Molecular electrostatic potentials by systematic molecular fragmentation

David M. Reid and Michael A. Collins
Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia

(Received 26 July 2013; accepted 7 October 2013; published online 14 November 2013)

A simple method is presented for estimating the molecular electrostatic potential in and around molecules using systematic molecular fragmentation. This approach estimates the potential directly from the electron density. The accuracy of the method is established for a set of organic molecules and ions. The utility of the approach is demonstrated by estimating the binding energy of a water molecule in an internal cavity in the protein ubiquitin. © 2013 AIP Publishing LLC.

[http://dx.doi.org/10.1063/1.4827020]

I. INTRODUCTION

The molecular electrostatic potential (ESP), $\Phi$, has been used extensively to characterise molecular interactions in systems of both chemical and biological importance. Perhaps its earliest application was in the prediction of electrophilic and nucleophilic additions and substitutions, but it has since been applied to the modelling of hydrogen bond geometries, $pK_a$ values, and catalysis by solvent, semiconductors, and enzymes. In addition, the dominant long range interactions that direct substrates to enzyme active sites are electrostatic in nature, and can therefore be modelled through accurate predictions of the potential. Thus for proteins in particular, the ESP is an invaluable tool in the modelling of interactions, with considerable medicinal and industrial applications.

$Ab$ $initio$ quantum chemistry computation time increases rapidly with increasing number of electrons. Until very recently, this has prohibited the direct calculation of energies and electrostatic potentials for systems any larger than a few hundred atoms. Hence, the ESP of large biological molecules was usually modelled using empirical charge distributions. For this reason, there has been considerable interest in recent years in a number of different approaches to approximating the ESP and other properties (including the energy) of large systems at reduced computational cost. These methods include the effective fragment potential method, fragment molecular orbital approaches, and energy-based fragmentation methods. The later involves breaking the molecule into fragments, evaluating the energy or property of the relatively small fragments, and then recombining the fragment energies or properties to estimate the total molecular values. Some methods have been applied to very large molecules. The computational time for these fragmentation methods scales linearly with the size of the molecule.

The first application of fragmentation methods to estimate the ESP was reported by Gadre and co-workers using the “molecular tailoring approach” (MTA). One of the first applications of the “molecular fragmentation by conjugate caps” (MFCC) approach of Zhang and co-workers was to the electron density and ESP of proteins. The electron density in some part of a large molecule is polarized by the charge distribution in other parts. In their “generalised energy-based fragmentation” (GEBF) method, Li and co-workers first accounted for this effect by performing the $ab$ $initio$ calculations of molecular fragments in the presence of “embedded charges” that represented the rest of the molecules. Various other fragmentation methods have been specifically applied to the evaluation of molecular electrostatic potentials. One purpose of recent work has been to develop $ab$ $initio$ based distributed charges and higher multipole moments that more accurately represent the molecular electrostatic potential than do empirical distributed charges. Substantial progress has been made towards developing charge distributions that accurately describe the ESP of proteins.

In this paper, we develop a related approach to obtaining the electrostatic potentials of general molecules using the systematic molecular fragmentation method. The total ESP is given directly by a sum of the ESP for relatively small molecular fragments. Following Le, Lee, and Bettens, the sum of contribution to the ESP is calculated on a three-dimensional grid of positions encompassing the molecule, using the electron density distribution for each fragment. The ESP may also be approximated using a multipole expansion of the fragment charge distributions. There are two advantages to a grid approach: It is simple to implement using readily available $ab$ $initio$ quantum chemistry packages, and it is valid in small cavities “within” molecules, where the multipole expansion may be unreliable. The accuracy of this approach is demonstrated by comparison of the approximate ESP with that evaluated directly by $ab$ $initio$ calculation on the whole molecule. These comparisons are made for a range of moderate-size molecules, for which $ab$ $initio$ calculations are feasible. The method is then applied to the estimation of the ESP for a structure of the protein ubiquitin. This is a 76 residue protein that is well conserved amongst all eukaryotic organisms from humans to yeast. Its predominant function is as a tag that is added to mark proteins for degradation. Though ubiquitin is for the most part tightly folded, a small cavity exists in the interior of the protein that is surrounded by hydrophobic side chains. This cavity measures several angstrom in diameter, making it large enough to accommodate a single molecule of water, though NMR experiments have as yet found no evidence of its presence. The interest in characterising the occupancy (or non-occupancy) of such
cavities stems from both the structural and enzymatic importance of bound water, and in recent years the subject has been strongly controversial. Studies conducted for similar cavities using both NMR and x-ray crystallography have found evidence for the presence of disordered (i.e., poorly defined spatially) water, though other groups have suggested alternative explanations that fit the data. Here we use ubiquitin as an example of how the ESP can be readily calculated by fragmentation and used to estimate the electrostatic energy of a water molecule in the cavity of interest. The utility of a fragmentation-based estimation of the ESP is made apparent in this application, whereby the energy of millions of possible locations for the water molecule can be estimated at negligible computational cost.

