Advancing Antibody Design: Integrating Protein Language Models for Enhanced Computational Strategies

by

Benyamin Jamialahmadi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Antibodies, or immunoglobulins, are integral to the immune response, playing a crucial role in recognizing and neutralizing external threats such as pathogens. The design of these molecules, however, is complex due to the limited availability of paired structural antibody-antigen data and the intricacies of structurally non-deterministic regions. In this thesis, we explore innovative approaches for computationally designing antibodies, addressing key challenges in traditional methods. Our focus is on overcoming the limitations of existing computational techniques in antibody design, which include limited structural data availability, CDR flexibility, and dependence on contextual information. We propose two novel solutions leveraging Protein Language Models (pLMs). The first employs a sequence-to-sequence model, analogous to language translation, utilizing data augmentation for semi-supervised training. The second approach integrates both sequential and structural antigen information into a pLM using specially designed adapter modules. These methods aim to efficiently utilize extensive sequence data, circumventing the challenges of limited structural data. Our models demonstrate promising results in the Rosetta Antibody Design benchmark, outperforming existing models and showcasing the potential of integrating pLMs in computational antibody design. This research contributes to enhancing the precision and applicability of antibody design, marking a significant advancement in therapeutic and diagnostic applications.

Acknowledgments

I would like to take this opportunity to extend my sincere gratitude to those who have been instrumental in the completion of my thesis.

First and foremost, my supervisors, Prof. Ali Ghodsi and Prof. Mohammad Kohandel, deserve my deepest gratitude for their endless support, guidance, and patience. Their knowledge and expertise have been invaluable to my research.

I am also very thankful to Prof. Mahmood Chamankhah for his essential advice and insights that have helped to shape my work.

A special thank you goes to my committee members, Prof. Ming Li and Prof. Anita Layton, for their thorough reviews and helpful feedback that have contributed to the improvement of my thesis.

Finally, I wish to express my appreciation to my friends, Sahand Ajami and Arman Hafizi, for their encouragement and support. Their help has been significant to my research.

To everyone who has supported me, your assistance and encouragement have been crucial to my success, and for that, I am forever grateful.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the field of molecular biology, understanding and designing proteins is crucial. Proteins are vital for biological systems due to their diverse roles, ranging from speeding up chemical reactions to providing structural support [6]. Notably, antibodies, also known as immunoglobulins, stand out for their essential role in the immune system, acting as defense agents against external biological threats like bacteria, viruses, and other harmful microorganisms, known as pathogens [26]. These pathogens can cause diseases, prompting an immune response in which antibodies are crucial in the body's defense system. They work by specifically identifying and attaching to parts of pathogens or foreign substances, such as toxins, a process fundamental in neutralizing or eliminating the threat.

Antibodies, shaped like a Y, are made up of two heavy and two light chains, play a crucial role in the immune defense system by targeting and binding to specific antigens, which are foreign substances that trigger an immune response [11]. The part of an antibody that binds to antigens, known as Complementarity-Determining Regions (CDRs), is key to this specificity. The third CDR of the heavy chain (H3[1]) is especially variable, allowing antibodies to bind to a wide variety of antigens [2]. The surface of the 3D conformation of the antibody protein, shaped by these CDRs, forms a specific topography that matches the topography of the target antigen's surface. When the antibody and antigen interact, these surfaces align like a lock and key mechanism [5]. This precise alignment is key for the immune system to distinguish between the body's own molecules and foreign ones, leading to the antigen's destruction and thus safeguarding the body from infection and disease [26].

[1] In this thesis, we refer to each CDR using the format [Chain][Number]; for example, H3 denotes the third CDR of the heavy chain

The significance of designing antibodies extends widely, notably in therapeutic and diagnostic fields [71]. In therapeutic applications, the design of antibodies is fundamental to creating targeted treatments, such as monoclonal antibodies, which have greatly improved the treatment of diverse diseases [11]. In diagnostics, engineered antibodies are essential for developing assays that accurately detect diseases, thereby enhancing the accuracy and precision of medical diagnostics [71]. The central task of antibody design is predicting the variable region sequences based on the structure and sequence of an antigen. This often includes focusing on CDRs, particularly H3, vital for antigen recognition.

Traditional methods for antibody design rely on complex physics-based calculations [45, 1]. However, the complexity and vast search space involved in this task can cause the traditional methods to be ineffective [38, 49]. Advancements in computational techniques have opened new avenues in antibody design, particularly through the adoption of deep learning methods. These methods focus on developing sophisticated neural networks to integrate both the sequence and the structural aspects of antibodies and antigens [29, 49, 37, 38]. However, despite their promise, these innovative models encounter substantial challenges. The more vital ones are the requirement for detailed contextual data [38] and the limited availability of training samples for model development [35]. Recognizing these challenges, our research is dedicated to overcoming these hurdles, aiming to fill a vital gap in the current landscape of antibody design.

The journey towards effective computational antibody design has faced notable challenges, which often slow down the creation of accurate and widely useful models. These challenges are as follows:

- **Limited availability of paired structural data**: One of the vital obstacles in antibody design is the shortage of extensive antibody-antigen structural data for training deep learning models. For instance, the SAbDab dataset [16], a widely-used resource containing paired antibody-antigen structures, includes only about 5,000 samples. This lack of data limits the models' overall efficacy and applicability by making it more difficult for them to learn from and generalize across a wide variety of antigens.
- **Impact of CDR flexibility on sequence prediction**: The structural flexibility of CDRs [28] presents a significant challenge in accurately predicting their sequences. Zheng et al. [77] emphasize that in such flexible regions, there is a weaker correlation between residue identities and their structural context. This aspect is particularly problematic in methods attempting to design both the sequence and structure of antibodies simultaneously, often leading to errors and resulting in the generation of inaccurate sequences.

- **Dependency on contextual information**: Existing models in antibody design often rely on additional data, like the structural arrangements of antibodies within their target environment (docked antibody frameworks) [49] or the specific shapes and conformations of antigenic epitopes [38]. Although this contextual information plays a vital role in enhancing the scalability and overall effectiveness of these methods, it is often challenging to obtain.

Out of the above challenges, the most pressing one is the scarcity of structural data. This shortage becomes particularly apparent when we contrast it with the abundance of available sequential data. For example, while the SAbDab dataset, offers only around 5,000 paired antibody-antigen structures, the Observed Antibody Space (OAS) [55] dataset contains about two billion single-chain as well as two million paired heavy and light chain antibody sequences. By utilizing this repository of sequential information, we can address the challenges posed by the lack of structural data, to develop more effective antibody design methods.

In addition, we were inspired by the successes of Protein Language Models (pLM) in various areas of protein modeling, including structure prediction [12] and design [77]. These models, structurally similar to language models such as BERT [14], are pre-trained on vast protein sequence databases and employ similar objectives, particularly masked language modeling. Through this training, pLMs learn the complex underlying patterns in protein sequences, understanding these biological structures [46].

To leverage pLMs for antibody design, we propose two solutions which are described as follows:

The first solution is a sequence-to-sequence model based on the transformer architecture developed by drawing an analogy to language translation. We treated antigens as the source, and antibody sequences as the target language. This perspective allows us to apply data augmentation techniques commonly used in neural machine translation (NMT), particularly the Back Translation to train our model in a semi-supervised manner [39]. This technique is especially beneficial in situations where there is a sparse or non-existent paired dataset between two languages. It utilizes large unpaired and unsupervised corpora of each language to train the translation model effectively [39].

Our first approach had a notable limitation: it could not utilize the structural information of antigens. To address this, we proposed a solution that could leverage both the sequential and structural information of antigens, as well as harness the capabilities of pLMs. In our second approach, we used a protein encoder to encode the antigen's information. This encoded information was then integrated into a pLM by specially designed

antigen adapter modules. These modules were crucial for enabling the pLM to generate antibody sequences conditioned on the specific antigen. By incorporating these adapter modules, the model could utilize the antigen's information in a targeted way, directly influencing the generation of corresponding antibody sequences.

An important aspect of both models is their ability to perform without reliance on contextual information. Additionally, focusing solely on antibody sequences, they are not affected by the structural flexibility of CDRs, which is a significant challenge in antibody design. These solutions aim to harness the available sequential data and the learning capabilities of pLMs, bridging the gap between extensive sequence information and limited structural data.

