

## 2D-3D Migration of Large Chemical Inventories with Conformational Multiplication. Application of the Genetic Algorithm

Ovanes Mekenyan,<sup>\*,†</sup> Todor Pavlov,<sup>†</sup> Vanjo Grancharov,<sup>†</sup> Milen Todorov,<sup>†</sup> Patricia Schmieder,<sup>‡</sup> and Gilman Veith<sup>§</sup>

Laboratory of Mathematical Chemistry, University "Assen Zlatarov", 8010 Bourgas, Bulgaria,  
USEPA, ORD, NHEERL, Mid-Continent Ecology Division, 6201 Congdon Boulevard,  
Duluth, Minnesota 55804, and International QSAR Foundation to Reduce Animal Testing,  
1501 West Knife River Road, Two Harbors, Minnesota 55616

Received May 6, 2004

Mathematical chemistry has afforded a variety of research areas with important tools to understand and predict the behavior of chemicals without having to consider the complexities of three-dimensional conformations of molecules. Predictive toxicology, an area of increasing importance to toxicity assessments critical to molecular design and risk management, must be based on more explicit descriptions of structure, however. Minimum energy conformations are often used for convenience due, in part, to the difficulty of computing a representative population of conformers in all but rigid structures. Such simplifying assumptions fail to reveal the variance of the stereoelectronic nature of molecules as well as the misclassification of chemicals which initiate receptor-based toxicity pathways. Because these errors impact both the success in discovering new lead and the identification of possible hazards, it is important that mathematical chemistry develop additional tools for conformational analysis. This paper presents a new system for automated 2D-3D migration of chemicals in large databases with conformer multiplication. The main advantages of this system are its straightforward performance, reasonable execution time, simplicity and applicability to building large 3D chemical inventories. The module for conformer multiplication within the 2D-3D migration system is based on a new formulation of the genetic algorithm for computing populations of possible conformers. The performance of the automated 2D-3D migration system in building a centralized 3D database for all chemicals in commerce worldwide is discussed. The applicability of the 3D database in assessing the impact of molecular flexibility on identifying active conformers in QSAR analysis and assessing similarity between chemicals is illustrated.

### INTRODUCTION

One of the principle challenges in QSAR is the quantification of the chemical structure itself. Some important properties of chemicals can be computed using only molecular topology or two-dimensional (2D) representations of structure along with empirically determined substituent factors. However, the interactions of chemicals with biological macromolecules involve a broad spectrum of chemical reactivity for which stereoelectronic models of structure are necessary. In turn, the electronic attributes of a chemical structure can only be estimated from the precise geometry of the structure. Thus, molecular interactions are three-dimensional (3D) in nature and molecular models must treat chemicals as 3D entities. A prevailing convention resulting more from computational barriers than empirical evidence has been to use 3D chemical structures derived from a computed minimum energy. However, flexible molecules can exist as hundreds of different 3D geometrical conformations, and the electronic properties (hence reactivity) of the different conformations of a single 2D structure can vary substantially. Minimum energy calculations often fail to identify conformers that, while possibly slightly less stable, have the required shape

and electronic properties to interact with receptors. This holds especially for enzyme mediated reactions where enzyme induced distortions in direction to the transition state drive the molecules even out of the local potential energy minima.

Conformers of an individual chemical which have free energy within the approximately 20 kcal/mol from the lowest energy structure (usually accepted threshold) often exhibited significant variation in potentially relevant electronic descriptors. For example, conformers of  $\beta$ -zeaxanol had a range of 0.45 eV for  $E_{\text{LUMO}}$ , 0.19 eV for  $E_{\text{HOMO}}$ , 0.42 eV for  $E_{\text{HOMO-LUMO}}$ , and 3.89 D for Dipole moment. Moreover, it has been found that the lowest energy conformer of  $\beta$ -zeaxanol was not the active one with respect to binding to estrogen receptor (ER). The observation that relatively small energy differences between conformers can result in significant variations in electronic structure highlighted the necessity of including all energetically reasonable conformers when defining common reactivity patterns.<sup>1,2</sup>

The identification of the "most active" conformers from among hundreds of possible representations requires development of complex algorithms that are based on the physical reality provided that the selection of active conformers is dependent on the specific interaction under investigation. Conformational flexibility appears to be a significant structural feature especially when receptor-mediated toxic endpoints are modeled. Capabilities need to be developed to

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

† University "Assen Zlatarov".

‡ USEPA.

§ International QSAR Foundation to Reduce Animal Testing.

represent chemicals as a distribution of plausible conformations, quantify the molecular descriptors as a function of conformation, and examine whether a chemical is flexible enough to conform to an "induced fit" by the receptor itself.

Of course, this capability for modeling hundreds of conformers is equally important in the discovery of QSARs themselves, in the context of specific adverse effects. Both the structural requirements for a specific chemical interaction and the molecular descriptors that best measure the intensity of the interaction could be missed if the model development focused on the wrong conformations. As it was already shown, the COREPA method has the potential to overcome these problems.<sup>3–5</sup>

Recently we have developed two approaches for conformer generation which were quite different with respect to the algorithm used as well as their performance. The first approach, 3DGEN,<sup>6</sup> is based on a combinatorial procedure for a systematic search of conformational space. The systematic approach, however, was found to provide good performance for relative small and rigid structures. A new approach for coverage of the conformational space by a limited number of conformers<sup>7</sup> was developed (called the GAS algorithm) to handle highly flexible chemicals. Instead of using the systematic search whose time-complexity increases exponentially with degrees of freedom, a genetic algorithm (GA) was employed to minimize 3D similarity among the generated conformers. This makes the problem computationally feasible even for large, flexible molecules. The 3D similarity of a pair of conformers is assumed reciprocal to the root-mean-square (RMS) distance between identical atomic sites in an alignment providing its minimum. Thus, in contrast to traditional GA, the fitness of a conformer is not quantified individually but only in conjunction with the population it belongs to. The approach handles the following stereochemical and conformational degrees of freedom: rotation around acyclic single and double bonds, inversion of stereocenters, flip of free corners in saturated rings, reflection of pyramids on the junction of two or three saturated rings. The latter two were particularly introduced to encompass structural diversity of polycyclic structures. However, they generally affect valence angles and can be restricted up to a certain level of severity of such changes. Stereochemical modifications are totally/selectively disabled when the stereochemistry is exactly/partially specified on input. For the chemicals under study, the stereochemistry of the active enantiomer is maintained during conformer generation. The reproducibility and robustness of GA runs and subsequent density of coverage of conformational space can be controlled by the ratio between parents and children.

