

4D-QSAR Analysis of a Set of Ecdysteroids and a Comparison to CoMFA Modeling

Malini Ravi,[†] Anton J. Hopfinger,[†] Robert E. Hormann,^{*,‡} and Laurence Dinan[§]

Laboratory of Molecular Modeling and Design (M/C-781), University of Illinois at Chicago, College of Pharmacy, 833 South Wood Street, Chicago, Illinois 60612-7231, RHeoGene, L.L.C., P.O. Box 949, Spring House, Pennsylvania 19477-0904, and Department of Biological Sciences, University of Exeter, Hatherly Laboratories, Prince of Wales Road, Exeter, Devon, EX4 4PS, U.K.

Received August 7, 2001

The ecdysteroid-responsive *Drosophila melanogaster* B_{II} cell line is a prototypical homologous inducible gene expression system. A training set of 71 ecdysteroids, for which the $-\log(\text{EC}_{50})$ potencies in the ecdysteroid-responsive B_{II} cell line were measured, was used to construct 4D-QSAR models. Four nearly equivalent optimum 4D-QSAR models, for two modestly different alignments, were identified ($Q^2 = 0.76-0.80$). These four models, together with two CoMFA models, were used in consensus modeling to arrive at a three-dimensional pharmacophore. The C-2 and C-22 hydroxyls are identified as hydrogen-bond acceptor sites which enhance activity. A hydrophobic site near C-12 is consistent with increasing activity. The side-chain substituents at C-17 are predicted to adopt semiextended "active" conformations which could fit into a cylinder-shaped binding pocket lined largely with nonpolar residues for enhanced activity. A test set of 20 ecdysteroids was used to evaluate the QSAR models. Two 4D-QSAR models for one alignment were identified to be superior to the others based on having the smallest average residuals of prediction for the prediction set (0.69 and 1.13 $-\log[\text{EC}_{50}]$ units). The correlation coefficients of the optimum 4D-QSAR models ($R^2 = 0.87$ and 0.88) are nearly the same as those of the best CoMFA model ($R^2 = 0.92$) determined for the same training set. However, the cross-validation correlation coefficient of the CoMFA model is less significant ($Q^2 = 0.59$) than those of the 4D-QSAR models ($Q^2 = 0.80$ and 0.80).

INTRODUCTION

Engineered chemically inducible gene expression systems (gene switches) are potentially useful in proteomics, cell-based high-throughput screening assays, toxicology screening, large-scale protein production, functional genomics, and gene therapy. The application of regulatable gene expression systems to gene therapy is particularly noteworthy. Gene therapy holds substantial potential for the treatment of a wide range of disorders including cancer, neurodegenerative, cardiovascular, respiratory, and autoimmune diseases,¹⁻⁷ notwithstanding outstanding issues.^{8,9}

In its simplest form, an inducible switch consists of a receptor capable of binding to a DNA response element upstream of the regulated gene and a specific small-molecule ligand with affinity for the receptor. More complex and variant systems involve additional homologous DNA-binding partner proteins and additional transcription factors and/or repressors. Desirable attributes of engineered inducible systems are those of regulatory systems in general: low basal expression level, rapid time response, high sensitivity, high inducer specificity, and a large dynamic range (dose-responsiveness and maximum level of induction), in addition to tissue selectivity and nontoxicity (nonimmunogenicity) of the expression system components.¹⁰ Recent approaches to gene therapy introduce inducible gene expression systems

which are actuated in a dose-dependent and reversible manner by a ligand.^{11,12}

Gene expression systems based on naturally occurring transcription components not only have some of the desired qualities of a good regulatory system but are also characterized by high basal expression, weak induction, and pleiotropic effects. These issues, in particular the concern for specificity, are addressed by chimeric receptors.¹³ To date, at least four distinct chimeric systems have been developed, each based on the responsiveness of the regulatory protein to a complementary ligand: tetracycline, progesterone, rapamycin, or 20-hydroxyecdysone.^{11,12} The last of these functions on the basis of affinity of the insect growth hormone 20-hydroxyecdysone (20E) for the ecdysteroid receptor (EcR) which, when bound to its homolog, the Ultraspiracle protein (USP), forms a functional receptor heterodimer.^{14,15} Both EcR and USP are members of the nuclear receptor superfamily, a class of proteins well-established to be modular in the function of discrete peptide domains.¹⁶⁻¹⁸ Improved EcR-based gene expression systems have been achieved by constructing chimeras through exchange of modular domains with those of other nuclear receptors.¹⁹ Furthermore, EcR-based switches are functional in heterologous mammalian cell environments.²⁰⁻²³ Nevertheless, it is notable that responsiveness to ecdysteroids in engineered mammalian cells is generally weaker than in insect cells.²⁴

Chimeric and heterologous EcR-based systems, in which the specificity of the EcR ligand-binding domain is perturbed, might be more confidently investigated if compared to EcR

* Corresponding author phone: (215)619-5690; fax: (215)619-1635; e-mail: rhormann@rheogene.com.

[†] University of Illinois at Chicago.

[‡] RHeoGene, L.L.C.

[§] University of Exeter.

ligand specificity in a completely homologous environment, namely, a natural insect system. An opportunity to do this is available through the *Drosophila melanogaster* B_{II} cell line.²⁵ The first goal of this study is to demonstrate the development of a highly receptor specific endpoint model for a prototypical inducer of a gene expression system. This is achieved through consensus multidimensional QSAR modeling.

The second purpose of this study is to develop pharmacophore hypotheses for ecdysteroid potency which would assist the design and identification of new ligand chemotypes. In the case of the structurally and physiologically related vertebrate steroid hormone receptors, several classes of nonsteroidal ligands with quite different chemistries have been identified.^{26,27} Therefore, it seems highly probable that nonsteroidal ligands can be found which interact with the ecdysteroid receptor complex as agonists or antagonists.²⁸ Indeed, one class of nonsteroidal ligands, the bisacylhydrazines, has been demonstrated to be agonists of 20-hydroxyecdysone^{29,30} and are documented to be efficacious actuators of ecdysteroid-regulatable gene induction systems.^{31–32}

A highly advantageous feature of expression systems based on EcR, in contradistinction to inducible systems derived from vertebrate receptors, is the lower probability of cross-responsiveness with native transcriptional control mechanisms. This is especially significant for gene therapy applications. Furthermore, available data suggest that ecdysteroids are nontoxic to mammals.³³ Thus, the nonmammalian origin of both the ligand and the receptor components of the ecdysteroid inducible system render it attractive for drug design.

This principle of non-cross-responsiveness also reveals the EcR gene expression system to be a highly attractive target for insect control in crop protection applications. Ecdysteroids are the steroid hormones of arthropods and are implicated in the control of molting, metamorphosis, reproduction, embryonic development, and diapause.³⁴ Since precisely controlled titers of ecdysteroids are essential for the temporal and spatial development of insects, and because ecdysteroids are structurally different from vertebrate steroids, the pursuit of crop protection agents which interfere with ecdysteroid signal transduction is a worthwhile endeavor from both an environmental and human health standpoint. Thus, the third purpose of this study is to develop a QSAR which might be used toward this goal. Moreover, ecdysteroid SAR could lead to strategies for the breeding or genetic modification of crop species to generate plants which contain enhanced levels of potent phytoecdysteroids, a strategy which would address the sometimes suboptimal physiochemical and metabolic characteristics of these substances.³⁵

To date a large number of comparative molecular field analyses, CoMFA, have been reported in this and other journals. However, there has not been the opportunity to perform a four-dimensional quantitative structure–activity relationship, 4D-QSAR, analysis^{36–38} on a data set also studied using CoMFA.^{39–41} The fourth purpose of this study is to evaluate the relative performance of each QSAR method and to determine how the two methods might complement one another. One factor driving the development of 4D-QSAR analysis is the need to take into account multiple conformations, alignments, and substructure groupings in constructing QSAR models. These degrees of freedom are

normally held fixed in a CoMFA analysis. Insofar as 4D-QSAR analysis can meaningfully predict “active” conformations and the preferred alignment for a training set, it may serve as a “preprocessor” for a subsequent CoMFA.

CoMFA has been performed on a training set of 71 ecdysteroids.²⁵ Here, this same training set has been examined using both 4D-QSAR analysis and CoMFA. Thus, in addition to providing QSAR models for mapping the receptor binding sites for the training set ecdysteroids as well as defining possible pharmacophores and predicting the activities of new hypothetical ecdysteroids, a comparison has been made between these two QSAR approaches.

METHOD

Training Set of Ecdysteroids. The training set of 71 ecdysteroids are given in Table 1, and their corresponding $-\log(\text{EC}_{50})$ values from the *Drosophila melanogaster* B_{II} cell line screen⁴² are given in Table 2. The $-\log(\text{EC}_{50})$ values of the training set span more than 5 orders of magnitude (4.00–9.51). Table 2 also contains the estimated hydrophobicity measure, MlogP,⁴³ taken directly from the CoMFA study of the ecdysteroids. The carbon atoms of the ecdysteroids are designated by conventional steroid numbering as given for ecdysone in Figure 1. Toward the end of the 4D-QSAR study of the training set, an additional 20 ecdysteroids were reported along with their EC_{50} values. These 20 analogues were adopted as a test set (Table 3) to evaluate the predictiveness of the QSAR models developed from the training set.

Comparative Molecular Field Analysis (CoMFA). CoMFA methodology has been described and reviewed extensively. CoMFA models were developed using Sybyl 6.5⁴⁴ by the general approach previously described.²⁵ The models depicted here differ from previous models in slight adjustments of several training set members and minor correction of three structures.⁴⁵ CoMFA fields were optimized by first examining standard steric and electrostatic fields (alone and in combination), indicator steric and electronic fields (alone and in combination), and hydrogen bonding fields, and optimizing as a function of energy cut-off, dielectric function, and minimum sigma. CoMFA region displacement, step size, and probe atom type had been thoroughly explored in the previous study.²⁵

4D-QSAR Analysis. The methodology and corresponding operations for performing a 4D-QSAR analysis have been presented elsewhere in detail.^{36–38} The analysis is begun by generation of a reference grid cell space and a 3D-structure for each compound in the training set. In this particular 4D-QSAR application, space was divided into a cubic grid cell lattice of 1 Å resolution. The crystal structure of the steroid ring system of ecdysone⁴⁶ was used as the core structure for building trial 3D-structures of each of the ecdysteroids. Hyperchem 5.01⁴⁷ was used to construct and refine the 3D-structures of the ecdysteroids.

The set of *interaction pharmacophore elements*, IPEs, given in Table 4, was next assigned to each compound followed by a conformational ensemble sampling of each compound to generate its conformational ensemble profile, CEP. The CEP provides information about the flexibility of a molecule and lays a foundation to identify the “binding” conformation of the molecule.

Table 1. Chemical Structures and Numbering Code for the 71 Ecdysteroids of the Training Set

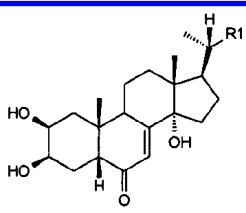
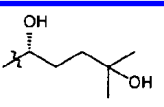
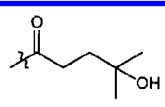
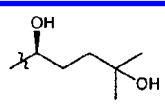
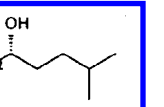
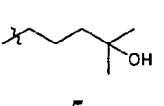
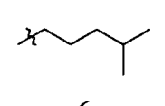
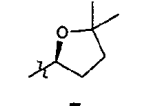
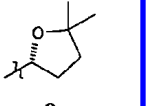
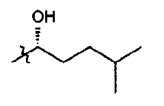
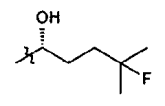
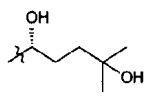
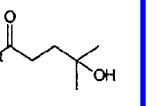
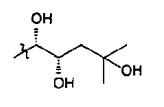
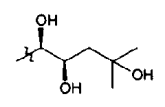
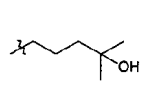
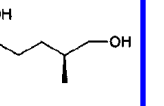
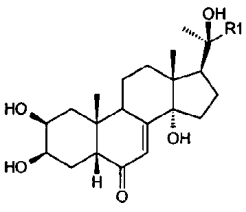
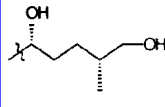
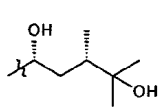
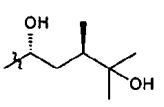
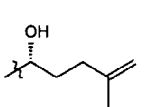
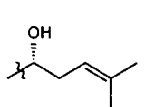
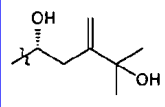
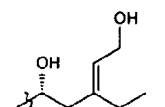
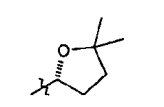
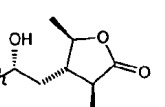
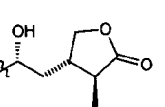
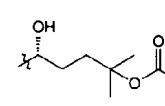
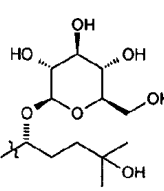
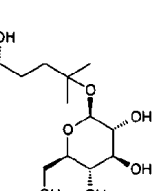
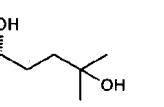
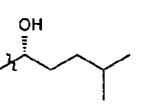
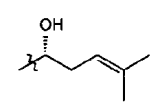
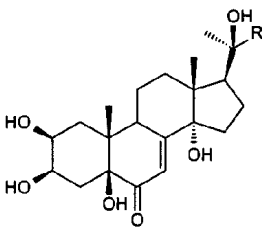
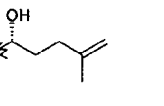
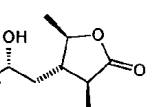
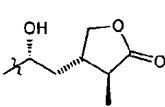
 <p>Ecdysone Core</p>	 1	 2	 3	 4
	 5	 6	 7	 8
	 9	 10	 11	 12
	 13	 14	 15	 16
 <p>20-Hydroxyecdysone Core</p>	 17	 18	 19	 20
	 21	 22	 23	 24
	 25	 26	 27	 28
	 29	 30	 31	 32
 <p>5, 20-Dihydroxyecdysone Core</p>	 33	 34	 35	

