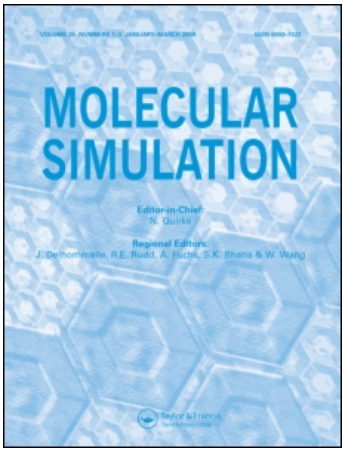

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

A Problem if Binary Molecular Substructural Variables are Used in Simulation Experiments

Peter P. Mager[a]
[a] Research Group for Pharmacochemistry, Institute of Pharmacology and Toxicology of the University, Leipzig, Saxony, Germany

Online publication date: 26 October 2010

To cite this Article Mager, Peter P.(2003) 'A Problem if Binary Molecular Substructural Variables are Used in Simulation Experiments', Molecular Simulation, 29: 2, 159 — 166
To link to this Article: DOI: 10.1080/0892702031000065818
URL: http://dx.doi.org/10.1080/0892702031000065818

Molecular Simulation, 2003 Vol. 29 (2), pp. 159–166

A Problem if Binary Molecular Substructural Variables are Used in Simulation Experiments

PETER P. MAGER*

Research Group for Pharmacochemistry, Institute of Pharmacology and Toxicology of the University, D-04107 Leipzig, Härtelstr. 16–18, Saxony, Germany

(Received June 2001; In final form July 2001)

Two different quantitative structure–activity relationship (QSAR) techniques were compared using a series of thromboxane A2 receptor antagonists (benzimidazole and imidazol[4,5-*b*]pyridine acid derivatives). In both cases, binary molecular substructural indicator variables were used for encoding the substituents of a lead structure. The first technique is based on the conventional Free–Wilson (FW) analysis (ordinary least-squares regression). The second technique is based on an optimized backpropagation neural network where non-additive substituent contributions can be analyzed, too. It is shown that the two approaches may lead to an overfitting.

Keywords: Optimized backpropagation neural network; Free–Wilson analysis; Thromboxane A2 receptor antagonists; Benzimidazole and imidazol[4,5-*b*]pyridine acid derivatives; Overfitting

INTRODUCTION

Novel benzimidazole and imidazol[4,5-*b*]pyridine acid derivatives were synthesized and biologically tested as thromboxane A2 receptor antagonists [1]. To investigate quantitative structure–activity relationships (QSARs), extrathermodynamic and linear free-energy parameters, and quantum chemical indices, were correlated against biological activities, and it was found that the parameters were insignificant. This focused the attention to the use of the so-called Free–Wilson (FW) analysis [2] which parameterizes substituents by binary substructural indicator variables. Usually, ordinary least-squares (OLS) multiple regression is applied to estimate the substituent contributions (the so-called *de novo* constants). Theoretically, it is assumed a linear combination between biological activity and physicochemical descriptors, that is, additive substituent effects. Another method is to apply one of the various artificial neural network models (FW-type neural network analysis). Based on the chosen neural network design, additive or non-additive effects [3] can be analyzed.

A drawback of both approaches is that one of the most important rules of a well planned QSAR design can be satisfied relatively seldom, the principle of a balanced dimension [4,5]. Verbally speaking, the rule states that a certain relationship must exist between the number of dependent variables, and the degrees of freedom due to hypothesis and error. If that principle cannot be fulfilled, there is the danger of overfitting, that is, there is a good agreement between the experimentally obtained biological activities and the theoretically calculated data although that interpretation is an artifact. The study gives an example and describes the reasons of such wrong conclusion.

METHOD

Chemistry

The synthetic routes were described previously [1], together with all data that verify the structures. Te substituents of the lead structure (Fig. 1) are listed in Table I. The code of the substructural binary indicator variables X_1–X_{17} is summarized in Table II (left side).

*E-mail: magp@medizin.uni-leipzig.de

ISSN 0892-7022 print/ISSN 1029-0435 online © 2003 Taylor & Francis Ltd
DOI: 10.1080/0892702031000065818

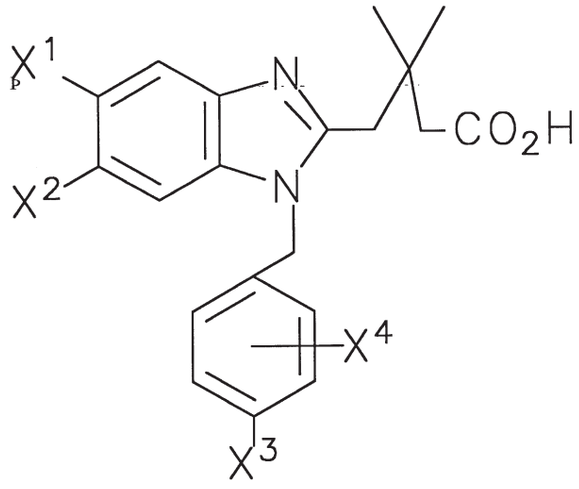

FIGURE 1 Lead structure of thromboxane A2 receptor antagonists.

