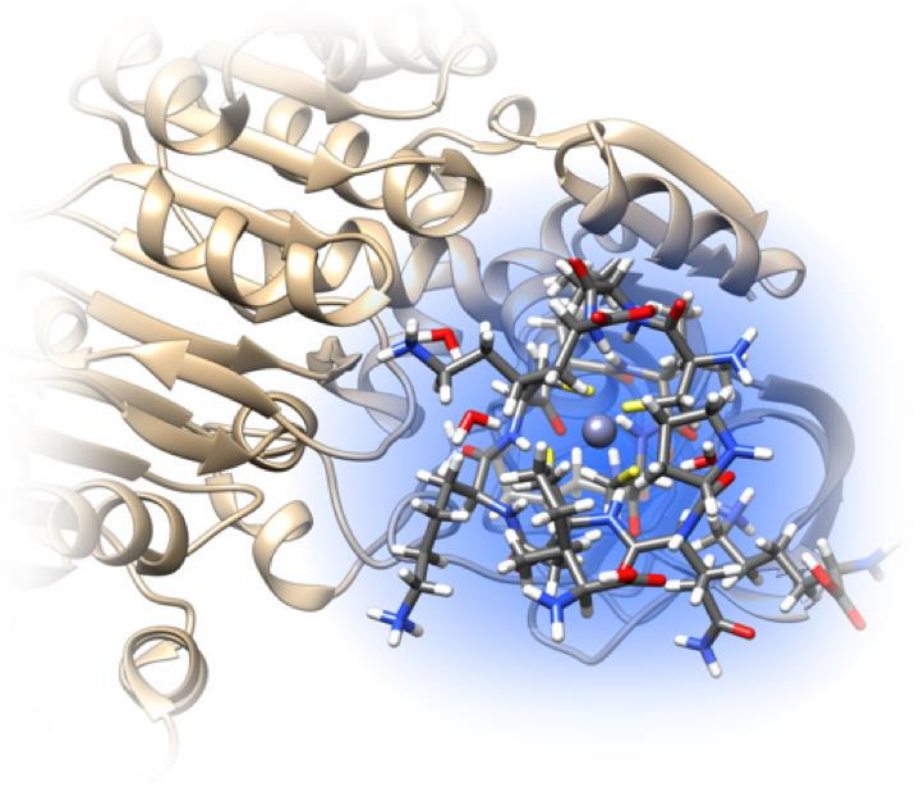
E-CAM

European HPC Centre of Excellence

E-CAM case study

The simulation of metal ions in protein-water systems using machine learning



With Dr. Francesco Fracchia, Scuola Normale Superiore di Pisa
Interviewer: Dr. Donal Mackernan, University College Dublin

www.e-cam2020.eu

Abstract

One quarter to one third of all proteins require metals to function but the description of metal ions in standard force fields is still quite primitive. In this case study an E-CAM pilot project[1] in collaboration with BiKi Technologies[2] to develop a suitable parameterisation using machine learning is described. The training scheme combines classical simulation with electronic structure calculations to produce a force field comprising standard classical force fields with additional terms for the metal ion-water and metal ion-protein interactions. The approach allows simulations to run as fast as standard molecular dynamics codes and is suitable for efficient massive parallelism scale-up.

Why is accurate modelling of metal ions interactions with soft matter important in general scientifically and from an industry perspective in particular?

Metal ions play a structural and catalytic role in many enzymes and an accurate description of the interaction with amino acids, nucleic acids, lipids, carbohydrates is essential for the study of biochemical processes. In the industrial field, the accurate description of the behaviour of metal ions in soft matter is critical for designing new drugs. A drug in several cases involves the interaction with a catalytic site of an enzyme, binding to it and inhibiting its function. When the catalytic site consists of a metal cofactor, the accurate description of the interaction between the metal ion and the ligand becomes decisive for making predictions about the effectiveness of potential new drugs.

How is E-CAM helping BiKi Technologies to solve this problem?

Since BiKi Technologies provides drug design services to pharmaceutical companies, including parameters and protocols for non-molecular dynamics experts, it is in their strategic interest to have tools to generate affordable force fields. Active sites of enzymes include frequently metal ion cofactors and an accurate description of these sites is fundamental to perform drug discovery analysis. Unfortunately, available force fields of metal ions proved not to be satisfactory for these applications. We tackled the issue by developing a novel procedure to parameterize metal ion force fields using ab initio data on model systems as reference, and the production of the software tools that exploit this procedure is currently in progress. The outcome of this task should be a novel and general approach to metalloprotein parameterization and a procedure for generating molecular mechanics parameters from quantum chemical potentials. BiKi has collaborated suggesting the protein systems exploited as test cases of the procedure, participating in key discussions and providing computational resources.

