The Annals of Applied Statistics
2025, Vol. 19, No. 1, 720–743
https://doi.org/10.1214/24-AOAS1985

INFERRING MECHANISTIC PARAMETERS OF SOMATIC HYPERMUTATION USING NEURAL NETWORKS AND APPROXIMATE BAYESIAN COMPUTATION

BY THAYER FISHER[1,a], KEVIN SUNG[2,c], NOAH SIMON[1,b], JULIA FUKUYAMA[3,e] AND FREDERICK A. MATSEN IV[2,d]

[1]*Department of Biostatistics, University of Washington,* [a]*thayerf@uw.edu,* [b]*nrsimon@uw.edu*
[2]*Computational Biology Program, Fred Hutchinson Cancer Center,* [c]*ksung2@fredhutch.org,* [d]*matsen@fredhutch.org*
[3]*Department of Statistics, Indiana University,* [e]*jfukuyam@indiana.edu*

Somatic hypermutation (SHM) is a critical enzyme-mediated process of the adaptive immune response in which antibodies acquire mutations to enhance antigen binding. Despite abundant research elucidating the biochemical basis of SHM, and substantial sequence data available for parameterization, previous computational models of SHM have not been explicitly mechanistic. In this paper we bridge this gap by developing a probabilistic latent variable model encapsulating a sequence of interacting steps, thus formulating the biochemical underpinnings of SHM in a mathematical framework. However, fitting this latent variable model is challenging. To navigate this complexity, we employ an approximate Bayesian computation strategy integrated with neural networks. We are able to estimate almost all of the parameters of the model to good accuracy but find that the parameters involving the boundaries of the nucleotide stripping process are slightly more challenging, given the type of data available.

1. Introduction. Researchers in many biological subfields have developed detailed models describing biochemical processes. These models are often qualitative, not quantitative: for example, the experiments underlying the stated biochemical process may have shown that a certain enzyme is necessary for one part the process; however, these models typically do not describe the rates or relative importances of the different pieces. Given appropriate data, one would like to infer rates and relative importances of the processes. Here we investigate somatic hypermutation (SHM), the biochemically-driven mutational process underlying antibody maturation, as an example of such a process.

Biological theories, such as the one describing SHM, are often given in forms that are easily translated to probabilistic models. Setting up such a model and performing parameter inference can be useful, allowing us to quantify parts of the process that are either not experimentally manipulable or not directly measurable. This can give us insight into the process under normal circumstances and facilitate comparisons of the process under different conditions. We will refer to a probabilistic model that is informed by what is known about the underlying biological processes as a “mechanistic model.” This is in contrast to more standard statistical models (e.g., the classic S5F model of Yaari et al. (2013)) that aim solely at good prediction accuracy and whose parameterizations are not necessarily related to the underlying biology.

To describe the biological context for this paper in a little more detail, we will now give a brief introduction to B cell receptors and affinity maturation. B cell receptors (BCRs), also known as immunoglobulins or antibodies in their secreted form, are proteins designed to bind and identify foreign molecules. These BCRs must be incredibly diverse because of the wide

Received June 2023; revised September 2024.
Key words and phrases. Bayesian statistics, likelihood-free inference, computational biology, adaptive immunology.

diversity of foreign molecules, such as viral proteins, that need to be bound by them. They are generated first by a random process of V(D)J recombination, which involves a BCR gene being created by concatenating randomly trimmed versions of one randomly selected V gene segment, one randomly selected D gene segment, and one randomly selected J gene segment. Random nucleotides are inserted between these segments. Because there are several dozen possible V and D segments, a handful of J gene segments, and many possibilities for trimming and random nucleotide insertion, this process allows for high diversity in BCRs with only a small number of genes.

If BCRs pass tests against self-reactivity and for functionality and are stimulated by binding foreign antigens, the corresponding B cells will enter specialized regions of the lymph nodes where they undergo a process called "affinity maturation." In affinity maturation, BCRs diversify according to a Darwinian process of mutation and selection, with the objective of improving binding to antigen. More specifically, BCRs which exhibit high binding affinity undergo positive selection. While V(D)J recombination increases BCR diversity through combining germline genes randomly, additional diversification through mutating the genes themselves is also necessary to adapt to new antigens.

Our focus is on somatic hypermutation (SHM), the mutation component of affinity maturation, which has a rate in the BCR-coding region of around a million-fold higher than the mutation rate of normal cellular division (Kleinstein, Louzoun and Shlomchik (2003)). This process is driven by specialized biochemical pathways, as described below. Accurate estimation of mutation probability during SHM is important for understanding which antibodies are accessible from a certain naïve progenitor cell. For example, it has been shown that specific low-probability mutations are important for broadly-neutralizing HIV antibody maturation (Bonsignori et al. (2017), Hwang et al. (2017), Wiehe et al. (2018)). Thus, to understand prospects for eliciting such antibodies through vaccination and how to design vaccine inserts rationally (Jardine et al. (2013)), we must estimate mutation probability accurately. Accurate mutation rate estimation is also important for natural selection assessment (Dunn-Walters and Spencer (1998), Hershberg et al. (2008), Uduman et al. (2011), Yaari, Uduman and Kleinstein (2012), Yaari et al. (2015), McCoy et al. (2015)) in which the mutation rate (the introduction rate of nucleotide changes) is compared to the substitution rate (the rate of such changes that survive natural selection).

Current research on SHM can be primarily divided into two approaches: biochemical and statistical. The biochemical approach uses gene knockout or knock-in (Langerak et al. (2007)) experiments in model organisms to elucidate mechanisms of somatic hypermutation. As a prototypical example, Zivojnovic et al. (2014) use a mouse with a specific DNA substrate spliced in to test the effect of the "Ung" pathway for generating mutations at adenine (A) and thymine (T) bases. They found that when the *Ung* gene was knocked out, the mutation patterns at A/T bases were significantly changed. This experiment clearly demonstrates the importance of the Ung pathway but does not provide an estimate about the frequency with which it is used.

The statistical approach, on the other hand, works to characterize the mutation process in humans and wild-type animals by developing predictive quantitative models. It typically works to understand the influence of local sequence context, namely, the surrounding DNA bases, on mutation probability. For example, if a cytosine base is contained in a "hotspot," such as AGCT, many studies have shown that this is sufficient for a greatly increased mutation rate (Yeap et al. (2015)). "Coldspots," such as CCC, have a lower rate of mutation. Early work analyzed these sequence contexts using a hypothesis-testing approach (e.g., Rogozin and Diaz (2004)) while more recent work has performed rate estimation for highly parameterized models (Yaari, Uduman and Kleinstein (2012), Cui et al. (2016)). Other work has included a penalized proportional hazards model (Feng et al. (2019)), models that include absolute position in the sequence (Spisak, Walczak and Mora (2020)), and deep neural networks on larger tracts of sequence (Tang, Krantsevich and MacCarthy (2022)). Such models are useful for predicting the path of affinity maturation (Dhar et al. (2018), Wiehe et al. (2018)) and analyzing natural diversification patterns (Yaari et al. (2015), Sheng et al. (2016), Vieira, Zinder and Cobey (2018)). Furthermore, many of these papers provide information that can clarify mechanistic understanding; for example Yaari et al. (2013) investigate DNA strand symmetry, while Tang, Krantsevich and MacCarthy (2022) interpret their mutability scores in terms of pathways. Many other papers use the pattern of mutations to describe the qualitative importance of the different processes in SHM (Dunn-Walters et al. (1998), Spencer, Dunn and Dunn-Walters (1999), Rogozin et al. (2001), Rogozin and Diaz (2004), Neuberger et al. (2005), Wilson et al. (2005), Wang, Rada and Neuberger (2010)). Nevertheless, this work has not resulted in rate estimates for the various repair pathways in vivo.

It would seem useful to be able to bring these two approaches to SHM research together by fitting parameters of a probabilistic model of SHM; such a model should, in principle, be able to be fit with less data by using prior mechanistic knowledge and should be useful by providing estimates of parameters with direct biological relevance. However, such a project faces several obstacles. First, there are a tremendous number of latent variable patterns that lead to identical mutation locations, and thus calculating a likelihood for a set of mutations, given a starting sequence, is not possible. Second, the mechanistic model may not be identifiable from the data: there may be many parameter settings leading to identical distributions of sequences of mutations. Third, although it is common to have tens of thousands or more mutated sequences from a given data set and each sequence is hundreds of bases long, mutations are relatively few, and the location of the mutations is stochastic. Thus, the signal for fitting a complex model may not be as strong as it initially appears.

In this paper we take on this challenge by using a method combining neural networks and strategies from Approximate Bayesian Computation (ABC) to infer parameters of a complex latent variable model of SHM (Figure 1). We draw parameter values from a prior distribution, simulate data given the parameter drawn from the prior, and compute summary statistics from the simulated data. Our simulator encodes a simplified version of the biological processes underlying SHM, and the constraints it imposes are important in overcoming the obstacles described in the previous paragraph. We then use the joint distribution of parameters and summary statistics as a training set for a neural network, which approximates the posterior expectation of the parameters given the summary statistics. This approach is similar to others who have used neural networks to estimate quantiles of the posterior distribution directly (Fisher et al. (2023)), but we use summary statistics to generate these values. Unlike some other approaches at using neural networks to improve the performance of ABC, which use the neural network to craft the summary statistics (Jiang et al. (2017)), we create the summary statistics ourselves and then use those statistics as inputs to a neural network. Others have attempted to use deep learning in the context of more classical statistical techniques (Jiang et al. (2017), Sheehan and Song (2016)), though generally these works use neural networks to generate informative statistics rather than generate parameter estimates from pre-selected summary statistics.

Although we do not have statistical guarantees of identifiability, we find in simulation that most parameters of the model can be readily identified. Our results also show some deficiencies of the core model of SHM and indicate what additional components are required to have a more complete description of the process. Our method focuses on the sequence-level mutation distribution, at the expense of accurately quantifying local context mutability. In order to guide further research in the area, we also review alternative procedures we tried, which were not successful, in the Discussion.

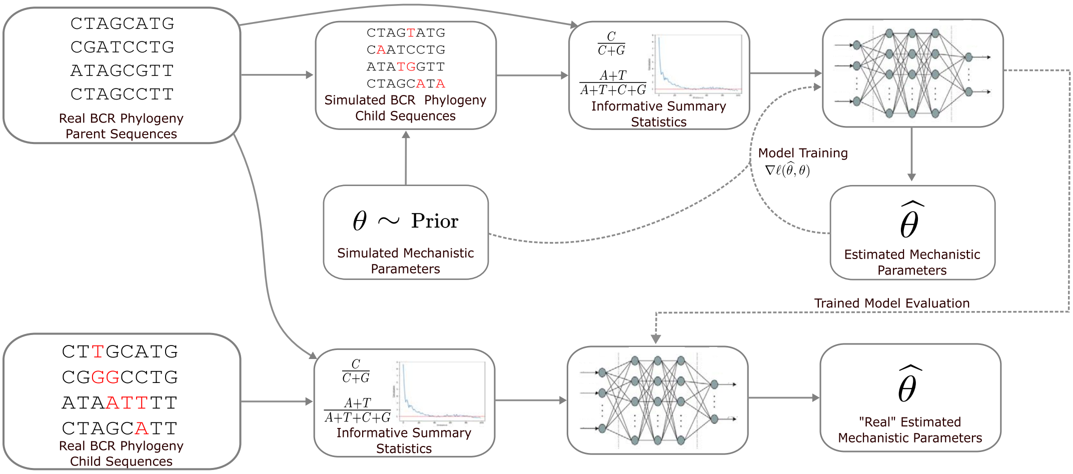
CTAGCATG
CGATCCTG
ATAGCGTT
CTAGCCTT
Real BCR Phylogeny Parent Sequences
CTAGTATG
CAATCCTG
ATATGGTT
CTAGCATA
Simulated BCR Phylogeny Child Sequences
$\frac{C}{C+G}$
$\frac{A+T}{A+T+C+G}$
Informative Summary Statistics
Model Training
$\nabla\ell(\widehat{\theta},\theta)$
$\theta \sim$ Prior
Simulated Mechanistic Parameters
$\widehat{\theta}$
Estimated Mechanistic Parameters
Trained Model Evaluation
CTTGCATG
CGGGCCTG
ATAATTTT
CTAGCATT
Real BCR Phylogeny Child Sequences
$\frac{C}{C+G}$
$\frac{A+T}{A+T+C+G}$
Informative Summary Statistics
$\widehat{\theta}$
"Real" Estimated Mechanistic Parameters

FIG. 1. *Model estimation framework. Parameters θ are drawn from the prior and used to generate simulated parameter-sequence pairs as mutant versions of the real parent sequences. These sequences are then processed into summary statistics, for example, shown here the number of mutations where the parent base is C divided by the total number of mutations where the parent base is C or G, a similar statistic involving the parent being A or T, and descriptions of spatial colocalization. The neural network is trained on these summary statistics to predict the parameters of the model. The trained model is then applied to real sequences to make estimates for the natural process.*

2. Methods.

2.1. *Overview of somatic hypermutation.* Now, we summarize, for the purposes of formalizing our model, current mechanistic understanding of the somatic hypermutation process. The model presented here is necessarily an oversimplification in order to obtain a model that can be fit. Further details on the simplifications employed are in the Discussion section, though others have described the intricacies of the process extensively (Pilzecker and Jacobs (2019)).