Table 1 (Continued)

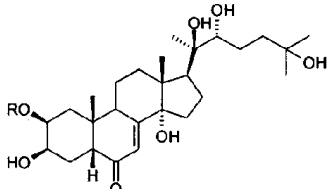

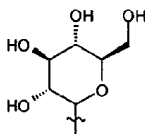
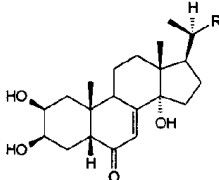
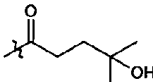
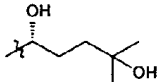
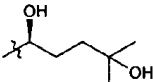
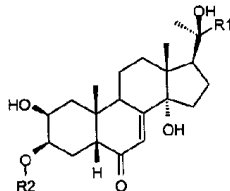
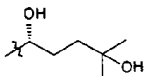
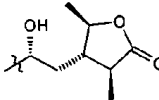
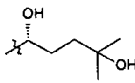


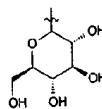
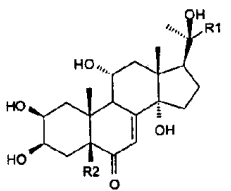
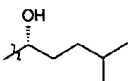
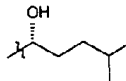
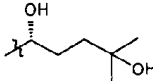
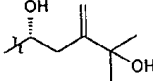
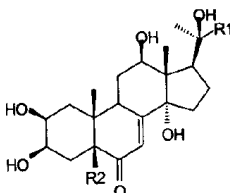
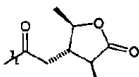
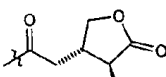
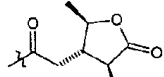
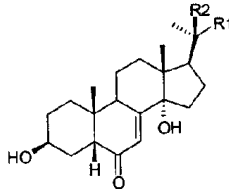
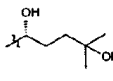
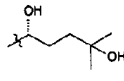
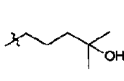
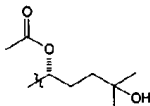
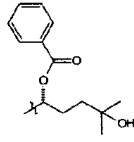
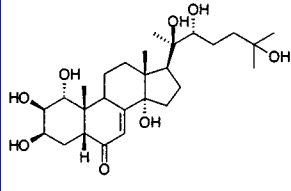
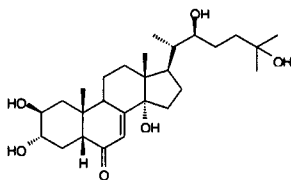
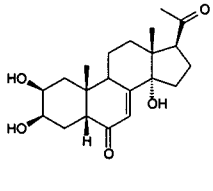
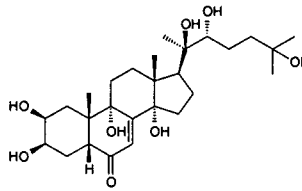
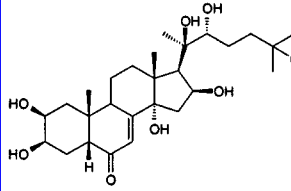
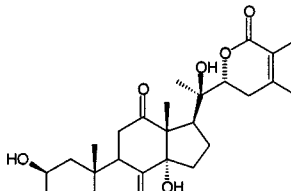
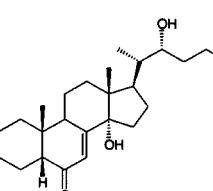
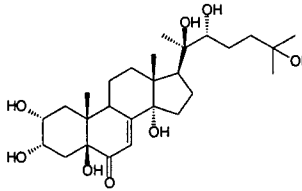
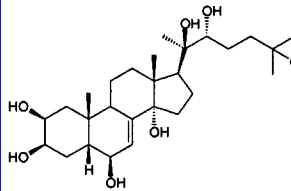
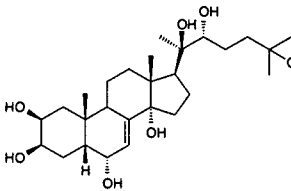
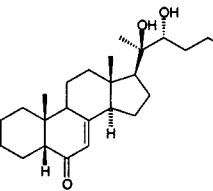
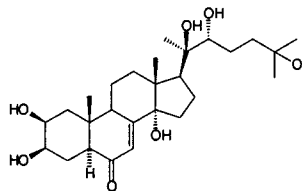
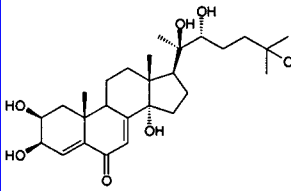
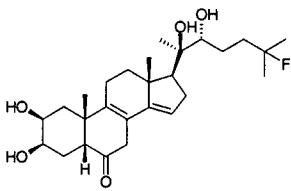
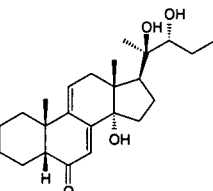
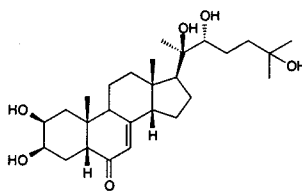
						
2-Substituted-20-hydroxyecdysone Core		36	37			
						
20-Iso-ecdysone Core		38	39	40		
	R1					
3-Substituted-20-hydroxyecdysone Core	R2					
		41	42	43		
	R1					
11-Hydroxy-20-hydroxyecdysone Core	R2	H	OH	H	H	
		44	45	46	47	
	R1					
12, 20-Dihydroxyecdysone Core	R2	OH	OH	H		
		48	49	50		
	R1					
2-Deoxyecdysone / 20-hydroxyecdysone Core	R2	H	OH	OH	H	OH
		51	52	53	54	55

Table 1 (Continued)

Core Modification and Unusual Substitution			
			
56	57	58	59
			
60	61	62	63
			
64	65	66	67
			
68	69	70	71

Currently, molecular dynamics simulation, MDS, is used to generate the CEP of each molecule in the training set. The MDSs have been performed using the MOLSIM package⁴⁸ with an extended MM2 force field⁴⁹ and a molecular dielectric of 3.5. The temperature for the MDS is set at 300 K with a simulation sampling time of 10 ps with intervals of 0.001 ps for a total sampling of 10 000 conformations of each ecdysteroid. The atomic coordinates of each conformation and its total potential energy sampled during the MDS are recorded every 0.1 ps for a total of 1000 "frames", or steps, in constructing the CEP of each compound.

Selection of the trial alignments used to compare the compounds in the training set is the next step in the 4D-QSAR analysis. To be clear, 4D-QSAR analysis does not "solve" the alignment problem by necessarily optimizing the model as a function of alignment. Rather, the 4D-QSAR scheme permits a rapid evaluation of individual trial align-

ments. Consequently, the alignment problem can be treated as a search and sample operation analogous to conformational profiling.

Three-ordered atom alignment rules^{36,50} are currently used in 4D-QSAR analysis. Alignments are usually selected to probe the main parts of a molecule and also to span across these main parts. The five alignments used in this study, reported in Table 5, reflect this strategy. Each conformation from the CEP of each compound is placed in the reference grid cell space according to the trial alignment. The grid cell occupancy descriptors, GCODs, are computed by recording how often a given IPE occupies a given grid cell over the CEP. The GCODs are the basis set of trial 4D-QSAR descriptors.

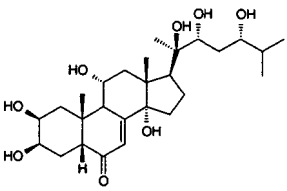
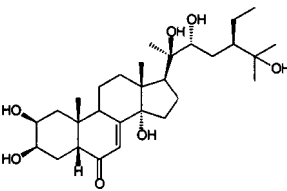
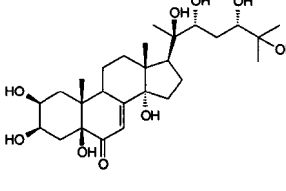
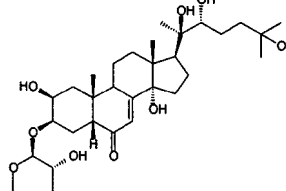
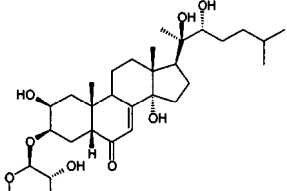
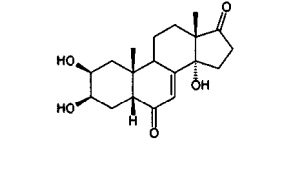
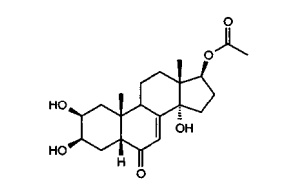
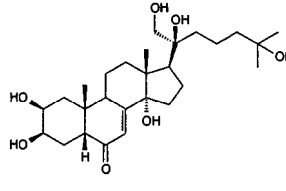
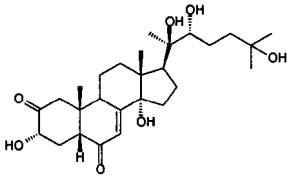
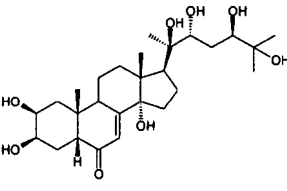
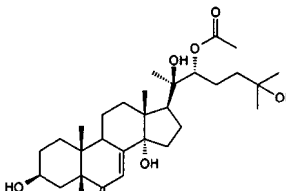
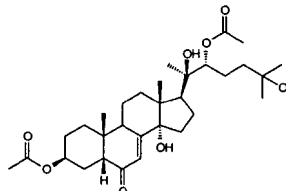
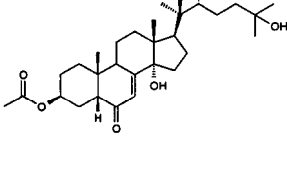
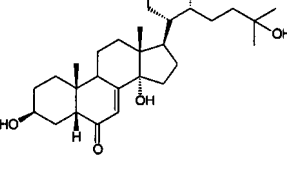
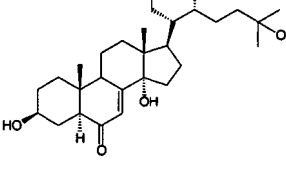
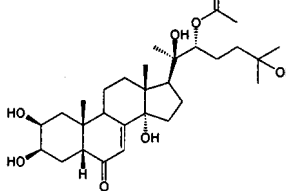
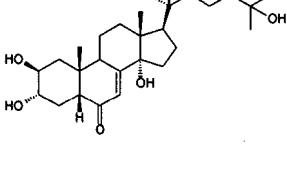
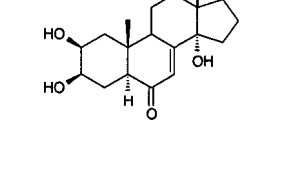
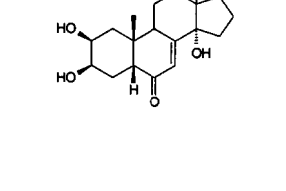
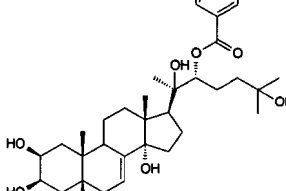
A 4D-QSAR analysis generates an enormous number of GCODs because of the large number of grid cells and the six IPE types. Partial least squares (PLS) regression⁵¹ is used to perform the data reduction fit between the observed

