

# Natural Bond Orbital Analysis in the ONETEP Code: Applications to Large Protein Systems

Louis P. Lee,<sup>\*[a]</sup> Daniel J. Cole,<sup>[a]</sup> Mike C. Payne,<sup>[a]</sup> and Chris-Kriton Skylaris<sup>[b]</sup>

First principles electronic structure calculations are typically performed in terms of molecular orbitals (or bands), providing a straightforward theoretical avenue for approximations of increasing sophistication, but do not usually provide any qualitative chemical information about the system. We can derive such information via post-processing using natural bond orbital (NBO) analysis, which produces a chemical picture of bonding in terms of localized Lewis-type bond and lone pair orbitals that we can use to understand molecular structure and interactions. We present NBO analysis of large-scale calculations with the ONETEP linear-scaling density functional theory package, which we have interfaced with the NBO 5 analysis program. In ONETEP calculations involving thousands of atoms, one is typically interested in particular regions of a nanosystem whilst accounting for long-range electronic effects from the entire system. We show that by transforming the Non-orthogonal

Generalized Wannier Functions of ONETEP to natural atomic orbitals, NBO analysis can be performed within a localized region in such a way that ensures the results are identical to an analysis on the full system. We demonstrate the capabilities of this approach by performing illustrative studies of large proteins—namely, investigating changes in charge transfer between the heme group of myoglobin and its ligands with increasing system size and between a protein and its explicit solvent, estimating the contribution of electronic delocalization to the stabilization of hydrogen bonds in the binding pocket of a drug-receptor complex, and observing, *in situ*, the  $n \rightarrow \pi^*$  hyperconjugative interactions between carbonyl groups that stabilize protein backbones. © 2012 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23150

## Introduction

The natural bond orbital (NBO) analysis method<sup>[1,2]</sup> as implemented by Weinhold and coworkers in the NBO 5 program<sup>[3]</sup> provides chemical insights from first principles calculations by transforming a ground-state quantum mechanical wavefunction into orbitals corresponding to the chemical notion of localized Lewis-type bond and lone pairs, with their associated formally vacant antibonding and Rydberg orbitals. The scheme involves successive transformations of the full ground-state single particle density matrix, first into a set of atom-centered orthogonal “natural atomic orbitals” (NAOs),<sup>[4]</sup> then into “natural hybrid orbitals” (NHOs) and, finally, NBOs. The populated NBOs, formed from the diagonalization of two-center density matrix blocks, are maximal in their occupancies (usually  $\sim 2.0$ ), and constitute the natural Lewis structure of the system. These orbitals are complemented by their formally unoccupied counterparts, whose finite occupancies represent deviation from the ideal Lewis chemical picture due to delocalization effects.

The NBO method has been applied in a wide variety of chemical studies, from rationalizing the electronic origin of nonpairwise additivity in co-operative hydrogen-bonded clusters,<sup>[5]</sup> describing electronic donation between ligand and transition metal orbitals,<sup>[6]</sup> to explaining molecular rotational barriers in terms of the energetics of NBO delocalization.<sup>[7–10]</sup> More recently, the NBO method has been applied to study biologically relevant systems, examples including charge determination via natural population analysis (NPA) along an enzymatic reaction co-ordinate,<sup>[11]</sup>  $n \rightarrow \pi^*$  backbone interactions in proteins,<sup>[12–14]</sup> hydrogen bonding in nucleic acid base pairs<sup>[15]</sup>

and observation of hydrogen bond co-operative strengthening in amides and peptides.<sup>[16]</sup> However, the applicability of first principles methods in determining the ground state properties of biomolecular assemblies is hindered by the scaling of the computational effort with the number of atoms in the system, which is often cubic or greater. In this article, we address this issue using a new generation of density functional theory (DFT) approach whose computational effort scales linearly with the number of atoms. Specifically, we interface the NBO 5 analysis package<sup>[3]</sup> with the linear-scaling DFT code ONETEP,<sup>[17]</sup> which has been developed for parallel computers using novel and highly efficient algorithms that allow calculations for systems up to tens of thousands of atoms.<sup>[18–20]</sup> ONETEP is unique as it achieves linear-scaling computational cost whilst preserving plane-wave accuracy, meaning that the energy and various computed properties can be systematically improved by increasing a single parameter equivalent to the kinetic energy

[a] L. P. Lee, D. J. Cole, M. C. Payne  
TCM Group, Cavendish Laboratory, 19 JJ Thomson Ave, Cambridge CB3 0HE,  
United Kingdom  
E-mail: lpl24@cam.ac.uk

[b] Chris-Kriton Skylaris  
School of Chemistry, University of Southampton, Highfield, Southampton  
SO17 1BJ, United Kingdom

Contract/grant sponsor: EPSRC; Contract/grant number: EP/F032773/1;  
Contract/grant sponsor: IRIDIS3 supercomputer of the University of  
Southampton; Contract/grant sponsor: Cambridge Commonwealth  
Trust; Contract/grant sponsor: EPSRC; Contract/grant sponsor: Royal  
Society for a University Research Fellowship.

© 2012 Wiley Periodicals, Inc.

cut-off of conventional pseudopotential plane-wave DFT implementations.

We begin the Methodology section by reviewing the NAO and ONETEP methods and describing our interface between ONETEP and the NBO 5 program<sup>[3]</sup> that allows us to compute NBOs and their properties from the ONETEP density matrix. We will also show how, by partitioning the density matrix obtained from a ONETEP calculation on a large system, it is possible to perform NBO analysis only in a small region of the large system in such a way that the NBOs obtained in this region are identical to those that would be obtained from NBO analysis on the entire system.

In Results, we first perform extensive validation of the properties of the NBOs derived from ONETEP calculations against the quantum chemistry code GAMESS<sup>[21]</sup> (which is officially supported and integrated with NBO 5), followed by several illustrative applications to large systems, where we use this new functionality in ONETEP to study local chemical properties of several large proteins, and how these are influenced by long-range interactions.

## Methodology

### NAOs

In the original NBO scheme,<sup>[2]</sup> the initial ground-state single particle density matrix  $\mathbf{P}_{\alpha\beta} = \langle \phi_\alpha | \hat{\rho} | \phi_\beta \rangle$ , represented in an atom-local basis  $|\phi_\alpha\rangle$ , is first transformed into an orthogonal NAO representation via symmetric weighted orthogonalization,<sup>[4]</sup> from which all other types of orbitals (NHOs, NBOs) are derived. The initial functions  $\{|\phi_\alpha\rangle\}$  are assumed to be spherical contracted Gaussians, as commonly used in quantum chemistry.

The full AO to NAO transformation  $\mathbf{T}_{\text{NAO}}$  can be summarized as<sup>[4]</sup>:

$$\mathbf{T}_{\text{NAO}} = \mathbf{N}_{\text{PNAO}} \mathbf{O}_S \mathbf{N}_{\text{Ryd}} \mathbf{O}_W \mathbf{N}_{\text{Red}} \quad (1)$$

The first step  $\mathbf{N}_{\text{PNAO}}$  involves the transformation of  $\{|\phi_\alpha\rangle\}$  into a set of “pre-orthogonal NAOs” (PNAOs) by diagonalizing the atom-local block of the (spin-averaged) density matrix  $\mathbf{P}^A$  of atom  $A$ . However, to preserve the invariance of the entire transformation  $\mathbf{T}_{\text{NAO}}$  to molecular orientation with respect to Cartesian rotation of  $\{|\phi_\alpha\rangle\}$ , an “ $lm$ -averaging” step is first performed on  $\mathbf{P}^A$  and the associated overlap matrix  $\mathbf{S}^A$  as:

$$\bar{\mathbf{P}}_{nl,n'l}^{(Al)} = \frac{1}{2l+1} \sum_{m=1}^{2l+1} \mathbf{P}_{nlm,n'l m}^{(Alm)}, \quad \bar{\mathbf{S}}_{nl,n'l}^{(Al)} = \frac{1}{2l+1} \sum_{m=1}^{2l+1} \mathbf{S}_{nlm,n'l m}^{(Alm)} \quad (2)$$

where  $\mathbf{P}^{(Alm)}$  and  $\mathbf{S}^{(Alm)}$  are block diagonals sharing the same  $m \in l$ . The symmetry-averaged  $\bar{\mathbf{P}}^{(Al)}$  and  $\bar{\mathbf{S}}^{(Al)}$  are then diagonalized to obtain the PNAOs  $\mathbf{N}^{(Al)}$ :

$$\bar{\mathbf{P}}_{\alpha\beta}^{(Al)} \mathbf{N}_{\alpha i}^{(Al)\beta} = \bar{\mathbf{S}}_{\alpha\gamma}^{(Al)} \mathbf{N}_{\alpha i}^{(Al)\gamma} f_i^{(Al)} \quad (3)$$

where we have generalized the  $nl$  indices in  $\bar{\mathbf{P}}_{nl,n'l}$  to Greek letters  $\bar{\mathbf{P}}_{\alpha\beta}$ , and the Einstein summation convention is implied.  $\bar{\mathbf{P}}$  and  $\bar{\mathbf{S}}$  are then expanded back into  $2l+1$  block-diagonals for all  $m \in l$  of a particular  $l$  (i.e.,  $\mathbf{N}^{(Alm)} = \mathbf{N}^{(Al)} \forall m \in l$ ). These

PNAOs are hence “natural” in their molecular environment by virtue of having the most condensed occupancies whilst retaining free-atom AO symmetries. This  $lm$ -averaging procedure requires free-atom angular symmetries in the initial basis, an issue that will be discussed in our adaptation of this method to ONETEP.

Subsequent steps include the division of PNAOs into “natural minimal basis” and “natural Rydberg basis” (NMB and NRB) sets constituting the formally occupied and unoccupied atomic orbitals in the ground state configuration. The NMB PNAO set then undergoes an occupancy-weighted symmetric orthogonalization procedure  $\mathbf{O}_W$ , which serves to orthogonalize orbitals with a weighted-preservation of their initial shapes by minimizing the Hilbert space distance:

$$\min \left( \sum_i f_i \| |\phi_i^{\text{PNAO}}\rangle - |\phi_i^{\text{PNAO}'}\rangle \|^2 \right) \Rightarrow \mathbf{O}_W = \mathbf{W}(\mathbf{W}\mathbf{S}\mathbf{W})^{-\frac{1}{2}}$$

where  $\langle \phi_i^{\text{PNAO}'} | \phi_j^{\text{PNAO}'} \rangle = \delta_{ij}$ ,  $\mathbf{S}_{ij} = \langle \phi_i^{\text{PNAO}} | \phi_j^{\text{PNAO}} \rangle$ ,  $\mathbf{W}_{ij} = \delta_{ij} f_j$

(4)

$\mathbf{O}_W$  reduces to Löwdin's symmetric orthogonalization  $\mathbf{O}_L = \mathbf{S}^{-\frac{1}{2}}$  when  $\mathbf{W} = 1$ . The weightings  $\mathbf{W}_{ij} = \delta_{ij} f_j$  are taken to be the symmetry-averaged occupancies  $f_i^{(Al)}$  from Eq. (3), and this ensures that highly occupied orbitals retain their character while low-occupancy ones are free to distort to achieve orthogonality.

Next, the NRBs  $\{|\phi_i^{\text{NRB}}\rangle\}$  are Schmidt-orthogonalized ( $\mathbf{O}_S$ ) with respect to the (now orthogonal) NMB space  $\{|\phi_i^{\text{NMB}'}\rangle\}$ :

$$|\phi_i^{\text{NRB}'}\rangle = |\phi_i^{\text{NRB}}\rangle - \sum_{j \in \text{NMB}} |\phi_j^{\text{NMB}'}\rangle \langle \phi_j^{\text{NMB}'} | \phi_i^{\text{NRB}} \rangle \quad (5)$$

followed by a restoration of the “natural” character of  $\{|\phi_i^{\text{NRB}'}\rangle\}$  by repeating  $\mathbf{N}_{\text{PNAO}}$  [Eq. (3)] only in the NRB space (denoted  $\mathbf{N}_{\text{Ryd}}$ ). The entire Hilbert space is then weighted-orthogonalized ( $\mathbf{O}_W$ ) using the new NRB occupancies, and the “natural” character of this final NAO set is once again restored by repeating  $\mathbf{N}_{\text{PNAO}}$  (denoted  $\mathbf{N}_{\text{Red}}$ ). The NMB/NRB orthogonalization step  $\mathbf{O}_S$  is crucial to ensure stability and rapid convergence of the total NAO population on each atom (the NPA charges) with respect to basis set expansion, by removing unnecessary weightings in  $\mathbf{O}_W$  given to the diffuse NRBs due to their potentially large overlaps with NMBs that have satisfactorily accommodated the electronic population.<sup>[4]</sup>

### ONETEP

ONETEP<sup>[17]</sup> is a linear-scaling DFT package based on the density matrix  $(\rho(\mathbf{r}, \mathbf{r}'))$  formalism, expressed equivalently in terms of atom-centered non-orthogonal generalized Wannier functions (NGWFs)  $\{\phi_\alpha(\mathbf{r})\}$ <sup>[22]</sup>:

$$\rho(\mathbf{r}, \mathbf{r}') = \phi_\alpha(\mathbf{r}) K^{\alpha\beta} \phi_\beta^*(\mathbf{r}') \quad (6)$$

For materials with band gaps,  $\rho(\mathbf{r}, \mathbf{r}')$  decays exponentially with distance between  $\mathbf{r}$  and  $\mathbf{r}'$ <sup>[23]</sup> and this property can be exploited to achieve linear-scaling cost by truncating the

density kernel  $K^{\alpha\beta}$  whenever the centers of  $\phi_\alpha(\mathbf{r})$  and  $\phi_\beta(\mathbf{r})$  exceed a predefined truncation radius.

