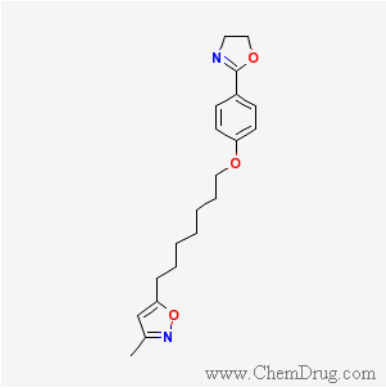
Supplementary material:

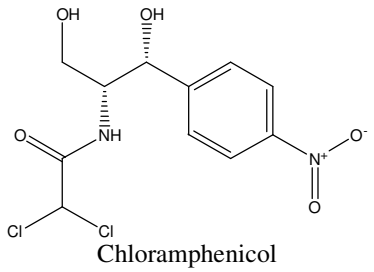
1. Optimizing the performance of ISE

ISE requires three sample sizes – the total size of the sample of conformations (for the "production" step) and the sizes of lowest and highest scorings (for the "analysis" step). The test described in the following section was used to obtain optimal values of these parameters.

Three questions were examined for the parameters used by ISE algorithm: 1. How does the sample size N, 2. how do upper and lower regions (N_H, N_L) of the scoring affect the final results? 3. How do factors of elimination evf_L and evf_H affect the final results?

Optimizations of the samples size N, N_H, N_L and elimination factors evf_L and evf_H were performed with two flexible molecules: WIN51711 (an Oxazole derivative) and Chloramphenicol shown in the figure below (figure S1)





Chloramphenicol

Figure S1. Ligands used for optimizing parameters for ISE

WIN51711 (Ligand W71 in three PDB structures 2RO4, 1PIV and 1D4M) is a highly flexible ligand demonstrated in those PDB structures to assume multiple conformations and multiple binding modes. It is a long molecule with 10 rotatable bonds (out of which 9 are consecutive, figure S1) which allows a huge number of free state conformations. In order to optimize the performance of ISE, we kept 4 bonds fixed and performed a full exhaustive search on the remaining six rotatable bonds, each at an increment of rotation of $30^{0,}$, thus representing $12^6 \sim 3x10^6$ conformations. The same procedure was applied to Chloramphenicol. In Choramphenicol we performed a full exhaustive search on all rotatable bonds, each at an increment of rotation of 45^0, representing $8^7 \sim 2x10^6$ conformations. The threshold for entering exhaustive search was 100,000 conformations for both molecules.

Computation times

Each iteration takes about 0.2 seconds computation time on a dual Xeon PC running GNU/Linux OS, not differing much with the sizes of samples analyzed. However energy calculation using Sybyl 7.0 package took about 25-55 seconds for 1000 conformations, increasing linearly with the number of conformations being analyzed.

The effect of sample size

ISE was applied with different total sample sizes (N=1,000 to 100,000, comprising between ~0.022-3.3% of the total conformations) and different sample sizes for the highest and lowest scoring function regions in each sample (N_H and N_L vary between 100 and 10,000, comprising 1-10% of the total sample, or 0.0022-0.83% of all

conformations) and the best set was compared to a similarly sized set of the best results from an exhaustive search. These sets included 100, 500 and 1000 best results, and the numbers of overlapping results from ISE with those of the exhaustive search are shown in columns of three groups (100, 500 and 1000 best exhaustive results) in figure S2 for WIN51711 and in figure S3 for Chloramphenicol. The higher the column in each group, the better the results.