The underlying mechanism of SHM involves an initial step of DNA damage by the ctivation-induced cytidine deaminase (AID) enzyme, which transforms a cytosine DNA base to a uracil. AID binds to one of the two DNA strands and traverses processively along the sequence, inducing damage at some fraction of the cytosines it encounters (Pham et al. (2003), Chelico et al. (2006), Mak et al. (2013)). This leads to spatial clustering of AID-induced damage. This defect is resolved by one of several pathways (Figure 2). The comprehensive review by Pilzecker and Jacobs (2019) describes five pathways: U synthesis, UNG2 TLS, ncMMR UNG2 hybrid, ncMMR, and UNG2 PCNA-Ub. The first three pathways lead to mutation only at the location of the lesion, while the last two can lead to mutation at the lesion and in a patch around the lesion.

Of the three pathways that lead to mutation only at the lesion, the major player is one that relies on base excision repair (what Pilzecker and Jacobs (2019), describe as the UNG2 TLS pathway). In this pathway the uracil is removed by the Uracil-DNA glycosylase (UNG) enzyme, and then special enzymes are used to read through the corresponding abasic site, with additional local mutation from translesion polymerase REV1 (Jansen et al. (2006)) and potentially from error-prone DNA polymerases. A mutation can also be introduced at the lesion site if the uracil is not recognized by any of the repair machinery and is used as a template when the DNA is replicated in preparation for cell division (what Pilzecker and Jacobs (2019), call "U synthesis"). This results in a C/G to T/A mutation at the lesion location.

The primary pathway that leads to mutations around the lesion is an error-prone version of the mismatch repair (MMR) pathway specific to B cells undergoing somatic hypermutation

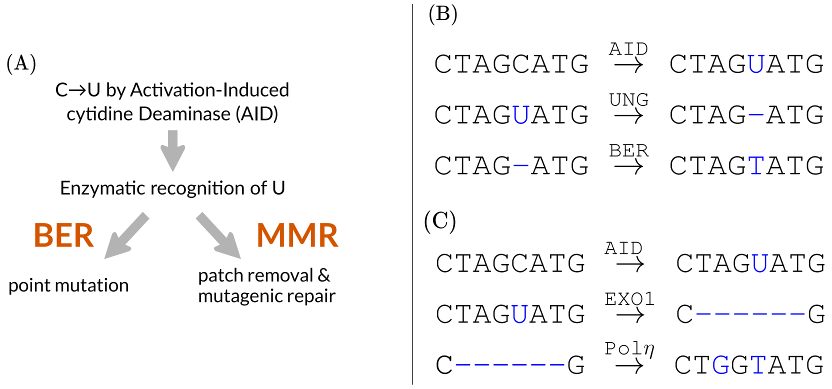

FIG. 2. *A simplified "consensus model" of the biochemical basis of somatic hypermutation* (*SHM*): (A) *AID lesion, recognition, and error-prone repair through the noncanonical base excision repair* (*BER*) *and mismatch repair* (*MMR*) *pathways.* (B) *Error-prone base excision repair* (*BER*) *induces a point mutation at a C site with some probability.* (C) *Error-prone mismatch repair* (*MMR*) *induces potentially multiple point mutations in a window of stochastic size, or exo stripping region, around a C site with some probability.*

("ncMMR" in Pilzecker and Jacobs (2019)). In this pathway, enzymes, EXO1 in particular, strip DNA around a region of the mismatch, which is filled in by error-prone polymerases. This region is referred to as the *exo stripping region*.

All of these processes occur on both strands of DNA. When DNA is replicated, the mutations from one strand are propagated to the other strand.

Given a collection of mutated BCR sequences, we have no way of discriminating between the two pathways that lead to mutations in a patch around the lesion or of discriminating among the three pathways that lead to mutation only at the location of the lesion. We will, therefore, combine the pathways in each category so that our model contains one for each type of pathway. All of the pathways that lead to mutations in a patch around the lesion will be aggregated into what we will call the "MMR pathway," and all of the pathways that lead to mutations at the lesion will be aggregated into what we will call the "BER pathway."

We have not attempted to model insertion-deletion mutations as part of somatic hypermutation (Briney, Willis and Crowe (2012), Lupo et al. (2022)) and process our sequences accordingly. As such, we can assume that every mutation is a point mutation on a sequence of fixed length. We also do not model certain minor pathways (e.g., Polζ; Saribasak et al. (2012)) and assume that their effects will be absorbed into our estimates of the two main pathways.

Some potential interactions between the repair machinery and the damage machinery are also not captured. Our simulator works by sampling lesions, sampling repair times and types for each lesion, and finally sampling repairs to each lesion. The repairs are then executed in order of their times, and each repair is completed before the next repair begins. Later repairs depend on earlier repairs only in the following way: if a lesion is in the exo stripping region of an earlier repair, it is considered already fixed, and it is removed from the list of sites to repair. This procedure keeps some of the true dependence between lesions and repairs (if a lesion is stripped out by exo, it no longer needs to be repaired). However, this model is a simplification in that it only allows for one "round" of mutation, that is, sampling of lesions followed by repair types and repairs. Therefore, we do not account for interaction between the AID machinery and the repair machinery, such as MMR causing a germline A/T/G being changed to a C, leaving it open for deamination by AID. On the other hand, one could have a site be mutated more than once via the MMR machinery.

We will fit probabilistic parameters to this simplified consensus model.

2.2. *Probabilistic graphical model formulation of somatic hypermutation.* Our mechanistic model translates a biochemical SHM model (Figure 2) into a probabilistic graphical model. To generate samples from this model, we first sample a set of AID lesions, which we will call "lesions" for short, then sample repair types for each lesion. The repair process is modeled by sampling a pathway to repair each of the lesions, with random variables governing recruitment probabilities of either the BER or MMR repair machinery to each site. A repaired sequence is then drawn given this choice, with errors introduced by respective error-prone machineries (Methot and Di Noia (2017)). As noted above, this corresponds to one "round" of somatic hypermutation, when in reality the process of lesion introduction and repair could happen multiple times. We simplify to one round of somatic hypermutation because, first, in real data our parent-child pairs don't tend to have very many mutations (the 0.5, 0.9, and 0.99 quantiles of the branch length distribution correspond to just under 1, 7.5, and 24.3 mutations expected per parent-child pair), and second, because of the difficulty in distinguishing between multiple rounds of mutation and higher base rates per round.

We now describe this forward model in detail.

2.2.1. *Forward model.* To model this process, we start with a sequence of length L, called $S^{(0)}$. These sequences could, in principle, be anything, but in this application $S^{(0)}$ will always be one of the inferred ancestral sequences from one of the clonal family trees estimated using IQ-TREE (Nguyen et al. (2014)) and described in more detail in Section 2.6. We will call this sequence the "parent" sequence because in phylogenetic terms the new sequence will descend from the parent sequence. Our random variables are:

- $\mathrm{FW} \in \{0, 1\}$ is an indicator of whether the forward ($\mathrm{FW} = 1$) or the reverse strand is targeted by AID.
- $A = (a_1, \dots, a_K)$, $a_k \in \{0, \dots, L-1\}$. We have K AID lesions, and a_k gives us the position along the sequence of the kth lesion. If $\mathrm{FW} = 1$, these correspond to positions on the forward strand' otherwise, they correspond to positions on the reverse strand.
- $T_{\mathrm{MMR}\,k}$ and $T_{\mathrm{BER}\,k} \in \mathbb{R}^+$ represent the time it takes for either the MMR or the BER machinery to be recruited to lesion k. If $T_{\mathrm{MMR}\,k} < T_{\mathrm{BER}\,k}$, then lesion k is repaired by MMR; otherwise, it is repaired by BER.
- E_k^l, E_k^r. E_k^l and E_k^r give the number of bases stripped out to the left and the right of the kth lesion. When BER is selected, these variables are 0.
- $S_i^{(k)}$, $i = 0, \dots, L-1$, $S_i^{(k)} \in \{\mathrm{A}, \mathrm{T}, \mathrm{C}, \mathrm{G}\}$. $S^{(0)}$ represents the parent sequence. $S^{(k)}$ represents the sequence after k of the lesions have been resolved, and $S^{(K)}$ represents the observed sequence (i.e., the sequence after all K lesions have been resolved). $S_i^{(k)}$ represents the nucleotide at position i in the mutated sequence $S^{(k)}$.

To describe the forward model, we also introduce some extra variables and notation. Let MMR_k be the indicator $\mathrm{Ind}(T_{\mathrm{MMR}\,k} < T_{\mathrm{BER}\,k})$, $T_k = \min(T_{\mathrm{MMR}\,k}, T_{\mathrm{BER}\,k})$ and $\mathrm{pos}(k)$ be the position along the sequence of lesion k. Let $\mathbf{V}$ denote the vector containing all of the random variables aside from those describing the sequence,

$$\mathbf{V} = \big(\mathrm{FW}, A, T_{\mathrm{MMR}\,1}, \dots, T_{\mathrm{MMR}\,K}, T_{\mathrm{BER}\,1}, \dots, T_{\mathrm{BER}\,K}, E_1^l, \dots, E_K^l, E_1^r, \dots, E_K^r\big).$$

Furthermore, define the function $\mathrm{stripped}(k, \mathbf{V})$ as

$$\begin{aligned}\mathrm{stripped}(k, \mathbf{V}) = \mathrm{Ind}\big(&\mathrm{pos}(k) \in \big[\mathrm{pos}(j) - E_{jl}, \mathrm{pos}(j) + E_{jr}\big] \\ &\text{for some } j \text{ s.t. } \mathrm{MMR}_j = 1 \cap T_j < T_k\big).\end{aligned}$$

The stripped function represents lesion k being stripped out before it was repaired and allows us to identify lesions that do not need to be resolved by either the BER or the MMR pathway.

