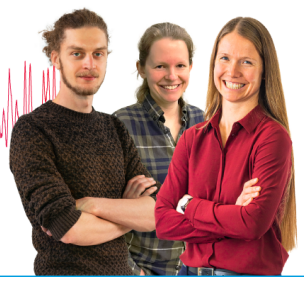
RESEARCH ARTICLE | NOVEMBER 02 2011

Optimization of a genetic algorithm for searching molecular conformer space

Zoe E. Brain; Matthew A. Addicoat

J. Chem. Phys. 135, 174106 (2011)
https://doi.org/10.1063/1.3656323



Optimization of a genetic algorithm for searching molecular conformer space

Zoe E. Brain[1] and Matthew A. Addicoat[2,a)]
[1]*Department of Computer Science, Australian National University, Canberra, ACT 0200, Australia*
[2]*Research School of Chemistry and Department of Computer Science, Australian National University, Canberra, ACT 0200, Australia*

(Received 25 August 2011; accepted 7 October 2011; published online 2 November 2011)

We present two sets of tunings that are broadly applicable to conformer searches of isolated molecules using a genetic algorithm (GA). In order to find the most efficient tunings for the GA, a second GA – a meta-genetic algorithm – was used to tune the first genetic algorithm to reliably find the already known *a priori* correct answer with minimum computational resources. It is shown that these tunings are appropriate for a variety of molecules with different characteristics, and most importantly that the tunings are independent of the underlying model chemistry but that the tunings for rigid and relaxed surfaces differ slightly. It is shown that for the problem of molecular conformational search, the most efficient GA actually reduces to an evolutionary algorithm. © *2011 American Institute of Physics*. [doi:10.1063/1.3656323]

I. INTRODUCTION

The always increasing speed of computers means that it is now possible to compute the electronic structure of molecules comprising up to a few dozen atoms using chemically accurate methods such as density functional theory (DFT) or MP2. However, an additional complication arises from the fact that such molecules readily flex, twist and bend producing a multitude of stable and meta-stable shapes, or conformers. The stability of these conformers may vary by tens or even hundreds of kJ mol^{-1}, so it is important to identify the global minimum energy conformer for subsequent calculation of thermochemical properties, which are generally desired with an accuracy of 1 kJ mol^{-1}. Furthermore, in many applications, the 3D shape of the molecule has a direct impact on the reactivity of the molecule with a given substrate.[1,2] This effect is particularly important in biological systems, including pharmaceutical design. Therefore, the first step in most computational investigations is necessarily to identify the lowest energy conformer.

A full conformational search of a molecule, if done strictly, requires the calculation of potential energy surfaces corresponding to simultaneous rotation about every bond in the molecule. Clearly, this is infeasible for all but the very smallest of molecules. In practice, bonds are rotated at a given resolution and the resultant structures are allowed to relax. Provided the resolution is sufficient, such an approach is unlikely to miss any stable structures and the lowest energy structure can be easily selected for further study. However, even with such practical approximations, the problem rapidly becomes infeasible to search exhaustively with increasing system size. For example; a molecule with 8 rotatable bonds, each mapped out at 120° resolution, yields $3^8 = 6561$ starting structures. Doubling the length of the molecule, however, yields $3^{16} = 43\,046\,721$ starting structures. Generally the problem scales as $(360^\circ/R)^N$, where R is the resolution and N is the number of rotatable bonds in the molecule.

The "combinatorial explosion" inherent in this problem is subject to very few constraints. Depending on the molecular connectivity, the constraint that two atoms should not be coincident or near so, will render some starting guesses infeasible. Similarly, the constraint that the molecular connectivity should remain constant (or else you have a different molecule), will also render a number of structures infeasible. This is quite unlike a problem such as molecular docking, where the notion that the two molecules must fit together, provides some constraint on the conformers of both molecules.

Several approaches to this combinatorial explosion problem inherent in locating the lowest energy conformer of a given molecule have been described in the literature over the last two decades and several excellent reviews and comparative studies[3–5] have been written. Approaches fall broadly into three classes; stochastic, systematic and deterministic methods, with a further distinction between methods that aim only to find the global minimum or some subset of low energy minima and methods that aim to find all minima. Stochastic methods are arguably the most popular method used for molecules of chemical interest as the size of such molecules is generally such that other methods are too computationally expensive to be routinely used. These methods, employ a "random walk" approach to exploring conformational space,[6,7] although they may optionally use *a priori* / known information about stable conformers to bias sampling. Examples include methods such as simulated annealing,[8] and Monte Carlo and molecular dynamics based methods.[9–12] The "Ant" algorithm, more commonly applied to the travelling salesman problem (TSP), has also been adapted to conformational search.[13] In addition, some "hybrid" methods, combining two or more methods, have been proposed and employed.[14]

a)Present address: Department of Chemistry, Nagoya University, Furo-cho, Chikusa-ku Nagoya 464-4602, Japan. Electronic mail: madd@rsc.anu.edu.au.

Systematic methods may either search the entire conformational space at some given resolution [such as TORK (Ref. 15) or SUMM (Refs. 16 and 17)] or attempt to reduce the conformational space in a deterministic manner. The most common approach is to divide the molecule into small fragments and determine their optimal structures independently. The whole molecule is then reassembled from these fragments with little, if any, consideration to the relative conformation of each fragment nor the possibility of fragments altering the optimum conformation of nearby fragments. This principle, termed "build-up" is the basis of such methods as A*,[18] sparse-matrix drive[19] and energy directed tree search.[20] An alternate systematic method, LMOD,[21] searches for all low-energy minima by following low-frequency vibrational modes, A third approach, so called "deterministic" methods use the techniques of interval arithmetic and various bounding algorithms in order to guarantee the global minimum is found.[22–24]

Genetic algorithms (GAs) are commonly considered to be an "intelligent" stochastic method as some degree of "learning" occurs through the repeated selection of the fittest individuals, analogous to the process of Darwinian evolution. GAs have been applied to various chemical problems including geometries of transition metal clusters,[25] geometries of molecular clusters,[26] ligand docking,[27] and molecular design.[28] A genetic algorithm begins with a randomly generated set (population) of *genomes*, each of which has an associated fitness score which is evaluated by some *fitness function*. The population of the next generation is generated by applying biological analog genetic operators such as random mutation and crossover (i.e., the synthesis of a new genome by matching complementary parts of two or more genomes). Individuals with a higher fitness score are given a higher chance of reproducing - i.e., passing on all or part of their genes to the next generation. This implicit retention of parts of the genome that contribute to high fitness, may be considered as a kind of "build-up." New generations are created until some predefined stopping criteria are met.

Applying this method to a conformation search, the first step is to represent the problem. Here, each genome is a string of length N, where N is the number of rotatable bonds in the molecule and each value corresponds to the torsional angle for the given bond. Given the $(360^\circ/R)^N$ scaling of the search space, encoding the torsional angle directly (i.e., at 1° resolution, is infeasible. However, provided one starts with an optimised structure, it is a reasonable approximation to consider sp^3 bonds at 120° resolution (i.e., $R = 3$) and sp^2 bonds at 180° resolution (i.e., $R = 2$). In the case that a given torsional angle was expected to be problematic, specification of $R = 6$ (i.e., resolution of 60°) or even higher would almost guarantee no minima were missed. The fitness function is simply an *ab initio* calculation of the conformer energy.

