Assessment of atomic charge models for gas-phase computations on polypeptides

T. Verstraelen,[a] E. Pauwels,[a] F. De Proft,[b] V. Van Speybroeck,[a] P. Geerlings,[b] M. Waroquier[a]

a) Center For Molecular Modeling, Ghent University, Technologiepark 903, 9050 Zwijnaarde, Belgium (Member of the QCMM Ghent-Brussels Alliance Group)

b) Department of General Chemistry (ALGC), Free University of Brussels - VUB, Pleinlaan 2, 1050 Brussels, Belgium (Member of the QCMM Ghent-Brussels Alliance Group)

Draft on Nov 11, 11

Abstract. The concept of the atomic charge is extensively used to model the electrostatic properties of proteins. Atomic charges are not only the basis for the electrostatic energy term in biomolecular force fields, but are also derived from quantum mechanical (QM) computations on protein fragments to get more insight into their electronic structure. Unfortunately there are many atomic charge schemes which lead to significantly different results, and it is not trivial to determine which scheme is most suitable for biomolecular studies. Therefore, we present an extensive methodological benchmark using a selection of atomic charge schemes (Mulliken, Natural, RESP, Hirshfeld-I, EEM and SQE) applied to two sets of penta-alanine conformers. Our analysis clearly shows that Hirshfeld-I charges offer the best compromise between transferability (robustness with respect to conformational changes) and the ability to reproduce electrostatic properties of the penta-alanine. The benchmark also considers two charge equilibration models (EEM and SQE), which both clearly fail to describe the locally charged moieties in the zwitterionic form of penta-alanine. This issue is analyzed in detail because charge equilibration models are computationally much more attractive than the Hirshfeld-I scheme. Based on the latter analysis, a straightforward extension of the SQE model is proposed, SQE+Q^0, that is suitable to describe biological systems bearing many locally charged functional groups.

Keywords. Proteins, Electrostatics, Population Analysis, ESP fitted charges, Charge Equilibration, Molecular Dipole.

1. Introduction

Electrostatic interactions determine many properties of proteins, such as the stability of protein folds,[1] the pKa shifts of acid and base residues,[2] the catalytic activity of enzymes,[3] compatibility with ligands or other proteins,[4] and so on. Protein electrostatics can be modeled at different levels of accuracy, ranging from a quantum-mechanical (QM) description of the electronic structure[5] to a coarse-grained approach where residues are modeled with just a few fixed point charges.[6,7]

Molecular Dynamics (MD) and Monte Carlo (MC) simulations based on force fields are essential tools to elucidate complex processes such as transitions between stable protein conformers,[8] protein folding,[9] and the interaction of proteins with various ligands.[10] For example, a straightforward MD simulation that follows the motion of a protein over a few nanoseconds can be analyzed with the Principal Component Analysis method to reveal the essential modes.[11] Extreme technological advances make it possible to run micro- and even millisecond MD simulations of proteins (with explicit solvent molecules) by following the motion of each atom with steps of 1 femtosecond. Two compelling examples are the Folding at Home project,[9] where distributed computing algorithms are used to study the folding of small proteins, and the Anton computer,[12] which is built from scratch to run protein MD simulations at an unparalleled efficiency. One of the most practical applications of protein MD is the in silico investigation of a pharmaceutical substance interacting with the active site of an enzyme, which allows entire databases of drug candidates to be screened before they are actually synthesized.[13] All MD applications given above require an accurate model for the non-covalent interactions. Hence, a detailed understanding and an accurate empirical description of protein electrostatics is of utter importance.

All-atom molecular mechanics force fields such as AMBER,[14] CHARMM[15] and OPLS-AA[16] are the most wide-spread models that approximate the potential energy surface of proteins. In such a force-field model, electrostatic interactions are treated by placing fixed effective point charges at the positions of the nuclei. In these fixed charge models, it is possible to approximate screening effects due to the polarizability of the protein and the surrounding solvent with a dielectric continuum background, where the dielectric constant inside the protein is different from the solvent. The electrostatic energy and nuclear forces in these dielectric continuum models can then be computed

with the Poisson-Boltzmann Method and the Boundary Element Method, or approximated with the Generalized Born and the Surface Generalized Born models.[17,18]

Instead of using a dielectric background for the protein and its environment, one may also treat electronic polarization explicitly. This can be done by introducing inducible dipoles at the nuclei of polarizable atoms, and by allowing atomic charges to fluctuate due to changes in the electrostatic potential. Recent examples of such models are PFF,[19?21] CHARMM with fluctuating charges,[22,23] polarizable AMBER[24] and SIBFA.[25] The main incentive for these models is a more accurate computation of the electrostatic interactions. A dielectric background can not take into account many-body effects, for example, the polarization of a hydrogen bond acceptor upon the formation of the hydrogen bond. For uniform systems such as liquid water, many-body polarization effects can be included in an averaged enhanced pair potential by rescaling the atomic charges and dispersion parameters. However, the proper amount of scaling is an empirical factor that is hard to determine, and the corrections are not simply transferable to non-uniform systems.[26] Polarizable force fields avoid these difficulties with a physically justified model.

Given the numerous applications of force-field models that use atomic charges to model electrostatic interactions, a proper assignment of atomic charges is an essential step in the calibration of force field parameters. For small model systems, it is possible to run reliable QM computations, and derive atomic charges from the distribution of the electronic density or the electrostatic potential (ESP) surrounding the molecule. Even beyond the realm of force fields, as QM computations become feasible on larger and larger polypeptide models (and eventually entire proteins), atomic charges derived from QM computations are an important tool to characterize the electronic structure. For example, it is shown that Natural[27] charges derived from DFT computations on the reduced cysteine residue and its direct protein environment correlate well with the pKa shift of the cysteine.[28] It is clear that atomic charges are often used as a tool to understand and model electrostatic interactions.

Despite the obvious meaning of an atomic charge for an experimental chemist, it is non-trivial to define charges for atoms in molecules (AIM) from the quantum-mechanical perspective. Many theoretical schemes were proposed in the literature over the past 60 years to derive AIM charges, often leading to incompatible results. Among the most

popular schemes, we have Mulliken population analysis,[29] Lowdin population analysis,[30] Hirshfeld partitioning,[31] Natural population analysis,[27] Bader's AIM scheme,[32] and Hirshfeld-I partitioning.[33] In addition to the AIM schemes, atomic charges can also be fitted to reproduce the electrostatic potential around a molecule,[34] with for example the Merz-Kollman[35], RESP[36], REPEAT[37] charges. Although ESP fitted charges can accurately reproduce the molecular multipole expansion and electrostatic interactions between molecules, they are often showing unexpected trends that make no chemical sense.[34] For the direct interpretation of the atomic charges, or for the calibration charge equilibration models, such unexplainable trends are prohibitive. In addition to charges derived from QM computations, one can also compute atomic charges with charge equilibration models such as the Electronegativity Equalization Method[38,39] (EEM) and the Split Charge Equilibration[40] (SQE). These models are calibrated to reproduce QM charges on relatively small molecules, however the transferability of the parameters to extended systems such as complete protein models needs to be tested carefully.[41,42]

Atomic charges derived from quantum-mechanical computations or charge equilibration models should in general satisfy two requirements: (i) they should give a reasonable estimate of the molecular electrostatic interactions, and (ii) the charges should be well-behaved. The latter means that the charges should only depend on the actual electronic distribution and should not be sensitive to purely methodological parameters, e.g. choice of the basis set, choice of the grid points used for the ESP fit, very small changes in the geometry, and so on. In the case of charge equilibration models, one should also carefully check that both requirements are fulfilled with a set of parameters that does not depend on the size of the molecule.

In this work, we conduct a series of critical tests, which will be referred to as benchmarks hereafter, to validate in how far a selection of charge schemes and charge equilibration models fulfill the two requirements given above. Ideally, such benchmarks should validate to what extent each atomic charge model is capable of reproducing experimentally observed electrostatic properties of proteins, taking into account the effect of the surrounding solvent. However, for reasons outlined below, results from MP2/Aug-cc-pVDZ//B3LYP/6-31G(d) computations on isolated penta-alanines, i.e. without surrounding solvent, will be used as a reference. A direct comparison of the atomic charge models with computational (as opposed to experimental) results allows a more in-depth analysis,

because a large amount of detailed information can be used for comparison. These reference computations do not include any solvent to simplify the analysis. The main effect of a solvent on the electrostatic interactions, e.g. dielectric screening and increased dipole moments, can be described in terms of charge transfer and polarization effects, which are included in all charge models in this work, except for contributions from induced atomic dipoles and higher-order multipoles. For example, the penta-alanine charges derived from the MP2 computations will show a response when explicit water or a dielectric continuum is added, because of the response of the MP2 electronic structure. Similarly, the EEM and SQE models also describe how atomic charges change when an external field is applied. Therefore, all these charge models should work for both gas-phase and condensed phase computations. However, in this study we limit the benchmarks to gas-phase computations because these already reveal several insights and problems that do not depend on the presence of a solvent, and need to be resolved first before performing more in depth studies on proteins in condensed phases.

The structure of the paper is as follows. The following section discusses all methodological aspects of the benchmarks: the generation of the penta-alanine conformers, the selection of charge models, and the specification of the benchmarks. In the third section we compare the performance of all charge schemes with respect to various benchmarks. Because the charge equilibration models show very specific errors observed for the first time, these errors are analyzed in detail, leading to an improved version of the SQE model. The last section summarizes the main conclusions.

2. Computational Methods

2.1. Penta-alanine model systems

The penta-alanine molecular system is used to conduct our methodological benchmarks. The limited number of atoms in this system makes it possible to perform an extensive sampling of the conformational space, and run MP2[43] single-point energy computations to properly take into account electron correlation effects. On the other hand, the system is sufficiently large to exhibit internal hydrogen bonds and polarization effects. Hence it features all types of electrostatic interactions that are found in larger polypeptides or proteins with thousands of atoms. Two

forms of the penta-alanine system are studied: the terminally blocked and the zwitterionic form, which will be referred to by shorthands TB and ZI in the remainder of the text. Both are shown in Figure 1. The ZI form has two charged residues at the end points, whereas the TB form has only neutral residues. Each structure is divided into fragments, which consist of a single residue each. In the TB form, the terminal blocks are also put in separate fragments.

For both the TB and ZI form of penta-alanine, more than 100 distinctive conformers were generated to examine how the electrostatic properties and the charge distribution depend on the geometry of the polypeptide. Both sets of conformers were generated with an autonomous algorithm that does not rely on any subjective human intervention. It consists of the following three steps:

1. A large set of trial geometries is generated, starting from a linear peptide chain, by randomly rotating the φ and ψ in the backbone over the interval [0, 2π]. Structures in which non-bonded atom pairs have an inter-atomic distance below 0.6 times the sum of their Van Der Waals radii, are excluded. *Non-bonded atom pairs* within the same molecule are defined as those pairs that are separated by at least four covalent bonds.

2. Each geometry is optimized at the B3LYP/6-31G(d)[44,45] level to the nearest local energy minimum. If the bond graph changes during the optimization, e.g. due to a proton transfer, the geometry is discarded. Duplicate geometries are also excluded.

3. On the remaining geometries MP2/Aug-cc-pVDZ[43,46] single-point computations are carried out. Geometries are rejected when the corresponding MP2 computation shows convergence problems in the Hartree Fock part.

At each step, this algorithm rejects some samples because they have an unrealistic geometry. Therefore, one has to start with a sufficiently large set of random geometries in the first step to obtain at least 100 samples for the two penta-alanine forms. In this specific application, using 109 and 1538 starting structures for the B3LYP/6-31G(d) optimizations, we ultimately ended up with 103 and 134 successful samples for system TB and system ZI,

respectively. Cartesian coordinates for all conformers are provided as additional information. Gaussian09[47] was used for all B3LYP and MP2 computations in this work.