The calculations were performed with a constant evf_L of 0.5 and evf_H of 2.0

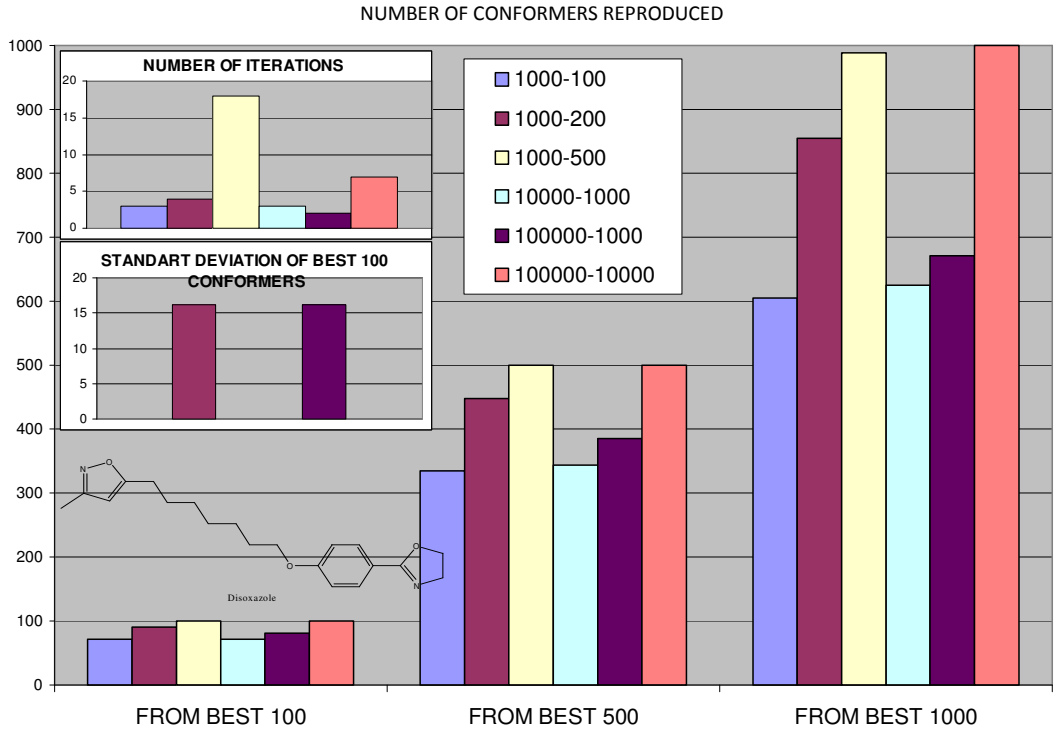


Figure S2. Reproduction of lowest energy conformers of WIN51711 from the full exhaustive search by the ISE method with different total and highest-lowest sample sizes

