The Alan Turing Institute

# Alan Turing Institute Data Science for Science Programme: Molecular Biology

## Call for pilot projects

The Alan Turing Institute (Turing) invites pilot projects from multidisciplinary teams of researchers working in partnership, to apply data science approaches to biology. These Guidance Notes contain information about the call as well as essential guidance for applicants.

As part of the Turing Institute's Data Science for Science Programme, we are developing research around areas of key importance including molecular biology. This call for pilot projects is fully funded by the UKRI's Strategic Priorities Fund programme through the AI for Science theme area. This research theme will develop the crucial role for Data science and Artificial Intelligence (AI) in answering many fundamental questions in genomics and biology regarding the structure and function of genomes, the molecular interplay within cells and organs, the role, shape and functions of proteins as well as advancing imaging techniques.

**Closing date for applications: 17th January 2020**

**Background**

In the ''Omics'' era of the life sciences, data represents information at various levels of biological systems, including the genome, transcriptome, epigenome, proteome, metabolome, molecular imaging, molecular pathways, different populations of people, etc.

Each of these fields has benefited from recent improvements to sensor technology, including the rise of DNA sequencing instruments, super-resolution digital microscopy and magnetic resonance imagery used for studying biological systems with greater resolution than ever before. The second major advance in the field has been the increased importance of large-scale multicore computing systems, advanced search and indexing algorithms as well as numerical optimization and modeling techniques.

As we become more capable at interpreting genomes and monitoring molecular changes, we will be able to answer many fundamental questions regarding the structure and function of genomes, the molecular interplay within cells and organs and the properties of entire species and ecosystems.

For this we need more robust, expressive, computable, quantitative, and precise ways to handle the big data. We need to overcome challenges related to the intrinsic complexity, biases and heterogeneity of genomic sequencing data due to error rates, experimental batch effects and different statistical models applied. We also need to develop novel techniques for data collection, organization and

integration of biological data that facilitate quick storage, easy re-processing and reproducible research as well as integrating hypothesis driven research with data driven methods.

## Scope

This call aims to catalyse new collaborations between biologists, statisticians and data scientists. It is designed to enable groups of researchers with complementary skills and expertise to explore opportunities at the nexus of biology and data science research. Pilot projects are invited from multi-disciplinary teams for short, 6-month scoping and prototyping projects. The scoping projects are expected to lay the foundation of more substantial proposals for funding to UKRI and other funding bodies.

Proposals must be innovative and interdisciplinary in nature, utilise existing data (that is fully consented and anonymised) and clearly explain the potential impacts of the work for computational biology. This call for pilot projects has five focus areas, namely:

· **Imaging**

Our current ability to process, visualize, integrate and analyse this new generation of dynamic and multidimensional imaging data has become a critical challenge in advancing cell biology. This gap creates an opportunity to develop new imaging analysis tools for cross model predictions; create data driven simulations for hypothesis testing; predict 3d structures from 2d images; automate data quality assurance and error correction; use machine vision to track single cells/molecules; and use full experimental information to inform clustering.

· **Genomics**

Recent technological advances have increased the mechanistic understanding of genome biology to an incredible degree. However, the complexity and sheer amount of information contained in DNA remain roadblocks to complete understanding of all functions and interactions of the genome. We are looking for projects that connect genotype to phenotype, improve the prediction of regulatory functions, create embeddings for genomic data and classify mutation types among others.

· **Single-cell biological data**

Recent technological developments have led to a rapid expansion in the size and range of available single-cell biological data. Single-cell techniques can generate data sets that are high-dimensional and also of very large sample size (large n and p). These data sets allow the detailed study of heterogeneity and noise in biological systems at the cellular level. However, separating the multiple sources of variation from technical artefacts is significant challenge, typically tackled using Bayesian hierarchical modelling. Understanding temporal dynamics and incorporating spatial information are especially challenging.

· **Protein shape and function**

Even though we now know the sequences and the structures of proteins from many important model organisms, less than 1% of these proteins have been experimentally characterized. In addition,

metagenomics initiatives are identifying millions of bacterial proteins found in human hosts such as the gut microbiome. We are looking to support projects that improve the prediction of the structure and functions of the proteins, build more and better quality 3D models, improve the accuracy of protein function predictions, classify metagenomic samples where there is lack of references; shed light in the energy landscape of protein dynamics (e.g. folding, catalysis, regulation, etc) as well as the various protein folding mechanisms.

· **Molecular systems biology**

Molecular systems biology is concerned with connecting fundamental biological mechanisms with cellular system function and behaviour. Advancing the field will require the development of coherent strategies for multi-omics data integration, as well as new methods for modelling and measuring system dynamics. Techniques for inferring static and dynamic networks of interaction are often utilised, and causal inference can be useful for identification of molecular mechanisms.

## Applicant guidance

Proposals should represent an inter-disciplinary team of biologists and data scientists and be a collaboration between at least two Institutions, either from the Turing University Network or from the Turing University Network and external Research organisations and scientific agencies. Please note that two teams from the same place would not be accepted as collaborators. All proposals must be led by two Investigators. The PI (Lead Applicant) must be either a Turing Fellow, Turing Research Fellow, Turing Visiting Researcher or situated at one of the Turing's university partners.
Additional Investigators (co-applicants) who are not directly involved with the Turing or its university partners may participate alongside a PI who is affiliated with Turing as described above.

