Electronic Supplementary Material (ESI) for ChemComm.

SUPPLEMENTARY INFORMATION

Massive quantum regions for simulations on bio-nanomaterials: Synthetic ferritin nanocages

Juan Torras,[*,1] *and Carlos Aleman*[1]

[1]Department of Chemical Engineering (EEBE) and Barcelona Research Center for Multiscale Science and Engineering

Universitat Politècnica de Catalunya,

C/Eduard Maristany 10-14, Ed I2, 08019, Barcelona, Spain

METHODS

maz-QM/MM-MD approach

The multiple active zone scheme for QM/MM-MD calculations (hereafter *maz*-QM/MM-MD approach), was recently developed to study systems with synergetic effects among different zones of interest at simulation time[1] and implemented into the PUPIL program.[2-4] The study of large systems with different regions that have to be treated at high ab initio level, such as large bioinspired nanoscale material systems, are potential targets to be applied by this kind of approach. The methodology is based in the hybrid QM/MM-MD strategy within the Born-Oppenheimer Molecular Dynamics (BOMD) approximation.

The QM/MM strategy is devoted to simplify the simulated system by partitioning the entire system (***S***) into an inner region (***I***) that is treated by quantum mechanics (QM) and the outer region (***O***) described within the Molecular Mechanics (MM) framework with a force field. The energy partition of the two main regions is modeled through the additive QM/MM scheme. Moreover, in the hybrid BOMD the electronic QM/MM equations are solved at each time-step to obtain the potential energy and forces acting on the nuclei of those atoms belonging to the QM region (hereafter active zone or AZ), whereas the energy and forces acting on the nuclei of classical particles are derived through MM force fields.

In the *maz*-QM/MM-MD scheme, the QM region is defined as the sum of several disjoint QM subregions (or active zones, AZs) but following the general QM/MM approach to obtain the energy and its gradients (forces). Thus, any particle within an AZ is subject to QM forces from electronic wave function. However, the interactions with the other AZs are treated by electrostatic interactions (forces from point charges), similarly to the rest of the MM region (see Figure S1). Point charges on those other AZ can be derived as Mulliken or

RESP charges. This procedure is similar to the one previously proposed by Kiyota *et al.*[5] Then, the energy partition of the QM region is formulated as follows,

$$E_{QM}(I) = \sum_{A} E_{QM}(I_A) + \frac{1}{2}\sum_{A}\sum_{B\neq A} E^{vdw}(I_A, I_B) + E^{el}(I_A, I_B) \qquad (1)$$

where the coupling term between atoms belonging to different unconnected AZs is reduced to the van der Waals and electrostatic interactions. More specifically, each atom of the AZ will be treated identically that in the ordinary QM/MM method but running concurrently. However, the electrostatic-embedding scheme will be responsible to handle the electrostatic interactions between the different AZs. The electronic density of each AZ will interact not only with the environment but with a grid of charges of the other AZs (QM subregions) instead of using their own electronic density. At each MD step, all energies and forces from the MM region and all the AZs will be merged to obtain the new system coordinates by integrating the equations of motion.

Compared to previous QM/MM-MD approaches, an important key to face the simulation of large systems using multiple AZs is the reduction of their QM subsystems. Thus, only a fraction of the MM particles (*i.e.* those that are within a cutoff radii from the QM region) are considered as point charges instead of considering the whole MM particles of the system. Indeed, the treatment of long-range interactions in conjunction with Periodic Boundary Conditions (PBC) is conventional for the prediction of condensed system properties but much less well-established for hybrid QM/MM-MD implementation. Long-range electrostatics based on the QM/MM-Ewald summation methodology described by Nam *et al.*[6] were recently incorporated in the PUPIL framework used in this work.[1] Specifically, a periodic correction term is added to the usual real-space electrostatic

interaction between QM and MM partitioning, thus reducing the computation time in each AZ.

Considering only those particles that are close to the QM subregion, the implementation and maintenance of a Verlet list on each AZ leads to a considerable reduction of computational time on the building of all QM subsystems to obtain the QM/MM coupling term at each MD step. All this work is performed in parallel over a distributed system, where each component of remote software has been assigned a unique AZ, taking place an exchange of information with the main simulation manager along simulation time.