Pharmacology

The affinity of compounds for washed human platelet thromboxane A2/prostaglandin H2 receptors was determined by radioligand studies. The percent of competition was determined at a dose of 10^{-7} M [1] and are listed in Table II (obtd).

Neural Network

The algorithm of the generalized backpropagation neural network was described elsewhere [6]. The sigmoidal backpropagation functions were solved by the nonlinear Levenberg–Marquardt algorithm. The learning rate momentum was equal to 0.8, the learning rate minimum and maximum were equal to 0.001 and 0.3. The squared multiple correlation coefficients (R^2) were statistically tested using the largest-root criterion [5]. Optimization was achieved by (i) estimating the global error vector prior to adjusting weights, and (ii) updating successively the weights until convergence was reached.

Regression Statistics

The underlying statistical approach is based on the MASCA model [5]. Briefly speaking, the model starts usually from a design constructed according to the rules of decision theory, collects the data to get summary statistics, tests statistical hypotheses using global and simultaneous statistical inference, applies diagnostic statistics for examining design and model validity, and uses the Bayes approach for subsequent procedures. The matrix computations are very stable by using suitable software (strong reduction of the number of cumulative rounding errors and of annulments of near-zero scores). Some modules

TABLE I Substituents of the lead structure (Fig. 1), and experimentally obtained and theoretically calculated biological activities. OLS1: traditional Free-Wilson analysis, complete model; OLS2: reduced Free-Wilson analysis; NA: non-additivity of substituent contributions using the neural network; AA: additivity of substituent contributions using the neural network

	Substituents					*Theoretical*			
Compound	X^1	X^2	X^3	X^4	*Obtd*	*OLS1*	*NA*	*AA*	*OLS2*
1	OMe	H	Cl	H	44	44	44	44	44
2	C1	H	Cl	H	74	72	74	73	74
3	Br	H	Cl	H	38	38	38	38	38
4	F	H	Cl	3-Cl	89	83	89	83	74
5	F	H	SMe	H	100	97	100	96	94
6	F	H	Br	2-F	98	90	98	89	89
7	F	H	Br	H	92	96	92	96	89
8	F	H	OMe	H	79	91	79	92	88
9	F	H	Cl	2-F	58	72	58	72	74
10	Cl	H	Cl	3-Cl	71	77	71	76	74
11	Cl	H	SMe	H	87	90	87	91	94
12	Cl	H	Cl	2-F	77	65	77	64	74
13	Cl	H	Br	2-F	74	83	74	84	89
14	Cl	H	Br	H	98	90	98	91	89
15	Cl	H	OMe	H	95	85	95	86	88
16	F	H	SOOMe	H	100	88	100	87	85
17	Cl	H	SOOMe	H	69	81	69	81	85
18	Cl	Cl	Cl	H	14	14	14	14	14
19	H	H	H	H	13	13	13	13	13
20	H	H	Me	H	52	52	52	52	52
21	H	H	F	H	66	66	66	66	66
22	H	H	OMe	H	89	87	89	87	88
23	H	H	Br	H	86	92	86	92	89
24	H	H	H	3-CF_3	49	49	49	49	49
25	H	H	NO2	H	83	83	83	83	83
26	H	H	OH	H	50	50	50	50	50
27	H	H	Br	2-F	88	85	88	85	89
Squared multiple correlation						0.92	1.00	0.92	0.90

TABLE II Positions of the substituents of the lead structure (Fig. 1), code of the indicator variables, and results of the total and reduced Free-Wilson analysis using additive substituent effects (*de novo* constants and corresponding test statistics TS)

Position	*Substituent*	*Code*	*Complete de novo*	*TS*	*Reduced de novo*	*TS*
X^1	OMe	X_1	−29.45	1.74	−29.80	2.45
	Cl	X_2	−1.41	0.16	0	
	Br	X_3	−35.45	2.09	−35.80	2.95
	F	X_4	4.74	0.53	0	
X^2	Cl	X_5	−58.04	3.94	−59.80	4.92
X^3	Cl	X_6	60.45	3.56	60.80	5.00
	SMe	X_7	78.83	4.71	80.50	5.93
	Br	X_8	78.54	5.41	76.33	6.37
	OMe	X_9	73.56	4.99	74.67	5.83
	SOOMe	X_{10}	69.83	4.17	71.50	5.26
	Me	X_{11}	39.00	2.33	39.00	2.49
	F	X_{12}	53.00	3.17	53.00	3.78
	NO_2	X_{13}	70.00	4.18	70.00	4.46
	OH	X_{14}	37.00	2.21	37.00	2.36
X^4	3-Cl	X_{15}	4.88	0.40	0	
	2-F	X_{16}	−6.64	0.82	0	
	3-CF_3	X_{17}	36.00	2.15	36.00	2.30
Critical quantile (5% level)				2.26		2.16

of the software are freely available (http://www.uni-leipzig.de/ ~ pharma/ppm2.htm).

