 \Pi_{zz} + C_6 D_1 + C_7 \Delta H_S + C_8 E_{SH} \\ & + C_9 E_H + C_{10} E_L + C_{11} E_{SL} + C_{12} Q_H \\ & + C_{13} \cos 2\Phi_H + C_{14} \sin 2\Phi_H + C_{15} \qquad (3) \\ N = 96, \quad & ; R^2 = 0.984, \quad Q^2 = 0.978, \\ F = 367.7, \quad & P = 4 \times 10^{-67}, \quad s = 0.06, \\ \Lambda = 3.46 & \end{aligned}$$

The regression coefficients and statistics for Equation (3) are presented in Table XI. F improved on flipping from 47.4 to 287.2. Noting that the signs of E_{SH} and E_H are opposite, as are those of E_{SL} and E_L, and their magnitudes are comparable, and also that Π_{xx}, Π_{yy} and Π_{zz} are of comparable size and the same sign, a simpler equation is suggested:

$$\begin{aligned}\log K_I = {} & C_1 \log P + C_2 \log I_S + C_3 D_1 + C_4 \Delta H_s \\ & + C_5 Q_H + C_6 \cos 2\Phi_H + C_7 \sin 2\Phi_H \mathrm{cr} \\ N = 96, \quad & R^2 = 0.982, Q^2 = 0.977, F = 454.9, \\ P = 3 \times 10^{-69}, & s = 0.11, \Lambda = 2.41 \end{aligned} \qquad (4)$$

Here $\Pi = (\Pi_{xx} + \Pi_{yy} + \Pi_{zz})/3$, $\Delta_L = E_{SL} - E_L$, and $\Delta_H = E_H - E_{SH}$

The regression coefficients and statistics for Equation (4) are presented in Table XII. Figure 3 shows a plot of the calculated against observed log K_I for Equation (4). Our previous best equation for this data[4] without flipping involved Π, Δ_H, Δ_L, and the dipole moment, and had an R^2 of 0.75.

This equation indicates that the sulfonylisoureas are less active than the sulfonylguanidines, and the sulfonylaminoguanidines are about as active. Higher activity is favored by a small solvation energy, a low lipophilicity, a smaller separation of charge, and a high polarizability. The latter reflects the influence of the bulky peptide tail on the very active compounds. All terms except 6 are of extremely

TABLE IX Sulfonylaminoguanidines **D1–D24** prepared in the present study, with their inhibition data against human thrombin and human trypsin

$$\mathrm{RSO_2NHN=C(NH_2)_2}$$
$$\mathbf{D1-D24}$$

R	Compound	K_I (nM)[a] Thrombin	Trypsin
p-F-C_6H_4-	**D1**	225	1025
p-Cl-C_6H_4-	**D2**	212	1100
p-Br-C_6H_4-	**D3**	203	1215
p-CH_3-C_6H_4-	**D4**	270	1775
p-O_2N-C_6H_4-	**D5**	166	990
m-O_2N-C_6H_4-	**D6**	170	1235
o-O_2N-C_6H_4-	**D7**	324	1800
3-Cl-4-O_2N-C_6H_3-	**D8**	154	1010
p-AcNH-C_6H_4-	**D9**	172	1025
p-H_2N-C_6H_4-	**D10**	91	1425
m-H_2N-C_6H_4-	**D11**	88	1400
C_6F_5-	**D12**	123	1350
o-HOOC-C_6H_4-	**D13**	205	1400
m-HOOC-C_6H_4-	**D14**	112	1520
p-HOOC-C_6H_4-	**D15**	97	1335
o-HOOC-C_6Br_4-	**D16**	213	1200
p-CH_3O-C_6H_4-	**D17**	227	1275
2,4,6-$(CH_3)_3$-C_6H_2-	**D18**	219	1345
4-CH_3O-3-H_2N-C_6H_3-	**D19**	219	1100
2-HO-3,5-Cl_2-C_6H_2-	**D20**	98	1100
4-Me_2N-C_6H_4-N=N-C_6H_4-	**D21**	139	1355
5-Dimethylamino-1-naphthyl-	**D22**	130	1200
1-Naphthyl	**D23**	125	1200
2-Naphthyl	**D24**	129	1285

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ± 5–10% of the shown values.

