Plagiarism – “Bagnall’s” sktime proposal

This report is about documented plagiarism found in Anthony Bagnall’s 2021 grant proposal “sktime”, in specific comparison with Franz Kiraly’s 2019 grant proposal “sktime”, on which Anthony Bagnall was a named contributor.

The “sktime” grant of Bagnall was awarded with the name “sktime” (EP/W030756/1), but later renamed by Bagnall and UKRI, so the name is no longer linked to the original (see section “renaming of the grant”).

Links to the proposals as submitted: (Bagnall 2021), (Kiraly 2019)

Executive Summary

The 2021 proposal - submitted by Bagnall - is basically a copy of Kiraly’s 2019 proposal, up to increased pointers to maturity of sktime in 2021 due to its growing user base, increased verbosity, less focus on technical detail, and minor variations in application cases described.

There is substantial rephrasing in the 2021 proposal, but the content is near-identical. The inherent plagiarism cannot be disputed, given a joint inspection of the two proposals.

Given that Kiraly is neither PI nor co-PI, there are no conceivable circumstances of Bagnall’s 2021 submission that could mitigate the inherent plagiarism.

Note about involvement/mention of Kiraly in 2021 grant

Kiraly is mentioned in the 2021 grant, but this is misleading:

- **Kiraly is not a PI or co-PI in the 2021 submission**, as one may be misled to believe by reading the body of the 2021 grant. What “counts” is the PI/co-PI field in the UKRI submission, which has only the three co-PIs – Renoult, Sami, Sambrook (see “original public record” at the bottom)
- Kiraly is mentioned with a minor role in the grant application, despite having been PI of the sktime project since its inception, see history of this file: https://github.com/sktime/sktime/blob/main/docs/source/get_involved/code_of_conduct.rst
- Kiraly would have been eligible as a PI or co-PI at the time of submission

What happened, according to Kiraly, is **a last-minute removal as a PI or co-PI**, and further alterations to minimize Kiraly’s role, after joint preparation of the grant text.

This would already be problematic even without substantial overlap between proposals.

Side-by-side comparison

Proposed research and context

2021 proposal	2019 proposal
Techniques for learning from time series have been developed in a wide range of disciplines, including: statistics; machine learning; signal processing; econometrics; and finance.	Learning with time series and temporal data is crucial to many applications, across wide areas of research in engineering, finance, health, the natural science, and many others.
Each discipline has a favoured set of tools and accepted workflows. Despite similarity between tasks, the development and evaluation of algorithms has traditionally been siloed.	Open source capabilities in dealing with such data is limited, leading to unnecessary replication of coding work,
Moreover, tasks are frequently reduced from one type of problem (e.g. forecasting) to another (e.g. regression, classification or clustering) and this commonly happens without reference to the state-of-the-art algorithms developed specifically for the new task.	or technically inappropriate (and therefore error-prone) reduction to cases off-shelf toolboxes can deal with (e.g., tabular data). […] for example, a forecasting strategy may be constructed by tabulating sliding windows and applying a time series regression method;
[…]	
But despite the ubiquity of time series data, no such framework exists for machine learning with time series.	Unlike for "classical" supervised learning, there is not a "one-interface-fits-all" approach,
Frameworks like sktime not only offer reusable functionality, but also provide overall structure to application code. They capture common design decisions and distil them into reusable templates that practitioners can copy. This reduces the number of decisions practitioners must take and allows them to focus on application specifics. Not only can practitioners write software faster as a result, but applications will have a similar structure.	and it necessitates development of a meta-language for model building and checking. […] The "optimal" solution to the issue would provide an interoperable system for modelling strategies and pipelines for each of the above, […] Can one define a (user-friendly) first-order type language for model building?
We provide a common platform to define and formalise multiple time series tasks such as forecasting, classification, clustering, regression, annotation, anomaly detection and change point detection as well as reduction approaches between them.	(i) Supervised learning with time series features, including time series classification and regression […] (ii) Time series annotation (supervised and unsupervised), including anomaly detection, segmentation […] (iii) Forecasting (supervised and unsupervised) […] Primary requirements for such a set of interoperable "time series" modelling toolbox modules are: (a) Availability of modelling atoms under a task-specific unified interface that exposes hyper-parameters and inference results […] (b) Abstract composition methodology for tuning, pipeline building. […] (c) Abstract reduction methodology, i.e., meta-learning that mutates the task. […]
Since its inception in 2018, sktime has become an established toolkit for time series analysis used world-wide by academics and industry alike. Whilst demand is steadily growing, sktime is increasingly facing bottlenecks in its maintenance activities. It has been without any dedicated funding since 2019	This proposal suggests to build on outcomes of the sktime project (Dec 2018 – May 2019) in order to complete the above vision, leveraging an expanding community of contributors, network of application case studies, co-development with SPF themes, and a consolidated code base that covers the basics of use case (i), supervised learning with time series features, in the

