CONDITIONALLY SITE-INDEPENDENT NEURAL EVOLUTION OF ANTIBODY SEQUENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Common deep learning approaches for antibody engineering focus on modeling the marginal distribution of sequences. By treating sequences as independent samples, these methods overlook affinity maturation as a rich and untapped source of information about the evolutionary process through which antibodies explore the underlying fitness landscape. In contrast, classical phylogenetic models explicitly represent evolutionary dynamics but lack the expressivity to capture complex epistatic interactions. We bridge this gap with **CoSiNE**, a continuous-time Markov chain parameterized by a deep neural network. Mathematically, we prove that CoSiNE provides a first-order approximation to the intractable sequential point mutation process, capturing epistatic effects with an error bound that is quadratic in branch length. Empirically, CoSiNE outperforms state-of-the-art language models in zero-shot variant effect prediction. We also introduce *Guided Gillespie*, a classifier-guided sampling scheme that steers CoSiNE at inference time, enabling efficient optimization of antibody binding affinity toward specific antigens.

1 INTRODUCTION

Antibodies are key effectors of the adaptive immune response, enabling humans and other vertebrates to recognize and neutralize an enormous diversity of molecular targets (antigens). This diversity is made possible by an accelerated evolutionary process operating within individuals, known as *affinity maturation*. During an immune response, B cells evolve in germinal centers, where their antibody genes undergo rapid somatic hypermutation (SHM) and are subsequently subjected to selection that favors variants with improved antigen binding.

Evolutionary processes are classically modeled using continuous-time Markov chains, but the immense state space of protein sequences renders direct modeling intractable. To mitigate this computational bottleneck, classical models of sequence evolution assume that sites evolve independently according to a context-agnostic substitution process. As a consequence, these models have limited expressivity and often produce unrealistic evolutionary trajectories. This limitation has restricted their applicability in antibody sequence design and optimization, especially in comparison to modern antibody language models, which implicitly capture complex intra-sequence dependencies (Graves et al., 2020; Ruffolo et al., 2021; Leem et al., 2021; Olsen et al., 2022b; Hie et al., 2023; Kenlay et al., 2024b; Olsen et al., 2024b; Shanker et al., 2024; Wang et al., 2025). However, these language models lack the ability to model the time-dependent *process* over which antibodies mature, instead learning a stationary distribution over all antibody sequences.

In this work, we aim to combine the strengths of these two paradigms by introducing a **co**nditionally **s**ite-**i**ndependent **n**eural **e**volution model (**CoSiNE**) that learns to simulate antibody affinity maturation while capturing epistatic interactions within the sequence (Figure 1). CoSiNE uses a neural network to parameterize site-specific rate matrices conditioned on the full sequence context, enabling a factorized transition likelihood that still captures dependencies among sites. We fit CoSiNE to $\sim$100k clonal trees and employ a principled Gillespie sampling algorithm to simulate clonal expansion. We also formulate a selection score produced by CoSiNE that outperforms other deep learning approaches in predicting variant effects, highlighting its ability to model antibody fitness. Finally, we propose a classifier guidance procedure to steer the functional properties of antibodies generated by CoSiNE and show successful optimization in-silico.

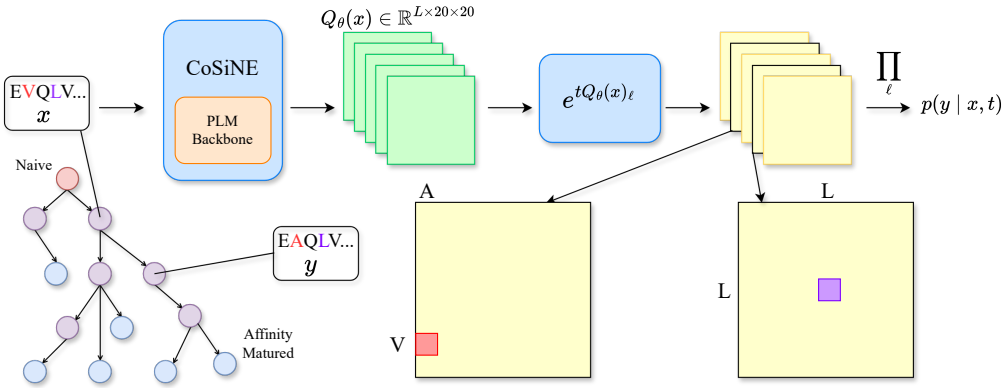


Figure 1: **Overview of CoSiNE**. Given an antibody sequence x, the framework calculates a full sequence transition probability $p(y \mid x, t)$.

2 Background

The evolutionary history of an antibody is represented by a clonal tree T_f. At the root of T_f is a naive antibody derived from genetic recombination. Over time, this naive sequence accumulates mutations, which are recorded in the tree as directed edges (x, y) with an associated branch length $t \in \mathbb{R}^+$. Informally, we call $\tau = (x, y, t)$ an evolutionary transition from the *parent* x to the *child* y over *time* t. We assume that t is calibrated to the expected number of mutations per-site between x and y.

Classical "Independent Sites" Models. Protein sequence evolution is most commonly modeled via continuous-time Markov chains (CTMCs). In general, given a discrete state space $\mathcal{S}$, a CTMC is completely defined by an initial distribution p_0 and a rate matrix $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$. Unfortunately, tractability is limited by the matrix exponential $\exp(t\mathbf{Q})$, which costs $O(|\mathcal{S}|^3)$ time. For protein sequences of length L with the typical 20 amino acid vocabulary $\mathcal{A}$, this step is problematic even for $L > 2$ since $|\mathcal{S}| = |\mathcal{A}|^L = 20^L$. To circumvent this issue, classical models of protein evolution like WAG Whelan & Goldman (2001) and LG Le & Gascuel (2008) make the independent-sites assumption to enable the factorization of the full sequence Markov chain into L independent chains, one for each site of the protein. The resulting transition probability $p(y \mid x, t) := \prod_{\ell=1}^{L} \exp(tQ_\ell)_{x_\ell, y_\ell}$ is thus tractable with a time complexity of $O(L|\mathcal{A}|^3)$. However, the independent-sites assumption leads to model misspecification because it fully ignores higher-order epistatic effects between sites.

Deep Learning Models Deep Amino acid Selection Model (DASM) predicts antibody fitness through per-residue selection scores with explicit disentanglement from mutational biases Matsen IV et al. (2025). However, DASM requires a manual clamping of selection scores to ensure that the resulting likelihoods represent a valid probability distribution. In contrast, CoSiNE derives a selection score through a log-likelihood ratio between a neural CTMC and a pre-trained SHM model Sung et al. (2025). This formulation ensures mathematical consistency, enabling CoSiNE to more accurately model the affinity maturation process. See Section A.3 for a detailed comparison of CoSiNE against DASM.

3 CoSiNE: Conditionally Site-Independent Neural Evolution Model

To address the limitations of classical models, we developed the CoSiNE model, which introduces two key modifications that mitigate model misspecification. First, we decouple rate estimation by learning site-specific rate matrices Q_ℓ for each site ℓ instead of a single unified matrix with scaling factors. Second, we model each evolutionary transition $\tau = (x, y, t)$ with its own set of rate matrices $\{Q_\ell^{(\tau)}\}_{\ell=1}^{L}$, which are inferred by conditioning on the parent sequence x, enabling the learning of

epistatic effects. Concretely, CoSiNE calculates transition probabilities as:

$$p_\theta(y \mid x, t) = \prod_{\ell=1}^{L} \exp\left(tQ_\theta(x)_\ell\right)_{x_\ell, y_\ell} \tag{1}$$

where $Q_\theta : \mathbb{R}^{|\mathcal{A}|^L} \to \mathbb{R}^{L \times |\mathcal{A}| \times |\mathcal{A}|}$ is parameterized by a neural network. In Section D.1, we provide theoretical grounding for CoSiNE by showing that CoSiNE constitutes a first-order approximation of a sequential point mutation process over the full sequence space. We provide an error upper bound that is quadratic with branch length, motivating CoSiNE's application to affinity maturation. We also propose a Gillespie procedure for CoSiNE that provably samples from $P(y \mid x, t)$ under certain conditions (Section D.2).

3.1 Inferring Antibody Fitness Landscapes from Affinity Maturation

Molecular evolution, including somatic evolution processes such as affinity maturation, can be viewed as a two-step process: First, mutations are introduced by an underlying mutational mechanism (SHM in the case of affinity maturation). Second, these mutations are filtered by selection according to the fitness advantages or disadvantages they confer. It is of great interest to explicitly learn the fitness of a given antibody sequence to, for example, enable the design of antibodies with desirable properties.

To estimate the fitness landscape from CoSiNE we must deconvolve the effects of SHM, which we model with Thrifty (Sung et al., 2025) and provide more details on how this is done in Section B.3.2. Following the mutation-selection framework introduced by Halpern & Bruno (1998), we decompose Q_{xy}, the observed transition rate from sequence x to sequence y, as:

$$Q_{xy} = k\, \mu_{xy}\, P_{\text{fix}}(x \to y), \tag{2}$$

where μ_{xy} is the transition rate under SHM, $P_{\text{fix}}(x \to y)$ is the probability of fixation, and k is an arbitrary scalar.

Using the small t approximation $p_\theta(y \mid x, t) = \exp(tQ_\theta)_{xy} \approx t(Q_\theta)_{xy}$ for $x \neq y$, we can manipulate Equation (2) to derive the following *selection score* that we use to evaluate CoSiNE on DMS assays:

$$\begin{aligned} \text{Score}(x \to y) &= \log p_\theta(y \mid x, t) - \log q(y \mid x, t) \\ &\approx \log\left(P_{\text{fix}}(x \to y)\right) + C, \end{aligned} \tag{3}$$

where $q(y \mid x, t)$ is the probability of sequence x mutating to y under Thrifty and $p_\theta(y \mid x, t)$ is the probability of the transition under CoSiNE. In Section D.3, we rely on standard population genetics theory (Kimura, 1962) to show that $P_{\text{fix}}(x \to y)$ is monotonic with respect to s_{xy}, the selective advantage of allele y over allele x.