We evaluated these models on the Rosetta Antibody Design (RAbD) benchmark [1]. The first model, with its translation-like approach, demonstrated performance on par with current state-of-the-art models. The second model, which integrated both structural and sequential data as well as antigen adapter utilization, was able to outperform the existing models in recovering antibody sequences. This success not only showcases the efficacy of our approaches in addressing data scarcity but also highlights the potential of our methods in enhancing the accuracy and applicability of computational antibody design.

Chapter 2

Background

This chapter lays the groundwork for this thesis, bridging biological insights with computational developments in antibody design. It is structured into two key sections: *Biological Overview* and *Related Works*. The first section delves into the molecular biology of proteins, focusing particularly on the structure and function of antibodies, their crucial role in the immune system, and the complexities of their interactions with antigens. The second section transitions to an extensive review of the computational field, examining past research, the progression of machine learning techniques in antibody design, and the implementation of sophisticated models such as transformers [68]. This chapter establishes the foundation for a detailed exploration of cutting-edge computational methods in the later parts of the thesis.

2.1 Biological Overview

2.1.1 Proteins

Proteins are essential biomolecules in all living organisms, serving a wide array of functions critical to life. They are involved in nearly every process within cells and can be found in all body tissues and fluids. Their roles include acting as enzymes to catalyze biochemical reactions, providing structure and support, transporting and storing molecules, and participating in immune responses.

At their core, proteins are polymers composed of amino acids, linked together by peptide bonds. There are 20 different amino acids that can be combined in various sequences to

form proteins [74]. The sequence of these amino acids, known as the primary structure, dictates the protein's final three-dimensional shape and function. Proteins fold into specific structures, driven by interactions among the amino acids. This folding process results in the protein's secondary (e.g., alpha-helices and beta-sheets), tertiary (the overall 3D structure of a single polypeptide chain), and in some cases, quaternary structure (the arrangement of multiple polypeptide chains) [74] (Figure 2.1).

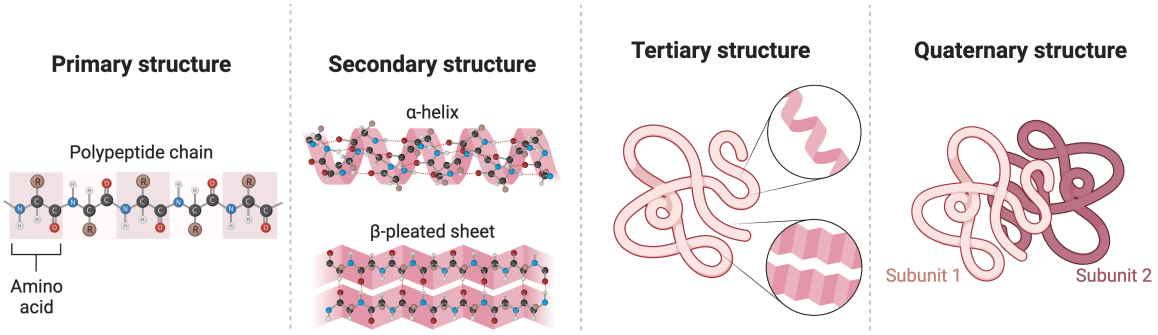

Figure 2.1: The four levels of protein structure. The primary structure is shown as a linear sequence of amino acids forming a polypeptide chain. The secondary structure is depicted with regions stabilized by hydrogen bonds, forming local structures such as alpha-helices and beta-pleated sheets. The tertiary structure represents the overall three-dimensional folding of a single polypeptide chain, influenced by interactions between the side chains of amino acids. Lastly, the quaternary structure is exemplified by the assembly of multiple subunits, each a polypeptide chain, coming together to form a functional protein complex. Adapted from "Protein Structure", by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates

2.1.2 Antibody

Antibodies, also known as immunoglobulins, are precise instruments of the immune system, adept at identifying and neutralizing specific pathogens. These proteins, with their distinctive Y shape, are produced by B cells, a type of white blood cell. They are essential for recognizing foreign molecules, or antigens, typically present on the surfaces of pathogens. The structure of an antibody includes two pairs of polypeptide chains: the larger heavy (H) chains and the smaller light (L) chains [3]. These chains feature constant regions, which define the antibody's class and initiate immune responses, and variable regions, which are crucial for antigen binding.

Antibodies are composed of constant and variable domains. In each antibody, the heavy chain includes three constant regions and one variable region, while the light chain has one constant and one variable region. Within each variable region, there are three CDRs, which collectively form the paratope—the specific binding site that attaches to an antigen's epitope. In this thesis, we refer to each CDR using the format [Chain][Number]; for example, H3 denotes the third CDR of the heavy chain (Figure Figure 2.2).

Surrounding these CDRs within the variable regions are sections called Frameworks (FWs). Each variable region is composed of four FWs (Figure Figure 2.2). Unlike the CDRs, which exhibit extreme variability, these framework regions are highly conserved, both in their sequence and structural conformation. This conservation plays a crucial role in maintaining the overall structure and stability of the antibody [53].

2.1.3 Antigen and Epitope

Antigens are substances that trigger an immune response, typically being a protein on the surface of pathogens. One key feature of these antigens is their epitopes - specific parts recognized by the immune system. Epitopes come in two forms: linear and conformational. Linear epitopes consist of amino acids arranged in a sequence, while conformational epitopes are composed of amino acids that assume a specific three-dimensional shape when the protein folds. Identifying and replicating conformational epitopes is more complex than linear ones because their recognition depends on the protein's tertiary structure. This structure can be altered during experimental procedures, making it a challenge to accurately determine and replicate the precise three-dimensional arrangement of amino acids in conformational epitopes. This complexity renders them elusive targets for immune responses and the development of antibody therapies.

2.1.4 Antibody Numbering Schemes

The structure of antibodies is quite complex, necessitating a universal language to accurately pinpoint and compare amino acid positions within their variable domains, which is essential for understanding antigen-binding specificity. The IMGT numbering system, introduced by Lefranc et al. [44], has become the standard for this purpose. It presents a systematic and uniform method, applicable across different species and antibody classes, assigning specific numbers to each amino acid and accounting for insertions and deletions. This enables a consistent description and analysis of all immunoglobulins. The