The paper is set out as follows: Sec. II presents the methods employed in this approach. The accuracy of the method is demonstrated in Sec. III for a selection of molecules, while Sec. IV presents results for the energy of a water molecule in the cavity of ubiquitin. Brief concluding remarks are contained in Sec. V.

II. METHODOLOGY

A. Fragmentation

The electrostatic potential of a molecule is estimated via the systematic molecular fragmentation by annihilation (SMFA) approximation. This approach is a variant on the earlier systematic molecular fragmentation (SMF) method, both of which have been described in detail previously. In simple terms, a molecule is described as collection of functional groups which are connected by single bonds. In SMFA, a molecule is decomposed into fragments by removing functional groups, in a sequence of steps that preserve the bonding environment of each group to some extent. The method has a systematic set of “Levels” which determine the proximity of eliminated groups, so that with increasing Level, a more extensive bonding environment is retained. The method is automated and applicable to any molecular structure, but the simple example of a chain-like structure is illustrative. For a chain of five groups, \( G_1G_2G_3G_4G_5 \), the molecule is decomposed as follows:

\[
G_1G_2G_3G_4G_5 \rightarrow G_1G_2 + G_2G_3 + G_3G_4 + G_4G_5 - G_2 - G_3 - G_4, \quad \text{(Level 1)}
\]

\[
\rightarrow G_1G_2G_3 + G_2G_3G_4 + G_3G_4G_5 - G_2G_3 - G_3G_4, \quad \text{(Level 2)}
\]

\[
\rightarrow G_1G_2G_3G_4 + G_2G_3G_4G_5 - G_2G_4, \quad \text{(Level 3), (2.1)}
\]

and so on for higher Levels. At Level 1, the interaction of each group with its \( \alpha \) substituents is included in the fragments. At levels, \( \beta \) substituents are included, and so on. For any general molecule, we can write the fragmentation of \( M \) as

\[
M \rightarrow \sum_{n=1}^{N_{\text{frag}}} c_n F_n, \quad \text{(2.2)}
\]

where the \( F_n \) represent “overlapping” fragments of the molecule, and the \( c_n \) are integers. When groups are eliminated in the fragmentation procedure, the remaining groups have unsatisfied valency. The normal valency of each atom is restored by appending hydrogen atoms along the original bond direction, as previously described. These are referred to as “hydrogen caps.”

We denote Eq. (2.2), as fragmentation at a given “Level.” As the value of Level increases, the fragments increase in size, and the corresponding estimate of a property, \( P \),

\[
P = \sum_{n=1}^{N_{\text{frag}}} c_n P(F_n), \quad \text{(2.3)}
\]

increases in reliability. In previous applications, the molecular total electronic energy has been the property of most interest, and for moderate-sized molecules, the energy can be estimated by Level 3 fragmentation to about 2 \( \text{kJ mol}^{-1} \), however, other properties, such as chemical shifts and molecular dipole moments, have also been estimated in this way by Zhang and co-workers, and by other groups.

The reliability of Eq. (2.3) is verified by systematically increasing the value of Level to obtain a systematic sequence of estimates for \( P \), which demonstrate convergence to some value.

