FLUFF-BALL, A Template-Based Grid-Independent Superposition and QSAR Technique: Validation Using a Benchmark Steroid Data Set

Samuli-Petrus Korhonen,* Kari Tuppurainen, Reino Laatikainen, and Mikael Peräkylä

Department of Chemistry, University of Kuopio, P.O. Box 1627, FIN-70211, KUOPIO, Finland

Received February 10, 2003

The Flexible Ligand Unified Force Field (FLUFF) is a molecular mechanistic superposition algorithm utilizing a template structure, on top of which the ligand(s) are superimposed. FLUFF enables a flexible semiautomatic superimposition in which the ligand and the template are allowed to seek the best common conformation, which can then be used to predict the biological activity by Boundless Adaptive Localized Ligand (BALL). In BALL, the similarity of the electrostatic and van der Waals volumes of the template and ligand is evaluated using the template-based coordinate system which makes the FLUFF-BALL invariant as to the rotations and translations of the global coordinate system. When tested using the CBG (corticosteroid binding globulin) affinities of 31 benchmark steroids, the FLUFF-BALL technique produced results comparable to standard 3D-QSAR methods. Supplementary test calculations were performed with five additional data sets. Due to its high level of automation and high throughput, the FLUFF-BALL is highly suitable for use in drug design and in scanning of large molecular libraries.

INTRODUCTION

Since the introduction of the Hansch equation[1] in the 1960s, the number of algorithms available for quantitative structure activity (QSAR) studies has increased explosively. The classical QSAR[2] relies on 2D representation of the molecule and therefore neglects considerable amount of stereochemical information. In the late 1980s the QSAR gained a new dimension in both theoretical and practical sense as closely related CoMFA[3] and GRID[4] techniques were introduced. These molecular field-based techniques addressed some of the major deficiencies inherent in classical QSAR techniques and were soon widely adopted. In these techniques the steric and electrostatic fields surrounding the molecule are described using Lennard-Jones and Coulomb potentials at the vertices of a 3D grid surrounding the molecule. The grid data is then analyzed using statistical methods such as PCR[5] or PLS[6,7] to uncover the correlation between interaction energies and observed activity.

However, the 3D-QSAR methods usually require an accurate superposition of structures, which has proven to be their greatest weakness, as this procedure usually requires considerable human intervention and is generally regarded to be the most arduous and time-consuming phase of the 3D-QSAR analysis.[8] This severely limits the efficiency of 3D-QSAR techniques when dealing with large libraries of molecules. Therefore a fully automated computational[9−11] "sieve", capable of rapid browsing through vast molecular libraries and eliminating the nonactive compounds, would greatly reduce the amount of modeling work needed.

For the above reasons, considerable effort has been directed into automation of the superposition process. The resulting plethora of algorithms can be roughly divided into point- and property-based approaches. In point-based algorithms pairs of atoms or pharmacophores are usually superposed using a least-squares fitting.[12,13] The greatest limitation of these techniques is the need for predefined anchor points because in case of dissimilar molecules the generation of these points becomes problematic even though algorithms for automatic detection of such points have been proposed.[12] On the other hand, the property-based algorithms offer a much wider choice of descriptors, which include various molecular properties such as molecular shape and volume,[14−17] electron density,[18,19] charge distribution,[20] and many more.[21−23] However, all these techniques generate a metric, usually called "similarity index", which describes the degree of overlap between the structures being superposed. The actual superposition is a process in which the similarity index is optimized by using an appropriate mathematical method.[15,17,20,24] Melani et al.[25] have compiled an extensive summary of the different superposition algorithms available.

An alternative approach tries to circumvent the superposition problem by using QSAR descriptors which are sensitive to the 3D structure of the molecule but do not require structural superposition. Alignment-free QSAR techniques proposed by Broto,[26] Gasteiger,[27] and Clementi[8] use autocorrelation functions and neural networks to create coordinate independent QSAR descriptor, while WHIM[28−30] and CoMMA[31] rely on inherently coordinate independent descriptors. Also many different quantum chemistry-based descriptors such as TQSI,[32] MQSM,[32] and QS-SM[9] have been developed, most of which are based on the molecular orbital (MO) approach.[33] Spectral information can also be used as a basis for QSAR analysis, as is demonstrated by the success of CoSA,[34] EVA,[35−38] and EEVA[39] approaches, which utilize IR, RAMAN, NMR, and MO energy spectra for QSAR. Many so-called E-state techniques[40,41] utilizing parameters derived from electrotopological structure of molecules have been proposed. All these alignment-free techniques could be called "21/2D-QSAR" as they are sensitive to the 3D-

* Corresponding author phone: +358-17-163275; fax: +358-17-163259; e-mail: Samuli-Petrus.Korhonen@uku.fi.

10.1021/ci034027o CCC: $25.00
Published on Web 08/27/2003

structure of the molecule, but they do not directly depend on the 3D-structure. Therefore, the descriptors and the results of the models are often more difficult to interpret than in the case of 3D-QSAR techniques.[8]

In addition to the problems caused by the superposition, 3D-QSAR techniques that depend on a global grid are also susceptible to errors rising from translation and rotation of structures.[42] Despite its limitations, the use of QSAR has spread. This has led to the development of new derivative techniques which include CoMSIA[43] and SOMFA.[44] Also so-called multidimensional 4D- and 5D-QSAR techniques which use an ensemble of conformations instead of one active conformation have been proposed.[45,46] The molecular similarity indices proposed by Carbo[47] and Hodgkin[48] have been also applied to QSAR.[11,49−51]

How different the QSAR techniques seem to be, the QSAR descriptors exhibit strong mutual correlations, and therefore only modest increase in predictive ability is achieved by combining different descriptors. It seems likely that a universally applicable QSAR technique is not to be found soon, despite intensive efforts.

In this paper, a template-based grid-independent QSAR method "FLUFF-BALL" is outlined. This novel technique is suitable for fast preliminary screening of molecule libraries. It consists of a semiautomatic superposition algorithm (FLUFF) and an accompanying QSAR technique (BALL). The algorithms are designed to provide maximum throughput with minimum amount of human intervention. The basis of the semiautomatic superposition algorithm is a novel field-fitting technique called Flexible Ligand Unified Force Field (FLUFF). By utilizing this technique, it is possible to superimpose a set of molecules on top of a given molecule, the template, as there is an energy contribution for the similarity of the steric and electrostatic fields of the ligand and the template. During superposition the molecules can be made fully flexible to allow maximal adaptation, but selected atoms can also be frozen. The Boundless Adaptive Localized Ligand (BALL) is a template-based local coordinate modification of the CoMFA-technique, which complements FLUFF by allowing the flexible superposition to be used as a basis for a structure−activity analysis.

DESCRIPTION OF THE METHOD

Semiautomatic Superposition. In the semiautomatic superimposition algorithm molecules are aligned with the template molecule in such a manner that the similarity of the van der Waals and electrostatic volume is maximized. For the FLUFF algorithm we must define the concepts of physical and logical molecules. Physical molecules are defined as a collection of atoms interconnected by bonds, whereas the logical molecules are an arbitrary collection of atoms. The template and the ligand used in FLUFF superposition are logical molecules, but there are some guidelines which must be observed when constructing these molecules. First and foremost an atom can belong to only one logical molecule at a time. It is also unadvisable to define two logical molecules within one physical molecule as the nonbonded interactions would be disrupted but the bonded interactions would be maintained. One can include only a part of a physical molecule in a logical molecule and left a part of it unassigned, but in that case the nonbonded interactions between the assigned and unassigned part of the molecule would be disrupted. However, this will not interfere with the FLUFF superposition. On the other hand, several physical molecules can be assigned to the same logical molecule without any adverse effects.

The superimposition force field was implemented by utilizing a modified Merck Molecular Force Field (MMFF-94).[52−56] This enables use of several well-documented and tested computational techniques of molecular mechanics. The actual superimposition is accomplished by performing a geometry optimization by using the superimposition force field. Alternatively molecular dynamics (MD) or Monte Carlo-search (MC) can be utilized.

The energy equation (eq 1) of MMFF94 can be divided into two separate components describing bonded and nonbonded interactions as follows (eq 2)

$$E_{\mathrm{MMFF94}} = E_{\mathrm{B}} + E_{\mathrm{NB}} \tag{1}$$

where

$$E_{\mathrm{B}} = \sum \mathrm{EB}_{ij} + \sum \mathrm{EA}_{ijk} + \sum \mathrm{EBA}_{ijk} + \sum \mathrm{EOOP}_{ijkl} + \sum \mathrm{ET}_{ijkl}$$

$$E_{\mathrm{NB}} = \sum \mathrm{EvdW}_{ij} + \sum \mathrm{EQ}_{ij} \tag{2}$$

In the superimposition, in addition to the bonded interactions and the nonbonded interaction terms within the logical molecule, an E_{SP}-term is generated to describe the similarity of van der Waals (ESvdw) and electrostatic volume (ESeel) of the ligand and the template (eq 3).

$$E_{\mathrm{SP}} = \sum \mathrm{ESvdw}_{ij} + \sum \mathrm{ESeel}_{ij} \tag{3}$$

The ESvdw_{ij} and ESeel_{ij} functions must not contain asymptotic points, and they must have an unambiguous derivative at all points. When these two conditions are met, the functions can be chosen arbitrarily. However, for the purposes of superposition and for optimization techniques in general, a sigmoidal shape generated by the function e^{x^n} is beneficial. These Gaussian type functions are also used in SEAL superposition algorithm[57] often used in conjunction with CoMFA.

The ESvdw Term. The following functional form is used for the ESvdw term (eq 4):

$$\mathrm{ESvdw}_{ij} = S_i S_j C_{\mathrm{vdw1}} e^{C_{\mathrm{vdw2}} r_{ij}{}^{n\mathrm{vdw}}}$$

$$D_{\mathrm{v}}'(r) = S_i S_j C_{\mathrm{vdw1}} e^{C_{\mathrm{vdw2}} r_{ij}{}^{n\mathrm{vdw}}} C_{\mathrm{vdw2}} n r^{n_{\mathrm{vdw}}-1} \tag{4}$$

where S_i and S_j are the scaling factors of the atoms i and j (usually $S_i = S_j = 1$), C_{vdw1} and C_{vdw2} are constants defined for interaction between atoms of type i and j (C_{vdw1} and $C_{\mathrm{vdw2}} < 0$), r_{ij} is the distance between atoms i and j, and n_{vdw} is the exponent defined for interaction between atoms of type i and j (usually $n_{\mathrm{vdw}} = 2$).