RESULTS AND DISCUSSION

Data Matrix

Figure 1 shows the lead structure, and Table I the substituents and experimentally obtained biological activity (obtd, output variable or regressand). The code of the descriptors is listed in Table II, left side (inputs variables or regressors). The complete FW-type matrix includes ones (they stand for the presence of a substituent) and zeroes (they parameterize the absence of a substructure). The matrix, together with the biological activities, can be downloaded (http://www.uni-leipzig.de/ ~ pharma/ppm2.htm, click ABSTRACT, file 11.LST) in order to reexamine the analysis.

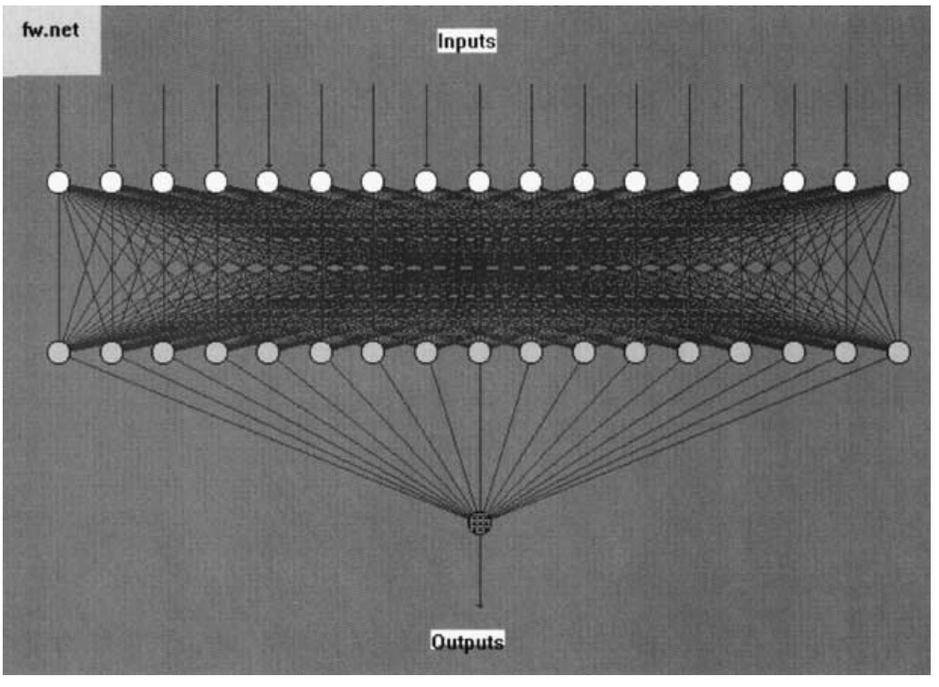

FIGURE 2 Network architecture of a Free-Wilson-type analysis of non-additive substituent effects.

Traditional Free–Wilson Analysis

The results of the conventional FW approach are summarized in Table II (right side), the theoretically calculated activities are listed in Table I (OLS). It can be seen that the wide-spread FW analysis gives a formally satisfactory result (the squared multiple correlation coefficient is $R^2 = 0.920$). The theoretically calculated data do not differ markedly from those obtained by experiment (the experimental error varies from ± 1 to ± 8). Each *de novo* constant that is equal to or larger than the critical quantile (2.26; 5% significance level) is formally significant (Table II).

Because of the well goodness-of-fit criterion, it may be believed at first glance that the analysis was correct. However, the principle of a balanced dimension was not satisfied: the degree of freedom due to hypothesis is 17, and the degree of freedom due to error is nine. Consequently, it might be expected that the good result is based on overfitting.

Neural Network Analysis

Neural networks have become powerful tools in QSAR to mimic problems of functional relationships, giving often better results than conventional statistical approaches. However, a mechanistic interpretation can hardly be made by using the neural network because the general problem is that structure–activity relationship "equations" are not developed by a black-box-like learning machine.

Non-additive Substituent Effects

The complete design consists of 27 compounds. As model parameters, the following layers of the neural network were used: one input layer with linear transfer function and 17 nodes (indicator variables); one hidden layer with sigmoidal transfer functions having each 17 nodes; and one output layer with sigmoidal transfer function and one node (biological activity). Full connections were analyzed (Fig. 2). Training information: iterations, 10,000; training error, 0.000029; learning rate, 0.006184; momentum factor, 0.800000; fast-propagation coefficient, 0; training patterns, 27; and test patterns, 0.

