INFRAEDI-02-2018

www.bioexcel.eu

BioExcel-2 Project Number 823830

D4.4 – Final report on dissemination and training

WP4: Dissemination and Training

Document Information

Deliverable Number	D4.4
Deliverable Name	Deliverable Name
Due Date	2022-04-30 (PM40)
Deliverable Lead	KTH
Authors	Michelle Mendonca (EMBL-EBI), Dana Cernoskova (EMBL-EBI), Marta Lloret Llinares (EMBL-EBI), Victoria E. Hill (EMBL-EBI), Alessandra Villa (KTH), Adam Hospital (IRB), Vera Matser (EMBL-EBI)
Keywords	Project Management
WP	WP4
Nature	Report
Dissemination Level	Public
Final Version Date	2022-04-25
Reviewed by	Dmitry Morozov (JYU), Emiliano Ippoliti (JUELICH), Richard Norman (NC)
MGT Board Approval	2022-04-28

Document History

Partner	Date	Comments	Version
EMBL-EBI	2022-03-24	First draft sent to the consortium for internal review	0.1
EMBL-EBI	2022-04-25	Version sent to MGT board for approval	0.2

Executive Summary

This document describes the dissemination and training activities that have been completed since July 2020, including an estimate of the impact that they had. The dissemination section presents information about the project website, social media channels, dissemination events and publications, as well as a report on the EMBO Workshop *Advances and Challenges in Biomolecular Simulations* organised in October 2021, the BioExcel flagship event. The training section presents the events organised and delivered in this period, the last updates in the BioExcel competency framework and a long-term impact assessment.

The EMBO workshop *Advances and Challenges in Biomolecular Simulations* was successfully delivered in a virtual format in October 2021, after being postponed due to the COVID-19 pandemic. The event included 30 speakers, both invited and selected from abstracts, and 98 posters. It was attended by 173 participants from 16 countries worldwide, who participated in lectures and networking sessions. We observed a very low dropout rate over the four days of the event. According to the results of the event feedback survey, the overall evaluation of the meeting was: 40.71% 'excellent', 45.13% 'very good', and 12.39% 'good'.

The BioExcel social media channels have seen an upward growth since June 2020, with 3000 new followers on Twitter and 1300 on LinkedIn. The BioExcel website has received over 54000 unique users. BioExcel has participated in a total of 121 events, where we estimate that we have reached a minimum of 8172 people. In addition, BioExcel partners have published 39 articles.

BioExcel converted the 2020 Summer School on Biomolecular Simulations to a remote event, thanks to the experience that we gained running virtual courses since 2019. Since then, we have run 3 more editions of the event, with improvements in each of them based on the experienced gained and the feedback from trainers and participants. In addition, we have run 17 training events, many of them in collaboration with other initiatives such as PRACE, other CoEs or National Competence Centres. The feedback received is very positive. All the events in this period were remote, but we plan to run a face-to-face Summer School in Sardinia in June 2022.

The BioExcel competency profile has been updated in collaboration with PerMedCoE, so that the computing and parallel computing competencies of both profiles are common. We created six career profiles that are available on the EMBL-EBI Competency Hub. Two learning pathways, collections of training resources, have been created to guide learners in the use of HPC machines and Jupyter notebooks.

The long-term impact of previous training activities shows that most respondents use the skills gained during the courses and many of them teach others, which broadens the impact that our training has in the community.

Contents

1. INTRODUCTION 6

2. DISSEMINATION 6

2. 1. General dissemination 6
2.1.1. Website 6
2.1.2. Newsletters 7
2.1.3. Social Media Channels 7
2.2. COVID-19 dissemination 9
2.3. Events and Publications 9
2.4. EMBO Workshop Advances and Challenges in Biomolecular Simulations 28
2.5. Pandemic 28
2.5.1. Original plans 29
2.5.2. Myia conference environment 29
2.6. Programme 31
2.6.1. High level overview of topics, speakers (abstract & invited), poster, social 32
2.6.2. Poster session 32
2.6.3. Networking 33
2.7. Statistics 34
2.8. Impact 35
2.8.1. Social media/communication 35

3. TRAINING 38

3.1. Events 39
3.1.1. BioExcel School on Biomolecular Simulations 43
3.1.2. Practical introduction to QM/MM using CP2K for biomolecular modelling 13 & 20 October 2020 48
3.1.3. QM/MM simulation with GROMACS and CP2K 22 & 23 April 2021 48
3.1.4. Computational biomolecular simulation workflows with BioExcel building blocks 49
3.1.5. Introduction to HPC for Life Scientists 51
3.1.6. Remote GROMACS courses 52
3.1.7. EMBO practical course - Integrative modelling of biomolecular interactions. 30 May - 5 June 2021 54
3.1.8. PRACE Autumn School: Fundamentals of Biomolecular Simulations and Virtual Drug Developments 20 - 24 September 2021 55
3.1.9. Training events for industry 56
3.1.10. Future training events 56
3.2. Impact assessment 56
3.2.1. Long-term feedback survey 6-12 months 58
3.2.2. Long-term feedback survey 2 years 60
3.3. Competency Hub 62
3.3.1. Competency profile version 3 62
3.3.2. Career profiles 64
3.3.3. Learning pathways 64

4. CONCLUSION 66

ANNEX I - THE BIOEXCEL COMPETENCY PROFILE V3 68

1. Introduction

This document follows the previous deliverable D4.3 – *Progress report and update on training and dissemination plan* and includes the dissemination and training activities carried out since June 2020.

Dissemination and training activities have the aim to raise awareness about the services and expertise within BioExcel among stakeholders and to increase the competence of end-users of the BioExcel supported software.

This period has still been affected by the COVID-19 pandemic, which is reflected in this document. We continued to deliver a remote training programme and we organised virtually the EMBO Workshop *Advances and Challenges in Biomolecular Simulations*, the BioExcel flagship event, which was initially postponed from its original date in October 2020.

2. Dissemination

Here we present an update on the dissemination strategy for BioExcel implemented since June 2020 using our digital marketing channels and the respective impact.

2. 1. General dissemination

2.1.1. Website

The BioExcel website (bioexcel.eu) acts as a main channel of dissemination. Through blog posts and relevant events, the website acts as a central point for BioExcel services and latest research in the sector. The website also hosts the newsletter which manages a steady stream of incoming users. A live feed of the BioExcel Twitter account is seen on the landing page.

Since June 2020, BioExcel has received over 54,000 unique users. The website has seen visitors from 175 different countries, with the major visiting countries being the United States, India, China and the United Kingdom. (Fig. 1)

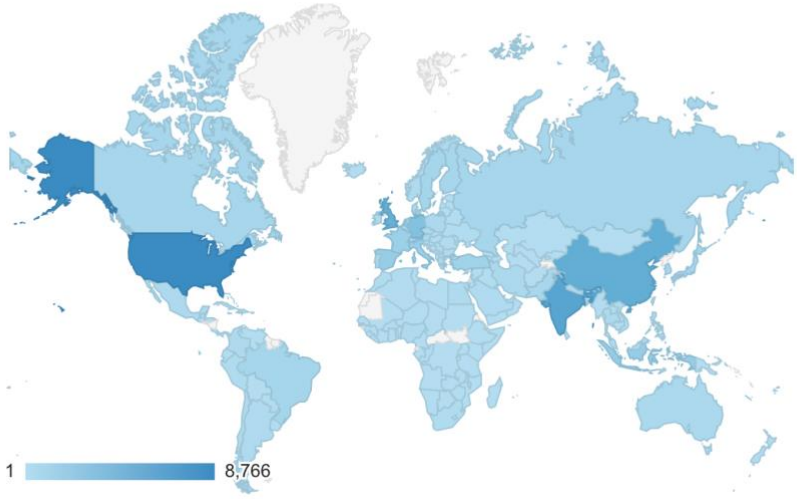


Figure 1. Google analytics with demographics

Acquisition analytics identify that the major source of traffic derives from organic web searches, followed by direct and social media referrals, for which BioExcel's Twitter channel remains dominant (Fig. 2).

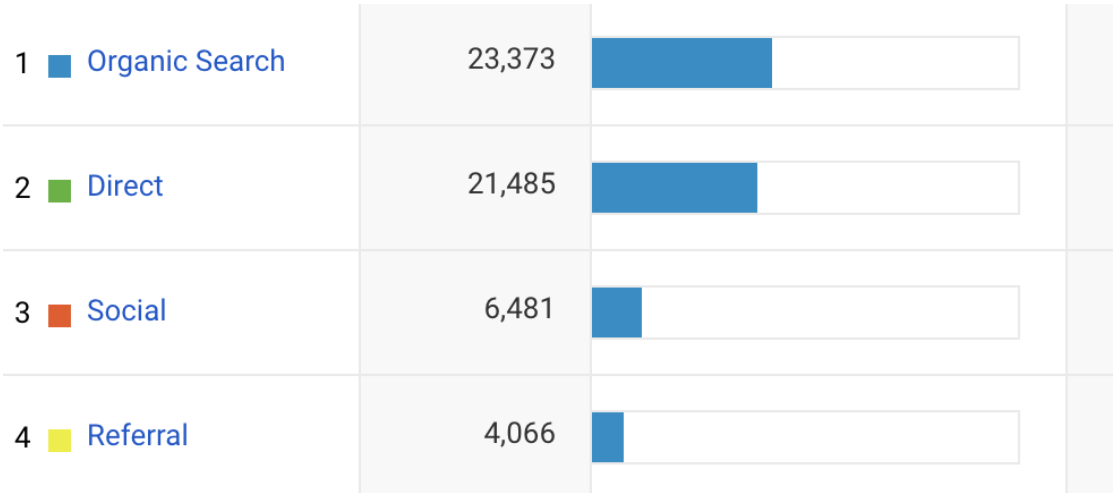


Figure 2. Google analytics with acquisition traffic

The most viewed content on the website is the BioExcel Summer and Winter School on Biomolecular Simulations followed by webinar related content (Fig. 3).

	Page	Pageviews	% Pageviews
1.	/	25,259	18.41%
2.	/events/summer-school-on-biomolecular-simulations/	6,610	4.82%
3.	/events/winter-school/	4,405	3.21%
4.	/category/webinar/	3,114	2.27%
5.	/events/bioexcel-school-on-biomolecular-simulations/	2,929	2.13%
6.	/research/projects/rational-drug-design/	2,235	1.63%

Figure 3. Google analytics with most viewed pages

2.1.2. Newsletters

Newsletters remain an integral opportunity to demonstrate expertise on the subject matter and develop trust with users. The BioExcel newsletter currently has over 2,000 subscribers which captures an audience not active on social media and that uses newsletters as their primary source of information.

Analytics show that with an average open rate of 30% and click rate 6%, we are meeting industry standards (As per campaign monitor's resource article; 15-25% open rate and 2.5% click-through rate).

2.1.3. Social Media Channels

Twitter - The BioExcel twitter channel has shown an upward growth with over 3000 followers. An example of a successful tweet advertising the EMBO Workshop

on Advances and Challenges in Biomolecular Simulations is shown below, which gained 8000 impressions. (Fig. 4)

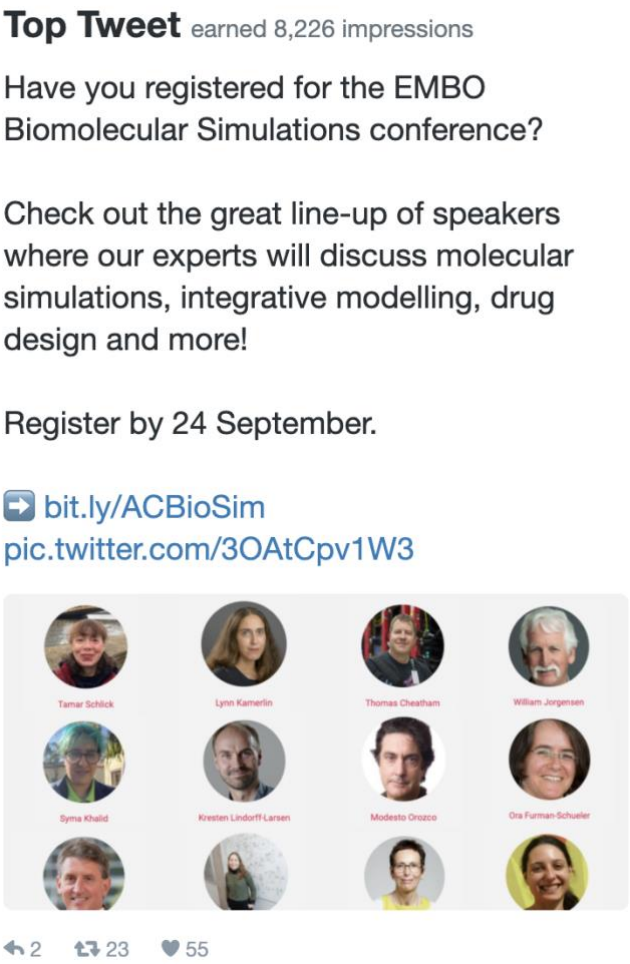


Figure 4. Top performing tweet

LinkedIn - The LinkedIn channel has seen an upward growth gaining over 1300 followers (A successful example of a post advertising the BioExcel Summer School with 1000 impressions is shown below (Fig. 5).

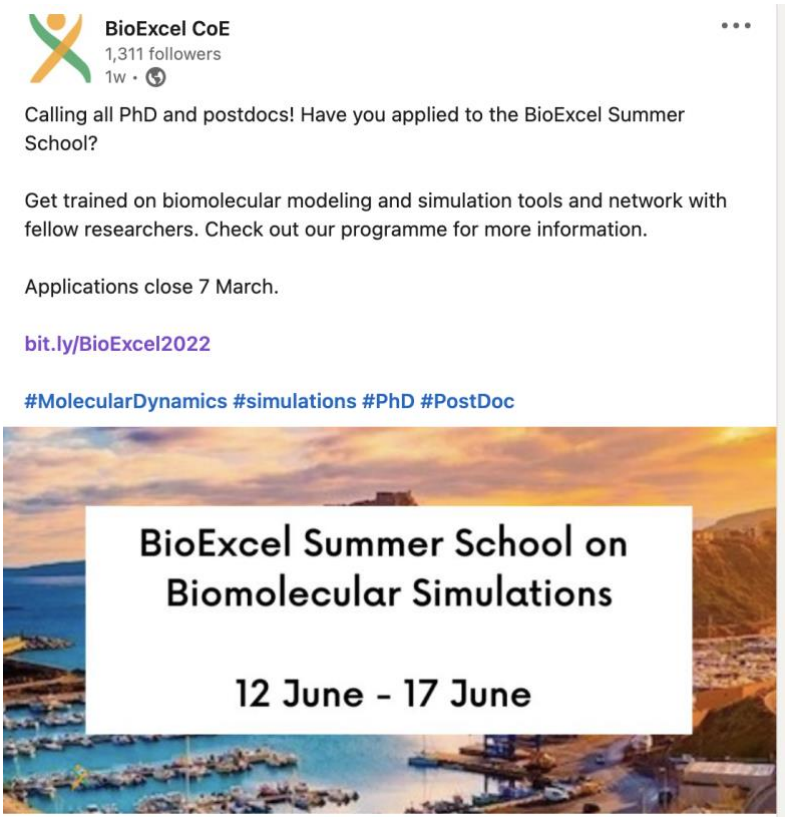


Figure 5. Top performing LinkedIn post

2.2. COVID-19 dissemination

BioExcel partners have launched a series of actions to support research into SARS-CoV-2 since the beginning of the COVID-19 crisis. Detailed information on COVID-19 work and dissemination can be accessed through the deliverables D6.3 - *High Performance Computing in support of COVID-19 Research* and D4.3 - *Progress report and update on training and dissemination plan* (under section 2.1 Dissemination during COVID-19).

2.3. Events and Publications

The following training and dissemination events shown in Table 1, 2 and 3 were carried out by BioExcel members from June 2020 to February 2022. We held a total of 121 events, with 93 classified as dissemination events and 36 classified as training events. A conservative estimate shows we have reached a minimum of 8172 people during this period.

Table 1. 2020 Training and Dissemination Events

Date	Event name	Location	Number of Participants
July 6-10 2020	SciPy 2020	Online	
July 7 2020	EOSC-Life Transferring face-to-face courses to remote delivery workshop	Online	65
July 10 2020	1st EU-ASEAN webinar on HPC-COVID19 related modeling and diagnostic	Online	40
July 13-16 2020	ISMB 2020	Online	
July 16 2020	CECAM Webinars: The importance of being H.P.C. Earnest	Online	
August 18-21 2020	Managing a bioinformatics core facility	Online	16
September 3 - 4 2020	Introduction to GROMACS (PRACE BioExcel)	Online	54
September 4 2020	Advances in computational modelling of cellular processes and high-performance computing	Online	110
September 16-17 2020	RES Users conference	Online	
September 10 2020	BioExcel Webinar #48: Computational biomolecular simulation workflows with BioExcel building blocks	Online	48
September 13-15 2020	Summer workshop: "Multiscale simulations and biological channels"	Online	20
September 15 2020	Student Webinar: Summer School 2020 Edition	Online	43
October 7 2020	HADDOCK lecture + tutorial at the CCPBioSim	Online	46
October 13 2020	AstraZeneca virtual training (session 1)	Online	12
October 20 2020	HADDOCK lecture at 2020 Virtual NMR Symposium	Online	115
October 22 2020	BioExcel Webinar #49: Molecular movies made easy with Molywood	Online	102
October 23 2020	AstraZeneca virtual training (session 2)	Online	15
October 13 and 20 2020	Practical introduction to QM/MM using CP2K for biomolecular modelling	Online	37
20 October 20 2020	TCBG Seminar	Online	20

October 30 2020	QM/MM Best Practice workshop: kickoff webinar - Gerrit Groenhof	Online	
November 9,11, 13 2020	Computational biomolecular simulation workflows with BioExcel building blocks	Online	18
November 1-6 2020	International Semantic Web Conference (ISWC)	Online	600
November 5 2020	BioExcel Webinar #50: Using competencies to guide training and professional development	Online	25
November 9 - 13 2020	ELIXIR biohackathon 2020	Online	400
November 26 2020	EBI structural bioinformatics course	Online	24
November 26 2020	EXDCI Workshop on AI and HPC convergence	Online	85
November 30 - December 4 2020	BioExcel Winter School on Biomolecular Simulations	Online	30
November 27 - December 4 2020	International FAIR Convergence Symposium	Online	300
December 2-4 2020	Online Thermofischer Winterschool on integrative structural biology	Online	60
December 8 2020	BioExcel Webinar #51: Multiscale QM/MM simulations: exploring chemical reactions using novel GROMACS/CP2K interface	Online	123
December 14 2020	VIII Bioinformatics and Genomics Symposium, special mini-symposium focused on bioinformatics and genomics of COVID-19. UPF, Spain	Online	
December 15 2020	Webinar on “Drug repurposing against SARS-Cov2 using HADDOCK” at Interdisciplinary consortia for the study of pandemics, CIC biomaGUNE Spain	Online	35
December 15 2020	Iktos virtual training (session 1) on Free Energy calculations	Online	9

