ARTICLE

Multiple Active Zones in Hybrid QM/MM Molecular Dynamics Simulations for Large Biomolecular Systems

Juan Torras,[a]

A new QM/MM molecular dynamics approach that can deal with the dynamics of large real systems involving several simultaneous active zones is presented. Multiple, unconnected but interacting quantum regions are treated independently in an ordinary QM/MM approach but in a manner which converges to a unique simulation. The multiple active zones in the hybrid QM/MM molecular dynamics methodology (*maz*-QM/MM MD) involves molecular dynamics that is driving the whole simulation with several parallel executions of energy gradients within the QM/MM approach that merge into each MD step. The Ewald-summation method is used to incorporate long-range electrostatic interactions among the active zones in conjunction with periodic boundary conditions. To illustrate and ascertain capabilities and limitations, we present several benchmark calculations using this approach. Our results show that *maz*-QM/MM MD method is able to provide simultaneous treatment of several active zones of very large proteins such as the Cu-4His-ΔC* cage, a self-assembly of a 24-mer cage-like protein ferritin.

Introduction

In the last decade, interest in simulating real systems, in which the environment plays an important role, has led to *ab initio* molecular dynamics (MD) and quantum mechanical/molecular mechanics (QM/MM) calculations on large molecular systems such as globular proteins, molecular aggregates, and large surface-molecule adducts of special interest to catalysis. Unfortunately, computing time and resources required for the electronic structure component grow very quickly with the size of the molecules. Therefore, the ability to simulate by means of ab initio calculations with approximately linear scaling with respect to system has become an important goal to be achieved in theoretical chemistry.

With recent advances in theory and computational techniques, several levels of theoretical methods have been extensively used. Thus, depending on system size, QM, MM, and coarse-grain modeling are widely used for the electronic structure, atomistic simulations, and mesoscopic systems, respectively. As each of these methodologies has its own time-scale, the existence of multiple time-scales in large systems led to the development of new hybrid methodologies such as QM/MM.[1] The hybrid QM/MM approach[2, 3] is based on the general idea that large systems may be partitioned into a chemically active zone where a high level treatment (QM) is required, whereas the remainder of the system only acts perturbatively so that a classical description (MM) is adequate. However, both the size of the QM region and the accuracy of the QM method used in it are constrained by the computational cost scaling, which is non-linear in the number of electrons. Thus, the description of large-sized QM regions requires an efficient low-cost approach for QM calculations. To deal with the scaling problems in *ab initio* methods, several molecular fragmentation approaches for electronic structure calculations have been developed over many years.[4] The basic idea is to divide the system into distinct fragments and obtain the total properties of the system through their fragments. The combination of QM/MM and molecular fragmentation methods has allowed study of larger systems by use of background point charges in the fragmentation quantum chemical method to treat the long-range electrostatics interactions. The combined strategy of fragmentation quantum chemistry and MM, named fragmentation QM/MM, in the electrostatics embedding model was implemented with some variants by several authors, *e.g.*, D&C-QM/MM,[5] FMO-QM/MM,[6] MoD-QM/MM,[7] and FMO-QM/MM MD[8] methodologies. These methods yield not only conformation energies and geometry optimizations, but also enable Born-Oppenheimer molecular dynamics (BOMD).

The logic of partitions may be extended. Most of real systems are so large that the overall system behavior is due to the evolution of several nearly independent active regions, *e.g.*, the study of complex allosteric effects in the protein-ligand field such as cooperative binding,[9] and the study of complex active sites in catalysis.[10] Concurrent treatment of the electronic structure in each one of these active zones (hereafter AZ) in combination with a lowlevel treatment of the environment following a QM/MM approach seems likely to be useful to deal with these problems.

A key point is to handle the interaction of the AZs properly. An initial approach to overcome this problem, was a multi-centered integrated QM/QM technique implemented by Hopkins and Tschumper[11] who used the so-called subtractive QM/MM scheme (mcQM/QM ONIOM). Later, the same authors allowed an overlap among all the centers in the OMC-ONIOM approach.[12] Then a hybrid approach (XO-ONIOM)[13] was proposed as an extended ONIOM method with the ability to combine any number of sub-systems described at any level of theories into a large calculation. Similarly, Kiyota and co-workers[14] proposed a multi-center approach but based on the additive QM/MM scheme (mcQM/MM). That type of approach showed good precision on the weak interactions such as hydrogen bonding and van der Waals for AZs separated by more than 3.5 Å. Recently, a different hybrid QM/QM scheme was proposed which mixes the classic OMC-ONIOM with the fragment molecular orbital (FMO) scheme as the lower layer in the scheme (OMC-ONIOM FMO).[15] This method yielded accurate estimations of intermolecular interaction energies between proteins and ligands. Most of those previous applications were extended to BOMD simulations. However, specific applications that connect MD programs with fragmentation-based QM programs have been being developed over the last decade, *e.g.*, the FMO-MD method developed by Komeiji *et al.*,[16] and, more recently, the FMO-QM/MM MD program of Nagaoka and co-workers.[8] Also, depending on the intended use, general-purpose interface programs such as ChemShell[17] and PUPIL[18-20] may be used to combine several MD and QM programs, even those that implements Fragmentation-based QM.

The focus of this work is on the methodology to address large systems with differentiated (*i.e.*, distinct) AZs that require special attention because their separate evolution is linked to the global system behavior. An example would be a protein with multiple metallic centers such as the ferritin. In it, the metallic ions have been shown to play an important role in protein-protein interactions to form metal-induced self-assembly cages.[21, 22] Another example might be the complex in NADH: ubiquinone oxidoreductase, which plays a major role in the respiratory electron transport chain from the NADH to ubiquinone across the membrane necessary for ATP synthesis.[23] In this process, distinct electron tunneling pathways between neighboring Fe/S clusters were identified and therefore are candidates for dynamical treatment as independent AZs. The proposed methodology might be also especially useful to study the synergies among different active sites of metalloenzymes such as Laccase. This enzyme presents multi-Cu sites that can catalyze the oxidation of a range of reducing substances with the concomitant reduction of O_2. They contain three Cu centers, two of them form a trinuclear Cu cluster onto which O_2 is reduced whereas the last Cu active site oxidizes the reducing substrate and transfers electrons to the former two Cu active sites.[24, 25]

To our knowledge, however, little progress has been made on the multicenter QM/MM MD simulation oriented to largesystem applications. Indeed, in real systems several AZs can coexist that may contribute to the overall system behavior. This paper presents a new methodology able to deal with those large systems. The descriptive name is Multiple Active Zones Quantum Mechanics/Molecular Mechanics Molecular Dynamics, hereafter *maz*-QM/MM MD. This new methodology is based on the QM/MM partition wherein the main interest zones of the simulated system are treated at the QM level. Each of the divided AZs is a distinct QM region of an independent QM/MM MD simulation (in the environment of the rest of the systems). This methodology also allows exploring reaction paths, free energy changes and transition states, similarly to the methods already used in usual hybrid QM/MM MD approach. *e.g.*, the steered molecular dynamics (SMD) of several replicas along the reaction coordinate with a posterior Jarzynski average to obtain the potential of mean force (PMF),[26] or using the umbrella sampling approach with a posterior data treatment to obtain the PMF.[27] However, using *maz*-QM/MM MD approach, the influence of different active sites might be included to explore the reaction path. Each one of these AZ is managed on the fly by a domain identifier module, with the resulting energy gradients of each QM/MM calculation combined into the whole system at every ab initio MD step. That opens the possibility of modifying the quantum regions on the fly during the simulation trajectory. Moreover, fragmentation-based methods might be applied within each AZ depending on its size and the QM program used.

The remainder of this paper is organized as follows; the theory of *maz*-QM/MM MD approach and the treatment of long-range electrostatics issues are outlined first. Next, the results obtained on three test systems using the *maz*-QM/MM MD are presented and discussed. Those systems are, the water-acetone solvation, Ace-Ala$_{16}$-NMe peptide in explicit water, and a fragment of Cu-Ferritin cage protein with several active zones. Finally, the paper concludes with a summary of the key results to figure out availabilities and limitations of this approach.

Theory

The overall theoretical treatment of multiple AZs which follows is in the context of QM/MM approach.[1] Thus, the solitary QM zone involved in the original partitioning scheme is extended to be a set of disjoint AZs, each treated at the QM level.
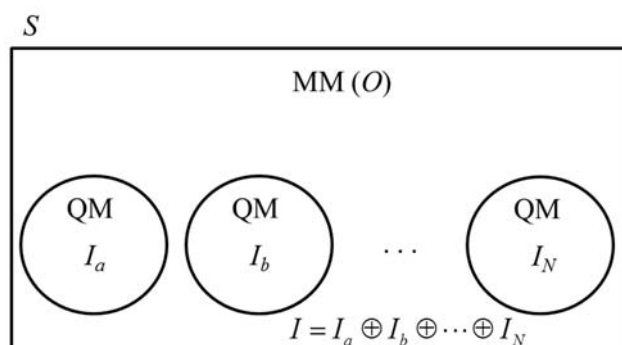

Fig. 1. Partition scheme of the system (S) in an outer region (O) and N unbounded inner regions (I)

QM/MM Partition of the Potential Energy

The entire system (S) is partitioned into the inner region (I) that is treated by means of quantum mechanics and the outer region (O) that is described by a force field, also referred to as QM and MM regions, respectively. A general scheme of system division into QM and MM parts is shown in Fig. 1. The energetics of the two main regions are modeled according to the basic energy expression for an additive QM/MM scheme, namely

$$E(S) = E_{QM}(I) + E_{MM}(O) + E_{QM-MM}(I,O) \qquad (1)$$

Here the subscript QM refers to the region of the system treated by means of QM methods, and conversely for the subscript

MM (region treated classically using force-field-based MM methods). The last term, E_{QM-MM}, collects the interaction terms between the QM and MM regions, which typically includes van der Waals, electrostatic, and bonded interactions across the region boundaries. The electrostatic QM/MM interactions are implemented according to the electrostatic-embedding scheme[28]

$$E_{QM-MM}(I,O) = E^{vdw}_{QM-MM}(I,O) + E^{el}_{QM-MM}(I,O) + E^{bonded}_{QM-MM}(I,O) \quad (2)$$

In the present approach, the QM region is defined as the sum of several disjoint QM subregions. Following the general QM/MM approach, the energy and gradients of any particle within a QM subregion are treated by QM but the interactions with the other QM subregions are treated the same as the other MM region. In the present implementation, QM particles from the other QM subregions are incorporated in the MM region as Mulliken charges. Note that the use of Mulliken charges is a technical choice, not an essential feature. This approach is similar to the one previously proposed by Kiyota et al.[14] However, in this work the sequential QM/MM optimization previously proposed in the bibliography is extended to a more general approach in a parallel execution of multiple QM/MM MD simulations of distinct AZs.

