Advancing computational methods for the functional characterisation of antibodies using structural and flexibility data



Fabian C. Spoendlin

Exeter College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2026

Acknowledgements

I am deeply grateful to Prof. Charlotte Deane for her invaluable guidance and feedback throughout my DPhil. Her mentorship was fundamental in shaping this thesis and my academic growth.

I would also like to thank my supervisors at Roche, Guy Georges, Catherine Wong and Alex Bujotzek, for sharing their expertise and insights to keep this work relevant for real world drug discovery.

I feel lucky to have been part of OPIG, which provided a stimulating, supportive, but most importantly fun environment, throughout my degree. I am especially thankful to all the people I shared office 2.17 with over the last four years and who provided me with daily discussions, insights and motivation. Special thanks go to Alex Greenshields-Watson, Oliver Turnbull, Alissa Hummer and Lewis Chinery, whom I bothered a disproportionate amount for advice as well as to the new 2025 office cohort, who have been essential in motivating me through the thesis writing process. I would also like to thank current and past OPIGlets Brennan Abanades, Matteo Cagiada and Matthew Raybould for offering me their guidance and supervision at many stages throughout this work.

Finally, I am eternally grateful to my family for their consistent support and encouragement throughout my entire academic journey. This DPhil would not have been possible without them.

Doing a DPhil is a challenging endeavour. Thanks to everyone who has contributed to making this such an exciting and rewarding experience.

Abstract

This doctoral thesis presents advancements in the computational characterisation of antibodies through the integration of structural and flexibility data. I first introduce SPACE2 (Chapter 2), a novel antibody epitope profiling method. In contrast to conventional approaches that establish functional equivalence based on sequence identity, I demonstrated that structural comparison enables the identification of a broader range of antibodies engaging the same epitope.

Findings from this initial project identified conformational diversity as a primary limitation in protein structure prediction. Consequently, the subsequent chapters focus on developing methods to accurately predict antibody structural flexibility at scale.

ITsFlexible, presented in Chapter 3, represents an initial approach to this problem. Rather than modelling structures of specific states, I employed a deep learning framework to predict whether or not an antibody CDR is likely to exchange between multiple conformations (exhibits structural flexibility).

Following on from this work, I focussed on methods to structurally characterise the antibody conformational ensemble. In Chapter 4, I describe Calvados 3-Fv, a high-throughput molecular dynamics (MD) workflow. Calvados 3-Fv accurately reproduced CDR ensemble properties at a significantly reduced computational cost compared to all-atom simulations while some limitations regarding the exact conformational landscapes covered remain. Leveraging this compute efficiency, I generated and released FlAbDab and FTCRDab, two databases containing more than 150,000 simulated antibody and TCR structures. These resources provide a substantial foundation for studying CDR flexibility and training the next generation of conformation-aware structure prediction models.

Finally, I present ABB4-STEROIDS (Chapter 5), a generative protein structure prediction model designed to sample the equilibrium distribution of antibody conformations. A key advantage of this framework is that it allows the training on a combination of data sources. I showed that pre-training on large-scale coarse-grained MD data, followed by fine-tuning on a small set of all-atom trajectories, significantly mitigates the biases inherent in coarse-grained data.

The thesis concludes (Chapter 6) by synthesising these contributions and outlining future avenues for research in antibody and protein conformation modelling.

Declaration on the use of generative AI

I hereby declare that the preparation of this thesis has involved the use of generative artificial intelligence (Gen AI) tools in accordance with the University's guidelines on research integrity and the safe and responsible use of AI. My use of these tools complies with all University, divisional, and departmental regulations.

The specific applications of Gen AI in this work are detailed below. Tools used include ChatGPT (versions 3.5-5.3) and Google Gemini (version 3).

1. Assistance in coding and technical development Gen AI was employed to assist in the development and refinement of coding scripts used throughout this research. All AI-assisted code was manually reviewed by me to ensure correctness and reproducibility. No agentic or fully autonomous coding systems were used to generate or edit entire codebases.

2. Literature research and subject understanding Gen AI tools were used to support literature research and to gain a general understanding of specific subject areas. However, I have maintained full critical engagement with the source material. Every claim and reference generated by AI was independently verified against primary sources before being included in the thesis.

3. Language refinement and editing Gen AI was used as an editing assistant to improve the grammar, wording and overall clarity of the text. This use was limited to the refinement of my own writing. No substantive original writing was performed by AI.

Consistent with the University's declaration of authorship, I pledge that the work submitted is entirely my own, except where otherwise indicated.

Publications

This DPhil led to the following publications:

- **Fabian C. Spoendlin**, Matteo Cagiada, King Ifashe, Odysseas Vavourakis, and Charlotte M. Deane. "Sampling antibody conformational ensembles with ABodyBuilder4-STEROIDS," Biorxiv (2026). https://doi.org/10.64898/2026.04.14.718378.

- Eric J. D. Wong, **Fabian C. Spoendlin**, Alexander Greenshields-Watson, Christopher R. Taylor, and Charlotte M. Deane. "Fast retrieval of structurally similar antibodies from large sequence databases with AbSLang," Biorxiv (2026). https://doi.org/10.64898/2026.04.14.718378.

- Alexander Greenshields-Watson, Parth Agarwal, Sarah A. Robinson, Benjamin Heathcote Williams, Gemma L. Gordon, Henriette L. Capel, Yushi Li, **Fabian C. Spoendlin**, et al. "ANARCII: A Generalised Language Model for Antigen Receptor Numbering," Communications Biology (2026). https://doi.org/10.1038/s42003-026-10186-z

- Matteo Cagiada*, **Fabian C. Spoendlin***, King Ifashe, and Charlotte M. Deane. "Uncovering the Flexibility of CDR Loops in Antibodies and TCRs through Large-Scale Molecular Dynamics," Biorxiv (2025). https://doi.org/10.1101/2025.11.10.687725.

- **Fabian C. Spoendlin**, Monica L. Fernandez-Quintero, Sai S. R. Raghavan, Hannah L. Turner, Anant Gharpure, Johannes R. Loeffler, Wing K. Wong, et al. "Predicting the Conformational Flexibility of Antibody and T-Cell Receptor CDRs." Nature Machine Intelligence (2025). https://doi.org/10.1038/s42256-025-01131-6.

- **Fabian C. Spoendlin**, and Charlotte M. Deane. "Structure-Based Epitope Profiling with the Structural Profiling of Antibodies to Cluster by Epitope 2 (SPACE2) Algorithm." In Epitope Mapping Protocols, edited by Johan Rockberg, Johan Nilvebrant, Magdalena Malm, and Niklas Berndt Thalén, 261–74. New York, NY: Springer US, 2025. https://doi.org/10.1007/978-1-0716-4591-8_16.

- Alexander Greenshields-Watson, Odysseas Vavourakis, **Fabian C. Spoendlin**, Matteo Cagiada, and Charlotte M. Deane. "Challenges and Compromises: Predicting Unbound Antibody Structures with Deep Learning." Current Opinion in Structural Biology 90 (2025). https://doi.org/10.1016/j.sbi.2025.102983.

- Charlotte Sweeney, Nele P. Quast, **Fabian C. Spoendlin**, Yee Whye Teh. "Estimating protein flexibility via uncertainty quantification of structure prediction models." NeurIPS 2024 Workshop: Machine Learning in Structural Biology (2024).

- **Fabian C. Spoendlin**, Wing Ki Wong, Guy Georges, Alexander Bujotzek, and Charlotte Deane. "AbFlex: Predicting the Conformational Flexibility of Antibody CDRs," ICML 2024 Workshop: Machine Learning for Materials and Life Sciences (2024). https://openreview.net/forum?id=or4tArwd5a.

- Jakob R. Riccabona*, **Fabian C. Spoendlin***, Anna-Lena M. Fischer, Johannes R. Loeffler, Patrick K. Quoika, Timothy P. Jenkins, James A. Ferguson, et al. "Assessing AF2's Ability to Predict Structural Ensembles of Proteins." Structure (2024). https://doi.org/10.1016/j.str.2024.09.001.

- Kit Gallagher, Maximilian A. Strobl, Derek S. Park, **Fabian C. Spoendlin**, Robert A. Gatenby, Philip K. Maini, and Alexander R. Anderson. "Mathematical Model-Driven Deep Learning Enables Personalized Adaptive Therapy." Cancer Research (2024). https://doi.org/10.1158/0008-5472.CAN-23-2040.

- **Fabian C. Spoendlin**, Brennan Abanades, Matthew I. J. Raybould, Wing Ki Wong, Guy Georges, and Charlotte M. Deane. "Improved Computational Epitope Profiling Using Structural Models Identifies a Broader Diversity of Antibodies That Bind to the Same Epitope." Frontiers in Molecular Biosciences 10 (2023).
  https://doi.org/10.3389/fmolb.2023.1237621

*Denotes equal contributions

Contents

List of Abbreviations **xv**

1 Introduction **1**
- 1.1 Motivation . . . 2
- 1.2 Contributions . . . 2
- 1.3 Antibodies . . . 3
  - 1.3.1 Antibody structure . . . 3
  - 1.3.2 Antibody sequence diversity . . . 5
  - 1.3.3 Antibodies in the immune system . . . 5
  - 1.3.4 Antibodies as therapeutics . . . 7
  - 1.3.5 Epitope profiling . . . 8
- 1.4 Statistical methods and machine learning . . . 9
  - 1.4.1 Cluster analysis . . . 9
  - 1.4.2 Deep learning . . . 12
- 1.5 Protein structure prediction . . . 17
- 1.6 Protein dynamics . . . 21
  - 1.6.1 Conformational flexibility of antibodies . . . 22
  - 1.6.2 Methods to study protein dynamics . . . 25
- 1.7 Thesis outline . . . 31

2 Computational epitope profiling using structural models **33**
- 2.1 Motivation . . . 34
- 2.2 Contributions . . . 34
- 2.3 Background . . . 35
- 2.4 Methods . . . 38
  - 2.4.1 Data sets . . . 38
  - 2.4.2 SPACE1 . . . 39
  - 2.4.3 SPACE2 . . . 40
  - 2.4.4 Numbering scheme and region definitions . . . 42
  - 2.4.5 Analysis of structural clusters . . . 42
- 2.5 Results . . . 45
  - 2.5.1 Evaluating an optimal clustering algorithm . . . 45

2.5.2 Examining the behaviour of agglomerative clustering across different structural similarity thresholds 47
2.5.3 Evaluating the optimal region for clustering 49
2.5.4 SPACE2 performs well on sets of antibodies against diverse targets 50
2.5.5 Advances in structure prediction improve structure-based computational epitope profiling 53
2.5.6 SPACE2 improves coverage compared to clonotyping 54
2.5.7 SPACE2 identifies functional convergence signals 56
2.6 Discussion 59

3 Predicting the conformational flexibility of CDRs 63
3.1 Motivation 64
3.2 Contributions 64
3.3 Background 65
3.4 Methods 68
3.4.1 ALL-conformations dataset 68
3.4.2 ITsFlexible 70
3.4.3 Baseline models 72
3.4.4 Alternative flexibility classifiers 73
3.4.5 Molecular dynamics simulations 74
3.4.6 Cryo-EM experimental protocol and data analysis 76
3.5 Results 79
3.5.1 ALL-conformations dataset 79
3.5.2 Predicting the flexibility of CDR-like protein loop motifs . . 80
3.5.3 ITsFlexible is highly predictive of CDR3 flexibility 82
3.5.4 ITsFlexible matches CDR flexibility in molecular dynamics simulations 85
3.5.5 Cryo-EM experiments confirm predicted flexibility 86
3.6 Discussion 89

4 Uncovering CDR flexibility through large-scale molecular dynamics 93
4.1 Motivation 94
4.2 Contributions 94
4.3 Background 95
4.4 Methods 98
4.4.1 CV3-Fv workflow 98
4.4.2 Datasets 99
4.4.3 Benchmarking 103

4.5 Results . . . 106
4.5.1 Characterisation of Antibody Fv using CALVADOS 3 . . . 106
4.5.2 Comparison with ensembles observed in experimental structures 110
4.5.3 Investigating the limitation of our flexibility predictions . . . 113
4.5.4 FlAbDab and FTCRDab . . . 115
4.5.5 Analysis of CDR dynamics . . . 116
4.6 Discussion . . . 119

5 Sampling conformational ensembles with deep learning 123
5.1 Motivation . . . 124
5.2 Contributions . . . 124
5.3 Background . . . 125
5.4 Methods . . . 128
5.4.1 Datasets . . . 128
5.4.2 All-atom MD simulations . . . 131
5.4.3 Metrics for ensemble analysis . . . 133
5.4.4 ABB4 model . . . 136
5.4.5 Inference of benchmarked models . . . 140
5.5 Results . . . 143
5.5.1 Overview of the ABodyBuilder4 models . . . 143
5.5.2 ABodyBuilder4-base serves as a strong base model for structure prediction . . . 144
5.5.3 ABodyBuilder4-STEROIDS accurately samples antibody conformational ensembles . . . 146
5.6 Discussion . . . 157

6 Conclusions and future directions 163
6.1 Conclusions . . . 164
6.2 Future work . . . 165

Appendices

A Computational epitope profiling using structural models 171
A.1 Supplementary Tables and Figures . . . 172

B Predicting the conformational flexibility of CDRs 179
B.1 Supplementary methods . . . 180
B.1.1 ITsFlexible models . . . 180
B.1.2 Amino acid encoding . . . 184
B.1.3 Molecular dynamics dataset . . . 185
B.1.4 Cryo-EM . . . 185

B.2 Supplementary results 189
B.2.1 CDR3 secondary structure 189
B.2.2 ALL-conformations 189
B.2.3 Extended evaluation of ITsFlexible performance 194
B.2.4 ABodyBuilder2 201
B.2.5 Case study antibodies for cryo-EM 205

C Uncovering CDR flexibility through large-scale molecular dynamics 207
C.1 Supplementary Figures 208

D Sampling conformational ensembles with deep learning 223
D.1 Supplementary methods 224
D.1.1 ABB4 model details 224
D.2 Supplementary results 229
D.2.1 Sampling of coarse-grained MD frames. 229
D.2.2 Increased CDRH3 flexibility in FlAbDab simulations started from predicted structural models 229
D.3 Supplementary figures 232

References 245

List of Abbreviations

AA . . . . . All-atom

ABB . . . . . ABodyBuilder

AF2 . . . . . AlphaFold2

AF3 . . . . . AlphaFold3

BCR . . . . . B-cell receptor

CDR . . . . . Complementarity determining region

CG . . . . . Coarse-grained

Cv3 . . . . . Calvados 3

Cv3-Fv . . . . Calvados 3-Fv

EGNN . . . . . Equivariant graph neural network

Fab . . . . . Fragment antigen binding

Fv . . . . . Fragment variable

GNN . . . . . Graph neural network

IDP . . . . . Intrinsically disordered proteins

IDR . . . . . Intrinsically disordered regions

Ig . . . . . Immunoglobulin

IPA . . . . . Invariant point attention

lDDT . . . . . Local distance difference test

mAb . . . . . Monoclonal antibody

MAE . . . . . Mean absolute error

MetaD . . . . Metadynamics

MD . . . . . Molecular dynamics

ML . . . . . Machine learning

MSA . . . . . Multiple sequence alignment

MSE . . . . . Mean squared error

plDDT . . . . Predicted local distance difference test

PR AUC . . . Area under the precision recall curve

RMSD . . . . . Root mean squared deviation

RMSE . . . . . Root mean squared error

RMSF . . . . . Root mean squared fluctuation

ROC AUC . . Area under the receiver operating characteristic curve

SE(3) . . . . . Special Euclidean group

SHM . . . . . Somatic hyper mutation

SO(3) . . . . . Special orthogonal group

TCR . . . . . T-cell receptor

1 Introduction

Contents

1.1 Motivation 2
1.2 Contributions 2
1.3 Antibodies 3
1.3.1 Antibody structure 3
1.3.2 Antibody sequence diversity 5
1.3.3 Antibodies in the immune system 5
1.3.4 Antibodies as therapeutics 7
1.3.5 Epitope profiling 8
1.4 Statistical methods and machine learning 9
1.4.1 Cluster analysis 9
1.4.2 Deep learning 12
1.5 Protein structure prediction 17
1.6 Protein dynamics 21
1.6.1 Conformational flexibility of antibodies 22
1.6.2 Methods to study protein dynamics 25
1.7 Thesis outline 31

1.1 Motivation

This chapter provides the background to understand the research in this thesis. I begin by introducing antibody structure and mechanisms giving rise to their sequence diversity before discussing their immunological function and uses as therapeutics. Next, I discuss the background of the key statistical and machine learning methods I used. I focus on cluster analysis techniques, geometric deep learning and generative modelling paradigms. I then introduce the field of protein structure prediction. Finally, I discuss the functional importance of conformational flexibility in antibodies and various computational and experimental techniques used to model protein dynamics.

1.2 Contributions

This chapter contains material reproduced from:

Fabian C. Spoendlin. "An Overview of Clustering Algorithms." April 6, 2023. https://www.blopig.com/blog/2023/04/an-overview-of-clustering-algorithms/.

Fabian C. Spoendlin, Monica L. Fernandez-Quintero, Sai S. R. Raghavan, Hannah L. Turner, Anant Gharpure, Johannes R. Loeffler, Wing K. Wong, et al. "Predicting the Conformational Flexibility of Antibody and T-Cell Receptor CDRs." Nature Machine Intelligence (2025).

Matteo Cagiada*, **Fabian C. Spoendlin***, King Ifashe, and Charlotte M. Deane. "Uncovering the Flexibility of CDR Loops in Antibodies and TCRs through Large-Scale Molecular Dynamics," Biorxiv (2025).

*Denotes equal contributions

1.3 Antibodies

Antibodies are important components of the adaptive immune system and have become a central technology in modern biotherapeutics (Lu et al. 2020; Khetan et al. 2022). As members of the immunoglobulin (Ig) superfamily, they belong to a diverse class of proteins specialised for molecular recognition and binding. Secreted by B-cells, antibodies identify pathogens to either neutralize them directly or recruit effector mechanisms that facilitate their clearance. Their immense sequence diversity allows for the recognition of a virtually limitless array of antigens with characteristic high affinity. This section provides an overview of antibody structure and sequence, their immunological roles, and their application as therapeutic agents.

1.3.1 Antibody structure

Antibodies are large proteins that adopt a characteristic Y-shape (Fig. 1.1). They consist of four polypeptide chains, two heavy and two light. Heavy chains contain four domains, three conserved and one variable, whereas light chains only contain two domains, one conserved and one variable. Each of the domains adopts an immunoglobulin fold. This is a highly conserved structure that consists of a 2-layer β-sandwich of 7 to 9 antiparallel β-strands (Chiu et al. 2019).

The two C-terminal heavy chain domains (CH2 and CH3) bind to each other and form the fragment crystallisable (Fc). The name 'fragment crystallisable' originates from the historical fact that this was the first part of the antibody that could be crystallised (Chiu et al. 2019). Biologically, the Fc functions to signal to and engage the immune system after the antibody binds to an antigen (Diebolder et al. 2014; Lu et al. 2018; Jafarshad et al. 2007; Lu et al. 2018).

The light chains together with the two N-terminal heavy chain domains (VH, CH1) form two fragment antigen binding arms (Fab). The VH and VL domains together constitute the fragment variable (Fv). Most parts of the Fv make up the framework region which adopts the immunoglobulin fold (Chiu et al. 2019). The six hyper-variable complementarity determining regions (CDRs, 3 on each chain) are located at the tip of the Fv and form the main interaction site with the antigen.

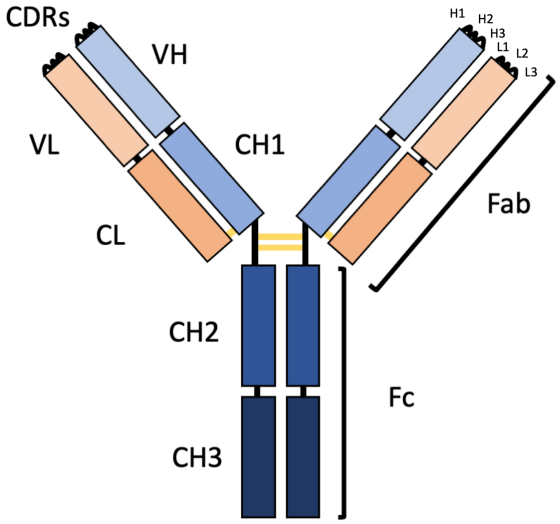


Figure 1.1: Schematic of an antibody structure. Standard human antibodies are formed of two light (orange) and two heavy chains (blue). Chains consist of a set of conserved (CL, CH1-3) and variable (VL, VH) domains. The variable domains together with CL and CH1 form the fragment antigen binding (Fab) that binds to a specific epitope on an antigen. The six complementarity-determining regions (CDRs) at the tip of each Fab form the main interaction site with the antigen and are responsible for the majority of antibody sequence and structural diversity. Domains CH2 and CH3 constitute the fragment crystallisable (Fc) which signals to the immune system after antigen binding.

Collectively all residues that interact with the antigen are called the paratope. The complementary region of antigen residues interacting with the antibody is named the epitope (Sela-Culang et al. 2013).

Five of the six CDRs are referred to as 'canonical' CDRs. Despite higher sequence diversity than the framework regions, these loops often adopt a limited number of structures referred to as canonical forms (e.g. Chothia et al. 1987; Chothia et al. 1989; Martin et al. 1996; North et al. 2011; Nowak et al. 2016). The CDRH3s are more diverse in sequence and tend to be longer than the canonical loops. For this reason, predicting the structures of canonical CDRs is possible with relatively high accuracy and modelling CDRH3s remains a challenge (e.g. Abanades et al. 2023; Kenlay et al. 2024; Leem et al. 2016). However, CDRH3 contributes disproportionately to the paratope and is therefore functionally and structurally often considered the most important region to model (Sela-Culang et al. 2013).

1.3.2 Antibody sequence diversity

The capacity for highly specific antigen recognition is enabled by the immense sequence diversity of antibodies. This diversity is generated through two main mechanisms: VDJ recombination and somatic hypermutation (SHM).

Antibodies are encoded by multiple gene segments distributed across the human genome. During B-cell development, individual segments are randomly spliced to form a single functional variable domain (Figure 1.2). The VH is assembled from three segments, Variable (V), Diversity (D), and Joining (J), whereas the VL is formed from only V and J segments (Tonegawa 1983). In humans, light chains are further categorised into two isotypes: kappa (κ) and lambda (λ). The combinatorial potential of this process is enormous: there are 60 V and 8 J gene segments for the λ locus, 79 V and 10 J segments for the κ locus and 111 V, 27 D, and 9 J segments for the heavy chain locus captured in the IMGT database (Lefranc et al. 2003). Beyond simple combination, further diversity is introduced through D-segment translation across all three reading frames and junctional diversity, where nucleotides are randomly deleted or inserted at the segment boundaries. This initial stage alone produces a theoretical combinatorial space of up to 10^{15} naive antibodies (Rees 2020). However, experimental estimates have suggested lower numbers around $10^7 - 10^9$ (Rees 2020). As the VDJ recombination site corresponds to CDRH3, this loop exhibits the highest degree of sequence and structural diversity (Tonegawa 1983).

A second level of diversity is introduced through SHM (detailed in the next section), which expands the accessible sequence space by several orders of magnitude. During affinity maturation B-cells undergo rapid proliferation during which point mutations are introduced into the variable region (Victora et al. 2012; Di Noia et al. 2007).

1.3.3 Antibodies in the immune system

The enormous diversity of the antibody immune repertoire enables the body to eliminate an almost unlimited range of antigens. Generation of the repertoire starts in the bone marrow, where, through VDJ recombination, B cells produce an initial

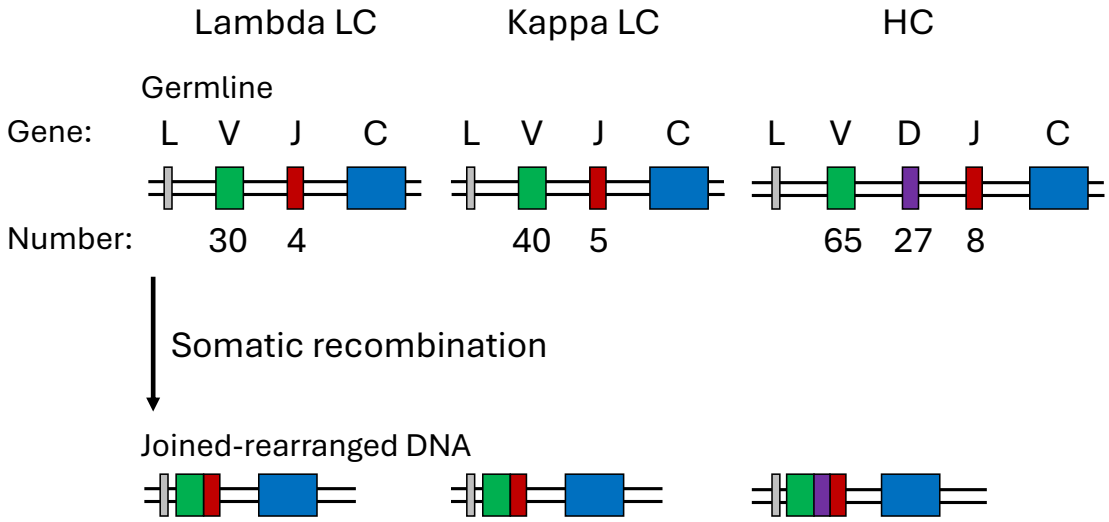


Figure 1.2: Schematic illustrating VDJ recombination. The germline encodes for a number of individual V, J and, for heavy chain, D genes which are flanked by a leader (L) and the constant (C) fragments. Somatic recombination at the DNA level combines one unique set of V, D and J segments in each B-cell with the flanking pieces to give rise to a diverse repertoire of antibodies.

diversity of immunoglobulins. In this early stage, the antibodies are said to be 'naive' and are expressed as membrane-bound receptors also referred to as B-cell receptors (BCRs) (Victora et al. 2012; Di Noia et al. 2007).

Once a naive antibody binds to a target in the secondary lymphoid organs, the B-cell is activated and migrates to a germinal centre to undergo affinity maturation. During this process, SHM introduces point mutations into the variable regions of the antibody genes resulting in further diversity of the immune repertoire. Rapid mutation rates create a pool of B-cell variants expressing antibodies with slightly altered binding properties. Only those variants with increased binding affinity are selected and differentiate into plasma cells. As these plasma cells mature, a transition occurs whereby the antibody is no longer expressed as a membrane-bound receptor but instead as a secreted antibody. After undergoing affinity maturation antibodies are called 'mature'. All antibodies that originate from a single progenitor cell are called a clonal lineage (Victora et al. 2012; Di Noia et al. 2007).

Once matured, antibodies serve the role of identifying extracellular pathogens

and toxins and facilitate their removal. Their most direct functional mechanism is neutralization, wherein the antibody binds to the surface of a pathogen or toxin, physically obstructing its ability to interact with components of the body. While this effectively prevents disease causing events, neutralization alone does not result in the elimination of the foreign particles from the body (Lu et al. 2018).

To achieve clearance, the antibody Fc domain engages other components of the immune system. Three primary downstream effector functions can be triggered in this way. First, pathogen bound antibodies can activate the complement pathway, a complex system involving over 30 plasma proteins. This system 'tags' the antigen for destruction, promotes local inflammation, and ultimately assembles the Membrane Attack Complex which directly kills the pathogen (Diebolder et al. 2014). Second, antibodies act as opsonins. This means they 'label' pathogens and attract phagocytic cells such as macrophages and neutrophils. These cells engulf target cells and digest them within their lysosomes (Lu et al. 2018). Finally, in cases where a target, such as a large parasite or a virally infected host cell, is too large for phagocytosis, antibodies trigger Antibody-Dependent Cellular Cytotoxicity (ADCC). In this process, Natural Killer (NK) cells bind to the antibodies and release cytotoxic granules, such as perforin and granzymes, which force the target cell to undergo programmed cell death (Jafarshad et al. 2007; Lu et al. 2018).

1.3.4 Antibodies as therapeutics

Owing to their high affinity and specificity, antibodies have received a lot of attention from the pharmaceutical and biotechnology industries. Therapeutic antibodies are the most rapidly growing class of drugs with 48 approved antibody therapeutics as of January 2025, 38 of which were approved in the previous three years (Carter et al. 2018; The Antibody Society 2026; Basu et al. 2019).

The development of therapeutic antibodies is an expensive and time consuming process, typically taking several years (Lu et al. 2020). The early stages of the process can be divided into two main steps (Chiu et al. 2016). The first step is usually referred to as lead discovery. The aim here is to discover antibodies that bind

to the target of interest. There are several techniques to find these initial binders. In animal immunisation approaches, the natural ability of the animal to produce antibodies against a target is exploited. Specifically, an animal is infected with the antigen of interest followed by sequencing of its immune repertoire (Hou et al. 2016). Alternatively, *in vitro* approaches such as phage display screen antibody libraries for target binding (Smith et al. 1997). Most recently, AI tools have been developed to design binders fully *in silico*. While promising their utility remains unclear at this point in time (Mille-Fragoso et al. 2025; Butcher et al. 2025; Chai Discovery et al. 2024; Isomorphic Labs 2026; Stark et al. 2025; Pacesa et al. 2024). The sequences obtained by these techniques have to be characterised to obtain a better understanding of how the antibodies interact with the antigen of interest. In a second step called lead optimisation, the initial binders are optimised to increase their binding affinity and improve safety and developability. For example, immunogenicity, instability, self-association, viscosity, polyspecificity or poor expression are properties that may prevent an antibody from being manufactured efficiently or cause toxic side effects (Raybould et al. 2019).

1.3.5 Epitope profiling

An important step during the lead discovery stage is to determine the epitope to which an antibody binds. After phage display or animal immunisation techniques are performed, a large number of immune receptor sequences are obtained, but little about their binding properties is known. In phage display only antibodies against the target antigen are generated. In contrast in animal immunisation, it is often inferred that expanded lineages are target antigen specific. However, neither technique provides information about precise epitopes (Hou et al. 2016; Smith et al. 1997).

Epitopes can be determined experimentally. However, these approaches come at a high cost (Thompson et al. 2020; Hastie et al. 2021; Beck et al. 2013). Computational tools are therefore used to assist epitope profiling. Predicting epitopes from antibody sequence directly was not possible until the very recent breakthroughs in co-folding models (Abramson et al. 2024; The OpenFold3 Team 2025; Passaro et al. 2025).

While these models can successfully predict the structures of antibody-antigen complexes, they currently require the sampling of thousands of candidate structures to achieve high accuracy, making this computationally expensive. Furthermore, their success rates decline for complexes with low similarity to the training data. Prior to the introduction of these co-folding frameworks, the predominant strategy for epitope characterisation was the clustering of antibodies into groups likely to bind the same epitope. This does not directly inform about the epitope but drastically reduces the experimental burden of characterising one representative antibody per cluster. Conventionally, antibodies are clustered by sequence similarity (Greiff et al. 2015; Yassai et al. 2009; Hershberg et al. 2015) as antibodies with high sequence similarity are thought to be clonotypes. A clonotype is a set of antibodies that share the same genetic origin and likely bind the same epitope (Greiff et al. 2015). However, sequence-based methods neglect that through convergent evolution antibodies with highly dissimilar sequences can adopt similar binding site structures and engage the same epitope (Robbiani et al. 2020; Robinson et al. 2021). It is often these sequence dissimilar antibodies that are of interest during lead discovery as they are likely to have distinct developability properties. Methods clustering antibodies by structural similarity have been developed recently and these overcome some of the limitations of sequence based clustering (Robinson et al. 2021; Spoendlin et al. 2023).

1.4 Statistical methods and machine learning

1.4.1 Cluster analysis

Cluster analysis refers to a group of unsupervised data analysis techniques. These methods have the objective of partitioning a dataset into smaller subsets (clusters) such that data points within a subset are more similar to each other than those in different subsets (Jaeger et al. 2023). There are several types of clustering algorithms, each offering different advantages. In this section, I introduce three popular paradigms: centroid-based, hierarchical and density-based clustering. In Chapter 2, these techniques are applied to group antibodies by structural similarity, with the aim of establishing functional equivalence.

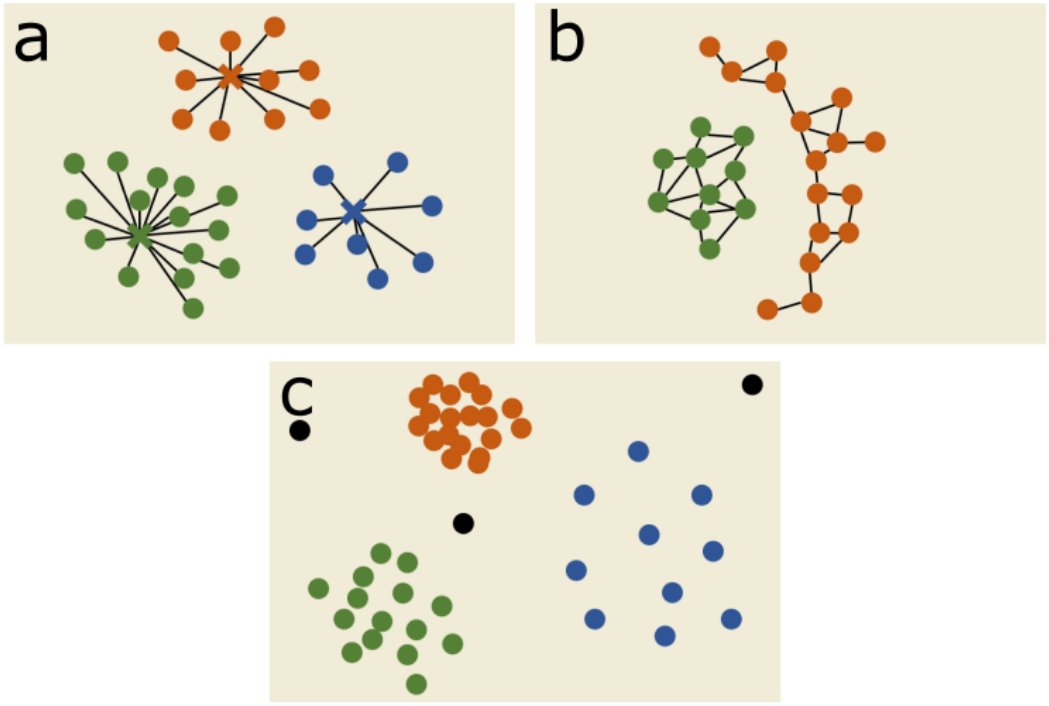


Figure 1.3: Schematic of clustering algorithms. a) Centroid-based clustering. b) Hierarchical clustering. c) Density-based clustering. Dots indicate data points and are coloured by the assigned cluster, black dots show outliers. Cluster centroids are highlighted with a cross.

Centroid-based clustering

Centroid-based clustering is one of the most common types of clustering algorithms (Figure 1.3a). Algorithms of this type choose a number of data points as centroids and link the remaining points to one of the selected centroids. The main challenge often is finding optimal centroids, the remaining data points are then generally assigned to the nearest one. Clusters produced by centroid-based clustering are 'circular' in shape.

K-means (MacQueen 1967) is an example of a centroid-based algorithm. To run the algorithm, the number of clusters (K) has to be pre-specified. K centroids are selected randomly and data points are initially assigned to one of them. The position of the centroids is iteratively updated to minimise the variance within clusters and data points are reassigned to a new centroid. A downside of K-means is that the number of clusters has to be known in advance.

Affinity propagation (Frey et al. 2007) offers a good alternative in case a suitable number of clusters is not known in advance. The algorithm considers all data points as potential centres and messages are passed between pairs of data points determining how well each point would serve as a centroid. Once an optimal set of centroids is found, the remaining points are assigned.

Hierarchical clustering

Hierarchical clustering algorithms (Figure 1.3b) not only produce a single clustering, but assign a cluster hierarchy to each data point. This type of algorithm is best explained by the example of agglomerative clustering (Murtagh et al. 2012). The algorithm starts with each data point assigned to its own cluster. During each iteration, the two clusters closest together are merged. Once the algorithm has finished, it generates a dendrogram that illustrates the hierarchical relationships between the clusters, indicating the iterations during which two clusters were merged. The dendrogram can be cut at any point to obtain one possible clustering. The point at which to cut the dendrogram is usually chosen based on a distance threshold between the linked clusters or a specified number of clusters. One advantage of hierarchical clustering over centroid-based clustering is that clusters can have any shape and are not necessarily circular.

Density-based clustering

Density-based clustering algorithms detect areas of the data space with a high concentration of data points surrounded by sparse areas (Figure 1.3c). The most popular density-based algorithm is DBSCAN (Schubert et al. 2017). This algorithm requires two parameters, a distance d and a scalar N. Circles of radius d are drawn around each data point to define a neighbourhood. Points are then labelled as core points, boundary points or outliers. Core points are defined as having at least N other points in their neighbourhood. Boundary points are in the neighbourhood of a core point but do not qualify as a core point themselves. The remaining points are labelled as outliers. Core points within each other's neighbourhoods or connected

through a chain of neighbourhoods are assigned to a cluster. Boundary points are added to the same cluster as the core points they are connected to.

OPTICS (Ankerst et al. 1999) is another commonly used density-based algorithm. In contrast to DBSCAN, the algorithm is able to detect clusters of varying densities.

An advantage of density-based algorithms is that outliers can be recognised and not all points are assigned to clusters.

1.4.2 Deep learning

Deep learning has revolutionised the analysis of complex and high-dimensional biological data. Neural networks learn the parametrisation of non-linear functions to model the relationships between data and outcomes of interest.

In this section, I introduce the deep learning frameworks central to current protein structure prediction research. I first discuss Graph Neural Networks (GNNs), a model type commonly used when working with representations of protein structures and which I used in Chapters 3 and 5. Then, I introduce generative modelling frameworks, such as diffusion and flow matching, that underpin state-of-the-art protein structure prediction models and form the basis of methods introduced in Chapter 5.

Graph neural networks and geometric deep learning

A graph is a discrete structure used to model objects and their pairwise relations, formally defined as an ordered pair $G = (V, E)$. Here, V represents a set of vertices and E represents a set of edges defining their relationships (e.g. Bishop et al. 2024b; Zhou et al. 2020). Graphs are well suited to model atomic structures as atoms can be represented as vertices and covalent bonds as edges (Duval et al. 2024).

To operate on such structures, a Graph Neural Network (GNN) must consider both node features and their connectivity for prediction. As graphs have no fixed ordering GNNs need to remain independent of the arbitrary indexing of nodes. This requirement is formalised through the concepts of invariance and equivariance (e.g. Bronstein et al. 2021).

A function $f : \mathcal{X} \to \mathcal{Y}$ is invariant under a set of transformations $\mathcal{S}$ if the output remains unchanged regardless of the transformation applied to the input:

$$f(s \cdot x) = f(x) \quad \forall s \in \mathcal{S},\, x \in \mathcal{X}$$

A function $f : \mathcal{X} \to \mathcal{Y}$ is equivariant under a set of transformations $\mathcal{S}$ if applying a transformation to the input x is equivalent to applying that same transformation to the output $f(x)$:

$$f(s \cdot x) = s \cdot f(x) \quad \forall s \in \mathcal{S},\, x \in \mathcal{X}$$

In a standard GNN, node-level updates must be permutation equivariant, while global graph-level predictions must be permutation invariant. Message Passing Neural Networks (MPNNs) achieve this by decomposing transformations into two operations. An 'aggregate' function computes an incoming message m_v for each node v based on its neighbourhood $\mathcal{N}(v)$. This function needs to be permutation invariant which is achieved by applying non-linear operations to node features separately followed by a pooling function independent of input order, e.g. taking the mean.

$$m_v^l = \text{AGGREGATE}_\phi\left(\{h_u^l : u \in \mathcal{N}(v)\}\right) = \frac{1}{|\mathcal{N}(v)|} \sum_{u \in \mathcal{N}(v)} \text{MLP}_\phi(h_u^l)$$

where h_v^l is the feature vector of node v in layer l, and MLP_ϕ is a learnable function (e.g. Bishop et al. 2024b).

An 'update' function subsequently computes a transformed node feature for the next layer, h_v^{l+1}, by integrating the current feature h_v^l with the aggregated message m_v^l. This function needs to be permutation equivariant. Permutation equivariance is preserved as the message m_v^l is invariant and therefore the update operation itself maintains the correspondence between the input and output node indices (e.g. Bishop et al. 2024b).

$$h_v^{l+1} = \text{UPDATE}_\phi\left(h_v^l,\ m_v^l\right) = \text{MLP}_\phi\left(h_v^l + m_v^l\right)$$

To operate on atomic structures, standard GNNs are extended into geometric GNNs. In these architectures, the vertex set V is associated with two distinct feature types. Scalar features $h \in \mathcal{H}$ correspond to standard node features while vector features $\vec{x} \in \mathcal{V}$ represent coordinates, such that $V = (h, \vec{x})$. These models are designed to respect the symmetries of the Euclidean group $SE(3)$, representing global rotations and translations in 3D space. Specifically, updates to scalar features h must be $SE(3)$-invariant, while updates to vector features $\vec{x}$ must be $SE(3)$-equivariant (e.g. Bronstein et al. 2021; Duval et al. 2024).

Vector-scalar GNNs provide a principal framework of such models. In these networks, vectors and scalars are updated by two distinct update functions. To guarantee the required symmetry properties, the interaction between features is restricted to a specific set of operations:

- **Addition:** Permitted only between features of the same type and the result maintains the type (e.g., $\vec{x}_1 + \vec{x}_2 \in \mathcal{V}$).
- **Multiplication:** Valid between different types according to specific algebraic rules. Scalar-scalar multiplication is permitted and results in a scalar ($a \times a \in \mathcal{H}$). Scalar-vector multiplication is permitted and results in a vector ($a \times \vec{x} \in \mathcal{V}$). Vector-vector multiplication via a dot product is permitted and results in a scalar ($\vec{x}_1 \cdot \vec{x}_2 \in \mathcal{H}$).
- **Non-linearities**: applied exclusively to scalars.

By adhering to these algebraic constraints, the network ensures that all scalar outputs are invariant and all vector outputs are equivariant to any transformation $s \in SE(3)$ (e.g. Bronstein et al. 2021; Duval et al. 2024).

This $SE(3)$ symmetry serves as a critical inductive bias for molecular modelling. For instance, the binding affinity of a protein-ligand complex must remain unchanged regardless of its orientation in space (invariance). Conversely, in tasks such as protein-ligand docking, the predicted ligand coordinates must rotate and translate

in accordance with the protein's orientation (equivariance). Consequently, vector-scalar architectures formed the backbone of many models for protein structure prediction (e.g. Jumper et al. 2021; Baek et al. 2021) and structure-based property prediction models (e.g. Abanades et al. 2022; Chinery et al. 2023; Hummer et al. 2023; Spoendlin et al. 2025c). However, the introduction of AlphaFold3 (Abramson et al. 2024) demonstrated that geometric invariances and equivariances can be learned through data augmentation, suggesting they do not necessarily need to be enforced as hard inductive biases within the model architecture.

Generative models

Generative modelling is a branch of deep learning aimed at creating new samples that resemble the training data. Rather than predicting labels, these models learn the underlying probability distribution of a dataset. Several generative paradigms exist and two of which, diffusion and Flow Matching, I will introduce in this section.

Diffusion Models: The basic principle of a diffusion model is to map a simple prior, e.g. Gaussian noise, to the complex data distribution. Sampling from noise from the prior distribution and passing this through the model allows new samples to be generated from the data distribution (e.g. Sohl-Dickstein et al. 2015; Ho et al. 2020; Song et al. 2019).

Diffusion models are trained according to the following methodology (Figure 1.4). A forward process is defined that gradually transforms data into noise. This is typically formulated through a diffusion kernel, where the distribution of the noisy sample $\mathbf{x}_t$ at any time $t \in [0, 1]$ is a Gaussian conditioned on the clean data $\mathbf{x}_0$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I})$$

where α_t is a predefined noise schedule. Under this formulation, $\mathbf{x}_t$ matches the data distribution at $t = 0$, while in the limit $t \to 1$, the distribution approaches a standard isotropic Gaussian, $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The generative task is to learn the reverse

process, $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The model f_θ is trained to predict the mean of this reverse transition, effectively "denoising" the sample:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I})$$

At inference time 'pure' noise is sampled from the prior distribution and iteratively denoised through many model passes until finally a 'clean' sample from the data distribution is obtained (e.g. Bishop et al. 2024a; Sohl-Dickstein et al. 2015; Ho et al. 2020).

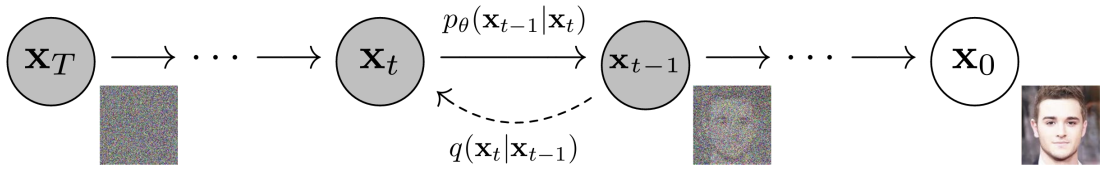


Figure 1.4: Schematic of diffusion process. The forward diffusion process (bottom) adds Gaussian noise to a data sample. For the reverse process (top) a model is trained to denoise a noisy sample and generate a clean data point. Figure reproduced from (Ho et al. 2020)

Flow Matching: A primary limitation of diffusion models is their high computational cost. The stochastic nature of the diffusion kernel requires many small inference steps to account for highly non-linear denoising paths. Flow Matching ameliorates this burden by providing a deterministic framework that encourages 'straighter' integration trajectories, requiring fewer steps during inference (Lipman et al. 2023).

Before diving into the principles of Flow Matching, I will point out that while diffusion conventionally defines $t = 1$ as noise and $t = 0$ as data, Flow Matching typically reverses this: $t = 0$ denotes the noise distribution p_0 and $t = 1$ denotes the data distribution p_1.

In Flow Matching the forward noising process is performed using a linear interpolation between data and Gaussian noise:

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1$$

where $\mathbf{x}_0 \sim p_0$ is a sample from the prior (noise) and $\mathbf{x}_1 \sim p_{\text{data}}$ is a data sample. As in diffusion a model is trained to reverse this process and predict how to denoise samples: $\mathbf{x}_{t+1} \sim \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$. However, in Flow Matching the objective is classically reformulated and expressed as a vector that points from the noise towards the data:

$$\frac{d}{dt}\mathbf{x}_t = v_t(\mathbf{x}_t, t)$$

During inference we integrate along this learned and deterministic vector field v_θ using an Ordinary Differential Equation (ODE) solver. As the Flow Matching objective encourages 'straighter' paths between noise and data than diffusion and no stochasticity is added during the integration process these models can generate highly accurate samples in significantly fewer steps than traditional diffusion (Lipman et al. 2023).

1.5 Protein structure prediction

Over the past five years, the field of protein structure prediction has undergone a fundamental revolution driven by the introduction of deep learning architectures. Prior to these breakthroughs, *de novo* structure prediction remained one of the most significant challenges in computational biology.

Homology modelling approaches were the standard way to predict protein structures before the rise of deep learning. These methods leverage the fact that proteins with high sequence identity fold into similar shapes. A model for a target protein structure can therefore be derived from the experimental structure of a homologous protein. The process typically takes the following steps (Martí-Renom et al. 2000; Hameduh et al. 2020):

1. **Template search:** find homologous template structures
2. **Sequence alignment:** align target with template sequences
3. **Model building:** build model of backbone structure by combining rigid bodies of several templates

4. **Loop modelling:** remodel loop regions using fragment databases and conformation search approaches

5. **Side chain addition:** add side chain atoms to the structure by searching rotamer libraries

6. **Model optimisation:** energy minimisation with molecular dynamics or Monte Carlo simulations

A key limitation of homology modelling is its dependence on the availability of high quality templates. Proteins for which template structures with more than 50% sequence identity are available, modelling is usually highly accurate. However, at lower sequence identity, especially when going below 30%, the quality of models drops off quickly as performing sequence alignments becomes challenging and the conservation of structure may no longer be guaranteed (e.g. Waterhouse et al. 2018; Kryshtafovych et al. 2019).

Homology modelling tools for general proteins (Waterhouse et al. 2018) and antibodies (Leem et al. 2016) are still available but their use has declined since the introduction of ML-based structure predictors.

Since the landmark release of AlphaFold2 (AF2) (Jumper et al. 2021), modelling structures of monomeric, globular proteins has become a routine task. High-quality structures can be obtained for the vast majority of such proteins and predictions are less dependent on sequence identity to previously solved structures. Several structure prediction models have been released following AF2 (e.g. Jumper et al. 2021; Lin et al. 2023; Baek et al. 2021; Ahdritz et al. 2024). In this section, I focus only on the AF2 architecture.

AF2 is a regression model that maps an amino acid sequence to its predicted 3D structure. Under the hood, the model executes a series of operations (Figure 1.5). First, sequence and structural searches are run to construct a Multiple Sequence Alignment (MSA) and identify homologous structural templates. The target sequence, MSA and templates serve as the three primary model inputs. The target sequence and MSA initialise the 'MSA representation', while residue

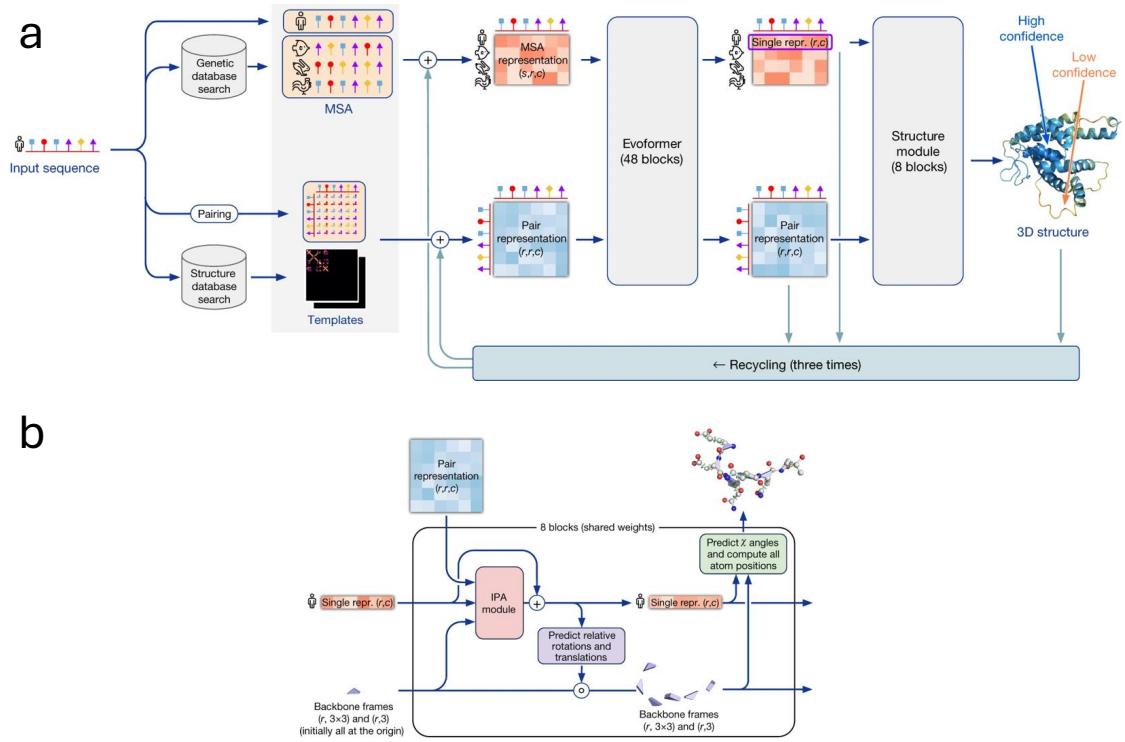


Figure 1.5: Overview of AF2 architecture. a) Schematic of the complete AF2 model. b) Detailed schematic of the AF2 structure module. Figure reproduced from (Jumper et al. 2021).

contacts are extracted from the template structures and used to initialise the 'pair representation'. These representations are iteratively updated through the Evoformer module, where the MSA representation detects co-evolutionary correlations to refine the residue contacts in the pair representation. Subsequently, the pair representation and the first row of the MSA, referred to as the 'single representation', are processed by the Structure Module to generate 3D atomic coordinates. This module utilises a 'frame representation', where each amino acid residue is treated as a rigid body defined by a position vector and a rotation matrix. This frame encodes the orientation of all backbone atoms. Simultaneously, the model predicts side-chain torsion angles, allowing for the reconstruction of an all-atom protein structure (Jumper et al. 2021).

While AF2 (Jumper et al. 2021) and its contemporaries (Lin et al. 2023; Baek et al. 2021; Ahdritz et al. 2024) achieved significant breakthroughs in predicting monomeric proteins, they remain inherently limited when modelling large protein

assemblies and complexes involving non-protein ligands (e.g. DNA or small molecules). These constraints are rooted in specific methodological choices. First, the AF2 Evoformer module operates on the entire MSA representation creating a major computational bottleneck that restricts system size. Second, the structure module can only represent amino acid residues.

A primary contribution of AlphaFold3 (AF3) (Abramson et al. 2024) and similar architectures (e.g. Wohlwend et al. 2024; Chai Discovery et al. 2024; The OpenFold3 Team 2025; Hayes et al. 2025) was to resolve these two limitations, enabling the prediction of diverse molecular assemblies. As illustrated in Figure 1.6, AF3 introduces several innovations. In addition to the standard AF2 inputs, a conformer library is searched to provide a fourth input relevant for small molecules. To alleviate the computational burden, the 'trunk' of the model is restructured. Intensive operations on the full MSA are confined to a smaller MSA module, while the majority of the processing occurs in the Pairformer, which operates only on pair and single representations. The Structure Module has been entirely replaced by a Diffusion Module. While the transition to a generative diffusion framework represents a fundamental shift from the deterministic approach of AF2, a parallel and equally critical development lies in the atomic representation. By representing all atoms directly through their 3D coordinates, the Diffusion Module achieves the geometric flexibility required to model non-protein molecules that was previously not possible. A further innovation is that the structure module utilises standard transformer layers, removing all inductive biases for geometric invariances. Instead, these properties are learned entirely through data augmentation.

Structure predictors like AF3 can be used to model any protein, however, a specialised lineage of antibody specific models has emerged (e.g. Abanades et al. 2022; Ruffolo et al. 2022a; Ruffolo et al. 2022b; Kenlay et al. 2024; Abanades et al. 2023). Because somatic hypermutation limits the utility of co-evolutionary information in antibodies, these models bypass the MSA completely which significantly accelerates prediction speeds.

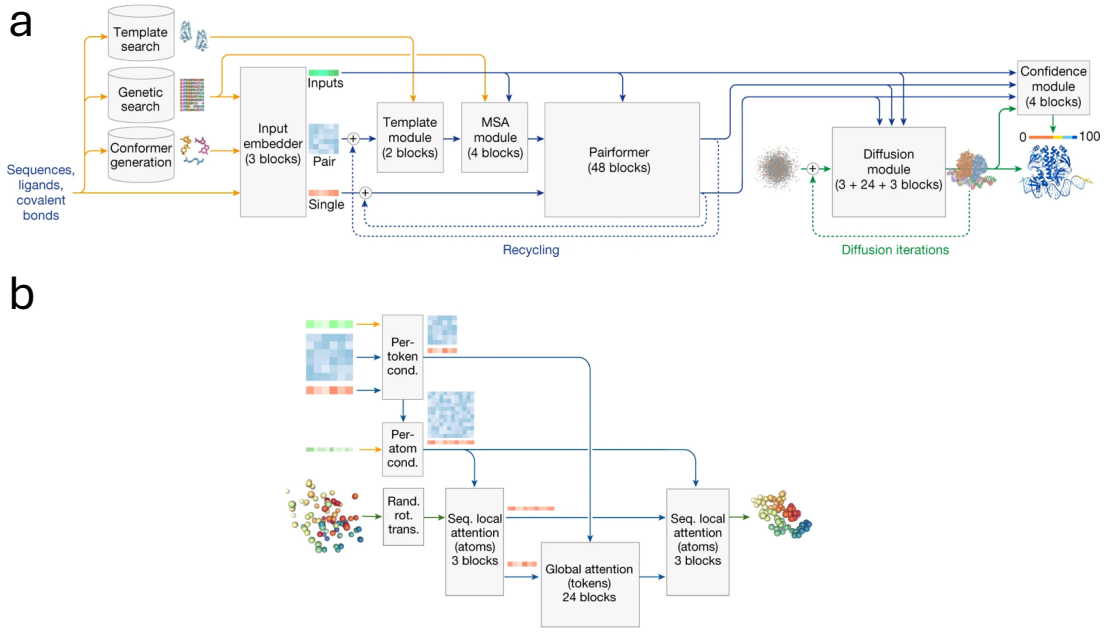


Figure 1.6: Overview of AF3 architecture. a) Schematic of the complete AF3 model. b) Detailed schematic of the AF3 structure module. Figure reproduced from (Abramson et al. 2024).

1.6 Protein dynamics

Many proteins are flexible molecules and change their 3-dimensional shapes in solution. These distinct structural configurations are termed conformations and transitions between them underpin nearly all biological activities (Teilum et al. 2009).

Structural flexibility allows proteins to sense and respond to changes in the cellular environment and the presence of other molecules. Conformational changes are central to signal transduction events. For example, the transition of a cellular receptor from a bound to an unbound state relays a signal across the membrane (Wang et al. 2018). Similarly, the regulation of enzyme function often relies on structural shifts. In allosteric protein regulation, binding of a molecule to a distant binding site triggers a conformational change that modulates the activity of the protein (Hilser et al. 2012). Furthermore, enzymatic activity may itself depend on rearrangement of residues to facilitate chemical interactions with the substrate (Teilum et al. 2009).

The temporal scale of these motions is remarkably broad, spanning approximately

15 orders of magnitude (Figure 1.7). At the fastest extreme, bond vibrations occur on the picosecond scale. Secondary and tertiary structure transitions, which are often functionally the most significant, happen within the microsecond to millisecond range. At the slowest end of the spectrum, global events such as protein folding may in some cases take several minutes or even hours (Teilum et al. 2009; Brennan et al. 2025).

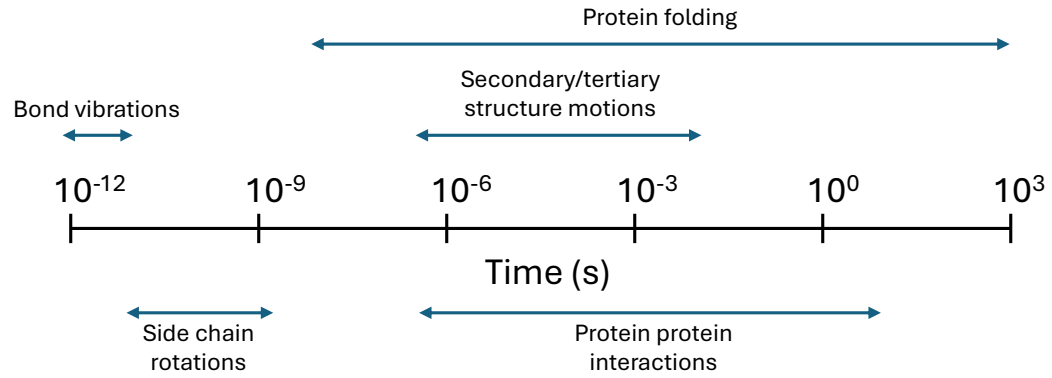


Figure 1.7: Timescales of protein dynamics. Schematic representation illustrating the temporal spectrum of molecular motions, ranging from bond vibrations (picoseconds) to collective secondary and tertiary structure rearrangements (microseconds to milliseconds) and protein folding (up to hours).

1.6.1 Conformational flexibility of antibodies

Antibodies exhibit several types of conformational changes across their domains which impact different functional properties. In this section, I will focus on the functional importance of structural dynamics in the Fv region. As discussed in Section 1.3.1, Fvs consist of a framework formed by a stable β-sandwich fold and CDR loops connecting the β-strands. While the frameworks remain more static to support stability (Julian et al. 2017), the majority of structural diversity is concentrated in the CDRs (Kroon et al. 2003; Fernández-Quintero et al. 2019b). CDR flexibility impacts antigen binding interactions and therefore has been linked to key functional properties (Fernández-Quintero et al. 2019c; Manivel et al. 2000; Krishnan et al. 2008; James et al. 2003; Jeliazkov et al. 2018).

CDRs tend to be more dynamic in the unbound state and some antibodies have been observed to adopt several conformations in solution (e.g. Fernández-Quintero

et al. 2019a). Upon antigen binding the flexibility is often reduced and the antibody tends to adopt a single conformation (e.g. Fernández-Quintero et al. 2019a). These observations provide support for conformational selection, a mechanism widely proposed to govern the function of diverse protein systems (Ma et al. 1999; Kumar et al. 2000). This model assumes that a protein in solution exists as a pre-existing equilibrium of multiple conformational states, including those that are binding-competent. Upon interaction with a ligand, the equilibrium shifts to favour the specific state required for binding. Conformational selection appears to be prevalent in antibodies; however, some observations suggest that induced fit remains an alternative mechanism used by antibodies in rare cases (Liu et al. 2024). Unlike conformational selection, the induced fit model implies that the binding-competent state is not significantly populated within the unbound ensemble. Instead, the initial interaction with the antigen is required to stabilise the final functional conformation.

CDR conformational flexibility can influence binding affinity (Mikolajek et al. 2022). From a thermodynamic perspective, dynamic residues possess higher conformational entropy than static ones. As antigen binding frequently rigidifies the CDR loops (Kroon et al. 2003; Fernández-Quintero et al. 2019a), it impacts the entropy of the antibody. It is often inferred that this loss of mobility imposes a significant entropic cost, thereby reducing the overall binding affinity. Supporting this view, several studies have identified the rigidification of CDRs and pre-configuration of specific conformational states as a natural mechanism used by antibodies to increase their affinity (Schmidt et al. 2013; Li et al. 2015). However, the relationship between flexibility and affinity can be more complicated. The total entropic change of the system must also account for the dynamics of the antigen and the surrounding solvent. A high entropic cost on the antibody side can be offset by an entropic gain in the antigen or through the release of ordered water molecules from the binding interface (Teilum et al. 2009; Mikolajek et al. 2022).

Polyspecificity and cross-reactive antigen recognition are further properties that have been linked to CDR flexibility (Guthmiller et al. 2020; Fernández-Quintero et al. 2019c). Although the molecular mechanisms underlying promiscuous binding remain

incompletely understood, several case studies suggest that structural flexibility is one of its drivers. For example, certain antibodies have been observed to bind to distinct antigens by adopting different conformations (James et al. 2003). MD studies have provided further support, demonstrating that in six case studies antibodies characterised by shallow, broad conformational landscapes were more promiscuous than those restricted to a narrow ensemble (Fernández-Quintero et al. 2019c). Paratope plasticity not only facilitates the recognition of distinct antigens but also allows antibodies to adapt to structural variations in imperfectly conserved epitopes. This flexibility is a key mechanism to accommodate mutational changes and enable the broad neutralisation of rapidly evolving viruses (Guthmiller et al. 2020).

The natural immune system may exploit CDR conformational flexibility to balance between polyspecific, low-affinity binding and specific, high-affinity interactions. In case studies of antibody lineages, affinity maturation was shown to introduce mutations that promote CDR rigidification and thereby narrow specificity whilst enhancing binding strength (Schmidt et al. 2013; Babor et al. 2009; Li et al. 2015; James et al. 2003; Blackler et al. 2022; Fernández-Quintero et al. 2019c). Thus, CDR flexibility may support a two-stage strategy: the initial promiscuous binding of a broad range of antigens by germline antibodies, which increases the effective size of the immune repertoire, followed by selection for more specific, high-affinity binders. Although conformational flexibility may be important in modulating functional properties of some antibodies, its functional relevance and applicability at the repertoire level remain unclear (Jeliazkov et al. 2018; Cagiada et al. 2025b).

Specificity and affinity are two critical properties of antibody therapeutics. In order to maximise target and minimise off-target interactions a therapeutic should have high affinity and specificity (Sormanni et al. 2018). This suggests therapeutics should have a preference towards rigidity. However, as conformational flexibility may allow better recognition of mutated antigen variants it could be desired when designing broadly neutralising antibodies (Li et al. 2022; Prigent et al. 2018).

1.6.2 Methods to study protein dynamics

Protein dynamics impact a range of functional properties and occur across a broad temporal scale. To study the full scope of conformational dynamics a suite of experimental and computational techniques have been developed (Wei et al. 2016).

Experimental methods

Conformational flexibility can be studied with several experimental methods. Nuclear magnetic resonance (NMR) spectroscopy (Kleckner et al. 2011), small-angle X-ray scattering (SAXS) spectroscopy (Gupta et al. 2022) and hydrogen-deuterium exchange mass spectrometry (HDX-MS) (Konermann et al. 2011) report on protein dynamics in solution although they do not typically provide atomically resolved flexibility information. X-ray crystallography and cryo-EM can capture conformational states with atomic resolution but are under explored (Ellaway et al. 2023; Zhong et al. 2021).

NMR directly reports on structural flexibility on a range of temporal scales. Changes in the chemical environment of a residue, as a result of conformational rearrangements, directly influence the line shape and chemical shifts in standard NMR experiments (Figure 1.8). In the slow exchange regime (1 s^{-1}), no exchange occurs during the duration of the experiment and signals from all conformations give rise to separate peaks. In the fast exchange regime ($100,000$ s^{-1}), rapid exchange occurs during the experiment which results in a single, sharp peak in between the chemical shifts of each of the states. For exchange on an intermediate timescale (1000 s^{-1}) 'exchange broadening' is observed, whereby the width of peaks gets larger and their amplitude smaller (Kleckner et al. 2011). More complicated pulse sequences, such as CPMG, have been developed to study protein dynamics in more detail (see Kleckner et al. (2011) for a review).

SAXS measures a scattering signal that relates to the pairwise distances of all electrons within a protein. From this signal, the radius of gyration and the maximum particle dimension can be determined, allowing the overall shape of a protein to be inferred. Because the experimental signal is averaged across all

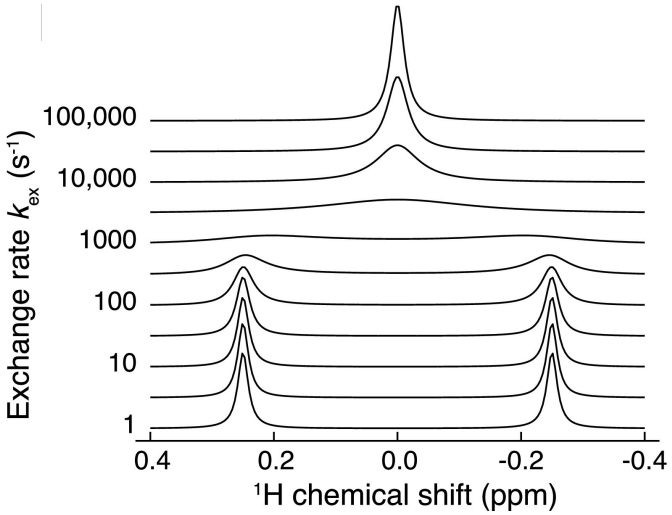


Figure 1.8: Effect of chemical exchange on NMR line shapes. The x-axis represents the hypothetical chemical shifts of a hydrogen nucleus which exchanges between two conformational states. Figure reproduced from (Waudby et al. 2023).

populated conformational states, it is possible to determine whether a specific conformational ensemble fits the experimental data (Gupta et al. 2022).

HDX-MS is based on the principle that hydrogen atoms exhibit varying rates of exchange when the protein is exposed to D_2O. Hydrogens in disordered regions have greater solvent accessibility and exchange faster than those in tightly folded structural elements (Konermann et al. 2011). Consequently, HDX-MS can be used to detect disordered protein regions in applications such as monitoring folding dynamics (Ferrari et al. 2025).

X-ray crystallography remains the gold standard for obtaining protein structures with atomic resolution. However, the crystallisation process generally requires highly homogeneous protein samples. This requirement limits the technique's ability to capture multiple conformational states within a single crystal. Solving separate structures of the same protein under different conditions is one way to identify distinct conformational states (Ellaway et al. 2023).

Cryo-EM has emerged as a powerful alternative for solving protein structures. While historically constrained by lower resolutions, methodological advances now enable the determination of structures with resolution comparable to crystallography (Kühlbrandt 2014). As the requirement for sample homogeneity is significantly

less stringent for cryo-EM diverse conformational states can be captured in a single imaging experiment. A range of recent works have focused on algorithms to reconstruct structures of multiple conformational states from heterogeneous particle images (Punjani et al. 2023; Zhong et al. 2021; Ducrocq et al. 2024).

Molecular dynamics simulations

Molecular dynamics (MD) simulations are a computational approach for modelling the time-dependent trajectories of atoms within a biomolecular system, such as a protein. Capturing these motions, MD enables the study of complex molecular behaviour, e.g. protein folding and ligand binding, at atomic resolution (Hollingsworth et al. 2018).

In classical all-atom simulations, the motion of individual atoms is described by Newton's second law, $F = ma$, where the force F acting on an atom is the product of its mass m and its acceleration a. In practice, MD algorithms calculate interatomic forces within a system to determine the acceleration acting on an atom and update its velocity. These velocities are subsequently used to compute atomic displacements over a discrete integration time step (Hollingsworth et al. 2018). These forces are defined by an underlying force field, with common variants including AMBER (Case et al. 2025) and CHARMM (Brooks et al. 2009). Potentials typically account for bonded interactions, such as covalent bond stretching and angle vibrations, as well as non-bonded interactions, including electrostatic and Van der Waals forces. As force field parameters are semi-empirical, derived from fits to experimental data, they remain approximations, which introduces an inherent degree of uncertainty into the simulations (Hollingsworth et al. 2018; Cagiada et al. 2026).