Table 2. 2021 Training and Dissemination Events

Date	**Event name**	**Location**	**Number of Participants**

8 January 2021	QM/MM Best Practice workshop - Janez Mavri	Online	73
12 January 2021	QM/MM Best Practice workshop - Carme Rovira	Online	95
12-15 January 2021	BioExcel/ENCCS GROMACS workshop	Online	30
13 January 2021	WorkflowsRI Community Summit	Online	40
18 January 2021	Student Webinar: Winter School 2020 Edition	Online	38
20 January 2021	HiPEAC conference 2021 workshop: The HPC CoE services and applications	Online	
26-27 January 2021	PATC - Programming Distributed Computing Platforms with COMPSs	Online	
26 January 2021	Iktos virtual training (session 2) on Free Energy methods application	Online	
29 January 2021	QM/MM Best Practice workshop - Panel Discussion	Online	
8-10 February 2021	CWL Virtual Mini-Conference	Online	
01-03 February 2021	CSC advanced workflows with GROMACS	Online	18
10 February 2021	Social media for scientists	Online	4
18 February 2021	BioExcel Webinar #52: What's new in GROMACS 2021	Online	96
February 2021	AstraZeneca workshop using proprietary data	Online	
February 2021	Malta Participation in Horizon2020	Online	
25 February - 25 March 2021	CECAM Workshop: Simulation of open systems in Chemistry, Pharma, Food Science and Immuno-diagnostics: Rare-event methods at constant chemical potentials including constant pH - an E-CAM Industry Scoping Work	Online	

3 March 2021	CINECA - Using Twitter to promote your project - beginner	Online	5
5 March 2021	CINECA - Using Twitter to promote your project - intermediate	Online	15
04-05 March 2021	Women in Science Symposium 2021	Online	
5 March 2021	CoE co-design workshop	Online	50
11 March 2021	BioExcel Webinar #53: A performance portable library for computing forces and energies of multi-particle systems	Online	41
17-18 March 2021	PATC: Introduction to Simulation Environments for Life Sciences	Online	30
21-26 March 2021	ISGC 2021 Taipei - with HADDOCK workshop	Online	
March/April 2021	EU-ASEAN HPC school	Online	
22-24 March 2021	PRACE/BioExcel Hands-on Introduction to HPC for Life Scientists	Online	20
30 March 2021	BioExcel Webinar #54: Applying the Accelerated Weight Histogram method to alchemical transformations	Online	91
30 March 2021 - 1 April 2021	Collaborations Workshop 2021	Online	
1-4 April 2021	Ga4GH Connect		50
7-9 April 2021	CINECA course "High Performance Molecular Dynamics"	Online	60
8 April 2021	CECAM Mixed-gen Session 4: Data Driven Science. Talk “BioExcel-CV19: a database of COVID-19 related Molecular Dynamics trajectories”	Online	100

12-13 April 2021	CROP-IB 2021 Breakthroughs in technology development, vegetable trait and crop improvement	Online	100
14 April 2021	EOSC-Life Exchange of Experience Workshop: Remote Access and User Training	Online	100
20-23 April 2021	Research Data Alliance plenary 17th	Online	
22-23 April 2021	QM/MM - GROMACS+CP2K workshop at EPCC	Online	29
23-24 April 2021	Huenfeld 2021: Workshop on Computer Simulation and Theory of Macromolecules	Online	
25-28 April 2021	Advanced Physical and Computational Techniques to Investigate Protein Dynamics	Online	
29 April 2021	UKRI-BBSRC Workshop on Computing in the Biosciences	Online	
4-5 May 2021	PDB50: A special symposium celebrating the 50th anniversary of the Protein Data Bank	Online	
19 May 2021	UKRI BBSRC Workshop on AI in the Biosciences	Online	70
30 May - 5 June 2021	Integrative modelling of biomolecular interactions	Online	24
9 June 2021	Molecular Modeling workshop in Oslo, Norway	Online	
6-9 June 2021	Journées Ouvertes de Bioinformatique, Informatique et Mathématiques	Paris	400
7-11 June 2021	BioExcel Summer school	Online	30

07-11 June 2021	CECAM Flagship Workshop: Open Databases Integration for Materials Design	Online	
15-17 June 2021	Free energy methods in drug design	Online	500
17 June 2021	Why social media matters for research now more than ever	Online	50
19 June 2021	SBGRID webinar - HADDOCK	Online	30
24 June - July 2021	The Deep Learning on Supercomputers Workshop	Online	50
28-30 June 2021	International Conference on ICT enhanced Social Sciences and Humanities 2021	Online	
29 June - 1 July 2021	International Society of Quantum Biology and Pharmacology (ISQBP) 2021 President's Meeting	Online	150
07-11 July, 2021	EU-ASEAN HPC Virtual School 2021: System Design and HPC Applications	Online	
25-29 July 2021	ISMB/ECCB	Online	
26-27 July or 28-29 July, 2021	Bioinformatics Open Source Conference (BOSC)	Online	
2-6 August 2021	New directions of AI in structural biology - CIRM (Centre International de Rencontres Mathématiques)	Online	50
6-9 Sept 2021	German Conference on Bioinformatics GCB 2021	Online	250
9-10 September 2021	GROMACS workshop: a collaboration between bioexcel and national EuroHPC competence center - portugal	Online	68

13-18 September2021	Mathematics of Life 2021	Hybrid (Hysaria, Bulgaria)	
14-17 September 2021	Expoquimia	Barcelona	
20 September 2021	FAIReScience workshop collocated at IEEE eScience conference	Virtual	
21 September 2021	Student Webinar: Summer School 2021 Edition	Online	19
24 September 2021	Instruct Software Developer webinar	Online	
27 September 2021	SBGRID Webinar - GROMACS	Online	30
27 September 2021	NHR Atomistic Simulations Lab seminar on BioExcel GROMACS-CP2K interface	Online	28
29 September - Oct 1 2021	GGMM (Groupe de Graphisme et de Modélisation Moléculaire)/SFCI (Socété Française de Chemoinformatique)	Lille, France	100
4-8 October 2021	Research to service: Planning and running a bioinformatics core facility	Barcelona	
12 October 2021	BioExcel Webinar #55: MDAnalysis, interoperable analysis of biomolecular simulations in Python	Online	181
11-15 October, 2021	Structural Bioinformatics EBI course	Online	30
18 - 21 October 2021	Advances and Challenges in Biomolecular Simulations	Online	166
19 - 21 October 1 2021	EGI conference 2021	Online	200

19-22 October 2021	Computational Approaches to Understanding and Engineering Enzyme Catalysis	Hybrid (Oulu, Finland)	30
26 October, 2021	BioExcel Webinar #56: 3dRS: a Web-Based Tool to Share Interactive Representations of 3D Biomolecular Structures and Molecular Dynamics Trajectories	Online	86
27 October, 2021	OSCU Open Science Symposium – Faculty of Science	Hybrid (Utrecht, The Netherlands)	79 (50% on location, 50% online)
9 November 2021	"FAIR Digital Objects" breakout at RDA 18th plenary	Online	43
8-12 November 2021	ELIXIR Biohackathon Europe	Hybrid (Barcelona, Spain)	150 F2F 150 remote
9 November 2021	BioExcel Webinar #57: Computationally designing therapeutic antibodies - combining immune repertoire data and structural information	Online	106
15 November 2021	WORKS Workshop, Supercomputing 2021	Online	
23 November 2021	BioExcel Webinar #58: CHARMM Force Field Development History, Features, and Implementation in GROMACS	Online	192
November - December 2021	bioBB virtual training nr 3	Online	12
9 December 2021	BioExcel Webinar #59: X3DNA-DSSR, a resource for structural bioinformatics of nucleic acids	Online	27
15 December 2021	5th International Symposium on Bioinformatics	Online	700

Table 3. 2022 Training and Dissemination Events

Date	Event name	Location	Number of Participants
31 January - 02 February 2022	PATC: Introduction to HPC for Life Scientists	Online	27
4 February 2022	HPC Fundamentals for End Users	Online	
7-9 February 2022	ONLINE: Advanced GROMACS workshop	Online	26
17 February 2022	BioExcel Webinar #60 - Social Media in Science Communication	Online	47
8 March 2022	BioExcel Webinar #61: What's new in GROMACS 2022	Online	101
9-10 March 2022	Virtual course on free energy calculation and alchemical transformation in GROMACS	Online	27
25 March - 1 April 2022	BioExcel School on Biomolecular Simulations	Online	30
2022	EMBO Practical Course: Practical integrative structural biology	Hamburg	TBC
23 - 25 May 2022	HADDOCK workshop	Budapest	TBC
12- 17 June 2022	BioExcel Summer School on Biomolecular Simulations	Pula (CA), Italy	TBC
November (1 year after planned 2020 US edition) OR could be 2022 during extension	Alchemical Free Energy workshop		

The following publications shown in Table 4, 5 and 6 were published by BioExcel members from June 2020 - February 2022.

Table 4. 2020 Publication list

Authors	Title	Journal Name	DOI	Open Access Link

Zivanovic, Sanja, Bayarri Genis, Colizzi Francesco, Moreno David, Gelpí Josep Lluis, Soliva Robert, Hospital Adam, and Orozco Modesto	Bioactive Conformational Ensemble Server And Database. A Public Framework To Speed Up In Silico Drug Discovery	J. Chem. Theory Comput.	10.1021/acs.jctc.0c00305	https://doi.org/10.1021/acs.jctc.0c00305
Zivanovic, Sanja, Colizzi Francesco, Moreno David, Hospital Adam, Soliva Robert, and Orozco Modesto	Exploring The Conformational Landscape Of Bioactive Small Molecules	J. Chem. Theory Comput.	10.1021/acs.jctc.0c00304	https://doi.org/10.1021/acs.jctc.0c00304
Wieczór, Miłosz, Hospital Adam, Bayarri Genis, Czub Jacek, and Orozco Modesto	Molywood: streamlining the design and rendering of molecular movies	Bioinformatics	10.1093/bioinformatics/btaa584	https://doi.org/10.1093/bioinformatics/btaa584
Moreno, David, Zivanovic Sanja, Colizzi Francesco, Hospital Adam, Aranda Juan, Soliva Robert, and Orozco Modesto	DFFR: A New Method for High-Throughput Recalibration of Automatic Force-Fields for Drugs	J. Chem. Theory Comput.	10.1021/acs.jctc.0c00306	https://doi.org/10.1021/acs.jctc.0c00306
Alexander Leitner, Alexandre M.J.J. Bonvin, Christoph H. Borchers, Robert J. Chalkley, Julia Chamot-Rooke, Colin W. Combe, Jürgen Cox, Meng-Qiu Dong, Lutz Fischer, Michael Götze, Fabio C. Gozzo, Albert J. R. Heck, Michael R. Hoopmann, Lan Huang, Yasushi Ishihama, Andrew R. Jones, Nir Kalisman, Oliver Kohlbacher, Karl Mechtler, Robert L. Moritz, Eugen Netz, Petr Novak, Evgeniy Petrotchenko, Andrej Sali, Richard A. Scheltema, Carla Schmidt, David Schriemer, Andrea Sinz, Frank Sobott, Florian Stengel, Konstantinos Thalassinos, Henning Urlaub, Rosa Viner, Juan A. Vizcaino, Marc R. Wilkins, Juri Rappsilber	Towards Increased Reliability, Transparency and Accessibility in Crosslinking Mass Spectrometry	ArXiv	https://arxiv.org/abs/2007.00383v1	https://arxiv.org/abs/2007.00383v1
Jorge Roel-Touris, Brian Jiménez-García, Alexandre M.J.J. Bonvin	Integrative Modeling of Membrane-associated Protein Assemblies	Nature Communications	10.1038/s41467-020-20076-5	https://www.nature.com/articles/s41467-020-20076-5
V. Tozzini , G. Palermo , A.M.J.J. Bonvin , M.	Editorial: Multiscale Modeling From	Frontiers in Molecular Biosciences	10.3389/fmolb.2020.0	https://doi.org/1

Dal Peraro and R.E. Amaro	Macromolecules to Cell: Opportunities and Challenges of Biomolecular Simulations		0194	0.3389/fmolb.2020.00194
Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R. Crusoe, Kristian Peters and Daniel Schober	FAIR Computational Workflows	Data Intelligence	10.1162/dint_a_00033	https://doi.org/10.1162/dint_a_00033
Nick Juty, Sarala M. Wimalaratne, Stian Soiland-Reyes, John Kunze, Carole A. Goble and Tim Clark	Unique, Persistent, Resolvable: Identifiers as the Foundation of FAIR	Data Intelligence	10.1162/dint_a_00025	https://doi.org/10.1162/dint_a_00025
Annika Jacobsen, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, Mercè Crosas, Michel Dumontier, Chris T. Evelo, Carole Goble, Giancarlo Guizzardi, Karsten Kryger Hansen, Ali Hasnain, Kristina Hettne, Jaap Heringa, Rob W.W. Hooft, Melanie Imming, Keith G. Jeffery, Rajaram Kaliyaperumal, Martijn G. Kersloot, Christine R. Kirkpatrick, Tobias Kuhn, Ignasi Labastida, Barbara Magagna, Peter McQuilton, Natalie Meyers, Annalisa Montesanti, Mirjam van Reisen, Philippe Rocca-Serra, Robert Pergl, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Juliane Schneider, George Strawn, Mark Thompson, Andra Waagmeester, Tobias Weigel, Mark D. Wilkinson, Egon L. Willighagen, Peter Wittenburg, Marco Roos, Barend Mons and Erik Schultes	FAIR Principles: Interpretations and Implementation Considerations	Data Intelligence	10.1162/dint_r_00024	https://doi.org/10.1162/dint_r_00024
Paul Groth, Helena Cousijn, Tim Clark and Carole Goble	FAIR Data Reuse – the Path through Data	Data Intelligence	10.1162/dint_a_00030	https://doi.org/10.1162/dint_a_00030

Gerrit Groenhof, Vaibhav Modi, Dmitry Morozov	Observe while it happens: catching photoactive proteins in the act with non-adiabatic molecular dynamics simulations	Current Opinion in Structural Biology	https://doi.org/10.1016/j.sbi.2019.12.013	https://www.sciencedirect.com/science/article/pii/S0959440X19301526
G.Portella, M.Orozco and M.Vendruscolo	Determination of a structural ensemble representing the dynamics of a G-quadruplex DNA	Biochemistry	10.1021/acs.biochem.9b00493	
A.Sridhar, S.E.Farr, G.Portella, T.Schlick, M.Orozco and R.Collepardo	Dependence of chromatin hierarchical looping on protein disorder and nucleosome asymmetry	Proc. Natl Acad. Sci. USA.	doi/10.1073/pnas.1910044117	www.pnas.org/cgi/doi/10.1073/pnas.1910044117
J.Walther, P.D.Dans, A.Balaceanu, A.Hospital, G.Bayarri and M.Orozco.	A Multi-Modal Coarse-Grain Model of DNA Flexibility Mappable to the Atomistic Level	Nucleic Acids Res.	doi: 10.1093/nar/gkaa015	https://doi.org/10.1093/nar/gkaa015
A.Sridhar, M.Orozco and R.Collepardo-Guevara.	Protein disorder-to-order transition enhances the nucleosome-binding affinity of H1.	Nucleic Acids Res.	doi: 10.1093/nar/gkaa285	https://doi.org/10.1093/nar/gkaa285
P.J.Dziubanska-Kusibab, H.Berger, F.Battistini, B.A.M.Bouwman, A.Iftekhar, R.Katainen, T.Cajuso, N.Crosetto, M.Orozco, L.A. Aaltonen and T.F.Meyer.	Collibactin DNA damage signature indicates mutational impact in colorectal cancer	Nature Medicine	10.1038/s41591-020-0908-2	https://doi.org/10.1038/s41591-020-0908-2
S. Sati, B. Bonev, Q. Szabo, D. Jost, P. Bensadoun, F. Serra, V. Loubiere, G. L. Papadopoulos, J.C. Rivera-Mulia, L. Fritsch, P. Bouret, D. Castillo, J.L. Gelpi, M. Orozco, C. Vaillant, F. Pellestor, F. Bantignies, M. A Marti-Renom, D. M Gilbert, J.M. Lemaitre and G. Cavalli.	4D Genome rewiring during oncogene-induced and replicative senescence	Mol.Cell	10.1016/j.molcel.2020.03.007	https://doi.org/10.1016/j.molcel.2020.03.007

Table 5. 2021 Publication list

Authors	**Title**	**Journal Name**	**DOI**	**Open Access Link**
R. Vargas Honorato, P.I. Koukos, B. Jimenez-Garcia, A. Tsaregorodtsev, M. Verlato, A. Giachetti, A. Rosato and A.M.J.J. Bonvin.	Structural biology in the clouds: The WeNMR-EOSC Ecosystem.	Frontiers Mol. Biosci.	10.3389/fmolb.2021.729513	
O.M.H. Salo-Ahen#, I. Alanko, R. Bhadane, A.M.J.J. Bonvin#, R. Vargas Honorato, S. Hossain, A.H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A. Støttrup Larsen, E. Lescrinier, P. Marimuthu, M.. Mirza, G. Mustafa, A. Nunes-Alves#, T. Pantsar, A. Saadabadi, K. Singaravelu, M. Vaeert.	Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development	Processes	https://doi.org/10.3390/pr9010071	
Brian Jimenez-Garcia, João Teixeira, Mikael Trellet, João Rodrigues, Alexandre M.J.J. Bonvin	PDB-Tools Web: A user-friendly interface for the manipulation of PDB files	PROTEINS: Structure, Function, and Bioinformatics	10.1002/prot.26018	https://onlinelibrary.wiley.com/doi/full/10.1002/prot.26018
Tobias Kuhn, Vincent Emonet, Haris Antonatos, Stian Soiland-Reyes, Michel Dumontier	Semantic micro-contributions with decentralized nanopublication services	PeerJ Computer Science	10.7717/peerj-cs.387	
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.	Ultrafast structural changes within a photosynthetic reaction centre	Nature	https://doi.org/10.1038/s41586-020-3000-7	https://www.nature.com/articles/s41586-020-3000-7#article-info