The standard deviation is 0.028, the bias is −0.00157, and the maximum error is 0.1. The squared multiple correlation coefficient is one (an apparently perfect fit). The resulting weights are collected in Table III, and Table II shows the theoretically calculated activities (NA).

Unfortunately, there is no doubt that the number of neural network weights of a FW-type analysis of non-additive substituent effects is very large. Consequently, the test power also is very low.

TABLE III Network weights and current adjustment deltas of the Free-Wilson-type analysis of non-additive substituent effects

Layer	*Node*	*Connection*	*Weight*	*Weight delta*
2	1	1	−0.20123	−0.000001
2	1	2	0.52662	−0.000002
2	1	3	−0.10704	−0.000001
2	1	4	−2.64644	0.000000
2	1	5	0.55805	−0.000001
2	1	6	−2.50383	−0.000004
2	1	7	−0.24736	0.000002
2	1	8	−0.40762	0.000001
2	1	9	−1.92839	0.000001
2	1	10	0.27626	0.000001
2	1	11	−0.50557	0.000000
2	1	12	−0.49821	0.000000
2	1	13	−1.01262	0.000000
2	1	14	−0.65157	0.000000
2	1	15	1.93174	0.000004
2	1	16	0.92768	0.000004
2	1	17	−0.41540	0.000000
2	2	1	−0.07838	0.000001
2	2	2	−0.39734	−0.000002
2	2	3	0.10147	0.000000
2	2	4	0.25275	0.000001
2	2	5	−0.54542	0.000000
2	2	6	0.49854	0.000002
2	2	7	0.58704	0.000000
2	2	8	0.52416	0.000000
2	2	9	0.38420	−0.000002
2	2	10	0.55800	0.000000
2	2	11	−0.01676	0.000000
2	2	12	0.30146	0.000000
2	2	13	0.65310	0.000000
2	2	14	−0.09442	0.000000
2	2	15	−0.05218	0.000000
2	2	16	−0.48209	−0.000002
2	2	17	0.09805	0.000000
2	3	1	0.09572	0.000000
2	3	2	0.23245	0.000000
2	3	3	−0.30546	0.000000
2	3	4	0.53728	0.000000
2	3	5	−0.09788	0.000000
2	3	6	0.21085	0.000000
2	3	7	0.48423	0.000000
2	3	8	0.12494	0.000000
2	3	9	0.59054	0.000000
2	3	10	0.10416	0.000000
2	3	11	−0.13372	0.000000
2	3	12	0.10667	0.000000
2	3	13	0.07759	0.000000
2	3	14	−0.02604	0.000000
2	3	15	0.07774	0.000000
2	3	16	0.29707	0.000000
2	3	17	−0.06530	0.000000
2	4	1	−0.11959	−0.000001
2	4	2	0.34948	0.000003
2	4	3	−0.05114	−0.000001
2	4	4	−0.04194	−0.000001
2	4	5	0.32562	0.000000
2	4	6	−0.45563	−0.000002
2	4	7	−0.27561	0.000000
2	4	8	−0.66816	−0.000001
2	4	9	−0.05546	0.000002
2	4	10	−0.54620	0.000000
2	4	11	−0.54401	0.000000
2	4	12	−0.60774	0.000000
2	4	13	−0.38158	0.000000
2	4	14	−0.48117	0.000000
2	4	15	−0.31186	0.000000
2	4	16	0.62710	0.000002
2	4	17	−0.07161	0.000000
2	5	1	0.10227	0.000000
2	5	2	0.18145	0.000001
2	5	3	−0.17706	0.000000