and its operations are currently entirely driven by volunteers.	form of a python toolbox compatible with the sklearn and pydata ecosystems.
This project will allow sktime to continue to sustain and grow its operations, by providing dedicated maintenance resources.	An expanded scale, and a sufficient amount of manpower, would put use cases (ii) - (iv) in reach of development (see "research").
It will also allow us to further enhance sktime's functionality and have impact on new scientific and industrial user communities.	In addition, we anticipate close interlinkage and co-development with projects in the health programme (see "impact").

Aims and objectives, programme and methodology

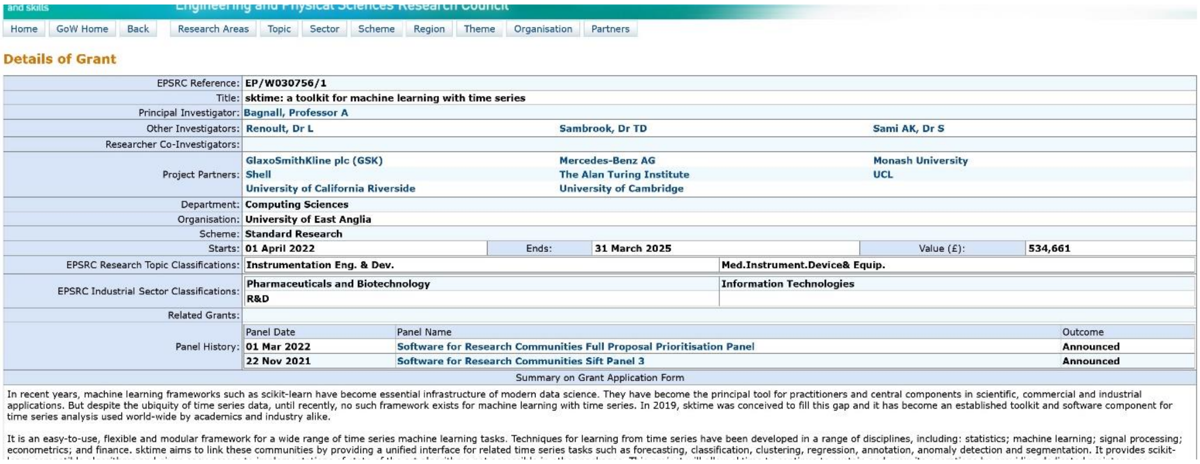
2021 proposal	2019 proposal
1. Maintenance and community building (WP1) to improve the process and speed of conducting essential maintenance activities and widen participation in the maintenance of sktime.	The sktime team are keen to develop better communication channels both internal and external to the Turing.
2. Extend functionality (WP2) to oversee and steer the development of new state-of-the-art functionality by the wider sktime community.	Our goal is for sktime to facilitate the rapid development of good solutions to a range of problems using state-of-the-art algorithms […] there are a number of research challenges within the tooling and methodology development domains that the project will need to address
3. Enhance scientific workflows (WP3) by applying sktime to problems arising in two specific research communities within the EPSRC remit.	[…] used to add value to existing research, and in turn inform development of the toolkit. We have a series of agreements with domain experts […] The toolkit will also play a central role in two EPSRC project proposals that will be submitted in 2019.
new user communities focused on medical and healthcare topics through dedicated companion packages	In addition, the health programme has confirmed it support for co-development and deployment for tasks
sktime already provides consistent interfaces for a number of Python libraries for time series analysis, including scikit-learn, statsmodels, tslearn, tsfresh and fbprophet. We collaborate with the maintainers of these libraries and will continue working towards defining standard interfaces for different learning tasks, with the aim of improving usability and interoperability of the ecosystem as a whole.	(1) Design of unified object oriented interfaces for modelling strategies: within the key tasks (i)-(iv), and across the tasks. Design of a data-task-strategy interface.
We will hold a series of events to help widen participation in the development and maintenance in academia, industry and the wider Python community. Academic participation will be encouraged through continued research collaborations, conference tutorials and publications of new results that showcase the functionality of sktime	We will instigate a formal mechanism for information sharing between development teams at the Turing to help foster a sense of community and spread best practice. We will set up regular surgery activities to allow ongoing projects to ask us what we could do for them, and to request guidance in dealing with time series data.
Industry applications will provide valuable feedback. The requirements and design of new features will be formalised through consultation with our industry partners. Their public support will encourage other	leveraging an expanding community of contributors, network of application case studies, co-development with SPF themes, and a consolidated code base that covers the basics of use case

