Conformational Coverage by a Genetic Algorithm: Saturation of Conformational Space[†]

Todor Pavlov,[‡] Milen Todorov,[‡] Galina Stoyanova,[‡] Patricia Schmieder,[§] Hristo Aladjov,[§]
Rossitsa Serafimova,[‡] and Ovanes Mekenyan*[,‡]

Laboratory of Mathematical Chemistry, University “Prof. As. Zlatarov”, 8010 Bourgas, Bulgaria, and USEPA,
ORD, NHEERL, Mid-Continent Ecology Division, 6201 Congdon Boulevard, Duluth, Minnesota 55804

Received January 15, 2007

The molecular modeling is traditionally based on analysis of minimum energy conformers. Such simplifying assumptions could doom to failure the modeling studies given the significant variation of the geometric and electronic characteristics across the multitude of energetically reasonable conformers representing the molecules. Moreover, it has been found that the lowest energy conformers of chemicals are not necessarily the active ones with respect to various endpoints. Hence, the selection of active conformers appears to be as important as the selection of molecular descriptors in the modeling process. In this respect, we have developed effective tools for conformational analysis based on a genetic algorithm (GA), published in *J. Chem. Inf. Comput. Sci.* (**1994**, *34*, 234; **1999**, *39* (6), 997) and *J. Chem. Inf. Model*. (**2005**, *45* (2), 283). This paper presents a further improvement of the evolutionary algorithm for conformer generation minimizing the sensitivity of conformer distributions from the effect of smoothing parameter and improving the reproducibility of conformer distributions given the nondeterministic character of the genetic algorithm (GA). The ultimate goal of the saturation is to represent the conformational space of chemicals with an optimal number of conformers providing a stable conformational distribution which cannot be further perturbed by the addition of new conformers. The generation of stable conformational distributions of chemicals by a limited number of conformers will improve the adequacy of the subsequent molecular modeling analysis. The impact of the saturation procedure on conformer distributions in a specific structural space is illustrated by selected examples. The effect of the procedure on similarity assessment between chemicals is discussed.

1. INTRODUCTION

Typically QSARs rely on a single conformer to represent the 3D molecular structure of the chemical under study, while all other energetically stable conformers are neglected. In the best case, the representative conformer is the one with the lowest potential energy for the isolated molecule or the one observed in crystal phase. The use of the lowest energy conformer is commonly accepted in SAR studies but inappropriate, because in complex systems such as biological tissues and fluids chemicals are likely to exist in a variety of conformational states. It has been found that conformational flexibility has a significant impact on molecular steric, electronic structure, and associated properties. Conformers for a given chemical, which are within the formation enthalpy range of 20 kcal/mol (the threshold between the energies of van der Waals and H-bonding interactions) from the lowest energy conformer, exhibit significant variation in potentially relevant electronic descriptors. For example, conformers of nitrofurazone (CAS 59-87-0) had a range of 0.9 eV for E_{LUMO}, 0.8 eV for E_{HOMO}, 1.2 eV for $E_{\mathrm{HOMO-LUMO}}$, and 6 D for dipole moment. The variation in descriptor values between conformers highlights the necessity of representing the molecular parameters as a finite range of values corresponding to energetically stable conformers, instead of a single point value.[1] Hence, the generation of the set of conformers representing the chemical 3D structure appears to be as important for the QSAR studies as the selection of descriptors associated with the endpoint under investigation. Apparently, the methods for exploring the conformational space of molecules have become an important issue in relating molecular structures to biological endpoints of respective chemicals.

Conformer generation methods could be divided into two major categories: deterministic (systematic) and nondeterministic (stochastic).[2,3] The complete set of low-energy conformations are produced by deterministic approaches. A systematic-search algorithm is the one included in the SYBYL molecular modeling package[4] (see also ref 5) where all cyclic moieties are considered rigid. A more versatile system for exhaustive conformational search is the internal coordinate tree-searching procedure by Lipton and Still[6] (see also refs 7 and 8). Of special note is that the tree-searching techniques resolve the conformational flexibility of cyclic saturated moieties, as opposed to other “brute-force” torsion angle search techniques where conformational degrees of freedom are restricted to rotations around noncyclic bonds, only. We have further developed the tree-searching technique,[7] initiating from the molecular topology and generating all conformers consistent with steric and stereochemical constraints. A practical drawback of all systematic algorithms is the “combinatorial explosion” due to the exponentially growing number of combinations of the rotational variables

[†] Dedicated to Professor Nenad Trinajstić on the occasion of his 70th birthday.

* Corresponding author e-mail: omekenya@btu.bg.

[‡] University “Prof. As. Zlatarov”.

[§] USEPA.

10.1021/ci700014h CCC: $37.00
Published on Web 04/28/2007

in flexible molecules. To confine the structural space populated by the conformers methods have been developed applying various heuristics using predefined libraries of small (predominantly cyclic) structural fragments or knowledge-based rules derived from experimental data. Examples of popular heuristic schemes are Omega[9] and Catalysts.[10−12] To assemble the initial model Omega combines predefined (or produced on the fly) fragment templates across the sigma backbone. Subsequently, Omega attempts to generate all possible (energetically reasonable) combinations of template conformations. The final set is formed by conformers that differ with respect to each other in terms of a root-mean-square deviation (rmsd) threshold.

Catalysts is a conformational space sampling system integrated with the Smellie et al. algorithm[11,12] repelling too similar conformers and thus maximizing diversity among generated conformers. It has been found[13,14] that the speed and performance of the algorithm to generate bioactive (experimentally observed) conformers were improved when an internal fragment library of ring conformations is used.

In spite of their successful applications[3,14,15] the heuristic modifications of the systematic approaches have the insufficiency to restrain (sometimes significantly) the space of accessible conformers due to the predefined templates (or other heuristics) they are using. For example, the combination of rigid templates along sigma bonds will limit capabilities of the models to reproduce the flexibility of the polycyclic saturated structures, such as the steroids.

Another way to circumvent the combinatorial complexity of systematic algorithms is the stochastic approach. Chang et al.[16] describe a Monte Carlo methodology for conformational search and demonstrate its efficiency in finding low-energy conformers. Mode-following and “minimum jumping” approaches (LMOD)[17] also apply random walks on the molecular energy surface. Of course, this cannot guarantee that the energy surface will be exhaustively and evenly visited. ICM model is another well reputable representation of Monte Carlo search algorithms.[18] Stochastic approaches are also used in molecular docking methods. Thus in the Gold system, a GA is used for conformational multiplication of ligands in the receptor cavity.[19,20] The program employs an island model, in which several subpopulations of chromosomes (binary string representing conformations) are created instead of one large population. Subsequently, the genetic operations—mutations and crossover—are applied to modify and/or combine the subpopulations. The stochastic approaches are usually combined with other techniques to improve their efficiency when a set of geometric and energetic constraints needs to be imposed on generated conformers. Thus, the Monte Carlo steps designed to improve the sampling quality by random “jumping” in the conformational space is followed by a deterministic run of molecular dynamics (QXP system).[21] The distance-geometry technique (DGEOM)[22] is using a random generator combined with a prescribed set of distance constraints controlling the produced interatomic distances.

The purpose of the present work is not the comparison of different methods for conformational analysis. We wanted to focus on some drawbacks of nondeterministic algorithms associated with the QSAR modeling. They were encountered when assessing the chemical reactivity as well as similarity between chemicals with respect to their 3D structures.

The nondeterministic algorithm we have analyzed was developed recently based on a GA.[23] The algorithm (called GAS) is maximizing the diversity between generated conformers. In contrast to traditional GA, the fitness of a conformer is not quantified individually but only in conjunction with the population it belongs to. To maximize the coverage of conformational space, besides the rotation around single bonds, new conformational degrees of freedom are handled, such as flip of free corners in saturated rings[24] and reflection of pyramids on the junction of two or three saturated rings.[23] The latter two were particularly introduced to encompass structural diversity of polycyclic structures. Later on, to improve the coverage of the conformational space the fitness function based on maximization of rmsd between conformers was combined with Shannon function accounting for evenness of conformer distribution across conformational space. For more details on the improved version of the algorithm one can consult ref 25.