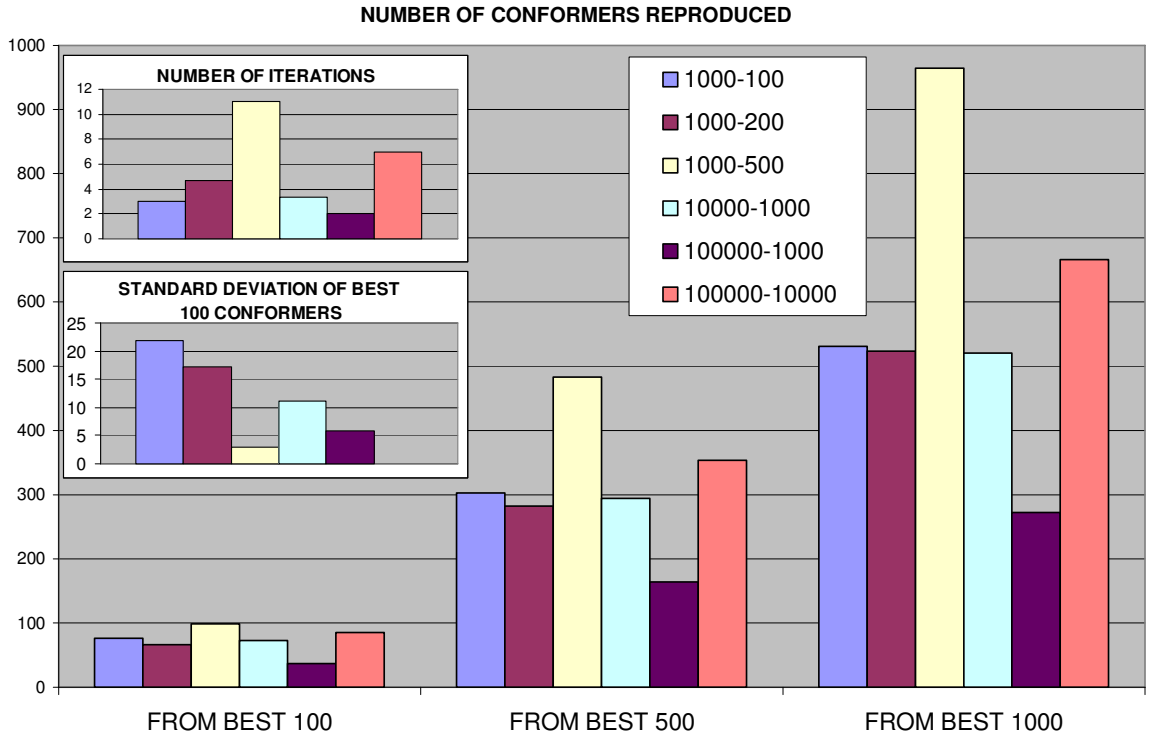


Figure S3. Reproduction of lowest energy conformers of Chloramphenicol from the full exhaustive search by the ISE method and highest-lowest sample sizes

For each bond rotation variable there are initially twelve and eight values of rotation angles respectively in the present applications. In a sample of 10,000, each of these angles is expected to appear some ~830 and 1250 times respectively, and the expectation value for the highest and lowest regions depend on the percentage of those regions in the full sample. Elimination decisions are made on the basis of these regions alone, and so it is crucial to have a representative number of each angle.

The 3 total sample sizes (1K, 10K and 100K) were all examined with N_H and N_L values of 10% while 1K was also examined with additional percentages of N_H and N_L. The need for the present analysis is clear once it is realized that the results for 20% and 50% of a total sample of 1K are better than those of 10% out of 10K and 100K, at all three groups.

Naturally, results do improve with sample size at each level, but the variations are smaller between the larger sets of 1K, 10K and 100K if one compares ISE results to a smaller set of 100 exhaustive results. If these 100 best results are sufficient to describe the ensemble of conformations (i.e., the spread of energy levels is large enough to eliminate the contribution of higher energy levels), then there is no need to spend much more time to compute large samples: 10K does nearly as well as 100K. Within 1K sample size, a bigger percentage of N_H and N_L achieves better results, which reflects the fact that making decisions to eliminate values is improved if bigger groups of best and worst results are examined, however at the expense of more iterations needed to complete the calculation..

In a total sample of 1000, examining a region of only a 100 best and worst scores suggests an expectation value for each angle of ~8 and 12.5 respectively, while it is ~80 and 125 in the samples of 10K/1000 or 100K/1000, with better significance. Much larger examined regions such as 100K/10K have expectation values of ~800 and 1.25K respectively for each angle, and the full sample is expected to have an expectation value of > 8K or 12.5K respectively, but decisions can not be made on the basis of expectations in the full sample, because these decisions depend on the contribution of particular variable values to the presence of results among the worst compared to the best scoring, not all across the scoring results. It thus seems that at both ends of N_H and N_L , enlarging their sizes will improve the results but will also take more computation time, and those values should ideally be optimized to balance elimination efficacy and computation cost.

If the best 100 conformations are retained, the overlap with full exhaustive results is between 60%-95%. For 500 conformations, it is ~50-80% and for 1000

conformations, ~40-70%. Retaining larger numbers after the ISE iterations and eliminations thus risks an increasing loss of best results.

We have seen in previous applications of ISE that the global minimum and other closely lying minima were retained in most problems in which ISE performance was compared to exhaustive results. In the present application global minimum and best 10 conformations were retained.

Examining the Stochastic effect of Random Number Generation

Since ISE is a stochastic search method which randomly samples conformations, a measure of consistency of results should be part of any optimization procedure. Each calculation was repeated three times with different seed numbers for the random generation of conformers. The standard deviation of the results of the three calculations is shown in the small graphs below the number of iterations (figures S4 and S5). Lack of a column in this graph means the standard deviation is zero i.e. all three results were similar.

Our previous tests for increasing the size of the total sample and increasing the percentage of N_H and $N_{L,}$ that were found to increase selectivity and improve the results, were also found to improve the consistency of results. This however is at the expense of increasing the number of iterations. The elimination procedure sometimes filters all high energy conformations and leaves at the end of the calculation samples where the energy difference between conformers is very small. We use increased percentages of N_H and N_L to handle such "flat" combinatorial surfaces – surfaces where energy differences between conformers are relatively small.