Having encoded the conformational search problem in such a way that it is amenable to solution by a genetic algorithm, a further problem arises; namely, what is the optimum GA to use? Population size, mutation rate and crossover probability are all freely variable parameters that may strongly affect the efficiency and reliability of the GA for the problem at hand. If one further considers effects of different forms of crossover, parental selection, mutation and measures such as elitism, the problem of optimising a GA quickly becomes strongly multidimensional. Pretsch and Brodmeier[29] examine the effects of population size, scaling function, mutation and crossover rates on a conformer search, however, they only consider a single implementation of each (population size = 30, exponential scaling, randomizing mutation, one-point crossover) and they do they specify optimum parameters.

One approach to this second problem, of optimizing the tunings of a GA is to use a Meta-Genetic Algorithm - a second GA which is used to optimise the parameters of the first GA. In this case the genome of the meta-GA is a string that encodes all the parameters of the GA that one wishes to optimize. For example, [10,0.2,0.5] could be read as a GA with a population size of 10, a mutation rate of 0.2 and a crossover probability of 0.5. The fitness function for the meta-GA is simply the number of genomes evaluated before finding the *a priori* known global minimum conformer. As the meta-GA approach necessarily involves solving the target problem many hundreds, if not thousands of times, the first step is to choose the largest possible example problem that is soluble by exhaustive search. The solution to this problem may then be used as a lookup table for the many GAs created by the meta-GA.

The meta-GA approach to GA optimization has been proposed previously;[30–32] however, most research has focussed on optimizing only one or two parameters and holding all others constant. This somewhat limited approach has been surprisingly successful across a variety of problems;[33,34] however, it seems intuitive that the largest degree of optimization can be achieved by optimizing all parameters simultaneously.

In this paper, we consider the problem of determining optimum parameters for a GA to locate the global optimum conformer of an isolated flexible molecule. The Hartree-Fock (HF) method paired with a small basis set is used to exhaustively search both rigid and relaxed surfaces for a series of molecules. The results are compared to previously published rigid surface computed using DFT. The good agreement between these results indicates that the derived GA tunings may be applied universally to single-molecule conformer searches, regardless of the underlying chemistry.

II. COMPUTATIONAL DETAILS

A. Determination of optimum GA parameters

The GA used for the conformer search problem is defined by seven parameters: population size, crossover operator, mutator operator, selection operator, crossover probability, mutation rate, and elitism. Population size refers to the number of individuals (conformers) present in each generation of the GA; therefore, the total number of conformers evaluated is equal to the population size × number of generations. The crossover operator (or method) defines the manipulation of the two parent genomes (selected by the selection operator) that occurs to produce a child genome and the crossover probability represents the chance that the given operator will be called. Similarly, the mutator operator defines the method used to mutate an individual gene and the

TABLE I. Genome defnition of the Meta-GA.

Parameter	Range	
Population size	5–1000	1–1000 with a floor of 5
Uniform Crossover		Exclusive with 1-pt,
Probability Density	1–1000	2-pt and None
One-Point Crossover		
Probability Density	1–1000	
Two-Point Crossover		
Probability Density	1–1000	
No Crossover		
Probability Density	1–1000	
Integer Range Mutator Probability Density	1–1000	
Integer Gaussian Mutator Probability Density	1–1000	
Swap Mutator Probability Density	1–1000	
Roulette Selector		Exclusive with
Probability Density	1–1000	Tournament, Uniform and Rank
Tournament Selector		
Probability Density	1–1000	
Uniform Selector		
Probability Density	1–1000	
Rank Selector		
Probability Density	1–1000	
Crossover		Divide by 10
Probability	1–1000	to yield %
Mutation Rate	1–1000	
Elitism?	1–1000	T/F

mutation rate represents the probability that such a mutation will occur. Elitism refers to always keeping the most fit individual as a member of the current population. The procedure used to determine the optimum GA parameters is the same as that described previously[35,36] and so is only briefly recapped here. Both the meta-GA and GA employ the PYEVOLVE (Ref. 38) genetic algorithm framework. The genome for the meta-GA was implemented as a set of integers between 1 and 1000 as shown in Table I. Once initial results indicated that population size and mutation rate were the most significant factors in determining GA efficiency, these were set at their optimized values and the remaining parameters re-optimized. A further re-optimization set the selection method.

The meta-GA itself, employed largely the default PyEvolve parameters (viz., parent selector: rank; tournament size = 2; mutation rate = 0.02; population size = 80; crossover: 1 Point; crossover rate = 0.5), with no optimization attempted. The number of generations for the meta-GA was set to 100 and convergence was confirmed by inspection. Both the meta-GA and the GA employed elitism.

For each combination of molecule and level of theory, the meta-GA was run 100 times, with randomly generated initial conditions, to generate 100 GA parameter sets. Each parameter set was then used 100 times to determine its efficiency, measured as the mean number of evaluations (i.e., number of generations × population size) and reliability (number of instances that located the minimum). In these calculations, each GA was terminated if it exceeded the number of evaluations equivalent to an exhaustive search.

B. Restart mechanisms

The five molecules of the first test set were chosen to be of a size that each desired surface (viz., UB3LYP, UHF, UHF relaxed, UHF solvated) could be searched exhaustively, meaning that the global minimum was known *a priori*. In the case of a larger molecule, it is not possible to know in advance the target conformer and consequently it is not possible to know in advance whether or not the GA has converged to the correct conformer or stalled in a local minimum. While an upper bound on the *ab initio* energy of the molecule may be provided by a calculation on a reference or guess structure, if the energy of the current best structure identified by the GA is below this threshold, it is generally not possible to determine whether the GA has converged to the global minimum or a local minimum. One approach to limit this uncertainty is to run multiple GA instances and examine the consensus of their output. A self-contained (within the GA run) approach to increase the confidence that the proposed minimum is indeed the global one, is to employ one or more restart mechanisms.[39,40]

In this work, two restart mechanisms were implemented, a Linear Search (LS) and a Cataclysmic Mutation (CM) restart. The Cataclysmic Mutation restart is a modified implementation of the restart employed in the well-known CHC genetic algorithm.[41] These two mechanisms were chosen to represent a "near search" and "far search," respectively. The linear search simply makes all possible one gene substitutions from the best genome identified by the GA. Therefore, the number of additional conformer evaluations, P_{LS} is given by

$$P_{LS} = \sum_{n=1}^{L} R_n - 1, \tag{1}$$

where L is the number of rotatable bonds in the molecule and R_n is the resolution applied to each rotatable bond. For the molecules in the Meta-GA test set, where 8 bonds (i.e., $L = 8$) are considered with a resolution of $R = 3$, this leads to 16 conformers being evaluated. The chemical rationale for employing this restart mechanism is a case where there are two (or more) possibly near-isoenergetic isomers that only differ by the rotation of one torsional angle and the GA identifies the wrong one of them as being the global minimum. This is most likely to reflect the rotation of end-groups, but could also occur in long aliphatic chains or molecules with intra-molecular hydrogen bonds. Somewhat arbitrarily, the LS restart is triggered when the best genome has not changed in $20L$ generations.