When selecting field constants C_{vdw1} and C_{vdw2} some care must be taken to ensure that the minimum occurs at the centers of the atoms, as too wide a fitting function can lead to a set of erroneous minima to be generated between the atoms. The phenomenon of false minima can be seen especially clearly when superimposing two benzene rings.

If the constant C_{vdw2} allows the fitting field to span a too great distance a minimum is created between the template atoms, and in the resulting fit ligand atoms are located exactly on the middle point between two template atoms.

The ESeel Term. For the term ESeel, the following functional form is used:

$$\mathrm{ESeel}_{ij} = S_i S_j q_i q_j C_{eel1} e^{C_{eel2} r_{ij}^{n_{eel}}}$$

$$D_e'(r) = S_i S_j q_i q_j C_{eel1} e^{C_{eel2} r_{ij}^{n_{eel}}} C_{eel2} n r_{ij}^{n_{eel}-1} \qquad (5)$$

where S_i, S_j, C_{eel1}, C_{eel2}, r_{ij}, and n_{eel} are analogous to the ES_{vdw} (eq 4) and q_i and q_j are the charges of atoms i and j.

This function exhibits an attractive behavior in case of two similar charges and a repulsive force when two opposite charges are placed in proximity. This can at times hamper efficient superposition, particularly in case of highly charged side chains, which are usually also flexible. To resolve the problem, a cutting option was included in the superposition algorithm, which removes all repulsive interactions by setting all positive energy terms to zero, but leaves all negative terms, i.e., attractive forces, unchanged. Also values of ESeel_{ij} function are highly dependent on the charges present in the molecules, which in part complicate balancing the strength of this function. In case of the ESvdw the per-atom strength of the function remains quite stable from ligand to ligand, but in case of ESeel it may occasionally be necessary to adjust the field strength by adjusting the constants C_{eel1} and C_{eel2}.

QSAR by Boundless Adaptive Localized Ligand (BALL). For full utilization of flexible superposition an accompanying QSAR method is needed as standard CoMFA-techniques are dependent on a global coordinate system, which means that all molecules must be aligned to the same spatial coordinates. In the flexible superposition the demand for a uniform global positioning cannot be met. Moreover, as the superposition may also be performed using a fully relaxed template, minor changes may also occur in the conformation of the template thus rendering standard CoMFA-techniques useless. To overcome these problems the Boundless Adaptive Localized Ligand (BALL) uses an internal coordinate system tied to the template. The grid vertices are placed at the atomic centers of the template molecule, thus rendering the internal coordinates immune to global translations and rotations. Thus minor changes in the template conformation do not necessarily have a major adverse effect on the accuracy of the model as anchor points of the local grid are tied to the template and transform with the template. In the case of extremely flexible molecules the changes in the template conformation cause a considerable cumulative error and reduce the accuracy of the model. In these cases the template can be locked, and the superposition can be performed using only a flexible ligand. When flexible side chains are connected to a rigid or semirigid body, the conformational changes are minor and a fully flexible superposition may be utilized.

The BALL model is based on two different functions. The first one describes the van der Waals similarity of the template and ligand molecules by generating three terms which are (i) the template atom's own volume by self-overlap, (ii) the common volume of the template atom and ligand atoms, and (iii) the residual ligand volume allocated to the template atom. The second function reflects the similarity of charges in the molecules by generating three electrostatic terms representing (i) the template electrostatic field at the center of the template atom, (ii) the difference of template's and ligand's electrostatic fields at the center of the template atom, and (iii) the residual field difference allocated from ligand atoms. Thus the complete descriptor consists of 6n terms, where n is the number of template atoms. The terms describing the template atom's own volume and the electrostatic field at the center of the template atom are included primarily for scaling purposes.

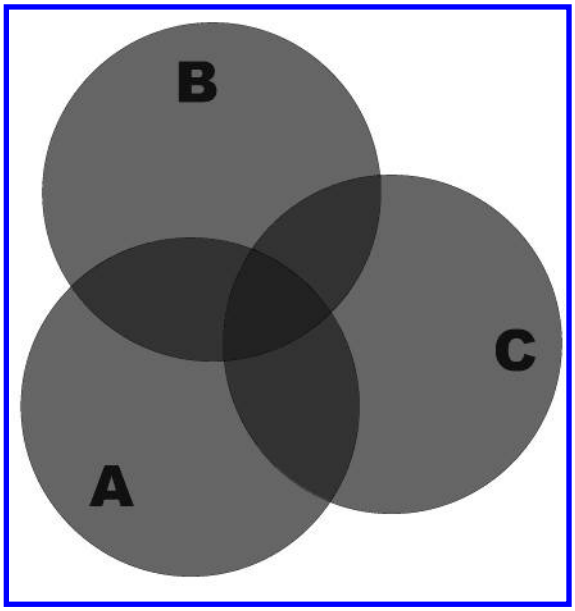

Figure 1. Schematic representation of overlapping atoms.

van der Waals Terms. The common volume, CV, of two atoms A and B defined by a Gaussian primitive (GTF) density function is

$$\mathrm{CV}_{AB} = \int\int\int D_A(x,y,z) \cdot D_B(x,y,z)\, dx\, dy\, dz = \int\int\int C_{A1} e^{C_{A2} r_a^2} \cdot C_{B1} e^{C_{B2} r_b^2}\, dx\, dy\, dz \qquad (6)$$

where C_{A1}, C_{B1}, C_{A2}, and C_{B2} are constants, and r_a and r_b are the distances from the centers of the atoms.

According to the Gaussian product rule, the product of two Gaussian primitives is a new Gaussian primitive. Thus the product of two primitives is

$$\mathrm{CV}_{AB} = C_{A1} C_{B1} \cdot e^{(C_{A2} C_{B2} r_{AB}^2 / C_{A2} + C_{B2})} \left(\frac{\pi}{-(C_{A2} + C_{B2})}\right) \sqrt{\frac{\pi}{-(C_{A2} + C_{B2})}}, \quad C_{A2} \text{ and } C_{B2} < 0 \qquad (7)$$

where r_{AB} is the distance between the centers of the atoms.

All higher order intersections can be returned to the basic intersection of two Gaussian primitives. For example the common volume of atoms A, B, and C (see Figure 1) can be evaluated by first generating the Gaussian function representing the common volume of atoms A and B and then by computing the common volume of this new function and atom C. Let us also examine a case where we are interested in the common volume of atom A in respect to atoms B and C. First we take the common volume of A and B and the

common volume of *A* and *C*. In addition a correction term is needed to take into account the common volume of *A*, *B*, and *C*, and thus $CV_{ABC} = A \cap B + A \cap C - A \cap B \cap C$. In a general case, the intersection terms which have an even number of atoms increase the common volume, and the terms which have an odd number of atoms decrease the common volume.

The volume of atom self-overlap can be computed by the following equation:

$$CV_{AA} = C_{A1}^{2} \cdot e^{(C_{A2}^{2} r_{AA}^{2}/2C_{A2})} \left(\frac{\pi}{-2C_{A2}}\right) \sqrt{\frac{\pi}{-2C_{A2}}}, \quad C_{A2} < 0 \text{ and } r_{AA} = 0$$

$$CV_{AA} = C_{A1}^{2} \left(\frac{\pi}{-2C_{A2}}\right) \sqrt{\frac{\pi}{-2C_{A2}}} \tag{8}$$

As the volume given by the self-overlap formula (eq 8) is different than the volume given by the volume integration, it becomes evident that the volume calculations are comparable only within the same order of intersection. Therefore it is necessary to create a conversion formula with which it is possible to generate comparable volumes of intersections. As the first order can also be computed using the self-overlap formula to yield the second-order equivalent of the atoms volume, the natural choice is to take the second order of intersection as the base level to which all the other intersections are converted. Approximation of a common volume term of the order *n*-1 is computed from the term of the order *n* as demonstrated by the equation (eq 9)

$$CV_{I} = \frac{\sum CV_{n-1}}{n} \cdot \left(\frac{CV_{n}}{\left(\frac{\sum CV_{n2}}{n} \right)} \right) \tag{9}$$

in which *n* is the order of the intersection, CV_{n-1} is the intersection of the order *n*-1 generated by excluding one atom at a time from the original set, CV_n is the *n*th-order intersection, and CV_{n2} is the pseudo-*n*th-order intersections generated from CV_{n-1} intersections by in turn including one atom twice in the intersection. This formula enables the conversion of the third- and fourth-order intersections to second order with a reasonable accuracy. However, in test runs it was observed that intersections higher than third order did not improve the accuracy of the model. Therefore the BALL algorithm only evaluates the first-, second-, and third-order intersections.

In evaluating BALL van der Waals terms the template atom's own volume is computed as a self-overlap by (eq 8). Then the common volume of the template atom and ligand atoms is computed as an intersection of template atom A and ligand atoms L_1-L_n using (eq 7) and (eq 9). The residual volume of the ligand atom means the volume of the atom not covered by the template atoms, and it is evaluated in a similar manner as the "free volume" of the template atom, but now the roles of the ligand and the template are inversed. This "residual volume" is allocated to the template atoms by separately evaluating (eq 10) for each template atom

$$F_{res} = F_{diff} e^{C_2 r_{tl}} \tag{10}$$

where F_{res} is the residual field allocated to template atom *t*, F_{diff} is the field difference at ligand atom *l*, C_2 is a constant, and r_{tl} is the distance between the template atom *t* and the ligand atom *l*.

To summarize, the following three van der Waals parameters are evaluated for each template atom in order to generate the QSAR descriptor: (1) template atoms own volume computed with self-overlap (eq 8), (2) the common volume of the template atom and ligand atoms, and (3) the residual volume of the ligand atom allocated using (eq 10).