A primary bottleneck of all-atom MD is the immense computational cost required to simulate biologically relevant events. To maintain numerical stability, an integration time step in the femtosecond range (10^{-15} s) has to be used, which is several orders of magnitude shorter than the timescales of functional protein dynamics (see Figure 1.7) (Hollingsworth et al. 2018). Even when using specialised supercomputing architectures (Shaw et al. 2021), constraints remain on both the

achievable simulation length and the total number of systems that can be studied. Consequently, even the largest all-atom datasets currently available, such as ATLAS (Vander Meersche et al. 2024) and mdCATH (Mirarchi et al. 2024), contain far fewer entries than the total number of experimentally determined protein structures.

To address the timescale limitations of all-atom simulations, several strategies have been described. Sampling conformational states separated by high free-energy barriers requires especially long simulations due to low transition probabilities. Enhanced sampling methods such as Metadynamics (Laio et al. 2002) overcome this by introducing a biasing potential to the force field. This bias effectively penalises previously sampled states, thereby 'pushing' the system to explore novel regions of the conformational landscape.

Coarse-grained (CG) simulations provide a more compute efficient alternative to all-atom MD. Rather than modelling every atom individually, these methods represent groups of atoms as single 'beads.' This reduction in the number of particles massively decreases the number of pairwise force calculations required, thereby accelerating the simulation significantly. However, CG simulations may suffer from reduced accuracy introduced by the approximation of the underlying force field (Kmiecik et al. 2016; Jin et al. 2022). One-bead-per-residue models, such as CALVADOS (Cao et al. 2024), have been used to create datasets as large as tens to hundreds of thousands of proteins (Tesei et al. 2024; Cagiada et al. 2025b).

Given their ability to capture structural fluctuations, MD simulations have been used widely to characterise antibody dynamics (e.g. Wong et al. 2011; Fernández-Quintero et al. 2019c; Guloglu et al. 2023).

Machine learning models

Building on the success of machine learning in protein structure prediction (Jumper et al. 2021; Abramson et al. 2024; Baek et al. 2021; Passaro et al. 2025), there has been growing interest in applying similar methodologies to predict conformational ensembles. Machine learning models offer the possibility to sample the equilibrium distribution of conformational states directly thereby bypassing the need for the

expensive computation of time-dependent trajectories. However, training ML models for this purpose remains a significant challenge. Predicting a diverse ensemble of conformational states is inherently more complex than predicting a single, rigid structure. Furthermore, much less data capturing conformational ensembles is available than 'static' protein structures in the PDB making the problem even more difficult to address.

Initial approaches to leverage ML models for predicting conformational ensemble focussed on modifying the AF2 (Jumper et al. 2021) inference procedure. AF2 inputs include an MSA generated for the sequence of the protein to be modelled (see Section 1.5 for details). The model learns co-evolutionary patterns from the MSA to infer structural constraints. In other words, if mutations in two residues correlate with each other they are likely to be in proximity. The MSA should, at least theoretically, encode such co-evolutionary signals not just for a single structural state but for all stable conformations. Several workflows have attempted to deconvolve this information in order to improve prediction of distinct conformational states with AF2. While the detailed implementation differs, these approaches generally manipulate the input MSA to encourage more diverse output structures (e.g. (del Alamo et al. 2022; Stein et al. 2022; Wayment-Steele et al. 2024; Sala et al. 2023b; Faezov et al. 2023; Sala et al. 2023a; Saldaño et al. 2022)). Strategies involve random subsampling (del Alamo et al. 2022), sequence-based clustering (Wayment-Steele et al. 2024) and introduction of artificial mutation to destabilise the dominant conformation (Stein et al. 2022). While these strategies generally succeed at increasing the diversity of predicted structures and reproducing conformational states for some proteins their general accuracy remains limited (Riccabona et al. 2024; Faezov et al. 2023; Sala et al. 2023a; Saldaño et al. 2022).

Following the release of AF3 (Abramson et al. 2024), the most recent generation of protein structure predictors has turned to generative modelling paradigms. While this has led to improvements in protein structure prediction (Abramson et al. 2024; Passaro et al. 2025; The OpenFold3 Team 2025), the generative framework also offers advantages for the conformation prediction task. In contrast to AF2-style

regression models, which learn a one-to-one mapping from sequence to structure, generative approaches model a distribution over possible outputs which is precisely the objective required for predicting multiple conformational states. State-of-the-art structure predictors are exclusively trained on crystal structure data and, therefore, despite the generative frameworks produce a limited diversity of outputs (Abramson et al. 2024; Wohlwend et al. 2024).

A range of studies have extended the ability of generative structure predictors to model conformational ensembles through the introduction of a training step on MD data (Lewis et al. 2025; Jing et al. 2024a; Janson et al. 2025; Passaro et al. 2025; Zhu et al. 2024; Zheng et al. 2024). The BioEmu model (Lewis et al. 2025) is among the best performing models currently. This model was trained in three stages. Initial training used the AF2 structure database (Varadi et al. 2024) which was shown to introduce significant diversity into the predictions. In a second stage a large number of MD simulations totalling more than 200 ms of simulation time were used to focus conformational sampling towards physically accurate states. In a final third stage, folding free energy measurements were used to improve the fraction of folded and unfolded samples. BioEmu has been shown to model various types of conformational changes including domain motions, local unfolding and cryptic pocket formation for certain proteins (Lewis et al. 2025).

Most recently, inference time modification of structure predictors have gained attention again. Guidance and steering of AF3-style model predictions towards lower-resolution experimental constraints has shown promising results for sampling desired conformations. For example, the ROCKET method (Fadini et al. 2026) optimises MSA profiles to fit output structures to cryo-EM maps of a target state. Similarly, the ConforMix method uses a constraint-based guidance function during the reverse diffusion process to sample from a conditional distribution (Richman et al. 2025)

Using generative models to predict MD trajectories, rather than direct sampling of equilibrium distributions, is another area of interest (Jing et al. 2024b; Xu et al. 2025). Time-resolved trajectories convey richer functional information, e.g. by

showing transition pathways. However, they inherit limitations from MD of time-correlated sampling, meaning that long simulations are required to approximate equilibrium distributions.

While data suggest that generative approaches have surpassed original AF2 MSA-subsampling strategies in accuracy, training a highly reliable and generalisable model remains an open challenge (Lewis et al. 2025; Janson et al. 2025; Jing et al. 2024a). Therefore, despite the rapid advancement of such machine learning methods, all-atom MD simulations continue to serve as the gold standard for the computational modelling of conformational states.

1.7 Thesis outline

This thesis describes my contributions to the development of computational methods for the functional characterisation of antibodies using data on protein structures and structural flexibility.

In Chapter 2, I describe the development of an epitope profiling algorithm which clusters antibodies by the similarity of structural models. I benchmarked against more conventional algorithms that perform clustering based on sequence similarity and demonstrated that the structure-based approach identifies a broader range of antibodies that bind to the same epitope.

Moving on from static structures, Chapters 3-5 introduce three methods to model the conformational flexibility of antibody Fvs. In Chapter 3, I describe an approach to search antibody and protein structure databases for evidence of conformational flexibility of CDRs and structurally related loop motifs. I then introduce a model to binary classify CDR3s by their ability to adopt a single (rigid loop) or multiple (flexible loop) conformational states.

Chapter 4 outlines the development of a high-throughput MD workflow for antibodies. A methodology based on a coarse-grained force field is tuned to reproduce CDR dynamics in all-atom simulations. I then used the workflow to simulate more than 150,000 antibodies.

In chapter 5, I introduce a generative protein structure prediction model trained on the coarse-grained MD data and a smaller set of all-atom simulations to enable sampling of multiple conformational states. Benchmarking against similar models not trained on large amounts of antibody data shows the best sampling of antibody conformational ensembles.

Finally, in Chapter 6 I summarise the main finding in this thesis and suggest avenues for further investigation.

2

Computational epitope profiling using structural models

Contents

2.1 Motivation 34
2.2 Contributions 34
2.3 Background 35
2.4 Methods 38
2.4.1 Data sets 38
2.4.2 SPACE1 39
2.4.3 SPACE2 40
2.4.4 Numbering scheme and region definitions 42
2.4.5 Analysis of structural clusters 42
2.5 Results 45
2.5.1 Evaluating an optimal clustering algorithm 45
2.5.2 Examining the behaviour of agglomerative clustering across different structural similarity thresholds 47
2.5.3 Evaluating the optimal region for clustering 49
2.5.4 SPACE2 performs well on sets of antibodies against diverse targets 50
2.5.5 Advances in structure prediction improve structure-based computational epitope profiling 53
2.5.6 SPACE2 improves coverage compared to clonotyping . . 54
2.5.7 SPACE2 identifies functional convergence signals 56
2.6 Discussion 59

2.1 Motivation

Determining what antigen and epitope antibodies engage is a key step when studying immune repertoires and designing antibody therapeutics. A range of computational techniques have been developed by the scientific community that group antibodies likely to engage the same epitope. These methods allow inference of antigen binding within a group if specificity for one of the members is known. Under the assumption that similar sequences lead to similar function, conventional algorithms cluster antibodies by sequence identity. While these methods produce highly consistent groupings, they are unable to cluster together the full breadth of antibodies against the same epitope as diverse sequences can exhibit similar binding site geometries and engage common targets.

To address these limitations, previous work in the group developed SPACE1, an algorithm that clustered antibodies based on structural similarity of homology models. In this chapter, I describe the development of SPACE2, which improves the original protocol through the usage of predicted structural models, which are more accurate than homology models, and systematic optimisation of the clustering algorithm.

2.2 Contributions

I performed the data preparation and the evaluation of the SPACE2 clustering method. The SPACE2 code was written in collaboration with Brennan Abanades. Matthew Raybould contributed to the benchmarking of the SPACE1 method. Roche performed the clustering for the three non-public datasets.

This chapter contains material reproduced from:

Fabian C. Spoendlin, Brennan Abanades, Matthew I. J. Raybould, Wing Ki Wong, Guy Georges, and Charlotte M. Deane. “Improved Computational Epitope Profiling Using Structural Models Identifies a Broader Diversity of Antibodies That Bind to the Same Epitope.” Frontiers in Molecular Biosciences 10 (2023).

2.3 Background

As discussed in Chapter 1, antibody function is closely linked to the epitope it engages, therefore studying epitopes is essential for the understanding of immunology and developpment of antibody drugs. Both experimental and computational approaches to determine epitopes are available.

Epitopes can be determined at high resolution by solving the structure of an antibody in complex with its antigen. However, structure determination methods are too resource intensive to be used to explore large data sets (Nilvebrant et al. 2018). Experimental epitope binning methods, such as competition assays (Abdiche et al. 2009), scale better, but it remains difficult to analyse very large data sets as costs grow at $O(n^2)$ with the number of antibodies (n) to be evaluated. Competition assays also only offer low resolution as they struggle to distinguish between antibodies that bind distinct sites but that overlap sterically.

Prior computational clustering of antibodies into functional groups that engage the same epitope can reduce the number of experiments that need to be run or even remove the need for experimental epitope determination entirely. Most computational epitope profiling methods group antibodies based on sequence similarity. Clonotyping, the most widely used method, attempts to link antibodies that originate from the same progenitor B-cell (Lopez-Santibanez-Jacome et al. 2019; Greiff et al. 2015). The exact definition of a clonotype varies across the literature. Commonly antibodies that originate from the same heavy chain V and J genes, match in CDRH3 length and exceed a threshold CDRH3 sequence identity are considered a clonotype. Threshold values between 80 and 100% have been reported. To introduce additional leniency the requirement for matching J genes can be neglected (Greiff et al. 2015). Clonal clustering is usually highly accurate, as antibodies within a cluster tend to engage the same epitope.

Clonotyping was originally intended to trace lineages of antibodies within an individual. Its use in functional clustering thus makes the assumption that antibodies against a given epitope must originate from progenitor B-cells with shared genetic origins. However, antibodies from different lineages and with highly dissimilar

sequences can adopt a similar binding site geometry and engage the same epitope (Robinson et al. 2021; Wong et al. 2021; Scheid et al. 2011; Rijal et al. 2019; Joyce et al. 2016). The ability to determine functional convergence is especially important when comparing the immune response of individuals, as different individuals exhibit personalised immunoglobulin gene usages (Briney et al. 2019). As clonotyping is not able to link antibodies from distinct genetic lineages, it loses power when analysing antibodies originating from different sources.

Alternative methods have been developed to try and identify functionally equivalent antibodies that are not similar in sequence. Clustering antibodies by sequence similarity across predicted paratope residues can link antibodies from different clonotypes (Richardson et al. 2021). However, methods that consider structural similarity to cluster antibodies are even better suited to detect less related sequences with functional convergence, because binding site structure provides more direct evidence of antibody function than its sequence. Several methods are available that attempt to functionally link antibodies based on a representation containing structural information in addition to physicochemical properties of paratope residues (Wong et al. 2021; Ripoll et al. 2021).

In a previous study, the SPACE1 method (Robinson et al. 2021) was described, which clusters antibodies based on structural similarity of homology models. The algorithm accurately clusters antibodies that bind the same epitope and is able to functionally link antibodies with diverse sequences. However, SPACE1 is limited by the coverage of homology modelling (in the original study only 73% of the data could be modelled to a usable standard) and its inaccuracies. The method was also only benchmarked at the level of domain consistency on one virus class. Recent progress in machine learning-based antibody structure prediction has led to more accurate structural models than those obtained with homology based approaches, especially in cases where no template with high sequence similarity is available (Jumper et al. 2021; Baek et al. 2021; Lin et al. 2022; Abanades et al. 2023; Abanades et al. 2022; Ruffolo et al. 2020; Ruffolo et al. 2022a; Ruffolo et al. 2022b). Higher accuracy

and higher confidence in structural models also allow increased coverage and have the potential to improve structure-based epitope profiling.

In this chapter, I present the Structural Profiling of Antibodies to Cluster by Epitope 2 (SPACE2) algorithm. SPACE2 builds on recent progress in machine learning-based antibody structure prediction and uses a novel clustering protocol systematically optimised and extensively benchmarked on epitope-resolution binding data. I show that SPACE2 outperforms SPACE1 by improving data coverage and identifying clusters even more diverse in sequences, genetic lineages and species origin. These results underline that structural data, which can now be rapidly and easily generated through structure prediction tools, contains orthogonal functional information to sequence and should be considered when investigating antibody function.

2.4 Methods

2.4.1 Data sets

Six data sets of antigen specific antibodies were used to analyse SPACE2 clustering performance.

The training set on which the clustering algorithm, thresholds and antibody region were set consisted of 3051 antibodies against the SARS-CoV-2 receptor binding domain (RBD). Antibodies were annotated with groups of overlapping epitopes originating from mutation escape profiling (Cao et al. 2023). I refer to this dataset as the Cao et al. (2023) training set throughout the paper.

CoV-AbDab (Raybould et al. 2021a), a data set of anti-lysozyme antibodies, a non-public data set of antibodies against Ebola viruses (EV) and two non-public data sets of antibodies against non-viral targets (NVA1 and NVA2) were used as additional data sets to evaluate SPACE2. CoV-AbDab is a database of antibodies against coronavirus antigens, such as those from SARS-CoV-2, SARS-CoV-1 and MERS-CoV. A version of CoV-AbDab timestamped the 3rd October 2022 was used containing 10,719 antibodies with sequence data. As CoV-AbDab is a collection of antibodies reported in the literature it contains the Cao et al. (2023) training set. When using CoV-AbDab as a test set (denoted as CoV-AbDab (test)) the training set was removed and only the remaining 7685 antibodies included. Epitope data in CoV-AbDab is reported as in the original publications and ranges from antigen to domain level.

A data set of anti-lysozyme antibodies was created from all 53 lysozyme specific antibodies in the structural antibody database (SAbDab) (Schneider et al. 2022; Dunbar et al. 2014) for which the antibody-antigen complex structure has been solved. Antibodies were grouped by their epitope using the Ab-ligity method (Wong et al. 2021) and annotated as binding to the same epitope if their Ab-ligity score was greater than a threshold of 0.1 (as in the original paper). Similarity of epitopes within an epitope group was confirmed by visual inspection.

The EV set contains 126 antibodies with epitope data ranging from antigen to domain level. The NVA1 set contains 31 antibodies with epitope data from competition assays. NVA2 contains 33 antibodies with epitope data from mutation escape profiling.

2.4.2 SPACE1

The original SPACE1 method clusters antibodies by the structural similarity of homology models. The algorithm was run as detailed in Robinson et al. (2021).

Homology models were produced using ABodyBuilder (Leem et al. 2016). ABodyBuilder uses structures from a database to build its models. In this study, I used quality-filtered SAbDab (Schneider et al. 2022; Dunbar et al. 2014) entries timestamped before 6$^{\text{th}}$ July 2022. Quality filtering restricted structures to those solved by X-ray crystallography, and excluded structures with a resolution of >2.5 Å and structures containing residues with a B-factor >80. In a standard ABodyBuilder run the method first attempts to model CDR loops with a template database search method (Choi et al. 2010). If no suitable template is found for a CDR, hybrid homology/*ab initio* modelling is performed (Leem et al. 2016). Only models for which homology templates for all six CDR loops were found are used for clustering in the SPACE1 method to keep the models as accurate as possible.

The remaining homology models are clustered by structural similarity of CDRs. Models are split into groups of antibodies with identical CDR lengths. Antibodies in each group are then clustered using a greedy clustering algorithm. The first antibody in the group was selected as the cluster centre and all antibodies with a CDR C_α RMSD smaller than a specified threshold after alignment of framework residues were added to the cluster. After all antibodies have been compared against the first cluster centre the algorithm selects the next unclustered antibody as a new cluster centre and cluster members are chosen as in the previous step. In addition to the RMSD threshold of 0.75 Å suggested by Robinson et al. (2021), I also assessed the performance at a 1.25 Å threshold.

2.4.3 SPACE2

The SPACE2 algorithm clusters antibodies by the similarity of models obtained from an ML-based structure prediction tool. The method functions in four main steps. Initially, a structural model of the antibody Fv is produced using ABodyBuilder2 (Abanades et al. 2023). ABodyBuilder2 is a deep-learning based tool for antibody structure prediction and was trained on SAbDAb structures timestamped up to 31st July 2021. The models are then split into groups of identical CDR lengths. Models in each group are then structurally aligned on the Cα of residues in framework regions and a pairwise distance matrix is computed of the Cα RMSDs of CDR loop residues. The antibodies are then clustered based on these distances.

Clustering algorithms

Eight different clustering algorithms were explored (agglomerative clustering, affinity propagation, DBSCAN, OPTICS-xi, OPTICS-DBSCAN, K-means, Butina clustering, greedy clustering). Agglomerative clustering (Murtagh et al. 2012), affinity propagation (Frey et al. 2007), DBSCAN (Schubert et al. 2017), OPTICS-xi, OPTICS-DBSCAN (Ankerst et al. 1999) and K-means (MacQueen 1967) were performed using the scikit-learn (Pedregosa et al. 2011) implementation. Butina clustering (Butina 1999) was performed using the RDKit implementation (Landrum 2006). A greedy clustering algorithm, which groups antibodies according to the algorithm described in Section 2.4.2, was implemented.

Parameters and evaluated ranges for each algorithm are shown in Table 2.1. The K-means algorithm requires an additional parameter (K) which corresponds to the predetermined number of clusters. K was set to the number of clusters obtained from agglomerative clustering using the best-performing parameters.

SPACE2-HC

A variation of the SPACE2 algorithm was implemented that clusters antibodies based on the structural similarity of heavy chains only (SPACE2-HC). The light chains were included for the modelling step, as ABodyBuilder2 (Abanades et al.

Table 2.1: Clustering algorithms and parameter ranges/values evaluated during optimisation.

Algorithm	Parameter	Range/values	Optimal value
Greedy clustering	RMSD threshold	0.5 - 10 Å	1.25 Å
Agglomerative clustering	RMSD threshold	0.5 - 10 Å	1.25 Å
	linkage criterion	complete, average, single	complete
Affinity propagation	Preferences	-5 - 4	Median(RMSD)
DBSCAN	RMSD threshold	0.5 - 5 Å	1 Å
	minimum samples	2, 5	2
OPTICS-xi	RMSD threshold	0.5 - 5 Å	1.5 Å
	minimum samples	2	2
	xi	0.005 - 0.5	$\leq$0.01
OPTICS-DBSCAN	RMSD threshold	0.5 - 2 Å	1 Å
	minimum samples	2	2
K-means	Initialisation	random, K-means++	K-means++
Butina clustering	RMSD threshold	0.5 - 5 Å	1 Å
	reordering	True, False	False

2023) requires sequences of both chains as an input. After this step, light chains were ignored. Antibodies were grouped based on the length of the heavy chain CDRs, aligned on heavy chain framework regions, and the C_α RMSD of CDRs H1-3 calculated. Agglomerative clustering with a 'complete' linkage criterion was used as the clustering algorithm of SPACE2-HC.

SPACE2-Paratope

A second variation of the SPACE2 algorithm was implemented that clusters antibodies based on the structural similarity of CDR loops which are predicted to form part of the paratope (SPACE2-Paratope). Structural models were produced with ABodyBuilder2 (Abanades et al. 2023). The Paragraph method (Chinery et al. 2023) with a classifier cut-off of 0.734, as suggested in the original paper, was then used to predict residues that are part of the paratope based on the models. All CDRs containing at least one paratope residue were then labelled as paratope CDRs. Antibodies were divided into groups containing the same set of paratope

CDRs. Antibodies in each group were further grouped based on the length of the paratope CDRs, aligned on heavy chain framework regions, and clustered based on the C_α RMSD of paratope CDRs. Agglomerative clustering with a 'complete' linkage criterion was used as the clustering algorithm of SPACE2-Paratope.

2.4.4 Numbering scheme and region definitions

IMGT numbering (Lefranc et al. 2003) and North CDR definitions (North et al. 2011) are used throughout.

2.4.5 Analysis of structural clusters

Domain/Epitope consistent clusters

Antibody clusters generated for the Cao et al. (2023) training set, the NVA1 set, the NVA2 set and the anti-lysozyme set were classified as 'epitope-consistent' or 'epitope-inconsistent'. 'Epitope-consistent' clusters of the Cao et al. (2023) training, NVA1 and NVA2 sets only contain antibodies that bind to the same epitope group as determined by experimental epitope binning. 'Epitope-consistent' clusters of the lysozyme data set only contain antibodies that bind to the same residue-level epitope determined using crystal structures.

Owing to the lower resolution of epitopes reported in the EV set and CoV-AbDab, clusters of these data sets were classified as 'domain-consistent' and 'domain-inconsistent'. EV set clusters were labelled as 'domain-consistent' if they only contain antibodies that engage the same antigen domain. CoV-AbDab clusters that satisfy the following rules, consistent with previous studies (Robinson et al. 2021), were determined to be 'domain-consistent':

1. clusters that only contain antibodies that bind to the same antigen and domain
2. clusters that contain antibodies binding to the same domain and others that bind to the same antigen without domain-level resolution
3. clusters that only contain antibodies that bind to the same antigen but do not have domain-level resolution of epitope data

4. clusters with internally consistent epitope data, e.g. a cluster of antibodies labelled to bind to the spike (S) protein N-Terminal Domain (NTD) and others labelled as S non-RBD binders, as S NTD is a subdomain of S non-RBD.

Performance metrics

Throughout this study seven metrics are used to analyse functional clustering. Two accuracy metrics, the fraction of epitope-consistent clusters (number of epitope-consistent multiple-occupancy clusters / number of multiple-occupancy clusters) and the fraction of clustered antibodies in epitope-consistent clusters (number of antibodies in epitope-consistent multiple-occupancy clusters / number of antibodies in multiple-occupancy clusters) were used. Two coverage metrics, the number of multiple-occupancy clusters and the number of antibodies in multiple-occupancy clusters were used. In order to examine accuracy and coverage with one measure I calculated the number of antibodies in consistent multiple-occupancy clusters. Two further metrics were used to assess the diversity of antibodies within clusters, the fraction of functionally consistent clusters containing antibodies from more than one clonotype as well as the mean CDRH3 sequence identity within functionally consistent clusters.

Random baseline

Random clustering was performed as a baseline. The distribution of cluster sizes obtained from the evaluated clustering algorithm with specific parameters was recorded. Clusters with an identical size distribution were then sampled randomly from the data set and performance metrics were calculated. Sampling was repeated 100 times and the metrics averaged.

Clonotyping

Clonotyping was performed using an in-house script. Lenient VH-clonotyping and Fv-clonotyping threshold conditions by community standards were used (Lopez-Santibanez-Jacome et al. 2019; Greiff et al. 2015). A VH-clonotype was defined as matching IGHV gene, length matched CDRH3 and >80% CDRH3 sequence

similarity. Fv-clonotypes were defined as matching VH-clonotype, matching of IG[K/L]V genes, length matched CDRL3 and >80% sequence identity of CDRL3.

2.5 Results

In this section, I provide a brief overview of the SPACE2 algorithm which clusters antibodies by the structural similarity of predicted structural models. The algorithm was extensively optimised and then benchmarked on several data sets of high resolution epitope data.

SPACE2 clusters antibodies in four main steps. Initially, structural models are produced using ABodyBuilder2 (Abanades et al. 2023). Models are then separated into groups of antibodies with identical lengths of the six CDRs, followed by the computation of a pairwise distance matrix of CDR C_α RMSDs. In the final step, a clustering algorithm divides the antibodies into structural clusters. Although some loops of different length can adopt similar structures, I decided to restrict structural comparison to antibodies with identical CDR lengths for the SPACE2 method as evidence suggests length independent structural similarities are infrequent (Nowak et al. 2016; Wong et al. 2019). Restricting structural comparison to CDRs of the same length also allows for more rapid computation as RMSDs do not have to be calculated between all pairs of antibodies within the set. Optimisation of the clustering protocol was performed on a training set of 3051 antibodies against the SARS-CoV-2 receptor binding domain (RBD) (Cao et al. 2023).

2.5.1 Evaluating an optimal clustering algorithm

I tested eight widely used clustering algorithms, greedy clustering, affinity propagation (Frey et al. 2007), Butina clustering (Butina 1999), DBSCAN (Schubert et al. 2017), OPTICS-DBSCAN, OPTICS-xi (Ankerst et al. 1999), agglomerative clustering (Murtagh et al. 2012) and K-means (MacQueen 1967), for their ability to correctly group functionally consistent antibodies in the Cao et al. (2023) training set. To assess the methods, I used the number of antibodies in epitope-consistent multiple-occupancy clusters as the target performance metric as it provides a trade-off between clustering accuracy and data set coverage. High accuracy or coverage metrics individually do not necessarily indicate a good epitope profiling method (Fig. 2.1). High accuracy can be achieved by dividing the data set into very small

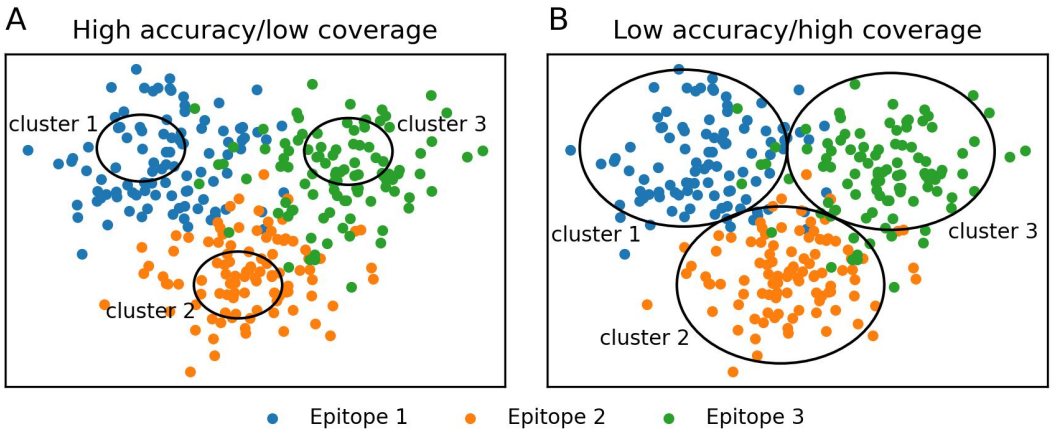


Figure 2.1: Illustration of the evaluation of clustering algorithms. Accuracy and coverage metrics were used to analyse clustering algorithms. Individually, these metrics do not necessarily indicate a good clustering algorithm, instead a trade-off between accuracy and coverage should be monitored. (A) High accuracy is achieved by making small clusters. These are likely to be epitope specific, however, most antibodies are not contained in a cluster. (B) High coverage is achieved by the formation of large clusters. These contain most of the antibodies in the data set but do not tend to be epitope specific.

clusters that are highly likely to be epitope/domain-consistent but do not cover the full diversity of antibodies able to engage a given epitope. Maximal coverage can be achieved by putting all antibodies into a single cluster, which does not provide any useful epitope information.

A parameter scan was carried out to find the optimal setting for each clustering method. Evaluated parameter ranges and optimal values are shown in Table 2.1. As expected, lenient parameters increased data set coverage whereas stringent parameters improve accuracy and the trade-off was maximised at intermediate values. The best performing algorithms, as defined by maximising the number of antibodies in epitope-consistent multiple-occupancy clusters, were agglomerative clustering (optimal parameters: linkage criterion = complete, RMSD distance threshold = 1.25 Å), OPTICS-xi (optimal parameters: xi $\leq$ 0.01, RMSD distance threshold = 2 Å) and K-Means (optimal parameters: initialisation method = K-means++) where K was set to the number of clusters obtained by agglomerative clustering with optimal parameters (Fig. 2.2). As K-means does not lead to an improvement over

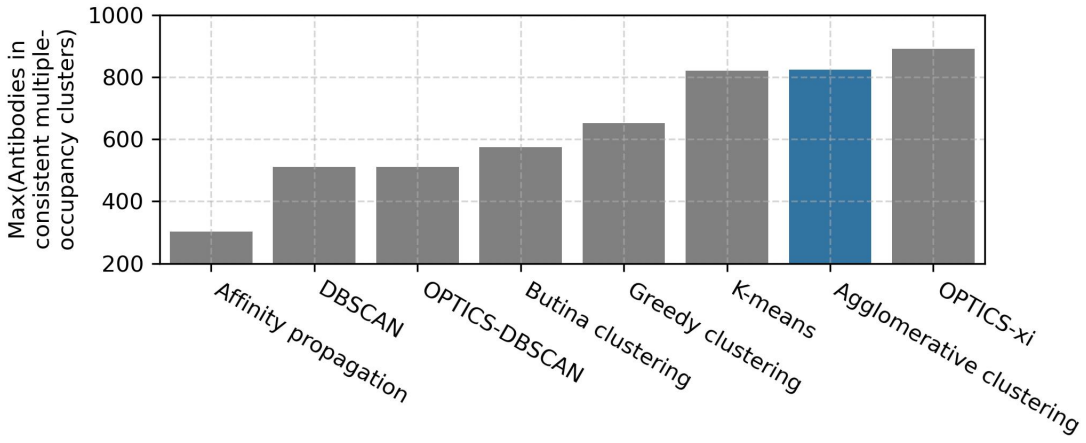


Figure 2.2: Examination of clustering algorithms. Parameter scans of eight clustering algorithms were performed using the Cao et al. (2023) training set. The performance of the clustering was measured in terms of the number of antibodies in epitope-consistent multiple-occupancy clusters (y-axis). The maximum value of this metric achieved by a specific algorithm across all evaluated parameters when clustering the Cao et al. (2023) training set is shown. Evaluated parameter ranges and optimal values are shown in Table 2.1. The agglomerative clustering algorithm selected for SPACE2 is highlighted in blue.

agglomerative clustering, it was disregarded for further analysis. A visualisation of the clustering obtained by the eight algorithms is shown in Figure A.1.

Agglomerative and OPTICS-xi clustering were compared in more detail (Table A.1). Both algorithms achieve a similar clustering accuracy and data set coverage. Agglomerative clustering produces larger clusters than OPTICS-xi, with mean cluster sizes of 3.0 versus 2.7 and maxima of 28 versus 11. When epitope consistent clusters are larger, it suggests they are better capturing the full diversity of the antibodies able to engage a given epitope. Therefore, agglomerative clustering was selected for use in SPACE2.

2.5.2 Examining the behaviour of agglomerative clustering across different structural similarity thresholds

The RMSD threshold parameter of agglomerative clustering determines the leniency of the algorithm as it sets the maximum distance between any two antibodies in a cluster. Small thresholds restrict clustering to highly similar structures, whereas larger values allow clusters to contain more dissimilar antibodies. I evaluated

agglomerative clustering for threshold values between 0.5 and 5 Å to assess how clustering results are affected.

Four metrics were monitored to assess the accuracy of clustering and the data set coverage. The fraction of epitope-consistent clusters (number of epitope-consistent multiple-occupancy clusters / number of multiple-occupancy clusters) and the fraction of clustered antibodies in epitope-consistent clusters (number of antibodies in epitope-consistent multiple-occupancy clusters / number of antibodies in multiple-occupancy clusters) were used as an accuracy measure. The number of multiple-occupancy clusters and the number of antibodies in multiple-occupancy clusters provide information on data set coverage.

Clustering accuracy and data coverage show a strong dependence on the RMSD threshold (Fig. 2.3). At thresholds ≤ 0.75 Å the clustering is highly accurate. More than 80% of clusters are epitope-consistent and approximately 80% of clustered antibodies are in epitope-consistent clusters. Increasing the threshold leads to a rapid drop in accuracy but improves data set coverage. The number of antibodies in multiple occupancy clusters starts to plateau at around 3 Å. The large changes of accuracy and data coverage as a function of threshold suggest that the threshold should be adjusted depending on the aim of the epitope profiling task. Optimal clustering is achieved at a value of 1.25 Å, as defined by maximising the number of antibodies in epitope-consistent multiple-occupancy clusters. However, the threshold can be set to any value between 0.75 and 3 Å to increase accuracy or coverage.

In all the analysis to this point, I reported only on clusters that are 100% epitope-consistent (i.e. only contain antibodies against the same epitope). To measure the inconsistency of the remaining clusters, I analysed the fraction of clusters in which at least 70% of the antibodies engage the same epitope. An additional 12% of clusters are >70% epitope-consistent and these clusters contain an extra 26% of all antibodies contained in multiple-occupancy clusters (Fig. A.2). This result indicates that even those clusters the standard performance metrics are marking as incorrect may contain large amounts of useful information.

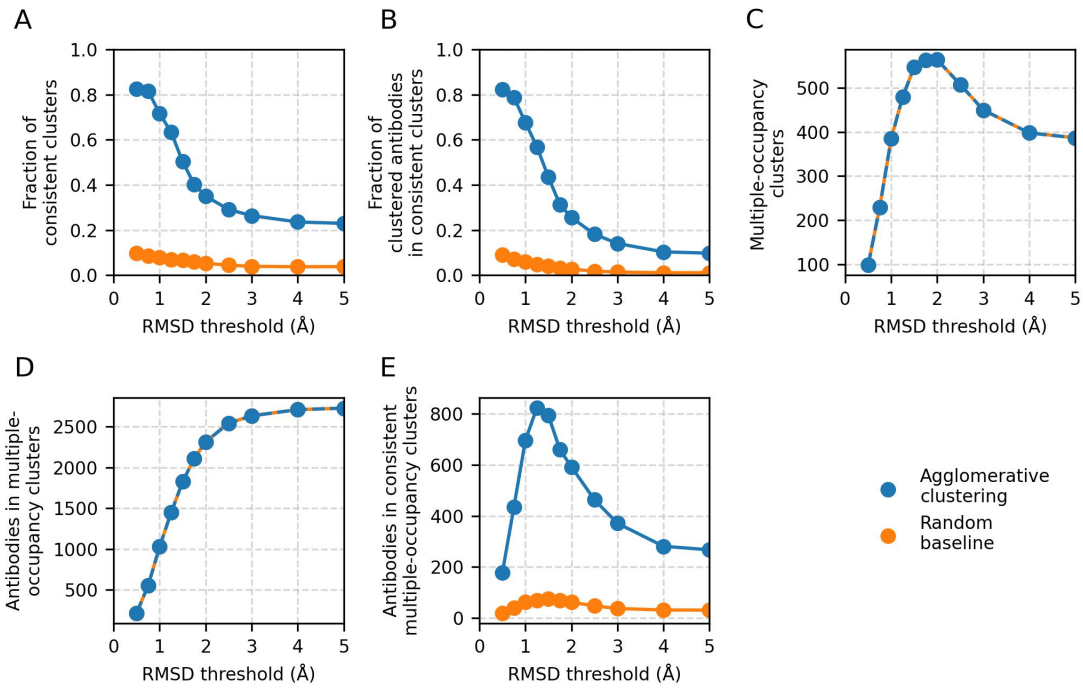


Figure 2.3: Results of agglomerative clustering as a function of RMSD threshold on the Cao et al. (2023) training set. Agglomerative clustering with a 'complete' linkage criterion was performed for threshold values between 0.5 Å and 5 Å. The values of the five performance metrics are plotted against evaluated threshold values: (A) fraction of epitope-consistent clusters, (B) fraction of clustered antibodies in epitope-consistent clusters, (C) number of multiple-occupancy clusters, (D) number of antibodies in multiple-occupancy clusters, (E) number of antibodies in epitope-consistent multiple-occupancy clusters. Results of a random clustering baseline (see methods) are shown for comparison. Values for the number of multiple-occupancy clusters and antibodies in multiple-occupancy clusters for the random baseline are matched to agglomerative clustering.

2.5.3 Evaluating the optimal region for clustering

The SPACE2 method calculates structural similarity of antibodies across all six CDRs. However, not all CDRs are equally involved in binding and I expected the structure of some CDRs to be more important in determining epitope specificity than the structure of others. Therefore, I investigated how the choice of CDRs over which RMSDs are calculated impacts clustering. I assessed two variations of SPACE2 that cluster based on subsets of CDRs.

In the first variation, the algorithm was adapted to consider structural similarity of heavy chain CDRs only (SPACE2-HC). This approach was motivated by sequence-

based methods such as clonotyping which often achieve good performance considering only the heavy chain sequence. In SPACE2-HC antibodies were grouped based on the length of the heavy chain CDRs, aligned on heavy chain framework regions, C_α RMSD of CDRs H1-3 calculated and clustered with an agglomerative clustering algorithm with a 'complete' linkage criterion. An RMSD threshold of 1.25 Å was found to optimise SPACE2-HC (Fig. A.3). SPACE2-HC performed worse than the standard SPACE2 algorithm as measured by a 33% drop in the trade-off metric of antibodies in epitope-consistent clusters (Table A.2). While SPACE2-HC slightly increased data set coverage, a substantial decrease in accuracy was observed.

A second variation of SPACE2 was implemented to cluster antibodies based only on the similarity of CDR loops that contain paratope residues (SPACE2-Paratope). Paratope residues were predicted using the Paragraph method (Chinery et al. 2023). Models were grouped by the combination of CDR loops that contain paratope residues (paratope CDRs). Models were then grouped again based on the length of paratope CDRs, aligned on framework regions and the C_α RMSD of paratope CDRs was calculated. An RMSD threshold of 1.5 Å was found to optimise agglomerative clustering for SPACE2-Paratope (Fig. A.3). Measured by the trade-off metric, SPACE2-Paratope performed worse than standard SPACE2 (Table A.2). A slight drop in both clustering accuracy and data coverage was observed.

The best clustering results were achieved by clustering based on the structural similarity of all six CDR loops. Therefore, the standard SPACE2 methods was chosen as the clustering protocol for further analysis.

2.5.4 SPACE2 performs well on sets of antibodies against diverse targets

SPACE2 was tested on five data sets of antigen specific antibodies using the clustering algorithm (agglomerative clustering) and parameter choices (complete linkage criterion, 1.25 Å RMSD threshold) defined on the Cao et al. (2023) training set. The test sets comprised a data set of anti-lysozyme antibodies, a non-public data set of anti-Ebola virus antibodies (EV), two non-public data sets of antibodies

Table 2.2: Performance of SPACE2 on test data sets. Values of the five performance metrics as well as the fraction of antibodies successfully modelled with ABodyBuilder2 are shown for each data set. For the two metrics of number of antibodies in multiple-occupancy clusters and number of antibodies in epitope-consistent multiple-occupancy clusters additionally a percentage is shown, indicating the percentage of antibodies in the data set. CoV-AbDab (test) denotes the subset of CoV-AbDab that is not contained in the Cao et al. (2023) training set (see methods). CoV-AbDab (test) was used for this analysis to prevent testing on training set antibodies.

Data set	**anti-lysozyme**	**CoV-AbDab (test)**	**EV**	**NVA1**	**NVA2**
mAbs in set	53	7685	126	31	33
Fac. of mAbs modelled	1.0	1.0	0.87	1.0	1.0
Frac. of consistent clust.	1.0	0.85	0.78	0.83	1.0
Frac. of clustered mAbs in consistent clust.	1.0	0.80	0.74	0.86	1.0
Multiple-occupancy clust.	5	1267	9	6	5
mAbs in multiple-occupancy clust.	50 (94%)	4188 (54%)	19 (15%)	14 (45%)	16 (48%)
mAbs in consistent multiple-occupancy clust.	50 (94%)	3353 (44%)	14 (11%)	12 (39%)	16 (48%)

against non-viral targets (NVA1 & NVA2) and CoV-AbDab (test), a version of CoV-AbDab with training set overlap removed (see methods) (Raybould et al. 2021a). An overview of results from clustering the test sets is shown in Table 2.2.

The anti-lysozyme data set contains antibodies against five distinct epitopes. SPACE2 clusters antibodies in this set with high accuracy, shown by 100% epitope-consistency of clusters. Good data coverage is observed and 50 of the 53 antibodies fall into multiple-occupancy clusters. SPACE2 divides the data set into eight clusters (Fig. 2.4). I observed three cases where antibodies binding a common epitope are separated into two clusters. Looking at these cases in more detail shows that despite engaging the same epitope the antibody structures do not overlay perfectly. In each case, I observed antibodies that bind the epitope in two different binding poses and these are separated into distinct clusters by SPACE2. These results show that SPACE2 groups antibodies with a high resolution.

SPACE2 also achieves a high clustering accuracy on CoV-AbDab (test), the EV set, the NVA1 set and the NVA2 set. 85%, 78%, 83% and 100% of clusters in the four sets are domain/epitope-consistent, respectively. Domain/epitope-consistent clusters comprise 80%, 74%, 86% and 100% of all antibodies grouped

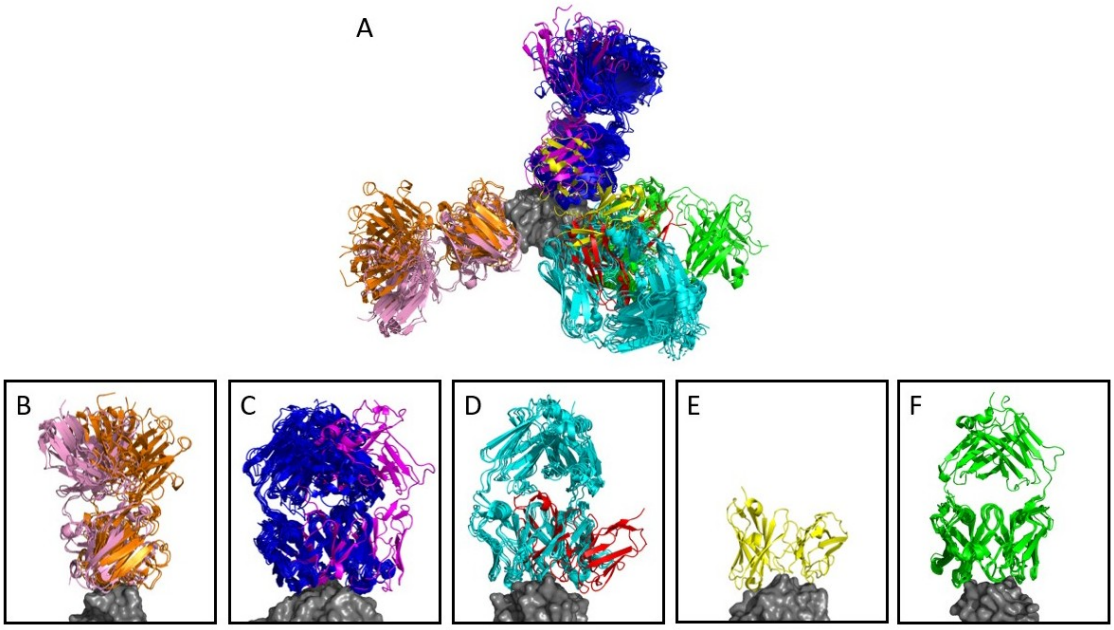


Figure 2.4: Anti-lysozyme antibodies. Crystal structures of 53 antibody-lysozyme complexes are shown aligned on the antigen structure (grey). Antibodies are coloured according to the clusters assigned by SPACE2. (A) Overlay showing all 53 antibody-lysozyme complexes. (B-F) Each panel shows all antibodies that bind to one of the five lysozyme epitopes as defined by Ab-ligity (Wong et al. 2021). Panels B, C and D each contain two sets of antibodies that do not overlay perfectly indicating a difference in binding pose. SPACE2 separates antibodies binding the same epitope in a different binding pose into distinct clusters as indicated by the colouring.

into multiple-occupancy clusters.

Data coverage differs for the four data sets (see methods for definition of coverage metrics). Coverage of CoV-AbDab (test), NVA1 and NVA2 is high with 54%, 45% and 48% of all antibodies contained within multiple-occupancy clusters. In comparison, only 19 of 126 EV set antibodies are grouped into multiple-occupancy clusters. The EV set is relatively small and contains antibodies against the Ebola virus glycoprotein, a large multi-domain protein with many potential epitopes (Lee et al. 2008). It is not expected to observe many antibodies engaging the same residue-level epitope in a small data set of antibodies against a target with many epitopes, which likely explains the low coverage.

Antibodies within the same epitope group in CoV-AbDab, the EV set, the NVA1 set and the NVA2 tend to be split across a large number of SPACE2 clusters.

This is explained by the low resolution of epitope labels in these data sets and antibodies annotated with the same epitope label likely bind to a large number of different residue level epitopes.

Overall, SPACE2 generalises well to the test sets. The algorithm achieves a high clustering accuracy on all five data sets and a good coverage on the CoV-AbDab (test), NVA1, NVA2 and the anti-lysozyme database. Coverage of the EV set is comparably low, indicating a challenge in clustering smaller data sets of epitope diverse antibodies.

2.5.5 Advances in structure prediction improve structure-based computational epitope profiling

I compared the performance of SPACE2 to SPACE1, the previous version of the algorithm. SPACE1 (Robinson et al. 2021) groups antibodies based on structural similarity of homology models produced with ABodyBuilder (Leem et al. 2016) followed by greedy structural clustering at an RMSD threshold of 0.75 Å.

The number of antibodies in epitope/domain-consistent multiple-occupancy clusters which balances both accuracy and coverage was again used as the metric for comparing performance. SPACE2 outperforms SPACE1 using its suggested threshold (RMSD threshold 0.75 Å) (Table A.3). As SPACE2 uses an RMSD threshold of 1.25 Å, I also explored a range of RMSD values to see whether the difference in RMSD threshold is the driver for the difference in performance. I found that a threshold of 1.25 Å improved SPACE1 clustering (Table A.3) but it was still significantly worse than SPACE2. SPACE1 with a 1.25 Å threshold results in an 18% and 9% decrease in antibodies in epitope/domain-consistent multiple-occupancy clusters on the two largest data sets, CoV-AbDab and the Cao et al. (2023) training set, respectively (Table 2.3). SPACE2's better performance is driven by better coverage while achieving a similar accuracy.

Modifications of the SPACE2 and SPACE1 methods reveal that the better performance of SPACE2 arises due to the larger number of antibodies modelled with ML-based structure prediction as compared to homology modelling and better

Table 2.3: Comparison of SPACE2, SPACE1 and clonotyping. The original SPACE1 algorithm was evaluated at an RMSD threshold of 1.25 Å. Two protocols were used for clonotyping (see methods). VH-clonotyping is restricted to genes and sequence of the heavy chain. Fv-clonotyping considers both heavy and light chains. For the two metrics of number of antibodies in multiple-occupancy clusters and number of antibodies in epitope-consistent multiple occupancy clusters additionally a percentage is shown, indicating the percentage of antibodies in the data set. The most important performance metric to consider when comparing different epitope profiling methods is the number of antibodies in epitope/domain-consistent multiple-occupancy clusters as high accuracy or coverage metrics individually may not indicate good performance. The fraction of epitope/domain-consistent clusters containing more than one VH-clonotype and the mean CDRH3 sequence identity observed within epitope/domain-consistent clusters are also given. The best result for each metric is highlighted in bold.

Data set	CoV-AbDab				Training set			
Method	SPACE2	SPACE1	Clonotyping		SPACE2	SPACE1	Clonotyping	
			VH	Fv			VH	Fv
Fract. of mAbs modelled	**1.0**	0.71	-	-	**1.0**	0.7	-	-
Fract. of consistent clust.	0.87	0.87	0.98	**0.99**	0.63	0.64	0.84	**0.83**
Fract. of clustered antibodies in consistent clust.	0.82	0.81	0.97	**0.99**	0.57	0.58	0.75	**0.79**
Multiple-occupancy clust.	**1811**	1165	1191	970	**480**	314	361	303
mAbs in multiple-occupancy clust.	**6271 (59%)**	4010 (37%)	4045 (38%)	2754 (26%)	**1446 (47%)**	935 (31%)	1060 (35%)	793 (26%)
mAbs in consistent multiple-occupancy clust.	**5126 (48%)**	3255 (30%)	3916 (37%)	2733 (25%)	**823 (27%)**	538 (18%)	797 (26%)	628 (21%)
Fract. of clusters containing >1 VH-clonotypes	**0.81**	0.71	0	0	0.55	**0.58**	0	0
Mean CDRH3 sequence identity	**0.54**	0.57	0.88	0.88	**0.66**	0.67	0.86	0.87

clustering with the agglomerative clustering protocol compared to greedy clustering. The higher quality of models obtained from ML-based structure prediction does not lead to clear improvements in clustering (Fig. 2.5 and Table A.4).

2.5.6 SPACE2 improves coverage compared to clonotyping

Clonotyping is the most commonly used epitope profiling method. It clusters antibodies based on sequence similarity. As clonotyping assumes that antibodies against a given epitope must originate from progenitor B-cells with shared genetic

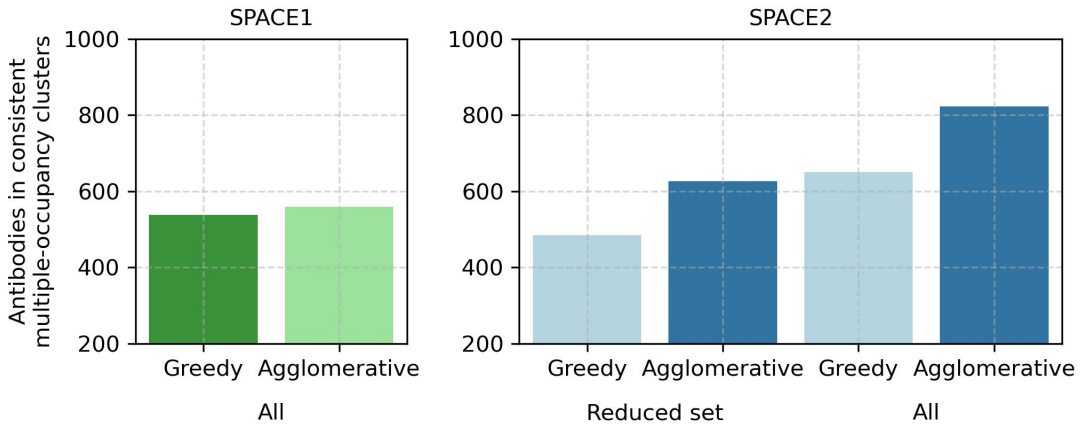


Figure 2.5: In-depth comparison of SPACE2 and SPACE1 performance on the Cao et al. (2023) training set. SPACE1 with a 1.25 Å RMSD threshold (green) and SPACE2 (blue) as well as an adaptation of SPACE1 (light green) using the default agglomerative clustering algorithm of SPACE2 (complete linkage criterion, 1.25 Å RMSD threshold) and an adaptation of SPACE2 (light blue) using the default greedy clustering algorithm of SPACE1 (1.25 Å RMSD threshold) were evaluated on the complete data set (all). Additionally, SPACE2 and its adaptation were evaluated on a reduced data set which only included the 2140 antibodies successfully modelled by homology modelling (reduced set).

origins it cannot detect functional convergence. Thus, the method is limited when clustering data sets of antibodies from different individuals or species. I compared SPACE2 to two lenient clonotyping protocols, VH and Fv-clonotyping (see methods), on the two largest data sets which contain antibodies from diverse sources. The Cao et al. (2023) training set consists of antibodies isolated from 165 human patients (Cao et al. 2023) and CoV-AbDab contains antibodies from a range of studies (∼450) and several species (Raybould et al. 2021a).

The performance of SPACE2 and the two clonotyping protocols are shown in Table 2.3. SPACE2 outperforms both VH and Fv-clonotyping in the key metric of antibodies in epitope/domain-consistent clusters on both data sets. Improvement in this metric is driven by increased data set coverage by SPACE2. I observed 33% and 21% more antibodies in multiple-occupancy clusters for SPACE2 than for Fv-clonotyping of CoV-AbDab and the Cao et al. (2023) training set, respectively. Data coverage by VH-clonotyping is better but still substantially lower than SPACE2. On the other hand, SPACE2 is less accurate than both clonotyping protocols. However,

the increase in coverage achieved by SPACE2 exceeds the drop in accuracy.

2.5.7 SPACE2 identifies functional convergence signals

I next analysed the diversity of antibodies within the SPACE2 clusters of the Cao et al. (2023) training set and CoV-AbDab, to see if functionally similar antibodies with very distinct sequences are being identified.

The majority of SPACE2 clusters contain antibodies belonging to several clonotypes, highlighting the ability to link antibodies from different genetic lineages. Specifically, 55% of epitope-consistent clusters from the Cao et al. (2023) training set and 81% of domain-consistent clusters from CoV-AbDab contain antibodies from more than one VH-clonotype (Table 2.3).

Moreover, I investigated the sequence similarity of antibodies within epitope / domain-consistent clusters. Clonotyping is limited to linking sequence similar antibodies as the method uses a CDRH3 sequence identity cut-off to cluster antibodies. Here, I used a lenient cut-off of 80%. I observed a mean CDRH3 sequence identity of 86% within epitope-consistent VH-clonotypes of the Cao et al. (2023) training set and 88% for domain-consistent VH-clonotypes of CoV-AbDab. In comparison, SPACE2 clusters tend to be highly diverse in sequence. Epitope/domain-consistent clusters have a mean CDRH3 sequence identity of 54% and 66% for CoV-AbDab and the Cao et al. (2023) training set, respectively (Table 2.3). A large number of CoV-AbDab clusters were observed with mean sequence identity below 40% (Fig. A.4) and some clusters even contain pairs of antibodies with no common CDRH3 residues.

Structural clustering is also able to functionally link antibodies from different organisms. For CoV-AbDab, SPACE2 produced 26 functionally consistent clusters containing antibodies from more than one species and was able to group antibodies from human, mouse and rhesus macaque origins. In comparison, optimised SPACE1 was only able to detect 18 domain-consistent inter-species clusters and clonotyping is unable to link antibodies from different species.

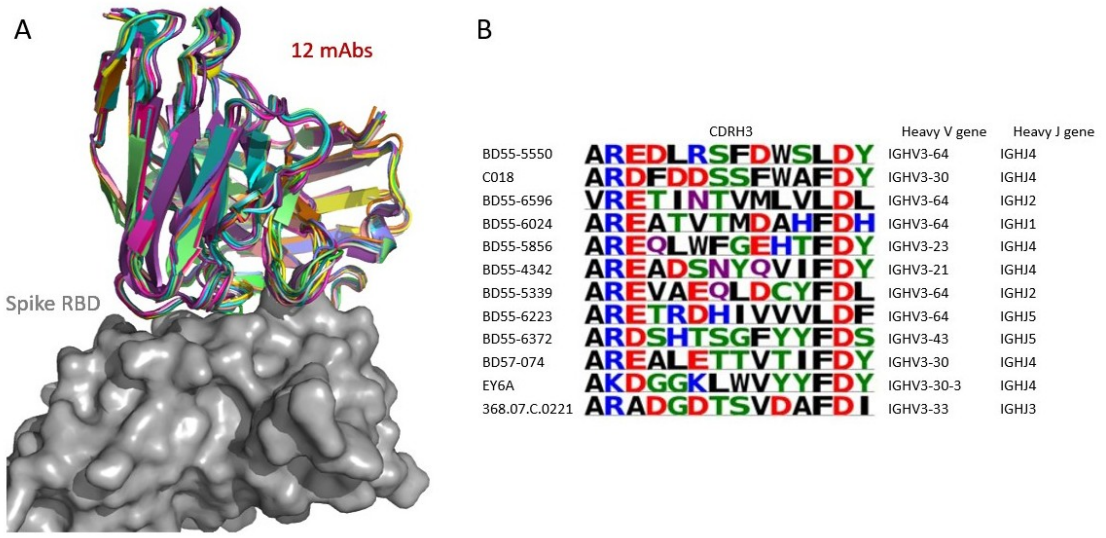


Figure 2.6: SPACE2 identifies functional convergence of antibodies. Twelve-membered CoV-AbDab cluster (368.07.C.0221, BD55-4342, BD55-5339, BD55-5550, BD55-5856, BD55-6024, BD55-6223, BD55-6372, BD55-6596, BD57-074, C018, EY6A) with a mean CDRH3 sequence identity of 33%. The crystal structure of EY6A in complex with its antigen is available (PDB 6ZCZ), SPACE2 suggests that the 11 remaining antibodies bind to the same residue-level epitope. (A) Structural models of the 12 members are overlaid with the crystal structure of EY6A (not shown) in complex with the spike protein RBD (grey). (B) CDRH3 sequence alignment of all 12 cluster members coloured by chemical properties of amino acid residues (produced with Logomaker (Tareen et al. 2020)) and heavy chain V and J genes.

SPACE2 informs on functional convergence of sequence dissimilar antibodies

I examined a SPACE2 cluster of the CoV-AbDab data set with 12 members (368.07.C.0221, BD55-4342, BD55-5339, BD55-5550, BD55-5856, BD55-6024, BD55-6223, BD55-6372, BD55-6596, BD57-074, C018, EY6A) (Fig. 2.6). Eleven of the antibodies engage the spike protein RBD, the final member is annotated as a spike protein binder with unknown domain. Clustering by SPACE2 suggests that these antibodies, determined to bind to the same domain of the spike protein, engage the same residue-level epitope.

The 12 antibodies are highly diverse in sequence and genetic lineage. The cluster shows a mean CDRH3 sequence identity of 33%. The antibodies possess a CDRH3 of length 12 and 8 of these residues differ on average. The most distant pair of antibodies in the cluster are BD55-6596 and EY6A which differ in 11 of

12 CDRH3 residues. The 12 antibodies originate from 7 different IgGH genes and fall into 12 separate VH-clonotypes.

The improvement of SPACE2 over SPACE1 can be seen when examining how these antibodies were clustered by SPACE1. Using SPACE1 with an optimised threshold only six of these antibodies (BD55-6024, BD55-6223, BD55-6596, BD57-074, C018, EY6A) were grouped together and even these six were part of a larger functionally inconsistent cluster with 44 members. BD55-5339 was in a separate functionally inconsistent SPACE1 cluster with 4 members and the remaining five antibodies were not placed in multiple-occupancy clusters.

SPACE2 identifies epitopes targeted by multiple species

As the CoV-AbDab database contains antibodies from multiple species (human, mouse, rhesus macaque), I examined whether SPACE2 can identify epitopes targeted by multiple species. There were 26 SPACE2 clusters of the CoV-AbDab database that contained antibodies from more than one species. I examined a SPACE2 cluster with seven members (368.02a.C.0049, B13, BD55-6574, BD57-092, DK15, Fab-160, SW186) (Fig. 2.7). The cluster contains six antibodies that engage the spike protein RBD and one spike specific antibody without domain-level epitope data. Five of the antibodies are human and originate from human patients, phage-display and transgenic mice. The remaining are murine and were found in immunised mice. SPACE2 suggests that these human and mouse antibodies engage the same residue-level epitope which highlights its ability to detect public epitopes targeted by multiple species.

SPACE1 with an optimised threshold was only able to link one of the two murine antibodies in this cluster to human structures, while clonotyping is unable to link mouse and human antibodies due to different gene usage.

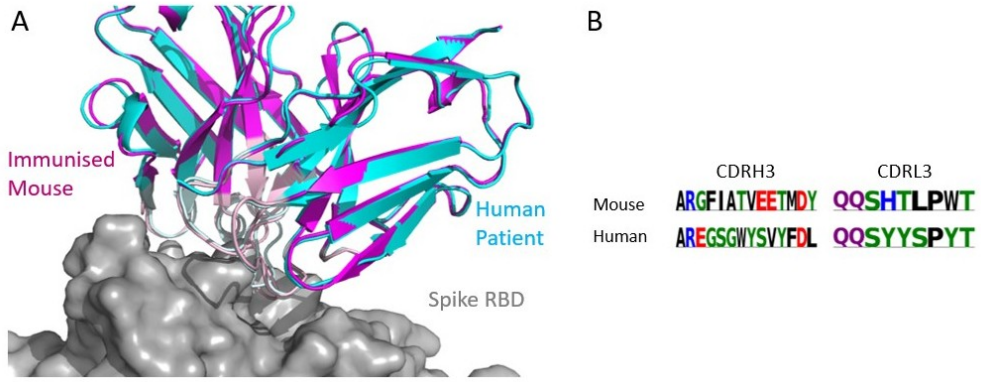


Figure 2.7: SPACE2 identifies epitopes targeted by multiple species. A SPACE2 cluster from CoV-AbDab containing antibodies from multiple species. Two representatives from a cluster comprising murine (SW186) and human (BD57-092) antibodies are shown. The crystal structure of SW186 is available (PDB 8DT3), SPACE2 suggests that BD57-092 binds to the same residue-level epitope. (A) The structural models of SW186 (pink) and BD57-092 (cyan) were overlaid with the crystal structure of SW186 (not shown) in complex with the spike protein RBD (grey). CDR regions of both antibodies are highlighted by the lighter colouring. (B) CDRH3 sequence alignment of the two antibodies coloured by chemical properties of amino acid residues (produced with Logomaker (Tareen et al. 2020)).

2.6 Discussion

Accurately identifying the epitope of antibodies is a key step in understanding immunology and biologic drug design. Such data are conventionally determined experimentally either by solving individual antibody-antigen crystal structures or by epitope binning methods such as competition binding assays. Prior computational clustering of antibodies into functional groups could reduce the number of experiments that need to be carried out or even remove the need for them entirely. Clonal clustering is most commonly used for this purpose, where antibodies are grouped by sequence identity and genetic lineage. But these types of methods will miss antibodies with low sequence identity that have functionally converged and target common epitopes (Robinson et al. 2021; Wong et al. 2021; Scheid et al. 2011; Rijal et al. 2019; Joyce et al. 2016; Shrock et al. 2023).

In a previous study, the SPACE1 method was introduced (Robinson et al. 2021), which clusters antibodies by structural similarity of their homology models. This

method showed that structure-based epitope profiling is better able to detect the full breadth of functional convergence. However, SPACE1 was limited by the coverage of template-based modelling and its inaccuracies. The method was also only benchmarked at the level of domain-consistency of antibodies against one virus class. In this chapter, I introduced SPACE2, an updated method which uses machine learning-based antibody structure prediction (Abanades et al. 2023) and a novel clustering protocol systematically optimised on epitope-resolution data.

I showed across six data sets that SPACE2 can accurately bin antibodies that engage the same epitope and achieve high data coverage. Available crystal structures of antibody-antigen complexes reveal that SPACE2 tends to group antibodies that bind to the same residue-level epitope in an identical binding pose. Epitope resolution of SPACE2 appears to be similar to that obtained from crystal structures and higher than data from epitope binning methods which struggle to distinguish between antibodies that bind distinct sites but overlap sterically.

SPACE2 outperforms the original SPACE1 tool (Robinson et al. 2021) and clonotyping when considering the number of antibodies in epitope-consistent multiple-occupancy clusters. Clonotyping is more accurate than SPACE2 but has far lower coverage. The lower accuracy of SPACE2 is explained by the fact that antibodies with similar CDR structures may engage different epitopes if chemical properties of the CDR residues are significantly different.

I also highlighted how SPACE2 allows the detection of functional convergence across populations of antibodies. Across functionally consistent clusters of the largest data set, CoV-AbDab (Raybould et al. 2021a), I detected a mean CDRH3 sequence identity as low as 54%. Furthermore, I observed 26 functional clusters containing antibodies from multiple species including human, mouse and rhesus macaque antibodies. In comparison, sequence-based epitope profiling such as clonotyping is severely restricted in grouping sequence diverse antibodies and is not able to link antibodies from different genetic origins and species (Lopez-Santibanez-Jacome et al. 2019; Greiff et al. 2015; Raybould et al. 2021b). While it is possible to cluster nanobodies with the SPACE2-HC implementation, when testing on CoV-AbDab

I was unable to detect functional convergence to antibodies. No clusters were detected containing both antibodies and nanobodies suggesting the two formats use different binding site structures to engage common epitopes.

SPACE2 clusters antibodies based on the length and structural similarity of all six CDRs. This approach may constrain the detection of functional convergence to some extent as it assumes antibodies require the same length of all six CDRs to engage the same epitope. I tried to address this issue by evaluating two adaptations of SPACE2 that reduce the number of CDRs required to have the same length. An implementation clustering antibodies based on heavy chain structural similarity (SPACE2-HC) caused a drastic decrease in clustering accuracy. This indicates that light chain structures are important for determining antibody binding specificity which is in line with previous findings on the functional selection of light chains (Jaffe et al. 2022) and their structural importance (Guloglu et al. 2023). Similarly, combining SPACE2 with information from paratope prediction (SPACE2-Paratope) (Chinery et al. 2023) and computing structural similarity only across CDRs predicted to contain paratope residues currently led to fewer functionally consistent clusters. Furthermore, some loops of different length can adopt similar structures (Nowak et al. 2016; Wong et al. 2019). Although evidence suggests that this is infrequent, future work could focus on being able to detect functional convergence across different length CDRs.

The ability to detect functional convergence of antibodies will provide valuable insights into the humoral immune response. SPACE2 is able to provide more complete information on public epitopes targeted by antibodies originating from different individuals and species. Whereas previous studies show several public epitopes are largely distinct between species (Shrock et al. 2023), I identified a number of inter-species clusters. Structural clustering of larger data sets of antibodies isolated from various species will further improve the understanding of differences in their immune responses.

While SPACE2 is computationally more expensive than sequence-based epitope profiling it is tractable for data sets of 10^4 antibodies, a typical number of sequences

obtained from methods such as 10X sequencing (Fig. A.5). The rate limiting step of SPACE2 is currently the prediction of antibody structures. Improvements in the speed of structure prediction tools as well as the release of antibody databases containing pre-modelled structures (Abanades et al. 2023) will contribute to reducing the computational cost of structure-based epitope profiling.

SPACE2 efficiently detects functional convergence of antibodies with highly diverse sequences, genetic lineage and species origins, illustrating that structural compared to sequence data contains additional information on antibody function. However, antibodies, especially the CDRs, can be dynamic and conformational flexibility has been shown to affect important functional properties such as affinity (Mikolajek et al. 2022) and specificity (James et al. 2003). Representing antibodies as structural ensembles rather than a single, static structure, therefore, promises to further improve the analysis of their function. As tools to model the dynamics of antibodies are limited, further method development is needed to effectively utilise flexibility data. In Chapter 3, I describe ITsFlexible, a deep learning model predicting the conformational flexibility of CDR loops.

3

Predicting the conformational flexibility of CDRs

Contents

3.1 Motivation 64
3.2 Contributions 64
3.3 Background 65
3.4 Methods 68
3.4.1 ALL-conformations dataset 68
3.4.2 ITsFlexible 70
3.4.3 Baseline models 72
3.4.4 Alternative flexibility classifiers 73
3.4.5 Molecular dynamics simulations 74
3.4.6 Cryo-EM experimental protocol and data analysis . . . 76
3.5 Results 79
3.5.1 ALL-conformations dataset 79
3.5.2 Predicting the flexibility of CDR-like protein loop motifs 80
3.5.3 ITsFlexible is highly predictive of CDR3 flexibility . . . 82
3.5.4 ITsFlexible matches CDR flexibility in molecular dynamics simulations 85
3.5.5 Cryo-EM experiments confirm predicted flexibility . . . 86
3.6 Discussion 89

3.1 Motivation

Structural flexibility of antibodies and TCR CDRs impacts key functional properties such as affinity and specificity. It remains difficult to study the dynamics of antibodies, and proteins more generally, as experimental data are scarce and computational costs of *in silico* methods are high.

In this chapter, I first present an approach to create a large dataset capturing conformational states of CDRs and related loop motifs by systematic parsing of protein structure databases. Then, I describe the development of a deep learning model, trained on this extensive dataset, that classifies whether CDR loops are rigid (adopt a single conformation) or exhibit flexibility (transition between multiple states).

3.2 Contributions

I created the datasets, performed the model development, benchmarking and analysis and wrote the chapter. The EGNN (Equivariant Graph Neural Network) architecture was adapted from implementations of Hummer et al. (2023) and Satorras et al. (2022). Cryo-EM experiments were carried out by Monica Fernandez-Quintero and colleagues at Scripps. Monica Fernandez-Quintero provided the molecular dynamics simulations and contributed to writing the methods section of the chapter.