J. Williams, O. Yefanov, S. Westenhoff, J. Davidsson, D. P. DePonte, S. Boutet, A. Barty, G. Katona, G. Groenhof, G. Branden, R. Neutze				
Martin Werner, Vytautas Gapsys, Bert L de Groot	One Plus One Makes Three: Triangular Coupling of Correlated Amino Acid Mutations	The Journal of Physical Chemistry Letters	10.1021/acs.jpclett.1c00380	https://pubs.acs.org/doi/abs/10.1021/acs.jpclett.1c00380
Yuriy Khalak, Gary Tresadern, Bert L de Groot, Vytautas Gapsys	Non-equilibrium approach for binding free energies in cyclodextrins in SAMPL7: force fields and software	Journal of computer-aided molecular design	10.1007/s10822-020-00359-1	https://link.springer.com/article/10.1007/s10822-020-00359-1
Charlotte W. van Noort, Rodrigo V. Honorato, Alexandre M.J.J. Bonvin	Information-Driven Modeling of Biomolecular Complexes	Current Opinion in Structural Biology	https://doi.org/10.1016/j.sbi.2021.05.003	https://arxiv.org/abs/2103.07508
G. Dudas, S.L. Hong, B.I. Potter, S. Calvignac-Spencer, F.S. Niatou-Singa, T.B. Tombolomako, T. Fuh-Neba, U. Vickos, M. Ulrich, F.H. Leendertz, K. Khan, A. Watts, I. Olendraite, J. Snijder, K.N Wijnant, A.M.J.J. Bonvin, P. Martres, S. Behillil, A. Ayouba, M.F. Maidadi, D.M. Djomsi, C. Godwe, C. Butel, A. Simaitis, M. Gabrielaite, M. Katenaite, R. Norvilas, L. Raugaite, R. Jonikas, I. Nasvytiene, Z. Zemeckiene, D. Gecys, K. Tamusauskaite, M. Norkiene, E. Vasiliunaite, D. Ziogiene, A. Timinskas, M. Sukys, M. Sarauskas, G. Alzbutas, D. Juozapaite, D. Naumovas, A. Pautienius, A. Vitkauskiene, R.	Travel-driven emergence and spread of SARS-CoV-2 lineage B.1.620 with multiple VOC-like mutations and deletions in Europe.	Nature Comm.	https://doi.org/10.1038/s41467-021-26055-8	https://doi.org/10.1038/s41467-021-26055-8

Ugenskiene, A. Gedvilaite, D. Cereskevicius, V. Lesauskaite, L. Zemaitis, L. Griskevicius, G. Baele.				
Michael R. Crusoe, Sanne Abeln, Alexandru Iosup, Peter Amstutz, John Chilton, Nebojša Tijanić, Hervé Ménager, Stian Soiland-Reyes, Carole Goble, The CWL Community	Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language	arXiv (Communications of the ACM)	https://doi.org/10.48550/arXiv.2105.07028	https://arxiv.org/abs/2105.07028
Rudolf Wittner, Cecilia Mascia, Francesca Frexia, Heimo Müller, Markus Plass, Clare Allocca, Fay Betsou, Tony Burdett, Ibon Cancio, Adriane Chapman, Martin Chapman, M\'elanie Courtot, Vasa Curcin, Johann Eder, Mark Elliot, Katrina Exter, Elliot Fairweather, Carole Goble, Martin Golebiewski, Andreas Kremer, Sheng Lin-Gibson, Anna Marsano, Marco Mattavelli, Josh Moore, Hiroki Nakae, Isabelle Perseil, Ayat Salman, James Sluka, Stian Soiland-Reyes, Caterina Strambio, Michael Sussman, Jason R. Swedlow, Kurt Zatloukal, Jörg Geiger	Towards a Common Standard for Data and Specimen Provenance in Life Sciences		https://doi.org/10.5281/zenodo.5093125	https://zenodo.org/record/5093125#.Ykvly4VByUk

Stian Soiland-Reyes, Genís Bayarri, Pau Andrio, Robin Long, Douglas Lowe, Ania Niewielska, Adam Hospital	Making Canonical Workflow Building Blocks interoperable across workflow languages	(Data Inteligence Journal)	https://doi.org/10.1162/dint_a_00135	https://direct.mit.edu/dint/article/doi/10.1162/dint_a_00135/109838/Making-Canonical-Workflow-Building-Blocks
Rehm et al.	GA4GH: International policies and standards for data sharing across genomic research and healthcare	Cell Genomics	https://doi.org/10.1016/j.xgen.2021.100029	
S. Nizamuddin, S. Koidl, T. Bhuiyan, T.V. Werner, M.L. Biniossek, A.M.J.J. Bonvin, S. Lassmann and H.Th.M. Timmers	Integrating quantitative proteomics with accurate genome profiling of transcription factors by greenCUT&RUN	Nucl. Acids Res	https://doi.org/10.1093/nar/gkab038	https://doi.org/10.1093/nar/gkab038
N.Renaud, C. Geng, S. Georgievska, F. Ambrosetti, L. Ridder, D.F Marzella, M.F. Réau, A.M.J.J. Bonvin and L.C. Xue	DeepRank: A deep learning framework for data mining 3D protein-protein interfaces	Nature Comm.	https://www.nature.com/articles/s41467-021-27396-0	https://www.biorxiv.org/content/10.1101/2021.01.29.425727v1
Z. Jandova, A.V. Vargiu and A.M.J.J. Bonvin	Native or Non-Native Protein–Protein Docking Models? Molecular Dynamics to the Rescue	J. Chem. Theo. and Comp.	https://doi.org/10.1021/acs.jctc.1c00336	https://doi.org/10.1021/acs.jctc.1c00336

C.J. Buchanan, B. Gaunt, P.J. Harrison, A. Le Bas, A. Khan, A.M. Giltrap, P.N. Ward, M. Dumoux, S. Daga, N. Picchiotti, M. Baldassarri, E. Benetti, C. Fallerini, F. Fava, A. Giliberti, P.I. Koukos, A. Lakshminarayanan, X. Xue, G. Papadakis, L.P. Deimel, V. Casablancas-Antràs, T.D.W. Claridge, A.M.J.J. Bonvin, Q.J. Sattentau, S. Furini, M. Gori, J.Huo, R.J. Owens, A. Renieri, GEN-COVID Multicenter Study, J.H. Naismith, A. Baldwin, B.G. Davis	Cryptic SARS-CoV2-spike-with-sugar interactions revealed by 'universal' saturation transfer analysis.	BioRxiv	https://doi.org/10.1101/2021.04.14.439284	https://doi.org/10.1101/2021.04.14.439284
P.I. Koukos, M.F. Réau and A.M.J.J. Bonvin	Shape-restrained modelling of protein-small molecule complexes with HADDOCK	J. Chem. Inf. and Mod.	https://doi.org/10.1021/acs.jcim.1c00796	https://doi.org/10.1021/acs.jcim.1c00796
T Neijenhuis, S.C. van Keulen and A.M.J.J. Bonvin	Interface Refinement of Low-to-Medium Resolution Cryo-EM Complexes using HADDOCK2.4	BioRxiv	https://doi.org/10.1101/2021.06.22.449462	https://doi.org/10.1101/2021.06.22.449462
M.F. Lensink, G. Brysbaert, T. Mauri, N. Nadzirin, S. Velankar, R.A.G. Chaleil, T. Clarence, P.A. Bates, R. Kong, B. Liu, G. Yang, M. Liu, H. Shi, X. Lu, S. Chang, R.S. Roy, F. Quadir, J. Liu, J. Cheng, A. Antoniak, C. Czaplewski, A. Giełdoń, M. Kogut, A.G. Lipska, A. Liwo, E.A. Lubecka, M. Maszota-Zieleniak, A.K. Sieradzan, R. Ślusarz, P.A. Wesołowski, K. Zięba, C.A. Del Carpio Muñoz, E. Ichiishi, A. Harmalkar, J.J. Gray, A.M.J.J. Bonvin, F. Ambrosetti, R. Vargas Honorato, Z. Jandova, B. Jiménez-García, P.I. Koukos, S. Van Keulen, C.W. Van Noort, M. Réau, J. Roel-Touris, S. Kotelnikov, D. Padhorny, K.A. Porter, A. Alekseenko, M. Ignatov, I. Desta, R. Ashizawa, Z. Sun, U. Ghani, N. Hashemi, S. Vajda, D. Kozakov, M. Rosell, L.A. Rodríguez-Lumbreras, J.	Prediction of protein assemblies, the next frontier: The CASP14-CAPRI experiment	Proteins: Struc. Funct. & Bioinformatics	http://doi.org/10.1002/prot.26222	http://doi.org/10.1002/prot.26222

Fernandez-Recio, A. Karczynska, S. Grudinin, Y. Yan, H. Li, P. Lin, S. Huang, C. Christoffer, G. Terashi, J. Verburgt, D. Sarkar, T. Aderinwale, X. Wang, D. Kihara, T. Nakamura, Y. Hanazono, R. Gowthaman, J.D. Guest, R. Yin, G. Taherzadeh, B.G. Pierce, D. Barradas-Bautista, Z. Cao, L. Cavallo, R. Oliva, Y. Sun, S. Zhu, Y. Shen, T. Park, H. Woo, J. Yang, S. Kwon, J. Won, C. Seok, Y. Kiyota, S. Kobayashi, Y. Harada, M. Takeda-Shitaka, P.J. Kundrotas, A. Singh, I.A. Vakser, J. Dapkūnas, . Olechnovič, Č. Venclovas, R. Duan, L. Qiu, X. Xu, S. Zhang, X. Zou, S.J. Wodak.				

Table 6. 2022 Publication list

Authors	**Title**	**Journal Name**	**DOI**	**Open Access Link**
Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, Carole Goble	Packaging research artefacts with RO-Crate	Data Science	https://doi.org/10.3233/DS-210053	https://www.researchobject.org/2021-packaging-research-artefacts-with-ro-crate/manuscript.html

2.4. EMBO Workshop Advances and Challenges in Biomolecular Simulations

The BioExcel flagship event was organised as an EMBO Workshop with the aim to establish itself as the main venue for bringing together researchers in the fields of biomolecular simulations, integrative modelling, free energy and drug design, workflows and automation and data integration. The workshop was sponsored by BioExcel and four of the six organisers are BioExcel PIs and/or work package leaders with the event logistics being run by the BioExcel training and dissemination work package. The workshop took place virtually 18 - 21 October 2021, with 173 attendees out of 16 countries worldwide and included 30 speakers, both invited and selected from abstracts.

After being postponed in 2020 due to the pandemic, the EMBO Workshop successfully took place virtually in 2021. The organising committee of the EMBO Workshop put significant effort into selecting an event platform and tailoring the programme to encourage discussion and networking between the attendees. The workshop took place in a dedicated virtual conference platform called Myia, which allowed participants to watch the live stream of the talks, and network with speakers and other participants in dedicated networking rooms. The workshop had two dedicated poster sessions and all posters were available during and after the workshop in the virtual platform. All attendees registered in Myia were able to contact each other through the private messaging system as well as a global chat where poster presenters were encouraged to introduce themselves and their research to entice participants to their posters and stimulate discussion. In addition to the event platform, an active social media campaign both before and during the workshop, and daily email highlights were used to increase engagement. An anonymous feedback survey was run directly by EMBO. In the feedback survey 99% of attendees rated the organisational aspects as excellent/very good/good (56% excellent, 38% very good, 5% good). 61% of respondents answered that the EMBO Workshop was better than other scientific conferences that they have attended virtually (21% the same, 3,5% worse and for 14% it was the first virtual event). The percentage of attendees who answered the feedback survey was 54%. The diversity of respondents were very good: 3% organisers, 6% invited speakers, 10% participants (selected for oral presentation), 61% participants (selected for poster presentation), 20% participants. Participants' positions were also well balanced: 23% senior academic (group leader/professor), 27% postdoctoral fellow, 46% PhD student, 4% undergraduate.

2.5. Pandemic

The EMBO workshop: Advances and Challenges in Biomolecular Simulations sponsored by BioExcel was originally due to take place 18-21 October 2020 in Brno, Czech Republic. The COVID-19 pandemic forced the organising committee to make alternative plans; after an initial postponement to October 2021, a decision was made in late 2020 to organise the workshop fully virtually. The move to a virtual format was taken in the interest of the safety of attendees, to promote inclusivity, to manage uncertainty and due to a continued reluctance to travel. The

programme of the workshop was adapted to better suit a virtual format, with the individual sessions shortened and the workshop running across an additional day (four days versus three-day face-to-face workshop).

2.5.1. Original plans

Originally the EMBO workshop was planned to take place in Brno in the Czech Republic and included a social visit to the Mendel Museum of Masaryk University and the Augustinian Abbey, the place of work of G.J. Mendel, for the workshop dinner. Figure 6 shows the original invitation poster and Figure 7 the website banner with the new design and date.

Figure 6. Original invitation for face-to-face workshop

Figure 7. New banner on the EMBO website with final dates and virtual workshop format

2.5.2. Myia conference environment

In order to deliver a high-quality event, a dedicated virtual event platform was chosen. A number of different platforms were evaluated, and the Myia conference environment was chosen. To give the participants a flavour of Brno, the virtual event platform used pictures provided by the Mendel Museum as the backdrop for the virtual lobby and also for the virtual meeting rooms (Fig. 8). A virtual tour of the Mendel Museum was offered as a social activity, with an appearance of G.J. Mendel himself. In addition, all participants received a unique glass vial with peas

from Mendel's garden in the workshop merchandise pack which was posted to them (Fig. 9).

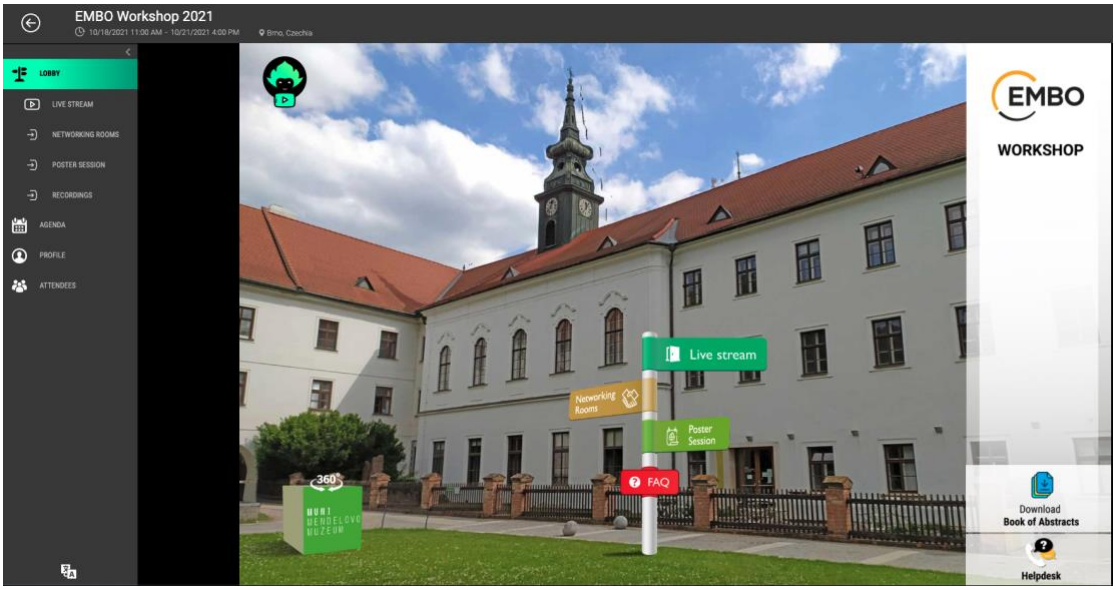


Figure 8. The virtual conference environment in the dedicated Myia event platform

Figure 9. Mendel peas sent as part of the merchandise package to participants

The Myia platform was customised by the Myia technical team to the needs of the EMBO workshop. To guarantee smooth running of the conference and to accommodate the international audience, the decision was made to pre-record the talks (all abstract talks and the majority of the invited talks) and hold live question and answer sessions. The live stream was recorded and made available immediately after the end of each day to allow participants from different time zones to watch the recording as soon as possible. Each talk was captioned to aid accessibility of the sessions. With approval of the speakers, individual talks remained available in the workshop platform to all registered participants until the end of 2021.

The Myia technical team ran the live stream during the conference, the schedule was well managed and the workshop ran to time which was confirmed in the participant feedback: the time management (e.g. keeping to time by the speakers, managing the discussions etc.) during the meeting: 80.53% sessions were kept on time, 11.50% most sessions were kept on time with only slight delays.

2.6. Programme

The programme was adapted to better suit the virtual meeting format and to accommodate the different time zones of the speakers. The programme included two keynote talks, talks from invited speakers, selected talks from abstract submissions as well as two poster sessions and a virtual guided tour of the Mendel Museum.

According to the participant feedback, 85% marked the programme content as the main priority when making the decision to attend the workshop, with the invited speakers (73%), and networking opportunities (34%) as second and third priority.

When asked if the meeting provided participants with new information in their field 95% answered "Yes". Scoring the overall evaluation of the programme, 97% of the participants answered *Excellent*, *Very good* or *Good* (44% *Excellent*, 38% *Very good*, 15% *Good*) and 96% of the participants evaluated the scientific talks as *Excellent*, *Very good* or *Good* (50 % Excellent, 38% *Very good*, 8% *Good*). The duration of the meeting was considered "just right" by 85% of participants, only 2% considered the meeting "too short" and the remaining 13% mentioned the meeting was "too long".

Participants from 16 countries and multiple different time zones attended the workshop. Figure 10 shows the number of participants who were watching the live stream each day of the workshop. From the usage statistics of the event platform we can see that each day of the workshop had a steady number of live attendees, with dropout between sessions being minimal. As expected for a multi-day (virtual) workshop there was a gradual dropoff of live attendees from day 1 to day 4 (9% from day 1-2, 4% from day 2-3 and 11% from day 3-4; 24% from day 1-4); which we consider to be very good and highlights that the workshop programme was able to keep the participants engaged throughout.

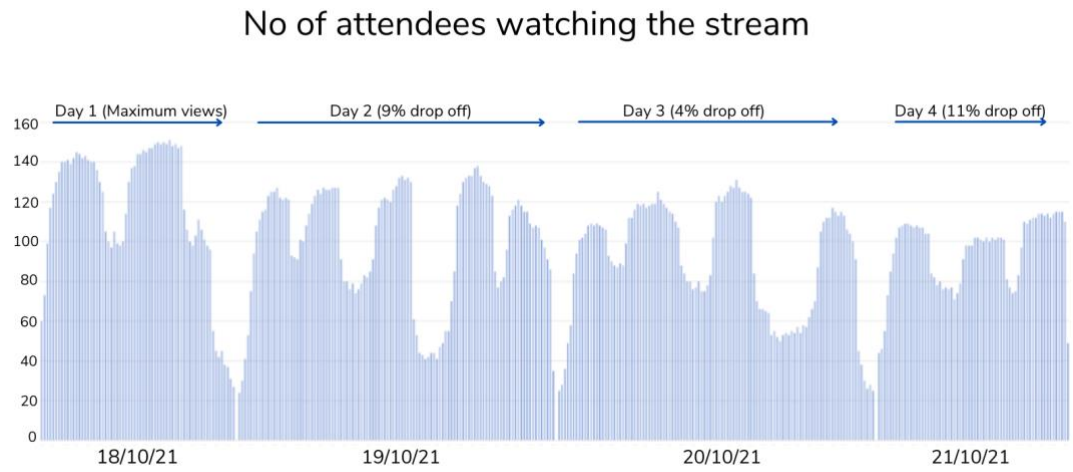


Figure 10. Live attendance figures watching the live stream during the workshop (graph statistics provided by the Myia platform analytics)

2.6.1. High level overview of topics, speakers (abstract & invited), poster, social

The EMBO Workshop Advances and Challenges in Biomolecular Simulations ran across 4-days and highlighted the scientific work in this area through 2 keynotes, 10 invited lectures, and 18 talks selected from abstracts. A high number of abstracts (124) were received, and all participants were encouraged to present a poster.

