Charles University
Faculty of Science

Study programme: Bioinformatics

Bc. Adam Král

Predicting Protein Thermostability With a Focus on Antibodies

Predikce termostability proteinů se zaměřením na protilátky

MASTER'S THESIS

Supervisor: Danny Asher Bitton, PhD

Prague 2025

I declare that I carried out this master's thesis on my own, and only with the cited sources, literature and other professional sources. This thesis, or any substantial part of it, has not been submitted for the award of any other or the same academic degree. Artificial intelligence tools were used solely to enhance the readability of my own text and to improve the resolution of figures reproduced from other sources.

In ..... date ....

Author's signature

I would like to thank my advisor, David Příhoda, for his positive outlook, valuable advice, and for guiding me toward this research direction. I especially appreciate his consistent support. I am also grateful to Kadina Johnston for her suggestions and help with proofreading. Finally, I would like to thank my supervisor, Danny Bitton, for giving me the opportunity to carry out my thesis in this field and for his support along the way.

Title: Predicting Protein Thermostability With a Focus on Antibodies

Author: Bc. Adam Král

Department: Department of Cell Biology

Supervisor: Danny Asher Bitton, PhD

Abstract: Thermostability is key to successful protein engineering and therapeutic antibody development; accurate prediction accelerates the identification of stable proteins and the design of robust variants. This thesis evaluates zero-shot and supervised machine learning approaches for thermostability prediction, with a focus on antibodies. We achieve state-of-the-art performance on the public AbProp dataset—483 antibodies with melting temperatures measured by differential scanning fluorimetry (DSF)—reaching a Spearman correlation of 0.49 in the zero-shot setting and 0.69 with supervised learning. We find that antibody-specific language models do not outperform general models in the zero-shot setting, which may be due to a wider distribution of thermostability in antibodies, resulting from possibly weaker evolutionary pressure for stability and the high variability introduced by their combinatorial V(D)J origin and somatic hypermutation. We further evaluate zero-shot generalization on the ProteinGym benchmark, showing that our models perform competitively on unrelated protein domains. We also analyze sequence positions contributing to supervised prediction and observe that antibodies with lambda light chains are, on average, more thermostable than those with kappa chains. Our results underscore the limitations of pretraining solely on antibody sequence data for zero-shot prediction. At the same time, they demonstrate that pretrained protein language models, which have previously been shown to perform well on general protein stability tasks, can also be successfully applied to antibody thermostability prediction in the zero-shot setting.

Keywords: antibody thermostability, protein language models, zero-shot prediction, ProteinGym, supervised learning

Contents

List of Abbreviations

T_m Melting Temperature, Denaturation Midpoint Temperature. 2, 5

BERT Bidirectional Encoder Representations from Transformers. 2, 30, 37, 43

CDR Complemetarity Determining Region. 2

CNN Convolutional Neural Network. 2

CV Cross Validation. 2, 59–61

DSF Differential Scanning Fluorimetry. 2

Fab Fragment Antigen-Binding. 2

Fv Fragment Variable. 2

GNN Graph Neural Network. 2

LC-MS/MS Liquid Chromatography with tandem Mass Spectrometry. 2, 21

mAb Monoclonal Antibody. 2, 35

MAE Mean Absolute Error. 2

MLM Masked Language Modeling. 2, 22

MLP Multilayer Perceptron. 2

MSE Mean Squared Error. 2, 36, 39

NaN Not a Number. 2, 38

NMA Normal Mode Analysis. 2, 36, 38

PLM Protein Language Model. 2, 7, 22, 23, 25, 26, 30, 31, 36, 37, 43, 48, 50

SCC Spearman Correlation Coefficient. 2, 34, 47, 51, 60, 70

TPP Thermal Proteome Profiling. 2, 21, 24

VH Variable Heavy Chain Domain. 2, 23

VL Variable Light Chain Domain. 2, 23

Introduction

Thermostability is a crucial property for proteins, especially those used in therapeutic and industrial applications. For therapeutic proteins like antibodies, stability is essential not only for maintaining efficacy but also for manufacturability, shelf life, and safety (Bailly et al., 2020). Proteins must withstand various environmental stresses during storage, shipping, and in-use handling, as well as withstand temperature fluctuations and freeze-thaw cycles.

Antibodies are remarkable proteins capable of targeting almost any molecule, including proteins involved in autoimmune diseases, viral infections, cancer, and other conditions. While the immune system typically avoids generating antibodies against the body's own proteins, such reactivity can occur in autoimmune disorders—or be deliberately exploited in therapeutic settings, where antibodies are engineered or selected against self-proteins. However, their stability remains a critical factor, influencing both the efficacy and safety of antibody-based therapeutics.

Typically, a stability profile of at least two years at 5 °C storage is required, and melting temperature (T_{m}) serves as a key indicator of long-term stability, as it is believed to reflect a protein's resilience under milder, but prolonged conditions (Bailly et al., 2020).

In addition, thermostability has been shown to correlate with expression levels, suggesting it plays a vital role in identifying viable drug candidates. Antibodies with higher thermostability tend to be more efficiently expressed in cell systems, which is expected to improve their suitability for large-scale production (Jain et al., 2017). Interestingly, biological factors also influence thermostability: for instance, antibodies with higher thermostability are often associated with increased expression on B cell receptors, although the process of somatic hypermutation, which enhances antibody specificity, tends to reduce thermostability, highlighting a trade-off between specificity and stability (Shehata et al., 2019).

Thermostability is equally important for enzymes used in industrial processes, where elevated temperatures are often employed to increase reaction rates, enhance reactant solubility, and decrease the risk of microbial contamination (Yu et al., 2017). However, higher temperatures also increase the risk of enzyme denaturation, which limits their functional lifespan (Daniel and Danson, 2013). For instance, mesophilic

enzymes like E. coli transketolase are limited by low stability at higher temperatures, which constrains their use in processes that would otherwise benefit from elevated conditions (Yu et al., 2017).

Applications like drug manufacturing benefit from multistep biocatalytic cascades, where multiple enzymes must operate stably at the same temperature within a single bioreactor. These cascades reduce waste, link reactions to overcome unfavorable equilibria, and prevent the buildup of unstable intermediates (Huffman et al., 2019). Thus, thermostability remains a key factor in enzyme engineering, enabling efficient and sustainable industrial applications where consistent performance across specific conditions is required.

Predictive models of thermostability play a valuable role in protein engineering and drug development. Such models can be applied in several ways:

1. Scoring for Discovery: In high-throughput discovery methods, such as binding assays for antibodies or enzyme library screenings, models that score sequences for thermostability allow researchers to prioritize candidates with promising stability profiles. This capability is especially useful in large datasets, where manual stability assessment is infeasible. Pearson or Spearman correlation, or MAE, is often reported when evaluating the predictions (Notin et al., 2023; Widatalla, Rafailov, et al., 2024).

2. Engineering Mutations: During antibody development or enzyme optimization, predictive models for thermostability help optimize lead sequences by identifying mutations that could enhance stability. This approach enables researchers to improve the stability of candidates that already show favorable binding or functional properties, making them more viable for development.

3. De Novo Protein Design: Another strategy is de novo protein design, where we aim to create proteins with a desired function, such as binding an antigen or catalyzing a reaction, while ensuring they are thermostable.

In this thesis, we focus on the first application: models that provide a thermostability score for a given protein sequence, as there are antibody-specific datasets we can evaluate on. We also discuss approaches for the second application—models that can identify stabilizing mutations within a sequence, although the evidence for these methods lies mainly in non-antibody proteins and there are currently no public antibody-specific datasets available for evaluation. These approaches can streamline protein engineering by highlighting promising candidates and identifying potential improvements in stability.

Recent efforts to predict protein thermostability increasingly rely on machine learning, using either supervised models trained on experimental data or zero-shot approaches based on pretrained protein language models (PLMs). These methods offer an

alternative to traditional physics-based tools such as Rosetta or FoldX, which, while grounded in physical modeling, are computationally demanding and require expert configuration. Recent benchmarks show that PLMs—including models like ESM-1v and ProteinMPNN—offer orders of magnitude faster inference than physics-based methods and often match or exceed their predictive accuracy in protein stability prediction. While supervised models perform best when sufficient labeled data is available, zero-shot PLMs offer a practical and scalable solution for early-stage screening across large sequence libraries. These approaches have also been extended to antibody thermostability, though limited public data makes this a more challenging setting. Nonetheless, both supervised and zero-shot PLM-based methods are emerging as powerful tools for guiding antibody engineering.

In this thesis, we focus on predicting the thermostability of antibodies, based on its variable region sequence. We explore both zero-shot and supervised methods and evaluate their performance on a public dataset of antibody sequences with experimentally measured thermostability.

In the zero-shot approach we develop per-antibody scores (models) using logits from a diverse set of pre-trained protein language models (PLMs) and compare them. We also evaluate these scores on the ProteinGym stability dataset which contains general proteins (however mostly small domains) to get a comparison of the performance among other state-of-the-art models, albeit on a different dataset, as there are only few studies exploring zero-shot methods on antibodies.

In the supervised setting, we train models on the same dataset and compare the performance of the two approaches and to related work taking the supervised approach.

We find that the zero-shot score derived from ESM Cambrian 300M PLM (ESM Team, 2024) and the ESM inverse folding model (ESM-IF) (Hsu et al., 2022) outperforms previously reported zero-shot results on antibody thermostability. Specifically, our score achieves a Spearman's rank correlation coefficient of 0.49, outperforming the best zero-shot result reported by Widatalla, Rafailov, et al. (2024), which reached ~0.35. To our knowledge, that study represents the only prior evaluation of zero-shot methods on public antibody thermostability data.

Additionally, we develop a supervised method that improves upon the previously best-performing model on the AbProp dataset, reported by Widatalla, Rollins, et al. (2023), achieving a Spearman correlation of 0.69 compared to their 0.62.

Background on antibodies, protein thermostability, PLMs, and related work are presented in Chapter 1. In Chapter 2 we describe the datasets and methods used in this thesis. We also provide a detailed description of the models and their training, as well as the evaluation process used to assess their performance. In Chapter 3 we present the results of our experiments, of both zero-shot and supervised approaches on the antibody thermostability dataset. For the zero-shot models we include results on a large ProteinGym stability benchmark of ~165,000 mutants, with general proteins. Finally we

compare the best zero-shot and supervised method. In Section 3.2.3 we summarize our findings and outline how both the models we developed and those we evaluated can be applied in antibody engineering.

Chapter 1

Biological and Computational Background

1.1 Antibodies

Immunoglobulins are a class of protein molecules produced by B cells. A secreted form of immunoglobulin is the *antibody* produced by terminally differentiated B cells (Murphy and Weaver, 2017, p.139). They serve to defend the host from pathogens, such as viruses and bacteria. They non-covalently bind to pathogen, potentially covering it whole and isolating it from the host; *neutralizing* the pathogen or toxin. The antibodies bound to a pathogen can also recruit immune cells to mount an immune response against it. In particular phagocytic immune cells can engulf and destroy e.g. the bacterium. This is called *opsonization* (making it 'tasty', from Greek, to the immune cells) (Murphy and Weaver, 2017, p.399).

The adaptive immune response, involving i.a. immunoglobulins, was acquired by jawed vertebrates (or *gnathostomata*) more than 450 million years ago and is found in all extant jawed vertebrate species from fishes to humans (Lefranc, 2014; Vadnais et al., 2017). The variability and the specificity of antibodies to diverse antigens is achieved through combinatorial V(D)J rearrangement (or recombination) of gene segments; even in evolutionarily distant sharks. Nonetheless, substantial differences in antibody genetics, structure, and function exist across species (Vadnais et al., 2017). In this work, we focus on mammalian antibodies—specifically human and rodent—since they form the basis of most therapeutic antibodies. In the following text, we discuss human antibodies.

Antibodies are composed of two identical heavy chains and two identical light chains, which are linked by disulfide bonds. The antibody can be divided into the *variable region* (V) that binds the antigen, and the *constant region* (Figure 1.1). The constant (C) region of the heavy chain can have only a few variants and determines

the immunoglobulin isotype and associated effector functions. Of the five mammalian immunoglobulin isotypes—IgD, IgM, IgG, IgA, and IgE—IgE is involved in allergic responses and parasitic immunity, while IgA is typically studied in the context of mucosal immunity and is secreted into mucosal surfaces, although it is also present in serum. Immunoglobulin G (IgG) is the most abundant in serum and is the dominant isotype used in therapeutic antibodies (Kang and C.-H. Lee, 2021). Consequently, we limit our discussion to the IgG isotype and its subclasses. The IgG subclasses differ substantially in their function—for instance, IgG3 promotes a strong inflammatory response by efficiently engaging effector immune cells, while IgG4 can neutralize allergens without triggering immune cell activation (Vidarsson et al., 2014).

We begin with essential genetics to illustrate how antibodies arise, what components they are made of, and how these contribute to their diversity and function. Antibodies are coded on *three* distinct loci—immunoglobulin heavy chain is coded in IGH locus on chromosome 14, kappa (κ) light chain in IGK locus on chromosome 2, and lambda (λ) light chain in IGL locus on chromosome 22 (Murphy and Weaver, 2017, p.177). Each B cell expresses only one of the two light chain types. The IGH locus consists of the Variable, Diversity and Joining (VDJ) genes, or sometimes *gene segments*, and multiple constant genes, coding the different isotypes (Figure 1.1). The IGK and IGL loci each contain V and J segments (no D), which is why the term V(D)J is used to encompass both heavy and light chain recombination. There are about thirty to forty V gene segments and 4–6 J segments in each locus, and 23 D segments in the IGH locus (Table 1.1). Ultimately, only one gene segment from each category—V, D, and J for the heavy chain, or V and J for the light chain—is selected to form a functional antibody. This process, which occurs during B cell development, is known as V(D)J recombination.

Segment	**κ Light Chain**	**λ Light Chain**	**Heavy Chain**
Variable (V)	34–38	29–33	38–46
Diversity (D)	0	0	23
Joining (J)	5	4–5	6
Constant (C)	1	4–5	9

Table 1.1 Number of functional gene segments in human immunoglobulin loci. When a B cell develops, one gene from each segment type (V, D, J) for the heavy chain or (V, J) for the light chain is selected during V(D)J recombination to produce a functional immunoglobulin chain. This combinatorial process is a key source of antibody diversity. A B cell expresses either κ or λ light chain. Adapted from Murphy and Weaver (2017).

In total, antibody diversity arises from four main mechanisms. First, combinatorial diversity is generated by the random recombination of V, D, and J gene segments, this is encoded in the genome. Second, junctional diversity is introduced at the joining sites through random addition and deletion of nucleotides during the V(D)J recombination.

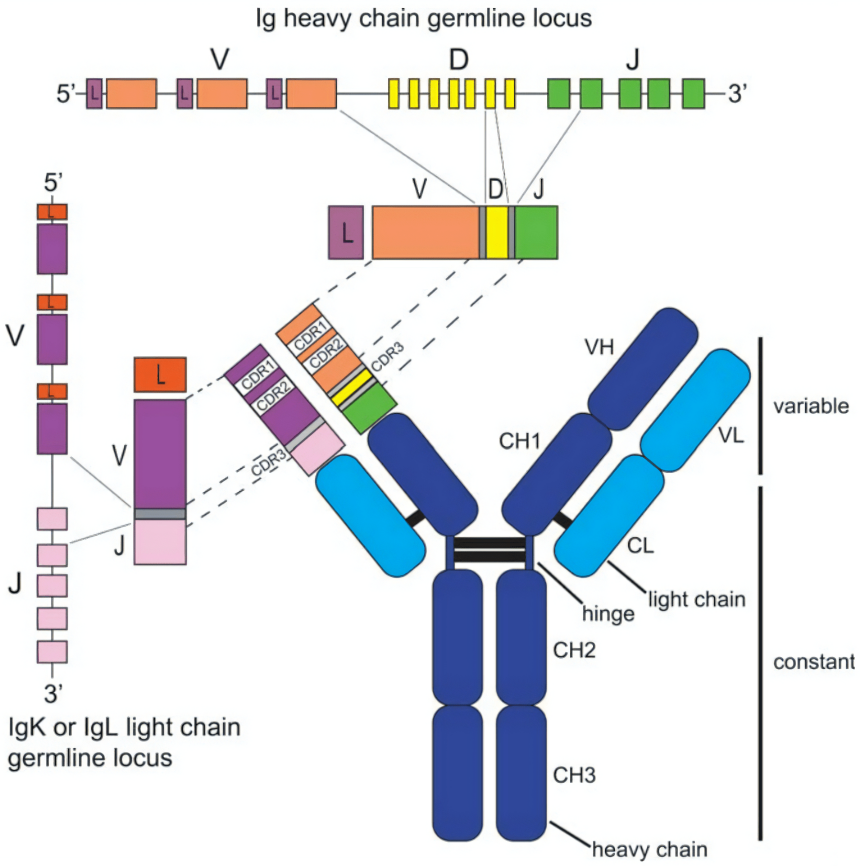

Figure 1.1 Schematic of antibody structure and genetic encoding. The lower part of the figure shows a full IgG antibody, consisting of two heavy and two light chains, each with variable and constant domains. The variable domains (VH and VL) are responsible for antigen recognition.

The top and left panels depict the germline (unrearranged) configuration of the immunoglobulin (Ig) loci. The heavy chain locus, shown at the top, contains tandem arrays of V, D, and J gene segments, while the kappa or lambda light chain locus on the left contains unrearranged V and J segments. Stepwise rearrangement of the germline DNA results in the joining of a heavy chain D and J gene segment, followed by joining of a V segment to the D-J product, to generate the DNA encoding the heavy chain variable region. In the process of rearrangement, the ends of the gene segments are subject to variable amounts of exonuclease digestion, and randomized nontemplated bases are added at the segment ends, to produce additional sequence diversity at the VDJ junctional region that encodes the complementarity-determining region 3 (CDR3) loop, which is often the region of the antibody heavy chain that has the greatest impact on antigen specificity. A similar process of V and J gene rearrangement with diversification of the VJ junction occurs in the light chain locus, to produce the rearranged light chain gene. The constant regions of the heavy and light chains (domains CH1, CH2, and CH3 for the heavy chain, and CL for the light chain) are encoded by downstream exons that are joined to the rearranged V(D)J gene by mRNA splicing. Disulfide bridges joining protein chains in the full antibody structure are shown with black line segments. Figure and caption adapted from Boyd and Joshi (2014).

Third, further combinatorial diversity results from the pairing of different heavy and light chains to form the antigen-binding site. Finally, somatic hypermutation, occurring after antigen exposure, introduces point mutations into the V-region genes, enabling affinity maturation, which makes the antibody a stronger binder of its antigen (Murphy and Weaver, 2017, p.184). The potential antigen receptor repertoire of each individual is estimated to comprise about 2×10^{12} different immunoglobulins (Lefranc, 2014).

A B cell rearranges first its heavy chain genes. After successful expression of the heavy chain, it divides into clones with the same heavy chain, and each clone independently and in parallel attempts light chain rearrangement. Not all heavy chains can pair with all light chains; certain VH–VL combinations fail to form a stable immunoglobulin molecule. In such cases, B cells may attempt further light-chain rearrangements, e.g. using the lambda locus after kappa rearrangements are unsuccessful (Murphy and Weaver, 2017; Luning Prak et al., 2011). [1] Nevertheless, most heavy and light chains are thought to be compatible and capable of forming functional antibodies (Murphy and Weaver, 2017, p.184–185). Typically, B cells express only one variant of the immunoglobulin. While rearrangement can proceed on both alleles (inherited from each parent), a mechanism known as allelic exclusion ensures that a B cell generally expresses only one version of each immunoglobulin chain (Murphy and Weaver, 2017; Luning Prak et al., 2011). [2] A developing B cell that fails to express a functional immunoglobulin ultimately undergoes apoptosis.

1.1.1 Antibodies as Therapeutic Molecules

Monoclonal antibodies (mAbs) are laboratory-produced molecules that recognize a single specific antigen, originally derived from a single B cell clone or generated using recombinant techniques such as phage display. Unlike polyclonal antibodies, which are a heterogeneous mixture produced by many B cells, mAbs are derived from a single genetic sequence, which enables high reproducibility and allows therapeutic antibodies to conform to the stringent specificity and safety standards of modern medicine. They can be applied to a wide range of diseases, including cancer, autoimmune, and infectious conditions.

Antibodies are used to target disease-driving molecules with precision—whether viral proteins or, interestingly, proteins normally present in the body—so-called endogenous antigens. While the immune system typically avoids targeting these to prevent

[1] Light and heavy chains can also undergo secondary rearragements, where the unexcised gene segments are used. Secondary rearrangements that alter the specificity of the immunoglobulin to avoid autoreactivity—in the bone marrow—are referred to as receptor editing (Luning Prak et al., 2011).

[2] This mechanism is essential for the proper functioning of the adaptive immune system. By enforcing allelic exclusion, the B cell exhibits a single antigen specificity (Murphy and Weaver, 2017, p.304), enabling accurate antigen recognition and effective coordination with T cells, which regulate B cell activation, selection, and isotype switching.

autoimmunity, this principle can be reversed in engineered antibody therapies—human-designed interventions that deliberately target such molecules. In diseases like cancer or autoimmunity, therapeutic antibodies can bind and neutralize signalling molecules or eliminate specific cell populations that express disease-associated targets. This deliberate targeting of otherwise tolerated molecules can lead to significant clinical benefits.

As of 2024, over 200 monoclonal antibodies (mAbs) have been approved and marketed as therapeutic agents (Crescioli et al., 2025). Rituximab, the world's first oncology antibody therapeutic, targets and depletes B cells expressing the CD20 antigen. It is a chimeric antibody with a murine variable region (Fv) and a human IgG1 Fc region, used in both B-cell lymphomas and autoimmune diseases (Pierpont et al., 2018). Tocilizumab targets the interleukin-6 receptor, preventing interleukin-6 binding and subsequent signaling, and is used in the treatment of rheumatoid arthritis (Mihara et al., 2011). Nivolumab illustrates the therapeutic use of immune-modulating antibodies that block inhibitory immune checkpoints without triggering immune effector functions. PD-1 (programed cell death 1) is an inhibitory receptor expressed on cytotoxic T cells that normally induces T cell anergy to prevent autoimmunity. Tumors exploit this mechanism by expressing PD-1 ligands, effectively inactivating T cells that would otherwise attack them. Nivolumab (anti-PD-1) blocks PD-1 and thus restores T cell activity against tumors. Its IgG4 backbone minimizes antibody-dependent cellular cytotoxicity (ADCC), reducing the risk of unintentionally depleting the reactivated T cells when the therapeutic goal is merely to block or neutralize the PD-1 receptor (Sundar et al., 2015).

Antibodies can be generated by immunizing animals and isolating antibody-producing B cells. In hybridoma technology, spleen B cells from an immunized mouse are fused with an immortal myeloma cell line, producing stable clones that secrete monoclonal antibodies Köhler and Milstein, 1975. Modern methods such as in vitro display technologies bypass the need for animal immunization (Hoogenboom, 2005). Phage display involves presenting antibody fragments on the surface proteins of bacteriophage capsids, while the DNA encoding each fragment is packaged inside the same particle. In this context, a *library* refers to a large, diverse collection of phages, each displaying a different antibody variant. This physical linkage between genotype and phenotype enables rapid screening of massive libraries against a target antigen. Binding phages are isolated, and their DNA is sequenced to recover the corresponding antibody genes. Human peripheral blood B cells can also be screened and sequenced (Pedrioli and Oxenius, 2021).

Most therapeutic antibodies belong to the IgG class due to their favorable serum half-life. Approximately three-fourths of these are of the IgG1 subclass, which activates immune effector functions and is thus well-suited for cancer therapies requiring cell killing. IgG2 and IgG4 are also commonly used. IgG4, in particular, has minimal effector function and is preferred for blocking or modulating immune pathways without

triggering strong immune responses. However, engineered variants of IgG subclasses can be designed to fine-tune or remove effector functions, including antibody-dependent cell-mediated cytotoxicity (ADCC) (Tang et al., 2021).

Beyond biological activity, antibody drug candidates must meet developability criteria—properties that determine how easily they can be manufactured, formulated, and administered. These include low aggregation tendency, thermal and colloidal stability, and compatibility with high-concentration formulations. Poor developability can lead to failure during production, storage or adverse immunogenicity in patients (Bailly et al., 2020).

1.1.2 Antibody Structure

The 3D structure of a full-length native antibody is shown in Figure 1.2. Limited digestion with the protease papain cleaves the antibody into three parts: two identical Fab fragments (fragment antigen binding), which contain the antigen-binding sites, and one Fc fragment (fragment crystallizable), which does not bind antigen but mediates interaction with effector molecules and cells. The cleavage occurs on the amino-terminal side of the inter-heavy chain disulfide bonds. The Fc fragment corresponds to the CH2 and CH3 domains and varies between heavy-chain isotypes. Both the Fab and Fc fragments are stable on their own (Murphy and Weaver, 2017, p.144).

In addition to these natural fragments, engineered single-chain fragment variable (scFv) constructs also exist. These consist of linked VH and VL domains and retain antigen-binding capacity. scFvs have become an established format and are approved as therapeutic agents (Ahmad et al., 2012; Ghaderi et al., 2023).

In this work, we focus on the variable region of antibodies and its relationship to thermostability. The variable domains of both heavy (VH) and light (VL) chains contain three hypervariable loops, which together form the antigen-binding site located at the tip of each arm of the antibody. These loops are known as complementarity-determining regions, or CDRs, because the surface they form is complementary to that of the antigen they bind. There are three CDRs from each of the heavy and light chains—namely, CDR1, CDR2, and CDR3. When the VH and VL domains are paired, the six CDRs create a single hypervariable site that determines the antigen specificity of the antibody (Figure 1.3). In most cases, CDRs from both VH and VL domains contribute to the antigen-binding site; thus, it is the combination of the heavy and light chain that usually determines the final antigen specificity. CDR1 and CDR2 are encoded within the V gene segment, while CDR3 is partly encoded by the D segment and contributes most to antigen-binding diversity (Murphy and Weaver, 2017, p.147).

The variable domains, as well as all other domains in the heavy and light chains, adopt a conserved structural fold known as the immunoglobulin fold. This fold consists of two antiparallel β-sheets packed together to form a "β-sandwich" (Figure 1.4). Each domain is further stabilized by an intradomain disulfide bond typically connecting the

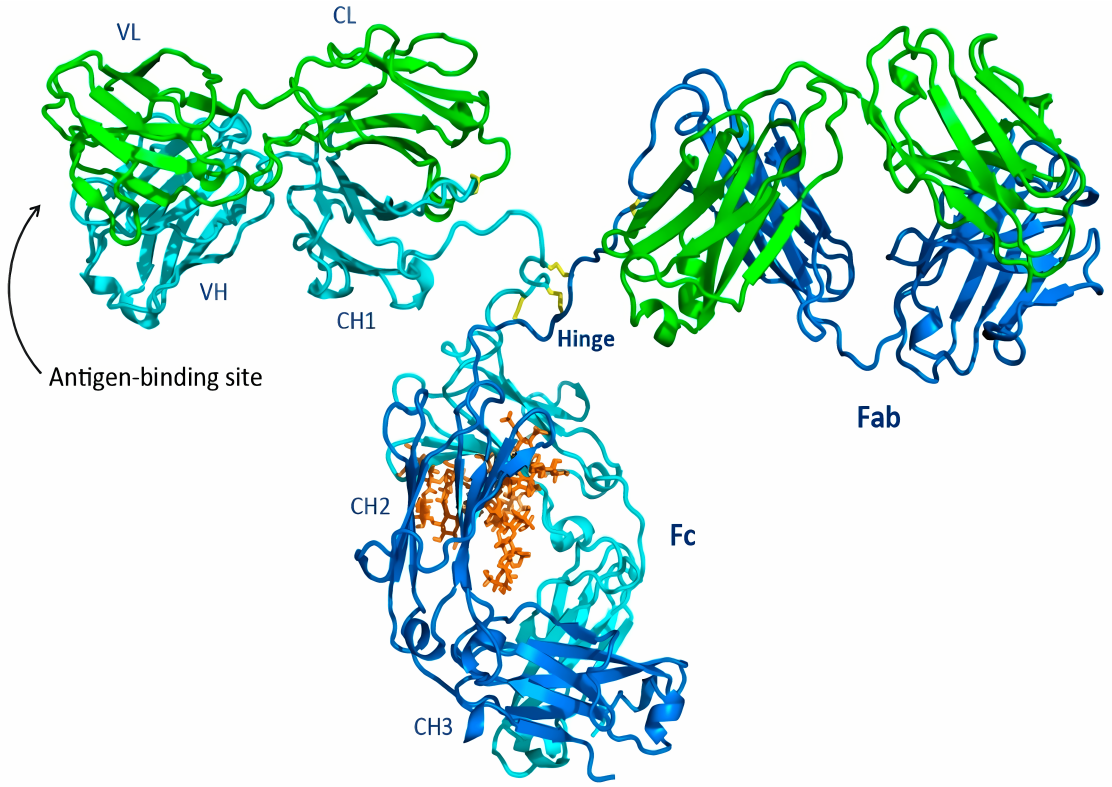

Figure 1.2 Cartoon representation of an intact IgG (PDB ID 1igt; Harris et al., 1997), which is a mouse IgG2a isotype. The light chains are shown in green, the heavy chains in cyan and blue, and the interchain disulfides in yellow sticks. The glycan, shown in orange sticks, is located in the Fc region and contributes to effector function and structural stability. The antigen-binding site is formed by the variable domains of the heavy (VH) and light (VL) chains. In this work, we focus on antibody thermostability as influenced by the VH and VL domains, with all other structural components kept constant. Adapted from Chiu et al. (2019).