industry supporters to become involved. They will provide feedback from using sktime in production in industry, guidance on new features and priorities and software domain expertise. Annual hackathons will help forge links between project partners and widen engagement.	
WP2 involves oversight of the broadening of the functionality of sktime to impact current and new user communities. The scope of the functional improvements is ambitious, but we envisage that the majority of the code will be provided by the open-source community	In addition, project development is open on GitHub, and we aim to further integrate with members of the pydata user and developer community, or Turing projects who consider project outputs useful.
The **forecasting module […]** The **classification module […]** The **transformation module […]** The **clustering module […]** The **regression module […]** The **annotation module […]** he new **change point detection module […]**	(i) Supervised learning with time series features, including time series classification and regression […] (ii) Time series annotation (supervised and unsupervised), including anomaly detection, segmentation […] (iii) Forecasting (supervised and unsupervised) […]
WP3 is concerned with impacting scientific communities that fall under the EPSRC remit to enhance their workflow by using sktime. This deepening of the reach of sktime will be achieved through collaborations with domain experts in two fields: signal processing for magnetoencephalography and electroencephalography (M/EEG) analysis; and exploratory analysis of data from healthcare technologies	Event classification in neuroscience, physics, object/motion recognition […] intensive care and medical monitoring, equipment health monitoring […] electronic health records, clinical studies with survival outcome, behaviour modelling, predictive maintenance
We will promote reproducible research and, by providing standard pipelines and access to state of the art algorithms, facilitate more effective and efficient research workflows.	In addition to that, reproducible practices projects (such as the Turing Way) are natural partners as modelling toolboxes providing the basis for reproducible analyses through standardized code and workflow components.
The MRC Cognition and Brain Sciences Unit of the University of Cambridge have accumulated a unique database of resting-state MEG data from approximately 150 patients with Mild Cognitive Impairment (MCI) – a potential prodromal stage of dementia – plus over a 150 age- and sex-matched controls.	MRC Cognition & Brain Sciences Unit of the University of Cambridge (LoS1, LoS2.pdf) Early detection of dementia from MEG/EEG – multivariate time series classification
The Collaborative Healthcare Innovation through Mathematics, EngineeRing and AI (CHIMERA) project [EP/T017791/1] is a collaborative hub based at UCL and partnered with the Turing, Great Ormond Street Hospital (GOSH) and University College London Hospitals NHS Foundation Trust.	Great Ormond Street Hospital (LoS3.pdf) Predictive modelling using Intensive Care Unit data – panel data Prediction

- ➔ Paraphrases, same idea for work packages
- ➔ Even same-identical collaboration partners