The Cataclysmic Mutation restart is employed as a final "line of defence" against premature convergence and is implemented after a Linear Search has failed to identify a lower energy conformer. In this restart mechanism, a new population of the same size as the parent GA is created and the current best genome is copied to every member of the population. Each genome is then mutated with probability, $p(M)$ and the

resultant population is evaluated. The mutation probability, $p(M)$ for each individual gene begins at 0.05 and increases by 0.05 every 5 generations. If $p(M)$ reaches 0.4 without a lower energy genome/conformer being found, then the GA is declared to have converged to the global minimum. In this sense, the CM restart may be thought of as a reverse simulated annealing. This is a considerably more expensive restart mechanism, with a cost of $35P$, where P is the population size chosen for the GA. This restart mechanism, however, is capable of locating optima that differ by more than one gene from the current best candidate.

In the case that either of these restart mechanisms locate a lower energy structure, it is inserted into the GA population and all restart counters are reset.

C. Molecular calculations

Preliminary work on this problem[36] employed the B3LYP density functional theory method and a moderate, 6-31+g(d,p) basis set to exhaustively search rigid surfaces. To study the effect of optimising each molecular conformer (i.e., a relaxed surface), a less expensive method was required. The Hartree-Fock (HF) method paired with a small, 6-31G basis set was thus chosen and both rigid and relaxed surfaces were generated by an exhaustive conformer search using this method. In the study of biological molecules, solvation effects are often significant, so each training set molecule was also optimised in a self consistent reaction field (SCRF), employing the polarizable continuum model and water as the solvent. For the rigid surfaces, any conformer with an interatomic distance <0.5 Å was excluded and assigned an energy of zero. When the molecular geometry is allowed to relax, the possibility of the molecular connectivity changing from the originally defined connectivity exists, leading to vastly different conformer energies. To avoid this, the connectivity of each conformer was tested against the original, optimized geometry of the molecule and those that did not match were excluded and also assigned an energy of zero. These surfaces were then saved and used as a "lookup table" in the meta-GA experiments. All molecular calculations were undertaken using the GAUSSIAN 09 (Ref. 42) program package.

To facilitate comparison with previous results, the five molecules chosen were as follows: carnosine and four molecules selected from x-ray crystal structures lodged in the Cambridge Structural Database;[43] DAWMOE, EZUDUY, WAVQAM, and WOBLII. The molecules were selected to be topologically and chemically distinct while being of similar size. Each molecule has 8 rotatable sp^3 bonds, leading to 6561 total conformers when considered at 120° resolution. Structures of the five molecules are shown in Figure 1, with the rotatable bonds indicated in bold.

In the application of the derived GA to larger molecules, the smallest 7 molecules (up to $N = 13$ rotatable bonds) were evaluated using B3LYP/6-31+g(d,p), but to reduce computational expense all "larger" molecules (by number of rotatable bonds, as it is the number of rotatable bonds that defines the size of the search space and thus the expected number of evaluations) were evaluated using HF/6-31g. All calculations were unrestricted. Each conformer was evaluated using only a single-point energy, in order to permit the energy of the x-ray structure to be used as an upper bound to the target energy.

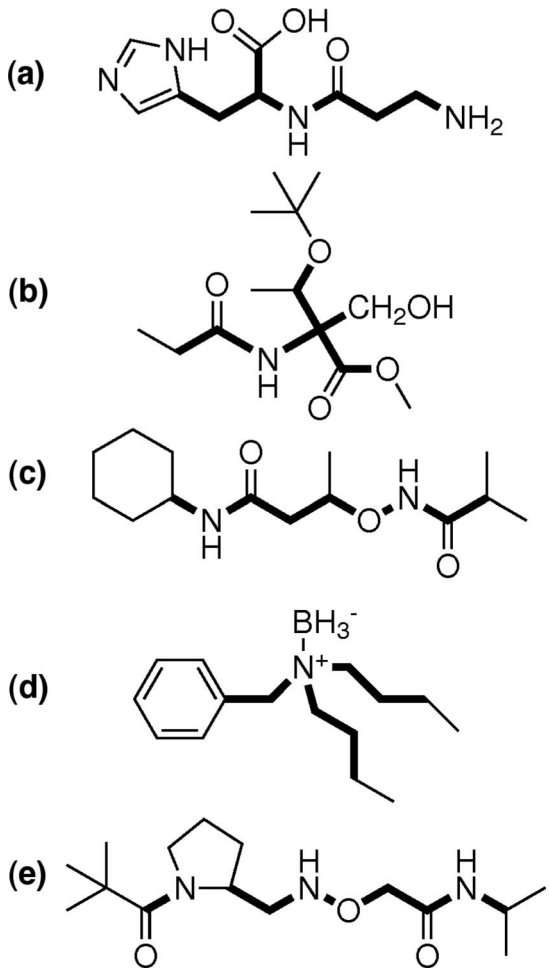


FIG. 1. The five molecules used to optimize GA parameters. Rotated bonds are indicated by bold lines.

III. RESULTS AND DISCUSSION

As in our preliminary[36] study, population size and mutation rate had the largest effect of the efficiency of the conformer-finding GA and so the efficiency of each optimized parameter set is graphed against only the population size and mutation rate ($\times 1000$). GA efficiency is defined as the mean number of conformer evaluations required to locate *a priori* known global minimum conformer of the molecule. In each GA run, the number of individual evaluations is capped at the cost of an exhaustive search (3^8 for all molecules presented here). The sole termination criterion used was the identification of the correct lowest energy conformer and any GA that failed to locate the global minimum 100% of the time is deemed to be unreliable. Full results for the test set of five molecules and rigid, relaxed and SCRF surfaces are included in the supplementary material.[37] Representative results for the DAWMOE molecule are shown in Figure 2.

In each of these graphs, three distinct regions were observed: For low mutation rates (LHS), performance is unreliable. The mutation rate is insufficient to absolutely guarantee