The workshop included talks around the following topics:

- Free Energy and Drug Design
- Enhanced sampling - Linking Experiments and Computations
- Molecular Simulations: this topic included hybrid methods (QMMM), classical simulations (MD) and methodological developments in force fields and molecular simulation techniques.
- Structural bioinformatics and integrative modelling.
- Cross-topics: various aspects of biomolecular simulations, including among others validation and reproducibility.

The keynote speakers were Tamar Schlick (New York University, USA) who spoke about *"Biomolecular modeling & simulation thriving in the age of technology"* and William L. Jorgensen (Yale University, USA) who presented work on *"FEP-Guided discovery of potent inhibitors of the main protease of SARS-CoV-2"*.

2.6.2. Poster session

A total of 98 posters were submitted, the workshop had two dedicated poster sessions, and all posters were available during and after the workshop. The top three rated posters were awarded a poster prize which has been shipped to their home. To facilitate discussion each poster had a dedicated virtual networking room which allowed live interaction with the abstract presenter as well as written messages (printscreen of the virtual poster area is in Figure 11). All attendees registered in Myia were able to contact each other through the private messaging

system as well as a global chat where poster presenters were encouraged to introduce themselves and their research to entice participants to visit their posters and stimulate discussion. In addition to the event platform, an active social media campaign both before and during the workshop, and daily email highlights were used to increase engagement.

Based on the statistics provided by Myia the posters were visited 691 times during the workshop. We can also tell that 98% of the poster discussion rooms had more than one attendee (assumption is that rooms with one attendee only had the presenter present), 49% had 5 or more, and 22% of the posters had 10 or more attendees. The most popular posters had more than 30 visitors. The poster sessions received mixed feedback in the survey and are recognisably one of the most difficult aspects to implement in a virtual event. We are very pleased that 80% of the respondents described the poster sessions as excellent/very good/good (19% excellent, 36% very good, 25% good), and several respondents commented that considering the limitations of a virtual format, this was the best virtual poster session they have ever attended.

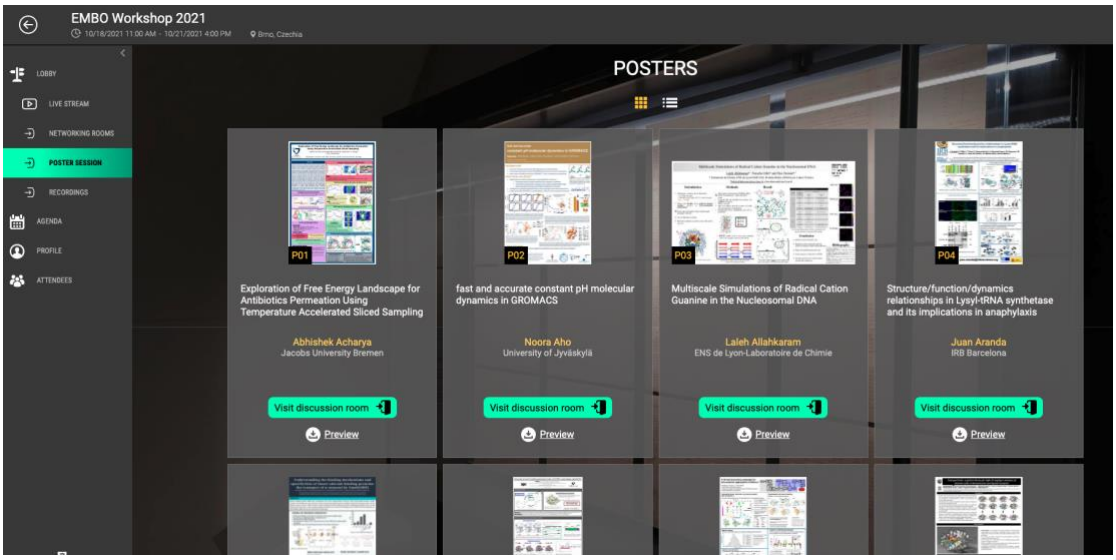


Figure 11. Myia event platform - poster session page

2.6.3. Networking

The Myia platform allowed participants to meet in the dedicated meeting rooms which were open during the whole week so participants could meet there at any time. The rooms were named according to the conference scientific topics and the same rooms were also used for the "meet the speaker session". The printscreen is below (Fig.12).

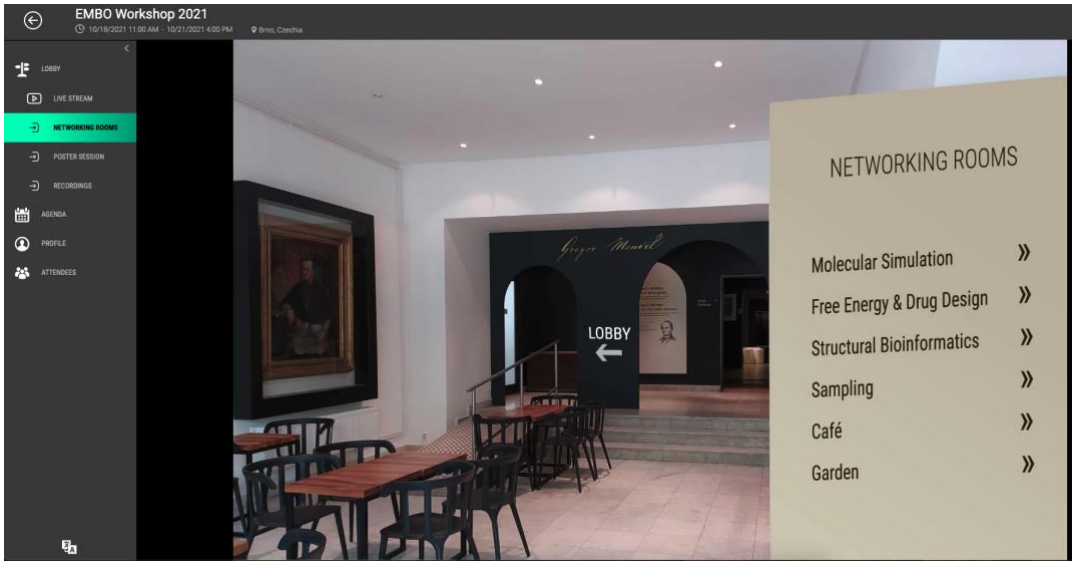


Figure 12. Myia Event platform - list of the breakout rooms, by clicking on the arrow participants joined the room

In Myia we were able to monitor the traffic of visitors to the meeting rooms as well as the poster sessions. From the received data we can confirm the most popular topic was Molecular Simulation (Table 7).

Table 7. Number of visitors in the breakout rooms

Room name	No of attendees
FAQ	39
Help Desk	20
Mendel Museum	39
Molecular Simulation	138
Free Energy & Drug Design	92
Structural Bioinformatics	85
Sampling	67
Café	96
Garden	99

93% of respondents felt that the number of participants at the meeting was just right, and respectively 89% and 81% of respondents thought there was sufficient time for discussion, and time to meet and network. We consider this a very good result for a virtual event. More than 40% of the respondents confirmed they acquired useful career contacts.

2.7. Statistics

In Table 8 we present the high-level statistics. The workshop has a good gender balance, with 43% female and 57% male speakers, and an equal number of female and male session chairs. A significant number of posters and abstracts were submitted which shows the successful engagement of the community.

Table 8. Statistics of participants

	Numbers	Percentage
Total number of participants	173	
Total number of speakers	30	
Total number of chairs	6	
Total number of organisers	8	
Number of female (participants only)	60	35%
Number of female speakers	13	43%
Number of female chairs	3	50%
Early Career Speakers	6	20%
Number of Nationalities participants	16	
Number of Nationalities (speakers)	11	
Number of posters in Myia	98	
Number of Abstracts submitted	124	
Number of Active users in Myia	209	95%

2.8. Impact

The EMBO Workshop Advances and Challenges in Biomolecular Simulations provided a platform for the biomolecular simulation community, especially early career researchers, to discuss, network and showcase their research. The virtual platform allowed us to bring people from all over the world thus making the event more inclusive and accessible. Through our online community building practices, we were able to generate a sense of excitement and belonging where people were able to share their experiences during the workshop. Many of the workshop attendees had previously attended BioExcel training courses, webinar or other types of events, showing how successful we have been in building a community around BioExcel Centre of Excellence.

2.8.1. Social media/communication

Twitter

A Twitter campaign was designed to promote the workshop and highlight the activities during this period. The campaign included tailored graphics and social media copy to promote speaker talks, networking activities and poster awards. We developed the #EMBOBioSim hashtag to track metrics and conversations across platforms. Twitter analytics displayed that we received 45.3K impressions and 1.4% engagement rate. A large amount of engagement was seen on twitter where attendees shared their workshop experience while also promoting their talks. These discussions helped foster the growing biomolecular research community, give a platform to early career researchers and boost their profile.

Example of tweets from the workshop are provided below (Fig. 13-15):

Figure 13. Tweet describing Day 1 of the conference

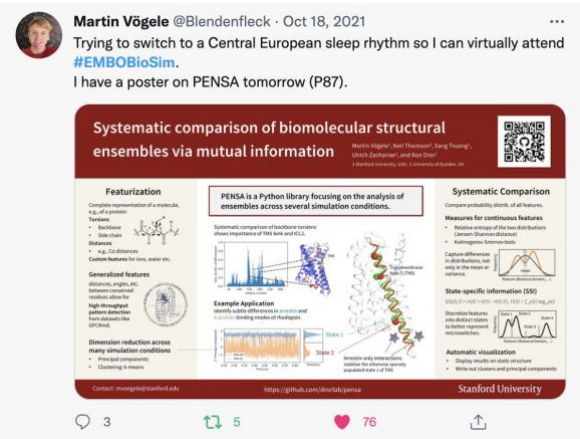


Figure 14. Tweet from participant sharing their poster

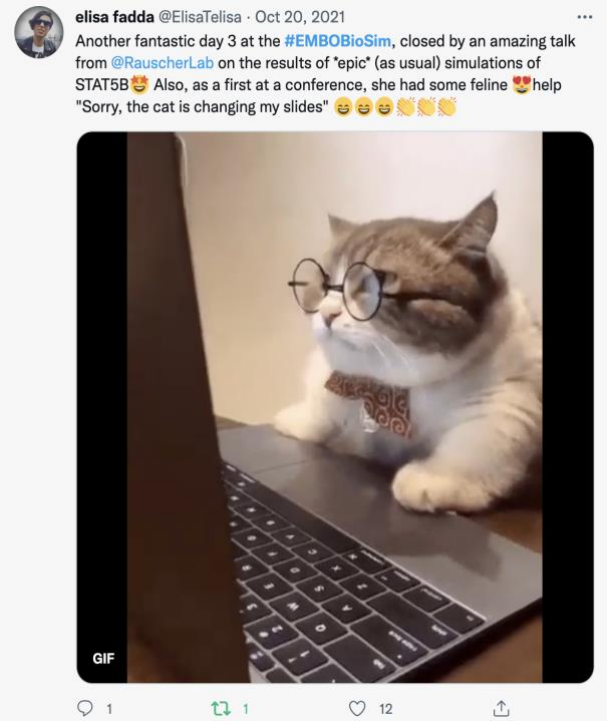


Figure 15. Tweet from the organiser sharing their Day 3 experience

LinkedIn

A LinkedIn campaign was designed alongside the twitter campaign to showcase workshop activities to a more industry focused audience. During this period, we received a 4.7% engagement rate. Similar to the twitter campaign, we were able to provide visibility to early career researchers to a new audience. An example of a LinkedIn post from the workshop is provided below in Figure 16.

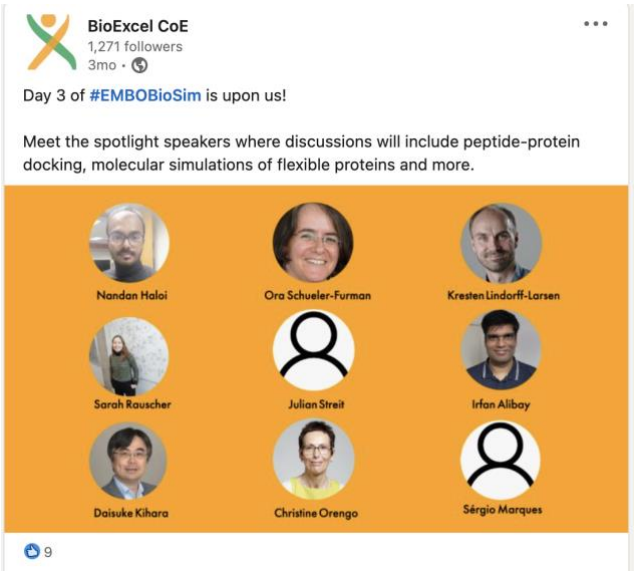


Figure 16. LinkedIn post highlighting Day 3 of the conference

In the feedback form we have asked how participants first learnt about this meeting? The most frequent responses were: >38% oral invitations (friend or colleague), around 16% got the information from email communication or from the EMBO website.

For communication purposes we created a dedicated email address acbiosim@bioexcel.com which has been used in the preparation phase as well as during the conference. To facilitate engagement with virtual participants we sent the highlighted information every day on top of standard workshop communication (Fig. 17).

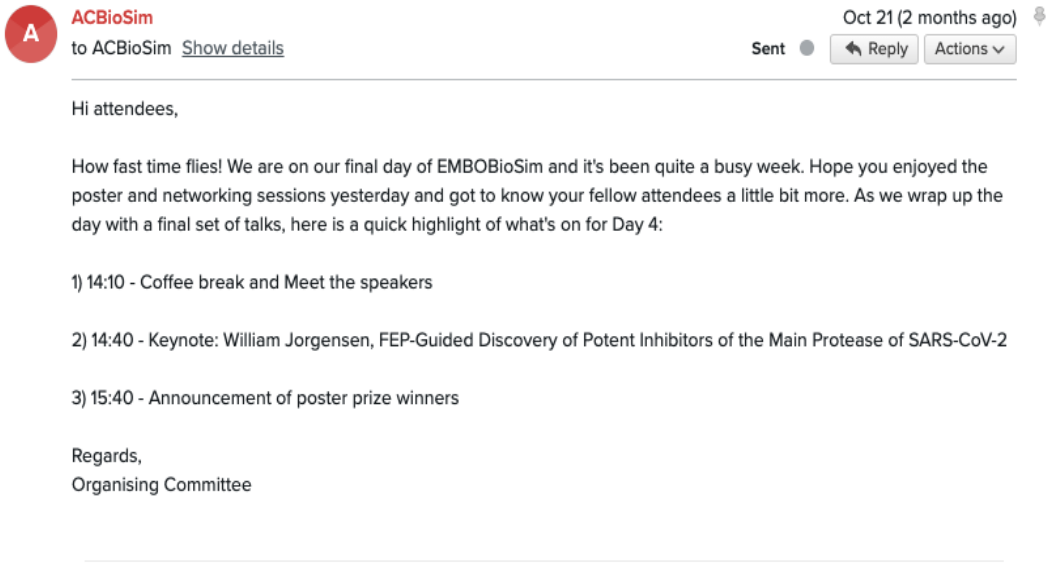


Figure 17. Example of the email which has been sent every day after the meeting as a wrap-up

3. Training

Here we present the training events carried out since June 2020, as the previous events with participation from BioExcel were reported in deliverable D4.3. - *Progress report and update on training and dissemination plan*. We include information about contents, participation and feedback. We present additional results from training impact assessment, including information collected in surveys sent out 6 months or 2 years after a training event.

In addition, the BioExcel training programme builds on the strong commitment of the BioExcel partners to deliver high quality training to the computational biomolecular community. Additional courses in which BioExcel partners have delivered training are presented in tables 1, 2 and 3 in section 2.3 (*Events and publications)* of this deliverable.

Finally, we show the most recent developments to the competency profile on which the BioExcel training programme is based. We have added career profiles

and learning pathways to help people in the computational biomolecular community decide on their next career steps and find training of interest.

3.1. Events

In table 9 we present summary data about number of participants and feedback received after the courses training events. In the sections below, we present further information regarding each course.

Table 9. Summary data from training events

Course	**Number of participants**	**% of female participants**	**% of participants who responded the feedback survey**	**% of feedback responses in *Very Good* & *Excellent***
Remote BioExcel Summer School on Biomolecular Simulations 2020	30	53.3%	80%	91.66%
Introduction to GROMACS - A SNIC/PRACE workshop in collaboration with BioExcel	54	Not collected	52%	93%
Practical introduction to QM/MM using CP2K for biomolecular modelling	37	18.92%	18.92%	86%
Computational biomolecular simulation workflows with BioExcel Building Blocks - 2020	18	44.44%	66.67%	100%
Remote BioExcel Winter School on Biomolecular Simulations	30	50%	93.33%	100%
BioExcel/ENCCS workshop – Advanced topics in simulations with GROMACS	30	33.33%	60%	100%
Introduction to HPC for Life Scientists – 2021	20	25%	60%	100%

QM/MM simulation with GROMACS + CP2K	29	17.24%	37.93%	100%
EMBO practical course - Integrative modelling of biomolecular interactions	24	50%	87.5%	100%
BioExcel Summer School on Biomolecular simulations 2021	29	58.62%	93.10%	100%
GROMACS workshop: a collaboration between bioexcel and national eurohpc competence centre - Portugal	68	39.71%	38.24%	96%
PRACE Autumn School: Fundamentals of Biomolecular Simulations and Virtual Drug	165	Not collected	22.4%	94.5%
Computational biomolecular simulation wokflows with BioExcel building blocks – 2021	12	25%	45.45%	100%
BioExcel/PerMedCoE PATC course on Introduction to HPC for Life scientists	27	62.96%	37%	100%
Advanced GROMACS workshop	26	7.69%	42%	98%
Virtual course on free energy calculation and	27	37.04%	40.74%	91%

alchemical transformation in GROMACS				
BioExcel School on Biomolecular Simulations	30	46.67%	66.67%	94%

3.1.1. BioExcel School on Biomolecular Simulations

In our previous deliverable, D4.3 - *Progress report and update on training and dissemination plan*, we explained that due to the pandemic of COVID-19, we were running the BioExcel Summer School in a remote format, but we did not report on the result of the course, which ran at the end of June 2020, when the deliverable was due. That was the first of four remote editions of the course. The main content and organisation have been the same for all four, but we have improved from one edition to the next based on the experience and feedback received. Here we present a general overview of the course and specific information about each of them.

The school has a duration of one week and is intended for researchers (primarily PhD and postdocs) using or planning to use biomolecular modelling and simulations in their everyday research. Familiarity with Linux and some basic knowledge of molecular modelling software is a requirement.

The remote course includes a combination of pre-recorded lectures, live Q&A sessions and live hands-on sessions. The topics included in the course are molecular dynamics simulations, biomolecular docking, BioExcel building blocks (Biobb), QM/MM, free energy calculations and advanced sampling methods (metadynamics). Participants have the possibility of asking questions through the ask.bioexcel forum, which was also used as a channel for social interaction (Fig. 18). For the hands-on sessions, we provide virtual machines for the participants, with all the software required already installed. During the hands-on computer practicals students use, among others, the BioExcel flagship software (GROMACS, HADDOCK, PMX and CP2K).

