D5.2

First report on verification and validation of codes and on the data analytics pilots

Nicola Marzari, Miki Bonacci, Pietro Bonfà, Michele Ceriotti, Stefaan Cottenier, Andrea Ferretti, Alessandro Laio, and Daniele Varsano

Due date of deliverable: 31/05/2020
Actual submission date: 01/06/2020
Final version: 01/06/2020

Lead beneficiary: EPFL (participant number 6)
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.:	D5.2

Authors: N. Marzari, M. Bonacci, P. Bonfà, M. Ceriotti, S. Cottenier, A. Ferretti, A. Laio, and D. Varsano

To be cited as: N. Marzari et al. (2020): First report on verification and validation of codes and on the data analytics pilots. Deliverable D5.2 of the H2020 project MAX (final version as of 01/06/2020). EC grant agreement no: 824143, EPFL, Lausanne, Switzerland.

Disclaimer:

D5.2 First report on verification and validation of codes and on the data analytics pilots

Content

1 Executive Summary

Key objectives of work package 5 ("Ecosystem for HPC, HTC and HPDA convergence") are the identification of protocols for the verification of materials science codes, and the development of algorithms and tools for high-performance data analytics in materials space. In this respect

- We have developed a protocol to verify any density-functional theory code under a broad range of diverse chemical environments, extending our previous work on elemental crystals and making it possible to span all possible oxidation states of every element, and comparing equations of state.

- we have developed a protocol identifying 500+ reference inorganic materials (insulators, semiconductors, metals, and magnetic or non magnetic) to verify CPU vs GPU implementations on the calculations of total energies, forces, and stresses

- we have developed a protocol to test high-accuracy pseudopotentials developed for excited state calculations that have the capability to reproduce all-electron results

- we have developed an automatic workflows to identify optimal parameters for the convergence of GW calculations and applied it to the GW100 dataset

- we have developed machine-learning models to predict the time-to-solution of electronic-structure codes

- we have developed an online interactive visualization and exploration library for high-performance data analytics called *chemiscope*, of which a demonstration version is available at https://chemiscope.org. This has been optimized for on-demand actions, so that additional data are fetched dynamically, and the applications is able to display, update, and interact with large dataset smoothly and efficiently

- we have identified a reference database of 85,000 stoichiometric inorganic materials that have been characterized experimentally, and we are calculating and disseminating computational data obtained with curated and reproducible AiiDA workflows

2 Verification and validation of codes

2.1 Verification for DFT (ground state) calculations for materials

In a major verification effort, a few years ago the precision of 40 different numerical implementations of density-functional theory (DFT) for materials has been assessed by a team effort directed by one of the present PIs (SC), involving more than 70 collaborators and using a test set of 71 elemental crystals[1] for which the equation of state of elemental solids was calculated independently using these 40 different codes. This so-called "Δ-project" led to the conclusion that the mainstream numerical methods and code implementations are in very good agreement with each other - which was not the case a decade before. Despite already being a large project by itself, this was only the first step of a long way towards eventually answering the question of reproducibility and precision in DFT calculations. The next key question is assessing to which extent the conclusions obtained from this small test set hold if there is much more chemical diversity, as can be probed by a larger and more diverse test set. That is what has been examined in the present project.

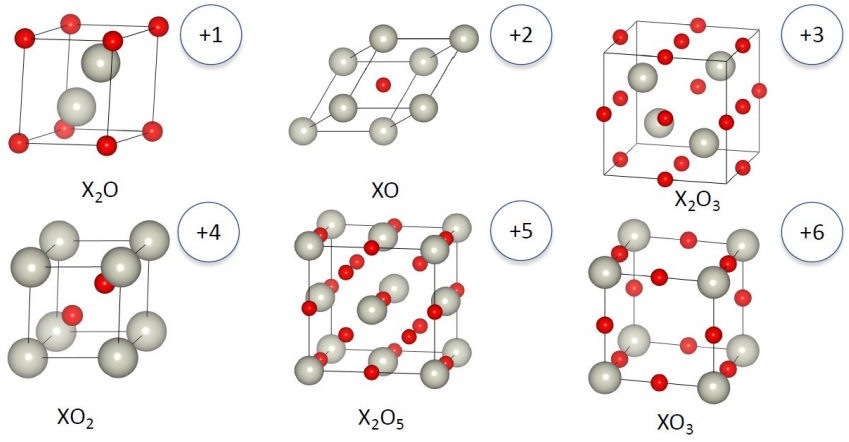


Figure 1: The 6 different oxides chosen as reference configurations, spanning all relevant oxidation states.

In order to scan chemical diversity in a systematic way, we examine for every element X (where X runs from hydrogen to curium) six different cubic oxides with stoichiometries that are chosen in such a way that the formal oxidation state of element X varies from +1 to +6. The six cubic oxides in ascending order of formal oxidation state have as representative chemical formulae X_2O, XO, X_2O_3, XO_2, X_2O_5 and XO_3. These formal oxidation states are used as a gauge for many different chemical environments, without claiming to represent the charge state of element X. By imposing a wide range of oxidation states, we ensure that a variety of chemical environments can be scanned for every element X. Not all of these environments might easily be found in real crystals, but DFT methods should be able to determine the numerically correct DFT solution.

Two codes were used in the first stage of this work: WIEN2k for the all-electron calculations, and VASP as a pseudopotential/PAW code. This is a temporary choice, meant to prepare and

[1] K. Lejaeghere et al., *Reproducibility in density functional theory calculations of solids*, Science 351, 25 (2016). (DOI: 10.1126/science.aad3000)

test the oxide data set. Once the test conditions are finalized, it is straightforward to run these benchmark calculations for other MAX codes using the AiiDA workflows for the equation of state. All 576 oxides -- including many that do not exist in reality and for which therefore no structural data are known -- are first geometry-optimized to determine the DFT-predicted equilibrium volume at the PBE level for the exchange-correlation functional. These approximate equilibrium volumes are frozen, and represent the "0% volumes" reference for each of the oxides. Then, for each system, 7 DFT total energies are calculated at volumes that differ ±6%, ±4%, ±2% and 0% from the frozen reference volume, and a Birch-Murnaghan equation of state is fit for every oxide. This is repeated for every DFT code. Ideally, the equation of state obtained by two DFT codes for the same oxide, should be identical. In reality, there is a small difference, which is expressed by a quantity called Δ_i (averaged over all 6 oxides for one given element i) or Δ (average over all Δ_i, i.e. over all elements).

