Available online at www.sciencedirect.com
SCIENCE DIRECT
Bioorganic & Medicinal Chemistry 14 (2006) 5178–5195
Bioorganic & Medicinal Chemistry

2D QSAR of PPARγ agonist binding and transactivation

Christoph Rücker,[a,*] Marco Scarsi[a] and Markus Meringer[b]

[a]*Biocenter, University of Basel, Klingelbergstrasse 50-70, CH-4056 Basel, Switzerland*
[b]*Kiadis B.V., Zernikepark 6-8, NL-9747 AN Groningen, The Netherlands*

Received 19 December 2005; revised 9 February 2006; accepted 4 April 2006
Available online 2 May 2006

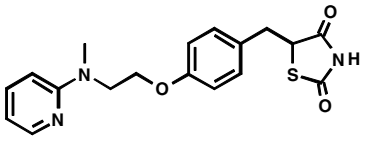
Abstract—Multilinear QSAR models are developed for the largest and most diverse set of PPARγ agonists treated hitherto. Binding of these small molecules to the human nuclear receptor PPARγ is described by models that are built on simple 2D molecular descriptors and nevertheless are of good quality and predictive power (e.g., 144 compounds, 10 descriptors, $r^2 = 0.79$, $r^2_{cv} = 0.76$). The models presented are thoroughly validated by crossvalidation, randomization experiments, bootstrapping, and training set/test set partitioning. They may therefore be helpful in the design of new antidiabetic drug candidates. For gene transactivation, the functional activity of the agonists, a corresponding model for a similarly diverse compound set is of somewhat lower statistical quality.

1. Introduction

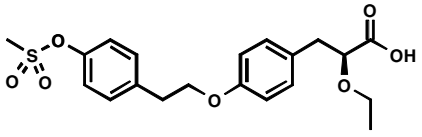
The peroxisome proliferator-activated nuclear receptors (PPARs) are a class of transcription factor proteins that play an important role in the regulation of lipid and glucose metabolism in vertebrates. They are linked to severe human diseases such as cardiovascular disease and type 2 diabetes.[1–5] The following simplified mechanism of action has been proposed: When binding a small molecule called an agonist, a PPAR is activated by undergoing a conformational change,[6] binds (in the form of a heterodimer with an RXR receptor) to a specific binding element in the DNA (response element located in a gene promotor sequence), thereby enhancing the transcription of specific genes that code for metabolic enzymes.

Three subclasses of PPARs are known, called PPARα, PPARγ, and PPARδ, that are coded by different genes and expressed at different levels in various tissues and are associated with various functions. Of these, PPARγ is mostly expressed in adipose tissue, where it is essential in adipocyte differentiation and controls fatty acid levels, increasing triglyceride synthesis and storage within adipocytes. Activation of PPARγ improves the condition of insulin resistance, and therefore PPARγ became a primary target in treatment of type 2 diabetes. Indeed, there is strong evidence that PPARγ regulates glucose homeostasis.[1–5]

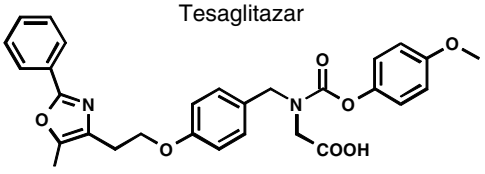
For PPARγ, several unsaturated fatty acids, in particular prostaglandins[7] and nitrolinoleic acids,[8] have been proposed as natural ligands. A few synthetic PPARγ agonists are approved drugs (e.g., rosiglitazone, a thiazolidinedione (TZD)) or under development as antidiabetics (e.g., tesaglitazar, an O-analogous tyrosine derivative, or muraglitazar).



Rosiglitazone



Tesaglitazar



Muraglitazar

Keywords: PPARγ agonists; 2D QSAR; Type 2 diabetes.
* Corresponding author. Tel.: +41 61 267 1469; fax: +41 61 267 1584; e-mail: christoph.ruecker@unibas.ch

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.04.005

The binding of a PPAR agonist to its receptor is measured in vitro and expressed numerically as the corresponding dissociation constant K_i (or its negative decadic logarithm pK_i). More interesting from a pharmaceutical point of view is a measure of the agonist's function, gene activation. This activity can be measured in a cell-based assay in vitro and is expressed as EC_{50} (or its negative decadic logarithm pEC_{50}), the concentration that causes half-maximal activation. Of course, the pharmaceutical effect in vivo, such as lowering of the lipid or glucose level in blood, is an even more valuable quantity to know. Unfortunately, both measurement and understanding are progressively more difficult for the three effects in the order mentioned, since the physicochemical phenomena dominant in the first case are more and more obscured by complex and poorly understood biological phenomena in the second and third cases.

Numerical values for the activities of many PPARγ agonists have been published, resulting from research in several pharmaceutical companies. The objective of the present work was to transform this wealth of data into information, with the aim to predict receptor binding and transactivation behavior of potential PPARγ agonists from their chemical structure alone.

Recently, several QSAR studies of agonist binding to human PPARγ were published.[9–14] Most of these used a 3D-QSAR field method such as CoMFA and CoMSIA, and the results therefore depend on alignment of the agonists to the conformations of rosiglitazone or farglitazar as found (X-ray) in their complexes with the PPARγ ligand binding domain.[15,16] Accordingly, each of these studies included rather limited numbers and types of PPARγ agonists, and use of the resulting models is restricted by the necessity of alignment and the required knowledge of the derived fields. A method not depending on molecular fields, but on assumed 3D conformations and quantum chemical descriptors, was used for a small series of PPARγ ligands.[14]

The so-called 2D-QSAR methods, on the other hand, are attractive for not depending on alignment or assumptions on conformations, therefore they can easily be applied to large compound sets, both in model building and in model application to new compounds. In such methods one has the choice among a wide variety of molecular descriptors independent of 3D conformation, for example, graph theoretical descriptors, simple molecular properties such as the molecular weight, and easily calculated physicochemical properties such as clog*P* or atomic partial charges. In any case, resulting models are described in the form of equations and therefore are easily portable.

As to the biological effect of PPARγ agonists, gene activation, CoMFA and CoMSIA methods were found unable to model experimental results,[11] while there is a 3D conformation-dependent study on one rather limited set of agonists,[14] and a 2D-QSAR study on another.[17]

The present study aimed at developing simple and easily portable 2D-QSAR models of broad applicability for both human PPARγ binding of and PPARγ-mediated gene activation by small molecules. We planned to include all known series of PPARγ agonists with appropriate experimental data available, and for the reasons given above 2D rather than 3D methods had to be used for such a diverse compound set.

2. Data set and methods

2.1. Compound set and experimental data

Various protocols are in use for measuring both receptor binding of and transactivation by PPARγ agonists,[18] and numbers obtained for the same compound under various protocols differ considerably.[7] For example, binding of rosiglitazone to human PPARγ was described by K_i values ranging from 47[19] to 230 nM.[20] For PPARγ-mediated human gene activation by rosiglitazone EC_{50} values between 18[21] and 730 nM[22] were published.

To test the sensitivity of numerical values for slight changes in the experimental protocol, we compared the results for binding of a series of agonists to human PPARγ. While data obtained in a scintillation proximity assay (Glaxo)[23] were published in journal articles,[24–26] the corresponding patent contains data obtained in a classical solution scintillation assay for an overlapping set of agonists.[27] The linear correlation between both series of numerical values (for the 61 compounds with enough data available) is no higher than $r^2 = 0.48$. Thus, either the two methods do not measure the same phenomenon, or, if they do, at least one does it in a rather unreliable manner. Therefore, we decided to include in our study human PPARγ data obtained under one and the same protocol only, that is, for receptor binding the scintillation proximity assay.[23] Data resulting from this method were successfully treated by independent research groups and thus seem to be more trustworthy.[9–12]

For gene activation we decided to use the data obtained from a transient cotransfection assay likewise developed at Glaxo.[28] Interestingly, again for a series of 65 agonists data are available both from the patent[27] and from the ensuing journal publications,[24–26] all obtained using this same assay. Nevertheless, there is a remarkable scattering in the data, as shown in Figure 1.

In this situation, we decided to use the data given in the journal publications, assuming that inconsistencies may at least in part be due to erroneous numbers in the patent being corrected in the later publications.

By the above, the synthetic agonists included in our study are essentially limited to those from Glaxo research. Specifically, along with tyrosine-based compounds and a few thiazolidinediones,[24–26] indole derivatives,[19] oxadiazole-substituted α-isopropoxyphenylpropanoic acids,[29] α,α-dimethyl-aminopropylphenoxyacetic acids,[30] tyrosine derivatives bearing small

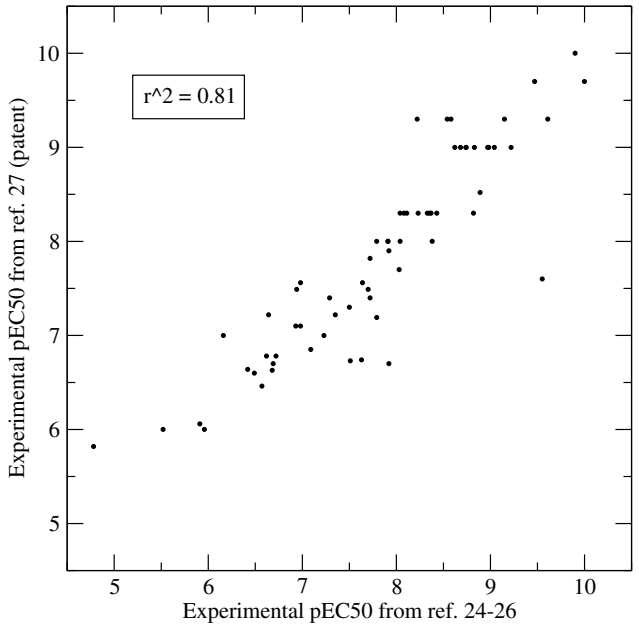


Figure 1. Experimental pEC_{50} values for gene transactivation by some PPARγ agonists, measured in a transient cotransfection assay and taken from Ref. 27 and from Refs. 24–26, respectively.

N-substituents,[21,31] fatty acids,[32] and thiazolidinedione-fatty acid hybrids[33] are included.

Published numerical values for the binding and transactivation behavior of chiral PPARγ agonists were obtained for pure *S* enantiomers in some cases, for racemates in others, though the activity is known to almost completely reside in the *S* enantiomers.[34,35] In order to render racemate data comparable to pure *S* enantiomer data, we added 0.3 (=$\log_{10}$ 2) to all pK_i and pEC_{50} values of racemates taken from the literature, which is equivalent to assuming that the concentration of a racemate required to obtain a certain effect is twice the concentration of the corresponding active enantiomer, and to ignoring any racemization that might occur to a pure enantiomer. The stereocorrection (0.3 log units) is small compared to the intrinsic scatter of the data, as seen in Figure 1 (obtained from stereocorrected data), where the worst differences between corresponding *x* and *y* value are 1–2 log units.