Thus, the energy partition of the QM region is formulated as follows,

$$E_{QM}(I) = \sum_a E_{QM}(I_a) + \frac{1}{2}\sum_a \sum_{b \neq a} E_{QM-QM}(I_a, I_b) \quad (3)$$

Since we are dealing with unconnected QM subregions, the coupling term between quantum zones will contribute only the van der Waals and electrostatic interactions between QM atoms of different subregions.

$$E_{QM-QM}(I_a, I_b) = E^{vdw}(I_a, I_b) + E^{el}(I_a, I_b) \quad (4)$$

Here, each one of the QM subregions will be treated identically to the ordinary QM/MM method. However, the electrostatic interactions between the different QM subregions are handled by the electrostatic-embedding scheme. Tus, several QM/MM calculations are performed concurrently, one for each QM subregion. It is polarized not only by the conventional MM region, but also by sets of point charges, each of which represent one of the other QM subregions. Thus, the electrostatic interaction between different subregions of eqn (4), $E^{el}(I_a, I_b)$, is approximated by the interaction between the electron density of one QM region with a grid of charges of the other (Q_b), e.g., Electro-Static Potential (ESP), RESP, and Mulliken charges, to simulate the charge polarization of the quantum subregion instead of using the electronic density itself. As long as two quantum regions have non-overlapping charge distributions, their classical interaction energy can be described exactly by an infinite expansion in terms of multipoles. That energy leaves out the anti-symmetrization of the electrons in the two regions but the assumption is that the regions are well-enough separated to make that quantum mechanical contribution negligible. However, the interaction energy of region I_b upon region I_a, involving an infinite expansion in terms of multipoles of $I_b(Q_b^\infty)$ interacting with the quantum system in $I_a(\rho_a)$, can be expressed by the eqn (5a).

$$E^{el}(I_a : \rho_a, I_b : \rho_b) \cong E^{el}(I_a : \rho_a, I_b : Q_b^\infty) \quad (5a)$$

$$\begin{aligned} &E^{el}(I_a : \rho_a, I_b : \rho_b) \\ &\cong \frac{1}{2}\left(E^{el}(I_a : \rho_a, I_b : Q_b) + E^{el}(I_a : Q_a, I_b : \rho_b)\right) \end{aligned} \quad (5b)$$

Nonetheless, in the present approach I_b subregion is represented with a finite number of multipoles (Q_b) and eqn (5a) is not symmetric in the interchange of I_a and I_b. Therefore the gradients produced in eqn (5a) are not the same upon interchange of I_a and I_b. To avoid this problem the symmetrization of eqn (5a) is required as it is shown in eqn (5b).

Finally, the system energy is taken as the energy of N independent QM subregions plus their corresponding QM/QM and QM/MM coupling terms. Eqn (1) can be reformulated taking into account the effective energy at each QM region independently,

$$E_{QM-QM}(I_a : \rho_a, I_b : Q_b) =$$

$$E^{vdw}(I_a, I_b) + \langle \psi_a | \sum_b \sum_i \frac{Q_b}{r_{bi}} | \psi_a \rangle + \sum_a \sum_b \frac{Z_a Q_b}{R_{ab}} \quad (6a)$$

$$\begin{aligned} E(S) = &E_{MM}(O) + \sum_a^N \left[E_{QM}(I_a) + E_{QM-MM}(I_a, O)\right] \\ &+ \sum_a^N \sum_{b>a}^N \frac{1}{2}\left[E_{QM-QM}(I_a : \rho_a, I_b : Q_b) + E_{QM-QM}(I_b : \rho_b, I_a : Q_a)\right] \end{aligned} \quad (6b)$$

the interaction energy component between a QM region denoted I_a on a QM region denoted I_b ($E_{QM-QM}(I_a, I_b)$) is different from the interaction energy $E_{QM-QM}(I_b, I_a)$. Thus, the symmetrization of those terms is considered in eqn (6b) analogous with the treatment in eqn (5b).

Notice that the whole approach has the hypothesis that two QM regions have to be enough separated in order that their charge distributions cannot overlap, or if there is any overlapping, this has to be low. In this case, the approximation in terms of a multipole expansion can be performed. Otherwise, this approach is not valid. Thus, only disjoint AZs are allowed in this approach. However, maz-QM/MM MD opens the opportunity to merge two different AZs when they are approaching so much, just a few angstroms (This feature is not yet implemented). Similarly, there is the possibility to split an AZ when there are parts that goes away from each other.

QM/MM Long-Range Electrostatics: Single Active Zone

The treatment of long-range interactions in conjunction with Periodic Boundary Conditions (PBC) is conventional for prediction of condensed system properties. This treatment is much less well-established for hybrid QM/MM MD implementations. However, some recent implementations have been incorporating long-range interactions, most of them being compatible with semi-empirical[29, 30] and DFT QM methods.[31] More recently, some modern software based on a modular approach to wrap several QM and MM packages have relied on a spherical truncation scheme, in which the electrostatic coupling term is neglected beyond a certain cutoff distance R_c.[20, 32] or through a solvent boundary potential to capture long-range electrostatics.[33]

On the other hand, long-range electrostatics based on the QM/MM-Ewald summation methodology were efficiently described by Nam et al.[29] by means of adding periodic

correction term for both QM and QM/MM interactions to the usual real-space electrostatic interaction between QM and MM partitioning. Laino *et al.* have implemented similar methodology based on the multigrid approach using DFTB on the QM region.[31] Similarly, Walker et al.[30] have used the partition proposed by Nam *et al.* for a broad set of semiempirical methods to treat long range electrostatics in a QM/MM MD approach within the Amber package.[34]

The charge distribution in a QM/MM calculation, under PBC, is partitioned between the charge distributions of the QM and MM regions, *i.e.*, the electron density (ρ) of the QM atoms (plus the core nuclear charges) on the former, and the classical MM point charges from the MM environment (q) on the latter charge distribution. Thus the long-range electrostatic interaction energy involving all charge distributions, $E^{PC}(\rho+q, \rho+q)$, is expressed as

$$E^{PC}(\rho + q, \rho + q) = E^{PC}(\rho,\rho) + E^{PC}(\rho,q) + E^{PC}(q,q) \tag{7}$$

where $E^{PC}(\rho,\rho)$ and $E^{PC}(\rho,q)$ terms are the most difficult to be treated (hereafter in the text superscript PC refers to the terms with periodic boundary conditions). However, *ab initio* programs may obtain these terms in a straightforward way in real-space (by means of using a cutoff to select environment point charges around active zone). Therefore, Nam *et al.* proposed an approximation where the periodic energy was obtained as a sum of the nonperiodic energy (as determined by conventional cutoff techniques), E^{RS}, plus a periodic boundary correction term, ΔE^{PC} ($E^{PC}(\rho,\rho)=E^{RS}(\rho,\rho)+ \Delta E^{PC}(\rho,\rho)$). The next key step was the approximation of electron density as a set of auxiliary point charges, Q, such as the Mulliken charges (eqn (8) and (9)). However, other charge partitions may be used in order to emulate the electrostatic potential that represents the full QM charge distribution at the distances between periodic images.

$$E^{PC}(\rho + q, \rho + q) =$$
$$[E^{RS}(\rho,\rho) + \Delta E^{PC}(\rho,\rho)] + [E^{RS}(\rho,q) + \Delta E^{PC}(\rho,q)] + E^{PC}(q,q) \cong$$
$$[E^{RS}(\rho,\rho) + \Delta E^{PC}(Q,Q)] + [E^{RS}(\rho,q) + \Delta E^{PC}(Q,q)] + E^{PC}(q,q) \tag{8}$$

where,

$$\Delta E^{PC}(Q,Q) =$$
$$E^{PC}(Q,Q) - \Delta E^{RS}(Q,Q) = E^{PC}(Q,Q) - \sum_{i}^{QM}\sum_{j>i}^{QM} Q_i Q_j \frac{1}{r_{ij}}$$
$$\Delta E^{PC}(Q,q) = \tag{9}$$
$$E^{PC}(Q,q) - \Delta E^{RS}(Q,q) = E^{PC}(Q,q) - \sum_{i}^{QM}\sum_{j}^{MM} Q_i q_j \frac{1}{r_{ij}}$$

The foregoing equation will leave the periodic boundary correction only for the point charges distributions, Q and q, whereas the exact quantum mechanical charge distribution is evaluated in the real space only.