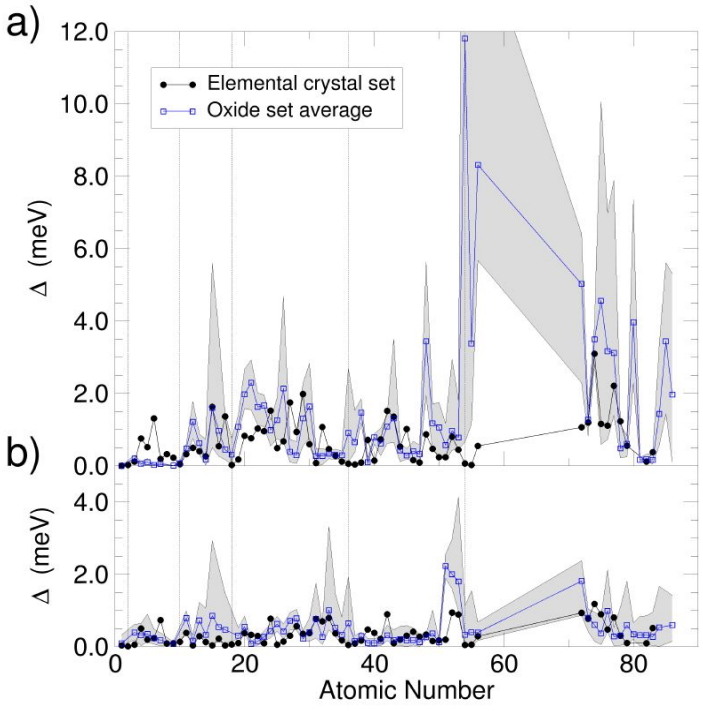


Figure 2: Panel a): calculated Δ_i-values for VASP (default 2018) vs. VASP (GW ready). The black symbols are the Δ_i for a test set of elemental crystals (only one crystal per element) from Ref. (1). The open blue symbols are the Δ_i averaged over the 6 oxides. The shaded area indicates the oxides with the smallest and largest value of Δ for each element. Panel b): calculated Δ_i-values for WIEN2k vs. VASP (GW ready).

Panel b) of figure 2 shows the calculated Δ_i-values for WIEN2k vs. VASP, the latter with a recent set of PAW potentials called "GW-ready". The black symbols are the Δ_i for a test set of elemental crystals (only one crystal per element) from ref. (1). The open blue symbols are the Δ_i averaged over the 6 oxides. The shaded area indicates the oxides with the smallest and largest value of Δ for each element. The order of magnitude of Δ_i is not significantly different between both sets of crystals, indicating that the VASP-GW potentials are a trustworthy representation of the all-electron behaviour in a wide variety of chemical environments.

Panel a) of Figure 2 shows a different situation. Here, the two codes that are compared are VASP with the GW-ready potentials and VASP with another set of potentials that was the default set until 2018. The elemental crystal test set (black symbols) showed low values for Δ_i only. The oxide test set (open blue symbols) reveals large values of Δ_i, and therefore

large differences between both calculations. This is evidence that the oxide test set is much more sensitive to subtle differences between codes, and is especially useful for testing the quality of pseudopotential libraries. A paper will be submitted soon in which the reasons for the observed differences are analysed, and in which the test set and protocol are made available. This will allow scientists from other code communities to run the benchmark for their code, and to assess the level of agreement between DFT codes at a next level of scrutiny.

In order to verify the results of the codes in CPU or GPU mode, we have identified a set of 550 structures with wide range of chemical and geometrical properties to act as a reference benchmark, and have used AiiDA to run the automatic self-consistent calculations using the PwBaseWorkChain for Quantum ESPRESSO for all structures for either the CPU and GPU version, finding excellent agreement (see Fig. 3).

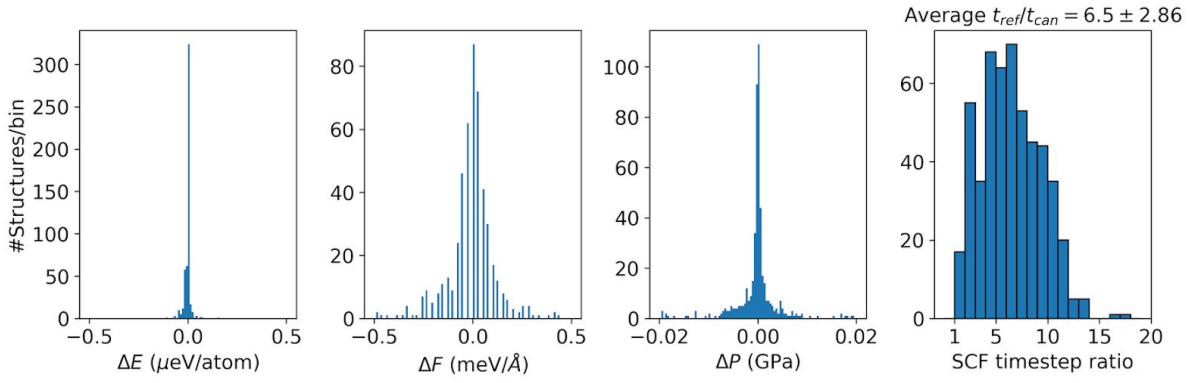


Figure 3: Agreement between energies, forces, and stresses for the SIRIUS-enabled GPU version of QE, showing a speedup of approximately 6.5 times compared to CPU only

