A Different Method for Steric Field Evaluation in CoMFA Improves Model Robustness

Traian Sulea,*[,†,§] Tudor Ionel Oprea,[‡,§] Sorel Muresan,[⊥] and Shek Ling Chan[†]

Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Avenue, Montréal, QC H4P 2R2, Canada, Astra-Hässle AB, S-431 83 Mölndal, Sweden, Agrotechnological Research Institute, Bornsesteeg 59, P.O. Box 17, NL-6700 AA Wageningen, The Netherlands, and FORSLYS: The Group for the Formal Study of Living Systems, P.O. Box 1120, RO-1900, Timisoara-14, Romania

Received March 10, 1997[⊗]

The all-grid probe Lennard-Jones 6−12 potential, typically used as steric descriptor in CoMFA, has been replaced by the volumes of van der Waals envelopes intersections between the probe-atom and the ligand molecules under investigation. The intersection volumes present a smoother distance dependence that overcomes the problems arising from formulation of a precise alignment and docking into the 3D lattice. A CoMFA-type application on a set of 78 steroid aromatase inhibitors with different grid and probe-atom characteristics suggests an improved model robustness for the new steric field in comparison with the classical 6−12 potential. Predefined cut-off and "minimum_sigma" values are not required. Systematic variation of the 3D lattice position leads to lower variations of cross-validated *r*-squared (q^2_{LOO}) at all levels of model complexity, which can be reduced with exclusion of some "interior" points. The relative simplicity and inexpensive computational demands make this steric field a promising alternative for routine CoMFA-based 3D-QSAR analyses.

INTRODUCTION

Comparative Molecular Field Analysis (CoMFA), one of the most appealing and successful 3D-QSAR methods, is based on computation of interaction energies of ligand molecules with a grid simulating receptor model.[1] This method attempts to find an empirical linear relationship between the experimentally determined free energy of binding and 3D-grid-based computed features for a set of compounds under investigation. After defining an alignment rule, steric (Lennard-Jones 6−12 potential), electrostatic (Coulombic potential), and/or hydrophobicity (hydrophobicity or lipophilicity potentials) interaction energies are calculated between a probe-atom with specified van der Waals radius and atomic charge, and each of the molecular structures, typically at the surrounding points of a predefined regularly spaced 3D lattice. In order to derive a QSAR equation from the resulting highly underdetermined matrices, a multivariate technique called partial least squares (PLS) is applied.[2,3] As it operates with latent variables, this method is not sensitive to possible colinearity of the underlying descriptor matrix. PLS analyses are usually performed in combination with cross-validation, in order to check for consistency of the model under consideration.[1,4] The Lennard-Jones 6−12 potential, typically used to compute steric fields, is characterized by a very steep increase in energy at short distances around molecular van der Waals surface and, therefore, requires an arbitrary cut-off value (usually at +30.0 kcal/mol).[5]

The strongly increasing steric potentials beyond van der Waals (vdW) contact interatomic distances produce certain difficulties in the CoMFA method. The acute steepness of the function increases the sensitivity of the statistic results to the precise docking of the aligned set of molecules into the 3D grid as well as to the accuracy of the alignment rule formulation. Moreover, the differences in slopes of the steric and electrostatic (and hydrophobicity) potentials lead to arbitrary settings for energy terms at those lattice points where one of the considered fields reaches its cut-off limit. This rather hazardous assignment becomes crucial for the quality of the QSAR model, since the PLS analysis is directed toward the points located in the vicinity of the compounds characterized by highest variance. Therefore, the definition of a very accurate alignment rule for the compounds with respect to the individual conformations as well as the relative orientation of the structure and the position of the points where the fields are probed around the aligned set of structures have proven to be of crucial importance for analyses involving the Lennard-Jones interaction energies.

Although the 3D lattice and the aligned set of investigated compounds are treated as rigid bodies in CoMFA methodology, in reality both ligand and receptor have certain degrees of flexibility. During ligand-receptor interaction, interatomic distances will never go much below van der Waals contact distances. Steric misfit (i.e., overlap of ligand molecule and binding site cavity van der Waals envelopes) is circumvented by local torsion angle deformation in both partners, at much lower energy costs. Our experience with the Minimum Steric/Topologic Difference (MTD) method[6] suggests that the potential energy increase, associated with circumvention of steric misfit, is roughly proportional to the volume of van der Waals envelope intersection between the ligand molecule and the (ideal shape of) receptor cavity. The MTD parameter is largely proportional to the overlapping volume of receptor and ligand van der Waals envelopes. As we already have pointed out,[7] it seems worthwhile to use more realistic, semiempirical assumptions in selection of the potential energy terms for the CoMFA method, suggesting replacement of the van der Waals attractive terms by hydrophobicity

* To receive all correspondence. E-mail address: traian.sulea@bri.nrc.ca.
† National Research Council of Canada.
‡ Astra-Hässle AB.
⊥ Agrotechnological Research Institute.
§ FORSLYS.
⊗ Abstract published in *Advance ACS Abstracts,* October 15, 1997.

S0095-2338(97)00422-8 CCC: $14.00

parameters, while repulsion should be represented by overlapping van der Waals volumes.

Alternative approaches have been developed in order to overcome the problems arising from the steepness of the Lennard-Jones potential at close proximity to the molecular surface and/or the differences in slope of the potentials used in actual CoMFA implementation. Klebe *et al.*[8] used similarity indices to compute molecular potential fields in Comparative Molecular Similarity Indices Analysis (CoMSIA). Using the rigid body molecular alignment program SEAL,[9] CoMSIA steric, electrostatic, and hydrophobic similarity indices replace the distance function of the Lennard-Jones and Coulomb-type potentials with Gaussian-type function, $\exp(-\alpha r^2)$, where the α term provides for a "local smearing" effect which places more weight on interactions close to the molecular surface, with a smooth transition to more distal points. The fields do not require cut-off values, and the net result is to lessen the effect of changes in the field descriptors associated with minor variations in molecular superposition or conformation. Cho and Tropsha[10] reduced the cross-validation results sensitivity to the grid orientation and spacing by using a CoMFA region optimization routine that focuses on regions in space characterized by maximal q^2. Somehow surprising results were obtained by Kroemer and Hecht[11] which improved model's consistency by replacing the Lennard-Jones potential with a rather steeper atom-based indicator variables field. The higher statistic quality of their indicator potential function seems to lie on signal-to-noise ratio improvement comparing with the original Lennard-Jones potential. Since the authors reported statistic results for only two grid positions, it is difficult to estimate the robustness of this steric potential function to a complete rotation of the 3D lattice around the aligned set of molecular structures. In order to avoid unjustified large parametric variance due to the steep increase of the steric field contribution at lattice points close to the molecules, some authors have suggested a truncation of the steric energies at 4.0 or 5.0 kcal/mol instead of the commonly used 30.0 kcal/mol level.[12−14] Using these cut-off values and an appropriate "minimum_sigma" to avoid an overproportional exclusion of columns, sometime superior models can be obtained.

Several works apply the CoMFA-derived fields beyond QSAR studies. The field fit minimization option[1,15] has two additional terms represented by grid-based CoMFA steric (Lennard-Jones 6−12) and electrostatic (Coulombic) fields. More recently, Cramer *et al.*[16,17] have shown that CoMFA steric field is among the most frequently useful descriptors yet found for neighborhood-based design of combinatorial libraries. The widespread applications of shape descriptors represented as steric 3D potentials require their realistic parametrization in order to achieve maximal results.