Internal hydrogen bonds are formed during the B3LYP/6-31G(d) optimization of the randomized structures. Because of the polarity and the dynamic nature of hydrogen bonds, they are an excellent tool to study internal polarization effects: when the donor (H) and the acceptor (O or N) approach, they can lower their potential energy by increasing the absolute value of their partial charges. To study such effects systematically in our dataset, we introduce a formal definition of all non-bonded O-H pairs that are prone to such a polarization. All pairs that have an interatomic distance below 2.5 Å, are called nearby non-bonded O-H pairs in the remainder of the text. The TB and ZI forms count on average 4.17 and 4.45 nearby non-bonded O-H pairs, respectively. An upper limit of 2.5 Å is chosen because it corresponds to a minimum in the histogram of the non-bonded O-H distances present in the samples of the two penta-alanine systems. (See Figure 2.)

2.2. Atomic charge models

Two types of atomic charge models are used: (i) a selection of atomic charge schemes applied to the MP2 single-point computations and (ii) charge equilibration models (EEM and SQE) that can be applied without prior electronic structure computation.

The first class of charge models derive atomic charges from the electron density, or the electrostatic potential, based on a quantum-mechanical electronic structure computation. We will refer to these methods in the remainder of the paper as *QM charges*. Such approaches have the advantage that the underlying quantum-chemical computation is accurate and transferable to a wide variety of chemical systems. Unfortunately, many algorithms for QM charges exist and they lead to very different results. Moreover, the scaling of the computational cost of an electronic structure computation in terms of the number of electrons limits the use of this approach on full proteins. Due to these limitations, our methodological assessment of a selection of charge schemes is limited to penta-alanine models. In this work we compute the Mulliken charges,[29] the Natural charges,[27] restrained electrostatic potential (RESP) fitted charges,[36] and Hirshfeld-I[33] charges for the MP2/Aug-cc-pVDZ single-point computations

on the 103 TB structures and 134 ZI structures. For the first two schemes Gaussian09 was used. RESP charges were computed with the RESP program from the Antechamber package,[48] using the default hyperbolic restraint on the heavy-atom charges with amplitude of 0.0005 a.u. For the Hirshfeld-I population analysis, we used our in-house code, HiPart.[49]

Charge equilibration models, on the other hand, can be applied to entire proteins with very modest computational requirements. The charge equilibration models considered in this work are the Electronegativity Equalization Method (EEM) and the Split Charge Equilibration (SQE). We will refer to the EEM and SQE charges in the remainder of the text as *EQ charges*. Parameters for both models were calibrated in a previous paper[50] based on Hirshfeld-I charges derived from MP2/Aug-cc-pVDZ computations on 500 small organic molecules. We use the HETS [**H**irshfeld-I, **E**EM, **T**rivial atom types, **S**tatic cost function] parameters for the EEM model and the HSFF [**H**irshfeld-I, **S**QE, **F**orce-field atom types, **F**ull cost function] parameters for the SQE model. Both sets of parameters were calibrated with a non-linear least squares procedure using the Hirshfeld-I charges from the 500 small molecules as reference data. An extensive analysis revealed that the introduction of force-field atom types does not lead to a significant improvement of the EEM model,[51,50] while it does make SQE model much more accurate. The EEM calibration relied only on the equilibrium charge distribution of the 500 molecules (Static cost function) and did not include the response of these charges to perturbations in the external field, simply because the EEM is not capable of reproducing response data. For the SQE calibration, both static and response data were used for the calibration (Full cost function). EEM and SQE charges are computed for all penta-alanine conformers, and two protein models: human HIV-2 protease[52] and bacterial 3 alpha, 20 beta-hydroxysteroid dehydrogenase.[53] The latter two protein structures were taken from the protein data bank (accession codes 1HSG and 1HSD, respectively), and protonated corresponding to a neutral pH. The generated initial structures were then optimized with the CP2K program[54] using the CHARMM force field.[55]

The MP2/Aug-cc-pVTZ level of theory was used in a previous paper for the calibration of the EEM and SQE models, and was reused in this work for the single-point computations on the penta-alanine conformers. This

facilitates the comparison of the EEM and SQE charges with the Hirshfeld-I charges. Such comparison is also a good test for the transferability to proteins of EEM and SQE calibrations based on small molecules.

2.3. Benchmark protocols

In order to answer the question posed in the title of this paper, two *general* benchmarks are carried out on the QM and the EQ charges of penta-alanine. A third *specific* benchmark is only applied to the EQ charges.

The first benchmark consists of a comparison of the MP2 dipole moments of the penta-alanine structures with the dipole moments derived from the atomic charges. Despite the simplicity of this test, it is a very effective tool to measure the electrostatic performance of an atomic charge scheme. The molecular dipole moment is the leading term in the multipole expansion of the molecular charge distribution, and is therefore the most significant quantity to determine the electrostatic interactions that the pentapeptide chain could have with other protein fragments or a solvent. Because both the TB and ZI set contain very different conformers - penta-alanine is very flexible - the dipole moment fluctuates heavily. Each conformer represents an alternative superposition of dipole moments and charges in the quasi rigid residues. Hence, an atomic charge scheme that can reproduce all the penta-alanine dipole moments, is an effective model from the electrostatic point of view.

In the first place, the dipole moments are compared by making scatter plots of the X, Y and Z components of the MP2 dipole moment versus the dipole moment components derived from the atomic charges. (The Cartesian frame axes will be specified below.) Second, for each charge model, each component of the dipole moment $\mathbf{p}_n$ of conformer n, and each penta-alanine form, the following statistical parameters are derived:

- The root-mean-square error (RMSE): $\sigma_{p_\alpha} = \sqrt{\frac{\sum_{n=1}^{N_{\mathrm{conf}}} \left(p_{n,\alpha}^{\mathrm{MODEL}} - p_{n,\alpha}^{\mathrm{MP2}}\right)^2}{N_{\mathrm{conf}}}}$, where N_{conf} is the number of conformers of the TB or ZI form, and α stands for the X, Y or Z component.

- The relative error: $\mathrm{RE}_{p_\alpha} = \frac{\sigma_{p_\alpha}}{\mathrm{RMS}_{p_\alpha}}$, where $\mathrm{RMS}_{p_\alpha} = \sqrt{\frac{\sum_{n=1}^{N_{\mathrm{conf}}} \left(p_{n,\alpha}^{\mathrm{MP2}}\right)^2}{N_{\mathrm{conf}}}}$ is the root-mean-square value of the MP2 reference dipole components.

- Slope and intercept: estimates for the parameters A_α (slope) and B_α (intercept) obtained solving the following set of equations in the least-squares sense:

$$A_\alpha \times p_{n,\alpha}^{\mathrm{MP2}} + B_\alpha = p_{n,\alpha}^{\mathrm{MODEL}} \quad \forall\, n = 1 \ldots N_{\mathrm{conf}}$$

The first two error measures are based on the hypothesis $p_{n,\alpha}^{\mathrm{MODEL}} \approx p_{n,\alpha}^{\mathrm{MP2}}$, which corresponds to what one would expect from the charge models: they should give a correct prediction of the dipole moment. The linear regression is only used to describe to nature of the errors, and is not used to compute the first two error measures. An additional advantage of this choice is that the error measures are not over-sensitive to larger values of the dipole moment in the dataset. The definitions of RMSE and Relative error only depend on difference between MODEL and MP2 values for the dipole moment. Data corresponding to relatively high or low values for the MP2 reference data have the same weight in these error measures. Because the slope and intercept are obtained from the least squares analysis, they tend to be more sensitive to those datapoints that correspond to the highest and the lowest values of the MP2 reference dipole moments. Therefore, the robustness of these parameters is verified, i.e. all fits were repeated without using the lowest and highest 10% of the reference data. In case the second fit results in similar parameters and Pearson R^2 values, one can conclude that the least squares analysis is robust.

In a second benchmark we investigate the geometry dependence of the charges. Although some fluctuations in the charges due to internal polarization are to be expected, such variations should remain limited because all structures have virtually the same bond lengths and valence angles. When the charges on a given atom vary to a large extent between different conformers, it is more likely a methodological issue, rather than a genuine trend.

Because the EEM and SQE parameters in this work are based on Hirshfeld-I charges derived from MP2/Aug-cc-pVDZ computations on 500 small molecules, one expects that the EEM and SQE charges for the

penta-alanine molecules should correlate well with the corresponding Hirshfeld-I charges computed in this work. In the third benchmark, we investigate the deviations of the EQ charges from the Hirshfeld-I charges.

2.4. Propagation of errors

In the first benchmark, dipole moments based on atomic charges are compared to the MP2 dipole moments. In order to analyze the origin of the errors on the dipole moments computed with the atomic charges, it is helpful to derive the propagation of errors on atomic charges to errors on the components of the molecular dipole moment. The error analysis assumes that the computed penta-alanine charges are perturbed by some random normal error (in practice caused by a numerical problems and the neglect of atomic dipoles) such that the correct molecular dipole moment is not reproduced. As a side-effect, this analysis also provides a unique Cartesian reference frame for each penta-alanine structure, and facilitates the comparison of the MP2 dipole moments with the charge-based dipole moments.

The dipole moment is computed from the partial charges as follows:

$$\mathbf{p} = \sum_{i=1}^{N} q_i \mathbf{r}_i \tag{1}$$

where q_i are the partial charges and $\mathbf{r}_i$ are the position vectors of the atoms. If the errors on the charges would be independent, the covariance matrix would take the following form:

$$\sigma_q^2 I, \tag{2}$$

where σ_q is the standard error on an atomic charge, e.g. 0.05e, I is the identity matrix. However, the errors on the charges can not be uncorrelated due to the total charge constraint. All charge schemes in this paper correctly reproduce the total charge, hence errors cancel out when taking the sum over all charges. Therefore we adopt the following covariance matrix for the partial charges:

$$C^{(q)} = \frac{\sigma_q^2}{1 - N^{-1}} \left(I - N^{-1} d d^T \right) \tag{3}$$

where d is a column vector with N elements all equal to one. The diagonal elements of this matrix are identical to those in eq 2, but the off-diagonal elements are modified such that the variance of the sum of all charges becomes zero. One could also construct other forms for the covariance matrix such that the covariance on the total charge is zero. However, this form is preferable because its spectrum is very similar to that of eq 2. One can show that vector d is the eigenvector of eq 3 with eigenvalue zero, and that all other eigenvalues are equal to $\frac{\sigma_q^2}{1-N^{-1}}$. The individual elements of the covariance matrix take the following form:

$$C_{ij}^{(q)} = COV[q_i, q_j] = \frac{\sigma_q^2}{1 - N^{-1}} \left(\delta_{ij} - N^{-1} \right) \tag{4}$$

The covariance of the components of the dipole vector can be derived from the covariance on the charges as follows:

$$COV[p_x, p_y] = \sum_{i=1}^{N} r_{i,x} \sum_{j=1}^{N} r_{j,y} COV[q_i, q_j] \tag{5}$$

After some trivial substitutions, one gets:

$$COV[p_x, p_y] = \frac{\sigma_q^2}{1 - N^{-1}} \sum_{i=1}^{N} (r_{i,x} - \langle r_x \rangle)(r_{i,y} - \langle r_y \rangle) \tag{6}$$

where $\langle r_x \rangle$ and $\langle r_y \rangle$ are Cartesian components of the geometric center of the atomic coordinates. The diagonalization of this dipole covariance matrix yields three eigenvalues ($\sigma_{p1} < \sigma_{p2} < \sigma_{p3}$) and three orthogonal eigenvectors that will be used as the new X, Y and Z axes. The direction of each basis vector is chosen such that the vector from the N to the C terminus of penta-alanine has positive coordinates in the new axes frame. Note that

the new axes must be computed for each conformer separately. All the dipole moments in the remainder of the paper are given with respect to the new axes.