The 3D QSAR modeling tool we have used to analyze the impact of GA based conformer generation algorithm on reactivity assessment is the COREPA probabilistic classification scheme which is trying to identify the **CO**mmon **RE**activity **PA**ttern between biologically similar chemicals.[26] The basic assumption in COREPA is that chemicals which elicit similar biological behavior through a common mode of action should possess a commonality in their stereoelectronic properties (i.e., reactivity pattern). Elucidation of this pattern requires examination of the conformational flexibility of the chemicals in an attempt to reveal areas in the multidimensional descriptor space which are most populated by the conformers of the biologically active molecules and less populated by conformers of inactive chemicals simultaneously. The COREPA approach has demonstrated its capabilities for modeling series of endpoints associated with the interaction of chemicals with the ER receptor,[27,28] DNA,[29] and protein.[30]

The common reactivity pattern in COREPA consists of the normalized sum of conformational distributions of biologically similar chemicals across a descriptor axis. These distributions, however, were found to be very sensitive to two factors. The first one is the values of the smoothing parameter which controls the shape of the continuous approximation of discrete conformer distributions. Our attempts to define an optimal value for this parameter showed that it depends strongly on the number of conformers representing conformational space and the degree of smoothing of the continuous approximation of conformer distribution. Large values of the smoothing parameter lead to highly detailed approximation which memorizes data without any generalization of reactivity pattern. Alternatively, small values of this parameter could bring such a generalization of reactivity pattern which could cause a bias in its representation.

The second factor affecting the conformer distribution is associated with the nondeterministic character of the conformer generating algorithm. The comparison of conformer distributions of a chemical (across predefined molecular descriptors) obtained by independent GAS runs showed that they could be quite different due to both the nondeterministic character of conformer generating algorithm and effect of the smoothing parameter. In turn, these effects could cause a bias in the reactivity patterns associated with biologically similar chemicals which could produce nonadequate QSAR models. Our attempt to improve the reproducibility of conformer distributions by optimizing the values of smoothing parameter failed. Hence, we adopted an alternative approach—at a constant (statistically recommended[31]) value of smoothing parameter, to increase the number of conformers representing conformational space until there is no significant variation in conformer distributions in two subsequent runs of the evolutionary algorithm for conformer generation.

In this respect, the aim of the present work is to improve the nondeterministic procedure for conformer generation based on evolutionary algorithm by developing a new method for optimal saturation of conformational space with a limited but appropriately generated set of conformers. The new coverage of conformational space will minimize the sensitivity of conformer distributions across molecular descriptors from the effects of smoothing parameter and nondeterministic nature of GA algorithm. The ultimate goal of the saturation is to represent the conformational space of chemicals with an optimal number of conformers providing a stable conformational distribution which will not be perturbed by the addition of new conformers. In turn, the generation of stable conformational distributions of chemicals by a limited number of conformers will improve the adequacy of the subsequent QSAR modeling and assessment of similarity between chemicals.

2. METHODS

2.1. Evolutionary Algorithm for Conformer Generation. The conformer multiplication procedure in this study is based on genetic algorithm (GA).[23] This algorithm is applied in an iterative procedure for collecting the structurally most diverse conformers using the GA core.

Basic Principles. Genetic algorithms for creating new structures typically begin with a random initial population of size $N_p$which is called the permanent population. The permanent population is extended by a number, N_c, of new individuals having a different structure. Out of this extended population with $N_p + N_c$ individuals, N_p representatives are selected based on the fitness criteria to form the next generation of chemical structures. The extension of a population, followed by its selective reduction to the size of the permanent population, N_p, forms a distinct evolution step in the algorithm. Additions to the population of structures are attained by both mutations and crossovers. Mutation algorithms produce random modifications of the atoms, and crossover algorithms use features of two existing structures to form a new structure. The evolution of the structures is an iterative process repeated until some ending criteria such as convergence is seen with respect to the minimal improvement over several iterations.

Conformational Variables. Five types of structural variables or changes in molecular conformations are used to represent the important characteristic encoded as a chemical "gene", and the genes are then combined into the chemical "chromosome": rotation around single and double bonds, inversion of stereocenters, flip of free corners in saturated rings, reflections of pyramids on the junction of two or three saturated rings, and flip of fragments.[25] Some of the structural modifications reflect stereochemical features rather than conformational degrees of freedom. Such genes are totally or selectively disabled from modification depending on input.

Cardinality of the Conformational Populations. The algorithm defines automatically the cardinality of population of conformers used to cover the conformational space. The number of structures in the permanent population (or population of parents) is determined according to the theoretical number of conformers which could be generated for the structure under consideration. In turn, the theoretical number of possible conformers depends on the flexibility of the structure as defined by the number of associated conformational variables.

Two modes of the algorithm can be applied using different criteria for evaluating of newly generated conformers. In both modes, the conformers are evaluated based on properties of the entire population. According to the first mode, the populations are selected that have the maximum rms atomic distances which tend to prevent forming of isolated clusters of similar conformers. The iterative process of generating the population of conformations is terminated when the average rms between generated conformers converges and there is no significant improvement in the population properties from additional conformations. At each iteration, the newly generated population can be characterized with an average rms distance not smaller than the distances corresponding to previous iterations. Generally, the average distance increases in each iteration step, because more "fitted" parents with large individual scores are selected. In the program implementation, two different ending criteria can be separately or jointly imposed. The first one fixes a limit for the number of iterations. The second one is a convergence test which requires that the average rms increase over several successive iterations drops below a user-defined threshold.

The second mode is based on the population entropy which is maximized using the Shannon entropy function, according to which the entropy S_j of conformer j could be expressed by the equation

$$S_j = -\sum_{\substack{j=1 \\ i\neq j}}^{J} P_{ij}\ln(P_{ij}) \tag{1}$$

where

$$P_{ij} = \frac{p_{ij}}{\sum_j p_{ij}} \tag{2}$$

p_{ij} is rms D between conformer i and j

The comparative analysis between fitness functions based on maximization of rmsd only and the use of this criterion in combination with the Shannon function showed that the entropy based fitness function tends to generate a more uniform distribution of conformers over the conformational space. In fact, the Shannon entropy function is applied in a combinatorial scheme to select the population of new parents (for the next evolution step of GA) providing the best coverage of the space among all generated conformers (parents and children) of the current GAS step.

Entropy of a whole generated set of conformers is

$$H = -\sum_{i} \sum_{\substack{j \\ i \neq j}} P_{ij} \ln(P_{ij}) \qquad (3)$$

To avoid the combinatorial problem for selection of parent population an algorithm was developed to identify the most "informative" conformers among generated ones using eq 1.

All generated conformers are screened for rejection criteria based on the degeneracy, energy, and quality of the structures. Degenerate conformers are those which are similar to other structures with respect to any of their torsion angles. The energy barriers for rotation around single bonds are 3—corresponding to 120° rotation; however, this torsion resolution is large enough to explain the flexibility in some polycyclic saturated chemicals. In this respect, a threshold of 60° is selected as appropriate (but it can be user defined). The energy of each conformation is then calculated by using a truncated force field energylike function, where the electrostatic terms are omitted (called pseudo molecular mechanics, PMM[7]). In fact, PMM is a strain-relief procedure. The calculated PMM energies are compared to a threshold value to eliminate highly strained structures. The quality filter is used to evaluate generated conformers in terms of violation of common bond lengths, valence and torsion angles, and nonbonded distances, etc. all based on force field parametrization.

Geometry optimization is further completed by quantum-chemical methods. MOPAC 93[32,33] is employed. Next, the conformers are screened to eliminate those whose heat of formation, ΔH_f^o, is greater than the ΔH_f^o associated with the conformer with absolute energy minimum by a user defined threshold—usually a range of 20 kcal/mol (or 15 kcal/mol) from the lowest energy conformers is analyzed. Subsequently, conformational degeneracy, due to molecular symmetry and geometry convergence, is detected within a user defined torsion angle resolution.

2.3. Conformational Distributions and Smoothing Parameter. The COREPA approach is based on Bayes theorem and provides a theoretically optimal decision rule.[34] The algorithm examines the distribution of all energetically reasonable conformations for the structure but selects active conformations on a case-by-case basis using the activity endpoints of specific studies. The probability distribution for the conformers is approximated from a Boltzman energy distribution. The probability of forming a specific conformer is $p(x|\mathrm{Csj})$, where Csj denotes the *j*th conformer of the chemical S. The probability of that chemical having a specific molecular descriptor value is denoted as $p(x|S_i)$ in eq 4

$$p(x|S_i) = \sum_{j=1}^{R_i} p(C_{ij}) p(x|C_{ij}) \qquad (4)$$

where S_i is the *i*th chemical in the data set, R_i is the number of conformers for the compound Si, and $p(C_{ij})$ is the probability to have the *j*th conformer of an *i*th compound.

The Boltzman probability $p(C_{ij})$ can be estimated by eq 5

$$p(C_{ij}) = \frac{e^{-\Delta E_j / k_B T}}{\sum_{m=1}^{N} e^{-\Delta E_m / k_B T}} \qquad (5)$$

where $\Delta E_j = E_j - E_{\min}$, and $E_{\min}$ is the energy of the conformer with minimal energy.