TABLE X Derivatives **D25**–**D32** obtained from sulfanilylaminoguanidine **D10** as lead, with their inhibition data against human thrombin and human trypsin

$$p - \mathrm{RNH{-}C_6H_4} - \underset{\textbf{D25–D32}}{\mathrm{SO_2NHN{=}C(NH_2)_2}}$$

R	Compound	K_I (nM)[a]	
		Thrombin	Trypsin
Cbz-*D*-Phe	**D25**	50	1315
Ts-*D*-Phe	**D26**	44	1340
Ts-*L*-Pro	**D27**	43	1530
Ts-*D*-PhePro	**D28**	10	1450
Cbz-*D*-PhePro	**D29**	11	1400
ts-GlyHis	**D30**	15	1555
ts-β-AlaHis	**D31**	10	1390
ts-*L*-ProGly	**D32**	15	1560

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors (data not shown) were ± 5–10% of the shown values. *Cbz = $PhCH_2OCO$; ts = p-$MeC_6H_4SO_2$-NHCO-; these groups acylate the amino-terminal H_2N moiety. When configuration is not specified, *L*-amino acid moieties were employed. The usual polypeptide formalism is used: the amino-terminal residue is written first (and it is always protected either by the Cbz or the ts moieties), whereas the carboxyterminal residue is acylating the sulfaguanidine N-4 amino group.

high statistical significance, 7 and 10 ($\sin 2\Phi_H$ and Π) being the highest.

Trypsin Inhibitors

The same compounds and same small pool of descriptors were employed as for thrombin inhibition. The equation obtained was:

$$\begin{aligned}\log K_I = {} & C_1 I_S + C_2 \Pi_{xx} + C_3 \Pi_{yy} + C_4 E_{SH} \\ & + C_5 E_H + C_6 E_L + C_7 \cos 2\Phi_H \\ & + C_8 \sin 2\Phi_L + C_9 \cos 4\Phi_L \\ & + C_{10} \sin 4\Phi_L + C_{11} \qquad (5)\end{aligned}$$

$$N = 96, \quad R^2 = 0.650, \quad Q^2 = 0.549,$$
$$F = 15.7, \quad P = 1 \times 10^{-15}, \quad s = 0.04,$$
$$\Lambda = 2.29$$

The regression coefficients and statistics for Equation (5) are presented in Table XIII. F improved on flipping from 4.0 to 23.2. Figure 4 shows a plot of the calculated against observed $\log K_I$ for Equation (5). Our previous best result for this data without flipping involved only Δ_H and Δ_L and had an R^2 of 0.31.[4] This correlation is similar to that for thrombin inhibition, but notably the dependence on polarizability has split; a large polarizability in the longest axis of the molecule is favorable, but in the medium axis it is detrimental. The term of highest statistical significance is 10 ($\sin 4\Phi_L$) although 1 and 3 are also high. While the R^2 and Q^2 for Equation (5) are poor, this reflects the fact that these compounds vary little in activity (their activity range is less than 0.3 log units). The standard error of estimate s is very good, and is comparable to the estimated experimental

TABLE XI The statistics of Equation (3)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
C	0.085	0.168	-8.4×10^{-4}	-1.44×10^{-3}	-1.92×10^{-3}	0.503	2.65×10^{-3}	0.158	−0.181	−0.119	0.110	−1.76	−0.018	0.148	−6.2
α	0.008	0.019	1.6×10^{-4}	1.8×10^{-4}	2.1×10^{-4}	0.089	3.2×10^{-4}	0.031	0.021	0.034	0.027	0.30	0.014	0.008	0.3
σ	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000	0.00069	0.00009	0.00000	0.22224	0.00000	0.31130

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a *t* test.

TABLE XII The statistics of Equation (4)

	1	2	3	4	5	6	7	8	9	10	11
C	0.0898	0.164	0.531	2.74×10^{-3}	−1.88	−0.008	0.156	0.128	−0.183	-3.9×10^{-3}	−5.92
α	0.0081	0.019	0.080	3.3×10^{-4}	0.28	0.014	0.008	0.023	0.021	1.7×10^{-4}	0.11
σ	0.00000	0.00000	0.00000	0.00000	0.00000	0.56452	0.00000	0.00000	0.00000	0.00000	0.00000

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a t test.

