JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

On the convergence of atomic charges with the size of the enzymatic environment

Danny E. P. Vanpoucke, Julianna Olah, Frank De Proft, Veronique Van Speybroeck, and Goedele Roos

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/ci5006417 • Publication Date (Web): 10 Feb 2015

Journal of Chemical Information and Modeling is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

On the Convergence of Atomic Charges with the Size of the Enzymatic Environment

Danny E.P. Vanpoucke[1,], Julianna Oláh[2], Frank De Proft[3], Veronique Van Speybroeck[1], Goedele Roos[3,4,*]*

[1] Center for Molecular Modeling (CMM), Ghent University Technologiepark 903, 9052 Zwijnaarde, Belgium

[2] Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, Szent Gellért tér 4, 1111 Budapest, Hungary

[3] Department of General Chemistry, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.

[4] Department of Structural Biology of the VIB and Structural Biology Brussels, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.

[*]Corresponding authors: Danny.Vanpoucke@UGent.be and groos@vub.ac.be

Abstract

Atomic charges are a key concept to give more insight into the electronic structure and chemical reactivity. The Hirshfeld-I partitioning scheme applied to the model protein human 2-cysteine peroxiredoxin thioredoxin peroxidase B is used to investigate how large a protein fragment needs to be in order to achieve convergence of the atomic charge of both, neutral and negatively charged residues. Convergence in atomic charges is rapidly reached for neutral residues, but not for negatively charged ones. This study pinpoints difficulties on the road towards accurate modeling of negatively charged residues of large biomolecular systems in a multiscale approach.

Introduction

Atomic charges are an important tool to study electronic structure and chemical reactivity in, for example, protein reaction mechanisms. The following examples illustrate their importance: molecular force fields use charges to model electrostatic interactions[1]; the equilibrium constant of acid dissociation (p*K*a) can be related to the charge of the conjugate base[2-6]. Further, atomic charges can be used to provide more insight into the reactivity via the atom-condensed Fukui function (in a finite difference approximation), indicating the reactivity towards a nucleophilic or electrophilic attack of a soft reagent on a particular (protein) site[7]. Fukui functions calculated on protein fragments have been used to understand and successfully predict the regioselectivity found in protein reaction mechanisms[4]. Most popular in such reactivity analysis are charges derived from quantum mechanical computations, such as the Mulliken population analysis[8-11] and the natural population analysis (NPA)[12]; these methods are sometimes denoted as wave function based methods. Unfortunately, the scaling of the computational cost in function of the number of electrons limits the routine calculation of these orbital-based charges to small protein fragments: the calculation of the NPA and Mulliken charges becomes computationally very expensive due to the procedure to obtain the molecular orbitals of large systems (>200 atoms) using localized basis sets. Alternatives to overcome these limitation are semi-empirical QM methods[13-16], as well as linear scaling QM codes[17, 18]. In contrast, the Hirshfeld-I (HI) atoms-in-molecules partitioning scheme[19] is purely electron density based, similar to Baders' Quantum Theory of Atoms In Molecules (QTAIM)[20], and as such can be performed as a grid-based, basis set independent charge scheme. As a result, it can be easily applied on larger systems once the electron density is generated [21, 22 23].

Previous benchmark studies on penta-alanine pointed out that the Hirshfeld-I charge scheme could reproduce the dipole moments correctly and that this charge scheme was robust with respect to geometrical changes [24]. Therefore, the Hirshfeld-I charge scheme is applied here to investigate how large a protein fragment needs to be in order to achieve convergence in the atomic charges of a central part of the system. In this contribution, we study the atomic charge of a catalytic cysteine residue (Cys51) in the redox protein human 2-cysteine peroxiredoxin thioredoxin peroxidase B (Tpx-B)[25] in spherical model systems of 3, 5, 7, 9, and 11 Å radius (Figure 1) around the Sγ atom of Cys51. Since Cys51 can be present either as a thiolate (S^-) or as a thiol (SH) in Tpx-B, using this particular cysteine based redox protein as model system allows the study of the charge convergence of both negatively charged and neutral residues in one model protein.

In this study, we find that convergence in atomic charges is rapidly reached for uncharged residues, but not for charged residues. These findings might have implications for the charge evaluation by (biomolecular) force fields, since the accurate representation of electrostatic interactions is crucial in any force field [1, 26].