The final agonist set consists of 176 compounds, 144 of which have measured pK_i values for binding to PPARγ, 150 of which have measured pEC_{50} values for transactivation, and 118 have both (Table 1). The pK_i range in the data set is from 4.68 to 9.16 (mean 7.52, standard deviation 1.24), the pEC_{50} values vary between 4.94 and 10.00 (mean 7.49, standard deviation 1.18, all logarithmic values derived from concentrations given in mol/L, stereocorrected values).

2.2. Molecular descriptors

We used a pool of molecular descriptors consisting of those supplied by the program MOE[36] (atom and bond counts, connectivity indices, partial charge descriptors, pharmacophore feature descriptors, calculated physical property descriptors, etc.) plus the MACCS keys, as implemented in an additional module for use within MOE.[37] This combination recently proved useful in a drug classification problem,[39] and detailed information on the descriptors is given in that work. Initially we also tried the descriptors from MOLGEN-QSPR,[40,41] but these yielded inferior results. For simplicity we did not use any quantum chemical descriptors. Because both the PPARγ agonists and the receptor itself are flexible, all descriptors depending on molecular conformation were excluded. Descriptor values were calculated for the compounds in the protonation state assumed to be predominant at pH 7, according to known pK_a values for important acidic and basic substructures. Descriptors exhibiting constant or nearly constant values in the respective compound sample were discarded. Likewise we removed one out of every pair of descriptors found to be collinear or anticollinear.

2.3. Descriptor selection

For selecting a good descriptor combination out of a large pool of descriptors we used both a genetic algorithm supplied as an additional module to MOE, and the step-up procedure provided by MOLGEN-QSPR.[42,43]

3. Results

3.1. Binding

We built two models for receptor binding of PPARγ ligands, one for the complete set of compounds, the other for a compound set obtained by training set/test set partition.

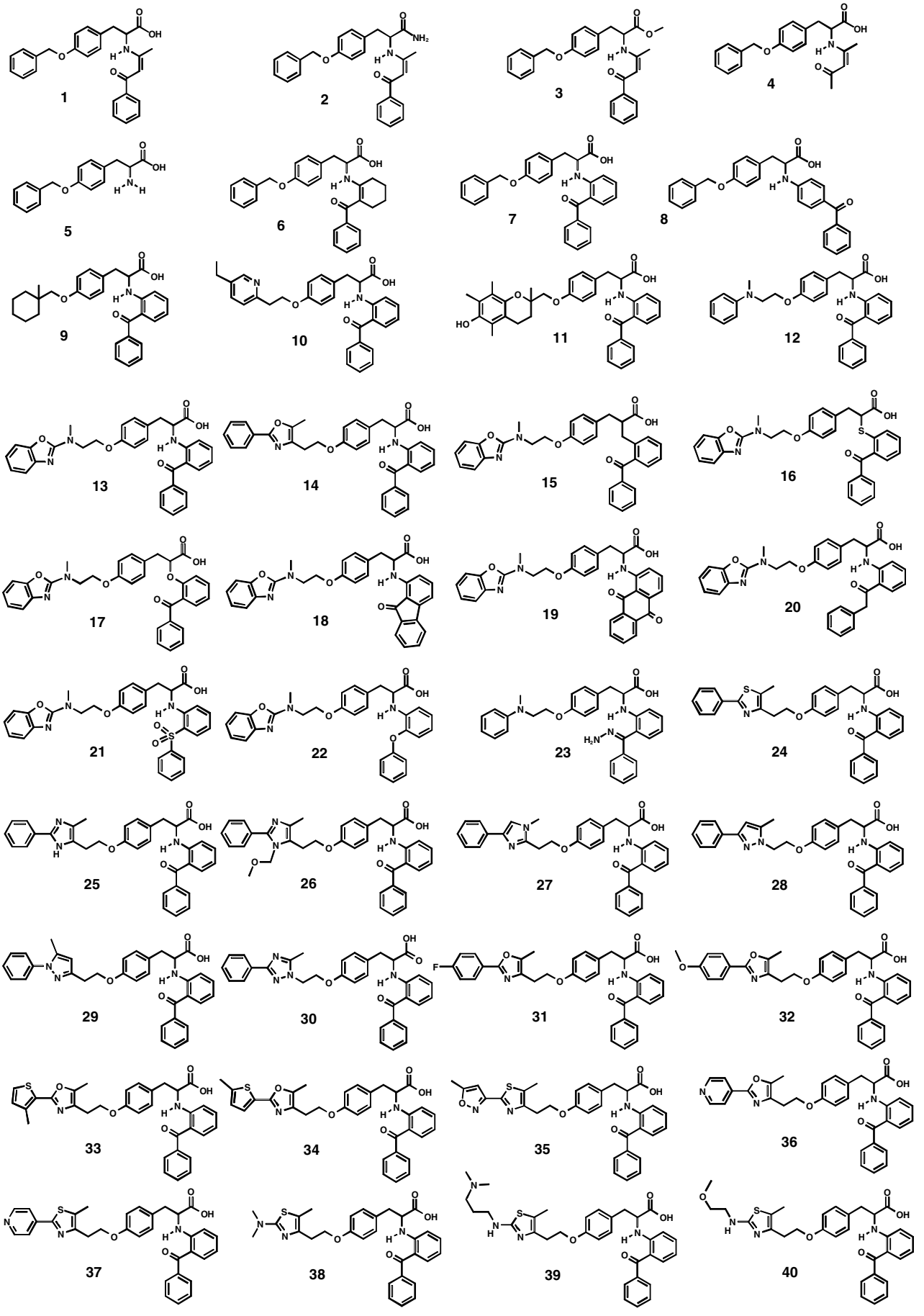
3.1.1. Complete compound set. For the 144 compounds with PPARγ binding data available, 230 descriptors remained after removal of collinear ones. For multilinear regression (MLR), the step-up procedure selected from this pool the following best 10-descriptor model m1. In the text we characterize a MLR model by the descriptors involved and by some statistics. For full models, see Tables 2 and 3.

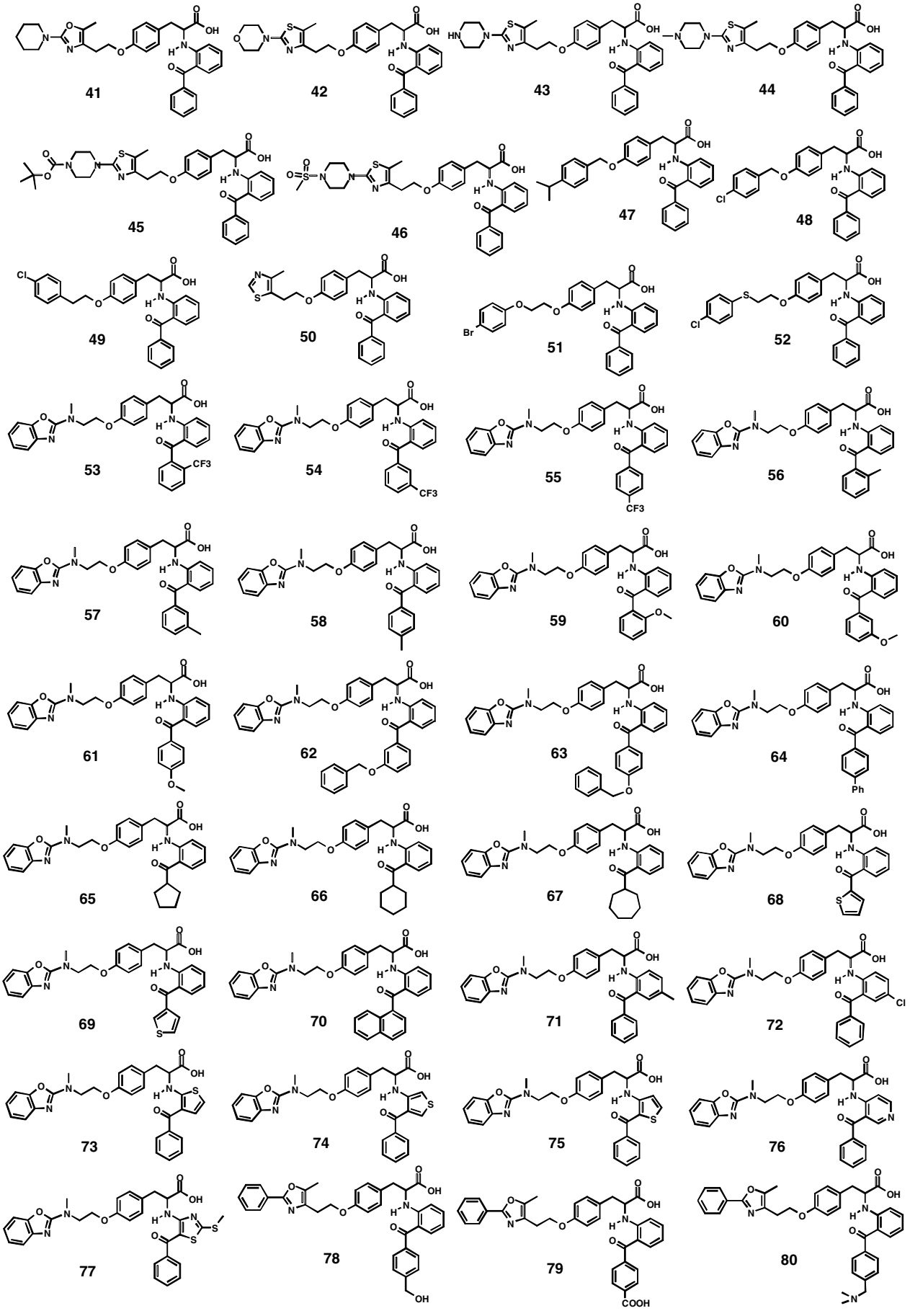
pK_i : VAdjEq PEOE_RPC- bpol sMR_VSA0 sMR_VSA3 sMR_VSA6 MACCS49 MACCS97 MACCS116 MACCS152