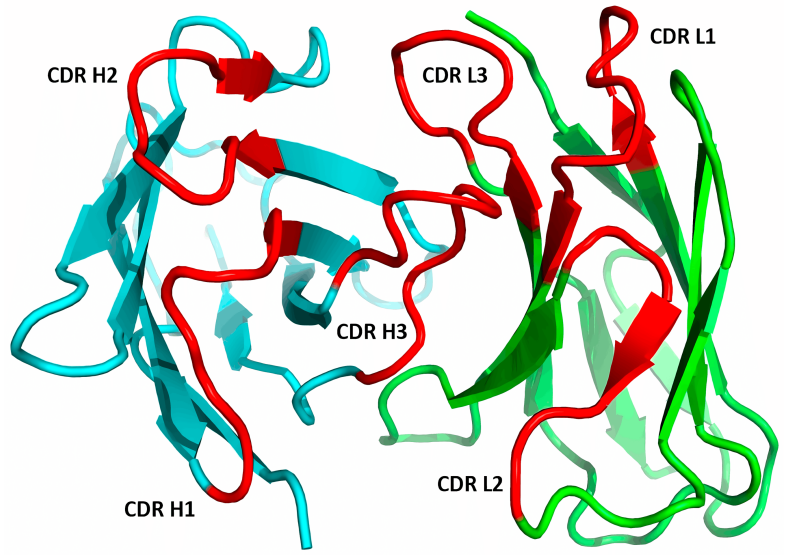

Figure 1.3 The antigen-binding site of an antibody, formed by the complementarity-determining regions (CDRs). Shown are the variable domains of the heavy (cyan) and light (green) chains with CDRs highlighted in red. Each chain contributes three CDRs: CDR1, CDR2, and CDR3. These six loops together create the antigen-binding site. (PDB ID 5i1a; Teplyakov et al., 2016; Chiu et al., 2019)

B and F β-strands. (Chiu et al., 2019).

The geometry of the binding site is modulated by the relative orientation of the VH and VL domains as shown in Figure 1.3. This interdomain arrangement varies between antibodies and has been proposed as an additional mechanism for expanding the antibody specificity repertoire. Mutations at framework positions in the VH–VL interface, outside the CDRs and distant from the binding site, have been shown to affect antigen affinity, suggesting that they act by altering the VH–VL orientation and, in turn, the geometry of the binding site (J. Dunbar et al., 2013).

In addition to their role in specificity, the VH–VL interface also influences antibody stability. T. Wang and Duan (2011) studied single-chain variable fragments (scFvs) under high-temperature molecular dynamics simulations and observed progressive disruption of the VH–VL interface at elevated temperatures. At 450 K (177 °C), native interdomain contacts declined gradually, while at 500 K (227 °C), near-complete dissociation occurred in most simulations. The contacts that were lost earliest involved weak hydrophobic or aromatic interactions, particularly near the CDRs. These findings indicate that disruption of the VH–VL interface can originate from the loss of such interactions, and that some of this dynamic behavior may relate to functional flexibility in the antigen-binding site. This is in line with the findings of Shehata et al. (2019), who showed that somatic hypermutation—which primarily introduces mutations in the CDRs to

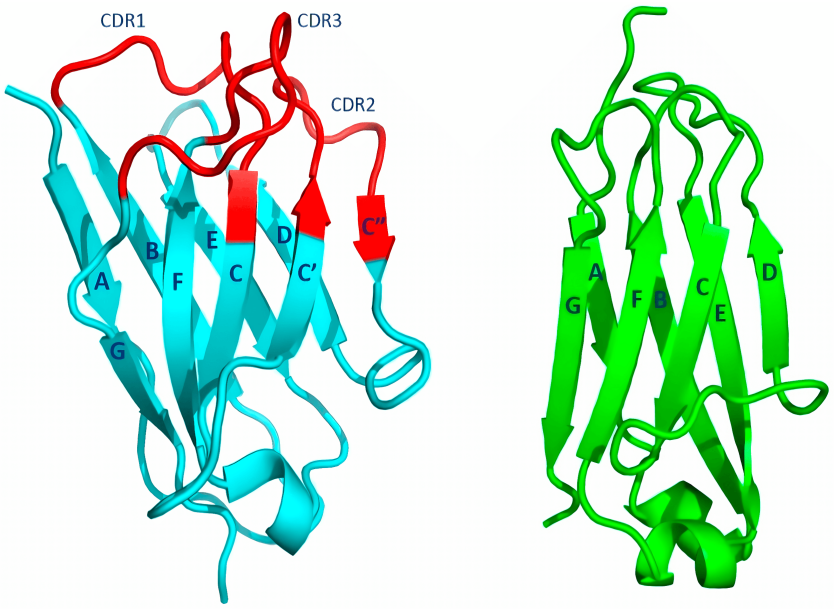

Figure 1.4 The immunoglobulin fold. The left cartoon image (cyan and red) shows the variable heavy-chain (VH) domain (PDB ID 5i1a; Teplyakov et al., 2016), with the complementarity-determining regions (CDRs) highlighted in red. The right cartoon image (green) shows the structurally similar constant domain (PDB ID 5i18; Teplyakov et al., 2016). Both adopt the immunoglobulin β-sandwich architecture composed of two antiparallel β-sheets. An intradomain disulfide bond forms between the B and F strands, stabilizing the domain structure (not shown; Chiu et al., 2019).

improve antigen binding—often compromises antibody stability, suggesting an inverse relationship between affinity maturation and thermostability.

1.1.3 Antibody Numbering and Alignment

Antibody variable domains are annotated using standardized numbering schemes that facilitate structural and functional comparisons. One of the most widely adopted is the IMGT numbering (Lefranc, Pommié, et al., 2003), which defines conserved framework region (FR) and CDRs across immunoglobulin and T cell receptor variable regions. This system ensures that conserved residues occupy consistent positions, such as cysteine at position 23, tryptophan at position 41, leucine at position 89, and cysteine at position 104, while allowing for variable length CDR loops.

The software package ANARCI (James Dunbar and Deane, 2016) enables automated numbering of variable domains by aligning them to pre-built species- and chain-specific profiles. It supports a range of numbering schemes and outputs consistent annotations across diverse sequences. Internally, ANARCI uses hidden Markov models (HMMs) trained on curated data from the IMGT/Gene Database. The numbering of each residue is then determined based on the rules defined by the selected scheme.

In our supervised models, we utilize the Honegger's numbering scheme (AHo) (Honegger and Plückthun, 2001), which numbers immunoglobulin chains—including heavy, light, and T cell receptor variable domains—based on structurally conserved positions identified from a set of reference structures. These structures were used to define the numbering scheme, common for all types of the mentioned immunoglobulins, but actual sequences are numbered using only the scheme's rules, without requiring structural information.

An illustration of the AHo numbering scheme, highlighting conserved structural elements, is presented in Figure 1.5.

By aligning sequences according to structurally or functionally conserved positions, these schemes allow the construction of meaningful multiple sequence alignments (MSAs) tailored to antibodies. Such MSAs serve a similar role as in general protein bioinformatics, providing a foundation for comparative analyses, identification of recurring sequence motifs—such as in CDR loops—and feature extraction in downstream modeling tasks.

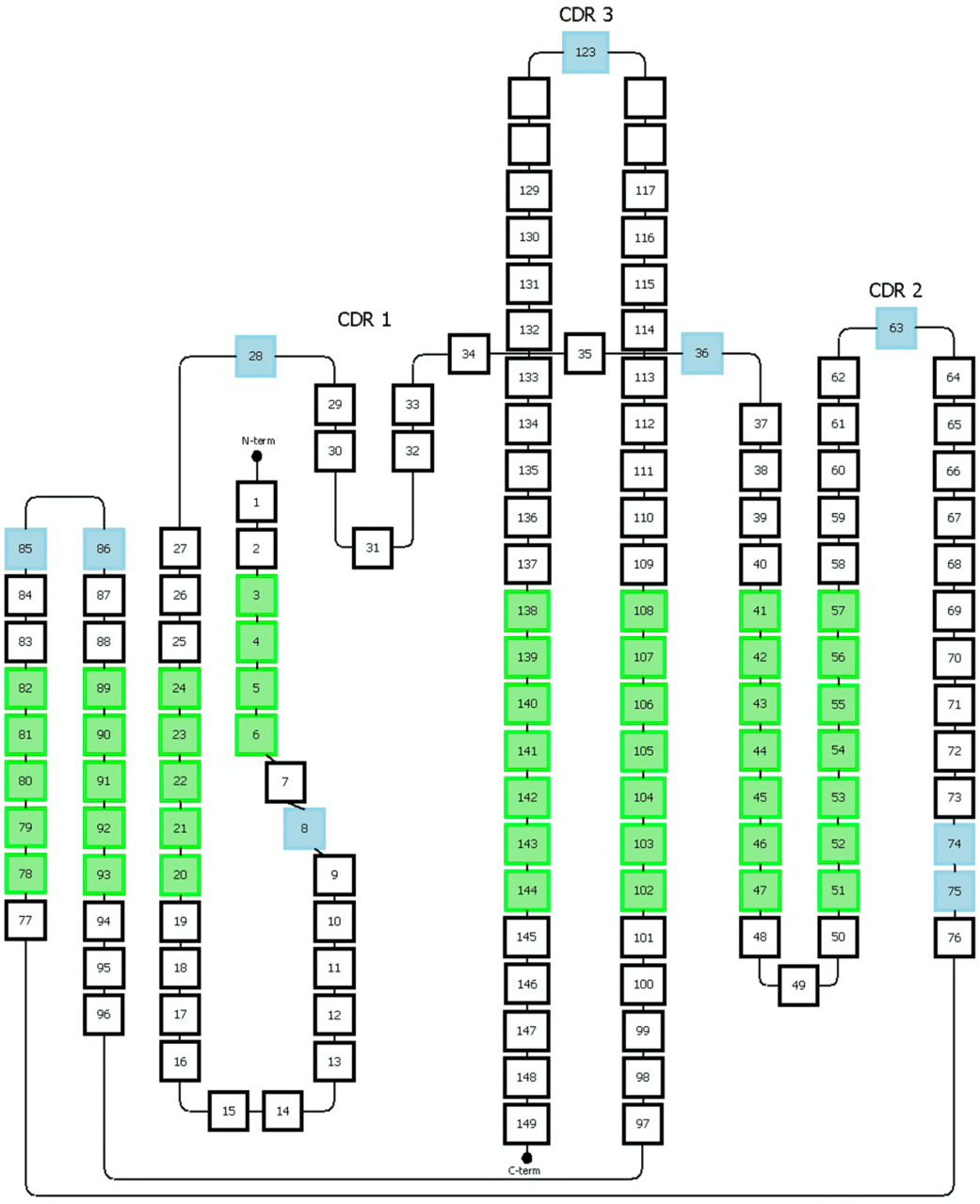

Figure 1.5 Representation of the Honegger's (AHo) numbering scheme for immunoglobulin variable domains. Green squares indicate structurally conserved positions. Blue squares represent positions that accommodate alignment gaps. The empty white squares within the CDR3 region reflect variable-length insertions and accommodate the diversity in CDR3 loop lengths. This schema enables structural superposition and comparison of antibody sequences. Figure from Dondelinger et al. (2018).

1.2 Thermostability Assays

Natural proteins are composed of chained amino acid building blocks and generally adopt a specific fold, 3D structure, the *native* fold or *conformation* that exhibits biological activity. The fold is stabilized by non-covalent interactions between amino acid residues or by covalent disulfide bridges between cysteines and in most cases spontaneously arises after the ribosome synthesizes the chain residue by residue. Protein *stability* is then the resilience of this folded state to unfolding.

Unfolding can occur partially and spontaneously even at equilibrium, since individual molecules fluctuate due to thermal motion and sample a distribution of energetic states. Externally, unfolding may be triggered by high temperature, chemical denaturants, or concentration effects. Therefore, the notion of stability is inherently context-dependent, as it reflects the resilience of the native fold under specific experimental conditions such as buffer composition, temperature, or denaturants. Each thermostability assay, in turn, defines and probes protein stability in its own way, depending on the principle and conditions it employs.

Commonly, protein thermostability is assessed by differential scanning fluorimetry (DSF) (Lavinder et al., 2009; Lang and Cole, 2017). In DSF, a hydrophobic-binding dye (e.g. SYPRO Orange) is added to the protein sample; as the temperature increases, unfolding exposes hydrophobic residues that bind the dye, causing fluorescence to rise. For proteins that undergo an approximately two-state transition, plotting fluorescence versus temperature yields a sigmoidal curve, and the inflection point (maximal slope) defines the melting temperature T_m, where half of the protein population is unfolded. In antibody thermostability studies (Jain et al., 2017; Shehata et al., 2019), the Fab T_m was determined using the first derivative of the raw fluorescence curve, identifying the peak of the derivative trace that corresponds to the inflection point of the unfolding transition.

1.2.1 High-Throughput Folding Stability via cDNA Display

The Mega-Scale dataset (Tsuboyama et al., 2023) comprises folding stability data (ΔG) for approximately 776,000 variants, including all single amino acid substitutions and selected double mutants across 331 natural protein domains and 148 de novo-designed domains (40–72 residues each). The ProteinGym (Notin et al., 2023) benchmark incorporates 64 domains from this Megascale dataset, which we will use alongside antibody datasets in our downstream evaluations.

In the cDNA display proteolysis assay, each mRNA molecule is covalently linked to its translated protein via a puromycin linker at the C-terminus, forming a stable protein–cDNA complex after reverse transcription. These complexes are incubated at room temperature with varying concentrations of a protease. Intact, folded proteins are resistant to cleavage and retain their N-terminal affinity tag, allowing selective

capture and sequencing of the associated cDNA. In contrast, unfolded proteins expose protease-sensitive sites, leading to cleavage and loss of the N-terminal tag—consequently, the associated cDNA is no longer captured. A simplified kinetic model is then fit to sequencing counts across the protease titration series Relative populations of unfolded ($[U]$) and folded ($[F]$) states are inferred, and stability is calculated by

$$\Delta G = -RT \ln \left(\frac{[U]}{[F]} \right)$$

This assay achieves remarkable scale, generating stability data for over 776,000 variants, both single and double mutants. However, it has both practical and methodological limitations: it is restricted to short, single-domain proteins due to DNA library synthesis constraints (domains under 72 residues), and the selection of double mutants was limited to structurally proximal residue pairs which precludes broader analyses of long-range mutational effects.

Although both cDNA display proteolysis and thermostability assays measure folding stability, the proteolysis approach infers ΔG at a fixed temperature under protease challenge, thus the spontaneous unfolding, whereas thermostability methods like DSF estimate the melting temperature (T_{m}), defined as the temperature at which the folded and unfolded states are equally populated.

Another commonly used method to assess thermostability is circular dichroism (CD) spectroscopy (Greenfield, 2006; Watson et al., 2023), which measures the differential absorption of circularly polarized light. Folded and unfolded proteins exhibit different absorption profiles due to changes in secondary structure, such as α-helices and β-sheets. CD is typically applied to purified, single protein samples and is not suitable for high-throughput screening.

Thermal Proteome Profiling (TPP) (Savitski et al., 2014; Jarzab et al., 2020) enables high-throughput measurement of thermostability across thousands of proteins in parallel. In this method, cells or lysates are exposed to a temperature gradient, and denatured proteins aggregate and precipitate, leaving only the soluble fraction. Protein abundances in each soluble fraction are quantified using Liquid Chromatography with tandem Mass Spectrometry (LC-MS/MS), and a melting curve is fitted to the protein's abundance measured across the temperature series, allowing estimation of the melting temperature (T_{m}) for each protein. While TPP allows proteome-wide assessment, its resolution is limited by the classical shotgun proteomics pipeline, which typically infers protein presence from only two or three peptide proxies (Sinitcyn et al., 2023). This poses challenges when analyzing proteins with high sequence similarity—such as antibodies—where peptide-level ambiguity may prevent reliable T_{m} assignment to a specific variant or isoform.

1.3 Protein Language & Inverse Folding Models

In recent years, methods inspired by natural language processing (NLP) have been successfully adapted to protein sequences, resulting in so-called protein language models (PLMs). These models are typically trained in a self-supervised manner, meaning that they learn internal representations of protein sequences without needing labeled data. One of the most common training objectives is Masked Language Modeling (MLM), where random amino acids are masked in a sequence and the model is tasked with predicting the correct residue at each masked position, leveraging the rest of the sequence. This approach, inspired by models like BERT (Devlin et al., 2019), enables the model to learn contextual dependencies and biochemical constraints that are implicit in natural protein sequences.

Protein language models such as ESM-2 (Lin, Akin, Rao, Hie, Zhu, Lu, Fazel-Zarandi, et al., 2022) and ProtBERT (Elnaggar et al., 2021) are trained on millions of sequences and produce embeddings that can be used for various downstream tasks, including mutation effect prediction, classification, and structure-related inference. These learned embeddings often implicitly capture structural and evolutionary features that correlate with biophysical properties like stability. The embeddings may be leveraged also as a basis for protein folding models as done in ESMFold (Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023).

Complementing protein folding models, inverse folding models aim to learn the mapping from structure to sequence. These models, such as ProteinMPNN (Dauparas et al., 2022) and ESM-IF (Hsu et al., 2022), take 3D structural information (backbone coordinates) as input and generate sequences likely to fold into the given structure. They are also called structure-informed language models, as they output the likelihood distribution over amino acids conditioned on structural context. Inverse folding models are commonly used for protein design, but their understanding of structural constraints also makes them promising for stability prediction tasks, especially when structural data or high-quality predictions are available.

Protein language models are typically trained on large protein sequence databases. UniProtKB (~252 million sequences, release 2025_01; The UniProt Consortium, 2023) contains both manually curated Swiss-Prot entries and unreviewed TrEMBL entries that are automatically annotated. In contrast, metagenomic resources like MGnify (~2.4 billion sequences, release 2024_04) contain protein sequences assembled from environmental sequencing samples, often covering microbial diversity not found in standard reference databases (Richardson et al., 2023). Clustering of protein sequences by sequence identity is often used to split the data into training, validation, and testing sets in order to evaluate generalization to sequences that differ significantly from the training data. Some clusters are assigned to training, others to validation, and the rest to testing, ensuring that closely related sequences are not shared across splits. Training can be performed then on all sequences within the clusters (e.g. ESM-1b; Rives et al.,

2021) or only on cluster representatives (Progen2; Nijkamp et al., 2022).

Language models can also be trained on specific protein classes. PLMs trained exclusively on a mixture of antibody Variable Heavy Chain Domain (VH) and Variable Light Chain Domain (VL) sequences include AntiBERTy (Ruffolo, Gray, et al., 2021). Newer models such as IgBert (Kenlay et al., 2024) and AbLang2 (Tobias H. Olsen, Moal, et al., 2024) are trained on paired VH–VL sequences originating from a single B cell, and therefore correspond to full antibody variable regions.

Most commonly, PLMs are encoder-only transformers (Vaswani et al., 2017) trained using the *masked language modeling* (MLM) objective, as in BERT. This is the case for ESM-2, AbLang2, and ProtBERT. In contrast, Progen2 (Nijkamp et al., 2022) is a decoder-only model trained with *causal language modeling* (next-token prediction), following the approach used in modern NLP systems.

Inverse folding models consist of an encoder that processes the backbone coordinates (i.e. without side chain atoms which would reveal amino acid identities) and a decoder that predicts the sequence likely to adopt the fold. ProteinMPNN (Dauparas et al., 2022) uses a graph neural network (GNN), specifically a message-passing neural network (MPNN) architecture (Bronstein et al., 2021). In the encoder, each residue is represented as a node whose hidden representation is initialized from the *structure* (i.e. backbone coordinates) and is then iteratively updated through message passing with the 48 nearest neighbor nodes. During decoding, the *sequence* information is added to the nodes via embedding. This information may come from previous token predictions (for autoregressive sampling), or from a fixed input sequence. In both cases, the model outputs residue likelihoods conditioned on the structure. We use the latter option to compute the likelihoods of native residues in antibody sequences, allowing us to score how well a sequence fits a given fold.

ESM-IF (Hsu et al., 2022) uses Geometric Vector Perceptron encoder layers, a type of GNN (Jing et al., 2021), to extract geometric features, followed by a generic autoregressive encoder-decoder transformer (Vaswani et al., 2017). As with ProteinMPNN, ESM-IF can both sample a new sequence autoregressively and score a supplied sequence as if it were its own prediction, analogously to teacher forcing used in transformer training.

AbMPNN (Dreyer et al., 2023) and AntiFold (Høie et al., 2024) are the variants of ProteinMPNN and ESM-IF fine-tuned on experimental and predicted antibody Fv structures.

The discrete token alphabet for protein language models can be extended to include tokenized structure, known as structure tokens. These are computed per residue and describe the local 3D conformation around that position. The space of conformations is discretized into a finite vocabulary. SaProt (Su, Han, et al., 2023), an ESM-2-style model, extends the vocabulary to the cross-product of standard amino acid tokens and Foldseek structure tokens (van Kempen et al., 2024).

ESM3 (Hayes et al., 2024) is another masked language modeling-trained model,

which includes its own discrete autoencoder to tokenize structural input. In addition to sequence and structure tokens, it also incorporates function tokens that describe biological activity, such as binding, enzymatic function, and domain or fold classifications derived from annotations in biological databases. Unlike the fixed 15% masking schedule used in BERT, ESM3 employs a variable noise schedule that enables iterative decoding: the model begins with a fully masked sequence and progressively predicts and unmasks tokens in successive decoding steps, combining aspects of masked and causal language modeling. With both sequence and structure modalities as inputs and outputs, and a sophisticated masked language modeling training scheme, ESM3 can function as both a folding and an inverse folding model.

1.4 Predictive Approaches to Thermostability

Accurately predicting protein thermostability from sequence is useful for a wide range of protein engineering applications. For instance, a researcher looking to identify natural enzyme sequences suitable for catalyzing a specific reaction in drug manufacturing may require the enzyme to remain operate at a specific temperature range, maintain functional stability over time in a bioreactor environment, and exhibit high catalytic activity. In large-scale sequence databases such as MGnify, which contains billions of entries, thermostability annotations are typically not available. A related proxy is the optimal growth temperature (OGT) of the organism from which the protein originates, as enzymes are often adapted to function near the host's physiological temperature.

Engqvist (2018) mined microbial OGT data from culture collection centers and matched it with enzyme optima reported in the BRENDA database (Chang et al., 2021). Comparing over 21,000 microbes and their enzyme optima, he observed a Pearson correlation of 0.75 between organismal growth temperature and the average optimal temperature of its enzymes.[3] This motivated subsequent work using machine learning to predict either the enzyme temperature optimum from sequence (G. Li et al., 2019), the organism's OGT from its proteome or sequence features (G. Li et al., 2019; Pudžiuvelytė et al., 2023), or to classify whether a protein originates from a thermophilic organism (Zhao et al., 2023).

Tthe Meltome Atlas dataset (Jarzab et al., 2020) probes protein thermostability via thermal proteome profiling (TPP; Section 1.2) across 13 organisms spanning bacteria, archaea, and eukaryotes. Meltome includes melting temperature (T_m) estimates for ~48,000 proteins, measured proteome-wide. This dataset enables the development of sequence-based thermostability predictors applicable to diverse proteins, including enzymes. As in the case of OGT-based inference, such predictors could support annotation and selection of enzyme candidates mined from metagenomic data for downstream

[3] Organisms with more than five reported enzyme temperature optima.

application. *DeepSTABp* is one such predictor, which embeds sequences using the ProtT5-XL encoder (Elnaggar et al., 2021), applies mean pooling over residues, and uses a multilayer perceptron (MLP) to predict T_m. The FLIP (Fitness Landscape Inference for Proteins) benchmark (Dallago et al., 2021) formalizes evaluation of thermostability prediction by defining three train/test splits over Meltome data: `mixed`, `human`, and `human-cell`, and provides baseline model performances for each split.

PRIME (Jiang et al., 2024) builds on this by pretraining an ESM-2-style model jointly on masked language modeling and organismal OGT prediction, curating a dataset of 96 million bacterial sequences with OGT labels from Engqvist (2018). This model is then fine-tuned on the FLIP thermostability task and shows improved performance over both ESM-1b and their ESM-2 baseline. For the `mixed` split, Spearman correlations were 0.680, 0.490, and 0.724 for ESM-1b, ESM-2, and PRIME respectively; for the `human-cell` split, the values were 0.750, 0.627, and 0.825. Note that FLIP baseline ESM-1b was computed by extracting embeddings, applying mean pooling, and training a separate MLP.

SaProt (Su, Han, et al., 2023) reports a Spearman of 0.724 on the `human-cell` split when fine-tuning the entire model. In the SaProtHub paper (Su, Z. Li, et al., 2024), they employ Low-Rank Adaptation (LoRA): they freeze all pretrained weights and learn small low-rank update matrices that are added to almost every weight matrix, providing a parameter-efficient alternative to full-model fine-tuning (Hu et al., 2021). However, SaProtHub's different train–test split prevents direct comparison. Mollon et al. (2025) attempt a controlled comparison of PLMs on the FLIP dataset. They evaluated SaProt and ESM-1v embeddings using the same MLP architecture, identical training hyperparameters, dynamic learning rate based on validation loss, and early stopping. SaProt embeddings yielded a higher Spearman correlation (0.697) than ESM-1v (0.670) on the `human-cell` set. The same study also evaluated the `human` set (with similar trends), but not the `mixed` split.

Interestingly, in the original FLIP benchmark, ESM-1v embeddings already achieved a Spearman of 0.74 on the `human-cell` split. In this light, it becomes difficult for us to attribute performance improvements observed in PRIME or SaProt solely to the incorporation of structure (in the case of SaProt) or to OGT-aware pretraining (in PRIME).

Besides these *absolute* thermostability prediction efforts, which aim to predict the melting temperature of any given protein, another important line of research focuses on predicting the *relative* thermostability of sequence variants of a single protein of interest, relative to the wild-type sequence. In this setting, the goal is to predict the change in melting temperature ΔT_{m} or $\Delta\Delta G$, where $\Delta\Delta G$ denotes the mutation-induced change in folding free energy (ΔG)—typically by a single or a few point mutations compared to the wild-type protein. This task is highly relevant for protein engineering, as it allows researchers to identify stabilizing mutations or eliminate destabilizing ones

from their design libraries.

Recently, Meier et al. (2021) showed that protein language models can predict the effect of variants in mutational assays. Deep mutational scanning (DMS; Fowler and Fields, 2014; Wei and X. Li, 2023) creates libraries of thousands—or more—of protein variants and returns for each a quantitative score of the phenotype of interest (e.g. stability or enzymatic activity). On a collection of 41 deep mutational scans assessing a diverse set of proteins on a variety of tasks, they observe PLMs such as ESM-1b or ProtBERT-BFD achieve across assays average Spearman correlation between 0.4 and 0.5, performing on par with then state-of-the art for unsupervised mutation prediction models EVMutation (Hopf et al., 2017) and DeepSequence (Riesselman et al., 2018), which need to be fit for each protein family separately with MSAs, while the general purpose PLMs can be used out-of-the-box for any protein. As the authors point out, unsupervised models like EVMutation and DeepSequence rely on evolutionary landscapes inferred from protein family-specific MSAs. In contrast, protein language models generalize from large-scale sequence data, capturing broad patterns shaped by billions of years of evolution across diverse proteins.

BERT-style protein language models return, for a given protein sequence of length L, an array of shape (L, V), where V is the size of the amino acid vocabulary. Each row corresponds to a position in the sequence and contains a probability distribution over the V amino acids, modeling $p(x_i = a)$ — the *likelihood* that amino acid a appears at position i, conditioned on the rest of the sequence. Taking the logarithm of this value yields the *log-likelihood* for a specific amino acid at a given position.