Here we describe the van der Waals envelopes intersection volumes as a steric potential field in 3D-QSAR CoMFA-based modeling of ligand−receptor interactions. This intersection volumes field (termed INVOL in what follows) varies smoothly with interatomic distances.[18] Thus, the large variations in steric potential associated with receptor grid spacing distances and the precise docking of ligand molecules will be strongly reduced. Although without a direct physical meaning, the field implicitly contains repulsive terms associated with the circumvention of steric misfit between the ligand and the receptor active binding site by local deformability of both partners to the noncovalent interaction. We apply the new steric field to derive a CoMFA-type 3D-QSAR model for an enlarged series of steroid aromatase inhibitors, a subset of which has been analyzed with the classical CoMFA steric field.[19] The sensitivity of the derived field to the grid orientation and spacing, "minimum_sigma" threshold, exclusion of "interior" points, and probe-atom characteristics as well as 3D location of the important regions of the final stdev*coeff contour maps were evaluated comparatively with the standard CoMFA implementation. In the present note we test the behavior of the steric function without inclusion of electrostatics and hydrophobicity.

METHODS

The program for calculating van der Waals intersection volumes is written in C++ programming language. All other computations were performed within SYBYL 6.2 program,[20] enhanced with SYBYL Programming Language (SPL) codes to integrate the use of INVOL into SYBYL. All computations were run on a Silicon Graphics workstation.

Data Set and Alignment Rule Formulation. We used the binding data for a set of steroid inhibitors to the human aromatase to validate the proposed steric field. A 50-compound set, previously analyzed with classical CoMFA method by Oprea and García,[19] was used as a starting point to which other 28 active steroid compounds were added. The modeling work and alignment procedure were kept the same. In brief, the molecular structures were optimized with the MOPAC 6.0 program,[21] using MNDO Hamiltonian and MNDO partial atomic charges. The starting geometries were based on three crystallographic structures from Cambridge Crystallographic Data Centre, and extended conformations were assigned for flexible side chains. The XED program[22] was used to identify the pharmacophoric pattern and to generate a preliminary alignment rule, using one compound (**21**) as template. Once the XED superimposition pattern was generated, the CoMFA alignment was constructed by an rms fit of all 17 carbons in the steroid skeleton. The biological activities of the entire set of steroids are given in Table 1, while Figure 1 summarizes the chemical structures of the 28 compounds added to the previously analyzed set of 50 steroids (see ref 19 for the chemical structures of the 50-steroid set and the references therein for the sources of biological data).

Calculation of the Intersection Volumes Steric Field. A schematic representation of the method is given in Figure 2. The ligand molecules were described within the "hard spheres" approximation. The van der Waals envelopes intersection volumes [Å^3] were computed using the Monte-Carlo integration technique.[23,24] An average density of 1000 randomly generated points per Å^3 was used to evaluate the volume of intersection between the probe-atom and the ligand atoms. This level of the volume estimation provides a good accuracy while keeping the CPU-time at a low cost.[24] Since the molecular surface of most organic compounds (in general) and biomacromolecules (in particular) is formed by hydrogen atoms, the probe-atom almost always has the hydrogen van der Waals radius. It should be noted that Bondi vdW radius[25] for hydrogen is 1.2 Å, while Tripos force field as implemented in SYBYL[20] makes use of an exceptionally large hydrogen atom with a van der Waals radius of 1.5 Å. We used both van der Waals radii for probe-atom

Table 1. Biological Activities of the Steroid Aromatase Inhibitors Used To Derive QSAR Analyses[a]

ID	pK_i	ID	pK_i	ID	pK_i	ID	pK_i
1	−1.04	**21**	1.89	**41**	1.92	**61**	1.66
2	−1.65	**22**	1.90	**42**	1.66	**62**	1.91
3	−1.11	**23**	0.60	**43**	1.51	**63**	1.56
4	−0.38	**24**	1.59	**44**	1.43	**64**	1.85
5	−0.74	**25**	1.30	**45**	1.68	**65**	−0.15
6	−1.18	**26**	2.30	**46**	1.20	**66**	1.06
7	−0.26	**27**	0.92	**47**	2.00	**67**	1.03
8	−0.48	**28**	0.08	**48**	2.29	**68**	1.21
9	0.00	**29**	0.77	**49**	1.21	**69**	1.02
10	−0.99	**30**	0.54	**50**	2.31	**70**	1.06
11	−0.15	**31**	0.64	**51**	1.43	**71**	0.76
12	0.92	**32**	0.34	**52**	1.17	**72**	2.15
13	−1.04	**33**	0.35	**53**	1.34	**73**	1.70
14	0.60	**34**	1.96	**54**	1.34	**74**	2.66
15	0.66	**35**	2.25	**55**	2.37	**75**	0.65
16	−0.72	**36**	2.85	**56**	2.14	**76**	1.46
17	−0.48	**37**	2.33	**57**	2.00	**77**	0.07
18	1.24	**38**	2.34	**58**	1.70	**78**	0.04
19	−0.30	**39**	2.17	**59**	1.74		
20	0.85	**40**	2.06	**60**	2.00		

[a] Experimental K_i (**p*K*i**) values are expressed as the $-\log_{10}$ of the micromolar concentrations. **ID** numbers correspond to those presented in Figure 1 for ID: **51**−**78** and ref 19 for ID: **1**−**50**. The sources of biological data for the first 50 compounds are given in ref 19. K_i values for the compounds **51**−**53** are taken from ref 27, **54**−**57** from ref 28, **58**−**65** from ref 29, **66**−**71** from ref 30, and **72**−**78** from ref 31.

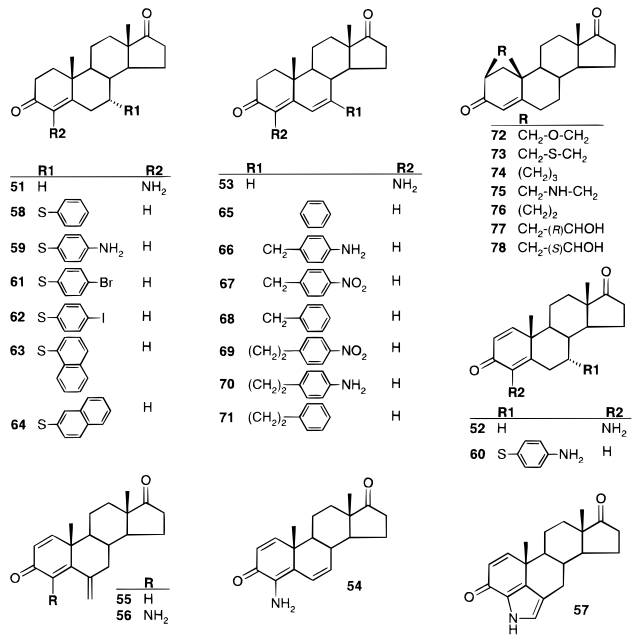

Figure 1. The 28 steroid structures added to enlarge the previously analyzed set of 50 steroid aromatase inhibitors.[19] The 78-steroid set was used to derive the 3D-QSAR models.

and hydrogen atoms of the ligands as an opportunity to test the sensitivity of the statistic results to these settings. The van der Waals radii for all other atom types correspond to the values used by Tripos force field as implemented in SYBYL program[20] (which correspond to the van der Waals radii presented by Bondi[25]).

