

BioExcel-2 Project Number 823830

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

WP3:
*Use Case Implementation &
In-Depth Support*



Copyright© 2019-2022 The partners of the BioExcel Consortium



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

Document Information

Deliverable Number	D3.6
Deliverable Name	Pre-Exascale showcase calculation and Use Case Final Report
Due Date	2022-02-28 (PM38)
Deliverable Lead	UEDIN
Authors	Arno Proeme (UEDIN) Alessandra Villa (KTH) Alexandre Bonvin (UU) Vytautas Gapsys (MPG) Rodrigo Vargas Honorato (UU) Adam Hospital (IRB) Esther Sala (NBD) Dmitry Morozov (JYU) Gerrit Groenhof (JYU) Mirko Paulikat (FZJ) Paolo Carloni (FZJ)
Keywords	
WP	WP3
Nature	Report
Dissemination Level	Public
Final Version Date	2022-02-25
Reviewed by	Vera Matser (EMBL-EBI) Ian Harrow (IHC) Stian Soiland-Reyes (UMAN) Aristarc Suriñach (NBD) Steve Robertshaw (AL)
MGT Board Approval	2022-02-28

Document History

Partner	Date	Comments	Version
UEDIN	2021-12-12	First draft (structure)	0.1
UU	2022-01-14	Added UC2	0.2
KTH, UU, MPG, IRB, NBD, JYU, FZJ	2022-02-07	First Draft	0.3
UU, FZJ, MPG, UEDIN	2022-02-25	Final Draft	0.4

Executive Summary

This deliverable is an update to the project half-time deliverable [D3.3 - Use Case Progress Report](#). It provides a final report on BioExcel-2's demonstrator research projects (the "Use Cases"), including a pre-exascale showcase calculation performed during the second half of the project. Final reports for each Use Case include a summary of work done, scientific results obtained, software and HPC resources used, and how the work has impacted the community. We demonstrate that the Use Cases serve as exemplars of how important challenges in biomolecular modelling and simulation can be tackled effectively using HPC resources and various combinations of the software developed and supported by the CoE. The Use Cases are at the core of demonstrating the impact of BioExcel's science-enabling software development to the benefit of the biomolecular modelling and simulation community. As well as providing valuable scientific results in their own right, they are a rich source of expertise and best practice methodologies which are drawn upon by the Centre to provide support and training to the community and to engage with industry.

The **pre-exascale showcase calculation** (Use Case 5 - UC5), demonstrates the **readiness** of BioExcel applications GROMACS and PMX to enable **state-of-the-art methods to obtain results of equivalent accuracy to leading commercial software**, and to do so with unprecedentedly fast turnaround time by using **pre-exascale HPC resources**. The simulations involved were performed over a period of **72 hours** using **480 nodes / 50k cores** - **93%** of a **3.5 Pflop/s** (peak) supercomputer ("[Raven](#)" hosted at MPCDF). This demonstrates the potential for large-scale rapid-turnaround virtual drug screening that would currently take weeks for an individual researcher to perform based on available compute time but which could become routine in the exascale era, with significant potential to massively accelerate drug discovery.

Use Cases 1 and 3 have demonstrated that computational pipeline protocols employing various combinations of the core BioExcel applications GROMACS, HADDOCK and PMX can be used for mutational analysis and to perform free energy and docking calculations for the design of antibody-based and small ligand therapeutics. The BioExcel-developed Building Blocks (BioBB) library for interoperable biomolecular simulation workflows in conjunction with the PyCOMPSs task-based execution manager has enabled many of these pipelines to execute efficiently on HPC resources, demonstrating how the ultimate goal of shortening the time and effort to develop new and better therapeutics may be achieved.

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

Use Case 2 has shown how parallel execution of containerised HADDOCK and the necessary data movement underpinning can be orchestrated by PyCOMPSs, demonstrating how (pre-)exascale HPC resources will be able to be used to study significant fractions of the interactions of biomolecules encoded by the human genome.

Use Cases 4a and 4b have demonstrated how QM/MM simulation with GROMACS, CP2K and CPMD using HPC resources to perform costly quantum calculations are set to improve electrospray ionization mass spectrometry - a key analytical technique in proteomics - and, combined with free energy calculations using PMX, are enabling the design of fluorescent proteins enabling the high-resolution monitoring of cellular functions, gene expression, protein-protein interactions, intra-cellular interactions in living systems as well as understanding and finding novel strategies to tackle disease.

Contents

Introduction	8
Use Case 1: Antibody Design	11
Final Report	12
Software	16
Exploitation of HPC resources	17
Impact	18
Use Case 2:	
High-throughput Modelling of Interactomes	21
Final Report	21
Software	21
Exploitation of HPC resources	22
Impact	24
Use Case 3: Rational Drug Design	25
Final Report	25
WF1: Moving mutational analysis into the structural field for drug design	26
WF2: Large-scale SARS-CoV2 mutation analysis using BioExcel HPC workflows	27
WF3: Quantitative predictions of binding affinity in lead optimization	28
WF4: Machine learning for efficient drug design	29
Software	30
Exploitation of HPC resources	31
Impact	32
Use Case 4a: Fluorescent Proteins	34
Final Report	34
Computational screening of Aequorea victoria GFP mutants	35
QM/MM MD simulations of photo-active proteins	35
Software	36
Exploitation of HPC resources	36
Impact	37
Use Case 4b: Proton Dynamics	40
Final Report	41
Classical MD Simulation of DNA in solutions and in water droplets	41
Proton Transfer in Model Systems by MiMiC QM/MM simulations	41
Ab initio Molecular Dynamics Simulations	42
Local Proton Transfer Profiles upon Desolvation by GROMACS/CP2K QM/MM simulation	43
Software	44

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

Exploitation of HPC resources	44
Impact	45
References	45
Use Case 5 (pre-exascale showcase calculation):	
High-throughput Drug Screening	47
Final Report	47
Setup	47
Results: calculation accuracy	49
References	51
Software	52
Exploitation of HPC resources	52
Impact	53
Publications	53
Other	53
Presentations	53
Appendix: Final status with reference to work plans	53
Use Case 1	54
Use Case 2	58
Use Case 3	59
Use Case 4a	60
Use Case 4b	61

Introduction

BioExcel's development of its core applications and the provision of associated support and training are driven by the ultimate goal of advancing the computational biomolecular research community's ability to use HPC to address grand challenges. To achieve this, BioExcel has undertaken a number of demonstrator research projects - so-called "Use Cases" - in which the software developed by BioExcel was used to solve concrete challenging problems at the forefront of scientific interest in both academic and industrial biomolecular research. These Use Cases have exercised and driven the development of BioExcel software and workflows and helped build expertise enabling BioExcel staff to effectively support and train the community in their usage. Results obtained and methodologies developed in the Use Cases have had impact in the community by serving as exemplars of what is possible using BioExcel-developed solutions, and they showcase how large-scale computing resources can be used effectively to solve life science challenges.

The BioExcel-2 Use Cases have served not only as exemplars of novel computational biomolecular research done using HPC, but also as targets to first ensure and then demonstrate that BioExcel-2 software development activities have impacted concrete application usage rather than merely theoretical performance or scaling. They have had an additional role as an originating mechanism for developing, documenting and disseminating best-practice examples of how to scale difficult problems to make effective use of increased parallelism, including through changes in methodology, such as advanced ensemble-based sampling simulations or high-throughput strategies, pointing the way at enabling effective use of (pre-)exascale computing resources. The Use Cases have thereby helped BioExcel address the needs of end users, by providing extensible templates for future research as well as good source material for training in modern effective computational biomolecular research done using HPC.

Tackling the Use Cases has also helped establish usage of new well-documented modular workflows composed of building blocks developed by BioExcel (BioBB) that interoperate between other BioExcel applications (GROMACS, PMX, HADDOCK, CP2K) as well as many others. These workflows have been made available as container images, aiding distribution, deployment and uptake, and the Use Case work has proven they can be adapted and executed efficiently on HPC resources using the PyCOMPSs execution framework. As well as benefiting existing academic and industrial users, this lowers barriers and facilitates exploitation of (pre-)exascale machines for biomolecular research by SMEs.

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

This deliverable is an update to the BioExcel-2 project half-time deliverable [D3.3 - Use Case Progress Report](#). It provides a final report on the progress made in the Use Cases described in D3.3.

Table 1: Overview of the Use Cases, including relevant BioExcel software employed and the BioExcel-2 partners primarily involved in the execution of each Use Case

	Topic	Software	Partners
UC1	Antibody Design	GROMACS HADDOCK PMX	KTH UU MPG
UC2	High-throughput Modelling of Interactomes	HADDOCK PyCOMPSs	UU BSC
UC3	Rational Drug Design	BioBB PyCOMPSs GROMACS PMX	IRB NBD BSC
UC4a	Fluorescent Proteins	GROMACS CP2K PMX	JYU
UC4b	Proton Dynamics	GROMACS CP2K MiMiC	FZJ
UC5	High-throughput Drug Screening (pre-exascale showcase calculation)	GROMACS PMX	MPG

Use Case 5 (UC5), briefly described as newly planned in D3.3, constitutes the BioExcel-2 **pre-exascale showcase calculation**. It demonstrates how BioExcel applications GROMACS and PMX can be used to screen hundreds of ligand modifications to benchmark the accuracy of computation against an experimental reference. We showed the **readiness of BioExcel software to enable usage of state-of-the-art methods to obtain results of equivalent accuracy to leading commercial software**, and to do so with unprecedentedly fast turnaround time by **using pre-exascale HPC resources**. The simulations involved were performed over a period of **72 hours using 480 nodes / 50k cores - 93%** of Max Planck Computing & Data Facility's [Raven supercomputer](#), capable of **3.5 Pflop/s** peak performance at 100% utilisation. This demonstrates the potential for large-scale rapid-turnaround virtual drug screening that would currently take weeks for an

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

individual researcher to perform based on available compute time but which could become routine in the exascale era, with significant potential to massively accelerate drug discovery.

Some of the COVID-19-related research work described in D3.3, including Use Case 6 (UC6), has been reported separately in a dedicated deliverable [D6.3 HPC Covid-19 Research Use Case](#). COVID-19 work related to the originally planned Workflow 2 (WF2) in UC3 however is described in this document.

For each Use Case this deliverable reminds readers of the background and rationale for the work, where relevant, and summarises the work done and scientific results obtained in the form of a final report. We describe the usage of BioExcel software, mentioning where relevant how the Use Case work was enabled by improvements to functionality and/or performance developed within BioExcel. We highlight the usage of HPC resources that were critical to enabling the work to be performed including, where possible, the hardware, numbers of cores or nodes used in individual jobs, and order-of-magnitude estimates of the total compute time used. Taken together, these details illustrate how each Use Case demonstrates the current and future enabling value to computational biomolecular research of large-scale HPC resources, like the (pre-)exascale machines becoming available in the EU.

We describe ways in which the Use Case work has had an impact on the community, both directly through dissemination of the results and/or methodologies via publications and presentations, and indirectly by feeding into BioExcel activities such as training and support and the production of tutorials and best practice guidance. Finally, the Appendix in this document provides a final status update for each Use Case on the level of all of the individual activities that made up the work plans originally described in [D3.1 - Use Case Work Plans](#).

Use Case 1: Antibody Design

Many diseases still lack treatment with conventional drugs. In recent years, antibodies have been developed as immunotherapeutics in the fight against diseases, due to their lower side effects compared with traditional, small-ligand-based drugs. An antibody drug is a large protein complex or semi-synthetic peptide that generally works by attaching itself to an antigen, which is a unique site of the pathogen and harnessing the immune system to directly attack and destroy it. Currently approved treatments with antibodies target many diseases including rheumatoid arthritis, multiple sclerosis and some types of cancer. Additionally, antibodies play an important role in disease or condition detection and diagnostics. For example, tests making use of such antibody-antigen reactions are used to detect pregnancy, prostate cancer, HIV or SARS-CoV-2 in the human body.

The therapeutic potential of antibodies lies in their high specificity and low immunogenicity. This is achieved by two unique structure encompassing regions; *fragment crystallizable region* (Fc region) and the *antigen-binding region* (Fab region). Fc region activates the immune response and is species specific, i.e. human Fc region should not evoke an immune response in humans (low immunogenicity). The Fab region needs to be highly variable to be able to bind to antigens of various nature (high specificity). More precisely, there are six loops, known as *complementarity-determining regions* (CDRs) or hypervariable loops that alter their sequence to complement different antigens. Moreover CDRs are greatly flexible too, as shown in Figure 1, where we see the hypervariable loops of the unbound state (U) in magenta and those of the bound state (B) in blue.

Antibody development is an expensive and labour-intensive process. Within BioExcel, we use computational tools to study the interactions between antibodies and antigens. Here, key challenges are to:

- Reliably predict 3D structures of antibodies, in particular the complementarity-determining regions (CDRs)
- Model their binding mode and understand how structure and dynamics are altered in this process
- Improve their binding affinity through systematic amino acid mutations

These challenges were addressed through a combination of the core BioExcel applications GROMACS (for MD simulations), HADDOCK (for docking simulation) and PMX (for free energy calculation) and synergy between three BioExcel partners (KTH, UU, MPG).

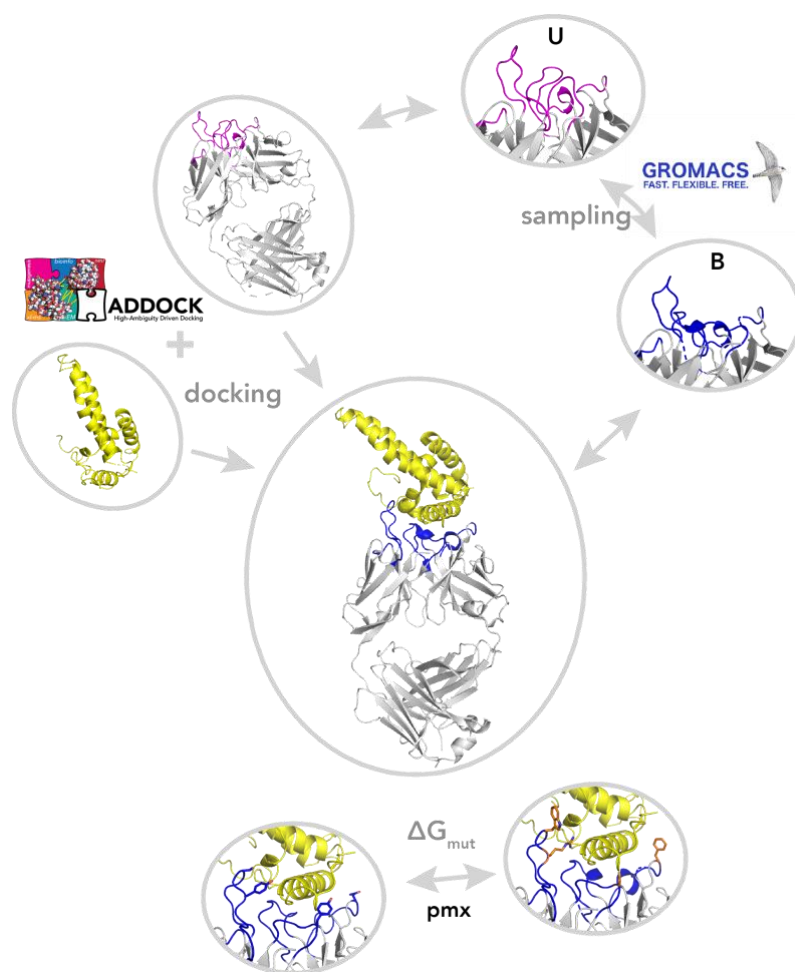


Figure 1: Three elements addressed by software packages in UC1. The prediction of the antibody-antigen structure is done by molecular docking using HADDOCK (UU), molecular sampling of the interface by GROMACS (KTH) and the free-energy calculations via pmx (MPG). This figure reports the antibody-antigen complex, PDB code: [3V6Z](#), where antibody is in grey, antigen in yellow, CDRs of the unbound antibody (U) in magenta and CDRs in the bound complex (B) in blue.