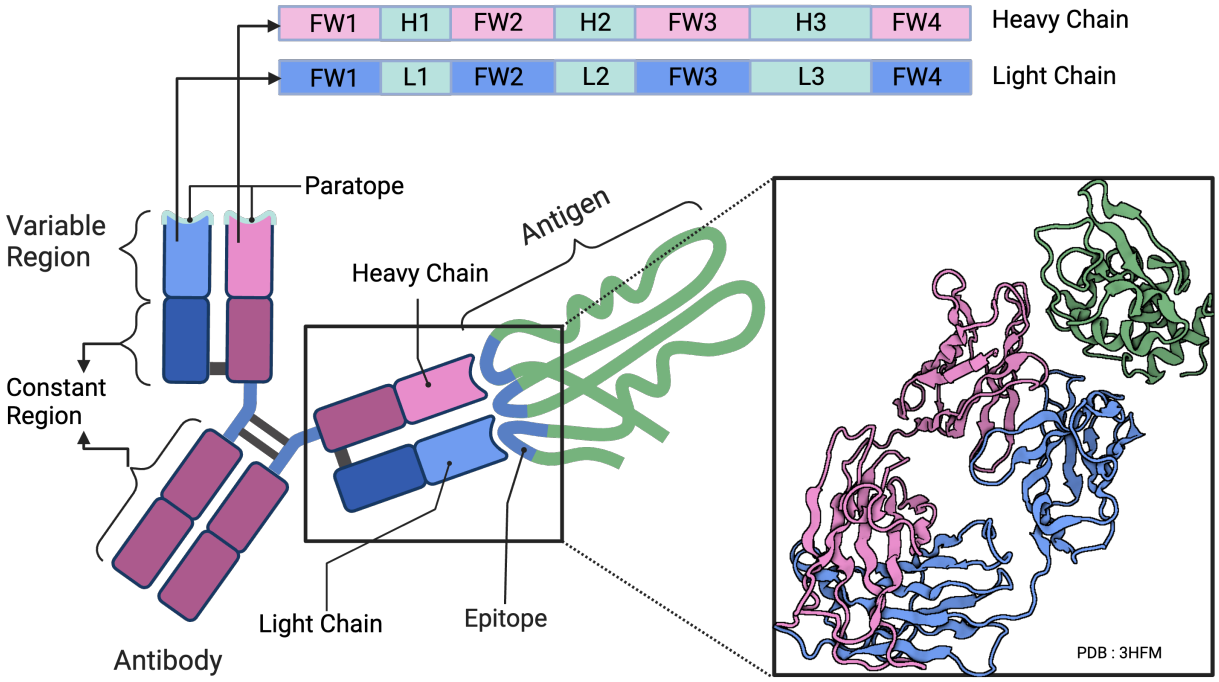

Figure 2.2: This schematic and structural representation displays the antibody's typical Y-shaped formation, consisting of two identical heavy chains and two identical light chains. Each chain is composed of alternating framework regions (FW) and complementarity-determining regions (CDRs). There are three CDRs (H1, H2, H3 for the heavy chain and L1, L2, L3 for the light chain) that are instrumental in binding to specific parts of the antigen, and four framework regions (FW1, FW2, FW3, FW4) that support the overall structure. The paratope, formed by the CDRs, engages the antigen's epitope in a precise lock-and-key manner. The right side of the image showcases the three-dimensional configuration of an antibody-antigen complex (PDB: 3HFM), highlighting the interaction between the antibody (in pink and blue) and the antigen (in green), with a focus on how the paratope's CDRs close contact with the antigen's epitope. Adapted from "Antigen Recognition by Antibodies", by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates

	FW1	CDR1	FW2	CDR2	FW3	CDR3	FW4
IMGT Position	1-26	27-38	39-55	56-65	66-104	105-117	118-128
Length	25-26	12 $\leq$	17	10 $\leq$	37-39	13 (or less or more)	10-11

Table 2.1: Summary of IMGT Position and Length Variability in Antibody Segments. This table presents the IMGT position ranges and corresponding length variability for the Framework (FW) and Complementarity-Determining Region (CDR) segments of antibodies.

IMGT system is especially valuable for uniformly identifying the CDRs in various antibody sequences. By offering a consistent framework for denoting precise positions within antibodies, the IMGT numbering enhances our capability to study, compare, and engineer antibodies accurately. The detailed IMGT position ranges and length variability for the FWs and CDRs of antibodies can be found in Table 2.1.

2.2 Related Work

2.2.1 Computational Antibody Design

Computational antibody design is an expanding field that uses various computational methods to predict antibody sequences for a given antigen. Traditional approaches often involve optimizing complex energy functions, a task complicated by the challenge of accurately simulating real-world protein interactions [45, 45, 42, 1, 73]. This complexity is due to the nature of protein-protein interactions which is too intricate to be captured entirely by statistical functions [21]. In response, deep learning has become increasingly prominent, branching into two main approaches: sequence-based methodologies and structure-sequence co-design.

Sequence-based models utilize deep learning to engineer antibodies by analyzing one-dimensional sequence data. The **EnsGrad** method, introduced by Liu et al. [48], employs gradient ascent alongside an ensemble of 24 neural networks to optimize antibody sequences. Meanwhile, Saka et al. [59] developed a generative LSTM model. Both models were trained using phage display libraries for a limited set of antigens, which restricts their ability to generalize to novel antigens. Another notable approach is **ReprogBert** [50], which reprograms a pre-trained English BERT model for protein sequence infilling, representing a novel application of language models in protein engineering.

while existing sequential approaches focus solely on antibody sequences without considering antigens, our first approach seeks to address this gap. We employ an encoder-decoder architecture, where the encoder encodes the antigen's information, and the decoder generates the corresponding binding antibody. This method involves fine-tuning Ankh [17], a T5 model [56] pre-trained on extensive protein sequence data. Our goal is to leverage the antigen's sequential information to guide antibody generation more effectively. Additionally, by incorporating transfer learning and data augmentation strategies, we aim to overcome the challenge posed by limited data availability.

The emergence of structure-sequence co-design methods marks a significant development in utilizing both sequence and structure data of proteins. **RefineGNN** [31], employing an autoregressive Graph Neural Network (GNN), predicts antibody sequence and structure, yet it doesn't design antibodies specific to antigens. Pioneering the field, **DiffAb** [49] successfully designs antigen-specific antibodies using generative diffusion models. **MEAN** [38] furthers this approach by applying GNN with E(3)-equivariant message passing, framing the task as graph translation. Building on this, Kong et al. [38] introduced **dyMEAN**, enhancing the model with an adaptive multi-channel equivariant encoder and a shadow paratope, allowing for a detailed consideration of each amino acid's atomic structure.

Our second approach addresses these limitations by taking a middle ground between sequential and co-design approaches. We employ graph modeling to encode both the sequence and structural information of the antigen. Then, we utilize sequence modeling to decode this information into the corresponding antibody sequence, effectively balancing the need for structural data with the advantages of sequence prediction.

2.2.2 Protein/Antibody Language Models

Protein language models such as ESM [46] and ProtTrans [18], which have been pretrained using the Masked Language Model (MLM) technique, have significantly enhanced our understanding of protein sequences. This advancement has inspired new developments in modeling antibody sequences, leading to the creation of specialized models like AbLang [66], AntiBERTa [43], AntiBERTy [57], and BALM [33]. AbLang, trained on the OAS database, specializes in filling in missing residues in B-cell repertoire sequences. AntiBERTa, utilizing a 12-layer transformer architecture, provides a nuanced numeric representation of antibody sequences, capturing essential elements of antibody functionality, and is adaptable for paratope prediction tasks. AntiBERTy, on the other hand, groups antibodies into clusters that mimic the process of affinity maturation. Meanwhile, BALM stands out

for its ability to predict both the function and structure of antibodies, exemplifying the significant role of machine learning in the field of immunology.

2.2.3 Protein Structural Encoding

The advancement of structure-based methodologies in computational biology has been significantly driven by innovations in encoding spatial characteristics of protein structures. Initially, these methodologies utilized 3D Convolutional Neural Networks (CNNs), as introduced by [13], and later incorporated the use of GNNs [20, 9, 32, 72, 8]. The field has recently seen a surge in exploring the potential of pre-training these structural encoders on extensive, unlabeled datasets. Pioneering contributions in this area, such as those by [25, 10, 23], have adopted approaches like contrastive learning, self-prediction, and denoising score matching. Additionally, Zhang el al. [75] have proposed the Geometry-Aware Relational Graph Neural Network (**GearNet**), utilizing relational graph convolutional layers and edge message passing with augmentation functions, offering a novel perspective in self-supervised learning for protein structures. These developments are expanding the boundaries of what can be achieved in the encoding of protein structural information.

2.2.4 Neural Machine Translation

Machine translation is a challenging task that involves converting text from one language to another, often using neural networks and encoder-decoder models. The encoder creates a vectorized representation of the source language sentence, which the decoder uses to generate the translated version. However, achieving high performance with these models requires large amounts of data, which can be difficult to obtain for low-resource languages. To address this challenge, researchers have explored unsupervised and semi-supervised learning methods to improve neural machine translation (NMT) performance. One such method is back translation [62], which involves translating monolingual text in both directions to generate synthetic bilingual data. The XLM model [40] has shown promising results with unsupervised data by leveraging back translation. Further research has aimed to refine the XLM model, including incorporating multilingual pre-trained BERT [78] and masked sequence-to-sequence techniques [63] to improve performance. Given the significant data requirements for antibody and antigen research, we explore the potential of semi-supervised NMT for antibody design in our first approach.

Chapter 3

Methodologies

This chapter details the methodologies employed to address the challenges of computational antibody design. The first proposed solution, a sequence-to-sequence model inspired by neural machine translation, utilizes a transformer architecture and transfer learning. Additionally, a back translation technique leverages unpaired data to improve performance. The second proposed solution leverages the capabilities of pLMs combined with antigen-specific adapter modules. This methodology employs a GNN protein encoder, pLM architecture, and antigen adapter modules to use both sequential and structural antigen information. Each methodology is comprehensively described, including its rationale, employed techniques, and parameters used.