1. Definition of bonding

The composition of the fragments in Eq. (2.2) depends on the Level of fragmentation and on the definition of bonds between atoms. As in previous applications, two atoms are connected by a single bond if the distance between those atoms is less than the sum of the covalent radii of the atoms (plus a small tolerance). Much shorter bonds are defined to be multiple bonds, so that such atoms are defined to be contained in the same functional group (for example, a carbonyl group). One can arbitrarily define some bonds as multiple bonds (for example, the CN bond in an amide group). In addition, one can define a hydrogen bond as a single bond connecting the donor and acceptor “heavy” atoms. Results will be presented below which indicate the effect of hydrogen bonding on the estimate of the molecular potential for peptides.

In this paper, we are concerned with the electrostatic potential, \( \Phi \), of a molecule at any arbitrary position in space,
denoted by the Cartesian coordinate vector \( \mathbf{x} \). Hence, we will evaluate
\[
\Phi(\mathbf{x}) = \sum_{n=1}^{N_{frag}} c_n \Phi(F_n(\mathbf{x})),
\]
(2.4)
for any given Level of fragmentation.

### B. Embedded charges

The ansatz leading to Eq. (2.3) implicitly assumed that interactions between functional groups in a molecule occur mainly via a “through-bond” mechanism: groups which are close in terms of bonded connectivity interact more strongly than groups which are separated by a longer sequence of bonds. However, “through-space” interactions are not negligible. Such interactions include electrostatic interactions, associated induction, and dispersion.\(^{41}\) In molecules, such as proteins, where some groups are formally charged, electrostatic interactions and induction may be very significant. Induction is a many-body effect; the polarisation of the electron density of a group is due to the net electric field at the group, and this field has contributions from the charge distributions of all other groups in the molecule. Hence, calculations of any property, based on molecular fragments, should account for the electric field enclosing those fragments.\(^{16,23,26,42,43}\) The utility of using embedded charges to model this electric field was demonstrated by Dahlke and Truhlar for water clusters,\(^{42}\) and by Li and co-workers\(^{26}\) for molecular fragmentation.

#### 1. Estimation of the potential

The electric field produced by a molecule (or a group) is due to its charge distribution (due to the electrons and nuclei). This molecular charge distribution can be described in terms of charge distributions on the atomic positions. These charges can be estimated in several ways, for example, using Mulliken charges, Natural Population Analysis (NPA) charges, or Stone’s distributed Multipole Analysis (DMA)\(^{44}\) of the \textit{ab initio} electronic wavefunction. Here, we have tested both NPA and DMA approaches. The charge distribution on each atom in the molecule is estimated via the fragmentation approximation, using Eq. (2.3), as reported previously.\(^{19}\) Briefly, if \( q(k) \) represents a charge, or multipole moment on atom \( k \), and \( q(k,n) \) represents the corresponding charge or moment on atom \( k \) in fragment \( n \), then Eq. (2.3) gives
\[
q(k) = \sum_{n=1}^{N_{frag}} c_n q(k, n). \tag{2.5}
\]

Equation (2.5) involves performing an \textit{ab initio} calculation on each fragment, \( F_n \), evaluating the charge or moment on each atom in the fragment, and then summing the fragment values [where \( q(k,n) \) is zero if atom \( k \) is not contained in fragment \( n \)]. Now, the \( q(k,n) \) in Eq. (2.5) is evaluated from an \textit{ab initio} calculation on each fragment \textit{in vacuo}. However, according to Eq. (2.5), all of the atoms have an associated charge distribution. Hence, the \textit{ab initio} calculation on each fragment, \( F_n \), ought to be evaluated in the electrostatic field of the charge distributions of atoms not contained in \( F_n \). Thus, Eq. (2.5) is replaced by
\[
q(k) = \sum_{n=1}^{N_{frag}} c_n q(k, n; \{ q(j), j \notin F_n \}). \tag{2.6}
\]

where \( q(k, n; \{ q(j), j \notin F_n \}) \) denotes the charge or moment on atom \( k \) in fragment \( n \), evaluated with \( F_n \) embedded in the field due to the charges and moments on all atoms, \( j \), which are not contained in \( F_n \). Equation (2.6) is evaluated iteratively to self-consistency: All atom-based charge distributions are evaluated by Eq. (2.5) (\textit{in vacuo}), then the \textit{ab initio} fragment calculations are repeated with the fragment embedded in the atom-based charge distributions to give new charges and moments on all atoms by Eq. (2.6), then the \textit{ab initio} fragment calculations are repeated using the new charges and moments, and so on, until the values produced by Eq. (2.6) do not change to within some tolerance. In this way, the self-consistent polarisation of each group in a molecule by every other group is accounted for.\(^{19}\)