The program has two additional options. One of them is used to generate a reduced grid (termed RED) which is the minimal grid that embeds the aligned set of compounds and is sufficient for intersection volume evaluations. The other one specifies the maximum volume of intersection allowed for PLS analysis and excludes all grid-points characterized

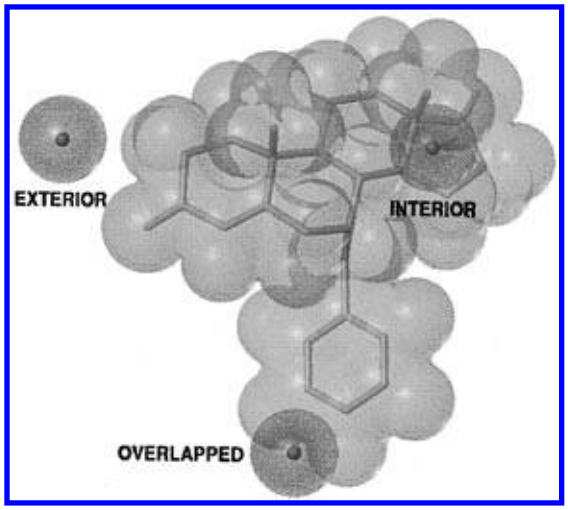

Figure 2. Schematic representation of the INVOL method. The volume of intersection between the van der Waals envelopes of the ligand molecule (green) and probe-atom (magenta) is calculated as steric 3D descriptor in the "overlapped" positions. "Exterior" and "interior" points are usually discarded.

by a higher volume of intersection ("interior" position in Figure 2). The last option was necessary to test the effect of the points "inside" a molecule *versus* the statistical quality of the generated models. The grid dimensions, grid spacing, probe-atom and/or hydrogen atoms van der Waals radii, the number of points used to estimate the volume of intersection, and the level for "interior" points exclusion are specified externally. The program reads the SYBYL mol2 files and writes the appropriate fields as ASCII files. These are subsequently read by an interface containing SPL macro commands that fills a CoMFA-type column into corresponding SYBYL Molecular Spreadsheet. The 3D lattices used to compute the field had either 2.0 or 1.0 Å spacing and 3D dimensions and locations corresponding to the following codes: AUTO, RED, USER, AUTO_2A (64 grids), and AUTO_1A (eight grids), as described below.

Calculation of the Lennard-Jones Steric Field. The Lennard-Jones steric fields were computed with a Csp^3 probe-atom using 3D lattices with either 2.0 or 1.0 Å spacing. The grid positions and dimensions used to derive the standard Lennard-Jones-based CoMFA were AUTO, RED_EXT, USER, AUTO_2A (64 grids), and AUTO_1A (eight grids) as described below. The Lennard-Jones steric interaction energies were truncated at 30.0 kcal/mol, and the points "inside" each molecule were not considered for the PLS analyses.

PLS Analyses and Cross-Validation Tests. The current PLS implementation in SYBYL/CoMFA was used. In order to avoid the randomness arising from test-group selection, cross-validations were performed by means of "leave-one-out" (LOO) procedure. "Minimum_sigma" standard deviation threshold was set to 2.0 kcal/mol for 6−12 potential-based analyses and to 0.0 $Å^3$ for INVOL-based analyses, unless otherwise specified. Preliminary analyses were carried out with a maximum of seven principal components (PC) and repeated using that number of components at which the difference in the cross-validated q^2_{LOO} value to the next principal component was less than 2%. The same number of PC was subsequently used to derive the final models, some of them contoured as stdev*coeff plots using standard conventions.

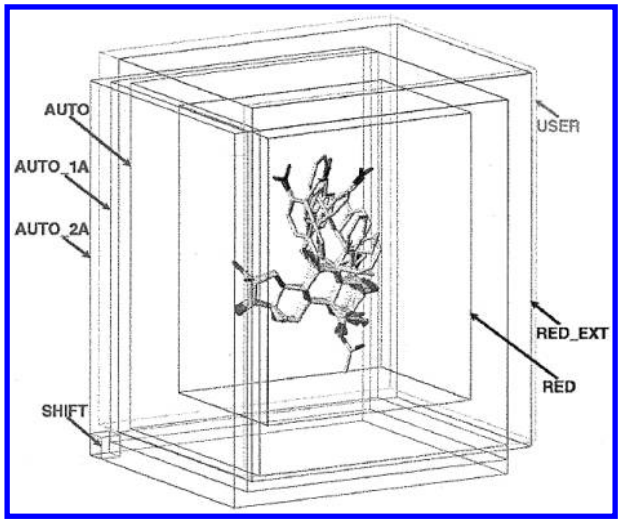

Figure 3. Grid positions used to perform various INVOL tests (see Methods section for descriptions).

Grid Settings and Grid-Dependent Tests. Various grid positions and spacings were used to test the dependence of the statistic results. The standard grid generated by SYBYL/CoMFA, which extends 4.0 Å beyond the aligned set of compounds along the principal axes of a Cartesian coordinate system, is termed AUTO. A reduced grid generated by the INVOL program, which is the minimal grid that embeds the aligned set of compounds at maximum, is called RED and was used only for the new steric field evaluations. To make the results comparable with those from the analyses involving Lennard-Jones potential, RED position was kept fixed, and its dimensions were extended with 3.0 Å in all Cartesian directions. The resulting grid, termed RED_EXT, was used only with Lennard-Jones fields. A user-specified region, called USER, different from the above ones, and used in the previous CoMFA analysis of a subset of steroidic aromatase inhibitors,[19] was applied for both fields. All the above mentioned 3D lattices were used with both 2.0 and 1.0 Å spacing.

In order to test the model's sensitivity to grid shifting, the AUTO grid was used as a starting point for generation of a 64-grid set with 2.0 Å spacing and an eight-grid set with 1.0 Å spacing. The two sets were generated as the lower corner of the AUTO grid was translated with 0.5 Å along X, Y, and Z axes in one (positive) direction, thus resulting in 64 different grid positions for 2.0 Å spacing and eight different grid positions for 1.0 Å spacing. Since some of the points in the original AUTO grid are lost during the grid movement, the coordinates at the lower corner of the AUTO grid were extended in opposite (negative) direction with 2.0 and 1.0 Å, for 2.0 Å spaced grids and 1.0 Å spaced grids, respectively, prior to grid shifting. The resulting grid sets, named AUTO_2A and AUTO_1A, were used for both steric fields evaluations.

For INVOL-based analyses performed within this test, the fields were computed with a hydrogen (and probe-atom) van der Waals radii of 1.5 Å. A graphic representation of the 3D lattices around the aligned set of molecules is given in Figure 3, and their corresponding numbers of points are listed in Table 2.

Other Tests. The influence of the probe-atom van der Waals radius was determined using various settings. Two of them consider the probe-atom as a hydrogen atom and set the hydrogen vdW radius to either 1.2 or 1.5 Å. Other cases with probe-atom/hydrogen atoms of the ligands vdW radii of 1.5 Å/1.2 Å, 1.7 Å/1.2 Å, and 1.7 Å/1.5 Å were also tested.

Table 2. Number of Points for Various Grids Used in This Study[a]

GS	GP	no. of points
2	AUTO	1170
	AUTO_2A	1540
	RED	693
	RED_EXT	1680
	USER	1560
1	AUTO	8550
	AUTO_1A	10374
	RED	4641
	RED_EXT	11286
	USER	11960

[a] **GS**: Grid spacing [Å]. **GP**: Grid position (see Methods section and Figure 3).