Electrostatic Terms. The electrostatic potential *V*, which is used to measure the electrostatic similarity, is

$$V = \frac{1}{4\pi\epsilon_0} \cdot \frac{Q}{r_p}$$

where ϵ_0 is permeability of vacuum ($8.85419 \cdot 10^{-12}$ F/M), *Q* is charge, and r_p is the distance between the charge and point *p*. However, when computing the QSAR descriptor, the constant term can be omitted, and thus

$$V = \frac{Q}{r_p}$$

which has one asymptotic point at $r_p = 0$, and it is therefore necessary to introduce a limiting factor which prevents the *r* from reaching zero. In this case a natural limiting factor is the van der Waals radius of the atom used as the point of origin

$$Sq_{ab} = \begin{cases} \left(\frac{q_b}{r_{ab}}\right) r_{ab} \geq r_a \\ \left(\frac{q_b}{r_a}\right) r_{ab} < r_a \end{cases} \tag{11}$$

where q_b is the charge of atom *b*, r_{ab} is the distance between the atoms *a* and *b*, and r_a is the van der Waals radius of atom *a*. The field projected by an arbitrary group of atoms *j* at the center of atom *a* is

$$Sq_{aj} = \sum_{b=j_1}^{b=j_n} Sq_{ab} \tag{12}$$

Electrostatic terms of the BALL descriptor are obtained by computing the electrostatic field of the template at the center of a template atom with eq 12. Then the difference of electrostatic potential is obtained by evaluating the field projected by ligand with eq 12 and computing the difference. The residual potential of the ligand atom means the difference in electrostatic potential projected by the ligand and by the template at the center of the ligand atom, and it is evaluated in a similar manner as the field difference terms of the template but now inversing the roles of the ligand and the template. This potential is then allocated to the nearest template atoms with (eq 10) as was done with the van der Waals terms.

To summarize, the following three electrostatic parameters are evaluated for each template atom in order to generate the QSAR descriptor: (1) the electrostatic potential projected at the center of the atom by the template molecule, (2) the

difference of potential projected by the template and by the ligand at the center of the atom, and (3) the residual electrostatic potential of the ligand atom allocated by equation (eq 10).

Implementation. The FLUFF algorithm and the BALL were implemented utilizing the MMS software, a molecular mechanics program running under Microsoft Windows originally developed for use with PERCH NMR software (www.perchsolutions.com) at the Department of Chemistry, University of Kuopio. Though FLUFF-BALL is implemented in the framework of the MMS, the algorithm is rather independent, enabling an easy transfer to other software packages and even to a standalone version. The FLUFF-BALL algorithm is a part of the 2003 release PERCH NMR software which is available free of charge for academic users. The program is also available upon request from the corresponding author.

The core of the program and all the novel algorithms described here were written in ANSI C++ using standard STL libraries in order to ensure that the code is easily portable to other environments. The molecular graphics are generated utilizing the OpenGL and GLUT libraries. The user interface is coded with MFC classes using Microsoft Visual C++ 6.0 SP5.

For ESvdw and ESeel, six user defined parameters C_{vdw1}, C_{vdw2}, n_{vdw}, C_{eel1}, $C_{eel2,}$ and n_{eel} (eqs 4 and 5) are required. These six parameters are given in a user editable text file, so that it is possible to specify unique values for each atom type pair or to specify values for a whole range of atom types. These values are read during initialization of the program, and they remain constant during the lifetime of the program instance.

The BALL requires four user defined parameters. The first two parameters are VdW_const and VdW_radiusIntensity, which correspond to the C_1 and C_2 constants of the GTF (eq 6) and control the behavior of the van der Waals similarity function (eq 7). The last two, VdW_dispersion and EEL_dispersion, control the allocation of orphan van der Waals and electrostatic density to the template atoms (eq 10). They both correspond to the C_2 constant of the GTF.

Statistical Methods. For the model building phase of PLS analyses the following abbreviations are used throughout this paper: CV = cross-validation, LOO = leave-one-out CV, PRESS = predictive residual sum of squares, S_{PRESS} = cross-validated standard error of prediction (eq 13), and Q^2 = cross-validated correlation coefficient (eq 14)

$$S_{\mathrm{PRESS}} = \sqrt{\frac{\mathrm{PRESS}}{n - c - 1}} \tag{13}$$

$$Q^2 = 1 - \frac{\sum (y_{\mathrm{obs}} - y_{\mathrm{pred}})^2}{\sum (y_{\mathrm{obs}} - y_{\mathrm{mean}})^2} = 1 - \frac{\mathrm{PRESS}}{\sum (y_{\mathrm{obs}} - y_{\mathrm{mean}})^2} \tag{14}$$

where n is the number of compounds, and c is the number of the principal components extracted, i.e., the S_{PRESS} value is weighted so that it penalizes models with a high number of principal components. For fitted models, R^2 = squared correlation coefficient, SE = standard error, and F = Fischer test for significance. For external test sets R^2_{ex} = squared correlation coefficient, $|\Delta_{av}|$ = mean absolute deviations, Pr-R^2 = predictive R^2-score (eq 15), and SDEP = standard error of prediction (eq 16)

Table 1. Benchmark Steroid Set[a]

code	compound	activity log[K]
M1	aldosterone	6.279
L2	androstanediol	5.000
L3	androstenediol	5.000
L4	androstenedione	5.763
L5	androsterone	5.613
H6	corticosterone	7.881
H7	cortisol	7.881
M8	cortisone	6.892
L9	dehydroepiandrosterone	5.000
H10	deoxycorticosterone	7.653
H11	deoxycortisol	7.881
M12	dihydrotestosterone	5.919
L13	estradiol	5.000
L14	estriol	5.000
L15	estrone	5.000
L16	etiocholanolone	5.255
L17	pregnenolone	5.255
L18	17-hydroxypregnenolone	5.000
H19	progesterone	7.380
H20	17-hydroxyprogesterone	7.740
M21	testosterone	6.724
H22	prednisolone	7.512
H23	cortisol 21-acetate	7.553
M24	4-pregnene-3,11,20-trione	6.779
H25	epicorticosterone	7.200
M26	19-nortesterone	6.144
M27	16α,17-dihydroxy-4-pregnene-3,20-dione	6.247
H28	17-methyl-4-pregnene-3,20-dione	7.120
M29	19-norprogesterone	6.817
H30	11β−17,21-trihydroxy-2α-methyl-4-pregnene-3,20-dione	7.688
M31	11β−17,21-trihydroxy-2α-methyl-9α-fluoro-4-pregnene-3,20-dione	5.797

[a] Code: L means low activity, M medium activity, and H high activity. Compound H11 was used as a template.

$$\text{Pr-}R^2 = \frac{\mathrm{SD} - \mathrm{PRESS}}{\mathrm{SD}} \tag{15}$$

$$\mathrm{SDEP} = \sqrt{\frac{\mathrm{PRESS}}{n}} \tag{16}$$

where SD is the sum of squared deviations between the activities of molecules in the test set and the mean affinity of the training set molecules.

RESULTS

Model Development and Validation. The FLUFF-BALL technique was validated by performing an extensive series of tests utilizing a widely used benchmark steroid set containing 31 molecules whose binding affinity for the CBG-protein is measured. Many authors[27,58,59] have pointed out that the majority of early works employing this data set contained incorrect structures. Therefore this work utilized a corrected set created for the evaluation of EEVA.[39] The benchmark set presented in Table 1 is divided into a training set consisting of 21 molecules (**1**−**21**) and a test set of 10 molecules (**22**−**31**). All structures were optimized with the semiempirical AM1 method using AMPAC software (QCPE No. 506, version 2.11). All charges were set to the values computed with the AM1 method.

Table 2. Summary of the PLS Runs

	Fix	*Flex*	*Mix*
Q^2 average ± SD[a]	0.738 ± 0.025	0.673 ± 0.033	0.737 ± 0.038
Q^2 minimun − Q^2 maximum	0.691−0.801	0.639−0.744	0.626−0.775
average optimum # of components[b] ± SD	2.176 ± 0.805	2.039 ± 1.126	4.277 ± 0.961
average SDEP[c] for 1 component models ± SD	0.708 ± 0.073	0.752 ± 0.062	0.725 ± 0.102
average SDEP for 2 component models ± SD	0.638 ± 0.033	0.743 ± 0.050	0.707 ± 0.083
average SDEP for 3 component models ± SD	0.702 ± 0.033	0.768 ± 0.052	0.698 ± 0.043
average SDEP for 4 component models ± SD	0.768 ± 0.046	0.824 ± 0.083	0.704 ± 0.035
average SDEP for 5 component models ± SD	0.827 ± 0.068	0.870 ± 0.083	0.709 ± 0.066

[a] Standard deviation. [b] Optimum number of components as reported by SYBYL. [c] Standard error of prediction.

To evaluate the performance of the FLUFF superposition algorithm, three separate sets of superposed molecules were generated. The first set was to be a benchmark set, and therefore a normal rigid superimposition using the SYBYL 6.5 software (Tripos, Inc.) was performed. As the molecules have a common steroid backbone, it was used for the generation of anchor atom pairs. This set will be referred to as the *Fix* set. For the FLUFF superposition the compound **H11** was selected to be the template structure. The second set was superimposed using the FLUFF superposition algorithm with the template locked to its AM1 optimized conformation and with the ligand free to adapt its conformation to the shape of the template. This set is referred to as the *Mix* set. The third set was also superimposed using the FLUFF, but this time both the template and the ligand were fully flexible, hence the name *Flex* set.

During the first tests of the semiautomatic superposition, it was discovered that in several cases the hydrogen atoms of the steroid molecules formed a "local minimum barrier" around the body of the molecule thus interfering with the optimal alignment of the molecules. Therefore the superposition was performed in two separate phases. In the first phase hydrogen atoms were excluded from the FLUFF field, which made them totally transparent to the atoms of a different logical molecule. The other atoms such as carbon and oxygen were allowed to see each other. In the second phase, all atoms were included in the FLUFF field. The two-phase optimization system was automated and is now a standard feature of the FLUFF algorithm. In the superposition of the *Mix* and *Flex* sets 2500 optimization steps were used for the first and second phases in order to ensure that the molecules have sufficient time to settle to their optimum conformations. In practice, the analysis of trajectory files showed that after 400−500 steps the energy gradient was less than 10^{-5}, and the changes in the conformation were neglible. Only the rough initial superposition was performed manually, and the FLUFF algorithm was allowed to find the optimum superposition without any human intervention.

The behavior of the BALL algorithm is controlled by four parameters, which affect the predictive ability: VdW_const and VdW_radiusIntensity define the shape and strength of the van der Waals density function, and VdW_dispersion and EEL_dispersion control how the field difference at ligand atoms is distributed to the template atoms. The VdW_const is the least significant as it controls only the intensity of the van der Waals function and not the shape of the function. This parameter is included only for scaling purposes, and its value can usually be set to 1. Naturally, the VdW_radiusIntensity parameter is limited to values between 0 and 1: zero means that the van der Waals field is immediately quenched and has no effective radius, a value of one means that the van der Waals field will propagate to infinity and will never dissipate. The dispersion terms can range from 0 to infinity where zero means that the dispersion field is constrained to a singular point and no real dispersion is performed and the infinite value results in a uniform dispersion field over the whole model space. Empirical tests showed that the values of the dispersion constants should not exceed 0.5, as the dispersion field will then become too diffuse to have any real value for prediction.

