Reviewed Preprint
v1 • January 5, 2026
Not revised

Reviewed Preprint
v2 • March 5, 2026
Revised by authors

For correspondence:
matsen@fredhutch.org

Competing interests: No competing interests declared

Funding: See page 25

Reviewing editor: Anne-Florence Bitbol, Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

© 2026, Matsen et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Separating selection from mutation in antibody language models

Frederick A Matsen[1,2,3,4] ✉, Will Dumm[1], Kevin Sung[1], Mackenzie M Johnson[1], David Rich[1], Tyler Starr[5], Yun S Song[6], Julia Fukuyama[7], Hugh K Haddox[1]

[1]Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, United States • [2]Department of Genome Sciences, University of Washington, Seattle, United States • [3]Department of Statistics, University of Washington, Seattle, United States • [4]Howard Hughes Medical Institute, Seattle, United States • [5]Department of Biochemistry, University of Utah, Salt Lake City, United States • [6]Computer Science Division and Department of Statistics, University of California, Berkeley, Berkeley, United States • [7]Department of Statistics, Indiana University, Bloomington, United States

eLife Assessment

This **fundamental** study introduces a new biology-informed strategy for deep learning models aiming to predict mutational effects in antibody sequences. It provides **convincing** evidence that separating selection from the nucleotide-level mutation process improves performance over the objectives of protein language models inspired by natural language processing. This paper should be of interest to computational immunologists, but also to the broader community interested in deep learning for biological sequence data and evolution.

https://doi.org/10.7554/eLife.109644.2.sa4

Abstract

Antibodies are encoded by nucleotide sequences that are generated by V(D)J recombination and evolve according to mutation and selection processes. Existing antibody language models, however, focus exclusively on antibodies as strings of amino acids and are fitted using standard language modeling objectives such as masked or autoregressive prediction. In this paper, we first show that fitting models using this objective implicitly incorporates nucleotide-level mutation processes as part of the protein language model, which degrades performance when predicting effects of mutations on functional properties of antibodies. To address this limitation, we devise a new framework: a Deep Amino acid Selection Model (DASM) that learns the selection effects of amino-acid mutations while explicitly factoring out the nucleotide-level mutation process. By fitting selection as a separate term from the mutation process, the DASM exclusively quantifies functional effects: effects that change some aspect of the function of the antibody. This factorization leads to substantially improved performance on standard functional benchmarks. Moreover, our model is an order of magnitude smaller and multiple orders of magnitude faster to evaluate than existing approaches, as well as being readily interpretable.

Introduction

Antibodies are remarkable molecules that can bind essentially any target with high affinity and specificity. Natural antibodies are generated through V(D)J recombination, and refined by somatic hypermutation and affinity-based selection in germinal centers. Antibodies are also important drugs, and improving binding and other important properties is thus an area of active research. A major goal in the field is to predict the effect of changing one amino acid for another at a given site of an antibody, both for antibody engineering and for understanding naturally occurring antibodies.

Antibody "foundation" language models are trained on large datasets of naturally occurring antibody sequences to assist with this and related problems [1, 2]. These models are trained using the masked objective, in which a site is masked from an antibody amino-acid sequence, and the model is trained to predict the masked amino acid given the remaining sequence. Recent models of this type have hundreds of millions or billions of parameters, and are trained on around a billion sequences [3].

While the masked objective has been very useful for modeling human language [4], this approach may not be ideal for learning functional effects of mutations to antibody sequences. To understand why, it is useful to consider: what could a language model learn in order to succeed under the masked modeling objective? It could memorize the germline genes [5,6] and learn about the probabilities of V(D)J recombination. It could learn the codon table, as according to this table some amino-acid mutations are much more likely than others. It could learn rates of somatic hypermutation, because codons containing mutable nucleotide sites are more likely to deviate from germline than those in sites that are less mutable [7]. It could also learn about the impact of amino acid mutations on antibody function through natural selection in the course of affinity maturation, which is the desired signal [8]. However, this desired signal is confounded by the preceding factors.

In this paper, we first demonstrate that masked language models learn all these factors shaping antibody sequences, despite most being irrelevant for functional prediction. Indeed, we show that conflating mutation and selection processes degrades performance on functional prediction tasks. We then develop a model that explicitly accounts for phylogenetic relationships, the codon table, and somatic hypermutation patterns, allowing it to focus exclusively on functional effects.

Our approach separates mutation and selection processes by encoding functional effects in a Deep Amino acid Selection Model (DASM) while explicitly modeling mutation using a separate fixed model trained on neutrally evolving data [9]. This fixed model uses convolutions on 3-mer embeddings to de-liver wide context sensitivity without needing a large number of parameters: the variant we use has around the same number of parameters as the classic S5F [10] 5-mer model. The DASM, trained on substantially less data, outperforms AbLang2 [5] and general protein language models including ESM2 [11,12] and ProGen2-small [13]. This outperformance holds on the largest benchmark datasets of the FLAb collection [8] and on recent high-throughput binding as-says [14, 15]. Unlike existing models, DASMs process complete sequences in a single pass and directly output selection factors for all possible mutations. DASMs are thus orders of magnitude faster than existing models, enabling in-tensive use on a laptop without a GPU. We provide a paired (heavy and light chain) model with open weights and associated code as part of our netam package *https://github.com/matsengrp/netam*.

Results

Antibody language models are biased by nucleotide-level mutation processes

We first sought to understand the extent to which processes such as neutral mutation rate and the codon table influence antibody language model prediction at masked sites. To do so, we used AbLang2 [5] as a case study. It has been implemented with more consideration of biology than other models, as it distinguishes between germline-encoded and non-germline-encoded sites in its training. To start, we examined a single naive BCR sequence obtained from a recent study that performed deep sequencing of human antibody repertoires [17]. We iterated over each site in this antibody's protein sequence, masked the site's amino acid, and used AbLang2 to compute the likelihood of each of the 19 alternative amino acids at that site, then normalized to get a probability. Next, using the antibody's nucleotide sequence, we split the alternative amino acids at a given site into two groups: those that can be encoded by a codon a single nucleotide change away from the original codon, and those that require multiple mutations.

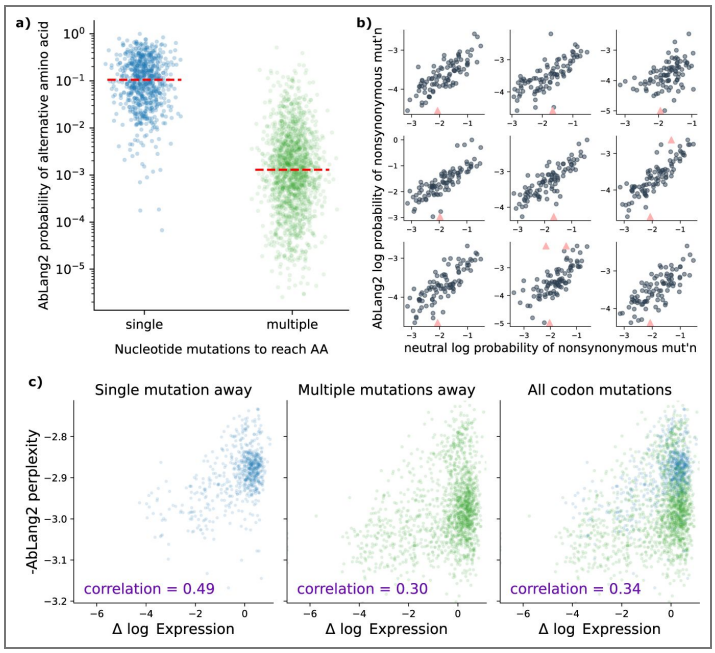


Figure 1. Nucleotide-level mutation processes distort protein language model predictions.

a: AbLang2 assigns almost 100× lower probabilities to amino acids requiring multiple nucleotide mutations compared to single-mutation variants. Each point represents a possible amino acid substitution at a single site in the amino acid sequence. **b**: AbLang2 probabilities correlate with neutral somatic hypermutation probabilities across the V-encoded portion of nine naive sequences, demonstrating how the model is strongly impacted by mutation bias. Each point represents a site in the sequence. Triangles are outliers that have been brought into the y range. **c**: AbLang2 functional prediction accuracy drops substantially for amino acids that are multiple (2 or 3) nucleotide mutations away from the wildtype codon. Data from [16].

We found a striking difference in the AbLang2 amino-acid probabilities (Figure 1a). The median probability for amino acids requiring multiple mutations to a codon was almost two orders of magnitude lower compared to those that only required one. This bias is consistent with the hypothesis that the AbLang2 predictions are influenced by mutation processes unrelated to selection.

The bias can be explained by the masked training procedure. It has already been established that language models trained with the masked objective memorize germline sequences [5, 6]. During somatic hypermutation (SHM), single-nucleotide-per-codon mutations to the germline sequence are much more likely than multi-nucleotide codon mutations. As a result, when predicting relative probabilities of alternative amino acids, language models would be expected to assign much higher probabilities to alternative amino acids that only require a single-nucleotide codon mutation (Figure 1a).

Given this observation, we also hypothesized that AbLang2 probabilities are influenced by differences in the rate of SHM between sites. SHM is a purpose-built enzymatic process that introduces mutations into BCR-coding DNA [18], which occurs independently of the process of natural selection in the germinal center. The rate of SHM varies by an order of magnitude or more between sites, and these biases are well characterized using probabilistic models [9, 10]. To test our hypothesis, we examined nine arbitrarily selected sequences from the Rodriguez *et al.* dataset [17]. For each sequence, we used a recent model of SHM [9] to compute per-nucleotide-site rates, and then used the strategy from our previous paper [19] to convert these rates into per-codon-site probabilities of nonsynonymous mutations.

We found a clear correlation between these neutral SHM probabilities and probabilities of mutations estimated using AbLang2 (Figure 1b). The Wald test for a nonzero slope reported a p-value below machine precision in each case.

We next sought to understand if these biases distorted predictions on a functional prediction task. To do so, we used the largest dataset [16] of the FLAb [8] collection of benchmarks, which measures the effect of single mutations on ex-pression in a phage display library. We scored each single-mutation variant using a score that, up to a linear transformation, is the difference between the log probability the model assigns to the variant amino acid and the log probability it assigns to the wild-type amino acid at that position (see masked-marginals pseudo-perplexity in Methods). We then computed the Pearson correlation between these AbLang2 scores and the experimentally measured expression effects, splitting by number of nucleotide mutations per codon as before.

We found a significant drop (from a correlation of 0.49 to 0.30) in predictive performance when going from amino acids that required one mutation to those that required multiple mutations (Figure 1c). Furthermore, when we considered all amino acids together, the correlation remained at 0.34. When we colored these data points by their probability of mutation under an SHM process (Figure S1), we can clearly see that the data points are spread out by the AbLang2 model according to their SHM rate, hindering functional prediction.

In summary, we found that nucleotide-level effects hamper the ability of AbLang2 to predict functional effects of mutations. We then developed an alternative modeling framework that directly factors out nucleotide-level effects.

Fitting a deep amino acid selection model (DASM)

We implemented a model to learn amino-acid preferences of antibodies without being influenced by germline genes, the codon table, or SHM biases (Figure 2). To do so, we extended our previous work estimating a single selection value for every site [19] to estimating a value for every alternative amino acid at every site. This extension is analogous to going from a dN/dS type estimate [20,21] to a sitewise mutation-selection model [22]. Our previous model only operated on antibody heavy chains. This new model operates on heavy chains, light chains, or heavy-light pairs.

As in this previous work, we performed clonal-family clustering, phylogenetic inference, and ancestral sequence reconstruction to generate collections of nucleotide "parent-child pairs" or PCPs (Figure 2a). This resulted in around 2 million PCPs that were used for training (Table S1). Instead of predicting the likelihood of nonsynonymous mutations, we predicted the likelihood of codon mutations using a deep-learning analog of a mutation-selection model. The neutral-probability model was inferred separately using out-of-frame data [9]. We also added a "multihit correction" (see Methods) to this neutral model which accounts for the spatial clustering of SHM [23], resulting in an elevated probability of multiple mutations in a given codon.

The selection component of the model estimates per-site per-amino-acid se-lection factors given an input amino-acid sequence via a transformer-encoder (Figure 2b). We trained DASMs of several sizes (~1M, ~4M, ~7M) using joint optimization of branch length t and parameters of the DASM (see Methods for details). We found that the ~4M parameter model performed the best according to our objective function. This DASM has 8 attention heads, 32 dimensions per head, a feedforward dimension of 1024, 5 transformer layers, and a dropout probability of 0.1; we will use this version for the rest of the paper. Hyperparameters other than model size were carried over from our previous work [19].

As an initial comparison between DASM and AbLang2, we used DASM to predict selection factors for the unmutated antibody sequence used in the Koenig benchmark [16]; we found that the selection factors were predictive of functional measurements irrespective of the number of mutations per codon (Figures 3a and S2). We were surprised by the strength of the correlation for amino-acid mutations that require multi-nucleotide codon mutations, given that training signal should be weaker for those rarer mutations. We also found that the selection factors showed very similar distributions for amino-acid mutations re-quiring single- vs. multi-nucleotide codon mutations (Figure S3). Furthermore, the DASM selection factors captured the alternating pattern of selective constraint on beta sheets evident in the expression data (red columns in Figures 3b and S4).

Thrifty+DASM accurately predicts affinity maturation

Next, we evaluated the ability of the DASM to predict the course of natural affinity maturation. To do so, we used PCPs from the Rodriguez data (Table S1), which were not used for training the DASM, nor for training AbLang2. For each PCP, we quantified each model's ability to predict both the location of nonsynonymous mutations in the child sequence and the identity of the mutant amino acid. For the DASM, we computed a mutation's probability by multi-plying the mutation's selection factor, predicted by the DASM, with its neutral mutation probability, predicted using the same fixed neutral mutation model used to train the DASM [9].

We found that the DASM-based approach was better than AbLang2 at predicting the location of nonsynonymous mutations observed in the PCPs (Figure S5ab). In fact, the overlap between observed and expected is substantially better than what we were able to obtain in a highly controlled mouse experiment [24] with a customized mutation model and a deep mutational scan that quantified how mutations affect binding to the relevant antigen [25].

We also found that the DASM-based approach was better than AbLang2 at predicting the identities of mutant amino acids. To compare the models at this task, we used a notion of "conditional perplexity". Specifically, for each PCP where the parent and child sequences differ on the amino-acid level, we calculated the perplexity of the residues in the child that are different from the parent, conditioning probabilities on there being a substitution (see condppl definition in Methods). We do this to give AbLang2 the best chance of succeeding: although it might have difficulty predicting the occurrence of mutations (Figure S5ab) it might still be able to predict the identity of mutations.

Indeed, the DASM had lower conditional perplexity than AbLang2 on 1,000 sequences from the Rodriguez [17] dataset (Figure S5c). The median value for DASM was 4.88, compared to 7.39 for AbLang2. In addition to having a lower median, the DASM also has fewer very large values.