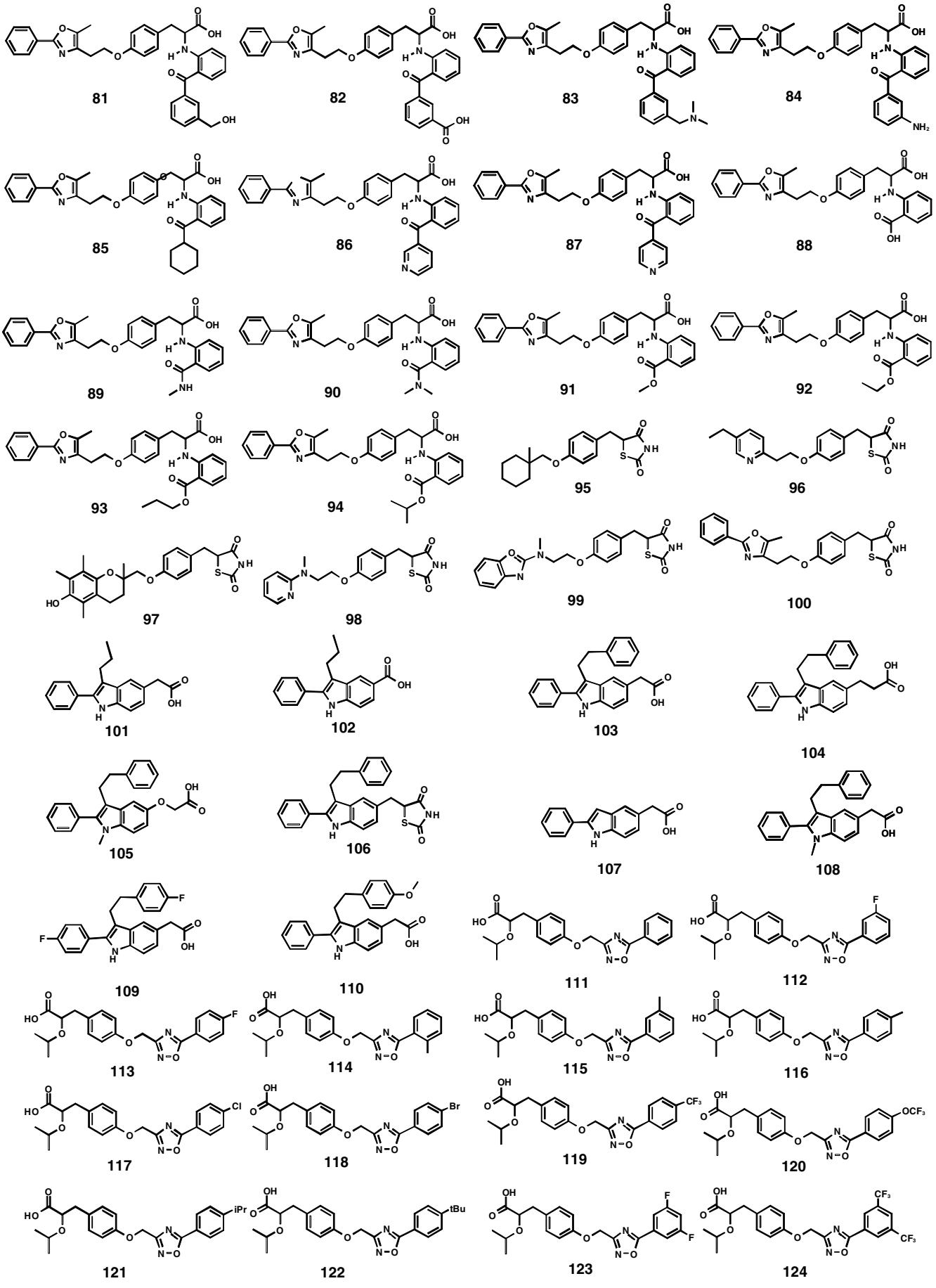
$$n = 144,\ r^2 = 0.7938,\ s = 0.5822,\ F = 51.20,\ r^2_{cv} = 0.7627,\ s_{cv} = 0.6246. \quad \text{(m1)}$$

Here subscript cv denotes quantities obtained by leave-one-out (LOO) crossvalidation. For m1, the crossvalidated statistics are close to those of fitting, an indication of model consistency. A calculated versus observed plot is given in Figure 2, showing both fitted (closed symbols) and LOO-crossvalidated values (open symbols).

Model m1 was subjected to additional validation procedures. Table 4 shows the result of a four times leave-one-


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40


41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80


81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124

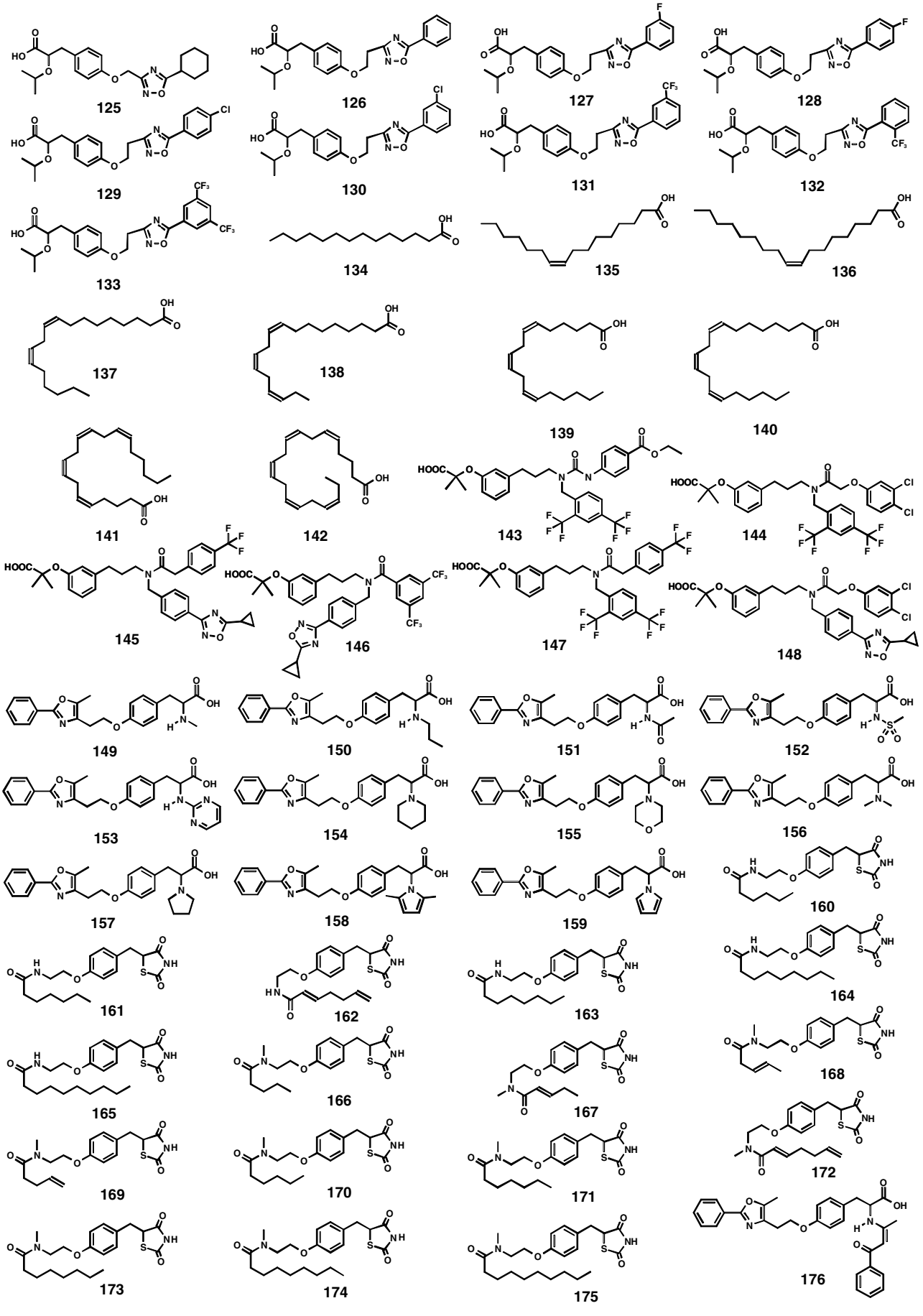

125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176

Table 1. Experimental and calculated pK_i and pEC_{50} values of compounds **1–176**

Compound	pK_i exp.	pK_i calcd (m1)	pK_i calcd (m2)	pEC50 exp.	pEC50 calcd (m3)	pEC50 calcd (m4)
1	7.93	6.78	6.60	6.64	5.21	6.99
2	5.88	6.44	6.56	6.31	5.80	5.09
3	6.12	6.03	6.63	6.16	7.01	5.37
4	5.71	6.33	6.17			
5				5.60	4.25	
6	6.10	6.61	6.75	4.99	5.71	5.02
7	7.09	7.08	6.97	5.08	6.10	6.02
8	6.20	6.33	6.13			
9	7.29	8.05	7.24	6.21	7.07	6.23
10	8.19	7.96	8.25	7.30	7.18	7.09
11	8.28	8.02	7.85	8.04	7.24	7.56
12	8.85	8.18	8.02	8.04	7.03	7.70
13	8.83	8.46	8.30	8.58	7.59	8.09
14	8.94	8.66	8.62	9.47	8.58	8.97
15	7.85	7.79	7.57	7.08	7.46	7.16
16	7.78	7.79	7.65	6.16	6.20	7.07
17	7.79	8.21	8.36	6.33	7.01	6.99
18	7.37	8.55	8.32	6.61	7.64	6.62
19	8.59	8.24	8.14	8.60	7.55	7.75
20	8.78	8.42	8.59	7.84	7.67	8.01
21	7.21	7.91	7.77	6.04	6.63	6.12
22	8.73	8.29	8.26	7.27	7.15	8.01
23	6.79	7.54	6.92	6.07	6.22	5.46
24	8.96	8.74	8.79	10.00	8.16	8.50
25	8.59	8.61	8.50	6.42	7.77	8.01
26	8.70	8.78	9.16	7.09	7.72	7.96
27	9.16	8.72	8.97	8.03	7.80	8.62
28	8.75	8.84	8.57	8.97	8.46	8.24
29	8.90	8.24	8.42	8.43	8.24	8.43
30	8.32	8.79	8.37	7.91	8.32	7.72
31	8.80	8.77	8.77	9.90	8.79	8.78
32	8.96	8.72	8.99	9.22	8.32	8.87
33	8.72	8.28	8.73	8.98	8.95	8.74
34	9.07	8.39	8.73	9.61	8.95	9.10
35	9.05	9.30	9.38	8.82	9.38	9.14
36	8.85	8.62	8.43	8.74	8.44	8.85
37	9.06	8.71	8.59	8.68	8.02	8.58
38	7.56	8.18	8.30	5.91	6.29	6.33
39	7.91	7.81	7.53	5.52	5.16	6.52
40	8.59	8.93	9.01	7.51	7.10	7.21
41	6.77	8.68	8.88	7.29	8.83	6.67
42	9.11	9.07	9.54	8.74	8.02	9.10
43	8.36	8.14	8.02	6.93	7.73	7.72
44	8.66	8.84	8.38	8.89	7.57	8.04
45	9.01	8.68	8.38	8.62	9.74	8.08
46	8.58	8.55	8.27	7.63	7.25	7.78
47	6.98	6.58	7.17	6.62	6.71	5.85
48	7.41	7.34	7.22	5.96	6.47	6.37
49	8.03	7.68	7.79	6.98	7.05	7.00
50	7.73	7.61	7.46	6.49	6.92	6.65
51	7.71	8.26	8.50	6.57	6.94	6.55
52	7.94	8.14	8.24	6.72	6.51	6.87
53	8.87	8.48	8.71	7.70	7.77	7.90
54	8.88	8.49	8.71	8.23	7.77	7.93
55	8.59	8.49	8.71	7.92	7.77	7.62
56	8.95	8.23	8.34	8.33	7.73	8.20
57	8.87	8.23	8.34	8.08	7.73	8.12
58	8.85	8.24	8.34	8.37	7.73	8.10
59	9.06	8.53	8.66	7.72	7.32	8.19
60	8.94	8.54	8.66	7.23	7.32	8.09
61	8.55	8.54	8.66	7.79	7.32	7.68
62	7.57	8.56	8.76	6.43	7.63	6.49
63	7.48	8.56	8.76	6.42	7.63	6.40
64	7.61	8.75	8.79	7.92	8.38	7.40
65	8.49	8.21	8.14	8.04	7.69	7.78