The effect of elimination factors

The effect of changing elimination factors was applied for a range of numbers (evf_L varied between 0.1 and 0.6; evf_H varied between 1.5 and 4) and the best set was compared to a similarly sized set of the best results from an exhaustive search. The total sample size and highest and lowest regions were kept fixed at 1000 and 100 conformations respectively. The results of these experiments as well as the random number generations are included in figures S4 and S5.

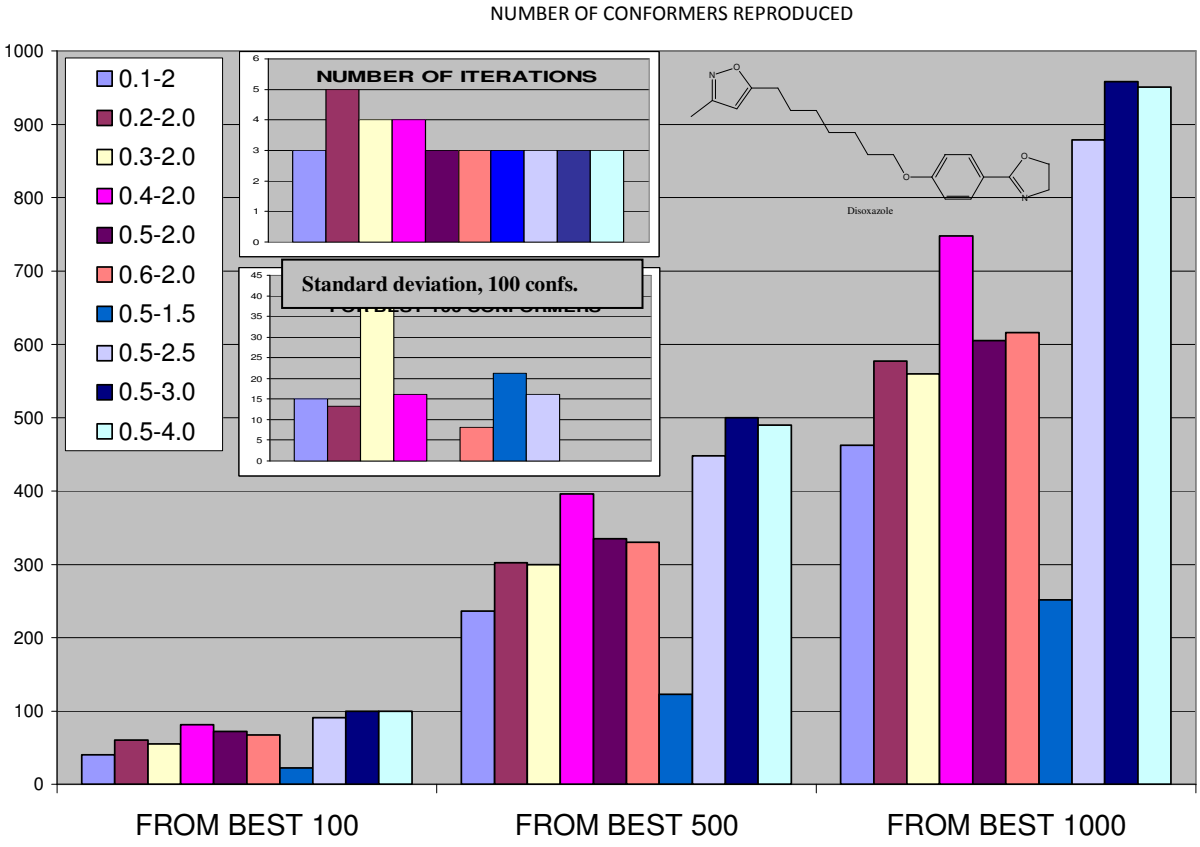


Figure S4. Reproduction of lowest energy conformers of WIN51711 from the full exhaustive search with different elimination factors. Results for different randomly generated numbers are given in the small upper left, lower block of "standard deviations".