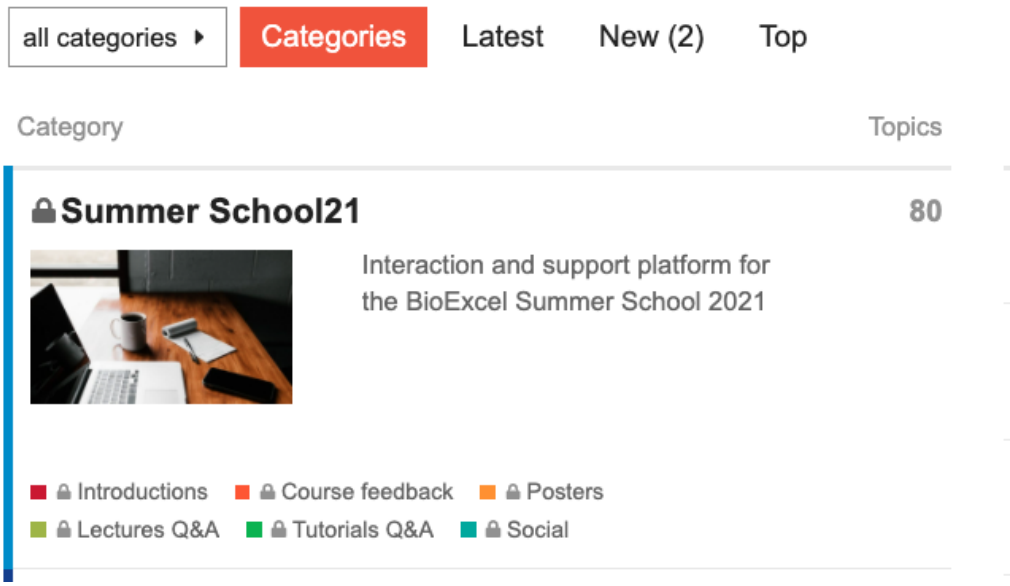


Figure 18. Screenshot of the ask.bioexcel forum showing the section for interactions during the summer school and the subsections in which participants and trainers can post messages: introductions, course feedback, posters, lectures Q&A, tutorials Q&A, and social. 80 conversations ('topics' on the image) were started in the forum.

In addition, the participants have the opportunity to present their research. This is organised as a combination of short presentations and posters that are available online during the entire course. Poster discussion is possible in the forum and other interactive tools. Three posters are selected during the course and their authors are invited to deliver a talk during a BioExcel webinar. Opportunities for social chat and networking are also provided in the course.

Below, we present specific information on each edition of the course focusing on the feedback received and the changes from the previous edition.

***BioExcel Summer School on Biomolecular Simulations* 22-26 June 2020**

30 participants attended this edition of the course, 16 of which were female. 80% of the participants responded to the feedback survey at the end of the course, where most of them rated the course as *Excellent* (70.83%) or *Very Good* (20.83%), with the rest rating it as *Good* (8.33%). All respondents would recommend the course and all of them considered that the balance between theoretical and practical sessions was about right. 91.67% of respondents stated that they would use the tools after the course. The rest replied "maybe" to that question.

In general, the course and most sessions were very well received, but in some of them, there was room for improvement and we took the feedback into account when designing the following editions of the course. Participants considered that it was good to have hands-on sessions to learn how to use the tool, but some of them can be improved. The session on CP2K was a demo and not a hands-on tutorial and some participants suggested that it should be a hands-on session next time. Some of the tutorials, especially the metadynamics one, were too long and people did not have time to finish it or they had to go too fast over it.

For the poster session, participants had 1 minute to present their poster on the first day. Then we gave participants time to look at the posters and ask questions on the ask.bioexcel.eu forum, but there was no live session where they could have a conversation with someone else. This was appreciated by some of them, because it led to some long conversations in some of the poster threads, with multiple views on the same topic, but some others missed the more direct interaction. The one minute presentation was considered too short by some of them.

In general, the different types of sessions and the online tools used during the course worked well, but some participants considered that the schedule was too tight and it implied spending too much time in front of the computer.

During the coffee breaks, we had chats about different topics with the participants. In the first one, we shared a poll with them to ask about possible topics for these chats and the one that interested them the most was careers outside academia. We invited Richard Norman and Ian Harrow, from Norman Consulting and Ian Harrow Consulting respectively, both partners in BioExcel, to participate in two of these breaks and tell participants about their experience in academia, in the industry and as consultants. This session was very successful, and we decided to take the idea to the following editions of the course, as an integral part of the programme.

***BioExcel Winter School on Biomolecular Simulations* 30 November - 4 December 2020**

After the success of the Summer School in June and with the possibility of reusing part of the material from June, we decided to run a winter edition of the course. We included some improvements based on the feedback received in June:

- There was some more time between sessions, so that people could get away from the screen
- Participants were given 2 minutes to present their posters
- The session on CP2K was a hands-on tutorial run in an HPC machine, where participants got access
- The tutorial on metadynamics was reduced to fit the scheduled time
- A career session was included in the programme, with three invited speakers that had followed different paths in their careers

In addition, as we already had a basic remote setup that worked, we decided to introduce one more tool, Gather, which allows for interaction with the people that you approach in a virtual environment. We ran two sessions in Gather, one to get to know the tool and the other participants, and another one where people could network and chat about the posters. The latter was right after the career session, so participants could use it to talk about career options. We added 6 posters selected by the trainers to Gather so that participants selected the best 3 ones. As Gather was the new tool in this course, we asked specifically for feedback about it, not only in the feedback survey at the end of the course, but also during the Gather sessions and the coffee breaks in Zoom. The new tool was very well received as an option to facilitate the interaction between participants.

More than 80 people applied to participate in the course, even though applications were open only for 6 weeks. 30 were selected, 15 males and 15 females. 93.3% of the participants responded to the feedback survey at the end of the course. All of them rated the course as *Excellent* (67.86%) or *Very Good* (32.14%). All respondents would recommend the course and all of them considered that the balance between theoretical and practical sessions was about right. 85.71% of respondents stated that they would use the tools after the course. The rest replied "maybe" to that question.

In general, all the sessions were well received and the ratings were more balanced than in the previous Summer School edition, where a few sessions were considered poorer than the rest. When asked about the best part of the course, some of the participants mentioned specific lectures or tutorials, but there were also some who talked about the good atmosphere in the course and the possibility to interact with other participants and trainers. This is a clear indication that we managed to build a full course experience and not a series of independent lectures and tutorials. Some tutorials can still be improved, so that people can follow them, especially the ones in the second half of the week (QM/MM, metadynamics, free energy calculations), as some participants might not have enough background knowledge on them. Our tutorial design takes into account that some people might only have one screen when following the course, but sometimes we forget it while delivering it, and it is necessary that we keep it in mind, so that participants have time to follow the tutorials themselves.

For the poster presentations, we gave participants 2 minutes on the first day and then we had two different types of sessions: one in the forum, one in the Gather, but we did not display all the posters in Gather. In general, the poster session is well appreciated. However, some people consider that it could be improved and some others consider that the forum part is not required, but some others appreciate having the possibility to ask questions in the forum. We think that maintaining both types of sessions is good to cater for different personality types. As one of the participants said: "I like the fact to have the forum, you can take time to respond to each question. And the Gather to complement it right". We can improve how we communicate this and we could include all the posters in Gather, so that it feels more like a poster session.

***BioExcel Summer School on Biomolecular Simulations* 4-11 June 2021**

The schedule was still a bit too tight in the Winter School and some people asked for a longer poster session in Gather, but this was not easy to accommodate. Therefore, we decided to start the course on a Friday. We spent the afternoon going through the introduction sessions and the participant presentations, which also allowed having more time between their presentations: we divided them into four groups, with 20 minutes for each group and a 10 minutes break between each of the groups.

We changed the poster session in Gather, which was longer and included all the posters. Unfortunately, there were some technical issues that day in Gather, so the first part of the session did not work well. Despite the technical difficulties, participants liked the use of Gather: "Gather was a great idea to socialize with the group. Although we had some trouble to connect Gather on the fourth day, the first day was perfect! I love the idea of talking with anybody who is close to you and also chatting as a group."

Another addition to this edition of the course was to include a coffee break with postdocs that could share their work and career path and respond to questions from participants. Some of them told us that it was a good idea, but it was too short, so we decided to allocate a longer break next time.

We received more than 150 applications for the course and selected 30 participants. One of them only attended the course for a very short time, so we had 29 participants, including 17 females and 12 males. 27 of them responded to the feedback survey, although only 23 responded to it completely. All of them rated the course as *Excellent* (73.91%) or *Very Good* (26.09%). All respondents considered that the balance between theoretical and practical sessions was about right and all but one would recommend the course (the other one responded "maybe"). 82.61% of respondents stated that they would use the tools after the course. The rest replied "maybe" to that question.

In general, all sessions were well received, including lectures, tutorials and networking sessions. To summarise it with the words of a participant: "the networking, posters and career sessions were very good. Kind of like a 'bonus' as the main aim was the taught material". When asked about the worst of the course,

many responses were about technical issues, these could be coming from the participant's internet connection or from other servers, such as the issue that we had with Gather for the poster session. We cannot do much about some of these issues, but we can try to minimise their effects on the participants' learning: for example, if we have detailed tutorial instructions, they will be able to follow the tutorial even if they get disconnected from the Zoom session.

***BioExcel School on Biomolecular Simulations* 25 March - 1 April 2022**

We ran a fourth edition of the remote course in March 2022, where we will follow the same schedule as in the previous Summer School in 2021, but we will provide some more time for the coffee break with postdocs, as participants think that this is a great opportunity for them.

For this edition, we used an online course handbook developed by EMBL-EBI for online courses (Fig. 19). From the feedback received, participants liked the course handbook and found it useful.

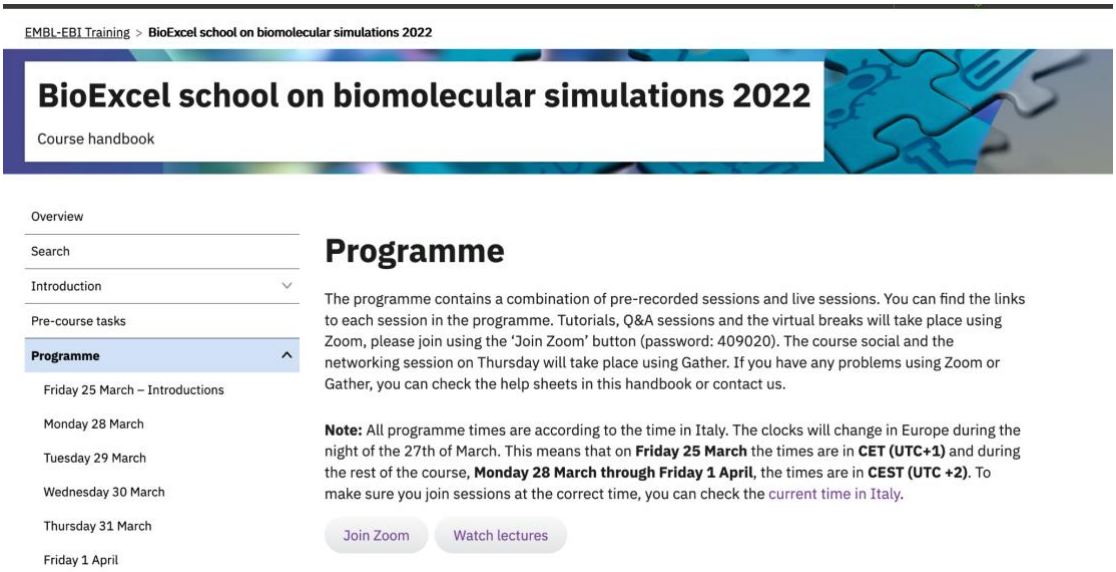


Figure 19. Screenshot of the programme page in the course handbook

30 participants were selected for this course from 71 applicants. One of them did not attend the live sessions because of the big time zone difference, but followed the recorded lectures and performed the tutorials on their own, and asked questions to the trainers in the BioExcel forum. Two other participants did not attend all the sessions due to illness. We received 20 responses (18 complete ones) in the feedback survey at the end of the course. Most respondents rated the course as *Excellent* (77.78%) or *Very Good* (16.67%), with the rest rating it as *Good* (5.56%). All respondents would recommend the course and all but one considered that the balance between theoretical and practical sessions was about right; the other one considered that it was too practical. 88.89% of respondents stated that they would use the tools after the course. The rest replied "maybe" to that question.

All sessions were well appreciated by the participants, although some of them would like to have more time for some of the tutorials. When asked about the best part of the workshop, most elements of it are mentioned by one or another respondent: lectures, tutorials, poster sessions, social elements. This indicates that we are delivering a complete learning experience with a variety of elements that can be valued by different people. One of the participants said: "Apart from well-selected topics, I liked the social parts such as the break talks and gather parts. In an online event such as this one, I value a lot these initiatives."

From the comments in the survey, we can see that a slow internet connection can be an issue for participants to be in Zoom and connect to the virtual machines or to connect to Gather; so this is something that we need to keep in mind, so that participants can still follow the course. It seems that the combination of pre-recorded lectures and live sessions and the fact that all the material is available for participants contributes to this, as even the participants that report internet issues give high ratings to the course.

3.1.2. Practical introduction to QM/MM using CP2K for biomolecular modelling 13 & 20 October 2020

This course was part of the series of courses *Roadmap to run your QM/MM simulations using CP2K and GROMACS* that started in June 2020 with the course *Preparing your systems for running QM/MM simulations in CP2K with AmberTools.* This series is organised jointly by BioExcel and ARCHER2 Training. The course in October 2020 consisted of two sessions of two hours of duration each.

This course aims to teach attendees how to prepare and run biochemical simulations in CP2K. Specifically, it covers how CP2K can be used to track the reaction dynamics of a system undergoing biological catalysis. In this course, the sessions focus on a Diels Alder transition taking place in an immunoglobulin. To start, the reaction site is described through a purely quantum-mechanical method. Then, the complexity of the system is increased to take into account the entire structure of the immunoglobulin (using a QM/MM approach), the solution, and finally, a ligand is added to the reaction. You can find more information on the event website:
https://bioexcel.eu/events/practical-introduction-to-qm-mm-using-cp2k-for-biomolecular-modelling/

The course had 37 attendees, 7 of which were female. 18.92% of them responded to the feedback survey. 86% of respondents rated the course as *Excellent* or *Very Good* and the rest as *Good.* Participants liked the content, organisation, trainers and hands-on sessions. Some respondents suggested including a more theoretical part in the course or suggesting some reading before the course.

3.1.3. QM/MM simulation with GROMACS and CP2K 22 & 23 April 2021

This course was part of the series of courses *Roadmap to run your QM/MM simulations using CP2K and GROMACS* that started in June 2020. This series is organised jointly by BioExcel and ARCHER2 Training. The course in April was a two day course. In response to the feedback received in the previous course, a more theoretical part was added in this course.

The target audience were people familiar with using GROMACS to perform molecular dynamics simulations and familiar with connecting to a remote machine using ssh. The course taught them how to perform hybrid QM/MM simulations using the molecular dynamics code GROMACS in combination with the quantum chemistry package CP2K through an interface developed by one of the instructors from the BioExcel Centre of Excellence. More information can be found on the event website:
https://bioexcel.eu/events/qm-mm-with-gromacs-cp2k/

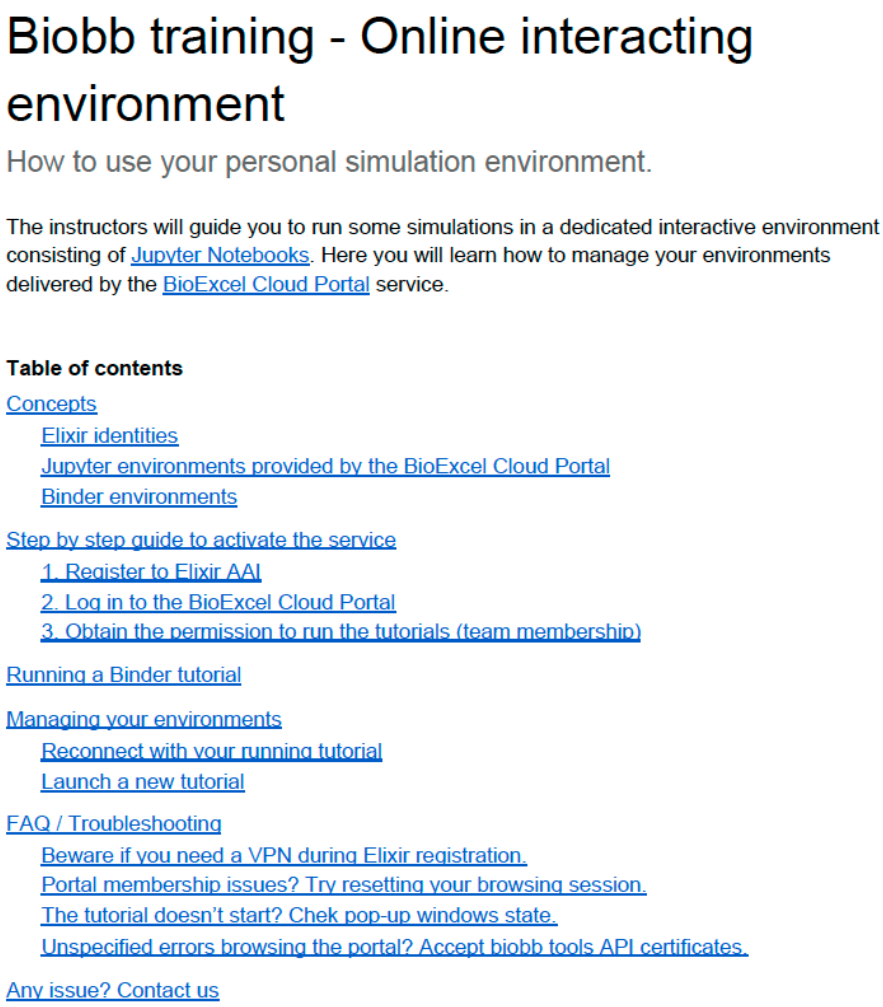
29 participants attended the course, of which 5 were females. 37.93% of them responded to the course feedback survey. All respondents rated the course as *Excellent* or *Very Good*. The comments received through the feedback survey were very positive, respondents liked the organisation of the course, the explanations, the tutorials, the expertise of the trainers and the interface that was demonstrated. On the other hand, some of them asked for more time for practicals and questions and answers sessions.

3.1.4. Computational biomolecular simulation workflows with BioExcel building blocks

After a first edition during our pilot phase of virtual courses in 2019, we ran two editions of this course, one in November 2020 and one in November 2021. Both editions consisted of three sessions of two hours each. This schedule was decided based on the feedback received in the first edition in 2019, which was organised in two sessions of three hours each.