3.2 Conditional Sequence Optimization via Guided Gillespie Sampling

Biologists often wish to design antibodies that strongly bind to a target of interest. Since the antigens associated with our training data are unknown, we cannot rely on the unconditional model alone. Instead, we propose a classifier guidance approach to sample from the posterior transition density $p(y \mid x, t, z)$, conditioned on the target antigen z. Following Nisonoff et al. (2025), this posterior is defined by scaling the unconditional rate matrix $\mathbf{Q}_{x,y}$ by the guidance term $[p(z \mid y)/p(z \mid x)]^\gamma$.

To make this guidance tractable, we model the predictor likelihood $p(z \mid y)$ as the probability that the binding affinity $r \sim \mathcal{N}(\mu_{\theta_z}(y), \sigma^2_{\theta_z}(y))$ between y and z exceeds an adaptive threshold $r_0 = \mu_{\theta_z}(x)$. To avoid the $O(L \cdot |\mathcal{A}|)$ cost of querying the predictor for every possible mutation y, we employ a first-order Taylor expansion to linearize the mean $\mu_{\theta_z}(y)$ around x and assume locally constant variance: $\sigma_{\theta_z}(y) \approx \sigma_{\theta_z}(x)$. This *Taylor-approximated guidance* (TAG) form allows us to estimate the guided rates for all neighbors in a single backward pass:

$$(\mathbf{Q}_z^{(\gamma)})_{x,y} \approx \left[2 \cdot \Phi\left(\frac{\nabla_x \mu_{\theta_z}(x)^\top (y - x)}{\sigma_{\theta_z}(x)}\right)\right]^\gamma \mathbf{Q}_{x,y}. \tag{4}$$

We apply TAG to Algorithm S1 to obtain *Guided Gillespie* sampling (Algorithm S2). Unlike guidance for flow matching or discrete diffusion, our formulation applies to the rate matrix of any sequential point mutation process. Crucially, this removes the need for boundary time conditions or training predictors on noisy sequences, enabling the direct use of predictors trained on raw experimental data.

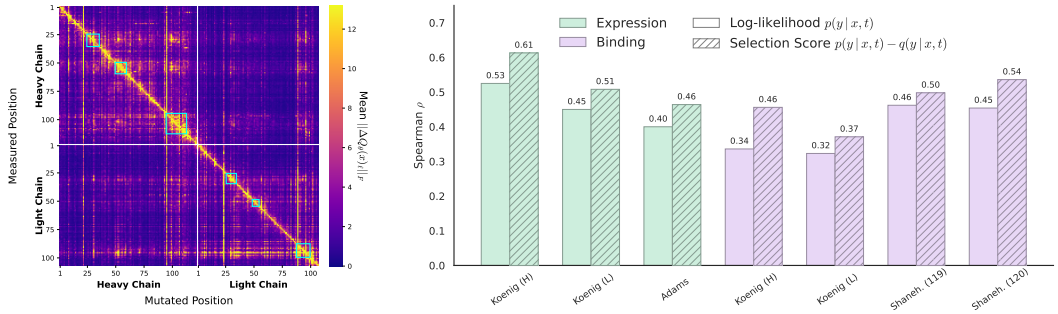


Figure 2: **CoSiNE sensitivity and selection score analysis.** (a) Categorical Jacobian for antibody 47D11. Cyan boxes denote CDRs. (b) Comparison of log-likelihood and selection score (Equation (3)) on DMS benchmarks.

4 Experiments

4.1 Fitting CoSiNE on a Clonal Tree Dataset

We fit CoSiNE on a dataset of $\sim$2 million evolutionary transitions constructed from 5 public sources (Jaffe et al., 2022; Tang et al., 2022; Vergani et al., 2017; Engelbrecht et al., 2025; Rodriguez et al., 2023), with train and test splits that match those of the DASM model (Matsen IV et al., 2025). We initialized CoSiNE from the 150M parameter ESM-2 checkpoint (Lin et al., 2023) and replaced the language modeling head with a custom output head. Additional details on model architecture and training are provided in Appendix A.

We found that CoSiNE fits this data remarkably well, achieving a test perplexity of 1.264 for heavy chain transitions in the Rodriguez et al. dataset (see Appendix A for our perplexity definition). In a head-to-head comparison with DASM+Thrifty, CoSiNE fits the test transitions with greater per-site likelihood in 62.3% of cases (Figure S6). Interestingly, this improvement is even more significant for long branch lengths ($t > 0.25$). To investigate epistasis, we compute the categorical Jacobian (Algorithm S3), inspired by Zhang et al. (2024), for a CoV-AbDab antibody (Raybould et al., 2020). Figure 2 shows that beyond expected local interactions (diagonal), we observe significant off-diagonal coupling within CDRs (cyan squares), consistent with structural coordination in the binding pocket. Notably, CoSiNE captures inter-chain epistasis: mutations in light chain CDR3 induce distinct sensitivity hotspots across all three heavy chain CDRs.

4.2 Zero-Shot Variant Effect Prediction

We evaluate CoSiNE's performance on DMS assays by measuring the Spearman correlation of our selection score (Equation (3)) with the experimental fitness values. To calculate the score, we set the parent x to be the wildtype sequence from which the mutants in the assay are derived. We fix $t = 0.2$ for all assays and VEP experiments.

Evaluation is performed on four DMS assays selected from the FLAb2 benchmark for their large sample sizes of sequences with the same length in an attempt to limit spurious correlations Chungyoun & Gray (2025). As baselines, we compare against AbLang-2, DASM, ProGen2 Small, ProGen2 Medium (best performing model for all FLAb2 binding datasets), ESM-2 150M (best performing model for all FLAb2 expression datasets), and ESM-2 650M (Olsen et al., 2024a; Matsen IV et al., 2025; Nijkamp et al., 2022; Lin et al., 2023). Further details on the DMS assays and evaluation with baseline models are provided in Section B.3.1.

The VEP results are shown in Table 1, where CoSiNE outperforms all other baselines on all datasets except one. Interestingly, CoSiNE substantially outperforms all other models on the Adams dataset, where the wildtype sequence is a mouse antibody. Although CoSiNE has never been trained on mouse antibodies, we suspect that the pretrained ESM2 backbone helps with generalization. As shown in Section 4.1, our selection score leads to increased correlation with fitness across all DMS datasets, implying it successful disentangles mutation and selection.

Table 1: **Spearman correlation on VEP benchmarks.** Koenig assays are split by heavy (H) and light (L) chain mutations; Shanehsazzadeh (Shaneh.) are split by heavy chain length (119 vs 120).

Model	Expression			Binding			
	Koenig (H)	Koenig (L)	Adams	Koenig (H)	Koenig (L)	Shaneh. (119)	Shaneh. (120)
AbLang-2	0.096	−0.127	−0.097	−0.090	−0.011	0.253	0.209
ESM2-150M	0.413	0.485	−0.112	0.112	0.266	0.236	0.205
ESM2-650M	0.326	0.429	0.124	0.063	0.265	0.227	0.360
ProGen2-Small	0.407	**0.513**	−0.024	0.098	0.332	0.119	0.070
ProGen2-Medium	0.392	0.408	0.231	0.085	0.235	0.299	0.319
DASM	<u>0.596</u>	0.474	<u>0.270</u>	<u>0.415</u>	<u>0.327</u>	<u>0.450</u>	**0.536**
CoSiNE (ours)	**0.613**	<u>0.508</u>	**0.464**	**0.456**	**0.371**	**0.498**	**0.536**

4.3 Guided Affinity Maturation from Naive Antibodies

We demonstrate the potential of CoSiNE to simulate affinity maturation towards high-affinity binders, starting from naive antibody sequences. We used predictors from (Jin et al., 2021) trained on the CoV-AbDab database, which employ an RNN encoder to transform heavy chain sequences into neutralization scores μ_{θ_z} against the SARS-CoV-1 receptor binding domains. We adopted MC dropout Gal & Ghahramani (2016) to estimate σ_{θ_z}. From here, we randomly picked naive sequences from the OAS database (Olsen et al., 2022a) and recursively sampled down the tree in Figure S1 using *Guided Gillespie*.

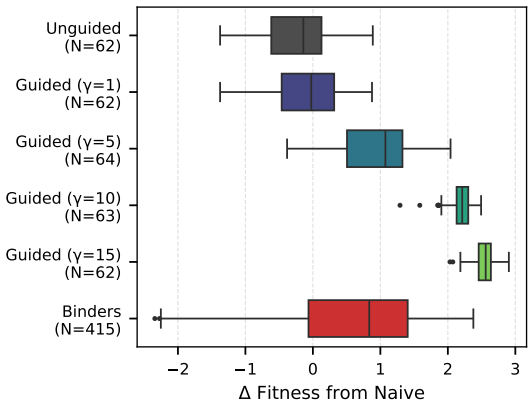


Figure 3: ***Guided Gillespie* consistently steers the predicted binding affinity of the sampled leaf sequences.** We plot the change in predicted binding affinity from the naive root sequence. Known binders from CoV-AbDab are in red.

We compare the affinity gain of our generated leaf sequences against 415 SARS-CoV-1 binders from the Cov-AbDab database. As shown in Figure 3, introducing guidance consistently shifts the unguided distribution towards higher affinity. While $\gamma \geq 10$ produced scores exceeding biological binders (likely exploiting oracle uncertainty), we noticed that $\gamma = 5$ generates affinity profiles that overlap with the real binders. We therefore adopted $\gamma = 5$ for further evaluation. Figure S9 demonstrates that these guided samples maintain structural plausibility (AbodyBuilder3 pLDDT, Kenlay et al. (2024a)) and humanness (OASis, Prihoda et al. (2021)) comparable to both unguided and natural sequences. These results hold across additional antigens and naive seed sequences (Section C.9), underscoring the generalizability of our approach.