Final Report

Over the course of the Use Case work through BioExcel-2 we tested a number of different possible pipelines in order to identify the best strategy to effectively combine docking and conformation sampling for the design of *ex-novo* antibody drugs. In total 18 *antibody-antigen* systems (Ab-Ag) were selected from [Docking Benchmark 5](#) and [Affinity Benchmark 2](#) to test and validate approaches combining HADDOCK, GROMACS, and PMX. The SARS-CoV-2 spike protein-antibody complex was also added to this set ([PDB ID 6W41](#), [Yuan et al., 2020](#)).

Below is a brief summary of the tested pipelines:

- HADDOCK antibody-antigen models, generated from crystallographic structures, were refined using standard MD simulation and/or enhanced sampling (reported and discussed in [D3.3 - Use Case Progress Report](#))
- HADDOCK antibody-antigen models were generated using MD-sampled conformations for antibody and antigen pairs (reported and discussed in [D3.3 - Use Case Progress Report](#))
- HADDOCK antibody-antigen models were generated using MD-sampled conformations for antibody (plus crystallographic of the apo form) and antigen experimental structure (Figure 2)

A summary of the progress and final status of work on this Use Case with reference to the original work plans described in [D3.1 - Use Case Work Plans](#) can be found in [Table 3](#) in the Appendix.

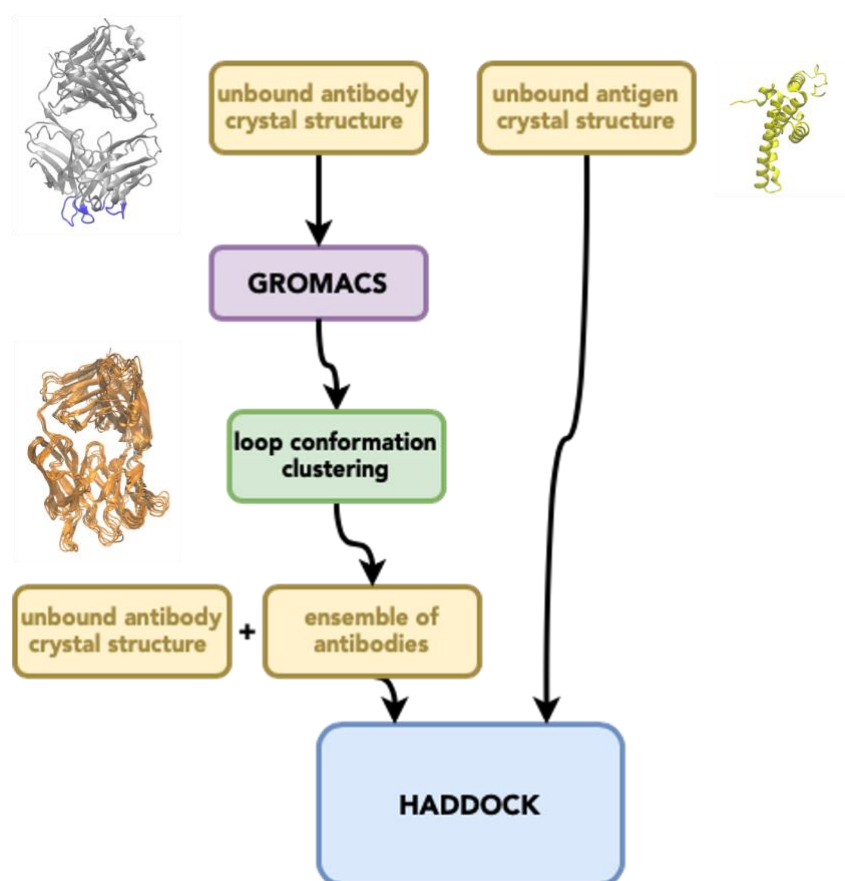


Figure 2: Workflow of antibody-antigen ensemble docking using conformations, sampled by MD and selected based on clustering of the hypervariable loops. Antibody crystal structure in grey, antigen crystal structure in yellow, and MD-sampled structures from the 20 higher populated clusters in orange.

In the last pipeline (Figure 2), antibody-antigen docking models were generated using a pool of conformations for the antibody (MD-sampled plus crystallographic structures) and the crystallography structure for antigen. The crystallographic structures of the apo form were used as input, while the holo form was used as reference for the validation of the pipeline. The antibody protein conformational space was sampled using standard MD simulations. To select the more relevant MD conformations, we clustered the structure based on the CDRs conformation and considered the 20 most populated clusters.

To assess the quality of docking models we used *DockQ* ([Basu et al., 2016](#)), which calculates a score based on the fraction of native contacts (Fnat), the ligand root mean squared deviation (LRMSD) and the interface root mean squared deviation (iRMSD). These are standard quality measures used in the Critical Assessment of PRedicted Interactions (CAPRI) community evaluation of protein-protein docking for structure prediction. DockQ scores range between 0 and 1, and classify interfaces into high (1-0.8), medium (0.8-0.49) or acceptable (0.49-0.23) quality models or incorrect ones (0.23-0). General practice is to use the available experimental structure as reference to assess the model quality. This required that we had to restrict the set of complexes used for the validation to 11 antibody-antigen complexes (for which both the apo and the holo structure are available, see for more details point A1.10 in [Appendix](#)).

Figure 3 shows the distribution of the DockQ vs HADDOCK score. The final pipeline (Figure 2) provides models of high quality and a large number of medium quality models compared with the HADDOCK original approach. Also using the final pipeline a linear trend is observed between the two docking scores: high HADDOCK score corresponds to a low DockQ value and low HADDOCK score to a high DockQ. This is the trend that is expected between the two scores.

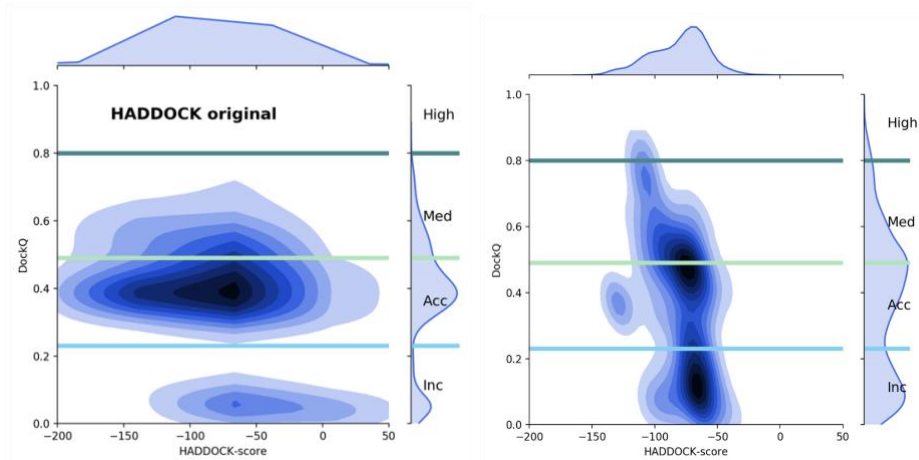


Figure 3: Distribution of model quality in DockQ vs HADDOCK score for 11 antibody-antigen complexes. On the left (labelled as original HADDOCK) values for docking models generated by HADDOCK from apo antibody and antigen pairs without prior

MD. On the right, values for docking models generated using the pipeline described in Figure 2.

Among the 11 complexes, we had some cases where no models were observed having a DockQ score higher than 0.6. We selected one of the complexes (PBDID: [3V6Z](#), shown in Figure 1) and developed a variant of the pipeline in Figure 2, where we use accelerated weight histogram (AWH) method to generate the ensemble of antibody structures. The AWH method was implemented in GROMACS during BioExcel-1, as described in deliverable [D1.5 - Final project release of pilot applications](#). In particular, the best HADDOCK model is selected and simulated. The approach was used to enhance the sampling of the antibody interface in presence of the antigen. The aim was to improve the pipeline by starting from MD-sampled structures with potentially improved CDR conformations thanks to the sampling performed in presence of the antigen instead of the sampling of an isolated antibody in solution. Indeed, the antibody-antigen models obtained using AWH-sampled structures in the pipeline showed a clear improvement in quality (Figure 4). In principle this approach can be applied to any antigen-antibody model, though it should first be validated on a large set of complexes.

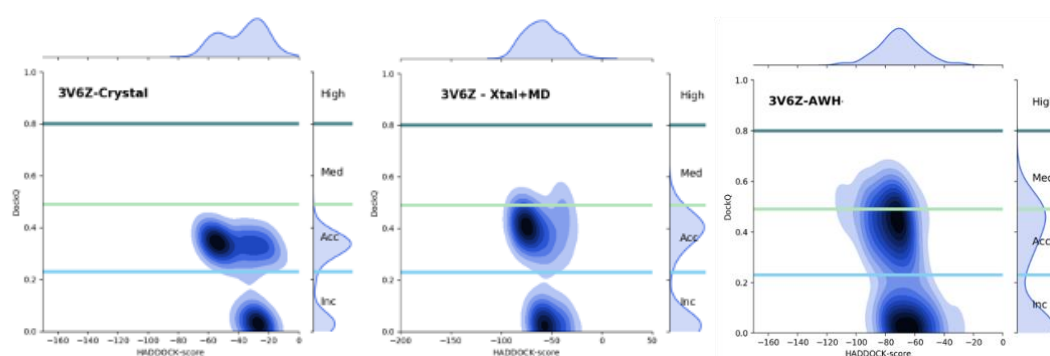


Figure 4: Distribution of model quality in DockQ vs HADDOCK score for 3V6Z antibody-antigen complex. On the left values for docking models generated by HADDOCK from apo antibody and antigen pairs without prior MD. In the middle, values for docking models generated using MD-sampled structures. On the right, values for docking models obtained using AWH-sampled structures.

Subsequently, a set of 4 antibody-protein complexes with a number of experimentally studied mutations were selected for further investigation. The PMX workflow relying on a rapid Rosetta-based screening protocol ([Rosetta FlexDDG](#), implemented in the scope of the [D6.3 HPC Covid-19 Research Use Case](#)) was applied to evaluate protein-protein binding free energy differences upon amino acid mutation. Comparison of the prediction accuracy for calculations started from the crystallographic structures to the accuracy obtained from those started with HADDOCK models would provide valuable information on the model

Through this project, our ultimate goal was to shorten the overall time and effort taken to develop new antibody-based therapeutics for patients. The pipeline in Figure 2 contributes to the early stages of achieving this goal.

A protocol for antibody-antigen modelling using HADDOCK version 2.4 with different levels of information was developed and published (Ambrosetti et al., [2020a](#) and [2020b](#)), together with an [online tutorial](#). We further developed a deep learning model implemented in a new [web server](#) to predict the antibody regions involved in the interaction ([Ambrosetti et al., 2020c](#)). To facilitate the use of this developed protocol through the HADDOCK2.4 web server we implemented an

option to automatically select all solvent accessible residues in the case when no information is available on the binding site of the antigen. Further, contributions from external users have improved the online tutorial by providing a machinery to automate the processing of antibodies and preparation of docking runs available from the public GitHub repository [HADDOCK-antibody-antigen](#).

GROMACS version 2019, 2020 as well as 2021 were used to perform molecular dynamics simulation. Use Case 1 takes advantage of recent GROMACS features and improvements developed within BioExcel. These include the AWH method, performance improvement in parallel runs (update groups), and SIMD acceleration of certain dihedral potentials, PME solve and update. For details see BioExcel-1 deliverable [D1.5 - Final project release of pilot applications](#) - and BioExcel-2 deliverable [D1.4 - Project release of core applications](#).

A free energy calculation setup based on the Rosetta FlexDDG protocol was implemented in PMX for an efficient rapid scan of mutation effects on protein-protein interactions. The protocol can be further combined in a multi-level screening workflow, as illustrated in deliverable [D6.3 - High Performance Computing in support of COVID-19 Research](#).

Exploitation of HPC resources

The MD-docking pipelines developed require microseconds worth of MD simulation of antibody systems in both apo and bound forms - if using the AWH approach - with system sizes ranging from 30,000 atoms for ligands to over 200,000 for complexes. In future these pipeline will be used in scenarios where some experimental data are available, but not the complex structure. We will have to account for n -fold mutations of the antibody CDRs and m -fold different binding poses, requiring $n+1$ simulations of the apo form and $n*m$ simulations of the bound form.

Currently, each MD simulation can be performed on 2 nodes (64 cores) Intel Xeon E5-2698 at 2.30 GHz with a performance of around 14 ns/day for systems with 197,000 atoms. If the sampled apo-structures are to be used for docking, a large number of parallel docking runs need to be launched from different antibody simulation snapshots. This can be done very effectively on HPC resources using the PyCOMPSs/HADDOCK Singularity container-based workflow developed under WP2 (see deliverable [D2.3 – First release of demonstration workflows](#)), which is also used in Use Case 2 described in this document. In this approach each docking run is assigned to a full node, and the entire set of dockings is managed by PYCOMPSs, which also takes care of necessary data movement, effectively appearing as one large HPC job. This demonstrates how future larger-scale HPC

resources could be exploited effectively to advance research through using the approach and tools developed within BioExcel.

While MD free energy screens of mutational effects are computationally particularly demanding, the Rosetta FlexDDG protocol offers a computationally less expensive solution. For the current use case, in total we screened 1342 mutations, performing calculations on a heterogeneous in-house cluster. Each calculation was performed on an 8-core CPU node and required at most 12 hours to finish. This amounted to **~130,000 core hours** for a large mutation screen, offering high throughput screening predictions which can then be brought to the next stage of computationally more expensive MD based calculations.

Impact

Use Case 1 work has been of support for development best practice guides for HADDOCK and GROMACS. Moreover a specific [antibody-antigen best practice](#) has already been developed for HADDOCK. The protocol identifying hypervariable loops of the antibodies was submitted and is already available to users on a [web server](#).

Use Case 1 work was presented at more than 15 conferences and workshops. It featured as an invited talk (*“Molecular Dynamics meets Docking: A use case on antibody design”*) at the course “High Performance Molecular Dynamics” held by CINECA in April 2021 (60 participants). It was also presented at the 2020 BioExcel Winter School (*“[When HADDOCK meets GROMACS](#)”*) and 2021 BioExcel Summer School (*“[MD meets Docking: An use case on antibody design](#)”*), attended by 60 participants in total. The recorded presentations of BioExcel Schools are available on the BioExcel YouTube channel and currently have more than 500 views.

