A Kernelized Stein Discrepancy for Biological Sequences

Alan N. Amin [1] **Eli N. Weinstein** [*2] **Debora S. Marks** [*1 3]

Abstract

Generative models of biological sequences are a powerful tool for learning from complex sequence data, predicting the effects of mutations, and designing novel biomolecules with desired properties. To evaluate generative models it is important to accurately measure differences between high-dimensional distributions. In this paper we propose the "KSD-B", a novel divergence measure for distributions over biological sequences that is based on the kernelized Stein discrepancy (KSD). The KSD-B can be evaluated even when the normalizing constant of the model is unknown; it allows for variable length sequences and can take into account biological notions of sequence distance. Unlike previous KSDs over discrete spaces the KSD-B (a) is theoretically guaranteed to detect convergence and non-convergence of distributions over sequence space and (b) can be efficiently estimated in practice. We demonstrate the advantages of the KSD-B on problems with synthetic and real data, and apply it to measure the fit of state-of-the-art machine learning models. Overall, the KSD-B enables rigorous evaluation of generative biological sequence models, allowing the accuracy of models, sampling procedures, and library designs to be checked reliably.

1. Introduction

Generative models of biological sequences have wide and growing application in protein design, phylogenetic analysis, clinical human genetics, epidemiology and beyond (Hopf et al., 2017; Riesselman et al., 2018; Russ et al., 2020; Shin et al., 2021; Frazer et al., 2021; Davidsen et al., 2019; Weinstein et al., 2022b; Thadani et al., 2022; Meier et al., 2021; Madani et al., 2020; Alley et al., 2019; Marcou et al., 2018). A central challenge in working with generative biological sequence models, as for all generative models, is model evaluation. How accurate are density estimates and how realistic are sequences sampled from the model?

Common existing evaluation techniques for generative sequence models do not provide an absolute and reliable measure of model fit. For example, generative sequence models are often evaluated based on their log likelihood; but this only addresses whether one model fits to the data better than another, not whether any model matches the data absolutely. Another evaluation strategy is to draw sequences from the model and test whether they match the data based on expertly chosen statistics or predictions of sequence properties (hydrophobicity, secondary structure, etc.). However, a generative model that passes such a test may still poorly fit the data outside of the chosen statistics.

In this article, we address the model evaluation problem by developing a divergence to compare distributions over sequences. We call this divergence the "KSD-B", as it is based on the kernelized Stein discrepancy (KSD) but applies to biological sequences (B) (Chwialkowski et al., 2016; Liu et al., 2016). The KSD-B allows us to answer whether or not the distribution over sequences produced by a generative model, p, equals the distribution of the data, q. That is, it enables a goodness of fit test for biological sequences. The KSD-B provides an absolute rather than a relative measure of model quality, and, given enough data, it is able to detect any difference between p and q.

KSDs compare distributions by (1) building a stochastic process that is stationary for p, (2) applying it to q, and (3) using a kernel to evaluate how much q changes. Much emphasis has been placed previously on the study of stochastic processes and kernels over Euclidean space, i.e. $\mathbb{R}^d$, such as diffusion processes, and Gaussian, inverse multiquadric or Matèrn kernels. The KSD-B, by contrast, is defined over sequence space $S = \cup_{L=1}^{\infty} \mathcal{B}^L$, the set of all finite length strings of an alphabet $\mathcal{B}$, where $\mathcal{B}$ is nucleotides for DNA or amino acids for proteins (Amin et al., 2021; Weinstein, 2022). The KSD-B uses a stochastic process based on substitutions, insertions and deletions (the sorts of mutations often seen in biological sequences) and uses kernels such as alignment kernels and kmer spectrum kernels (which capture biological notions of sequence similarity).

Despite the shift in setting, we build the stochastic process and kernel carefully so that the KSD-B shares with the orig-

[*]Equal contribution [1]Harvard Medical School [2]Columbia University [3]Broad Institute of Harvard and MIT. Correspondence to: Alan N. Amin <alanamin@g.harvard.edu>, Eli N. Weinstein <ew2760@columbia.edu>, Debora S. Marks <debbie@hms.harvard.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

inal Euclidean KSD a number of desirable properties and theoretical guarantees. First, it faithfully measures whether p and q are equal for any p and q, i.e. KSD-B$_p(q) = 0$ if and only if $p = q$. This means the KSD-B can detect any difference between p and q. Second, the KSD-B detects convergence and non-convergence: KSD-B$_p(q_n)$ converges to 0 as $n \to \infty$ if and only if q_n converges to p (Gorham & Mackey, 2017). This protects against big mistakes by the KSD-B in practice: if KSD-B$_p(q)$ is very close to zero (say, within the noise of estimation), q must at least be very similar to p, not completely different. Third, KSD-B$_p(q)$ can be computed using only unnormalized probabilities from p and samples from q. This means the KSD-B can be used even when the normalizing constant of the model is intractable, as in energy-based generative models. Finally, the KSD-B can be efficiently estimated in realistic settings.

As a computable divergence, the KSD-B is a broadly useful tool for more than evaluating model fit. For instance, it can be used to evaluate model samples. Often it is only possible to draw *approximate* samples from a posterior, for instance in the context of semi-supervised protein design or ancestral sequence reconstruction. Since the KSD-B requires only unnormalized probabilities, and can detect convergence and non-convergence, it can be used to determine whether the distribution of approximate samples matches the model (Gorham & Mackey, 2017). The KSD-B can also be used to design libraries based on generative models. Since the KSD-B can detect convergence and non-convergence, we can find a set of samples that are representative of p by minimizing KSD-B$_p(q)$ with respect to an empirical distribution of samples q (Chen et al., 2018; 2010).

Sec. 2 provides background on kernelized Stein discrepancies. Sec. 3 introduces a novel class of discrete Stein discrepancies. Sec. 4 defines the KSD-B. Sec. 5 proves that it is faithful and detects both convergence and non-convergence. Sec. 6 describes how the KSD-B can be estimated efficiently. Sec. 7 develops kernels for the KSD-B. Sec. 8 demonstrates the KSD-B empirically. Sec. 9 concludes.

Related Work Computable Stein discrepancies were first developed for Euclidean space (Gorham & Mackey, 2015; Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017). There have been a number of generalizations to discrete space (Shi et al., 2022; Yang et al., 2018a; Han et al., 2020; Hodgkinson et al., 2020). However, these methods only apply to finite discrete spaces, and so, biologically, they are only appropriate in the special case where all sequences have the same length. Our method differs in that it applies to the infinite discrete setting of S, and thus allows for arbitrary length variation.

