QSAR of PPARγ Agonist Binding and Transactivation: Some Typical Problems Encountered in QSAR

Christoph Rücker,*[,§] Marco Scarsi,[§] and Markus Meringer[#]

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

Received ...

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 can be described by models that contain simple 2D molecular descriptors and nevertheless are of good quality and predictive power. On the other hand, modeling of gene transactivation, the functional activity of the agonists, turned out to be much more difficult. The models presented are thoroughly validated by crossvalidation, randomization experiments, bootstrapping, and training set/test set partitioning. Problems encountered that are typical for QSAR studies are discussed in some detail.

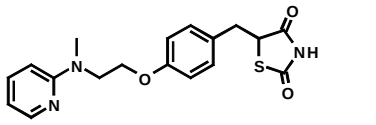
INTRODUCTION

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.[1-4] 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,[5] 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.

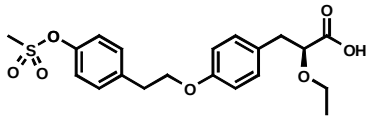
*Corresponding author phone: +41-61-267-1469; fax: +41-61-267-1584;
e-mail: christoph.ruecker@unibas.ch
[§]Biocenter, University of Basel
[#]Kiadis B.V.

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.[1-4]

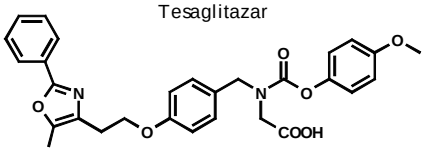
For PPARγ, several unsaturated fatty acids, in particular prostaglandins[6] and nitrolinoleic acids,[7] have been proposed as natural ligands. A few synthetic PPARγ agonists are approved drugs (e.g. rosiglitazone, a thiazolidinedione (TZD)) or under development in several pharmaceutical companies as antidiabetics (e.g. tesaglitazar, an O-analogous tyrosine derivative, or muraglitazar).



Rosiglitazone



Tesaglitazar



Muraglitazar

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), or as IC_{50}, the concentration that results in 50% binding. 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, it 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 is 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 case.

Numerical values for the activities of many PPARγ agonists have been published, resulting from research in several pharmaceutical companies. It would be very valuable to transform this wealth of data into information, with the goal 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.[8-12] While the first of these dealt with thiazolidinediones,[8] others[9-12] treated one or a few series of tyrosine derivatives originating from Glaxo Wellcome research and published in 1998.[13-15] The more advanced of these studies,[11,12] using the CoMFA and CoMSIA methods, 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.[16,17] The resulting models are reported to be of high quality (though a few compounds were excluded before and after the analyses), but they include rather limited numbers and types of PPARγ agonists, and their use is restricted by the necessity of alignment. The same is true for a recent CoMFA study on binding of dual PPARα/PPARγ activators.[18]

As to human gene activation by PPARγ agonists, using the above methods good models for EC_{50} could not be obtained.[11] For a few agonists a QSAR equation between mouse PPARγ transactivation and a docking scoring function was given.[19] For two rather limited series of human PPARγ agonists, QSAR equations for transactivaton were published recently.[20,21]

The aim of the present study was to develop simple and easily portable models of broader applicability for both human PPARγ binding of and PPARγ-mediated gene activation by small molecules. We aimed at including all known series of PPARγ agonists with appropriate experimental data available.

METHODS

Experimental Data. Unfortunately, various protocols are in use for measuring both receptor binding of and transactivation by PPARγ agonists,[22] and numbers obtained for the same compound under various protocols differ considerably. For example, for rosiglitazone we found published IC_{50} values for binding to human PPARγ varying from 36 nM[23] to 465 nM.[24] Alternatively, the same phenomenon was described by K_i values ranging from 47 nM[25] to 230 nM.[26] For PPARγ-mediated human gene activation by rosiglitazone EC_{50} values covering anything between 18 nM[27] and 730 nM[28] 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[23] were published in references 13-15, in the corresponding patent[29] data were disclosed that were obtained in a classical solution scintillation assay for an overlapping set of agonists. For the 61 compounds with both kinds of numerical values available, data are displayed in Figure 1.

From Figure 1 we have to conclude that the two methods either 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, i.e. 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.[10-12]

For gene activation we decided to use the data obtained from a transient cotransfection assay likewise developed at Glaxo.[30] Interestingly, again for a series of 65 agonists data are available both from the patent[29] and from the ensuing publications,[13-15] seemingly all obtained using this same assay. Nevertheless there is a remarkable scattering in the data, as shown in Figure 2.

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

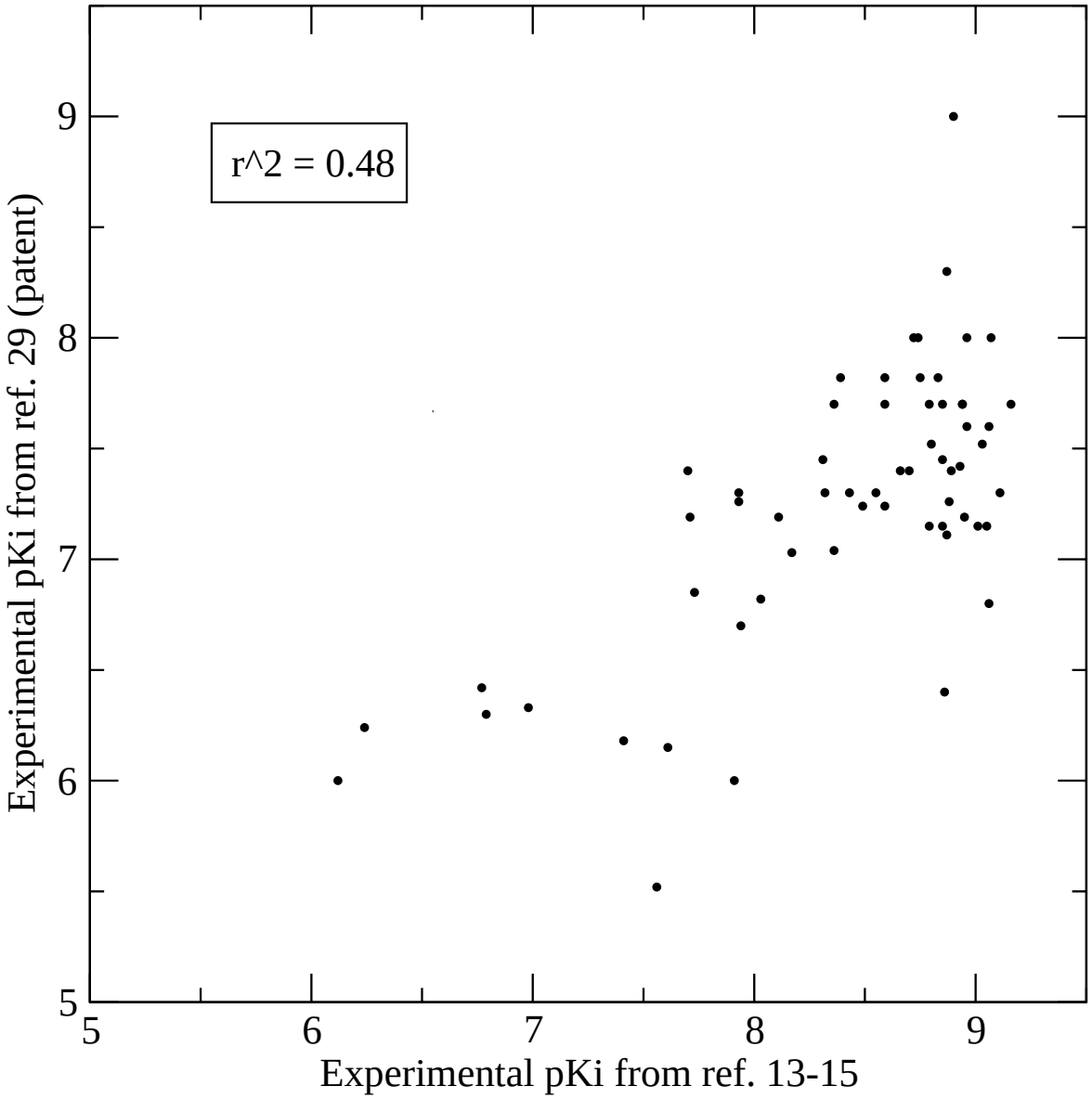


Figure 1. Experimental pK_i values for receptor binding of some PPARγ agonists, measured in a classical solution scintillation assay (reference 29) and in a scintillation proximity assay (references 13-15).

publications. Obviously, the quality of any data treatment result is limited by the quality of the input data.

By the above, the synthetic agonists included in our study are essentially limited to those from Glaxo research. Specifically, along with the tyrosine-based compounds and the few thiazolidinediones from references 13-15, indole derivatives,[25] oxadiazole-substituted α-isopropoxyphenylpropanoic acids,[31] α,α-dimethyl-aminopropylphenoxyacetic

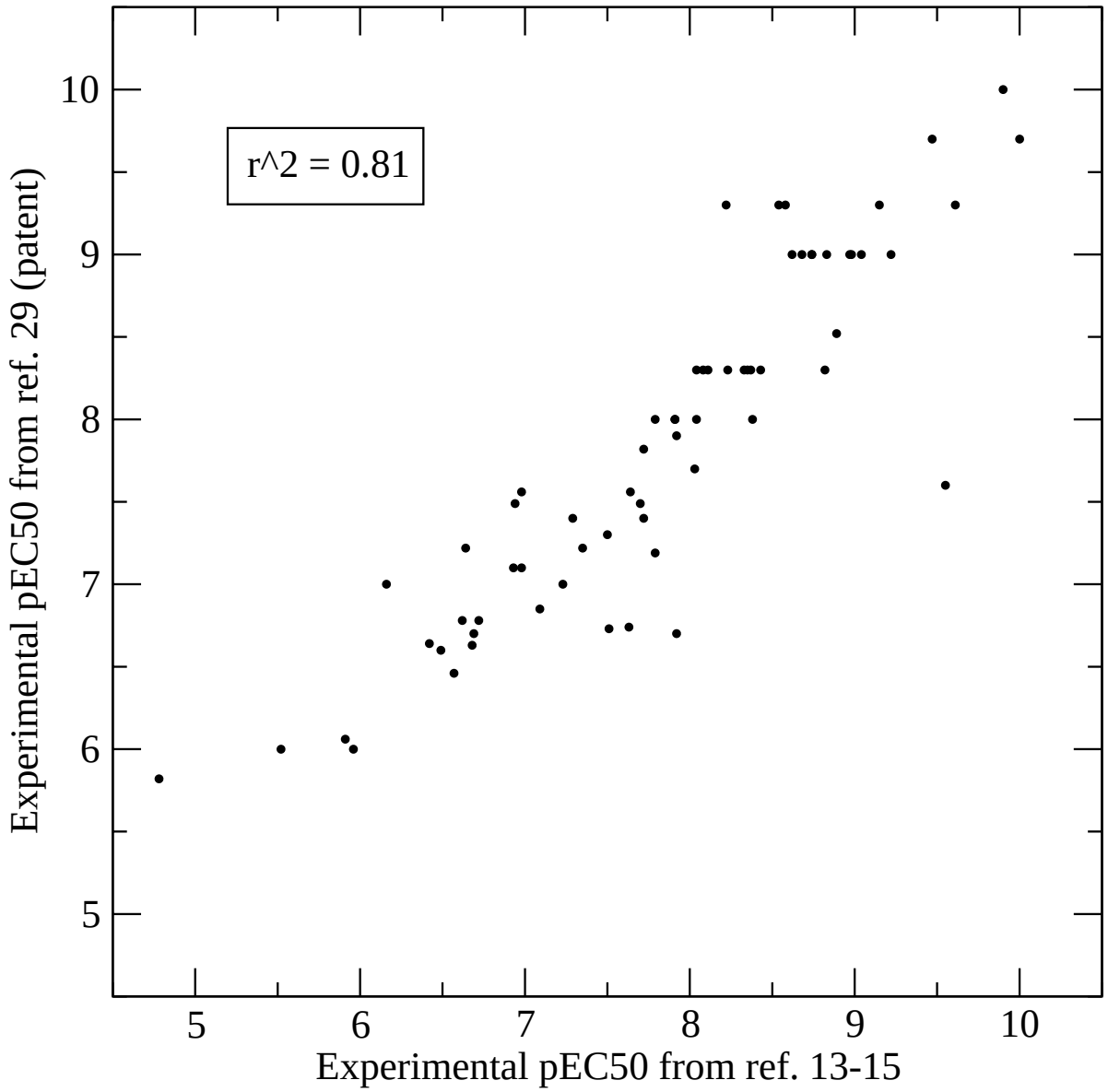


Figure 2. Experimental pEC_{50} values for gene activation by some PPARγ agonists, measured in a transient cotransfection assay and taken from reference 29 and from references 13-15, respectively.

