D6.2

Definition and planning of new MaX Demonstrators

Andrea Ferretti, Alberto García, Luigi Genovese, Stefano de Gironcoli, Ivan Marri, Elisa Molinari, Pablo Ordejón, Deborah Prezzi, Daniele Varsano

Due date of deliverable:	31/05/2020
Actual submission date:	31/05/2020
Final version date:	31/05/2020
Revised version date	15/01/2021
Revised version submission date	19/02/2021
Lead beneficiary:	ICN2 (participant number 3)
Dissemination level:	PU - Public

Document information

Project acronym:	MaX
Project full title:	Materials Design at the Exascale
Research Action Project type:	European Centre of Excellence in materials modelling, simulations and design
EC Grant agreement no.:	824143
Project starting / end date:	01/12/2018 (month 1) / 30/11/2021 (month 36)
Website:	www.max-centre.eu
Deliverable No.:	D6.2

Authors: A. Ferretti, A. García, L. Genovese, S, de Gironcoli, I. Marri, E. Molinari, P. Ordejón, D. Prezzi, D. Varsano

To be cited as: Ferretti et al., (2020): Definition and planning of new MaX Demonstrators. Deliverable D6.2 of the H2020 project MaX (final version as of 31/05/2020). EC grant agreement no: 824143, ICN2, Barcelona, Spain.

Changes and contribution history

	Change	Author	Notes
1	Sections 3.1 - 3.2: info requested included.	P. Ordejón, D. Prezzi, L. Genovese	Included information on performance goals and resources needed and secured for the demonstrators in sections 3.1 to 3.4.
2	Section 3.3: demonstrator project modified.	S. de Gironcoli, P. Ordejón, A. Ferretti	We have modified the third demonstrator. We removed the one previously proposed on synthetic redox-active metallo-peptides (which, regarding the computational aspects was somewhat similar to the demonstrators in Sections 3.1 and 3.2), and added a new one on "High-Throughput Machine-Learned DFT-quality Potentials", which connects with our efforts in bridging our simulations with methods based on artificial intelligence approaches.
3	New references and figures included	P. Ordejón, L. Genovese, S. de Gironcoli	Numbering of references changed. Added new references (1, 2, 3, 16 and 17) and figures (4, 5 and 8). Former references 13 and 14 have been removed.

D6.2 Definition and planning of new MaX Demonstrators

Content

1 Executive Summary

This deliverable reports on three new demonstration projects which follow from the progress of the previous Demonstrators being deployed in MaX, and from the research being carried out by different MaX participants in collaboration with researchers in different scientific fields: (i) analyzing the interactions between the SARS-CoV-2 main protease and compounds candidate for inhibiting its function; (ii) electronic processes in electrode-electrolyte interfaces; and (iii) the development of high-throughput machine-learned potentials with DFT quality. All three will make full use of the capabilities of MaX flagship codes in the context of pre-exascale infrastructures available during MaX.

2 Introduction

The progress done in the previous Demonstrators being deployed during the first half of MaX was presented in Deliverable D6.1. From that work, combined with the new capabilities provided by the technical advances made in the MaX codes and their integration with AiiDA, and from the research being carried out by different MaX participants in collaboration with researchers in different scientific fields, new demonstration projects have been devised, as explained in this document.

The first project, on SARS-CoV-2 proteins, reflects the response of the MaX community to the COVID-19 worldwide crisis. Notwithstanding that the main focus of the MaX flagship codes is centered on materials science, the study of SARS-CoV-2 proteins, due to the size and complexity of the system, offers the opportunity to put into value the power of the MaX codes in the context of pre-exascale infrastructures and to provide valuable information about the interaction of proteins relevant for the biological function of SARS-CoV-2 (in particular, the main protease) from its electronic structure. In particular, we will study the interactions of the protein with compounds which are candidates for inhibiting the function of the protein, therefore being potentially drugs to combat the disease. The project stems from the interaction and collaboration of MaX groups with the community of biologists trying to develop antiviral drugs using in-silico approaches.

The second project deals with several aspects of the interaction of electrolytes with solid state electrodes. The efficiency of the MaX flagship codes (in particular Quantum ESPRESSO, SIESTA and Yambo) in GPU-accelerated platforms will be exploited to be able to reach the large systems sizes and long simulation times necessary to address problems of great relevance for energy conversion and storage, and for the prevention of corrosion in exposed materials. Also, some unique features of the MaX codes (as the capability of SIESTA to deal with non-equilibrium situations in which a voltage is applied and an electrical current is established) will be key to address some of the problems in this field. The project is framed in the participation of MaX's groups in international collaboration projects, such as the large-scale European project BIG-MAP, which is part of the BATTERY 2030+ initiative and orchestrates automated and artificial intelligence solutions for the accelerated research and discovery of battery materials.

Finally, a third project focuses on the obtention of affordable, accurate and widely applicable interatomic potentials for the modelling of materials. We integrate techniques based on the generation of machine-learned potentials, with high throughput protocols for massive DFT calculations to obtain the data to train the potentials. The demonstration combines machine-learning techniques, high-throughput protocols, code performance estimators, and massive and efficient DFT calculations. The objective is to obtain reliable potentials for challenging applications in materials science.

In addition to the three projects, described below in more detail, new scientific projects enabled by the evolution of the MaX codes will be considered during the next period, taking also into account the feedback and the input from the community of code users and the actual availability of pre-exascale computational resources.