During preparation of this manuscript, Baum et al. (2022) proposed a method similar to the KSD-B, that is applicable to infinite discrete spaces such as S. We go further by proving strong theoretical guarantees for the KSD-B, and showing how the method of Baum et al. (2022) can fail to satisfy these guarantees. We also introduce a rigorous approximation method that is substantially more computationally efficient than that of Baum et al. (2022) in practice.

There are a small number of other approaches to nonparametric testing of generative biological sequence models. Amin et al. (2021) develop a Bayesian goodness of fit test (BEAR), but it depends on access to normalized likelihoods from the model and does not provide guarantees for convergence detection. Amin et al. (2023) develop a maximum mean discrepancy (MMD) two-sample test with guarantees for convergence detection, but it requires samples from the model. Even when these alternatives are applicable, we find empirically that the KSD-B can outperform both.

2. Background

The KSD-B builds on and extends the notion of a Stein discrepancy, which is a type of integral probability metric.

Integral Probability Metrics (IPMs) IPMs are a general method for measuring the difference between two distributions p and q. They compute the maximum difference in expectation between p and q over a set of test functions $\mathcal{F}$, i.e. $\sup_{f\in\mathcal{F}} |E_q f - E_p f|$. Here, $E_q f = \sum_{X\in S} f(X)q(X)$ is the expectation with respect to q of a function f on sequences. Many popular divergences can be written as IPMs, e.g. choosing $\mathcal{F}$ to be the set of bounded functions gives the total variation distance. In general, the success of an IPM at detecting if $p \neq q$ depends on the size of $\mathcal{F}$, with larger function families offering better discrimination.

IPMs are an especially useful choice of discrepancy for biological sequence models, where the ultimate goal is often to synthesize and test samples from a model p in the laboratory. In particular, let f^* denote the mapping from sequence to phenotype of interest (protein stability, binding, etc.). In general, f^* is not known before performing extensive experiments; however, if $\mathcal{F}$ is large, it can be reasonable to assume $f^* \in \mathcal{F}$. Then, a small IPM guarantees that samples from p will have similar phenotypes to samples from q, as $|E_q f^* - E_p f^*| \leq \sup_{f\in\mathcal{F}} |E_q f - E_p f|$ (Weinstein et al., 2022b;a). This guarantees, for instance, that samples from a generative model have similar phenotypes to the natural sequences it was trained on.

Stein Discrepancies To evaluate an expectation such as $E_p f$, one would typically require samples or normalized probabilities from p. However, these are not always available, for instance if p is an energy-based model, or the posterior of a semi-supervised model. The Stein discrepancy solves this problem by constructing a family of test functions $\mathcal{F}$ for which $E_p f$ is exactly zero, but which is still sufficiently large to detect any difference between p and q.

The basic idea is to use a continuous time Markov process with p as its stationary distribution. In particular, consider a generator $\mathcal{L}_p$, defined as $(\mathcal{L}_p f)(X) =$

$\lim_{t\to 0} \frac{1}{t}(E[f(X^t)] - f(X))$, where X^t is the position of the Markov process initialized at X after evolving for time t (Barbour, 1990; Gorham et al., 2019; Shi et al., 2022). Now, $E_q \mathcal{L}_p f$ describes the amount that the expectation of f changes as q evolves under the Markov process. If $q \neq p$, then evolving q will change it to be closer to p, so intuitively, if $\mathcal{F}$ is large enough, there must exist some function f that changes expectation, i.e. $\sup_{f\in\mathcal{F}} |E_q \mathcal{L}_p f| > 0$. However, if $q = p$, the expectation of f will not change at all since p is stationary, i.e. $\sup_f |E_p \mathcal{L}_p f| = 0$. Thus if we set $\mathcal{F} = \mathcal{L}_p(\mathcal{F}')$ for a set of functions $\mathcal{F}'$, the IPM becomes $\sup_{f\in\mathcal{F}} |E_q f - E_p f| = \sup_{f\in\mathcal{F}'} |E_q \mathcal{L}_p f|$, so it no longer depends on an expectation with respect to p, but can still detect if $q \neq p$ for $\mathcal{F}'$ sufficiently large. This IPM is a "Stein divergence", and $\mathcal{L}_p$ a "Stein operator". Markov chain Monte Carlo (MCMC) methods are a useful tool for building Stein operators, as they allow construction of a Markov processes with stationary distribution p even when the normalizing constant of p is unknown.

Diffusion Stein Discrepancies On Euclidean space, a classic choice of operator is the Langevin Stein operator, which is derived from a Langevin sampler (Liu et al., 2016; Chwialkowski et al., 2016; Gorham & Mackey, 2017). The Langevin Stein operator is an instance of a diffusion Stein discrepancy, which are derived from Itô diffusions, and have strong theoretical properties, in particular they detect non-convergence (Gorham et al., 2019; Barp et al., 2019).

The properties of diffusion Stein discrepancies stem in part from a subtle but important extension of the discrepancy $\sup_{f\in\mathcal{F}} |E_q \mathcal{L}_p f|$ to a larger set of test functions. When the generator $\mathcal{L}_p$ comes from a diffusion, it can be written in the form $\mathcal{L}_p = \mathcal{T}_p \nabla$, where $\mathcal{T}_p$ is another operator and ∇ is the gradient operator. This gives a Stein discrepancy $\sup_{f\in\mathcal{F}} |E_q \mathcal{T}_p \nabla f| = \sup_{g\in\nabla\mathcal{F}} |E_q \mathcal{T}_p g|$, i.e. we can think of the discrepancy as maximizing over the set of gradients of functions in $\mathcal{F}$. Now, $\mathcal{F}$ is a set of functions from $\mathbb{R}^d$ to $\mathbb{R}$, so $\nabla\mathcal{F}$ is a set of functions from $\mathbb{R}^d$ to $\mathbb{R}^d$, i.e. each $g \in \nabla\mathcal{F}$ is a vector field rather than scalar field. However, not all vector field functions can be written as the gradient of scalar field functions. One can therefore consider expanding the class of test functions to the larger set of all vector fields, $\mathcal{G} \supset \nabla\mathcal{F}$, making the Stein discrepancy $\sup_{g\in\mathcal{G}} |E_q \mathcal{T}_p g|$. Since the set of test functions is larger, diffusion Stein discrepancies can more easily detect whether or not q matches p. Critically, the properties of Itô diffusions guarantee that we still have $\sup_{g\in\mathcal{G}} |E_p \mathcal{T}_p g| = 0$ (Gorham et al., 2019).