3.1 First Approach: Antibody T5 (AbT5)

In this section, we delve into the details of our first approach, Antibody T5 (AbT5). We begin by introducing the key components that form the foundation of our approach. First, we present three distinct datasets that we employed for training our model, addressing the challenge posed by the shortage of available antigen-antibody paired dataset. Next, we explore the architecture of the Ankh model, a pre-trained pLM specifically optimized for protein-related tasks, which serves as the cornerstone of our approach. Lastly, we describe the loss functions involved in training, including supervised translation with paired data and unsupervised translation with unpaired datasets, utilizing back translation to enrich the training set and enhance the model's performance.

3.1.1 Datasets

Antigen sequences are generally much more diverse than antibody sequences, posing a significant challenge for modeling. Therefore, we chose to focus solely on the linear epitope region of the antigen sequences. This approach not only reduces the data complexity but also prioritizes the crucial information relevant to antibody design.

Our training methodology for Ankh uses three distinct datasets:

- Paired Dataset:

  This dataset provides the core foundation for supervised learning. Sourced from the SAbDab database [16], it consists of paired antigen and antibody complexes. To ensure data quality and prevent test data leakage, we clustered the samples based on the H3 region with a minimum sequence identity of 40%. Clusters containing RAbD benchmark data were excluded to prevent bias. The remaining clusters were then split into a training set (90%) and a validation set (10%). Importantly, only the epitope section of the antigen sequences was used, focusing on the top 20 closest positions to the antibody in their 3D structure. This approach prioritizes the relevant interaction region while minimizing irrelevant sequence information.

- Unpaired Antibody Dataset:

  To leverage the abundant sequential data for back translation, we utilized the OAS dataset [55]. This dataset contains approximately two million paired sequences of human heavy and light chain antibodies. To reduce redundancy and ensure data diversity, we clustered the sequences with a 60% sequence identity threshold and selected representative sequences from each cluster. This process yielded approximately 290,000 unique antibody sequences, significantly enriching the unpaired data pool for back translation.

- Unpaired Antigen Dataset:

  We used the Immune Epitope Database (IEDB) [70] for our unpaired antigen dataset. This database contains over 200,000 linear epitope sequences from 10,000 antigens. After removing duplicate and overlapping entries, we obtained a refined dataset of 40,000 non-redundant linear epitope sequences. These sequences, in conjunction with the unsupervised antibody dataset, enabled us to effectively utilize back translation for improved model learning.

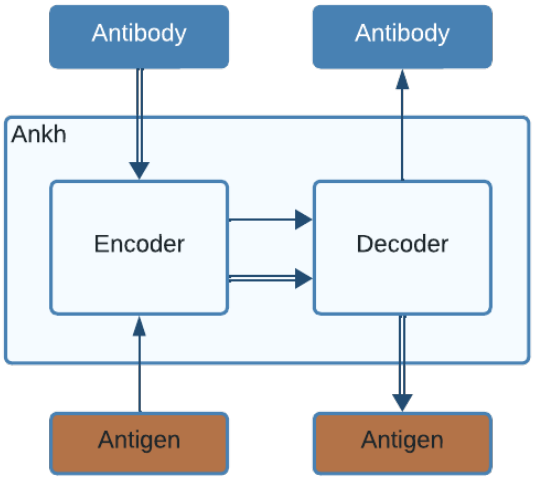

Figure 3.1: This schematic represents the encoder-decoder architecture used to model antibody design as a neural machine translation task, utilizing the Ankh model. The encoder processes the antigen sequence as input, encoding the information, which is then passed to the decoder. The decoder, in turn, generates the corresponding antibody sequence. The Ankh model, based on the T5 architecture, comprises a deep stack of encoder layers and autoregressive decoder layers pre-trained on a vast database of protein sequences. This structure allows Ankh to leverage pre-existing protein sequence knowledge, facilitating effective transfer learning for the specialized task of antibody design.

This combination of carefully curated datasets, encompassing both paired and unpaired data, provided the necessary foundation for the robust training of our model for antibody design.

3.1.2 Architecture

To model antibody design effectively as a neural machine translation task, we recognized the necessity of adopting an encoder-decoder architecture. In this architectural framework, the encoder takes the source sequence and encodes it, while the decoder generates the target sequence. Moreover, as highlighted in the introduction chapter, we were keen on harnessing the capabilities offered by pre-trained models. Given this aspiration, we decided to integrate **Ankh** [17] as the primary building block of our model.

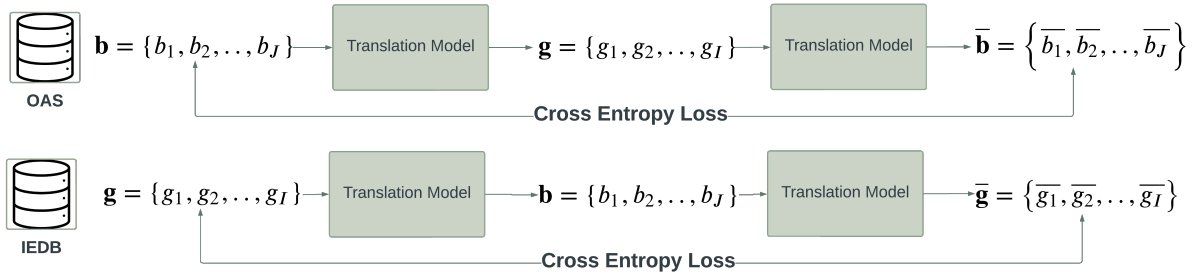

Figure 3.2: Back translation diagram for antibody-epitope sequence modeling, employing the IEDB dataset of linear epitopes and the OAS dataset of antibody sequences. In this process, the model initially translates epitopes from the IEDB dataset into corresponding antibody sequences. These generated sequences are then used as input to back-translate and recreate the original epitope sequences. The cross-entropy between the original and regenerated epitope sequences is used to determine the reconstruction's loss value. A similar procedure is conducted with the OAS dataset, where antibody sequences are first translated into epitopes and then back-translated to antibodies.

The Ankh model adheres closely to the T5 model, which relies on an encoder-decoder paradigm rooted in the transformer architecture. Specifically, our implementation utilizes *Ankh-base*, featuring 48 encoder layers and 24 autoregressive decoder layers. This configuration enhances Ankh's capacity to meet the demands of our computational antibody design task effectively.

In terms of pre-training, Ankh has undergone extensive pre-training on the Uniref50 [65] dataset, encompassing 45 million protein sequences. This pre-training process leveraged MLM techniques, enabling Ankh to acquire a deep understanding of protein sequences. As a result, Ankh is well-prepared for transfer learning, greatly improving its ability to handle new tasks such as antibody design. This adaptability is achieved without the need for excessive amounts of task-specific data, making it highly versatile.

3.1.3 Finetuning and Loss Functions

Our methodology is motivated by the analogy of viewing antigens and antibodies as different languages composed of amino acids. Consequently, we represent antibodies and antigens as sequences $\mathbf{g}$ and $\mathbf{b}$, respectively, denoted as:

$$\mathbf{g} = g_1^I = (g_1, g_2, ..., g_I), \text{ where } g_i \in \mathcal{A} \tag{3.1}$$

$$\mathbf{b} = b_1^J = (b_1, b_2, ..., b_J), \text{ where } b_j \in \mathcal{A} \tag{3.2}$$

Here $\mathcal{A}$ is the set of all 20 naturally occurring amino acids. This idea allows us to leverage the power of sequence-to-sequence models, traditionally used in language translation, to address the challenges of computational antibody design.

