Title

Protein C-GeM: A Coarse-Grained Electron Model for Fast and Accurate Protein Electrostatics Prediction

Journal

Journal of Chemical Information and Modeling, 61(9)

ISSN

1549-9596

Authors

Guan, Xingyi
Leven, Itai
Heidar-Zadeh, Farnaz
et al.

Publication Date

2021-09-27

DOI

10.1021/acs.jcim.1c00388

Protein C-GeM: A coarse-grained electron model for fast and accurate protein electrostatics prediction

Xingyi Guan,[†,‡] Itai Leven,[†,¶] Farnaz Heidar-Zadeh,[†,§] and Teresa Head-Gordon[*,†,‡,‖]

†*Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California, Berkeley 94720*

‡*Chemical Sciences Division, Lawrence Berkeley National Laboratory*

¶*Chemical Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720*

§*Department of Chemistry, Queen's University, Kingston, Ontario K7L 3N6, Canada*

‖*Departments of Bioengineering and Chemical and Biomolecular Engineering, University of California, Berkeley, California 94720*

E-mail: thg@berkeley.edu

Abstract

The electrostatic potential (ESP) is a powerful property for understanding and predicting electrostatic charge distributions that drive interactions between molecules. In this study, we compare various charge partitioning schemes including fitted charges, density-based QM partitioning, charge equilibration methods, and our recently introduced coarse-grained electron model, C-GeM, to describe the ESP for protein systems. When benchmarked against high quality Density Functional Theory calculations of

the ESP for tripeptides and the crambin protein, we find that the C-GeM model is of comparable accuracy to *ab initio* charge partitioning methods, but with orders of magnitude improvement in computational efficiency since it does not require either the electron density nor the electrostatic potential as input.

INTRODUCTION

The electrostatic potential (ESP) is fundamental for understanding and predicting biomolecular recognition between molecules[1,2] For proteins in particular, the ESP is often crucial for predicting contact sites of protein-protein association,[3] and the electrostatic complementarity between protein and small molecule ligands or peptide therapeutics is considered critically important to obtain optimal affinity and selectivity in structure-based drug discovery.[4,5]

An ESP is generated by evaluating the work to move a unit charge probe from infinity to an area of interest on or near the protein surface. Numerically this is achieved by defining a grid, either on the molecular surface of the protein or by drawing equipotential contours in the region around the protein.[6] At each surface point $\mathbf{r}$, the ESP energy of the probe is calculated and the molecular surface is then displayed to indicate regions of negative or positive electrostatic potential of the protein molecule. An accurate way of obtaining the molecular ESP is through *ab initio* calculations, for which the ESP is defined as

$$V(\mathbf{r}) = \sum_A \frac{Z_A}{|\mathbf{R}_A - \mathbf{r}|} - \int \frac{\rho(\mathbf{r'})d\mathbf{r'}}{|\mathbf{r'} - \mathbf{r}|} \tag{1}$$

where Z_A and $\mathbf{R}_A$ are the charge and position of nucleus A, and $\rho(\mathbf{r'})$ is the electronic density at position $\mathbf{r'}$. However, the computational cost of a full quantum mechanical (QM) ESP increases rapidly with the number of atoms, and becomes prohibitive for systems such as large macromolecules.

Instead a large macromolecule can be partitioned in such a way that the electrostatic potential can be reproduced by assigning partial charges to every atom in a molecule, $\{q_A\}_{A=1}^{N_{\text{atoms}}}$,

i.e.,

$$V(\mathbf{r}) \sim \sum_{A=1}^{N_{\text{atoms}}} \frac{q_A}{|\mathbf{r} - \mathbf{R}_A|} \tag{2}$$

Atomic charges derived from fitting a classical Coulomb model to reproduce the *ab initio* molecular electrostatic potentials (so called ESP-charges) are frequently used in simulations of macromolecules,,[7–10] and they are the main electrostatic description used for all major fixed charge force fields such as AMBER[11,12] and CHARMM,[13] and utilized in large molecule ESP solvers using the Poisson-Boltzmann equation such as APBS.[14] One widely used ESP charge is the AM1-BCC model,[15] which captures the underlying features of the electron distribution including formal charge and delocalization using the semi-empirical AM1 method, and applies bond charge corrections (BCCs) that are fitted to *ab initio* ESP. While more cost-effective than full QM, the ESP-charges are numerically ill-conditioned such as being overly sensitive to conformational changes and restricted to applications where the electron density changes are relatively small.[16] While ESP-fitted charge models such as CHELPG[17] can be more robust, and can accurately reproduce the molecular ESP, they are not competitors to the prediction application because they require the ESP as its input.

Alternatively, QM-based partitioning methods divide a molecule into atomic subsystems by partitioning either the molecular wave-function in Hilbert space (i.e., orbital-based methods) or molecular electron density in real space (i.e., density-based methods). The first and most prevalent orbital-based partitioning method is the Mulliken[18] scheme which divides each molecular orbital into its atomic pieces. The original Mulliken partitioning suffered from excessive basis-set sensitivity, but subsequent refinement alleviated this shortcoming by defining atomic pieces in more sophisticated ways.[19–21] Unfortunately, the orbital-based charges are generally inferior for reproducing the electrostatic potential, as compared to density-based partitionings.[22]

The density-based QM partitioning exhaustively divide the molecular electron density

distribution, $\rho(\mathbf{r})$, between its constituent atoms according to

$$\rho_A(\mathbf{r}) = \sum_{A}^{N_{\text{atoms}}} w_A(\mathbf{r})\rho(\mathbf{r}) \tag{3}$$

$$\sum_{A}^{N_{\text{atoms}}} w_A(\mathbf{r}) = 1 \quad \text{and} \quad w_A(\mathbf{r}) \geq 0$$

where the electron density of atom A at point $\mathbf{r}$ in space, $\rho_A(\mathbf{r})$, is dictated by its share $w_A(\mathbf{r})$ at that point. Subsequently, the atomic charge of atom A is computed by,

$$q_A = Z_A - \int \rho_A(\mathbf{r}) d\mathbf{r} \tag{4}$$

The quality of these charges in reproducing the electrostatic potential heavily depends on the definition of atomic weights, $w_A(\mathbf{r})$. The atomic weights used in density-based methods are either binary as in Bader's Quantum Theory of Atoms in Molecules (QTAIM)[23] or fuzzy as developed in the Hirshfeld partitioning schemes and its variants.[16,22,24–29] Among these, the latter results in nearly-spherical atomic regions, so they have rapidly converging atomic multipole expansions and give a good approximation of $V(\mathbf{r})$ based on Eq. (2).

The Hirshfeld-family of methods use a set of proatom atomic densities $\{\rho_A^0(\mathbf{r})\}_{A=1}^{N\text{atoms}}$ to assign the atomic weights through,[24,30–32]

$$w_A(\mathbf{r}) = \frac{\rho_A^0(\mathbf{r})}{\sum_{B=1}^{N_{\text{atoms}}} \rho_B^0(\mathbf{r})} \tag{5}$$

The original Hirshfeld[24] method uses neutral proatom densities as the reference; this choice is arbitrary and results in very small atomic charges. To fix these shortcomings, various Hirshfeld-inspired methods have been developed to select optimal proatom densities.[16,22,25,32] The first, and most prevalent, method is Iterative Hirshfeld (HI),[25] which refines the proatoms self-consistently so that they have the same charges as the atoms. Two more recent and promising methods are the Minimal Basis Iterative Stockholder (MBIS)[22] and Additive

Variational Hirshfeld (AVH)[16,32] which variationally optimize the proatom densities so that they best reproduce the molecular density. When the atomic partial charges are determined from the population of these atomic subsystems, the more accurate reproduction of $V(\mathbf{r})$ is a measure of the partitioning scheme's quality and utility,[16,22] and thus we consider them here. Of course there is a very large and extensive number of *ab initio* charge partitioning methods that we have not considered here, and the interested reader can refer to a recent review by Martin and co-workers[33] to learn more about these approaches. While QM based partitioning approaches have the advantage of being physically grounded and generally applicable to a wide range of systems of interest, and have proven effective for modelling intermolecular interactions,[34] they still suffer from the underlying expense of QM calculations and thus are not extensible to large systems such as proteins.

The electronegatitivy equalization methods (EEM) is an alternative approach that straddles the boundary of empirical fitting but formulated within the QM foundations of atomic hardness and electronegativity.[35,36] It has been used as the electrostatic model for reactive force fields[37] and has been adapted for fast electrostatic screening applications for large molecular databases as well as protein electrostatics applications due to its relative efficiency.[38–40] However, although elegant, EEM has some significant shortcomings including unphysical long-range charge transfer, non-integer molecular charge at large molecular separations, lack of out-of-plane polarization, poor parameterization, and lack of transferability that makes EEM methods less accurate than desired but which are analyzed here for completeness.[41–43]

Hence, an accurate but fast method for protein ESP prediction is still highly desirable. In this study, we evaluate the coarse-grained electron force field model, C-GeM for which atoms are represented by a positive core and an electron shell described by Gaussian charge distributions.[44] Integration of the Coulombic interactions of the Gaussian densities yields an analytical form for the electrostatic energy between arbitrary core-core, core-shell, and shell-shell interactions. By minimizing the electronic shell positions in the field of atomic

core positions, the model can provide accurate electrostatic properties of molecules and their interactions. A schematic of this process is shown in Figure 1.