3. Discrete Vector Field Stein Discrepancies

In this section we describe Stein discrepancies for distributions on discrete spaces. (Our discussion in this section is not specific to sequence space S; it applies to any infinite discrete space.) On discrete spaces, diffusion Stein discrepancies are not applicable, since they depend on gradients. We introduce a novel approach to expanding the set of test functions on discrete spaces, which will be essential for guaranteeing that our new discrepancy, the KSD-B, detects non-convergence and convergence. The idea is analogous to diffusion Stein discrepancies, with finite differences replacing gradients, and the property of detailed balance replacing the condition that the Markov process is an Itô diffusion.

To construct a Stein operator for a discrete space S, we first consider a generic continuous time Markov process (Shi et al., 2022). Let $T_{p,X\to Y}$ denote the transition rate of the process from sequence X to sequence Y. We assume this transition rate is zero except to a finite number of sequences Y that are near X, i.e. "mutants" of X; we write YMX if Y is a mutant of X, where M is a relation on S. Then, the Stein operator is,

$$(\mathcal{L}_p f)(X) = \sum_{Y\in S|YMX} T_{p,X\to Y}(f(Y) - f(X)),$$

since in a continuous time Markov process, the transition rate out of X must equal the total transition rate to other states, i.e. $T_{p,X\to X} = -\sum_{YMX} T_{p,X\to Y} = -\text{flux}_p(X)$. We can think of the quantity $f(Y) - f(X)$ for YMX as a discrete analogue of the gradient of f, as it looks at the difference between the value of f at adjacent points Y and X. We therefore define ∇f for any function f on S as $\nabla f(X, Y) = f(Y) - f(X)$ for YMX, following Chow et al. (2018). Now, we can write $\mathcal{L}_p f$ in the form $\mathcal{T}_p \nabla f$, so the discrete Stein discrepancy is $\sup_{f\in\mathcal{F}} |E_q \mathcal{T}_p \nabla f|$.

Assume the Markov process also satisfies detailed balance, i.e. $T_{p,X\to Y} p(X) = T_{p,Y\to X} p(Y)$. We can rearrange the Stein discrepancy (Proposition B.8), so $E_q \mathcal{T}_p \nabla f =$

$$\frac{1}{2} \sum_{YMX} q(X) T_{p,Y\to X} \left(\frac{p(Y)}{p(X)} - \frac{q(Y)}{q(X)} \right) \nabla f(X, Y), \quad (1)$$

where we have used the fact that $\nabla f(X, Y) = -\nabla f(Y, X)$. This equation gives some intuition for the discrete Stein discrepancy: the term $p(Y)/p(X) - q(Y)/q(X)$ compares the likelihood ratios of q and p at nearby points X and Y rather than their likelihoods at a single point, and so does not depend on the normalizing constant of p or q. If $p = q$, the difference in likelihood ratios is zero, so the entire equation is zero regardless of the value of ∇f.

We now expand the set of test functions. Instead of considering the Stein discrepancy $\sup_{f\in\mathcal{F}} |E_q \mathcal{T}_p \nabla f| = \sup_{g\in\nabla\mathcal{F}} |E_q \mathcal{T}_p g|$, we consider $\sup_{g\in\mathcal{G}} |E_q \mathcal{T}_p g|$, where $\mathcal{G} \supset \nabla\mathcal{F}$ is a set of functions from $S \times S$ to $\mathbb{R}$ which satisfy the anticommutative property $g(X, Y) = -g(Y, X)$. We refer to such functions g as "vector fields", following Chow et al. (2018). Thanks to the anticommutative property, Eqn. 1 still holds, with ∇f replaced by g. Using $\mathcal{G}$ instead of $\nabla\mathcal{F}$, we can have, for example, test functions g that are non-zero for just one edge (X, Y) in M; this is impossible for gradients ∇f, and allows the Stein discrepancy to test for more subtle differences between p and q. Moreover, Eqn. 1 shows that despite expanding the class of functions, the Stein discrepancy is still zero at $p = q$, i.e. $\sup_{g\in\mathcal{G}} |E_p \mathcal{T}_p g| = 0$.

We will refer to discrete Stein discrepancies that use vector fields, $\sup_{g\in\mathcal{G}} |E_q \mathcal{T}_p g|$, as "vector field Stein discrepancies", in contrast to the "scalar field Stein discrepancies" $\sup_{f\in\mathcal{F}} |E_q \mathcal{T}_p \nabla f|$ that have been studied previously (Shi et al., 2022; Baum et al., 2022). We will find that vector field Stein discrepancies have distinct theoretical and practical advantages over scalar field Stein discrepancies in the context of biological sequence data.

4. A Stein Discrepancy for Biological Sequences

In this section we define the KSD-B, a kernelized vector field Stein discrepancy for biological sequences.

Mutation We start by considering the relation M, that is, what pairs of sequences should be considered nearby. Typically in biology, two sequences are considered similar if they differ by a small number of mutations. We therefore say YMX if Y differs from X by a single mutation, either a substitution (changing a single letter of X), insertion (adding a single letter anywhere in X) or deletion (removing a single letter anywhere in X). This choice of M ensures that there are only a relatively small number of neighbors of each sequence, but one can still reach any point in S from any other by jumping from neighbor to neighbor.

Zanella Stein Operator To construct the Markov process over sequences, we use the framework of locally informed proposals, which is a general strategy for building Markov chain Monte Carlo methods on discrete spaces that satisfy detailed balance (Zanella, 2020; Shi et al., 2022). Consider any continuous non-negative, non-decreasing, and non-zero function χ that satisfies $\chi(t) = t\chi(1/t)$ for all $t > 0$ and $\chi(0) = 0$; examples are $\chi(t) = \sqrt{t}$ and $\chi(t) = \min\{t, 1\}$. For any YMX, define the transition rate from X to Y, $T_{p,X\to Y}$ as

$$\chi\left(\frac{p(Y)}{p(X)}\right) \times \#\{\text{single mutations taking } X \text{ to } Y\}, \quad (2)$$

with $T_{p,X\to Y} = \infty$ if $p(X) = 0$. The first term depends on the difference in probability under p of X and Y; in the case $\chi(t) = \min\{t, 1\}$, we can recognize it as the Metropolis-Hastings-Rosenbluth correction. The second term accounts for variation in sequence length, which creates situations where different mutations to X can lead to the same Y; for example, if $X = AA$, we can reach $Y = A$ by deleting either the first or the second position of X, so $\#\{\text{single mutations taking } X \text{ to } Y\} = 2$. With this construction, the Markov process satisfies detailed balance, i.e. $T_{p,X\to Y} p(X) = T_{p,Y\to X} p(Y)$ where we define $\infty \times 0 = 0$ throughout. The resulting Stein operator $\mathcal{T}_p\nabla$ is called the "Zanella Stein operator" (Hodgkinson et al., 2020). With this operator, we can define not only a scalar field Stein discrepancy but also a vector field Stein discrepancy, $\sup_{g\in\mathcal{G}} |E_q \mathcal{T}_p g|$.