TABLE III – *continued*

Layer	*Node*	*Connection*	*Weight*	*Weight delta*
2	5	4	−0.13000	−0.000001
2	5	5	0.00824	0.000000
2	5	6	−0.31544	−0.000001
2	5	7	−0.46361	0.000000
2	5	8	0.00301	0.000000
2	5	9	0.06365	0.000001
2	5	10	−0.08361	0.000000
2	5	11	0.15424	0.000000
2	5	12	0.09088	0.000000
2	5	13	0.15409	0.000000
2	5	14	0.10880	0.000000
2	5	15	−0.13809	0.000000
2	5	16	−0.00443	0.000001
2	5	17	−0.08619	0.000000
2	6	1	0.16443	−0.000001
2	6	2	−0.29377	0.000000
2	6	3	−0.03754	−0.000001
2	6	4	−1.38232	−0.000005
2	6	5	0.80793	0.000000
2	6	6	−0.30369	0.000006
2	6	7	−0.68412	0.000002
2	6	8	−1.75901	−0.000001
2	6	9	−1.41982	0.000000
2	6	10	−0.41996	0.000001
2	6	11	−0.61824	0.000001
2	6	12	−0.73346	0.000001
2	6	13	−0.95508	0.000001
2	6	14	−0.64872	0.000001
2	6	15	0.67060	0.000001
2	6	16	−1.63607	−0.000006
2	6	17	−0.38222	0.000001
2	7	1	0.02965	0.000001
2	7	2	2.47826	0.000012
2	7	3	−0.18021	0.000000
2	7	4	−1.80698	−0.000003
2	7	5	−0.86564	0.000000
2	7	6	−1.08748	−0.000003
2	7	7	1.52998	0.000000
2	7	8	−0.04859	−0.000011
2	7	9	0.72705	0.000003
2	7	10	2.04618	−0.000002
2	7	11	0.02424	−0.000002
2	7	12	0.57540	−0.000002
2	7	13	0.53168	−0.000002
2	7	14	0.05700	−0.000002
2	7	15	2.62129	−0.000003
2	7	16	1.98164	0.000011
2	7	17	0.27392	−0.000002
2	8	1	−0.05352	0.000000
2	8	2	0.19573	0.000000
2	8	3	−0.18565	0.000000
2	8	4	0.10779	0.000000
2	8	5	0.03675	0.000000
2	8	6	−0.09311	0.000000
2	8	7	0.01429	0.000000
2	8	8	0.15197	0.000000
2	8	9	0.03118	0.000000
2	8	10	−0.30596	0.000000
2	8	11	0.04391	0.000000
2	8	12	−0.01414	0.000000
2	8	13	0.19215	0.000000
2	8	14	0.16492	0.000000
2	8	15	0.20592	0.000000
2	8	16	0.24394	0.000000
2	8	17	−0.20245	0.000000
2	9	1	0.22793	0.000001
2	9	2	0.44570	0.000000
2	9	3	0.15547	0.000001
2	9	4	0.21883	0.000002
2	9	5	−0.69451	0.000000
2	9	6	−0.00667	0.000000
2	9	7	0.56496	0.000000

TABLE III – *continued*

Layer	*Node*	*Connection*	*Weight*	*Weight delta*
2	9	8	1.02484	0.000001
2	9	9	0.60371	−0.000001
2	9	10	0.29748	0.000000
2	9	11	0.44076	0.000000
2	9	12	0.60977	0.000000
2	9	13	0.72184	0.000000
2	9	14	0.46322	0.000000
2	9	15	0.28347	0.000000
2	9	16	0.73041	0.000003
2	9	17	0.30933	0.000000
2	10	1	−0.24472	0.000000
2	10	2	−0.05214	0.000000
2	10	3	−0.24152	0.000000
2	10	4	0.13144	0.000000
2	10	5	−0.05568	0.000000
2	10	6	−0.16758	0.000000
2	10	7	0.34281	0.000000
2	10	8	0.23690	0.000000
2	10	9	0.10984	0.000000
2	10	10	−0.05764	0.000000
2	10	11	0.17491	0.000000
2	10	12	0.22097	0.000000
2	10	13	0.02333	0.000000
2	10	14	−0.27436	0.000000
2	10	15	0.21776	0.000000
2	10	16	0.07500	0.000000
2	10	17	0.05688	0.000000
2	11	1	0.25183	0.000000
2	11	2	0.14557	0.000000
2	11	3	0.14997	0.000000
2	11	4	0.12512	0.000000
2	11	5	0.10638	0.000000
2	11	6	−0.07722	0.000000
2	11	7	0.19834	0.000000
2	11	8	0.05409	0.000000
2	11	9	0.22141	0.000000
2	11	10	0.16564	0.000000
2	11	11	0.05033	0.000000
2	11	12	−0.30295	0.000000
2	11	13	−0.07336	0.000000
2	11	14	0.21768	0.000000
2	11	15	−0.16319	0.000000
2	11	16	−0.12507	0.000000
2	11	17	−0.14807	0.000000
2	12	1	0.01190	0.000000
2	12	2	−0.35217	−0.000001
2	12	3	−0.12001	0.000000
2	12	4	0.27546	0.000001
2	12	5	0.03602	0.000000
2	12	6	0.08940	0.000001
2	12	7	0.44326	0.000000
2	12	8	0.19995	0.000000
2	12	9	0.16320	−0.000001
2	12	10	0.15591	0.000000
2	12	11	0.04883	0.000000
2	12	12	0.34633	0.000000
2	12	13	0.05269	0.000000
2	12	14	−0.02897	0.000000
2	12	15	0.16096	0.000000
2	12	16	−0.32190	−0.000001
2	12	17	0.33006	0.000000
2	13	1	0.26101	0.000000
2	13	2	0.08448	0.000000
2	13	3	0.04784	0.000000
2	13	4	0.07836	0.000000
2	13	5	−0.30614	0.000000
2	13	6	−0.21157	−0.000001
2	13	7	−0.23429	0.000000
2	13	8	0.21988	0.000000
2	13	9	0.01022	0.000001
2	13	10	0.01923	0.000000
2	13	11	−0.23340	0.000000