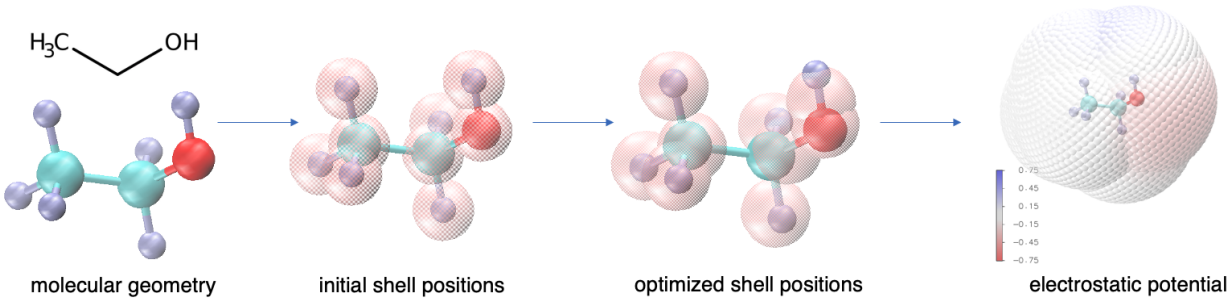


Figure 1: Schematic illustration of how C-GeM generates the electrostatic potential from given molecular geometry.

Previous models which share similarities to C-GeM include the core-shell model developed to account for polarization in ionic crystals,[45] the PQEq method which utilizes a Gaussian Drude oscillator model together with charge equilibration,[46] and the ACP method which partitions the electron density according to the core and valance shell electrons.[47] C-GeM differs from these previous models through its unique ability to predict permanent electrostatics, polarization, and charge transfer without having to perform computationally expensive ab-initio calculations. While the C-GeM model has been previously parameterized for the atomic elements carbon, hydrogen, oxygen and chloride,[44] in this work we have expanded the C-GeM parameterization for the nitrogen and sulfur atomic elements for a complete protein level chemistry. When optimized with tripeptide and small molecule training data, C-GeM is found to perform better than ESP-fitted charges, EEM, and Hirshfeld charges in reproducing the ESP of the protein test set that is comprised of tripeptides of different sequences and the crambin protein. To improve accuracy of the C-GeM model further we also introduce atom typing, i.e. optimization of different parameters for aliphatic and polar carbon and hydrogen atoms and for primary, secondary and tertiary amines for nitrogen. This atom typing approach thus makes the C-GeM model as accurate as HI charges and competitive with MBIS and AVH density partitioning methods when evaluated on the protein test set. Altogether the C-GeM model offers a new way to do high-throughput electrostatic screening

with *ab initio* accuracy with orders of magnitude less computational expense since it does not require the electron density or ESP but instead predicts these quantities.

Theory

The C-GeM model divides atoms into positive cores and negative shells, both of which are represented as Gaussian distributed charges. The properties of a core depend on its atom type i, while all of the electrons (shells) are treated equivalently. The charge density of a core of atom type i ($\rho_{i,c}$) and that of a generic shell (ρ_s) is given by the following functional form

$$\rho_{i,c}(\mathbf{r}) = q_{i,c}(\frac{\alpha_{i,c}}{\pi})^{3/2} e^{-\alpha_{i,c}(|\mathbf{r}-\mathbf{r}_{i,c}|^2))} \tag{6}$$

$$\rho_s(\mathbf{r}) = q_s(\frac{\alpha_s}{\pi})^{3/2} e^{-\alpha_s(|\mathbf{r}-\mathbf{r}_s|^2))} \tag{7}$$

where ($\mathbf{r}$) is an arbitrary position in space and $\mathbf{r}_{i,c}$ and $\mathbf{r}_s$ are position vectors for the core and shell centers, respectively. The shell charge (q_s) is always set to -1, while the core charge($q_{i,c}$) is usually set to +1 but can vary based on the chemical conditions of charge as we illustrate below. The width of a Gaussian charge is controlled by $\alpha_{i,c}$ for cores and α_s for shells:

$$\alpha_{i,c} = \frac{\lambda}{2R_{i,c}^2} \quad \alpha_s = \frac{\lambda}{2R_s^2} \tag{8}$$

where λ is a global fitting parameter, $R_{i,c}$ is the atomic covalent radius[9] of atom type i that is further fine tuned to reflect the atomic radii in actual molecules, and R_s is the effective radius of the shells.

The Coulombic interaction between two elements (core-core, core-shell and shell-shell) can be expressed as the integration over two Gaussian densities, which has the following

analytical form:

$$\begin{aligned} E_{ij}^{elec}(r_{ij}) &= \int\int \frac{\rho_i(\mathbf{r}_i)\rho_j(\mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|} d\mathbf{r}_i d\mathbf{r}_j \\ &= \frac{q_i q_j}{r_{ij}} erf(\sqrt{\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}} r_{ij}) \end{aligned} \tag{9}$$

where r_{ij} is the distance between the two elements. In the limit of $r_{ij} \to 0$, the pairwise Coulombic interaction can be rewritten as

$$\lim_{r_{ij}\to 0} E_{ij}^{elec}(r_{ij}) = \frac{2q_i q_j}{\sqrt{\pi}} (\sqrt{\frac{\alpha_i \alpha_j}{\alpha_i + \alpha_j}}) \tag{10}$$

In addition to electrostatics, a Gaussian energy term is used that reflects the strength of core-shell or shell-shell interaction, taking into account the electronegativity of specific atom types:

$$E_{ij}^{gauss}(r_{ij}) = \beta_i e^{-\gamma_i r_{ij}^2} + P(r_{ij}) \tag{11}$$

where β_i is a parameter accounting for the magnitude of the interaction energy, $P(r_{ij})$ is a penalty term for shell-shell distances that are too close, and γ is a parameter that controls the radial range of the interaction, which is defined as

$$\gamma_{i,c} = \frac{\omega_c}{2R_{i,c}} \tag{12}$$

for core-shell Gaussian interactions, controlled by a global parameter ω_c and atomic parameter $R_{i,c}$ for atom type i. The radial range for shell-shell interaction is controlled by global parameter γ_s.

With the theoretical idea that the C-GeM energy between a core of atom type i and its shell j should match the ionization potential of that atom type (χ_i), we demand that

$$\chi_i = E_{ij}^{elec}(r_{ij} = 0) + E_{ij}^{gauss}(r_{ij} = 0) \tag{13}$$

where χ_i is the ionization potential of atom type i. In the case of a shell-shell interaction, we use a global fitting parameter χ_{shell} to represent the effective shell-shell interaction energy that leads to following definition for the magnitude of Gaussian interaction β_i:

$$\begin{aligned}\beta_i &= \lim_{r_{ij}\to 0} \frac{\chi_i - E_{ij}^{elec}}{e^{-\gamma_i r_{ij}^2}} \\ &= \chi_i - \frac{2q_i q_j}{\sqrt{\pi}}\left(\sqrt{\frac{\alpha_i\alpha_j}{\alpha_i+\alpha_j}}\right)\end{aligned} \tag{14}$$

To avoid shell configurations that optimize to the exact same position and become inseparable, we introduced a penalty term for shell-shell interaction at very short range. This term effectively help shells avoid each other so that they experience distinct forces at all time.

$$P(r_{ij}) = \begin{cases} 10e^{-200r_{ij}}, & \text{if } i \in \text{shells and } j \in \text{shells} \\ 0, & \text{otherwise} \end{cases} \tag{15}$$

The total C-GeM energy of a given system with fixed cores involves an optimization of the shell positions to minimize the energy,

$$E_{CGeM} = \sum_i \sum_{j<i} E_{ij}^{elec}(r_{ij}) + E_{ij}^{gauss}(r_{ij}) \tag{16}$$

as per a usual Born-Oppenheimer assumption. The resulting shell configurations is used to generate the electrostatic potential on a set of given points using the following equation:

$$V(\mathbf{r}) = \sum_{i\in\text{cores}} \frac{q_i}{(|\mathbf{r}-\mathbf{r}_i|)} + \sum_{i\in\text{shells}} \frac{q_i}{(|\mathbf{r}-\mathbf{r}_i|)} \tag{17}$$

where all of the core and shell Gaussian charges are approximated by point charges at their center to speed up the ESP evaluation.

METHODS

To address both neutral and charged systems, we require an identification of the formal charge on each atom. All neutral atoms are initialized with a +1 core and a -1 shell at atomic center; a negatively charged atom receives an additional -1 charge shell based on its formal charge, and these additional shells are randomly displaced within $10^{-3}\mathring{A}$ distance to avoid overlaps; a positively charged atom is initialized with an incremented core charge (q_c = 1 + formal charge) and a -1 shell at the atomic center.

C-GeM training and testing protocol. There are five global parameters (λ ,ω_{core}, γ_{shell}, χ_{shell} and R_{shell}) and two atom-specific parameters per atom type (χ_i and R_i) in the C-GeM model. These parameters are fitted by minimizing the average mean absolute error (MAE_{avg}) over the training set with respect to *ab initio* ESP, where the MAE for one molecule is computed as:

$$MAE = \frac{1}{n}\sum_{i}^{n} |V_{C-GeM}(\mathbf{r}_i) - V_{DFT}(\mathbf{r}_i)| \tag{18}$$

where n is the total number of grid points, $V_{C-GeM}(\mathbf{r}_i)$ and $V_{DFT}(\mathbf{r}_i)$ are the C-GeM and the DFT ESP computed for a grid point at position $\mathbf{r}_i$.