Renaming of the grant “sktime”, EP/W030756/1

Original public record



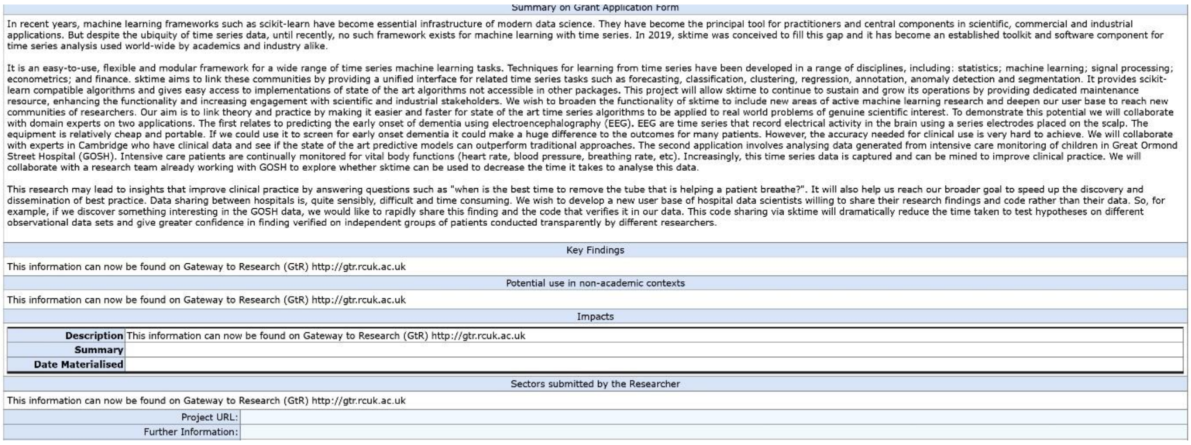
Details of Grant

EPSRC Reference:	EP/W030756/1		
Title:	sktime: a toolkit for machine learning with time series		
Principal Investigator:	Bagnall, Professor A		
Other Investigators:	Renoult, Dr L	Sambrook, Dr TD	Sami AK, Dr S
Researcher Co-Investigators:			
Project Partners:	GlaxoSmithKline plc (GSK) Shell University of California Riverside	Mercedes-Benz AG The Alan Turing Institute University of Cambridge	Monash University UCL
Department:	Computing Sciences		
Organisation:	University of East Anglia		
Scheme:	Standard Research		
Starts:	01 April 2022	Ends: 31 March 2025	Value (£): 534,661
EPSRC Research Topic Classifications:	Instrumentation Eng. & Dev.	Med.Instrument.Device& Equip.	
EPSRC Industrial Sector Classifications:	Pharmaceuticals and Biotechnology R&D	Information Technologies	
Related Grants:			
Panel History:	Panel Date	Panel Name	Outcome
	01 Mar 2022	Software for Research Communities Full Proposal Prioritisation Panel	Announced
	22 Nov 2021	Software for Research Communities Sift Panel 3	Announced

Summary on Grant Application Form

In recent years, machine learning frameworks such as scikit-learn have become essential infrastructure of modern data science. They have become the principal tool for practitioners and central components in scientific, commercial and industrial applications. But despite the ubiquity of time series data, until recently, no such framework exists for machine learning with time series. In 2019, sktime was conceived to fill this gap and it has become an established toolkit and software component for time series analysis used world-wide by academics and industry alike.

It is an easy-to-use, flexible and modular framework for a wide range of time series machine learning tasks. Techniques for learning from time series have been developed in a range of disciplines, including: statistics; machine learning; signal processing; econometrics; and finance. sktime aims to link these communities by providing a unified interface for related time series tasks such as forecasting, classification, clustering, regression, annotation, anomaly detection and segmentation. It provides scikit-



Summary on Grant Application Form

In recent years, machine learning frameworks such as scikit-learn have become essential infrastructure of modern data science. They have become the principal tool for practitioners and central components in scientific, commercial and industrial applications. But despite the ubiquity of time series data, until recently, no such framework exists for machine learning with time series. In 2019, sktime was conceived to fill this gap and it has become an established toolkit and software component for time series analysis used world-wide by academics and industry alike.