The application of the formula for the kernel density estimate to $p(x|C_{ij})$ gives

$$p(x|C_{ij}) = \frac{1}{N_{ij} h} \sum_{k=1}^{N_{ij}} \varphi\left(\frac{x - x_{ijk}}{h}\right) \qquad (6)$$

where N_{ij} is the number of values of descriptor *x* for the *j*th conformer of *i*th chemicals.

Substitution of eqs 5 and 6 into eq 4 allows the calculation of a conformational distribution of a chemical across a descriptor *x*.

$$p(x|S_i) = \sum_{j=1}^{R_i} \frac{e^{-\Delta E_j / k_B T}}{\sum_{m=1}^{N} e^{-\Delta E_j / k_B \mathrm{T}}} \left[\frac{1}{N_{ij} h} \sum_{k=1}^{N_{ij}} \varphi\left(\frac{x - x_{ijk}}{h}\right) \right] \qquad (7)$$

To create a probability distribution for each value of descriptor *x*, a kernel density function[31] is superimposed on each individual data point, and these data density kernels are summed and normalized to give an overall probability distribution. The kernel density function, $\varphi(x)$, provides a bounded symmetrical probability distribution function for estimation of the class-conditional probability distribution as shown in eq 8

$$p(x) = \frac{1}{nh} \sum_{k=1}^{n} \varphi\left(\frac{x - x_k}{h}\right) \qquad (8)$$

where *h* is a smoothing parameter. The smoothing function can be optimized through cross-validation; however, COREPA sets the initial smoothing as $h = 1.059\sigma\sqrt[5]{n}$, σ being the standard deviation of the data set and *n* being the number of data points.[31]

Our attempts to define an optimal value for the smoothing parameter in order to maximize the reproducibility of conformer distributions generated in two subsequent applications of the nondeterministic GAS algorithm failed. As shown below, small values of this parameter produce highly detailed continuous approximation of the conformer distribution which memorizes data without any generalization of reactivity pattern (Figure 1a). Alternatively, very large values of the smoothing parameter could bring such a fuzziness of reactivity pattern which could cause a bias in the derived reactivity pattern (Figure 1b)

On the other hand, it is intuitive that the reproducibility of conformer distributions will increase with the increase of the number of generated conformers. Hence, a solution of this problem seemed to be a procedure increasing stepwise the number of conformers until no significant change in the derived conformer distributions is observed. The purpose of such an algorithm is to generate an optimal number of conformers to represent conformational space producing a stable conformer distributions not affected by the addition of new conformers. Such an algorithm was expected to be convergent given the convergence of the algorithm for covering conformer space.

2.4. Procedure for Optimal Saturation of Conformer Space. Three approaches for saturation of conformational

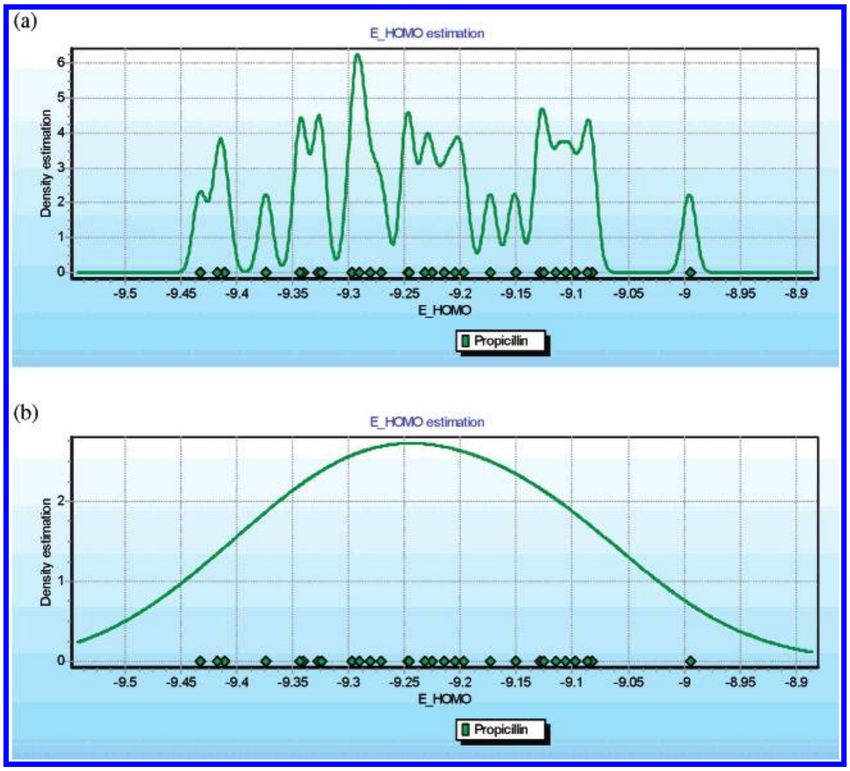


Figure 1. Probabilistic conformer distribution for propicilin (CAS 551-27-9) across E_{HOMO} at (a) $h = 0.1$ and (b) $h = 1.5$.

space were investigated, in the present work. All of them are based on the independent application of the evolutionary algorithm (GA) for conformer generation. This algorithm is applied consecutively a number of times until the conformational space is saturated within a user defined population density. During the saturation process, the system is trying to optimize the number of conformers used to cover the conformational space by eliminating part of the generated conformers in the subsequent application of GA and preserving some of them mostly contributing to the coverage of the conformational space. The differences in the proposed three approaches for saturation of the conformational space are in the mechanism of removal of these “useless” conformers. The removal schemes could be summarized as follows:

Removal Accounting for Degeneracy of Conformers. The less informative conformers as evaluated by the Shannon entropy formula[25] are removed until one of the closest neighbors (most similar conformers as evaluated by rms) needs to be eliminated. Then, the elimination procedure stops, and degenerate conformers are removed at a user defined degeneracy threshold for torsion angle resolution (60 degrees, by default).

Removal of Conformers Deviating by More Than 1σ from the Population Mean (1σ − Approach). The conformer distribution across the individual Shannon entropies is built. If there are conformers deviating from the mean, on its left site, with more than one sigma (one standard deviation), then the conformers which appear to be the farthest reaches are removed. The above procedure is repeated, with recalculated individual entropies, until no conformer appears apart from the mean by one sigma.

Removal of Conformers Deviating by More Than 2σ from the Population Mean (2σ − Approach). This approach is similar to the second one; however, the deviations of the conformers from the mean (from its left side) are identified within 2 standard deviations (2σ).

The stop criterion for adding new conformers is defined by the difference in the Hellinger distance between the conformer distributions[31,34] of two consecutive applications of GA. In the present version of the procedure, this difference is set to not exceed 0.005 in two consecutive applications of the GA algorithm. In the case when this criterion is not reached in 30 iterations the saturation procedure stops. Experiments with a selected set of chemicals showed that a larger difference in the Helinger distance could result in quite different distributions (such exercises have been performed with ΔHD = 0.05). We also decided to rely on the more stringent 2σ-approach for eliminating uninformative conformers during the saturation procedure. This approach generates a larger collection of conformers which, however, still can be handled by our software.

To reach stable distributions with respect to molecular geometry and electronic structure, it is recommended that the saturation procedure for conformer generation be applied independently for two global molecular descriptors: geometric—sum of the geometric distances in the molecule (Geom Winer)—and an electronic—difference in the frontier orbitals (Egap) (for molecular descriptors used in the OASIS system one can consult ref 1). This means that the stop criteria of the saturation procedure have to be fulfilled simultaneously for both parameters. The selection of molecular descriptors with respect to which the saturation procedure is applied is context dependent; in other words, the specificity of the subsequent modeling task should define this selection. As it will be shown in the Results section, the saturation with one of the parameters does not necessarily lead to saturation with respect to the other parameter(s).