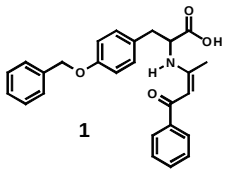
Stereochemistry. Published numerical values for the binding and transactivation behavior of chiral PPARγ agonists were obtained for pure *S* enantiomers in some cases, for racemates in others, though the activity is known to almost completely reside in the *S* enantiomers.[36,37] In order to render racemate data comparable to pure *S* enantiomer data, we added

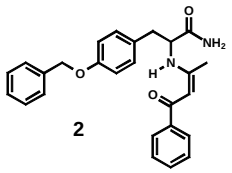
0.3 (=$\log_{10}$ 2) to all pK_i and pEC_{50} values of racemates taken from the literature, which is equivalent to assuming that the concentration of a racemate required to obtain a certain effect is twice the concentration of the corresponding active enantiomer, and to ignoring any racemization that might occur to a pure enantiomer. These assumptions are discussed in the Discussion section. The stereocorrection (0.3 log units) is small compared to the intrinsic scatter of the data. For example, Figures 1 and 2 above were obtained from stereocorrected data. In the worst cases the difference between corresponding x and y value in Figure 1 is about 2 log units (e.g. **37**, **38**, **39**, **68**), and 1-2 log units in Figure 2 (e.g. **46**, **64**, **75**, **85**). Corresponding plots obtained from data not stereocorrected look essentially the same (not shown).

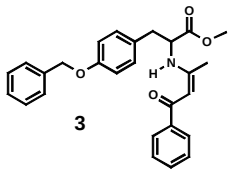
Compound set. The final agonist set consists of 176 compounds (Scheme 1), 144 of which have measured pK_i values for binding to PPARγ, 150 of which have measured pEC_{50} values for transactivation, and 118 have both (Table 1). The pK_i range in the data set is from 4.68 to 9.16 (mean 7.52, standard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

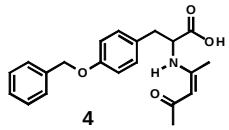
Descriptors. In what follows we used a pool of molecular descriptors consisting of those supplied by the program MOE[38] plus the MACCS keys, as implemented in an additional module for use within MOE.[39] This combination had been found useful for a drug classification problem.[40] Initially we also tried the descriptors from MOLGEN-QSPR,[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 known to be highly 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.