Table 2. EC₅₀ and -log(EC₅₀) Values of the Training Set Compounds against the *Drosophila* BII Cell Line^a

no.	steroid	EC ₅₀ (M)	-log(EC ₅₀)	MLog P
1	ecdysone	1.1×10^{-6}	5.96	2.20
2	22-dehydroecdysone	4.5×10^{-8}	7.35	2.10
3	22- <i>epi</i> -ecdysone	$3.5, 5.3 \times 10^{-6}$	5.36 ^d	2.20
4	25-deoxyecdysone	1.0×10^{-8}	8.00	2.97
5	22-deoxyecdysone	$\sim 10^{-5}$	5.30 ^b	2.97
6	22,25-dideoxyecdysone	$\sim 10^{-6}$	6.30 ^b	3.76
7	(22S)20-(2',2'-dimethylfuranyl)ecdysone	7.3×10^{-6}	5.14	2.97
8	(22R)20-(2',2'-dimethylfuranyl)ecdysone	1.0×10^{-5}	5.00	2.97
9	ponasterone A	3.1×10^{-10}	9.51	2.20
10	25-fluoroponasterone A	5.1×10^{-9}	8.29	2.30
11	20-hydroxyecdysone	7.5×10^{-9}	8.12	1.44
12	22-dehydro-20-hydroxyecdysone	1.7×10^{-7}	6.77	1.35
13	geradiasterone	4.0×10^{-7}	6.40	0.69
14	22,23- <i>epi</i> -geradiasterone	4.0×10^{-5}	4.40	0.69
15	22-deoxy-20-hydroxyecdysone	1.4×10^{-8}	7.85	2.20
16	(25S)-inokosterone	2.7×10^{-7}	6.57	1.44
17	(25R)-inokosterone	1.5×10^{-7}	6.82	1.44
18	24- <i>epi</i> -makisterone A	2.2×10^{-7}	6.66	1.64
19	makisterone A	1.3×10^{-8}	7.89	1.64
20	25,26-didehydroponasterone A	1.5×10^{-7}	6.82	2.10
21	stachysterone C ($\Delta 24[25]$)	1.4×10^{-8}	7.85	2.10
22	24(28)-dehydromakisterone A	4.0×10^{-9}	8.40	1.54
23	24(28)-dehydroamarasterone B	5.2×10^{-7}	6.28	1.74
24	shidasterone	$1.5, 4 \times 10^{-6}$	5.56 ^d	2.20
25	cyasterone	$1.2, 5.3 \times 10^{-8}$	7.49 ^d	1.80
26	29-norcyasterone	1.2×10^{-8}	7.92	1.61
27	viticosterone E	1×10^{-7}	7.00	1.80
28	20-hydroxyecdysone 22- β -D-glucopyranoside	4.7×10^{-5}	4.33	-0.66
29	20-hydroxyecdysone 25- β -D-glucopyranoside	8.5×10^{-6}	5.07	-0.66
30	polypodine B	1.0×10^{-9}	9.00	0.69
31	25-deoxypolypodine B	1.0×10^{-8}	8.00	1.44
32	5 β -hydroxystachysterone C ^c	3.5×10^{-8}	7.46	1.35
33	5 β -hydroxy-25,26-didehydroponasterone A ^c	6.3×10^{-8}	7.20	1.35
34	sengosterone	9.0×10^{-8}	7.05	1.06
35	29-norsengosterone	1.3×10^{-7}	6.89	0.87
36	20-hydroxyecdysone 2-acetate	4×10^{-7}	6.40	1.80
37	20-hydroxyecdysone 2- β -D-glucopyranoside	2×10^{-5}	4.70	-0.66
38	22-dehydro-20- <i>iso</i> -ecdysone	3.0×10^{-6}	5.52	2.10
39	20- <i>iso</i> -ecdysone	1.0×10^{-4}	4.00	2.20
40	20- <i>iso</i> -22- <i>epi</i> -ecdysone	1.0×10^{-4}	4.00	2.20
41	20-hydroxyecdysone 3-acetate	4.7×10^{-7}	6.33	1.80
42	Cyasterone 3-acetate	4.3×10^{-7}	6.37	2.17
43	20-hydroxyecdysone 3- β -D-glucopyranoside	1.3×10^{-5}	4.89	-0.66
44	ajugasterone C	$3.0, 8.0 \times 10^{-8}$	7.26 ^d	1.44
45	muristerone A	2.2×10^{-8}	7.66	0.69
46	turketerone	$3.0 \times 10^{-7}, 1.3 \times 10^{-6}$	6.10 ^d	0.69
47	atrotosterone C	3.0×10^{-6}	5.52	0.80
48	22-dehydro-12-hydroxysengosterone	2.7×10^{-6}	5.57	0.24
49	22-dehydro-12-hydroxynorsengosterone	1.3×10^{-5}	4.89	0.05
50	22-dehydro-12-hydroxycyasterone	1.3×10^{-6}	5.89	0.97
51	2-deoxyecdysone	5.0×10^{-5}	4.30	2.97
52	2-deoxy-20-hydroxyecdysone	6.6×10^{-7}	6.18	2.20
53	2,22-dideoxy-20-hydroxyecdysone	$\sim 5 \times 10^{-5}$	4.60 ^b	2.97
54	2-deoxyecdysone 22-acetate	1.4×10^{-6}	5.85	3.31
55	2-deoxy-20-hydroxyecdysone 22-benzoate	6.3×10^{-6}	5.20	2.67
56	1- <i>epi</i> -integristerone A	2.5×10^{-7}	6.60 ^d	0.69
57	3- <i>epi</i> -22- <i>iso</i> -ecdysone	4.4×10^{-6}	5.36	2.20
58	poststerone	$\sim 2 \times 10^{-5}$	4.98 ^b	1.62
59	9,20-dihydroxyecdysone	1.6×10^{-5}	4.80	0.69
60	malacosterone	9.0×10^{-6}	5.05	0.69
61	ajugalactone	1.6×10^{-7}	6.80	1.63
62	3-dehydroecdysone	6.0×10^{-6}	5.22	2.10
63	rapisterone D	1.0×10^{-9}	9.00	0.69
64	6 β -hydroxy-20-hydroxyecdysone	1.7×10^{-7}	6.77	1.55
65	6 α -hydroxy-20-hydroxyecdysone	2.0×10^{-6}	5.70	1.55
66	14-deoxy(14 α -H)-20-hydroxyecdysone	3.0×10^{-8}	7.52	2.20
67	(5 α -H)20-hydroxyecdysone	3.3×10^{-6}	5.48	1.44
68	4-dehydro-20-hydroxyecdysone	2.8×10^{-7}	6.55	1.35
69	2 β ,3 β ,20R,22R-tetrahydroxy-25-fluoro-5 β -cholest-8,14-dien-6-one	7.2×10^{-9}	8.14	2.98
70	5-deoxykaladasterone	5.2×10^{-10}	9.28	2.10
71	14-deoxy(14 β -H)-20-hydroxyecdysone	8.3×10^{-7}	6.08	2.20

^a The Mlog P values are also given. These data are taken directly from ref 20. ^b Midpoint between concentrations of observed activity and absence of activity. ^c Assignment of **32** vis-à-vis **33** based on analogy to HPLC elution order of **20** and **21**. ^d Average between two concentrations of observed activity.

Table 3. Ecdysteroid Test Set for 4D-QSAR and CoMFA Analysis

			
72	73	74	75
			
76	77	78	79
			
80	81	82	83
			
84	85	86	87
			
88	89	90	91

dependent variable, $-\log(\text{EC}_{50})$, and the corresponding set of GCODs. Before performing PLS, three different types of data filtering can also be performed to reduce the size of the descriptor pool. The most highly weighted GCODs (currently 200) from the PLS regression analysis were used to form the trial basis set of GCOD descriptors for deriving the best 4D-QSAR models.

The final 4D-QSAR models are generated using a genetic algorithm, GA. The 200 most highly weighted PLS determined descriptors, plus any non-GCOD descriptors that the investigator wishes to consider, are used as the model building basis set. Any functional forms of the descriptors can be used in 4D-QSAR model construction. In this application only linear forms of the GCOD were used, and

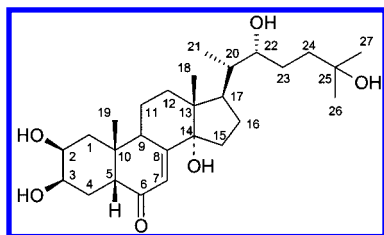


Figure 1. The carbon atom numbering system of ecdysteroid molecules, using ecdysone as an example.

Table 4. Interaction Pharmacophore Elements, IPEs, Used in This 4D-QSAR Analysis

IPE description	IPE no.
any type of atom	0
nonpolar atom	1
polar atom of positive charge	2
polar atom of negative charge	3
hydrogen bond acceptor	4
hydrogen bond donor	5

Table 5. Three-Ordered Atom Alignment Rules Used in the 4D-QSAR Analysis

atom numbers			alignment no.
6	12	16	1
20	22	25 ^a	2
2	3	6	3
17	20	22	4
3	17	25 ^a	5

^a This atom number varied for steroids such as **25**, **26**, **24**, **25**, **42**, and **61** where the carbon which neighbors the ring carbonyl group was selected.

models were built both with and without inclusion of MlogP. The specific GA currently used in the 4D-QSAR program⁵² is a modification of the genetic function approximation, GFA.^{53,54} The optimal number of descriptors in the QSAR models was explored by varying the smoothing factor of Friedman's lack-of-fit measure, LOF,⁵⁴ which, in turn, is the direct optimization metric used in model building.

Various diagnostic measures are currently used to analyze the resultant 4D-QSAR models generated by the GFA including descriptor usage as a function of crossover operation, linear cross-correlation among descriptors and/or dependent variables, $-\log(\text{EC}_{50})$, number of distinct and significant models, and corresponding measures of model significance. The measures of model significance currently being used are the LOF, the correlation coefficient, and the leave-one-out cross-validation correlation coefficient. Outlier analysis is also performed. An analogue of the training set is considered an outlier when the difference in the predicted and observed dependent variable values is more than 2.0 standard deviations from the mean.

Identification of the global optimum 4D-QSAR model is difficult for this training set because the regression equations contain a large number of terms (GCODs) resulting in the generation of many nearly equivalent models during the GFA optimization step. Correspondingly, the optimization convergence of the 4D-QSAR models for the ecdysteroid training set is sensitive to the choice of smoothing factor and the number of crossovers permitted in the GFA analysis. The number of crossovers required for convergence to optimum 4D-QSAR models is in the range of 73 000–75 000, and the smoothing factor⁵⁴ is in the 1.5–1.7 range

in order to evolve 4D-QSAR models which maximize fitting significance for a minimum number of descriptors. However, the resultant MLR equations (4D-QSAR models) are statistically significant and stable but hard to find in descriptor space.

The 4D-QSAR model building process described above can be repeated for each trial alignment. In this study five different three ordered-atom alignments were considered. Once the desired set of alignments are evaluated, the "best" 4D-QSAR model with respect to alignment, conformational sampling, IPE, and grid cell sizes is identified. While a single best model can be identified using one, or more, measures of significance, added information and understanding comes from comparing a family of best models. In essence, the best, and distinct, set of 4D-QSAR models from GA analysis is used in composite to build a *manifold* 4D-QSAR model. GA optimization offers the potential of uncovering the set of ways in which the data of a training set can be packaged in a good statistical model. The manifold of the 4D-QSAR models captures these multiple "explanations" and can be used to make consensus predictions of activity on test compounds.

Distinct, that is to say, statistically independent QSAR models are defined as models whose residuals of fit are poorly cross-correlated to one another. The idea behind this condition is that if the residuals are poorly correlated to one another, the models are "explaining" the training set data differently, and, therefore, the models are different from one another. The correlation coefficients of the residuals of fit for the CoMFA and 4D-QSAR models have been used to estimate the similarity between the models predicted by these two methods.

The linear cross-correlation matrix of the GCODs within a model is constructed to determine which pairs, if any, of GCODs are highly correlated to one another. Members of highly correlated pairs of QSAR descriptors are often assumed as providing redundant information. However, a highly cross-correlated pair of GCODs is not necessarily viewed as requiring one of two GCODs to be removed from the model. A test is made to see if both highly correlated GCODs are necessary to retain high statistical significance of the model. If one of the GCODs can be removed without major loss in significance, this GCOD is removed from the descriptor trial basis set and 4D-QSAR model optimization is repeated. If both GCODs are needed to retain a good model, neither GCOD is removed from the model, but an analysis is attempted to discover the source of the high cross-correlation. In a number of cases it has been seen that GCODs are coupled to one another in an "allosteric" fashion through the conformational behavior of a molecule.^{36,50,52} The occupancy of one GCOD increases or decreases as the other decreases or increases, respectively, because of the conformational behavior of the molecule and the selected alignment.

