1-1-2015

Genetic Algorithm Optimization of Point Charges in Force Field Development: Challenges and Insights

Maxim Vadimovich Ivanov
Marquette University

Marat R. Talipov
Marquette University, marat.talipov@marquette.edu

Qadir K. Timerghazin
Marquette University, qadir.timerghazin@marquette.edu

Accepted version. *Journal of Physical Chemistry A,* Vol. 119, No. 8 (2015): 1422-1434. DOI. © 2015 American Chemical Society. Used with permission.

Genetic Algorithm Optimization of Point Charges in Force Field Development: Challenges and Insights

Maxim V. Ivanov

*Department of Chemistry, Marquette University,
Milwaukee, WI*

Marat R. Talipov

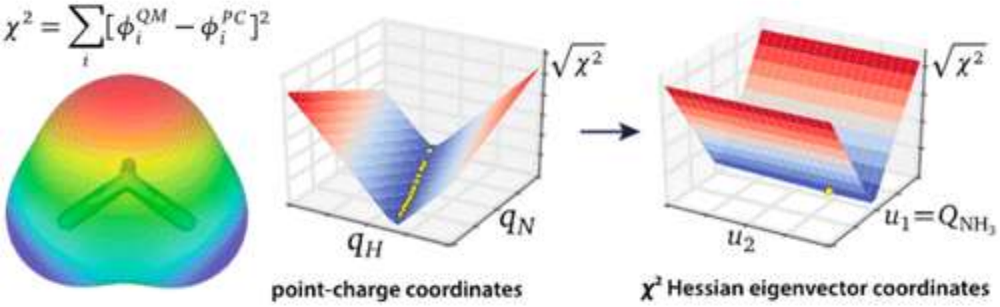
*Department of Chemistry, Marquette University,
Milwaukee, WI*

Qadir K. Timerghazin

*Department of Chemistry, Marquette University,
Milwaukee, WI*

Journal of Physical Chemistry A, Vol 119, No. 8 (2015): pg. 1422-1434. DOI.

Abstract


$\chi^2 = \sum_i [\phi_i^{QM} - \phi_i^{PC}]^2$
$\sqrt{\chi^2}$
q_H
q_N
point-charge coordinates
$\sqrt{\chi^2}$
u_2
$u_1 = Q_{NH_3}$
χ^2 Hessian eigenvector coordinates

Evolutionary methods, such as genetic algorithms (GAs), provide powerful tools for optimization of the force field parameters, especially in the case of simultaneous fitting of the force field terms against extensive reference data. However, GA fitting of the nonbonded interaction parameters that includes point charges has not been explored in the literature, likely due to numerous difficulties with even a simpler problem of the least-squares fitting of the atomic point charges against a reference molecular electrostatic potential (MEP), which often demonstrates an unusually high variation of the fitted charges on buried atoms. Here, we examine the performance of the GA approach for the least-squares MEP point charge fitting, and show that the GA optimizations suffer from a magnified version of the classical buried atom effect, producing highly scattered yet correlated solutions. This effect can be understood in terms of the linearly independent, natural coordinates of the MEP fitting problem defined by the eigenvectors of the least-squares sum Hessian matrix, which are also equivalent to the eigenvectors of the covariance matrix evaluated for the scattered GA solutions. GAs quickly converge with respect to the high-curvature coordinates defined by the eigenvectors related to the leading terms of the multipole expansion, but have difficulty converging with respect to the low-curvature coordinates that mostly depend on the buried atom charges. The performance of the evolutionary techniques dramatically improves when the point charge optimization is performed using the Hessian or covariance matrix eigenvectors, an approach with a significant potential for the evolutionary optimization of the fixed-charge biomolecular force fields.

1 Introduction

Molecular dynamics (MD) simulations is a powerful tool to study structure and function of biological macromolecules at the atomic level.[1-3] The accuracy of MD simulations is highly dependent on the molecular mechanics force field used—its functional form, as well as its empirical parameters. In traditional macromolecular all-atom force fields, the bonded parameters include equilibrium bond distances, bond and dihedral angles, along with the corresponding force constants and rotation barriers, while nonbonded interactions are typically described by atom-centered point charges and Lennard-Jones parameters. These bonded and nonbonded force field parameters are fitted against either experimental data or, more commonly, data obtained from electronic structure calculations. Generally, force field parametrization involves separate optimization of the bonded and nonbonded parameters, as it is common in parametrization of the classical force field models such as CHARMM,[4-6] AMBER,[7-9] GROMOS,[10] and OPLS,[11,12] as well as in more recent developments.[13-17] For instance, in parametrization of the nonbonded terms in the popular AMBER family of force fields,[7,18,19] the point charges are fitted to the reference molecular electrostatic potential (MEP) of the molecule, while Lennard-Jones parameters are fitted to reproduce the experimental bulk properties. However, simultaneous fitting of several parameters describing intermolecular interactions (point charges, Lennard-Jones parameters, and in the case of polarizable force fields, atomic polarizabilities) may significantly improve the accuracy of force field description.[20,21] These simultaneous optimizations of different force field terms can take advantage of extensive training sets that can be easily generated using electronic structure calculations and may include data on the intermolecular interaction energies.[22-26] Moreover, in this approach the fitted interaction energy would implicitly include the polarization effects, even staying within the fixed point-charge force field framework.[9,27,28] However, such simultaneous force field fitting represents a technically challenging multiobjective optimization of the parameters of different physical nature.

Among various optimization algorithms available for this purpose, evolutionary methods such as genetic algorithms (GAs) provide a powerful technique that can efficiently deal with complex and poorly understood search space.[29-33] GAs have been successfully used in force field development, including fitting of dihedral angle[34,35] and van der Waals[17,25] parameters, atomic polarizabilities,[16] parametrization of coarse-grained[36] and reactive[37,38] force fields, and applied in numerous *ad hoc* force field parameter optimizations.[39-43] Interestingly, although the assignment of the fixed point charges is a critical part of many force fields, the application of GAs and other evolutionary/stochastic optimization techniques to the MEP point-charge fitting has not been explored, to the best of our knowledge.

The traditional approach for determining point charges in the force field development, usually referred to as the ESP (electrostatic potential) method,[44] is to fit the point charges against the reference quantum mechanical (QM) MEP φ^{QM} by minimizing the sum of squared residuals $\varphi^{QM}-\varphi^{PC}$ calculated over the *N* point on a grid:

$$\chi^2 = \sum_i^N [\varphi^{QM}(\vec{R}_i) - \varphi^{PC}(\vec{R}_i)]^2 \quad (1)$$

where φ^{PC} is the potential produced by the point charges:

$$\varphi^{PC}(\vec{R}) = \sum_j^M \frac{q_j}{|\vec{R} - \vec{r}_j|} \quad (2)$$

Examples of different implementations of this method include Merz–Kollman,[45,46] CHELP,[47] and CHELPG,[48] which mainly differ by the choice of the reference grid. These approaches typically employ Lagrange multipliers to impose a constraint on the overall molecular charge and, sometimes, on the molecular dipole moment. Alternatively, the χ^2 function can be minimized directly using gradient-based methods with restraint on the total charge and dipole moment.[49]

Although the atom-centered MEP-derived point charges provide a clear interpretation of the electrostatic properties and are computationally inexpensive, they can poorly reproduce the anisotropic electronic features (e.g., lone pairs, π-systems),[50,51] and also suffer from several technical difficulties. The optimized values of the point charges not only depend on the grid density and size, or the spatial orientation of the molecule relative to the Cartesian axes,[48,52-56] they also can be inconsistent even across very similar molecules, at odds with the fundamental chemical concept of the transferability of atomic properties. Not only the MEP-fitted charges for atoms of a common functional group in chemically similar molecules may be very different, the charges obtained for the conformers of the same molecule often vary by more than one electron unit. Stouch and Williams reported[57,58] that the disparate charges obtained for directly connected atoms in different conformers seem to linearly correlate with each other with high variation (~1.3 e^-) of the charge values on the interior, buried atoms (mostly aliphatic carbon atoms), while the exterior atoms (mostly hydrogens) vary in a much smaller range (~0.3 e^-). Later, the large variations of charge values have been rationalized by the low statistical contribution of the buried carbons to the overall electrostatic potential.[59] Furthermore, the ill-conditioned character of the MEP fitting problem seems to be exacerbated by the introduction of the total charge constraint using Lagrange multipliers that leads to the rank deficiency of the least-squares (LS) matrix.[53,60]

The conformational dependence of the MEP-derived point charges has been significantly reduced in the restrained electrostatic potential (RESP) method by Bayly et al.[59,61] that uses an external hyperbolic restraint to force the buried carbon atoms to have small point charges, thus decreasing the charge variations across different conformers. Although several alternative methods of charge derivation have been proposed,[47,53,60,62,63] restraining the charges of buried atoms to prevent the optimization from converging toward unreasonable values and/or to reduce conformational dependence of the charges became the most popular in force field development.[64-77] In most of these methods, besides a constraint on the total charge of the molecule, an additional restraining function is added to the LS sum (eq 1) to keep the buried atom charges close to some predefined values, despite its possible negative effect on the dipole moment values and the overall quality of MEP.[53,78]

Considering the challenges presented by the relatively straightforward single-objective point charge fitting against the MEP, simultaneous optimization of point charges along with other force field parameters against a diverse training set could be expected to present even more pitfalls. Therefore, in this work we investigate the performance of the GA techniques when applied to the MEP point charge fitting problem in a case of small model molecules with the emphasis on the convergence properties of the algorithm.

2 Details of Charge Fitting and Analysis Procedures

Reference MEP

All geometry optimizations were performed at the B3LYP/aug-cc-pVDZ level,[79,80] as implemented in the Gaussian 09 package.[81] Reference MEPs were generated as cubic grids with linear density of 2.8 points/Å, followed by removal of the points outside of 1.4–2.0 van der Waals radii range around each atom. This sampling procedure covers the solvent-accessible region of the molecule, in line with common charge fitting procedures.[59,60]

ESP Point Charge Fitting

In the ESP method the solution is obtained by minimizing the LS sum (eq 1) that can be rewritten in a more compact algebraic form:

$$\chi^2 = |\vec{\varphi} - \mathbf{A}\vec{q}|^2 = |\vec{\varphi}|^2 + \vec{g} \cdot \vec{q} + \vec{q}^{\,\mathrm{T}}\mathbf{H}\vec{q} \quad (3)$$

$$\vec{g} = -2\mathbf{A}^{\mathrm{T}}\vec{\varphi} \quad (4)$$

$$\mathbf{H} = \mathbf{A}^{\mathrm{T}}\mathbf{A} \quad (5)$$

where the vector $\vec{\varphi} = (\varphi^{\mathrm{QM}}(\vec{R}_1) \ldots \varphi^{\mathrm{QM}}(\vec{R}_N))$ consists of the reference electrostatic potential calculated at each point of the grid; $\vec{q} = (q_1 \ldots q_M)$ is a set of point charges; **A** is the LS matrix with the elements corresponding to the inverse distance $1/r_{ij}$ between point i of the grid and point charge j in the molecule; vector $\vec{g}$ and matrix **H** are gradient vector and Hessian matrix of the LS sum, correspondingly.

Because of the quadratic dependence of the LS sum on the charge vector $\vec{q}$ the solution to the LS problem can be found by setting partial derivatives of χ^2 with respect to each point charge to zero, which results in the system of linear equations, known as normal equations:[82]

$$\mathbf{A}^{\mathrm{T}}\mathbf{A}\vec{q}^{*} = \mathbf{A}^{\mathrm{T}}\vec{\varphi} \quad (6)$$

where q^{*} is the solution to the problem which is further referred to as ESP charges and used as the reference to compare against the GA-optimized values. No additional constraints or restraints have been imposed to these charges, except for the atom equivalence due to the symmetry of the molecule.

Table 1. Parameters and the Genetic Operators Used in the GA Fitting of the MEP Point Charges

parameter	description	value
maximum number of generations	convergence criterion	100
population size	number of chromosomes in the population	20–200
variable range	range of charge values used to generate a chromosome	[−1; 1]

operator	binary-coded	real-coded	probability
crossover	two-point[83]	BLX-α,α = 0.5[83]	0.90
mutation	flip bit[29,30]	random[31]	0.03
selection	proportional selection[29,30]		

Point Charge Fitting with Genetic Algorithms (GAs)

In the GA approach, each candidate solution is referred to as a chromosome or an individual. A set of chromosomes, called population, is evolving during a GA run through an iterative application of genetic operators of selection, crossover, and mutation.[29,30] Each chromosome in the population has an associated fitness function value, or a fitness score, that measures how close this candidate solution is to the desired optimum solution. The algorithm starts by randomly generating the initial population of the chromosomes, followed by evaluation of their fitness function values. These scores are then used to select chromosomes for further crossover and mutation that produce the next generation of the chromosomes. When the number of generations reaches a predefined maximum, the algorithm stops and the chromosome with the best fitness score in the final population is taken as the solution to the optimization problem.

The GA parameters used here for the point charge fitting against a reference MEP are given in Table 1. Each chromosome encoded a set of atom-centered point charges either in a traditional binary or real number representation. We found that, as in several other cases,[84,85] the real-number coding requires a smaller population size than the binary coding to achieve the results of the same quality (Figures S1–S2 in the Supporting Information). Therefore, the real-coded chromosomes were used throughout this work.

All point charges have been fitted within the −1 to +1 *e* range, with no additional restraints, unless stated otherwise. The root-mean square error (RMSE) was used as the fitness function:

$$f = \mathrm{RMSE} = \sqrt{\frac{\sum_{i}^{N} [V^{\mathrm{QM}}(\vec{R}_i) - V^{\mathrm{PC}}(\vec{R}_i)]^2}{N}} = \sqrt{\frac{\chi^2}{N}} \quad (7)$$

Thus, the chromosome with the lowest fitness score in the last generation was considered as the solution being sought. RMSE has been chosen as the fitness function because of its clear statistical meaning; however, using either the RMSE or the LS sum χ^2 (eq 1) as the fitness function in the GA optimizations gives very similar results. The average fitness score ⟨*f*⟩ of a population of size *S* calculated at each generation was used to characterize the convergence of a single GA run, while the standard deviation σ_f was used to characterize how diverse or localized are the chromosomes in the population:

$$\langle f \rangle = \frac{1}{S} \sum_{i}^{S} f_i \quad (8)$$

$$\sigma_f = \sqrt{\frac{1}{S} \sum_{i}^{S} (f_i - \langle f \rangle)^2} \quad (9)$$

Covariance Matrix Analysis of GA Solutions

Because of the stochastic nature of the algorithm, several independent GA runs were used to assess the quality/scatter of the obtained solutions. In most cases, several runs converged to a set of widely dispersed solutions. To understand the nature of this dispersion and reveal possible correlations between optimized parameters, we computed variance-covariance (or covariance) matrices Σ for each set of the obtained GA solutions. The diagonal elements of the covariance matrix contain the variances of the charges (eq 10) and the off-diagonal elements contain the covariances between each pair of charges (eq 11):

$$\mathrm{var}(q_j) = \frac{1}{N-1}\sum_{i}^{N}(q_{ij} - \langle q_j \rangle)^2 \quad (10)$$

$$\mathrm{cov}(q_j, q_k) = \frac{1}{N-1}\sum_{i}^{N}(q_{ij} - \langle q_j \rangle)(q_{ik} - \langle q_k \rangle) \quad (11)$$

where *N* is the number of GA runs, q_{ij} is the charge on atom *j* from *i*th GA run, $\langle q_j \rangle$ is charge on atom *j* averaged over all GA runs. Eigenvectors of the covariance matrix form an eigenbasis $\tilde{\Sigma}$ consisting of the orthonormal vectors $\vec{s}_i$ (principal components), along which the data are changing with the variance defined by the corresponding eigenvalue σ_i^2:

$$\Sigma \vec{s}_i = \sigma_i^2 \vec{s}_i \quad (12)$$

$$\tilde{\Sigma} = (\vec{s}_1 \ldots \vec{s}_M) \quad (13)$$

where $\tilde{\Sigma}$ is the square matrix of size *M*, defined by the number of point charges; σ_i is standard deviation along eigenvector $\vec{s}_i$.