TABLE III – *continued*

Layer	*Node*	*Connection*	*Weight*	*Weight delta*
2	13	12	−0.10610	0.000000
2	13	13	−0.04909	0.000000
2	13	14	−0.13200	0.000000
2	13	15	0.27083	0.000000
2	13	16	0.17666	0.000000
2	13	17	−0.11932	0.000000
2	14	1	−0.02161	0.000000
2	14	2	−0.05679	−0.000001
2	14	3	0.21749	0.000000
2	14	4	−0.13742	0.000000
2	14	5	−0.33144	0.000000
2	14	6	−0.07772	0.000000
2	14	7	0.27416	0.000000
2	14	8	0.46878	0.000000
2	14	9	0.36587	0.000000
2	14	10	0.04733	0.000000
2	14	11	0.16405	0.000000
2	14	12	0.19572	0.000000
2	14	13	−0.16037	0.000000
2	14	14	0.21479	0.000000
2	14	15	−0.27802	0.000000
2	14	16	−0.31828	0.000000
2	14	17	0.20563	0.000000
2	15	1	0.01164	0.000000
2	15	2	0.18378	−0.000002
2	15	3	−0.28717	0.000000
2	15	4	−0.07634	0.000001
2	15	5	−0.46819	0.000000
2	15	6	0.20715	0.000001
2	15	7	0.53740	0.000000
2	15	8	0.59949	0.000000
2	15	9	−0.06718	−0.000001
2	15	10	0.55764	0.000000
2	15	11	0.29647	0.000000
2	15	12	0.17838	0.000000
2	15	13	0.24989	0.000000
2	15	14	0.07663	0.000000
2	15	15	0.19957	0.000000
2	15	16	−0.14329	0.000000
2	15	17	0.33242	0.000000
2	16	1	0.05117	−0.000002
2	16	2	1.15666	0.000012
2	16	3	0.17181	−0.000002
2	16	4	0.12595	0.000001
2	16	5	0.75070	−0.000001
2	16	6	−0.49919	0.000000
2	16	7	−1.14500	0.000001
2	16	8	−1.76874	−0.000010
2	16	9	−0.31291	0.000003
2	16	10	−0.91169	0.000000
2	16	11	−0.85647	0.000000
2	16	12	−0.40290	0.000000
2	16	13	−0.85233	0.000000
2	16	14	−0.62656	0.000000
2	16	15	−0.22282	−0.000001
2	16	16	1.28846	0.000010
2	16	17	−0.82592	0.000000
2	17	1	0.11250	0.000001
2	17	2	0.08401	−0.000002
2	17	3	−0.30938	0.000000
2	17	4	−0.02705	0.000001
2	17	5	−0.36169	0.000000
2	17	6	0.27283	0.000002
2	17	7	0.64158	0.000000
2	17	8	0.46710	0.000000
2	17	9	0.40651	−0.000001
2	17	10	0.65289	0.000000
2	17	11	0.24960	0.000000
2	17	12	0.58326	0.000000
2	17	13	0.64386	0.000000
2	17	14	0.29528	0.000000
2	17	15	0.48835	0.000000
2	17	16	−0.28934	−0.000001
2	17	17	0.21914	0.000000
3	1	1	−4.47685	−0.000004
3	1	2	1.25313	0.000002
3	1	3	0.23410	0.000000
3	1	4	−1.79806	−0.000001
3	1	5	−0.57199	0.000000
3	1	6	−3.64026	−0.000002
3	1	7	4.41536	0.000005
3	1	8	−0.37465	0.000000
3	1	9	1.59001	0.000000
3	1	10	0.04857	0.000000
3	1	11	−0.25438	0.000001
3	1	12	0.50844	0.000001
3	1	13	−0.41634	0.000000
3	1	14	0.33652	0.000000
3	1	15	0.86695	0.000000
3	1	16	−3.46559	−0.000005
3	1	17	1.27811	0.000000

Additive Substituent Effects

As model parameters, the following layers of the neural network were used: one input layer with linear transfer function and 17 nodes (indicator variables); one hidden layer with sigmoidal transfer functions having one node; and one output layer with sigmoidal transfer function and one node (biological activity). Full connections were analyzed (Fig. 3). Training information: iterations, 10,000; training error, 0.055; learning rate, 0.3; momentum factor, 0.800000; fast-propagation coefficient, 0; training patterns, 27; and test patterns, 0.