The final step in the 4D-QSAR analysis is to hypothesize the "active conformation" of each compound in the training set. This is achieved by first identifying all conformer states in the CEP of a molecule that are within ΔE (currently set at 2 kcal/mol) energy units of the global minimum of the CEP. Each member of the resulting set of low-energy conformations is individually evaluated using the best 4D-QSAR model. Since only a *single* conformation is used, the grid cell occupancy is either zero or one for each GCOD of

Table 6. Statistical Summary for the 4D-QSAR Models of the Three Best Alignment Rules and for the Corresponding Best CoMFA Models

	4D-QSAR alignment/M log P						CoMFA	
	1	1/MlogP	3	3/MlogP	5	5/MlogP	E, electrostatics ^a	ES, electrostatics/sterics ^b
no. of training set members	67 ^c	67 ^c	67 ^c	67 ^c	67 ^c	67 ^c	67 ^d	67 ^e
no. of descriptors	14	14	15	15	15	16	200	476
no. of components							5	5
r^2	0.846	0.872	0.834	0.884	0.714	0.768	0.901	0.923
standard error of estimate							0.426	0.370
Q^2	0.755	0.799	0.736	0.800	0.593	0.619	0.564	0.593
standard error of prediction							0.893	0.848

^a Relative contributions: electrostatics: = 1.0 (indicator region-focused), F values = 110.7, $\sigma = 0.5$. ^b Relative contributions: steric: 0.366 (sterics indicator, nonregion-focused), electrostatics: 0.634 (electrostatic indicator, region-focused), F values = 145.3, $\sigma = 0.45$. A similar model with logP was constructed, resulting in some degradation of Q^2 . ^c Outliers: 22, 30, 43, 63. ^d Outliers: 15, 22, 30, 60. ^e Outliers: 30, 53, 60, 63.

the model. The single conformation within ΔE which predicts the highest "activity" ($-\log[\text{EC}_{50}]$) is selected as the active conformation. The postulated active conformations can be used as structure design templates, which includes their deployment as the molecular geometries of each ligand of the training set for a structure-based binding study.⁵⁵ The postulated active conformations and the best alignment can also be used to perform a CoMFA 3D-QSAR analysis of the training set. In the case at hand, the predicted active conformations have been compared to those used in the CoMFA analysis prepared independent of the 4D-QSAR investigation.

Other useful structure and activity features can be estimated from a 4D-QSAR model. The difference between the predicted activity (using the active conformation) and the observed activity of a molecule can be viewed as the loss in possible activity of the molecule owing to its flexibility. That is, the active conformation tends to have IPEs that completely occupy activity-enhancing GCODs, while the activity-decreasing GCODs are not occupied. In contrast, the CEP of the molecule leads to both the activity-enhancing and activity-decreasing GCODs being partially occupied. Put another way, the difference between predicted activity using the active conformation and the observed activity is the loss in activity owing to ligand conformational entropy.

The 4D-QSAR model can also be used as a *virtual high throughput screen*, VHTS, to predict the activities of the members of a virtual library of "similar" compounds to those of the training set.^{56,57} Although VHTS of a library of hypothetical ecdysteroids was NOT considered as part of this study, the $-\log(\text{EC}_{50})$ values of a test set of 20 compounds (test library) were predicted and compared to the corresponding observed values as a validation test of each of the 4D-QSAR and CoMFA models.

RESULTS

Optimum CoMFA Models. The electrostatics CoMFA model (model E, Table 6) was developed by region-focusing an initial electrostatic indicator field using discriminant power (exponential weight = 0.4). The analysis of the resultant focused field without MLogP was optimized on Q^2 by elimination of outliers. Inclusion of MLogP at this stage gave an inferior Q^2 value. The most satisfactory PLS analysis was obtained by elimination of ecdysteroids **15**, **22**, **30**, and **60**. This analysis was chosen, and further improvement was attempted by focusing a second time after elimination of the

outliers ($Q^2 = 0.564$; five components; $s_{\text{PRESS}} = 0.893$; 200 CoMFA columns; weighting: discriminatory power = 0.3). A non-cross-validated PLS analysis was performed, and the final parameters and statistics appear in Table 6 as model E.

The electrostatic indicator/steric indicator CoMFA model (model ES) was developed in like fashion to the electrostatic model by initial exploration of field types and variables and resulted in an optimal model based on an initial set of 71 ecdysteroids using a region-focused electrostatic indicator field and a steric indicator field (nonregion-focused) ($Q^2 = 0.475$; 476 descriptors; $s_{\text{PRESS}} = 1.01$; 5 components). Inclusion of MlogP degraded the model. This analysis was then optimized on Q^2 by elimination of outliers. The most satisfactory PLS analysis yielded $Q^2 = 0.593$ (476 columns; five components; $s_{\text{PRESS}} = 0.848$) with elimination of steroids **30**, **53**, **60**, and **63** (Table 6). Among a family of possible models, these PLS analyses described in Table 6 were chosen to represent this particular CoMFA approach applied to the ecdysteroid dataset.

To recapitulate, PLS analysis utilizing the electrostatics indicator field alone (model E) resulted in $Q^2 = 0.564$ and $R^2 = 0.901$ with a standard error of estimate = 0.426, when four training set outliers were removed ($n = 67$). PLS analysis utilizing the steric indicator field together with the electrostatics indicator field (model ES) resulted in $Q^2 = 0.593$ and $R^2 = 0.923$ with a standard error of estimate = 0.370, when four training set outliers were removed.

The electrostatics contour maps for model E and the electrostatics/steric contour map for model ES appear in Figures 2 and 3, respectively, with ponasterone A shown as a representative molecule. For model E, a large region where positive charge is favored (blue) appears wrapping around the side chain. Important regions where positive charge is disfavored (red) appear at O-20 and between C-26 and C-27. Less significant red regions appear on the α -side of C-7 and C-8 and fairly distant from the molecule on the α -side of C-14. For model ES, the situation is similar. In the ES model, regions where steric bulk is favored appear distal to C-21 and at the terminus of the side chain. Regions where steric bulk is disfavored (yellow) appear around the perimeter of the side chain.

Optimum 4D-QSAR Models. The R^2 and Q^2 for the optimum models determined for alignments 1, 3, and 5 are given in Table 6. The correlation equations representing the best 4D-QSAR models are given in Table 7. The cross-correlations of the residuals of fit among all pairwise

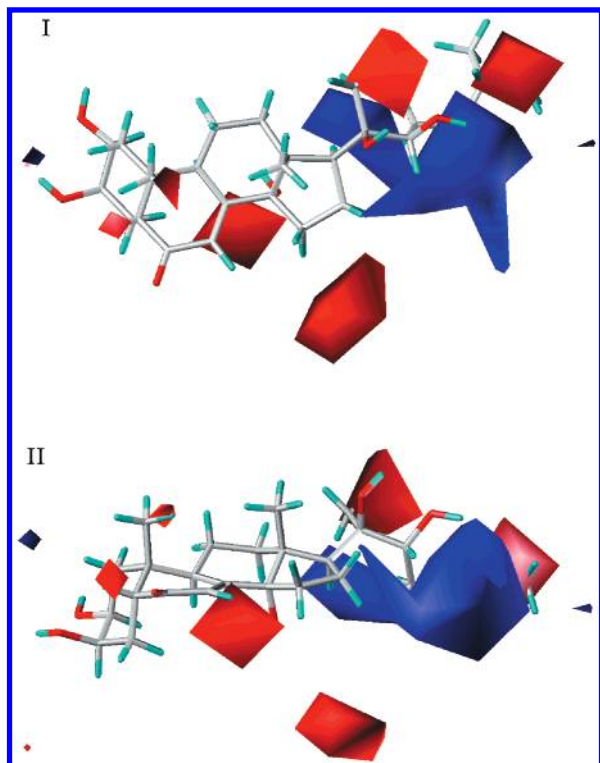


Figure 2. CoMFA electrostatic field contour plot ($SD \times \text{coef}$) for CoMFA model E (electrostatic indicator field). Blue polyhedra represent regions where positive charge is favorable; contribution level = 85%. Red polyhedra represent regions where negative charge is favorable; contribution level = 10%. Ponasterone A **9**, one of the most active ecdysteroids, is depicted. Panel I is a view from the β -face; panel II is a view toward the C-6/C-14 edge of the steroid.

combinations of these and the two CoMFA models are presented in Table 8. The residuals of all pairs of models are weakly cross-correlated to one another. This finding suggests that each of the 4D-QSAR models, and the two CoMFA models, are all relatively *independent* of one another.

Table 6 includes the optimum 4D-QSAR models with and without Mlog P included as a basis set descriptor in the GFA model building and optimization process. Alignments 2 and 4 were not considered for the entire training set because poor models were found using these alignments for subsets of the training set.

Alignment 5 does not yield a significant 4D-QSAR model for the entire training set based upon both its R^2 and Q^2 values (Table 6). Overall, both alignments 1 and 3 have been retained as possibly the preferred alignment. Both of these alignments largely involve placing the steroid ring system of the compounds upon one another, in contradistinction to the other alignments which involve overlap of substituent atoms and/or steroid ring system atoms. The inclusion of MlogP into the development of the 4D-QSAR models meaningfully increases both R^2 and Q^2 for the best models of alignments 1 and 3.

The residuals of fit of the top 10 4D-QSAR models identified for both alignments 1 and 3 are all highly cross-correlated to one another. Thus, each alignment is considered to have only one distinct 4D-QSAR model associated with it. Each of the best models for the two alignments has 14–15 GCODs. An evaluation of the cross-correlation matrix

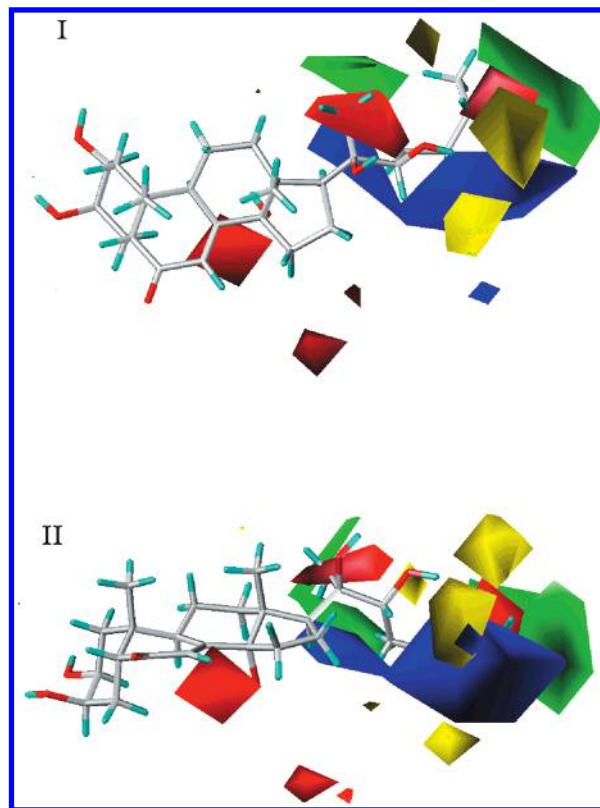


Figure 3. CoMFA electrostatic field contour plot ($SD \times \text{coef}$) for model ES (electrostatic and steric indicator fields). Blue polyhedra represent regions where positive charge is favorable; contribution level = 85%. Red polyhedra represent regions where negative charge is favorable; contribution level = 8%. Green polyhedra represent sterically favored regions; contribution level = 80%. Yellow polyhedra represent sterically disfavored regions; contribution level = 25%. Ponasterone A **9**, one of the most active ecdysteroids, is depicted. Panel I is a view from the β -face; panel II is a view toward the C-6/C-14 edge of the steroid.

of the GCODs of each of the best models indicates that no pair of GCODs are highly cross-correlated to one another. This finding suggests that each GCOD is making a distinct data fitting contribution to the 4D-QSAR model.

The majority of GCODs in all four 4D-QSAR models of Table 7 have either “any” or “nonpolar” IPE atom types. Moreover, the “any” IPE types often have negative regression coefficients, suggesting that these regions of space, for the given alignment, cannot be occupied by any type of ecdysteroid atom without a loss in activity. That is to say, these sites are predicted to be occupied by the receptor. The regression coefficients of the nonpolar IPE GCODs are both positive and negative. The negative regression coefficient for a given nonpolar GCOD probably means that some polar or hydrogen-bonding atom type should be at this GCOD site, but not enough data are available in the training set to make a selection of the specific IPE type. The positive regression for a nonpolar GCOD is taken to mean that this GCOD is near a hydrophobic region of the receptor binding site.