(*continued on next page*)

Table 1 (*continued*)

Compound	pK_i exp.	pK_i calcd (m1)	pK_i calcd (m2)	pEC50 exp.	pEC50 calcd (m3)	pEC50 calcd (m4)
66	8.39	8.17	8.21	8.54	7.87	7.65
67	7.70	8.13	8.29	8.41	8.05	6.90
68	8.86	8.42	8.37	7.64	7.83	8.13
69	8.93	8.30	8.37	7.50	7.83	8.20
70	8.79	8.70	8.67	7.37	8.15	7.98
71	8.17	8.23	8.34	6.94	7.73	7.39
72	8.36	8.67	8.55	7.35	7.93	7.61
73	7.93	8.30	8.37	7.72	7.83	7.15
74	8.89	8.31	8.37	8.38	7.83	8.16
75	8.31	8.43	8.37	8.22	7.83	7.57
76	7.11	8.42	8.10	5.33	6.27	6.25
77	7.67	8.30	8.71	5.69	5.66	6.80
78	8.68	8.44	7.85	8.06	7.83	8.52
79	6.49	6.70	6.83	5.42	6.83	5.87
80	8.11	6.90	7.27	6.69	7.26	7.96
81	8.77	8.44	7.85	7.91	7.83	8.61
82	6.39	6.70	6.83	6.98	6.83	5.77
83	6.24	6.90	7.27	6.68	7.26	6.00
84	8.79	8.69	8.42	8.35	7.65	8.70
85	8.79	8.37	8.54	9.55	8.86	8.83
86	9.03	8.61	8.43	8.83	8.44	9.03
87	8.74	8.62	8.43	9.04	8.44	8.74
88				5.61	6.44	
89	8.11	7.85	7.80	7.79	6.76	8.07
90	6.90	7.75	7.57	6.55	7.25	6.81
91	8.43	8.37	8.23	9.15	9.08	8.43
92	8.52	8.71	8.28	9.04	9.37	8.51
93	8.62	8.67	8.65	9.52	9.53	8.58
94	9.01	8.44	8.23	9.24	9.40	9.00
95	5.81	6.60	6.45	4.94	5.13	5.14
96	6.21	6.77	7.46	6.53	6.92	5.47
97	6.82	7.32	7.06	6.57	7.16	6.49
98	7.63	6.94	7.24	7.35	6.77	6.87
99	7.87	7.42	7.51	8.25	7.71	7.54
100	8.67	7.75	7.83	8.80	8.59	9.14
101	5.41	5.33	5.53	5.15	5.78	5.49
102	5.43	5.15	5.34			
103	6.83	6.07	5.79	6.51	6.23	6.84
104	5.14	6.14	5.86	5.52	6.44	5.05
105	6.67	6.71	7.25			
106	5.72	6.50	6.17			
107	5.35	4.90	5.40			
108	6.26	6.09	6.31	5.25	6.42	6.25
109	6.31	6.47	6.09	6.47	6.67	6.23
110	7.32	6.28	6.74	7.36	6.00	7.24
111				7.40	7.29	
112				7.30	7.50	
113				7.00	7.50	
114				7.30	7.42	
115				7.60	7.42	
116				8.10	7.42	
117				8.00	7.62	
118				7.79	7.68	
119				7.90	7.61	
120				7.50	7.50	
121				8.70	7.78	
122				8.82	7.83	
123				7.30	7.71	
124				7.79	7.85	
125				6.94	7.48	
126				8.19	8.17	
127				9.00	8.38	
128				7.70	8.38	
129				7.90	8.50	
130				8.70	8.50	
131				8.52	8.46	

Table 1 (*continued*)

Compound	pK_i exp.	pK_i calcd (m1)	pK_i calcd (m2)	pEC50 exp.	pEC50 calcd (m3)	pEC50 calcd (m4)
132				7.10	8.46	
133				8.70	8.69	
134	4.68	4.59	5.29			
135	5.19	5.12	5.43			
136	5.39	5.41	5.57			
137	5.21	5.55	5.56			
138	5.22	5.23	5.55			
139	5.66	5.68	5.55			
140	5.62	5.89	5.69			
141	5.80	6.03	5.68			
142	5.80	5.72	5.67			
143				9.30	8.27	
144				7.52	8.23	
145				7.09	6.63	
146				6.49	6.82	
147				8.40	7.95	
148				6.77	6.91	
149	5.68	6.35	6.04	5.30	5.93	5.94
150	7.28	6.33	6.11	6.62	6.79	7.59
151	5.59	6.77	6.19	5.46	7.21	5.69
152	6.21	5.79	5.79	5.59	5.78	6.30
153	6.55	7.61	7.08	6.23	6.12	6.65
154	6.32	6.55	6.41	6.86	7.35	6.59
155	6.80	6.80	6.90	7.62	7.17	7.55
156	6.07	6.26	5.67	6.39	5.51	6.36
157	6.44	6.53	6.34	6.41	7.19	6.75
158	6.01	7.17	7.22	6.15	7.84	6.21
159	8.16	7.44	7.42	8.33	7.58	8.49
160	5.30	6.21	6.34			
161	6.15	6.29	6.42			
162	5.47	6.15	6.10			
163	6.84	6.36	6.49			
164	6.87	6.42	6.56			
165	7.19	6.46	6.63			
166	6.26	6.60	6.78			
167	6.22	6.32	6.47			
168	6.22	6.74	6.76			
169	6.52	6.73	6.76			
170	7.05	7.14	6.85			
171	7.52	7.20	6.92	6.39	6.50	6.62
172	8.05	6.61	6.60	6.59	6.27	7.15
173	7.62	7.26	6.99	6.85	6.68	6.70
174	8.05	7.30	7.06	6.85	6.85	7.12
175	8.00	7.32	7.14	6.84	7.03	7.05
176				9.55	7.71	

quarter-out crossvalidation, using the same descriptor combination as in the original model. For details of the procedure, see Katritzky.[44]

In Table 4, the average r^2 of fitting is close to the original value of m1, and the average r^2 of prediction is close to that of the LOO crossvalidation. Thus, the model is not overly dependent on a few particular experimental values.

y-Randomization, also called y-scrambling or permutation test, was said to be 'probably the most powerful validation procedure'.[45] We performed 25 independent y-randomization experiments, and the respective best r^2 values were between 0.2218 and 0.4193, mean 0.3025, standard deviation 0.0423. Thus, not a single best r^2 (nor a r^2_{cv} ($=q^2$)) value from these experiments came close to the corresponding number of the original model.

Selection of a set of descriptors out of a larger descriptor pool implies the risk of chance correlation.[46,47] In order to avoid this risk and to strictly judge the statistical significance of model m1, we generated for our 144 compounds the values of 230 pseudodescriptors made of random numbers, and tried to describe the original target pK_i using a combination of 10 from these, by applying the same descriptor selection procedure as above. In 25 independent such experiments the respective best models had r^2 values between 0.3115 and 0.4574, mean 0.3859, standard deviation 0.0358. The r^2 value of m1 (0.7938) is separated from the mean best random r^2 by about eleven standard deviations and thus is not expected to have arisen by chance.

Finally, bootstrapping was performed as a further diagnostic to get an impression of the variability of

Table 2. Full multilinear QSAR models m1–m4

$$\begin{aligned} \mathrm{p}K_\mathrm{i} = &-24.9625(\pm3.2665)\cdot \mathrm{VAdjEq} + 10.1544(\pm3.7646)\cdot \mathrm{PEOE\_RPC\text{-}} - 0.0777(\pm0.0176)\cdot \mathrm{bpol} \\ &- 0.0272(\pm0.0042)\cdot \mathrm{sMR\_VSA0} - 0.0633(\pm0.0158)\cdot \mathrm{sMR\_VSA3} + 0.0168(\pm0.036)\cdot \mathrm{sMR\_VSA6} \\ &+ 1.1317(\pm0.1898)\cdot \mathrm{MACCS49} + 0.7718(\pm0.1073)\cdot \mathrm{MACCS97} + 0.4512(\pm0.1046)\cdot \mathrm{MACCS116} \\ &+ 0.5177(\pm0.0998)\cdot \mathrm{MACCS152} + 17.0750(\pm1.5651) \end{aligned} \quad \text{(m1)}$$

$$\begin{aligned} \mathrm{p}K_\mathrm{i} = &-0.0233(\pm0.0109)\cdot \mathrm{b\_single} + 0.3635(\pm0.0789)\cdot \mathrm{slog}P + 0.0206(\pm0.0040)\cdot \mathrm{slog}P\mathrm{\_VSA3} \\ &+ 0.8444(\pm0.258)\cdot \mathrm{MACCS49} + 0.4500(\pm0.1103)\cdot \mathrm{MACCS93} + 0.8388(\pm0.0876)\cdot \mathrm{MACCS97} \\ &+ 0.2932(\pm0.1253)\cdot \mathrm{MACCS132} - 0.5876(\pm0.1294)\cdot \mathrm{MACCS140} - 0.2855(\pm0.1591)\cdot \mathrm{MACCS141} \\ &+ 0.5910(\pm0.0980)\cdot \mathrm{MACCS152} + 3.8594(\pm0.4163) \end{aligned} \quad \text{(m2)}$$