Details of Implementation

All charge-fitting procedures were implemented using Python programming language within *fftoolbox* and *genetica* modules with the source code available online at the GitHub repository. The *fftoolbox* module extracts molecular geometry and the reference electrostatic potential from the Gaussian cube file and performs a calculation of the LS sum over the points in the grid. Besides the atom-centered point charges, *fftoolbox* also supports the optimization of the extra points placed out of the atomic centers. The ESP method (eqs 3–6) is implemented as a part of *fftoolbox* with the normal equation solved using the *numpy* library,[86] and the gradient-based optimization of the point charges is implemented using the Broyden, Fletcher, Goldfarb, Shanno (BFGS) quasi-Newton method using the *scipy* library.[86] GA optimization routines are implemented in the *genetica* module using either binary or real-number chromosome representation. The point charge optimization can be performed in three coordinate systems: point charges, multipole moments, or in the eigenbasis of the LS-sum Hessian matrix. Besides a single-objective minimization, *genetica* also supports vector-valued FFs using Vector Evaluated[87] GA (VEGA)—an extension of the single-objective GA method to support multiobjective optimizations. Covariance matrix adaptation evolution strategy (CMA-ES) optimizations were performed using the *cma* Python library;[88-91] in these optimizations, all values of the initial solution were set to zero and the initial standard deviation was set to 0.1. Covariance matrix calculations as well as all matrix eigendecompositions were performed using the *numpy* library. Graphical representation of the results is supported by the *matplotlib* library.[92]

3 GA Charge Fitting for Small Models

First, we examine the performance of GAs for the MEP point charge fitting in a straightforward case of several small molecules with only two symmetry-independent charges, but vastly different electrostatic properties: water, ammonia, benzene, and methane. For these systems, a single GA run with a small population size (<40 chromosomes) converges to a localized set of solutions within 25–50 generations, after which the population stabilizes with only small fluctuations of the charge values/fitness scores (Figure S1 in the Supporting Information). Surprisingly, although all GA runs demonstrate robust convergence, independent runs converge to vastly different solutions for the same molecule (Figure 1). For instance, 200 GA runs for CH_4 produced solutions with charges on the carbon atom q_C varying from −0.99 to 0.95 *e*, while the charge on the hydrogen varied from −0.24 to 0.25 *e*. Similar scatter of the small-population GA-derived charge values is observed for other molecules. In the case of H_2O, NH_3, and CH_4 the charges of the central, "buried" atoms show much larger deviations than the hydrogen atom charges. Although highly dispersed, the GA solutions tend to cluster around the solutions that correspond to the charges derived with the ESP method, eq 6 (shown as yellow dots in Figure 1). Increase of the population size decreases the scatter: GA runs with populations greater than 50 chromosomes yield solutions within ±0.01 *e* of the ESP values.

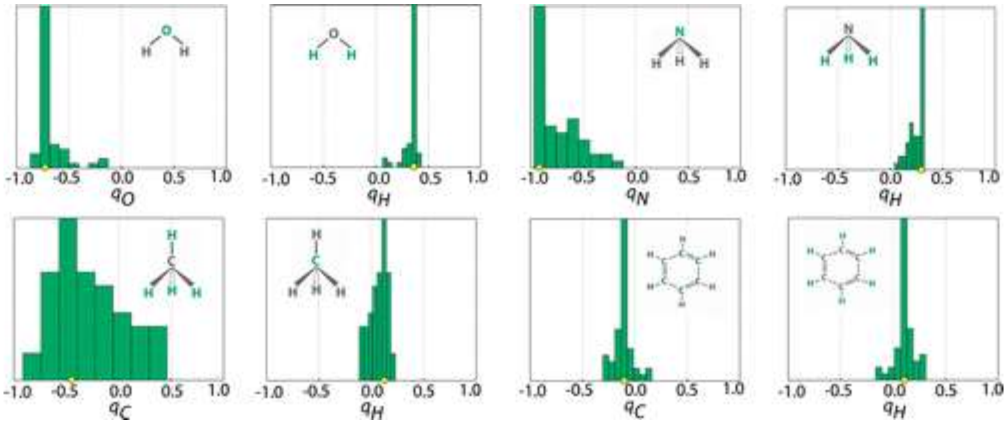


Figure 1. Distributions of the GA-optimized charges for the model molecules with two symmetry-independent charges, obtained from 200 GA runs with 20 chromosomes in the population. Yellow dots indicate the solutions obtained with the ESP method.

At first glance, these results simply suggest that the MEP point charge fitting with GAs is highly inefficient and requires larger population sizes. It is, however, intriguing why the small-population GA runs quickly converge to nonoptimal solutions that cannot be improved upon any further, even in hundreds of additional generations (premature convergence). In other words, what is the origin of these nonoptimal solutions that trap small-population GA runs? Further investigation revealed that there is a perfect (R^2 = 1.00) linear correlation between the pairs of q_X (X = O, N, or C) and q_H values produced from different GA runs (Figure 2). For each correlation, the slopes correspond to the number of hydrogen atoms per atom X in the molecule, while the intercept correspond to the overall charge Q = 0.0 *e* of the molecule:

$$Q = n_X q_X + n_H q_H \quad (14)$$

$$q_X = -\frac{n_H}{n_X} q_H + \frac{1}{n_X} Q \quad (15)$$

where n_X is the number of X atoms, and n_H/n_X is the number of hydrogen atoms per atom X. Indeed, although the GA runs converge to dispersed solutions, the zero total charge is always reproduced, with standard deviation σ = 0.001–0.01 *e*.

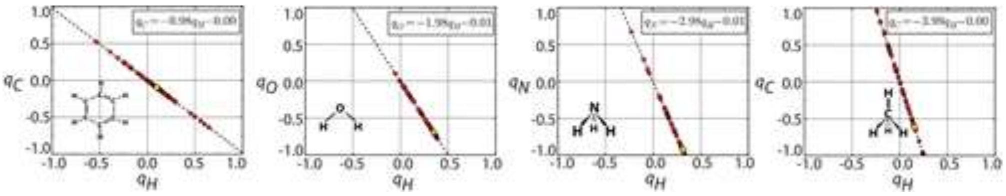


Figure 2. Correlations between the GA-optimized charges for the two-charge model molecules obtained from 200 independent GA runs with 20 chromosomes in the population; all trend lines have correlation coefficient R^2 = 1.00. Yellow dots indicate the solutions obtained with the ESP method.

We further investigated the GA-fitting performance for molecules with three symmetry-independent charges on the example of mono- and disubstituted methane derivatives CH_3X, X = F, Cl, O^-, and CH_2X_2, X = F, Cl. Similarly to the two-charge systems, multiple small-population GA runs (<100 chromosomes) yield highly scattered solutions, which tend to cluster around the ESP values as the population sizes increase. However, only GA runs with greater than 100 chromosomes yield consistent results that match the ESP charges within ±0.01 *e*. The scatter is the largest in the case of the charges on the carbon atoms q_C; for example, 200 30-chromosome GA runs for CH_3Cl produce q_C values covering the entire −1 to +1 *e* range, while the charge on hydrogen and chlorine vary in much smaller ranges (−0.1 to 0.3 *e* and −0.3 to −0.1 *e*, respectively).

Table 2. Average Values and the Standard Deviations (in Parentheses) of the Monopole and Dipole Moments Computed from the GA-Optimized Point Charges for CH_3X, CH_2X_2 (X = F, Cl), and CH_3O^- Molecules along with the Reference Values from DFT Calculations

molecule	**monopole (au)**	**dipole (au)**	**DFT dipole (au)**
CH_3F	0.002(0.001)	0.782(0.005)	0.771
CH_3Cl	0.000(0.002)	0.827(0.042)	0.794
CH_2F_2	−0.002(0.002)	0.814(0.087)	0.803
CH_2Cl_2	−0.001(0.003)	0.712(0.047)	0.667
CH_3O^-	−0.9674(0.006)	0.847(0.018)[a]	0.772[a]

[a] In the case of a charged CH_3O^- molecule, the dipole moment was calculated using the standard orientation of the spatial coordinates, as implemented in Gaussian 09.

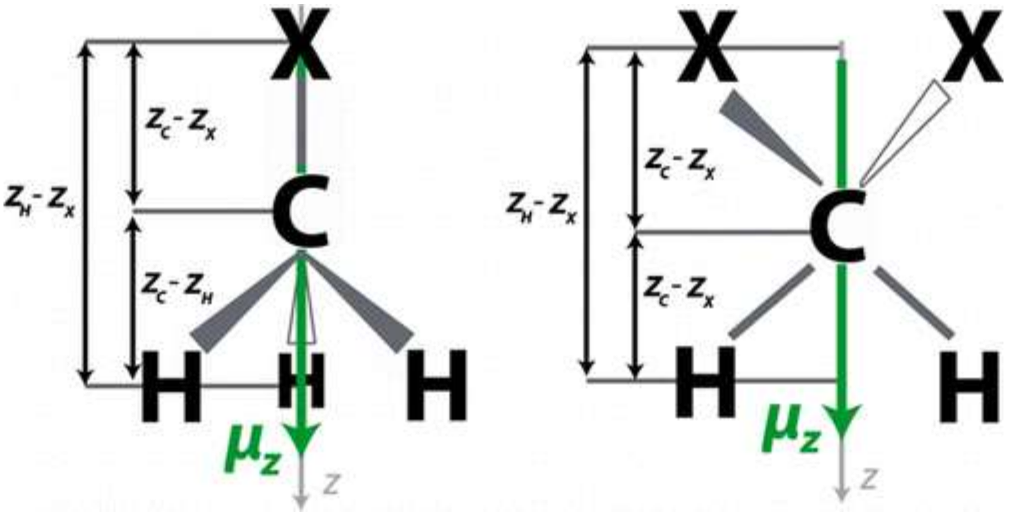

X
Z_c-Z_x
Z_H-Z_x
C
Z_c-Z_H
H
H
H
μz
z
X
X
Z_c-Z_x
Z_H-Z_x
C
Z_c-Z_x
H
H
μz
z

Figure 3. Coordinate system for the CH_3X and CH_2X_2 molecules used in eqs 17 to 19.

Unlike the two-charge systems, the GA solutions for CH_3X, and CH_2X_2 not only reproduce the correct total charge, but also produce constant dipole moment values, which are close to the reference DFT values (Table 2): the standard deviation σ is in 0.001–0.006 *e* range for the total charge and in 0.005–0.087 au range for the dipole moment. Thus, regardless of the population size, the GA-optimized point charges satisfy the eqs 16 and 17 for the first two terms of the multipole expansion: the monopole/total charge and the dipole moment. These equations can be written as dot products between the charge vector $\vec{q}$ and the corresponding vector $\vec{u}_i$:

$$Q = n_X q_X + n_C q_C + n_H q_H = \vec{u}_1 \cdot \vec{q} \quad (16)$$

$$\mu_Z = n_X z_X q_X + n_C z_C q_C + n_H z_H q_H = \vec{u}_2 \cdot \vec{q} \quad (17)$$

where n_A is the stoichiometric number of the atom *A* in the molecule, z_A is its coordinate along the *z* axis (oriented along the symmetry axis as shown in Figure 3), and q_A is its point charge. Geometrically, these equations define two planes with the vectors $\vec{u}_1$ and $\vec{u}_2$ which are orthogonal to the corresponding plane. The GA solutions align along a three-dimensional line formed by the intersection of these two planes (Figure 4A) which is defined by the cross product vector $\vec{u}_3 = \vec{u}_1 \times \vec{u}_2$:

$$\vec{q} = \vec{q}_0 + t\vec{u}_3 \quad (18)$$

where t is a free parameter, the vector $\vec{q}_0$ is a set of point charges that satisfies eqs 16 and 17. Projections of this three-dimensional line give three pairwise linear relationships between each pair of the atomic charges (Figure 4B, Figure S3 in the Supporting Information); for example, a projection on the (q_C, q_H) plane results in a linear correlation between q_C and q_H. These pairwise correlations can be derived using the geometric parameters (Figure 3) and dipole moment values:

$$q_C = -\frac{n_H}{n_C}\frac{z_H - z_X}{z_C - z_X}q_H + \frac{1}{n_C}\frac{\mu_z - Qz_X}{z_C - z_X} \quad (19)$$

Importantly, there is a good numerical agreement between the correlations obtained analytically using the DFT dipole moments and from the linear fitting of the scattered GA solutions (Table S2 in the Supporting Information). Thus, the linear relationships observed for the two- and three-independent charge systems arise because all GA solutions satisfy the constant total charge and (for the three-charge systems) the dipole moment requirements, while the higher multipole moments produced by these solutions are scattered.

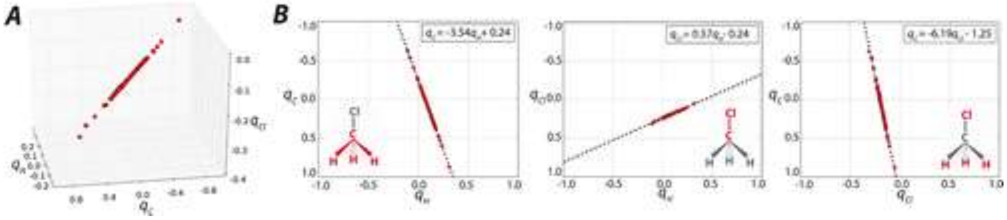


Figure 4. Correlation between the chloromethane point charges obtained from 200 independent GA runs shown in three dimensions (A) and as two-dimensional projections, i.e., pairwise correlations between charges (B).

4 Covariance Matrix Analysis of GA Results

In the trivial case of the two- and three-independent charge systems, the scattered nature of the small-population GA-optimized point charges can be interpreted using a simple correlation analysis (Figures 2 and 4). However, understanding the results for larger, more realistic molecules would require a more general approach, such as the analysis of the eigenvectors of the covariance matrix Σ computed for a set of GA solutions. We tested this approach by re-examining the small-population GA results for the two- and three-charged model systems discussed above.

For the two-charge molecules, the covariance matrix diagonalization (Table S3 in the Supporting Information) yields one vector with almost negligible variance/eigenvalue ($\sigma_1^2 < 10^{-5}$) and one vector with much higher variance ($\sigma_2^2 = 0.06–0.19$). The first vector $\vec{s}_1$, along which the data does not vary, numerically corresponds to the normalized vector $\vec{u}_1$ that defines the total charge and is determined by the stoichiometry of the molecule:$Q = n_X q_X + n_H q_H = \vec{u}_1 \cdot \vec{q}$ (20)where $\vec{u}_1 = (n_X\ n_H)$ and $\vec{q} = (q_X\ q_H)$. The second vector $\vec{s}_2$, that is, the vector along which the data show a significant variation, numerically corresponds to a normalized vector $\vec{u}_2 = (n_H\ -n_X)$, also determined by the stoichiometry. Thus, the eigenbasis of the covariance matrix $\tilde{\Sigma}$ can be represented as

$$\tilde{\Sigma} = (\vec{s}_1\ \vec{s}_2) = \begin{pmatrix} \dfrac{\vec{u}_1}{|\vec{u}_1|} & \dfrac{\vec{u}_2}{|\vec{u}_2|} \end{pmatrix} \quad (21)$$

The dramatic difference in the data variation along the two covariance eigenvectors suggests that the fitness function has very different curvatures along these two directions. This curvature of the fitness function can be examined explicitly by computing and diagonalizing its Hessian matrix or, for simplicity, the Hessian of the LS sum **H** (eq 5):[93]

$$\mathbf{H}\vec{h}_i = \kappa_i \vec{h}_i \quad (22)$$

$$\tilde{\mathbf{H}} = (\vec{h}_1 \ldots \vec{h}_M) \quad (23)$$

As can be seen from Figure 5 and Table S3 in the Supporting Information, the Hessian eigenbases $\tilde{\mathbf{H}}$ computed for all four two-charge molecules are numerically identical to the corresponding covariance matrix eigenbases $\tilde{\Sigma}$ and the basis of normalized vectors: $\vec{u}_i$, $\tilde{\mathbf{U}}$:

$$\tilde{\Sigma} = \tilde{H} = \tilde{U} = \begin{pmatrix} \frac{\vec{u}_1}{|\vec{u}_1|} & \frac{\vec{u}_2}{|\vec{u}_2|} \end{pmatrix} \quad (24)$$

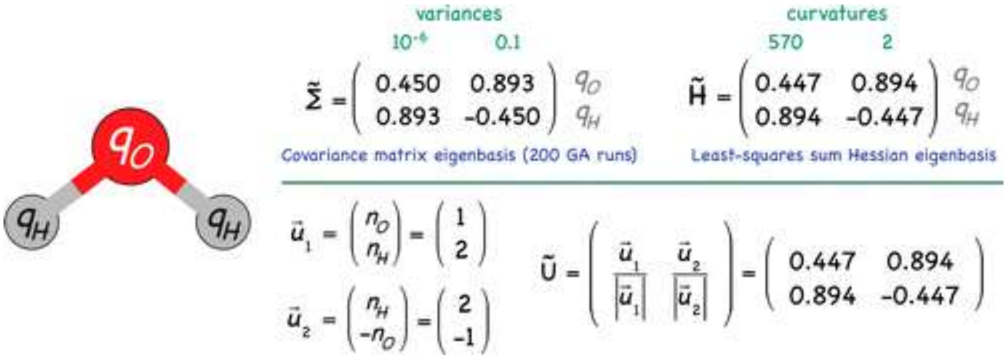

variances
10^{-6} 0.1
$\tilde{\Sigma} = \begin{pmatrix} 0.450 & 0.893 \\ 0.893 & -0.450 \end{pmatrix}$ q_O q_H
Covariance matrix eigenbasis (200 GA runs)
curvatures
570 2
$\tilde{H} = \begin{pmatrix} 0.447 & 0.894 \\ 0.894 & -0.447 \end{pmatrix}$ q_O q_H
Least-squares sum Hessian eigenbasis
q_O
q_H
q_H
$\vec{u}_1 = \begin{pmatrix} n_O \\ n_H \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$
$\vec{u}_2 = \begin{pmatrix} n_H \\ -n_O \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \end{pmatrix}$
$\tilde{U} = \begin{pmatrix} \frac{\vec{u}_1}{|\vec{u}_1|} & \frac{\vec{u}_2}{|\vec{u}_2|} \end{pmatrix} = \begin{pmatrix} 0.447 & 0.894 \\ 0.894 & -0.447 \end{pmatrix}$

Figure 5. Numerical equivalence of the eigenvectors of the covariance matrix for the results of 200 GA runs, the eigenvectors of the least-squares sum Hessian matrix, and the normalized vectors $\vec{u}_1$ and $\vec{u}_2$, on the example of a water molecule; for other model molecules, see Tables S3 and S4 in the Supporting Information.