The errors on the components of the dipole vector due to the errors on the partial charges are uncorrelated in the new coordinates, with the largest error along the Z axis and the smallest error along the X axis. This covariance matrix also has a geometrical interpretation: when the factor $\frac{\sigma_q}{1-N^{-2}}$ is set to one, eq 6 becomes the covariance of the atomic coordinates with respect to the geometric center. This means that the molecule is the most elongated along the new Z axis and the most compact along the new X axis.

3. Results and Discussion

3.1. General benchmarks

The first two benchmarks outlined in the previous section are applied to both the terminally blocked (TB) and the zwitterionic (ZI) penta-alanine. The first benchmark compares the dipole moments derived from the atomic charges, using different charge schemes, with the MP2 dipole moments. The second benchmark tests the sensitivity of the atomic charges to conformational changes of the penta-alanine model.

For the first benchmark, the X, Y and Z components of the charge-model dipole moment are plotted versus the MP2 dipole moment in Figure 3, for all charge models considered in this paper: Mulliken, Natural, RESP, Hirshfeld-I, EEM and SQE. The statistical parameters of the scatter plots are given in Table 1. The root-mean-square error of the dipole moment components derived from different charge schemes varies over two orders of magnitude. In order of increasing RMSE we get: RESP, Hirshfeld-I, Natural, SQE, EEM, and Mulliken. The results for the ZI system are comparable to the TB system, except for the SQE model where the errors on the dipole moment of system ZI are clearly larger than for system TB. Considering the relative errors, only the RESP and the Hirshfeld-I schemes give a quantitative description of the penta-alanine dipole moment in terms of atomic partial charges.

The least squares analysis of the errors in Table 2 reveals similar results. The intercepts are somewhat ill-defined for the Mulliken scheme, which is due to the magnitude and the unpredictable nature of the errors. For all other schemes, the parameters do not change significantly when only the *central 80%* of the data is used, confirming the robustness of the corresponding least squares parameters. The most important observation for further analysis is that the Natural scheme systematically overestimates the dipole moments (slope > 1), while the EEM and SQE schemes underestimate the dipole moments (slope < 1). Table 2 also shows that the Pearson R^2 value is close to 100% for all fits related to the Natural, RESP and Hirshfeld-I data, and also for some fits based on the SQE data. This does not imply that all these models predict the dipole moments with the same accuracy, but rather that their errors can all be *explained* to a comparable extent with a linear model.

Mulliken charges manifestly fail in reproducing the electrostatic properties of the penta-alanine, which is not surprising. The method has known weaknesses,[27,56] which will be illustrated in detail in the second benchmark. Natural charges slightly overestimate the dipole moment, which is in line with earlier work,[50] where it was found for a set of 500 small organic molecules that Natural charges overestimate the amplitude of the electrostatic potential. The RESP charges perfectly reproduce the dipole moment. This is not surprising as these charges are fitted to reproduce the ESP on a set of grid points surrounding the penta-alanine, and these ESP grid data are mainly determined by the first terms in the multipole expansion of the molecular system at hand. It is therefore trivial that any type of ESP-fitted charges will reproduce the molecular dipole moment.[36] The Hirshfeld-I charges are only slightly worse than the RESP charges in this test. This observation conforms to earlier work,[57] where it was found that Hirshfeld-I gives in general a good prediction of the ESP surrounding the molecule. This can be understood as follows: the Hirshfeld-I scheme partitions the molecular electron density into nearly-spherical atomic contributions. Due to their sphericity, the atomic densities have relatively small dipoles and higher order multipoles, which results in a good reproduction of the ESP by truncating each atomic multipole expansion after the monopole.

The dipole moments predicted by both the EEM and SQE model underestimate the MP2 reference data. The correlation is clearly worse compared to the RESP and Hirhsfeld-I dipoles. The EEM and SQE parameters are

derived from Hirshfeld-I charges from a set of 500 small molecules, and are shown to be transferable to similar small molecules.[50] However, the results in Table 1 and Figure 3 show that these parameters are not suitable to compute dipole moments on larger systems such as the penta-alanine models in this work. Especially the large errors of the SQE model for the Zwitterionic system prompt for a detailed analysis, which will be given in the following subsections.

The EEM model has one major weakness that is reported extensively in the literature: it dramatically overestimates the polarizability in the limit of large molecules.[40,41,58] One of the side effects is that the EEM scheme allows (in large systems) a metallic internal polarization that pushes the leading multipole moments towards zero. This issue is illustrated with an example in Figure 4, where the electrostatic potential due to the EEM and SQE charges are visualized for two proteins: the human HIV-2 protease and the Bacterial 3 alpha, 20 beta-hydroxysteroid dehydrogenase, which will be referred to further on by their pdb accession codes, 1HSG and 1HDC. In order to visualize the trends in the ESP at the nanometer scale, local effects due to individual atoms are blurred by a convolution of the ESP with a Gaussian function with $\sigma = 5\text{Å}$. The blurred ESP function is plotted with a color scale on the cartoon visualization of both proteins. This visualization shows that the ESP due to the EEM charges is virtually constant, except for local atomic fluctuations. Such behavior is typically associated with metallic objects, while proteins are conventionally modeled as insulators with a relative dielectric constant between 2 and 20.[59] This discrepancy shows that the EEM is clearly not applicable to large molecular systems. The SQE model is proposed to overcome this metallic behavior,[40] which is clearly visible in Figure 4. The gradient of the electrostatic potential in the alpha helices due to the alignment of hydrogen bonds is correctly reproduced. Nevertheless, the discrepancies between the SQE dipole moments and the MP2 reference data in figure 3 imply that the SQE potential maps in figure 4 still contain other qualitative errors that would hamper a direct comparison with MP2 or experimental reference data.

The purpose of the second benchmark is to determine the robustness of the charges obtained with various schemes with respect to conformational changes. Because the conformers of systems TB and ZI do not show large variations in bond lengths and valence angles, we do not expect very large differences in the net atomic charges. If

one of the charge schemes does show such excessive fluctuations, these are rather due to methodological issues instead of genuine changes in the electron distribution.[60] Figure 5 depicts the variations in charge for each atom in the systems TB and ZI. The figures immediately reveal that the Mulliken and RESP charges show an exorbitant geometry dependence, which can be seen in the large range of the fluctuations on the charges. All other charge schemes, are much more well-behaved in this test.

The lack of robustness of the Mulliken and RESP charges are detrimental for further statistical applications. For example the calibration of EEM or SQE parameters based on such poorly defined charges can not be successful, simply because the noise on the reference data will result in noise on the estimated parameters. Also a direct (chemical) interpretation of such charges is simply impossible and not reproducible: a change of conformation can result in completely different charges without a clear physical explanation.

The noisy nature of the Mulliken and RESP charges is illustrated in more detail in Figure 6. For this figure, a relaxed potential energy surface scan was performed along the ψ angle of the alanine dipeptide molecule, using the B3LYP/6-31G(d) level. At each stationary point of the relaxed scan, independent MP2/Aug-cc-pVTZ single-point computations were carried out. The charge on the α carbon (obtained with each of the six atomic charge schemes in this work) is plotted as function of the dihedral angle ψ in the top panel of Figure 6. The Mulliken charge exhibits large fluctuations and the RESP charge shows a stochastic behavior, which renders both schemes useless for a direct interpretation. All four other charge schemes show much smaller and more deterministic variations with the dihedral angle. The small fluctuations in the α carbon charge can be correlated with large geometric changes at three points along the relaxed scan. In the bottom panel in Figure 6, the Cartesian RMSD between two subsequent stationary points is plotted as function of ψ. RESP charges may still be useful if one is only interested in reproducing the electrostatic potential for a fixed geometry.

The weaknesses of the Mulliken and RESP charges are well-known, although the relation with an erroneous geometry dependence is rarely demonstrated.[61,62] The problem of the Mulliken scheme is that the charges are directly derived from the expansion coefficients of the density matrix in an atomic orbital basis. The so-called net

atomic charges are trivial to assign, but the contributions to the density due to pairs of basis functions on different atoms are divided over both atomic populations by an ad hoc 50/50 rule.[29] When diffuse basis functions are present, degeneracies easily occur in the expansion of the density matrix. This means that two virtually equivalent ground states may be written with different expansions, and hence result in different Mulliken charges.[63]

ESP-fitted charges are typically ill-defined due to the rank-deficiency of the ESP cost function. This results in a large sensitivity of ESP fitted charges to small influences such as the choice of the grid points, the orientation of the molecule with respect to the grid, and (as we also observe here) the molecular geometry.[34]

Only the Hirshfeld-I scheme has a good performance in both benchmarks: these charges both reproduce the dipole moments and are robust with respect to geometrical changes. With the Hirshfeld-I scheme it is indeed possible to rationalize protein electrostatics in terms of atomic charges: these charges give a quantitatively correct picture of the ESP and are suitable for a chemical interpretation. The latter is possible because Hirshfeld-I charges are less affected by methodological defects which are present in the RESP and Mulliken schemes. It is noteworthy that RESP is still extensively used for the calibration of atomic charge parameters in the AMBER force field.[64] Using the RESP scheme, one has to fit charges using ESP grid data from a large number of conformers of a given target molecule in order to obtain a set of charges that work for all these conformers,[65] while Hirshfeld-I charges only need to be computed for one conformer and are inherently transferable to other conformers. We conclude, based on the above benchmarks, that the Hirshfeld-I scheme is to be preferred over the RESP scheme to determine atomic charge parameters in force-field models.

Despite the good performance of the Hirshfeld-I method, it is not generally applicable to full proteins with current-day computing power, because Hirshfeld-I depends on an accurate quantum-mechanical electronic structure computation. The EQ models, which can be applied to proteins, are also robust but do no give quantitatively correct dipoles. In addition to the obvious computational advantages of EQ charges, they are also valuable because they explain the charges in terms of constant and local parameters such as the atomic electronegativity, the atomic hardness and the bond hardness. We must understand why the EQ models fail to reproduce the dipole moment

correctly. While the EEM has some missing ingredients that may explain the observed errors, the SQE model (an extension of the EEM) should normally perform better. Although the SQE dipole moments correlate better with the MP2 data for the TB penta-alanine, the correlation becomes worse for the ZI form. This striking aberration in performance is most likely a fundamental limitation of the SQE model. The main differences between the TB and ZI form is that the latter bears two opposite charges in the end groups. Apparently the SQE model fails to reproduce the charge distribution when such locally charged groups are present. The relation between the large observed errors and the split charge formalism will be analyzed carefully in the remainder of the paper. Such analysis is essential for further advances in the field of charge equilibration models.

3.2. Specific benchmark for charge equilibration models

In order to gain more insight into the poor description of the penta-alanine dipole moment by the EEM and SQE model, we conduct a specific benchmark that is only applicable to the EQ models. Figure 7 shows the difference between the EQ and Hirshfeld-I charges. The color coding is the same as in Figure 5. Three trends are immediately visible: (i) the EEM charges deviate more from the Hirshfeld-I charges than the SQE charges, (ii) the data for the SQE model show that the average error on the charge for each atom is generally larger than the spread on the error, (iii) the errors in the ZI system are more pronounced at the terminal residues, while the TB system does not show this trend. The last observation is even more pronounced when plotting the error on the total charge of each fragment, as depicted in Figure 8.

The improved accuracy of the SQE model compared to the EEM in terms of atomic charges is in line with the benchmarks performed earlier.[50] It is somewhat disappointing that, despite the small root-mean-square error (RMSE) between the EQ and Hirshfeld-I charges (given in the first row of Table 3), the EQ dipole moments in Table 1 show large errors. The three last rows in Table 3 contain RMSE estimates on the EQ dipole moments, derived with eq 6 from the RMSE between the EQ and Hirshfeld-I charges. The order of magnitude of these estimates is in line with the corresponding numbers for EEM and SQE in Table 1. Hence, the deviation of the EQ from Hirshfeld-I charges must be further reduced in order to get quantitatively correct dipole moments.