Meier et al. (2021) propose scoring mutations using the difference in log-likelihood between the mutant and wild-type residues at each mutated position. The language model is given the full protein sequence, with the mutated positions masked:

$$\sum_{i \in M} \log p(x_i = x_i^{\text{mut}} \mid x_{\setminus M}) - \log p(x_i = x_i^{\text{wt}} \mid x_{\setminus M})$$

where:

- M is the set of mutated positions,
- x_i^{mut} is the mutated amino acid at position i,
- x_i^{wt} is the wild-type amino acid at position i,
- $x_{\setminus M}$ is the sequence with all positions in M masked,
- $p(x_i \mid x_{\setminus M})$ denotes the probability assigned by the language model to amino acid x_i, conditioned on the sequence context.

Notin et al. (2023) expand on this work by introducing ProteinGym, a comprehensive benchmark of 217 DMS assays, including 66 targeting thermostability. They evaluate

79 models in the zero-shot setting (as of version v1.2), using metrics such as Spearman correlation to assess performance. ProteinGym also includes comparisons supervised learning approaches under consistent evaluation protocols. In the supervised setting, a subset of each assay is used for training and the remaining data for testing. This allows to compare these two approaches—how well models learn from experimental data versus how much they can generalize in a zero-shot fashion.

For zero-shot *absolute* stability prediction, Cagiada et al. (2024) report that, across 265 proteins—most of them small domains from the Mega-Scale dataset (Tsuboyama et al., 2023; see Section 1.2)—the native sequence likelihood from an inverse folding model achieves a Spearman correlation of 0.63 with experimental stability data. Importantly, this evaluation was performed only on the wild-type sequences of these domains, rather than on the many mutant variants available in the Mega-Scale dataset. This distinguishes it as a zero-shot *absolute* folding stability prediction. In their setup, the sequence likelihood is computed using the inverse folding model ESM-IF (Hsu et al., 2022) as:

$$\sum_{i=1}^{L} p(x_i = x_i^{\text{native}} \mid \text{structure})$$

where:

- L is the sequence length,
- x_i^{native} is the native amino acid at position i,
- $p(x_i \mid \text{structure})$ is the inverse folding model's predicted probability of observing x_i at position i, given the structure predicted using AlphaFold 2 (Jumper et al., 2021).

Reeves and Kalyaanamoorthy (2024) compare the performance of PLMs against traditional methods, including Rosetta Cartesian DDG (Kellogg et al., 2011; Park et al., 2016) and the statistical potential-based method KORPM (Hernández et al., 2023), using the FireProtDB dataset, which contains experimentally measured ΔT_{m} and $\Delta\Delta G$ values from protein mutation experiments (Stourac et al., 2021). They evaluate models such as ESM-1v, ProteinMPNN, and ESM-IF in the zero-shot setting. When averaging performance by assay (i.e. per protein), Rosetta CartDDG achieves the highest Spearman correlation coefficient (SCC) of 0.421, with ProteinMPNN and ESM-IF close behind at 0.384 and 0.380, respectively. The authors conclude that Rosetta is marginally better for ranking all mutations within a protein (i.e. in weighted-average Spearman ranking).

However, when the evaluation task is adjusted to prioritize stabilizing mutations—measured via weighted Normalized Discounted Cumulative Gain (wNDCG)—or to distinguish stabilizing from destabilizing mutations—measured via weighted Area Under the Precision-Recall Curve (wAUPRC)—three out of four ProteinMPNN variants outperform Rosetta. Moreover, since Rosetta and KORPM had parameters fitted on a

large portion of FireProtDB, the authors refer to these as supervised methods. When mutations overlapping with their training data are removed, both methods suffer significant drops in performance, particularly on the ΔT_m subset (1006 unique mutations). In contrast, PLMs maintain their performance, indicating better generalization.

This pattern holds on a separate mutation set of 41 proteins, each with no more than 25% sequence identity to the supervised methods' training data. On this set, physics-based methods fall behind PLMs, and the advantage of inverse folding models narrows—sequence-only models perform slightly better overall.

Despite comparable predictive accuracy, ProteinMPNN offers vastly superior throughput. According to Dutton et al. (2024), ProteinMPNN achieves a million-fold speedup over physics-based Rosetta or FoldX (Schymkowitz et al., 2005) for mutation scanning tasks when run on similar hardware.

1.4.1 Antibody Thermostability

Compared to general proteins, thermostability prediction for antibodies is a less explored area, largely due to the scarcity of public datasets. To our knowledge, only a single small-scale dataset of antibody thermostability is publicly available. Nevertheless, both supervised and zero-shot approaches have been studied.

Harmalkar et al. (2023) investigate thermostability prediction of single-chain variable fragments (scFvs), which are antibody constructs where the VH and VL domains are linked by a flexible peptide. Using a proprietary dataset of approximately 2,700 scFv sequences derived from 17 projects targeting different antigens, they quantify thermostability using the TS50—the temperature at which 50% of binding is lost after heat stress. Their best-performing model consists of a multilayer perceptron (MLP) classifier, [4] built on top of frozen AntiBERTy embeddings (Ruffolo, Gray, et al., 2021), where residue embeddings are downprojected and concatenated. This approach achieves a Spearman correlation of 0.71 on held-out data. However, performance on an out-of-distribution[5] TS50 test set drops below 0.2, and on another test set of only 9 scFvs with nanoDSF-measured melting temperatures (T_m), the model achieves a high correlation of 0.9. They also evaluate ESM-1v and AntiBERTy in a zero-shot setting using pseudo-log-likelihood scoring (Section 2.2.3), and conclude that zero-shot predictions

[4]The classifier outputs probabilities across four temperature bins (e.g. 50 °C to 60 °C). To report a continuous TS50 value, they treat the classifier as a regressor by computing a weighted average of the mean temperature in each bin, weighted by the predicted probabilities. We report results for the resulting continuous TS50 values.

[5]The authors describe: "We collated 2,700 scFv sequences from 17 projects that target different antigens (further referred to as experimental sets) to constitute the sequence data. Additionally, sequences from another scFv study (currently under clinical trials) and an isolated scFv dataset form out-of-distribution, blind test sets. Out-of-distribution refers to the fact that the out-of-distribution sequences are blind to the algorithm." However, it is not clear to us why exactly the held-out data are not also considered blind to the algorithm. For further details, we refer the reader to the original paper.

do not correlate well with thermostability, whether evaluated on all TS50 or other out-of-distribution data.

Widatalla, Rollins, et al. (2023) curate a dataset of 483 antibodies with T_m values of measured by differential scanning fluorimetry (DSF). Using a graph attention network (GAT), they process embeddings from the AbLang language model (Tobias H Olsen et al., 2022), while allowing fine-tuning AbLang's last n layers, n being a hyperparameter. On a held-out set of 73 examples, their model reaches a Spearman correlation of 0.62. More recently, Widatalla, Rafailov, et al. (2024) fine-tune ESM-IF on the Mega-Scale stability dataset to produce ProteinDPO and evaluate it in a zero-shot setting on the full AbProp T_m dataset, achieving a Spearman correlation of ~0.35, improving over the ~0.25 they report for the vanilla ESM-IF model.

In this work, given the limited availability of antibody thermostability data, we examine both zero-shot models and simple supervised models such as linear regression and decision tree-based methods.

Chapter 2

Methods

2.1 Antibody Thermostability Dataset

We use the AbProp T-mid dataset (Widatalla, Rollins, et al., 2023) curated from (Jain et al., 2017; Shehata et al., 2019), to our knowledge the largest consistent public antibody thermostability dataset. It consists of a concatenation of measured T_{m} values from 137 clinical-stage antibody sequences and 346 human Fv antibody sequences, all cloned into the same IgG1 isotype and measured as Fab fragments. The resulting antibodies differ only in their variable domains, enabling a controlled comparison of Fv influence on thermostability.

We predict structures of the Fv regions using ABodyBuilder2 (Abanades et al., 2023) with default settings, which includes minimization of the predicted structure with Amber ff14SB force field (Maier et al., 2015). The output is a PDB file containing heavy (H) and light (L) chains with IMGT numbering. The structures are used as inputs to pre-trained structure-informed language models in both the zero-shot setting, and the supervised setting. The predicted structures are also used to compute features for the supervised models described later.

2.2 Zero-Shot Prediction

2.2.1 Evaluated Protein Language Models

We evaluate a diverse set of 14 protein language models, including 4 antibody-specific models, 5 structure-informed PLMs, 9 Bidirectional Encoder Representations from Transformers (BERT)-style, and 5 autoregressive models.

The antibody-specific models include IgBERT (Kenlay et al., 2024), AbLang2 (Tobias H. Olsen, Moal, et al., 2024), AbMPNN (Dreyer et al., 2023), and AntiFold (Høie et al., 2024). The general-purpose masked language models comprise ESM Cambrian

(ESMC) 300M and 600M (ESM Team, 2024), ESM3-open (Hayes et al., 2024), ESM2-150M and ESM2-650M (Lin, Akin, Rao, Hie, Zhu, Lu, Fazel-Zarandi, et al., 2022), ProtBert and ProtBert-BFD (Brandes et al., 2022). The autoregressive models include ProGen2-base (Nijkamp et al., 2022), ESM-IF (Hsu et al., 2022), and ProteinMPNN (Dauparas et al., 2022).

An overview of model architectures, input modalities, and pretraining datasets is provided in Figure 3.5.

2.2.2 Input Representation

The models are fed the variable domain sequences (heavy and light chains), or the structure. For general purpose PLMs we concatenate the light and heavy chain in that order with a 16-mer (GGGS)4 linker. Antibody-specific models natively allow input of the heavy and light chains. For structure-informed models, we provide the predicted structure generated by ABodyBuilder2 (Abanades et al., 2023).

ESM-IF was not trained on multi-chain structures, however they provide a function `score_sequence_in_complex` which concatenates residue coordinates of all chains, using the default padding (10 positions), representing missing coordinates in the structure, and scores a passed single sequence (shortcutting the autoregressive prediction with already known previous targets, analogous to teacher-forcing). It returns residue likelihood scores of the native sequence of the one chain. It has no information about the sequence of the other chain, except its backbone coordinates. We score both heavy and light chain with this method. Additionally we implement a similar function that scores a sequence also in the presence of the sequence of the other chain, this time joining chains with a (GGGS)4 linker, and score light and heavy chains after the model has been 'forced' to predict the sequence of the other chain, we term this setting '`ESM-IF score_second_sequence`'. Therefore in this setting, chain L precedes chain H, when scoring chain H, and vice versa.

ESM3 also supports structure tokens on the input, which are automatically generated from a PDB file. We explore this in `with_structure` and `only_structure` settings. ESM3 was also trained with a chain break token, so besides the ordinary linker variant, we join light and heavy chain sequence with the chain break in `chain_break` setting.

2.2.3 Score Computation

We compute the per-residue likelihoods for each sequence in the dataset using the PLMs. The models return logits in the shape of `(L,V)`, where `L` is the sequence length and `V` is the vocabulary size (including 20 amino acids, besides special tokens). We drop positions corresponding to special tokens (such as the beginning of sequence token),

and keep logits of the sequence, perhaps with the linker.[1] For the majority of models (except four that will be listed below), we process the logits on each position in the following ways, and the overall score is computed as mean over all positions:

- `ll_score` — log-likelihood of the native residue
- `logit_score` — logit (before softmax) of the native residue
- `ll_score_linker` — same as `ll_score` but we do not drop linker positions
- `logit_score_linker` — same as `logit_score` but we do not drop linker positions
- `joint_ll_score` — sum of log-likelihoods of all amino acids
- `mean_logit_score` — mean of the logits for all amino acids
- `entropy_score` — entropy of amino acid probability distribution
- `top5_entropy_score` — entropy-like sum of the top 5 amino acids likelihoods: $-\sum_{i=1}^{5} p_i \log p_i$ where p_i denotes the i-th largest value in the softmax-normalized amino acid probability distribution at a given position.

The first two scores are computed discarding the linker positions, and others with the linker positions included, if applicable. For ESM3 we always drop the chain break token. For ProGen2, which can be run autoregressively in two directions (left-to-right and right-to-left), we compute two variants of the scores. In `lrrl_scoremean` we compute the scores above using the logits from left-to-right and right-to-left runs and average the scores. While in `lrrl_logitsmean` we first average the logits from both runs and then compute the scores.

For ProteinMPNN (or AbMPNN), ESM-IF (or AntiFold), the inverse-folding models, score computation was done a bit differently (we did it before trying out the other models). For ProteinMPNN we use options `--conditional_probs_only 1`, `--seed 37`, and `--num_seq_per_target 10`; otherwise, defaults are used. The first option specifies we condition on the previous native sequence targets when predicting next position (as opposed to sampling a new sequence). The second option sets the random seed for decoding order. The third option specifies the number of sampled decoding orders. ProteinMPNN follows the numbering in the PDB structure. As our PDB structures are IMGT numbered, there may be gaps in the numbering. ProteinMPNN's preprocessing inserts new residue positions at these numbering gaps, with unknown

[1] We keep the vocabulary dimension as is, i.e. we do not filter just for amino acids tokens before the softmax. However, for some scores—such as `joint_ll_score` (description follows)—it could be beneficial.

coordinates, and a sequence token `X`, representing missing residue coordinates in the structure, we term this the `default` setting. As these gaps don't have biological or practical meaning for a predicted structure with all coordinates, we also experiment with stripping these gaps in the preprocessed files, and term this setting `nogap`. We disregard the logits for newly inserted positions in the `default` setting. We gather likelihoods and log-likelihoods of the native sequence and mean them over the decoding replicates, resulting in a vectors *ml*, and *mll* of shape `(L,)`. Finally we compute the scores by averaging or summing the likelihoods or log-likelihoods over all positions, CDR positions, and framework positions (as defined by the IMGT numbering). In total, due to the combinations of the two settings (default and nogap) and the two types of scores (likelihoods and log-likelihoods), summing or averaging over positions and CDR or framework positions, we have 24 different scores. Similarly, for `ESM-IF` `score_second_sequence` setting we have 6 different scores for each chain, as there is only single decoding order. For the default ESM-IF setting we simply compute 2 scores—the average and sum of log-likelihoods over all positions. Finally, we simply sum these heavy and light chain scores to obtain per-antibody score for downstream analysis.

We run ESMC 300M also in pseudo-log-likelihood (PLL) setting (Salazar et al., 2020), requiring L forward passes, where L is the length of the sequence, masking in each pass a single residue and recording the returned probability distribution on that position; this is done over each position. Finally, the likelihood distributions obtained only from masked positions are concatenated to return a array of shape `(L,V)` where `V` is the vocabulary size.

2.2.4 Evaluation Strategy

We evaluate each model score (around 250 total) on the entire AbProp dataset. We employ 10,000 times bootstrap resampling to compute Spearman correlation coefficients between the scores and experimental T_m values, obtaining a distribution of SCCs for each score. Using the same set of dataset resamples for each score allows a statistical comparison of the performance between the scores, or models; paired bootstrap test (Koehn, 2004), which however is not strictly a test and does not return a p-value for a statistic distributed under the null hypothesis. For testing hypotheses whether a model A has the same performance as model B—or not—we use the permutation test.[2] In case our hypothesis involves multiple comparisons, we use the Holm-Bonferroni method (Holm, 1979) to adjust the p-values.

We implement the permutation test in the following way: under null hypothesis, where performance of model A and model B is the same, it does not matter whether a prediction for sample i is from model A or model B. We randomly assign the model

[2]Implemented according to `https://ufal.mff.cuni.cz/~straka/courses/npfl129/2223/slides/?13#96`

predictions to each dataset sample (we 'permute' the model predictions, permutation of size 2—two models; in fact we randomly choose one of the models for each sample), and compute the SCC between the predictions and experimental T_m values. We repeat this 100,000 times, obtaining an approximate distribution of SCCs under the null hypothesis. For a two-tailed test—not assuming which model is better—we then compute the p-value as the fraction of SCCs in the null distribution that are as extreme (to either direction) as the observed SCC between model B and experimental T_m (or by symmetry equivalent to choosing model A). We use the Numba just-in-time compiler (Lam et al., 2015) for Python to speed up the computation. The null distribution we compute is approximate, as we did not evaluate all possible configurations of model assignments to samples (exponential in dataset size), however with 100,000 repeats we observe little variation in the returned p-values and for simplicity treat them as the p-value of the null hypothesis. Before performing the permutation test we standardize the scores.

2.2.5 Score Clustering and Combination

To compare the best-performing models (top 5% scores in absolute median SCC with T_m) in terms of prediction or *score* similarity, we cluster the prediction vectors using average linkage clustering (UPGMA) as implemented in `scikit-learn 1.5.1` (Pedregosa et al., 2011). The distance matrix used for hierarchical clustering is computed such that each element at index (i, j) contains the Spearman Correlation Coefficient (SCC) between score vectors i and j, where each vector spans the length of the dataset (i.e. quantifying the similarity of *scores*, not the quality of T_m prediction). This clustering and the subsequent score combination (described below) are performed on the AbProp training portion (Section 2.3.1) to obtain an unbiased estimate of the combined scores' performance on the holdout portion. The bootstrapping used to select the top 5% of models (by absolute median SCC) is performed on the training set to ensure consistency with subsequent evaluation steps.

We then form flat clusters using `fcluster` with the `maxclust` criterion, specifying three and four clusters, and select the best score (by absolute median SCC with T_m) from each cluster as its representative. The selected scores are standardized (z-scored), and we compute a weighted combination using a weight vector of shape `(3,)` or `(4,)`, resulting in the combined scores ZS3 (3xESMC) score and ZS4 (ESMC+ESM-IF) score, respectively. The former includes three scores from ESMC, while the latter adds one score from ESM-IF.

For ZS3 score, the weights correspond to the median SCC of each score, ensuring that all components are positively correlated with the target T_m prior to summation. [3]

[3]For the ProteinGym evaluation, we inadvertently used the element-wise cubed weight vector instead of the original. Based on performance on the training set (SCC 0.465 for the original weights vs 0.462 for the cubed; 410 samples) and on the full set (SCC 0.481 for the original vs 0.479 for the cubed), the impact appears to be negligible, which is well within the uncertainty range as estimated from bootstrapped

For ZS4 score, the weight vector is analogous but intentionally cubed element-wise.

We experimented with the following element-wise transformations of the weight vector: sign, identity, and exponentiation by 3 and 5, choosing the best performing on the train dataset. The final definition of ZS3 score [3] is a dot product of the following three standardized ESMC 300M scores: `joint_ll_score`, `logit_score`, and `ll_score`, using the weight vector $[0.42, -0.37, 0.34]$. ZS4 score uses the same three ESMC scores and additionally includes the average log-likelihood from ESM-IF in the `score_second_sequence` setting, with the full weight vector $[0.42, -0.37, 0.34, 0.31]^3$.

We evaluate scores on both the full dataset and the holdout subset to obtain an unbiased performance estimate.

2.2.6 ProteinGym Evaluation

We evaluate the second best zero-shot score, ZS3 (3xESMC) score, and all other scores computed from ESMC 300M logits Section 2.2.3, on the ProteinGym stability dataset (Notin et al., 2023), comprising of large-scale mutation scanning results for 66 proteins. With double mutants in 51 of the 66. In total around 69,000 single and 65,000 double mutants. We predict logits with ESMC 300M for ~135,000 sequences, and compute the scores as described in Section 2.2.3, namely the ZS3 score. We are using the data and repository from v1.2 release, which contains predictions by 79 models, 30 sequence-only, 30 structure-informed, and 19 MSA-only models. We evaluate the performance of these existing models together with our newly computed ESMC scores, analyzing single and double mutants separately. This separation by mutation depth (single vs. double) and assay type (stability) at the same time was not included in the original aggregated ProteinGym results. For simplicity, we report only the Spearman correlation, although the benchmark includes additional evaluation metrics.

2.3 Supervised Learning

2.3.1 Dataset Split

Widatalla, Rollins, et al. (2023) split the AbProp T_m dataset (Section 2.1) randomly, as the monoclonal antibodies (mAbs) originate from random B cells or distinct clinical-stage antibodies, and report that 99.9% of sequence pairs exhibit less than 92.9% sequence identity. The dataset is provided as a CSV file containing MSA of the light and heavy chains, along with target and split annotations. A total of 410 sequences were used for training and 73 for evaluation (labeled as *holdout*). We keep the same dataset split as in

confidence intervals and beyond the precision we report.

the original work. [4]

We number the heavy and light chain sequences using the structure-based Honegger (AHo) numbering scheme (Honegger and Plückthun, 2001), implemented via the tools ANARCI (version 2020.04.23) (James Dunbar and Deane, 2016) and abnumber (version 0.3.2). The numbered sequences are then aligned into a multiple sequence alignment (MSA), where each column corresponds to a unique position across the union of all AHo-numbered positions observed in the dataset. The resulting MSA contains a total of 279 columns (positions) spanning both heavy and light chains.

Our preliminary analysis showed that the closest training-set sequence to each sequence in the holdout set is, on average, 83.0% identical in the MSA. Furthermore, all but one holdout sequence differ from their closest training sequence by at least 20 mutations, with a median difference of 48. These results support the claim that the training and holdout sets are sufficiently dissimilar.

For hyperparameter search we use two times repeated 5-fold cross-validation on the training set, and select the best hyperparameters based on the Mean Squared Error (MSE) between predicted and experimental T_{m} values. The splitting in folds is random although with the same random seed for all types of models. Final models are trained on the entire training set, and evaluated on the holdout set.

2.3.2 Input Features

For the supervised models, we use three types of features: (i) one-hot encoding of the amino acid sequence, (ii) amino acid logit scores from PLMs, and (iii) computed B-factors using Normal Mode Analysis (NMA) from the predicted structures.

PLM logits are computed using the ESM-IF structure-informed PLM (Hsu et al., 2022), AbMPNN structure-informed PLM (Dreyer et al., 2023), and the ESMC 300M sequence-only PLM (ESM Team, 2024). These logits are the output of the last layer of the language models, and are an array of shape `(L,V)` where `L` is the sequence length and `V` is the vocabulary size (including 20 amino acids, besides special tokens).

For the two structure-informed PLMs, which take as input only the structure—the backbone coordinates—we use the predicted structures generated by ABodyBuilder2 (Abanades et al., 2023) . Both models are autoregressive, so the logits computed for a position depend on the previous positions (predicted or supplied ground-truth). We supply the native, ground-truth, sequence for both models, and the logits are computed for the entire sequence.

[4]The CSV file contains split labels: train (326), test (84), and holdout (73). We verified in the original repository that the concatenation of the train and test splits was indeed used for training (with cross-validation), and the final model was evaluated on the 73-sequence holdout set (*AbPROP/Src/Cv_train_ablang_gnn.Py* 2024; *AbPROP/Src/Ensemble.Py* 2024). Following the same protocol, we refer to the combined train and test splits as the training set, and evaluate on the holdout set.

ProteinMPNN (AbMPNN) supports natively multiple chains in the input. While ESM-IF was not trained on protein complexes (or multi-chain structures), it allows joining two chains with a series of masked residue coordinates. We left this padding left as default to 10 'residues'. Logits for these padded residues are discarded in further processing. In hindsight this is shorter than the commonly used 15-mer (GGGGS)3 or 16-mer (GGGS)4 linker for scFvs, and linkers under 12 residues were found to not form a functional Fv domain (Hudson and Kortt, 1999), which we were striving for—in the model internal representation. This probably does not affect the performance, since in the zero-shot setting we employ both the 16-mer linker and 10-mer padding, and the difference is negligible (Figure 3.1), nevertheless it could be better aligned with biology by using a longer padding in future work.

For ESMC (ESM Team, 2024), presumably a BERT-style PLM, we provide the Fv sequence as a concatenation of light and heavy chains with a 16-mer (GGGS)4 linker. The logits are computed for the entire sequence in one pass encoder-only style without iterative decoding. The model is fed the full unmasked native sequence and we score how well the model 'reconstructs' this sequence. This seems trivial. However, due to the likely BERT-style training procedure, the model has learned not to fully rely on each input residue, as part of the positions that are 'masked', and actually contribute to the reconstruction loss, are not replaced by a mask token (80% of the time in predecessor model ESM2 (Lin, Akin, Rao, Hie, Zhu, Lu, Fazel-Zarandi, et al., 2022)), but by a random token (10% of the time) or kept as the ground-truth token (10% of the time). This means that the model has learned to not fully rely on the input sequence and can still provide meaningful level of confidence for the amino acids on the positions. Logits for linker positions are discarded from further processing as features.

Logits are then processed in the following way:

- For ESM-IF we softmax the logits and gather only likelihoods for the amino acid present in the native sequence, obtaining a vector of shape `(L,)` where `L` stands for sequence length. In case of these per-position likelihoods we use ProteinDPO weights (Widatalla, Rafailov, et al., 2024), which are the weights for the ESM-IF model fine-tuned on the Mega-Scale (Tsuboyama et al., 2023) thermostability dataset.

- For ESMC we do the same processing as in Section 3.1, but we keep the sequence dimension. In summary, we generate the following features with shape `(L,)`; this is how the features are computed on *each* position:

  - `ll` — log-likelihood of the native residue
  - `joint_ll` — sum of log-likelihoods of all amino acids
  - `logit` — logit (before softmax) of the native residue
  - `mean_logit` — mean of the logits for all amino acids

- `entropy` — entropy of amino acid probability distribution
- `top5_entropy` — entropy-like sum of the top 5 amino acids likelihoods $-\sum_{i=1}^{5} p_i \log p_i$ where p_i denotes the i-th largest value in the softmax-normalized amino acid probability distribution at a given position.

In addition, as global features, we include two global logit-based scores—global meaning per-antibody example, as opposed to per-position—derived from the ESMC 300M and ESM-IF models. These are the two best scores from the zero-shot approach, ZS4 (ESMC+ESM-IF) score and ZS3 (3xESMC) score, and are described in Section 2.2. The composite scores were chosen on the training set and are completely blind to the test set.

B-factors are computed with Normal Mode Analysis (NMA) using a Coarse-Grained Elastic Network Atom Contact Model (ENCoM) (Frappier and R. J. Najmanovich, 2014) from NRGTEN Python package (Mailhot and R. Najmanovich, 2021). Predicted and minimised structures from ABodyBuilder2 are used as input. We compute the B-factors as in the `ENCoM.compute_bfactors()` method, without scaling them up by 1000. B-factors are computed from eigenvalues and eigenvectors of the Hessian matrix, which is a square matrix of size $(3N)^2$ where N is the number of residues in the protein. Each eigenvalue describes a mode of oscillation in the modeled protein—its frequency—and its corresponding eigenvector of shape $3N$ describes the involvement of the residue in this mode of oscillation in each `x`, `y`, and `z` coordinate. The B-factor for a given residue is computed as a sum of its contributions across all modes, where each contribution is proportional to corresponding eigenvector element and inversely proportional to the mode's eigenvalue, this is summed over the three `x`, `y`, `z` dimensions in the eigenvector. The lower frequency modes—lower eigenvalues—thus cause greater displacement and have greater impact on the B-factor. These B-factors characterize the expected fluctuation of each residue around its equilibrium (energy-minimized) position. Faster than all-atom NMA, the coarse-grained ENCoM defines a force field that is still amino-acid specific. From this force field, the Hessian matrix is computed as the matrix of the second spatial derivatives of the potential energy, and diagonalized to obtain the eigenvalues and eigenvectors used in B-factor computation (Bauer et al., 2019).

Per-position features for each antibody are aligned to the respective numbered positions in the MSA, resulting in a matrix of shape (examples, MSA columns = 279). We assign Not a Number (NaN) to positions with gaps in a given sequence. We use the AHo numbering scheme, which aligns structurally equivalent positions across antibodies, enabling consistent per-position comparison despite sequence length variability. In contrast, naively padding sequences would misalign features and would make it difficult or impossible for traditional ML models to keep track of which features correspond to which structural positions. This is especially relevant for thermostability, as the antibody structure is largely conserved and it is the interactions between residues that

help maintain the folded state and prevent thermal unfolding.

We organize the model and feature processing in an `sklearn.Pipeline`. It performs one-hot encoding of amino acids if this type of feature is present, and the imputation of missing values (we use feature mean imputation for Lasso, and fill a constant 1e7 for tree-based models, except for HistGradientBoostingRegressor which supports missing values.) Therefore it is fit only for the training data (every time a subset in case of cross-validation).

We experiment with different subsets of positional features—198 MSA positions without any missing values; present in all sequences (`0g`), 213 MSA positions with under 2 gaps (`2g`) and all 279 positions (`All Positions`). In case of a subset of positions used, we experiment with adding global features for each of the six CDRs, by averaging the numerical positional features there (and we define these CDRs for our purposes as follows, using the AHo numbering, they are the AHo regions without a conserved structure (AHo defines only the regions with conserved structure, not CDRs explicitly), but smaller, and we also exlude the positions that were included as per-position features (`CDR agg`): CDR1: (32, 41), CDR2: (58, 68), CDR3: (109, 138). This applies to both the heavy and light chains, as the AHo numbering is equivalent for both.