There is an inverse relationship between the eigenvalues of the fitness function/LS sum Hessian and the covariance matrices: the Hessian eigenvector $\vec{h}_2$ with near-zero eigenvalue/curvature corresponds to the covariance eigenvector $\vec{s}_2$ with a large variance; at the same time, the Hessian eigenvector $\vec{h}_1$ with a large curvature corresponds to the covariance eigenvector $\vec{s}_1$ with near-zero variance. The latter high-curvature/small-variance vector is also the vector that defines the total charge of the molecule, $\vec{u}_1$ (eq 20). Thus, the linear correlations observed for the GA solutions (Figure 2) arise due to a high curvature of the fitness function with respect to the deviation of the total charge from the optimal value (zero for the studied molecules).

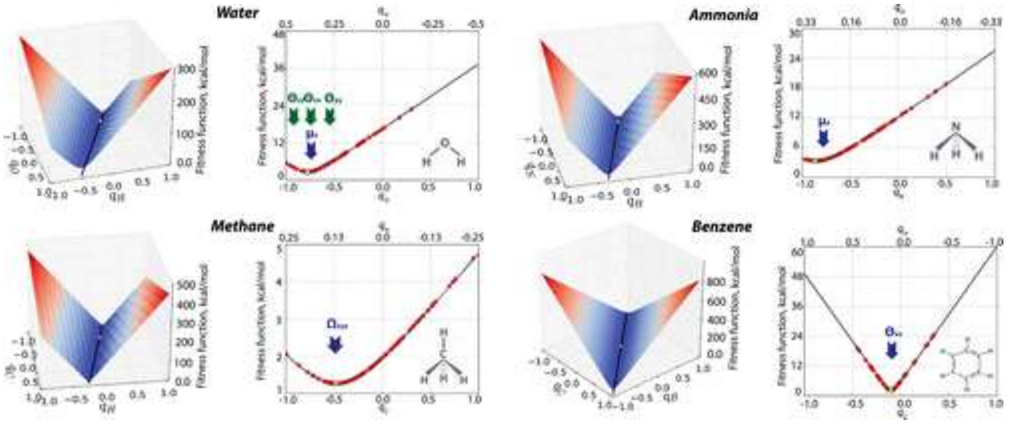


Figure 6. Fitness function profiles for the two-charge model molecules: full profiles (3D plots) and the profiles along the zero total charge line (2D plots). Red dots show the solutions obtained from GA optimizations (200 runs), and the yellow dots indicate the ESP solutions. The arrows indicate the solutions that reproduce the reference values of the corresponding nonvanishing multipole moment components from the reference DFT calculations.

The fitness function plots indeed show a dramatic difference in the curvatures (Figure 6): when plotted against q_X and q_H, the fitness function has a characteristic "V"-like shape, with the line of zero total charge going through the bottom of the valley (eq 20). As evident from the 3D plots, changing the central atom charge q_X from −1 to 1 *e* can result in up to 300–800 kcal/mol increase of the fitness function. At the same time, 2D profiles along the zero total charge line show 1–2 orders smaller variation of the fitness function values (<60 kcal/mol, note the difference in scales for the 3D and 2D plots in Figure 6). The actual minimum of the fitness function is determined by the next nonvanishing multipole moment(s) indicated by the positions of the arrows in Figure 6.

In the case of the three-charge model molecules CH_3X and CH_2X_2, diagonalization of the covariance matrices Σ of the scattered GA solutions yields two vectors, $\vec{s}_1$ and $\vec{s}_2$, along which the variance is negligible ($\sigma_{1,2}^2 < 10^{-5}$), and the third $\vec{s}_3$ with much larger variation of the data ($\sigma_3^2 = 0.1$–0.2). As the GA solutions conserve both the total charge Q and the dipole moment μ_z, we can expect that the $\vec{s}_1$ and $\vec{s}_2$ vectors correspond to the vectors $\vec{u}_1 = (n_X\ n_C\ n_H)$ and $\vec{u}_2 = (n_Xz_X\ n_Cz_C\ n_Hz_H)$, eqs 16 and 17, in which case the third vector $\vec{s}_3$ should be collinear with the cross product $\vec{u}_3 = \vec{u}_1 \times \vec{u}_2$, along which the GA solutions are distributed. Unlike the $\vec{s}_1$ and $\vec{s}_2$ vectors, the $\vec{u}_1$ and $\vec{u}_2$

vectors are generally not orthogonal, but their orthogonality can be achieved by appropriately shifting the coordinate origin:

$$\vec{u}_1 \cdot \vec{u}_2 = 0 \quad (25)$$

$$n_{\mathrm{X}}^2(z_{\mathrm{X}} - z_0) + n_{\mathrm{C}}^2(z_{\mathrm{C}} - z_0) + n_{\mathrm{H}}{}^2(z_{\mathrm{H}} - z_0) = 0 \quad (26)$$

$$z_0 = \frac{n_{\mathrm{X}}{}^2 z_{\mathrm{X}} + n_{\mathrm{C}}{}^2 z_{\mathrm{C}} + n_{\mathrm{H}}{}^2 z_{\mathrm{H}}}{n_{\mathrm{X}}{}^2 + n_{\mathrm{C}}{}^2 + n_{\mathrm{H}}{}^2} \quad (27)$$

where z_0 is the coordinate of the new origin along the z axis.
As expected, the set of the three orthogonal vectors $u\vec{}_i$:

$$(\vec{u}_1 \;\; \vec{u}_2 \;\; \vec{u}_3) = \begin{pmatrix} n_{\mathrm{X}} & n_{\mathrm{X}}(z_{\mathrm{X}} - z_0) & n_{\mathrm{C}} n_{\mathrm{H}}(z_{\mathrm{C}} - z_{\mathrm{H}}) \\ n_{\mathrm{C}} & n_{\mathrm{C}}(z_{\mathrm{C}} - z_0) & n_{\mathrm{H}} n_{\mathrm{X}}(z_{\mathrm{H}} - z_{\mathrm{X}}) \\ n_{\mathrm{H}} & n_{\mathrm{H}}(z_{\mathrm{H}} - z_0) & n_{\mathrm{X}} n_{\mathrm{C}}(z_{\mathrm{X}} - z_{\mathrm{C}}) \end{pmatrix} \quad (28)$$

numerically matches, after normalization, with the eigenbasis of the corresponding covariance matrix of the GA solutions Σ and the eigenbasis of the LS sum Hessian matrix **H** (Table S4 in the Supporting Information):

$$\hat{\Sigma} = \hat{\mathrm{H}} = \hat{\mathrm{U}} = \begin{pmatrix} \frac{\vec{u}_1}{|\vec{u}_1|} & \frac{\vec{u}_2}{|\vec{u}_2|} & \frac{\vec{u}_3}{|\vec{u}_3|} \end{pmatrix} \quad (29)$$

Thus, analysis of the covariance matrix provides a convenient and general method to understand the nature of the premature convergence of the small-population GA point charge optimizations that yields highly dispersed suboptimal solutions.

5 Rotation of the Optimization Coordinates

As we've seen, GA optimizations of point charges tend to quickly converge with respect to the leading terms of the multipole expansion associated with large curvature of the LS sum, but have difficulty navigating toward the minima along the other directions defined by the Hessian eigenvectors associated with small curvatures. Thus, the

Hessian/covariance matrix eigenvectors provide a set of linearly independent, natural coordinates expressed as linear combinations of the point charge coordinates. The latter, on the other hand, represent a linearly dependent set of coordinates for the fitness function minimization problem.

In fact, optimization in a rotated coordinate system is known to dramatically deteriorate the GA convergence.[94] This can be illustrated on the example of minimization of a simple function of two variables (Figure 7A) that has a low curvature along the *x*-axis and much higher curvature along the *y*-axis, resulting in a "V"-shaped surface similar to the fitness function of the two-charge systems (Figure 6). This model function does not present a problem for GA optimization in terms of the linearly independent parameters *x* and *y*, as written in Figure 7A: all GA runs quickly converge to the true minimum (zero standard deviation of the GA solutions). However, if the coordinate system is rotated by angle θ relative to the original axes (Figure 7B), the GA performance significantly deteriorates, as is evident from the increasing standard deviation, which reaches the maximum for θ = 45° (Figure 7C).

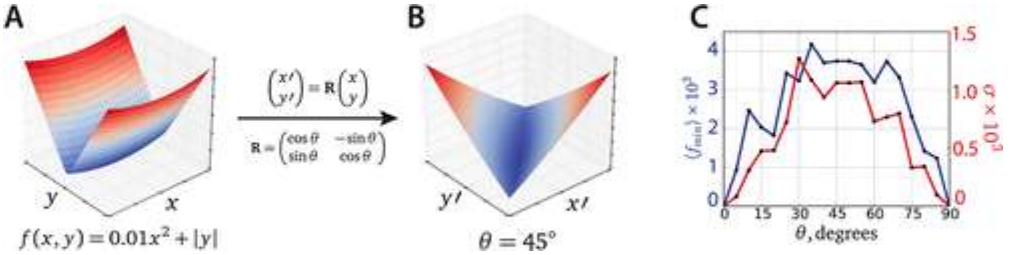


Figure 7. Effect of coordinate rotation on the convergence of GA minimizations on the example of a simple model function *f* of two variables associated with highly different curvatures: the model function plotted in the original coordinate system (A) and in the coordinate system rotated by 45° (B); the average $f_{\min}$ values obtained from 50 GA minimization runs (blue) and the corresponding standard deviations (red) vs the rotation angle θ.

This effect can be understood in terms of the high selective pressure along the high-curvature component *y*. The first chromosome to reach the minimum along *y*, that is, the line at the bottom of the valley, will quickly dominate the entire GA population; any new chromosome that even slightly deviates in the high-curvature direction incurs high fitness penalty and is not propagated to the next generation. In the original nonrotated coordinate system, the population is free to explore various values of the low-curvature parameter *x* without straying away from the bottom of the valley along the coordinate *y*. However, in the case of a rotated coordinate system, the population would produce a viable offspring in the direction of the global minimum only if both linearly dependent variables *x*′ and *y*′ change in a precise way to stay at the bottom of the valley. Since this is a low-probability event for a small population, the population stops changing once it reaches the minimum along the high-curvature direction, even though it may be far from the minimum along the low-curvature direction.

Table 3. Charge Fitting for Two- and Three-Charge Model Molecules: Average Fitness Scores with Standard Deviations (in Parentheses) for the GA Optimizations Using the Point Charge Coordinates vs the Coordinates Defined by the LS-Sum Hessian Eigenvectors, along with the Fitness Scores of the Reference ESP Solutions; All Units Are in kcal/mol

	200 GA runs, 30 chromosomes		
molecule	**point-charge coordinates**	**eigenvector coordinates**	**ESP**
H_2O	2.91(1.02)	$2.66(5.34 \times 10^{-6})$	2.66
NH_3	3.89(1.06)	$3.34(1.06 \times 10^{-5})$	3.34
C_6H_6	2.82(1.83)	$2.15(1.50 \times 10^{-5})$	2.15
CH_4	1.66(0.57)	$1.27(1.30 \times 10^{-6})$	1.27
CH_3Cl	2.46(0.41)	$2.14(4.84 \times 10^{-2})$	2.14
CH_2Cl_2	2.79(0.42)	$2.46(1.95 \times 10^{-5})$	2.46
CH_3F	2.26(0.41)	$1.89(3.06 \times 10^{-5})$	1.89
CH_2F_2	2.35(0.64)	$1.84(2.17 \times 10^{-5})$	1.84
CH_3O^-	4.71(0.98)	$3.61(5.02 \times 10^{-5})$	3.61

This population stagnation/premature convergence of the GA optimizations in rotated coordinate systems can be overcome by using large populations and/or higher mutation rates, which can lead to a significant computational cost. A more appealing solution is to perform the optimization in linearly independent coordinates defined by the eigenbasis of the LS-sum Hessian $\tilde{\mathbf{H}}$. In this case, the chromosomes encode a vector $\vec{n}$ of *M* real numbers—the optimization coordinates in the basis $\tilde{\mathbf{H}}$, while the fitness function is still evaluated in terms of the point charges $\vec{q}$ (eq 7) obtained using a linear transformation:

$$\vec{q} = \tilde{\mathbf{H}}\vec{n} \quad (30)$$

We tested this approach for the same two- and three-charge model molecules discussed above. With other GA parameters kept unchanged, optimizations in the new coordinate system demonstrated a much more robust convergence, as they require less than a half of the population size to achieve results of the same accuracy. For example, in the case of the three-charge CH_3X and CH_2X_2 molecules, 30 chromosomes were sufficient to converge to solutions that match the ESP charges within ±0.01 *e*, and to completely eliminate the linear correlations observed for the direct point charge optimizations (Table 3, Figure S4 in the Supporting Information).

Thus, the efficiency of the point charge fitting using GAs can be dramatically improved by rotating the optimization coordinates using the eigenvectors of the LS-sum Hessian. This finding, however, seems of little practical value by itself. Indeed, more efficient methods, such as ESP, exist for simple point charge fitting against the MEP. On the other hand, in a more complex case of simultaneous optimization of the point charges along with other force field parameters, evaluation of the fitness function Hessian could be much more problematic. However, as we already discussed, the covariance matrix of the GA solutions is numerically equivalent to the Hessian, and, in fact, this useful property of the covariance matrices is utilized in some recently developed advanced evolutionary methods such as the covariance matrix adaptation evolution strategy (CMA-ES) approach.[88-91]

Like other evolutionary strategy (ES) techniques, CMA-ES differs from less sophisticated classical GA methods in the implementation of the crossover and mutation operations; in some cases (CMA-ES included), new candidate solutions/offspring are sampled from the multivariate normal distribution, rather than produced by the traditional crossover operator. However, the most important CMA-ES feature in the context of this discussion is that a new set of solutions is generated using an approximate covariance matrix, which is updated at every step of the optimization. In this respect, CMA-ES is highly reminiscent of the quasi-Newton optimization techniques that use an approximate Hessian matrix which is updated at every step. Thus, although the classical GA approaches do not seem to hold much promise for simultaneous fitting of the force field parameters together with point charge values, more sophisticated evolutionary methods like CMA-ES may prove successful in this endeavor.

6 Real-Life Example: Charge Fitting for 1-Chlorobutane

We tested the performance of the GA and CMA-ES methods for the point-charge fitting problem in the case of five conformers of 1-chlorobutane, a more realistic example than the two- and three-charge models discussed so far. In line with the assumptions made in the force field development, the hydrogen atoms within each methyl and methylene group were considered equivalent, giving 9 point charge values overall to optimize for each conformer; the point charges were fitted separately for each conformer. In each case, 200 GA runs with populations of 200 chromosomes expectedly produced highly scattered solutions with the average fitness score significantly higher than that of the reference ESP solutions (Table 4). However, just like in the case of the small models, the GA solutions consistently reproduce the total charge and the magnitude of the dipole moment (Table S5 in the Supporting Information); also, there is a very good correspondence between the eigenvectors of the covariance matrix of the GA solutions and the LS sum Hessian (Table S6 and Figure S5 in the Supporting Information).

The eigenvector that corresponds to the highest curvature (~3600) and the smallest variance (~10^{-6}) corresponds to the total charge; it is identical for all conformers. While in the case of a large molecule such as 1-chlorobutane it is less straightforward to derive analytical expressions for the other high-curvature/low-variance eigenvectors, they seem to correspond to the leading multipole moments—the correspondence which is especially clear for the second highest-curvature vector (curvature ~200; variance ~10^{-5}) that defines the main dipole moment component (Figure S5 in the Supporting Information).[95] As the curvature decreases, the physical interpretation of the associated eigenvectors becomes less clear, and the similarity between the eigenvectors calculated for different conformers decreases, reflecting different electrostatic properties of these conformers. The last four eigenvectors have curvatures in the 0.3–0.03 range and correspondingly large variances, ~10^{-2}–10^{-1}.

These low-curvature/high-variance coordinates have a small contribution to the overall MEP, do not seem to be associated with particular multipole moments, and primarily depend on the charges of the buried carbon atoms (Figure S5 in the Supporting Information).

Table 4. Charge Fitting for 1-Chlorobutane Conformers: Average Fitness Scores with Standard Deviations (in Parentheses) for the GA Optimizations Using Two Coordinate Systems, Along with the Fitness Scores of the CMA-ES and ESP Solutions; All Units Are in kcal/mol

	200 GA runs, 200 chromosomes			
conformation	**point-charge coordinates**	**eigenvector coordinates**	**CMA-ES**	**ESP**
anti 1	3.05(0.43)	2.58(0.20)	2.09	2.09
anti 2	3.02(0.41)	2.57(0.19)	2.13	2.13
gauche 1	3.06(0.45)	2.61(0.21)	2.12	2.12
gauche 2	3.08(0.45)	2.58(0.19)	2.10	2.10
gauche 3	3.15(0.49)	2.62(0.18)	2.14	2.14

The GA optimizations in terms of the variables defined by the LS-sum Hessian eigenvectors yielded solutions with much better fitness scores (Table 4) and significantly decreased the scatter of the solutions (Figure S6 in the Supporting Information). At the same time, multiple CMA-ES runs converged to the identical solutions, which are also equal—within more than five decimal places—to the ESP values. The superb performance of CMA-ES method in this test case suggests that it could be a promising global-search evolutionary technique for force field development; a detailed discussion of the CMA-ES performance for simultaneous optimization of nonbonded force field parameters for several model systems will be reported elsewhere.