The NGWFs are represented in a basis of highly localized periodic cardinal sine (psinc) functions (otherwise known as Fourier-Lagrange mesh functions)  $D_k(\mathbf{r})$ , which are, as a result of the cardinality property, only nonzero on a particular grid point  $k$ <sup>[24]</sup>:

$$\phi_\alpha(\mathbf{r}) = \sum_k D_k(\mathbf{r}) C_{k,\alpha} \quad (7)$$

Localization is achieved by confining the NGWFs to be nonzero only within a certain radius from their centers  $r_c$ . The psinc expansion allows the NGWFs to be optimized during calculation in addition to the density kernel, thereby reducing the size of the density kernel by using a minimal set of NGWFs centered on each atom whilst maintaining high accuracy. The psinc functions, in which all quantities are ultimately expressed, are related to plane waves via a Fourier transform, meaning that systematic improvement is possible through adjustment of the psinc grid spacing, analogous to converging the kinetic energy cutoff in traditional  $\mathcal{O}(N^3)$  plane-wave DFT implementations.

The NGWFs need to be initialized on the psinc grid. Possible starting choices are contracted Gaussian basis functions from the STO-3G basis, with optional added polarization functions from the 6-31G\* set, or pseudoatomic orbitals (PAOs) confined to the localization region (fireballs).<sup>[25]</sup> The latter approach is preferable as the NGWFs will be initialized closer to the ground state of the molecular environment.

The density matrix formulation of ONETEP based on atom-centered NGWFs lends itself naturally to the NBO methodology without the need to project the ground-state wavefunctions onto an atom-centered atomic basis set<sup>[26]</sup> as is commonly done in plane-wave codes, thereby avoiding several difficulties such as charge spilling<sup>†</sup> and the arbitrary choice of the projection basis. We have implemented the full NAO transformation scheme [Eq. (1)]<sup>[4]</sup> within the ONETEP program, adapted to accept NGWFs as the starting orbitals, which allows us to re-cast various ground-state wavefunction properties in terms of the NAO basis.

It should be emphasized that ONETEP differs from traditional quantum chemistry packages in the way the orbitals are treated—in ONETEP, a minimal set of NGWFs is often used, but spatially optimized on a real-space grid to provide the best pos-

sible representation within its molecular environment. Such optimization distorts the NGWFs from their initial shapes (usually PAOs), potentially reducing their resemblance to true atomic orbitals. We will show that there is in fact a strong preservation of the initial angular character, making optimized NGWFs an admissible orbital set representation for the NBO method. In addition, the expansion of NGWFs on the psinc grid, as with the plane wave approach, leads to difficulties in describing the highly oscillatory wavefunction close to the nucleus, necessitating the use of pseudopotentials. We use norm-conserving pseudopotentials, ensuring that valence charges are completely described in the pseudized core region—this has the effect of excluding core orbitals from the NBO analysis, although, as we will show in our validation tests, the inherent chemical picture remains consistent with all-electron calculations.

## NBOs

The NBOs,<sup>[1–3,27]</sup> as developed by Weinhold and coworkers in the NBO 5 analysis package,<sup>[3]</sup> are constructed individually for each pair of chemically bonded atoms  $A$  and  $B$  from the density matrix in the NAO representation. This is achieved by diagonalizing a pair-block of the density matrix  $\mathbf{P}^{AB}$  spanning all NAOs centered on atoms  $A$  and  $B$ :

$$\mathbf{P} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots \\ \dots & p^{AA} & \dots & p^{AB} & \dots \\ & \vdots & \ddots & \vdots & \\ p^{BA} & & & p^{BB} & \\ \vdots & & & \vdots & \ddots \end{pmatrix} \rightarrow \mathbf{P}^{AB} = \begin{pmatrix} p^{AA} & p^{AB} \\ p^{BA} & p^{BB} \end{pmatrix} \quad (8)$$

$$\mathbf{P}^{AB} \mathbf{b}_i^{AB} = \mathbf{b}_i^{AB} \epsilon_i$$

to obtain the two-center orbitals  $\{\mathbf{b}_i^{AB}\}$  that have maximal occupancies within the  $A - B$  subspace. In practice, the single-center blocks  $\mathbf{P}^{AA}$  are first diagonalized to extract a set of highly occupied atom-centered orbitals  $\{\mathbf{h}_i^{AA}\}$  which correspond to unhybridized core orbitals and lone pair hybrids, defined when their occupancies exceed a threshold  $n_{\min}$  (1.90 e in NBO 5). To avoid eigenvector mixing due to possible near-degeneracies between the occupancies of the aforementioned orbitals with occupied bond orbitals,<sup>[27]</sup> core and lone pairs are projected out of each  $\mathbf{P}^{AB}$  block, resulting in the depleted matrix  $\tilde{\mathbf{P}}^{AB}$  which is then diagonalized as in Eq. (8) to obtain the set of orbitals  $\{\mathbf{b}_i^{AB}\}$  with occupancies greater than  $n_{\min}$ . These orbitals are composed of their atom-centered hybrids  $\{\mathbf{h}_i^{AA}\}$  and  $\{\mathbf{h}_i^{BB}\}$ :

$$\mathbf{b}_i^{AB} = c_i^{AA} \mathbf{h}_i^{AA} + c_i^{BB} \mathbf{h}_i^{BB} \quad (9)$$

with polarization coefficients  $c_i^{A,B}$ . The set of core, lone pairs, and bond-participating hybrids  $\{\mathbf{h}_i^{AA}, \mathbf{h}_i^{BB}\}^\ddagger$  on each atom are

<sup>†</sup>Charge spilling as defined by Sánchez-Portal et al. <sup>[26]</sup>,  $S_\Omega$  for a plane-wave basis:

$$S_\Omega = \frac{1}{N_k N_x} \sum_k \sum_x |\psi_x(k)| (1 - P(k)) |\psi_x(k)|$$

$$P(k) = \sum_i |\phi_i(k)\langle\phi_i(k)|$$

where  $N_k$  and  $N_x$  denote the number of Brillouin Zone points and plane-wave eigenstates being considered, respectively,  $|\psi_x(k)\rangle$  the plane-wave eigenstates, and  $|\phi_i(k)\rangle$  the (non-orthogonal) atom-centered orbitals (in reciprocal space) onto which  $|\psi_x(k)\rangle$  are projected. Spilling is the result of both plane-wave and atom-centered Hilbert spaces not fully spanning one another.

<sup>‡</sup>Pre-orthogonal NHOs are NHOs where the NAO basis is replaced by PNAOs while retaining the NAO coefficients, hence they are not the  $\mathbf{h}_i^{AA}$  in Eq. (9), despite being non-orthogonal.<sup>[3]</sup> Similar definitions apply to pre-orthogonal NBOs<sup>[1]</sup>.

subsequently Löwdin-orthogonalized to obtain the NBOs  $\{\mathbf{n}_i^A, \mathbf{h}_i^A\}$ ,<sup>[27]</sup> which are directional along chemical bonds within the molecule, and provide optimal (for the chemical environment of each atom) hybrid atomic orbitals familiar from traditional chemistry concepts.

Each  $2 \times 2$  density matrix block in the  $\{\mathbf{h}_i^A, \mathbf{h}_i^B\}$  basis [related to Eq. (9)]  $\mathbf{P}^{\mathbf{h}_i^A, \mathbf{h}_i^B}$  of bond  $i$  for atoms  $A$  and  $B$  is then re-diagonalized to obtain the final NBO  $\sigma_i^{AB}$  and its antibonding counterpart  $\sigma_i^{AB*}$ :

$$\sigma_i^{AB} = c_i^A \mathbf{h}_i^A + c_i^B \mathbf{h}_i^B \quad \sigma_i^{AB*} = c_i^B \mathbf{h}_i^A - c_i^A \mathbf{h}_i^B \quad (10)$$

If an extended basis (with NRBS) is used, the NBOs  $\{\sigma_i^{AB}\}$  will be augmented with a set of single-center low-occupancy Rydberg orbitals, constructed from the residual density matrix in the NAO basis after depleting all Lewis-type (lone pairs and bond) NBOs, forming a set of orthogonal orbitals to complete the variational space spanned by the original system.<sup>[28]</sup>

The final set of  $N_e/2$  ( $N_e$  = the total number of electrons) bonding NBOs represents the natural Lewis structure of the molecule, providing an optimal representation of the classical Lewis picture of bonding in the molecule. The NBO occupancies are often close to 2.0, and their sum commonly includes more than 99% of  $N_e$  in simple molecules.<sup>[2]</sup> Residual occupancies in the formally vacant Rydbergs and antibonds on the other hand represent irreducible delocalization in the ground-state wavefunction, and provide a measure of the deviation from an ideal Lewis picture.<sup>[2]</sup>

To perform NBO analysis in ONETEP, we first transform (internally) all relevant matrices from the NGWF into the NAO representation, and write them out into a formatted file (FILE.47) which serves as the input for the NBO 5 program.

### Donor/Acceptor Interaction

One of the goals of interfacing the NBO method with ONETEP is to provide chemical insights into regions of large systems by studying effects such as electronic delocalization.<sup>[1,29]</sup> An example is the  $n \rightarrow \sigma^*$  secondary hyperconjugation interaction of a hydrogen bond involving intermolecular delocalization (charge transfer) between a lone pair donor  $n$  and an antibond acceptor  $\sigma^*$  of an adjacent molecule. Such interactions are prevalent in biological systems, stabilizing protein and nucleic acid structures and also regulating their interactions with their environment.<sup>[12,15,16]</sup>

In the NBO formalism, this delocalized charge transfer represents deviation from an ideal Lewis description. Such noncovalent contributions are introduced to the overall bonding picture via the relaxation of electronic localization within bonding NBOs by considering delocalization to antibonds that results in variational energetic lowering.<sup>[2]</sup> This energetic lowering can be estimated via second order perturbation theory,<sup>[12,15,16]</sup> where the total Hamiltonian is decomposed into a zeroth-order part  $\mathbf{H}^{(0)}$  containing the diagonal elements  $H_{ii} = \langle \sigma_i | \hat{H} | \sigma_i \rangle$  and perturbative part  $\mathbf{H}'$  containing all off-diagonal components:

$$\mathbf{H} = \mathbf{H}^{(0)} + \lambda \mathbf{H}' \quad (11)$$

$$\mathbf{H}^{(0)} = \begin{pmatrix} H_{11} & 0 & \cdots & 0 \\ 0 & H_{22} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & H_{nn} \end{pmatrix}, \quad \mathbf{H}' = \begin{pmatrix} 0 & H_{12} & \cdots & H_{1n} \\ H_{21} & 0 & \cdots & H_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{n1} & H_{n2} & \cdots & 0 \end{pmatrix} \quad (12)$$

$$\Delta E_i^{1st} = f_i \langle \sigma_i | \hat{H}' | \sigma_i \rangle = 0 \quad (\text{since } H'_{ii} = 0) \quad (13)$$

$$\Delta E_i^{2nd} = \sum_{j \neq i} f_j \frac{\langle \sigma_i | \hat{H}' | \sigma_j \rangle^2}{E_i - E_j} \quad (\text{since } H_{ij} = H'_{ij} \text{ for } i \neq j) \quad (14)$$

where the sum over  $j$  in  $\Delta E_i^{2nd}$  is over all occupied and unoccupied NBOs. The first-order energy correction of  $|\sigma_i\rangle$ ,  $\Delta E_i^{1st}$ , vanishes by virtue of the perturbative decomposition, leaving  $\Delta E_i^{2nd}$  as the first nonzero energetic correction due to off-diagonal couplings. We can hence estimate the stabilization effects of  $\sigma_i \rightarrow \sigma_j^*$  bond-antibond NBO interactions by inspecting particular  $\Delta E_{i-j}^{2nd} = f_i \frac{\langle \sigma_i | \hat{H}' | \sigma_j^* \rangle^2}{E_i - E_j^*}$  elements.<sup>[1,29]</sup>

### Selective NBO generation

In simulations of nanoscale structures, we are often interested in small, well-defined regions, such as the active site of a protein or a particular face of a nanoparticle. However, a full quantum mechanical treatment of the extended system may be necessary to obtain accurate values for properties such as forces and binding energies, as these may converge slowly with the size of the quantum mechanical subsystem.<sup>[30]</sup> Therefore, after performing a DFT-level calculation on the full system, we wish to perform the NBO transformation and analysis only for the subregion of interest, as this produces a much simpler summary containing only interactions that we intend to study. An additional practical reason is that NBO 5 has a technical limit of 999 atoms for proper functionality.

