# Hybrid dynamics simulation engine for metalloproteins

Manuel Sparta[a], Feng Ding[b], David Shirvanyants[b], Nikolay V. Dokholyan[b], and Anastassia N. Alexandrova[a]

[a]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095
[b]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599

**Supporting Information**

## I. Details of the QM/DMD method

As described in the main text, QM/DMD is an iterative procedure in which DMD dynamical simulation, and QM relaxation alternate to convergence.

### a. Initial geometry

The procedure accepts as input a single PDB file that contains all the chemical entities in the system (the protein itself, metals, substrates, cofactors, structural water molecules). Although DMD normally adopts an implicit representation of the non-polar hydrogen for amino acids, it can be waived (have a full atomic description) for substrates and cofactors.

### b. Constrains

The Discrete Molecular Dynamics (DMD) part of the simulation runs within the following constrains:

1. The position of the metal atom is fixed.
2. Any atom involved in a bond with a metal center is frozen.
3. Specific bond lengths within the QM domain may be frozen.
4. It is possible to impose that selected interatomic distances remain in an interval defined during the QM calculation.

While the primary scope of these constrains is to impose a feedback from the quantum mechanical calculation of the QM-DMD region into the dynamical simulation of the protein, they also allow for the simulation of metals (or specific oxidation states) that are not explicitly parameterized.

Our implementation of the QM/DMD method is highly automated, but the insight and control provided by the user is fundamental for the success and accuracy of the simulation. If a change in the connectivity occurs during the QM optimization, or if there is a significant discrepancy between QM and DMD structures, these effects must be addressed by the user via the imposition of additional constrains that effect the DMD run. In the current implementation, constrains available to retain the chemistry determined at the QM level of theory include:

- freezing cartesian coordinates of some atoms
- freezing interatomic distance
- modify the nature of an amino acid (protonating His, Glu, Asp or deprotonating Cys)
- modify the effective charge on a residue.

For example, if the carboxylate ion of a side chain of Glu was found to extract a proton from a given substrate, the user could:

- impose that the carboxylate ion and the group donating the proton retain their relative position.
- eliminate the negative charge from the carboxylate ion and add it to the substrate

The proper use of additional constrains makes possible to deal with cases where the DMD description and the QM description differ significantly. Once again, the key for a correct set up is the user's chemical intuition, e.g. during the QM optimization the user should anchor (by freezing specific atoms) long and mobile part of the QM-DMD region that could be placed in

unphysical position due to the fact that the QM-DMD region is "separate" from the rest of the protein, or limit the freedom of certain torsions during the DMD phase to respect the binding motif around the metal (for instance imposing the coplanarity of the metal and the imidazole group of an bound histidine).

### c. Discrete Molecular Dynamics

DMD underwent several modifications. Protonation states and proton donor/acceptor behavior often imposed on amino acids by proximal metal centers, or expected during chemical reactions were introduced. Specifically, Cys, Ser, and Tyr became capable of being deprotonated and serving as H-bond acceptors, Asp and Glu became capable of serving as H-acceptors, and His became capable of becoming doubly protonated. In addition, the square wells describing the bonded interactions in the QM-DMD region were made narrower and centered at the quantum mechanical minima, rather than at the averaged positions estimated from all the protein structures available in the Protein Data Bank, as in traditional DMD.

### d. Simulations details

Simulations are iterative. First, DMD is run on all but the QM-only region (dark grey in Fig. 1A), for 13,500 DMD time units (t.u.) each corresponding to 50 fs, together with an annealing profile. During each iteration of DMD, a temperature profile depicted in Figure S1 is adopted, to maximize the amount of information that can be extracted from a relatively short DMD run. Initially, we perform a short (1000 DMD time unit (t.u.); each time unit corresponds to ~50 fs) simulation at temperature T=0.10 Kcal/(mol•$k_B$) with a fast heat exchange to coupled virtual bath (10 t.u.$^{-1}$) using Andersen thermostat,[1] which was found to be an efficient method to remove eventual clashes in the pdb files.[2] Subsequently, the temperature is ramped to T=0.20 kcal/(mol•$k_B$) and decreased stepwise (consisting of 5 steps, 500 t.u. each) to equilibrate the structure. Finally the data collection occurs over at least 10000 t.u. calculated at T=0.10 kcal/(mol•$k_B$) with the same exchange rate.

This temperature program was found to be efficient in providing well-equilibrated and extensively represented ensembles. We note that, in absolute terms, T=0.10 kcal/(mol•$k_B$) is a low temperature for standard DMD (~50 Kelvin). It is adopted here, because we are interested in sampling a relatively small basin around the equilibrium geometry of the protein.
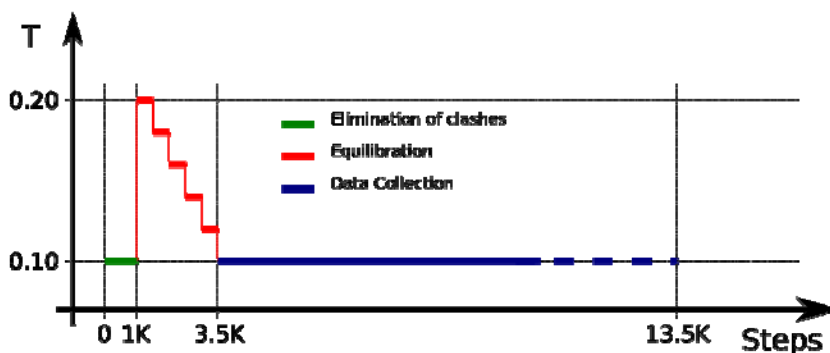


**Fig S1**. Temperature profile during each of the DMD simulation.

### e. Clustering

During the data collection, system snapshot is saved every 10 DMD t.u. for further analysis, hence during each iterations, at least 1000 snapshots are saved. The Kabsch[3] RMSD is computed for all pairwise snapshot structures to quantify their geometric similarity. Based on the

RMSD matrix, a hierarchical clustering algorithm(4) is used to organize all the snapshots in distinct clusters. The user can specify either the number of clusters, or the maximum relative distance (hence the amplitude) within each cluster. For each cluster, a specific representative structure is selected, and one can either choose the structure in the subset closest to the centroid, or the structure with the lowest DMD energy. In the study of Rubredoxin and its mutant we requested the number of clusters to be 5, and both the structure closest to the centroid, and the one with the lowest energy were used as representatives.

### f. Extraction of the QM-DMD region

For each of the representatives, the QM-DMD domain is extracted, and its QM energy is calculated (see the main text for the QM/DMD partitioning scheme). This step involves the following automated passages:

1. Metals, substrates, cofactors, and any amino acids label to be in the quantum mechanical domain are extracted from the specific snapshot. The QM-DMD region is user-specified.

2. While DMD takes advantage of an implicit description of non-polar hydrogens on the amino acids, all atom must be present for the QM calculations. We used a functionality of Chimera to restore all the hydrogen atoms not explicitly represented during the DMD simulation.