So far the terms $E^{RS}(\rho,\rho)$, and $E^{RS}(\rho,q)$ of eqn (8) are obtained from the ab initio program, while the term $E^{PC}(q,q)$ is derived directly from the MD program used. What is left to be determined are the terms $\Delta E^{PC}(Q,Q)$ and $\Delta E^{PC}(Q,q)$, which can be derived using Ewald summation.[35] Thus, the periodic energy of a set of point charges distribution such as the Mulliken charges (Q) from a quantum zone is given by,

$$E^{PC}(Q,Q) = \frac{1}{2}\sum_{i,j}^{QM}\sum_{\mathbf{n}=0}^{\infty}{}' \frac{Q_i Q_j}{|\mathbf{r}_{ij} + \mathbf{n}|} \tag{10}$$

where $\mathbf{r_{ij}} = \mathbf{r_i} - \mathbf{r_j}$ are the difference between charge positions in the periodic unit cell, and the summation over $\mathbf{n}$ is over all integer translations of the real space lattice vectors $\mathbf{n} = n_1 \mathbf{a_1} + n_2 \mathbf{a_2} + n_3 \mathbf{a_3}$. The prime indicates omission of terms with $i = j$ when $\mathbf{n} = 0$.

Ewald summation is a well-known technique to sum the long-range interactions between their infinite periodic images efficiently. In fact, the sum of the conditionally convergent series in eqn (10) is transformed into the sum of two rapidly convergent series plus a constant term. The three parts, namely, the direct (real) space sum ($E^{direct}(Q,Q)$), the reciprocal (imaginary) sum ($E^{recip}(Q,Q)$), and the constant term, known as the self-term ($E^{self}(Q,Q)$), are expressed in three convergent sums:

$$E^{PC}(Q,Q) = E^{direct}(Q,Q) + E^{recip}(Q,Q) + E^{self}(Q,Q)$$
$$= \frac{1}{2}\sum_{\substack{i,j \\ i\neq j}}^{QM}\sum_{\mathbf{n}}^{|\mathbf{n}|=n_{max}} Q_i Q_j \frac{erfc(\kappa|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|} + \tag{11}$$
$$\frac{2\pi}{V}\sum_{i,j}^{QM}\sum_{\mathbf{k}\neq 0}^{|\mathbf{k}|=k_{max}} Q_i Q_j \frac{\exp(-k^2/4\kappa^2)}{k^2}\cos(\mathbf{k}\cdot\mathbf{r}_{ij}) + \sum_{i}^{QM}Q_i^2\frac{\kappa}{\sqrt{\pi}}$$

where V is the volume of the simulation box, *erfc(x)* is the complementary error function, and $\mathbf{n}$ was defined earlier. $\mathbf{k}$ is the reciprocal lattice vector, $\mathbf{k} = 2\pi\mathbf{m}$, and $\mathbf{m}$ sums over all integer translations of the reciprocal lattice $\mathbf{m} = m_1 \mathbf{a_1}^* + m_2 \mathbf{a_2}^* + m_3 \mathbf{a_3}^*$, being $\{\mathbf{a_i^*}\}$ the set of reciprocal base vectors.

Three parameters, namely n_{max}, k_{max}, and κ control the convergence of the sums in eqn (11). The integer, n_{max}, defines the range of the real-space sum, the integer k_{max}, does the same for reciprocal-space, and κ is the Ewald convergence parameter which determines the relative rate of convergence between the real and reciprocal sums. Usually, κ is chosen such that only the ($\mathbf{n} = 0$) term is required in eqn (11) to obtain the desired level of accuracy, namely the minimum image convention within a spherical cutoff (R_c).[36]

Thus, the boundary periodic corrections terms $\Delta E^{PC}(Q,Q)$ and $\Delta E^{PC}(Q,q)$ can be formulated based on the Ewald summations as follows,[30]

$$\Delta E^{PC}(Q,Q) =$$
$$E^{recip}(Q,Q) + E^{self}(Q,Q) - \frac{1}{2}\sum_{\substack{i,j \\ i\neq j}}^{QM}\sum_{\mathbf{n}}^{|\mathbf{n}|=n_{max}} Q_i Q_j \frac{erf(\kappa|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|}$$
$$= E^{recip}(Q,Q) + E^{self}(Q,Q) + \Delta E^{direct}_{cuttof}(Q,Q) \tag{12}$$
$$\Delta E^{PC}(Q,q) = E^{recip}(Q,q) - \sum_{i}^{QM}Q_i \sum_{j}^{MM}q_j \sum_{\mathbf{n}}^{|\mathbf{n}|=n_{max}} \frac{erf(\kappa|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|}$$
$$= E^{recip}(Q,q) + \Delta E^{direct}_{cuttof}(Q,q)$$

where *erf(x)* is the error function. Notice; that because an atom cannot be both a QM and a MM atom, there is no self-term in $\Delta E^{PC}(Q,q)$ of eqn (12). Also, the presence of pairs of disjoint sets of point charges (such as Q and q) makes it difficult to directly apply the simplification proposed by Sangster and Dixon for an uniform set of charges, *e.g.* Q,[37] which made the

reciprocal space calculation considerably more efficient by re-ordering the terms of $E^{recip}(Q,Q)$ in eqn (11), namely

$$E^{recip}(Q,Q) = \frac{2\pi}{V} \sum_{i,j}^{QM} \sum_{\mathbf{k}\neq 0}^{|\mathbf{k}|=k_{max}} Q_i Q_j \frac{\exp(-k^2/4\kappa^2)}{k^2} \cos(\mathbf{k}\cdot\mathbf{r}_{ij}) =$$

$$\frac{2\pi}{V} \sum_{\mathbf{k}\neq 0}^{|\mathbf{k}|=k_{max}} \frac{\exp(-k^2/4\kappa^2)}{k^2} \left\{ \left[\sum_{i=1}^{QM} Q_i \cos(\mathbf{k}\cdot\mathbf{r}_i) \right]^2 + \left[\sum_{i=1}^{QM} Q_i \sin(\mathbf{k}\cdot\mathbf{r}_i) \right]^2 \right\} \quad (13)$$

However, similar expression may be obtained by trigonometric manipulation of the $E^{recip}(Q,q)$ term, where the disjoint point charge sets Q and q (from QM and MM regions, respectively) can be easily transformed from a double sum over i and j of order $o(n^2)$, into two single sums of order $o(n)$ at each one of the set of point charges, by eliminating all the non-crossed trigonometric terms between the two sets of point charges.

$$E^{recip}(Q,q) = \frac{2\pi}{V} \sum_{i}^{QM} \sum_{j}^{MM} \sum_{\mathbf{k}\neq 0}^{|\mathbf{k}|=k_{max}} Q_i q_j \frac{\exp(-k^2/4\kappa^2)}{k^2} \cos(\mathbf{k}\cdot\mathbf{r}_{ij})$$
$$= E^{recip}(Q+q,Q+q) - E^{recip}(Q,Q) - E^{recip}(q,q) \quad (14)$$

where $(Q+q)$ is referring to a full set containing all QM and MM point charges. Thus, all the reciprocal terms involved in eqn (14) can be obtained by applying eqn (13) at each uniform set of point charges.

The Ewald summation algorithm has a final complexity of about $o(k_{max}^3 n) \approx o(n^{1.5})$. Usually, hybrid QM/MM MD methodology uses ab initio codes that consume roughly 99% of the time used for each ab initio MD step, depending on the quantum region. There are faster implementations of long-range electrostatics, for instance, the particle-mesh Ewald (PME)[38] or cartesian treecode Ewald (CTE),[39] both with $o(n \log n)$ complexity. The use of these algorithms may become crucial in classical MD or even in semiempirical QM/MM MD, though less important due to the difference in computation time involving MM and QM codes. However, the use of standard Ewald summation to compute the QM/MM electrostatic interactions is justified when high-level ab initio QM methods are used, since the latter spends several orders of magnitude more of computing time than the former.

Using eqn (11-14), the long-range electrostatic energy in eqn (8) can be rewritten as,

$$E^{PC}(\rho+q,\rho+q) \cong E^{RS}(\rho,\rho) + E^{RS}(\rho,q) +$$
$$E^{recip}(Q,Q) + E^{self}(Q,Q) + \Delta E^{direct}_{cuttof}(Q,Q) +$$
$$E^{recip}(Q,q) + \Delta E^{direct}_{cuttof}(Q,q) + E^{PC}(q,q) \quad (15)$$

where the first two terms of eqn (15) come from the QM part of the calculation, $i.e.$, the interactions among QM atoms in the direct space and the interaction of QM atoms with MM atoms that are within the cutoff distance of any QM atom. Also, from the third up to seventh terms are the interactions of QM atoms, represented by a Mulliken charge approximation, with their periodic QM and MM images. The last term is the periodic interaction energy among MM particles that comes from the MD part of the calculation.