2.3 Verification for G_0W_0 (excited state) calculations for materials

Reproducibility of quasiparticle calculations was recently addressed by comparing different implementations in three GPL codes for the GoWo approximations[2]. This effort involved 16 researchers, developers and users of the Abinit code, BerkeleyGW and the MaX flagship code Yambo, which analyzed the reason of the discrepancies that are often found in literature for instance in the prediction of quasiparticle gap of semiconductors, that can account up to 1eV for some metal-oxides depending on the code used. GW calculations typically consist of different steps that include the starting ground state usually calculated at DFT level, the evaluation of the exchange self energy, the screened dynamical coulomb interaction, the construction of the correlation self-energy and the solution of the quasiparticle equation. Each of these steps involve different approximations as the exchange

[2] T. Rangel, M. Del Ben, D. Varsano, et al., Validating GoWo codes for solids, Comput. Phys. Comm. in press (2020). DOI: 10.1016/j.cpc.2020.107242

and correlation potential used for the starting point, the treatment of the Coulomb divergence, the modeling of the dynamical part of the screened potential, often treated using the plasmon-pole approximation, different schemes for the quasi-particle equation solution and truncation of the sum-over-state in the evaluation of the polarizability and GW convolution for the correlation self-energy.

A systematic analysis of all of these approximations implemented in the different codes was carried out using as test cases prototypical systems as the very well studied bulk silicon, gold as paradigmatic metallic system where difficulties arise due to convergence issues, the non-negligible influence of semicore orbitals and the role of relativistic effects, and two metal-oxides: rutile TiO2, and wurtzite ZnO, a challenging and controversial system for GW where the quasiparticle band gap values reported in literature may differs by more than 1eV. The study allowed to trace back the primary origin of major discrepancies between codes reported in prior literature to be the different implementations of the Coulomb divergence in the Fock exchange term and the frequency integration scheme of the GW self-energy. Different techniques to treat the Coulomb divergence were benchmarked, and several effective approaches were identified (Fig. 4). A source of large discrepancies was also ascribed to the used plasmon-pole model, which for some systems can lead to large

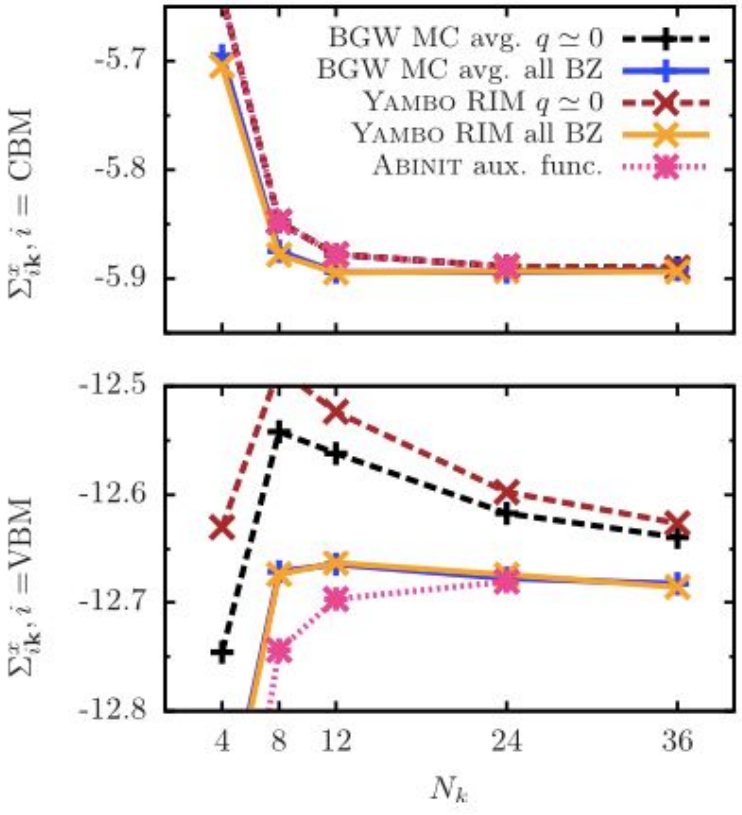


Figure 4: Convergence of the matrix elements of the exchange part of the Self-Energy for the Valence Band Maximum (VBM) and Conduction Band Minimum (CBM) at the Γ point for silicon, with respect to the number of k-points Nk ×Nk ×Nk. In the different codes, several techniques are used to treat the Coulomb singularity. When effective approaches such as the RIM for Yambo, MC average for BerkeleyGW and the use of auxiliary functions in Abinit are used, a perfect agreement among the codes is found.

deviations (>0.5 eV) from full frequency calculations. Importantly, it was shown that within judicious choices of approximations and the same

pseudopotentials, the converged GW quasiparticle energies calculated with the different codes agree within less than 0.1 eV. These results comprise an important verification of codes using the GW method for systems in the condensed phase, showing that different implementations can agree numerically at a level much greater than the known accuracy of the GW approximation and the underlying approximate Kohn–Sham eigensystem. Moreover, they provide a framework for users and developers to validate and document the precision of new applications and methodological improvements relating to GW codes.

2.4 Verification and validation for G_0W_0 (excited state) using the GW100 test set

As mentioned, the success of density-functional theory (DFT) in predicting material properties is clearly explained by the number and the relevance of publications that have been produced since the formulation of the theory, with the verification efforts discussed above. At the same time, many-body perturbation theory (MBPT) has been adopted from the community as the standard tool to describe electronic and optical properties of materials. Reflecting the DFT case, a lot of codes are available and first attempts of cross-validation were done in the past[3]. Among them, Yambo, one of MaX flagship codes, is a popular MBPT package which has been largely cross-validated with other softwares for what concerns GW calculations in solids (see also Sec. 2.3). On the other hand, a formal and complete validation over molecular systems is still lacking. In the past, some MBPT codes have been quantitatively compared on the GW100 set[4]: a group of 100 molecules used as a benchmark of the GW method. For all molecules of the set, computation of the ionization potential and electron affinity is done by means of GW approximation.