Then FW, A, $T_{\mathrm{MMR}\,k}$, $T_{\mathrm{BER}\,k}$, E_k^l, and E_k^r are generated as follows (the PGCP is described below):

$$\begin{aligned}
\mathrm{FW} &\sim \mathrm{Bernoulli}(p_{\mathrm{FW}}),\\
A &\sim \mathrm{PGCP}(\mu, \sigma, \ell),\\
T_{\mathrm{MMR}\,k} &\sim \mathrm{Exponential}\big((1-p_{\mathrm{BER}})/p_{\mathrm{BER}}\big),\\
T_{\mathrm{BER}\,k} &\sim \mathrm{Exponential}(1),\\
E_k^l &\sim \begin{cases}\mathrm{Geometric}(1/\mu_{\mathrm{el}}) & \mathrm{MMR}_k = 1,\\ 0 & \text{otherwise},\end{cases}\\
E_k^r &\sim \begin{cases}\mathrm{Geometric}(1/\mu_{\mathrm{er}}) & \mathrm{MMR}_k = 1,\\ 0 & \text{otherwise}.\end{cases}
\end{aligned}$$

To describe the distribution of $S^{(K)}$, we will assume that $\mathrm{FW} = 1$. If not, $S^{(K)}$ is generated by the same process happening on the reverse strand,

$$\begin{aligned}
&S_i^{(k)} \mid S_i^{(k-1)}, \mathbf{V}\\
&\quad\sim \begin{cases}
S_i^{(k-1)} \text{ w.p. } 1 & \text{if } \mathrm{stripped}(k, \mathbf{V}) = 1 \cup i \notin [\mathrm{pos}(k) - E_k^l, \mathrm{pos}(k) + E_k^r],\\
(\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C})_{\mathrm{Categorical}(\pi^{(M)}_{S_i^{(k-1)}})} & \text{if } \mathrm{MMR}_k = 1 \cap i \in [\mathrm{pos}(k) - E_k^l, \mathrm{pos}(k) + E_k^r],\\
(\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C})_{\mathrm{Categorical}(\pi^{(B)})} & \text{if } \mathrm{MMR}_k = 0 \cap i \in [\mathrm{pos}(k) - E_k^l, \mathrm{pos}(k) + E_k^r].
\end{cases}
\end{aligned}$$

For each lesion, each position on the sequence either doesn't need to be resolved (because the lesion was stripped out previously or because the position is not targeted by the repair machinery) or should be resolved by the MMR or the BER machinery. Note that $S^{(0)}, \dots, S^{(K)} \mid \mathbf{V}$ is a Markov process, where $S^{(1)}, \dots, S^{(K-1)}$ are hidden states. However, the standard techniques for hidden Markov models do not apply here, as we do not have an observation for each hidden state in the process; furthermore, there are other latent variables that are not part of the Markov process.

One important feature of our lesion vector A is that the lesion locations should be spatially correlated. To achieve this, the lesions are the result of thinning what we call *prelesions*. At each site k, there is some prelesion probability p_k. These probabilities are first sampled independently across sites to obtain the prelesion set. Following this, we thin the prelesions according to a Gaussian process with a particular mean, variance, and lengthscale. In particular, we include all prelesions only at sites where $G_k > 0$, where G_k is the value of the Gaussian process at site k. We choose to do this in order to enable strong spatial correlation due to AID moving in a processive manner along the sequence (Pham et al. (2003), Chelico et al. (2006), Mak et al. (2013)) and mesoscale-sequence effects on AID deamination deriving from local DNA sequence flexibility (Wang et al. (2023)). Lesions at sites which are close together in a sequence are positively correlated. This distribution for the lesions is called a *Probit Gaussian Cox Process* (PGCP), a modification of the sigmoidal Gaussian Cox process (Møller, Syversveen and Waagepetersen (1998)). A more detailed definition of a PGCP can be found in Section B.3 of the Appendix (Fisher et al. (2025)), which can be found in a separate document. Instead of applying the sigmoid function to the Gaussian process to obtain the thinning probability, as in a sigmoidal Gaussial Cox process, prelesions are thinned if and only if the Gaussian random variable associated with them is negative.

The parameters are:

- $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$, $\ell \in \mathbb{R}$: Parameters controlling the overall rate and correlation of the lesion generating Cox process. For the purposes of identifiability, we will assume that μ is fixed at -10.0.
- p_{BER}: Parameter controlling the relative rate of BER recruitment to MMR recruitment.
- μ_{el}, μ_{er}: The lengths of the exo stripping region to the left and the right of the lesion is modeled as a geometric random variables with means μ_{el} and μ_{er}.
- p_{fw}: Parameter controlling the relative rate of transcription of the forward strand.
- $\pi^{(M)}$: A 4×4 transition probability matrix, assumed known, describing the transition between nucleotides after MMR. $\pi_A^{(M)}$ represents the row of the matrix corresponding to transition away from A and, similarly, for the other bases. The matrix has a nonzero diagonal, allowing for error-free repair by the MMR pathway.
- $\pi^{(B)}$: A length-4 vector, assumed known, giving the probability of each base being incorporated given repair by the BER pathway. All elements of this vector are nonzero, allowing for error-free repair by the BER pathway.

Our estimand is, therefore, $\theta = \{\sigma, \ell, p_{\mathrm{BER}}, \mu_{\mathrm{el}}, \mu_{\mathrm{er}}, p_{\mathrm{FW}}\}$.

For our choice of $\pi^{(B)}$, we reference work from Krijger et al. (2009). They find that Ung, the enzyme largely responsible for BER, induces mutations in a particular pattern. When examining mice who were deficient in Msh2, the enzyme believed to be responsible for mutations on the MMR pathway, they found that 75.4% of Guanine mutations were to Adenine, and 77.9% of Cytosine mutations were to Thymine. These results simplify our model greatly, as they allow the source pathway of an observed mutation to be determined with higher confidence.

For $\pi^{(M)}$, we estimated the transition matrix using a strategy based on examining neighboring mutations; see Section B.1 in the Appendix (Fisher et al. (2025)) for details and justification.

2.3. *Estimating model parameters with neural network ABC.* One of the problems with estimating parameters in settings with latent variables, such as A, MMR, and E here, is that we cannot easily calculate $\mathbb{P}(\theta \mid S)$, even via stochastic techniques (e.g., MCMC) which only require calculating $\mathbb{P}(S \mid \theta)$. The difficulty in calculating $\mathbb{P}(S \mid \theta)$ stems from having to integrate over all latent states which could have generated S, which are too numerous to be computationally tractable.

In order to estimate $\mathbb{E}(\theta \mid S)$, we use a neural network-based estimation procedure which is similar to approximate Bayesian computation (ABC); however, in our case the target is posterior mean estimates of the parameters rather than a full approximate posterior. A visual description of the estimation framework can be seen in Figure 1. In ABC-based methods, one generates data under a candidate parameter set θ'. The parameters, θ', are then accepted or weighted by some metric ρ of similarity of S' to observed data. Typically, ρ is based not on the high-dimensional observed data S but on some low-dimensional summary $f(S)$. In our framework, rather than needing to specify ρ to obtain an entire approximate posterior, we instead input $f(S)$ into a neural network, which then outputs an estimate of θ'. Our proposed method differs from ABC in that we are not attempting to estimate/sample from the posterior distribution but rather trying to estimate the posterior mean. It is similar in that both approaches use summary statistics to deal with high-dimensional data. One of the reasons why we have employed this strategy is that the summary statistics we use here are high-dimensional, ($>$100-dimensional), meaning that even some of the sophisticated likelihood-free methods such as ABC-Markov chain Monte Carlo (Marjoram et al. (2003)) would struggle.

It is important to note that the posterior mean is the true minimizer of mean-squared error with respect to the joint distribution of θ, $f(S)$, where $f(S)$ are the summary statistics of sequences S simulated under mechanistic parameters θ, that is,

$$\mathbb{E}(\theta \mid f(S)) = \arg\min_{g} \mathbb{E}((g(f(S)) - \theta)^2),$$

where the minimization is over measurable functions. Here the expectation is taken with respect to the joint distribution of θ, $f(S)$. By generating many pairs of parameter sets and corresponding simulated sequences, we can create a large pool of training data, $\{\theta_i', f(S_i')\}_{i=1}^N$. We can then train a neural network to approximate $\mathbb{E}(\theta \mid f(S))$ using mean-squared error as our loss function. Provided the neural network is sufficiently rich, and our function $f(\cdot)$ captures most of the relevant information about θ present in observed sequences, we will train a good estimator of $\mathbb{E}(\theta \mid S)$.

This framework allows us to move from a high-dimensional latent-state Bayesian inference setting to a low-dimensional supervised learning setting. Provided we can simulate a large enough set of sequences, we can estimate $\mathbb{E}(\theta \mid S)$ without ever estimating the latent states which define the association between θ and S. Most of the difficulty in estimation with this framework lies in specifying the correct neural network architecture and constructing informative summary statistics for S.

In addition to estimating $\mathbb{E}(\theta \mid S)$, we also work to quantify uncertainty via residual estimation. To do this, we also predict the log squared residual and minimize $(\exp(\hat{R}) - (\hat{\theta} - \theta)^2)^2$, where $\hat{\theta}$ is our estimate of the parameter set and $\hat{R}$ is the log-squared residuals estimate.

2.4. *Summary statistics.* We will now describe the various summary statistics we use, with their (Nicknames) and how they aid in inferring key mechanistic parameters.

2.4.1. *Spatial colocalization* (Colocalization). One of the key components of our set of summary statistics is the spatial colocalization (Spisak, Walczak and Mora (2020)), which describes the observed clustering of mutations in SHM. Unlike local models of mutability, such as S5F (Yaari et al. (2013)), our mechanistic framework allows for an elevated frequency of close-together mutations, as seen in real sequences. Local models use several bases on either side of a particular site to model the probability of mutation at that site; the most common is a 5-mer model, which uses two bases on either side. These lack the complexity to model correlation in mutations. A visualization of the shortcomings of a 5-mer model in this respect can be seen in Figure 3. Other works have incorporated wider sequence context in probabilistic

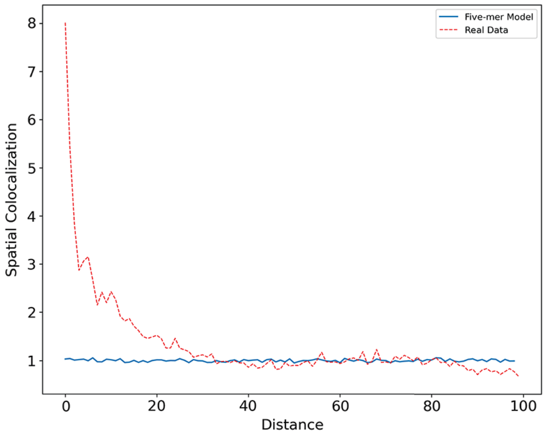

FIG. 3. *True spatial colocalization vs. 5-mer simulated colocalization. Mutations happen much closer together than the local context model accounts for.*

modeling, but attempts to mimic the bunches of mutations seen in real sequences generally involve a second round of "follow-up" mutations which explicitly occur near existing mutation (e.g., Spisak, Walczak and Mora (2020)). However, such models are phenomenological rather than explicitly mechanistic and still render the likelihood intractable.

More formally, spatial colocalization measures the degree to which mutations occur closer together than what one might expect were they spatially independent. The spatial colocalization of Spisak, Walczak and Mora (2020), for a pair of sites indexed by i, $i + d$, is defined as

$$\frac{p(i, i+d)}{p(i)p(i+d)}, \tag{1}$$

where $p(i)$ is the probability of a mutation at site i. So if the two sites are independent, the colocalization value will be 1. We are interested in the colocalization averaged over all pairs distance d apart. This statistic quantifies the spatial effect on co-mutation over an entire dataset/repertoire.

Estimating this quantity for real sequence data presents some challenges. If all observed sequences were globally aligned and we had unlimited data, we could estimate this fraction empirically for every site. However, in practice, it is possible to have many different parent sequences from which observed sequences mutate, and these sequences do not fit neatly into a multiple sequence alignment. It is also possible that we have few mutations per site, making the variance of our estimates of the pairwise colocalization high. These issues make estimation of equation (1) infeasible for the data we will be engaging with in this work.

