# Somatic Hypermutation Informed Vocabulary Encoder Representations

**Chiho Im**[1] **Artem Mikelov**[2] **Ryan Zhao**[1] **Anshul Kundaje**[1,3] **Scott D. Boyd**[2]

[1]Department of Computer Science, Stanford University
[2]Department of Pathology, Stanford University
[3]Department of Genetics, Stanford University
{chihoim, amikelov, ryanzhao, akundaje, sboyd1}@stanford.edu

## Abstract

Somatic hypermutations (SHMs) acquired during affinity maturation of memory B cell receptors (mBCRs) carry important immunological signals, but remain challenging for protein language models (PLMs) to capture effectively. We introduce SHIVER, a mutation-aware antibody language model that treats each amino acid substitution as a distinct token, allowing the model to directly encode the context-dependent impact of SHMs. Trained on paired heavy and light chain sequences from human mBCR repertoires, SHIVER incorporates a tailored vocabulary, a subsampling strategy for data augmentation, and a mutation-focused masking scheme to better model the dynamics of affinity maturation. We evaluate SHIVER on the task of predicting mBCR binding to influenza antigens and find that it outperforms both general and antibody-specific PLMs using a simple logistic head. Our results suggest that explicitly modeling SHMs improves biological relevance and generalization of learned representations.

## 1 Introduction

Protein language models (PLM) such as ESM-2 [11], ProtTrans [3], ProteinBERT [2], pre-trained with self-supervised learning, produce meaningful embeddings that capture biologically relevant semantic information from input sequences. Specialized protein language models such as IgLM [22], AbLang2 [18], and mBLM [26] that produce representations of antibody sequences facilitate diverse downstream applications such as heavy-light chain pairing [5] and antigen binding prediction [6, 7, 26], potentially leading to drastic improvements in cost efficiency for the development of new monoclonal antibody-based therapeutics.

Current antibody-centered language models, however, often overlook a critical aspect of antibody biology: somatic hypermutation (SHM) coupled with antibody affinity maturation, a process which serves as biological "fine-tuning" of the initial antibody-coding sequences derived from their germline origins. SHM introduces mutations which are not random but shaped by evolutionary pressures to increase affinity of antibody binding and functional constraints from a structural perspective.

To better capture the immunological relevance of SHM, we introduce somatic hypermutation informed vocabulary encoder representations (SHIVER), a novel antibody language model that incorporates explicit mutation-aware tokens. Rather than treating mutated residues identically to germline-derived residues, our approach encodes each mutation from the germline sequence as a distinct token (e.g., K $\rightarrow$ R becomes "K_R"). This design allows the model to directly learn the functional implications of specific amino acid changes in the context of affinity maturation. Our model predicts the interaction

The 20th Conference on Machine Learning in Computational Biology. PMLR, 2025.

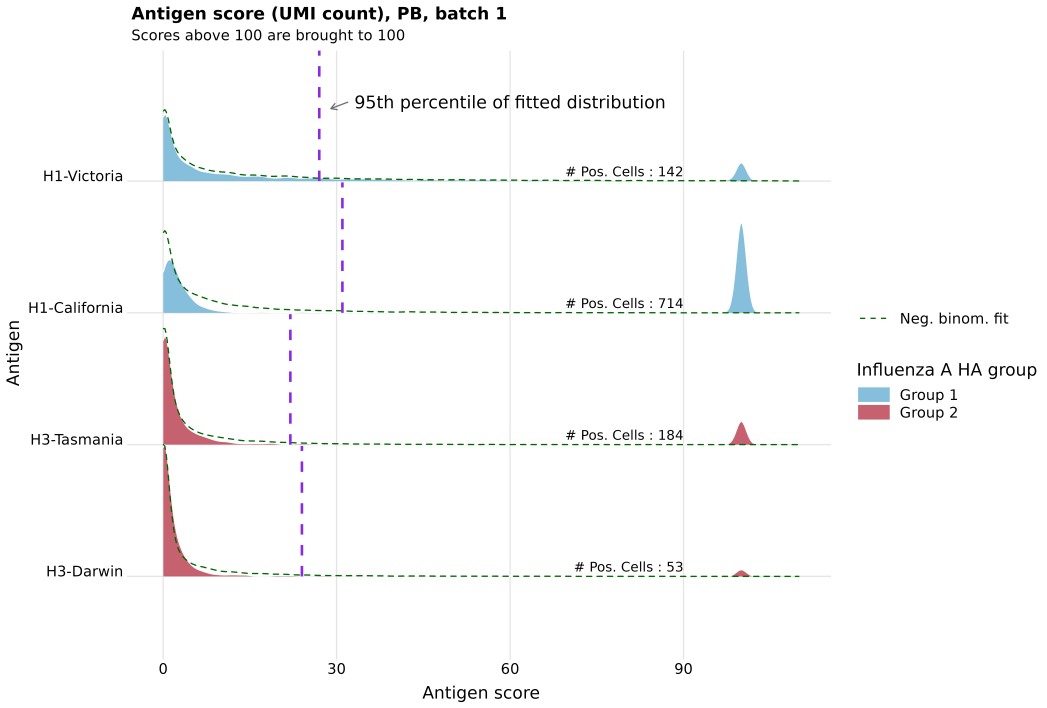


Figure 1: Distribution of raw UMI count scores in the peripheral blood (PB) memory B cell receptors for a representative batch. Dashed green light represents modeled distribution of noise (blue), and dashed purple line represents the chosen threshold at 95th percentile.