Graphical representations of the 4D-QSAR models relative to active and inactive analogues are given in Figures 4 and 5, respectively. Overall, an inspection of the visual representations of the models reveals that the GCODs are near C-2, C-3, C-11, C-12, C-20, and C-22 (Figure 4) and quite dense in the space of the substituent, R1. The active conformations of ponasterone A **9** (highest observed activity;

Table 7. Best 4D-QSAR Models as Represented by Their Respective Multidimensional, Linear Regression Equations

Part a	
Alignment 1 without Mlog P	
$-\log(\text{EC}_{50}) = 4.20\text{GCOD1}(-3,6,9,1) + 2.99\text{GCOD2}(-2,6,10,1) + -3.69\text{GCOD3}(1,6,10,0) +$	
$1.94\text{GCOD4}(-2,-1,3,1) + 4.26\text{GCOD5}(1,5,8,3) + -1.05\text{GCOD6}(1,2,9,1) +$	
$7.98\text{GCOD7}(-2,-4,0,3) + 3.94\text{GCOD8}(-2,8,10,0) + 4.85\text{GCOD9}(0,0,6,1) +$	
$3.10\text{GCOD10}(0,2,8,0) + -6.36\text{GCOD11}(0,7,9,4) + 3.19\text{GCOD12}(3,4,7,5) +$	
$2.09\text{GCOD13}(1,-2,2,0) + -1.62\text{GCOD14}(0,4,8,1) + -0.527$	
$N = 67$	eq 1
Alignment 1 with Mlog P	
$-\log(\text{EC}_{50}) = 2.53\text{GCOD1}(-2,6,10,1) + -2.18\text{GCOD2}(0,-1,3,0) + 6.47\text{GCOD3}(-1,7,10,1) +$	
$5.52\text{GCOD4}(-3,8,10,0) + 4.30\text{GCOD5}(-4,5,8,1) + -8.15\text{GCOD6}(-2,-4,2,0) +$	
$1.23\text{GCOD7}(1,5,9,3) + 5.74\text{GCOD8}(-2,-5,2,0) + -2.04\text{GCOD9}(2,4,6,0) +$	
$-2.42\text{GCOD10}(0,-1,7,5) + -6.54\text{GCOD11}(0,7,9,4) + 1.48\text{GCOD12}(3,4,7,2) +$	
$-2.35\text{GCOD13}(0,4,8,1) + 1.52\text{GCOD14}(-2,6,8,1) + 6.50$	
$N = 67$	eq 2
Part b	
Alignment 3 without Mlog P	
$-\log(\text{EC}_{50}) = -3.66\text{GCOD1}(-1,-1,-2,0) + -4.07\text{GCOD2}(8,8,-6,0) + -7.22\text{GCOD3}(3,7,-4,1) +$	
$-3.60\text{GCOD4}(1,2,-3,1) + -4.32\text{GCOD5}(1,1,2,1) + -5.60\text{GCOD6}(6,6,-4,0) +$	
$6.36\text{GCOD7}(2,5,-2,1) + 4.89\text{GCOD8}(9,7,-2,0) + 2.59\text{GCOD9}(7,6,-4,0) +$	
$3.86\text{GCOD10}(7,8,-4,0) + 2.36\text{GCOD11}(9,6,-3,1) + 2.20\text{GCOD12}(0,3,-3,0) +$	
$3.99\text{GCOD13}(6,8,-5,0) + 4.55\text{GCOD14}(0,-1,-1,4) + 1.34\text{GCOD15}(0,1,3,1) + 4.51$	
$N = 67$	eq 3
Alignment 3 with Mlog P	
$-\log(\text{EC}_{50}) = -7.74\text{GCOD1}(8,8,-6,0) + -7.72\text{GCOD2}(9,6,-2,1) + 5.55\text{GCOD3}(5,5,-4,1) +$	
$-7.91\text{GCOD4}(5,5,-2,1) + 3.82\text{GCOD5}(-2,-1,2,0) + 8.10\text{GCOD6}(9,8,-5,1) +$	
$-5.83\text{GCOD7}(9,7,-3,1) + -9.99\text{GCOD8}(1,1,2,1) + -8.97\text{GCOD9}(6,6,-4,0) +$	
$5.89\text{GCOD10}(4,5,-5,0) + 12.09\text{GCOD11}(9,6,-3,0) + 5.17\text{GCOD12}(3,7,0,0) +$	
$3.38\text{GCOD13}(3,5,-1,1) + 4.98\text{GCOD14}(7,7,-2,1) + 7.39\text{GCOD15}(0,-1,-1,3) + 4.59$	
$N = 67$	eq 4

Table 8. Cross-Correlation Coefficients of the Residuals of Fit for Each of the Four Best 4D-QSAR and Two Best CoMFA Models for the Training Set

alignment 1						
alignment 1 with MlogP	0.67					
alignment 3	0.42	0.31				
alignment 3 with MlogP	0.082	0.27	0.31			
CoMFA E	0.34	0.16	0.32	0.049		
CoMFA E/S	0.44	0.27	0.41	0.29	0.68	
	alignment 1	alignment 1 with MlogP	alignment 3	alignment 3 with MlogP	CoMFA E	CoMFA E/S

Figure 4) and that of inactive 20-*iso*-22-*epi*-ecdysone **40** (Figure 5) have been specifically used as reference structures in the display the GCODs of the 4D-QSAR models.

An inspection of Figure 4 for the four 4D-QSAR models shown relative to ponasterone A **9** in its postulated active conformation for each respective model reveals GCODs near the C-2 hydroxyl group. Furthermore, these GCODs indicate that occupancy by a polar acceptor or hydrogen-bonding acceptor increases activity, or occupancy by the "any" atom type decreases activity. This configuration of GCODs suggests that the C-2 hydroxyl acts as a hydrogen-bond acceptor with respect to the receptor cavity and that there is a strict space requirement in order for it to do so. GCODs around the C-3 hydroxyl suggest the need for substitution at this position, and perhaps preferably, a polar negative atom. Three of the four models have GCODs near C-12. Two of these GCODs have a nonpolar IPE type with a positive coefficient, and one has a hydrogen bond donor IPE type with a negative coefficient. These features suggest that the C-12 region fits to a hydrophobic area of the receptor. The region near C-11 is more difficult to interpret.

The active conformations of ponasterone A **9** from the models developed for alignment 1 place a positive occupancy hydrogen-bond donor and a polar positive GCOD near the

C-20 hydroxyl hydrogen as well as a GCOD corresponding to maintaining an open ligand steric space near this hydroxyl group. Therefore, the alignment 1 models would suggest that the C-20 hydroxyl is involved in hydrogen-bond donation. In contrast, the predicted active conformations of alignment 3 correspond to the C-20 hydroxyl not being located near any type of GCOD. There is no GCOD associated with O-6 in any of the models, and the hydroxyl of C-14 is largely devoid of any nearby GCODs. Thus, in alignment 3, O-6 as well as the C-20 and C-14 hydroxyl groups are, *as a consequence of the selection of alignment 3*, predicted to be either (1) not involved in receptor binding or (2) involved in receptor binding to the same extent across all compounds of the training set.

Finally, the remainder of the GCODs from all four models are distributed about the R1 substituent. The best way to characterize these GCODs is by first noting that the associated IPEs are either "any" or "nonpolar". Second, the predicted active conformations of ponasterone A **9**, for all four 4D-QSAR models, have R1 substituents as largely extended conformations directed away from the steroid ring system. The GCODs bundled about R1 map out constraints in space for optimizing activity which lead to nonfolded conformations of the substituent. However, the active

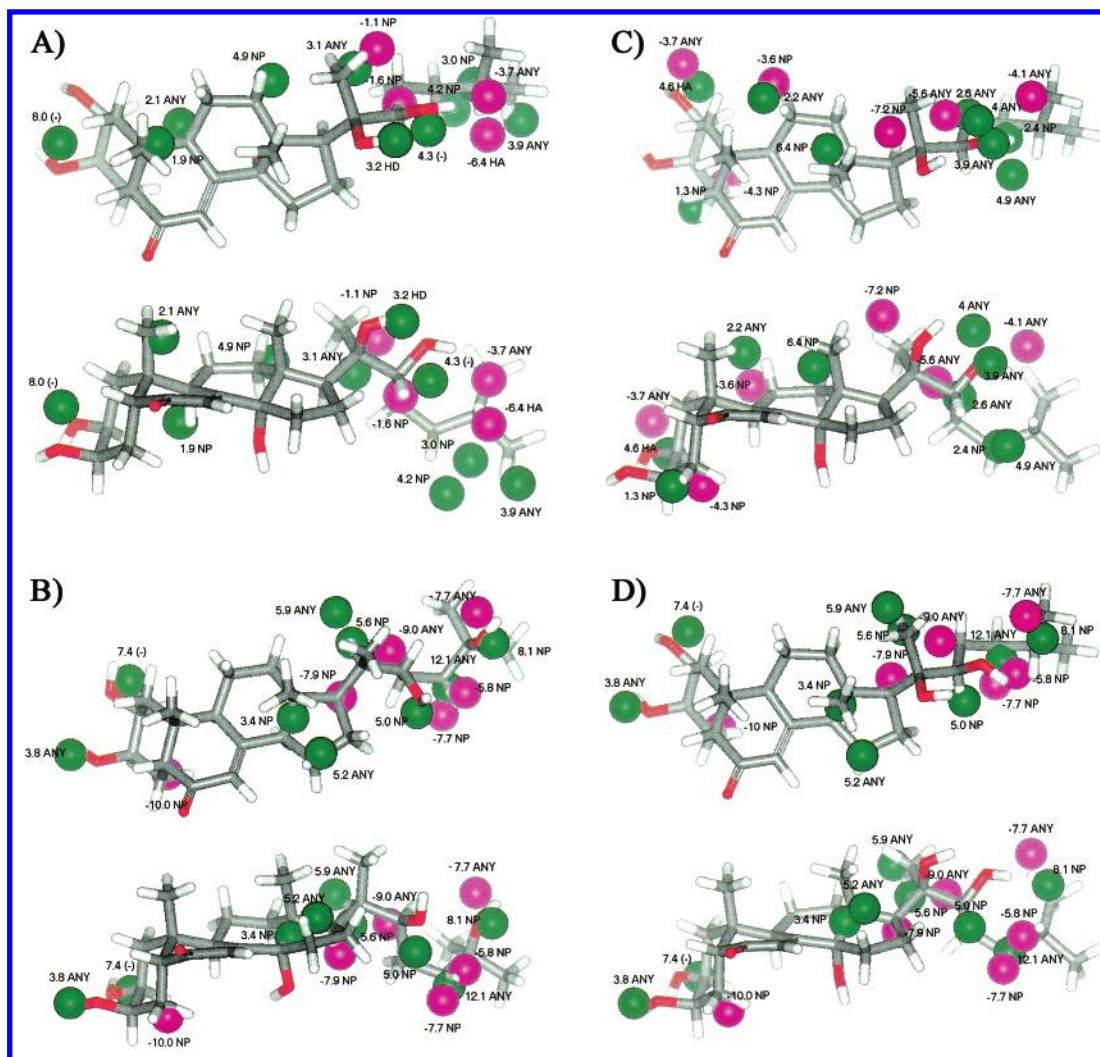


Figure 4. Representations of the four 4D-QSAR models relative to Ponasterone A **9** ($-\log[EC_{50}] = 9.5$) in its postulated active conformation with respect to each model: A) alignment 1 without MlogP, B) alignment 1 with MlogP, C) alignment 3 without MlogP and D) alignment 3 with MlogP. The GCODs (grid cell occupancy descriptors) are shown as spheres. Positive values of a GCOD indicates occupancy by the appropriate IPE increases activity, while a negative value indicates occupancy decreases activity. IPE type abbreviations: NP = nonpolar, (+) = polar positive, (−) = polar negative, HA = hydrogen bond acceptor, HD = hydrogen bond donor. The top structure in each panel is a view from the β -face; the bottom structure in each panel is a view toward the C-6/C-14 edge of the steroid.

conformations are not simply fully extended geometries. The conformation about bond C-20–C-22, for example, places the two hydroxyl groups midway between gauche and cis to one another. This local conformation seems to be stabilized by a weak hydrogen bond between the two hydroxyls, in which the C-20 hydroxyl donates and the C-22 hydroxyl accepts (however, molecular dynamics indicates role reversal is possible). Overall, the predicted active conformations of R1 for ponasterone A **9** from all four 4D-QSAR models suggest a receptor cavity for this substituent which is a cylinder-shaped hole largely lined with nonpolar residues. However, there is a hydrogen-bond donor and an acceptor near the top of this receptor cylinder which hydrogen-bond to one or other of the C-20 and C-22 hydroxyls, depending on the precise orientation of these two hydroxyls with respect to each other.