Kernelization To make the discrepancy tractable and ensure $\mathcal{G}$ is sufficiently large to detect when $q \neq p$, we turn to reproducing kernel Hilbert spaces (RKHSs). As with the original kernelized Stein discrepancy, we let the set of test functions $\mathcal{G}$ be a unit ball in an RKHS $\mathcal{H}_k$ with kernel k, i.e. $\mathcal{G} = \{g : \|g\|_k \leq 1\}$ where $\|\cdot\|_k$ is the norm on $\mathcal{H}_k$ (Gorham & Mackey, 2017; Liu et al., 2016). To ensure that $\mathcal{H}_k$ only contains vector fields, i.e. functions $g : S \times S \to \mathbb{R}$ that satisfy anticommutivity, it is sufficient to use a kernel that also satisfies anticommutivity,

$$k((X, Y), (X', Y')) = -k((Y, X), (X', Y')). \quad (3)$$

Note, since kernels are symmetric, the analogous equation holds flipping X', Y'. We will see in Sec. 7 and Appx. B.4.3 how to construct vector field kernels that capture biological notions of sequence similarity.

Starting from a kernel $k : S\times S \to \mathbb{R}$ describing scalar fields on S, one can derive a kernel k^∇ satisfying Eqn. 3 such that the functions in $\mathcal{H}_{k^\nabla}$ are the gradients of the functions in $\mathcal{H}_k$ (Appx. B.3.2). We can thus obtain kernelized scalar field Stein discrepancies as a special case of kernelized vector field Stein discrepancies, using a "scalar field kernel", instead of a more general "vector field kernel".

KSD-B The KSD-B is defined as $\text{KSD-B}_{p,k}(q) = \sup_{\|f\|_k\leq 1} |E_q \mathcal{T}_p f|$. We can compute the supremum analytically (proof in Appx. B.3.1).

Proposition 4.1. *Say k is a vector field kernel and q is a p, k-integrable distribution on S, meaning $E_{X\sim q} \sum_{YMX} T_{p,Y\to X} \sqrt{k((X, Y), (X, Y))} < \infty$. Now, $\text{KSD-B}_{p,k}(q)^2 = (\sup_{\|g\|_k\leq 1} |E_q \mathcal{T}_p f|)^2$ is equal to*

$$E_{X,X'\sim q} \sum_{YMX, Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k((X, Y), (X', Y')).$$

If p is p, k-integrable, then for all $f \in \mathcal{H}_k$, $E_p \mathcal{T}_p f = 0$. (4)

5. Detecting convergence and non-convergence

In this section we establish the key theoretical properties of the KSD-B that makes it useful for goodness of fit tests, evaluating sample quality, and other applications. Our proofs are inspired by those for Euclidean space KSDs in Gorham & Mackey (2017); Gorham et al. (2019).

KSD-B is Faithful For the KSD-B to be useful as a measure of goodness of fit, it must be able to detect if a model distribution p matches a data distribution q. To do so, the divergence must be faithful: $\text{KSD-B}_{p,k}(q) \to 0 \iff p = q$. Faithfulness holds if the set of test functions is sufficiently large. For KSDs on Euclidean spaces, faithfulness is usually guaranteed by using a kernel that is universal, meaning $\mathcal{H}_k$ is dense on some function space (such as L^p space).

Over a discrete space such as S, we can use a set of test functions that is, in some sense, even larger. More precisely, there are kernels on S (but not $\mathbb{R}^d$) that have discrete masses, meaning their RKHS $\mathcal{H}_k$ includes delta functions (Def. B.11) (Jorgensen & Tian, 2015). Kernels with discrete

masses are always universal, but not the other way around (Amin et al., 2023). If we use a kernel with discrete masses, the KSD-B is faithful (proof in Appx. B.3.4).

Proposition 5.1. *Assume the support of p is connected, i.e. $\mathrm{supp}(p)$ is a connected set in the graph with vertices S and edges M. If either (a) k is a vector field kernel with discrete masses or (b) k is a scalar field kernel with discrete masses on S and $E_q \sum_{YMX} T_{p,Y\to X} < \infty$, then $\text{KSD-B}_{p,k}(q) = 0$ if and only if $p = q$.*

KSD-B Detects Convergence and Non-convergence Now, rather than consider a fixed distribution q, we consider a sequence of distributions $q_1, q_2, \ldots$. We are interested in showing the KSD-B can detect convergence and non-convergence of q_n to p, meaning $\text{KSD-B}_{p,k}(q_n) \to 0$ if and only if q_n converges to p in distribution, or in some closely related metric. This is useful for goodness of fit testing, because it says that if $\text{KSD-B}_{p,k}(q)$ is close to but not exactly zero, the difference between q and p cannot be very large, suggesting that we are unlikely to make a big mistake if our estimate of $\text{KSD-B}_{p,k}(q)$ is slightly off in practice. It is also useful for evaluating sample quality; in this case, q_n corresponds to the empirical distribution of n samples, and we are interested in whether the sample distribution converges to p (Gorham & Mackey, 2017). Another setting in which detecting convergence and non-convergence is important is when we are optimizing q to match p using $\text{KSD-B}_{p,k}(q)$ as the objective. In this case, we want lower values of $\text{KSD-B}_{p,k}(q_n)$ to correspond to distributions q_n closer to the target p.