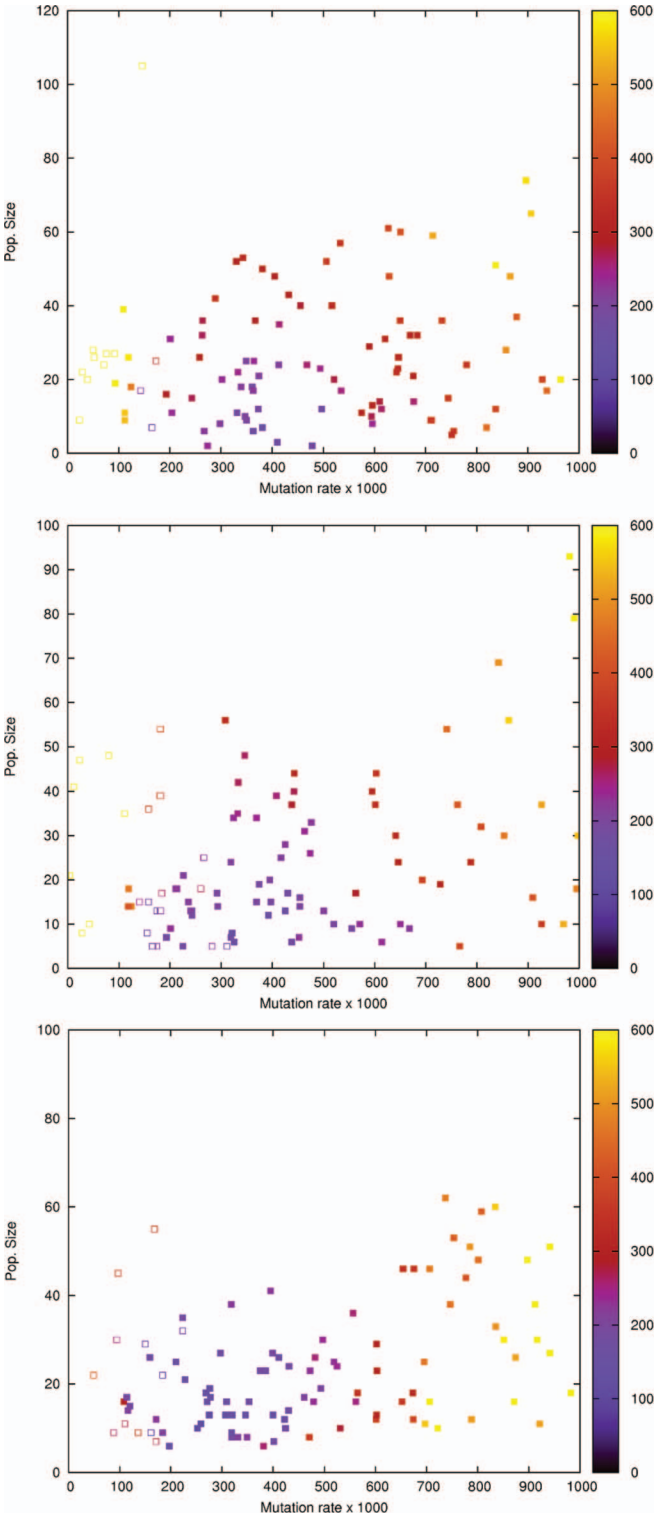


FIG. 2. GA efficiency for the DAWMOE molecule on (a) HF/6-31G rigid, (b) relaxed, and (c) SCRF surfaces as a function of population size and mutation rate for parameter sets located by the meta-GA. Unreliable parameter sets are shown as hollow squares. Results for all molecules are included in the supplementary material (Ref. 37).

that the solution found will be a global, rather than a local, optimum. Increasing population size did not ameliorate this situation sufficiently, compared with an increased mutation rate. A high population size also increased computational load when stairclimbing, after the global hill had been located. For high mutation rates, the combination of elitism and a high

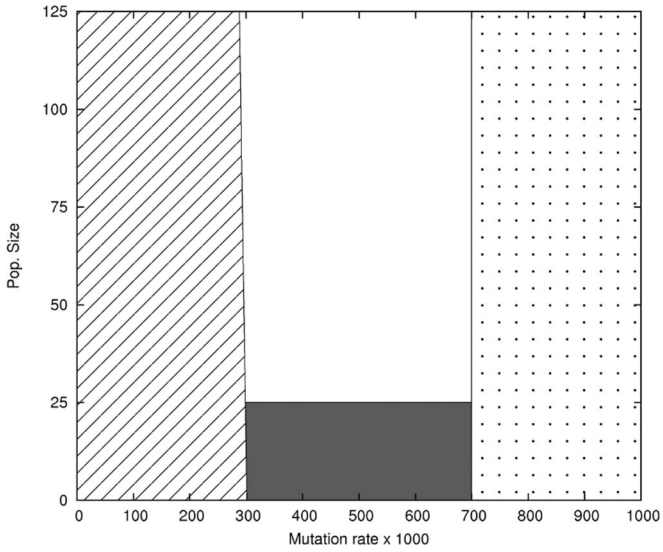


FIG. 3. Regions corresponding to good GA parameter sets (dark grey), high mutation rate (dots) and unreliable regions (diagonal stripes). Parameter sets in the remaining unshaded region are reliable but inefficient.

mutation rate means the global "hill" is quickly found, but the stairclimbing within that hill is slow as it relies on small mutations, which are improbable. Mutation rate in the PYEVOLVE tool is a square-root function: a mutation rate of 0.6 means that 0.6 of the genes are chosen randomly, and each has a 0.6 probability of mutating — thus 0.36 on average will mutate, with a minimum of 0 and a maximum of 0.6. For this reason, mutation rates above 0.7 are considered as having essentially random mutation.

For mutation rates around 0.3–0.5, adequate coverage of the surface, and adequate stairclimbing were both possible with a relatively small population. Within this region, a somewhat broad "sweet spot" of optimal tunings was located. These three regions are illustrated in Figure 3.

Some qualitative conclusions with regard to the underlying chemical problem are also evident from Figure 2: The HF rigid surfaces yield very similar optimum parameters to the previous B3LYP/6-31G(d,p) results, suggesting that in the context of molecular conformer searches, neither the level of theory or the basis set applied to the molecule change the optimum GA tunings. Therefore, parameters derived at an inexpensive level of theory can be used for other (more accurate and expensive) levels of theory.

There is, however, a striking difference between the results for the rigid and relaxed Hartree-Fock surfaces. Primarily, the region associated with unreliable parameter sets extends up to a mutation rate of 0.3 (compared to 0.2 for rigid surfaces) and the region of most efficient parameters is shifted toward higher mutation rates by a similar amount. The optimum population size does not change significantly. The SCRF results (which include optimization of the conformer) do not differ significantly from the gas-phase relaxed results.

This difference in ideal mutation rate may be explained in terms of the underlying surface. When each conformer is permitted to relax, several initial conformers may relax to a single molecular geometry and thus, energy. The relaxed surface therefore, comprises several steps or "plateau"

TABLE II. Mean, minimum and maximum population sizes and mutation rates for the 10 best genetic algorithms found for each molecule and model chemistry combination. The mean population size and mutation rate is also shown for each model chemistry.

	Lowest Mean Evaluations	Highest Mean Evaluations (n points=10)	Pop Size			Mutation Rate × 1000		
			Mean	Min[a]	Max	Mean	Min	Max
B3LYP rigid surface[a]								
Carnosine	218.52	279.5	22.7	2	43	353.5	225	463
DAWMOE	127.44	161.6	11.2	8	16	368.1	215	513
EZUDUY	146.88	192.1	9.8	6	17	268.8	188	384
WAVQAM	120.06	163.84	10.9	5	16	332.1	198	436
WOBLII	95.4	122.87	8.5	3	12	243.6	152	429
			12.62			313.22		
HF rigid surface								
Carnosine	94.36	139.7	8.9	4	14	258.9	188	381
DAWMOE	168.84	207.25	10.5	2	25	389.3	331	497
EZUDUY	151.12	195.78	8.2	3	13	310.2	195	434
WAVQAM	117.12	159.72	7.6	4	14	297.7	210	425
WOBLII	95.88	113.04	7.8	4	12	247.5	125	360
			8.6			300.72		
HF relaxed surface								
Carnosine	425	465.93	12.8	2	31	333.475	503	900
DAWMOE	156.56	190.6	9.4	5	17	331.1	193	438
EZUDUY	482.86	538.4	24.1	7	42	517.6	340	903
WAVQAM	69.12	73.7	6.2	5	10	685.7	468	865
WOBLII	88.25	119.6	6.9	5	14	237.2	136	436
			11.88			421.015		
HF SCRF relaxed surface								
Carnosine	176.54	207.09	10.7	4	14	435.5	201	597
DAWMOE	116.056	158.21	13.6	6	19	279.6	197	400
EZUDUY	197.26	334.02	14.9	5	21	420.9	280	559
WAVQAM	79.14	97.3	10.4	4	17	646.8	369	931
WOBLII	282.595	392.4	15.1	6	32	359.1	189	550
			12.94			428.38		

[a]B3LYP results are taken from Ref. 36.
[b]An artificial floor of 5 is applied to the population size.

regions of varying sizes and "heights." Without the "gradient" information implicit in a more smoothly varying surface, a higher mutation rate is required to move from one step to another.