Finally we concatenate the positional features with the global features (not necessarily all features for each model though), and use them as input to the models.

2.3.3 Model Types, Training Procedure and Evaluation

We employ two types of models—decision tree-based models and Lasso regression (linear regression with L1 regularization). Python 3.12 is used with the `scikit-learn 1.5.1` library (Pedregosa et al., 2011). We train the model types with varying combinations of input features (Section 2.3.2). We perform hyperparameter search, Grid Search for Lasso, or Random Search for decision trees, and select the best hyperparameters based on the average MSE reported for the two times repeated 5-fold cross-validation. We retrain the models with best hyperparameters on the entire training set and evaluate them on the holdout set. These are the models we report results for. In addition, we ensemble all tree-based models by averaging their predictions.

List of Lasso hyperparameters (Grid Search employed):

- `alpha` — regularization strength, values: `np.logspace(-8, 2, num=50)`
- `max_iter` — maximum number of iterations, 1e4
- `positive` — True/False. If True, restricts coefficients to be positive.

We evaluate three types of tree-based regression models:

- `RandomForestRegressor` — ensemble of decision trees trained via bagging

  - Hyperparameters tuned using randomized search over 20 configurations:
    * `max_depth`: [5, 7, 9, 11, 13, 15, 17, 19]
    * `n_estimators`: [10, 50, 100, 200, 500]

- `GradientBoostingRegressor` — standard gradient-boosted decision trees (GBDT).
  - Randomized search over 2500 configurations:
    * `n_estimators`: [100, 200, 300]
    * `learning_rate`: [0.01, 0.05, 0.1, 0.2]
    * `max_depth`: [3, 4, 5, 6]
    * `min_samples_split`: [2, 5, 10]
    * `min_samples_leaf`: [1, 2, 4]
    * `subsample`: [0.6, 0.8, 1.0]
    * `max_features`: ['sqrt', 'log2', None]
    * `loss`: ['squared_error', 'huber']

- `HistGradientBoostingRegressor` — histogram-based GBDT optimized for large datasets, supporting natively missing values.
  - Randomized search over 15 configurations:
    * `max_iter`: [70, 100, 150]
    * `learning_rate`: [0.01, 0.05, 0.1]
    * `max_depth`: [3, 5, 7]
    * `l2_regularization`: [0.0, 1.0, 10.0, 15.0]
    * `max_leaf_nodes`: [15, 31, 63]
  - Additionally, we evaluated a single fixed configuration to assess runtime and performance (Model K in Figure 3.5):
    * `max_leaf_nodes`: 15
    * `max_iter`: 100
    * `max_depth`: 7
    * `learning_rate`: 0.1
    * `l2_regularization`: 10.0

2.3.4 Feature Importance

We employ permutation feature importance (Breiman, 2001; *Scikit-Learn Guide* 2025) to assess the importance of each feature in the model. This method evaluates the impact

of shuffling a feature on the model's performance on unseen data, reporting the change in prediction error. This has several advantages over using impurity importance or Lasso coefficients. Notably, it is computed on the test set and actually reflects how features affect the model's performance. We find that especially tree-based models overfit the training set (there is a performance gap between holdout and training prediction error), and this necessarily leads to the model using features that are only useful for the train set. It is applicable to any model type, and it is directly possible to compare the importance number, the impact on prediction error, across different model types. A disadvantage is that it is computationally expensive, as it requires (features × shuffle_repeats) evaluations of the model on the test set, while the impurity importance or Lasso coefficients are easily obtained from the trained model alone.

Permutation feature importance is computed with scikit-learn's `inspection.permutation_importar` with 10 permutations of each feature. First, the baseline score (R^2) is computed using the original features. Then, the output is the difference between the baseline score and the R^2 after shuffling, yielding a sample of 10 values for each repeat for each feature.

To visualize important positions in a 3D structure, we selected the Fab structure of certolizumab, which is included in the AbProp dataset. Certolizumab was chosen because it has CDR lengths representative of the dataset and a high-quality X-ray crystal structure of the Fab (PDB ID: 5wuv, J. U. Lee et al., 2017) with a resolution of 1.95 Å and an R_{free} of 0.179.

For plots regarding model performance or feature importance we use `matplotlib 3.9.1` (Hunter, 2007) and `seaborn 0.13.2` (Waskom, 2021). For visualization of important positions in a 3D structure we use Mol* and MolViewSpec (Sehnal et al., 2021; Bittrich et al., 2024).

2.3.5 Light Chain Type Determination

To determine whether a given variable light (VL) domain originates from a kappa or lambda light chain, we analyzed the amino acid at AHo position L146, which corresponds to the third-to-last residue in the IGKJ or IGLJ gene segment which is the terminal segment in VL sequence. For both human and mouse functional genes, this residue is highly conserved: IGKJ functional genes encode a glutamate (E) or aspartate (D), while IGLJ functional genes encode a threonine (T). Germline sequences and alignments were obtained from the IMGT database[5] (Lefranc, Giudicelli, et al., 2009).

2.3.6 Comparison of the Zero-Shot and Supervised Approaches

To compare the predictions of zero-shot and supervised models, we evaluate the best zero-shot scores also on the holdout set, which is unseen by both zero-shot and su-

[5] `https://www.imgt.org/IMGTrepertoire/Proteins/index.php#h2_16`

pervised models. We employ 10,000x bootstrap resampling to compute Spearman correlation coefficients between model and experimental T_m values. We employ permutation testing, as described above (Section 2.2.3), to compare the models.

We also include three literature models—ESM-IF, ProteinDPO, and SaprotHub-Thermostability. The first two are zero-shot methods whose results on the full AbProp dataset were reported in (Widatalla, Rafailov, et al., 2024). We re-compute these results to provide confidence intervals and to evaluate performance on our holdout subset. For both, we use the average log-likelihood score as described in Section 2.2.3; since ProteinDPO is simply a fine-tuned ESM-IF, only the weights differ and the processing pipeline remains identical. We are using the ProteinDPO `paired` variant (as referred to in the paper), as this was the only variant with published weights.

SaprotHub-Thermostability (Su, Z. Li, et al., 2024) was trained on the `human_cell` portion of Meltome using the FLIP preprocessing pipeline (Dallago et al., 2021) (but using a different train/test split than defined by FLIP), as detailed in Section 1.4.[6] Although trained on thermostability, it has not been exposed to AbProp examples. We ran the Google Colab notebook[7] specifying the regression adapter `SaProtHub/Model-Thermostability-650M` and supplying ABodyBuilder2-predicted Fv structures. Because SaprotHub requires single-chain inputs, we modified the PDBs to include a single chain id for both the original heavy and light chains, and renumbered the residues.[8]

[6]The model adapter is available at `https://huggingface.co/SaProtHub/Model-Thermostability-650M`.

[7]`https://colab.research.google.com/github/westlake-repl/SaprotHub/blob/main/colab/SaprotHub_v2.ipynb`

[8]We also added a 16-residue gap in the numbering between the two original chains. Depending on implementation of SaprotHub's preprocessing, this may or may not have an effect.

Chapter 3

Results and Discussion

3.1 Zero-Shot Prediction

We evaluate a diverse set of 14 pre-trained protein language models on the task of antibody thermostability prediction. We do not perform any additional training—thus operating in a zero-shot setting—and use the models as they are. Protein language models return for each position of the protein a probability distribution over the 20 amino acids (plus special tokens). We aggregate these per-position distributions into a single score for the entire protein sequence. The tested language models take as input either the protein sequence, the structure, or both. We evaluate the resulting scores on the entire AbProp dataset. Given the relatively small size of the dataset, we use bootstrapping with 10,000 dataset resamples to better assess the variability of performance estimates and to construct confidence intervals. For statistical comparison between models, we additionally apply a permutation test.

Additionally, we cluster the best-performing scores based on their pairwise correlation and combine the least-correlated ones into a single score, with the goal of producing a more informative and robust measure.

Finally, we evaluate the combined score on a complementary ProteinGym (Notin et al., 2023) stability benchmark with ~135,000 single and double mutants from 66 non-antibody proteins.

3.1.1 Performace of Tested Models

We evaluate a carefully selected set of protein language models including 4 antibody-specific models, 5 structure-informed PLMs, 9 BERT-style, and 5 autoregressive models. The most important distinction between the models, as all employ state-of-the-art deep learning architectures, is perhaps the training data and the training objective. We include models that were trained on UniRef (Suzek, Y. Wang, et al., 2015) sequences, large metagenomic sequence databases, models that were trained on antibody sequences, and

models that were trained on protein structures experimental or predicted, single-chain or complexes. Overview of the models is provided in Table 3.1.

Model	Variants / Size	Training Data	Input	Pre-training Task	Architecture
ESMC	300M, 600M	Uniref + MGnify + JGI; 83M, 372M, and 2B clusters @70%	Seq	Masked LM	Transformer
ESM3	open 1.4B	UniRef + MGnify90, 70–90 sampling	Seq + Struct + Fun	Masked LM	Transformer
ESM-2	150M, 650M	UniRef50 (+ UniRef90 expansion) ~60M	Seq	Masked LM	Transformer
ProtBert	420M	UniRef100 (216M)	Seq	Masked LM	Transformer
ProtBert-BFD	420M	UniRef100 + BFD100 (2.1B)	Seq	Masked LM	Transformer
ProGen2	base 764M	UniRef90 + BFD30 (1/3 of UniRef90's size)	Seq	Causal LM	Transformer decoder
ProteinMPNN	—	High-res PDB structures, PDB30 weighted sampling (23K clusters)	Struct	Inverse Folding	GNN
AbMPNN	—	SAbDab(CDR90) 1.7K + OASp–ImmuneBuilder(CDR90) 86K	Struct	Inverse Folding	GNN
ESM-IF	124M	CATH40 (16K), AlphaFold2 predicted UniRef50 (<500 amino acids) 12M	Struct	Inverse Folding	GNN + Transformer
AntiFold	124M	*Identical to AbMPNN*	Struct	Inverse Folding	GNN + Transformer
AbLang2	45M	OASu(CDR3:100)95 + OASp95 (35.6M unpaired, 1.26M paired)	Seq	Masked LM	Transformer, (ESM-2 35M-like)
IgBert	420M	OAS95 (1.42B unpaired, 1.31M paired)	Seq	Masked LM	Transformer (ProtBert)

Table 3.1 Overview of protein language models evaluated for antibody thermostability prediction. ESMC stands out with its training data diversity as it contains 2 billion sequence clusters from Joint Genome Institute (JGI) protein sequences which were clustered at 70% identity level, however exact training procedure is yet to be released in an upcoming pre-print (ESM Team, 2024). The antibody-specific models are IgBert, AbLang2, AbMPNN, and AntiFold, while the remaining models are general-purpose. In the "Training Data" column, clustering identity thresholds (with or without percentage signs) and dataset sizes are shown where available. OAS — Observed Antibody Space (Tobias H. Olsen, Boyles, et al., 2022); OASp — paired sequences only; OASu — unpaired sequences; SAbDab — Structural Antibody Database (James Dunbar, Krawczyk, et al., 2014); BFD — Big Fantastic Database (Jumper et al., 2021); CATH (Sillitoe et al., 2021); JGI (Nordberg et al., 2014).

We find that structure-informed language models rank among the top performers on our benchmark (Figure 3.1), consistent with results from the ProteinGym benchmark (Notin et al., 2023). However, the best-performing model overall is ESM Cambrian (ESMC) (ESM Team, 2024), a sequence-only model. It significantly outperforms all other models at the 0.05 significance level (adjusted for 21 comparisons) with the exception of a few: specifically, the two structure-based ESM-IF models, where adjusted p-values are 0.053 and 0.041—borderline cases where the null hypothesis is not rejected under the Holm sequential procedure. Additionally, ESMC's larger variant (ESM 600C) and ESM3, another model from the same group, are not significantly outperformed. Interestingly, ESM3 achieves better performance when using only sequence input rather than both sequence and structure.

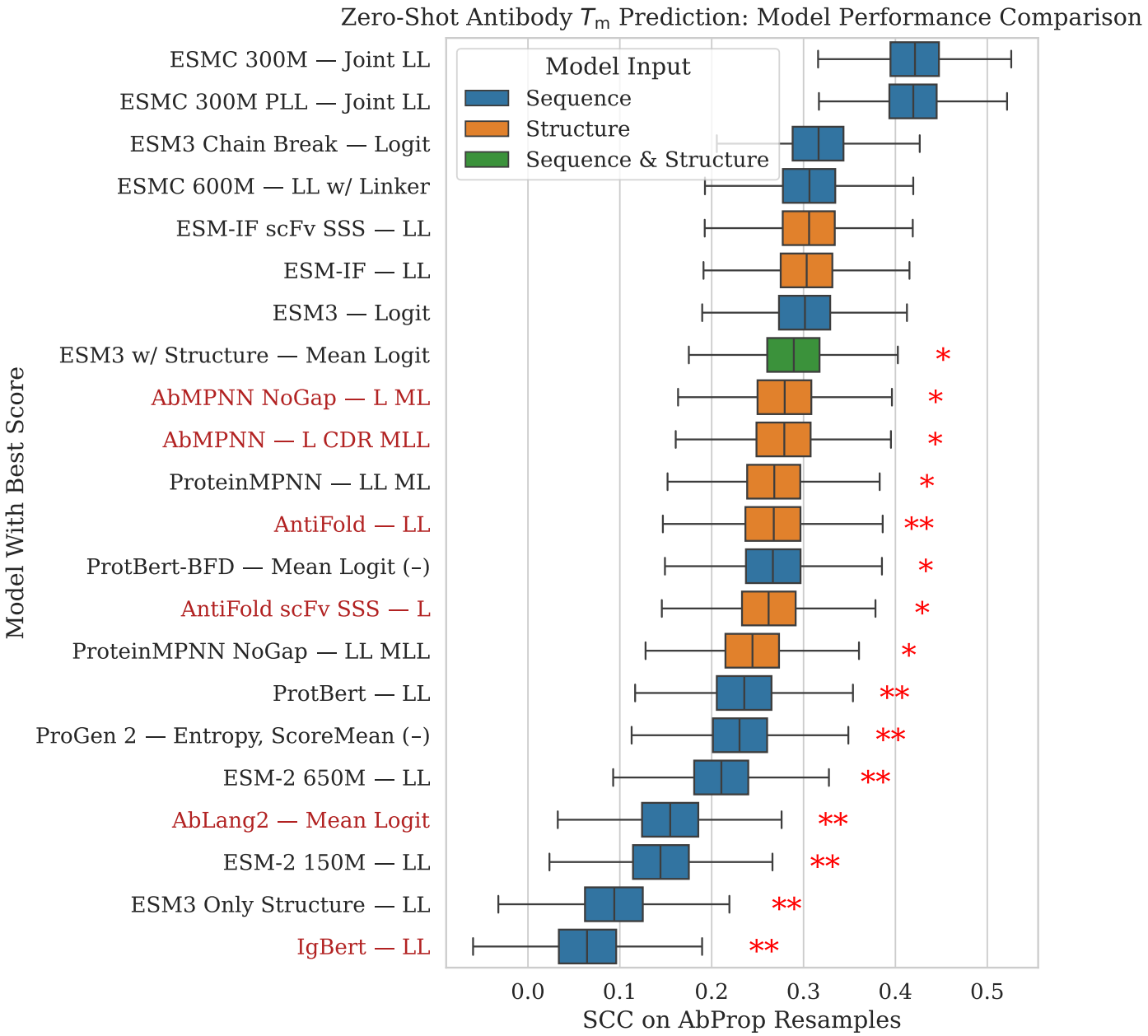

Figure 3.1 Zero-shot prediction of antibody thermostability by diverse pre-trained protein language models. Antibody-specific models are highlighed in red. For each antibody in the AbProp dataset, per-residue probability distributions are aggregated into a single score. Scores that originally exhibited a negative correlation with T_m were inverted (indicated by a '(–)') so that higher values consistently denote better performance. For models with multiple available scores, only the best score is shown (the score reported after '—'). SCC on 10,000 bootstrap resamples of the entire AbProp dataset is reported. Pairwise comparisons were performed with a permutation test using Holm adjustment for multiple comparisons (n=21). Significance markers (* and **) denote adjusted p-values below 0.05 and 0.01, respectively, indicating that the top model ESMC 300M with the joint log-likelihood score significantly outperforms most of the other models with the exception of ESM-IF and the new ESM-family models—its larger variant ESMC 600M, and ESM3.

Abbreviations: PLL = pseudo-log-likelihood; SSS = score second sequence; LL = log likelihood; ML (MLL) = mean residue likelihoods (log-likelihoods) over decoding replicates; ScoreMean = mean score from left-to-right and right-to-left decoding. For details see Section 2.2.3.

We hypothesize the unexpected lead of the three sequence-only models could be due to 1) the predicted structures of antibody Fv are not accurate enough. Whereas in the ProteinGym benchmark—where structure-informed PLMs consistently occupy the top-performing positions—the proteins for the stability benchmark were selected from PDB (Tsuboyama et al., 2023; Notin et al., 2023), and a reasonable assumption is they are high quality structures given the small size of the domains and the selection process.[1] Therefore, the key methodological difference between our setting and the ProteinGym benchmark is that we predict structure from sequence and then, using inverse folding models, try to predict back or *score* the original sequence—introducing error in the structure prediction step. While in the ProteinGym benchmark, (likely) high-quality experimental structures are available, and the single wild-type structure is used to score near (single or double) mutants with the inverse folding model. With antibody-specific structure prediction, we aim to capture the inductive biases and learned knowledge of the antibody folding model, which may benefit the inverse folding step—especially given that structure is generally more conserved than sequence, potentially allowing for better generalization even in inverse folding models not trained specifically on antibodies. In this way, we first leverage the knowledge embedded in the antibody-specific folding model and then use its predicted structural representation—shown to be effective for stability prediction in ProteinGym—to score the sequences. 2) While the description of ESMC model is not available yet, ESM3 employs two thresholds of clustering the training sequences (Hayes et al., 2024), 70% (outer level) and 90% (inner level), where it samples an outer level cluster and then a random sequence from the inner level [2]. This could allow the distingushing of (evolutionary stability) of sequences 70–90% identical, and could learn the evolutionary stability on the protein-variant level,[3] as opposed to the predecessor ESM2 which was trained on UniRef50[4], essentially boiling down the protein variant distribution at the 70–90% identity level to a single or no example. This 70-90% identity matches the diversity in the AbProp dataset. Perhaps

[1]Although the structures of the small domains were ultimately predicted again using AlphaFold2 (Jumper et al., 2021) prior to downstream analyses—one of the reasons they give is to trim the protein to the stable fold, without e.g. flexible loops—we assume the default settings were used. In that case, AlphaFold2 could have used the available PDB structure as a template, with nothing preventing it from closely reproducing it during prediction. Or it could simply reproduce the folded structure as seen during training, since the original structures were likely in its training data.

[2]They limit the number of sequences (inner clusters) within an outer cluster to 20. It is not reported why, however it might reduce the sampling bias which we discuss later.

[3]This raises question why would the 90% clustering be not enough and why there would a benefit from the outer level (70%) cluster training data distribution weighing.

[4]Lin, Akin, Rao, Hie, Zhu, Lu, Fazel-Zarandi, et al. (2022) decribe "To increase the amount of data and its diversity, we sampled a minibatch of UniRef50 sequences for each training update. We then replaced each sequence with a sequence sampled uniformly from the corresponding UniRef90 cluster." We understand this as sampling from the *single* UniRef90 cluster that represents the UniRef50 cluster, i.e. sampling sequences 90-100% identical, not 50-90%. This supports our narrative, however the phrasing is not entirely clear and the meaning would change if it was meant to be in plural like "UniRef90 *clusters*".

a similar procedure was employed for the newer ESMC developed by the same team.

Meier et al. (2021) examined the effect of the clustering level of pre-training data by retraining the ESM-1b model architecture on UniRef with 30%, 50%, 70%, 90%, and 100% clustering thresholds. On a validation set of 10 single-mutation DMS assays, they show that 90% clustering is optimal (average Spearman correlation across assays: 0.564), outperforming both 50% (0.537; used in ESM-1b) and 30% (0.456). Interestingly, they also observe a degradation of performance for UniRef100 (0.458), where performance drops earlier with increased training steps.

Using a test set of 31 assays, they show that clustering the training data at 90% identity improves the model's variant prediction performance (naming this model ESM-1v) from 0.424 (ESM-1b) to 0.457. It should be noted, however, that while they label this improvement as significant on their test set, it is not reproduced on the ProteinGym benchmark: ESM-1v (single) performs worse than ESM-1b on ProteinGym, with scores of 0.374 vs. 0.394 across all 217 assays. The same trend is observed for the subset of 66 stability assays (0.437 for ESM-1v (single) vs. 0.500 for ESM-1b).

In any case, applying 30% clustering leads to a performance drop approximately fourfold larger, supporting the general hypothesis that lower clustering thresholds degrade model generalization. The reason behind the performance drop for UniRef100 is unclear. We suspect that increased sampling bias in the dataset at this level could be a contributing factor. For example, the largest cluster in UniRef90 in 2015 contains 43,060 sequences of 'Ribulose bisphosphate carboxylase large chain' (RuBisCO) enzymes, mainly from eukaryotes (Suzek, Y. Wang, et al., 2015). These would contribute 43,060 training examples for a UniRef100-based model, but only one for UniRef90. Vidyasagar et al. (2004) show that RuBisCO large chain amino acid identity varies from 85–96% among higher plants, 52–88% among algae, and 37–82% among bacteria, concluding that RuBisCO large chains are highly conserved among higher plants. This suggests the largest RuBisCO cluster is dominated by higher plant sequences. A perhaps more compelling example is the largest UniRef90 cluster in 2007, which was formed by over 3,600 Matrix 1 proteins of Influenza A virus (Suzek, Huang, et al., 2007). Taken together, the fact that the largest UniRef90 clusters include mostly higher-plant sequences and Influenza proteins indicates a strong human bias in data collection. (However, depending on how the bias is defined, plants do in fact comprise approximately 80% of Earth's biomass (Bar-On et al., 2018).) As of April 2025, we report that 6 out of the 8 largest UniRef90 clusters are Influenza A virus proteins, with the largest containing ~63,000 sequences. For RuBisCO large chain, the two largest UniRef90 clusters have ~28,000 and ~17,000 members, respectively, spanning species such as Soybean, Common Sunflower, *Arabidopsis thaliana*, and *Coffea arabica* (though this warrants a more detailed taxonomic analysis).

Additionally, while clustering at the high 90% level appears to improve *single-mutant* variant effect prediction (as shown by (Meier et al., 2021), but not reproduced in the

ProteinGym benchmark), it may not translate to *multiple-mutant* scenarios. In the AbProp dataset, for example, the average sequence identity is 63.0%, which corresponds to approximately 106 mutations between pairs of sequences.

We do not observe a performance degradation when comparing a UniRef50-trained model (ESM-2 650M) to a UniRef100-trained model (ProtBert), with a p-value of 0.57 for the null hypothesis that both models perform equally.

All in all, the explanation for the good performance of ESM3 and ESMC models may be that clustering at 70% identity, for the training data distribution weighing, helps to reduce the sampling bias, and keeping the actual training examples diverse at 90% identity helps to learn the distribution of (evolutionarily) stable sequences at this 70–90% level.[5]

3.1.2 Comparing Antibody-Specific Models with General Models

We find that antibody-specific PLMs perform perhaps counterintuitively unremarkably in this antibody stability task (Figure 3.1). We find that out of 11 sequence-only general PLMs, antibody sequence models, AbLang2 and IgBert show no statistically significant improvement over the remaining nine general models (n=22 comparisons, none significant). In the opposite direction, however, out of total 22 comparisons whether the general models outperform the antibody-specific models, 10 comparisons are statistically significant (adjusted bootstrap test likelihood <0.05) in favor of the general models.

While a vast dataset of antibodies successfully expressed in human or animal cells could in principle support thermostability prediction—given that expression levels have been shown to correlate with thermostability (Jain et al., 2017)—the binary nature of the available information may be limiting. Successful expression results in inclusion in the dataset, whereas sequences that fail to express are eliminated during B cell development and remain unobserved. Perhaps a continuous measure of expression level from cells would be needed to capture more nuanced information relevant to thermostability.

A reason why general PLMs perform better could be that the proteins in the evolutionary space could be selected for their stability in the heritable evolutionary process, whereas there is inherent randomness in the antibody gene rearrangement which gives rise to the antibody sequences which precludes to some extent the evolutionary selection between the antibody sequences. Therefore, antibody-only datasets may exhibit a wide distribution of thermostability, which means the likelihood scoring with models that merely replicate this distribution is less informative about the actual magnitude of thermostability (since any value is essentially permissible). In contrast, models trained on general protein sequences—which can undergo evolutionary selection for thermosta-

[5]We assume that ESMC follows a training procedure similar to ESM3, as no preprint is currently available.

bility and may therefore reflect a narrower, stability-biased distribution—could provide more informative likelihood scores regarding thermostability.

This, however, does not explain why the structure-informed models fine-tuned on antibodies—AntiFold (a fine-tuned ESM-IF) and AbMPNN (ProteinMPNN)—do not exhibit significant performance degradation. AbMPNN, in its default setting, shows a small but non-significant boost of 0.01 (bootstrap test likelihood = 0.394, AbMPNN superior to ProteinMPNN), while AntiFold shows a slightly larger but also non-significant decrease of 0.036 (bootstrap test likelihood = 0.114, AntiFold inferior to ESM-IF) in median Spearman correlation. One possible explanation is that both models were pre-trained on general protein datasets (e.g. PDB, UniRef50) and retain useful representations from that training. Their fine-tuning on antibody data was performed with a small learning rate, potentially limiting the extent of adaptation and avoiding overwriting general protein knowledge.

Indeed, AntiFold was explicitly fine-tuned using a layer-wise learning rate decay strategy, where earlier layers were updated more conservatively to mitigate catastrophic forgetting (Høie et al., 2024). The authors report that initializing from random weights—as opposed to pre-trained ESM-IF weights—results in markedly worse performance in antibody structure prediction (CDR-H3 amino acid recovery: 31.5% vs. 58.4%), supporting the importance of pretraining.

The sequence-only model AbLang2, by comparison, was randomly initialized and trained only on antibody sequences, which may contribute to its relatively weaker performance in thermostability prediction. While IgBERT was initialized from a general protein model (ProtBERT) and further trained on approximately 1.4 billion unpaired antibody sequences, AntiFold and AbMPNN were fine-tuned on a much smaller dataset of ~148,000 paired VH–VL chains and ~3,500 antibody-antigen complexes. The 10,000 times larger dataset and likely much greater number of training steps for IgBERT may have led to increased specialization, potentially reducing the retention of general protein features learned during pretraining. In contrast, AntiFold and AbMPNN were fine-tuned more conservatively on smaller antibody datasets, likely retaining much of the general knowledge learned from pretraining.

As a further test of this hypothesis, one could compare the ESM3-open (1.4B) model we used—which was trained without the antibody OAS dataset—with the closed variant of ESM3 (1.4B), which includes OAS sequences. If the closed model performs worse, it could support the idea that including antibody sequences with a broad, less stability-constrained distribution has a "normalizing" effect that reduces the model's ability to capture thermostability-related signals.