It is an easy-to-use, flexible and modular framework for a wide range of time series machine learning tasks. Techniques for learning from time series have been developed in a range of disciplines, including: statistics; machine learning; signal processing; econometrics; and finance. sktime aims to link these communities by providing a unified interface for related time series tasks such as forecasting, classification, clustering, regression, annotation, anomaly detection and segmentation. It provides scikit-learn compatible algorithms and gives easy access to implementations of state of the art algorithms not accessible in other packages. This project will allow sktime to continue to sustain and grow its operations by providing dedicated maintenance resource, enhancing the functionality and increasing engagement with scientific and industrial stakeholders. We wish to broaden the functionality of sktime to include new areas of active machine learning research and deepen our user base to reach new communities of researchers. Our aim is to link theory and practice by making it easier and faster for state of the art time series algorithms to be applied to real world problems of genuine scientific interest. To demonstrate this potential we will collaborate with domain experts on two applications. The first relates to predicting the early onset of dementia using electroencephalography (EEG). EEG are time series that record electrical activity in the brain using a series electrodes placed on the scalp. The equipment is relatively cheap and portable. If we could use it to screen for early onset dementia it could make a huge difference to the outcomes for many patients. However, the accuracy needed for clinical use is very hard to achieve. We will collaborate with experts in Cambridge who have clinical data and see if the state of the art predictive models can outperform traditional approaches. The second application involves analysing data generated from intensive care monitoring of children in Great Ormond Street Hospital (GOSH). Intensive care patients are continually monitored for vital body functions (heart rate, blood pressure, breathing rate, etc). Increasingly, this time series data is captured and can be mined to improve clinical practice. We will collaborate with a research team already working with GOSH to explore whether sktime can be used to decrease the time it takes to analyse this data.

This research may lead to insights that improve clinical practice by answering questions such as "when is the best time to remove the tube that is helping a patient breathe?". It will also help us reach our broader goal to speed up the discovery and dissemination of best practice. Data sharing between hospitals is, quite sensibly, difficult and time consuming. We wish to develop a new user base of hospital data scientists willing to share their research findings and code rather than their data. So, for example, if we discover something interesting in the GOSH data, we would like to rapidly share this finding and the code that verifies it in our data. This code sharing via sktime will dramatically reduce the time taken to test hypotheses on different observational data sets and give greater confidence in finding verified on independent groups of patients conducted transparently by different researchers.

Key Findings

This information can now be found on Gateway to Research (GtR) http://gtr.rcuk.ac.uk

Potential use in non-academic contexts

This information can now be found on Gateway to Research (GtR) http://gtr.rcuk.ac.uk

Impacts

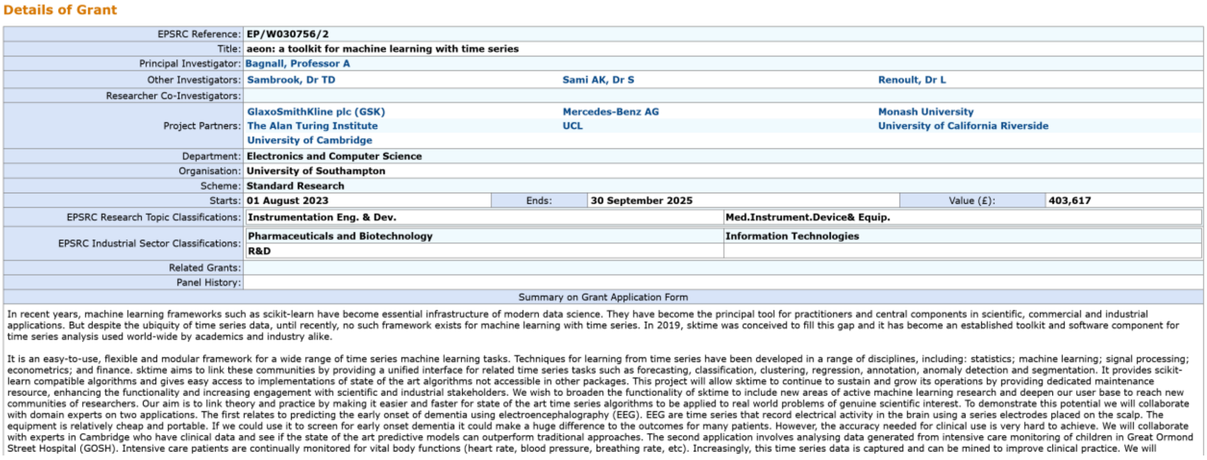
Description	This information can now be found on Gateway to Research (GtR) http://gtr.rcuk.ac.uk
Summary	
Date Materialised	

Sectors submitted by the Researcher

This information can now be found on Gateway to Research (GtR) http://gtr.rcuk.ac.uk

Project URL:	
Further Information:	

