Article

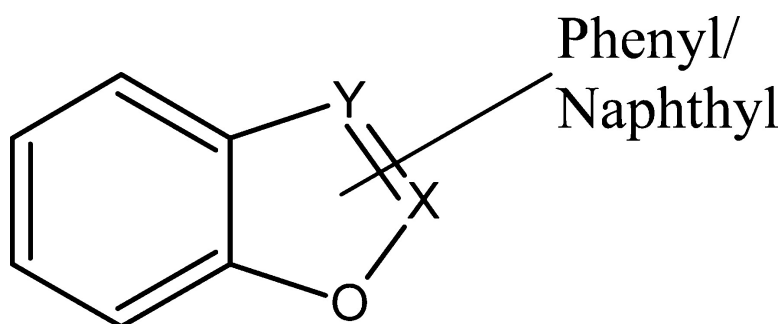# Identification of the Structural Requirements of the Receptor-Binding Affinity of Diphenolic Azoles to Estrogen Receptors ☐ and β by Three-Dimensional Quantitative Structure–Activity Relationship and Structure–Activity Relationship Analysis

Assia Demyttenaere-Kovatcheva, Mark T. D. Cronin, Emilio Benfenati, Alessandra Roncaglioni, and Elena LoPiparo

X=C; Y=N (Oxazole)
X=N; Y=C (Isoxazole)

More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 2 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

Identification of the Structural Requirements of the Receptor-Binding Affinity of Diphenolic Azoles to Estrogen Receptors α and β by Three-Dimensional Quantitative Structure–Activity Relationship and Structure–Activity Relationship Analysis

Assia Demyttenaere-Kovatcheva,*,† Mark T. D. Cronin,† Emilio Benfenati,‡ Alessandra Roncaglioni,‡ and Elena LoPiparo‡

School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England, and Istituto di Ricerche Farmacologiche "Mario Negri" Milano, Via Eritrea 62, 20157 Milano, Italy

Three-dimensional (3D) quantitative structure–activity relationship (QSAR) and structure–activity relationship (SAR) analyses were applied concurrently to a data set of highly selective estrogen receptor β (ERβ) agonists. The data set consisted of diphenolic azoles characterized by similar structural skeletons but with different binding modes to the estrogen receptor site. Models were developed separately with respect to the relative binding affinities (RBAs) to ERα and ERβ. Steric and electrostatic fields were calculated for a training set of 72 compounds using comparative molecular field analysis (CoMFA). The model developed for ERα RBA yielded R^2 of 0.91 and q_{cv}^2 of 0.60. The model developed for ERβ RBA yielded R^2 of 0.95 and q_{cv}^2 of 0.40. Both models were validated successfully using an external test set of 32 compounds. A new concept of test set evaluation based on the variability of the biological response due to the variability of the living organism has been introduced. The CoMFA analysis was supported by a SAR study. In addition to the most favorable steric and electrostatic regions identified by CoMFA, a number of structural descriptors were identified as being important for binding. These are the number of substituents attached to the main skeleton of each compound, the largest distance between the oxygen atoms of each molecule, and the angle defined by the planes that split the phenyl or the naphthyl and the benzisoxazole or the benzoxazole moiety in a morphometrically longitudinal way.

Introduction

The mechanism of interaction between the estrogen receptor (ER) and its potential ligands has been a topic of profound research for many years. It has resulted in the identification of two ER isoforms,[1,2] ERα and ERβ, and synthesis of multiple series of novel compounds.[3–9] The ligand binding domains (LBDs) for both receptor subtypes have been identified and compared as to their overall structure and hormone-binding cavity.[10,11] The high sequence identity between the receptors was considered to be a logical reason for the high similarity in the tertiary and quaternary structures of both receptor isoforms.[11] However, an alignment of the LBD of ERα (rat, mouse, and human) and ERβ (rat) showed that, along with various regions of conservation, there are also segments that are not conserved.[1] The findings mentioned above expanded the number of new challenges for research. For example, the following topics still lack sufficient information: (1) the importance of each receptor subtype for the natural function of the tissues where it occurs, (2) the interaction of both receptor subtypes reflecting their biological response to a certain ligand, and (3) the mechanism of activation of each receptor subtype.

The hypothesis that the RBA of a ligand to the ER receptor may be related to the actual biological response triggered by the receptor–ligand interaction instigated synthesis of novel ligands showing a preferential selectivity in their binding affinity to only one of the receptor subtypes.[3–9] However, the mobility and plasticity of the ER ligand-binding cavity, in particular, have been identified as one of the most important factors allowing the binding of compounds of different structural types to the receptor site.[11,12] This fact is most probably behind the ability of most of the natural and synthetic ligands to bind to both ERα and ERβ. However, these ligands may have a different magnitude of selectivity in vitro and may not trigger any biological response in vivo.[9]

Recent crystal structures of the ER–ligand complexes[9] demonstrate that compounds of very similar structural type bind to the receptor site in a completely different mode. This fact is essentially important for the analysis of ER–ligand complexes because three-dimensional (3D) quantitative structure–activity relationship (QSAR) theories presented up until the present time have been built using a common pharmacophore[13] of the compounds under investigation or results were obtained by theoretical docking approaches.[14,15] Due to the identification of the two ER receptor subtypes, theoretical research focused on the modeling of the relative binding

---

* To whom correspondence should be addressed. E-mail: A.Kovatcheva@livjm.ac.uk. Tel: +44 151 231 2066. Fax: +44 151 231 20170.
† Liverpool John Moores University.
‡ Istituto di Ricerche Farmacologiche "Mario Negri" Milano.

affinity to each of the receptors separately. A study in relation to the present work has been developed on a set of structurally diverse ligands using the comparative molecular field analysis (CoMFA)[16] methodology. The analysis of the CoMFA models and associated contour plots indicated close similarity between ERα and ERβ in terms of mode of binding, thus, providing a rational basis for ligand selectivity.[13]

A basic problem in the correctness and predictive power of the QSAR models relates to the numerical value of the biological response under investigation. The biological response (also termed the endpoint) covers a range of numerical values depending on gender, age, health condition, etc. Thus, the variability of the biological response is dependent on the variability in the living organism or in a test system.

The aim of this paper was to investigate the structural features contributing to the binding affinity to ERα and ERβ receptors of a set of ligands with very similar structural skeleton and experimentally proven different modes of action.[9] To this end, a 3D-QSAR study applying CoMFA methodology, in combination with a SAR investigation, has been performed. Since the assumption about the relationship between RBA and the biological response triggered by the ER−ligand complex in the whole organism has been shown to deliver controversial results in vivo and in vitro, it is important to note that the results relate only to the endpoint RBA of both ER subtypes. Another aspect of the present work addresses the importance of data quality in the area of QSAR. A new way of evaluating the predictive power of QSAR models was introduced. This allowed the biological variability of the data to be incorporated into the modeling process.