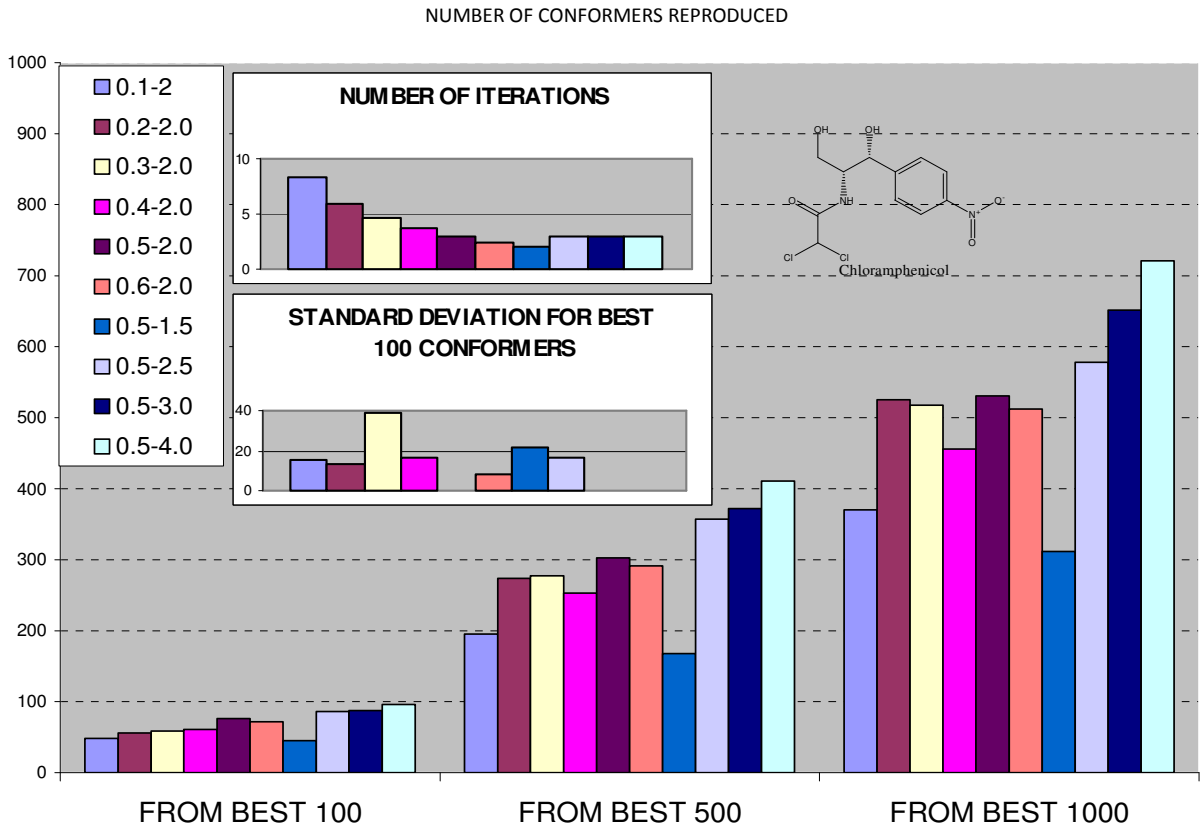


Figure S5. Reproduction of lowest energy conformers of Chloramphenicol from the full exhaustive search by the ISE method with different elimination factors. Results for different randomly generated numbers are given in the small upper left, lower block of "standard deviations".

The global minimum and best 10 conformations were almost always retained except when using very extreme conditions for elimination of values, such as a very small factor for elimination from the high energy group ($evf_H = 1.5$). This means that values that appear in the high energy group 1.5 times the expected number in the total sample will be evicted, so the selectivity is very low and parameters such as this are not usually used except for optimization of other parameters.

For evf_L the optimal values are 0.4 and 0.5 for WIN51711 and Chloramphenicol respectively. Lower evf_L values should increase selectivity, however the number of iterations also increases. This in turn increases the possibility of elimination of low energy conformers, therefore hampering the quality of the results. Higher evf_L values have low selectivity as expected.

For evf_H higher values generally give better results. The effect of increasing the number of iterations needed to complete the conformational search, observed with low evf_L was not seen with high evf_H i.e. the number of iterations remained the same, therefore only the selectivity effect was demonstrated and improved the results.