$$\begin{aligned} \mathrm{pEC}_{50} = &\ 27.4408(\pm4.1820)\cdot \mathrm{PEOE\_VSA\_FPPOS} + 0.5377(\pm0.0752)\cdot \mathrm{slog}P - 0.0208(\pm0.0057)\cdot \mathrm{slog}P\mathrm{\_VSA0} \\ &+ 0.0334(\pm0.0050)\cdot \mathrm{sMR\_VSA6} - 1.4341(\pm0.1900)\cdot \mathrm{MACCS22} + 0.9581(\pm0.2781)\cdot \mathrm{MACCS49} \\ &- 1.2896(\pm0.3508)\cdot \mathrm{MACCS64} - 0.5947(\pm0.1405)\cdot \mathrm{MACCS80} + 0.8876(\pm0.1679)\cdot \mathrm{MACCS94} \\ &+ 0.3394(\pm0.1068)\cdot \mathrm{MACCS97} - 0.9398(\pm0.1647)\cdot \mathrm{MACCS106} - 0.5726(\pm0.1911)\cdot \mathrm{MACCS109} \\ &+ 1.6799(\pm0.4170)\cdot \mathrm{MACCS125} + 0.1848(\pm0.0874)\cdot \mathrm{MACCS137} + 2.1372(\pm0.5284) \end{aligned} \quad \text{(m3)}$$

$$\begin{aligned} \mathrm{pEC}_{50} = &\ 1.0470(\pm0.0663)\cdot \mathrm{p}K_\mathrm{i} + 0.5246(\pm0.1461)\cdot \mathrm{PEOE\_PC\text{-}} + 0.6195(\pm0.1343)\cdot \mathrm{MACCS57} \\ &+ 0.1795(\pm0.0283)\cdot \mathrm{MACCS62} + 0.1969(\pm0.4695) \end{aligned} \quad \text{(m4)}$$

Numbers in parentheses are standard errors. For explanation of descriptors, see Table 3.

the regression coefficients and to detect any pathologies in the data.[48] For model m1, the result of 10^6 runs on bootstrap samples was $r^2_{\mathrm{bs}} = 0.8067$, standard deviation 0.0319, values that do not point to any problem with model m1.[48] The mean regression coefficients and the intercept resulting from these 10^6 runs were all within 4% of those found for the original model, except that of MACCS116 which deviates by 5.3%.

3.1.2. Training set/test set partition. Since the predictive ability of a model can be assessed only from the result of predictions, we randomly partitioned the 144 compounds with pK_i data available into a training set (90%) and a test set (10%). According to their origin from the references, the compounds are naturally partitioned into groups, that is, tyrosine derivatives group 1 **1**–**23**,[24] group 2 **24**–**52**,[25] group 3 **53**–**94**,[26] thiazolidinediones **95**–**100**,[24] indoles **101**–**110**,[19] fatty acids **134**–**142**,[32] tyrosine derivatives bearing a small N-substituent **149**–**159**,[31] and thiazolidinedione-fatty acid hybrids **160**–**175**.[33] To represent these groups in the training and test sets in a balanced manner, we randomly selected 10% of the compounds from each group. The test set so obtained, containing compounds **8**, **12**, **28**, **43**, **50**, **63**, **70**, **81**, **92**, **96**, **105**, **134**, **156**, **170**, and **174**, with pK_i values well distributed over the whole activity range, was set aside. For the remaining 129 compounds the best model found, m2, was obtained using the genetic algorithm variable selection module of MOE.

$$\begin{aligned} &\mathrm{p}K_\mathrm{i} : \mathrm{b\_single}\quad \mathrm{slog}P\quad \mathrm{slog}P\mathrm{\_VSA3}\quad \mathrm{MACCS49} \\ &\mathrm{MACCS93}\quad \mathrm{MACCS97}\quad \mathrm{MACCS132}\quad \mathrm{MACCS140} \\ &\mathrm{MACCS141}\quad \mathrm{MACCS152} \\ &\qquad n = 129,\ r^2 = 0.7909,\ s = 0.5887, \\ &\qquad F = 44.6,\ r^2_{\mathrm{cv}} = 0.7471,\ s_{\mathrm{cv}} = 0.6475 \end{aligned} \quad \text{(m2)}$$

For the complete model, see Table 2. Three of the ten descriptors in m2 are also in m1 (MACCS49, MACCS97, and MACCS152). In the calculated versus observed plot (Figure 3) the training set compounds are represented by closed symbols.

y-Randomization (25 independent experiments) resulted in a mean best r^2 of 0.3217 (min 0.2305, max 0.3989, standard deviation 0.0439), with not a single best r^2 (or q^2) coming close to those of model m2.

Similarly, description of the original pK_i data by 10 out of 230 random pseudodescriptors (25 independent experiments) yielded a mean best r^2 of 0.4337 (min 0.3694, max 0.4903, standard deviation 0.0327). Thus, the original r^2 is separated from the mean best random r^2 by eleven standard deviations, and m2 is therefore not a chance correlation.

Application of m2 to the 15 test set compounds resulted in $r^2_{\mathrm{pred}} = 0.6998$. The predicted p$K_\mathrm{i}$ values for the test set compounds are underlined in Table 1 and included in Figure 3 (open symbols).

Models m1 and m2 result in similar calculated pK_i values, as given in Table 1 and shown in Figure 4. The average absolute difference for pK_i calculated by m1 and m2 is 0.22.

3.1.3. Combining structural classes. Our study is the first to treat together PPARγ ligands as structurally diverse as tyrosine derivatives, TZDs, indoles, fatty acids, etc., and a priori it is not obvious that all these behave similarly. Therefore, in Figure 5a, the data for model m1 (the fit part of Figure 2) are displayed once more, this time resolved with respect to subgroups. Figure 5b shows corresponding plots for the subgroups. While each subgroup does not cover the whole activity range, each subgroup is assigned by m1 its proper place within the pK_i scale. Furthermore, all subgroups exhibit the same trend as the combined sample.

Table 3. Descriptors used in the final models[a]

b_single	Number of single bonds including bonds to H atoms
VAdjEq	Vertex adjacency information index (equality)
PEOE_PC-	Sum of negative partial charges of atoms, where partial charges are calculated using the PEOE method
PEOE_RPC-	Relative negative partial charge; the smallest negative partial charge divided by the sum of negative partial charges (PEOE partial charges)
PEOE_VSA_FPPOS	Fractional positive polar vdW surface area; sum of vdW surface areas of atoms whose partial charge is greater than 0.2, divided by the total surface area
bpol	Sum of bond polarizabilities; sum over all bonds of differences between atom polarizabilities
slog*P*	log *P* calculated by the atom type contribution method[b]
slog*P*_VSA0	Sum of vdW surface areas of atoms whose contribution to slog*P* is less than or equal to −0.4[b]
slog*P*_VSA3	Sum of vdW surface areas of atoms whose contribution to slog*P* is between 0 and 0.1[b]
sMR_VSA0	Sum of vdW surface areas of atoms whose contribution to sMR is less than or equal to 0.11, where sMR is the molar refraction calculated by the atom type contribution method[b]
sMR_VSA3	Sum of vdW surface areas of atoms whose contribution to sMR is between 0.35 and 0.39[b]
sMR_VSA6	Sum of vdW surface areas of atoms whose contribution to sMR is between 0.485 and 0.56[b]
MACCS22	Number of atoms in 3-membered rings
MACCS49	1 if molecule is formally charged, 0 otherwise
MACCS57	Number of O atoms in rings
MACCS62	Number of ring atoms vicinal to a non-ring bond that immediately connects rings
MACCS64	Number of non-ring S atoms attached to a ring
MACCS80	Number of N atoms separated by 4 bonds
MACCS93	Number of methylated heteroatoms
MACCS94	Number of N atoms bonded to at least one non-C heavy atom
MACCS97	Number of O atoms 4 bonds away from an N atom
MACCS106	Number of atoms bonded to at least 3 non-C heavy atoms
MACCS109	Number of O–CH_2 bonds
MACCS116	Number of CH_2 groups 3 bonds from a CH_3
MACCS125	1 if there are at least 2 aromatic rings, 0 otherwise
MACCS132	Number of CH_2 groups 2 bonds away from an O atom
MACCS137	Total number of heteroatoms in rings
MACCS140	Number of O atoms decreased by 3 if there are more than 3 O; 0 otherwise
MACCS141	Number of CH_3 decreased by 2 if there are more than 2 CH_3; 0 otherwise
MACCS152	Number of C atoms bonded to 2 or more C atoms and 1 O atom

[a] For details, see Refs. 36 and 39.
[b] See Wildman, S. A.; Crippen, G. M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868.

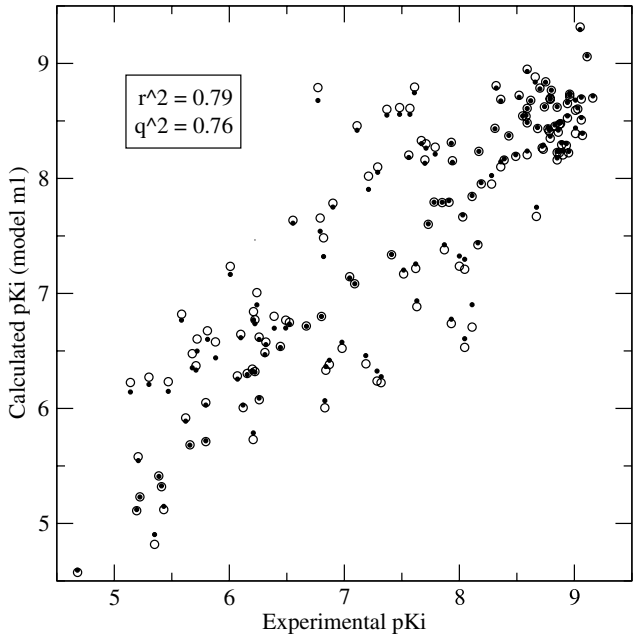


Figure 2. Calculated and observed pK_i values for receptor binding of PPARγ agonists (model m1). Closed symbols represent fit, open symbols represent LOO-crossvalidated values.

Table 4. Results of leave-one-quarter-out crossvalidation of model m1

Set to fit	r^2 (fit)	Set to predict	r^2 (pred)
1, 2, and 3	0.8180	4	0.7037
1, 2, and 4	0.7951	3	0.7650
1, 3, and 4	0.7867	2	0.7874
2, 3, and 4	0.7882	1	0.7983
Average	0.7970		0.7636

3.2. Transactivation

For gene transactivation by the compounds under study we built two models, a QSAR model in the literal sense and an activity–activity relation.