The positive feedback and interest of the community in combining the two techniques and the two software packages led us to import the pipeline into the format of a Jupyter notebook, which can in future be used in academic and industrial training. [This notebook](#) (which is not yet finalised and still under active development) will make the developed workflow more accessible to interested users and make it easier to apply the pipeline in the study of other complexes. The approach we have developed and made available should aid the search for rapid treatments and reliable diagnostic tools. The importance of this is clear in these times of a global fight against COVID-19, and also given the need to improve preparedness for potential future pandemics.

Finally the work on Use Case 1 has already contributed to the development of online tutorial material both for HADDOCK ([HADDOCK-antibody-antigen tutorial](#)

and [HADDOCK2.4 Antibody - Antigen tutorial](#)) and for GROMACS ([the GROMACS tutorials!](#)). Furthermore a tutorial for the pmx-based rapid mutational screens employing the Rosetta FlexDDG protocol is under development.

Success Stories and publications:

- F. Ambrosetti, Z. Jandova and A.M.J.J. Bonvin. [A protocol for information-driven antibody-antigen modelling with the HADDOCK2.4 webserver](#). *ArXiv*, 2005.03283 (2020).
- F. Ambrosetti, T.H. Olsed, P.P. Olimpieri, B. Jiménez-García, E. Milanetti, P. Marcatilli and A.M.J.J. Bonvin. [proABC-2: PRediction Of AntiBody Contacts v2 and its application to information-driven docking](#). *Bioinformatics*, 36, 5107–5108 (2020). <https://doi.org/10.1093/bioinformatics/btaa644>
- F. Ambrosetti, B. Jiménez-García, J. Roel-Touris and A.M.J.J. Bonvin. [Modeling Antibody-Antigen Complexes by Information-Driven Docking](#). *Structure*, 28, 119-129 (2020). <https://doi.org/10.1016/j.str.2019.10.011> ([Preprint](#) available).
- R.A. Norman, F. Ambrosetti, A.M.J.J. Bonvin, L.J. Colwell, S. Kelm, S. Kumar and K. Krawczyk. [Computational approaches to therapeutic antibody design: established methods and emerging trends](#). *Briefings in Bioinformatics*, bbz095 (2019). <https://doi.org/10.1093/bib/bbz095>

Conferences, workshops and courses:

- Alexandre Bonvin - lecture - “Integrative modeling of biomolecular complexes”. GIDRM meeting on Computational methods and NMR spectroscopy: A powerful synergy for chemistry, materials science and biology. Pisa, Italy, December 10th, 2019.
- Alexandre Bonvin - lecture - “Integrative modeling of biomolecular complexes”. Applied Bioinformatics in Life Sciences, Leuven, Belgium, February 13-14, 2020.
- Alexandre Bonvin - lecture + tutorial - “Integrative modeling of biomolecular complexes”. Online BioExcel Summer School on biomolecular simulations. July 22-26, 2020.
- Alexandre Bonvin - lecture - “Integrative modeling of biomolecular complexes”. North Jersey ACS NMR Topical Group - 2020 Virtual NMR Symposium. Oct. 20, 2020.
- Alexandre Bonvin - lecture + tutorial - “Exploring protein docking with HADDOCK”, Structural Bioinformatics course. EMBL-EBI, Hinxton UK, Nov. 26, 2020.
- Zuzana Jandova - lecture - “When HADDOCK meets GROMACS”. Online BioExcel Winter School on biomolecular simulations. Nov. 30 – Dec. 4, 2020.

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

- Alexandre Bonvin - lecture + tutorial - “Integrative modeling of biomolecular complexes”. Online BioExcel Winter School on biomolecular simulations. Nov. 30 – Dec. 4 , 2020.
- Alexandre Bonvin - lecture + tutorial - “Integrative modeling of biomolecular complexes”. Online HADDOCK workshop at the International Symposium on Grid and Cloud computing. Taipei, March 22nd, 2021.
- Alexandre Bonvin - lecture - “Integrative modeling of biomolecular complexes”. Online webinar series for the 50 years of the Molecular Biophysics and 100 year birth anniversary of Prof. G.N. Ramachandran at IIT Bangalore, April 13nd, 2021.
- Alexandre Bonvin - lecture + tutorial - “Integrative modeling of biomolecular complexes”. EMBO practical course on Integrative modelling of biomolecular complexes. Online, May 30 – June 5, 2021
- Alexandre Bonvin - lecture + tutorial - “Integrative modeling of biomolecular complexes”. Online BioExcel Summer School on biomolecular simulations. June 7-11, 2021.
- Alexandre Bonvin - lecture - “Integrative modeling of biomolecular complexes”. Webinar for the PhD school at SISSA, Italy. June 16th, 2021
- Alexandre Bonvin - lecture + tutorial - “Integrative modelling of biomolecular complexes”. PRACE autumn School on fundamentals of biomolecular simulations and virtual drug development. Sept. 20-24, 2021
- Alexandre Bonvin - lecture + tutorial -- “Exploring protein docking with HADDOCK”, Structural Bioinformatics online course. EMBL-EBI, Hinxton UK, Oct. 14, 2020.
- Alexandre Bonvin - keynote lecture: “Integrative modelling of biomolecular complexes”. 5th international Symposium on Bioinformatics. Dec. 15-17, 2021

Webinars:

- Alexandre Bonvin - “[The HADDOCK2.4 server: New Features and Guided Demo](#)”. BioExcel webinar, June 21st, 2020..
- Alexandre Bonvin - “[Integrative modelling of biomolecular complexes with HADDOCK](#)”. SBGrid webinar, June 19th, 2021.

Use Case 2:

High-throughput Modelling of Interactomes

Final Report

In Use Case 2, we developed tools and workflows for high-throughput modelling of interactomes using HADDOCK. To achieve this pre-exascale task we first identified possible bottlenecks: 1) number of files generated, 2) distribution of tasks and 3) scaling. We then proceeded with relevant optimizations. Furthermore, we created a benchmark dataset containing a large number of both true-positive and false-positive interactions that served as the basis for testing HADDOCK's capability to model interactomes and correctly distinguish positive interactions from false ones. A small-scale proof of concept was done to validate the viability of using HADDOCK to categorize the interactions and reported in [D3.3 - Use Case Progress Report](#). Further efforts focused on code and workflow engine developments to enable this pre-exascale task.

Software

Reaching exascale, which will be required for e.g., large scale simulations of interactomes or drug screening campaigns, will require efficient workflows and job management machinery, but also consideration about how and where to run the tasks and efficiently manage the huge amount of data being generated.

The number of files generated and how it can impact the simulations was investigated via a [benchmark containing 1,000,000 files](#) – docking models in PDB format (plain text file) - generated by HADDOCK. A typical docking run with HADDOCK generates several thousands to tens of thousands of files. Each of these files contains critical information in its header to calculate a docking score and rank the models. In an exascale scenario, hundreds to thousands of docking runs should run in parallel, which means that potentially several million files might be generated. The I/O was benchmarked considering the most exhaustive tasks: 1) extract the energy terms from the header of the files, 2) calculate the HADDOCK score and 3) write a sorted list which represents the final ranking.

The timings were investigated in detail as reported in deliverable **D1.2 - Exascale co-design benchmark cases** and have shown that the processing of these 1M files could take from up to 2:15 hours (real-time) depending on the machine settings. These results have further galvanized our efforts to reduce the number of files

generated by HADDOCK and minimize network traffic during the computations. The focus of UC2 has been so far related to HADDOCK v2.4 and its Python3 implementation (v2.5 – not yet released). But we are also tackling this challenge in the development of the new modular version of HADDOCK v3.0 (see deliverable [D1.5 – Long-term hardware and application roadmap](#)). While the compute intensive parts of the HADDOCK workflow are an example of parallel execution with perfect or near-perfect parallel efficiency, the pre- and post-processing steps are sequential. The compute intensive parts also generate a large amount of rather small files, which can stress both network and filesystem. For this reason and in response to user requests we have redesigned HADDOCK's internal job handling machinery, adding options to send self-contained jobs that can run fully locally on node and only return data upon completion, in a similar way as what is currently done in HTC/grid submission mode.

Exploitation of HPC resources

The distribution of tasks has been prototyped in collaboration with the PyCOMPSs Team as was successfully demonstrated using the MareNostrum4 supercomputer and described in deliverable [D1.3 - Updated software roadmaps](#), and with an *in-house* scheduler prototype capable of handling very large volumes of simulations, custom built to optimize HADDOCK execution.

The scheduler prototype ([GitHub - haddock-pilot](#)) is an MPI-based HADDOCK resource manager that distributes the workload on all allocated nodes in an HPC system. In the default running mode HADDOCK starts locally on a login/master node and takes responsibility for submitting the workloads to the local batch system (or grid submission). For large scale simulations this might generate a very high load on the queuing system scheduler which has to handle thousands of submissions. In the new pilot scheduler, the pilots are started on all allocated nodes in an HPC system through a single submission to the batch system. Each allocated node will run the pilot, which pulls in and processes HADDOCK workloads locally, minimizing network communication during the computations and only returning the data once the run has completed. The pilots automatically stop once the work units have all been processed.

Using this pilot, we investigated how HADDOCK would benefit from an increasing number of resources. Its scalability goals are to deliver, in the shortest time possible, models of all predicted interactions, which is a strong scaling problem. By increasing the number of nodes on which the computations are distributed and minimizing network traffic we can reduce the time to solution. HADDOCK main use is in HTC mode. The submissions via the HADDOCK web portal for example make heavy use of EOSC HTC resources. First benchmarking results with the

haddock-pilot mechanism indicate a strong scaling, linear with the number of nodes allocated (Table 2).

Table 2: HADDOCK Scaling performance

Nodes	Cores	Wall Time [min]	Speedup
1	96	1050	1,0
2	192	529	2,0
5	480	193	5,4
10	960	100	10,5
50	4800	20	52,5

The reported scaling benchmark task consists of modelling 100 biomolecular complexes of similar size (which defines the computational requirements per complex), running the full HADDOCK workflow consisting of the sampling of 10,000 models per complex and refinement of the best 200, based on a combination of short molecular dynamics and energy minimization as implemented in HADDOCK (see [HADDOCK2.4 manual - the docking protocol](#) for details). Complexes are taken from Docking Benchmark 5, available from [GitHub - haddocking/BM5-clean](#) (Vreven et al., 2015).

In another, complementary, effort the current benchmarking scripts developed over the years by multiple group members for HADDOCK v2.4 have been streamlined into a single pipeline (see [GitHub - haddocking/benchmark-tools](#)) which has already being used successfully by Master students, demonstrating its ease of use. It will be further integrated with the haddock-pilot code.

All this work done mainly in the context of HADDOCK v2.4 has directly impacted our efforts in the development of the modular version of HADDOCK v3. The execution routines of HADDOCK v3 ([GitHub - haddocking/haddock3](#)) have been developed with scaling, reproducibility, and accessibility in mind. HADDOCK v3 features a fully-fledged HPC submission routine (batch mode), option to make the run self-contained so it can be shared/executed in a different machine, e.g., cloud or locally on a node of an HPC system, minimizing in that way network traffic. Benchmarking capabilities are available directly from the command-line interface. Further optimizations such as the reduction of network traffic when running in a HPC environment and reduction of number of files generated are ongoing.

Impact

The code changes driven by this use case consolidated HADDOCK as a viable tool for the study of a large volume of interactions. Several performance bottlenecks were identified and addressed, these changes have been implemented in the source code of HADDOCK v2.4 and are a driving force in the development of HADDOCK v3. We have also secured a grant from the Netherlands eScience Center - eTEC 2020 - Virtual Research Environment for Integrative Modeling, that will allow us to build a virtual research platform to set up, execute (both in HTC and HPC modes) and analyze HADDOCK3 workflows: [Four pioneering projects: 2020 eTEC winners revealed - eScience Center](#)

Presentations:

- Alexandre Bonvin - lecture + tutorial - *"Integrative modelling of biomolecular complexes - toward interactome modelling"*. [HPC EU-ASEAN virtual Summer School](#). July 5-9, 2021

Use Case 3: Rational Drug Design

The **Rational Drug Design** Use Case, introduced in deliverable [D3.1 - Use Case Work Plans](#), was focused on the structural study and design of drug molecules able to bind a target. With an obvious link to the pharmacological field, the project wanted to demonstrate the predictive power of the BioExcel key applications assisted by HPC workflows and how these could help in reducing the number of molecules to be randomly tested. The use case was designed as a collaborative project effort, with **NBD** having an important role as a proposal originator but also as a link to the pharmaceutical area, helped by **IRB** as project leaders, **BSC** as HPC workflow developers and **MPG** and **KTH** as key applications (PMX, GROMACS) experts.

The Use Case was presented as a collection of 4 separate projects, covering different high impact computer-aided drug design techniques in biomolecular workflows, all of them relevant to the pharmaceutical industry, such as structure-based drug design or machine learning techniques applied to drug discovery. Each study was visualized as a dedicated workflow (WF) (see deliverable [D3.1 - Use Case Work Plans](#)). This plan was slightly affected by the appearance of the global **COVID-19 pandemic**, which caused us to move effort from the originally proposed WF2 project to the one presented in this document (*Large-scale SARS-CoV2 mutation analysis using BioExcel HPC workflows*), which was developed during initial lockdown months, briefly introduced in [D3.3 – Use Case Progress Report](#) and also reported in deliverable [D6.3 - HPC for COVID-19 research](#).

The final list of developed projects within the use case is then:

- *WF1: Moving mutational analysis into the structural field for drug design*
- *WF2: Large-scale SARS-CoV2 mutation analysis using BioExcel HPC workflows¹*
- *WF3: Quantitative predictions of binding affinity in lead optimization*
- *WF4: Machine learning for efficient drug design*

This document presents the final reports for each of these 4 projects, including links to the scientific paper drafts and workflows source code. A final updated Use Case work plan can be found in the [Appendix](#).

¹ (replacing originally planned WF2 project
“Tackling mutations inactivating tumor suppressors”)

Final Report

WF1: Moving mutational analysis into the structural field for drug design

Mutations in the kinase domain of the Epidermal Growth Factor Receptor (EGFR) can be drivers of cancer and cause drug resistance in patients under treatment. First-generation ATP-competitive inhibitors lead to treatment resistance; second-generation inhibitors display limited efficacy in circumventing the gatekeeper mutation, and third-generation inhibitors are mutant-selective, but still prone to eventually develop resistance. There is thus a pressing need for anticipating the consequences of such mutations in drug development and clinical practice. We have devised a workflow for classifying EGFR mutations by their impact on binding FDA-approved drugs for predicting drug sensitivity/resistance patterns for clinically relevant EGFR mutations. Our method could have a clear impact in personalized medicine against oncogenic mutations in EGFR and could also be used for other targets where resistance is an issue.

The project focuses on oncogenic missense mutations that confer a selective advantage to a cancer cell and drive cancer progression. The project studied 36 different clinically annotated mutations, classified as either resistance-causing (mutation-induced treatment resistance) or activating mutations (enhancing clinical effectiveness). Mutations are checked on 5 FDA-approved drugs: [Erlotinib](#), [Lapatinib](#), [Gefitinib](#), [Osimertinib](#) and [Icotinib](#). Equilibrium MD simulations (10 replicas) of 100ns each were calculated for all combinations (apo-WT, holo-WT, apo-Mut and holo-Mut), with an accumulated time of up to $\sim 100\mu\text{s}$. Snapshot ensembles taken from these trajectories were used then for the free energy calculations, and a final $\Delta\Delta G$ was computed for all cases. The list of final $\Delta\Delta G$ s were finally compared to the clinical records to determine the capacity of prediction of the method.