The course is an introduction to the BioExcel building blocks, a fully interoperable software library comprising a collection of Python wrappers on top of biomolecular simulation tools such as GROMACS, Ambertools, Openbabel or ACPYPE. It was aimed at people starting computational biomolecular simulations and building computational biomolecular simulations workflows. Familiarity with molecular dynamics and minimum Linux skills were required to join.

The two editions of the course ran in the BioExcel Cloud Portal, using an exclusive instance of BinderHub installed in the EMBL-EBI OpenStack instance, the Embassy Cloud (see D2.5 - Provision of a Workflow Environment at BioExcel portal). The infrastructure behind the service, consisting of a Kubernetes cluster with a total of 96 cores and 192GB of RAM memory, allowed the deployment and execution of a Jupyter Notebook per user, with enough dedicated resources for each of them to run the whole workflow in real time. The platform was specifically designed to run training and workshops where users are expected to interact with Jupyter Notebooks. Participants received instructions on how to join the Portal before the event, so that they could follow the live hands-on sessions (Figure 20). Support from EMBL-EBI was available during the course to make sure that participants did not have any issue with the infrastructure.



Biobb training - Online interacting environment

How to use your personal simulation environment.

The instructors will guide you to run some simulations in a dedicated interactive environment consisting of Jupyter Notebooks. Here you will learn how to manage your environments delivered by the BioExcel Cloud Portal service.

Table of contents

Concepts
- Elixir identities
- Jupyter environments provided by the BioExcel Cloud Portal
- Binder environments

Step by step guide to activate the service
- 1. Register to Elixir AAI
- 2. Log in to the BioExcel Cloud Portal
- 3. Obtain the permission to run the tutorials (team membership)

Running a Binder tutorial

Managing your environments
- Reconnect with your running tutorial
- Launch a new tutorial

FAQ / Troubleshooting
- Beware if you need a VPN during Elixir registration.
- Portal membership issues? Try resetting your browsing session.
- The tutorial doesn't start? Chek pop-up windows state.
- Unspecified errors browsing the portal? Accept biobb tools API certificates.

Any issue? Contact us

Figure 20. Screenshot of the instructions sent to participants

Demonstration workflows implemented in Jupyter notebooks using the library were presented and explained step by step (e.g. protein-ligand complex GROMACS MD setup). The first steps on how to build and run complex workflows in command line were explored during the course.

Twenty participants were given a place in each edition of the course on a first-come, first-served basis. In total, 30 participants attended the course. The feedback responses are very similar for the two editions of the course, so we will report them together. In total, there were 18 responses. All of them rated the course as very good or excellent and everyone would recommend the course. When asked about the best of the course, most of them talked about the hands-on sessions and the interactivity of the course. Other responses included good organisation, instructions and good explanations. Among the worst parts of the course, some of them mentioned technical issues, while some others said that there was nothing bad. One of them said that it would have been good to have more time.

3.1.5. Introduction to HPC for Life Scientists

This is a course aimed at life scientists with little or no experience in the use of High-Performance Computing (HPC). The course is an introduction to HPC, focusing on the aspects that are most important for those new to this technology. It helps researchers judge how HPC can best benefit their work and equips them to go on to successfully and efficiently make use of HPC facilities in future. The course covers basic concepts in HPC hardware, software, user environments, file systems, and programming models. It also provides an opportunity to gain hands-on practical experience and assistance using an HPC system through examples drawn from the life sciences, such as biomolecular simulations.

This course had been run previously face-to-face; the last edition was in October 2019 and it was reported in D4.2 - *Training plan*. Since then, we have run two remote editions:

- Introduction to HPC for Life Scientists, 22 - 24 March 2021, organised together with PRACE and delivered in collaboration with ARCHER2 - the UK national supercomputing service.
- Introduction to HPC for Life Scientists, 31 January - 2 February 2022, organised in collaboration with PerMedCoE and PRACE. In this case, we changed parts of the programme to add examples from cell simulation tools developed by PerMedCoE, but the focus was still on the HPC aspects that are relevant for researchers.

In the following sections, we report on how these courses went.

Introduction to HPC for Life Scientists, March 2021

The course was attended by 20 participants, 5 of which were female. 60% of them responded to the feedback survey, where everyone rated the course as *Excellent* or *Very Good*. When asked what they liked about the course, some of them talked about specific sessions, such as QM simulations or GROMACS, but most of them focused on more general aspects of the course, especially on the materials and the practicals. Some also mentioned the schedule of the course and the breaks, and the approachability of the trainers. Among the aspects of the course that could be improved, the main one is the limited time for practicals, it would be good to have some more. Some respondents consider that there could be more practicals and that they could have more detailed instructions. In general, the feedback about the course was very positive, but we can improve in the course timings and how the practicals are delivered, so that everyone can follow.

Introduction to HPC for Life Scientists, January-February 2022

A total of 27 participants (17 females) followed the course and 10 of them (38%) completed the course feedback survey afterwards. On a scale where 5 was the maximum, all respondents rated the course overall as 4 or 5, and above 90% of the respondents answered 4 or 5 when asked about the event organisation, how much it had met their expectations, and how useful they had found the materials/exercises of the course for their work. All respondents considered that the length of the course, the depth of the content and the hands-on sessions were

adequate, and 80% considered that the pace of teaching was adequate. All of them would recommend the course to their colleagues and networks. Some of them suggested clarifying some of the basics at the beginning of the course. The use of a living document was appreciated by the participants and sticking to the time schedule was highlighted as important.

3.1.6. Remote GROMACS courses

We have run several courses and workshops aimed at current or future GROMACS users. Some courses are oriented to beginners, others to advanced users. All the courses have been run online (due to the pandemic) and most were organised in collaboration with PRACE, Euro-HPC center and European supercomputer centers. Most of the speakers are selected among active GROMACS developers.

The courses are a combination of lectures and online tutorials and cover a maximum of 3 topics (expected for some advanced course). The aim is to provide to the participants the tools to identify the best simulation methods to address their research questions and to set up simulations using GROMACS in the most appropriate and efficient way. The courses are characterized by having selected moments where the participants had the possibility to interact directly with the speakers. In online courses this has been successfully achieved using living documents (Google Docs or HackMD).

Introduction to GROMACS - A SNIC/PRACE workshop in collaboration with BioExcel, September 2020

The workshop aimed to provide the basis to perform molecular dynamics simulation. After attending the workshop, participants should have the knowledge and skills to be able to perform molecular dynamics simulations and simple free energy calculations using GROMACS.

The course was attended by 54 participants. 52% of them responded to the feedback survey, where the majority (93%) rated the course as *Excellent* or *Very Good*. When asked what they liked about the course, the majority wrote the tutorials and the jupyter-notebook implementation of the tutorials, some pointed to some specific lecturer and others enjoyed the opportunity they had to ask questions to the speakers. Among the aspects of the course that could be improved, some asked for more tutorials, others for a longer course covering more topics (e.g. on metadynamics, replica exchange methods and binding free energy of complex systems) and one asked for a better description of the target audience. In general, the feedback about the course was very positive, but we can improve by providing more courses covering a wider range of topics.

BioExcel/ENCCS workshop – Advanced topics in simulations with GROMACS, January 2021

In collaboration with EuroCC National Competence Centre Sweden, BioExcel gave a GROMACS workshop on umbrella sampling simulations, replica exchange molecular dynamics, accelerated weight histogram methods and GPU performance The workshop targeted advanced GROMACS users, required a pre-knowledge in performing simple simulations with GROMACS and provided the knowledge to perform simulations using advanced techniques using GROMACS. In

the course we introduced the practice of using quiz or polls during the lecture to increase the attendees' involvement.

The course was attended by 30 participants, of which 10 were female. 60 % of them responded to the feedback survey, where everybody rated the course as *Excellent* or *Very Good*. When asked what they liked best about the course, the majority wrote the lecture materials and the jupyter-notebook implementation of the tutorials, some enjoyed the use of "quiz" during the lectures/tutorials, others the break-out rooms to discuss deeper some topics with the trainers. Among the aspects of the course that could be improved, most asked for workshops covering other simulation topics, like QM/MM methods, metadynamics, electric field simulation and alchemical transformation using accelerated weight histogram method. But when they were asked what topic they found least interesting, the majority answered that all the topics were very interesting. Concerning what can be improved, some asked for more discussion time and other more guidelines for the breakout rooms, but overall all the comments were very positive.

***Advanced GROMACS workshop @CSC*, February 2021**

The course gave a practical introduction on how to run GROMACS MD simulations efficiently on modern hardware including both CPUs and GPUs. In addition, workflow automation and advanced sampling techniques were discussed. The event was organised in collaboration with BioExcel and CSC and supported by PRACE.

The workshop targeted advanced GROMACS users, and focused on hands-on exercises using CSC's supercomputers. The course was attended by 18 participants. Overall, the participants were extremely happy with the workshop, so we performed a similar event in February 2022. In general, the participants found very useful the material provided in advance (tutorial web pages, pre-recorded lectures, etc). Concerning what can be improved, some asked for more examples/exercises ("not just hit-enter-sequence") and others for more time for them to run the program and perform the exercises.

***GROMACS Workshop: A collaboration between BioExcel and national Euro-HPC competence center - Portugal*, September 2021**

The workshop covered an introduction to MD simulation, to accelerated weight histogram method and to GPU performance. At the end the participants should be familiar with the features of GROMACS code.

The workshop was attended by 68 participants, of which 27 were female. 40 % of them responded to the feedback survey, where the majority (96 %) rated the course as *Excellent* or *Very Good*. When asked what were the most positive aspects of the workshop, we got answers like: "I wanted to get a broad overview of Gromacs, which I achieved in the workshop" and "The possibility to ask live questions and the jupyter notebooks". Since we got positive feedback in the previous workshops about the tutorial in Jupyter Notebook format, we asked if they were able to run the tutorial during the workshop both online (via mybinder.org or the BioExcel portal) and offline (this required a local installation of miniconda). 55% answered positively, 18% answered that they had not tried yet.

***ONLINE: Advanced GROMACS workshop*, February 2022**

The workshop covered several topics that improve efficiency on using GROMACS: enhanced sampling techniques, QM/MM with CP2K, performance optimization for CPU and GPU, automation via BioBB-workflows and GROMACS Python-API. The workshop was organized as a collaboration between BioExcel CSC, ENCCS and it was supported by EuroCC.

The workshop was attended by 26 participants, of which 2 were female. 42 % of them responded to the feedback survey, where the majority (98 %) rated the course as *Excellent* or *Very Good*. The workshop targeted advanced GROMACS users. To avoid problems with inexperienced attendees, we provided introduction material on the workshop topics. Interestingly, 80% of participants rate his/her knowledge of the topic with a 3 on a scale 1 to 5. The workshop was run using an open on-demand interface that allowed easy access to the Jupyter Notebook environments and to the HPC resources for attendees and trainers. It was the first time that CSC used this setting. Some attendees found that the technical problems (that occurred) slowed down the workshop. But overall the attendees found the hands-on easy to follow. The attendees got the opportunity to have their own reservation on a CSC machine for the workshop days, but none of them made use of it. The attendees did not take the opportunity to ask questions live, but they preferred to be active in the living document (HackMD platform was used).

***Virtual course on free energy calculation and alchemical transformation in GROMACS*, March 2022**

This BioExcel short course (2 half days - morning) covered two methods to calculate the free energy associated with an alchemical transformation in GROMACS: Bennett acceptance ratio and accelerated weighted histogram methods. It alternates lectures and hands-on, based on Jupyter Notebook. Each participant had access to a virtual machine to run tutorials during the hands-on and to run some exercise in the afternoon. Some of them made use of the opportunity. Also, most of the attendees attended the hour reserved in the afternoon for discussion.

The course was attended by 27 participants, 10 of which were female. 41% replied to the feedback survey, where 91% of them rated the course as *Excellent* or *Very Good* and said that they would recommend the course. When asked about what was best about the course, most of them mentioned the tutorials, either regarding the balance between lectures and tutorials or the Jupyter Notebook as an easy way to follow them. Regarding things that we could improve, they suggested that there could be more time between the morning and the afternoon session, so that they can try more things. Some also asked for more content or in-depth topics. This could be done in a longer course.

3.1.7. EMBO practical course - Integrative modelling of biomolecular interactions. 30 May - 5 June 2021

This EMBO practical course on integrative modelling was sponsored by BioExcel. Initially it was planned to run in Izmir (Turkey) in May 2020. Due to the COVID-19 pandemic, it was postponed twice and finally ran remotely in 2021.

The course combined lectures and practical sessions on docking-based computational methods and related bioinformatics approaches, aiming at predicting how proteins interact with other biomolecules or ligands. Lectures provided the theoretical background on state-of-the-art algorithms for sampling and scoring docking models and described the use of low- and high-resolution information, and conservation- and coevolution-based interface prediction methods. The course also covered protein-peptide docking, data-driven molecular simulations, and other relevant topics in the field, including modelling of large assemblies and membrane proteins. During the practicals, participants worked with integrative modelling tools using real biological problems. To encourage interaction between all participants and stimulate discussions, students had the opportunity to present their own research, both in flash presentations and poster sessions, and bring their own research problems to dedicated troubleshooting sessions.

The remote course used a living document to provide information to participants: course information, software requirements, links to materials, lecture notes, etc. Slack was used during the course as a communication tool. The poster sessions were organised in Gather, to provide an interactive platform where people could talk to each other and discuss the posters in small groups.

The course was attended by 24 participants, of which 12 were female. 21 of them responded to the feedback survey. All respondents would recommend the course and all of them rated it as excellent or very good. Both the organisation and the different types of sessions were very well rated by most of the respondents. Lectures and practicals were mostly rated as *excellent.* Respondents found that the course was useful and instructive and that the instructors were enthusiastic and approachable. They valued the social activities and considered that the poster sessions in Gather allowed them to present their poster and discuss their research.

When asked about the most positive aspects of the course, participants included, among others: the lectures, the practicals, the relation among the lectures, the poster sessions and the "impressive organisation", in the words of one of them. There were also suggestions of topics to add, such as machine learning. One respondent mentioned that it would be better if the trainers visited the poster and discussed it with them. This could be improved by encouraging trainers to visit the posters.

3.1.8. PRACE Autumn School: Fundamentals of Biomolecular Simulations and Virtual Drug Developments 20 - 24 September 2021

This course was organised by BioExcel Centre of Excellence, STFC Daresbury Laboratory, NCSA Bulgaria, and Sofia University and in partnership with AstraZeneca and NostrumBiodiscovery. The course provided an introduction to the fundamentals of modern methods for biomolecular simulations - molecular dynamics (MD), 3D QSAR, integrative modelling, free energy and hybrid QM/MM calculations. Widely used software applications, including BioExcel software (GROMACS, HADDOCK, PMX, ChemShell) were presented, and their usage on HPC facilities was demonstrated. The target audience were early career scientists with a background in pharmacy, biochemistry and biophysics with little experience in

computational methodologies and High-Performance Computing (HPC). More information can be found on the event website:
https://events.prace-ri.eu/event/1222/

There were 165 attendees, of which 37 (22.4%) responded to the feedback survey. 94.6% of respondents rated the course as *Excellent* or *Very good*; the rest rated it as *Good*. 81% of respondents considered that the balance between theory and practice was good while 16.2% considered that the course was too theoretical and the rest thought that it was too practical. 89.2% of respondents would recommend the course; the rest replied *Maybe* to that question. Respondents consider that there could be more hands-on sessions in this course.

3.1.9. Training events for industry

BioExcel has engaged in three pilot projects with companies: AstraZeneca, Iktos and Adaptimmune. These pilot projects combined several products that are included in the BioExcel service catalogue. A summary of the projects can be found in D5.5 - *Final business plan*. They included training sessions on BioExcel Building Blocks and free energy calculations with PMX. These events built on the training experience developed in BioExcel through WP4. They were run as part of task 3.3 - *In-depth support*, so a detailed report about the events will be included in D3.7 - *User community support and engagement support (final)* that is due in June 2022.

3.1.10. Future training events

Two years after the COVID-19 pandemic started and forced us to do all of our events online, we are planning face-to-face events in the last months of BioExcel-2:

- HADDOCK workshop, organised in collaboration with the Pázmány Péter Catholic University, in Budapest, Hungary. 23-25 May 2022.
- BioExcel Summer School on Biomolecular Simulations, organised in Sardinia, Italy, with the support of Sardegna Ricerche. 12-17 June 2022.

3.2. Impact assessment

To ensure the quality of the training programme BioExcel routinely collects course feedback for all BioExcel training activities. The responses provide us with valuable information on how to improve our events and the direction of the training programme. However, these responses are generally collected on the last day of the event and therefore do not provide us with a lot of information about the long-term impact of the BioExcel training programme and tend to focus on immediate concerns (e.g. a problem with a tutorial, a session was too short or too difficult). To assess the impact of our training programme activities, we want to know if attending our training courses has had a positive impact on their research/career and whether or not they are actively using the tools that were introduced to them during the training event.

To assess the long-term impact, BioExcel sends two long-term feedback surveys to all delegates who have attended BioExcel training courses. The first is sent approximately 6-12 months after the training event and the second survey is sent after 2 years. The surveys collect data on whether participants are using the tools covered during the training course, whether they have established collaborations,

and whether they have taught others. In addition, we ask what impact the course has had on their research and the research of others, as well as what the best and worst parts of the course was.

Since the last deliverable (D4.3. – *Progress report and update on training and dissemination plan*), we have collected long-term feedback data from courses delivered between June 2018 and June 2021. We sent the first long-term feedback survey (6 months after the course) to the participants of the BioExcel Summer Schools of 2020 and 2021, the BioExcel Winter Schools of 2020, the Introduction to HPC for Life Sciences course of 2019 and 2021. The second long-term feedback survey (2 years after the course) was sent to the participants of the BioExcel Summer Schools on Biomolecular Simulations of 2018 and 2019, The BioExcel/PRACE Seasonal School of 2019, and the Introduction to HPC for Life Scientists course of 2019.

The response rate was as follows:

Table 10. Response rate to long-term feedback surveys, both 6 months (6m) and 2 years (2y) as indicated.

Course	**Number of participants who received the survey**	**Response rate**
BioExcel Winter School 2020 (6m)	30	50.0%
BioExcel Summer School 2020 (6m)	30	46.7%
Introduction to HPC for Life Sciences 2021 (6m)	20	25%
BioExcel Summer School 2021 (6m)	29	44.8%
Introduction to HPC for Life Sciences 2019 (6m)	12	16.7%
BioExcel Summer School on Biomolecular simulations 2018 (2y)	26	26.9%
BioExcel Summer School on Biomolecular Simulations 2019 (2y)	27	14.8%
BioExcel/PRACE Seasonal School 2019 (2y)	53	5.7%
Hands-on Introduction to HPC for Life Scientists 2019 (2y)	12	16.7%

The above response rates refer to the number of people that responded to at least one question of the survey and the true number of responses per question may vary.

3.2.1. Long-term feedback survey 6-12 months

Here we present the results of the survey analysing all responses to all surveys combined.