What are the chief difficulties that accurate simulation has to face?

The accurate description of the behaviour of a metal ion in a biological environment is difficult even using quantum mechanical methods. In fact, in particular for transition metals, only

[1] Pilot Project on Quantum Mechanical Parameterisation of Metal Ions in Proteins (https://www.e-cam2020.eu/pilot-project-biki-2/)
[2] BiKi Technologies (http://www.bikitech.com)

multi-reference methods that include relativistic effects would provide high quality results. However, these methods generally have too high computational cost to be applied to large-sized systems such as a protein. A larger field of applicability can be obtained through QM/MM methods, using DFT for the QM part. Even in these cases, at present the cost of computation is too high to conduct large-scale dynamic simulations in order to obtain structural, thermodynamic and kinetic properties of the systems of interest. Computational costs can be drastically reduced by conducting dynamics with classic models, so it is necessary to develop force fields. For metal ions two alternatives are possible: bonded and non-bonded models. Bonded models can be built to reproduce the desired structural proprieties of the environment of the metal ion but they are not useful to describe the chemistry of the site. In a drug design perspective this is a significant limitation.

Non-bonded models can lead to chemistry, however the parametrization of accurate force fields of this type is more difficult because the greater structural flexibility of the model can give rise to incorrect coordination of the metal ion. Our work is focused on this particular aspect of the problem: parameterization of non-bonded force fields for metal ions using statistical learning techniques.

Can you can elaborate a little bit more on your approach and explain to what extent it is built on what was there before and to what extent there are new things?

The parametrization of the metal ion force fields is generally conducted trying to reproduce the experimental properties of the ions in water. This is the case, for example, of the force fields produced by Merz and collaborators in numerous works[3], which use the radial distribution functions and the free hydration energies as references. This approach suffers from two difficulties: 1) limited availability of experimental data, often only available for water and other simple polar solvents; 2) little possibility of exploring the space of the parameters of the force-fields. In this case, you can only use a limited functional form for the force-field. If the force fields is not flexible enough, the performance will not be high.

The use of references calculated with ab initio methods to perform the fitting of the force fields allows to:

- explore a much larger parameter space, also using more flexible functional forms than simpler models such as Lennard-Jones
- to exercise greater control over the types of systems to be used as a reference, in particular to generate molecular structures representative of systems similar to those in which force fields will be used.

In our methodology we use ab initio energies and forces of model systems as reference quantities. The model systems are clusters in which the metal ion is coordinated by molecules that are different according to the situations that we want to study (solvents or amino acids). Clusters are generated by applying a combinatorial algorithm that maximizes the dissimilarity of configurations. The fitting is then performed by introducing in the algorithm: input data (geometries of the cluster), output data (ab initio quantities) and the functional form of the force field model. We tackle the supervised learning problem using linear and non-linear

[3] Li, Roberts, Chakravorty, Merz *J. Chem. Theory Comput.* **9** (2013) 2733

parameters simultaneously by combining linear ridge regression and cross-validation techniques with the differential evolution optimization algorithm.

The clusters that are used to generate different environments can be selected to suit the needs of the user. In the present case, to develop and test the procedure we have used metal ions in water as study cases.