The pipeline is able to correctly classify 23 out of 31 mutations by their impact on the binding affinity of FDA-approved drugs. The classification model achieved an accuracy of 0.74, which could be improved to an accuracy of 0.8 when removing charge-changing mutations. When calculating on HPC infrastructure, one free energy prediction for a mutation-drug pair can be obtained in only four hours. Given the accuracy and speed of our workflow, free energy calculations can keep up with the pace of industrial drug discovery projects and can have a positive impact, for example, by identifying other selective kinase inhibitors to which a mutant kinase of a given patient population is susceptible.

The draft paper describing this study is available: [Automatically predicting drug sensitivity and resistance patterns for clinically relevant EGFR mutations](#), and includes supplementary information. The HPC workflows used in the project

were extensively presented in deliverable [D2.3 - First release of demonstration workflows](#), and publicly available as a BioExcel GitHub repository ([GitHub - biobb hpc workflows](#)).

WF2: Large-scale SARS-CoV2 mutation analysis using BioExcel HPC workflows

The BioExcel project has been deeply involved in COVID-19 research since the pandemic appeared. In response to the disease, BioExcel performed [research SARS-COV-2](#), an important part of which consisted of studies performed using HPC and our advanced software applications and expertise, mostly related to the spike protein and the way in which SARS-CoV-2 invades human cells. We also launched a [series of actions support the wider community engaging in SARS-CoV-2 research](#)

In this context, the set of workflows designed and developed for the WF1 were quickly modified and updated to tackle a set of particularly interesting (and challenging) questions: a) understanding the mechanism of virus entrance into the cell and the adaptation of the virus to different host species, b) understanding the different sensitivity to the virus (beyond the age) of individuals, c) predict the next mutations of the virus and how it might adapt to be even more infectious and d) understanding how the virus has evolved by comparing its structure/genome to other coronavirus strains in different species including RaTG13, Pangolin, SARS-CoV as well as the US-variant. The mechanism of entrance of the virus is based on the capsid protein spike, which is recognized by an extracellular protease (ACE2). The main objective was to determine the impact of genetic changes in the viral spike Receptor Binding Domain (RBD) and in the ACE2/RBD complex for the recognition of the virus. This can be achieved (similarly to the previously presented WF1) using free energy calculations to trace the impact of the mutational landscape on the binding of RBD to the host receptor proteins.

The most remarkable results of the project outline a unique evolutionary process in which SARS-CoV-2 gained the ability to directly infect humans after having developed an unusually high RBD affinity for horseshoe bat, *Rhinolophus affinis* ACE2 ([RaTG13](#)), its natural host receptor, and a highly optimized spike structure and dynamics. These features rule out the requirement of SARS-CoV-2 adaptation in an intermediate host, where other animal species, if any, were acting as vectors in SARS-CoV-2 transfer from bats to humans due to its broad host tropism.

The draft paper describing this study is available: [Evolutionary Path and Host-selection Mechanism of SARS-CoV-2](#), and includes supplementary information. The project is also an important part of the BioExcel review: [HPC- approaches for Molecular Dynamics. Covid-19 research: a use case](#), to be submitted soon. The HPC workflows used in the project were extensively presented in the [D2.3 -](#)

[First release of demonstration workflows](#), and are available as a BioExcel GitHub repository (see [GitHub - biobb hpc workflows](#)).

WF3: Quantitative predictions of binding affinity in lead optimization

Lead optimization is the phase where leads are optimized for better interactions with the target protein. Usually, it involves only small changes in the core of the lead, or addition of substituents at selected positions. Provided that changes at this stage are rather conservative among a given chemical series, free energy methods can give useful insights into favourable affinity changes. A short but important peptide was chosen as a test case to take advantage of the expertise achieved with the previous workflows and explore the capacity of predictions for amino acid mutations in peptide ligands. Pre-generated amino acid libraries allow a fast mutation and screening process. Large-scale free energy calculations using relevant protein conformations from MD were carried out, followed by the selection of the most promising candidates.

The chosen peptide was the duodecimal peptide PMI that is known to compete with p53 for binding to the Murine Double Minute 2 (MDM2) or Murine Double Minute X (MDMX) protein homologs. p53 is critical for maintaining genetic stability and preventing cancer. The E3 ubiquitin ligase MDM2 and its homologue MDMX act as negative regulators of p53. Designing inhibitors of MDM2 or MDMX is an attractive strategy for enhancing p53 activity and thus achieving the desired antitumoral therapeutic effect.

The affinity of the peptide PMI is roughly two orders of magnitude higher than that of the same length p53 peptide. The experimental K_d of PMI is known to be 3.2×10^{-9} M for MDM2 and 8.5×10^{-9} M for MDMX. Experiments obtaining K_d values for the twelve Alanine scanning mutants of PMI are also available, with values ranging between 10^{-4} and 10^{-10} M and some of the mutated peptides being extremely potent.

The study started with high-resolution crystal structures of both proteins complexed with PMI and p53 (PDB identifiers: 3EQS, 3EQY, 1YCR, 3DAB respectively). A simulated Alanine scanning was performed, and resulting MD trajectories were used to compute the binding free energy values.

Preliminary results show a very good correlation between experimental data and calculated data, with a mean absolute error lower than 2 kcal/mol for all the systems, improving to 1.0 kcal/mol if an outlier is not considered. In particular, for the PMI systems (MDM2-PMI, MDMX-PMI), an R^2 higher than 0.9 is achieved,

whereas for the p53 systems (MDM2-p53, MDMX-p53), the obtained R^2 is around 0.7, maybe due to the higher number of ionizable (charged) residues.

The HPC workflows used in the project were extensively presented in deliverable [D2.3 - First release of demonstration workflows](#), and is available as a BioExcel GitHub repository (see [GitHub - biobb hpc workflows](#))

WF4: Machine learning for efficient drug design

Essential biological processes such as DNA replication, transcription, splicing, and repair are driven by protein-DNA interactions. Protein residues in the DNA-binding site enclose crucial information for the binding mechanism; furthermore, DNA-sequence preferences are crucial for forming the DNA-complex and performing the biological activity. The identification of these DNA preferences has been a popular field of study in the last decade. Different types of protein-DNA binding site predictors exist, mainly based on sequence and/or structural data. The BioExcel approximation, taking advantage of the HPC resources and pre-exascale MD workflows, is a new predictor that uses a different type of descriptors: DNA conformational and flexibility observables. The developed Machine Learning (ML) workflow predicts protein-DNA binding affinity based on experimental binding data and computationally derived DNA flexibility features. The input data depends on the experimental technique that has been used to detect protein-DNA sequence selection (uPBM, gcPBM and HT-SELEX).

The development of the project led us to a set of conclusions: efficient predictive algorithms have to solve a series of intrinsic problems. On the one hand, the concept of transcription factor binding site (TFBS) is not uniquely defined as it deeply depends on the experimental technique used to detect it, generating noise and making it impossible to create a universal predictor. On the other hand, transcription factors use a repertoire of mechanisms for selecting target DNA sequences, and the most informative parameters describing these mechanisms largely depend on the sequence variability explored by the experiment. The complexity of the problem increases even more if in vitro predictions are tried to be extrapolated in vivo, where other factors than transcription factor affinity play a role.

On the whole, the final workflow provides excellent results, outperforming all available methods when predicting in vitro transcription factor binding sites irrespective of the experiment used for validation. The workflow trained on in vitro data has an excellent ability to detect the binding sites of the same transcription factor in vivo. Although it predicts many potential binding sites where no experimental evidence of in vivo binding is detected, a grand majority

of these seemingly false positives are trivially explained by chromatin structure and nucleosome occupancy. Combining this workflow with simple nucleosome maps we were able to locate in vivo TFBS with a high accuracy.

The draft paper describing this study is available: [DNAffinity: A Machine-Learning Approach To Predict Dna Binding Affinities Of Transcription Factors](#), and includes supplementary information. The workflows used in the project were presented in deliverable [D2.4 – Development of a framework for the combination of HPC and HPDA operations](#).

Software

The first three projects rely strongly on the **GROMACS** and **PMX BioExcel** key applications. Workflows were built using the **BioExcel Building Blocks** (BioBB) library, taking advantage of its high interoperability and reproducibility features. Besides, the compatibility of the BioBB workflows with the **PyCOMPSs** workflow manager allowed an efficient use of HPC resources. This was particularly useful in the fast-growth free energy calculations, which needed up to 1,000 independent simulations to compute a single free energy difference. These were easily distributed to execute using HPC resources using PyCOMPSs (see following section).

The last project, involving **Machine Learning** (ML) methods, was firstly designed and implemented with the **Python Scikit Learn** and **Tensorflow** libraries, and the workflows were later cloned using the BioBB ML building blocks, gaining reproducibility and easy access to HPC resources. In addition, DNA MD simulations and helical parameters used as observables for the training process were also calculated with BioBB workflows.

The collection of BioBB HPC workflows used in the use case are openly available as a source code repository in the [BioExcel GitHub](#), with information about how to install and reproduce them in HPC machines through a **Conda Pack** environment. Some of the workflows and sub-workflows used for the use case are also available, in a reduced or lite version, as Jupyter Notebooks. They can be found in the demonstration workflows section of the **BioBB** website (e.g. [Protein MD setup](#), [mutation free energy](#), [structural DNA helical parameters](#)).

The use case with its different projects and workflows has been the perfect test case for the **BioBB** library. The iterative process followed during the development and testing of the workflows allowed great improvements in the **BioBB** library, with at the same time reaching a great efficiency in HPC calculations and achieving promising results in less time. It is, in our opinion, a perfect example of how a

scientific use case can help in the design and development of workflows and tools, and how in turn these tools can help in finding good results.

A paper presenting the BioExcel Building Blocks library was published: [*BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows \(Nature Sci Data 6, 169, 2019\)*](#), and includes supplementary information.

Exploitation of HPC resources

The Rational Drug Design use case would not have been possible without HPC resources. All workflows were designed to be launched in HPC supercomputers, taking advantage of the work distribution and parallel execution, using the combination of **BioBB** library and **PyCOMPSs** workflow manager.

For the first 3 workflows, equilibrium MD simulations were computed in **MPI** regime, using hundreds to thousands of cores in each of the jobs (**GROMACS**). Binding free energies were computed using 1,000 independent thermodynamic integration simulations (500 forward, 500 reverse) with a BioBB workflow combining **GROMACS** and **PMX**. This calculation, which would take months in a single desktop, was run in minutes using thousands of supercomputer cores (see deliverable [D2.3 - First release of demonstration workflows](#)).

These HPC workflows were used to run a scalability study that was presented to the European Commission as part of a Scalability Fitness Check report ([Fitness Check 2021 – BioExcel CoE Scalability Report of Flagship codes](#)). Weak and strong scalability studies were run, with jobs launched in the BSC Marenostrum supercomputer with up to 6,000 cores. The scalability study is now being extended in the [Petascale supercomputer Discoverer](#) (Sofia Tech Park, Bulgaria), with jobs using up to 40,960 cores in parallel (draft paper in preparation).

HPC resources were important in this use case not just for the computation, but also for the amount of disk space needed for the projects. As an example, 200 TB of active storage were used to compute all the differences in binding free energy for the SARS-CoV-2 mutants with the hACE2 protein. Once the HPC projects were closed, archival storage (e.g. storage tapes) was also very important, to store and save all generated data for later queries.

The COVID-19 project (WF2) was presented to the [PRACE COVID-19 Fast Track Call for Proposals](#), and the study named “**Exploring Covid19 Infectious Mechanisms and Host Selection Process**”, was awarded with **6 million core hours** on Joliot-Curie supercomputer (CEA/GENCI, France) and was published as

a [PRACE success story](#). The same project has been also recurrently presented to the [Red Española de Supercomputación \(RES\) call for proposals](#), being granted in three periods during the years 2020 and 2021, with a total accumulated time of 15 million core-hours on the BSC Marenostrum supercomputer (BCV-2020-2-0003, BCV-2020-3-0011 and BCV-2021-2-0028).

The draft paper describing the BioExcel HPC workflows is available: [Enabling the execution of large scale workflows for molecular dynamics simulations](#), and includes supplementary information. The BioExcel scalability study document is available at: [Fitness Check 2021 – BioExcel CoE Scalability Report of Flagship codes](#).

Impact

HPC workflows designed and developed for this use case have been presented in different international activities and events. In particular, the work done in the COVID-19 research field has been internationally recognized. Here we list the main activities linked to the project that demonstrate the impact of the work in the field:

Success Stories:

- ICEI/FENIX Success Story: [Using Icei Resources For Atomistic Molecular Dynamics Simulations On Covid-19 Related Research](#)
- PRACE Success Story: [Investigating The Impact Of Mutations In SARS-CoV-2](#)

Webinars:

- CECAM, The importance of being H.P.C. Earnest: [Bioexcel Building Blocks and HPC. A Test Case In Covid Research](#).
- CECAM, CoVid-19: challenges and responses in simulation, modeling and beyond: [HPC And Big Data Approaches In Covid-19 Research](#)

Conferences, workshops and courses:

- VIII Jornada de Bioinformàtica i Genòmica: [Big Data and HPC to fight COVID-19](#)
- SC 2021, PAW-ATM 2021: The 4th Annual Parallel Applications Workshop, Alternatives to MPI+X: [Towards an Efficient Use of Exa-Scale High-Performance Computing](#)

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

- 17th IEEE eScience 2021: [Towards an Efficient use of Exascale High-Performance Computing](#)
- COVID-19 PRACE days 2021, Scientific Parallel Track COVID-19: [Exploring Covid19 Infectious Mechanisms and Host Selection Process](#)

HPC Resources:

- PRACE COVID-19 Fast Track Call for Proposals: **Exploring Covid19 Infectious Mechanisms and Host Selection Process**
- Red Española de Supercomputación (RES) Proposals: **Exploring Covid19 Infectious Mechanisms and Host Selection Process** (BCV-2020-2-0003, BCV-2020-3-0011 and BCV-2021-2-0028).

Use Case 4a: Fluorescent Proteins

Fluorescent proteins are the backbone for high-resolution biological imaging, but designing suitable proteins for specific experimental conditions is difficult. The goal of this UC was to unlock the predictive power of MD simulations for optimizing fluorescent proteins (FPs). We thus have developed a user-friendly protocol for automatically computing the relevant properties of such proteins and their mutants based on an established atomistic molecular dynamics model. The protocol combines (i) force field MD simulations with GROMACS, (ii) PMX and free-energy calculations, and (iii) QM/MM calculations to predict the thermostability (protein folding and oligomerization affinity) and photochemical properties (absorption spectrum and emission spectrum) of fluorescent proteins.

Final Report

We integrated the various calculations into a Fluorescent Protein Computer-Aided Design (FluProCad, Fig. 6) workflow with the following features: (i) construction of a solvated structural model with user-specified mutation(s); (ii) automatic creation of hybrid topologies of the mutating amino acid for free energy calculations (iii) free energy calculations with PMX (iv) QM/MM calculations of optical spectra. We validated the workflow by comparing the solution structure, thermal stabilities, oligomerization affinity, absorption spectra, and emission spectra for mutants of *Aequorea victoria* Green Fluorescent Protein (avGFP) and of rsGreen0.7. These results are briefly discussed below. Within the context of this UC, we also performed QM/MM MD simulations with the GROMACS/CP2K, developed in WP1, to retrieve additional properties of specific fluorescent proteins, including activation barriers, as discussed below.