The errors on the EQ charges are mainly geometry independent, which can be fixed by calibrating very specific first-order parameters for the penta-alanine system. The current calibration, which is applicable to a very broad set of molecules, does not contain sufficiently specific atom types when one is only interested in describing polypeptides.

The absolute values of the total charge on the terminal residues of the ZI penta-alanine are underestimated by the EEM and SQE model. Apparently, both models can not effectively describe systems that are locally charged. In case of the EEM, the metallic polarization causes charged functional groups to be neutralized by opposite charges in surrounding atoms. Although the SQE model solves the metallic problem with split charges that have a bond hardness, the split charges also introduce a new artifact. When a functional group is locally negatively charged, a series of split charges must be present to connect this group with a locally positively charged group. As will be discussed below, the SQE model can not handle such situations with parameters that are transferable from small to large molecules.

3.3. Statistical Breakdown of the errors on the SQE charges

Thus far we observed quite serious deficiencies in the SQE results, both in terms of the dipole moment of the penta-alanine system and the atomic partial charges. In this subsection, the relation between both types of errors is examined. Below, ad hoc (non-transferable) corrections are added to the SQE charges to see how such corrections could fix the errors on the dipole moments. The proposed ad hoc corrections are not meant as an actual improvement to the SQE model, but are rather used to provide useful feedback for an improved SQE model.

Two corrections are proposed. The major correction is atom-specific, but geometry-independent. These corrected charges will be referred to as SQE' charges. After making the first correction, an additional minor correction is useful on the TB system, which is related to a small error in the polarization of nearby non-bonded O-H pairs (including but not limited to hydrogen bonds). For statistical reasons outlined below, we can not test the second correction on the ZI form. The combination of the first and the second correction will be referred to as the SQE" charges.

The specific benchmarks on the charge equilibration models show that a large part of the error on the partial charge of an atom is due to an average error between the SQE charge and Hirshfeld-I charge, while the geometry dependence of the error is limited. The first correction consists of a geometry-independent set of atom-specific parameters that will be added to the SQE charges. For each atom, the correction is such that the average SQE' charge is equal to the average Hirshfeld-I charge. The fluctuations on the SQE' charges due to changes in geometry are still the same as in the original SQE charges.

Figure 9 shows the correlation between the MP2 and SQE' dipole moments, which is greatly improved compared to the original SQE dipoles. The statistical parameters of the comparison are given in Table 4. As mentioned above, this ad hoc correction can be implemented in the SQE framework by calibrating improved first-order parameters. Note that the major correction is much larger for the zwitterionic system. The errors related to charged functional groups in the SQE model are also fixed in this step.

After the first correction, there are still some deviations left between the MP2 and SQE' dipole moments. After some testing it was found that the residual error on the dipole moment correlates well with the sum of vectors connecting O and H atoms in nearby (<2.5Å) non-bonded O-H pairs. A linear fit reveals that an ad hoc transfer of +0.066e from the donor (H) to the acceptor (O) yields an optimal correction to the SQE' dipole moments, which corresponds to a reduction of the polarity of the O-H pairs. This correction has only noticeable effects in the TB system. We assume that the correction is also useful for the ZI form, but that it is too small compared to the statistical errors on the first correction.

The sum of the major and the minor correction leads to the SQE" charges. The correspondence between the MP2 and SQE" dipole moments, shown in Figure 9 and Table 4, is again improved compared to the SQE' dipole moments. This correction is geometry dependent, and is therefore related to a small error in the polarization of non-bonded O-H pairs in the SQE model. It is recommended to pay special attention to the calibration of the second order SQE parameters related to hydrogen bond donors and acceptors, e.g. in the form of very specific training data and hydrogen-bond specific benchmarks.

Although these corrections lead to a satisfactory correspondence between the SQE'' with the MP2 dipole moment, we stress again that they are only used to gain more insight in the errors in the SQE results. The corrections are not simply transferable to other systems.

3.4. SQE model with reference charges

It is not yet clear why charged functional groups can not be treated properly with the SQE model, and how this issue can be fixed. In this subsection we introduce a coarse-grained charge-equilibration model of a zwitterionic system to facilitate the analysis.

Consider a linear chain, as depicted in Figure 10a. The intermediate beads are neutral and the end points of the chain are oppositely charged functional groups, each bearing a formal integer charge. The formal charge is typically the result of an ionic dissociation reaction. In the case of the penta-alanine model (or any conjugate base or acid group in the sidechain of a residue in a protein in general), the charged moieties are formed by proton addition and abstraction, and are not caused by charge transfer from a positively to a nearby negatively charged group.

The scheme depicted in Figure 10a is an idealized representation of the chain molecule with integer formal charges to facilitate the description of this zwitterion in terms of split charges in figure 10b. Effective charges in a realistic system are not fundamentally different: due to the finite hardness of the charged functional groups, a small amount of charge may 'leak' to neighboring beads. Figure 10b shows the split-charge representation of this idealized zwitterion, in which each bond between the terminal beads is polarized. It is questionable whether such a split-charge configuration is reasonable in the SQE model. After all, the amount of energy required to polarize all bonds is proportional to the chain length, and the SQE model does not contain any other chain-length dependent terms to compensate for this. In order to get a better understanding of the energetics and behavior of zwitterions in the SQE model, the linear chain model is treated numerically below. We expect that the following intuitive properties should be reproduced by a proper charge equilibration model:

1. In the limit of long chains, the net charge on each terminal bead should converge to the integer formal charge. Some charge may leak to neighboring beads due to the finite hardness of the beads. Therefore one should always consider the sum of the charge on the terminal bead and its neighbors to recover the formal charge.

2. The parameters of a charge equilibration model should not depend on the chain length, i.e. they should be transferable between different chain lengths.

Below we show that EEM and SQE both fail to meet these expectations. A straightforward extension of the SQE model, SQE+Q^0, is proposed that meets both intuitive expectations.

In order to compute partial charges in the coarse-grained zwitterion model, we introduce an SQE model for the linear chain. Let N be the number of beads, η the hardness of the beads, and r_0 the distance between two neighboring beads. All beads have an electronegativity parameter of χ_{ref}, except for the first and last bead, which have electronegativity $\chi_{\mathrm{ref}} + \Delta\chi$ and $\chi_{\mathrm{ref}} - \Delta\chi$, respectively. All bonds between the beads have the same bond hardness, κ. The N charges in this model are linked with $N-1$ split charges through the following equations (see figure 10b):

$$\begin{aligned} q_0 &= p_{0,1} \\ q_i &= p_{i,i+1} - p_{i-1,i} \quad \forall i \in 1, \ldots, N-2 \\ q_{N-1} &= -p_{N-2,N-1} \end{aligned} \tag{7}$$

The energy of the coarse-grained charge equilibration model becomes:[40]

$$E = \sum_{i=0}^{N-1} \left(\chi_i q_i + \frac{1}{2} \eta_i q_i^2 \right) + \sum_{i=0}^{N-2} \frac{1}{2} \kappa_i p_{i,i+1}^2 + \sum_{i=0}^{N-1} \sum_{j=0}^{i-1} \frac{q_i q_j}{r_0 (i-j)} \tag{8}$$

When κ is set to zero, the model reduces to the EEM. The last term represents the Coulomb interaction between the beads. The equilibrium charge distribution is found by minimizing the energy, using the split charges, $p_{i,i+1}$, as

independent degrees of freedom. There is no need to constrain the total charge with a Lagrange multiplier because the total charge constraint is satisfied implicitly by eq 7.

The charge distribution is computed in a series of linear zwitterions for N going from 2 to 100. Table 5 contains the parameters for the EEM and SQE model that were used to compute the charge distributions. These parameters were chosen in an attempt to reproduce the behavior of a Zwitterion, while still keeping the numbers in a realistic order of magnitude. One may use other parameters, but they lead to essentially the same observations. Figure 11a, b and c depict the essential results. In Figure 11a, the dipole moment is plotted as function of the chain length. In Figure 11b, the sum of the charge in the second half of the chain is shown. Figure 11c depicts the charge distribution in the last 10 beads of a chain that consists of 100 beads.

The dipole moment of the chain computed with the EEM scheme scales sublinearly in Figure 11a, i.e. the slope in the loglog plot converges to 0.9 for large N, which is due to internal polarization effects. Figure 11a also shows that the dipole moment derived with the SQE model becomes constant for large N, which is completely wrong for the zwitterion model. Both trends can be easily understood when studying the charge distribution over the beads as function of the chain length in Figure 11b. The total charge in the second half of the chain decreases with larger N (slowly in the EEM case and rapidly in the SQE case) instead of converging to the formal charge. Figure 11c shows that the partial charge of the last bead is compensated by opposite charges in nearby neighboring beads in both the EEM and SQE model. In the case of EEM, this is due to excessive internal polarization. In the case of SQE, the electronegativity of the terminal beads is not sufficient to polarize all split charges over the entire chain. Instead, only the first and last few split charges can be polarized, which forces the chain molecule to be locally neutral.

The deficiencies of the SQE that are apparent in the linear chain model explain why the SQE parameters calibrated in earlier work on a set of 500 small organic molecules (including some molecular ions and zwitterions)[50] can not simply describe locally charged functional groups in much larger systems such as the zwitterionic penta-alanine. The formal charges in figure 10a can only be obtained with the SQE model by polarizing all intermediate split charges, as shown in figure 10b. Because the energy required to polarize all intermediate beads is proportional to $N-1$

$N-1$, a correct dipole moment for long zwitterions is only possible when the electronegativity parameters of the terminal beads are also proportional to $N-1$. The numerical example also shows that deviations from local neutrality are nearly impossible in the SQE model, while this is still possible with the EEM. This chain model explains the very poor reproduction of molecular dipoles by the SQE model for the ZI form of penta-alanine in the first benchmark.

A simple extension of the SQE model, hereafter called SQE+Q^0, can easily overcome the observed shortcomings of the SQE model. The current form of the SQE model assumes that atoms in a molecule can only obtain a partial charge by transferring this charge (over covalent bonds) from nearby atoms. This assumption is inherently present in the relation between the atomic charges and the split charges:

$$q_i = \sum_{j \in B_i} p_{j,i} \tag{9}$$

where q_i is the charge of atom i, $p_{j,i}$ is the charge transferred from atom j to atom i, and the sum runs over the atoms j that are bonded to atom i. (We refer the reader to ref. [50] for more details.) However, charge transfer over bonds is not the only physical route for an atom to become partially charged. Another major mechanism, which is missing in the SQE model, is that charged functional groups originate from an ionic bond dissociation reaction. We propose an additional term, a constant integer, in the relation between atomic charges and split-charges:

$$q_i = \sum_{j \in B_i} p_{j,i} + q_i^0 \tag{10}$$

The reference state of the atoms in the molecule, when all split charges are zero, consists not only of neutral atoms; we allow some atoms to bear an integer charge in the reference state. This reference atomic charge corresponds the conventional notion of a formal charge in chemistry. Reference charges are similar to the precharges, which are used to augment bond-charge-increments (BCI's) in fixed-charge models, and are also needed in that context to allow local deviations from neutrality.[66] Note that this extension is not meant to describe ionic reactions as a process. The extension can only handle the products of such reactions.