5 Conclusion and Limitations

We presented CoSiNE, a framework that bridges deep protein language models with phylogenetic substitution models to capture the temporal dynamics of antibody affinity maturation. By introducing a first-order approximation for sequential point mutation CTMCs and a selection score that disentangles mutability and fitness, we achieved state-of-the-art results in variant effect prediction. Furthermore, to the best of our knowledge, we established a novel connection between discrete diffusion and classical evolution models, enabling guided sampling that integrates epistatic interactions with time-dependent evolution for protein design. However, certain limitations remain. Our reliance on a first-order approximation introduces model misspecification, and the current framework does not account for insertions or deletions. While these factors are less critical for rapid antibody evolution, they pose challenges for generalizing to broader protein classes. Despite this, CoSiNE serves as a robust foundation for future generative models of protein evolution.

IMPACT STATEMENT

This work aims to advance the field of machine learning by developing principled models for learning and simulating biological sequence evolution. By bridging modern deep learning with classical evolutionary modeling, our approach contributes new tools for understanding antibody affinity maturation and, more broadly, molecular evolution. Potential positive impacts include improved computational methods for studying immune responses, informing antibody engineering, and supporting downstream applications in biomedical research such as vaccine design and therapeutic antibody development. These advances could ultimately contribute to better diagnostics and treatments. We view this work as a methodological contribution to machine learning and computational biology, with societal implications that are consistent with and comparable to prior advances in biological sequence modeling.

REFERENCES

Rhys M Adams, Thierry Mora, Aleksandra M Walczak, and Justin B Kinney. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife*, 5:e23156, 2016.

Michael Chungyoun and Jeffrey J Gray. Fitness landscape for antibodies 2: Benchmarking reveals that protein ai models cannot yet consistently predict developability properties. *bioRxiv*, pp. 2025–12, 2025.

François Ehrenmann, Quentin Kaas, and Marie-Paule Lefranc. Imgt/3dstructure-db and imgt/domaingapalign: a database and a tool for immunoglobulins or antibodies, t cell receptors, mhc, igsf and mhcsf. *Nucleic Acids Research*, 38(suppl_1):D301–D307, 2010. doi: 10.1093/nar/gkp946.

Eric Engelbrecht, Oscar L Rodriguez, William Lees, Zach Vanwinkle, Kaitlyn Shields, Steven Schultze, William S Gibson, David R Smith, Uddalok Jana, Swati Saha, et al. Germline polymorphisms in the immunoglobulin kappa and lambda loci underpinning antibody light chain repertoire variability. *Nature Communications*, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. URL https://arxiv.org/abs/1506.02142.

Cade Gordon, Aniruddh Raghu, Peyton Greenside, and Hunter Elliott. Generative humanization for therapeutic antibodies, 2024. URL https://arxiv.org/abs/2412.04737.

Jordan Graves, Jacob Byerly, Eduardo Priego, Naren Makkapati, S. Vince Parish, Brenda P. Medellin, and Monica Berrondo. A review of deep learning methods for antibodies. *Antibodies*, 9, 2020. URL https://api.semanticscholar.org/CorpusID:218469594.

Aaron L Halpern and William J Bruno. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Molecular biology and evolution*, 15(7):910–917, 1998.

Brian L. Hie, Varun R. Shanker, Duo Xu, Theodora U. J. Bruun, Payton A.-B. Weidenbacher, Shaogeng Tang, Wesley Wu, John E. Pak, and Peter S. Kim. Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology*, 42:275 – 283, 2023. URL https://api.semanticscholar.org/CorpusID:263364541.

David B Jaffe, Payam Shahi, Bruce A Adams, Ashley M Chrisman, Peter M Finnegan, Nandhini Raman, Ariel E Royall, FuNien Tsai, Thomas Vollbrecht, Daniel S Reyes, et al. Functional antibodies exhibit light chain coherence. *Nature*, 611(7935):352–357, 2022.

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

Henry Kenlay, Frédéric A. Dreyer, Daniel Cutting, Daniel A. Nissley, and Charlotte M. Deane. Abodybuilder3: improved and scalable antibody structure predictions. *Bioinformatics*, 40, 2024a. URL https://api.semanticscholar.org/CorpusID:270199272.

Henry Kenlay, Frédéric A. Dreyer, Aleksandr Kovaltsuk, Dom Miketa, Douglas E. V. Pires, and Charlotte M. Deane. Large scale paired antibody language models. *PLOS Computational Biology*, 20, 2024b. URL https://api.semanticscholar.org/CorpusID:268691578.

Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713, 1962.

Patrick Koenig, Chingwei V Lee, Benjamin T Walters, Vasantharajan Janakiraman, Jeremy Stinson, Thomas W Patapoff, and Germaine Fuh. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences*, 114(4):E486–E495, 2017.

Si Quang Le and Olivier Gascuel. An improved general amino acid replacement matrix. *Molecular biology and evolution*, 25 7:1307–20, 2008. URL https://api.semanticscholar.org/CorpusID:14748147.

Jinwoo Leem, Laura S. Mitchell, James H.R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3, 2021. URL https://api.semanticscholar.org/CorpusID:244060395.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Frederick A Matsen IV, Will Dumm, Kevin Sung, Mackenzie M Johnson, David Rich, Tyler Starr, Yun S Song, Julia Fukuyama, and Hugh K Haddox. Separating selection from mutation in antibody language models. *bioRxiv*, pp. 2025–10, 2025.

Kristin S Midelfort, Hugo H Hernandez, Stephanie M Lippow, Bruce Tidor, Catherine L Drennan, and K Dane Wittrup. Substantial energetic improvement with minimal structural perturbation in a high affinity mutant antibody. *Journal of Molecular Biology*, 343(3):685–701, 2004. doi: 10.1016/j.jmb.2004.08.019.

Yasukazu Nakamura, Takashi Gojobori, and Toshimichi Ikemura. Codon usage tabulated from international dna sequence databases: status for the year 2000. *Nucleic Acids Research*, 28(1):292, 2000. doi: 10.1093/nar/28.1.292. URL https://www.kazusa.or.jp/codon/. Codon frequencies from Homo sapiens table: https://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=9606&aa=1.

Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, and Bui Quang Minh. Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32:268 – 274, 2014. URL https://api.semanticscholar.org/CorpusID:16191489.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models, 2022. URL https://arxiv.org/abs/2206.13517.

Hunter Nisonoff, Junhao Xiong, Stephan Allenspach, and Jennifer Listgarten. Unlocking guidance for discrete state-space diffusion and flow models, 2025. URL https://arxiv.org/abs/2406.01572.

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

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

Tobias Hegelund Olsen, Iain H. Moal, and Charlotte M. Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2, 2022b. URL https://api.semanticscholar.org/CorpusID:246226399.

Tobias Hegelund Olsen, Iain H. Moal, and Charlotte M. Deane. Addressing the antibody germline bias and its effect on language models for improved antibody design. *Bioinformatics*, 40, 2024b. URL https://api.semanticscholar.org/CorpusID:267575279.

David Prihoda, Jad Maamary, Andrew B. Waight, Verónica Juan, Laurence Fayadat-Dilman, Daniel Svozil, and Danny A. Bitton. Biophi: A platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *mAbs*, 14, 2021. URL https://api.semanticscholar.org/CorpusID:236971421.

Duncan K. Ralph and Frederick Albert Matsen IV. Inference of b cell clonal families using heavy/light chain pairing information. *PLOS Computational Biology*, 18, 2022. URL https://api.semanticscholar.org/CorpusID:247596752.

Matthew I. J. Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M. Deane. Cov-abdab: the coronavirus antibody database. *Bioinformatics*, 2020. URL https://api.semanticscholar.org/CorpusID:218764844.

Oscar L Rodriguez, Yana Safonova, Catherine A Silver, Kaitlyn Shields, William S Gibson, Justin T Kos, David Tieri, Hanzhong Ke, Katherine JL Jackson, Scott D Boyd, et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nature communications*, 14(1):4419, 2023.

Jeffrey A. Ruffolo, Jeffrey J. Gray, and Jeremias Sulam. Deciphering antibody affinity maturation with language models and weakly supervised learning. *ArXiv*, abs/2112.07782, 2021. URL https://api.semanticscholar.org/CorpusID:245144689.

Amir Shanehsazzadeh, Sharrol Bachas, Matt McPartlon, George Kasun, John M Sutton, Andrea K Steiger, Richard Shuai, Christa Kohnert, Goran Rakocevic, Jahir M Gutierrez, et al. Unlocking de novo antibody design with generative artificial intelligence. *BioRxiv*, pp. 2023–01, 2023.

Varun R. Shanker, Theodora U. J. Bruun, Brian L. Hie, and Peter S. Kim. Unsupervised evolution of protein and antibody complexes with a structure-informed language model. *Science*, 385:46 – 53, 2024. URL https://api.semanticscholar.org/CorpusID:270962763.

Kevin Sung, Mackenzie M Johnson, Will Dumm, Noah Simon, Hugh Haddox, Julia Fukuyama, and Frederick A Matsen IV. Thrifty wide-context models of b cell receptor somatic hypermutation. *Elife*, 14:RP105471, 2025.

Catherine Tang, Artem Krantsevich, and Thomas MacCarthy. Deep learning model of somatic hypermutation reveals importance of sequence context beyond hotspot targeting. *Iscience*, 25(1), 2022.

Stefano Vergani, Ilya Korsunsky, Andrea Nicola Mazzarello, Gerardo Ferrer, Nicholas Chiorazzi, and Davide Bagnara. Novel method for high-throughput full-length ighv-dj sequencing of the immune repertoire from bulk b-cells with single-cell resolution. *Frontiers in immunology*, 8:1157, 2017.

Meng Wang, Jonathan Patsenker, Henry Li, Yuval Kluger, and Steven H. Kleinstein. Supervised fine-tuning of pre-trained antibody language models improves antigen specificity prediction. *PLOS Computational Biology*, 21, 2025. URL https://api.semanticscholar.org/CorpusID:277464195.

Simon Whelan and Nick Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18 5:691–9, 2001. URL https://api.semanticscholar.org/CorpusID:44418374.

Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proceedings of the National Academy of Sciences*, 121(45):e2406285121, 2024.