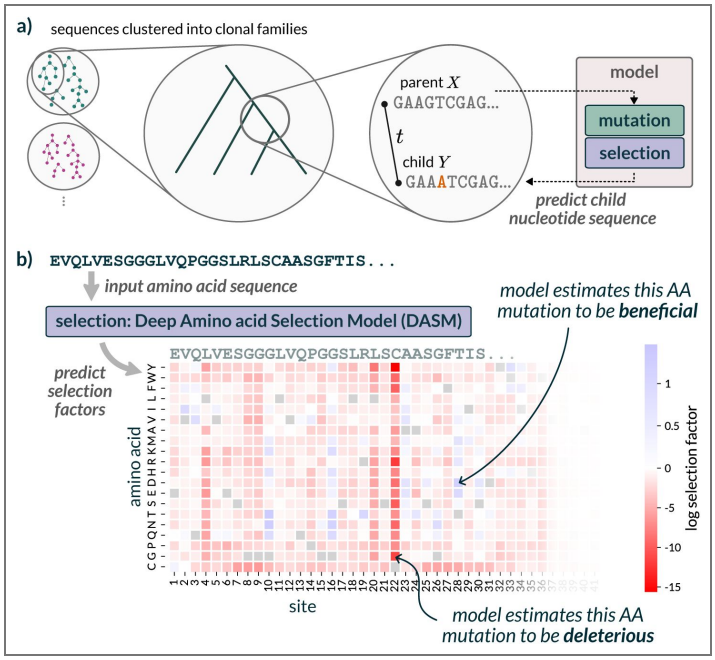


Figure 2. Our model separates mutation from selection to predict functional effects without nucleotide-level biases.

a: Our model combines a fixed mutation component (trained on non-functional data) with a learned selection component (DASM transformer). Training uses inferred parent-child sequence pairs from reconstructed B cell phylogenies to predict natural affinity maturation after a jointly-inferred time *t*. **b**: The DASM directly predicts selection factors for all amino acid substitutions at every position in a single forward pass. Postive factors indicate beneficial changes, and negative factors indicate deleterious changes.

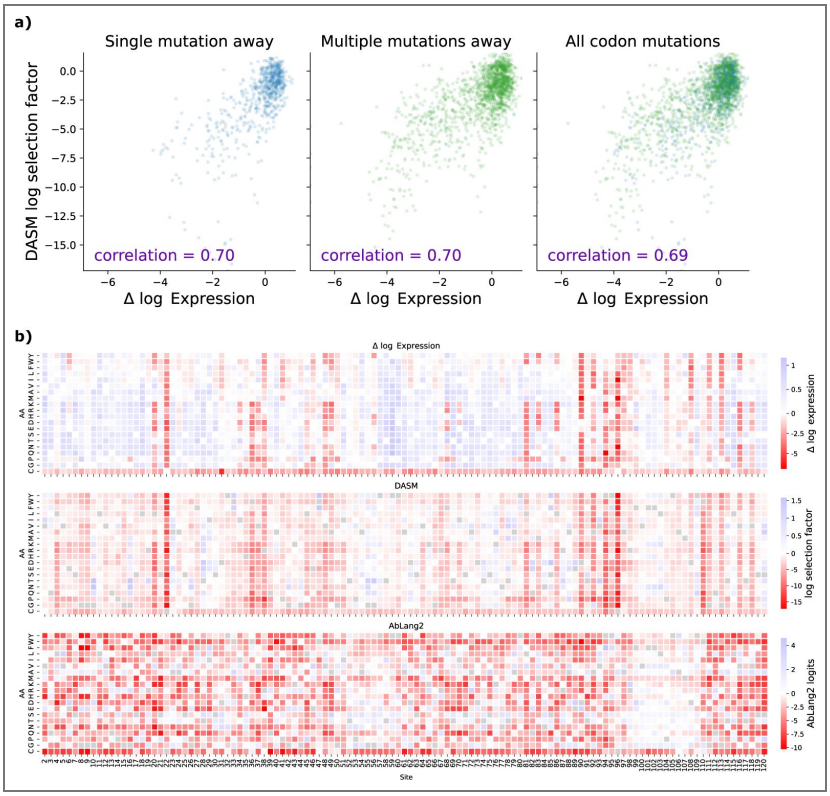


Figure 3. Comparing model predictions with experimentally measured effects of mutations on antibody expression from [16].

a: The DASM maintains high predictive accuracy on functional effects of mutations regardless of codon accessibility. The correlation is equally high for amino-acid mutations that only require a single-nucleotide mutation (left plot) vs. amino-acid mutations that require multi-nucleotide mutations (center plot), demonstrating successful separation of mutation bias from functional effects. Compare Figure 1c. **b**: DASM predictions mimic patterns in the expression data. For additional heatmap comparisons see Figure S4.

The DASM outperforms masked and autoregressive models at predicting functional effects of mutation

We next performed a more comprehensive characterization of the ability of models to predict mutational effects. Because the DASM is trained exclusively on evolutionary trajectories rather than functional measurements, evaluation on expression and binding benchmarks is strictly zero-shot with no risk of data leakage. We expanded our suite of comparator models to include ESM2 [11, 12], a general protein language model trained with a masked objective, and ProGen2 [13], an autoregressive language model. We used the small general variant of ProGen2, which the FLAb benchmarking paper found to be the best overall model [8], and the 650M variant of ESM2. To predict the plausibility of mutant sequences under these models, we followed the FLAb protocol and calculated each sequence's pseudo-perplexity for AbLang2 [5] and ESM2 [26], scoring each variant in its own sequence context (Methods), and each sequence's perplexity for ProGen2, then compared the model outputs to experimentally measured mutational effects using Pearson correlation. To predict mutational effects using the DASM, we simply fed the DASM with the unmutated sequence from the DMS experiment and used the output selection factors as predicted effects. We cannot use perplexity because selection factors alone do not give likelihoods.

We based our evaluations on the April 2024 version of the FLAb benchmarking project [8], which preceded our work and thus was not subject to selection bias by us. We also benchmarked high-throughput binding data (more recent than FLAb) from the Whitehead lab [14,15] that provided affinity measurements across many variants and antigens.

To start, we evaluated these models on the two largest datasets from the FLAb collection. The first data set, from Koenig et al. [16], is a deep mutational scanning experiment performed on the Fab of the anti-VEGF antibody G6.31. The authors generated single-site saturated mutagenesis libraries incorporating all possible amino-acid mutations to both the variable heavy and variable light domains. Using phage display, they then performed two independent selection steps: one in which they selected for Fab variants that were stably expressed, and another in which they selected for Fab variants that could bind VEGF. They quantified mutational effects on these properties by computing enrichment ratios of each mutation in selected versus unselected pools, as quantified using deep sequencing.

We found that the DASM outperformed the other models on both binding and expression (Table 1, Figure S4). Scaling ESM2 from 650M to 3B parameters did not improve performance: indeed, the larger model showed slightly degraded correlations, particularly for light chain predictions (Table S2). This is consistent with recent observations that medium-sized protein language models can outperform larger ones on transfer learning tasks [27]. Thus we only used the 650M model in the rest of our analyses.

We next turned to the second-biggest dataset of the FLAb benchmark, from Shanehsazzadeh et al. [28]. In this study, the authors took an anti-HER2 antibody, trastuzumab, and used deep generative models to redesign the heavy chain complementarity determining regions (HCDRs). They generated libraries of HCDR variants, including both HCDR3-only designs and HCDR123 designs. Unlike the first dataset described above, many variants had multiple mutations relative to the unmutated antibody sequence. The experiments expressed antibody variants in E. coli, and sorted by FACS based on binding signal. We restricted our attention to sequences with the two most common heavy chain amino acid lengths (119 and 120) as the other lengths did not have enough data to make a full comparison.

We assessed the ability of models to predict these binding measurements. For the language models, we computed the perplexity of each variant, as before. For the DASM, we calculated a consensus sequence for each sequence length, and then used the DASM to compute selection factors for that sequence. Then, we computed a score for each variant by summing the log selection factors for each amino-acid mutation in the variant relative to the consensus sequence. As above, the resulting correlation is substantially higher for the DASM compared to the other models (Table 2).

Table 1. Correlation of models with predicting effects of single mutations on antibody expression and antigen binding, as measured in [16].

Masked language models scored using per-sequence pseudo-perplexity following the FLAb protocol [8].

	Binding		Expression	
Model	Heavy	Light	Heavy	Light
AbLang2	−0.114	−0.108	0.153	−0.109
DASM	**0.335**	**0.316**	**0.688**	**0.674**
ESM2	0.009	0.243	0.384	0.416
ProGen2	0.156	0.276	0.559	0.568

Table 2. Correlation of models with binding measurements on the data of [28], which typically involves multi-mutant variants.

See Figure S6 for scatterplots.

Seq. length	AbLang2	DASM	ESM2	ProGen2
119	0.263	**0.458**	0.191	0.074
120	0.166	**0.518**	0.308	0.052

For our third validation, we sought to see how our model could predict affinity in the context of large-scale yeast display experiments using MAGMA-seq [14]. We were especially interested in this assay because the measurement is affinity-only rather than a conflation of affinity and expression as in the previous data sets. Because none of these models are trained in an epitope-specific way this will be a challenge for the models. Specifically, MAGMA-seq determines quantitative monovalent binding affinities through deep sequencing of barcoded Fab libraries sorted at multiple antigen concentrations in a manner analogous to Tite-Seq [29]. These data are not present in the FLAb benchmarks. We used the largest datasets in the original MAGMA-seq paper [14] which test variants of influenza antibodies. We also used MAGMA-seq data from [15] which combinatorially applied mutations from development trajectories of mature human antibodies against SARS-CoV-2 spike protein.

We found that the DASM performed substantially better than other models for this task (Table 3, Figure S7). Nearly all correlation coefficients were small, pointing to the difficulty in predicting effects of mutations on antigen-specific binding affinities. Nevertheless, the DASM was the only model to have a positive correlation for most datasets. The UCA 002-S21F2 lineage proved challenging for all models, with all models having a negative correlation. This may not be surprising given that 002-S21F2 uses a rare VH5-51/VK1-33 gene combination found in only 3 out of 5252 SARS-CoV-2 antibodies in the CoV-AbDab database [30].

DASM models are orders of magnitude faster than competing models, and much smaller

We next compared the computational requirements for doing these evaluations. We imagined two settings: one where a scientist wishes to evaluate a small collection of sequences on their laptop without a GPU, and the other where they wish to evaluate a larger collection of sequences on a GPU server. Our timing script ran the DASM model, as well as AbLang2 and ESM2 using stepwise masking to obtain amino acid likelihoods. The local machine was a MacBook Pro with a M2 Max chip and 96 GB of memory. The GPU server had an AMD EPYC 75F3 32-core processor with 995 GB of memory and an NVIDIA A100 80GB GPU.

The DASM achieved dramatic computational efficiency gains compared to masked language models (Table 4). On a CPU, the DASM evaluates sequences over 1,000 times faster than AbLang2 and more than 10,000 times faster than ESM2. On a GPU, the DASM is 100 times faster than AbLang2 and over 1,000 times faster than ESM2. This speed advantage comes from the DASM providing predictions for all variants in a single pass, eliminating the need for iterative masking procedures required by masked language models. (We note that it is possible to make predictions from masked language models without masking, although this is not considered best practice because of the influence of the unmasked residue.)

This evaluation was actually generous to the masked language models, as it assumes that we are evaluating likelihoods of masked sites and combining them to evaluate multi-mutant variants. That is in contrast to what was actually done in the performance evaluation, which is direct evaluation of the sequence perplexity for each sequence individually (which should be more accurate). Pro-Gen2 was not included in this comparison because it is categorically slower for this use case: autoregressive models for a panel of variants require running each variant individually through the model.

The DASM benchmarked here is also significantly smaller than alternate models. It has 4M parameters, compared to 45M for AbLang2, 650M for ESM2, and 151M for ProGen2. We suspect that this efficiency is enabled by DASM learning selection effects only, rather than selection and mutation-level effects as for existing models.

In summary, we imagine end-users will find it convenient to download a small (23 MB) weights file and run the DASM on their laptop, as compared to existing models requiring more powerful hardware.

Table 3. Correlations between model predictions and binding affinity.

The "Petersen" data is from an experiment probing the rules of recognition for influenza neutralizing antibodies [14] and "Kirby" data is from combinatorial libraries applying combinations of mutations on the path from naive to mature SARS-CoV-2 antibodies [15]. See Figure S7 for corresponding scatter plots.

Source	Antibody	AbLang2	DASM	ESM2	ProGen2
Petersen [14]	222-1C06	0.041	**0.248**	−0.039	−0.030
	319-345	0.079	**0.279**	0.199	0.222
Kirby [15]	002-S21F2	−0.246	**−0.165**	−0.283	−0.353
	Ab_2-15	−0.325	**0.094**	−0.265	−0.069
	C118	−0.624	**0.276**	−0.293	−0.159

Table 4. Computational efficiency comparison on sequences from the MAGMA-seq experiments.

10 sequences were run on CPU, and 100 on the GPU server.

Model	CPU (s/seq)	GPU (s/seq)
DASM	0.0097	0.0053
AbLang2	11.0529	0.5196
ESM2	112.5989	7.6090

Discussion

We have motivated and developed a new direction for protein language modeling. To motivate this approach, we showed how masked language modeling implicitly learned the codon table and somatic hypermutation rates, properties that are orthogonal to antibody function. This incorporation degrades model performance for predicting effects of mutations on antibody expression and binding.

To address this limitation, we introduced a new framework: a Deep Amino Acid Selection Model (DASM), which models selection effects separately from mutation effects. By modeling these components independently, the DASM more accurately predicts the functional consequences of amino-acid mutations while having dramatically reduced computational requirements. The success of our approach highlights the importance of incorporating biological realities into machine-learning models for protein engineering. This aligns with recent calls for deeper dialogue between machine learning and evolutionary biology to address phylogenetic biases in biological foundation models [31].

Thrifty+DASM can be viewed as a neural-network extension of the models rooted in the work of Halpern and Bruno [22]. There, the probability of a codon mutation is expressed as the product of that mutation under a neutral nucleotide process, times a term representing the selection for or against the corresponding amino-acid substitution at that site. Although the original model expressed selection coefficients in a parameter-sparse way using equilibrium frequencies, further model elaborations allowed for per-site-per-amino acid selection factors, as described here, either in a fixed-effects [32, 33] or random-effects [34] framework. Other related work includes developing evolutionary models using deep mutational scanning (DMS) experiments [35,36], and models of sequence-structure compatibility [37]. Thrifty+DASM also extends our previous work that estimates a single value of natural selection at every site [19]; such a model is not comparable to an antibody foundation model because it does not estimate per-amino-acid probabilities.