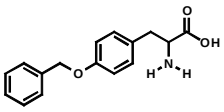
Scheme 1

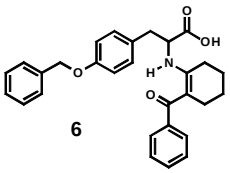

1

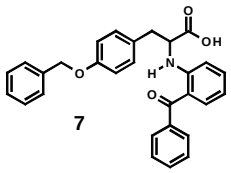

2

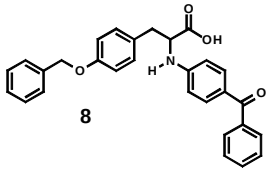

3

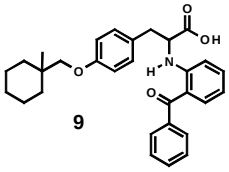

4

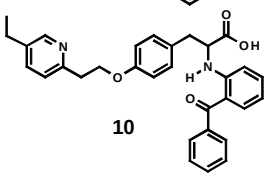

5

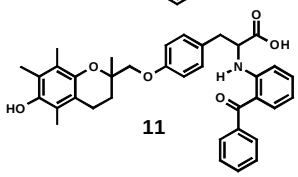

6

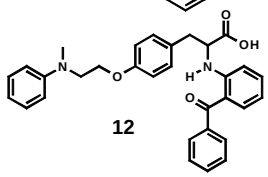

7

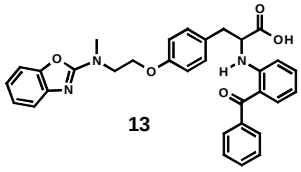

8

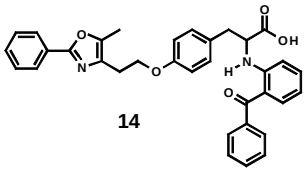

9

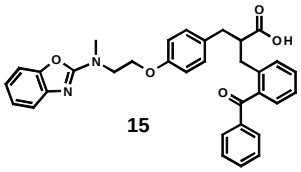

10

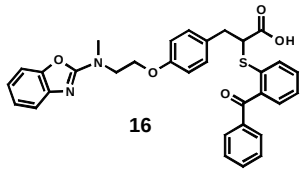

11

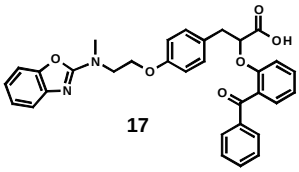

12

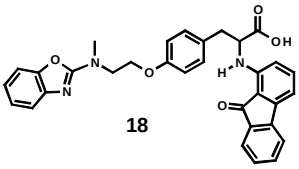

13

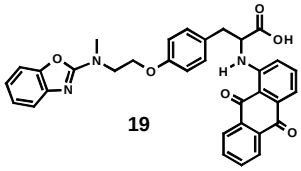

14

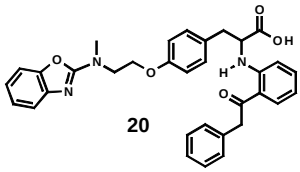

15

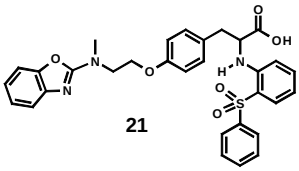

16

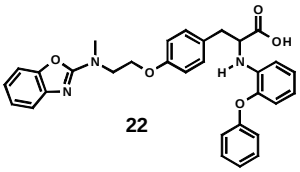

17

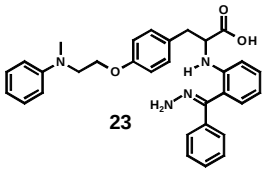

18

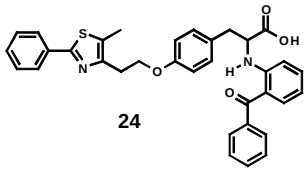

19

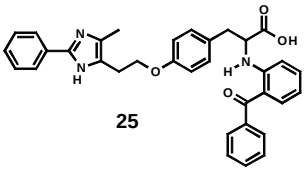

20

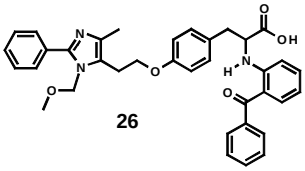

21

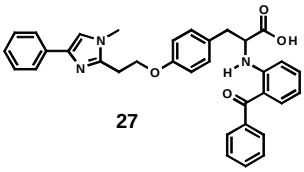

22

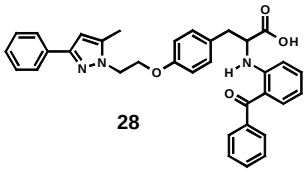

23

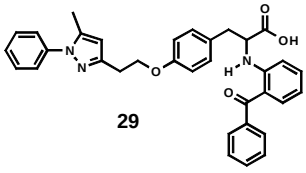

24

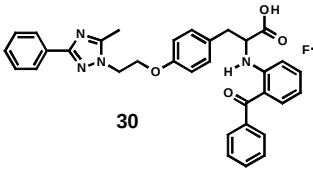

25

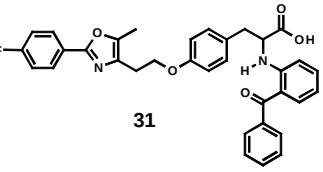

26

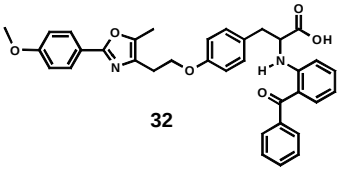

27

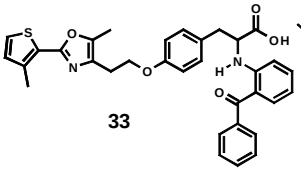

28

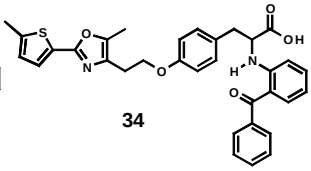

29

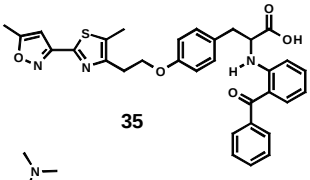

30

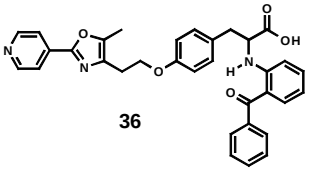

31

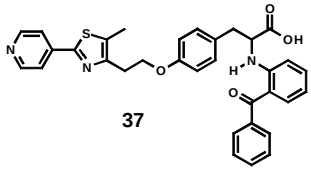

32

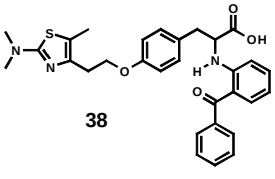

33

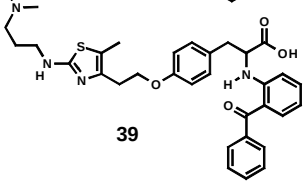

34

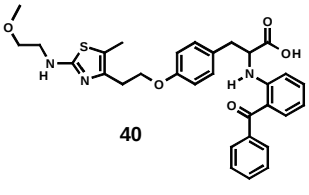

35

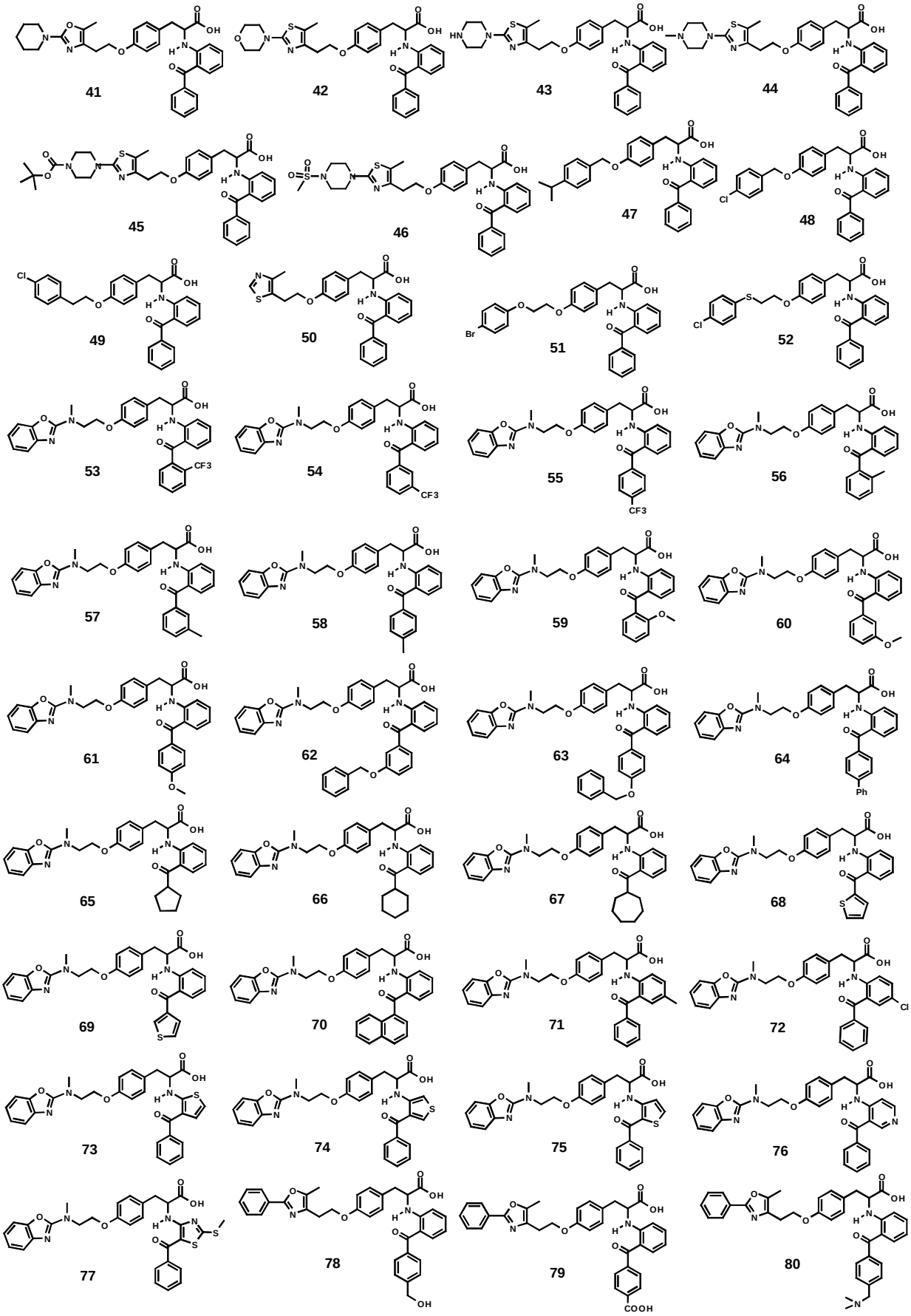

36

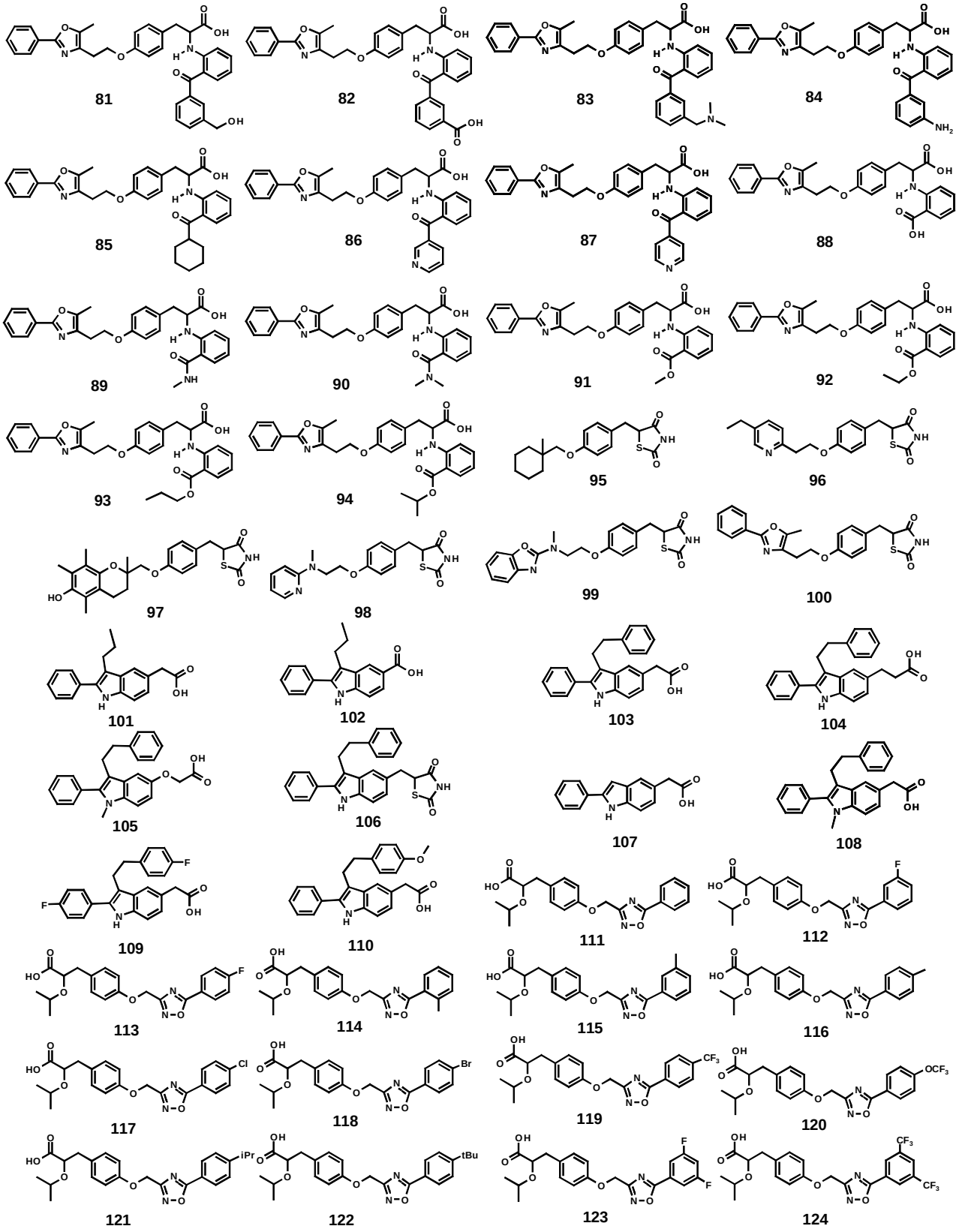

37


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

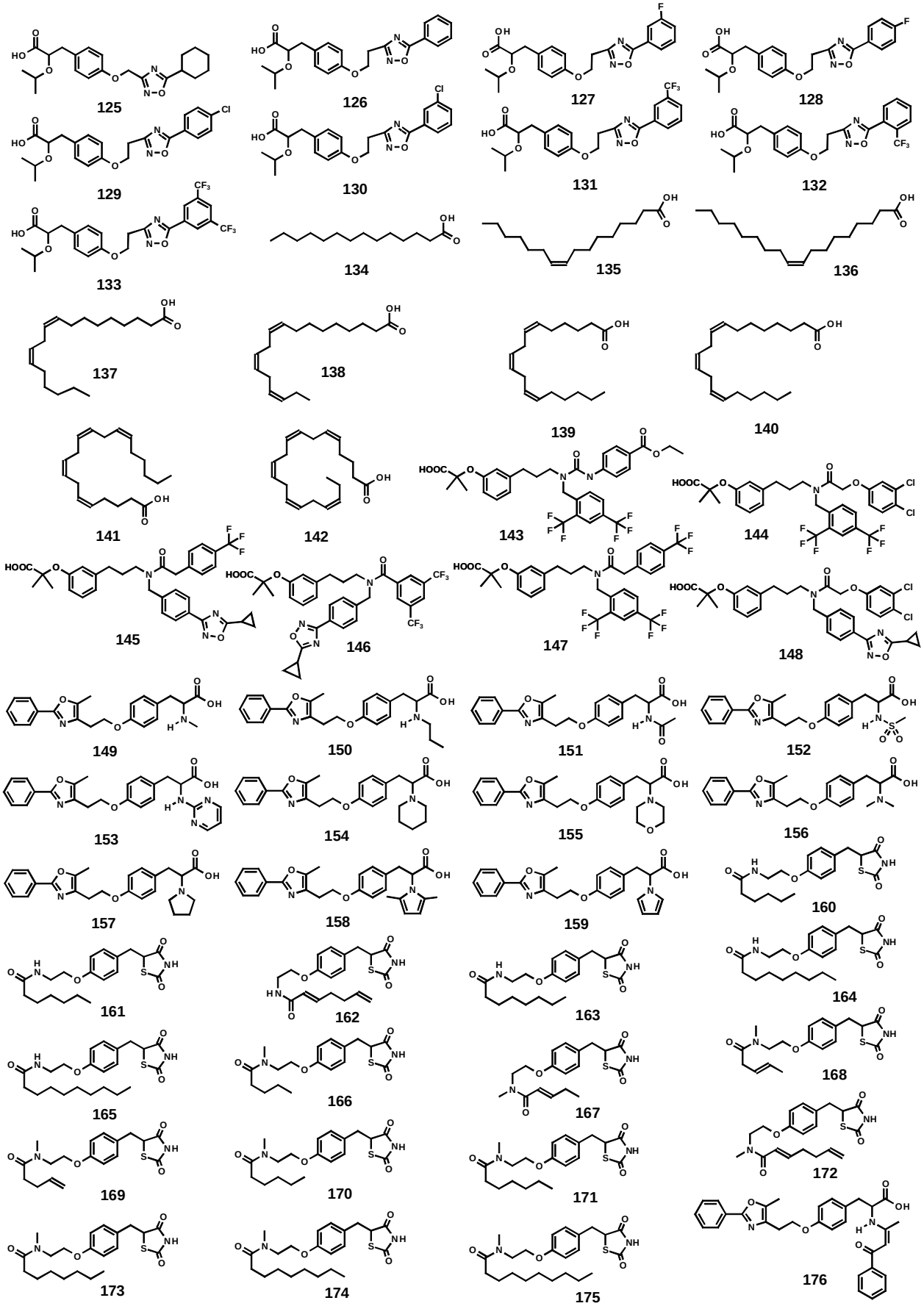

125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176

Table 1. Experimental and Calculated pK_i and pEC_{50} Values of Compounds **1-176**.

ID	pKi exp.	pKi calc. (m1)	pKi calc. (m2)	pEC50 exp.	pEC50 calc. (m3)	pEC50 calc. (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
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	
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	

Descriptor selection. Often one does not know in advance which descriptor or combination of descriptors are relevant for the problem at hand. Commercial statistics packages provide methods for more or less automatically selecting a good descriptor combination. Such procedures select a near-best descriptor combination out of a large pool of descriptors essentially by screening many combinations selected by some heuristic, and always keeping the best one according to a preset criterion. We for this purpose used both a genetic algorithm supplied as an additional module to MOE, and the step-up procedure provided by MOLGEN-QSPR.[42]

RESULTS

A. Binding. The best multilinear regression (MLR) equation we were able to find for PPARγ agonist binding is model m1, made of 10 descriptors selected by the step-up procedure from the pool of 230 descriptors that survived for this sample of 144 compounds. (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 (m1)

$n = 144$, $r^2 = 0.7938$, $s = 0.5822$, $F = 51.20$, $r^2_{cv} = 0.7627$, $s_{cv} = 0.6246$,

where subscript cv denotes quantities obtained by leave-one-out (LOO) crossvalidation. A calculated vs observed-plot is shown in Figure 3, showing both fitted (closed symbols) and LOO-crossvalidated values (open symbols).[43]

Though $r^2 = 0.79$ is not overly comfortable, 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, one may conclude that, if valid, it may be useful. A superficial glance at the statistics is encouraging: The crossvalidated statistics are not much worse than those of fitting. The absolute t values of all descriptors and the intercept in m1 are above 2.6, and F = 51 is far above the critical tabulated value 1.90 (α = 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 $9.36 \cdot 10^{16}$ 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 issue, and the conventional tabulated F values are not relevant. The problem has been known since 1972 at least,[44] it is now termed descriptor selection bias and was rediscussed recently.[45] Essentially, the problem is that given a large number of descriptors, there are probably some combinations that describe a data set relatively well purely by chance, i.e. even if the descriptors consist of random numbers. If many descriptor combinations are screened, and the "best" combination is kept, then the risk is high to keep one of these chance correlations.

Validation. For validation one should have an independent test set of compounds of the same kind as those in the training set and with experimental values available obtained by the same measurement protocol. However, following Hawkins,[46] we had decided to use all available compounds for training, in order to obtain a model built on as diverse structures as possible. Therefore, in the absence of a test set, we spent some effort in additional validation procedures.

Table 2. Full Models[a,b]

pK_i = -24.9625(±3.2665)·VAdjEq +10.1544(±3.7646)·PEOE_RPC-
-0.0777(±0.0176)·bpol -0.0272(±0.0042)·sMR_VSA0
-0.0633(±0.0158)·sMR_VSA3 +0.0168(±0.036)·sMR_VSA6
+1.1317(±0.1898)·MACCS49 +0.7718(±0.1073)·MACCS97
+0.4512(±0.1046)·MACCS116 +0.5177(±0.0998)·MACCS152
+17.0750(±1.5651) (m1)

pK_i = -0.0233(±0.0109)·b_single +0.3635(±0.0789)·slogP
+0.0206(±0.0040)·slogP_VSA3 +0.8444(±0.258)·MACCS49
+0.4500(±0.1103)·MACCS93 +0.8388(±0.0876)·MACCS97
+0.2932(±0.1253)·MACCS132 -0.5876(±0.1294)·MACCS140
-0.2855(±0.1591)·MACCS141 +0.5910(±0.0980)·MACCS152
+3.8594(±0.4163) (m2)

pEC_{50} = 27.4408(±4.1820)·PEOE_VSA_FPPOS +0.5377(±0.0752)·slogP
-0.0208(±0.0057)·slogP_VSA0 +0.0334(±0.0050)·sMR_VSA6
-1.4341(±0.1900)·MACCS22 +0.9581(±0.2781)·MACCS49
-1.2896(±0.3508)·MACCS64 -0.5947(±0.1405)·MACCS80
+0.8876(±0.1679)·MACCS94 +0.3394(±0.1068)·MACCS97
-0.9398(±0.1647)·MACCS106 -0.5726(±0.1911)·MACCS109
+1.6799(±0.4170)·MACCS125 +0.1848(±0.0874)·MACCS137
+2.1372(±0.5284) (m3)

pEC_{50} = 1.0470(±0.0663)·pK_i +0.5246(±0.1461)·PEOE_PC-
+0.6195(±0.1343)·MACCS57 +0.1795(±0.0283)·MACCS62 +0.1969(±0.4695) (m4)

[a]Numbers in parentheses are standard errors.
[b]For explanation of descriptors involved see Table 3.