3 New Demonstration Projects

3.1 Addressing SARS-CoV-2 molecular constituents from electronic structure: fragment analysis with BigDFT and SIESTA.

Scientific Framework: Given the critical pandemic situation that the world is currently facing, an unprecedented effort is being devoted to the study of SARS-CoV-2 by researchers from different scientific communities and groups worldwide. From the biomolecular standpoint, particular focus is being devoted on the protease and on the spike protein. The protease is found within the virus core along with the nucleocapsid protein and RNA. It is an essential enzyme for the life-cycle of the virus as it produces structural and functional proteins that are required for the maturation and replication of the virus. As such, it is an important potential antiviral drug target: if its function is inhibited, the virus remains immature and non-infectious. Using fragment-based screening, researchers have identified a number of small compounds that bind to the active site of the protease and can be used as a starting point for the development of protease inhibitors.

The work proposed in this new Demonstrator builds upon the advances made already within MaX using both SIESTA (as described in MaX Deliverable D6.1, section 3.1.1) in performing structural optimizations of the experimental structure of the main protease, and BigDFT (as described in MaX Deliverable D6.1, section 3.2) in reducing the complexity of large scale systems and the analysis in terms of coarse grain fragments.

We have demonstrated that SIESTA is currently able to perform short molecular dynamics simulations and structural optimizations of the monomer of the SARS-CoV-2 main protease (M^{pro}) in a water environment, even with standard diagonalization. The usage of algorithms with reduced scaling like PEXSI or the linear scaling solvers developed within MaX will only improve this situation, making the calculations feasible for even larger systems (like the dimer structure of M^{pro}, relevant for its biological function) and for longer simulation times. The same holds for BigDFT, thanks to the development of its linear scaling approach. We now have the possibility to model the electronic structure of the protease in contact with a potential docked inhibitor, and provide new insights on the interactions between them by selecting specific amino-acids that are involved in the interaction and characterizing their polarities. This new approach we propose is complementary to the docking methods used up

to now and based on in-silico research of the inhibitor. We have started a series of calculations, taking advantage of the PDB structures available. In this section, our main objective is to show a simplified demonstration of a computational approach based on our fragmentation scheme that would be accessible to other scientific communities, like biologists or medicinal chemists, who may be able to extract new ideas from data presented as follows. Although the fragment analysis approach has already been implemented and tested within BigDFT, we will adapt it also to SIESTA.

Biological systems are naturally composed of fragments such as amino-acids in proteins or nitrogenous bases in DNA. We show in this example the SARS-CoV-2 main protease (PDB ID 6LU7) in complex with an N3 Inhibitor. Such a structure is made of a dimer of two identical subunits, each one with a docked inhibitor. We made our calculation by presenting only one monomer, made of 4732 atoms, depicted in Fig. 1. Such a biological system is made of two chains: one associated with the amino acids which belong to the enzyme, and the second associated with the inhibitor.

With our approach we are able to evaluate whether the amino acid-based fragmentation is consistent with the electronic structure resulting from the QM computation. Such a system has already been analyzed at a QM level of theory with fragment molecular orbital technique, where the fragmentation of the system has to be imposed beforehand following chemical intuition. Here, to evaluate the reliability of the model we have at our disposal the purity indicator that gives us, for each fragment, the level of confidence with which such a fragment can be considered as an independent unit of the system. This is an important indicator for the end-user, as it enables to evaluate the quality of the information associated with a given fragment. Usually, a cutoff of 0.05 for the purity indicator is employed, which has proven to provide meaningful physico-chemical results in most circumstances. Such values are presented in Fig. 2, with a color code showing their fulfillment with the cutoff value mentioned above. We see that the Guanine fragments, in particular, are not classified as pure in our scheme. This would imply that such fragments should be merged with their neighbors in order to be considered as meaningful QM entities. This is a typical situation as the Guanine is the only amino-acid that does not have a lateral structure, therefore it wants to connect itself to the other. In the same figure, a refragmentation of the system that fulfills the cutoff of 0.05 is presented.

We have shown in the Laccase example in MaX Deliverable D6.1 how a coarse grained view of the system is important. We obtain QM observables on the system's fragments, which are based on a population analysis of electronic density of the system, projected on the amino-acid. A novelty that our approach enables is the possibility of quantifying the strength of the chemical interaction between the different fragments. It is possible to select a target region and identify which fragments of the systems interact with this region by sharing electrons with it. We can reconstruct the fragmentation of the system in such a way as to focus on an active site in a specific portion of the protein. In this example, we will focus around the inhibitor. We show in Fig. 3 which are the sections of the amino-acidic sequence that have a non-negligible interaction with the fragments that belong to the chain of the inhibitor. Such representation can be transformed in a graph-like view like in the case shown in MaX Deliverable D6.1, where the interacting fragments may also be characterized by their QM charge.