Results

CoMFA and SAR analysis have been applied to the log(RBAs) of a data set of diphenolic azoles. The results are presented with respect to the statistical significance and mechanistic interpretation of the models derived.

Statistical Analysis. For each endpoint (RBAα and RBAβ), several CoMFA models have been developed separately as shown in Figures 1 and 2. The best model for each endpoint has been selected according to the following criteria: (i) the lowest number of components included in the final model, (ii) the best statistical results after partial least-squares (PLS) analysis, and (iii) the greatest predictive power for an external test set. A summary of the PLS analysis is given in Table 4.

For each endpoint, several compounds appeared to be outliers. Compounds **4**, **31**, and **32** appear as outliers in both analyses. Compound **4** lacks an OH group attached to the single phenyl ring. This indicates the importance of the OH group and the hydrogen bond interaction between this molecular moiety and both the ERα and ERβ receptors. Another representative of the training set lacking an OH group is compound **34**. Compound **34** is larger than compound **4** due to the naphthyl moiety. Obviously, the interactions caused by the lipophilic features of the larger naphthyl moiety are of greater importance than the hydrogen bond between the remote OH group attached to the same moiety and the receptor site.
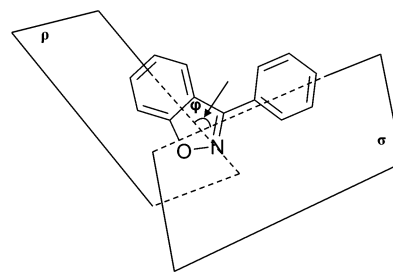
Along with the CoMFA modeling, SAR investigations have been performed. It was found that the distance
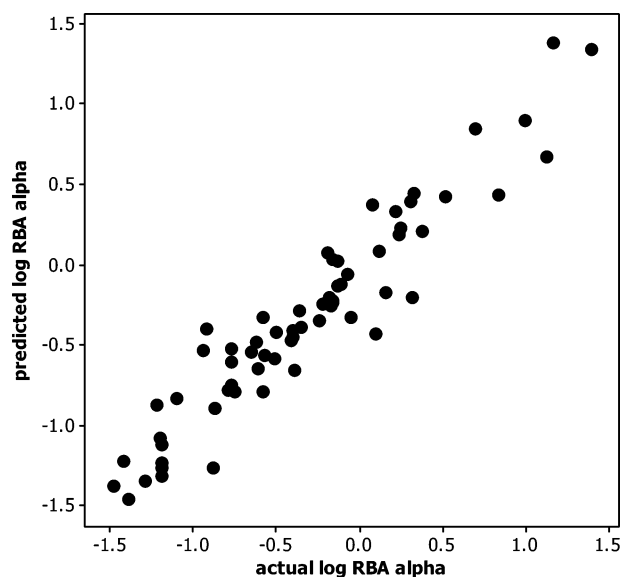
Figure 1. CoMFA−scatter plot of the actual versus predicted log(RBA) (ERα) for the training set.
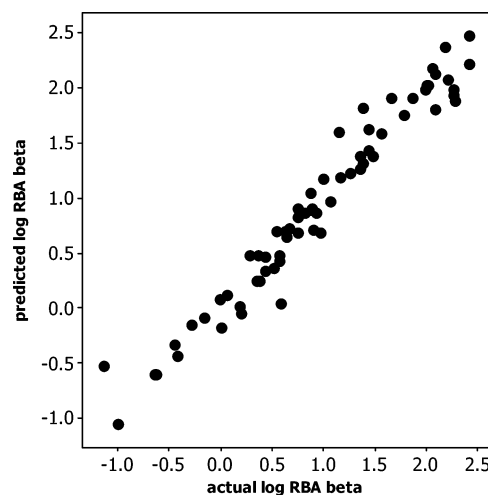
Figure 2. CoMFA−scatter plot of the actual versus predicted log(RBA) (ERβ) for the training set.

Figure 3. Representation of the angle φ built by the planes that split the benzisoxazole moiety (plane ρ) and the phenyl ring (plane σ) using the main skeleton of the subset of molecules.

between the O atoms of the OH groups attached to the benzisoxazole and phenyl moieties respectively was 12 ± 0.3 Å for compounds **24**, **26**, **31**, **32**, **35**, **36**, and **38**. The same distance for the rest of the molecules, except molecules **39** and **40**, was 11 ± 0.4 Å, and the angle φ (Figure 3) was 120° ± 30°. This may explain the fact that compound **34**, which has no OH group in the naphthyl moiety, and compounds **24**, **35**, **36**, and **38**, which contain in the naphthyl moiety one OH group and

Table 1. Structures of the Data Set Consisting of Benzoxazole and Benzisoxazole Derivatives

Cpd 1: Genistein

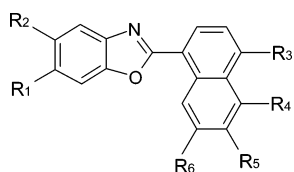| compd	R_1	R_2	R_3	R_4	R_5	R_6	compd	R_1	R_2	R_3	R_4	R_5	R_6
1							13	OH	H	CH_3	H	OH	H
2	OH	H	OH	H	OH	H	14	OH	H	$CH_2CH_2CH_3$	H	OH	H
3	H	H	OH	H	OH	H	15	OH	H	H	$CH_2CH_2CH_3$	OH	H
4	OH	H	OH	H	H	H	16	OH	H	OH	CH_3	OH	H
5	OH	H	H	H	OH	H	17	OH	H	CH_2CN	H	OH	H
6	H	OH	OH	H	OH	H	18	OH	H	CH=NOH	H	OH	H
7	H	OH	H	H	OH	H	19	OH	H	OH	H	OH	Br
8	OH	H	H	Cl	OH	H	20	OH	H	OH	H	OH	Cl
9	OH	H	CN	H	OH	H	21	OH	H	OH	H	OH	CH_3
10	OH	H	H	Br	OH	H	22	OH	H	CH_2CH_3	H	OH	$CH_2CH_2CH_3$
11	OH	H	H		OH	H	23	OH	H	$CH_2CH_2CH_3$	H	OH	$CH_2CH_2CH_3$
12	OH	H	H	CH_3	OH	H							
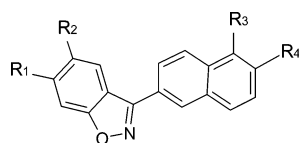
compd	R_1	R_2	R_3	R_4	compd	R_1	R_2	R_3	R_4
24	OH	H	H	OH	26	OH	H	Br	OH
25	OH	H	OH	H	27	H	OH	H	OH
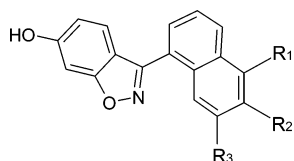
compd	R_1	R_2	R_3	compd	R_1	R_2	R_3	compd	R_1	R_2	R_3
28	H	OH	H	30	H	H	OH	29	OH	H	H