3.1.3 Score Clustering and Combination

We select the top 5% zero-shot scores in absolute median SCC with T_m (Supplementary

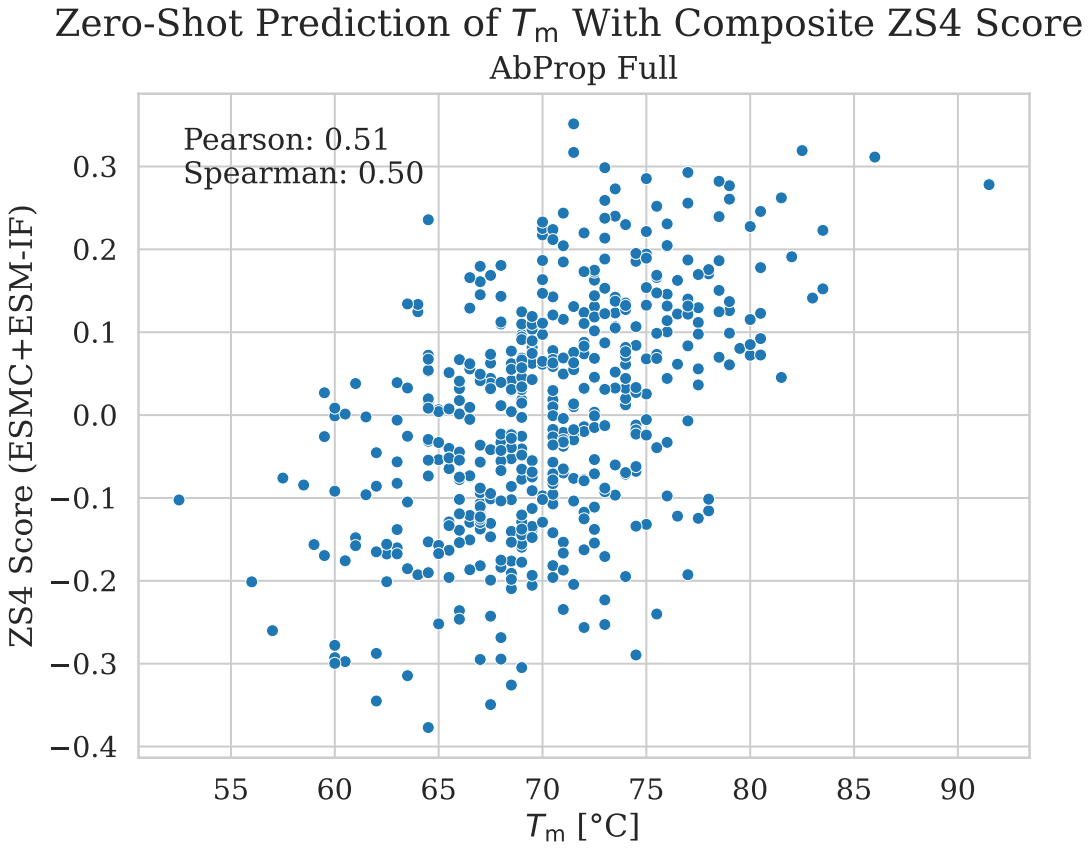

Figure 3.2 Scatter plot of predicted vs. experimental T_m using the composite ZS4 score on the entire AbProp dataset. For comparison, the ESMC-only ZS3 score achieves a SCC of 0.48 and PCC of 0.49 on the full set (not shown). The best zero-shot result as we know of in literature is for the ProteinDPO model (Widatalla, Rafailov, et al., 2024), who report SCC of 0.35. For details, see the main text.

We rank and select scores for combination, as described above, on the AbProp training set (Section 2.3.1) to ensure an unbiased evaluation of the composite scores on the holdout set. For completeness, we also report results on the full AbProp dataset—since all scores remain zero-shot; training is performed—but note that these values reflect rather "validation" performance, as the composite score constituents were selected based on their performance on a subset of the full dataset.

ZS4 score achieves 0.49 SCC on the full AbProp dataset (Figure 3.2, closely followed by the ESMC-only ZS3 score with 0.48 SCC. On the holdout set, ZS4 score achieves 0.57 SCC and ZS3 score 0.58 SCC.

The best zero-shot result reported in the literature comes from ProteinDPO, a fine-tuned version of the ESM-IF model fine-tuned on the Mega-Scale dataset, which

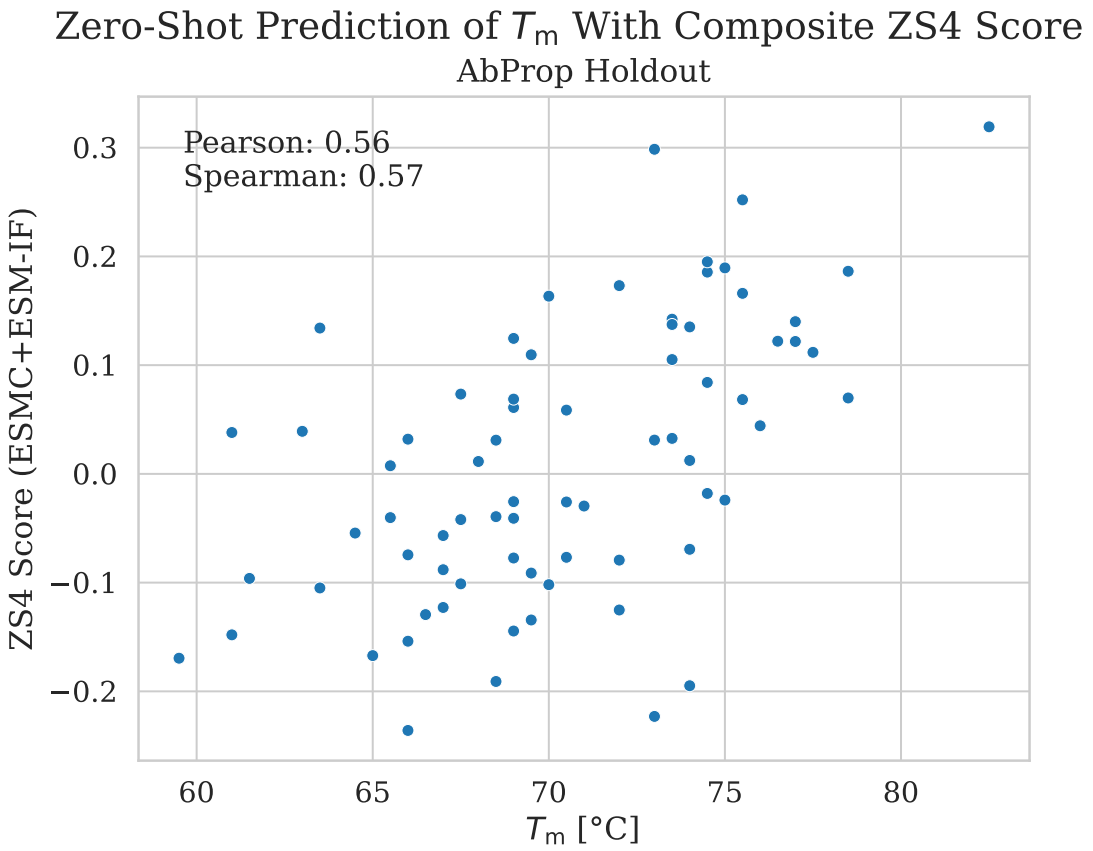

Figure 3.3 Scatter plot of predicted vs. experimental T_m using the composite ZS4 score on the holdout subset of AbProp dataset. For comparison, the ESMC-only ZS3 score achieves a SCC of 0.58 and PCC of 0.57 (not shown).

achieves an SCC of ~0.35 on the full AbProp dataset. In comparison, vanilla ESM-IF reaches ~0.25 (Widatalla, Rafailov, et al., 2024). Our composite scores surpass both ProteinDPO and ESM-IF on the full AbProp dataset. While this comparison may be slightly favorable to our method due to the score selection being performed on the same dataset (the training subset), we address this by reproducing their predictions and providing a fair evaluation on the holdout set in Section 3.2.3.

Additionally, we include an independent benchmark using ProteinGym, where our antibody-derived score ZS3 score ranks as the top-performing method in the sequence-only category for predicting double mutant stability.

3.1.4 ProteinGym Evaluation

We evaluate the performance of our composite ZS3 (3xESMC) score on the ProteinGym stability benchmark, which contains 66 proteins with a total of 65,000 single and 69,000 double mutants, alongside 79 models already included in the v1.2 version of the benchmark and 5 simple scores we derived from ESMC (Section 2.2.3). The benchmark is in many ways orthogonal to our primary antibody thermostability dataset, as it contains small domains with a median length of 57.5 amino acids (2 exceptions with lengths 220 and 770, the latter being a three-domain protein), while the AbProp T_{m} dataset consists only of 483 antibodies with a median length of 255 amino acids, comprised of two domains. ProteinGym stability benchmark contains only single and double mutants, while the AbProp dataset has a mean sequence identity between pairs 63.0%, corresponding to about 106 mutations between two sequences. We break down the model performances—in total 84 zero-shot scores—by model type and the mutation depth. For consistency and simplicity, we employ Spearman correlation, although the benchmark evaluation includes also other metrics.

We find that an ensemble score of models we selected based on performance on the antibody thermostability dataset—ZS3 (3xESMC) score—achieves the best performance among all sequence-only models on the double mutant stability prediction task (SCC 0.26, next best sequence only ESM2-150M with SCC 0.23 Figure 3.4). For single mutants, however, ZS3 score is outperformed by the standard log-likelihood score of the native sequence from the ESMC model (SCC 0.26 vs SCC 0.37; Supplementary Figure 14).

When ranking all 85 evaluated model scores by the difference in performance between double and single mutants, four of our scores—including ZS3 score—occupy the top four positions. This indicates greater robustness to multiple mutations compared to the rest of the ProteinGym v1.2 benchmark (79 models present). These four scores—ZS3 score, `joint_ll_score`, `entropy_score`, and `top5_entropy_score` (see Section 2.2.3)—incorporate the likelihood distribution over residues at each position, rather than focusing solely on the likelihood of the native residue (the residue present in the sequence; wild-type or mutated). This may explain their relatively weaker

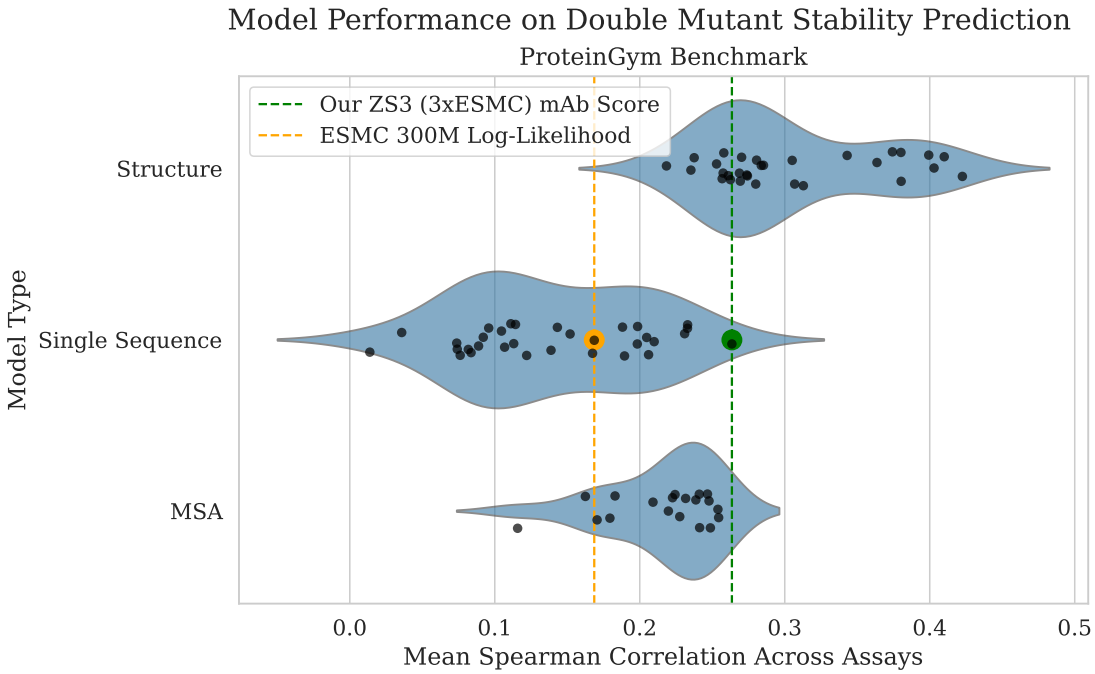

Figure 3.4 Performance of the ZS3 score, we derived for antibody thermostability, on the complementary ProteinGym stability benchmark. The score is evaluated on a total of 69,000 double mutants across 51 distinct proteins, alongside 79 other models included in the ProteinGym v1.2 benchmark. Our score (shown in green) outperforms all other sequence-only models in predicting double mutant stability, hinting at increased robustness when applied to multiple mutant stability prediction. For comparison, we also show in orange the commonly employed log-likelihood score of the native sequence, derived from the same ESMC model (and being one of the constituents of ZS3 score).

performance on single mutants, yet greater robustness in the presence of multiple mutations, even when limited to just double mutants.

Future work could explore deriving similar global scores from the top-performing models in ProteinGym and evaluating them on a multiple mutant benchmark.

3.2 Supervised Learning

Following the exploration of existing protein language models, we turn to the supervised learning task of antibody thermostability prediction. We use the AbProp dataset training/holdout split as in (Widatalla, Rollins, et al., 2023) and train a variety of models, including tree-based models and Lasso (linear regression with L1 regularization). We also explore the effect of different feature sets on the model performance, including one-hot encoding of the amino acids, the use of the zero-shot model residue likelihoods, residue B-factors computed from predicted structures and global zero-shot scores we developed in Section 3.1. Hyperparameters were optimized using repeated 5-fold cross-validation on the training set, selecting configurations that minimized mean squared error (MSE) between predicted and experimental T_{m} values. Details of the models and hyperparameters and the rationale for the train/holdout split are provided in Section 2.3.

The best model, according to validation performance (i.e. the average performance on the validations folds), Gradient Boosting Tree (Model M) reaches 0.64 SCC on the holdout set (Figure 3.5 and table 3.2). An ensemble of all tree-based models improves SCC to 0.69 SCC and is the best also in other test metrics, outperforming the best existing model on the AbProp dataset, a Graph Attention Network (0.62 SCC) from (Widatalla, Rollins, et al., 2023).

Model	Val MSE	MSE	SCC	PCC	R^2	MAE
Gradient Boosting Tree (M)	19.5	14	0.64	0.61	0.38	2.9
Random Forest (H)	19.5	13	0.68	0.63	0.39	2.7
Lasso (C)	19.6	15	0.60	0.58	0.33	3.0
Random Forest (I)	19.7	14	0.65	0.62	0.39	2.9
Random Forest (G)	19.7	13	0.69	0.63	0.40	2.7
Random Forest (J)	19.9	13	0.67	0.63	0.40	2.8
Random Forest (F)	20.1	14	0.65	0.62	0.38	2.9
Hist Gradient Boosting Tree (L)	20.5	13	0.67	0.64	0.40	2.7
Lasso (E)	20.5	15	0.60	0.58	0.31	3.2
Lasso (D)	20.5	15	0.60	0.58	0.31	3.2
Hist Gradient Boosting Tree (K)	21.7	13	0.68	0.65	0.42	2.7
Lasso (A)	22.3	17	0.51	0.50	0.24	3.1
Lasso (B)	22.8	18	0.47	0.46	0.21	3.2

Table 3.2 Performance of individual models trained on the AbProp dataset sorted by cross validation MSE (Val MSE). Val MSE was also used to select the best hyperparameters, and the models here already employ the selected hyperparameters and differ either in set of included features (see Figure 3.5) or the model type. Remaining metrics are computed on the holdout set of 73 examples. Notably, while the Lasso models achieve comparable performance on the validation set, they underperform on the holdout set compared to the tree-based models. The zero-shot scores (or other global features) are included in all models except Lasso A—which uses only the aligned one-hot encoding of the sequence, and thus can be considered as a baseline—and Lasso B, which uses per-position log-likelihoods of the native residues.

Figure 3.5 (Caption next page.)

Figure 3.5 (Previous page.) Predicted vs. experimental T_m values for supervised models on the AbProp dataset. Each panel shows the performance of a supervised regression model trained with various feature combinations and algorithms, including Lasso regression, Random Forests, Histogram-Based Gradient Boosting Trees, standard Gradient Boosting Trees, and their ensemble. Scatterplots compare predicted and experimental T_m, along with Spearman correlation and other metrics on the holdout set.

Models differ in the input feature sets used. ZS denotes zero-shot features developed in Section 3.1—ZS4 score and ZS3 score. Positional features include: AA (one-hot amino acids), ESM (log-likelihoods from ESMC 300M), ESM_O (other ESMC scores), IF (ESM-IF likelihoods), BF (B-factors computed from predicted structures), and ABM (AbMPNN likelihoods). '2g' and '0g' indicate models trained on positional features restricted to positions with fewer than 2 or 0 gaps in the dataset-wide multiple sequence alignment (MSA), respectively. 'CDR agg' refers to positional features averaged over each of the complementarity-determining regions (see Section 2.3.2).

All models were trained using repeated 5-fold cross-validation on the training portion of the dataset, optimizing hyperparameters to minimize mean squared error (MSE). Final evaluation was performed on the independent holdout set defined in Widatalla, Rollins, et al. (2023). Cross-validation results used for model selection are reported in Table 3.2.

The best-performing model on the validation folds was a Gradient Boosting Tree using 0g-filtered ESM features (Model M), while an ensemble of all tree-based models (Model N) achieved the strongest holdout performance overall, with 0.69 Spearman correlation, outperforming the best previously reported model on AbProp (SCC 0.62). Despite competitive validation results, Lasso models consistently underperformed on the holdout set, likely due to their linear nature and conservative bias toward the mean. In contrast, Random Forests and boosting methods generalized better, particularly in capturing the dynamic range of T_m values.

Models trained using repeated cross-validation showed similar validation performance between Lasso regression and Random Forest models, yet Random Forests demonstrated better generalization to the independent holdout set Table 3.2. The conservative nature of Lasso, characterized by its linearity and regularization, provided stable predictions on cross-validation folds. However, perhaps the Random Forests' ability to capture nonlinear relationships resulted in improved performance on the holdout set, which contains a more balanced range of T_m values (interquartile range 7 °C vs 6 °C on the train) Figure 3.6. It seems thus that Random Forests may suffer from larger variance, impacting the ranking in a crowded narrow region, however, performing well in the distinction between good and bad values. While Lasso is biased to the mean and cannot predict well the samples with low or high T_m, being overly conservative for extreme value samples (bad, good) and thus having limited practical applicability in filtering out the bad samples or selecting the good ones.

Later, after reflecting on the results, we decided to examine replacing the CV MSE scoring (which is equivalent to the default R2 scoring) with our goal metric Spearman. We retrained Model G (Random Forest employing a wide set of features) using cross-

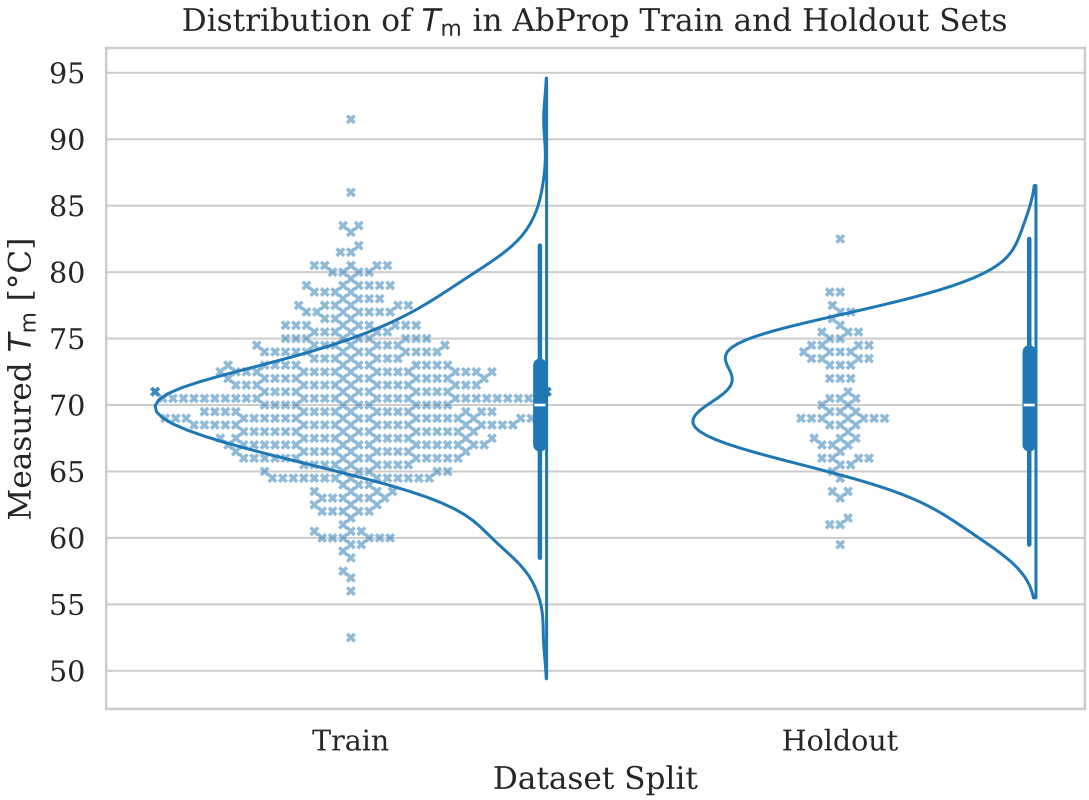

Figure 3.6 Distribution of melting temperatures T_m—the target variable—in the AbProp training and holdout datasets. Boxplots indicate the interquartile range (IQR). The IQR is slightly lower in the training set (6.0 °C) than in the holdout set (7.0 °C) which corresponds to the wider peak shown, while the standard deviation is higher in the training set (5.2 °C vs. 4.7 °C), reflecting the presence of more extreme outliers. We hypothesize Cross Validation (CV) with SCC as the scoring metric leads to rather focusing on fine-tuning the predictions in the crowded peak region of the training set, while the employed MSE forces the model to learn the global scale prediction, as larger errors, often seen in the examples with extreme values of measured T_m (Figure 3.5), are weighted more, which may benefit the prediction on the holdout set which exhibits less crowding around the mean.

validation with Spearman scoring and obtained SCC 0.63 and MAE 2.9, a worse result than original Model G with SCC 0.69, and MAE 2.7. Average SCC on validation folds was 0.51, which may reflect the difficulty of the local distinguishing thermostability in the crowded unimodal distribution. MSE which weights more the outliers helps the prediction by perhaps focusing more on the larger incosistencies rather than fine-tuning the prediction by small amounts to improve the spearman in the most dense region of the training data (Figure 3.6). And the usage of MSE could contribute to why our model outperforms the best existing model (SCC 0.62) by Widatalla, Rollins, et al. (2023) which was obtained with CV using Spearman scoring.

3.2.1 Feature Importance

We assess feature importance using permutation importance (Breiman, 2001; *Scikit-Learn Guide* 2025), which quantifies the decrease in model performance when the values of a single feature are randomly shuffled, thereby breaking the relationship between that feature and the target variable. Unlike impurity-based importance or Lasso coefficients, this method evaluates impact directly on the test set, offering a more faithful estimate of a feature's contribution to generalization. This is especially valuable for tree-based models, which we observe to overfit, and may rely on features that are only informative during training. The approach is model-agnostic and yields comparable importance scores across different types of models. However, it is computationally intensive, requiring multiple evaluations per each feature. We use scikit-learn's `permutation_importance` with 10 shuffles per feature, reporting the average decrease in R^2 relative to the baseline computed on unshuffled data.

We analyze feature importance for each Lasso and Random Forest model. [6] We 1) sum importances for each feature type to compare the importance of zero-shot scores, per-position residue likelihoods or one-hot encoded amino acids as a whole, 2) we sum the feature importances for position-based features to find interesting positions 3) we visualize feature importances for all positions-based features, and 4) we find features that the models agree on.

We find that for Random Forests and Lassos C and D—which incorporate the zero-shot scores—the most important feature is the global zero-shot score ZS4 (ESMC+ESM-IF), with above 0.4 in R^2 importance in both types of models. Importance of the very similar ZS3 (same 3 out of 4 components) is negligible in comparison, or zero in Lassos. For Lassos that is the expected result, since the ZS4 supersedes ZS3 while both are highly correlated. All other global features, such as log-likelihoods averaged over CDRs (see Section 2.3.2), have negligible importance (importance below 0.001, not shown), even lower than the top 30 positions shown in Figure 3.7.

[6] We skip this analysis for the Gradient Boosting Tree models, as it is computationally expensive, and they achieve similar performace as Random Forests.

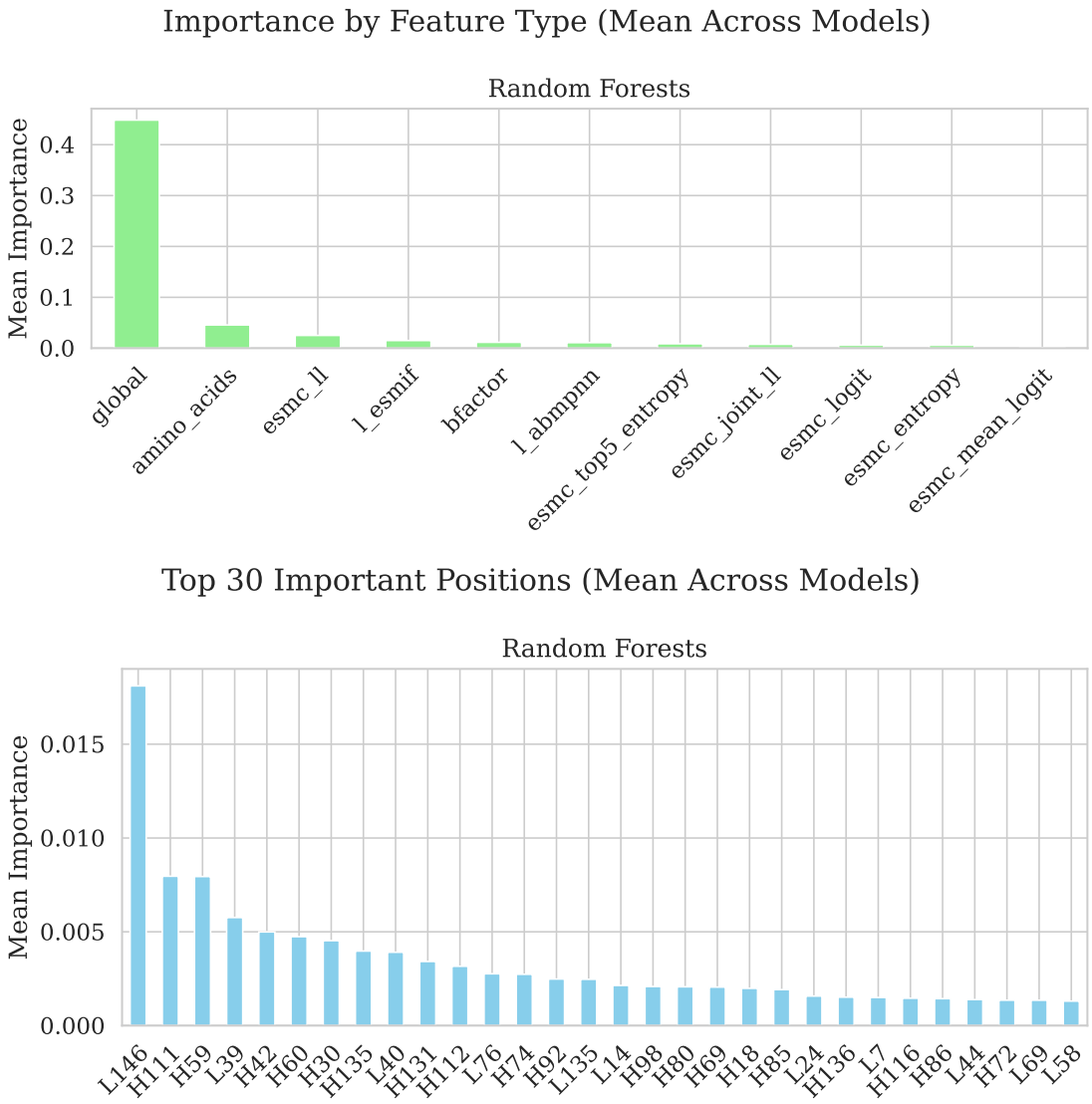

Figure 3.7 Feature importance for Random Forest models. Importance is defined as the average drop in holdout R^2 across 10 permutations when a feature is randomly shuffled across samples. The top plot shows importance aggregated by feature type: for each model, importances are summed across all positions (e.g. all amino acid identity features), and the result is then averaged across models that include the given feature type. Global features—including ZS4 and ZS3 scores—are grouped into a single bar. The bottom plot shows importance aggregated by position (AHo numbering): for each model, importances are summed across all feature types at a given position (e.g. all features associated with L146), and then averaged across models that include that position. Feature types include global descriptors (e.g. ZS4 and ZS3), one-hot encoded amino acids, `esmc_ll` (log-likelihood of the native residue from the ESMC model), `l_esmif` (likelihood from thermostability-fine-tuned ESM-IF—ProteinDPO), B-factors from predicted structures, and others described in Section 2.3.2. The mean importance is computed across only those models where the given feature or position was explicitly included, ensuring comparability across models with varying input configurations. Notably, global features dominate the overall importance, while position L146—associated with light-chain type—stands out among individual positions. We show the top 30 positions out of 279, of which 150 have mean importance above zero.