QM/MM Long Range Electrostatics: Multiple Active Zones

The charge distribution in a QM/MM calculation, involving several AZ and with PBC, can be partitioned following a framework similar to that in eqn (7) for a single AZ. Thus, the long-range electrostatic interaction energy, with all charge distributions from N active zones, is expressed as

$$E^{PC}(\rho+q,\rho+q) = \sum_a^N \left[E^{PC}(\rho_a,\rho_a) + E^{PC}(\rho_a,q) \right]$$
$$+ \frac{1}{2} \sum_a^N \sum_{b\neq a}^N \left[E^{PC}(\rho_a,\rho_b) \right] + E^{PC}(q,q) \quad (16)$$

where ρ is the electronic density of the whole quantum region which in turn is partitioned in N disjoint AZs, each with its own specific electronic density (ρ_i). The first summation is the long-range electrostatic energy extended to each active zone, the second is the coupling term among AZs, and the last term is the interaction energy of all the point charges in the MM region. The periodic energy may be obtained following the same scheme proposed by Nam $et. al.$ for a single AZ (eqn (8) and (9)) but extended to each one of these disjoints AZs. Similarly, the electron density in each AZ can be approximated by a set of auxiliary point charges, Q_i, $e.g.$ the Mulliken charges. The long-range electrostatic energy then can be rewritten as

$$E^{PC}(\rho+q,\rho+q) \cong$$
$$\sum_a^N \left[E^{RS}(\rho_a,\rho_a) + E^{RS}(\rho_a,q) + \Delta E^{PC}(Q_a,Q_a) + \Delta E^{PC}(Q_a,q) \right]$$
$$+ \frac{1}{2} \sum_a^N \sum_{b\neq a}^N \left[E^{RS}(\rho_a,Q_b) + \Delta E^{PC}(Q_a,Q_b) \right] + E^{PC}(q,q) \quad (17)$$

where $E^{RS}(\rho_a,\rho_a)$, $E^{RS}(\rho_a,q)$ and $E^{RS}(\rho_a,Q_b)$ terms are obtained from the QM code in real-space (by means of using a cutoff to select environment point charges around AZ denoted as a) and $\Delta E^{PC}(Q_a,Q_a)$, $\Delta E^{PC}(Q_a,Q_b)$ and $\Delta E^{PC}(Q_a,q)$ terms are obtained by applying eqn (12). However, eqn (17) might be simplified by joining all the $\Delta E^{PC}(Q_i,Q_j)$ terms in a unique calculation involving the whole set of point charges defined on the QM region ($Q = \sum_a Q_a$) at the end of calculation,

$$E^{PC}(\rho+q,\rho+q) \cong \sum_a^N \left(E^{RS}(\rho_a,\rho_a) + E^{RS}(\rho_a,q) + \Delta E^{PC}(Q_a,q) \right)$$
$$+ \frac{1}{2} \sum_a^N \sum_{b\neq a}^N \left[E^{RS}(\rho_a,Q_b) \right] + \Delta E^{PC}(Q,Q) + E^{PC}(q,q) \quad (18)$$

where ρ is the electronic density of the whole quantum region, Q is the whole set of charges on the QM region representing electron density ($e.g.$ Mulliken charges), and q is the whole set of point charges on the MM region. The indices a and b denote the electron density and the set of point charges for each one of the defined AZs.

Technical Details

The maz-QM/MM MD methodology summarized above has been implemented in the PUPIL program.[18-20] It is a generic interface for building a QM/MM MD simulation framework which fully supports the linking of any MD and QM programs in a systematic way. Thus, the main tasks handled in PUPIL are to build the coupling between two external programs through three basic steps; first defining the system partition, second, building the calculations of QM and QM/MM coupling terms

by using either mechanical or electrostatic embedding, and finally, recovering the coupling energy and gradients for combination with their classical counterparts in the MD program. All these tasks are performed automatically and independently of those MD and QM programs involved in the QM/MM MD simulation.
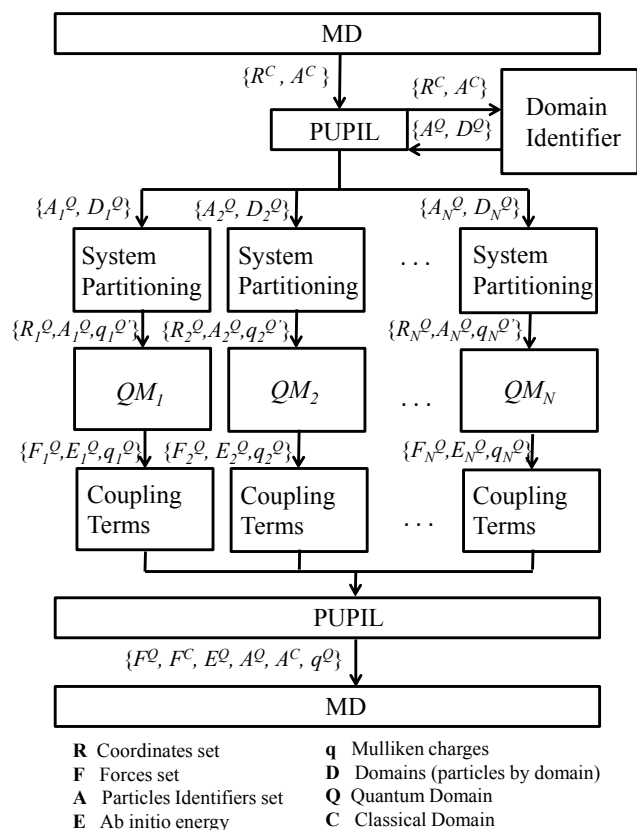

Fig. 2. Flow chart of the maz-QM/MM MD simulation for the distributed processes at a given *ab initio* MD step.

Fig. 2 shows the flow chart for an ab initio MD step which uses the *maz*-QM/MM MD methodology. Initially, when the PUPIL manager receives the classical system coordinate from the MD program the system partition is started. Then, the classical coordinates and the particle-type identifiers are redirected to an external program (Domain Identifier, DI). It is responsible for describing the special extent of each quantum zone as well as how many AZs will be used. (The DI module could be upgraded to a more sophisticated determination of the quantum zone, involving either merger or splitting of AZs on the fly. This feature is not yet implemented.)

One of the main advantages of *maz*-QM/MM MD methodology is easy parallelization. In fact, all QM calculations in the AZs are executed in parallel. Thus, after initial AZ determination, the PUPIL manager starts a set of parallel threads in which the system preparation and the coupling terms calculations (eqn (2-3), (6a), and (18)) determination are conducted independently for each AZ. The $\Delta E^{PC}(Q,Q)$ term in eqn (18) is calculated by the PUPIL manager after completion of the calculations in all the AZs, whereas the $E^{PC}(q,q)$ term is calculated by the MD engine. To provide good performance, most of the terms involved in the Ewald summation are calculated through a parallel execution using Java Native Interface (JNI)[40] combined with native C code.

The Mulliken charges of QM particles are recalculated in each AZ at each QM/MM MD step. These are stored to be reused in the next iteration for calculating the coupling term of eqn (6a) and the long-range electrostatics terms of eqn (18). Following this approach, the electrostatic embedding between different AZs is synchronized because all of the Mulliken charges of QM particles at all times polarize the electron density of the other AZs. This enables the process to be self-consistent along the simulation trajectory.

At each step in each AZ, energy gradients are obtained from the QM engine for both the QM particles and the point charges involved in the electrostatic embedding. Finally, all energy gradients derived for the AZs (QM engine) and for the MM region (MD engine) are combined and returned to the MD program in order to generate new atomic coordinates of the system, whereupon there occurs the next set of QM calculations.

Computational Details

To test and verify the *maz*-QM/MM MD methodology, three different test cases were considered. They provide variety in the size and total charge of each AZ. First, two different systems were used to test the energy stability of *maz*-QM/MM MD simulations; the Cu-Ferritin monomer in a droplet of water and the Ace-Ala$_{16}$-NMe peptide (hereafter A$_{16}$) in a droplet of water and in a simulation box of explicit water under PBC. Second, a QM/MM MD simulation of four neutral molecules of acetone in explicit water (four AZ's) was conducted. The final test was involving several active zones of a large fragment of the Cu-Ferritin cage protein in a box of explicit waters.

System Preparation.

Classical MD simulations were performed to equilibrate all the test systems, *i.e.* acetone, A$_{16}$, and Cu-Ferritin monomer in explicit water. In all cases, the solvent was described using the TIP3P model[41] whereas, all force field parameters for the solute molecules and protein residues were extrapolated from the Generalized AMBER Force Field (GAFF)[42] and the ff03[43, 44] force field, respectively. All MD trajectories were obtained using the AMBER 12 software package.[34]

Initially, all systems in a box of solvent were minimized, heated to 298 K, and equilibrated using the NPT ensemble for 0.5 ns at 1 atm and 298 K (2 fs time steps). Finally, a production run using the same parameters as for the previous equilibration were conducted during 10 ns on the two first tests. All of these steps were performed using the SHAKE algorithm[45] to keep the bond lengths involving hydrogen atoms at their equilibrium distance. The atom pair distance cutoffs were applied at 14 Å to compute van der Waals interactions. Electrostatic interactions were computed using the nontruncated electrostatic potential by means of Ewald summations.

The A$_{16}$ and Cu-Ferritin systems solvated in a droplet of water were prepared similarly to those within a box of solvent. Specifically, a time step of 1.0 fs was used for all simulations and the nonbonded interactions were not truncated. The system energies first were minimized, then the systems were heated, and equilibrated for 1 ns at MM level with Langevin dynamics at 298 K using a collision frequency of 3 ps^{-1}.

The positions of all atoms belonging to what would become QM regions, which involve metallic ions, were restrained by a harmonic potential with a force constant of 20 kcal mol^{-1} Å^{-2} during the equilibration to avoid diffusion toward system boundaries. The last snapshots from all production runs were

used as initial configurations for the subsequent *maz*-QM/MM MD simulations.

Hybrid *maz*-QM/MM MD Simulations.

All hybrid *maz*-QM/MM MD calculations were run using the PUPIL interface,[18, 20] which supports linking, among others, of QM calculations from the NWChem[46] and Gaussian 09[47] programs with MD simulations from the AMBER12[34] program. All atoms involved in the AZs were changed to a QM description in the form of DFT with the B3LYP exchange-correlation functional.[48,49] The basis set were 6-31G, 6-31+G* and an effective core potential LANL2DZ (Los Alamos National Laboratory 2 Double-Zeta).[50] The last-named basis was used on the metal ions, whereas the Pople type basis set was used on all other atoms. On the classical side, all simulation parameters and force fields used for the *maz*-QM/MM MD simulations were the same as those used for system preparation.