3. The QM-DMD region is then protonated, to satisfy all dangling valencies. The most common approach is to truncate included amino acid side chains at C$\alpha$, but our scheme is fully flexible and any portion of the protein can be selected to be included in the QM-DMD region.

4. The structure is then converted into a suitable format to be installed in a single point calculation template.
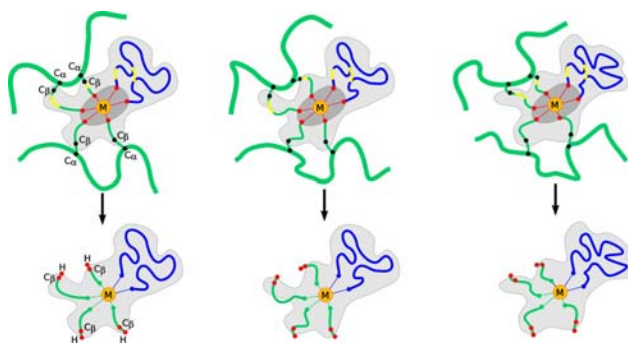


**Fig S2.** Graphical representation of the extraction of the QM-DMD region from the full system. All ligands and molecules of interest are retained, the amino acids are chopped at the C$\alpha$ and C$\beta$ bond and the valence is saturated with hydrogen. While in the dynamical simulation the metal center was frozen, during the quantum mechanical calculation the C$\beta$ and the link H are kept frozen in the position of the parental structure.

### g. QM calculation

Each of the structures prepared in the previous step undergoes a quantum mechanical single point energy calculation. In principle, any QM model can be used to compute the energy, but it should be kept in mind that, depending on the number of clusters computed in step D, this can become the bottleneck of the entire procedure. Furthermore, it is possible to perform (partial) optimization of each structure before computing its single point energy. Although this would be

beneficial in the next step (i.e. computing of the scoring function), preliminary results suggested that the single point energy is fairly correlated with the energy of the optimized structures. Hence, the more cost-effective, single point calculation approach is used. In the present implementation of the method the QM calculation is performed in the gas phase approximation, but more sophisticated embedding schemes can be used.

**h. Scoring Index (SI)**

It is now necessary to reduce the set of systems collected in the previous steps into a single candidate whose active center is going to be optimized at QM level. In a rather pragmatic way, we included both the QM energy of the QM-DMD region and the DMD energy of the whole system in the scoring index (SI), which, for the i-th system, reads:

$SI^i = n(E^i_{DMD} - E^{min}_{DMD}) + (1-n)(E^i_{QM} - E^{min}_{QM})$,

where $E^i_{DMD}$ and $E^i_{QM}$ are the DMD energy of the whole system and the QM energy of the QM-DMD region for the i-th structure, respectively, $E^{min}_{QM}$ is the minimum QM energy in the set of different QM-DMD regions, and $E^{min}_{DMD}$ is the minimum DMD energy in the set of whole systems. $n$ is number between 0 and 1, in our simulations set to 0.5. The SI can be seen as composed by two terms, the first deals with the energetic penalty computed for the QM-DMD region (QM energy), the second is related to the DMD energy of the whole system. The structure that minimized the sum of these two terms (weighted by the pair n, 1-n) is selected for further minimization. The analysis of the individual terms entering the SI may be useful when troubleshooting failing, unstable, or not-converging simulations by tracing the QM and DMD components. The follow-up structural and electronic analysis can reveal the nature of the problem, and lead to possible solutions, such as changing the size of the QM-DMD region, adopting a different QM method, removing or introducing additional constraints, or changing $n$.

During the implementation and preliminary testing of the method, we explored other form of the scoring index, including scaling the QM and DMD components by the number of atoms in each subsystem, and exploring the effect of $n$ of the performance of the method. For rubredoxin, the current form was found to provide an optimal balance between the information gathered during the DMD and QM part of the simulation. Different value of $n$ may be needed for other proteins, and preliminary testing should be done in each particular case. The structure with the lowest *SI* is selected for further QM optimization of the QM-DMD region.

The best structure (having the smallest *SI*) is chosen, and its QM-DMD region is partially optimized quantum mechanically.

**i. QM Optimization of the QM-DMD region**

The selected structure is optimized with the desired quantum mechanical method. At this stage, the position of the Cβ (or any other atom that bridge the QM-DMD region with the rest of the protein) and saturating Hydrogen atoms are frozen in order to retain the features imposed by the protein. At the same time, this insures that the optimized QM-DMD region can be plugged back into the protein without introducing excessive strains. Figure S3 shows a schematic representation of the process. As for the single point energy calculations, we perform the optimization within the gas phase approximation but, in principle, any Quantum Mechanical method and any solvent embedding scheme can be used.

In addition, it was realized that this step is generally the most time consuming for each iteration, therefore, to speed up the process, we decided to perform only a loose optimization of the QM-DMD region (by limiting the number of iteration cycles to 100) before reinstalling it into the protein, while the complete optimization was continued outside of QM/DMD procedure, for later analysis. We tested a convergence criterion based on the norm of the gradient, but such

procedure was found to use a consistently larger amount of optimization cycles and was not found to provide significant improvements over the preset number of optimization cycles.

The optimized QM region is then uncapped and reinstalled in the protein, and the QM-DMD boundary shrinks back for the dark grey region. In the course of the following iteration, in DMD stage, the structural information from the optimized QM-DMD region can propagate to the rest of the protein.
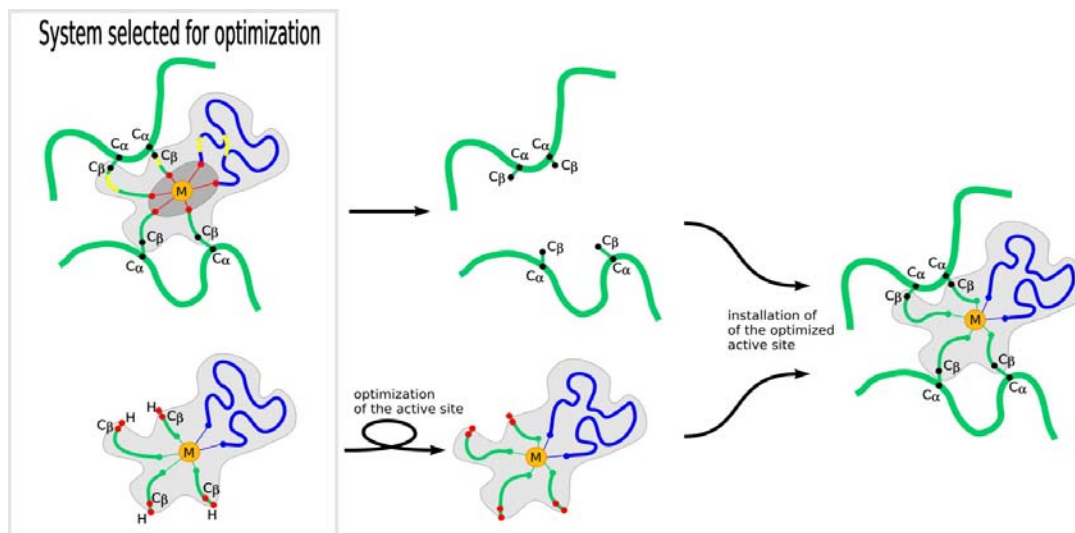


**Fig. S3.** The system with the lowest SI is selected for further optimization of its QM-DMD region. During the optimization, the atoms that bridge the QM-DMD region with the rest of the protein, as well as the saturating H atoms are kept fixed (red dots). After optimization, the site is reinstalled into the protein.