between several antigens and memory B cell receptors with higher accuracy compared to other PLMs. By explicitly modeling the trajectory of SHMs, SHIVER bridges the gap between sequence representation and functional insight. This framework paves the way for more accurate modeling of immune memory and rational antibody design guided by evolutionary signals.

## 2 Dataset & Preprocessing

### 2.1 MiXCR Processing of Single-cell RNA Sequencing Reads of MBC Receptors

Memory B cells (MBC) are long-lived cells of the adaptive immune system that persist after an initial infection and rapidly respond upon re-exposure to the same antigen. They express B cell receptors (BCRs) on their surface that share antigen specificity with the antibodies they secrete, if they become plasma cells. These receptors are often refined through somatic hypermutation and affinity maturation. Memory B cell receptor (mBCR) repertoires demonstrate distinct characteristics from other B cell subsets as they carry marks of antigen exposure and selection [4, 13].

Thus, we used publicly available single cell sequencing data of human mBCR repertoires [9, 19, 20] to obtain a pre-training dataset of paired heavy and light chain amino acid sequences. Sequencing reads were processed using MiXCR [1, 14], which enables robust error correction, ambient RNA-derived contamination removal, and reliable heavy and light chain pairing (Section A.1). Importantly, detailed annotation for each paired sequence allowed extraction of SHMs, which was instrumental for utilizing mutation-aware amino-acid vocabulary, a key feature of the proposed approach. This resulted in a total of 210,473 de-duplicated paired mBCR sequences for pre-training.

### 2.2 mBCR Dataset Labeled with Flu Antigen Binding

To demonstrate the utility of the proposed antibody language model, we later evaluate our model on the task of antibody-antigen binding prediction due to the importance of this application for drug development and diagnostics. We curated a dataset of paired heavy- and light-chain sequences from

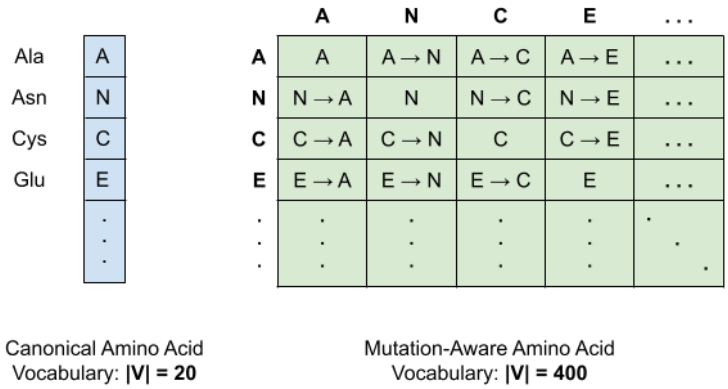


Figure 2: Illustration of our mutation-aware vocabulary. Rather than using the 20 canonical amino acids, we use *directional pairs* of amino acids as part of the vocabulary for training our protein language model, characterizing every possible substitution-based amino acid mutation.

*unseen* mBCR receptors from two different donor samples: 1) peripheral blood (PB) B cells and 2) splenic B cells. Each receptor sequence is annotated with raw binding scores to several antigens; each score represents a number of antigen molecules bound to a cell with this receptor. Antigens comprise a set of influenza hemagglutinins (HAs) from seasonal vaccine strains (Section A.2).

To obtain binary labels for binding and non-binding cells based on antigen scores, we developed the following procedure applied independently to each sample. Similarly to the approach used in [27], we modeled a background (noise) distribution of raw scores for each antigen by fitting a negative binomial distribution, excluding all cells with raw scores greater than 200. The thresholds were then defined as the 95th percentile of the fitted distributions, with cells exceeding this threshold assigned as positive for that antigen. This procedure was applied separately for each antigen and each sample, thus accounting for possible batch effects from the experiments (Figure 1).