We first show that the KSD-B detects non-convergence, i.e. $\text{KSD-B}_{p,k}(q_n) \to 0$ implies $q_n \to p$ in distribution. In the Euclidean case, KSDs based on diffusions can only detect non-convergence if the stochastic process converges quickly enough, which occurs if p is not heavy tailed (Gorham et al., 2019; Gorham & Mackey, 2017). In sequence space, the "tail" of a distribution describes how it depends on sequence length (Amin et al., 2021). For the KSD-B to detect non-convergence, we also need to control the tail of p; Prop. B.15 shows what can go wrong otherwise. We assume that p has uniformly quickly decreasing tails (Asm. B.3). By constructing a Lyapunov function for stochastic processes on sequences, we show this implies sufficiently fast convergence of the stochastic process $\mathcal{L}_p$ (Thm. B.4). Our tail assumption is reasonable for many generative models of sequences trained on real data sets; for example, we prove that profile HMMs, a widely used model of protein domains, always satisfy the assumption, regardless of the data they are trained on (for $\chi(t) = t \wedge 1$; Sec. B.2.5). Note the second term in Eqn. 2 is crucial for ensuring our tail assumption implies fast convergence of $\mathcal{L}_p$.

Another important condition needed in the Euclidean case is that the kernel k is heavy tailed (Gorham & Mackey, 2017). In Props. B.16 and B.17 we show that the KSD-B, too, may fail to detect non-convergence if we allow k to have thin tails. The crucial issue is that to detect non-convergence of distributions q_n that become more and more spread out as $n \to \infty$, there has to exist a test function $\mathcal{T}_p g$, for $g \in \mathcal{H}_k$, that has thick tails. To guarantee this, we assume k has heavy tails (Appx. B.4.3).

Theorem 5.2. *Say p is a distribution on S obeying Asm. B.3 and k is either a vector field kernel with discrete masses obeying Asm. B.18A or a scalar field kernel with discrete masses obeying Asm. B.18B. Say $(q_n)_n$ is a sequence of distributions on S. If $\text{KSD-B}_{p,k}(q_n) \to 0$ then q_n converges to p in distribution.*

Proof in Appx. B.3.5. Next, we show that the KSD-B detects convergence, i.e. if $q_n \to p$ then $\text{KSD-B}_{p,k}(q_n) \to 0$ (proof in Appx. B.3.5). In the Euclidean case, the KSD is guaranteed to detect convergence in a reweighted Wasserstein metric (Gorham & Mackey, 2017). We show that the KSD-B detects convergence in a reweighted total variation metric.

Proposition 5.3. *Say k is a vector field kernel and $p, q_1, q_2, \ldots$ are p, k-integrable distributions on S. Call $A(X) = \sum_{YMX} T_{p,Y\to X} \sqrt{k((X,Y),(X,Y))}$. If $\sum_X |p(X) - q_n(X)| A(X) \to 0$ then $\text{KSD-B}_{p,k}(q_n) \to 0$.*

Scalar Field Case If we use a scalar field kernel, rather than a vector field kernel, we can still guarantee detection of non-convergence, but only if we assume that the kernel is unbounded (Asm. B.18B), implying that $k(X, X)$ grows arbitrarily large as the length of X increases (Prop. B.16). This in turn implies that the discrepancy will consider longer sequences arbitrarily more important than shorter sequences when judging the similarity between p and q. Biologically, this judgment rarely makes sense. It is also contrary to the common practice of normalizing kernels so that $k(X, X) = 1$ for all $X \in S$ (Saigo et al., 2004).

Scalar field discrepancies can also detect convergence, but with an unbounded kernel, they cannot do so very well. In particular, Prop. 5.3 still holds (since scalar field kernels are a special case of vector field kernels), but $A(X)$ is now unbounded. It is thus harder to achieve $\sum_X |p(X) - q_n(X)| A(X) \to 0$, and so there are fewer cases in which the discrepancy can detect convergence.

6. Approximating the KSD-B

In this section we develop an efficient stochastic approximation to the KSD-B, improving its ability to scale to long sequences. Our approach centers on reducing the cost of evaluating each of the terms,

$$\sum_{XMY, X'MY'} T_{p,X\to Y} T_{p,X'\to Y'} k((X,Y),(X',Y')), \quad (5)$$

which appear inside the expectation in Eqn. 4. Evaluating these terms exactly is expensive for longer sequences, since the number of possible single mutations of X, that is $|\{Y \in S \,|\, YMX\}|$, scales linearly with sequence length $|X|$.

Baum et al. (2022) study a discrepancy that is essentially, in our terminology, a KSD-B with a scalar field kernel. They propose to reduce the computational cost of Eqn. 5 by reducing the size of the neighborhood around each sequence, shrinking M to $M_{(\tau)}$ where τ is a parameter controlling the size of the graph. For example, they have $YM_{(\tau)}X$ only if Y differs from X by a single substitution, with only substitutions within a distance τ allowed; distance between letters in $\mathcal{B}$ is measured by assigning each letter to an integer between 1 and $|\mathcal{B}|$, then computing their difference modulo $|\mathcal{B}|$. This approach has weaknesses in that (a) removing connections between sequences of different lengths results in a discrepancy that is no longer faithful, and (b) biologically, there is no canonical ordering of amino acids or nucleotides.

Instead of shrinking neighborhoods deterministically, we approximate Eqn. 5 stochastically, by sampling mutants of X. We do so by taking a single step of a discrete time Markov process initialized at X, with transition matrix K, where $K_{X\to Y} = T_{p,X\to Y}/\mathrm{flux}_p(X)$ if $X \neq Y$ and $K_{X\to Y} = 0$ if $X = Y$ (Appx. B.2).

Proposition 6.1. *Let p be a distribution on S, and $(q_n)_n$ a sequence of distributions on S with $\sup_n E_{q_n}\mathrm{flux}_p < \infty$. Say k is a bounded vector field kernel. Let $(N_{n,X})_{n,X\in S}$ be a set of numbers. For each X, n, let $(Y^n_{X,m})_{m=1}^{N_{n,X}}$ be a set of iid samples, each drawn by taking a single step of a Markov chain with the transition matrix $K_{X\to Y}$ initialized at X. Define the approximation $\widehat{\text{KSD-B}}^n_{p,k}(q_n)^2$, as*

$$E_{X,X'\sim q}\big[\mathrm{flux}_p(X)\mathrm{flux}_p(X')\times \frac{1}{N_{n,X}N_{n,X'}}\sum_{m,m'} k((X,Y^n_{X,m}),(X',Y^n_{X',m'}))\big].$$

If $N_{n,X}/(\log(n)+|X|)\to\infty$ then almost surely

$$\left|\text{KSD-B}_{p,k}(q_n) - \widehat{\text{KSD-B}}^n_{p,k}(q_n)\right| \to 0.$$

The proof is in Appx. B.3.7 and is roughly based on the use of a sub-Gaussian concentration inequality as in Thm. 4 of Gorham et al. (2020). The result shows that we can accurately approximate the KSD-B by sub-sampling mutants. Note it requires that the kernel is bounded, which is impossible for scalar field kernels that detect non-convergence. Thus, a further advantage of vector field over scalar field kernels is access to a good approximation of the KSD-B.