In our approach, we define the conditional probability distribution of translating antigens into antibodies using our auto-regressive model as:

$$P_{b\to g}(\mathbf{b}|\mathbf{g}) = \prod_{j=1}^{J} P_{b\to g}(b_j|b_1^{j-1}, c(\mathbf{g})), \tag{3.3}$$

where $P_{b\to g}(b_j|b_1^{j-1}, c(\mathbf{g}))$ represents the probability of generating j_{th} amino acid b_j in the antibody sequence, conditioned on the previously generated amino acids b_1^{j-1} and the context $c(\mathbf{g})$ derived from the encoded antigen sequence $\mathbf{g}$. This formulation captures the sequential nature of the translation process and dependency on the preceding amino acids and the antigen sequence. Similarly, we define the conditional probability distribution of antibodies into antigens as:

$$P_{g\to b}(\mathbf{g}|\mathbf{b}) = \prod_{i=1}^{I} P_{g\to b}(g_i|g_1^{i-1}, c(\mathbf{b})). \tag{3.4}$$

Considering this formulation, we incorporate two training strategies into our fine-tuning process, combining supervised and unsupervised learning, to mitigate the challenge posed by the limited availability of paired data.

Supervised Translation with Paired Data: We train Ankh with the paired antigen-antibody sequence dataset under a supervised framework. The primary objective here is to minimize the supervised cross-entropy loss between the model's predicted sequences and the actual target ones, defined as:

$$\mathcal{L}_{b\to g}^{sup} = \mathbb{E}[-\log P_{g\to b}(\mathbf{g}|\mathbf{b})], \tag{3.5}$$

$$\mathcal{L}_{g\to b}^{sup} = \mathbb{E}[-\log P_{b\to g}(\mathbf{b}|\mathbf{g})]. \tag{3.6}$$

This training allows Ankh to accurately grasp the correlations between antigen and antibody sequences, thereby enhancing the precision of its translational capabilities.

Unsupervised Translation with Unpaired Datasets: Leveraging the extensive unpaired datasets previously mentioned, we employ back translation as an unsupervised

learning technique (Figure 3.2). Let $v^*(\mathbf{g})$ be the antibody sequence inferred from the translation of antigen sequence $\mathbf{g}$ such that $v^*(\mathbf{g}) = \text{argmax}\, P_{g\to b}(v|\mathbf{g})$. Similarly, we denote $u^*(\mathbf{b})$ as the antigen sequence resulted from the translation of antibody sequence $\mathbf{b}$ such that $u^*(\mathbf{b}) = \text{argmax}\, P_{b\to g}(u|\mathbf{b})$. We consider $(v^*(\mathbf{g}), \mathbf{g})$ and $(\mathbf{b}, u^*(\mathbf{b}))$ as synthetic paired data to construct the cross-entropy loss functions for the unsupervised translation process as:

$$\mathcal{L}_g^{back} = \mathbb{E}[-\log P_{b\to g}(\mathbf{g}|v^*(\mathbf{g}))], \tag{3.7}$$

$$\mathcal{L}_b^{back} = \mathbb{E}[-\log P_{g\to b}(\mathbf{b}|u^*(\mathbf{b})]. \tag{3.8}$$

It should be noted that although we need the model to only translate from antigen to antibody, we train the model to translate in both directions. This helps the model in producing high-quality synthetic samples with back translation. We update the model's parameters using gradient descent by minimizing each loss separately. Additionally, to achieve more precise antibody prediction, we keep the pre-trained encoder's parameters frozen and only train the decoder's weights during backward passes. This allows the model to focus on task-specific decoding and also retain the encoder's understanding of protein sequences.

In the fine-tuning process for Ankh, we also implemented two key features to enhance its performance in antibody design. We created special tokens for each language — one for antibodies and one for antigens. These tokens act as prompts for the decoder's input, guiding the model to accurately translate the sequences according to the specified language. This addition is crucial for ensuring that the model correctly identifies and translates each sequence based on whether it represents an antibody or an antigen.

3.2 Second Approach: Conditional-BALM (CBALM)

3.2.1 Task Formulation

The input antigen is denoted as G, represented by a sequence-structure tuple $(s_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$, where $s_i \in \mathcal{A}$ is the type of amino acids at position i, and $\mathbf{x}_i \in \mathbb{R}^3$ is its alpha carbon atom coordinates. Our objective is to predict the antigen's corresponding antibody sequence B, consisting of heavy (H) and light (L) chains, given the antigen information G. The task aims to learn parameters θ to maximize the conditional probability:

$$\max_{\theta} P(B_{\text{masked}}|G, B_{\text{known}}) \tag{3.9}$$

In specific cases, part of the antibody sequence, typically the framework, might already be identified, indicated as B_{known}. Our goal is to accurately predict the unspecified segments of the antibody sequence, referred to as B_{masked}. For the case of full antibody prediction, no section is known, hence $B_{\text{known}} = \varnothing$.

3.2.2 Architecture

Informed by our previous discussions, we elected to concentrate solely on the design of the antibody sequence. Our objective is to gain a deeper understanding of the antigen to effectively create a complementary antibody sequence. To achieve this, we adopted an encoder-decoder architecture.

The encoder in our framework is tasked with capturing the antigen's structural and sequential details comprehensively. The decoder, on the other hand, uses this information to produce the corresponding antibody sequence, systematically incorporating all essential aspects of the antigen. We utilized GearNet, a structural encoder pre-trained on the AlphaFold dataset (Figure 2A), for antigen encoding. This pre-training process equipped GearNet with a robust understanding of protein structures, making it adept at accurately encoding the information of antigens. The encoder function ENC(G) produces H_G, which is in the shape of $n \times d_{\text{GearNet}}$, where n is the length of the antigen sequence and d_{GearNet} is the dimensionality of the encoded space. H_G is then passed through an adapter module, undergoing a linear transformation with weights W_{AdaptEnc} to produce E_G such that $E_G = W_{\text{AdaptEnc}} \cdot H_G$, and we parameterize the dimensionality of E_G with d_{AdaptEnc}. The transformed representation E_G will be passed to the decoder, serving as the foundation to generate corresponding antibody sequences and translating the complex antigen information into a format usable for antibody design.

Our decoding component, the Bio-inspired Antibody Language Model (BALM), leverages a transformer-based self-attention mechanism with 30 layers and rotary positional encoding to understand the unique and conserved properties of antibodies (Figure 2B). Initially trained on a large dataset of unlabeled antibody sequences, BALM effectively captures contextual embeddings essential for inferring binding functions. Notably, BALM employs a unique antibody positional encoding method based on the IMGT numbering scheme, which provides a consistent framework for identifying amino acid positions, thereby enhancing its ability to generate meaningful embeddings.

In our methodology, the antibody sequence representation provided to BALM adopts a structure similar to ESM-2: it commences with a [CLS] token, is followed by the heavy chain sequence, and concludes with an [EOS] token. To adapt BALM for conditional

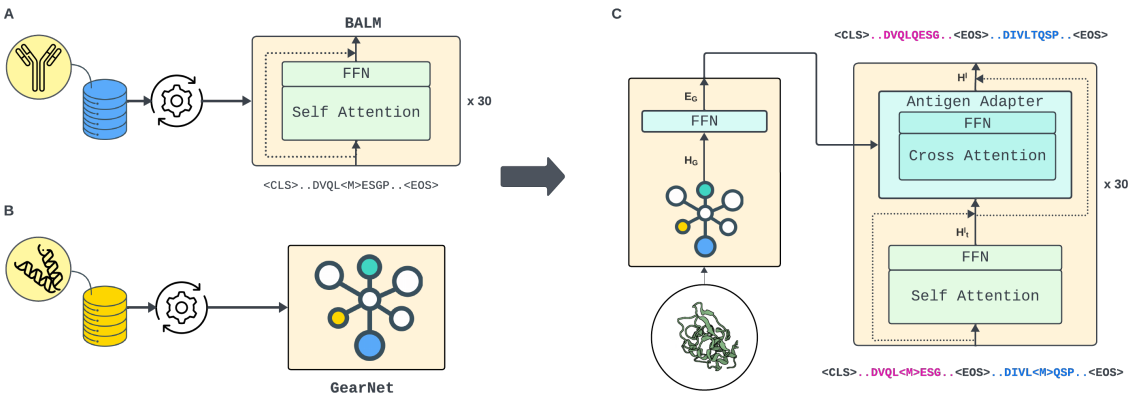

Figure 3.3: This diagram showcases our proposed encoder-decoder model for antibody sequence prediction, combining the strengths of two pre-trained networks. Panel A features BALM, a transformer-based model pre-trained on a vast dataset of single-chain antibodies using a Masked Language Model (MLM) objective for sequence processing. In Panel B, GearNet is presented as a general-purpose protein structure encoder, pre-trained on the comprehensive AlphaFold dataset to encode structural data into graph form. Panel C unifies these components: GearNet's graph-encoded antigen information is channeled into BALM via an antigen adapter, using cross-attention mechanisms to translate intricate antigen structures into corresponding antibody sequences.

generation crucial to our approach, we integrated antigen adapter layers after each self-attention layer within BALM. This integration facilitates the efficient incorporation of the encoded antigen information into the decoding process. Each antigen adapter layer consists of a cross-attention layer followed by a feed-forward layer. In the cross-attention operation denoted as CrossAtt(Q, K, V), the output of each block in adapted BALM is given by

$$H^l = W_{\text{FFN}} \cdot \text{CrossAtt}(H_t^l, E_G, E_G) + H_t \tag{3.10}$$

where l signifies the layer number, E_G represents the encoded antigen information, and W_{FFN} are the weights of the feed-forward layer, and H_t is the output from the previous feed-forward and self-attention layers. This refined structure ensures a synergistic interplay between the antigen's structural and sequential information, aligning with our objective of generating corresponding antibody sequences based on the encoded antigen information.