DMD sampling and clustering again provides QM with a finite set of structures. The alternating QM and DMD jobs with the breathing QM-DMD boundary are performed for 20 to hundreds of iterations, depending on the task. More specific details of QM/DMD, in general, as well as in application to our test case, can be found in the SI.

### II. Computational Details
The initial pdb structure for Rubredoxin (1FHM)(5) as obtained at the RCSB protein data bank. The metal and the four coordinated sulfur atoms constitute the QM-only region, hence they were frozen during the DMD simulation, an additional set of constraints was imposed during the DMD phase to maintain the structural features determined by the QM optimizations: the interatomic distances between the metal and CB of the coordinated cysteines were allowed to vary in the interval defined by the QM optimized values +/- 0.01 Å; the same type of constraint, was imposed for each of the bonds along the backbone of the QM-DMD region.

Each DMD simulation followed the temperature program as previously described, for the data collection, a 10000 t.u. run at a temperature equivalent to 0.10, was performed, and a frame out of ten was saved for further analysis. The 1000 frames were divided into 5 clusters, for each cluster, both the minimal DMD energy structure and the structure closer to the centroid were used in the QM screening, (10 Single Points in total).

The QM-DMD region considered during the QM calculations consisted of the following

amino acids: CYS6, THR7, VAL8, CYS9, GLY10, TYR11, VAL38, CYS39, PRO40, LEU41, CYS42, GLY43, VAL44. The peptide chain was initiated using the C and O atoms of THR5 and VAL38 while the Cα-C bonds were truncated. On the other end, the peptide chains were terminated saturating the Cα of GLY45 and ILE12 with a methyl group. Due to its directionality, the side chain of TYR11 was truncated at the Cβ-Cγ bond.

Whenever a C-C bond was truncated, the C included in the QM region was saturated with a Hydrogen atom positioned along the original C-C bond at a distance equal to $0.7052*d(C-C)$. All atoms added to saturate chopped amino acids, as well as the oxygen atoms belonging to THR5 and VAL38 were frozen during the QM optimization of the QM-DMD region, to retain the geometry imposed by the rest of the protein.

During the QM/DMD iterations, for the QM calculation we used Density Functional Theory, in the specific the BP86 (6-10) formulation of the Exchange and Correlation functional, and the double $\zeta$ quality basis set (def2-SV(P)).(11) Resolution of Identity (RI) and Multipole Accelerated Resolution of Identity (MARI-J) as implemented in Turbomole were exploited to speed up the calculations. Empirical dispersion correction for DFT calculations were included both in the energy and gradient evaluations.

During the QM/DMD simulation, we focused on the reduced form of the protein, i.e. iron was considered to be in its formal 2+ oxidation state; accordingly, the QM-DMD region had a net charge of -2, and the high spin state (quintet). The structure of each QM-DMD region was then reoptimized in the oxidized form, and in this case the net charge of the QM-DMD region was -1 (the quartet state).

Subsequently, for each of the optimized structures, a single point calculations were carried out using TPPS (6,8,12,13) as functional, and the def2-TZVPP basis set.(12) To account for the effect of the solvent, the Conductor-like Screening Model (COSMO) continuum solvation model, was used. The cavity in the solvent was constructed considering not only the atoms in the QM-DMD region, but also the remaining portion of the protein. The permittivity, $\varepsilon$, that represented the solvent was set equal to 78.4.

### III. Additional results
### WT Rubredoxin

Six independent replica of the QM/DMD simulation were carried out for the WT rubredoxin, and a total of 101 structures were collected and analyzed. As shown in the main article, the QM/DMD simulation leads the protein slightly away from the original initial structure, in Fig. S4, we show that the RMSD computed on the backbone atoms stabilizes around 1.4 Å, furthermore, the blue line depict a simulation where a slightly distorted structure was used as input (see main article for details).
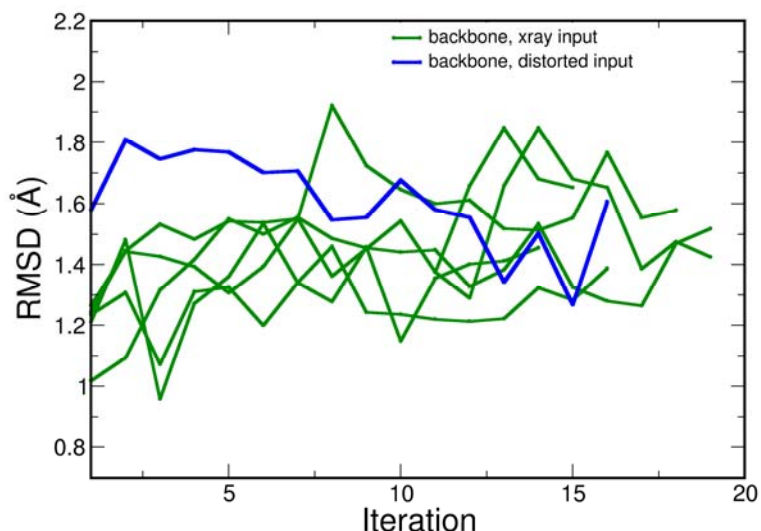
**Fig. S4.** RMSD computed on the backbone atoms of the protein as function of the number of iteration in the QM/DMD simulation. The blue line depicts the recovery of a slightly distorted QM-DMD region (see main article for details).

In Fig. S5, the total QM energy for the QM-DMD region of the protein in the reduced and oxidized forms of Rubredoxin are reported. The two energy sets appear to be fairly correlated, a low energy value in the reduced form correspond to a low energy value in the oxidized form.

In order to calculate the reduction potential, we used the following strategy: As mentioned, during the QM/DMD simulations, we followed the evolution of the reduced form of the protein only. After that, the structures of the oxidized counterpart were obtained by reoptiminzing the QM-DMD region with Fe(III), subject to the same constrains imposed by the rest of the protein. The "adiabatic" electron detachment energies were calculated as the energy differences between the Fe(II) snapshot coming from QM/DMD, and the Fe(III) forms resulting from the optimization of that specific snapshot structure. We justify this strategy by (1) the fact that the size of the QM-DMD region used in this study was fairly large, and (2) that the DMD engine is formally unaware of the oxidation state of the metal, but only respond to the moiety "metal and four coordinated sulfur atoms". We therefore expected that the a posteriori optimization of the oxidized form would yield comparable structures to those obtainable with QM/DMD simulations on the oxidized form. To test this hypothesis, we performed QM/DMD simulations on the oxidized protein. We found that indeed the structural parameters and energetic for the protein obtained in QM/DMD, and a posteriory optimization agreed. This confirmed the validity of our approximation.