The standard deviation is 6.86, the bias is −0.0199, and the maximum error is 13.78. The squared multiple correlation coefficient is 0.919. The weights are collected in Table III, and Table II shows the theoretically calculated activities (AA). The results are comparable with those found by the conventional FW analysis (Table IV).

Reduction in Dimensionality of Molecular Substructural Variables

Selecting those variables for which statistics are relevant and deleting the least important variables is a straightforward way to improve the test power (principle of parsimony [5]).

Reduced Free–Wilson Analysis

Using the so-called reduced FW analysis, the insignificant binary substructural variables are omitted, and the procedure is repeated once more [4,7]. Returning to Table II, it can be seen that the test statistics corresponding to the variables X_2, X_4, X_{15}, and X_{16} tend statistically to zero. The omission of these variables leads to a squared correlation coefficient of $R^2 = 0.898$. The regression coefficients

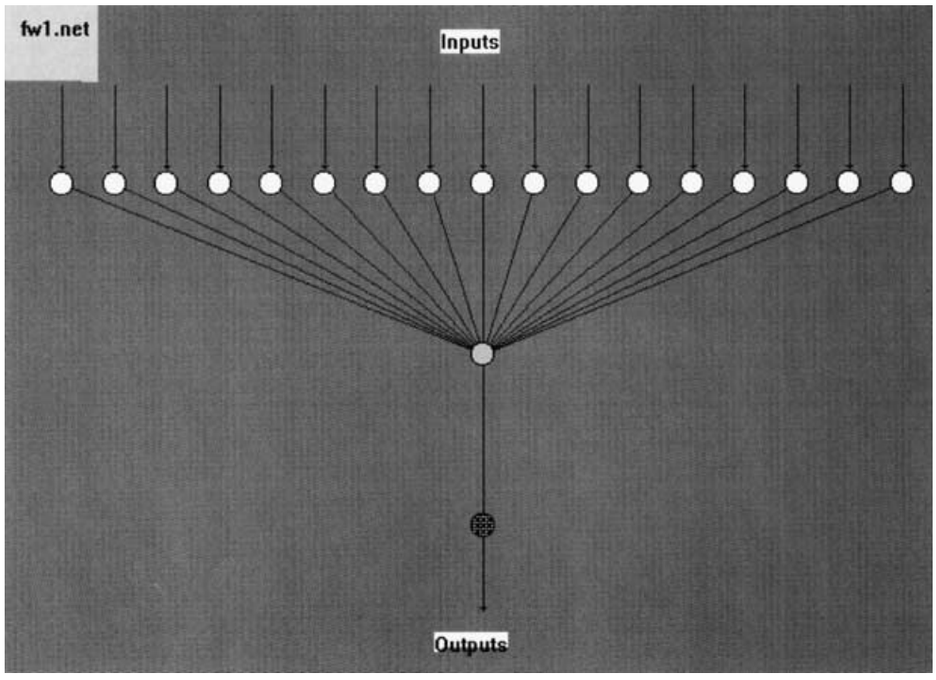

FIGURE 3 Network architecture of a Free-Wilson-type analysis of additive substituent effects.

and the fitted data are listed in Tables I and II. The degree of freedom due to hypothesis is 14, and the degree of freedom due to error is 12. The principle of a balanced dimension is not satisfied although the "degree of the principal dilemma" is lowered.

Metameters

Another approach is based on the estimation of metameters of binary molecular substructural variables by using suitable mathematical transformations. The resulting metameters should be arranged in decreasing order, according to their capacity to express the information content, and are used then in subsequent analyses instead of the original descriptors [5]. Such subsequent analyses are certain regression approaches (principal-component, non-least-squares, latent-root, ridge regression) and the various types of neural networks (mainly back-propagation and generalized-regression neural network analysis). The common underlying background is that the binary molecular substructural variables are multicollinear, and their metameters have a decreasing rank order related to the biological activity.

TABLE IV Network weights and current adjustment deltas of the Free-Wilson-type analysis of additive substituent effects

Layer	*Node*	*Connection*	*Weight*	*Weight delta*
2	1	1	0.78895	−0.000002
2	1	2	0.03953	0.000000
2	1	3	0.96457	−0.000001
2	1	4	−0.14677	0.000001
2	1	5	2.02767	0.000024
2	1	6	−2.11353	−0.000027
2	1	7	−2.62667	−0.000021
2	1	8	−2.62230	−0.000022
2	1	9	−2.46971	−0.000023
2	1	10	−2.31548	−0.000027
2	1	11	−1.53864	−0.000030
2	1	12	−1.89083	−0.000029
2	1	13	−2.34269	−0.000026
2	1	14	−1.48659	−0.000030
2	1	15	−0.08483	−0.000001
2	1	16	0.22266	−0.000003
2	1	17	−1.46024	−0.000030
3	1	1	−7.66776	−0.000091