3.2.1. QSAR model. For the 150 compounds with data available (pEC_{50} values) for gene transactivation by the PPARγ/ligand complex, 229 descriptors remained after removal of collinear ones. The best MLR equation, selected by the step-up procedure, is 14-descriptor model m3:

$$\begin{aligned} &pEC_{50} : \text{PEOE\_VSA\_FPPOS} \quad \text{slog}P \quad \text{s log } P\text{\_VSA0} \\ &\text{sMR\_VSA6} \quad \text{MACCS22} \quad \text{MACCS49} \quad \text{MACCS64} \\ &\text{MACCS80} \quad \text{MACCS94} \quad \text{MACCS97} \quad \text{MACCS106} \\ &\text{MACCS109} \quad \text{MACCS125} \quad \text{MACCS137} \\ &\quad n = 150,\ r^2 = 0.6487,\ s = 0.7335, \\ &\quad F = 17.80,\ r^2_{\text{cv}} = 0.5727,\ s_{\text{cv}} = 0.8089. \end{aligned} \tag{m3}$$

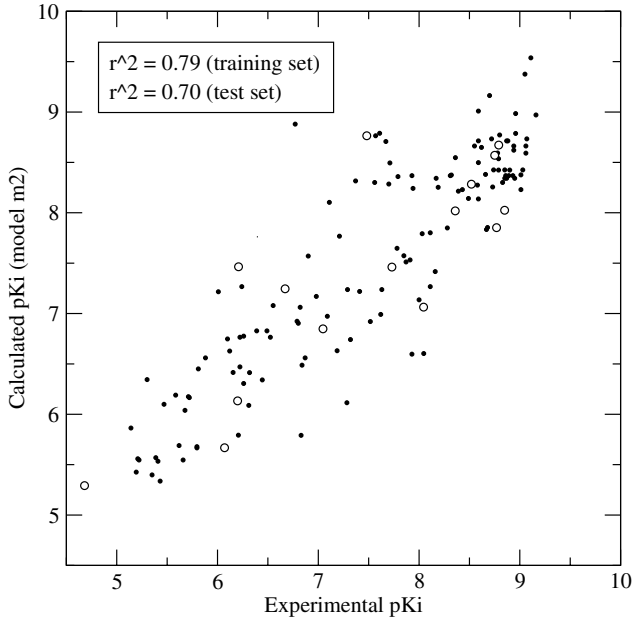


Figure 3. Calculated and observed pK_i values for receptor binding of PPARγ agonists (model m2). Closed symbols represent fit values for the training set, open symbols represent predictions for the test set.

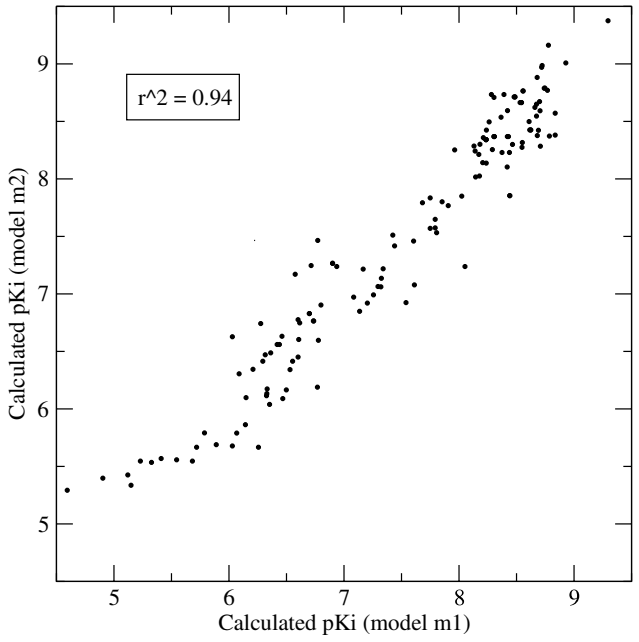


Figure 4. pK_i values for receptor binding of PPARγ agonists calculated using models m1 and m2.

For the full model, see Table 2. A calculated versus observed plot is given in Figure 6 for both fit and LOO-crossvalidated data (closed and open symbols, respectively).

Of the 14 descriptors appearing in m3 three are also in m1 (sMR_VSA6, MACCS49, and MACCS97) and three are also in m2 (slog*P*, MACCS49, and MACCS97). This seems to be more than coincidence: binding is a prerequisite for transactivation.

The result of a four times leave-one-quarter-out cross-validation is shown in Table 5. The average r^2 of fitting is similar to the original one of m3, and the average r^2 of prediction is acceptable, demonstrating the robustness of m3.

In 25 independent y-randomization experiments, the mean best r^2 was 0.3469 (min 0.2870, max 0.4514, standard deviation 0.0435). Not a single best r^2 or r^2_{cv} (q^2) value from the y-randomization experiments came close to the original r^2 or q^2 of m3.

We also generated for our 150 compounds the values of 229 pseudo-descriptors made of random numbers, and tried to describe the target pEC_{50} by a combination of these, applying the same descriptor selection procedure as above. In 25 independent such experiments the mean best r^2 was 0.4543 (min 0.3966, max 0.5288, standard deviation 0.0368). The real $r^2 = 0.6487$ (model m3) is five standard deviations away from the mean and m3 thus is not expected to have arisen by chance.

The result of 10^6 bootstrap runs is $r^2_{bs} = 0.6849$, standard deviation 0.0400, values that do not point to any problem with model m3.[48] The mean regression coefficients resulting from these 10^6 runs are all within 5% of those found for the original model. Thus, all validation procedures agree that m3, notwithstanding its lower quality compared to m1, is still statistically valid.

3.2.2. Activity–activity model. Not surprisingly, there is a rather high correlation between pEC_{50} values (transactivation) and pK_i values (binding) in our data set ($r^2 = 0.6153$, $n = 118$). It should therefore be possible to establish an activity–activity relationship. Obviously, such a relation would allow prediction of pEC_{50} for those compounds that have a pK_i available. The best four-descriptor model found is m4:

$$\begin{aligned} pEC_{50} : pK_i \quad & \text{PEOE\_PC-} \quad \text{MACCS57} \quad \text{MACCS62} \\ & n = 118,\ r^2 = 0.7618,\ s = 0.6087, \\ & F = 90.34,\ r^2_{cv} = 0.7385,\ s_{cv} = 0.6378. \end{aligned} \qquad \text{(m4)}$$

Compare $s = 0.61$ with the standard deviation of pEC_{50} in this compound population, 1.23 log units. As seen in Table 2, in m4 the regression coefficient of pK_i is close to unity, and the intercept may well be zero. Bootstrapping (10^6 runs) resulted in $r^2_{bs} = 0.7686$, standard deviation 0.0332, not pointing to any problem with model m4.[48] The mean regression coefficients and intercept resulting from these 10^6 runs are all within 5% of those found originally.

4. Discussion

4.1. Binding

Our QSAR models for PPARγ ligand binding represent a significant step forward. Comparing the statistics of m1 and m2 to those of published models for PPARγ binding,[9–14] it should be noted that (i) our models are derived

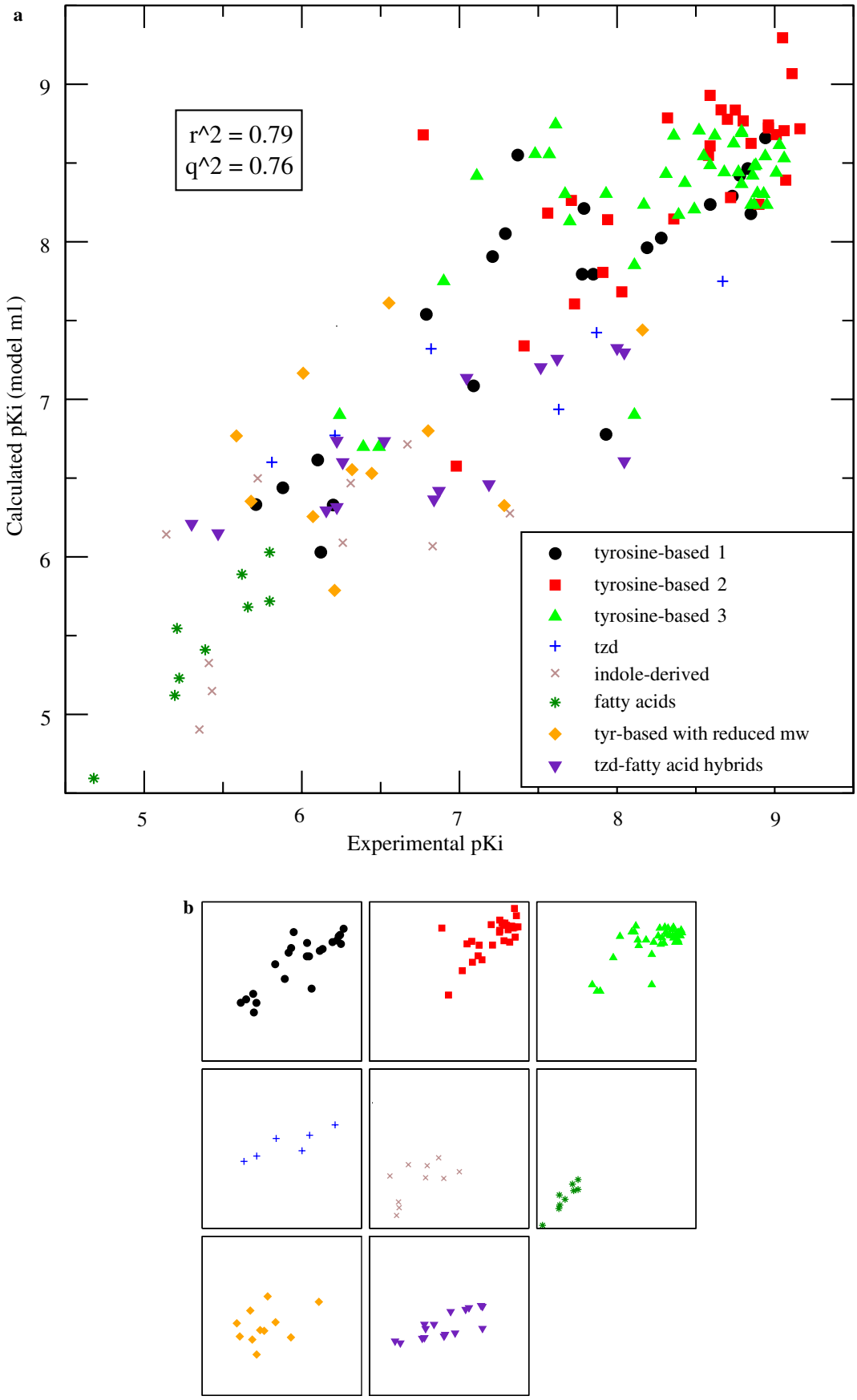


Figure 5. (a) Calculated pK_i (by model m1) versus experimental pK_i, broken down by subgroups. For explanation of, symbols see inset. (b) The same data displayed separately for the subgroups.