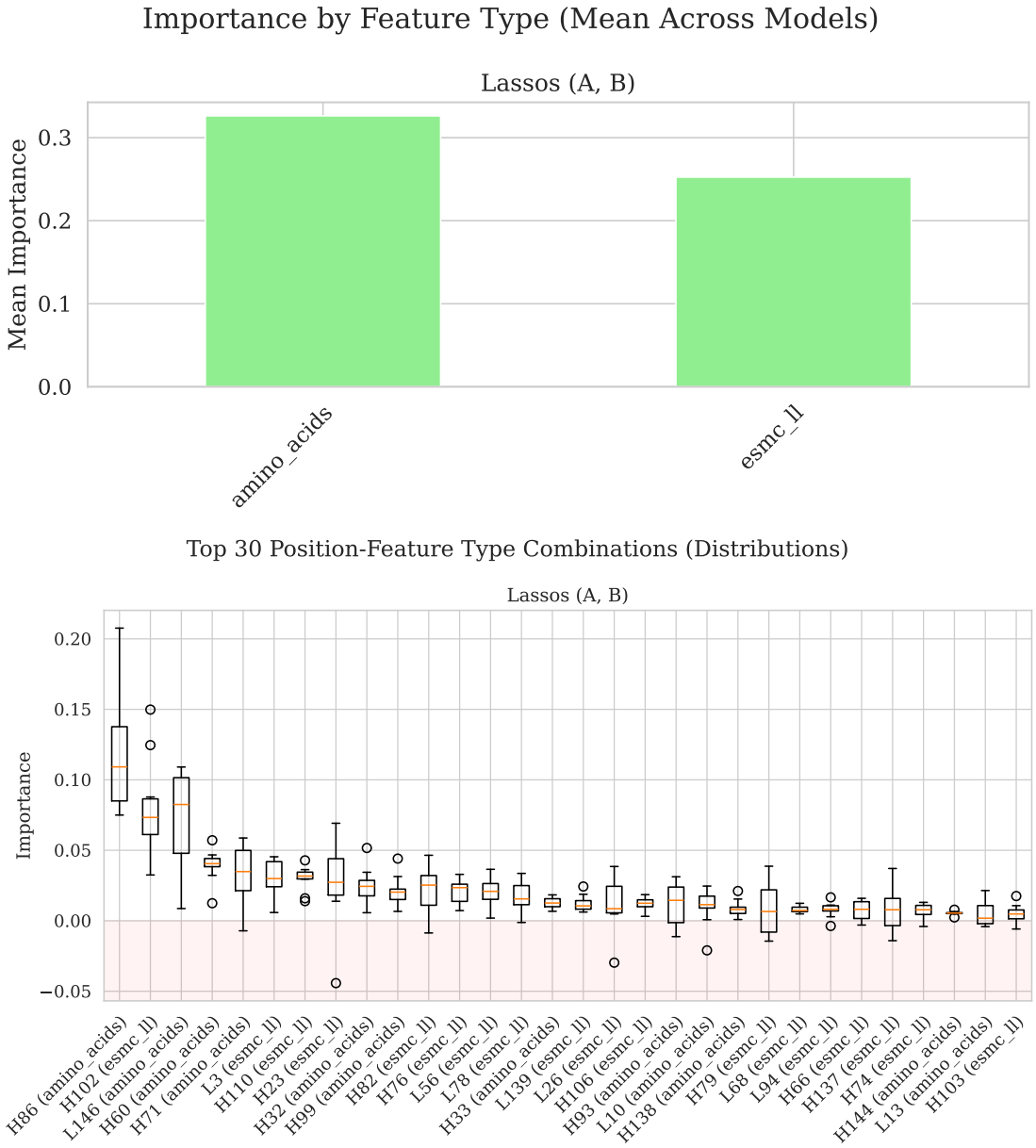

Figure 3.8 Feature importance for models using only positional features. Shown are two Lasso models (A and B) trained without global features, using either one-hot encoded amino acids or ESMC-derived log-likelihood scores (`esmc_ll`). Top plot: feature importance aggregated by feature type, based on permutation drop in holdout R^2. Bottom plot: top 30 most important position–feature combinations (AHo numbering). Each box shows the distribution of importance values obtained from 10 permutations of the corresponding feature in its respective model—values are not averaged across models, as each model uses a distinct feature set. Of the 213 total positions, 49 show nonzero importance. Amino acid identity features contribute most strongly overall, aligning with the better performance of the amino acid model. The shaded area below zero highlights no positive effect on prediction accuracy observed.

Random Forest models rely on a total of 150 positions with nonzero importance, whereas the position-only Lasso models (A and B)—using either ESMC log-likelihood or amino acid identity features—depend on only 45 positions combined.

We find that the most important position in Random Forest models is L146 (AHo numbering). This position is also the second resp. the third most important position in Lassos (C, D) resp. (A, B). L146 is located at the end of the variable region, before the first constant domain, and is dictated by the light chain type—kappa or lambda (Section 2.3.5).

Finally, we select top 50 positions from Random Forests and top 50 from non-global Lassos (A,B) and arrive at the intersection of 15 positions (Supplementary Table 6). If additionally intersected with the 10 global Lassos positions, we get two positions—L146 and H86 (mean importance in the three types: 0.022 and 0.021).

We visualize the 15 consensus positions on the certolizumab Fab structure (PDB ID: 5wuv; J. U. Lee et al., 2017, Figure 3.9), and additionally show the top 30 positions from Lasso models A and B (Supplementary Table 7)—relying solely on positional features—in Figure 3.10. Key positions are discussed in the context of this structure.

We analyzed the structural distribution of important residues identified by our models, specifically evaluating if these residues are located at the VH–VL interface as T. Wang and Duan (2011) found in a computational study on scFv that during thermal unfolding the interface contacts are the first ones affected. [7] We use the certolizumab Fab structure (PDB ID: 5wuv J. U. Lee et al. (2017)) in our analyses. Among interface residues, we found positions L105, L139, and H137 to participate in water-mediated interactions bridging the heavy and light chains, however these positions are around ten times less important than the most important positions. Such water bridges have previously been identified as crucial factors in stabilizing polypeptide conformations in aqueous environments (Petukhov et al., 1999). Position L139—buried at the VH–VL interface—although conserved and invariant in our dataset (phenylalanine), still shows importance via the ESMC log-likelihood feature. This reflects an interesting shift from modeling residue variability to evaluating the compatibility of the surrounding sequence context with an invariant residue.

Residue H60 is particularly notable, being buried and hydrogen-bonded with residues H31 (CDR-H1) and H34 in the certolizumab structure. The proximity and hydrogen bonding suggest that H60 structurally stabilizes the CDR-H1 loop. Residue H60 also exhibits high sequence variability across our dataset.

Overall, the most significant positions consistently identified across models are H86 (H84 in IMGT) and L146, the former being solvent exposed, and distant from the VH–VL interface, and the latter being crucial for determining the light-chain type (kappa in the shown structure).

[7] With full-length antibodies, interfaces with constant domains also can play a part (Tadokoro et al., 2024).

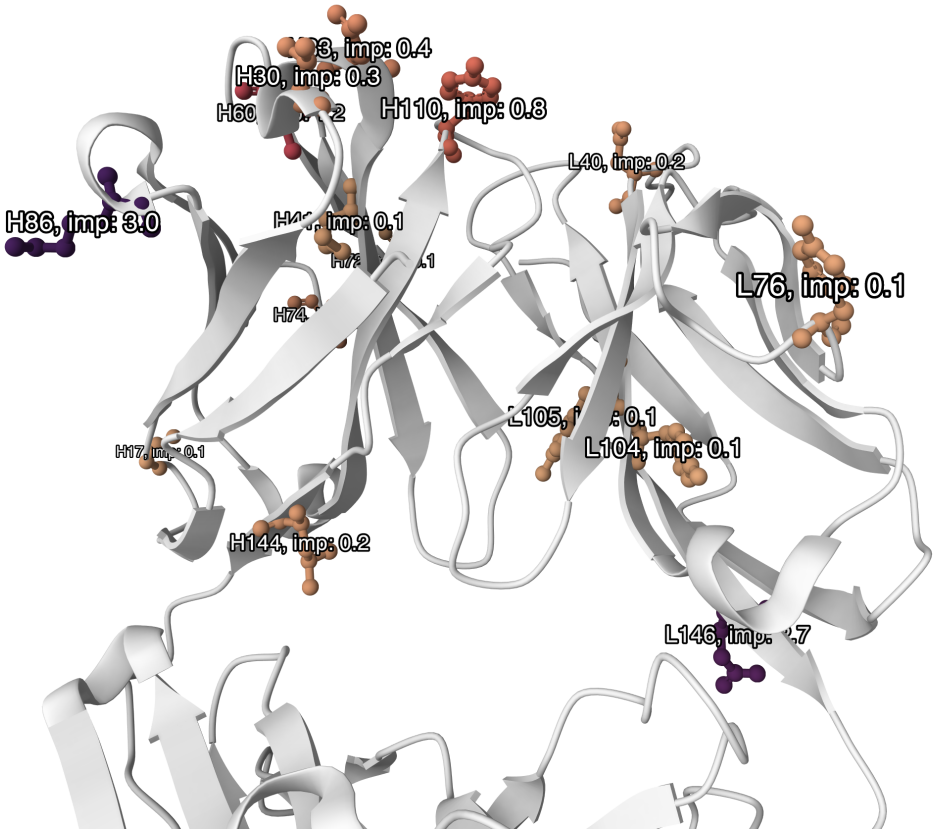

Figure 3.9 Positions identified by consensus of the top 50 important positions from the Random Forest models and from the non-global Lasso A and Lasso B models, mapped onto a certolizumab Fab structure (PDB ID: 5wuv; J. U. Lee et al. (2017)). On the left VH domain, on the right VL, constant domains CH1 and CL at the bottom. Importances (scaled in figure by 100) highlight the influence of these positions on antibody thermostability prediction. The most important residues are H86 (importance 0.030) and L146 (0.027), with L146 determining the antibody light-chain type (kappa in this structure). Residue H60 (0.012), a buried threonine, forms stabilizing hydrogen bonds with backbone oxygen of residue H34 (distance between oxygens: 2.6 Å) and possibly also H31 (3.3 Å), thus supporting the conformation of CDR-H1. We found a single position on the VH–VL interface, L105 (0.001), interacting indirectly with the VH chain via water bridges, however it displays rather low importance.

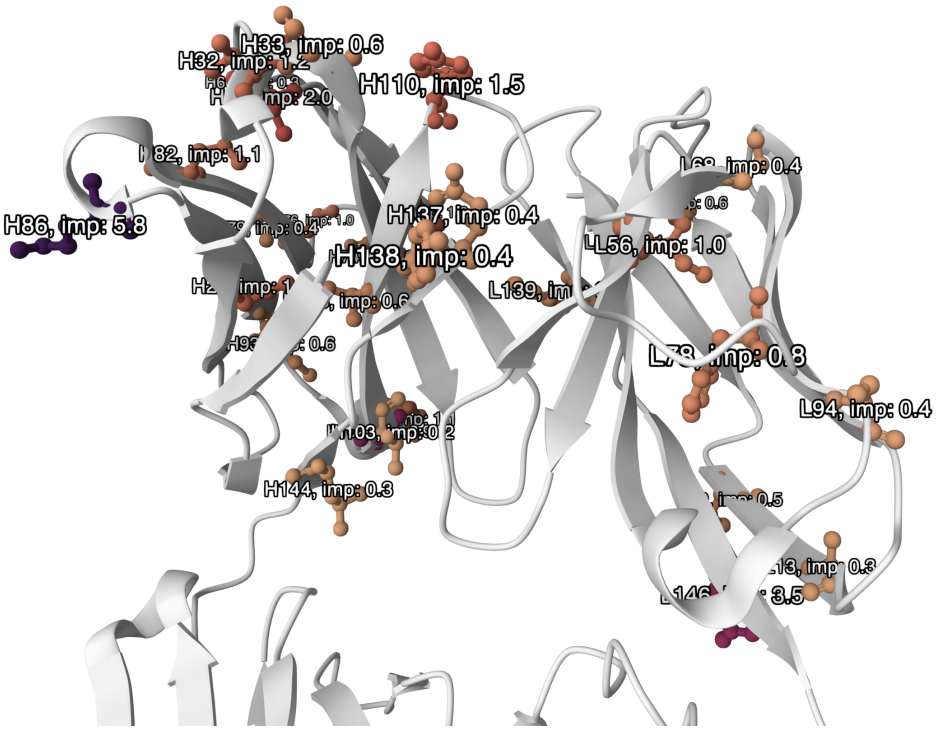

Figure 3.10 Top 30 positions identified by the Lasso (A, B) models trained exclusively on positional features (amino acid identities and ESMC log-likelihood scores), visualized on the certolizumab Fab structure (PDB ID: 5wuv; J. U. Lee et al. (2017)). As before, VH is shown on the left, VL on the right, with CH1 and CL at the bottom. Importances (scaled by 100 for visualization) reflect the drop in holdout R^2 upon feature permutation. The highest importance is observed at position H86 (mean importance 0.058), followed by buried residue H102 (0.039), L146 (0.035), and buried residue H60 (0.019). Notably, interface residues L139 (0.006) and H137 (0.004) participate in water-mediated bridging interactions between chains. Position L139 is invariant in our dataset (phenylalanine), and its detected importance likely results from ESMC features capturing the compatibility with surrounding residues rather than direct sequence variability. H102 (at the bottom of VH) is a conserved and a rather buried position forming a beta sheet interaction; again this is from ESMC and may reflect the surroundings, perhaps the beta sheet stability.

While preliminary literature comparisons were performed, future work could more thoroughly align the important positions found here with antibody engineering studies. Future work could also extend feature importance analyses to the zero-shot prediction setting.

For antibody engineering purposes—specifically, to prioritize positions for mutation—we consider the important positions identified for Lasso A (Figure 3.8) as the most informative. There are two reasons for this: first, Lasso A is based on amino acid features, meaning that the identified positions reflect how specific residues affect thermostability directly, rather than indirectly through neighboring structural context, as seen in Lasso B, where an invariant position (L139) was ranked as important. Second, Lasso A does not rely on global zero-shot scores, which already encode some thermostability information; using them could mask true position-specific effects.

3.2.2 Antibodies With Lambda Light Chains Are More Stable Than Those With Kappa Light Chains

We found that consistently in models, the terminal position of the variable region of the light chain (VL) is one of the most important features (Section 3.2.1). As this position is dictated by the light chain type (Section 2.3.5), we decided to investigate the effect of the light chain type on thermostability.

A two-sample, two-tailed Student's t-test assuming equal variances was used to compare the stability of antibodies with lambda versus kappa light chains. We chose a two-tailed test as we did not assume beforehand if kappa or lambda was more stable. The results showed a significant difference between the groups ($t = 3.77$, $df = 477$, $p = 1.84 \times 10^{-4}$). The mean difference was 2.08 °C corresponding to Cohen's d of 0.41, with group variances of and 25.9 (lambda) 23.1 (kappa). The sample sizes were $n_\lambda = 372$ and $n_\kappa = 107$. Therefore, lambda light chains were significantly more stable than kappa light chains. The distributions are shown in Figure 3.11.

It is generally accepted that the *kappa* light chain is more developable (Raybould et al., 2024). On the other hand, we show that thermostability, an important aspect of developability, is significantly higher in *lambda* chains, although with a rather moderate effect size. Here, we do not attempt to explain the mechanism, but it is a relevant question for future work.

Although about 35% of human antibodies use lambda light chains, only about 10% of clinical-stage therapeutic antibodies (as of 2019) are lambda-based (Raybould et al., 2024). This under-representation may stem from both discovery biases, such as mouse-derived screening libraries, which have kappa:lambda ratios as high as 20:1, and developability concerns. Lambda CDR3 tend to be longer and more hydrophobic than kappa, increasing the risk of aggregation and potentially contributing to lower preclinical success rates (Raybould et al., 2024), and lambda and kappa CDR1 and CDR3

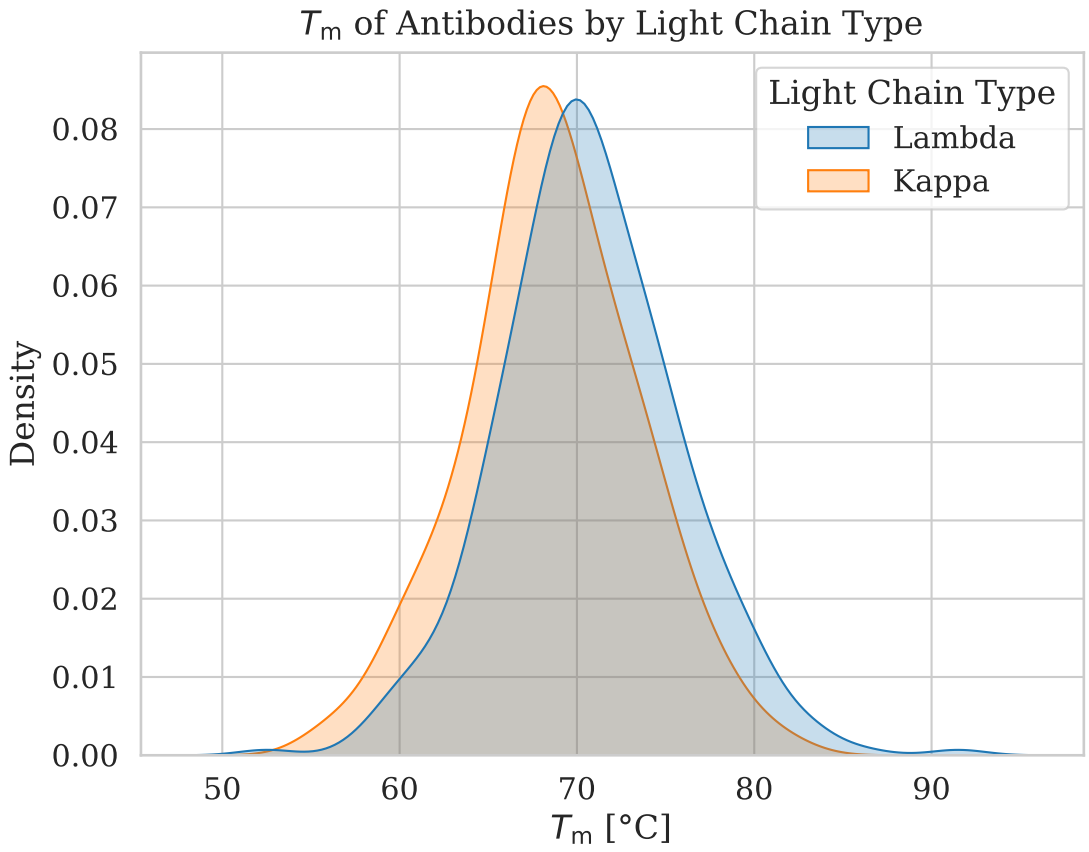

Figure 3.11 Distribution of antibody thermostability by light chain type. The mean T_m of lambda light chains (70.7 °C) is significantly higher than that of kappa light chains (68.6 °C, $p = 1.84 \times 10^{-4}$). All 483 antibodies from the AbProp dataset (Widatalla, Rollins, et al., 2023) were used, with the exception of 4 that were not classified to either kappa (107) or lambda (372) light chain type.

differ in the propensity for secondary structure composition (Townsend et al., 2016). However, lambda antibodies are increasingly found to preferentially engage a growing number of epitopes and, although they carry a higher average developability risk than kappa antibodies, a sizeable proportion exhibit lower-risk profiles, making them more tractable therapeutic candidates (Raybould et al., 2024).

3.2.3 Comparison of the Zero-Shot and Supervised Approaches

In this section, we compare zero-shot and supervised approaches to antibody thermostability prediction on the AbProp dataset. The main comparison between supervised and zero-shot methods is carried out on the holdout set. In addition, we use the full dataset to compare our zero-shot models with previously published zero-shot approaches. We include statistical tests to assess the significance of observed differences.

We compare our best-performing zero-shot score, ZS4 (ESMC+ESM-IF) score, with our best-performing supervised model—ensemble of tree-based predictors (Model N above)—on the AbProp holdout set, unseen by both models. As detailed later, the supervised ensemble achieves a Spearman correlation coefficient (SCC) of 0.69, with a 95% confidence interval of [0.56, 0.78] based on 10,000 bootstrap resamples. In comparison, ZS4 score reaches an SCC of 0.57, with a 95% confidence interval of [0.39, 0.70]. A permutation test yields a p-value of 0.010 for the hypothesis that ZS4 score performs equally or better than the supervised method, providing statistical evidence for the superiority of the tree-based ensemble (rejected at the 0.05 significance level).

Since published results for ProteinDPO and ESM-IF were only available in approximate form (e.g. as bar plots) and reported only for the full AbProp dataset, we reproduced these models using publicly available weights and our standardized evaluation workflow (see Section 2.3.6). This enabled direct and fair comparison on the holdout set, which had not been separately evaluated in the original publication (Widatalla, Rafailov, et al., 2024).

As discussed in Section 3.1.3, the ZS4 score was constructed using a portion of the AbProp dataset for model selection, making the holdout set the only unbiased estimate of its zero-shot performance. On the holdout set, ZS4 significantly outperforms the previously best-performing zero-shot method reported in the literature, ProteinDPO. Our reproduced version of ProteinDPO reaches an SCC of 0.24 on the holdout set, while ZS4 reaches 0.57. A permutation test for the hypothesis that both models perform equally yields a p-value = 0.004, indicating a highly significant difference. We also report performance on the full dataset: reproduced ProteinDPO (0.30), ZS4 (0.49); the permutation test on the full set gives a p-value of 2×10^{-5}; this result is reported for completeness, as the training portion of the AbProp dataset was used for model selection. The permutation testing protocol is described in Section 2.2.4.

Performance across both the full and holdout sets is summarized in Tables 3.3

Method	AbProp (Full) Spearman
ESM-IF (Lit.)	0.25
ESM-IF (Rep.)	0.30 ± 0.04
ProteinDPO (Lit.)	0.35
ProteinDPO (Rep.)	0.30 ± 0.04
SaprotHub Thermo	0.26 ± 0.04
Our ZS3 (3xESMC)	0.48 ± 0.04
Our ZS4 (ESMC+ESM-IF)	**0.49 ± 0.04**

Table 3.3 Comparison of zero-shot prediction performance on the full AbProp dataset (Spearman correlation). The ± values represent the standard deviation of the Spearman correlation computed over 10,000 bootstrap resamples. Rows marked as *(Lit.)* contain performance values reported in the literature (approximated from bar plots), while rows marked as *(Rep.)* show results we reproduced using the published weights and the workflow applied in this thesis (see Section 2.3.6). ProteinDPO was fine-tuned on the Mega-Scale thermostability dataset, and SaprotHub Thermo is the SaProt 650M model fine-tuned on a subset of the Meltome dataset and applied by us to antibody thermostability. Our methods (composite scores ZS3 and ZS4) achieve the best performance, though the scores they are composed of were selected using part of the AbProp dataset. An unbiased evaluation on a holdout set is presented in Table 3.4.

and 3.4. We include the current state-of-the-art model for absolute thermostability prediction, SaProt 650M fine-tuned on a portion of Meltome dataset, referred to as SaprotHub Thermo in this work (Sections 1.4 and 2.3.6). We observe fluctuations in performance between the two subsets; for example, SaprotHub Thermo reaches 0.26 on the full set and 0.40 on the holdout, while ESM-IF scores 0.30 on the full set and 0.23 on the holdout. Our zero-shot models (ZS3 and ZS4) remain consistently strong across both settings. This variation does not necessarily indicate overfitting or differences in generalization ability; rather, it likely reflects sampling variability between the full dataset (483 antibodies) and the smaller holdout subset (73 antibodies). This interpretation is supported by the width of the confidence intervals estimated via bootstrapping and the standard deviations of the bootstrapped SCC distributions reported in the tables.

The reproduced results for ProteinDPO and ESM-IF differ slightly from those reported in the original publication (Table 3.3). While we followed the general evaluation strategy described in the paper and used our standardized workflow, several implementation details were not publicly specified. For instance, although both evaluations rely on average log-likelihood, it is unclear whether the original authors used a linker between the heavy and light chains, or how their sequence formatting was handled. Additionally, they used IgFold (Ruffolo, Chu, et al., 2023) to generate structures, whereas we used ABodyBuilder2, which may affect structure-based scoring. For ProteinDPO

Method	AbProp (Holdout) Spearman
ESM-IF (Rep.)	0.23 ± 0.11
ProteinDPO (Rep.)	0.24 ± 0.11
SaprotHub Thermo	0.40 ± 0.10
Our ZS3 (3xESMC)	**0.57 ± 0.08**
Our ZS4 (ESMC+ESM-IF)	**0.57 ± 0.08**
OHE baseline	0.51 ± 0.10
GAT AbProp (Lit.)	0.62
Our Tree Ensemble	**0.69 ± 0.06**

Table 3.4 Performance comparison on the AbProp holdout set (Spearman correlation) between supervised and zero-shot approaches, including both our models and previously published state-of-the-art methods. The upper section shows zero-shot models, with *(Rep.)* indicating that the results were reproduced using the published weights and the workflow applied in this thesis. This enables direct assessment of holdout performance, which had not been evaluated separately in previous work, where only performance on the full AbProp dataset was reported. Our composite models (ZS3 and ZS4) achieve the highest zero-shot performance. The lower section presents supervised models. *(Lit.)* indicates that performance values were taken directly from the original publication. We include an OHE-based baseline (referred to earlier as Lasso A), the original Graph Attention Network (GAT) model—the best-performing model in the AbProp original paper, and our tree ensemble, which achieves the best overall performance. For this supervised method, the difference is not large, and given the small size of the holdout dataset, it may not be statistically significant (we did not test for significance, as we did not reproduce the GAT model). The zero-shot ZS4 score is significantly better than the reproduced ProteinDPO, which was the previously best-performing zero-shot method reported on the AbProp dataset (permutation test, p-value 0.004; see main text for details).

in particular, only weights for the `paired` version of the model (reaching a Spearman correlation of approximately 0.32) were publicly available; the `weighted` version that reached 0.35 in the original report was not released.

Conclusion

In this thesis, we addressed the problem of predicting antibody thermostability using modern machine learning approaches. Specifically, we explored both zero-shot inference from pretrained protein language models and supervised learning models trained on a public dataset of antibody melting temperatures. The main goal of this thesis was to come up with and evaluate approaches for predicting antibody thermostability, and to identify sequence or structural positions that most influence it. In addition, we examined whether language models pretrained on general proteins differ from those pretrained on antibody-specific data in their predictive capabilities, and whether incorporating predicted protein structure can improve performance.

We proposed a composite zero-shot score that combines outputs from two pretrained models, ESM-IF and ESM Cambrian 300M, and demonstrated that it achieves state-of-the-art results on the AbProp dataset, outperforming previously published methods. We further evaluated this method on the ProteinGym stability benchmark to assess generalization beyond antibodies, and compared the results against antibody-specific models. In addition, we developed a supervised ensemble model based on decision trees that achieves superior performance on the same dataset. The following paragraphs briefly summarize our findings and outline how the models developed here can be applied in antibody engineering.

We found that in the zero-shot setting, models incorporating structural information (e.g. inverse folding models) tend to rank highly, as observed on the ProteinGym benchmark for general proteins. However, the best performance in our evaluations for antibody thermostability was achieved by a newly released sequence-only model, ESM Cambrian 300M (ESMC).

Our approach by combining four likelihood scores from two protein language models (PLMs) establishes the state-of-the-art on AbProp dataset, improving the previous best reported Spearman correlation from 0.35 (ProteinDPO; Widatalla, Rafailov, et al., 2024) to 0.49 in the zero-shot setting, using an ensemble score derived from ESMC protein sequence-only language model and a structure-informed language model ESM-IF. An ensemble score that uses logits derived from ESMC model only ZS3 score achieves similar performance (0.48). However, since the selection of individual scores contributing to the ensemble was performed on the training portion of the dataset, this

result may be biased. To provide an unbiased estimate, we evaluated our models on a holdout subset that was not used during score construction or model selection. On this holdout set, our best method (ZS4 score) achieves a Spearman correlation of 0.57 and significantly outperforms the reproduced ProteinDPO model.

Although developed for antibody thermostability, our composite ZS3 (3xESMC) score method generalizes well to stability prediction tasks on general, non-antibody proteins. On the ProteinGym benchmark—which differs substantially from our antibody dataset in protein length, domain count, and mutation depth—ZS3 score achieves the highest performance among all sequence-only models for double mutant prediction. For single mutants, however, it is outperformed by the native log-likelihood score from the ESMC model. Notably, ProteinGym evaluates mutations at a much lower depth (1–2 mutations), whereas sequences in the AbProp T_{m} dataset differ by an average of 106 mutations. The relative strength of ZS3 score appears to lie in its robustness to multiple mutations—an especially important property in our setting, where mutational distances are substantially higher.

We find that antibody-specific language models do not outperform general PLMs in predicting antibody thermostability—a somewhat counterintuitive result given their domain-specific focus. A likely explanation lies in the nature of antibody sequence generation: the process involves extreme randomness through V(D)J recombination and somatic hypermutation, which is optimized primarily for antigen binding rather than thermostability. In fact, somatic hypermutation has been shown to sometimes compromise stability. As a result, antibody datasets may exhibit a wide and unconstrained distribution of thermostability, making it difficult for models trained solely on this data to learn predictive stability patterns. In contrast, general PLMs are trained on evolutionarily selected proteins, where stability is more likely to be conserved, potentially making them more informative for thermostability prediction. Structure-informed models like AbMPNN and AntiFold do not show a significant drop in performance, possibly because they retain useful representations from pretraining on general protein datasets.