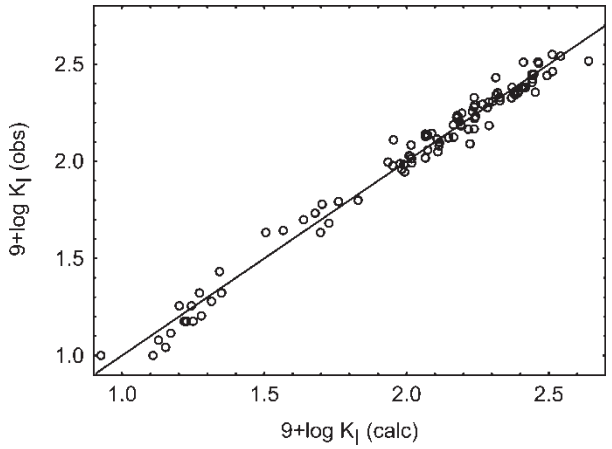


FIGURE 3. Plot of observed versus calculated (Equation 3) log inhibition constant for inhibitors of thrombin.

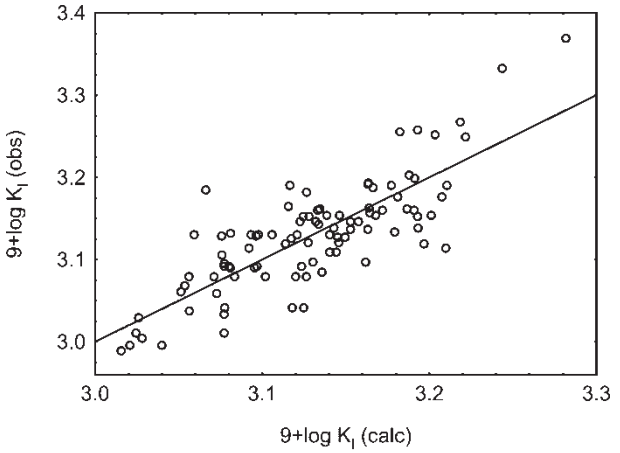


FIGURE 4. Plot of observed versus calculated (Equation 5) log inhibition constant for inhibitors of trypsin.

error, which is similar in magnitude for all of the classes of drugs.

Clostridium histolyticum Collagenase Inhibitors

We have not previously reported a QSAR on these compounds, the preparation and testing of which we described recently.[5] For the compounds in Table XIV we calculated the orbital parameters and two other sets of descriptors. The first of these was the lipophilicity and volume of the substituents on the five substitution sites on the benzene ring, and the second the atomic charges of the sulfonamide group, the solvation energy, the local dipole index. Indicator variables I_{HX} (zero for the carboxylic acids and 1 for the hydroxamates), and I_{Cl} (zero for the nitrobenzene derivatives and 1 for the chlorobenzenes) were also included. Two separate cases were considered: all of the compounds pooled, and only the hydroxamates. All four sets of calculations gave satisfactory results, but we present only two: the volume/lipophilicity data with the pooled data, and the charge-polarizability-solvation energy data with the hydroxamates. For the first of these flipping was carried out on the substituent data and the angle data, and for the second, of course, only on the angle data. The first set gave the equation:

$$\begin{aligned} \text{Log}\,K_I = {} & C_1 I_{HX} + C_2\pi_3 + C_3\pi_6 + C_4 V_2 + C_5 V_3 \\ & + C_6 V_4 + C_7 V_6 + C_8 E_{SL} + C_9 \cos 4\Phi_L \\ & + C_{10} \sin 4\Phi_L + C_{11} \\ N = 102, \quad & R^2 = 0.978, \quad Q^2 = 0.973, \\ F = 404.4, \quad & P = 8 \times 10^{-71}, \quad s = 0.20, \\ \Lambda = {} & 1.51 \end{aligned} \tag{6}$$

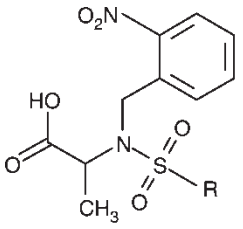
F improved on flipping from 108.4 to 235.1. The regression coefficients and statistics for Equation (6) are presented in Table XV. Figure 5

TABLE XIII The statistics of Equation (5)

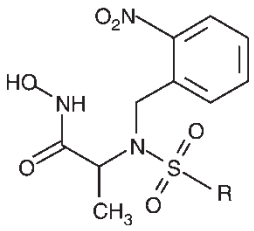
	1	2	3	4	5	6	7	8	9	10	11
C	0.059	-4.03×10^{-4}	6.06×10^{-4}	0.065	−0.046	0.059	0.0194	0.0177	0.045	0.0417	−5.61
α	0.010	1.0×10^{-4}	1.2×10^{-4}	0.020	0.014	0.016	0.0103	0.0063	0.012	0.0053	0.18
σ	0.00000	0.00009	0.00000	0.00206	0.00109	0.00067	0.06361	0.00587	0.00051	0.00000	0.00000

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a t test.