Given that NBOs are manifestly local, it should be possible to generate them only for the region of interest from the density matrix of the entire system. This is accomplished by constructing a partial density matrix  $\mathbf{P}^{\text{Part}}$  from pair-blocks of the full density matrix  $\mathbf{P}$  in the NAO representation as in Eq. (8), and passing this, plus all other associated partial matrices such as the Hamiltonian to NBO 5:

$$\mathbf{P} = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots \\ \cdots & p_{XX} & \cdots & p_{XY} & p_{XZ} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & p_{YX} & \cdots & p_{YY} & p_{YZ} & \vdots \\ p_{ZX} & p_{ZY} & p_{ZZ} & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix} \rightarrow \mathbf{P}^{\text{Part}} = \begin{pmatrix} p_{XX} & p_{XY} & p_{XZ} & \cdots \\ p_{YX} & p_{YY} & p_{YZ} & \vdots \\ p_{ZX} & p_{ZY} & p_{ZZ} & \vdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \quad (15)$$

where  $\{X, Y, \dots\}$  denotes all atoms from the region of interest and  $\{Z, \dots\}$  the shell of "fringe" atoms connecting the region of interest with the rest of the system.

The local nature of the NBO generation, which, for example, for a pair of atoms  $A$  and  $B$  requires operations only on the density matrix pair block  $\mathbf{P}^{AB}$ , ensures that the resultant bond orbitals on the  $\{X, Y, \dots\}$  set will be identical to those constructed from  $\mathbf{P}^{\text{Full}}$  (provided the same NBO search parameter  $n_{\text{min}}$ , as described

in Methodology, is used. The inclusion of the  $\{Z, \dots\}$  set is necessary as, without them, atoms at the spatial fringe of the partition would have their bonds to the rest of the system severed, and the atom-centered hybrids  $\mathbf{h}_i^A$  that would otherwise participate in bond formation become dangling lone pairs with different eigenvectors instead, which would in turn would affect all  $\mathbf{h}_i^A$  centered on the atom  $A$  during the symmetric orthogonalization procedure used to produce the final set of orthogonal NHOs. It should be emphasized that  $\mathbf{P}^{\text{Part}}$  incorporates all the effects of the full system, such as electronic polarization at the DFT level, in the region being partitioned.

Provided the number of atoms in the region of interest  $\{X, Y, \dots\}$  is constant with increasing system size, the cost of the selective generation of NBOs from  $\mathbf{P}^{\text{Part}}$  is also constant with system size. However, since  $\mathbf{P}^{\text{Part}}$  (and other associated matrices such as the partial Hamiltonian) is in the NAO representation, we require the NGWF to NAO transformation on the whole system, an operation which scales cubically with system size due to the need to compute  $\mathbf{S}^{-\frac{1}{2}}$ . This is however a one-off postprocessing step with modest computational cost for practical calculations on systems with thousands of atoms. For example, analysis of a 8006-atom drug-receptor complex presented later in Results, using the internal NGWF to NAO implementation in ONETEP, followed by the output of  $\mathbf{P}^{\text{Part}}$  and its associated matrices for analysis in NBO 5 for a  $\sim 210$ -atom subregion required approximately 1.1 h on 192 Intel Westmere cores, utilizing the scalAPACK package for matrix operations. The single point DFT calculation itself required 85 h on the same number of cores. We note that while it is possible to circumvent the NAO transformation and pass the partial matrices in the NGWF representation, complications would arise from NGWF interatomic overlaps being neglected during partitioning, necessitating further convergence tests on the necessary inclusion of extra buffer/fringe atoms surrounding the region of interest, possibly limiting our ability to study larger partitions.

## Results

### Computational methods

We have implemented both the NAO transformation procedure, to cast the NGWFs as a set of orthogonal NAOs, and our density matrix partitioning method for selective NBO generation in ONETEP, allowing direct calculation of NAO-based properties such as NPA charges, while further NBO-based analysis is made possible by passing required (either for the whole sys-

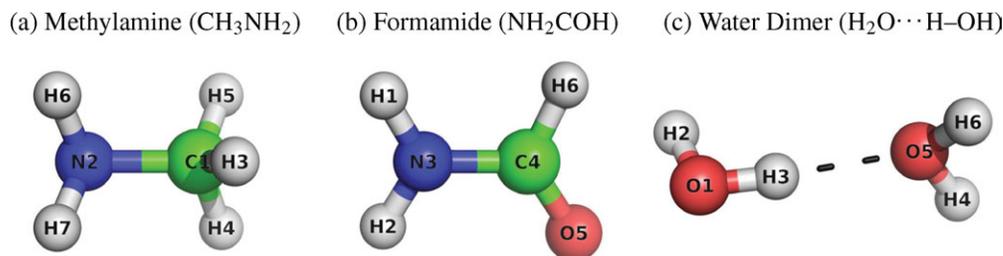


Figure 1. Small system test molecules and their atomic labeling. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

tem or for a portion of it) matrices in the NAO basis to the NBO 5 analysis package. In the examples presented here, ONETEP calculations were performed with the PBE<sup>[31]</sup> exchange–correlation functional using norm-conserving pseudopotentials. Unless otherwise stated, simulations were performed in a cubic periodic supercell with sides of 25 Å, with a psinc energetic cutoff of 1200 eV and NGWF cutoff radii of 10.0  $a_0$ . In ONETEP, NGWF sets labeled “ONETEP Minimal” correspond to a minimal number of NGWFs (1s for 1 NGWF on H, 2s2p for 4 NGWFs on C, N, O) initialized as pseudo-atomic orbitals (PAO) obtained from solving the Kohn–Sham equation for the free atom within the confined NGWF cutoff region,<sup>[25]</sup> using identical pseudopotentials as used in the full calculation (bare Coulomb for H, pseudized 1s<sup>2</sup> core for second row atoms C, N, and O), while “ONETEP Extended 1” and “ONETEP Extended 2” correspond to PAO sets with more polarization/Rydberg functions (1s2p for four NGWFs on H, 2s2p3d for nine NGWFs on C, N, O in “Extended 1” and 1s2s2p3s3p for nine NGWFs on H, 2s2p3s3p3d4s4p for 17 NGWFs on C, N, O in “Extended 2”).

Validation was performed against the *ab initio* quantum chemistry code GAMESS-US<sup>[21]</sup> compiled with the NBO 5 module, using the PBE exchange–correlation functional, all electron calculations and, unless otherwise stated, the aug-cc-pVQZ basis set. Small molecules used for validation and their atom labeling are shown in Figure 1.

### NAOs

As a result of their *in situ* optimization, the final NGWFs in ONETEP do not possess pure angular symmetries. Therefore, the AO to NAO transformation as applied to them, specifically steps involving  $lm$ -averaging (Eq. 2), is no longer rigorously defined. However, in all test cases, the final optimized set of NGWFs were observed to retain much of the angular character of their initial states, > 98% for the highly occupied minimal NGWFs, with very little charge spilling per orbital (0.01  $e$ ), especially when initialized as PAOs (Table 1). Using extended PAO-initialized orbital sets results in stronger preservation of initial NGWF angular ( $l$ ) character, most likely due to the increase in variational freedom available in the density kernel which can then be used to accommodate more of the wavefunction energetic minimization compared to real-space NGWF optimization. In fact, the resemblance of the final NGWFs to their initial PAO state as determined by  $\langle \phi_{\text{Initial}}^{\text{NGWF}} | \phi_{\text{Final}}^{\text{NGWF}} \rangle$  is remarkable, often exceeding 99%, due to the fact that the initial PAOs were generated close to

**Table 1.** ONETEP NGWF angular character analysis for N2 and H3 ( $\phi_{\text{Final}}^{\text{NGWF}}$ ) in Methylamine (Fig. 1a), after the same-center Löwdin orthogonalization.

Atom	NGWF #	Initial AO	%s	%p	%d	%f	Spilling (e)	$\langle \phi_{\text{Initial}}^{\text{NGWF}}   \phi_{\text{Final}}^{\text{NGWF}} \rangle$
ONETEP Extended 1								
N2	1	s	99.34	0.115	0.163	0.377	0.0021	0.994
	2	$p_y$	0.159	99.398	0.187	0.255	0.0033	0.994
	3	$p_z$	0.000	99.337	0.488	0.174	0.0050	0.991
	4	$p_x$	0.092	99.522	0.246	0.139	0.0036	0.995
	5	$d_{xy}$	0.014	0.049	99.750	0.186	0.0027	0.995
	6	$d_{yz}$	0.000	0.049	99.840	0.111	0.0016	0.997
	7	$d_{z^2}$	0.037	0.059	99.765	0.139	0.0018	0.996
	8	$d_{xz}$	0.000	0.013	99.871	0.116	0.0024	0.997
	9	$d_{x^2-y^2}$	0.198	0.054	99.518	0.230	0.0045	0.993
H3	1	s	99.592	0.175	0.140	0.093	0.0021	0.996
	2	$p_y$	0.354	99.391	0.122	0.134	0.0056	0.993
	3	$p_z$	0.000	99.813	0.099	0.089	0.0022	0.996
	4	$p_x$	0.127	99.684	0.110	0.079	0.0030	0.996
ONETEP Minimal								
N2	1	s	98.597	0.404	0.443	0.557	0.0026	0.989
	2	$p_y$	0.496	98.727	0.355	0.421	0.0037	0.988
	3	$p_z$	0.000	98.013	1.650	0.336	0.0080	0.972
	4	$p_x$	0.296	98.708	0.495	0.501	0.0062	0.987
H3	1	s	98.663	0.443	0.646	0.249	0.0033	0.991

$|\phi_{\text{Initial}}^{\text{NGWF}}\rangle$  is the PAO as in column 3, which is the initial state of each NGWF.

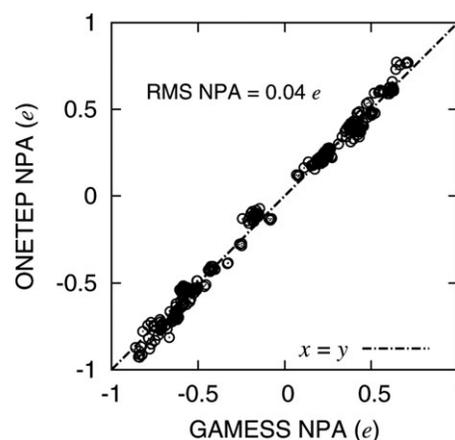
their ground-state to begin with. Therefore, we label the NGWFs according to their initial PAO  $lm$ -symmetry for the  $lm$ -averaging step in the NAO transformation, assuming that the angular impurities of the final NGWFs are sufficiently small that averaging over angular momentum channels will only cause minor mixing of nonrelevant angular components and hence negligible errors in the overall NAO transformation.

**Table 2.** Comparison between NPA charges on methylamine ( $\text{CH}_3\text{NH}_2$ ), formamide ( $\text{NH}_2\text{COH}$ ), and water dimer ( $\text{H}_2\text{O}\cdots\text{H}-\text{OH}$ ).

Atom	GAMESS	ONETEP		
	aug-cc-pVQZ	ONETEP Minimal	ONETEP Extended 1	ONETEP Extended 2
<b><math>\text{CH}_3\text{NH}_2</math></b>				
C1	-0.427	-0.510	-0.452	-0.366
N2	-0.840	-0.920	-0.869	-0.810
H3	0.172	0.192	0.177	0.150
H4	0.198	0.227	0.206	0.177
H5	0.198	0.227	0.206	0.177
H6	0.350	0.393	0.366	0.335
H7	0.350	0.393	0.366	0.336
RMS dQ with respect to GAMESS:		0.052	0.017	0.030
<b><math>\text{NH}_2\text{COH}</math></b>				
H1	0.390	0.429	0.400	0.372
H2	0.394	0.431	0.400	0.378
N3	-0.780	-0.878	-0.842	-0.757
C4	0.440	0.464	0.482	0.502
O5	-0.544	-0.574	-0.548	-0.568
H6	0.101	0.129	0.105	0.073
RMS dQ with respect to GAMESS:		0.050	0.031	0.032
<b><math>\text{H}_2\text{O}\cdots\text{H}-\text{OH}</math> dimer</b>				
O1	-0.944	-1.086	-1.034	-0.950
H2	0.451	0.531	0.525	0.456
H3	0.474	0.512	0.466	0.480
H4	0.473	0.551	0.535	0.480
O5	-0.927	-1.058	-1.017	-0.947
H6	0.473	0.551	0.535	0.480
RMS dQ with respect to GAMESS:		0.098	0.070	0.010

Despite retaining much of their initial AO character, the final NGWFs are in general still non-orthogonal with respect to the same atomic center. In most cases, overlap magnitudes of  $< 0.01$  are observed. Moreover, since the NGWFs are ultimately represented on a real-space grid of finite resolution, their orthogonality is only approximate even in the initial AO state. As the  $lm$ -averaging step only mixes between density matrix components of nonoverlapping orbitals  $m \in l$ , we perform atom-centered Löwdin orthogonalization on the final NGWFs to preserve their shapes while achieving an AO-like orthogonal character—this transformation is observed to have negligible impact on the angular character or resemblance of the optimized NGWFs to their initial state.