By separating out selection from mutation and the effect of evolutionary contingency, DASMs provide direct interpretability: the selection factors indicate which amino-acid mutations are beneficial or deleterious at each position. This interpretability could prove valuable for antibody engineering, where under-standing which specific mutations would drive improved expression or binding is crucial for rational design. We have dedicated a companion paper to lever-aging this interpretability to provide new perspectives on the operating rules of affinity maturation [19]: that work provides a nuanced sitewise perspective on natural selection in antibodies that challenges classical oversimplified views of selection patterns. For users wishing to explore the relationship between DASM selection factors and structure for human antibodies in the SAbDAb database [38], we have made interactive 3D visualizations of DASM selection factors using dms-viz [39] available at *https://matsen.group/dasm-viz/v1/*. The current generation of DASM model does not use any antigen-labeled training data. The signal that it leverages to infer some limited ability to predict binding comes from natural affinity maturation. This affinity maturation comes through natural repertoires and so represents a mix of all of the antigens to which the sampled individuals have been exposed.

In future work we will explore fine-tuning of these models with antigen-specific data. While DASM provides better-than-state-of-the-art zero-shot performance on binding benchmarks, recent work for binding focuses on using protein embeddings as inputs to subsequent classifiers [40], or fine-tuning for binding prediction [3, 41]. Because the DASM framework separates mutation from functional properties, a fine-tuned DASM may improve on these previous efforts. We also have plans to extend the DASM framework to estimate the effect of natural selection on insertion and deletion events.

We are beginning the process of extending DASMs to other settings. Al-though antibodies form an interesting first application of DASMs, and have many properties that make DASMs a useful tool, DASMs are not restricted to the antibody case. Viral sequences also exist in abundance and may be amenable to the DASM approach. Rather than many clonal families for antibodies, we will fit a DASM with many separate alignments for related but distinct evolutionary histories.

Taking this idea to its logical extent, we would like to train a DASM model analogous to ESM [26] for all proteins. However, this will be a major computational undertaking. Even setting training such a model aside, doing phylogeny and ancestral sequence reconstruction on all protein alignments will require careful planning. If we followed the steps used here for phylogeny and ancestral sequence reconstruction on the 2.6 million sequence alignments used for the MSA Transformer [42] it would take around 3,000 CPU-years.

DASMs represent a new paradigm for deep models of protein function that leverage evolutionary information directly. Even in this initial implementation, DASMs outperform existing foundation models while requiring orders of magnitude less computation, parameters, and training data. This efficiency suggests that evolutionary structure provides a powerful approach for understanding protein function.

Methods

Data

BCR sequence data was processed with partis [43] to cluster into clonal families and infer germlines. Inferred insertions or deletions were reversed, so that all sequences align to the naive sequence without gaps. We selected clonal families with at least two productive sequences; a sequence is considered productive if the canonical cysteine and tryptophan codons that flank the CDR3 are in the same frame as the start of the V segment (although they can be mutated), and there are no stop codons. We excluded sequences with stop codons. Following the training of other LLMs (e.g. [5]) we also excluded sequences with mutated conserved “signature” cysteines, in contrast to our previous work [19].

As in our previous work, tree inference and ancestral sequence reconstruction were performed per clonal family with the K80 substitution model using the naive sequence as outgroup, allowing mutation rate heterogeneity across sites using a 4-category FreeRate model using IQ-Tree [44]. However, for paired data we used the *edge-linked-proportional* partition model in IQ-Tree that allowed the heavy and light chains to evolve at overall different rates [45]. Because these clonal families are independent these phylogenetic inferences were run in parallel.

Once this was done, we had a set of parent-child pairs (PCPs) that correspond to the pairs of parent and child sequences on the edges of the phylogenetic tree (Figure 2a). We used these PCPs to train the model.

We denote pairs of parent and child sequences as (X, Y), where X is the parent sequence and Y is the child sequence. We use X acid sequence corresponding to X. to denote the amino

Model

Formulation and loss

Assume we are given a parent codon sequence X and a child sequence Y . We will use $\bar{c}$ to denote the amino acid sequence of codon c, and use $\bar{Z}$ to denote the amino acid sequence of codon sequence Z. We will use j to denote codon sites. As before [19], we model the neutral probability of a mutation to codon c as the product of per-nucleotide-site mutation probabilities: specifically, the probability of mutating from the parent codon to codon c is the product of the probabilities of each constituent nucleotide mutation (or non-mutation) under the neutral model. This results in the probability $p_{j,c}(t, X)$ of a mutation to codon c at site j after time t.

The selection term $f_{j,\bar{c}}(\bar{X})$ is a “selection factor” that quantifies the natural selection happening at site j if it was to be changed to $\bar{c}$. If this value is greater than 1, it predicts that the mutation would be beneficial in the course of affinity maturation, while if it is less than 1, it predicts that the

mutation would be deleterious.

We parameterize the DASM f using the standard transformer-encoder architecture [46]: an amino-acid embedding, sinusoidal positional encodings, and PyTorch's TransformerEncoder module. The only non-standard component to this architecture is a custom "wiggle" activation function to the output layer that prevents extreme selection factors as previously described [19]. This function asymptotes to zero for highly deleterious mutations and grows sub-linearly for beneficial ones.

We will now define the likelihood $\ell_{j,c}(t, X)$ of codon c at site j given time t and parent sequence X. When c is a non-wildtype codon it is

$$\ell_{j,c}(t,X) := \begin{cases} p_{j,c}(t,X) f_{j,\bar{c}}(\bar{X}) & \text{if } c \text{ codes for an amino acid} \\ 0 & \text{if } c \text{ codes for stop.} \end{cases} \quad (1)$$

where $f_{j,a}(\bar{X}) := 1$ when a is the amino acid for the wildtype codon. We then take the likelihood of the wildtype codon to be 1 minus the sum of these non-WT codons.

The overall likelihood for a parent-child pair is then

$$\prod_j \ell_{j,y_j}(t,X)$$

where y_j is the jth codon of Y. We optimize this likelihood jointly over the branch lengths t and parameters of the selection model f. This joint optimization is performed cyclically, in which a complete cycle consists of neural network optimization followed by branch length optimization for every parent-child pair. The parent sequence and the child sequence are pre-estimated, fixed, and used as training data. The branch lengths are independent and so are optimized in parallel.

Because $\ell_{j,c}(t,X) = p_{j,c}(t,X)$ when c codes for the wildtype amino acid, this approach effectively fixes the selection factors for neutral substitutions to be 1 for the purposes of branch length optimization. This is useful because it gives us a "gauge" that eliminates a potential unidentifiability in the model.

Especially early in training it can happen that $p_{j,c}(t,X) f_{j,\bar{c}}(\bar{X})$ is greater than one, or that the sum of such terms is greater than one. In these cases, the sum is clamped to be a little less than 1.

Perplexity

Perplexity is the standard way of evaluating the plausibility of a sequence according to a model: it is the across-site geometric mean of the inverse probability of the observed amino acid. These inverse probabilities can be interpreted as the effective number of tokens that the model considers plausible at each position. This is equivalent to the exponential of the average negative log probability of each token given its preceding context:

$$\text{ppl}(x) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log p(x_i \mid x_{<i})\right),$$

where $x = (x_1, x_2, \ldots, x_n)$ is a sequence of n tokens, and $x_{<i}$ represents all tokens preceding position i.

For encoder-only models like ESM2 and AbLang2, which mask tokens rather than using autoregressive prediction, we follow the authors of these models and instead calculate a "pseudo-perplexity" using the probability of each token given all other tokens in the sequence:

$$\text{pseudo-ppl}(x) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n} \log p(x_i \mid x_{\setminus i})\right),$$

where $x_{\setminus i}$ denotes the sequence with token i masked out.

An alternative "masked-marginals" approach scores variants using only wild-type context. For a wild-type sequence w, masked-marginals computes $p(a \mid w_{\setminus i})$ for all amino acids a at each position i once, then uses these wild-type-derived probabilities to compute pseudo-perplexity for any variant x:

$$\text{mm-pseudo-ppl}(x) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log\left(p\left(x_i \mid w_{\setminus i}\right)\right)\right).$$

For a single-mutation variant x that differs from wild-type w only at position j, all terms except position j cancel when comparing to wild-type, giving

$$\begin{aligned}&\log(p(x_j \mid w_{\setminus j})) - \log(p(w_j \mid w_{\setminus j}))\\ &\quad = n\left[\log(\text{mm-pseudo-ppl}(w)) - \log(\text{mm-pseudo-ppl}(x))\right].\end{aligned}$$

Thus, the log-probability difference between variant and wild-type amino acids equals, up to an additive constant n log(mm-pseudo-ppl(w)) that depends only on the wild-type sequence, negative n times the log pseudo-perplexity of the variant.

For Figure 1c on the single-mutant Koenig dataset, we found that this approach gave a substantially higher correlation for AbLang2 and so used it in that figure. For benchmarking comparisons (Table 1), we followed standard practice and used per-sequence pseudo-perplexity.

We also use a notion of "conditional" perplexity: for each parent-child pair (x, y) of sequences that differ on the amino acid level, calculate the perplexity of the residues that are different from the parent, conditioning probabilities on there being an amino acid substitution. That is, if we let $S_{x,y}$be the sites that differ between x and y, then

$$\text{cond-ppl}(x, y) = \exp\left(-\frac{1}{|S_{x,y}|}\sum_{i\in S_{x,y}}\log p(\bar{y}_i \mid x;\ \bar{y}_i \neq \bar{x}_i)\right)$$

where bar represents amino acid sequence as before.

Multihit correction

In the SHM process, having a mutation at a site increases the probability of mutations at nearby sites [23]. Although a general solution to this problem can easily lead down the path of intractability, we wanted to incorporate this phenomenon into our work.

Inspired by the work of [47], we added several simple "multihit" rate multipliers to our neutral model. These multipliers account for the varying probabilities of mutations based on the number of nucleotide changes required.

We found that, on neutrally-evolving out-of-frame data, such a correction substantially improved model fit (Figure S9). Before correction, the model systematically underestimates multi-hit mutations. After training the multihit model with correction factors for 1, 2, and 3 mutations per codon, the observed-expected agreement improved substantially for these classes.

This multihit correction was trained on the same separate out of frame data as used in training our neutral model [9] and was incorporated into the neutral mutation probabilities for DASM training.

Heavy and light rates

Light chains mutate at a lower rate than heavy chains during affinity maturation. To quantify this relative rate, we used IQ-Tree's rate partition feature during phylogenetic inference on paired heavy-light chain data (see Data section). This allowed us to estimate separate substitution rates for heavy and light chains within each clonal family. We found the median relative rate of light to heavy chains across all clonal families to be 0.63, which we used as our fixed light chain rate adjustment parameter. This estimate is broadly concordant with previous observations that light chains evolve at roughly half the rate of heavy chains [48].

This relative rate is incorporated into the likelihood calculation by scaling the neutral mutation rates for light chain sequences by the light chain rate adjustment factor before combining them with selection factors. Specifically, when evaluating paired heavy-light chain sequences, the per-site neutral mutation rates $p_{j,c}(t, X)$ for light chain sites are multiplied by this adjustment factor, while heavy chain rates remain unscaled.

Model implementation, training, and evaluation

This model was implemented in PyTorch 2.5 [49]. Models were trained using the RMSprop optimizer, with 4 cycles, each consisting of branch length optimization then neural network optimization.

We used the following software: Altair [50,51], BioPython [52], Matplotlib [53], pandas [54], pytest [55], Seaborn [56], and Snakemake [57].

LLM model score computation

ESM2 scores were calculated using the FLAb paper methodology, separately evaluating heavy and light chains then averaging the pseudo-perplexities, while AbLang2 directly evaluated paired heavy-light sequences. ProGen2 scores were computed following the same FLAb methodology of separately evaluating heavy and light chains then averaging perplexities.

Koenig evaluation

We obtained the Koenig et al. deep mutational scanning data from the FLAb benchmark repository [8] (commit 67738ee, April 17, 2024). For DASM evaluation, we used the wild-type G6.31 heavy and light chain sequences as input to predict selection factors for each possible amino acid substitution at each site. The model's log selection factors were compared against the experimental log enrichment ratios via Pearson correlation. We evaluated performance separately for heavy and light chains across both binding and expression datasets.

Shanehsazzadeh evaluation

We obtained the Shanehsazzadeh et al. binding affinity data from the FLAb benchmark repository [8] (commit 67738ee, April 17, 2024). We subset to the "zero-shot" component of the dataset. For DASM evaluation, we partitioned the data by heavy chain amino acid length (119 and 120 residues) due to the diverse nature of the designed sequences. For each length group, we computed a site-by-site consensus sequence to serve as the reference sequence for model predictions. The aggregate log selection factor for each variant was calculated as the sum of log selection factors for each position that differed from the consensus sequence. Model performance was evaluated through Pearson correlation analysis between the aggregate log selection factors and experimental binding affinities.

MAGMA-seq evaluation

We combined data from two complementary studies using the MAGMA-seq methodology: Petersen et al. [14] providing CDR-targeted mutagenesis data around mature influenza antibodies and Kirby et al. [15] providing UCA to mature antibody evolution trajectories for SARS-CoV-2 antibodies. All sequence variants were assigned to their corresponding antibody systems using reference matching with sequence similarity thresholds. To ensure data quality, experimental replicates were aggregated using geometric mean in $\log_{10}$ space, and high-variance measurements (coefficient of variation > 0.5) were filtered out as in the original papers. After deduplication and quality filtering, the unified dataset contained 1,128 sequences across 6 antibody systems: 4 Kirby UCAs and 2 Petersen mature antibodies. The 1 20 antibody system was dropped from the Kirby data as it only had 7 assigned sequences. Reference sequences were used as inputs for DASM score calculations. Model performance was evaluated using Pearson correlation with binding affinity ($-\log_{10} K_D$).

Supplementary Materials

purpose	name	samples	clonal families	PCPs	median mutations
train	JaffePairedCC	4	50,776	209,599	7
train	TangCC	21	45,267	651,899	2
train	VanwinkleheavyTrainCC	149	21,269	124,985	4
train	VanwinklelightTrainCC1m	330	2,658	1,000,000	2
test	RodriguezCC	51	3,592	38,050	5

Table S1. Data used in this paper. CC means that PCPs with mutated cysteines are excluded from the data. JaffePairedCC is paired data from [59] sequenced using 10X. TangCC data is heavy chain data from [58, 60] and was sequenced using the methods of [60]. VanwinkleheavyTrainCC1m is a subset of the heavy chain data from [61] sequenced using Takara 5'RACE BCR kit. VanwinklelightTrainCC1m is a 1M subset of the light chain data from [61] sequenced using Takara 5'RACE BCR kit. RodriguezCC data is the 5' RACE heavy chain data from [17] and is used only for testing. The "samples" column is the number of individual samples in the dataset; in these datasets, each sample is from a distinct individual. "Clonal families" is the number of clonal families in the dataset. "PCPs" is the number of parent-child pairs in the dataset. "Median mutations" is the median number of mutations per PCP in the dataset.

Table S2. Comparison of ESM2 model sizes on the Koenig benchmark using the masked-marginals approach.

The larger ESM2-3B model (2.8B parameters) performs slightly worse than the 650M variant, particularly on light chain predictions.

	Binding		Expression	
Model	Heavy	Light	Heavy	Light
ESM2-650M	−0.001	0.308	0.418	0.524
ESM2-3B	−0.025	0.283	0.418	0.469

Figure S1. Conflating mutation and selection hinders functional prediction.

