Engineering and Physical Sciences Research Council

Brunel University of London

1 (a) Title: Dr Surname: Dantu First name and middle initials: Sarath C

(b) Department name: Computer Science

(c) Have you supervised Vacation Internships before?

No. I have supervised three interns: two funded through departmental summer internships and one through PDC.

RESEARCH PROJECT

2 (a) Title of project: (no more than 220 characters)

Discovering the Building Blocks of Protein Dynamics with AI

(b) Description of the proposed project (no more than 700 words) outlining:

i) Background to the project;
ii) Aims and objectives. Any key hypotheses to be tested or questions to be asked. What you hope to achieve during the period of research;
iii) Methods experimental design and methods;
iv) Brief outline of a timetable of work.
Please note that continuation of undergraduate projects will not be considered.

Engineering and Physical Sciences Research Council

Brunel University of London

Background

Proteins are dynamic molecular machines, and their biological function often depends on coordinated internal motions rather than static structure alone. Molecular dynamics (MD) simulations provide a route to study these motions, but they generate large, high dimensional datasets that are difficult to interpret manually. This project will contribute to the development of a computational framework for rational protein design by identifying fundamental building blocks of structural dynamics in proteins. We refer to these local recurrent dynamic environments as Protein Dynamic Units (PDUs). This preliminary data analysis will support future **EPSRC and BBSRC Responsive mode** applications on AI enabled rational protein design.

Aims and objectives

The main aim is to test whether local dynamic environments in proteins can be represented in a way that reveals reusable structural dynamic motifs across different protein families. The specific objectives are to:

1. extract local PDU candidates from mdCATH trajectories;
2. compute structural and dynamic features for each PDU;
3. convert PDUs into vector embeddings suitable for AI analysis;
4. train autoencoders to learn compact latent representations;
5. use UMAP and clustering methods to visualise and identify recurrent PDU groups.

Methods

The project will use the mdCATH dataset, a large public resource of MD simulations of CATH protein domains (5398 domains, 62ms of simulation, 134,950 trajectories, 3.3Tb size). The internship will focus on developing a prototype pipeline to extract PDU candidates from mdCATH trajectories, convert them into vector embeddings, and use unsupervised AI methods to identify recurring patterns in local protein dynamics.

The student will use Python tools such as MDAnalysis or MDTraj to read MD trajectories and extract residue centred local environments using a fixed spatial radius. For each PDU candidate, the student will calculate descriptors such as residue identity, secondary structure, pairwise contacts, local distances, flexibility, solvent exposure, and trajectory derived dynamic features. These descriptors will be standardised and converted into vector embeddings.

The second stage will involve training simple autoencoder models to compress the PDU embeddings into a latent space. The student will then apply UMAP and clustering methods to explore whether distinct classes of PDUs emerge. The pipeline will be run on the Brunel BUSTER high performance computing system, giving the student practical experience of Linux based command line workflows, job submission, managing data on HPC systems, and running scalable computational analyses.

Indicative timetable:

Week 1: Project induction, background reading on protein dynamics and PDUs, Linux command line training, BUSTER HPC access, Python environment setup, and familiarisation with mdCATH trajectory data.
Week 2: Implement initial data handling workflow using MDAnalysis or MDTraj; test reading structures and trajectories; define residue centred PDU extraction criteria.
Week 3: Extract PDU candidates from a small mdCATH subset; validate extracted local environments and refine extraction parameters.
Week 4: Calculate structural and dynamic features for each PDU, including residue identity, contacts, distances, flexibility and trajectory derived descriptors.
Week 5: Standardise features and convert PDU candidates into vector embeddings suitable for machine learning analysis.
Week 6: Train initial autoencoder models to learn compact latent representations of PDUs; run selected jobs on BUSTER HPC where appropriate.
Week 7: Apply UMAP and clustering methods to the learned embeddings; generate exploratory figures and assess whether recurrent PDU groups emerge.
Week 8: Consolidate scripts, document the pipeline, prepare preliminary figures and write a short technical report summarising outputs for future BBSRC responsive mode development.

Engineering and Physical Sciences Research Council

Brunel University of London

The expected outputs are open-source pipeline for, a small processed PDU dataset, exploratory UMAP and clustering figures, and preliminary evidence to inform future grant applications. This is a new research project and is not a continuation of an undergraduate dissertation or existing undergraduate project.

Q3 **What techniques/training will the scholarship provide?** (no more than 150 words)

The internship will provide training in computational structural biology, MD trajectory analysis, AI based data analysis, and high-performance computing. The student will learn how to work with protein structure and trajectory files, use Python libraries such as NumPy, pandas, MDAnalysis or MDTraj, and extract local residue environments from MD data. Student will gain experience in feature engineering, vector embeddings, data normalisation, autoencoders, UMAP, and clustering. The student will also run the pipeline on the Brunel BUSTER HPC system, gaining practical skills in Linux command line use, remote computing, job submission, environment management, and handling large datasets. Additional training will include software development using Git version control, reproducible research workflows, code documentation, scientific figure preparation, and communicating results through a short report and presentation.

Q4 **How does this research relate to work being carried out in the supervisor's laboratory?** (no more than 100 words)

My research group aims to understand the rules governing protein evolution, dynamics and function to develop computational models for rational protein design. A major challenge is that current AI tools predict structures well but still struggle to predict how mutations alter protein dynamics, stability and function. This internship extends our work on DyNoPy and PyCoM by testing whether recurrent local dynamic environments can be identified as Protein Dynamic Units. Establishing such units would provide a new representation of protein dynamics, supporting future AI models for designing proteins with improved industrial, therapeutic and environmental applications.

Q5 **Please provide a short statement outlining the arrangements that will be put in place to supervise the student** (no more than 200 words)

The student will be supervised directly by Dr Sarath C Dantu in the Department of Computer Science. The internship will begin with an induction covering the project aims, expected outputs, mdCATH data, research integrity, data management, and safe use of computing resources. Early supervision will include practical support for setting up Python environments, using Linux command line tools, accessing the Brunel BUSTER HPC system, and submitting computational jobs.

The student will have a scheduled one to one supervision meeting each week to review progress, troubleshoot technical issues, and agree realistic targets for the following week. Shorter informal check ins will be available as needed, especially during the initial data handling, HPC, and coding stages. The student will be integrated into the supervisor's research environment and encouraged to attend relevant group discussions.

The student will maintain a research log documenting datasets, code, parameters, and results. Code will be developed in a shared version-controlled repository, allowing regular feedback. The final stage will focus on consolidating scripts, preparing figures, and producing a short report and presentation summarising the pipeline and preliminary findings.

Project Supervisor's Signature:  Date: 27 May 2026