Table 2 shows a comparison between NPA charges calculated using various ONETEP NGWF sets and GAMESS using the aug-cc-pVQZ basis set, while Figure 2 shows the NPA charge



**Figure 2.** Correlation between GAMESS cc-pVTZ and 'ONETEP Minimal' NPA charges for four different  $\sim 100$ -atom protein fragments from the RAD51-BRC4 (PDB:1N0W) complex.

**Table 3.** NPA and Mulliken charges for a boron-doped 64-atom bulk silicon cubic lattice showing the stability of NPA charges as a function of NGWF set size in *ONETEP*.

Basis	Population ( <i>e</i> )			
	B		Si	
	NPA	Mulliken	NPA	Mulliken
<i>ONETEP</i> Minimal	−0.780	−0.756	0.202	0.202
<i>ONETEP</i> Extended 1	−0.780	−0.182	0.190	0.090
CASTEP <sup>[33]</sup>	–	−0.78	–	0.20

The ‘*ONETEP* Extended 1’ set for Si includes NGWFs initialized as 3s3p3d PAOs with pseudized 1s2s2p core.

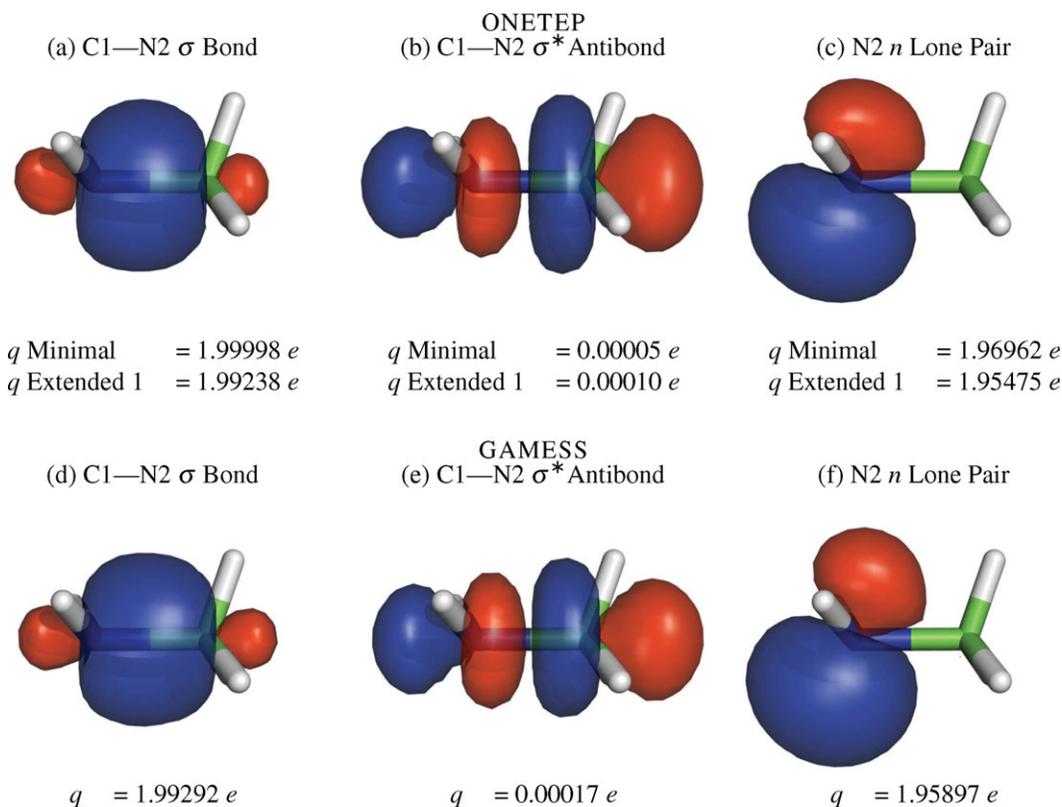
correlation for four ~100 atom protein fragments extracted from the RAD51-BRC4 protein–protein complex [Protein Data Bank (PDB): 1NOW], between *ONETEP* and *GAMESS* calculations using the ‘*ONETEP* Minimal’ and cc-pVTZ basis respectively. The protein fragment simulations in *ONETEP* were performed using a 0.45  $a_0$  psinc grid spacing (~1000 eV plane-wave cut-off) and using the spherical Coulomb cut-off approach to eliminate periodic image interactions,<sup>[32]</sup> in a cubic simulation cell commensurate with the Coulomb cut-off radius. Considering the difference in the treatments of core electrons, cell periodicity, and basis functions in the two codes, the agreement (< 0.1 *e*) is good, even for the computationally inexpensive ‘*ONETEP* Minimal’ NGWF set and may be improved by increasing the number of NGWFs in *ONETEP*.

The NAO transformation can be also be applied to periodic systems,<sup>[28]</sup> where periodicity is described by the density matrix. The *ONETEP* approach, where optimization of the NGWFs and density kernel takes place in real-space, has a major advantage over *k*-space methods in treating systems without inherent symmetry. Such an example is the study of crystalline defects, where a sufficiently large primitive unit cell is required to avoid interactions with periodic images. Table 3 shows a calculation on a periodic 64-atom B-doped crystalline Si supercell performed

in *ONETEP* using both the ‘*ONETEP* Minimal’ and the ‘*ONETEP* Extended 1’ NGWF sets. The NPA charges on B and its neighboring Si show agreement with an equivalent calculation<sup>[33]</sup> performed using the plane-wave DFT code CASTEP<sup>[34]</sup> using valence pseudo-atomic orbital basis set projection with Mulliken population analysis.<sup>[35,36]</sup> It is worth noting the stability of the NPA charges with respect to increasing NGWF set size as compared to Mulliken charges, which suffers due to the overrepresentation of the molecular wavefunction when highly overlapping diffuse functions are included in the calculation.

## NBOs

The density matrix for the methylamine system was transformed internally within *ONETEP* into the orthogonal NAO basis and written as an input file for *NBO 5*,<sup>[3]</sup> from which the NBOs were obtained. Figure 3 compares a number of example NBOs in methylamine obtained from *ONETEP* with those from *GAMESS* and their respective occupancies, showing the expected double occupancy of Lewis-type bonding orbitals and vacancy of their antibonding counterparts. Despite the differences in computational approach and our assumption of the NGWFs having pure AO symmetry, an identical Lewis picture with good agreement in occupancy is obtained between the two codes.



**Figure 3.** Examples of NBOs obtained from the NAOs generated internally in *ONETEP* from the final optimized NGWFs, using the ‘*ONETEP* Extended 1’ NGWF set. NBOs have been normalized to unity and plotted with an isosurface value of  $\pm 0.05$  a.u. (red  $-$ , blue  $+$ ). Orbital occupancies *q* are given below each figure. *GAMESS* NBOs (plotted for the triple valence- $\zeta$  basis, occupancies from aug-cc-pVQZ calculation) are qualitatively identical, as expected from LCAO theory. Plots were generated from Gaussian Cube files, obtained directly in *ONETEP*, and via *NBO2molden*<sup>[37]</sup> and *molden 5.0*<sup>[38]</sup> in *GAMESS*. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

**Table 4.** Water dimer  $n_{\text{O}} \rightarrow \sigma_{\text{O-H}}^*$  interaction attributes for different ONETEP and GAMESS orbital sets.

Basis	# Functions	Energies (kcal/mol)		Charges (e)		
		$E_{\text{Dimer}}$	$\Delta E^{2\text{nd}}$	$dQ_{\text{H}_2\text{O}(1 \rightarrow 2)}$	$\Delta \sigma_{\text{O-H}}^*$	$\Delta n_{\text{O}}$
ONETEP						
ONETEP Minimal	12	-5.19	15.34	0.0429	0.0434	0.0438
ONETEP Extended 1	34	-4.86	15.20	0.0520	0.0426	0.0369
ONETEP Extended 2	70	-5.06	7.82	0.0139	0.0179	0.0165
GAMESS						
STO-3G	14	-4.98	16.13	0.0461	0.0475	0.0437
4-31G	26	-8.12	12.77	0.0304	0.0305	0.0306
aug-cc-pVQZ	430	-4.75	7.29	0.0185	0.0174	0.0158

The last two columns are changes in occupancies of the  $\sigma_{\text{O-H}}^*$  and  $n_{\text{O}}$  orbitals between monomer and dimer systems, illustrating the dominance of the  $n_{\text{O}} \rightarrow \sigma_{\text{O-H}}^*$  donor-acceptor interaction in describing the net charge transfer between monomers  $dQ_{\text{H}_2\text{O}(1 \rightarrow 2)}$ .

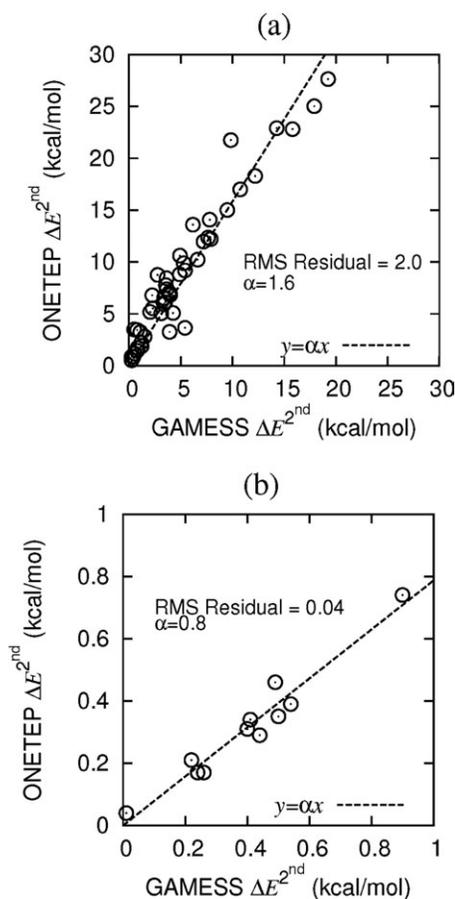
A second example, dealing with the  $n \rightarrow \sigma^*$  hydrogen bond donor-acceptor interaction in water dimer (Fig. 1c) is presented in Table 4, where electronic delocalization due to intermolecular hyperconjugation between an occupied orbital ( $n$ ) from one molecule and an antibond of the other ( $\sigma^*$ ) contributes to significant stabilization of the hydrogen bond interaction by overcoming strong steric repulsion between the monomers, allowing them to approach more closely, penetrating the van der Waals contact radius.<sup>[2]</sup> The dimer geometry was first optimized in ONETEP and used in single-point energy calculations in both GAMESS and ONETEP. Dimerization energy (monomers in dimer geometries),  $E_{\text{Dimer}}$  in ONETEP was -4.86 kcal/mol (ONETEP Extended 1), compared to -4.75 kcal/mol in GAMESS (aug-cc-pVQZ). All GAMESS dimer energies were calculated using BSSE counterpoise correction by Boys and Benardi.<sup>[39]</sup> In ONETEP, the optimization of NGWFs independently in both monomer and dimer geometries eliminates the need for such corrections.<sup>[40]</sup>

In NBO 5, all possible bond-antibond  $\Delta E^{2\text{nd}}$  energies are evaluated.<sup>[3]</sup> Our calculations reveal one prominent interaction between the oxygen lone pair  $n_{\text{O}}$  of the first monomer and the O-H antibond  $\sigma_{\text{O-H}}^*$  of the second, whose donor-acceptor interaction describes a significant portion of the total charge transfer between monomers. This picture is consistent across both GAMESS and ONETEP using different basis/NGWF sets, and with previous NBO studies.<sup>[2,41]</sup> Quantitatively,  $\Delta E^{2\text{nd}}$  of the hyperconjugation interaction was observed to be highly dependent on the size of the basis/NGWF sets, especially for the more compact ones, both in GAMESS using fixed AOs and in ONETEP using optimized NGWFs (Table 4) regardless of the high accuracy of the dimerization energy in the latter.