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
slogP	logP calculated by the atom type contribution method[b]
slogP_VSA0	sum of vdW surface areas of atoms whose contribution to slogP is less or equal to -0.4[b]
slogP_VSA3	sum of vdW surface areas of atoms whose contribution to slogP 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 references 38 and 40.
[b]see Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.

Thus, first we perfomed an additional crossvalidation, i.e. four times leave-one-quarter-out (using the same descriptor combination as in the original model, for details see Katritzky[47]), though this has in principle the same limitations as leave-one-out crossvalidation.[48] The result shown in Table 4 seems reasonable for a real model.

Y-randomization, also called y-scrambling or permutation test, was said to be "probably the most powerful validation procedure".[49] In this method 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 are expected to be of far lower quality than the real model.[50,51] In fact, y-randomization is an approximation of the action of chance. In 25 independent such y-randomization experiments, the mean best r^2 was 0.3025 (min 0.2218, max 0.4193, 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.

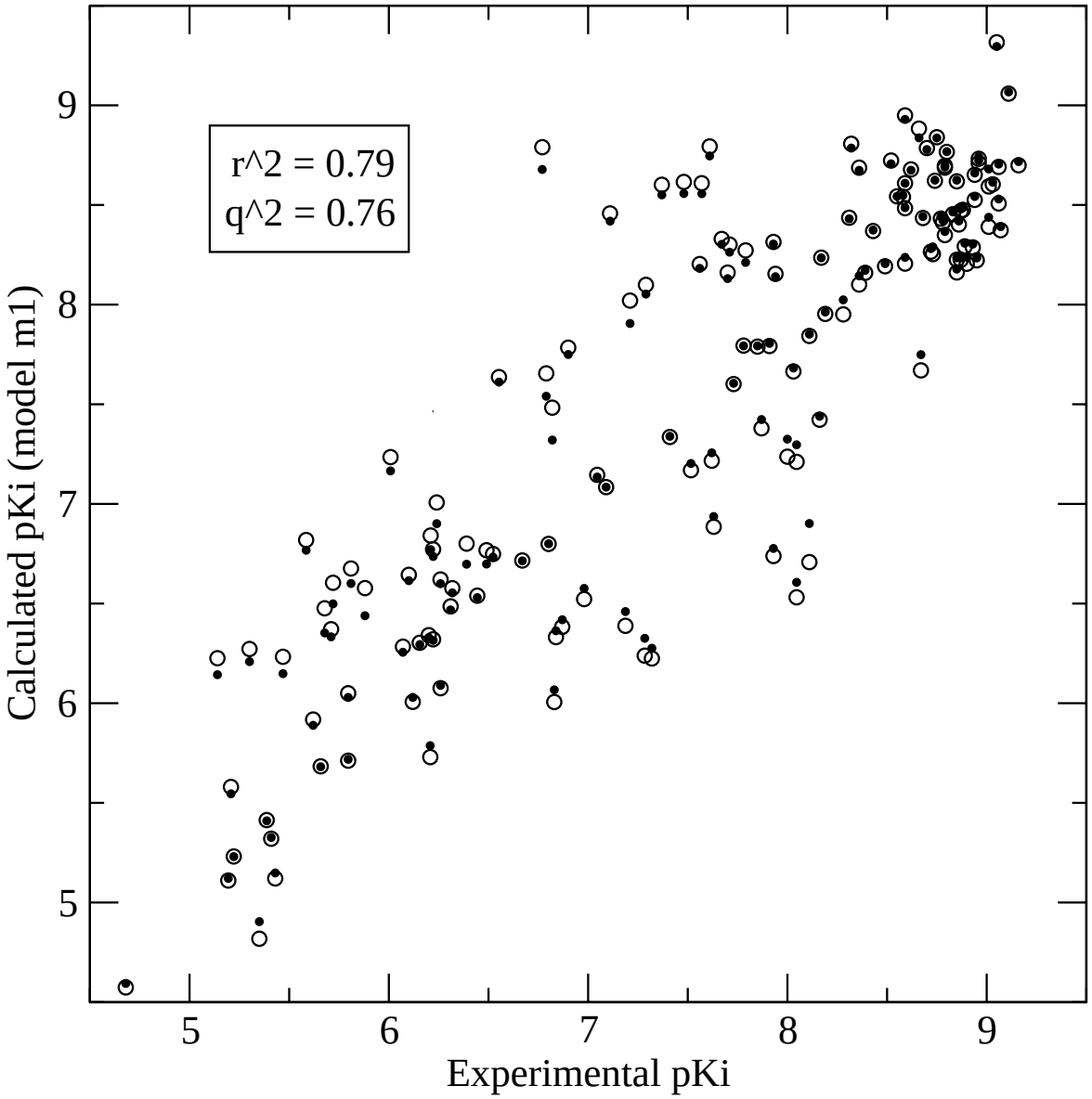


Figure 3. Calculated and observed pK_i values for receptor binding of PPARγ agonists (model m1). Closed symbols represent fit, open symbols represent LOO-crossvalidated values.

In order to strictly judge the statistical significance of model m1 we generated for our 144 compounds the values of 230 pseudo-descriptors 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

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

not expected to arise by chance under these conditions. This is equivalent to rejecting the null hypothesis that all regression coefficients should be zero.

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 does not, as such, render a model useless, since important for multilinear regression is not in what two descriptors agree, but in what they differ.[42a,52] There may be some further multicollinearity in the descriptors that is not easily detected and will, among other things, lead to inflated uncertainty in the regression coefficients. Bootstrapping was performed as a further diagnostic to get an impression of the variability of the regression coefficients and to detect any pathologies in the data.[53] The result of 10^6 runs on bootstrap samples was r^2_{bs} = 0.8067, standard deviation 0.0319. According to reference 53, this value does not point to any problem with model m1. 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%.

Training set/test set partition. The predictive ability of a model can be assessed only from the result of predictions. We therefore randomly partitioned the 144 compounds with pK_i data available into a training set (90%) and a test set (10%). The compounds are naturally partitioned into groups, tyrosine derivatives group 1 **1-23**,[13] group 2 **24-52**,[14] group 3 **53-94**,[15] thiazolidinediones **95-100**,[13] indoles **101-110**,[25] fatty acids **134-142**,[34] tyrosine derivatives bearing a small N-substituent **149-159,**[33]

and thiazolidinedione-fatty acid hybrids **160**-**175**,[35] according to their origin from the references. These groups should be represented in the training and test sets in a balanced manner. Hence 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**, **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.

pK_i: b_single slogP slogP_VSA3 MACCS49 MACCS93 MACCS97 MACCS132 MACCS140 MACCS141 MACCS152 (m2)

n = 129, r^2 = 0.7909, s = 0.5887, F = 44.6, r^2_{cv} = 0.7471, s_{cv} = 0.6475

Three of the ten descriptors in m2 are also in m1. 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.

In the calculated vs observed-plot (Figure 4) 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. Likewise, 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 thus not expected to be a chance correlation.

Application of m2 to the 15 test set compounds resulted in r^2_{pred} = 0.6998. The predicted pK_i values for the test set compounds are underlined in Table 1 and included in Figure 4 (open symbols). Models m1 and m2, as expected, result in similar calculated pK_i values, as given in Table 1 and shown in Figure 5.

Application of models to low-activity compounds. For some low-activity compounds an upper bound of binding affinity to PPARγ (lower bound of K_i, upper bound of pK_i) is given in the source publications. These are tyrosine derivatives **5**, **177**-**181**,[13] and **88**,[15] indole derivatives **182**-**184**,[25]

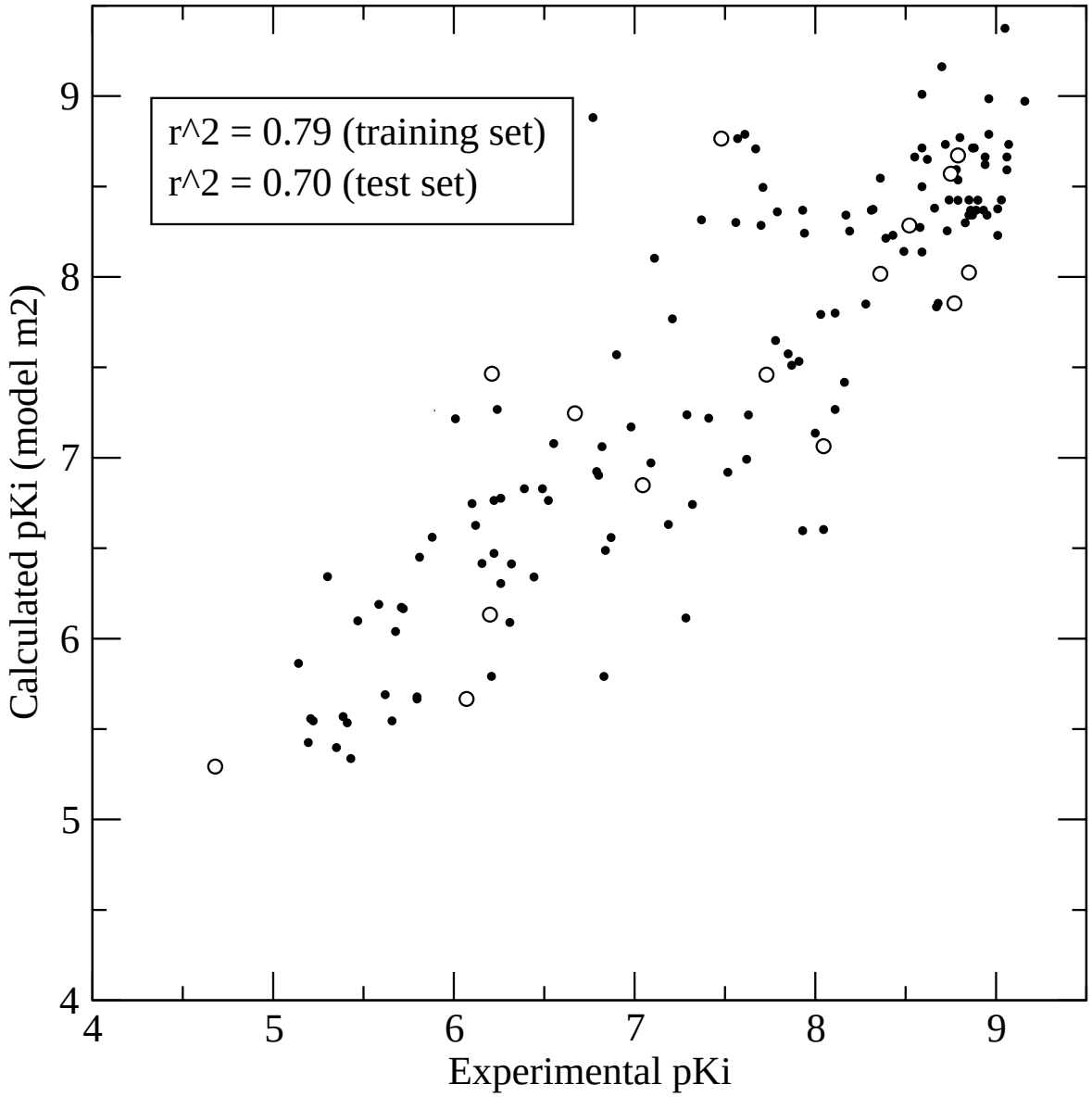
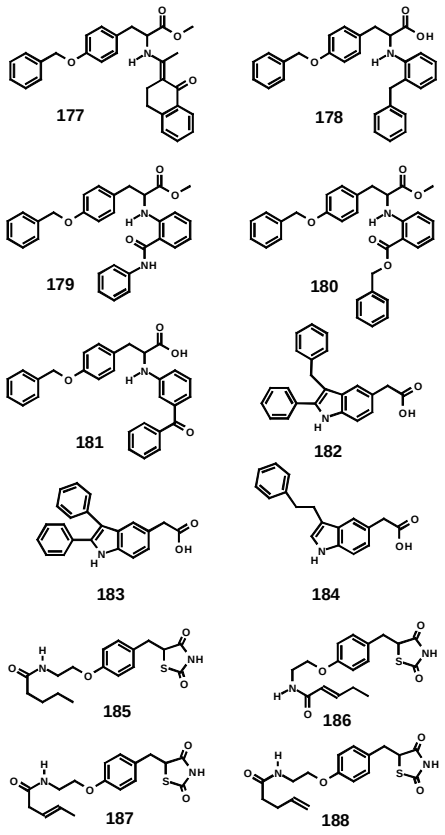


Figure 4. Calculated and observed pK_i values for receptor binding of PPARγ agonists (model m2). Closed symbols represent fit values for the training set, open symbols represent predictions for the test set.

the fatty acids capric, lauric, palmitic, stearic, arachidic, behenic, and erucic acid,[34] and the TZD-fatty acid hybrids **185**-**188**.[35] For these compounds pK_i values were predicted using models m1 and m2, the results are shown in Table 5.