This chapter contains material reproduced from:

Fabian C. Spoendlin, Monica L. Fernandez-Quintero, Sai S. R. Raghavan, Hannah L. Turner, Anant Gharpure, Johannes R. Loeffler, Wing K. Wong, et al. “Predicting the Conformational Flexibility of Antibody and T-Cell Receptor CDRs.” Nature Machine Intelligence (2025).

3.3 Background

Structural flexibility of CDRs has been linked to key functional properties of antibodies and TCRs, as discussed in detail in Chapter 1. For example it has been associated with specificity (Guthmiller et al. 2020; James et al. 2003) and binding affinity (Mikolajek et al. 2022). Furthermore, for some antibodies, conformational changes are known to be required for antigen recognition through an induced fit mechanism (Liu et al. 2024).

Computational prediction of a single, static structure of a protein from its sequence is now considered a routine task (e.g. Jumper et al. 2021; Lin et al. 2023; Baek et al. 2021). However, predicting flexibility and structures of multiple conformational states remains challenging. One factor which has limited progress in conformation prediction is the availability of sufficient training data. High-resolution experimental data on conformational states are scarce (Kay 2011; Konermann et al. 2011) and simulation approaches such as molecular dynamics are computationally too expensive to run at a large scale (Vander Meersche et al. 2024).

Evidence on the conformational flexibility of proteins can be obtained from solved crystal structures. While they do not directly report on dynamics, several structures of the same protein solved under different conditions can capture distinct conformational states. As multiple structures must be available, the number of proteins for which flexibility can be assessed from crystallographic data is much smaller than the total number of solved structures. Crystal structures have been used to explore the flexibility of specific loop types (Marks et al. 2018) or for case studies of full length proteins (Riccabona et al. 2024; del Alamo et al. 2022; Stein et al. 2022) but there has been little systematic mining of the protein data bank (PDB) for all instances of the same sequence representing alternative conformations (Ellaway et al. 2023).

Despite data challenges, several methods have been developed that attempt to predict structures of protein conformational ensembles. As discussed in Chapter 1, recent approaches have concentrated on the subsampling of MSAs during the

AF2 inference procedure (e.g. del Alamo et al. 2022; Stein et al. 2022; Wayment-Steele et al. 2024; Sala et al. 2023b; Faezov et al. 2023; Sala et al. 2023a; Saldaño et al. 2022) and generative structure prediction models trained on MD data (e.g. Lewis et al. 2025; Zhu et al. 2024; Zheng et al. 2024; Lu et al. 2024; Huguet et al. 2024; Jing et al. 2024a; Janson et al. 2025). Evaluation of conformation prediction tools has been limited by the data, but available evidence suggests that reliable predictions are not yet possible.

In this chapter, I first introduce an approach to create a dataset capturing the conformational flexibility of CDR3s. Specifically, I searched antibodies and TCR structure databases (Dunbar et al. 2014; Leem et al. 2018) for molecules with multiple solved structures. To increase the amount of available data, I also mined the PDB (Berman 2000) for motifs with the same secondary structure pattern as CDR3s, defined as loops bounded by two antiparallel β-strands, across all proteins. The resulting **A**ntibody **L**ike **L**oop conformations (ALL-conformations) dataset contains 1.2 million crystal structures of loops with 100,000 unique sequences, capturing all experimentally observed conformations of loop motifs found between pairs of antiparallel β-strands including both antibody and TCR CDR3s (Figure 3.1, Panel A). I then analysed the structural flexibility in the ALL-conformations set and labelled more than 20,000 unique loop sequences by their ability to undergo conformational changes (Figure 3.1, Panel B). Finally, I built the **I**mmunoglobulins and **T**CR**s** **Flexib**ility c**l**assifi**e**r (ITsFlexible), a method that classifies whether CDR3s are rigid (adopt a single conformation) or exhibit flexibility (transition between multiple states) (Figure 3.1, Panel C). ITsFlexible predicts the flexibility of CDR3s evaluated in ensembles of crystal structures with high accuracy and achieves state-of-the-art performance. The model also generalises effectively to a test set derived from MD simulations. Furthermore, ITsFlexible was used to predict the flexibility of three CDRH3 loops with no solved structures and cryo-EM was used to experimentally determine their conformations. These experiments showed that two of the three model predictions were correct.

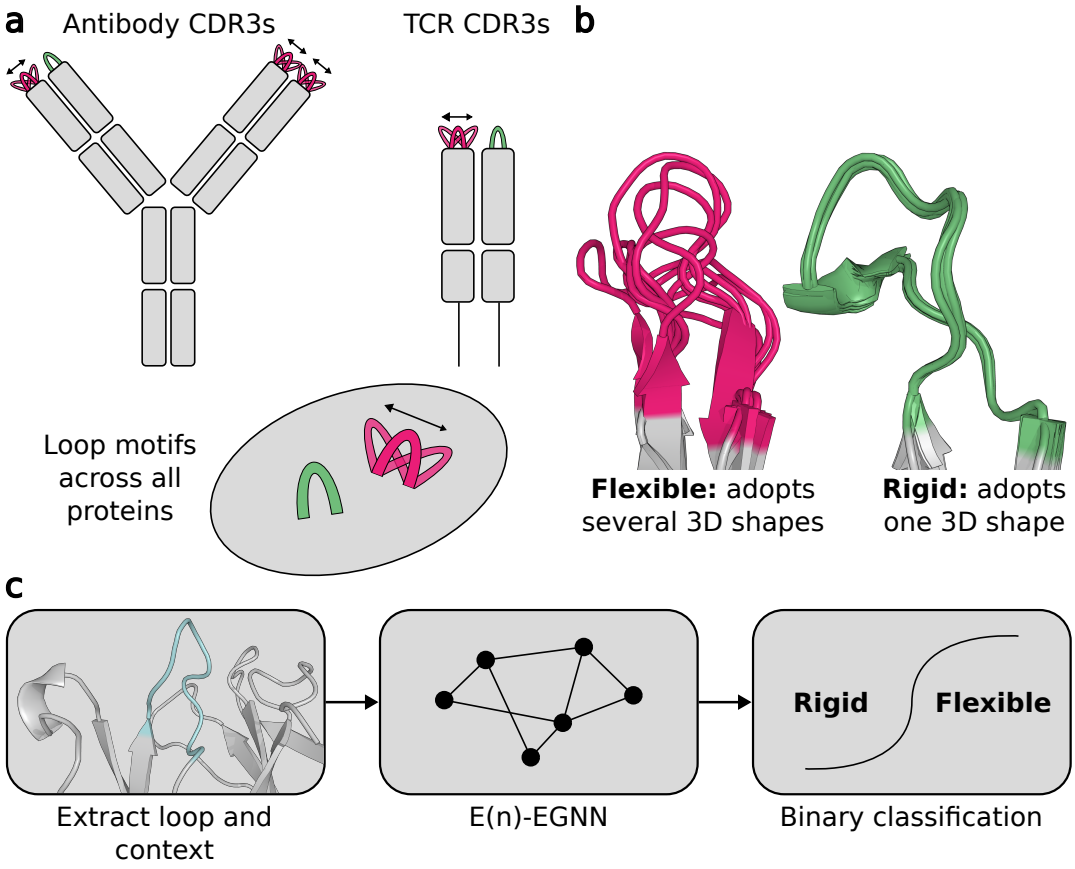


Figure 3.1: Overview of ALL-conformations and ITsFlexible. A) ALL-conformations is a dataset that contains crystal structures of antibody CDR3s, TCR CDR3s and CDR-like loop motifs across all proteins. The dataset captures all observed conformational states of such loops. B) Loops are labelled as either flexible, if they are observed in more than one conformation, or rigid, if evidence suggests they adopt a single conformation. A conformation is defined by structural similarity. An RMSD threshold of 1.25 Å is used to separate states. C) Flowchart detailing the ITsFlexible method predicting the conformational flexibility of CDR loops. The structure and sequence of a loop (cyan) and its context (grey) are extracted from a PDB file and a graph representation is generated. A graph neural network (GNN) classifies loops as conformationally flexible or rigid.

ALL-conformations is released on Zenodo (`doi.org/10.5281/zenodo.15032263`) (Spoendlin et al. 2025b) and ITsFlexible is available on Github (`github.com/oxpig/ITsFlexible`) (Spoendlin et al. 2025a).

3.4 Methods

3.4.1 ALL-conformations dataset

CDR3s of proteins in the immunoglobulin superfamily share a common secondary structure, they are formed by a loop that connects two antiparallel β-strands (Figure B.4). I created **A**ntibody **L**ike **L**oop-conformations (ALL-conformations), a dataset consisting of five subsets that captures all observed conformational states of antibody CDRH3s and CDRL3s, TCR CDRB3s and CDRA3s and loop motifs between antiparallel β-strands across all proteins in the PDB. To generate the datasets, I implemented a systematic approach to search protein structure databases (PDB (Berman 2000), SAbDab (Dunbar et al. 2014), STCRDab (Leem et al. 2018)) for all solved structures of the five protein motifs.

The PDB set

I mined all the protein structures deposited in the PDB before November 22nd 2023 (Berman 2000) for loop motifs sitting between two adjacent antiparallel β-strands. I used the DSSP algorithm (Kabsch et al. 1983) to assign secondary structure to amino acid residues. I then identified all antiparallel β-strands labelled by the algorithm and extracted the regions (loops) in between the two strands. The obtained loop structures were quality filtered for those solved by X-ray crystallography with resolution under 3.5 Å and no unresolved loop residues. A maximum of three residues forming secondary structure elements of β-strand or α-helix within each loop were allowed.

Antibody and TCR CDR3 sets

All antibody Fv structures were extracted from the Structural Antibody Database (SAbDab) (Schneider et al. 2022; Dunbar et al. 2014). Both standard Fvs and single chain Fvs were included. All TCR Fv structures were extracted from the Structural T-Cell Receptor Database (STCRDab) (Leem et al. 2018). Multiple copies of antibodies in the same PDB structure were extracted as CDRs can adopt distinct conformations. Furthermore, some structures contain residues with alternate atom

coordinates. These states were separated and individually added to the dataset. Structures were filtered for those solved by X-ray crystallography with a resolution below 3.5 Å, presence of a complete Fab (both heavy and light chain present in antibody structures; alpha and beta or gamma and delta chains for TCRs) and no unresolved residues in any of the CDRs.

In this chapter, CDR3 loops are defined as the IMGT numbered (Lefranc et al. 2003) residues 107-116 which differs from the standard definitions. This choice was made to ensure that the definition of a CDR3 is more consistent with the way loops are defined by their secondary structure in the PDB set. The standard definition of a CDR3 loop in the IMGT numbering scheme comprises residues 105-117. However, an analysis of all antibody structures in SAbDab (Dunbar et al. 2014) showed that the residues at start and end of IMGT definitions tend to be part of the β-strands on either side of the loop (Figure B.4). Positions 107-116 correspond better with the residues that sit between the two β-strands and we, therefore, used this range to define CDR3s throughout this manuscript.

Conformational flexibility

Structures in the ALL-conformations were grouped into sets of loops considered to be identical. For the PDB set, structures were grouped by sequence identity of the loop and all structures within a group were considered to depict the same loop irrespective of the sequence of the rest of the protein. Antibody and TCR structures were grouped by sequence identity of the entire Fv rather than simply the CDR3s. In this way, CDR3s are only considered to be the same if the entire domain is identical.

For loops with multiple available structures the conformational flexibility was analysed by calculating the number of accessible conformations. A conformation was defined based on structural similarity of a loop, using the RMSD of Cα atoms. Sequence identical loops were clustered using an agglomerative clustering algorithm with complete linkage (Murtagh et al. 2012) and 1.25 Å distance threshold. This enforces that any two structures within a conformation have a maximum RMSD of 1.25 Å. This clustering approach was chosen as it is known to provide a good

functional clustering of CDR loops in antibodies (Spoendlin et al. 2023). The structural similarity of loops can be calculated in multiple ways (see Appendix Section B.2.2 for another choice).

Each loop was assigned a label of 'flexible', 'rigid' or 'unknown'. Loops for which multiple conformations (clusters) were found were assigned a flexible label. Loops for which only a single conformation was observed were not automatically assigned a 'rigid' label. The absence of multiple observed conformations does not prove that a loop cannot adopt multiple conformations. It is possible that alternative states were not captured by the limited number of crystal structures available. The more structures that are available of a loop in which it adopts a single conformation the higher is the confidence for a lack of flexibility. Therefore, a requirement was introduced that a loop needs to adopt a single conformation in at least five separate PDB files to be labelled as rigid. I chose the requirement of five separate PDB files instead of simply five occurrences as the same loop can occur several times in the same PDB file due to multiple copies of a protein within a crystal unit cell. Loops neither labelled as 'flexible' nor 'rigid' were assigned to the 'unknown' group.

3.4.2 ITsFlexible

ITsFlexible is a binary classifier. It takes as input a structural representation of a CDR loop motif and classifies it as flexible, able to adopt multiple conformations, or rigid, occupying a single stable conformation. Model architecture and training procedure are outlined below and details provided in the SI. The model is available on Github (`github.com/oxpig/ITsFlexible`) (Spoendlin et al. 2025a).

Model architecture

The ITsFlexible model is a graph neural network (Figure B.1) and consists of three equivariant graph convolutional layers (EGCLs) (Satorras et al. 2022). The layers take input of node features, coordinates and edge features. Node features are iteratively updated and the last layer of node embeddings are pooled. A linear layer with sigmoid activation function is applied for binary classification. The

chosen model architecture makes predictions invariant to transformations of the group E(3) (translations, rotations, reflections), therefore orientations and absolute positions of the input protein structure can be ignored. Predictions are dependent only on relative residue distances.

Model inputs

A loop and its structural context were encoded as a residue level graph. The context was provided by all residues within 10 Å of any loop residues. For simplicity in the PDB set only residues located on the same protein chain as the loop were selected as context. For antibodies and TCRs residues located on both immunoglobulin chains were included, as both chains influence conformational flexibility (Guloglu et al. 2023). Node features were a 22-dimensional vector consisting of a one-hot encoding of amino acid type (1 class for each of the 20 amino acids plus an additional class for unknown residues) and a one-hot encoding (1 class) of whether the residue is located in the loop or structural context. Non-standard amino acids closely related to a standard amino acid were encoded as such, others were encoded as unknown residues (see Table B.1). The amino acid encoding was concatenated with a one-hot encoding of the residue being located in the loop or the context. Coordinates for each node were taken as the position of the Cα atom of a residue. Nodes were locally connected with edges using a 10 Å distance threshold. Edge features are 9-dimensional providing a one-hot encoding of the presence of a covalent bond between two residues and a Cα distance encoding. The distance encoding was produced by 8 Gaussian radial basis functions (RBFs) equally distributed between 0 Å and 10 Å.

Training

ITsFlexible was trained on the ALL-conformations PDB set using a 70-15-15 training, validation and test split. For length matched loops a maximal sequence identity of 80% was allowed between the splits. Additionally, all loops with more than 80% aligned sequence identity to any loop (not restricted to matching loop length) in the ALL-conformations CDRH/L/A/B3 sets were removed from the training and validation sets. For the training set five structures per loop were sampled randomly

to ensure the stability of predictions to small changes in atom coordinates. For the validation set one structure per loop was sampled at random.

ITsFlexible was trained with a binary cross-entropy loss using the Adam optimiser (Kingma et al. 2014) with a learning rate of $2 \cdot 10^{-4}$ and weight decay of 10^{-6}. During training edges were dropped at random with a probability of 0.2. The validation area under the precision-recall curve (PR AUC) was monitored and training stopped when converged. Ten models were trained and the one with the best validation PR AUC was selected. All models were trained on an Nvidia Quadro RTX 6000 GPU, in approximately 2 GPU hours per model.

Evaluation

ITsFlexible performance was evaluated using structural inputs derived from crystal structures and predicted structural models. When evaluating on crystal structures, an ITsFlexible prediction was made for each available experimental structure containing a loop and took the mean prediction score to calculate performance metrics. To evaluate ITsFlexible with input derived from structural models, antibody and TCR structures were predicted with ImmuneBuilder (IB) and AF2. ImmuneBuilder predictions are made from paired Fv sequences using the ABodyBuilder2 (for antibodies) and TCRBuilder2 (for TCRs) models with default parameters. AF2 predictions of antibodies and TCRs were made using the Colabfold implementation of AlphaFold-Multimer (Mirdita et al. 2022) with default parameters.

3.4.3 Baseline models

A set of three baseline models were created to classify loop flexibility based on simple biophysical input features. The first baseline was formed by a logistic regression classifier fit to inputs of loop length. Longer loops contain more bonds around which they can rotate and are therefore expected to be more flexible in conformation than shorter ones (Guloglu et al. 2023; Marks et al. 2018). I found moderate correlation between loop length and flexibility in the datasets (Table B.4). The second baseline was a logistic regression classifier fit to inputs of the solvent exposure of a loop.

I approximated solvent exposure by the number of residues located within a 10 Å radius around the loop. Loops with higher solvent exposure have less steric hindrance restricting conformational rearrangements and are expected to be more flexible (Guloglu et al. 2023). A final baseline model was formed by a logistic regression classifier fit to input of both length and solvent exposure. All baseline models were fit on the training split of the PDB set.

3.4.4 Alternative flexibility classifiers

Three alternative workflows were implemented to predict CDR flexibility. These were based on protocols designed to model protein conformational ensembles and confidence metrics of protein structure prediction tools.

AlphaFold2 pLDDT

AF2 returns pLDDT scores for predicted structures which can be interpreted as a residue level confidence measure of the prediction. Antibody and TCR structures were predicted using the Colabfold implementation of AlphaFold-Multimer (Mirdita et al. 2022) with default parameters. The pLDDT score of the highest ranked model was extracted. The mean pLDDT of residues located in a CDR was used as input of a logistic regression model to classify flexibility.

AlphaFold2 MSA subsampling

An AF2 MSA subsampling workflow, based on an established protocol (Riccabona et al. 2024; del Alamo et al. 2022), was used to predict structural ensembles. AlphaFold-Multimer (Evans et al. 2021) was run using the Colabfold (Mirdita et al. 2022) implementation to model antibody and TCR structures. The maximum MSA depth was set to 64, extra sequences to 128 and the number of recycles to 1 (default 3 in AF2). Low values of all three parameters increase the diversity of predicted structures and improve sampling of alternative conformations. Setting parameters too low can lead to the generation of unfolded structures (Riccabona et al. 2024). I chose these parameter values as they are within ranges described in previous studies (Riccabona et al. 2024; del Alamo et al. 2022) and they led to maximal diversity

while limiting the occurrence of unfolded structures for the antibodies in the test set. Colabfold was run with 8 seeds resulting in 40 models (5 models are produced per seed) for each protein. The MSA subsampling protocol took approximately 3 minutes on a Nvidia A100 GPU. Owing to the computational cost, MSA subsampling was only performed for the smaller CDR sets and not the full PDB set.

I used the magnitude of structural diversity observed across the 40 predicted structures as a zero-shot classifier of CDR flexibility. Structural diversity was calculated by the Cα RMSD of the CDR loop residues. The mean structural diversity was used as input of a logistic regression model to classify flexibility.

ABB2 RMSPE

The ABB2 predicted error, a residue level confidence metric (Abanades et al. 2023), was used to classify flexibility of antibody CDRs. Antibody structures were predicted using a retrained version of ABB2. There is a large overlap of antibodies in the CDR test set I used for evaluation of flexibility prediction and the training set of ABB2. To avoid data leakage I retrained ABB2 removing all antibodies with 100% CDRH3 or CDRL3 sequence identity to any CDR in the test sets. This reduced the training set to 4469 antibodies compared to the 5669 in the original training set. This version of ABB2 retains good accuracy at antibody structure prediction evaluated on a benchmark (Table B.10). For details on the ABB2 retraining as well as an evaluation of the original ABB2 for flexibility prediction see Appendix Section B.2.4.

Antibody Fv structures were predicted using ABB2 with the retrained weights, for all remaining parameters the ABB2 default values were used. The predicted error of modelled antibodies was extracted. The root mean square predicted error (RMSPE) of CDR residues was used as input of a logistic regression model to classify flexibility.

3.4.5 Molecular dynamics simulations

Molecular dynamics (MD) simulations were performed following the protocol described in Fernández-Quintero et al. (2019c). The 19 investigated antibodies (Table B.2) were prepared at pH 7.4 using the Protonate3D tool in MOE (Chemical

Computing Group ULC 2024; Labute 2009). C-termini were capped with N-methyl groups. Each structure was solvated in a cubic TIP3P water box with a minimum wall distance of 12 Å from the protein (Jorgensen et al. 1983; Gapsys et al. 2020; Fischer et al. 2024), and all systems were parametrized using the AMBER ff19SB force field. A uniform background charge was applied to neutralize the system and enable accurate calculation of long-range electrostatics (Sabri Dashti et al. 2012).

To broaden the exploration of conformational space, enhanced sampling was performed using well-tempered metadynamics (WT-metaD), implemented in GROMACS with PLUMED 2 (Barducci et al. 2008; Domene et al. 2015; Tribello et al. 2014; The PLUMED consortium 2019; Bonomi et al. 2009; Bauer et al. 2022). The collective variables were defined as a linear combination of the sine and cosine of the ψ torsion angles of either the CDRL3 and CDRH3 loops or all CDR loops. WT-metaD simulations were run for 1000 ns per system, with a Gaussian height of 10 kJ/mol, a deposition interval of 5000 steps, and a bias factor of 10. To reduce any bias introduced by enhanced sampling and to obtain trajectories representing unbiased dynamics, a two-step procedure was used. First, the WT-metaD trajectories were aligned on the Cα-atoms of the variable domains and clustered on the CDR loops using average linkage hierarchical clustering (RMSD cut-off 1.2 Å) in cpptraj (Roe et al. 2013). Representative cluster structures were used as starting points for independent unbiased MD simulations (100 ns per replicate), providing an aggregated sampling of >10 µs per system.

Classical MD simulations were carried out in the AMBER 22 simulation package (Case et al. 2023), using the AMBER ff19SB force field and TIP3P solvent. Simulations were performed in an NpT ensemble using pmemd.cuda (Salomon-Ferrer et al. 2013), with temperature maintained at 300 K via Langevin dynamics (collision frequency 2 ps^{-1}) (Adelman et al. 1976; Doll et al. 1975) and pressure maintained with a Monte Carlo barostat (one volume change per 100 steps) (Aqvist et al. 2004). Bonds involving hydrogen atoms were constrained with SHAKE, enabling a 2 fs time step.

To extract representative conformations, hierarchical clustering of the unbiased MD trajectories was performed using average linkage clustering (an RMSD cut-off of 2.5 Å for the CDR loops after alignment on the Fv domains), as implemented in cpptraj.

To determine binary labels of 'flexible' and 'rigid' for the CDRs of the simulated antibodies, I performed analysis identical to the crystal structure datasets (see methods section: ALL-conformations datasets). The representative MD structures were clustered by Cα RMSD of CDR3 loops using an agglomerative clustering algorithm with complete linkage (Murtagh et al. 2012) and a 1.25 Å distance cutoff. The number of conformations observed for the CDR3s of each antibody is shown in Table B.2. CDRH3s and CDRL3s were then assigned labels flexible, if multiple conformations were observed in the simulations, or rigid, in case of a single conformation.

The flexibility of simulated antibodies was predicted by ITsFlexible. A prediction was made for each representative structure of an antibody and the maximum prediction score used. I also tried using the mean prediction score. This however resulted in a narrower distribution of scores across the dataset which is less optimal for classification tasks.

3.4.6 Cryo-EM experimental protocol and data analysis

The Patent and Literature Antibody Database (PLAbDab) (Abanades et al. 2024) was searched for antibodies that target influenza H1N1 hemagglutinin. Antibody Fv structures were predicted with a default run of ABB2 (Abanades et al. 2023). Flexibility was classified with ITsFlexible from inputs of the obtained structural models. Antibodies with a predicted probability score lower than 0.1 were considered as rigid examples and antibodies with a score above 0.5 as flexible examples. Three antibodies (1 rigid, 2 flexible) with good developability properties predicted with TAP (Raybould et al. 2019) were selected for cryo-EM experiments.

Sample preparation

Three HA (HA/California/07/2009) in complex with one Fab each (AEL31302/AEL31311 (antibody 1), AMB38310/AMB38599 (antibody 2) and AMB38442/AMB38568 (antibody 3) were prepared by mixing Fab with a 1:1 molar ratio of Fab:HA protomer and incubated for approximately 4 hours at 4°C. 0.25 µL of 4 mM CHAPSO detergent was added to the complex to aid in particle tumbling. The final concentrations for the complexes of HA with each Fab were 1.14 mg/mL, 1.11 mg/mL and 1.02 mg/mL on the grid respectively. The samples were added to glow discharged 1.2/1.3 Au Ultrafoil 300 mesh grids and subjected to vitrification using a Vitrobot Mark IV system. The settings were as follows: 3 µL of sample, temperature inside the chamber was 4°C, humidity was 90%, blotting force was 1, wait time was 3 s. The sample was blotted off for 4.5 s and the grids were plunge-frozen into liquid-nitrogen-cooled liquid ethane.

Data collection, processing, and model building

Cryo grids of the complexes were imaged at 190,000x nominal magnification using a Falcon 4i camera on a Glacios microscope at 200 kV. Automated image collection was performed using EPU from ThermoFisher. Micrographs for antibody 1 were collected untilted and micrographs for antibody 2 were collected with a 30° tilt. Images were aligned, dose-weighted, and Contrast Transfer Function (CTF)-corrected in the CryoSPARC Live™ software platform, with automated image collection also performed using Smart EPU software (ThermoFisher). Data processing for all three datasets was carried out in CryoSPARC v4.5.3 (Punjani et al. 2017). Blob particle picking was performed on all micrographs with a minimum particle diameter of 100 Å and a maximum of 200 Å. Particles extracted at 480 pixels box size were used to perform 2D classification for antibody 1 and antibody 2, which were then used to generate a 3D reference model from ab-initio refinement, followed by heterogeneous refinement to obtain one good class that was further non-uniform (NU) heterogeneous refined. For antibody 3, particles at 512 pixels were extracted and Fourier cropped to 256 pixels box size to perform 2D classification,

followed by ab-initio, heterogenous and NU refinements. Gold-Standard Fourier Shell Correlation (GSFSC) resolution was calculated to be 3.77 Å for antibody 1 and 3.91 Å for antibody 3. The models were docked into the cryo-EM density map in UCSF ChimeraX (Goddard et al. 2018). The structural model was built iteratively with COOT followed by real-space refinement in the PHENIX package (Afonine et al. 2018). The Kabat numbering system (Kabat et al. 1979) was used for antibodies and the H3 numbering scheme for HA.

For antibody 2 initially, no high-resolution map could be obtained. The antibody reveals a high degree of variability contributing to conformational changes in the head and stem of the HA. The workflow to visualise the conformational changes in the complex is shown in Figure B.2. The 243000 particles and the map resulting from the NU refine job downsized to 128 pixels were used as input for the 3D flex (Punjani et al. 2023) data preparation job. For the 3D Mesh Prep job, a solvent mask was generated by low pass filtering the consensus map from 3D flex data prep job and segmented as shown in Figure B.2 and a base number of 40 tetrahedral cells was used in combination with a minimum rigidity weight of 0.5. The resulting 3D flex mesh was used to run 3D flex train with a number of latent dimensions of 3. 3D Flex generator was used to produce a volume series of 41 frames using the consensus map showing the areas of high flexibility. The movie depicting the conformational flexibility based on the volume series was generated using UCSF ChimeraX (Goddard et al. 2018).

Table 3.1: Number of structures and sequences in ALL-conformations

Dataset	Structures	Unique loop/loop + anchors/Fv sequences	Loop/loop + anchors/Fv seqs with multiple structures
PDB set	1,208,000/-	99,000/162,000/-	69,000/112,000/-
Antibodies			
CDRH3	6401	2190/2190/2477	1372/1372/1460
CDRL3	6401	1762/1776/2366	1147/1150/1432
TCRs			
CDRB3	691	197/197/234	137/137/144
CDRA3	691	192/196/233	134/137/151

3.5 Results

3.5.1 ALL-conformations dataset

Antibody **L**ike **L**oop conformations (ALL-conformations) is a dataset that captures the conformational flexibility of loop motifs bounded by two antiparallel β-strands which includes five subsets, antibody CDRH3s and CDRL3s, TCR CDRB3s and CDRA3s and loop motifs across all proteins in the PDB (Figure B.4). The dataset collects available crystal structures of such loops and contains the structures of all experimentally observed conformational states.

Extracting CDR3 structures from SAbDab (Dunbar et al. 2014) and STCRDab (Leem et al. 2018) and loop motifs from the PDB (Berman 2000) I obtained a total of more than 1.2 million examples with more than 100,000 unique sequences (Table 3.1). The PDB set contains loops between 1 and 87 residues, however, it is heavily enriched for shorter loops (Figure B.5). Length distributions are different for CDR3s with a peak around 10 to 15 residues and no loops shorter than 4 amino acids (Figure B.6). I found that a large number of loop sequences are represented by multiple crystal structures (Figure B.5 & B.6).

The loops were classified as rigid, flexible or unknown (see methods). Loops for which multiple conformations are observed in crystal structures were labelled as 'flexible'. A conformation was defined as a cluster in which the pairwise RMSD

of any member is below a threshold of 1.25 Å. This approach was chosen because it was previously shown to provide a good functional clustering of antibodies (Spoendlin et al. 2023). A conformation should, therefore, reflect a functionally distinct structural state. It is not possible to guarantee that all loops only observed in a single conformation are indeed 'rigid' as there always remains a possibility that additional conformational states have not yet been captured. To address this issue, I only label loops as 'rigid' if they adopt the same conformation in more than five structures. This should ensure that the set is heavily enriched for loops with a single accessible conformation. Using these definitions, I identified more than 16,000 'rigid' and 4,000 'flexible' loops (Table B.5). For an evaluation of the impact of the RMSD threshold on the proportion of the two classes see Figure B.10 and Appendix Section B.2.3.

3.5.2 Predicting the flexibility of CDR-like protein loop motifs

The ALL-conformations set was used to train **I**mmunoglobulins and **T**CR**s** **Flexib**ility c**l**assifi**e**r (ITsFlexible), a model that predicts protein loop flexibility. ITsFlexible is a graph neural network (GNN) (Figure B.1) that performs binary classification of loops according to whether they can occupy multiple conformations (flexible) or adopt a single stable state (rigid) from inputs encoding the sequence and structure of a loop and its structural context (see methods for details).

I trained and evaluated ITsFlexible on the PDB set of ALL-conformations, containing loop motifs observed across any protein. The data split was performed based on sequence identity and test set loops had a maximum of 80% sequence identity with length matched loops in the training and validation sets. Classifier performance was compared to random classification, three baseline models and a zero-shot flexibility prediction workflow based on AF2 pLDDT (Jumper et al. 2021) (Figure 3.2). Random classification shows validation metrics when randomly classifying a dataset with the given proportion of labels. The baselines predict flexibility from biophysical features of length and solvent exposure for which have

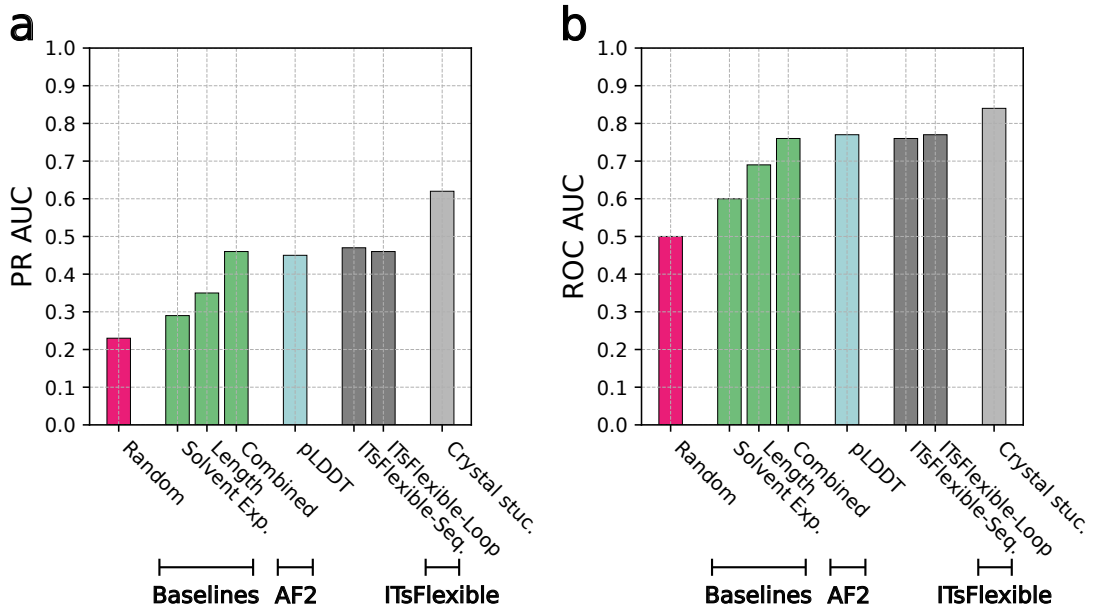


Figure 3.2: ITsFlexible performance evaluated on the PDB test set. Classification on the PDB test set containing 2845 loop motifs is evaluated with metrics of (a) PR AUC and (b) ROC AUC. ITsFlexible performance (light grey) is compared to random classification (red), three biophysical baselines (green), an AF2 pLDDT-based model (blue) and two ITsFlexible versions with input ablations (dark grey).

previously been shown to affect loop dynamics (Guloglu et al. 2023; Marks et al. 2018). Long loops contain more bonds around which they can rotate and solvent exposure reduces steric hindrance which restricts conformational rearrangements. pLDDT is a residue level confidence score and provides a good indicator of disordered protein regions (Jumper et al. 2021). As the test set contains several structures of each loop, an ITsFlexible prediction was made for all of them and metrics were calculated based on the mean prediction score.

ITsFlexible is able to classify loop flexibility (area under the precision-recall curve (PR AUC) of 0.62 and receiver-operating characteristic area under the curve (ROC AUC) of 0.84) and outperforms the baselines and zero-shot classifier. I investigated features predictive of loop flexibility through ablation of model inputs. ITsFlexible-loop is similar to the default model but inputs are reduced to the structure and sequence of only the loop itself (see Appendix Section B.1.1). ITsFlexible-sequence is a CNN-based model trained on a sequence encoding loops (see Appendix Section B.1.1). ITsFlexible-loop and ITsFlexible-sequence achieve a similar performance to

one another. Both outperform the relevant baseline (only baseline-length is relevant as solvent exposure depends on the structural context which both models do not consider) but are less predictive than ITsFlexible (Figure 3.2).

These results give some indication of factors that determine the conformational flexibility of protein loops. It is well known that longer and solvent exposed loops tend to be more flexible than shorter and buried ones (Marks et al. 2018; Guloglu et al. 2023). While findings agree with these general trends, I also showed that loop sequence is more predictive of flexibility than length alone, suggesting the sequence of a loop impacts its ability to adopt multiple conformations. A large boost in performance was achieved by encoding the structural context of the loop motif. In line with previous evidence from MD studies (Guloglu et al. 2023), this highlights that interactions of a loop with its context within the protein are important determinants of its conformational dynamics.

3.5.3 ITsFlexible is highly predictive of CDR3 flexibility

I next investigated the ability of ITsFlexible, trained on general protein loops, to predict the flexibility of antibody and TCR CDR3s. The model was evaluated on the CDR3 sets of ALL-conformations which were designed to have no overlap, defined as more than 80% aligned sequence identity, with the training and validation sets.

I initially evaluated ITsFlexible using crystal structures as input. In addition to the baselines and the pLDDT-based predictor introduced in the previous section, I evaluated two additional zero-shot flexibility predictors. The first workflow predicts flexibility based on the diversity in CDRs modelled by multiple AF2 runs with subsampled MSAs (see methods). The second workflow uses the residue level confidence score (RMSPE) of ABodyBuilder2 (ABB2), an antibody specific structure predictor (Abanades et al. 2023).

ITsFlexible proved to be highly predictive of antibody and TCR CDR3 flexibility (Figure 3.3). The method outperforms the biophysical baselines and zero-shot classifiers in nearly all CDR test sets. The only exception is the CDRA3 set where the MSA subsampling approach surpasses ITsFlexible by a narrow margin. However,

ITsFlexible substantially outperforms all other methods on CDRH3s, the largest and therefore most representative set, and is the only method that consistently achieves high predictive accuracy across all four test sets. I also point out that ITsFlexible was evaluated with the most stringent train-test split compared to the zero-shot models. Specifically, CDRs in the ITsFlexible training set were filtered by 80% sequence identity to the test sets, whereas 100% sequence identity was used for ABB2 and the AF2 training set may even contain test set overlaps.

I conducted a more detailed investigation of the ABB2 confidence score. This analysis revealed that RMSPE is potentially more indicative of the number of times a certain antibody occurs in the training data rather than flexibility (see Appendix Section B.2.4).

Comparing classification performance between the four CDR sets (using ROC AUC), I observed the best results for L3s and B3s with slightly lower values for H3s and A3s. Antibody H3s have higher sequence diversity than L3s which is expected to make classification more difficult. The lower performance for A3s could be linked to findings of a different pattern of genetic diversity compared to B3s (Quast et al. 2025). However, I also highlight the small size of the A3 set which may result in a less accurate estimation of performance.

Furthermore, I analysed the consistency of ITsFlexible predictions when using different structures containing the same loop as model input. The results show that variations in prediction scores are generally small, demonstrating consistency (Figure B.7).

As structural data are only available for a subset of known antigen receptor sequences (Olsen et al. 2022; Raybould et al. 2024), ITsFlexible was additionally tested using predicted structural models from ImmuneBuilder (IB) and AF2 as inputs. On the antibody CDR sets, similar performance is observed for predictions from crystal structures and IB models and slightly worse predictions are made from AF2 models. While there is no clear correlation between model quality and the error in the ITsFlexible prediction, I observed that higher prediction errors tend to occur at lower model quality (Figure B.8). Difference in classification performance

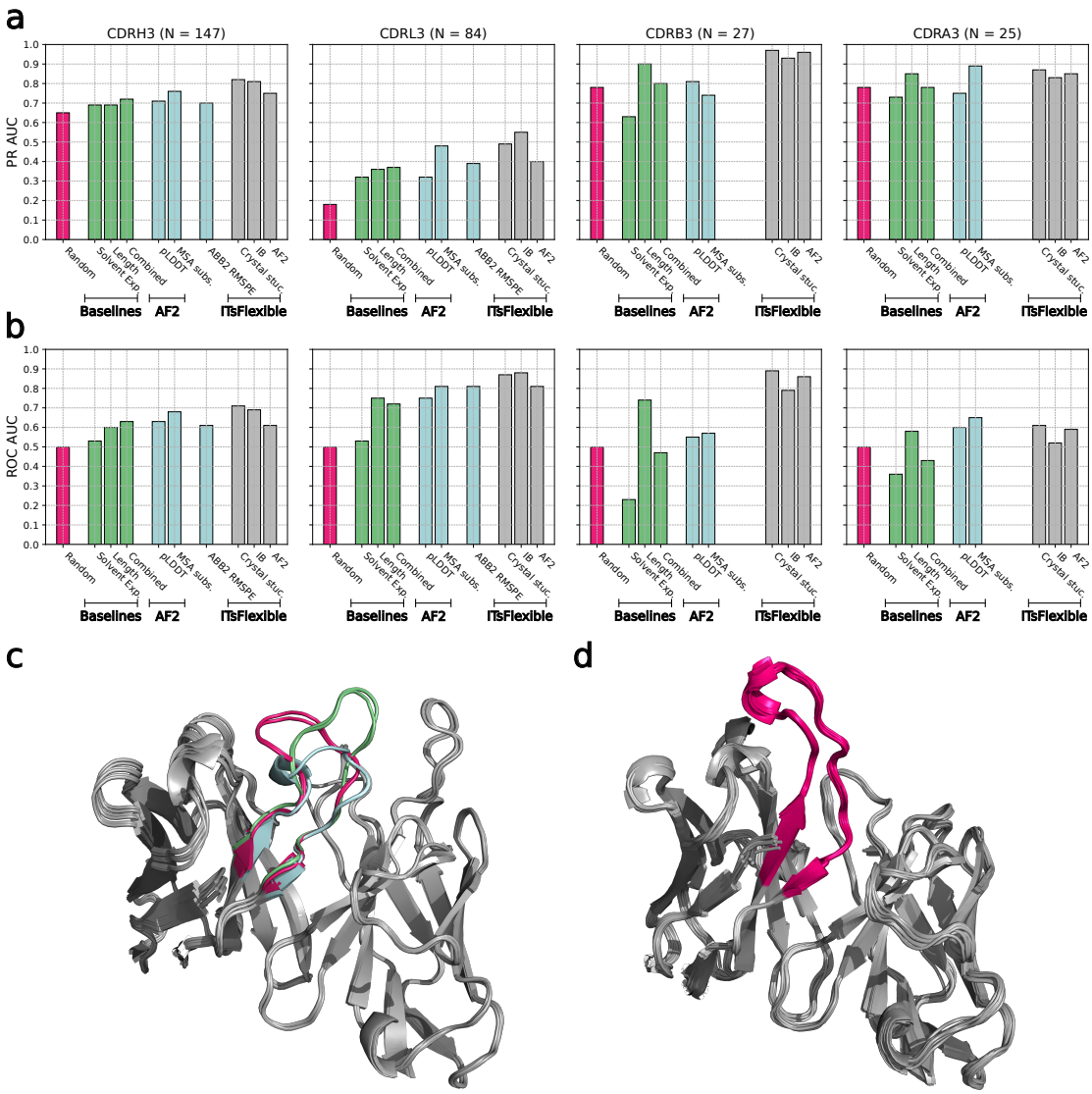


Figure 3.3: ITsFlexible performance evaluated on the CDR test sets. a-b) Classification on the four test sets is evaluated with metrics of (a) PR AUC and (b) ROC AUC. ITsFlexible performance from inputs of crystal structures, ImmuneBuilder (IB) and AF2 models (light grey) is compared to random classification (red), three biophysical baselines (green) and three zero-shot models based on outputs of protein structure prediction tools (blue). c-d) Example of antibodies predicted to be flexible and rigid by ITsFlexible. c) Overlay of 6 structures of the same antibody Fv with CDRH3 predicted to be flexible. Crystal structures indicated that the CDRH3 (highlighted in colour) adopts three different conformations (red, blue, green). d) Overlay of 22 structures of the same antibody Fv with CDRH3 predicted to be rigid. The CDRH3 (red) occupies the same conformation in all 22 structures.

may, therefore, be tied to the lower quality of the AF2 compared to the IB models of the test set (Table B.7). On the TCR sets, predictions from AF2 models are marginally better in PR AUC than from IB models with a larger difference in ROC AUC. As for the antibody sets, I observed a lower model accuracy and a larger error in ITsFlexible prediction score for AF2 (Table B.7). However, for TCRs this does not deteriorate the classification metrics.

Throughout the analysis in this manuscript, I made the following assumptions to determine CDR flexibility. CDR3s are defined as IMGT residues 107-116 (see methods and Figure B.4), structural diversity is calculated as the loop RMSD after alignment on the loop residues and a conformation is defined as a structural cluster with pairwise RMSD between any member below 1.25 Å. Additional analysis was performed on datasets with CDR3s defined by their exact secondary structure (Table B.8), flexibility calculated by alignment on the framework residues (Figure B.9) and different choices of RMSD threshold (Figure B.10). I found that ITsFlexible's high predictive accuracy remains qualitatively similar for the first two experiments whereas the choice of threshold does impact classification to a minor extent.

3.5.4 ITsFlexible matches CDR flexibility in molecular dynamics simulations

While crystallographic data can reveal conformational states of CDRs, it does not directly measure flexibility, and there always remains the possibility that additional conformational states exist but have not been captured. A limitation of the ALL-conformation dataset, therefore, is the lower confidence of 'rigid' compared to 'flexible' labels. Although requirements were introduced to address this limitation (see methods), the true extent of flexibility may well be underestimated.

In an alternative approach, I used MD simulations to classify CDR flexibility. I simulated a set of 19 antibodies and labelled the flexibility of CDR3s (Table B.2). As expected, I observed a higher proportion of flexible CDR3s in MD (84% for CDRH3, 37% for CDRL3) than in the crystal structure data (64% for CDRH3, 18% for CDRL3). ITsFlexible achieved near perfect separation for CDRH3s, with

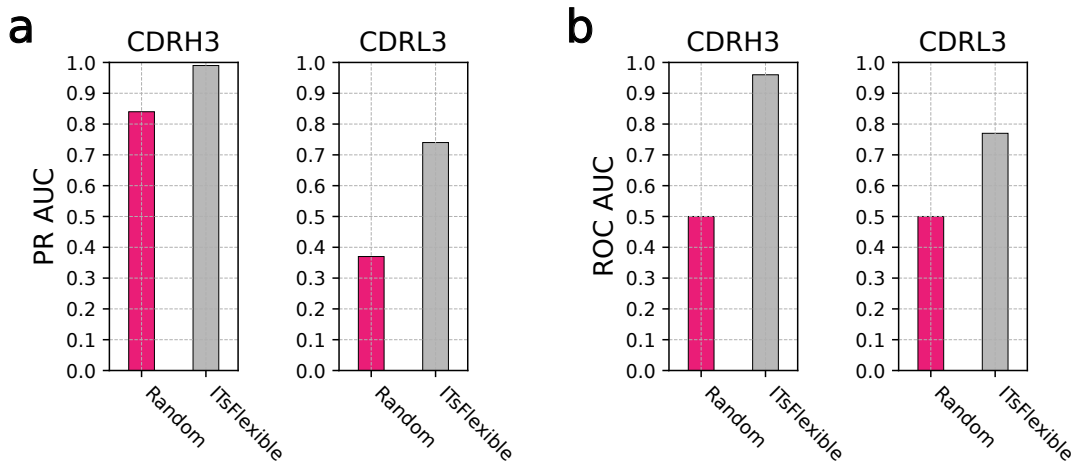


Figure 3.4: ITsFlexible performance evaluated on the MD test set containing 19 antibodies. Classification of CDRH3s and CDRL3s is evaluated with metrics of (a) PR AUC and (b) ROC AUC. A prediction was made for each representative structure extracted from MD and classification performed based on the maximum ITsFlexible score observed across the ensemble.

slightly worse performance for CDRL3s (Figure 3.4), suggesting that ITsFlexible also detects signals promoting flexibility in physics-based molecular simulations.

3.5.5 Cryo-EM experiments confirm predicted flexibility

I chose three antibodies as challenging test cases for experimental validation using cryo-EM. Antibodies in the Patent and Literature Antibody Database (PLAbDab) (Abanades et al. 2024) against the influenza H1N1 hemagglutinin with ITsFlexible score below 0.1 (rigid examples) or above 0.5 (flexible examples) were considered. Three antibodies with low sequence identity to any examples in the training set and loop lengths opposing trends observed in ALL-conformations (long loop for rigid case, shorter loops for flexible cases) were selected. These were imaged in complex with the antigen and the heterogeneity in the density maps, indicative of conformational variety (Punjani et al. 2023), was analysed.

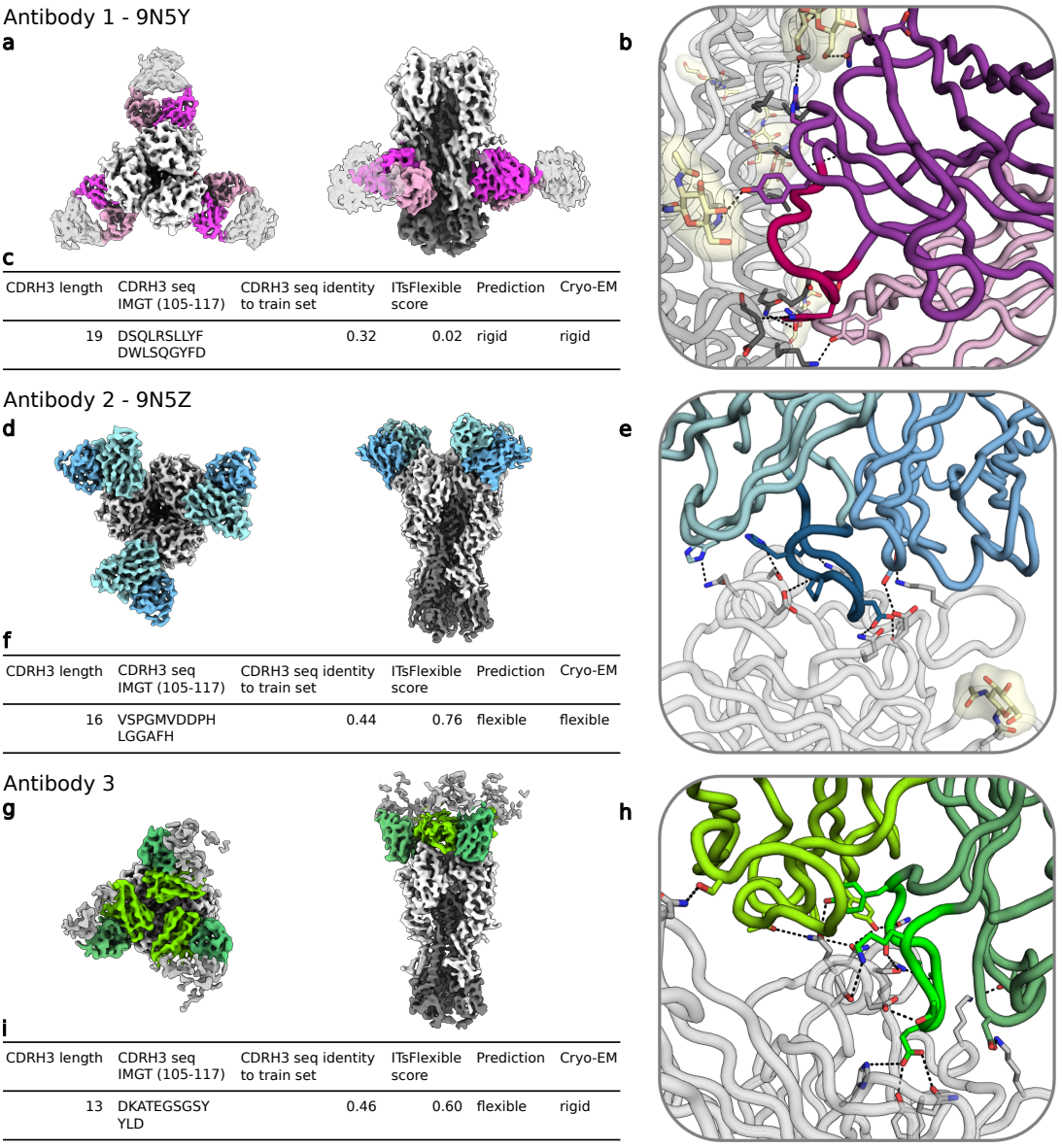
Antibody 1 is an example of a long CDRH3 (length 19, longer than 86% of CDRH3s in ALL-conformation) predicted to be rigid with high confidence (ITsFlexible score of 0.02). The cryo-EM data revealed that the majority of good particles selected from 2D classification adopt one homogenous 3D class, which was

used to build a high-resolution consensus structure (Figure 3.5). In agreement with model prediction these data show that the CDR occupies a single conformation.

Antibody 2 is an example with a slightly shorter CDRH3 loop (length 16) predicted to be flexible with high confidence (ITsFlexible score of 0.76). The initial data processing showed high flexibility in the binding interface of the antibody and the antigen albeit at low resolution. With further 3D classification and data collection, a high resolution structure was obtained of one of the binding states (Figure 3.5). The resolution of the data did not allow building models of distinct conformational states which makes it hard to localize the flexibility to the CDRH3 loop. However, the density maps show clear conformational heterogeneity at the binding interface and induced by the antibody (Supplementary Movie 1 & 2). Flexibility of CDRH3 is a likely cause for this observation.

Antibody 3 is a further example of a CDRH3 predicted to be flexible. The CDRH3 is even shorter than in antibody 2 (length 13) and the ITsFlexible score of 0.60 suggests a lower prediction confidence. The cryo-EM data disagreed with the prediction and revealed no heterogeneity (Figure 3.5).

A potential limitation of the experimental setup is that antibodies were imaged in complex with the antigen. This choice was made as the minimum particle size required for cryo-EM complicates the study of free antibodies. Binding can rigidify residues as additional molecular interactions with the antigen impose extra constraints (Mikolajek et al. 2022) and I expect ITsFlexible predictions to better match the flexibility of the unbound antibody (see discussion). This limitation could be a reason for the absence of flexibility observed in experiments for antibody 3. In particular, a glutamate residue at the tip of the CDRH3 forms 3 hydrogen bond interactions with the antigen (Figure 3.5). Accordingly, an argument could also be made that the unbound state of antibody 1 is more flexible than observed. The data, nevertheless, provide clear evidence for reduced flexibility of antibody 1 compared to antibody 2 when imaged under the same conditions, matching the predictions.



Antibody 1 - 9N5Y

CDRH3 length	CDRH3 seq IMGT (105-117)	CDRH3 seq identity to train set	ITsFlexible score	Prediction	Cryo-EM
19	DSQLRSLLYF DWLSQGYFD	0.32	0.02	rigid	rigid

Antibody 2 - 9N5Z

CDRH3 length	CDRH3 seq IMGT (105-117)	CDRH3 seq identity to train set	ITsFlexible score	Prediction	Cryo-EM
16	VSPGMVDDPH LGGAFH	0.44	0.76	flexible	flexible

Antibody 3

CDRH3 length	CDRH3 seq IMGT (105-117)	CDRH3 seq identity to train set	ITsFlexible score	Prediction	Cryo-EM
13	DKATEGSGSY YLD	0.46	0.60	flexible	rigid

Figure 3.5: Case study antibodies selected for cryo-EM experiments. High-resolution structures of the three antibodies in complex with influenza H1N1 hemagglutinin were solved. a, d, g) Top and side view of the cryo-EM map of antibodies (colour, heavy chain dark, light chain bright) in complex with the antigen (grey). In each structure, three symmetrically arranged copies of the antibody were captured. Antibody 1 binds to the hemagglutinin stem and antibodies 2 and 3 to the hemagglutinin head. b, e, h) Cartoon representation of antibody-antigen binding interface. CDRH3s are shown in different shades of colouring and binding interactions are highlighted in stick representation. c, f, i) Summary tables showing the CDRH3 length and sequence, sequence identity to the closest example in the ITsFlexible training set, the ITsFlexible prediction score (a higher value indicates a higher likelihood of being flexible), a binary prediction of loop flexibility based on the ITsFlexible score and the flexibility determined with the cryo-EM experiments. Additional metadata for the three antibodies is shown in Table B.12.

3.6 Discussion

Conformational changes are fundamental to functional properties of many proteins (Wei et al. 2016; Teilum et al. 2009). In antibodies and TCRs, flexibility, especially of the highly important CDR3s, impacts key properties such as affinity and specificity (Mikolajek et al. 2022; Schmidt et al. 2013; Guthmiller et al. 2020; James et al. 2003). However, prediction of protein flexibility remains difficult and current structure prediction tools are unable to model multiple conformational states of proteins and even more so antigen receptor CDRs reliably (Riccabona et al. 2024; Jing et al. 2024a). One factor that has limited the progress in conformation prediction is the absence of large datasets necessary to train and evaluate such machine learning models.

In this chapter, I focused on the computational prediction of antigen receptor CDR3 flexibility. To help overcome the lack of data, I collected structural motifs with the same secondary structure pattern, defined as a loop connecting two consecutive antiparallel β-strands, across all proteins. Mining the PDB, I created ALL-conformations, a dataset containing more than 1.2 million crystal structures of such loops with more than 100,000 unique sequences. Analysing the conformational flexibility in ALL-conformations, I was able to label 20,000 loops with high confidence as either flexible, adopting multiple conformations, or rigid, occupying a single state.

Using the ALL-conformations dataset, I developed ITsFlexible, a method for the classification of CDR loops as either flexible or rigid. The model was trained on the subset of loop motifs from general proteins in the PDB and evaluated on its ability to predict CDR flexibility in an out-of-distribution setting. ITsFlexible achieved state-of-the-art performance on crystal structure test sets and generalised effectively to MD simulation data. In addition, I also showed that the model achieves similar performance when using inputs of structural models rather than crystal structures, highlighting its applicability to antibodies without experimentally solved structures.

An ablation of ITsFlexible inputs indicated biophysical factors that influence the flexibility of CDRs. In line with previous MD studies (Guloglu et al. 2023),

I identified the arrangement of residues in the structural context of CDRs as a key factor driving flexibility.

Furthermore, the analysis also showed that uncertainty scores of protein structure prediction tools and diversity in structural ensembles generated with established workflows are not reliable predictors of CDR flexibility. The AF2 pLDDT score has been described as a good indicator for disordered regions of proteins (Jumper et al. 2021), however, I found that it was not highly predictive of CDR flexibility. Similarly, the predicted error of the antibody specific ABB2 correlates more strongly with the number of times a particular sequence was present in the training set than flexibility. I also assessed the diversity of CDR ensembles produced by AF2 MSA subsampling as a predictor of flexibility. MSA subsampling is a popular workflow to model multiple conformational states of proteins (del Alamo et al. 2022; Stein et al. 2022; Wayment-Steele et al. 2024; Sala et al. 2023b; Faezov et al. 2023; Sala et al. 2023a; Saldaño et al. 2022). While this approach is more predictive than confidence scores, the workflow again does not capture CDR flexibility accurately.

ITsFlexible's predictive performance was experimentally evaluated using three antibodies chosen as challenging test cases. The antibodies had low sequence identity to the training set and their loop lengths opposed the trends observed in the data (long loop for rigid case, shorter loops for flexible cases). Using cryo-EM, I imaged the antibodies in complex with their antigen and analysed heterogeneity in the density maps (Zhong et al. 2021; Punjani et al. 2023). For two of the three cases the experimental evidence confirmed the predictions. Among the correctly predicted cases, one rigid and one flexible, the former was three residues longer, further demonstrating that the model detects features beyond loop length associated with flexibility. The third antibody featured an even shorter CDRH3 predicted to be flexible although with lower confidence and experiments showed no evidence of conformational heterogeneity. Overall, cryo-EM experiments provide strong evidence that ITsFlexible is predictive of real CDR3 conformational dynamics observed in solution.

One limitation of this work is the approach of identifying flexibility. I label a CDR as flexible if multiple conformational states were observed irrespective of whether the antigen receptor is bound to antigen or in its free state. Previous studies showed that the bound conformation is frequently included in the unbound ensemble suggesting that conformational selection is a more common mechanism of antibody binding than induced fit (Liu et al. 2024; Blackler et al. 2022; Fernández-Quintero et al. 2019b). I, therefore, expect that ITsFlexible predictions are representative of the flexibility of unbound CDRs. This assumption is supported by the agreement with MD simulations of unbound antigen receptors. As some functional properties, such as binding affinity, are dependent on the balance of flexibility in bound and unbound states (Mikolajek et al. 2022), I suggest the prediction of changes in CDR flexibility upon binding as an important future direction. Owing to the limited amount of data capturing CDR conformational states, however, predicting flexibility conditional on the antigen is currently challenging.

In this chapter, I presented ALL-conformations and ITsFlexible. ALL-conformations captures the full range of experimentally observed conformational diversity of loops between antiparallel β-strands, and allows the investigation of CDR dynamics. ITsFlexible accurately predicts whether CDR3s are conformationally flexible or rigid. CDR flexibility has been linked to reduced binding affinity (Mikolajek et al. 2022; Schmidt et al. 2013) and increased polyspecificity (Guthmiller et al. 2020; James et al. 2003), both of which are typically undesirable when engineering antigen receptors for therapeutic use (Sormanni et al. 2018). By allowing the rapid screening of candidate molecules, ITsFlexible can help identify receptors with more favourable therapeutic profiles. However, using crystallographic data to study conformational flexibility has several limitations. Firstly, only proteins with several available crystal structures can be investigated which drastically limits the set of antibodies that can be studied. Secondly, the observed ensembles are incomplete as there is no guarantee that there are not additional conformations that have not been captured. In Chapter 4, I present an alternative approach to investigate CDR dynamics at

scale. I introduce a high throughput MD workflow tuned for antibodies and TCRs which I then used to simulate more than 150,000 immune receptors.

4

Uncovering CDR flexibility through large-scale molecular dynamics

Contents

4.1 Motivation 94
4.2 Contributions 94
4.3 Background 95
4.4 Methods 98
4.4.1 CV3-Fv workflow 98
4.4.2 Datasets 99
4.4.3 Benchmarking 103
4.5 Results 106
4.5.1 Characterisation of Antibody Fv using CALVADOS 3 . 106
4.5.2 Comparison with ensembles observed in experimental structures 110
4.5.3 Investigating the limitation of our flexibility predictions 113
4.5.4 FlAbDab and FTCRDab 115
4.5.5 Analysis of CDR dynamics 116
4.6 Discussion 119

4.1 Motivation

CDR dynamics impact key functional properties of immune receptors. In Chapter 3, I showed that information on CDR conformational states and flexibility can be obtained from crystallographic data. However, the approach is limited to a small set of antibodies for which multiple structures are available. Furthermore, there is no guarantee that experimental structures capture the complete ensemble of accessible conformations and they do not inform on the time-scales of motions or the relative populations of states. All of these issues can potentially be addressed using MD simulations.

In this chapter, I present a high throughput, coarse-grained MD workflow to simulate antibodies and TCRs. I used the method to simulate a set of more than 150,000 immune receptors and performed preliminary analysis of their structural dynamics.

4.2 Contributions

I co-led this project together with Matteo Cagiada. Both of us contributed to the method development, benchmarking and wrote the manuscript on which this chapter is based. I set up the large-scale simulations. King Ifashe performed the all atom MD simulations.

This chapter contains material reproduced from:

Matteo Cagiada*, **Fabian C. Spoendlin***, King Ifashe, and Charlotte M. Deane. "Uncovering the Flexibility of CDR Loops in Antibodies and TCRs through Large-Scale Molecular Dynamics," Biorxiv (2025).

*Denotes equal contributions

4.3 Background

As discussed in Chapter 1, CDR flexibility is thought to balance antibody binding affinity and specificity. Therefore, it has been suggested that the natural immune system exploits CDR conformational flexibility to balance between promiscuous (polyspecific), low-affinity binding and specific, high-affinity interactions. In case studies of antibody lineages, affinity maturation was shown to introduce mutations that promote CDR rigidification and thereby narrow specificity whilst enhancing binding strength. (Schmidt et al. 2013; Babor et al. 2009; Li et al. 2015; James et al. 2003; Blackler et al. 2022; Fernández-Quintero et al. 2019c). Thus, CDR flexibility may support a two-stage strategy: the initial promiscuous binding of a broad range of antigens by germline antibodies, which increases the effective size of the immune repertoire, followed by selection for more specific, high-affinity binders. Although conformational flexibility may be important in modulating functional properties of some antibodies, its functional relevance and applicability at the repertoire level remain unclear (Jeliazkov et al. 2018). To study the functional importance at the repertoire level, high throughput methods to characterise CDR conformational dynamics are needed.

MD simulations are widely used to study protein dynamics and have been applied to study antibody flexibility (Wong et al. 2011; Fernández-Quintero et al. 2019c; Guloglu et al. 2023). However, the high computational cost of all-atom MD limits long-timescale sampling and the number of proteins that can be simulated. Even the largest all-atom simulation datasets, such as ATLAS (Vander Meersche et al. 2024) and mdCATH (Mirarchi et al. 2024), contain a number of simulations far lower than the number of unique experimentally determined protein structures. Coarse-grained (CG) models offer a more scalable alternative, as they reduce the number of particles in the system and thereby increase sampling efficiency. Recent studies have demonstrated that CG models can efficiently simulate datasets of up to 30,000 proteins (Tesei et al. 2024), offering accurate molecular descriptions within a reasonable computational time.

High quality experimentally determined structures are available for more than 3,000 unique antibodies (Dunbar et al. 2014; Schneider et al. 2022). If machine learning-based predictions are also considered a far greater number of antibody structures are available, e.g. Abanades et al. (2023) released a set of 150,000 predicted structures. Coarse-grained MD models offer the most viable approach to extensively investigate the dynamics of a dataset of this size. However, simulating antibody Fvs with CG models comes with several difficulties. While frameworks support antibody stability (Julian et al. 2017), the CDR loops are structurally diverse and potentially dynamic (Kroon et al. 2003; Fernández-Quintero et al. 2019b). A suitable model, therefore, must accurately describe both 'folded' and flexible regions. A few such models have been proposed, including different versions of the Martini force field (Benayad et al. 2021; Thomasen et al. 2022; Thomasen et al. 2024) and more recently CALVADOS 3 a residue-based model specifically designed for multi-domain proteins with disordered linkers (Cao et al. 2024; von Bülow et al. 2025).

In this chapter, we leveraged the efficiency of CALVADOS 3, to simulate and characterise the flexibility of antibody and TCR Fvs. The Fv region of immunoproteins is an ideal candidate for CALVADOS 3 because it consists of two immunoglobulin domains, each of which has alternating CDR loops (disordered regions) and a layer of seven to nine antiparallel β-strands (folded regions). We optimized our simulation protocol to replicate results obtained from all-atom MD simulations and demonstrated that our setup achieves high accuracy in reproducing ensemble metrics. We were able to characterize the flexibility of 3,140 non-redundant, quality-filtered Fv structures from the SAbDab (Dunbar et al. 2014; Schneider et al. 2022) and 280 TCRs from the STCRDab (Leem et al. 2018).

Due to the low computational cost of this approach (two CPU hours per simulation), we incorporated predicted antibody structures into our dataset and, ultimately, simulated over 150,000 immune receptors. We openly released our simulations as new databases in the form of the Flexibility Antibody Database (FlAbDab) and the Flexibility TCR Database (FTCRDab). These large datasets will provide a resource to investigate the functional role of CDR conformational

flexibility and enable the training of machine learning models on a more information rich representation of antibody structure.

4.4 Methods

4.4.1 CV3-Fv workflow

For each query Ab and TCR, the FV structures to use as input for the simulation (VH and VL domains) were extracted from the PDB file and the remainder of the structure, including conserved domains and antigens, was discarded. Then, we ran a short energy minimization of the Fv structure using OpenMM (Eastman et al. 2024). A protocol identical to the one used as part of AF2 (Jumper et al. 2021) was used. The protocol used AMBER ff99SB force field and additional harmonic constraints with a spring constant of 0.7 N/m were placed on all non-hydrogen atoms. Minimization was performed until the system reached a tolerance of 2.39 kcal/mol.

We performed molecular dynamics simulations in the NVT ensemble at 25 °C and pH 7 using the CALVADOS suite (von Bülow et al. 2025). Simulations used a Langevin integrator with a time step of 10 fs and a friction coefficient of 0.01^{-1} ps. Salt-screened electrostatic interactions were modelled using a dielectric constant of 74.2 and an ionic strength of 0.15 M. Simulations were conducted in a periodic box with dimensions of 25 × 25 × 150 nm for a total of 100.5 ns. The 100 ns length was chosen as a compromise between trajectory convergence and computational efficiency (Fig. C.14)

In the first instance, we used the residue model CALVADOS 3 using parameters from the official release (Cao et al. 2024), following the functional forms described in the original CALVADOS publication. Here, regions outside the IMGT-defined CDR loops were treated as folded domain and restrained using a harmonic potential with a force constant (k) of 700 N/m. We named this setup CALVADOS 3-IMGT (CV3-IMGT). Given the poor agreement with all-atom simulation in predicting the flexibility of CDR region, we introduced additional harmonic restraints to selected residue pairs within loops under two conditions: (i) if a hydrogen bond was present between the atoms of the pair in the crystal structure or (ii) if their side-chain centres of mass were in close proximity. Restraints were applied using a harmonic potential with a force constant of k = 175 N/m for the former and k = 350 N/m for the latter.

Hydrogen bonds and side-chain contacts within CDRs were identified using the DSSP algorithm. A hydrogen bond was defined as present if the DSSP energy score was lower than -0.41 kJ/mol for a given acceptor-donor pair. Side-chain interactions were defined as any residue pair with a centre-of-mass distance below 4.5 Å (using the Cα atom for glycine). We named this setup CALVADOS 3-Fv (CV3-Fv).

Coarse-grained simulations were converted to all-atom trajectories using the cg2all package (Heo et al. 2024) with the 'ResidueBased' option selected to convert the residues from beads to all atoms.

The representative ensemble for each entry in the FlAbDab and FTCRDab databases includes frames randomly selected from the simulation trajectories. The total number of frames in each ensemble is proportional to the flexibility of CDRH3, ranging from a minimum of 5 structures for rigid entries (RMSD $<$ 0.5 Å) to a maximum of 15 structures for highly flexible entries (RMSD $>$ 3 Å).

4.4.2 Datasets

All atom MD simulations

The all atom simulations of TCR PDB 7f5k (chains A and B) were obtained from a previous study (Quast et al. 2025) where a single replicate of a 2 μs unrestrained simulation was performed at 300K and pH 7.5.

We performed simulations for antibodies anti-hen egg white lysozyme (PDB: 2Q76) and anti-HBV e-antigen (PDB: 2Q76) as well as Titin Epitope in HLA-A1 (PDB 5bs0) as outlined below.

We performed complete all-atom MD simulations of mouse anti-hen egg white lysozyme antibody F10.6.6 Fab fragment (PDB: 2Q76), monoclonal anti-HBV e-antigen Fab fragment (e6), unbound (PDB: 3V6F), and the Titin Epitope in HLA-A1 (PDB: 5BS0). All non-protein residues were removed from the file. The original anti-HBV e-antigen Fab fragment PDB file contained insertion codes in chains H and L, which were renumbered to follow the sequential IMGT numbering scheme, i.e., assigning consecutive residue numbers based on the IMGT antibody numbering convention. The original F10.6.6 Fab fragment PDB file contained

missing electron density for residues 130–134 of chain B, which was modelled using Chimera X. The structures were protonated at the pH they were crystallised at using H++ (Anandakrishnan et al. 2012), and the AMBER compatible file was extracted from this.