Among the GW data already present in the GW100 repository there are many full frequency GW calculations complemented by some data obtained using the plasmon-pole approximation (PPA) suggested by Hybertsen and Louie (HL-GPP). No results are available, though, using the so-called Godby-Needs PPA to model the dynamic screening matrix. Since this is the PPA flavour implemented in Yambo, our results will also give us the possibility to critically discuss the accuracy of the GN-PPA as compared to other PPAs or full frequency approaches. Therefore, the main objective of this work are:

- To analyze the 100 molecules belonging to the GW100 dataset using the Yambo code by means of high-throughput techniques to implement automatic convergence workflows. The analysis concerns the computation of the ionization potential and electron affinity from the quasiparticle energies of the HOMO and LUMO orbitals, obtained by means of the G0W0 calculations. In doing so, we will make use of the GN-PPA model, as implemented in Yambo, to describe the frequency dependence of the response function. By doing so, we will be able to complement the GW100

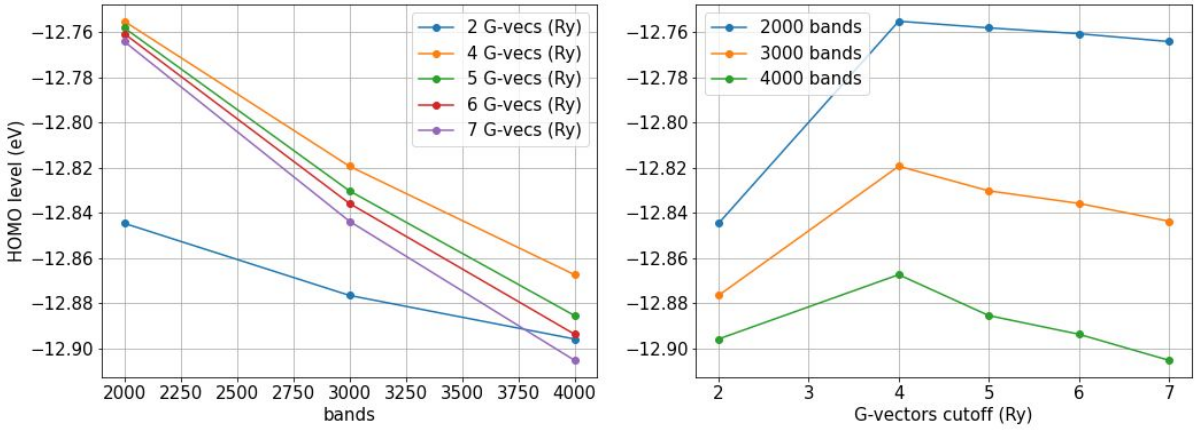
[3] M. J. van Setten, M. Giantomassi, X. Gonze, G.-M. Rignanese, and G. Hautier, "*Automation methodologies and large-scale validation for gw: Towards high-throughput gw calculations*", Phys. Rev. B **96**, 155207 (2017).

repository with data calculated using this flavour of PPA, currently missing. In fact, in the first GW100 study, only the HL-GPP as implemented in the BerkeleyGW code has been addressed.

- To verify the automatic convergence workflow for GW calculations that we have recently developed. As the computational resources increase, routine GW calculations on a large number of systems become possible. In this respect, having validated workflows automatically performing, say, tens of GW runs to extract converged quasiparticle energies, is of fundamental importance.

Importantly, tight convergence parameters need to be used in order to produce reference results for the absolute position of the HOMO and LUMO levels for the molecules in the GW100 set. In turn this requires a very detailed and accurate convergence study concerning the main parameters of the GW calculations (some of which are also interdependent). A few years ago, some algorithms were proposed in order to make such convergence procedures automatic and compatible with unsupervised high-throughput computing. Within the AiiDA framework we have further developed a convergence workflow for GW calculations and implemented it to be used with Yambo (https://github.com/yambo-code/yambo-aiida). In the present work we have used this automatic workflow to perform the GW convergence tests required and to precisely control the accuracy of the obtained results.

Moreover, the large number of GW calculations performed have been run on the recently deployed Marconi100 cluster at CINECA, a large scale GPU-accelerated machine, thereby also demonstrating high-throughput calculations driven by AiiDA and using Quantum ESPRESSO and Yambo on a GPU-empowered machine. More details about this demonstration aspect can be found in D6.1 .



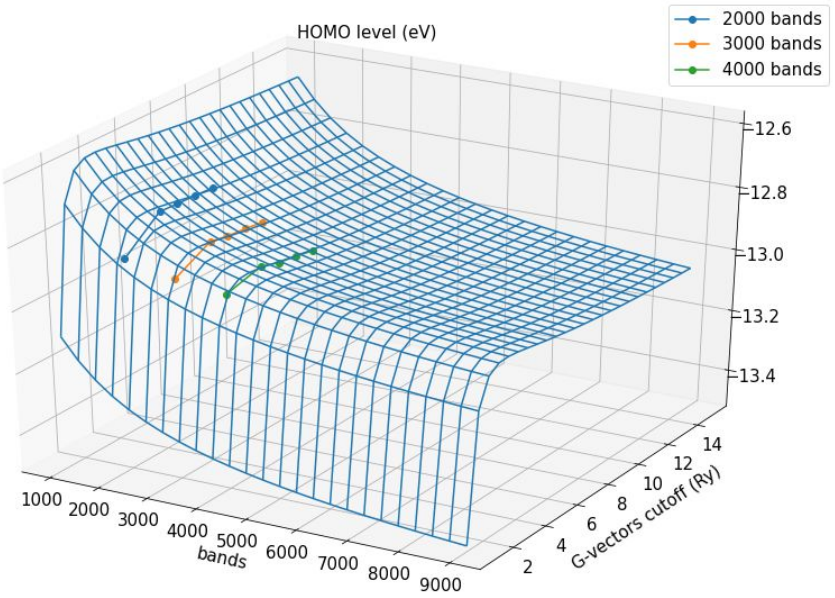


Figure 5: 1D and 2D convergence plots with respect to bands and kinetic energy cutoff used to represent the response function Xo, for the CH4 molecule in the GW100 set. For each molecule, the extrapolation of the GW results for the HOMO QP correction is performed on the basis of 15-20 GW calculations done using different convergence parameters, as shown in the graphs above. A lot of calculations are needed, due to the strong interdependence of the parameters involved.