A new load-balancing system was implemented to reach a satisfactory resource distribution on the fly to minimize the accumulated time of computation and the largest computational time of an AZ at each MD step. The latter becomes the bottleneck at each MD step due to the necessity to merge all AZs calculations (QM subsystems) to obtain the QM/MM coupling term at a given iteration. The load-balancing methodology is based in the evolutionary algorithm[7] implemented in the Opt4J open source Java-based framework.[8] The phenotype in this work is mapped as a binary interpretation of different ways to group available resources (processors) in a finite number of groups (QM subsystems).

Molecular system

The simulated system is a modified human ferritin,[9] named 4His-ΔC*, with selective response to metal ions, Cu^{2+} controlling the self-assembly necessary for the monomer-to-cage conversion. Thus, 4His-ΔC* implements a novel engineering strategy by means of a metal-templated interface redesign, which allows to transforms a natural protein–protein interface into one that only engages in selective response to a metal ion. More specifically, this system is a recombinant ferritin variant of human H-maxiferritin protein that presents

two additional symmetrically related Cu^{2+} coordination sites in the C_2 interface, is devoid of any cysteine residues, and presents the K86Q mutation for facilitating crystallization.[9] In this work, the structural behavior of the modified interface has been studied using two different transition metal ions in the coordination centers, Cu^{2+} and Ni^{2+}.

Moreover, two different system sizes have been considered in this work: the C_2 interface and the full cage-like protein ferritin. The former consists of two four-helix bundle monomers linked by four metallic coordination sites involving residues of both monomers, while the latter is made by 24 four-helix bundle monomers arranged in octahedral (432) symmetry. Thus, two C_2 interface dimers, here after denoted Cu-dimer and Ni-dimer, were built using Cu^{2+} and Ni^{2+} as coordination transition metals, respectively. Each system was solvated with explicit water molecules under periodically boundary conditions in a box containing 17273 molecules. Similarly, the complete 4His-ΔC* cage, hereafter Cu-cage, was built using Cu^{2+} as coordinating transition metal and solvated with 88638 explicit water molecules under periodic boundary conditions. Coordinates for the 4His-ΔC* cage were taken from 4DYX PDB file.

Molecular Dynamics Simulation Protocol

Classical MD simulations were performed to equilibrate the three modeled systems (*i.e.* Cu-dimer, Ni-dimer, and Cu-cage) in explicit water. In all cases, the solvent was described using the TIP3P model,[10] while all force field parameters for the solute molecules and protein residues were extrapolated from the ff03.r1 libraries.[11, 12] All MD trajectories were conducted using the AMBER 14 software package.[13] Initially, all systems in a box of solvent were minimized, heated to 298 K, and equilibrated using the NPT ensemble for 2 ns at 1 atm and 298 K (2 fs of time steps) up to constant density. Finally, another equilibration

run using NVT ensemble with parameters similar to those of the previous equilibration was conducted during 10 ns for all systems. In the first equilibration both the protein and metallic ions were kept close to the starting structure, meanwhile in the second one only metallic ions and their coordination sphere were restrained using a quadratic harmonic potential with a constant force of 20 kcal mol^{-1}. The SHAKE algorithm[14] was used to keep the bond lengths involving hydrogen atoms at their equilibrium distance. The atom pair distance cutoffs were applied at 10 Å to compute van der Waals interactions. Long-range electrostatic interactions were computed by means of Ewald summations.

maz-QM/MM-MD Simulation Protocol

All hybrid *maz*-QM/MM-MD[1] calculations were run using the PUPIL interface,[2-4] which supports linking, among others, of QM calculations using the NWChem[15] and Gaussian 09[16] programs, and MD simulations from the AMBER14[13] package. All metallic coordination sites were changed to a QM description in the form of DFT with the B3LYP exchange–correlation functional.[17, 18] The basis set was 6-31G and an effective core potential LANL2DZ (Los Alamos National Laboratory 2 Double-Zeta).[19] The last-named basis was used on the metal ions, whereas the Pople type basis set was used on all other atoms. The rest of the system were considered at the classical level, all simulation parameters and force fields used for the *maz*-QM/MM MD simulations were the same as those used for system preparation.