Thus, one must use an estimator for colocalization rather than attempting to compute it directly. If we have N sequences, sequence i has length ℓ_i, m_i^r is the number of pairs of mutations a distance r apart from each other, and $m_i^+ = \binom{\text{\# mutations in sequence } i}{2}$ is the total number of pairs of mutations in sequence i, the "SWM" (Spisak, Walczak and Mora (2020)) colocalization estimate is

$$\text{coloc}_{\text{SWM}}(r) := \frac{1}{N} \sum_{i=1}^{N} \frac{m_i^r / m_i^+}{(\ell_i - r)/\binom{\ell_i}{2}}. \tag{2}$$

This version of spatial colocalization can be thought of as the fraction of pairs of mutations distance r apart, compared to the fraction of total sequence pairs which are distance r apart. Critically, this quantity is estimated per-sequence and does not require us to estimate the per-site mutation frequency anywhere. The result is that the estimates of the spatial colocalization are more stable in smaller samples and in samples where there are only a small number of mutated sequences for each parent sequence. We will use a vector of 100 of these spatial colocalizations as a summary statistic, as described below.

2.4.2. *Mutation frequencies* (Base Freq, AT Frac, *and* C / C+G). We will use several mutation frequencies as informative summary statistics. The first is the overall baseline mutation frequency (Base Freq), or the fraction of total bases which differ from the parent sequence. We care about the overall mutation frequency because it tells us about the relative rate of prelesion thinning. The second is fraction of total mutations for which the parent base is A or T (nicknamed AT Frac). We care about this fraction because A/T mutations can only occur on the MMR pathway. The final mutation fraction of interest is the total number of mutations where the parent base is C, divided by the total number of mutations where the parent base is C or G (nicknamed C / C+G). This is because the BER pathway gives rise to mutations at C sites if the lesion was on the forward strand and gives rise to mutations at G sites if the lesion was on the reverse strand.

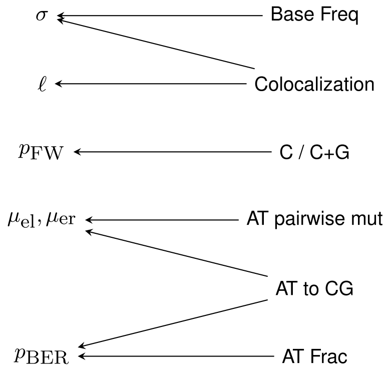

FIG. 4. *The collection of parameters* (*left*) *and the summary statistics designed to infer them* (*right*). *Arrows designate how we intended the summary statistics to inform the parameter estimation. However, as described in Section* 3.2, *the actual influence is more complex.*

2.4.3. *Distance between mutation-associated sites* (AT to CG *and* AT pairwise mut). Finally, we include two distance metrics in our set of summary statistics. The first is the average distance of a mutation at an A/T site from the nearest C/G site (AT to CG). The idea behind this summary statistic is that, for a larger exo stripping region, A/T mutations will be slightly farther away from C/G sites on average. The second distance metric is the pairwise A/T mutation distance (AT pairwise mut). This measures the mean pairwise distance between mutations at A/T sites. We include this summary statistic because the size of the exo stripping region impacts the clustering of A/T mutations. For instance, if we know that multiple A/T mutations happen in an exo stripping region, then we expect them to be closer together in a small exo stripping region in contrast to a large exo stripping region.

2.5. *Parameters of interest.* We will now describe parameters of interest in our model, their priors, and how the aforementioned summary statistics allow us to infer them (Figure 4).

2.5.1. *Gaussian process variance* (σ). The Gaussian process variance controls the overall prelesion thinning probability by governing the variability of the underlying Gaussian process. We take a uniform prior from 5.0 to 15.0 on the standard deviation of the Gaussian process. This includes the entire range of reasonable values, where the smallest value corresponds to almost all the prelesions being thinned, and the largest value corresponds to a thinning rate of 75%, out of a minimum of 50% of the prelesions being thinned. The lower bound of 50% is due to the technical details of the Probit Gaussian Cox Process (PGCP).

Along with the base rate, the variance of the Gaussian process affects the overall fraction of C sites which are deaminated by AID as follows. Recall that the mean of the Gaussian process is fixed at -10.0. This means that the probability that a prelesion is thinned is proportional to $\int_{10/\sigma}^{\infty} e^{\frac{-x^2}{2}}$.

Thus, in terms of summary statistics, a large value of σ, along with a high base rate, contributes to a high overall mutation frequency. Furthermore, a smaller value of σ corresponds to a larger colocalization magnitude (higher colocalization when $d = 1$).

Therefore, we include distance 1 spatial colocalization as well as the overall mutation frequency, with the idea that it will help estimate the variance of the Gaussian process.

2.5.2. *Gaussian process lengthscale* (ℓ). The Gaussian process lengthscale controls the spatial scale of correlation in deamination probability. We have a uniform prior from -12.0 to -2.0 for the log of the lengthscale. The lengthscale affects the rate of decrease of the colocalization. A log-lengthscale of -12 corresponds to a small amount of correlation between neighboring sites ($\sim$0.4), dropping off to near zero for sites separated by two or more bases. In contrast, log-lengthscale of -2 corresponds to a correlation of $>$0.99 for sites separated by up to 15 bases and a correlation of $>$0.9 for sites separated by up to 50 bases. Thus, prior covers the entire range of plausible scenarios.

To help estimate the lengthscale, we simply include the spatial colocalization vector from distance 1 to distance 100 using the SWM estimator described above. The rate at which the colocalization decays to 0 is used by the neural network to estimate the lengthscale.

2.5.3. *BER probability* (p_{BER}). This parameter describes the probability that the BER repair pathway will be recruited to a lesion. If not repaired by BER, a lesion is repaired by MMR. We have a uniform prior from 0 to 1 for this parameter. Notably, MMR allows for mutations in a neighborhood around the lesion, rather than just at lesion itself. Summary statistics which contain information about this statistic include the fraction of mutations occurring at A/T sites as well as the average distance of A/T mutations from a C/G site. These are informative for the BER probability parameter because a higher frequency of these mutations, as well as these mutations occurring closer to C/G sites, suggests a lower fraction of lesions which are repaired by BER.

2.5.4. *Forward probability* (p_{FW}). This parameter controls the probability that the given round of mutation comes from the forward strand. We have a uniform prior from 0 to 1 for this parameter. A higher forward strand probability parameter implies that we see a disproportionate number of sequences originating from the forward strand.

The main summary statistic that informs this parameter is the fraction of C/G mutations which are at C sites. Repair of a lesion on the forward strand gives rise to a C mutation, while repair of a lesion on the reverse strand gives rise to a G mutation. Therefore, the fraction of C/G mutations at C sites gives us information about whether the forward or the reverse strand is more likely to be targeted.

2.5.5. *Exo left/right stripping means* (μ_{el}, μ_{er}). These two parameters control the mean size of the exo stripping region on either side of the lesion when the MMR pathway is recruited. The parameter is the mean length of the stripping region, which is geometrically distributed. We have a uniform prior from 1 to 20 for the mean stripping region size, that is, $\mu_{\text{el}} \sim \text{Unif}(1, 20)$. and $\mu_{\text{er}} \sim \text{Unif}(1, 20)$. This prior is based on the biologically plausible range suggested in the literature (Unniraman and Schatz (2007)).

Summary statistics, which contain information on these parameters, include the average distance to the left and to the right of C/G sites from A/T mutations. These parameters are the most difficult to estimate in this model, because it is difficult to determine from the observed data if two A/T mutations exist in the same exo stripping region or if they originated from different lesions.

2.6. *Data*. In order for our model fit to reflect exclusively the mechanistic properties of somatic hypermutation, it is critical that we train it on data which are free of the impact of natural selection. If we attempted to estimate the mechanistic parameters on sequences which were subject to the full process of affinity maturation, it would be impossible to determine which patterns observed were due to selective pressure and which were due to SHM mechanisms. Therefore, we engage with out-of-frame sequence data, sequences which are nonfunctional due to codons being out of frame after the process of V(D)J recombination. Specifically, we use the out-of-frame clonal family clustering and multiple sequence alignments using data from Briney et al. (2019) and analyzed by Spisak, Walczak and Mora (2020), who use additional filtering criteria to make sure that the data represents neutral evolution.

From these, we filter out all sequences containing insertions or deletions, as these are not a part of our mechanistic model. For each of these clonal families, we use IQ-TREE (Nguyen et al. (2014)) to infer the phylogeny, or evolutionary history of the clonal family. Ancestral sequences are inferred using the K80 substitution model (Kimura (1980)), with inferred naive sequences from Spisak, Walczak and Mora (2020). Following phylogenetic inference, we consider each edge of the tree to be an observation, with the mechanistic SHM process acting independently on each edge. In total, our biological dataset contained 84,322 parent-child pairs.

2.7. *Training setup.* We trained our neural network on simulated data generated from our model of somatic hypermutation using parent sequences drawn from the biological data set. To generate each simulated dataset, we sampled 2000 parents at random from phylogenies inferred from the out-of-frame sequence data as well as a set of mechanistic parameters (θ from its prior distribution above). For each parent sequence, we simulate a single round of the somatic hypermutation process, which can generate multiple mutations in the sequence. Our estimation procedure ensures that the mutation frequency is approximately equal to that in real data. We experimented with incorporating branch lengths and extra rounds of mutation, but it did not appear to influence the accuracy of parameter estimation and added a substantial computational cost. After simulating a sequence for each parent according to the forward model, we calculated the summary statistics described above and input them into our neural network, which then output an estimate for the posterior mean of θ. This is the procedure outlined in Figure 1, with the goal of minimizing mean-squared error. The neural network was trained on a set of 10,000 such simulated datasets, and evaluated on a held-out test set of 1000 such datasets. In order to avoid the relative values of the summary statistics adversely affecting our estimation, we standardized each summary statistic to have mean zero and unit variance.

The neural network used for training was a simple feedforward neural network (Svozil, Kvasnicka and Pospichal (1997)) with five hidden layers and 64, 32, 16, 8, and 4 nodes per hidden layer, respectively. Training was done with Adam (Kingma and Ba (2014)) in PyTorch (Paszke et al. (2019)) using mean-squared error as the loss. We used mini-batches of size 200 and trained for 1000 iterations with step size 10^{-2}.

Our implementation of the forward model as well as all the code used for inference is available at https://github.com/thayerf/shmpy.

3. Results.

3.1. *Performance on simulated data.* We assessed how accurately our neural network recovered the mechanistic parameters by testing it on 1000 held-out simulated datasets (Figure 5). We were able to explain approximately 75% of the prior variance with our chosen summary statistics and neural network described above. We recover the value of the length-scale, the Gaussian process variance σ, the forward probability, and BER probability with high accuracy. We struggle to estimate the size of the exo stripping region to the left and right of the excised base with the same level of accuracy. However, when taking the sum, as in Figure 5, we find that we can recover the size of the exo stripping region with reasonably high accuracy. This suggests that the direction of the exo window is difficult to identify in some cases.

We also assessed our ability to quantify uncertainty (Figure 6). Our estimates of the squared error are weakly correlated with the true value but are much noisier than our estimates of the parameters themselves.