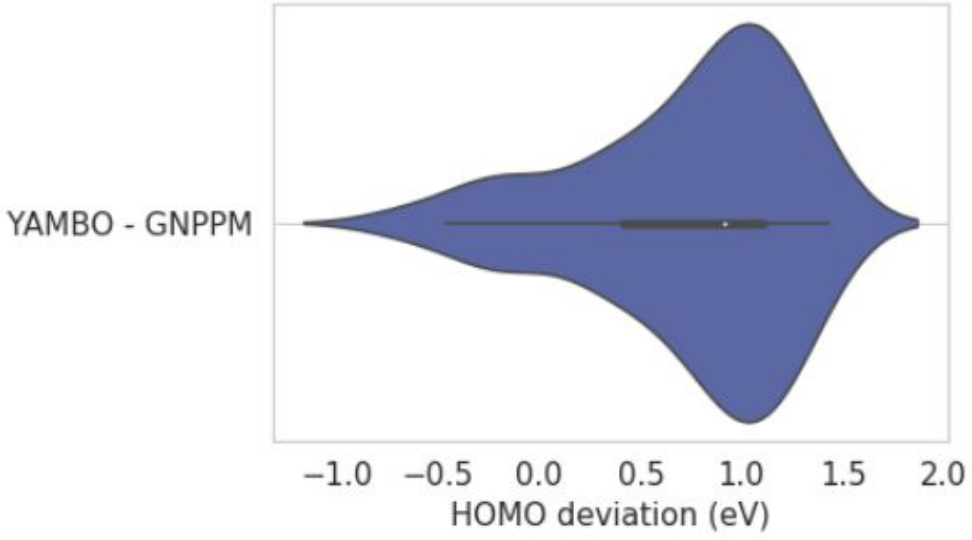


Figure 6: Violin plot reporting the error distribution for the GW100 data computed using the Godby-Needs plasmon-pole model (GN-PPM). Data are preliminary, meaning that further refinement of the procedures used for the extrapolation need to be performed.

For what concerns the production of reference GW100 data, upon extensive testing, we have adopted a rather large unit cell (40 x 40 x 40 $Bohr^3$), pseudopotentials from the pseudo-dojo portal[4]and the related cutoff energies, and performed several GW calculations by considering 2000, 3000, and 4000 bands for the sum-over-states and 2, 4, 5, 6, 7, 8(when possible) Ry for the kinetic energy cutoff used to represent the response function Xo. Given these sampling configurations, we have then performed an extrapolation following the indications of Rangel et al in Ref. [4] . For the sake of the discussion, prototype one dimensional plots and two dimensional fittings are shown in Fig. 5. Overall, the extrapolated results are collected in a "violin" plot and shown in Fig. 6 in order to display the error distribution obtained over the GW100 set. It is important to remark that this activity is still ongoing and the results of Fig. 6 are just preliminary, being still subject to more numerical checking, especially for what concerns the fitting procedures.

On the validation side (i.e. the agreement between physical theories and experimental results) the current assessment is that the mean average errors of different flavours of GW in predicting the HOMO of molecules is of the order of 0.2-0.4 eV.

3 High-performance data analytics pilots

A. Pilot 1: Predicting code performance

An accurate prediction of the time-to-solution required by massively parallel scientific codes would be extremely beneficial not only for scientists, that could better program and allocate their computational tasks but also from a HPC resource and, in turn, an environmental perspective, since resource waste due to suboptimal execution parameters could be easily detected. An important step in this direction has been obtained with machine learning techniques for DFT-based materials science codes. A recent work by Pittino et al[5], presented at PASC 19, shows how accurate predictions obtained with machine learning approaches can outperform parametrized analytical performance models made by domain experts.

The rise of heterogeneous HPC systems, where the standard central processing unit (CPU) is accompanied by one or more accelerators (eg GPUs), and the drastic increase in the number of cores per node requires the adoption of different algorithmic strategies for the same problem and inflates the number of options for parallel or accelerated execution of scientific codes. This, in turn, makes the (parallel) execution of such applications more complex and

[4] http://www.pseudo-dojo.org

[5] F. Pittino, P. Bonfà, A. Bartolini, F. Affinito, L. Benini, and C. Cavazzoni. Prediction of Time-to-Solution in Material Science Simulations Using Deep Learning. In Proceedings of the PASC 19 Conference, CH. ACM, New York, NY, USA. DOI: 10.1145/3324989.3325720

their performance harder to predict. Today, the primary effect from the user standpoint is that suboptimal execution schemes are often adopted since a complete exploration of the complicated and interdependent set of execution options is a lengthy and hardly automatable task. In order to tackle this problem, machine learning techniques have been used to predict the time required by the large set of algorithms utilized in a self-consistent field iteration of the Density Functional Theory method, using only input details and runtime options as descriptors.

In this work the authors have proven that out-of-the-box Machine Learning models can predict surprisingly accurately the time-to-solution of complex scientific applications, like Quantum ESPRESSO. In particular, they have revealed that Deep Learning algorithms, in this case a Fully Connected Neural Network, achieve the best performance, which corresponds to a relative error lower than 100% for 99% of the simulations, with a distribution peak and median at about 10% relative error. On the other hand they have also shown that a full-custom semi-analytical model specifically tailored to solve this task, whose few free parameters have been optimised on this dataset, exhibits a lower performance than that of the Neural Network.