To quantify the optimum population size and mutation rate, the parameter sets were ranked in order of the mean number of evaluations required to locate the correct conformer, with unreliable parameter sets excluded. The mean, minimum and maximum population size and mutation rate were calculated from the ten most efficient parameter sets. The results are shown in Table II.

Several conclusions are evident from this table. Firstly, that the number of evaluations required to determine the lowest energy conformer depends more on the geometry of the molecule (and thus the conformer landscape), than the level of theory used. Secondly, it is evident that the ten most efficient GA parameter sets tended to have efficiencies within 25% of the single best parameter set suggesting that slight variations in GA parameters from the optimal, will have only a small effect on the GA efficiency.

A. Effect of selection method

To further investigate the more subtle effects of selection and crossover methods on the efficiency of the GA, the meta-GA was run a further 100 times on each molecular surface, with the population size and mutation rate fixed at the values determined in Table II (i.e., population size = 10 and mutation rate = 0.4 or 0.5 for the rigid and relaxed surfaces respectively). The swap mutator was found to be deleterious in effect and so was not used. Conversely, elitism was found to be universally helpful and so was always used in the GA.

The results for all five molecules showed that Rank selection, which always selects the best individual present in the population, clearly outperformed all other selection methods. Tournament selection may have some utility on relaxed surfaces. PYEVOLVE contains two implementations of Tournament selection, the "Alternate Tournament" method does not rely on a Roulette wheel for initialization, no difference was seen between the performance of these two implementations. Roulette and Uniform selection, both of which involve

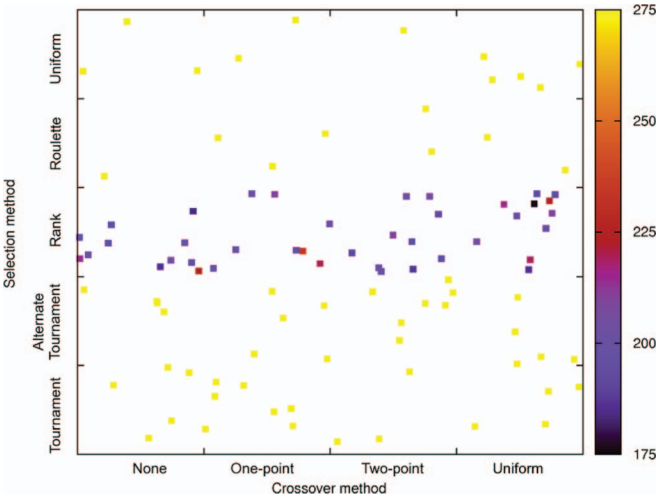


FIG. 4. GA efficiency vs. selection method and crossover method for the molecule DAWMOE on the HF/6-31G relaxed surface with population size and mutation rate held constant at values indicated in the text. Results for all molecules are included in the supplementary material (Ref. 37).

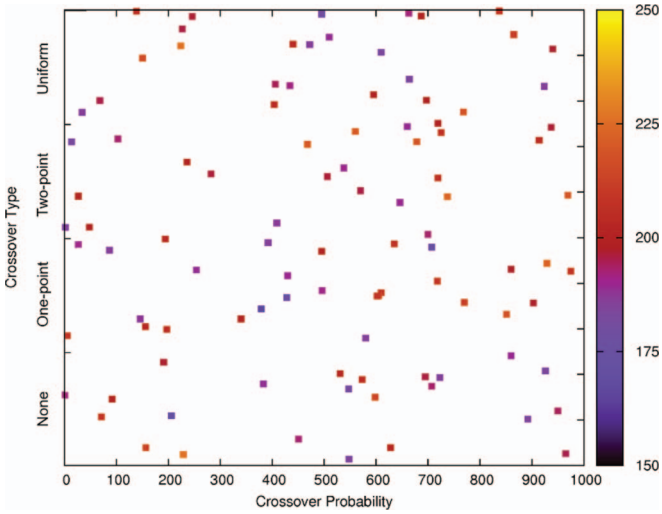


FIG. 5. GA efficiency vs. crossover method and crossover probability for the molecule DAWMOE on the HF/6-31G relaxed surface with population size and mutation rate held constant at values indicated in the text. Results for all molecules are included in the supplementary material (Ref. 37).

a random element, performed poorly. Representative results for the WOBLII molecule are shown in Figure 4.

B. Effect of crossover method

In neither the original meta-GA optimizations nor the constrained optimization, with population size and mutation rate fixed was any effect of crossover type or probability seen. To further show that the conformer search problem is insensitive to crossover, the meta-GA was run another 100 times on each molecular surface, this time fixing all variables except crossover type and probability (i.e., in addition to the constraints used in Sec. III A, the Rank selection method was fixed).

Again, the results for all molecules were consistent, however, no clear relationship between crossover and GA efficiency was observed. Results for the carnosine molecule are shown in Figure 5.

This is perhaps a surprising result, despite searching for the best *genetic algorithm*, the insensitivity to crossover indicates that an *evolutionary algorithm* is the most appropriate for conformer-searching an isolated molecule. The superior performance of the Rank selector indicates that the most efficient search of a molecular potential energy surface involves locating the global optimum "hill" and then stairclimbing by means of small mutations applied to the most fit individual in the population.

The effectiveness of crossover, which is the key step that differentiates a GA from an EA, is dependent on the underlying fitness landscape (Potential Energy Surface in this case). In turn, the shape of the fitness landscape is dependent on the representation used to define it. Rothlauf and Goldberg[44] analysed the effect of the representation on the efficiency of a GA and found that efficiency correlated with the locality afforded by the representation. In other words, fitness landscapes where the change in fitness due to a one gene change in the genome is at least moderately predictable, are more amenable to traversal by a GA. In a later paper,[45] Rothlauf considered the problems that can arise when using a redundant representation - where two or more genotypes represent the same phenotype. This situation is particularly likely to arise when the conformer is allowed to relax in the function evaluation. Regardless of redundancy, the crossover operator can produce offspring that bear little resemblance to its parents. In addition to considering the representation, the crossover operator itself can be designed, dependent on the representation, to constrain the difference between parent and child phenotypes. The cluster cut-and-splice operator implemented by Assadollahzadeh and co-workers[25] is an example of such a designed operator that is able to largely restrict the degree of change inflicted by crossover.