The SQE+Q^0 model can be applied to the coarse-grained zwitterion model by rewriting the relation between the charges and the split charges:

$$\begin{aligned} q_0 &= -Q^0 + p_{0,1} \\ q_i &= p_{i,i+1} - p_{i-1,i} \qquad \forall i \in 1, \ldots, N-2 \\ q_{N-1} &= Q^0 - p_{N-2,N-1} \end{aligned} \tag{11}$$

where $-Q^0$ and Q^0 are the reference charge for the first and the last atom, respectively. As shown in Figure 10c, one may set reference charges equal to the formal charges of the terminal groups in the chain molecule such that no split charges are required anymore to recover the charge distribution of the zwitterion. The SQE+Q^0 parameters are given in the last column of Table 5. In order to turn the chain into a zwitterion, a reference charge of +1 and -1 is assigned to the terminal beads. This configuration is similar to the zwitterionic form of the penta-alanine, where the terminal groups bear a formal charge caused by proton abstraction or addition in the end groups. The electronegativity parameters of the terminal beads are set to zero in the SQE+Q^0 approach because they are no longer needed to induce charge transfer between the end points.

The corresponding results are plotted in red in Figure 11a, b and c. Figure 11a and b show that the SQE+Q^0 reproduces the expected behavior: the dipole moment scales linearly with the chain length, and the total charge on the second half of the chain converges to unity. Figure 11c shows that the formal charge on the terminal bead may leak to some extent to neighboring beads.

4. Conclusions

Our benchmarks on two sets of penta-alanine conformers show that the Hirshfeld-I scheme is the most attractive method to derive atomic charges from quantum-mechanical computations on organic systems. These charges are chemically intuitive, reproduce electrostatic properties and are robust with respect to conformational changes. Several other charge schemes (such as Mulliken, RESP, EEM and SQE) show serious deficiencies, which have severe consequences on the development (polarizable) force fields for polypeptides.

To support these conclusions, we conducted two *general* benchmarks of several atomic charge schemes. Each scheme was applied to a set of 103 terminally blocked (TB) and 134 zwitterionic (ZI) stable isolated penta-alanine conformers. A solvent model is not included to simplify the analysis and because profound insight in the gas phase systems is prerequisite before conducting more elaborate studies in a more complex molecular environment. The investigated charge schemes can be divided into two categories: (i) quantum mechanical charges (Mulliken, Natural, RESP and Hirshfeld-I) based on MP2/Aug-cc-pVTZ electronic densities, and (ii) equilibration charges (Electronegativity Equalization Method [EEM] and Split Charge Equilibration [SQE]) for which no prior electronic structure computation is required. The first benchmark illustrates how well each charge model can reproduce the MP2 dipole moments for the sets of the TB and ZI penta-alanine conformers. The second benchmark tests the robustness of the charges in the same set of conformers. Although one should ultimately test how well each charge model is capable of reproducing experimental observations on proteins, taking into account solvent effects, our current investigation (in which high-level gas-phase computations are used as a reference) already reveals several insights and weaknesses. Only the Hirsfheld-I scheme gives satisfactory results in both benchmarks: it reproduces the MP2 dipole moments quantitatively, and the charges have a minimal sensitivity to conformational changes. We conclude that Hirsfheld-I charges are transferable between different conformers, and that they do not exhibit unrealistically large fluctuations like Mulliken or RESP charges. The latter amenity implies that Hirshfeld-I charges are more suitable for the development of biomolecular force fields compared to e.g. the RESP charges, which are currently used for the development of AMBER.

Although the Hirshfeld-I method performs well in the two general benchmarks, it is not simply applicable to large biosystems such as proteins with a surrounding solvent. Hirshfeld-I charges depend on an accurate electronic structure computation, which is computationally not attractive or even feasible for systems with many thousands of atoms. Because charge equilibration models are a computationally feasible alternative with current computer hardware, we investigated in detail the origins of the errors on the dipole moments derived from EEM and SQE charges. The poor performance of the EEM can be traced back to the incorrect scaling of the polarizability. Although the SQE model should fix this issue, both the mathematical form of the SQE model and the

calibrated parameters must be improved to reach quantitative accuracy. For a correct description of charged functional groups, an extension of the standard SQE model is proposed, SQE+Q^0, which introduces atomic reference charges for atoms bearing a formal charge. Furthermore, one has to calibrate first-order parameters that are specific for peptides, and one must carefully test the polarization of weakly covalent interactions such as hydrogen bonds. The proposed SQE+Q^0 model offers a reliable model to optimize the correlation between the quantummechanically obtained dipole moments and those predicted by the computationally more attractive charge equilibration models for use in extended biosystems.

Ackowledgements. This work is supported by the Fund for Scientific Research - Flanders (FWO), the Research Board of Ghent University (BOF) and BELSPO in the frame of IAP/6/27. Funding was also received from the European Research Council under the European Community's Seventh Framework Programme (FP7(2007-2013) ERC grant agreement number 240483). The authors would also like to thank the Ghent University for the computational resources (Stevin Supercomputer Infrastructure). PG en FDP wish to acknowledge the Research Foundation-Flanders (FWO) and the Free University of Brussels (VUB) for continous support to their research group.

Supporting information. The Cartesian coordinates of the 103 terminally blocked and 134 zwitterionic penta-alanine conformers. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Vizcarra, C. L.; Mayo, S. L. *Curr. Op. Chem. Biol.* **2005**, *9*, 622-626.
(2) Olsson, M.; Søndergaard, C.; Rostkowski, M.; Jensen, J. *J. Chem. Theory Comput.* **2011**, *7*, 525-537.
(3) McCammon, J. A. *P. Natl. Acad. Sci. USA* **2009**, *106*, 7683 -7684.
(4) Zhang, Z.; Witham, S.; Alexov, E. *Phys. Biol.* **2011**, *8*, 035001.
(5) Roos, G.; Loverix, S.; De Proft, F.; Wyns, L.; Geerlings, P. *J. Phys. Chem. A* **2003**, *107*, 6828-6836.
(6) Kim, Y. C.; Hummer, G. *J. Mol. Biol.* **2008**, *375*, 1416-1433.
(7) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. *J. Chem. Theory Comput.* **2008**, *4*, 819-834.
(8) Hess, B. *Phys. Rev. E* **2002**, *65*, 031910+.
(9) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. *J. Am. Chem. Soc.* **2010**, *132*, 1526-1528.

(10) Wang, J.; Morin, P.; Wang, W.; Kollman, P. A. *J. Am. Chem. Soc.* **2001**, *123*, 5221-5230.
(11) Hess, B. *Physical Review E* **2000**, *62*, 8438–8448.
(12) Dror, R. O.; Jensen, M. Ø.; Borhani, D. W.; Shaw, D. E. *J. Gen. Physiol.* **2010**, *135*, 555 -562.
(13) Kuhn, B.; Gerber, P.; Schulz-Gasch, T.; Stahl, M. *J. Med. Chem.* **2005**, *48*, 4040-4048.
(14) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
(15) Brooks, B. R.; Brooks III, C. L.; Mackerell Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545-1614.
(16) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657-1666.
(17) Neves-Petersen, M. T.; Petersen, S. B. In *Biotechnology Annual Review*; El-Gewely, M. R., Ed.; Elsevier: Amsterdam, The Netherlands, 2003; Vol. 9, pp. 315-395.
(18) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B* **1998**, *102*, 10983-10990.
(19) Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694–715.
(20) Kaminski, G.; Stern, H.; Berne, B.; Friesner, R.; Cao, Y.; Murphy, R.; Zhou, R.; Halgren, T. *J. Comput. Chem.* **2002**, *23*, 1515–1531.
(21) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2004**, *108*, 621–627.
(22) Patel, S.; Mackerell, A. D.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1504–1514.
(23) Patel, S.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1–16.
(24) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781–790.
(25) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.
(26) Liu, Y. P.; Kim, K.; Berne, B. J.; Friesner, R. A.; Rick, S. W. *J. Chem. Phys.* **1998**, *108*, 4739–4755.
(27) Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985**, *83*, 735–746.
(28) Roos, G.; Loverix, S.; Geerlings, P. *J. Phys. Chem. B* **2006**, *110*, 557-562.
(29) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.
(30) Lowdin, P.-O. *J. Chem. Phys.* **1950**, *18*, 365-375.
(31) Hirshfeld, F. L. *Theoret. Chem. Acta.* **1977**, *44*, 129–138.
(32) Bader, R. F. W. *Phys. Rev. B* **1994**, *49*, 13348.
(33) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbo-Dorca, R. *J. Chem. Phys.* **2007**, *126*, 144111.
(34) Francl, M. M.; Chirlian, L. E. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Boyd, D. B., Eds.; John Wiley & Sons, Inc., 2000; Vol. 14, pp. 1–31.
(35) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.
(36) Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. *J. Phys. Chem.* **1993**, *97*, 10269–10280.
(37) Campañá, C.; Mussard, B.; Woo, T. K. *J. Chem. Theory Comput.* **2009**, *5*, 2866-2878.
(38) Mortier, W.; Ghosh, S.; Shankar, S. *J. Am. Chem. Soc.* **1986**, *108*, 4315–4320.
(39) Van Genechten, K. A.; Mortier, W. J.; Geerlings, P. *J. Chem. Phys.* **1987**, *86*, 5063–5071.
(40) Nistor, R. A.; Polihronov, J. G.; Müser, M. H.; Mosey, N. J. *J. Chem. Phys.* **2006**, *125*, 094108.
(41) Warren, L. G.; Davis, J. E.; Patel, S. *J. Chem. Phys.* **2008**, *128*, 144110.
(42) Verstraelen, T.; Bultinck, P.; Van Speybroeck, V.; Ayers, P. W.; Van Neck, D.; Waroquier, M. *J. Chem. Theory Comput.* **2011**, *7*, 1750-1764.
(43) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
(44) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
(45) Ditchfield, R. *J. Chem. Phys.* **1971**, *54*, 724.
(46) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.

(47) 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.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09, Revision A.1*; Gaussian Inc.: Wallingford CT, 2009.

(48) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graph. Mod.* **2006**, *25*, 247-260.

(49) Verstraelen, T. HiPart: A Hirshfeld partitioning program. http://molmod.ugent.be/code (accessed Jul 18, 2011).

(50) Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. *J. Chem. Phys.* **2009**, *131*, 044127.

(51) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. *J. Phys. Chem. A* **2002**, *106*, 7887–7894.

(52) Chen, Z.; Li, Y.; Chen, E.; Hall, D. L.; Darke, P. L.; Culberson, C.; Shafer, J. A.; Kuo, L. C. *J. Biol. Chem.* **1994**, *269*, 26344–26348.

(53) Ghosh, D.; Erman, M.; Wawrzak, Z.; Duax, W. L.; Pangborn, W. *Structure* **1994**, *2*, 973-980.

(54) CP2K project homepage. http://cp2k.berlios.de/ (accessed Jul 22, 2011).

(55) MacKerell; Bashford, D.; Bellott; Dunbrack; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.

(56) Streitwieser, A.; Collins, J. B.; McKelvey, J. M.; Grier, D.; Sender, J.; Toczko, A. G. *Proc. Natl. Acad. Sci. USA* **1979**, *76*, 2499 -2502.

(57) Van Damme, S.; Bultinck, P.; Fias, S. *J. Chem. Theory Comput.* **2009**, *5*, 334–340.

(58) Nistor, R. A.; Müser, M. H. *Phys. Rev. B* **2009**, *79*, 104303.

(59) Simonson, T.; Brooks, C. L. *J. Am. Chem. Soc.* **1996**, *118*, 8452-8458.

(60) Bader, R. F. W. *J. Chem. Phys.* **1972**, *56*, 3320.

(61) Hu, H.; Lu, Z.; Yang, W. *J. Chem. Theory Comput.* **2007**, *3*, 1004–1013.

(62) Stouch, T. R.; Williams, D. E. *J. Comput. Chem.* **1992**, *13*, 622-632.

(63) Politzer, P. *J. Chem. Phys.* **1971**, *55*, 5135.