This plot is analogous to the rightmost panel of Figure 1c, but colored by mutability according to the Thrifty [9] model.

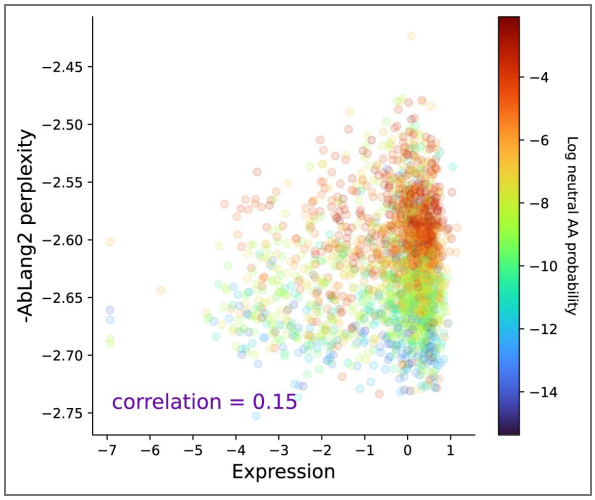


Figure S2. DASM removes codon bias in light chain functional predictions.

Equivalent plot as Figure 3 but for light chain.

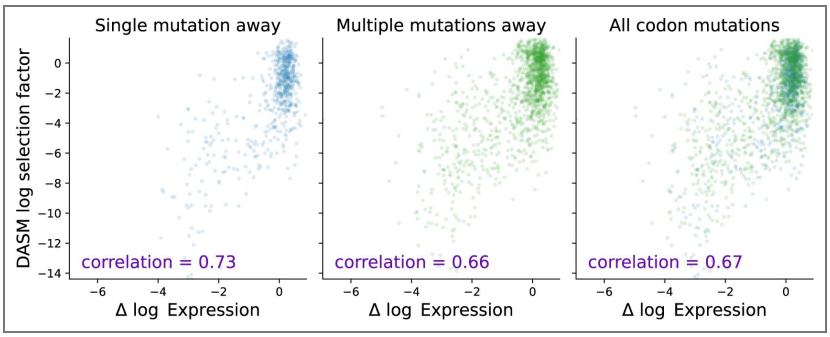


Figure S3. DASM selection factors match the codon-table pattern seen in experimental measurements, while masked language models show artifacts from the codon table.

The experimental data (left two panels) show a slight decrease in median scores for amino acids requiring multiple nucleotide mutations ("multiple") versus single mutations ("single"). DASM captures this pattern, showing similar distributions for both categories. In contrast, AbLang and ESM assign radically lower scores to multi-nucleotide amino acid substitutions, consistent with the masked language modeling objective learning codon-level mutation probabilities as described in the main text (Figure 1a). Comparison done on the heavy chain of the Koenig [16] wildtype sequence.

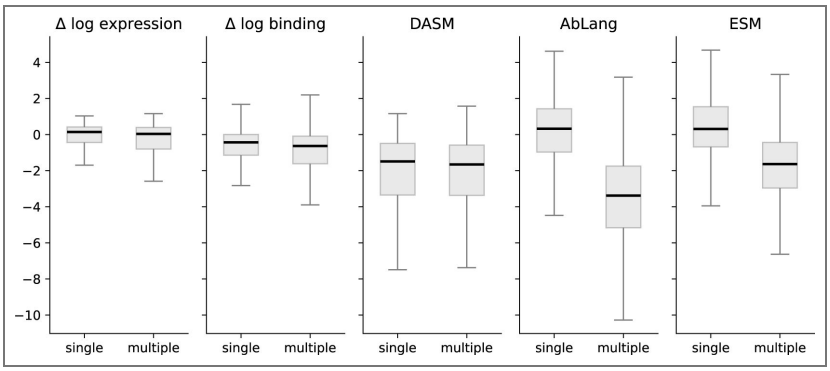
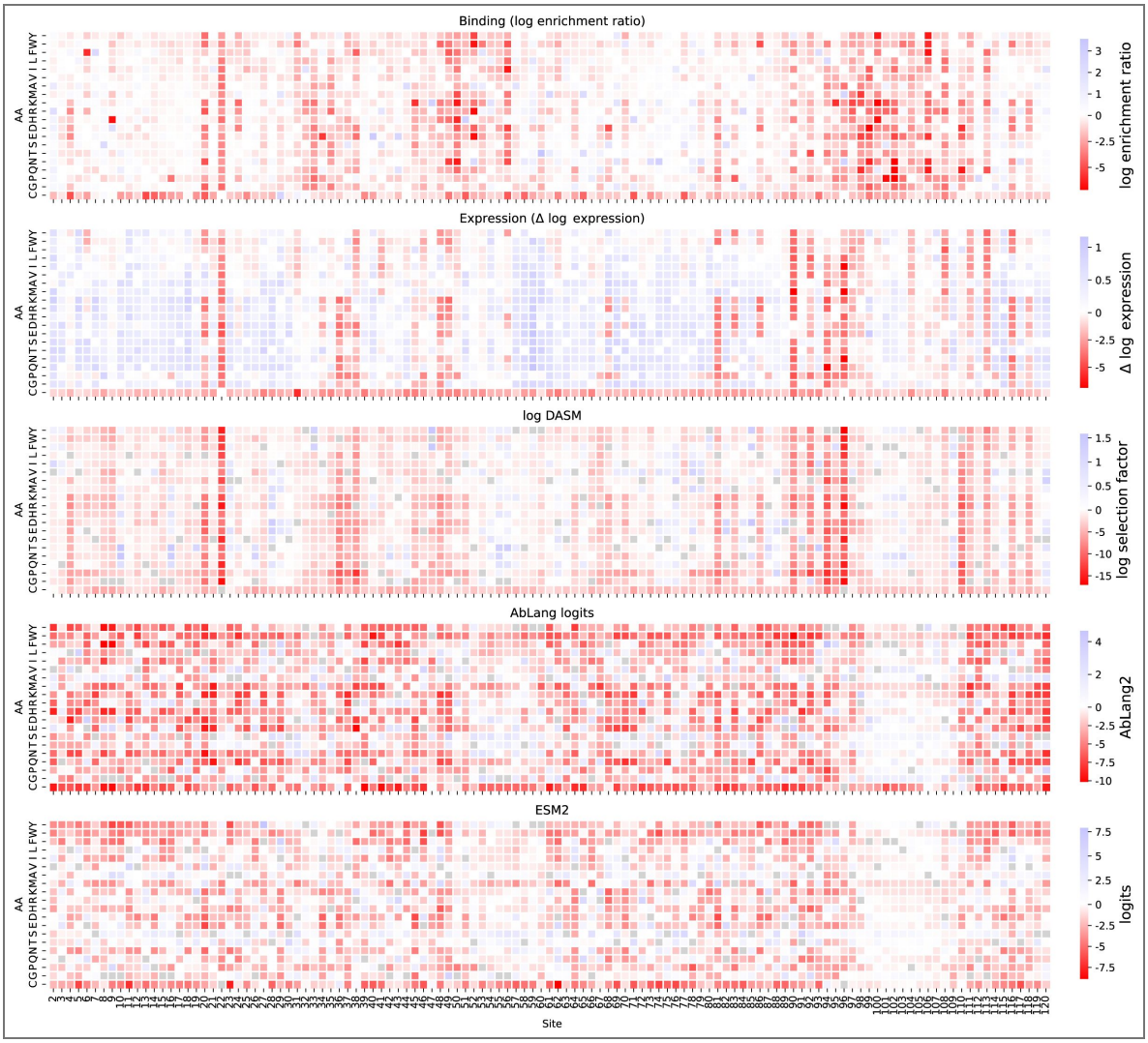


Figure S4. Heatmap showing the heavy-chain data of [16] along with model predictions.



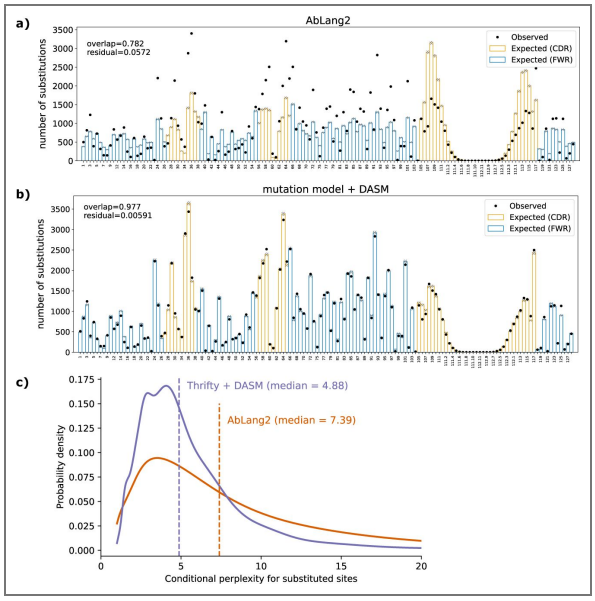


Figure S5. The DASM with the Thrifty neutral model accurately assesses probabilities of natural affinity maturation paths.

a, b: The DASM is better than AbLang2 at predicting the location of nonsynonymous mutations observed in PCPs withheld from model training. c: The DASM achieves lower conditional perplexity (median 4.88 vs 7.39) when predicting amino acid identity at mutated sites, with fewer extreme outliers. The conditional perplexity is the perplexity of the child amino acid, conditioned on there being a mutation at that site. Note that due to the inherent stochasticity of affinity maturation, there is a lower limit to this conditional perplexity that is substantially greater than 1.

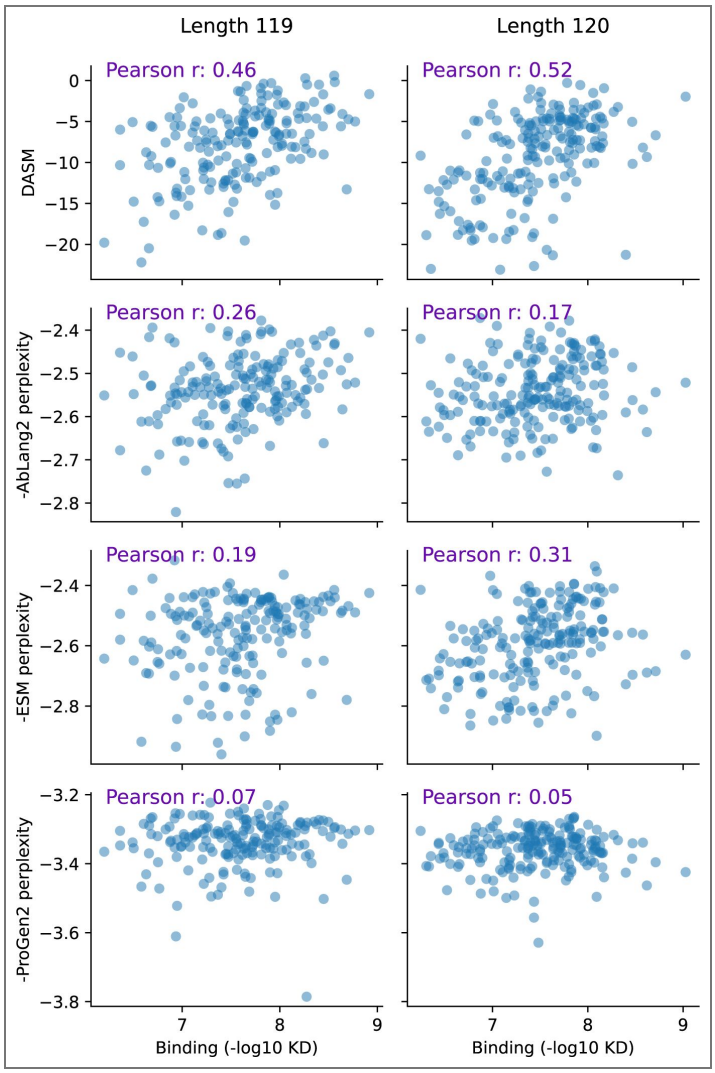


Figure S6. **Scatterplots for the data of [28] zero-shot data set, partitioned by sequence length.**

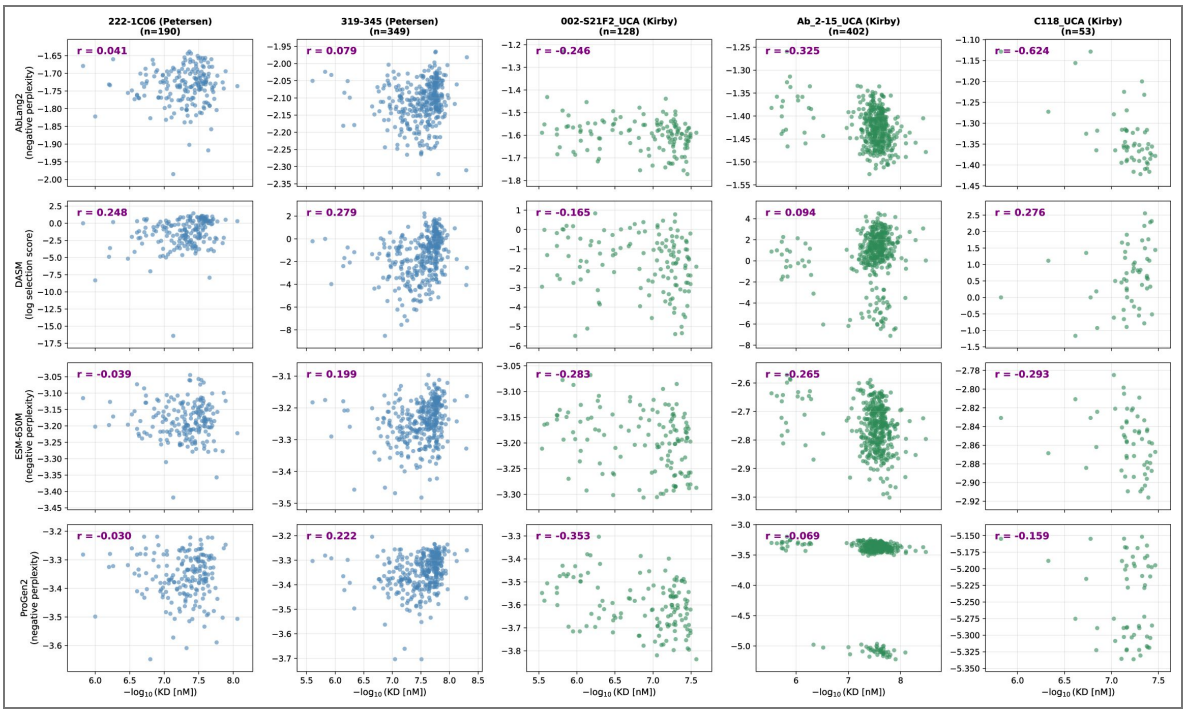


Figure S7. Model predictions versus experimentally measured binding affinity for five antibodies using the MAGMA-seq protocol.