Once, the self-consistent charge distributions have been obtained, any molecular property can then be evaluated within this embedded-charge approach. Thus, Eq. (2.3) is replaced by
\[
P = \sum_{n=1}^{N_{frag}} c_n P(F_n; \{ q(j), j \notin F_n \}), \tag{2.7}
\]
where each fragment is embedded in the charge distribution of all atoms that are not contained in that fragment. To apply Eq. (2.7) to the molecular energy, a correction for “double counting” of interactions between the embedded charges must be included.\(^{26}\)

#### 2. Computational details

Each hydrogen cap occurs in a subset of the fragments in Eq. (2.2) with coefficients, \( c_n \), which sum to zero (as these H atoms do not exist in \( M \)). Ideally, the corresponding sum in Eq. (2.6) for the charge distributions on these hydrogen caps should sum to zero. However, in practice the caps retain some residual charge, and for low levels of fragmentation, such charges can be of the order of \( 10^{-3} \) to \( 10^{-2} \) a.u. To remove these small spurious cap moments from the charge distribution, while preserving the correct overall charge, the charge on each cap is added to the charge on the atom to which it is bonded (and the cap charge is set to zero), while higher order moments (dipoles, quadrupoles, etc.) on the caps are set to zero.

For molecules that contain no formally charged groups, we have found that NPA charges given by Eq. (2.5) do not change significantly with iteration. Previous studies\(^{19,23}\) using DMA charge distributions as an embedding potential have shown that to obtain a sufficiently accurate representation of the electric potential requires the use of multipoles up to second order (charge, dipole, and quadrupole) distributed on each atomic site. The \textit{ab initio} calculations reported herein were performed with the GAUSSIAN03 program package.\(^{45}\) In this case, it is necessary to represent embedded point dipoles and quadrupoles by a cluster of point charges located very close...
to the atomic site. The location of these charges, and their relation to the dipole and quadrupole moments are contained in Table S1 of the supplementary material.46

In an ab initio calculation for any fragment, capping hydrogen atoms are close to charges and multipoles located on atoms that are bonded to the fragment in the whole molecule (typically at a distance of about 0.5 Å). The question arises as to whether such closely adjacent charges accurately represent the appropriate embedding potential. Wang and Truhlar, in the context of quantum mechanics (QM)/molecular mechanics (MM) calculations, have moved such adjacent charges further from the molecule which is treated ab initio.47 Section III presents the variation of the calculated molecular ESP with the position of these closely adjacent charges. If, \( \mathbf{X}_H \) represents the position of a hydrogen cap, and \( \mathbf{X}_C \) represents the position of the adjacent charge, then the charge is moved to an alternative position, \( \mathbf{X}_C' \), given by

\[
\mathbf{X}_C' = \mathbf{X}_C + \lambda (\mathbf{X}_C - \mathbf{X}_H). \quad (2.8)
\]

The parameter \( \lambda \) determines the magnitude of the displacement of the adjacent charge.

\section*{C. Test cases}

The accuracy of a fragmentation estimate of a molecular ESP is reported here for a set of medium-sized neutral organic molecules and ions. These molecular structures have been chosen from the Cambridge Structural Database to include a range of organic functional groups, eight neutral molecules (including peptides that contain hydrogen bonds, denoted as Set 1), and eight ions and zwitterions (denoted as Set 2). These 16 structures contain from 27 to 110 atoms, with an average of 68. The coordinates of all these structures are contained in Table S2 of the supplementary material.46 These molecules are sufficiently small that it was feasible to perform ab initio calculations for each entire molecule at the Hartree-Fock (HF) level of theory with the 6-31G, 6-31G(d,p) and 6-31++G(d,p) Pople-type basis sets.48–50 Ab initio calculations for the whole molecules were also carried out using the Møller-Plesset perturbation theory (MP2)51 method with the 6-31G and 6-31G(d,p) basis sets.