As the optimum values of the BALL parameters were not known, an extensive optimization procedure was performed. The value of VdW_const was left at one in all validation runs because earlier testing had indicated that no scaling is needed. For VdW_radiusIntensity values of 0.950, 0.900, 0.850, 0.800 0.750, 0.700, 0.650, 0.600, 0.550, 0.500, 0.250, 0.125, 0.075, 0.050, and 0.025 were evaluated. For VdW_dispersion and EEL_dispersion values of 0.500, 0.250, 0.125, 0.075, and 0.050 were evaluated. Altogether, this yields a total of 375 unique combinations of parameters. The BALL descriptors were calculated for all three data sets using the aforementioned parameter groups resulting in 1125 different descriptors. With the SYBYL 6.5 software, a PLS fit was performed for all descriptors using a Leave-One-Out (LOO) cross-validation without any scaling of the data. The scaling methods offered by SYBYL were also tested, but these runs resulted in inferior models. The maximum number of principal components was set at 5 as it conforms with the generally accepted one-quarter rule, i.e., the number of PCs should not exceed $n/4$, where n is the number of molecules.

The results of the PLS runs are summarized in Table 2. In all of the models a high Q^2 value was achieved, while the optimum number of components ranged from 1 to 5. The average Q^2 values of the *Mix* and *Fix* sets were almost identical, while the *Flex* set produced a markedly lower value. The *Mix* model produced a systematically higher number of components than the *Fix* and *Flex* models, and the number of components would have been considerably higher if the limit of five components had been removed.

The Q^2 values of the models are presented in Figures 2−4 as a function of the BALL parameters VdW_radiusIntensity and VdW_dispersion. The parameter EEL_dispersion was omitted because it had only minor effect on the Q^2 values. The most striking feature is the remarkable similarity between the *Fix* and the *Flex* models. The overall saddle shape of the Q^2 surface is nearly identical. On the other hand, the *Mix* model produced a totally different profile. The high Q^2 values form a plane in the lower end of the VdW_radiusIntensity values instead of a saddle shape in the high end.

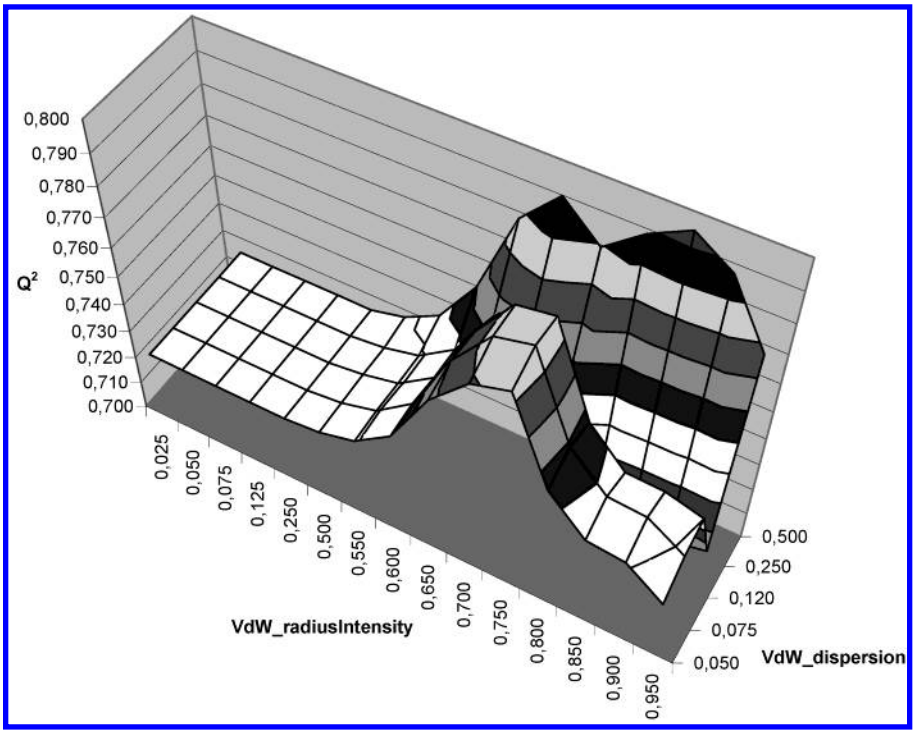

Figure 2. Q^2 values of the *Fix* model plotted against parameters VdW_RadiusIntensity and VdW_dispersion.

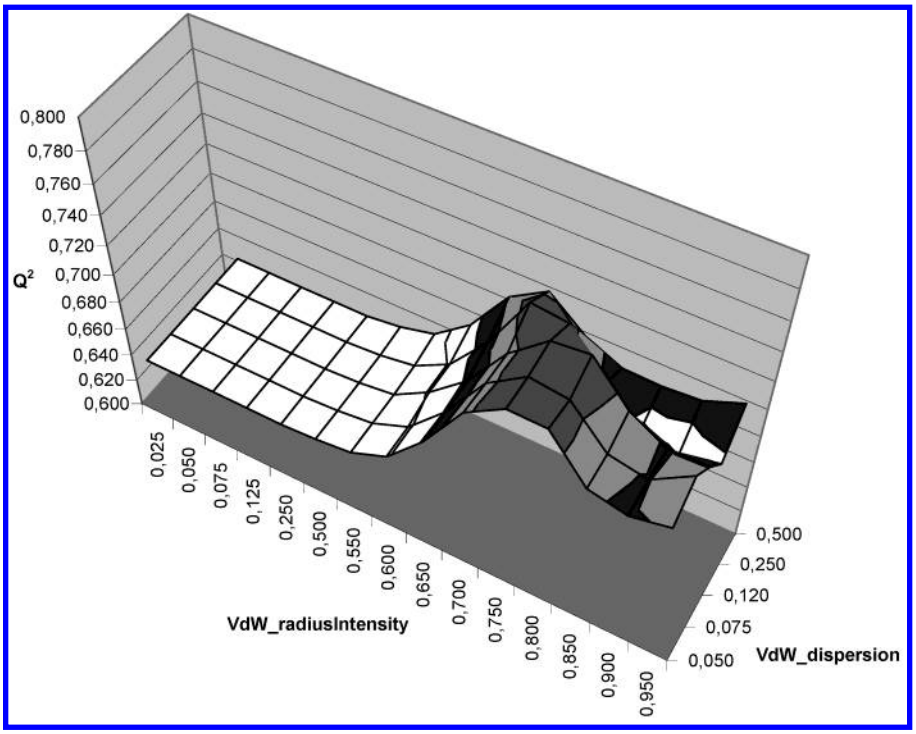

Figure 3. Q^2 values of the *Flex* model plotted against parameters VdW_RadiusIntensity and VdW_dispersion.

A parameter set was selected from the *Mix*, *Fix*, and *Flex* sets to be used in further studies. The properties of the models selected are listed in Table 3. Each of these three models was evaluated by first running a scrambling test in which the observed data are randomized. With this procedure it is possible to eliminate the internal correlation of the descriptor. All tests were run using the MATLAB program release 12 (The MathWorks, Inc.) by using an in-house written script. The scrambling runs with 2500 different randomizations resulted in an average Q^2 ($\pm$SD) value of -0.226 (± 0.214) for the *Mix* set with the maximum of five components. The average Q^2 ($\pm$SD) values of -0.153 (± 0.175) and -0.124 (± 0.159) were computed for the *Fix* and *Flex* sets, respectively. As expected, the BALL descriptors lose correlation with biological activity when the observed data is scrambled, thus indicating that the correlation with the correct data is not fortuitous.

In the standard steroid set molecules **1**−**21** constitute the training set and molecules **22**−**31** form the test set. The statistical stability of the model can be further evaluated by generating random sets of 21 molecules to be the training sets and 10 molecule sets for test sets. This procedure yields

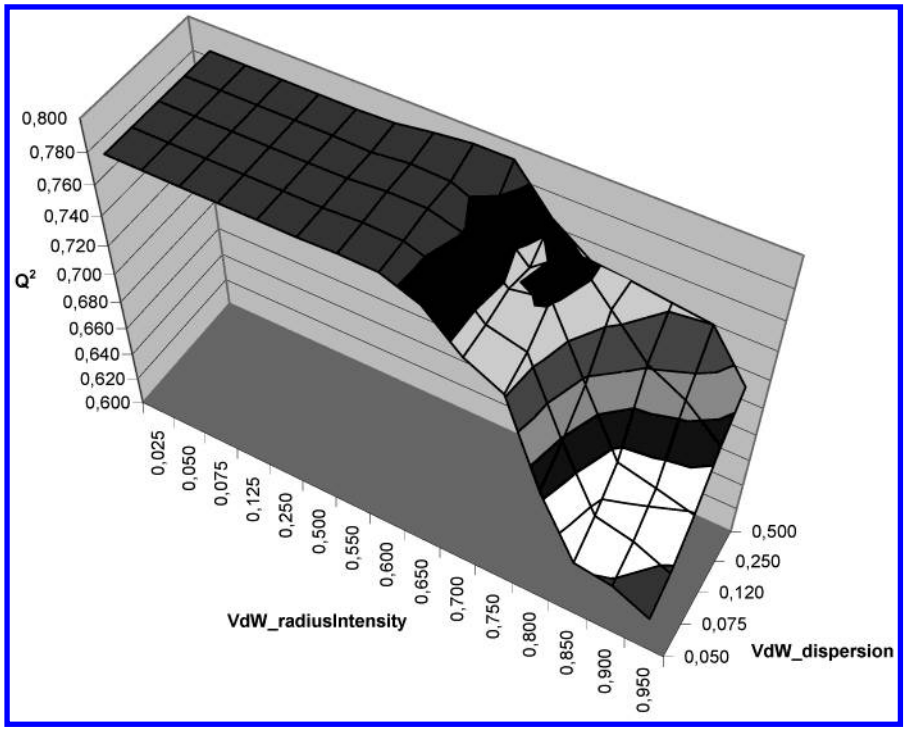

Figure 4. Q^2 values of the *Mix* model plotted against parameters VdW_RadiusIntensity and VdW_dispersion.