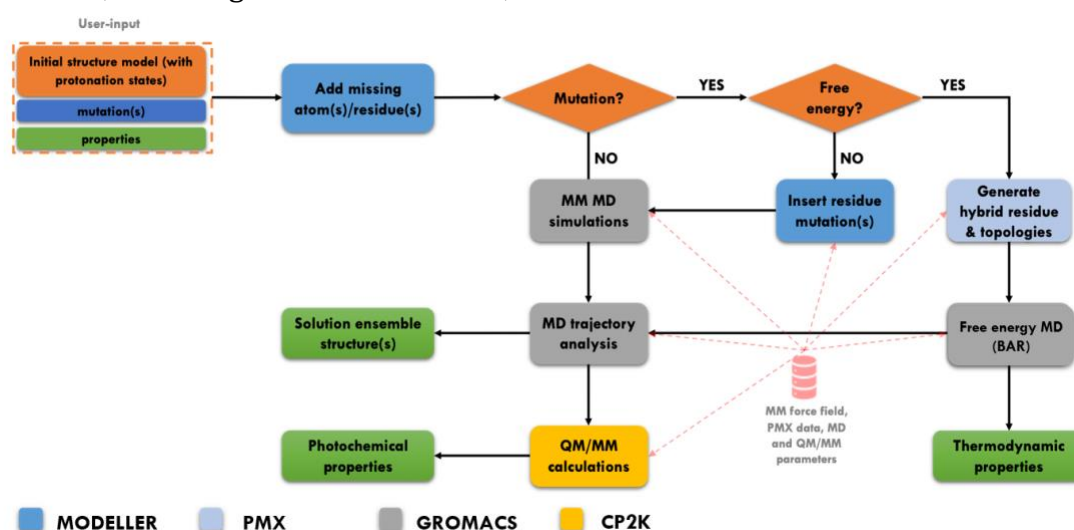


Figure 6: Fluorescent Protein Computer-Aided Design (FluProCad) workflow for computing thermodynamic and spectral properties of fluorescent proteins.

Computational screening of *Aequorea victoria* GFP mutants

We used the FluProCad workflow to investigate the effect of specific mutations in GFP (i.e. S65T, F64L, A206K, and S65T+F64L) on the optical spectra, protein stability and dimerization affinity. Our results suggest that the F64L mutation improves folding while the A206K mutation reduces dimerization affinity, in line with experiment. Stokes shifts computed from the simulations also agree with experiment. In collaboration with the group of Prof. Peter Dedecker from the KU Leuven, we further validated the workflow by blindly predicting the solution structures of several rsGreen0.7 variants, and comparing the solution ensembles to ensembles of MD simulations starting from the unpublished x-ray structures of these proteins. These results will be submitted for publication. The FluProCad workflow, including the relevant force field and QM/MM parameters for simulations of GFP variants, is available for download from the public repository [bioexcel/FluProCad-2.0](https://bioexcel.eu/FluProCad-2.0).

QM/MM MD simulations of photo-active proteins

In collaboration with the group of Prof. Jasper van Thor at Imperial College, we used the CP2K/GROMACS QM/MM interface developed in WP1, to elucidate the isomerization mechanism of the chromophore in the photo-switchable fluorescent protein rs-Kiir. As the rate at which the chromophore isomerizes between the dark *trans* and fluorescent *cis* isomers, is an important parameter in nanoscopy experiments with such proteins, detailed atomic-level information about the mechanism, and in particular about the effect of the protein environment on the barrier is essential to improve the performance of such proteins. We thus computed the free energy profile for this process by combining QM/MM MD with the Accelerated Weight Histogram (AWH) enhanced sampling method. The results suggest that isomerization occurs via a hula-twist mechanism with a barrier of 124 kJ/mol. Under the reasonable assumption that crossing this barrier is the rate-limiting step, this result is in good agreement with the rate estimates from the experiment.

In collaboration with Prof. Janne Ihalainen at the University of Jyväskylä, we have investigated the activation mechanism in the Phytochrome photo-receptor, which are promising templates for the development of fluorescent proteins that can emit in the near-infrared. However, because the biliverdin chromophore enters a photocycle upon light absorption, the fluorescence quantum yield is very low. Improving the latter by means of mutations requires a detailed understanding of the activation mechanism first. To provide atomistic insights into that mechanism we performed a series of QM/MM MD simulations of all relevant steps in both the electronic excited state and the ground state. To address the initial response to

photon absorption, we used non-adiabatic QM/MM simulations. The results of these simulations suggest an ultrafast (sub-ps) photo-isomerization of the biliverdin cofactor. To track structural changes after the photo-isomerization step we performed enhanced QM/MM molecular dynamics with the CP2K program and extracted the free energy profile for the further relaxation of the biliverdin chromophore. The results of these simulations suggest a conformational change in the chromophore into a conformation that matches the x-ray structure of the activated state. The barrier associated with this relaxation is estimated at 67 kJ/mol, in agreement with timescales extracted from time-resolved spectroscopic measurements.

Software

The following software packages were used in this use case:

- GROMACS for performing molecular dynamics and free-energy simulations
- PMX for setting up and analyzing free-energy simulation of dimerisation and folding of fluorescent proteins
- CP2K and GROMACS/CP2K interface for performing QM/MM simulations of absorption/emission spectra and isomerization free energy profiles of photoactive proteins.

Exploitation of HPC resources

The FluProCad workflow is parallel on a large scale and suitable for high-throughput scanning of the effects of mutations on the key properties of these proteins at a wide range of modern HPC platforms. For FluProCad simulations, we used the following resources:

- In total we computed ~3.5 microseconds of MD simulations for all GFP systems in order to calculate the effect of mutations on the folding stability and dimerization affinity. These computations consumed **1,500,000 core hours** (300,000 per system). A potential limitation is that the analysis of the results requires post-processing a very large number of files.
- The QM/MM calculations performed to model the Stokes shift involved 100 independent single-point excited state calculations to evaluate both the absorption and emission spectra peaks. The calculations are generated from 100 independent snapshots from the MD trajectories and can be executed at the same time using 1 node/snapshot. In total, spectra simulation for each GFP variant consumed approximately 10,000 core hours, for a total of **50,000 core hours** for all studied mutants.

Also the other applications in this use case relied on HPC:

- Enhancing the sampling in our QM/MM simulations by means of Umbrella Sampling or AWH is parallel with perfect or near-perfect efficiency. For umbrella sampling simulations of the isomerization of phytochrome fluorescent protein we performed 0.8 ns of MD simulation for 16 umbrellas. Each umbrella trajectory was computed on 4 nodes (512 cores), with approximate performance of 6-8 ps per day. The total costs for the umbrella sampling therefore amounted to **2,500,000 core hours**. For the AWH simulations aimed at calculating the free energy profile associated with the chromophore isomerization in the rsKiir0 fluorescent protein, we used 6 walkers (individual simulations) each running on 2 nodes or 256 AMD EPYC cores (12 nodes in total) with an overall performance of ~8 ps/day. For a total of 0.5 ns simulation time we used **400,000 core hours**. All QM/MM simulations were performed on the Mahti cluster at CSC (Finland). A potential limitation for such computations, on other platforms, would be a rather large memory requirement for the QM/MM calculations. For example the QM/MM calculations on the Phytochrome with 91 atoms inside the QM region, requires up to 1 Gb of memory per core (128 Gb per node). Increasing the size of the QM subsystem will consequently further increase memory demands.

Impact

The end users of the FluProCad workflow are experimentalists who wish to optimize the properties of fluorescent proteins. To test this, we have shared the workflow with collaborators at KU Leuven, who could successfully predict the solution structures of a series of rsGreen0.7 mutants. These results are shared in the GitHub repository [bioexcel/FluProCad-2.0](https://github.com/bioexcel/FluProCad-2.0).

Publications:

- S. Mustalahti, D. Morozov, H. L. Luk, R. R. Pallerla, P. Myllyperkiö, M. Pettersson, P. Pihko, G. Groenhof, "[Photoactive Yellow Protein Chromophore Photoisomerizes around a Single Bond if the Double Bond Is Locked](#)", *The Journal of Physical Chemistry Letters*, 2020
<https://doi.org/10.1021/acs.jpcllett.0c00060>
- E. Claesson, W. Y. Wahlgren, H. Takala, S. Pandey, L. Castillon, V. Kuznetsova, L. Henry, M. Panman, M. Carrillo, J. Kuebel, R. Nanekar, L. Isaksson, A. Nimmrich, A. Cellini, D. Morozov, M. Maj, M. Kurttila, R. Bosman, E. Nango, R. Tanaka, T. Tanaka, L. Fangjia, S. Iwata, S. Owada, K. Moffat, G. Groenhof, E. A. Stojkovic, J. A. Ihalainen, M. Schmidt, S. Westenhoff, "[The primary structural photoresponse of phytochrome](#)"

- [proteins captured by a femtosecond X-ray laser](https://doi.org/10.7554/eLife.53514)”, *eLife*, 2020
<https://doi.org/10.7554/eLife.53514>
- Gerrit Groenhof, Vaibhav Modi, Dmitry Morozov,
<https://doi.org/10.1016/j.sbi.2019.12.013>, *Current Opinion in Structural Biology*, 2020 <https://doi.org/10.1016/j.sbi.2019.12.013>
 - R. Dods, P. Bath, D. Morozov, V. Ahlberg Gagner, D. Arnlund, H.-L. Luk, J. Kuebel, M. Maj, A. Vallejos, C. Wickstrand, R. Bosman, K. R. Beyerlein, G. Nelson, M. Liang, D. Milathianaki, J. Robinson, R. Harimoorthy, P. Berntsen, E. Malmerberg, L. Johansson, R. Andersson, S. Carbajo, E. Claesson, C. E. Conrad, P. Dahl, G. Hammarin, M. S. Hunter, C. Li, S. Lisova, A. Royant, C. Safari, A. Sharma, G. J. Williams, O. Yefanov, S. Westenhoff, J. Davidsson, D. P. DePonte, S. Boutet, A. Barty, G. Katona, G. Groenhof, G. Branden, R. Neutze, [“Ultrafast structural changes within a photosynthetic reaction centre”](https://doi.org/10.1038/s41586-020-3000-7), *Nature*, 2021 <https://doi.org/10.1038/s41586-020-3000-7>
 - Elina Multamäki, Rahul Nanekar, Dmitry Morozov, Topias Lievonon, David Golonka, Weixiao Yuan Wahlgren, Brigitte Stucki-Buchli, Jari Rossi, Vesa P. Hytönen, Sebastian Westenhoff, Janne A. Ihalainen, Andreas Möglichen & Heikki Takala, [“Comparative analysis of two paradigm bacteriophytochromes reveals opposite functionalities in two-component signaling”](https://doi.org/10.1038/s41467-021-24676-7), *Nature Communications*, 2021
<https://doi.org/10.1038/s41467-021-24676-7>
 - Morozov, D., Mironov, V., Moryachkov, R. V., Shchugoreva, I. A., Artyushenko, P. V., Zamay, G. S., Kichkailo, A. S., [The role of SAXS and molecular simulations in 3D structure elucidation of a DNA aptamer against lung cancer](https://doi.org/10.1016/j.omtn.2021.07.015). *Molecular Therapy-Nucleic Acids*, 2021
<https://doi.org/10.1016/j.omtn.2021.07.015>
 - Mironov, V., Shchugoreva, I. A., Artyushenko, P. V., Morozov, D., Borbone, N., Oliviero, G., Groenhof G., Kichkailo, A. S., et.al., [Structure and Interaction Based Design of Anti-SARS-CoV-2 Aptamers](https://doi.org/10.1002/chem.202104481). *Chemistry Europe*, 2022 <https://doi.org/10.1002/chem.202104481>

Conferences, workshops and courses:

- [Workshop in Multi-Scale Modelling](#), Lorentz Center Leiden, 2019
- CECAM workshop [“Frontiers in multi-scale molecular modeling of photo-receptors”](#)
- [Finnish Computational Chemistry Days 2019](#)
- 17th International Congress on Photobiology: [Observe while it happens: catching photoreceptors in the act with free electron lasers and computer simulations](#)
- [“9th International Symposium on Photochromism”](#)

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

- CECAM workshop on simulation of open systems in Chemistry, Pharma, Food Science and Immuno-diagnostics: [Molecular dynamics simulations at constant pH](#)
- Virtual Workshop: [Best Practices in QM/MM Simulation of Biomolecular Systems](#)
- [Computational Approaches to Understanding and Engineering Enzyme Catalysis](#)

Webinars:

- BioExcel webinar: [Multiscale QM/MM simulations: exploring chemical reactions using novel GROMACS/CP2K interface](#)
- PC2 online seminar: [Using Novel Gromacs/CP2K Interface to Perform Multi-scale QM/MM Simulations](#)

Tutorials:

- Advanced GROMACS workshop: [QM/MM simulations with GROMACS and CP2K](#)
- [QM/MM with GROMACS + CP2K workshop at EPCC](#)
- FEBS Course Computational Approaches to Understanding and Engineering Enzyme Catalysis: [An introduction into molecular dynamics simulations of proteins](#)

Use Case 4b: Proton Dynamics

This Use Case applied the two QM/MM interfaces developed within BioExcel, namely the GROMACS/CP2K interface developed as part of BioExcel-2 and the MiMiC GROMACS/CPMD interface developed as part of BioExcel-1. We focused on the simulation of proton transfer events in mass spectrometry experiments of a DNA oligo, namely a heptanucleotide with sequence d(GpCpGpApApGpC) in the presence of ammonium counterions. Our work addressed the question whether ammonium ions may be involved in proton transfer with DNA.

Whereas applications of QM/MM have so far been limited mostly to single molecules only, this Use Case provided the unprecedented opportunity to study intermolecular proton transfer phenomena and the impact of the latter on the structure, dynamics and energetics of a biomolecular complex. The project built on the wide expertise in DNA simulation of Prof. Orozco's group at IRB [Rueda et al. 2003; Rueda et al. 2005; Arcella et al. 2012; Porrini et al. 2017] and on previous collaborations regarding computational mass spectrometry of DNA carried out jointly by the FZJ and IRB groups [Arcella et al. 2015].

The choice of the DNA oligo as a specific system of interest for the study of proton transfer in mass spectrometry replaced the proteins originally planned to be studied. This choice was made in part because experimental data regarding the main charge of the complex in the gas phase [Arcella et al. 2015] became readily available, enabling comparison with simulation. As in the originally proposed project, the system of interest is a complex between two (bio)molecules, thereby equally meeting the goals of novelty and innovation with regards to use of QM/MM simulation for this purpose.

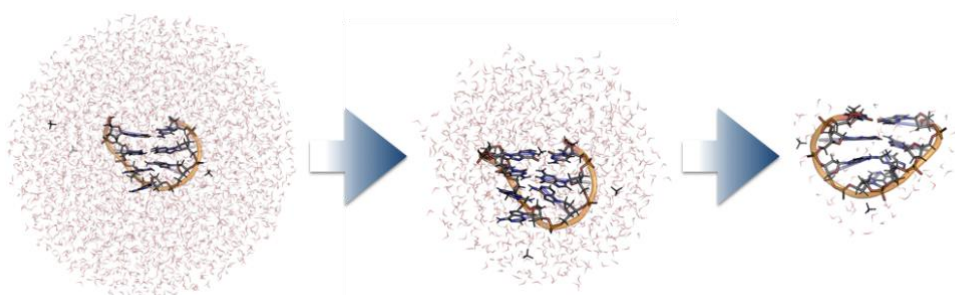


Figure 7: Illustration of the desolvation process of the d(GpCpGpApApGpC) heptanucleotide in a water droplet within mass spectrometry conditions.