A ADDITIONAL DETAILS OF COSINE

A.1 DATA COLLECTION AND MODEL TRAINING

To train the COSINE model, we compiled B-cell receptor (BCR) sequencing datasets from five sources (Jaffe et al., 2022; Tang et al., 2022; Vergani et al., 2017; Engelbrecht et al., 2025; Rodriguez et al., 2023). We adopted the data processing and phylogenetic inference protocol described by Matsen IV et al. (2025). Briefly, sequences were clustered into clonal families and naive germlines were inferred using `partis` (Ralph & Matsen IV, 2022). We retained families with at least two productive sequences, defining productivity by the absence of stop codons and the presence of canonical cysteine and tryptophan codons flanking the CDR3 in the same reading frame as the V segment start. Consistent with recent large language model training pipelines (e.g., AbLang-2), we further excluded sequences with mutated conserved signature cysteines. Insertions and deletions were reversed to align all sequences to their naive ancestor without gaps.

For phylogenetic reconstruction, we performed tree inference and ancestral sequence reconstruction (ASR) using IQ-TREE (Nguyen et al., 2014) under the K80 substitution model, using the naive sequence as an outgroup. We accounted for mutation rate heterogeneity across sites via a 4-category FreeRate model. For paired heavy and light chain data, we specifically employed the edge-linked-proportional partition model to allow the chains to evolve at distinct overall rates. This pipeline yielded a final training set of parent-child pairs (PCPs) extracted from the edges of the resulting phylogenetic trees, comprising approximately $\sim$2 million transitions from $\sim$120,000 clonal families collected from 555 individual donors. We defer further details to the Matsen IV et al. paper.

COSINEwas initialized from the 150M parameter ESM-2 checkpoint (Lin et al., 2023) and the the language modeling head was replaced with a randomly initialized output head that uses the `softplus` activation function to estimate the off-diagonal rates of $Q_\theta(x)_\ell$. To satisfy the properties of a valid rate matrix, we set the diagonal entries of $Q_\theta(x)_\ell$ to the negative sum of their respective rows. We also inserted a chain-break token between the heavy and light sequences of paired antibodies, enabling simultaneous reasoning over both chains.

Models were trained with mixed precision (BF16) for a maximum of 1 million steps, using a batch size of 16 with gradient accumulation over 3 steps. We employed the AdamW optimizer with a learning rate of 2.5×10^{-4}, utilizing 5,000 warmup steps followed by a polynomial decay schedule with a power of 2.0. To prevent overfitting, we applied a weight decay of 0.01 specifically to parameters with two or more dimensions (e.g., weights and embeddings), while excluding biases and layer normalization parameters. Gradients were clipped at a norm of 1.0, and training employed early stopping with a patience of 50 intervals based on validation loss. In practice, the loss converged after about 1 day on a single A100 GPU.

A.2 GILLESPIE AND *Guided Gillespie* SAMPLING

We provide the pseudocode for both unconditional Gillespie sampling and guided Gillespie sampling using COSINE. Differences for the guided version are highlighted in red. Notice that only a single evaluation of the predictor is required per step.

A.3 COMPARISON OF COSINE AGAINST DASM+THRIFTY

During training, DASM (Matsen IV et al., 2025) explicitly factorizes the affinity maturation process into the product of a somatic hypermutation (SHM) process $q(y \mid x, t)$, with a selection function $F(y)$. Inferring both q and F simultaneously from data is not possible since there is now an extra degree of freedom. Instead, DASM utilizes a frozen Thrifty SHM model q_ϕ, trained on out-of-frame transitions, and proposes to learn $F_\theta(y)$ via MLE of the observed transitions.

$$p(y \mid x) = \prod_{\ell=1}^{L} p(y_\ell \mid x, t) = \prod_{\ell=1}^{L} q(y_\ell \mid x, t) F_\theta(y_\ell \mid x) \quad \text{s.t.}$$
$$p(x_\ell \mid x, t) = 1 - \sum_{a \in \mathcal{A}} q(a \mid x, t) F_\theta(a \mid x)$$

Algorithm S1 Gillespie Sampling

input Model Q_θ, start sequence x, branch length t
output Samples $y \sim P(\cdot \mid x, t)$

$t' \leftarrow 0$
while $t' < t$ **do**
  $\lambda_x \leftarrow -\sum_{\ell=1}^{L} Q_\theta(x)_{x_\ell, x_\ell}$
  $\tau \sim \text{Exp}(\lambda_x)$
  if $t' + \tau > t$ **then**
    Return $y \leftarrow x$
  end if
  Sample (ℓ^*, a^*) with
    $P(\ell, a) = (Q_\theta(x)_\ell)_{x_\ell, a} / \lambda_x$
  $x_{\ell^*} \leftarrow a^*$
  $t' \leftarrow t' + \tau$
end while
Return $y \leftarrow x$

Algorithm S2 Guided Gillespie Sampling

input Model Q_θ, start sequence x, time t, predictor $\mu_\theta, \sigma_\theta$, scale γ
output Samples $y \sim P(\cdot \mid x, t, z)$

$t' \leftarrow 0$
while $t' < t$ **do**
  $g \leftarrow \nabla_x \mu_\theta(x)$
  $\tilde{Q}_{x,y} \leftarrow Q_{x,y} \cdot [2\Phi(g^\top (y - x)/\sigma_\theta(x))]^\gamma$
  $\lambda_x \leftarrow \sum_{y \neq x} \tilde{Q}_{x,y}$
  $\tau \sim \text{Exp}(\lambda_x)$
  if $t' + \tau > t$ **then**
    Return $y \leftarrow x$
  end if
  Sample (ℓ^*, a^*) with $P(\ell, a) = (\tilde{Q}_\theta(x)_\ell)_{x_\ell, a} / \lambda_x$
  $x_{\ell^*} \leftarrow a^*$
  $t' \leftarrow t' + \tau$
end while
Return $y \leftarrow x$

The second constraint essentially subsumes the normalizing constant into the probability of no-transition. The authors clamp the sum on the right-hand side of the constraint to be less than 1 in order to maintain a valid probability distribution.

Instead of using this formulation, which requires manual clamping of the selection scores, CoSiNE infers $F_\theta(y \mid x)$ at inference time using a log-likelihood ratio between our pre-trained affinity maturation and somatic hypermutation models (Equation (3)).

B EXPERIMENTAL DETAILS

B.1 SYNTHETIC EXPERIMENTS ON SINGLE CODONS

To empirically validate Proposition D.1 and Lemma D.2, we utilized a computationally tractable state space that allows for manipulation of the full rate matrix. Specifically, we modeled short DNA sequences with length $L = 3$ over the vocabulary $\mathcal{D} = \{A, G, C, T\}$, thus obtaining a state space that corresponds exactly to the 64 standard codons. To study processes with different levels of epistasis, we constructed ground truth rate matrices using linear interpolation between two extremes:

$$\mathbf{Q}_{\text{true}} = (1 - \varepsilon)\mathbf{Q}_{\text{factorized}} + \varepsilon\mathbf{Q}_{\text{state-dep}}$$

where $\varepsilon \in [0, 1]$ is the epistasis strength parameter. For $\mathbf{Q}_{\text{factorized}}$ (no epistasis), we sampled site-independent base rates: for each site $\ell \in \{1, 2, 3\}$, we drew a 4×4 rate matrix with off-diagonal entries from Uniform$(0, 2)$ and diagonal entries set to the negative row sum. The global 64×64 matrix $\mathbf{Q}_{\text{factorized}}$ assigns rate $r_{\ell,a,b}$ to transitions differing only at site ℓ with substitution $a \to b$. For $\mathbf{Q}_{\text{state-dep}}$ (maximum epistasis), each of the $64 \times 9 = 576$ transitions with a Hamming distance equal to 1 receives an independent rate drawn from Uniform$(0, 2)$. At $\varepsilon = 0$, rates are perfectly factorizable; at $\varepsilon = 1$, every transition is state-dependent.

For each $\mathbf{Q}_{\text{true}}$ that we drew in this way, we generated 2.5M training samples by drawing branch lengths $b \sim \text{Exp}(\lambda = 0.5)$, sampling start states x uniformly, and sampling end states from $P(\cdot \mid x, t) = \exp(t\mathbf{Q}_{\text{true}})_{x,\cdot}$. We compared three estimators: (1) **Full MLE**, which directly parameterizes all 576 transition rates and optimizes via gradient descent on the exact likelihood; (2) **Factorized**, a neural model that outputs context-dependent site-level 4×4 rate matrices and assumes site-independence during training; and (3) **Factorized SNR**, which uses the same architecture as (2) but applies SNR weighting (Section C.1). All models were trained for up to 1000 epochs with Adam (lr $= 0.01$ for factorized models, lr $= 0.1$ for Full MLE) and early stopping (patience=50). We tested $\varepsilon \in \{0, 0.25, 0.5, 0.75, 1.0\}$ with 3 replicates per level, measuring estimation error as the relative difference in Frobenius norm of the estimated and ground truth rate matrices: $\|\mathbf{Q}_{\text{est}} - \mathbf{Q}_{\text{true}}\|_F / \|\mathbf{Q}_{\text{true}}\|_F$.

Using the trained factorized models, we compare Gillespie sampling (Algorithm S1) against per-site matrix exponentiation (Equation (1)). To evaluate each sampling method without sampling noise, we compute exact transition probability distributions analytically. For Gillespie, we achieve this by first reconstructing the full 64×64 estimated rate matrix by querying the model at all 64 states, then computing the transition probability matrix. We compare both methods for the factorized and SNR-weighted models (4 curves total) by measuring KL divergence to the ground truth transition probability across 30 log-spaced branch lengths from $t = 0.01$ to $t = 10.0$, averaged uniformly over all 64 starting states.

B.2 CALCULATING THE CATEGORICAL JACOBIAN VIA PERTURBATION

To quantify the epistatic interactions learned by COSINE, we approximate the Jacobian by exhaustively computing all single point mutations. This procedure is described in Algorithm S3 with an example output sensitivity matrix depicted in Section 4.1.

B.3 VARIANT EFFECT PREDICTION DETAILS

B.3.1 DMS ASSAYS AND BASELINE MODELS

The results shown in Table 1 came from four DMS assays from the FLAb2 benchmark (Chungyoun & Gray, 2025). We provide more details on these datasets in Table S1. Following Matsen IV et al. (2025), we obtain the Koenig and Shanehsazzadeh (Shaneh.) datasets from commit 67738ee (April 17, 2024) of the FLAb Github repository. The Adams dataset is taken from commit 3453aeb (September 1, 2025). The Adams dataset contains multiple fitness measurements per mutant sequence, so we correlate antibody model predictions with the average fitness per mutant during VEP evaluation.