$\Delta E^{2\text{nd}}$ , the net charge transfer between monomers  $dQ_{\text{H}_2\text{O}(1 \rightarrow 2)}$ , and the occupancies of  $\sigma_{\text{O-H}}^*$  and  $n_{\text{O}}$  were observed to converge with increasing basis/NGWF set size. The restricted dimensionality of the smaller orbital sets causes electronic populations to be condensed into a limited set of NBOs, possibly forcing the antibonding orbitals to completely describe delocalization, whereas in an extended set of functions, the single-center diffuse Rydbergs derived from the NRBs take on some of this role, as seen by the systematic lowering of  $\sigma_{\text{O-H}}^*$  occupancy with increasing number of basis

functions. In addition, delocalization into intramolecular Rydbergs could also explain the suppression of  $dQ_{\text{H}_2\text{O}(1 \rightarrow 2)}$ . Subsequently, the  $n \rightarrow \sigma^*$  interaction alone plays a lesser role in intermolecular delocalization. We note that a similar basis set dependence of the energetic lowering estimate of such intermolecular charge

transfer delocalization has been observed by others.<sup>[42]</sup> Although extended basis sets are required for converged  $\Delta E^{2\text{nd}}$ , a reasonable qualitative picture may be obtained for such interactions in realistic systems employing just a minimal NGWF set in ONETEP, as illustrated in Figure 4a—by comparing the  $\Delta E^{2\text{nd}}$  hyperconjugation estimates of  $n_{\text{O}} \rightarrow \sigma_{\text{O-H}}^*$  interactions in hydrogen bonds present in the four protein fragments



**Figure 4.** Correlation between GAMESS CC-pVTZ and 'ONETEP Minimal'  $\Delta E^{2\text{nd}}$  estimates for (a)  $n_{\text{O}} \rightarrow \sigma_{\text{O-H}}^*$  hydrogen bond and (b)  $n_{\text{O}} \rightarrow \pi_{\text{C=O}}^*$  Bürgi-Dunitz type interactions from the same calculations of the four  $\sim 100$ -atom protein fragments as in Figure 2. RMS residuals were obtained from a  $y = \alpha x$  linear regression.

of Figure 2, a reasonable correlation with a root mean square (RMS) residual (from linear regression) of 2.0 kcal/mol is observed between the ONETEP calculations using a minimal NGWF set and GAMESS calculation at the cc-pVTZ level, the latter of which is converged with respect to  $\Delta E^{2\text{nd}}$ . We conclude that despite the inherent overestimation when using smaller orbital sets,  $\Delta E^{2\text{nd}}$  of  $n \rightarrow \sigma^*$  interactions are sufficiently well-correlated with converged values, allowing us to reliably identify and qualitatively compare their strengths despite using a minimal NGWF set in ONETEP. On a similar note, we have also investigated the correlation of Bürgi–Dunitz type  $n_{\text{O}} \rightarrow \pi_{\text{C=O}}^*$  interactions between consecutive backbone carbonyl groups in the same samples (Fig. 4b), which appear to be slightly underestimated in ONETEP for our limited sample of points, but otherwise well-correlated, with an RMS residual of 0.04 kcal/mol.

### Illustrative applications to large protein systems

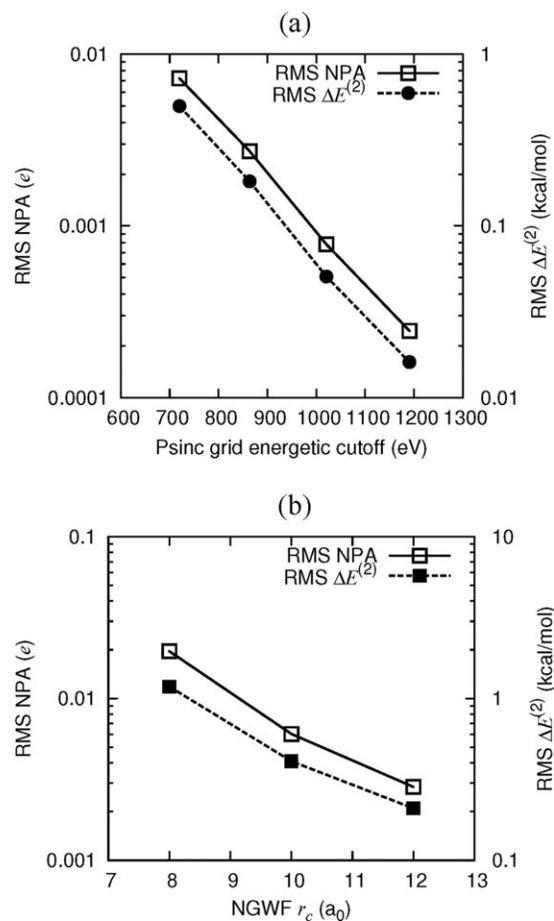
With our ONETEP-NBO 5 interface and the ability to analyze subregions of large systems by selectively passing required partial matrices to NBO 5, we have performed verification tests and sample studies on local chemical properties obtainable via the NBO transformation, ranging from simple NPA analysis on the charge transfer to a bound ligand and between a protein and its solvent, to probing stabilizing hyperconjugative interactions in small localized regions of large systems such as the avidin–biotin binding pocket and the backbone-stabilizing  $n \rightarrow \pi^*$  interaction.<sup>[12]</sup> The effect of having a full quantum treatment of the entire system on these local properties was also investigated.

### Set up for large proteins

X-ray crystal structures for all large protein systems studied in sections entitled “Charge Transfer Between Protein and Surface Water,” “The Avidin–Biotin Complex,” and “ $n \rightarrow \pi^*$  Interactions in the RAD51-BRC4 Complex” were obtained from the PDB (1K7K, 1AVD, and 1N0W respectively). Hydrogen atoms were added with pdb2pqr<sup>[43,44]</sup> and crystallographic water molecules were retained. The structures were minimized with the AMBER 10 molecular dynamics package<sup>[45]</sup> and the AMBER ff99SB biomolecular force field<sup>[46]</sup> with the heavy atoms of the protein frozen in their crystallographic positions.

Convergence tests for NPA charges and  $n \rightarrow \sigma^*$   $\Delta E^{2\text{nd}}$  hyperconjugation estimates of hydrogen bonds were performed with respect to the psinc grid spacing and NGWF cutoff radius using a small 80-atom fragment from the RAD51-BRC4 (PDB: 1N0W) protein–protein complex, where additional hydrogen atoms were added to terminate broken bonds. The “ONETEP Minimal” NGWF representation was used, and a spherical cut-off approach for Coulomb potentials was used to eliminate interactions of the fragment with its periodic image,<sup>[32]</sup> in a cubic simulation cell commensurate with the Coulomb cut-off radius.

RMS deviations of NPA charges and  $\Delta E^{2\text{nd}}$  interactions were both observed to converge rapidly with increasing NGWF radius and psinc energy cutoff (Fig. 5). Based on these calibration tests we have chosen to use (unless stated otherwise in subsequent sections)  $0.45 a_0$  for the psinc spacing (on a cubic



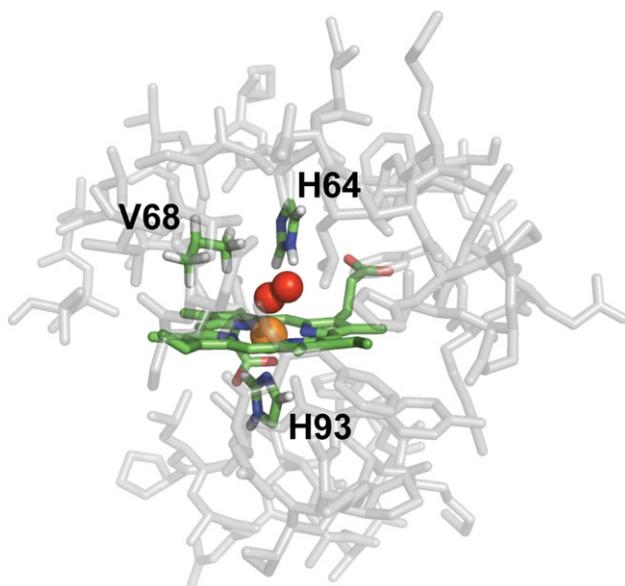
**Figure 5.** (a) Convergence of RMS NPA and selected hydrogen bond  $n_{\text{O}} \rightarrow \sigma_{\text{X-H}}^* \Delta E^{2\text{nd}}$  interactions with respect to psinc kinetic energy cutoff (relative to approximately 1370 eV) for the 80-atom RAD51 BRC4 fragment. (b) As in (a), but for the NGWF radius cutoff (relative to  $15 a_0$ ).

grid) corresponding to a 1000 eV plane-wave energy cutoff in each lattice vector direction, and NGWF radii of  $10 a_0$  for our large protein studies.

## Myoglobin

Myoglobin (Mb) is a small heme protein that is responsible for storing oxygen in muscle tissues. The heme group's Fe 3d orbitals are energetically well-aligned with the  $\pi^*$  acceptor orbitals in  $\text{O}_2$  and CO, and it is capable of strongly binding to these gaseous molecules. The Mb protein reduces the heme group's natural preference for CO binding over  $\text{O}_2$ <sup>[47]</sup>—the  $\pi^*$  acceptor orbitals on  $\text{O}_2$  are lower in energy than on CO, resulting in a greater charge transfer from the Fe 3d orbitals, and hence, a stronger interaction with the surrounding protein, particularly the distal histidine residue (H64, Fig. 6).

To determine how large-scale polarization affects charge transfer to the two ligands, and hence the function of the Mb protein, we use NPA on systems of increasing size using our internal NAO routine in ONETEP. Our systems consist of the heme group, bound to either  $\text{O}_2$  or CO, and surrounded by 1, 3, and 53 Mb residues, with the largest system containing 1007 atoms. Each system has been geometry-optimized as



**Figure 6.** The 53-residue myoglobin fragment used in this study. The 3-residue subset is depicted in green and contains the heme group, a H93 residue co-ordinated to Fe, and two distal residues identified as being important in ligand discrimination between O<sub>2</sub> and CO, H64 and V68. The 1-residue system contains only the heme group and H93. Fe and O<sub>2</sub> are represented as orange/red spheres. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

discussed elsewhere,<sup>[48]</sup> by treating the full system using DFT within ONETEP.

The NPA charge on O<sub>2</sub> is observed to vary depending on the fragment size, converging to a value of  $-0.46 e$  in the 53-residue complex (Table 5), in good agreement with a previous CASSCF/molecular mechanics (MM) study,<sup>[49]</sup> whereas the total charge on CO remains mostly unaffected. The change in the total charge of O<sub>2</sub> is mostly localized on the oxygen atom closest to the H64 residue, to which it forms a hydrogen bond. The charges of the N–H group in H64 also appear to be strongly affected by the inclusion of the protein environment, varying by as much as  $0.1 e$  and resulting in a more positive net charge on the N–H bond in the larger fragment. Both results indicate the necessity of including a significant proportion of the environment when investigating local properties that are strongly influenced by charge transfer and polarization.

### Charge transfer between protein and surface water

Proteins are typically found in a solvated environment and, as such, the presence of water on the surface of protein systems is a crucial component in investigations of their function. A recent study by Ufimtsev et al.<sup>[50]</sup> illustrates the importance of treating surface waters explicitly in a QM calculation, where a pronounced charge transfer ( $\sim 3 e$ ) was observed from the Bovine pancreatic trypsin inhibitor

protein to the surface water, reducing the net charge of the protein from its gas-phase value of  $+6 e$ . Such quantum phenomenon would not have been accounted for in classical or quantum mechanics/molecular mechanics (QM/MM) approaches. The explicit surface water was also shown to be involved in compensating net charges that would otherwise develop around neutral residues in the gas phase due to intra-protein polarization/charge transfer.

To investigate the segregation of charges in a similar solvated biomolecular system, we have performed single point calculations on a protein with increasing number of enveloping water molecules followed by NPA using ONETEP. We have selected an inosine triphosphate pyrophosphatase enzyme from *Escherichia coli* (PDB: 1K7K) for this study on the basis of its size (3125 protein atoms after addition of hydrogen atoms), its high-resolution crystal structure (1.5 Å), and its net charge ( $-7 e$ ).<sup>[51]</sup> The protein was solvated by a box of TIP3P water molecules, equilibrated at 300 K using the AMBER 10 molecular dynamics package with frozen protein atoms, and three systems were extracted for analysis, containing the closest 500, 1000, and 2000 water molecules to the protein. We label the first 500 water molecules layer 1, the next 500 layer 2, and the next 1000 layer 3.

The net charge of the protein in the presence of solvent is drastically reduced—the addition of 500 molecules of explicit water causes a charge transfer of  $7.7 e$  from protein to water, leaving the protein slightly positively charged. Further addition of water molecules does not affect this charge transfer, though there is a significant redistribution of charge between the water layers. Even after the addition of 2000 water units, individual molecules in the outermost layer may still carry significant charge (up to  $0.1 e$ ), though the distribution of molecular charges are very similar to that found near the protein (Fig. 7b).

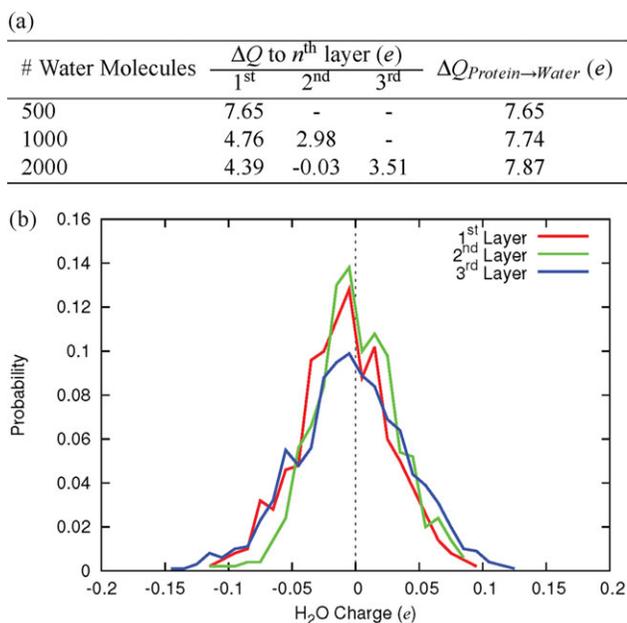
Population transfer was observed to be primarily from the negatively charged residues (Fig. 8), while little change was observed in the average charge distributions for the neutral and positively charged counterparts. Figure 8c shows that the largest charge transfer is from residues at the protein surface, as expected since this is where charged residues are most likely to be located. Such asymmetry is not accounted for in classical force-field simulations, where net residue charges are typically fixed to integer values.