When strained conformers are obtained by any of the algorithms, the possible violations of imposed geometric constraints are corrected with a strain-relief procedure (pseudo molecular mechanics; PMM) based on a truncated force field energy-like function, where the electrostatic terms are omitted.<sup>6</sup> In fact, the PMM force field involves additive interatomic interactions for bond lengths, valence angles, dihedral angles, out-of-plane bends of  $sp^2$  conjugated sites and Lennard-Jones repulsions of nonbonded sites. The basic form and parametrization of the interatomic interactions mentioned above was taken from the Chem-X force field.<sup>8,9</sup> Geometry optimization is further completed by quantum-chemical methods. Usually, MOPAC 93<sup>10,11</sup> is employed by

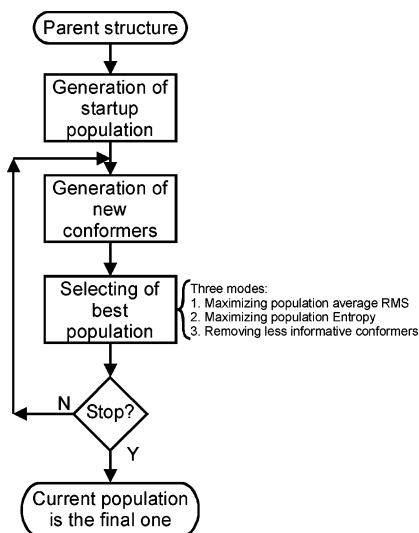
making use of the AM1 Hamiltonian. Next, the conformers are screened to eliminate those, whose heat of formation,  $D\Delta H_f^\circ$ , is greater from the  $\Delta H_f^\circ$  associated with the conformer with absolute energy minimum by user defined threshold. Usually, 20 kcal/mol (or 15 kcal/mol) threshold is employed based on experimental evidence that the free energy of binding to some steroid hormones is in the range of  $-10$  to  $-20$  kcal/mol,<sup>12,13</sup> which would provide the necessary energy to elevate conformers from the low(est) energy state during binding. Subsequently, conformational degeneracy, due to molecular symmetry and geometry convergence is detected within a user defined torsion angle resolution.

Few insufficiencies, however, have been encountered when applying both algorithms for conformer multiplication of large chemical inventories. First, the systematic algorithm was producing a huge number of conformers even for reasonably small structures; moreover, most of the generated conformers were found to be strained structures requiring significant time to get optimized. On the other hand the genetic algorithm was not applicable to rigid chemicals meeting difficulties to generate an initial population and trigger the evolution. In this respect, we decided to apply the systematic algorithm in cases where the combinatorial problem can be handled, in terms of reasonable timing of program performance. Alternatively, the GA was applied for flexible chemicals. Second, we have found that conformers produced by the original formulation of GA<sup>7</sup> do not meet our intuitive understanding for coverage of conformational space. Thus, often clusters of structurally close conformers were generated which were separated by large RMS distances (the larger number of small distances between conformers constituting the clusters was compensated by the large distances between clusters). That means that in case of straight chain alkenes we will see the extreme cases of most stretched and most bended (compact) conformers rather than the whole range of structurally different conformers. To fix this problem the fitness function based only on maximization of RMS distance between conformers (earlier formulation of the method) had to be combined with another function accounting also for distribution of conformers across conformational space. Next, we have realized that some of the structural variables used by the original formulation need to be generalized in order to expend the conformational diversity of generated structures. Finally, a new procedure has been developed for automated determination of the number of conformers needed for reasonably good coverage of conformational space. Examples are listed for the performance of the modified GA algorithm. The combination of systematic and refined GA algorithms has been used in a new automated system for 2D-3D migration with conformational multiplication of chemicals. The performance of this system for building large 3D chemical inventories is exemplified and discussed.

## METHODS

The core of the modified GAS algorithm is presented in the forthcoming along with the general scheme of its application.

**Genetic Algorithm for Conformer Multiplication. Basic Principles.** Genetic algorithms for creating new structures



**Figure 1.** The schematic presentation of the GAS core algorithm.

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 fitness criteria to form the next generation of chemical structures. The extension of a population, followed by its selective reduction to the size of 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 genes representing various structural features, 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 where minimal improvement over several iterations is seen.

The general scheme of GA for conformer multiplication is presented in Figure 1.

**Conformational Variables.** Five types of structural variables or changes in molecular conformation are used to represent the important characteristic encoded as a chemical “gene”, and the genes are then combined into the chemical “chromosome”. The earlier formulation of the algorithm<sup>7</sup> handles the following degrees of freedom: rotation around acyclic single and double bonds, inversion of stereocenters, flip of free corners in saturated rings, reflection of pyramids on the junction of two or three saturated rings. The new structural modification, called “*Flip of fragments*”, is described, here. This modification is a generalization of the flip of free corners analyzed in the original formulation. The flipped fragments are defined by pairs of atoms that if we remove from the molecule the latter will get split to at least two disconnected fragments that are incident to both atoms. Atoms A and B in Scheme 1 define such a general flip. They divide molecules to Fragment\_1 and Fragment\_2. If angles  $\alpha$  and  $\beta$  are equal we can just rotate Fragment\_2 around axis AB at angle  $2 \times \alpha$ . But because both angles can be different the flipping operation is more complex and defined as follows:

1. Generate the mirror image of Fragment\_2 with respect to the plane defined by A, B, F1a and F1b. In case these

points are not coplanar the middle point (F1M) of F1a and F1b is defined and a mirror image is produced with respect to the plane defined by A, B and F1M.

2. Restore the original image of Fragment\_2 by reflecting the fragment into itself (reflecting to the plane defined by A, B and F2M).

If atoms A and B define another one fragment (Fragment\_3) in addition to Fragment\_1 and Fragment\_2, the same operations applied to Fragment\_2 are applied to Fragment\_3 as well.

In case Fragment\_2 contains one atom only (i.e., F2a coincides with F2b) the flip of the fragment is transformed into a flip of free corner.

**Analysis of Conformer Populations. Cardinality.** The proposed 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. The suggested ratio between the theoretical number of conformers and cardinality of population of conformers which will be generated in an attempt to cover conformational space is listed in Table 1.