Table 3. Properties of the Parameter Sets Selected for Further Study

set	VdW_Radius-Intensity	VdW_ dispersion	EEL_ dispersion	Q^2	SDEP	Opt #[a]
Fix	0.700	0.500	0.500	0.801	0.552	2
Flex	0.750	0.120	0.250	0.733	0.678	4
Mix	0.125	0.120	0.120	0.772	0.647	5

[a] Opt # means the optimum number of components as reported by SYBYL.

a considerably better estimate of the models predictive ability than the validation based on a single set. For testing of the FLUFF-BALL 2500 random sets were generated and evaluated again using an in-house MATLAB script. The average statistical descriptors are presented in Table 4 with their standard deviations. As can be readily seen from the results the *Mix* model gives the best average prediction and the *Flex* set gives the worst results. Again the *Mix* set generates the highest amount of components, and its optimum number of components is clearly limited by the maximum of five components. When the same run was performed with the maximum number of components set to 20, there are 1390 models for which the optimum number of components was higher than five, and the highest number of components generated is 11. Distribution of the optimum number of components in the *Mix* set is shown as a histogram in Figure 5. A most interesting pattern of two separate clusters of optimum components is observed. When the scrambling run is done to the *Mix* set with the maximum number of

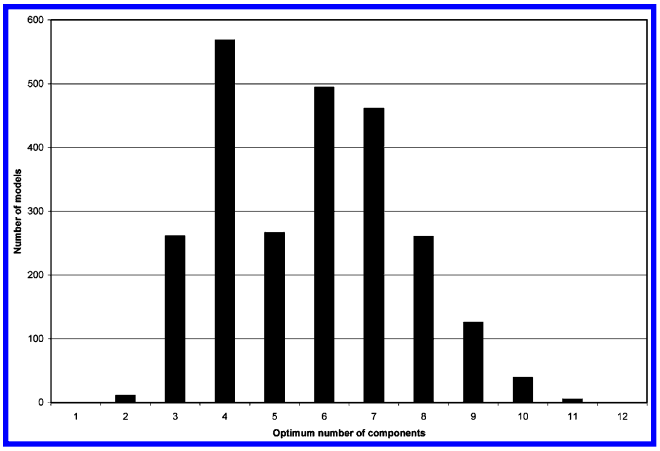

Figure 5. Histogram of the optimum number of components for scrambled *Mix C20* models.

Table 4. Average Statistical Descriptors of the Models Generated from 2500 Random Teaching and Prediction Sets

	Fix	*Flex*	*Mix C5*[a]	*Mix C20*[b]
$S_{PRESS} \pm$ SD	0.738 ± 0.110	0.780 ± 0.108	0.698 ± 0.070	0.672 ± 0.084
$Q^2 \pm$ SD	0.573 ± 0.134	0.540 ± 0.138	0.664 ± 0.084	0.710 ± 0.100
Opt # ± SD	2.415 ± 0.702	2.960 ± 0.851	4.303 ± 0.762	5.725 ± 1.826
$R^2 \pm$ SD	0.742 ± 0.109	0.772 ± 0.126	0.915 ± 0.053	0.943 ± 0.056
SE ± SD	0.569 ± 0.137	0.534 ± 0.147	0.337 ± 0.090	0.261 ± 0.126
F-score ± SD	29.29 ± 22.44	27.87 ± 18.17	52.56 ± 26.52	160.95 ± 125.95
$R^2_{ex} \pm$ SD	0.619 ± 0.175	0.570 ± 0.172	0.695 ± 0.149	0.730 ± 0.145
$\Delta_{av} \pm$ SD	0.544 ± 0.130	0.593 ± 0.119	0.510 ± 0.130	0.478 ± 0.138
SDEP ± SD	0.709 ± 0.170	0.753 ± 0.153	0.618 ± 0.143	0.586 ± 0.153
Pr-$R^2 \pm$ SD	0.543 ± 0.261	0.490 ± 0.251	0.662 ± 0.166	0.694 ± 0.162

[a] *Mix* set run with the maximum number of components set to 5. [b] *Mix* set run with the maximum number of components set to 20.

Table 5. Statistical Descriptors of the Models Generated from the Standard Steroid Set[a]

	Fix	*Flex*	*Mix C5*	*Mix C20*
S_{PRESS}	0.627	0.740	0.686	0.680
Q^2	0.758	0.682	0.726	0.815
Opt #	3	4	4	9
R^2	0.894	0.907	0.902	0.988
SE	0.414	0.399	0.410	0.173
F-score	47.86	39.15	36.91	100.7
R^2_{ex}	0.141 (0.561)	0.155 (0.710)	0.072 (0.068)	0.425 (0.180)
Δ_{av}	0.765 (0.579)	0.599 (0.412)	0.643 (0.594)	0.411 (0.415)
SDEP	1.009 (0.687)	0.863 (0.502)	0.712 (0.712)	0.481 (0.492)
Pr-R^2	−0.103 (0.534)	0.193 (0.751)	0.451 (0.573)	0.749 (0.761)

[a] Values in parentheses represent predictions without compound M31.

components set to 20 a Q^2 (±SD) value of −0.230 (±0.226) is observed. When the limit of maximum components is raised to 20, for *Fix* and *Flex* sets, no new models with a high number of optimum components are discovered.

Standard Benchmark Test. The statistical descriptors of the standard steroid set in which the teaching set consists of the molecules **1**−**21** and the prediction set of molecules **22**−**31** are listed in Table 5, and the predictive results are shown in Table 6 and Figure 6. Unfortunately the standard partition of the steroid data happens to be one of the sets in which the *Mix* model generates a very high number of components. The optimum number of components is nine if the maximum number of components is 20.

It is noteworthy that the compounds **M27** and **M31** are systematically predicted to have too high an activity. Only the *Mix C20* model predicts the activities of these molecules with a reasonable accuracy. The reason for the anomalous activities of these molecules can be attributed to their structure. Molecule **M31** has a fluorine in the 9α-position, and it is the only molecule in the steroid set that is *endo*-substituted. Because the prediction results of the compound **M31** were systematically poor, it was excluded from the prediction set and a new set of prediction runs was performed. The results of these runs are shown in parentheses in Table 5. In the case of molecule **M27** there are several adjacent hydrogen bond-forming groups which bind to each other in a way that the model cannot fully imitate, thus creating a marked error in the predicted values. In fact, Kubinyi[60] has emphasized that the molecules in the standard training set do not cover all structural features found in the test set. In general the *Mix C20* model tends to predict too low activities for the very active molecules, and the *Fix* and *Flex* sets tend to predict too high activities for the same molecules.

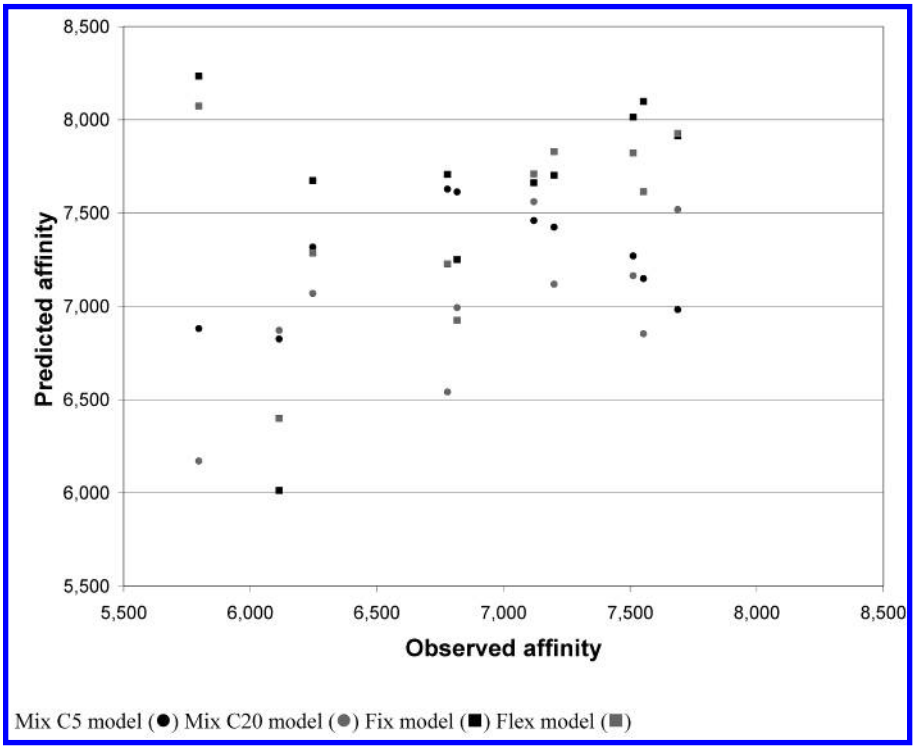

Figure 6. Observed versus predicted CBG affinity of the 10 steroids.

Table 6. Prediction Results for the Standard Steroid Set

compound	observed	*Fix* (Δ)	*Flex* (Δ)	*Mix C5* (Δ)	*Mix C20* (Δ)
H22	7.512	8.015 (0.503)	7.822 (0.310)	7.270 (−0.242)	7.164 (−0.348)
H23	7.553	8.098 (0.545)	7.614 (0.061)	7.149 (−0.404)	6.853 (−0.700)
M24	6.779	7.707 (0.928)	7.227 (0.448)	7.628 (0.849)	6.541 (0.238)
H25	7.200	7.702 (0.502)	7.829 (0.629)	7.424 (0.224)	7.119 (0.081)
M26	6.114	6.013 (−0.101)	6.399 (0.285)	6.825 (0.711)	6.871 (0.757)
M27	6.247	7.674 (0.162)	7.285 (−0.227)	7.318 (−0.194)	7.069 (−0.443)
H28	7.120	7.663 (0.543)	7.710 (0.590)	7.460 (0.340)	7.561 (0.441)
M29	6.817	7.251 (0.434)	6.926 (0.109)	7.614 (0.797)	6.993 (0.176)
H30	7.688	7.914 (0.226)	7.927 (0.239)	6.983 (−0.705)	7.519 (−0.169)
M31	5.797	8.234 (2.437)	8.073 (2.276)	6.881 (1.084)	6.171 (0.374)