For each molecule in this set, the electronic structure has been evaluated exactly at the above levels of ab initio theory (that is, calculated for the whole molecule) and evaluated via the fragmentation approximation. Thus the error in the fragmentation estimate of the ESP can be directly calculated for these cases.

\section*{D. The electrostatic potential}

The ESP at each grid point was evaluated using the CUBEGEN facility of the GAUSSIAN03 program.45 The same grid was used for the whole molecule and for each and every fragment, so that Eq. (2.13) was evaluated for each grid point.

\section*{III. TEST RESULTS}

Table I presents a summary of the errors in the calculated ESP for the neutral test molecules of Set 1. The corresponding errors for the eight individual molecules are shown in Table S3 of the supplementary material.46

When no background charges are employed, Table I shows that the rms error in the ESP declines with increasing level of fragmentation to just under 0.002 a.u. To put this in perspective, 0.002 a.u. is the potential due to a unit charge at a distance of about 250 Å. The error is further significantly reduced if hydrogen bonds are included in the bonding definition, so that such interactions are included within fragments. However, for many of these molecules, hydrogen bonds significantly increase the size (and computational cost) of the fragment ab initio calculations. For these neutral molecules, background charges (where employed) have been modelled as simple NPA point charges at the nuclear positions. Iteration of the NPA charges, according to Eq. (2.15) was found to lead to negligible changes in these charges. Table I shows that inclusion of background charges reduces the rms error to 0.001–0.002 a.u. The rms error depends relatively weakly on the location of the closely adjacent charges, as determined by different values of the \( \lambda \) parameter, but completely removing these adjacent charges leads to a larger rms error. Each row of Table I indicates that the accuracy of the calculated ESP improves rapidly with increasing level of fragmentation.

Table II presents a summary of the errors in the calculated ESP for the eight molecules that contain formal charges (Set 2). The corresponding errors for the eight individual molecules are shown in Table S4 of the supplementary material.46 Comparing the first two rows of Table II (corresponding to no background charges) with Table I clearly shows that the use of background charges is essential for an accurate estimate of the ESP for these ions and zwitterions.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Calculation & Level 1 & Level 2 & Level 3 \\
\hline
Basic & 4.40 & 2.37 & 1.87 \\
H-bonding & 4.33 & 1.71 & 0.83 \\
Delete & 10.2 & 5.36 & 4.73 \\
Kept & 9.25 & 3.63 & 2.13 \\
\hline
\hline
\end{tabular}
\caption{The average rms error (multiplied by 10^3) in the ESP (au) for the eight neutral molecules (Set 1) is shown for several related calculations. Calculations were performed in the absence of background charges, with direct treatment of hydrogen bonding (H-bonding) or without (Basic). Calculations were also performed with background charges, where junction charges are deleted (Delete), retained (Kept) or moved using values of \( \lambda \) in Eq. (2.8). All calculations were performed for the HF/6-31G model chemistry.}
\end{table}
TABLE II. The average rms error (multiplied by 10^3) in the ESP (au) for the eight molecules which contain formal charges (Set 2) are shown for several related calculations. Calculations were performed in the absence of background charges, with direct treatment of hydrogen bonding (H-bonding) or without (Basic). Calculations were also performed with background charges, where junction charges are deleted (Delete), retained (Kept) or moved using values of λ in Eq. (2.8). All calculations were performed for the HF/6-31G model chemistry.

<table>
<thead>
<tr>
<th>Calculation</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic</td>
<td>9.19</td>
<td>8.27</td>
<td>6.42</td>
</tr>
<tr>
<td>H-bonding</td>
<td>7.23</td>
<td>5.57</td>
<td>3.12</td>
</tr>
<tr>
<td>Delete</td>
<td>4.31</td>
<td>2.57</td>
<td>2.15</td>
</tr>
<tr>
<td>Kept</td>
<td>3.43</td>
<td>2.24</td>
<td>1.90</td>
</tr>
<tr>
<td>λ = 2.5</td>
<td>3.38</td>
<td>2.22</td>
<td>1.86</td>
</tr>
<tr>
<td>λ = 3.0</td>
<td>3.44</td>
<td>2.22</td>
<td>1.86</td>
</tr>
<tr>
<td>λ = 3.5</td>
<td>3.47</td>
<td>2.21</td>
<td>1.86</td>
</tr>
<tr>
<td>λ = 3.0 iterated once</td>
<td>3.08</td>
<td>1.89</td>
<td>1.52</td>
</tr>
<tr>
<td>λ = 3.0 iterated twice</td>
<td>3.07</td>
<td>1.87</td>
<td>1.48</td>
</tr>
</tbody>
</table>