Furthermore, since the structure and dynamics of the protein changes very subtle upon the change of the oxidation state (as expected from Marcus theory for electron transporter proteins), we assumed that $\Delta G$ can be approximated by $\Delta E$, and the entropic contribution can be ignored. This is described in detail in the main text.

These data allow us to compute the energy difference associate with the reaction $Ox + e^- \rightarrow$ Red. Once averaged over the 101 structures, our calculated value is -17.9 kcal/mol with and the

standard deviation of 1.3 kcal/mol. Clearly, this is not a good estimate of the $\Delta G$ of the redox reaction, preliminary calculations show that the lack of solvent effect is the major source of error. However, the structure of the protein predicted by QM/DMD agrees well with the experimentally available one, down to the subtlest details.

Implicit solvation treatment was added *a posteriori* on the set of structures collected in the QM/DMD simulations. In the presence of the implicit solvent, the $\Delta E$ associated with the RedOx reaction equals -82.5 with a standard deviation of 1.4 kcal/mol, confirming the importance of the inclusion of solvent corrections.

As mentioned in the main article, the fine tuning of the RedOx potential of Rubredoxin and its mutants was initially correlated with an increase in the strength of the Hydrogen bond between the amide hydrogen of Residue 44 and the sulfur of Cys42. Fig. S6 shows the H-bond distance between the amide hydrogen of Val44 and the sulfur of Cys42. As it was previously observed, the H-bond distance is shorter in the reduced form of the protein, and calculated average value in this case is 2.63 Å with a standard deviation of 0.09 Å. For the oxidized state, it is possible to notice a few outliers lying well above 3.40 Å, it was found that in these structures, the amide hydrogen of Val44 is engaged in a bond with the oxygen on the backbone of Cys42 rather than with the sulfur. By considering only the structures with the N-H—S bond, we computed an average H---S distance of 2.97 Å with a standard deviation of 0.10 Å, a value in good agreement with previous results.
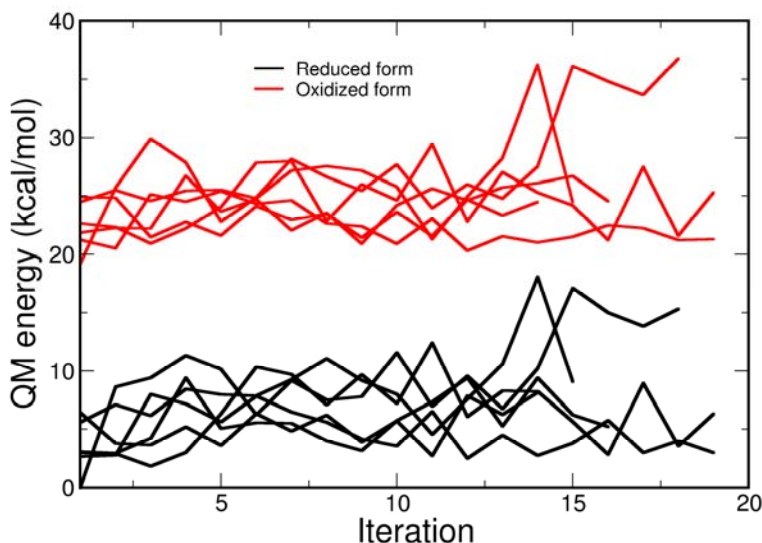


**Fig. S5.** Total energy of the QM-DMD region for the reduced and oxidized forms of Rubredoxin relative to the most stable conformer in the reduced form.
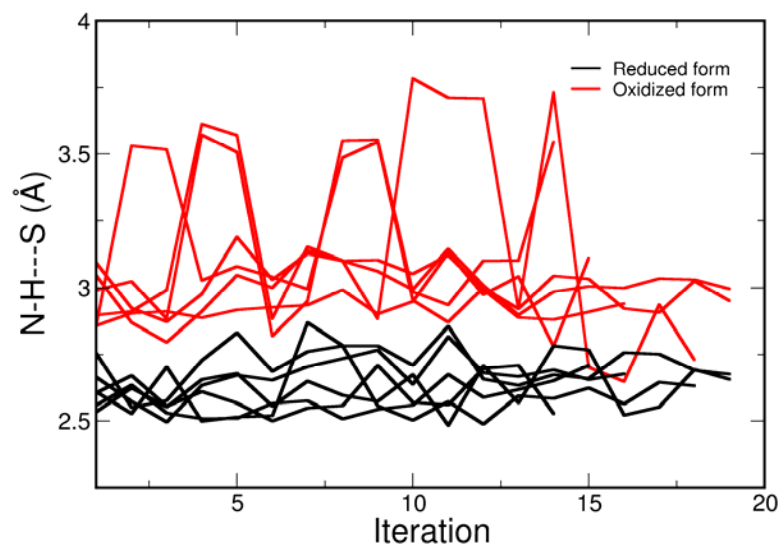
**Fig. S6.** H-bond distance between the amide hydrogen of Val44 and the sulfur of Cys42 for the reduced and oidized form of wt Rubredoxin.

**Val44Ile Rubredoxin**

Contrary to Val, Ala and Gly, the side chain of Ile can adopt two distinct conformations. In the first one, depicted in Fig. S7(A), the longer branch of the side chain points away from the core of the protein, in the second one (Fig. S7(B)) the longer branch accommodates into a pocket towards the core of the protein.
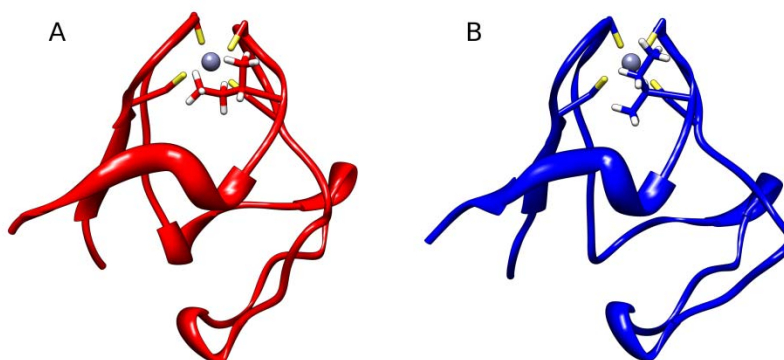


**Fig. S7.** Two possible conformations of the side chain of Ile44 in the V44I mutant of Rubredoxin: C$\delta$ points either towards the core of the protein (A) or away from the core (B).

For each of the conformation of the side chain of Ile44 we generated a suitable input and ran a single QM/DMD simulation. For this mutant, the structure used in input was updated only if the iteration provides an QM-DMD region with lower QM energy.