The relationships of the predicted active conformations of 20-*iso*-22-*epi*-ecdysone **40**, which is relatively inactive, to the GCODs of the corresponding 4D-QSAR models (Figure 5) are markedly different from those of ponasterone A **9**. As a consequence of the alignment rule, the GCODs near the C-2 and C-3 hydroxyls of **9** are of course retained

for **40**. However, the C-22 hydroxyl of **40** are not near the polar and/or hydrogen bonding IPE type GCODs important for defining the role of the C-22 hydroxyl of **9**. Furthermore, three of the four predicted active conformations of the R1 substituent correspond to semifolded chains such that most of the substituent is distant from the collection of GCODs that define the conformational requirements of the R1 substituent for activity. Many of these GCODs, if occupied, are predicted to enhance activity but are obviously not occupied for the semifolded active conformations. The predicted active conformation of the alignment 3 without MlogP model is a semiextended conformation similar to those of ponasterone A **9**. However, this conformation corresponds to some of the substituent atoms occupying two GCODs whose occupancies diminish activity.

Thus, it appears that the R1 side chain of 20-*iso*-22-*epi*-ecdysone **40** cannot adopt a predicted active conformation for any of the four 4D-QSAR models which does not occupy GCODs which diminish activity, while at the same time occupying GCODs which enhance activity. Moreover, there is no GCOD in any of the models that can be positively related to the presence of the terminal hydroxyl group of

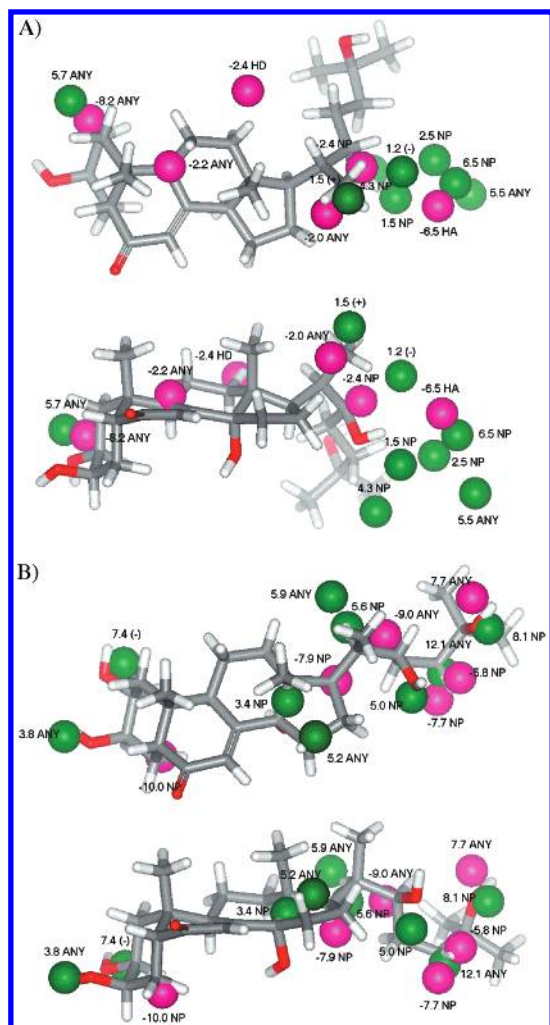


Figure 5. Same as Figure 4, but for 20-iso-22-epi-ecdysone **40** ($-\log[\text{EC}_{50}] = 4.0$) in its postulated active conformation: A) alignment 1 with MlogP and B) alignment 3 with MlogP. The top structure in each panel is a view from the β -face; the bottom structure in each panel is a view toward the C-6/C-14 edge of the steroid.

R1 for steroid **40**. The overall result is a compound predicted to have a low EC_{50} , as is observed.

Outliers and Test Set Predictions. There are four outliers in the training set which span across all four 4D-QSAR models. These outliers are compounds **22**, **30**, and **63**, all of which are quite active, and **43**, which is relatively inactive. Each of these outliers has some distinct structural features relative to those of the entire training set. Compound **22** has added bulk at C-24 as well as a hydroxyl at C-25 which may impart additional activity relative to that specified by any of the 4D-QSAR models. Both steroids **30** and **63** have C-5 and C-25 hydroxyls. These are the only two steroids in the training set with this particular combination of hydroxyl groups. Steroid **43** is the only compound in the training set with a glucosyl group at C-3. This group could be the source of low activity observed for this compound.

The observed, predicted, and residual activities of the 20 steroids in the test set using the four 4D-QSAR models constructed from alignments 1 and 3 with, and without, MlogP as a GFA basis set descriptor are given as part of Table 9. The last row of Table 9 contains the average residual of prediction for 17 of the 20 test compounds. On the basis

that the smallest average residual of fit corresponds to the most significant model, the 4D-QSAR models for alignment 1, including and excluding MlogP, are equally the best overall models. The average residuals of fit for the test compounds of the alignment 1 4D-QSAR models is about twice as large as the average residuals of fit for the training set (av resid. fit: align. 1 (0.69 vs 0.36), align. 1/MlogP (0.69 vs 0.35)). For the alignment 3 models and for the CoMFA models, this ratio is greater (av resid. fit: align. 3 (0.89 vs 0.41), align. 1/MlogP (1.13 vs 0.35), CoMFA E (1.47 vs 0.32), CoMFA ES (1.47 vs 0.27)).

Comparison of the 4D-QSAR and CoMFA Models. A statistical summary comparison of the best CoMFA and 4D-QSAR models for the training set is presented in Table 6. Statistical significance of the CoMFA and 4D-QSAR models are quite similar as measured by R^2 , with the CoMFA models being slightly better, but the Q^2 values of the 4D-QSAR models are appreciably higher than those from the CoMFA study. The statistical outcome of a CoMFA model can be quite sensitive to training set alignment and conformation. With respect to the ecdysteroid training set, a completely satisfactory alignment is very difficult to achieve due to the substantial structural variability in the side chains.

Cross-correlation coefficients of the residuals of fit among pairs of the CoMFA and/or the 4D-QSAR models are given in Table 8. Perhaps most striking in Table 8 are the generally low absolute values of the cross-correlation coefficients. From these one may infer that no pair of models “explains” the structure–activity data in the same way. The highest cross-correlation coefficient between 4D-QSAR and CoMFA models is 0.44. Thus, while all of the best QSAR models have reasonably similar R^2 values, their respective individual predictions are significantly different from one another. This behavior of near equal model quality, but variability in individual model predictions, suggests that consensus modeling, using *all* the best QSAR models, may be the optimum route to take to realize the most robust predictions.

Both the CoMFA and (alignment 1) 4D-QSAR models identify the C-20 and C-22 hydroxyl groups as being significant. Regions near these groups are included as negative charge favored regions by the CoMFA models, while the 4D-QSAR models favor the occupation of GCODs in this area by polar atoms and/or hydrogen bond donors/acceptors. In some respects, the 4D-QSAR models are more structure–activity specific than CoMFA models. The 4D-QSAR models require that not just any atom be attached to C-20 (GCOD9 of alignment 1 with MlogP), but rather that this atom must be bonded to a polar positive atom (e.g., GCOD12 of alignment 1 with MlogP). Moreover, this polar positive atom must be a hydrogen-bond donor (GCOD12 of alignment 1 without MlogP). Also specified is a polar negative atom at the C-22 site (GCOD5 of alignment 1 without MlogP and GCOD7 of alignment 1 with MlogP). It is generally well-known that ecdysteroid activity is sensitive to the oxidation state and configuration at the C-20 and C-22 centers. Although both the CoMFA and 4D-QSAR models do not favor negative charge density along the distal portion of the side chain, the 4D-QSAR models also specifically disfavor occupation by a hydrogen-bond acceptor at the C-25 position (GCOD11 of alignment 1 without MlogP).

Both CoMFA and 4D-QSAR models show sensitivity to changes in substituent geometry at both C-20 and C-22, as

Table 9. Test Set Ecdysteroids and the Observed $-\log(\text{EC}_{50})$ Values, Predicted $-\log(\text{EC}_{50})$ Values, and Residuals of Prediction, in “()” for the Four Best 4D-QSAR Models and CoMFA Models^a

no.	compound	obsd activity $-\log(\text{EC}_{50})$	predicted activity					
			align. 1	align. 1 with Mlogp	align. 3	align. 3 with Mlogp	CoMFA E	CoMFA ES
72	punisterone	6.08	6.08 (-0.93)	7.01 (-0.95)	7.89 (-0.81)	7.27 (-1.19)	5.87 (0.21)	8.36 (-2.28)
73	makisterone C	6.70	6.31 (0.39)	6.55 (0.15)	6.92 (-0.22)	7.34 (-0.64)	7.37 (-0.67)	8.85 (-2.15)
74	5 β -hydroxyabutasterone	7.64	6.17 (1.49)	7.00 (0.64)	7.92 (-0.28)	8.66 (-1.02)	7.31 (0.33)	7.96 (-0.32)
75	20-hydroxyecdysone 3 β -D-xylopyranoside	5.8	6.42 (-0.62)	6.00 (-0.20)	7.52 (-1.72)	7.29 (-1.49)	3.18 (2.62)	4.99 (0.81)
76	ponasterone A 3 β -D-xylopyranoside	5.8	6.31 (-0.51)	5.58 (0.22)	7.08 (-1.28)	7.89 (-2.09)	2.85 (2.95)	3.53 (2.27)
77	rubrosterone	< 4	3.78	5.26	6.07	5.00	1.02	1.96
78	dihydorubrosterone 17-acetate	< 4	3.43	4.43	6.16	4.37	5.58	7.28
79	(20R)22-deoxy-20,21-dihydroxyecdysone	6.70	6.85 (-0.15)	5.75 (0.95)	6.71 (-0.01)	7.00 (-0.30)	5.60 (1.10)	7.05 (-0.35)
80	2-dehydro-3- <i>epi</i> -20-hydroxyecdysone	6.40	5.52 (0.88)	5.79 (0.61)	8.48 (-2.08)	9.07 (-2.67)	9.13 (-2.73)	5.07 (1.33)
81	24- <i>epi</i> -abutasterone	5.92	6.09 (-0.17)	5.43 (0.49)	6.32 (-0.41)	8.11 (-2.19)	7.49 (-1.57)	7.96 (-2.04)
82	2-deoxy-20-hydroxyecdysone 22-acetate	4.62	6.34 (-1.72)	6.10 (-1.48)	6.26 (-1.64)	4.86 (-0.24)	4.53 (0.09)	7.49 (-2.87)
83	2-deoxy-20-hydroxyecdysone 3,22-diacetate	< 4	6.97	5.89	6.46	2.74	3.15	6.13
84	2-deoxy-20-hydroxyecdysone 3-acetate	5.37	5.819 (-0.44)	4.58 (0.79)	5.71 (-0.34)	6.49 (-1.12)	6.18 (-0.81)	6.93 (-1.56)
85	2-deoxy-21-hydroxyecdysone	5.37	5.46 (-0.09)	5.06 (0.31)	6.13 (-0.76)	6.26 (-0.89)	2.71 (2.66)	4.98 (0.39)
86	(5 α -H)2-deoxy-21-hydroxyecdysone	4.02	6.15 (-2.13)	6.28 (-2.26)	4.35 (-0.33)	3.74 (0.28)	3.56 (0.46)	4.89 (-0.87)
87	20-hydroxyecdysone 22-acetate	5.40	5.36 (0.04)	5.45 (-0.05)	7.11 (-1.71)	6.44 (-1.04)	5.50 (-0.10)	8.25 (-2.85)
88	3- <i>epi</i> -20-hydroxyecdysone	6.89	6.50 (0.39)	6.30 (0.59)	6.40 (0.49)	6.43 (0.46)	7.84 (-0.95)	6.74 (0.15)
89	(5 α -H)dihydorubrosterone	5.25	3.09 (2.16)	5.47 (-0.24)	5.88 (-0.63)	3.41 (1.84)	2.31 (2.94)	3.12 (2.13)
90	(5 β -H)dihydorubrosterone	≤ 5.0	5.24 (-0.24)	4.83 (0.17)	5.63 (-0.63)	3.29 (1.71)	3.23	3.54
91	20-hydroxyecdysone 22-benzoate	8.37	7.01 (1.36)	6.86 (1.51)	7.63 (0.74)	9.02 (-0.65)	4.98 (3.39)	7.21 (1.16)
average residuals of fit			0.69	0.69	0.89	1.13	1.47	1.47

^a The last row of the table contains the average residuals of fit for the predictions of each of the four models.

seen by comparing the predicted $-\log(\text{EC}_{50})$ values for steroids **2** and **38**. The only difference between 22-dehydroecdysone **2** and 22-dehydro-20-*iso*-ecdysone **38** is the 20R- and 20S-carbon configuration, respectively, which results in a lower predicted $-\log(\text{EC}_{50})$ for **38**. The CoMFA models suggest that loss in potency results from oxygen substitution at C-24, C-26, and C-27. This is corroborated by the 4D-QSAR models which favor occupation of GCODs in this region by nonpolar entities. Generally, both CoMFA and 4D-QSAR models place restrictions on steric bulk around the length of the side chain. However, when compared to CoMFA, occupancy of favorable GCODs in the 4D-QSAR models is limited to small regions of space around the side chain, which can be attained by few conformations of a given molecule. Among the five alignments explored in this 4D-QSAR study, alignment 1 closely resembles the CoMFA alignment, which is based on rms fit of the atoms comprising the steroid core. 4D-QSAR alignment 3 bears a near resemblance to the CoMFA alignment. It is these two alignments which result in the most meaningful 4D-QSARs. This coincidence of 4D-QSAR alignments and CoMFA alignment is consistent with the viewpoint that the ecdysteroid align-

ment in the ligand binding domain of *EcR* is an alignment in which the steroid core atoms overlap and the chain atoms take on more disparate conformations. Thus, the CoMFA results concur with 4D-QSAR analysis in suggesting a side-chain cavity which is substantially lipophilic and which requires a fairly tight side-chain fit, but the CoMFA and 4D-QSAR models differ in localized electrostatic requirements.

