"A.I. Predicts the Shape of Nearly Every Protein Known to Science" - While this *New York Times* headline would have been met with incredulity by most biologists only a few years ago, DeepMind's AlphaFold has already become one of the most important computational methods for structural biology today. However, despite AlphaFold undoubtedly being a milestone achievement, the structures it predicts still lack critical biological context: Instead of existing in isolation, proteins often interact with small chemical compounds that are essential to their function. Small chemical compounds are also the most commonly employed class of drug used to modulate protein function. More advanced models for predicting protein-ligand interactions are therefore needed to fully leverage the power of current machine learning methods for drug discovery against novel targets. In addition, the recent success in protein structure prediction has been followed by significant advancements in generative modelling of entirely new protein structures. Models incorporating cutting-edge research from geometric deep learning and natural language processing have already shown promising success for designing proteins with desired features, yet there is still a lot of potential for pushing the frontier. Combining a background in Biochemistry with more than 2.5 years of computational research experience, I believe I am ideally suited to tackle such research questions. As a PhD student in the Columbia Computer Science program, I therefore plan to focus on developing novel machine learning methods for predicting molecular interactions and designing new proteins with tailored properties.

My undergraduate studies have given me a fundamental understanding of the problems I am now trying to solve with machine learning. After graduating high school as best student in my year with an award from the German Chemical Society, I joined the prestigious Biochemistry program at Heidelberg University as one of only 25 selected students, where I learned about the inner workings of proteins and molecular causes of diseases. In my first research project in Professor Maike Bublitz's group at the University of Oxford, I used X-ray crystallography to determine the structure of a membrane protein in complex with novel inhibitors. Now, I am using such experimental protein structures as training data for machine learning models, and my previous experience with crystallography has proven a tremendous advantage for constructing bespoke high-quality datasets.

Working on molecular docking simulations with Professor Rebecca Wade for my bachelor's thesis at the Heidelberg Institute for Theoretical Studies (HITS), I experienced firsthand how computational algorithms can efficiently generate plausible mechanistic hypotheses. Since then, I have used my master's studies to specialize in computational biology with coursework such as *Bioinformatics/Molecular Dynamics*, and *Foundations of machine learning and high-dimensional data analysis*. Aiming to get an understanding of diverse problems in the field, I chose to work on a breadth of research projects, ranging from physics-based protein design in Professor Bruno Correia's group at École Polytechnique Fédérale de Lausanne

(EPFL) to analyzing high-dimensional transcriptomics data in Dr. Judith Zaugg's group at the European Molecular Biology Laboratory (EMBL). Exploring different research fields has taught me how to quickly assimilate new knowledge through self-study, for example by getting up to speed with the programming language R in just a few weeks when starting at EMBL. My research has been productive, with parts of my bachelor's thesis integrated into a journal submission to *Circulation*, and two other manuscripts currently in writing.

I have long been interested in machine learning methods and their ability to easily grasp complex patterns in high-dimensional data. This motivated me to join Roche for a 6-month project on probabilistic modeling of antibody sequences from display datasets. Display campaigns involve the vast screening of antibody libraries against target epitopes over multiple rounds in order to enrich strongly binding sequences. Given a large-scale dataset of millions of antibody variants from such a campaign, I worked on developing models for predicting the affinity of experimentally unobserved sequences using Gaussian processes (GPs). Studying the literature surrounding stochastic processes proved challenging at first, though I quickly came to appreciate the rigorous uncertainty estimates provided by Bayesian models. Compared to the previously used random forests, my GP models improved predictive performance and were able to much better quantify their ability to generalize to out-of-distribution sequences. In order to streamline the future application of those models, I wrote the "BayDisplay" Python library for working with Gaussian processes on display datasets, which is now internally used at Roche. Additionally, I presented my findings as a poster at the *Protein Engineering Summit Boston* (PEGS Boston 2023).

My master's thesis in the AlQuraishi laboratory at Columbia has allowed me to fully merge my background in computational biochemistry with my interest in cutting-edge machine learning methods. The main goal of my thesis is to expand AlphaFold into a model that is fully capable of protein-ligand co-folding, bridging the previously mentioned gap between protein structure prediction and drug discovery. Leveraging my past experience in working with protein structures and large-scale sequence datasets, I first created a novel high-quality dataset of protein-ligand complexes and also computed new multiple sequence alignments, required as inputs to AlphaFold, which have become part of a publication accepted at *NeurIPS 2023 Datasets and Benchmarks*. By adding ligand representations and customized loss functions to the network, I successfully implemented the first version of the co-folding model "OpenBind". First experiments with OpenBind have shown promising results on smaller test sets, and the model is currently being trained at large scale on the full training data. In my PhD I am hoping to continue working with Professor AlQuraishi towards pushing the frontier of machine learning for molecular modeling.

Delving deep into AlphaFold's source code made me appreciate how inter-disciplinary approaches

are key for tackling grand challenges such as the protein structure prediction problem. AlphaFold's architecture borrows from multiple different fields, combining graph transformers, SE(3)-equivariant attention, and masked language modeling, and recent protein design models have taken a similarly integrated approach. I therefore look forward to further strengthening my understanding of specialized research areas in machine learning and am excited about potential learning and collaboration opportunities with Columbia's strong and diverse Computer Science faculty. I am intrigued by Professor Knowles' work on Bayesian models for genomics, and I believe that probabilistic modeling also holds great potential for protein design. In addition, the recent success of protein language models piqued my interest in natural language processing, and I have been particularly impressed by the research of Professor McKeown and Professor Collins.

In conclusion, my previous academic experience has uniquely prepared me for pursuing graduate research in developing models for molecular modeling and protein design. As I look to the future, I aspire to become an independent scientist, leading a research group in either academia or industry. The Columbia Computer Science program would not only offer me an academic foundation that will guide me towards this goal, but also a platform to create meaningful and lasting impacts at the interface of machine learning and biology.