The total QM energy of the QM-DMD region for the two conformational families in their

9

reduced and oxidized forms are presented in Fig S8. The four sets of data are well clustered as the standard deviation for all set is ca. 2 kcal/mol. For both the reduced and oxidized forms, the conformer B ($Ile_{C\delta}$ points away from the core of the protein) is predicted to be slightly more stable accotding to QM (4±3 kcal/mol), whereas the DMD energies for the two variants are degenerate.
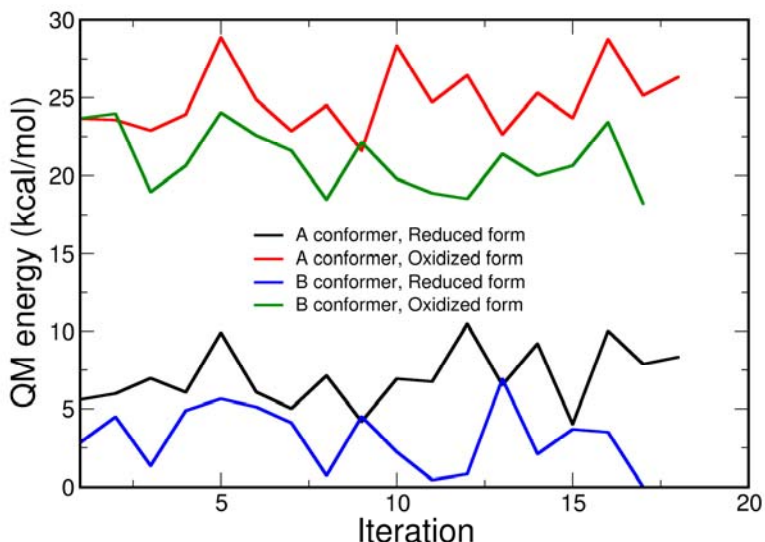


**Fig. S8.** Total energy of the QM-DMD region for the reduced and oxidized forms of V44I Rubredoxin relative to the most stable conformer in the reduced form. Two conformational families shown in Fig. S7 are considered.

If one considers the $\Delta E$ associated with the RedOx reaction, both conformational families present a value of 17.4 with a standard deviation of 1.4 kcal/mol. After correcting for the solvent effect, the two values split by 0.2 kcal/mol. Given the small energy difference, we will consider only the average over all structures of both conformers (-81.1±1.7 kcal/mol), in the analysis of the effect of mutation.

On the other hand, if one considers the length of the hydrogen bonds characterizing the QM-DMD region, (see main article, Table 1), the most significant difference is found for the $S_{Cys42}$-$H_{Val44}$ hydrogen bond. For conformer A, the bond length is equal to 2.75±0.13 and 3.03±0.12 Å, in the reduced and oxidized form respectively, whereas the corresponding values for conformer B are 2.49±0.06 and 2.77±0.07 Å.

**Val44Ala Rubredoxin**

Two independent runs of QM/DMD were used to characterize the Val44Ala mutant of Rubredoxin. The total QM energies of the QM-DMD region for the reduced and oxidized forms are presented in Fig S9. The $\Delta E$ associated with the RedOx reaction for the V44I mutant is found to be -18.0 with a standard deviation of 0.8 kcal/mol. The inclusion of solvent corrections led to $\Delta E$ = -82.1±1.1 kcal/mol.

In the case of the Val44Ala mutant, the H-bond distance between the amide hydrogen of Ala44 and the sulfur of Cys42 is found to be 2.33±0.02 Å in the reduced form, and 2.55±0.05 Å
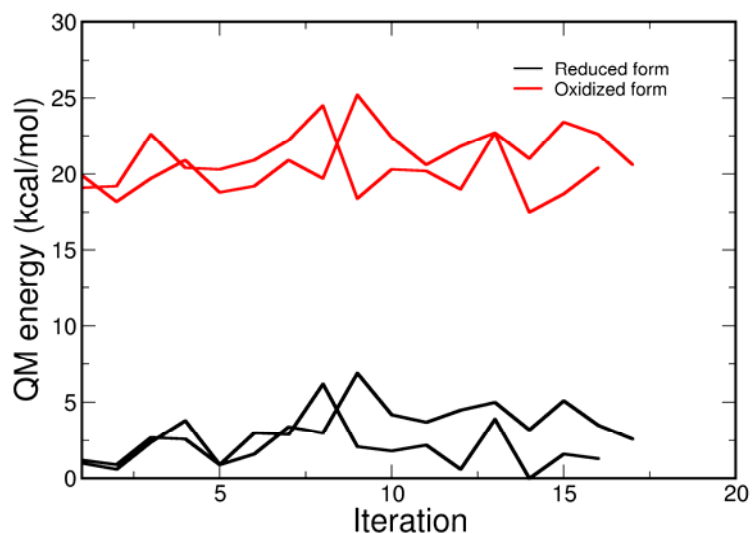
in the oxidized form.



**Fig. S9.** Total QM energy of the QM-DMD region for the reduced and oxidized form of V44A Rubredoxin relative to the most stable conformer in the reduced form.

The structures collected during the QM/DMD simulation of the V44A variant of Rubredoxin were found in excellent agreement with the available X-ray structure (PDB ID 1C09), as shown in Fig S10.
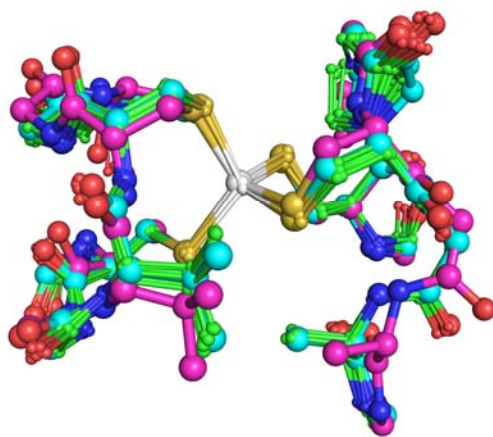


**Fig. S10.** Superposition of the experimental QM-DMD region (PDB ID 1C09, cyan, large spheres, thick sticks) and 10 representatives of the simulated structures for the V44A-rubredoxin (green, small spheres, thin sticks). The structure used in input (obtained by simple replacement of the side chain of Res44) is included for comparison (magenta, large spheres, thick sticks). Hydrogen atoms are removed for clarity.