Outlier features illustrate the individual 4D-QSAR analysis and CoMFA tendencies. Steroids 24(28)-dehydromakisterone A **22**, polypodine B **30**, and rapisterone D **63** are outliers in all 4D-QSAR models and at least one CoMFA model and are underestimated. Polypodine B **30** and rapisterone D **63** share the feature of hydroxyl substitution at C-5. Low predictions for dehydromakisterone A **22** may result from methodological causes, such as poor choice of the CoMFA conformations, or fortuitously inappropriate 4D-QSAR conformational sampling, but these factors are difficult to isolate. Other outlying structures are instructive because of differences in the way CoMFA and 4D-QSAR models interpret them. Malacosterone **60**, for example, is seriously overestimated by both CoMFA models but accurately predicted by the 4D-QSAR models. The CoMFA models do not detect

the unique 16-OH group, while the 4D-QSAR method may perceive the nonbonding effects of the 16-OH group on the conformational space of the side chain. Occupation of the regions of space mapped out by the 4D-QSAR models by favorable and/or unfavorable pharmacophores is often dictated by the preferred conformations of the molecule. Conformational preferences can, therefore, influence the predicted $-\log(\text{EC}_{50})$ values, thereby resulting in poor correlation of the 4D-QSAR residuals with those of CoMFA models where molecular conformation is fixed. As a case in point, the predicted active conformations for ponasterone A **9** and 5-deoxykaladasteone **70** by the 4D-QSAR models closely resemble the static conformations used in the CoMFA models, and both are similar to the crystal structure of ecdysone itself, differing by only slight adjustments in the torsions defining the terminal four or five heavy atoms of the R1-chain. On the other hand, the active conformation of 20-*iso*-22-*epi*-ecdysone **40** differs quite substantially from that used in the CoMFA study. The CoMFA conformation places the length of the chain in approximately the same position as all other congeners but swaps the C-21 and O-20 positions. 4D-QSAR analysis predicts an active conformation in which C-21 and O-20 assume similar positions and the chain points off in an atypical direction, the "typical" conformation being more than 2 kcal/mol less stable than the minimum energy conformer state.

Both the CoMFA and the 4D-QSAR models predict the low $-\log(\text{EC}_{50})$ of **40** reasonably well but for very different reasons. In the CoMFA case, the fit is accurately low because of the failure to occupy an important negative charge region, and in the 4D-QSAR case, the prediction is appropriately low owing to the failure to occupy a favorable hydrogen-bond acceptor/polar negative region AND avoid detrimental steric regions at the same time. The active conformation of **40** is merely one snapshot in which the lesser of the two negative features is implied.

DISCUSSION

A qualitative form of *consensus* QSAR modeling has been implicitly employed to develop the salient features of the structure-activity relationship embedded in the training set. Common GCOD features are sought across all four of the optimum 4D-QSAR models and compared to the field features of the CoMFA models.

A good active candidate molecule is one that has features in common with steroids **9** and **70**. In particular, it appears that the C-2/C-3 as well as the C-20/C-22 hydroxyl groups are necessary for favorable activity, while a C-25 hydroxyl is unfavorable. Conformational and/or steric restrictions along the R1 side chains are found in each of the 4D-QSAR models. Each of the four 4D-QSAR models make unique contributions to the composite set of side-chain pharmacophore GCODs while at the same time sharing common pharmacophore elements, that is, GCODs.

The structure-activity behavior at positions C-20/C-22 is perhaps most emphasized by the 4D-QSAR models. GCOD12 of both models based on alignment 1 is particularly interesting. Steroids **1**–**8**, having the ecdysone core in common, all have a hydrogen at the C-20 position. Among these structures, 22-dehydroecdysone **2** and 25-deoxyecdysone **4**

exhibit moderately high $-\log(\text{EC}_{50})$ values, whereas the other compounds of this subset are not very active. Steroid **4** is unique in this group because it is the only one containing a C-22 hydroxyl but no hydroxyl substitution at the C-25 position. Focusing for the moment on the C22 position, steroid **2** contains a keto oxygen at this site. Perhaps the C-20 hydrogen interacts favorably with the keto oxygen of steroid **2** as well as the hydroxyl oxygen of steroid **4**. Both the keto and the hydroxyl oxygens occupy GCOD5 of alignment 1 without MlogP and GCOD7 of alignment 1 including MlogP models, making the C-20 hydrogen a "pseudo hydrogen-bond donor" occupying GCOD12 of alignment 1. By contrast, C-20 of 22-dehydro-2-*iso*-ecdysone **38** has an *S*-configuration, preventing occupation of GCOD12 by the hydrogen. The only difference between steroid **2** and **38** is the 20*R*- and 20*S*-carbon configurations, respectively, accounting for the very low activity of steroid **38** in comparison with compound **2**. Similar C-20 configurational differences between steroids **1** and **39** and steroids **3** and **40** result in the lower activities of steroids **39** and **40** relative to steroids **1** and **3**, respectively. In each case the *S*-configuration at C-20 does not allow a favorable interaction between the C-20 hydrogen and the C-22 oxygen. Another pair of steroids, 2-deoxyecdysone **51** and 2-deoxy-20-hydroxyecdysone **52**, also emphasize the crucial activity roles of the C-20 hydrogen-bond donor, as characterized by GCOD12 of alignment 1. Steroid **52** contains the C-20 hydroxyl and is more active than steroid **51**, in which this group is absent.

Overall, the 4D-QSAR models of alignment 1, with and without MlogP, appear to give the best predictions for the test set molecules as reported in Table 9 for both the best 4D-QSAR and CoMFA models.

The CoMFA models have a large amount of electrostatic character governing the prediction of $-\log(\text{EC}_{50})$. Conversely, the 4D-QSAR models are rich in nonpolar and "any atom" GCODs. This apparent dichotomy may have its origins in the manner in which electrostatic and hydrogen bonding descriptors are represented in the two methodologies. CoMFA uses a Coulomb-like function to model electrostatic and hydrogen bonding fields.^{39–41} The distance-dependence (r) between a test molecule atom and the field probe is $(1/r)$. This functional relationship leads to *large regions of space, or many field probe grid cells*, to represent the field potentials of polar and/or hydrogen bonding atoms. Conversely, a 4D-QSAR analysis identifies *a single, or at most, two GCODs*, to represent polar and/or hydrogen bonding IPEs. In some cases an additional GCOD having a negative regression coefficient and an "any atom" IPE is located close to a polar and hydrogen bonding GCOD. This grouping of GCODs occurs in some of the 4D-QSAR models of this study and is interpreted as a requirement for free space about a polar and hydrogen bonding atom in order to complete the interaction with the receptor. The salient point, however, is that 4D-QSAR analysis employs a very few number of GCOD descriptors to represent polar and hydrogen bonding sites, while CoMFA, because of the form of the field function, employs a large number of field cell descriptors. On the other hand, 4D-QSAR analysis tends to use a relatively large number of GCODs to represent nonpolar (hydrophobic) surface sites on a ligand. The actual number of GCODs is usually proportional to the size of the nonpolar

site. In this study, for example, a large fraction of the GCODs in the 4D-QSAR models define the surface and corresponding shape for the often nonpolar R1 side chain.

Outliers common to both the 4D-QSAR and CoMFA models are revealing. Polypodine B **30** and rapisterone D **63** share the feature of hydroxyl substitution at C-5, and, in addition, the C-2 and C-3 configurations of rapisterone D are inverted. Since the two independent QSAR approaches of 4D-QSAR and CoMFA similarly underpredict these structures, this result suggests that these two steroids bind to the *EcR* LBD in a different orientation relative to the other ecdysteroids and/or that associated molecular events affected by ligand binding, such as heterodimerization or repressor/coactivator exchange, take place more efficiently. By analogy with mammalian nuclear receptors, this effect might be mediated through the AF-2 region and helix 12.⁵⁸

A final exploration of the relationship between the CoMFA and 4D-QSAR analysis for this training set consists of two studies. One study involves the use of the predicted active conformations and preferred alignment from the 4D-QSAR analysis in another CoMFA study. The objectives are to see if a better CoMFA model is generated and to what extent the new model differs from the original CoMFA model. Toward this goal, CoMFA models were explored based on alignment 1 (rms fit of carbon atom 6, 12, and 16) and the predicted active conformations for the alignment 1 4D-QSAR model. Electrostatic, steric, and hydrogen-bonding fields were explored, both in the standard and indicator format. Without removal of outliers, on the basis of Q^2 , the best model obtained for the 71 member training set yielded $R^2 = 0.966$, but $Q^2 = 0.423$ for five PLS components. Region-focused indicator electrostatic and steric fields as well as hydrogen-bonding fields were utilized, each in roughly equal proportions on a 1 Å grid with C(sp³+) as the probe atom. The model was not at all sensitive to removal of outliers. Even this (almost certainly overfitted) model supports the generally recognized observation that the CoMFA methodology is extremely sensitive to precise alignment of structures. The 4D-QSAR alignment rule applied here as the CoMFA alignment, but more particularly the scatter in 4D-QSAR active conformations, may not constitute a suitable CoMFA "starting point". Better results may perhaps be attained by application of wider grid spacing and/or use of smoother Gaussian functional forms of the electrostatic and steric potentials inherent in the CoMSIA approach.⁵⁹

A second study would be to use the GCODs from the 4D-QSAR analysis as external descriptors in the revised CoMFA study to see if a combination of CoMFA and 4D-QSAR descriptors could lead to yet a better and more insightful model. This study has not been done to date.

Very recently, Wurtz et al.^{60,61} have proposed two models for agonist interaction with the ligand-binding domain of a dipteran (*Chironomus tentans*) *EcR*. The interactions are derived from docking studies of 20E and the diacylhydrazine RH-5849 with homology-modeled ligand binding domains, which, in turn, are based on the crystal structures for the ligand binding domains of the human retinoic acid receptor (hsRAR γ) and the human vitamin D receptor (hsVDR).^{62,63} Between these two models, the orientation of 20E differs by roughly 180°, thereby altering the interactions between the functional groups on 20E and the residues lining the ligand-binding pocket. These models partially explain the

observed biological activities of limited sets of ecdysteroid or bisacylhydrazine analogues. Comparison of the 4D-QSAR and CoMFA ecdysteroid models with these homology models is instructive. Together, the receptor-independent 4D-QSAR and CoMFA models imply a sterically restrictive hydrophobic cylinder around the steroid side chain. This is consistent with 20E-liganded homology models based on either the retinoic acid receptor (RAR) or the vitamin D receptor (VDR) as well as the generally hydrophobic LBD characteristics of known nuclear-receptor LBDs on which these models are based.⁶⁰ Another significant 4D-QSAR consensus conclusion is the hydrogen-bond acceptor characteristic of 22-OH. In the 20E-liganded homology model based on the vitamin D receptor (VDR), this hydroxyl is also a hydrogen-bond acceptor. Yet a third 4D-QSAR conclusion is the hydrogen-bond donor characteristic of 20-OH. The 20E-liganded homology model based on the retinoic acid receptor (RAR) predicts hydrogen bonding to Met368 S, which by necessity would render the 20-OH a hydrogen-bond donor, in agreement with the 4D-QSAR models. In the VDR homology model, the 20-OH group is a hydrogen-bond acceptor with Asn422-NH₂, but 20-OH may also participate with Asn422-C γ O as donor. Both Met368 and Asn422 are conserved residues in the *Drosophila EcR*. In making these 4D-QSAR/homology model comparisons, it is important to bear in mind that although a specific hydrogen bonding role can be deduced from 4D-QSAR GCODs, the intra- vs intermolecular nature of the hydrogen bonding interaction is not defined, and the possibility of a dual role for a given hydroxyl group may not be excluded.