The genetic algorithm can be applied in two modes—using two different fitness criteria for evaluating new conformers. These evaluated criteria are based on properties of the entire population of conformers. In the first mode, the populations are selected that have the maximum RMS atomic distances which tends to prevent isolated clusters of similar conformers. The iterative process for 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 generation population can be characterized with an average RMS distance not smaller than 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 is trivial and 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.

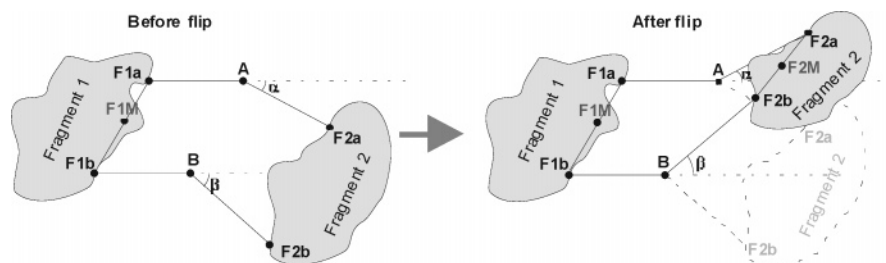
According to the second mode, the entropy of the population is maximized using the Shannon entropy function

$$S_j = - \sum_{\substack{j=1 \\ i \neq j}}^J P_{ij} \ln_2(P_{ij}) \quad (1)$$

where

$$P_{ij} = \frac{p_{ij}}{\sum_j p_{ij}} \quad (2)$$

Scheme 1



**Table 1.** Ratio between the Theoretical Number of Conformers and Cardinality of the Population Which Will Be Generated To Cover Conformational Space, Used by Default in the Present Formulation of the Approach

theoretical number	permanents	candidates
< 2E3	5	3
2E3–1E5	10	4
1E5–1E10	15	7
> 1E10	30	10

The entropy 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 next evolution step of GA) providing the best coverage of the space among all generated conformers (parents and children) of the current GA step:

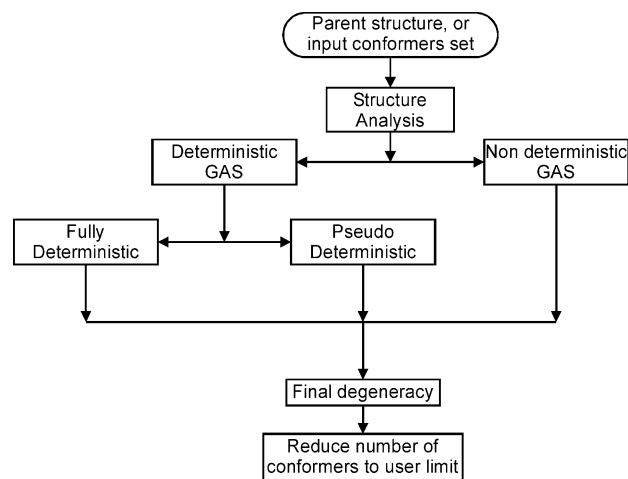
$$H = - \sum_i \sum_{j \neq i} P_{ij} \ln(P_{ij}) \quad (3)$$

To avoid the combinatorial problem for selection of the parent population an algorithm was developed to identify the most “informative” conformers among generated ones using eq 1. This algorithm could be considered as a simplified version of the second mode for evaluating populations of generated conformers.

Any of the generated conformers is 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. A threshold of 60° is appropriate but can be user defined. The energy of each conformation is then calculated using PMM and compared to a threshold value to eliminate highly strained structures. The quality filter is used to evaluate generated conformers in terms of a violation of common bond lengths, valence and torsion angles, nonbonded distances, etc. (details on the filtering process are given in subsequent sections).

**General Scheme for Conformer Multiplication.** The conformer generation procedure is preceded by quantum chemical optimization of the single 3D structure produced by 2D-3D conversion of input chemical. Our experience shows that this preliminary optimization improves the quality of generated conformers.

**Modes of Conformer Generation.** The general scheme for conformer generation is used to define the applicability domains of two modes for conformer generation—deterministic and nondeterministic (Figure 2). Due to its combinatorial nature, the first mode is applied in case of a less flexible structure. The specified threshold for this selection is based on a theoretical number of generated conformers. The



**Figure 2.** General scheme of automated conformer generation.

structure is analyzed, and if this number is less than a user defined threshold (2000 by default), then the system applies the deterministic mode, otherwise—the nondeterministic algorithm is turned on.

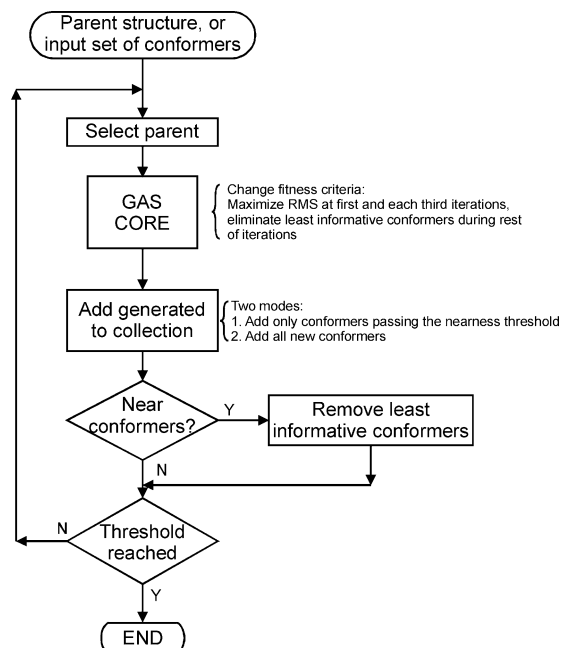
**Deterministic Conformer Generation.** The deterministic mode systematically explores all possible combinations of conformational variables aiming to produce exhaustively all possible conformers. The algorithm is based on the same toolbox of chromosomal expression of a structural variable as in a genetic algorithm, however, without using genetic operations to produce conformers.

Some of the chromosomes depend on the order they have been applied; e.g., application of one of them could make unfeasible the modification of other variables and vice versa. Such order dependent chromosomes are the flips of pyramids, free corners and fragments. Because of their order dependence, these conformational variables are applied in random order, and because of that the conformer multiplication is called pseudo deterministic. To reduce indeterminateness of the procedure the same chromosome operator is applied several times (5 by default) on the parent structure.