As seen from Table 5, predictions are good for **5**, and reasonable for capric and lauric acids and **184** only. In judging the other more or less incorrect (i.e. too high) predictions we should keep in mind that models m1 and m2 were not built on such low-activity compounds, i.e. all these predictions are extrapolations with respect to activity. As long as the values of descriptors that appear in the model are within the range spanned by the original compound set, models tend to predict activity values also in the range spanned by the original compounds.

Tyrosine derivatives **88** and **177**-**181** are predicted among the 18% lowest active tyrosines in the original compound set by both models. Indoles **182** and **183** are predicted among the lowest-active 60% of the original indoles by m1, and among the lowest 50% by m2. The TZD-fatty acid hybrids **185**-**188** are predicted among the lowest-active 31% (m1) or 38% (m2) of their kind. However, a few fatty acids are wrongly predicted among the most active within their class.

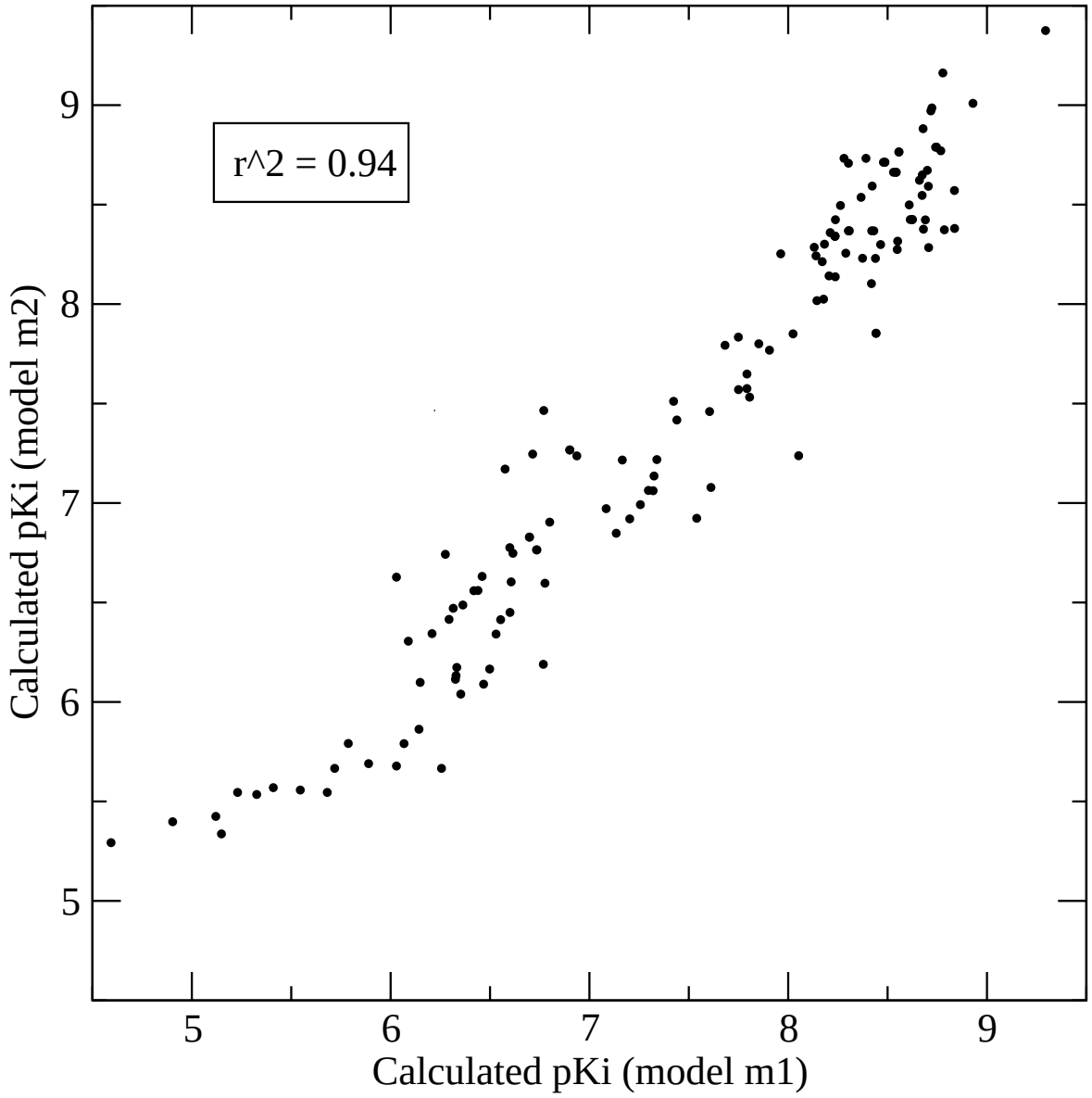


Figure 5. pK_i Values for receptor binding of PPARγ agonists calculated using models m1 and m2.

The explanation usually given for a bad prediction is that a compound is unique in some respect, and as such was not represented in the training set. In fact,

177 is the only compound with a tetralone side chain,

178 is the only example bearing a CH_2 instead of a keto group in the side chain,

179 is the only anilide instead of a ketone (side chain),

180 is the only benzyl ester instead of a ketone (side chain),

Table 5. pK_i Prediction for Low-Activity Compounds.

Compound	pK_{iexptl}	pK_{ipred} (m1)	pK_{ipred} (m2)
5	<5.5	3.47	4.03
88	<5.5	7.38	7.36
177	<5.5	6.13	6.63
178	<5.5	6.17	6.21
179	<5.5	5.91	6.37
180	<5.5	6.20	6.71
181	<5.5	6.33	6.13
182	<4	5.99	5.79
183	<5.0	6.30	5.89
184	<5.0	4.92	5.30
$H_3C(CH_2)_8COOH$ (capric)	<4.5	3.32	5.00
$H_3C(CH_2)_{10}COOH$ (lauric)	<4.5	4.05	5.15
$H_3C(CH_2)_{14}COOH$ (palmitic)	<4.5	4.99	5.44
$H_3C(CH_2)_{16}COOH$ (stearic)	<4.5	5.28	5.58
$H_3C(CH_2)_{18}COOH$ (arachidic)	<4.5	5.47	5.72
$H_3C(CH_2)_{20}COOH$ (behenic)	<4.5	5.60	5.87
$H_3C(CH_2)_7CH{=}CH(CH_2)_{11}COOH$ (erucic)	<4.5	5.74	5.86
185	<5	6.11	6.27
186	<5	5.82	5.97
187	<5	6.24	6.26
188	<5	5.79	6.26

181 is the only example of a *meta*-substituted tyrosine N-phenyl group, and **88** is unique in bearing another carboxylic acid function in the side chain instead of the benzoyl group. Further, **177**, **179**, and **180** are methyl esters rather than carboxylic acids, whereas there is only one methyl ester (**3**) in the training set. Finally, indoles **182** and **183** are the only ones bearing a benzyl or a phenyl group in position 3 of the indole nucleus, rather than the phenethyl group present in most of the other indols. These structural differences easily perceived by a chemist are probably not adequately reflected by the descriptors used. In fact, scrutinizing the descriptor values one finds that compounds **177**-**184** and **88** are numerically within the ranges spanned by compounds **1**-**176** for all descriptors, except that **180** is out-of-range for PEOE_VSA+5, and **177**-**180** are out-of-range for Q_VSA_FPNEG. These two descriptors are not included

in models m1 or m2. Similarly, for compounds **185**-**188** all descriptor values are within the original range, except a few descriptors not appearing in the models.

The fatty acids palmitic acid through erucic acid, being saturated or nearly saturated acyclic compounds, strongly differ in structure from most training set compounds. This is easily seen both by inspection and by their out-of-range values of several descriptors (none of which entered the models). This clearly suggests that it is inadequate to predict these compounds' activities based on models m1 or m2.

B. Transactivation. The best multilinear regression equation we were able to find for gene activation by PPARγ activated by agonists (pEC_{50} values) is model m3, made of 14 descriptors selected by the step-up procedure from a pool of 229 descriptors:

pEC_{50}: PEOE_VSA_FPPOS slogP slogP_VSA0 sMR_VSA6 MACCS22
MACCS49 MACCS64 MACCS80 MACCS94 MACCS97 MACCS106
MACCS109 MACCS125 MACCS137 (m3)

n = 150, r^2 = 0.6487, s = 0.7335, F = 17.80, r^2_{cv} = 0.5727, s_{cv} = 0.8089.

A calculated vs observed-plot is given in Figure 6 for both fit and LOO-crossvalidated data (closed and open symbols, respectively).

Note that of the 14 descriptors appearing in m3 three are also in m1, and three are also in m2. This seems to be more than coincidence: Binding is reasonably considered a prerequisite for transactivation. All pairwise descriptor intercorrelations in m3 have $r^2 < 0.35$.

Though $r^2 = 0.65$ is at the lower limit of what may be considered useful, m3 has s = 0.73 log units, appreciably smaller than the standard deviation of the experimental data, 1.18 log units. The differences between the crossvalidated and fitted statistics are not excessively large. The absolute t values of all descriptors and the intercept in m3 are above 2.1, and F = 17.80 is far above the tabulated critical value 1.77 (α = 5%) for 150 data points and 14 descriptors. However, the descriptor combination in m3 was selected from in principle $8.34 \cdot 10^{21}$ combinations of 14 out of 229 descriptors, and the model quality as described by r^2, r^2_{cv} and F is lower than that of m1 or m2. The possibility of chance correlation therefore here seems to be worse than before.