An additional step we must take when calculating the COSINE selection score for these assays is determining the underlying nucleotide sequence for the wildtype antibody so that we can calculate the transition likelihood under an SHM model (see Section B.3.2 for more details). For the Koenig

Algorithm S3 Categorical Jacobian Computation

Input: Antibody sequence x of length L, Vocabulary $\mathcal{A}$ ($|\mathcal{A}| = 20$)
Output: Sensitivity Matrix $\mathbf{S} \in \mathbb{R}^{L \times L}$
Initialize $\mathbf{S} \leftarrow \mathbf{0}_{L \times L}$
Compute wildtype rate matrices: $Q = Q_\theta(x)$
for $i = 1$ **to** L **do**
  for $a \in \mathcal{A}$ **such that** $a \neq \mathbf{x}_i$ **do**
    $x' \leftarrow x$
    $x'_i \leftarrow a$ {Mutate residue at position i to a}
    $Q' \leftarrow Q_\theta(x')$
    for $j = 1$ **to** L **do**
      $\mathbf{S}_{i,j} \leftarrow \mathbf{S}_{i,j} + \|\mathbf{Q}_j - \mathbf{Q}'_j\|_F$ {Calculate shift in output at position j}
    end for
  end for
  $\mathbf{S}_{i,:} \leftarrow \mathbf{S}_{i,:}/(|\mathcal{A}| - 1)$ {Average over all possible mutations}
end for
Return S

and Shaneh. datasets, we use `IMGT/DomainGapAlign` to map the wildtype amino acid sequence to the closest germline V and J genes (Ehrenmann et al., 2010). We then obtain the V- and J-segment nucleotide sequences from HG38 and backtranslate remaining mismatches and junction regions to the codon with the highest frequency in the human genome ((Nakamura et al., 2000)). The Adams dataset uses mouse antibody 4-4-20 scFv as its wildtype, and its nucleotide sequence was obtained from Addgene plasmid pCT302 (Midelfort et al., 2004).

Table S1: **Overview of Deep Mutational Scanning (DMS) Datasets.** *Avg. Subs* denotes the average number of amino acid substitutions per mutant relative to wildtype.

Original Source	Assay Type	Dataset Name	Mutated Chain	Chain Length		Num. Seqs	Avg. Subs
				Heavy	Light		
Koenig et al. (2017)	Expression	Koenig (H)	Heavy	120	108	2261	1
		Koenig (L)	Light	120	108	2014	1
	Binding (VEGF)	Koenig (H)	Heavy	120	108	2261	1
		Koenig (L)	Light	120	108	2014	1
Adams et al. (2016)	Expression	Adams	Heavy	117	112	2803	1.98
Shanehsazzadeh et al. (2023)	Binding (HER2)	Shaneh. (119)	Heavy	119	107	184	5.22
		Shaneh. (120)	Heavy	120	107	201	5.97

We evaluate zero-shot VEP with six other baseline models, which can be divided into three categories based on their architectures and how they are evaluated.

1. Masked language models (**AbLang-2, ESM-2 150M, ESM-2 650M**) are evaluated via pseudo-perplexity in accordance with the FLAb2 benchmark.
2. Autoregressive models (**ProGen2 Small, ProGen2 Medium**) are evaluated via perplexity, also in accordance with the FLAb2 benchmark.
3. The **DASM** model is evaluated by summing its log selection factors as described in Matsen IV et al. (2025).

B.3.2 Calculating Sequence Likelihoods with Thrifty SHM Model

To estimate sequence transition likelihoods under SHM, we use `ThriftyHumV0.2-59-hc-tangshm`, which provides per-codon transition probabilities with a multihit correction described in Matsen IV et al. (2025). Since the Thrifty model operates in the state space of codons (64 states) and CoSiNE operates in the state space of amino acids (20 states), we must sum over all possible codons that could code for the observed alternate allele when calculating sequence likelihoods.

More formally, let x_ℓ denote the amino acid identity at position ℓ in the wildtype sequence for a DMS assay and $c(x_\ell)$ denote its underlying nucleotide codon (see Section B.3.1 for details on how this is determined). Let y_ℓ denote the corresponding amino acid in the mutant sequence y and $C(y_\ell)$ denote the set of possible codons that could code for y_ℓ. We calculate the likelihood of sequence y under Thrifty as

$$q(y \mid x, t) = \prod_{\ell}^{L} q(y_\ell \mid x_\ell, t) \quad \text{where}$$

$$q(y_\ell \mid x_\ell, t) = \begin{cases} q(c(x_\ell) \mid c(x_\ell), t) & \text{if } y_\ell = x_\ell \\ \sum_{c \in C(y_\ell)} q(c \mid c(x_\ell), t) & \text{if } y_\ell \neq x_\ell. \end{cases}$$

B.4 Guided Affinity Maturation from Naive Antibodies

We randomly sampled naive antibodies from a subset of the OAS database Olsen et al. (2022a) containing only heavy chain sequences with the IgM isotype from human donors. To guide and score the sampled sequences, we utilized the SARS-CoV-1 and SARS-CoV-2 neutralization predictors from Jin et al., with weights downloaded from `https://github.com/wengong-jin/RefineGNN`. These oracles are only trained on heavy chain sequences. To parameterize the variance σ_{θ_z}, we used MC dropout Gal & Ghahramani (2016) with 10 fixed masks. For the numerator in Equation (4), we set the model to evaluation mode and computed the gradient of the output $\mu_{\theta_z}(x)$ with respect to the one-hot input sequence x. Known binders were curated from the CoV-AbDab database. We set the naive antibody at the root of the clonal tree in Figure S1 and recursively sampled down its nodes in breadth first order. We repeated this procedure 5 times for each guidance setting and collected the leaf sequences for comparison. The selected tree has 13 leaves, so a non-repetitive sampler would produce 65 unique leaf sequences.

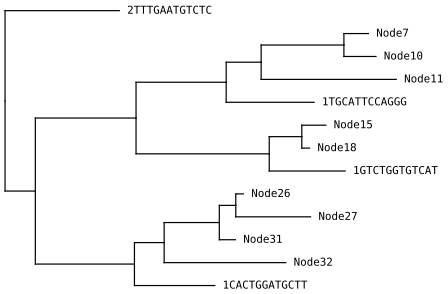


Figure S1: **Clonal Tree used for Guided Affinity Maturation Experiment in Section 4.3**

B.5 Guided Optimization of Antibody CDRs

All methods were tasked with optimizing a SARS-CoV-1 binder from the CovAbDab database. Optimization was restricted to the CDRs, identified via the IMGT numbering scheme Ehrenmann et al. (2010), with a strict budget of at most 5 mutations from the seed sequence. We generated 1,000 variants per method. We used the same SARS-CoV-1 neutralization predictor as Section B.4. To ensure a fair comparison, all methods (excluding the reference Greedy baseline) were constrained to use ≤ 5 oracle calls per generated sequence. We compared CoSiNE against the following baselines:

Genetic Algorithm (GA). We evolved a population of 1,000 sequences over 9 generations. In each generation, the top 50% of sequences were selected based on fitness, and the remaining 50% were regenerated by applying random mutations to the selected parents. To respect the mutation budget without inefficient rejection sampling, we restrict changes to mutated positions if the budget of 5 mutations has been met. This allows for lateral moves or reversions while satisfying constraints.

Product of Experts (ESM-2 and AbLang-1). Following Gordon et al., we sampled variants using protein language models with a product of experts (PoE) strategy to incorporate oracle guidance. This method uses Gibbs sampling from a joint distribution $P(x) \propto P_{\text{MLM}}(x) \cdot \exp(\lambda F(x))$, where P_{MLM} is a masked language model prior. We evaluated both a general protein language model (ESM-2 150M) and an antibody-specific model (AbLang-1). We used AbLang-1 instead of AbLang-2 because the latter strictly requires paired antibody sequences instead of a single VH chain. We set the guidance strength to $\lambda = 50.0$. To respect the computational budget of ≤ 5 oracle calls per sample, we implemented a caching approximation. Before sampling, we pre-computed the fitness effects of all possible single mutations in CDRs. During the Gibbs sampling process, oracle scores

were retrieved from this additive cache rather than re-evaluated using the mutated sequence. This approximates the fitness landscape as locally additive.

Greedy Search. As a performance upper bound, we implemented a stochastic hill-climbing strategy. At each step, the algorithm evaluates all possible single mutations ($L \times 19$ variants), selects the top $K = 15$ candidates by fitness gain, and samples the next step using a softmax over their fitness improvements ($\Delta F/T$, with $T = 1.0$). This process repeats for up to 5 steps. While effective, this method requires orders of magnitude more compute ($\sim$2700 oracle calls per sequence) and serves only as a reference for the maximum achievable affinity under the given constraints.

Evaluation Metrics. Performance was evaluated across three axes: **Fitness**, measured by the mean and maximum improvement in predicted binding affinity; **Diversity**, quantified by the average pairwise distance within the generated samples; and **Naturalness**, measured using the OASis score.

C SUPPLEMENTARY RESULTS

C.1 ESTIMATION ERROR OF $Q_\theta(x)_\ell$ VIA SNR-WEIGHTED MLE

Under the transition likelihood of the CoSiNE model in Equation (1), maximum likelihood inference of $Q_\theta(x)_\ell$ is performed by minimizing the negative log-likelihood of the observed evolutionary transitions

$$\mathcal{L}(\theta) = -\sum_{\tau=(x,y,t)} \sum_{\ell=1}^{L} \log \exp(tQ_\theta(x)_\ell)_{x_\ell, y_\ell}$$

Unfortunately, directly training with this objective will generally fail to satisfy the assumption in Proposition D.1, which relies on $(Q_\theta(x)_\ell)_{x_\ell,\cdot}$ approximating the instantaneous non-zero rates in $\mathbf{Q}_{x,\cdot}$. Indeed, on branches with large t, this factorized objective encourages $Q_\theta(x)_\ell$ to learn *effective* rates that account for unobserved epistatic interactions and intermediate states between x and y. This results in a time scale dependent bias where the inferred rates for long branches may diverge from $\mathbf{Q}$.