In summary, there are two computationally intensive steps in approximating the KSD-B. Given a fixed N_n and maximum sequence length L, we need to (1) calculate the likelihood under p of all mutational neighbours of all n sequences and (2) perform $(n\times N_n)^2$ evaluations of k. So, in principle, the computational cost scales as $O(n\times L\times[p \text{ scaling with } L] + n^2\times N_n^2\times[k \text{ scaling with } L])$. The second term usually dominates.

7. Kernels for the KSD-B

In this section we describe kernels for the KSD-B. The challenge is to construct kernels that simultaneously satisfy the requirements of our theoretical guarantees and capture biological notions of sequence similarity. Common approaches to measuring biological sequence similarity include (1) comparing sequences position by position (e.g. Hamming kernels), (2) comparing sequences based on pairwise alignments (e.g. alignment kernels), (3) comparing sequences based on their kmer content (e.g. kmer spectrum kernels), and (4) comparing sequences using learned embeddings into Euclidean space. Amin et al. (2023) develop kernels on S that use these biological notions of sequence similarity but are also highly flexible, having discrete masses.

To build kernels for the KSD-B, we first extend these scalar field kernels to vector fields, while preserving the discrete mass property. General techniques for doing so are developed in Proposition B.34; here we introduce two concrete examples, one based on Hamming distance and the other based on pairwise alignment distance. To define the kernels, we give their value for just one ordering of each pair of sequences related by M. More precisely, let σ be a function that takes every pair of sequences (X,Y) satisfying XMY to $\{-1,1\}$, with the restriction that $\sigma(X,Y) = -\sigma(Y,X)$ and $\sigma(X,Y) = 1$ if $|X| > |Y|$. Define $M^\sigma = \{(X,Y)\in M \mid \sigma(X,Y) = 1\}$. In Prop. B.33 we show that any kernel k defined on M^σ can be uniquely extended to a vector field kernel on M by applying the anticommutivity property (Eqn. 3). Now, let $d_H(X,Y)$ be the Hamming distance between X and Y. The exponential Hamming kernel is $\exp(-\lambda d_H(X,Y))$ with $\lambda > 0$; it has discrete masses by Thm. 21 of Amin et al. (2023). Now, the exponential Hamming vector field kernel (**Exp-H**) is,

$$(\exp(-\lambda d_H(X,X')) + \exp(-\lambda d_H(Y,Y')))^2,$$

for $(X,Y),(X',Y')\in M^\sigma$. Similarly, if k_{ali} is an alignment kernel with discrete masses (Amin et al., 2023, Thm. 23), we can construct the vector field alignment kernel (**Ali**),

$$\left(r^{|X|}k_{\text{ali}}(X,X')r^{|X'|} + r^{|Y|}k_{\text{ali}}(Y,Y')r^{|Y'|}\right)^2$$

for $r > 0$ sufficiently small. The **Ali** and **Exp-H** kernels have discrete masses, guaranteeing the KSD-B will be faithful (Prop. 5.1).

Discrete masses are not sufficient, however, to guarantee the kernel can be used to detect non-convergence; for this we need kernels that are, roughly speaking, heavy tailed. We consider kernels $k((X,Y),(X',Y'))$ of the form,

$$\breve{k}(Y,Y')\mathbb{1}(|X|\geq|Y|)\mathbb{1}(|X'|\geq|Y'|)$$

for $(X,Y),(X',Y')\in M^\sigma$. Setting $\breve{k}(Y,Y') = (C + d_H(Y,Y'))^{-\beta}$ gives an inverse multiquadric Hamming vector field kernel (**IMQ-H**). It is heavy tailed in the sense that

it decays as a power law of Hamming distance, rather than exponentially. We also consider setting $\breve{k}(X, X')$ to

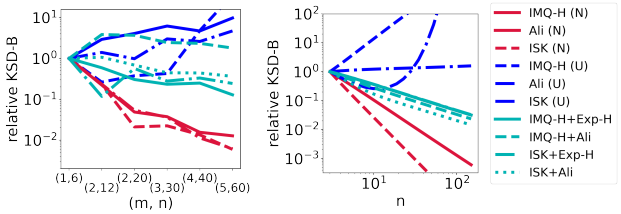
$$|X|^{-3/2}\left(\sum_{V\in S}\#\{V \text{ in } X\}\#\{V \text{ in } X'\}\right)|X'|^{-3/2},$$

where $\#\{V \text{ in } X\}$ is the number of occurrences of the substring (kmer) V in X. This gives an infinite kmer spectrum vector field kernel (**ISK**), which decays as a power law of sequence length.

By adding together a kernel with discrete mass k_δ and one with heavy tails k_{HT} we can build a kernel $k_\delta + k_{\mathrm{HT}}$ that meets the requirements of Thm. 5.2. In particular, Prop. B.39 shows that if p is a pHMM and $\chi(t) = \min\{t, 1\}$, we can add either of the discrete mass kernels (**Exp-H** or **Ali**) to either of the heavy tail kernels (**IMQ-H** or **ISK**) and satisfy all the assumptions of Thm. 5.2.

Embedding Kernels Another approach to constructing kernels for biological sequences is to leverage representations. Consider in general an embedding function $F : S \to \mathbb{R}^D$ that maps sequences to a low-dimensional Euclidean space. For example, F may come from a deep generative model trained on a large data set of biological sequences; we use UniRep64 (Alley et al., 2019). We can apply a Euclidean kernel k_E to the embedding space to build a scalar field kernel: for $X, Y \in S$, $k_{F,\mathrm{Emb}}(X, Y) = k_E(F(X), F(Y))$ (Yang et al., 2018b; Amin et al., 2021). This approach allows for learned, rather than hand-crafted, notions of sequence similarity. It also allows for fast evaluation of the KSD-B. The cost of using the Hamming kernel scales as $O(n^2 \times N_n^2 \times L)$, and that of the alignment kernel as $O(n^2 \times N_n^2 \times L^2)$. With an embedding kernel, using an F defined by an autoregressive sequence model, we can (1) embed all sequences and their mutants in $O(n \times N_n \times L)$ time (since the cost of evaluating F scales linearly with L), and then (2) calculate the kernel k_E applied to the embeddings in $O(n^2 \times N_n^2)$ time, as the embedding space is independent of L. Though embedding kernels are less theoretically tractable than the other kernels we have considered, our theory can nonetheless help guide their design. First, we extend scalar field embedding kernels to vector fields. Then, we construct an embedding kernel that is likely to have discrete masses by rescaling F, following the recommendation of Amin et al. (2023). We then add to it an embedding kernel that uses a heavy tailed Euclidean kernel (Appx. C.10).