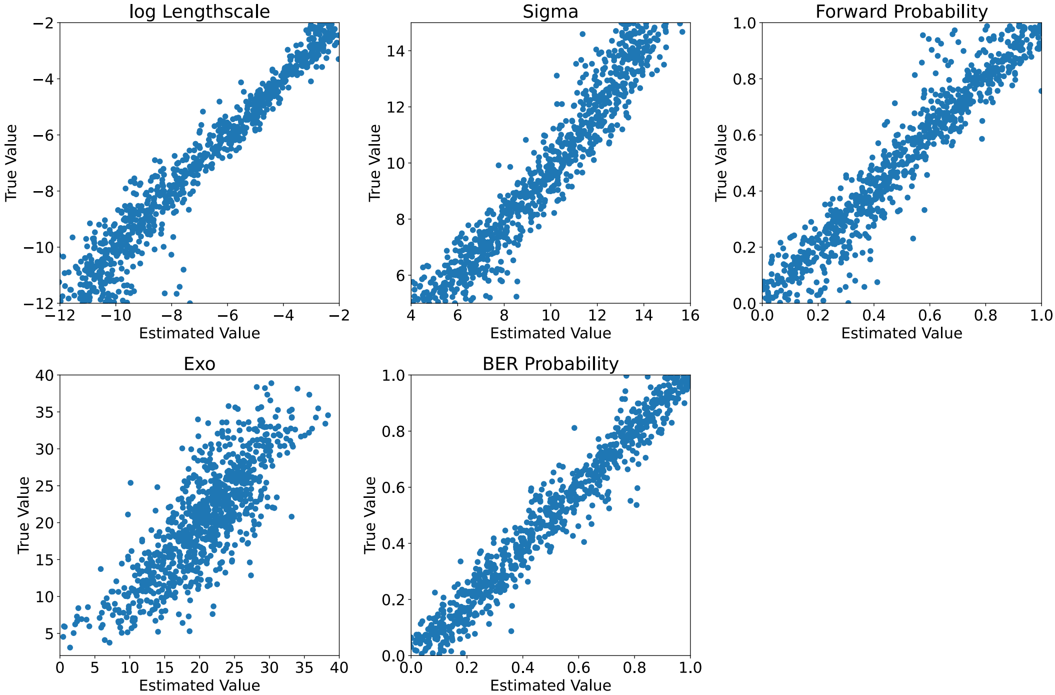

FIG. 5. *True vs. estimated parameters for* $\log \ell$ (log Lengthscale), σ (Sigma), p_{FW} (Forward Probability), $\mu_{\mathrm{el}} + \mu_{\mathrm{er}}$ (Exo), *and* p_{BER} (BER Probability). *Despite the inclusion of several summary statistics which inform the exo window size, the approximate posterior variance is slightly higher. We show the sum* $\mu_{\mathrm{el}} + \mu_{\mathrm{er}}$ *of the exo stripping region parameters because this is recovered more accurately than either of the two directional components* μ_{el} *and* μ_{er}. *Other parameters are estimated accurately.*

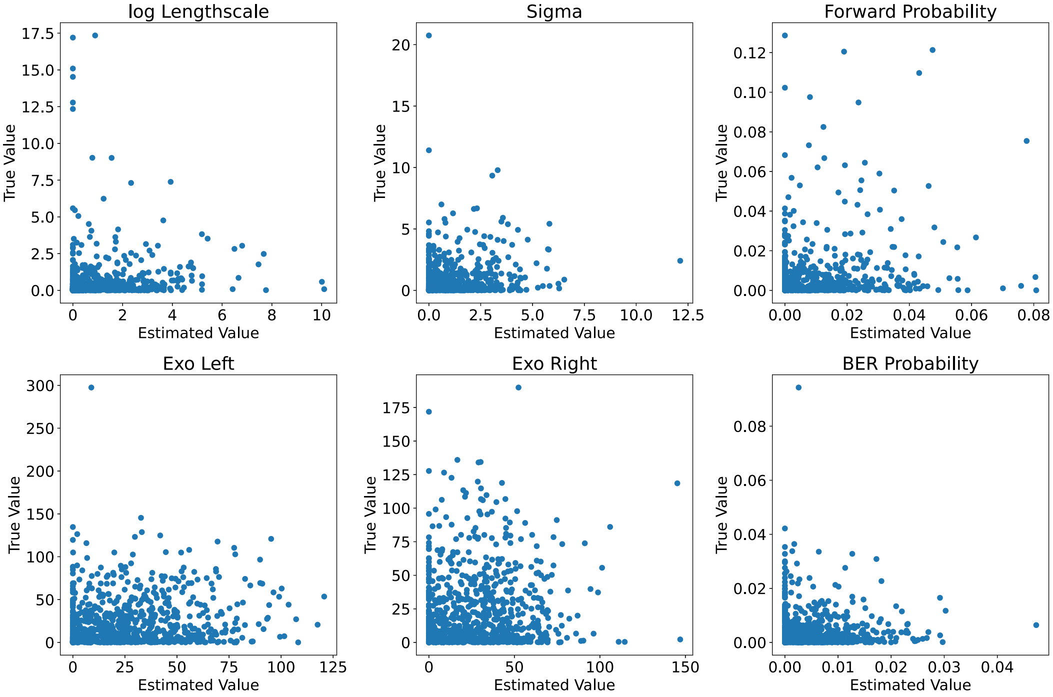

FIG. 6. *Neural network test set squared error estimate vs. true value for* $\log \ell$ (log Lengthscale), σ (Sigma), p_{FW} (Forward Probability), μ_{el} (Exo Left), μ_{er} (Exo Right), *and* p_{BER} (BER Probability).

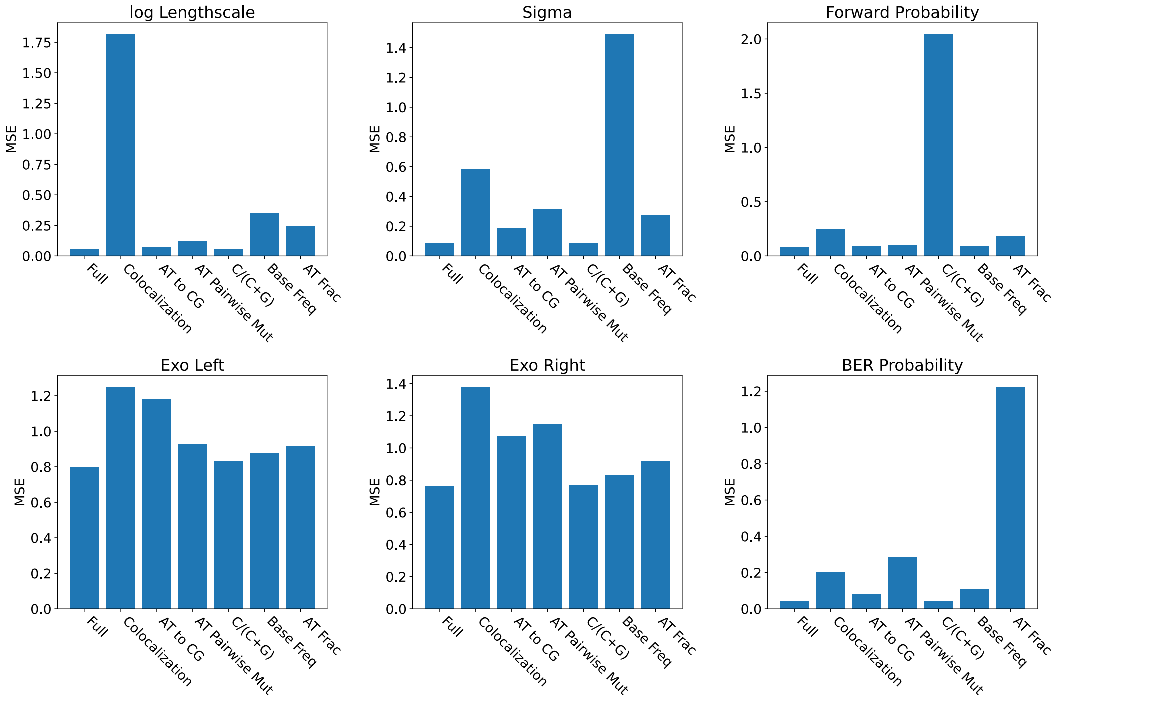

FIG. 7. *Neural network test set mean squared error (MSE) vs. which summary statistic was shuffled, stratified by parameter. "Full" denotes the full model without shuffling. A higher MSE relative to the full model implies that a particular summary statistic is informative for that parameter.*

3.2. *Importance of summary statistics for parameter inference.* In order to explore the importance of each summary statistic in predicting each parameter, we performed an experiment in which we shuffled the values of each summary statistic in our test set among simulation replicates. After shuffling one of the summary statistics, we then recorded the MSE for each parameter and compared it to the MSE of the full model. Summary statistics, which are important for estimating a given parameter, will greatly increase the MSE for that parameter when permuted. The mean squared error vs. the summary statistic removed can be seen in Figure 7. We found that colocalization is the most informative summary statistic for the lengthscale, the base mutation frequency is most informative for the variance, the fraction of C/G mutations which occur at C sites is most informative for the forward probability, and the A/T mutation fraction is most informative for the BER probability. For the exo window sizes, it appears that all of the summary statistics contribute relatively equally. However, a large part of this reduction in mean squared error is due to strong predictive performance in parameters related to the spatial colocalization of AID lesions.

3.3. *Posterior predictive checks.* We also wanted to confirm that summary statistics from real mutated sequences fell in a range which could be credibly generated by our forward model. All but one of the summary statistics lie in the high-density range of the simulated sequences (Figure 8). The notable exception is the distance from an A/T mutation to the nearest C/G site. Real A/T mutations are farther away from C/G sites on average than can be explained by our forward model. One possible explanation is the faithful repair of the central deaminated cytosine in the MMR pathway, which we do not account for. Additionally, we conducted posterior predictive checks verifying that sequences generated under the estimated parameter set, when run through our estimation procedure, recover the same set of parameters. Details of these simulations can be found in Appendix C (Fisher et al. (2025)).

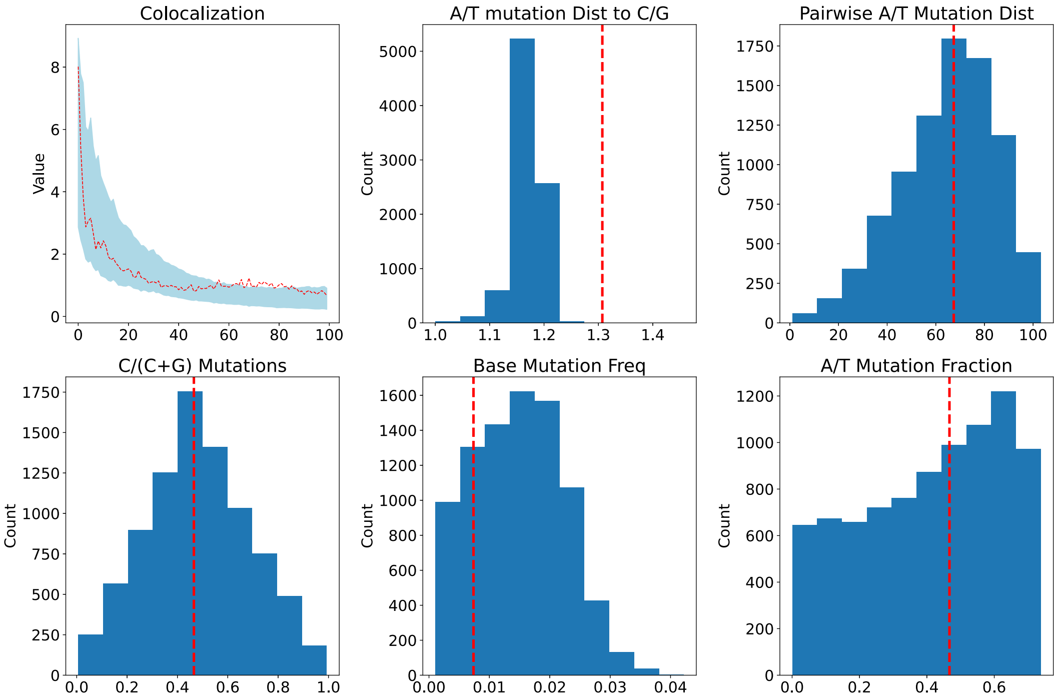

FIG. 8. *Real data summary statistics compared to their simulated data equivalent. The real sequence summary statistics* (*red dashed line*) *fall into a range that can be credibly explained by our forward simulation framework* (*histogram*), *except for the A/T mutation distance to a C/G site. For the colocalization plot*, *the shaded region represents the central* 90% *of colocalization values at each distance*.