Final Report

Classical MD Simulation of DNA in solutions and in water droplets

We simulated the d(GpCpGpApApGpC) heptamer in aqueous solution. The molecule has a charge of -6 . We used ammonium ions, which are present under experimental conditions as counterions [Arcella et al 2015]. We carried out 500 ns classical MD simulations within the NPT ensemble at 300 K and 1 bar. During the simulation, the oligonucleotide retains its hairpin structure. This consists of a short B-DNA fragment (nucleobases G_1-C_7 and C_2-G_6) and a d(G₃A₄A₅)-triloop.

Selected snapshots from these simulations were used to create water droplets with a 30 Å radius around the DNA molecule. We considered water droplets containing two, three and four ammonium ions to control the total system charges to be -4 , -3 (main charge state) and -2 , respectively, as experimentally observed. For each droplet, we have run 300 x 500 ps MD simulations at 300 K. Dissociated water molecules were removed from the system after each run. After this procedure, less than 50 water molecules were left over. Ammonium ions formed hydrogen bonds either with a single phosphate group or with two neighbouring phosphate groups. The first type of interaction is predominant in the early stage of the desolvation process, while the second one is more present when more water is released, in which deformations of the hairpin structure can be observed. In the time span of 100–150 ns the ammonium–water coordination number drops linearly from 8 (regular in aqueous solution) to 2–4. MD snapshots with the two interactions were selected for this partially hydrated period.

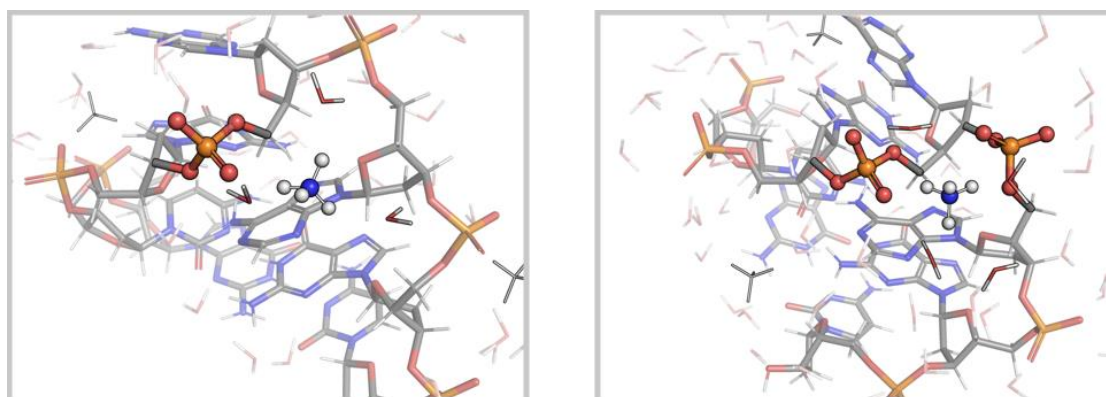


Figure 8: MD snapshots from the simulated desolvation process of the DNA molecule in a water droplet containing ammonium ions. Left: Interaction of an ammonium ion with one phosphate group. Right: Interaction of an ammonium with two phosphate groups building a bridging hydrogen bonding network.

Proton Transfer in Model Systems by MiMiC QM/MM simulations

We carried out dimethyl phosphate as a model for the DNA backbone to study the proton transfer from ammonium ions within a partially hydrated environment.

Systems contained 3, 4, and 5 water molecules. The QM part (treated at the DFT-PBE level of theory) consists of the ammonium ion and dimethyl phosphate. The MM part consists of water molecules. We used the MiMiC interface [Olsen et al 2019]. Umbrella sampling with WHAM analysis was used to determine the proton distribution. The difference of the N–H and H–O distances determines the collective variable: negative values represent ion pair configurations, while positive values describe neutral molecules. For each system, 11 equidistant windows in the interval $[-1.0 \text{ \AA}; 1.0 \text{ \AA}]$ were sampled for 14.5 ps to calculate the free energy profile. The proton transfer profiles are found to be very sensitive to the number of water molecules. The ion pair is stabilised by 5 water molecules, showing a repulsive curve upon proton transfer. Moving to a system with 4 water molecules, the flat potential indicates a shared proton, while the neutral state is stabilized with three water molecules.

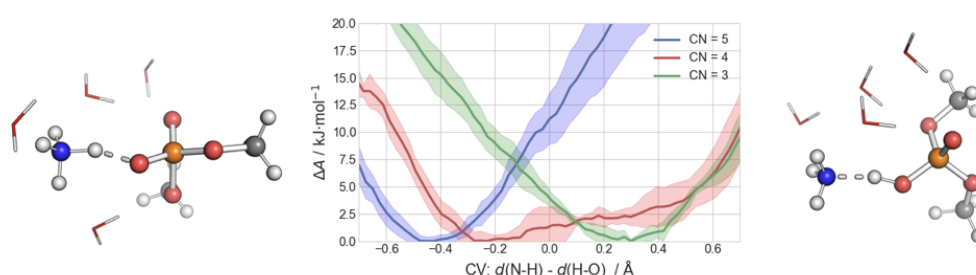


Figure 9: Proton transfer profiles between an ammonium ion and dimethyl phosphate solvated by 3 (green), 4 (red) and 5 (blue) water molecules. Negative CV values describe the ion pair (left), while positive values indicate the neutral state (right) of the molecules.

Ab initio Molecular Dynamics Simulations

We selected three MD snapshots of the main charge state (−3) for ab initio MD simulations to compare the strength of a specific ammonium- phosphate H-bond (the one formed with A₄ phosphate) at different degrees of solvation. The snapshots contained 82, 67 and 32 water molecules with 483, 438 and 333 atoms in total, respectively.

The CPMD program was used for ab initio MD calculations at the BLYP-D level. As a compromise between accuracy and computational cost, we applied an energy cutoff of 70 Ry (density cutoff of 280 Ry). Reflecting wall potentials were required to limit the box size, since water molecules can easily move around under gas phase conditions. The temperature of the systems was first damped in 150 steps and then reheated in 5000 steps to 300 K (0.48 fs/step) to ensure a smooth transition from MM to QM potentials. The simulations were then continued at a constant temperature of 300 K. Up to this point, we carried out 11000, 16000 and 20000 steps of ab initio MD in the order of decreasing system size. For the higher

solvated systems (87 and 67 water molecules) we observe relatively weak hydrogen bonds between the ammonium ion and the A₄-phosphate group with average O–H distances of 1.8 Å. In contrast, the system with less water molecules shows a stronger hydrogen bond of 1.7 Å, also with the tendency to share the proton.

Local Proton Transfer Profiles upon Desolvation by QM/MM simulation

We selected 11 snapshots of the main charge state (–3) for MD-QM/MM simulations, for which we employed the novel QM/MM interface (GROMACS/CP2K) developed within BioExcel-2. We investigated the proton transfer for interactions of an ammonium ion to single phosphate groups at C₂ and A₄ (for each 4 snapshots) and to two neighbouring phosphate groups at A₄–A₅ (3 snapshots). The QM regions comprised the ammonium ion and the DNA backbone of the involved phosphate groups including adjacent deoxyribose moieties. PBE-D/DZVP was used for the QM part. Covalent bonds C4'–C5' and C1'–N_{base} were cut and saturated with H-link atoms. The umbrella sampling protocol, described above for the model system, was used with the modification that 10 windows in the interval [–1.0; 0.8] Å were sampled for 25 ps. Typical free energy profiles are depicted in Figure 10.

The resulting profiles reveal an important feature for proton transfer in the desolvation process of DNA in mass spectrometry. While we observe both preferred ionic states and configuration suitable for proton transfer for the interaction with a single phosphate group (Figure 10a and 10b, respectively), the interaction with a second phosphate stabilises the ionic state in all our simulations. As recognized from our classical MD simulations, the former type of interactions are frequently observed in the early stage of the desolvation process. The latter one arises mostly when the DNA molecule is almost desolvated, causing structural changes due to the strong ionic interactions.

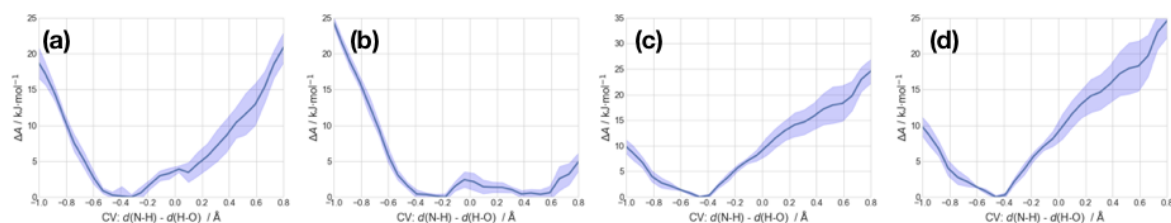


Figure 10: Proton transfer profiles between an ammonium ion and the phosphate groups of the DNA oligomer. Negative CV values describe the ion pair, while positive values indicate the neutral state. Profiles (a) and (b) were calculated for the interaction of an ammonium ion with a single phosphate group, while profiles (c) and (d) correspond to a proton transfer with two interacting phosphate groups.

In conclusion, in spite of the relative simplicity of the system used, the scenarios emerging from the calculations are quite complex. Indeed, the ammonium-phosphate interactions depend strongly on solvation and on the number of interacting phosphate groups.

Software

The simulations were carried out with the following software:

- GROMACS [Abraham et al. 2015]: classical MD simulation for the aqueous solution structure of the oligonucleotide as well as for the desolvation process within water droplets in gas phase
- [CPMD](#): *ab initio* MD simulations for the partially hydrated oligonucleotide
- MiMiC [Olsen et al 2015] (CPMD/GROMACS interface) and GROMACS/CP2K interface patched with PLUMED: MD-QM/MM simulations employing the umbrella sampling protocol to determine the proton distribution between ammonium ions and phosphate groups

Exploitation of HPC resources

In total, we carried out 1.85 μ s of classical MD simulations of the DNA heptamer in aqueous solution and in a water droplet in the gas phase. The 500 ns MD simulation of the biomolecule in solution (39,714 atom) was carried out on 4 nodes (48 cores/node) with a performance of \sim 280 ns/day. The desolvation process of the DNA molecule in a water droplet was simulated with 3 replicas for each charge state (-4 , -3 and -2) in 300 x 500 ps batches on one node. Hereby, the performance strongly depends on the stage of the simulation. Initially, the water droplets contained \sim 11,200 atoms while at the end of the computational protocol \sim 330 atoms were left over, leading to a performance of \sim 30 and \sim 900 ns/day.

Ab initio MD simulations were carried out for three snapshots with large system sizes of 483, 438 and 333 QM atoms. We used the highly scalable CPMD program package, running on 32 nodes (48 cores/node). We observe a performance of 1.3–2.1 ps/day, with which we accumulated 22.7 ps simulations in total. This task required **500,000 core hours** of HPC resources on the SuperMUC-NG cluster at LRZ.

Intermolecular proton transfer phenomena between ammonium ions and DNA phosphate groups were studied for the two types of interactions with different degrees of solvation. In total, we determined 14 free energy profiles using umbrella sampling with 10 windows per profile. Each of the 140 independent runs required 25 ps simulations to give reasonably converged results. In view of the

high computational demand of ab initio MD calculations, this task had to be carried out through MD-QM/MM simulations, for which we used the CP2K/GROMACS interface developed in BioExcel-2. The QM regions comprised 23/28 and 41 atoms for the interaction of the ammonium ion with a single phosphate and two phosphate groups, respectively. The QM/MM calculations were done with higher precisions settings compared to the more demanding ab initio MD simulations (density cutoff of 500 Ry vs. 280 Ry), running on 2 nodes (48 cores/node). We observed a performance of ~3.5–6.0 ps/day dependent on the size of the QM region. This task required 3.6 ns QM/MM simulation time in total, for which we used **1,760,000 core hours** of HPC resources on the SuperMUC-NG cluster at LRZ.

Impact

Mass spectrometry can predict structural data of biomolecules at low resolution. Providing models by simulation has a great impact in structural biology. Indeed, mass spectrometry requires far less samples than any other higher resolution structural biology technique, from NMR to cryoEM, to X-ray crystallography. However, the interpretation of the data is non-trivial. One of the main difficulties in interpreting the data is the presence of quantum phenomena associated with proton transfer [Arcella et al. 2015] [Li et al. 2017]. This study provides novel and important information on such phenomena in nucleic acids and it will be of invaluable help to interpret mass spectrometry data on DNA. We will submit a paper for publication before the end of BioExcel-2 (June 2022). The results have been disseminated in a seminar at the Italian Institute of Technology (Genova, Italy – 24.08.2021) given by the PI of the project, Prof. Paolo Carloni.

References

- [Rueda et al. 2003] Manuel Rueda, Susana G Kalko, F Javier Luque, Modesto Orozco (2003):
The structure and dynamics of DNA in the gas phase.
Journal of the American Chemical Society **125**(26):8007–8014
<https://doi.org/10.1021/ja0300564>
- [Rueda et al. 2005] Manuel Rueda, F Javier Luque, Modesto Orozco (2005):
Nature of Minor-Groove Binders–DNA Complexes in the Gas Phase.
Journal of the American Chemical Society **127**(33):11690–11698
<https://doi.org/10.1021/ja0422110>
- [Arcella et al. 2012] Annalisa Arcella, Guillem Portella, Maria Luz Ruiz, Ramon Eritja, Marta Vilaseca, Valérie Gabelica, Modesto Orozco (2012):
Structure of Triplex DNA in the Gas Phase. *Journal of the American Chemical Society* **134**(15):6596–6606 <https://doi.org/10.1021/ja209786t>
- [Porrini et al. 2017] Massimiliano Porrini, Frédéric Rosu, Clémence Rabin, Leonardo Darré, Hansel Gómez, Modesto Orozco, Valérie Gabelica (2017):
Compaction of Duplex Nucleic Acids upon Native Electrospray Mass Spectrometry.
ACS Central Science **3**(5):454–461 <https://doi.org/10.1021/acscentsci.7b00084>

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

- [Arcella et al. 2015] Annalisa Arcella, Jens Dreyer, Emiliano Ippoliti, Ivan Ivani, Guillem Portella, D Valérie Gabelica, Prof. Paolo Carloni, Modesto Orozco (2015):
Structure and Dynamics of Oligonucleotides in the Gas Phase. *Angewandte Chemie* **54**(2) 467–471 <https://doi.org/10.1002/anie.201406910>
- [Olsen et al. 2019] Jógvan Magnus Haugaard Olsen, Viacheslav Bolnykh, Simone Meloni, Emiliano Ippoliti, Martin P. Bircher, Paolo Carloni, Ursula Rothlisberger (2019):
MiMiC: A Novel Framework for Multiscale Modeling in Computational Chemistry. *Journal of Chemical Theory and Computation* **15**(6):3810–3823
<https://doi.org/10.1021/acs.jctc.9b00093>
- [Abraham et al. 2015] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, Erik Lindahl (2015):
GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**:19–25
<https://doi.org/10.1016/j.softx.2015.06.001>
- [Li et al. 2017] Jinyu Li, Wenping Lyu, Giulia Rossetti, Albert Konijnenberg, Antonino Natalello, Emiliano Ippoliti, Modesto Orozco, Frank Sobott, Rita Grandori, Paolo Carloni (2017):
Proton Dynamics in Protein Mass Spectrometry. *The Journal of Physical Chemistry Letters* **8**(6):1105–1112 <https://doi.org/10.1021/acs.jpclett.7b00127>

Use Case 5 (pre-exascale showcase calculation): High-throughput Drug Screening

Nowadays drug design projects benefit from highly accurate protein-ligand binding free energy predictions based on molecular dynamics simulations. While such calculations have been computationally expensive in the past, we now demonstrate that workflows built on open source software packages can efficiently leverage pre-exascale compute resources to screen hundreds of compounds in a matter of days. In this use case we report our results of free energy calculations on a large set of pharmaceutically relevant targets assembled from industrial drug discovery projects.