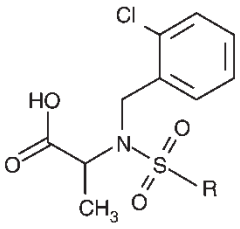
TABLE XIV Inhibition of *Clostridium histolyticum* collagenase (ChC) with the carboxylic acids **E1–E25, G1–G25** and the corresponding hydroxamates **F1–F25, H1–H25**



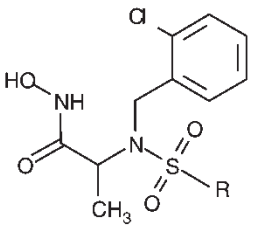
E1-E25



F1-F25



G1-G25



H1-H25

R	Compound	K_I[a] (μM)	Compound	K_I[a](nM)
C_6H_5-	**E1**	27	**F1**	59
4-F-C_6H_4-	**E2**	12	**F2**	30
4-Cl-C_6H_4-	**E3**	11	**F3**	32
4-Br-C_6H_4-	**E4**	10	**F4**	33
4-CH_3-C_6H_4-	**E5**	17	**F5**	41
4-O_2N-C_6H_4-	**E6**	5.0	**F6**	13
3-O_2N-C_6H_4-	**E7**	5.2	**F7**	10
2-O_2N-C_6H_4-	**E8**	4.4	**F8**	11
3-Cl-4-O_2N-C_6H_3-	**E9**	3.1	**F9**	9
4-AcNH-C_6H_4-	**E10**	3.3	**F10**	12
4-BocNH-C_6H_4-	**E11**	2.6	**F11**	9
3-BocNH-C_6H_4-	**E12**	2.5	**F12**	8
4-Ac-C_6H_4-	**E13**	2.0	**F13**	10
C_6F_5-	**E14**	0.3	**F14**	5
3-CF_3-C_6H_4	**E15**	0.4	**F15**	5
2,5-$Cl_2C_6H_3$	**E16**	3.4	**F16**	14
4-CH_3O-C_6H_4-	**E17**	5.7	**F17**	19
2,4,6-$(CH_3)_3$-C_6H_2-	**E18**	6.0	**F18**	21
4-CH_3O-3-BocNH-C_6H_3-	**E19**	2.5	**F19**	9
2-HO-3,5-Cl_2-C_6H_2-	**E20**	2.7	**F20**	10
3-HOOC-C_6H_4-	**E21**	2.1	**F21**[b]	8
4-HOOC-C_6H_4-	**E22**	1.9	**F22**[b]	7
1-Naphthyl	**E23**	1.6	**F23**	6
2-Naphthyl	**E24**	1.8	**F24**	8
5-Me_2N-1-naphthyl-	**E25**	2.1	**F25**	9
Quinolin-8-yl	**E26**	2.0	**F26**	9
C_6H_5-	**G1**	27	**H1**	59
4-F-C_6H_4-	**G2**	12	**H2**	30
4-Cl-C_6H_4-	**G3**	11	**H3**	32
4-Br-C_6H_4-	**G4**	10	**H4**	33
4-CH_3-C_6H_4-	**G5**	17	**H5**	41
4-O_2N-C_6H_4-	**G6**	5.0	**H6**	13
3-O_2N-C_6H_4-	**G7**	5.2	**H7**	10
2-O_2N-C_6H_4-	**G8**	4.4	**H8**	11
3-Cl-4-O_2N-C_6H_3-	**G9**	3.1	**H9**	9
4-AcNH-C_6H_4-	**G10**	3.3	**H10**	12
4-BocNH-C_6H_4-	**G11**	2.6	**H11**	9
3-BocNH-C_6H_4-	**G12**	2.5	**H12**	8
C_6F_5-	**G13**	0.3	**H13**	5
3-CF_3-C_6H_4	**G14**	0.4	**H14**	5
2,5-$Cl_2C_6H_3$	**G15**	3.4	**H15**	14
4-CH_3O-C_6H_4-	**G16**	5.7	**H16**	19