After renaming by Bagnall and UKRI



Details of Grant

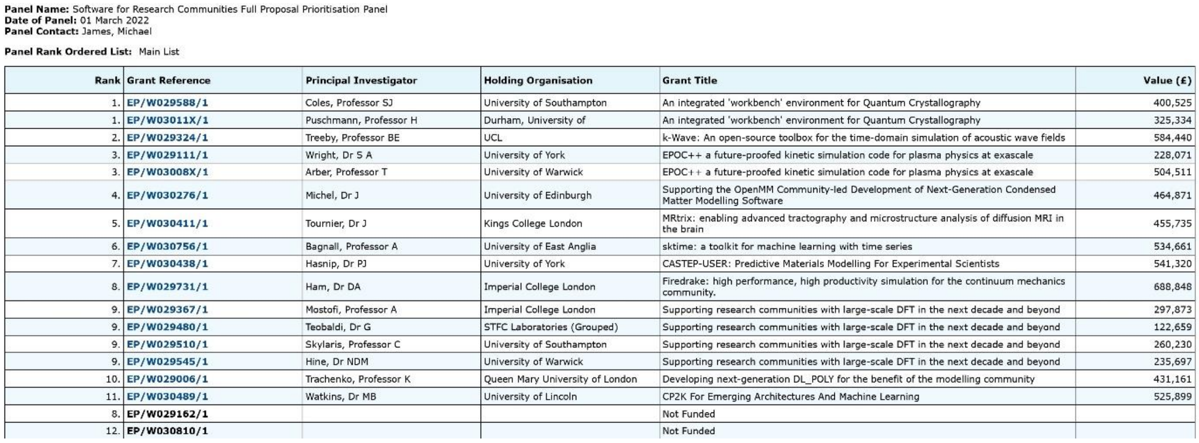
EPSRC Reference:	EP/W030756/2		
Title:	aeon: a toolkit for machine learning with time series		
Principal Investigator:	Bagnall, Professor A		
Other Investigators:	Sambrook, Dr TD	Sami AK, Dr S	Renoult, Dr L
Researcher Co-Investigators:			
Project Partners:	GlaxoSmithKline plc (GSK) The Alan Turing Institute University of Cambridge	Mercedes-Benz AG UCL	Monash University University of California Riverside
Department:	Electronics and Computer Science		
Organisation:	University of Southampton		
Scheme:	Standard Research		
Starts:	01 August 2023	Ends: 30 September 2025	Value (£): 403,617
EPSRC Research Topic Classifications:	Instrumentation Eng. & Dev.	Med.Instrument.Device& Equip.	
EPSRC Industrial Sector Classifications:	Pharmaceuticals and Biotechnology R&D	Information Technologies	
Related Grants:			
Panel History:			

Summary on Grant Application Form

In recent years, machine learning frameworks such as scikit-learn have become essential infrastructure of modern data science. They have become the principal tool for practitioners and central components in scientific, commercial and industrial applications. But despite the ubiquity of time series data, until recently, no such framework exists for machine learning with time series. In 2019, sktime was conceived to fill this gap and it has become an established toolkit and software component for time series analysis used world-wide by academics and industry alike.

It is an easy-to-use, flexible and modular framework for a wide range of time series machine learning tasks. Techniques for learning from time series have been developed in a range of disciplines, including: statistics; machine learning; signal processing; econometrics; and finance. sktime aims to link these communities by providing a unified interface for related time series tasks such as forecasting, classification, clustering, regression, annotation, anomaly detection and segmentation. It provides scikit-learn compatible algorithms and gives easy access to implementations of state of the art algorithms not accessible in other packages. This project will allow sktime to continue to sustain and grow its operations by providing dedicated maintenance resource, enhancing the functionality and increasing engagement with scientific and industrial stakeholders. We wish to broaden the functionality of sktime to include new areas of active machine learning research and deepen our user base to reach new communities of researchers. Our aim is to link theory and practice by making it easier and faster for state of the art time series algorithms to be applied to real world problems of genuine scientific interest. To demonstrate this potential we will collaborate with domain experts on two applications. The first relates to predicting the early onset of dementia using electroencephalography (EEG). EEG are time series that record electrical activity in the brain using a series electrodes placed on the scalp. The equipment is relatively cheap and portable. If we could use it to screen for early onset dementia it could make a huge difference to the outcomes for many patients. However, the accuracy needed for clinical use is very hard to achieve. We will collaborate with experts in Cambridge who have clinical data and see if the state of the art predictive models can outperform traditional approaches. The second application involves analysing data generated from intensive care monitoring of children in Great Ormond Street Hospital (GOSH). Intensive care patients are continually monitored for vital body functions (heart rate, blood pressure, breathing rate, etc). Increasingly, this time series data is captured and can be mined to improve clinical practice. We will