Final Report

Setup

The overall project workflow is summarized in Figure 11. We initialize the procedure with the protein-ligand complexes provided with the publication of the benchmark set assembled in Merck KGaA [Schindler et al. 2020]. This step of system assembly and cleaning, followed by the careful modelling of ligands is highly important. Introducing ligand poses that do not reliably reflect actual ligand binding preferences would have severe consequences on the final free energy (ΔG) estimate accuracy. There are also numerous decisions required at this step: protein and ligand protonation states, protein starting structure selection, if needed, reconstruction of missing atoms, residues and various additional aspects. In the current work we started with this step readily accomplished by [Schindler et al. 2020] and continued our procedure with the topology generation.

This way, in the first step, for each of the considered complexes we created GROMACS compatible topologies for various force fields. Proteins were represented by means of AMBER99SB*ILDN and CHARMM36m. To parameterize the ligands we chose three different force fields: GAFF version 2.11, CGENFF v3.0.1 and OpenFF v1.2.0 Parsley. As at this step we did not employ high level quantum chemical calculations the step only takes several minutes per ligand. If a more elaborate parameterization is desired, it may become more time efficient to perform the computationally costly QM calculations on an in-house cluster or at an HPC facility.

Afterwards, in the second step of the procedure, we created hybrid structures and topologies for the ligand pairs using the PMX software [Gapsys et al. 2015]. This step is not computationally demanding and can be performed sequentially in a matter of minutes or hours even for a large set of perturbations. The generated hybrid structures were then assembled together with the protein structures and

a standard GROMACS procedure of system solvation and addition of salt was performed.

Up to this point, the prepared systems are agnostic to the specific free energy protocol, i.e. they can be used for the free energy perturbation (FEP), discrete, slow-growth or non-equilibrium thermodynamic integration or any other alchemical protocol of interest. Here, based on our experience in a previous investigation [Gapsys et al. 2020] we have chosen to use the non-equilibrium free energy calculation procedure.

As a preparatory step, we have performed an energy minimization and a brief equilibration of the system for 100 ps (step 3 in Figure 11) on an in-house cluster. In principle this step could be merged with the following main calculation performed on the HPC Supercomputer. For the current project, however, we decided to carry out initial short simulations on the in-house compute cluster. This way we ensured that the prepared systems were stable and ready to be transferred to the HPC Supercomputer Raven for the further free energy calculations. Since in an everyday application this step would be a part of the next step (step 4 in Figure 11), its timing is of no particular importance, as it constitutes only a minor fraction of the full free energy computation.

The fourth step in Figure 11 is the main point of the computations in this use case highlighting the scaling capabilities for such calculations. While the GROMACS simulation engine itself offers high throughput in terms of generated trajectory time, the employed free energy calculation protocol further allows for essentially perfectly (or near-perfectly) efficient parallelization of the jobs.

In the final step, the generated output was transferred from the HPC facility and analysed on the local workstation by means of the PMX software. The accuracies of the predicted free energies were further explored by comparison to experiment and previous calculations.

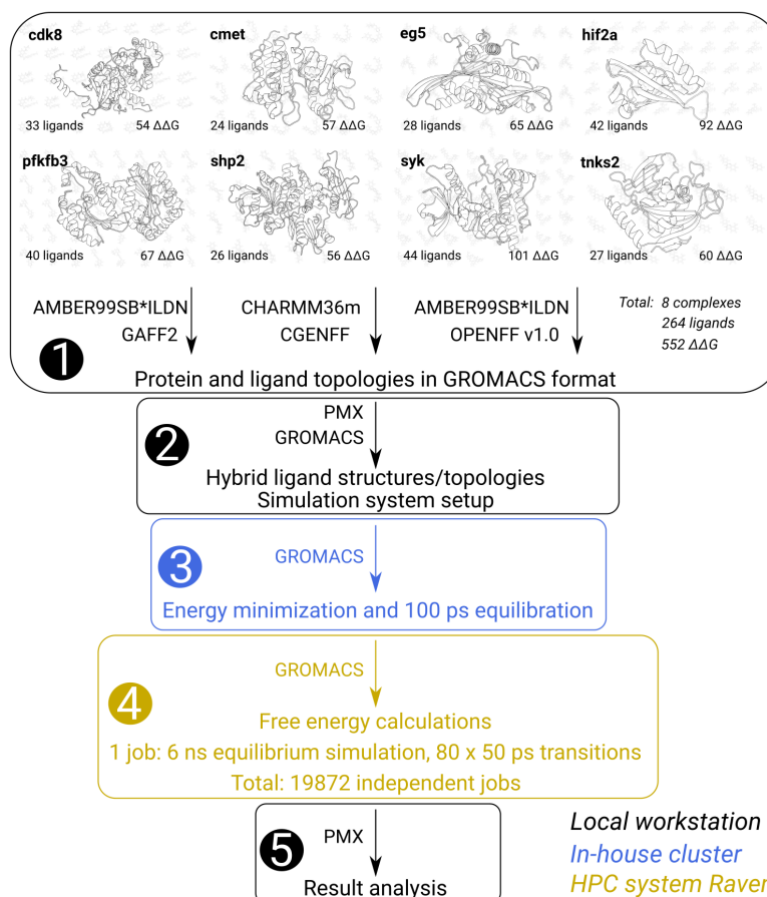


Figure 11. (1) We start with the protein-ligand complexes prepared and made public in [Schindler et al. 2020]. The protein topologies are prepared in AMBER99SB*ILDN and CHARMM36m force fields, while for the ligands we used GAFF 2.11, CGENFF 3.0.1 and OpenFF 1.2.0 force fields. (2) Hybrid ligand structures and topologies for alchemical calculations are created with the PMX software and further the systems are assembled and prepared for the simulations with GROMACS. (3) Energy minimization and a brief 100 ps equilibration was performed on an in-house cluster. For the further automation of the workflow, this procedure could be merged with the following step and run in an HPC facility. (4) The main step where the simulations were performed on the Raven Supercomputer. We were able to parallelize the calculations with perfect (or near-perfect) efficiency by dividing the whole set into individual jobs as detailed in the text. (5) After finalizing the simulations on the HPC machine, the data was retrieved and analyzed locally.

Results: calculation accuracy

Overall, the calculation accuracy matches well our earlier observations for a different protein-ligand benchmark set [Gapsys et al. 2020]. Relying on the earlier experience [Gapsys et al. 2020], we have constructed the consensus approach combining results from two different family force fields - GAFF and CGENFF. In turn, this yields better accuracy in terms of agreement with the experimentally measured values than the force fields considered individually when comparing

predicted $\Delta\Delta G$ with experimental measurements (Figure 12). The consensus calculations (AUE 1.11 ± 0.5 kcal/mol, Pearson correlation 0.59 ± 0.04) also approach the performance of the commercial software FEP+ (AUE 1.06 ± 0.04 , Pearson correlation 0.66 ± 0.03).

Individually, GAFF 2.11 and Openff v.1.2.0 achieved comparable accuracy, while CGENFF performed slightly worse. It is possible that the results obtained with the CGENFF force field could be further improved by employing a newer force field version, as currently we relied on an older parameter set (3.0.1). The situation of the OpenFF force field might be similar to CGENFF: here we have benchmarked an early version (v1.2.0 Parsley) of the force field. At the time when the calculations were performed, this OpenFF version had not yet undergone Lennard-Jones parameter reparameterization and other fixes. Recently, OpenFF v2.0 has been released and some preliminary calculations indicate its improved accuracy in predictions of differences between free energies ($\Delta\Delta G$). Therefore, in the future it would be interesting to probe how much the accuracy would improve by employing the updated force field versions.

In the bottom panels of Figure 12 we show the breakdown of the calculated $\Delta\Delta G$ values by protein-ligand complex. The performance of the individual force fields depends on the system simulated and is often strongly influenced by large outliers, e.g. the overall well behaved GAFF force field shows a reduced accuracy for the shp2 complex mainly due to a few poor predictions. The consensus approach often suppresses the largest deviations from the experimental measurements. Modelling of the initial ligand pose also plays an important role for the result accuracy. For example, for the cmet protein-ligand complex, Schindler et al. (2020) reported the results after probing several modelled poses (personal communication). In the current work we used a single pose which in some cases was suboptimal for the cmet system, in turn yielding more outliers and lowering prediction accuracy.

As it was not the main aim of the current use case to investigate all the particular details of the predicted $\Delta\Delta G$ values and their force field dependence, we are continuing our work on the in depth analysis of the data. In the next step, we are going to incorporate the data generated in this scan into a larger benchmark study comprising protein-ligand complexes assembled from numerous benchmark sets and compare free energy predictions from multiple force fields and their different versions.

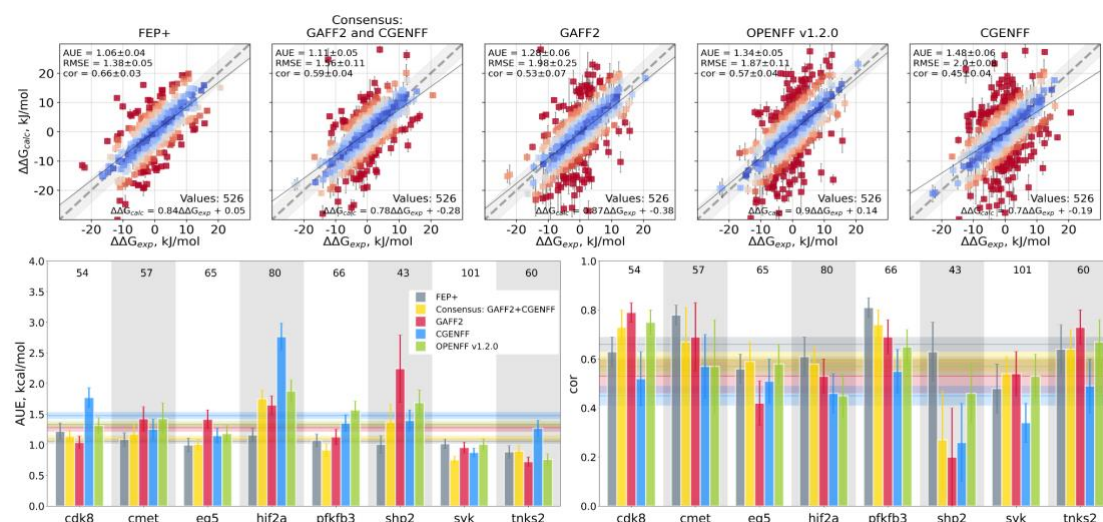


Figure 12. Comparison of the computed $\Delta\Delta G$ values to the experimental measurements. Top row: scatter plots of all the values that have been reported in [Schindler et al. 2020] computed with the commercial software FEP+. In adherence with the best practice, the directions of the $\Delta\Delta G$ edges have been retained exactly the same as in [Schindler et al. 2020]. The first panel reports FEP+ results by Schindler et al. (2020), while the other panels present results from the current work. Bottom row: average unsigned error (AUE) and Pearson correlation (cor) for each protein-ligand complex separately. The horizontal lines denote mean values. The numbers in the panels report on the free energy differences calculated for each system.

References

[Schindler et al. 2020] Christina E. M. Schindler, Hannah Baumann, Andreas Blum, Dietrich Böse, Hans-Peter Buchstaller, Lars Burgdorf, Daniel Cappel, Eugene Chekler, Paul Czodrowski, Dieter Dorsch, Merveille K. I. Eguida, Bruce Follows, Thomas Fuchß, Ulrich Grädler, Jakub Gunera, Theresa Johnson, Catherine Jorand Lebrun, Srinivasa Karra, Markus Klein, Tim Knehans, Lisa Koetzner, Mireille Krier, Matthias Leiendecker, Birgitta Leuthner, Liwei Li, Igor Mochalkin, Djordje Musil, Constantin Neagu, Friedrich Rippmann, Kai Schiemann, Robert Schulz, Thomas Steinbrecher, Eva-Maria Tanzer, Andrea Unzue Lopez, Arielle Viacava Follis, Ansgar Wegener, and Daniel Kuhn. [Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects](https://doi.org/10.1021/acs.jcim.0c00900). *J. Chem. Inf. Model.*, **60**(11):5457–5474, 2020 <https://doi.org/10.1021/acs.jcim.0c00900> [preprint available]

[Gapsys et al. 2015] Vytautas Gapsys, Servaas Michielssens, Daniel Seeliger, and Bert L. de Groot. Pmx: [Automated protein structure and topology generation for alchemical perturbations](https://doi.org/10.1002/jcc.23804). *J. Comput. Chem.*, **36**(5):348–354, 2015. <https://doi.org/10.1002/jcc.23804>

[Gapsys et al. 2020] Vytautas Gapsys, Laura Pérez-Benito, Matteo Aldeghi, Daniel Seeliger, Herman van Vlijmen, Gary Tresadern, and Bert L. de Groot. [Large scale relative protein ligand binding affinities using non-equilibrium alchemy](https://doi.org/10.1039/C9SC03754C). *Chem. Sci.*, **11**(4):1140–1152, 2020. <https://doi.org/10.1039/C9SC03754C>

Software

The whole procedure of the large scale scan was performed with open source software, including core BioExcel applications.

Simulation system preparation and all the simulations were performed with the GROMACS package. Open source software was used to automate ligand topology preparation: AmberTools, ACPYPE, MATCH, OpenFF toolkit, Parmed.

Ligand hybrid structures and topologies for the alchemical free energy calculations were prepared with the PMX package. PMX was also used for the result analysis.

Exploitation of HPC resources

The overall scan comprised 19872 independent jobs: 552 $\Delta\Delta G$ calculations in 3 force fields for 2 thermodynamic branches (water, protein-ligand) each of which requires 2 simulations (one for forward and one for backward direction) and 3 independent replicas for each calculation. In total, this sums up to ~200 microseconds of simulation trajectory generated in the scan.

The work was accomplished in **72 hours**, leveraging resources allocated during the pre-production phase of the Max Planck Computing & Data Facility's Raven Supercomputer, enabling simultaneous usage of **480 nodes** hosting Intel Xeon Cascade Lake-AP processors with 96 cores (192 threads) each, corresponding to **93% of the entire machine**. In total, this amounted to **~3.4 million core hours** of compute time.