C. Application to larger molecules

To validate the optimized GA, it was tested on several larger molecules. A set of 12 molecules were selected from a search of the Cambridge Structural Database for molecules with the formula: $C_{20-40}H_{25-50}S_0P_0$ with all other atoms permitted in any quantity. The molecules spermine and tris(3-aminopropyl)amine were added as they were of local interest. As the aim was to validate the efficiency of the optimized GA, rather than undertake any specific chemistry, single point energies were used to evaluate each conformer. In only one case, HEGFOP, was a lower energy conformer identified by the GA: a structure 0.2 mEh below the energy of the CSD reference structure and differing only in the orientation of one bond was identified by each GA instance. Five GA instances were run for each molecule and the results are shown in Table III.

In order to judge the effect of molecular complexity on the rate of GA convergence, random conformers of the *n* −alkanes, $C_{13}H_{28}$, $C_{18}H_{38}$, $C_{20}H_{42}$, $C_{23}H_{48}$ were constructed and subjected to the GA. Co-ordinates of the initial guess and final optimized structures are shown in the supplementary material.[37] DIYSOU contains one branch along its backbone as

TABLE III. GA performance on larger molecules. Molecule names in small caps indicate structure key names from the CSD Database. *N* indicates the number of rotatable bonds in each molecule; *x*, *y* indicate bonds restricted to 2-way, 3-way rotation respectively, E represents the energy of the minimum structure in Hartree (target energy, see text for further details), N_{gen} is the generation in which the global minimum was located in each run.

Molecule	*N*	E	N_{gen}					Mean N_{gen}	Size of Search space	Search space searched (%)
REFNOF	1,9	−904.304391	68	56	117	105	69	83	39366	2.05
$C_{13}H_{28}$	10	−512.332013	27	75	48	219	13	76	59049	1.29
DIYSOU	10	−941.406353	37	67	56	143	78	76	59049	1.29
HEGFOP	11	−863.258002	69	61	128	93	98	90	177147	0.51
GAPTAS	12	−891.828153	313	210	269	38	152	196	531441	0.37
Tris(3-aminopropyl)amine	12	−576.4469428	111	107	105	66	64	109	531441	0.17
spermine	13	−615.773061	113	76	49	30	55	65	1.59×10^6	0.04
MEJMEU	14	−1101.169396	293	491	568	592	178	424	4.78×10^6	8.87×10^{-2}
$C_{18}H_{38}$	15	−703.489386	809	252	430	182	654	465	1.43×10^7	3.24×10^{-2}
FABPUU	15	−1075.473261	296	815	739	1403	551	761	1.43×10^7	5.30×10^{-2}
FEQVAY	16	−1651.556526	1012	355	664	257	798	617	4.30×10^7	1.43×10^{-2}
$C_{20}H_{42}$	17	−781.5248878	400	757	441	643	806	609	1.29×10^8	4.72×10^{-3}
GOYHEH	17	−1510.866832	1029	891	1414	629	953	983	1.29×10^8	7.61×10^{-3}
POWFIQ	17	−1741.434989	1750	567	491	835	900	909	1.29×10^8	7.04×10^{-3}
XOHRIV	18	−1549.966843	701	184	241	794	1008	586	3.87×10^8	1.51×10^{-3}
$C_{23}H_{48}$	20	−898.579203	786	1246	1861	1392	1115	1280	3.49×10^9	3.67×10^{-4}
AFIHON	20	−1310.234742	959	1585	1161	1463	1505	1335	3.49×10^9	3.83×10^{-4}
CLAMPL	23	−2523.10295	1171	1298	1717	1368	1934	1498	9.41×10^{10}	1.59×10^{-5}

well as ether and amide functional groups, but does not take significantly longer to converge than the equivalent length alkane. In larger molecules, however, the effect of molecular complexity was quite readily evident. Both GOYHEH and POWFIQ molecules are branched molecules with a number of heteroatoms, each took over 900 generations to converge, i.e., approximately 150% of the number of generations required to identify the global minimum of $C_{20}H_{42}$. By contrast, AFIHON is essentially a linear structure and it is therefore unsurprising that its global minimum is located in a similar number of generations to $C_{23}H_{48}$.

It is an assumption of the meta-GA procedure that the target problem is scalable and that a larger problem of the same type can be solved in the same way (i.e., using the same GA) as the smaller problem. Table II indicates that within the "good" region of Figure 3, the GA performance should be close to optimum. To confirm that the optimized GA parameters are more efficient than non-optimized parameters for a large molecule, the rigid FABPUU molecule, using the AM1 (Ref. 46) semi-empirical method, was submitted to 10 instances of both the optimized GA and a second GA, defined by the PyEvolve default settings. No restart mechanisms were employed in this test. With over 14 million possible conformers, it is not feasible to exhaustively search the entire surface, even at a low level of theory and so, consensus scoring must be employed. Nine instances of the optimized GA converged to the same conformer, $E_{(AM1)} = -0.110723$, requiring a mean of 364 generations (3635 evaluations) to do so. In contrast, none of the default GA instances identified the correct conformer, even when allowed over 80,000 evaluations.

D. Convergence behaviour

The relaxed FABPUU molecule was chosen to study in further detail the convergence behaviour of the GA. The three branches extending from the central nitrogen atom are the same, leading to the possibility of a highly symmetric geometry being the global minimum. In addition the –NH and =O atoms on each branch have the capacity to form intramolecular hydrogen bonds, thus suggesting a complex potential energy surface with strongly bound local minima that represent a "trap" for the GA. As in Sec. III C, five individual instances of the GA were used. In this case, where the global minimum is entirely unknown, multiple GA instances provide consensus on the identified global minimum as well as average efficiency.

All five instances of the GA converged to a structure possessing pseudo-C_3 symmetry with each of the three branches hydrogen bonded to both other branches. We therefore propose this structure, E(HF/6-31g) = −1075.473674, as the gas phase global minimum. Only one GA instance converged directly to this minimum, the other four initially found the second-lowest energy structure, which represents a "false" minimum, 1.5 mEh above the global minimum, but eventually progressed to the global minimum. In this 'false' minimum structure, only two of the three branches are hydrogen-bonded; there are, therefore, three equivalent structures of this type, suggesting an increased likelihood of locating this structure. In three of the four runs that located this structure first, convergence was reached only after either linear search or cataclysmic restart. Figures 6(a)–6(c) shows the reference structure, the false minimum and the global minimum respectively.

The progress of each GA instance can be seen in Figure 7. All five instances made regular replacements of the candidate best genome for approximately the first 450 generations, the time spent "stuck" in the false minimum is evident after this point. This result suggests that more stringent criteria could be used to trigger resort to restart mechanisms later in the evolution.