# 3 Methods

## 3.1 Mutation-Aware Vocabulary

Traditional protein language models have been trained using the vocabulary of 20 canonical amino acids [2, 11, 18, 22, 26]. As a result, such protein language models often struggle to directly capture the effect of mutations in a protein sequence. To address this, SHIVER is trained using a new vocabulary of amino acids augmented with mutation information (Figure 2). Using the canonical amino acids only, the size of vocabulary (excluding the unknown amino acid token X and other special tokens) is 20. With the additional mutation tokens, we now have a total of $20 + \frac{20!}{18!} = 400$ tokens, which characterize every possible substitution-based amino acid mutation that can occur in a protein sequence. Such mutation-aware vocabulary serves as an efficient alternative method to tokenize pairs of mutated/germline antibody sequences (Figure 3A).

## 3.2 Model Architecture and Training Details

SHIVER is implemented as a RoBERTa-style Transformer encoder [12], consisting of 6 layers with 12 attention heads per layer, a hidden embedding dimension of 768, and intermediate feedforward layers of dimension 3072. The model supports sequences up to 256 tokens, sufficient to accommodate the concatenated variable regions (Fv) of heavy and light antibody chains. Training is performed using a masked language modeling (MLM) objective, where tokens are masked and predicted based on their surrounding context. Pre-training is conducted on 210,473 paired memory B cell receptor (mBCR) sequences derived from the OAS database, using the mutation-aware vocabulary to guide learning toward biologically meaningful variation. While the architecture is smaller than large-scale PLMs such as ESM-2, SHIVER's inductive biases – specifically its mutation-centric vocabulary and paired-chain representation – enable it to learn specialized embeddings that outperform larger models

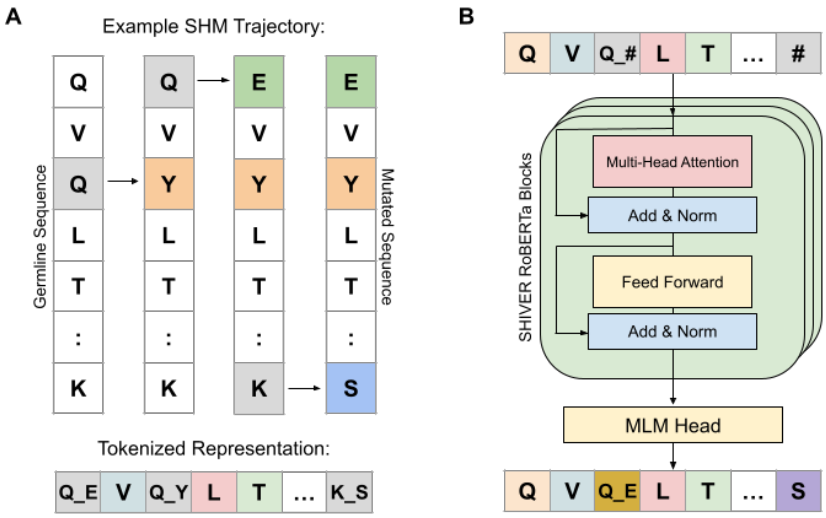


Figure 3: A) An example of tokenized representation of mBCR sequence labeled with mutation information. B) Architecture of SHIVER. During training, for a given masked input sequence of paired antibody heavy and light chains, the model tries to predict mutation tokens for partially masked tokens and canonical amino acid tokens for regular mask tokens.

on downstream binding prediction tasks. The diagram of the model architecture is shown in Figure 3B. The masking strategy for masking mutation tokens is described in Section 3.4.

### 3.3 Data Augmentation via Mutation Subsampling

Due to the limited availability of publicly accessible human mBCR sequence data (∼200,000 sequences), we introduce a sampling-based augmentation strategy to increase training diversity and promote generalization. Each sequence in our dataset is annotated with a list of somatic hypermutations relative to the corresponding germline sequence. To create augmented examples, we stochastically subsample subsets of these mutations, effectively simulating intermediate stages along the B-cell affinity maturation trajectory.

Concretely, for a given mutated sequence containing $k$ mutations, we uniformly sample an integer $q \in \{1, ..., k\}$ and randomly select $q$ mutations from the original set. We then construct an augmented sequence by retaining the selected mutations and reverting the remaining $k - q$ residues to their germline (unmutated) identities.

This procedure is motivated by the following properties of SHM process. First, mutations accumulate accumulate progressively and iteratively during SHM process. Second, the majority of non-synonymous SHMs that pass are enriched for mutations that improve or maintain antigen binding. [16]. To control the extent of augmentation, we introduce a sampling probability hyperparameter $\alpha \in [0, 1]$. For each sequence, with probability $\alpha$, we retain all mutations (i.e., no augmentation is applied). With probability $1 - \alpha$, we apply the mutation subsampling procedure described above.

This approach allows the model to learn from both fully mutated and partially reverted sequences, providing diverse mutational contexts during training. By exposing the model to intermediate mutational states, we aim to improve its ability to learn meaningful representations of SHM patterns. For our experiments, we use $\alpha = 0.5$, which was the empirically determined optimal value.