"Petersen" data are from an experiment probing the rules of recognition for influenza neutralizing antibodies [14] and "Kirby" data are from combinatorial libraries applying combinations of mutations on the path from naive to mature [15]. Each row shows a different model (DASM, ESM2, AbLang2, ProGen2) and each column shows a different antibody lineage.

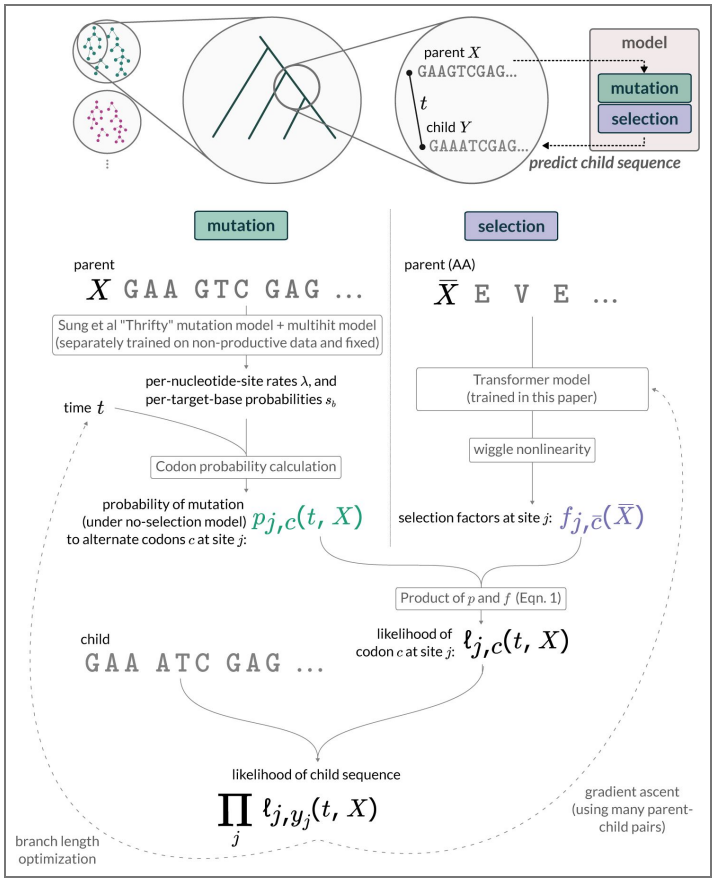


Figure S8. The model is trained to predict the probability of a child sequence given the parent sequence.

It is divided into mutation and selection components. Mutation: given a parent sequence *X*, probability of mutation to alternate codons after time *t* is calculated using a model of SHM [9] and a "multihit model" (see Methods) and then aggregated into codons as in [19] to obtain $p_{j,c}(t, X)$. Selection: given an amino acid translation $\bar{X}$ of the parent, the transformer-encoder gives per-site selection factors $f_{j,\bar{c}}(\bar{X})$. These are then multiplied (1) and summed to give the probability of the observed child sequence at every site. This gives a likelihood for a parent-child pair. The algorithm maximizes the likelihood across branch lengths *t* for each parent-child pair as well as across the parameters of the transformer model for all parent-child pairs in the dataset (dashed lines).

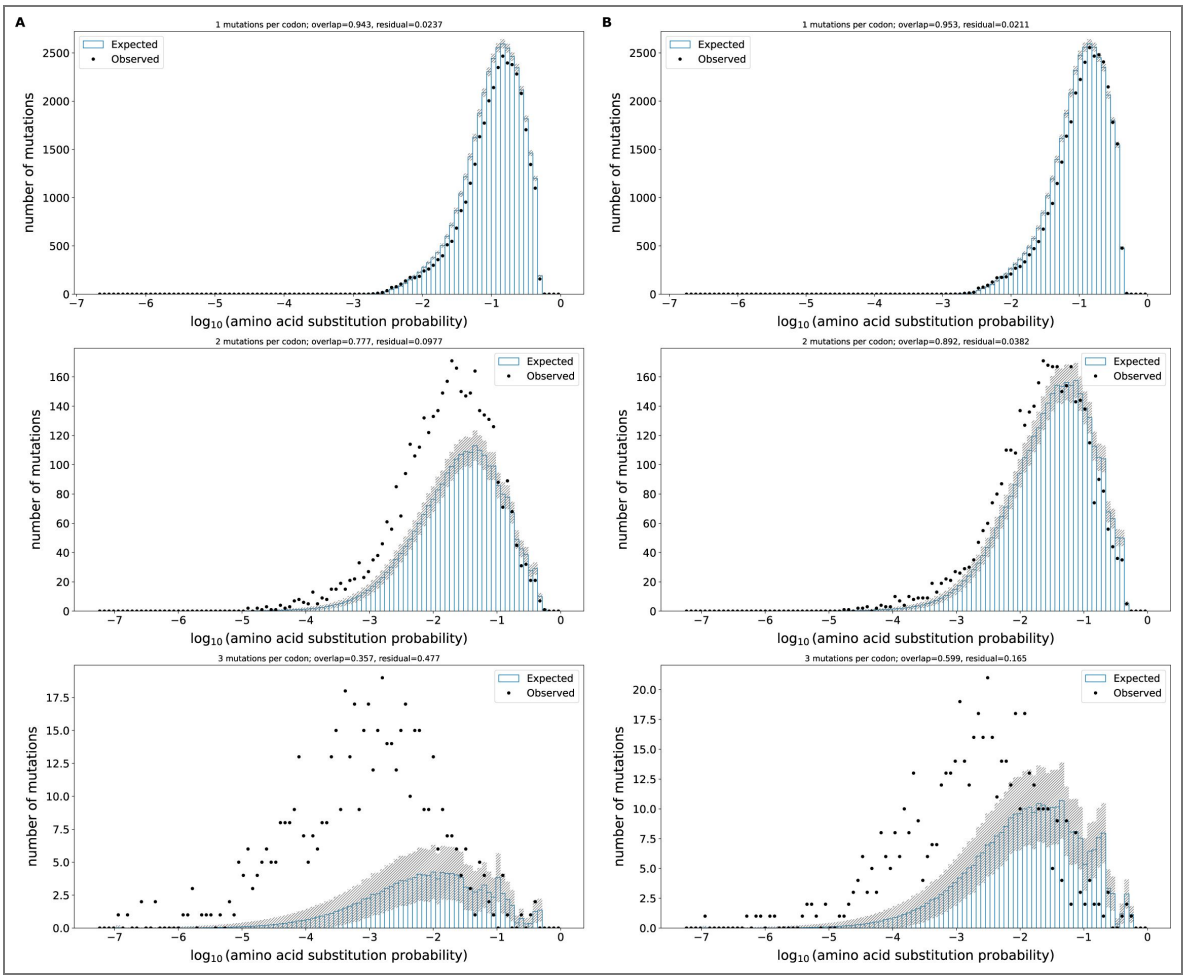


Figure S9. Models that allow for multiple hits per codon (right column) show better model fit than without (left column) on the out-of-frame data from [58] as prepared in [9].

These observed-versus-expected plots compare the observed number of mutations to the computationally predicted number across bins of mutation probability. For each bin, the expected number of mutations is calculated as the sum of mutation probabilities for all sites in that bin, while the observed count represents the actual mutations found in those sites. The overlap metric quantifies the area of overlap between observed and expected distributions divided by their average area, while the residual metric measures the normalized root-mean-square difference between observed and expected counts. These plots are faceted by the number of mutations per codon.

Data availability

Models and inference code can be found at *https://github.com/matsengrp/netam*, including a simple means of accessing the pretrained model demonstrated in the notebooks/dasm_demo.ipynb notebook. Our reproducible experiments are available at *https://github.com/matsengrp/dasm-experiments*. Relevant preprocessed data has been uploaded to Zenodo at *https://doi.org/10.5281/zenodo.17322891*.

Acknowledgements

We are grateful to the authors of [58] and the lab of Corey Watson for sharing pre-processed data. We thank Antoine Koehl for helpful discussions, Michael Chungyon and Timothy Whitehead for help with data processing, and Aakarsh Vermani for identifying our use of masked-marginals perplexity for the Koenig dataset. This work supported by NIH grants R01-AI146028 (PI Matsen), R56-HG013117 and R01-HG013117 (PI Song), DP2-AI177890 (PI Starr), and Searle Scholars Program (PI Starr).

Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685. Frederick Matsen is an investigator of the Howard Hughes Medical Institute.

ChatGPT and Claude AI models were used to draft code and text for this manuscript.

Additional information

Reproducing figures

The following figures were made in notebooks, which can be found in the notebooks/dasm_paper directory of the experiments repository.

- Figure 1ab is made in nt_process_in_llms.ipynb.
- Figure 1c and Figure S1 are made in koenig_masked_marginals.ipynb.
- Table 1 and Figures 3, S2, S3, and S4 are made in koenig.ipynb.
- Figure S5 is made by the Snakemake pipeline and in perplexity.ipynb.
- Table 2 and Figure S6 are made in shanehsazzadeh.ipynb.
- Table S1 is made in data_summaries.ipynb.
- Table S2 values are from *https://github.com/matsengrp/dasm-experiments/pull/2*.

Figure S9 is made in multihit_model_exploration.ipynb in the *https://github.com/matsengrp/thrifty-experiments-1* repository.

Our repository file data/whitehead/MAGMA_PIPELINE_STRUCTURE.md describes how Figure S7 and Table 3 are made.

Table 4 is made by timing_direct_gpu.py and make_timing_table.py.

Funding

Funder	Grant reference number	Author
HHS | National Institutes of Health (NIH)	R01-AI146028	Frederick A Matsen David H Rich Mackenzie M Johnson Julia Fukuyama
HHS | National Institutes of Health (NIH)	R56-HG013117	Yun S Song
HHS | National Institutes of Health (NIH)	R01-HG013117	Yun S Song
HHS | National Institutes of Health (NIH)	DP2-AI177890	Tyler N Starr
Searle Scholars Program (SSP)		Tyler N Starr
Howard Hughes Medical Institute (HHMI)		Will Dumm

Howard Hughes Medical Institute (HHMI)	Kevin Sung
Howard Hughes Medical Institute (HHMI)	Hugh Haddox
Howard Hughes Medical Institute (HHMI)	Frederick A Matsen

Author ORCID iDs

Frederick A Matsen: https://orcid.org/0000-0003-0607-6025
Will Dumm: https://orcid.org/0000-0002-8617-476X
Kevin Sung: https://orcid.org/0000-0002-7289-845X
Mackenzie M Johnson: https://orcid.org/0000-0002-3915-2023
David Rich: https://orcid.org/0009-0005-2501-4032
Tyler Starr: https://orcid.org/0000-0001-6713-6904
Yun S Song: https://orcid.org/0000-0002-0734-9868
Julia Fukuyama: https://orcid.org/0000-0002-7590-5563
Hugh K Haddox: https://orcid.org/0000-0001-8324-8324

References

[1] **Ruffolo Jeffrey A,** Gray Jeffrey J, Sulam Jeremias (2021) Deciphering antibody affinity maturation with language models and weakly supervised learning. *arXiv* http://arxiv.org/abs/2112.07782

[2] **Leem Jinwoo,** Mitchell Laura S, Farmery James H R, Barton Justin, Galson Jacob D (2022) Deciphering the language of anti-bodies using self-supervised learning. *Patterns* **3**:e100513 https://doi.org/10.1016/j.patter.2022.100513

[3] **Barton Justin,** Gaspariunas Aretas, Yadin David A, Dias Jorge, Nice Francesca L, Minns Danielle H, Snudden Olivia, Povall Chelsea, Tomas Sara Valle, Dobson Harry, *et al.* (2024) A generative foundation model for antibody sequence understanding. *bioRxiv* https://doi.org/10.1101/2024.05.22.594943

[4] **Devlin Jacob,** Chang Ming-Wei, Lee Kenton, Toutanova Kristina (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv* http://arxiv.org/abs/1810.04805

[5] **Olsen Tobias H,** Moal Iain H, Deane Charlotte M (2024) Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics* btae618 https://doi.org/10.1093/bioinformatics/btae618

[6] **Ng Karenna,** Briney Bryan (2025) Focused learning by antibody language models using preferential masking of non-templated regions. *Patterns* **6**:e101239 https://doi.org/10.1016/j.patter.2025.101239

[7] **Sheng Zizhang,** Schramm Chaim A, Kong Rui, NISC Comparative Sequencing Program, Mullikin James C, Mascola John R, Kwong Peter D, Shapiro Lawrence (2017) Gene-Specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hyper-mutation. *Front. Immunol* **8** https://doi.org/10.3389/fimmu.2017.00537

[8] **Chungyoun Michael,** Ruffolo Jeffrey, Gray Jeffrey (2024) FLAb: Bench-marking deep learning methods for antibody fitness prediction. *bioRxiv* https://doi.org/10.1101/2024.01.13.575504

[9] **Sung Kevin,** Johnson Mackenzie M, Dumm Will, Simon Noah, Haddox Hugh, Fukuyama Julia, Matsen Frederick A (2025) Thrifty wide-context models of B cell receptor somatic hypermutation. *eLife* https://doi.org/10.7554/elife.105471.1

[10] **Yaari Gur,** Vander Heiden Jason A, Uduman Mohamed, Gadala-Maria Daniel, Gupta Namita, Stern Joel N H, O'Connor Kevin C, Hafler David A, Laserson Uri, Vigneault Francois, *et al.* (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol* **4** https://doi.org/10.3389/fimmu.2013.00358

[11] **Rao Roshan,** Meier Joshua, Sercu Tom, Ovchinnikov Sergey, Rives Alexander (2020) Transformer protein language models are unsupervised structure learners.