TABLE XIV – *continued*

R	Compound	K_I[a] (μM)	Compound	K_I[a](nM)
2,4,6-$(CH_3)_3$-C_6H_2-	**G17**	6.0	**H17**	21
4-CH_3O-3-BocNH-C_6H_3-	**G18**	2.5	**H18**	9
2-HO-3,5-Cl_2-C_6H_2-	**G19**	2.7	**H19**	10
3-HOOC-C_6H_4-	**G20**	2.1	**H20**[b]	8
4-HOOC-C_6H_4-	**G21**	1.9	**H21**[b]	7
1-Naphthyl	**G22**	1.6	**H22**	6
2-Naphthyl	**G23**	1.8	**H23**	8
5-Me_2N-1-naphthyl-	**G24**	2.1	**H24**	9
Quinolin-8-yl	**G25**	2.0	**H25**	9

[a] K_I values were obtained from Dixon plots using a linear regression program, from at least three different assays. Errors were around ± 10% (from at least three determinations). [b] The C_6H_4-COOH moiety transformed into C_6H_4-CONHOH.

TABLE XV The statistics of Equation (6)

	1	2	3	4	5	6	7	8	9	10	11
C	−2.35	−0.292	−0.584	-6.5×10^{-3}	-1.97×10^{-3}	-5.8×10^{-3}	7.5×10^{-3}	0.853	−0.080	0.351	−8.20
α	0.04	0.113	0.144	2.7×10^{-3}	7.7×10^{-4}	8.1×10^{-4}	2.3×10^{-23}	0.079	0.030	0.034	0.04
σ	0.00000	0.01139	0.00010	0.01900	0.01223	0.00000	0.00175	0.00000	0.00781	0.00000	0.00000

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a *t* test.

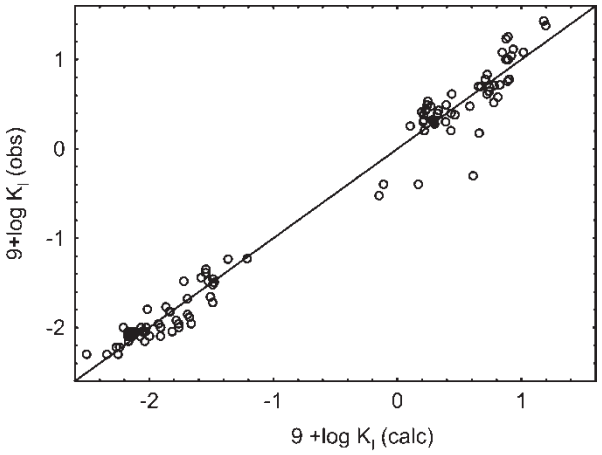


FIGURE 5. Plot of observed versus calculated (Equation 6) log inhibition constant for carboxylic acid and hydroxamate inhibitors of *C. histolyticum* collagenase.

shows a plot of the calculated against observed $\log K_I$ for Equation (6). We previously obtained without flipping an R^2 of 0.97 in an 11-descriptor equation involving also charges on the meta and para carbon atoms.[17]

The term of highest statistical significance is 1 (I_{HX}) as would be expected, as this term groups the data into the very active hydroxamates and the less active carboxylic acids. Terms 8 and 10 (E_{SL} and $\sin 4\Phi_L$), approximately of equal significance, are the next most important terms.

The second set (the hydroxamates only) gave the equation:

$$\begin{aligned}\log K_I = {} & C_1\Pi_{yy} + C_2\Pi_{zz} + C_3Q_S + C_4D_1 \\ & + C_5\Delta H_S + C_6E_{SL} + C_7\log P + C_8I_{C1} \\ & + C_9\cos 2\Phi_H + C_{10}\sin 2\Phi_H \\ & + C_{11}\cos 4\Phi_L + C_{12}\sin 4\Phi_L + C_{13} \qquad (7)\end{aligned}$$

$$N = 51, \quad R^2 = 0.918, \quad Q^2 = 0.860,$$

$$F = 35.4, \quad P = 6\times10^{-17}, \quad s = 0.09,$$

$$\Lambda = 2.82$$

F improved on flipping from 9.58 to 35.3. The regression coefficients and statistics of Equation (7) are presented in Table XVI. We previously obtained without flipping an R^2 of 0.84 in an equation that also involved the charges on the benzene-ring carbon atoms.[17] Figure 6 shows a plot of the calculated against observed $\log K_I$ for Equation (7). The term