Many BioExcel training events teach participants how to use certain computational tools, so a direct impact that we can measure is whether they actively use these tools in their research. All respondents said they have applied the tools they learned about in their own research and have indicated which ones they have used, as shown in Figure 21 (note that they could provide more than one answer).

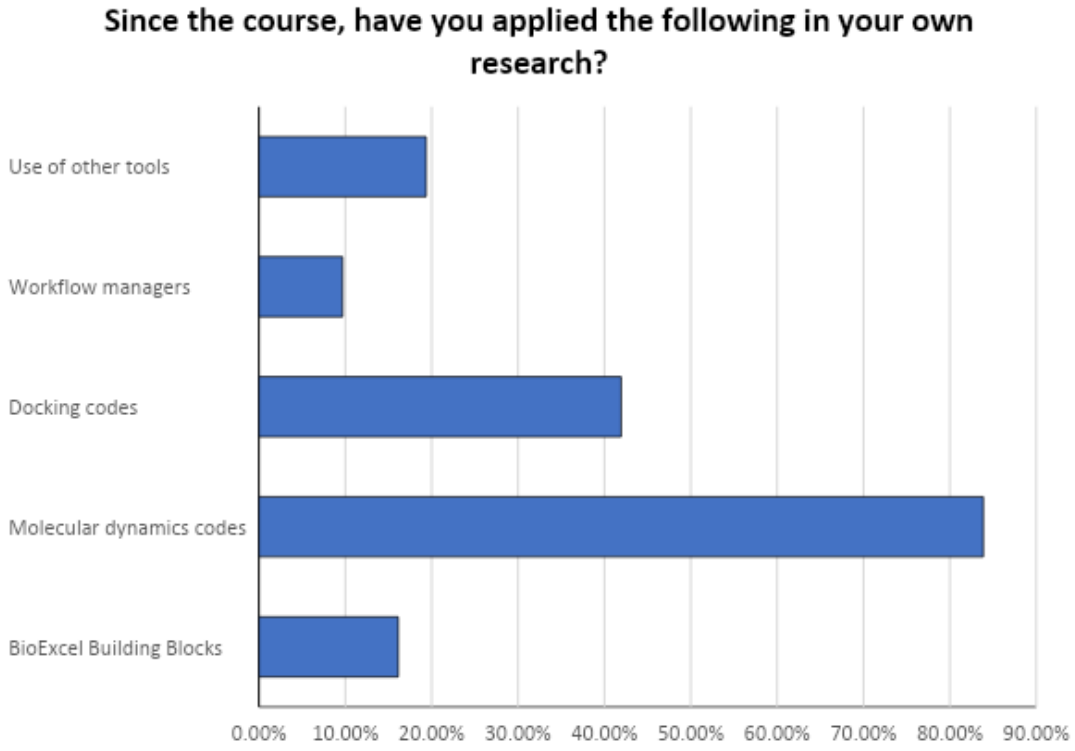


Figure 21. Percentage of respondents choosing the indicated answers to the question "During your course you will have learnt and used a variety of methods and resources. Since the course, have you applied the following in your own research?" Note that more than one answer may be selected.

One way in which our courses have an impact in the research community is that the participants share what they learned with colleagues and students. Most respondents (77%) have taught someone (Fig. 22). This means that our training reaches more people than just to those that attend our courses.

Figure 22. Percentage of respondents choosing the indicated answers to the question "How many others have you taught the skills and/or knowledge that you learnt during the course this year?"

Our training courses also promote networking, as they include time for participants to interact with each other and with the trainers during breaks and social activities. These interactions might lead to collaborations that will directly impact the research activities of the participants. In our long-term feedback survey we asked "Did you establish any collaborations with other participants / BioExcel researchers during this course?" Only 6% of respondents replied positively to the question, despite 47% of respondents replying positively to the same question in the previous deliverable (D4.3. – *Progress report and update on training and dissemination plan*). These numbers are not totally comparable because we added a new question to the survey: "Have you interacted with other participants or BioExcel researchers after this course?" 55% responded positively to this question, and some of them might have responded positively to the question about collaborations if this one had not been included. There could also be an impact on networking because of delivering the course virtually during the pandemic. In fact, one comment to the question was 'No but only due to real-life circumstances that prevented it; the scientific will to do so was there from both parties.'

We also ask our former participants if they would recommend our course or not with one option being if they have already recommended it. 97% of respondents to this survey would recommend the course they attended, with one person selecting 'maybe' and 74% have already recommended it to someone else. This speaks extremely well about the positive experience that they had during the BioExcel courses.

In addition to several pre-formulated questions, we also include the following in our survey: "Please comment on the impact that the course has had on your research and the research of others". To analyse the impact statements we try to

find common themes across the responses. We score each statement across the commonly identified themes. The same themes were used as in deliverable "D4.6 - Final report on dissemination and training" of the previous funding phase of BioExcel. Not all the themes in the list were matched by responses to the present survey.

Some participant statements contain more than one theme. The most commonly reported themes are "Gaining knowledge/skills" with 21 counts and "Essential to research/actively using skills" with 14 counts (Fig. 23). These two related themes indicate that our training is successful in transmitting participants important skills for their research. Following this, participants also indicated that they "taught others" as a major outcome from the course.

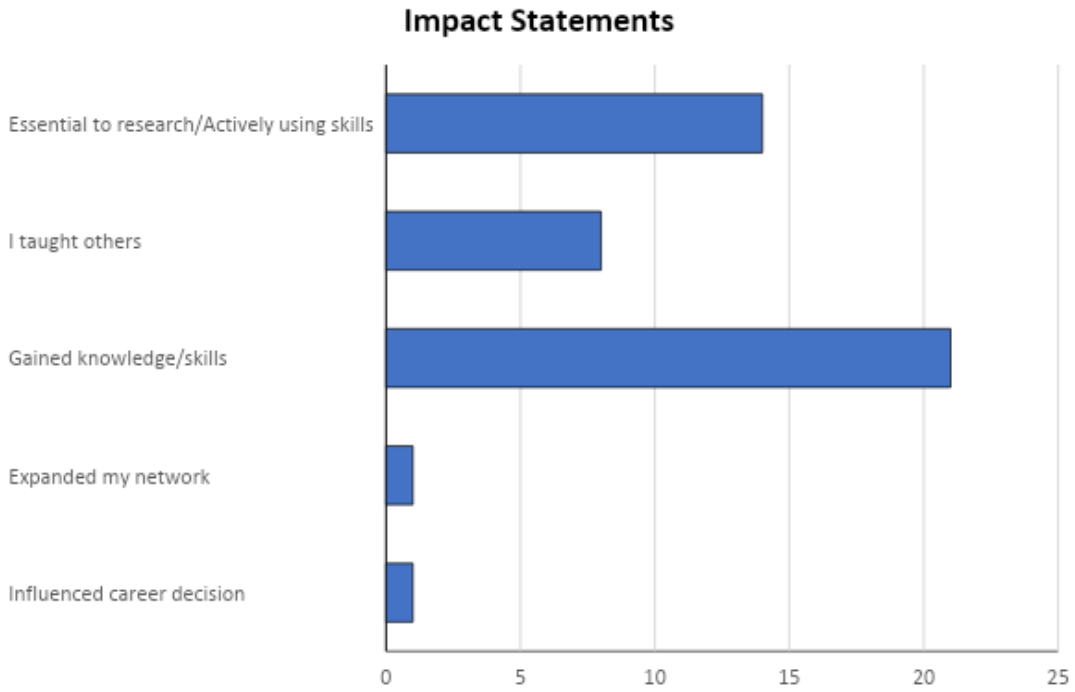


Figure 23. Analysis of the impact statements collected in the 2 year long-term feedback survey. Number of statements from participants that match the indicated themes.

3.2.2. Long-term feedback survey 2 years

We analysed the results of the 2 years impact survey collectively due to the low response rates of the surveys. A total of 15 completed surveys were analysed.

One person hadmoved away from their research area, whilst all other respondents (94%) had used the tools covered during the courses in their own research (Fig. 24). All of them would recommend the course and 88% have already done so. The majority of respondents (94%) have taught others the skills learnt during the course, showing that once again the training reaches beyond those that attend our courses and continues to do so some time after termination of the course.

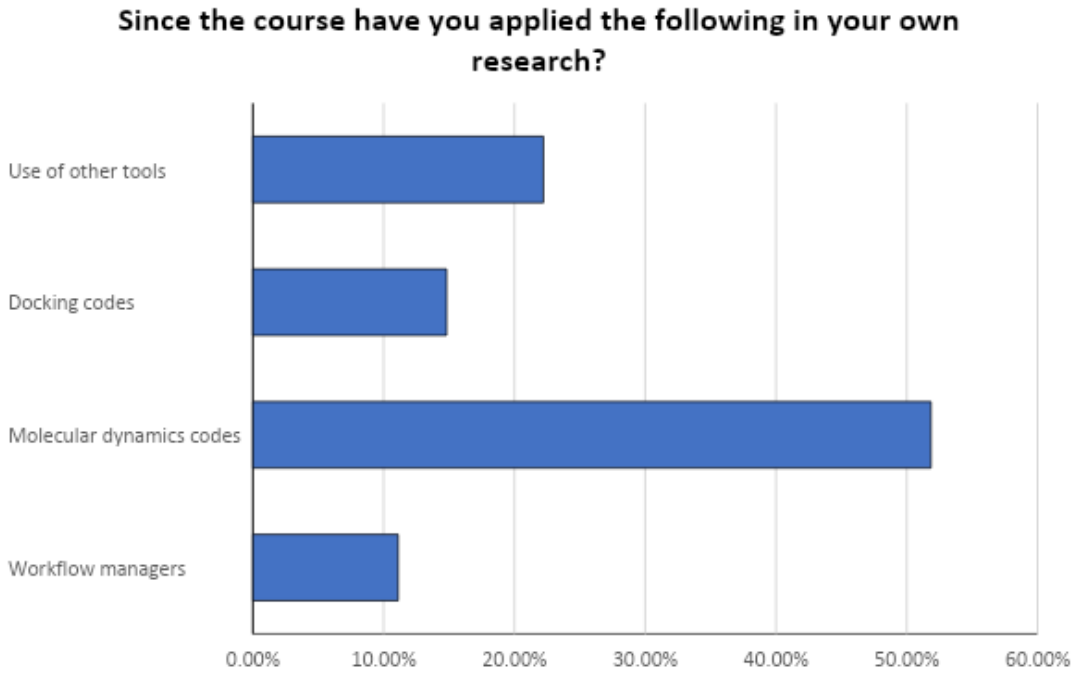


Figure 24. Percentage of respondents choosing the indicated answers to the question "During your course you will have learnt and used a variety of methods and resources. Since the course, have you applied the following in your own research?" Note that more than one answer may be selected.

As this feedback is sent after a longer time, we are interested in whether the participation in the course had an impact on their research. In addition to collaborations, here we ask about other outcomes (e.g. writing a dissertation, applying for a grant, writing a scientific article). 87% of the respondents answered positively to the question "Have you interacted with other participants or BioExcel researchers after this course?" and one third of respondents have established collaborations as a result of the course, which shows that networking during our courses can have this long-standing impact. 13% of them wrote a dissertation after their participation in the course and 53% published a paper. Regarding other possible outcomes, 27% replied *yes* to "Did you receive/apply for a grant after attending this course?".

In this survey, we also asked participants to comment on the impact on their research and this was analysed in the same way as the 6–12-month survey impact responses. The most commonly reported themes are "Essential to research/actively using skills" and "Gaining knowledge/skills" (Fig. 25).

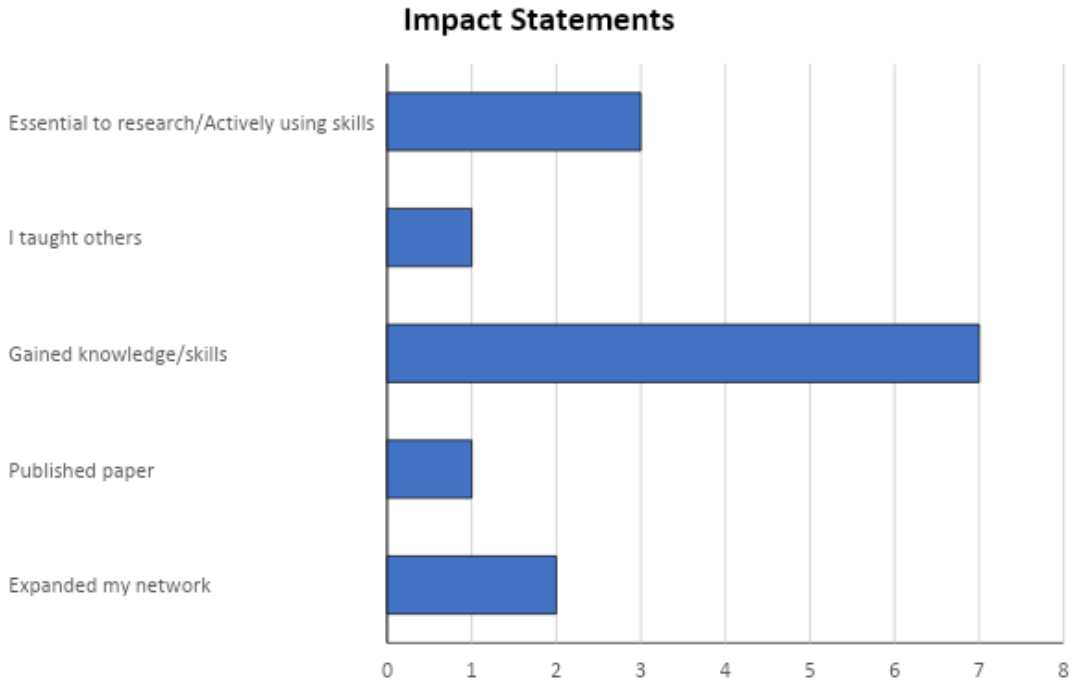


Figure 25. Analysis of the impact statements collected in the 2 years long-term feedback survey. Number of statements from participants that match the indicated themes.

3.3. Competency Hub

The BioExcel Training programme is based on a competency profile. This allows us to define the competencies that users of BioExcel need to develop to enable them to fully exploit the applications and services provided by BioExcel. The website developed at EBI to host competency profiles was called Competency Mapper. The name has been changed to Competency Hub, as the functionalities that it hosts have extended beyond the initial lists of competencies and training resources. In the last two years, BioExcel has developed a new version of the competency profile and added new content associated with it: career profiles and learning pathways.

3.3.1. Competency profile version 3

The BioExcel competency profile has been updated in collaboration with PerMedCoE, another Centre of Excellence focused on computational solutions for the life sciences. In this new version, there is a common set of 'computing' and 'parallel computing' competencies that address the competence requirements of both the BioExcel and the PerMedCoE user communities; and a domain specific group that is tailored to meet the distinct requirements of each user community and that has not been updated.

The BioExcel competency profile version 3 can be viewed in the Competency Hub. It includes one new competency in the 'computing domain': "Package and distribute software". The other competencies in the 'computing' and 'parallel computing' domains have been reviewed and edited so that they can be used by both BioExcel and PerMedCoE (Table 11). In addition, we improved the text of the competencies according to the feedback received from users of the competency

profile. The changes include rewording of some competencies and the rewording, addition or removal of the attributes included in them: knowledge, skills and attitudes. The complete profile with competencies and attributes is presented in Annex I. A list of changes can be found in the release notes of version 3.0 in the Competency Hub.

Table 11. BioExcel Competency profile, including 3 domains and 17 competencies.

Scientific domain
Apply expertise in formal & natural sciences appropriate to the discipline, and follow best practice in experimental design
User-driven service provision and support
Search for, assess and compile appropriate literature and data sets to address specific research questions
Comprehension of, and compliance with, best practice in data management / organization / archiving and storage and data management planning
Comprehension of how data-driven science, data analysis and computational modelling can be combined to generate and test hypotheses (e.g. machine learning, data mining, pattern recognition)
Computing domain
Evaluate the ability of a program running in a specific computing environment to perform a simulation (e.g. define algorithmic time and hardware resources required to solve a problem)
Operate effectively within a Linux environment
Write or adapt scripts and computer programs (software development) to perform simulations in compliance with good programming practice
Install or deploy pre-built software on a desktop or server computer
Acknowledge, and comply with, licensing policy
Monitor application execution
Package and distribute software
Parallel computing domain
Use a batch job system
Use computational workflow systems, understanding their potential benefits and limitations
Write parallel programs
Assess advantages and limitations for deploying, executing and optimising computations in a cloud/grid/HPC environment

Use performance profiling to identify bottlenecks and optimise the code

3.3.2. Career profiles

In the previous deliverable, D4.3. - *Progress report and update on training and dissemination plan*, we reported on the development of a module on the Competency Hub website that allowed us to create career profiles, in order to help people with their career development. These profiles list the background and activities of a reference persona in a specific role and rate them against the BioExcel competencies. At that time, we had published the first profile: junior software research engineer. Since then, we have added more career profiles to the BioExcel competency framework (Fig. 26):

- Senior research software engineer
- Junior computational chemist
- Senior computational chemist
- Research associate in biomolecular modelling
- PhD student in biomolecular simulations
- Principal investigator

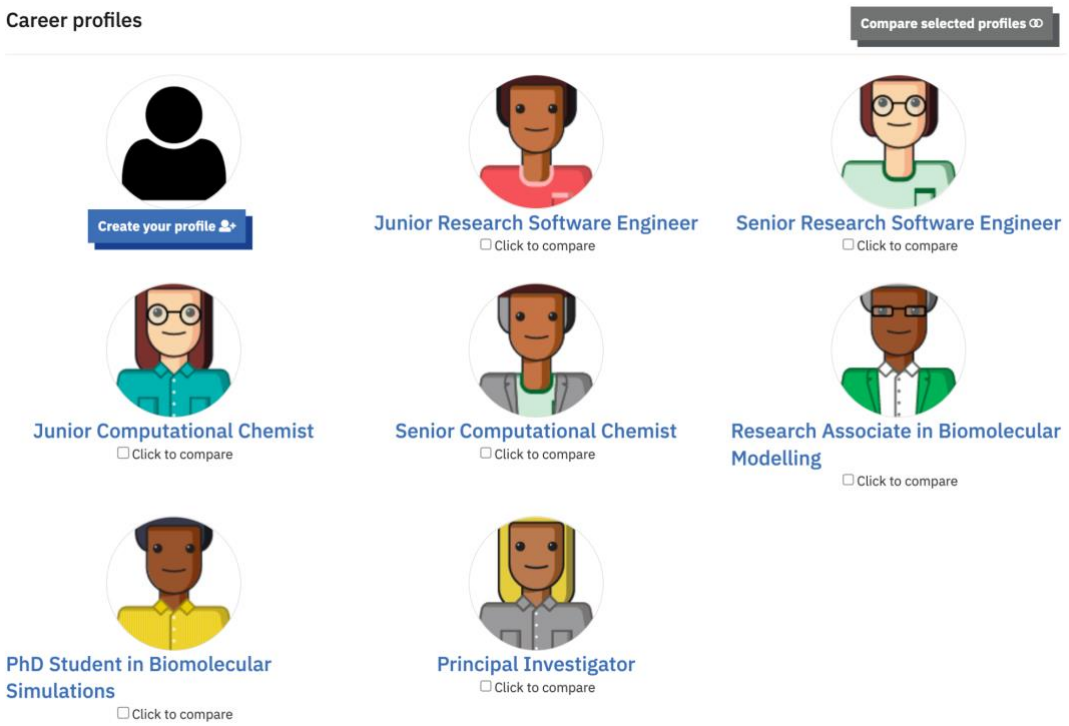


Figure 26. Screenshot of the career profiles page on the Competency Hub.