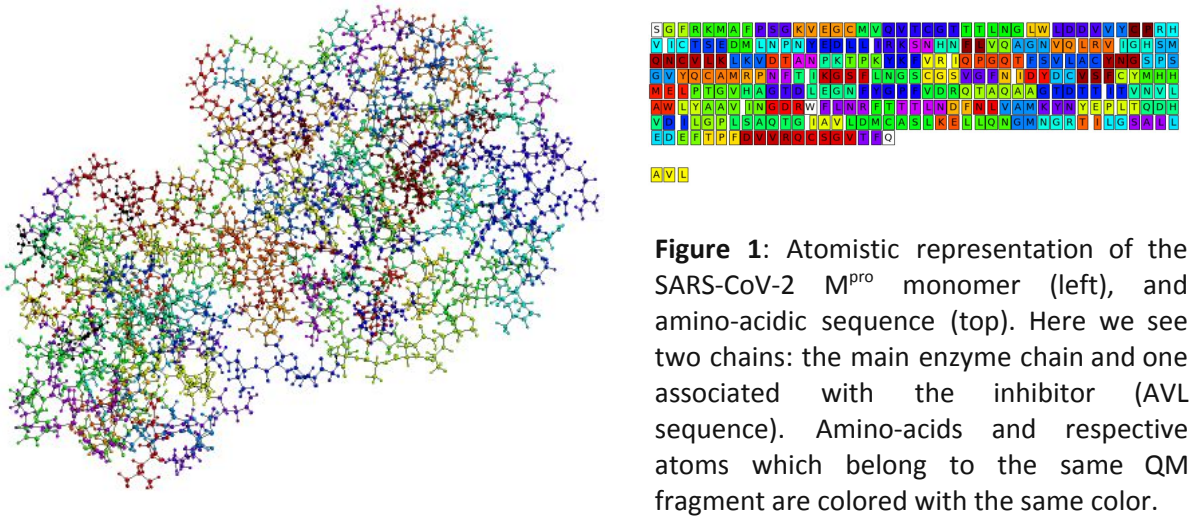


Figure 1: Atomistic representation of the SARS-CoV-2 M^{pro} monomer (left), and amino-acidic sequence (top). Here we see two chains: the main enzyme chain and one associated with the inhibitor (AVL sequence). Amino-acids and respective atoms which belong to the same QM fragment are colored with the same color.

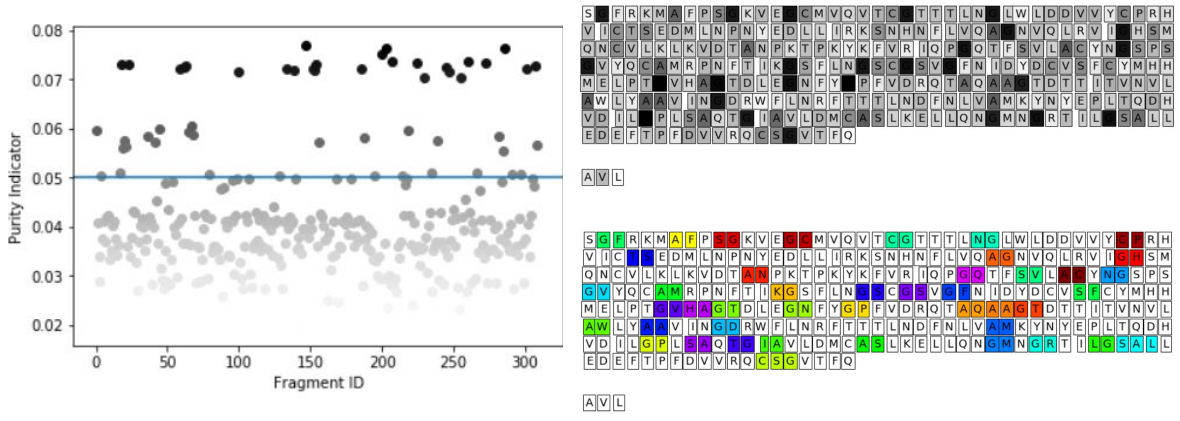


Figure 2: Pertinence of the QM fragmentation based on the amino-acids. Each amino-acid is associated with a fragment, and the value of the purity indicator is extracted (left). These values can then be mapped to the sequence chains of the system (right top). Fragments which have a value larger than the chosen cutoff are then re-fragmented with neighboring fragments and merged together. The fragmentation (right bottom) is made such that each fragment has a purity indicator below the desired cutoff. Amino-acids which belong to the same QM fragment are colored with the same color, whereas white amino-acids are already associated with pure fragments.

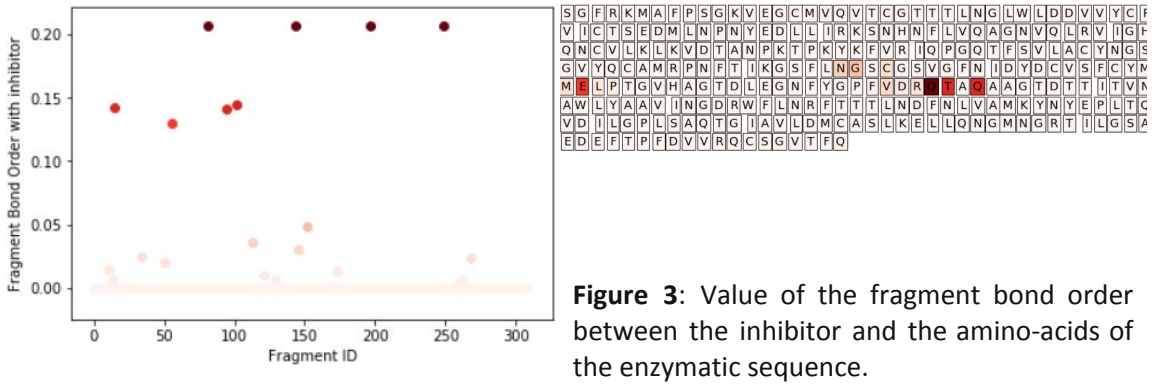


Figure 3: Value of the fragment bond order between the inhibitor and the amino-acids of the enzymatic sequence.