### 3.4 Mutation-Focused Masking Strategy

Traditionally, in masked language modeling, we use a single mask token to mask the positions in an input sequence for which the model would like to predict the original token's identity. However, recent methods have used hybrid tokens which allowed the usage of multiple types of masked tokens. For example, SaProt is a protein language model that uses a hybrid "structure-aware" token for amino

| MODEL | EMBEDDING SIZE | H1-VICTORIA | H1-CALIFORNIA | H3-TAZMANIA | H3-DARWIN |
|---|---|---|---|---|---|
| ESM-2 | 1280 (640*2) | 0.673 | 0.737 | 0.727 | 0.724 |
| ESM-2 AB | 1280 (640*2) | <u>0.687</u> | <u>0.756</u> | <u>0.736</u> | <u>0.739</u> |
| ESM-C | 1920 (960*2) | 0.681 | 0.742 | 0.714 | 0.701 |
| IGLM | 512 | 0.677 | **0.759** | 0.728 | 0.722 |
| ABLANG-2 | 480 | 0.679 | 0.709 | 0.709 | 0.712 |
| MBLM | 768 | 0.682 | 0.748 | 0.711 | 0.729 |
| SHIVER (OURS) | 768 | **0.719** | 0.755 | **0.739** | **0.749** |

Table 1: AUROC scores for binder predictions against different human influenza HA's using embeddings from various protein language models. ESM2-Ab denotes ESM-2 finetuned on SHIVER's training dataset. (**bold** indicates best and <u>underline</u> indicates second best)

acids that capture both the identity of the amino acid as well as the discrete structural features [25]. Correspondingly during training, they use a set of *partial* mask tokens where the model tries to predict only the amino acid identity and not the structural feature. Similarly, when we are masking the tokens for SHIVER, we are more interested in predicting the result of the mutation compared to its germline origin. Thus, we used 20 partial mask tokens, $\{A\_\#, C\_\#, ...\}$, in addition to the regular mask token to mask both the mutation tokens as well as the canonical amino acid tokens, respectively.

In a given sequence, the expected ratio of mutation tokens on average is around 10% of all tokens. This means that with random assignment of masks, the mutation tokens will be masked more infrequently compared to regular canonical amino acid tokens. However, SHMs play an important role in characterizing the binding affinity and specificity of immune receptors to particular antigens. Just as language modeling efforts that emphasize non-templated regions – such as CDR3s in antibodies – have led to improved semantic understanding of antigen recognition [15, 23], we hypothesize that increasing the focus on mutation sites will similarly help the model learn their functional relevance. Therefore, we enforce a higher ratio of mutation tokens to be masked so that our model is able to learn the distribution of mutations in the context of surrounding amino acids.

Given a paired (H/L chains) memory B-Cell receptor sequence of length $n$ containing $k$ mutation tokens ($k \leq n$), we want to compute the probabilities $p_r$ and $p_m$, which denote the probability of masking each **r**egular token and **m**utation token, respectively. Following standard training procedures for BERT-based models, we would like to mask roughly 15% of the tokens in expectation for a given sequence. We denote hyperparameter $q \in [0, 1]$ as the fraction of mutation tokens that were masked out of all masked tokens. Then, we parameterize $p_r$ and $p_m$ such that the fraction of mutation tokens masked follows a negative binomial distribution centered at $q$:

$$(p_r, p_m) = \begin{cases} (0.15, 0) & k = 0 \\ \left(0.15 \cdot \frac{qn}{k},\ 0.15 \cdot \frac{(1-q)n}{n-k}\right) & k > 0 \end{cases}$$

## 4 Results

As mentioned in Section 2.2, there are two separate sets of mBCR sequences used for evaluation. In Sections 4.1 and 4.2, the pre-trained SHIVER model is evaluated using the peripheral blood (PB) mBCR dataset, labeled with binding information to select influenza HAs. In Section 4.3, the pre-trained SHIVER model is first supervised fine-tuned on the PB mBCR dataset, and is evaluted on a separate splenic mBCR dataset labeled with binding information to a subset of the above influenza HAs.

### 4.1 SHIVER Enables Sequence-based Prediction of mBCR-Flu Antigen Binding

As many of the self-supervised protein language models contain biologically important semantic information, we wanted to test whether they can be used to predict interactions against human Flu antigens without further finetuning. We formulate the problem as a binary classification task, where given an input sequence (with mutation annotation) the model tries to predict its binary interaction

| Model | Embedding Size | H1-Victoria | H1-California | H3-Tazmania | H3-Darwin |
|---|---|---|---|---|---|
| Random Embedding | 768 | 0.497 | 0.473 | 0.492 | 0.467 |
| Random Embedding | 1024 | 0.513 | 0.500 | 0.511 | 0.508 |
| One-Hot AA Encoding | 40 (20*2) | 0.607 | 0.646 | 0.621 | 0.601 |
| Canonical AA Vocab | 768 | 0.683 | 0.733 | 0.724 | 0.735 |
| SHIVER (H-chain only) | 768 | 0.705 | 0.710 | 0.688 | 0.691 |
| SHIVER | 768 | 0.719 | 0.755 | 0.739 | 0.749 |

Table 2: AUROC scores for binder predictions against different human Flu Antigens using randomly generated embeddings, one-hot encodings, RoBERTa embeddings trained with canonical AA vocabulary, and SHIVER embeddings with/without light chain data.

label. We used logistic regression on the embeddings generated by each model to predict the binary labels on a panel of human Flu antigens (H1 and H3 subtypes of hemagglutinin, or HA, a surface glycoprotein on the influenza virus), using 5-fold cross validation to evaluate performance. The results are shown in Table 1.