Acetone in water. The acetone in water tests used four acetone molecules solvated by a simulation box of 2048 molecules of TIP3P[41] rigid three-site point-charge water molecules. Each acetone molecule was considered as an independent AZ (four AZ's), while all the water molecules remained within the MM framework. The systems were allowed to relax, using the NWChem-PUPIL-Amber[51] interface for 2 ps, followed by a production run of 8 ps (16000 steps, 0.5 fs time step) in an NVT ensemble at 298 K using Langevin dynamics. PBC were applied in the preparation of the NWChem input so as to wrap neighboring point charges around all AZs. Long-range electrostatics of the quantum regions (AZs) were handled as discussed above and with a spherical cutoff scheme to evaluate the van der Waals interactions, whereas the full simulation box was considered for the real-space of electrostatic interactions within the QM/MM coupling methodology. The reciprocal-space cutoff value was set to 16 (maximum integer translation of the reciprocal lattice).
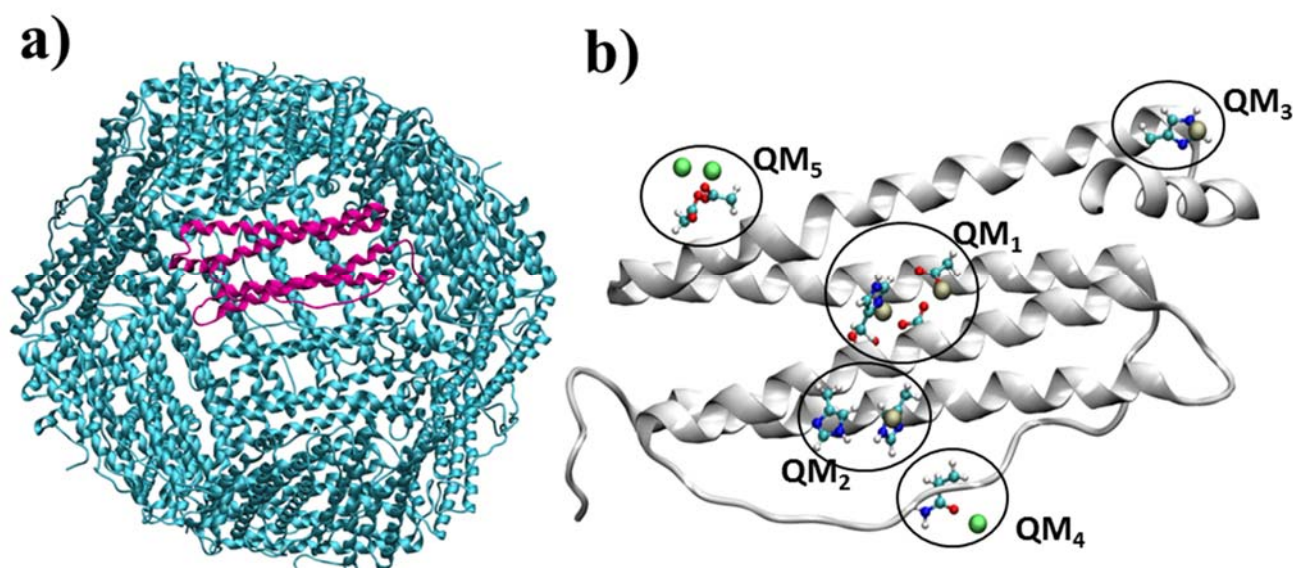

Fig. 3. Cu-Ferritin cage (a) with a selected building block monomer (magenta). (b) Detailed location of the active zones in a monomer of the Cu-Ferritin cage.

Ace-Ala16-NMe peptide. Ace-Ala16-NMe model peptide (A_{16}) was built by imposing an initial secondary structure of 3_{10}-helix. The final system was obtained by solvating A_{16} peptide in two different ways: using a rectangular box with 2113 water molecules and a water droplet with a soft half-harmonic potential beyond 30 Å of radius (3450 water molecules). Two different AZ's were tested: one side chain alone of alanine residue in the QM region, and two alanine side chains separated by a distance of ~16 Å as two independent QM regions. The NWChem-PUPIL-Amber[51] interface with a time step of 0.5 fs was used for all QM/MM MD simulations with bond distances to hydrogen atoms constrained using the SHAKE algorithm.[45] Nonbonded interactions were not truncated for the water droplet simulations. Simulations using PBCs were performed with a cutoff of 14 Å for the real-space, whereas nonbonded interactions and the PME algorithm[38] were considered to account for long-range electrostatics beyond the cutoff (MM only). Long-range electrostatics of the AZs were treated with Ewald summations considering the full simulation box for the real-space of electrostatic interactions (QM/MM only).

Table 1. Description of AZs Considered on the *maz*-QM/MM MD Simulation within a Building Block of Cu-Ferritin Cage.

AZ	Residues	QM Atoms
QM_1	Glu_{27} Glu_{62} Glu_{107} His_{65} Cu_{203} Cu_{204}	31
QM_2	His_{63} His_{67} Cu_{201}	23
QM_3	His_{173} Cu_{202}	12
QM_4	Gln_{86} Ca_{205}	12
QM_5	Asp_{131} Glu_{134} Ca_{206} Ca_{207}	14

Cu-Ferritin Cage Protein. Protein cages such as ferritin[21] are made from a small number of protein building blocks, which by means of their self-interacting surfaces produce highly cooperative, symmetrical structures. Fig. 3a illustrates the Cu-4His-ΔC* cage structure that was obtained by Huard *et al*.,[22] a self-assembly of the 24-mer cage-like protein ferritin controllable by divalent copper binding (4DYX is the Protein Data Bank reference, hereafter Cu-Ferritin cage). The third test on the *maz*-QM/MM MD approach was performed in a real system made of one monomer building block alone from the 24-mer Cu-Ferritin cage solvated in a droplet of water (6154

water molecules) and within a simulation box of explicit waters (12541 water molecules). The Cu-Ferritin monomer contains a total of four Cu^{2+} and three Ca^{2+} metallic ions bonded to several residues of the protein located in five separated regions (Fig. 3b). Thus, five AZs were considered with a total of 92 quantum atoms, as listed in Table 1. Only the side chain of each amino acid residue bonded to a metal ion was turned on as a quantum atom in the AZ.

The periodically bounded systems were allowed to relax within the *maz*-QM/MM MD methodology, using the Gaussian-PUPIL-Amber interface,[20] for 5 ps (10000 steps, 0.5 fs time step) in an NVT ensemble at 298 K with the same parameters

used in the classical MD simulations discussed above. The coordinates along the trajectories in the last 3 ps were saved for subsequent analysis. PBC were applied in the preparation of the Gaussian 09 input so as to wrap neighboring point charges around all the quantum regions.

The Cu-Ferritin system solvated in a droplet of water was allowed to relax using the NWChem-PUPIL-Amber interface,[51] for 5 ps in a microcanonical (NVE) ensemble with a time step of 0.5 fs. Non-bonded interactions were not truncated. Energy drifts were obtained from a linear regression of total energies along the trajectory.
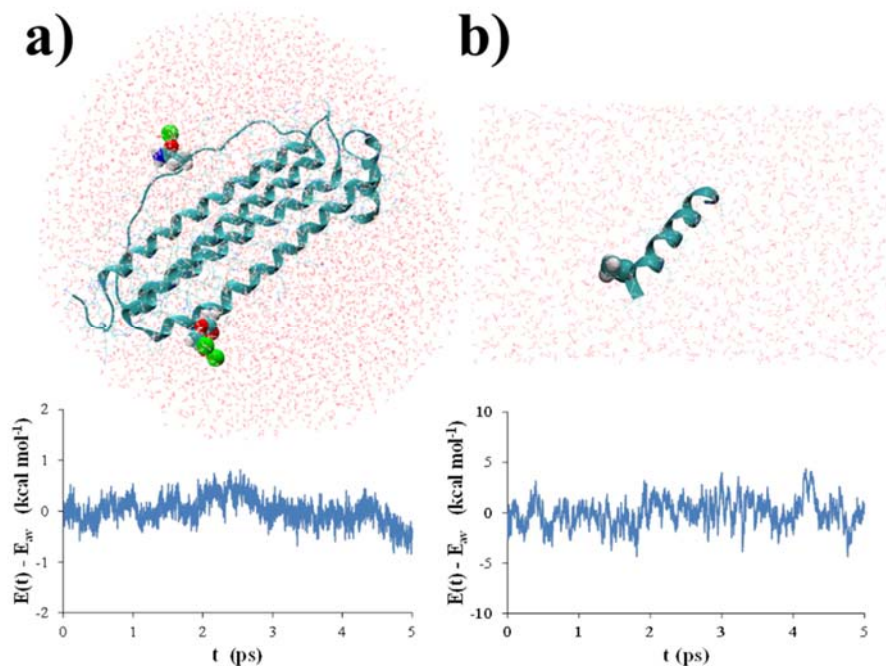

Fig. 4. Energy conservation during constant energy *maz*-QM/MM MD simulations using (a) two active zones of Cu-Ferritin monomer in a droplet of 6154 TIP3P water molecules and (b) Ace-Ala$_{16}$-NMe peptide in a periodic box of 2113 TIP3P water molecules. The QM regions are highlighted and consist of (a) the previously defined QM$_3$ and QM$_4$ active zones in Table 1, and (b) the side chain of the Ala$_3$ residue.

Results and Discussion

Hybrid QM/MM MD Energy Conservation

The energy stability of the *maz*-QM/MM MD method was tested by microcanonical (NVE) ensemble MD simulations for two different systems: a monomer of Cu-Ferritin in a droplet of water, and the Ace-Ala$_{16}$-NMe peptide (hereafter A$_{16}$) under a droplet of water and in a water box with PBC. All simulations were initiated after an initial equilibration with classical MD at 298 K with trajectories of 5.0 ps and a time step of 0.5 fs. The AZ's that were switched to quantum regions in the Cu-Ferritin are defined on Table 1. Thus, the NVE simulations involving one and two active zones within the Cu-Ferritin monomer correspond to the QM$_4$ and QM$_{3,4}$ regions, respectively. On the other hand, the QM region of A$_{16}$ involves the side chains of alanine residues. Hydrogen link atoms are used to saturate the dangling bonds in the QM region. All calculations were performed with the NWChem-PUPIL-Amber interface.