Table 7. Optimal Statistical Descriptors of the Models Generated from *HALO*, *MCF*, *PCDD*, and *PCDF* Sets

set	VdW_ Radius-Intensity	VdW_ dispersion	EEL_ dispersion	Q^2	SDEP	Opt #
HALO Fix	0.850	0.500	0.120	0.659	18.703	13
HALO Flex	0.750	0.500	0. 500	0.717	17.643	15
HALO Mix	0.800	0.500	0.050	0.643	18.830	12
MCF log K_a Fix	0.800	0.500	0.050	0.339	0.979	4
MCF log K_a Flex	0.850	0.250	0.050	0.544	0.824	5
MCF log K_a Mix	0.850	0.250	0.050	0.521	0.845	5
MCF pEC_{50} Fix	0.800	0.500	0.050	0.431	1.094	4
MCF pEC_{50} Flex	0.800	0.500	0.050	0.469	1.057	4
MCF pEC_{50} Mix	0.850	0.500	0.050	0.458	1.067	4
PCDD Fix	0.025	0.050	0.500	0.688	0.883	4
PCDDFlex	0.850	0.250	0.500	0.728	1.004	7
PCDDMix	0.850	0.250	0.500	0.728	1.004	7
PCDF Fix	0.075	0.500	0.500	0.727	0.871	7
PCDFFlex	0.850	0.050	0.500	0.752	1.104	7
PCDFMix	0.850	0.050	0.500	0.752	1.104	7

Table 8. Average Statistical Descriptors of the Models Generated from 2500 Random *HALO* Teaching and Prediction Sets

	Fix	*Flex*	*Mix*
$S_{PRESS} \pm$ SD	25.435 ± 2.992	22.978 ± 3.511	23.526 ± 2.947
$Q^2 \pm$ SD	0.433 ± 0.145	0.495 ± 0.177	0.482 ± 0.166
Opt # ± SD	9.572 ± 1.032	9.349 ± 1.087	9.464 ± 1.027
$R^2 \pm$ SD	0.942 ± 0.086	0.923 ± 0.084	0.862 ± 0.121
SE ± SD	7.074 ± 2.960	8.064 ± 3.140	11.041 ± 3.861
F-score ± SD	49.19 ± 32.02	41.14 ± 25.68	19.11 ± 12.43
$R^2_{ex} \pm$ SD	0.529 ± 0.256	0.571 ± 0.235	0.528 ± 0.231
$\Delta_{av} \pm$ SD	15.635 ± 5.178	15.924 ± 5.075	16.580 ± 5.822
SDEP ± SD	20.006 ± 6.941	19.585 ± 7.042	20.753 ± 8.495
Pr-$R^2 \pm$ SD	0.279 ± 0.505	0.285 ± 0.655	0.222 ± 0.819

The statistical descriptors of *Mix C5* and *Fix* models are very similar, but the predictive results are highly dissimilar. The results given by the *Flex* model fall between the results of *Mix C5* and *Fix* models. The relatively poor performance of *Fix* and *Flex* models can partially be explained by the very poor prediction of compound **M31**'s activity. The best overall prediction results are achieved by the *Mix C20* model. However this model uses a high number of components and is therefore not optimal for predictive use. If the compound **M31** is excluded from the prediction set, the statistical descriptors of *Mix C20* and *Flex* models become similar, and both models give good predictive results. The very good performance of the *Flex* model, when compared to the *Fix* model, suggests that the conformational adaptation can significantly aid the construction of a QSAR model. The descriptors of *Mix C5* and *Fix* models are also similar, but the predictive results of these models are inferior to the results of *Mix C20* and *Flex* models. In Table 11 the predictive results of the FLUFF-BALL algorithm are compared to 13 other widely used QSAR methods. While the FLUFF-BALL does not yield the best overall result, its performance is nevertheless comparable to those of most previous QSAR methods.

Additional Validation Sets. Further validation of the FLUFF-BALL's performance was carried out utilizing five additional data sets. The *HALO* set[61] contains 44 halogenated estradiol derivatives whose affinity for the estrogen receptor is measured using receptor binding assay. The *MCF* set[62] contains 42 estradiol-17β analogues for which the K_a values of the receptor−ligand complex and the MCF-7 cell growth response EC_{50} data are available. Both biological activities reported for the *MCF* set were processed as in the original article, thus generating two separate data sets, the *MCF log K_a* and the *MCF pEC_{50}*. To gain comparable results molecule **1** (estratriene) was excluded from the MCF data as was also done in the original article. The *PCDD* and *PCDF* sets[63,64] respectively contain 25 halogenated dibenzo-*p*-dioxin congeners and 34 chlorinated dibenzofuran congeners for which the receptor binding data for cytosolic aromatic hydrocarbon (*Ah*) receptor is known.

All molecules were AM1 optimized as was done in the case of the benchmark steroid set. For *HALO* and *MCF* data sets estradiol was used as a template, and the maximum number of components was set to 15 in the case of *HALO* sets and to 11 when computing the *MCF* data. For *PCDD* and *PCDF* sets the base compounds dibenzo-*p*-dioxin and

Table 9. Average Statistical Descriptors of the Models Generated from 2500 Random *MCF log K_a* and *PEC_{50}* Teaching and Prediction Sets

	log K_a Fix	*log K_a Flex*	*log K_a Mix*	*pEC_{50} Fix*	*pEC_{50} Flex*	*pEC_{50} Mix*
$S_{PRESS} \pm$ SD	1.103 ± 0.101	1.030 ± 0.109	1.066 ± 0.111	1.183± 0. 088	1.172 ± 0.102	1.890 ± 0.100
$Q^2 \pm$ SD	0.208 ± 0.141	0.336 ± 0.147	0.295 ± 0.155	0.375 ± 0.127	0.429 ± 0.109	0.416 ± 0.109
Opt # ± SD	3.481 ± 1.228	4.702 ± 1.246	4.831 ± 1.478	4.011 ± 1.089	5.790 ± 1.720	5.928 ± 1.673
$R^2 \pm$ SD	0.539 ± 0.174	0.664 ± 0.121	0.643 ± 0.149	0.661 ± 0.114	0.778 ± 0.121	0.779 ± 0.121
SE ± SD	0.822 ± 0.160	0.723 ± 0.125	0.742 ± 0.143	0.860 ± 0.129	0.703 ± 0.141	0.703 ± 0.133
F-score ± SD	9.88 ± 7.10	11.18 ± 7.56	10.43 ± 9.50	16.17 ± 6.31	20.47 ± 7.33	19.74 ± 6.57
$R^2_{ex} \pm$ SD	0.235 ± 0.156	0.348 ± 0.181	0.303 ± 0.177	0.398 ± 0.264	0.471 ± 0.249	0.458 ± 0.252
$\Delta_{av} \pm$ SD	0.899 ± 0.156	0.805 ± 0.167	0.835 ± 0.172	0.925 ± 0.256	0.878 ± 0.255	0.892 ± 0.255
SDEP ± SD	1.087 ± 0.175	0.995 ± 0.190	1.032 ± 0.192	1.175 ± 0.297	1.107 ± 0.298	1.118 ± 0.295
Pr-$R^2 \pm$ SD	0.120 ± 0.270	0.251 ± 0.325	0.199 ± 0.294	0.228 ± 0.506	0.264 ± 0.589	0.231 ± 0.590

Table 10. Average Statistical Descriptors of the Models Generated from 2500 Random *PCDD* and *PCDF* Teaching and Prediction Sets

	PCDD Fix	*PCDD Flex*	*PCDD Mix*	*PCDF Fix*	*PCDF Flex*	*PCDD Mix*
$S_{PRESS} \pm$ SD	1.009 ± 0.150	1.137 ± 0.196	1.137 ± 0.196	0.898 ± 0.101	0.926 ± 0.120	0.934 ± 0.119
$Q^2 \pm$ SD	0.648 ± 0.078	0.643 ± 0.079	0.643 ± 0.079	0.700 ± 0.070	0.657 ± 0.086	0.657 ± 0.085
Opt # ± SD	4.496 ± 0.533	4.938 ± 0.246	4.941 ± 0.241	4.595 ± 0.739	4.667 ± 0.567	4.663 ± 0.570
$R^2 \pm$ SD	0.901 ± 0.049	0.925 ± 0.038	0.926 ± 0.038	0.865 ± 0.037	0.883 ± 0.058	0.884 ± 0.060
SE ± SD	0.507 ± 0.121	0.476 ± 0.131	0.472 ± 0.136	0.598 ± 0.072	0.507 ± 0.138	0.504 ± 0.143
F-score ± SD	34.89 ± 48.89	59.87 ± 304.09	78.09 ± 407.32	20.99 ± 12.88	41.17 ± 38.25	42.03 ± 39.42
$R^2_{ex} \pm$ SD	0.687 ± 0.248	0.628 ± 0.254	0.619 ± 0.262	0.683 ± 0.223	0.637 ± 0.223	0.623 ± 0.232
$\Delta_{av} \pm$ SD	0.849 ± 0.389	0.814 ± 0.250	0.822 ± 0.265	0.655 ± 0.168	0.711 ± 0.205	0.718 ± 0.211
SDEP ± SD	1.052 ± 0.567	0.966 ± 0.281	0.973 ± 0.295	0.787 ± 0.198	0.865 ± 0.232	0.875 ± 0.238
Pr-$R^2 \pm$ SD	0.113 ± 2.446	0.458 ± 0.399	0.438 ± 0.483	0.231 ± 0.307	0.549 ± 0.289	0.522 ± 0.356

Table 11. Comparison of FLUFF-BALL with Other QSAR Techniques for Test Set 22−31[b]

method	R^2_{ex}	Δ_{av}	SDEP	Pr-R^2
COMPASS[67]	0.16 (0.69)	0.46 (0.29)	0.70 (0.34)	0.46 (0.89)
MS-WHIM[30]	0.28 (0.63)	0.44 (0.30)	0.66 (0.41)	0.52 (0.83)
PARM[68]	0.33 (0.30)	0.52 (0.56)	0.71 (0.74)	0.45 (0.45)
TQSAR[32]	0.16 (0.36)	0.59 (0.46)	0.76 (0.56)	0.37 (0.69)
SOMFA[44]	0.20 (0.62)	0.43 (0.32)	0.58 (0.36)	0.63 (0.87)
EVA[35−38]	0.36 (0.34)	0.42 (0.39)	0.53 (0.51)	0.69 (0.74)
CoMFA[3]	0.25 (0.75)	0.46 (0.30)	0.71 (0.40)	0.45 (0.84)
GRIND[8]	−(0.88)[a]	−(0.23)[a]	−(0.26)[a]	−(0.93)[a]
MFTA[69]	0.87 (0.82)	0.21 (0.23)	0.30 (0.31)	0.90 (0.90)
COMSA[58]	0.09 (0.41)	0.52 (0.38)	0.70 (0.44)	0.47 (0.81)
MEDV[70]	0.45 (0.57)	0.54 (0.48)	0.65 (0.59)	0.54 (0.66)
QS-SM[9]	0.36 (0.22)	0.47 (0.42)	0.54 (0.49)	0.68 (0.76)
EEVA[39]	0.36 (0.58)	0.41 (0.30)	0.58 (0.40)	0.64 (0.85)
FLUFF-BALL (*Fix*)	0.14 (0.56)	0.77 (0.58)	1.01 (0.69)	−0.10 (0.53)
FLUFF-BALL (*Flex*)	0.16 (0.71)	0.60 (0.41)	0.86 (0.50)	0.19 (0.75)
FLUFF-BALL (*Mix C5*)	0.07 (0.07)	0.64 (0.59)	0.71 (0.71)	0.45 (0.57)
FLUFF-BALL (*Mix C20*)	0.43 (0.18)	0.41 (0.42)	0.48 (0.49)	0.75 (0.76)