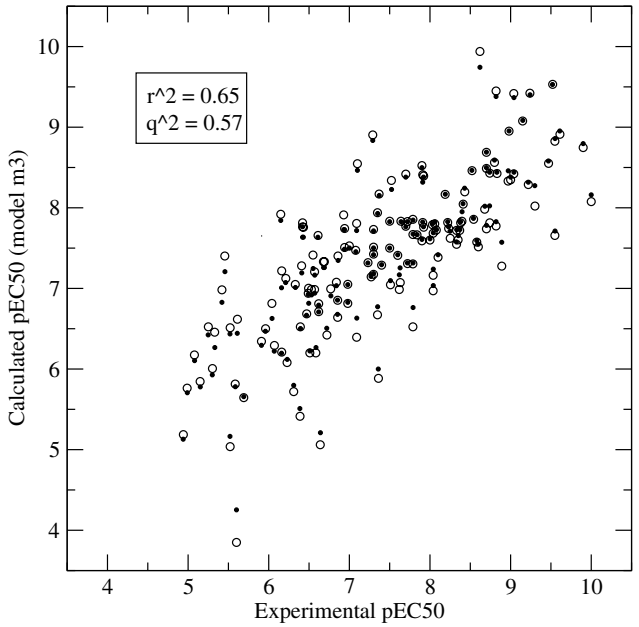


Figure 6. Calculated and observed pEC_{50} values for transactivation by PPARγ agonists (model m3). Closed symbols represent fit, open symbols represent LOO-crossvalidated values.

Table 5. Results of leave-one-quarter-out crossvalidation of model m3

Set to fit	r^2 (fit)	Set to predict	r^2 (pred)
1, 2, and 3	0.6747	4	0.4547
1, 2, and 4	0.6719	3	0.5038
1, 3, and 4	0.6640	2	0.5596
2, 3, and 4	0.6247	1	0.6512
Average	0.6588		0.5423

from observations of a far larger and more diverse set of compounds, (ii) we did not exclude any observations before or after the analysis ('outliers'), and (iii) our models do not suffer from the large $r^2 - r^2_{cv}$ gap that is often observed for CoMFA and similar models.[10–13]

In statistical data treatment, it is always problematic to combine observations on heterogeneous samples, since a correlation found for the combined sample may vanish within subgroups, or trends present in the subgroups may be obscured or even reversed in the combined sample. In order to enable predictions for a broad range of compounds, a combined treatment of the compound classes considered here was highly desirable. On the other hand, in the beginning of this work it was not at all clear whether this was possible. The result shown in Figure 5 therefore is important. We tentatively interpret this finding as follows: from the receptor's point of view, the variation in the central part of the ligand structure seems to be of minor importance compared to the general structural pattern that may be roughly described as 'anionic head group linked to flexibly interconnected hydrophobic and preferably unsaturated (aromatic) moieties'.

There are conflicting opinions on how a statistical model should be validated. Tropsha argues that a training set/test set partition should always be done.[49,50] On the other hand, Hawkins emphasizes that reserving a considerable part of the experimental observations just for validation may be a waste of information compared to using all data for model building.[51,52] We therefore decided to use both options, thus obtaining alternative models m1 and m2.

Model m1 is based on all 144 experimental observations. The absolute t values of all descriptors and the intercept in m1 are above 2.6 (Table 2). There is one pair of highly intercorrelated descriptors included in model m1, VAdjEq and PEOE_RPC- ($r^2 = 0.79$). All other pairwise descriptor intercorrelations have $r^2 < 0.68$. A high intercorrelation of two descriptors is not automatically prohibitive, since important for multilinear regression is not in what two descriptors agree, but in what they differ.[53–55]

Though its $r^2 = 0.79$ appears low compared to earlier studies, m1 has $s = 0.58$ log units, which is appreciably smaller than the standard deviation of the experimental data, 1.24 log units. Given the facts that m1 applies to the broadest variety of PPARγ agonists treated so far, and that it uses simple descriptors, it may be a useful model of wide applicability.

A superficial glance at the statistics of m1 is encouraging, and in particular $F = 51$ is far above the critical tabulated value 1.90 ($\alpha = 5\%$) for 144 data points and 10 descriptors. However, it is important to realize that the descriptor combination in m1 was selected from in principle 9.36×10^{16} combinations of 10 out of 230 descriptors. The number of descriptors in the pool was 1.6 times the number of compounds. In such a situation the possibility to obtain chance correlations is a serious and often overlooked issue, and the conventional tabulated F values are not relevant. The problem has been known since 1972 at least,[46,47] it was termed descriptor selection bias and was rediscussed recently.[56,57] We addressed this issue by performing y-randomization and by using random pseudodescriptors. In y-randomization the target activity values are randomly permuted, leaving all descriptor values untouched, and for the permuted y values the best QSAR model is built using the same descriptor selection procedure that led to the original model. This is repeated several times. Since the link between structure and activity is deliberately destroyed, the resulting models should be of far lower quality than the real model.[58,59] In fact, y-randomization is an approximation of the action of chance. The full action of chance was probed in our further experiments using random pseudodescriptors. The result shows that chance produces models of far lower quality than m1, though of somewhat higher quality than those obtained by y-randomization. While such experiments can demonstrate the statistical significance of a model as a whole, information on the significance of particular descriptors present in a model cannot be extracted therefrom.

Model m2 is based on 129 compounds in a training set. The absolute t values of all descriptors and the intercept in m2 are above 1.7. The highest pairwise descriptor

intercorrelation in m2 is that of MACCS141 and MACCS152, $r^2 = 0.69$. As judged by simple statistics, y-randomization and use of random pseudodescriptors, m2 is of a quality comparable to m1. Predictions of pK_i for the remaining 15 compounds (test set) are in reasonable agreement with experimental values ($r^2_{pred} = 0.70$), so that m2 can be considered predictive. For most compounds, pK_i values calculated according to m1 or m2 are similar, though only 3 out of the 10 descriptors in each model are shared. This fact alone is a strong caveat against (over)interpretation of the presence of particular descriptors in the models. On the other hand, this fact indicates some predictive ability of m1 also.

Both models share descriptors MACCS49, MACCS97, and MACCS152. MACCS49 is an indicator of formal charge, and regarding the binding mode as revealed by X-ray crystallography,[15,16,21,32] the carboxylate charge is in fact expected to play an important role. The other two MACCS descriptors are substructure counts whose role is not as easily understood.

4.2. Transactivation

After a previous attempt at describing the transactivation behavior of even a rather homogeneous series of PPARγ agonists by 3D-QSAR met with failure,[11] and after other relevant 3D-QSAR publications were silent on this issue,[9,10,12] treatment of transactivation by 2D-QSAR for a large and diverse compound set was a challenge. The result was not unexpected. In model m3, for 150 diverse compounds 14 descriptors were required to achieve $r^2 = 0.65$. Still, model m3 has $s = 0.73$ log units, appreciably smaller than the standard deviation of the experimental data, 1.18 log units. Differences between crossvalidated and fitted statistics are reasonably small. All pairwise descriptor intercorrelations in m3 have $r^2 < 0.35$. The absolute t values of all descriptors and the intercept in m3 are above 2.1. However, with the descriptor combination in m3 selected from in principle 8.34×10^{21} combinations, and with the rather low r^2 the risk of chance correlation here seemed worse than before. Therefore, again thorough validation was obligatory. As shown in the results part, validation procedures did not reveal any special problems with m3.

Model m3 shares with model m1 descriptors sMR_VSA6, MACCS49, and MACCS97, and with model m2 slog*P* and again MACCS49 and MACCS97. Both these MACCS descriptors are essential in the binding models, and slog*P*, an index of hydrophobicity, may emphasize the role of hydrophobic contacts in binding and in membrane permeation. Notably, in model m4 these descriptors do not appear, since reasonably their effects are included in pK_i already.

The difficulties encountered in building QSAR models for transactivation may be a result of the biological processes involved being much more complicated than the physicochemical process of binding. Thus, transactivation is the result of at least three steps each depending on ligand structure, (a) diffusion or transport of the ligand through membranes into the cell and into the nucleus, (b) ligand binding to the receptor, and (c) conformational change of the receptor induced by ligand binding. For the overall process, one cannot expect high correlation with a few descriptors, since this would require all steps to depend on the same descriptors in a similar manner. Moreover, the experimental EC_{50} data may not be sufficiently comparable between the various compounds, in that probably not all important confounders were identified and kept constant for all measurements. For these reasons we did not devote further work to models for transactivation such as m3.

Prediction of pEC_{50} from a given pK_i would be an easy exercise if a high-quality activity–activity relation could be found. However, such a relation will be difficult to establish and to validate for the same reasons as in the case of m3. Further, in an activity–activity model such as m4 (in contrast to a model built exclusively on molecular structure) one of the prerequisites for linear regression is violated, the assumption of negligible errors in the independent variables.

5. Conclusion

Portable and easy-to-use predictive models were derived for receptor binding for the largest and most diverse set of PPARγ ligands treated so far. Allowing advance estimation of binding of prospective ligands to PPARγ, these models should become valuable tools in drug design. Corresponding models for transactivation by a similarly diverse set of PPARγ ligands were of lower quality. Possible reasons for the latter fact are the low quality of experimental data and the inadequacy of a few simple descriptors to model a biological phenomenon as complex as transactivation.

Acknowledgments

We thank Professors Urs A. Meyer, Torsten Schwede, and Joseph Gut for providing various kinds of support, and Hubert Hug and Robert Dannecker for critical discussions. This research was funded by the Swiss Commission for Technical Innovation (KTI/CTI, Grant 6570.2 MTS-LS).

References and notes

1. Willson, T. M.; Brown, P. J.; Sternbach, D. D.; Henke, B. R. *J. Med. Chem.* **2000**, *43*, 527.
2. Willson, T. M.; Lambert, M. H.; Kliewer, S. A. *Ann. Rev. Biochem.* **2001**, *70*, 341.
3. Rangwala, S. M.; Lazar, M. A. *Trends Pharmacol. Sci.* **2004**, *25*, 331.
4. Henke, B. R. *J. Med. Chem.* **2004**, *47*, 4118.
5. Wang, M.; Tafuri, S. *J. Cell. Biochem.* **2003**, *89*, 38.
6. Kallenberger, B. C.; Love, J. D.; Chatterjee, V. K. K.; Schwabe, J. W. R. *Nat. Struct. Biol.* **2003**, *10*, 136.