The true deterministic algorithm (not pseudo deterministic one) is applied in case of quite rigid structures. All combinations of orderings of conformational variables are applied on each chromosomal operator. Because of time complexity of this combinatorial algorithm the system selects this mode only when a chemical has up to 8 order dependent conformational variables and the sum of all conformation variables is not more than 8 either.

**Nondeterministic Conformer Generation.** This algorithm is applied for an iterative collecting of most different and equidistant conformers using the GAS core, as illustrated in Figure 3. The algorithm takes into account the fact that the number of conformers needed to cover conformational space





**Figure 3.** Schematic illustration of nondeterministic application of GAS algorithm.

is not known at the beginning. On the other hand the population of conformers generated by GA is determined rigidly (Table 1). Hence, the GA core is applied iteratively. The conformers generated at each GA run are combined with an already generated conformer population. The conformers with very small RMS (i.e., structurally not differing with other conformers) are removed from the collection. The *Neighborhood (nearness) threshold* is used to define structurally similar conformers. This threshold is determined as a function of the maximal RMS for the collection (user defined % of the maximal RMS).

**Nondeterministic Mode of Conformer Generation.** The following steps can be identified in a nondeterministic mode of conformer generation:

**Adding of Conformers to Population.** To expand the structural diversity of generated conformers, the GAS core is applied several times and produced conformer sets are combined. When a new set of conformers is generated by GAS core there are two methods of adding it to a previous conformer collection:

1. Add all new conformers and then start removing until all collected conformers meet the nearness threshold.
2. Add new conformers one by one if they meet the nearness threshold. There is a special case of this mode—if the new conformer enlarges maximum RMS for the collection it is added even if it does not meet the nearness threshold. If after adding a mode a new maximal RMS is obtained, the process of removing of the conformers is turned on until all collected records meet the nearness threshold requirement.

In the same program implementation, the GAS core is applied alternatively to maximize either RMS or the evenness of the coverage of conformation space (producing equidistant conformers). Presently, the maximization of RMS is performed at every third core call (could be changed optionally); the rest of the calls are used to collect most informative individuals in the population i.e., to maximize the coverage of conformation space. Due to the combinatorial

character of the GAS core used for maximizing RMS the cardinality of the population of parents is reduced twice to ease the computational task.

**Removing of Conformers.** The removal of conformers starts if there is at least one pair of conformers for which the RMS distance is less than the nearness threshold. Removed are the conformers with minimum individual entropy according to eq 1. The removal is continuous until all of the conformers meet the nearness threshold. The conformers providing the maximal RMS for the collection are not removed to preserve the structures with the highest steric difference (e.g., most stretched and most compact conformers).

**Ending Criteria.** Iterations with the GAS core are ended if one of the following thresholds is met (default values are listed although they could be modified optionally):

1. The increase of the maximal RMS for the collection is less than 5%. This threshold is enabled after first 20 iterations.
2. Less than 2 new conformers are added during the last 5 iterations.
3. More than 250 conformers are collected
4. More than 100 iterations are performed.

As defined only first one of the above criterion ends with a removal procedure to provide most even coverage of conformational space. To make all ending criteria consistent, the application of other criteria (2, 3 and 4) is also followed by the removal process.

After ending conformer generation (no matter which criterion is applied) some final operations are performed, including the following:

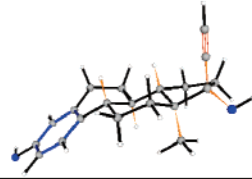
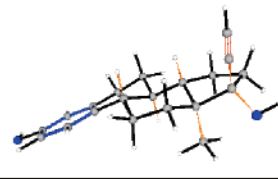
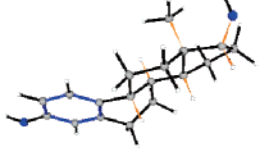
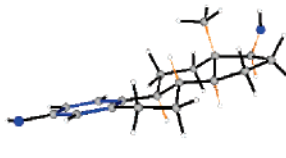
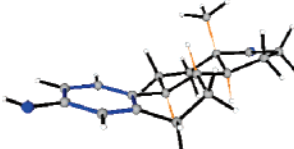

1. Removal of degenerated conformers. These are conformers not differing from the rest with respect to any of their torsion angles. The degeneracy threshold is 60° by default (user defined). If collected conformers are less than 100 the degeneracy threshold could be reduced — 30° by default.
2. Reducing the final population. It is applied if the population cardinality exceeds a user defined threshold (30 by default). The reduction is based on sequentially removing a less informative conformer till reaching the threshold. This reduction is imposed due to the time complexity of subsequent quantum chemical calculations and data storage problems emerging when large chemical inventories are conformationally multiplied.

To make conformer multiplication of large inventories feasible a time limit (60 min by default) is imposed for multiplication of a single chemical.

The nondeterministic scheme of conformer generation is applied to overcome one of the limitations of the genetic algorithm and namely the *ad hoc* determination of the number of conformers which will be used to cover the conformational space. Second, this scheme allows combining both criteria for selecting conformers based on their dissimilarity and ability to cover evenly the conformational space.

The detailed description of the “Converter” module of the OASIS program suites for automated 2D-3D migration of chemicals with their conformer multiplication is available as Supporting Information.

**Table 2.** Steroid Chemicals for Which the Deterministic Mode of the Conformer Generation Algorithm Is Applied

NAME, CAS	Conformer Variable	Theoretical Number of Conformers	Conformer 1	Conformer 2
Ethynylestradiol CAS 57-63-6	6	96		
17 $\beta$ -Estradiol CAS 50-28-2	5	32		
Estrone CAS 53-16-7	5	32		

## RESULTS AND DISCUSSION

The deterministic mode of application of the algorithm for conformer generation is illustrated in Table 2.