**Val44Gly Rubredoxin**

Eight independent QM/DMD simulations were done for the Val44Gly mutant of Rubredoxin, in total ca. 120 structures were generated. The QM energies for in the reduced and oxidized forms are shown in Fig. S6. The average energy difference between the two states was found to be -17.5 kcal/mol with a standard deviation of 1.4 kcal/mol. The inclusion of *a posteriori* solvent correction led to a $\Delta E$ =-81.1 ±1.7 kcal/mol.
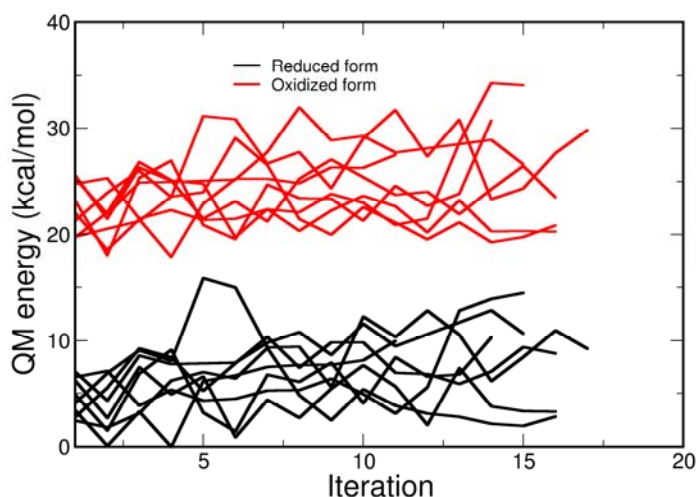


**Fig. S11.** Total energy of the QM-DMD region for the reduced and oxidized form of the Val44Gly mutant of Rubredoxin, relative to the most stable conformer in the reduced form.
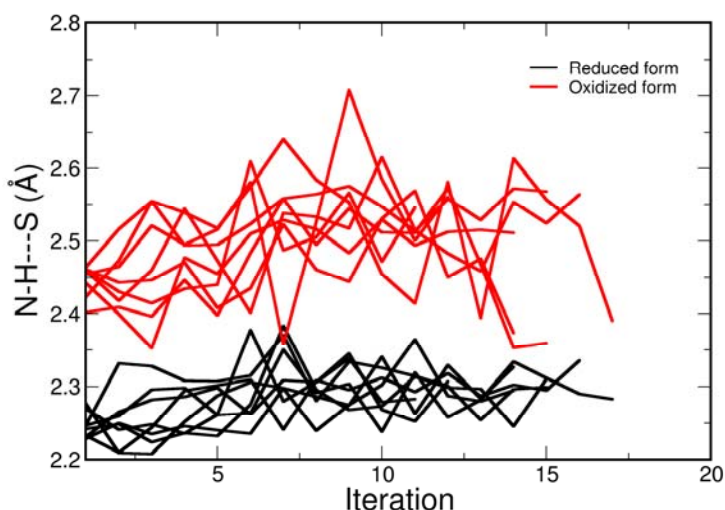


**Fig. S12.** H-bond distance between the amide hydrogen of Gly44 and the sulfur of Cys42 for the reduced and oxidized form of V44G Rubredoxin.

**Table S1.** Calculated and experimentally-derived H-bond lengths between the $S_{Cys}$ atoms bound to Fe and the backbone N-H groups, in the two forms of the WT rubredoxin and its mutants, in Ångstroms.

| | $Fe^{2+}$ | | | | $Fe^{3+}$ | | | |
|---|---|---|---|---|---|---|---|---|
| | **WT** | **V44I** | **V44A** | **V44G** | **WT** | **V44I** | **V44A** | **V44G** |
| $S_{Cys42}$-$H_{Res44}$ | 2.63±0.09 | 2.75±0.13[a] 2.49±0.06[a] | 2.33±0.02 | 2.29±0.04 | 3.07±0.24 2.97±0.20[b] 2.87[c] (3.01)[d] | 3.03±0.12[a] 2.77±0.07[a] 2.75[c] | 2.55±0.05 2.55[c] (2.60)[e] | 2.50±0.07 2.36[c] |
| $S_{Cys9}$-$H_{Tyr11}$ | 2.61±0.16 | 2.56±0.04[a] 2.45±0.06[a] | 2.50±0.05 | 2.62±0.12 | 2.90±0.20 | 2.89±0.10[a] 2.77±0.06[a] | 2.85±0.07 | 3.01±0.22 |
| $S_{Cys6}$-$H_{Cys9}$ | 2.58±0.08 | 2.56±0.13[a] 2.52±0.07[a] | 2.56±0.06 | 2.69±0.06 | 2.49±0.05 | 2.48±0.04[a] 2.54±0.06[a] | 2.52±0.06 | 2.62±0.09 |
| $S_{Cys6}$-$H_{Val8}$ | 2.35±0.04 | 2.32±0.03[a] 2.36±0.03[a] | 2.35±0.04 | 2.35±0.03 | 2.55±0.06 | 2.52±0.04[a] 2.55±0.02[a] | 2.56±0.04 | 2.59±0.05 |
| $S_{Cys39}$-$H_{Cys42}$ | 2.52±0.08 | 2.51±0.05[a] 2.50±0.05[a] | 2.54±0.05 | 2.61±0.08 | 2.61±0.12 | 2.65±0.08[a] 2.64±0.06[a] | 2.70±0.08 | 2.78±0.14 |
| $S_{Cys39}$-$H_{Leu41}$ | 2.35±0.05 | 2.35±0.04[a] 2.37±0.03[a] | 2.34±0.04 | 2.30±0.04 | 2.49±0.06 | 2.50±0.05[a] 2.54±0.04[a] | 2.53±0.05 | 2.54±0.08 |