Compared to standard practice, in the QM/MM calculations, tightened SCF convergence criteria (energy, density and gradient $< 10^{-10}$) and an increased accuracy of the numerical quadrature grid for the XC potential were used. The NWChem program was modified to get the energy and forces from the output files with higher than customary precision (error introduced by parsing low precision real numbers in the output text files). Thus, the summation of all the forces on the QM particles and point charges were kept to an acceptably low value ($\sum F_i < 10^{-9}$). Also, the initial guess Fock matrix was obtained from scratch using atomic densities at each MD step in order to avoid systematic bias that could lead to an additional energy drift.

Fig. 4 typifies the energy fluctuations observed for the microcanonical ensemble MD simulations. The total energy drift is assumed linear with time, so it was quantified by fitting the total energy to a linear function, the slopes of which are reported as a stability measure in Table 2. We also tabulate the root mean square fluctuations in the energy after removal of the linear drift term. This RMS term, namely, energy noise, provides a measure of the energy fluctuations that arise solely

from the finite time step. The energy drift should be interpreted as zero unless the product of the drift rate and the total simulation time is significantly larger than the energy noise.[52]

The results show that total energy drift of Cu-Ferritin simulation on a droplet using two AZs remains close to the energy obtained with a single AZ during the simulated time. Also, notice that the energy drifts obtained using standard MD simulations with the same test system and under the same conditions as *maz*-QM/MM MD simulation were -0.076 and -0.053 kcal mol⁻¹ for the Cu-Ferritin and A_{16} on a droplet of water, respectively. Those values are very close to the hybrid simulations values listed in Table 2. Moreover, the energy drifts obtained here are in agreement with those reported with similar QM/MM MD interfaces.[32, 53] Consequently, the inclusion of an additional active zone does not effect to the usual stability of QM/MM MD simulations without PBC. However, the inclusion of infinite images under PBC causes an increase in the total energy drift, such as is observed in the A_{16} system simulation. Nevertheless, the energy noise also increases, thereby making the noise-drift ratios similar to those for simulations without PBC. Observe that inclusion of long-range electrostatic interactions makes the drift increase modestly. The differences of the observed energy drift can be attributed to the unbalanced forces between MM and QM particles when PBC are used. However, for many MD applications using a thermostat an observed small energy drift on the NVE dynamics can be tolerated.

Table 2. Total Energy Fluctuations for 5 ps of NVE QM/MM MD trajectories. Number of Active Zones (AZ) and Long-range Electrostatic interactions on the QM region (LRE) are also shown.

System	AZ's	LRE	ΔE(kcal mol⁻¹)		^{a}kT dof⁻¹ps⁻¹
			Noise	ΔEps⁻¹	
Cu-Ferritin Cap	1	-	0.09	-0.063	1.7×10^{-6}
Cu-Ferritin Cap	2	-	0.25	-0.056	1.5×10^{-6}
A16 Cap	1	-	0.11	-0.079	4.2×10^{-6}
A16 Cap	2	-	0.06	-0.090	4.8×10^{-6}
A16	1	no	1.23	0.168	1.5×10^{-5}
A16	1	yes	1.33	0.342	3.0×10^{-5}

^{a} dof is referring to the degrees of freedom

Results Testing Neutral Active Zones.

The acetone-water test is a conventional MD calculation wherein each of the acetone molecules defines an independent QM subregion within the *maz*-QM/MM MD approach. Fig. 5 shows the temperature and total energy evolution during the last 6 ps of *maz*-QM/MM MD (hereafter quantum simulation) and MD (hereafter classical simulation) trajectories in a NVT ensemble after an equilibration period of 4 ps. No significant differences between the two simulations were observed. Fig. 5b shows that the *maz*-QM/MM MD method provides total energy conservation during the simulation with a standard deviation of 77.8 kcal mol⁻¹, whereas classical MD has a standard deviation of 77.0 kcal mol⁻¹. Similar behavior is observed for the temperature with averaged values of 297.9±4.0 K and 297.2±3.7 K for the quantum and classical simulations, respectively.
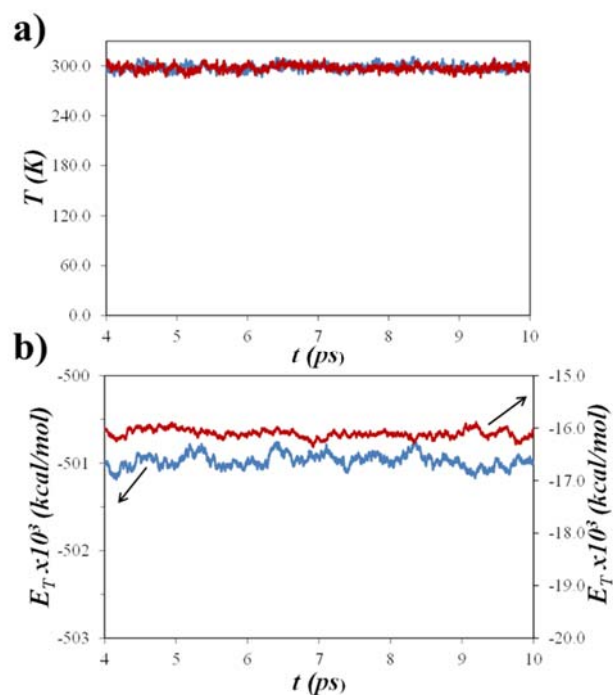

Fig. 5. Time evolution of the (a) temperature and (b) total energy of the acetone-water system. Data were derived from *maz*-QM/MM MD (blue lines) and MD (brown lines).

Fig. 6a displays the radial distribution function (RDF) of the primary acetone molecule relative to all other acetone molecules in the system. There is obvious similarity between quantum and classical RDFs. Indeed, two well defined maxima are obtained at 10.0 and 19.0 Å from the quantum simulation, while the maxima for the classical simulation are located at 11.7 and 19.0 Å. Nevertheless, the quantum simulation shows a close and more pronounced maximum at 10 Å when compared with classical simulation, thus presents a bit more distributed location of acetone molecules.

The RDF of water-hydrogen atoms around the carbonyl oxygen ($g_{C=O\cdots H}(r)$) is given in Fig. 6b. The RDF shows a large environment polarization in the *maz*-QM/MM MD simulation with two clear solvation layers around the carbonyl moiety. The first maximum is located at 1.85 Å, with two secondary and less pronounced maxima at 3.13 and 5.53 Å, respectively. Different behavior is observed in the classical simulation. It gives one clear solvation layer (d_{OH} = 1.90 Å) but much less pronounced secondary solvation layers. The coordination numbers to the carbonyl moiety for both trajectories are 2.5 and 2.0 hydrogen atoms for the quantum and classical simulations, respectively. Those figures indicate a larger number of hydrogen bonding interactions in the quantum simulation. Similarly, the hydrogen bonding analysis (not shown), gives a larger occupancy percentage for the quantum simulation. In fact, the averages of distances of hydrogen bonds, which corresponds to the highest occupancy (>95%) during simulations, are 1.94 and 2.07 Å for the quantum and classical simulations, respectively. Observe that the former value is in agreement with the 1.961 Å reported for high level ab initio calculation on acetone-water clusters.[54]
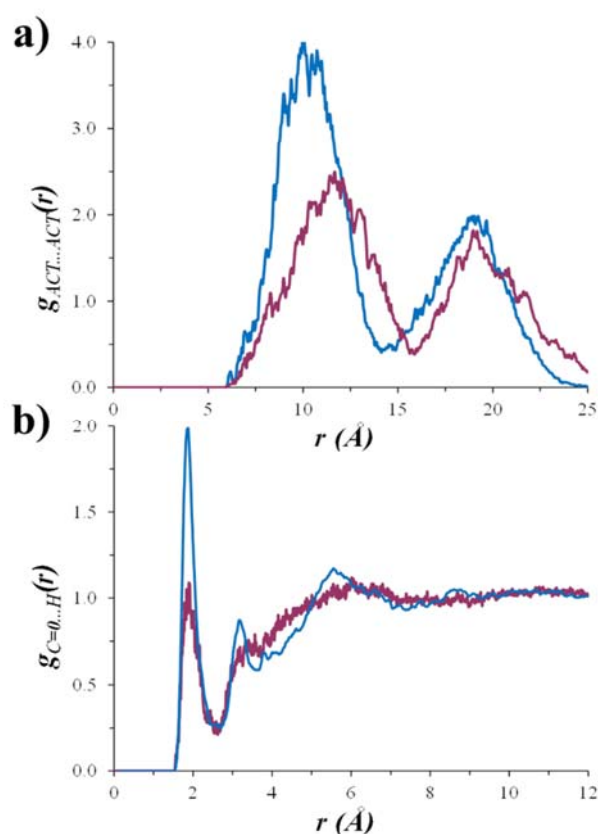
Fig. 6. Radial distribution functions for distances (a) between acetone molecules and (b) between carbonyl oxygen atom and water hydrogen atoms. Data were derived from *maz*-QM/MM MD (blue line) and MD (brown line) trajectories.