3.2.3 Training

The training objective in our model is the Causal Masked Language Modeling (CMLM) objective [19], facilitating the non-autoregressive generation of antibody sequences conditioned on antigen information. Under this setup, a set of tokens from B (the antibody sequence) is replaced with the [MASK] token, with the number of masked tokens being uniformly sampled from 1 to $|B|$, where $|B|$ is the length of the antibody sequence. To enhance the model's focus on learning to predict the more challenging Complementarity-Determining Regions (CDR) of the antibody, as opposed to the relatively constant framework regions, we employ Focal Loss for our training loss. Focal Loss, commonly used in object detection tasks, is adept at handling unbalanced data and helps in redirecting the model's focus toward learning the intricate CDR regions, which is pivotal for our task.

The training of our model was conducted using the Adam optimizer, paired with an inverse square root learning rate scheduler. This scheduler featured a warm-up phase of 1,000 steps, after which the learning rate peaked at 10^{-4}. To preserve the integrity of the pre-trained models, we initially kept all pre-trained weights frozen during the first five epochs. After this period, we unfroze the pre-trained weights and set their learning rate to a tenth of that of the newly added layers. This cautious approach allowed for end-to-end training, providing the model with the opportunity to gradually adjust to the specific characteristics of our dataset.

Chapter 4

Experiments

To validate our model, we employed the RAbD benchmark, consisting of 60 antigen-antibody complexes, providing a robust framework for evaluating the model's performance in real-world scenarios. In our assessment of the model's predictive performance, we focused on three main metrics: Amino Acid Recovery (AAR), Contact Amino Acid Recovery (CAAR), and Root Mean Square Deviation (RMSD). AAR evaluates how accurately the model predicts the amino acid sequence of antibodies, showing the percentage of residues it correctly identifies. CAAR takes a more specific approach, assessing the model's ability to predict residues that interact with the antigen — crucial for determining binding affinity and specificity. As our models do not directly predict protein structure, we utilized the dyMEAN structure prediction model, available on its official GitHub repository, to predict the complex tertiary structure for the generated antibody sequences. We then used RMSD to measure structural accuracy, calculating the average deviation between the backbone atoms of the predicted and actual protein structures post-alignment using the Kabsch algorithm. A lower RMSD indicates higher structural fidelity. By combining AAR and CAAR for sequence accuracy and RMSD for structural accuracy through dyMEAN predictions, we comprehensively evaluate our model's efficacy in replicating both the antibody sequences and their functional conformations.

4.1 Single CDR Design

In this experiment, we address the challenge of predicting a single CDR of the heavy chain, provided the light chain and the remaining domains of the heavy chain are given. In our first approach, we utilize the sequence preceding the target CDR as a prompt for the

Model	H1 AAR	H1 CAAR	H1 RMSD
Diffab	68.00%	60.02%	0.55
AbODE	70.50%	-	0.65
dyMEAN	76.55%	63.19%	0.56
AbT5	69.72%	56.29%	-
CBALM	**80.56%**	**69.44%**	**0.47**

Model	H2 AAR	H2 CAAR	H2 RMSD
Diffab	54.64%	42.77%	**0.41**
AbODE	55.70%	-	0.73
dyMEAN	69.52%	63.50%	0.48
AbT5	70.08%	68.87%	-
CBALM	**71.25%**	**66.46%**	0.48

Model	H3 AAR	H3 CAAR	H3 RMSD
Diffab	37.47%	20.88%	2.08
AbODE	39.80%	-	1.73
dyMEAN	43.20%	27.83%	**1.58**
AbT5	37.72%	27.78%	-
CBALM	**45.68%**	**31.57%**	1.63

Table 4.1: Comparative Performance of AbT5, CBALM, Diffab, and dyMEAN in single CDR design. The tables showcase results for single CDR design of heavy antibody's heavy chain, with Amino Acid Recovery (AAR), Contact Amino Acid Recovery (CAAR), and Root Mean Square Deviation (RMSD) metrics calculated specifically for the target CDR.

decoder, asking the model to generate the sequence following this prompt. Our second approach leverages a masking and unmasking technique where the target CDR is initially masked to obscure its sequence from the model; subsequently, the model engages in a prediction task, trying to accurately unmask and generate the sequence for the obscured CDR.

Results As indicated in Table 4.1, the AbT5 model has shown its ability to perform on par with other established methods in predicting antibody sequences. Notably, AbT5 achieves this without incorporating any structural information, relying solely on sequence-based data. This achievement is significant, demonstrating the strength and potential of our sequence-to-sequence model in computational antibody design. Additionally, our CBALM model demonstrated remarkable proficiency in predicting the sequence of a single CDR of the heavy chain, achieving this without relying on additional contextual information. This success highlights CBALM's ability to effectively use context from conserved antibody regions and to integrate targeted antigen information through adapter modules, enabling precise engineering of the variable CDR regions critical for antigen recognition and binding. Notably, CBALM outshone models like DiffAb, AbODE, and dyMEAN in design-

Model	L1 AAR	L2 AAR	L3 AAR	H1 AAR	H2 AAR	H3 AAR
Diffab	52.02%	56.03%	47.88%	68.97%	54.64%	39.82%
dyMEAN	**75.89%**	**83.65%**	52.66%	76.13%	68.48%	37.51%
CBALM	63.20%	75.40%	**61.49%**	**79.76%**	**70.23%**	**44.98%**

	Overall		
Model	AAR	CAAR	RMSD
Diffab	51.26%	40.35%	1.95
dyMEAN	60.36%	50.30%	1.35
CBALM	**63.00%**	**53.67%**	**1.01**

Table 4.2: Results for multi-CDR design, where all six CDRs of the antibody sequence are generated simultaneously, with RMSD calculated for the entire antibody structure.

ing CDR-H3, showcasing its advanced sequence prediction capabilities. This indicates the benefits of focusing exclusively on sequence information. Impressively, CBALM exceeded all state-of-the-art models in sequence-based metrics and most in structure-based metrics. However, it is essential to consider that for structure-based metrics, the employment of an external structure prediction tool might have introduced some inaccuracies. Therefore, while CBALM's performance in structure-based metrics is significant, it should be evaluated with an understanding of the potential limitations associated with the structure prediction method employed.

4.2 Multi CDR Design

In this experiment we tackled a more complex challenge: predicting the sequences of all six CDRs simultaneously—three from the heavy chain and three from the light chain. This task was executed given the antigen structure and the framework sequence of the antibody. Our approach in this scenario demonstrates the CBALM model's robustness in handling a comprehensive design task, where multiple variable regions are predicted in concert. This not only tests the model's ability to integrate and process intricate antigen information but also its capacity to simultaneously manage multiple CDRs.