7 Variance of the Least-Squares Solution, Hessian Eigenvalues, and the Buried Atom Effect

Besides their importance for the application of evolutionary methods in the force field development, the insights into the severe convergence problems of the point charge fitting using classical GA methods can also be useful to revisit some of the well-known issues with the ESP method. The ESP charges can vary depending on the grid setup, and often are highly inconsistent for even slightly different conformers of the same molecule; the variation is especially large for the carbon atoms of methyl and methylene groups—the *buried atom* effect. These difficulties, commonly ascribed to the rank-deficient character of the LS matrix,[53,60] can be understood in a new light once we recognize that the variation of the ESP solutions has the same underlying factors as the much larger scatter of the GA solutions.

In fact, all LS fitting problems, not just the ESP, produce slightly different solutions from the LS matrices **A** that differ by the number of grid points, type of the grid, its density, etc. The covariance of these solutions, q^{*}, has been shown to be proportional to the inverse of the Hessian matrix:[96]

$$\mathrm{cov}(\overrightarrow{q^*}) \propto \mathbf{H}^{-1} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1} \quad (31)$$

Since a matrix inversion does not change the corresponding eigenvectors, this covariance matrix also shares the eigenbasis $\tilde{\mathbf{U}}$ with the covariance matrix of the GA solutions (e.g., eq 29). Thus, the variance/scatter of the ESP and GA solutions are related to the same fundamental properties of the LS-sum Hessian matrix, whose eigenvectors $\vec{h}_1$ define the natural, linearly independent coordinates for the MEP fitting problem. This provides a convenient framework to discuss the ill-conditioned nature of the ESP problem, and the buried atom effect associated with it.

The numerical instabilities observed for the standard ESP implementations can be related to the LS-sum Hessian eigenvectors with the highest and the lowest curvatures. For any molecule, the first eigenvector $\vec{h}_1$ defines the total charge coordinate, and the curvature along this coordinate is orders of magnitude larger than the curvatures along other coordinates. Hence, a very strong total charge restraint is naturally built into the ESP problem. Nevertheless, most of the ESP implementations introduce an additional total charge constraint using Lagrange multipliers,[45-48] a redundancy that leads to the known rank-deficiency of the resulting LS matrix.[53,55,60,62] On the other hand, optimization in the eigenmode coordinates with the coordinate along the $\vec{h}_1$ vector set to a desired value (e.g., 0 or −1 *e*) provides a straightforward and natural way to ensure the exact overall charge of the molecule.

On the other hand, the vexing problem of the buried atoms arises as a natural consequence of the high-variance coordinates with curvatures many orders of magnitude smaller than the curvatures of the coordinates associated with the leading multipole moments. These low-curvature/high-variance coordinates have a small contribution to the MEP and do not significantly affect the overall fitness of a solution. Thus, several solutions can have very similar fitness scores because they have the same positions along the high-curvature coordinates, although their positions along the low-curvature coordinates could be quite different. Yet, these very similar solutions would appear very different when expressed in terms of the linearly dependent point-charge coordinates.

Importantly, the lowest-curvature/highest-variation eigenvectors have the dominant contributions from the charges on the buried carbon atoms, as can be seen in the case of the CH_3X and CH_2X_2 molecules and the 1-chlorobutane conformers (Tables S4 and S6, and Figure S5 in the Supporting Information). As a result, these carbon atoms show the highest variation of the point charges—either ESP or GA-optimized—which is further amplified by the hydrogen/carbon stoichiometric ratios for the CH_3 and CH_2 groups, when the charge equivalence is applied to the hydrogen atoms. The usual approach to prevent the wide variation of the ESP charges on the buried carbon atoms is to use additional restraints to keep these charges close to a predefined value, such as zero,[59] or simply to constrain them to zero[55] or some chemically reasonable value.[57] This, however, can negatively affect the overall dipole moment values produced by the fitted point charges, as well as the overall quality of the fit;[53,55,78] a better strategy may involve restraining or constraining the values along the low-curvature Hessian eigenmode coordinates.

8 Conclusions

Motivated by the idea of using evolutionary approaches for the simultaneous optimizations of several types of force field parameters—including point charges, we explored the performance of the genetic algorithm (GA) approach for a simpler problem of point-charge fitting against the reference molecular electrostatic potential (MEP). We find that unless unreasonably large population sizes are used, the GA optimizations produce highly scattered, but correlated, solutions. Analysis of the covariance matrices for these scattered sets of GA solutions revealed a remarkable correspondence between the covariance matrices and the fitness function Hessian matrix, which share the same set of the eigenvectors. This eigenbasis represents a linearly independent set of coordinates that are natural for the MEP point-charge fitting problem, unlike the linearly dependent point charge coordinates. Some of the Hessian/covariance matrix eigenvectors define the coordinates related to the leading terms of the multipole expansion (the total charge/monopole, dipole moment components); these coordinates are associated with high curvature of the fitness function and thus negligible variation of the GA solutions. On the other hand, other eigenvectors are associated with negligible fitness function curvatures and thus large variance.

The huge disparity between the curvatures of the Hessian eigenvector coordinates causes premature convergence of the GA optimizations performed in terms of the linearly dependent point-charge coordinates, because of the high fitness penalty for even a slight deviation from the minimum along the high-curvature direction that effectively prevents the GA population from exploring the fitness profile along the low-curvature direction. This leads to a variety of GA solutions with highly scattered point charge values and moderately low, but not always optimal fitness scores. The severe scatter of the GA solutions can be seen as an exaggerated version of the well-known buried atom effect, the variation of the ESP charges of the buried carbon atoms observed for different grid setups and/or for different conformers.[97] This effect arises from the coordinates defined by the low-curvature Hessian eigenvectors and the fact that the point charges are inappropriate, highly linearly dependent (and also redundant)[55] coordinates for the MEP fitting problem. Thus, MEP fitting in coordinates defined by the fitness function/LS-sum Hessian eigenbasis is essential when using evolutionary methods. In this respect, the most promising approach is to take advantage of the correspondence between the eigenvectors of the covariance matrix of the solutions and the fitness function Hessian matrix, as it is done in advanced evolutionary techniques such as covariance matrix adaptation evolution strategy (CMA-ES).

Besides not being proper quantum mechanically observed parameters, atom-centered point charges are not even proper variables for the classical MEP fitting problem. At the same time, the simplicity and efficiency of the point charge model ensures its continuing survival in the field of the biomolecular simulations, at least in the short term.[9,27,28,98] Thus, the insights revealed by the analysis of the GA performance for the point charge fitting problem could prove useful for the further development and parametrization of the biomolecular force fields using evolutionary methods, as well as other optimization techniques.

The authors declare no competing financial interest.

Acknowledgment

Q.K.T. is a recipient of the National Science Foundation (NSF) CAREER award CHE-1255641, and Marquette University Way-Klinger Young Scholar Award. M.V.I. is a recipient of the Bournique Memorial Fellowship (Marquette University). This work used the high-performance computing cluster *Père* at Marquette University funded by NSF awards OCI-0923037 and CBET-0521602, and the Extreme Science and Engineering Discovery Environment (XSEDE) supported by NSF Grant ACI-1053575.

References

[1] Sim, A. Y. L.; Minary, P.; Levitt, M. Modeling Nucleic Acids *Curr. Opin. Struct. Biol.* 2012, 22, 273–278

[2] Kamerlin, S. C. L.; Vicatos, S.; Dryga, A.; Warshel, A. Coarse-Grained (Multiscale) Simulations in Studies of Biophysical and Chemical Systems *Annu. Rev. Phys. Chem*. 2011, 62, 41–64

[3] Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules *Nat. Struct. Biol.* 2002, 9, 646–652

[4] Lopes, P. E.; Huang, J.; Shim, J.; Luo, Y.; Li, H.; Roux, B.; MacKerell, A. D.Polarizable Force Field for Peptides and Proteins Based on the Classical Drude Oscillator *J. Chem. Theory Comput*. 2013, 9, 5430–5449

[5] Mackerell, A. D.; Feig, M.; Brooks, C. L. Extending the Treatment of Backbone Energetics in Protein Force Fields: Limitations of Gas-Phase Quantum Mechanics in Reproducing Protein Conformational Distributions in Molecular Dynamics Simulations *J. Comput. Chem*. 2004, 25, 1400–1415

[6] MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S. A. All-atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins *J. Phys. Chem. B* 1998, 102, 3586–3616

[7] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules *J. Am. Chem. Soc.* 1995, 117, 5179–5197

[8] Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Field for Simulations of Proteins and Peptides *J. Comput. Chem.* 2006, 27, 781–790

[9] Cerutti, D. S.; Swope, W. C.; Rice, J. E.; Case, D. A. Ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins *J. Chem. Theory Comput.* 2014, 10, 4515–4534

[10] Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. Validation of the 53A6 GROMOS Force Field *Eur. Biophys. J.* 2005, 34, 273–284

[11] Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins Via Comparison with Accurate Quantum Chemical Calculations on Peptides *J. Phys. Chem. B* 2001, 105, 6474–6487

[12] Jorgensen, W. L.; Tirado-Rives, J. The OPLS [optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin *J. Am. Chem. Soc.* 1988, 110, 1657–1666

[13] Khoury, G. A.; Thompson, J. P.; Smadbeck, J.; Kieslich, C. A.; Floudas, C. A. Forcefield_PTM: Ab Initio Charge and AMBER Forcefield Parameters for Frequently Occurring Post-translational Modifications *J. Chem. Theory Comput.* 2013, 9, 5653–5674

[14] Wang, J.; Cieplak, P.; Cai, Q.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Luo, R. Development of Polarizable Models for Molecular Mechanical Calculations. 3. Polarizable Water Models Conforming to Thole Polarization Screening Schemes *J. Phys. Chem. B* 2012, 116, 7999–8008

[15] Wang, J.; Cieplak, P.; Li, J.; Wang, J.; Cai, Q.; Hsieh, M.; Lei, H.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations II: Induced Dipole Models Significantly Improve Accuracy of Intermolecular Interaction Energies *J. Phys. Chem. B* 2011, 115, 3100–3111

[16] Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations I:

Parameterization of Atomic Polarizability *J. Phys. Chem. B* 2011, 115, 3091–3099

[17] Wang, J.; Cieplak, P.; Li, J.; Cai, Q.; Hsieh, M.-J.; Luo, R.; Duan, Y. Development of Polarizable Models for Molecular Mechanical Calculations. 4. Van Der Waals Parametrization *J. Phys. Chem. B* 2012, 116, 7088–7101

[18] Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. Lipid14: The Amber Lipid Force Field *J. Chem. Theory Comput.* 2014, 10, 865–879

[19] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-phase Quantum Mechanical Calculations *J. Comput. Chem.* 2003, 24, 1999–2012

[20] Wang, L.-P.; Martínez, T. J.; Pande, V. S. Building Force Fields—An Automatic, Systematic and Reproducible Approach *J. Phys. Chem. Lett.* 2014, 5, 1885–1891

[21] Wang, L.-P.; Chen, J.; Van Voorhis, T. Systematic Parametrization of Polarizable Force Fields From Quantum Chemistry Data *J. Chem. Theory Comput.* 2012, 9, 452–460

[22] Li, W.; Grimme, S.; Krieg, H.; Möllmann, J.; Zhang, J. Accurate Computation of Gas Uptake in Microporous Organic Molecular Crystals *J. Phys. Chem. C* 2012, 116, 8865–8871

[23] Fischer, M.; Kuchta, B.; Firlej, L.; Hoffmann, F.; Fröba, M. Accurate Prediction of Hydrogen Adsorption in Metal- Organic Frameworks with Unsaturated Metal Sites via a Combined Density-Functional Theory and Molecular Mechanics Approach *J. Phys. Chem. C* 2010, 114, 19116–19126

[24] McDaniel, J. G.; Yu, K.; Schmidt, J. R. Ab Initio, Physically Motivated Force Fields for CO_2 Adsorption in Zeolitic Imidazolate Frameworks *J. Phys. Chem. C* 2012, 116, 1892–1903

[25] Chen, L.; Morrison, C. A.; Du ren, T. Improving Predictions of Gas Adsorption in Metal–Organic Frameworks with Coordinatively Unsaturated Metal Sites: Model Potentials, ab Initio Parameterization, and GCMC Simulations *J. Phys. Chem. C* 2012, 116, 18899–18909

[26] Mackerell, A. D. Empirical Force Fields for Biological Macromolecules: Overview and Issues *J. Comput. Chem.* 2004, 25, 1584–1604

[27] Cerutti, D. S.; Rice, J. E.; Swope, W. C.; Case, D. A. Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization *J. Phys. Chem. B* 2013, 117, 2328–2338

[28] Götz, A. W.; Bucher, D.; Lindert, S.; McCammon, J. A. Dipeptide Aggregation in Aqueous Solution from Fixed Point-Charge Force Fields *J. Chem. Theory Comput.* 2014, 10, 1631–1637

[29] Holland, J. H. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*; Massachusetts Institute of Technology: Cambridge, MA, 1992.

[30] Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley Publishing Co., Inc.: Boston, MA, 1989.

[31] Michalewicz, Z. *Genetic Algorithms+ Data Structures= Evolution Programs*; Springer-Verlag: Berlin Heidelberg, 1996.

[32] Konak, A.; Coit, D. W.; Smith, A. E. Multi-objective Optimization Using Genetic Algorithms: A Tutorial *Reliab. Eng. Syst. Saf.* 2006, 91, 992–1007

[33] Crepinsek, M.; Liu, S.-H.; Mernik, M. Exploration and Exploitation in Evolutionary Algorithms: A Survey *ACM Comput. Surv*. 2013, 45, 35

[34] Wang, J.; Kollman, P. A. Automatic Parameterization of Force Field by Systematic Search and Genetic Algorithms *J. Comput. Chem*. 2001, 22, 1219–1228

[35] Betz, R. M.; Walker, R. C. Paramfit: Automated Optimization of Force Field Parameters for Molecular Dynamics Simulations *J. Comput. Chem.* 2015, 36, 79–87

[36] Leonarski, F.; Trovato, F.; Tozzini, V.; Leś, A.; Trylska, J. Evolutionary Algorithm in the Optimization of a Coarse-Grained Force Field *J. Chem. Theory Comput*. 2013, 9, 4874–4889

[37] Pahari, P.; Chaturvedi, S. Determination of Best-Fit Potential Parameters for a Reactive Force Field Using a Genetic Algorithm *J. Mol. Model.* 2012, 18, 1049–1061

[38] Larsson, H. R.; van Duin, A. C. T.; Hartke, B. Global Optimization of Parameters in the Reactive Force Field ReaxFF for SiOH *J. Comput. Chem.* 2013, 34, 2178–2189

[39] Strassner, T.; Busold, M.; Herrmann, W. A. MM3 Parametrization of Four- and Five-coordinated Rhenium Complexes by a Genetic Algorithm—Which Factors Influence the Optimization Performance? *J. Comput. Chem.* 2002, 23, 282–290

[40] Tafipolsky, M.; Schmid, R. Systematic First Principles Parameterization of Force Fields for Metal-Organic Frameworks Using a Genetic Algorithm Approach *J. Phys. Chem. B* 2009, 113, 1341–1352

[41] Courcot, B.; Bridgeman, A. J. Optimization of a Molecular Mechanics Force Field for Polyoxometalates Based on a Genetic Algorithm *J. Comput. Chem.* 2011, 32, 240–247

[42] Cundari, T. R.; Fu, W. Genetic Algorithm Optimization of a Molecular Mechanics Force Field for Technetium *Inorg. Chim. Acta* 2000, 300, 113–124

[43] Courcot, B.; Bridgeman, A. J. Optimization of a Molecular Mechanics Force Field for Type-II Polyoxometalates Focussing on Electrostatic Interactions: A Case Study *J. Comput. Chem*. 2011, 32, 1703–1710

[44] Cox, S. R.; Williams, D. E. Representation of the Molecular Electrostatic Potential by a Net Atomic Charge Model *J. Comput. Chem*. 1981, 2, 304– 23

[45] Singh, C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules *J. Comput. Chem*. 1984, 5, 129–145

[46] Besler, B. H.; Merz, K. M.; Kollman, P. A.Atomic Charges Derived From Semiempirical Methods *J. Comput. Chem*. 1990, 11, 431–439

[47] Chirlian, L. E.; Francl, M. M. Atomic Charges Derived From Electrostatic Potentials: A Detailed Study *J. Comput. Chem*. 1987, 8, 894–905

[48] Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles From Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis *J. Comput. Chem.* 1990, 11, 361–373

[49] Momany, F. A. Determination of Partial Atomic Charges From ab Initio Molecular Electrostatic Potentials. Application to Formamide, Methanol, and Formic Acid *J. Phys. Chem*. 1978, 82, 592–601

[50] Cardamone, S.; Hughes, T. J.; Popelier, P. L. A. Multipolar Electrostatics *Phys. Chem. Chem. Phys*. 2014, 16, 10367–10387

[51] Kramer, C.; Spinn, A.; Liedl, K. R. Charge Anisotropy: Where Atomic Multipoles Matter Most *J. Chem. Theory Comput*. 2014, 10, 4488–4496

[52] Woods, R. J.; Khalil, M.; Pell, W.; Moffat, S. H.; Smith, V. H. Derivation of Net Atomic Charges From Molecular Electrostatic Potentials *J. Comput. Chem*. 1990, 11, 297–310

[53] Sigfridsson, E.; Ryde, U. Comparison of Methods for Deriving Atomic Charges From the Electrostatic Potential and Moments *J. Comput. Chem.* 1998, 19, 377–395

[54] Tsiper, E. V.; Burke, K. Rules for Minimal Atomic Multipole Expansion of Molecular Fields *J. Chem. Phys*. 2004, 120, 1153–1156

[55] Jakobsen, S.; Jensen, F. Systematic Improvement of Potential-Derived Atomic Multipoles and Redundancy of the Electrostatic Parameter Space *J. Chem. Theory Comput*. 2014, 10, 5493–5504

[56] The dependence of the fitted charges on the molecule orientation can be eliminated by complete grid sampling, e.g., by integrating over isodensity surface, see refs 54 and 55.