We note that the results of this section are derived using NPA charges, and different population analysis methods, such as electrostatic potential fitting and density-based partitioning approaches (e.g., Bader<sup>[52]</sup> and Hirshfeld<sup>[53]</sup> atoms-in-

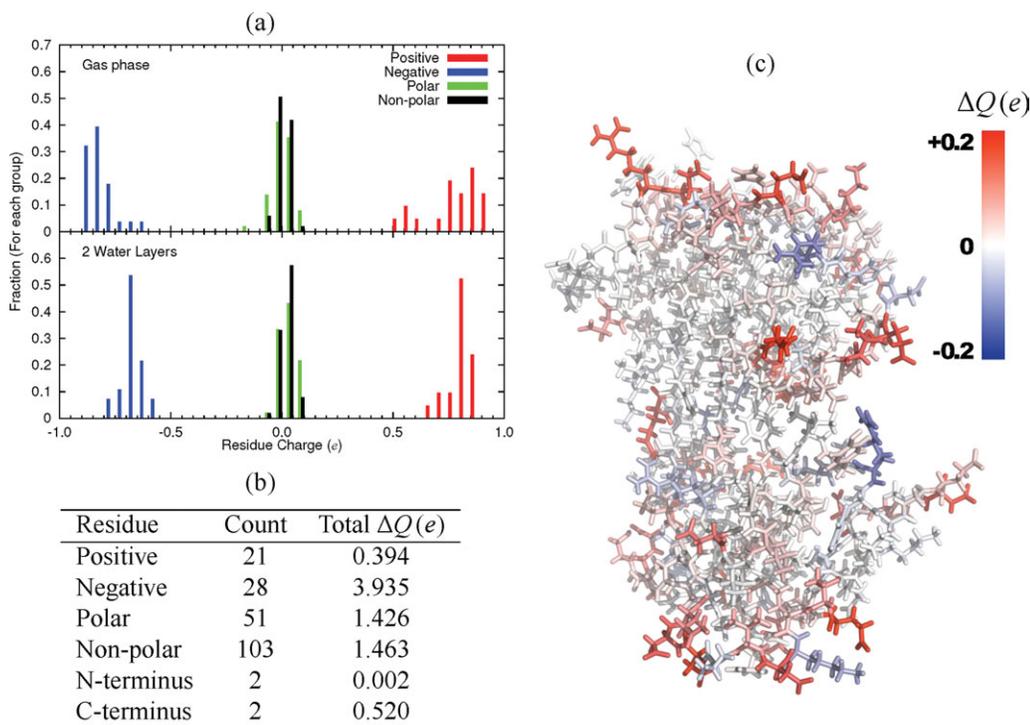
**Table 5.** NPA charges in Fe, O<sub>2</sub>, and CO, and the N–H group of H64 closest to the ligand as a function of system size.

# Residues	Mb-CO				Mb-O <sub>2</sub>			
	Fe	CO	N <sub>H64</sub>	H <sub>H64</sub>	Fe	O <sub>2</sub>	N <sub>H64</sub>	H <sub>H64</sub>
1	0.95	-0.05 (0.45/-0.50)	–	–	1.21	-0.30 (-0.12/-0.18)	–	–
3	0.96	-0.09 (0.42/-0.51)	-0.43	0.42	1.13	-0.37 (-0.18/-0.19)	-0.35	0.38
53	0.39	0.01 (0.50/-0.49)	-0.30	0.43	0.82	-0.46 (-0.20/-0.26)	-0.26	0.37

Values in brackets are charges broken down into atoms close to (Fe/H64).



**Figure 7.** (a) Cumulative charge transfer from protein to water for different number of water molecules surrounding the protein. (b)  $\text{H}_2\text{O}$  charge distribution in the 2000 water molecules system, segregated by layers as defined in (a). Despite having different cumulative charges, all three water layers exhibit similar charge distribution per  $\text{H}_2\text{O}$  unit. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]



**Figure 8.** (a) Histograms of the charge distributions of different residue classes for the protein in gas phase and solvated (1000 water molecules) calculations. The four terminal residues have been excluded. (b) Cumulative charge difference for each residue class between gas phase and solvated systems as in (a),  $\Delta Q = Q^{\text{Solvated}} - Q^{\text{GasPhase}}$ , once again indicating prominent charge transfer from the negatively charged residues. Similar observations are made for terminal residues, where charge transfer from the negatively charged C-terminus is significantly larger than their positively charged N-terminus counterparts. (c) Visual representation of  $\Delta Q$  as defined in (b), colored by residue. Water molecules not shown. Significant changes in residue charges occur close to the protein surface. Terminal residues have been excluded.

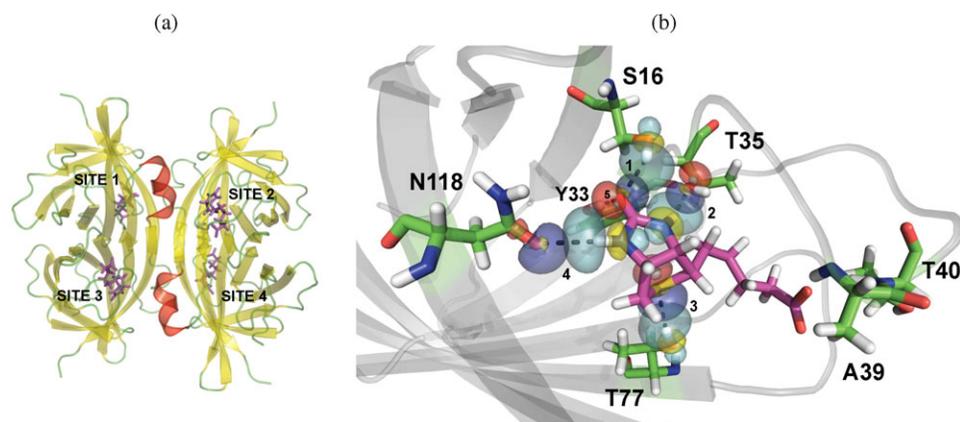
molecules), would most likely give quantitatively different results. However, we would expect the charge distribution trend to remain consistent, provided the method is robust in computing charges of buried atoms, and stable with respect to variables such as basis set size.

### The Avidin-Biotin Complex

Avidin-biotin is one of the most strongly bound reversible protein-ligand complexes in nature, with a binding affinity of  $\sim 20$  kcal/mol,<sup>[54,55]</sup> making investigations into its binding mechanism an active area in computational chemistry. Several studies have highlighted the presence of a strong hydrogen bond network between the biotin molecule and the avidin residues at the binding site, although there is still debate as to whether these are the main contributor to the unusually strong binding affinity of the complex, or merely serve as a recognition pocket.<sup>[55]</sup> In addition, a recent MM investigation<sup>[56]</sup> using fragment-based QM calculations to assign polarized protein-specific charges has demonstrated the importance of electrostatic polarization in stabilizing hydrogen bonds.

In the simulation of avidin-biotin, we are therefore interested only in a small well-defined region of the system, but require accurate long-range electrostatics to determine the correct interactions in the binding pocket as it would appear

in the real system. In this light, we have performed a single DFT calculation on the full 8006-atom complex, followed by internal NPA analysis in ONETEP to compare charges on the biotin molecules, and NBO analysis using NBO 5, focusing only on the geometrically distinct binding pockets 1 and 3 (Fig. 9a), using the selective NBO generation framework as described in Methodology. Our sub-region of interest being passed to NBO 5 consists of the biotin molecule and all nearby residues as depicted in Figure 9b, and the electronic delocalization energetic lowering of individual hydrogen bonds was gauged via second-order perturbation estimates  $\Delta E^{2\text{nd}}$  of  $n \rightarrow \sigma^*$  hyperconjugations between NBO lone pairs  $n$  and acceptor  $\text{X-H}$  antibond  $\sigma^*$  for each donor/acceptor group. The



**Figure 9.** (a) Avidin–biotin tetramer (PDB: 1AVD) showing the four binding sites. Avidin as cartoon, biotin as magenta sticks. Sites 2 and 4 are related to sites 1 and 3, respectively, by rotational symmetry. (b) The third avidin–biotin binding site. Numerical indices label hydrogen bonds. Donor (blue/red) and acceptor (cyan/yellow) NBOs of the strongest  $n \rightarrow \sigma^*$  interaction pair derived from the full system for hydrogen bonds 1–4, plotted with isosurface values of  $\pm 0.06$  a. u. [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

resulting (summed)  $n \rightarrow \sigma^*$  interaction strengths are shown in Table 6 for the donor–acceptor pairs identified in a previous QM study,<sup>[56]</sup> with selected NBOs plotted in Figure 9b. In addition, to investigate environmental influences, fragment calculations were performed on models of sites 1 and 3, containing the biotin molecule and seven of its nearest-neighbor residues truncated along peptide bonds and appropriately terminated with hydrogen atoms [except for Y33 and T35 (Fig. 9b), where the connecting residue T34 was included].

We found that of all possible protein–ligand hyperconjugations at the binding site,  $n \rightarrow \sigma^*$  interactions were the most prevalent, ranging in strength from 4 to 26 kcal/mol. The

strengths of these interactions were strongly dependent on the donor–acceptor distance, and the two binding pockets showed very different behavior, highlighting the inequivalent geometries of the two subunits.<sup>[57]</sup> The hyperconjugation between N118 and the urea group of biotin is observed to be consistently the weakest. Environmental effects are clearly important here, especially for the  $n \rightarrow \sigma^*$  interactions involving T77, N118, and Y33, all of which are observed to increase in the presence of the entire protein. These observations are in good agreement with QM studies of small clusters, where the inclusion of residue D13 was found to significantly enhance the N118–biotin interaction.<sup>[55]</sup> Moreover, environmental influence is also observed in the total NPA charge of biotin, which systematically increases with system size from its isolated value of  $-1.0 e$  (Table 6, bottom), indicating significant changes in electrostatics as the molecules are embedded in the protein.

We emphasize the limitations of the second-order perturbative treatment in estimating hydrogen bond strengths, which was discussed earlier, and the fact that steric effects between analogous  $n$  and  $\sigma$  orbitals, in addition to classical electrostatics<sup>[2,41]</sup> are not accounted for. Therefore,  $\Delta E^{2nd}$  of  $n \rightarrow \sigma^*$  only provides an estimate of the energy-lowering charge transfer component of hydrogen bonds. A more accurate charge transfer energy can be obtained by performing NBO deletion<sup>[1]</sup> on specific or groups of  $\langle n | \hat{H} | \sigma^* \rangle$  Hamiltonian couplings, but this remains an area for future work in ONETEP. Despite these caveats, this analysis is a computationally cheap post-processing step that allows us to conveniently derive an intuitive picture of the types of chemical interactions occurring at local protein–ligand binding sites whilst retaining all long-range electrostatic interactions described by a DFT calculation performed on the whole protein. Observations, such as the weak N118–biotin  $n \rightarrow \sigma^*$  interaction, will make NBO methods more generally applicable in drug-lead optimizations,<sup>[58]</sup> where knowledge of the relative interaction strengths of chemical groups may identify candidates for mutation. Finally, our observations of changes in hydrogen bond hyperconjugation strength and NPA charges in biotin with system size suggest that the inclusion of polarization effects or protein-specific charges in MM force fields<sup>[56]</sup> could improve their transferability.