Overall, SHIVER appears to outperform the existing protein/antibody language models across most of the Flu antigens, as measured by AUROC scores. Notably, SHIVER achieves the highest AUROC on three out of the four tested HA antigens, suggesting that the representations it learns capture relevant features for predicting antibody-antigen binding interactions. For H1-California, where SHIVER does not yield the top score, its performance is still highly competitive – the AUROC differs only marginally from that of the best-performing alternative, indicating that SHIVER remains robust across antigens with varying sequence properties. These results highlight SHIVER's strong generalization capabilities and suggest that its learned embeddings through mutation-informed vocabulary are well-aligned with immunological signals relevant to HA-mBCR interactions.

### 4.2 Model Ablations Reveal the Impact of Chain Pairing and Mutation-Specific Encoding

In this section, we investigate the components of SHIVER that contribute to its strong performance on antibody-antigen binding prediction. First, we assess the importance of light chain information by training SHIVER on heavy chain sequences only, using the same pre-training and fine-tuning datasets. There is an inherent scarcity of paired mBCR sequence data both in public and proprietary datasets. For example, in public datasets (i.e. OAS), there are 517,858 paired chain mBCR's sequenced as opposed to 13,478,858 heavy chain mBCR's (as of 03.28.2025). Thus, several antibody language models have separately trained models for heavy and light chains due to the lack of information regarding their native pairing [10, 17]. As shown in Table 2, removing light chain input results in a noticeable drop in AUROC scores across all antigens, underscoring the contribution of native heavy-light chain pairing to accurate binding representation. This aligns with immunological evidence that antigen recognition is shaped by structural and functional complementarity between both chains.

Beyond chain pairing, we also evaluate the value of SHIVER's mutation-aware vocabulary by comparing performance against a model trained with canonical amino acid tokens. Using the same architecture and training procedure, the canonical vocabulary variant consistently underperforms relative to the mutation-aware version, demonstrating that explicitly modeling amino acid substitutions yields more biologically informative embeddings. This improvement is likely due to the model's ability to capture the directionality and context of specific mutations, which are critical for interpreting affinity maturation trajectories. Moreover, we observe that naive encoding strategies, such as one-hot vectors or randomly initialized embeddings, indeed fail to approach the performance of RoBERTa model trained with canonical amino acid vocabulary – highlighting the necessity of learned contextual representations. While embedding size can affect performance, SHIVER achieves competitive or superior results despite using a smaller embedding dimension (768) compared to models like ESM-C (1920), suggesting that the inductive bias introduced by the mutation-aware design can outweigh increases in representational capacity alone.

Together, these ablation experiments confirm that SHIVER's gains stem not only from architectural scaling, but from principled design choices that reflect immunological structure – namely, leveraging paired chain information and tokenizing mutations directly to capture the semantic implications of somatic hypermutation.

### 4.3 Fine-tuning SHIVER Enables Generalization Across Different Immune Contexts

To further assess the transferability and robustness of SHIVER representations, we fine-tuned the model on memory B cell receptor (mBCR) sequences derived from peripheral blood (PB) samples of individuals vaccinated with the seasonal tetravalent inactivated influenza vaccine in Fall 2022 (Section 2.2). These paired heavy-light chain sequences were annotated with binding affinities to a panel of influenza hemagglutinin (HA) antigens, providing a supervised signal to adapt SHIVER embeddings to real-world antigen exposure contexts. We then evaluated the fine-tuned model on an independent dataset of mBCR sequences from stimulated splenic B cell cultures, representing a distinct immunological environment not seen during training. As shown in Table A.3, fine-tuning resulted in modest but consistent improvements in AUROC scores. We hypothesize that the relatively small size of the fine-tuning dataset – comprising only $\sim$2,000 labeled examples – may limit the magnitude of performance gains.

Despite the limited supervision, SHIVER's strong baseline performance and its ability to generalize across donor- and tissue-specific repertoires suggest that it captures immunologically meaningful features that persist across diverse immune contexts. The model's robustness to distributional shifts between peripheral blood and splenic compartments – each with distinct clonal structures and activation histories – demonstrates the biological relevance of its mutation-aware representations. These results highlight SHIVER's potential as a foundation model for immunological applications, where fine-tuning on additional antigen-binding data could further enhance its utility for modeling context-specific immune responses.