compd	R_1	R_2	R_3	R_4	R_5	R_6	compd	R_1	R_2	R_3	R_4	R_5	R_6
31	OH	H	H	OH	H	H	35	OH	H	H	Br	OH	H
32	OH	H	H	H	OH	H	36	OH	F	H	H	OH	H
33	OH	H	OH	H	H	H	37	H	OH	H	H	OH	H
34	OH	H	H	H	H	H	38	OH	H	H	OH	H	CH_3
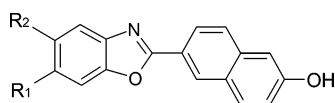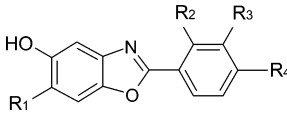
compd	R_1	R_2	compd	R_1	R_2
39	H	OH	40	OH	H
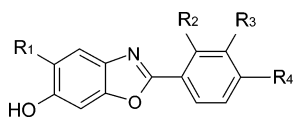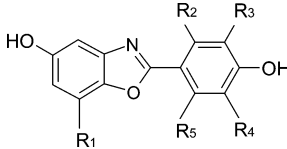
compd	R_1	R_2	R_3	R_4	compd	R_1	R_2	R_3	R_4	compd	R_1	R_2	R_3	R_4
41	H	OH	H	OH	44	H	H	OH	OH	47	H	H	H	OH
42	H	H	H	OH	45	H	H	F	OH	48	Cl	H	CCH_3	OH
43	H	OH	OH	H	46	H	H	Cl	OH	49	H	H	O-n-C_4H_9	OH
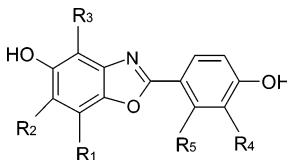
Table 1. (Continued)

compd	R_1	R_2	R_3	R_4	compd	R_1	R_2	R_3	R_4	compd	R_1	R_2	R_3	R_4
50	H	H	H	OH	**53**	H	Cl	H	OH	**56**	Cl	H	F	OH
51	H	H	F	OH	**54**	H	OH	H	OH	**57**	Br	H	F	OH
52	H	H	Cl	OH	**55**	Cl	H	H	OH	**58**	H	OH	OH	H

compd	R_1	R_2	R_3	R_4	R_5	compd	R_1	R_2	R_3	R_4	R_5	compd	R_1	R_2	R_3	R_4	R_5
59	OMe	H	H	H	H	**73**	$CONH_2$	H	H	H	H	**87**	2-Br-vinyl	H	H	H	H
60	Br	H	H	H	H	**74**	CO_2H	H	H	H	H	**88**	2-Br-vinyl	H	H	H	F
61	Br	H	H	F	H	**75**	ethyl	H	H	H	H	**89**	vinyl	H	H	H	F
62	Br	H	H	CF_3	H	**76**	propyl	H	H	H	H	**90**	vinyl	H	H	H	F
63	Br	H	H	H	F	**77**	isopropyl	H	H	H	H	**91**	vinyl	F	H	H	F
64	Br	H	H	H	CH_3	**78**	butyl	H	H	H	H	**92**	vinyl	H	F	H	F
65	CN	H	H	H	H	**79**	ethynyl	H	H	H	H	**93**	vinyl	H	F	F	CH_3
66	CN	H	H	F	H	**80**	allyl	H	H	H	H	**94**	vinyl	H	H	H	H
67	CN	H	H	H	F	**81**	allyl	H	H	F	H	**95**	phenyl	H	H	H	H
68	CH_2Br	H	H	H	H	**82**	allyl	H	H	H	F	**96**	2-furyl	H	H	H	H
69	CH_2CN	H	H	H	H	**83**	vinyl	H	H	H	H	**97**	2-furyl	H	H	H	H
70	CHO	H	H	H	H	**84**	vinyl	H	H	F	H	**98**	2-thienyl	H	H	H	H
71	CO_2Me	H	H	H	H	**85**	2-F-vinyl	H	H	F	H	**99**	2-thiazole	H	H	H	H
72	COEt	H	H	H	H	**86**	2-Me-vinyl	H	H	H	H	**100**	cyclopentane	H	H	H	H

compd	R_1	R_2	R_3	R_4	R_5	compd	R_1	R_2	R_3	R_4	R_5
101	vinyl	H	Br	F	H	**103**	OMe	H	Br	F	H
102	vinyl	Br	Br	F	H	**104**	OMe	Br	Br	F	H

one lipophilic substituent, fit in the final model. For compounds **39** and **40**, the distance between the O atoms participating in the OH groups attached to the benzisoxazole and phenyl moieties respectively was >13 Å. However, the angle φ (Figure 3) for these two compounds is ~180°, which is the favorable geometrical feature with respect to the entire data set. Compounds **5**, **14**, **18**, and **100** should reflect the differences in ERα and ERβ. Compounds **4** and **5** indicate that for smaller molecules alone two OH groups are not enough for good binding to the pocket of the receptor site (as compared to genistein). Compounds **14** and **100** are outliers, possibly due to unfavorable steric fields surrounding the propyl and cyclopentyl substituents. This may be due to the lack of knowledge of their exact conformation. Compound **18** indicates that the substitution at this carbon atom for the ERβ is challenging to both steric and electrostatic contributions.

In addition, the predictive power of the models was tested using an external test set of 32 compounds. The predictions of the values for the test set compounds are given in section b of Tables 2 and 3. For both models, more than 60% of the compounds had predicted activities falling in the "very good" category. More than 80% of the compounds had predicted activities that were "good" or "very good".

Contour Plot Analysis. The results derived from the statistical analyses can be confirmed visually on the contour plots (Figures 4 and 5). Positive steric contributions are represented in green, while negative contributions are yellow. Positive electrostatic contributions are represented in blue, while negative contributions are red. The numbers on the map are associated with the region over which they are located.

The contour map for the model created using RBAβ is presented in Figure 5. Comparison with Figure 4 shows that most of the regions are similar in nature but with different intensities for both receptors. This may explain the difference in the important locations for the binding affinity to both receptors.