The effect of the "interior" points exclusion was tested only for the two cases with the same van der Waals radius for both probe-atom and hydrogen atoms of the ligands (i.e., 1.2 Å/1.2 Å and 1.5 Å/1.5 Å). For these settings, the maximum intersection volumes, corresponding to the probe-atom vdW volumes, are 7.2 $Å^3$ and 14.3 $Å^3$ for 1.2 Å and 1.5 Å vdW radii, respectively. The maximum INVOL used for PLS analyses was set to various values between 1.0 and 7.0 $Å^3$ in the first case and to values from 1.0 to 14.0 $Å^3$ in the second case, with a 1.0 $Å^3$ step in both cases. All points that exceed this cut-off were excluded from the PLS analyses.

Various minimum_sigma values were assigned during PLS analyses for both steric fields. In the case of INVOL field, minimum_sigma was set to various values within 0.0−2.0 $Å^3$ interval for 1.2 Å hydrogen vdW radius and within 0.0−3.0 $Å^3$ interval for 1.5 Å hydrogen vdW radius, with a step of 0.1 $Å^3$ for both cases. For Lennard-Jones field, minimum_sigma was set within 0.0−4.0 kcal/mol interval with a step of 0.1 kcal/mol.

RESULTS AND DISCUSSIONS

We have previously proposed the use of the van der Waals envelopes intersection volumes as steric field in CoMFA based on the reduced steepness of the function in comparison with the classical Lennard-Jones potential.[18] Indeed, if the two functions are seen as different powers of the interatomic distance, i.e., r^{-3} (INVOL) and r^{-12} (conventional vdW), it is obvious that the former is much smoother at small distances. In addition, we have pointed out that the new function has the advantage of eliminating the artificial cut-off and have predicted an increased stability of the QSAR statistical parameters on various grid properties. In this section we comparatively present the results of the tests performed with van der Waals envelopes intersection volumes and Lennard-Jones 6−12 potentials as steric fields in CoMFA-type 3D-QSAR on the same data set.

Grid Location and Spacing. An important observation is that INVOL can replace the CoMFA standard steric field and produce models that are much less grid-positon and spacing dependent. This is illustrated in Figure 4, where the q^2_{LOO} values for the first seven principal components are represented as obtained with both functions using two grid sets with systematically modified grid position.

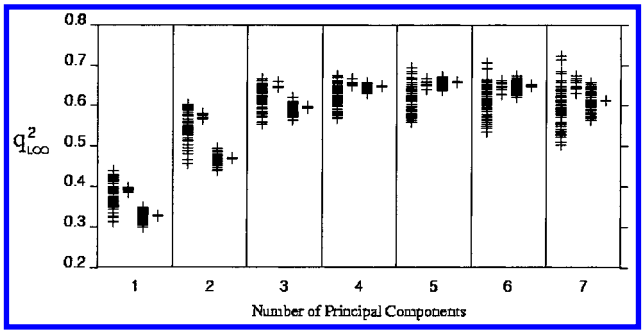

Figure 4. LOO cross-validation results from the grid position and spacing sensitivity test. q^2_{LOO} values by number of principal components (PC). From left to right, for each PC: LJ 6−12 at 2.0 Å grid spacing, LJ 6−12 at 1.0 Å grid spacing, INVOL at 2.0 Å grid spacing, and INVOL at 1.0 Å grid spacing. All INVOL-based analyses were performed with 1.5 Å van der Waals radius for both probe-atom and hydrogen atoms of the ligands. A minimum_sigma of 0.0 Å^3 without exclusion of "interior" points was applied for PLS analyses with INVOL columns. PLS correlations with Lennard-Jones columns were carried out using standard settings.

Table 3. Statistical Results for the Grid Position and Spacing Sensitivity Test[a]

field	GS	*n*	PC	$q^2_{LOO-min}$	$q^2_{LOO-max}$	Δq^2_{LOO}	$q^2_{LOO-mean}$	SD
INVOL	2	64	1	0.300	0.349	0.049	0.327	0.012
			2	0.439	0.496	0.057	0.469	0.014
			3	0.563	0.621	0.058	0.591	0.011
			4	**0.630**	**0.656**	**0.026**	**0.642**	**0.006**
			5	0.637	0.671	0.034	0.652	0.008
			6	0.619	0.673	0.054	0.644	0.014
			7	0.563	0.658	0.095	0.614	0.023
	1	8	1	0.326	0.328	0.002	0.327	0.001
			2	0.469	0.471	0.002	0.471	0.001
			3	0.594	0.597	0.003	0.595	0.001
			4	**0.646**	**0.649**	**0.003**	**0.647**	**0.001**
			5	0.657	0.659	0.002	0.657	0.001
			6	0.648	0.651	0.003	0.650	0.001
			7	0.612	0.614	0.002	0.613	0.001
LJ 6-12	2	64	1	0.312	0.438	0.126	0.383	0.031
			2	0.456	0.604	0.148	0.549	0.034
			3	**0.554**	**0.667**	**0.113**	**0.617**	**0.026**
			4	0.569	0.674	0.105	0.622	0.028
			5	0.558	0.694	0.136	0.618	0.033
			6	0.534	0.707	0.173	0.609	0.040
			7	0.502	0.723	0.221	0.597	0.047
	1	8	1	0.384	0.396	0.012	0.391	0.004
			2	0.567	0.581	0.014	0.572	0.005
			3	**0.645**	**0.659**	**0.014**	**0.648**	**0.004**
			4	0.650	0.666	0.016	0.655	0.004
			5	0.641	0.666	0.025	0.653	0.007
			6	0.627	0.663	0.036	0.647	0.010
			7	0.629	0.674	0.045	0.653	0.014

[a] **Field**: Steric field (either INVOL or LJ 6−12) used to perform CoMFA analyses. **GS**: Grid spacing [Å]. ***n***: The number of grid positions. **PC**: The number of principal components. For each principal component, $\boldsymbol{q^2_{LOO-min}}$ is the minimum q^2_{LOO}; $\boldsymbol{q^2_{LOO-max}}$ is the maximum q^2_{LOO}; Δq^2_{LOO} represents $q^2_{LOO-max} - q^2_{LOO-min}$; $\boldsymbol{q^2_{LOO-mean}}$ is the average q^2_{LOO}; and SD is the standard deviation of q^2_{LOO}. The optimum number of components is bolded in each case.

Statistic analysis of these results (Table 3) reveals that there is a minimum in the absolute spread of the q^2_{LOO} values *versus* the numbers of principal components used for cross-validation, for both potential functions. The position of this minimum is probably associated with the optimum number of PC, as selected using the 2% q^2_{LOO} increase rule. Indeed, with INVOL potential at 2.0 Å spacing, the absolute q^2_{LOO} variance gets down to 0.026 q^2_{LOO} units at an optimum of 4 PC, while with the 6−12 potential Δq^2_{LOO} reaches 0.113 q^2_{LOO} units for an optimum of 3 PC (minimum of 0.105 q^2_{LOO} units at 4 PC). As seen from Figure 5a, the absolute q^2_{LOO} variance is reduced with 48.7% (3 PC) to 75.2% (4 PC) by using INVOL instead of classical Lennard-Jones. In general, 1.0 Å spacing substantially reduces the grid sensitivity for both fields (see Table 3). However, although the statistics for 1.0 Å spaced grids is based on only eight grids, the use of INVOL field drops the absolute q^2_{LOO} spread with 78.6% (3 PC) to 95.6% (7 PC).