[a] The values predicted for the two families of conformations are reported. [b] Structures where $S_{Cys42}$-$H_{Val44}$ are not engaged in a Hydrogen-Bond are discarded from the calculation of the average. [c] NMR data from Ref. 25. [d] X-ray derived data from Ref. 28. [e] X-ray data from Ref. 28.
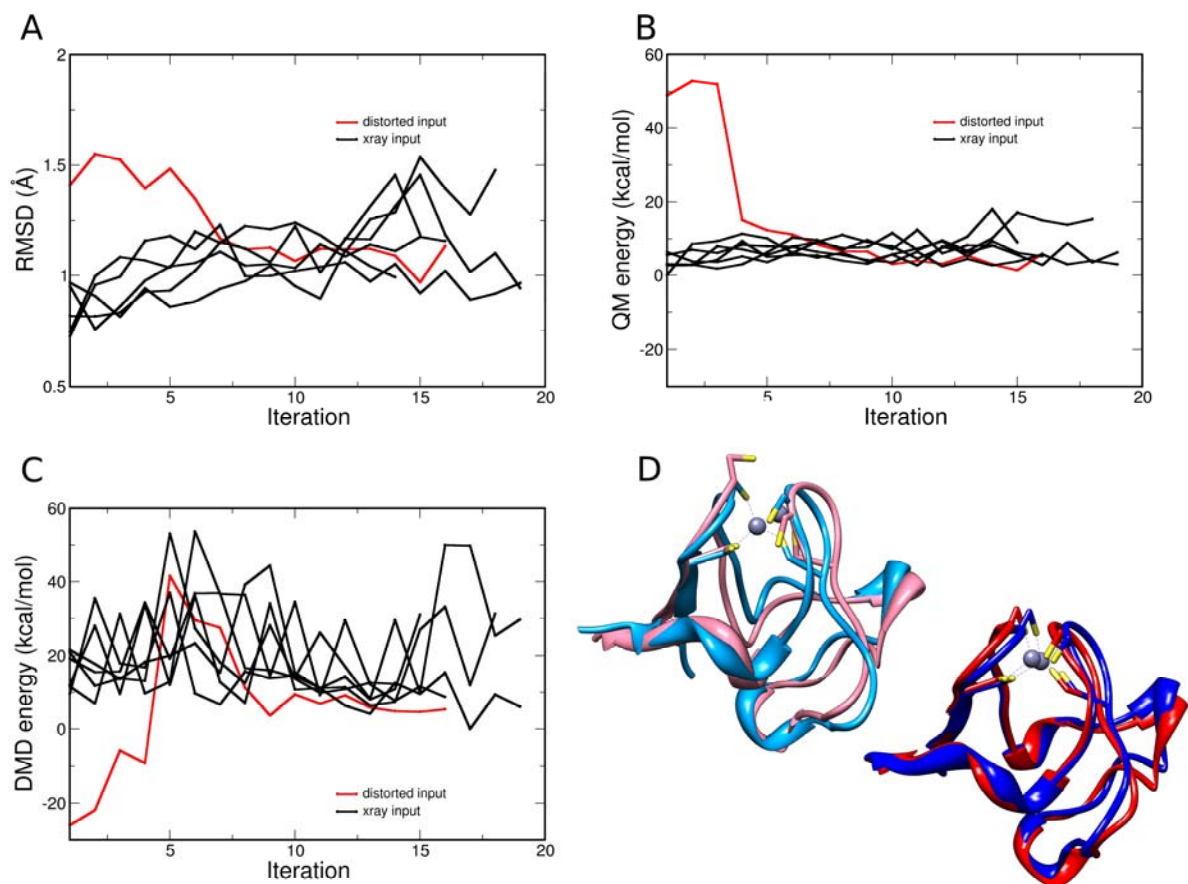
**Fig. S13.** The large level of QM/DMD testing: structure recreation and recovery for rubredoxin in the $Fe^{2+}$ form. The stability of the backbone and QM-DMD region is illustrated by (A) the conserved RMSD values from the X-ray structure (black lines), (B) the stable QM energies of the QM-DMD region (black lines), and (C) the stable DMD energies of the entire protein minus the QM-only region (black lines). In addition, the WT structure distorted from the equilibrium is returned to it in ca. 8 iterations of QM/DMD (red lines in A-C). (D) The overlay of the original X-ray structure of rubredoxin (light blue) and the manually distorted structure (pink), and the overlay of representative QM/DMD equilibrated structures, starting from X-ray (dark blue), and starting from the distorted structure (red).

**Reduction potential as a function of mutations in the position 44 in rubredoxin**

**Table S2.** Experimental and calculated properties: RedOx potentials (mV), derived $\Delta G$ and $\Delta\Delta G$ values (kcal/mol), lengths of H-bonds between the backbone N-H group of residue 44 and $S_{Cys42}$ (Å), calculated $\Delta E$ and $\Delta\Delta E$ (kcal/mol) for the RedOx reaction in gas phase and with the implicit solvent correction.

| | Experiment | | | | | Gas phase | | | Solvent | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Protein** | **RedOx** | **NH—S Ox.** | **NH—S Red.** | **$\Delta G^{b}$** | **$\Delta\Delta G$** | **NH—S Ox.** | **$\Delta E$** | **$\Delta\Delta E$** | **$\Delta E$** | **$\Delta\Delta E$** |
| **WT** | -77 | 2.87 | 2.63±0.09 | -100.6 | 0.0 | 2.97±0.20 | -17.9±1.3 | 0.0±1.8 | -82.5±1.4 | 0.0±1.9 |
| **V44I** | -54 | 2.75 | 2.75±0.13[a] | -101.2 | -0.6 | 3.03±0.12[a] | -17.8±1.4[a] | 0.1±2.0[a] | -82.2±1.9[a] | 0.3±2.4[a] |
| | | | 2.49±0.06[a] | | | 2.77±0.07[a] | -17.9±1.4[a] | 0.0±2.0[a] | -82.0±1.4[a] | 0.5±2.0[a] |
| **V44A** | -23 | 2.55 | 2.33±0.02 | -101.8 | -1.2 | 2.55±0.05 | -18.0±0.8 | -0.1±1.2 | -82.1±1.1 | 0.4±1.3 |
| **V44G** | 0 | 2.36 | 2.29±0.04 | -102.4 | -1.8 | 2.50±0.07 | -17.5±1.5 | 0.5±2.1 | -81.1±1.7 | 1.4±2.2 |

[a] The values predicted for the two families of conformations are reported. [b] The values are calculated with the absolute potential of the standard hydrogen electrode being -102.4 kcal/mol, as recommended by IUPAC.

The numbers above were obtained without any kind of electrostatic embedding, because embedding schemes do not guarantee more accurate results, as was shown by Ryde et al. (Ref 47, main text). In order to convince ourselves that this is indeed the case, we performed simulations with the embedding, organized as follows: The active site was embedded into a point charge distribution (in place of the rest of the protein) obtained by placing a charge in the position of each DMD-only atom. The atomic charges were assigned as in the OPLS-AA force field, except for the carbon atoms bonded to a non-polar hydrogen(s), for which OPLS-UA force field was used. This setup mimicked closely the DMD representation of the system. The total system (actual QM atoms and point charges) was then solvated using the implicit solvation model, COSMO, with ε=80, and the solvent cavity including the entire protein, not just the active site. The use of different ε for the actual solvent, and the protein in calculations of reduction potentials of rubredoxin was found important in earlier works (Ref. 46b, main text). The reduction potential was recalculated for this model. The $\Delta G$ values corresponding to the reduction potentials obtained in this way read as -76.3 ± 3.4, -75.1 ± 3.5, -76.3 ± 3.2, and -75.4 ± 3.0 kcal/mol, for WT, V44I, V44A, and V44G, respectively. Evidently, they are further apart from the experimental values (Table S2), as compared to the results obtained without embedding. We therefore did not use any embedding in the results reported in the main text. However, in principle, any advanced embedding scheme can be coupled to QM/DMD.

## $Pd^{2+}$ substituted rubredoxin
We showcase the features and potentials of QM/DMD with an additional simulation that demonstrates how the flexibility of a QM calculation combined with the efficient sampling of the backbone provide an unprecedented tool for the accurate modeling of metallo-protein.

We simulated a $Pd^{2+}$ containing rubredoxin, where the low spin singlet configuration was imposed on Pd. Low spin $Pd^{2+}$, being a $d^8$ metal, is known to prefer square planar coordination. Clearly the starting tetrahedral coordination environment provided for the metal by the structure of the protein was not optimal for $Pd^{2+}$. At the same time, the structure of the protein is obviously evolutionarily optimized so as to provide the tetrahedral, and not square planar coordination sphere for the metal. We expected QM/DMD to yield the structure of Pd-rubredoxin that would strike a compromise between these two opposing structural trends.