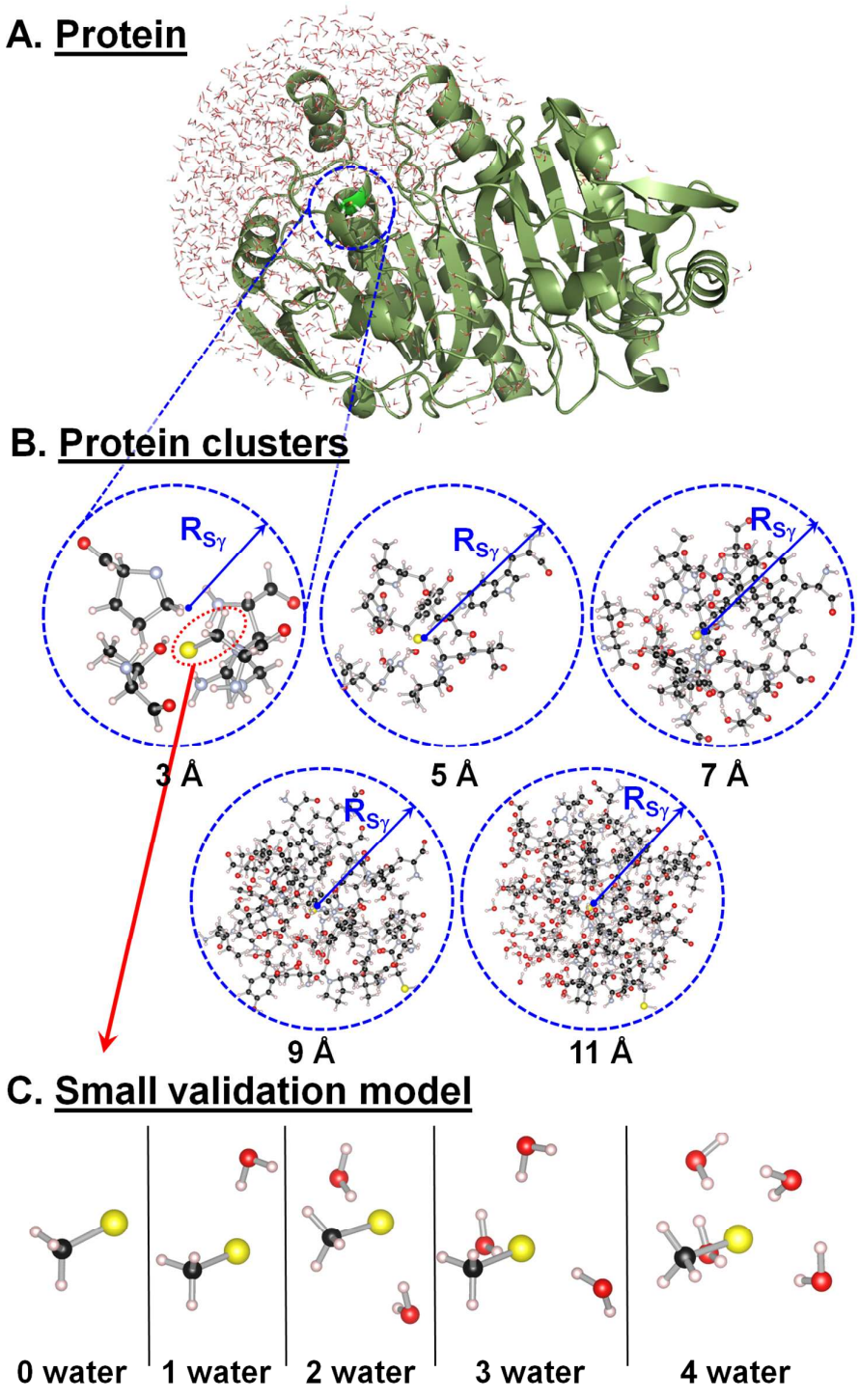

Figure 1: Studied systems: A. Cys51 in the dimeric redox protein human 2-cysteine peroxiredoxin thioredoxin peroxidase B (Tpx-B). Chains A and B of the X-Ray structure of Tpx-B are shown, solvated within a 60 Å water sphere with pre-equilibrated water molecules, represented by the TIP3P model, centered on the Sγ sulfur of Cys 51 (Chain A). B. Spherical protein + water clusters of 3, 5, 7, 9, and 11 Å radius around the Sγ atom of Cys51, cut out from the final structure of a 200ps

MD run. Here, the clusters with Cys51 in the thiolate form are shown. Similar clusters with Cys51 in the thiol form are also considered (Table1). C. Small validation test model.

Computational methods

Protein systems In the calculations, chains A and B of the X-Ray structure of human 2-cysteine peroxiredoxin thioredoxin peroxidase B (Tpx-B)[25] (PDB ID: 1QMV) are used, as the two chains together form the active dimeric form of the protein (Figure 1A). Hydrogen atoms are added to the initial X-Ray structure and their positions are optimized. The protonation states of acid and basic residues are predicted by using the PROPKA program[27] and all titratable residues are modeled in the natural protonation states. The hydrogen-bonding environment of all histidine residues is checked in order to account for the possible hydrogen bonds surrounding them. His 83 and His 168 are modeled as δ-protonated and His 197 as ε-protonated. The Sγ sulfur atom of Cys51 of Chain A is used as the centre of the system. Cys51 can be present in the thiolate (S^-) or in the neutral thiol (SH) form. Both protonation states are considered here and the structures for each protonation state are prepared independently from each other. The structures are solvated within a 60 Å box, centered on the Sγ sulfur of Cys 51 (Chain A), with 8000 pre-equilibrated water molecules, represented by the TIP3P model. Water molecules farther than 25 Å from the Sγ sulfur of Cys 51 (Chain A) are removed. The added water is then equilibrated by stochastic boundary MD at 300 K over 20 ps with respect to the protein structure and minimized. Then, all atoms within a 25 Å sphere around Sγ sulfur of Cys 51 (Chain A) are structurally optimized. Due to the size of the protein (10079 atoms) structural optimization is performed using the CHARMM27 force field [28] for protein and water molecules, while the capping N-terminal N-carbamoyl-alanine (NCB) residue is described by a custom CHARMM topology file, for which atom typing and assignment of parameters and charges are taken from an analogous residue. Details of the topology file and of the parameters of NCB can be found in the supporting information of a previous publication by some of the present authors [29]. All modeling calculations are carried out using the CHARMM software package [30]. The S^- form of cysteine is created using a previously published patch residue[29]. Parameters for the S^- form were previously published[31, 32].

The optimization step is followed by stochastic boundary MD simulation[33] of the whole system. Atoms farther than 25 Å from the Sγ sulfur of Cys 51 (Chain A) are fixed throughout the simulations, according to ref. [34]. All systems are heated to 300 K over 60 ps followed by a 300 ps long equilibration of the system. A subsequent MD run at 300 K is carried out over 200 ps.

Starting from the final structures of the 200 ps MD runs of the S^- form and the SH form, cuts are made around Sγ sulfur of Cys 51 (chain A), using sphere sizes with radius = 3, 5, 7, 9, and 11 Å (Table 1, Figure 1B). Residues with at least one atom within the sphere cutoff distance are completely kept, and resulting terminal structures are capped with hydrogen atoms. This resulted in 10 systems, which are further divided into two sets: in the first set, all atoms inside the sphere are included in the calculation (protein and water molecules), as indicated in Figure 1B, while in the second set, only the protein part is considered without the water molecules. To differentiate between the two, the presence of

water molecules is indicated by a superscript *w* (Table 1). The charge density grids for these 20 systems with different spheres sizes (radius = 3, 5, 7, 9, and 11 Å) around the Sγ atom of Cys51 are obtained with density functional theory (DFT) calculations (see next paragraph).