## 5 Discussion & Future Work

Proteins accumulate mutations gradually throughout evolution, often one amino acid at a time. In contrast, memory B cells in humans experience rapid and extensive sequence diversification through SHMs, introducing multiple amino acid substitutions within days. These hypermutations occur in a targeted manner, particularly within the variable regions of the immunoglobulin heavy and light chains, and play a critical role in antibody affinity maturation. Modeling this dynamic, context-dependent process requires more than a static representation of amino acid identity—it demands a vocabulary that can reflect the evolutionary and functional consequences of mutation events themselves.

In this work, we introduced a mutation-aware vocabulary that treats each amino acid substitution as a distinct token. This design enables the model to directly capture the semantics of specific mutations and their biological impact, moving beyond treating a mutated residue as indistinguishable from its unmutated counterpart. As shown in Figure 4, the empirical coverage of pairwise substitution mutations in our dataset reflects known immunological constraints and selection pressures. For example, mutations introducing non-canonical cysteines are notably rare, consistent with their potential to cause structural disruption via misfolded disulfide bonds – mutations that are typically purged from the repertoire through negative selection [21]. The model's exposure to such constraints enables it to implicitly learn which mutation patterns are biologically plausible and which are likely to be deleterious.

Our findings suggest that incorporating mutation-aware tokens leads to improved representations of antibody sequences, especially in tasks that involve functional interpretation, such as predicting antigen binding. SHIVER demonstrates strong capabilities in identifying mBCR binders against diverse influenza antigens, outperforming several existing antibody-specific language models as well as general protein language models finetuned on our same pre-training dataset (Table 1). This suggests that explicitly modeling SHM injects inductive bias that aligns closely with the immune system's evolutionary strategy. Although this work focuses on substitution-based mutations, future directions include incorporating insertion and deletion mutations as well as providing codon-level DNA sequence context. Expanding in these directions may further enhance the biological fidelity of the model and broaden its applicability to more diverse immunological modeling tasks.

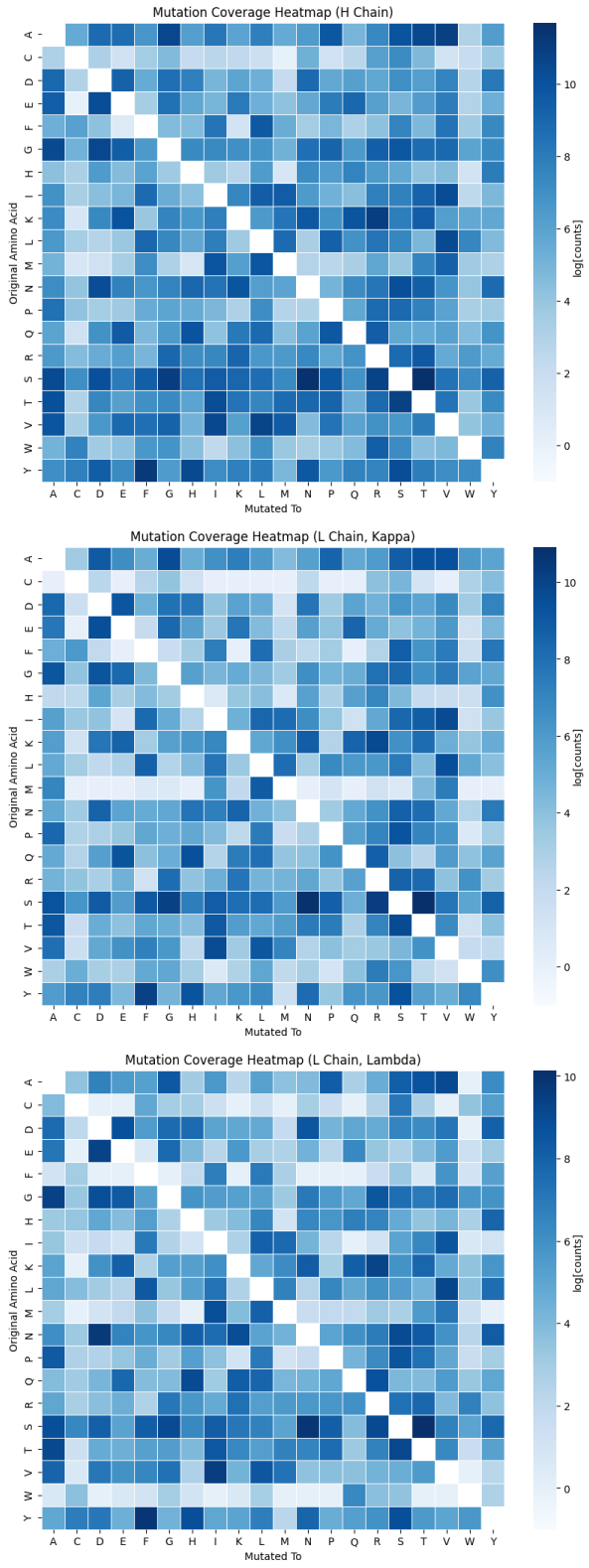


Figure 4: Coverage of substitution mutations across heavy and light chain sequences of human memory B cell receptors in the OAS. The light chain sequences (middle and bottom) were grouped by chain type: Kappa/Lambda