Case Description: In more specific terms, we focus our attention on two different systems:

- The SARS-CoV-2 main protease (Fig. 1) is being analyzed and fragmented with the same complexity reduction algorithms employed in D6.1. A large number of configurations of M^{pro} are being simulated, with the substrates being a number of tentative inhibitors of the enzyme's main functional process. The interaction map of the inhibitors in contact with M^{pro} can then be extracted from the interactions between the fragments. This approach enables us to identify interaction signatures for each of the inhibitors. We are collaborating with two different groups, in order to identify descriptors that can be built out of these indicators (see [1] for a first preprint under review).

  For the Mpro simulation we are presently able to simulate with BigDFT a single snapshot with 24 node hours on the IRENE TGCC supercomputer in France. We obtained a TGCC Grand Challenge of 15 Million CPU hours, and a new project has been recently granted to perform similar simulations in the Fugaku supercomputer. Taking into account the CPU to node ratio of Fugaku, we expect to be able to perform similar simulations as in IRENE in about 100 node hours. All these resources will enable us to complete a larger exploration of different inhibitors with long molecular dynamics runs.

- We are also applying that scheme to investigate antibody/antigen (AA) assemblies. As the size of these assemblies can be very large (exceeding 1000 residues), we have developed a sequential multi-scale Molecular Modeling/Quantum Mechanic (mMM/QM) scheme, where BigDFT is coupled to a MM approach [2]. This enables us to efficiently simulate on modern supercomputing architectures systems up to tens of thousands quantum atoms and million classical atoms, respectively. The mMM approach is used to sample the potential energy surface whereas the QM one is used to assess/refine the quality of the mMM approach and to identify AA domains pivotal to understand the AA assembly stability. We note that the domains can be far from the main AA contact regions. This project has been recently granted the SANOFI I-tech award.

  Fig. 4 show a picture of one the systems considered: 1YNT, a complex between the monomeric form of Toxoplasma gondii surface antigen 1 (SAG1) and a monoclonal antibody that mimics the human immune response [3], which contains 20 thousand

atoms and 1400 residues. Also shown is the obtained heatmap of the interactors between the epitope and paratope regions of the complex.

We have performed extensive benchmarks for these systems with the production version of the code that was available in early 2020 (BigDFT 1.9.0). Fig. 5 shows some tests for the 1YNT system obtained in the IRENE Supercomputer in France and the ARCHER Supercomputer in the UK. We observe that the performances can be improved on IRENE in terms of the communication patterns, due to load unbalance and communication of small data. We expect that solving these issues will allow a gain of about 30% of computation walltime in these systems. We plan to employ the Fugaku computing hours in order to develop solutions for this problem.

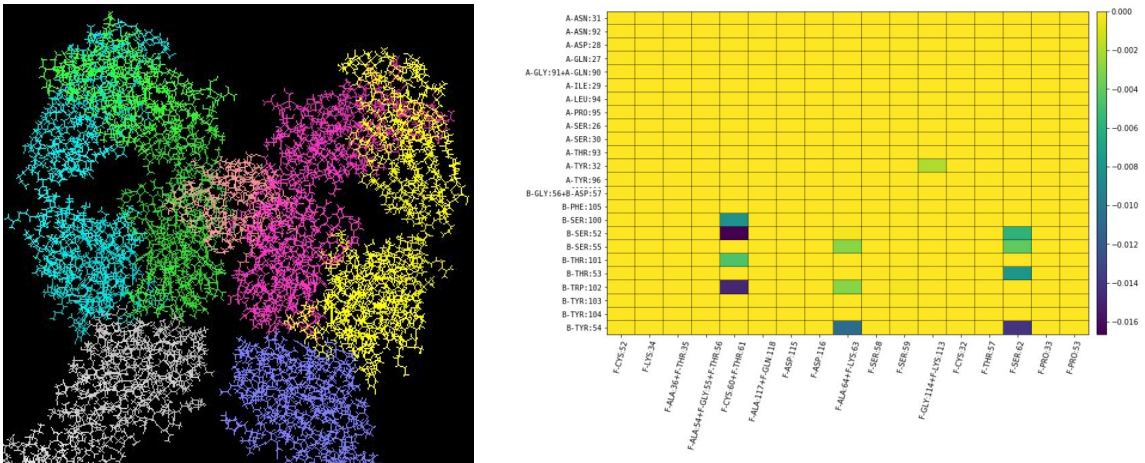


Figure 4. Left: 1YNT antibody/antigen assembly. Top: Heat map of interactions between the epitope and paratope regions of the complex.