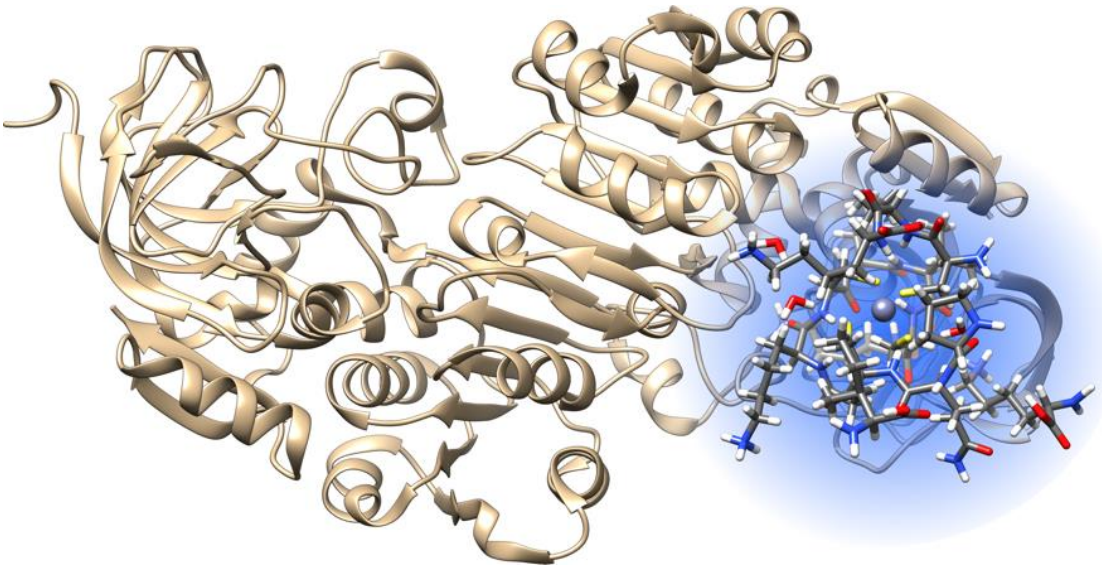


Figure 1: Ab-initio calculations on cluster systems are used as references for the parametrization of the metal ions force fields. The clusters are built cutting protein configurations selected by GRASP sampling.

This approach then should scale very well on a massively parallel machine, as the ab-initio code can be run on completely independent configurations?

Yes, it is true. Also, the combinatorial algorithm (that is the GRASP sampling code) that chooses the training set can be parallelized. The largest cost of the procedure is the calculation of ab-initio quantities and it can indeed be run independently exploiting fully the scaling properties of state-of-the-art electronic structure codes.

Are your codes being tested for real protein/drug environments? Are there plans for tests by a pharmaceutical company?

Tests on real proteins including zinc ion cofactors are in progress. We are focused on the alcohol dehydrogenase system because includes a zinc ion coordinated by four cysteine residues, that is a recurring motif in proteins. The uniform environment of the ion allows to extend gradually the work performed in water to a heterogeneous system. We have also collected data for the carbonic anhydrase II, thermolisin, L-rhamnose isomerase and the zinc finger motif. In this phase we are having difficulty in reproducing the stability of the conformations of the studied metal ions environments. When the generated force fields meet quality requirements in terms of agreement with the experimental data we will pass the complete code to BiKi.

What do you think are the future prospects and likely impact of this project?

In the case of metal ion force field parameters, results have so far only been published for the case of ions in water. The application of the procedure to more complex and more interesting contexts (such as proteins) is still in progress.

Further developments from a methodological point of view will be the production of force fields with more flexible functional forms.

Although the method has been developed to perform parametrizations of force fields of metal ions, it is much more general. It can be used to parameterize force fields of anions and solvents and intramolecular force fields. I should stress that the methodology is a statistical learning process that optimizes the parameters of pre-established functional forms, so that the resulting model is able to reproduce a dataset used for reference. The method can be used to optimize parameters of models of any kind, a problem that is often encountered both in the physical and social sciences. It is necessary to point out that the accuracy of the estimates is strictly dependent on the flexibility of the model and on his suitability to reproduce a given kind of data.

Are there any publications which describe the method in more detail?

The statistical method is illustrated in the seminal work "Force Field Parametrization of Metal Ions from Statistical Learning Techniques"[4]. Currently, we have introduced two small variations in the application of the method to proteins: 1) the clusters systems used as reference models in this case must be saturated with -H o -OH moieties; 2) a more efficient metaheuristic algorithm has been introduced to perform the optimization of the hyper-parameters, namely the basic version of differential evolution[5] has been replaced by Success-history based parameter adaptation for differential evolution[6]. These improvements will be described in future publications.

[4] Fracchia, Del Frate, Mancini, Rocchia, Barone *J. Chem. Theory Comput.* **14** (2018) 255
[5] Storn, Kenneth *J. Glob. Optim.* **11** (1997) 341
[6] Tanabe, Fukunaga In Evolutionary Computation (CEC), IEEE Congress on pp. 71-78 (2013)

Funded by the European Union under grant
Number 676531 (Project E-CAM)