We also wanted to make sure that our estimated mechanistic parameters for the real sequences fell in the prior range. A comparison of the prior distribution to the estimated parameter values for real mutated sequences can be seen in Figure 9. We can see that our prior contains feasible values for real sequences.

Finally, we were interested in how our model performed on a per-site mutability level. To investigate this, we wished to compare loss on a logistic regression model which estimates the mutation frequency at a given site using only the 5-mer located there, analogous to the classic S5F model (Yaari et al. (2013)). However, it is difficult to calculate the per-site mutation frequency of our mechanistic model because, unlike regression-based approaches, we do not estimate per-site probabilities directly. In order to obtain per-site mutability estimates, we forward simulated 10,000 sequences each from 1000 parent sequences according to our fit mechanistic parameters. We then averaged over these 10,000 simulated sequences to get a per-site estimate of mutability for each of the 1000 parent sequences. We calculated an average per-site log loss by comparing these estimates to vectors of mutation indicators for the true 1000 child sequences. The same per-site calculation was conducted for the 5-mer model. A 5-mer model outperforms our mechanistic model with respect to per-site mutability, with a per-site log loss of 0.315 for the 5-mer model compared to 0.345 for the mechanistic model. This suggests that our mechanistic model, which was designed to capture global mutation patterns, is unable to attain the local accuracy of a 5-mer model.

3.4. *Cross-colocalization.* Both to check on our model's performance on a measure that was not used for training and to get insight into processes our model is missing, we looked at a measure of colocalization that takes base identity into account. We refer to this quantity as the *cross-colocalization*: the cross-colocalization for bases b_1 and b_2 at a distance r is defined as the ratio between the fraction of pairs of mutations containing bases b_1 and b_2 at distance

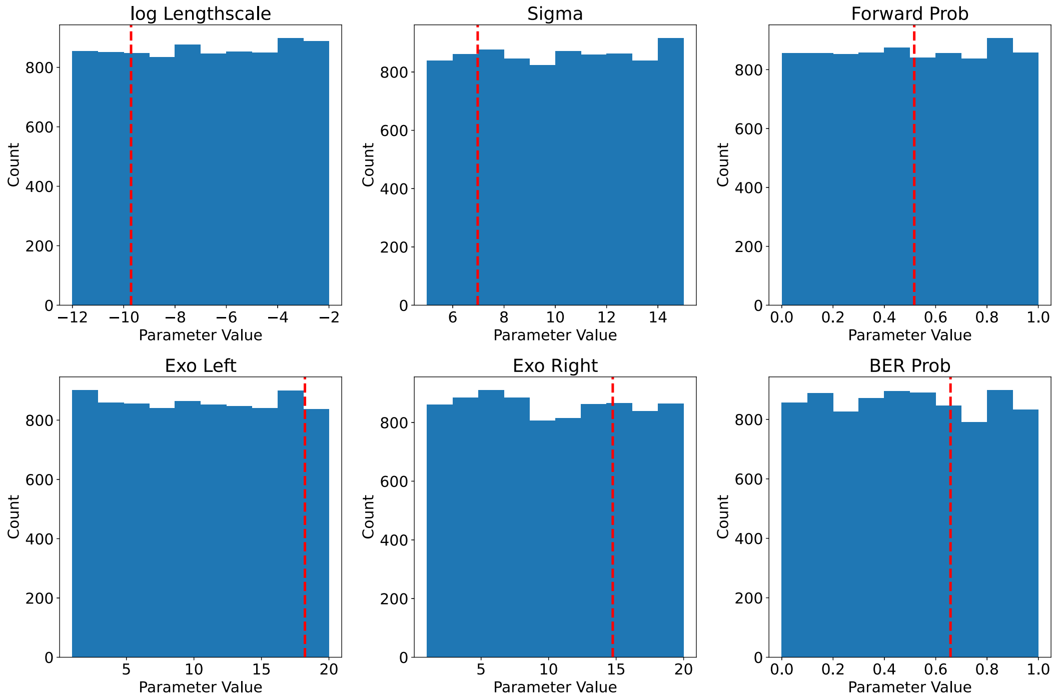

FIG. 9. *Real data estimated parameter values (red dashed lines) vs. histogram of prior distribution (blue bars). All of the estimated governing parameters for the real data fall into the domain of our priors, suggesting that our forward model is calibrated to have realistic relative rates of the various pathways.*

r and the fraction of pairs of locations on the sequence containing bases b_1 and b_2 at distance r (see equation (2)). This quantity is a generalization of the colocalization defined by Spisak, Walczak and Mora (2020) and is described in more detail in Appendix Section B.2 (Fisher et al. (2025)). Measures with a similar goal have been described in other papers (Krantsevich, Tang and MacCarthy (2020)), but they require more information to compute. Specifically, they require per-site estimates of mutation probability, which would be available to us only if we used a parametric model or restricted ourselves to a subset of the sequences.

We perform a comparison of the cross-colocalizations between the real and simulated data in Figure 10. As expected, the cross-colocalizations fall off as a function of r in both the real and the simulated sequences. The cross-colocalizations involving C or G are overall more comparable between the real and simulated sequences than the AA/TT/AT cross-colocalizations. The primary difference between the cross-colocalizations involving C or G occurs at distance 1 for C/G pairs: that cross-colocalization is substantially higher in the real data than in the simulated data. This large difference in cross-colocalization at distance 1 is evidence in favor of a specific process not included in our model that only acts on C/G pairs at distance 1. Two such processes that have been suggested in the literature are overlapping hotspots (Wei et al. (2015)) and tandem mutations (Saribasak et al. (2012)). This particular result gives more credence to the overlapping hotspots explanation, as the difference in cross-colocalization occurs for only one pair of bases, and the tandem mutation pathway is not hypothesized to be specific to a certain set of bases.

The other main difference between the cross-colocalizations computed on the real and simulated data is that the AA, TT, and AT cross-colocalizations are higher overall and drop off more slowly in the simulated data than in the real data. Similar to cross-colocalization results for pairs of bases including C or G, these results show that the colocalization pattern for AA/AT/TT pairs can not be explained without resorting to explanations involving overlapping hotspots or extra pathways not included in the model. This suggests that there may

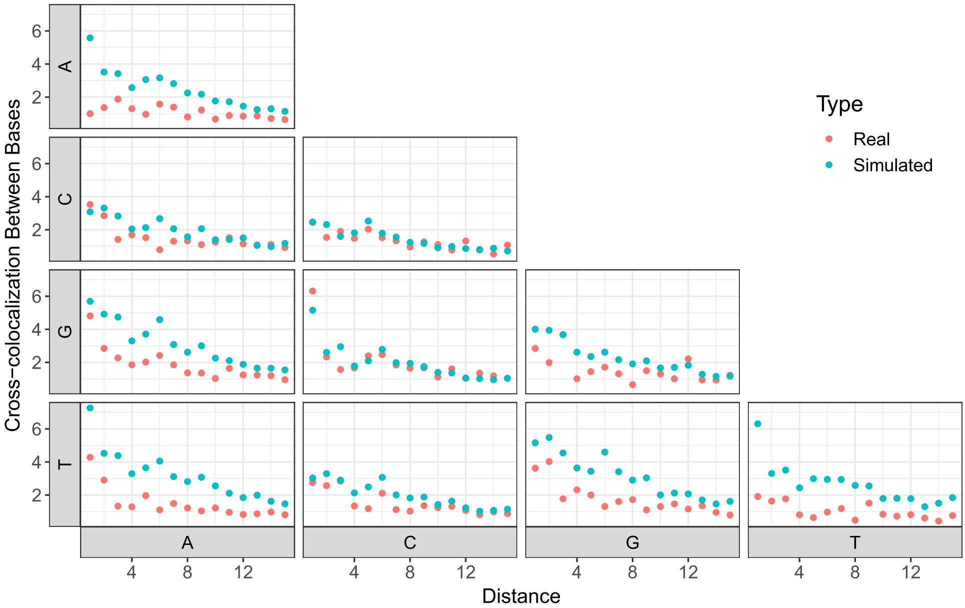

FIG. 10. *Cross-colocalization for data from Spisak, Walczak and Mora* (2020) *and simulated sequences. This quantity is the colocalization between sites with bases as written in the row and column labels.*

be additional layers of complexity not included here which contribute to certain mutation patterns.

4. Discussion. The forward simulation and estimation framework we describe is, to our knowledge, the first work to try to fit mechanistic parameters to the consensus model for somatic hypermutation. While previous work has focused on modeling mutability through local sequence context or isolating mechanistic pathways via knockout experiments, we present a method which estimates relative rates of known mechanistic pathways. Here we have built a simplified forward model and developed means of estimating its mechanistic parameters using a neural-network-based ABC-like framework.

Other work has employed deep learning in ABC. Jiang et al. (2017) use deep neural networks to estimate the posterior mean for parameters in a stochastic genetic oscillator and then use that posterior mean as a summary statistic for traditional ABC. This strategy was not effective on our data. Adapting their framework to our situation involves using a neural net to estimate the posterior mean from the raw sequence data, but due to the high-dimension and sparsely-distributed information in our sequence data, we could not get good enough estimates of the posterior mean for this approach to work in our model. Fearnhead and Prangle (2012) generate candidate summary statistics automatically via regression. Fisher et al. (2023) attempt to estimate posterior quantiles directly via neural network estimation and forward simulation. The method proposed here is similar to these previous works but uses the neural network itself to determine the relationship between user-generated summary statistics and the estimand.

There are several elements of real sequences which are recovered faithfully with our estimation framework. The rate at which spatial colocalization decays to 1, governed primarily by the lengthscale in our model, is estimated with high accuracy. More notably, the size of the exo stripping region estimated here is similar to the range discussed in previous experiments (Unniraman and Schatz (2007)), though others have proposed significantly larger windows (Kadyrov et al. (2006)). Crucially, despite using a noninformative prior, our procedure finds the size of the exo stripping region to be about 35 bases. This aligns with previous work showing that mutations can occur under SHM in the middle of long stretches of A/T bases (Unniraman and Schatz (2007)). We reproduce that feature in this work. However, it is worth noting that the exo parameters were estimated the least accurately of all our estimands. This is likely due to the large number of possible explanations for a given mutation pattern as well as the sparse distribution of information about the MMR pathway throughout our simulated sequences.

Some features estimated in the model differ from or were not well characterized in previous in vitro or animal studies. Both in vitro (Pham et al. (2019)) and in vivo (Pilzecker and Jacobs (2019)) studies have shown that AID deaminates C's on both DNA strands. The in vitro data study suggests about equal AID targeting to each strand, while the in vivo data suggests about 1.5× more deaminations on the coding strand. We estimate a forward strand bias of about 1.22×, meaning that we do not see as dramatic of a strand bias for AID targeting in our estimation framework. It is possible this is due to differences between human and mouse sequence data, since knockout experiments which isolate individual pathways are not possible for human data.

Our estimated lengthscale parameter, which controls how likely lesions are to occur near each other, corresponds to a correlation of 0.9 for neighboring pairs of positions, 0.7 for pairs of positions with one base in between them, decreasing to 0.1 for for pairs of positions with four bases in between them. This fairly steep drop-off in the correlation corresponds to less processivity of AID than is observed in vitro (e.g., Senavirathne et al. (2015), show that AID can scan along a 70-nucleotide ssDNA construct). However, it is a potential explanation for the observation that AID action is much reduced in vivo compared with in vitro (Shen et al. (2006), Pilzecker and Jacobs (2019)).