(64) Ponder, J. W.; Case, D. A. In *Protein Simulations*; Daggett, V., Ed.; Elsevier, 2003; Vol. Volume 66, pp. 27-85.

(65) Reynolds, C. A.; Essex, J. W.; Richards, W. G. *J. Am. Chem. Soc.* **1992**, *114*, 9075-9079.

(66) Bush, B. L.; Bayly, C. I.; Halgren, T. A. *J. Comput. Chem.* **1999**, *20*, 1495-1516.

(67) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33-38.

(68) Persistence of Vision Raytracer (Version 3.6) http://www.povray.org/ (accessed Jul 19, 2011).

Tables

Table 1. Error measures for the reproduction of the MP2 dipole moments of the penta-alanine samples with atomic charges obtained with several charge models.

	Terminally blocked (TB)		**Zwitterion (ZI)**	
	$\sigma_{p\#}$ [D]	Rel. Err. [%]	$\sigma_{p\#}$ [D]	Rel. Err. [%]
	Mulliken		*Mulliken*	
X	7.98	259	12.24	298
Y	5.02	142	6.65	113
Z	8.00	136	9.90	57
	Natural		*Natural*	
X	1.08	35	1.07	26
Y	1.11	32	1.02	17
Z	1.53	26	1.04	6
	RESP		*RESP*	
X	0.03	1	0.04	1
Y	0.03	1	0.05	1
Z	0.04	1	0.09	1
	Hirshfeld-I		*Hirshfeld-I*	
X	0.46	15	0.60	15
Y	0.28	8	0.26	4
Z	0.36	6	0.38	2
	EEM		*EEM*	
X	1.59	52	3.11	76
Y	2.21	63	4.33	73
Z	4.39	75	14.66	85
	SQE		*SQE*	
X	0.86	28	3.96	96
Y	1.15	33	6.18	105
Z	2.58	44	19.40	113

Table 2. Results from the least squares analysis of the errors between the MP2 dipole moments of the penta-alanine samples and the dipole moments obtained with several atomic charge models. (A=slope, B=intercept)

	Terminally blocked (TB)						**Zwitterion (ZI)**					
	All data			**80 % of the data**			**All data**			**80 % of the data**		
	A [1]	**B [D]**	**R^2 [%]**	**A [1]**	**B [D]**	**R^2 [%]**	**A [1]**	**B [D]**	**R^2 [%]**	**A [1]**	**B [D]**	**R^2 [%]**
	Mulliken			*Mulliken*			*Mulliken*			*Mulliken*		
x	1.54	2.10	29.80	1.43	2.08	27.45	1.69	3.75	22.18	1.96	3.94	29.27
y	0.97	0.29	31.51	1.04	-0.22	37.62	0.86	-3.34	60.97	0.87	-2.74	59.37
z	1.40	-1.97	58.36	1.41	-2.02	57.73	0.84	-0.45	70.32	0.75	-1.93	65.67
	Natural			*Natural*			*Natural*			*Natural*		
x	1.32	-0.18	99.15	1.32	-0.16	99.11	1.20	0.34	97.48	1.19	0.34	97.45
y	1.27	-0.12	98.53	1.26	-0.07	98.22	1.17	0.48	99.25	1.18	0.49	99.40
z	1.20	-0.71	99.44	1.20	-0.74	99.48	1.07	0.88	99.85	1.07	0.82	99.84
	RESP			*RESP*			*RESP*			*RESP*		
x	1.00	0.00	99.99	1.00	-0.01	99.99	1.00	0.00	99.99	1.00	0.00	99.99
y	1.00	0.01	99.99	1.00	0.01	99.99	1.00	0.01	99.99	1.00	0.01	99.99
z	1.00	0.01	100.00	1.00	0.00	100.00	1.00	0.01	100.00	1.00	0.01	100.00
	Hirsfheld-I			*Hirsfheld-I*			*Hirsfheld-I*			*Hirsfheld-I*		
x	0.99	0.01	97.80	1.00	0.02	97.66	0.99	-0.16	98.05	0.97	-0.17	98.08
y	1.02	0.05	99.42	1.02	0.06	99.43	1.02	-0.03	99.84	1.02	-0.03	99.84
z	1.02	0.08	99.68	1.02	0.06	99.67	1.01	-0.10	99.96	1.01	-0.06	99.96
	EEM			*EEM*			*EEM*			*EEM*		
x	0.55	0.32	86.89	0.54	0.31	86.66	0.44	0.86	58.71	0.46	0.88	61.06
y	0.40	0.20	86.75	0.40	0.18	85.81	0.33	0.48	82.21	0.33	0.51	83.68
z	0.27	0.20	79.91	0.27	0.26	80.77	0.19	0.76	70.90	0.17	0.58	68.89
	SQE			*SQE*			*SQE*			*SQE*		
x	0.74	0.00	98.24	0.73	0.00	98.03	0.47	1.82	48.61	0.51	1.86	51.68
y	0.68	-0.08	98.49	0.69	-0.08	98.41	0.33	2.72	61.53	0.37	2.91	66.54
z	0.64	-1.57	97.80	0.64	-1.56	98.02	0.13	5.19	70.49	0.13	5.09	71.88

Table 3. Estimate of the RMSE between the EEM/SQE and MP2 dipole moments based on RMSE between EEM/SQE and Hirshfeld-I charges. This estimate assumes that the errors on the charges are uncorrelated, except for the total charge constraint.

	Terminally blocked		**Zwitterion**	
	EEM	**SQE**	**EEM**	**SQE**
σ_q [e]	0.07	0.04	0.08	0.08
σ_{p1} [D]	3.69	2.34	3.50	3.80
σ_{p2} [D]	5.25	3.33	5.63	6.12
σ_{p3} [D]	10.68	6.79	8.43	9.16

Table 4. Error measures for the reproduction of the MP2 dipole moments of the penta-alanine samples with the corrected SQE charges.

	Terminally blocked (TB)		**Zwitterion (ZI)**	
	$\sigma_{p\#}$ [D]	Rel. Err. [%]	$\sigma_{p\#}$ [D]	Rel. Err. [%]
	SQE '		*SQE* '	
X	0.70	23	1.41	34
Y	0.80	23	1.45	25
Z	1.63	28	1.88	11
	SQE ''			
X	0.75	24		
Y	0.52	15		
Z	0.66	11		

Table 5. Parameters for the computation of the charge distribution in the coarse-grained zwitterion model.

	EEM	**SQE**	**SQE+Q⁰**
$\Delta\chi$ [eV]	5	5	0
η [eV]	10	10	10
κ [eV]	0	5	5
q_0 [e]	-	-	1

Figures

Figure 1. The two penta-alanine forms studied in this paper: (a) terminally blocked and (b) zwitterionic. The numbering of the atoms and fragments, as they are used in the remainder of the paper, is indicated in both cases.

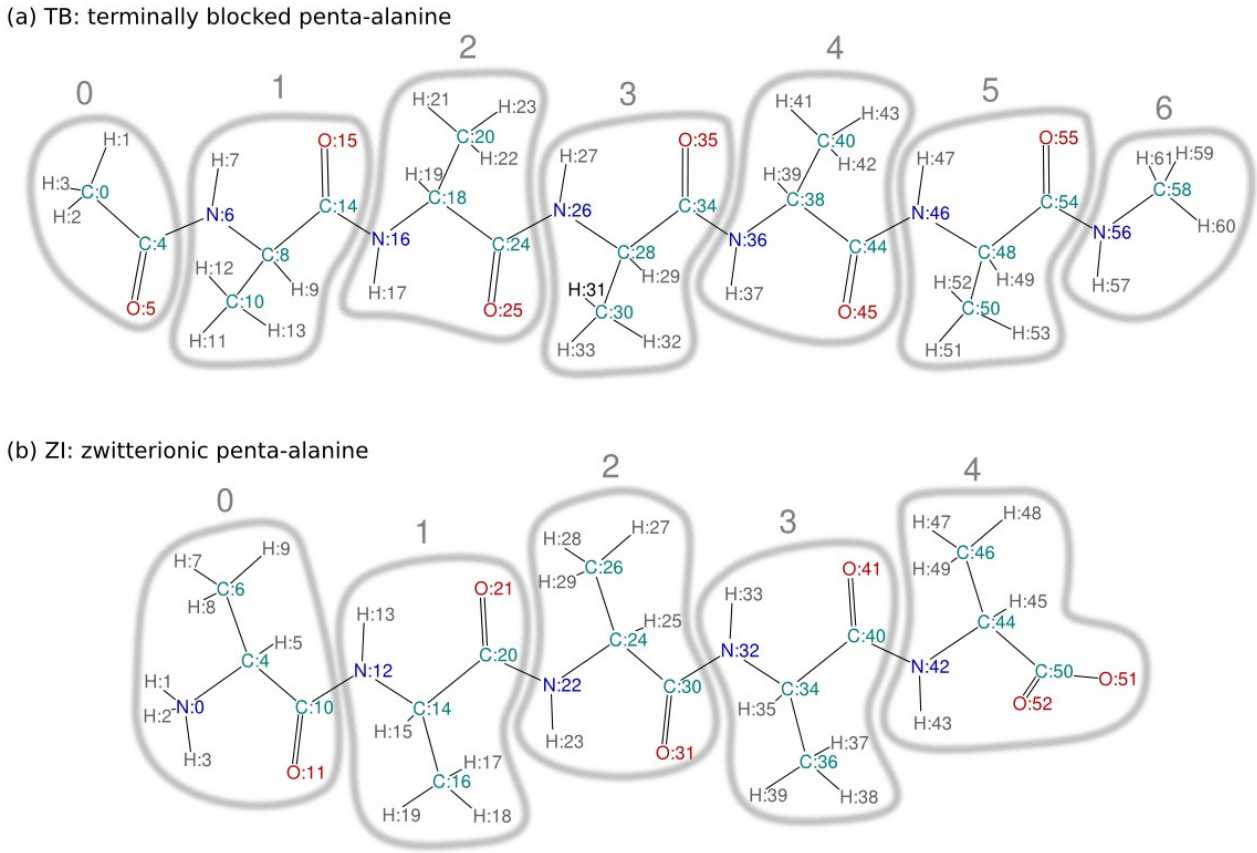


Figure 2. Histogram of the O-H distances in all generated penta-alanine conformers. All pairs with a distance below 2.5 Å are referred to as nearby non-bonded O-H pairs in this paper.

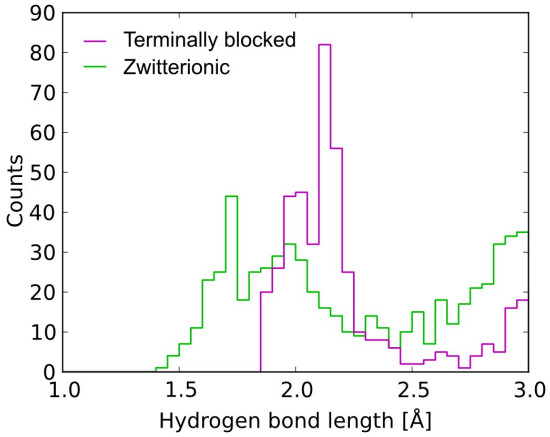


Figure 3. Scatter plots showing the correlation between the MP2 dipole moments and the dipole moments computed with the charge-models. The colors red, green and blue are used for the X, Y and Z components of the dipole moment, respectively. Linear fits are included through each dataset. The first bisector is plotted in gray. All dipole moments are given in units of Debye [D].