We ran all simulations with AMBER24 (Case et al. 2023) using the ff19sb force field (Tian et al. 2020) with the OPC3 water model (Izadi et al. 2016), and all systems were prepared using tleap from the AmberTools package (Case et al. 2023). For each structure, we manually defined disulphide bonds between the appropriate cysteine residues to maintain a proper tertiary structure. We prepared the solvated systems through a multistep process in which the protein charge was first neutralised, followed by solvation in a truncated octahedral OPC water box with a buffer distance of 10.0 Å from the protein surface to the edges of the box. We finally added additional randomised sodium and chloride ions to achieve a physiological ionic strength of approximately 150 mM, with ion concentrations adjusted based on the initial protein charge state.

All solvated systems underwent multistage relaxation and thermal equilibrium prior to production MD. We performed initial energy minimisation for 1000 cycles using a hybrid approach with 30 cycles of steepest descent followed by conjugate gradient optimisation. We applied strong harmonic positional restraints of 100 kcal mol^{-1} $Å^{-2}$ to all protein atoms while allowing freedom for solvent molecules and ions to optimise their positions around the constrained protein structure. We used a 10 Å cutoff for nonbonded interactions during this initial minimisation phase.

We then heated the systems from 100 K to 298 K over 1 ns (1,000,000 steps with 1 fs time steps) using canonical ensemble (NVT) MD with Langevin temperature regulation. We used Langevin thermostat with a collision frequency of 1.0 ps^{-1} to provide efficient temperature coupling while maintaining the same 100 kcal mol^{-1} $Å^{-2}$ restraints on protein atoms. Temperature ramping was implemented using the nmropt facility with linear scaling from the initial temperature of 100 K to the target temperature of 298 K over the entire 1 ns duration.

Following thermal equilibration, we performed isothermal-isobaric (NPT) equilibration on the system at 298 K and 1 atm pressure for 1 ns using the Monte Carlo barostat algorithm with the same protein restraints maintained. We additionally performed an additional 1 ns NPT equilibration with reduced protein restraints of 10 kcal mol^{-1} $Å^{-2}$ to allow initial structural relaxation while maintaining overall protein integrity.

We carried out a second phase of energy minimisation consisting of 10,000 cycles with selective application of restrictions only to the protein backbone atoms (C_α, N and C atoms) at 10 kcal mol^{-1} $Å^{-2}$, allowing side chain optimisation while preserving the secondary structure. The nonbonded cut-off was reduced to 8 Å for all subsequent simulation stages.

We then achieved a progressive restraint reduction through a series of 1 ns NPT equilibration runs with backbone restraints that steadily decreased from 10 kcal mol^{-1} $Å^{-2}$ to 1 kcal mol^{-1} $Å^{-2}$ to 0.1 kcal mol^{-1} $Å^{-2}$, allowing gradual increase in protein flexibility and conformational sampling. Each stage used Langevin dynamics for temperature control with a collision frequency of 1.0 ps^{-1} and the Monte Carlo barostat for pressure regulation.

In the final equilibration stage we performed 1 ns of completely unrestrained NPT MD to achieve complete system relaxation and establish baseline dynamics prior to production simulations. All equilibration simulations employed SHAKE constraints to fix bond lengths involving hydrogen atoms, particle-mesh Ewald summation for long-range electrostatic interactions, and periodic boundary conditions. We saved trajectory coordinates every 10 ps for analysis, with system energies and properties monitored every 1 ps to ensure stability throughout the equilibration protocol.

Following equilibration, we ran three replicates for antibody systems and four for TCR systems, performing MD production simulations for 500 ns under NPT conditions at 298 K and 1 atm pressure. The production simulations employed a 2-fs integration timestep with SHAKE constraints applied to all hydrogen-containing bonds to enable a larger timestep while maintaining numerical stability. The temperature regulation was maintained using the Langevin thermostat with a

collision frequency of 1.0 ps^{-1}, and pressure control was achieved through the Berendsen weak coupling algorithm. The cutoff for non-bonded interactions was extended to 9 Å for production simulations to improve the accuracy of long-range interactions. Trajectory coordinates were saved every 10 ps (5000 steps) for subsequent analysis, with system energies and thermodynamic properties recorded at the same frequency. We generated checkpoint files every 100 ps to enable simulation continuation and recovery. We did not apply any positional restrictions during production simulations, allowing complete conformational freedom for all components of the system, including protein, solvent, and ions.

Metadynamics simulations

We used a dataset of 9 antibodies (PDBs: 2q76, 3v6f, 3hi6, 3eoa, 1jps, 5i18, 5i15, 3g6d, 2y07) simulated with all-atom metadynamics simulations released by Spoendlin et al. (2025c) for evaluation. Simulation details from the published protocol are reported (Fernández-Quintero et al. 2019c). In a first step, well-tempered metadynamics was used for enhanced sampling. A linear combination of the sine and cosine ψ torsion angles of CDR loops were defined as collective variables. In a second step, representative structures from the metadynamics simulations were used as starting points for short classical MD simulations to obtain unbiased trajectories. Classical MD simulations were clustered based on CDR RMSD with an average linkage hierarchical clustering algorithm and RMSD cut off of 2.5 Å to extract the representative structures we used for evaluation.

Experimental structures

We ran simulations for all non-redundant and quality filtered antibody Fvs, scFvs and TCR Fvs with experimentally resolved structures. We extracted all antibody and scFv structures from the Structural Antibody Database (SAbDab) on May 29th 2024 and TCR structures from the Structural TCR Database (STCRDab) on April 10th 2024. We filtered structures looking for those with a resolution lower than 3.5 Å and no missing residues between IMGT numbered positions 5 and 128. In case of multiple available structures, we selected one structure for each Fv sequence.

We preferred unbound structures and those solved by X-ray crystallography. This resulted in a set of 3033 antibodies, 132 scFvs and 281 TCRs. 3009 antibodies, 131 scFvs and 280 TCRs were successfully simulated.

Predicted structural models

In addition to the experimental structures we simulated a larger dataset of predicted structural models. We used the dataset released by Abanades et al. (2023) of 148,000 non-redundant paired antibody sequences from the OAS (Olsen et al. 2022) predicted with ABodyBuilder2 (Abanades et al. 2023).

4.4.3 Benchmarking

Evaluation against all atom and metadynamics simulation

We used four metrics to compare the CV3-Fv protocol against all-atom and metadynamics simulations: frame-pairwise RMSD, $\text{RMSD}_{\text{mean}}$, residue-level RMSF and frame coverage.

Frame-pairwise RMSD measures the pairwise RMSD between all frames in a simulation. For each CDR, we independently aligned simulation frames to the mean simulation coordinates using the Cα atoms of CDR residues. We then computed the full (number of frames x number of frames) pairwise Cα RMSD matrix for each CDR.

$\text{RMSD}_{\text{mean}}$ evaluates the distance of each frame to the mean structure in the simulation. First, we aligned the simulation frames to the mean Cα coordinates across a simulation. We performed separate alignments and evaluations for the two chains (H and L for the antibodies and alpha and beta for the TCRs). We then evaluated the RMSD of each CDR to the mean coordinates for each frame, using the Cα atoms of each CDR residue.

Residue-level RMSF measures the displacement of Cα atoms across the simulation. We aligned the simulation frames to the mean Cα coordinates across a simulation. We performed separate alignments and evaluations for the two chains (H and L for the antibodies and alpha and beta for the TCRs). The standard formula was used to calculate RMSF of all Cα atoms of Fv residues.

Frame coverage measures coverage of the conformational space explored in all atom simulation by the CV3-Fv simulations. We compute the pairwise distance matrix of all frames in the all atom simulations to all frames in the CV3-Fv simulations. We used Cα RMSD of CDR residues after alignment on Cα atoms of CDR residues. We defined the coverage of all atom frames as the minimum distance between an all atom frame and any CV3-Fv frame, and we analysed the fraction of all atom frames covered below a given distance threshold.

Evaluation against experimental ensembles

We evaluated simulations against structural ensembles observed in crystal structures. First, we collected antibody and scFv structures in the experimental structures dataset and then we filtered for structures solved by X-ray crystallography. We then grouped the filtered structures by Fv sequence identity.

For each Fv sequence, we calculated the number of observed conformations for each of the six CDRs individually. We used pairwise Cα RMSD of CDR residues after alignment on Cα atoms of CDR residues. We defined a conformation cluster by agglomerative clustering with a complete linkage criterion. We used RMSD thresholds of 0.5 Å and 1.25 Å as a more narrow and a wider definition of a conformation. This selected clustering algorithm means that the RMSD of any two structures within a cluster is below the threshold.

For each CDR, we selected all antibodies with more than one experimental conformation (0.5 Å threshold) and calculated a metric we called alternative conformation coverage. Alternative conformation coverage evaluates how well the CV3-Fv simulations cover the experimentally observed conformational space by measuring the minimum distance between an experimental conformation and our simulations. Specifically, for each cluster (conformation), we calculated the pairwise Cα RMSD of CDR residues after alignment on Cα atoms of CDR residues from all cluster members to all simulation frames. We then selected the minimum distance for each member and defined the mean across all members as a conformation coverage metric. We reported the coverage of all conformations except the one

containing the structure from which the simulations were started as our metric of alternative conformation coverage. Lastly, we analysed the fraction of alternative conformations covered below a given distance threshold.

To compare the flexibility observed in the CV3-Fv simulations against the experimental data, we followed the analysis used in (Spoendlin et al. 2025c). Each CDR was assigned a label of 'flexible', 'rigid' or 'unknown'. CDRs with more than one conformation (1.25 Å threshold) were labeled as 'flexible'. CDRs with a single conformation observed in more than 5 separate PDB structures were labeled as 'rigid'. The remaining antibodies were assigned to the 'unknown' group. We then assessed how predictive flexibility measures of CV3-Fv simulations are of flexible and rigid labels. We then calculated the frame-pairwise RMSD (Cα RMSD of loop residues after alignment on Cα atoms of loop residues) of simulations. The 95th percentile of the frame-pairwise RMSD was then used as an input for binary classification of loops as 'rigid' or 'flexible' with a logistic regression. The 95th percentile rather than the maximum frame-pairwise RMSD was used to avoid inflation of the value by rare outliers. To assess the accuracy, we used the area under the precision-recall curve (PR AUC) as evaluation metric for the classification.

We further compared the changes in VH-VL orientation in our simulations against the distributions observed in experimental structures. VH-VL orientation was characterized by 5 angles (HL, HC1, HC2, HL1, HL2) and a distance (dc) as described in a previous work (Dunbar et al. 2013). We then evaluated the standard deviation of the six metrics within a structural ensemble for all Fvs with multiple available structures.

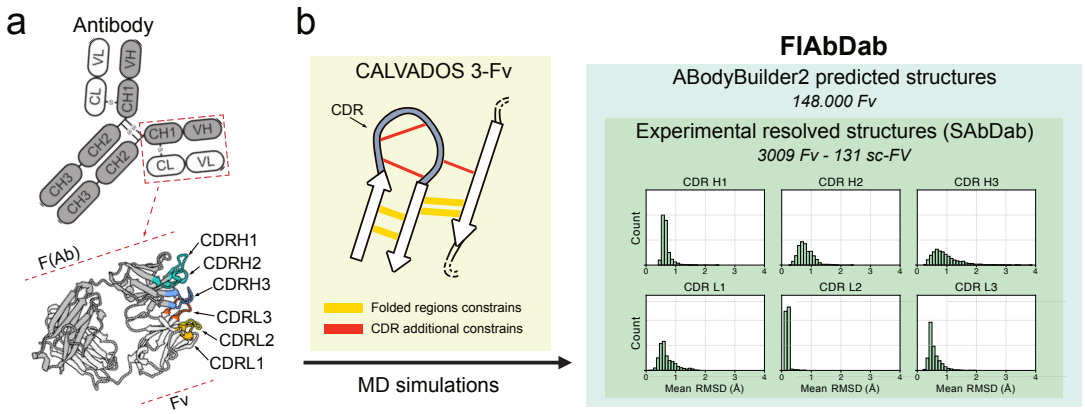


Figure 4.1: Schematic illustration of our approach to characterise flexibility in antibody Fv regions and generate conformational ensembles using molecular dynamics simulations. (a) We characterised the flexibility of a fragment variable (Fv) region of an antibody, especially looking at complementarity-determining regions (CDRs). (b) We performed molecular dynamics simulations based on CALVADOS 3-Fv, an antibody *ad-hoc* setup for the residue-level model CALVADOS 3. In CALVADOS 3-Fv, additional restraints are introduced to better characterise the flexibility of CDR residues. This enabled more accurate modelling of Fv loop flexibility at a lower computational cost. (c) We used CALVADOS 3-Fv to create our new database, FlAbDab, which can be used to investigate Fv dynamics and train new models. FlAbDab contains over 150,000 simulated Fvs, including those with experimentally solved and computationally predicted structures. We also used CALVADOS 3-Fv simulations to characterise all the unique experimental structures in SAbDab.

4.5 Results

To assess the structural flexibility of CDR loops, we used CG MD simulations to generate and characterise conformational ensembles of Fvs for over 3000 experimentally resolved and 150,000 computationally predicted antibody structures (Fig. 4.1), as well as 280 TCR Fvs.

4.5.1 Characterisation of Antibody Fv using CALVADOS 3

We first evaluated the accuracy of the CALVADOS 3 model for simulations of Fv. We simulated Fv regions with CALVADOS 3 by applying a harmonic potential to folded regions outside CDR loops and leaving the CDR loops free from additional constraints. The CDR loop boundaries were defined using IMGT definitions (Lefranc et al. 2003) after the structures were IMGT numbered using ANARCI (Dunbar

et al. 2016). Hereafter, we refer to this simulation setup as CV3-IMGT.

To benchmark the accuracy of the CV3-IMGT setup, we used two different Fvs: the first from a mouse anti-hen egg white lysozyme antibody (F10.6.6, PDB code 2Q76), and the second from an anti-HBV e-antigen monoclonal antibody (anti-HBeAg e6, PDB code 3V6F). For both Fvs, we used as comparison in-house all-atom MD simulations (Fig. C.1) and cluster structures from previously published metadynamics (metaD) simulations (Spoendlin et al. 2025c) . We found that, in this setup, CALVADOS 3 simulations of the Fv tended to overestimate the flexibility of all CDRs in both chains (Fig. C.2).

We hypothesised that this excessive flexibility observed was caused by misrepresenting CDR loops as disordered regions. CDR loops often include secondary structure elements (Fig. C.3), such as extended β-strands or α-helices (Chothia et al. 1987; Weitzner et al. 2015) and show a significant number of non-covalent interactions, which are typical of folded regions, between their residues as well as with residues in framework regions or in different CDRs. To capture these additional CDR interactions in our CALVADOS 3 simulations, we applied additional constraints to the CDR loops using information extracted from the Fv input crystal structures. We named this setup CALVADOS 3-Fv (CV3-Fv).

We evaluated the accuracy of our new setup by testing it on the two antibody systems used to evaluate the CV3-IMGT setup, as well as two additional TCR systems (PDB 5BS0 and 7F5K). First, we observed an improved agreement in CDR flexibility predictions between our CV3-Fv simulations and all-atom MD simulations, compared to CV3-IMGT simulations (Fig. C.2). We also used the harmonic ensemble similarity metric from the ENCORE package (Tiberti et al. 2015) to evaluate the similarity of the CV3-Fv ensembles to the all-atom simulation ensemble. We confirmed that, for all simulated antibody and TCR systems and CDRs, the CV3-Fv setup gave greater agreement with the all-atom simulation than with the CV3-IMGT simulations (Fig. C.2e,f, Fig. C.12).

We then used 'frame coverage', defined as the minimum distance between each all-atom frame and any CV3-Fv frame, to assess the sampling of conformational

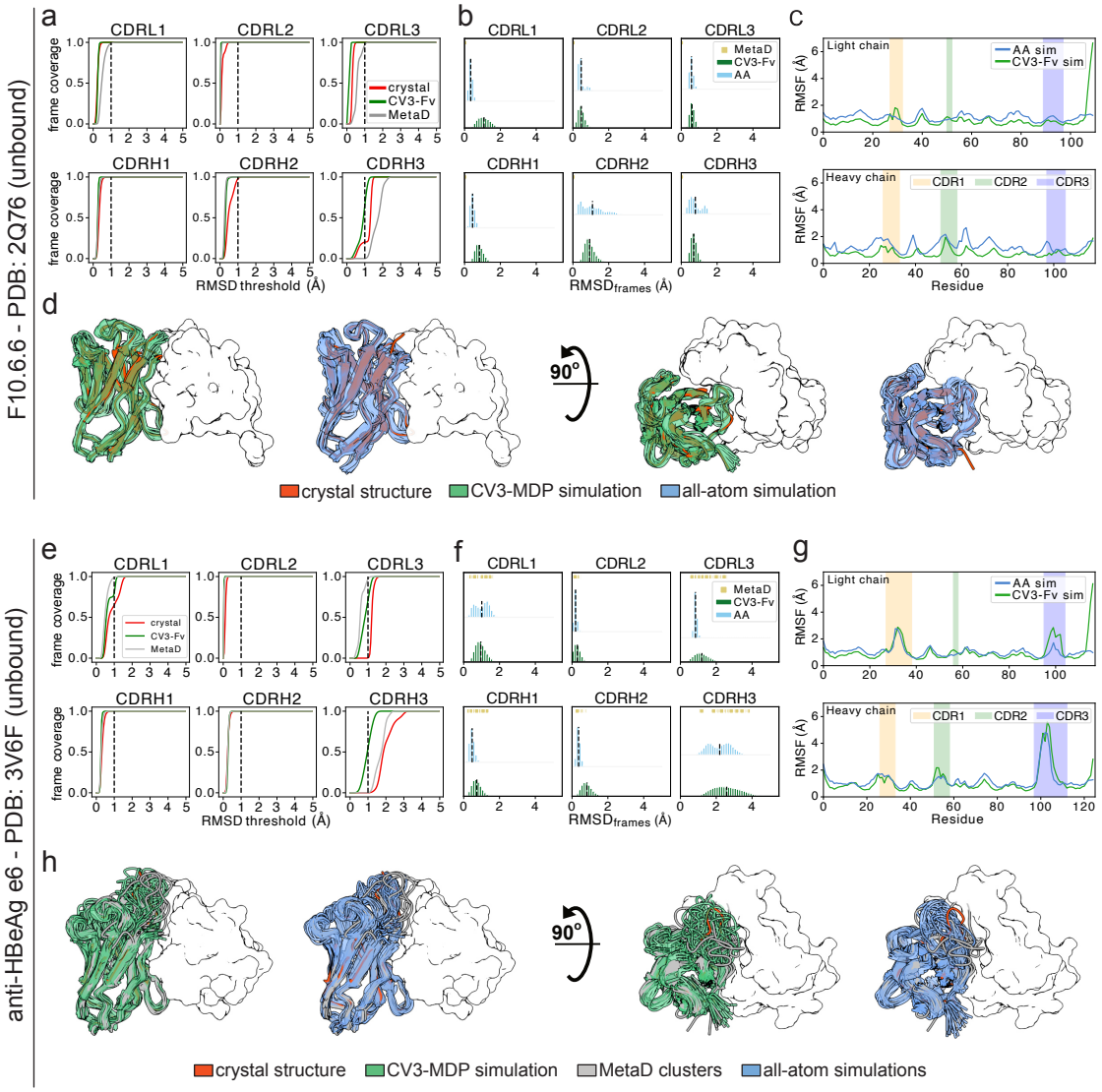


Figure 4.2: Benchmarking the accuracy of CV3-Fv simulations against all-atom and metadynamics simulations. Comparison of the dynamical properties of Fv regions, particularly the CDR loops, for two monoclonal antibodies: anti-hen egg white lysozyme (PDB: 2Q76) and anti-HBV e-antigen (PDB: 2Q76). (a,e) 'Frame coverage' showing the proportion of all atom frames covered at a given threshold by any CV3-Fv frame (green) as well as for the metaD cluster structures (blue) and the starting crystal structure (red). The fraction of all-atom covered frames is reported on the y-axis and the corresponding threshold distance on the x-axis. A black dashed line indicates the 1Å threshold. (b, f) Framewise root-mean-square deviation (RMSD) distribution for all six CDR loops, comparing CV3-Fv (green), all-atom simulations (light blue), and metaD cluster structures (yellow). Mean RMSD values are indicated with black dashed lines . (c, g) Per-residue RMSF values from CV3-Fv (green) and all-atom (blue) simulations. Shaded backgrounds mark different CDR loops. (d, h) Structural ensembles of the Fv region: light chains are shown as white surfaces and heavy chains as cartoons. For the heavy chain, CV3-Fv (green) and all-atom (blue) representative structures are overlaid on top of the starting crystal structure (red) and metaD clusters. Left and right panels show side and top views, respectively.

states (4.2a,e and C.11a,f). In the antibody systems, CV3-Fv simulations captured over 80% of the conformational space observed in the all-atom simulations across all CDR loops within a 1Å threshold. At the individual level, frame coverage exceeded 90% for loops L1, H1, L2, H2 and L3; and reaches approximately 60% for H3. We observed similar results for the TCR systems, with high coverage for many of the CDRs, with nearly 100% of frames covered below 1Å by the CV3-Fv protocol. For CDRs A2 and A3 of 5bs0 and B3 of TCR 7f5k, we observed slightly lower values falling to a minimum of 66%. These results, together with the correlations observed in the ensemble similarity metrics described above, suggest that CV3-Fv simulations achieve a close sampling similarity to all-atom simulations.

As conformational coverage alone does not provide insight into the overall flexibility of the ensemble, we looked at commonly used flexibility metrics such as frame-pairwise root-mean-square deviation (RMSD) and root-mean-square fluctuations (RMSF) to collect additional information. We found that the CV3-Fv setup accurately captures the all-atom dynamics of both rigid and flexible loops in antibodies and TCR systems, with RMSD and RMSF values that closely match those from all-atom simulations (Figs. 4.2b–c, f–g and C.11b–c, g–h). For highly flexible loops, such as CDR H3 of the anti-HBeAg e6 Fv, the dynamics remain limited compared to those sampled using enhanced sampling techniques like metadynamics as these methods can explore conformational states separated by high energy barriers, regions that a residue-level model like CV3-Fv may not fully access. We also observed that the frame-wise RMSD distributions from the CV3-Fv simulations lacked the multimodal characteristics occasionally seen in all-atom simulation profiles, displaying instead more Gaussian-like profiles. This suggests that CV3-Fv simulations may not accurately reproduce either the balance between distinct conformational states or the energy barriers that separate them, limiting the accuracy of the predicted kinetics. However, this behaviour is not surprising, given that coarse-grained models are known to smooth the protein energy landscape (Kmiecik et al. 2016; Jin et al. 2022), often due to a lack of convergence.

We further extended our CV3-Fv benchmark by analysing simulations for four additional Fvs: three from antigen-bound crystal structures and one from an unbound Fv. Weighted cluster structures from metaD simulation were available for each system for comparison. We found good agreement between our CV3-Fv simulations and the metaD structures in terms of predicted flexibility for each CDR (Figs. C.5 and C.4). However, we observed that, in all simulations (including those of our first two benchmark systems), the CDRH2 loops exhibited greater conformational diversity in our simulations than in the metaD clusters, suggesting a possible overestimation of its flexibility. Nevertheless, given the short length outside the framework of CDRH2 (Kelow et al. 2022) and its limited functional role, the impact of this overestimation is expected to be minimal.

4.5.2 Comparison with ensembles observed in experimental structures

As multiple experimental structures are available for several antibodies, often capturing distinct conformational states, we examined how well the CDR conformational space, flexibility and the distribution of VH-VL orientation angles observed in our CV3-Fv simulations match the experimental data.

We investigated the covered conformational space. We extracted all Fvs with multiple conformations in any of the CDRs from SAbDAb (numbers for each CDR are shown in Fig. 4.3a) and assessed their coverage in CV3-Fv simulation using 'alternative conformation coverage'. This metric measures the minimum distance from any simulation frame to each experimental conformation except for the conformation from which the simulation was started (Fig. 4.3a). We compared against a baseline of the same metric computed from only the starting structure and found that our simulations consistently outperformed this measure. Specifically, CV3-Fv simulations sampled over 80% of experimentally observed CDRH1 and L1 conformations and over 90% of H2, L2 and L3 conformations below 1Å. For CDRH3 coverage was slightly lower, with coverage of around 60% of conformations below 1Å.

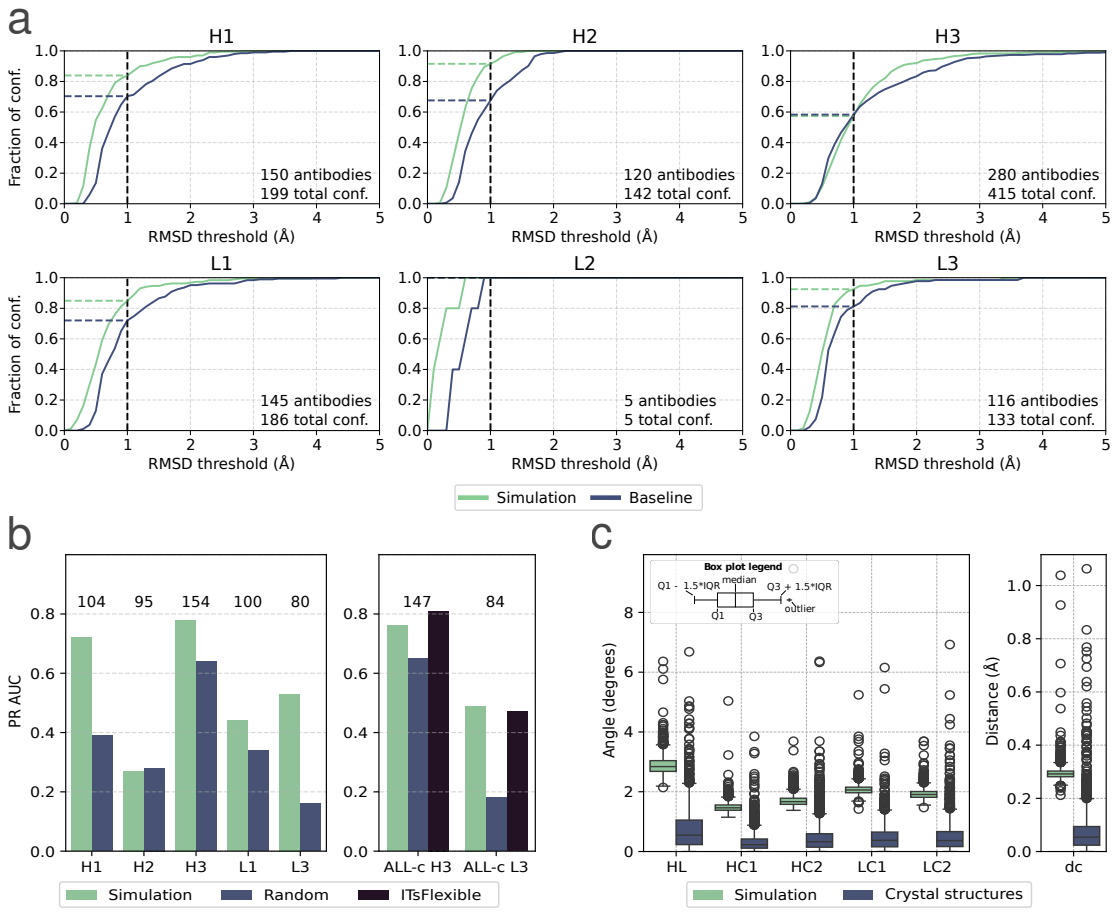


Figure 4.3: Comparison of simulations against experimental structural ensembles. (a) Alternative conformation coverage of CV3-Fv simulations. The plot shows the percentage of alternative conformational states (all conformations excluding the conformation of the starting structure) covered by CV3-Fv at each distance threshold (green). The baseline (blue) provides a comparison by calculating the same metric based on only the starting structure. (b) Prediction of CDR flexibility from RMSD observed in CV3-Fv simulations. PR AUC for classification of rigid and flexible CDRs (determined in crystal ensembles) from frame pairwise RMSD in CV3-Fv simulations (green). Analysis was performed for five CDRs calculating RMSDs using the standard IMGT CDR definitions (H1, H2, H3, L1, L3). Performance was also analyzed on the ALL-conformations CDRH3 and CDRL3 datasets which use a modified CDR definition. The Classification is compared to a random baseline (blue) and the ITsFlexible method (black). The number of samples is reported above each bar. (c) Distribution of VH-VL orientation angles (HL, HC1, HC2, LC1, LC2) and inter domain distance (dc). The distribution of the metrics in CV3-Fv simulations (green, 1764 samples plotted) and crystal structure ensembles (blue, 3140 samples plotted) was calculated and the standard deviation is shown.

To compare the CDR flexibility in the simulations to experimental data, we used a previously described approach (Spoendlin et al. 2025c). CDRs of Fv structures from SAbDAb with multiple available structures were assigned binary labels (Spoendlin et al. 2025c). Each CDR was labelled as either 'rigid' or 'flexible', depending on whether a conformational change of more than 1.25 Å was observed between different crystal structures (see Methods for details). This approach does not guarantee that all CDRs labelled 'rigid' do not exhibit conformational flexibility as it is possible that some conformational states have not been captured. However, the set should be heavily enriched for CDRs that lack flexibility. We analysed if the magnitude of the frame-pairwise RMSD in the simulation agrees with these labels. Specifically, we used RMSD as an input of a logistic regression and monitored the area under the precision-recall curve (PR AUC; Fig. 4.3b). As no CDRL2s were labelled as 'flexible' we only performed analysis for the remaining five CDRs (the number of datapoints is indicated in the figure). PR AUC predicted from CV3-Fv RMSDs exceeds a random baseline for four of the five CDRs. For CDRH2, we observed a value close to random, which may be explained by our previous observation that the CV3-Fv protocol overestimates CDRH2 flexibility. We further analysed performance on ALL-conformations (Spoendlin et al. 2025c) which allows for comparison against ITsFlexible (Spoendlin et al. 2025c). ITsFlexible is a deep learning tool specifically trained for this binary classification task and was benchmarked on the ALL-conformations data set. ALL-conformations contains CDRH3 and CDRL3 loops assigned 'rigid' and 'flexible' labels as in the data sets introduced in this chapter with a small difference in the IMGT residues over which RMSD is calculated. On both datasets, CV3-Fv achieves results comparable to ITsFlexible. Taken together, these results indicate a good agreement between flexibility in CV3-Fv simulations and experimental data.

Lastly, we compared the distribution of VH-VL orientation angles (HL, HC1, HC2, LC1, LC2) and the inter domain distance (dc) to observations in crystal structure ensembles. Our simulations tend to sample a slightly wider distribution of angles (1-2 degrees) and distances (0.3 Å) than observed in the experimental

ensembles (Fig. 4.3c). In absolute terms these are small deviations and overall the method captures the ensemble well.

4.5.3 Investigating the limitation of our flexibility predictions

Monoclonal antibodies can undergo a range of conformational changes upon binding to their specific antigens. The extent and nature of these changes depend on several factors, including affinity maturation and the structural constraints imposed by the antigen (Zhao et al. 2019). Moreover, for the purpose of crystallization, antibodies are often bound to small molecules, peptides or partial fragments of the antigen. As a result, paratopes may exhibit significantly different conformations compared to the unbound state, with smaller buried surface areas and fewer hydrogen bonds compared to antibody–antigen complexes (Lee et al. 2022). Therefore, we assessed how the choice of starting structure affects simulations using three Fvs. For each of these, crystal structures representing different starting conditions (unbound, bound, bound in alternative conformation) were available.

First, we examined Efalizumab Fv, which we found in three different conformations in the PDB: two from a bound state (PDB ID: 3EOA) and one from an unbound state (PDB ID: 3EO9). The CDRs in these structures were moderately similar, with RMSDs consistently below 1 Å for all CDRs. We found consistent flexibility predictions for all CDRs across all three simulations, regardless of the initial structure (Fig. C.8) including CDRH3 (Fig. C.8c). When we compared our results with cluster structures from metaD simulations, we found that the CV3-Fv simulations depict CDR H3 as a more rigid loop than the metaD conformations suggest. It is interesting to note that, despite the metaD clusters sampling a broader and more flexible ensemble, they do not include conformations close to the initial unbound crystal structures (Fig. C.8c,e).

As small conformational changes between starting structures did not impact the outcome of simulations, we continued our investigation with Fvs showing larger

differences (RMSD > 1 Å) in CDR conformation. We used Fvs of antibodies D3h44 and AL-57 as two example systems.

For D3h44 Fv, we collected four distinct conformations: one from an unbound structure (PDB ID: 1JPT) and three from structures bound to two different antigens – another Fv (PDB ID: 1PG7) and a tissue factor (PDB ID: 1JPS). The conformations for the CDRH3 loops differ between the bound and unbound state, but also between the two complexes, with RMSD over 1 Å. We found good agreement between flexibility metrics across simulations starting from the different crystal conformations (Fig. C.6a,b). As in previous cases, we also found that CV3-Fv predicted flexibility of CDRH3 was underestimated compared to the metaD conformation. However, when we inspected the conformational ensembles, in particular for CDRH3 , we observed significant differences between the ensembles (Fig. C.6c,d), with ensembles generated from the unbound structure and the Fv-Fv complex differing significantly from those generated from the complex with the tissue factor. We observed the same pattern in the metaD simulations, which were initiated from the tissue factor–complex crystal structure and did not sample conformations close to the unbound state.

We saw similar patterns in the second system analysed, the Fv of AL-57, which had three available crystal structures — two in complex with an antigen (PDB: 3HI6) and one unbound (PDB 3HI5). Here, the difference in CDRH3 conformations between the bound and unbound structures is over 1.5 Å. As before, we observed minimal differences in the flexibility metrics obtained for the different CV3-Fv simulations (Fig. C.7a, b), regardless of the initial structure. However, when we examined the structural ensemble of CDRH3 (Fig. C.7c,d,e), we found that simulations initiated from the bound state did not sample the conformational space of the unbound crystal structure. Again, this observation was consistent with the cluster centroids from the metaD simulations. We then examined the torsion angles of the CDRH3 loop and identified a small number of residues with distinct distributions of these angles for specific residues, particularly Asp109 and Phe115 (Fig. C.7f), located near the termini of the loop and close to the characteristic CDR 'kink' turn (Weitzner et al. 2015). Here, simulations initiated from the bound

conformation explored a torsional space markedly different from that of the unbound state and were unable to access it. The inability of the bound simulations, even in the structures from metaD simulations, to reach the unbound state suggests that these torsion angles may be separated by high energy barriers. This, combined with differences in the harmonic restraints typically applied to CDR residues (which depend on the starting crystal conformation), may further restrict their ability to explore alternative conformations.

Overall, these results suggest that starting from a bound or unbound conformation might lead to different conformational bundles, even when flexibility metrics remain consistent across simulations. This is particularly evident when bound structures are used as starting points, as CDR loops may undergo substantial conformational changes that place them in regions of the energy landscape separated from the unbound state by significant energy barriers.

4.5.4 FlAbDab and FTCRDab

Having established that our CV3-Fv protocol captures the flexibility of CDR loops, we created the Flexibility Antibody Database (FlAbDab) and the Flexibility TCR Database (FTCRDab), two large databases of simulated antibodies and TCRs. We initially simulated all non-redundant and quality filtered experimentally solved structures before a cutoff date (May 29th 2024 for antibodies, April 10th 2024 for TCRs). This led to simulations of 3140 antibody Fvs (including standard Fvs and single chain Fvs) and 280 TCR Fvs. We further increased the size of FlAbDab by simulating an additional dataset of 148,000 antibody structures predicted with ABodyBuilder2 and released as part of their original publication (Abanades et al. 2023). The CV3-Fv protocol runs with a low computational cost of approximately 2 hours on a standard CPU for a single antibody or TCR. This allowed us to create the two databases using approximately 300,000 hours of CPU computation.

4.5.5 Analysis of CDR dynamics

Although the primary focus of this work was the generation of a new dataset we also carried out a brief characterisation of the CDR flexibility from the simulations of antibodies and TCRs started from experimentally determined structures. This set contained 3140 antibody and 280 TCR Fvs.

We used RMSD and RMSF metrics to characterize the dynamics of antibody CDRs. Specifically, we used the RMSD to the mean simulation structure ($RMSD_{mean}$) and residue-level RMSF. As expected the highest $RMSD_{mean}$ is observed for CDRH3 with a median of 0.9 Å (Figure 4.4), as it tends to be the longest CDR. However, even for this loop the majority of samples show only moderate flexibility with only a few outliers exhibiting RMSD values of above 2 Å. CDRH2 shows the second highest $RMSD_{mean}$, which may seem surprising as H2s tend to be shorter than most other CDRs. This finding may be linked to CV3-Fv's tendency to overestimate H2 flexibility. However, nearly all residues defined as CDR2 by IMGT are actually loop residues. By contrast, IMGT-defined CDR1 and CDR3 loops frequently include residues located in the β-strands at either end of the loop. This structural difference is likely to be responsible for the relatively high RMSD values observed for CDR2. CDRs H1, L1 and L3 show similar magnitudes of flexibility with mean $RMSD_{mean}$ of around 0.6 Å. A very low $RMSD_{mean}$ is observed for CDRL2, which is explained by an average CDRL2 length of 3 residues in our dataset. We observed a strong correlation between the loop $RMSD_{mean}$ and CDR length (Pearson R 0.77) as well as more specifically for CDRH3 $RMSD_{mean}$ and length (Pearson R 0.77).

We further analyzed the relationship between CDR $RMSD_{mean}$ and residue-level RMSF and saw two regimes of loop movements. In the first, we observed high RMSD and low RMSF values (Figure 4.4c) which indicates a large structural change with little fluctuation over time, representing slow transitions between dissimilar conformational states. This regime appears most frequently in CDRH3. For the other five CDRs, we more commonly observed a second regime described by low RMSD and high RMSF values which describes higher-frequency oscillations around a narrower conformational state.

We then investigated whether our data shows a connection between antibody maturity and flexibility. Several studies (Schmidt et al. 2013; Babor et al. 2009; Li et al. 2015; James et al. 2003; Blackler et al. 2022; Fernández-Quintero et al. 2019c) have suggested CDR rigidification upon maturation while others argue against it (Jeliazkov et al. 2018). First, we confirmed that our simulations (Fig. C.9) had successfully characterised the rigidification of CDR loops previously reported in specific antibody maturation lineages (Fernández-Quintero et al. 2020). We then investigated if antibodies more distant from the germline show lower flexibility. We could not establish a correlation between $\mathrm{RMSD}_{\mathrm{mean}}$ and the sequence identity to the V and J genes of the corresponding germline for any of the CDRs (Figure C.10a & b). As lower sequence identity does not strictly mean an antibody is more mature, we also examined whether germline antibodies in general are more flexible than matured antibodies. We labelled an antibody chain as 'naive' if both V and J genes were identical to the germline and 'mature' if either gene contained more than 3 mutations. Comparing $\mathrm{RMSD}_{\mathrm{mean}}$ of mature and naive antibodies, we observed that naive antibodies were on average more flexible than mature ones. However, the differences were small and we could not establish statistical significance. Taken together, our data does not show a clear trend of rigidification with maturation.

We also analysed the TCR simulations. CDRs B3 and A3 show the highest $\mathrm{RMSD}_{\mathrm{mean}}$ values in our data (Figure C.13). These tend to be the longest loops in TCRs and exhibit the highest sequence diversity (Quast et al. 2025). On average, we found slightly lower RMSD values compared to the antibody set. As we again observed a correlation between length and flexibility, this finding can be attributed, at least in part, to the shorter length of TCR CDRs. We also highlight that most TCR CDRs exhibit low RMSD and high RMSF, indicating faster oscillations around a narrower conformational state.

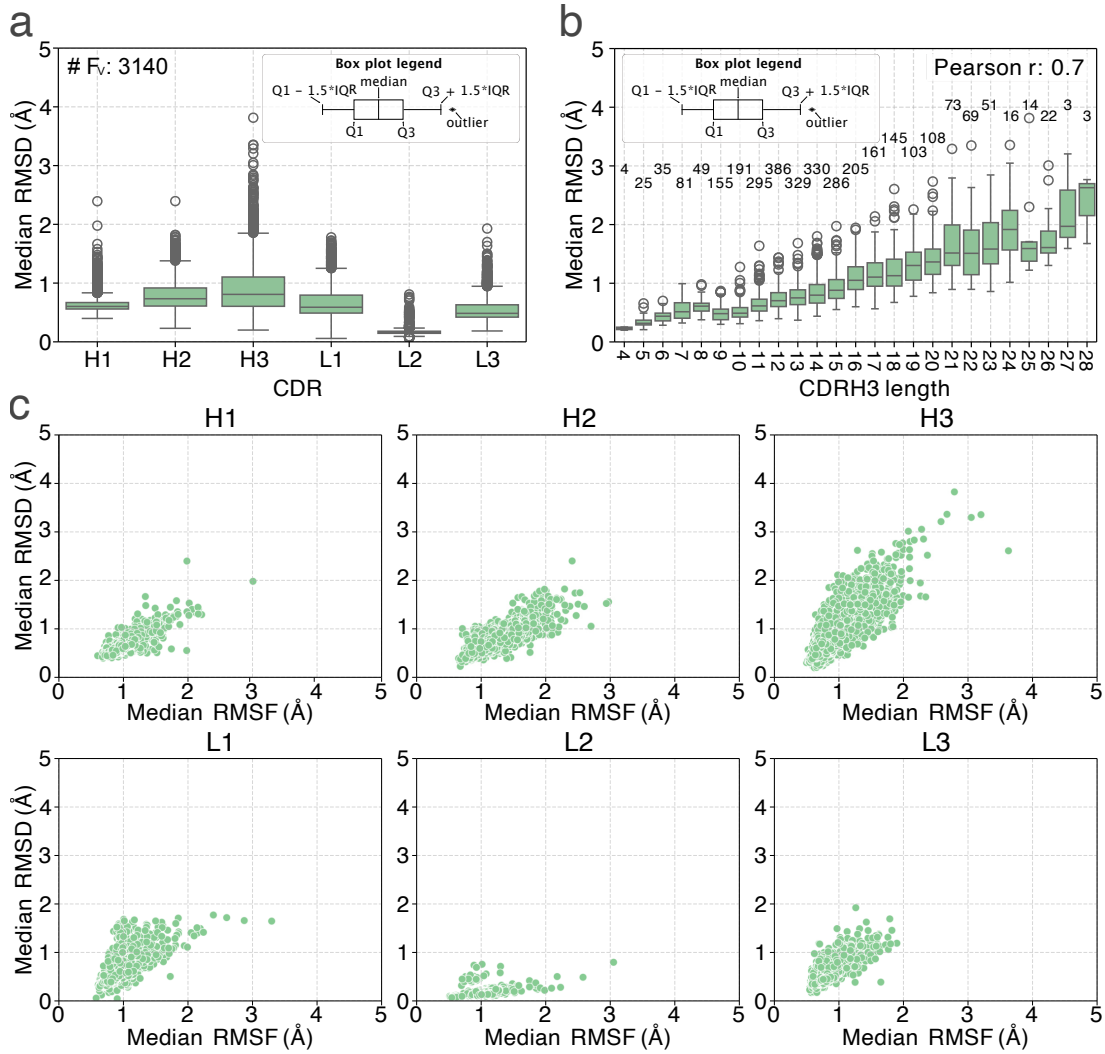


Figure 4.4: CDR dynamics of antibody with experimental structures. CV3-Fv simulations were run for 3140 non-redundant and quality filtered antibody structures from the SAbDab. a) Box plot showing the median RMSD across all frames to the average loop structure ($RMSD_{mean}$) for each of the CDRs. Each box plot shows values of all 3140 samples. b) Box plot showing the median CDRH3 $RMSD_{mean}$ as a function of CDR length. A Pearson correlation of RMSD against CDRH3 length was run and the correlation coefficient indicated. The number of samples is reported on the top of each box plot. c) Scatter plot showing the median $RMSD_{mean}$ of each loop against the median residue-level RMSF.

4.6 Discussion

Structural dynamics of antibodies, particularly in the Fv region and CDR loops, are thought to be essential for their function. This flexibility is tightly linked to binding affinity (Mikolajek et al. 2022; Li et al. 2015; Schmidt et al. 2013; Haidar et al. 2014) and may enable the structural adaptability required to recognise diverse antigens (James et al. 2003; Fernández-Quintero et al. 2019d; Guthmiller et al. 2020). Therefore, flexibility is a key factor in the discovery and design of novel therapeutics. However, the limited availability of data on Fv dynamics makes it difficult to study these effects systematically and even more challenging to train predictive models to address the millions of uncharacterised antibody sequences.

Recent advances in computational power and coarse-grained modelling have made it possible to use MD simulations to generate large datasets and explore protein dynamics on a large scale (Von Bülow et al. 2025). In this chapter, we examined the suitability of CALVADOS 3, a residue-level, coarse-grained model initially designed for multi-domain proteins with folded domains and intrinsically disordered regions, for simulating the dynamics of antibody Fvs.

We found that simulating Fvs by considering the CDR loops to be entirely intrinsically disordered regions led to an overestimation of their flexibility in CALVADOS 3, so we introduced additional simulation restraints within the CDR loops based on the interaction patterns observed in the initial crystal structure. This new setup, called CV3-Fv, generated ensembles that showed greater agreement in terms of flexibility and conformational sampling with all-atom and enhanced sampling simulations. Although flexibility metrics correlate well with the atomistic data, CV3-Fv simulations tend to underestimate distinct conformational states and the barriers between them, particularly in highly flexible loops. Consequently, we sometimes found that the relative populations of macrostates were different for more flexible systems compared to all-atom simulations. This limitation is not surprising, given that coarse-grained models inherently smooth the underlying protein energy landscape (Kmiecik et al. 2016; Jin et al. 2022).

We also investigated the effect of using bound or unbound crystal structures as starting points for the simulations. We found that structures of bound Fv, particularly those of Fvs bound to antigen fragments or peptides, can lead to distinct conformational ensembles, even when flexibility metrics report similar values. This divergence is particularly apparent when the initial structures differ in terms of the backbone torsion angles within the CDRs, which results in different local constraints during the simulation. Consequently, certain regions of the conformational space, especially those near the native unbound state, may not be fully sampled. While this issue could stem from our strategy for introducing constraints into the model, we observed similar behaviour in metadynamics simulations, which also failed to sample the unbound region when initiated from a bound crystal structure. Recent studies (Zhao et al. 2019; Lee et al. 2022) have shown that some CDR loops can undergo substantial conformational changes upon antigen binding. These shifts may result in states separated by high energy barriers from the unbound conformation, which makes it difficult, or even impossible, for simulations such as CV3-Fv to recover the unbound ensemble. While not all complexes exhibit this behaviour, it is important to note that over 90% of antibody structures in SAbDab are in the bound form. This suggests that this limitation may affect a significant proportion of the simulated Fvs.

We used CV3-Fv to create FlAbDab, the first database of molecular dynamics simulations that covers all non-redundant, experimentally resolved antibody Fvs (over 3000 simulations) in addition to a set two orders of magnitude larger of predicted structures (150,000 simulations). We expanded our work to create FTCRDab which comprises MD simulations for all non-redundant, experimentally resolved TCR structures (280 simulations). The release of FlAbDab and FTCRDab opens the possibility for several downstream applications. The databases will allow the study of CDR dynamics and its functional importance in a large set of immune receptors and provide a new data source for training machine learning models.

We performed a preliminary analysis on the subset of antibodies and TCRs with experimentally resolved structures. For the majority of samples, we observed moderate magnitudes of flexibility with a median $\text{RMSD}_{\text{mean}}$ below 1 Å for all

CDRs and only a few outliers showed values higher than 2 Å. A correlation between loop length and flexibility highlights length as a key factor in determining loop dynamics, although previous studies suggest that length alone is not sufficient to explain flexibility (Spoendlin et al. 2025c; Guloglu et al. 2023). Our large datasets may allow the identification of sequence or structural motifs with the potential to more rationally engineer CDR flexibility (Kouba et al. 2025).

While the functional importance of CDR flexibility has been investigated in case studies (Mikolajek et al. 2022; Li et al. 2015; Schmidt et al. 2013; Haidar et al. 2014; James et al. 2003; Fernández-Quintero et al. 2019d; Guthmiller et al. 2020), the extent to which flexibility is exploited in natural immune repertoires to balance between affinity and specificity remains poorly understood. It has been suggested that naive antibodies tend to be more flexible to promote promiscuous, low-affinity binding and thereby increase the effective size of the immune repertoire and that during maturation antibodies become more conformationally restricted (Schmidt et al. 2013; Babor et al. 2009; Li et al. 2015; James et al. 2003; Blackler et al. 2022; Fernández-Quintero et al. 2019c). In contrast, a previous large-scale study has not found any evidence that maturation promotes rigidity (Jeliazkov et al. 2018). Although our simulations reproduced the phenomenon of rigidification in lineages known to exhibit it, we found no clear, universal trend of increased CDR rigidity when comparing naïve and mature antibodies. However, we highlight several caveats in our analysis. Firstly, analysis was performed on SAbDab antibodies which do not present a natural immune repertoire and patterns of flexibility may differ. Secondly, our analysis was limited to statistics across a large number of simulations and, as shown in previous studies, rigidification may well be relevant to individual antibody lineages in a more nuanced manner. The investigation of the functional role of CDR flexibility at the repertoire level using our databases of antibody MD simulations will be an important future direction.

Another application of our databases is their use for training machine-learning models. Despite the recent developments (Lewis et al. 2025; Zhu et al. 2024; Zheng et al. 2024; Lu et al. 2024; Huguet et al. 2024; Jing et al. 2024a; Janson et al. 2025),

accurately predicting protein conformational states remains an open challenge. Although few models have been developed to predict dynamics independently of MD data (Cagiada et al. 2025a; Invernizzi et al. 2025), most existing approaches follow a similar principle: they employ generative structure prediction frameworks that include at least one training phase based on MD-generated ensemble structures. However, despite promising results indicating that conformational dynamics can be learned, the amount of available MD data appears to be at least one of the limiting factors. The largest previously available MD simulation databases are one to two orders of magnitude smaller than the number of solved protein structures and do not contain antibodies or TCRs (Vander Meersche et al. 2024; Mirarchi et al. 2024; Tesei et al. 2024). Together, FlAbDab and FTCRDab offer new possibilities for the systematic exploration of the sequence-structure relationship in antibodies and TCRs, as well as the principles behind adaptive immunity, and facilitate the development of next-generation computational models.

In Chapter 5, I introduce a generative protein structure prediction model to sample antibody conformational ensembles. This model is pre-trained on FlAbDAb simulations. To further improve the sampled energy landscape and overcome some of the biases observed in the CV3-Fv simulations, I performed an additional fine-tuning step on a smaller set of all atom simulations.

5

Sampling conformational ensembles with deep learning

Contents

5.1 Motivation 124
5.2 Contributions 124
5.3 Background 125
5.4 Methods 128
5.4.1 Datasets 128
5.4.2 All-atom MD simulations 131
5.4.3 Metrics for ensemble analysis 133
5.4.4 ABB4 model 136
5.4.5 Inference of benchmarked models 140
5.5 Results 143
5.5.1 Overview of the ABodyBuilder4 models 143
5.5.2 ABodyBuilder4-base serves as a strong base model for structure prediction 144
5.5.3 ABodyBuilder4-STEROIDS accurately samples antibody conformational ensembles 146
5.6 Discussion 157

5.1 Motivation

In Chapter 4, I introduced Calvados 3-Fv, a high throughput MD model for immune receptors, and the two databases FlAbDab and FTCRDab, containing 150,000 simulated trajectories. I also described several biases observed in these simulations. Generative protein structure prediction models can be trained to sample conformational ensembles and promise several advantages. Firstly, they can be even more compute-efficient as they allow directly sampling of equilibrium distributions rather than time-dependent trajectories. Secondly, models can be pre-trained on coarse-grained simulations and fine-tuned on all atom simulations or even experimental data to improve their accuracy.

In this chapter, I introduce the ABB4 series of generative structure prediction models to sample conformational ensembles of antibodies.

5.2 Contributions

I created the datasets and performed most of the model development and all of the training and benchmarking. The ABB4 code base builds on the FrameFlow repository (Yim et al. 2023a; Yim et al. 2024) and has contributions from Odysseas Vavourakis. King Ifashe performed the all-atom MD simulations and wrote the corresponding methods section.

This chapter contains material reproduced from:

Fabian C. Spoendlin, Matteo Cagiada, King Ifashe, Odysseas Vavourakis, and Charlotte M. Deane. "Sampling antibody conformational ensembles with ABodyBuilder4-STEROIDS," Biorxiv (2026).

5.3 Background

As discussed in Chapter 1, conformational changes are fundamental to the function of many proteins (Wei et al. 2016; Teilum et al. 2009) including antibodies (Kroon et al. 2003; Fernández-Quintero et al. 2019b).

Molecular dynamics (MD) simulations are a powerful tool to study conformational changes. However, standard all-atom MD is constrained by huge computational costs. Even with specialised supercomputers (Shaw et al. 2021) or enhanced sampling (Laio et al. 2002) it remains challenging to simulate large number of proteins over long timescales. Coarse-grained simulations provide a more efficient alternative. However, these can suffer from reduced accuracy introduced by the approximation of the underlying force field (Kmiecik et al. 2016; Jin et al. 2022).

Machine learning methods offer the potential of even more efficient sampling of conformational ensembles. Initial uses of ML in the flexibility space focussed on stochastic subsampling of the multiple sequence alignment (MSA) used by AlphaFold2(Jumper et al. 2021) during inference to increase output diversity (del Alamo et al. 2022; Stein et al. 2022; Wayment-Steele et al. 2024; Sala et al. 2023b; Faezov et al. 2023; Sala et al. 2023a; Saldaño et al. 2022; Riccabona et al. 2024). More recent structure predictors employ generative models which are better suited for sampling distributions of conformational states in theory. However, state-of-the-art generative structure predictors such as AlphaFold3 (Abramson et al. 2024) and Boltz-1 (Wohlwend et al. 2024) are trained on crystal structure data only and have been shown to produce a limited diversity of outputs (Abramson et al. 2024; Wohlwend et al. 2024). A range of specialised models trained on MD data predict structures with higher diversity and that better match conformational ensembles (Lewis et al. 2025; Janson et al. 2025; Jing et al. 2024a; Zheng et al. 2024). These methods differ slightly. AlphaFlow (Jing et al. 2024a) and BioEmu (Lewis et al. 2025) are regular structure predictors that map an input sequence to an ensemble of structures, whereas aSAM (Janson et al. 2025) takes both initial structure and sequence as input to produce alternative conformations.

While ML-based conformation predictors have shown promising results in modelling general proteins, no existing method was trained and evaluated on an extensive corpus of antibody data. To address this gap, I present **A**ntibody**B**uilder**4** - **S**tructure predictor **T**uned on **E**nsembles of complementary determining **R**egions **O**bserved **I**n molecular **D**ynamics **S**imulations (ABB4-STEROIDS), a generative structure prediction model that samples conformational states of antibodies in this chapter (Figure 5.1a). ABB4-STEROIDS is a flow matching model pre-trained on 4.2 million representative frames of $\sim 136,000$ coarse-grained antibody MD simulations from the FlAbDab (Cagiada et al. 2025b) and fine-tuned on a new set of 83 all-atom MD simulations, which is released alongside the model. ABB4-STEROIDS achieves state-of-the-art performance at reproducing ensemble metrics observed in the MD simulations and matching the flexibility and conformational space of antibodies observed in experimental data. The model is openly available on Github (github.com/oxpig/ABB4).

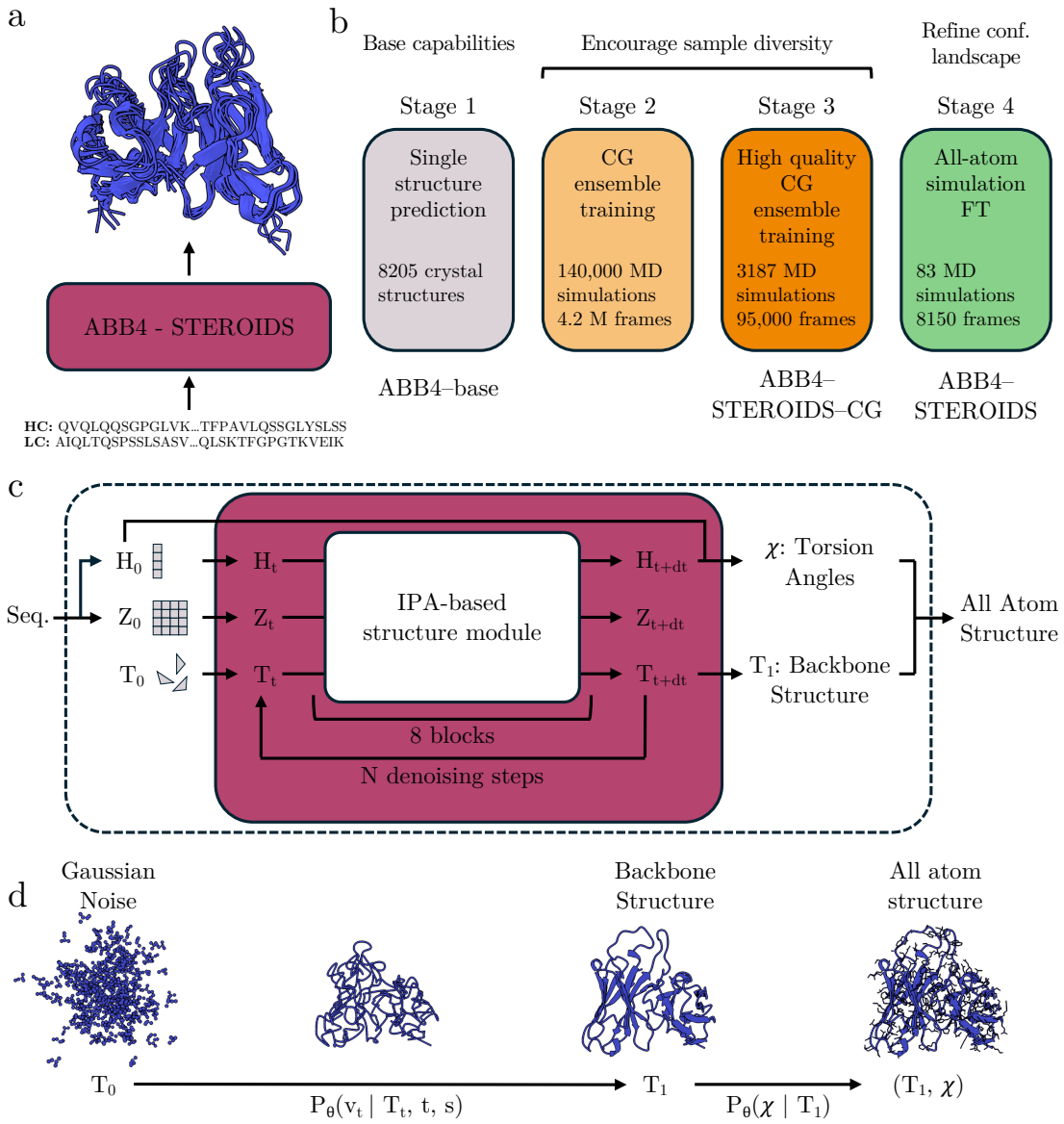


Figure 5.1: Overview of the ABB4 models. a) ABB4-STEROIDS generates an ensemble of antibody structures for a given input sequence. b) Summary of the four-stage training procedure. c) Illustration of the model architecture. d) Diagram providing an overview of the flow matching methodology. H: single/node representation. Z: pair/edge representation. T: backbone frames χ: Torsion angles. Subscripts denote the flow matching time step.

5.4 Methods

5.4.1 Datasets

Single structure prediction data set

A dataset of experimentally solved structures was created for the single structure prediction task. I used the data set released with ABB3 (Kenlay et al. 2024) and supplemented this set with structures released after its cutoff date.

The ABB3 data set contains 9042 antibody structures. This corresponds to all experimentally solved structures before 13/12/2023 filtered by the following criteria: resolution better than 3.5 Å, removal of outliers that fall outside of 3.5 standard deviations from the mean for any of the six summary statistics given by ABangle (Dunbar et al. 2013), ultra-long CDRH3 loops with more than 30 residues and origin from species which occur less than 15 times in the SAbDAb (Dunbar et al. 2014; Schneider et al. 2022).

To supplement this set, I extracted all antibodies from the SAbDab (Dunbar et al. 2014; Schneider et al. 2022) released between 13/12/2023 and 25/03/2025 and applied the same filtering criteria. I further removed any structures with missing residues between IMGT numbered residues 5 and 120. This resulted in an additional 927 antibody structures, increasing the total size of the data set to 9969 structures.

Data splits

For evaluation, I created a test set of 100 antibodies with no overlap to the training sets of any method benchmarked in this paper (ABB3, Boltz-1, AF2, aSAM, AlphaFlow and Bioemu). Antibodies in the single structure prediction data set released after 31/01/2023, which corresponds to the AlphaFlow (Jing et al. 2024a) training cutoff and is the latest training cutoff of all benchmarked methods, were considered to be included in the test set. Structures with more than 80% identity of the concatenated CDR sequence to any antibody released before the 31/01/2023 with resolution below 9 Å were removed. In this step, I only use a 9 Å resolution cutoff as this is the only filter used in the Boltz-1 training set and corresponds to the most lenient quality filtering by any of the benchmarked methods. In addition, I required

test set structures to have a resolution below 2.5 Å and selected the 100 antibodies with the latest release date as my full test set resulting in a test cutoff on 08/11/23. As ABB3 was not trained with date-based data splits its training set overlaps with the full test set. I, therefore, removed 9 antibodies from the full test set resulting in a reduced data set of 91 antibodies for comparison against ABB3. A full list of PDB codes of test set antibodies is provided on GitHub (github.com/oxpig/ABB4).

The remaining antibodies in the single structure prediction data set were clustered at 80% identity of the concatenated CDR sequence using MMseqs2 (Steinegger et al. 2017). 150 antibodies with no overlap to the test set were selected as the validation set and the remaining 8205 antibodies with no overlap to test and validation sets were used for training. A total of 1264 antibody structures remained unassigned.

Coarse-grained MD data set

For ensemble training and evaluation, I created two coarse-grained MD simulation datasets using the Flexible Antibody Database (FlAbDab) (Cagiada et al. 2025b) and the associated CALVADOS 3-Fv method (Cagiada et al. 2025b).

I obtained simulations for all non-redundant antibodies in the single structure dataset from FlAbDab. As the database was built using an earlier data cutoff (29/05/2024) than my datasets I used CALVADOS 3-Fv to simulate as many of the remaining antibodies as possible. Antibodies were assigned the same same data split as in the single structure prediction data set resulting in 100 test, 150 val and 3187 training simulations. This set contains a much smaller number of training structures as the antibodies were filtered for sequence redundancy which was not performed for the single structure prediction data set. I refer to this set as high quality simulations.

As part of FlAbDab, a set of $\sim 150,000$ antibody simulations started from predicted structural models was released. A filtered subset of these simulations was used as an additional low quality data set for model training. Simulations were filtered by 80% concatenated CDR sequence identity to the test and validation sets as well as outliers in global, heavy chain, light chain, CDRH1, CDRH2, CDRH3,

CDRL1, CDRL2 and CDRL3 RMSD and VH-VL distance. An outlier was defined as a deviation by more than 3.5 standard deviations from the mean. This resulted in a dataset of $\sim 136,000$ additional lower quality training simulations.

The coarse-grained MD trajectories in FlAbDab (Cagiada et al. 2025b) contain 1000 frames. As it is computationally expensive to work with this many frames, I extracted a random subset of 30 frames from each trajectory as a representative ensemble. I made this choice as I observed that RMSF and RMSD values start to converge for most simulations when sampling around 20 frames at random (see SI Section D.2.1 for details).

All-atom MD data set

A total of 83 antibodies were simulated using the protocol outlined in the methods below. Before performing any analysis and further data processing of these simulations the first 60 ns of each replica were discarded. I extracted a representative ensemble of 100 frames for each antibody to create the dataset for model training. Specifically, I sampled 25 frames with a constant time interval of 9.6 ns from each replica.

Datasets of experimental CDR ensembles

To benchmark against experimental data, I created datasets of antibody structural ensembles observed in crystal structures with an approach introduced in (Spoendlin et al. 2025c; Cagiada et al. 2025b). I largely follow the methodology described in Cagiada et al. (2025b) but make a small modification to the RMSD calculation.

All antibodies in the SAbDab on May 29th 2024 were collected and filtered for structures solved by X-ray crystallography, a resolution lower than 3.5 Å and no missing residues between IMGT numbered positions 5 and 128. Structures were grouped by Fv sequence identity to define an experimental ensemble. For each experimental ensemble the number of observed conformations was calculated for each of the six CDRs individually. Pairwise Cα RMSD of CDR residues after alignment on Cα atoms of framework residues of the same chain was calculated (note in (Cagiada et al. 2025b) structures were aligned on Cα atoms of CDR residues). A conformation was defined as a cluster resulting from agglomerative clustering with

Table 5.1: Number of data points in the experimental CDR flexibility datasets

Dataset	CDRH1	CDRH2	CDRH3	CDRL1	CDRL2	CDRL3
Total labelled CDRs	133	136	220	137	83	110
Multiple conf.	117	117	205	117	48	87
Single conf.	16	19	15	20	35	23

a complete linkage criterion and an RMSD thresholds of 1.25 Å. This approach means that the RMSD of any two structures within a cluster is below 1.25 Å. This threshold was chosen as it was shown to result in good functional clustering (Spoendlin et al. 2023). Binary labels of flexibility ('flexible', 'rigid') were assigned to each of the CDRs depending on whether the data suggest a CDR has multiple or a single accessible conformation. CDRs with more than one conformation were labelled as 'flexible' and CDRs with a single conformation observed in more than five separate PDB structures were labelled as 'rigid'. The remaining antibodies were left unlabelled. The final number of data points is shown in Table 5.1.

5.4.2 All-atom MD simulations

To run all-atom MD simulations, starting structures were obtained from the PDB (Berman 2000) and each file was cleaned by removing all non-amino-acid records. Structures were processed with PDBFixer (v1.12) to identify and replace non-standard residues, remove heterogens, model missing residues and add missing heavy atoms, including terminal oxygen atoms at chain C-termini. Before adding hydrogens, the OpenMM Topology.createDisulfideBonds method was used to detect disulphide bonds based on Sγ-Sγ proximity so that cysteines were correctly typed as CYX. Hydrogen atoms were added at pH 7.0 using OpenMM Modeller (v8.4.0) (Eastman et al. 2024) and N-terminal and C-terminal patches explicitly applied to the first and last protein residues of each chain.

The protein was parametrised with the AMBER ff19SB force field and the TIP3P-FB water model was used (Tian et al. 2020; Wang et al. 2014). Each system was solvated in a truncated octahedral box with a minimum padding of 1.2 nm from the solute to the box edge and the ionic strength was set to 0.15 M by adding Na^+

and Cl^- ions. Long-range electrostatics were treated using Lennard-Jones particle mesh Ewald with an error tolerance of 5×10^{-4} and a real-space cutoff of 0.9 nm. Hydrogen mass repartitioning was applied by redistributing hydrogen masses to 1.5 amu to enable a 4 fs integration time step. Covalent bonds involving hydrogen were constrained using SHAKE/SETTLE and water molecules were kept rigid. The equations of motion were integrated with the LangevinMiddle integrator using a collision frequency of 1 ps^{-1} and a constraint tolerance of 10^{-6}, and centre-of-mass motion was removed at regular intervals.

Energy minimisation was performed in three stages using a convergence tolerance of 10 kJ/mol/nm throughout: first, all non-solvent, non-hydrogen atoms were restrained with 100 kcal/mol/Å^2 for 5,000 iterations; second, restraints were reduced to 10 kcal/mol/Å^2 for 2,500 iterations; and third, all restraints were removed and minimised for a further 2,500 iterations.

Each system was equilibrated using an eight-stage protocol totalling 7 ns. After minimisation, velocities were assigned from a Maxwell-Boltzmann distribution at 100 K and a 20 ps NVT simulation was run with a 1 fs time step and friction 5 ps^{-1}. The system was then heated from 100 K to 298 K over 1 ns under NVT using a linear temperature ramp divided into 200 blocks, using a 2 fs time step, friction 5 ps^{-1}, and heavy-atom positional restraints of 50 kcal/mol/Å^2. NPT was initiated by introducing a Monte Carlo barostat (1 bar, attempt frequency every 25 steps) and an initial 50 ps NPT settling period was run at 2 fs with friction 5 ps^{-1} and a reduced barostat attempt frequency (every 100 steps). 1 ns of NPT was then run at 4 fs with heavy-atom restraints of 100 kcal/mol/Å^2, followed by 1 ns of NPT with heavy-atom restraints reduced to 10 kcal/mol/Å^2. Next, heavy-atom restraints were removed and backbone (N, Cα, C) restraints of 10 kcal/mol/Å^2 were applied, an intermediate energy minimisation was performed (5,000 iterations), and NPT was continued for 1 ns with backbone restraints of 10 kcal/mol/Å^2. Successive 1 ns NPT segments were then run with backbone restraints reduced to 1 kcal/mol/Å^2 and 0.1 kcal/mol/Å^2, and the protocol was finished with 1 ns of unrestrained NPT. For each NPT equilibration segment, a NaN-recovery protocol was used: if numerical

instabilities occurred, they were rolled back to the most recent checkpoint, a brief stabilisation was performed with a reduced time step and restraints were increased, systems re-minimised, and the segment retried.