Hirshfeld(-I) calculations The DFT calculations to generate electron densities are performed within the projector augmented wave method as implemented in the Vienna ab initio Package (VASP) program using both the local density approximation (LDA) as parameterized by Ceperley and Alder and the generalized gradient approximation functional as constructed by Perdew, Burke, and Ernzerhof (PBE) [35-40]. For the second row elements (C, N, and O) only the *2s* and *2p* electrons are considered as valence electrons, while for S only the 3s and 3p electrons are considered. The plane wave kinetic energy cutoff is set to 500 eV, and the Brillouin zone is sampled using only the Γ-point[41]. Due to the periodic nature of the code, a vacuum region of 15 Å is included to prevent the periodic copies of the molecular fractions from interacting (Figure S1). In addition, also dipole corrections are included to prevent the possible interaction of (spurious) dipoles. The atomic charges of the systems are calculated using our grid based implementation of the Hirshfeld-I partitioning scheme in the HIVE code [19, 21, 22, 42]. Atom centered spherical integrations are performed using Lebedev-Laikov grids of 1202 grid points per shell and a logarithmic radial grid [43, 44]. The iterative scheme is considered converged when the largest difference in charge of every system atom is less than $1.0x10^{-5}$ electron in two consecutive iterations. To generate accurate reference densities for anions the R4 method presented elsewhere[21, 22], is used in this work. The reader is referred to section S3 of the SI for more details on this method.

CH_3S^- test systems To test the validity of the results obtained for the large protein clusters using the methodology explained in the previous paragraph, model calculations on the following representative small test systems are performed (Figure 1C): CH_3S^- anion surrounded by 0 to 4 water molecules. The CH_3S^- clusters are optimized at the B3LYP/6-311++G(d,p) level before a charge calculation is conducted. Frequency calculations are performed to check if the geometry is a minimum using the opt + freq keyword. All presented minimum energy conformers correspond to structures having no imaginary frequencies. HI charges are calculated using the same method employed for the biomolecular clusters, while NPA charges are obtained at the PBE/6-311++G(d,p) level, for the sake of comparison with the larger biomolecular systems. The later QM calculations and all structure optimizations for these small systems are performed using Gaussian09 [45].

Table 1. Number of atoms (N) and formal charge (Q) for the different protein clusters under study (Figure 1B). The superscript w indicates the protein clusters in which the water molecules (number of present water molecules given in brackets) are present as indicated in Figure 1B. N and Q refer to the clusters in which the water molecules are omitted.

Sphere size	Thiolate (S^-)				Thiol (SH)			
	N	Q	N^w	Q^w	N	Q	N^w	Q^w
3 Å	69	0	na	na	62	+1	na	na
5 Å	166	0	175 (3)	0	203	0	206 (1)	0
7 Å	321	-1	342 (7)	-1	351	0	366 (5)	0

9 Å	580	-1	646 (22)	-1	636	0	678 (14)	0
11 Å	779	-2	917 (46)	-2	782	-1	890 (36)	-1

Results

Hirshfeld(-I) charges are calculated using the electron densities generated within DFT, using the projector augmented wave (PAW) method as implemented in the Vienna ab initio Package (VASP). It might be remarked that the use of a solid state physics code, centered on the PAW methodology and periodic boundary conditions is not conventional to simulate an isolated molecule. However, in this case where the size of the biomolecular systems of interest (see Table 1 and Figure 1 for the scale of the systems) is rather large this setup was rather efficient to generate the input required to obtain reliable Hirshfeld charges. The use of uniform grids instead of atom centered grids, and plane waves instead of gaussian type orbitals allows for these large systems to be treated at a surmountable computational cost. As such, this type of setup is used to provide the quality of input required to obtain reliable Hirshfeld charges. Not using linear scaling nor empirical fitting, for the current systems, it was only possible to obtain densities of the protein clusters using the 3-21G basis set within a reasonable computational time.

To check the functional independence of the results, the electron densities are generated with both the LDA and PBE functional. Because the goal of our work is to check how the charge, which is a partitioning of the electron density, behaves as function of the protein cluster size, this choice of functionals suffices. It might be expected that results obtained with hybrid functionals would give qualitatively the same trend as they contain large portions of the here tested functionals.

Although the absolute values are influenced by the used functional, the trends of the calculated HI charges using LDA and PBE are exactly the same, as is shown in Figure S2. Since PBE is better suited to describe the density inhomogeneity in a molecular system, only the PBE based values are presented in this work. In all the calculations, the protein clusters are considered as neutral .This represents the situation in gas phase in which all titrable groups are charge neutral, although formally, in solvent, several clusters don't have a formal charge equal to 0 (Table 1). We found that the use of the formal charges of these clusters did not modify the atomic charges significantly (as can be seen in Table S3 of the SI).

Figure 2 shows the Hirshfeld-I charge to be much larger (in absolute value) than the Hirshfeld charge. This is due to a well-known issue of Hirshfeld charges. By construction, Hirshfeld charges are as similar as possible to the reference ionic densities (i.e. the starting guess for the atoms-in-molecules atomic charge). This has two consequences: 1) too small (in absolute value) atomic charges are obtained since the reference densities most generally used are those of neutral atoms, and 2) the use of different sets of reference ions give rise to different Hirshfeld charges for the same atom in the same system. The Hirshfeld-I scheme was

especially developed to alleviate this latter issue[19]. Through its iterative nature, the Hirshfeld-I scheme leads consistently to exactly the same atomic charges independent of the starting guess for the atomic charges. In addition, this also resolved the similarity issue of the original Hirshfeld scheme, giving rise to atomic charges that are consistently larger in size than Hirshfeld charges. As such we will only report the Hirshfeld-I charges in this work, the Hirshfeld charges are presented in the supplementary information.