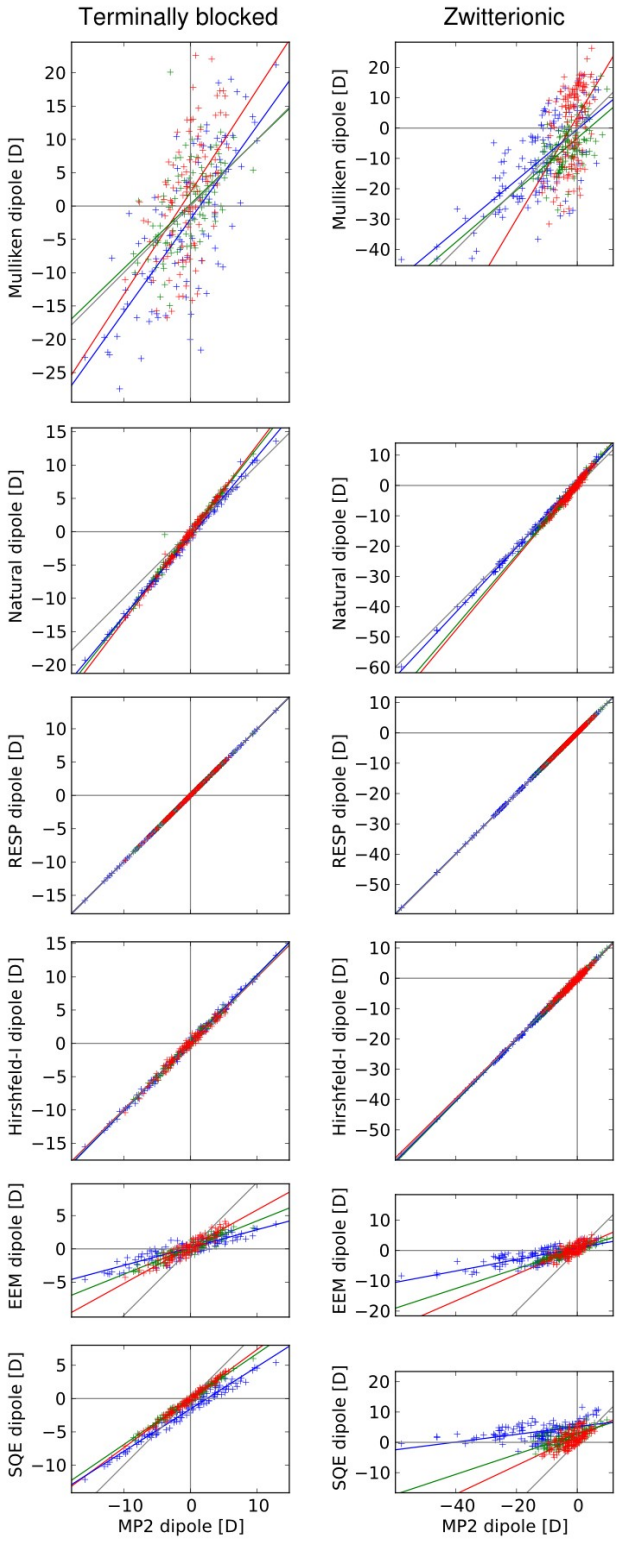


Figure 4. Visualization of the electrostatic potential (ESP) for two proteins: 1HDC (only one of the four symmetry-related chains is shown) and 1HSG. The protonation of acid and base groups is done at pH=7. Geometries are optimized with NAMD, using the CHARMM force field. The ESP is derived from EEM and SQE charges for both systems. The cartoon representation is colored according to the ESP convoluted with a Gaussian function with $\sigma = 5\text{Å}$. Protein images are rendered with VMD[67] and POVray.[68]

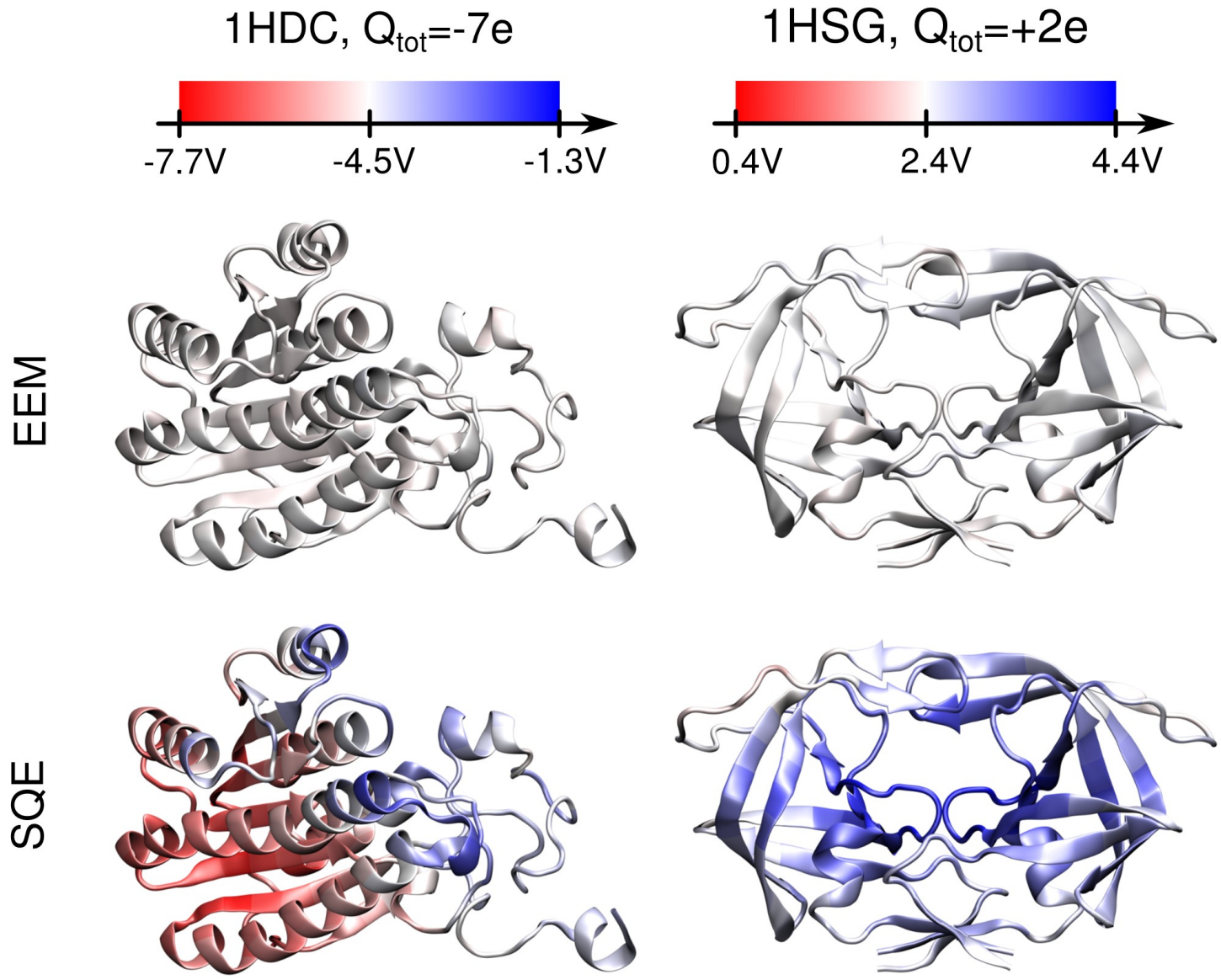


Figure 5. Distribution of the atomic charges obtained with different charge models for all conformers of the terminally blocked and zwitterionic penta-alanine. The atom indices and colors correspond to the labels in Figure 1: gray = H, cyan = C, dark blue = N and red = O. Dashed lines separate atoms belonging to different fragments. Each charge inside one conformer is plotted as one colored dash. A thick black dash corresponds to the average of an atom charge over all geometries.

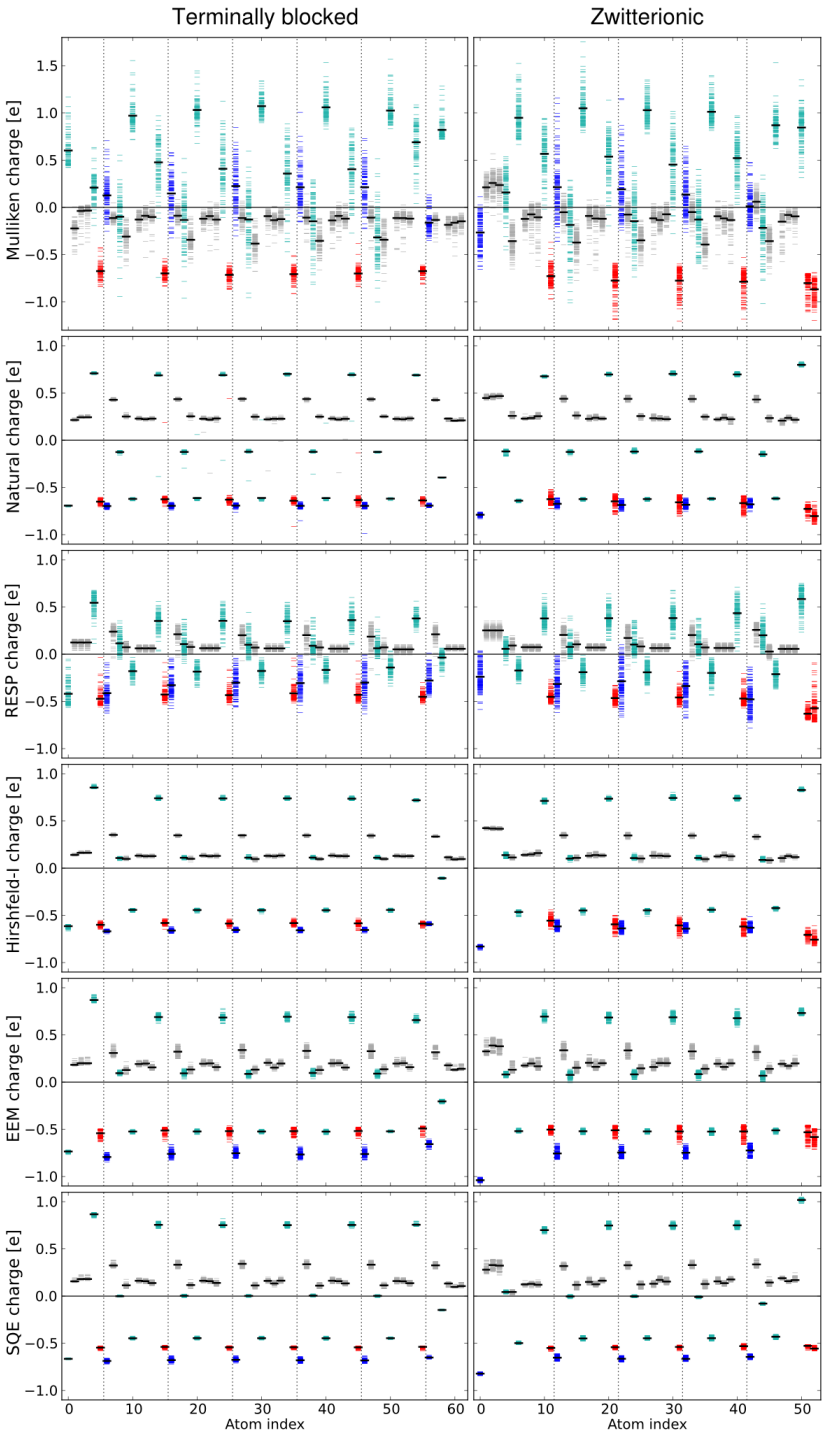


Figure 6. Variation of calculated charge as a function of conformational change. The top panel contains the charge on the α carbon in alanine dipeptide computed with different schemes during a relaxed scan along the ψ angle. The dashed vertical lines correspond to large geometric changes during the scan, as can be seen in the RMSD between subsequent structures plotted in the lower panel.