The favorable electrostatic interactions are associated with an increase of the positive or negative charge area. The more favorable electrostatic interactions for the phenyl or naphthyl substructure (see regions 1 and 4 in Figures 4 and 5) do not correspond to the substitution of compounds **4**, **24**, **31**, and **32** (see region 1 in Figures 4 and 5). The favorable steric interactions encompass the contribution of a bulky substituent. Obviously compounds **14**, **18**, and **100** have steric restrictions in region 5, which is in a closer vicinity to the unfavorable areas. Several large areas were identified to be specific for each receptor. These areas may have contributed to

Table 2. Actual, Calculated, and Residuals for the Training Set[a] and Actual, Predicted, and Residual for the Test Set for log(RBAα) CoMFA Model

a. Training Set

compd	actual	calcd	resid	compd	actual	calcd	resid	compd	actual	calcd	resid
2	1.12	0.68	0.44	45	−1.42	−1.22	−0.20	73	−1.48	−1.37	−0.11
7	−0.41	−0.47	0.06	46	−1.19	−1.23	0.04	74	−1.19	−1.12	−0.07
8	0.07	0.38	−0.31	47	−0.94	−0.53	−0.41	75	−0.22	−0.24	0.02
10	0.24	0.23	0.01	48	−1.19	−1.26	0.07	77	−0.57	−0.56	−0.01
11	0.30	0.40	−0.10	49	−1.29	−1.34	0.05	78	−0.19	0.08	−0.27
12	−0.13	0.03	−0.16	50	−0.58	−0.32	−0.26	79	−0.18	−0.20	0.02
13	−0.17	−0.25	0.08	52	−1.22	−0.87	−0.35	80	−0.36	−0.28	−0.08
17	−0.35	−0.39	0.04	54	0.23	0.19	0.04	82	0.51	0.43	0.08
18	−0.87	−0.89	0.02	55	−0.16	−0.22	0.06	87	−0.16	−0.23	0.07
20	0.99	0.9	0.09	56	−0.75	−0.79	0.04	88	0.21	0.34	−0.13
21	0.69	0.85	−0.16	57	−0.58	−0.79	0.21	89	0.15	−0.17	0.32
22	1.39	1.34	0.05	58	−1.20	−1.07	−0.13	90	−0.17	−0.25	0.08
23	1.16	1.39	−0.23	60	0.31	−0.20	0.51	91	0.11	0.09	0.02
26	−0.05	−0.32	0.27	61	0.09	−0.43	0.52	92	−1.10	−0.83	−0.27
27	−0.16	0.04	−0.20	62	−0.77	−0.52	−0.25	95	−0.61	−0.64	0.03
33	−0.77	−0.60	−0.17	63	0.83	0.44	0.39	96	−0.40	−0.41	0.01
34	−0.77	−0.75	−0.02	65	−0.11	−0.12	0.01	97	−0.79	−0.78	−0.01
35	−0.07	−0.06	−0.01	66	−0.65	−0.54	−0.11	98	−0.51	−0.58	0.07
36	−0.13	−0.13	−0.00	67	0.37	0.21	0.16	99	−0.62	−0.48	−0.14
37	0.32	0.45	−0.13	70	−0.92	−0.40	−0.52	100	−0.50	−0.42	−0.08
40	−0.39	−0.65	0.26	71	−1.19	−1.31	0.12	102	−0.40	−0.45	0.05
44	−0.88	−1.26	0.38	72	−1.39	−1.46	0.07	104	−0.24	−0.34	0.10

b. Test Set

name	actual mean	resid	pred	mean − SD	mean + SD	resid from range	name	actual mean	resid	pred	mean − SD	mean + SD	resid from range
1	−0.09	0.15	−0.24	0.01	−0.14	0.10	51	−0.69	0.18	−0.87	−0.67	−0.7	0.17
3	−1.28	−1.65	0.37	−1.44	−1.16	−1.53	53	0.36	0.59	−0.23	0.36	0.37	0.59
6	−1.1	−0.88	−0.21	−1.07	−1.12	−0.90	59	−0.9	−0.30	−0.59	−0.63	−1	−0.03
9	0.08	0.34	−0.26	0.02	0.12	0.28	64	0.86	0.97	−0.11	0.84	0.88	0.95
15	0.18	0.55	−0.37	0.05	0.27	0.42	68	−0.97	−0.58	−0.38	−0.49	−1.1	−0.10
47	−0.2	−0.49	0.29	−0.25	−0.18	−0.47	69	−1.19	−0.78	−0.40	−1.36	−1.08	−0.67
19	1.03	0.09	0.93	1.2	0.96	0.02	76	−0.09	−0.25	0.16	−0.22	0.01	−0.15
25	1.12	1.58	−0.46	7.34	1.07	1.53	81	−0.54	−0.26	−0.27	−0.4	−0.59	−0.12
28	0.81	1.36	−0.55	0.65	0.93	1.20	83	−0.15	0.45	−0.60	0	−0.2	0.40
29	1.73	2.13	−0.40	1.56	1.85	1.96	84	−0.58	0.29	−0.87	−0.38	−0.66	0.21
30	−0.73	−0.36	−0.36	7.34	−0.63	−0.26	85	−0.07	0.51	−0.58	0.1	−0.14	0.44
38	0.29	1.09	−0.80	0.2	0.35	1.00	86	−0.38	−0.01	−0.36	−0.41	−0.37	0.00
39	−1.07	−0.79	−0.27	0.16	−1.25	0.00	93	−1.49	−0.13	−1.35	−1.66	−1.38	−0.02
40	0.59	1.08	−0.49	7.34	0.2	0.69	94	−0.54	−0.12	−0.41	7.34	−0.8	0.00
42	−0.45	0.23	−0.68	−0.32	−0.51	0.17	101	−0.51	−0.14	−0.36	−0.4	−0.56	−0.03
43	−0.87	0.34	−1.21	−0.92	−0.84	0.29	103	−0.92	−0.53	−0.38	−0.83	−0.96	−0.44

	residual			residual from range	
	no.	%		no.	%
very good	17	53		22	69
good	9	28		5	16
bad	6	19		5	15

[a] Outliers are excluded from the list.

the differences in the fold selectivity of a compound to each receptor site. Region 6 is different for both endpoints. For RBAα, region 6 depicts unfavorable electrostatic interactions, while for RBAβ it depicts favorable steric interactions.

Discussion

The aim of this study was to investigate the three-dimensional surroundings of a series of benzisoxazole derivatives for which the following information was established experimentally: (i) the RBA to both ERα and ERβ and (ii) the binding mode to the receptor site. The data set was split into a training and test set, which were used for a CoMFA study. The final models were built after the exclusion of several outliers. The internal (using a training set) and external (using an external test set) predictive power of these models was evaluated

by high values of q^2 and R^2, a low value of predictive sum of squares (PRESS) and standard error of estimate (SEE), and over 80% predictions for an external test set classified as "good" and "very good". In addition, the models were successfully tested on the compound genistein, which does not belong to the chemical space covered by the training test.