The training set consists of 54 small molecules and 38 tripeptides, with an additional 19 tripeptides defining the validation set. The protein analogs are small molecules that represents the chemistry of amino acids, and a list of these molecule is provided in Supplementary Table S3 and Table S4. The 57 larger and more complex tripeptides are formulated by fragmentation of larger proteins culled from the PDB,[48] and uniformly sampled by amino acid residue types to capture the diversity of peptides. Three models are trained by minimizing the mean MAE of all of the small protein analogs and 2/3 of the tripeptides using the Nelder-Mead algorithm,[49] with one set of 19 tripeptides used as a validation set. The final model is obtained by the average of parameters from these three training models. The parameters for charge related atom types C_{+1}, N_{+1}, H_C, C_C and O_A are optimized while fixing all other parameters obtained from the neutral model with a charged training set (Ta-

ble S5) of 17 molecules including small charged molecules and tripeptides, and tested on a charged test set (Table S6) of 18 tripeptides that are positively charged, negatively charged or zwitterionic. Finally we also test the various models on the crambin protein (PDB ID 1CRN[50]), whose hydrogens are added using the Reduce (3.23) software.[51]

ESP generated by Gaussian charges vs point charges. There are two approaches to generate the ESP from a set of core and shell positions. One is the Gaussian charge approach, where the ESP is computed by

$$V(\mathbf{r}) = \sum_{i}^{cores} E_{ik}^{elec}(|\mathbf{r} - \mathbf{r}_i|) + \sum_{j}^{shells} E_{jk}^{elec}(|\mathbf{r} - \mathbf{r}_j|) \tag{19}$$

where a point(k) in space is treated as a fictitious core with $q_k = +1$ and $\alpha_k = 1569.8$, which is a Gaussian sharply peaked at position $\mathbf{r}$. This approach is the natural approach arise from the Gaussian definition of cores and shells. The other approach is to treat all cores and shells as point charges when calculating the ESP:

$$V(\mathbf{r}) = \sum_{i \in \text{cores}} \frac{q_i}{(|\mathbf{r} - \mathbf{r}_i|)} + \sum_{i \in \text{shells}} \frac{q_i}{(|\mathbf{r} - \mathbf{r}_i|)} \tag{20}$$

Note that this treatment of approximate cores and shells as point charges at their Gaussian center is only done in the process of generating the ESP, not in the optimization of shell positions. The two approaches gives essentially indistinguishable prediction in ESP as shown in Figure 2a), where the mean ESP generated with Gaussian charges aligns perfectly (R^2 = 0.99999993) with that generated with point charges. The average MAE between ESP generated with Gaussian charges and ESP generated with point charges is only $3.97 * 10^{-4}$ eV, which is trivial compared to the average magnitude of ESP at 0.755 eV. However, the point charge approach is advantageous in terms of calculation speed as it avoids the relatively expensive operation of erf function evaluation. This is demonstrated in Figure 2b), where the ESP time (the time to compute ESP from fixed core and shell positions) is plotted against the number of grid points for all molecules used in for training and testing

process for parameter optimization. The ESP time for the point charge treatment is clearly faster than the Gaussian charge treatment by roughly a factor of 10. When the number of points is a large number, this difference can be significant to influence the efficiency of ESP evaluation. As both methods provide essentially the same accuracy and the point charge treatment is clearly faster, in the following discussion of this paper, we will adopt the point charge approach to calculate the ESP. In the cases where the ESP grid points of interest is closer to the atomic center, we switch back to the Gaussian charge implementation.

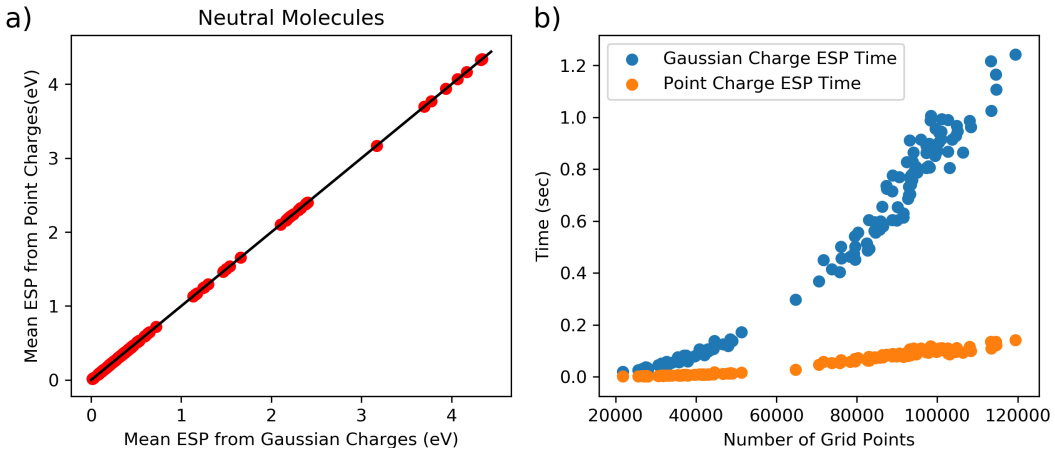


Figure 2: **a)** The mean ESP generated with Gaussian charges aligns perfectly with that generated with point charges. **b)** The time to compute ESP from core and shell positions with respect to number of grid points for different molecules using point charge and Gaussian charge treatment.

DFT reference and other methods for computing the ESP. The reference *ab initio* ESP for all molecules except crambin are generated with the Q-Chem 5.2 software package[52] using the ωB97X-V functional[53] with the def2-QZVPP basis set; the ESP of crambin is generated using ωB97X-V with the cc-pVDZ basis set. We also compare the results of C-GeM with other available methods including EEM, AM1-BCC, Hirshfeld, Iterative Hirshfeld, MBIS and AVH. The EEM-derived charges are obtained from the LAMMPS[54] ReaxFF[55] implementation of EEM using the peptide and protein parameters.[56] The AM1-BCC charges are obtained from antechamber tool part of AMBERTools. The Hirshfeld, Iterative Hirshfeld, MBIS and AVH charges are computed with IOData,[57] ChemTools[58] and HORTON 2.1.1 software packages.[59]

In addition, the electrostatic potential for crambin has been performed with continuum electrostatic calculations using the Adaptive Poisson–Boltzmann Solver (APBS) v3.0.0.[14] For APBS the hydrogen added crambin structure was prepared with PDB2PQR v3.1.0[60] using the AMBER force field, and enabling the generation of pqr files with atomic charges and radii. APBS computations were carried out with the linearized PB equation with a 1.0 dielectric constant for solvent and solute (protein) to mimic the vacuum condition in other calculations. Temperature was set to 298.15 K, and a single Debye–Hückel boundary condition was applied. The grid dimension was set to 353 x 353 x 353 such that the grid spacing is 0.149 x 0.125 x 0.150 Å, similar to the grid spacing in the molecular surface grid we used in the *ab initio* calculations. The ESP generated with APBS was mapped onto the molecular surface grid through the multivalue utility in APBS software package.

Grid resolution and timing metrics. The grid points on which electrostatic potentials are evaluated are generated following the Merz-Singh-Kollman (MK) scheme[7] on 10 evenly distributed layers of range from 1.4-2.54 vdW radii distance. Here we report the ESP generated on an average of 37,000 grid points for small protein analogs (average 12.4 atoms) and 90,000 grid points for tripeptides (average 37.4 atoms). The computation times are measured with the timeit python module on a single core Intel XEON Gold 6230 CPU unless otherwise mentioned. The times for DFT calculations are obtained from QChem output files.

RESULTS AND DISCUSSIONS

Neutral small protein analogs and tripeptides

In this study, we trained three protein C-GeM models that share common global parameters ω_{core}, γ_{shell}, λ, R_{shell} and χ_{shell}: 1) C-GeM without atom typing, where each element (H, C, N, O, S, Cl) has its own atomic parameters for the ionization potential and atomic radius. 2) C-GeM with C and H atom typed, where C is classified into polar carbon (C_A) and aliphatic carbon (C_B) based on whether it has an electronegative neighboring atom (N,O,S,Cl), and

H is classified into polar hydrogen (H_A) and aliphatic hydrogen(H_B) in the same way. 3) C-GeM with C, H and N atom typed, where on top of model 2, we further classify nitrogen according to the number of H neighbors it has, into N_A for N with 2 H neighbors, N_B for N with 1 H neighbor and N_C for N with no H neighbor. These three models are referred to as CGem, CGem_CH and CGem_CHN respectively, and their parameters are shown in Table 1.