The analysis of the generated conformers after the saturation procedure showed that for relative small and rigid molecules, some of the produced conformers are structurally very similar. Because of that, at the end of each saturation step a new elimination criterion is added: the shortest rms distance between corresponding atoms of each pair of conformers to be larger than a threshold (e.g., 0.1 Å). If during the conformer multiplication process no saturation is reached after a user defined number of iteration steps, then the above discriminating distance is decreased by user a defined percent allowing the description of the conformational space with more similar conformers (i.e., allowing less stringent coverage constraints). Two schemes have been developed for eliminating identical conformers. The first one, called "minimizing", is aimed at minimizing the total number of conformers during the elimination process; e.g., if we have three conformers—the first one similar to the second and the second one similar to the third, then the algorithm eliminates the first and third conformers preserving the second as a representative one. This is not the case in the second scheme, called "maximizing". Here, the system will preserve the first and third conformers and will eliminate the second one. The first scheme corresponds to the idea of minimizing the number of conformers encompassing the conformational space. The second scheme, however, appears to be effective in case of more rigid structures, where the "minimizing" algorithm significantly reduces the number of conformers. The user can choose one of these schemes. If saturation is not reached by a selected scheme, then the system automatically switches to the other one. If no saturation is reached either by the alternatively selected scheme, then the scheme with the smaller number of conformers is preserved (in case the number of conformer is not 1) and the system produces a warning message that no saturation is reached. Structures with a small number of conformational variables are multiplied by the deterministic application of the conformer multiplication algorithm, and because of that no saturation procedure is performed.

3. RESULTS AND DISCUSSION

The capability of the conformational space saturation procedure is demonstrated in Figure 2, where the conformer distribution of methyldimethylaminoazobenzene (Figure 2a) (CAS 58-80-1) across Egap is derived by applying the saturation procedure.

As seen, the Hellinger distance (HD) between the first and second run of GA is 0.014, and the procedure needs 9 iterations to meet the stop criterion—HD = 0.005 in two consecutive applications of the GA algorithm. The first conformer distribution curve is formed by 26 conformers (Figure 2b), the second one—by 46 conformers, and finally—122 conformers represent the saturated conformational space, at the ninth iteration (Figure 2i). A threshold of HD = 0.005 in two consecutive runs of GA is assumed to provide convergence of conformer distributions and is an indication that the subsequent addition of new conformers will not affect the generated conformer distribution.

All generated conformers for methyldimethylaminoazobenzene used to illustrate the saturation procedure are quantum-chemically optimized (MOPAC 93, AM1, PRECISE[32,33]). These are the energetically reasonable isomers of the molecule within the formation enthalpy range of 20 kcal/mol from the lowest energy conformer. The same requirements for the energy of the generated conformers are applied for the other case studies analyzed in this work. In this respect, Figure 2 also demonstrates the critical importance of the conformational analysis for molecular modeling. As seen, the range of Egap values for energetically reasonable conformers of methyldimethylaminoazobenzene exceeds 1 eV.

The similar saturation of conformational space of methyldimethylaminoazobenzene with respect to the global geometric descriptor—sum of the steric distances between non-hydrogen atoms in the molecule (Geom Wiener), required three iteration steps, only, to reach convergence (Figure 3). Apparently, this is due to relatively low flexibility of the studied molecule. A comparison with Figure 2 shows that the procedure is far from convergence at the third iteration step, with respect to the Egap parameter.

The saturation of conformational space of methyldimethylaminoazobenzene with respect to Geom Wiener and Egap, simultaneously, apparently requires 9 iterations (as shown in Figure 2). The reproducibility of the conformer distribution curves across Geom Wiener, during the saturation process where the convergence is estimated with respect to Geom Wiener and Egap simultaneously, is shown in Figure 4.

The performance of the procedure for saturation of conformational space was also analyzed for a flexible alkane—decane ($C_{10}H_{22}$). The results are presented in the Supporting Information.

The saturation of conformational space of the molecule required a significant increase in the number of conformers as compared with those generated by a single application of GA. This increase for the studied molecules and molecular descriptors is illustrated in Table 1.

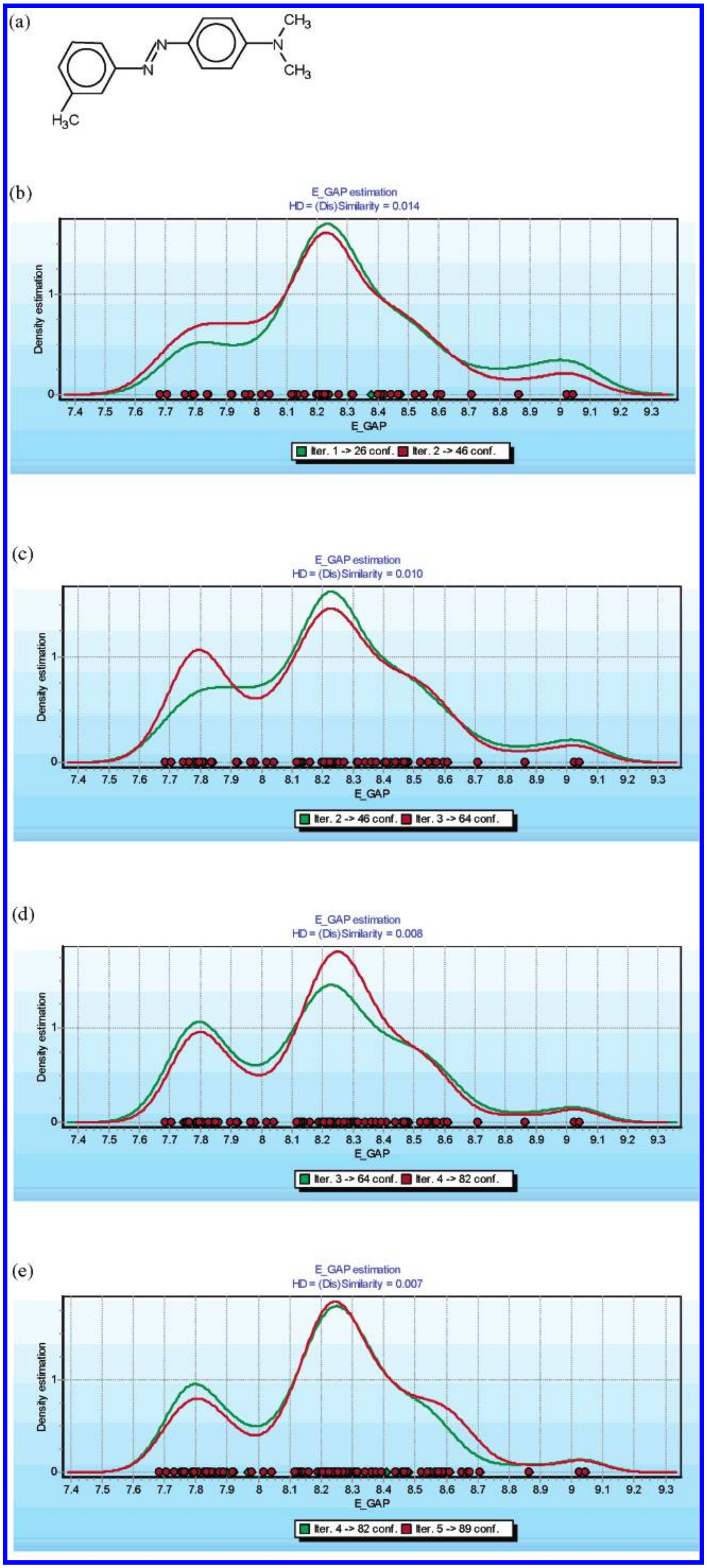
As seen, when the convergence of the saturation process for methyldimethylaminoazobenzene is estimated with respect to both parameters, for one of them the saturations is reached much earlier (third step for Geom Wiener).

The saturation of conformational space of the molecule, however, required a significant increase in the number of conformers as compared with those generated by a single application of GA. This increase for methyldimethylaminoazobenzene and decane (see also the Supporting Information) associated with different molecular descriptors is illustrated in Table 1.

As seen, the number of conformers for methyldimethylaminoazobenzene and decane increase from 18 to 78 and 26 to 64, respectively, when the conformational analysis is across Geom Wiener. Similarly, the increase with respect to Egap is from 18 to 79 and from 26 to 122, respectively. Apparently, the increase for the electronic parameter is larger than the increase associated with the geometric parameters. This could be related to the higher sensitivity of the studied electronic parameter (Egap) to conformational changes as compared with the geometric parameter (Geom Wiener). We assume this conclusion could be generalized to other electronic and geometric molecular descriptors.

As expected the saturation with respect to both the electronic and geometric descriptors, simultaneously, requested a higher number of conformers—122 and 85 for methyldimethylaminoazobenzene and decane, respectively.