[57] Stouch, T. R.; Williams, D. E. Conformational Dependence of Electrostatic Potential-Derived Charges: Studies of the Fitting Procedure *J. Comput. Chem*. 1993, 14, 858–866

[58] Stouch, T. R.; Williams, D. E. Conformational Dependence of Electrostatic Potential Derived Charges of a Lipid Headgroup: Glycerylphosphorylcholine *J. Comput. Chem*. 1992, 13, 622–632

[59] Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP Model *J. Phys. Chem*. 1993, 97, 10269–10280

[60] Francl, M. M.; Carey, C.; Chirlian, L. E.; Gange, D. M. Charges Fit to Electrostatic Potentials. 11. Can Atomic Charges Be Unambiguously Fit to Electrostatic Potentials? *J. Comput. Chem*. 1996, 17, 367–383

[61] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollmann, P. A. Application of RESP Charges to Calculate Conformational Energies, Hydrogen Bond Energies, and Free Energies of Solvation *J. Am. Chem. Soc*. 1993, 115, 9620–9631

[62] Hinsen, K.; Roux, B. A Potential Function for Computer Simulation Studies of Proton Transfer in Acetylacetone *J. Comput. Chem*. 1997, 18, 368–380

[63] Simmonett, A. C.; Gilbert, A. T.; Gill, P. M. An Optimal Point-Charge Model for Molecular Electrostatic Potentials *Mol. Phys*. 2005, 103, 2789–2793

[64] Burger, S. K.; Schofield, J.; Ayers, P. W. Quantum Mechanics/Molecular Mechanics Restrained Electrostatic Potential Fitting. *J. Phys. Chem. B* 2013.

[65] Arnautova, Y. A.; Jagielska, A.; Scheraga, H. A. A New Force Field (ECEPP-05) for Peptides, Proteins, and Organic Molecules *J. Phys. Chem. B* 2006, 110, 5025–5044

[66] Zeng, J.; Duan, L.; Zhang, J. Z. H.; Mei, Y. A Numerically Stable Restrained Electrostatic Potential Charge Fitting Method *J. Comput. Chem.* 2013, 34, 847–853

[67] Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on ab Initio Target Data *J. Chem. Theory Comput*. 2013, 9, 3543–3556

[68] Rai, B. K.; Bakken, G. A. Fast and Accurate Generation of ab Initio Quality Atomic Charges Using Nonparametric Statistical Regression *J. Comput. Chem.* 2013, 34, 1661–1671

[69] Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: A Generalizable Biomolecular Force Field. Carbohydrates *J. Comput. Chem.* 2008, 29, 622–655

[70] Seo, M.; Castillo, N.; Ganzynkowicz, R.; Daniels, C. R.; Woods, R. J.; Lowary, T. L.; Roy, P.-N. Approach for the Simulation and Modeling of Flexible Rings: Application to the A-d-Arabinofuranoside Ring, a Key Constituent of Polysaccharides from *Mycobacterium tuberculosis J. Chem. Theory Comput.* 2008, 4, 184–191

[71] Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the Multimolecule and Multiconformational RESP Methodology to Biopolymers Charge Derivation for DNA, RNA, and Proteins *J. Comput. Chem.* 1995, 16, 1357–1377

[72] Dupradeau, F.-Y.; Pigache, A.; Zaffran, T.; Savineau, C.; Lelong, R.; Grivel, N.; Lelong, D.; Rosanski, W.; Cieplak, P. The R.E.D. Tools: Advances in RESP and ESP Charge Derivation and Force Field Library Building *Phys. Chem. Chem. Phys*. 2010, 12, 7821–7839

[73] Bereau, T.; Kramer, C.; Monnard, F. W.; Nogueira, E. S.; Ward, T. R.; Meuwly, M. Scoring Multipole Electrostatics in Condensed-Phase Atomistic Simulations *J. Phys. Chem. B* 2013, 117, 5460–5471

[74] Hofmann, F. D.; Devereux, M.; Pfaltz, A.; Meuwly, M. Toward Force Fields for Atomistic Simulations of Iridium-Containing Complexes *J. Comput. Chem.* 2013, 35, 18–29

[75] Laio, A.; VandeVondele, J.; Rothlisberger, U. D-RESP: Dynamically Generated Electrostatic Potential Derived Charges From Quantum Mechanics/Molecular Mechanics Simulations *J. Phys. Chem. B* 2002, 106, 7300–7307

[76] Laio, A.; Gervasio, F. L.; VandeVondele, J.; Sulpizi, M.; Rothlisberger, U. A Variational Definition of Electrostatic Potential Derived Charges *J. Phys. Chem. B* 2004, 108, 7963–7968

[77] Graen, T. M. D.; Hoefling, M.; Grubmüller, H. AMBER-DYES: Characterization of Charge Fluctuations and Force Field Parameterization of Fluorescent Dyes for Molecular Dynamics Simulations *J. Chem. Theory Comput*. 2014, 10, 5505–5512

[78] Vöhringer-Martinez, E.; Verstraelen, T.; Ayers, P. W. The Influence of Ser-154, Cys-113, and the Phosphorylated Threonine Residue on the Catalytic Reaction Mechanism of Pin1 *J. Phys. Chem. B* 2014, 118, 9871–9880

[79] Becke, A. D. Density-Functional Thermochemistry. III. The Role of Exact Exchange *J. Chem. Phys.* 1993, 98, 5648–5652

[80] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields *J. Phys. Chem*. 1994, 98, 11623–11627

[81] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Gaussian 09, revision C.01; Gaussian, Inc.: Wallingford, CT, 2010.

[82] Lawson, C.; Hanson, R. *Solving Least Squares Problems*; Prentice-Hall, Inc.: Englewood Cliffs, NJ, 1974.

[83] Eshelman, L. J.; David, S. Real-Coded Genetic Algorithms and Interval-Schemata. In *Foundations of Genetic Algorithms*; Whitely, D., Ed.;

Morgan Kaufmann Publishers, Inc.: San Manteo, CA, 1993; pp 187–202.

[84] Herrera, F.; Lozano, M.; Verdegay, J. L. Tackling Real-Coded Genetic Algorithms: Operators and Tools for Behavioural Analysis *Artif. Intell. Rev.* 1998, 12, 265–319

[85] Goldberg, D. E. Real-Coded Genetic Algorithms, Virtual Alphabets, and Blocking *Complex Syst*. 1991, 5, 139–167

[86] Van Der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation *Comput. Sci. Eng.* 2011, 13, 22–30

[87] Schaffer, D. Proceedings of the 1st International Conference on Genetic Algorithms; Institution of Engineering and Technology: London, 1985; pp 93–100

[88] Igel, C.; Hansen, N.; Roth, S. Covariance Matrix Adaptation for Multi-objective Optimization *Evol. Comput*. 2007, 15, 1–28

[89] Hansen, N.; Müller, S. D.; Koumoutsakos, P. Reducing the Time Complexity of the Derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES) *Evol. Comput*. 2003, 11, 1–18

[90] Hansen, N.; Ostermeier, A. Completely Derandomized Self-Adaptation in Evolution Strategies *Evol. Comput*. 2001, 9, 159–195

[91] Hansen, N. *The CMA Evolution Strategy: A Comparing Review. In Towards a New Evolutionary Computation*; Lozano, J.; Larrañaga, P.; Inza, I.; Bengoetxea, E., Eds.; Springer: Berlin Heidelberg, 2006; pp 75–102.

[92] Hunter, J. D. Matplotlib: A 2D Graphics Environment *Comput. Sci. Eng.* 2007, 9, 0090–0095

[93] The Hessian matrices of the fitness function used, i.e., the RMSE (eq. 11), and of the LS sum (eq. 1) have identical eigenvectors, but numerically different eigenvalues. These differences, however, do not affect the discussion.

[94] Salomon, R. Re-evaluating Genetic Algorithm Performance under Coordinate Rotation of Benchmark Functions. A Survey of Some Theoretical and Practical Aspects of Genetic Algorithms *BioSystems* 1996, 39, 263–278

[95] The correspondence between the largest-curvature LS-sum Hessian eigenmodes and the total charge and the dipole moment components has also been noted by Rothlisberger and co-workers, ref 75.

[96] Hansen, P. C. *Least Squares Data Fitting with Applications*; Johns Hopkins University Press: Baltimore, MD, 2013.

[97] The fluctuations of the ESP charges calculated using QM/MM methods along MD trajectories, which have been ascribed to the polarization effects (ref 77) may in fact originate, to a significant degree, from the buried atom effect.

[98] Debiec, K. T.; Gronenborn, A. M.; Chong, L. T. Evaluating the Strength of Salt Bridges: A Comparison of Current Biomolecular Force Fields *J. Phys. Chem. B* 2014, 118, 6561–6569

Supporting Information

GA convergence, the population size effect; GA convergence, binary vs real-number chromosome coding; analysis of the GA point-charge fitting results; GA point charge fitting in terms of the rotated/eigenvector coordinates; 1-chlorobutane point charge fitting. This material is available free of charge via the Internet at http://pubs.acs.org.

Genetic Algorithm Optimization of Point Charges in Force Field Development: Challenges and Insights

Supporting Information

Maxim V. Ivanov, Marat R. Talipov, and Qadir K. Timerghazin*

Department of Chemistry, Marquette University, P.O. Box 1881, Milwaukee, Wisconsin 53201-1881, United States

Table of contents

1. GA Convergence: The Population Size Effect

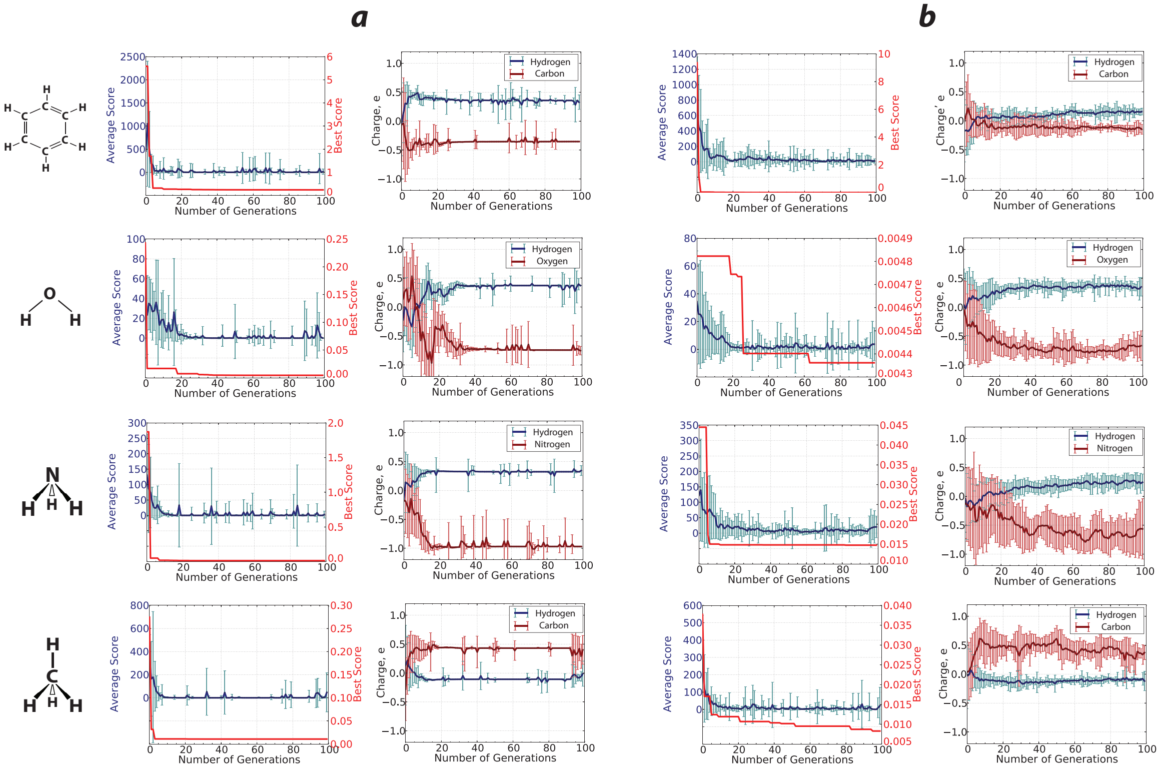


Figure S1. GA convergence with 20 chromosomes in the population (a) as compared to 50 chromosomes in the population (b).

2. GA Convergence: Binary vs. Real-Number Chromosome Coding

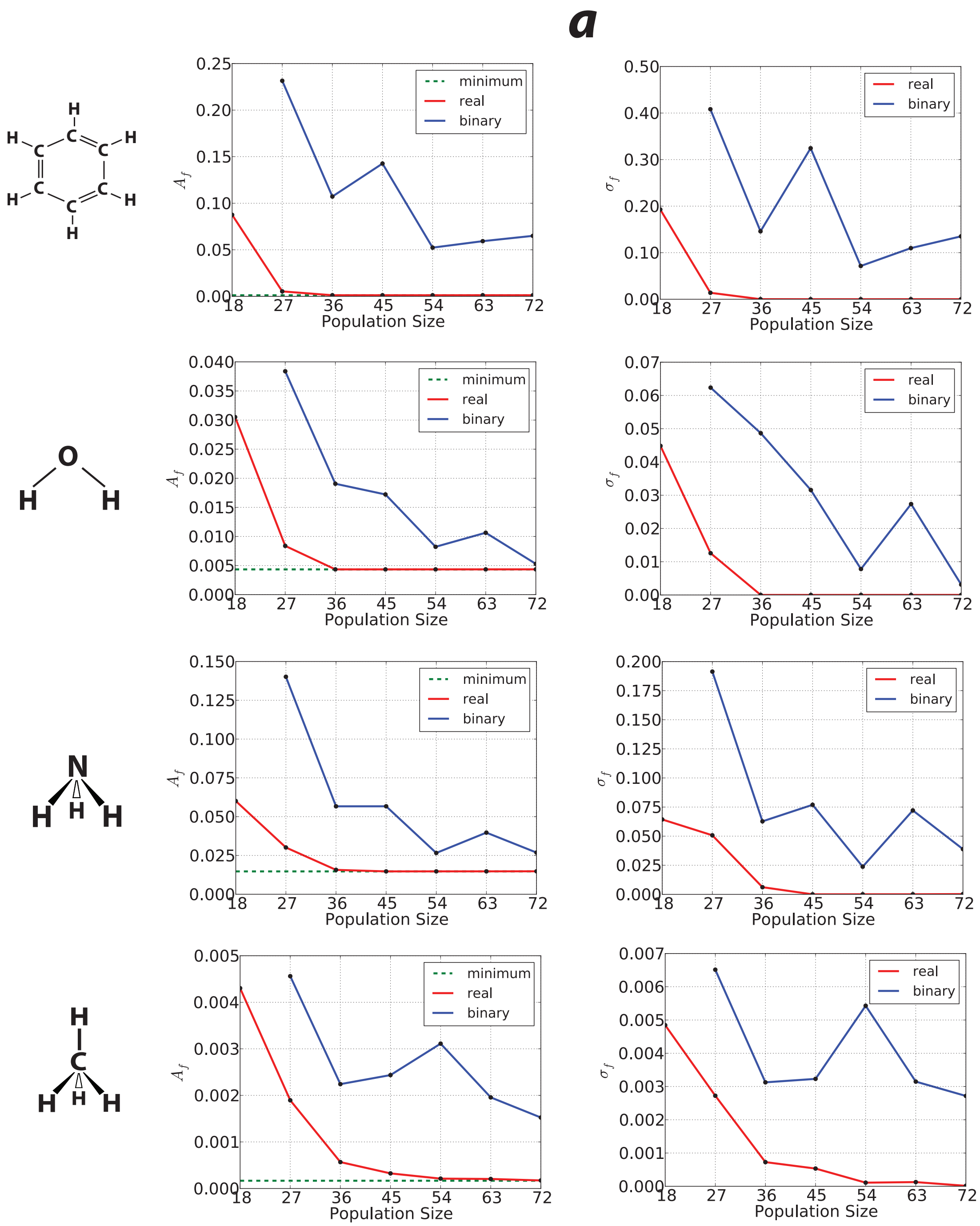


Figure S2a. Average fitness scores A_f and their standard deviations σ_f for 200 GA runs as functions of the population size f the model molecules with two symmetry independent charges. Real-number representation is compared with binary representation of chromosomes. Green dashed line corresponds to the solution found by ESP method.