Figure 1b shows the six quantum regions considered on Cu- and Ni-dimers (C_2 pore interface) with a total amount of 188 quantum atoms. The six QM regions were made of two coordination sites involving Cu^{2+} or Ni^{2+} transition metal ions (QM_{C2_4H}), two sites coordinating Ca^{2+} ions (QM_{C2_Ca}), and two coordinated metal regions in the center of each

monomer (QM_{cnt}). On the other hand, the large simulated system involving the full cage-like protein ferritin, Cu-cage, would include a total of 86 quantum regions and 2830 quantum atoms. The full system was comprised of 12 C_2 dimer interfaces (6 QM regions each, Figure 1b), 8 C_3 pore interfaces, and 6 C_4 pore interfaces (Figure 1c and 1d).

Cu-dimer and Ni-dimers were allowed to relax within the *maz*-QM/MM MD methodology, using the NWChem-PUPIL-Amber interface,[20] for 6 ps (12000 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 4 ps were saved for subsequent analysis. The last part of the trajectory presents a variation of ±0.15 Å in the root mean square displacements of those protein residues involved in the AZs. PBC were applied in the preparation of the NWChem input so as to wrap neighboring point charges around all the quantum regions. Long-range electrostatics of the quantum regions were treated with Ewald summations considering a cutoff of 25 Å around each quantum region for the real-space of electrostatic interactions within the QM/MM coupling methodology, meanwhile the reciprocal-space cutoff value was set to 24 (maximum integer translation of the reciprocal lattice). Similarly, the Cu-cage system with a massive number of quantum regions were allowed to relax for 150 fs (300 steps, 0.5 fs time step) in an NVT ensemble at 298 K but using Gaussian09-PUPIL-Amber interface.[4] Ewald summations were considered at each QM/MM coupling term to treat long-range electrostatics interactions of quantum regions with a cutoff of 45 Å and a reciprocal cutoff value of 6 Å. The coordinates along the trajectories in the last 100 fs were saved for subsequent analysis.

1. J. Torras, *Phys. Chem. Chem. Phys.*, 2015, **17**, 9959-9972.
2. J. Torras, E. Deumens and S. B. Trickey, *J. Comput. Aided Mater. Des.*, 2006, **13**, 201-212.
3. J. Torras, Y. He, C. Cao, K. Muralidharan, E. Deumens, H.-P. Cheng and S. B. Trickey, *Comput. Phys. Commun.*, 2007, **177**, 265-279.
4. J. Torras, G. d. M. Seabra, E. Deumens, S. B. Trickey and A. E. Roitberg, *J. Comput. Chem.*, 2008, **29**, 1564-1573.
5. Y. Kiyota, J.-Y. Hasegawa, K. Fujimoto, B. Swerts and H. Nakatsuji, *J. Comput. Chem.*, 2009, **30**, 1351-1359.
6. K. Nam, J. Gao and D. M. York, *J. Chem. Theory Comput.*, 2004, **1**, 2-13.
7. T. Bäck, *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*, Oxford University Press, 1996.
8. M. Lukasiewycz, M. Glaβ, F. Reimann and J. Teich, in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, ACM, Dublin, Ireland, 2011, pp. 1723-1730.
9. D. J. E. Huard, K. M. Kane and F. A. Tezcan, *Nat. Chem. Biol.*, 2013, **9**, 169-176.
10. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *J. Chem. Phys.*, 1983, **79**, 926-935.
11. 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.
12. M. C. Lee and Y. Duan, *Proteins: Struct., Funct., Bioinf.*, 2004, **55**, 620-634.

13. D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. C. III, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, R. C. Walker, J. Wang, R. M. Wolf, X. Wu and P. A. Kollman, *AMBER 14*, (2014) University of California, San Francisco.

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

15. 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.

16. 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.

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

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

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

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

Table S1. Averaged distances (in Å) for the metal-residue coordination from *maz*-QM/MM-MD simulations on Cu- and Ni-dimers. Relevant interactions involving the $QM_{C2\_4H,}$ QM_{C2_Ca} and QM_{cnt} sites of the C_2 interface are shown. Data for the active Zones (AZ) involved at each interface, which were calculated at high computational level, have been compared with experimental values.