One can judge for the performance of the algorithm with conformers generated for steroidal structures (references for rigid chemicals), as listed in Table 2. For three steroidal structures (retaining the stereochemistry of the natural enantiomer) we have presented the number of conformer variables, the theoretical number of conformers and structures of the ultimately obtained conformers after quantum-chemical optimization and elimination of degenerate structures. As seen, besides the crystal-phase structure with semichair and chair conformations for B and C rings, respectively (conformer II) a new conformer is obtained with boat and chair configurations for the same rings, respectively (conformer I). Steroid conformer interconversions were assessed as kinetically and thermodynamically feasible.<sup>14</sup>

The next improvement of the present GAS algorithm, in terms of the coverage of conformational space, is illustrated in Figure 4. Here, the generated conformers for *n*-decane by the original version of the algorithm (user defined number is generated) are compared with conformers generated by the new formulation of algorithm. As seen, the conformers produced by an earlier version of the system tend to produce two clusters of structures similar to both extremes—most stretched (linear) and most bended (compact) conformers. In opposite, the structures generated by a modified algorithm provide coverage of conformational space corresponding to our intuitive vision—populating the area between both extreme conformers. The entropy fitness function as defined by equations (1)–(3) seems to quantify reasonably well the intuitive vision for uniform coverage of conformational space. As a limitation of this measure for assessing the conformer population evenness one could point out the underlying metric for defining the distance between two conformers. Thus, two conformers of a chemical could be very similar according to the intuitive vision and still the

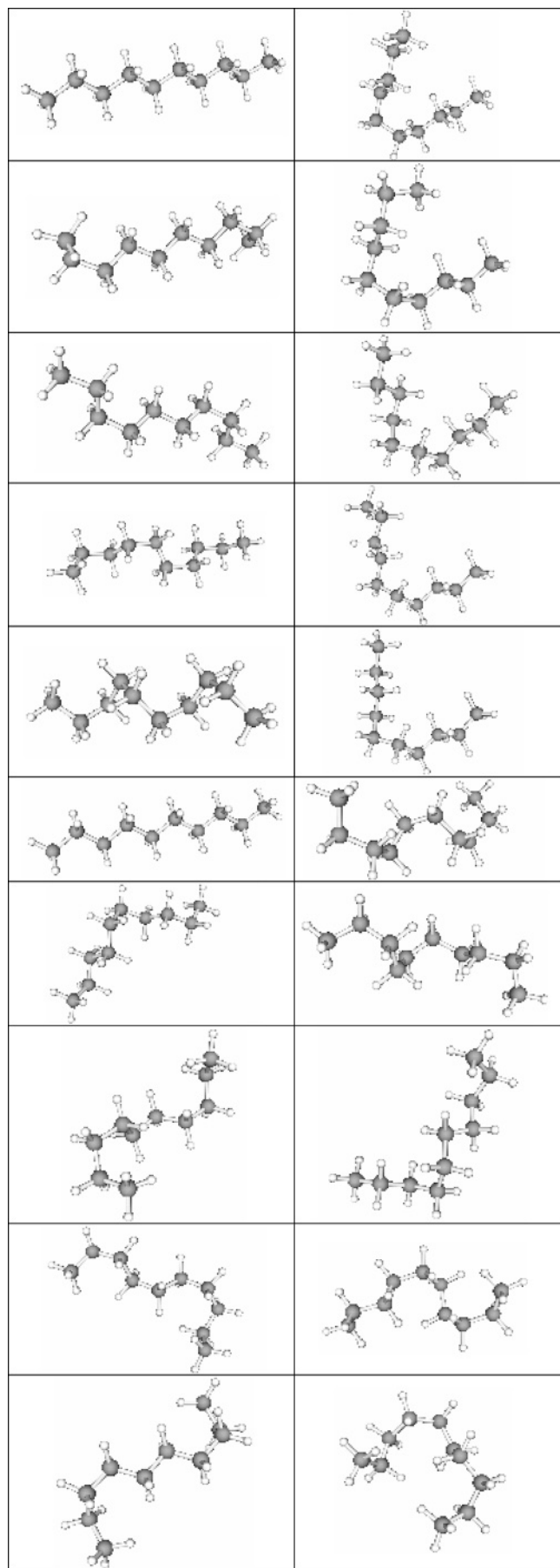
RMS value assessing the structural dissimilarity (distance) between them be high (e.g., obtained as a sum of small increments). At this point, however, no better replacement of this distance measure is available.

The impact of a newly introduced conformational variable on the structural diversity of generated conformers could be best illustrated for saturated cyclic structures. For the structure illustrated in Figure 5, the number of conformers and their maximal RMS are shown for original and new formulations of the method.

As seen, the use of the new structural modifications allowed describing more exhaustively molecular flexibility. The total number of generated conformers for selected saturated cyclic chemicals increased significantly—from 37 to 81 conformers for the five chemicals under investigation (Figure 5). Another indication for a better description of molecular flexibility is the variation of the maximum RMS values (which is used as measure for the structural difference between conformers). As seen, higher values are reached when the fragment flips are used as a conformational variable for studied chemicals.

The introduction of the new conformational variable affects also the range of variation of molecular descriptors across generated conformers. Thus, the ranges of variation of  $E_{\text{HOMO}}$  and  $E_{\text{LUMO}}$  for chemical #2 (Figure 5) are from  $-9.23$  to  $-9.18$  eV and from  $0.45$  to  $0.49$  eV, respectively, when the new variable is not included (Figure 6a). These ranges significantly increased when the new structural modification is turned on: from  $-9.23$  to  $-9.08$  eV and from  $0.43$  to  $0.60$  eV, respectively (Figure 6b). This apparently could be related to the missing conformers in the original formulation and uneven coverage of the conformational space due to the lacking conformational variable.

To give some assessment for time complexity of the 2D-3D migration system, we decided to handle 100 randomly selected structures from the Centralized 3D database for existing chemicals (see next subsection), with a size between 10 and 15 atoms. The chemicals were 2D-3D migrated 10