To mitigate this issue, we sought to leverage the insight that the signal-to-noise (SNR) ratio of our first-order approximation scales as $O(1/t)$. We therefore proposed a SNR-weighted loss function:

$$\mathcal{L}_{SNR}(\theta) = \sum_{\tau=(x,y,t)} \frac{1}{\delta + t} \mathcal{L}(\theta; x, y, t)$$

where $\delta \in [0, 1]$ is a hyper-parameter that controls the magnitude of the weighting. In the synthetic codon experiments (detailed in Section B.1), we were able to show that in a data-rich setting, SNR weighting indeed reduces the estimation error compared to an unweighted MLE objective, especially at high epistasis levels for the ground truth rate matrix (Figure S2). In addition, we found that setting $\delta = 1 - \varepsilon$ yielded the strongest results, which is intuitive since re-weighting is meant to compensate for epistatic effects in the underlying process.

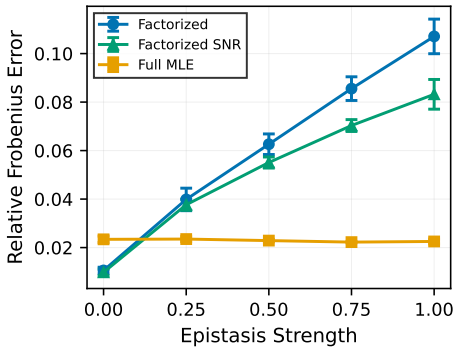


Figure S2: **As ε increases, SNR-weighting (green) reduces the relative Frobenius norm error between the estimated and true rate matrices compared to unweighted MLE (blue), while the Full MLE (yellow) is constant.**

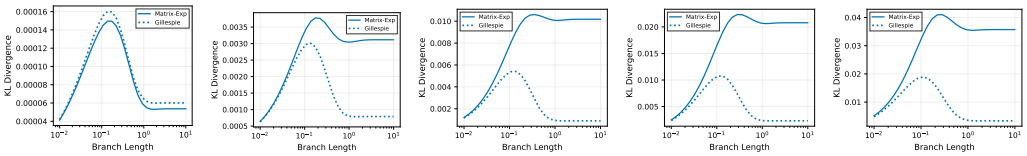
Motivated by this result, we trained CoSiNE on the full clonal dataset using this SNR-weighted objective at different values of $\delta \in \{0.0, 0.5, 1.0\}$. However, we could not identify improvements for these models in comparison to a CoSiNE model trained with the un-weighted loss. We hypothesize that this is due to a relatively data-poor setting for the antibody environment, given that our state space is much larger than the synthetic environment and we have less training data.

C.2 SAMPLING ERROR IN SYNTHETIC CODON EXPERIMENT

Using the synthetic codon environment detailed in Section B.1, we sought to quantify the difference between the sampling distributions of CoSiNE and the true transition probability distribution of the underlying process. For each level of epistasis $\varepsilon \in \{0.0, 0.25, 0.5, 0.75, 1.0\}$, we generated a ground truth rate matrix and trained a CoSiNE model on simulated transitions (see Section B.1). We evaluate performance by measuring the KL divergence between the ground truth transition distribution $P(y \mid x, t)$ and the distributions induced by either Gillespie sampling (Algorithm S1) or the per-site matrix exponentiation approach (Equation (1)).

We found that Gillespie sampling is superior at all epistasis levels and branch lengths. As expected, both sampling methods perform comparably when $\varepsilon = 0$, since the ground truth rate matrix is also site-independent (Figure S3a). However, at $\varepsilon = 1.0$, notice that the KL divergence of the matrix exponential scales quadratically in t before plateauing at a high error, indicating a failure to capture the correct stationary distribution (Figure S3e). In contrast, Gillespie sampling maintains consistently lower divergence and recovers the correct stationary distribution. This is an important result since it empirically validates Proposition D.1 and supports the claim in Lemma D.2. As ε increases, we

notice that the advantage of Gillespie sampling over the factorized matrix exponential approach also becomes more significant.

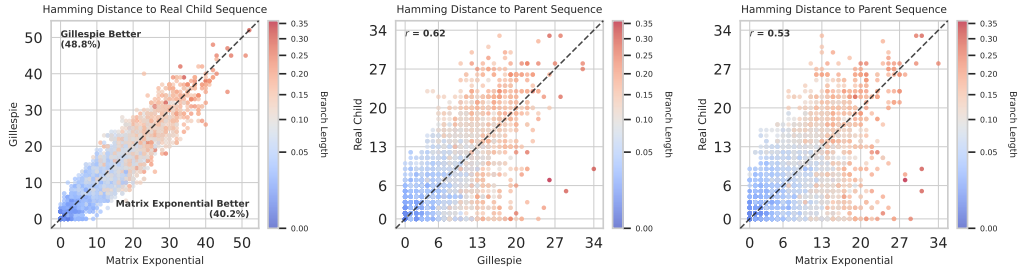


(a) $\varepsilon = 0.00$ (b) $\varepsilon = 0.25$ (c) $\varepsilon = 0.50$ (d) $\varepsilon = 0.75$ (e) $\varepsilon = 1.00$

Figure S3: **Gillespie samples are more consistent with the true transition probability distribution than samples from the factorized CoSiNE likelihood at all levels of epistasis.** The difference increases smoothly with the epistasis strength ε in the underlying rate matrix. Ground truth matrix generation and estimation protocols are in Section B.1.

C.3 Sampling Error in Real Antibody Experiment

Although we cannot obtain the transition likelihood for Gillespie samples in the large antibody sequence state space, we sought to validate the synthetic codon results (Section C.2) in the real antibody environment using auxiliary metrics. We selected all clonal trees in the test split with ≥ 4 leaves and sampled new sequences for each tree using the model trained in Section 4.1, conditional on the root sequence. For each leaf node in each tree, we collected the corresponding sampled sequence from both Gillespie and factorized matrix exponential approaches and compared their hamming distance to the real leaf sequence at that node. Assuming that a lower hamming distance to the real leaf sequence indicates lower sampling error, we observe in Figure S4a that leaves simulated with Gillespie sampling are closer to the real leaf sequence in $\sim$49% of cases, whereas the opposite is true only $\sim$40% of the time. Next, we computed the hamming distance from the root sequence to every real and simulated leaf sequence. Under the assumption that this distance should be somewhat consistent for corresponding real and simulated leaves, we scatterplot their correlation in Figure S4b for Gillespie samples and Figure S4c for factorized matrix exponential samples. Once again, we find that Gillespie has superior performance, achieving a Pearson correlation of 0.62 while the per-site matrix exponentiation approach only obtains 0.53. Altogether, these results further consolidate our results in the synthetic codon environment and strongly suggest that Gillespie sampling is a principled and superior sampling approach for CoSiNE.



(a) (b) (c)

Figure S4: **Gillespie sampling exhibits lower sampling error and better preserves evolutionary distances on real antibody clonal families.** a) Comparison of sampling fidelity: Gillespie samples are closer (in Hamming distance) to the true held-out leaf sequence in 48.8% of cases, compared to 40.2% for the factorized matrix exponential. b, c) Correlation between the root-to-leaf Hamming distances of real versus simulated leaf sequences. Gillespie sampling (b) achieves a higher Pearson correlation ($r = 0.62$) with the observed evolutionary distances than the factorized matrix exponential (c) approach ($r = 0.53$).

C.4 Log-Likelihood versus Selection Score for DMS Performance

Figure S5 illustrates the effect of our correction method on the Koenig Light Chain expression dataset. In the left plot, there is a clear separation in the data according to the number of nucleotide edits between the wildtype and mutant sequences. This indicates that the model has learned that for a fixed amount of elapsed time, transitions with fewer nucleotide edits are more probable than transitions with more nucleotide edits. This makes sense considering our chosen branch length is somewhat short ($t = 0.2$) and the model is trained to maximize the likelihood of observed transitions. However, if we naively correlate the model's predicted likelihood for the sequence with selection, we introduce the bias that mutations with few nucleotide edits are higher fitness than mutations with more nucleotide edits, which is clearly erroneous. In the plot on the right, we can see that taking the ratio with the likelihood under the Thrifty model removes this edit distance bias as indicated by the fact that the point clouds corresponding to each edit distance now roughly occupy the same space.

The selection score correction does more than just correct for this edit distance bias. For example, certain nucleotide substitutions are more likely to occur than others and without accounting for this variation, those biases will be interpreted as improvements in fitness. In Figure S5 we see that even among mutants with the same nucleotide edit distance, the correlation improves when adding the SHM correction.

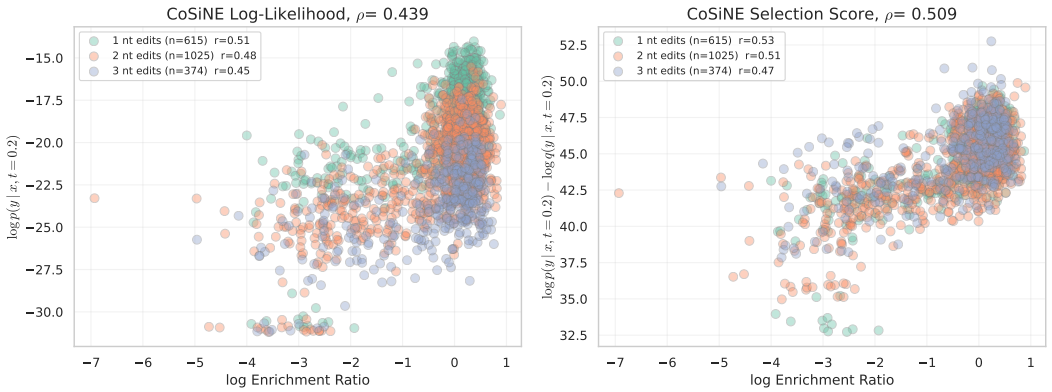


Figure S5: **Analysis of CoSiNE scoring methods against log Enrichment Ratios for the Koenig DMS assay measuring expression (Light Chain). Left**: Log-Likelihood performance ($\rho = 0.439$). **Right**: Selection Score performance ($\rho = 0.509$). Points are colored by the edit distance from wildtype codon to mutant codon. We assume the mutant codon is the one which requires the smallest number of nucleotide edits.