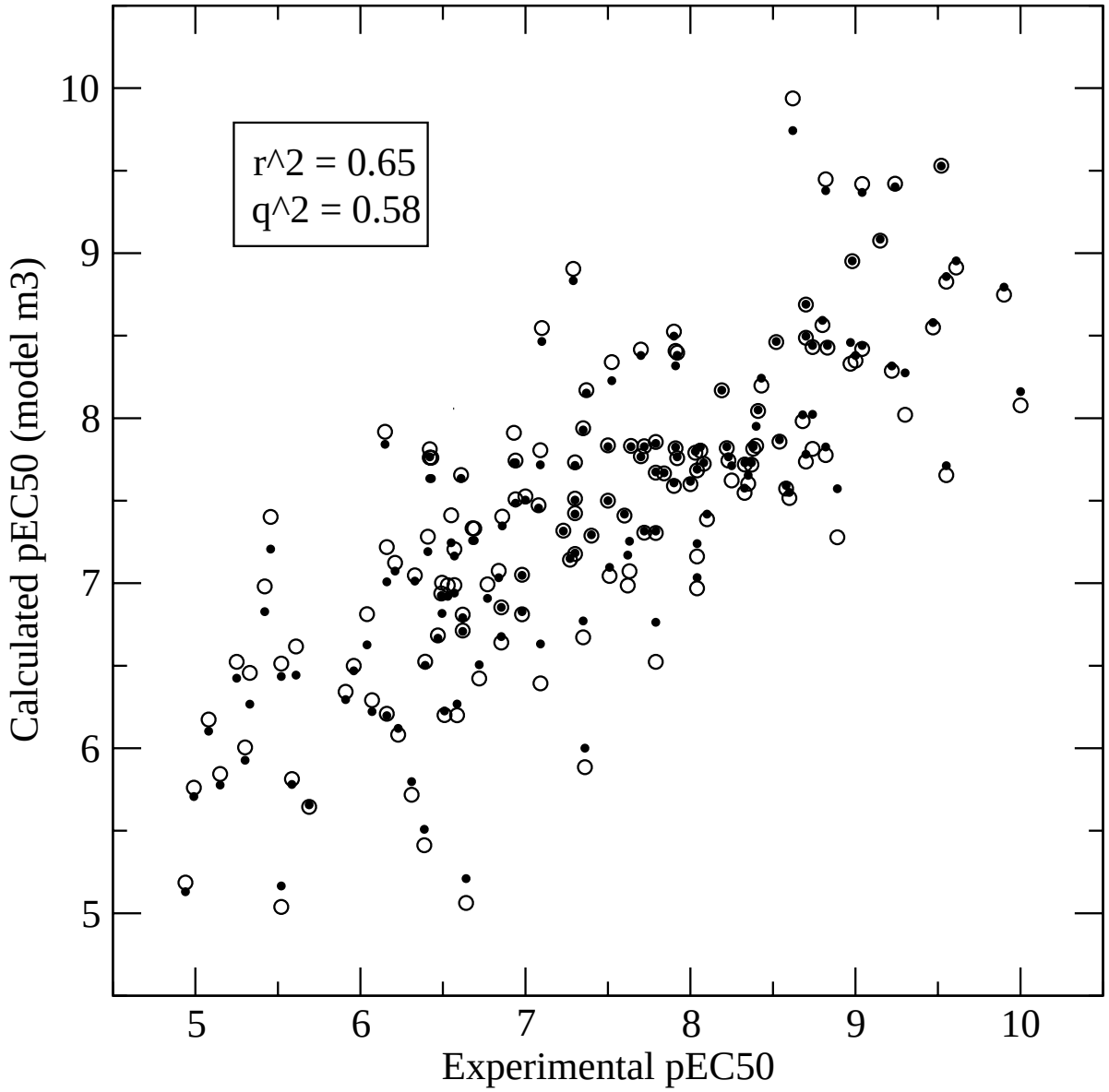


Figure 6. Calculated and observed pEC_{50} values for transactivation by PPARγ agonists (model m3). Closed symbols represent fit, open symbols represent LOO-crossvalidated values.

Validation. The result of an additional crossvalidation, i.e. four times leave-one-quarter-out, is shown in Table 6.

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

Table 6. 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

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 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. According to reference 53, this value does not point to any problem with model m3. 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 demonstrate that m3, notwithstanding its lower quality compared to m1, is still statistically valid.

Transactivation using binding activity as a descriptor. 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 only that have a pK_i available. Further, it should be kept in mind that one of the assumptions for applicability of linear regression is violated here, the assumption of negligible errors in the independent variables.

The best four-descriptor model we found is m4:

pEC_{50}: pK_i PEOE_PC- MACCS57 MACCS62 (m4)

n = 118, r^2 = 0.7618, s = 0.6087, F = 90.34, r^2_{cv} = 0.7385, s_{cv} = 0.6378.

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

DISCUSSION

Our models were subjected to (and passed) more validity tests than are usually performed, because we considered such tests necessary in view of the multitude of problems encountered. In our opinion, these problems are typical for many QSAR studies, and therefore we discuss them here in some detail.

Experimental data. In many QSAR studies one has to rely on a single given set of data, with no information on their quality available. The present study is an exception in that luckily we have some evidence on data quality (Figures 1 and 2). This evidence unfortunately points to low data quality. Dealing with biological phenomena, one probably cannot expect highly reproducible data. To enhance reproducibility and comparability, we included only data obtained under one and the same measurement protocol. On the other hand, the receptor binding data span molar concentrations differing by 4.5 orders of magnitude, and the transactivation data even span 5 orders of magnitude in molar concentrations. This alone leads us to expect a considerable scatter in the data. Who would expect reliable data of lengths ranging from one micrometer to ten centimeters, all measured *with the same instrument*?

In the case of transactivation, there is another fundamental problem. An EC_{50} value as published is defined as the concentration at which the respective compound produces 50% of maximal gene transactivation. However, the maximal gene transactivation is not usually defined unambiguously. In some publications it is the maximal activation that can be observed using the respective compound, in others it is the maximal effect seen with some standard compound, e. g. rosiglitazone (**98**). Unless standardized, EC_{50} values do not describe concentrations resulting in the same effect, they thus compare uncomparable things. This problem may be one of several factors contributing to the difficulties we and others had

in attempting to correlate pEC_{50} values with structure.[11,20,21] One may even question that a QSAR treatment of such data makes any sense at all.

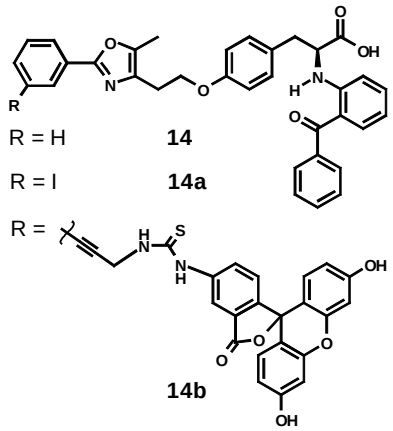
Compound integrity. In most cases it is not assured that the measured effect is due to the agonist in the form applied, as opposed to some derivative or metabolite. For example, it is known that methyl esters are hydrolysed under physiological conditions.

An additional source of uncertainty in experimental data is stereochemistry. Many but not all of the agonists treated here are chiral, e.g. thiazolidinediones and tyrosine derivatives. Thiazolidinediones racemize under physiological conditions, e.g., for rosiglitazone racemization $t_{1/2}$ at pH 7.2 was found to be 3 h.[36] Consequently, there is no difference between individual enantiomers or the racemate in antidiabetic activity or in gene activation, effects that require days or >12 h, respectively, to be measured. Measurement of receptor binding, on the other hand, requires less time, and consequently IC_{50} for PPARγ binding of (*S*)-rosiglitazone was found 70 times lower than that of the *R* enantiomer.[36] For tyrosine derivatives or their oxygen analogs racemization generally is not expected. In fact, the *S* enantiomer of an α-benzyloxy-phenylpropanoic acid was found 100-fold more potent in vivo than its *R* isomer. However, surprisingly for the corresponding α-methoxy acid an as yet unexplained unidirectional isomerization of *R* into *S* was observed in a cell-based assay, and no enantiomer differentiation in vivo. Again, for the same compound in a PPARγ binding experiment IC_{50} of the *R* enantiomer was 20-fold higher than that of the *S* isomer.[37] Thus both TZDs and tyrosine derivatives are more active in *S* configuration. This is in accord with X-ray results of rosiglitazone/PPARγ complexes, in which rosiglitazone was found to have *S* configuration.[16,17]
These experimental findings suggest not to stereocorrect racemate data obtained under racemizing conditions, or even to up-correct activity data obtained for pure *S* compounds measured under racemizing conditions. However, for most compounds it is not known whether or not (or how rapidly) racemization occurs under the particular experimental conditions.

Combining observations. In statistical data treatment, it is always problematic to combine observations on heterogenous samples, since a

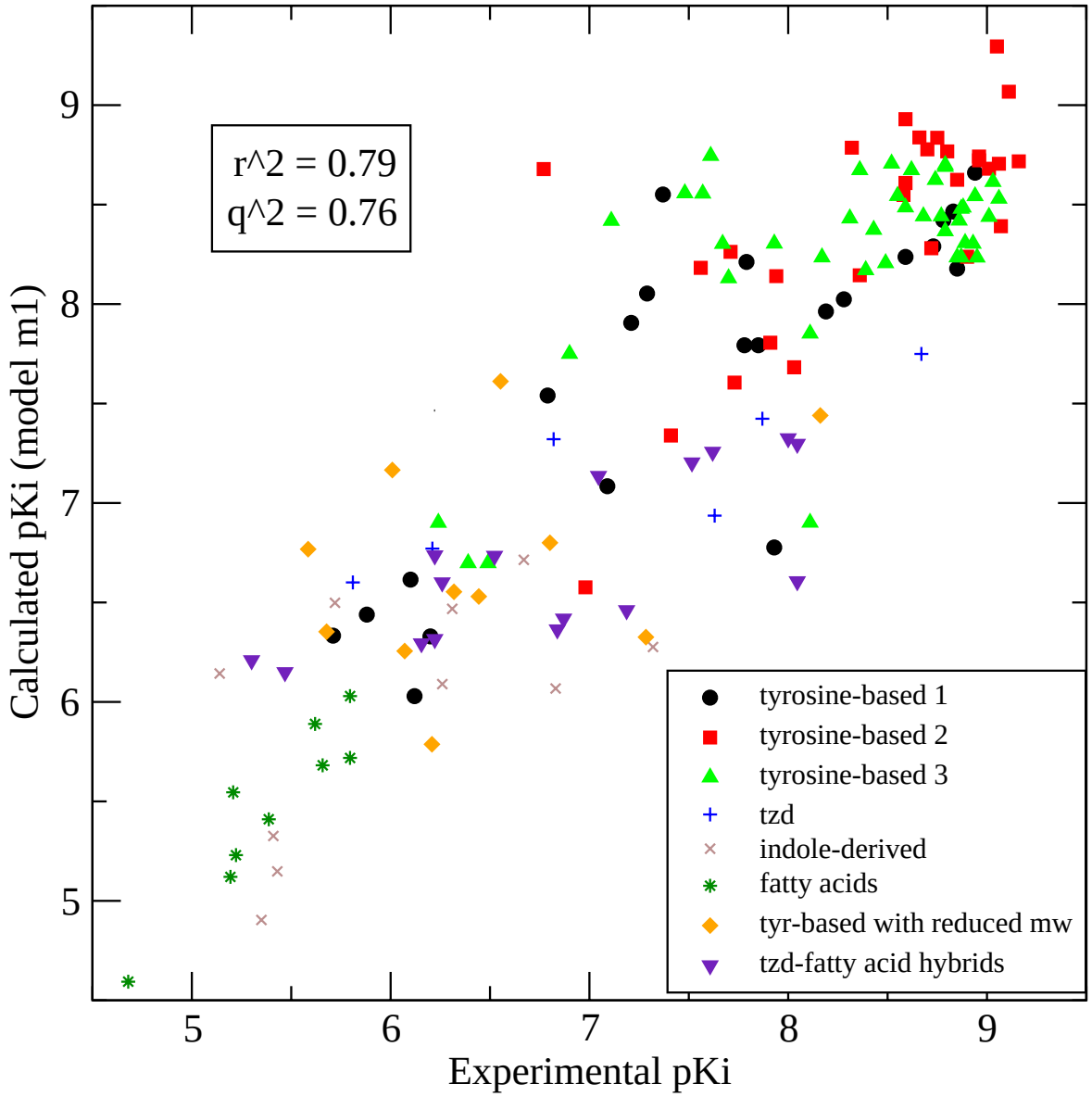
correlation found for the combined sample may vanish within subgroups, or trends present in the subgroups may be obscured or even reversed in the combined sample. The compounds in our study form structurally distinct subgroups, such as tyrosine derivatives, TZDs, indoles, fatty acids, etc.. Therefore in Figure 7a the data for model m1 (the fit part of Figure 3) are displayed once more, this time resolved with respect to subgroups. Figure 7b shows corresponding plots for the subgroups. It is obvious that the only subgroup covering most of the activity range is tyrosine derivatives group 1. The most important result of this analysis is reassuring: Though the r^2 value for the complete data set is certainly overoptimistic with respect to the subgroups, at least in all subgroups the same trend as in the combined sample is apparent. One may say that 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 described as "carboxylic acid or thiazolidinedione group linked to flexibly interconnected unsaturated (aromatic) moieties".

Descriptors. All the available molecular descriptors take into consideration the whole molecule. As such they are very useful for describing simple physical properties that depend on the whole molecule, such as boiling point, solubility, logP, etc. For the same reason such descriptors are not well-suited to describe phenomena that depend on specific (but unknown) parts of the molecule, such as binding to a receptor. In other words, a large part of a molecule may be completely irrelevant with respect to such a phenomenon. For illustration consider the work of DeGrazia,[54] in which a fluorescent derivative of farglitazar **14** was taylored in such a manner as not to compromise binding to the receptor. This was achieved by linking the fluorescein moiety covalently to farglitazar in a position known from the X-ray structure of the **14**/ PPARγ complex to point outside the binding pocket. The resulting compound **14b** had a K_d value of 61 nM, very similar to the K_i value of intermediate iodide **14a** (50 nM), whereas most molecular descriptors will be largely different for **14b** and **14a**.