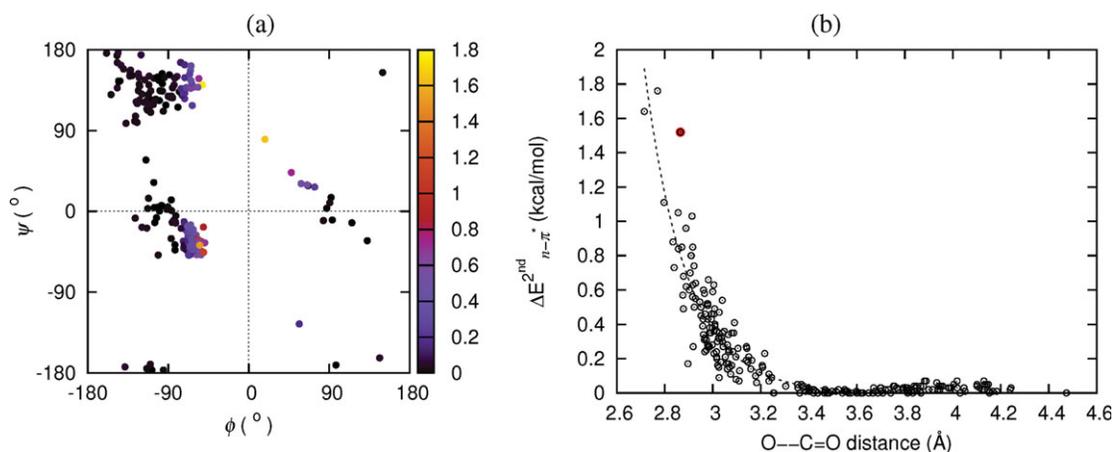
#### $n \rightarrow \pi^*$ interactions in the RAD51–BRC4 complex

We have demonstrated, in the previous section, how hydrogen bond hyperconjugation interactions involving the lone pair NBO,  $n$ , of a donor and the anti-bonding NBO,  $\sigma^*$ , of the acceptor group may be ranked according to their strength

**Table 6.** (Top) Avidin–biotin hydrogen bond strength based on NBO second-order perturbation estimates, summed over all lone pairs  $n$  participating in  $n \rightarrow \sigma^*$  interactions of each donor–acceptor group.

Interaction	$\Delta E^{2nd}$ (kcal/mol)					
	Site 1			Site 3		
	Full	Fragment	$d$ (Å)	Full	Fragment	$d$ (Å)
1	24.4	24.4	3.07	8.5	8.6	2.75
2	12.6	12.0	2.62	23.1	22.9	2.97
3	–	–	–	23.6	20.7	3.17
4	10.7	9.7	3.45	4.3	3.9	3.06
5	26.1	24.4	2.94	14.5	13.6	2.73
Total charge on biotin ( $e$ )						
Site	Isolated Biotin	Fragment	Full			
1		–0.79	–0.64			
3	–1.0	–0.75	–0.68			

Donor–acceptor distance  $d$  of  $X \cdots H - Y$  is measured between  $X$  and  $Y$ . Hydrogen bonds as labeled in Figure 9b. Interaction 3 is absent from site 1 due to the geometric orientation of the  $-OH$  group of T77. (Bottom) Total NPA charge on biotin as a function of system size. Calculations were performed using the static geometry of the full system. The valeric acid side chain in isolated biotin is unprotonated as in the bound system.



**Figure 10.** (a) Ramachandran plot for RAD51-BRC4 with points colored by  $n_O \rightarrow \pi_{C=O}^*$  interaction strength (kcal/mol) from a single DFT calculation on the entire protein complex. (b) Plot of  $\Delta E_{n \rightarrow \pi^*}^{2nd}$  as a function of the O...C=O distance between neighboring amides, with a fitted decaying exponential trendline. The interaction highlighted in red corresponds to the RAD51-BRC4 hotspot interaction depicted in Figure 11a.

estimated via second-order perturbation analysis. Recently, an analogous interaction, involving electronic delocalization from the *p*-rich lone pair of oxygen into the anti-bonding  $\pi^*$  orbital between consecutive protein backbone carbonyl groups, the  $n \rightarrow \pi^*$  interaction, has been shown to influence conformational preferences of small model peptides.<sup>[13]</sup> Furthermore, a survey of the PDB revealed widespread backbone conformations in regions of the Ramachandran plot that correspond to favorable  $n \rightarrow \pi^*$  interactions, potentially stabilizing secondary structures by several kcal/mol.<sup>[12]</sup> However, little is known of the environmental effects on these interactions.

As an example system, we consider here the protein–protein interface between the BRC4 repeat of the BRCA2 protein and RAD51.<sup>[59]</sup> RAD51 is implicated in the error-free repair of double-stranded DNA breaks and is hence necessary for the maintenance of genome stability and ultimately the prevention of cancer.<sup>[60,61]</sup> Our RAD51–BRC4 system (3778 atoms) contains 241  $n \rightarrow \pi^*$  interactions, identified based on geometry alone, specifically the nucleophilic attack angle and distance between the carbonyl oxygen of one residue to the carbonyl carbon of the subsequent residue.<sup>[12]</sup>

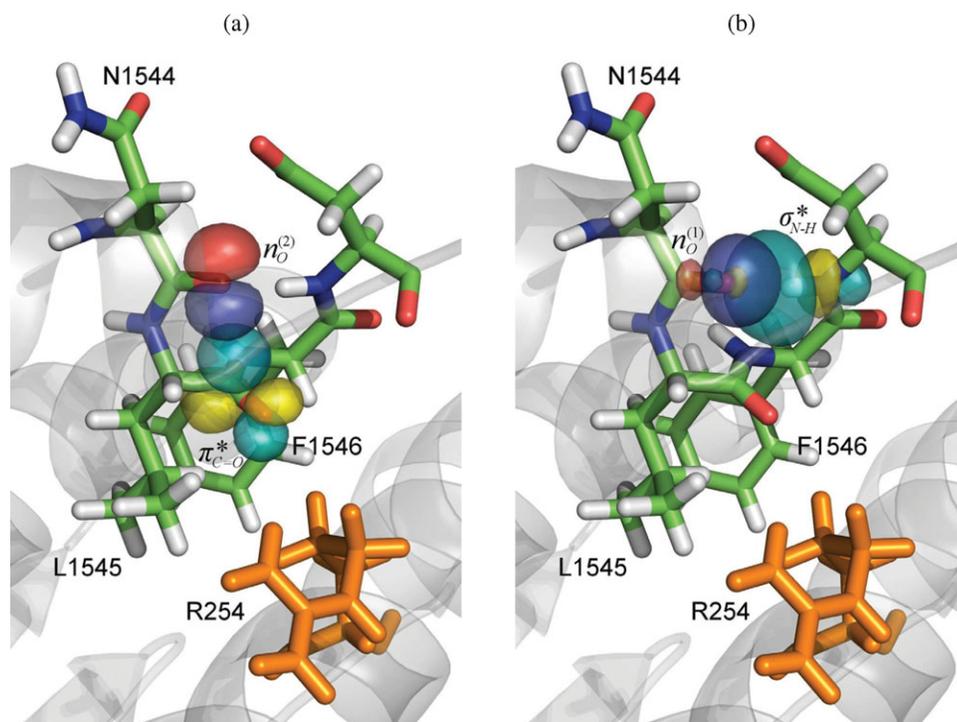
To demonstrate the utility of our NBO 5 interface, we have performed a DFT calculation on the full RAD51–BRC4 system, followed by selective NBO generation for analysis using NBO 5 on regions containing backbone fragments in order to extract the relevant  $n_O \rightarrow \pi_{C=O}^*$  interaction strength estimates,  $\Delta E_{n \rightarrow \pi^*}^{2nd}$ , via second-order perturbation analysis. Through a single linear-scaling DFT calculation on the entire complex, we were able to obtain energetic information at the QM level for each of these interactions separately whilst fully accounting for all long-range interactions with the rest of the protein.

Figure 10a shows the  $\Delta E_{n \rightarrow \pi^*}^{2nd}$  estimates of the strengths of the 241  $n_O \rightarrow \pi_{C=O}^*$  interactions from our DFT calculation of the RAD51–BRC4 complex as a function of the backbone dihedral angles. RAD51–BRC4 contains a high proportion of residues in common secondary structure elements ( $\alpha$ -helices and  $\beta$ -sheets) and the backbone conformations of a number of residues are oriented for substantial stabilization by  $n_O \rightarrow$

$\pi_{C=O}^*$  interactions, in good agreement with previous theoretical predictions.<sup>[12]</sup> We should emphasize that this demonstration is meant to show the potential of NBO-based analysis in ONETEP in decomposing interactions within proteins into recognizable chemical concepts, and not a comprehensive study of the BRC4–RAD51 complex. Hence, for simplicity, solvation effects are ignored.

To further our analysis, we plotted  $\Delta E_{n \rightarrow \pi^*}^{2nd}$  against the O...C=O distance between adjacent amides (Fig. 10b) to identify interactions that are stronger than expected based on their geometry alone, which could indicate an environmental influence. Interestingly, one of the interactions that deviates significantly from the average trend in its interaction strength (highlighted red in Fig. 10b) is located in one of the two identified BRC4 binding hotspots, which contains the tetramer 1545-LFDE-1548.<sup>[62]</sup> The hydrophobic residues L1545 and F1546 are found in a hydrophobic binding pocket in RAD51, yet neighbor a positively charged residue R254. Figure 11a shows the orientation of the hydrophobic L1545 and F1546 residues, the residue R254 on RAD51, and the orbitals corresponding to the  $n_O^{(2)} \rightarrow \pi_{C=O}^*$  backbone interaction between residues N1544 and L1545. The  $n_O^{(2)} \rightarrow \pi_{C=O}^*$  interaction on BRC4 will be strengthened in the presence of R254 if the positively charged residue enhances the acceptor characteristics of the L1545  $\pi_{C=O}^*$  anti-bonding orbital. To test this hypothesis, a fragment representing the BRC hotspot region and the residue R254 (as depicted in sticks in Fig. 11) was extracted from the full RAD51–BRC4 complex and terminated with hydrogen atoms. DFT calculations were performed on this fragment, and one where R254 was separated by an additional 5 Å displacement away from the N1544–L1545–F1546 unit to suppress its influence on the  $n_O^{(2)} \rightarrow \pi_{C=O}^*$  interaction.  $\Delta E_{n(2) \rightarrow \pi^*}^{2nd}$  estimates obtained from the original and separated fragments

<sup>5</sup>There are two lone pairs on a backbone O, an *s*-rich  $n_O^{(1)}$  that participates in backbone–backbone hydrogen bonding, and a *p*-rich  $n_O^{(2)}$  that participates in  $n \rightarrow \pi^*$  interactions.<sup>[12]</sup>



**Figure 11.** The 1545-LFDE-1548 RAD51-BRC4 binding hotspot. L1545 and F1546 of BRC4 reside in a hydrophobic pocket in RAD51. The positively charged R254 residue is positioned close to the BRC backbone and may modulate  $n_{\text{O}}^{(2)} \rightarrow \pi_{\text{C}=\text{O}}^*$  interactions through Coulombic interaction. (a)  $n_{\text{O}}^{(2)} \rightarrow \pi_{\text{C}=\text{O}}^*$  NBOs, plotted with an isosurface value of  $\pm 0.05$  a.u. Donor orbital in red/blue, acceptor in yellow/cyan ( $-/+$ ). (b) As in (a), but for the backbone-backbone (intra-BRC4) hydrogen bond  $n_{\text{O}}^{(1)} \rightarrow \sigma_{\text{N-H}}^*$ . [Color figure can be viewed in the online issue, which is available at [wileyonlinelibrary.com](http://wileyonlinelibrary.com).]

were 1.69 and 1.40 kcal/mol respectively, indicating that R254 does enhance the acceptor character of the L1545  $\pi_{\text{C}=\text{O}}^*$  antibonding orbital and may be a contributor to the strong binding affinity of the RAD51-BRC4 interaction hotspot. Meanwhile, the counterpart  $n_{\text{O}}^{(1)} \rightarrow \sigma_{\text{N-H}}^*$  interaction (Fig. 11b) was virtually unaffected, with  $\Delta E_{n(1)}^{2\text{nd}} \rightarrow \sigma^*$  estimates of 6.01 and 6.03 kcal/mol in the original and separated fragments, respectively.

Whether hyperconjugation is mainly responsible for the optimal orientation of the BRC4 protein toward the RAD51 binding pocket, and whether similar co-operative interactions are found elsewhere in nature are questions for further investigation. It is worth noting, however, that R254, the residue responsible for  $n_{\text{O}} \rightarrow \pi_{\text{C}=\text{O}}^*$  enhancement, is evolutionarily conserved (as positively charged arginine or lysine) across RAD51 orthologues in eukaryotes and archaea, which share a similar domain organization. In eukaryotes, regulators like BRCA2 may bind at this hotspot. Interestingly, R254 is not conserved in the E-coli RAD51 orthologue RecA, which shares a different domain organization and neither binds BRC repeats nor has an obvious BRCA2 orthologue.<sup>[62]</sup> These observations are supportive of our findings from the ONETEP/NBO 5 interface that R254 may be important in BRCA2 binding.

## Conclusions

First principles electronic structure calculations provide very accurate descriptions of matter at the electronic and atomistic

level but are computationally very demanding. Approaches such as NBO analysis which extract physical and chemical properties from the quantum description of the system can provide valuable insights which are not directly available from the electronic wavefunction. In the case of the linear-scaling DFT code ONETEP, this information is contained in a set of *in situ* optimized NGWFs, each of which extends over many atoms, and the density kernel, which is the representation of the density matrix in the duals of the NGWFs. Our interface between ONETEP and NBO 5 software allows transformation of the ground state quantum mechanical wave function into localized chemical orbitals, each of which can be identified as Lewis-type bond or lone pairs according to the classical notion of chemical bonding, supplemented by

their formally vacant antibond and Rydberg counterparts. Properties associated with working in a localized chemical basis, such as NPA charges and hyperconjugation interactions become accessible and greatly enhance the information content of a single DFT calculation.