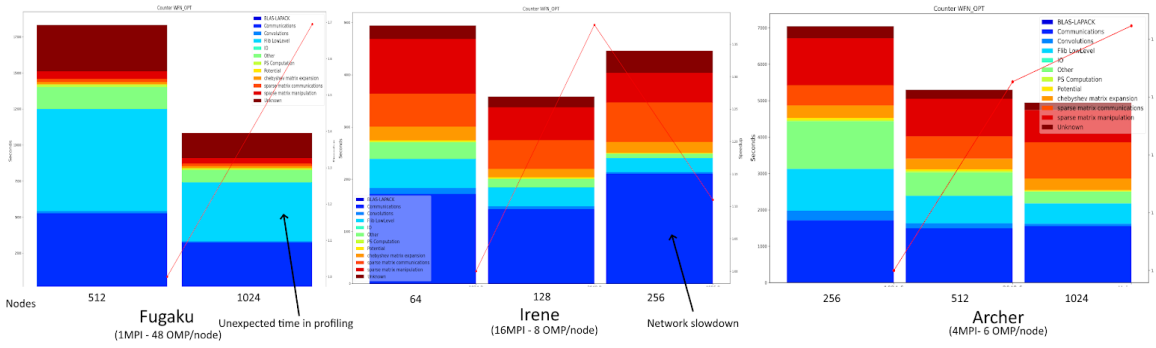


Figure 5. Benchmarks of runs for the 1YNT antibody/antigen assembly in Fugaku, IRENE and Archer.

3.2 Ab-initio investigation of electrode-electrolyte interfaces

Scientific Framework: In this demonstrator we aim at tackling the study of the properties of prototypical electrode/electrolyte interfaces (EEI) with accurate, highly predictive theories. Indeed, composition and electro-chemical processes taking place at the EEI are essential for many technologically relevant problems, such as the performance and stability in Li-ion batteries [4] and other (photo)electrochemical devices for energy conversion and storage, or processes involved in the corrosion of exposed surfaces of materials, to mention just a few. In view of their predictive power, fully ab-initio simulations are well-suited methodologies to approach the problem, especially when addressing novel or non-conventional electrode compositions and morphologies as well as the concurrent search of optimized electrolytes or the design of new materials for coatings. However, an accurate investigation of the whole EEI has been so far hindered by the prohibitive computational cost. In fact, theoretical studies have accurately addressed so far either the electronic structure of novel electrode materials [5-8], though neglecting the effects induced by the interaction with the electrolyte, or the energetic levels of different electrolytes, despite considering the isolate electrolyte components only, i.e. without including the explicit interaction between the solvent and the electrode surface [9]. Studies of the whole EEI are much scarcer, and so far reduced to simple electrodes (typically noble metals) and electrolytes (aqueous solutions). Moreover, in most of the cases, the interface is investigated neglecting the effect of an applied bias, thus not allowing to study the system in-operando conditions.

The difficulty in dealing with these systems at the DFT level lays in the first place on their complexity, as they involve the interface of the liquid electrolyte with an often already complex solid electrode surface. Very large systems (with many thousands of atoms) are usually necessary to model these interfaces in a minimally realistic manner. The liquid state of the electrolyte also forces one to perform very long molecular dynamics simulations in order to extract meaningful physical quantities like free energies. A second difficulty is that many of the relevant processes occur when the electrode is subject to an external voltage, and in the presence of electric current. This implies a non-equilibrium situation which is usually out of the scope of usual DFT technologies. Although some schemes have been devised to treat at least the situation in which an external voltage is applied and induces an electrified EEI [10-12], they have not been sufficiently tested and do not provide robust and general solutions to address these non-equilibrium situations. In addition, even at equilibrium (with no voltage or currents), the use of local and semilocal approximations for the DFT functionals, widely employed to simulate larger/realistic systems with affordable computational cost, in many cases is not sufficient to obtain an accurate quantitative (and sometimes also qualitative) description of the electronic properties of the EEI components, able to provide useful information for the experimental search.

This demonstrator arises in the context of several projects and collaborations in which the involved MaX groups participate. A particularly important one is the large-scale European project BIG-MAP, which is part of the BATTERY 2030+ initiative and sees the involvement of three MaX's groups, i.e. CNR, ICN2 and EPFL.

MaX can provide an impulse to this important field. In the context of pre-exascale and exascale infrastructures, the MaX codes are ready to make an impact, in several ways:

- We will take advantage of the **GPU porting of the relevant codes implementing ab initio molecular dynamics (AIMD)**: cp.x for Car-Parrinello molecular dynamics (CPMD) and pw.x for ground state DFT from Quantum ESPRESSO, and SIESTA for ground state DFT and AIMD. The demonstrated performance of the codes (see D2.1, D2.2, D4.2, D4.3, D6.1) will allow extensive ab-initio molecular dynamics simulations in systems of the complexity needed to describe the EEI properly.

- SIESTA has the **capability to perform calculations in systems where a voltage bias is imposed between two electrodes**, using Non-Equilibrium Green's Functions, as implemented in the TranSIESTA method [13,14]. We have already advanced in using this capability to study electrified EEI's between noble metals and aqueous solutions. As an example, and for demonstration purposes, Fig. 6 shows results of a simulation of water between two gold electrodes, which indicate how the dipoles of the H_2O molecules orient according to the voltage applied to the electrodes [15]. Fig. 7 shows the diffusion and drift of Na+ and Cl- ions in solution, with and without the presence of a bias potential between the electrodes.