Finally, our estimated BER probability of about 0.65 suggests that there is a slight bias toward the BER machinery being recruited over the MMR machinery (or, more precisely, a slight bias toward machinery that repairs only the lesion over machinery that can introduce mutations in a patch around the lesion). To our knowledge, the relative rates of these pathways have not been characterized in the literature.

There are several shortcomings of the model proposed here. While the colocalization falls within the central 90% of values at every distance, the distance 2 colocalization is much higher in real sequences than our model can account for. This difference can possibly be attributed to tandem mutations caused by Polζ (Saribasak et al. (2012)) or by overlapping hotspots (Wei et al. (2015)). Our cross colocalization results suggest overlapping hotspots are a more likely explanation. We also found that a simpler 5-mer model had lower log loss than our mechanistic model on real data, presumably due to mutation hotspots not being modeled in our framework. Attempts to incorporate mutability into the prelesion probabilities made little difference, likely due to the relatively high marginal probability that a prelesion is thinned. Further attempts should be made to incorporate local sequence context into the pathways downstream of AID targeting as well as into lesion probability. Doing so will allow for hotspots which are as highly mutable relative to the base mutation frequency as what we see in real sequence data.

Although our model was able to recover the total size of the exo stripping region, it struggled to identify the mean size of the region to the left and right of the region (μ_l and μ_r). The issue seems to be that there is no observable quantity that is very informative about these parameters (see Figure 7, where all the parameters aside from Exo left and Exo right have a dominant summary statistic). If we imagine what sort of statistic we would need to get good estimates of these quantities, we would want something that measures the signed distance between a mutation and the lesion that produced it. The problem is that this is not observable (we don't know where the lesions were, and even if we did we wouldn't know the relationship between a mutation and the lesion that was responsible for it), and in any individual sequence there is not enough information to even get a good estimate of where the lesions were. These quantities may only be estimable in controlled systems, for instance, transgenic mice containing a region where there is only one potential target for AID (Unniraman and Schatz (2007)).

Many recent theories and uncertainties about the mechanisms governing SHM are not incorporated into our model. Recent work has suggested that AID can return to an exo-stripped patch, leading to additional SHM (Green et al. (2011)). Other work has suggested that the exo-stripped patch can be longer than any sequence we engage with here, potentially several thousand base-pairs long (Kadyrov et al. (2006)). Overall, one must carefully balance the seemingly limitless complexity of somatic hypermutation mechanisms with the difficulty of high-dimensional Bayesian inference. While a model that incorporates every minor element of the pathways governing SHM might seem ideal, the sparse information in the mutated sequences as well as the exponentially growing number of explanations for the mechanistic history of a sequence renders such a model intractable. In fact, while estimation in the simple proposed model is accurate, it is not clear that a model which incorporates several additional pathways as well as several downstream context-sensitive transition probabilities would even be identifiable. The complexity of future mechanistic models should be gradually increased to ensure accurate recovery of the relevant pathways.

In order to guide future work, we will now review other approaches that we tried. As other analysts have had success in training neural networks on raw sequence data (Flagel, Brandvain and Schrider (2019)), we initially attempted to predict parameters directly from the hot-encoded sequences, but it was not computationally feasible. Information about the mechanistic parameters is still sparsely distributed throughout the entire set of sequences, and it is also not clear how to effectively combine information from thousands of hot-encoded sequences into a single parameter estimate. Our initial attempt was to use a convolutional neural network to output a low-dimensional vector for each sequence. Following this, a recurrent neural network would combine these outputs into a single parameter estimate. Not only was this computationally expensive, but performance was poor.

Since it is much easier to find the parameters that maximize the complete-data likelihood than to find parameters that maximize the observed-data likelihood, our next estimation attempt was to use an importance sampling EM procedure. We attempted to sample from an approximation of the conditional distribution of the latent variables, given the observations, and find the values of the parameters that maximized the complete data log likelihood, using importance weights to correct for the fact that we were sampling from an approximation of the conditional distribution of the latent variables, given the data, and not the exact conditional distribution. This failed because our approximation did not match the truth closely enough, leading to high variance in the weights and often a single latent variable draw per observation having a weight that dominated all the other draws for that observation.

Finally, we attempted to incorporate an attention-based embedding of the one-hot-encoded sequence as a learned summary statistic. This embedding was then used as a multidimensional summary statistic alongside the other summary statistics. We were motivated to try this because attention-based models have exhibited strong performance in other settings (Vaswani et al. (2017)). However, likely due to computational issues relating to properly training the model to convergence, this method did not improve performance over the method proposed in this paper. It is possible that, given more training time or parameter tuning, this approach would show better performance. More information about how we incorporated attention in this attempt can be found in Appendix A (Fisher et al. (2025)).

Our original hypothesis in this work was that by leveraging mechanistic knowledge, we would be able to fit more accurate models. This strategy has worked in other settings, including our recent work on V(D)J recombination (Russell et al. (2023)). However, in this context, this has not resulted in more accurate mutability predictions. For somatic hypermutation it seems possible that the biochemical process is simply too complex to gain statistical efficiency using this approach. The path forward in this setting may be for more sophisticated mechanism-free models as in (Spisak, Walczak and Mora (2020), Tang, Krantsevich and MacCarthy (2022)).

On the other hand, this modeling exercise has revealed some features of the data that support or sharpen current theories about the mechanisms governing SHM. Some parameters governing a forward model of mechanistic pathways can be recovered faithfully. We show that known pathways which induce mutations are sufficient to produce the level of spatial correlation we see in mutation sites in vivo. Although we were not able to provide more accurate predictions than mechanism-free models, we still strongly believe in the merit of mathematically formalizing biochemical models and engaging with parameter estimation from real data. As experiments isolating different pathways are not available for humans, attempting to infer parameters jointly from sequence data is the only way to get parameter estimates for biochemical models; even when isolating those pathways, getting parameter estimates is difficult.

Future work into refinement of the proposed forward model should be considered. While this work showed that our neural network estimation procedure can estimate the parameters governing forward simulation of the two major pathways which drive SHM, there are additional layers of complexity which could be added. However, with the current formulation, such additions would also necessitate additional hand-crafted, informative summary statistics to estimate parameters associated with the new pathways. It is not guaranteed that joint accuracy in estimation, or even identifiability, would be maintained were the model made significantly more complex. Our results suggest that the high dimension and sparsely distributed information contained in BCR sequence data may indicate limits on how much we can learn about biochemical mechanisms from sequencing experiments alone.

Acknowledgments. The authors would like to thank to David Schatz and Thomas MacCarthy for informative discussions, Mackenzie Johnson and Saoirse Courts for suggestions on the manuscript as well as Natanael Spisak, Thierry Mora, and Aleksandra Walczak for sharing preprocessed data from (Spisak, Walczak and Mora (2020)).

Frederick A. Matsen IV is also affiliated with the Howard Hughes Medical Institute, the Department of Genome Sciences, University of Washington, and the Department of Statistics, University of Washington.

Authors Frederick A. Matsen IV and Julia Fukuyama wish to be considered co-senior authors.

Funding. NIH Grants R01 AI146028 (PI Matsen). Scientific Computing Infrastructure at Fred Hutch funded by ORIP Grant S10OD028685. Frederick Matsen is an investigator of the Howard Hughes Medical Institute.

SUPPLEMENTARY MATERIAL

Supplement to "Inferring mechanistic parameters of somatic hypermutation using neural networks and approximate Bayesian computation" (DOI: 10.1214/24-AOAS1985SUPPA; .pdf). Appendix A: Transformer Modeling. Modeling extension for parameter estimation using transformer neural networks. Appendix B: Additional Sequence Information. Fixed elements of the forward simulation framework and their sources. Appendix C: Posterior Predictive Checks. Verification that our estimator reconstructs a fixed parameter set under which sequences are simulated.

Supplementary code (DOI: 10.1214/24-AOAS1985SUPPB; .zip). Code used to conduct these analyses, including forward simulation and estimations as well as figure generation.

REFERENCES

Bonsignori, M., Kreider, E. F., Fera, D., Meyerhoff, R. R., Bradley, T., Wiehe, K., Alam, S. M., Aussedat, B., Walkowicz, W. E. et al. (2017). Staged induction of HIV-1 glycan-dependent broadly neutralizing antibodies. *Sci. Transl. Med.* **9**. https://doi.org/10.1126/scitranslmed.aai7514

Briney, B., Inderbitzin, A., Joyce, C. and Burton, D. R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566** 393–397. https://doi.org/10.1038/s41586-019-0879-y

Briney, B. S., Willis, J. R. and Crowe, J. E. Jr (2012). Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun.* **13** 523–529. https://doi.org/10.1038/gene.2012.28

Chelico, L., Pham, P., Calabrese, P. and Goodman, M. F. (2006). APOBEC3G DNA deaminase acts processively $3' \to 5'$ on single-stranded DNA. *Nat. Struct. Mol. Biol.* **13** 392–399. https://doi.org/10.1038/nsmb1086

Cui, A., Di Niro, R., Vander Heiden, J. A., Briggs, A. W., Adams, K., Gilbert, T., O'Connor, K. C., Vigneault, F., Shlomchik, M. J. et al. (2016). A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J. Immunol.* **197** 3566–3574. https://doi.org/10.4049/jimmunol.1502263

Dhar, A., Davidsen, K., Matsen, F. A. 4th and Minin, V. N. (2018). Predicting B cell receptor substitution profiles using public repertoire data. *PLoS Comput. Biol.* **14** e1006388. https://doi.org/10.1371/journal.pcbi.1006388

Dunn-Walters, D. K., Dogan, A., Boursier, L., MacDonald, C. M. and Spencer, J. (1998). Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *J. Immunol.* **160** 2360–2364.

Dunn-Walters, D. K. and Spencer, J. (1998). Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology* **95** 339–345.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 419–474. MR2925370 https://doi.org/10.1111/j.1467-9868.2011.01010.x

Feng, J., Shaw, D. A., Minin, V. N., Simon, N. and Matsen, F. A. IV (2019). Survival analysis of DNA mutation motifs with penalized proportional hazards. *Ann. Appl. Stat.* **13** 1268–1294. MR3963571 https://doi.org/10.1214/18-AOAS1233

Fisher, T., Luedtke, A., Carone, M. and Simon, N. (2023). Deep learning for marginal Bayesian posterior inference with recurrent neural networks. https://doi.org/10.5705/ss.202020.0348

Fisher, T., Sung, K., Simon, N., Fukuyama, J. and Matsen IV, F. A. (2025). Supplement to "Inferring mechanistic parameters of somatic hypermutation using neural networks and approximate Bayesian computation." https://doi.org/10.1214/24-AOAS1985SUPPA, https://doi.org/10.1214/24-AOAS1985SUPPB

Flagel, L., Brandvain, Y. and Schrider, D. R. (2019). The unreasonable effectiveness of convolutional neural networks in population genetic inference. *Mol. Biol. Evol.* **36** 220–238. https://doi.org/10.1093/molbev/msy224

Green, B., Belcheva, A., Nepal, R. M., Boulianne, B. and Martin, A. (2011). The mismatch repair pathway functions normally at a non-AID target in germinal center B cells. *Blood* **118** 3013–3018. https://doi.org/10.1182/blood-2011-03-345991

Hershberg, U., Uduman, M., Shlomchik, M. J. and Kleinstein, S. H. (2008). Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int. Immunol.* **20** 683–694. https://doi.org/10.1093/intimm/dxn026

Hwang, J. K., Wang, C., Du, Z., Meyers, R. M., Kepler, T. B., Neuberg, D., Kwong, P. D., Mascola, J. R., Joyce, M. G. et al. (2017). Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc. Natl. Acad. Sci. USA* **114** 8614–8619. https://doi.org/10.1073/pnas.1709203114

Jansen, J. G., Langerak, P., Tsaalbi-Shtylik, A., van den Berk, P., Jacobs, H. and de Wind, N. (2006). Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. *J. Exp. Med.* **203** 319–323.