TABLE XVI The statistics of Equation (7)

	1	2	3	4	5	6	7	8	9	10	11	12	13
C	-4.87×10^{-3}	-2.88×10^{-3}	−2.92	0.0194	0.575	−0.157	−0.157	−0.507	0.084	0.037	−0.143	−0.201	−4.85
α	7.0×10^{-4}	6.1×10^{-4}	0.77	0.0046	0.054	0.041	0.041	0.058	0.022	0.034	0.022	0.025	0.79
σ	0.00000	0.00003	0.00054	0.00016	0.00000	0.00045	0.00045	0.00000	0.00043	0.33317	0.00000	0.00000	0.00001

The σ are the standard errors of estimate of the parameters and α the corresponding probabilities of obtaining the observed values if the true value were zero, as determined by a *t* test.

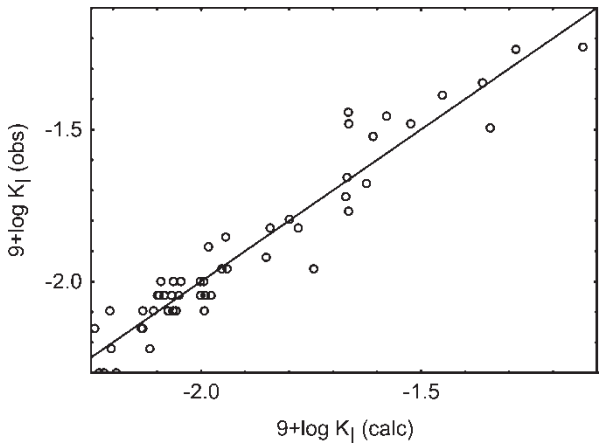


FIGURE 6 Plot of observed versus calculated (Equation 7) log inhibition constant for hydroxamate inhibitors of *C. histolyticum* collagenase.

of highest significance here is 5 (ΔH_S), but 1, 8 and 12 are almost as good.

Randomization Trial

To test the stability of the solutions to the problems described above a randomization trial was conducted. For each problem, with all variables included, the flip procedure was carried out with the dependent variable randomly reassigned in 1000 trials. The mean and standard deviation of the Fishers F and the maximum F, and also the fraction of runs with F greater than F for the non-randomized data were calculated for each system. The results are presented in Table XVII.

As would be expected high formal significances are obtained with randomized data. The results confirm that the significances obtained with the unmodified data are very much higher, and the randomization procedure provides a check on whether or not the results are valid. Only for the CA inhibition data are any F values obtained that are greater than those for the nonrandomized data, and even with that data, the fraction of such events is less than the 0.05 required for statistical significance. The CA inhibition included only 27 compounds, and it is clear that this is at or below the minimum number of compounds for which useful results can be obtained.

DISCUSSION

Satisfactory equations were obtained for both CA I and CA II. The number of compounds studied here, 27, is close to the minimum for which this technique could be applied. Table II indicates that in most cases, those flips are significant where the symmetry of the compound does not prohibit this, although the number of compounds is too small for complete confidence. Equation (1) has, by the Λ value, an appreciable colinearity, but is otherwise satisfactory. Surprisingly, the colinearity involves not the two large and opposite contributions from Q_H and Q_S ($R^2 = 0.25$), but a mixture of Q_H, log P, and $\sin 4\Phi_L$. Equation (2) is in general not as good as Equation (1), but it has almost no colinearity. A comparison of the two equations indicates an approach to attaining selectivity for inhibitors of one isozyme over the other: they are oppositely affected by log P, and also by the direction of the nodes in both the HOPO and LUPO. For this small number of compounds the randomization test indicates instability by the relatively large values of the mean F and also the large standard deviation. The results with the other systems are far more satisfactory, probably because of the large number of cases.

In all cases the orbital descriptors, that is the energies of the four orbitals resembling the HOMO and LUMO of benzene, and the orientation of their nodes, are the best or close to the best predictors. Unlike the case of the hallucinogens,[1] other descriptors also make a very significant contribution. The other descriptors employed here are mostly very easy to interpret physically.

