[redacted]

[redacted]

SABS: R3 Admissions Team
University of Oxford
OX1 2JD
United Kingdom

Subject: Statement of Purpose accompanying Application to EPSRC Centre for Doctoral Training Sustainable Approaches to Biomedical Science - Responsible and Reproducible Research (996 words)

Dear SABS: R3 Admissions Team,

I wish to pursue a DPhil in protein informatics, at the intersection of machine learning, physics and biology. I recently completed my MEng in biomedical engineering from [redacted] achieving First Class honours consistently throughout my degree, and have spent the 18 months since refining my expertise in computational science and machine learning within industry and at the [redacted] Research School in [redacted].

I would like to continue my academic progression as a DPhil student at Oxford because of its excellence in research and teaching and the networking and cohort building opportunities offered. My research interest is protein structure prediction and design using modern machine learning, and the SABS:R3 programme is my preferred route to contributing research to this field.

The SABS:R3 CDT provides access to excellent faculty, and is closely linked to the OPIG, which I would ideally seek to join throughout my DPhil. In fact, I became aware of the SABS:R3 programme while investigating opportunities to contribute to the work led by Prof. Charlotte Deane and Prof. Garrett Morris. Upon further investigating the CDT, I realised how well the programme caters to my chosen path of academic research and how complementary my experience and background is. The initial phase of teaching will provide me with the opportunity to refresh my medical knowledge in particular, while the computational approach means I can contribute the skill-set I have developed. Further, the engineering theory I enjoyed and learnt throughout my Masters, including control theory, machine learning and data science, would find useful application. The ties to industry within SABS:R3 ensure grounding of my research; it is important to me to know the real-world application and contribution of the work I conduct. In summary, I am currently in the crucial knowledge-building stage of my academic career, and I believe particularly the initial phase of the SABS:R3 programme will provide a strong foundation upon which to build my future research.

I have gained and actively sought exposure to protein design and relevant fields throughout my academic and professional career. I selected my Masters project to gain exposure to molecular physics and simulation and I attended elective modules including pattern recognition, machine learning and neural computation, synthetic biology, tissue engineering and control theory, which included extensive coursework projects in which I implemented a variety of machine learning models. Methods I have used include reinforcement learning, neural networks, unsupervised learning, principal component analysis, linear discriminant analysis and Bayesian inference[1].

In order to explore the field of molecular physics and simulation, I chose a Masters project building a simulation of DNA tile self-assembly using statistical physics. The objective was to gain insights into the thermodynamics of DNA tile self-assembly: before its successful experimental implementation, it was widely thought that self-assembly would fail due to incorrect bonds and aggregation during assembly. As part of [redacted] group I built a stochastic simulation of the one- pot self-assembly reaction of single-stranded DNA segments, and validated the results with thermodynamic theory and combinatorics. I wrote the source code from scratch in C and earned high marks for my final report. The source code is published and open source[2]. The project required substantial research beyond the scope of my degree and I enjoyed being part of [redacted] research group.

While successfully completing my Masters project, the results of CASP13 inspired me to further investigate machine learning [1]. In order to generate valuable research on protein structure prediction I sought greater understanding of the incredibly powerful algorithms used in bioinformatics. I achieved this with a six month internship at a machine learning and robotics start-up in Berlin, where I learnt more about the practical application of data science in commerce and industry. I researched and prototyped algorithms for industrial robots and computer vision, completing two successful client projects. I gained experience with software workflow and maintenance, skills which I continue to apply in research.

While in Berlin, I was offered a PhD position at the international [redacted], where I work on developing reinforcement learning agents to attain Theory of Mind. This position has enabled me to discover and produce scientific work at the cutting edge of machine learning research, projects which are ongoing and are planned for publication this summer. Despite my enthusiasm for the novel methods I am developing, I realised that my interests are more aligned with medically applied research, grounded in science, rather than human-computer interfaces. Therefore, I wish to redirect my energy to medical research.

Now is an exciting time to work on protein design, the combination of advances in sequencing, experimental imaging and computing power has created opportunities from previously intractable problems [2, 3, 4]. Findings from modern machine learning can be leveraged to develop powerful protein design

[1] [redacted]

[2] [redacted]

software, and have been shown to accelerate physical simulations by function approximation [1, 5]. Work making use of graph neural networks (GNN) to generate novel proteins has recently been published [6, 7], an area which I think holds promise for general protein design due to GNNs demonstrated ability to encode complex hierarchical relationships [8]. I am also interested in going beyond existing data to generate *de novo* structures in under-sampled regions. As a DPhil student I would love the opportunity to conduct research in this space, which I believe will substantially contribute to humanity's medical capabilities in the coming decades.

When I told my friends I was applying to SABS:R3 to pursue protein design research, they groaned and exclaimed "finally!". They never understood why I hadn't immediately applied to do the research I've been talking about for the last four years. In the meantime I have developed my machine learning expertise and gained research experience enabling me to support the research undertaken by the faculties of computational biomedical science at Oxford.

Kind regards,

[redacted]

*References

[1] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W.R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 2020.

[2] Enrique Marcos and Daniel Adriano Silva. Essentials of de novo protein design: Methods and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(6), 2018.

[3] David Baker. What has de novo protein design taught us about protein folding and biophysics? *Protein Science*, 28(4), 2019.

[4] Antonella Paladino, Filippo Marchetti, Silvia Rinaldi, and Giorgio Colombo. Protein design: from computer models to artificial intelligence. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(5), 2017.

[5] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to Simulate Complex Physics with Graph Networks. *ICML*, 2020.

[6] Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep Generative Models for 3D Linker Design. *Journal of chemical information and modeling*, 60(4), 2020.

[7] Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M. Kim. Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Systems*, 11(4), 2020.

[8] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv*, 2018.