Given the significant effect of the imposed electric field in such cases, the question arises as to how accurately the field should be estimated for each fragment \textit{ab initio} calculation. To investigate this question, the charge distribution surrounding each fragment has been estimated in two ways, using NPA point charges at the atomic sites or using multipoles up to second order (charge, dipole, and quadrupole) distributed on each atomic site, calculated using Stone’s distributed multipole analysis (DMA) method.\textsuperscript{44} Perhaps somewhat surprisingly, the rms error in the ESP was found to be little changed when NPA point charges were replaced by distributed multipoles. However, the maximum error in the ESP over the grid was found to be somewhat (about 20%) smaller using distributed multipoles rather than NPA point charges. Hence, the errors reported in Table II (after the first two rows) are those obtained using distributed multipoles. As for the neutral molecules, Table II shows that the rms error is only weakly dependent on the location of the charges adjacent to the fragments, so long as these charges are not completely removed. For a range of values of the λ parameter, the rms error in the ESP lies below 0.002 a.u. for Level 3 fragmentation. This table also indicates that, for these molecules containing charges, iteration of Eq. (2.6) for one or two iterations does improve the accuracy of the ESP (though further iterations have negligible effect).

Tables I and II indicate that the ESP is accurately estimated on average using systematic molecular fragmentation, including the use of embedded charges for molecules that contain formally charged groups. There are, of course, a distribution of errors. There is a strong correlation between the absolute error at some location and the distance of that location to the nearest atom in the molecule. Not surprisingly, the absolute error is higher for grid points closer to nuclei. However, it is important to realise that the relative error is not very high, even when the ESP is high. Table S5 of the supplementary material presents the maximum error in the ESP over the grid for the eight neutral molecules. Figure 2 shows, for each grid point using distributed multipoles and λ = 3, the estimated ESP versus the exact value for SUNXUU, corresponding to Fig. 1. This figure clearly shows that the relative error in the ESP is large only when the value of the ESP is relatively negligible. Hence, the ESP surface for the molecule is quantitatively and qualitatively described by this approach.

IV. UBIQUITIN

As discussed in the Introduction, a small cavity exists in the interior of ubiquitin that is surrounded by hydrophobic side chains, as illustrated in Figure 3. This cavity is

FIG. 1. For each grid point in the cube enclosing the SUNXUU molecule, the figure shows the absolute error in the ESP, for Level 3 fragmentation and λ = 3, versus the minimum distance (d) of the grid point to an atom in the molecule, divided by the van der Waals (VdW) radius of the atom. Calculations were performed at the HF/6-31G level of \textit{ab initio} theory.

FIG. 2. For each grid point in the cube enclosing the SUNXUU molecule, the figure shows the fragmentation (Level 3 fragmentation with λ = 3) estimate of the ESP versus the exact HF/6-31G value.
large enough to accommodate a single molecule of water. The general interest in characterising the occupancy (or non-occupancy) of such cavities stems from both the structural and enzymatic importance of bound water. Part of the difficulty in verifying (or disproving) the presence of water is that disordered molecules contribute little to the electron density determined in x-ray crystallography, particularly at high resolution. It has been argued that the lack of H-bonding and the limited spatial extent of the cavity in ubiquitin ensure that occupancy is both energetically and entropically unfavourable. In addition, thermodynamic studies have concluded that stable occupation only occurs for considerably larger hydrophobic cavities, capable of accommodating multiple, mutually H-bonded water molecules.