Descriptor selection and risk of chance correlation. Livingstone and Salt[45a] extensively simulated multilinear regressions using random number dependent and independent variables for many combinations of numbers of compounds (n), numbers of descriptors in a QSAR equation (p), and number of descriptors considered for inclusion in a model (k). They fitted their simulation results by a highly nonlinear equation for calculation of a critical F value. This equation, however, is of no use in our case, since near the margins of the (n,p,k)space covered in reference 45a and outside that range errors may be high, and the characteristics of our models (n = 144, p = 10, k = 230 (binding), and n = 150, p = 14, k = 229 (transactivation)) are far outside that range. Therefore we had to perform our own experiments using nonsense (random) descriptors as described above.

The risk of chance correlation due to descriptor selection in QSAR was pointed out 33 years ago by Topliss et al.[44a,b] These papers were ahead of their time: Before 1979 most of the molecular descriptors now available, computer programs for generating descriptor values, and programs for descriptor selection were nonexisting, so that the problem then was a theoretical one. Now it has become urgent, as demonstrated by several publications in which descriptors are naively selected from a pool several times as large as the number of compounds. Since the size of the final descriptor pool is seldom reported in such studies, the risk of a chance correlation in most cases cannot be assessed by the reader.


r^2 = 0.79
q^2 = 0.76
tyrosine-based 1
tyrosine-based 2
tyrosine-based 3
tzd
indole-derived
fatty acids
tyr-based with reduced mw
tzd-fatty acid hybrids
Calculated pKi (model m1)
Experimental pKi
5
6
7
8
9

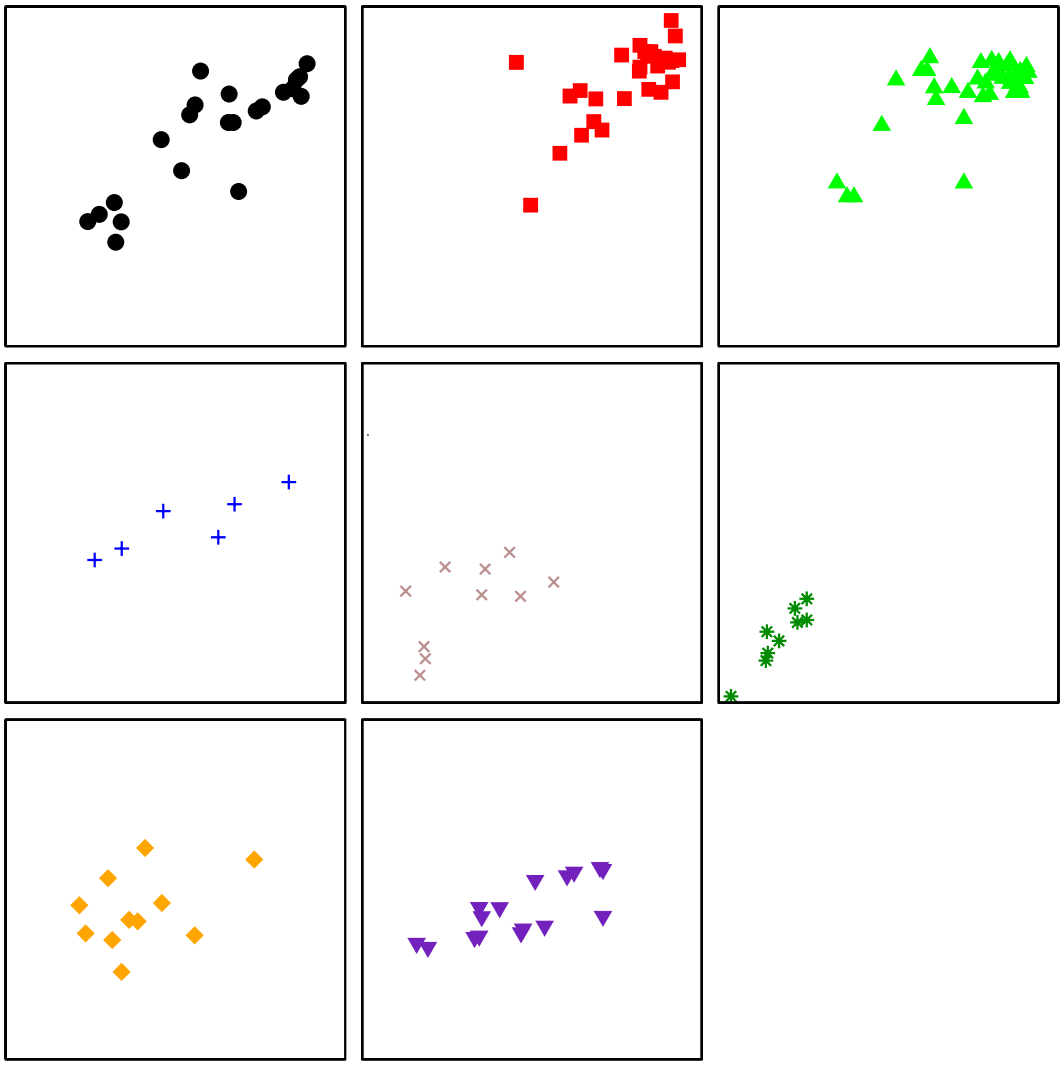


Figure 7. Fit data from Figure 3, broken down by subgroups. Circles – tyrosine derivatives group 1; squares – tyrosine derivatives group 2; triangles – tyrosine derivatives group 3; crosses – thiazolidinediones; oblique crosses – indole derivatives; asterisks – fatty acids; diamonds – tyrosine derivatives of low molecular weight; reversed triangles – thiazolidinedione-fatty acid hybrids.

In our opinion, the quote by Estrada "*It is desirable to have as many as possible molecular descriptors to characterize molecular structure but to include as few as possible into the QSAR/QSPR model*"[55] has to be amended. To have a large number of descriptors at our disposition is not simply a blessing, rather we should use them very cautiously. There is a real dilemma. In order not to fall victim of chance correlations we should include as few descriptors as possible in the pool to be considered by a selection algorithm. These, of course, should be those

that have a high probability of being useful, while at the same time we must not exclude the really relevant ones. That means we preferably should have some (hypothetical) knowledge of the solution before algorithmically treating a particular problem. This situation, however, is common in science: We can test hypotheses, but without a hypothesis we should not expect a good (or even the best) solution to result out of nothing. Obviously, lacking a hypothesis on the relevance of descriptors, critical validation of models is the more imperative.

The descriptor selection bias problem, though numerically considered mainly in the context of multilinear regression,[44,45] is expected to occur generally whenever the optimum combination (for a certain purpose) of descriptors out of a large descriptor pool is selected, e.g. in the application of classification and regression trees (recursive partitioning), as well.

Domain of applicability. The foremost purpose of a QSAR model is prediction. For this to work reliably, it is important to define the model's domain of applicability.[56,57] That range obviously can span compounds only that are similar to those used for modeling both structurally and with respect to target activity. We included in our models the broadest possible selection of structures, i.e. all compounds with appropriate experimental values available, so that our compounds are more diverse than those treated previously. Nevertheless their diversity is far lower than one would like it to be. On the other hand, we should keep in mind the diversity/predictivity antagonism: The broader the structure range, the lower the quality of prediction. So when applying a model to predict a new compound's activity, the chemist has to carefully judge the structural similarity, which is not always an easy task. There is a permanent temptation to use a model for prediction outside its range of applicability, e.g. trying to find in silico a new class of compounds with desired activity. Unless taken merely as a heuristic, this would be a systematic misuse of the model.

There is another problem associated with the range of applicability. Often in papers reporting biological activities, for some compounds an upper bound of activity is given, either because the biological test was unable to produce meaningful results above a certain concentration, or because the authors were not interested in compounds exhibiting low activity. In the absence of a real test set, it is tempting to apply a

model to such compounds, in the hope that the model will place them below, or at least in the lowest part of the range of activity considered. While with our models this worked reasonably, often it does not, for simple reasons. First, whereas these compounds may seem similar to those used for modeling with respect to structure, they are by definition not similar with respect to activity. The model is not calibrated, and consequently cannot be used for an activity range well below that of the compounds included. Second, there are cases where the out-of-range activity (i.e. the inactivity) of a compound cannot be explained by the values of some descriptors being out of range. In such cases, to maintain the central dogma of QSPR/QSAR (similar structures result in similar activities) we have to conclude that these compounds' structures differ from those included in the model in a more subtle manner, e.g. in that they may be out-of-range with respect to a *combination* of descriptors, or with respect to a descriptor not considered.

From this it follows that for many cases of wrong prediction an "excuse" can be found in the uniqueness of the respective structure: Each chemical entity is unique in some respect, and it is common practice to argue (as we did above) that a particular mispredicted compound's pecularity was not represented in the training set. This is, however, a weak argument in that it could be applied to many well-predicted compounds also, and even worse, it will apply to most compounds encountered in the future, as well.

Comparison between typical QSPR and typical QSAR situations. Most of the problems discussed above that are encountered in typical QSAR studies are of far less importance in QSPR. Thus, in QSPR experimental data are often more reliable, in that measurements are less difficult, and often values for the same compound measured by several independent research groups are available. Stereochemistry and flexible conformations are less of a problem in physical properties than in specific receptor-ligand binding. Physical properties are determined by the entirety of a molecular structure, whence whole-molecule descriptors are adequate for most QSPR problems. The fundamental assumption of linearity, risky as it is in MLR-QSPR, is highly dangerous in QSAR for several reasons: First, target variables in QSAR, unlike in QSPR, typically vary over several orders of magnitude. Second, biological effects, in particular those

measured in intact cells or even whole organisms, are often regulated by complex nonlinear mechanisms (regulatory circuits), so that we can expect an optimum value to exist for many descriptors. As a result of all this, r^2 and F values are typically higher in QSPR than in QSAR, even if models contain fewer descriptors, so that descriptor intercorrelation and descriptor selection bias (chance correlation) are less of a problem in QSPR. Finally, a model's range of applicability is often easily defined in QSPR, where the input compounds are selected according to structural criteria, whereas in QSAR the compounds included are defined by some biological activity and thus may represent various structural classes.

ACKNOWLEDGMENT

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) (a) Willson, T. M.; Brown, P. J.; Sternbach, D. D.; Henke, B. R. The PPARs: From Orphan Receptors to Drug Discovery. *J. Med. Chem.* **2000**, *43*, 527-550. (b) Willson, T. M.; Lambert, M. H.; Kliewer, S. A. Peroxisome Proliferator-Activated Receptor γ and Metabolic Disease. *Ann. Rev. Biochem*. **2001**, *70*, 341-367.

(2) Rangwala, S. M.; Lazar, M. A. Peroxisome Proliferator-Activated Receptor γ in Diabetes and Metabolism. *Trends Pharmacol. Sci.* **2004**, *25*, 331-336.

(3) Henke, B. R. Peroxisome Proliferator-Activated Receptor α/γ Dual Agonists for the Treatment of Type 2 Diabetes. *J. Med. Chem.* **2004**, *47*, 4118-4127.

(4) Wang, M.; Tafuri, S. Modulation of PPARγ Activity with Pharmaceutical Agents: Treatment of Insulin Resistance and Atherosclerosis. *J. Cell. Biochem*. **2003**, *89*, 38-47.

(5) Kallenberger, B. C.; Love, J. D.; Chatterjee, V. K. K.; Schwabe, J. W. R. A Dynamic Mechanism of Nuclear Receptor Activation and its Perturbation in a Human Disease. *Nature Struct. Biol.* **2003**, *10*, 136-140.

(6) Ferry, G.; Bruneau, V.; Beauverger, P.; Goussard, M.; Rodriguez, M.; Lamamy, V.; Dromaint, S.; Canet, E.; Galizzi, J.-P.; Boutin, J. A. Binding of Prostaglandins to Human PPARγ: Tool Assessment and New Natural Ligands. *Eur. J. Pharmacol.* **2001**, *417*, 77-89.

(7) Schopfer, F. J.; Lin, Y.; Baker, P. R. S.; Cui, T.; Garcia-Barrio, M.; Zhang, J.; Chen, K.; Chen, Y. E.; Freeman, B. A. Nitrolinoleic Acid: An Endogenous Peroxisome Proliferator-Activated Receptor γ Ligand. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2340-2345.