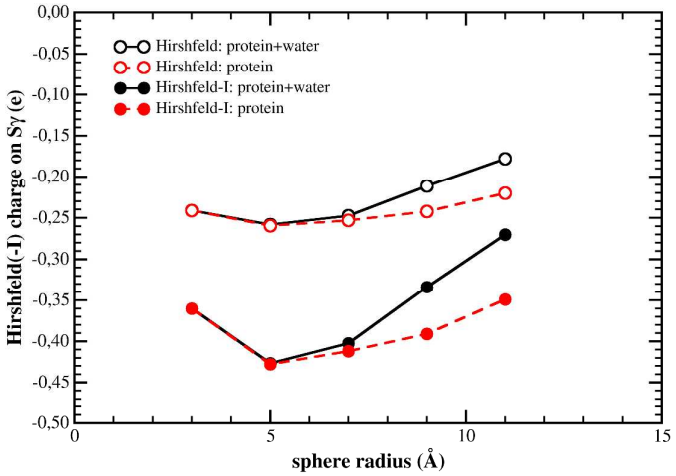

Figure2: Comparison of Hirshfeld and Hirshfeld-I charges of Sγ in the protein clusters (Figure 1b) in which Cys51 is present in the thiolate form as function of the sphere radius. Solid curves are for systems containing both a protein fraction and water molecules, while dashed lines indicate systems containing only the protein fraction.

Hirshfeld(-I) atomic charges of the Cys51 thiolate(S⁻) as function of the protein cluster size

Hirshfeld(-I) charges are calculated on the 3-11 Å protein spheres of Tpx-B (Figure 1B) with Cys51 present in the thiolate (deprotonated) form. Figure 3 shows the Hirshfeld-I charges obtained for the atoms of the Cys51 residue as function of the sphere radius (the Hirshfeld charges can be found in Figure S3 of the Supplementary Material). Already for a sphere of 5 Å, the atomic Hirshfeld(-I) charges are converged for all atoms but the Sγ atom. For this atom it is clear that even at 11Å, the obtained charge is not yet converged (see Figure 2 and yellow line in Figure 3 and S3). Instead, a continuous increase of the charge (the charge on Sγ is becoming less negative) with the size of the surrounding protein sphere is found. The presence of water molecules has only little influence on the charge of the atoms of the Cys51 residue, except for Sγ. The negative charge on the Sγ atom is significantly reduced by the presence of the water molecules (Figures 2, 3, and S3). In combination with the trend in the Sγ charge as function of the cluster size, this indicates that the Sγ charge is rather delocalized, allowing additional atoms of the larger clusters and the water molecules to distract part of the charge from the Sγ atom (Figures 2, 3b, and S3b). This is consistent with the results obtained for the CH_3S^- and CH_3S test systems embedded in small water clusters (*cf.* below).

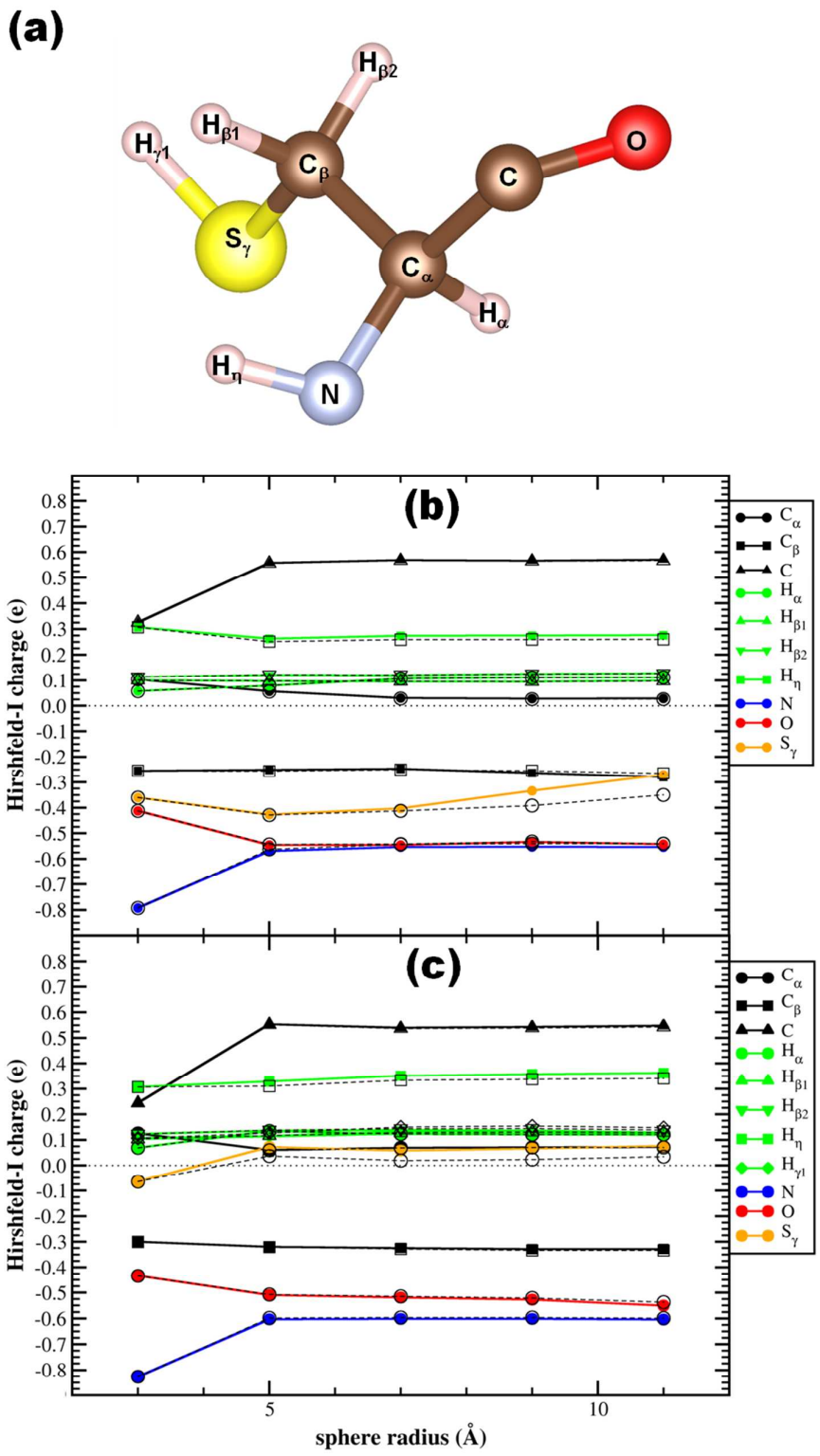

Figure 3: (a) Ball-and-stick model of the Cys51 residue in the neutral thiol form showing the atomic labels to identify the individual atoms; the same labeling is used for the thiolate form. Hirshfeld-I charges for all Cys51 atoms in the thiolate (b) and thiol (c) forms as function of the sphere radius, calculated using the PBE exchange-correlation functional**.** Solid curves indicate systems containing both the protein fraction and the water molecules, while the black dashed curves indicate systems containing the protein fraction only, without water molecules.

