JANE M. DOE

Proposed Field: Life Sciences — Bioinformatics and Computational Biology

University of California, Berkeley • jmdoe@berkeley.edu

PERSONAL, RELEVANT BACKGROUND AND FUTURE GOALS STATEMENT

Introduction and Career Aspirations

My career goal is to lead an academic research group developing computational tools that integrate machine learning and structural biology to address complex diseases. Motivated by the limitations of traditional drug discovery during my grandfather's battle with cancer, I realized that data-driven modeling is essential to accelerate therapeutics. I plan to pursue a Ph.D. in Bioinformatics at UC Berkeley, focusing on deep learning methods for predicting protein dynamics.

Prior Research Experience

Research Project 1: Graph Neural Networks for Protein Structure Analysis

Advisor: Dr. Sarah Jenkins, UC Berkeley Department of Bioengineering (*Sept 2024 – Present*)

At UC Berkeley, I developed a dynamic Graph Neural Network (GNN) to model flexible protein-protein interfaces, representing structures as residue-level graphs with physical-distance edges. This architecture achieved a 12% improvement in binding affinity prediction accuracy over state-of-the-art rigid baseline models. This work led to a first-author manuscript currently under review at the *Journal of Computational Biology* and a poster at ISMB 2025.

Intellectual Merit: This study is the first application of dynamic edge-convolution to flexible docking, demonstrating that message-passing architectures can successfully learn transient physical interactions in biomolecular complexes. **Broader Impacts:** I released my code as the open-source library `ProGraph` (800+ downloads) and mentored two undergraduate women in my laboratory, guiding them through basic graph theory and PyTorch implementations to support diversity in STEM.

Research Project 2: Deep Learning for Genomic Variant Callers

Advisor: Dr. Robert Chen, Lawrence Berkeley National Laboratory (*June 2023 – Aug 2024*)

As a research assistant at LBNL, I designed a deep convolutional neural network (CNN) that treats genomic read alignments as multi-channel images to detect rare single-nucleotide variants (SNPs) and short indels. Our model reduced false-positive indel calls by 18% in low-coverage regions, and was integrated directly into LBNL's production pipeline. I presented these findings at the 2024 ASHG Annual Meeting.

Intellectual Merit: The project proved that convolutional neural networks can learn to differentiate sequencing artifacts from true biological variants using multi-dimensional spatial alignment features without hand-crafted heuristics. **Broader Impacts:** We deployed this model to public cloud infrastructure, making high-accuracy calling accessible to resource-constrained labs. I also organized a workshop at a local high school, introducing 25 underrepresented students to Python genomics.

Future Goals and Preparation

A Ph.D. at UC Berkeley will equip me with advanced training in statistical genomics. Beyond research, I am committed to public outreach. I will serve as a graduate mentor for the Berkeley Summer Research Program, guiding students from underrepresented backgrounds through computational biology projects, fostering an inclusive scientific community.

JANE M. DOE

Proposed Field: Life Sciences — Bioinformatics and Computational Biology

University of California, Berkeley • jmdoe@berkeley.edu

GRADUATE RESEARCH PLAN STATEMENT

Project Title: Graph-Based Deep Learning for Predicting Protein Interactions in Cancer Pathways

Introduction and Background

Protein-protein interactions (PPIs) mediate most cellular processes, and their dysregulation is a hallmark of cancer. However, predicting PPI interfaces is challenging due to the large search space and binding site flexibility. Current docking methods are expensive and struggle with transient interactions. I propose a graph deep learning framework to rapidly and accurately predict PPI interfaces in cancer-associated pathways.

Proposed Methodology

I will develop a multi-scale graph neural network (GNN) that operates on both sequence and structure levels to predict PPIs.

Task 1 (Graph Construction): Construct residue-level graphs with edges weighted by physical distances. Node features will include evolutionary conservation scores and physical properties (charge, hydrophobicity).

Task 2 (GNN Architecture): Design an architecture combining message-passing neural networks (MPNNs) with cross-attention to learn representations of individual proteins and predict interacting residue pairs.

Task 3 (Validation): Train the model on the Protein Data Bank (PDB), validate on cancer pathway databases (Reactome, ChEMBL), and validate key predictions in collaboration with the wet lab.

Timeline and Key Milestones

The proposed timeline is structured across three years of the fellowship, balancing coursework, software development, and experimental validation.

Year 1	Complete core coursework; construct residue-level graphs and extract features (Task 1); present at ISMB.
Year 2	Develop, train, and optimize the multi-modal GNN architecture (Task 2); release code on GitHub; submit manuscript.
Year 3	Perform computational screening of the p53 pathway and validate predictions in wet lab (Task 3); defend proposal.

Review Criteria

Intellectual Merit: The proposed research introduces a novel multi-modal GNN architecture that integrates sequence-level evolutionary history with structure-level physics, reducing prediction times from hours to seconds to enable genome-wide screening.

Broader Impacts: The resulting software will be released as an open-source tool, facilitating therapeutic discovery for labs worldwide. Furthermore, I will integrate this work into a modular "AI in Medicine" curriculum for UC Berkeley's summer camp for underrepresented high school students.