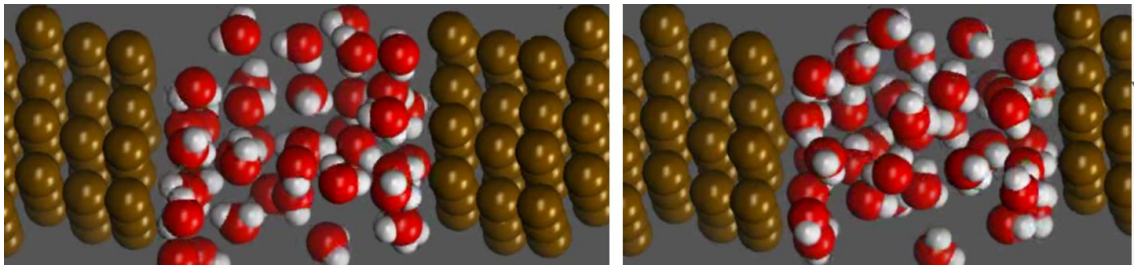
- **Excited state electronic properties** can be addressed at the GW level with Yambo, which was also **demonstrated in GPU-accelerated infrastructure** (see D2.2, D4.2, D4.3, D6.1).

- These three codes have been fully **integrated through the AiiDA platform** (see D5.3), for which we have already demonstrated an automated functionality for "on-the-fly" calculations (see D6.1).

Case description. While specific cases will be chosen in collaboration with the partners as the projects start, we will initially address prototypical electrolyte-electrode interfaces specific for the two research lines described below.

As for Li-ion batteries, we will combine ab-initio molecular dynamics (AIMD) and DFT approaches to investigate redox reactions at the electrode-electrolyte interface by considering an explicit solvent model to describe a simple aqueous electrolyte environment and by adopting a graphitic material as anode. The ionic mobility at the electrode/electrolyte interface will be studied by using AIMD simulations under the presence of a bias included using the TranSIESTA code. The electronic properties of the interface will be then investigated within the GW approximation on selected snapshots of the dynamics. This very demanding computational step is fundamental to produce an accurate line-up energy profile between electronic levels of electrolyte molecules and the anode chemical potential, which is of key importance to determine the electrochemical window and thus the stability of the EEI, and represents a preliminary step toward a more exhaustive multiscale and multiphysics modeling of a complete device.

The second line of research will focus on the design of novel corrosion inhibitors for metallic surfaces by means of AIMD simulations under the presence of a bias (TranSIESTA code). Indeed the chromate coatings commonly used are being banned due to the toxicity and high carcinogenicity owed to the hexavalent chromium they contain, and there is a high quest for new, non-toxic coatings. This project is done in collaboration with the group of Prof. Ivan Cole at RMIT (Australia), and linked to industrial contracts.



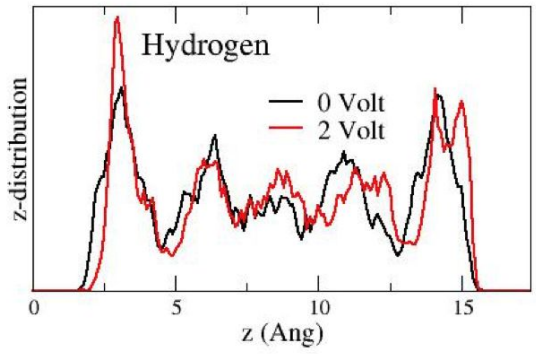


Figure 6. Top panel: snapshots of AIMD simulations of water between two gold electrodes, for two different values of the voltage applied between electrodes: 0 Volt (left) and 2 Volt (right). The orientation of the hydrogen atoms towards the right electrode is apparent, to align the dipoles with the electric field. Left panel: distribution of H atoms along the simulation cell, showing the changes induced by the application of the voltage between the electrodes. Results obtained with SIESTA [15].

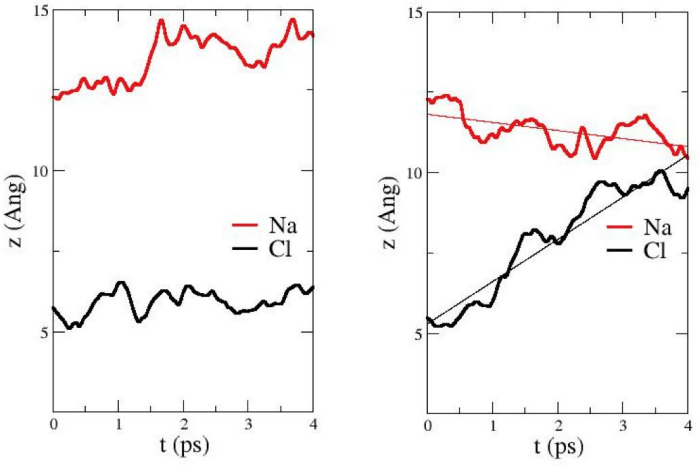
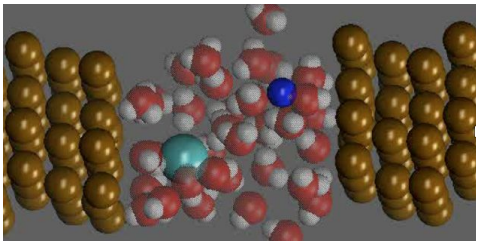


Figure 7. Top: Position of a Na+ cation and Cl- anion along the z direction (perpendicular to the surface of the electrodes) during an AIMD simulation of the electrolyte in contact with gold electrodes (Right). In the top-left panel, no voltage is applied, while the top-right panel shows the drift driven by a voltage applied between the electrodes. Results obtained with SIESTA [15].