Table 1: Parameters for C-GeM models CGem, CGem_CH and CGem_CHN. H_A for polar hydrogen, H_B for aliphatic hydrogen, H_C for hydrogen directly bonded to positive atoms; C_{+1} for carbon with a positive formal charge, C_A for polar carbon, C_B for aliphatic carbon, C_C for carbon directly bonded to positive atoms; N_{+1} for nitrogen with a positive formal charge, N_A for N with 2 H neighbors, N_B for N with 1 H neighbor and N_C for N with no H neighbor; O_A for oxygens in negatively charged acetate group

global parameters

$\omega_{core}(\mathring{A}^{-1})$	$\gamma_{shell}(\mathring{A}^{-2})$	λ	$R_{shell}(\mathring{A})$	χ_{shell} (eV)
0.152	5.220	2.103	0.708	19.956

C-GeM atomic parameters

	CGem		CGem_CH		CGem_CHN	
atom type	R(Å)	χ (eV)	R(Å)	χ (eV)	R(Å)	χ (eV)
H	0.67	-16.33	-	-	-	-
C	0.59	-19.12	-	-	-	-
N	0.44	-21.85	0.55	-23.08	-	-
O	0.34	-24.26	0.54	-22.83	0.51	-23.35
S	0.66	-21.28	0.86	-18.43	0.84	-19.29
Cl	0.31	-25.43	0.63	-21.73	0.56	-22.87
H_A	-	-	0.22	-12.97	0.20	-13.79
H_B	-	-	0.68	-16.42	0.65	-16.95
H_C	0.81	-15.49	0.52	-13.35	0.57	-13.52
C_A	-	-	0.77	-15.12	0.75	-15.74
C_B	-	-	0.57	-19.48	0.57	-19.49
C_C	0.60	-19.52	0.71	-13.26	0.72	-14.32
N_A	-	-	-	-	0.54	-23.65
N_B	-	-	-	-	0.61	-20.04
N_C	-	-	-	-	0.50	-24.43
O_A	0.62	-22.78	0.58	-23.50	0.60	-23.57
C_{+1}	0.55	-31.93	0.68	-30.15	0.74	-31.50
N_{+1}	0.88	-27.99	0.73	-38.02	0.72	-39.13

To sample the protein chemistry space, we developed the models with data from small protein analogs that covers the basic functional groups and scaffolds for peptides, along with tripeptides that describe actual protein chemistry but are small enough for high quality *ab initio* computation. The performance of these models was evaluated with respect to MAE_{avg} and RMSE_{avg} between the ESP of the reference ωB97X-V/def2-qzvpp theory and the ESP generated by the various models, as well as the dipole error obtained as the norm of the difference vector by subtracting the *ab initio* reference dipole from the approximate dipole. In Figure 3 we present the MAE_{avg} and mean dipole error of the C-GeM models as well as empirically derived partial charge method EEM and QM-calculation-based atomic partial charge methods Hirshfeld, HI, MBIS and AVH. The statistics of the results including RMSE_{avg} are listed in Table S1. Among the C-GeM models, atom typing hydrogen and carbon improves the MAE_{avg} from 0.067 eV to 0.059 eV and RMSE_{avg} from 0.094 eV to 0.082 eV in terms of ESP quality for small protein analogs, and improves the MAE_{avg} from 0.122 eV to 0.084 eV and RMSE_{avg} from 0.166 eV to 0.117 eV in terms of ESP quality for tripeptides. Nitrogen atom typing further improves the ESP MAE_{avg} and RMSE_{avg} down to 0.053 eV and 0.077 eV for small protein analogs, and 0.070 eV and 0.101 eV for tripeptides respectively. Oxygen atom typing were explored, but it showed minimal improvements on the training molecules while introducing an overfitting problem that degrades the results for the validation set. Therefore, oxygen was kept as its elemental type.

All three C-GeM models significantly outperform the EEM model (0.094 eV MAE_{avg}, for small protein analogs and 0.185 eV MAE_{avg} for tripeptides), which is the only method of comparable computational cost to C-GeM. The C-GeM models are also more accurate for tripeptides than the AM1-BCC charges (0.96 eV MAE_{avg}) that relies on the semi-empirical AM1 method and thus is computationally slower than C-GeM. All C-GeM models are significantly better than Hirshfeld (0.106 eV and 0.176 eV MAE_{avg} respectively), whereas the best CGem_CHN model also outperforms the AVH method(0.086 eV and 0.100 eV) by 30%, and slightly outperforms the HI method (0.058 eV and 0.080 eV), while MBIS (0.040 eV

and 0.053 eV) remains the best among all methods, albeit with much greater expense and thus not affordable for proteins.

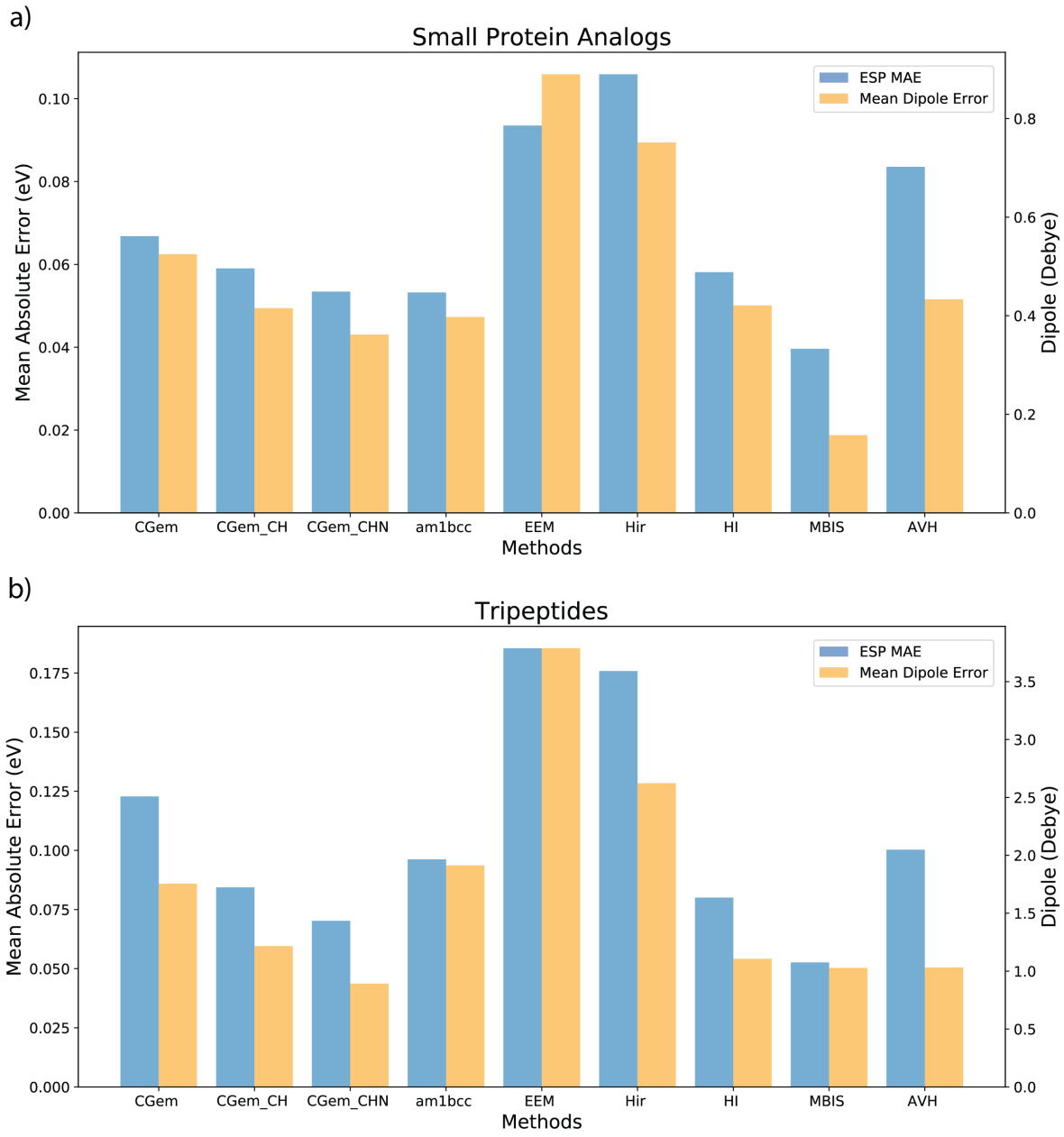


Figure 3: Average mean absolute error electrostatic potential and average dipole error of different atom typed C-GeM models, EEM, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ωB97X-V/ def2-qzvpp reference for a) 54 small protein analogs and b) 57 tripeptides, labeled with the average error and standard deviation within the set.

While producing an accurate ESP description that is comparable to *ab initio* calculation

based charges, the C-GeM models produce excellent prediction for molecular dipoles, which is a property that the model was not parameterized for, but arises naturally from off-centered shell positions. Atom typing improves the mean dipole errors for C-GeM models, from 0.525 Debye in CGem to 0.415 Debye in CGem_CH to 0.362 Debye in CGem_CHN for the small protein analogs set, and from 1.755 Debye in CGem to 1.216 Debye in CGem_CH to 0.892 Debye in CGem_CHN for the tripeptide set, similar in trend as to how atom typing improves the ESP MAE_{avg} and RMSE_{avg}. In the small protein analogs set, the mean dipole error of CGem_CHN is only inferior to that of MBIS (0.158 Debye), and superior to all other QM based or empirically derived methods including EEM (0.890 Debye), AM1-BCC (0.397 Debye), Hirshfeld (0.751 Debye), HI (0.421 Debye) and AVH (0.433 Debye). For the tripeptide set, CGem_CHN produces the best mean dipole error among all of the methods including MBIS (1.028 Debye), HI (1.106 Debye), AVH (1.032 Debye), whereas EEM (3.788 Debye), Hirshfeld (2.623 Debye) and AM1-BCC (1.914 Debye), have errors larger than all of the C-GeM models.