Site	AZ	Bond[a]	Cu-dimer	Exp.	Ni-dimer
QM_{C2_4H}	QM1	M(1)-H56'	2.08 ± 0.06	2.11	2.09 ± 0.06
		M(1)-H60'	2.05 ± 0.07	2.08	2.09 ± 0.04
		M(1)-H63	2.14 ± 0.06	2.35	2.10 ± 0.06
		M(1)-H67	2.03 ± 0.07	2.09	2.08 ± 0.07
	QM4	M(4)-H56	2.15 ± 0.11	2.11	2.25 ± 0.09
		M(4)-H60	2.07 ± 0.06	2.08	2.01 ± 0.05
		M(4)-H63'	2.08 ± 0.07	2.09	2.17 ± 0.07
		M(4)-H67'	2.11 ± 0.07	2.35	2.08 ± 0.07
	QM1-QM4	M(1)-M'(4)	10.32 ± 0.16	10.06	10.87 ± 0.35
QM_{C2_Ca}	QM3	Ca(3)-D84'	2.48 ± 0.07	2.36	2.61 ± 0.07
		Ca(3)-D84'	2.50 ± 0.07	2.69	2.68 ± 0.16
		Ca(3)-Q86	2.32 ± 0.07	2.28	2.42 ± 0.12
	QM6	Ca(6)-D84	2.50 ± 0.12	2.27	2.45 ± 0.06
		Ca(6)-D84	2.61 ± 0.11	2.69	3.43 ± 0.59
		Ca(6)-Q86'	2.38 ±0.09	2.28	2.29 ± 0.07
	QM3-QM6	M(3)-M'(6)	9.12 ± 0.31	8.06	9.93 ± 0.70
QM_{Cnt}	QM2	M_1(2)-E27 o1	1.90 ± 0.04	2.01	1.93 ± 0.04
		M_1(2)-E62 o1	1.96 ± 0.08	2.10	1.98 ± 0.06
		M_1(2)-H65	2.00 ± 0.06	2.16	2.01 ± 0.04
		M_2(2)-E106 o1	2.00 ± 0.06	2.11	2.05 ± 0.05
		M_2(2)-E106 o2	2.19 ± 0.09	2.29	2.13 ± 0.09
		M_2(2)-E62 o2	1.97 ± 0.06	1.88	2.02 ± 0.07
		M_1(2)-M_2(2)	4.20 ± 0.32	3.40	3.77 ± 0.06

	M_1(5)-E27' o1	1.91 ± 0.08	2.01	1.92 ± 0.04
	M_1(5)-E62' o1	1.95 ± 0.08	2.10	1.97 ± 0.05
	M_1(5)-H65'	2.09 ± 0.08	2.16	2.15 ± 0.07
QM5	M_2(5)-E106' o1	1.93 ± 0.05	2.11	1.97 ± 0.06
	M_2(5)-E106' o2	2.00 ± 0.06	2.29	2.16 ± 0.09
	M_2(5)-E62' o2	2.09 ± 0.07	1.88	2.04 ± 0.09
	M_1(5)-M_2(5)	4.34 ± 0.33	3.40	3.81 ± 0.17

[a] M is referring to Cu^{2+} or Ni^{2+} metal ion for the Cu- and Ni-dimer, respectively. Single quotes sign is referring to the second ferroxidase monomer. The AZ that belongs each metal ion is noted between parentheses.

Table S2. Occupancy (occ, in %) of hydrogen bonds involving C_2 interface residues. Data obtained from *maz*-QM/MM-MD trajectory for Cu- and Ni-dimer of Cu_4His-ΔC* ferritin are listed. Average H···O distance (d_{OH}, in Å) and average hydrogen bonding angle (α, in º) are also shown. Hydrogen bonds relevant for the stability of the cage are highlighted in red.