This simple test, beside being a basic proof of principle, allowed us to determine the processes that become computational bottlenecks in the *maz*-QM/MM MD approach. Fig. 7 presents the time-consumption partition for each of the subprocesses in one step of QM/MM MD. One sees that on average 93.7% of the time consumption is consumed in the QM calculation, as expected, while the rest is used mainly building the quantum zone embedding. A more detailed analysis of percentages used in the different subprocesses, regardless of the QM calculation time, shows that the main bottleneck lies directly in the construction of the system environment (83.2%) while the electrostatic calculation consumes only about 2.8 % of the time. Thus, an increase of performance at each cycle should focus upon greater efficiency in building the quantum zone model.
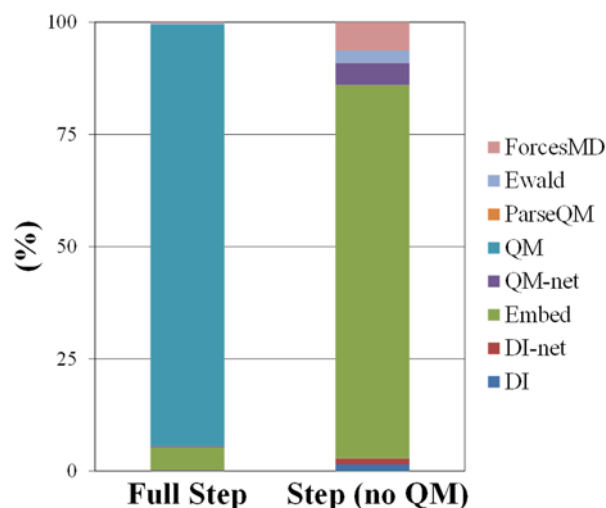

Fig. 7. Percentage distribution of time spent for each subprocess involved within a full QM/MM MD step (first column). Time distribution details of all subprocesses without considering the QM calculation time are also shown (second column).

Table 3. Averaged Distances of the Metal-Residue Coordination of Cu-Ferritin Monomer in Explicit Water (in Å). Experimental Distances from Crystallographic Data are also shown.

AZ	Bond	QM/MM	exp.	Bond	QM/MM	exp.
QM$_1$	Cu$_{203}$ − His$_{65}$@ND1	2.04±0.07	2.165	Cu$_{204}$ − Glu$_{62}$@OE2	1.91±0.05	1.882
	Cu$_{203}$ − Glu$_{27}$@OE1	2.20±0.21	2.006	Cu$_{204}$ − Glu$_{107}$@OE1	2.02±0.08	2.290
	Cu$_{203}$ − Glu$_{62}$@OE1	1.92±0.07	2.096	Cu$_{204}$ − Glu$_{107}$@OE2	3.39±0.13	2.114
QM$_2$	Cu$_{201}$ − His$_{63}$@NE2	2.06±0.07	2.078	Cu$_{201}$ − His$_{67}$@NE2	2.03±0.06	2.113
QM$_3$	Cu$_{202}$ − His$_{173}$@NE2	2.06±0.06	2.189			
QM$_4$	Ca$_{205}$ − Gln$_{86}$@OE1	2.43±0.08	2.280			
QM$_5$	Ca$_{206}$ − Glu$_{134}$@OE1	3.13±0.28	2.406	Ca$_{207}$ − Asp$_{131}$@OD1	2.52±0.21	2.398
	Ca$_{206}$ − Asp$_{131}$@OE2	2.46±0.10	4.897	Ca$_{207}$ − Glu$_{134}$@OE1	2.35±0.08	3.471

Results Testing Cu-Ferritin protein.

For the last simulation a metalloprotein with several well-located AZs was tested. In this system, the metal ions help to stabilize the Cu-Ferritin cage structure. Thus, a simple relaxation that involves up to five active zones with metal ions of Cu-Ferritin was conducted as a proof of principle of the maz-QM/MM MD methodology.

Table 3 lists the averaged distances between the metal ions and the atoms from the coordination group. Results are compared with the X-ray crystallographic data of the Cu-Ferritin cage.[22] Upon relaxation in explicit water, a little reorganization of some of the metal-coordination centers was observed. Indeed, the QM$_2$, QM$_3$, and QM$_4$ AZs did not change much when are compared with experimental coordinates. However, the QM$_1$

and QM$_5$ AZs show some reorganization. Specifically, Cu$_{204}$ in the QM$_1$ region reduces the number of ligands (see Table 3) by means of an increase in the averaged distance to Glu$_{107}$@OE2 atom (d = 3.30 Å). Also, the two calcium ions in the QM$_5$ region are reorganized by exchanging ligands, so that ultimately the Ca$_{206}$ is bonded with the Asp$_{131}$ residue and Ca$_{207}$ is bonded with Glu$_{134}$ and Asp$_{131}$ residues.

Table 4. Water Molecules Coordinated to the Metal Ions of Cu-Ferritin.

AZ	Metal	$d_{max}{}^a$	CN^a	L^b	this work				exp.^{d}	
					$d_{M-O}{}^c$				d_{M-O}	
QM$_1$	Cu$_{203}$	1.80	1	3	1.835				1.878	2.262
	Cu$_{204}$	1.82	2.3	2	1.823	1.834			1.960	1.991
QM$_2$	Cu$_{201}$	1.96	4.0	2	1.934	1.936	2.021	2.191	-	
QM$_3$	Cu$_{202}$	1.86	4.0	1	1.863	1.892	1.913	2.025	2.611	2.727
QM$_4$	Ca$_{205}$	2.47	7.0	1	2.472 2.558	2.493 2.561	2.512 2.616	2.514	-	
QM$_5$	Ca$_{206}$	2.48	7.0	1	2.486 2.562	2.510	2.524	2.542	-	
	Ca$_{207}$	2.51	4.9	2	2.495	2.576			2.373	

^{a} Maximum peak of probability (d_{max}, Å) and coordination number (CN) are derived from radial distribution functions $g_{M-O}(r)$; ^{b} Number of proteins residues as ligands to metal ion; ^{c} Averaged distances between metal ion and water oxygen atom (d_{M-O}, Å) corresponds to a full occupancy of water molecules with $d_{M-O} < 3$ Å during *maz*-QM/MM MD trajectories. ^{d} Reference 22

Generally speaking, solvation effects on the active sites lead to stronger bonds with the main copper ligands and to weaker interactions with calcium ligands. This is evidenced by shortening and elongation of metal-ligand distances on the copper and calcium ions, respectively. The aqueous environment is playing an important role in the stability of the ions and metal-ligand distances. In fact, the solvation effect allows a larger number of coordination waters, in comparison with the crystallographic data, that stabilize the metal ions in addition to the amino acids side chains acting as ligands.

Table 4 lists the coordination number (CN) and the number of water molecules that are situated within a cut-off of 3 Å to a given metal ion during the whole production simulation. The water molecules in the neighborhood of copper ions show distances of ~2 Å and lower. Also, a different coordination is observed, mainly due to steric hindrance, between the tetra-coordinated copper ions on the QM$_1$ AZ and the hexa-coordinated copper ions on the QM$_2$ and QM$_3$ AZ's. Observed d_{M-O} distances are close to the solvation structure of the free copper ion in water, which have a reported CN value of 6 with a distorted octahedral structure (r_{eq} = 1.94 - 2.1 Å, and r_{ax} = 2.27 - 2.60 Å).[55] Different behavior is observed for the calcium ions, which present a higher coordination with the neighbor waters than copper ions. In fact, the total coordination number of the calcium ions with only one protein ligand, *i.e.* Ca$_{205}$ and Ca$_{206}$, is 8. However, steric hindrance on the Ca$_{207}$ with two protein ligands leads a reduction to its CN. The CN values and distances from calcium ion to water oxygen atom are in agreement with the experimental data that show a high CN (6 - 10) with reported distances of 2.39 - 2.46 Å.[56] In addition to examining the coordination of copper ions in the central QM$_1$ active zone we have further investigated the reduced electron density gradient (Fig. 8) using the NCIPlot program.[57, 58] This methodology is able to explain and identify easily the regions with strong and weak electron pairing. What it is shown is a large interaction between the Cu$_{203}$ and the Glu$_{27}$, Glu$_{62}$ and

His$_{65}$ residues, whereas Cu$_{204}$ exhibits two strong interactions with the Glu$_{62}$ and Glu$_{107}$ residues. These interactions are in agreement and further illustrate the results of averaged coordination distances list in Table 3. In future studies, additional research around the coordination's stability and the role that different metal ions might play in the stability of the whole Cu-Ferritin cage structure, would be of interest.
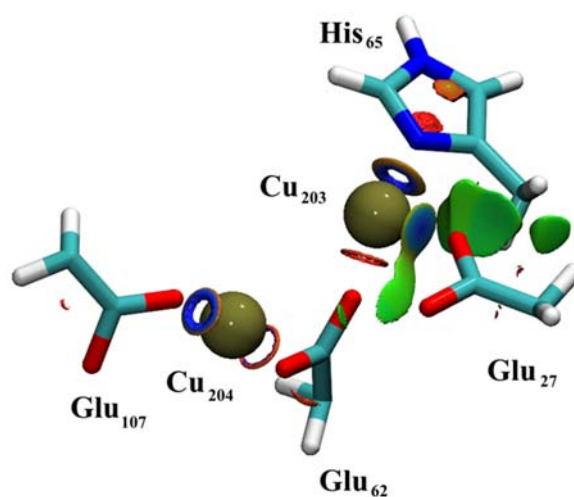

Fig. 8. Noncovalent interactions of the averaged structure of QM$_1$ active zone. Attractive (blue surface), weak (green surface), and repulsive (red surface) interactions are shown.

Conclusions

In this study, a feasible approach to run hybrid QM/MM molecular dynamics of very large molecular systems involving several interdependent active zones has been proposed and tested. The new *maz*-QM/MM MD scheme is based on the QM/MM MD method but instead of a large QM region, the system is partitioned into several unconnected QM active zones. A general molecular dynamics approach drives the simulation but parallel execution of multiple QM/MM techniques, one for each of the AZs are performed at each molecular dynamic step.

The *maz*-QM/MM MD scheme also incorporates the treatment of long-range electrostatic interactions among the AZs in conjunction with PBC by means of the Ewald summations methodology. These interactions have been shown to be necessary to predict the properties of condensed systems, especially those involving electrically charged AZs. Observe that the additional long-range interactions treatment does not introduce an excessive computational cost (2.8% of total ab initio step without the QM calculation).