(8) Kulkarni, S. S.; Gediya, L. K.; Kulkarni, V. M. Three-Dimensional Quantitative Structure Activity Relationships (3-D-QSAR) of Antihyperglycemic Agents. *Bioorg. Med. Chem.* **1999**, 7, 1475-1485.

(9) Rathi, L.; Kashaw, S. K.; Dixit, A.; Pandey, G.; Saxena, A. K. Pharmacophore Identification and 3D-QSAR Studies in N-(2-Benzoylphenyl)-L-tyrosines as PPARγ Agonists. *Bioorg. Med. Chem.* **2004**, *12*, 63-69.

(10) Liao, C.; Xie, A.; Shi, L.; Zhou, J.; Lu, X. Eigenvalue Analysis of Peroxisome Proliferator-Activated Receptor γ Agonists. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 230-238.

(11) Liao, C.; Xie, A.; Zhou, J; Shi, L.; Li, Z.; Lu, X. 3D QSAR Studies on Peroxisome Proliferator-Activated Receptor γ Agonists Using CoMFA and CoMSIA. *J. Mol. Model.* **2004**, *10*, 165-177.

(12) Hyun, K. H.; Lee, D. Y.; Lee, B.-S.; Kim, C. K. Receptor-based 3D QSAR Studies on PPARγ Agonists Using CoMFA and CoMSIA Approaches. *QSAR Comb. Sci.* **2004**, *23*, 637-649.

(13) 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. N-(2-Benzoylphenyl)-L-tyrosine PPARγ Agonists. 1. Discovery of a Novel Series of Potent Antihyperglycemic and Antihyperlipidemic Agents. *J. Med. Chem.* **1998**, *41*, 5020-5036.

(14) 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. N-(2-Benzoylphenyl)-L-tyrosine PPARγ Agonists. 2. Structure-Activity Relationship and Optimization of the Phenyl Alkyl Ether Moiety. *J. Med. Chem.* **1998**, *41*, 5037-5054.

(15) 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. N-(2-Benzoylphenyl)-L-tyrosine PPARγ Agonists. 3. Structure-Activity Relationship and Optimization of the N-Aryl Substituent. *J. Med. Chem.* **1998**, *41*, 5055-5069.

(16) 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. Ligand Binding and Co-Activator Assembly of the Peroxisome Proliferator-Activated Receptor-γ. *Nature* **1998**, *395*, 137-143.

(17) 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. Asymmetry in the PPARγ/RXRα Crystal Structure Reveals the Molecular Basis

of Heterodimerization among Nuclear Receptors. *Mol. Cell* **2000**, *5*, 545-555.

(18) Khanna, S.; Sobhia, M. E.; Bharatam, P. V. Additivity of Molecular Fields: CoMFA Study on Dual Activators of PPARα and PPARγ. *J. Med. Chem.* **2005**, *48*, 3015-3025.

(19) Xu, X.; Cheng, F.; Shen, J.; Luo, X.; Chen, L.; Yue, L.; Du, Y.; Ye, F.; Jiang, S.; Zhu, D.; Jiang, H.; Chen, K. Agonist-PPARγ Interactions: Molecular Modeling Study with Docking Approach. *Int. J. Quantum Chem.* **2003**, *93*, 405-410.

(20) Soni, L. K.; Gupta, A. K.; Kaskhedikar, S. G. 2D-QSAR Analysis of Oxadiazole Substituted α–Isopropoxyphenylpropionic Acids as PPAR-α & PPAR-γ Agonists. *E-J. Chem.* **2004**, *1*, 170-177; http://www.websamba.com/ejchem/cas%20issue%203/170-177.pdf.

(21) Hemalatha, R.; Soni, L. K.; Gupta, A. K.; Kaskhedikar, S. G. QSAR Analysis of 5-Substituted 2-Benzoylaminobenzoic Acids as PPAR Modulator. *E-J. Chem.* **2004**, *1*, 243-250; http://www.websamba.com/ejchem/5%20issue/243-250.pdf.

(22) Yu, C.; Chen, L.; Luo, H.; Chen, J.; Cheng, F.; Gui, C.; Zhang, R.; Shen, J.; Chen, K.; Jiang, H.; Shen, X. Binding Analyses Between Human PPARγ-LBD and Ligands. Surface Plasmon Resonance Biosensor Assay Correlating with Circular Dichroic Spectroscopy Determination and Molecular Docking. *Eur. J. Biochem.* **2004**, *271*, 386-397.

(23) Nichols, J. S.; Parks, D. J.; Consler, T. G.; Blanchard, S. G.; Development of a Scintillation Proximity Assay for Peroxisome Proliferator-Activated Receptor γ Ligand Binding Domain. *Anal. Biochem.* **1998**, *257*, 112-119.

(24) Binggeli, A.; Boehringer, M.; Grether, U.; Hilpert, H.; Maerki, H.-P.; Meyer, M.; Mohr, P.; Ricklin, F. Carboxylic Acid Substituted Oxazole Derivatives for Use as PPAR-alpha and –gamma Activators in the Treatment of Diabetes. US Patent 6642389 (2001), priority date May 15, 2001; CA 137:370079.

(25) 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. Synthesis and Biological Activity of a Novel Series of Indole-Derived PPARγ Agonists. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 3329-3334.

(26) Davis, R. G.; Anderegg, R. J.; Blanchard, S. G. Iterative Size-Exclusion Chromatography Coupled with Liquid Chromatography Mass Spectrometry to Enrich and Identify Tight-Binding Ligands from Complex Mixtures. *Tetrahedron* **1999**, *55*, 11653-11667.
(27) 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. Structural Determinants of Ligand Binding Selectivity between the Peroxisome Proliferator-Activated Receptors. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 13919-13924.
(28) Yanagisawa, H.; Takamura, M.; Yamada, E.; Fujita, S.; Fujiwara, T.; Yachi, M.; Isobe, A.; Hagisawa, Y. Novel Oximes Having 5-Benzyl-2,4-thiazolidinedione as Antihyperglycemic Agents: Synthesis and Structure-Activity Relationship. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 373-375.
(29) 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. Substituted 4-Hydroxy-phenylalcanoic Acid Derivatives with Agonist Activity to PPAR-gamma. US Patent 6294580 (2001), priority date Feb 28, 1996; CA 127:278064.
(30) Lehmann, J. M.; Moore, L. B.; Smith-Oliver, T. A.; Wilkison, W. O.; Willson, T. M.; Kliewer, S. A. An Antidiabetic Thiazolidinedione Is a High Affinity Ligand for Peroxisome Proliferator-Activated Receptor γ (PPARγ). *J. Biol. Chem.* **1995**, *270*, 12953-12956.
(31) 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. Identification of a Series of Oxadiazole-Substituted α-Isopropoxy Phenylpropanoic Acids with Activity on PPARα, PPARγ, and PPARδ. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2385-2388.
(32) (a) 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. Identification of a Series of PPARγ/δ Dual Agonists Via Solid-Phase Parallel Synthesis. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 2959-2962. (b) 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.).
(33) 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. Synthesis and Biological Activity of L-Tyrosine-Based

PPARγ Agonists with Reduced Molecular Weight. *Bioorg. Med. Chem. Lett.* **2001**, *11*, 3111-3113.

(34) (a) 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. Molecular Recognition of Fatty Acids by Peroxisome Proliferator-Activated Receptors. *Mol. Cell* **1999**, *3*, 397-403. (b) In this publication, numerical values for binding are given as IC_{50}. Under the measurement conditions the difference between pIC_{50} and pKi 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.

(35) Tomkinson, N. C. O.; Sefler, A. M.; Plunket, K. D.; Blanchard, S. G.; Parks, D. J.; Willson, T. M. Solid-Phase Synthesis of Hybrid Thiazolidinedione-Fatty Acid PPARγ Ligands. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2491-2496.

(36) Parks, D. J.; Tomkinson, N. C. O.; Villeneuve, M. S.; Blanchard, S. G.; Willson, T. M. Differential Activity of Rosiglitazone Enantiomers at PPARγ. *Bioorg. Med. Chem. Lett.* **1998**, *8*, 3657-3658.

(37) 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. Non-thiazolidinedione Antihyperglycaemic Agents. Part 3: The Effects of Stereochemistry on the Potency of α-Methoxy-β-phenylpropanoic Acids. *Bioorg. Med. Chem.* **1999**, *7*, 821-830.

(38) Molecular Operating Environment, version 2004.03; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada.

(39) (a) Also available within MOE are electrotopological state indices.[39b] These, however, could not be used due to errors in the particular implementation provided within MOE. (b) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: San Diego, 1999.

(40) Xue, L.; Bajorath, J. Accurate Partitioning of Compounds Belonging to Diverse Activity Classes. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 757-764.

(41) (a) Braun, J.; Kerber, A.; Meringer, M.; Rücker, C. Similarity of Molecular Descriptors: The Equivalence of Zagreb Indices and Walk Counts. *MATCH Commun. Math. Comput. Chem.* **2005**, *54*, 163-176. (b) Rücker, C.;

Braun, J.; Kerber, A.; Laue, R. The Molecular Descriptors Computed with MOLGEN. http://www.mathe2.uni-bayreuth.de/molgenqspr.

(42) (a) Rücker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Example of Haloalkane Boiling Points. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 2070-2076. (b) Rücker, C.; Meringer, M.; Kerber, A. QSPR Using MOLGEN-QSPR: The Challenge of Fluoroalkane Boiling Points. *J. Chem. Inf. Model.* **2005**, *45*, 74-80.

(43) Leave-one-out was done after descriptor selection, as usual. This practice was criticized by Wold[43a] and by Hawkins.[46b] However, the alternative suggested by these authors seems rather impracticable. (a) Wold, S. Validation of QSAR's. *Quant. Struct.-Act. Relat.* **1991**, *10*, 191-193.

(44) (a) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem*. **1972**, *15*, 1066-1068. (b) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238-1244.

(45) (a) Livingstone, D. J.; Salt, D. W. Judging the Significance of Multiple Linear Regression Models. *J. Med. Chem.* **2005**, *48*, 661-663. There are errors in equation 2 in this paper; D. J. Livingstone kindly sent the correct equation via e-mail to C. R.. (b) Livingstone, D. J.; Salt, D. W. Variable Selection – Spoilt for Choice? *Rev. Comput. Chem.* **2005**, *21*, 287-348.

(46)(a) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579-586. (b) Hawkins, D. M. The Problem of Overfitting, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1-12.

(47) Katritzky, A. R.; Fara, D. C.; Karelson, M. QSPR of 3-Aryloxazolidin-2-one Antibacterials. *Bioorg. Med. Chem.* **2004**, *12*, 3027-3035.

(48) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Model.* **2002**, *20*, 269-276.

(49) Kubinyi, H. QSAR in Drug Design. Chapter X.4.2 in Gasteiger, J. (Ed.) Handbook of Chemoinformatics, volume 4, pp 1532-1554, Weinheim, 2003.

(50) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In: van de Waterbeemd, H. (Ed.) *Chemometric Methods in Molecular Design*, Weinheim, 1995, pages 309-318.

(51) Baumann, K.; Stiefl, N. Validation Tools for Variable Subset Regression. *J. Computer-Aided Mol. Des.* **2004**, *18*, 549-562.
(52) (a) Randić, M. Orthogonal Molecular Descriptors. *New J. Chem.* **1991**, *15*, 517-525. (b) Randić, M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672-687. (c) Peterangelo, S. C.; Seybold, P. G. Synergistic Interactions among QSAR Descriptors. *Int. J. Quantum Chem.* **2004**, *96*, 1-9.
(53) Cramer, R. D., Bunce, J. D.; Patterson, D. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, 7, 18-25.
(54) DeGrazia, M. J.; Thompson, J.; Vanden Heuvel, J. P.; Peterson, B. R. Synthesis of a High-Affinity Fluorescent PPARγ Ligand for High-Throughput Fluorescence Polarization Assays. *Bioorg. Med. Chem.* **2003**, *11*, 4325-4332.
(55) Estrada, E.; Delgado, E. J.; Alderete, J. B.; Jana, G. A. Quantum-Connectivity Descriptors in Modeling Solubility of Environmentally Important Organic Compounds. *J. Comput. Chem.* **2004**, *25*, 1787-1796.
(56) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest. Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69-77.
(57) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model*. **2005**, *45*, 839-849.