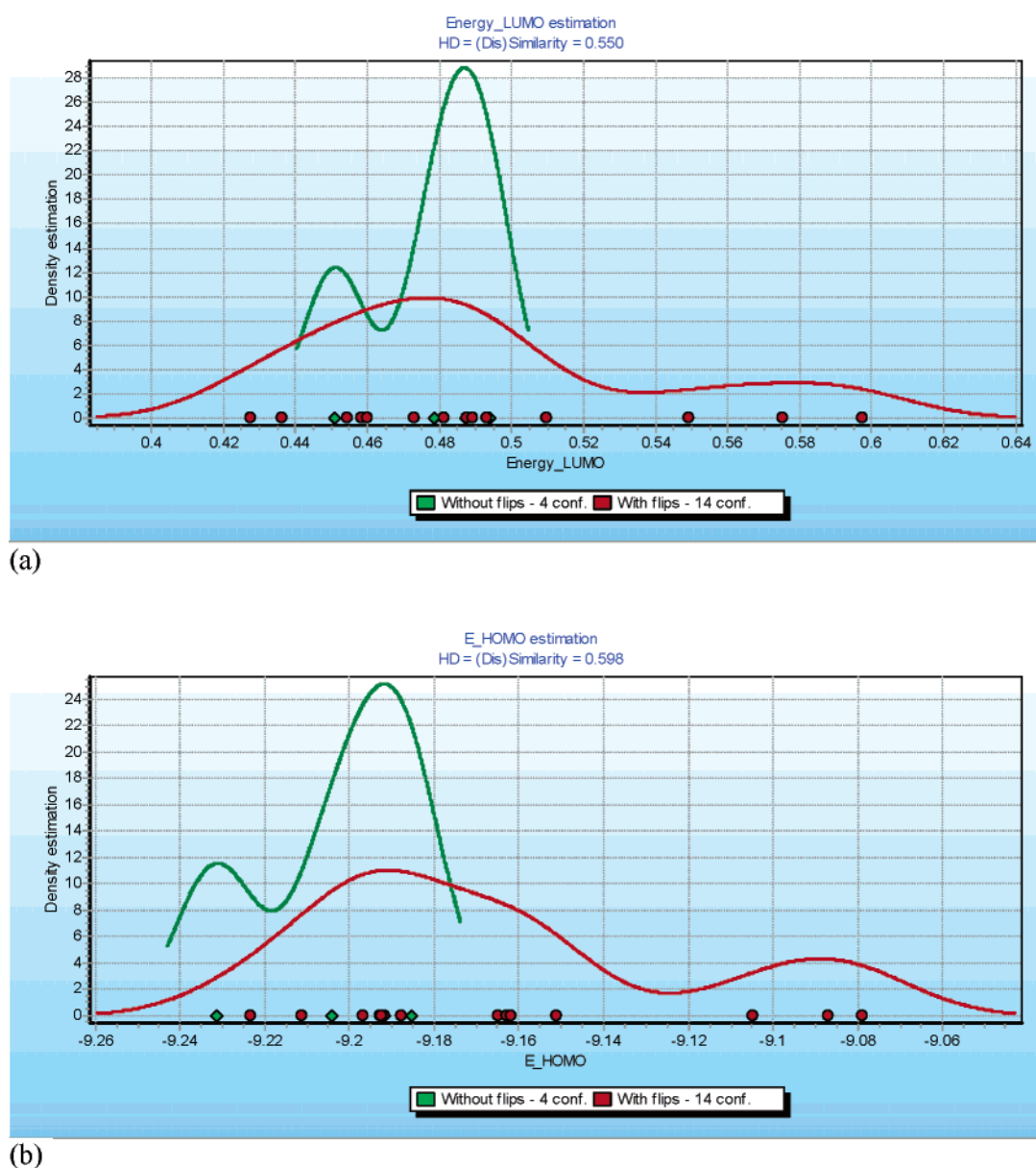
**Figure 4.** Application of the original (a) and new (b) formulations of the GAS algorithm for conformer multiplication of *n*-decane. User defined number of conformers (10) are produced.

#	Structure	No fragment flips	With fragment flips
1		conformers 6 max RMS 0.697	conformers 10 max RMS 2.238
2		conformers 4 max RMS 0.393	conformers 14 max RMS 2.706
3		conformers 2 max RMS 0.015	conformers 12 max RMS 0.384
4		conformers 21 max RMS 1.947	conformers 30 max RMS 2.392
5		conformers 5 max RMS 0.713	conformers 15 max RMS 3.098

**Figure 5.** The number of conformers and their maximal RMS as produced by the original and new formulation of the genetic algorithm when the flip of fragment is included as a conformational variable.

times thus simulating artificially 1000 chemicals. The calculations have been performed on a single PC P4 at 2GHz. The number of generated conformers was confined up to 30. The system generated in total 8164 conformers in total for 255 min; i.e., ~3 min per chemical (which included also quantum-chemical optimization of conformers at PRECISE mode).

**Applicability of the 2D-3D Conversion Module for Building 3D Databases.** The performance of the algorithm for conformer generation was evaluated within the global system for 2D-3D migration. The software implementation of this system was used for an automated 2D-3D migration of the chemicals in Centralized 3D Database of Existing Chemicals. The individual databases of regulatory agencies in North America and Europe, including IUCLID of European Chemicals Bureau (with 61428 chemicals), Danish EPA (159445 chemicals), TSCA (56884 chemicals), HPVC (8571 chemicals) and pesticides active/inactive ingredients of US EPA (3297), and DSL of Environment Canada (11441 chemicals) and Japanese METI (14289) were combined in CD-EC. The total number of unique chemicals from all these databases exceeded 183 000. These chemicals were submitted to conformational multiplication and quantum-chemical evaluation using the above-described system for 2D-3D migration. To overcome time complexity of the calculations we have applied so-called distributed computing. All PCs of the Laboratory of Mathematical Chemistry (about 25 of the class P4 at 2GHz—in average) are working in an intranet—connected with a server where the database is located. Computers are taking not-treated chemicals from the database and working on them all the time they are not occupied by other tasks; because of that this regime of applying the 2D-3D migration procedure was named “idle”. Anytime one can see the performance of all computers in a graph (how many structures they have calculated; which



**Figure 6.** Conformer density population across frontier orbitals  $E_{\text{LUMO}}$  (a) and  $E_{\text{HOMO}}$  (b) for chemical #2, in Figure 5. Density population for each parameter is shown for the original formulation of the algorithm and a new version with a fragment flip accounted for as a conformational variable.