QM/DMD performed fully up to our expectations (Figure S14). Pd-rubredoxin adopted a clear compromise structure in which the protein (including the backbone) got distorted in response to the presence of Pd, so as to bring its coordination geometry closer to square planar. With respect to the native structure, the two protein loops that contain the Cys responsible for the metal coordination get closer to each other. At the same time, this increases the steric clashes between the side chains, e.g. Val8 and Val44. As a side note, this result, in fact, suggests that mutation of Val8 and Val44 to smaller residues could improve the affinity of this protein to metals that prefer a square planar coordination.

Compared to the Fe based proteins, the metal-sulfur bond lengths were found to be elongated by ca 0.08 Å, in agreement with the larger ionic radius of Pd. At the same time, the coordination sphere of Pd evolved away from tetrahedral and toward square planar, but the square planar coordination was not fully reached, being counterbalanced by the structure of the rest of the protein. In the Figure S14, the native structure (large sphere, tick sticks) is superposed to the simulated QM-DMD region bound to Pd (thin sticks, small spheres), to illustrate this change in the coordination sphere of the metal.
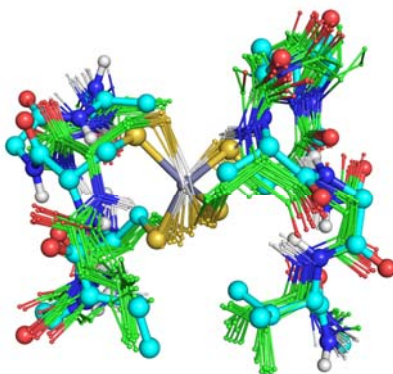


**Figure S14:** Superposition of the native QM-DMD region (large spheres, thick sticks) and the simulated structures for the Pd-rubredoxin (small spheres, thin sticks). Non polar hydrogen atoms are removed for clarity.

Of course, this is a purely academic exercise, and we do not claim that Pd-rubredoxin can be stable enough to be obtained. However, this test shows that QM/DMD provides a balanced description of metalloproteins, and can be used to predict structural changes upon metal replacement even in such challenging situations where the coordination of the metal has to change drastically.

## Ca²⁺ substituted COMT

To further illustrate the need from QM/DMD sampling in resolving the structure of the $Ca^{2+}$ substituted COMT, we extracted the active site of COMT directly from the crystal structure, replaced the cation, and optimized the complex without any adjustment of the position of the backbone. The resultant structure of the active site is essentially analogous to that of the $Mg^{2+}$ form, as shown in Figure S15A. The metal does not acquire the additional ligand, and no rearrangement of the H-bonding network is observed. The structural differences in the active site can be found only after the extensive sampling allowing for the repacking of the protein. The ensemble of structures resultant from the QM/DMD simulation is shown in Figure S15B for comparison.
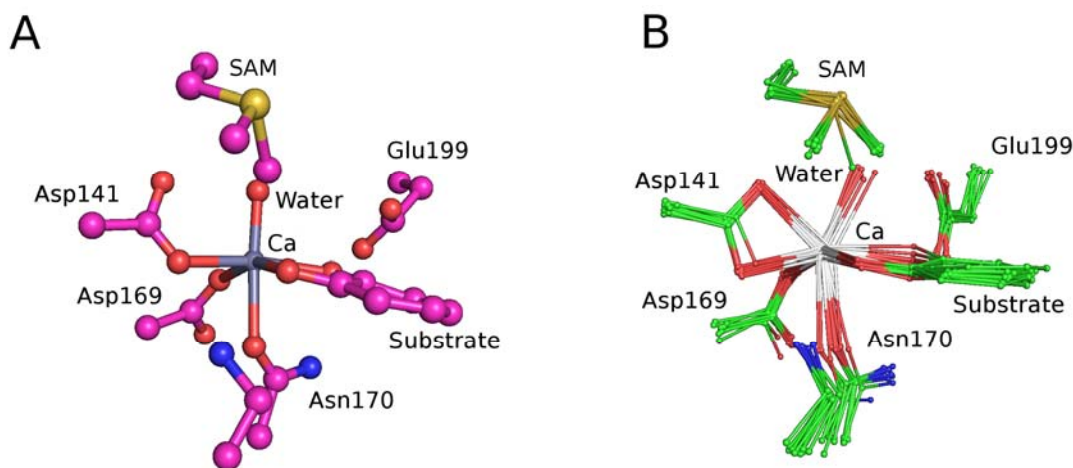


**Figure S15:** Comparison of the active site of the $Ca^{2+}$ form of COMT obtained via (A) replacement of the metal in the crystal structure and subsequent optimization at the fixed position of the protein backbone, and (B) QM/DMD sampling.

## References

1. Andersen, H. C. 1980. Molecular dynamics at constant pressure and/or temperature. J. Chem. Phys. 72:2384-2393.

2. Ramachandran, S., P. Kota, F. Ding. and N. V. Dokholyan. 2011. Automated Minimization of Steric Clashes in Protein Structures. Proteins 79:261-270.

3. Kabsch, W., 1976. A solution of the best rotation to relate two sets of vectors. Acta Crystallogr. A 32:922-923.

4. Barton, G.J. (1993, 2002) "OC - A cluster analysis program", University of Dundee, Scotland, UK; http://www.compbio.dundee.ac.uk/downloads/oc.

5. Min T., C. E. Ergenekan, M. K. Eidsness, T. Ichiye, and C. Kang. 2001. Leucine 41 is a gate for water entry in the reduction of Clostridium pasteurianum rubredoxin. Protein Sci. 10:613-621.

6. Dirac P. A. M. 1929. Quantum Mechanics of Many-Electron Systems. Proc. Royal Soc. (London) A 123:714-733.

7. Slater J. C. 1951. A Simplification of the Hartree-Fock Method. Phys. Rev. 81:385-390.

8. Vosko S. H., L. Wilk, and M. Nusair. 1980. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: a Critical Analysis. Can. J. Phys. 80:1200-1211.

9. Becke A. D. 1988. Density-functional exchange-energy approximation with correct asymptotic behavior. Phys. Rev. A 38:3098-3100.

10. Perdew J. P. 1986. Density-functional approximation for the correlation energy of the inhomogeneous electron gas. Phys. Rev. B 33:8822-8824.

11. Weigend F., and R. Ahlrichs. 2005. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design an assessment of accuracy. Phys. Chem. Chem. Phys. 7:3297-3305.

12. Perdew J. P., and Y. Wang. 1992. Accurate and simple analytic representation of the electron-gas correlation energy. Phys. Rev. B 45:13244-13249.

13. Tao J., J. P. Perdew, V. N. Staroverov, and G. E. Scuseria. 2003. Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solid. Phys. Rev. Lett. 91:146401-4.