It is uncertain whether it is the orbital energies themselves that determine activity, or the differences Δ_H and Δ_L. Both are plausible. Orbitals on two species, drug and receptor, interact most strongly when their energies are nearly equal. This could be reflected in a dependence of activity on orbital energies of the drug, if these are not too close to those of the receptor. Alternatively, the differences in energy, Δ_H and Δ_L, may reflect the difficulty of distorting the orbitals of one participant to match the other – zero in the case of benzene itself.

TABLE XVII Results of 1000 flip regressions on enzyme inhibitor systems with activities randomised

System	Original F	Mean F (randomised)	Std. dev. F (randomised)	Maximum F (randomised)	Fraction $F_{rand} > F_{orig}$
Carbonic Anhydrase I	25.02	7.09	5.37	65.22	0.015
Carbonic Anhydrase II	38.86	7.19	5.75	75.37	0.004
Thrombin	287.21	14.83	2.76	27.48	0.000
Trypsin	23.25	10.83	2.21	23.02	0.000
Collagenase (all)	235.13	11.90	1.88	19.69	0.000
Collagenase (Hydroxamates)	35.30	7.15	2.57	19.13	0.000

In QSAR work we can never be entirely confident that we have isolated the most relevant descriptors. Thus with the thrombin and trypsin inhibitors here we give much weight to polarizability. Much of the variability in this is due to the bulky peptide tails attached to the most active compounds, and it is impossible to be sure that some other variable related to these is not responsible for the effect. Regression analysis works best in designed data sets. In chemistry, however, these are by the nature of chemical molecules unachievable. We do not even approach the next best situation, which is of randomly selected molecules, but use intuitively guided selection, which may be to some degree imposing our intuitions on the results. Thus, as with the collagenase inhibitors, we can get disparate sets of descriptors that explain the data equally well, and either could be taken as the physical model.

As the randomization trial shows, formal statistical significances obtained by flip regression cannot be taken at face value. The procedure results in high apparent significance for random data, and only the decrease in significance on randomization is an indicator of validity. The likelihood of apparent significance of randomized results increases rapidly as the number of drugs falls below 30, and it seems probable that 30 is the practical minimum number of compounds required for valid results. The flip procedure needs to be used with caution until its properties are better understood.

References

[1] Clare, B.W. (2002) *J. Comput.-Aided Mol. Des.* **16**, 611–633.
[2] File nodangle.zip from site http://www.chem.uwa.edu.au/research/bclare
[3] Supuran, C.T. and Clare, B.W. (2001) *SAR and OSAR in Environmental Research* **12**, 17–29.
[4] Supuran, C.T., Scozzafava, A., Briganti, F. and Clare, B.W. (2000) *J. Med. Chem.* **43**, 1793–1806.
[5] Clare, B.W., Scozzafava, A. and Supuran, C.T. (2001) *J. Med. Chem.* **44**, 2253–2258.
[6] Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P. (1985) *J. Amer. Chem. Soc.* **107**, 3902–3909.
[7] Barone, V. and Cossi, M. (1998) *J. Phys. Chem.* **A102**, 1995–2001.
[8] MOPAC 93.00 (1993) Stewart, J.J.P., Fujitsu Ltd., Tokyo, Japan, also Stewart, J.J.P. MOPAC93 Release 2. *QCPE Bull* (1995) **15**, 1314 (Copyright Fujitsu 1993, all rights reserved).
[9] Besler, B.H., Merz, K.M. and Kollman, P.A. (1990) *J. Comput. Chem.* **11**, 431–439.
[10] Furnival, G.M. and Wilson, R.W. (1974) *Technometrics* **16**, 499–511.
[11] Chatterjee, S. and Price, B. (1977) *Regression Analysis by Example*, 1st Ed. (Wiley, New York), pp 199–200.
[12] File martha.zip from site http://www.chem.uwa.edu.au/research/bclare
[13] Supuran, C.T., Nicolae, A. and Popescu, A. (1996) *Eur. J. Med. Chem.* **31**, 431–438.
[14] Supuran, C.T., Popescu, A., Ilisiu, M., Constandache, A. and Banciu, M.D. (1996) *Eur. J. Med. Chem.* **31**, 439–447.
[15] Supuran, C.T. and Clare, B.W. (1998) *Eur. J. Med. Chem.* **33**, 489–500.
[16] Edwards, A.L. (1985) *Multiple Regression and the Analysis of Variance and Covariance*, 2nd Ed. (Freeman, New York), pp 49–50.
[17] Supuran, C.T. and Clare, B.W. unpublished data.