For all of the methods, the tripeptides are more challenging to predict than the small protein analogs because of their intrinsically larger size. For instance, the mean absolute ESP value is 0.190 eV for small protein analogs and 0.415 eV for tripeptides, and the mean dipole magnitude is 1.569 Debye for small protein analogs but 7.389 Debye for tripeptides. However, the relative performance among the methods we compared is quite stable across the two different datasets. The fact that the performance of the C-GeM models are relatively stable compared to QM based charge partitioning method, which are general methods that does not distinguish sizes or specific protein chemistry, reflects that C-GeM is transferable with respect to system size for protein like molecules.

While having similar accuracy, the C-GeM models are orders of magnitude faster than the QM charge partitioning approaches that are based on high quality *ab initio* calculations. The DFT benchmark calculation (ωB97X-V/def2-qzvpp) on average takes 8.4 minutes and 6.9 hours per molecule for small protein analogs and tripeptides, respectively, even after taking

advantage of OpenMP parallelization techniques, and the QM based charge partitioning methods require additional steps to partition atomic densities on top of the QM calculation. The AM1-BCC method takes 8.38 seconds for small protein analogues and 5.77 minutes for tripeptides using antechamber program, and although more efficient than the QM-based methods, is also 2-3 orders of magnitude slower than our C-GeM models.

Table 2: Computation time per molecule of C-GeM, CH atomtyped C-GeM, CHN atomtyped C-GeM on small protein analogs and tripeptides. Charge time is the time to initialize and optimize shell positions for C-GeM models, and ESP time is the time to map C-GeM cores and shells onto predefined grid points for electrostatic potential.

Small Protein Analogs			
	Charge Time (sec)	ESP Time (sec)	Total Time (sec)
CGeM	0.053	0.010	0.064
CGeM_CH	0.047	0.009	0.057
CGeM_CHN	0.044	0.009	0.053
Tripeptides			
	Charge Time (sec)	ESP Time (sec)	Total Time (sec)
CGeM	0.126	0.081	0.207
CGeM_CH	0.133	0.080	0.213
CGeM_CHN	0.121	0.080	0.201

By contrast, the C-GeM models can predict the ESP on the order of tenth of a second on a single core of Intel XEON Gold 6230 CPU, and all C-GeM models have very similar computational timings for the ESP, about 0.01 seconds for small protein analogs and about 0.08 seconds for tripeptides (Table 2), which is comparable to the EEM class of methods. The actual timing comparisons between EEM and C-GeM models are not directly comparable because the EEM times were obtained with a C++ code in LAMMPS and the C-GeM times were obtained with our in-house Python code, but EEM is the same order of magnitude for the system sizes we've investigated until this point; we return to timings again later in the crambin protein case. The internal comparisons among C-GeM models shows that atom typing did not slow down the calculation, despite adding additional step to classify the atoms. The charge time decreases in the order of CGem, CGem_CH, and CGem_CHN in

both the small protein analogs set and the tripeptides set, which suggests that atom typing of C,H and N helps the shell optimization process to converge faster.

To demonstrate that the C-GeM model can deal with the conformational variations of a molecule, we compute C-GeM ESP for a tripeptide molecule randomly selected from the PEPCONF[61] dataset. The error of C-GeM models relative to the ωB97X-V / def2-qzvpp reference on the 6 conformations of tripeptide LEU_TYR_GLN(Figure S1) are shown in Table 3, which supports the fact that all C-GeM models yield stable predictions on varied conformations of the same molecule.

Table 3: Mean absolute error (MAE) in eV on electrostatic potential (ESP) of different atom typed C-GeM models with respect to ωB97X-V / def2-qzvpp reference on 6 conformations of tripeptide LEU_TYR_GLN.

molecule	CGem MAE	CGem_CH MAE	CGem_CHN MAE
CONF_1	0.087	0.072	0.063
CONF_2	0.108	0.067	0.083
CONF_3	0.105	0.073	0.075
CONF_4	0.109	0.077	0.078
CONF_5	0.088	0.064	0.061
CONF_6	0.114	0.075	0.081

Charged small protein analogs and tripeptides

In the previous section, we demonstrated that C-GeM models can predict the ESP of molecules at accuracy comparable to *ab initio* generated charges but orders of magnitude faster for neutral small protein analogs and tripeptides. However, proteins under physiological conditions have residues that are charged under neutral pH, which would need specialized treatment in the C-GeM models. We considered two residues that are negative under neutral pH, aspartic acid (Asp) and glutamic acid (Glu), and two residues that are positive under neutral pH, arginine (Arg) and lysine (Lys). For the negatively charged residues, an extra shell is added onto the negatively charged atom (O_A for negative oxygen) as shown in Figure 4(a), which creates a net charge of -1 localized around the negatively charged atom. For

positively charged residues, the idea is to assign the core of the charged atom a +2 charge and mark it as a different atom type(C_{+1} for carbon and N_{+1} for nitrogen as shown in Table 1), while still having a shell on that atom to allow for shell movements. As shown in Figure

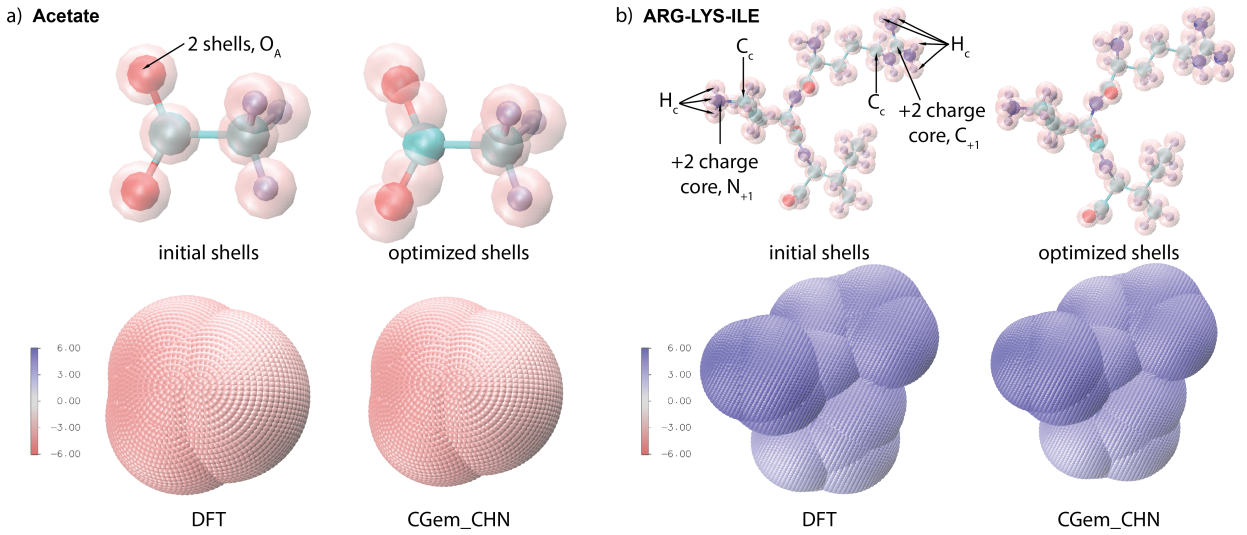


Figure 4: Demonstration for C-GeM on charged molecules **a)** Methylammonium (net -1 charge) **b)** Tripeptide ARG-LYS-ILE (net +2 charge)

4(b) for Lys, the positive nitrogen carries a +2 core, and for Arg, we placed the +2 core on the guanidino carbon instead of the formally charged nitrogen to account for the equivalence of the two guanidino nitrogens. We also find it useful to have separate atom type for the hydrogens (H_C) and carbons (C_C) that are directly bonded to the positive atoms.

With this protocol, we trained the parameters for the charged atoms, and fixing all of the parameters we obtained from the neutral model, using a training set of 17 molecule consisting of 4 small side chain analog molecules and 13 tripeptides that are positive, negative of zwitterionic (Table S5). The resulting models were tested on another 18 tripeptides (Table S6) with charged residues that the model has not seen. The MAE_{avg} and mean dipole errors C-GeM models on the charged dataset compared to QM based charges are presented in Figure 5 and Table S2. EEM charges are not included because the LAMMPS implementation of EEM fails to deal with non-zero charges. The charged molecules in general have larger mean ESP values (2.26 eV for the charged training set and 2.24 eV for the charged test set)

and much larger dipoles (78.1 Debye and 137.1 Debye for the charged training and test set, respectively), which could make the prediction more difficult.

In general Figure 5 and Table S2 show that the C-GeM models exhibit a stable performance on these difficult charged molecules that are not too far from their corresponding performance on the neutral molecules. The basic CGeM model yields 0.102 eV MAE_{avg} for the charged training set and 0.116 eV MAE_{avg} while the best CGem_CHN model yields 0.081 eV MAE_{avg} for the charged training set, and 0.076 eV MAE_{avg} for the charged test set. This is a significant improvement in MAE_{avg} for charge training and test set, respectively, over Hirshfeld charges (0.154 eV and 0.174 eV), and comparable to HI (0.060 eV and 0.071 eV), MBIS (0.060 eV and 0.065 eV) and AVH (0.078 eV and 0.086 eV). The dipole errors exhibit a similar trend: the best C-GeM model CGem_CHN reports a dipole error of 1.34 Debye for the charged training set and 1.15 Debye for charged test set, which is comparable to MBIS (1.33 Debye and 1.36 Debye) and significantly improved over Hirshfeld (1.96 Debye and 2.50 Debye), but worse than HI (0.88 Debye and 1.10 Debye) and AVH (0.88 Debye and 0.88 Debye). Both the trend and the numbers are very similar across the charged training set and testing set, which shows the generality of the models. The AM1-BCC charges are relatively accurate in the neutral molecule case, but clearly have some difficulty in predicting charged protein chemistry molecules, giving rise to a MAE of 0.161 eV and 0.149 eV, and dipole error at 3.67 Debye and 3.36 Debye for the charged training and test set, respectively.