[12] **Rives Alexander,** Meier Joshua, Sercu Tom, Goyal Siddharth, Lin Zeming, Liu Jason, Guo Demi, Ott Myle, Lawrence Zitnick C, Ma Jerry, *et al.* (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U. S. A*

118:e2016239118 https://doi.org/10.1073/pnas.2016239118

[13] **Nijkamp Erik,** Ruffolo Jeffrey, Weinstein Eli N, Naik Nikhil, Madani Ali (2022) ProGen2: Exploring the boundaries of protein language models. *arXiv* http://arxiv.org/abs/2206.13517

[14] **Petersen Brian M,** Kirby Monica B, Chrispens Karson M, Irvin Olivia M, Strawn Isabell K, Haas Cyrus M, Walker Alexis M, Baumer Zachary T, Ulmer Sophia A, Ayala Edgardo, *et al.* (2024) An integrated technology for quantitative wide mutational scanning of human antibody fab libraries. *Nat. Commun* **15**:3974 https://doi.org/10.1038/s41467-024-48072-z

[15] **Kirby Monica B,** Petersen Brian M, Faris Jonathan G, Kells Siobhan P, Sprenger Kayla G, Whitehead Timothy A (2025) Retrospective SARS-CoV-2 human antibody development trajectories are largely sparse and permissive. *Proc. Natl. Acad. Sci. U. S. A* **122**:e2412787122 https://doi.org/10.1073/pnas.2412787122

[16] **Koenig Patrick,** Lee Chingwei V, Walters Benjamin T, Janakiraman Vasantharajan, Stinson Jeremy, Patapoff Thomas W, Fuh Germaine (2017) Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proc. Natl. Acad. Sci. U. S. A* https://doi.org/10.1073/pnas.1613231114

[17] **Rodriguez Oscar L,** Safonova Yana, Silver Catherine A, Shields Kaitlyn, Gibson William S, Kos Justin T, Tieri David, Ke Hanzhong, Jackson Katherine J L, Boyd Scott D, *et al.* (2023) Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nature Comm* **14**:4419 https://doi.org/10.1038/s41467-023-40070-x

[18] **Pilzecker Bas,** Jacobs Heinz (2019) Mutating for good: DNA damage responses during somatic hypermutation. *Front. Immunol* **10** https://doi.org/10.3389/fimmu.2019.00438

[19] **Matsen Frederick A,** Sung Kevin, Johnson Mackenzie M, Dumm Will, Rich David, Starr Tyler, Song Yun S, Bradley Philip, Fukuyama Julia, Haddox Hugh K (2025) A sitewise model of natural selection on individual antibodies via a transformer-encoder. *Mol. Biol. Evol* msaf186 https://doi.org/10.1093/molbev/msaf186

[20] **Goldman N,** Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol* **11**:725-736 PubMed

[21] **Muse S V,** Gaut B S (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol* **11**:715-724 PubMed

[22] **Halpern A L,** Bruno W J (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol. Biol. Evol* **15**:910-917 PubMed

[23] **Spisak Natanael,** Walczak Aleksandra M, Mora Thierry (2020) Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic Acids Res* **48**:10702-10712 https://doi.org/10.1093/nar/gkaa825

[24] **DeWitt William S,** Vora Ashni A, Araki Tatsuya, Galloway Jared G, Alkutkar Tan-wee, Bortolatto Juliana, Castro Tiago B R, Dumm Will, Jennings-Shaffer Chris, Jia Tongqiu, *et al.* (2025) Replaying germinal center evolution on a quantified affinity land-scape. *bioRxiv* https://doi.org/10.1101/2025.06.02.656870

[25] **Johnson Mackenzie M,** Sung Kevin, Haddox Hugh K, Vora Ashni A, Araki Tatsuya, Victora Gabriel D, Song Yun S, Fukuyama Julia, Matsen Frederick A (2025) Nucleotide context models outperform protein language models for predicting antibody affinity maturation. *bioRxiv* https://doi.org/10.1101/2025.06.16.659977

[26] **Lin Zeming,** Akin Halil, Rao Roshan, Hie Brian, Zhu Zhongkai, Lu Wenting, Smetanin Nikita, Verkuil Robert, Kabeli Ori, Shmueli Yaniv, *et al.* (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**:1123-1130 https://doi.org/10.1126/science.ade2574

[27] **Luiz C Vieira,** Morgan L Handojo, Claus O Wilke (2025) Medium-sized protein language models perform well at transfer learning on realistic datasets. *Sci. Rep.* **15**:21400 https://doi.org/10.1038/s41598-025-05674-x

[28] **Shanehsazzadeh Amir,** Bachas Sharrol, McPartlon Matt, Kasun George, Sutton John M, Steiger Andrea K, Shuai Richard, Kohnert Christa, Rakocevic Goran, Gutierrez Jahir M, *et al.* (2023) Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv* https://doi.org/10.1101/2023.01.08.523187

[29] **Adams Rhys M,** Mora Thierry, Walczak Aleksandra M, Kinney Justin B (2016) Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife* **5** https://doi.org/10.7554/eLife.23156

[30] **Kumar Sanjeev,** Patel Anamika, Lai Lilin, Chakravarthy Chennareddy, Valanparambil Rajesh, Reddy Elluri Seetharami, Gottimukkala Kamalvishnu, Davis-Gardner Meredith E, Edara Venkata Viswanadh, Linderman Susanne, *et al.* (2022) Structural insights for neutralization of omicron variants BA.1, BA.2, BA.4, and BA.5 by a broadly neutralizing SARS-CoV-2 antibody. *Sci. Adv.* **8**:eadd2032 https://doi.org/10.1126/sciadv.add2032

[31] **York Ryan,** Avasthi Prachee (2025) Phylogenies and biological foundation models. *Arcadia Science* https://doi.org/10.57844/arcadia-znum-bm22

[32] **Tamuri Asif U.,** Reis Mario dos, Goldstein Richard A. (2012) Estimating the distribution of selection coefficients from phylogenetic data using sitewise mutation-selection models. *Genetics* **190**:1101-1115 https://doi.org/10.1534/genetics.111.136432

[33] **Tamuri Asif U,** Goldman Nick, Reis Mario dos (2014) A penalized-likelihood method to estimate the distribution of selection coefficients from phylogenetic data. *Genetics* **197**:257-271 https://doi.org/10.1534/genetics.114.162263

[34] **Rodrigue Nicolas,** Lartillot Nicolas (2014) Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics* **30**:1020-1021 https://doi.org/10.1093/bioinformatics/btt729

[35] **Bloom Jesse D** (2017) Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol. Direct* **12** https://doi.org/10.1186/s13062-016-0172-z

[36] **Hilton Sarah K,** Doud Michael B (2017) and Jesse D Bloom. phydms: software for phylogenetic analyses informed by deep mutational scanning. *PeerJ* **5**:e3657 https://doi.org/10.7717/peerj.3657

[37] **Robinson Douglas M,** Jones David T, Kishino Hirohisa, Gold-man Nick, Thorne Jeffrey L (2003) Protein evolution with dependence among codons due to tertiary structure. *Mol. Biol. Evol* **20**:1692-1704 https://doi.org/10.1093/molbev/msg184

[38] **Dunbar James,** Krawczyk Konrad, Leem Jinwoo, Baker Terry, Fuchs Angelika, Georges Guy, Shi Jiye, Deane Charlotte M (2014) SAbDab: the structural antibody database. *Nucleic Acids Res* **42**:D1140-6 https://doi.org/10.1093/nar/gkt1043

[39] **Hannon William W,** Bloom Jesse D (2024) Dms-viz: Structure-informed visualizations for deep mutational scanning and other mutation-based datasets. *J. Open Source Softw.* **9**:6129 https://doi.org/10.21105/joss.06129

[40] **Wang Meng,** Patsenker Jonathan, Li Henry, Kluger Yuval, Kleinstein Steven H (2024) Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. *Nucleic Acids Res* **52**:548-557 https://doi.org/10.1093/nar/gkad1128

[41] **Wang Meng,** Patsenker Jonathan, Li Henry, Kluger Yuval, Kleinstein Steven H (2025) Supervised fine-tuning of pre-trained antibody language models improves antigen specificity prediction. *PLoS Comput. Biol* **21**:e1012153 https://doi.org/10.1371/journal.pcbi.1012153

[42] **Rao Roshan,** Liu Jason, Verkuil Robert, Meier Joshua, Canny John F, Abbeel Pieter, Sercu Tom, Rives Alexander (2021) MSA transformer. *bioRxiv* https://doi.org/10.1101/2021.02.12.430858

[43] **Ralph Duncan K,** Matsen Frederick A (2022) 4th. Inference of B cell clonal families using heavy/light chain pairing information. *PLoS Comput. Biol* **18**:e1010723

[44] **Minh Bui Quang,** Schmidt Heiko A, Chernomor Olga, Schrempf Dominik, Woodhams Michael D, von Haeseler Arndt, Lanfear Robert (2020) Iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol* **37**:1530-1534 https://doi.org/10.1093/molbev/msaa015

[45] **Chernomor Olga,** von Haeseler Arndt, Minh Bui Quang (2016) Terrace aware data structure for phylogenomic inference from supermatrices. *Syst Biol* **65**:997-1008 https://doi.org/10.1093/sysbio/syw037

[46] **Vaswani Ashish,** Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N, Ukasz Kaiser L-, Polosukhin Illia (2017) Attention is all you need. In: Guyon I, Luxburg U V, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (Eds). *Advances in Neural Information Processing Systems* **30** Curran Associates, Inc.. pp. 6000-6010

[47] **Lucaci Alexander G,** Wisotsky Sadie R, Shank Stephen D, Weaver Steven (2021) Extra base hits: Widespread empirical support for instantaneous multiple-nucleotide changes. *PLoS One* **16**:e0248337 https://doi.org/10.1371/journal.pone.0248337

[48] **Jensen Cole G,** Sumner Jacob A, Kleinstein Steven H, Hoehn Kenneth B (2024) Inferring B cell phylogenies from paired H and L chain BCR sequences with dowser. *J. Immunol* **212**:1579-1588 https://doi.org/10.4049/jimmunol.2300851

[49] **Paszke Adam,** Gross Sam, Massa Francisco, Lerer Adam, Brad-bury James, Chanan Gregory, Killeen Trevor, Lin Zeming, Gimelshein Natalia, Antiga Luca, *et al.* (2019) PyTorch: An imperative style, high-performance deep learning library. *arXiv* https://doi.org/10.48550/arXiv.1912.01703

[50] **Satyanarayan Arvind,** Moritz Dominik, Wongsuphasawat Kanit, Heer Jeffrey (2017) Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics* **23**:341-350

[51] **VanderPlas Jacob,** Granger Brian, Heer Jeffrey, Moritz Dominik, Wongsuphasawat Kanit, Satyanarayan Arvind, Lees Eitan, Timofeev Ilia, Welsh Ben, Sievert Scott (2018) Altair: Interactive statistical visualizations for python. *Journal of Open Source Software* **3**:e01057 https://doi.org/10.21105/joss.01057

[52] **Cock Peter J. A.,** Antao Tiago, Chang Jeffrey T., Chapman Brad A., Cox Cymon J., Dalke Andrew, Friedberg Iddo, Hamelryck Thomas, Kauff Frank, Wilczynski Bartek, *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**:1422-1423 https://doi.org/10.1093/bioinformatics/btp163

[53] **Hunter J. D.** (2007) Matplotlib: A 2d graphics environment. *Computing in Science & Engineering* **9**:90-95 https://doi.org/10.1109/MCSE.2007.55

[54] **McKinney Wes** (2010) Data Structures for Statistical Computing in Python. In: Proceedings of the 9th Python in Science Conference. pp. 56-61 https://doi.org/10.25080/Majora-92bf1922-00a

[55] **Krekel Holger,** Oliveira Bruno, Pfannschmidt Ronny, et al (2004) pytest. GitHub. version: v.8.1.1 https://github.com/pytest-dev/pytest

[56] **Waskom ML** (2021) seaborn: statistical data visualization. *Journal of Open Source Software* **6**:e03021 https://doi.org/10.21105/joss.03021

[57] **Mölder Felix,** Jablonski Kim Philipp, Letcher Brice, Hall Michael B, Tomkins-Tinch Christopher H, Sochat Vanessa, Forster Jan, Lee Soohyun, Twardziok Sven O, Kanitz Alexander, *et al.* (2021) Sustainable data analysis with snakemake. *F1000Res* **10** https://doi.org/10.12688/f1000research.29032.2

[58] **Tang Catherine,** Krantsevich Artem, MacCarthy Thomas (2022) Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience* **25**:e103668 https://doi.org/10.1016/j.isci.2021.103668

[59] **Jaffe David B,** Shahi Payam, Adams Bruce A, Chrisman Ashley M, Finnegan Peter M, Raman Nandhini, Royall Ariel E, Tsai Funien, Vollbrecht Thomas, Reyes Daniel S, *et al.* (2022) Functional antibodies exhibit light chain coherence. *Nature* **611**:352-357 https://doi.org/10.1038/s41586-022-05371-z

[60] **Vergani Stefano,** Korsunsky Ilya, Mazzarello Andrea Nicola, Fer-rer Gerardo, Chiorazzi Nicholas, Bagnara Davide (2017) Novel method for high-throughput full-length IGHV-D-J sequencing of the immune repertoire from bulk b-cells with single-cell resolution. *Front. Immunol* **8**:1157

https://doi.org/10.3389/fimmu.2017.01157

[61] **Engelbrecht Eric,** Rodriguez Oscar L, Lees William, Vanwinkle Zach, Shields Kaitlyn, Schultze Steven, Gibson William S, Smith David R, Jana Uddalok, Saha Swati, *et al.* (2025) Germline polymorphism in the immunoglobulin kappa and lambda loci explain variation in the expressed light chain antibody repertoire. *bioRxiv* https://doi.org/10.1101/2025.05.28.656470

Peer reviews

Reviewer #1 (Public review):

Summary:

Matsen et al. describe an approach for training an antibody language model that explicitly tries to remove effects of "neutral mutation" from the language model training task, e.g. learning the codon table, which they claim results in biased functional predictions. They do so by modeling empirical sequence-derived likelihoods through a combination of a "mutation" model and a "selection" model; the mutation model is a non-neural Thrifty model previously developed by the authors, and the selection model is a small Transformer that is trained via gradient descent. The sequence likelihoods themselves are obtained from analyzing parent-child relationships in natural SHM datasets. The authors validate their method on several standard benchmark datasets and demonstrate its favorable computational cost. They discuss how deep learning models explicitly designed to capture selection and not mutation, trained on parent-child pairs, could potentially apply to other domains such as viral evolution or protein evolution at large.

Overall, we think the idea behind this manuscript is really clever and shows promising empirical results. Two aspects of the study are conceptually interesting: the first is factorizing the training likelihood objective to learn properties that are not explained by simple neutral mutation rules, and the second is training not on self-supervised sequence statistics but on the differences between sequences along an antibody evolutionary trajectory. If this approach generalizes to other domains of life, it could offer a new paradigm for training sequence-to-fitness models that is less biased by phylogeny or other aspects of the underlying mutation process.

Future versions of the work can consider extending the ideas to additional datasets, species, definitions of fitness, or even different proteins entirely.

Comments on revisions:

We thank the authors for addressing our points and have no remaining questions.

https://doi.org/10.7554/eLife.109644.2.sa3

Reviewer #2 (Public review):

Summary:

Endowing protein language models with an ability to predict the function of antibodies would open a world of translational possibilities. However, antibody language models have yet to achieve the breakthrough success, which large language models have achieved for the understanding and generation of natural language. This paper elegantly demonstrates how training objectives imported from natural language applications lead antibody language models astray on function prediction tasks. Training models to predict masked amino acids teaches models to exploit biases of nucleotide-level mutational processes, rather than protein biophysics. Taking the underlying biology of antibody diversification and selection seriously allows disentangling these processes, through what the authors call deep amino acid

selection models. These models extend previous work by the authors (Matsen MBE 2025) by providing predictions not only for the selection strength at individual sites, but also for individual amino acids substitutions. This represents a practically important advance.

Strengths:

The paper is based on a deep conceptual insight, the existence of multitude of biological processes that affect antibody maturation trajectories. The figures and writing a very clear, which should help make the broader field aware of this important but sometimes overlooked insight. The paper adds to a growing literature proposing biology-informed tweaks for training protein language models, and should thus be of interest to a wide readership interested in the application of machine learning to protein sequence understanding and design.