Production simulations were run for 300 ns per replica under NPT conditions (298 K, 1 bar) with all restraints removed and a 4 fs time step, giving a total of 1.2 μs of simulation. Trajectory frames were saved every 15,000 steps (60 ps) and thermodynamic observables (potential energy, kinetic energy, total energy, temperature, volume, and density) recorded at the same interval. Checkpoints were written every 25,000 steps to enable restarts and three independent replicas were run for each antibody structure. All simulations were performed using CUDA-accelerated OpenMM (v8.4.0) with mixed-precision arithmetic on NVIDIA H200 GPUs on the Isambard-AI facility HPCs.

Production trajectories were processed using cpptraj in AmberTools24 package (Roe et al. 2013; Case et al. 2023). For each replica, the equilibrated structure was used as the reference topology, trajectories were centred on the protein centre of mass, unwrapped across periodic boundaries, and re-imaged using autoimage. Solvent molecules and ions were stripped to generate protein-only trajectories and frames were aligned to the first frame by root-mean-square fitting of protein heavy atoms. All trajectory processing was performed on an 8-core CPU allocation with 150 GB memory.

5.4.3 Metrics for ensemble analysis

Metrics were calculated for each antibody region separately. IMGT numbering was used for region definitions (Lefranc et al. 2003) and numbering was carried out using the ANARCII software (Greenshields-Watson et al. 2025a).

Evaluation of predicted ensembles against MD data

I used several metrics to compare sampled structures against the MD ensembles. Distance-based metrics recall, precision and distributional accuracy are based on the distance between predicted structures and individual frames from the

MD trajectories. I calculated the pairwise RMSD matrix $C \in \mathbb{R}^{m \times n}$ between the m structures in the sampled ensemble $\{x_i\}_{i=1}^{m}$ and the n representative MD frames $\{y_j\}_{j=1}^{n}$.

$$C_{ij} = \mathrm{RMSD}(x_i, y_j) \tag{5.1}$$

Pairwise RMSD is defined by my standard protocol as the Cα RMSD of a domain after alignment on all Cα atoms of the chain the domain is located on. For 'global' metrics, structures are aligned on Cα atoms of both chains. Recall is defined as the mean minimum distance between all MD frames and any sampled structure.

$$\mathrm{Recall} = \frac{1}{n} \sum_{j=1}^{n} \min_{1 \leq i \leq m} C_{ij} \tag{5.2}$$

Precision is defined as the mean minimum distance between all sampled structures and any MD frame.

$$\mathrm{Precision} = \frac{1}{m} \sum_{i=1}^{m} \min_{1 \leq j \leq n} C_{ij} \tag{5.3}$$

Distributional accuracy is defined as the 1-Wasserstein distance between the distributions of sampled structures and MD frames, using C as the ground cost matrix. Assigning uniform weights $\hat{\mu}_i = \frac{1}{m}$ and $\mu_j = \frac{1}{n}$, the distance is

$$W_1(\hat{\mu}, \mu) = \min_{\pi \in \Pi(\hat{\mu}, \mu)} \sum_{i=1}^{m} \sum_{j=1}^{n} C_{ij}\, \pi_{ij}, \tag{5.4}$$

where $\Pi(\hat{\mu}, \mu)$ is the set of all couplings with marginals $\hat{\mu}$ and μ.

Ensemble diversity metrics measure the diversity within a predicted ensemble and compare the magnitude of diversity to the MD data. I used the mean absolute error (MAE) of ensemble RMSD and the root mean squared error (RMSE) of ensemble RMSF as two such metrics. To calculate MAE RMSD, I computed the pairwise RMSD matrix $R^{sampled} \in \mathbb{R}^{m \times m}$ between the m sampled structures of an antibody and the pairwise RMSD matrix $R^{MD} \in \mathbb{R}^{l \times l}$ between all l frames in the MD trajectories.

$$R_{ij}^{sampled} = \mathrm{RMSD}(x_i, x_j), \qquad R_{ij}^{MD} = \mathrm{RMSD}(l_i, l_j), \tag{5.5}$$

The MAE between the two RMSD matrices over K antibodies is defined as

$$\mathrm{MAE(RMSD)} = \frac{1}{K}\sum_{k=1}^{K}\left|\mathrm{mean}_{i,j}\left(R_{ijk}^{\mathrm{sampled}}\right) - \mathrm{mean}_{i,j}\left(R_{ijk}^{\mathrm{MD}}\right)\right|. \tag{5.6}$$

To calculate RMSE of RMSF, I computed the RMSF of residue r in the sampled ensemble $RMSF_r^{sampled}$ and l frames of the MD trajectory $RMSF_r^{MD}$. The RMSE of RMSF is defined per antibody as

$$\mathrm{RMSE(RMSF)} = \sqrt{\mathrm{mean}_r\left(\mathrm{RMSF}_r^{\mathrm{sampled}} - \mathrm{RMSF}_r^{\mathrm{MD}}\right)^2}. \tag{5.7}$$

I used time-lagged independent component analysis (tICA) to visually compare the predicted ensemble against the all-atom MD simulations. To perform the tICA, the all-atom MD simulations were subsampled at 240 ps intervals resulting in 4000 frames per antibody. tICA was performed on the ϕ and ψ backbone torsion angles of each CDR individually.

Evaluation of predicted ensembles against experimental data

I used two metrics, 'flexibility correlation' and 'conformation coverage', to compare sampled structures against the experimental ensembles.

Flexibility correlation measures the correlation between RMSD in sampled ensembles and flexibility of the experimental ensemble. I calculated the pairwise RMSD matrix $R^{sampled} \in \mathbb{R}^{m\times m}$ between the m sampled structures of an antibody. For each antibody, I summarized its sampled structural variability by the median pairwise RMSD

$$V_{RMSD} = \mathrm{median}\left(\left\{\, R_{ij}^{sampled} : 1 \le i < j \le m \,\right\}\right). \tag{5.8}$$

Flexibility correlation is defined as the point biserial correlation coefficient (r) between V_{RMSD} and the binary flexibility labels of the experimental ensemble, where $Y = 1$ indicates a flexible antibody and $Y = 0$ indicates a rigid antibody. Flexibility correlation was calculated for all antibodies that were labelled as either 'flexible' or 'rigid'.

Conformation coverage measures the coverage of conformations in the experimental ensemble by the sampled structures. I compute the pairwise RMSD matrix

$C \in \mathbb{R}^{m \times n}$ between the m sampled structures of an antibody and the n structures in the experimental ensemble. As a conformational cluster $\mathcal{E}_k = \{y_j : j \in J_k\}$ may contain more than one structure, I first define its mean distance $D \in \mathbb{R}^{m \times k}$ to a sampled structure x_i as

$$D_{ik} = \frac{1}{|J_k|} \sum_{j \in J_k} C_{ij}. \tag{5.9}$$

The conformation coverage score is then defined as the minimum distance to any sampled structure

$$\text{Coverage}_k = \min_{1 \leq i \leq m} D_{ik}. \tag{5.10}$$

Conformation coverage was only calculated for CDRs with multiple conformations.

5.4.4 ABB4 model

ABB4 is a generative protein structure prediction model using flow matching. The generative process, model architecture and training and inference procedure are briefly outlined below and details provided in the SI.

Model formulation

Protein structure is parametrised, following AlphaFold2 (Jumper et al. 2021), by a backbone rigid-body transformation $T \in SE(3)$ together with a set of side chain torsion angles χ for each amino acid residue. I formulated structure prediction as a generative process over backbone frames, modelling a conditional distribution $p_\theta(T_1 \mid T_0, s)$, which interpolates between an initial noisy frame T_0 and the final protein backbone configuration T_1 conditioned on the sequence s. Torsion angles are predicted via $p_\theta(\chi \mid T_1)$, conditioned on the final backbone frame.

Backbone frames $T = (x, r)$ consist of a translation vector $x \in \mathbb{R}^{3 \times N}$ and a rotation matrix $r \in SO(3)^N$. I modelled the generative process using the flow-matching formulation on $SE(3)$ introduced in FrameFlow (Yim et al. 2023a). In brief, translations and rotations are denoised independently through a learned vector field $p_\theta(v_t \mid T_t, t, s)$, where $v = (v_x, v_r)$. In practice, the model is parametrised to

predict clean frames $p_\theta(\hat{x}_1, \hat{r}_1 \mid T_t, t, s)$, from which I construct the corresponding vector field and integrate it to obtain the next sample in the trajectory $\hat{T}_{t+1}$,

$$v_x = \frac{\hat{x}_1 - x_t}{1-t}, \qquad v_r = \frac{\log_{r_t}(\hat{r}_1)}{1-t}, \tag{5.11}$$

where $\log_{r_t}(\hat{r}_1)$ denotes the logarithmic map on $SO(3)$, corresponding to the geodesic direction from r_t to $\hat{r}_1$. For further details on flow matching in $SE(3)$ see FrameFlow (Yim et al. 2023a).

The model is trained with direct losses on the backbone translations $\mathcal{L}_{R(3)}$ and rotations $\mathcal{L}_{SO(3)}$ as in FrameFlow (Yim et al. 2023a) and auxiliary FAPE $\mathcal{L}_{FAPE}$, torsion angle $\mathcal{L}_\chi$ and torsion norm $\mathcal{L}_{\|\chi\|}$ losses on all-atom coordinates and torsion angle predictions as defined in AlphaFold2 (Jumper et al. 2021). Details on individual loss terms are provided in SI Section D.1.1. The total loss is defined as:

$$\mathcal{L}_{total} = 8 \times \mathcal{L}_{R(3)} + 0.8 \times \mathcal{L}_{SO(3)} + 0.08 \times \mathcal{L}_{FAPE} + 0.4 \times \mathcal{L}_\chi + 0.02 \times \mathcal{L}_{\|\chi\|}, \tag{5.12}$$

with weights set to match the magnitude of individual terms after losses plateaued. All of these losses have the objective to reconstruct a single structure.

I additionally implemented a loss that acts across an ensemble of generated structures $\mathcal{L}_{ensemble}$ with the objective to fit the RMSF to the expectation in the MD simulations (for details see SI Section D.1.1). The main ABB4 models do not use the ensemble loss but I evaluated a model variant trained with this loss. For this model the ensemble loss is added to the total loss with a weight of 0.05.

I used a global sequence-conditioned optimal transport strategy for training. In contrast to previous work, which employed unconditional, mini-batch optimal transport (Bose et al. 2024), I implemented a variant tailored to the task of ensemble prediction. For each protein sequence s, I computed a global optimal transport coupling π_s^* over all noise-structure pairs associated with that sequence, and sample from it during training. Accordingly, the following objective is optimised:

$$\mathbb{E}_{(T_0,T_1|s)\sim\pi_s^*}\Big[\mathcal{L}_{total}(T_0,T_1)\Big]. \tag{5.13}$$

This approach preserves the benefits of optimal transport while conditioning on biological sequence identity, which ensures consistency across all conformations of a given protein sequence. For more details see SI Section D.1.1.

Model architecture

The ABB4 model builds on invariant point attention (IPA) first introduced in AlphaFold2 (Jumper et al. 2021) and largely follows the architecture introduced in FrameDiff (Yim et al. 2023b).

The model takes as input: node features H_0, edge features Z_0 and initial backbone frames T_0. The node features are constructed from an amino acid embedding, a positional embedding, a time-step embedding and an antibody region embedding. The antibody region embedding is a 14-dimensional one-hot vector indicating which region (CDRH1, CDRH2, CDRH3, CDRL1, CDRL2, CDRL3, FwH1, FwH2, FwH3, FwH4, FwL1, FwL2, FwL3, FwL4) the residue belongs to. For an antibody of r residues, these are embedded into an initial node feature encoding $H_0 \in \mathbb{R}^{r\times 256}$. Edges encode a distogram of backbone frame translations and a relative positional encoding. These are embedded into an initial edge feature encoding $Z_0 \in \mathbb{R}^{r\times r\times 256}$.

Node features, edge features and backbone frames are updated by 8 layers of a structure module (Figure 5.2). The final layer of the model outputs the backbone rotations and translation as well as side chain torsion angles which are used to reconstruct an all-atom structure. ABB4 has a total of 19.6 million trainable parameters.

Training

I trained three variants of the ABB4 model, which I refer to as ABB4-base, ABB4-STEROIDS-CG, and ABB4-STEROIDS. These differ by the set of training stages applied. In total, I used four distinct stages to train the models. During stage 1, I trained on the single structure prediction dataset for the task of single structure

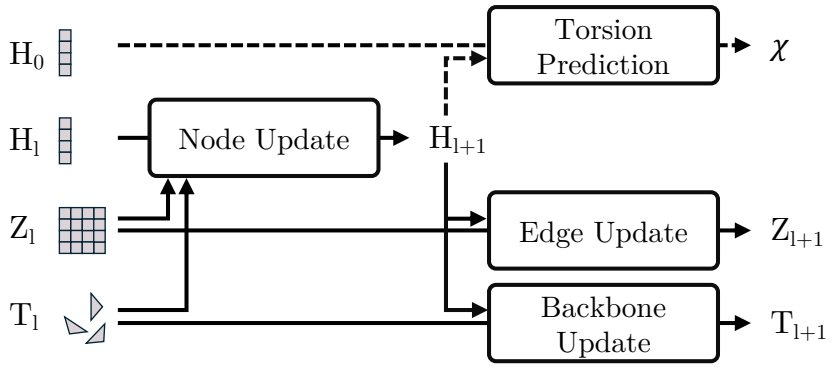


Figure 5.2: Single layer of the ABB4 model. Each layer l has the current node representations H_l, edge representations Z_l and backbone frames T_l and initial node representations H_0. Representations are processed with node, edge and backbone update blocks to prepare the input for layer $l+1$. In the final layer (indicated by the dashed line), torsion angles χ are additionally predicted. Details on the architecture of the four model blocks are shown in Figure D.4.

prediction. I refer to the resulting model as ABB4-base. In the following two stages, the base model was trained on representative ensembles from the coarse-grained MD dataset for the task of ensemble prediction. In stage 2, I trained on all available simulations including the high quality set and the 136,000 lower quality simulations started from predicted models. In stage 3, only the high quality set of simulations was used. I refer to the model trained with stages 1 to 3 as ABB4-STEROIDS-CG. In stage 4, I further fine-tuned the model on a smaller number of all-atom MD simulation and obtained the final ABB4-STEROIDS model. The total dataset sizes used in each stage are shown in Table 5.2.

Models were trained with a batch size of 32 (stage 1) and 40 (stage 2, 3 & 4). The AdamW optimiser was used with weight decay of 0.01. A cosine annealing with warm restarts learning rate schedule was used with an initial learning rate of 1e-4, a minimum learning rate of 1e-5 and 50-epoch interval between restarts.

I used early stopping based on validation set metrics. During stage 1, I performed a mini roll-out with 10 model evaluations to generate one structure for each validation set antibody and compute the RMSD to the crystal structure. I stopped training when the CDRH3 RMSD converged. During stages2, 3 and 4, I rolled out an

Table 5.2: Training dataset sizes

Stage	Dataset	Training		Validation		Test	
		mAbs	Struc.	mAbs	Struc.	mAbs	Struc.
1	Single structure pred.	8205	8205	139	139	100	100
2	CG MD all	136,000	4,180,000	1480	4440	100	3000
3	CG MD high quality	3187	95,000	1480	4440	100	3000
4	All-atom MD	73	7150	4	400	6	600

mAbs: number of unique antibodies, Strucs: total number of crystal structures/MD frames, CG: coarse-grained.

ensemble of 30 structures for each validation set antibody and compute RMSD within the generated ensemble. I stopped model training when the error in RMSD between the generated ensemble compared to the full MD trajectory converged for CDRH1 (stage 2) and CDRH3 (stage 3 & 4). I found this approach using mini roll-outs results in much improved performance over monitoring of step-level validation losses.

Inference

Unless stated otherwise, inference was performed with 100 model evaluations. An ensemble of 100 structures was sampled for each antibody during evaluation.

5.4.5 Inference of benchmarked models

In this section, details on how inference was run with any of the benchmarked models are provided. When evaluating a model for single structure prediction (ABB3, Boltz-1), a single structure was generated for each antibody. For evaluation of ensemble prediction (Boltz-1, BioEmu, AlphaFlow, aSAM, AF2 MSA subsampling), 100 structures were sampled for each antibody.

ABB3, Boltz-1 and AF2 allow for inputs with multiple chains and can, therefore, be used directly to predict antibody structures. In contrast, BioEmu, AlphaFlow and aSAM accept inputs with only a single chain. For these methods, I placed heavy and light chain sequences on a single chain connected by six repeats of a glycine and serine linker ($6 \times GGGGS$).

ABB3

ABB3 inference was performed according to the authors' instructions (Kenlay et al. 2024). Input embeddings were produced using the default T5 pLM model distributed with the package.

Boltz-1

Inference was performed with Boltz-1 weights (Wohlwend et al. 2024) and default parameters. MSAs were generated through Boltz-1 using the MMseqs2. For single structure prediction, I set the number of diffusion samples to 1 and for ensemble prediction to 100.

BioEmu

Backbone structures were sampled from the BioEmu (Lewis et al. 2025) model with default parameters. Side chains were reconstructed using the provided script. MD equilibration was not performed.

AlphaFlow

The AlphaFlow-MD distilled model with default parameters was used for inference (Jing et al. 2024a). The distilled model was chosen over the base model to reduce the computational cost. MSAs were generated using the provided script to query the ColabFold (Mirdita et al. 2022) servers.

aSAM

aSAM (Janson et al. 2025) does not perform structure prediction and requires a starting structure as input. Three structures of each antibody were generated with Boltz-1 and the highest ranked structure was selected as the input structure for aSAM. I used the ATLAS model for inference with default parameters.

AlphaFold2 MSA subsampling

AF2 was run through ColabFold (Mirdita et al. 2022). For MSA subsampling, the max MSA parameter was set to 64:128, the number of recycles to 1 and the number of seeds to 20 (5 models are generated per seed which results in 100 total structures). These parameters have previously been shown to generate antibody ensembles with maximal diversity while limiting the occurrence of unfolded structures (Spoendlin et al. 2025c).

5.5 Results

5.5.1 Overview of the ABodyBuilder4 models

ABB4-STEROIDS is a generative structure prediction model trained to sample conformational ensembles of antibodies (Figure 5.1a). In this section, an overview of the model architecture and training procedure is presented (for details see methods and SI).

The architecture builds on the AF2 (Jumper et al. 2021) structure module (Figure 5.1c). An input sequence gets embedded into a single representation (H), encoding residue level features, and a pair representation (Z), encoding inter-residue features. Protein structures are represented by a frame parametrisation (T) where all backbone atoms of a residue are modelled by a single rotation matrix and translation vector. The three representations are updated by 8 blocks of a structure module using Invariant Point Attention (IPA) at its core and side chain torsion angles (χ) are predicted with an additional head (Figure 5.2).

ABB4 is trained as a flow matching model over backbone frames largely following the FrameFlow (Yim et al. 2023a) methodology (Figure 5.1d). During training, a target structure and a time step are sampled and the corresponding noise applied to the backbone frames to obtain a partially noised structure (T_t). Conditioned on noised backbone frames (T_t), sequence (s) and time step (t) the model is trained to predict the clean backbone (T_1) and torsion angles (χ). During inference, initially pure noise is sampled (T_0). From model predictions a vector field (v_t) is constructed that interpolates between T_t and T_{t+dt} and denoises the samples. After a full model rollout a clean backbone structure is reconstructed from the final frame (T_1). Torsion angles (χ) predicted during the final denoising step are used to build an all-atom structure.

I implemented a novel four stage training strategy to optimise conformation sampling (Figure 5.1b). In stage 1, a model was trained on experimental structures to establish basic single structure prediction capabilities. I refer to the resulting model as ABB4-base which serves as a foundation for subsequent stages. In stages

2 and 3, the base model was pre-trained on a large corpus of coarse-grained MD simulation data deposited in the FlAbDab (Cagiada et al. 2025b). The objective of these stages was to encourage the model to sample diverse conformational ensembles; however, the model may inherit biases of the coarse-grained force field. More specifically, Stage 2 utilised the complete dataset, comprising over 4.2 million structural frames extracted from $\sim 136,000$ antibody simulations, including both high-quality trajectories initiated from experimental structures and lower-quality simulations initiated from predicted structural models. In stage 3, the training was restricted exclusively to the high quality subset. The resulting model from stages 1 to 3 is designated ABB4-STEROIDS-CG. In stage 4, I obtained the final ABB4-STEROIDS model by fine-tuning ABB4-STEROIDS-CG on the smaller set of 83 all-atom simulations. The primary objective of this stage was to mitigate biases introduced through the coarse-grained data and to refine the sampled landscape and relative populations of conformational states.

5.5.2 ABodyBuilder4-base serves as a strong base model for structure prediction

As described above, the ABB4-base model was trained on experimental structures. I benchmarked the model for single structure prediction against state-of-the-art open-source antibody specific and general protein structure prediction models.

I evaluated against ABodyBuilder3 (ABB3) (Kenlay et al. 2024), an antibody structure predictor based on the AF2 architecture and trained as a regression model. Similar to ABB4-base, antibody structures in isolation (agnostic of the antigen) were used for training. Extensive benchmarking showed that ABB3 achieved performance comparable to AF2. I also used Boltz-1, an open-source representative of the most recent generation of AF3-style generative protein structure predictors (Wohlwend et al. 2024), which was trained on all proteins in the PDB including protein complexes. AF3-style models, such as Boltz-1, offer improved accuracy over AF2 in modelling antibody-antigen complexes (Abramson et al. 2024; Passaro et al. 2025). However, there remains less definitive evidence regarding

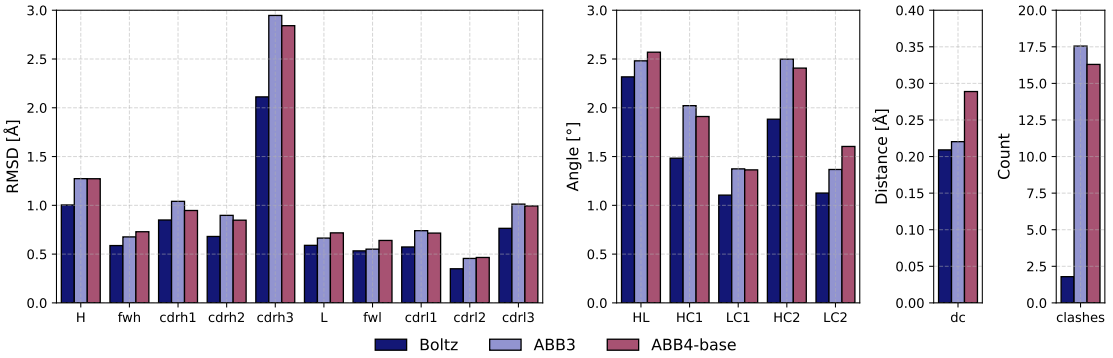


Figure 5.3: Evaluation of ABB4-base single structure prediction against Boltz-1 and ABB3. Single structure prediction was evaluated on a reduced test set of 91 antibody structures. Predictions were assessed using RMSD of antibody regions (H: heavy chain, fwh: heavy chain framework, L: light chain, fwl: light chain framework), deviations in VH-VL orientation angles (HL, HC1, LC1, HC2, LC2) and distance (dc) as defined in (Dunbar et al. 2013) and the number of atomic clashes defined as an overlap in van der Waals radii of any backbone or side chain atoms.

their performance in antigen-agnostic antibody structure prediction (Dreyer et al. 2025; Hitawala et al. 2024).

I compared the performance of the three methods on a reduced test set of 91 antibodies (Figure 5.3). The complete test set of 100 antibodies was curated to exclude any antibody with more than 80% sequence identity, calculated across concatenated CDRs, to structures released before the training cutoff of benchmarked methods, with the exception of ABB3 which did not use a cutoff-based data split (see Methods for details). Consequently, the ABB3 training set overlaps with the full test set; overlapping entries were removed when comparing against this specific method. A comparison of ABB4-base and Boltz-1 on the full test set is provided in Figure D.5.

I evaluated the predictive accuracy of the methods using the RMSD across the different antibody regions, errors in the VH-VL orientations and physical validity as determined by atomic clashes. Across all metrics, I observed a general trend of ABB4-base showing small improvements over ABB3 and Boltz-1 providing a more substantial advancement. These results highlight that while ABB4-base does not achieve state-of-the-art accuracy at single structure prediction, it serves as a reliable base model which is competitive with contemporary structure predictors.

5.5.3 ABodyBuilder4-STEROIDS accurately samples antibody conformational ensembles

In a next step, I trained the ABB4-base model on MD data to obtain ABB4-STEROIDS-CG and ABB4-STEROIDS. I evaluated the models on held-out MD test sets and compared against a set of experimentally observed ensembles.

I benchmarked my models against established structure and ensemble prediction methods. From the former group I included Boltz-1 (Wohlwend et al. 2024). I also evaluated AF2 with MSA subsampling, a workflow that modifies AF2 during inference by reducing the MSA depth, thereby generating higher-diversity samples (e.g. (Wayment-Steele et al. 2024)). From the latter group, I used aSAM (Janson et al. 2025), AlphaFlow (Jing et al. 2024a) and BioEmu (Lewis et al. 2025). These are generative models that were trained on MD data to improve ensemble prediction capabilities. To evaluate ensemble prediction, I generated a total of 100 structures for each antibody with my models and the five alternatives.

Accurate reproduction of MD ensemble metrics

In a first step, I evaluated the ensemble prediction capabilities of the benchmarked methods against the coarse-grained MD data on the full test set of 100 antibodies and against the all-atom MD data on a test set of 6 antibodies. I focused the analysis on ensemble diversity metrics (ensemble RMSD & RMSF) to compare the magnitudes of flexibility in the predicted and MD ensembles. I further report metrics based on the distances between predicted and MD frames (recall, precision and distributional accuracy).

Ensembles generated by the seven methods show large differences in the spread of sampled structures (flexibility) (Figure 5.4a & b and Figure D.7a & b). The ABB4-STEROIDS-CG model generates the most diverse ensembles, whereas the ABB4-STEROIDS model predicts slightly reduced flexibility in CDR1s and CDR3s, with a markedly more pronounced reduction observed in CDR2s. All other methods generally underestimate the flexibility of canonical CDRs. In particular, Boltz-1 ensembles display very low ensemble RMSD and RMSF values. Across all models,

higher diversity is predicted for CDRH3s, and in some cases this even exceeds the flexibility observed in MD simulations. Dissociation of the VH and VL domains is rare, even for methods not explicitly trained on multi-domain proteins (such as AlphaFlow and aSAM), with the exception of BioEmu, where dissociation occurs more frequently (Figure D.6).

I computed errors in ensemble RMSD and RMSF to quantify how well predicted flexibility matches the MD (Figure 5.4c & d and Figure D.7c & d). Overall, Boltz-1 and AF2 MSA subsampling exhibit the largest errors on both the coarse-grained and all-atom test sets, whereas aSAM, AlphaFlow, and BioEmu tend to perform better. ABB4-STEROIDS-CG consistently shows the smallest errors in both metrics across all antibody domains on the coarse-grained test set. On this set, ABB4-STEROIDS achieves similar errors for CDR1s and CDR3s, but larger errors for CDR2. The increased CDR2 errors may stem from a known limitation of the coarse-grained MD force field, which tends to overestimate CDR2 flexibility (Cagiada et al. 2025b). On the all-atom test set, ABB4-STEROIDS shows the lowest errors for the majority of CDRs, including CDRH3, with ABB4-STEROIDS-CG showing comparable performance except for CDRH2. This observation suggests that the brief fine-tuning on all-atom data shifts the CDRH2 distribution substantially toward the all-atom MD reference, mitigating biases learned during coarse-grained training.

I further evaluated a recall metric to provide an indication of how well each method samples the conformational space. On the coarse-grained test set, ABB4-STEROIDS-CG outperforms all other methods on canonical CDRs, while ABB4-STEROIDS also ranks near the top (Figure 5.4d). For CDRH3, all methods exhibit lower recall values, and several outperform the ABB4-STEROIDS models. In the all-atom test set, I observed a similar trend in which ABB4-STEROIDS performs top or near top for canonical CDRs with lower coverage for CDRH3. I point out three issues that should be considered when analysing the recall metric. First, the metric does not penalise excessive conformational entropy. Consequently, models like BioEmu, which predict broad and highly diverse CDRH3 ensembles, achieve high recall. Second, the distances are calculated based on a global frame of reference (CDR

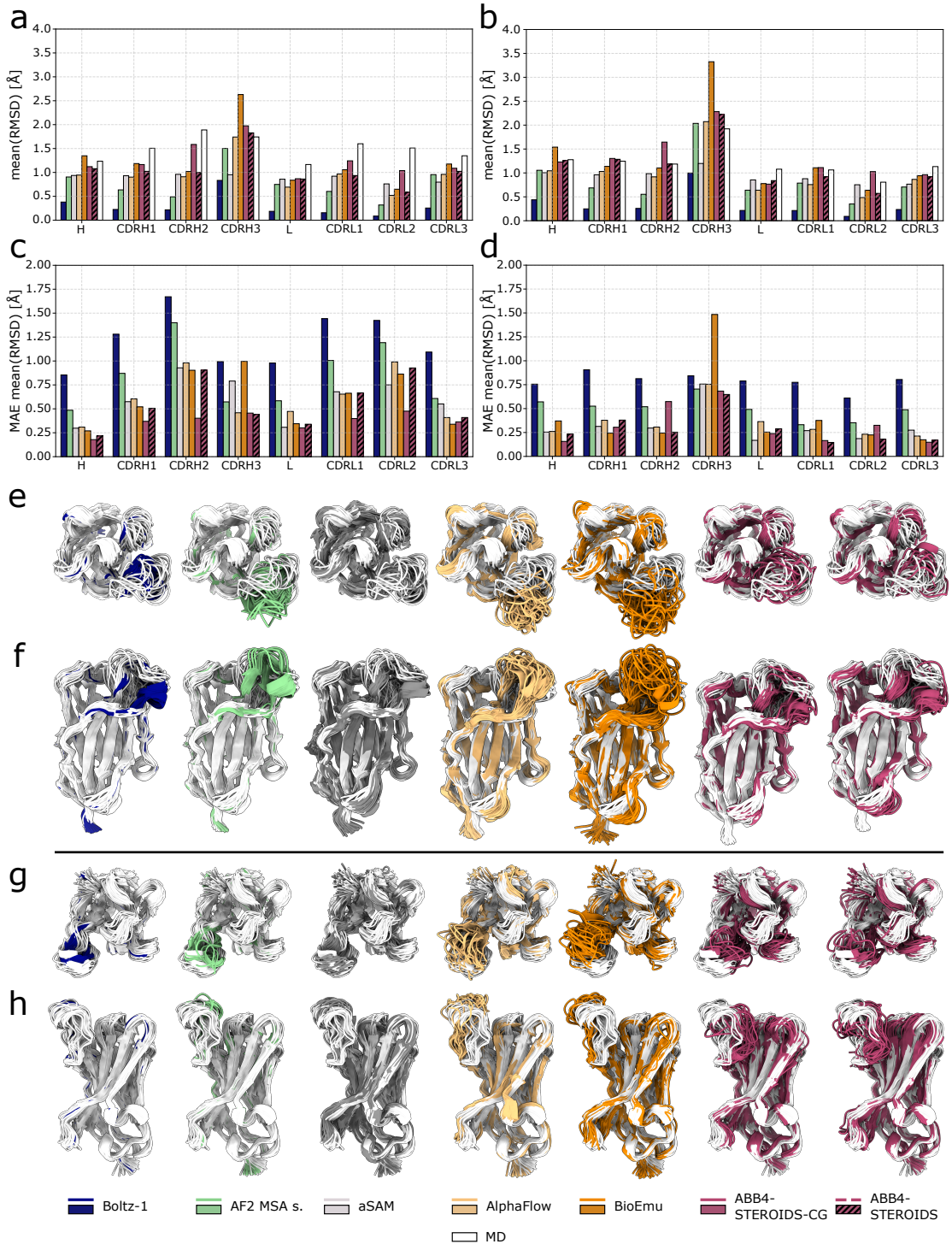


Figure 5.4: Evaluation of predicted structural ensembles against MD data. Mean ensemble RMSD in predictions and MD trajectories (MD) across all antibodies in the coarse-grained MD (a) and all-atom (b) test sets. Values computed across the heavy chain (H), light chain (L) and individual CDRs. Mean absolute error (MAE) of ensemble RMSD in the predictions of the coarse-grained (c) and all-atom (d) test sets. Top (f) and side (e) view of predicted ensembles for antibody PDB 9DS1 overlaid with coarse-grained MD frames. Only antibody heavy chains are shown, light chains were removed for visualisation. Top (g) and side (h) view of predicted ensembles for antibody PDB 8VVB overlaid with all-atom MD frames.

RMSD after alignment on framework region) which emphasises inaccuracies towards the bottom of the loop. Third, the metric is agnostic to the temporal dynamics of the MD trajectories. High-frequency structural fluctuations are weighted equally to the slower transitions that typically characterize functional states.

For completeness, I computed metrics of precision and distributional accuracy on the coarse-grained test set (Figure D.8). However, I note weaknesses in using these metrics to evaluate ensemble prediction. Precision is defined as the minimum distance from each structure of the predicted ensemble to any MD frame and does not provide any information on how accurately an ensemble is reproduced. Distributional accuracy indicates the Wasserstein distance between the predicted and MD distributions. Both metrics are dominated by Boltz-1 and AF2 MSA which are highly accurate structure predictors but generate ensembles of low flexibility which do not match the simulations.

Overall, these results indicate that the ABB4-STEROIDS models accurately reproduce the diversity of structural ensembles observed in MD simulations, as evidenced by ensemble RMSD and RMSF metrics. As expected, ABB4-STEROIDS-CG demonstrates superior performance on the coarse-grained test set, while the ABB4-STEROIDS model exhibits higher accuracy on the all-atom test set. The large size of the coarse-grained set provides a more robust estimate of performance, albeit with the caveat that the underlying conformational distributions may inherit known biases, particularly for CDRH2 (Cagiada et al. 2025b). In contrast, the all-atom set is more reliable but necessitates a cautious interpretation of results due to the limited size. Finally, the recall metric indicates high performance for canonical CDRs, although I acknowledge certain methodological limitations and provide a more detailed analysis of the sampled conformational space in the next section.

Improved sampling of conformations and physical validity through all-atom MD fine-tuning

The coarse-grained MD data used to pre-train my models has inherent limitations. First, the CALVADOS 3-Fv parametrisation that was used to generate these data was optimised to reproduce all-atom ensemble averages. Consequently, certain

conformational states may be under-sampled or their conformational distributions may differ slightly from those of more accurate all-atom simulations (Cagiada et al. 2025b; Kmiecik et al. 2016; Jin et al. 2022). Second, the force field models each residue as one bead. Coordinates of individual atoms are reconstructed with an additional model which may introduce interatomic clashes (Cagiada et al. 2025b).

To prevent the final ABB4-STEROIDS model from inheriting these biases, I introduced stage 4 in the training procedure where I fine-tuned the model pre-trained on extensive coarse-grained MD data with a small set of all-atom simulations. In this section, I highlight that this stage enhances the sampling of conformational distributions, allowing ABB4-STEROIDS to outperform all other benchmarked methods in reproducing the all-atom MD conformational landscape, and improves the physical validity of samples.

To compare the distribution of sampled conformational space with the all-atom MD simulations, I performed time-lagged independent component analysis (tICA). tICA reports on the slowest structural transitions in the simulations (Pérez-Hernández et al. 2013) which tend to be biologically the most meaningful and operates in a local frame of reference. This approach overcomes some of the limitations I highlighted for the recall metric used in the previous section.

While both ABB4-STEROIDS-CG and ABB4-STEROIDS generate ensembles with accurate diversity, a direct comparison of the conformational space they sample highlights improvements as a result of fine-tuning on all-atom simulations. An improved conformational landscape is evident across several case studies of test set antibody CDRs shown in Figure 5.5a. For the CDRH3 of PDB 8r1d, the coarse-grained model predicts structures largely outside the MD ensemble. In contrast, the fine-tuned version shifts a significantly larger fraction of its samples into the MD-occupied space. For the CDRH3 of PDB 8w83, although ABB4-STEROIDS-CG samples within the MD space, it over-represents a rare state on the left of the tICA plot. ABB4-STEROIDS corrects this by prioritising the more populated conformation. In the case of CDRL1 of PDB 8f2t, I observed that the fine-tuned model ceases to sample the rare states occupied by the coarse-grained

model, but more accurately reflects the dominant MD conformation. Finally, for the CDRH2 of PDB 8f2t, both models stay within the MD landscape. However, ABB4-STEROIDS-CG frequently samples rare conformations, while ABB4-STEROIDS better matches the relative populations. In all cases, the ensembles generated by ABB4-STEROIDS more closely mirror the relative populations and equilibrium densities observed in the MD simulations.

A comparison of predictions across all seven benchmarked models suggests that ABB4-STEROIDS provides the closest match to the all-atom MD distributions in tICA space. The conformational landscape sampled by several case studies antibodies is shown in Figure 5.5c. Identical plots for all six test set antibodies are shown in Figures D.9-D.14. For the CDRH3 of PDB 8f2t all-atom MD reveals four distinct states. Boltz-1, aSAM, and AlphaFlow primarily sample a single state. While AF2 MSA subsampling and BioEmu produce greater diversity, many of their samples fall outside the MD landscape. In contrast, ABB4-STEROIDS is the only model that effectively populates two of the dominant conformations with high probability. For CDRH3 of PDB 8oni and CDRH3 of PDB 8r1d most benchmarked methods predict ensembles largely disjoint from the MD distribution. ABB4-STEROIDS consistently recovers the largest fraction of the ground-truth conformational space. Notably, for PDB 8r1d, both ABB4-STEROIDS and AlphaFlow successfully place a significant portion of their samples within the MD ensemble; for PDB 8oni, only ABB4-STEROIDS does so. Finally, for the CDRH1 of PDB 8vvb, whereas most models are restricted to a single conformational state, ABB4-STEROIDS samples a secondary state represented on the right of the tICA plot.

I further assessed the physical validity of the model predictions by quantifying inter-atomic clashes. A comparison between the two ABB4 variants reveals that all-atom fine-tuning reduces the number of clashes in ABB4-STEROIDS relative to ABB4-STEROIDS-CG. In the broader benchmark, ABB4-STEROIDS exhibits a clash profile similar to BioEmu, while AlphaFlow displays the highest frequency of steric violations. As expected, the static structure prediction tools, Boltz-1 and AF2 MSA subsampling, produce the smallest number of clashes. Among the

ensemble-based methods, aSAM is the only one to achieve a clash count similar to that of static structure predictors. However, it should be noted that aSAM requires an input structure, for which I used a Boltz-1 prediction, and I did not investigate how the quality of this initial template influences the resulting physical validity.

Consistency of predictions with experimental evidence of flexibility

After extensive evaluation against MD ensembles, I compared predictions with experimental data. To this end, I collected and compiled structural ensembles of antibodies derived from multiple experimental structures.

Although comparison with these ensembles provides additional validation of my model, their scope may be limited, as experimental structures do not necessarily resolve all physiologically relevant conformational states. Experimental ensembles also do not represent temporal trajectories, which restricts the applicability of the methods I have previously used, such as tICA analysis. I, therefore, focussed on assessing the overall predicted flexibility and coverage of the conformational states found in experimental ensembles using a set of metrics introduced in previous works (Spoendlin et al. 2025c; Cagiada et al. 2025b) .

I correlated the flexibility in predictions with the experimental data. As calculating average RMSD values across the incomplete experimental ensembles is inaccurate, I used the following approach. I collected all antibodies in the SAbDab (Dunbar et al. 2014; Schneider et al. 2022) for which multiple experimental structures are available and assigned binary labels of 'rigid' and 'flexible' depending on whether the data shows evidence for a single or multiple accessible conformations (see methods for details). I measured the agreement between the RMSD in predicted ensembles and the binary labels of the experimental data using a point-biserial correlation (Figure 5.6a). ABB4 models achieve the highest correlation of all methods on all CDRs. Again I observed a big improvement for CDRH2 of ABB4-STEROIDS over the ABB4-STEROIDS-CG model (Cagiada et al. 2025b).

I further assessed the coverage of experimental conformations by the predictions. I used all antibodies in the SAbDab with structures showing multiple conformations

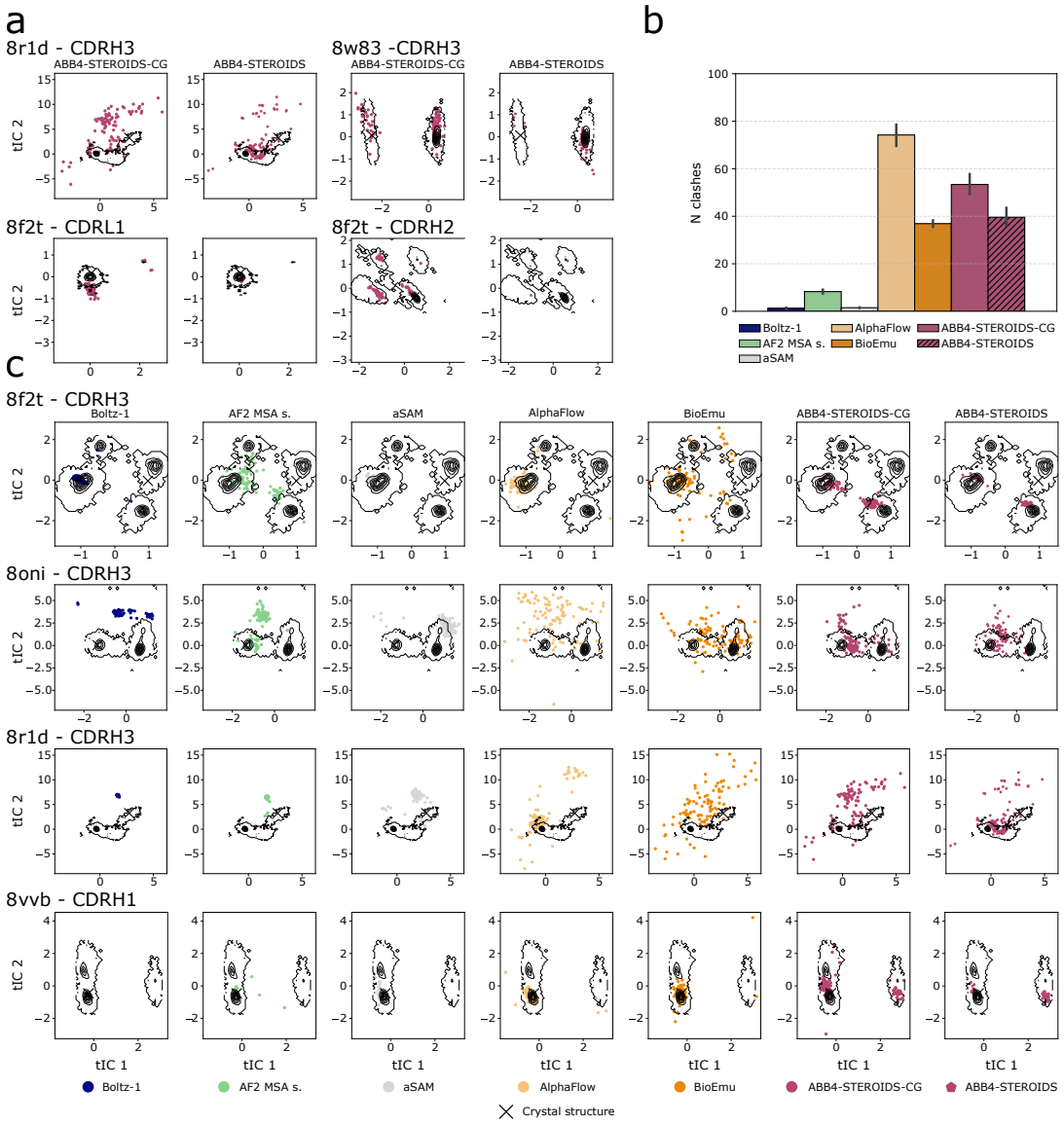


Figure 5.5: Sampling of conformational states and physical validity of predictions. a) Comparison of the conformational ensemble sampled by ABB4-STEROIDS-CG and ABB4-STEROIDS using tICA. b) Average number of atomic clashes in predicted structural models of the test set. c) Comparison of the conformational ensemble sampled by all seven benchmarked methods. a & c) The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all-atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross.

('flexible' antibodies in the previous paragraph). I calculated the coverage of experimental conformations by the predictions and plotted the fraction of conformation against their coverage (Figure 5.6b). Boltz-1 effectively covers a small fraction of conformations with very low RMSD (even below the experimental error margins) showing that it efficiently reproduces some of the crystal structures. However, its limited diversity prevents it from capturing the broader ensemble. Consequently, aSAM, AlphaFlow, BioEmu and ABB4 models outperform Boltz-1 at covering larger fractions of conformations. Among these, the ABB4 models consistently define the Pareto front, demonstrating the best balance between structural accuracy and diversity. More specifically, ABB4 models cover the highest number of conformations for canonical loops at 2 Å and for CDRH3 at 3 Å.

Ablation studies and model variants

Following the evaluation of the main ABB4-STEROIDS and ABB4-STEROIDS-CG models, I assessed several model ablations and variants. To ensure results are robust, I conducted these evaluations on the larger coarse-grained MD rather than the small all-atom test set. All performance metrics are benchmarked against the ABB4-STEROIDS-CG model as the primary baseline.

I first examined the impact of training stage 2, which incorporates a large volume of potentially lower quality coarse-grained MD simulations initialized from predicted structural models. In Figure D.15, I compared ABB4-STEROIDS-CG against a variant trained exclusively with stages 1 and 3 (ABB4-STEROIDS-1,3). The ablated model performs slightly better in validations against the MD data. I attribute this lack of improvement in the full model, despite a 50-fold increase in training data, to subtle differences in the patterns of CDR dynamics in the simulations initialised from structural models compared to those from experimental structures (see SI Section D.2.2). However, including stage 2 leads to clear benefits against the experimental benchmark (Figure D.15). This suggests that the larger, more diverse dataset may prevent the model from overfitting to specific coarse-grained MD distributions and enhance generalization to experimental ensembles.

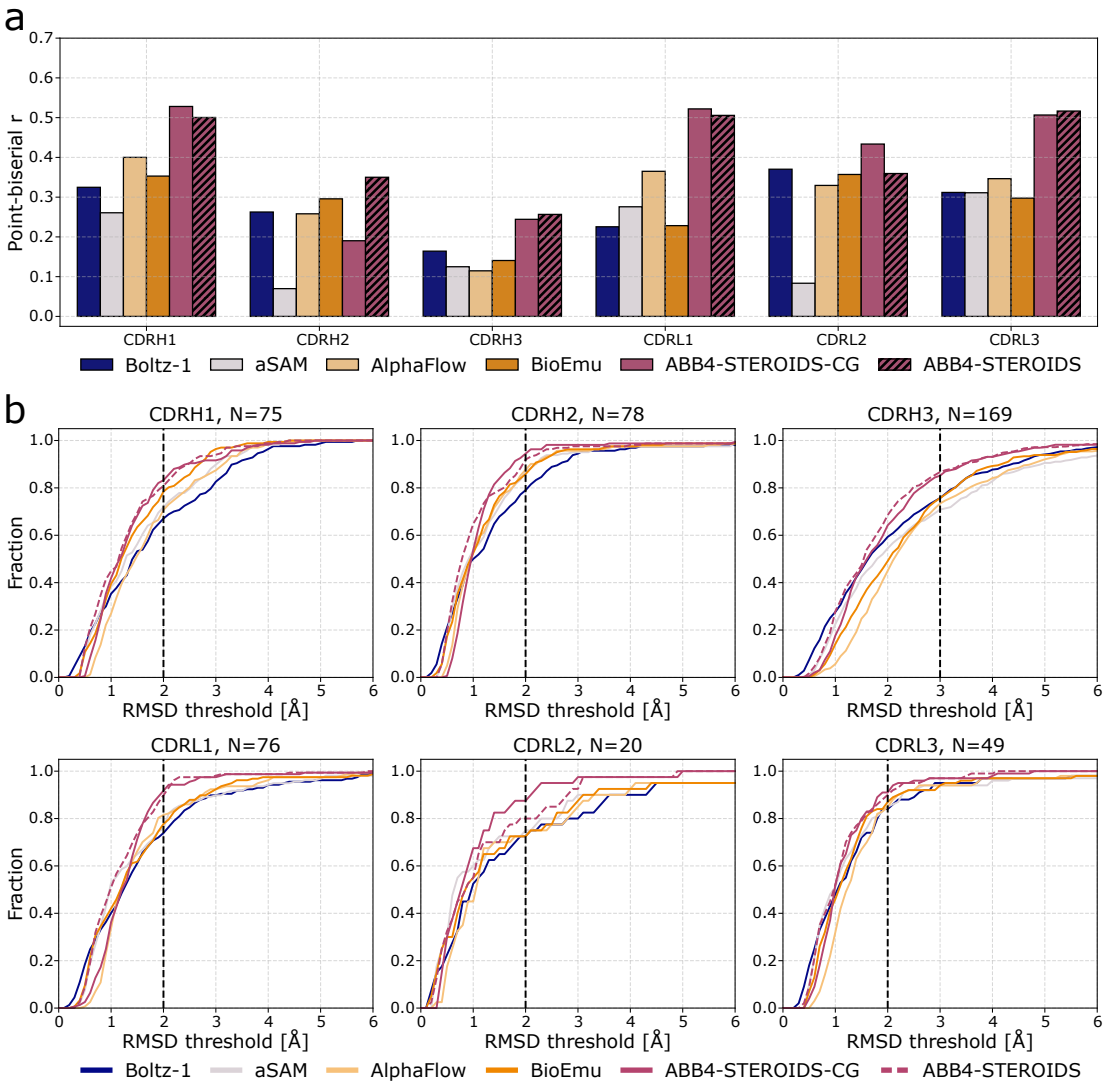


Figure 5.6: Consistency of predictions with experimental evidence of flexibility. a) Flexibility correlation. Point-biserial correlation coefficient (r) between binary flexibility labels assigned to experimental ensemble and RMSD in predicted ensembles for all CDR loops. b) Conformation coverage. Fraction of conformations observed in experimental ensembles covered as a function of distance. Dashed lines highlight ABB4-STEROIDS/-CG covering the highest number of conformations at given distance thresholds.

I further examined the relationship between inference steps and model performance (Figure D.16). While the default setup uses 100 steps, I found that reducing this to 50 steps has a negligible impact on performance. A slight decrease in accuracy is observed at 20 and 10 steps, followed by a loss of performance when the number of inference steps is reduced to five. If computational efficiency is a priority, the inference process can be shortened to 50 steps. Further reductions are possible, though they incur a cost to model performance.

Finally, I explored a model variant incorporating an additional loss function targeted at ensemble properties. Models for conformational sampling are typically trained using standard structure prediction losses applied to individual frames. While training frames are sampled from MD trajectories to capture dynamics, the objective is computed on a single predicted structure at a time. To better align the training objective with the task of ensemble prediction, I introduced an ensemble-based loss function designed to minimize the error in residue-level RMSF (see SI Section D.1.1). I observed some training instability when the difference between predicted and ground-truth RMSF was high, though this was partially mitigated by employing a χ^2-style loss instead of a standard Mean Squared Error (MSE). The additional loss improved performance in ensemble metrics for canonical CDRs (Figure D.17). However, performance for CDRH3 was slightly reduced, likely due to the inherent difficulty in predicting the RMSF of this loop, which increased the training instabilities associated with this loss. These results suggest that while the ensemble loss is effective for domains where RMSF can be predicted with reasonable accuracy, further refinement is needed for regions with higher uncertainties like CDRH3.

5.6 Discussion

Predicting a single, static structure is now possible at high accuracy for most globular and monomeric proteins. However, it remains more difficult to predict the multiple conformational states of proteins. Several approaches have been explored to predict these conformational ensembles. One set of methods has focussed on inference-time modification of common structure predictors such as MSA subsampling (Wayment-Steele et al. 2024) or guidance and steering techniques (Fadini et al. 2026; Richman et al. 2025). Alternative approaches have trained specialised ensemble prediction models using MD data (Janson et al. 2025; Jing et al. 2024a; Lewis et al. 2025). In this chapter, I introduced ABB4-STEROIDS, a generative structure prediction tool that samples conformational ensembles of antibodies with high accuracy.

In a first step, I trained ABB4-base as a standard antibody structure predictor. Similar to previous ABodyBuilder models (ABB2, ABB3) (Kenlay et al. 2024; Abanades et al. 2023), I exclusively trained on antibody structures (removing antigens when present), which is in contrast to recent protein structure predictors such as Boltz-1 that include antibody-antigen complexes in their training data (Wohlwend et al. 2024). The generative framework of ABB4-base resulted in small improvements over ABB3 (Kenlay et al. 2024). However, Boltz-1 achieves a higher overall accuracy, a finding consistent with a previous study indicating that AF3-style models may outperform AF2-style models for antigen-agnostic antibody structure prediction (Hitawala et al. 2024). Major improvements have not been possible across multiple generations of models trained on antibody structures in the absence of antigens (ABB2, ABB3, ABB4-base, AF2) (Jumper et al. 2021; Abanades et al. 2023; Kenlay et al. 2024). I, therefore, speculate that training on complexes may be informative for predicting antibody structures, especially considering that most antibodies in the PDB are resolved with their antigen. Nevertheless, ABB4-base serves as an accurate base model for further training.

I then trained the base models on a large dataset of antibody specific coarse-grained MD simulations to obtain ABB4-STEROIDS-CG. For these training stages, I used simulations deposited in the FlAbDab (Cagiada et al. 2025b) from which

I extracted over 4.2 million representative trajectory frames capturing $\sim 136,000$ antibodies. In a final training stage, the model was fine-tuned on a new set of 83 antibody all-atom MD simulations to obtain ABB4-STEROIDS.

ABB4-STEROIDS-CG outperforms all alternative methods I assessed (Boltz-1 (Wohlwend et al. 2024), AF2 MSA subsampling (del Alamo et al. 2022; Wayment-Steele et al. 2024; Stein et al. 2022) , aSAM (Janson et al. 2025), AlphaFlow (Jing et al. 2024a), BioEmu (Lewis et al. 2025)) at reproducing CDR flexibility evaluated on a test set of 100 coarse-grained MD trajectories. Performance gains are most pronounced for canonical CDRs where alternative methods substantially underestimate flexibility. Improvements on CDRH3 are smaller, which I link to the magnitude of CDRH3 flexibility being on a similar scale to current prediction accuracy. This increases output diversity across all models and makes accurate prediction of CDRH3 conformational states particularly challenging. The fine-tuned ABB4-STEROIDS maintains good performance for the majority of CDRs on the coarse-grained set. A drop in performance is observed for CDR2s which traces back to slight differences in patterns of CDR2 flexibility modelled by the coarse-grained modelling protocol (CALVADOS 3-Fv) in comparison to all-atom simulations (Cagiada et al. 2025b).

I further benchmarked the models on a smaller test set of all-atom simulations comprising six antibodies. While these results should be interpreted with caution given the limited sample size, I highlight two observations. First, the ABB4-STEROIDS model maintains robust performance on evaluations on the all-atom test set, demonstrating that its accuracy is not confined to coarse-grained distributions. Second, ABB4-STEROIDS shows significantly improved performance over ABB4-STEROIDS-CG for CDRH2, suggesting that all-atom MD fine-tuning mitigates biases in CDRH2 modelling that are introduced by the coarse-grained potential of CALVADOS 3-Fv (Cagiada et al. 2025b).

Other than CDRH2 flexibility, the coarse-grained MD data exhibits two further inherent limitations. First, coarse-grained models are known to smooth the protein energy landscape (Kmiecik et al. 2016; Jin et al. 2022), which can lead to an

imprecise modelling of distributions of conformational states. Second, individual atomic coordinates within a residue are not modelled explicitly and need to be reconstructed which can result in interatomic clashes (Cagiada et al. 2025b). The introduction of a short fine-tuning stage on a small volume of all-atom simulations appears to mitigate both of these issues. ABB4-STEROIDS shows both superior sampling of conformational distributions and fewer steric violations compared to the untuned ABB4-STEROIDS-CG model.

Comparing the conformational landscape sampled by ABB4-STEROIDS against the other benchmarked models reveals that my model most accurately reflects the conformational landscape observed in all-atom MD simulations. However, I also show that specific conformational states can be missed and relative population densities may not be accurate. These findings suggest that sampling conformational ensembles with structure prediction models remains a challenging task.

As MD may have its own inaccuracies and biases I further evaluated predictions against experimental data which serves as the most critical benchmark. Following a previously established protocol (Spoendlin et al. 2025c), I curated a dataset from SAbDab (Dunbar et al. 2014; Schneider et al. 2022) consisting of antibodies with multiple resolved structures. Initially, I categorized these antibodies using binary labels to distinguish between those with single versus multiple accessible conformations. Correlating these labels with the predicted ensemble diversity showed that the ABB4 models most accurately mirror experimental observations of flexibility. In a secondary analysis, I quantified coverage of experimentally observed conformational states by each of the methods. While Boltz-1 recovers a small subset of states with the highest accuracy, the ABB4 models consistently sample larger fractions of conformations at the lowest distance threshold. This suggests that while conventional structure prediction models excel at high-resolution predictions of a limited number of states, ABB4 models provide a more comprehensive representation of the accessible structural landscape.

In an ablation study, I showed that training on a large amount of potentially lower quality data prevented overfitting to the MD distributions. The FlAbDab

(Cagiada et al. 2025b) contains $\sim$ 3000 high-quality CALVADOS 3-Fv simulations started from experimental structures and a much larger set of $\sim 150,000$ simulations started from predicted structural models. Including this second set in the training procedure enhanced generalisation to experimental ensembles.

Furthermore, I evaluated a novel auxiliary loss function specifically targeted at ensemble properties. Generally, structure prediction models, even those designed for conformational sampling, rely on loss functions applied to individual frames. While BioEmu (Lewis et al. 2025) recently introduced a loss calculated over multiple predictions to regress folded versus unfolded proportions, I propose a new formulation that minimizes the error in residue-level RMSF. Unlike the BioEmu ensemble-based loss, which requires model rollouts during training and thus incurs significant computational overhead, this loss is integrated into the regular loss computation introducing no additional cost. This approach led to substantial improvements in the predictions for canonical CDRs. However, CDRH3 did not experience a similar performance boost, a result I attribute to training instabilities caused by high RMSF uncertainty in this region. Due to a slight worsening in CDRH3 predictions, I decided not to include the loss in the final model. Nonetheless, these findings suggest that an ensemble-targeted loss may be an effective strategy for optimizing structural distributions.

A limitation of ABB4-STEROIDS is that I exclusively trained on simulations of unbound antibodies. I expect predictions to correspond to the ensemble of the unbound state. As it has been shown that bound conformations are commonly contained in the unbound ensemble (Liu et al. 2024), ABB4-STEROIDS should be able to sample bound conformations. This argument is consistent with my observations of good coverage of experimental conformations, the majority of which are resolved in antigen-bound form. Antigen binding generally tends to rigidify CDRs (Fernández-Quintero et al. 2019b; Kroon et al. 2003), therefore, I do not expect predictions to match the magnitude of flexibility observed in the bound state. Despite this limitation, the unbound antibody ensemble remains highly relevant for drug design as it influences critical properties such as affinity and specificity.

Furthermore, antibodies are administered in the unbound state and properties such as viscosity and stability must be optimised for this state (Greenshields-Watson et al. 2025b). The prediction of conformational changes induced by antigen binding has been studied (Dreyer et al. 2025). However, predicting conformational ensembles of the bound state remains an important direction for future work, as it would provide deeper insight into the modulation of functional properties by structural flexibility. For example, the balance of conformational flexibility in both the bound and unbound states has been shown to influence binding affinity (Mikolajek et al. 2022).

In summary, ABB4-STEROIDS is a deep learning model capable of sampling antibody conformational ensembles with high accuracy. The model was trained on a large volume of coarse-grained simulations of antibodies and fine-tuned on a smaller set of all-atom simulations created as part of this study. ABB4-STEROIDS and the dataset of all-atom antibody simulations are released open-source. These resources enable improved and large-scale investigations of antibody flexibility with the potential to accelerate rational design of flexibility-modulated properties in therapeutic development.

In Chapter 6, I present the main conclusions derived from the research conducted in this thesis and outline prospective areas for future work.

6

Conclusions and future directions

Contents

6.1 Conclusions 164
6.2 Future work 165

6.1 Conclusions

This DPhil thesis details contributions toward methods for the functional characterisation of antibodies using structural and flexibility data. In Chapter 2, I introduce SPACE2, a novel epitope profiling tool. Rather than assuming sequence similarity is required for functional equivalence, I demonstrated that structural similarity provides a powerful orthogonal approach. Comparison at the structural level enables the identification of a broader range of antibodies against a target antigen. Identification of a sequence-diverse set of antibodies in this way is of high interest in drug development due to their distinct developability properties.

This initial project identified conformational dynamics as a primary limitation in protein structure prediction at the time. Consequently, the remainder of this thesis focuses on the accurate, high-throughput prediction of antibody dynamics. While dynamics have been linked to key functional properties such as affinity and specificity in individual case studies (Fernández-Quintero et al. 2019c; Manivel et al. 2000; Krishnan et al. 2008; James et al. 2003; Jeliazkov et al. 2018; Mikolajek et al. 2022), their importance at the repertoire level remains unclear. Prior to this work, reliable modelling of these dynamics was restricted to computationally expensive all-atom MD simulations.

In Chapter 3, I present ITsFlexible, an initial approach to predict the flexibility of antibody CDRs. Rather than modelling conformational states directly, this method addresses a problem of reduced complexity. ITsFlexible classifies CDRs as 'flexible' (multiple accessible conformations) or 'rigid' (a single stable state). For this work, I developed a strategy to mine protein structure databases for evidence of protein loop flexibility which I then used to train the ITsFlexible model. Upon release, ITsFlexible represented the most reliable method for predicting CDR flexibility.

Chapters 4 and 5 focus on methods that directly predict structures of conformational states. Chapter 4 describes Calvados 3-Fv (Cv3-Fv), a coarse-grained (CG) MD workflow that accurately models CDR dynamics at a low computational cost. Cv3-Fv builds on the Calvados 3 (Cv3) force-field, a one-bead-per-residue model originally built for multi-domain proteins with intrinsically disordered regions (IDRs).

Through the introduction and tuning of additional harmonic constraints within the Cv3 force-field, an accurate model for CDRs was obtained. Cv3-Fv reproduces ensemble properties of all-atom simulations well, albeit with some limitations in the exact energy landscapes sampled. Using Cv3-Fv, I generated FlAbDab and FTCRDab, two databases which comprise over 150,000 simulated antibodies and TCRs. At the time of publication, these were the largest standardised datasets of protein simulations available, providing a foundation for studying CDR dynamics and offering a vast data source for pre-training machine learning models.

Finally, in Chapter 5, I introduce ABB4-STEROIDS, a generative protein structure prediction model that samples antibody conformational ensembles. Pre-trained on a large body of CG MD data and fine-tuned on a much smaller set of 83 all-atom trajectories, this model outperforms existing generative frameworks in accuracy. In comparison to Cv3-Fv, ABB4-STEROIDS enables direct sampling of the equilibrium distribution without requiring the simulations of time-dependent trajectories which leads to a further reduction in computational cost. Furthermore, as the structure prediction framework allows for integration of multiple data sources during training, including CG and all-atom simulations, some of the inherent biases of the CG data can be mitigated.

Taken as a whole, Chapters 3-5 introduce three methods that facilitate the large-scale study of CDR flexibility. These can be used to study mechanisms by which CDR plasticity gives rise to certain functional properties. They also enable the screening of antibodies based on CDR rigidity, a trait thought to be desirable in drug discovery to improve affinity and specificity, and provide several structural hypotheses for downstream applications, such as antibody-antigen docking.

6.2 Future work

Despite the advances introduced in this thesis, large-scale modelling of antibody conformational ensembles remains a challenge, mirroring persistent hurdles in concurrent research on general protein dynamics.

While the results in Chapters 4 and 5 demonstrate that both Cv3-Fv and ABB4-STEROIDS accurately capture ensemble properties, such as average RMSD and RMSF values, limitations in predicting exact structures of conformational states and their relative occupancies persist. To address these issues, I propose two primary strategies for further model improvements.

The first and more trivial strategy involves the scaling of all-atom MD data through larger-scale simulation efforts. As demonstrated in Chapter 5, fine-tuning ABB4-STEROIDS on a modest set of all-atom trajectories significantly improved the sampled conformational landscapes. Expanding this data by an order of magnitude, a task now rendered feasible by advancements in GPU acceleration, is likely to yield further substantial improvements in predictive accuracy.

A second strategy involves the integration of additional experimental data sources. While atomic-resolution dynamics at femtosecond timescales, as obtained from MD, remain difficult to measure experimentally, several methods provide time- and ensemble-averaged insights. These methods include NMR, SAXS and HDX-MS which are described in detail in Section 1.6.2. Recent work, such as BioEmu (Lewis et al. 2025), demonstrated an initial strategy of fine-tuning ensemble predictors on folding free energies. However, as free energy provides a relatively sparse signal, a single value for an entire protein, and only informs on global folding dynamics, training on richer and denser experimental data has the potential to further enhance model performance. NMR (Wayment-Steele et al. 2025) and HDX-MS (Ferrari et al. 2025) datasets of suitable size to implement such a strategy have been released over the last year.

Beyond data acquisition, a critical avenue for future research is the design of robust confidence metrics for ensemble prediction, analogous to the pLDDT used in AlphaFold2 (Jumper et al. 2021). Currently, no comparable metric exists for ensemble predictors like ABB4-STEROIDS. When such a model predicts a diverse set of conformations for a given protein it remains difficult to distinguish between true diversity predicted due to inherent structural flexibility and diversity predicted as a result of general model uncertainty. In Sweeney et al. (2024), we

explored the application of established uncertainty quantification techniques to derive a suitable confidence metric. While this represents a preliminary approach, it underscores the necessity of a rigorous framework for assessing the reliability of predicted conformational ensembles in future iterations of these models.

Appendices

A

Computational epitope profiling using structural models

A.1 Supplementary Tables and Figures

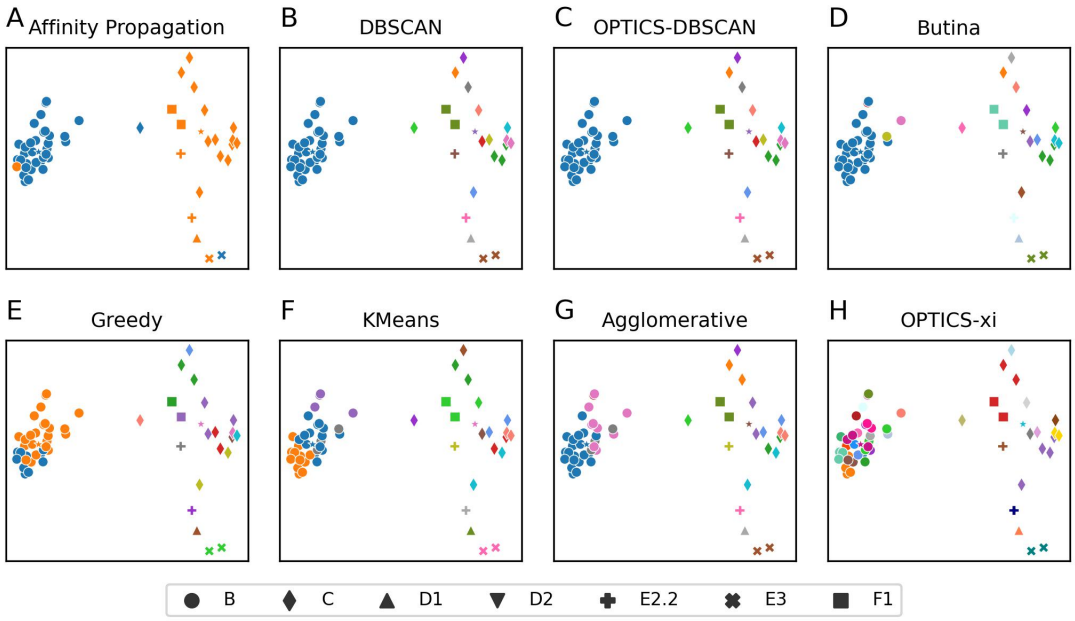


Figure A.1: Visualisation of the clustering obtained by the eight algorithms. Principal coordinate analysis of the RMSD matrix for a group of 62 antibodies with identical CDR lengths in the Cao et al. (2023) training set (length CDRH1: 13 residues, H2: 10, H3: 16, L1: 12, L2: 8, L3: 9). The data points corresponding to the antibodies are coloured by the cluster assigned by one of the algorithms (affinity propagation (A), DBSCAN (B), OPTICS-DBSCAN (C), Butina clustering (D), greedy clustering (E), k-means (F), agglomerative clustering (G), OPTICS-xi (H)) using optimal parameters. The marker style indicates the epitope group the corresponding antibody engages.