Grant award panel

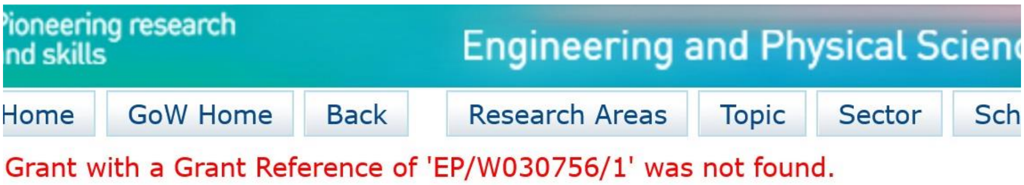


Panel Name: Software for Research Communities Full Proposal Prioritisation Panel
Date of Panel: 01 March 2022
Panel Contact: James, Michael

Panel Rank Ordered List: Main List

Rank	Grant Reference	Principal Investigator	Holding Organisation	Grant Title	Value (£)
1.	EP/W029588/1	Coles, Professor SJ	University of Southampton	An integrated 'workbench' environment for Quantum Crystallography	400,525
1.	EP/W03011X/1	Puschmann, Professor H	Durham, University of	An integrated 'workbench' environment for Quantum Crystallography	325,334
2.	EP/W029324/1	Treeby, Professor BE	UCL	k-Wave: An open-source toolbox for the time-domain simulation of acoustic wave fields	584,440
3.	EP/W029111/1	Wright, Dr S A	University of York	EPOC++ a future-proofed kinetic simulation code for plasma physics at exascale	228,071
3.	EP/W03008X/1	Arber, Professor T	University of Warwick	EPOC++ a future-proofed kinetic simulation code for plasma physics at exascale	504,511
4.	EP/W030276/1	Michel, Dr J	University of Edinburgh	Supporting the OpenMM Community-led Development of Next-Generation Condensed Matter Modelling Software	464,871
5.	EP/W030411/1	Tournier, Dr J	Kings College London	MRtrix: enabling advanced tractography and microstructure analysis of diffusion MRI in the brain	455,735
6.	EP/W030756/1	Bagnall, Professor A	University of East Anglia	sktime: a toolkit for machine learning with time series	534,661
7.	EP/W030438/1	Hasnip, Dr PJ	University of York	CASTEP-USER: Predictive Materials Modelling For Experimental Scientists	541,320
8.	EP/W029731/1	Ham, Dr DA	Imperial College London	Firedrake: high performance, high productivity simulation for the continuum mechanics community.	688,848
9.	EP/W029367/1	Mostofi, Professor A	Imperial College London	Supporting research communities with large-scale DFT in the next decade and beyond	297,873
9.	EP/W029480/1	Teobaldi, Dr G	STFC Laboratories (Grouped)	Supporting research communities with large-scale DFT in the next decade and beyond	122,659
9.	EP/W029510/1	Skylaris, Professor C	University of Southampton	Supporting research communities with large-scale DFT in the next decade and beyond	260,230
9.	EP/W029545/1	Hine, Dr NDM	University of Warwick	Supporting research communities with large-scale DFT in the next decade and beyond	235,697
10.	EP/W029006/1	Trachenko, Professor K	Queen Mary University of London	Developing next-generation DL_POLY for the benefit of the modelling community	431,161
11.	EP/W030489/1	Watkins, Dr MB	University of Lincoln	CP2K For Emerging Architectures And Machine Learning	525,899
8.	EP/W029162/1			Not Funded	
12.	EP/W030810/1			Not Funded	

Renaming occurred on or around Aug 27, 2023



Grant with a Grant Reference of 'EP/W030756/1' was not found.