Hirshfeld(-I) atomic charges of the Cys51 thiol(SH) as function of the protein cluster size In addition to the Cys51 thiolate, also the charge convergence of the neutral Cys51 thiol is investigated as function of the protein cluster size. For the neutral thiol, the Hirshfeld(-I) charges for the atoms converge rapidly with the cluster's sphere size (Figure 3c, S3c). Unlike the thiolate case, the charge on the protonated Sγ atom in the Cys51 thiol converges already for a sphere size of 5 Å (Figure 3c, S3c, yellow line). The charges of the atoms of the neutral and deprotonated Cys51 residues are remarkably similar, with the obvious exception of the Sγ atom (Table 2). Note that for these calculations, due to the size of the system, and the limitations of present ab-initio codes, the accuracy of the atomic charges is estimated to be of the order of 0.01e [22].

Table 2 Hirshfeld-I charges (Q) for the 11Å clusters with Cys51 in the thiolate (S^-) and thiol (SH) form, calculated in the presence of water molecules. Δ represents the charge difference between thiol and thiolate.

	Q_{S-}	Q_{SH}	$\Delta (Q_{SH} - Q_{S-})$
C	0.57	0.55	-0.02
C_α	0.01	-0.06	-0.07
C_β	-0.34	-0.39	-0.04
H_α	0.13	0.14	+0.01
$H_{\beta1}$	0.11	0.14	+0.03
$H_{\beta2}$	0.15	0.15	0.00
H_η	0.28	0.38	+0.09
N	-0.56	-0.62	-0.06
O	-0.53	-0.54	-0.01
S_γ	-0.22	0.10	+0.32
$H_{\gamma1}$	--	0.13	(+0.13)
Total	-0.41	-0.02	+0.38

Charge calculated on the CH_3S^- test systems To check the validity of the trends obtained for the protein clusters, the density based Hirshfeld-I charges and the orbital based NPA charges are calculated on the CH_3S^- validation systems, surrounded by 0 up to 4 water molecules (Figure 1c and Table 3). For these systems, we find that the water molecules distract negative charge from the S atom of CH_3S^- to form a charged water cluster, similar to what happens in the protein clusters with Cys51 present as a thiolate (Figure 2). In these small systems, convergence of the charge of the S atom is reached when three water molecules are present as solvation shell. NPA and Hirshfeld-I charges show the same trends, which validates the Hirshfeld-I charges obtained on the protein systems.

Table 3 Hirshfeld-I (left) and NPA (right) charges calculated at the PBE/6-311++G(d,p) level for CH_3S^- surrounded by 0-4 water molecules. N^w indicates the number of water molecules. Q(S), Q(CH_3S) and Q(waters) indicates respectively the charge on the S atom, the charge on CH_3S and the charge on the water molecules. Q(Av. waters) indicates the average charge per water molecule.

	CH_3S^- (Hirshfeld-I)					CH_3S^- (NPA)				
N^w	0*	1*	2	3	4	0	1	2	3	4
Q(S)	(-0.64)	(-0.65)	-0.56	-0.44	-0.43	-0.75	-0.68	-0.64	-0.63	-0.63
Q(CH_3S)	(-0.75)	(-0.78)	-0.67	-0.54	-0.52	-1.00	-0.90	-0.83	-0.79	-0.78
Q(waters)	NA	(-0.18)	-0.33	-0.46	-0.48	NA	-0.10	-0.17	-0.21	-0.22
Q(Av. waters)	NA	(-0.18)	-0.17	-0.15	-0.12	NA	-0.10	-0.08	-0.07	-0.05

Footnotes to Table 3

* For the Hirshfeld-I calculations, the grid-based electron density was obtained from a periodic code using a plane wave basis set. For (negatively) charged systems this can lead to electron density which is not allocated on the molecular system [22]. This delocalization is not observed for (standard) atom centered Gaussian basis sets, since here, the electrons are artificially bound through the basis set. For the two systems indicated, the total integrated charge in the system (i.e. all electrons close to the CH_3S^- (+ water molecule) is less negative than should be expected as a result of this delocalization. For the system with 0 water molecules the total charge of the system is only -0.75e, while for the system with a single water molecule the total charge found is -0.96e, instead of the expected -1e. Despite this complication, the results further support the already present trend of decreasing negative charge with system size.

Discussion

According to quantum mechanics, atoms are smeared out and their charge is shared among nearby atoms. Since the atomic charge is not a physically measurable property, there is no unique way to assign electrons to atoms. Nevertheless, atomic charges are a useful concept. As such, atoms in molecules partitioning schemes are aimed at providing charges which are as transferable as possible and at the same time do not result in counterintuitive atomic charges. Hirshfeld-I charges have been shown to be very transferable, and provide atomic charges which are very reasonable in light of chemical intuition[19, 21, 22, 24, 42]. Here, in the presented results, the transferability is shown in two aspects: firstly, the fast convergence noticed of the Cys51 atomic charges with cluster size, and secondly, the calculated atomic charges of the equivalent atoms in the thiol and thiolate cluster are equal (within the given accuracy and with the obvious exception of the Cys51 S_γ atom). Furthermore, the charges produced by applying the Hirshfeld-I partitioning scheme on the protein clusters are in agreement with chemical intuition. An example of this is the charge obtained for the Cys51S_γ atom, which is negative when Cys51 is in the deprotonated thiolate form and ~0 when Cys51 is in the neutral thiol form.