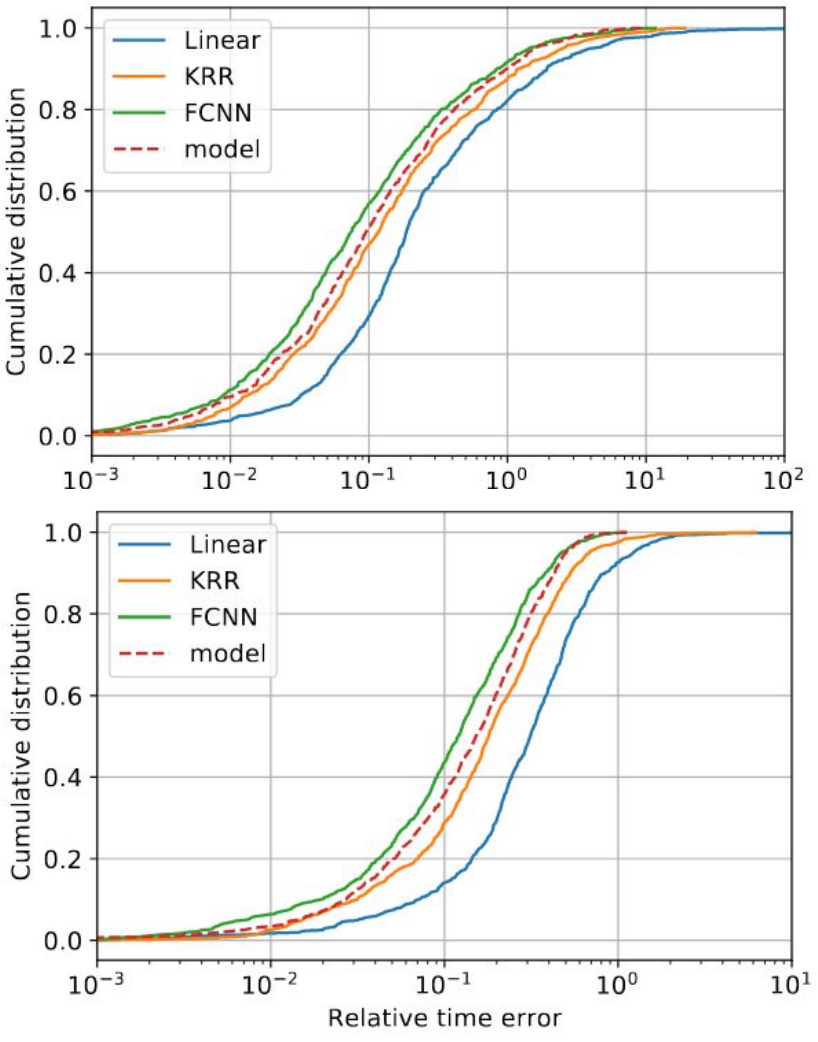


Figure 7. Distribution of the timing errors (absolute and relative) for three different algorithms - Linear Regression, Kernel Ridge Regression (KRR), Fully Connected Neural Network (FCNN) - compared to the analytical model.

It should be noted, however, that all models described in this paper have been trained using very similar versions of one scientific application, all in the same major release cycle. Once a new major version of the code is released, it is therefore highly probable that the models will need some retraining to retain their accuracy. The investigation on how to generalise the models to multiple codes and multiple versions of the same code,

together with a thorough cross-validation of the architecture and hyperparameters, will be the subject of our future work.

This work paves the way to the development of very accurate models for predicting in advance the properties of scientific appli- cations. It can then be extended to predict not only the time per iteration, but also the number of iterations or other properties of the application execution. It serves as a valuable tool for an accurate scheduling of the applications, but it can also be used to provide an a-posteriori evidence to the user of an issue on the execution when the predicted execution time is very different from the actual one.

B. Pilot 2: Configuration explorer/data explorer toolkit

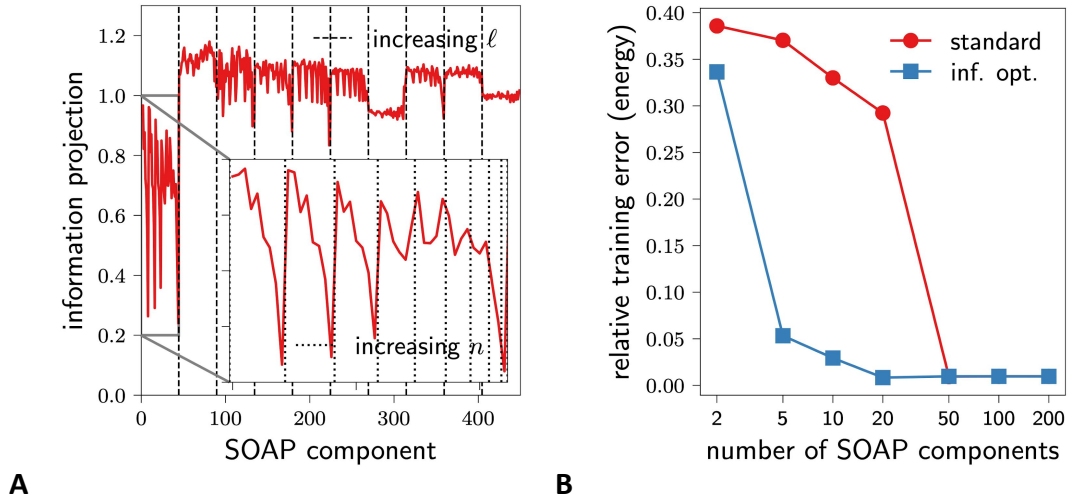


Figure 8: Application of the test to choose the most informative components of a SOAP vector. Panel a): Information projection of each SOAP component, the lowest the projection the most informative is the component. Panel b): training energy error achieved with a kernel regression using the first SOAP component (red curve) vs the most informative SOAP components (blue curve).