Tackling these two projects will require calculations with an extremely large number of atoms, which are not feasible for purely QM methods. We will use a hybrid QM/MM approach developed by us [16] where part of the system is solved with QM and the rest with classical interatomic potential (molecular mechanics MM), and implemented in SIESTA. During the last few months we have optimized the efficiency of the MM module for parallel execution. Concerning the TranSIESTA module, work will be done within WP1-4 to produce a massively parallel, GPU enabled code. Although currently TranSIESTA has two levels of parallelization (MPI parallelism over the number of points in the energy integration contour, and OpenMP threading), this only allows for efficient execution in up to around 480 threads (e.g., 10 nodes in MareNostrum4). We aim at including an extra level of MPI parallelism and support for GPUs using parallel linear algebra libraries (ELPA and ELSI) the same way we did for the diagonalization solvers in WP1 (see D1.1, D2.1, D2.2, D4.3). Using this three-level parallelism we expect to be able to use around 200 nodes of MareNostrum4 (Intel Xeon CPUs), and 250 nodes in Marconi100 (IBM Power CPUs and NVIDIA GPUs) in each single run (while several runs can be executed simultaneously in a high-throughput mode to sample different configurations).

For the excited state calculations at the GW level, a minimal initial goal is to study a prototype system with ~500 atoms (a very large size for this type of calculations). This was estimated as affordable in the currency available HPC infrastructures by considering both the GPU porting and very good performance demonstrated for the Yambo code on Marconi100 for related systems, the limitations set by the size of the matrices to invert (see discussion in D6.1), and the current availability of resources (only up to 250 nodes on Marconi100, i.e. 1000 GPU cards). By considering the timings reported in D6.1 and further tests on a supercell with 64 water molecules, we estimated that we need 750k hours on Marconi100 for the single GW run to reach the above mentioned system sizes (estimated by considering a N^4 scaling formula).

As a second step, we plan to further push the system size to the extreme affordable case. As mentioned in D6.1, the memory footprint is a critical issue, making it of fundamental importance to test the GPU porting of the code against large systems. A second, key point is related to the lack of a parallel distributed linear algebra library for dense matrices on GPUs, which determines the maximum size of the matrices to invert. In a continuous feedback-loop with the technical WP1-4, we plan to test any further development along these lines. Moreover, depending on the actual need with respect to the goals, we might ask to prioritise specific development actions, e.g. the development of a workaround for matrix inversion or the use of a more efficient interpolation strategy for matrix evaluation on the k-point grid. Another development that might become of key importance in this field is a GW-enabled polarizable continuum model (PCM) approach, which allows one to reduce the number of atoms (and electrons) considered explicitly in the GW calculation.

To start this demonstration activity, we have already obtained about 1.5M hours in MareNostrum4 (where the demonstrations presented in [15] were executed) and 1.5M hours on Marconi100. Moreover, we applied for a further project in MareNostrum4 (3.8M hours), and plan to apply for a PRACE project (6M hours) in the next call.

3.3 High-Throughput Machine-Learned DFT-quality Potentials

A new Demonstrator is proposed that integrates the development of techniques aimed at the generation of machine-learned potentials trained on DFT data for challenging applications in materials science (Li-ion batteries, silica degradation due to radiation damage, thermal properties of complex materials, to name a few possible applications of current interest). Availability of affordable and widely applicable interatomic potentials is key to unlock the riches of modern materials modeling, as they will allow simulations with an accuracy similar to that of DFT used to train the potentials, but at a very small fraction of the computational workload of an actual DFT calculation.

The development of machine-learned potentials will be enabled by high throughput protocols and code performance estimators for the massive and efficient calculation of the needed DFT training energies. A sketch of the proposed self-consistent scheme to generate an accurate and transferable neural network potential (NNP) with DFT accuracy is shown in Fig. 8 [17]. The initial step to start the process (yellow arrow) can be performed with a classical force-field as shown here, or with a first NNP model (light blue triangle) generated from any comprehensive dataset of structures such as the ones present in publicly accessible repositories (Aflowlib, Materials Genome Initiative, Nomad, Materials Cloud…). Once an initial model potential is selected it is refined with the following steps: i) crystal structure exploratory techniques (Evolutionary Algorithms, Random Sampling, Particle Swarm Optimization or other methods) are employed to generate a diverse set of energetically and structurally relevant candidate structures; ii) a subsequent clustering-based pruning of structures ensures that no single polymorph biases the dataset since, at each self-consistent step, only novel structures (red and blue disks for the particular step highlighted in the Fig. 8) are to be considered, further refined, and added to the dataset of candidate structures; iii) subsequent Molecular Dynamics (MD) simulations sample the potential energy surface of the different polymorphs around their thermal equilibrium, always with the inexpensive NNP; iv) finally expensive DFT calculations (yellow box) are performed on a sparse subset of independent MD-sampled structures (red and blue porous discs) and added to the ab initio dataset obtained thus far (other colors porous discs); v) the thus augmented dataset is used to train the next NNP model (a darker blue triangle), starting the next cycle of the self-consistent scheme until no new structures are found in the process and the final NNP is obtained. In our implementation the NNP training will be performed either with the PANNA (Properties from Artificial Neural Network Architectures) code, developed in SISSA in collaboration with researchers in Harvard, ENS (Paris) and Toulouse, or the DeepMD code, developed at Princeton.