The trends observed for the Hirshfeld(-I) charges calculated on the protein clusters is confirmed in small test systems (Table 3), both by the density based Hirshfeld-I and orbital based NPA partition schemes. Validation using the NPA charge on the protein clusters is more difficult, since these are computationally more expensive and could only be calculated using the moderate 6-31+G(d,p) basis set for the cluster having a 3 and 5 Å radius; for the larger clusters the small 3-21G basis set needs to be used (Figure S4, and section S2 in SI). While NPA and Hirshfeld(-I) charges show the same trends in the test systems, a discrepancy

is found in the charge convergence of the Sγ atom of the Cys51 thiolate in the protein clusters. The NPA charges calculated on the Sγ atom of the Cys51 thiolate converge with the cluster size, in contrast to the Hirshfeld(-I) charges (*cf.* section S1 in SI). Essentially, the difference is due to the fact that these are two different partitioning methods, and illustrates some arbitrariness in the determination of the atomic charge. NPA charges are calculated from the natural populations present in natural atomic orbitals (NAOs) centered on the atom of interest. To obtain the NAOs, the wavefunction is transformed into a localized form. As a result, the NPA partition scheme assigns charge to an atomic center based on the total electron density in the basis functions located at that center. The convergence of the NPA charge of the Sγ atom of the Cys51 thiolate is consistent with the convergence of the Mulliken charge (results not shown here) – which is also a partitioning scheme based on orbital occupancy. The Hirshfeld(-I) partition scheme, on the other hand, compares the electron density of a pro-molecule built from non-interacting atoms, with the density found in the actual molecule, resulting in a weighted partitioning of the density in each point in space over all atoms in the system. As such, no localized wave functions are involved and delocalized electrons are treated differently. In the orbital picture they are assigned to the atom providing the basis function while in the density picture the real space location of the density leads to this assignment.

NPA charges obtained using the very small 3-21G basis set are almost identical to the charges obtained with the 6-31+G(d,p) basis set in both the small (3 and 5 Å sphere) protein clusters and in the test systems (Table 3, Figure S4, section S1). Although our results show that the NPA charges are not very basis set dependent, the 3-21G basis set might not be large enough to correctly describe the anisotropic environment of the protein cluster, the loosely bound electrons of the thiolate anion and thus the long range electrostatics. While these effects might be less severe in the test systems and small protein clusters, these effects might be more pronounced in the large protein clusters, for which diffuse functions in the basis set are needed. Therefore, the less computationally demanding density based Hirshfeld(-I) charges constitute a very valuable alternative over orbital based charges as NPA or Mulliken, which are computationally more demanding and are proven not to meet the transferability or chemical intuition criteria all the time (for example, Mulliken charges are very basis set dependent).

The very quick convergence of the NPA charges and charges obtained with the Hirshfeld(-I) scheme, for all atoms but the Sγ atom of the Cys51 thiolate, shows that the atomic interactions are limited in range (< 5Å) . This is in line with a recent benchmark study [46] towards the convergence of calculated protein-ligand interaction energies E_{int}. This study pointed out that the correct ranking of E_{int} is already achieved for a cutoff distance of a sphere with 7Å radius around the ligand (note that the effective distance around the ligand is about 10-12 Å, since in the set up, all residues with at least one atom within the cutoff distance were kept completely). The sizable contributions beyond a distance of 7-10Å seem to be rather uniform and contribute similar in all cases. As such, according to this study and to our results obtained for the atomic charge, embedding models can be limited to relatively small regions that need to be tracked QM without losing predictive power.

For the Sγ atom of the Cys51 thiolate, the story is more complex. From the Hirshfeld(-I) calculations it is clear that its interaction is extremely long-ranged. This is shown by the clear non-convergence of the atomic charge for both the bare protein cluster, and the cluster including water molecules (Figure 3b). Here, the negatively charged thiolate is studied as a model, but most likely, these results can be extended to every negatively charged atom in a certain protein residue (for example, the oxygen atoms of Asp and Glu). The presence of water molecules significantly influences the charge of the Sγ atom of the Cys51 thiolate (Figure 2), by distract charge from the Sγ atom. This behavior could be reproduced by the test calculations on the CH_3S^- molecule surrounded by small water clusters (Table 3). From Figure 2 it is clear that this behavior is only present with the Sγ atom of the Cys51 thiolate, bearing the negative charge, and not for the other atoms of the Cys51 residue. Therefore, there is a clear distinction between the Sγ atom bearing the negative charge in the deprotonated form of Cys51 and the Sγ atom of the neutral Cys51 thiol (and by extension also all other atoms of both the Cys51 thiol and thiolate and all atoms of every neutral residue of a (protein) cluster). For all these atoms, the charge rapidly converges with the size of the cluster (Figure 3c) and solvation does not have a significant influence. This clearly shows that protonation blocks the interaction of the Sγ atom with the rest of the environment. In essence, the (in)activity of a centre is shown by the (non)convergence of its atomic Hirshfeld(-I) charge and by the influence of solvation. This observation, together with the discrepancy between different partitioning schemes, indicates some limitations for the construction of embedding models, since the long range charge dependence of negatively charged residues contrasts strongly with the use of small QM regions. This is in agreement with earlier studies showing that QM/MM energies have convergence problems, unless the QM-MM junctions are moved away from the active-site residues.[47] In addition, for electrostatic-only embedding potentials, the already known problem of spurious charge leakage from the QM region to the environment region (the so-called spill-out effect) is expected to become more pronounced, both due to the poor charge convergence and the long range charge dependence[48].

In light of the Hirshfeld(-I) results, protonation might be considered to take the screening effect from the environment to the limit. Table 2 shows that when Cys51 is present as a thiol, the Cys51 residue as a whole becomes roughly neutral, in line with the increasing (less negative) charge of the Cys51 thiolate with the system size. As such, when the surrounding system is large enough, the total charge on Cys51 Sγ will converge to a small value. The SH-group of the Cys51 thiol has a positive charge of 0.24e, in contrast to the negative charge of -0.22e for the Sγ atom in the Cys51 thiolate. This indicates that the hydrogen atom bound to the Sγ atom in the neutral Cys51 thiol pushes charge away from the sulfur atom, which leads to a slightly higher negative charge on the remaining atoms of the Cys51 thiol. As such the protonation of the Sγ atom deactivates the site by reallocating the former charge away from the Sγ.