Cu-dimer						Ni-dimer					
Acceptor	**DonorH**	**Donor**	**occ**	**d_{OH}**	**α**	**Acceptor**	**DonorH**	**Donor**	**occ**	**d_{OH}**	**α**
D44@OD1	S6'@HG	S6'@OG	100	2.60	166	Q83@O	Y32'@HH	Y32'@OH	100	2.56	167
D44'@OD1	S6@HG	S6@OG	100	2.65	165	I85'@O	I85@H	I85@N	100	2.82	168
Q83'@O	Y32@HH	Y32@OH	98	2.70	162	I85@O	I85'@H	I85'@N	98	2.87	155
D44'@OD2	Q7@HE22	Q7@NE2	97	2.84	162	D84'@OD1	K87@H	K87@N	97	2.87	161
D44@OD1	R79'@HH22	R79'@NH2	95	2.81	147	D44'@OD1	R79@HH22	R79@NH2	92	2.80	152
D44@OD1	R79'@HH12	R79'@NH1	95	2.80	149	Q83'@O	Y32@HH	Y32@OH	83	2.74	152
Y39'@OH	H67@HD1	H67@ND1	87	2.84	161	**Y39'@O**	**N74@HD21**	N74@ND2	**83**	**2.77**	**149**
D42@OD2	**N74'@HD22**	N74'@ND2	85	2.85	162	**Y39@O**	**N74'@HD21**	N74'@ND2	**77**	**2.88**	**154**
Q83@O	Y32'@HH	Y32'@OH	85	2.78	165	D45@OD1	R79'@HH21	R79'@NH2	70	2.77	149
D44'@OD1	R79@HH22	R79@NH2	85	2.83	158	D42'@OD2	**N74@HD22**	N74@ND2	68	2.88	164
D84@OD1	K87'@H	K87'@N	76	2.86	160	**Y39'@OH**	H67@HD1	H67@ND1	62	2.86	145
D84'@OD1	K87@H	K87@N	75	2.84	164	D45'@OD1	R79@HH21	R79@NH2	62	2.81	144
I85@O	I85'@H	I85'@N	74	2.85	161	D44@OD2	Q7'@HE22	Q7'@NE2	59	2.86	163
I85'@O	I85@H	I85@N	72	2.82	160	D44'@OD1	S6@HG	S6@OG	58	2.64	157
K71'@HB2	**Y39@HH**	Y39@OH	59	2.81	153	D44'@OD1	R79@HH12	R79@NH1	44	2.80	151
D44@OD2	Q7'@HE22	Q7'@NE2	58	2.87	159	D44'@OD2	S6@HG	S6@OG	42	2.65	156
D45'@OD2	R79@HH21	R79@NH2	51	2.74	144	D84@OD1	K87'@H	K87'@N	41	2.87	167
Y39@HH	K71'@HB2	K71'@CB	45	2.89	163	D84@OD2	Q86'@HE21	Q86'@NE2	38	2.80	158
D44@OD2	R79'@HH21	R79'@NH2	42	2.79	149	D44@OD1	R79'@HH22	R79'@NH2	36	2.91	165
Y39@O	**N74'@HD21**	N74'@ND2	**38**	**2.90**	**150**	D44'@OD2	R79@HH22	R79@NH2	35	2.86	149
D44'@OD1	R79@HH12	R79@NH1	32	2.89	147	D44@OD1	S6'@HG	S6'@OG	26	2.83	159
Y39'@O	**N74@HD21**	N74@ND2	**25**	**2.93**	**151**	Q83'@OE1	K87@HZ3	K87@NZ	16	2.84	152

H67@O	**Y39'@HH**	Y39'@OH	15	2.76	145	H67@O	**Y39'@HH**	Y39'@OH	11	2.66	141
D44'@CG	S6@HG	S6@OG	15	2.96	146						
D44@OD1	R79'@HH21	R79'@NH2	14	2.84	148						
D42'@HB3	**N74@HD22**	N74@ND2	11	2.84	144						
N74'@HD22	**Y39@HD1**	Y39@CD1	11	2.93	158						
N74@HD22	D42'@HB3	D42'@CB	10	2.82	141						
Y39'@HH	K71@HB2	K71@CB	10	2.94	141						