To explain the presence of outliers, a deeper insight into the structural geometry of the molecules has been obtained by focusing on the distance between the O atoms and the angle φ defined by the planes that split the naphthyl and the benzisoxazole moiety morphometrically along the longitudinal axis of each compound. Three-dimensional contour plots supported the analyses of the statistical results. A pharmacophore has been defined according to which the largest distance between two O atoms should be preferably 11 ± 0.3 Å when the

Table 3. Actual, Calculated, and Residuals for the Training Set[a] and Actual, Predicted, and Residual for the Test Set for log(RBAβ) CoMFA Model

a. Training Set

compd	actual	calcd	resid	compd	actual	calcd	resid	compd	actual	calcd	resid
2	2.01	2.02	−0.01	**44**	0.54	0.7	−0.16	**72**	0.28	0.48	−0.21
5	0.82	0.87	−0.04	**45**	0.97	0.69	0.28	**73**	0.58	0.04	0.53
7	0.89	0.90	−0.00	**46**	−0.29	−0.15	−0.14	**74**	−1.14	−0.53	−0.61
8	1.56	1.59	−0.03	**47**	0.36	0.48	−0.12	**75**	1.44	1.62	−0.18
10	1.48	1.38	0.10	**48**	−0.65	−0.6	−0.04	**77**	0.64	0.64	0.01
11	1.65	1.91	−0.26	**49**	−1.01	−1.05	0.05	**78**	0.66	0.72	−0.06
12	1.06	0.97	0.10	**50**	0.87	1.04	−0.17	**79**	1.38	1.82	−0.43
13	0.99	1.18	−0.19	**52**	0.18	0.02	0.16	**80**	1.44	1.43	0.02
14	1.26	1.22	0.03	**54**	1.16	1.19	−0.04	**81**	2.08	2.13	−0.05
17	0.57	0.48	0.10	**55**	1.35	1.26	0.10	**87**	0.90	0.71	0.19
20	2.26	1.93	0.33	**56**	0.75	0.69	0.06	**88**	1.35	1.38	−0.02
21	1.86	1.91	−0.06	**57**	0.93	0.86	0.07	**89**	2.28	1.88	0.40
22	2.00	2.03	−0.03	**58**	−0.43	−0.44	0.01	**90**	1.99	1.99	−0.00
23	2.06	2.18	−0.11	**60**	2.26	1.98	0.27	**91**	2.21	2.07	0.15
24	2.41	2.22	0.19	**61**	2.08	1.81	0.27	**92**	0.74	0.90	−0.16
26	0.62	0.70	−0.08	**62**	0.34	0.25	0.08	**95**	0.19	−0.05	0.24
27	0.51	0.36	0.15	**63**	2.41	2.47	−0.06	**96**	0.43	0.47	−0.05
33	−0.63	−0.61	−0.02	**65**	1.78	1.76	0.02	**97**	0.06	0.12	−0.06
34	−0.46	−0.34	−0.12	**66**	1.14	1.6	−0.46	**98**	0.57	0.43	0.14
35	0.43	0.34	0.09	**67**	2.18	2.37	−0.19	**99**	−0.01	0.08	−0.09
36	0.89	0.89	0.00	**70**	0.79	0.87	−0.09	**102**	0.37	0.25	0.12
37	1.38	1.32	0.07	**71**	0.00	−0.18	0.19	**104**	0.75	0.83	−0.08
40	−0.17	−0.09	−0.08								

b. Test Set

name	actual mean	resid	pred	mean − SD	mean + SD	resid from range	name	actual mean	resid	pred	mean − SD	mean + SD	resid from range
1	1.56	0.16	1.40	1.57	1.52	0.12	**51**	0.74	−0.09	0.82	0.70	0.84	0.00
3	−0.30	−2.67	2.37	−0.34	−0.16	−2.53	**53**	0.79	−0.18	0.96	0.63	7.30	0.00
6	0.42	−1.76	2.18	0.39	0.50	−1.67	**59**	0.79	−0.89	1.68	0.82	0.70	−0.85
9	1.48	0.26	1.22	1.41	1.70	0.19	**64**	2.41	0.14	2.27	2.34	2.70	0.06
15	1.16	−0.13	1.29	1.12	1.26	−0.03	**68**	0.95	0.27	0.68	0.88	1.26	0.20
47	1.04	−1.14	2.18	0.98	1.22	−0.96	**69**	−0.46	−1.14	0.68	−0.67	7.30	0.00
19	2.30	0.38	1.93	2.44	2.07	0.14	**76**	1.51	0.34	1.18	1.49	1.60	0.31
25	1.18	−0.45	1.63	1.16	1.22	−0.47	**81**	1.41	0.42	0.99	1.34	1.70	0.35
28	0.87	−1.03	1.89	1.03	0.61	−0.87	**83**	2.01	0.84	1.17	2.00	2.05	0.83
29	1.78	0.09	1.68	1.94	1.52	0.00	**84**	1.86	0.98	0.87	1.76	2.30	0.89
30	−0.68	−2.36	1.68	−0.63	−0.79	−2.31	**85**	2.05	2.13	−0.08	1.98	2.30	2.06
38	1.78	2.93	−1.15	1.67	2.30	2.83	**86**	0.40	−0.52	0.92	0.35	0.60	−0.32
39	−0.16	−0.46	0.30	−0.90	7.30	0.00	**93**	0.25	0.25	0.00	0.35	0.07	0.07
40	2.08	0.86	1.22	2.11	2.00	0.78	**94**	1.12	−0.15	1.28	1.28	0.87	0.01
42	0.86	−0.01	0.87	0.90	0.76	0.00	**101**	1.16	0.95	0.21	1.14	1.22	0.93
43	0.30	−0.06	0.36	0.27	0.38	0.00	**103**	0.84	−0.13	0.97	0.80	0.96	0.00

	residual			residual from range	
	no.	%		no.	%
very good	18	56		19	63
good	6	19		7	22
bad	8	25		6	15

[a] Outliers are excluded from the list.

angle φ is ∼120°. This condition was found to be very important when the main skeleton has only two substituents represented by OH groups. For the same geometrical conditions, three substituents can guarantee successful binding to the receptor site. The OH groups should preferably be attached to the benzisoxazole moiety and to the phenyl ring attached by single bond to the benzisoxazole moiety. If the largest distance between two O atoms is more than 11.40 Å, the angle φ should preferably be closer to 180° to guarantee significant binding.