It is worth noting that the overall dipole error is amplified due to the large magnitude, as we are defining dipole error as the norm of the difference vector between C-GeM or partial charge derived dipoles and the DFT dipole $|\boldsymbol{\mu}_{DFT} - \boldsymbol{\mu}_{C-GeM}|$, which captures both the magnitude and directional information. Hence with a large dipole, a small deviation in the angle θ between $\boldsymbol{\mu}_{DFT}$ and $\boldsymbol{\mu}_{DFT}$ can result in large errors in the norm of the difference vector, even if the error in magnitude $||\boldsymbol{\mu}_{DFT}| - |\boldsymbol{\mu}_{C-GeM}||$ is small. For instance, the 1.23 Debye dipole error in CGem_CHN can be decomposed into 0.65 Debye error in magnitude and 0.91 $^\circ$ in θ.

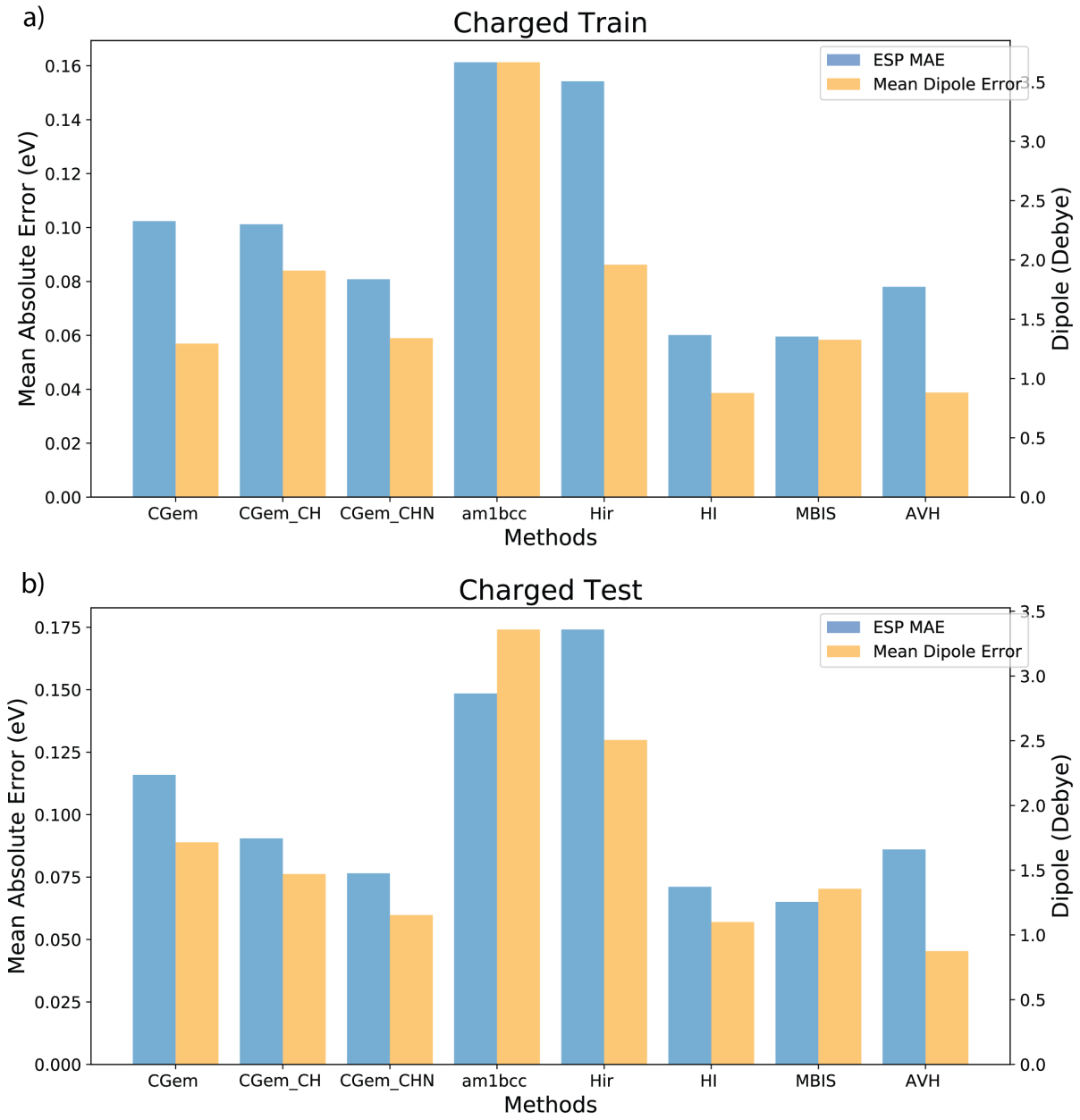
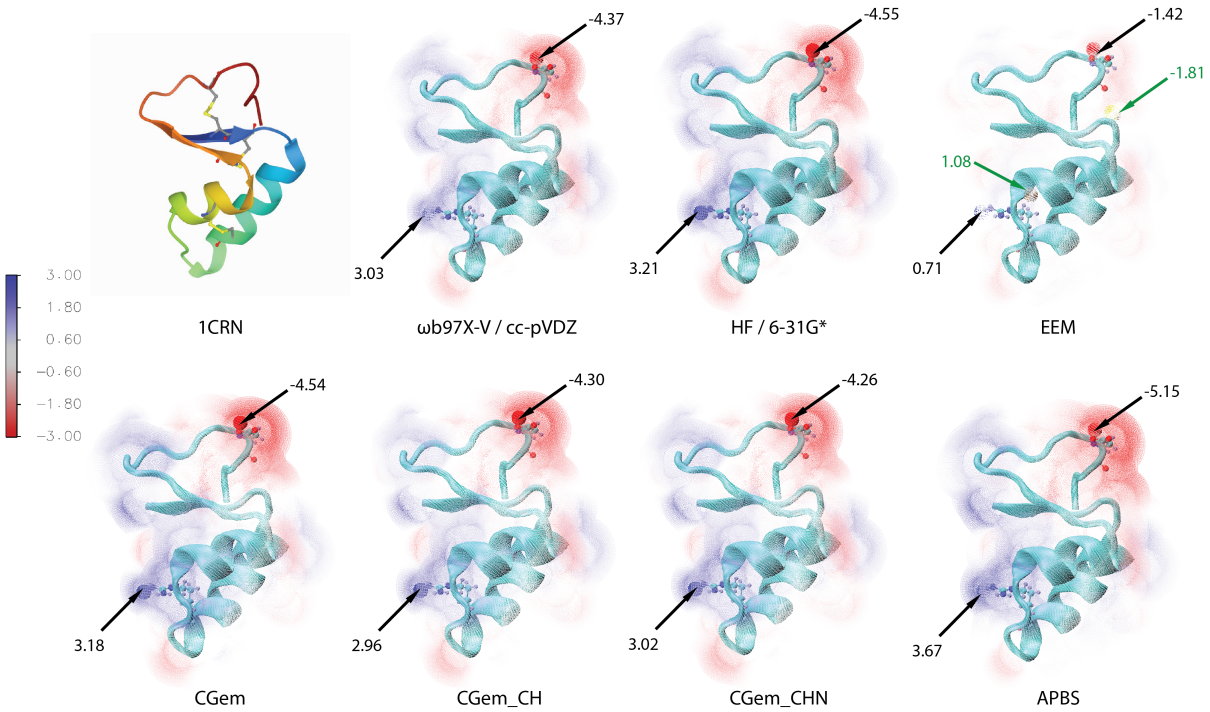


Figure 5: Mean absolute error (MAE) on electrostatic potential (ESP) and dipole error of different atom typed C-GeM models, AM1-BCC, Hirshfeld, iterative Hirshfeld, MBIS and AVH partial charges with respect to ωB97X-V / def2-qzvpp reference for a) training set for charged side chains b) testing set for charged side chains.

Evaluation of C-GeM model on the crambin protein

As the C-GeM models worked well to reproduce the DFT benchmark for the ESP and dipole directions in both the neutral and the charged cases of small molecules and protein fragments, we precede to examine C-GeM models on a full protein, crambin, which is difficult in terms of resources for the QM based partial charge methods, but totally accessible for C-GeM models as they are orders of magnitudes faster.