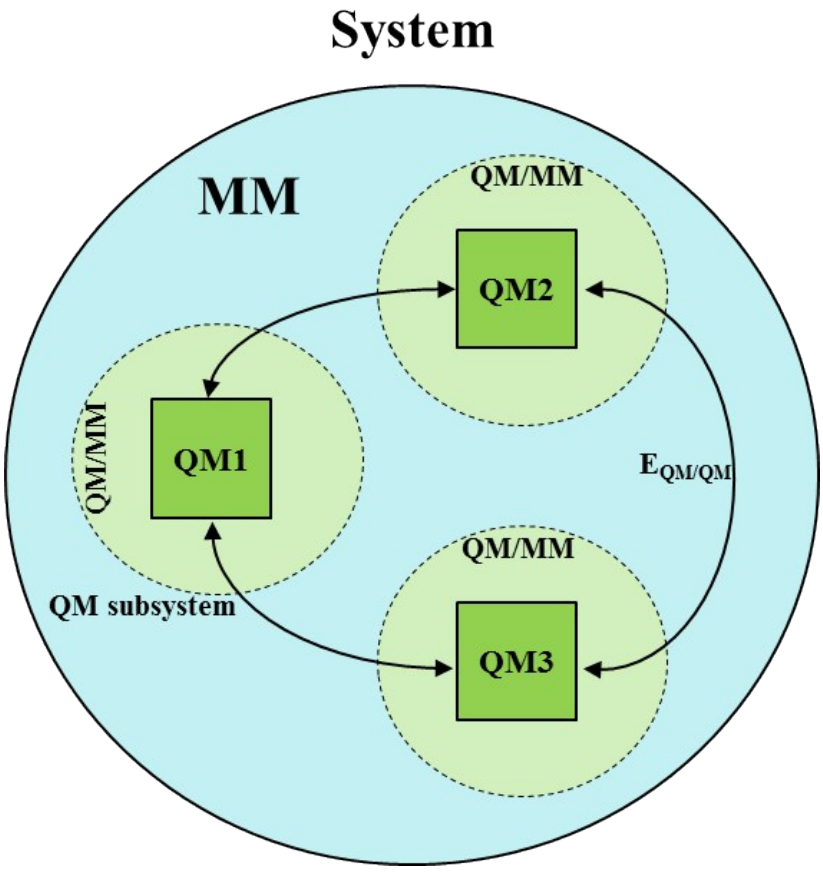

Figure S1. Partition scheme for the modelling of a large system. Each AZ is modeled by a disjoint QM subregion (QM_i) within the QM/MM approach forming a QM subsystem that will be coupled with the other QM subsystems through the $E_{QM/QM}$ coupling term using non-bonding interactions.

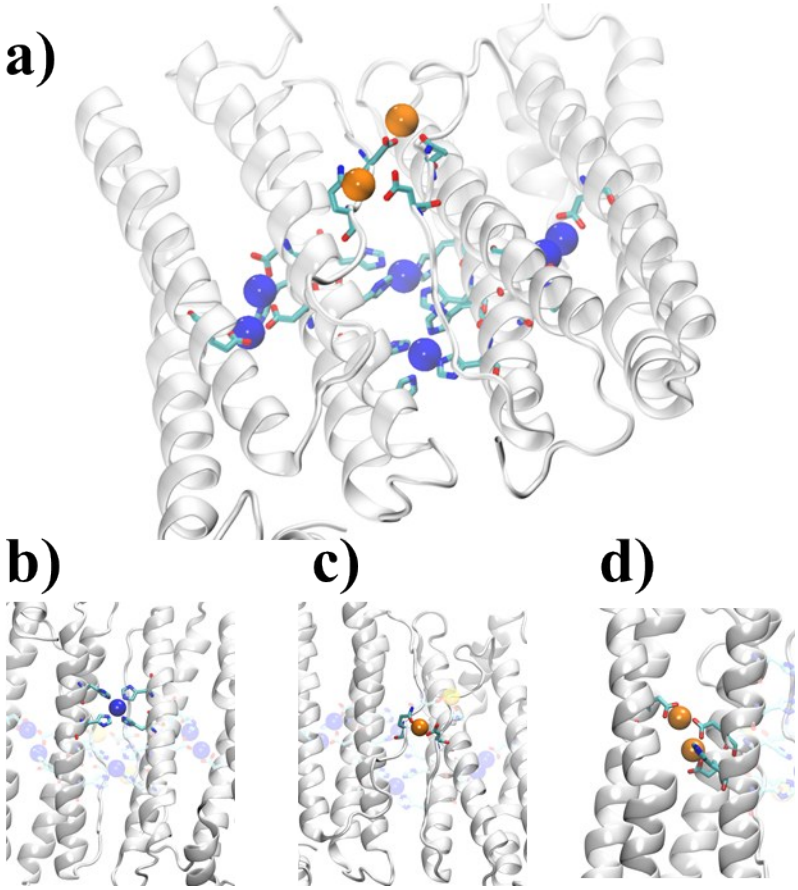

Figure S2. Detail of the quantum regions in the C_2 dimer interface; (a) General view of C_2 pore interface, and details of (b) Cu^{2+}/Ni^{2+} (blue color) coordination site of interface through histidine residues (QM_{C2_4H}), (c) Ca^{2+} (orange color) coordination sites of C_2 interface (QM_{C2_Ca}), and (d) central monomer with two copper coordination (QM_{cnt}).