Scalar Field Kernels We will compare our vector field kernels to scalar field alternatives. We develop unbounded versions of the inverse multiquadric Hamming kernel, **IMQ-H (U)**, alignment kernel, **Ali (U)**, and infinite kmer spectrum kernel, **ISK (U)**, which have discrete masses and are guaranteed to detect non-convergence (Prop. B.38). We also consider scalar field embedding kernels that are likely to have discrete masses.



(a) $q_{m,n} \not\to p$ (b) $q_n \to p$

Figure 1. **Detecting Convergence and Non-convergence** Vector field kernels are shown in turquoise, unbounded scalar field kernels in blue, and bounded scalar field kernels in red. The y-axis gives the KSD-B, normalized to its value at $q_{1,6}$ or q_3.

8. Empirical Results

In this section we examine the empirical performance of the KSD-B on simulated and real data. Details are in Appx. C.

Detecting Convergence and Non-convergence We first illustrate our theoretical results on detecting convergence and non-convergence. To start, we consider a simple example model with a single letter in the alphabet, $|\mathcal{B}| = 1$, and $p(X) \propto e^{-|X|}$. We consider a sequence of distributions defined by $q_{m,n}(X) \propto |X|^{-1}\mathbb{1}(m \leq |X| < n)$. As $m, n \to \infty$, $q_{m,n}$ does not converge to p. In line with our theoretical results (Thm. 5.2), the KSD-B with vector field kernels does not converge to zero, nor does the KSD-B with unbounded scalar field kernels (Fig. 1(a)). However, if we use bounded versions of the scalar field kernels, normalized to have $k(X, X) = 1$ for all X (**IMQ-H (N)**, **Ali (N)**, and **ISK (N)**), we find that the KSD-B fails to detect non-convergence (Prop. B.16).

Next, consider the heavy tailed distribution $p(X) \propto |X|^{-1.4}$, and $q_n(X) \propto |X|^{-1.4}\mathbb{1}(|X| \leq n)$. Now, q_n converges to p as $n \to \infty$. In line with our theoretical results (Prop. 5.3), the KSD-B with vector field kernels converges to zero, as does the KSD-B with bounded scalar field kernels (Fig. 1(b)). However, with an unbounded scalar field kernel, it does not. In short, vector field Stein discrepancies enable detection of both convergence and non-convergence, while bounded scalar field discrepancies cannot reliably detect non-convergence, and unbounded scalar field discrepancies cannot reliably detect convergence.

Goodness of Fit Testing In this section we evaluate the ability of the KSD-B to detect mismatches between models and data. We start with a generative biological sequence model p, then perturb its parameters to form q, and draw samples; we then evaluate the goodness of fit of p on the samples from q, for differing perturbation strengths. We are interested in how well the KSD-B can detect small perturbations. To construct a hypothesis test, we bootstrap the KSD-B as in Liu et al. (2016), and set a significance threshold of 0.1. In the following examples, we use the DNA alphabet and $N_n = 20$ samples in the KSD-B approximation. We draw 100 samples from q to form each data set,

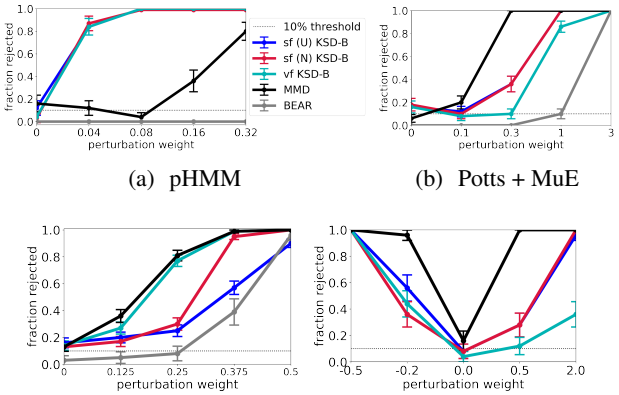
and report the mean and standard error of the test's rejection rate (power) across independent samples of the entire data set. We compare tests based on a vector field kernel, IMQ-H+Exp-H (labelled **vf KSD-B** in Fig. 2), unbounded scalar field kernel, IMQ-H (U) (**sf (U) KSD-B**), and normalized scalar field kernel, IMQ-H (N) (**sf (N) KSD-B**). Where possible, we also compare to an MMD two-sample test with the IMQ-H (N) kernel (**MMD**), which requires samples from p (Gretton et al., 2012; Amin et al., 2023), and a nonparametric Bayesian goodness of fit test (**BEAR**), which requires normalized likelihoods for p (Amin et al., 2021). In principle, the KSD-B test is most powerful when the "slopes" of p and q differ (Eqn. 1), while the MMD and BEAR tests focus on differences in the likelihood of p and q. We expect the KSD-B and MMD tests to be powerful when the differences between p and q lead to large changes in the chosen kernel, while we expect the BEAR test to be powerful when the differences between p and q lead to large changes in the parameters of an autoregressive model fit to each distribution.[1]

First, we consider testing profile hidden Markov models (pHMMs). We let p be a pHMM with latent sequence length 20, and with a high probability of generating the letter C at position 5. We then perturb the model by decreasing the probability of the C at position 5. The KSD-B test should easily detect the presence of sequences which are one mutation away from a much more likely sequence. We see the KSD-B indeed has high power to detect this perturbation, as compared to the MMD and BEAR tests (Fig. 2(a)).

Next we consider a Potts model with a mutational emission (MuE) distribution (Marks et al., 2011; Weinstein & Marks, 2021). Potts models are often used for evolutionary protein families, and have been applied to design novel proteins and predict 3D structure and mutational effects. The MuE adds insertions and deletions to samples from the Potts model. We set the Potts model length to 15. We start with no interaction energies between sites in the Potts model, and perturb by adding stronger and stronger interactions. Here, we find the MMD test outperforms the KSD-B test, which in turn outperforms the BEAR test (Fig. 2(b)).

Next we examine an autoregressive model; such models have been used, for example, to design novel proteins (Shin et al., 2021; Amin et al., 2021). We set p to be a linear autoregressive model of lag 2 which generates sequences that range in length from 1 to 60. We perturb p by adding a nonlinear term. Here, we find that the vector field KSD-B test and the MMD test are the most powerful (Fig. 2(c)).