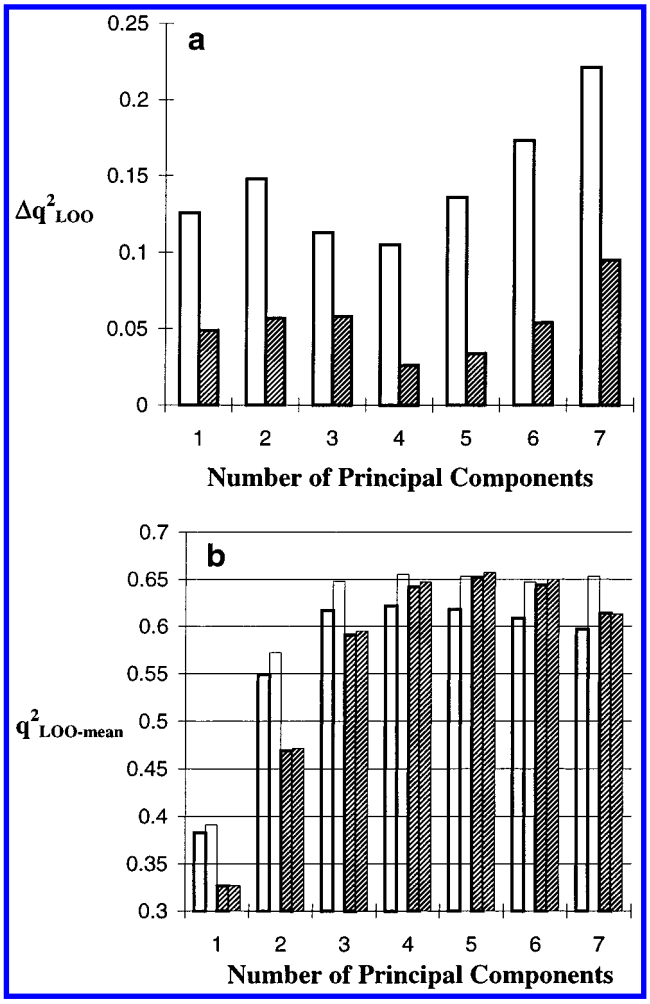

Figure 5. Statistic analysis of the results obtained from the grid position and spacing sensitivity test. a. Absolute q^2_{LOO} variance (Δq^2_{LOO}) at each level of model complexity. Open bars are used for LJ 6−12-based results and striped bars for INVOL-based results. Only the results with 2.0 Å spaced grids are shown. b. Average q^2_{LOO} values for each level of model complexity. Open bars are used for LJ 6−12-based results and striped bars for INVOL-based results. Bolded bars are used for 2.0 Å spaced grids and thin bars for 1.0 Å spaced grids.

Looking at the q^2_{LOO} values averaged for each number of PC, it can be noticed that intersection volumes field leads to very similar results at both 2.0 and 1.0 Å spacings. In contrast, $q^2_{LOO-mean}$ values for 6−12 potential are generally higher at 1.0 Å comparing with 2.0 Å spacing (Figure 5b). It is interesting to point out the fact that the average q^2_{LOO} found for an optimum of 4 PC with INVOL (0.642 at 2.0 Å and 0.647 at 1.0 Å spacing) is in good agreement with that found with an optimum of 3 PC with LJ 6−12 at 1.0 Å spacing (0.648), while the conventional field gives a lower mean value at 2.0 Å spacing (0.617). Although the classical 6−12 potential leads to higher q^2_{LOO} faster, the $q^2_{LOO-mean}$ obtained with INVOL at 4 PC is higher than maximal

Table 4. LOO Cross-Validation and Final Results for the van der Waals Radii Sensitivity Test[a]

field	PAr/Hr	GS	GP	OPC	q^2_{LOO}	SEP	r^2	SEE
INVOL	1.2/1.2	2	AUTO	3	0.604	0.688	0.711	0.588
			RED	4	0.654	0.647	0.769	0.529
			USER	4	0.617	0.681	0.742	0.559
		1	AUTO	4	0.648	0.653	0.760	0.539
			RED	4	0.648	0.653	0.760	0.539
			USER	4	0.648	0.653	0.760	0.539
	1.5/1.5	2	AUTO	4	0.635	0.665	0.746	0.555
			RED	4	0.648	0.653	0.763	0.536
			USER	4	0.643	0.658	0.756	0.544
		1	AUTO	4	0.647	0.653	0.758	0.542
			RED	4	0.647	0.654	0.757	0.542
			USER	4	0.647	0.654	0.758	0.542
	1.5/1.2	2	AUTO	4	0.629	0.671	0.737	0.564
			RED	5	0.644	0.662	0.764	0.538
			USER	5	0.630	0.674	0.754	0.550
	1.7/1.2	2	AUTO	5	0.645	0.661	0.756	0.548
			RED	5	0.629	0.675	0.751	0.553
			USER	5	0.634	0.671	0.750	0.554
	1.7/1.5	2	AUTO	4	0.639	0.661	0.748	0.552
			RED	5	0.649	0.656	0.768	0.534
			USER	4	0.644	0.657	0.756	0.544
LJ 6-12	-	2	AUTO	3	0.611	0.682	0.733	0.565
			RED_EXT	3	0.620	0.674	0.740	0.558
			USER	3	0.653	0.644	0.772	0.522
		1	AUTO	3	0.646	0.650	0.761	0.534
			RED_EXT	3	0.637	0.659	0.757	0.539
			USER	3	0.655	0.642	0.768	0.527

[a] **Field**: Steric field (either INVOL or LJ 6−12) used to perform CoMFA analyses. **PAr/Hr**: Probe-atom/hydrogens van der Waals radii [Å]. **GS**: Grid spacing [Å]. **GP**: Grid position. **OPC**: Optimum number of principal components. **SEP**: Standard error of predictions. **SEE**: Standard error of estimates.

averaged q^2_{LOO} obtained with the Lennard-Jones potential at 2.0 Å.

Standard deviations (SD, computed with a biased *n*-based formula for the entire data population) of the q^2_{LOO} calculated for each number of principal components, clearly associate the minimum SD with the optimum number of PC (Table 3). The minimum SD values are located at 4 PC for INVOL (SD = 0.006 at 2.0 Å and 0.001 at 1.0 Å spacing, the last one being constant over all number of PC) and at 3 PC for LJ 6−12 (SD = 0.026 at 2.0 Å and 0.004 at 1.0 Å spacing, the second value also found at one and four PC). Using INVOL at 2.0 Å spacing, the q^2_{LOO} standard deviation decreases with 51.1% (7 PC) to 78.6% (4 PC), and for 1.0 Å spacing it gets down with 75% (at 1, 3, and 4 PC) to 92.9% (7 PC).

Overall better statistic results are obtained with van der Waals envelopes intersection volumes function over Lennard-Jones potential, at a cost of an increased model complexity with one principal component. A fine-tuning of the model's statistical quality could be achieved by selecting different hydrogen van der Waals radii, as discussed below.

van der Waals Radii. Various settings for the van der Waals radii of the probe-atom and hydrogen atoms of the ligands were tested. The results from the analyses involving three different grid positions (AUTO, RED - RED_EXT for comparison with the Lennard-Jones potential - and USER) and two grid spacings (2.0 and 1.0 Å) are summarized in Table 4.

