

Preface

PROTEIN SCIENCE HAS MADE TREMENDOUS PROGRESS since the early studies in the mid-twentieth century. The landmark paper on the structure of myoglobin by John Kendrew, followed by the structure of hemoglobin published by Max Perutz, opened the doors to structural biology and deepened our understanding of protein function. Around the same time, Frederick Sanger's sequencing of insulin provided a glimpse into the primary structure of a protein. Together, these discoveries laid the foundation for protein science, establishing the importance of both protein structure and amino acid sequence, and sparking fundamental questions about the relationship between the two.

In the decades that followed, advances in sequencing technologies, structural biology, and bioinformatics allowed researchers to investigate proteins at an unprecedented scale. Large sequence and structure databases, high-throughput experimental methods, and computational modeling techniques transformed protein science into a data-rich field. With this influx of data came both the opportunity and challenge of extracting meaningful patterns, predicting function from sequence, and engineering new proteins with desired functions. The past few years have seen particularly impactful breakthroughs at the intersection of machine learning and protein science, beginning with the introduction of transformer-based protein language models in 2019 and 2020. These models, inspired by natural language processing, demonstrated an ability to learn meaningful representations of proteins from large-scale protein sequence data, capturing evolutionary, functional, and structural signals. This was soon followed by the introduction of AlphaFold 2, which set a new standard for protein structure prediction. These developments have fundamentally changed the field of protein science, with applications in protein function annotation, variant effect prediction, and generative protein design.

This book explores the rapidly evolving intersection of machine learning and protein science. The first chapters (Chapters 1 and 2) introduce machine learning approaches for learning representations of proteins, including applications to antibody comprehension. Subsequent chapters cover statistical models of coevolution (Chapter 3) and large-scale homology searches (Chapter 4), which have implications for protein structure prediction. The middle chapters examine machine learning applications in functional annotation and evolution, including variant effect prediction (Chapter 5) and the fundamental question of whether protein novelty is predictable (Chapter 6). We then explore generative models for both protein sequence and structure (Chapters 7–9). The final chapter (Chapter 10) reflects on the environmental impact of applying large-scale machine learning in protein science and engineering, acknowledging the need to balance technological advancement with sustainable computational practices.

PETER K. KOO
CHRISTIAN DALLAGO
ANANTHAN NAMBIAR
KEVIN K. YANG