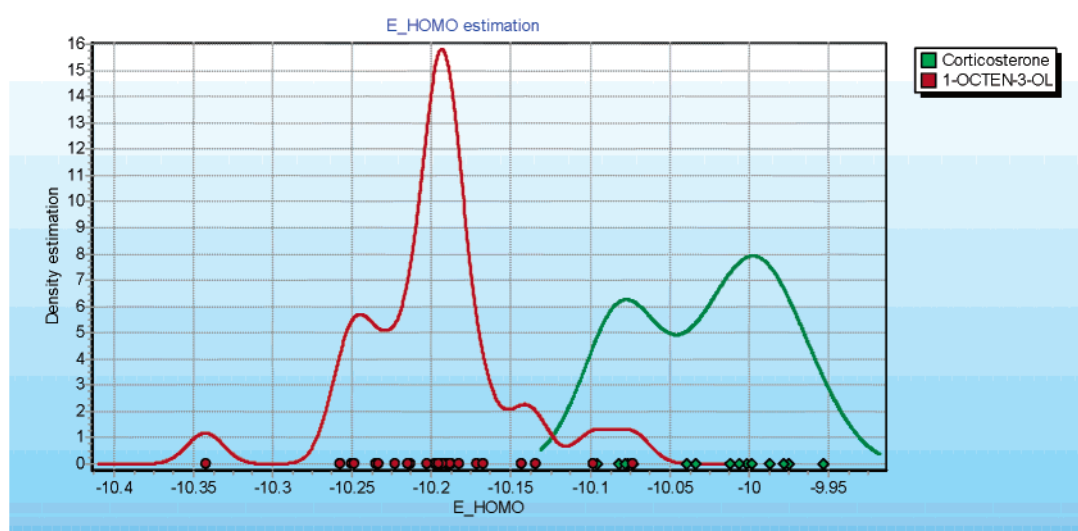
structure is presently treated, etc.). The distributed computing (idle regime) accelerates significantly the computing job. The enhanced computation effectiveness allowed the task of 2D-3D migration, conformational multiplication and quantum chemical assessment (in PRECISE mode) of all existing chemicals to be accomplished in several months.

Once chemicals are conformationally multiplied and quantum chemically evaluated, one can infer a hypothesis for structural conditioning of the studied endpoints. For that purpose the Centralized database provides a flexible searching where one could logically combine 2D and 3D structural queries.

**Selection of Active Conformers for QSAR Analysis.** The algorithm for conformer multiplication allows selection of active conformers for QSAR analysis. As it was already mentioned the steric and electronic properties vary significantly across the conformers of one chemical. This is exemplified in Figure 6. Hence, it is obvious that the selection

of active conformers of chemicals for QSAR analysis is as important as the selection of suitable molecular descriptors. Existing 2D-3D converting packages generate different conformers for same the chemical—one of the low energy minima. As a result, there is no reproducibility of the QSAR models applying the same molecular descriptors if they are produced by different 2D-3D converting software. Subsequently, one needs to perform conformational analysis and select active conformers for QSAR analysis. The algorithm proposed in the present work provides a straightforward scheme for conformer generation. Further, generated 3D representatives of chemicals could be used for selecting active conformers according to the approach of Mekenyan et al.<sup>15</sup> There are a variety of selection schemes, which can be applied after examining the conformer distributions. Thus, one could select conformers providing the extreme (maximum or minimum) values of the parameters. For example, one can select most electrophilic conformers taking those





**Figure 7.** The of probabilistic conformer distribution for chemicals corticosterone (CAS 50226) and 1-octen-3-ol (CAS 3391864) across  $E_{\text{HOMO}}$ . The similarity between these chemicals with respect to  $E_{\text{HOMO}}$  was assessed by the Hellinger distance of 1.67 (maximum value for dissimilarity is 2.00).

having minimum values of  $E_{\text{LUMO}}$ ; alternatively, most nucleophilic conformers could be selected taking those having maximum values of  $E_{\text{HOMO}}$ ; one could hypothesize that most stable conformers are active ones taking those having minimum of calculated heats of formation ( $\Delta H_f$ ), etc. For each selection, conformers are considered as distinct structural representatives of the chemical, and each of them is associated with the dependent variable (endpoint under investigation) for the ‘parent’ two-dimensional compound. The identification of proposed ‘active’ conformers is based on an evaluation of the regression statistics associated with QSARs derived using the different conformation samples.

The dynamic QSAR technique for selection of active conformers was successfully employed for modeling toxicity of unsaturated alcohols,<sup>15</sup> semicarbazides,<sup>16</sup> and  $\alpha$ -terthienyls;<sup>17</sup> it was also used for developing QSARs for aryl hydrocarbon receptor (AhR)<sup>18</sup> and estrogen receptor (ER)<sup>19</sup> activity of congeneric chemicals.

**Assessing Similarity between Chemicals Accounting for Their Conformational Flexibility.** According to conventional methods, the similarity between two chemicals is assessed univocally by comparing their single 3D structural representatives. The algorithms for that purpose are based either on aligned of both structures or assessing differences in numerical values of structural descriptors. The situation, however, is getting complicated when chemicals are compared accounting for their conformational flexibility. For that purpose one could use the proposed algorithm for 2D-3D migration. In this case each chemical is represented by a set of conformers. This is illustrated in Figure 7 where the discrete distributions of two chemicals across  $E_{\text{HOMO}}$  parameter axis are presented.

For the global molecular descriptors, each conformer is represented by a single point value of the parameter. For atomic parameters, several descriptor values associated with various local sites (atoms) of the conformer are allocated across the parameter axis.

As seen from Figure 7, both chemicals are very similar in terms of their  $E_{\text{HOMO}}$ , according to some of the conformers having not significantly differing values for this parameter,

whereas the same two chemicals are highly dissimilar according to the  $E_{\text{HOMO}}$  values of other conformers (e.g., see outmost left and outmost right conformers).

To evaluate similarity between chemicals with respect to a molecular parameter we have developed a method<sup>3–5</sup> where conformer distributions of chemicals across a specified parameter are compared rather than the very structures. Thus, the alignment problem is avoided and conformer flexibility is accounted for.

The conformer distributions of chemicals are created as probability distributions. For each value of descriptor  $x$  a kernel density function<sup>20</sup> is superimposed on each individual data point, and these data density kernels are summed and normalized to give an overall probability distribution,  $p(x)$  (see Figures 6 and 7).

The similarity between chemicals can be quantified using the distance between probability distributions, such as Hellinger the Kullback-Leibler divergence, etc.<sup>20–24</sup> For example, the Hellinger distance between two probabilistic distributions across parameter  $x$  is calculated with equation 10:

$$HD_{1,2}^2 = HD(p_1(x), p_2(x)) = \int (\sqrt{p_1(x)} + \sqrt{p_2(x)})^2 dx \quad (4)$$

The minimum value of the Hellinger distance is zero, and it is reached when two probability density functions are the same. The maximum value of the Hellinger distance is two, and it is reached when two probability density functions are most distinct. Higher Hellinger distance values mean a high dissimilarity between chemical probability distributions.