Virtually all applications of numerical algorithms in materials physics rely on the possibility of encoding the structure of the system studied into a vector of numbers, referred to as “representations”, “descriptors", or “fingerprints”. Different candidates have been proposed, with notable examples being the Atomic Symmetry Functions (ASFs) and the Smooth Overlap of Atomic Positions (SOAP). The choice of the optimal representation is system dependent, and at present this choice is often left entirely to trial and error and to physical and algorithmic intuition.

A simple statistical test was developed to assess the relative information content of different numerical representations of a given material dataset. The test allows recognizing if a metric (built using a subset of features) is more or less informative than another, and if a variable can be safely neglected. This test can be used, for example, to decide if a metric built using

ASMs is more or less informative than SOAP vectors. On several synthetic datasets the test proved to be capable of finding the most informative representation out of a set of candidates, also providing the right hierarchy of the information content of the representations.

The test developed can be of great use in the search of the optimal representation for applications in materials physics. Figure 6 provides a first example of this usage on a database of carbon structures at high pressure. Panel a) shows the relative information content of each component of a SOAP vector, (the lower the number on the y-axis the higher the information content). Interesting and entirely nontrivial patterns can be observed as a function of the SOAP truncation parameters *n* and *l*. The most informative SOAP components can then be used for a variety of numerical applications. For instance, panel b) shows how they can provide compact representations useful for machine learning of structure energies. The results obtained will be published in the near future, and the code will be made publicly available through GitHub, as is custom for all our deliverables.

C. Pilot 2: Chemiscope - interactive exploration of large datasets

We developed a first version of an online interactive visualization and exploration library for HPDA called chemiscope, of which a demonstration version is available at https://chemiscope.org. It is a graphical tool for the interactive exploration of materials and molecular databases, correlating local and global structural descriptors with the physical properties of the different systems. The default interface is composed of two panels (Figure 9). The left panel consists of a 2D or 3D scatter plot, in which each point corresponds to a chemical entity. The axes, color, and style of each point can be set to represent a property or a structural descriptor to visualize structure-property relations directly. The right panel displays the three-dimensional structure of the chemical entities, possibly including periodic repetition for crystals. Visualizing the chemical structure can help finding an intuitive rationalization of the layout of the dataset and the structure-property relations.

Chemiscope is built with a modular design, and does not compute directly the structural descriptors. These can be obtained from one of the many codes implementing such descriptors such as librascal (https://github.com/cosmo-epfl/librascal) or QUIP (https://github.com/libAtoms/QUIP) for example. Since the most common descriptors can be very high dimensional, it can be convenient to apply a dimensionality reduction algorithm that maps them to a lower-dimensional space for easier visualization. The resulting point representing individual structures or atomic environments can be visualized in 2D or 3D spaces. For example the SketchMap[6] algorithm was used with the Smooth Overlap of Atomic

[6] M. Ceriotti, G. A. Tribello, and M.Parrinello, *Simplifying the representation of complex free-energy landscapes using sketch-map*, PNAS 108, 13023 (2011). DOI: 10.1073/pnas.1108486108

Positions descriptor[7] to generate the visualization in Figure 9 below. Integration with the efforts described in the previous section shall provide a complete end-to-end solution for data analytics and visualization.

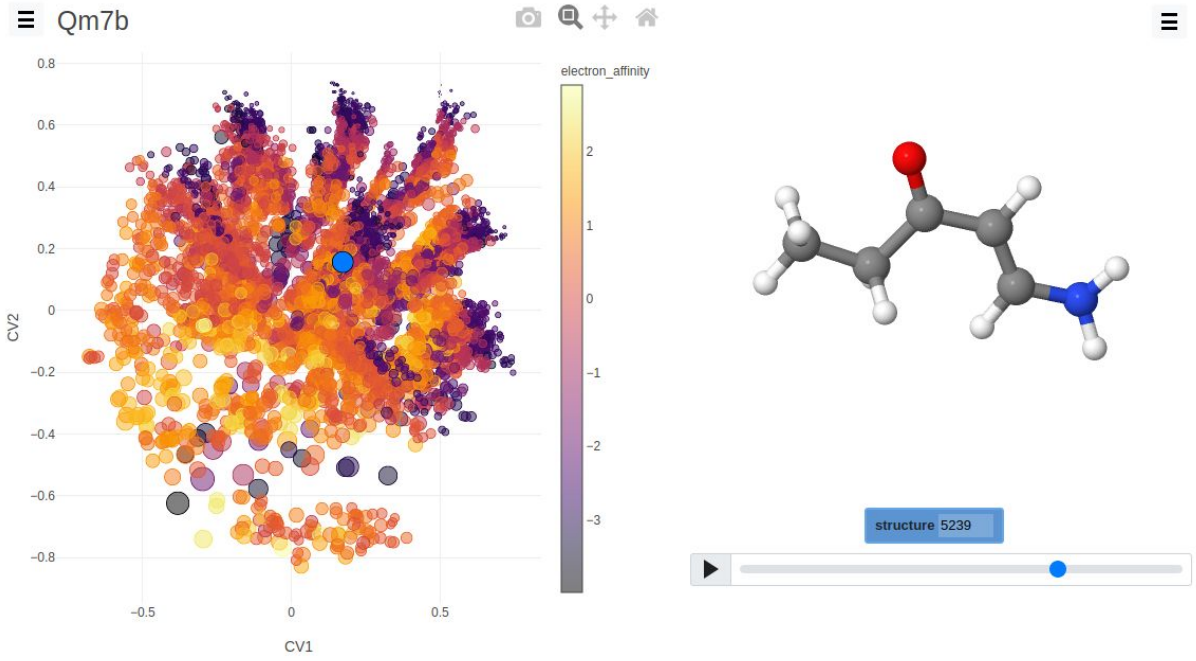


Figure 9: Visualisation of a dataset of small organic molecules using chemiscope.