In the past decade, in a number of papers we have formulated the basic concepts of the OASIS modeling approach, according to which the molecular flexibility needs to be accounted for in QSAR analysis and similarity assessment. In this respect, it could be very informative to


(a)
CH3
N
CH3
N
N
H3C
(b)
E_GAP estimation
HD = (Dis)Similarity = 0.014
Density estimation
0
1
7.4 7.5 7.6 7.7 7.8 7.9 8 8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9 9 9.1 9.2 9.3
E_GAP
Iter. 1 -> 26 conf.
Iter. 2 -> 46 conf.
(c)
E_GAP estimation
HD = (Dis)Similarity = 0.010
Density estimation
0
1
7.4 7.5 7.6 7.7 7.8 7.9 8 8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9 9 9.1 9.2 9.3
E_GAP
Iter. 2 -> 46 conf.
Iter. 3 -> 64 conf.
(d)
E_GAP estimation
HD = (Dis)Similarity = 0.008
Density estimation
0
1
7.4 7.5 7.6 7.7 7.8 7.9 8 8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9 9 9.1 9.2 9.3
E_GAP
Iter. 3 -> 64 conf.
Iter. 4 -> 82 conf.
(e)
E_GAP estimation
HD = (Dis)Similarity = 0.007
Density estimation
0
1
7.4 7.5 7.6 7.7 7.8 7.9 8 8.1 8.2 8.3 8.4 8.5 8.6 8.7 8.8 8.9 9 9.1 9.2 9.3
E_GAP
Iter. 4 -> 82 conf.
Iter. 5 -> 89 conf.

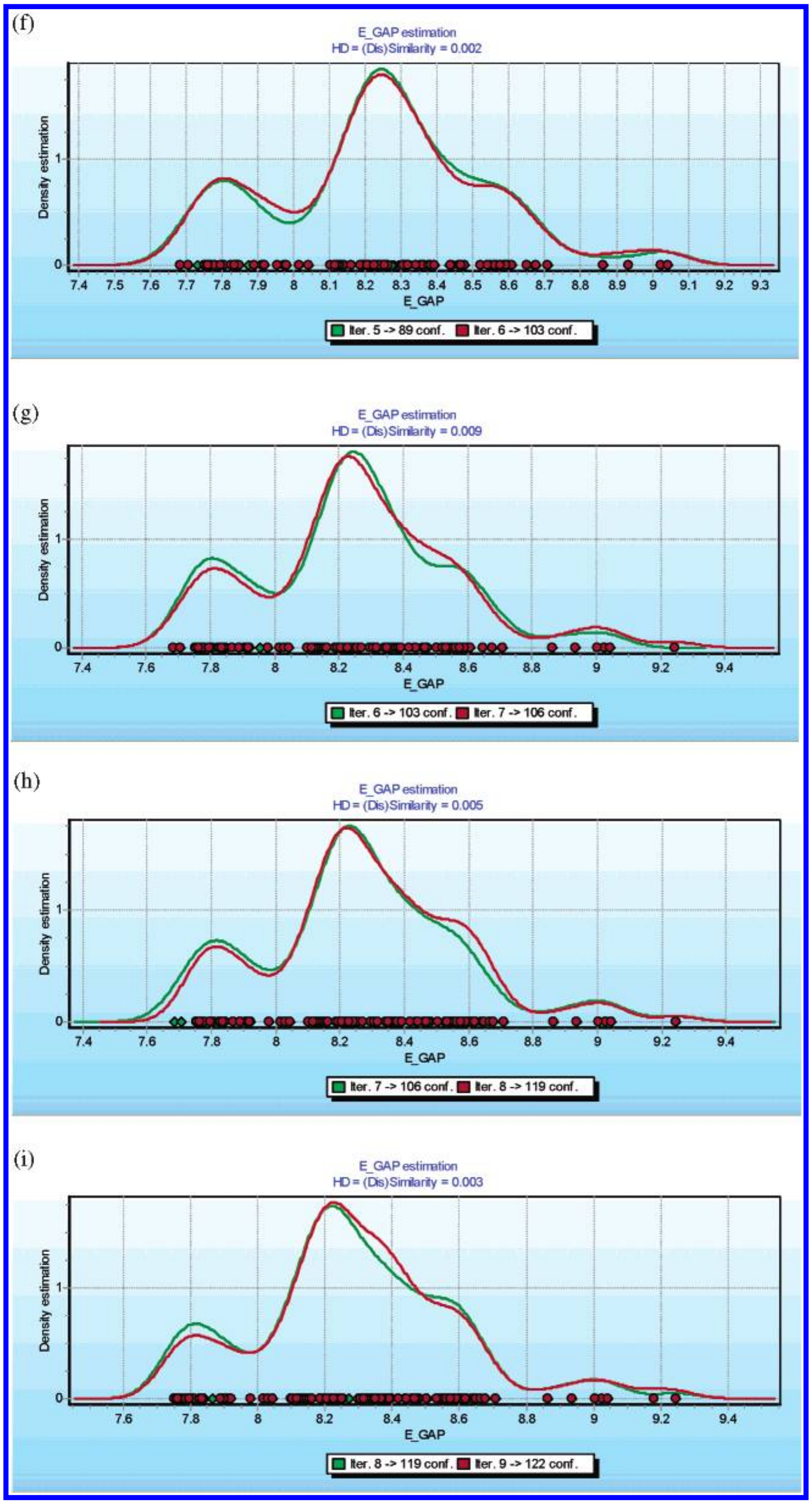


Figure 2. The subsequent steps of the conformer saturation procedure of methyldimethylaminoazobenzene (CAS 58-80-1) (a) across Egap ((b)−(i)) when the convergence of the saturation process is estimated with respect to the same parameter. In this and subsequent case studies, the stop criterion of the convergency procedure is based on the Hellinger distance (HD) between two probabilistic distributions. The imposed HD threshold for convergence is 0.005 in two consecutive applications of GA. In the analyzed example saturation is reached at the ninth iteration step.

compare the conformer distributions of cyclohexamide (CAS 66-81-9) and glycol-ether tetraacetic acid (CAS 67-42-5) across Egap. Conformers are produced by two generation procedures: the individual run of the evolutionary algorithm (Figure 5) and by applying the conformational space saturation procedure (Figure 6). As in the preceding examples, all generated conformers are quantum-chemically optimized (MOPAC 93, AM1, PRECISE); energetically reasonable isomers are selected, only, falling within the formation enthalpy range of 20 kcal/mol from the lowest energy conformer.

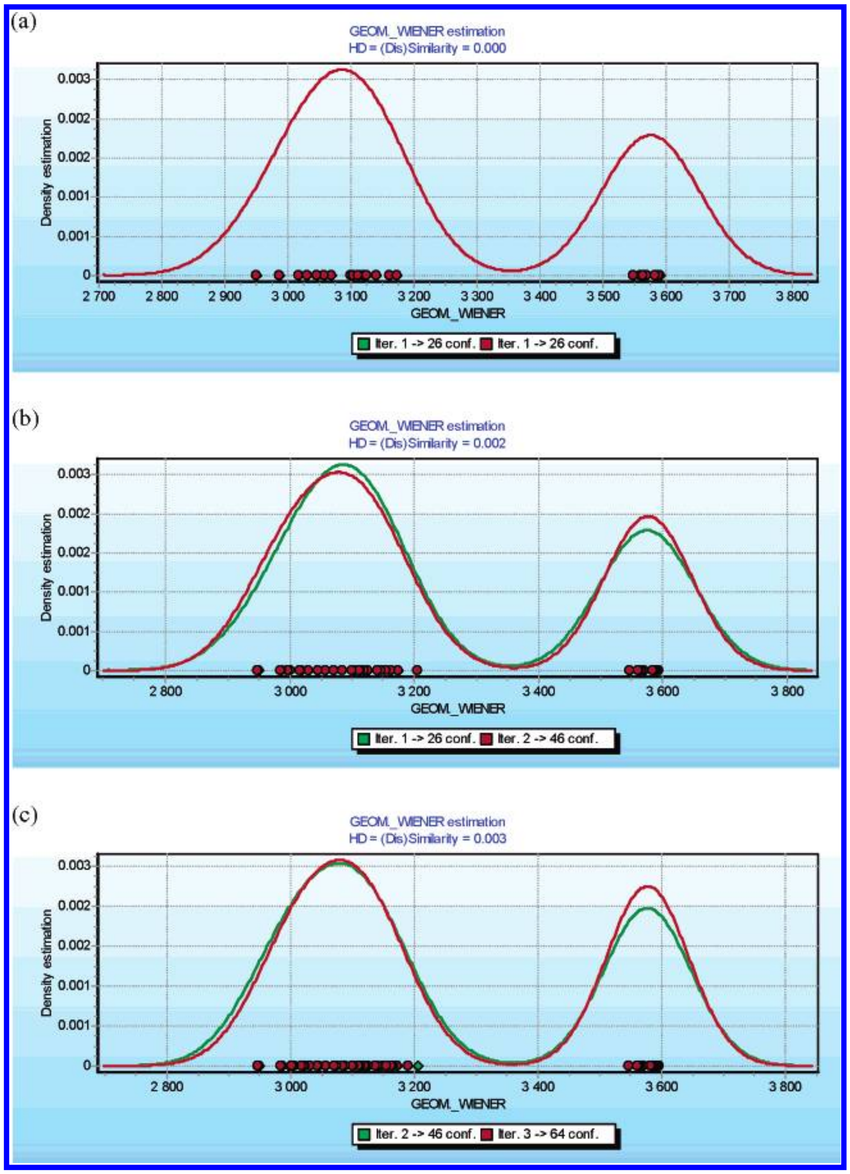


Figure 3. The subsequent steps of the conformer saturation procedure for methyldimethylaminoazobenzene (CAS 58-80-1) across Geom Wiener ((a)−(c)) when the convergence of the saturation process is estimated with respect to Geom Wiener parameter, only.