The of probabilistic conformer distribution for chemicals corticosterone (CAS 50226) and 1-octen-3-ol (CAS 3391864) across  $E_{\text{HOMO}}$  are shown in Figure 7. The overlap between distributions leads to a similarity with respect to this parameter of 1.67 assessed as a Hellinger distance.

For more details on assessing similarity between chemicals with accounting for conformer flexibility one could see Mekenyan et al.<sup>5</sup> The described approach to assessing similarity has been used to derive the COREPA probabilistic scheme for classification of chemicals.

## SUMMARY AND CONCLUSIONS

A system for automated 2D-3D migration of chemicals in large databases with conformer multiplication is proposed. The main advantage of this system is its straightforward performance, reasonable time, simplicity and applicability for building large 3D chemical inventories. The module for conformer multiplication of the 2D-3D migration system is based on a new formulation of the genetic algorithm for generating a final number of structures to represent the conformational space of chemicals. A new fitness function based on Shannon entropy formulas is introduced in the algorithm to improve the coverage of the conformational space of the molecules. This function is applied in a combination with earlier goodness criterion maximizing RMS distances between generated conformers, only. The joint application of both functions provides a better coverage of conformational space by generated structures. Next, a new conformer variable is introduced to better reflect the flexibility of saturated polycyclic structures. Finally, the algorithm automatically determines the number of conformers needed for reasonably good coverage of conformational space based on number of flexibility of chemicals.

Another novelty of the module for conformational multiplication of chemicals is its deterministic application for relatively rigid molecules. The combinatorial domain of the deterministic applications of conformer multiplication algorithm is defined automatically according to the number of conformational variables of chemicals and time constraints.

The automated 2D-3D migration system was used for building a centralized 3D database with all existing chemicals across regulatory agencies. The application of the molecular flexibility concept for identifying active conformers in QSAR analysis and assessing similarity between chemicals is illustrated.

## ACKNOWLEDGMENT

This work was supported by USA EPA (CR-83199501-0) and by the European Union (project ALARM, GOCE-CT-2003-50667).

**Supporting Information Available:** Detailed description of the "Converter" module of the OASIS program suites for automated 2D-3D migration of chemicals with their conformer multiplication. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- 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.
- 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.
- Mekenyan, O. G.; Ivanov, J.; Karabunarliev, S.; Bradbury, S.; Ankley, G.; Karcher, W. A Computationally-Based Hazard Identification Algorithm That Incorporates Ligand Flexibility. *Environ. Sci. Technol.* **1997**, *31*, 3702–3711.
- Mekenyan, O. G.; Nikolova, N.; Karabunarliev, S.; Bradbury, S.; Ankley, G.; Hansen, B. New Developments in a Hazard Identification Algorithm For Hormone Receptor Ligands. *Quant. Struct.-Act. Relat.* **1999**, *18*, 139–153.
- 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.
- 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.
- Mekenyan, O. G.; Dimitrov, D.; Nikolova, N.; Karabunarliev, S. Conformational Coverage by a Genetic Algorithm. *Chem. Inf. Comput. Sci.* **1999**, *39/6*, 997–1016.
- White, D. N. J. Molecular Mechanics Calculations. *Spec. Rep. Chem. Soc.* **1987**, *6*, 38–63.
- Davies, E. K.; Murrall, N. W. How Accurate a Force Field Need Be? *Comput. Chem.* **1989**, *13* (2), 149–156.
- Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.
- Stewart, J. J. P. 1993. MOPAC 93. Fujitsu Limited, 9–3, Nakase 1-Chome, Mihama-ku, Chiba-city, Chiba 261, Japan, and Stewart Computational Chemistry. 15210 Paddington Circle, Colorado Springs, CO 80921, USA.
- Anstead, G. M.; Carlson, K. E.; Katzenellenbogen, J. A. The estradiol pharmacophore: ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* **1997**, *62*, 268–303.
- Wiese, T.; Brooks, S. C. Molecular modeling of steroidal estrogens: novel conformations and their role in biological activity. *J. Steroid. Biochem. Mol. Biol.* **1994**, *50*, 61–72.
- Ivanov, J. M.; Mekenyan, O. G.; Bradbury, S. P.; Schuurmann, G. A Kinetic Analysis of the Conformational Flexibility of Steroids. *Quant. Struct.-Act. Relat.* **1998**, *17*, 437–449.
- Mekenyan, O. G.; Ivanov, J. M.; Veith, G. D.; Bradbury, S. P. DYNAMIC QSAR: A New Search For Active Conformations and Significant Stereoelectronic Indices. *Quant. Struct.-Act. Relat.* **1994**, *13*, 302–307.
- Mekenyan, O. G.; Schultz, T. W.; Veith, G. D.; Kamenska, V. B. "Dynamic" QSAR for semicarbazide-induced mortality in frog embryo. *J. Appl. Toxicol.* **1996**, *16*, 355–363.
- Veith, G. D.; Mekenyan, O. G.; Ankley, G. T.; Call, D. J. QSAR evaluation of  $\alpha$ -terthienyl phototoxicity. *Environ. Sci. Technol.* **1995**, *29*, 1267–1272.
- Mekenyan, O. G.; Veith, G. D.; Call, D. J.; Ankley, G. T. A QSAR evaluation of Ah receptor binding of halogenated aromatic xenobiotics. *Environ. Health Perspect.* **1996**, *104*, 1302–1309.
- Bradbury, S. P.; Mekenyan, O. G.; Ankley, G. T. Quantitative Structure-Activity Relationships For Polychlorinated Hydroxy-biphenyl Estrogen Receptor Binding Affinity: An Assessment of Conformer Flexibility. *Environ. Chem. Toxicol.* **1996**, *15*, 1945–1954.
- Silverman, B. W. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall: **1986**.
- Devroye, L.; Györfi, L.; Lugosi, G. A probabilistic Theory of Pattern Recognition, Springer: **1996**.
- Haykin, S.; Neural Networks: A Comprehensive Foundation, Prentice Hall: **1998**.
- McLachlan, G. Discriminant Analysis and Statistical Pattern Recognition, John Wiley and Sons: **1992**.
- Scott, D. W. Density Estimation, Rice University, Houston, TX (<http://rice.edu>).

CI0498463