Supporting information for:

Polarized Protein-Specific Charges from
Atoms-in-Molecule Electron Density Partitioning

Louis P. Lee,† Daniel J. Cole,*‡† Chris-Kriton Skylaris,¶ William L. Jorgensen,‡
and Mike C. Payne†

TCM Group, Cavendish Laboratory, 19 JJ Thomson Ave, Cambridge CB3 0HE, United Kingdom,
Department of Chemistry, Yale University, New Haven, Connecticut 06520-8107, United States,
and School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, United
Kingdom

*To whom correspondence should be addressed
†University of Cambridge
‡Yale University
¶University of Southampton
Supplementary methods

Simulations of bulk water

Bulk water supercells were generated using the tleap module in AMBER11 by solvating a single water molecule with the TIP3P water model in a cubic box. Each supercell structure was minimized then heated to 300K in 300 ps followed by equilibration for 5 ns in the NPT ensemble. The last snapshot from each supercell structure was used for single point DFT calculations in ONETEP under periodic boundary conditions, using a psinc spacing of 0.45 Bohr corresponding to a kinetic energy cutoff of $\sim 1000$ eV, NGWF localization radii $r_{NGWF}$ of 10 Bohr initialized as in Section 2.4, and the PBE exchange-correlation functional. A separate post-processing analysis was then performed on the converged electronic density to compute DDEC/ONETEP charges.

Relaxing simulation parameters

As DDEC charges are relatively insensitive to the quality of the underlying ab initio calculation such as basis set size, it is possible to relax convergence parameters without significantly impacting the accuracy of computed charges. The aforementioned calculations were performed using 12 density kernel optimization (LNV) steps per NGWF iteration. Using relaxed convergence parameters of 800 eV psinc energy cutoff, 8.0 bohr NGWF localization radii, and variable 4–8 LNV steps per NGWF optimization cycle reduced the calculation time of the 2502-atom bulk water system to three hours on the same system and number of cores. The mean absolute deviation of the resulting DDEC charges is less than 0.01 $e$ from the calculation shown in Figure 1(b).

Simulations of phenol

Phenol structures for DDEC analysis were generated using AMBER11. The phenol molecule was modeled using the GAFF force field, and was solvated in a periodic box by TIP4P water molecules. The system was equilibrated in the NPT ensemble and five snapshots were extracted at 300 K comprising the phenol molecule and 20 nearest water neighbors. The lysozyme L99A/M102E
X-ray crystal structure (3GUO) was used as the starting point for simulations of phenol in the protein binding pocket. The protein was modeled using the ff99SB biomolecular force field and the system was equilibrated at 300 K in a periodic water box. Five snapshots were extracted for DDEC analysis, comprising phenol, five nearby residues (I78, L84, A99, E102, V103) and one water molecule. RESP charges for phenol were calculated based on a HF/6-31G* electronic structure calculation of the optimized structure in Gaussian09 and used the antechamber module of Amber11.

Simulations of proteins

ONETEP calculations for proteins were performed in the same way as the 25 small molecules, as described in the main text, except that implicit solvent was used to describe the aqueous environment. Using a smeared-ion formalism, the molecular Hartree energy is obtained not in reciprocal space, but rather by solving the Poisson equation (homogeneous in vacuum, inhomogeneous in solution) in real space, under open boundary conditions. This is achieved via a multigrid approach. A 10 Bohr gap was left between the electron density and the edge of the multigrid and the ion smearing width set to 0.8 Bohr. The converged electronic density obtained in vacuum was used to generate the density-dependent dielectric cavity in solution. The values of the solvation parameter $\beta$ and electronic density threshold $\rho_0$ were 1.3 and $7.8 \times 10^{-4}$ a.u. respectively. The relative dielectric permittivity of the solvent was set to 80.0 for all implicit solvent calculations. DDEC charges were computed in a post-processing manner using the electronic density of the molecule converged in solvent. That is, the density from which the DDEC charges are derived have been polarized by the solvent.

MM models were prepared from the initial crystal structures by solvation in TIP3P water and neutralization by counterions followed by energy minimization in a cubic box with minimum distance from protein atoms of 10 Å. The system was heated from 0 to 300 K over a period of 100 ps followed by equilibration for 1.2 ns. NPT production runs were performed for 10 ns with a 1 fs timestep at 1.0 atm, 300 K, using Langevin dynamics with a collision frequency of 1.0 ps$^{-1}$. Tra-
jectories were saved every 50 fs for analysis. The ff99SB force-field was used throughout. For simulations with DDEC charges, the ff99SB charges were replaced prior to minimization.

**Convergence of DDEC charges**

Figure S1: DDEC/ONETEP charge convergence for nitroethane (CH$_3$CH$_2$NO$_2$), measured as the RMS difference from charges calculated using the finest tested parameter value (in brackets), for: (a) ONETEP psinc energy cutoff (1600 eV), (b) $r_{NGWF}$ (14.0 Bohr), (c) $r_{max}$ (11.0 Bohr), and (d) $N_{Rad}$ (160 shells corresponding to a radial shell spacing of 0.0625 Bohr). Charges are converged to much less than 0.01 $e$ for the parameters chosen.
Comparison between PBE and B3LYP functionals

Figure S2: Correlation between DDEC/CHARGEMOL charges from PBE and B3LYP DFT calculations in GAUSSIAN09 for the 25 molecule benchmark set.

Phenol charge comparison

Figure S3: Phenol charge comparison between (a) RESP HF/6-31G* in vacuum, (b) RESP B3LYP/6-31G* with PCM solvation, and (c) DDEC/ONETEP charges as in Figure 4(b). RMS charge deviation with respect to (c) is 0.04 e for both (a) and (b).
Comparison with the AMBER ff03 force field

![Comparison graph](image)

| Protein | Total Charge (e) | $\Delta V$ (kcal/mol) | $|\boldsymbol{\mu}_{DFT} - \boldsymbol{\mu}|$ (D) |
|---------|-----------------|------------------------|---------------------------------|
| 1UBQ    | 0               | 10.06                  | 46.3                            |
| 1MHN    | -4              | 7.68                   | 21.4                            |
| 6LYT    | +8              | 7.88                   | 23.2                            |

Figure S4: (Top) Comparison between DDEC/ONETEP partial atomic charges and the AMBER ff03 force field for ubiquitin. Correlation coefficient = 0.943, MAD = 0.095 e. (Bottom) RMS error ($\Delta V$) between the ab initio DFT potential and Coulombic point charge potentials derived from AMBER ff03 charges for three proteins as calculated using Eq. 11. Also shown are errors in the protein dipole moment vectors, which have been calculated relative to the center of mass of each protein to remove ambiguity for charged systems.
Comparison with IH and ISA charges

Figure S5: (a) Convergence profile for DDEC and ISA charges for the 2502 atom bulk water system in ONETEP. Charges are considered converged when their maximum absolute change for any atom between three consecutive iterations is $< 2 \times 10^{-5}$ e. The slow convergence of ISA charges is observed to be a general feature of the method, which can be attributed to the shallow optimization landscape presented by the density distribution around embedded atoms. The addition of IH reference densities into the optimization alleviates this issue. (b) Atomic and molecular charge distributions for the DDEC and ISA method in ONETEP for various bulk water supercells. The ISA charge distribution has a much greater spread, indicating its environmental sensitivity to atoms in embedded regions. (c) Comparison between experimental and calculated backbone N–H order parameter $S^2$ for 1UBQ using the ISA and IH methods in ONETEP for a 10 ns trajectory. IH charges reproduce experimental data more closely, with a root-mean-squared difference of 0.052, compared with 0.078 for ISA.