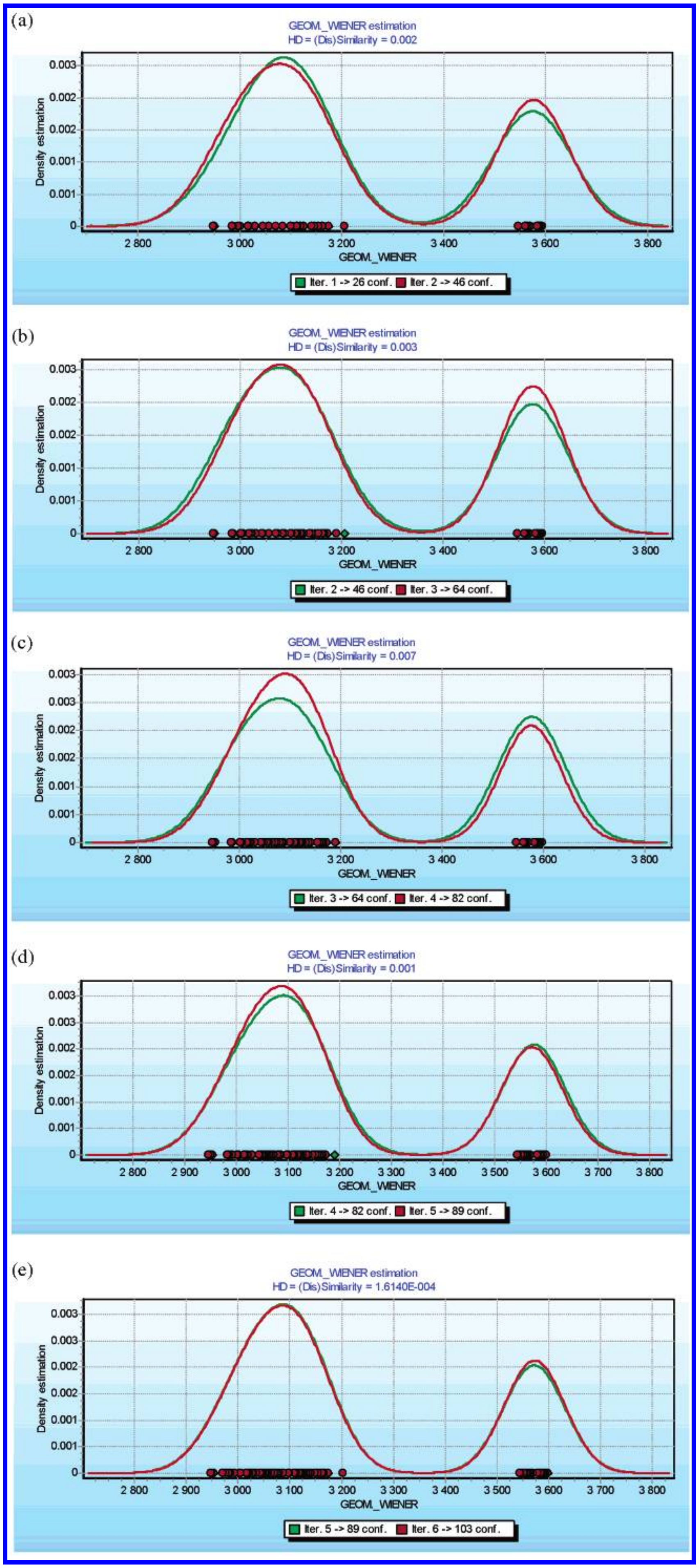
If the goal of the current molecular modeling exercise is the estimation of similarity between both chemicals with respect to Egap, one should compare the values of this parameter for compared structures. As seen from the discrete conformer distributions in Figures 5a and 6a, however, it is very difficult to make such a comparison given the multiplicity of stable conformers across the selected parameter. Recently, a new formalism to define similarity between chemicals has been introduced based on comparison continuous conformer distributions of the chemicals instead of the traditional overlapping of their 3D (or 2D) structures.[26,35] Given the probabilistic character of conformer distributions one could determine the Hellinger distance (HD) between these distributions as a measure for their similarity with respect to the selected molecular descriptor(s). According to the formalism of this metrics, HD = 0 when the conformational distributions fully coincide and HD = 2 in the opposite case. The first extreme case (HD=0) corresponds to the highest possible similarity of chemicals with respect to the selected molecular descriptor.

As seen, HD has significantly different values depending on whether conformer distributions are based on individual runs of an evolutionary algorithm (Figure 5b) or on the saturation procedure for conformer generation (Figure 6b). In other words, the estimated similarity between chemicals is higher when the conformational space is saturated (HD=1.32) as compared to conformer generation based on individual GA runs (HD=1.83). This example demonstrates the need of saturation of conformational space for adequate estimation of similarity between molecules.

One should keep in mind that although the saturation of conformational space requires the generation of more conformers than the individual runs of the evolutionary algorithm, still a limited number of conformers are used to represent the conformational space.

SUMMARY AND CONCLUSIONS

In a series of preceding works we have shown that conformational flexibility of chemicals should be taken into account in molecular modeling and similarity analysis. In


(a)
GEOM_WIENER estimation
HD = (Dis)Similarity = 0.002
Density estimation
0.003
0.002
0.002
0.001
0.001
0
2 800
3 000
3 200
3 400
3 600
3 800
GEOM_WIENER
🟩 Iter. 1 -> 26 conf. 🟥 Iter. 2 -> 46 conf.
(b)
GEOM_WIENER estimation
HD = (Dis)Similarity = 0.003
Density estimation
0.003
0.002
0.002
0.001
0.001
0
2 800
3 000
3 200
3 400
3 600
3 800
GEOM_WIENER
🟩 Iter. 2 -> 46 conf. 🟥 Iter. 3 -> 64 conf.
(c)
GEOM_WIENER estimation
HD = (Dis)Similarity = 0.007
Density estimation
0.003
0.003
0.002
0.002
0.001
0.001
0
2 800
3 000
3 200
3 400
3 600
3 800
GEOM_WIENER
🟩 Iter. 3 -> 64 conf. 🟥 Iter. 4 -> 82 conf.
(d)
GEOM_WIENER estimation
HD = (Dis)Similarity = 0.001
Density estimation
0.003
0.003
0.002
0.002
0.001
0.001
0
2 800
2 900
3 000
3 100
3 200
3 300
3 400
3 500
3 600
3 700
3 800
GEOM_WIENER
🟩 Iter. 4 -> 82 conf. 🟥 Iter. 5 -> 89 conf.
(e)
GEOM_WIENER estimation
HD = (Dis)Similarity = 1.6140E-004
Density estimation
0.003
0.003
0.002
0.002
0.001
0.001
0
2 800
2 900
3 000
3 100
3 200
3 300
3 400
3 500
3 600
3 700
3 800
GEOM_WIENER
🟩 Iter. 5 -> 89 conf. 🟥 Iter. 6 -> 103 conf.