TABLE V Eigenvalues (EV), percentages of the eigenvalues related to the information content of the binary molecular substructural variables (%), cumulative percentages (cum), test statistics (TS), and critical Chi squared quantiles (5% significance level)

Order	*EV*	*%*	*Cum (%)*	*TS*	χ^2 *quantile*
1	2.026	11.92	11.92	132.96	181.77
2	1.722	10.13	22.05	9.22	27.59
3	1.487	8.75	30.80	6.62	26.30
4	1.400	8.23	39.03	4.67	25.00
5	1.110	6.53	45.56	4.50	23.68
6	1.080	6.35	51.91	1.76	22.36
7	1.059	6.23	58.14	1.85	21.03
8	1.038	6.11	64.25	2.08	19.68
9	1.038	6.11	70.36	2.43	18.31
10	1.038	6.11	76.47	3.29	16.92
11	1.038	6.11	82.58	4.70	15.51
12	1.038	6.11	88.69	7.31	14.07
13	0.897	5.28	93.97	13.08	12.59
14	0.652	3.83	97.80	20.06	11.07
15	0.203	1.20	99.00	31.85	9.49
16	0.140	0.81	99.81	6.86	7.81
17	0.033	0.19	100.00	12.70	5.99

TABLE VI Principal-component regression analysis applied to diagnostic statistics

Order	*Coefficient*	*TS*
1	−11.83	5.10
2	−6.63	2.86
3	−3.26	1.40
4	−10.56	4.55
5	−2.26	0.98
6	−1.77	0.76
7	1.82	0.78
8	4.22	1.88
9	2.98	1.28
10	−2.20	0.95
11	−1.31	0.56
12	−0.46	0.20
13	−8.13	3.50
14	−1.74	0.75
15	−2.45	1.06
16	−4.43	1.91
17	−10.57	4.55
Critical quantile (5%)		2.26

To examine this assumption, the ordered eigenvalues of the correlation matrix of the binary molecular substructural variables are calculated and statistically tested [5]. Each test statistic that is equal to or larger than the critical quantile is significantly different from zero. Only those principal components (PCs) that correspond to the significant eigenvalues are relevant for subsequent analyses.

The results are listed in Table V. It can be seen that PC13–PC15, and PC17 are formally meaningful. However, they summarize only $(5.28 + 3.83 + 1.20 + 0.19)\% = 10.5\%$ of the information content of the binary molecular substructural variables. This implies that each subsequent analysis using metameters is quite senseless with respect to the given example. The second evidence for this conclusion is based on principal-component regression (Table VI). The term "order" is related to the principal components that are ordered according to their corresponding eigenvalues. The results indicate clearly that there is no monotonic reduced-rank property. For example, the least important component PC17 that is associated with the near-zero eigenvalue EV17 = 0.33 (Table V), is formally significant although it contributes very rarely (0.19%) to the information content of the binary descriptors.

In summary, there is no way to use statistical alternatives that are based on metameters. This conclusion is also valid for partial-least squares regression.

CONCLUSIONS

Theoretical values of biological activity can be calculated and predicted by using binary molecular substructural variables. Two approaches can be applied, the traditional FW analysis (complete and reduced model) that is based on ordinary least-squares regression, and neural network techniques. Theoretically speaking, the latter ones consider also non-additive substituent contributions but with the limitation that the number of weights are growing exponentially.

Unfortunately, there is no way to reduce the dimensionality by a general algorithm that works well for all designs in daily practice. As a consequence, the danger of overfitting must be taken into account.

References

[1] Nicolai, E., Goyard, J., Benchetrit, T., Teulon, J.-M., Caussade, F., Virone, A., Delchambre, C. and Cloarec, A. (1993) "Synthesis and structure–activity relationships of novel benzimidazole and imidazol[4,5-*b*]pyridine acid derivatives as thromboxane A_2 receptor antagonists", *J. Med. Chem.* **36**, 1175–1187.
[2] Free, S.M. and Wilson, J.W. (1964) "A mathematical contribution to structure–activity studies", *J. Med. Chem.* **7**, 395–399.
[3] Schaper, K.-J. (1999) "Free–Wilson-type analysis of non-additive substituent effects on THPB dopamine receptor affinity using artificial neural networks", *Quant. Struct.–Act. Relat.* **18**, 354–360.
[4] Mager, P.P. (1988) Multivariate Chemometrics in QSAR: a Dialogue (Wiley, New York).
[5] Mager, P.P. (1991) Design Statistics in Pharmacochemistry (Wiley, New York), pp. 19–337.
[6] Zupan, J. and Gasteiger, J. (1993) Neural Networks for Chemists (VCH, Weinheim).
[7] Mager, P.P. (1984) Multidimensional Pharmacochemistry (Academic Press, New York).