The contour plots demonstrate a quite similar distribution of the favorable and unfavorable steric and electrostatic fields, which, however, differ in their intensity. The areas that are characteristic only for a given receptor site were associated with the fold selec-tivity of the compounds to the corresponding receptor site.

Conclusion

The present study deals with several important aspects for successful QSAR modeling. It has been demonstrated that the CoMFA alignment of the compounds can be reliable when considered with regard to the information of their binding mode to the receptor site. Until the current time, no automated approach for alignment is known that can recognize the differences in the binding mode of compounds with a very similar structural skeleton. Thus, such an alignment relies on experimental results derived from crystal structures.

Another aspect of this study shows that, unlike classical QSAR, the concept of variability of the biologi-
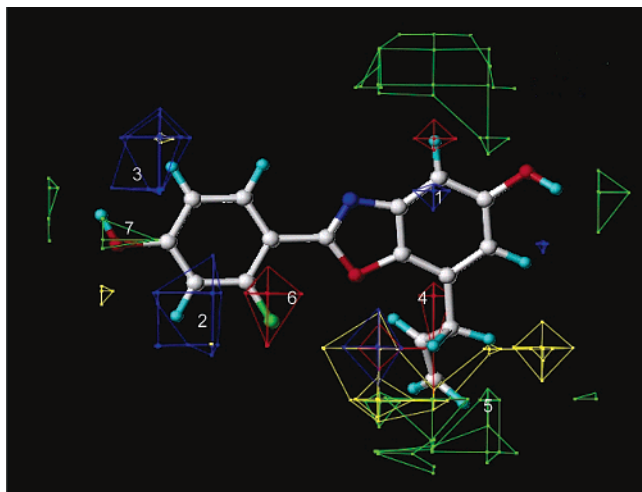
Figure 4. CoMFA−contour plot visualizing the steric and electrostatic contributions of the training set for log(RBA) (ERα). Positive steric contributions are represented in green, while negative contributions are yellow. Positive electrostatic contributions are represented in blue, while negative contributions are red.
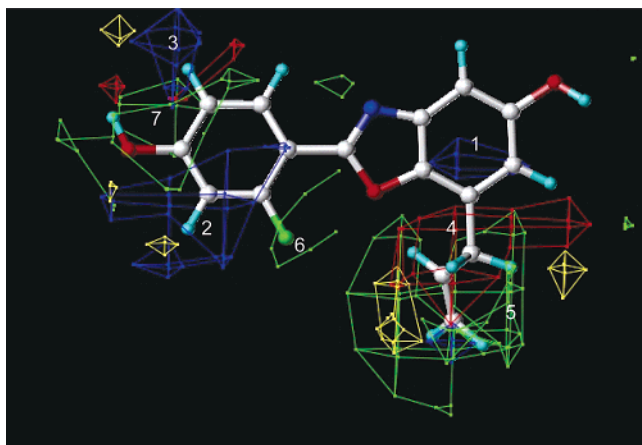
Figure 5. CoMFA−contour plot visualizing the steric and electrostatic contributions of the training set for log(RBA) (ERβ). Positive steric contributions are represented in green, while negative contributions are yellow. Positive electrostatic contributions are represented in blue, while negative contributions are red.

Table 4. Summary of CoMFA Results for ERα and ERβ[a]

	ERα	ERβ
optimal number of components	6	10
q_{cv}^2	0.60	0.40
R^2	0.91	0.95
SEE	0.21	0.20
F values	94	117
probability	0	0
steric contributions	0.40	0.49
electrostatic contributions	0.60	0.51
outliers	**4, 5, 14, 24, 31, 32**	**4, 18, 31, 32, 100**

[a] Probe atom is C (sp³, +1); cutoff value for steric and electrostatic fields is 30.

cal response is incorporated in the development and testing of QSAR models. The statistical results derived from internal and external validation indicate successful modeling of the relative binding affinity to both ERα and ERβ for in vitro measurements. It is demonstrated that the "quality" of an appropriate data set for QSAR should not be judged on well-performed measurements

solely. All steps starting from molecular modeling, going through careful consideration of the methods applied to the particular QSAR modeling, and ending up with well-argued data analysis are important for the development of a mechanistically interpretable model. In addition, the identification of outliers is shown to be crucial for the explanation of hidden information with respect to the entire data set. Finally, to introduce a meaningful mechanistic interpretation of the QSAR models, the SAR investigation is also a very helpful tool.

Materials and Methods

Data Set. The compounds of the data set belong to a series of recently synthesized highly selective ERβ agonists.[9] From this series, 104 compounds were selected according to the following criteria: (i) accuracy of the available published chemical structure and (ii) determination of the IC_{50} value for both ERα and ERβ. The structures of all compounds of the data set are given in Table 1.

Biological Activity. The human ERα and ERβ ligand binding domains (LBD) were assessed separately. The RBA, expressed as IC_{50}, was determined in a competitive radioligand-binding assay.[9] According to the published experimental data, IC_{50} values were the mean of at least two experiments. Thus, the standard deviation, SD, was also provided. Values without a standard deviation were established for a single determination only.[9] The value of IC_{50} has been converted to RBA according to the formula in eq 1:

$$RBA = \frac{IC_{50}(\text{estradiol})}{IC_{50}(\text{competitor})} \times 100 \qquad (1)$$

The logarithm(10) of the observed values was used as dependent variable for the present QSAR study. The RBAα and RBAβ values for the training and test sets are given in Tables 2, sections a and b, and 3, sections a and b, respectively.

The SD associated with the measurement of a certain biological endpoint allows the definition of a range of values within which the endpoint is "acceptable". In these terms, the SD reflects the variability in the response of a single individual experiment toward the endpoint under consideration. In the present work, the residual between actual and predicted by CoMFA models RBA was calculated with respect to the lower and upper limits of the range of the RBA, defined when the SD has been subtracted from or added to the mean RBA, respectively. Following this concept, the predictive power of the model has been evaluated on an external test set by the number of compounds for which the predictions were "very good", "good", and "poor". If the residual between the closer limit of the actual range defined by the mean log(RBA) and the corresponding log(SD) and the predicted log(RBA) value is between 0 and 0.5, the prediction of the compound is considered to be "very good". If the residual is between 0.5 and 1, the prediction of the compound is considered to be "good". If the residual is greater than 1, the prediction of the compound is considered to be "poor".