## References


[1] Dmitriy A Bolotin, Stanislav Poslavsky, Igor Mitrophanov, Mikhail Shugay, Ilgar Z Mamedov, Ekaterina V Putintseva, and Dmitriy M Chudakov. Mixcr: software for comprehensive adaptive immunity profiling. *Nature methods*, 12(5):380–381, 2015.

[2] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

[3] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.

[4] Marie Ghraichy, Valentin von Niederhäusern, Aleksandr Kovaltsuk, Jacob D Galson, Charlotte M Deane, and Johannes Trück. Different b cell subpopulations show distinct patterns in their igh repertoire metrics. *Elife*, 10:e73111, 2021.

[5] Dongjun Guo, Deborah K Dunn-Walters, Franca Fraternali, and Joseph CF Ng. Immunomatch learns and predicts cognate pairing of heavy and light immunoglobulin chains. *bioRxiv*, pages 2025–02, 2025.

[6] Brian L Hie, Varun R Shanker, Duo Xu, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, Wesley Wu, John E Pak, and Peter S Kim. Efficient evolution of human antibodies from general protein language models. *Nature biotechnology*, 42(2):275–283, 2024.

[7] Chiho Im, Ryan Zhao, Scott D Boyd, and Anshul Kundaje. Sequence-based tcr-peptide representations using cross-epitope contrastive fine-tuning of protein language models. In *International Conference on Research in Computational Molecular Biology*, pages 34–48. Springer, 2025.

[8] David Jung and Frederick W Alt. Unraveling v (d) j recombination: insights into gene regulation. *Cell*, 116(2):299–311, 2004.

[9] Hamish W King, Nara Orban, John C Riches, Andrew J Clear, Gary Warnes, Sarah A Teichmann, and Louisa K James. Single-cell analysis of human b cell maturation predicts how antibody class switching shapes selection dynamics. *Science immunology*, 6(56):eabe6291, 2021.

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

[11] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[13] Artem Mikelov, Evgeniia I Alekseeva, Ekaterina A Komech, Dmitry B Staroverov, Maria A Turchaninova, Mikhail Shugay, Dmitriy M Chudakov, Georgii A Bazykin, and Ivan V Zvyagin. Memory persistence and differentiation into antibody-secreting cells accompanied by positive selection in longitudinal bcr repertoires. *Elife*, 11:e79254, 2022.

[14] Artem Mikelov, George Nefediev, Alexander Tashkeev, Oscar L Rodriguez, Diego Aguilar Ortmans, Valeriia Skatova, Mark Izraelson, Alexey N Davydov, Stanislav Poslavsky, Souad Rahmouni, et al. Ultrasensitive allele inference from immune repertoire sequencing data with mixcr. *Genome Research*, 34(12):2293–2303, 2024.

[15] Karenna Ng and Bryan Briney. Focused learning by antibody language models using preferential masking of non-templated regions. *Patterns*, 2025.

[16] Valerie H Odegard and David G Schatz. Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6(8):573–583, 2006.

[17] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[18] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40(11):btae618, 2024.

[19] Ganesh E Phad, Dora Pinto, Mathilde Foglierini, Murodzhon Akhmedov, Riccardo L Rossi, Emilia Malvicini, Antonino Cassotta, Chiara Silacci Fregni, Ludovica Bruno, Federica Sallusto, et al. Clonal structure, stability and dynamics of human memory b cells and circulating plasmablasts. *Nature immunology*, 23(7):1076–1085, 2022.

[20] Akshaya Ramesh, Ryan D Schubert, Ariele L Greenfield, Ravi Dandekar, Rita Loudermilk, Joseph J Sabatino Jr, Matthew T Koelzer, Edwina B Tran, Kanishka Koshal, Kicheol Kim, et al. A pathogenic and clonally expanded b cell transcriptome in active multiple sclerosis. *Proceedings of the National Academy of Sciences*, 117(37):22932–22943, 2020.

[21] Zizhang Sheng, Chaim A Schramm, Rui Kong, NISC Comparative Sequencing Program, James C Mullikin, John R Mascola, Peter D Kwong, and Lawrence Shapiro. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Frontiers in immunology*, 8:537, 2017.

[22] Richard W Shuai, Jeffrey A Ruffolo, and Jeffrey J Gray. Iglm: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989, 2023.