Results In analyzing the results of Table 4.2 the multi CDR design experiment, the CBALM model's performance distinctly highlights its exceptional capability in handling complex antibody design tasks. The model's ability to predict all six CDRs simultaneously, with less dependency on the known antibody sections than other methods, is particularly

noteworthy. This indicates CBALM's advanced proficiency in not only processing detailed antigen information but also in effectively managing the intricate interactions of multiple variable regions. The results from this experiment, showcasing higher AAR and CAAR along with a significantly lower RMSD, clearly demonstrate CBALM's superiority in comprehensive and accurate antibody sequence prediction

4.3 Whole Variable Region Prediction

In this critical phase of our research, we extended the scope of CBALM's capabilities by challenging it to predict the entire variable region of antibodies. This includes both the heavy and light chains, encompassing all the CDRs and framework regions. Unlike the previous tasks that focused on specific CDRs, this task requires the model to generate a more extensive and complete sequence, reflecting the full complexity of an antibody's variable region. This comprehensive prediction task tests CBALM's ability to synthesize complex antigen information and apply it across a broader sequence range, embodying a significant step towards practical applications in antibody design and development.

Model	AAR	CAAR
dyMEAN	70.35%	40.02%
AbT5	59.88%	30.41%
CBALM	**72.07%**	**42.69%**

Table 4.3: Results of whole Antibody Prediction

Results In the results of the whole variable region prediction experiment (Table 4.3), our models were compared against dyMEAN, the only other method capable of addressing this extensive task. Remarkably, CBALM surpassed dyMEAN in all metrics for predicting the entire variable region of the antibody. This includes both heavy and light chains, showcasing CBALM's superior ability to synthesize and apply complex antigen information across a broader sequence. However, AbT5 did not exhibit satisfactory performance due to its inability to use antibody numbering information to guide the antibody generation.

Chapter 5

Conclusion

Our study presents two approaches in the field of computational antibody design, each addressing the intricate challenges of data scarcity and the complex nature of antibody structures.

The first methodology, AbT5, is a sequence-to-sequence model that draws inspiration from neural machine translation. Utilizing a transformer architecture and transfer learning, AbT5 demonstrates a unique approach to antibody sequence prediction. This model particularly excels in leveraging unpaired data through back translation, enhancing its performance and ability to handle the intricacies of antibody design.

The second methodology, CBALM, builds upon the strengths of pLMs, integrating these with antigen-specific adapter modules. CBALM employs a GNN protein encoder alongside a pLM architecture. This combination enables the effective use of both sequential and structural antigen information, marking a significant step in the evolution of antibody design methodologies.

Together, these two approaches provide a comprehensive framework for tackling the challenges in computational antibody design. AbT5 offers a robust solution for sequence prediction, while CBALM expands the frontier by incorporating structural data into the prediction process. Both methodologies, with their distinct yet complementary capabilities, pave the way for significant advancements in the field, enhancing the precision and applicability of computational techniques in antibody design.

5.1 Future Work

Our research has opened several promising pathways for future exploration and enhancement in the field of computational antibody design. Building upon the foundations laid by our two novel approaches, AbT5 and CBALM, we propose the following directions for future work:

A significant area for future development in our first approach, AbT5, involves the incorporation of structural data into the back translation process. Currently, AbT5 excels in sequence prediction but integrating structural information could further refine its predictive capabilities. Research into methodologies that can seamlessly blend structural insights with sequence data in the back translation process could lead to a more holistic model, potentially improving the accuracy and applicability of antibody predictions.

The second approach, CBALM, is uniquely modular, allowing for the integration of more advanced protein structure encoders. Future work could explore the use of cutting-edge protein structure encoding techniques to enhance CBALM's performance. By updating or replacing current encoders with more sophisticated alternatives, we can expect to see improvements in the model's ability to process and utilize structural data, leading to more accurate antibody design predictions.

Both AbT5 and CBALM, while effective, highlight the need for better metrics in evaluating the generated antibodies. Current assessment methods may not fully capture the nuances and complexities of antibody-antigen interactions. Future research should focus on developing more refined and comprehensive metrics that can accurately assess the efficacy and specificity of predicted antibodies.

In summary, the potential enhancements in incorporating structural data in AbT5, the integration of advanced protein structure encoders in CBALM, and the development of improved evaluation metrics represent crucial steps toward the evolution of computational antibody design. These advancements could lead to more precise, effective, and clinically relevant antibody therapies, ultimately contributing significantly to biomedical research and healthcare.

References

[1] Jared Adolf-Bryfogle, Oleks Kalyuzhniy, Michael Kubitz, Brian D. Weitzner, Xiaozhen Hu, Yumiko Adachi, William R. Schief, and Roland L. Dunbrack. Rosettaantibodydesign (rabd): A general framework for computational antibody design. *PLOS Computational Biology*, 14:e1006112, 04 2018.

[2] Rahmad Akbar, Habib Bashour, Puneet Rawat, Philippe A Robert, Eva Smorodina, Tudor-Stefan Cotet, Karine Flem-Karlsen, Robert Frank, Brij Bhushan Mehta, Mai Ha Vu, et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *MAbs*, 14(1):2008790, 2022.

[3] Rahmad Akbar, Philippe A. Robert, Cédric R. Weber, Michael Widrich, Robert Frank, Milena Pavlović, Lonneke Scheffer, Maria Chernigovskaya, Igor Snapkov, Andrei Slabodkin, Brij Bhushan Mehta, Enkelejda Miho, Fridtjof Lund-Johansen, Jan Terje Andersen, Sepp Hochreiter, Ingrid Hobæk Haff, Günter Klambauer, Geir Kjetil Sandve, and Victor Greiff. In silico proof of principle of machine learning-based antibody design at unconstrained scale. *mAbs*, 14:2031482, 2022.

[4] Bissan Al-Lazikani, Arthur M. Lesk, and Cyrus Chothia. Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology*, 273(4):927–948, 1997.

[5] Mohammed M. Al Qaraghuli, Karina Kubiak-Ossowska, Valerie A. Ferro, and Paul A. Mulheran. Antibody-protein binding and conformational changes: identifying allosteric signalling pathways to engineer a better effector response. *Scientific Reports*, 10(1):13696, 2020.

[6] B Alberts, A Johnson, J Lewis, et al. *Molecular Biology of the Cell*. Garland Science, New York, 4 edition, 2002.

[7] N. N. Author. Suppressed for anonymity, 2021.

[8] Sarp Aykent and Tian Xia. Gbpnet: Universal geometric representation learning on protein structures. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 4–14, New York, NY, USA, 2022. Association for Computing Machinery.

[9] Federico Baldassarre, David Menéndez Hurtado, Arne Elofsson, and Hossein Azizpour. Graphqa: protein model quality assessment using graph convolutional networks. *Bioinformatics*, 37(3):360–366, 2021.

[10] Can Chen, Jingbo Zhou, Fan Wang, Xue Liu, and Dejing Dou. Structure-aware protein self-supervised learning. *Bioinformatics*, 39(4):btad189, 2023.

[11] Mark L Chiu, Dennis R Goulet, Alexey Teplyakov, and Gary L Gilliland. Antibody structure and function: the basis for engineering therapeutics. *Antibodies*, 8(4):55, 2019.

[12] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M. Church, Peter K. Sorger, and Mohammed AlQuraishi. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

[13] Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, 06 2018.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[15] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.

[16] James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M. Deane. Sabdab: the structural antibody database. *Nucleic Acids Research*, 42(D1):D1140–D1146, 2014.

[17] Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv preprint arXiv:2301.06568*, 2023.

[18] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Wang Yu, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021.

[19] Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models, 2019.

[20] Vladimir Gligorijević, P Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C Taylor, Ian M Fisk, Hera Vlamakis, et al. Structure-based protein function prediction using graph convolutional networks. *Nature communications*, 12(1):3168, 2021.

[21] Jordan Graves, Jacob Byerly, Eduardo Priego, Naren Makkapati, S. Parish, Brenda Medellin, and Monica Berrondo. A review of deep learning methods for antibodies. *Antibodies*, 9:12, 04 2020.

[22] Junliang Guo, Zhirui Zhang, Linli Xu, Hao-Ran Wei, Boxing Chen, and Enhong Chen. Incorporating bert into parallel sequence decoding with adapters, 2020.

[23] Yuzhi Guo, Jiaxiang Wu, Hehuan Ma, and JunZhou Huang. Self-supervised pre-training for protein embeddings using tertiary structures. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 6801–6809. AAAI Press, 2022.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Pedro Hermosilla and Timo Ropinski. Contrastive representation learning for 3d protein structures, 2022.