Both this and the previous test were preformed using the OR option for the two values, of evf_H and evf_L. Using the AND option prevents the elimination of low energy conformations, but on the other hand requires more iterations and does not always converge, therefore could not be used for optimizing parameters.

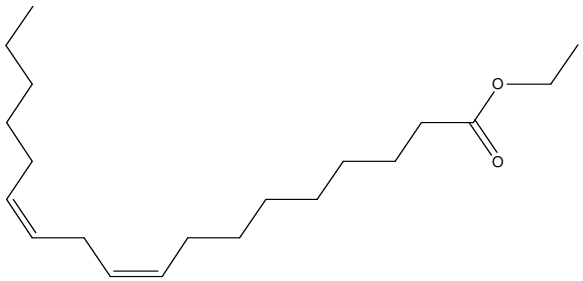
No procedure to correct erroneous elimination was used for these tests because we wanted to examine the net effect of the parameters. When adding the correction procedures, as described in the METHODS section of the paper, all best conformers were retained. In the next section we will use the correction procedures with large combinatorial spaces.

Erroneous elimination in large combinatorial spaces:

As the calculation proceeds, the expectation values for each variable value are growing as a result of the elimination of variable values because less angles will be sampled for each rotating bond. If the total sample size and those of highest and lowest regions are not reduced with each iteration, the chance for erroneous elimination becomes smaller due to the increased relative (relative for each value)

sample size. Therefore the problem of erroneous elimination exists predominantly in the initial iterations of large combinatorial spaces (>10^{15} possible combinations). We constructed a returning function for erroneously eliminated conformers and checked it against ensembles of low energy conformers. For molecules of the present study, WIN51711 and chloramphenicol, exhaustive searches are difficult or impossible to perform. Therefore we made use of the effect of repeated criteria checks mentioned in the methods section and shown in figure 2 of the methods section, to form lowest energy ensembles. The effect of repeated criteria check before elimination (CR) is similar to increasing sample size but in less computation time and with reduced file sizes. It also shows which would be the next eliminations if we were to decrease further the number of possible combinations. By increasing the number of repeated criteria checks before elimination we generated ensembles of lower energy conformers until those could no longer be lowered. We used these results as indications for the best conformers and measured the ability to reproduce them with the returning function described in section 2.5, also comparing the best energies obtained with each calculation.

Figure S6 shows consistency reproduction results for the lowest energy conformers that could not be lowered further. Ethyl linoleate is a flexible molecule which has 16 rotatable bonds giving a conformational space of about 1.85*E17 conformations at an increment of rotation of 30^0.



ethyl linoleate

Figure S6. The effect of double bonds on the range of generated conformers, see text below.

The ends of the hydrocarbon chain may be coiled up on the backbone of the molecule, forming very high energy conformers or may instead be fully extended forming low energy conformers. In either case the middle dihedral angles are the same and their frequent appearance in the high energy group may cause their elimination and loss of the low energy conformers associated with them. Therefore this molecule in which the middle dihedral angle values contribute to both high and low energy groups represents a challenge to ISE and could be solved only with the correction procedure mentioned in the methods section 2.5. This procedure returns dihedral angle values which have high appearance in both high and low energy groups and low appearance in the middle energy group. At the beginning of the calculation the middle group contains high energy conformations with conformational clashes, so these conformations cannot be considered middle energy conformations. As the calculation progresses, high energy conformations that have only high energy contributors are eliminated and more conformations having both high and low energy contributors are sampled. This causes a significant reduction of energy in the middle group (see column C in figure S7 below).

A	B	C	D
1	31.878	1.74E+07	1.20E+12
2	17.528	1.57E+07	1.13E+12
3	30.621	2.37E+07	9.61E+11
4	31.663	1.39E+07	1.97E+12
5	19.498	2.53E+07	1.12E+12
6	26.809	2.07E+07	1.75E+12
7	22.666	8.94E+06	1.11E+12
8	24.816	2.72E+07	1.18E+12
9	25.997	1.54E+07	1.22E+12
10	23.418	1.27E+07	1.20E+12
11	27.214	7.50E+06	9.73E+11
12	26.352	1.08E+06	1.40E+12
13	21.196	247161	1.20E+12
14	18.013	66941.6	1.02E+12
15	20.37	36921.7	6.73E+11
16	22.272	20233.1	9.27E+11
17	25.744	6209.84	6.21E+11
18	21.377	4214.15	5.08E+11
19	18.249	2262.76	1.22E+12
20	21.018	685.253	8.78E+11
21	22.059	1123.14	1.12E+12
22	16.56	739.085	7.53E+11
23	20.994	787.056	6.14E+11

Figure S7. Mean energies of low conformers (column B), middle conformers (column C) and high energy conformations (column D) as the calculation proceeds. Column A shows the ation number. All values are in Kcal/mol.