[23] Rohit Singh, Chiho Im, Yu Qiu, Brian Mackness, Abhinav Gupta, Taylor Joren, Samuel Sledzieski, Lena Erlach, Maria Wendt, Yves Fomekong Nanfack, et al. Learning the language of antibody hypervariability. *Proceedings of the National Academy of Sciences*, 122(1):e2418918121, 2025.

[24] Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

[25] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] Yiquan Wang, Huibin Lv, Qi Wen Teo, Ruipeng Lei, Akshita B Gopal, Wenhao O Ouyang, Yuen-Hei Yeung, Timothy JC Tan, Danbi Choi, Ivana R Shen, et al. An explainable language model for antibody specificity prediction using curated influenza hemagglutinin antibodies. *Immunity*, 57(10):2453–2465, 2024.

[27] Perry T. Wasdin, Alexandra A. Abu-Shmais, Michael W. Irvin, Matthew J. Vukovich, and Ivelin S. Georgiev. Negative binomial mixture model for identification of noise in antibody–antigen specificity predictions from single-cell data. *Bioinformatics Advances*, 4(1):vbae170, 2024.

# A Appendix

## A.1 Pre-processing of Single-cell RNA Sequencing data

Raw paired-end FASTQ files (Read 1 and Read 2) generated from 10x Genomics single-cell immune profiling experiments were processed using MiXCR v4.7.0 with the `10x-sc-xcr-vdj` preset. This preset is optimized for 10x Genomics $5'$ V(D)J single-cell sequencing libraries and leverages the known read structure in which Read 1 contains cell barcodes and Unique Molecular Identifiers (UMIs), and Read 2 contains the full-length cDNA sequences spanning the immune receptor variable regions. MiXCR performs demultiplexing of single-cell data by parsing cell barcodes embedded in Read 1 and associating each read with its corresponding cell of origin. This enables the grouping of all reads derived from the same cell, allowing for accurate pairing of native receptor heavy-light chains. Following barcode parsing, MiXCR performs alignment of Read 2 sequences against a reference database of V, D, J, and C gene segments (derived from the IMGT database) using a modified Smith-Waterman algorithm [24]. This algorithm performs local pairwise alignment while accounting for common recombination events and somatic hypermutations (SHMs) typically observed in adaptive immune receptors.

In this study, we specifically considered only substitution-based mutations located within the V gene region. Insertions and deletions, which constitute less than 10% of all observed mutations, were excluded from analysis due to their relative rarity and the added complexity they introduce in defining mutation positions. Furthermore, only mutations in the V gene were retained because both D and J gene segments contribute to CDR3, whose boundaries are imprecise due to random exonuclease activity at the junctions during V(D)J recombination [8]. Such exonuclease-mediated trimming of the ends of V and J gene segments complicates accurate mutation mapping in CDR3, making it challenging to distinguish somatic hypermutations from germline-encoded variation in these regions.

## A.2 Antigen-specific B cell Sorting and single-cell RNA Sequencing

Magnetically isolated peripheral blood and stimulated splenic B cells were stained with a set of antigen proteins, including hemagglutinins (HA) from the following influenza strains: A/Victoria/2570/2019 (H1-Victoria), A/California/07/09 (H1-California), A/Tazmania/503/2020 (H3-Tazmania) and A/Darwin/9/2021 (H3-Darwin), each carrying a fluorophore label and DNA tag, containing a barcode sequence indicating the antigen, and a unique molecular identifier (UMI) – a stretch of random nucleotides labeling each individual molecule. To increase specificity, an additional set of the same antigen proteins, but carrying a different fluorophore was added to the cells. Then cells were stained with a panel of fluorescently labeled antibodies specific to cell surface markers in order to distinguish cell populations. Antigen-binding MBCs were isolated using fluorescence activated cell sorting on BD FACSAria Fusion (BD Biosciences) based on being positive for both antigen-associated fluorophores and surface markers. Isolated cells were single-cell sequenced using Chromium Next GEM Single Cell 5' Reagent Kits v2 with human BCR amplification and Feature Barcode technology (10x Genomics). Sequencing data was processed using Cell Ranger v.7.0.1 using multi pipeline, allowing single cell gene expression recovery and UMI enumeration for each of the DNA tags associated with the antigens, resulting in the raw counts characterizing antigen binding. VDJ annotation was performed using MiXCR v4.7.0 using a built-in preset `10x-sc-xcr-vdj` and merged with the Cell Ranger output using cell barcodes.

| MODEL | H1-CALIFORNIA | H3-TAZMANIA |
|---|---|---|
| SHIVER | 0.830 | 0.746 |
| SHIVER (FINE-TUNED) | 0.831 | 0.748 |

Table A.3: Comparison of Flu antigen binding predictions of splenic B cell dataset using SHIVER as is (top), supervised fine-tuned on peripheral blood mBCR sequences labeled with binding affinities to the Flu antigens (bottom).