Weaknesses:

Proponents of the state-of-the-art protein language models might counter the claims of the paper by appealing to the ability of fine-tuning to deconvolve selection and mutation-related signatures in their high-dimensional representation spaces. Leaving the exercise of assessing this claim entirely to future work somewhat diminishes the heft of the (otherwise good!) argument. In the context of predicting antibody binding affinity, the modeling strategy only allows prediction of mutations that improve affinity on average but not those which improve binding to specific epitopes.

Comments on revisions:

We thank the authors for clarifying the description of the methods and for adding additional discussion of important directions for future work.

https://doi.org/10.7554/eLife.109644.2.sa2

Reviewer #3 (Public review):

Summary:

This work proposes DASM, a new transformer-based approach to learning the distribution of antibody sequences which outperforms current foundational models at the task of predicting mutation propensities under selected phenotypes, such as protein expression levels and target binding affinity. The key ingredient is the disentanglement, by construction, of selection-induced mutational effects and biases intrinsic to the somatic hypermutation process (which are embedded in a pre-trained model).

Strengths:

The approach is benchmarked on a variety of available datasets and for two different phenotypes (expression and binding affinity). The biologically informed logic for model construction implemented is compelling and the advantage, in terms of mutational effects prediction as well as computational efficiency, is clearly demonstrated via comparisons to state-of-the-art models.

Weaknesses:

While all the main points are well addressed and supported, it could have been interesting to strengthen the claim of gain in interpretability by investigating it explicitly in relation to the functional effects studied in this paper.

Comments on revisions:

I thank the authors for clarifying a few points I had flagged up and I appreciate much better that the content of the companion paper was precisely covering model selection and structural interpretability.

Regarding my first point (references for language models for antibodies), I feel that the parenthetical citation format shouldn't be a problem (but the editors might advise here). Antiberta2 is this paper: *https://www.biorxiv.org/content/10.1101/2023.12.12.569610v1.full.pdf* (yet, I understand if the authors want to focus on models purely sequence-based). A couple of additional references could be: *https://academic.oup.com/bioinformatics/article/40/11/btae659/7888884*; *https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1012646*; *https://www.pnas.org/doi/10.1073/pnas.2418918121*; *https://arxiv.org/abs/2506.13006*.

A very minor comment: could one add some p-value (it could be a supplementary table) for the Pearson correlation coefficients? The comparison between methods is rather clear, but for some correlations it's a bit unclear whether they should be considered significant. It would be important to understand the extent to which in different datasets one might expect functional prediction power based on an evolutionary objective function alone.

https://doi.org/10.7554/eLife.109644.2.sa1

Author response:

The following is the authors' response to the original reviews

> ***Public Reviews:***
>
> ***Reviewer #1 (Public review):***
>
> *Summary:*
>
> *Matsen et al. describe an approach for training an antibody language model that explicitly tries to remove effects of "neutral mutation" from the language model training task, e.g. learning the codon table, which they claim results in biased functional predictions. They do so by modeling empirical sequence-derived likelihoods through a combination of a "mutation" model and a "selection" model; the mutation model is a non-neural Thrifty model previously developed by the authors, and the selection model is a small Transformer that is trained via gradient descent. The sequence likelihoods themselves are obtained from analyzing parent-child relationships in natural SHM datasets. The authors validate their method on several standard benchmark datasets and demonstrate its favorable computational cost.*
>
> *They discuss how deep learning models explicitly designed to capture selection and not mutation, trained on parent-child pairs, could potentially apply to other domains such as viral evolution or protein evolution at large.*
>
> *Strengths:*
>
> *Overall, we think the idea behind this manuscript is really clever and shows promising empirical results. Two aspects of the study are conceptually interesting: the first is factorizing the training likelihood objective to learn properties that are not explained by simple neutral mutation rules, and the second is training not on self-supervised sequence statistics but on the differences between sequences along an antibody evolutionary trajectory. If this approach generalizes to other domains of life, it could offer a new paradigm for training sequence-to-fitness models that is less biased by phylogeny or other aspects of the underlying mutation process.*

Thank you for your kind words.

> *Weaknesses:*
>
> *Some claims made in the paper are weakly or indirectly supported by the data. In particular, the claim that learning the codon table contributes to biased functional effect predictions may be true, but requires more justification.*

Thank you for this comment, which made us realize that we had not adequately explained the key insight of Figure S3. We have expanded the caption of Figure S3 to clarify:

"DASM selection factors match the pattern seen in experimental measurements, while masked language models show artifacts from the codon table.

The experimental data (left two panels) show a slight decrease in median scores for amino acids requiring multiple nucleotide mutations ("multiple") versus single mutations ("single").

DASM captures this pattern, showing similar distributions for both categories.

In contrast, AbLang and ESM assign radically lower scores to multinucleotide amino acid substitutions, consistent with the masked language modeling objective learning codon-level mutation probabilities as described in the main text (Figure 1a)."

This figure directly supports our claim: the experimental fitness data show similar distributions for single-mutation vs multiple-mutation amino acids, yet AbLang2 and ESM assign dramatically different scores to these groups, while DASM does not.

> *Additionally, the paper could benefit from additional benchmarking and comparison to enhanced versions of existing methods, such as AbLang plus a multi-hit correction.*

It's an interesting idea to consider enhancing existing models. However, this approach faces some challenges. Most fundamentally, it is difficult to recast AbLang and other such models in an evolutionary framework: the masked language objective is simply not an evolutionary one. We have written a whole paper working to do this (*https://doi.org/10.1371/journal.pcbi.1013758*) and the results were middling despite our best efforts. Specifically regarding multihit, the effects of multihit are minor compared to the codon table effects, and those require the structure of codon-based evolutionary model.

> *Further descriptions of model components and validation metrics could help make the manuscript more readable.*

We have clarified several aspects of the model in the revision: we now describe the Thrifty neutral model in the introduction, clarify the transformer architecture and wiggle activation function in the Methods, and explain the joint branch-length optimization procedure.

In the introduction we now describe Thrifty:

"This fixed model uses convolutions on 3-mer embeddings to deliver wide context sensitivity without needing a large number of parameters: the variant we use has around the same number of parameters as the classic S5F 5-mer model."

In the Methods we clarify the architecture:

"We parameterize the DASM *f* using the standard transformer-encoder architecture: an amino-acid embedding, sinusoidal positional encodings, and PyTorch's TransformerEncoder module.

The only non-standard component to this architecture is a custom "wiggle" activation function to the output layer that prevents extreme selection factors as previously described.

This function asymptotes to zero for highly deleterious mutations and grows sub-linearly for beneficial ones."

And the joint optimization:

"This joint optimization is performed cyclically, in which a complete cycle consists of neural network optimization followed by branch length optimization for every parent-child pair.

The parent sequence and the child sequence are pre-estimated, fixed, and used as training data.

The branch lengths are independent and so are optimized in parallel."

> ***Reviewer #2 (Public review):***
>
> *Summary:*
>
> *Endowing protein language models with the ability to predict the function of antibodies would open a world of translational possibilities. However, antibody language models have yet to achieve breakthrough success, which large language models have achieved for the understanding and generation of natural language. This paper elegantly demonstrates how training objectives imported from natural language applications lead antibody language models astray on function prediction tasks. Training models to predict masked amino acids teaches models to exploit biases of nucleotide-level mutational processes, rather than protein biophysics. Taking the underlying biology of antibody diversification and selection seriously allows for disentangling these processes through what the authors call deep amino acid selection models. These models extend previous work by the authors (Matsen MBE 2025) by providing predictions not only for the selection strength at individual sites, but also for individual amino acid substitutions. This represents a practically important advance.*
>
> *Strengths:*
>
> *The paper is based on a deep conceptual insight, the existence of a multitude of biological processes that affect antibody maturation trajectories. The figures and writing a very clear, which should help make the broader field aware of this important but sometimes overlooked insight. The paper adds to a growing literature proposing biology-informed tweaks for training protein language models, and should thus be of interest to a wide readership interested in the application of machine learning to protein sequence understanding and design.*

Thank you for your kind words.

> *Weaknesses:*
>
> *Proponents of the state-of-the-art protein language models might counter the claims of the paper by appealing to the ability of fine-tuning to deconvolve selection and mutation-related signatures in their high-dimensional representation spaces. Leaving the exercise of assessing this claim entirely to future work somewhat diminishes the heft of the (otherwise good!) argument.*

This is an interesting idea! However, it seems to us that this approach has some fundamental limitations. Existing models operate on amino acid sequences with no nucleotide representation, so while they can be implicitly biased by the codon table, they have no signal to separate selection from effects related to the codon table and SHM rates.

We interpret this comment as proposing that we could use fine-tuning on functional data to pull out the selection components (that would only affect the functional data) versus the mutation component. That sounds like an interesting research project. We would be concerned that there are correlations between mutability and selective effects (e.g., CDRs are both more mutable and under different selection), creating identifiability problems unless separate data sources are used as we do here.

Additionally, the fine-tuning approaches we are aware of are taskspecific: they require labeled data from a specific assay (binding to antigen X, expression in system Y) that may or may not relate to the general evolutionary selection signal. Also, such approaches are limited to the specific data used and may not do a good job of guiding the model to a signal that is not present in the training data.

By structuring the model as we do, we obtain the evolutionary interpretation directly from phylogenetic signal without requiring taskspecific supervision.

In the context of predicting antibody binding affinity, the modeling strategy only allows prediction of mutations that improve affinity on average, but not those which improve binding to specific epitopes.

We agree, and this is fundamental to any general purpose model. Predictions of binding patterns for a specific target requires information about that target to be specified in the training data. We look forward to developing such task-specific models in the future.

We have added a paragraph to the Discussion clarifying this limitation:

“The current generation of DASM model does not use any antigen-labeled training data.

The signal that it leverages to infer some limited ability to predict binding comes from natural affinity maturation.

This affinity maturation comes through natural repertoires and so represents a mix of all of the antigens to which the sampled individuals have been exposed.”

> ***Reviewer #3 (Public review):***
>
> *Summary:*
>
> *This work proposes DASM, a new transformer-based approach to learning the distribution of antibody sequences which outperforms current foundational models at the task of predicting mutation propensities under selected phenotypes, such as protein expression levels and target binding affinity. The key ingredient is the disentanglement, by construction, of selection-induced mutational effects and biases intrinsic to the somatic hypermutation process (which are embedded in > a pre-trained model).*
>
> *Strengths:*
>
> *The approach is benchmarked on a variety of available datasets and for two different phenotypes (expression and binding affinity). The biologically informed logic for model construction implemented is compelling, and the advantage, in terms of mutational effects prediction, is clearly demonstrated via comparisons to state-of-the-art models.*

Thank you.

> *Weaknesses:*
>
> *The gain in interpretability is only mentioned but not really elaborated upon or leveraged for gaining insight.*

We are also excited about the ability of these models to provide interpretable predictions. We have dedicated an entire paper to this direction: "A Sitewise Model of Natural Selection on Individual Antibodies via a Transformer-Encoder" in MBE (*https://doi.org/10.1093/molbev/msaf186*). The interpretations offered by that paper overturn some of the oversimplified dogma about how natural selection works in antibodies (purifying in FWK and diversifying in CDR), giving a more nuanced sitewise perspective. The paper also highlights the importance of specific structural features of the antibodies.

This eLife paper, on the other hand, is focused on comparison to antibody language models and benchmarking zero-shot prediction on functional tasks.

We have better highlighted this new paper in our revision with:

"We have dedicated a companion paper to leveraging this interpretability to provide new perspectives on the operating rules of affinity maturation (Matsen et al., MBE 2025): that work provides a nuanced sitewise perspective on natural selection in antibodies that challenges classical oversimplified views of selection patterns."

> *The following aspects could have been better documented: the hyperparametric search to establish the optimal model; the predictive performance of baseline approaches, to fully showcase the gain yielded by DASM.*

We appreciate the concern and the desire to reveal all the factors that lead to a strong performance result. For this particular paper, we feel that this is less of a concern because we are optimizing according to an evolutionary objective function and then evaluating according to a functional one. We now describe how other than model size, hyperparameters stayed the same as in our previous paper (Matsen et al., MBE 2025).

Regarding baseline approaches, our previous paper includes comparisons to simpler models for the evolutionary objective. Here we focus on comparison to antibody language models for functional prediction. Comparing between state-of-the-art models is the standard practice for papers in this field.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations for the authors):***
>
> *We recommend modest amounts of revision, discussed below:*
>
> *Major comments:*
>
> *(1) In the first section of the results, there is extensive discussion on shortcomings of existing antibody language models like AbLang2 that seems to associate all of the performance gap with the inability to separate non-synonymous mutations separated by 1 or 2+ substitutions.*
>
> *In reality, some of the lower likelihoods in the 2+ substitution case could actually reflect real fitness deficits (while others could indeed be rarer occurrences in the training data). The authors should either moderate these claims or do an analysis that leverages antibody deep mutational scanning data to show that, conditioned on the fitness of the antibody (probably expression) being the same (either all high or all low), AbLang2 still artefactually considers rarer-training/less-codon-accessible variants to be less fit.*

As described above, we believe that this is addressed by Figure S3, but if not please correct us.

> *(2) Some in the machine learning for antibody community might view the set of benchmarked datasets to be incomplete and somewhat arbitrarily selected, though we do think this is a good start, and the results are promising. A dataset commonly used in this field that is missing from this paper is from Shehata et al. (https://pubmed.ncbi.nlm.nih.gov/31553901/). A binding affinity experiment that is also commonly used in the field is from Phillips et al. (https://elifesciences.org/articles/71393) - this dataset measures combinatorial changes of framework regions on binding, which may be especially relevant here.*

We're glad to have the opportunity to clarify this, thanks.

We based our evaluations on the April 2024 version of the FLAb benchmarking project (*https://doi.org/10.1101/2024.01.13.575504*) which preceded our work and thus was not subject to selection bias by us. We took the largest data sets in that repository. After this we became aware of the rich data sets offered by the Whitehead lab that provided binding measurements for many variants for a number of antigens, and added that to the evaluation set.

We have clarified this in the manuscript:

“We based our evaluations on the April 2024 version of the FLAb benchmarking project, which preceded our work and thus was not subject to selection bias by us.

We also benchmarked high-throughput binding data (more recent than FLAb) from the Whitehead lab that provided affinity measurements across many variants and antigens.”

The Shehata dataset is interesting but doesn't fit so much in the DASM mold: it is a survey of biophysical properties across many independent antibodies rather than a deep investigation of point mutants of a smaller collection of focal antibodies.

FLAb has grown to include the Phillips dataset. We are working full-tilt on the next version of DASM and will be including many other datasets in our paper on DASM2. Thanks for the tip!

> *(3) Similar to the above comment, we were also extremely curious as to why the authors did not test data from DeWitt et al. (https://pubmed.ncbi.nlm.nih.gov/40661619/). Instead, the authors only make a cryptic reference to this study on lines 201-6, but we could not even find a figure describing the results discussed on these lines. It would be great to actually include this data.*

We agree, however, our model is for human rather than mouse. We would like to train a mouse model in the future but have not yet lined up the appropriate data.