[a] Compound M31 is reported as an outlier and is excluded. [b] Values in parentheses represent models without compound M31.

dibenzofuran were used as templates and the maximum number of components was set to 7. All five sets were then subjected to the same extensive test validation procedure as the benchmark steroid set, and the optimal statistical descriptors obtained are presented in Table 7. The Q^2 values of the *HALO* set (0.643−0.717) are fully comparable to the ones obtained in the original article (0.566−0.767) using SYBYL field fitting and CoMFA. The *MCF log* K_a models also yielded Q^2 values (0.339−0.544) which are comparable to the CoMFA models obtained using RMS fit of the steroid backbone (0.395−0.583) and also to the results of SEAL fit (0.426−0.597). The *MCF* pEC_{50} set produced slightly lower values (0.431−0.469), and the difference for the CoMFA models obtained using RMS fit of the steroid backbone (0.463−0.624) or the SEAL fit (0.424−0.582) was considerable. In general one should note that the *HALO* set generated a high number of components, and when the maximum number of allowed components was raised several models generated up to 20 components. However these data were disregarded because such a high number of components indicate a considerable over-fitting. The *PCDD* set yielded Q^2 values (0.688−0.728) that were slightly lower but still comparable to the values reported in the literature[27,63−66] (0.715−0.862). The *PCDF* set generated slightly better Q^2 results (0.727−0.752) which are also closer to the values reported in the literature[63−66] (0.742−0.795). In general one should note that the *Flex* models generated the best Q^2 values closely followed by the *Mix* models, and the *Fix* models generated the worst models. In the case of *PCDD* and *PCDF* the *Flex* and *Mix* resulted in almost identical superpositions, but in the *Fix* set the halogen substituents were slightly offset because of the small changes in the optimal backbone conformation.

As the original articles did not provide separate training and test sets, a division of 29 molecules for the training set and 15 for the test set was used when randomly generating the teaching and test sets for *HALO* data. For both *MCF* sets the similar division was 28 and 14 molecules, respectively, and for PCDD and *PCDF* 5 compounds were separated for the test set leaving 21 and 29 compounds, respectively, for the teaching set. The maximum number of components was set at 10 for both *HALO* and *MCF* data and at 5 for *PCDD* and *PCDF*. The results of the 2500 runs are shown in Tables 8−10. The *HALO* and *MCF* sets, which were known to be computationally difficult, generated a wide spectrum of models as is indicated by the high standard deviations of the statistical descriptors. However, all models were clearly predictive as indicated by positive Pr-R^2 values. Especially in the case of the *HALO* the instabilities observed most likely stem from the uneven distribution of the observed values. The *Flex* and *Mix* cases of the *PCDD* and *PCDF* sets generated models with high average Q^2 values and high predictivity as indicated by Pr-R^2 values of 0.438−0.549.

DISCUSSION

The complete FLUFF-BALL analysis is performed in the following four steps: (1) The molecular models are generated and the template and ligands are assigned. (2) The initial superposition is performed. (3) The molecules are superimposed using FLUFF force-field and a suitable computational technique such as geometry optimization. (4) The Ball descriptors are evaluated and the PLS model is built.

Of the steps above only step 2 needs human intervention as the molecular structures can be imported from a molecular database and the computations can be performed automatically. The human intervention is needed in step 2 as the FLUFF superposition technique is capable of correctly superimposing the molecules of the test sets when given only a rough initial superposition. For all of the cases the FLUFF algorithm needed a proper alignment of the backbones, as the simple geometry optimization used is incapable of finding the global minimum, because it cannot flip the whole molecule to a new orientation once it has found a local minimum. The initial geometry is not very important as flexible molecules soon find the MMFF94 optimum geometry. Naturally the initial conformation dictates the local minimum which the molecule adopts when optimized, but in case of the test sets the AM1 optimized and manually drawn structures adopted the same conformation.

For the superposition results, it must be noted that *Mix* and *Flex* sets of the *PCDD* and *PCDF* generated almost identical results with very precise alignment of the molecules, whereas in the *Fix* sets there was still some difference in the orientation of the substituents. The BALL results of the *PCDD* and *PCDF* sets clearly reflect this as the *Fix* sets

generated slightly lower Q^2 values and the optimum components differed from the optimum values of the *Flex* and *Mix*. For the steroid sets one must note that the steroid backbone is rather rigid, and its conformational changes were small. Therefore, the main task was the finding of correct alignment for the side chains. In the *Fix* sets the only criteria for the alignment was the root-mean-square-deviation of the backbone carbon atoms. This resulted in good alignment of the backbones and considerable differences in the side chain conformations. In the *Mix* sets there were some differences in the optimum backbone conformation between the MMFF94 optimized ligand and the AM1 optimized template. This effect was especially clear in the A- and D-rings of the steroid backbone. However, the side chains were generally well aligned. For the *Flex* sets it was clear that the steroids were assuming a common optimum geometry, and the quality of the superposition was more balanced as there were only small errors in both backbone and in side chain alignments. As the FLUFF is a force field -based technique, the superposition can be performed by geometry optimization, as was done in this work, but any other method for finding minimum energy can also be used.

When all test sets were run through the BALL algorithm, a set of highly predictive QSAR descriptors was obtained. The BALL algorithm uses a local coordinate system which enables the use of fully flexible superposition techniques, as it allows some changes in the template's conformation without adverse effects on the predictive power of the QSAR descriptor. When the parameter spreads were evaluated it became evident that the main parameter affecting the predictive power is the vdW_radiusIntensity, while vdW_dispersion had a lesser effect. The impact of the EEL_dispersion was increased if the structure contained highly charged atoms such as halogens, but in many cases the charge dispersion did not affect the Q^2 value of the model.

The optimum parameters for the *Fix* and *Flex* sets of the standard benchmark were similar, while the *Mix* set was markedly different from the other two sets (Table 4). All FLUFF-BALL models produced a valid model with high Q^2 values (0.626−0.801). The *Mix* set proved to be the most problematic of the three sets. When the 2500 randomly selected teaching and test sets were run through, the *Mix* set occasionally generated a model with a high number of PCs. Unfortunately the standard partition of the molecules was one of those cases. This behavior may be related to the fact that in the *Mix* set, two separate minima are competing with each other. The template is in AM1 optimized conformation, and the FLUFF algorithm tries to force a ligand with MMFF94 optimum on top of it. However, on average the *Mix* set generated the best models especially if the number of components was allowed to rise above 5 (Table 5). The predictions of standard benchmark set by FLUFF-BALL can be divided into two categories. First there are *Fix* and *Mix C5* which have a poor predictive power and to the *Flex* and *Mix C20* which have a considerably higher predictive power. As the *Mix C20* is discarded because of its high number of PCs, it becomes evident that the fully flexible superposition produced the highest predictive power. In the *HALO* and *MCF* sets and the *Flex* and *Mix* cases of the *PCDD* and *PCDF* the optimal parameters are located in a small area around VdW_RadiusIntensity of 0.800. The VdW_dispersion and EEL_dispersion parameters have a higher variance, but overall they have a lesser impact on the Q^2 value. All in all the BALL parameter optimization can be focused on the area of VdW_RadiusIntensity 0.700−0.900, and if the two dispersion parameters are included in the optimization this means that 125 unique sets are to be evaluated. With a powerful desktop PC this optimization can be performed in less than 20 min. Furthermore this optimization can be performed without human intervention so it only demands computer time. But if the optimization is to be omitted, a BALL parameter set of VdW_RadiusIntensity 0.800, VdW_dispersion 0.500, and EEL_dispersion 0.500 should provide a reasonable Q^2 value. However one should note that this optimum is found using high connectivity structures and may not be universally applicable.

When the FLUFF-BALL's results for the standard benchmark are compared with the results of 13 other QSAR techniques (Table 11), one can observe that the predictive power of FLUFF-BALL is near to the average of these techniques. When comparing the results of FLUFF-BALL to the results of the other methods, one must keep in mind that the only human intervention needed to obtain the FLUFF-BALL prediction is a rough initial superposition of the template and ligand. All other stages of prediction can be automated, although a quick visual inspection of the superposition results is advisable in order to detect possible misalignments.

CONCLUSIONS

The results of this work show that FLUFF-BALL is capable of generating robust predicting models for several different data sets and biological activities. Because the FLUFF-BALL technique can easily be automated, and it is computationally simple, it makes a useful and quite fast "molecular sieve". Despite this design emphasis the FLUFF-BALL produced results comparable to the other techniques listed in Table 11. Also one must bear in mind that the majority of them require an extensive amount of user involvement to produce the results and are therefore less suitable for fast screening applications.

ACKNOWLEDGMENT

Mikael Peräkylä would like to the Academy of Finland for its support (Grant #48577). Samuli-Petrus Korhonen would like to thank The National Graduate School of Informational and Structural Biology (ISB) for financial support.

REFERENCES AND NOTES

(1) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616−1626.

(2) Hansch, C.; Leo, A. *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(3) 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.

(4) Goodford, P. J. A Computational Procedure for Determining Energetically Favourable Binding Sites on Biologically Important Molecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(5) Jackson, J. E. *A User's Guide to Principal Components*; Wiley & Sons: 1991.