Jardine, J., Julien, J.-P., Menis, S., Ota, T., Kalyuzhniy, O., McGuire, A., Sok, D., Huang, P.-S., MacPherson, S. et al. (2013). Rational HIV immunogen design to target specific germline B cell receptors. *Science* **340** 711–716. https://doi.org/10.1126/science.1234150

Jiang, B., Wu, T.-Y., Zheng, C. and Wong, W. H. (2017). Learning summary statistic for approximate Bayesian computation via deep neural network. *Statist. Sinica* **27** 1595–1618. MR3701500

Kadyrov, F. A., Dzantiev, L., Constantin, N. and Modrich, P. (2006). Endonucleolytic function of MutLa in human mismatch repair. *Cell* **126** 297–308. https://doi.org/10.1016/j.cell.2006.05.039

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16** 111–120. https://doi.org/10.1007/BF01731581

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KLEINSTEIN, S. H., LOUZOUN, Y. and SHLOMCHIK, M. J. (2003). Estimating hypermutation rates from clonal tree data. *J. Immunol.* **171** 4639–4649.

KRANTSEVICH, A., TANG, C. and MACCARTHY, T. (2020). Correlations in somatic hypermutation between sites in IGHV genes can be explained by interactions between AID and/or Polη hotspots. *Front. Immunol.* **11** 618409. https://doi.org/10.3389/fimmu.2020.618409

KRIJGER, P. H. L., LANGERAK, P., VAN DEN BERK, P. C. M. and JACOBS, H. (2009). Dependence of nucleotide substitutions on Ung2, Msh2, and PCNA-Ub during somatic hypermutation. *J. Exp. Med.* **206** 2603–2611. https://doi.org/10.1084/jem.20091707

LANGERAK, P., NYGREN, A. O., KRIJGER, P. H., VAN DEN BERK, P. C. and JACOBS, H. (2007). A/T mutagenesis in hypermutated immunoglobulin genes strongly depends on PCNAK164 modification. *J. Exp. Med.* **204** 1989–1998.

LUPO, C., SPISAK, N., WALCZAK, A. M. and MORA, T. (2022). Learning the statistics and landscape of somatic mutation-induced insertions and deletions in antibodies. *PLoS Comput. Biol.* **18** e1010167. https://doi.org/10.1371/journal.pcbi.1010167

MAK, C. H., PHAM, P., AFIF, S. A. and GOODMAN, M. F. (2013). A mathematical model for scanning and catalysis on single-stranded DNA, illustrated with activation-induced deoxycytidine deaminase. *J. Biol. Chem.* **288** 29786–29795. https://doi.org/10.1074/jbc.M113.506550

MARJORAM, P., MOLITOR, J., PLAGNOL, V. and TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. USA* **100** 15324–15328.

MCCOY, C. O., BEDFORD, T., MININ, V. N., BRADLEY, P., ROBINS, H. and MATSEN IV, F. A. (2015). Quantifying evolutionary constraints on B-cell affinity maturation. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370** 20140244. https://doi.org/10.1098/rstb.2014.0244

METHOT, S. P. and DI NOIA, J. M. (2017). Chapter two—Molecular mechanisms of somatic hypermutation and class switch recombination. In *Advances in Immunology* (F. W. Alt, ed.) **133** 37–87. Academic Press, San Diego.

MØLLER, J., SYVERSVEEN, A. R. and WAAGEPETERSEN, R. P. (1998). Log Gaussian Cox processes. *Scand. J. Stat.* **25** 451–482. MR1650019 https://doi.org/10.1111/1467-9469.00115

NEUBERGER, M. S., DI NOIA, J. M., BEALE, R. C. L., WILLIAMS, G. T., YANG, Z. and RADA, C. (2005). Somatic hypermutation at A:T pairs: Polymerase error versus dUTP incorporation. *Nat. Rev., Immunol.* **5** 171–178.

NGUYEN, L.-T., SCHMIDT, H. A., VON HAESELER, A. and MINH, B. Q. (2014). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32** 268–274. https://doi.org/10.1093/molbev/msu300

PASZKE, A., GROSS, S., MASSA, F., LERER, A., BRADBURY, J., CHANAN, G., KILLEEN, T., LIN, Z., GIMELSHEIN, N. et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**.

PHAM, P., BRANSTEITTER, R., PETRUSKA, J. and GOODMAN, M. F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424** 103–107. https://doi.org/10.1038/nature01760

PHAM, P., MALIK, S., MAK, C., CALABRESE, P. C., ROEDER, R. G. and GOODMAN, M. F. (2019). AID-RNA polymerase II transcription-dependent deamination of IgV DNA. *Nucleic Acids Res.* **47** 10815–10829. https://doi.org/10.1093/nar/gkz821

PILZECKER, B. and JACOBS, H. (2019). Mutating for good: DNA damage responses during somatic hypermutation. *Front. Immunol.* **10** 438. https://doi.org/10.3389/fimmu.2019.00438

ROGOZIN, I. B. and DIAZ, M. (2004). Cutting edge: DGYW/WRCH is a better predictor of mutability at G:C bases in Ig hypermutation than the widely accepted RGYW/WRCY motif and probably reflects a two-step activation-induced cytidine deaminase-triggered process. *J. Immunol.* **172** 3382–3384.

ROGOZIN, I. B., PAVLOV, Y. I., BEBENEK, K., MATSUDA, T. and KUNKEL, T. A. (2001). Somatic mutation hotspots correlate with DNA polymerase eta error spectrum. *Nat. Immunol.* **2** 530–536. https://doi.org/10.1038/88732

RUSSELL, M. L., SIMON, N., BRADLEY, P. and MATSEN, F. A. IV (2023). Statistical inference reveals the role of length, GC content, and local sequence in V(D)J nucleotide trimming. *eLife* **12** e85145. https://doi.org/10.7554/eLife.85145

SARIBASAK, H., MAUL, R. W., CAO, Z., YANG, W. W., SCHENTEN, D., KRACKER, S. and GEARHART, P. J. (2012). DNA polymerase ζ generates tandem mutations in immunoglobulin variable regions. *J. Exp. Med.* **209** 1075–1081.

SENAVIRATHNE, G., BERTRAM, J. G., JASZCZUR, M., CHAURASIYA, K. R., PHAM, P., MAK, C. H., GOODMAN, M. F. and RUEDA, D. (2015). Activation-induced deoxycytidine deaminase (AID) co-transcriptional scanning at single-molecule resolution. *Nat. Commun.* **6** 10209.

SHEEHAN, S. and SONG, Y. S. (2016). Deep learning for population genetic inference. *PLoS Comput. Biol.* **12** e1004845. https://doi.org/10.1371/journal.pcbi.1004845

SHEN, H. M., TANAKA, A., BOZEK, G., NICOLAE, D. and STORB, U. (2006). Somatic hypermutation and class switch recombination in Msh6-/- Ung-/- double-knockout mice. *J. Immunol.* **177** 5386–5392.

SHENG, Z., SCHRAMM, C. A., CONNORS, M., MORRIS, L., MASCOLA, J. R., KWONG, P. D. and SHAPIRO, L. (2016). Effects of Darwinian selection and mutability on rate of broadly neutralizing antibody evolution during HIV-1 infection. *PLoS Comput. Biol.* **12** e1004940. https://doi.org/10.1371/journal.pcbi.1004940

SPENCER, J., DUNN, M. and DUNN-WALTERS, D. K. (1999). Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators. *J. Immunol.* **162** 6596–6601.

SPISAK, N., WALCZAK, A. M. and MORA, T. (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

SVOZIL, D., KVASNICKA, V. and POSPICHAL, J. (1997). Introduction to multi-layer feed-forward neural networks. *Chemom. Intell. Lab. Syst.* **39** 43–62.

TANG, C., KRANTSEVICH, A. and MACCARTHY, T. (2022). Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *iScience* **25** 103668. https://doi.org/10.1016/j.isci.2021.103668

UDUMAN, M., YAARI, G., HERSHBERG, U., STERN, J. A., SHLOMCHIK, M. J. and KLEINSTEIN, S. H. (2011). Detecting selection in immunoglobulin sequences. *Nucleic Acids Res.* **39** W499–W504. https://doi.org/10.1093/nar/gkr413

UNNIRAMAN, S. and SCHATZ, D. G. (2007). Strand-biased spreading of mutations during somatic hypermutation. *Science* **317** 1227–1230. https://doi.org/10.1126/science.1145065

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L. and POLOSUKHIN, I. (2017). Attention is all you need.

VIEIRA, M. C., ZINDER, D. and COBEY, S. (2018). Selection and neutral mutations drive pervasive mutability losses in long-lived anti-HIV B-cell lineages. *Mol. Biol. Evol.* **35** 1135–1146. https://doi.org/10.1093/molbev/msy024

WANG, M., RADA, C. and NEUBERGER, M. S. (2010). Altering the spectrum of immunoglobulin V gene somatic hypermutation by modifying the active site of AID. *J. Exp. Med.* **207** 141–153. https://doi.org/10.1084/jem.20092238

WANG, Y., ZHANG, S., YANG, X., HWANG, J. K., ZHAN, C., LIAN, C., WANG, C., GUI, T., WANG, B. et al. (2023). Mesoscale DNA feature in antibody-coding sequence facilitates somatic hypermutation. *Cell* **186** 2193–2207. https://doi.org/10.1016/j.cell.2023.03.030

WEI, L., CHAHWAN, R., WANG, S., WANG, X., PHAM, P. T., GOODMAN, M. F., BERGMAN, A., SCHARFF, M. D. and MACCARTHY, T. (2015). Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc. Natl. Acad. Sci. USA* **112** E728–E737. https://doi.org/10.1073/pnas.1500788112

WIEHE, K., BRADLEY, T., RYAN MEYERHOFF, R., HART, C., WILLIAMS, W. B., EASTERHOFF, D., FAISON, W. J., KEPLER, T. B., SAUNDERS, K. O. et al. (2018). Functional relevance of improbable antibody mutations for HIV broadly neutralizing antibody development. *Cell Host Microbe* **23** 759–765. https://doi.org/10.1016/j.chom.2018.04.018

WILSON, T. M., VAISMAN, A., MARTOMO, S. A., SULLIVAN, P., LAN, L., HANAOKA, F., YASUI, A., WOODGATE, R. and GEARHART, P. J. (2005). MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A:T mutations in antibody genes. *J. Exp. Med.* **201** 637–645.

YAARI, G., BENICHOU, J. I. C., HEIDEN, J. A. V., KLEINSTEIN, S. H. and LOUZOUN, Y. (2015). The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **370**. https://doi.org/10.1098/rstb.2014.0242

YAARI, G., UDUMAN, M. and KLEINSTEIN, S. H. (2012). Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40** e134. https://doi.org/10.1093/nar/gks457

YAARI, G., VANDER HEIDEN, J. A., UDUMAN, M., GADALA-MARIA, D., GUPTA, N., STERN, J. N. H., O'CONNOR, K. C., HAFLER, D. A., LASERSON, U. et al. (2013). Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* **4** 358. https://doi.org/10.3389/fimmu.2013.00358

YEAP, L.-S., HWANG, J. K., DU, Z., MEYERS, R. M., MENG, F.-L., JAKUBAUSKAITE, A., LIU, M., MANI, V., NEUBERG, D. et al. (2015). Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell* **163** 1124–1137. https://doi.org/10.1016/j.cell.2015.10.042

ZIVOJNOVIC, M., DELBOS, F., ZUBANI, G. G., JULÉ, A., ALCAIS, A., WEILL, J.-C., REYNAUD, C.-A. and STORCK, S. (2014). Somatic hypermutation at A/T-rich oligonucleotide substrates shows different strand polarities in Ung-deficient or -proficient backgrounds. *Mol. Cell. Biol.* **34** 2176–2187. https://doi.org/10.1128/MCB.01452-13