With a van der Waals radius of 1.2 Å for both probe-atom and hydrogen atoms of the investigated molecules, the q^2_{LOO} spans 0.050 q^2 units at 2.0 Å, but it is constant at 1.0 Å spacing, having practically the same value (0.648) to that identified in the previous test. Δq^2_{LOO} exceeds the range found in the grid characteristics test, partly because that test was carried out with a 1.5 Å vdW radius for both probe-atom and hydrogens, and partly due to the fact that one of the three q^2_{LOO} values reported here corresponds to a three-component model. The results from computations carried out with a 1.5 Å vdW radius for both probe-atom and hydrogens became more stable. The absolute q^2_{LOO} variance is reduced to 0.013 q^2 units for 2.0 Å and zero at 1.0 Å spacing (both of them being well in the range suggested by the grid characteristics test). It can be concluded that 1.2 Å vdW radius is not the best choice for INVOL-based computations, the resulting field being too sparse and somehow resembling to the Lennard-Jones potentials from this point of view. The enhanced stability at 1.5 Å vdW radius seems to lie in a smoother distance dependence and the ability of the larger probe-atom to efficiently investigate a wider space, since the hydrogens of the ligands have a greater van der Waals radius too.

Even though the results show that a vdW radius of 1.5 Å for both probe-atom and hydrogen atoms of the compounds is a good alternative to 1.2 Å, leading to superior statistic results in the cross-validation test and a reduced sensitivity to the grid location, we further investigated other settings: 1.5 Å/1.2 Å, 1.7 Å/1.2 Å, and 1.7 Å/1.5 Å for probe-atom/hydrogens vdW radii. The models show an enhanced instability in terms of both complexity and q^2_{LOO} values. We note that the probe-atom and the hydrogens should have identical van der Waals radii in order to get consistent results. Standard error of prediction (SEP) shows a similar trend with q^2_{LOO} but in opposite direction. Final models were also carried out for all cases, at optimum number of PC, and the results indicate that the behaviors of r^2 and standard error of estimate (SEE) follow those of q^2_{LOO} and SEP, respectively. The Lennard-Jones-based analyses were performed with the same grid locations and spacings, as a control. The results obtained in the grid characteristics test were well reproduced, suggesting a reduced complexity with the cost of a higher instability, even at 1.0 Å grid spacing.

For the tests discussed up to here, all columns were considered in the PLS analyses with INVOL field. We will show below that excluding some "interior" points, the optimum number of principal components can be substantially reduced.

"Interior" Points. Exclusion of "interior" points (points "inside" a molecule, see Figure 2) for computations with volume intersection field highlighted a new aspect (Figure 6). A progressive elimination of the "interior" points with highest volume of intersection decreased the statistic quality of the models. However, the consistency was improved with exclusion of those points characterized by an intersection volume higher than approximately 50% of the maximal INVOL value. The model generated with 1.5 Å vdW radius for probe-atom and hydrogen atoms reaches a q^2_{LOO} value of 0.65 with only two principal components, outperforming both LJ 6−12-based models and INVOL-based analyses without exclusion of "interior" points.

Therefore, optimal quality is achieved when only half of the probe-atom is "buried" into compounds under investigation. We believe that "interior" points introduce more noise, without having a specific physical meaning to the ligand−

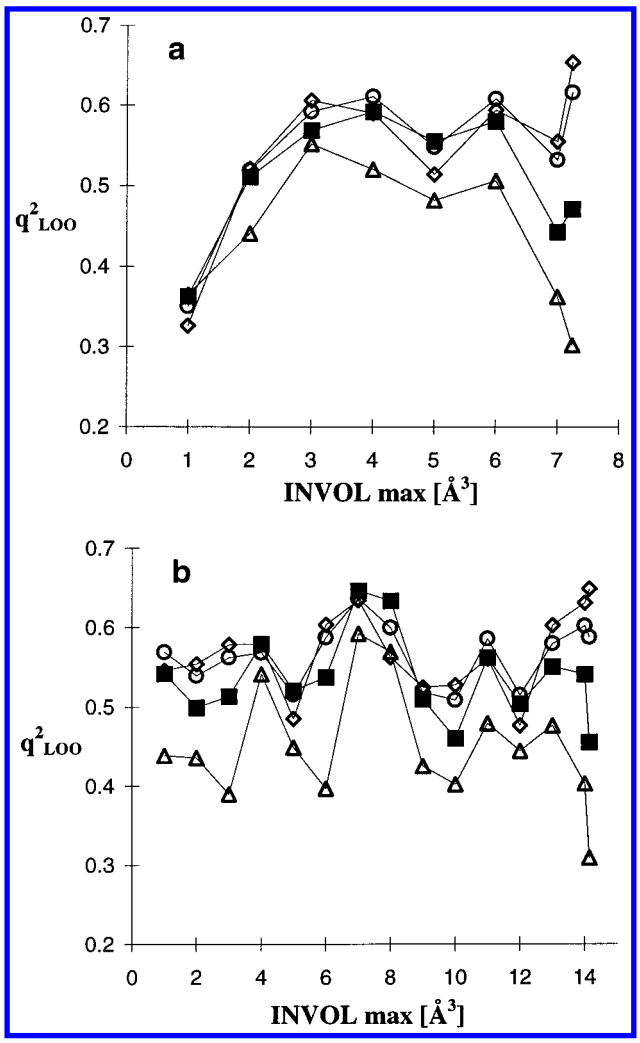

Figure 6. LOO cross-validation results from the "interior" points exclusion sensitivity test. 1 PC - open triangles; 2 PC - closed squares; 3 PC - open circles; 4 PC - open diamonds. Both probe-atom and hydrogen atoms of the ligands had a. 1.2 Å and b. 1.5 Å van der Waals radius. All analyses were performed with 0.0 Å^3 minimum_sigma and grid RED.

receptor interactions. The slopes of the two steric function are still very different upon the interval corresponding to 0−50% overlapping volume,[18] suggesting that INVOL field is relatively stable for different grid locations even with this truncation applied. Interestingly, when all points were excluded but those with an intersection volume lower than 1.0 Å^3, the models maintain some internal predictive power. In these cases, the probe "feels" the ligands by just "touching" them, resulting in similar approximations to the models employing Lennard-Jones potential and being characterized by the same instability to the grid location (data not shown).