This is the activation stage for the returning function which begins measuring the number of appearances of dihedral angle values when the mean energy in the middle group is about 50 times the size of the mean energy in the low group. From this stage on appearances are added and the lowest possibilities are subtracted and saved for the next iteration, hence this function learns more low energy possibilities and becomes more efficient at its job – returning conformations which are low in the middle group and high in both low and high energy groups.

The possibility of loosing contributors to both high and low energy conformers exists mainly in the stage before the returning function is activated and in the initial stages of its activation when the returning function has not learnt enough possibilities so the

low possibilities used for returning high-low contributors are not close to the lowest. Therefore the returning function is implemented with relatively high values for repeated criteria check before elimination (CR >= 10). See figure S8 for a graphic summary.

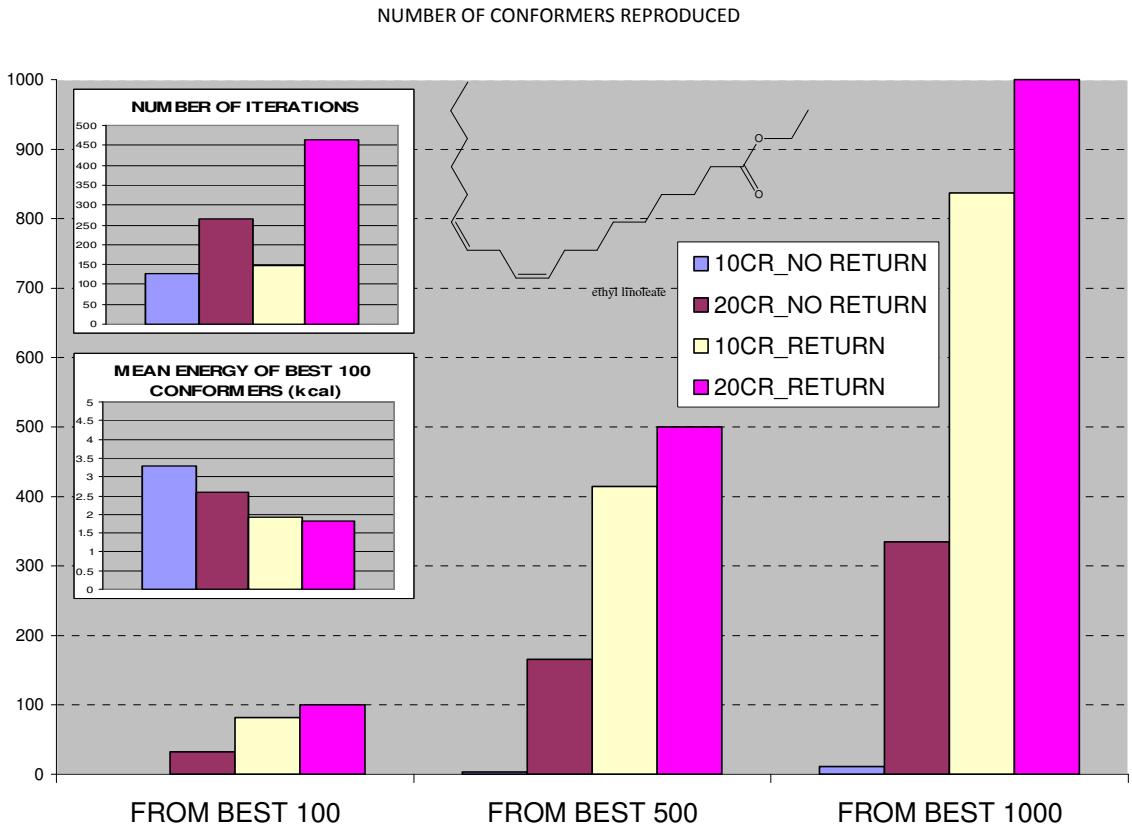


Figure S8. Reproduction of lowest energy conformers of ethyl linoleate from by the ISE method with repeated elimination criteria and erroneous elimination return