Eligible additional co-applicants will be based at Higher Education Institution (HEI) or Independent Research Organisation (IRO) and must be eligible to receive UKRI funding (https://www.ukri.org/funding/how-to-apply/eligibility/). Organisations who are not eligible to receive UKRI funding can participate as a project partner providing they cover all of their own expenses.

Applications should demonstrate how the complementary expertise and experience in the two institutions are relevant to the project. In this first call, up to two 6 month projects are intended to be supported. An award of up to £50,000 will be offered to each selected proposal. We plan to fund further projects in subsequent years.

## Support available

Applicants should provide a costing covering 100% of direct costs for projects of up to 6 months in duration. Part of the aim of the call is to generate results quickly. Awards may include:

- Up to 20% of investigator time (e.g. 4 investigators at 5%) costed as Directly Incurred for the purposes of this grant.
- Other Directly Incurred staff salaries (e.g. existing academic staff, research assistants or other research staff, technicians and other support staff). o Due to the short time frame of the projects and the requirement that projects will end on 30th September 2020, we would recommend naming available researchers at the point of application to allow projects to start quickly. o PhD studentships cannot be supported through this call.
- Overhead rate of £32,500 per 1 researcher FTE, per annum.
- Research equipment essential for the project.
- Travel and subsistence, and other meeting costs where relevant.

## Application process

The call opens on 15th November 2019. Please discuss all applications with your host institution/department ahead of submission. Typically institutions have an internal deadline of at least 5 working days prior to the funder deadline so take this into consideration when preparing an application.

Applications should be completed and submitted by 17:00 GMT, 17th January 2020 to the Turing's FlexiGrant portal: https://ati.FlexiGrant.com with the following documents:

- A copy of your proposed research case for support (two pages maximum, plus a maximum of one additional page for references)
    - This should include background and a description of the proposed research, aims and objectives, tools and methods, relevance and beneficiaries
- A fully completed costing using your Institution's costing tool
    - This should include a budget table (template provided) and narrative justification of resources.

## Additional FlexiGrant sections:

a) Project summary: Please provide details of your proposed research project including its objectives, key research questions, impact and relevance, and how the project would be developed into a long-term collaboration between the two universities/Institutions. (Max. 300 words)

b) Project team: List the team members, who should represent different disciplines and areas of expertise. Specify the expected contributions from each person (i.e. who will do what?) and the strategy for achieving synergy across the project (Max. 300 Words)

c) Expected Outputs and Impact: Please provide information on the outputs expected from this project. These should include items such as anticipated papers, international conference presentations, applications made to external funds (including suggestions of possible sources) and impact. (Max 400 words)

## Details of the project team

Although projects must be led by at least two Investigators, the application form should be submitted from the FlexiGrant account of the Lead Applicant. Co-applicants, collaborators and research management staff can be added as collaborators to FlexiGrant.

Applications will be assessed by peer review at the Alan Turing Institute and approved by the Strategic Priorities Fund Management Board. Applicants will be notified of their award in August 2019.

Key dates

Opening date for applications 15th November 2019 (via FlexiGrant system)
Closing date for applications 17th January 2020
Review panel meeting 12th February 2020
Results communicated 18th February 2020*
Project start date No later than 1st April 2020 *

Depending on the level of interest received, it may not be possible to offer detailed feedback on unsuccessful applications.

## Assessment Criteria

All pilot projects will be assessed by a multidisciplinary panel, chaired by Professor Jon Rowe and should consider each of the following areas.

**Research excellence**

- Does the proposal demonstrate novelty and added value in relation to the current state of knowledge within the field?
- Does the proposal make use of an appropriate methodology, acknowledging that this may not be your own preferred methodology?
- Are the activities detailed within the proposal well planned and scientifically feasible?

**Research team**

- Does the lead applicant and, where relevant, the broader research team possess the necessary capability to deliver the proposed project?
- Are the track record(s) appropriate and does the team possess the necessary balance of skills?
- Are all team members named and ready to commence the project without recruitment delays?

**Impacts and outcomes**

- Has the proposal demonstrated plans for engagement within and outside of academia? Do the researchers have a plan for impactful outputs?

- Are the impact activities clearly described and are the potential outcomes and impacts realistic and appropriate to the type of research undertaken and the scope of the project?

- Are any opportunities for future collaboration and leveraged funding outlined and, if so, how realistic are they?
- Note that outputs and impacts do not have to be academic papers, and broader impact activities are encouraged.

**Resources and value for money**

- Are the resources requested to carry out the proposal clearly defined and justified?
- Are they realistic and appropriate given the scope of the project?
- Does the research team have guaranteed and immediate access to any datasets and/or facilities essential to the quick commencement and ultimate success of the project?

**Contact**

For questions regarding the application process or other elements of the call, please contact: datascienceforscience@turing.ac.uk.

*This work is supported by The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1*

UK Research and Innovation