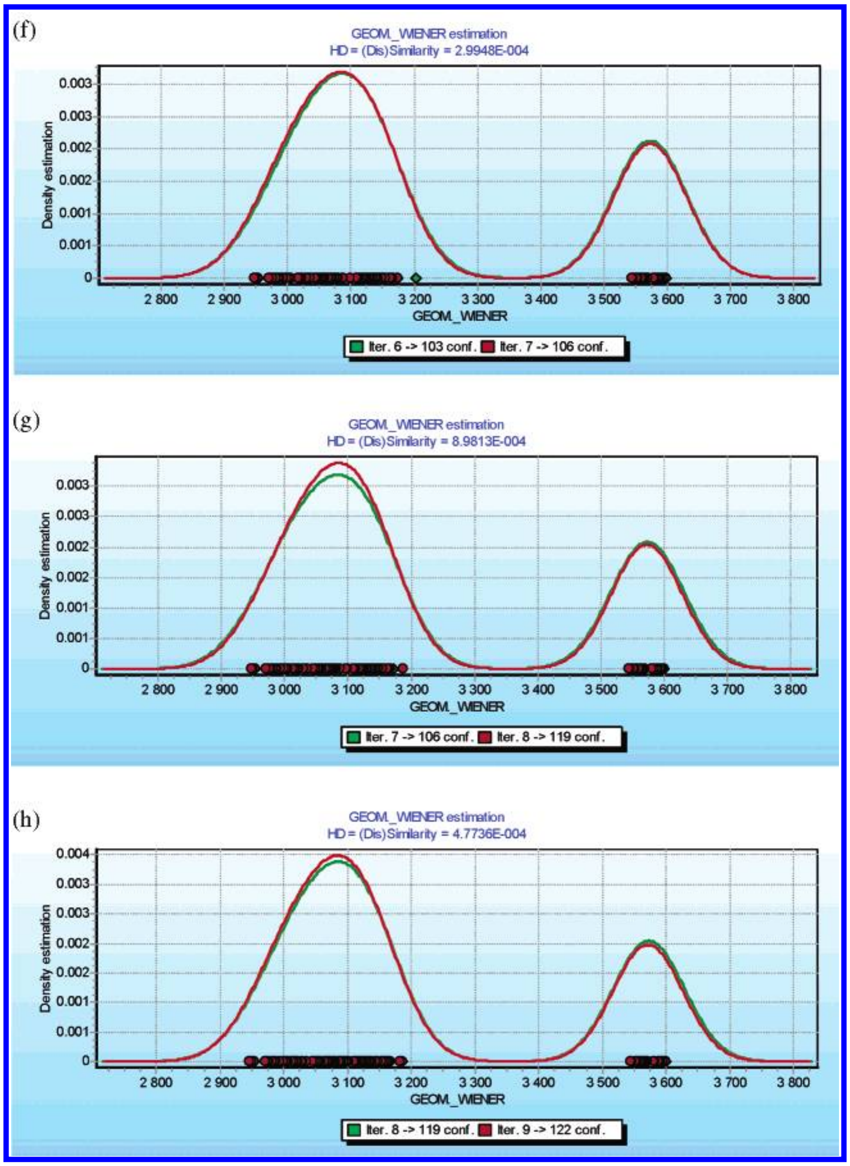
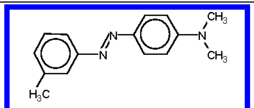
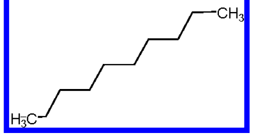


Figure 4. The subsequent steps of the conformer saturation procedure for methyldimethylaminoazobenzene (CAS 58-80-1) across Geom Wiener when the convergence of the saturation process is estimated with respect to Geom Wiener and Egap simultaneously ((a)−(h)).

Table 1. Comparison between the Number of Conformers of Methyldimethylaminoazobenzene and n-Decane Generated by the Saturation Procedure and Individual GA Runs Across Egap and Geom Wiener Parameters, Respectively

			GA		saturation		
CAS	name	structure	GW	Egap	GW	Egap	GW+Egap
55-80-1	methyldimethylamino-azobenzene		18	18	78	79	85
124-18-5	n-decane		26	26	64	122	122

this respect, significant efforts are focused toward developing effective algorithms for adequate conformational analysis. Further, conformer distributions in a selected structural space could be used for similarity analysis or deriving a common reactivity pattern for biologically similar chemicals for a subsequent screening and prioritization process.

This paper represents a new improvement of the procedure for conformational analysis based on GA. The GA based procedure we have developed so far was found to have two major drawbacks: lack of reproducibility of generated conformers (and their distributions) due to the nondeterministic character of the evolutionary algorithm and sensitivity of conformer distributions from the effect of the smoothing parameter. These insufficiencies could cause a bias in similarity estimates and derived reactivity patterns associated with biologically similar chemicals which, in turn, could produce nonadequate QSAR models. Our attempt to improve the reproducibility of conformer distributions by optimizing

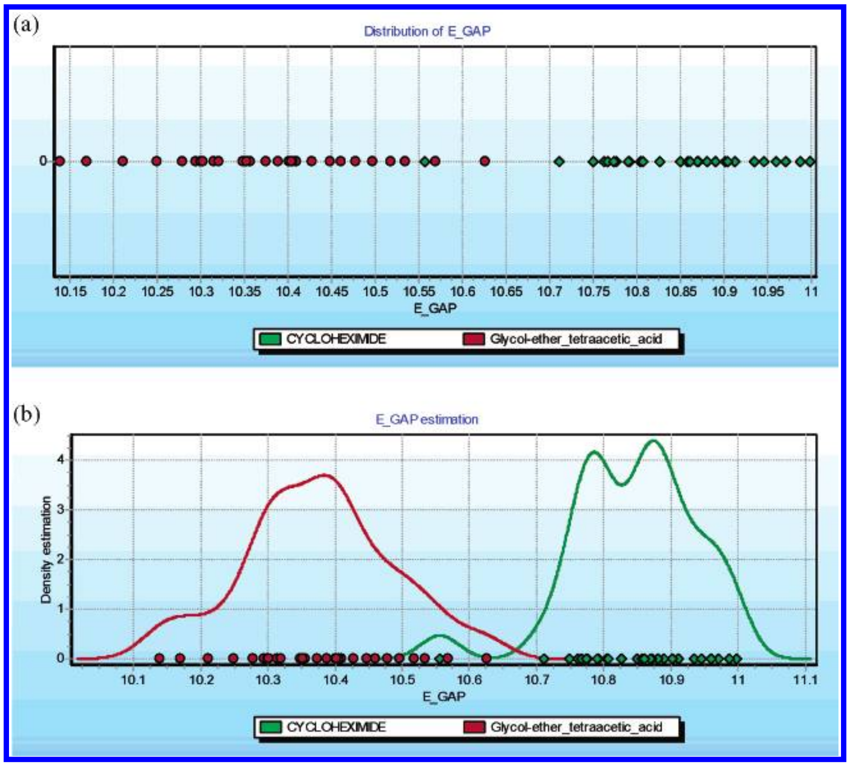


Figure 5. Discrete (a) and continuous (b) conformer distributions of cyclohexamide (CAS 66-81-9) and glycol-ether tetraacetic acid (CAS 67-42-5) across Egap. Conformers are generated by individual runs of the evolutionary algorithm (GA).

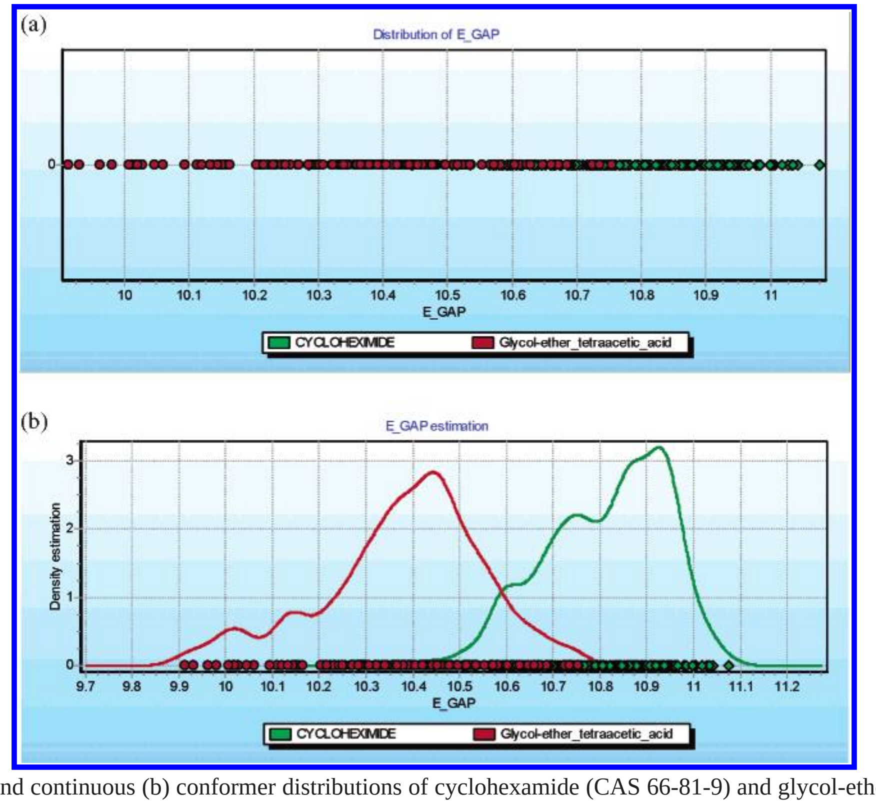


Figure 6. Discrete (a) and continuous (b) conformer distributions of cyclohexamide (CAS 66-81-9) and glycol-ether tetraacetic acid (CAS 67-42-5) across Egap. Conformers are generated by applying conformational space saturation procedure.