Figure S8 shows the reproduction of best 100, 500, and 1000 conformations found when results reached convergence and could not further be improved. Each calculation was repeated 3 times for consistency. The returning function lowered the mean energy of the lowest 100 conformations giving the best results at the highest repeated criteria check (CR=20). This however was achieved at the cost of about 450 iterations. Recalling that iteration time is small (about 0.02 seconds per iteration) these conditions are not unreasonable, and the whole calculation takes about 24 hours, depending on force-field calculation time.

We also note that differences between energies in different final ensembles is minimal (less than 1 kcal even when all conformers in the ensemble are different).

2. Choice of charge computation scheme for computing dipole moments

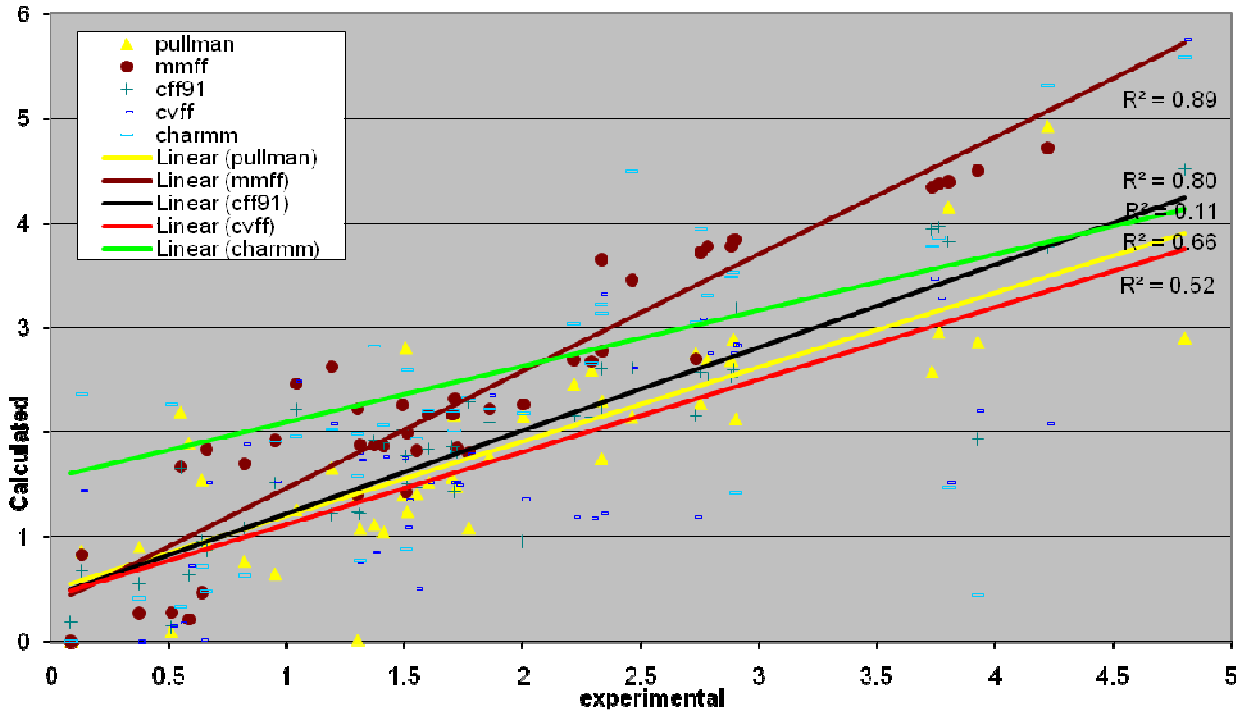


Figure S9: Correlation of dipole moments of non-flexible molecules computed using different charge methods to the experimental dipole moment values. The best line, correctly predicting about 90% of the dipole moments, was found to be associated with MMFF charges.