3.3.3. Learning pathways

The Competency Hub has also been updated to allow for the creation of learning pathways, collections of training resources that help people overcome a specific challenge or learn about a specific topic. The website development has been performed by EMBL-EBI in collaboration with PerMedCoE. BioExcel was the first

user of the Competency Hub to add a learning pathway: Using HPC infrastructure to perform biomolecular simulations (Fig. 27).

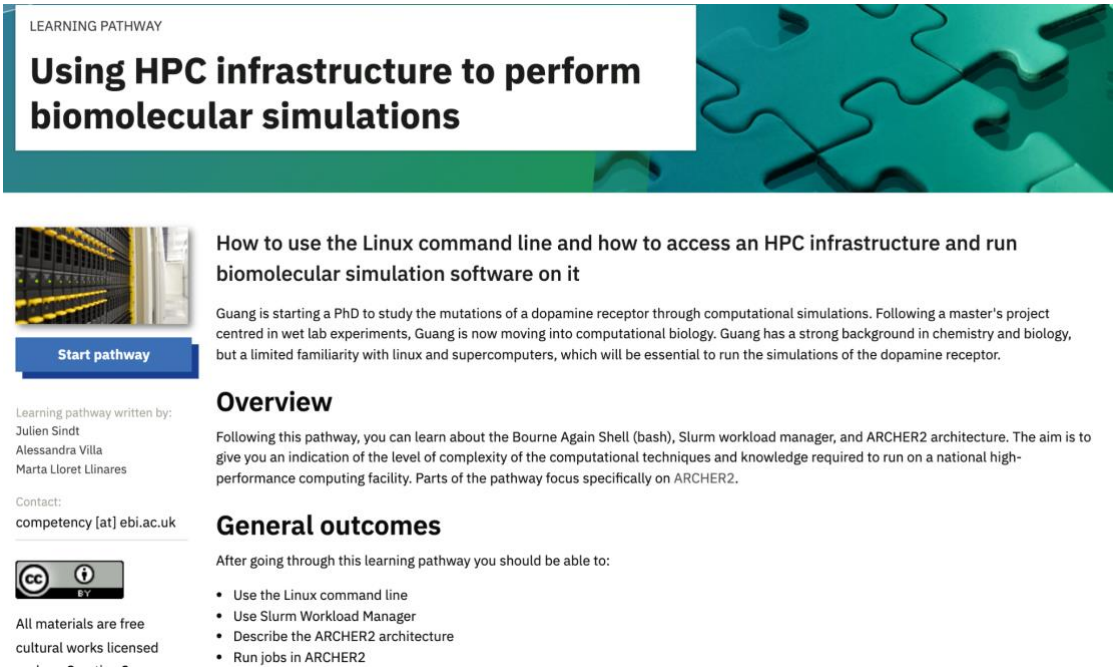


Figure 27. Cover page of the learning pathway "Using HPC infrastructure to perform biomolecular simulations" developed by BioExcel.

This is a learning pathway aimed at people who have never accessed an HPC infrastructure. It consists of three modules: The Unix Shell, Introduction to the Slurm workload manager (Fig. 28), and ARCHER2 system and modules. Some parts of the content are specific for the ARCHER2 machine, but most of it is independent of the machine. Learners can start the pathway in the place that is more interesting for them. It contains links to other resources and quizzes to check the knowledge. As it is hosted in the Competency Hub, it is possible to check which BioExcel competencies are related to each of the training resources included in the pathway.

A user feedback session on this learning pathway was organised with participants of BioExcel courses. They liked the idea of the learning pathway, a site with all the resources and quizzes needed to learn about a specific topic. The general information on the website was clear and easy to understand. However, the navigation of the site was not straightforward and in some cases there were two different places with the same link, which confused the participants. This was improved in the next iteration of our website development.

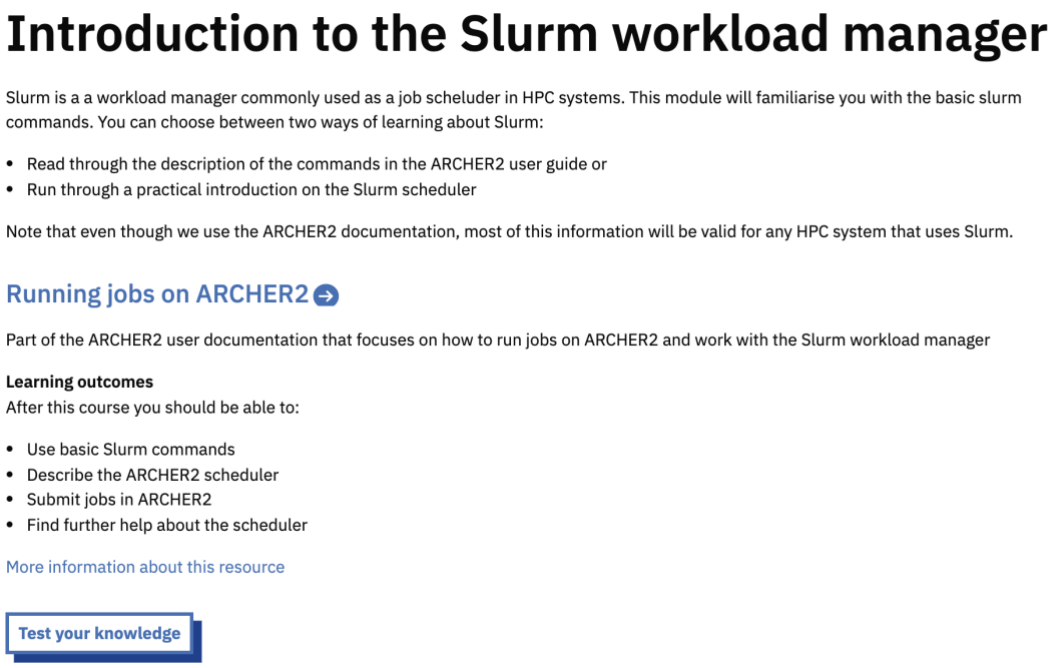


Figure 28. Screenshot of one of the modules of the learning pathway on using HPC to run simulations

We have now published a second learning pathway: Using Jupyter notebooks for biomolecular research. This pathway is aimed at early career researchers with no experience of Jupyter Notebook and aims to introduce the student to the basics of using notebooks, IPython, and Markdown; how to share the notebooks with others; and good practice when using the application. There are branches in the pathway for different programming languages and notebook sharing options so that the learner can tailor the content to their background and experience. Finally, there is a focus on showing how they can be useful for their research by including a live BioBB tutorial within a Jupyter Notebook. Most of the content itself is hosted within a Jupyter notebook so that the learner will get hands-on practice as they proceed through the pathway which will enhance their learning.

4. Conclusion

Between June 2020 and March 2022, BioExcel continued to engage with the biomolecular simulations community through dissemination and training activities. Our dissemination strategy allowed us to increase our number of followers and promote our research and events. In October 2021, we successfully delivered the BioExcel flagship event: the EMBO Workshop *Advances and Challenges in Biomolecular Simulations*.

The BioExcel training programme continued to be delivered in a virtual format and we kept gaining experience to run courses where participants can interact with each other and with the trainers, in addition to learning during the lectures and the hands-on tutorials. Part of our courses have been organised in collaboration with other initiatives, such as PRACE, PerMedCoE or some of the National Competence Centres for HPC. This allows us to reach a wider audience.

In addition to the training events, BioExcel contributes to the development of the skills in the community through the publication of training material, the creation of learning pathways and the addition of training resources and reference profiles to the BioExcel competency framework.

Annex I - The BioExcel competency profile v3

The tables below show version 3.0 of the BioExcel competency profile. Each of the tables corresponds to a domain: scientific competencies, computing competencies and parallel computing competencies.

Scientific competencies			
Competencies	**Knowledge**	**Skills**	**Attitudes**
Apply expertise in formal & natural sciences appropriate to the discipline, and follow best practice in experimental design	- Has a deep comprehension of biological problems - Comprehends that a biological theory is not necessarily true - Fundamental scientific knowledge (biology, chemistry, physics, mathematics) - Comprehends the need for positive and negative controls, and replicates - Comprehends that models require experimental validation	-Has an interdisciplinary view - Asks relevant, hypothesis driven, well-defined biological questions - Accurately judges the validity of his/her results	- Takes a comprehensive approach to problems - Looks for prior work (e.g. literature and public datasets) - Displays scientific humbleness (own and other results) - Allows for unexpected results - Comprehends that results might be difficult to interpret
User-driven service provision and support	- Recognises different types of users, their interests, motivation, activities and workflows - Awareness of customer support practices and training best practice - Knowledge of how own service fits in with broader service landscape - Deep understanding of data protection principles and implementation protocols	- Manages expectations effectively - User experience skills, including user requirements/needs collection and appreciation of how to differentiate requirements from needs - User training/facilitation/support skills	- Is proactive in discovering different users and their motivation - Actively facilitates gathering of use cases and user requirements, anticipates user problems - Seeks out and acts on user feedback to improve the user experience and increase satisfaction - Service/User-oriented mindset

Scientific competencies			
Competencies	**Knowledge**	**Skills**	**Attitudes**
Search for, assess and compile appropriate literature and data sets to address specific research questions	- Is aware of the concept of unique identifiers - Comprehends the need to question / validate data	- Assesses quality (e.g. methods, data, results) Judges fit-for-purpose - Tracks where data came from - Combines data from different sources Is able to find relevant literature and datasets	- Advertises the existence of good quality datasets to others - Contributes to public datasets when appropriate
Comprehension of, and compliance with, best practice in data management / organization / archiving and storage and data management planning	- Knows and understands the data formats people are using - Is aware of the backup policy of the institution/compute resource - Is able to judge what data needs to be kept for storage - Comprehends the need for standards, ontologies and metadata - Comprehends the challenge surrounding legacy data (e.g. inaccessibility, resourcing, IP)	- Demonstrates knowledge of the file system structure - Tracks ever step in a process (traceability) - Automates data analyses where appropriate - Structure data, where applicable submit to databases - Safely and efficiently moves data	- Actively promotes and uses existing standards, ontologies, and metadata annotation - Documents and creates backups for his/her work in non-redundant ways - Has a version control system in place - Adds his/her data in supplementary information of journal articles (e.g SMILES, InChI for compounds)
Comprehension of how data-driven science, data analysis and computational modelling can be combined to generate and test hypotheses (e.g. machine learning, data mining, pattern recognition)	- Statistical and mathematical modelling methods, and key scientific and statistical analysis software packages - General data science approaches to life science problems, such as machine learning and artificial intelligence - Experimental design to ensure the statistical validity of high-throughput experiments	- Is conscious of the risks of overfitting and of appropriate methods for cross-validation and control of overfitting - Designs appropriately powered experiments to answer the research question	- Appreciates the importance of statistics in experimental design, data analysis and interpretation - Approaches problems with a systems-based, data-driven approach to scientific discovery

Computing competencies (C)			
Competencies	**Knowledge**	**Skills**	**Attitudes**
Evaluate the ability of a program running in a specific computing environment to perform a simulation (e.g. define algorithmic time and hardware resources required to solve a problem)	- Capabilities and limitations of computer-based systems, processes, components and programs - Software requirements of the program - Optimal hardware to run the program - Where to find information about the level of trust of the system and its suitability for work with sensitive data - How to describe algorithmic performance and complexity (e.g. Big O notation)	- Evaluates applicability, optimisation and scope of different tools before choosing - Is able to benchmark a program to get an estimate of the required computing time - Identifies bottleneck in the program (network, CPU, RAM, etc) - Assesses when to move data, and when to run code elsewhere	- Keeps up-to-date with emerging techniques and applications - Actively searches for available sources of support (training material, forum, helpdesk, etc.) - Prototypes/implements the design solution and verifies performance against specification - Checks output correctness according to software documentation - Identifies new technological opportunities
Operate effectively within a Linux environment	- How to efficiently navigate their way around the Operating System (OS), including the file system - Where to find the location of important configuration files & applications - Most common CLI tools / programs (grep, find, du, etc.)	- Creates and manages files and directories in a system - Is able to use Unix/Linux features like pipes & redirection - Changes access permissions when required - Searches effectively for files and content - Reads and edits files without a GUI - Writes scripts to automate and/or facilitate actions when appropriate	- Searches proactively for support on the web - Creates backup files - Uses access permissions appropriately - Is up to date with new bug fixes and security patches from periodic Linux updates
Write or adapt scripts and computer programs (software development) to perform simulations in compliance with good programming practice	- Common programming concepts like loops and function calls - The pros and cons of different scripting and programming languages - Existing tools and libraries to reuse - Where to find examples or written guidelines regarding best practice in their field - The importance of writing an optimised code	- Reads, adapts and debugs existing programs or scripts - Judges when a task should be automated or scripted - Compiles code when appropriate - Designs and structures efficient and portable scripts and programs - Writes and runs appropriate tests - Uses a compiled / high-performance language (e.g. C++, Julia) in contrast to	- Understands the need for best practice - Conforms, and inspires others, to best practices for writing reusable code - Uses revision control, including code review - Uses an editor with support for programming - Uses appropriate scripting or programming language

		scripting/interpreted language (Python/Bash) when needed	
Install or deploy pre-built software on a desktop or server computer	- Existing repositories and revision control systems (e.g. Git, SVN, mecurial-versioning) - The stages of software release cycles - How to install software and dependencies - The dependencies of the software - The difference between kinds of binaries & source - The different types of account (e.g. user, admin) and when they should be used - Command line use	- Selects appropriately packaged code - Uses package managers - Is able to revert a system to a known state - Debugs (Mission Control, VisualVM,..) and tests software - Accesses servers remotely	- Consults the manual - Checks licensing before installing or running software - Appreciates the impact of installing or running software on other users of the machine - Debugs and tests non-commercial software that is built collaboratively on a volunteer basis
Acknowledge, and comply with, licensing policy	- Types of licensing and differences between them - Significance and potential ambiguity of licensing depending on where you work (e.g. academia/industry, funding) - The difference between open and closed licenses	- Chooses appropriate license for their own software - Identifies tricky or unclear licenses and seeks expert help - Keeps track of licensing terms of dependencies	- Always checks the license information for software - Questions ambiguity in license information
Monitor application execution	- The relevant metrics to monitor (e.g. queue monitoring, storage used) - Which metrics require special permissions to monitor - A variety of monitoring tools and their usage model (e.g. has to be started with the application, only targets a single process)	- Recognises bottlenecks in the program (Network, CPU, RAM, etc.) - Identifies the current state of a process - Checks the resource utilisation (e.g., memory, CPU) for a process	- Uses resources efficiently - Actively considers impact on the platform and on other users
Package and distribute software	- The different uses for source and binary distributions - Several packaging technologies (e.g. RPM, wheels, conda-packages) and distribution channels (e.g. distribute from your own GitHub, create a conda channel, submit code to some	- Submits a package to some upstream repository - Automates their software packaging - Designs a build pipeline	- Writes a how-to-do guide from the user perspective - Writes software with packaging requirements in mind and focus on reusability

	upstream repository) and how to package software for them - The components required to set up a pipeline for automatic software packaging - Main points when writing a tool documentation or tutorial - When to package as a library vs. stand-alone executable - Container technologies (e.g. Docker, singularity)	- Writes clear documentation providing installation instructions and adequate examples to illustrate the use of their tool	

Parallel computing competencies (PC)			
Competencies	**Knowledge**	**Skills**	**Attitudes**
Use a batch job system	- Concept of queues and the runtime environment - Relevant CLI parameters of the available queue system - Versions, incompatibilities and interdependencies of installed software - How to monitor job progress and be notified of events - Basics of computer architecture	- Integrates queue system commands in those scripts with high computational demands - Can estimate what resources are required per job and verify or adjust the requirements after a completed job - Avoids usage which taxes the batch job system too much - Can estimate which software is responsible for the observed issues (e.g. is it my software or theirs which is causing the issue?) - Identifies storage requirements and utilises reusability of the intermediate data objects/analysis step results	- Seeks optimal utilisation of the available HPC resources - Monitors resource consumption, checks allocation on smaller tasks before submitting a big one
Use computational workflow systems, understanding their	- The existence and functionality of workflow systems	- Evaluates and selects an appropriate workflow system - Installs workflow managers	- Seeks out and makes use of appropriate existing workflows and workflow components

potential benefits and limitations	- Advantages and disadvantages of different workflow systems - When (and when not) to implement workflow tools - How workflow systems affect system usability and stability	- Is able to run a workflow - Writes and modifies a workflow - Identifies the best resources to run the workflow (e.g. cpu, gpu, hybrid)	- Keeps in mind portability and reusability when developing a workflow
Write parallel programs	- Common parallel programming paradigms and technologies (e.g. shared memory vs distributed memory, task-based parallelism, OpenMP, MPI, PGAS) - Common bottlenecks in parallel programming (e.g. false sharing, static scheduling) - Different system architectures and how they impact parallel programming	- Recognises (independent) units of work and potential parallelism and any dependencies - Uses appropriate tools to debug, profile, refactor parallel code - Selects the appropriate level of parallelism abstraction considering available frameworks and native language constructs - Writes the code avoiding the most typical bottlenecks	- Makes use of automated benchmarking and regression test to monitor the performance on different types of architecture - Actively uses profiling tools to detect parallel bottlenecks
Assess advantages and limitations for deploying, executing and optimising computations in a cloud/grid/HPC environment	- Concepts related to virtualisation - Use cases that are limited by computational infrastructure and would benefit from HPC and/or HTC technologies - Upsides and downsides of the different environments	- When needed, undertakes a cost-benefit analysis to understand whether cloud resources are the best value - Accurately judges added value of HPC/HTC technologies for different scenarios and user groups - Is able to package software/data and deploy to the cloud - Selects an appropriate grid or cloud provider - Can build and provision a Virtual Machine (VM)	- Seeks out and makes use of appropriate existing virtual machines that are suitable for the scientific problem to be solved - Evaluates whether a specific use case can benefit from the use of HPC or HTC
Use performance profiling to identify bottlenecks and optimise the code	- Tools to help measure performance - How to recognise the most common types of performance bottlenecks (memory, CPU or IO bound)	- Is able to profile a code - Refactors code in order to remove bottlenecks - Is able to decide when to optimise an existing algorithm or when to change the algorithm itself	- Runs programs using appropriate resources (number of processors, etc.) - Actively collects and analyses logging & performance information - Foresees possible bottlenecks during the development

	- How to interpret the results of performance profiling tools to modify the original code to diminish the impact of identified bottlenecks		