(6) Rännar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS Kernel Algorithm for Data Set With Many Variables and Fewer Objects. Part 1: Theory and algorithm. *J. Chemom.* **1994**, *8*, 111−125.
(7) Höskuldsson, A. PLS Regression methods. *J. Chemom.* **1988**, *2*, 211−228.
(8) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-Independent Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.
(9) Amat, L.; Besalu, E.; Carbo-Dorca, R. Identification of Active Molecular Sites Using Quantum-Self-Similarity Measures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 978−991.
(10) Bohl, M. MTD Approach to Quantitative Structure−Activity Relatioship for Cardiotonic Steroids. *Z. Naturforsch.* **1985**, *40*, 858−862.
(11) Allen, M. S.; LaLoggia, A. J.; Dorn, L. J.; Martin, M. J.; Costantino, G.; Hagen, T. J.; Koehler, K. F.; Skolnick, P.; Cook, J. M. Predictive Binding of Beta-Carboline Inverse Agonists and Antagonists via the CoMFA/GOLPE Approach. *J. Med. Chem.* **1992**, *35*, 4001−4010.
(12) Miller, M. D.; Sheridan, R. P.; Kearsley, S. K. SQ: A Program for Rapidly Producing Pharmacophorically Relevent Molecular Superpositions. *J. Med. Chem.* **1999**, *42*, 1505−1514.
(13) Mcmartin, C.; Bohacek, R. S. QXP: Powerful, Rapid Computer Algorithms for Structure-Based Drug Design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.
(14) Masek, B. B.; Merchant, A.; Matthew, J. B. Molecular Shape Comparison of Angiotensin I1 Receptor Antagonists. *J. Med. Chem.* **1993**, *36*, 1230−1238.
(15) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: A Molecular-Field Matching Program Exploting Applicability of Molecular Similarity Approaches. *J. Comput. Chem.* **1997**, *18*, 934−954.
(16) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.
(17) Grant, J. A.; Gallardo, M. A.; Pickup, B. T. A Fast Method of Molecular Shape Comparison: A Simple Application of a Gaussian Description of Molecular Shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.
(18) Nissink, J. W. M.; Verdonk, M. L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of molecules: Electron Density Fitting by Application of Fourier Transforms. *J. Comput. Chem.* **1997**, *18*, 638−645.
(19) Constans, P.; Amat, L.; Carbo-Dorca, R. Toward a Global Maximization of the Molecular Similarity Function: Superposition of Two Molecules. *J. Comput. Chem.* **1997**, *18*, 826−846.
(20) Parretti, M. F.; Kroemer, R. T.; Rothman, J. H.; Richards, W. G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem.* **1997**, *18*, 1344−1353.
(21) Mills, J. E. J.; Perkins, T. D. J.; Dean, P. M. An Automated Method for Predicting the Positions of Hydrogen-Bonding Atoms in Binding Sites. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 229−242.
(22) Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties. A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−205.
(23) Kubinyi, H.; Folkers, G.; Martin, Y. C. *3D QSAR in Drug Design, Volume 2: Ligand Protein Interactions and Molecular Similarity*; ESCOM: 1998.
(24) Chae, Chong Hak; Oh, Dong Gweon; Shin, Wanchul Flexible Molecular Superposition: Development of a Combined Similarity Index and Application of the Constrained Optimization Technique. *J. Comput. Chem.* **2001**, *22*, 888−900.
(25) Melani, F.; Gratteri, P.; Adamo, M.; Bonaccini, C. Field Interaction and Geometrical Overlap: A New Simplex and Experimental Design Based Computational Procedure for Superposing Small Ligand Molecules. *J. Med. Chem.* **2003**, *46*, 1359−1371.
(26) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. Autocorrelation Descriptor. *Eur. J. Med. Chem.* **1984**, *19*, 66−70.
(27) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular Surface Properties for Modeling Corticosteroid Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(28) Todeschini, R.; Vighi, M.; Provenzani, R.; Finizio, A.; Gramatica, P. Modeling and Prediction by Using WHIM Descriptors in QSAR Studies: Toxicity of Heterogeneous Chemicals on Daphnia Magna. *Chemosphere* **1996**, *32*, 1527−1545.
(29) Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E. Weighted Holistic Invariant Molecular Descriptors. Part 2. Theory Development and Applications on Modelling Physicochemical Properties of Polyaromatic Hydrocarbons. *Chemom. Intell. Lab. Syst.* **1995**, *27*, 221−229.
(30) Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zaliani, A. MS-WHIM, New 3D Theoretical Descriptors Derived from Molecular Surface Properties: A Comparative 3D QSAR Study in a Series of Steroids. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 79−92.
(31) Silverman, D. B. The Thirty-one Benchmark Steroid Revisited: Comparative Molecular Moment Analysis (CoMMA) with Principal Component Regression. *QSAR* **2000**, *19*, 237−246.
(32) Robert, D.; Amat, L.; Carbo-Dorca, R. Three-Dimensional Quantitative Structure−Activity Relationship from Tuned Molecular Quantum Similarity Measures: Prediction of the Corticosteroid-Binding Globulin Binding Affinity fro a Steroid Family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(33) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev.* **1996**, *96*, 1027−1044.
(34) Bursi, R.; Dao, T.; van Wijk, T.; de Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative Spectra Analysis (CoSA): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−868.
(35) Turner, D. B.; Willett, P. The EVA Spectral Descriptor. *Eur. J. Med. Chem.* **2000**, *35*, 365−375.
(36) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of Novel Infrared Range Vibration-Based Descriptor (EVA) for QSAR Studies: General Application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409−422.
(37) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of Novel Infrared Range Vibration-Based Descriptor (EVA) for QSAR Studies: 2. Model Validation Using a Benchmark Steroid Dataset. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 271−296.
(38) Turner, D. B.; Willett, P. Evaluation of the EVA descriptor for QSAR studies: 3. The Use of a Genetic Algorithm to Search for Models with Enhanced Predictive Properties (EVA_GA). *J. Comput.-Aided Mol. Des.* **2000**, *14*, 1−21.
(39) Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Peräkylä, M. Evaluation of Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 607−613.
(40) Maw, H. H.; Hall, L. H. E-State Modelling of Corticosteroids Binding Affinity Validation of Model for Small Data Set. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1248−1254.
(41) Kellogg, G. E.; Kier, L. B.; Gillard, P.; Hall, L. H. E-State fields: Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.
(42) Cho, S. J.; Tropsha, A. Cross-Validated R2-guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.
(43) 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.
(44) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure−Activity Studies. *J. Med. Chem.* **1999**, *42*, 573−583.
(45) Hopfinger, A. J.; Wang, S.; Tokarski, J. S.; Baiqiang, J.; 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.
(46) Vedani, A. 5D-QSAR: The Key for Simulating Induced Fit. *J. Med. Chem.* **2002**, *45*, 2139−2149.
(47) Carbo, R.; Leyda, L.; Arnau, M. How Similar is a Molecule to Another? An Electron Density Measure of Similarity Between Two Molecular Structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(48) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian Function for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.
(49) Good, A.; So, S.-S.; Richards, G. Structure−Activity Relationships from Molecular Similarity Matrices. *J. Med. Chem.* **1993**, *36*, 433−438.
(50) Good, A. C.; Peterson, S. J.; Richards, W. G. QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.
(51) Benigni, R.; Cotta-Ramusino, M.; Giorgi, F.; Gallo, G. Molecular Similarity Matrixes and Quantitative Structure−Activity Relationship: A Case Study with Methodological Implications. *J. Med. Chem.* **1995**, *38*, 629−635.
(52) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(53) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17*, 520−552.

(54) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553−586.
(55) Halgren, T. A. Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94. *J. Comput. Chem.* **1996**, *17*, 587−615.
(56) Halgren, T. A. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. *J. Comput. Chem.* **1996**, *17*, 616−641.
(57) Kearsley, S. K.; Smith, G. M. An Alternate Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Methodol.* **1990**, *3*, 615−633.
(58) Polanski, J.; Walczak, B. The Comparative Molecular Surface Analysis (COMSA): a Novel Tool for Molecular Design. *Comput. Chem.* **2000**, *24*, 615−625.
(59) Coats, E. A. The CoMFA steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discovery Des.* **1998**, *12*−*14*, 199−213.
(60) van de Waterbeemd, H.; Testa, B.; Folkers, G. Computer-Assisted Lead Finding and Optimization: Current Tools for Medicinal Chemistry. In *A General View on Similarity and QSAR Studies*; Kubinyi, H., Ed.; VHCA: Basel, Switzerland, 1997; pp 7−28.
(61) Gantchev, T. G.; Ali, H.; van Lier, J. E. Quantitative Structure−Activity Relationship/Comparative Molecular Field Analysis (QSAR/CoMFA) for Receptor-Binding Properties of Halogenated Estratriol Derivatives. *J. Med. Chem.* **1994**, *37*, 4164−4176.
(62) Wiese, T. E.; Polin, L. A.; Palomino, E.; Brooks, S. C. Induction of the Estrogen Specific Mitogenic Response of MCF-7 Cells by Selected Analogues of Estradiol-17β: A 3D QSAR Study. *J. Med. Chem.* **1997**, *40*, 3659−3669.
(63) Tuppurainen, K.; Ruuskanen, J. Electronic Eigenvalue (EEVA): a New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies. A QSAR Approach to Ah Receptor Binding Affinity of Polychlorinated Biphenyls (PCBs), Dibenzo-p-Dioxins (PCDDs) and Dibenzofurans (PCDFs). *Chemosphere* **2000**, *41*, 843−848.
(64) So, S.-S.; Karplus, M. Three-Dimensional Quantitative Structure−Activity Relationship from Molecular Similarity Matrices and Neural Networks. 2. Applications. *J. Med. Chem.* **1997**, *40*, 4360−4371.
(65) Waller, C. L.; McKinney, J. D. Comparative Molecular Field Analysis of Polyhalogenated Dibenzo-*p*-Dioxins, Dibenzofurans, and Biphenyls. *J. Med. Chem.* **1992**, *35*, 3660−3666.
(66) Poso, A.; Tuppurainen, K.; Ruuskanen, J.; Gynther, J. Binding of Some Dioxins and Dibenzofurans to the Ah Receptor. A QSAR Model Based on Comparative Molecular Field Analysis (CoMFA). *J. Mol. Struct.: THEOCHEM* **1993**, *282*, 259−264.
(67) Jain, A. N.; Koile, K.; Chapman, D. Compass: Predicting Biological Activities from Molecular Surface Properties. *J. Med. Chem.* **1994**, *37*, 2315−2327.
(68) Chen, H.; Zhou, J.; Xie, G. PARM: A Genetic Evolved Algorithm to Predict Bioactivity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243−250.
(69) Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S. Molecular Field Topology Analysis Method in QSAR Studies of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 659−667.
(70) Liu, S.-S.; Yin, C.-S.; Li, Z.-L.; Cai, S.-X. QSAR Study of Steroid Benchmark and Dipeptides Based on MEDV-13. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 321−329.

CI034027O