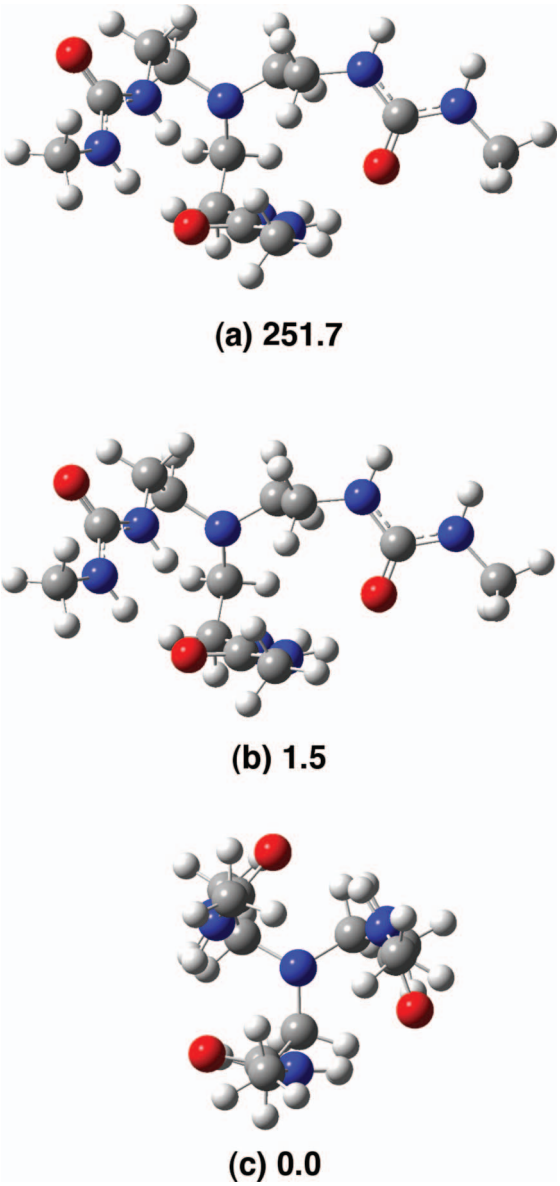


FIG. 6. Three conformers of the FABPUU molecule: (a) The CSD reference structure, reoptimized. (b) The false minimum. (c) The pseudo-C_3 global minimum. Δ E (mEh), relative to the proposed minimum, (c) is indicated below each structure.

A second consideration when examining convergence is the distribution of energies at the conclusion of the GA instance. This is important for two reasons; firstly, many chemical problems, such as prediction of thermochemical properties, require consideration of multiple minima and secondly, in the case where a GA instance is terminated prematurely and the global minimum has not been discovered, an understanding of the energy distribution permits an estimate of the relative energy (to the unknown global minimum) of the candidate structure. At the termination of each instance, the distribution of non-zero energies calculated is the same (ANOVA, P = 0.05). Further details can be seen in the supplementary material.[37] Table IV shows the distribution of energies above the global minimum energy for the non-zero evaluations (i.e., valid molecular co-ordinates) encountered by each instance. The lowest 1% of energies evaluated cover an average range

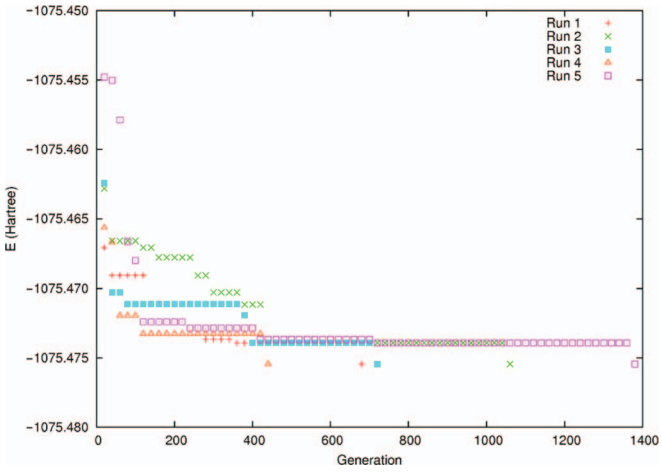


FIG. 7. Energy (Hartree) of best conformer by GA generation for 5 independent GA runs on the relaxed FABPUU molecule using HF/6-31g. Between generations 450 and 700, several lines are coincident. The generation size is 10.

of approximately 5 milliHartree, from the false minimum to 6.7 mEh, however, the actual range covered varies from 2.8 mEh (Instance 1) to 7.1 mEh (Instance 4). Considering the lowest 5% of evaluated energies removes the majority of this variability and consistently covers a range of approximately 15 mEh above the global minimum, 25% fall within 25 mEh above the global minimum and 90% of evaluated energies are within 50 mEh. This relatively consistent clustering of energies suggests that sampling the evaluated conformers at the termination of an instance would provide a useful snapshot of the conformational landscape.

E. Applicability

The profusion of conformational search methods and the fact that new methods are still an active area of research, indicates that conformer search is not yet a globally solved problem. A recent research overview[4] answered the question in the affirmative (i.e., that conformational searching is a solved problem) only after applying the restriction of considering druglike molecules with ≤10 rotatable bonds. Most methods

TABLE IV. Distribution of conformer energies, expressed as ΔE (mEh) w.r.t. the energy of the proposed global minimum, E = −1075.473674. N_{eval} is the number of non-zero molecule energies evaluated in the GA instance.

	Instance					
Percentile	1	2	3	4	5	Mean
0.01	4.3	6.6	6.6	8.6	7.7	6.7
0.02	8.0	9.1	10.3	11.4	10.8	9.9
0.05	13.7	14.3	14.3	15.4	15.1	14.6
0.10	17.8	18.2	18.6	19.0	19.0	18.5
0.20	22.7	23.2	23.3	23.5	23.7	23.3
0.25	24.5	25.1	25.4	25.2	25.5	25.2
0.50	31.5	32.0	32.6	32.1	33.2	32.3
0.75	40.6	41.3	42.0	41.1	42.0	41.4
0.90	49.4	49.8	50.5	49.3	50.1	49.8
(N_{eval})	3433	5684	4496	4959	7240	

have some limitation in applicability and define their search space differently, which makes comparison between methods difficult, however, some comparisons are possible. The largest test molecules studied using the EDTS (Ref. 20) systematic method (which employs a similar torsional encoding to that used here) have 11 rotatable bonds and depending on the molecular topology, search between 0.1 and 0.6% of the search space, compared with 0.51% required for a similarly flexible molecule in this study. Several variants of a Tabu-search based conformational search[47] located the global minimum of [Met^5]enkephalin (which would have 20 rotatable bonds in this study) in 35–2610 steps, but the repeatability of the search also ranged from 0–87%. The EA produced by this study has been shown to perform well with search spaces of 10^6–10^{11}. For larger search spaces, one would expect the scaling to improve, but reliance on restart mechanisms would likely increase with increasingly complex molecular topology.

IV. CONCLUSIONS

The meta-GA approach is useful for scalable problems, where two criteria can be met: Firstly that a "small" model problem is soluble by exhaustive search and secondly that such a model problem contains all the "features" of a larger problem. The first criterion provides the test data for the meta-GA and the second ensures that the parameters derived for the model problem are applicable for similar problems that are too large to solve exhaustively. The use of such an automated tool allows for many variables to be optimized simultaneously and is not subject to the vagaries of human intuition.