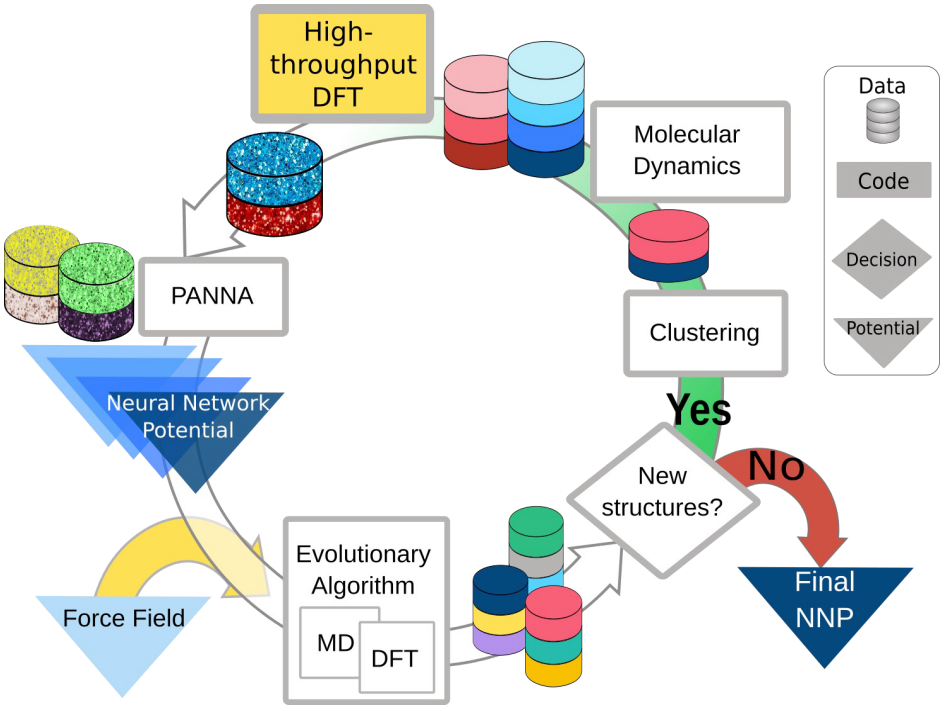


Figure 8. Self-consistent scheme for neural network potential generation (adapted from ref [17]).

The only significantly expensive step in the procedure outlined above is the calculation of the DFT energy for the new set of candidate structures at each iteration of the self consistent cycle. These sets will comprise thousands of structures with possibly different numbers of atoms and geometries around the thermal equilibrium of the different polymorphs. Efficient evaluation of these energies exploiting future exascale resources can be obtained via high throughput protocols as each configuration is independent from the others and they can all be simultaneously evaluated in parallel; for instance, the AiiDA infrastructure, developed and maintained within the MaX project, could be used to manage the resulting massive computational work-flows and data book-keeping. This is a marked (manyfold) advantage with respect to the brute force approach where the system of interest is directly computed via ab-initio methods: on one hand the number of total system configurations to be computed is set by the need of properly sampling the system diversity but not by the time-scale of the simulation that may entail long “uninteresting” periods where the system simply vibrates; on the other hand, the dimension of the systems needed to describe the relevant structural diversity might be much smaller than the one needed to display the physical phenomenon of interest; finally as already mentioned, the needed DFT calculations are many but all independent and can therefore exploit the high-throughput paradigm efficiently, at variance with long MD simulations that are intrinsically serial and can face severe scalability issues.

For instance the generation of the general purpose NN potential for carbon described in our recent work [17] required only a total of ~60k DFT single point calculations in cells containing 16-32 atoms extracted from hundreds of trial MD runs performed with inexpensive NN potentials (roughly 15-25000 for each generation in the potential generation scf cycle). A

total of about 300k CPU core-hours where required by the DFT calculations. This allows inexpensive simulations in cells with thousands of atoms for long simulation times that would require computational resources 2 to 3 orders of magnitude larger if tackled directly at the DFT level.

Efficiency in the evaluation of the energy of each individual configuration is also important and the optimization of the computational parameters (required resources, adopted parallelization strategies, availability of novel hybrid architectures) involved in the code execution will be important to achieve overall cost effectiveness. The adoption of code performance prediction tools targeted in WP5 will help address this issue.

4 References

[1] L. Genovese et al., (2020) Preprint. https://doi.org/10.26434/chemrxiv.12924974.v2
[2] http://biodev.cea.fr/polaris/
[3] Structure taken from https://www.rcsb.org/structure/1ynt
[4] M. Gauthier et al., J. Phys. Chem. Lett. 6, 4653 (2015)
[5] Y. Shaidu et al. J. Phys. Chem. C 122, 20800 (2018)
[6] F. Zhou et al., Appl. Surf. Sci. 463, 610 (2019)
[7] J. Zhuang et al., Adv. Materials 29, 1606716 (2017)
[8] P. Nie et al., Small 14, 1800635 (2018)
[9] O. Borodin et al., J. Phys. Chem. C 117, 8661 (2013)
[10] S. Surendralal et al., Phys. Rev. Lett. 120, 246801 (2018)
[11] R. Khatib et al., arXiv:1905.11850.
[12] C-Y. Li et al., Nature Materials 18, 697 (2019)
[13] M. Brandbygeet al., Phys. Rev. B 65, 165401 (2002).
[14] N. Papior et al., Computer Physics Communications 212, 8 (2017).
[15] P. Ordejón et al. (to be published).
[16] C. Sanz-Navarro et al., Theor Chem Acc 128, 825 (2011).
[17] Y. Shaidu et al., arXiv:2011.04604v1; npj Comput Mater to appear (2021).