Quantum mechanical predictions of the interaction energy of water with ubiquitin may shed some additional light on the subject. A practical difficulty relates to the positioning and orientation of the water molecule. A thorough study would require calculations with water in a large number of locations in the cavity. Even with fragmentation, this would be time consuming. As a compromise, the approach here is to perform calculations to determine the ESP within the cavity, and to use this ESP to approximate the electrostatic interactions with water in many locations. This can serve as an approximate guide to the favourable geometries of water within the cavity that can then be adopted in further single point calculations of the energy.

The geometry of ubiquitin was obtained from the Protein Data Bank. Formal charges were assigned to each group on the basis of the bonding, the number of valence electrons, and the octet rule. There are 24 formally charged groups, 12 with charge +1 and 12 with charge −1. Given the results in the tables above, the molecule was fragmented at Level 3, and the ESP for each fragment was evaluated in the presence of the charges of the other fragments (represented by distributed multipoles), with $\lambda = 3$. The ESP was calculated on a grid which completely enclosed the posited cavity in ubiquitin: a cubic box of length 5 Å on each side, with grid points separated by 0.05 Å on each axis, centered on the position shown in Fig. 3. This position was taken to be the location that was nearly equidistant from residues ILE23, LEU67, and VAL26, and as far as possible from the nearest atoms in ubiquitin. Calculations of the ESP were performed at the HF/6-31G, HF/6-31G(d,p), HF/6-31++G, HF/6-31+G(d), HF/6-31++G(d,p), and MP2/6-31++G(d,p) model chemistries to determine the effects of polarisability, diffuse functions, and correlation. Comparison of these different levels indicated that the ESP changed from the HF to MP2 level by about 0.007 a.u. on average. The variation of the ESP with basis set, at the HF level, was of similar magnitude. In the results reported below, the more reliable MP2/6-31++G(d,p) ESP has been used.

The electrostatic interaction of ubiquitin with a water molecule in the cavity has been evaluated using a set of distributed multipoles for the water molecule, as follows. For the MP2/6-31++G(d,p) electron density for water, the DMA model has been used to evaluate a charge, $q$, dipole, $\mu$, and quadrupole moment, $\Theta$, sited on each atom. The electric field, $\mathbf{F}$, at the grid points in the cavity has been estimated by finite difference of the ESP, $\Phi$:

$$
\mathbf{F}_\alpha(x) = -\frac{\Phi(x + a_\alpha) - \Phi(x - a_\alpha)}{2|a_\alpha|}, \quad \alpha = 1, 2, 3. \quad (4.1)
$$

The gradient of the electric field, $\frac{\partial \mathbf{F}_\alpha(x)}{\partial x_\beta}$, is also estimated by finite difference as

$$
\frac{\partial \mathbf{F}_\alpha(x)}{\partial x_\beta} = \frac{\mathbf{F}_\alpha(x + a_\beta) - \mathbf{F}_\alpha(x - a_\beta)}{2|a_\beta|}, \quad \alpha, \beta = 1, 2, 3. \quad (4.2)
$$

Here, $|a_\alpha| = |a_\beta| = 0.05 \text{ Å}$, and $a_\alpha$ is a vector in the $\alpha$ direction.

The ESP, field, and field gradient at each nuclear position in the water molecule were obtained by interpolation from the corresponding values at the nearest and next nearest grid points in each Cartesian direction. The electrostatic energy of the water molecule was then given simply by $E$:
Equation (4.3) is easily evaluated for any position of the water molecule in the volume spanned by the grid of ESP values. To sample this energy at a large number of positions of the water molecule, a Monte Carlo simulation was performed. Each Monte Carlo move consisted of three rotations for the water molecule [about x, y, and z with random angles in the range [−π, π] and whole molecule translations [in x, y, and z with random size in the range (−0.5,0.5) Å]. The accessible positions were limited by ensuring that each of the waters’ atoms was within the grid, and, in addition, that the minimum of their separation from protein nuclei, divided by the sum atoms was within the grid, and, in addition, that the minimum of their separation from protein nuclei, divided by the sum of their VdW radii, was greater than 0.8. The Monte Carlo simulation was carried out until 10⁶ accessible configurations were compiled. Many of the configurations with low energy were observed to be concentrated in the vicinity of a particular configuration. This geometry and eight related structures were selected.