The ESP map of crambin is shown in Figure 6 with their minimum and maximum ESP value labeled. The C-GeM models give qualitatively correct predictions for the ESP compared to the DFT reference computed ESP with ωB97X-V / cc-pVDZ, with MAE of 0.13, 0.12 and 0.11 for CGem, CGem_CH and CGem_CHN respectively. These predictions are superior to the EEM method (0.49 eV MAE), which fails to describe the ESP qualitatively correctly due to unphysical long-range charged transfer, and APBS (0.25 eV MAE), which essentially is due to the AMBER ESP fitted partial charges (the dielectric constant was set to 1 to account for protein in vaccuum of all methods). The CGem_CHN predicts -4.26 eV and 3.02 eV as minimum and maximum on the ESP surface, which is very close to -4.37 eV and 3.03 eV predicted by the DFT reference at the same position in space. By contrast the EEM method yields a more featureless ESP, predicting a minimum and maximum of -1.42 eV and 0.71 eV, whereas the APBS result exaggerates the extremes with -5.15 eV and 3.67 eV for the minimum and maximum, respectively. Finally, the best C-GeM model CGem_CHN also gives a relatively acceptable dipole error of 6.45 Debye compared to the total 37.8 Debye for crambin as determined by the DFT benchmark. In this case the EEM dipole moment is egregiously incorrect.



	MAE (eV)	RMSE(eV)	Dipole error (Debye)
HF / 6-31G*	0.06	0.08	1.77
CGem	0.13	0.17	7.03
CGem_CH	0.12	0.16	7.53
CGem_CHN	0.11	0.15	6.45
EEM	0.49	0.65	33.19
APBS	0.25	0.32	-

Figure 6: Predicted ESP figure for crambin(1CRN) with ωB97X-V / cc-pVDZ, HF/6-31G*, EEM, CGem, CGem_CH, CGem_CHN and APBS. The electrostatic potential (in eV) at points with maximum and minimum ESP value for ωB97X-V / cc-pVDZ are labeled. The table presents the MAE and RMSE on ESP and the dipole error of these methods with respect to ωB97X-V / cc-pVDZ reference for crambin.

The advantage in computational efficiency for the C-GeM models is very significant in the case of this larger molecule of more than 600 atoms. The C-GeM models can predict the ESP on more than 500,000 grid points within 20 seconds, which is five orders of magnitude faster than the ωB97X-V / cc-pVDZ reference. The C-GeM models are also faster than APBS at the same grid resolution, noting that the speed of APBS suffer from first computing the ESP on a full-space grid of similar spacing, and then interpolation onto the molecular surface

grid. The best C-GeM model, CGem_CHN (13.7 sec) is faster than CGem (15.6 sec) and CGem_CH (17.4 sec) despite it requiring additional steps of atom typing, which again shows that atom typing speeds up the convergence of the shell position in the optimization cycles.

CONCLUSIONS AND OUTLOOK

The ability to generate accurate electrostatic potential surfaces for predicting protein binding motifs with high computational efficiency for high-throughput screening of drug molecules is an important area for structural based drug discovery. At present this dual goal of accuracy and efficiency has been difficult to achieve. Here we have introduced a new method for generating the ESP that is both accurate and fast using the C-GeM approach. We have shown that it offers accuracy comparable to the expensive *ab initio* methods with orders of magnitude reduction in expense, and is far more accurate than cheaper computational alternatives such as EEM or PBE approaches.

We have also shown that the EEM model and the density partitioning Hirshfeld schemes are the least competitive in regards accuracy, which is not surprising, but are compared here because of their continued popularity. The AM1-BCC model, usually thought of as an efficient method, was found to be inferior to the C-GeM models in both efficiency and accuracy, and is found to be unstable when computing charged protein fragments. While more first principle approaches such as HI, MBIS, or AVH are relatively accurate, they are computationally expensive and thus unsuitable for high-throughput computation on large proteins or for the many molecules required for high throughput screening applications.

In summary, the C-GeM force field accuracy comes in part from eliminating unphysical long-range charge transfer, by accounting for out-of-plane polarization, and charges are not required to be centered on atoms, thereby accounting for electrostatic features such as sigma holes that define important binding motifs for biomolecules. The C-GeM model is light-weight in parameters compared to other many-body force fields such as AMOEBA,[62]

which has many more atom types and many more parameters such as the atomic multipoles up through quadrupoles, atomic polarizability parameters, and damping functions. By contrast the protein C-GeM model has at most 15 atom types each with 2 atomic parameters that represent the electronegativity and ionization potential, and a common set of 5 global parameters for all atoms. In future development work we will advance C-GeM further to account for more complicated solvent environments and physiological salt conditions that are important for biomolecular recognition, and apply the model to more diverse applications beyond ESP predictions.

ACKNOWLEDGMENTS

The methods work was supported by the National Science Foundation under grant CHE-1955643 and the application area by the C3.ai Digital Transformation Institute. We also thank the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231.

SUPPORTING INFORMATION

Numerical values of ESP errors for all molecules and methods, and a file containing all SMILES string representation of all molecules in this study.

DATA AND SOFTWARE AVAILABILITY

All data and software for C-GeM can be accessed from thglab.berkeley.edu

DECLARATION OF INTERESTS

The authors declare no competing interests.

References

(1) Weiner, P. K.; Langridge, R.; Blaney, J. M.; Schaefer, R.; Kollman, P. A. Electrostatic Potential Molecular Surfaces. *Proceedings of the National Academy of Sciences* **1982**, *79*, 3754–3758.

(2) Scrocco, E.; Tomasi, J. The Electrostatic Molecular Potential as a Tool for the Interpretation of Molecular Properties. New Concepts II. 1973; p 95–170.

(3) McDonald, N. Q.; Lapatto, R.; Rust, J. M.; Gunning, J.; Wlodawer, A.; Blundell, T. L. New Protein Fold Revealed by a 2.3-Å Resolution Crystal Structure of Nerve Growth Factor. *Nature* **1991**, *354*, 411–414.

(4) Rathi, P. C.; Ludlow, R. F.; Verdonk, M. L. Practical High-Quality Electrostatic Potential Surfaces for Drug Discovery Using a Graph-Convolutional Deep Neural Network. *Journal of Medicinal Chemistry* **2020**, *63*, 8778–8790.

(5) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of Accurate Charges in Molecular Docking: Quantum Mechanical/Molecular Mechanical (QM/MM) Approach. *Journal of Computational Chemistry* **2005**, *26*, 915–931.

(6) Murray, J. S.; Politzer, P. The Electrostatic Potential: an Overview. *WIREs Computational Molecular Science* **2011**, *1*, 153–163.

(7) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *Journal of Computational Chemistry* **1984**, *5*, 129–145.

(8) Woods, R. J.; Chappelle, R. Restrained Electrostatic Potential Atomic Partial Charges for Condensed-Phase Simulations of Carbohydrates. *Theochem* **2000**, *527*, 149–156.

(9) Cordero, B.; Gómez, V.; E. Platero-Prats, A.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent Radii Revisited. *Dalton Transactions* **2008**, *0*, 2832–2838.

(10) Chen, D.-L.; Stern, A. C.; Space, B.; Johnson, J. K. Atomic Charges Derived from Electrostatic Potentials for Molecular and Periodic Systems. *The Journal of Physical Chemistry A* **2010**, *114*, 10225–10233.

(11) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

(12) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *WIREs Computational Molecular Science* **2013**, *3*, 198–210.

(13) Kumar, A.; Yoluk, O.; MacKerell Jr, A. D. FFParam: Standalone Package for CHARMM Additive and Drude Polarizable Force Field Parametrization of Small Molecules. *Journal of Computational Chemistry* **2020**, *41*, 958–970.

(14) Jurrus, E. et al. Improvements to the APBS Biomolecular Solvation Software Suite. *Protein Science* **2018**, *27*, 112–128.

(15) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *Journal of Computational Chemistry* **2002**, *23*, 1623–1641.

(16) Heidar-Zadeh, F.; Ayers, P. W.; Verstraelen, T.; Vinogradov, I.; Vöhringer-Martinez, E.; Bultinck, P. Information-Theoretic Approaches to Atoms-in-Molecules: Hirshfeld Family of Partitioning Schemes. *The Journal of Physical Chemistry A* **2018**, *122*, 4219–4245.

(17) Breneman, C. M.; Wiberg, K. B. Determining Atom-Centered Monopoles from Molecular Electrostatic Potentials. The Need for High Sampling Density in Formamide Conformational Analysis. *Journal of Computational Chemistry* **1990**, *11*, 361–373.

(18) Mulliken, R. S. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. IV. Bonding and Antibonding in LCAO and Valence-Bond Theories. *The Journal of Chemical Physics* **1955**, *23*, 2343–2346.

(19) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural Population Analysis. *The Journal of Chemical Physics* **1985**, *83*, 735–746.

(20) Lu, W. C.; Wang, C. Z.; Schmidt, M. W.; Bytautas, L.; Ho, K. M.; Ruedenberg, K. Molecule Intrinsic Minimal Basis Sets. I. Exact Resolution of ab initio Optimized Molecular Orbitals in Terms of Deformed Atomic Minimal-Basis Orbitals. *Journal of Chemical Physics* **2004**, *120*, 2629–2637.

(21) Knizia, G. Intrinsic Atomic Orbitals: An Unbiased Bridge between Quantum Theory and Chemical Concepts. *Journal of Chemical Theory and Computation* **2013**, *9*, 4834–4843.

(22) Verstraelen, T.; Vandenbrande, S.; Heidar-Zadeh, F.; Vanduyfhuys, L.; Van Speybroeck, V.; Waroquier, M.; Ayers, P. W. Minimal Basis Iterative Stockholder: Atoms in Molecules for Force-Field Development. *Journal of Chemical Theory and Computation* **2016**, *12*, 3894–3912.