With supervised models we achieve state-of-the-art performance for antibody thermostability prediction. With our ensemble of tree-based model, we achieve 0.69 Spearman's correlation coefficient on the AbProp holdout set, an increase over 0.62 in the previous work. We analyze the most important features and discover that lambda light chains are significantly more thermostable than kappa light chains, with a moderate effect size. Our supevised model can be directly applied to prioritize lead sequences for drug development by predicted thermostability. For introducing stabilizing mutations, however, we recommend using the zero-shot models with standard likelihood scoring, as 1) our supervised model uses a variant of a zero-shot score developed for absolute, not relative thermostability prediction of close variants, and it smooths the potential effect of mutations, 2) zero-shot models can efficiently score the effects of all possible

point mutations in a single pass, providing an in silico approximation of site-saturation mutagenesis.

Bibliography

Bailly, Marc et al. (Apr. 5, 2020). "Predicting Antibody Developability Profiles Through Early Stage Discovery Screening". In: *mAbs* 12.1, p. 1743053. DOI: 10.1080/19420862.2020.1743053. PMID: 32249670. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7153844/ (visited on 11/02/2024).

Jain, Tushar et al. (Jan. 31, 2017). "Biophysical Properties of the Clinical-Stage Antibody Landscape". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.5, pp. 944–949. ISSN: 1091-6490. DOI: 10.1073/pnas.1616408114. PMID: 28096333.

Shehata, Laila et al. (Sept. 24, 2019). "Affinity Maturation Enhances Antibody Specificity but Compromises Conformational Stability". In: *Cell Reports* 28.13, 3300–3308.e4. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2019.08.056. URL: https://www.sciencedirect.com/science/article/pii/S2211124719311040 (visited on 11/05/2024).

Yu, Haoran et al. (Feb. 1, 2017). "Two Strategies to Engineer Flexible Loops for Improved Enzyme Thermostability". In: *Scientific Reports* 7.1, p. 41212. ISSN: 2045-2322. DOI: 10.1038/srep41212. URL: https://www.nature.com/articles/srep41212 (visited on 11/05/2024).

Daniel, Roy M. and Michael J. Danson (Sept. 2, 2013). "Temperature and the Catalytic Activity of Enzymes: A Fresh Understanding". In: *FEBS Letters*. A Century of Michaelis - Menten Kinetics 587.17, pp. 2738–2743. ISSN: 0014-5793. DOI: 10.1016/j.febslet.2013.06.027. URL: https://www.sciencedirect.com/science/article/pii/S0014579313004857 (visited on 11/05/2024).

Huffman, Mark A. et al. (Dec. 6, 2019). "Design of an in Vitro Biocatalytic Cascade for the Manufacture of Islatravir". In: *Science* 366.6470, pp. 1255–1259. DOI: 10.1126/science.aay8484. URL: https://www.science.org/doi/10.1126/science.aay8484 (visited on 11/05/2024).

Notin, Pascal et al. (Dec. 8, 2023). "ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction". In: *bioRxiv: The Preprint Server for Biology*, p. 2023.12.07.570727. DOI: 10.1101/2023.12.07.570727. PMID: 38106144.

Widatalla, Talal, Rafael Rafailov, et al. (May 21, 2024). *Aligning Protein Generative Models with Experimental Fitness via Direct Preference Optimization*. DOI: 10.1101/2024.

05.20.595026. URL: https://www.biorxiv.org/content/10.1101/2024.05.20.595026v1 (visited on 06/03/2024). Pre-published.

ESM Team, _ (Apr. 12, 2024). *ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning*. URL: https://www.evolutionaryscale.ai/blog/esm-cambrian (visited on 03/31/2025).

Hsu, Chloe et al. (Apr. 10, 2022). *Learning Inverse Folding from Millions of Predicted Structures*. DOI: 10.1101/2022.04.10.487779. URL: https://www.biorxiv.org/content/10.1101/2022.04.10.487779v1 (visited on 03/31/2025). Pre-published.

Widatalla, Talal, Zachary Rollins, et al. (2023). "AbPROP: Language and Graph Deep Learning for Antibody Property Prediction". In: ICML. CompBio.

Murphy, Kenneth M. and Casey Weaver (2017). *Janeway's Immunobiology*. 9th ed. New York : Garland Science. ISBN: 978-0-8153-4505-3. URL: http://lib.ugent.be/catalog/rug01:002339990.

Lefranc, Marie-Paule (Feb. 5, 2014). "Immunoglobulin and T Cell Receptor Genes: IMGT® and the Birth and Rise of Immunoinformatics". In: *Frontiers in Immunology* 5, p. 22. ISSN: 1664-3224. DOI: 10.3389/fimmu.2014.00022. PMID: 24600447. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3913909/ (visited on 04/16/2025).

Vadnais, Melissa L. et al. (2017). "Antibodies from Other Species". In: *Protein Therapeutics*. John Wiley & Sons, Ltd, pp. 85–112. ISBN: 978-3-527-69912-4. DOI: 10.1002/9783527699124.ch4. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527699124.ch4 (visited on 04/16/2025).

Kang, Seung Hyun and Chang-Han Lee (June 2021). "Development of Therapeutic Antibodies and Modulating the Characteristics of Therapeutic Antibodies to Maximize the Therapeutic Efficacy". In: *Biotechnology and Bioprocess Engineering* 26.3, pp. 295–311. ISSN: 1226-8372, 1976-3816. DOI: 10.1007/s12257-020-0181-8. URL: https://link.springer.com/10.1007/s12257-020-0181-8 (visited on 04/18/2025).

Vidarsson, Gestur et al. (Oct. 20, 2014). "IgG Subclasses and Allotypes: From Structure to Effector Functions". In: *Frontiers in Immunology* 5. ISSN: 1664-3224. DOI: 10.3389/fimmu.2014.00520. URL: https://www.frontiersin.orghttps://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2014.00520/full (visited on 04/16/2025).

Boyd, Scott D. and Shilpa A. Joshi (Sept. 19, 2014). "High-Throughput DNA Sequencing Analysis of Antibody Repertoires". In: *Microbiology Spectrum* 2.5. Ed. by James E. Crowe Jr. et al., p. 2.5.23. ISSN: 2165-0497. DOI: 10.1128/microbiolspec.AID-0017-2014. URL: https://journals.asm.org/doi/10.1128/microbiolspec.AID-0017-2014 (visited on 04/16/2025).

Luning Prak, Eline T. et al. (Jan. 2011). "B Cell Receptor Editing in Tolerance and Autoimmunity". In: *Annals of the New York Academy of Sciences* 1217, pp. 96–121. ISSN: 0077-8923. DOI: 10.1111/j.1749-6632.2010.05877.x. PMID: 21251012. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3077556/ (visited on 04/17/2025).

Crescioli, Silvia et al. (Dec. 2025). "Antibodies to Watch in 2025". In: *mAbs* 17.1, p. 2443538. ISSN: 1942-0870. DOI: 10.1080/19420862.2024.2443538. PMID: 39711140.

Pierpont, Timothy M. et al. (June 4, 2018). "Past, Present, and Future of Rituximab—The World's First Oncology Monoclonal Antibody Therapy". In: *Frontiers in Oncology* 8, p. 163. ISSN: 2234-943X. DOI: 10.3389/fonc.2018.00163. PMID: 29915719. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5994406/ (visited on 04/18/2025).

Mihara, Masahiko et al. (Feb. 25, 2011). "Tocilizumab, a Humanized Anti-Interleukin-6 Receptor Antibody, for Treatment of Rheumatoid Arthritis". In: *Open Access Rheumatology : Research and Reviews* 3, pp. 19–29. ISSN: 1179-156X. DOI: 10.2147/OARRR.S17118. PMID: 27790001. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5074778/ (visited on 04/18/2025).

Sundar, Raghav et al. (Mar. 2015). "Nivolumab in NSCLC: Latest Evidence and Clinical Potential". In: *Therapeutic Advances in Medical Oncology* 7.2, pp. 85–96. ISSN: 1758-8340. DOI: 10.1177/1758834014567470. PMID: 25755681. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4346216/ (visited on 04/18/2025).

Köhler, G. and C. Milstein (Aug. 7, 1975). "Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity". In: *Nature* 256.5517, pp. 495–497. ISSN: 0028-0836. DOI: 10.1038/256495a0. PMID: 1172191.

Hoogenboom, Hennie R. (Sept. 2005). "Selecting and Screening Recombinant Antibody Libraries". In: *Nature Biotechnology* 23.9, pp. 1105–1116. ISSN: 1087-0156. DOI: 10.1038/nbt1126. PMID: 16151404.

Pedrioli, Alessandro and Annette Oxenius (Dec. 1, 2021). "Single B Cell Technologies for Monoclonal Antibody Discovery". In: *Trends in Immunology* 42.12, pp. 1143–1158. ISSN: 1471-4906, 1471-4981. DOI: 10.1016/j.it.2021.10.008. PMID: 34743921. URL: https://www.cell.com/trends/immunology/abstract/S1471-4906(21)00213-1 (visited on 04/18/2025).

Tang, Yu et al. (Nov. 11, 2021). "Impact of IgG Subclass on Molecular Properties of Monoclonal Antibodies". In: *mAbs* 13.1, p. 1993768. ISSN: 1942-0862. DOI: 10.1080/19420862.2021.1993768. PMID: 34763607. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8726687/ (visited on 04/18/2025).

Ahmad, Zuhaida Asra et al. (2012). "scFv Antibody: Principles and Clinical Application". In: *Clinical and Developmental Immunology* 2012, p. 980250. ISSN: 1740-2522. DOI: 10.1155/2012/980250. PMID: 22474489. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3312285/ (visited on 04/17/2025).

Ghaderi, Hajarossadat et al. (Nov. 1, 2023). "Recombinant Antibody Fragment Therapeutics: Current Status and Future Prospects of scFv, Nanobody, and Mimotopes". In: *Journal of Drug Delivery Science and Technology* 89, p. 105009. ISSN: 1773-2247. DOI: 10.1016/j.jddst.2023.105009. URL: https://www.sciencedirect.com/science/article/pii/S1773224723008614 (visited on 04/17/2025).

Harris, Lisa J. et al. (Feb. 1, 1997). "Refined Structure of an Intact IgG2a Monoclonal Antibody," in: *Biochemistry* 36.7, pp. 1581–1597. ISSN: 0006-2960. DOI: 10.1021/bi962514+. URL: https://doi.org/10.1021/bi962514+ (visited on 04/17/2025).

Chiu, Mark L. et al. (Dec. 3, 2019). "Antibody Structure and Function: The Basis for Engineering Therapeutics". In: *Antibodies* 8.4, p. 55. ISSN: 2073-4468. DOI: 10.3390/antib8040055. PMID: 31816964. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6963682/ (visited on 03/24/2024).

Teplyakov, Alexey et al. (Aug. 17, 2016). "Structural Diversity in a Human Antibody Germline Library". In: *mAbs* 8.6, pp. 1045–1063. ISSN: 1942-0862. DOI: 10.1080/19420862.2016.1190060. PMID: 27210805. URL: https://doi.org/10.1080/19420862.2016.1190060 (visited on 04/17/2025).

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. ISSN: 1741-0134. DOI: 10.1093/protein/gzt020. PMID: 23708320.

Wang, Ting and Yong Duan (Sept. 2011). "Probing the Stability-Limiting Regions of an Antibody Single-Chain Variable Fragment: A Molecular Dynamics Simulation Study". In: *Protein Engineering, Design and Selection* 24.9, pp. 649–657. ISSN: 1741-0126. DOI: 10.1093/protein/gzr029. PMID: 21729946. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3160207/ (visited on 07/15/2024).

Lefranc, Marie-Paule, Christelle Pommié, et al. (Jan. 1, 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. ISSN: 0145-305X. DOI: 10.1016/S0145-305X(02)00039-3. URL: https://www.sciencedirect.com/science/article/pii/S0145305X02000393 (visited on 04/17/2025).

Dunbar, James and Charlotte M. Deane (Jan. 15, 2016). "ANARCI: Antigen Receptor Numbering and Receptor Classification". In: *Bioinformatics* 32.2, pp. 298–300. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv552. URL: https://doi.org/10.1093/bioinformatics/btv552 (visited on 04/04/2025).

Honegger, A. and A. Plückthun (June 8, 2001). "Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool". In: *Journal of Molecular Biology* 309.3, pp. 657–670. ISSN: 0022-2836. DOI: 10.1006/jmbi.2001.4662. PMID: 11397087.

Dondelinger, Mathieu et al. (Oct. 16, 2018). "Understanding the Significance and Implications of Antibody Numbering and Antigen-Binding Surface/Residue Definition".

In: *Frontiers in Immunology* 9, p. 2278. ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02278. PMID: 30386328. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6198058/ (visited on 03/22/2025).

Lavinder, Jason J. et al. (Mar. 25, 2009). "High-Throughput Thermal Scanning: A General, Rapid Dye-Binding Thermal Shift Screen for Protein Engineering". In: *Journal of the American Chemical Society* 131.11, pp. 3794–3795. ISSN: 0002-7863. DOI: 10.1021/ja8049063. PMID: 19292479. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2701314/ (visited on 04/18/2025).

Lang, Brian E. and Kenneth D. Cole (May 2017). "Differential Scanning Calorimetry and Fluorimetry Measurements of Monoclonal Antibodies and Reference Proteins: Effect of Scanning Rate and Dye Selection". In: *Biotechnology Progress* 33.3, pp. 677–686. ISSN: 1520-6033. DOI: 10.1002/btpr.2464. PMID: 28371560.

Tsuboyama, Kotaro et al. (Aug. 2023). "Mega-Scale Experimental Analysis of Protein Folding Stability in Biology and Design". In: *Nature* 620.7973, pp. 434–444. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06328-6. URL: https://www.nature.com/articles/s41586-023-06328-6 (visited on 11/12/2024).

Greenfield, Norma J. (2006). "Using Circular Dichroism Collected as a Function of Temperature to Determine the Thermodynamics of Protein Unfolding and Binding Interactions". In: *Nature protocols* 1.6, pp. 2527–2535. ISSN: 1754-2189. DOI: 10.1038/nprot.2006.204. PMID: 17406506. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2752288/ (visited on 04/19/2025).

Watson, Joseph L. et al. (Aug. 2023). "De Novo Design of Protein Structure and Function with RFdiffusion". In: *Nature* 620.7976, pp. 1089–1100. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06415-8. URL: https://www.nature.com/articles/s41586-023-06415-8 (visited on 04/19/2025).

Savitski, Mikhail M. et al. (Oct. 3, 2014). "Tracking Cancer Drugs in Living Cells by Thermal Profiling of the Proteome". In: *Science* 346.6205, p. 1255784. DOI: 10.1126/science.1255784. URL: https://www.science.org/doi/10.1126/science.1255784 (visited on 11/07/2024).

Jarzab, Anna et al. (May 2020). "Meltome Atlas—Thermal Proteome Stability across the Tree of Life". In: *Nature Methods* 17.5, pp. 495–503. ISSN: 1548-7105. DOI: 10.1038/s41592-020-0801-4. URL: https://www.nature.com/articles/s41592-020-0801-4 (visited on 08/13/2024).

Sinitcyn, Pavel et al. (Dec. 2023). "Global Detection of Human Variants and Isoforms by Deep Proteome Sequencing". In: *Nature Biotechnology* 41.12, pp. 1776–1786. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01714-x. URL: https://www.nature.com/articles/s41587-023-01714-x (visited on 11/19/2024).

Devlin, Jacob et al. (May 24, 2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. DOI: 10.48550/arXiv.1810.04805. arXiv: 1810.

04805 [cs]. URL: http://arxiv.org/abs/1810.04805 (visited on 04/18/2025). Pre-published.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Maryam Fazel-Zarandi, et al. (July 2022). "Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction". In.

Elnaggar, Ahmed et al. (May 4, 2021). *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning*. DOI: 10.1101/2020.07.12.199554. URL: https://www.biorxiv.org/content/10.1101/2020.07.12.199554v3 (visited on 04/23/2024). Pre-published.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. (Mar. 17, 2023). "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model". In: *Science* 379.6637, pp. 1123–1130. DOI: 10.1126/science.ade2574. URL: https://www.science.org/doi/10.1126/science.ade2574 (visited on 04/18/2025).

Dauparas, J. et al. (June 4, 2022). *Robust Deep Learning Based Protein Sequence Design Using ProteinMPNN*. DOI: 10.1101/2022.06.03.494563. URL: https://www.biorxiv.org/content/10.1101/2022.06.03.494563v1 (visited on 03/31/2024). Pre-published.

The UniProt Consortium (Jan. 6, 2023). "UniProt: The Universal Protein Knowledgebase in 2023". In: *Nucleic Acids Research* 51.D1, pp. D523–D531. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1052. URL: https://doi.org/10.1093/nar/gkac1052 (visited on 04/19/2025).

Richardson, Lorna et al. (Jan. 6, 2023). "MGnify: The Microbiome Sequence Data Analysis Resource in 2023". In: *Nucleic Acids Research* 51.D1, pp. D753–D759. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1080. URL: https://doi.org/10.1093/nar/gkac1080 (visited on 04/19/2025).

Rives, Alexander et al. (Apr. 13, 2021). "Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences". In: *Proceedings of the National Academy of Sciences* 118.15, e2016239118. DOI: 10.1073/pnas.2016239118. URL: https://www.pnas.org/doi/10.1073/pnas.2016239118 (visited on 04/19/2025).

Nijkamp, Erik et al. (June 27, 2022). *ProGen2: Exploring the Boundaries of Protein Language Models*. DOI: 10.48550/arXiv.2206.13517. arXiv: 2206.13517 [cs]. URL: http://arxiv.org/abs/2206.13517 (visited on 04/04/2025). Pre-published.

Ruffolo, Jeffrey A., Jeffrey J. Gray, et al. (Dec. 14, 2021). *Deciphering Antibody Affinity Maturation with Language Models and Weakly Supervised Learning*. DOI: 10.48550/arXiv.2112.07782. arXiv: 2112.07782 [q-bio]. URL: http://arxiv.org/abs/2112.07782 (visited on 04/19/2025). Pre-published.

Kenlay, Henry et al. (Mar. 26, 2024). *Large Scale Paired Antibody Language Models*. DOI: 10.48550/arXiv.2403.17889. arXiv: 2403.17889 [q-bio]. URL: http://arxiv.org/abs/2403.17889 (visited on 04/04/2025). Pre-published.

Olsen, Tobias H., Iain H. Moal, et al. (Feb. 7, 2024). *Addressing the Antibody Germline Bias and Its Effect on Language Models for Improved Antibody Design*. DOI: 10.1101/2024.02.02.578678. URL: https://www.biorxiv.org/content/10.1101/2024.02.02.578678v1 (visited on 04/04/2025). Pre-published.

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (visited on 04/19/2025).

Bronstein, Michael M. et al. (May 2, 2021). *Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges*. DOI: 10.48550/arXiv.2104.13478. arXiv: 2104.13478 [cs]. URL: http://arxiv.org/abs/2104.13478 (visited on 04/19/2025). Pre-published.

Jing, Bowen et al. (May 16, 2021). *Learning from Protein Structure with Geometric Vector Perceptrons*. DOI: 10.48550/arXiv.2009.01411. arXiv: 2009.01411 [q-bio]. URL: http://arxiv.org/abs/2009.01411 (visited on 04/19/2025). Pre-published.

Dreyer, Frédéric A. et al. (Oct. 30, 2023). *Inverse Folding for Antibody Sequence Design Using Deep Learning*. DOI: 10.48550/arXiv.2310.19513. arXiv: 2310.19513 [cs, q-bio]. URL: http://arxiv.org/abs/2310.19513 (visited on 03/31/2024). Pre-published.

Høie, Magnus Haraldson et al. (May 6, 2024). *AntiFold: Improved Antibody Structure-Based Design Using Inverse Folding*. DOI: 10.48550/arXiv.2405.03370. arXiv: 2405.03370 [q-bio]. URL: http://arxiv.org/abs/2405.03370 (visited on 04/04/2025). Pre-published.

Su, Jin, Chenchen Han, et al. (Oct. 13, 2023). "SaProt: Protein Language Modeling with Structure-aware Vocabulary". In: The Twelfth International Conference on Learning Representations. URL: https://openreview.net/forum?id=6MRm3G4NiU (visited on 04/19/2025).

Van Kempen, Michel et al. (Feb. 2024). "Fast and Accurate Protein Structure Search with Foldseek". In: *Nature Biotechnology* 42.2, pp. 243–246. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01773-0. URL: https://www.nature.com/articles/s41587-023-01773-0 (visited on 04/19/2025).

Hayes, Thomas et al. (Dec. 31, 2024). *Simulating 500 Million Years of Evolution with a Language Model*. DOI: 10.1101/2024.07.01.600583. URL: https://www.biorxiv.org/content/10.1101/2024.07.01.600583v2 (visited on 04/04/2025). Pre-published.

Engqvist, Martin K. M. (Nov. 6, 2018). "Correlating Enzyme Annotations with a Large Set of Microbial Growth Temperatures Reveals Metabolic Adaptations to Growth

at Diverse Temperatures". In: *BMC Microbiology* 18, p. 177. ISSN: 1471-2180. DOI: 10.1186/s12866-018-1320-7. PMID: 30400856. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219164/ (visited on 04/20/2025).

Chang, Antje et al. (Jan. 8, 2021). "BRENDA, the ELIXIR Core Data Resource in 2021: New Developments and Updates". In: *Nucleic Acids Research* 49.D1, pp. D498–D508. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1025. URL: https://doi.org/10.1093/nar/gkaa1025 (visited on 04/20/2025).

Li, Gang et al. (June 21, 2019). "Machine Learning Applied to Predicting Microorganism Growth Temperatures and Enzyme Catalytic Optima". In: *ACS Synthetic Biology* 8.6, pp. 1411–1420. DOI: 10.1021/acssynbio.9b00099. URL: https://doi.org/10.1021/acssynbio.9b00099 (visited on 05/09/2024).

Pudžiuvelytė, Ieva et al. (Mar. 28, 2023). *TemStaPro: Protein Thermostability Prediction Using Sequence Representations from Protein Language Models*. DOI: 10.1101/2023.03.27.534365. URL: https://www.biorxiv.org/content/10.1101/2023.03.27.534365v1 (visited on 03/19/2024). Pre-published.

Zhao, Jianjun et al. (Jan. 2023). "DeepTP: A Deep Learning Model for Thermophilic Protein Prediction". In: *International Journal of Molecular Sciences* 24.3 (3), p. 2217. ISSN: 1422-0067. DOI: 10.3390/ijms24032217. URL: https://www.mdpi.com/1422-0067/24/3/2217 (visited on 04/20/2025).

Dallago, Christian et al. (Nov. 11, 2021). *FLIP: Benchmark Tasks in Fitness Landscape Inference for Proteins*. DOI: 10.1101/2021.11.09.467890. URL: https://www.biorxiv.org/content/10.1101/2021.11.09.467890v1 (visited on 08/12/2024). Pre-published.

Jiang, Fan et al. (Nov. 27, 2024). "A General Temperature-Guided Language Model to Design Proteins of Enhanced Stability and Activity". In: *Science Advances* 10.48, eadr2641. DOI: 10.1126/sciadv.adr2641. URL: https://www.science.org/doi/10.1126/sciadv.adr2641 (visited on 04/20/2025).

Su, Jin, Zhikai Li, et al. (May 28, 2024). *SaprotHub: Making Protein Modeling Accessible to All Biologists*. DOI: 10.1101/2024.05.24.595648. URL: https://www.biorxiv.org/content/10.1101/2024.05.24.595648v1 (visited on 05/30/2024). Pre-published.

Hu, Edward J. et al. (Oct. 16, 2021). *LoRA: Low-Rank Adaptation of Large Language Models*. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs]. URL: http://arxiv.org/abs/2106.09685 (visited on 04/24/2025). Pre-published.

Mollon, Manuel F. et al. (Jan. 30, 2025). *Exploring Large Protein Language Models in Constrained Evaluation Scenarios within the FLIP Benchmark*. DOI: 10.48550/arXiv.2501.18223. arXiv: 2501.18223 [cs]. URL: http://arxiv.org/abs/2501.18223 (visited on 04/20/2025). Pre-published.

Meier, Joshua et al. (July 10, 2021). *Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function*. DOI: 10.1101/2021.07.09.450648.

URL: https://www.biorxiv.org/content/10.1101/2021.07.09.450648v1 (visited on 01/29/2024). Pre-published.

Fowler, Douglas M. and Stanley Fields (Aug. 2014). "Deep Mutational Scanning: A New Style of Protein Science". In: *Nature Methods* 11.8 (8), pp. 801–807. ISSN: 1548-7105. DOI: 10.1038/nmeth.3027. URL: https://www.nature.com/articles/nmeth.3027 (visited on 01/29/2024).

Wei, Huijin and Xianghua Li (Jan. 12, 2023). "Deep Mutational Scanning: A Versatile Tool in Systematically Mapping Genotypes to Phenotypes". In: *Frontiers in Genetics* 14. ISSN: 1664-8021. DOI: 10.3389/fgene.2023.1087267. URL: https://www.frontiersin.orghttps://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2023.1087267/full (visited on 04/24/2025).

Hopf, Thomas A. et al. (Feb. 2017). "Mutation Effects Predicted from Sequence Co-Variation". In: *Nature Biotechnology* 35.2 (2), pp. 128–135. ISSN: 1546-1696. DOI: 10.1038/nbt.3769. URL: https://www.nature.com/articles/nbt.3769 (visited on 03/08/2024).

Riesselman, Adam J. et al. (Oct. 2018). "Deep Generative Models of Genetic Variation Capture the Effects of Mutations". In: *Nature Methods* 15.10, pp. 816–822. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0138-4. URL: https://www.nature.com/articles/s41592-018-0138-4 (visited on 04/20/2025).

Cagiada, Matteo et al. (Mar. 15, 2024). *Predicting Absolute Protein Folding Stability Using Generative Models*. DOI: 10.1101/2024.03.14.584940. URL: https://www.biorxiv.org/content/10.1101/2024.03.14.584940v1 (visited on 03/23/2024). Pre-published.

Jumper, John et al. (Aug. 2021). "Highly Accurate Protein Structure Prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: https://www.nature.com/articles/s41586-021-03819-2 (visited on 04/06/2025).

Reeves, Shawn and Subha Kalyaanamoorthy (Sept. 2024). "Zero-Shot Transfer of Protein Sequence Likelihood Models to Thermostability Prediction". In: *Nature Machine Intelligence* 6.9, pp. 1063–1076. ISSN: 2522-5839. DOI: 10.1038/s42256-024-00887-7. URL: https://www.nature.com/articles/s42256-024-00887-7 (visited on 04/26/2025).

Kellogg, Elizabeth H. et al. (2011). "Role of Conformational Sampling in Computing Mutation-Induced Changes in Protein Structure and Stability". In: *Proteins: Structure, Function, and Bioinformatics* 79.3, pp. 830–838. ISSN: 1097-0134. DOI: 10.1002/prot.22921. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22921 (visited on 04/20/2025).

Park, Hahnbeom et al. (Dec. 13, 2016). "Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules". In:

Journal of Chemical Theory and Computation 12.12, pp. 6201–6212. ISSN: 1549-9626. DOI: 10.1021/acs.jctc.6b00819. PMID: 27766851.

Hernández, Iván Martín et al. (Jan. 1, 2023). "Predicting Protein Stability Changes upon Mutation Using a Simple Orientational Potential". In: *Bioinformatics* 39.1, btad011. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btad011. URL: https://doi.org/10.1093/bioinformatics/btad011 (visited on 04/20/2025).

Stourac, Jan et al. (Jan. 8, 2021). "FireProtDB: Database of Manually Curated Protein Stability Data". In: *Nucleic Acids Research* 49.D1, pp. D319–D324. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa981. URL: https://doi.org/10.1093/nar/gkaa981 (visited on 04/20/2025).

Dutton, Oliver et al. (June 19, 2024). *Improving Inverse Folding Models at Protein Stability Prediction without Additional Training or Data*. DOI: 10.1101/2024.06.15.599145. URL: https://www.biorxiv.org/content/10.1101/2024.06.15.599145v2 (visited on 04/20/2025). Pre-published.

Schymkowitz, Joost et al. (July 1, 2005). "The FoldX Web Server: An Online Force Field". In: *Nucleic Acids Research* 33 (Web Server issue), W382–W388. ISSN: 0305-1048. DOI: 10.1093/nar/gki387. PMID: 15980494. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1160148/ (visited on 04/20/2025).