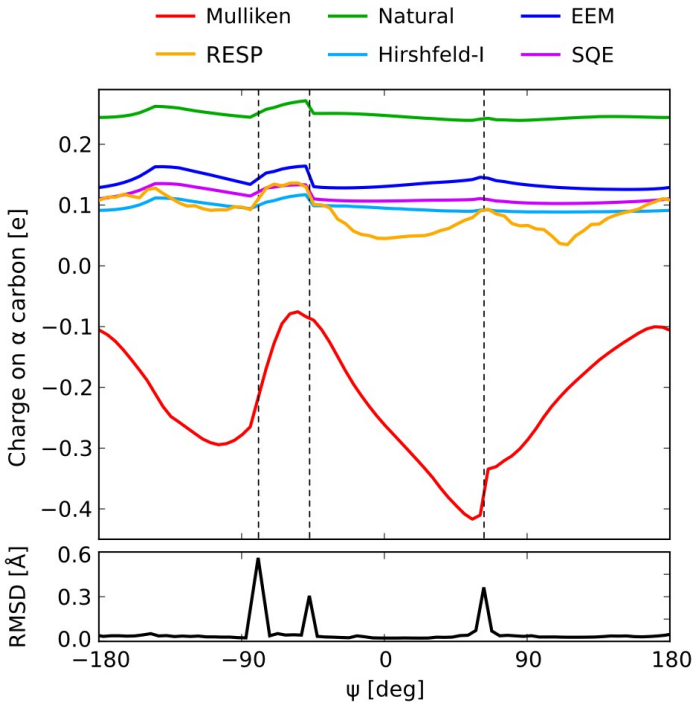


Figure 7. Distribution of the differences between the EEM or SQE and the Hirshfeld-I charges. The same conventions are used as in Figure 1: gray = H, cyan = C, dark blue = N and red = O.

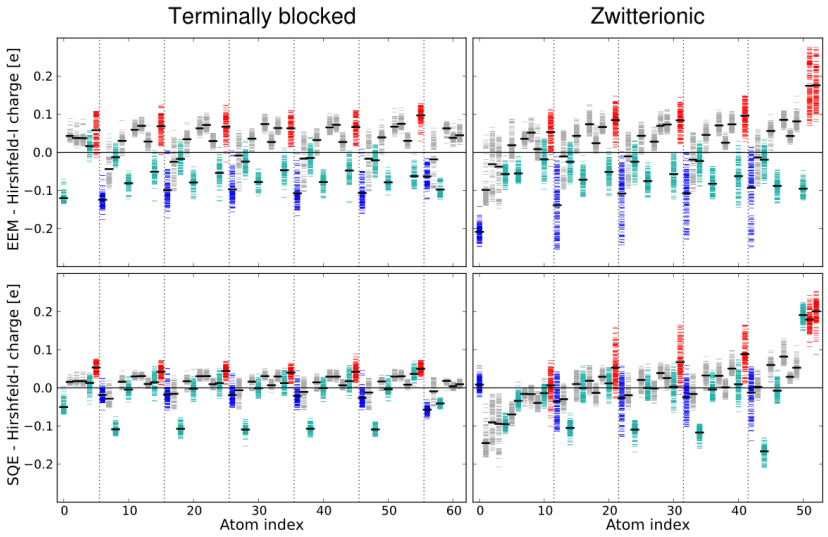


Figure 8. Distribution of the differences between the EEM or SQE and the Hirshfeld-I fragment charges.

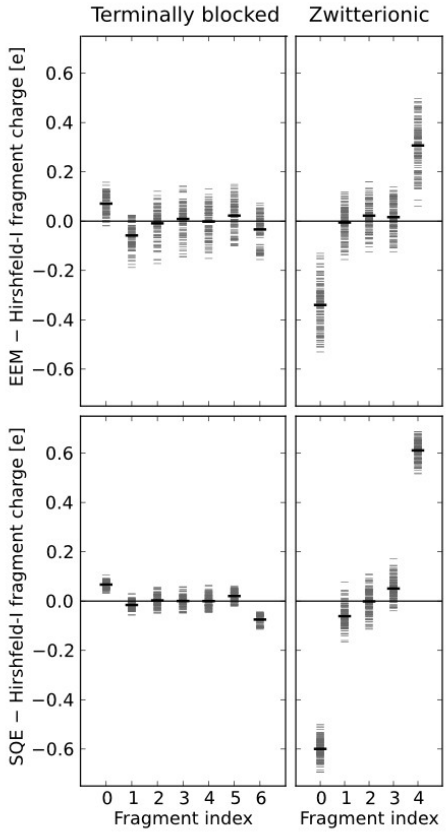


Figure 9. Scatter plots showing the correlation between the MP2 dipole moment and the dipole moment computed with the corrected SQE charges. The colors red, green and blue are used for the X, Y and Z components of the dipole moment, respectively. Linear fits are included through each dataset. The first bisector is plotted in gray.

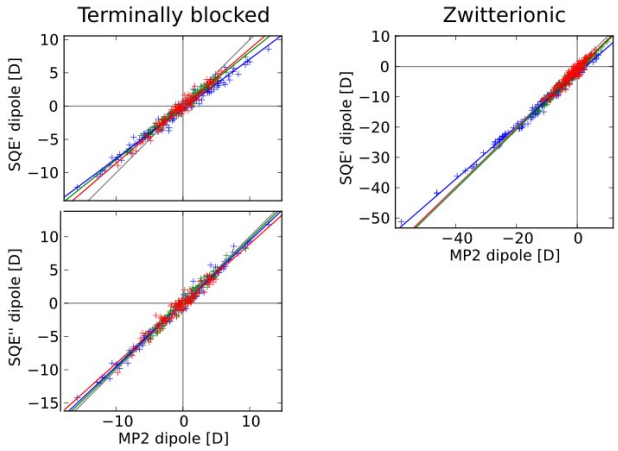


Figure 10. A coarse-grained model of a linear zwitterion. The formal charge distribution in the zwitterion is written with two models: SQE and SQE+Q0. (See text.)

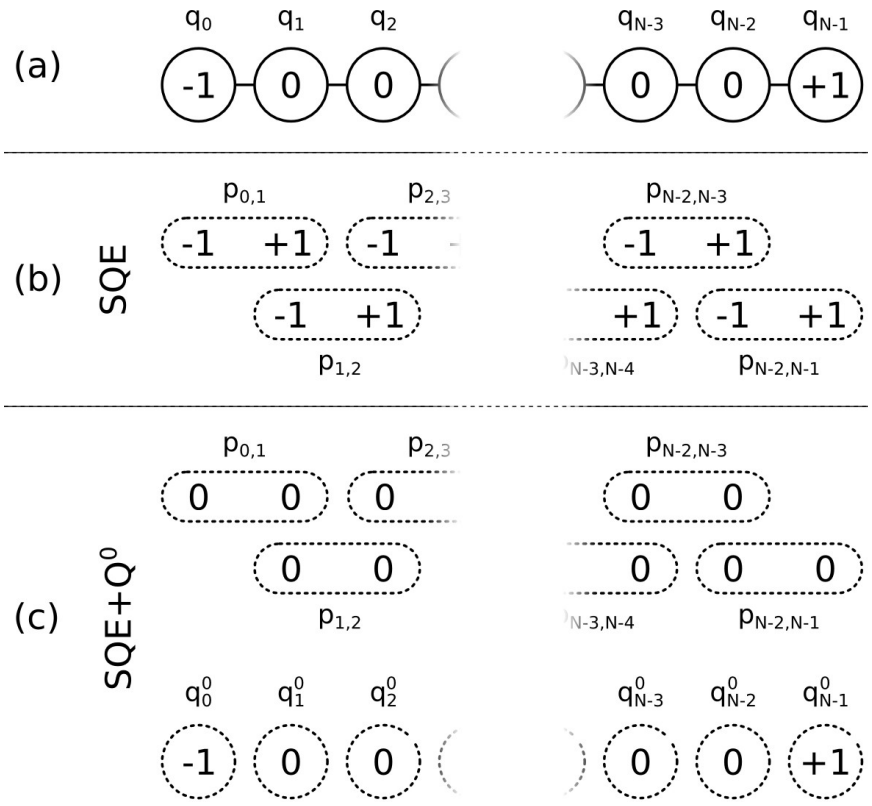
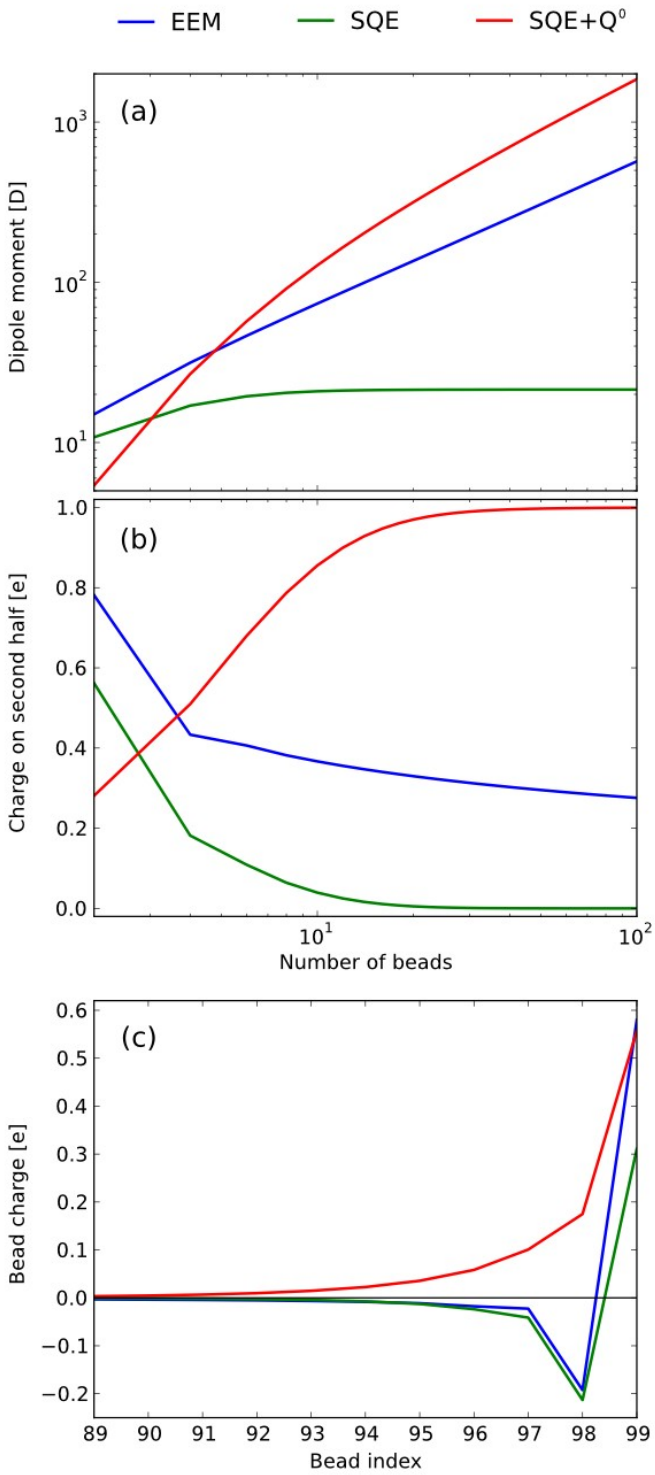


Figure 11. Effective charge distribution in the coarse-grained model of the linear zwitterion, computed with three charge equilibration models: EEM, SQE and SQE+Q^0. (See text.) (a) The dipole moment of the chain as function of the chain length. (b) The total charge in the second half of the chain as function of the chain length. (c) The charge on the last 10 beads in the chain that consists of 100 beads.



For Table of Contents Use Only (9cm by 3.5cm)