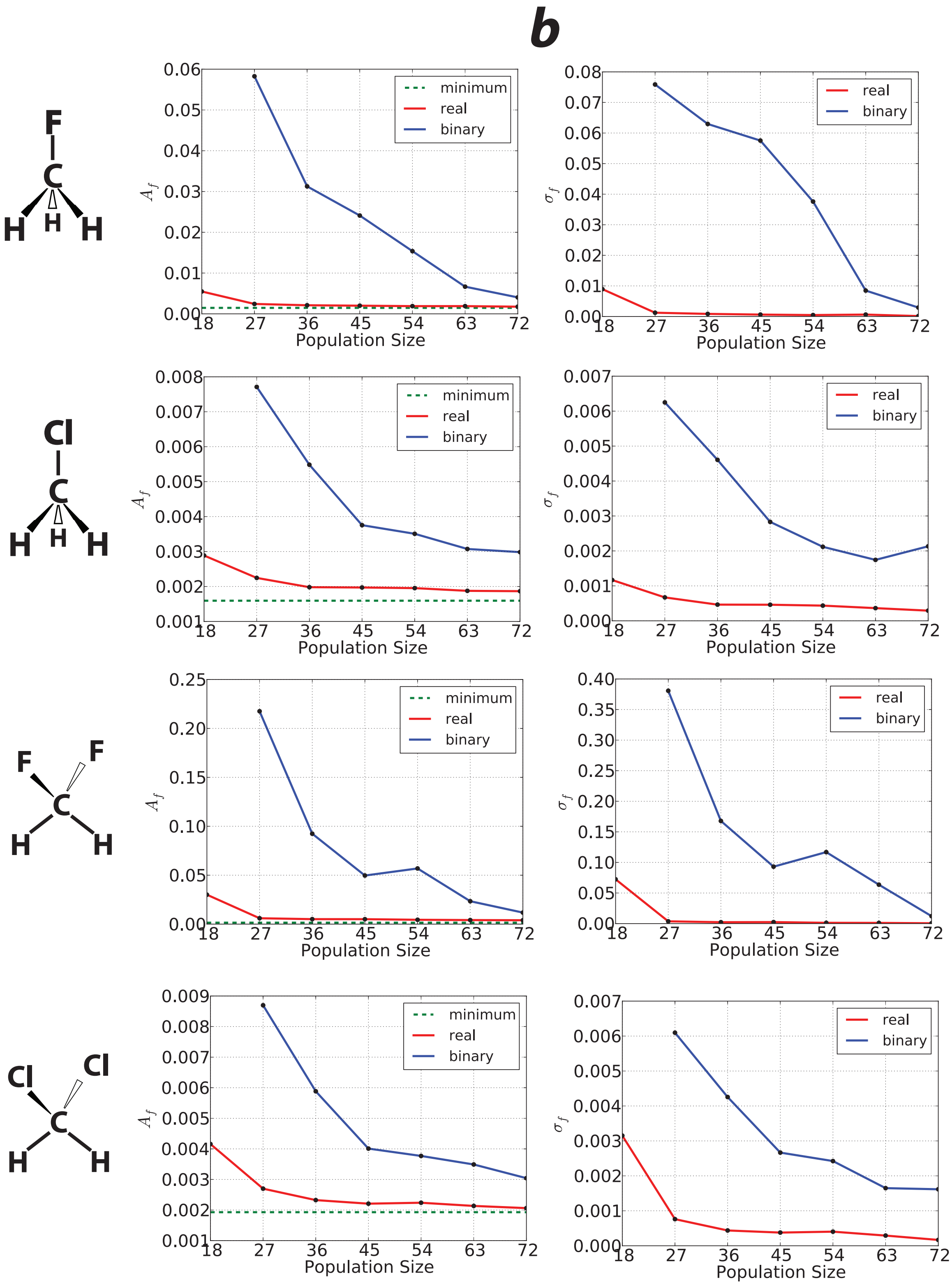


Figure S2b. Average fitness scores A_f and their standard deviations σ_f for 200 GA runs as functions of the population size for the model molecules with three symmetry independent charges. Real-number representation is compared with binary representation of chromosomes. Green dashed line corresponds to the solution found by ESP method.

3. Analysis of the GA Point-Charge Fitting Results

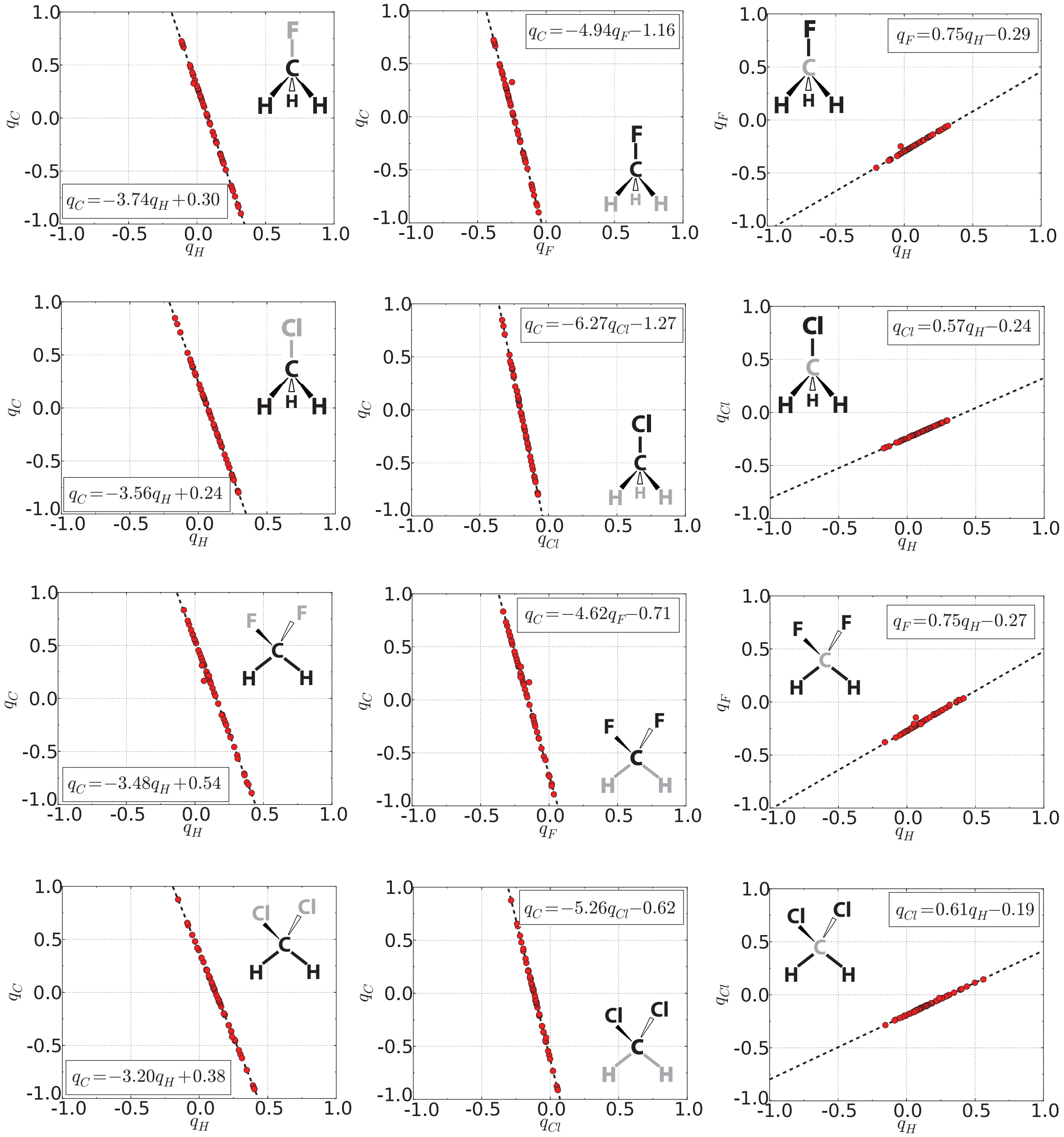


Figure S3. Correlations between the GA-optimized charges in CH_3X, CH_2X_2 (X = F, Cl) molecules obtained from 200 independent GA runs. All trend lines have correlation coefficient R^2 = 1.00. All optimizations were performed with 30 chromosomes in the population.

Table S1. Best and average $<q>$ values with corresponding standard deviations σ of point charges and their fitness scores f obtained from 200 GA runs using charges and Hessian eigenvectors as optimization coordinates. Results are compared with the solutions found by ESP method. All values are in atomic units.

			200 GA Runs					
			Point-Charge Coordinates			Eigenvector Coordinates		
Molecule		ESP	Best	$<q>$	σ	Best	$<q>$	σ
H_2O	q_O	-0.758	-0.758	-0.529	0.256	-0.758	-0.758	0.000
	q_H	0.379	0.379	0.264	0.129	0.379	0.379	0.001
	$f \times 10^3$	2.935	2.935	7.582	5.523	2.935	0.000	0.000
NH_3	q_N	-0.970	-0.970	-0.588	0.413	-0.970	-0.970	0.001
	q_H	0.324	0.324	0.196	0.139	0.324	0.324	0.001
	$f \times 10^3$	3.686	3.686	8.494	5.126	3.686	3.694	0.039
C_6H_6	q_C	-0.125	-0.124	-0.099	0.215	-0.125	-0.121	0.009
	q_H	0.125	0.125	0.098	0.219	0.125	0.121	0.009
	$f \times 10^4$	9.577	9.580	92.772	91.687	9.577	12.163	5.208
CH_4	q_C	-0.616	-0.615	-0.024	0.535	-0.616	-0.602	0.026
	q_H	0.154	0.154	0.006	0.134	0.154	0.151	0.007
	$f \times 10^4$	3.387	3.387	16.757	12.480	3.387	3.805	0.848
CH_3F	q_C	-0.038	0.026	0.129	0.476	-0.038	-0.039	0.007
	q_H	0.090	0.072	0.045	0.127	0.090	0.090	0.002
	q_F	-0.229	-0.242	-0.263	0.097	-0.229	-0.228	0.003
	$f \times 10^3$	1.462	1.473	1.965	0.538	1.462	1.468	0.042
CH_3Cl	q_C	-0.560	-0.554	-0.030	0.483	-0.560	-0.560	0.000
	q_H	0.225	0.223	0.077	0.136	0.225	0.225	0.001
	q_{Cl}	-0.114	-0.115	-0.200	0.076	-0.114	-0.114	0.000
	$f \times 10^3$	1.593	1.593	2.309	0.716	1.593	1.593	0.000
CH_2F_2	q_C	0.251	0.261	0.023	0.127	0.251	0.129	0.039
	q_H	0.084	0.080	0.148	0.441	0.084	0.100	0.220
	q_F	-0.209	-0.210	-0.158	0.097	-0.209	-0.163	0.078
	$f \times 10^3$	1.482	1.484	2.254	1.141	1.482	2.830	2.472
CH_2Cl_2	q_C	-0.599	-0.553	-0.042	0.500	-0.599	-0.599	0.001
	q_H	0.305	0.291	0.131	0.155	0.305	0.305	0.000
	q_{Cl}	-0.005	-0.014	-0.111	0.096	-0.005	-0.005	0.000
	$f \times 10^3$	1.930	1.934	2.699	0.739	1.930	1.930	0.000

Table S2. Pairwise linear correlations between the point charges obtained from 200 GA optimizations for the CH_3X, CH_2X_2 (X = F, Cl) and CH_3O^- molecules compared with the analytically derived relationships (eqs. 18-19 in the main text). All values are in atomic units.

Molecule	From GA	Analytical
CH_3F	$q_C = -3.74q_H + 0.30$	$q_C = -3.75q_H + 0.29$
	$q_C = -4.94q_F - 1.16$	$q_C = -4.99q_F - 1.17$
	$q_F = 0.75q_H - 0.29$	$q_F = 0.76q_H - 0.29$
CH_3Cl	$q_C = -3.56q_H + 0.24$	$q_C = -3.55q_H + 0.23$
	$q_C = -6.27q_{Cl} - 1.27$	$q_C = -6.41q_{Cl} - 1.26$
	$q_{Cl} = 0.57q_H - 0.24$	$q_{Cl} = 0.56q_H - 0.23$
CH_2F_2	$q_C = -3.48q_H + 0.54$	$q_C = -3.50q_H + 0.53$
	$q_C = -4.62q_F - 0.71$	$q_C = -4.67q_F + 0.71$
	$q_F = 0.75q_H - 0.27$	$q_F = 0.75q_H - 0.27$
CH_2Cl_2	$q_C = -3.20q_H + 0.38$	$q_C = -3.22q_H + 0.36$
	$q_C = -5.26q_{Cl} - 0.62$	$q_C = -5.28q_{Cl} - 0.59$
	$q_{Cl} = 0.61q_H - 0.19$	$q_{Cl} = 0.61q_H - 0.18$
CH_3O^-	$q_C = -4.07q_H - 0.24$	$q_C = -4.11q_H - 0.29$
	$q_C = -3.69q_O - 2.92$	$q_C = -3.71q_O - 2.92$
	$q_H = 0.91q_O + 0.66$	$q_H = 0.90q_O + 0.64$

Table S3. Numerical equivalence between the eigenbasis of the covariance matrix calculated for 200 independent 20-chromosome GA runs , $\widetilde{\boldsymbol{\Sigma}}$, the eigenbasis of the LS-sum Hessian matrix, and the analytically generated orthonormal basis $\widetilde{\mathbf{U}}$ (eq. 28 in the main text). Eigenvalues of the covariance matrix correspond to the variance (in atomic units, e^2) along each eigenvectors; eigenvalues of the Hessian correspond to the curvatures (in atomic units, $1/a_0^2$) along corresponding eigenvectors.

	Methane		Ammonia		Water		Benzene	
				$\widetilde{\boldsymbol{\Sigma}}$				
Variance	5.75E-09	0.19	7.18E-07	0.13	1.15E-06	0.07	1.85E-07	0.06
q_X	0.244	0.970	0.319	0.948	0.450	0.893	-0.717	0.697
q_H	0.970	-0.244	0.948	-0.319	0.893	-0.450	-0.697	-0.717
				$\widetilde{\mathbf{H}}$				
Curvature	2094.94	0.07	1116.93	1.05	570.33	2.13	9470.55	22.29
q_X	0.244	0.970	0.318	0.948	0.447	0.894	-0.718	0.696
q_H	0.970	-0.244	0.948	-0.318	0.894	-0.447	-0.696	-0.718
				$\widetilde{\mathbf{U}}$				
q_X	0.243	0.970	0.316	0.949	0.447	0.894	0.707	0.707
q_H	0.970	-0.243	0.949	-0.316	0.894	-0.447	0.707	-0.707

4. GA Point Charge Fitting in Terms of the Rotated/Eigenvector Coordinates

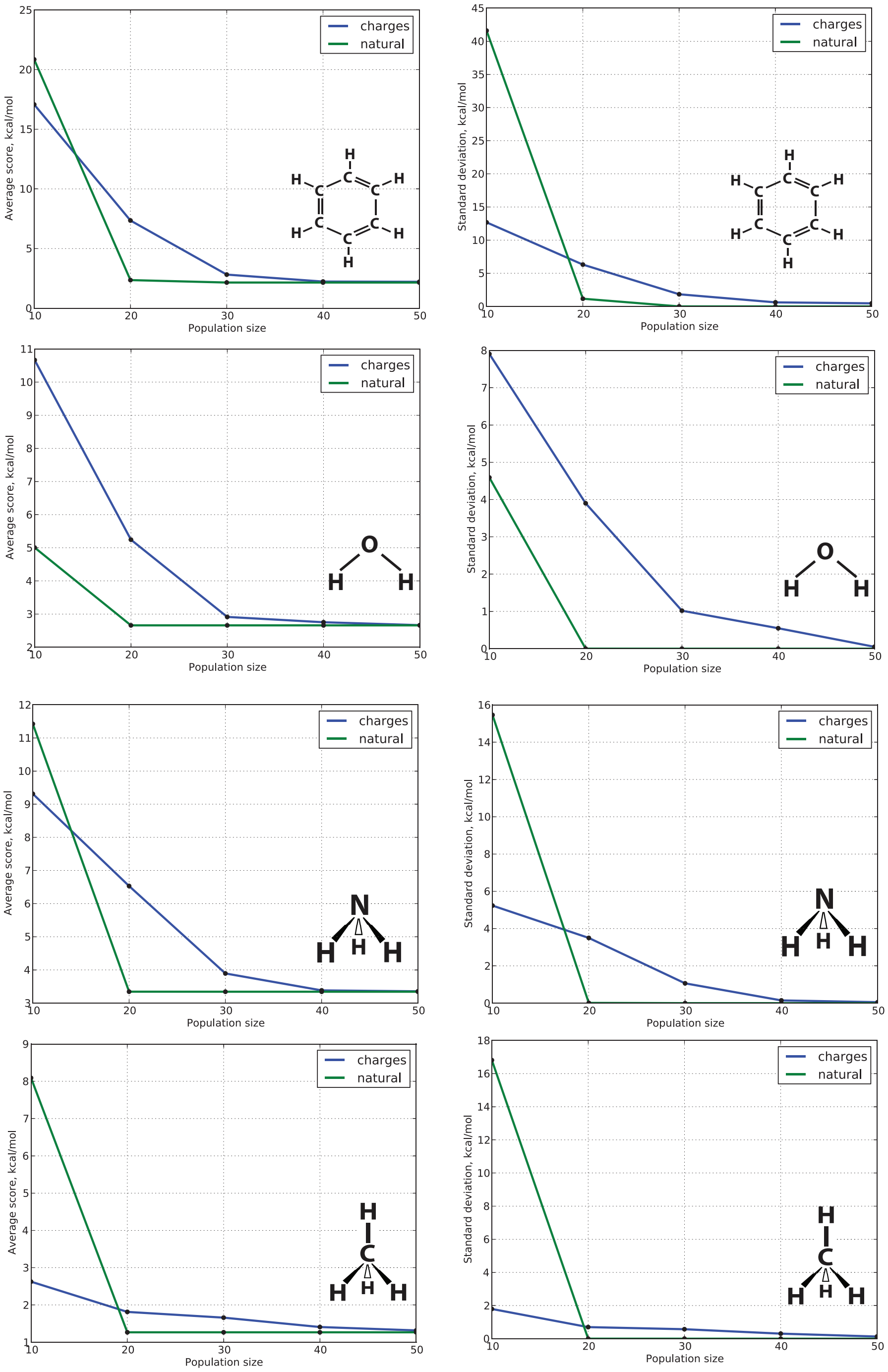


Figure S4a. Average fitness score and its standard deviation for 200 GA runs performed using the point charge values as the optimization coordinates vs. the coordinates defined by the eigenbasis of the LS-sum Hessian matrix; two-charge models.