This study is performed on a particular protein as model system, but might highlights some issues to consider for accurate modeling of (non-)charged residues in a multiscale approach in general. Furthermore, it highlights inconsistencies in charge partitioning schemes[49] of

negatively charged residues and as such, pinpoints extreme difficulties that might arise when modeling highly negatively charged systems *e.g.* nucleic acids.

Conclusions

In this work, the influence of the size of the enzymatic environment on the atomic charges has been investigated, with specific focus on the S_γ atom of the Cys51 residue in the model protein human 2-cysteine peroxiredoxin thioredoxin peroxidase B. We have shown that the behavior of the charge convergence of negatively charged residues depends on its protonation state and on the used partitioning scheme. The Hirshfeld(-I) scheme indicates that the S_γ atom of the negatively charged (deprotonated) Cys51 thiolate shows long-range interactions leading to the non-convergence of the atomic charge of this specific atom. The presence of the solvent environment (i.e. water molecules) even significantly lowers its negative charge (i. e. the charge becomes more positive). In contrast, for atoms which do not bear the negative charge, the atomic charge converges fairly quickly (interaction radius < 5 Å) and the presence of water molecules has little to no influence. This behavior is also seen for all atoms, including S_γ of the Cys51 thiolate, when the atomic charges are calculated with the NPA partitioning scheme. This discrepancy between different population analysis schemes together with the non-convergence of atomic charges complicates the construction of accurate embedding models. We have thus shown that protonation is important in the behavior of the calculated charge in the model protein Tpx-B from first principles results. We expect that these results can be generalized, highlighting problematic issues for accurate modeling of negatively charged residues in a multiscale approach.

Acknowledgement

The authors acknowledge financial the support from the Research Board of the Ghent University (BOF)Calculations were carried out using the Stevin Supercomputer Infrastructure at Ghent University. DEPV and GR thank the Research Foundation Flanders (FWO) for postdoctoral fellowships. JO acknowledges the receipt of a Bolyai János Research Fellowship. FDP wishes to acknowledge the Research Foundation-Flanders (FWO) and the Vrije Universiteit Brussel (VUB) for their financial support, especially mentioning the Strategic Research Program awarded to the ALGC group by the VUB which started on January 1, 2013

ASSOCIATED CONTENT

Supporting Information Available: Section S1-S6 including extra information and Figure S1-S4, Table S1-S4. This material is available free of charge via the Internet at http://pubs.acs.org.

References

1. Vanommeslaeghe, K.; Raman, E. P.; MacKerell, A. D., Jr., Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **2012**, 52, 3155-68.
2. Gross, K. C.; Seybold, P. G.; Peralta-Inga, Z.; Murray, J.; Politzer, P., Comparison of Quantum Chemical Parameters and Hammett Constants in Correlating pKa Values of Substituted Anilines. *J. Org. Chem.* **2001**, 66, 6919-6925.
3. Roos, G.; Loverix, S.; Geerlings, P., Origin of the pKa Perturbation of N-terminal Cysteine in alpha- and 3(10)-Helices: a Computational DFT Study. *J. Phys. Chem. B* **2006**, 110, 557-562.
4. Roos, G.; Geerlings, P.; Messens, J., Enzymatic Catalysis: The Emerging Role of Conceptual Density Functional Theory. *J. Phys. Chem. B* **2009**, 113, 13465-13475.
5. Roos, G.; Foloppe, N.; Van Laer, K.; Wyns, L.; Nilsson, L.; Geerlings, P.; Messens, J., How Thioredoxin Dissociates its Mixed Disulfide. *PLOS Comput. Biol.* **2009**, 5, e1000461.
6. Ugur, I.; Marion, A.; Parant, S.; Jensen, J. H.; Monard, G., Rationalization of the pKa Values of Alcohols and Thiols Using Atomic Charge Descriptors and Its Application to the Prediction of Aminoacid pKa's. *J. Chem. Inf. Model.* **2014**, 54 2200-2213.
7. Geerlings, P.; De Proft, F.; Langenaeker, W., Conceptual Density Functional Theory. *Chem. Rev.* **2003**, 103, 1793-873.
8. Mulliken, R. S., Electronic Population Analysis on LCAO-MO Molecular Wave Functions. II. *J. Chem. Phys.* **1955**, 23, 1841-1846.
9. Mulliken, R. S., Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, 23, 1833-1840.
10. Mulliken, R. S., Electronic Population Analysis on LCAO-MO Molecular Wave Functions. III. *J. Chem. Phys.* **1955**, 23, 2338-2342.
11. Mulliken, R. S., Electronic Population Analysis on LCAO-MO Molecular Wave Functions. IV. *J. Chem. Phys.* **1955**, 23, 2343-2346.
12. Reed, A. E.; Weinstock, R. B.; Weinhold, F., Natural Population Analysis. *J. Chem. Phys.* **1985**, 83, 735-746.
13. Wick, C. R.; Hennemann, M.; Stewart, J. J. P.; Clark, T., Self-consistent Field Convergence for Proteins: a Comparison of Full and Localized-Molecular-Orbital Schemes. *J. Mol. Model.* **2014**, 20, 2159.
14. Stewart, J. J. P., Application of the PM6 Method to Modeling Proteins. *J. Mol. Model.* **2009**, 15, 765-805.
15. Gaus, M.; Goez, A.; M. Elstner, M., Parametrization and Benchmark of DFTB3 for Organic Molecules. *J. Chem. Theory Comput.* **2013**, 9, 338-354.
16. Antony, J.; Grimme, S., Fully Ab Initio Protein-Ligand Interaction Energies with Dispersion Corrected Density Functional Theory. *J. Comput. Chem.* **2012**, 33, 1730-1739.
17. Lee, L. P.; Cole, D. J.; Payne, M. C.; Skylaris, C.-K., Natural Bond Orbital Analysis in the Onetep Code: Applications to Large Protein Systems. *J. Comput. Chem.* **2013**, 34, 429-444.
18. Dunnington, B. D.; Schmidt, J. R., Generalization of Natural Bond Orbital Analysis to Periodic Systems: Applications to Solids and Surfaces via Plane-Wave Density Functional Theory. *J. Chem. Theory Comput.* **2012**, 8, 1902-1911.
19. Bultinck, P., Critical Analysis of the Local Aromaticity Concept in Polyaromatic Hydrocarbons. *Faraday Discuss.* **2007**, 135, 347-365.
20. Bader, R. F. W., *Atoms in Molecules: A Quantum Theory*. Oxford University Press: Oxford, 1990.