Minimum_Sigma. The effect of minimum_sigma gave surprising results in analyses involving LJ 6-12 potential: q^2_{LOO} was almost invariant at various settings of minimum_sigma between 0.0 and 4.0 kcal/mol with 0.1 kcal/mol step. For analyses where intersection volumes were used as steric field, low levels for minimum_sigma yield better cross-validation results, as seen in Figure 7. The behavior of the 6−12 potential-based models seems to be a particularity of the investigated series of steroids, with a rather extended common core framework. However, we expect intersection volumes function to be more sensitive to this parameter, due

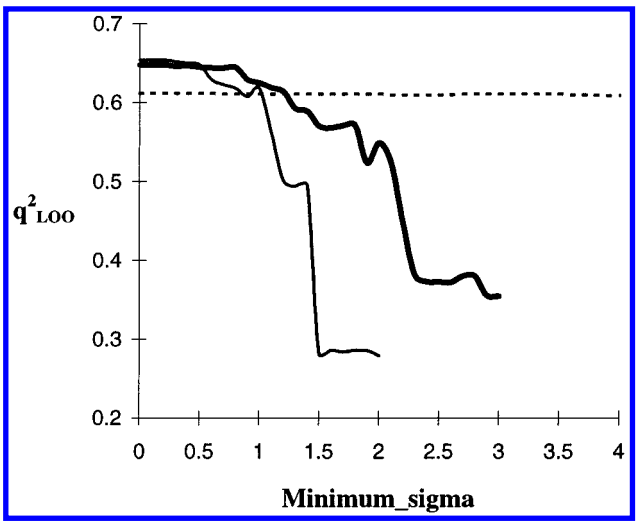

Figure 7. LOO cross-validation results from the minimum_sigma sensitivity test. LJ 6−12-based results are plotted as dotted line. INVOL-based results are represented as thin line for 1.2 Å and thick line for 1.5 Å van der Waals radius of hydrogen and probe-atom. All analyses were performed without exclusion of "interior" points at optimum number of principal components (4 PC for INVOL and 3 PC for LJ 6−12). INVOL computations are based on grid RED and LJ 6−12 computations on grid AUTO.

to its higher discriminative power in respect to the field values associated to distinct vertices of the 3D lattice. This result, in fact, is a reflection of the pronounced steepness of the Lennard-Jones potential and highlights the differences in slopes of the two functions. Moreover, using INVOL the number of points that contain information is increased compared with Lennard-Jones potential, since increased minimum_sigma values (which exclude more points with signal) reduce the statistical quality of the models.

Final stdev*coeff Contour Maps. Final 3D-QSAR models were carried out for all analyses involved in the van der Waals radii sensitivity test. Since the grid RED gave consistent results with INVOL field in respect to all tests performed, in Figure 8 we present the stdev*coeff contour plots for the 3D-QSAR models based on intersection volumes field computed with grid RED and Lennard-Jones 6−12 potential computed with grid RED_EXT. The INVOL-based model was carried out at 2.0 Å spacing, minimum_sigma of 0.0 Å^3, probe-atom and hydrogens vdW radii of 1.5 Å, no exclusion of "interior" points, and four PC. LJ 6−12-based model was obtained at 2.0 Å spacing, standard SYBYL/CoMFA conditions, and 3 PC. The overall similar disposition of the major features that govern favorable and unfavorable steric interactions confirms the fact that the proposed steric field can be a substitute for the classical one. However, subtle differences can be noticed between the two maps. Since the final models are often used to predict the biological activities for new compounds, the exact location of these features seems to play an important effect.

CONCUSIONS

The van der Waals envelopes intersection volumes (INVOL) field has proven to be an alternative replacement for the Lennard-Jones 6−12 potential commonly used to quantify the steric properties of a series of bioactive compounds and to identify relevant features that govern their interaction with corresponding biological targets within a CoMFA-type 3D-QSAR methodology.

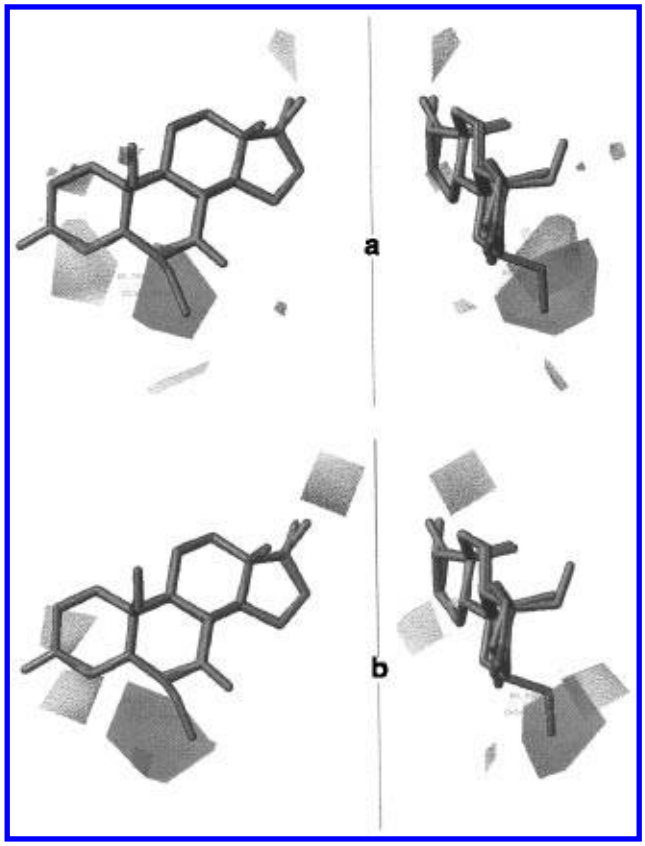

Figure 8. Final stdev*coeff contour plots (orthogonal views, hydrogens omitted for clarity) obtained with a. INVOL field and b. LJ 6−12 field. Field contours surround regions of high (green) and low (yellow) steric bulk tolerance. The most potent inhibitor (**36**: $pK_i = 2.85$) is shown in magenta and the most inactive one (**2**: $pK_i = -1.65$) in red (see Results and Discussions section for conditions).

The reported field presents a few notable advantages in comparison with classical Lennard-Jones potential: (1) lower sensitivity to the grid orientation and grid spacing, with an overall effect of 50−75% at 2.0 Å spacing and 75−90% at 1.0 Å spacing, for the series of compounds investigated; (2) reduced number of points of the 3D lattice and overall better statistic results; (3) no predefined cut-off and minimum_sigma threshold required; and (4) reduced model complexity by exclusion of some "interior" points. In addition, we consider that the circumvention of steric misfit by local deformability of both receptor binding site and ligand during ligand−receptor interaction represent a more realistic way to quantify this process. The relative simplicity and inexpensive computational demands make this steric field a promising alternative for routine CoMFA-based 3D-QSAR analyses. Integration of the field in other techniques that make use of the CoMFA-type steric description of molecular shape (e.g., 3D search of combinatorial libraries,[16,17] field-fit-based alignment,[1,15] docking techniques, to name just a few) could improve the accuracy of the results.

The optimal conditions that we suggest for CoMFA analyses involving INVOL steric field, based on the tests performed in this work, can be summarized as follows: grid position - minimal grid that embeds the compounds with a maximum of space economy; grid spacing - 2.0 Å (to save CPU-time) or 1.0 Å (for enhanced stability of the results to grid location); probe-atom and hydrogen atoms of the ligands van der Waals radii - 1.5 Å; "interior" points exclusion - INVOL higher than 7.0 Å^3; minimum_sigma - 0.0 Å^3. However, we recommend that some or all of the tests presented here should be performed for the particular set of compounds under investigation.

It should be pointed out that the field completely lacks the attractive terms and atom-specific properties other than van der Waals volume. In addition, INVOL still maintains the regularity of the 3D partition of the space surrounding the aligned set of ligands (a better partitioning is obtained with a hypermolecular description of 3D space in MTD approach).

Efforts are currently involved for inclusion of electrostatics with a van der Waals surface-directed cut-off, desolvation fields, and HINT[26] hydrophobicity fields along with intersection volumes field as well as validation of its predictive power through applications on diverse sets of bioactive flexible compounds, also including test sets.