For obtaining the lowest energy conformer of an isolated molecule, on both rigid and relaxed surfaces, a population size of between 10 and 15 is ideal. In the case of a rigid surface, a mutation rate of between 0.3 and 0.5 is most efficient, however, when the surface is relaxed, a mutation rate of at least 0.4 is required. To ensure reliability of the GA, we therefore recommend a mutation rate of 0.4 for a rigid surface and 0.5 for a relaxed surface, using a population size of 10 in both cases. The crossover operators tested were not found to be helpful, resulting in an Evolutionary Algorithm rather than a Genetic Algorithm.

The set of molecules investigated show that the "essence" of the conformational search problem is unaffected by molecular connectivity (numbers of tertiary and quaternary carbon atoms), functional groups or ionic character. It is also shown that the ideal GA tunings are invariant to the quantum chemical method and basis set used in the evaluations of each conformer. Therefore the two sets of parameters presented here are applicable to the conformational search of any single molecule undertaken at any desired level of theory.

ACKNOWLEDGMENTS

The authors gratefully acknowledge a grant of computer time from the National Computational Infrastructure (NCI). The assistance of Dr. Alister Page in reading early versions of the manuscript is also gratefully acknowledged.

[1] B. K. Shoichet, I. D. Kuntz, and D. L. Bodian, J. Comput. Chem. **13**, 380 (1992).
[2] W. Cai, X. Shao, and B. Maigret, J. Mol. Graph. Model. **20**, 313 (2002).
[3] N. Foloppe and I. Chen, Curr. Med. Chem. **16**, 3381 (2009).
[4] I.-J. Chen and N. Foloppe, Drug Dev. Res. **72**, 85 (2011).
[5] M. Saunders, K. N. Houk, Y. D. Wu, W. C. Still, M. Lipton, G. Chang, and W. C. Guida, J. Am. Chem. Soc. **112**, 1419 (1990).
[6] M. Saunders, J. Am. Chem. Soc. **109**, 3150 (1987).
[7] M. Saunders, J. Comput. Chem. **12**, 645 (1991).
[8] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, Science **220**, 671 (1983).
[9] S. B. Ozkan and H. Meirovitch, J. Phys. Chem. B **107**, 9128 (2003).
[10] C. Baysal and H. Meirovitch, J. Chem. Phys. **105**, 7868 (1996).
[11] S. B. Ozkan and H. Meirovitch, J. Phys. Chem. B **107**, 9128 (2003).
[12] G. Chang, W. C. Guida, and W. C. Still, J. Am. Chem. Soc. **111**, 4379 (1989).
[13] F. Daeyaert, M. De Jonge, L. Koymans, and M. Vinkers, J. Comput. Chem. **28**, 890 (2007).
[14] Y. Sakae, T. Hiroyasu, M. Miki, and Y. Okamoto, J. Comput. Chem. **32**, 1353 (2011).
[15] C.-E. Chang and M. K. Gilson, J. Comput. Chem. **25**, 1987 (2003).
[16] I. Kolossváry and W. C. Guida, J. Comput. Chem. **14**, 691 (1993).
[17] J. M. Goodman and W. C. Still, J. Comput. Chem. **12**, 1110 (1991).
[18] A. R. Leach and K. Prout, J. Comput. Chem. **11**, 1193 (1990).
[19] P. Tuffery, C. Etchebest, S. Hazout, and R. Lavery, J. Comput. Chem. **14**, 790 (1993).
[20] E. I. Izgorodina, C. Y. Lin, and M. L. Coote, Phys. Chem. Chem. Phys. **9**, 2507 (2007).
[21] I. Kolossváry and W. C. Guida, J. Am. Chem. Soc. **118**, 5011 (1996).
[22] C. Lavor, Int, J. Quantum Chem. **95**, 336 (2003).
[23] C. D. Maranas and C. A. Floudas, J. Chem. Phys. **100**, 1247 (1994).
[24] Y. Lin and M. A. Stadtherr, J. Comput. Chem. **26**, 1413 (2005).
[25] B. Assadollahzadeh, P. R. Bunker, and P. Schwerdtfeger, Chem. Phys. Lett. **451**, 262 (2008).
[26] J. L. Llanio-Trujillo, J. M.C. Marques, and F. B. Pereira, J. Phys. Chem. A **115**, 2130 (2011).
[27] J. Fuhrmann, A. Rurainski, H.-P. Lenhof, and D. Neumann, J. Comput. Chem. **31**, 1911 (2010).
[28] P. Pfeffer, T. Fober, E. Huellermeier, and G. Klebe, J. Chem. Inf. Model. **50**, 1644 (2010).
[29] T. Brodmeier and E. Pretsch, J. Comput. Chem. **15**, 588 (1994).
[30] J. J. Grefenstette, IEEE Trans. Syst. Man Cybern. **16**, 122 (1986).
[31] B. Friesleben and M. Hartfelder, in *Artificial Neural Networks and Genetic Algorithms*, edited by R. Albrecht, C. Reeves, and N. Steele (Springer-Verlag, Heidelberg, 1993), pp. 392–399.
[32] W. A. de Landgraaf, "Parameter calibration using meta-algorithms," Master's thesis, Artificial Intelligence Vrije Universiteit, Amsterdam, 2006.
[33] R. Haupt, in *Antennas and Propagation Society International Symposium, 2000. IEEE*, Vol. 2 (2000), pp. 1034 –1037.
[34] Y. Zhang, M. Sakamoto, and H. Furutani, in *Natural Computation, 2008. ICNC '08. Fourth International Conference on*, Vol. 1 (2008), pp. 70–75.
[35] Z. E. Brain and M. A. Addicoat, in *Genetic and Evolutionary Computation Conference, GECCO 2010, Proceedings* (2010), pp. 823–824.
[36] Z. Brain and M. Addicoat, in *Artificial Life XII, Twelfth International Conference on the Synthesis and Simulation of Living Systems* (2010), pp. 378–385.
[37] See supplementary material at http://dx.doi.org/10.1063/1.3656323 for graphs showing optimised GAs for each training set molecule.
[38] C. S. Perone, SIGEVOlution **4**, 12 (2009).
[39] W. Lin and T. Chen, Neurocomputing **69**, 2301 (2006).
[40] M. Magdon-Ismail and A. Atiya, Neural Comput. **12**, 1303 (2000).
[41] L. J. Eshelman, "The CHC adaptive search algorithm: How to have safe search when engaging in nontraditional genetic recombination," in *Foundations of Genetic Algorithms*, edited by G. J. E. Rawlins (Morgan Kauffman, San Mateo, CA, 1991), Vol. 1, pp. 265–283.
[42] M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, GAUSSIAN 09 Revision A.2, 2009, Gaussian Inc., Wallingford, CT.
[43] D. Fletcher, R. McMeeking, and D. Parkin, J. Chem. Inf. Comput. Sci. **36**, 746 (1996).
[44] F. Rothlauf and D. Goldberg, in *Parallel Problem Solving from Nature PPSN VI* (Springer, Berlin, 2000), pp. 395–404.
[45] F. Rothlauf and D. Goldberg, Evol. Comput. **11**, 381 (2003).
[46] M. Dewar, E. Zoebisch, E. Healy, and J. Stewart, J. Am. Chem. Soc. **107**, 3902 (1985).
[47] C. Grebner, J. Becker, S. Stepanenko, and B. Engels, J. Comput. Chem. **32**, 2245 (2011).