[26] C. Janeway and P. Travers. *Immunobiology: The Immune System in Health and Disease*. Current Biology Limited, 1994.

[27] Charles A. Janeway, Paul Travers, Mark Walport, and Mark J. Shlomchik. *Immunobiology: The immune system in health and disease*. Garland Science, 5 edition, 2001.

[28] Jeliazko R. Jeliazkov, Adnan Sljoka, Daisuke Kuroda, Nobuyuki Tsuchimura, Naoki Katoh, Kouhei Tsumoto, and Jeffrey J. Gray. Repertoire analysis of antibody cdr-h3 loops suggests affinity maturation does not typically result in rigidification. *Frontiers in Immunology*, 9, 2018.

[29] W. Jin, J. Wohlwend, R. Barzilay, and T. Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.

[30] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Antibody-antigen docking and design via hierarchical structure refinement. In *International Conference on Machine Learning*, pages 10217–10227. PMLR, 2022.

[31] Wengong Jin, Jeremy Wohlwend, Regina Barzilay, and Tommi Jaakkola. Iterative refinement graph neural network for antibody sequence-structure co-design. *arXiv preprint arXiv:2110.04624*, 2021.

[32] Bowen Jing, Stephan Eismann, Pratham N. Soni, and Ron O. Dror. Equivariant graph neural networks for 3d macromolecular structure, 2021.

[33] Hongtai Jing, Zhengtao Gao, Sheng Xu, Tao Shen, Zhangzhi Peng, Shwai He, Tao You, Shuang Ye, Wei Lin, and Siqi Sun. Accurate prediction of antibody function and structure using bio-inspired antibody language model. *bioRxiv*, pages 2023–08, 2023.

[34] M. J. Kearns. *Computational Complexity of Machine Learning*. PhD thesis, Department of Computer Science, Harvard University, 1989.

[35] Jisun Kim, Matthew McFee, Qiao Fang, Osama Abdin, and Philip M Kim. Computational and artificial intelligence-based methods for antibody development. *Trends in Pharmacological Sciences*, 2023.

[36] Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation, 2022.

[37] Xiangzhe Kong, Wenbing Huang, and Yang Liu. Conditional antibody design as 3d equivariant graph translation, 2023.

[38] Xiangzhe Kong, Wenbing Huang, and Yang Liu. End-to-end full-atom antibody design, 2023.

[39] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[40] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining, 2019.

[41] P. Langley. Crafting papers on machine learning. In Pat Langley, editor, *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pages 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

[42] Gideon D. Lapidoth, Dror Baran, Gabriele M. Pszolla, Christoffer Norn, Assaf Alon, Michael D. Tyka, and Sarel J. Fleishman. Abdesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins*, 83:1385–1406, 08 2015.

[43] Jinwoo Leem, Laura S. Mitchell, James H.R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022.

[44] Marie-Paule Lefranc, C Pommié, M Ruiz, V Giudicelli, E Foulquier, L Truong, V Thouvenin-Contet, and G Lefranc. Imgt unique numbering for immunoglobulin and t cell receptor variable domains and ig superfamily v-like domains. *Developmental and Comparative Immunology*, 27(1):55–77, 2003.

[45] Tong Li, Robert J. Pantazes, and Costas D. Maranas. Optmaven – a new framework for the de novo design of antibody variable region models targeting specific antigen epitopes. *PLoS ONE*, 9:e105954, 08 2014.

[46] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[47] G. Liu, H. Zeng, J. Mueller, B. Carter, Z. Wang, J. Schilz, G. Horny, M.E. Birnbaum, S. Ewert, and D.K. Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36(7):2126–2133, 2020.

[48] Ge Liu, Haoyang Zeng, Jonas Mueller, Brandon Carter, Ziheng Wang, Jonas Schilz, Geraldine Horny, Michael E Birnbaum, Stefan Ewert, and David K Gifford. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 36:2126–2133, 11 2019.

[49] Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *bioRxiv*, 2022.

[50] Igor Melnyk, Vijil Chenthamarakshan, Pin-Yu Chen, Payel Das, Amit Dhurandhar, Inkit Padhi, and Devleena Das. Reprogramming pretrained language models for antibody sequence infilling, 2023.

[51] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, editors. *Machine Learning: An Artificial Intelligence Approach, Vol. I.* Tioga, Palo Alto, CA, 1983.

[52] T. M. Mitchell. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

[53] Veronica Morea, Arthur M. Lesk, and Anna Tramontano. Antibody modeling: Implications for engineering and design. *Methods*, 20(3):267–279, 2000.

[54] A. Newell and P. S. Rosenbloom. Mechanisms of skill acquisition and the law of practice. In J. R. Anderson, editor, *Cognitive Skills and Their Acquisition*, chapter 1, pages 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

[55] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022.

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[57] Jeffrey A Ruffolo, Jeffrey J Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv preprint arXiv:2112.07782*, 2021.

[58] K. Saka, T. Kakuzaki, S. Metsugi, D. Kashiwagi, K. Yoshida, M. Wada, H. Tsunoda, and R. Teramoto. Antibody design using lstm based deep generative model from phage display library for affinity maturation. *Scientific reports*, 11(1):1–13, 2021.

[59] Koichiro Saka, Taro Kakuzaki, Shoichi Metsugi, Daiki Kashiwagi, Kenji Yoshida, Manabu Wada, Hiroyuki Tsunoda, and Reiji Teramoto. Antibody design using lstm based deep generative model from phage display library for affinity maturation. *Scientific Reports*, 11, 03 2021.

[60] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):211–229, 1959.

[61] Harry W. Schroeder and Lisa Cavacini. Structure and function of immunoglobulins. *The Journal of Allergy and Clinical Immunology*, 125(2):S41–S52, 2010.

[62] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[63] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936, 2019.

[64] Y Sun and Y Shen. Structure-informed protein language models are robust predictors for variant effects. *Res Sq*, Aug 3 2023. Preprint.

[65] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 11 2014.

[66] Iain H. Moal Tobias H. Olsen and Charlotte M. Deane. Ablang: An antibody language model for completing antibody sequences. *bioRxiv*, 2022.

[67] C.-L. Towse and V. Daggett. When a domain isn't a domain, and why it's important to properly filter proteins in databases: Conflicting definitions and fold classification systems for structural domains makes filtering of such databases imperative. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 34(12):1060, 2012.

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[69] Yogesh Verma, Markus Heinonen, and Vikas Garg. Abode: Ab initio antibody design using conjoined odes, 2023.

[70] Randi Vita, Swapnil Mahajan, James A Overton, Sandeep Kumar Dhanda, Sheridan Martini, Jason R Cantrell, Daniel K Wheeler, Alessandro Sette, and Bjoern Peters. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343, 10 2018.

[71] Thomas A. Waldmann. Monoclonal antibodies in diagnosis and therapy. *Science*, 252(5013):1657–1662, 1991.

[72] Limei Wang, Haoran Liu, Yi Liu, Jerry Kurtin, and Shuiwang Ji. Learning hierarchical protein representations via complete 3d graph networks, 2023.

[73] Shira Warszawski, Aliza Borenstein Katz, Rosalie Lipsh, Lev Khmelnitsky, Gili Ben Nissan, Gabriel Javitt, Orly Dym, Tamar Unger, Orli Knop, Shira Albeck, Ron Diskin, Deborah Fass, Michal Sharon, and Sarel J. Fleishman. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLOS Computational Biology*, 15:e1007207, 08 2019.

[74] Malcolm Watford and Guoyao Wu. Protein. *Adv. Nutr.*, 9(5):651–653, September 2018.

[75] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining, 2023.

[76] Zuobai Zhang, Minghao Xu, Arian Jamasb, Vijil Chenthamarakshan, Aurelie Lozano, Payel Das, and Jian Tang. Protein representation learning by geometric structure pretraining. In *International Conference on Learning Representations*, 2023.

[77] Zaixiang Zheng, Yifan Deng, Dongyu Xue, Yi Zhou, Fei YE, and Quanquan Gu. Structure-informed language models are protein designers. In *International Conference on Machine Learning*, 2023.

[78] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*, 2020.

[79] Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation, 2020.