ACKNOWLEDGMENT

The authors would like to thank Professor Zeno Simon (University of Medicine and Pharmacy, Timisoara, Romania) for many helpful discussions and Dr. Enrico Purisima and Hervé Hogues (Biotechnology Research Institute, Montréal, Canada) for technical assistance. This is National Research Council of Canada publication number 39962.

Supporting Information Available: The C++ program for calculating INVOL field, SPL codes for integrating the fields into SYBYL, and 3D coordinates with MNDO partial charges (SYBYL mol2 format) of the aligned set of 78 steroids are available on the World Wide Web. See any current masthead page for ordering information.

REFERENCES AND NOTES

(1) Cramer, R. D.; 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.
(2) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J. The Covariance Problem in Linear Reggresion. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.
(3) Dunn, W. J.; Wold, S.; Edlung, U.; Hellberg, S.; Gasteiger, J. Multivariate Structure-Activity Relationship Between Data from a Battery of Biological Tests and an Ensemble of Structure Descriptors: The PLS Method. *Quant. Struct.-Act. Relat.* **1984**, *3*, 131−137.
(4) Cramer, R. D.; Bunce, J. D.; Patterson, D. E.; Franke, I. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies. *Quant. Struct.-Act. Relat.* **1988**, *7*, 18−25.
(5) Cramer, R. D.; DePriest, S. A.; Patterson, D. E.; Hecht, P. In *3D-QSAR in Drug Design. Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 443−485.
(6) Simon, Z.; Chiriac, A.; Holban, S.; Ciubotariu, D.; Mihalas, G. I. *Minimum Steric Difference. The MTD Method for QSAR Studies*; Res. Stud. Press, Letchworth and Wiley: New York, 1984.
(7) Oprea, T. I.; Ciubotariu, D.; Sulea, T.; Simon, Z. Comparison of the Minimal Steric Difference (MTD) and Comparative Molecular Field Analysis (CoMFA) Methods for Analysis of Binding of Steroids to Carrier Proteins. *Quant. Struct.-Act. Relat.* **1993**, *12*, 21−26.
(8) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules To Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.
(9) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Method.* **1990**, *3*, 615−633.
(10) Cho, S. J.; Tropsha, A. Cross-Validated R^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.
(11) Kroemer, T. R.; Hecht, P. Replacement of Steric 6−12 Potential-Derived Interaction Energies by Atom-Based Indicator Variables in CoMFA Leads to Models of Higher Consistency. *J. Comput.-Aided Mol. Design* **1995**, *9*, 205−212.
(12) Kim, K. H.; Martin, Y. C. Direct Prediction of Dissociation Constants (pK_a) of Clonidine-like Imidazoles, 2-substituted Imidazoles and 1-Methyl-2-substituted-imidazoles from 3D Structures using a Com-

parative Molecular Field Analysis (CoMFA). *J. Med. Chem.* **1991**, *34*, 2056−2060.
(13) Greco, G.; Novellino, E.; Silipo, C.; Vittoria, A. Comparative Molecular Field Analysis on a Set of Muscarinic Agonists. *Quant. Struct.-Act. Relat.* **1991**, *10*, 298−299.
(14) Klebe, G.; Abraham, U. On the Prediction of Binding Properties of Drug Molecules by Comparative Molecular Field Analysis. *J. Med. Chem.* **1993**, *36*, 70−80.
(15) Clark, M.; Cramer, R. D.; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative Molecular Field Analysis (CoMFA). 2. Toward its Use with 3D-Structural Databases. *Tetrahedron Comput. Method.* **1990**, *3*, 47−59.
(16) Cramer, R. D.; Clark, R. D.; Patterson, D. E.; Freguson, A. M. Bioisosterism as a Molecular Diversity Descriptor: Steric Fields of Single "Topometric" Conformers. *J. Med. Chem.* **1996**, *39*, 3060−3069.
(17) Patterson, D. E.; Cramer, R. D.; Freguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behaviour: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(18) Muresan, S.; Sulea, T.; Ciubotariu, D.; Kurunczi, L.; Simon, Z. Van der Waals Intersection Envelope Volumes as a Possible Basis for Steric Interaction in CoMFA. *Quant. Struct.-Act. Relat.* **1996**, *15*, 31−32.
(19) Oprea, T. I.; García A. E. Three-Dimensional Quantitative Structure-Activity Relationships of Steroid Aromatase Inhibitors. *J. Comput.-Aided Mol. Design* **1996**, *10*, 186−200.
(20) The program SYBYL 6.2 is available from Tripos Associates, 1699 South Hanley Road, St. Louis, MO 63144.
(21) (a) Stewart, J. J. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Design* **1990**, *4*, 1−105. (b) The program MOPAC (Molecular Orbital Package) 6.0 is available from QCPE Creative Arts, Bldg. 181, Indiana University, Bloomington, IN, No. 455.
(22) Vinter, J. Extended Electron Distributions Applied to the Molecular Mechanics of Some Intermolecular Interactions. *J. Comput.-Aided Mol. Design* **1994**, *8*, 653−668.
(23) Ciubotariu, D.; Gogonea, V.; Iorga, I.; Deretey, E.; Medeleanu, M.; Muresan, S.; Bologa, C. New Shape Descriptors for Quantitative Treatement of Steric Effects. II. The Molecular Van der Waals Volume: Two Monte Carlo Algorithms. *Chem. Bull. Tech. Univ. (Timisoara)* **1993**, *38*, 83−92.
(24) Muresan, S.; Bologa, C.; Medeleanu, M.; Gogonea, V.; Dragos, D.; Ciubotariu, D. New Shape Descriptors for Quantitative Treatement of Steric Effects. III. Statistical Validation of the Monte Carlo Algorithm MON_CAVE_1. *Chem. Bull. Tech. Univ. (Timisoara)* **1994**, *39*, 47−56.
(25) Bondi, A. *J. Phys. Chem.* Van der Waals Volumes and Radii. **1964**, *68*, 441−451.
(26) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT: A New Method of Empirical Hydrophobic Field Calculation for CoMFA. *J. Comput.-Aided Mol. Design* **1991**, *5*, 545−552.
(27) Di Salle, E.; Giudici, D.; Ornati, G; Briatico, G; D'Alessio, R.; Villa, V; Lombardi, P. 4-Aminoandrostenedione Derivatives: A Novel Class of Irreversible Aromatase Inhibitors. *J. Steroid Biochem. Molec. Biol.* **1990**, *37*, 369−374.
(28) Di Salle, E.; Briatico, G.; Giudici, D.; Ornati, G; Zaccheo, T.; Buzzetti, F.; Nesi, M.; Panzeri, A. Novel Aromatase and 5alpha-Reductase Inhibitors. *J. Steroid Biochem. Molec. Biol.* **1994**, *49*, 289−294.
(29) Brueggemeier, R. W.; Moh, P. P; Ebrahimian, S; Darby, M. V. Steroidal Inhibitors as Chemical Probes of the Active Site of Aromatase. *J. Steroid Biochem. Molec. Biol.* **1993**, *44*, 357−365.
(30) Li, P.-K.; Brueggemeier, R. W. Synthesis and Biochemical Studies of 7-Substituted 4,6-Androstadiene-3,17-diones as Aromatase Inhibitors. *J. Med. Chem.* **1990**, *33*, 101−105.
(31) Johnston, J. O.; Wright, C. L.; Burkhart, J. P.; Peet, N. P. Biological Characterization of A-ring Steroids. *J. Steroid Biochem. Molec. Biol.* **1993**, *44*, 623−631.

CI9704221