> *(4) The authors should comment on potential data leakage if the SHM trajectories used in training have a similar sequence or antigen similarity to the benchmark expression/binding datasets.*

This is a good question that we should clarify. Our model is trained only on evolutionary trajectories and not functional data. Evaluation is then done on functional data without fine-tuning. Because these evaluation data are categorically different from the training data and thus data leakage is not a problem. Recall that our model is zero-shot: it only considers evolutionary trajectories and not functional data as such. In a similar way, other self-supervised models such as MLMs do not exclude seeing an antibody in the training data when they are doing functional prediction.

We have clarified this in the manuscript with

“Because the DASM is trained exclusively on evolutionary trajectories rather than functional measurements, evaluation on expression and binding benchmarks is strictly zero-shot with no risk of data leakage.”

> *Relatedly, what happens if this approach is applied to completely de novo antibodies?*

We direct this reviewer to the Shanehsazzadeh dataset that involves antibodies that were suggested by an AI algorithm rather than observed in nature.

If the reviewer is referring to completely synthetic antibody molecules, such as those generated by inverse folding, we have not attempted this.

> *(5) It makes sense that you included the multihit correction as a response to your earlier instantiation (without this correction) underestimating the probabilities of multiple mutations in a codon associated with a single amino acid substitution (lines 476-477).*
>
> *However, this could potentially make for a somewhat unfair comparison to existing methods: if, say, we took AbLang (or another comparator) and also applied a multi-hit correction (even in some naive way at inference time), how would that compare to DASM? If this comparison favors DASM, it would show that models need more than just such a correction on top of existing methods to do good sequence scoring--which would only amplify the impact of the results.*

Thank you for this suggestion. We believe that we have addressed it in the response to the public reviews, but please let us know if not.

> *Minor comments:*
>
> *(1) It would be worth explicitly defining/summarizing the mutation model used in the study, e.g. giving an overview of Thrifty in the introduction or where it first appears.*

Thanks, we have done this:

“Our approach separates mutation and selection processes by encoding functional effects in a Deep Amino acid Selection Model (DASM) while explicitly modeling mutation using a separate fixed model trained on neutrally evolving data.

This fixed model uses convolutions on 3-mer embeddings to deliver wide context sensitivity without needing a large number of parameters: the variant we use has around the same number of parameters as the classic S5F (Yaari et al., 2013) 5-mer model.”

> *(2) Paragraph starting on line 58: it sounds like you're suggesting that masked deep learning models will learn certain features of genomes in a certain order. We suggest that you weaken the language, giving examples of various things the model could learn, not implying that such models will necessarily learn the most useful features after the less useful ones.*

We have fixed this by removing the "First... Second... Third... Finally" ordering:

“It could memorize the germline genes and learn about the probabilities of V(D)J recombination.

It could learn the codon table, as according to this table some aminoacid mutations are much more likely than others. It could learn rates of somatic hypermutation...

It could also learn about the impact of amino acid mutations on antibody function through natural selection in the course of affinity maturation, which is the desired signal.

However, this desired signal is confounded by the preceding factors."

> *(3) Line 72: You make a strong claim that existing models conflate mutation and selection without knowing for sure that they didn't successfully learn these components separately (it seems this would require a lot of mechanistic interpretability). The language could be softened here.*

We believe that we have addressed this in the response to public reviews, but please let us know if not.

> *(4) Line 79: Say a bit more about the separate fixed mutation model here. Why shouldn't we worry about this choice (especially the word "fixed") biasing your results? Does the empirical performance of your method suggest this doesn't really matter?*

We have added to the description of the fixed mutation model, as described above.

As described in the public response, training SHM models on out-of-frame sequences is an established methodology for characterizing mutation in the absence of selection. In principle one could jointly train a model of SHM and selection, but one could have identifiability problems as there is a correlation between more mutable sites (e.g. in the CDRs) and those under relaxed selection. Using out-of-frame sequences gives a clean an independent description of the SHM process.

> *(5) Line 81: on what benchmarks does it outperform? State briefly.*

Great suggestion. Done:

"The DASM, trained on substantially less data, outperforms AbLang2 and general protein language models including ESM2 and ProGen2-small. This outperformance holds on the largest benchmark datasets of the FLAb collection and on recent high-throughput binding assays."

> *(6) Paragraph starting on line 90: The topic sentence reads a bit vague to us. Do you mean that you want to learn the extent to which models are regurgitating nucleotide similarity of AAs in determining the scores associated with AAs at masked sites?*

Thank you. We have updated to

"We first sought to understand the extent to which processes such as neutral mutation rate and the codon table influence antibody language model prediction at masked sites."

> *(7) Paragraph starting on line 108: feels speculative and maybe better for the discussion...*

We appreciate this comment, but we have decided to keep the content where it is. Although this would make sense as a Discussion item we feel like it fits well here right next to the evidence, and the structure of our Discussion doesn't really have a place for it.

> *(8) Paragraph starting on line 116: don't say "sequences from [12]" or "method of [15]." Explain what these are before giving the citation.*

Whoops! Thanks. We have fixed these.

> *(9) Line 134: Consider giving a brief definition of perplexity?*

Thanks. We added our favorite definition:

"Perplexity (as defined in the Methods) is the standard way of evaluating the plausibility of a sequence according to a model: it is the acrosssite geometric mean of the inverse probability

of the observed amino acid."

> *(10) Line 154: A citation here could be useful to support the claim that these models are learning phylogeny.*

We have replaced with the more clearly established "codon table":

"We implemented a model to learn amino-acid preferences of antibodies without being influenced by germline genes, the codon table, or SHM biases."

> *(11) Lines 161-162: Given that phylogenetic inference methods can be tough to scale, we're curious how you managed to get 2 million PCPs from the data? Did you construct a bunch of different phylogenies (in > parallel)?*

Indeed! We now clarify in the methods section that these trees were run in parallel across clonal families:

"As in our previous work, tree inference and ancestral sequence reconstruction were performed per clonal family with the K80 substitution model...

Because these clonal families are independent these phylogenetic inferences were run in parallel."

> *(12) Line 173-174: Can you say more about the joint optimization of the branch lengths? Are you conditioning on a phylogenetic tree topology only, and leaving the branch lengths unknown? Do you account for the fact that these branch lengths in the same phylogenetic tree aren't independent?*

Thanks for pointing out the need to clarify these points. We have done so in the methods section and provided a pointer to the methods section in the main text.

In the main text we now say:

"We trained DASMs of several sizes (~1M, ~4M, ~7M) using joint optimization of branch length t and parameters of the DASM (see Methods for details)."

And in the Methods:

"This joint optimization is performed cyclically, in which a complete cycle consists of neural network optimization followed by branch length optimization for every parent-child pair.

The parent sequence and the child sequence are pre-estimated, fixed, and used as training data.

The branch lengths are independent and so are optimized in parallel."

> *(13) Line 358: Yes, in a trivial sense, separating mutation and selection means that we know exactly how each of those two components has been learned. We would be curious if you could say anything about mechanistic interpretability within the deep learning selection model. If not, could this be a future research direction?*

We believe that we have addressed this in the response to public reviews, but please let us know if not.

> *(14) Lines 384-386--indeed. Do you have any proposals for how a phylogeny could be constructed at this scale?*

As above this is not one big phylogeny but many, which invites parallelization.

> ***Reviewer #2 (Recommendations for the authors):***

> *(1) I agree that a full study of fine-tuning strategies for all possible alternative models is beyond the scope of the paper. However, a little bit of fine-tuning would go a long way to demonstrate how easy (or hard) it is to extract the relevant signal from a general protein language model embedding.*

As described in our response to the public reviews, we appreciate this point but have decided to focus on the core novelty of the paper and leave fine-tuning experiments to future work.

> *(2) The authors might want to add some discussion about what signals their models capture with regard to binding affinity (averages), and how this limitation might be addressed in future work.*

As described in our response to the public reviews, we have added a paragraph to the Discussion clarifying this limitation.

> ***Reviewer #3 (Recommendations for the authors):***
>
> *(1) Introduction: I think more references have to be provided re: Antibody "foundation" language models, e.g. adding AntiBERTy and the two versions of AntiBERTa.*

We have added citations to those two models, although we weren't sure what the second version of AntiBERTa was. There are very many antibody language models. If we could use number ranges we would cite a dozen or more, but I hesitate to add many of them in the eLife format, which has parenthetical citations. If there are others that you consider essential don't hesitate to suggest them.

> *(2) A key point of the approach is the disentanglement of "mutation" and "selection", as mentioned in the introduction. However, the explanation of what the authors mean by mutation and selection comes only later. I would anticipate it in the introduction for clarity.*

This is a great point. The revised intro has this in the second sentence:

"Natural antibodies are generated through V(D)J recombination, and refined by somatic hypermutation and affinity-based selection in germinal centers."

and the "While the masked..." paragraph now more clearly calls out selection.

> *(3) Line 133: expression of what? Could the authors also explain mechanistically why expression should be impacted by a mutation? In what conditions do these data sample expression?*

We have clarified that it is expression in a phage display library:

"To do so, we used the largest dataset of the FLAb collection of benchmarks, which measures the effect of single mutations on expression in a phage display library."

> *(4) Line 142: Clarify that 0.49 and 0.3 are correlation coefficients. Also, what type of correlation coefficient is this?*

Thanks for the catch! They are Pearson correlations as we now describe.

> *(5) Line 173: The hyperparametric search should have been more documented (with a description of how it was carried out and plots).*

As described in our response to the public reviews, we are optimizing according to an evolutionary objective function and then evaluating according to a functional one. Other than

model size, hyperparameters stayed the same as in our previous paper (Matsen et al., MBE 2025).

> *(6) Line 358: The authors say that 'DASMs provide direct interpretability'. However, this is not really inspected. A valuable addition would be to show how such interpretability is made possible, how it can recapitulate existing biological knowledge or provide hints for antibody engineering.*

As described above, this is addressed in detail in our previous paper.

> *(7) Line 398: 'Inferred insertions or deletions were reversed, so that all sequences align to the naive sequence without gaps.' Could the authors comment on whether this is a limitation of the approach, why it wasn't dealt with and whether it could be the direction of future work?*

Funny you should mention this! We have been planning out such an extension in detail recently. We have added a sentence in the discussion:

“We also have plans to extend the DASM framework to estimate the effect of natural selection on insertion and deletion events.”

> *(8) Line 430-431: Could the authors clarify 'shared' over what? Also, I believe these two lines really describe the DASM architecture. This should be spelt out more clearly and tied to the description provided in lines 173-175. A diagram of the architecture would be a valuable addition to provide a full picture of the model (this could be added to the general diagram of the modelling approach of Figure S8).*

We have clarified in the text that this is indeed a description of the DASM architecture -- thanks for the catch:

“We parameterize the DASM f using the standard transformer-encoder architecture: an amino-acid embedding, sinusoidal positional encodings, and PyTorch's TransformerEncoder module.

The only non-standard component to this architecture is a custom “wiggle” activation function to the output layer that prevents extreme selection factors as previously described.”

The architecture is very “stock” - just the default torch TransformerEncoder, so I don't think that it merits a diagram. We have expanded our discussion of the simple architecture in the revision. This sits in contrast to the setup for the loss function, which is quite custom and is the subject of Figure 2 and Figure S8.

> *(9) Another general remark is that, to fully showcase the predictive advantage offered by DAMS with all the modelling choices entailed, one could show the performance of simpler models, like the mutation model alone (with no selection factors), or models where selection factors are just learnt independently for each site, or are learnt with a simple linear layer instead of a transformer (these are just ideas of some simpler approach that can set baselines over which DASM improvement can be shown).*

This is a great suggestion. The primary focus of this paper is in comparing to alternate antibody language models in terms of functional prediction.

These simpler models could be used for comparing the evolutionary objective, which we did in our previous paper (*https://doi.org/10.1093/molbev/msaf186*). We note that a sitewise model with fixed sites cannot really be appropriately formulated due to sequences being of different lengths.

Additional changes

In addition to the reviewer-requested changes, we added a comparison of ESM2 model sizes (650M vs 3B parameters) on the Koenig benchmark. We found that scaling ESM2 from 650M to 3B parameters did not improve performance. Indeed, the larger model showed slightly degraded correlations, particularly for light chain predictions. This is consistent with recent observations that medium-sized protein language models can outperform larger ones on transfer learning tasks (Vieira et al., Sci. Rep. 2025). We added Table S2 documenting these results and cite this finding in the main text to justify our use of the 650M model throughout the analyses. After doing this, we realized for the Shanehsazzadeh evaluation we had accidentally used ESM2-3B instead of ESM2-650M. The corrected ESM2-650M values are slightly lower (0.191 and 0.308 for sequence lengths 119 and 120, respectively, compared to the previous values of 0.248 and 0.337). This correction does not affect our conclusions, as DASM substantially outperforms ESM2 on this benchmark before and after the change.

We also realized in the course of revision that we had been scoring AbLang2 using the masked-marginals pseudo-perplexity approach for the single-mutant Koenig dataset (Figure 1c), rather than the standard persequence pseudo-perplexity used elsewhere in the paper. For maskedmarginals, probabilities are computed using only wild-type context, whereas standard pseudo-perplexity uses each variant's own context.

The masked-marginals approach has a simple interpretation: for singlemutation variants, it is a linear transformation of the log ratio of the variant amino acid probability to the wild-type amino acid probability, both evaluated under wild-type context. This log-odds ratio directly measures how much the model prefers the mutation over the original residue.

We found that masked-marginals performed better for AbLang2 on this dataset, so we continued using it for Figure 1c. However, for the benchmarking table (Table 1), we switched to per-sequence pseudoperplexity as for the other comparisons in the paper, following the standard benchmarking protocol defined in FLAb (Chungyoun et al., 2024). We document both approaches in the Methods section:

“An alternative “masked-marginals” approach scores variants using only wild-type context.

For a wild-type sequence *w*, masked-marginals computes $p(a \mid w_{\setminus i})$. for all amino acids *a* at each position *i* once, then uses these wild-type-derived probabilities to compute pseudoperplexity for any variant *x*...

For a single-mutation variant *x* that differs from wild-type *w* only at position *j*, all terms except position *j* cancel when comparing to wild-type, giving $\log\left(p(x_j \mid w_{\setminus j})\right) - \log\left(p(w_j \mid w_{\setminus j})\right) = n[\log(\textmm - pseudo - ppl(w)) - \log(\textmm - pseudo - ppl(x))]$. Thus, the log-probability difference between variant and wild-type amino acids equals, up to an additive constant $\log\left(p(x_j \mid w_{\setminus j})\right) - \log\left(p(w_j \mid w_{\setminus j})\right) = n[\log(\textmm - pseudo - ppl(w)) - \log(\textmm - pseudo - ppl(x))]$ that depends only on the wild-type sequence, negative *n* times the log pseudo-perplexity of the variant.

For Figure 1c on the single-mutant Koenig dataset, we found that this approach gave a higher correlation for AbLang2 and so used it in that figure.

For benchmarking comparisons (Table 1), we followed standard practice and used per-sequence pseudo-perplexity.”

https://doi.org/10.7554/eLife.109644.2.sa0