Harmalkar, Ameya et al. (2023). "Toward Generalizable Prediction of Antibody Thermostability Using Machine Learning on Sequence and Structure Features". In: *mAbs* 15.1. ISSN: 1942-0862. DOI: 10.1080/19420862.2022.2163584. PMID: 36683173. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9872953/ (visited on 07/15/2024).

Olsen, Tobias H et al. (Jan. 1, 2022). "AbLang: An Antibody Language Model for Completing Antibody Sequences". In: *Bioinformatics Advances* 2.1, vbac046. ISSN: 2635-0041. DOI: 10.1093/bioadv/vbac046. URL: https://doi.org/10.1093/bioadv/vbac046 (visited on 04/20/2025).

Abanades, Brennan et al. (May 29, 2023). "ImmuneBuilder: Deep-Learning Models for Predicting the Structures of Immune Proteins". In: *Communications Biology* 6.1, pp. 1–8. ISSN: 2399-3642. DOI: 10.1038/s42003-023-04927-7. URL: https://www.nature.com/articles/s42003-023-04927-7 (visited on 04/02/2025).

Maier, James A. et al. (Aug. 11, 2015). "ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB". In: *Journal of Chemical Theory and Computation* 11.8, pp. 3696–3713. ISSN: 1549-9618. DOI: 10.1021/acs.jctc.5b00255. URL: https://doi.org/10.1021/acs.jctc.5b00255 (visited on 04/04/2025).

Brandes, Nadav et al. (Apr. 12, 2022). "ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function". In: *Bioinformatics* 38.8, pp. 2102–2110. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btac020. URL: https://doi.org/10.1093/bioinformatics/btac020 (visited on 04/04/2025).

Salazar, Julian et al. (2020). "Masked Language Model Scoring". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712. DOI: 10.18653/v1/2020.acl-main.240. arXiv: 1910.14659 [cs]. URL: http://arxiv.org/abs/1910.14659 (visited on 04/19/2025).

Koehn, Philipp (July 2004). "Statistical Significance Tests for Machine Translation Evaluation". In: *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2004. Ed. by Dekang Lin and Dekai Wu. Barcelona, Spain: Association for Computational Linguistics, pp. 388–395. URL: https://aclanthology.org/W04-3250/ (visited on 04/15/2025).

Holm, Sture (1979). "A Simple Sequentially Rejective Multiple Test Procedure". In: *Scandinavian Journal of Statistics* 6.2, pp. 65–70. ISSN: 0303-6898. JSTOR: 4615733. URL: https://www.jstor.org/stable/4615733 (visited on 04/04/2025).

Lam, Siu Kwan et al. (Nov. 15, 2015). "Numba: A LLVM-based Python JIT Compiler". In: *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*. LLVM '15. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-4005-2. DOI: 10.1145/2833157.2833162. URL: https://dl.acm.org/doi/10.1145/2833157.2833162 (visited on 04/15/2025).

Pedregosa, Fabian et al. (2011). "Scikit-Learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85, pp. 2825–2830. ISSN: 1533-7928. URL: http://jmlr.org/papers/v12/pedregosa11a.html (visited on 04/15/2025).

AbPROP/Src/Cv_train_ablang_gnn.Py (Jan. 23, 2024). URL: https://github.com/Merck/AbPROP/blob/6da6bb5a1437f1bfef4847afb2dae35bf6f5c5ca/src/cv_train_ablang_gnn.py#L190C1-L190C118 (visited on 04/01/2025).

AbPROP/Src/Ensemble.Py (Jan. 23, 2024). URL: https://github.com/Merck/AbPROP/blob/6da6bb5a1437f1bfef4847afb2dae35bf6f5c5ca/src/ensemble.py#L42 (visited on 04/01/2025).

Hudson, Peter J and Alexander A Kortt (Dec. 10, 1999). "High Avidity scFv Multimers; Diabodies and Triabodies". In: *Journal of Immunological Methods* 231.1, pp. 177–189. ISSN: 0022-1759. DOI: 10.1016/S0022-1759(99)00157-X. URL: https://www.sciencedirect.com/science/article/pii/S002217599900157X (visited on 04/02/2025).

Frappier, Vincent and Rafael J. Najmanovich (Apr. 24, 2014). "A Coarse-Grained Elastic Network Atom Contact Model and Its Use in the Simulation of Protein Dynamics and the Prediction of the Effect of Mutations". In: *PLOS Computational Biology* 10.4, e1003569. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1003569. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003569 (visited on 04/02/2025).

Mailhot, Olivier and Rafael Najmanovich (Oct. 11, 2021). "The NRGTEN Python Package: An Extensible Toolkit for Coarse-Grained Normal Mode Analysis of Proteins, Nucleic Acids, Small Molecules and Their Complexes". In: *Bioinformatics* 37.19,

pp. 3369–3371. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab189. URL: https://doi.org/10.1093/bioinformatics/btab189 (visited on 04/02/2025).

Bauer, Jacob A. et al. (Sept. 10, 2019). "Normal Mode Analysis as a Routine Part of a Structural Investigation". In: *Molecules* 24.18, p. 3293. ISSN: 1420-3049. DOI: 10.3390/molecules24183293. PMID: 31510014. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6767145/ (visited on 02/20/2024).

Breiman, Leo (Oct. 1, 2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: https://doi.org/10.1023/A:1010933404324 (visited on 04/03/2025).

Scikit-Learn Guide (2025). scikit-learn. URL: https://scikit-learn/stable/modules/permutation_importance.html (visited on 04/03/2025).

Lee, Jee Un et al. (Jan. 2017). "Molecular Basis for the Neutralization of Tumor Necrosis Factor ? By Certolizumab Pegol in the Treatment of Inflammatory Autoimmune Diseases". In: *International Journal of Molecular Sciences* 18.1 (1), p. 228. ISSN: 1422-0067. DOI: 10.3390/ijms18010228. URL: https://www.mdpi.com/1422-0067/18/1/228 (visited on 04/14/2025).

Hunter, John D. (May 2007). "Matplotlib: A 2D Graphics Environment". In: *Computing in Science & Engineering* 9.3, pp. 90–95. ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55. URL: https://ieeexplore.ieee.org/document/4160265 (visited on 04/15/2025).

Waskom, Michael L. (Apr. 6, 2021). "Seaborn: Statistical Data Visualization". In: *Journal of Open Source Software* 6.60, p. 3021. ISSN: 2475-9066. DOI: 10.21105/joss.03021. URL: https://joss.theoj.org/papers/10.21105/joss.03021 (visited on 04/15/2025).

Sehnal, David et al. (July 2, 2021). "Mol* Viewer: Modern Web App for 3D Visualization and Analysis of Large Biomolecular Structures". In: *Nucleic Acids Research* 49.W1, W431–W437. ISSN: 0305-1048. DOI: 10.1093/nar/gkab314. URL: https://doi.org/10.1093/nar/gkab314 (visited on 04/15/2025).

Bittrich, Sebastian et al. (2024). "Describing and Sharing Molecular Visualizations Using the MolViewSpec Toolkit". In: *Current Protocols* 4.7, e1099. ISSN: 2691-1299. DOI: 10.1002/cpz1.1099. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.1099 (visited on 04/15/2025).

Lefranc, Marie-Paule, Véronique Giudicelli, et al. (Jan. 2009). "IMGT, the International ImMunoGeneTics Information System". In: *Nucleic Acids Research* 37 (Database issue), pp. D1006–1012. ISSN: 1362-4962. DOI: 10.1093/nar/gkn838. PMID: 18978023.

Suzek, Baris E., Yuqi Wang, et al. (Mar. 15, 2015). "UniRef Clusters: A Comprehensive and Scalable Alternative for Improving Sequence Similarity Searches". In: *Bioinformatics* 31.6, pp. 926–932. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu739. URL: https://doi.org/10.1093/bioinformatics/btu739 (visited on 04/07/2025).

Olsen, Tobias H., Fergus Boyles, et al. (Jan. 2022). "Observed Antibody Space: A Diverse Database of Cleaned, Annotated, and Translated Unpaired and Paired Antibody Sequences". In: *Protein Science: A Publication of the Protein Society* 31.1, pp. 141–146. ISSN: 1469-896X. DOI: 10.1002/pro.4205. PMID: 34655133.

Dunbar, James, Konrad Krawczyk, et al. (Jan. 1, 2014). "SAbDab: The Structural Antibody Database". In: *Nucleic Acids Research* 42.D1, pp. D1140–D1146. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1043. URL: https://doi.org/10.1093/nar/gkt1043 (visited on 03/19/2025).

Sillitoe, Ian et al. (Jan. 8, 2021). "CATH: Increased Structural Coverage of Functional Space". In: *Nucleic Acids Research* 49.D1, pp. D266–D273. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa1079. PMID: 33237325.

Nordberg, Henrik et al. (Jan. 2014). "The Genome Portal of the Department of Energy Joint Genome Institute: 2014 Updates". In: *Nucleic Acids Research* 42 (Database issue), pp. D26–31. ISSN: 1362-4962. DOI: 10.1093/nar/gkt1069. PMID: 24225321.

Vidyasagar, P.B. et al. (2004). "Conserved Oligopeptides in the Rubisco Large Chains". In: *Life in the Universe*. Ed. by Joseph Seckbach et al. Red. by Joseph Seckbach. Vol. 7. Dordrecht: Springer Netherlands, pp. 133–134. ISBN: 978-94-007-1003-0. DOI: 10.1007/978-94-007-1003-0_26. URL: http://link.springer.com/10.1007/978-94-007-1003-0_26 (visited on 04/07/2025).

Suzek, Baris E., Hongzhan Huang, et al. (May 15, 2007). "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters". In: *Bioinformatics* 23.10, pp. 1282–1288. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btm098. URL: https://doi.org/10.1093/bioinformatics/btm098 (visited on 04/07/2025).

Bar-On, Yinon M. et al. (June 19, 2018). "The Biomass Distribution on Earth". In: *Proceedings of the National Academy of Sciences* 115.25, pp. 6506–6511. DOI: 10.1073/pnas.1711842115. URL: https://www.pnas.org/doi/10.1073/pnas.1711842115 (visited on 04/07/2025).

Tadokoro, Takashi et al. (Feb. 12, 2024). *Thermostability and Binding Properties of Single-Chained Fv Fragments Derived from Therapeutic Antibodies*. DOI: 10.1101/2024.02.09.577534. URL: https://www.biorxiv.org/content/10.1101/2024.02.09.577534v1 (visited on 07/15/2024). Pre-published.

Petukhov, Michael et al. (Jan. 1999). "Local Water Bridges and Protein Conformational Stability". In: *Protein Science* 8.10, pp. 1982–1989. ISSN: 0961-8368, 1469-896X. DOI: 10.1110/ps.8.10.1982. URL: https://onlinelibrary.wiley.com/doi/10.1110/ps.8.10.1982 (visited on 04/14/2025).

Raybould, Matthew I. J. et al. (Jan. 8, 2024). "Contextualising the Developability Risk of Antibodies with Lambda Light Chains Using Enhanced Therapeutic Antibody Profiling". In: *Communications Biology* 7.1, pp. 1–13. ISSN: 2399-3642. DOI: 10.1038/s42003-023-05744-8. URL: https://www.nature.com/articles/s42003-023-05744-8 (visited on 03/26/2025).

Townsend, Catherine L. et al. (Sept. 27, 2016). “Significant Differences in Physicochemical Properties of Human Immunoglobulin Kappa and Lambda CDR3 Regions”. In: *Frontiers in Immunology* 7, p. 388. ISSN: 1664-3224. DOI: 10.3389/fimmu.2016.00388. PMID: 27729912. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5037968/ (visited on 03/26/2025).

Ruffolo, Jeffrey A., Lee-Shin Chu, et al. (Apr. 25, 2023). “Fast, Accurate Antibody Structure Prediction from Deep Learning on Massive Set of Natural Antibodies”. In: *Nature Communications* 14.1, p. 2389. ISSN: 2041-1723. DOI: 10.1038/s41467-023-38063-x. URL: https://www.nature.com/articles/s41467-023-38063-x (visited on 04/24/2025).

Appendix

Model	Score	Median SCC	CI 95% lo	CI 95% hi
ESMC 300M PLL	joint_ll_score	0.41	0.33	0.49
ESMC 300M	joint_ll_score	0.41	0.32	0.48
ESMC 300M PLL	logit_score	0.38	0.30	0.46
ESMC 300M PLL	logit_score_linker	0.38	0.29	0.46
ESMC 300M	logit_score_linker	-0.35	-0.44	-0.26
ESMC 300M	logit_score	-0.34	-0.43	-0.25
ESMC 300M	ll_score_linker	0.33	0.24	0.42
ESMC 300M	ll_score	0.33	0.24	0.41
ESMC 300M PLL	ll_score_linker	0.33	0.23	0.41
ESMC 300M PLL	ll_score	0.32	0.23	0.41
ESM3 Chain Break	logit_score	0.32	0.23	0.41
ESMC 300M	top_5_entropy_score	-0.32	-0.40	-0.23
ESM-IF scFv SSS	ll_avg	0.32	0.23	0.41
ESMC 300M	entropy_score	-0.32	-0.40	-0.23

Table 5 Top 5% of zero-shot scores with metrics computed on the bootstrapped AbProp training set. The table shows the median Spearman correlation coefficient (SCC) and the 95% confidence interval (CI) for each score. The scores are sorted by their absolute median SCC. For comparison of the ranking on the full dataset, see the boxplot Figure 3.1, however, note that the boxplot shows only the best score for each model. Two minor differences are apparent in the ranking: on the full set, ESMC 600M is ranked marginally above ESM3, here it did not make the top 5% cut. The second difference is that ESMC 300M in the pseudo-log-likelihood setting (PLL) here ranks marginally above the standard setting, while on the full set it is the other way around. Note that PLL setting is computationally expensive—it requires L runs of the model, where L is the length of the sequence—while all other settings are single-pass. The computational cost of the PLL setting is not justified by the marginal improvement in performance, here not visible due to rounding. Therefore, when combining the scores, we simply use the standard score instead. PLL—pseudo-log-likelihood, SSS—`score_second_sequence`

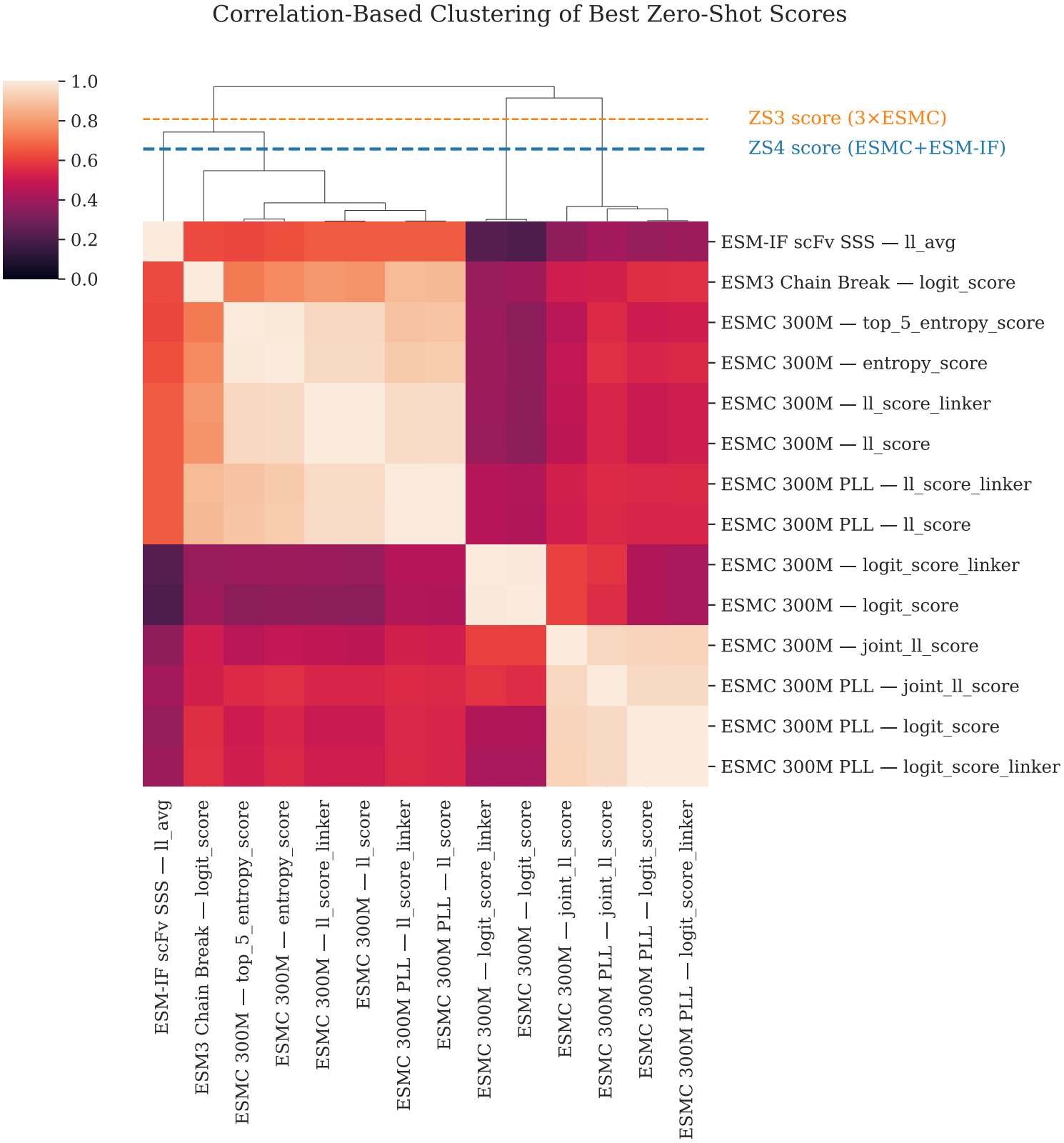

Figure 12 Correlation-based hierarchical clustering of zero-shot scores. Top 5% scores (in absolute median SCC with T_m; on AbProp training set) were selected. The heatmap shows the absolute pairwise SCCs between these scores. The top dendrogram was computed using UPGMA (average linkage clustering) based on $1 - |\text{SCC}|$ as the distance metric. A dashed blue line marks a cut in the dendrogram at a certain height, partitioning the metrics into four visually distinct clusters. Representatives from each of these flat clusters were combined into a composite score, ZS4 (ESMC+ESM-IF) score (see Section 2.2.3). Additionally, a second cut was performed to define three clusters comprising *only* representative scores from the *ESMC model*, which were merged into ZS3 (3xESMC) score. These composite scores were evaluated in the main text.

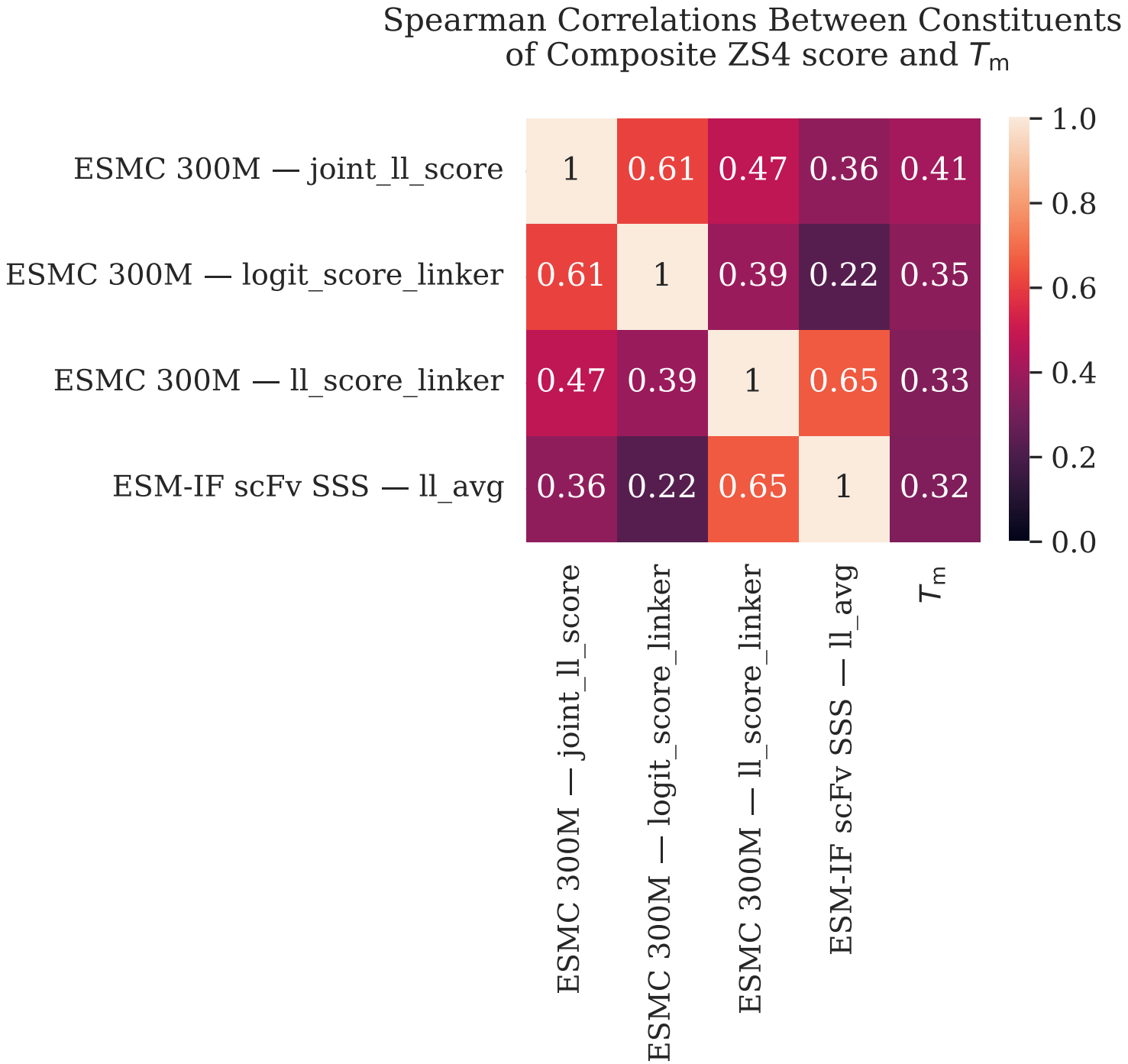

Figure 13 Pairwise Spearman correlations between the constituent scores of the composite ZS4 score and the target variable T_m. The composite score includes three scores from ESMC 300M (joint log-likelihood, logit score with linker, and log-likelihood with linker) and one from ESM-IF (average log-likelihood in the `score_second_sequence` setting). Correlations with T_m are shown in the last column.

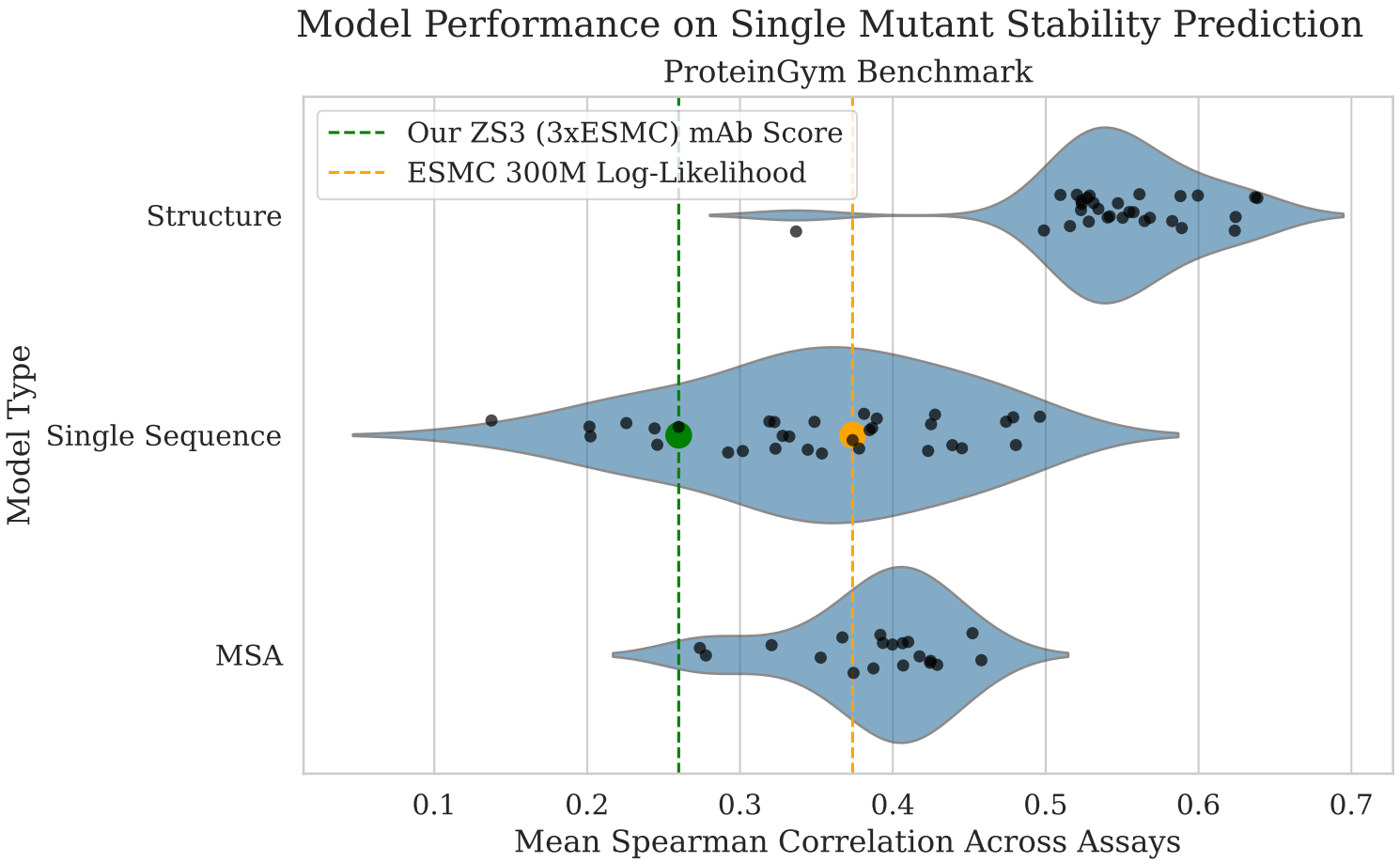

Figure 14 Performance of the antibody-derived zero-shot ZS3 score (green), composed of three ESMC-based components, on the ProteinGym single mutant stability benchmark. The score is evaluated across assays involving single amino acid substitutions in diverse proteins, and compared to 79 models grouped by input type: structure-based, single-sequence, and MSA-based, already present in the benchmark, and the ESMC log-likelihood. While ZS3 score beats all single-sequence models in the double mutant stability prediction, it underperforms relative to the standard ESMC log-likelihood score (orange) in single mutant effect predition, suggesting that aggregating the whole amino acid distribution—smoothing out the potential effect of individual mutations—may reduce predictive power in the single mutant setting.

	Lassos A, B	Random Forests	Mean
H86	0.058	0.001	0.030
L146	0.035	0.018	0.027
H60	0.020	0.005	0.012
H110	0.015	0.001	0.008
H33	0.006	0.001	0.004
H74	0.004	0.003	0.003
H30	0.001	0.005	0.003
L40	0.000	0.004	0.002
H144	0.003	0.001	0.002
L76	0.000	0.003	0.001
H17	0.001	0.001	0.001
H41	0.000	0.001	0.001
H72	0.000	0.001	0.001
L105	0.000	0.001	0.001
L104	0.000	0.001	0.001

Table 6 15 positions identified by intersecting the top 50 positions selected by Random Forests and by the non-global Lasso (A and B) models. Importance scores represent the average drop in R^2 upon feature shuffling, summed across all feature types at each position and averaged across models. The final "Mean" column shows the average of the two model-specific importance scores. Visualized in Figure 3.9.

Position	Importance	Position	Importance	Position	Importance
H86	0.058	H82	0.011	H138	0.004
H102	0.039	H76	0.010	H79	0.004
L146	0.035	L56	0.010	L68	0.004
H60	0.020	L78	0.008	L94	0.004
H71	0.016	H33	0.006	H137	0.004
H110	0.015	L139	0.006	H74	0.004
H23	0.014	L26	0.006	H66	0.003
H32	0.012	H106	0.006	H144	0.003
L3	0.011	H93	0.006	L13	0.003
H99	0.011	L10	0.005	H103	0.002

Table 7 Top 30 positions with the highest importance scores from Lasso models A and B, considering only positional features. Importance values reflect the drop in R^2 upon shuffling, summed across feature types and averaged between the two models. These positions are visualized in structural context in Figure 3.10, and key sites are further discussed in relation to the mapped structure.