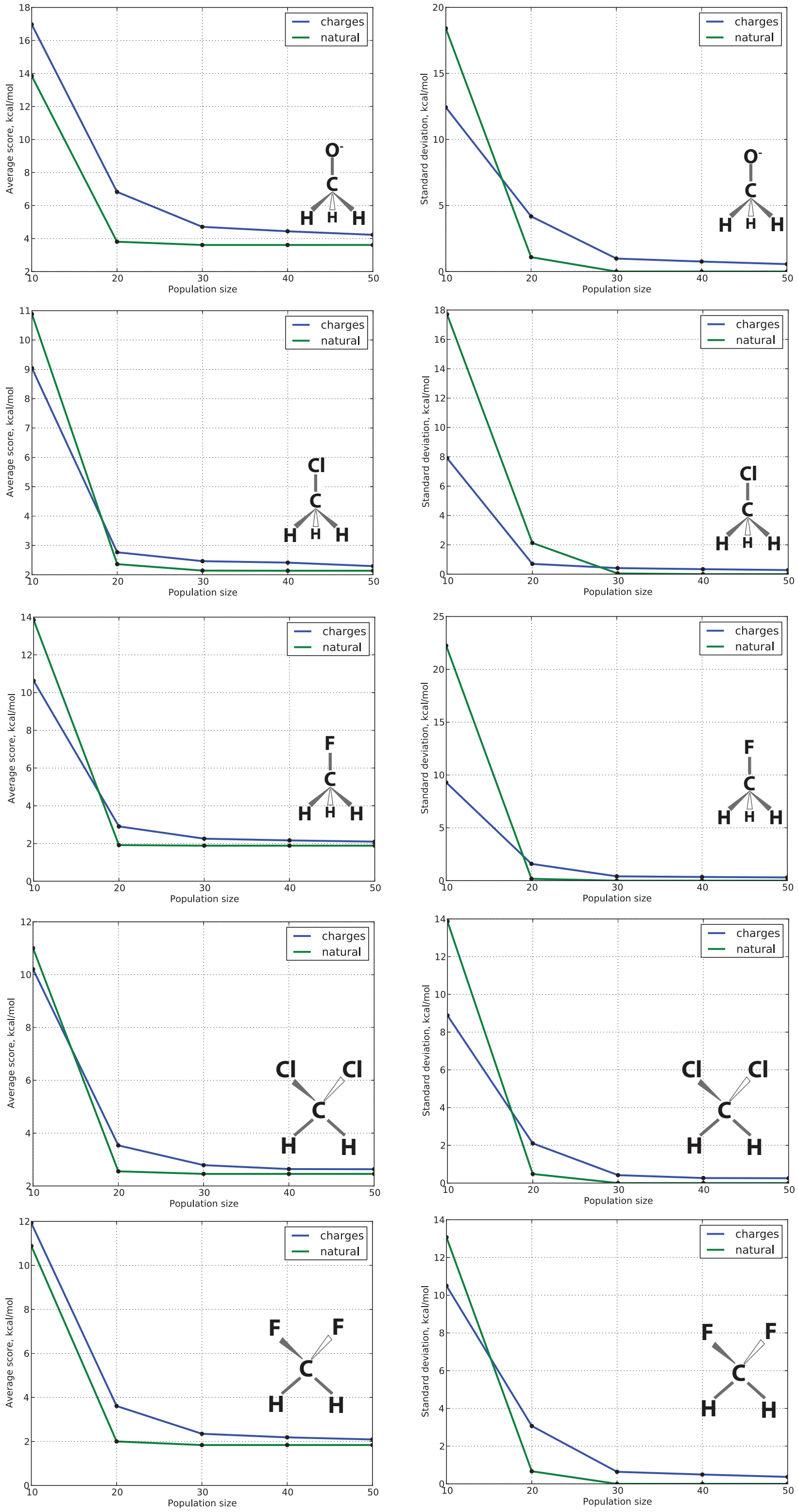


Figure S4b. Average fitness score and its standard deviation for 200 GA runs performed using the point charge values as the optimization coordinates vs. the coordinates defined by the eigenbasis of the LS-sum Hessian matrix; three-charge models.

Table S4. Covariance matrix and Hessian matrix eigenbases $\widetilde{\mathbf{\Sigma}}$ and $\widetilde{\mathbf{H}}$ compared with the orthonormal basis $\widetilde{\mathbf{U}}$ (eqs. 28-29 in the main text) . Eigenvalues of the covariance matrix correspond to the variance (in atomic units, e^2) along each eigenvectors; eigenvalues of the Hessian correspond to the curvatures (in atomic units, $1/a_0^2$) along the corresponding eigenvectors. Covariance matrices are calculated for 200 GA runs with 30 chromosomes in the population.

	Chloromethane			Fluoromethane			Methoxide			Dichloromethane			Difluoromethane		
								$\widetilde{\mathbf{H}}$							
Curvature	1555.96	18.23	0.10	1374.20	14.45	0.09	1375.88	15.79	0.12	1382.94	27.35	0.14	1125.74	22.21	0
q_C	0.303	0.059	0.951	0.303	0.101	0.948	0.302	0.165	0.939	0.337	-0.074	0.939	0.336	-0.043	0.
q_X	0.292	0.944	-0.151	0.293	0.936	-0.194	0.293	0.921	-0.256	0.657	0.732	-0.178	0.662	0.722	-0
q_H	0.907	-0.324	-0.269	0.907	-0.336	-0.254	0.907	-0.352	-0.230	0.674	-0.677	-0.296	0.670	-0.691	-0
								$\widetilde{\mathbf{\Sigma}}$							
Variance	8.40E-08	8.73E-06	0.12	6.63E-08	5.95E-06	0.11	3.69E-07	4.55E-05	0.14	1.62E-07	8.59E-06	0.11	9.79E-08	9.30E-06	0
q_C	0.302	0.064	0.951	0.302	0.102	0.948	0.303	0.163	0.939	0.335	-0.084	0.938	0.336	-0.043	0.
q_X	0.274	0.950	-0.151	0.288	0.938	-0.193	0.291	0.922	-0.254	0.676	0.715	-0.177	0.664	0.719	-0
q_H	0.913	-0.306	-0.269	0.909	-0.331	-0.254	0.908	-0.350	-0.232	0.656	-0.694	-0.297	0.668	-0.694	-0
								$\widetilde{\mathbf{U}}$							
q_C	0.3015	0.058	0.9519	0.3015	0.0968	0.9488	0.3015	0.1511	0.9399	0.3333	-0.0604	0.9398	0.3333	-0.041	0.
q_X	0.3015	0.9785	-0.1497	0.3015	0.9397	-0.1884	0.3015	0.9206	-0.2534	0.6667	0.7603	-0.1781	0.6667	0.7098	-0.
q_H	0.9045	-0.3213	-0.2674	0.9045	-0.369	-0.2535	0.9045	-0.3514	-0.2288	0.6667	-0.6362	-0.2918	0.6667	-0.688	-0.

5. 1-Chlorobutane Point Charge Fitting

Chart 1. 1-Chlorobutane Conformers Considered, with Atom Numbering

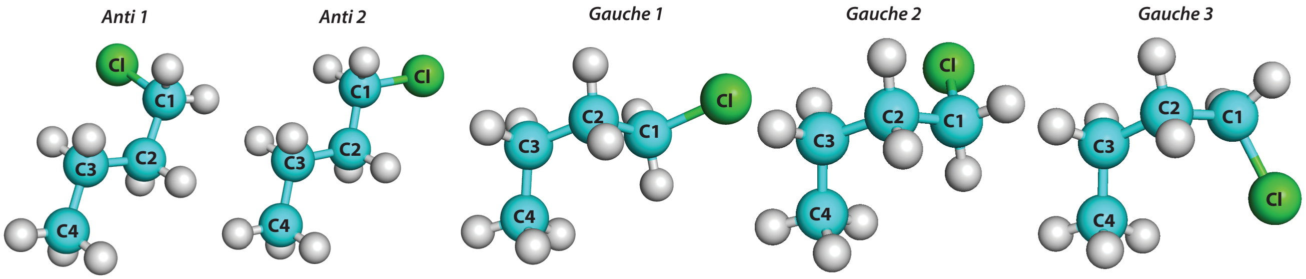


Table S5. Point charges with corresponding dipole moment and total charge obtained with CMA-ES/ESP methods as compared with the average and the standard deviation σ of the point charges, total charge and dipole moment obtained from 200 independent runs for the five conformers of 1-chlorobutane in point charge coordinates and Hessian eigenvectors coordinates.

	q_{C4}	q_{H4}	q_{C3}	q_{H3}	q_{C2}	q_{H2}	q_{C1}	q_{H1}	q_{Cl}	Dipole moment, au	Total charge, au	Fitness, kcal/mol
	CMA-ES/ESP											
Anti 1	-0.257	0.058	0.186	-0.024	0.052	0.014	-0.141	0.105	-0.193	0.918	0.009	2.094
Anti 2	-0.229	0.054	0.163	-0.027	0.051	0.015	-0.104	0.096	-0.206	0.995	0.008	2.126
Gauche 1	-0.134	0.032	0.110	-0.018	0.050	0.011	-0.060	0.088	-0.216	1.006	0.008	2.119
Gauche 2	-0.203	0.047	0.172	-0.027	0.015	0.007	0.005	0.066	-0.213	0.903	0.009	2.136
Gauche 3	-0.119	0.027	0.148	-0.038	0.094	-0.010	-0.030	0.072	-0.211	0.906	0.009	2.143

… continued on the next page

Table S5, Continued.

		q_{C4}	q_{H4}	q_{C3}	q_{H3}	q_{C2}	q_{H2}	q_{C1}	q_{H1}	q_{Cl}	Dipole moment, au	Total charge, au	Score, kcal/mol
		Point-Charge Coordinates											
Anti 1	<A>	-0.030	0.004	0.037	0.009	0.002	0.026	0.005	0.064	-0.215	0.922	0.008	3.054
	σ	0.380	0.100	0.365	0.113	0.369	0.116	0.371	0.113	0.065	0.087	0.008	0.435
Anti 2	<A>	-0.030	0.008	0.025	0.003	0.027	0.021	0.001	0.069	-0.225	1.015	0.007	3.016
	σ	0.371	0.100	0.341	0.107	0.384	0.123	0.352	0.110	0.060	0.108	0.008	0.405
Gauche 1	<A>	-0.028	0.006	0.058	-0.007	0.007	0.025	-0.017	0.078	-0.223	1.010	0.008	3.059
	σ	0.379	0.103	0.354	0.106	0.370	0.114	0.364	0.114	0.063	0.110	0.008	0.446
Gauche 2	<A>	-0.006	0.000	0.036	0.002	0.022	0.011	0.022	0.061	-0.214	0.905	0.009	3.079
	σ	0.374	0.101	0.367	0.111	0.355	0.107	0.351	0.109	0.063	0.081	0.008	0.455
Gauche 3	<A>	0.041	-0.007	-0.016	0.006	0.036	0.015	0.008	0.065	-0.212	0.921	0.010	3.148
	σ	0.378	0.101	0.360	0.109	0.344	0.106	0.375	0.118	0.062	0.096	0.009	0.489
		Eigenvector Coordinates											
Anti 1	<A>	-0.224	0.051	0.138	-0.014	0.081	0.009	-0.147	0.106	-0.192	0.921	0.009	2.576
	σ	0.237	0.059	0.242	0.066	0.243	0.063	0.243	0.069	0.039	0.068	0.006	0.196
Anti 2	<A>	-0.224	0.053	0.153	-0.024	0.062	0.013	-0.116	0.099	-0.204	1.003	0.008	2.574
	σ	0.254	0.063	0.254	0.064	0.238	0.069	0.207	0.062	0.037	0.082	0.006	0.194
Gauche 1	<A>	-0.121	0.030	0.091	-0.013	0.046	0.013	-0.060	0.088	-0.214	1.009	0.009	2.605
	σ	0.257	0.064	0.252	0.065	0.240	0.065	0.210	0.064	0.035	0.080	0.006	0.207
Gauche 2	<A>	-0.225	0.059	0.370	-0.100	-0.276	0.026	0.056	-0.166	-0.290	0.902	0.130	2.394
	σ	0.285	0.071	0.295	0.077	0.266	0.074	0.090	0.065	0.042	0.113	0.046	0.141
Gauche 3	<A>	-0.081	0.017	0.123	-0.033	0.107	-0.013	-0.035	0.074	-0.213	0.922	0.009	2.621
	σ	0.253	0.063	0.238	0.060	0.246	0.063	0.213	0.061	0.036	0.076	0.006	0.178

Table S6. Numerical representation of the eigenbases of the LS-sum Hessian and the covariance matrix for the 200 GA runs for the five conformers of 1-chlorobutane. Eigenvalues of the covariance matrix correspond to the variance (in atomic units, e^2) along each eigenvector; eigenvalues of the Hessian correspond to the curvatures (in atomic units, $1/a_0^2$) along the corresponding eigenvectors.

Anti 1

$\widetilde{\mathbf{H}}$

Curvature	3612.01	231.01	58.41	31.07	15.20	0.20	0.15	0.06	0.04
q_{C4}	-0.197	-0.182	-0.004	-0.048	-0.005	0.348	-0.463	0.575	0.510
q_{H4}	-0.582	-0.656	-0.150	-0.367	0.048	-0.132	0.149	-0.141	-0.110
q_{C3}	-0.202	-0.021	0.175	0.165	-0.054	0.644	-0.311	-0.171	-0.598
q_{H3}	-0.404	-0.004	0.723	0.416	-0.188	-0.289	0.079	0.033	0.117
q_{C2}	-0.200	0.106	-0.128	0.187	0.013	0.492	0.406	-0.460	0.528
q_{H2}	-0.400	0.191	-0.605	0.574	0.061	-0.246	-0.147	0.099	-0.112
q_{C1}	-0.194	0.252	0.000	-0.118	0.036	0.219	0.637	0.606	-0.249
q_{H1}	-0.382	0.599	-0.026	-0.515	-0.377	-0.097	-0.226	-0.156	0.050
q_{Cl}	-0.189	0.258	0.201	-0.137	0.901	-0.053	-0.129	-0.082	0.029

$\widetilde{\mathbf{\Sigma}}$

Variance	1.24E-06	1.81E-05	8.78E-05	1.53E-04	3.00E-04	2.27E-02	3.05E-02	8.06E-02	1.18E-01
q_{C4}	0.196	-0.181	0.006	0.053	-0.011	-0.533	0.314	0.672	-0.303
q_{H4}	0.580	-0.652	-0.098	0.395	0.019	0.190	-0.090	-0.156	0.057
q_{C3}	0.202	-0.024	0.157	-0.183	-0.039	-0.658	0.095	-0.363	0.570
q_{H3}	0.404	-0.005	0.671	-0.515	-0.129	0.286	0.019	0.072	-0.130
q_{C2}	0.202	0.102	-0.144	-0.158	0.036	-0.362	-0.602	-0.282	-0.573
q_{H2}	0.406	0.182	-0.663	-0.493	0.143	0.179	0.229	0.060	0.104
q_{C1}	0.193	0.252	0.014	0.118	0.012	0.033	-0.629	0.529	0.457
q_{H1}	0.381	0.595	0.017	0.455	-0.454	0.017	0.236	-0.140	-0.109
q_{Cl}	0.180	0.284	0.234	0.231	0.868	-0.004	0.133	-0.072	-0.055

Table S6, Continued.