C.5 Additional VEP Experiments

In Table S2 we evaluate CoSiNE on the same datasets as Section 4.2 when conditioning on different amounts of context. Surprisingly, conditioning on just a single chain results in a higher correlation on four out of seven datasets. However, when paired conditioning outperforms, it does so by an average Spearman correlation of **0.121**, compared to just **0.006** when single chain conditioning outperforms. This suggests that inter-chain interactions are crucial for modeling fitness but only under certain contexts. What differentiates these contexts is a direction for future work. It's also worth noting that the majority of the training data for CoSiNE is single chains, suggesting this result could be an artifact of training data composition.

Table S2: **Comparison of CoSiNE on zero-shot VEP with both the heavy and light chains provided as context (Paired) versus just the chain with the mutations (Single).** Blue indicates datasets with mutations on the heavy chain, and red indicates mutations on the light chain.

Model	Expression			Binding			
	Koenig (H)	Koenig (L)	Adams	Koenig (H)	Koenig (L)	Shaneh. (119)	Shaneh. (120)
CoSiNE-Paired	**0.613**	0.508	**0.464**	**0.456**	0.371	0.498	0.536
CoSiNE-Single	0.545	**0.509**	0.234	0.390	**0.375**	**0.504**	**0.549**

In Table S3, we report results from the same experiment as Table 1, but use Pearson correlation instead of Spearman.

Table S3: **Comparison of deep protein models on VEP benchmarks across expression and binding landscapes as measured by Pearson correlation.**

Model	Expression			Binding			
	Koenig (H)	Koenig (L)	Adams	Koenig (H)	Koenig (L)	Shaneh. (119)	Shaneh. (120)
AbLang-2	0.153	−0.109	−0.096	−0.114	−0.108	0.263	0.166
ESM2-150M	0.476	0.539	−0.119	0.044	0.266	0.215	0.197
ESM2-650M	0.384	0.416	0.097	0.009	0.243	0.191	0.308
ProGen2-Small	0.559	0.568	−0.043	0.156	0.276	0.074	0.052
ProGen2-Medium	0.553	0.579	0.209	0.123	0.253	0.296	0.275
DASM	**0.688**	0.674	0.221	0.335	0.316	0.458	0.518
CoSiNE (ours)	0.687	**0.696**	**0.409**	**0.367**	**0.345**	**0.502**	**0.521**

C.6 Per-Site Test Likelihood of CoSiNE versus DASM+Thrifty

For each evolutionary transition $\tau = (x, y, t)$ in the Rodriguez test set, we compared the per-site likelihood $\frac{1}{L}p(y \mid x, t) =$ obtained with CoSiNE against that obtained with DASM+Thrifty (Figure S6). A higher value indicates a better fit on the test data. As expected, we see that transitions with lower branch lengths are easier to fit, as they have fewer mutations between x and y. As the branch length increases, both models have lower per-site likelihoods.

Overall, CoSiNE outperforms DASM+Thrifty in 62.3% of transitions and is significantly better on transitions with long branch lengths. We attribute this improvement to the increased expressiveness of our model compared to older approaches like DASM+Thrifty, which do not deal with the full sequence state space.

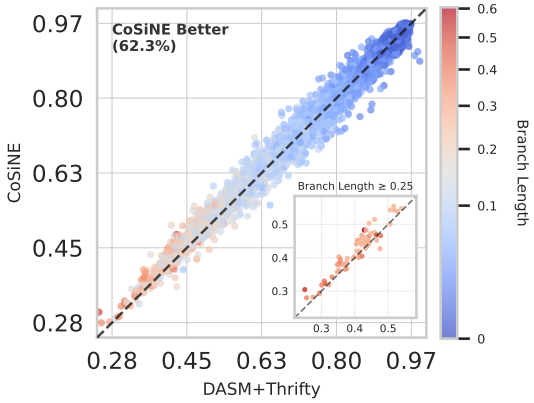


Figure S6: **Mean per-site likelihood of CoSiNE versus DASM+Thrifty on held out evolutionary transitions from the test set**. CoSiNE achieves better model fit, especially on transitions with longer branch lengths ($t \geq 0.25$).

C.7 Taylor-Series Approximated Guidance versus Exact Guidance

To evaluate the utility of our first-order Taylor series approximation for oracle guidance, we compared the fitness improvements and computational costs of exact guidance versus Taylor series approximated guidance. Using the same SARS-CoV-1 oracle as before (Section B.4), we sampled sequences at $t = \{0.01, 0.05, 0.10\}$ with guidance strength $\gamma = 2.0$. We generated 5 samples from each of 3 randomly selected seed antibody sequences, yielding 15 samples per condition.

As shown in Figure S7 (left), both guidance methods produced similar fitness improvements across all branch lengths. Two-sided t-tests confirmed that these differences were not statistically significant at any branch length ($p = 0.571$, $p = 0.918$, and $p = 0.695$), indicating that Taylor approximation does not compromise the quality of *Guided Gillespie* samples.

In contrast, the computational costs differed dramatically between methods (Figure S7, right). Exact guidance requires evaluating the oracle on all single-amino-acid mutants at each Gillespie step, compared to a single call for TAG guidance. This achieves speedups of 488× (0.01), 928× (0.05), and 916× (0.10) across the three branch lengths.

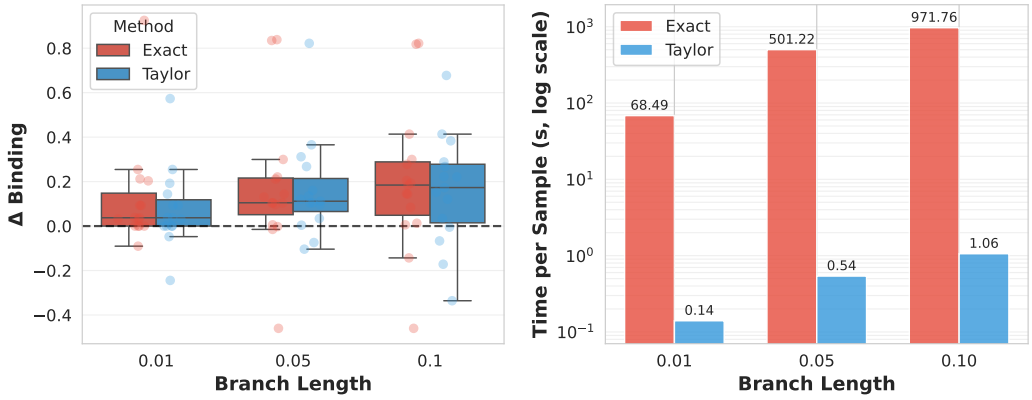


Figure S7: **TAG guidance matches exact guidance performance with 500–900× speedup**. (Left) Fitness improvements and (right) runtime comparison across branch lengths. No significant fitness differences ($p > 0.05$), but TAG is orders of magnitude faster.

C.8 Local Antibody CDR Optimization Results

Beyond simulating affinity maturation over long branch lengths, we evaluated whether CoSiNE could perform constrained local optimization. We tasked the model with refining a SARS-CoV-1 binder using a strict budget of five mutations within the CDR regions. Instead of conditioning on a particular branch length, we fixed the number of Gillespie steps taken, which ensures that the mutation ceiling is respected. We benchmarked against a genetic algorithm (GA) and a Product-of-Experts (PoE) sampler inspired by Gordon et al. (2024) that steers protein language models (ESM-2 and AbLang-1) using the oracle. Among these methods, CoSiNE achieved the highest gains in predicted binding affinity while maintaining high humanness (Table S4) as seen in Table S4. A full description of this experiment is provided in Section B.5.

Table S4: **Diversity, naturalness, and binding affinity metrics for constrained CDR optimization of SARS-CoV-1 binder.**

Method	Unique	E_{dist}	IntDiv ↑	OASis ↑	Δ Bind$_{\text{avg}}$ ↑	Δ Bind$_{\text{max}}$ ↑	N_{oracle} ↓
Greedy*	0.996	5.00	6.21	0.807	0.385	0.466	2756
Genetic Algo.	0.991	4.47	7.30	0.789	<u>0.196</u>	0.363	4.81
AbLang-1-*PoE*	<u>0.999</u>	4.97	<u>7.71</u>	**0.849**	0.167	<u>0.390</u>	0.580
ESM2-*PoE*	**1.00**	4.98	**8.47**	0.802	0.149	0.355	0.580
CoSiNE (ours)	0.988	4.56	6.83	<u>0.822</u>	**0.198**	**0.395**	5.00

C.9 Additional Guided Affinity Maturation Results

We provide additional results for the guided affinity maturation sampling experiment using other naive sequences from the OAS database. Notice that in all cases, guidance effectively steers the predicted binding affinity of the generated leaf sequences. Furthermore, the structural quality versus humanness of leaf sequences in Figure 3 are shown below in Figure S9.

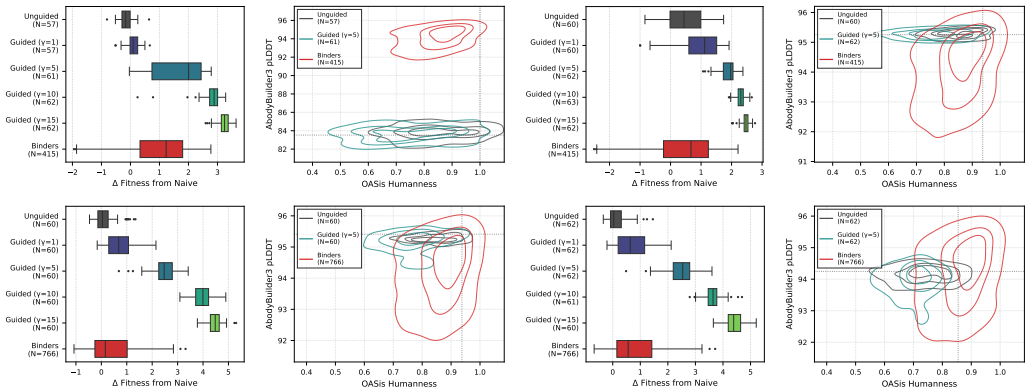


Figure S8: **Guided affinity maturation results for additional naive antibody sequences using CoV-1 (top) and CoV-2 (bottom) oracles.** We randomly selected naive IgM heavy sequences from the OAS database and recursively sampled down the tree in Figure S1.