(23) Bader, R. F. W. A Quantum Theory of Molecular Structure and its Applications. *Chemical Reviews* **1991**, *91*, 893–928.

(24) Hirshfeld, F. L. Bonded-atom Fragments for Describing Molecular Charge Densities. *Theoretica chimica acta* **1977**, *44*, 129–138.

(25) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical Analysis and Extension of the Hirshfeld Atoms in Molecules. *The Journal of Chemical Physics* **2007**, *126*, 144111.

(26) Lillestolen, T. C.; Wheatley, R. J. Atomic Charge Densities Generated using an Iterative Stockholder Procedure. *The Journal of Chemical Physics* **2009**, *131*, 144101.

(27) Manz, T. A.; Sholl, D. S. Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials. *Journal of Chemical Theory and Computation* **2010**, *6*, 2455–2468.

(28) Verstraelen, T.; Ayers, P.; Van Speybroeck, V.; Waroquier, M. The Conformational Sensitivity of Iterative Stockholder Partitioning Schemes. *Chemical Physics Letters* **2012**, *545*, 138–143.

(29) Verstraelen, T.; Ayers, P. W.; Van Speybroeck, V.; Waroquier, M. Hirshfeld-E Partitioning: AIM Charges with an Improved Trade-off between Robustness and Accurate Electrostatics. *Journal of Chemical Theory and Computation* **2013**, *9*, 2221–2225.

(30) Nalewajski, R. F.; Parr, R. G. Information Theory, Atoms in Molecules, and Molecular Similarity. *Proceedings of the National Academy of Sciences* **2000**, *97*, 8879–8882.

(31) Heidar-Zadeh, F.; Ayers, P. W.; Bultinck, P. Deriving the Hirshfeld Partitioning using Distance Metrics. *The Journal of Chemical Physics* **2014**, *141*, 094103.

(32) Heidar-Zadeh, F.; Ayers, P. W. How Pervasive is the Hirshfeld Partitioning? *The Journal of Chemical Physics* **2015**, *142*, 044107.

(33) Cho, M.; Sylvetsky, N.; Eshafi, S.; Santra, G.; Efremenko, I.; Martin, J. M. L. The Atomic Partial Charges Arboretum: Trying to See the Forest for the Trees. *ChemPhysChem* **2020**, *21*, 688–696.

(34) Vandenbrande, S.; Waroquier, M.; Speybroeck, V. V.; Verstraelen, T. The Monomer Electron Density Force Field (MEDFF): A Physically Inspired Model for Noncovalent Interactions. *Journal of Chemical Theory and Computation* **2017**, *13*, 161–179.

(35) Mortier, W. J.; Van Genechten, K.; Gasteiger, J. Electronegativity Equalization: Application and Parametrization. *Journal of the American Chemical Society* **1985**, *107*, 829–835.

(36) Mortier, W. J.; Ghosh, S. K.; Shankar, S. Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *Journal of the American Chemical Society* **1986**, *108*, 4315–4320.

(37) van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. ReaxFF: A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A* **2001**, *105*, 9396–9409.

(38) Geidl, S.; Bouchal, T.; Raček, T.; Svobodová Vařeková, R.; Hejret, V.; Křenek, A.; Abagyan, R.; Koča, J. High-quality and Universal Empirical Atomic Charges for Chemoinformatics Applications. *Journal of Cheminformatics* **2015**, *7*, 59.

(39) Ionescu, C.-M.; Geidl, S.; Svobodová Vařeková, R.; Koča, J. Rapid Calculation of Accurate Atomic Charges for Proteins via the Electronegativity Equalization Method. *Journal of Chemical Information and Modeling* **2013**, *53*, 2548–2558.

(40) Ouyang, Y.; Ye, F.; Liang, Y. A Modified Electronegativity Equalization Method for Fast and Accurate Calculation of Atomic Charges in Large Biological Molecules. *Physical Chemistry Chemical Physics* **2009**, *11*, 6082–6089.

(41) Bertels, L. W.; Newcomb, L. B.; Alaghemandi, M.; Green, J. R.; Head-Gordon, M. Benchmarking the Performance of the ReaxFF Reactive Force Field on Hydrogen Combustion Systems. *Journal of Physical Chemistry A* **2020**, *124*, 5631–5645.

(42) Boes, J. R.; Groenenboom, M. C.; Keith, J. A.; Kitchin, J. R. Neural Network and ReaxFF Comparison for Au Properties. *International Journal of Quantum Chemistry* **2016**, *116*, 979–987.

(43) Lee Warren, G.; Davis, J. E.; Patel, S. Origin and Control of Superlinear Polarizability Scaling in Chemical Potential Equalization Methods. *The Journal of Chemical Physics* **2008**, *128*, 144110.

(44) Leven, I.; Head-Gordon, T. C-GeM: Coarse-Grained Electron Model for Predicting the Electrostatic Potential in Molecules. *The Journal of Physical Chemistry Letters* **2019**, *10*, 6820–6826.

(45) Mitchell, P. J.; Fincham, D. Shell Model Simulations by Adiabatic Dynamics. *Journal of Physics: Condensed Matter* **1993**, *5*, 1031–1038.

(46) Naserifar, S.; Brooks, D. J.; Goddard, W. A.; Cvicek, V. Polarizable Charge Equilibration Model for Predicting Accurate Electrostatic Interactions in Molecules and Solids. *The Journal of Chemical Physics* **2017**, *146*, 124117.

(47) Voityuk, A. A.; Stasyuk, A. J.; Vyboishchikov, S. F. A Simple Model for Calculating Atomic Charges in Molecules. *Phys. Chem. Chem. Phys.* **2018**, *20*, 23328–23337.

(48) Li, J.; Bennett, K. C.; Liu, Y.; Martin, M. V.; Head-Gordon, T. Accurate Prediction of Chemical Shifts for Aqueous Protein Structure on "Real World" Data. *Chemical Science* **2020**, *11*, 3180–3191.

(49) Gao, F.; Han, L. Implementing the Nelder-Mead Simplex Algorithm with Adaptive Parameters. *Computational Optimization and Applications* **2012**, *51*, 259–277.

(50) Teeter, M. M. Water Structure of a Hydrophobic Protein at Atomic Resolution: Pentagon Rings of Water Molecules in Crystals of Crambin. *Proceedings of the National Academy of Sciences* **1984**, *81*, 6014–6018.

(51) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: using Hydrogen Atom Contacts in the Choice of Side-Chain Amide Orientation; Edited by J. Thornton. *Journal of Molecular Biology* **1999**, *285*, 1735–1747.

(52) Shao, Y. et al. Advances in Molecular Quantum Chemistry Contained in the Q-Chem 4 Program Package. *Molecular Physics* **2015**, *113*, 184–215.

(53) Mardirossian, N.; Head-Gordon, M. ωB97X-V: A 10-parameter, Range-separated Hybrid, Generalized Gradient Approximation Density Functional with Nonlocal Correlation, Designed by a Survival-of-the-fittest Strategy. *Physical Chemistry Chemical Physics* **2014**, *16*, 9904–9924.

(54) Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics* **1995**, *117*, 1–19.

(55) Aktulga, H. M.; Fogarty, J. C.; Pandit, S. A.; Grama, A. Y. Parallel Reactive Molecular Dynamics: Numerical Methods and Algorithmic Techniques. *Parallel Computing* **2012**, *38*, 245–259.

(56) Monti, S.; Corozzi, A.; Fristrup, P.; Joshi, K. L.; Shin, Y. K.; Oelschlaeger, P.; Duin, A. C. T. v.; Barone, V. Exploring the Conformational and Reactive Dynamics of Biomolecules in Solution using an Extended Version of the Glycine Reactive Force Field. *Physical Chemistry Chemical Physics* **2013**, *15*, 15062–15077.

(57) Verstraelen, T. et al. IOData: A python library for reading, writing, and converting computational chemistry file formats and generating input files. *Journal of Computational Chemistry* **2021**, *42*, 458–464.

(58) Heidar-Zadeh, F.; Richer, M.; Fias, S.; Miranda-Quintana, R. A.; Chan, M.; Franco-PÃ©rez, M.; GonzÃlez-Espinoza, C. E.; Kim, T. D.; Lanssens, C.; Patel, A. H.; Yang, X. D.; VÃ¶hringer-Martinez, E.; CÃ¡rdenas, C.; Verstraelen, T.; Ayers, P. W. An

Explicit Approach to Conceptual Density Functional Theory Descriptors of Arbitrary Order. *Chemical Physics Letters* **2016**, *660*, 307–312.

(59) Verstraelen, T.; Tecmer, P.; Heidar-Zadeh, F.; González-Espinoza, C. E.; Chan, M.; Kim, T. D.; Boguslawski, K.; Fias, S.; Vandenbrande, S.; Berrocal, D.; Ayers, P. W. `http://theochem.github.com/horton/`,.

(60) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Research* **2004**, *32*, W665–W667.

(61) Prasad, V. K.; Otero-de-la Roza, A.; DiLabio, G. A. PEPCONF, a Diverse Data Set of Peptide Conformational Energies. *Scientific Data* **2019**, *6*, 180310.

(62) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. Current Status of the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B* **2010**, *114*, 2549–2564.