Training and Test Sets. The data set was divided into a training set and an external test set in such way that two-thirds of the compounds were assigned to the training set and one-third to the test set. Thus, the training set consisted of 72 compounds and the test set consisted of 32 compounds. To perform a comparison of the results obtained for both receptors, the initial training and test sets included identical compounds for both ERα RBA and ERβ RBA endpoints. Several methods have been developed to divide a data set into training and test sets.[17,18] In the current study, these methods were not considered to be appropriate due to the presence of original measurements with standard deviation of the IC_{50} greater than 50%. Moreover, a particular aim of the present study was to test the CoMFA models on a compound out of the chemical space covered by the compounds of the training

set. For this reason, the test set compounds (cf. section b of Tables 2 and 3) were selected according to the following criteria: (i) lower accuracy of the IC$_{50}$ values, that is, the standard deviation of the IC$_{50}$ associated with each compound was greater than 50% of the corresponding mean IC$_{50}$ value for at least one of both subtypes, (ii) insufficient information about the binding mode, or (iii) structural geometry that is not bounded by the descriptor space formed by the training set. All compounds of the test set, except compounds **1**, **28**, **29**, and **30**, were selected according to the first criterion. Compounds **28**, **29**, and **30** were selected according to the second criterion. Compound **1**, genistein, was selected according to the third criterion.

Only compounds with an activity with a small standard deviation were included in the training set (cf. section a, Tables 2 and 3). Thus, for these compounds, the range of activity determined by the standard deviation over which the IC$_{50}$ value can spread has been neglected, and the biological activity was represented by a single value corresponding to the experimentally established mean value IC$_{50}$.[9] Due to the lower accuracy of determination of the IC$_{50}$ value of the compounds of the test set, the whole range including the mean of the IC$_{50}$ and the corresponding standard deviation has been taken into account. If the standard deviation was estimated to be larger than the corresponding mean IC$_{50}$ value, the resulting lower limit of the IC$_{50}$ range was calculated as a negative value, which is possible from a statistical point of view but impossible practically. Thus, for such cases, the lower limit of the IC$_{50}$ range has been approximated to 0.

Molecular Representation. The atomic coordinates of the compound 2-(3-fluoro-4-hydroxyphenyl)-7-vinyl-1,3-benzoxazol-5-ol[8] were extracted from the crystal complex with the accession code 1X7b in the Protein Data Bank (PDB).[19] The optimized geometry served as a template to derive the skeleton of the data set compounds. Three-dimensional structures of the molecules were built using the SKETCH module of the Sybyl7.0 molecular modeling software.[20]

The atomic coordinates of genistein were extracted from the crystal complex with the accession code 1X7J[22] in the PDB. The structure has been optimized subsequently. All geometry optimizations were performed using a conjugate gradient of 0.001 kJ/mol and the Tripos force field with Gasteiger−Hückel charges within Sybyl7.0.[20]

Alignment Rules. The alignment of the compounds is the most important step to develop a true and reliable CoMFA model. Since the release of the protein−ligand complexes in the Protein Data Base was not permitted during the work on this paper, the alignment of the compounds of the data set is based on the information about the binding mode provided in the original source.[9]

Although the data set consists of compounds of a similar structural type, the phenyl benzisoxazoles, the crystallized complexes between ERβ and the ligands **2**, **24**, **31**, **50**, and **84** showed differences in their binding mode to the receptor site.[9] Each of these compounds represents a subset of molecules characterized by an identical main skeleton and identical binding mode but with different substituents (cf. Table 1 and ref 9). Thus, the benzisoxazole substructure of molecules **2**, **24**, **31**, **50**, and **84** served as a template for the remaining compounds of the corresponding subsets.

However, to ensure alignment reproducibility, certain rules were established. The template molecules **2**, **24**, **31**, **50**, and **84** were aligned according to the rules described below.

For molecules **2** and **31**, the binding mode determines that the benzisoxazole substructure interacts with amino acids Glu305 and Arg346 of the ERβ.[9] However, the benzisoxazole substructures of both molecules are not identical. The rest of the molecular structure of compound **2** is represented by a phenyl ring and for compound **31** by a naphthyl substructure. Both the phenyl ring and the naphthyl substructure are bound to a different carbon atom from the benzisoxazole substructure. Thus, the common template for these two compounds includes the phenyl ring associated with the bond to the benzisoxazole

substructure together with the bond to the benzisoxazole substructure.

For compounds **24**, **50**, and **84**, the binding mode determines that the benzisoxazole substructure interacts with the His475 and Ile373 of the ERβ receptor site, which, in turn, corresponds to Met421 of the ERα receptor site.[9] Thus, the common template for these compounds includes the single phenyl ring or the naphthyl substructure, respectively, and the phenyl ring associated with the benzisoxazole substructure.

Generation of CoMFA Models. The aligned molecules were placed in a three-dimensional grid space with the following characteristics: grid space 1.5 Å, grid size $22 \times 22 \times 16$ grid units; number of grid points 2816.

An absolute maximal value of 30 kcal/mol for the steric and electrostatic energies calculated at each grid point was established experimentally. The CoMFA descriptors in terms of van der Waals (steric) and Coulombic (electrostatic) interactions were calculated using an sp^3 carbon probe atom with a +1.0 charge. In the present study, equal weights were assigned to steric and electrostatic fields using the CoMFA standard scaling procedure implemented in SYBYL, which implies partial least-squares (PLS) analysis.[20] To establish the maximal number of components contributing to the CoMFA model with the lowest standard error of estimate (SEE), PLS analysis was applied in combination with leave-one-out (LOO) cross-validation. LOO cross-validation implies exclusion of each compound of the training set and the prediction of its activity by the model developed using the remaining compounds of the training set. The cross-validated coefficient q^2 can be established using the formula in eq 2.

$$q^2 = 1 - \frac{\sum (Y_{\text{pred}} - Y_{\text{obsd}})^2}{\sum (Y_{\text{obsd}} - Y_{\text{mean}})^2} \qquad (2)$$

where Y_{pred}, Y_{obsd}, and Y_{mean} are predicted, actual, and mean of the relative binding affinity (log(RBA)), respectively, and $\sum (Y_{\text{pred}} - Y_{\text{osbd}})^2$ is the predictive sum of squares known as PRESS. In addition, the statistical significance of the models was described by the SEE and F and probability value computed according to the definitions in SYBYL. If the probability $p = 0$, the results are not obtained by chance and the explanatory variables are truly uncorrelated.

For each model, the LOO cross-validated predictions were examined. Compounds with the highest absolute value of the residual between the observed and the predicted relative binding affinity were identified as outliers. After establishing the outliers and the optimal number of components, the PLS procedure was repeated without cross-validation while being given, as input, the exact number of components contributing to the final model. To reduce the noise of the analysis, a minimum filtering value of 3.00 kcal/mol was used. Along with the LOO cross-validation, the models were tested on an external test set. In the evaluation of the predictive power of the models using an external test set, the SD of the endpoint was taken into account.