7. Ferry, G.; Bruneau, V.; Beauverger, P.; Goussard, M.; Rodriguez, M.; Lamamy, V.; Dromaint, S.; Canet, E.; Galizzi, J.-P.; Boutin, J. A. *Eur. J. Pharmacol.* **2001**, *417*, 77.
8. Schopfer, F. J.; Lin, Y.; Baker, P. R. S.; Cui, T.; Garcia-Barrio, M.; Zhang, J.; Chen, K.; Chen, Y. E.; Freeman, B. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 2340.
9. Rathi, L.; Kashaw, S. K.; Dixit, A.; Pandey, G.; Saxena, A. K. *Bioorg. Med. Chem.* **2004**, *12*, 63.
10. Liao, C.; Xie, A.; Shi, L.; Zhou, J.; Lu, X. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 230.
11. Liao, C.; Xie, A.; Zhou, J.; Shi, L.; Li, Z.; Lu, X. *J. Mol. Model.* **2004**, *10*, 165.
12. Hyun, K. H.; Lee, D. Y.; Lee, B.-S.; Kim, C. K. *QSAR Comb. Sci.* **2004**, *23*, 637.
13. Khanna, S.; Sobhia, M. E.; Bharatam, P. V. *J. Med. Chem.* **2005**, *48*, 3015.
14. Hemalatha, R.; Soni, L. K.; Gupta, A. K.; Kaskhedikar, S. G. *E.-J. Chem.* **2004**, *1*, 243, http://www.websamba.com/ejchem/5%20issue/243-250.pdf.
15. Nolte, R. T.; Wisely, G. B.; Westin, S.; Cobb, J. E.; Lambert, M. H.; Kurokawa, R.; Rosenfeld, M. G.; Willson, T. M.; Glass, C. K.; Milburn, M. V. *Nature* **1998**, *395*, 137.
16. Gampe, R. T.; Montana, V. G.; Lambert, M. H.; Miller, A. B.; Bledsoe, R. K.; Milburn, M. V.; Kliewer, S. A.; Willson, T. M.; Xu, H. E. *Mol. Cell* **2000**, *5*, 545.
17. Soni, L. K.; Gupta, A. K.; Kaskhedikar, S. G. *Eur. J. Chem.* **2004**, *1*, 170, http://www.websamba.com/ejchem/cas%20issue%203/170-177.pdf..
18. Yu, C.; Chen, L.; Luo, H.; Chen, J.; Cheng, F.; Gui, C.; Zhang, R.; Shen, J.; Chen, K.; Jiang, H.; Shen, X. *Eur. J. Biochem.* **2004**, *271*, 386.
19. Henke, B. R.; Adkison, K. K.; Blanchard, S. G.; Leesnitzer, L. M.; Mook, R. A., Jr.; Plunket, K. D.; Ray, J. A.; Roberson, C.; Unwalla, R.; Willson, T. M. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 3329.
20. Davis, R. G.; Anderegg, R. J.; Blanchard, S. G. *Tetrahedron* **1999**, *55*, 11653.
21. Xu, H. E.; Lambert, M. H.; Montana, V. G.; Plunket, K. D.; Moore, L. B.; Collins, J. L.; Oplinger, J. A.; Kliewer, S. A.; Gampe, R. T., Jr.; McKee, D. D.; Moore, J. T.; Willson, T. M. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 13919.
22. Yanagisawa, H.; Takamura, M.; Yamada, E.; Fujita, S.; Fujiwara, T.; Yachi, M.; Isobe, A.; Hagisawa, Y. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 373.
23. Nichols, J. S.; Parks, D. J.; Consler, T. G.; Blanchard, S. G. *Anal. Biochem.* **1998**, *257*, 112.
24. Henke, B. R.; Blanchard, S. G.; Brackeen, M. F.; Brown, K. K.; Cobb, J. E.; Collins, J. L.; Harrington, W. W., Jr.; Hashim, M. A.; Hull-Ryde, E. A.; Kaldor, I.; Kliewer, S. A.; Lake, D. H.; Leesnitzer, L. M.; Lehmann, J. M.; Lenhard, J. M.; Orband-Miller, L. A.; Miller, J. F.; Mook, R. A., Jr.; Noble, S. A.; Oliver, W., Jr.; Parks, D. J.; Plunket, K. D.; Szewczyk, J. R.; Willson, T. M. *J. Med. Chem.* **1998**, *41*, 5020.
25. Collins, J. L.; Blanchard, S. G.; Boswell, G. E.; Charifson, P. S.; Cobb, J. E.; Henke, B. R.; Hull-Ryde, E. A.; Kazmierski, W. M.; Lake, D. H.; Leesnitzer, L. M.; Lehmann, J.; Lenhard, J. M.; Orband-Miller, L. A.; Gray-Nunez, Y.; Parks, D. J.; Plunkett, K. D.; Tong, W.-Q. *J. Med. Chem.* **1998**, *41*, 5037.
26. Cobb, J. E.; Blanchard, S. G.; Boswell, E. G.; Brown, K. K.; Charifson, P. S.; Cooper, J. P.; Collins, J. L.; Dezube, M.; Henke, B. R.; Hull-Ryde, E. A.; Lake, D. H.; Lenhard, J. M.; Oliver, W., Jr.; Oplinger, J.; Pentti, M.; Parks, D. J.; Plunket, K. D.; Tong, W.-Q. *J. Med. Chem.* **1998**, *41*, 5055.
27. Willson, T. M.; Mook, R. A.; Kaldor, I.; Henke, B. R.; Deaton, D. N.; Collins, J. L.; Cobb, J. E.; Brackeen, M.; Sharp, M. J.; O'Callaghan, J. M.; Erickson, G. A.; Boswell, G. E. Patent WO 9731907 (1997) [=US 6294580 (1998)], *Chem. Abstr.* 127:278064.
28. Lehmann, J. M.; Moore, L. B.; Smith-Oliver, T. A.; Wilkison, W. O.; Willson, T. M.; Kliewer, S. A. *J. Biol. Chem.* **1995**, *270*, 12953.
29. Liu, K. G.; Smith, J. S.; Ayscue, A. H.; Henke, B. R.; Lambert, M. H.; Leesnitzer, L. M.; Plunket, K. D.; Willson, T. M.; Sternbach, D. D. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2385.
30. Liu, K. G.; Lambert, M. H.; Leesnitzer, L. M.; Oliver, W., Jr.; Ott, R. J.; Plunket, K. D.; Stuart, L. W.; Brown, P. J.; Willson, T. M.; Sternbach, D. D. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2959, In this publication the identity of monomer building block alcohol **f** is erroneous. The building block actually used was [4-(5-cyclopropyl-1,2,4-oxadiazol-3-yl)phenyl]methanol (e-mail message of D. D. Sternbach to C. R.).
31. Liu, K. G.; Lambert, M. H.; Ayscue, A. H.; Henke, B. R.; Leesnitzer, L. M.; Oliver, W. R., Jr.; Plunket, K. D.; Xu, H. E.; Sternbach, D. D.; Willson, T. M. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 3111.
32. Xu, H. E.; Lambert, M. H.; Montana, V. G.; Parks, D. J.; Blanchard, S. G.; Brown, P. J.; Sternbach, D. D.; Lehmann, J. M.; Wisely, G. B.; Willson, T. M.; Kliewer, S. A.; Milburn, M. V. *Mol. Cell* **1999**, *3*, 397, In this publication numerical values for binding are given as IC_{50}. Under the measurement conditions the difference between pIC_{50} and pK_i is less than 0.1 log units (e-mail message of S. G. Blanchard to M. S.). We therefore treated these IC_{50} values as if they were K_i.
33. Tomkinson, N. C. O.; Sefler, A. M.; Plunket, K. D.; Blanchard, S. G.; Parks, D. J.; Willson, T. M. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2491.
34. Parks, D. J.; Tomkinson, N. C. O.; Villeneuve, M. S.; Blanchard, S. G.; Willson, T. M. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3657.
35. Haigh, D.; Allen, G.; Birrell, H. C.; Buckle, D. R.; Cantello, B. C. C.; Eggleston, D. S.; Haltiwanger, R. C.; Holder, J. C.; Lister, C. A.; Pinto, I. L.; Rami, H. K.; Sime, J. T.; Smith, S. A.; Sweeney, J. D. *Bioorg. Med. Chem.* **1999**, *7*, 821.
36. Molecular Operating Environment, version 2004.03; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada.
37. Also available within MOE are electrotopological state indices.[38] These, however, could not be used due to errors in the particular implementation provided within MOE.
38. Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: San Diego, 1999.
39. Xue, L.; Bajorath, J. *J. Chem. Inf. Comput. Sci* **2002**, *42*, 757.
40. Braun, J.; Kerber, A.; Meringer, M.; Rücker, C. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 163.
41. Rücker, C.; Braun, J.; Kerber, A.; Laue, R. http://www.mathe2.uni-bayreuth.de/molgenqspr.
42. Rücker, C.; Meringer, M.; Kerber, A. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2070.
43. Rücker, C.; Meringer, M.; Kerber, A. *J. Chem. Inf. Model* **2005**, *45*, 74.
44. Katritzky, A. R.; Fara, D. C.; Karelson, M. *Bioorg. Med. Chem.* **2004**, *12*, 3027.
45. Kubinyi, H. QSAR in Drug Design. In *Handbook of Chemoinformatics*, Gasteiger, J., Ed.; Wiley-VCH: Weinheim, 2003; Vol. 4, Chapter X.4.2, pp 1532–1554.
46. Topliss, J. G.; Costello, R. J. *J. Med. Chem.* **1972**, *15*, 1066.

47. Topliss, J. G.; Edwards, R. P. *J. Med. Chem.* **1979**, *22*, 1238.
48. Cramer, R. D.; Bunce, J. D.; Patterson, D. E.; Frank, T. E. *Quant. Struct.-Act Relat.* **1988**, *7*, 18.
49. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Model* **2002**, *20*, 269.
50. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69.
51. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579.
52. Hawkins, D. M. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1.
53. Randić, M. *New J. Chem.* **1991**, *15*, 517.
54. Randić, M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672.
55. Peterangelo, S. C.; Seybold, P. G. *Int. J. Quantum Chem.* **2004**, *96*, 1.
56. Livingstone, D. J.; Salt, D. W. *J. Med. Chem.* **2005**, *48*, 661.
57. Livingstone, D. J.; Salt, D. W. *Rev. Comput. Chem.* **2005**, *21*, 287.
58. Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 309–318.
59. Baumann, K.; Stiefl, N. *J. Comput. Aided Mol. Des.* **2004**, *18*, 549.