21. Vanpoucke, D. E. P.; Van Driesche, I.; Bultinck, P., Reply to 'Comment on "Extending Hirshfeld-I to Bulk and Periodic Materials" *J. Comput. Chem.* **2013**, 34, 422-427.
22. Vanpoucke, D. E. P.; Bultinck, P.; Van Driesche, I., Extending Hirshfeld-I to Bulk and Periodic Materials. *J. Comput. Chem.* **2013**, 34, 405-417.
23. Vanpoucke, D. E. P.; Cottenier, S.; Van Speybroeck, V.; Van Driessche, I.; Bultinck, P., Tetravalent Doping of CeO2: The Impact of Valence Electron Character on Group IV Dopant Influence. *J. Am. Ceram. Soc.* **2014**, 97, 258-266.
24. Verstraelen, T.; Pauwels, E.; De Proft, F.; Van Speybrouck, V.; Geerlings, P.; Waroquier, M., Assessment of Atomic Charge Models for Gas Phase Computations on Polypeptides. *J. Chem. Theory Comput.* **2011**, 8, 661-676.
25. Schroder, E.; Littlechild, J. A.; Lebedev, A. A.; Errington, N.; Vagin, A. A.; Isupov, M. N., Crystal Structure of Decameric 2-Cys Peroxiredoxin from Human Erythrocytes at 1.7 angstrom Resolution. *Struct. Fold. Des.* **2000**, 605-615.
26. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Could, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A Second Generation Force Field for Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, 117, 5179-5197.
27. Olsson, M. H.; Søndergard, C. R.; Rostkowski, M.; Jensen, J. H., PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa predictions. *J. Chem. Theory Comput.* **2011**, 7, 525-537.
28. MacKerell, A. D., Jr.; Banavali, N.; Foloppe, N., Development and Current Status of the CHARMM Force Field for Nucleic Acids. *Biopolymers* **2000**, 56, 257-65.
29. Olah, J.; van Bergen, L.; De Proft, F.; Roos, G., How does the protein environment optimize the thermodynamics of thiol sulfenylation? Insights from model systems to QM/MM calculations on human 2-Cys peroxiredoxin. *J. Biomol. Struct. & Dyn.* **2014**, 33, 584-596.
30. Brooks, B. R.; Brooks, C. L., 3rd; Mackerell, A. D., Jr.; 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., CHARMM: the Biomolecular Simulation Program. *J. Comp. Chem.* **2009**, 30, 1545-614.
31. Foloppe, N.; Sagemark, J.; Nordstrand, K.; Berndt, K. D.; Nilsson, L., Structure, Dynamics and Electrostatics of the Active Site of Glutaredoxin 3 from Escherichia coli: Comparison with Functionally Related Proteins. *J. Mol. Biol.* **2001**, 310, 449-70.
32. Foloppe, N.; Nilsson, L., The glutaredoxin -C-P-Y-C- Motif: Influence of Peripheral Residues. *Structure* **2004**, 12, 289-300.
33. Brooks III, C. L.; Karplus, M., Deformable Stochastic Boundaries in Molecular Dynamics. *J. Chem. Phys.* **1983**, 79, 6312.
34. van der Kamp, M. W. PhD thesis, Modelling Reactions and Dynamics of Claisen Enzymes, Chapter 4.6. University Bristol, 2008.
35. Blöchl, P. E., Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, 50:24, 17953-17979.
36. Kresse, G.; Joubert, D., From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, 59:3, 1758-1775.
37. Kresse, G.; Hafner, J., Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, 47:1, 558-561.
38. Kresse, G.; Furthmüller, J., Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **1996**, 54, 11169-11186.
39. Ceperley, D. M.; Alder, B. J., Ground State of the Electron Gas by a Stochastic Method. *Phys. Rev. Lett.* **1980**, 45:7, 566-569.

40. Perdew, J. P.; Burke, K.; Ernzerhof, M., Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, 77, 3865-3868.
41. Monkhorst, H. J.; Pack, J. D., Special Points for Brillouin-Zone Integrations. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1976**, 13, 5188-5192.
42. Vanpoucke, D. E. P., Hive, version 2.1, http://users.ugent.be/~devpouck/. **2011**.
43. Lebedev, V. I.; Laikov, D. N., Quadrature Formula for the Sphere of 131th Algebraic Order of Accuracy. *Dokl. Akad. Nauk* **1999**, 366, 741-745.
44. Becke, A. D., A Multicenter Numerical-Integration Scheme for Polyatomic-Molecules. *J. Chem. Phys.* **1988**, 88, 2547-2553.
45. M. J. Frisch, G. W. T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, In; Gaussian, Inc.: Wallingford CT, 2009.
46. Yilmazer, N. D.; Korth, M., Comparison of Molecular Mechanics, Semi-Empirical Quantum Mechanical, and Density Functional Theory Methods for Scoring Protein-Ligand Interactions. *J. Phys. Chem. B* **2013**, 117, 8075-8084.
47. Hu, L.; Söderhjelm, P.; Ryde, U., On the Convergence of QM/MM Energies. *J. Chem. Theory Comput.* **2011**, 7, 761-777.
48. Laio, A.; Van de Vondele, J.; Rothlisberger, U., A Hamiltonian Electrostatic Coupling Scheme for Hybrid Car-Parrinello Molecular Dynamics Simulations. *J. Chem. Phys.* **2002**, 116, 6941-6947.
49. Henriques, J.; Costa, P. J.; Calhorda, M. J.; Machuqueiro, M., Charge Parametrization of the DvH-c3 Heme Group: Validation using Constant-(pH,E) Molecular Dynamics Simulations. *J. Phys. Chem. B* **2013**, 117, 70-82.

TOC graphics