	Anti 2								
Curvature	3614.47	230.35	64.53	30.14	9.96	0.19	0.14	0.03	0.06
					$\tilde{\mathbf{H}}$				
q_{C4}	0.195	-0.184	-0.005	-0.046	0.012	0.355	0.417	0.539	-0.579
q_{H4}	0.578	-0.666	0.092	-0.374	-0.046	-0.134	-0.137	-0.116	0.145
q_{C3}	0.200	-0.017	-0.149	0.184	0.052	0.628	0.337	-0.616	0.115
q_{H3}	0.398	0.008	-0.634	0.499	0.300	-0.274	-0.096	0.122	-0.019
q_{C2}	0.201	0.105	0.145	0.169	-0.062	0.507	-0.367	0.485	0.517
q_{H2}	0.405	0.184	0.637	0.502	-0.186	-0.256	0.143	-0.097	-0.116
q_{C1}	0.196	0.252	0.001	-0.126	-0.011	0.222	-0.675	-0.231	-0.569
q_{H1}	0.389	0.562	-0.269	-0.425	-0.433	-0.115	0.242	0.052	0.132
q_{Cl}	0.185	0.312	0.259	-0.312	0.824	-0.029	0.123	0.019	0.091
					$\tilde{\mathbf{\Sigma}}$				
Variance	1.23E-06	1.96E-05	7.95E-05	1.29E-04	4.07E-04	2.31E-02	3.43E-02	1.27E-01	6.31E-02
q_{C4}	0.196	0.184	-0.002	0.043	-0.008	-0.480	0.384	-0.561	0.483
q_{H4}	0.580	0.672	-0.143	0.345	0.029	0.165	-0.127	0.121	-0.116
q_{C3}	0.203	0.005	0.177	-0.154	-0.028	-0.614	0.299	0.635	-0.182
q_{H3}	0.408	-0.037	0.713	-0.386	-0.282	0.267	-0.071	-0.128	0.054
q_{C2}	0.201	-0.107	-0.118	-0.192	0.079	-0.404	-0.420	-0.438	-0.599
q_{H2}	0.401	-0.180	-0.541	-0.593	0.219	0.235	0.168	0.083	0.155
q_{C1}	0.193	-0.250	-0.033	0.124	0.004	-0.246	-0.680	0.221	0.558
q_{H1}	0.384	-0.569	0.159	0.508	0.391	0.118	0.244	-0.058	-0.121
q_{Cl}	0.177	-0.288	-0.327	0.210	-0.844	0.019	0.120	-0.014	-0.090

Table S6, Continued.

	Gauche 1								
Curvature	3635.17	179.83	82.72	38.41	10.68	0.23	0.13	0.03	0.07
					$\widetilde{\mathbf{H}}$				
q_{C4}	0.196	-0.185	0.036	-0.032	0.016	0.404	0.412	0.516	-0.570
q_{H4}	0.585	-0.663	0.280	-0.253	0.038	-0.154	-0.137	-0.111	0.141
q_{C3}	0.198	-0.028	-0.185	0.144	0.005	0.552	0.424	-0.618	0.196
q_{H3}	0.391	-0.074	-0.589	0.604	0.193	-0.234	-0.136	0.126	-0.058
q_{C2}	0.199	0.155	-0.122	-0.120	-0.097	0.493	-0.305	0.479	0.576
q_{H2}	0.396	0.397	-0.411	-0.587	-0.260	-0.240	0.121	-0.088	-0.155
q_{C1}	0.197	0.223	0.168	0.060	0.001	0.345	-0.663	-0.285	-0.489
q_{H1}	0.397	0.406	0.536	0.406	-0.356	-0.178	0.226	0.069	0.106
q_{Cl}	0.186	0.351	0.201	-0.146	0.870	-0.042	0.121	0.027	0.078
					$\widetilde{\mathbf{\Sigma}}$				
Variance	1.28E-06	2.55E-05	7.25E-05	1.24E-04	4.32E-04	2.06E-02	3.47E-02	6.61E-02	1.28E-01
q_{C4}	0.195	0.185	0.037	0.034	0.014	0.389	-0.439	0.541	-0.538
q_{H4}	0.578	0.662	0.295	0.253	0.054	-0.141	0.150	-0.118	0.132
q_{C3}	0.199	0.029	-0.190	-0.137	0.013	0.494	-0.497	-0.606	0.214
q_{H3}	0.399	0.076	-0.599	-0.580	0.216	-0.214	0.166	0.123	-0.068
q_{C2}	0.200	-0.155	-0.119	0.122	-0.109	0.503	0.288	0.450	0.597
q_{H2}	0.396	-0.385	-0.398	0.595	-0.275	-0.252	-0.106	-0.079	-0.161
q_{C1}	0.198	-0.223	0.168	-0.066	-0.028	0.420	0.601	-0.308	-0.494
q_{H1}	0.397	-0.386	0.515	-0.428	-0.374	-0.209	-0.209	0.078	0.107
q_{Cl}	0.190	-0.387	0.224	0.150	0.850	-0.046	-0.106	0.033	0.071

Table S6, Continued.

	Gauche 2								
Curvature	3620.41	171.98	84.31	39.25	7.82	0.29	0.13	0.07	0.04
					$\tilde{\mathbf{H}}$				
q_{C4}	0.198	-0.184	0.031	-0.034	-0.021	-0.278	-0.525	-0.491	0.575
q_{H4}	0.589	-0.667	0.259	-0.255	-0.022	0.124	0.171	0.124	-0.125
q_{C3}	0.199	-0.022	-0.183	0.143	-0.012	-0.436	-0.538	0.144	-0.636
q_{H3}	0.396	-0.055	-0.576	0.620	-0.178	0.181	0.200	-0.059	0.127
q_{C2}	0.197	0.161	-0.121	-0.129	0.064	-0.436	0.080	0.707	0.451
q_{H2}	0.390	0.411	-0.412	-0.576	0.278	0.228	-0.024	-0.202	-0.090
q_{C1}	0.195	0.228	0.173	-0.002	-0.158	-0.602	0.554	-0.406	-0.142
q_{H1}	0.390	0.400	0.556	0.405	0.405	0.182	-0.122	0.047	0.020
q_{Cl}	0.189	0.334	0.209	-0.131	-0.835	0.216	-0.192	0.106	0.012
					$\tilde{\boldsymbol{\Sigma}}$				
Variance	1.49E-06	3.26E-05	6.07E-05	1.60E-04	8.51E-04	2.18E-02	4.63E-02	7.51E-02	1.55E-01
q_{C4}	-0.203	-0.180	0.026	-0.026	-0.010	-0.244	-0.592	-0.445	0.563
q_{H4}	-0.611	-0.657	0.242	-0.248	-0.029	0.112	0.187	0.112	-0.120
q_{C3}	-0.198	-0.012	-0.180	0.155	0.009	-0.407	-0.518	0.134	-0.673
q_{H3}	-0.394	-0.023	-0.561	0.638	-0.174	0.166	0.203	-0.056	0.143
q_{C2}	-0.191	0.164	-0.120	-0.129	0.082	-0.447	0.060	0.719	0.423
q_{H2}	-0.373	0.408	-0.429	-0.572	0.283	0.234	-0.015	-0.204	-0.081
q_{C1}	-0.190	0.231	0.170	-0.016	-0.154	-0.632	0.506	-0.439	-0.095
q_{H1}	-0.382	0.420	0.577	0.375	0.388	0.200	-0.104	0.049	0.010
q_{Cl}	-0.183	0.335	0.184	-0.150	-0.841	0.204	-0.187	0.120	-0.004

Table S6, Continued.

	Gauche 3								
Curvature	3613.95	196.26	79.13	37.21	10.82	0.29	0.13	0.07	0.03
					$\widetilde{\mathbf{H}}$				
q_{C4}	0.199	0.182	-0.042	0.004	0.030	0.301	0.512	-0.568	0.499
q_{H4}	0.594	0.654	-0.310	0.133	0.177	-0.131	-0.162	0.139	-0.109
q_{C3}	0.198	0.028	0.193	-0.121	-0.071	0.554	0.453	0.262	-0.569
q_{H3}	0.392	0.061	0.623	-0.554	-0.225	-0.238	-0.153	-0.074	0.108
q_{C2}	0.197	-0.141	0.112	0.159	-0.007	0.524	-0.321	0.481	0.543
q_{H2}	0.391	-0.351	0.371	0.698	0.060	-0.239	0.109	-0.121	-0.112
q_{C1}	0.195	-0.227	-0.161	-0.057	-0.026	0.397	-0.565	-0.559	-0.303
q_{H1}	0.385	-0.567	-0.397	-0.376	0.373	-0.170	0.198	0.144	0.064
q_{Cl}	0.193	-0.136	-0.379	0.067	-0.877	-0.105	0.099	0.072	0.037
					$\widetilde{\mathbf{\Sigma}}$				
Variance	1.35E-06	2.25E-05	6.83E-05	1.30E-04	3.82E-04	1.38E-02	3.63E-02	5.75E-02	1.36E-01
q_{C4}	0.201	0.176	0.045	-0.014	0.031	0.291	-0.502	-0.579	0.504
q_{H4}	0.603	0.635	0.306	-0.162	0.199	-0.128	0.163	0.136	-0.114
q_{C3}	0.197	0.031	-0.179	0.134	-0.072	0.508	-0.523	0.305	-0.530
q_{H3}	0.389	0.083	-0.548	0.613	-0.270	-0.212	0.176	-0.084	0.100
q_{C2}	0.196	-0.144	-0.135	-0.150	0.011	0.523	0.311	0.484	0.543
q_{H2}	0.389	-0.349	-0.460	-0.641	0.064	-0.250	-0.094	-0.130	-0.113
q_{C1}	0.192	-0.236	0.160	0.042	-0.017	0.460	0.519	-0.522	-0.355
q_{H1}	0.377	-0.582	0.406	0.354	0.365	-0.196	-0.187	0.130	0.080
q_{Cl}	0.195	-0.145	0.391	-0.141	-0.863	-0.098	-0.085	0.073	0.045

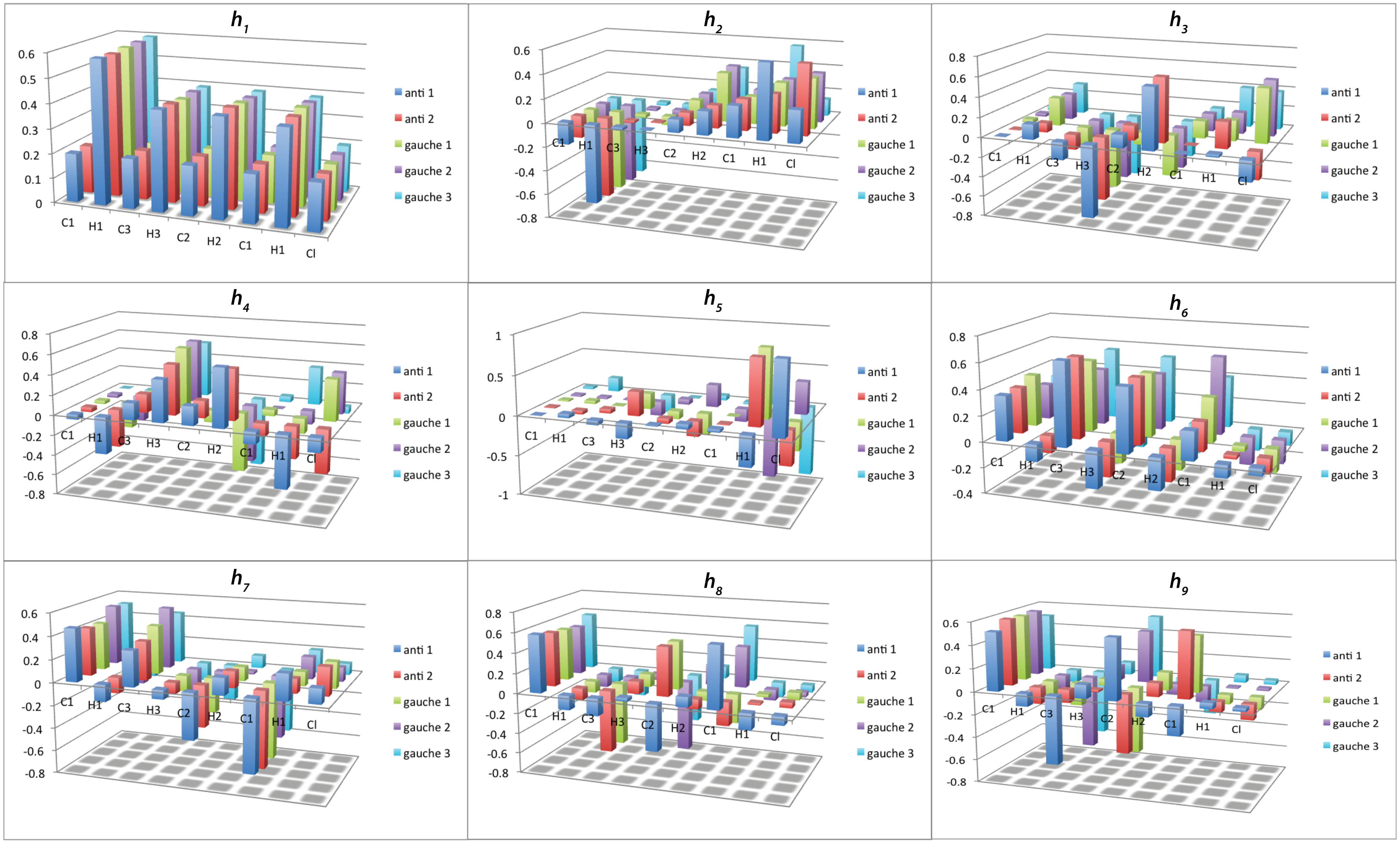


Figure S5. Bar-chart representation of the eigenvectors of the Hessian matrix for five conformers of 1-chlorobutane.

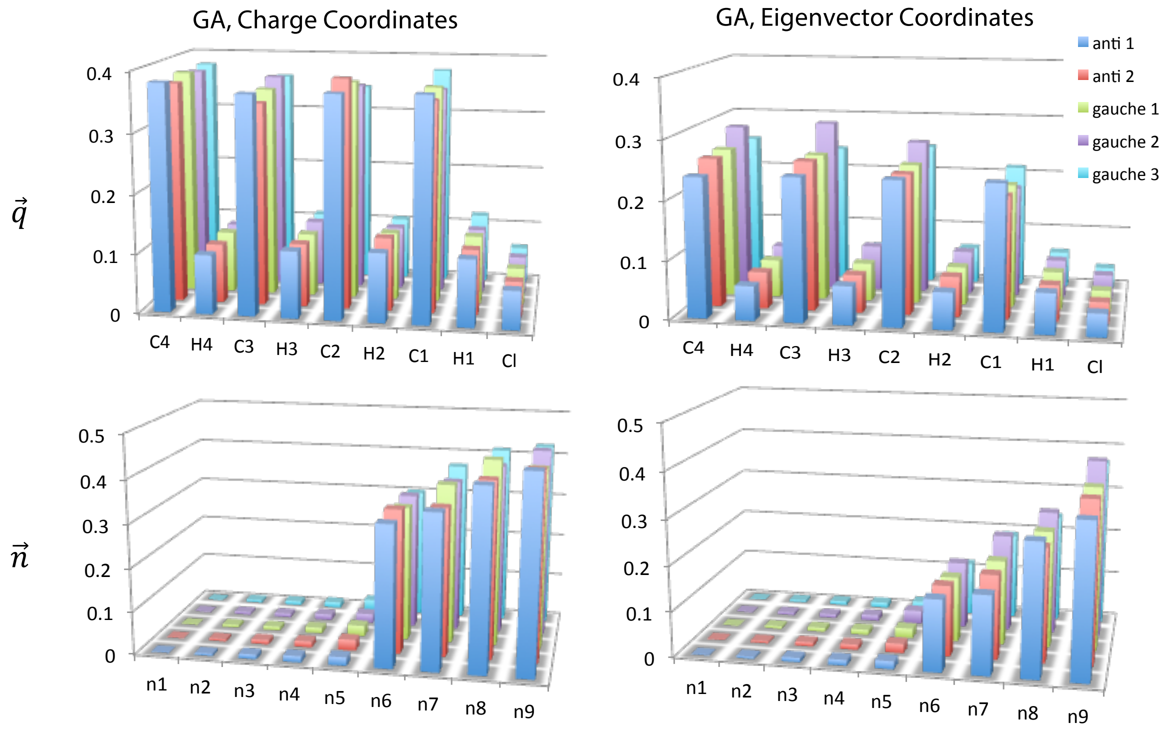


Figure S6. Standard deviations σ of the charges $\vec{q}$ and the corresponding coordinates defined by the LS-sum Hessian eigenvectors $\vec{n}$ obtained from the solutions of 200 GA performed in terms of the charge coordinates, and in terms of the LS-sum Hessian eigenvector coordinates (right).