The *in situ* optimization of the NGWFs, initialized as either Gaussian-type orbitals or PAOs, which is required for the accurate description of the density matrix with a minimal set of local orbitals, causes them to depart from their pure AO angular symmetry state, an attribute mandated by the NBO method. However, we have shown that the NGWFs do in fact retain much of their AO character, typically  $> 99\%$ , especially when initialized as PAOs, allowing us to adapt the NBO method for use with standard ONETEP calculations.

We have implemented the NGWF-to-NAO transformation internally in ONETEP, allowing us to perform NPA on systems with thousands of atoms. Selectively passing relevant partial matrices, after transformation into the NAO basis, to NBO 5 allows us to conveniently study chemical interactions within a local region whilst retaining all long-range effects from the full system treated at the DFT level. The NAOs and NPA charges obtained from our NGWF-to-NAO transformation are comparable to similar calculations in GAMESS using a high-quality aug-cc-pvQZ basis, as were the NBOs derived from them.

We have studied the behavior of  $n \rightarrow \sigma^*$  intermolecular hyperconjugation interactions via second-order perturbation estimates  $\Delta E^{2\text{nd}}$ , which were observed to be basis-set dependent, both in GAMESS and ONETEP but in both cases converge with

increasing size of the orbital sets. Despite the non-convergence of the electronic delocalization energetic estimate in hydrogen bonds when using a minimal number of NGWFs in ONETEP, a strong correlation is observed with converged  $\Delta E^{2nd}$  from a cc-pVTZ calculation in GAMESS, allowing us to compare relative hyperconjugation strengths with confidence.

Finally, we have demonstrated some of the potential applications of the ONETEP/NBO 5 interface with our investigation of several large, > 1000 atom protein systems. We have shown the need for considering a significant proportion of the protein environment when calculating properties that are influenced by the electrostatics of the surroundings, such as NPA charges of ligands bound to the heme group of myoglobin or the  $n \rightarrow \sigma^*$  hyperconjugations between biotin and the avidin binding pocket. In addition, we have shown that the presence of surface water on a protein induces a pronounced redistribution of charges. We have also performed an *in situ* calculation of  $n_O \rightarrow \pi_{C=O}^*$  hyperconjugation interactions in the protein-protein complex between RAD51 and BRC4, confirming expectations of strong interactions in areas of the Ramachandran plot corresponding to common secondary structures. We have determined a possible stabilization mechanism at one of the RAD51-BRC4 energetic hotspots via enhancement of the electron acceptor character of a BRC4 backbone  $\pi_{C=O}^*$  NBO by a charged arginine residue on RAD51, which in turn strengthens the hyperconjugative interactions in the BRC backbone. Such structural stabilization mechanisms are avenues of further research.

Through calculations such as these, NBO analysis can be used to provide chemical insights from large-scale DFT calculations on entire biomolecules, nanostructures, and other materials of high technological interest comprising thousands of atoms.

## Acknowledgments

The authors thank Prof. Weinhold, Prof. Glendening, and their associates for their technical support in updating routines in the NBO 5 analysis package, which facilitated our interface with ONETEP. They are also grateful to Eeson Rajendra for helpful discussions and Nicholas Hine for ONETEP-related technical consultations. Computational resources were provided by the Cambridge HPC Service.

**Keywords:** NBO · ONETEP · linear-scaling DFT · biomolecule · hyperconjugation

How to cite this article: L. P. Lee, D. J. Cole, M. C. Payne, C.-K. Skylaris, *J. Comput. Chem.* **2013**, *34*, 429–444. DOI: 10.1002/jcc.23150

- [1] A. D. MacKerell, Jr., B. Brooks, C. L. Brooks, III, L. Nilsson, B. Roux, Y. Won, M. Karplus, In *Encyclopedia of Computational Chemistry*, Vol. 3; P. v. R. Schleyer, N. A. Allinger, P. A. Kollman, T. Clark, H. F. Schaefer III, J. Gasteiger, Eds.; Wiley: Chichester, **1998**; Chapter 'Natural Bond Orbital Methods,' pp. 1792–1811.
- [2] A. E. Reed, L. A. Curtiss, F. Weinhold, *Chem. Rev.* **1988**, *88*, 899.
- [3] E. D. Glendening, J. K. Badenhoop, A. E. Reed, J. E. Carpenter, J. A. Bohmann, C. M. Morales, F. Weinhold, NBO 5.9 (<http://www.chem.wisc.edu/~nbo5>) & the NBO 5.9 Manual, Theoretical Chemistry Institute, University of Wisconsin: Madison, WI, **2011**.
- [4] A. E. Reed, R. B. Weinstock, F. Weinhold, *J. Chem. Phys.* **1985**, *83*, 735.
- [5] B. F. King, F. Weinhold, *J. Chem. Phys.* **1995**, *103*, 333.
- [6] C. Adamo, F. Leij, *J. Chem. Phys.* **1995**, *103*, 10605.
- [7] F. Weinhold, *Nature* **2001**, *411*, 539.
- [8] V. Pophristic, L. Goodman, *Nature* **2001**, *411*, 565.
- [9] H. Roohi, A. Ebrahimi, F. Alirezapoor, M. Hajealirezahi, *Chem. Phys. Lett.* **2005**, *409*, 212.
- [10] J.-W. Song, H.-J. Lee, Y.-S. Choi, C.-J. Yoon, *J. Phys. Chem. A* **2006**, *110*, 2065.
- [11] F. Claeysens, K. E. Ranaghan, N. Lawan, S. J. Macrae, F. R. Manby, J. N. Harvey, A. J. Mulholland, *Org. Biomol. Chem.* **2011**, *9*, 1578.
- [12] G. J. Bartlett, A. Choudhary, R. T. Raines, D. N. Woolfson, *Nature Chem. Biol.* **2010**, *6*, 615.
- [13] C. E. Jakobsche, A. Choudhary, S. J. Miller, R. T. Raines, *J. Am. Chem. Soc.* **2010**, *132*, 6651.
- [14] M. L. DeRider, S. J. Wilkens, M. J. Waddell, L. E. Bretscher, F. Weinhold, R. T. Raines, J. L. Markley, *J. Am. Chem. Soc.* **2002**, *124*, 2497.
- [15] A. Mohajeri, F. F. Nobandegani, *J. Phys. Chem. A* **2008**, *112*, 281.
- [16] R. Ludwig, *J. Mol. Liq.* **2000**, *84*, 65.
- [17] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, M. C. Payne, *J. Chem. Phys.* **2005**, *122*, 084119.
- [18] N. D. M. Hine, P. D. Haynes, A. A. Mostofi, C.-K. Skylaris, M. C. Payne, *Comput. Phys. Comm.* **2009**, *180*, 1041.
- [19] D. J. Cole, C.-K. Skylaris, E. Rajendra, A. R. Venkitaraman, M. C. Payne, *Europhys. Lett.* **2010**, *91*, 37004.
- [20] D. J. Cole, E. Rajendra, M. Roberts-Thomson, B. Hardwick, G. J. McKenzie, M. C. Payne, A. R. Venkitaraman, C.-K. Skylaris, *PLoS Comp. Bio.* **2011**, *7*, e1002096.
- [21] M. W. Schmidt, K. K. Baldrige, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.* **1993**, *14*, 1347.
- [22] C.-K. Skylaris, A. A. Mostofi, P. D. Haynes, O. Dieguez, M. C. Payne, *Phys. Rev. B* **2002**, *66*, 035119.
- [23] E. Prodan, W. Kohn, *Science* **2005**, *102*, 11635.
- [24] A. A. Mostofi, P. D. Haynes, C.-K. Skylaris, M. C. Payne, *J. Chem. Phys.* **2003**, *119*, 8842.
- [25] A. Ruiz-Serrano, N. D. M. Hine, C.-K. Skylaris, *J. Chem. Phys.* **2012**, *136*, 234101.
- [26] D. Sánchez-Portal, E. Artacho, J. M. Soler, *J. Phys.: Condens. Matter* **1996**, *8*, 3859.
- [27] J. P. Foster, F. Weinhold, *J. Am. Chem. Soc.* **1980**, *102*, 7211.
- [28] B. D. Dunington, J. R. Schmidt, *J. Chem. Theory Comput.* **2012**, *8*, 1902.
- [29] C. Landis, F. Weinhold, In *Valency and Bonding*; CUP: Cambridge, **2005**.
- [30] I. Solt, P. Kulhánek, I. Simon, S. Winfield, M. C. Payne, G. Csányi, M. Fuxreiter, *J. Phys. Chem. B* **2009**, *113*, 5728.
- [31] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.
- [32] N. D. M. Hine, J. Dziedzic, P. D. Haynes, C.-K. Skylaris, *J. Chem. Phys.* **2011**, *135*, 204103.
- [33] D. J. Cole, M. C. Payne, L. C. Ciacchi, *Surf. Sci.* **2007**, *601*, 4888.
- [34] S. J. Clark, M. D. Segall, C. J. Pickard, P. J. Hasnip, M. I. J. Probert, K. Refson, M. C. Payne, *Z. Kristallogr.* **2005**, *220*, 567.
- [35] M. D. Segall, *Mol. Phys.* **1996**, *89*, 571.
- [36] M. D. Segall, R. Shah, C. J. Pickard, M. C. Payne, *Phys. Rev. B* **1996**, *54*, 16317.
- [37] W. Zou, nbo2molden, <http://people.smu.edu/wzou/program/n2m.zip> (accessed January 9, 2011).
- [38] G. Schaftenaar, J. Noordik, *J. Comput. Aided Mol. Des.* **2000**, *14*, 123.
- [39] S. Boys, F. Bernardi, *Mol. Phys.* **1970**, *19*, 553.
- [40] P. D. Haynes, C.-K. Skylaris, A. A. Mostofi, M. C. Payne, *Chem. Phys. Lett.* **2006**, *422*, 345.
- [41] E. D. Glendening, *J. Phys. Chem. A* **2005**, *109*, 11936.
- [42] E. D. Glendening, A. Streitwieser, *J. Chem. Phys.* **1994**, *100*, 2900.
- [43] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, N. A. Baker, *Nucl. Acids Res.* **2004**, *32*, W665.

- [44] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, N. A. Baker, *Nucl. Acids Res.* **2007**, *35*, W522.
- [45] D. A. Case, T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, M. Crowley, R. C. Walker, W. Zhang, K. M. Merz, B. Wang, S. Hayik, A. Roitberg, G. Seabra, I. Kolossvary, K. F. Wong, F. Paesani, J. Vanicek, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, D. H. Mathews, M. G. Seetin, C. Sagui, V. Babin, P. A. Kollman; AMBER 10; University of California: San Francisco. **2008**
- [46] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* **2006**, *65*, 712.
- [47] J. S. Olson, G. N. Phillips, Jr., *J. Biol. Inorg. Chem.* **1997**, *2*, 544.
- [48] D. J. Cole, D. D. O'Regan, M. C. Payne, *J. Phys. Chem. Lett.* **2012**, *3*, 1448.
- [49] H. Chen, M. Ikeda-Saito, S. Shaik, *J. Am. Chem. Soc.* **2008**, *130*, 14778.
- [50] I. S. Ufimtsev, N. Luehr, T. J. Martinez, *J. Phys. Chem. Lett.* **2011**, *2*, 1789.
- [51] A. Savchenko, M. Proudfoot, T. Skarina, A. Singer, O. Litvinova, R. Sanishvili, G. Brown, N. Chirgadze, A. F. Yakunin *J. Mol. Biol.* **2007**, *374*, 1091.
- [52] R. F. W. Bader, *Chem. Rev.* **1991**, *91*, 893.
- [53] F. L. Hirshfeld, *Theor. Chim. Acta* **1977**, *44*, 129.
- [54] N. M. Green, *Biochem. J.* **1963**, *89*, 585.
- [55] J. DeChancie, K. N. Houk, *J. Am. Chem. Soc.* **2007**, *129*, 5419.
- [56] Y. Tong, Y. Mei, Y. L. Li, C. G. Ji, J. Z. H. Zhang, *J. Am. Chem. Soc.* **2010**, *132*, 5137.
- [57] L. Pugliese, A. Coda, M. Malcovati, M. Bolognesi, *J. Mol. Biol.* **1993**, *231*, 698.
- [58] W. L. Jorgensen, *Acc. Chem. Res.* **2009**, *42*, 724.
- [59] L. Pellegrini, D. S. Yu, T. Lo, S. Anand, M. Lee, T. L. Blundell, A. R. Venkitaraman, *Nature* **2002**, *420*, 287.
- [60] S. C. West, *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 435.
- [61] A. R. Venkitaraman, *Annu. Rev. Pathol.* **2009**, *4*, 461.
- [62] E. Rajendra, A. R. Venkitaraman, *Nucl. Acids Res.* **2010**, *38*, 82.

---

Received: 12 July 2012  
Revised: 31 August 2012  
Accepted: 7 September 2012  
Published online on 15 October 2012