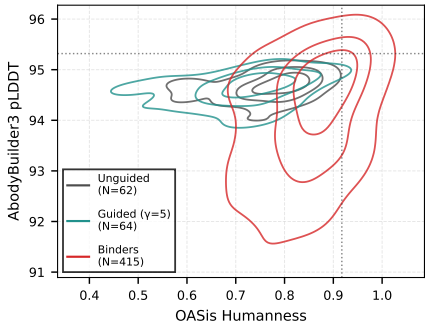


Figure S9: **Antibodies sampled by CoSiNE under guidance ($\gamma = 5$) maintain high structural quality (AbodyBuilder3 pLDDT) and humanness (OASis).** We compare against unguided leaf samples (black) and CoV-AbDab binders (red).

D THEORETICAL RESULTS

D.1 COSINE IS A FIRST-ORDER APPROXIMATION OF SEQUENTIAL POINT MUTATION PROCESS

Let $\mathbf{Q} \in \mathbb{R}^{|\mathcal{A}|^L \times |\mathcal{A}|^L}$ be the generator rate matrix of a sequential point mutation process over protein sequences of length L with amino acid vocabulary $\mathcal{A}$. The vector of transition probabilities to other states after some time t is given by indexing the transition matrix as follows:

$$P(\cdot \mid x, t) = \left(e^{t\mathbf{Q}}\right)_{x,\cdot} \in \mathbb{R}^{|\mathcal{A}|^L}$$

Proposition D.1. *Assume the per-site rate matrices $Q_\theta(x)_\ell$ are parameterized such that*

$$\left(Q_\theta(x)_\ell\right)_{x_\ell, y_\ell} = \mathbf{Q}_{x,y}$$

for all x, y with Hamming distance $d(x, y) = 1$ and ℓ is the unique site where x and y differ. Then, the error between the transition probability vectors is bounded such that

$$\|P(\cdot \mid x, t) - p_\theta(\cdot \mid x, t)\|_1 \leq (\lambda t)^2 = O(t^2) \tag{5}$$

where $\lambda = \max_x\{-\mathbf{Q}_{x,x}\}$ is the maximum exit rate of any given state.

Proof. Notice that we can reframe the transition probabilities from the per-site rate matrices using Kronecker products as follows

$$p_\theta(\cdot \mid x, t) = \left(e^{tQ_\theta(x)_1} \otimes \cdots \otimes e^{tQ_\theta(x)_L}\right)_{x,\cdot}$$

Using the identity $e^{tA} \otimes e^{tB} = e^{t(A \oplus B)}$ we can define a **Kronecker-sum generator matrix**

$$\mathbf{Q}_\theta(x) := Q_\theta(x)_1 \oplus \cdots \oplus Q_\theta(x)_L$$

This makes a comparison between the full state space transition probabilities and per-site factorized transition probabilities more convenient

$$P(\cdot \mid x, t) = \left(e^{t\mathbf{Q}}\right)_{x,\cdot} \tag{6}$$

$$p_\theta(\cdot \mid x, t) = \left(e^{t\mathbf{Q}_\theta(x)}\right)_{x,\cdot} \tag{7}$$

Now, we can uniformize both transition kernels by picking $\lambda = \max_z\{-\mathbf{Q}_{z,z}\}$ (which is also equal to $\max_z\{-\mathbf{Q}_\theta(x)_{z,z}\}$) and defining embedded DTMCs

$$R = I + \mathbf{Q}/\lambda \tag{8}$$

$$S = I + \mathbf{Q}_\theta(x)/\lambda \tag{9}$$

Uniformization tells us

$$e^{t\mathbf{Q}} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} R^n, \quad e^{t\mathbf{Q}_\theta(x)} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} S^n,$$

Using the matching assumption for single-site mutants stated in the theorem, we know that

$$\mathbf{Q}_{x,\cdot} = \mathbf{Q}_\theta(x)_{x,\cdot} \implies R_{x,\cdot} = S_{x,\cdot}$$

Therefore, the $n = 0$ and $n = 1$ terms in the uniformization series match exactly for row x, meaning the differences start only at $n \geq 2$

$$\left(e^{t\mathbf{Q}} - e^{t\mathbf{Q}_\theta(x)}\right)_{x,\cdot} = \sum_{n=2}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} \left(R^n - S^n\right)_{x,\cdot}$$

Because $(R^n)_{x,\cdot}$ and $(S^n)_{x,\cdot}$ are valid probability distributions, $\| (R^n - S^n)_{x,\cdot} \|_1 \leq 2$, allowing us to bound the L_1 error between the true transition probability vector and per-site factorized probability vector as

$$\|P(\cdot \mid x,t) - p_\theta(\cdot \mid x,t)\|_1 = \left\| \left(e^{t\mathbf{Q}} - e^{t\mathbf{Q}_\theta(x)} \right)_{x,\cdot} \right\|_1 \leq 2 \sum_{n=2}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} = 2\left(1 - e^{-\lambda t}(1+\lambda t)\right)$$

Recognizing that the series $1 - e^{-u}(1+u) = u^2/2 - u^3/6 + \cdots \leq u^2/2$ for $u \geq 0$ allows us to simplify the bound to*

$$\|P(\cdot \mid x,t) - p_\theta(\cdot \mid x,t)\|_1 \leq (\lambda t)^2 = O(t^2)$$

□

Intuitively, the theorem shows that our model offers a principled way of capturing first-order mutation dynamics while restricting the approximation error to epistatic effects that grow quadratically in the branch length. Moreover, this approximation is particularly well-suited for affinity maturation because high-frequency clonal selection typically constrains lineages to short branch lengths. In this regime, the first-order evolutionary signal captured by our model $O(\lambda t)$ dominates the quadratic approximation error $O(\lambda^2 t^2)$.

D.2 Exactness of Gillespie Sampling

Although the factorized transition probability in Equation (1) allows for trivial sampling, it incurs an error with respect to the more expressive sequential point mutation process that scales quadratically with the branch length. To address this limitation, we propose a Gillespie sampling algorithm adapted to the instantaneous rates of the CoSiNE model. In Lemma D.2, we show that this procedure is principled and provably samples from $P(\cdot \mid x,t)$ under certain conditions.

Lemma D.2. *Let $x_0, \ldots, x_{t_{N-1}}, x_{t_N}$ be the trajectory of sequences sampled from the Gillespie Algorithm S1, using branch length t and starting sequence x_0. For all $x \in \{x_0, \ldots, x_{N-1}\}$, assuming that*

$$(Q_\theta(x)_\ell)_{x_\ell, y_\ell} = \mathbf{Q}_{x,y}$$

holds for all sequences y with Hamming distance $d(x,y) = 1$, then $x_{t_N} \sim P(\cdot \mid x_0, t)$.

Proof. The proof follows from the fact that a continuous-time Markov chain is uniquely characterized by its holding time distributions and jump chain probabilities. By construction, the algorithm computes the total exit rate $\lambda_x \leftarrow -\sum_{\ell=1}^{L} Q_\theta(x)_{x_\ell, x_\ell}$ to sample the holding time. Under the lemma's condition that $(Q_\theta(x)_\ell)_{x_\ell, y_\ell} = \mathbf{Q}_{x,y}$ for all single residue mutants y, this sum is exactly equal to the target exit rate $-\mathbf{Q}_{x,x}$. Subsequently, the algorithm selects the next state y, corresponding to mutation (ℓ^*, a^*), with probability $P(\ell, a) = (Q_\theta(x)_\ell)_{x_\ell, a}/\lambda_x$. Due to the previous set of assumptions, this is identical to the transition probability $\mathbf{Q}_{x,y}/-\mathbf{Q}_{x,x}$ of the target process. Since both the exponential holding times and discrete transition probabilities match those defined by the generator $\mathbf{Q}$ at every step, the simulated trajectory is a statistically exact realization of the true process, ensuring x_{t_N} is distributed according to $P(\cdot \mid x_0, t)$. □

The validity of Proposition D.1 and Lemma D.2 relies on $(Q_\theta(x)_\ell)_{x_\ell, y_\ell}$ approximating the instantaneous non-zero rates in the point mutation process rate matrix $\mathbf{Q}_{x,y}$. However, because CoSiNE is trained with the approximate transition probability in Equation (1), its estimated rates will incur some bias due to the marginalization of unobserved intermediate states on long branches. We acknowledge this theoretical limitation of CoSiNE and study its effect thoroughly in Section C.1. In practice, we show that the transition probability error of our model indeed scales quadratically with the branch length and that Gillespie sampling produces final state distributions that match those of the true process much more closely (Sections C.2 and C.3).

*The constant λ corresponds to the maximum total substitution rate out of any sequence state. Under standard sequential point-mutation models, this rate scales at most linearly with sequence length and is bounded by the sum of per-site mutation rates.

D.3 Relation Between Fixation Probability and Relative Fitness

Following Kimura (1962), we define P_{fix} for a haploid population (which we use to describe affinity maturation due to asexual reproduction of B cells and allelic exclusion, which ensures that only one allele is expressed per B cell) as

$$P_{\text{fix}}(x \to y) = \frac{1 - e^{-2s_{xy}}}{1 - e^{-2N_e s_{xy}}}, \tag{10}$$

where x is the wildtype allele, y is a newly introduced allele, $s_{xy} = F_y - F_x$ is the selective advantage of allele y over allele x, and N_e is the effective population size. Crucially, the fixation probability expression is monotonic with respect to s_{xy}, which explains why the selection score calculated by Equation (3) shows strong Spearman correlation with empirical relative fitness measurements.