SAR Measurements. Several CoMFA models have been developed using known data sets of estrogens.[16,21] The CoMFA models provided models with good statistical characteristics and predictive power. However, the information derived from the CoMFA plots was not sufficient for the identification of the structural requirements of the data set under investigation. In this study, the information from CoMFA modeling is supported by SAR investigations to obtain a deeper insight into the complex interactions between ligands and ERα and ERβ. Due to the similarity in the chemical structure of the compounds, the distances between the O atoms in each of the molecules of the subseries and the angle defined by the planes that split the naphthyl and the benzoxazole moiety morphometrically along a longitudinal axis (Figure 3) were taken into consideration. All measurements were performed using Sybyl7.0.[20]

References

(1) Kupier, G. G.; Enmark, E.; Pelto-Huikko, M.; Nilsson, S.; Gustafsson, J. A. Cloning of a Novel Estrogen Receptor Expressed in Rat Prostate and Ovary. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5925−5930.

(2) Mosselman, S.; Polman, J.; Dijkema, R. ERβ: Identification and Characterization of a Novel Human Estrogen Receptor. *FEBS Lett*. **1996**, *392*, 49−53.

(3) Sun, J.; Meyers, M. J.; Fink, B. E.; Rajendran, R.; Katzenellenbogen, J. A.; Katzenellenbogen B. S. Novel Ligands That Function as Selective Estrogens or Antiestrogens for Estrogen Receptor-Alpha or Estrogen Receptor-Beta. *Endocrinology* **1999**, *140*, 800−804.

(4) Mortensen, D. S.; Rodriguez, A. L.; Carlson, K. E.; Sun, J.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Synthesis and Biological Evaluation of a Novel Series of Furans: Ligands Selective for Estrogen Receptor Alpha. *J. Med. Chem.* **2001**, *44*, 3838−3848.

(5) Angelis, M. D.; Stossi, F.; Carlson, K. A.; Katzenellenbogen, B. S.; Katzenellenbogen, J. A. Indazole Estrogens: Highly Selective Ligands for the Estrogen Receptor β. *J. Med. Chem.* **2005**, *48*, 1132−1144.

(6) Hoekstra, W. J.; Patel, H. S.; Liang, X.; Blanc, J.-B. E.; Heyer, D. O.; Willson, T. M.; Iannone, M. A.; Sue H. Kadwell, S. H.; Miller, L. A.; Pearce, K. H.; Simmons, C. A.; Shearin, J. Discovery of Novel Quinoline-Based Estrogen Receptor Ligands Using Peptide Interaction Profiling. *J. Med. Chem.* **2005**, *48*, 2243−2247.

(7) Renaud, J.; Bischoff, S. F.; Buhl, T.; Floersheim, Ph.; Fournier, B.; Geiser, M.; Halleux, C.; Kallen, J.; Keller, H.; Ramage, P. Selective Estrogen Receptor Modulators with Conformationally Restricted Side Chains. Synthesis and Structure−Activity Relationship of ERα-Selective Tetrahydroisoquinoline Ligands. *J. Med. Chem.* **2005**, *48*, 364−379.

(8) Manas, E. S.; Unwalla, R. J.; Xu, Z. B.; Malamas, M. S.; Miller, C. P.; Harris, H. A.; Hsiao, C.; Akopian, T.; Hum, W.-T.; Malakian, K.; Wolfrom, S.; Bapat, A.; Bhat, R. A.; Stahl, M. L.; Somers, W. S.; Alvarez, J. C. Structure-Based Design of Estrogen Receptor-β Selective Ligands. *J. Am. Chem. Soc.* **2004**, *126*, 15106−15119.

(9) Malamas, M. S.; Manas, E. S.; McDevitt, R. E.; Gunawan, I.; Xu, Z. B.; Collini, M. D.; Miller, C. P.; Dinh, T.; Henderson, R. A.; Keith, J. C., Jr.; Harris, H. A. Design and Synthesis of Aryl Diphenolic Azoles as Potent and Selective Estrogen Receptor-β Ligands. *J. Med. Chem.* **2004**, *47*, 5021−5040.

(10) Brzozowski, A.; Pike, A.; Dauter, Z.; Hubbard, R.; Bonn, T.; Engström, O.; Öhman, L.; Greene, G. L.; Gustafsson, J.-Å.; Carlquist, M. Molecular Basis of Agonism and Antagonism in the Oestrogen Receptor. *Nature* **1997**, *389*, 753−758.

(11) Pike, A.; Brzozowski, A.; Hubbard, R.; Bonn, T.; Thorsell, A.-G.; Engström, O.; Ljunggren, J.; Gustafsson, J.-Å.; Carlquist, M. Structure of the Ligand-Binding Domain of Estrogen Receptor Beta in the Presence of a Partial Agonist and a Full Antagonist. *EMBO J.* **1999**, *18*, 4608−4618.

(12) Wagner, R. L.; Apriletti, J. W.; McGrath, M. E.; West, B. L.; Baxter, J. D.; Fletterick, R. J. A Structural Role for Hormone in the Thyroid Hormone Receptor. *Nature* **1996**, *378*, 690−697.

(13) Tong, W.; Perkins, R. QSAR Models for Binding of Estrogenic Compounds to Estrogen Receptor α and β Subtypes. *Endocrinology* **1997**, *138*, 4022−4025.

(14) Sippl, W.; Höltje, H.-D. Structure-Based 3D-QSAR-Merging the Accuracy of Structure-Based Alignments With the Computational Efficiency of Ligand-Based Methods. *J. Mol. Struct.* **2000**, *503*, 31−50.

(15) Sippl, W. Binding Affinity Prediction of Novel Estrogen Receptor Ligands Using Receptor-Based 3-D QSAR Methods. *Bioorg. Med. Chem.* **2002**, *10*, 3741−3755.

(16) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(17) Golbraikh, A.; Tropsha, A. Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 357−369.

(18) Kovatcheva, A.; Golbraikh, A.; Ohloff, S.; Xiao, Y.-D.; Zheng, W.; Wolschann, P.; Buchbauer, G. and Tropsha, A. Combinatorial QSAR of Ambergris Fragrance Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582−595.

(19) http://www.rcsb.org, accessed in March 2005.

(20) http://www.tripos.com, accessed in March 2005.

(21) Waller, C. L. A Comparative QSAR Study Using CoMFA, HQSAR, and FRED/SKEYS Paradigms for Estrogen Receptor Binding Affinities of Structurally Diverse Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 758−765.

(22) Manas, E. S.; Xu, Z. B.; Unwalla, R. J.; Somers, W. S. Understanding the Selectivity of Genistein for Human Estrogen Receptor-Beta Using X-ray Crystallography and Computational Methods. *Structure* **2004**, *12*, 2197.

JM050604J