On the other hand, the homology models predict that all six hydroxyls and the C-6 carbonyl of 20E are implicated in H-bonding, whereas the receptor-independent 4D and CoMFA models are more selective. Specifically, the homology models predict hydrogen-bond participation of the 14-OH group, whereas both 4D-QSAR and CoMFA models do not recognize this position in any significant way, despite structural variation at C-14 in the training set. Both homology models also consider 25-OH as a hydrogen-bond participant, but 4D-QSAR specifically disfavors a hydrogen-bond acceptor at this position, and the CoMFA models lack an electrostatic region at 25-OH, but rather one is shifted to C-26. Since the presence of a C-25 hydroxyl is known to reduce the biological potency of ecdysteroids in all insect and crustacean systems studied to date, it is difficult to reconcile this finding with either of the homology models. Other positions such as 6-C=O, 2-OH, and 3-OH are hydrogen-bond participants in one or more of the QSAR and homology models but differ in the acceptor/donor characteristics.

ACKNOWLEDGMENT

The UIC authors would like to acknowledge our coauthors for bringing the data set used in this study to our attention. Resources of The Laboratory of Molecular Modeling and Design at UIC were employed in performing this research. We appreciate the helpful and stimulating discussions with Dr. José Duca and Jane Tseng of UIC over the course of this study. L.D. wishes to acknowledge the financial support of the Biotechnology and Biological Sciences Research Council, Rohm & Haas Co. and EU-INTAS (Contract No. 96-1291), the excellent technical assistance of Pensri Whit-

ing, and the many researchers who provided samples of ecdysteroid analogues.²⁵

REFERENCES AND NOTES

- Gomez-Navarro, J.; Curiel, D. T.; Douglas, J. T. Gene therapy for cancer. *Eur. J. Cancer* **1999**, *35*, 2039–2057.
- Costantini, L. C.; Bakowska, J. C.; Breakefield, X. O.; Isacson, O. Gene therapy in the CNS. *Gene Ther.* **2000**, *7*, 93–109.
- Yla-Herttuala, S.; Martin, J. F. Cardiovascular gene therapy. *Lancet* **2000**, *355*, 213–222.
- Flotte, T. R. Gene therapy for cystic fibrosis. *Curr. Opin. Mol. Ther.* **1999**, *1*, 510–518.
- Tsokos, G. C.; Nepom, G. T. Gene therapy in the treatment of autoimmune diseases. *J. Clin. Invest.* **2000**, *106*, 181–183.
- Mountain, A. A. Gene therapy: the first decade. *Trends Biotechnol.* **2000**, *18*, 119–128.
- French Anderson, W. Human gene therapy. *Nature* **1998**, *392* (6679, suppl.), 25–30.
- Bestor, T. H. Gene silencing as a threat to the success of gene therapy. *J. Clin. Invest.* **2000**, *105*, 409–411.
- Pouton, C. W. Biological barriers to gene transfer. *Drug Targeting Delivery* **1999**, *10*, 65–102.
- Zuo, J.; Chua, N.-H. Chemical-inducible systems for regulated expression of plant genes. *Curr. Op. Biotech.* **2000**, *11*, 146–151.
- Agha-Mohammadi, S.; Lotze, M. T. Regulatable systems: applications in gene therapy and replicating viruses. *J. Clin. Invest.* **2000**, *105*, 1177–1183.
- Clackson, T. Regulated gene expression systems. *Gene Therapy*, **2000**, *7*, 120–125.
- Allgood, V. E.; Eastman, E. M. Chimeric receptors as gene switches. *Curr. Opin. Biotechnol.* **1997**, *8*, 474–479.
- Lezzi, M.; Bergman, T.; Mouillet, J.-F.; Henrich, V. C. The ecdysone receptor puzzle. *Arch. Insect Biochem. Physiol.* **1999**, *41*, 99–106.
- Yao, T.-P.; Forman, B. M.; Jiang, Z.; Cherbas, L.; Chen, J.-D.; McKeown, M.; Cherbas, P.; Evans, R. M. Functional ecdysone receptor is the product of EcR and Ultraspiracle genes. *Nature* **1993**, *366*, 476–479.
- Kumar, R.; Thompson, E. B. The structure of the nuclear hormone receptors. *Steroids* **1999**, *64*, 310–319.
- Moras, D.; Gronemeyer, H. The nuclear receptor ligand binding domain: structure and function. *Curr. Opin. Cell Biol.* **1998**, *10*, 384–391.
- Weatherman, R. V.; Fletterick, R. J.; Scanlan, T. S. Nuclear-Receptor ligands and ligand-binding domains. *Annu. Rev. Biochem.* **1999**, *68*, 559–581.
- No, D.; Yao, T.-P.; Evans, R. M. Ecdysone-inducible gene expression in mammalian cells and transgenic mice. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 3346–3351.
- Lüers, G. H.; Jess, N.; Franz, T. Reporter-linked monitoring of transgene expression in living cells using the ecdysone-inducible promoter system. *Eur. J. Cell Biol.* **2000**, *79*, 653–657.
- Gingrich, J. R.; Roder, J. Inducible gene expression in the nervous system of transgenic mice. *Annu. Rev. Neurosci.* **1998**, *21*, 377–405.
- Saez, E.; Nelson, M. C.; Eshelman, B.; Banayo, E.; Koder, A.; Cho, G. J.; Evans, R. M. Identification of ligands and coligands for the ecdysone-regulated gene switch. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 14512–14517.
- Pastorino, J. G.; Chen, S. T.; Tafani, M.; Snyder, J. W.; Farber, J. L. The overexpression of Bax produces cell death upon induction of the mitochondrial permeability transition. *J. Biol. Chem.* **1998**, *273*, 7770–7775.
- The EC₅₀ values for ponasterone A and muristerone A, several of the more active ecdysteroids in gene-switch assays, are generally 10–1000× higher in mammalian cells than in *Drosophila* cells.
- Dinan, L.; Hormann, R. E.; Fujimoto, T. An extensive ecdysteroid CoMFA. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 185–207.
- McDonnell, D. P.; Vegeto, E.; Gleeson, M. A. G. Nuclear hormone receptors as targets for new drug discovery. *Bio/Technology* **1993**, *11*, 1256–1261.
- Rosen, J.; Day, A.; Jones, T. K.; Jones, E. T. T.; Nadzan, A. M.; Stein, R. B. Intracellular receptors and signal transducers and activators of transcription superfamilies: novel targets for small-molecule drug discovery. *J. Med. Chem.* **1995**, *38*, 4855–4874.
- Dinan, L.; Savchenko, T.; Whiting, P.; Sarker, S. D. Plant natural products as insect steroid receptor agonists and antagonists. *Pesticide Sci.* **1999**, *55*, 331–335.
- Wing, K. D. RH 5849, a nonsteroidal ecdysone agonist: effects on a *Drosophila* cell line. *Science* **1988**, *241*, 467–469.
- Wing, K. D.; Slawewski, R. A.; Carlson, G. R. RH 5849, a nonsteroidal ecdysone agonist: effects on larval Lepidoptera. *Science* **1988**, *241*, 470–472.
- Martinez, A.; Sparks, C.; Hart, C. A.; Thompson, J.; Jepson, I. Ecdysone agonist inducible transcription in transgenic tobacco plants. *Plant J.* **1999**, *19*, 97–106.
- Carlson, G. R.; Cress, D. E.; Dhadialla, T. S.; Hormann, R. E.; Le, D. P. Nonsteroid ligands for modulating the expression of exogenous genes via an ecdysone receptor complex; Eur. Pat. Appl. EP 99–304444, 1999.
- Sláma, K.; Lafont, R. Insect hormones – ecdysteroids: their presence and actions in vertebrates. *Eur. J. Entomol.* **1995**, *92*, 355–377.
- Ecdysone: From chemistry to mode of action*; Koolman, J., Ed.; Thieme: Stuttgart, 1989.
- Dinan, L. Ecdysteroid structure and hormonal activity. In *Ecdysone: from chemistry to mode of action*; Koolman, J., Ed.; Thieme Verlag: Stuttgart, 1989; pp 345–354.
- Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Jin, B.; Albuquerque, M.; Madhav, P. J.; Duraiswami, C. Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J. Am. Chem. Soc.* **1997**, *119*, 10509–10524.
- Hopfinger, A.; Tokarski, J. 3D-QSAR analysis. In *Practical Application of Computer-Aided Drug Design*; Charifson, P. S., Ed.; Dekker: New York, 1997; p 105.
- Hopfinger, A. J.; Duca, J. S. Extraction of pharmacophore information from high-throughput screens. *Curr. Opin. Biotech.* **2000**, *11*, 97–103.
- 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.
- Cramer, R. D., III.; Depriest, S.; Patterson, D.; and Hecht, P. The developing practice of comparative molecular field analysis. In *3D QSAR in Drug Design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 443–485.
- 3D QSAR in Drug Design: Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998.
- Clément, C. Y.; Bradbrook, D. A.; Lafont, R.; Dinan, L. Assessment of a microplate-based bioassay for the detection of ecdysteroid-like or antiecdysteroid activities. *Insect Biochem. Mol. Biol.* **1993**, *23*, 187–193.
- Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol water partition coefficient. *Chem. Pharm. Bull.* **1992**, *40*, 127–130. MlogP implemented in SYBYL programming language by Blake, J. Pfizer Inc.: Version 1.2, Nov 8, 1994.
- Sybyl 6.5; Tripos Associates: 1699 South Hanley Road, St. Louis, MO 63144.
- Structural corrections to previous CoMFA model:²⁵ stachysterone **C 21** (added missing H on O-20), shidasterone **24** (added missing H on O-20), and 2-deoxy-20-hydroxyecdysone 22-benzoate **55** (correction of inverted configuration at C-22 and removal of superfluous keto group at C-12).
- Huber, R.; Hoppe, W. Zur Chemie des Ecdysons, VII: Die Kristall- und Molekülstrukturanalyse des Insektenverpuppungshormons Ecdyson mit der automatisierten Faltmolekülmethode. *Chem. Ber.* **1965**, *98*, 2403–2424.
- HyperChem Program Release 5.01 for Windows*; Hypercube, Inc.: 1996.
- MOLSIM User's Guide V3.0; D. C. Doherty and The Chem21 Group, Inc.: 1780 Wilson Drive, Lake Forest, IL. 1998.
- Tokarski, J. S.; Hopfinger, A. J. Constructing protein models for ligand–receptor binding thermodynamic simulations: an application to a set of peptidomimetic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 779–791.
- Albuquerque, M. G.; Hopfinger, A. J.; Barreiro, E. J.; de Alencastro, R. B. Four-dimensional quantitative structure–activity relationship analysis of a series of interphenylene 7-oxobicycloheptane oxazole thromboxane A(2) receptor antagonists. *J. Chem. Inf. Computer Sci.* **1998**, *38*, 925–938.
- Glen, W. G.; Dunn, W. J., III.; Scott, D. R. Linear regression and the partial least-squares, PLS, regression method. *Tetrahedron Comput. Methods* **1989**, *2*, 349–354.
- 4D-QSAR Analysis User's Guide, V2.0; The Chem21 Group, Inc.: 1780 Wilson Drive, Lake Forest, IL, 2000.
- Holland, J. *Adaptation in artificial and natural Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
- Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure–activity relationships and quantitative structure–property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
- Tokarski, J. S.; Hopfinger, A. J. Prediction of ligand–receptor binding thermodynamics by free energy force field (FEFF) 3D-QSAR analysis: application to a set of peptidomimetic renin inhibitors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 792–811.

- (56) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening – an overview. *Drug Discovery Today* **1998**, 3, 160–194.
- (57) Hopfinger, A. J.; Reaka, A.; Venkatarangan, P.; Duca, J. S.; Wang, S. Construction of a virtual high throughput screen by 4D-QSAR analysis: application to a combinatorial library of glucose inhibitors of glycogen phosphorylase b. *J. Chem. Inf. Comput. Sci.* **1999**, 39, 1151–1160.
- (58) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, 95, 927–937.
- (59) Klebe, G. Comparative molecular similarity indices analysis: COMSIA. In *3D QSAR in Drug Design: Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: Dordrecht, 1998; pp 87–104.
- (60) Wurtz, J.-M.; Guillot, B.; Fagart, J.; Moras, D.; Tietjen, K.; Schindler, M. A new model for 20-hydroxyecdysone and dibenzoylhydrazine binding: a homology modeling and docking approach. *Protein Sci.* **2000**, 9, 1073–1084.
- (61) Wurtz, J.-M.; personal communication.
- (62) Renaud, J. P.; Rochel, N.; Ruff, M.; Vivat, V.; Chambon, P.; Gronemeyer, H.; Moras, D. Crystal structure of the RAR- γ ligand-binding domain bound to all-trans retinoic acid. *Nature* **1995**, 378, 681–689.
- (63) Rochel, N.; Wurtz, J.-M.; Mitschler, A.; Klaholz, B.; Moras, D. The crystal structure of the nuclear receptor for vitamin D bound to its natural ligand. *Molecular Cell* **2000**, 5, 173–179.

CI010076U