This division of simulations into separate jobs was dictated by the available resources and could be easily modified to match a specific HPC architecture. For example, having access to a particularly large compute facility one could further separate every short 50 ps transition into an individual job, enabling the execution of ~1.6 million small jobs in parallel, thus further reducing the waiting time to prediction.

Impact

In this showcase use case we highlight that rapid high throughput sampling of protein-ligand binding affinities is readily achievable using open source BioExcel-developed and supported software. Provided that sufficient computational resources are available, large scale alchemical protein-ligand binding free energy predictions can be efficiently run solely relying on open source software in a routine fashion to guide drug discovery projects. Screening hundreds of derivatives of an initial hit or lead compound can be achieved in a matter of days while obtaining the high accuracy of alchemical free energy calculations. Our results show how the accuracy of prediction versus experiment differs with each force field for the same free energy calculation approach.

Publications

- Gapsys, V.; Hahn, D. F.; Tresadern, G.; Mobley, D. L.; Rampp, M.; de Groot, B. L. Pre-exascale computing of protein-ligand binding free energies with open source software for drug design. *Journal of Chemical Information and Modeling*, 2022, in press
- Hahn, D. F.; and Bayly, C. I.; Bruce Macdonald, H. E.; Chodera, J. D.; Gapsys, V.; Mey, A. S. J. S.; Mobley D. L.; Perez Benito, L.; Schindler C. E. M.; Tresadern, G.; Warren, G. L. [Best practices for constructing, preparing, and evaluating protein-ligand binding affinity benchmarks](#). *arXiv 2105.06222*, 2021

Other

- An illustrative workflow/tutorial for preparing and running free energy calculations: [GitHub - pmx tutorials](#)
- Data set with ligand and protein structures, topologies, and calculated relative free energy values for the Merck protein-ligand set: [GitHub - rel ddG MerckDataSet ICIM](#)

Presentations

- AstraZeneca: Computational Chemistry seminar. July 6, 2021. Remote. "Computational Alchemy and Absolute Protein-Ligand Binding Free Energies for Drug Design".

Appendix: Final status with reference to work plans

In this appendix we summarise the final status of the planned Use Cases with explicit reference to the activities that make up the work plans described in deliverable [D3.1 - Use Case Work Plans](#).

Use Case 1

Table 3: Use Case 1 Work Plan

ID	Activity	Final Status
A1.1	Prepare antibody/antigen model for all-atom simulation	<p>CHARMM36m has been considered for the molecular dynamics simulations. The approach can also be extended to other biomolecular force fields to avoid bias due to the choice of a specific set of parameters to describe the molecular system.</p> <p>Structures of apo and/or bound forms for 15 antibodies (PDB codes: 1BJ1, 1KTZ, 1VFB, 1DQJ, 2W9E, 3EO1, 3HI6, 3G6D, 3V6Z, 4DN4, 1E6J, 1JPS, 1NSN, 4G6J, 6W41, 3HMX, 3L5W, 3RVW) have been pre-processed.</p> <p>Both experimental structures and docking models were considered. During pre-processing hydrogen atoms are added to the structure, here special attention has been given to the hydrogen position to protonable residues at the interface (in some cases manual assignment has been necessary). Simulation parameters files compatible with the force fields have been defined and shared with UseCase 3 (partner KTH).</p>
A1.2	Initial sampling of antibody conformations before docking	<p>100 ns data production has been used to initially sample the antibody conformations. For the final protocols, we focus on 11 antibodies (see A1.10) for which both the apo and holo structures were experimentally available. Representative</p>

		antibody conformations have been extracted using a RMSD-based cluster analysis and used as input in the docking procedure. Two approaches for loop clustering have been used. We also performed 100 ns MD data production for antigen structures. Note this was not in the original plan, but we have decided to include MD sampled structure for antigen to have consistency (structure generated with the same force field) among the structures used in A1.8. (partner KTH)
A1.3	Employ advanced sampling schemes for sampling of antibody conformations before docking	For antibody 3V6Z (that presents a loop rearrangement upon antigen binding) extra 0.5 microseconds have been considered. Based on these results, we use accelerated weight histogram (AWH) to enhance the sampling of the antibody loops at the antibody-antigen interface (partner KTH). Distance between the antibody and antigen c.o.m. was used as reaction coordinate, and the system was sampled for 100 ns.
A1.4	Refinement of statically modelled antibody/antigen interactions through Molecular Dynamics simulation	HADDOCK models of antibody-antigen complexes for all the selected antibodies have been refined with a series of 100 ns molecular dynamics simulations. MD simulations have been performed for a representative set of antibody-antigen models generated by HADDOCK using experimental and simulated (see A1.2 and A1.3) antibody/antigen structures as input. In cases where large conformational changes are experimentally observed at the interface, as 3V6Z, advanced sampling techniques such as AWH have been applied to enhance the conformational sampling of the antibody/antigen interface. Simulations are followed by analysis of the interaction patterns and interface quality. (partners KTH/UU)
A1.5	Automate and document molecular dynamics simulation setup, production run and analysis	The first steps (definition of the parameters, tools, and setting) toward the automatisation of the pre-processing phase/ initial MD sampling/ post-processing have been taken and collected in the form of a Jupyter notebook. Issues like

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

		hydrogen atom assignments and post processing strategy/setting to select the most representative conformers are not yet resolved. Simulation parameters files compatible with the force fields have been documented and shared with UseCase 3 . (partner KTH)
A1.6	Benchmarking alchemical free energy calculations for antibody-antigen stabilization	The effort from the task was largely redirected to Use Case 5. We have benchmarked the quality of the pmx workflow based on the Rosetta FlexDDG protocol by calculating changes in free energy differences upon amino acid mutations. Analysis of the calculated changes in binding affinity reveals that the Rosetta FlexDDG protocol is able to accurately predict changes in binding affinity, however, sensitivity of the method is system dependent and further MD based calculations may be necessary depending on the antibody-antigen system being studied.
A1.7	Optimizing antibody-antigen interactions by alchemical free energy calculations after docking and MD sampling	The effort from the task was largely redirected to Use Case 5. For the current case, we evaluated the quality of HADDOCK docked models by means of pmx workflow based on the Rosetta FlexDDG protocol. Comparison of the calculated changes in binding affinity to the experimental measurements hint that predictions based on the HADDOCK models are only slightly less accurate than those relying on the crystallographic structures.
A1.8	Defining the baseline performance of HADDOCK for antibody-antigen modelling under different scenarios	HADDOCK's MD stage can well account for the dynamic nature of CDRs. Despite finding at least acceptable quality models in most cases, correct model ranking seems to be problematic. Thus MD is recommended to distinguish native from non-native models. (UU)
A1.9	Improved modelling of antibody-antigen modelling by using antibody conformations sampled by MD	Preliminary benchmarking of different scenarios for antibody-antigen docking with HADDOCK using antibody and antigen conformations sampled by molecular dynamics simulation have been done to measure if these lead to improved docking results. It was observed that antigen sampling has a minor effect on results, mostly probably due to the rigid nature of the antigen protein. Thus we use only MD sampled

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

		conformations for the antibody in our final protocols. On average 20 conformers have been selected from 100ns-MD sampling simulations for each of the 11 antibodies. The crystallographic structure of the apo has been added to the pool of structure before docking (partner UU/KTH)
A1.10	Demonstration of antibody-antigen modelling on real experimental data	To validate the final MD-docking protocol, we have used 11 antibody-antigen complexes (1DQJ, 2W9E, 3E01, 3G6D, 3HI6, 3HMX, 3L5W, 3RVW, 3V6Z, 4DN4, 4G6J) from Docking Benchmark 5 and Affinity Benchmark 2, for which experimental structures were available both for the apo (input) and holo structure (reference). We demonstrate that depending on a particular antibody-antigen system, accurate predictions of binding affinity changes upon amino acid mutation are readily feasible. We delivered the MD-docking protocol in the form of a Jupyter notebook, which will allow the community to use it in their research projects.

Use Case 2

Table 4: Use Case 2 Work Plan

	Activity	Final Status
A2.1	Reliability analysis of protein-protein interaction energy prediction through all-atom molecular dynamics simulation	A proof of concept experiment was designed and conducted to determine if it is possible to determine true-positive from true negative interactions. The results showed that such determination is possible (see deliverable D3.3 - Use Case Progress Report).
A2.2	Distinguishing native from non-native docking models by MD	Done and published Z. Jandova, A.V. Vargiu and A.M.J.J. Bonvin. Native or non-native protein-protein docking models? Molecular dynamics to the rescue . <i>J. Chem. Theo. and Comp.</i> 17, 5944–5954 (2021).
A2.3	Investigating suitable mechanism of launching a large number of docking simulations as single job on HPC resources	Done – two solutions offered (via PyCOMPs and the haddock-pilot machinery)
A2.4	Optimize the identification of near-native docking models using machine learning methods	<p>Done and published (related to A2.2). In addition, a deep learning model was developed in a related project and recently published</p> <p>Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M. F., Bonvin, A. M. J. J. & Xue, L. C. DeepRank: a deep learning framework for data mining 3D protein-protein interfaces. <i>Nat Commun</i> 12, (2021).</p> <p>M.F. Réau, N.Renaud, L.C. Xue and A.M.J.J. Bonvin. DeepRank-GNN: A Graph Neural Network Framework to Learn Patterns in Protein-Protein Interfaces. <i>BioRxiv</i> 10.1101/2021.12.08.471762 (2021)</p>

Use Case 3

Table 5: Use Case 3 Work Plan

	Activity	Final Status
A3.1	Cancer mutations filtering & identification of pathological mutations	Done (WF1). A manually curated final dataset of 23 interesting pathological mutations were chosen for the massive free energy study.
A3.2	Conformational ensembles from protein variants	Done (WF1, WF2, WF3). Ensembles are extracted from equilibrium MD simulations and used as input for the free energy (thermodynamic integration) calculations.
A3.3	Conformational ensembles from small molecules	Workflow up and running, but finally not used in this use case.
A3.4	Ensemble Docking	Done (WF1). Best poses chosen for two of the ligands studied (osimertinib, icotinib).
A3.5	Alchemical free energy calculations of relative ligand binding free energy differences	Done (WF1, WF2, WF3). Relative binding free energy difference upon protein residue mutation workflow up and running, prepared for HPC execution.
A3.6	Molecular Dynamics trajectories analyses	Done (WF1, WF2, WF3, WF4). Workflows up and running. Quality control analyses (WF1, WF2, WF3) and DNA-specific analyses, including helical parameters (WF4).
A3.7	Molecular Dynamics trajectories storage	Effort moved towards the https://bioexcel-cv19.bsc.es/ database and associated web server. The core infrastructure is being extended and generalized to deposit any kind of MD trajectory, and in particular, the use case 3 generated MD trajectories.
A3.8	Peptide Mutational Scans	Done (WF3). Workflows up and running simulating an Alanine Scanning.
A3.9	High Performance Data Analytics	Done (WF3). Workflows up and running extracting flexibility parameters from a massive amount of DNA trajectories in HPC resources.
A3.10	Enabling advanced use of	Done (WF1, WF2, WF3). Workflows developed are

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

	GROMACS, improve usability & implement advanced functionality within the use case	available through the BioExcel GitHub repository with instructions on how to easily deploy a Conda Pack to reproduce the calculated values. https://github.com/bioexcel/biobb_hpc_workflows/tree/condapack
--	--	--

Use Case 4a**Table 6: Use Case 4a Work Plan**

	Activity	Final Status
A4.1	Inquiring needs and document interface to apply forces from QM simulations during molecular dynamics simulation	On the basis of performed QM/MM simulations with CP2K a set of parameters for reliable QM/MM modelling of fluorescent proteins has been determined. That parameters have been embedded into the GROMACS-CP2K interface, developed within WP1
A4.2	Prepare fluorescent proteins and mutants models for simulation	5 models of avGFP protein mutants and 6 models of rsGreen0.7 protein mutants have been prepared for simulation of folding and dimerisation free-energies
A4.3	Extensive sampling of fluorescent proteins conformations using classical molecular dynamics	Sampling for up to 100ns have been performed for each fluorescent protein and their mutants using GROMACS
A4.4	Free energy simulations of folding and dimerization of different fluorescent protein mutants.	For each system, the unfolded, folded-monomeric and folded-dimeric states were evaluated to compute the folding and dimerization free energies, accumulating up to 630 ns per model
A4.5	Calculation of isomerization profiles of chromophores in protein environments	QM/MM umbrella sampling simulations of the isomerization of phytochrome fluorescent protein performed in total for 0.8 ns of dynamics over 16 individual windows that were simulated in parallel.
A4.6	Modelling of spectral properties of fluorescent proteins and their mutants	The absorption and emission spectra computed for 5 avGFP variants using hybrid QM/MM and MD sampling approach replicates the stroke-shift behavior of the experimentally measured data.
A4.7	Supporting massively parallel QM/MM simulations of fluorescent proteins	We have performed preliminary benchmarks with ensemble-based methods of free energy profiles for the chromophore isomerisation in the proteins. Both Umbrella sampling and AWH methods showed a good scaling and could be parallelized over a large number of nodes.

Use Case 4b

As mentioned in the [UC4b summary](#) in this document, the choice was made to study proton transfer in mass spectrometry of a DNA oligo instead of the protein study originally planned. This was done because experimental data regarding the main charge of this DNA oligo complex in the gas phase became available. Not only did this enable comparison of simulation with experiment, it accelerated the investigation relative to the system proposed originally, whilst retaining the goal of novelty and innovation with regards to the use of QM/MM simulation with BioExcel codes for this type of study.

Table 7: Use Case 4b Work Plan

	Activity	Final Status
A4.7	Prepare the DNA heptamer models for molecular dynamics simulations in water²	The DNA heptamer with ammonium counterions was set up in aqueous solution and simulated classically for 500 ns. The desolvation of the oligonucleotide in a water droplet was simulated for three total charge states and analysed to obtain configurations suitable for proton transfer.
A4.8	QM and QM/MM investigations of proton transport between ammonium and the oligonucleotide: calculations of structural properties³	<i>Ab initio</i> molecular dynamics simulations were carried out for the oligonucleotide-ammonium complex at three different degrees of solvation, sampling in total for 22.7 ps. QM/MM simulations for 14 configurations were done in preparation for the free energy calculations.
A4.9	QM and QM/MM investigation of proton transport between ammonium and the oligonucleotide: free energy calculations⁴	QM/MM umbrella sampling simulations of the proton transfer between ammonium and the oligonucleotide were carried out for 14 free energy profiles, each consisted 10 individual windows of 25 ps simulation time that were simulated in parallel.

² Replaces “Prepare the beta-lactoglobulin protein models for molecular dynamics simulations”

³ Replaces “QM and QM/MM investigation of proton transport in beta-lactoglobulin protein by CPMD and MiMiC interface”

⁴ Replaces “QM and QM/MM investigation of proton transport in beta-lactoglobulin protein by CP2K and CP2K/GROMACS QM/MM interface”

D3.6 – Pre-Exascale showcase calculation and Use Case Final Report

A4.10	Writing of the paper summarizing the research A4.7 - A4.9⁵	in progress as part of the no-cost BioExcel-2 extension period. It is expected that the postdoc, along with the FZJ and IRB teams, will complete the paper by the end of the BioExcel-2 project extension (June 2022).
-------	--	--

⁵ Replaces “Supporting massively parallel QM/MM simulations of protein dynamics”