The interaction energies of these nine water configurations with ubiquitin were evaluated using Level 3 fragmentation with account of non-bonded interactions, as previously described. In this case, since the water molecule is not covalently bonded, or hydrogen bonded, to the ubiquitin molecule, and the geometry is not optimised, the interaction energy is simply the sum of ab initio interaction energies of the water molecule with nearby fragments of the protein plus electrostatic and dispersion interactions between the water and the protein. Figure 4 shows the geometry of the nine water molecules together with the nearby atoms of the protein. Table III presents the ESP-derived binding energies of these nine water molecules together with the corresponding ab initio fragmentation-derived values. The ESP-derived values (column 3) might be directly compared to the ab initio values in column 1, which comprise the ab initio binding energies due to the fragments of ubiquitin in relatively close contact with the water molecule plus electrostatic interactions with all the more distant fragments of ubiquitin. In addition, there is a dispersion, or van der Waals, interaction of the water molecule with all the more distant fragments of ubiquitin. This dispersion interaction has been evaluated and included in the total ab initio binding energy of column 2.

We note that the ESP-derived approximation provides a reasonable estimate of the binding energy. The ESP-derived approximation neglects the exchange-repulsion energy in the interaction of the water molecule with neighbouring fragments of ubiquitin, which likely accounts for the lower binding energy in column 1, compared to column 3. However, the ESP-derived approximation also neglects long range attractive dispersion interactions which increase the total binding energy, as shown in column 2.

On the basis of these calculations, we might estimate the binding energy of water in the cavity of the crystal structure for ubiquitin as about 10 kJ mol⁻¹. Optimisation of the structure, with and without water in the cavity, might change this estimate, but such a calculation is beyond the scope of this paper. We note that the MP2/6-31++G(d,p) binding energy of a water molecule in water, might be estimated from the binding energy in water clusters as about 48 kJ mol⁻¹. On this basis, it seems unlikely that a water molecule would reside in the cavity in ubiquitin, rather than in the surrounding water solvent.

V. CONCLUDING REMARKS

The ESP of a set of organic molecules and ions has been estimated using a simple application of systematic molecular fragmentation. The accuracy of the approach, and its convergence with increasing Level of fragmentation has been demonstrated. For molecules that contain formally charged groups, we have found that ab initio estimates of the ESP of molecular fragments should be evaluated in the presence of distributed charges that model the electrostatic environment of the whole molecule. It has also been seen that the accuracy of the estimated ESP does not depend sensitively

![FIG. 4. The geometries of 9 possible positions for a water molecule are shown together with all atoms of the ubiquitin structure that are within 4.5 Å of the centre of the cavity.](image-url)
on the precise location of these distributed charges at “capping” sites. This simple approach calculates the ESP directly from the electron wavefunction, rather than from a distributed multipole expansion of the molecular charge distribution. Hence, the estimated ESP might be reasonably accurate close to molecular nuclei, where the multipole expansion must fail.

The approach has been demonstrated by estimating the optimum location of a water molecule inside an internal cavity in the structure of ubiquitin. In this case, the water molecule is in close proximity with the atoms of the protein. Since the computational cost of estimating the energy using the ESP is low, it was possible to investigate of order $10^6$ possible locations of the water molecule via a Monte Carlo simulation. The accuracy of the ESP approach was also demonstrated by the fact that $ab$ initio estimates of the binding energy (also using fragmentation) were in reasonable agreement with the ESP estimates. Hence, this approach may provide a simple, and cost effective, method for estimating molecular interaction energies.

It should be noted that all the calculations for ubiquitin have been carried out in the absence of any solvent. Just as the electron density in ubiquitin is polarised by the charge distribution of the molecule itself, so any surrounding water would be polarised. The electric field due to this polarised solvent would also make a contribution to the total field. Hence, in a complete description of ubiquitin in water, one should account for the polarization of the solvent: either by explicitly considering individual water molecules and or by approximating the polarization of distant water molecules by a “reaction field.” In the case studied herein, where the cavity in ubiquitin is well-separated from the solvent, such effects may not be significant.

ACKNOWLEDGMENTS

The authors wish to thank our colleague, Professor Gottfried Oetting, for suggesting the problem of water in ubiquitin, and for helpful discussions.