As proofs of principle, we have tested our approach on three different systems. Energy stability was obtained in a NVE dynamics without PBC, whereas a little drift was observed for those systems when periodic boundary condition is considered. However, in many simulations a small drift on the hybrid simulations can be tolerated by means of using thermostats.

A single building block of the Cu-4His-ΔC* cage structure, a self-assembly of a 24-mer cage-like protein ferritin, has been successfully relaxed in explicit water by means of *maz*-QM/MM MD methodology. This method thus becomes a good candidate for addressing the dynamics of real systems in which there is an interrelationship among different AZs. The synergies among different active zones can be more easily explored using this approach. Concretely, it could facilitate the investigation of the steric effects and long-range effects on the protein environment, revealing the influence of conformational flexibility of proteins on the reaction mechanism, and helping to understand the effects of mutations on the enzymes structure, mechanisms and flexibility. The present study might be used for analyzing the stability of the Cu-4His-ΔC* cage, which holds about 120 AZs with about 50 of them involved on the self-assembly of protein.

Acknowledgements

Notes and references

a Department of Chemical Engineering, EEI, Universitat Politècnica de Catalunya, Av. Pla de la Massa 8, Igualada 08700, Spain. E-mail: joan.torras@upc.edu

1. A. Warshel and M. Levitt, *J. Mol. Biol.*, 1976, **103**, 227-249.
2. H. Lin and D. Truhlar, *Theor. Chim. Acta*, 2007, **117**, 185-199.
3. H. M. Senn and W. Thiel, *Angew. Chem. Int. Edit*, 2009, **48**, 1198-1229.
4. M. S. Gordon, D. G. Fedorov, S. R. Pruitt and L. V. Slipchenko, *Chem. Rev.*, 2011, **112**, 632-672.
5. B. Wang and K. M. Merz, *J. Chem. Theory Comput.*, 2005, **2**, 209-215.
6. H. Li, W. Li, S. Li and J. Ma, *J. Phys. Chem. B*, 2008, **112**, 7061-7070.
7. A. Ghysels, H. L. Woodcock, J. D. Larkin, B. T. Miller, Y. Shao, J. Kong, D. V. Neck, V. V. Speybroeck, M. Waroquier and B. R. Brooks, *J. Chem. Theory Comput.*, 2011, **7**, 496-514.
8. T. Okamoto, T. Ishikawa, Y. Koyano, N. Yamamoto, K. Kuwata and M. Nagaoka, *Bull. Chem. Soc. Jpn.*, 2013, **86**, 210-222.
9. A. Christopoulos and T. Kenakin, *Pharmacological Reviews*, 2002, **54**, 323-374.
10. F. Zaera, *The Journal of Physical Chemistry Letters*, 2010, **1**, 621-627.
11. B. W. Hopkins and G. S. Tschumper, *J. Comput. Chem.*, 2003, **24**, 1563-1568.
12. B. W. Hopkins and G. S. Tschumper *, *Mol. Phys.*, 2005, **103**, 309-315.
13. W. Guo, A. Wu and X. Xu, *Chem. Phys. Lett.*, 2010, **498**, 203-208.
14. Y. Kiyota, J.-Y. Hasegawa, K. Fujimoto, B. Swerts and H. Nakatsuji, *J. Comput. Chem.*, 2009, **30**, 1351-1359.
15. N. Asada, D. G. Fedorov, K. Kitaura, I. Nakanishi and K. M. Merz, *The Journal of Physical Chemistry Letters*, 2012, **3**, 2604-2610.
16. Y. Komeiji, Y. Inadomi and T. Nakanob, *Comput. Biol. Chem.*, 2004, **28**, 155-161.
17. P. Sherwood, A. H. de Vries, M. F. Guest, G. Schreckenbach, C. R. A. Catlow, S. A. French, A. A. Sokol, S. T. Bromley, W. Thiel, A. J. Turner, S. Billeter, F. Terstegen, S. Thiel, J. Kendrick, S. C. Rogers, J. Casci, M. Watson, F. King, E. Karlsen, M. Sjovoll, A. Fahmi, A. Schafer and C. Lennartz, *J. Mol. Struc.-THEOCHEM*, 2003, **632**, 1-28.
18. J. Torras, E. Deumens and S. B. Trickey, *J. Comput. Aided Mater. Des.*, 2006, **13**, 201-212.
19. J. Torras, Y. He, C. Cao, K. Muralidharan, E. Deumens, H.-P. Cheng and S. B. Trickey, *Comput. Phys. Commun.*, 2007, **177**, 265-279.
20. J. Torras, G. d. M. Seabra, E. Deumens, S. B. Trickey and A. E. Roitberg, *J. Comput. Chem.*, 2008, **29**, 1564-1573.
21. X. Liu and E. C. Theil, *Acc. Chem. Res.*, 2005, **38**, 167-175.
22. D. J. E. Huard, K. M. Kane and F. A. Tezcan, *Nat. Chem. Biol.*, 2013, **9**, 169-176.
23. T. Hayashi and A. A. Stuchebrukhov, *Proceedings of the National Academy of Sciences*, 2010, **107**, 19157-19162.
24. F. Xu, J. J. Kulys, K. Duke, K. Li, K. Krikstopaitis, H.-J. W. Deussen, E. Abbate, V. Galinyte and P. Schneider, *Appl. Environ. Microbiol.*, 2000, **66**, 2052-2056.
25. K. Piontek, M. Antorini and T. Choinowski, *J. Biol. Chem.*, 2002, **277**, 37663-37669.
26. J. Torras, G. d. M. Seabra and A. E. Roitberg, *J. Chem. Theory Comput.*, 2008, **5**, 37-46.
27. Z. Futera and J. V. Burda, *J. Comput. Chem.*, 2014, **35**, 1446-1456.
28. D. Bakowies and W. Thiel, *J. Phys. Chem.*, 1996, **100**, 10580-10594.
29. K. Nam, J. Gao and D. M. York, *J. Chem. Theory Comput.*, 2004, **1**, 2-13.

30. R. C. Walker, M. F. Crowley and D. A. Case, *J. Comput. Chem.*, 2008, **29**, 1019-1031.

31. T. Laino, F. Mohamed, A. Laio and M. Parrinello, *J. Chem. Theory Comput.*, 2006, **2**, 1370-1378.

32. A. W. Götz, M. A. Clark and R. C. Walker, *J. Comput. Chem.*, 2014, **35**, 95-108.

33. E. Boulanger and W. Thiel, *J. Chem. Theory Comput.*, 2012, **8**, 4527-4538.

34. D. A. Case, T. A. Darden, T. E. Cheatham, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, R. C. Walker, W. Zhang, K. M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. Seabra, J. Swails, A. W. Goetz, I. Kolossváry, K. F. Wong, F. Paesani, J. Vanicek, R. M. Wolf, J. Liu, X. Wu, S. R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G. Cui, D. R. Roe, D. H. Mathews, M. G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T. Luchko, S. Gusarov, A. Kovalenko and P. A. Kollman, *AMBER 12*, (2012) University of California, San Francisco.

35. P. P. Ewald, *Annalen der Physik*, 1921, **369**, 253-287.

36. A. Y. Toukmaji and J. A. Board Jr, *Comput. Phys. Commun.*, 1996, **95**, 73-92.

37. M. J. L. Sangster and M. Dixon, *Advances in Physics*, 1976, **25**, 247-342.

38. T. Darden, D. York and L. Pedersen, *J. Chem. Phys.*, 1993, **98**, 10089-10092.

39. Z.-H. Duan and R. Krasny, in *Proceedings of the 2003 ACM symposium on Applied computing*, ACM, Melbourne, Florida, 2003, pp. 172-177.

40. S. Liang, *Java Native Interface: Programmer's Guide and Reference*, 1st edn., Addison-Wesley Longman Publishing Co., Inc., 1999.

41. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926-935.

42. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *J. Comput. Chem.*, 2004, **25**, 1157-1174.

43. Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. Wang and P. Kollman, *J. Comput. Chem.*, 2003, **24**, 1999-2012.

44. M. C. Lee and Y. Duan, *Proteins: Struct., Funct., Bioinf.*, 2004, **55**, 620-634.

45. J.-P. Ryckaert, G. Ciccotti and H. J. C. Berendsen, *J. Comput. Phys.*, 1977, **23**, 327-341.

46. M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. de Jong, *Comput. Phys. Commun.*, 2010, **181**, 1477-1489.

47. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.

48. A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372-1377.

49. C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785.

50. P. J. Hay and W. R. Wadt, *J. Chem. Phys.*, 1985, **82**, 299-310.

51. J. G. Warren, G. Revilla-López, C. Alemán, A. I. Jiménez, C. Cativiela and J. Torras, *J. Phys. Chem. B*, 2010, **114**, 11761-11770.

52. J. M. Herbert and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3269-3275.

53. C. M. Isborn, A. W. Götz, M. A. Clark, R. C. Walker and T. J. Martínez, *J. Chem. Theory Comput.*, 2012, **8**, 5092-5106.

54. K. Coutinho, N. Saavedra and S. Canuto, *J. Mol. Struc.-THEOCHEM*, 1999, **466**, 69-75.

55. H. Ohtaki and T. Radnai, *Chem. Rev.*, 1993, **93**, 1157-1204.

56. T. Megyes, T. Grósz, T. Radnai, I. Bakó and G. Pálinkás, *J. Phys. Chem. A*, 2004, **108**, 7261-7271.

57. E. R. Johnson, S. Keinan, P. Mori-Sánchez, J. Contreras-García, A. J. Cohen and W. Yang, *J. Am. Chem. Soc.*, 2010, **132**, 6498-6506.

58. J. Contreras-García, E. R. Johnson, S. Keinan, R. Chaudret, J.-P. Piquemal, D. N. Beratan and W. Yang, *J. Chem. Theory Comput.*, 2011, **7**, 625-632.