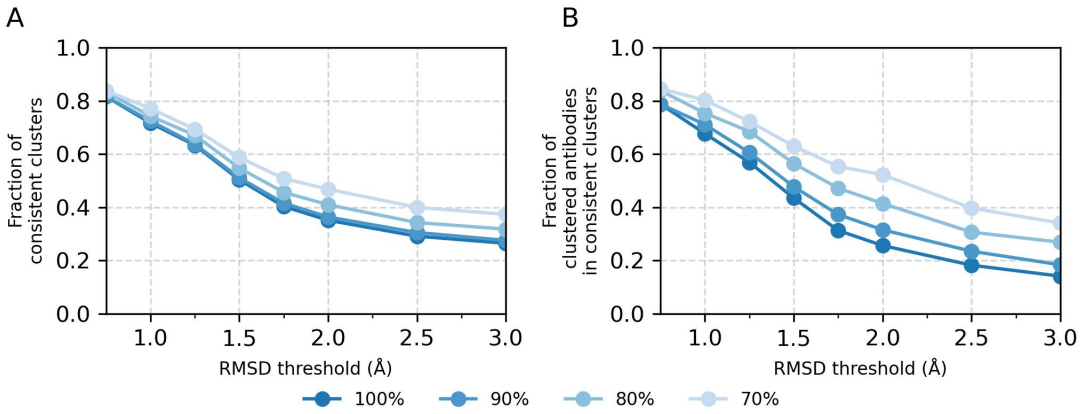


Figure A.2: Extent of cluster inconsistency. Accuracy metrics used in this study only capture clusters that exclusively contain antibodies against the same epitope (100% consistency). The two accuracy metrics, (A) the fraction of epitope-consistent clusters and (B) the fraction of clustered antibodies in epitope-consistent clusters, were recalculated by considering clusters in which at least 90%, 80% and 70% of members engage the same epitope as epitope-consistent. This provides a measure of the extent of inconsistency in clusters not captured by the standard performance metrics. A large number of clusters are observed where most antibodies engage the same epitope, but contain a few incorrect antibodies.

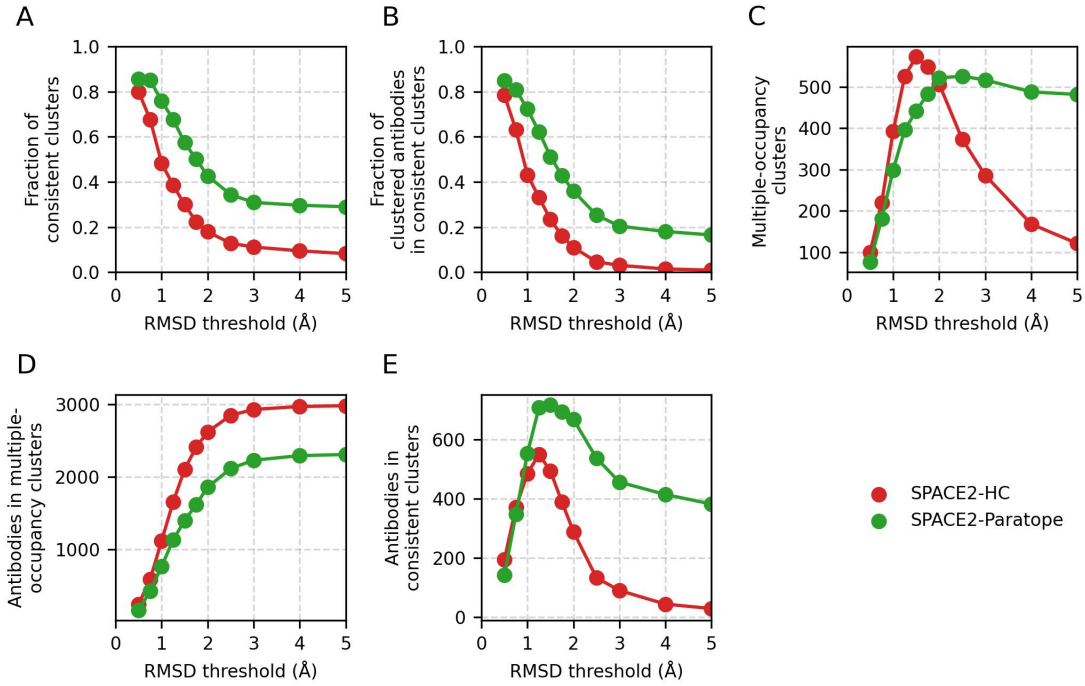


Figure A.3: Parameter optimisation of SPACE2 variations. Versions of the algorithm clustering antibodies based on the similarity of heavy chain CDRs (SPACE2-HC) and the similarity of CDRs predicted to contain paratope residues (SPACE2-Paratope) were implemented. As in the standard SPACE2 method, both SPACE2-HC and SPACE2-Paratope cluster antibodies using an agglomerative clustering algorithm with a 'complete' linkage criterion. A scan of the RMSD threshold parameter was performed to find optimal values. Optimisation was carried out on the Cao et al. (2023) training set. The values of the five performance metrics are plotted against evaluated threshold values: (A) fraction of epitope-consistent clusters, (B) fraction of clustered antibodies in epitope-consistent clusters, (C) number of multiple-occupancy clusters, (D) number of antibodies in multiple-occupancy clusters, (E) number of antibodies in epitope-consistent multiple-occupancy clusters. The best clustering results are achieved at a threshold value of 1.25 Å for SPACE2-HC and 1.5 Å for SPACE2-Paratope as defined by the maximum number of antibodies in epitope-consistent clusters.

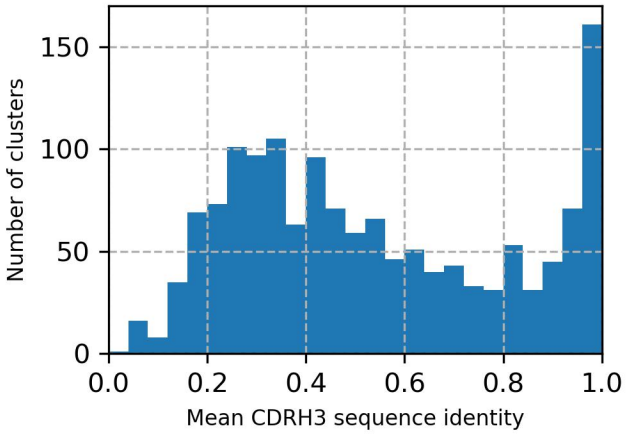


Figure A.4: Histogram of the mean CDRH3 sequence identity within domain-consistent SPACE2 clusters from CoV-AbDab. The distribution shows two peaks. The peak close to 1.0 indicates a group of clusters exclusively containing antibodies with almost identical sequences. A second peak at 0.3 indicates the presence of a large number of clusters containing antibodies that are highly diverse in sequence.

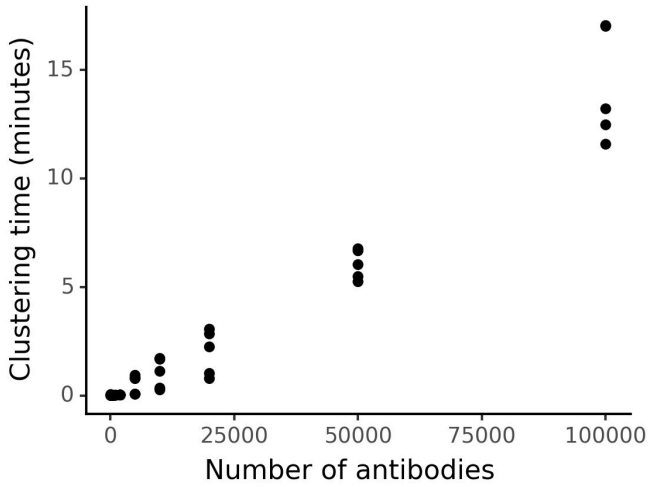


Figure A.5: SPACE2 clustering speed. The SPACE2 algorithm functions in four steps: structural modelling of antibodies, grouping by CDR length, computation of the RMSD matrices and agglomerative clustering. Structural modelling with ABodyBuilder2 is currently the rate-limiting step, which take around 5 seconds per structure using a Tesla P100 GPU (Abanades et al. 2023). To determine the computational cost of the remaining three steps, we clustered a database of 148,000 paired natural antibody sequences from the Observed Antibody Space database pre-modelled with ABodyBuilder2 (Abanades et al. 2023; Olsen et al. 2022). The graph shows the time required to run the remaining three steps of the clustering code parallelised over 12 CPUs plotted against the number of antibodies to be clustered. These steps scale roughly at $O(n^{1.5})$ with the number of antibodies (n). It is possible to cluster a data set of 10,000 antibodies with SPACE2 within half a day (including all steps of the algorithm). In comparison, clonotyping has a lower computational cost. Our in-house script for Fv-clonotyping takes approximately 2 seconds to cluster 10,000 antibodies on a single CPU.

Table A.1: Performance of agglomerative and OPTICS-xi clustering on train set.

Metric	Agglomerative clustering	OPTICS-xi
Fract. of consistent clust.	0.63	0.59
Fract. of clustered antibodies in consistent clust.	0.57	0.55
Multiple-occupancy clust.	480	584
mAbs in multiple-occupancy clust.	1446	1615
mAbs in epitope-consistent multiple-occupancy clust.	823	884
Mean clust. size	3.0	2.7
Max clust. size	28	11

Table A.2: Comparison of standard SPACE2, SPACE2-HC and SPACE2-Paratope performance metrics on the Cao et al. (2023) training set. The best results for each metric are highlighted in bold.

Metric	Standard SPACE2	SPACE2-HC	SPACE2-Paratope
Fract. of consistent clust.	**0.63**	0.36	0.57
Fract. of clustered mAbs in consistent clust.	**0.57**	0.33	0.51
Multiple-occupancy clust.	480	**526**	442
mAbs in multiple-occupancy clust.	1446	**1655**	1339
mAbs in epitope-consistent multiple-occupancy clust.	**823**	550	717

Table A.3: Comparison of SPACE1 evaluated at different RMSD thresholds on the Cao et al. (2023) training set. The lower threshold leads to better metrics of clustering accuracy. However, a larger threshold improves data coverage and the key metric of antibodies in epitope-consistent multiple-occupancy clusters. The best results achieved by SPACE1 for each metric are highlighted in bold. Values achieved by SPACE2 are shown for comparison.

Metric	SPACE1 0.75 Å	SPACE1 1.25 Å	SPACE2
Fract. of consistent clust.	**0.75**	0.64	0.63
Fract. of clustered mAbs in consistent clust.	**0.68**	0.58	0.57
Multiple-occupancy clust.	251	**314**	480
mAbs in multiple-occupancy clust.	723	**935**	1446
mAbs in epitope-consistent multiple-occupancy clust.	494	**538**	823
Fract. of clust. containing >1 VH-clonotypes	0.55	**0.58**	0.55
Mean CDRH3 sequence identity	0.70	**0.67**	0.66

Table A.4: In depth comparison of SPACE2 and SPACE1 performance on the Cao et al. (2023) training set. SPACE1 with a 1.25 Å RMSD threshold, SPACE2, an adaptation of SPACE1 using an agglomerative clustering algorithm (complete linkage criterion, 1.25 Å RMSD threshold) and an adaptation of SPACE2 using a greedy clustering algorithm (1.25 Å RMSD threshold) were evaluated on the complete data set (all). Additionally, SPACE2 and its adaptation were evaluated on a reduced data set which only included the 2140 antibodies successfully modelled by homology modelling (reduced set).

	SPACE1		SPACE2			
Data set	All		Reduced set		All	
Clustering algorithm	Greedy	Agglomerative	Greedy	Agglomerative	Greedy	Agglomerative
mAbs modelled	2140	2140	2140	2140	3046	3046
Fract. of consistent clust.	0.64	0.74	0.55	0.63	0.57	0.63
Fract. of clustered mAbs in consistent clust.	0.58	0.7	0.44	0.57	0.44	0.57
Multiple-occupancy clust.	314	292	288	358	389	480
mAbs in multiple-occupancy clust.	935	802	1109	1104	1451	1446
mAbs in consistent multiple-occupancy clust.	538	559	485	627	651	823

B

Predicting the conformational flexibility of CDRs

B.1 Supplementary methods

B.1.1 ITsFlexible models

ITsFlexible

ITsFlexible is an equivariant graph neural network (EGNN). The model takes a geometric graph as input that contains node features, coordinates and edge features and outputs a probability score that indicates the probability of a data point belonging to the positive class (overview in Figure B.1 and details in Algorithm 1). Initially, node features are embedded into a 128 dimensional vector. Node embeddings are then updated through 3 E(n)-equivariant graph convolutional layers (EGCLs) (Satorras et al. 2022):

$$
\begin{aligned}
m_{ij} &= \phi_m(h_i^l, h_j^l, \|x_i^l - x_j^l\|^2, a_{ij}) \\
x_i^{l+1} &= x_i^l + C \sum_{j \neq i} (x_i^l - x_j^l)\ \phi_x(m_{ij}) \\
m_i &= \sum_{j \neq i} m_{ij} \\
h_i^{l+1} &= \phi_h(h_i^l, m_i)
\end{aligned}
$$

where h^l are the node embeddings, x^l are the coordinates and a_{ij} are the edge features. The node embeddings from the last layer are max-pooled, and a linear layer with a sigmoid activation function is applied for binary classification. Coordinate updates in the EGCL are equivariant with respect to the group E(3) and node embedding updates are invariant. Classification of ITsFlexible is therefore invariant to translations and rotations of the input structure.

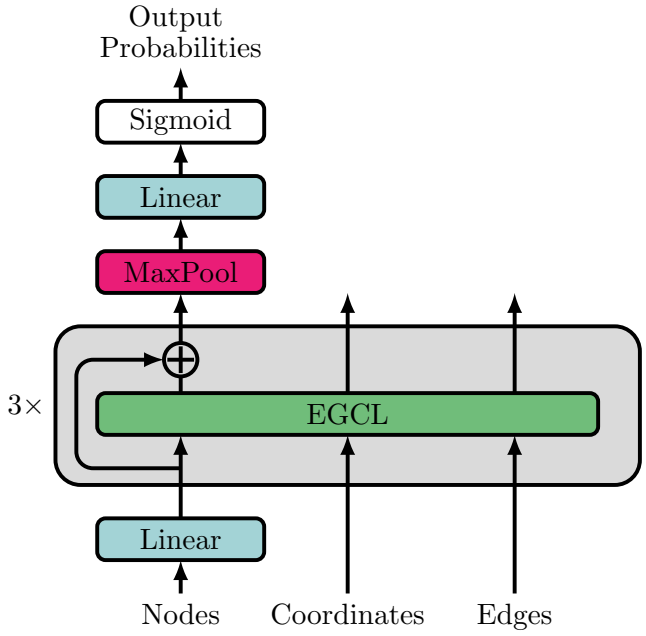


Figure B.1: ITsFlexible model architecture. ITsFlexible is a GNN that consists of 3 E(n)-equivariant graph convolutional layers (EGCLs). The final node features are max pooled and a classification head with linear layer and sigmoid activation is applied.

Algorithm 1 ITsFlexible

Require: Node features $n \in \mathbb{R}^{N_{nodes} \times 22}$, node coordinates $x \in \mathbb{R}^{N_{nodes} \times 3}$, edges $a \in \mathbb{R}^{N_{edges} \times 9}$

def ITsFlexible(n, x, a)
  $h_n^0 \leftarrow \text{Linear}(n_n)$ $\quad h_n \in \mathbb{R}^{128}, n \in \mathbb{R} \cap [0, N_{nodes}]$
  for $l = 1$ **to** 3
    $a^l = \text{DropoutEdge}(a)$
    $h^l, x^l = \text{EGCL}(h^{l-1}, x^{l-1}, a^l)$
    $h^l = h^l + h^{l-1}$
  end for
  $h^4 = \text{MaxPool}(h^3)$ $\quad h^4 \in \mathbb{R}^{128}$
  $o = \text{Sigmoid}(\text{Linear}(h^4))$ $\quad o \in \mathbb{R} \cap [0, 1]$
  return o

ITsFlexible-Loop

ITsFlexible-Loop is an EGNN very similar in architecture and training to ITsFlexible. The main difference is the model input. The ITsFlexible-Loop input is a graph encoding of only loop residues, the structural context is not provided. The coordinates and edge features of the graph are generated as described in the methods section in Chapter 3. As all nodes are located in the loop, node features are reduced to a 21-dimensional one-hot encoding of amino acid type. The remaining hyperparameters are the same as for the ITsFlexible model.

ITsFlexible-Sequence

ITsFlexible-Sequence is a CNN-based model which classifies loops from inputs of a sequence representation. Algorithm 2 shows details of the model. The model input is a one-hot encoding of the amino acid sequence. Each residue is encoded as 21 dimensional vector (1 class for each of the 20 amino acids plus an additional class for unknown residues). Non-standard amino acids closely related to a standard amino acid are encoded as such, others are encoded as unknown residues (Table B.1). Encodings are padded with zeros to a standard length of 51 residues. This enforces a length limit of 51 residues for input sequences, which corresponds to the longest loop observed in the PDB (Berman 2000). A sequence embedding is produced with two 1D-convolutional blocks with 256 channels. The final embedding is flattened to a vector of dimensions 51×256. Two linear layers followed by a sigmoid activation function are used for binary classification from the sequence embedding.

ITsFlexible-Sequence was trained on the PDB set with identical splits as ITsFlexible (see methods). ITsFlexible-Sequence was trained with a binary cross-entropy loss using the Adam optimiser (Kingma et al. 2014) with a learning rate of $2 \cdot 10^{-4}$. During training dropout of 0.2 was used in the convolutional layers and dropout of 0.05 in the linear layers. The validation area under the precision-recall curve (PR-AUC) was monitored and training stopped when converged. Ten models were trained and the one with the best validation PR-AUC was selected.

Algorithm 2 ITsFlexible-Sequence

Require: Loop sequence encoding $s \in \mathbb{R}^{N_{residues} \times 21}$

def ITsFlexibleSequence(s)
  $i = \text{ZeroPad}(s)$ $i \in \mathbb{R}^{51 \times 21}$
  $h_n^0 \leftarrow i_n$ $n \in \mathbb{R} \cap [0, 51]$
  for $l = 1$ **to** 2
    $h^l \leftarrow \text{ReLu}(\text{Conv1d}(h^{l-1}))$ $h^l \in \mathbb{R}^{51 \times 256}$
    $h^l = \text{Dropout}(h^l)$
    $h^l = \text{MaxPool1d}(h^l)$
  end for
  $h^3 \leftarrow \text{Flatten}(h^2)$ $h^3 \in \mathbb{R}^{(51 \times 256)}$
  $h^4 \leftarrow \text{Dropout}(\text{ReLu}(\text{Linear}(h^3)))$ $h^4 \in \mathbb{R}^{512}$
  $o \leftarrow \text{Sigmoid}(\text{Linear}(h^4))$ $o \in \mathbb{R} \cap [0, 1]$
  return o

B.1.2 Amino acid encoding

When encoding ITsFlexible inputs, non-standard amino acids were represented either as a unknown residue or as the closest related standard amino acid. Non-standard amino acids that occur frequently in the test sets and are closely related to one of the 20 standard amino acids were encoded as such. Table B.1 shows a map of the selected non-standard amino acids and their encoding.

Table B.1: Encoding of non-standard amino acids

Non-standard amino acid	Encoded as
Beta-L-aspartic acid (IAS)	Aspartic acid (ASP)
S-hydroxycysteine (CSO)	Cysteine (CYS)
4-methyl-histidine (HIC)	Histidine (HIS)
N-dimethyl-lysine (MLY)	Lysine (LYS)
Selenomethionine (MSE)	Methionine (MET)
Phosphoserine (SEP)	Serine (SER)
Topaquinone (TPQ)	Tyrosine (TYR)
3-amino-6-hydroxy-tyrosine (TYQ)	Tyrosine (TYR)

B.1.3 Molecular dynamics dataset

19 antibodies were simulated with the molecular dynamics protocol described in the methods section of Chapter B. The number of conformations observed for CDRH3s and CDRL3s is shown in Table B.2.

Table B.2: Number of conformations observed in molecular dynamics simulations

Antibody	N CDRH3 conformations	N CDRL3 conformations
1BFO	3	1
1JPS	3	1
1MLB D44	3	2
2Q76 D44	1	1
2VXT	1	2
2Y06	3	1
2Y07	3	1
2Y36	3	1
3EOA	5	1
3G6D	3	4
3HI6	4	2
3RVW	6	2
3V6F	9	6
4KMT	2	1
5I15	6	1
5I18	9	1
5I1A	3	1
7G12	3	2
7G12 mature	1	1

B.1.4 Cryo-EM

Table B.3: Cryo-EM data collection, refinement and validation statistics

	antibody 1 - 9N5Y - AEL31302/AEL31311	antibody 2 - 9N5Z - AMB38310/AMB38599
Data collection and processing		
Microscope	TFS Glacios	TFS Glacios
Voltage (keV)	200	200
Camera	TFS Falcon 4i	TFS Falcon 4i
Collection mode	Counting	Counting
Magnification	190,000x	190,000x
Pixel size at detector (Å)	0.718	0.718
Total electron exposure (e–/Å2)	45	44.84
Number of EER frames	40	40
Defocus range (μ m)	-0.8 to -1.6	-0.8 to -1.6
Automation software	EPU	EPU
Micrographs collected (no.)	4,628	3,976
Micrographs used (no.)	4,457	3,295
Initial particle images (no.)	626,964	641,551
Final particle images (no.)	85,692	142,446
Symmetry	C1	C3
Map resolution (masked/unmasked Å)	3.7/3.8	3.0/3.1
FSC threshold	0.143	0.143
Refinement		
Initial model used (PDB code)	7T3D	7T3D
Refinement package	Phenix RSR	Phenix RSR
Model resolution (Å)	3.77	3.16
FSC threshold	0.5	0.5
EMRinger score	3.08	3.32
CC (mask)	0.80	0.77
Model composition		
Non-hydrogen atoms	13,754	13,702
Protein residues	1728	1724
Ligands	10	14
R.m.s. deviations		
Bond lengths (Å)	0.005	0.002
Bond angles (°)	0.996	0.583
Validation		
MolProbity score	1.39	1.61
Clashscore	3.37	12.41
Poor rotamers (%)	0.00	0.00
Ramachandran plot		
Favored (%)	96.14	98.3
Allowed (%)	3.86	1.70
Disallowed (%)	0.00	0.00
Cβ outliers (%)	0.00	0.00
CaBLAM outliers (%)	2.95	2.6

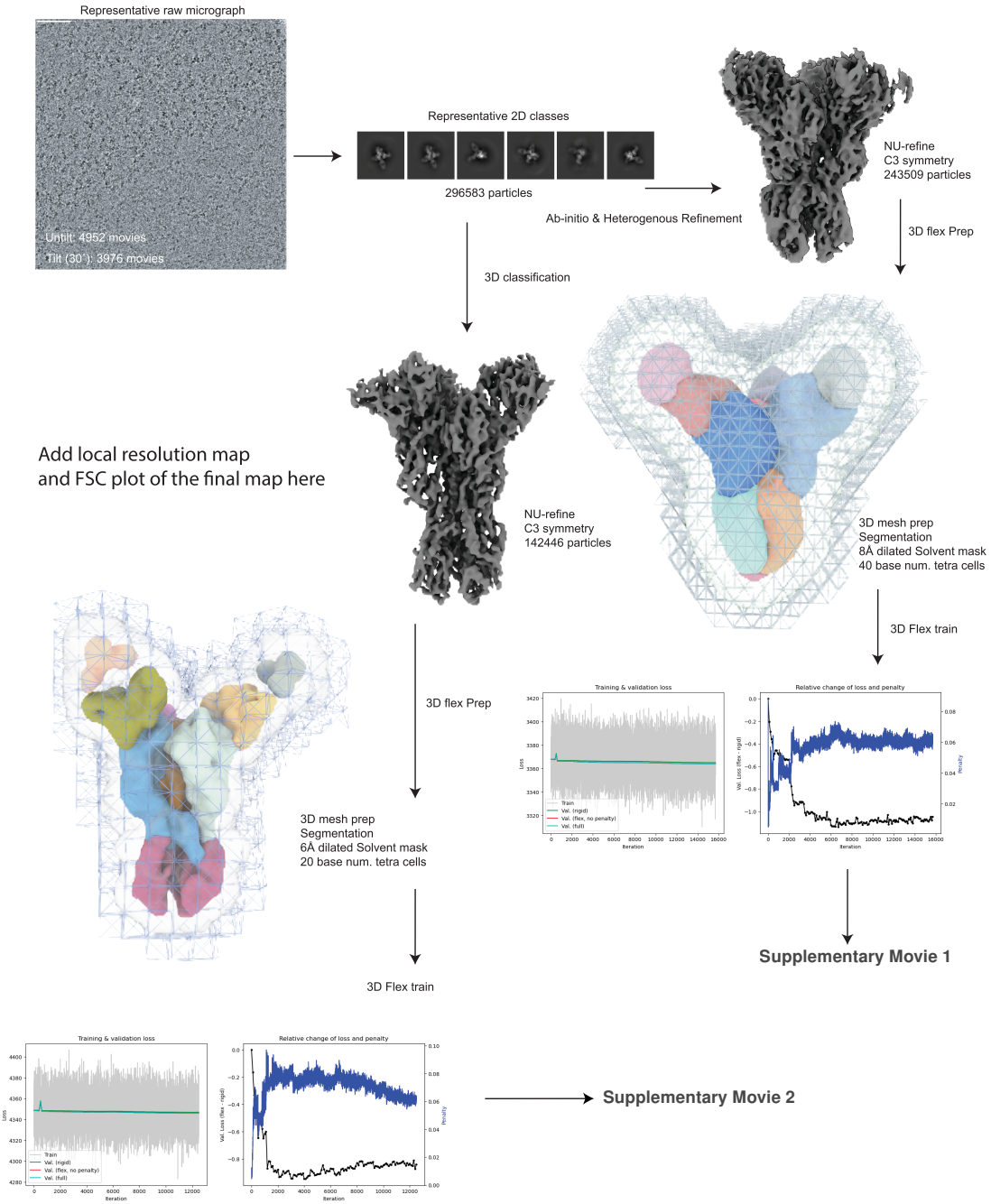


Figure B.2: Workflow of the flexibility analysis in cryo-EM experiments. Figure detailing the 3D flex workflow described in the methods section of Chapter 3.

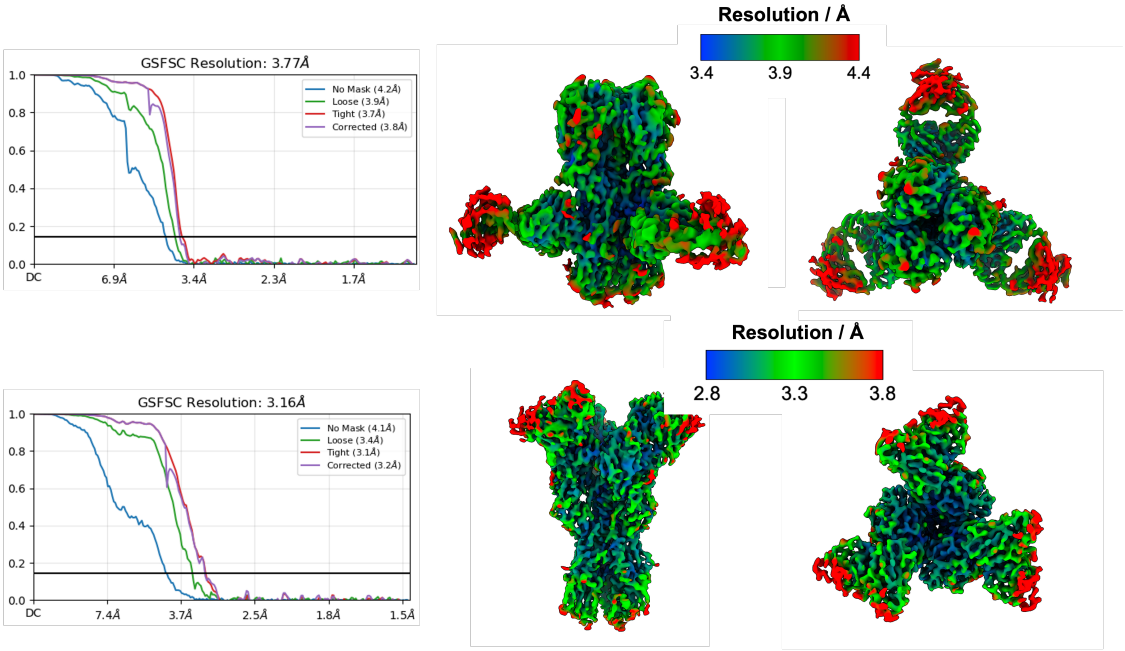


Figure B.3: Local resolution maps for antibody 1 / 9N5Y (top) and antibody 2 / 9N5Z (bottom).

B.2 Supplementary results

B.2.1 CDR3 secondary structure

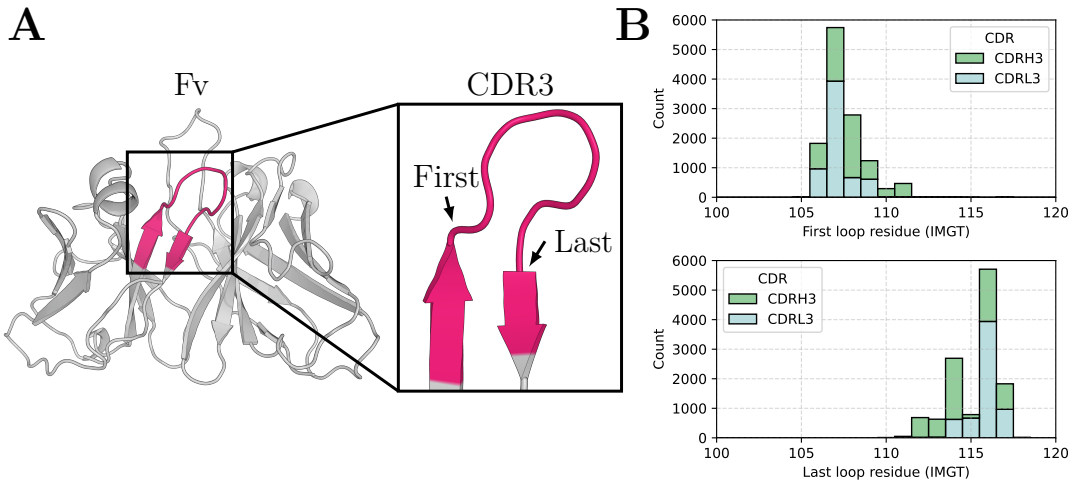


Figure B.4: Antibody and TCRs CDR3s are loops bounded by two antiparallel β-strands. A) The variable domains (Fvs) adopt an immunoglobulin fold which consists of 9 antiparallel β-strands. CDR3s, defined by IMGT residues 105 - 117 (highlighted in pink), connect the two C-terminal strands of the Fv. The start (105) and end (117) residues are located on the strands and the majority of the CDR3 forms a loop motif. The first and last residue in the loop motif are highlighted. B) Histogram showing the distribution of antibody variable domain residues (IMGT numbered) that occur as the first and last residue of the loop motif. The majority of loops start at IMGT residue 107 and end at IMGT residue 116. Unless stated otherwise, a CDR3 is defined as IMGT residue 107-116 throughout the paper.

B.2.2 ALL-conformations

ALL-conformations statistics

This section provides additional statistics for ALL-conformations. The distributions of loop length and the number of structures available for loops are shown in Figure B.5 & B.6. Correlation between flexibility and loop length is shown in Table B.4. The number of loops labelled as 'flexible' and 'rigid' in each set is shown in Table B.5.

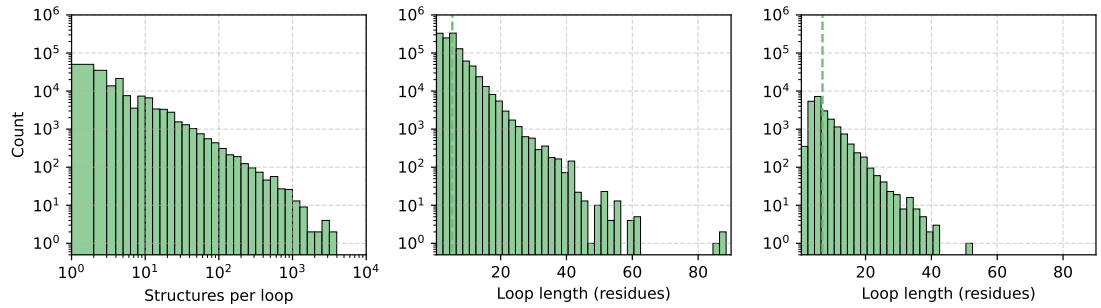


Figure B.5: Distributions of ALL-conformations PDB set. The left column shows a histogram of the number of available structures for each loop in ALL-conformations (N=1.2e6). The middle column shows the distribution of loop length for all loops in ALL-conformations (N=1.2e6). The right column shows the distribution of loop length for only those loops with determined flexibility (i.e. rigid and flexible in Table B.5, N=20,814). The vertical dashed lines indicate the mean of a distribution.

Table B.4: Correlation of CDR length and flexibility

Set	Original dataset (alignment on CDR)	Alternative dataset (alignment on Fv)
CDRH3	0.13	0.11
CDRL3	0.34	0.21
CDRB3	0.36	0.16
CDRA3	0.12	-0.03

R values of the point biserial correlation of the number of residues in the CDR and binary labels rigid (0) and flexible (1).

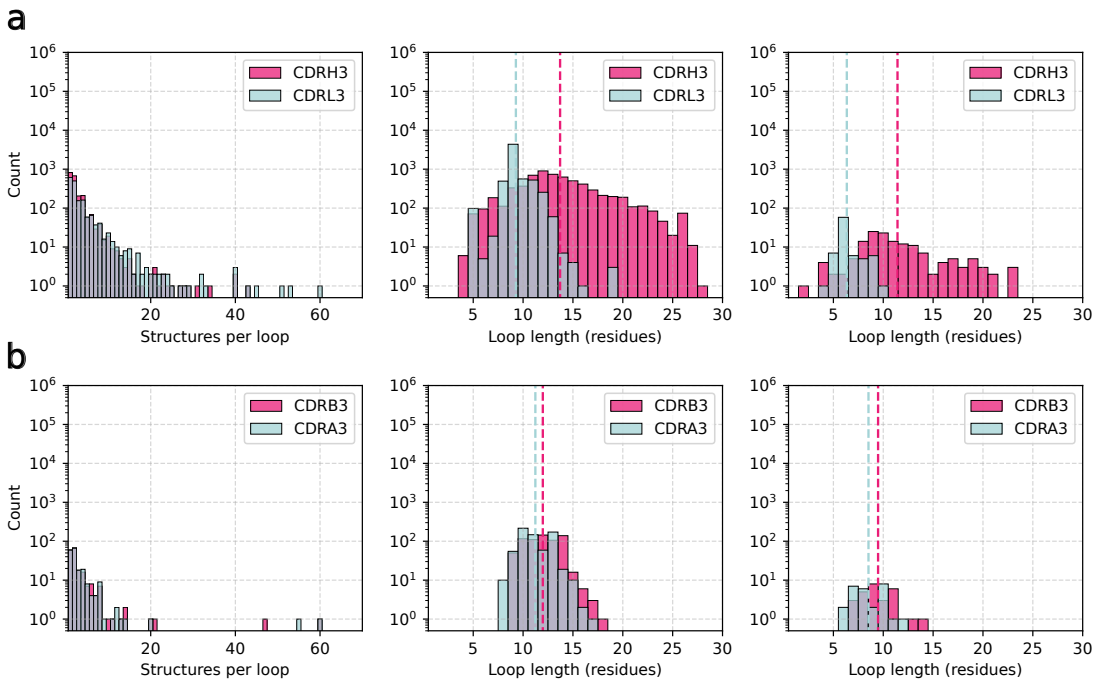


Figure B.6: Distributions of ALL-conformations CDR sets. a) Antibody CDRH3 and CDRL3 sets, b) TCR CDRB3 and CDRA3 sets. The left column shows a histogram of the number of available structures for each loop in ALL-conformations. The middle column shows the distribution of loop length for all loops in ALL-conformations. The right column shows the distribution of loop length for only those loops with determined flexibility (i.e. rigid and flexible in Table B.5). The vertical dashed lines indicate the mean of a distribution. The number of plotted data points is shown in Tables 3.1 & B.5.

Table B.5: Conformational flexibility in ALL-conformations

Dataset	Unknown flexibility	Flexible - multiple conformations	Rigid - single conformation	Total loops with determined flexibility
PDB set	78,000	4289	16,525	20,814
Antibodies				
CDRH3	2325	97	55	152
CDRL3	2282	15	69	84
TCRs				
CDRB3	207	21	6	27
CDRA3	206	21	6	27

Alternative definition of loop flexibility

In Chapter 3, I made several choices to determine the conformational flexibility of loops in ALL-conformations. Structures in the PDB set were initially grouped by sequence identity of the loop, whereas structures in the CDR3 sets were grouped by sequence identity of the Fv. Structural similarity was then calculated by the RMSD of loop residues after structural alignment on the loop.

In this section, a different approach to determine the conformational flexibility was explored. In the PDB set structures were grouped by sequence identity of the loop and two additional anchor residues on either of its sides. Structural similarity was then calculated by the RMSD of the loop after structural alignment on the anchor residues.

Antibody and TCR structures were grouped by Fv identity as in the original dataset. To calculate structural similarity, structures were aligned on the framework regions of the chain in which the CDR is located. I chose to align on the entire framework region rather than anchor residues (as for the PDB set) to reduce noise in the alignment. In certain cases the β-strand located on the C-terminal side of the CDR3 is not resolved well in crystal structures due to its proximity to the end of the chain. When choosing to align on anchor residues only the alignment can therefore show inaccuracies.

A conformation was defined, as described in the methods section, using agglomerative clustering with complete linkage and a 1.25 Å distance threshold. I recalculated the flexibility of loops in ALL-conformations using this alternative definition of flexibility. Summary statistics are shown in Table B.6.

Table B.6: Conformational flexibility in ALL-conformations using alternative definitions

Dataset	Unknown flexibility	Flexible - multiple conformations	Rigid - single conformation	Total loops with determined flexibility
PDB set	128,000	18,526	15,927	34,453
Antibodies				
CDRH3	2241	187	49	236
CDRL3	2239	66	61	127
TCRs				
CDRB3	193	35	6	41
CDRA3	197	31	5	36

B.2.3 Extended evaluation of ITsFlexible performance

Consistency of predictions from multiple structures of the same loop

Throughout Chapter 3, ITsFlexible performance is evaluated by making a prediction for each available crystal structure containing the loops in the test set. The mean prediction score across all structures containing the same loop is then used to calculate performance metrics. In this section, I investigated the consistency of the predictions when using different structures of the same loop as input. Figure B.7 shows the standard deviation of the ITsFlexible prediction score across all structures containing the same loop. This value is generally small suggesting consistent predictions.

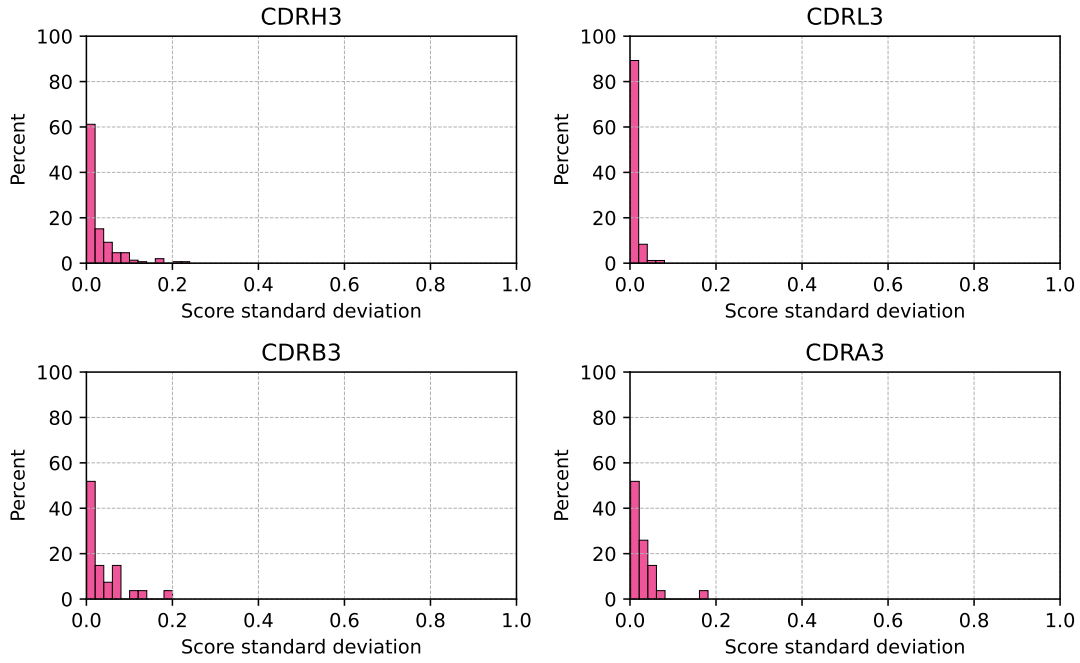


Figure B.7: Standard deviation of ITsFlexible prediction scores when making predictions using multiple crystal structures containing the same loop as input. Number of data points plotted for CDRH3: N=147, CDRL3: N=84, CDRB3: N=27, CDRA3: N=27.

Predictions from structural models

Table B.7: Comparison of AF2 and IB model accuracy and impact on ITsFlexible predictions

	CDRH3 (N=147)		CDRL3 (N=84)		CDRB3 (N=27)		CDRA3 (N=27)	
	RMSD	MAE score	RMSD	MAE score	RMSD	MAE score	RMSD	MAE score
Xtal v AF2	3.2 Å	0.073	1.4 Å	0.014	4.9 Å	0.064	5.3 Å	0.120
Xtal v IB	1.8 Å	0.049	0.8 Å	0.011	1.6 Å	0.051	1.8 Å	0.067

The RMSD columns indicate the average C-alpha RMSD of CDR residues after alignment on framework residues between crystal structure and structural models across the CDR test sets. The MAE score columns show the mean absolute error in the ITsFlexible prediction when using a structural model compared to the crystal structure as input.

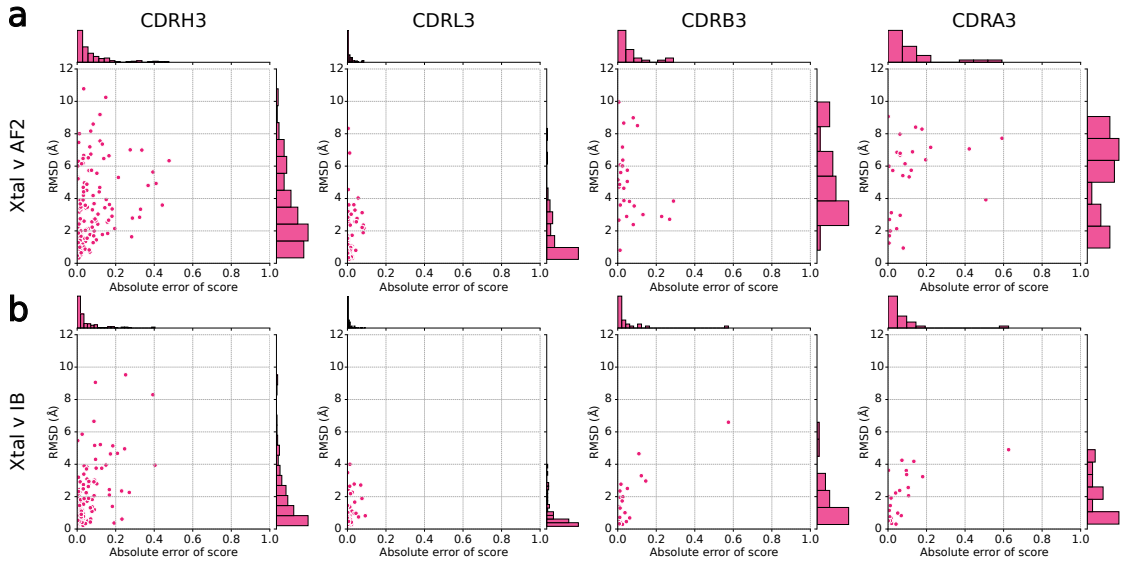


Figure B.8: Impact of model accuracy on ITsFlexible prediction. Analysis for (a) AF2 and (b) IB predictions. For the four CDR test sets the RMSD of a structural model is plotted against the absolute error of the ITsFlexible prediction score (i.e. the absolute difference in prediction score when using a structural model instead of the crystal structure as input). Number of data points plotted for CDRH3: N=147, CDRL3: N=84, CDRB3: N=27, CDRA3: N=27.

Performance with secondary structure-based CDR3 definitions

Throughout the analysis in the main sections of this chapter, I defined CDR3s as IMGT residues 107-116. This range was shown to align with the secondary structure of the loop for the majority of antibodies (Figure B.4). In this section,

I assessed the impact of defining each CDR3 exactly by its secondary structure. I used the DSSP algorithm (Kabsch et al. 1983) to identify the exact numbering of the first residue after the immunoglobulin F-strand and the last residue before the G-strand and extracted the region in between as the CDR3. I refer to this definition as the DSSP CDR3.

CDRH3 and CDRL3 test sets were created with conformational flexibility recalculated over DSSP CDR3 definitions and ITsFlexible performance was evaluated (Table B.8). Similar performance is achieved for the DSSP and the standard CDR3 definitions. Due to the simplicity of CDR3s being defined by the same IMGT numbers in all antibodies across the dataset, I chose to use the 107-116 definition throughout Chapter 3.

Table B.8: DSSP defined CDR test set performance.

Method	CDRH3		CDRL3	
	DSSP (N=141)	IMGT 107-116 (N=147)	DSSP (N=84)	IMGT 107-116 (N=84)
Random	0.62	0.65	0.19	0.18
Baselines				
Solvent Exposure	0.66	0.69	0.25	0.32
Length	0.64	0.69	0.29	0.36
Combined	0.73	0.72	0.38	0.37
ITsFlexible				
Crystal structure	**0.82**	**0.82**	**0.52**	**0.49**

The performance of the methods on CDR test sets is evaluated by the area under the precision-recall curve (PR AUC). The best performance achieved for each test set is highlighted in bold.

Performance with alternative definition of flexibility

ITsFlexible was retrained on the dataset with alternative definitions of flexibility as described in Section B.2.2. This model was parametrised similarly to the model described in the methods section. The context distance threshold was increased to 30 Å and the edge threshold distance to 20 Å to provide the model with information about a larger structural context.

Model performance was evaluated on the PDB and the CDR test sets (Figure B.9). The data splits were created based on sequence identity. The PDB test set did not contain any loops with more than 80% sequence identity across the loop and anchor residues to length matched loops in the training and validation sets. The CDR test sets did not contain any loops with more than 80% aligned sequence identity to any loop (not restricted to matching loop length) in the training and validation sets. Baselines and alternative workflows were performed as outlined in the methods. I made a small modification of the AF2 MSA subsampling to make the workflow more consistent with the definition of flexibility used here. Flexibility in the structural ensembles was calculated by loop RMSD after alignment on the framework regions of the corresponding chain. ITsFlexible achieves good performance on all five datasets, indicating the model's ability to detect flexibility signals independent of choices made when calculating structural flexibility.

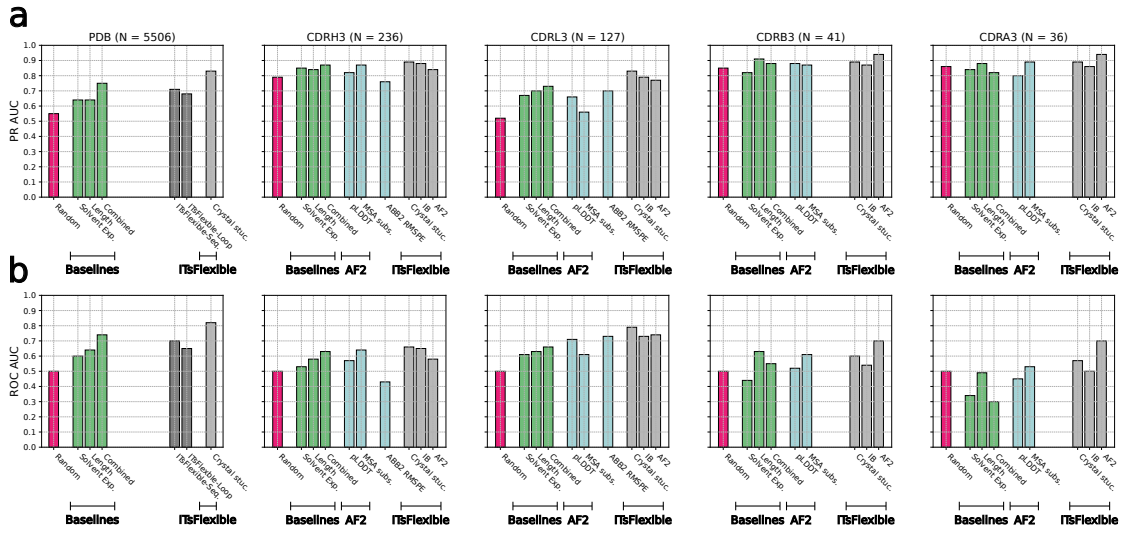


Figure B.9: ITsFlexible performance with alternative definitions of flexibility. Classification on the five test sets is evaluated with metrics of (a) PR AUC and (b) ROC AUC. ITsFlexible performance from inputs of crystal structures, ImmuneBuilder (IB) and AF2 models (light grey) is compared to random classification (red), three biophysical baselines (green) and three zero-shot models based on outputs of protein structure prediction tools (blue). Additionally, for the PDB test set I evaluated two ITsFlexible versions with input ablations (dark grey).

Impact of varying RMSD thresholds on datasets and model performance

In 3, I defined a loop conformation as a structural cluster where the pairwise RMSD between any two members is below a threshold of 1.25 Å. This threshold was chosen based on previous work suggesting 1.25 Å as an optimal threshold for functional clustering. The definition of a conformation should, therefore, most accurately reflect distinct functional states.

In this section, I analysed how different values of the threshold impact the proportion of labels in the datasets. At low thresholds (1 Å) the highest number of flexible and lowest number of rigid loops are observed (Figure B.10). Increasing the threshold leads to a large reduction of flexible loops accompanied by a small increase in rigid loops. At a value of 2.5 Å all CDR test sets, except for H3, contain less than five flexible examples.

In a next step, I investigated how varying RMSD threshold and the consequent proportion of labels affects the predictive performance of ITsFlexible. I chose 1 Å and 2 Å as two additional values to evaluate the model. For values below 1 Å the resolution of the experimental structure may make the determined labels less reliable and for values above 2 Å the limited number of flexible loops in the test set would hinder detailed analysis of predictive performance. I relabelled the datasets using these new RMSD thresholds and retrained and evaluated ITsFlexible on these labels. The model architecture and training procedure were kept identical for this experiment.

When using the new RMSD thresholds, ITsFlexible retains high accuracy at predicting loop flexibility (Figure B.10). Measured by PR AUC the model outperforms the biophysical baselines and zero-shot classifiers in nearly all CDR test sets. The only exception is the CDRA3 set where AF2 based zero-shot predictors come top in both cases. As observed for the standard RMSD threshold (1.25 Å), ITsFlexible is the only method that consistently achieves high predictive accuracy across all test sets.

When comparing the classifier performance at the two thresholds, I observed slightly better results at 2 Å. The metrics achieved by ITsFlexible tend to show

a larger increase over random classification than at the 1 Å threshold (Figure B.10). Looking at the distribution of labels in the datasets, this may perhaps be counterintuitive. A better balance of rigid and flexible loops in the 1 Å datasets would, in theory, be more favourable for model training. However, I speculate that labels are less noisy at higher thresholds which may improve classification.



Figure B.10: Impact of varying RMSD thresholds on datasets and model performance. a-c) Number of loops labelled as rigid and flexible as a function of the RMSD threshold used to define a conformation. The PDB set is shown in (a), antibody CDRs in (b) and TCR CDRs in (c). d-e) ITsFlexible performance evaluated on the test sets using a 1 Å RMSD threshold in metrics of PR AUC (d) and ROC AUC (e). f-g) ITsFlexible performance evaluated on the test sets using a 2 Å RMSD threshold in metrics of PR AUC (f) and ROC AUC (g). ITsFlexible performance from inputs of crystal structures (grey) is compared to random classification (red), three biophysical baselines (green) and three zero-shot models based on outputs of protein structure prediction tools (blue).

B.2.4 ABodyBuilder2

ABB2 retraining

The original ABB2 (Abanades et al. 2023) was trained on a version of SAbDab (Schneider et al. 2022) downloaded in July 2021. As SAbDab was also used to create ALL-conformations and consequently the antibody CDR flexibility test sets, there is a substantial overlap with the ABB2 training set (Figure B.9). To avoid data leakage, ABB2 was retrained using the original training protocol on a dataset with the ALL-conformations overlap removed. Antibodies with 100% sequence identity in CDRH3 or CDRL3 regions to a data point in the CDRH3 or CDRL3 flexibility test sets were removed from the training set. This reduced the number of training set antibodies from 5669 to 4469. The validation and test set were not changed. The accuracy of the retrained ABB2 is compared to the original ABB2 on a benchmark of 34 non-redundant antibody structures from the SAbDab (Table B.10). The retrained ABB2 retained high accuracy at predicting antibody structures.

Table B.9: Overlap of antibody CDR test sets and ABB2 training set

	CDRH3	CDRL3
Total in test set	147	84
ABB2 training overlap*	127	76

*Number of CDR3 sequences that also appear in the ABB2 training set.

Table B.10: Comparison of ABB2 and retrained ABB2 on benchmark

Method	CDRH1	CDRH2	CDRH3	FwH	CDRL1	CDRL2	CDRL3	FwL
ABB2 original	0.86	0.74	2.61	0.74	0.52	0.38	0.87	0.56
ABB2 retrained	0.87	0.75	2.93	0.76	0.54	0.35	0.92	0.58

The mean RMSD (in Å) of the prediction to the crystal structure across the benchmark set is shown for the six CDRs and the framework regions according to IMGT definitions.

Flexibility classification with ABB2

In this section, I show results of using the original ABB2 (trained by Abanades et al. (2023)) for flexibility classification and highlight that due to data leakage and dataset biases the performance is not representative. Flexibility classification based on RMSPEs obtained from structures predicted with the original ABB2 performs substantially better than using the retrained ABB2 (Table B.11). This boost in performance is linked to biases in the CDR flexibility test sets. When labelling CDRs as rigid or flexible, I introduced a requirement that a rigid CDR needs to appear in at least five separate PDB structures (see methods). Therefore, CDRs labelled as rigid tend to be represented by more copies in the ABB2 training set than those labelled to be flexible (Figure B.11).

I further observed a negative correlation between the number of times a CDR occurs in the ABB2 training set and its RMSPE (Figure B.11). This finding can be explained by the way that ABB2 calculates the PE score. ABB2 predicts four structures for each antibody. The PE score is calculated as the diversity between these structures (Abanades et al. 2023). The more often an antibody appears in the training set, the more it is weighted in the training loss. This makes it more likely that the antibody structure is predicted with high accuracy and reduces the diversity between the four structures. The correlation disappears for the retrained ABB2 which shows that the correlation is indeed an artifact of the training set.

To highlight the importance of this bias, I predicted flexibility simply based on the number of times an antibody appears in the original ABB2 training set (Table B.11). This predictor, based on data bias only, performs closer to the original ABB2 than to the random baseline.

Table B.11: Flexibility classification with ABB2

Method	CDRH3 (n = 147)	CDRL3 (n = 84)
Random	0.65	0.18
ITsFlexible	0.82	0.49
ABB2 rmspe		
Retained	0.70	0.39
Original	0.81	0.62
N structures in ABB2 training set*	0.74	0.41

*Logistic regression model with input corresponding to the number of times a given CDR occurs in the ABB2 training.
The performance of the methods on CDR test sets is evaluated by the area under the precision-recall curve (PR AUC).

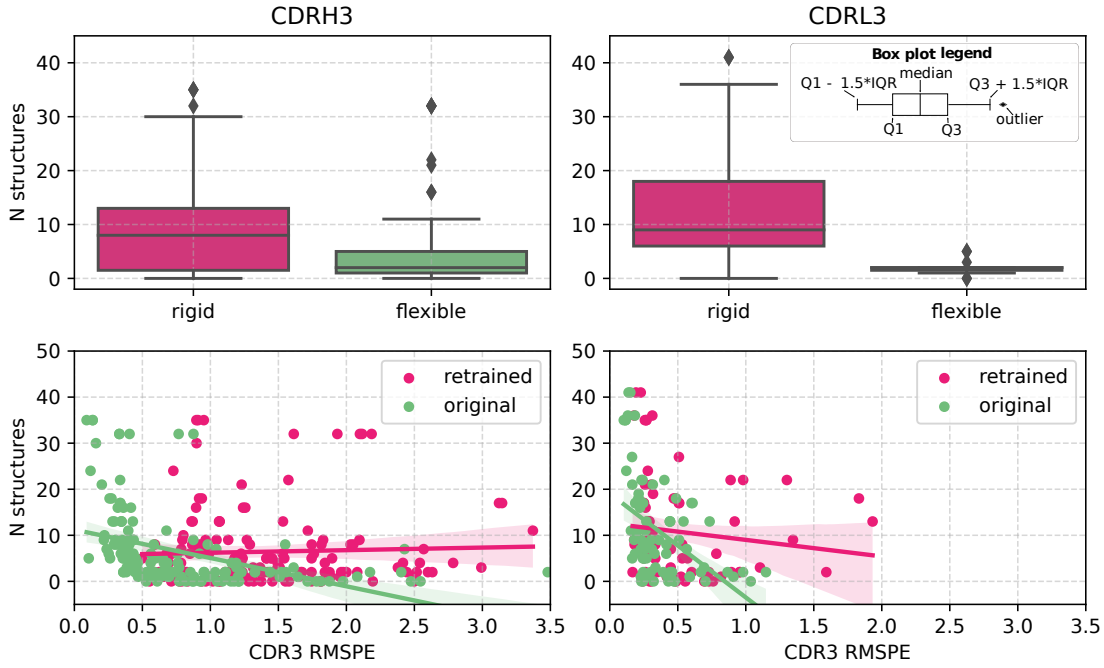


Figure B.11: Biases in the antibody CDR3 flexibility classification test sets. top) Boxplot of the number of structures appearing in the original ABB2 training set for CDRs labelled as flexible and rigid. Rigid CDRs tend to be represented more often in the training set than flexible structures. The box shows the quartiles of the distribution, the whiskers extend to 1.5 times the inter-quartile range (IQRs) of the lower and upper quartile and outliers are shown as points. Number of data points plotted for CDRH3-rigid: N=59, CDRH3-flexible: N=54, CDR3-rigid: N=69, CDRL3-flexible: N=4. bottom) The number of structures in the original ABB2 training set is plotted against the CDR3 RMSPE for the retrained and the original ABB2. A linear regression is fit to the data points. For the original ABB2 there is a moderate negative correlation (CDRH3: Pearson $R = -0.42$, CDRL3: Pearson $R = -0.47$) between the number of training set structures and the RMSPE of a CDR. For the retrained ABB2 there is no correlation (CDRH3: Pearson $R = 0.06$, CDRL3: Pearson $R = -0.12$). Data points plotted for CDRH3: N=113, CDRL3: N=73.

B.2.5 Case study antibodies for cryo-EM

Table B.12: ITsFlexible and zero-shot classifier predictions of cryo-EM case study antibodies

ID	ITsFlexible prediction	ITsFlexible score		AF2 pLDDT		AF2 MSA RMSD		ABB2 RMSPE	
		Value	Pctile	Value	Pctile	Value	Pctile	Value	Pctile
1 - 9N5Y	rigid	0.02	29	71	73	3.13	100	0.90	33
2 - 9N5Z	flexible	0.76	100	68	81	2.80	100	2.20	89
3	flexible	0.60	97	71	67	0.51	6	1.42	54

Pctile indicates the percentile of the value compared to CDRs with the same label across the ALL-conformations CDRH3 test set. The percentile is calculated so that a higher number indicates a more flexible prediction. For the ITsFlexible score, AF2 MSA RMSD and ABB2 RMSPE higher values are associated with increased flexibility and the 100th percentile is set to the highest value observed in the test set. For the AF2 pLDDT a lower value is associated with increased flexibility and the 100th percentile is set to the lowest value observed in the test set.

C

Uncovering CDR flexibility through large-scale molecular dynamics

C.1 Supplementary Figures

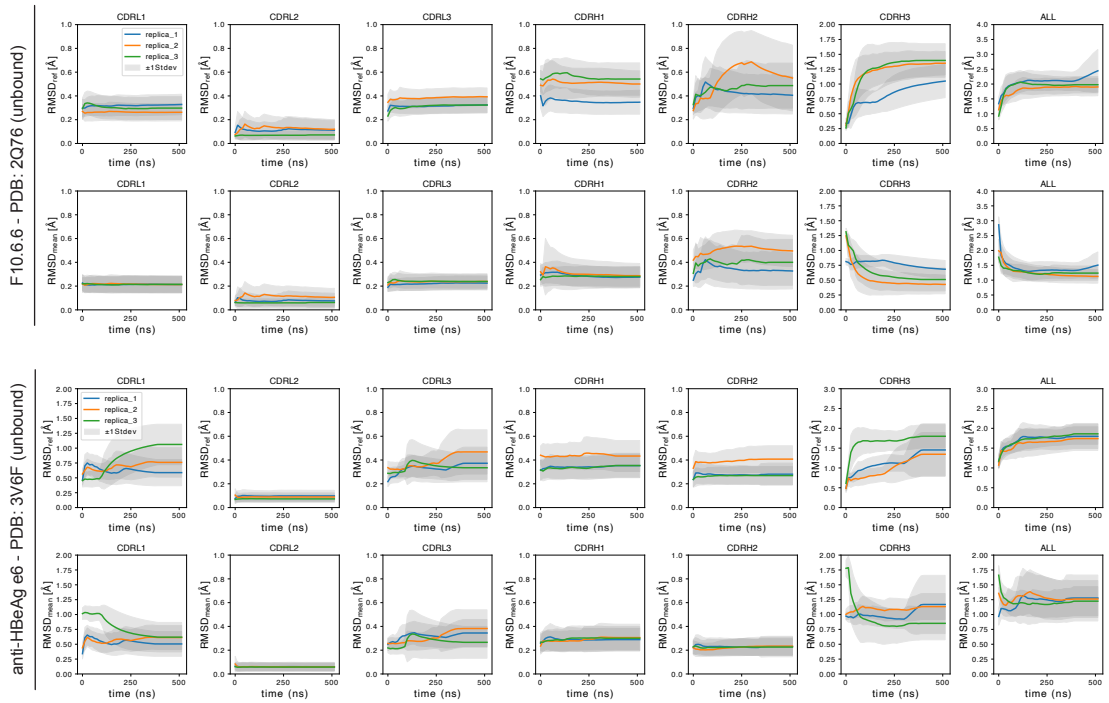


Figure C.1: Convergence curves for all-atom simulations of benchmark antibodies. For each antibody system, the upper panels show the cumulative mean RMSD relative to the starting structure for each CDR, while the lower panels show the cumulative mean RMSD relative to the mean atomic coordinates. Each simulation replica is represented by a different colour, and shaded areas indicate the standard deviation around the cumulative mean.

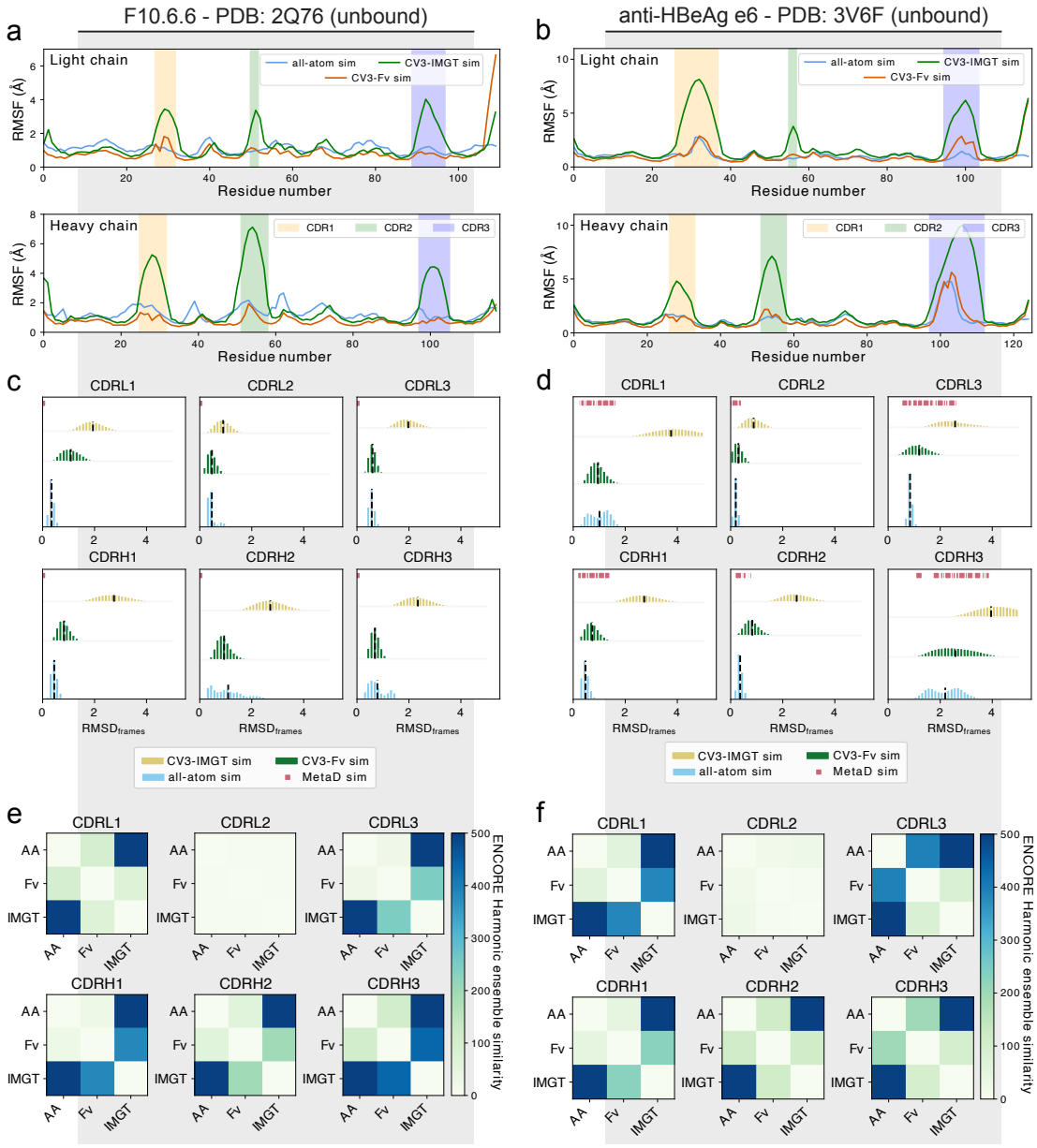


Figure C.2: Optimising CALVADOS 3 for Fv simulations Comparison of the flexibility of Fv regions, particularly the CDR loops, in CV3-IMGT and CV3-Fv simulations for two monoclonal antibodies: anti-hen egg white lysozyme (PDB: 2Q76) and anti-HBV e-antigen (PDB: 2Q76). (a,b) Residue-level RMSF values from CV3-IMGT (dark green), CV3-Fv (red) and all-atom (blue) simulations. Shaded backgrounds mark different CDR loops. (c,d) Frame-wise RMSD distribution for all six CDR loops, comparing CV3-IMGT (yellow), CV3-Fv (green), all-atom simulations (light blue), and metaD cluster structures (red). Mean RMSD values are indicated with black dashed lines. (e,f) Harmonic Ensemble Similarity (HES) matrices obtained from the ENCORE package for each CDR in the two benchmark systems, comparing all-atom (AA), CV3-IMGT (IMGT) and CV3-Fv (Fv) simulation ensembles. The lower the HES score, the higher the agreement between the two ensembles.

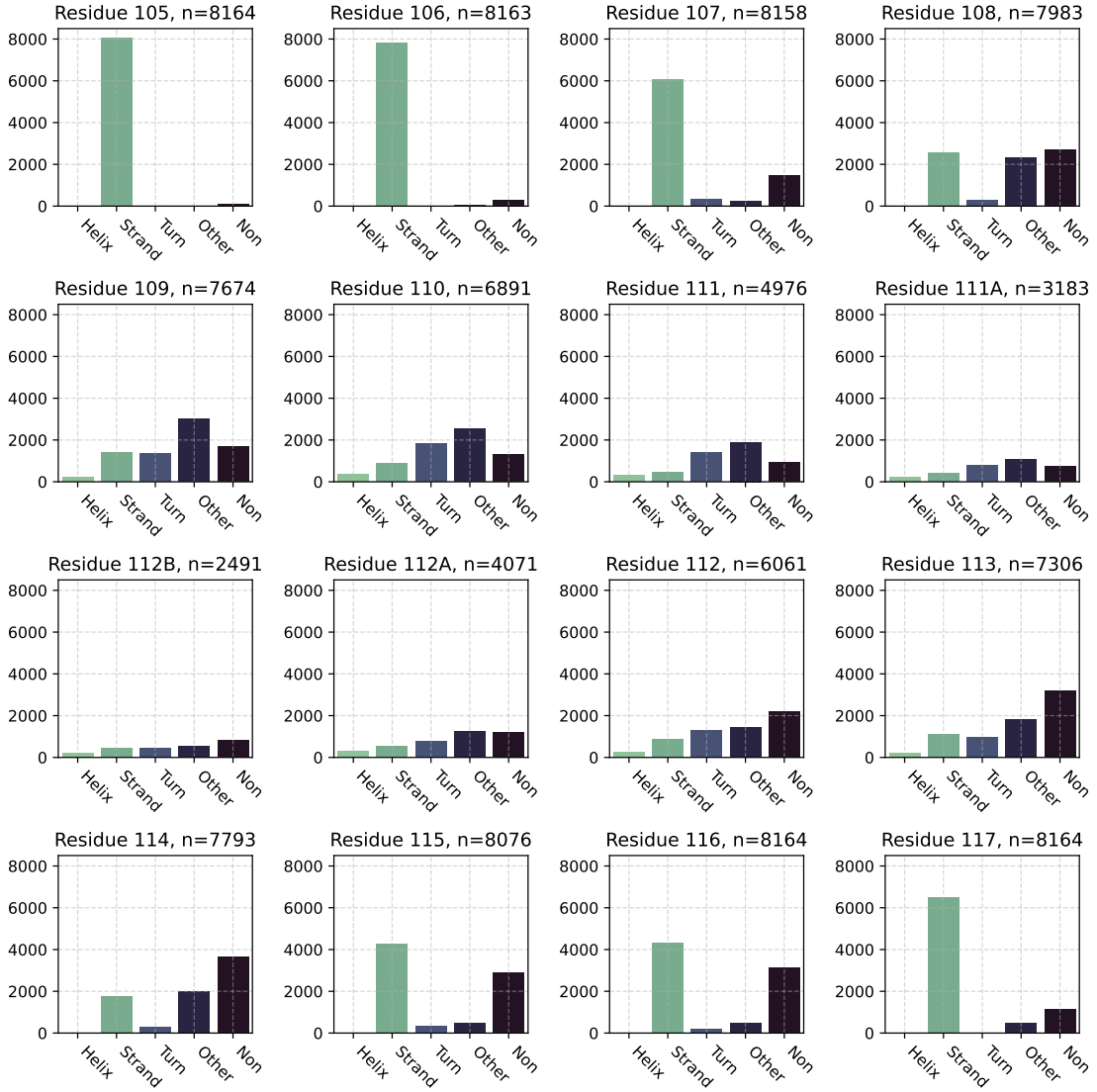


Figure C.3: Secondary structure elements in CDRH3 loops The figure shows the type and number of secondary structures present in each IMGT-defined position (one per plot) in CDRH3, as characterised by CV3-Fv in over 3000 experimental structures.