the values of the smoothing parameter failed. An alternative approach has been developed in the present work based on the following—at a statistically recommended constant value of smoothing parameter, the number of conformers representing conformational space is increased until there is no significant variation in conformer distributions for two subsequent runs of the evolutionary algorithm for conformer generation. The proposed saturation procedure generates stable conformer distributions with respect to selected molecular descriptors, which are not affected by the addition of new conformers to the generated representative set. The "cost" of the saturation procedure is the generation of a larger number of conformers to represent the conformational space of the molecules as compared with the number of conformers generated by the individual runs of GA. Still, the procedure associates a limited number of isomers with the studied chemicals. Moreover, the number of conformers needed to saturate the conformational space is optimized in a way to not significantly exceed this one produce by the individual GA runs.

The saturation of conformational space provides more adequate similarity estimates between molecules and more reliable 3D QSAR models accounting for the conformational flexibility of chemicals.

In the present work, the saturation procedure is exemplified by two molecular descriptors, Egap and Geom Wiener, selected to represent the electronic and geometric features of molecules. The preliminary studies showed that not always does the saturation with respect to selected electronic descriptor ensure saturation with respect to all other electronic descriptors; the same holds true with geometric parameters. In case of no such a relationship, the saturation procedure should be applied to get convergence with respect to the list of descriptors rather than a single descriptor. What is the "cost" in terms of an increase in the number of conformers to provide an optimal coverage of conformational space with respect to a larger list of descriptors needs to be investigated. Additional work to analyze this relationship is in progress.

ACKNOWLEDGMENT

Research associated with this paper was funded in part through the ReProTect Integrated Project funded by the EC under the Sixth Research and Technological Development Framework Programme VI RTD and EPA cooperative research agreement (CR83199501). This paper has been subjected to review by the National Health and Environmental Effects Research Laboratory and approved for publication. Approval does not signify that the contents reflect the views of the Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Supporting Information Available: Performance of the procedure for saturation of conformational space for decane. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Mekenyan O. G.; Nikolova N.; Schmieder P. Dynamic 3D QSAR techniques: application in toxicology. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 147−165.

(2) Lipton, M.; Still, W. C. The Miltiple Minimum Problem in Molecular Modeling. Tree Searching Internal Coordinate Conformational Space. *J. Comput. Chem.* **1988**, *9*, 343−355.

(3) Loferer, M. J.; Kolossovary, I.; Aszodi, A. Analyzing the performance of conformational search programs on compound databases. *J. Mol. Graphics Modell.* In press.

(4) *SYBYL*; Tripos Associates Inc.: St. Louis, MO.

(5) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained Search of Conformational Hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3−21.

(6) Lipton, M.; Still, W. C. The Miltiple Minimum Problem in Molecular Modeling. Tree Searching Internal Coordinate Conformational Space. *J. Comput. Chem.* **1988**, *9*, 343−355.

(7) Ivanov, J. M.; Karabunarliev, S. H.; Mekenyan. O. G. 3DGEN: A system For an Exhaustive 3D Molecular Design. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 234−243.

(8) Leach, A. R.; Kuntz, I. D. Conformational Analysis of Flexible Ligands in Macromolecular Receptor Sites. *J. Comput. Chem.* **1992**, *13*, 730−748.

(9) *Omega, Version 2.0*; Openeyes Scientific Software: Santa Fe, NM, 2006.

(10) *Catalyst, Version 4.11*; Accelrys: San Diego, CA, 2006.

(11) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage. 2. Applications of conformational models. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (2), 295−304.

(12) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of conformational coverage. 1. Validation and estimation of coverage. *J. Chem. Inf. Comput Sci.* **1995**, *35* (2), *285−94.*

(13) Kirchmair, J.; Laggner, C.; Wolber, G.; Langer, T. Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithms. *J. Chem. Inf. Model.* **2005**, *45* (2), 422−430.

(14) Kirchmair, J.; Wolber, G.; Laggner, C.; Langer, T. Comparative Performance Assessment of the Conformational Model Generators Omega and Catalyst: A Large-Scale Survey on the retrieval of Protein-Bound Ligand Conformations. *J. Chem. Inf. Model.* **2006**, *46* (4), 1848 -1861.

(15) Burkert, O.; Allinger, N. L. *Molecular mechanics*; ACS Monographs, Washington, DC, 1982.

(16) Chang, G.; Guida, W. C.; Still, W. C. An Internal Cordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379−4386.

(17) Kolossvary, I.; Guida, W. C. Low-mode conformational search elucidated: Application to C39H80 and flexible docking of 9-Deazaguanine inhibitors into PNP. *J. Comput. Chem.* **1999**, *20*, 1671−1684.

(18) Abagyan, R.; Totrov, M. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J. Mol. Biol.* **1994**, *235*, 983−1002.

(19) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(20) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46* (1), 401−415,

(21) McMartin, C.; Bohacek, R. S. QXP: powerful, rapid computer algorithms for structure-based drug design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 333−344.

(22) Spellmeyer, D. C.; Wong, A. K.; Bower, M. J.; Blaney, J. M. Conformational analysis using distance geometry methods. *J. Mol. Graphics Modell.* **1997**, *15*, 18−36.

(23) Mekenyan, O. G.; Dimitrov, D.; Nikolova, N.; Karabunarliev, S. Conformational Coverage by a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 997−1016.

(24) Payne, A. W. R.; Glen, R. C. Molecular recognition using a binary genetic search algorithm. *J. Mol. Graphics Modell.* **1993**, *11*, 74−91.

(25) Mekenyan, O. G.; Pavlov, T.; Grancharov, V.; Todorov, M.; Schmieder, P.; Veith, G. 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm. *J. Chem. Inf. Model.* **2005**, *45*, 283−292.

(26) Mekenyan, O. G.; Nikolova, N.; Schmieder, P.; Veith, G. D. COREPA-M: A Multi- Dimensional Formulation of COREPA. *QSAR Comb. Sci.* **2004**, *23*, 5−18.

(27) Bradbury, S.; Kamenska, V.; Schmieder, P.; Ankley, G.; Mekenyan, O. A Computationally-Based Identification Algorithm for Estrogen Receptor Ligands.Part I. Predicting hER Binding Affinity. *Toxicol. Sci.* **2000**, *58*, 253−269.

(28) Mekenyan, O. G.; Kamenska, V.; Schmieder, P.; Ankley, G.; Bradbury, S. A Computationally-Based Identification Algorithm for Estrogen Receptor Ligands. Part II. Evaluation of a hER Binding Affinity Model. *Toxicol. Sci.* **2000**, *58*, 270−281.

(29) Mekenyan, O. G.; Dimitrov, S.; Pavlov, T.; Veith, D. G. A Systems Approach to Simulating Metabolism in Computational Toxicology. I. The TIME Heuristic Modelling Framework. *Curr. Pharm. Des.* **2004**, *10* (11), 1273−1293.

(30) Dimitrov, D.; Low, K. L; Patlewicz, G.; Kern, P.; Dimitrova, G.; Comber, M.; Philips, R.; Niemela, J.; Bailey, P.; Mekenyan, O. Skin Sensitization: Modeling Based on Skin Metabolism Simulation and Formation of Protein Conjugates. *Int. J. Toxicol.* **2005**, *24*, 189−204.

(31) Gibbs, A. L; Su. F. E. On choosing and bounding probability metrics. *Intl. Stat. Rev.* **2002**, *7* (3) 419−435.

(32) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

(33) *MOPAC 93*; Fujitsu Limited: Japan, and Stewart Computational Chemistry: Colorado Springs, CO, U.S.A., 1993

(34) Duda R. O.; Hart, P. E.; Stork, D. *Pattern Classification*, 2nd ed.; John Wiley & Sons: 2000; pp 538−542.

(35) Nikolov, N.; Grancharov, V.; Stoyanova, G.; Pavlov, T.; Mekenyan, O. Representation of Chemical Information in OASIS Centralized 3D Database for Existing Chemicals. *J. Chem. Inf. Model.* **2006**, *46* (6), 2537−2551.

CI700014H