The library is implemented using web technologies (WebGL, JavaScript/TypeScript, HTML, CSS) and runs inside web browsers. The use of web technologies makes chemiscope usable from different operating systems without the need to develop, maintain and package the code for each operating system. It also means that we can provide an online service allowing users to visualize their own dataset without installing anything. Chemiscope is implemented as a library of reusable components linked together via callbacks. This makes it easy to modify the default interface to generate more elaborate visualizations: for example displaying multiple maps generated with different parameters of a dimensionality reduction algorithm. The code is available on Github (https://github.com/cosmo-epfl/chemiscope) under the permissive BSD license, while depending on the LGPL-licenced Jmol for displaying atomic structures. Both user and developer documentation are available online at https://chemiscope.org/docs/.

[7] A.P. Bartók, R. Kondor, and G. Csányi, *On representing chemical environments*, Phys. Rev. B 87, 184115 (2013). DOI: 10.1103/PhysRevB.87.184115

We are currently integrating chemiscope within the Materials Cloud, to allow direct visualization and analysis of curated datasets, and provide a complete HPDA solution for the different high-throughput screening efforts within and outside the CoE.

D. Pilot 2: High-performance data analytics

When dealing with very large datasets, two factors are limiting our ability to do interactive data analysis: the initial time to load and display the dataset to the user, and the time between user input and when changes are displayed on screen. Concerning the first point, load times around 10 to 30 seconds are usually acceptable. For interactive exploration to feel natural, the system should have a response to user input in less than 100ms ideally, not more than 500ms.

Chemiscope uses Plotly.js (https://plot.ly/javascript/) to render and animate 2D and 3D plots using WebGL. This allows it to render datasets using GPU (Graphical Processing Units) to compute the position and color of pixels on screen quickly. In 2D mode, with 100 000 points, updates to the displayed map (zoom, translation, rotations) are under 10ms; and with 1 000 000 points it goes up to 100ms. The initial rendering time is around a few seconds for a million points. This loading delay occurs each time a visualization setting changes (color, size, or shape of points). Initial loading time is kept low by only loading the minimal amount of data on the first page load, and then dynamically fetching additional data as needed. In particular, the structures can take a lot of bandwidth and thus time to load, so chemiscope offers an option to developers deploying it to load them on-demand, when the user requests a new structure to be displayed. After a structure has been loaded, it is cached for faster access if it is needed again later. Using these strategies, Chemiscope is able to display, update, and interact with large datasets smoothly and efficiently.

E. Pilot 3: Dissemination of highly-curated computational materials data

We have imported all the inorganic materials contained in the experimental databases ICSD, COD, and Pauling file, obtaining around 84,000 inorganic stoichiometric compounds (see Fig. 10; 77,000 with unit cells smaller than 100 atoms), for which we are now deploying our refined automated workflows to calculated with tight protocols all the fundamental properties (see Fig. 11 for an example).

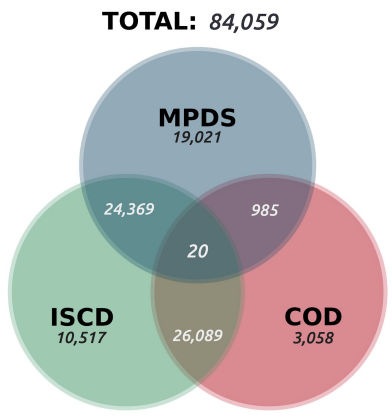


Figure 10: the unique stoichiometric inorganic compounds contained in the 3 major experimental databases of crystal structures (ICSD, COD, and Pauling File/MPDS).

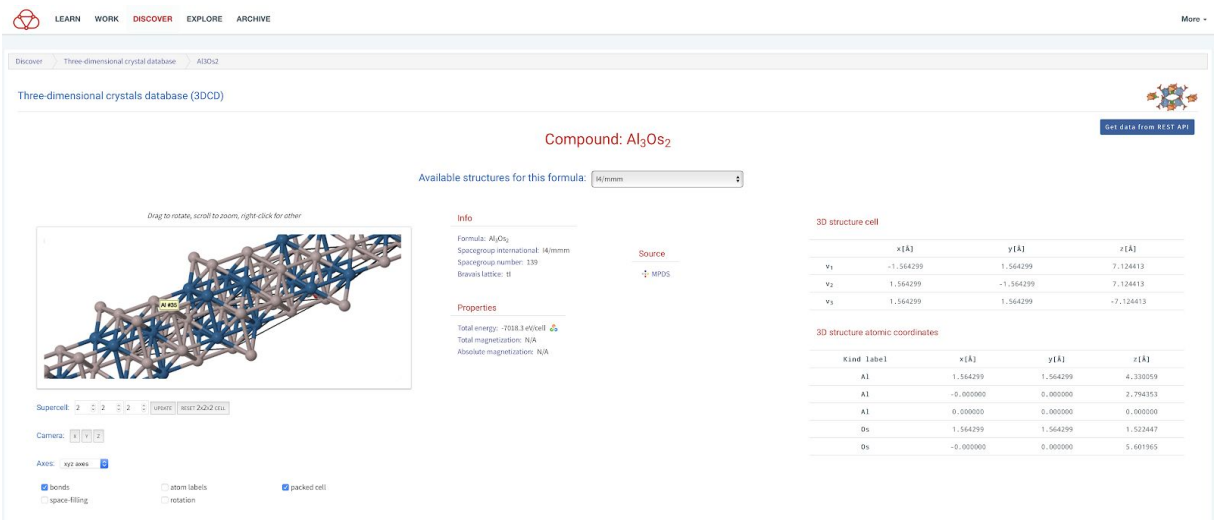


Figure 11: dissemination of computational, curated properties for the reference database of inorganic structures identified above (Figure 10).

F. Pilot 4: Edge computing

The last pilot will take part in the second half of the MaX project timeline.