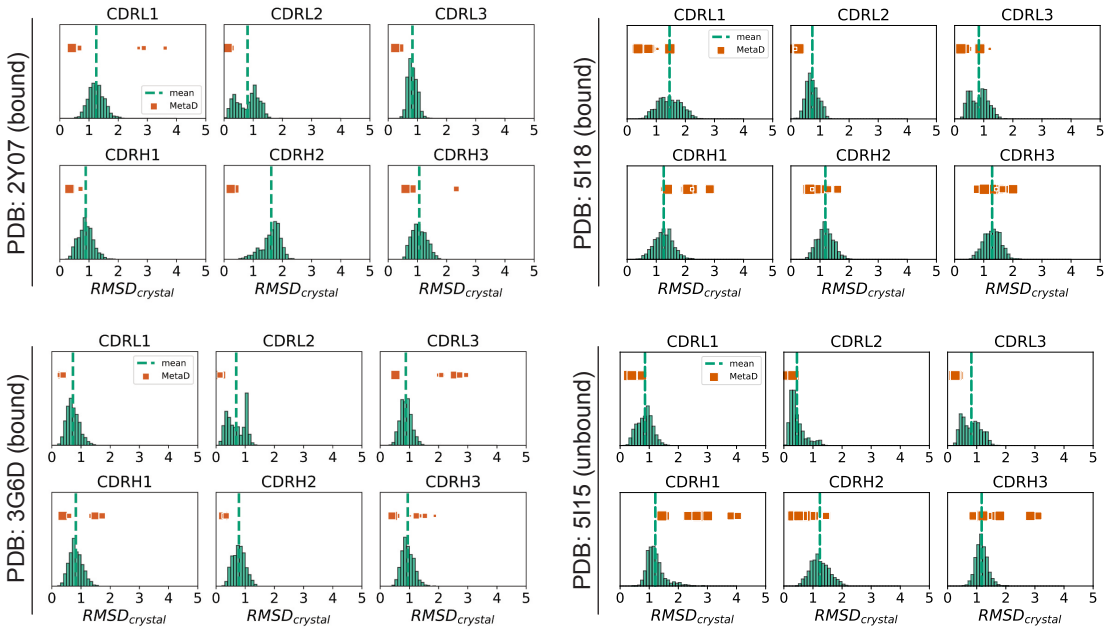


Figure C.4: CDR loop RMSD to crystal for CV3-Fv and metaD clusters. The figure shows the flexibility comparison using RMSD from initial structure between CV3-Fv simulations and the metaD cluster structures for four benchmark mAbs, reported here with their crystal structure PDB code. For each system, panels show the RMSD from initial structure distributions for each of the six CDRs. CV3-Fv simulations are shown as green distributions, and metaD cluster values are shown as red dots. The mean RMSD values are shown as green dashed vertical lines in all the plots.

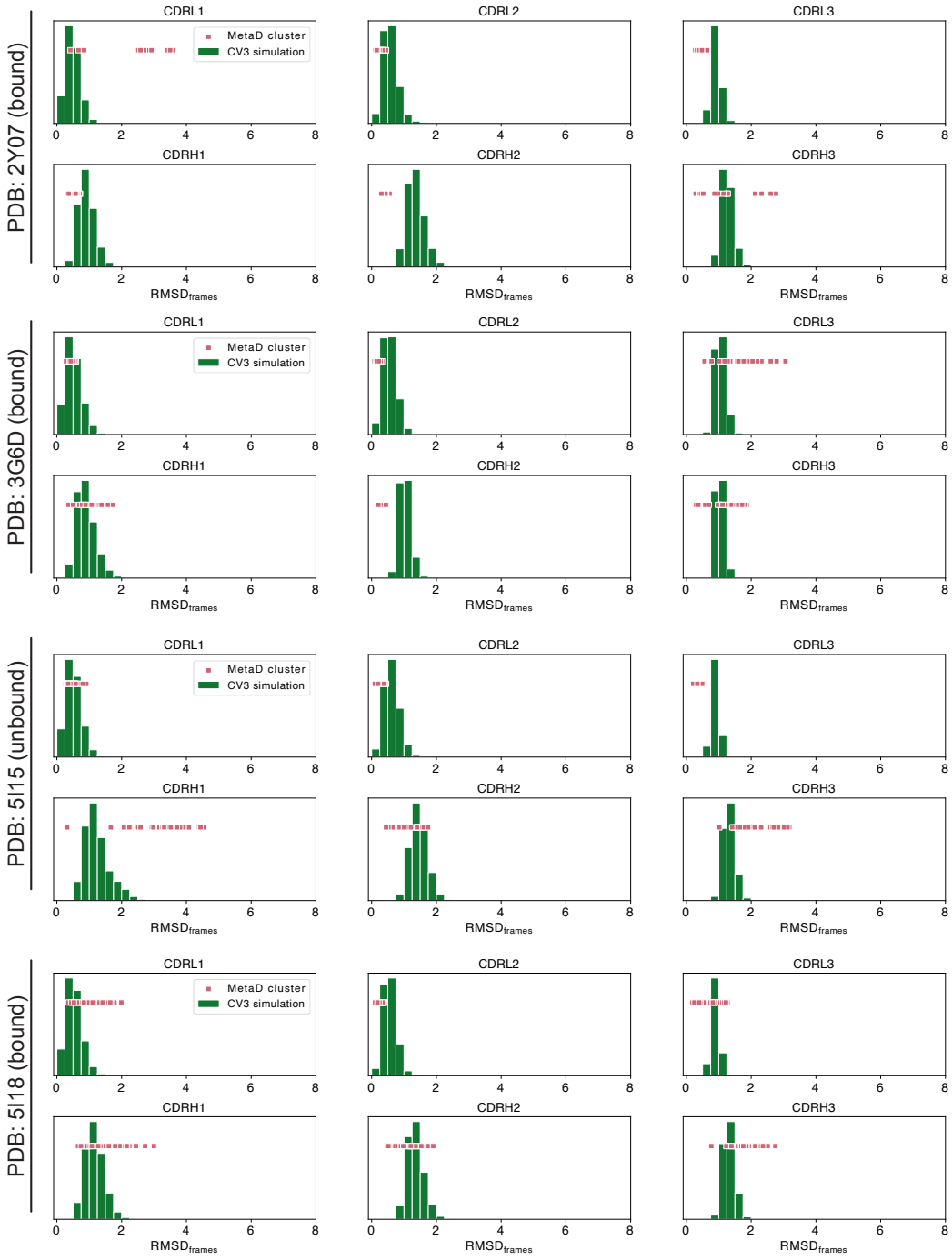


Figure C.5: CDR flexibility in CV3-Fv simulations The figure shows the flexibility comparison using RMSD from the mean simulation coordinates between CV3-Fv simulations and the metaD cluster structures for four benchmark mAbs, reported here with their crystal structure PDB code. For each system, each panel shows the RMSD from the mean simulation coordinate distributions for one of the six CDRs. CV3-Fv simulations are shown as green distributions, and metaD cluster values are shown as red dots. The mean RMSD values are shown as green dashed vertical lines in all the plots.

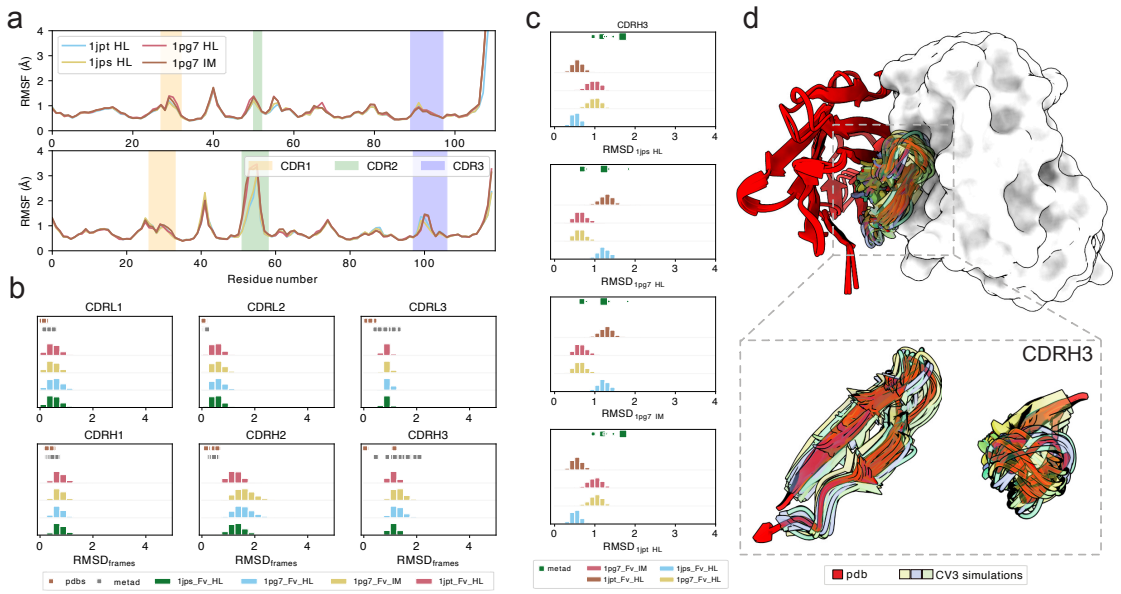


Figure C.6: Analysis of the flexibility prediction for the CV3-Fv simulations of the D3h44 Fv from different initial conformations. (a) RMSF values for each residue in the CV3-Fv simulations. Each curve corresponds to a simulation started from a different available crystal structure. Background colours highlight the different CDR loop regions. (b) Frame pairwise root-mean-square deviation (RMSD) distributions for all six CDR loops, comparing CV3-Fv simulations started from different crystal structures (green, yellow, light blue, and red) to metaD cluster structures (gray) and the crystal structures themselves (brown). (c) RMSD distributions from the starting crystal structures for CDRH3, comparing CV3-Fv simulations (coloured as in panel b) to metaD cluster structures (green). (d) Structural illustration of the Fv region. The light chain is shown as a white surface; heavy chains are shown as cartoons. The top panel displays the entire heavy chain from the crystal structure (red) and the CDRH3 loop from CV3-Fv simulations started from different crystal structures (yellow, purple, green). The bottom panel shows a zoomed-in view of the CDRH3 loop.

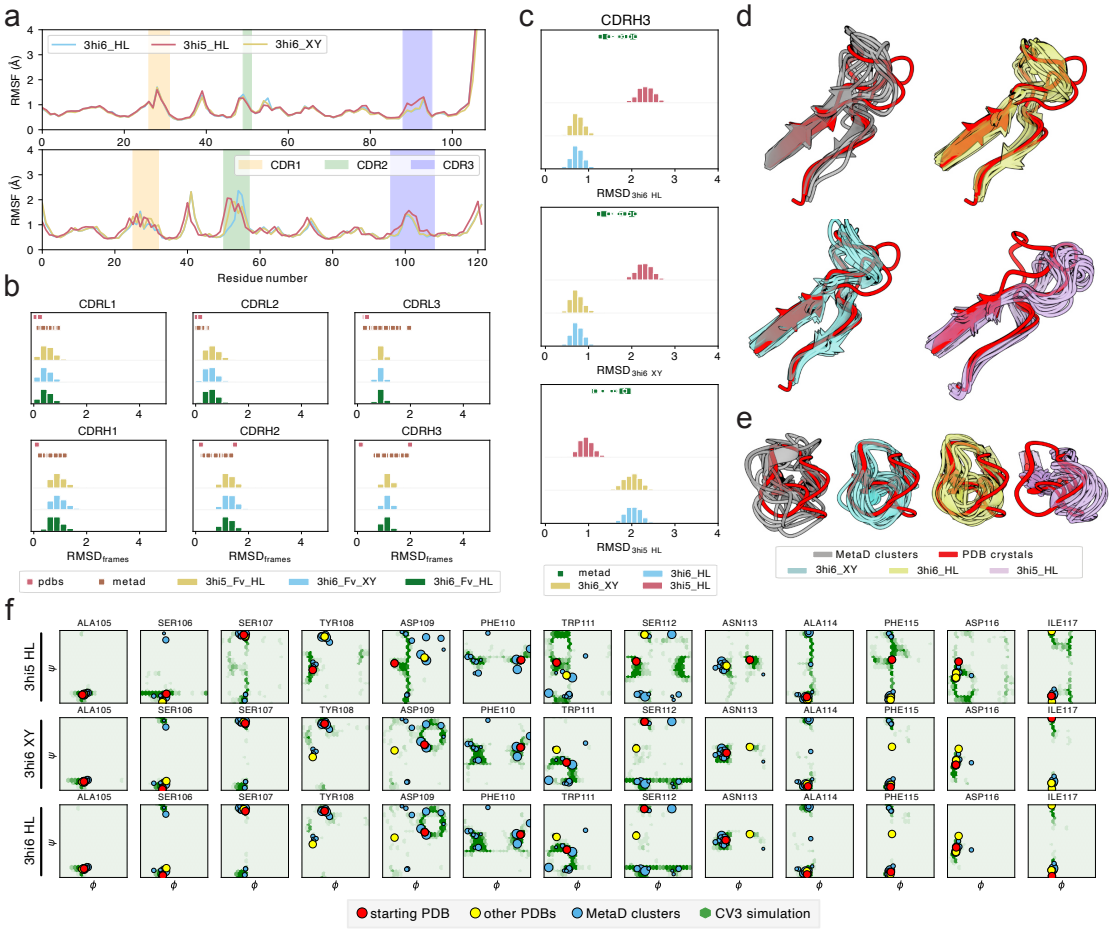


Figure C.7: Analysis of the flexibility prediction for the CV3-Fv simulations of the AL-57 Fv from different initial conformations. (a) RMSF values for each residue in the CV3-Fv simulations. Each curve corresponds to a simulation started from a different available crystal structure. Background colours highlight the different CDR loop regions. (b) Frame pairwise root-mean-square deviation (RMSD) distributions for all six CDR loops, comparing CV3-Fv simulations started from different crystal structures (green, yellow, light blue, and red) to metaD cluster structures (brown) and the crystal structures themselves (gray). (c) RMSD distributions from the different starting crystal structures for CDRH3, comparing CV3-Fv simulations initiated from each structure (green, yellow, light blue, and red) to metaD cluster structures (green). (d,e) Structural ensembles of the CDRH3 loop from CV3-Fv simulations and metaD cluster structures. The starting crystal structure is shown in red. Panels (d) and (e) display top and side views of the ensembles, respectively. (f) Torsional backbone space for each residue in CDRH3. Green hex-bin plots show the torsion angles sampled by each residue in our CV3-Fv simulations. Each row corresponds to a different simulation, initiated from a different crystal structure. In each plot, red dots indicate the torsion angles of the starting crystal structure, yellow dots represent alternative crystal structures, and blue dots denote metaD cluster structures.

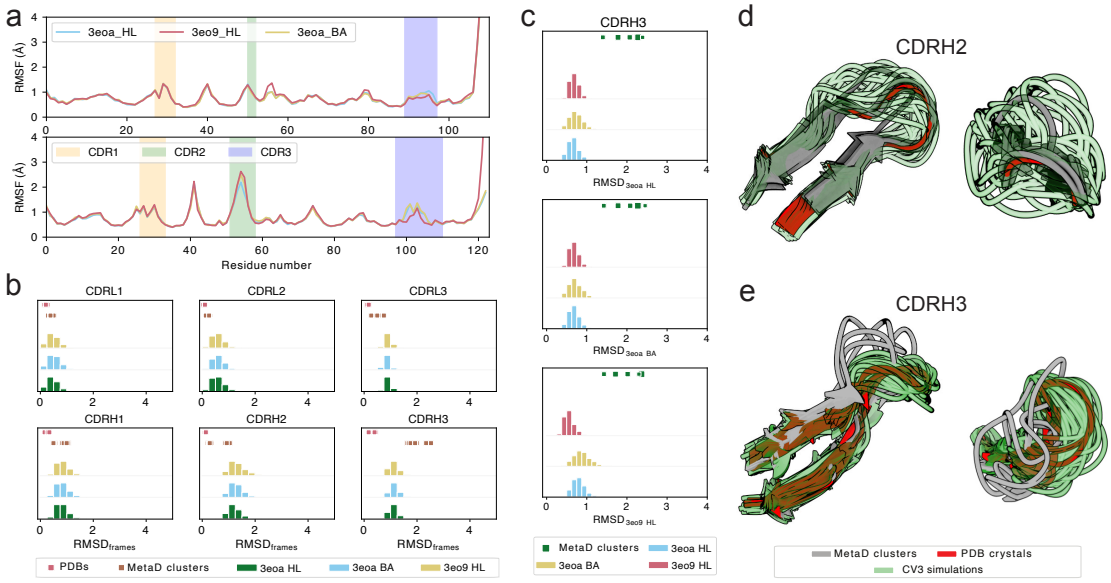


Figure C.8: Analysis of the flexibility prediction for the CV3-Fv simulations of the Efalizumab Fv from different initial conformations. (a) RMSF values for each residue in the CV3-Fv simulations. Each curve corresponds to a simulation started from a different available crystal structure. Background colours highlight the different CDR loop regions. (b) Frame pairwise root-mean-square deviation (RMSD) distributions for all six CDR loops, comparing CV3-Fv simulations started from different crystal structures (green, yellow, light blue, and red) to metaD cluster structures (brown) and the crystal structures themselves (red). (c) RMSD distributions from the starting crystal structures for CDRH3, comparing CV3-Fv simulations (coloured as in panel b) to metaD cluster structures (green). (d,e) Representative structural ensembles of the CDRH2 and CDRH3 loops from CV3-Fv simulations (green) and metaD cluster structures. The starting crystal structure is shown in red. Panels on the left and right display top and side views of the ensembles, respectively.

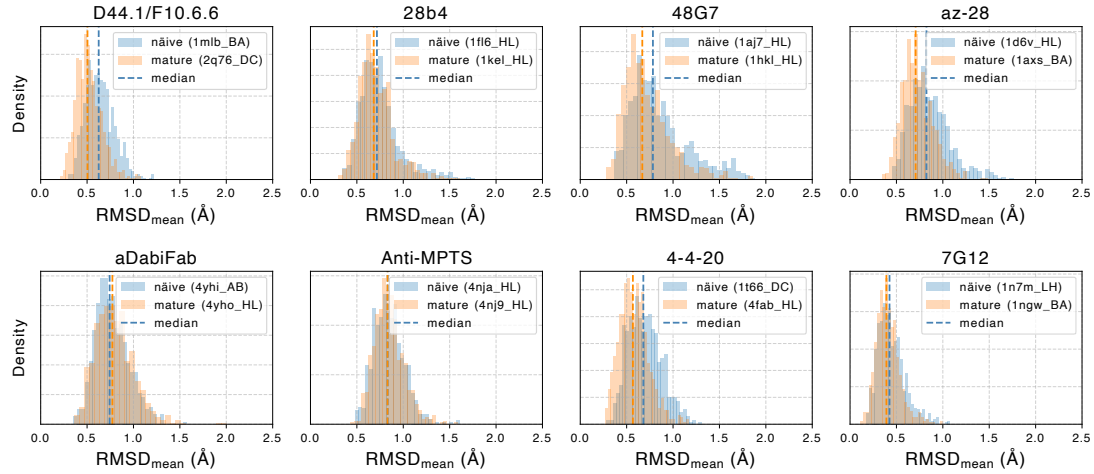


Figure C.9: Rigidification of antibody lineages during maturation. Comparison of naive and mature antibodies from the same lineage. Each plot shows the root mean square deviation (RMSD) from the mean coordinates of each frame for the paired naïve (light blue) and mature (orange) antibodies of different lineage types. The vertical lines indicate the median RMSD value.

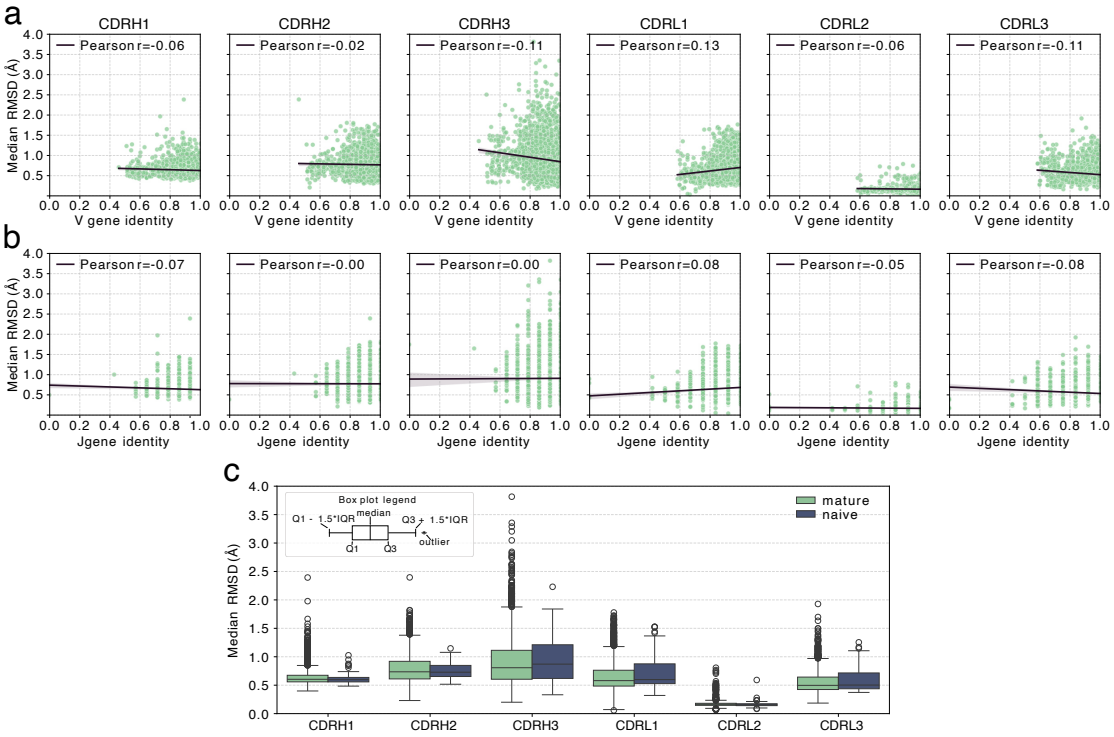


Figure C.10: CDR flexibility and antibody maturation. Median CDR $\text{RMSD}_{\text{mean}}$ plotted against (a) V and (b) J gene identity to the corresponding germline. A linear regression was fit to the data (black line) and the 95% confidence interval is shown in grey shading. Pearson's correlation coefficient is reported in each panel. c) Distribution of $\text{RMSD}_{\text{mean}}$ of CDRs in naive (blue) and mature (green) antibody chains. Antibody heavy and light chains were separately classified as naive or mature. The box plot shows samples of 57 naive and 2467 mature heavy chain CDRs (CDRH1-3) and samples of 117 naive and 2020 mature light chain CDRs (CDRL1-L3).

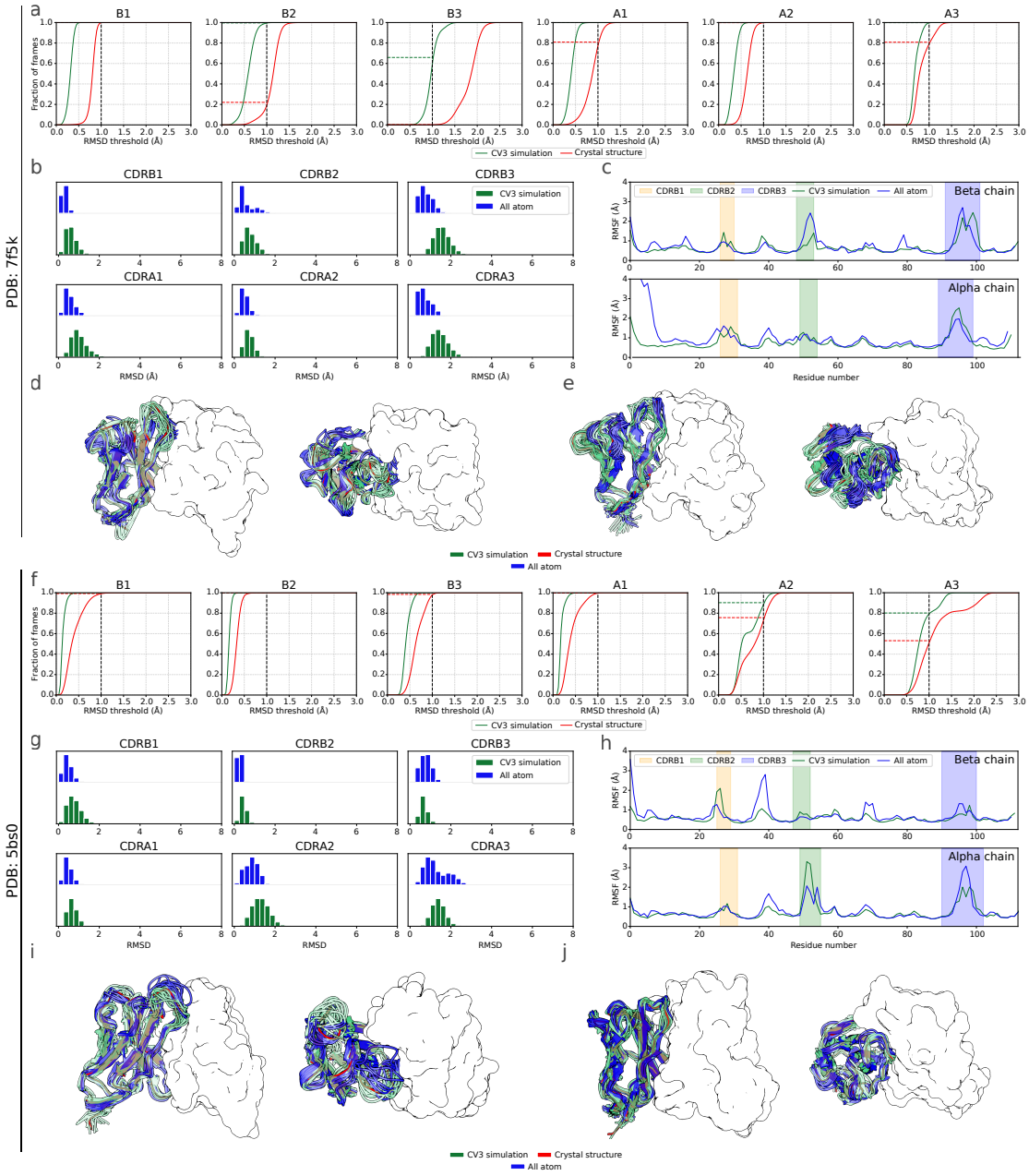


Figure C.11: Benchmarking the accuracy of CV3-Fv simulations for TCRs against all-atom simulations. Comparison of the dynamical properties of the CDR loops for two TCRs: PDB 7f5k and PDB 5bs0. (a,e) Frame coverage showing the proportion of all atom frames covered at a given threshold by any CV3-Fv frame (green) and the starting crystal structure (red). A black dashed line indicates the 1Å threshold. (b, f) Frame pairwise RMSD distribution for all six CDR loops, comparing CV3-Fv (green) and all-atom simulations (blue). (c, g) Per-residue RMSF values from CV3-Fv (green) and all-atom (blue) simulations. Shaded backgrounds mark different CDR loops. (d, i) Structural ensembles of the Fv region. Alpha chains are shown as white surfaces and beta chains as cartoons. For the beta chain CV3-Fv (green) and all-atom (blue) representative structures are overlaid on top of the starting crystal structure (red). Left and right panels show side and top views, respectively.

Figure C.11 (continued). (e, j) Structural ensembles of the Fv region. Beta chains are shown as white surfaces and alpha chains as cartoons. For the alpha chain CV3-Fv (green) and all-atom (blue) representative structures are overlaid on top of the starting crystal structure (red). Left and right panels show side and top views, respectively.

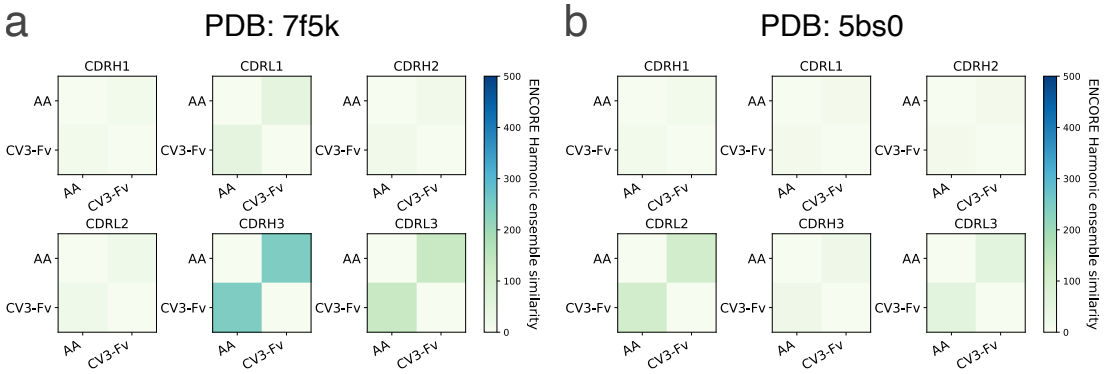


Figure C.12: Benchmarking the accuracy of CV3-Fv simulations for TCRs against all-atom simulations with Harmonic Ensemble Similarity (HES). (a,b) HES matrices obtained from the ENCORE package for each CDR in 7f5k (a) and 5bs0 (b), comparing all-atom simulations (AA) and CV3-Fv (Fv). The lower the HES score, the higher the agreement between the two ensembles.

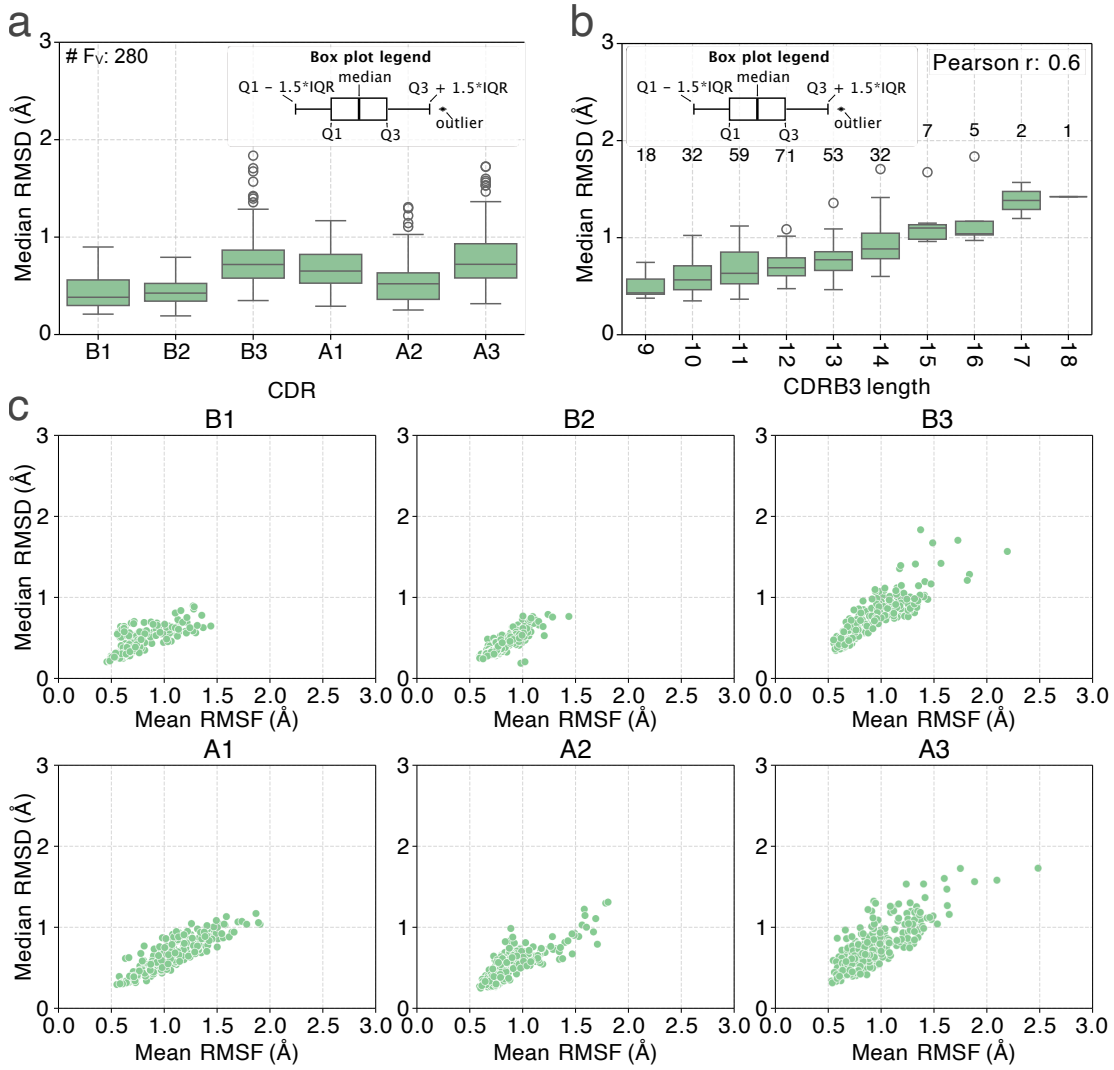


Figure C.13: CDR dynamics of TCRs with experimental structures. CV3-Fv simulations were run for 280 non-redundant and quality filtered TCR structures from the STCRDab. a) Box plot showing the median RMSD across all frames to the average loop structure ($RMSD_{mean}$) for each of the CDRs. Each box plot shows values for 280 samples. b) Box plot showing the median CDRH3 $RMSD_{mean}$ as a function of CDR length. A Pearson's correlation of RMSD against CDRH3 length was run and the correlation coefficient indicated. The number of samples is reported on the top of each box plot. c) Scatter plot showing the median $RMSD_{mean}$ of each loop against the median residue-level RMSF.

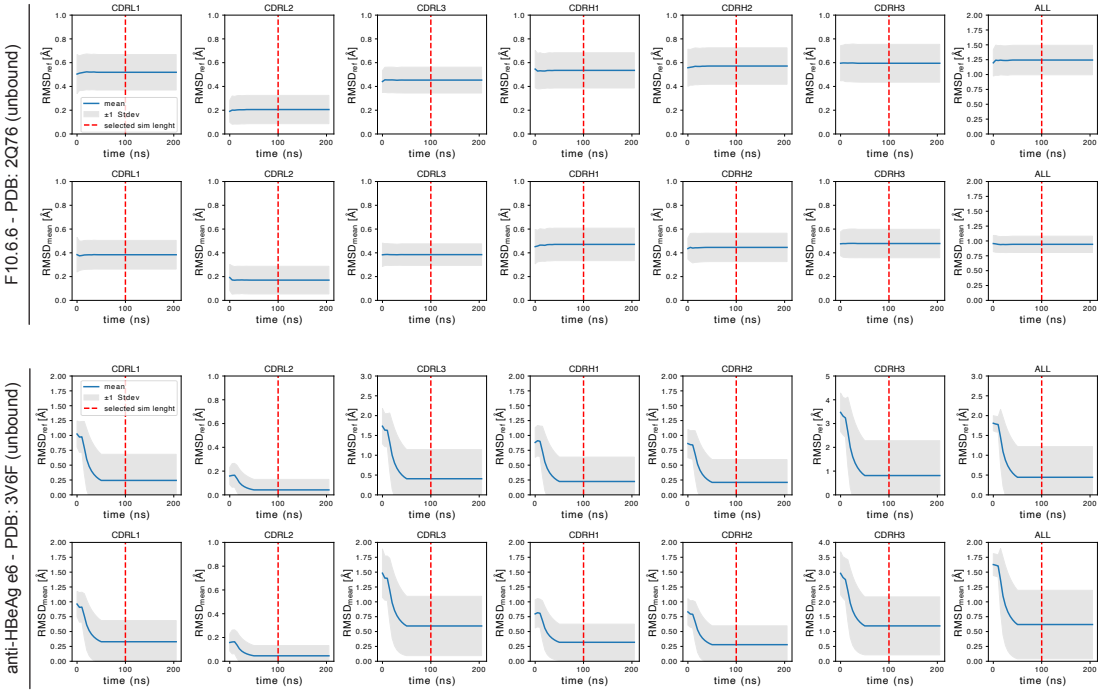


Figure C.14: Convergence curves for coarse-grained simulations of benchmark antibodies. For each antibody system, the upper panels show the cumulative mean RMSD relative to the starting structure for each CDR, while the lower panels show the cumulative mean RMSD relative to the mean coordinates. Each simulation replica is represented by a different colour, and shaded areas indicate the standard deviation around the cumulative mean. The vertical red threshold line shows the final simulation length selected for all simulated systems.

D

Sampling conformational ensembles with deep learning

D.1 Supplementary methods

D.1.1 ABB4 model details

Loss functions

ABB4 is trained using a loss composed of five terms. The model predicts clean backbone frames $\hat{T}_1 = (\hat{x}_1, \hat{r_1})$ and torsion angles $\hat{\chi}$, which together are used to reconstruct all-atom coordinates $\hat{a}$. I optimised the model under an $SE(3)$ flow matching objective following the formulation of FrameFlow (Yim et al. 2023a), using separate losses on the translational and rotational components in $R(3)$ and $SO(3)$, respectively:

$$\mathcal{L}_{R(3)} = \mathbb{E}_{(T_0,T_1|s)\sim\pi_s^*,\, s\sim\mathcal{D},\, t\sim\mathrm{U}(0,1)} \left[\frac{1}{(1-t)^2} \sum_{i=1}^{N_{frames}} \left(\left\| \hat{x}_1^i(T_t, t, s) - x_1^i \right\|_{\mathbb{R}^3}^2 \right) \right], \quad \text{(D.1)}$$

$$\mathcal{L}_{SO(3)} = \mathbb{E}_{(T_0,T_1|s)\sim\pi_s^*,\, s\sim\mathcal{D},\, t\sim\mathrm{U}(0,1)} \left[\frac{1}{(1-t)^2} \sum_{n=1}^{N_{frames}} \left(\left\| \log_{r_t^i}\left(\hat{r}_1^i(T_t, t, s)\right) - \log_{r_t^i}\left(r_1^i\right) \right\|_{SO(3)}^2 \right) \right], \quad \text{(D.2)}$$

where π_s^* denotes the optimal-transport coupling (for details see SI Section D.1.1), $U(0,1)$ denotes the uniform distribution over time and $log_{r_t}(r_1)$ is the Riemannian logarithm on $SO(3)$, corresponding to the tangent-space geodesic direction at r_t pointing toward $\hat{r}_1$.

To stabilise the flow over backbone frames and to introduce side chain prediction capabilities, I included auxiliary losses on the reconstructed all-atom coordinates $\hat{a}$ and torsion angles $\hat{\chi}$. First, I used the frame-aligned point error (FAPE) (Jumper et al. 2021):

$$\mathcal{L}_{\mathrm{FAPE}} = \mathbb{E}_{(T_0,T_1|s)\sim\pi_s^*,\, s\sim\mathcal{D},\, t\sim\mathrm{U}(0,1)} \left[\frac{1}{N_{\mathrm{frames}} \times N_{\mathrm{atoms}}} \sum_{i=1}^{N_{\mathrm{frames}}} \sum_{j=1}^{N_{\mathrm{atoms}}} \min\left(\left\| \hat{T}_1^{i^{-1}}(T_t, t, s) \circ \hat{a}_1^j(T_t, t, s) - T_1^{i^{-1}} \circ a_1^j \right\|_{\mathbb{R}^3}^2,\ d_{\mathrm{clamp}} \right) \right], \quad \text{(D.3)}$$

where T^{-1} denotes the inverse rigid transformation and d_{clamp} clamps the deviation at 10Å.

Following AlphaFold2 (Jumper et al. 2021), torsion angles χ are parametrised as points on the unit circle, $\chi_n^i \in \mathbb{R}^2$ with $\|\chi^{i,n}\| = 1$. The torsion parametrisation exhibits a 180°-rotation-symmetry, meaning that each torsion angle corresponds to two equivalent unit-circle vectors. AlphaFold2 supervises the prediction against either of these two symmetric targets. I simplified the torsion loss by supervising only a single representative orientation, which I found does not affect performance.

The torsion angle loss is

$$\mathcal{L}_\chi = \mathbb{E}_{(T_0,T_1|s)\sim\pi_s^*,\, s\sim\mathcal{D},\, t\sim\mathrm{U}(0,1)}\left[\frac{1}{N_{\text{frames}} \times N_{\text{torsions}}} \sum_{i=1}^{N_{\text{frames}}} \sum_{n=1}^{N_{\text{torsions}}} \left(\left\| \frac{\hat{\chi}^{i,n}(T_t,t,s)}{\|\hat{\chi}^{i,n}(T_t,t,s)\|} - \chi^{i,n} \right\|^2 \right)\right], \tag{D.4}$$

and I additionally regularised the predicted torsion-vector norm to encourage predictions to remain on the unit circle:

$$\mathcal{L}_{\|\chi\|} = \mathbb{E}_{(T_0,T_1|s)\sim\pi_s^*,\, s\sim\mathcal{D},\, t\sim\mathrm{U}(0,1)}\left[\frac{1}{N_{\text{frames}} \times N_{\text{torsions}}} \sum_{i=1}^{N_{\text{frames}}} \sum_{n=1}^{N_{\text{torsions}}} \left(\left| \|\hat{\chi}^{i,n}(T_t,t,s)\| - 1 \right| \right)\right]. \tag{D.5}$$

RMSF ensemble loss

In addition to per-structure losses, I introduced an ensemble loss with the objective to match the RMSF of ensembles generated by the model to the RMSF observed in MD simulation directly. Specifically, for each protein sequence s, I predict an ensemble $\hat{E} = \{\hat{T}_1^1, \ldots, \hat{T}_1^K\}$ of K structures and compute RMSF of C_α atoms across the ensemble. I then penalise the deviation between the predicted RMSF and the reference (MD) RMSF using a normalized χ^2-style loss. This gives the objective

$$\hat{E}(E_T, t, s) = \{\hat{x}_1^1(T_t^1, t, s)), \ldots, \hat{x}_1^K(T_t^K, t, s))\}, \tag{D.6}$$

$$\mathcal{L}_{ensemble} = \mathbb{E}_{(E_0,E_1|s)\sim\pi_s^*,\, s\sim\mathcal{D},\, t\sim\mathrm{U}(0,1)}\left[\frac{1}{N_{\text{frames}}}\sum_{i=1}^{N_{\text{frames}}} \left(\frac{\left\|RMSF^i(\hat{E}(E_T, t, s)) - RMSF^i(E)\right\|^2}{(\sigma^i)^2}\right)\right], \tag{D.7}$$

where x is the backbone frame translation which corresponds to the C_{alpha} coordinate, $\mathrm{RMSF}^i(\cdot)$ denotes the root-mean-square fluctuation of residue i across the ensemble, and $(\sigma^i)^2$ is a per-residue variance normalisation term. In the experiments I set $\sigma^i = 0.1 \times \mathrm{RMSF}^i$. Using a χ^2-style (variance-normalised) loss rather than a standard MSE stabilises training by accounting for heteroscedastic fluctuation amplitudes across different residues, so that highly flexible regions do not dominate the loss.

Global sequence-conditional optimal transport

In this setting, there are two probability distributions of noise $p(T_0)$ and protein structures $p(T_1)$. The goal of optimal transport (OT) is to move mass according to $p(T_0)$ so that it becomes distributed as $p(T_1)$, while minimising a given cost of movement $c(x, y)$. A transport plan π is a joint distribution over (x, y) with marginals $p(T_0)$ and $p(T_1)$ and I denote the set of all such couplings by $\Pi(p(T_0), p(T_1))$. Formally, the OT problem is given by (Villani 2009):

$$\pi^* = \arg\min_{\pi\,\in\,\Pi(p(T_0),\,p(T_1))} \mathbb{E}_{(x,y)\sim\pi}\Big[c(x, y)\Big]. \tag{D.8}$$

By constructing simpler integration paths, optimal transport leads to faster and more stable training and was shown to improve performance on certain tasks (Bose et al. 2024). In recent protein structure design models, OT couplings were computed unconditionally $(p(T_0), p(T_1)) \sim \pi^*$ and couplings were approximated over a mini-batch to limit the computational cost Bose et al. 2024; Yim et al.

2023a. In protein structure prediction, structures are sampled conditional on a sequence $p(T_1 \mid s)$ in contrast to protein design where I sample from an unconditional distribution $p(T_1)$. I therefore construct a sequence-conditional optimal coupling $(p(T_0), p(T_1) \mid s) \sim \pi_s^*$. As the sequence conditionality reduces the number of samples over which the coupling is calculated, I can determine the exact global OT rather than approximating over a mini-batch. Importantly, this computation can be done outside the training loop, so it does not add overhead during training.

Concretely: let $(T_1^1, \ldots, T_1^K)$ denote all available structures (conformations) for sequence s. I sample translation noise $x_0^k \sim \mathcal{N}(0, I_3)^N$ and rotation noise $r_0^k \sim \mathcal{U}(SO3)^N$, then compute the cost matrix $c^{i,j}(T_0^i, T_1^j)$ using Euclidean distance after alignment in $R(3)$

$$x^i_{0,aligned} = align(x^i_0, x^i_1) \tag{D.9}$$

$$c^{i,j}_{R(3)} = \|x^i_1 - x^j_{0,aligned}\|, \tag{D.10}$$

and the relative angle in $SO(3)$

$$r^{i,j}_{rel} = r^i_1 (r^j_0)^T \tag{D.11}$$

$$c^{i,j}_{SO(3)} = \arccos\Big(\frac{\mathrm{Tr}(r^{i,j}_{\mathrm{rel}}) - 1}{2}\Big), \tag{D.12}$$

summing to a total cost of

$$c^{i,j}_{SE(3)} = c^{i,j}_{R(3)} + 10 \times c^{i,j}_{SO(3)}, \tag{D.13}$$

with weights chosen to balance the magnitude of both terms (Bose et al. 2024).

By contrast to FoldFlow and FrameFlow, which sample initial rotations from the isotropic-Gaussian noise distribution ($r_0^i \sim \mathcal{IG}(SO3)^N$), I sample from a uniform distribution ($r_0^i \sim \mathcal{U}(SO3)^N$). In FrameFlow, sampling from $\mathcal{IG}(SO3)^N$ was described to improve performance. I speculate that this improvement arises

because sampling from $\mathcal{IG}(SO3)^N$ effectively ensures that rotational noise is always optimally coupled. As FoldFlow only computes OT coupling over $R(3)$ explicitly, this sampling results in a separate OT coupling over $SE(3)$. This may lead to undesirable behaviour during inference when rotational and translational noise components are sampled that were coupled to different output structures during training. Under the formulation to perform OT jointly over $SE(3)$, sampling from $\mathcal{IG}(SO3)^N$ or $\mathcal{U}(SO3)^N$ is equivalent.

D.2 Supplementary results

D.2.1 Sampling of coarse-grained MD frames.

Simulations in FlAbDab produced with CV3-Fv (Cagiada et al. 2025b) consist of 1000 trajectory frames. As it is computationally expensive to work with such a large number of frames, I used a set of representative frames for model training and parts of the analysis. To determine the optimal number of frames to use as representatives, I computed how well randomly sampled ensembles of N frames reproduce ensemble metrics calculated across the entire trajectory. I observed that the error in RMSD and RMSF, when computed from randomly sampled frames compared to values across the entire trajectory, starts to converge at an ensemble size of $N = 20$ (Figure D.1 & D.2). To leave a margin for errors, I chose to sample 30 frames from each trajectory as a representative ensemble.

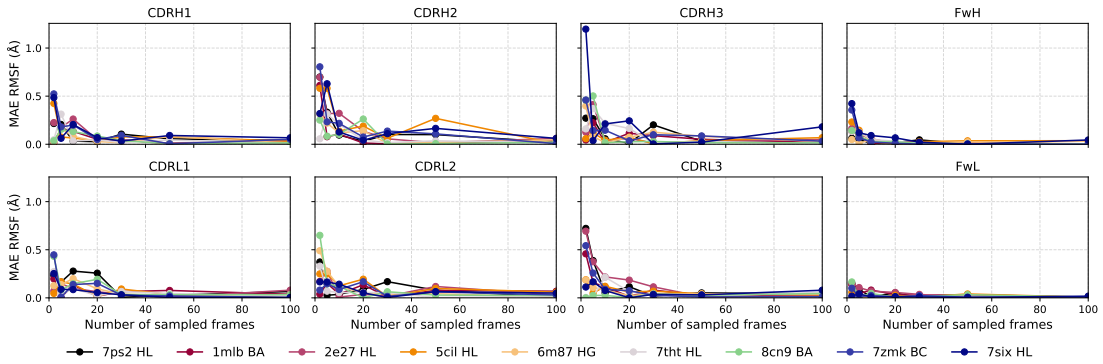


Figure D.1: Error in RMSD of sampled MD trajectory frames. The mean absolute error in RMSD of an ensemble of sampled frames compared to the entire MD trajectory is plotted against the number of frames sampled for each ensemble. Values are shown for 9 randomly sampled antibody systems.

D.2.2 Increased CDRH3 flexibility in FlAbDab simulations started from predicted structural models

FlAbDab (Cagiada et al. 2025b) contains a set of simulations started from experimental structures and a much larger set of simulations started from predicted structural models with ABodyBuilder2 (Abanades et al. 2023). To assess the quality of these simulations, I compared the average RMSF and RMSD of the

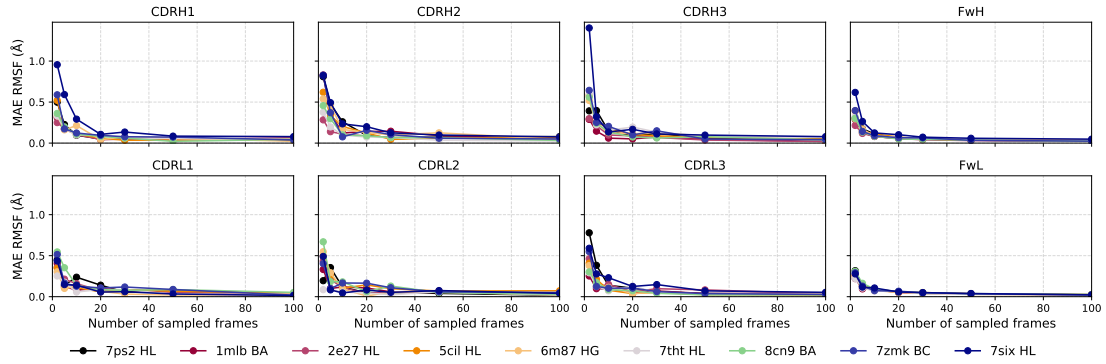


Figure D.2: Error in RMSF of sampled MD trajectory frames. The mean absolute error in RMSF of an ensemble of sampled frames compared to the entire MD trajectory is plotted against the number of frames sampled for each ensemble. Values are shown for 9 randomly sampled antibody systems.

two sets (Figure D.3) and observed that the flexibility of CDRH3s in simulations started from predicted structural models tends to be increased. This could be linked to findings that ABB2 predicts CDRH3s with increased solvent exposure (Raybould et al. 2019) which might in turn lead to reduced steric constraints and increased flexibility in simulations.

When training ABB4-STEROIDS in stage 2, I observed that a reduction in the training loss does not correspond to a better reproduction of CDRH3 RMSD and RMSF values evaluated on the validation and test sets. I linked this finding to the increased flexibility of CDRH3s of antibodies in the stage 2 training set (mainly simulations started from predicted models) compared to the validation and test set (only containing high quality simulations started from experimental structures).

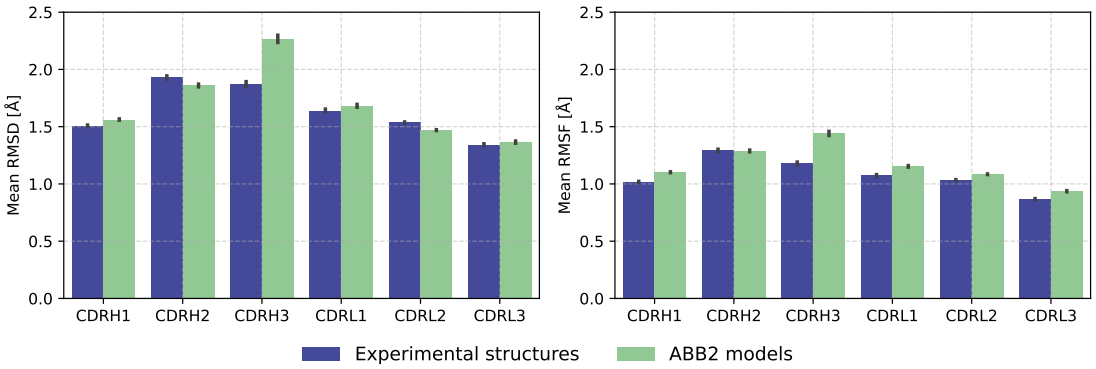


Figure D.3: Comparison of CDR flexibility in FlAbDab simulations started from experimental and predicted structural models. Flexibility was assessed in all 4220 simulations started from experimental structures and a sample of 4220 randomly sampled simulations started from structural models. Comparison of mean RMSD values (left) and mean RMSF values (right) for all CDR loops.

D.3 Supplementary figures

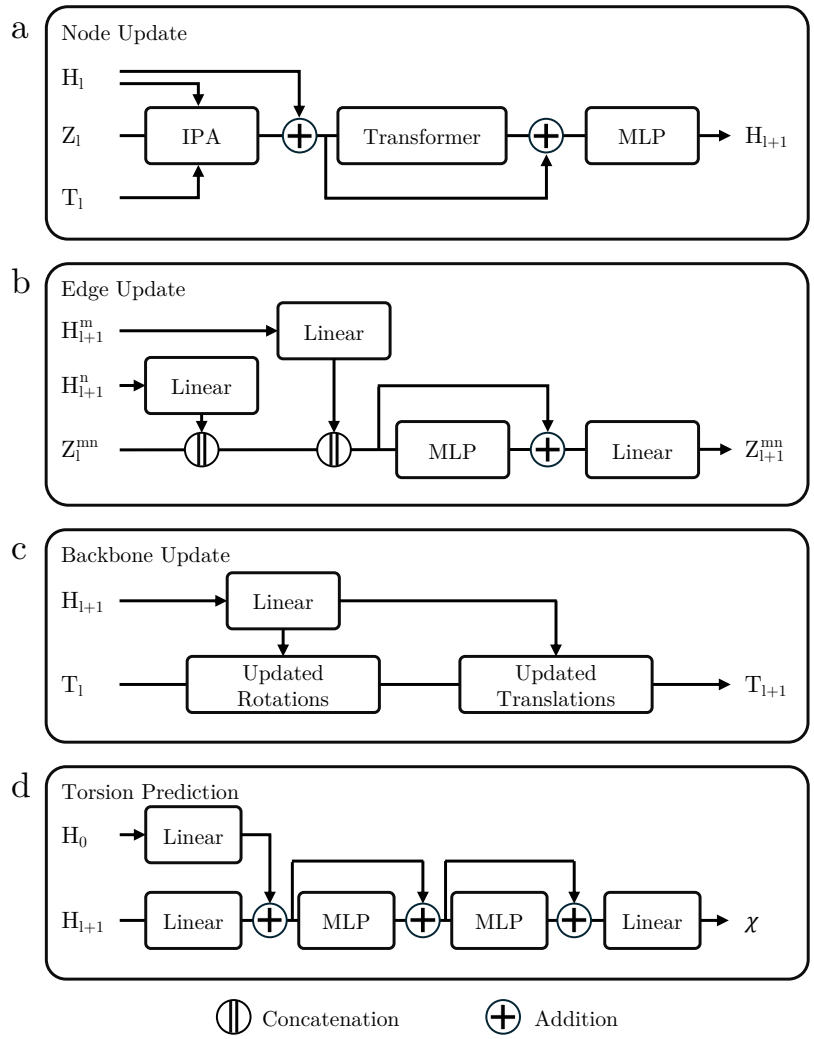


Figure D.4: Detailed architecture of the ABB4 models. ABB4 consists of four main blocks: a node update block (a), an edge update block (b), a backbone update block (c) and a torsion prediction block (d). Node features H, edge features Z and backbone frames T get updated and torsion angles χ are predicted in the corresponding blocks. Subscripts denote the layer l of the model and superscripts the residue indices. IPA indicates an invariant point attention module, MLP a multilayer perceptron and Transformer a transformer encoder module. Updated translations and rotations indicate direct operations on the translation vectors and rotation matrices.

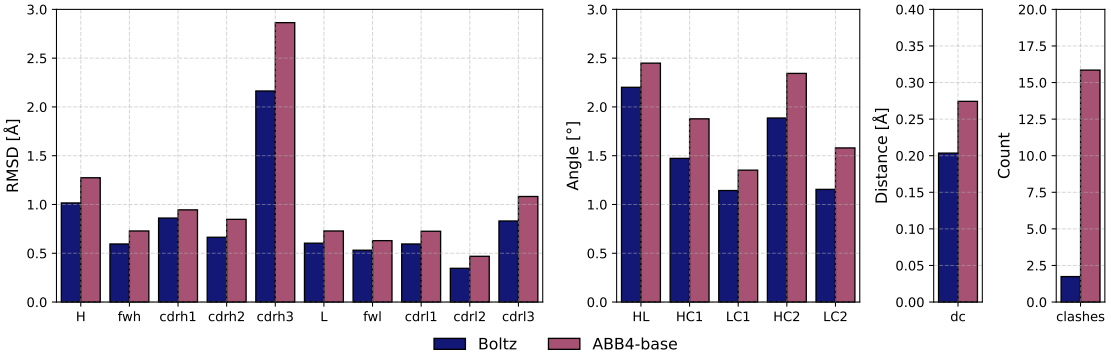


Figure D.5: Evaluation of ABB4-base single structure prediction against Boltz-1. Single structure prediction was evaluated on a test set of 100 antibody structures. Predictions were assessed using RMSD of antibody regions (H: heavy chain, fwh: heavy chain framework, L: light chain, fwl: light chain framework), deviations in VH-VL orientation angles (HL, HC1, LC1, HC2, LC2) and distance (dc) as defined in (Dunbar et al. 2013) and the number of atomic clashes defined as an overlap in van der Waals radii of any backbone or side chain atoms.

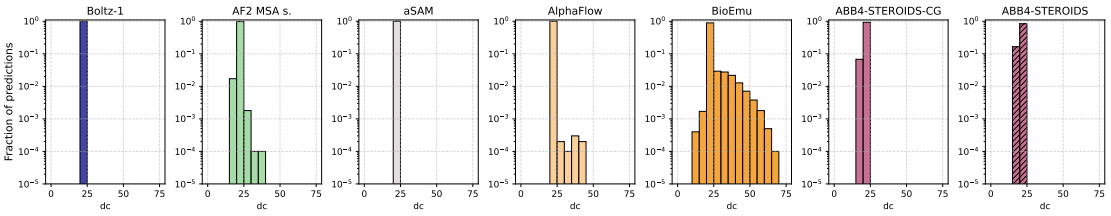


Figure D.6: Dissociation of VH and VL domains in predictions. Distributions of VH-VL distance (dc), defined as the distance between the centre of mass of VH Cα atoms and VL $C\alpha$ atoms). A distance larger than 30 is indicative of the VH and VL domains dissociating. Fractions are reported from a total of 100 predictions of the full test set of 100 antibodies (10,000 models in total).

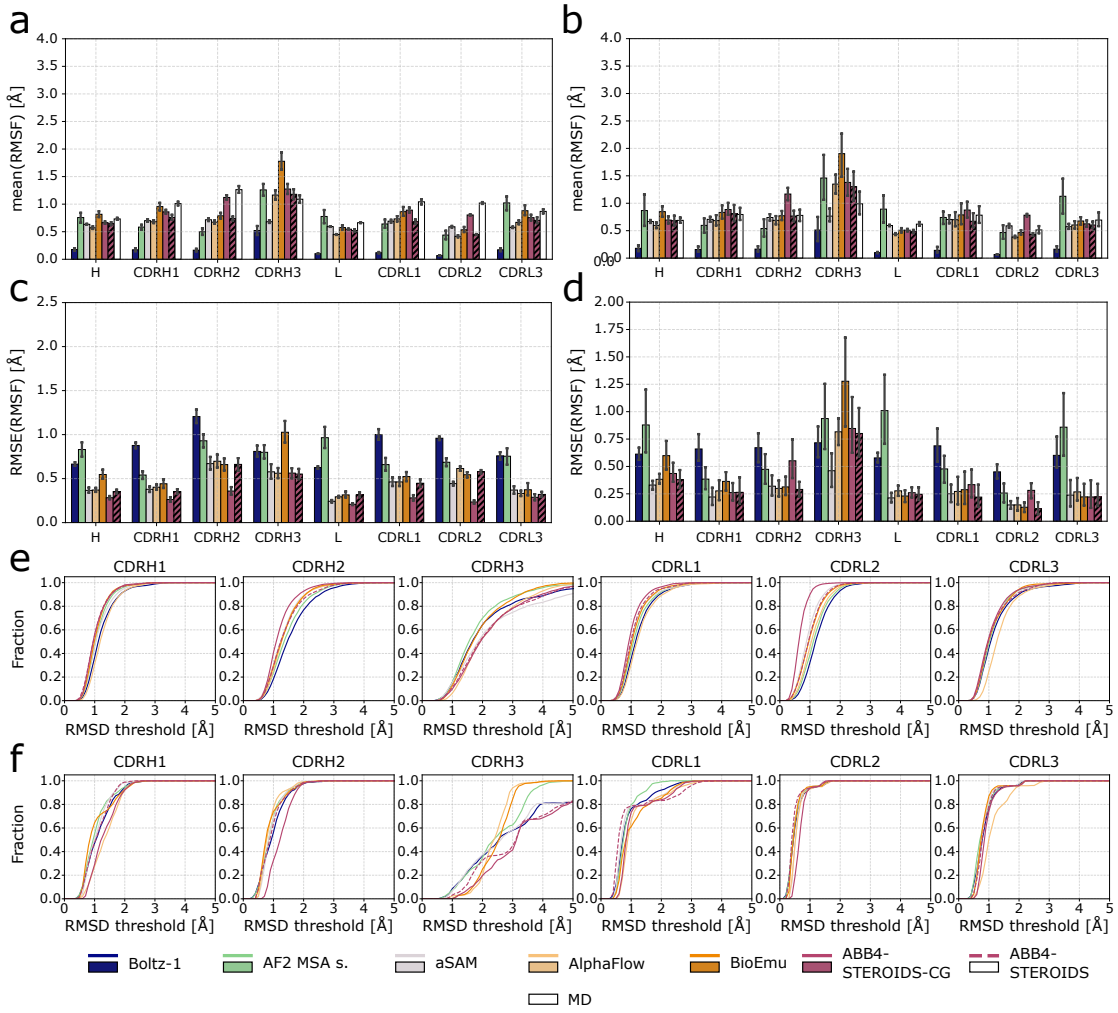


Figure D.7: Extended evaluation of predicted structural ensembles against MD data. Mean ensemble RMSF in predictions and MD trajectories (MD) across all antibodies in the coarse-grained MD (a) and all-atom (b) test sets. Values computed across the heavy chain (H), light chain (L) and individual CDRs. Root mean square error (RMSE) of ensemble RMSF in the predictions of the coarse-grained (c) and all-atom (d) test sets. e) Coverage of MD conformational frames as a function of distance threshold for the coarse-grained (e) and all-atom (f) test sets.

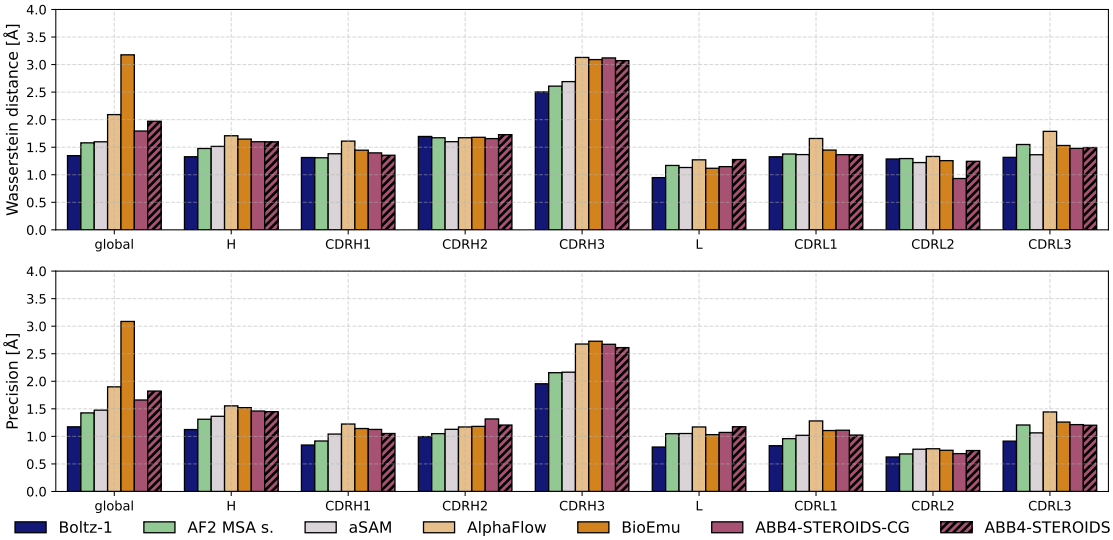


Figure D.8: Evaluation of predicted structural ensembles against MD trajectories using distance-based metrics. Wasserstein distance (a) and precision (b) between predicted ensembles and representative frames from MD simulations. Values computed across the entire antibody (global), heavy chain (H), light chain (L) and individual CDRs.

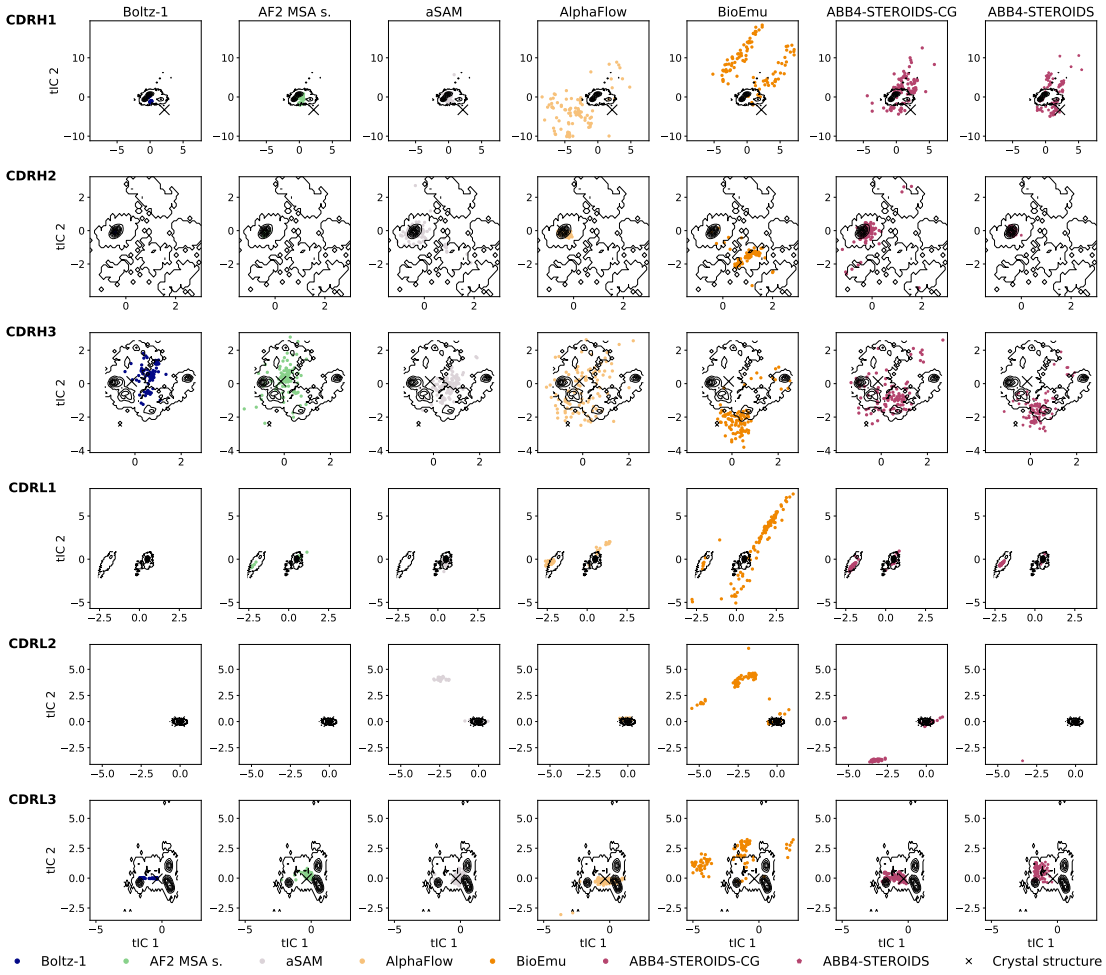


Figure D.9: Comparison of the conformational ensemble sampled for test set antibody PDB 8d01. The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross.

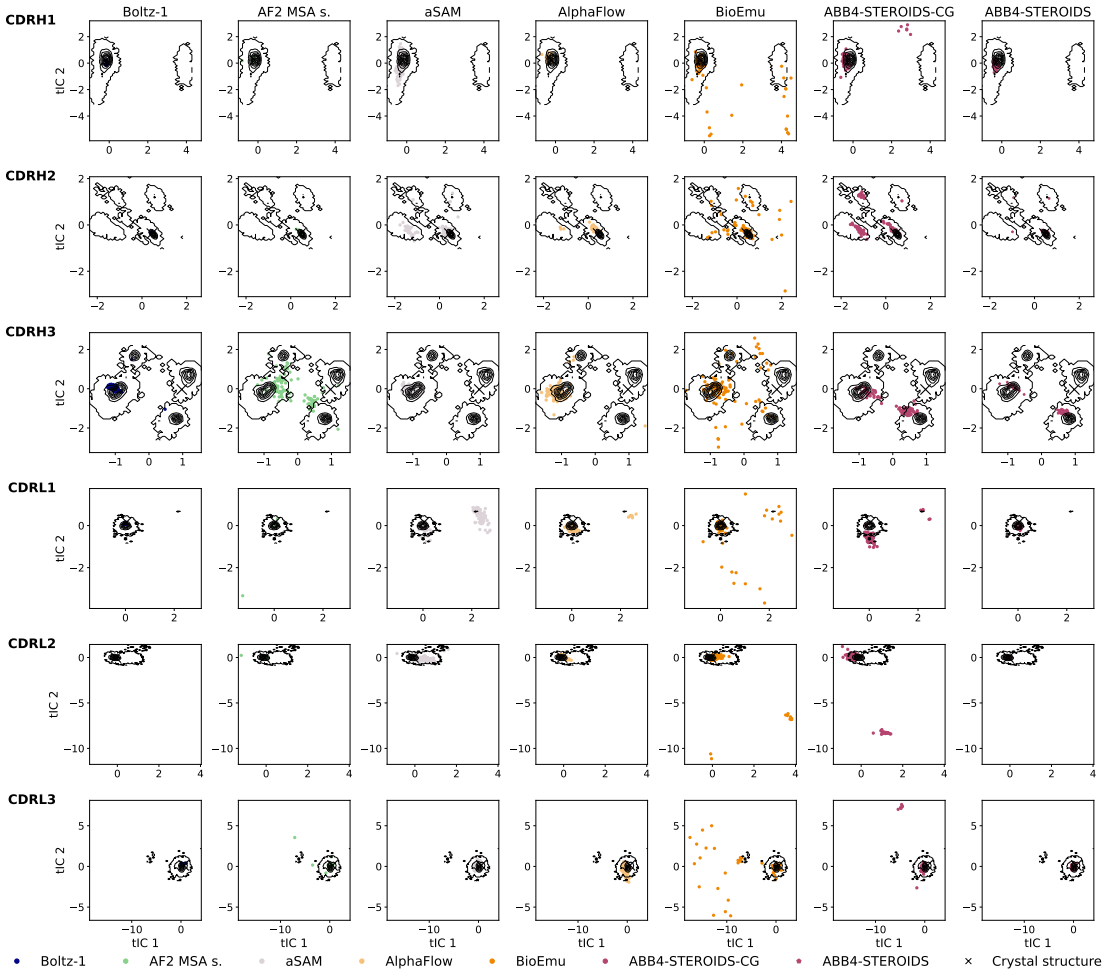


Figure D.10: Comparison of the conformational ensemble sampled for test set antibody PDB 8f2t. The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross

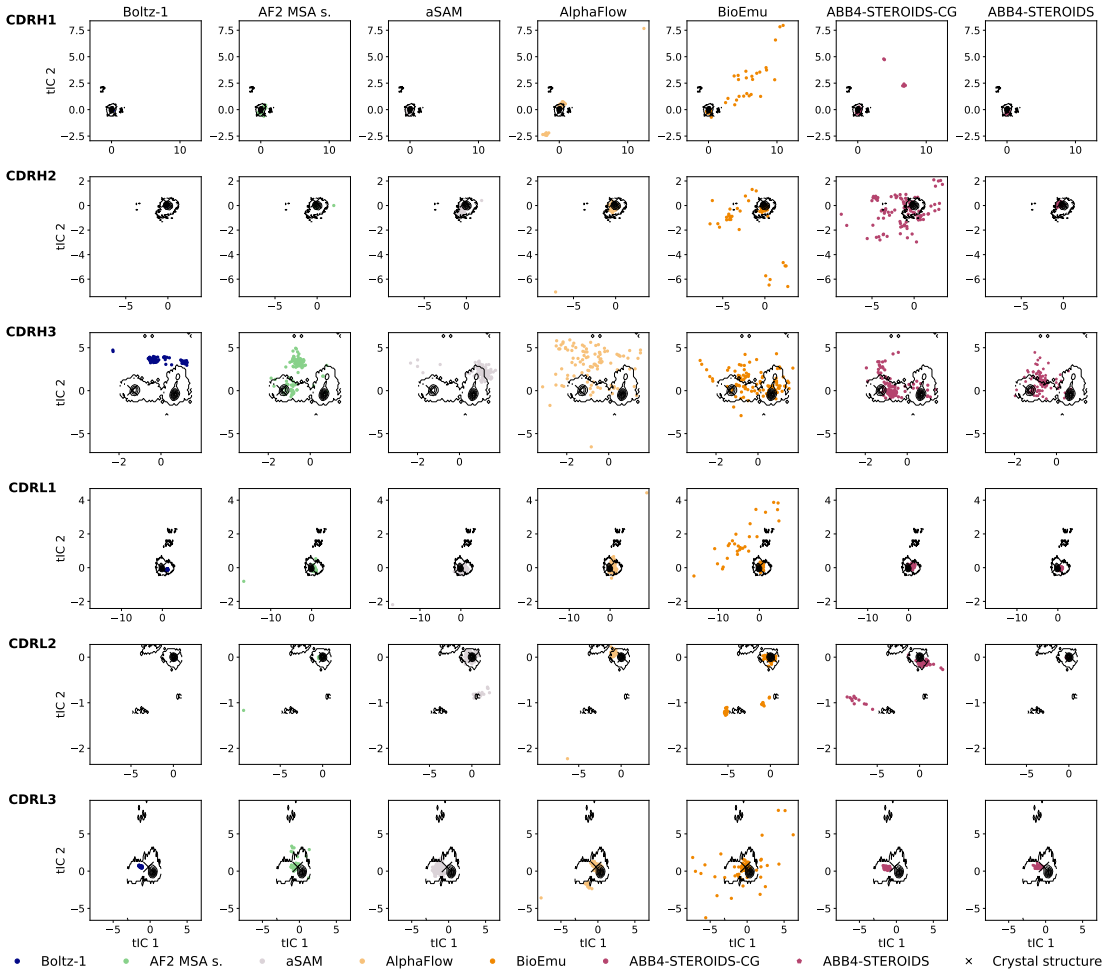


Figure D.11: Comparison of the conformational ensemble sampled for test set antibody PDB 8oni. The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross

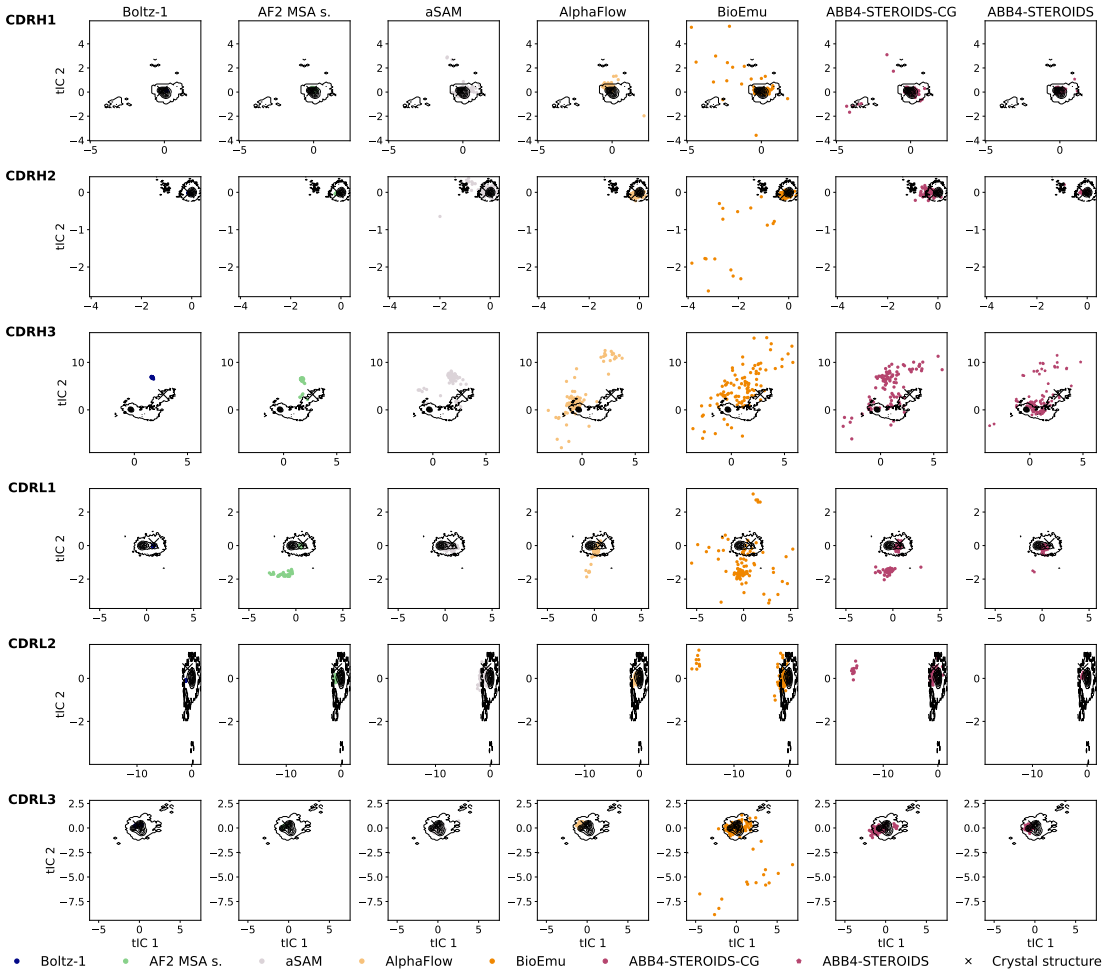


Figure D.12: Comparison of the conformational ensemble sampled for test set antibody PDB 8r1d. The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross

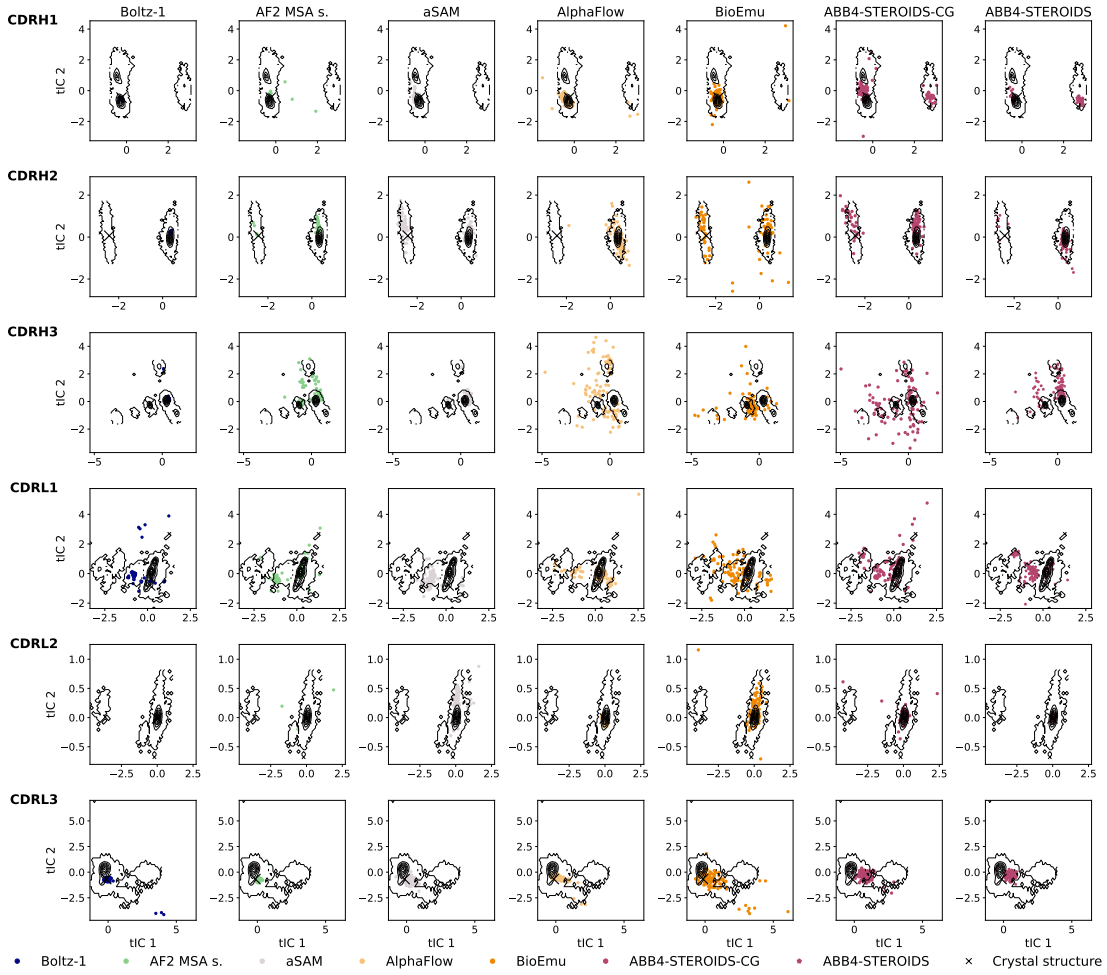


Figure D.13: Comparison of the conformational ensemble sampled for test set antibody PDB 8vvb. The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross

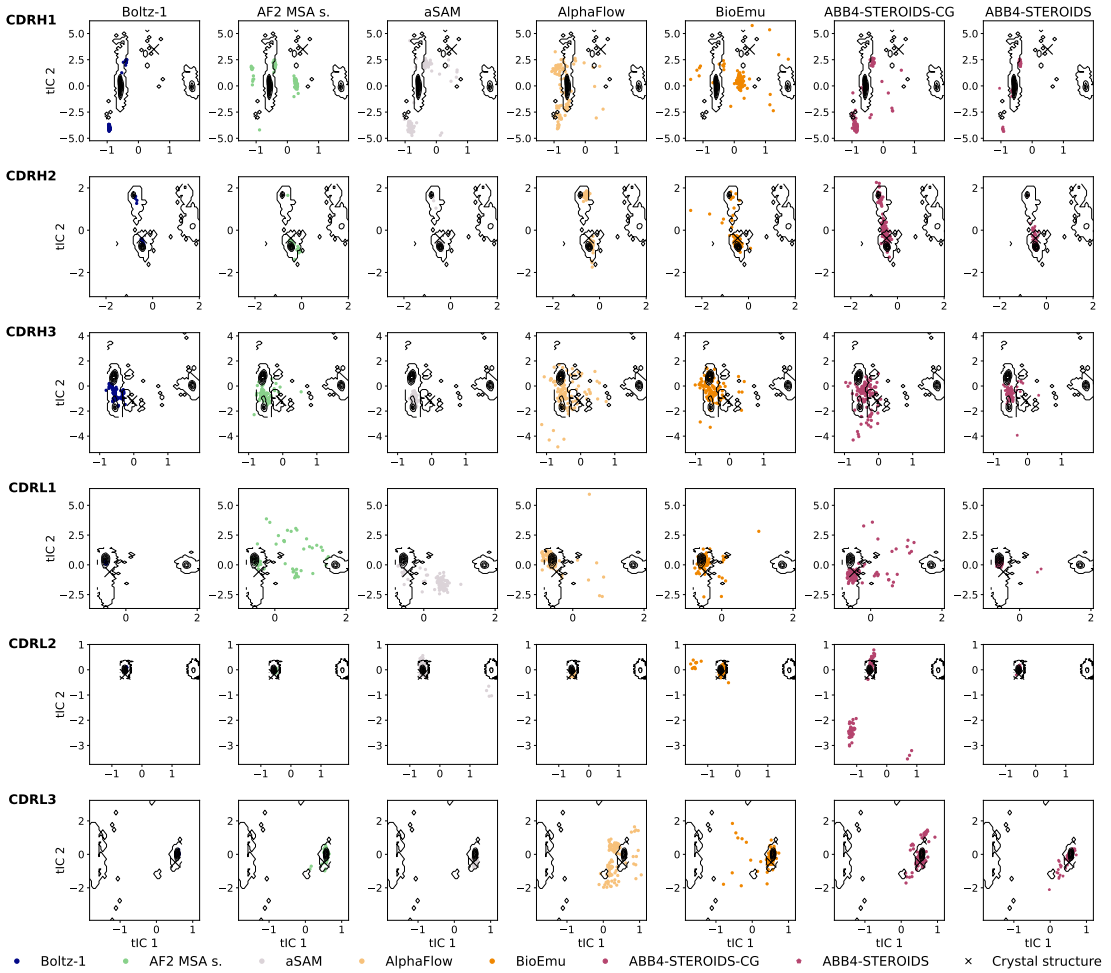


Figure D.14: Comparison of the conformational ensemble sampled for test set antibody PDB 8w83. The first two time-lagged independent components (tICs) are plotted. The black contour indicates the all atom MD distributions. Predicted structures were mapped to the tICs and are represented by the coloured dots. The crystal structure from which the simulation was started is shown by the black cross

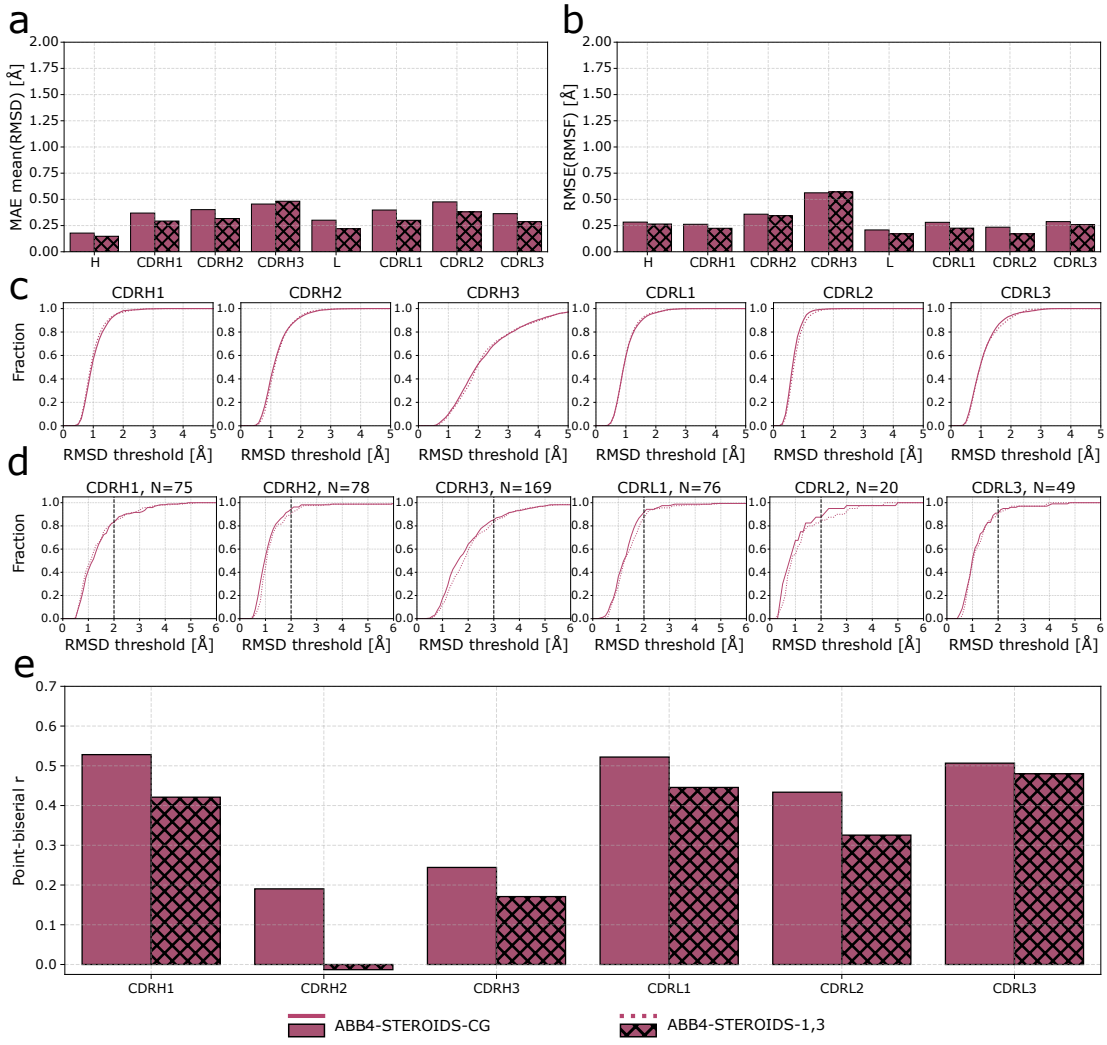


Figure D.15: Evaluation of training stage ablation. In contrast to ABB4-STEROIDS-CG, trained with stages 1 to 3, ABB4-STEROIDS-1,3 was only trained with stages 1 and 3. Mean absolute error (MAE) of RMSD (a) and root mean square error (RMSE) of RMSF (b) in the predicted ensembles compared to the MD simulations. Values computed across the heavy chain (H), light chain (L) and individual CDRs. c) Coverage of MD conformational frames as a function of distance threshold. d) Conformation coverage. Fraction of conformations observed in experimental ensembles covered as a function of distance. Dashed lines highlight ABB4-STEROIDS/-CG covering the highest number of conformations at given distance thresholds. e) Flexibility correlation. Point-biserial correlation coefficient (r) between binary flexibility labels assigned to experimental ensembles and RMSD in predicted ensembles for all CDR loops.

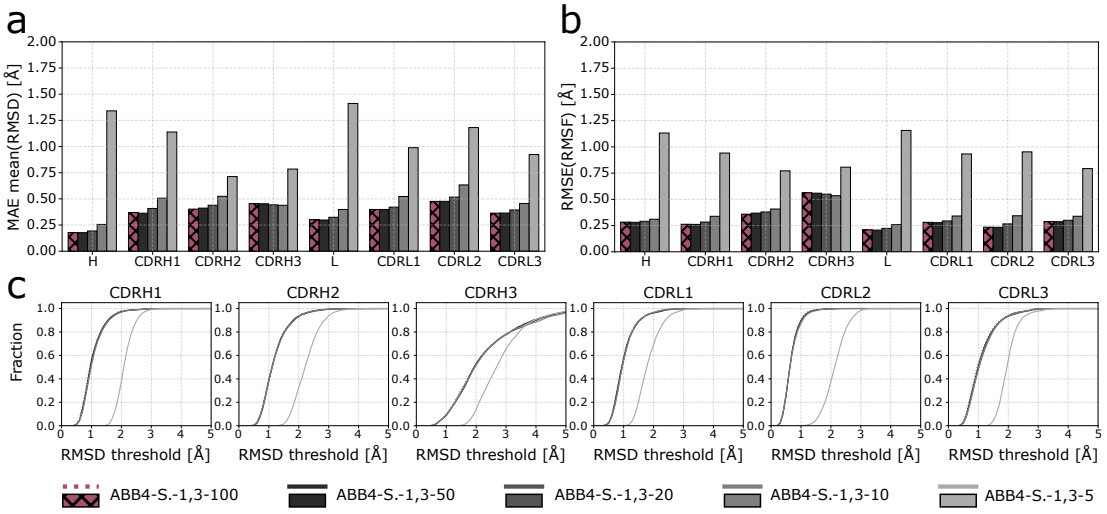


Figure D.16: Evaluation of inference steps on model performance. ABB4-STEROIDS-1,3 with default inference procedure of 100 steps (ABB4-S.-1,3-100) and a reduced number of 50 (ABB4-S.-1,3-50), 20 (ABB4-S.-1,3-20), 10 (ABB4-S.-1,3-10) and 5 steps (ABB4-S.-1,3-5). Mean absolute error (MAE) of RMSD (a) and root mean square error (RMSE) of RMSF (b) in the predicted ensembles compared to the MD simulations. Values computed across the heavy chain (H), light chain (L) and individual CDRs. c) Coverage of MD conformational frames as a function of distance threshold.

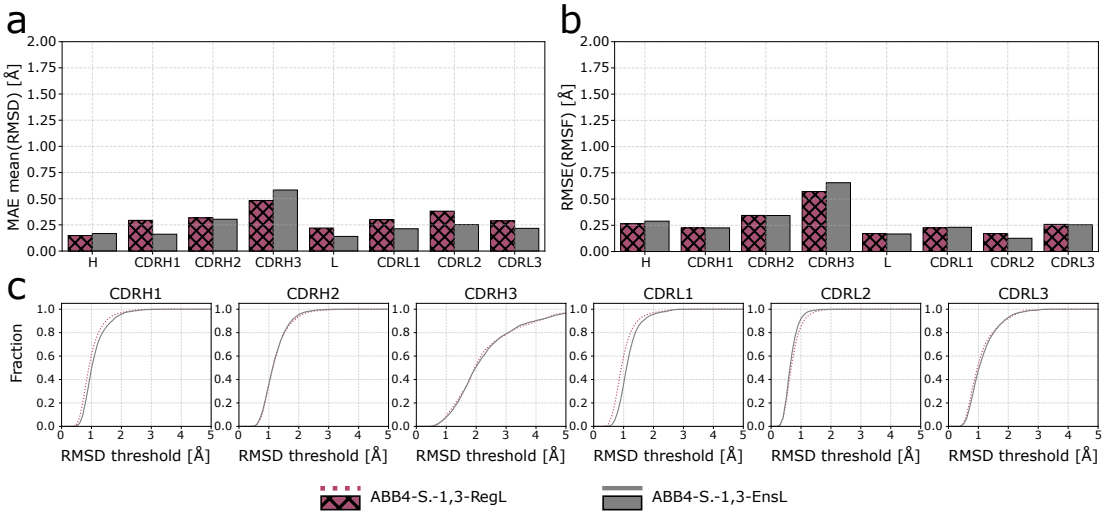


Figure D.17: Evaluation of ensemble loss function on model performance. ABB4-STEROIDS-1,3 trained with the default loss (ABB4-S.-1,3-RegL) and ensemble loss (ABB4-S.-1,3-EnsL). Mean absolute error (MAE) of RMSD (a) and root mean square error (RMSE) of RMSF (b) in the predicted ensembles compared to the MD simulations. Values computed across the heavy chain (H), light chain (L) and individual CDRs. c) Coverage of MD conformational frames as a function of distance threshold.

References

Abanades, Brennan et al. (Mar. 2022). "ABlooper: fast accurate antibody CDR loop structure prediction with accuracy estimation". In: *Bioinformatics* 38.7, pp. 1877–1880.

Abanades, Brennan et al. (May 2023). "ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins". In: *Communications Biology* 6, p. 575.

Abanades, Brennan et al. (Jan. 2024). "The Patent and Literature Antibody Database (PLAbDab): an evolving reference set of functionally diverse, literature-annotated antibody sequences and structures". In: *Nucleic Acids Research* 52.D1, pp. D545–D551.

Abdiche, Yasmina N. et al. (Mar. 2009). "Exploring blocking assays using Octet, ProteOn, and Biacore biosensors". In: *Analytical Biochemistry* 386.2, pp. 172–180.

Abramson, Josh et al. (May 2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature*, pp. 1–3.

Adelman, S. A. and J. D. Doll (Mar. 1976). "Generalized Langevin equation approach for atom/solid-surface scattering: General formulation for classical scattering off harmonic solids". In: *The Journal of Chemical Physics* 64.6, pp. 2375–2388.

Afonine, Pavel V. et al. (June 2018). "Real-space refinement in *PHENIX* for cryo-EM and crystallography". In: *Acta Crystallographica Section D Structural Biology* 74.6, pp. 531–544.

Ahdritz, Gustaf et al. (Aug. 2024). "OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization". In: *Nature Methods* 21.8, pp. 1514–1524.

Anandakrishnan, R., B. Aguilar, and A. V. Onufriev (July 2012). "H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations". In: *Nucleic Acids Research* 40.W1, W537–W541.

Ankerst, Mihael et al. (June 1999). "OPTICS: ordering points to identify the clustering structure". In: *ACM SIGMOD Record* 28.2, pp. 49–60.

Aqvist, Johan et al. (Jan. 2004). "Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm". In: *Chemical Physics Letters* 384.4-6, pp. 288–294.

Babor, Mariana and Tanja Kortemme (June 2009). "Multi-constraint computational design suggests that native sequences of germline antibody H3 loops are nearly optimal for conformational flexibility". In: *Proteins: Structure, Function, and Bioinformatics* 75.4, pp. 846–858.

Baek, Minkyung et al. (Aug. 2021). "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557, pp. 871–876.

Barducci, Alessandro, Giovanni Bussi, and Michele Parrinello (Jan. 2008). "Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method". In: *Physical Review Letters* 100.2, p. 020603.

Basu, Koli et al. (Dec. 2019). "Why recombinant antibodies - benefits and applications". In: *Current Opinion in Biotechnology* 60, pp. 153–158.

Bauer, Paul, Berk Hess, and Erik Lindahl (Feb. 2022). *GROMACS 2022 Source code*.

Beck, Alain et al. (Jan. 2013). "Characterization of Therapeutic Antibodies and Related Products". In: *Analytical Chemistry* 85.2, pp. 715–736.

Benayad, Zakarya et al. (Jan. 2021). "Simulation of FUS Protein Condensates with an Adapted Coarse-Grained Model". In: *Journal of Chemical Theory and Computation* 17.1, pp. 525–537.

Berman, H. M. (Jan. 2000). "The Protein Data Bank". In: *Nucleic Acids Research* 28.1, pp. 235–242.

Bishop, Chirsopher M and Hugh Bishop (2024a). "Diffusion Models". In: *Deep Learning, Foundations and Concepts*. Cham: Springer International Publishing, pp. 581–608.

— (2024b). "Graph Neural Networks". In: *Deep Learning, Foundations and Concepts*. Cham: Springer International Publishing, pp. 581–608.

Blackler, Ryan J. et al. (May 2022). "Antigen binding by conformational selection in near-germline antibodies". In: *Journal of Biological Chemistry* 298.5, p. 101901.

Bonomi, Massimiliano et al. (Oct. 2009). "PLUMED: A portable plugin for free-energy calculations with molecular dynamics". In: *Computer Physics Communications* 180.10, pp. 1961–1972.

Bose, Avishek Joey et al. (Apr. 2024). *SE(3)-Stochastic Flow Matching for Protein Backbone Generation*. arXiv:2310.02391 [cs].

Brennan, Patrick C. et al. (Aug. 2025). "Capturing protein dynamics across timescales with site-directed spin labeling electron paramagnetic resonance spectroscopy". In: *Current Opinion in Structural Biology* 93, p. 103073.

Briney, Bryan et al. (Feb. 2019). "Commonality despite exceptional diversity in the baseline human antibody repertoire". In: *Nature* 566.7744, pp. 393–397.

Bronstein, Michael M. et al. (2021). *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*.

Brooks, B. R. et al. (July 2009). "CHARMM: The biomolecular simulation program". In: *Journal of Computational Chemistry* 30.10, pp. 1545–1614.

Butcher, Jasper et al. (Sept. 2025). *De novo Design of All-atom Biomolecular Interactions with RFdiffusion3*.

Butina, Darko (July 1999). "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets". In: *Journal of Chemical Information and Computer Sciences* 39.4, pp. 747–750.

Cagiada, Matteo and Charlotte M. Deane (Apr. 2026). "Moving the antibody: Molecular dynamics for molecular mechanisms and developability". In: *Current Opinion in Structural Biology* 97, p. 103225.

Cagiada, Matteo et al. (Apr. 2025a). *AF2χ: Predicting protein side-chain rotamer distributions with AlphaFold2*.

Cagiada, Matteo et al. (Nov. 2025b). *Uncovering the flexibility of CDR loops in antibodies and TCRs through large-scale molecular dynamics*.

Cao, Fan et al. (Nov. 2024). "A coarse-grained model for disordered and multi-domain proteins". In: *Protein Science* 33.11, e5172.

Cao, Yunlong et al. (Feb. 2023). "Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution". In: *Nature* 614.7948, pp. 521–529.

Carter, Paul J. and Greg A. Lazar (Mar. 2018). "Next generation antibody drugs: pursuit of the 'high-hanging fruit'". In: *Nature Reviews Drug Discovery* 17.3, pp. 197–223.

Case, David A. et al. (Oct. 2023). "AmberTools". In: *Journal of Chemical Information and Modeling* 63.20, pp. 6183–6191.

Case, David A. et al. (Aug. 2025). "Recent Developments in Amber Biomolecular Simulations". In: *Journal of Chemical Information and Modeling* 65.15, pp. 7835–7843.

Chai Discovery et al. (Oct. 2024). *Chai-1: Decoding the molecular interactions of life.*

Chemical Computing Group ULC (2024). *Molecular Operating Environment (MOE).* 910-1010 Sherbrooke St. W., Montreal, QC H3A 2R7.

Chinery, Lewis et al. (Jan. 2023). "Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors". In: *Bioinformatics* 39.1, btac732.

Chiu, Mark L and Gary L Gilliland (June 2016). "Engineering antibody therapeutics". In: *Current Opinion in Structural Biology* 38, pp. 163–173.

Chiu, Mark L. et al. (Dec. 2019). "Antibody Structure and Function: The Basis for Engineering Therapeutics". In: *Antibodies* 8.4, p. 55.

Choi, Yoonjoo and Charlotte M. Deane (May 2010). "FREAD revisited: Accurate loop structure prediction using a database search algorithm: Loop Structure Prediction Using a Database Search". In: *Proteins: Structure, Function, and Bioinformatics* 78.6, pp. 1431–1440.

Chothia, Cyrus and Arthur M. Lesk (Aug. 1987). "Canonical structures for the hypervariable regions of immunoglobulins". In: *Journal of Molecular Biology* 196.4, pp. 901–917.

Chothia, Cyrus et al. (Dec. 1989). "Conformations of immunoglobulin hypervariable regions". In: *Nature* 342.6252, pp. 877–883.

Del Alamo, Diego et al. (Mar. 2022). "Sampling alternative conformational states of transporters and receptors with AlphaFold2". In: *eLife* 11, e75751.

Di Noia, Javier M. and Michael S. Neuberger (June 2007). "Molecular Mechanisms of Antibody Somatic Hypermutation". In: *Annual Review of Biochemistry* 76.1, pp. 1–22.

Diebolder, Christoph A. et al. (Mar. 2014). "Complement Is Activated by IgG Hexamers Assembled at the Cell Surface". In: *Science* 343.6176, pp. 1260–1263.

Doll, J. D., L. E. Myers, and S. A. Adelman (Dec. 1975). "Generalized Langevin equation approach for atom/solid-surface scattering: Inelastic studies". In: *The Journal of Chemical Physics* 63.11, pp. 4908–4914.

Domene, Carmen, Paolo Barbini, and Simone Furini (Apr. 2015). "Bias-Exchange Metadynamics Simulations: An Efficient Strategy for the Analysis of Conduction and Selectivity in Ion Channels". In: *Journal of Chemical Theory and Computation* 11.4, pp. 1896–1906.

Dreyer, Frédéric A. et al. (July 2025). *Conformation-Aware Structure Prediction of Antigen-Recognizing Immune Proteins.* arXiv:2507.09054 [q-bio].

Ducrocq, Gabriel et al. (Oct. 2024). "cryoSPHERE: Single-Particle HEterogeneous REconstruction from cryo EM". In.

Dunbar, J. et al. (Oct. 2013). "ABangle: characterising the VH-VL orientation in antibodies". In: *Protein engineering, design & selection: PEDS* 26.10, pp. 611–620.

Dunbar, James and Charlotte M. Deane (Jan. 2016). "ANARCI: antigen receptor numbering and receptor classification". In: *Bioinformatics* 32.2, pp. 298–300.

Dunbar, James et al. (Jan. 2014). "SAbDab: the structural antibody database". In: *Nucleic Acids Research* 42.D1, pp. D1140–D1146.

Duval, Alexandre et al. (Mar. 2024). *A Hitchhiker's Guide to Geometric GNNs for 3D Atomic Systems*. arXiv:2312.07511 [cs, q-bio, stat].

Eastman, Peter et al. (Jan. 2024). "OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials". In: *The Journal of Physical Chemistry B* 128.1, pp. 109–116.

Ellaway, Joseph I. J. et al. (July 2023). *Automated Pipeline for Comparing Protein Conformational States in the PDB to AlphaFold2 Predictions*. preprint.

Evans, Richard et al. (Oct. 2021). *Protein complex prediction with AlphaFold-Multimer*.

Fadini, Alisia et al. (Apr. 2026). "AlphaFold as a prior: experimental structure determination conditioned on a pretrained neural network". In: *Nature Methods*, pp. 1–11.

Faezov, Bulat and Roland L. Dunbrack (July 2023). *AlphaFold2 models of the active form of all 437 catalytically-competent typical human kinase domains*.

Fernández-Quintero, Monica L et al. (Dec. 2019a). "Conformational selection of allergen-antibody complexes—surface plasticity of paratopes and epitopes". In: *Protein Engineering, Design and Selection* 32.11, pp. 513–523.

Fernández-Quintero, Monica L. et al. (Aug. 2019b). "CDR-H3 loop ensemble in solution – conformational selection upon antibody binding". In: *mAbs* 11.6, pp. 1077–1088.

Fernández-Quintero, Monica L. et al. (Jan. 2019c). "Characterizing the Diversity of the CDR-H3 Loop Conformational Ensembles in Relationship to Antibody Binding Properties". In: *Frontiers in Immunology* 9, p. 3065.

Fernández-Quintero, Monica L. et al. (Nov. 2019d). "Transitions of CDR-L3 Loop Canonical Cluster Conformations on the Micro-to-Millisecond Timescale". In: *Frontiers in Immunology* 10, p. 2652.

Fernández-Quintero, Monica L. et al. (Oct. 2020). "Antibodies exhibit multiple paratope states influencing VH–VL domain orientations". In: *Communications Biology* 3.1, p. 589.

Ferrari, Állan J. R. et al. (Mar. 2025). *Large-scale discovery, analysis, and design of protein energy landscapes*.

Fischer, Anna-Lena M. et al. (Mar. 2024). "The Role of Force Fields and Water Models in Protein Folding and Unfolding Dynamics". In: *Journal of Chemical Theory and Computation* 20.5, pp. 2321–2333.

Frey, Brendan J. and Delbert Dueck (Feb. 2007). "Clustering by Passing Messages Between Data Points". In: *Science* 315.5814, pp. 972–976.

Gapsys, Vytautas and Bert L De Groot (Aug. 2020). "On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size". In: *eLife* 9, e57589.

Goddard, Thomas D. et al. (Jan. 2018). "UCSF ChimeraX: Meeting modern challenges in visualization and analysis". In: *Protein Science* 27.1, pp. 14–25.

Greenshields-Watson, Alexander et al. (Apr. 2025a). *ANARCII: A Generalised Language Model for Antigen Receptor Numbering*.

Greenshields-Watson, Alexander et al. (Feb. 2025b). "Challenges and compromises: Predicting unbound antibody structures with deep learning". In: *Current Opinion in Structural Biology* 90, p. 102983.

Greiff, Victor et al. (Nov. 2015). "Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires". In: *Trends in Immunology* 36.11, pp. 738–749.

Guloglu, Bora and Charlotte M. Deane (July 2023). "Specific attributes of the VL domain influence both the structure and structural variability of CDR-H3 through steric effects". In: *Frontiers in Immunology* 14, p. 1223802.

Gupta, Akshita et al. (2022). "Experimental techniques to study protein dynamics and conformations". In: *Advances in Protein Molecular and Structural Biology Methods*. Elsevier, pp. 181–197.

Guthmiller, Jenna J. et al. (Dec. 2020). "Polyreactive Broadly Neutralizing B cells Are Selected to Provide Defense against Pandemic Threat Influenza Viruses". In: *Immunity* 53.6, 1230–1244.e5.

Haidar, Jaafar N. et al. (Apr. 2014). "Backbone flexibility of CDR3 and immune recognition of antigens". In: *Journal of Molecular Biology* 426.7, pp. 1583–1599.

Hameduh, Tareq et al. (2020). "Homology modeling in the time of collective and artificial intelligence". In: *Computational and Structural Biotechnology Journal* 18, pp. 3494–3506.

Hastie, Kathryn M. et al. (Oct. 2021). "Defining variant-resistant epitopes targeted by SARS-CoV-2 antibodies: A global consortium study". In: *Science* 374.6566, pp. 472–478.

Hayes, Thomas et al. (Feb. 2025). "Simulating 500 million years of evolution with a language model". In: *Science* 387.6736, pp. 850–858.

Heo, Lim and Michael Feig (Jan. 2024). "One bead per residue can describe all-atom protein structures". In: *Structure* 32.1, 97–111.e6.

Hershberg, Uri and Eline T. Luning Prak (Sept. 2015). "The analysis of clonal expansions in normal and autoimmune B cell repertoires". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1676, p. 20140239.

Hilser, Vincent J., James O. Wrabl, and Hesam N. Motlagh (June 2012). "Structural and Energetic Basis of Allostery". In: *Annual Review of Biophysics* 41.1, pp. 585–609.

Hitawala, Fatima N. and Jeffrey J. Gray (Sept. 2024). *What has AlphaFold3 learned about antibody and nanobody docking, and what remains unsolved?*

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). *Denoising Diffusion Probabilistic Models*.

Hollingsworth, Scott A. and Ron O. Dror (Sept. 2018). "Molecular Dynamics Simulation for All". In: *Neuron* 99.6, pp. 1129–1143.

Hou, X.-L. et al. (Apr. 2016). "Current status and recent advances of next generation sequencing techniques in immunological repertoire". In: *Genes & Immunity* 17.3, pp. 153–164.

Huguet, Guillaume et al. (May 2024). *Sequence-Augmented SE(3)-Flow Matching For Conditional Protein Backbone Generation*. arXiv:2405.20313 [cs, q-bio].

Hummer, Alissa M. et al. (May 2023). *Investigating the Volume and Diversity of Data Needed for Generalizable Antibody-Antigen $\Delta\Delta G$ Prediction*. preprint.

Invernizzi, Michele et al. (Oct. 2025). *Advancing Protein Ensemble Predictions Across the Order–Disorder Continuum*.

Isomorphic Labs (Feb. 2026). *Accurate Predictions of Novel Biomolecular Interactions with IsoDDE*.

Izadi, Saeed and Alexey V. Onufriev (Aug. 2016). "Accuracy limit of rigid 3-point water models". In: *The Journal of Chemical Physics* 145.7, p. 074501.

Jaeger, Adam and David Banks (May 2023). "Cluster analysis: A modern statistical review". In: *WIREs Computational Statistics* 15.3, e1597.

Jafarshad, Ali et al. (Mar. 2007). "A Novel Antibody-Dependent Cellular Cytotoxicity Mechanism Involved in Defense against Malaria Requires Costimulation of Monocytes FcγRII and FcγRIII". In: *The Journal of Immunology* 178.5, pp. 3099–3106.

Jaffe, David B. et al. (Nov. 2022). "Functional antibodies exhibit light chain coherence". In: *Nature* 611.7935, pp. 352–357.

James, Leo C., Pietro Roversi, and Dan S. Tawfik (Feb. 2003). "Antibody Multispecificity Mediated by Conformational Diversity". In: *Science* 299.5611, pp. 1362–1367.

Janson, Giacomo, Alexander Jussupow, and Michael Feig (Mar. 2025). *Deep generative modeling of temperature-dependent structural ensembles of proteins*.

Jeliazkov, Jeliazko R. et al. (2018). "Repertoire Analysis of Antibody CDR-H3 Loops Suggests Affinity Maturation Does Not Typically Result in Rigidification". In: *Frontiers in Immunology* 9.

Jin, Jaehyeok et al. (Oct. 2022). "Bottom-up Coarse-Graining: Principles and Perspectives". In: *Journal of Chemical Theory and Computation* 18.10, pp. 5759–5791.

Jing, Bowen, Bonnie Berger, and Tommi Jaakkola (Feb. 2024a). *AlphaFold Meets Flow Matching for Generating Protein Ensembles*. arXiv:2402.04845 [cs, q-bio].

Jing, Bowen et al. (Sept. 2024b). *Generative Modeling of Molecular Dynamics Trajectories*. arXiv:2409.17808 [cs, q-bio].

Jorgensen, William L. et al. (July 1983). "Comparison of simple potential functions for simulating liquid water". In: *The Journal of Chemical Physics* 79.2, pp. 926–935.

Joyce, M. Gordon et al. (July 2016). "Vaccine-Induced Antibodies that Neutralize Group 1 and Group 2 Influenza A Viruses". In: *Cell* 166.3, pp. 609–623.

Julian, Mark C. et al. (Mar. 2017). "Efficient affinity maturation of antibody variable domains requires co-selection of compensatory mutations to maintain thermodynamic stability". In: *Scientific Reports* 7.1, p. 45259.

Jumper, John et al. (Aug. 2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589.

Kabat, Elvin Abraham, Tai Te Wu, and Howard Bilofsky (1979). *Sequences of immunoglobulin chains: tabulation and analysis of amino acid sequences of precursors, V-regions, C-regions, J-chain and 2-microglobulins*. National Institute of Health.

Kabsch, W. and C. Sander (Dec. 1983). "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features". In: *Biopolymers* 22.12, pp. 2577–2637.

Kay, Lewis E. (Dec. 2011). "NMR studies of protein structure and dynamics". In: *Journal of Magnetic Resonance* 213.2, pp. 477–491.

Kelow, Simon et al. (Oct. 2022). *A penultimate classification of canonical antibody CDR conformations*.

Kenlay, Henry et al. (May 2024). *ABodyBuilder3: Improved and scalable antibody structure predictions*. arXiv:2405.20863 [cs, q-bio].

Khetan, Rahul et al. (Dec. 2022). "Current advances in biopharmaceutical informatics: guidelines, impact and challenges in the computational developability assessment of antibody therapeutics". In: *mAbs* 14.1, p. 2020082.

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*.

Kleckner, Ian R and Mark P Foster (Aug. 2011). "An introduction to NMR-based approaches for measuring protein dynamics". In: *Biochimica et biophysica acta* 1814.8, pp. 942–968.

Kmiecik, Sebastian et al. (July 2016). "Coarse-Grained Protein Models and Their Applications". In: *Chemical Reviews* 116.14, pp. 7898–7936.

Konermann, Lars, Jingxi Pan, and Yu-Hong Liu (2011). "Hydrogen exchange mass spectrometry for studying protein structure and dynamics". In: *Chem. Soc. Rev.* 40.3, pp. 1224–1234.

Kouba, Petr et al. (Feb. 2025). *Learning to engineer protein flexibility*. arXiv:2412.18275 [q-bio].

Krishnan, Lavanya et al. (Feb. 2008). "Role of Antibody Paratope Conformational Flexibility in the Manifestation of Molecular Mimicry". In: *Biophysical Journal* 94.4, pp. 1367–1376.

Kroon, Gerard J.A. et al. (July 2003). "Changes in structure and dynamics of the Fv fragment of a catalytic antibody upon binding of inhibitor". In: *Protein Science* 12.7, pp. 1386–1394.

Kryshtafovych, Andriy et al. (Dec. 2019). "Critical assessment of methods of protein structure prediction (CASP)-Round XIII". In: *Proteins* 87.12, pp. 1011–1020.

Kühlbrandt, Werner (Mar. 2014). "The Resolution Revolution". In: *Science* 343.6178, pp. 1443–1444.

Kumar, Sandeep et al. (Jan. 2000). "Folding and binding cascades: Dynamic landscapes and population shifts". In: *Protein Science* 9.1, pp. 10–19.

Labute, Paul (Apr. 2009). "Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures". In: *Proteins* 75.1, pp. 187–205.

Laio, Alessandro and Michele Parrinello (Oct. 2002). "Escaping free-energy minima". In: *Proceedings of the National Academy of Sciences* 99.20, pp. 12562–12566.

Landrum, Greg (2006). *RDKit: Open-source cheminformatics*.

Lee, Jeffrey E. et al. (July 2008). "Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor". In: *Nature* 454.7201, pp. 177–182.

Lee, Jessica H. et al. (July 2022). "Structural Features of Antibody-Peptide Recognition". In: *Frontiers in Immunology* 13.

Leem, Jinwoo et al. (Oct. 2016). "ABodyBuilder: Automated antibody structure prediction with data-driven accuracy estimation". In: *mAbs* 8.7, pp. 1259–1268.

Leem, Jinwoo et al. (Jan. 2018). "STCRDab: the structural T-cell receptor database". In: *Nucleic Acids Research* 46.D1, pp. D406–D412.

Lefranc, Marie-Paule et al. (Jan. 2003). "IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains". In: *Developmental & Comparative Immunology* 27.1, pp. 55–77.

Lewis, Sarah et al. (Aug. 2025). "Scalable emulation of protein equilibrium ensembles with generative deep learning". In: *Science* 389.6761, pp. 850–858.

Li, Ruofan et al. (Aug. 2022). "Conformational flexibility in neutralization of SARS-CoV-2 by naturally elicited anti-SARS-CoV-2 antibodies". In: *Communications Biology* 5.1, pp. 1–15.

Li, Tong et al. (July 2015). "Rigidity Emerges during Antibody Evolution in Three Distinct Antibody Systems: Evidence from QSFR Analysis of Fab Fragments". In: *PLOS Computational Biology* 11.7, e1004327.

Lin, Zeming et al. (July 2022). *Language models of protein sequences at the scale of evolution enable accurate structure prediction*.

Lin, Zeming et al. (Mar. 2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637, pp. 1123–1130.

Lipman, Yaron et al. (Feb. 2023). *Flow Matching for Generative Modeling*. arXiv:2210.02747 [cs, stat].

Liu, Chu'nan et al. (Dec. 2024). "Do antibody CDR loops change conformation upon binding?" In: *mAbs* 16.1, p. 2322533.

Lopez-Santibanez-Jacome, Laura, S. Erendira Avendano-Vazquez, and Carlos Fabian Flores-Jasso (Apr. 2019). "The Pipeline Repertoire for Ig-Seq Analysis". In: *Frontiers in Immunology* 10, p. 899.

Lu, Jiarui et al. (Mar. 2024). *Str2Str: A Score-based Framework for Zero-shot Protein Conformation Sampling*. arXiv:2306.03117 [cs, q-bio].

Lu, Lenette L. et al. (Jan. 2018). "Beyond binding: antibody effector functions in infectious diseases". In: *Nature Reviews Immunology* 18.1, pp. 46–61.

Lu, Ruei-Min et al. (Jan. 2020). "Development of therapeutic antibodies for the treatment of diseases". In: *Journal of Biomedical Science* 27.1, p. 1.

Ma, Buyong et al. (Sept. 1999). "Folding funnels and binding mechanisms". In: *Protein Engineering, Design and Selection* 12.9, pp. 713–720.

MacQueen, J. (Jan. 1967). "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* 5.1, pp. 281–298.

Manivel, Venkatasamy et al. (Nov. 2000). "Maturation of an Antibody Response Is Governed by Modulations in Flexibility of the Antigen-Combining Site". In: *Immunity* 13.5, pp. 611–620.

Marks, Claire, Jiye Shi, and Charlotte M Deane (Mar. 2018). "Predicting loop conformational ensembles". In: *Bioinformatics* 34.6, pp. 949–956.

Martí-Renom, Marc A. et al. (June 2000). "Comparative Protein Structure Modeling of Genes and Genomes". In: *Annual Review of Biophysics and Biomolecular Structure* 29.1, pp. 291–325.

Martin, Andrew C.R. and Janet M. Thornton (Nov. 1996). "Structural Families in Loops of Homologous Proteins: Automatic Classification, Modelling and Application to Antibodies". In: *Journal of Molecular Biology* 263.5, pp. 800–815.

Mikolajek, Halina et al. (Aug. 2022). "Correlation between the binding affinity and the conformational entropy of nanobody SARS-CoV-2 spike protein complexes". In: *Proceedings of the National Academy of Sciences* 119.31, e2205412119.

Mille-Fragoso, Luis S. et al. (Sept. 2025). *Efficient generation of epitope-targeted de novo antibodies with Germinal*.

Mirarchi, Antonio, Toni Giorgino, and Gianni De Fabritiis (Nov. 2024). "mdCATH: A Large-Scale MD Dataset for Data-Driven Computational Biophysics". In: *Scientific Data* 11.1, p. 1299.

Mirdita, Milot et al. (June 2022). "ColabFold: making protein folding accessible to all". In: *Nature Methods* 19.6, pp. 679–682.

Murtagh, Fionn and Pedro Contreras (Jan. 2012). "Algorithms for hierarchical clustering: an overview". In: *WIREs Data Mining and Knowledge Discovery* 2.1, pp. 86–97.

Nilvebrant, Johan and Johan Rockberg (2018). "An Introduction to Epitope Mapping". In: *Epitope Mapping Protocols*. 3rd. Methods in Molecular Biology. New York, NY: Springer, pp. 1–10.

North, Benjamin, Andreas Lehmann, and Roland L. Dunbrack (Feb. 2011). "A New Clustering of Antibody CDR Loop Conformations". In: *Journal of Molecular Biology* 406.2, pp. 228–256.

Nowak, Jaroslaw et al. (May 2016). "Length-independent structural similarities enrich the antibody CDR canonical class model". In: *mAbs* 8.4, pp. 751–760.

Olsen, Tobias H., Fergus Boyles, and Charlotte M. Deane (Jan. 2022). "Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences". In: *Protein Science* 31.1, pp. 141–146.

Pacesa, Martin et al. (Oct. 2024). *BindCraft: one-shot design of functional protein binders*.

Passaro, Saro et al. (June 2025). *Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction*.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830.

Pérez-Hernández, Guillermo et al. (July 2013). "Identification of slow molecular order parameters for Markov model construction". In: *The Journal of Chemical Physics* 139.1, p. 015102.

Prigent, Julie et al. (May 2018). "Conformational Plasticity in Broadly Neutralizing HIV-1 Antibodies Triggers Polyreactivity". In: *Cell Reports* 23.9, pp. 2568–2581.

Punjani, Ali and David J. Fleet (June 2023). "3DFlex: determining structure and motion of flexible proteins from cryo-EM". In: *Nature Methods* 20.6, pp. 860–870.

Punjani, Ali et al. (Mar. 2017). "cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination". In: *Nature Methods* 14.3, pp. 290–296.

Quast, Nele P. et al. (Mar. 2025). "T-cell receptor structures and predictive models reveal comparable alpha and beta chain structural diversity despite differing genetic complexity". In: *Communications Biology* 8, p. 362.

Raybould, Matthew I J et al. (May 2021a). "CoV-AbDab: the coronavirus antibody database". In: *Bioinformatics* 37.5, pp. 734–735.

Raybould, Matthew I. J., Anthony R. Rees, and Charlotte M. Deane (Jan. 2021b). "Current strategies for detecting functional convergence across B-cell receptor repertoires". In: *mAbs* 13.1, p. 1996732.

Raybould, Matthew I. J. et al. (Mar. 2019). "Five computational developability guidelines for therapeutic antibody profiling". In: *Proceedings of the National Academy of Sciences* 116.10, pp. 4025–4030.

Raybould, Matthew I.J. et al. (Sept. 2024). "The Observed T Cell Receptor Space database enables paired-chain repertoire mining, coherence analysis, and language modeling". In: *Cell Reports* 43.9, p. 114704.

Rees, Anthony R (Jan. 2020). "Understanding the human antibody repertoire". In: *mAbs* 12.1, p. 1729683.

Riccabona, Jakob R. et al. (Sept. 2024). "Assessing AF2's ability to predict structural ensembles of proteins". In: *Structure* 32, 2147–2159.e2.

Richardson, Eve et al. (Jan. 2021). "A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies". In: *mAbs* 13.1, p. 1869406.

Richman, Daniel D. et al. (Dec. 2025). *Unlocking hidden biomolecular conformational landscapes in diffusion models at inference time*. arXiv:2512.03312.

Rijal, Pramila et al. (Apr. 2019). "Therapeutic Monoclonal Antibodies for Ebola Virus Infection Derived from Vaccinated Humans". In: *Cell Reports* 27.1, 172–186.e7.

Ripoll, Daniel R., Sidhartha Chaudhury, and Anders Wallqvist (Mar. 2021). "Using the antibody-antigen binding interface to train image-based deep neural networks for antibody-epitope classification". In: *PLOS Computational Biology* 17.3, e1008864.

Robbiani, Davide F. et al. (Aug. 2020). "Convergent antibody responses to SARS-CoV-2 in convalescent individuals". In: *Nature* 584.7821, pp. 437–442.

Robinson, Sarah A. et al. (Dec. 2021). "Epitope profiling using computational structural modelling demonstrated on coronavirus-binding antibodies". In: *PLOS Computational Biology* 17.12, e1009675.

Roe, Daniel R. and Thomas E. Cheatham (July 2013). "PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data". In: *Journal of Chemical Theory and Computation* 9.7, pp. 3084–3095.

Ruffolo, Jeffrey A., Jeremias Sulam, and Jeffrey J. Gray (Feb. 2022a). "Antibody structure prediction using interpretable deep learning". In: *Patterns* 3.2, p. 100406.

Ruffolo, Jeffrey A. et al. (July 2020). "Geometric potentials from deep learning improve prediction of CDR H3 loop structures". In: *Bioinformatics* 36, pp. i268–i275.

Ruffolo, Jeffrey A. et al. (Apr. 2022b). *Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies*.

Sabri Dashti, Danial, Yilin Meng, and Adrian E. Roitberg (Aug. 2012). "pH-Replica Exchange Molecular Dynamics in Proteins Using a Discrete Protonation Method". In: *The Journal of Physical Chemistry B* 116.30, pp. 8805–8811.

Sala, D. et al. (Aug. 2023a). "Modeling conformational states of proteins with AlphaFold". In: *Current Opinion in Structural Biology* 81, p. 102645.

Sala, Davide, Peter W. Hildebrand, and Jens Meiler (2023b). "Biasing AlphaFold2 to predict GPCRs and kinases with user-defined functional or structural properties". In: *Frontiers in Molecular Biosciences* 10.

Saldaño, Tadeo et al. (May 2022). "Impact of protein conformational diversity on AlphaFold predictions". In: *Bioinformatics* 38.10, pp. 2742–2748.

Salomon-Ferrer, Romelia et al. (Sept. 2013). "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald". In: *Journal of Chemical Theory and Computation* 9.9, pp. 3878–3888.

Satorras, Victor Garcia, Emiel Hoogeboom, and Max Welling (Feb. 2022). *E(n) Equivariant Graph Neural Networks*. arXiv:2102.09844 [cs, stat].

Scheid, Johannes F. et al. (Sept. 2011). "Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding". In: *Science* 333.6049, pp. 1633–1637.

Schmidt, Aaron G. et al. (Jan. 2013). "Preconfiguration of the antigen-binding site during affinity maturation of a broadly neutralizing influenza virus antibody". In: *Proceedings of the National Academy of Sciences* 110.1, pp. 264–269.

Schneider, Constantin, Matthew I J Raybould, and Charlotte M Deane (Jan. 2022). "SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker". In: *Nucleic Acids Research* 50.D1, pp. D1368–D1372.

Schubert, Erich et al. (July 2017). "DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN". In: *ACM Transactions on Database Systems* 42.3, 19:1–19:21.

Sela-Culang, Inbal, Vered Kunik, and Yanay Ofran (2013). "The Structural Basis of Antibody-Antigen Recognition". In: *Frontiers in Immunology* 4.

Shaw, David E. et al. (Nov. 2021). "Anton 3: twenty microseconds of molecular dynamics simulation before lunch". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. St. Louis Missouri: ACM, pp. 1–11.

Shrock, Ellen L. et al. (Apr. 2023). "Germline-encoded amino acid–binding motifs drive immunodominant public antibody responses". In: *Science* 380.6640, eadc9498.

Smith, George P. and Valery A. Petrenko (Apr. 1997). "Phage Display". In: *Chemical Reviews* 97.2, pp. 391–410.

Sohl-Dickstein, Jascha et al. (2015). *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*.

Song, Yang and Stefano Ermon (2019). *Generative Modeling by Estimating Gradients of the Data Distribution*.

Sormanni, Pietro, Francesco A. Aprile, and Michele Vendruscolo (2018). "Third generation antibody discovery methods: *in silico* rational design". In: *Chemical Society Reviews* 47.24, pp. 9137–9157.

Spoendlin, Fabian C. and Oliver M Turnbull (Aug. 2025a). *fspoendlin/ITsFlexible: v1.0.1*.

Spoendlin, Fabian C. et al. (2023). "Improved computational epitope profiling using structural models identifies a broader diversity of antibodies that bind to the same epitope". In: *Frontiers in Molecular Biosciences* 10, p. 1237621.

Spoendlin, Fabian C. et al. (2025b). *ALL-conformations / ITsFlexible*.

Spoendlin, Fabian C. et al. (Oct. 2025c). "Predicting the conformational flexibility of antibody and T cell receptor complementarity-determining regions". In: *Nature Machine Intelligence*, pp. 1–13.

Stark, Hannes et al. (Nov. 2025). *BoltzGen: Toward Universal Binder Design*.

Stein, Richard A. and Hassane S. Mchaourab (Aug. 2022). "SPEACH_AF: Sampling protein ensembles and conformational heterogeneity with Alphafold2". In: *PLOS Computational Biology* 18.8, e1010483.

Steinegger, Martin and Johannes Söding (Nov. 2017). "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature Biotechnology* 35.11, pp. 1026–1028.

Sweeney, Charlotte et al. (Dec. 2024). *Estimating protein flexibility via uncertainty quantification of structure prediction models*.

Tareen, Ammar and Justin B Kinney (Apr. 2020). "Logomaker: beautiful sequence logos in Python". In: *Bioinformatics* 36.7, pp. 2272–2274.

Teilum, Kaare, Johan G. Olsen, and Birthe B. Kragelund (July 2009). "Functional aspects of protein flexibility". In: *Cellular and Molecular Life Sciences* 66.14, pp. 2231–2247.

Tesei, Giulio et al. (Jan. 2024). "Conformational ensembles of the human intrinsically disordered proteome". In: *Nature*, pp. 1–8.

The Antibody Society (2026). *Therapeutic monoclonal antibodies approved or in review in the EU or US*.

The OpenFold3 Team (2025). *OpenFold3-preview*.

The PLUMED consortium (Aug. 2019). "Promoting transparency and reproducibility in enhanced molecular simulations". In: *Nature Methods* 16.8, pp. 670–673.

Thomasen, F. Emil et al. (Apr. 2022). "Improving Martini 3 for Disordered and Multidomain Proteins". In: *Journal of Chemical Theory and Computation* 18.4, pp. 2033–2041.

Thomasen, F. Emil et al. (Aug. 2024). "Rescaling protein-protein interactions improves Martini 3 for flexible proteins in solution". In: *Nature Communications* 15.1, p. 6645.

Thompson, Michael C., Todd O. Yeates, and Jose A. Rodriguez (July 2020). "Advances in methods for atomic resolution macromolecular structure determination". In: *F1000Research* 9, p. 667.

Tian, Chuan et al. (Jan. 2020). “ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution”. In: *Journal of Chemical Theory and Computation* 16.1, pp. 528–552.

Tiberti, Matteo et al. (Oct. 2015). “ENCORE: Software for Quantitative Ensemble Comparison”. In: *PLOS Computational Biology* 11.10, e1004415.

Tonegawa, Susumu (Apr. 1983). “Somatic generation of antibody diversity”. In: *Nature* 302.5909, pp. 575–581.

Tribello, Gareth A. et al. (Feb. 2014). “PLUMED 2: New feathers for an old bird”. In: *Computer Physics Communications* 185.2, pp. 604–613.

Vander Meersche, Yann et al. (Jan. 2024). “ATLAS: protein flexibility description from atomistic molecular dynamics simulations”. In: *Nucleic Acids Research* 52.D1, pp. D384–D392.

Varadi, Mihaly et al. (Jan. 2024). “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences”. In: *Nucleic Acids Research* 52.D1, pp. D368–D375.

Victora, Gabriel D. and Michel C. Nussenzweig (Apr. 2012). “Germinal Centers”. In: *Annual Review of Immunology* 30.1, pp. 429–457.

Villani, Cedric (2009). “Cyclical monotonicity and Kantorovich duality”. In: *Optimal Transport: Old and New*. Vol. 338. A Series of Comprehensive Studies in Mathematics. Berlin: Springer.

Von Bülow, Sören, Kristoffer E. Johansson, and Kresten Lindorff-Larsen (Oct. 2025). *AF-CALVADOS: AlphaFold-guided simulations of multi-domain proteins at the proteome level*.

Von Bülow, Sören et al. (2025). *Software package for simulations using the coarse-grained CALVADOS model*.

Wang, Lee-Ping, Todd J. Martinez, and Vijay S. Pande (June 2014). “Building Force Fields: An Automatic, Systematic, and Reproducible Approach”. In: *The Journal of Physical Chemistry Letters* 5.11, pp. 1885–1891.

Wang, Yong et al. (Feb. 2018). “Role of protein dynamics in transmembrane receptor signalling”. In: *Current Opinion in Structural Biology* 48, pp. 74–82.

Waterhouse, Andrew et al. (July 2018). “SWISS-MODEL: homology modelling of protein structures and complexes”. In: *Nucleic Acids Research* 46.W1, W296–W303.

Waudby, Christopher A. and Ignacio Alfonso (Dec. 2023). “An introduction to one- and two-dimensional lineshape analysis of chemically exchanging systems”. In: *Journal of Magnetic Resonance Open* 16-17, p. 100102.

Wayment-Steele, Hannah K. et al. (Jan. 2024). “Predicting multiple conformations via sequence clustering and AlphaFold2”. In: *Nature* 625.7996, pp. 832–839.

Wayment-Steele, Hannah K. et al. (Mar. 2025). *Learning millisecond protein dynamics from what is missing in NMR spectra*.

Wei, Guanghong et al. (June 2016). “Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell”. In: *Chemical Reviews* 116.11, pp. 6516–6551.

Weitzner, Brian D., Roland L. Dunbrack, and Jeffrey J. Gray (Feb. 2015). “The Origin of CDR H3 Structural Diversity”. In: *Structure* 23.2, pp. 302–311.

Wohlwend, Jeremy et al. (Nov. 2024). *Boltz-1 Democratizing Biomolecular Interaction Modeling*.

Wong, Sergio E., Ben D. Sellers, and Matthew P. Jacobson (Mar. 2011). "Effects of somatic mutations on CDR loop flexibility during affinity maturation". In: *Proteins: Structure, Function, and Bioinformatics* 79.3, pp. 821–829.

Wong, Wing Ki, Jinwoo Leem, and Charlotte M. Deane (2019). "Comparative Analysis of the CDR Loops of Antigen Receptors". In: *Frontiers in Immunology* 10.

Wong, Wing Ki et al. (Jan. 2021). "Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope". In: *mAbs* 13.1, p. 1873478.

Xu, Yaoyao et al. (Nov. 2025). *TEMPO: Temporal Multi-scale Autoregressive Generation of Protein Conformational Ensembles*. arXiv:2511.05510.

Yassai, Maryam B. et al. (July 2009). "A clonotype nomenclature for T cell receptors". In: *Immunogenetics* 61.7, pp. 493–502.

Yim, Jason et al. (Oct. 2023a). *Fast protein backbone generation with SE(3) flow matching*. arXiv:2310.05297 [q-bio].

Yim, Jason et al. (2023b). "SE(3) diffusion model with application to protein backbone generation". In.

Yim, Jason et al. (July 2024). *Improved motif-scaffolding with SE(3) flow matching*. arXiv:2401.04082 [cs, q-bio, stat].

Zhao, Jun, Ruth Nussinov, and Buyong Ma (Jan. 2019). "Antigen binding allosterically promotes Fc receptor recognition". In: *mAbs* 11.1, pp. 58–74.

Zheng, Shuxin et al. (May 2024). "Predicting equilibrium distributions for molecular systems with deep learning". In: *Nature Machine Intelligence* 6, pp. 558–567.

Zhong, Ellen D. et al. (Feb. 2021). "CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks". In: *Nature Methods* 18.2, pp. 176–185.

Zhou, Jie et al. (Jan. 2020). "Graph neural networks: A review of methods and applications". In: *AI Open* 1, pp. 57–81.

Zhu, Junjie et al. (May 2024). *Precise Generation of Conformational Ensembles for Intrinsically Disordered Proteins via Fine-tuned Diffusion Models*.