Finally, we consider an ancestral sequence reconstruction model; such models have been used, for example, to resurrect ancient proteins (Pillai et al., 2020). We consider a star-shaped phylogeny, in which an ancestral sequence X is drawn from a pHMM prior, $\pi(X)$, and descendants Y_i are drawn by evolving X for time t according to a stochastic mutational process $\kappa(Y \mid X, t)$, parameterized by a MuE distribution. We are interested in the posterior over ancestors, $p(X \mid Y_1, \ldots, Y_5) \propto \pi(X) \prod_{i=1}^{5} \kappa(Y_i \mid X, t)$. We set p to the posterior with $t = 1$, then perturb p by modifying t. In this example, the normalizing constant of p is unavailable, and sampling from p requires approximation methods. To have a point of comparison, we applied the MMD to samples from p drawn by a long run of an efficient discrete MCMC method (Sun et al., 2022). We find the KSD-B can accurately detect model-data mismatches, without samples or normalized likelihoods (Fig. 2(d)).



(a) pHMM (b) Potts + MuE

(c) autoregressive (d) ancestral reconstruction

Figure 2. **Goodness of Fit Testing** We evaluate the power of goodness of fit tests, based on the KSD-B and alternatives, to reject the null hypothesis that $p = q$. In each plot, the x-axis corresponds to different values of q, which come from perturbing p by different amounts. We also plot the 10% significance threshold.

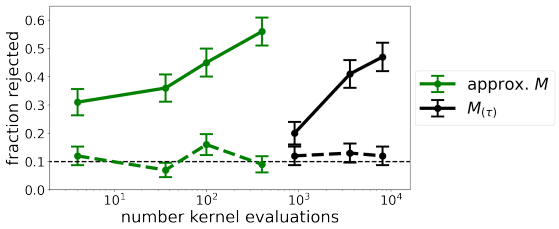


Figure 3. **Approximating the KSD-B** We compare the power of a goodness of fit test using our stochastic approximation (green) to one using neighborhood reduction (black). We set $\tau \in \{1, 2, 3\}$ for the neighborhood size and $N_n \in \{2, 4, 10, 20\}$ for the number of mutation samples. The solid lines are for tests performed on samples from a perturbed distribution and the dotted lines are for tests performed on samples from p. In the latter case, the test matches the 10% significance threshold, showing good calibration.

Approximating the KSD-B Next we evaluate our stochastic KSD-B approximation strategy, and compare its efficiency to the reduced neighborhood approach of Baum et al. (2022) (Sec. 6). We set p to a pHMM, with $\mathcal{B}$ the amino

[1]Note also the computational cost of the KSD-B and MMD tests scales quadratically in the number of datapoints, while that of the BEAR test scales roughly linearly; hence the BEAR test is typically more tractable for massive data sets (Amin et al., 2021).

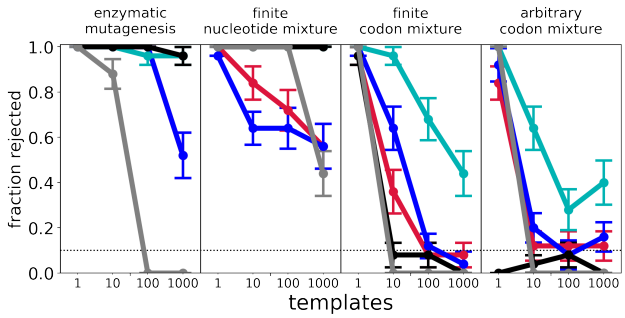


Figure 4. **Evaluating Variational Synthesis Models** Goodness of fit tests comparing a target model p to synthesis models q. Each subplot is a different DNA synthesis technology. The x-axis shows the number of templates, which corresponds roughly to the complexity and cost of the synthesis procedure, with larger numbers of templates allowing better matches to p (Weinstein et al., 2022b). Legend is identical to Fig. 2

acids. The approach of Baum et al. (2022) requires assigning an ordering to the amino acids; we do so on the basis of hydrophobicity. We then consider perturbations of p that increase the probability of hydrophobic residues. Since the reduced graph $M_{(\tau)}$ does not have connections between strongly hydrophilic and strongly hydrophobic residues (even when those amino acids are similar in another respect, such as size), tests based on $M_{(\tau)}$ can struggle to detect this perturbation.

We compare the power of tests using reduced neighborhoods to those using stochastic sub-sampling, as a function of computational cost (solid lines, Fig. 3). Cost is measured by the total number of kernel evaluations required for each pair of sequences in the data set. Our approach yields an order-of-magnitude decrease in kernel evaluations while achieving the same power. It also maintains good calibration, as we confirm by evaluating the test on samples from the unperturbed model p (dashed lines).

Evaluating Synthesis Strategies Next we consider an application of the KSD-B to a specific library design problem where goodness of fit tests have been used previously, variational synthesis (Weinstein et al., 2022b). Here, the aim is to design a stochastic synthesis procedure which produces approximate samples from a target generative model in the laboratory at very large scale. As a target p, we consider a pHMM trained on a data set of human T cell receptor CDR3 sequences, which range in length from 10 to 27 amino acids (10x Genomics, 2022). We optimize synthesis models q based on different synthesis technologies: finite nucleotide mixtures, enzymatic mutagenesis, finite codon mixtures, and arbitrary codon mixtures. Previously, BEAR tests were used to evaluate the match between the synthesis model q and target p; here, we apply the KSD-B. We find that the KSD-B test can still detect mismatches even when the BEAR test cannot, and that vector field kernels outperform scalar field kernels and the MMD test.

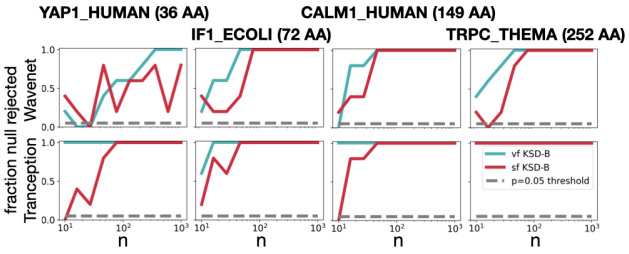


Figure 5. **Evaluating Large Models Fit to Protein Families** We perform a goodness of fit test for two deep generative models, Wavenet (top row) and Tranception (bottom row), using scalar field (turquoise) and vector field (red) embedding kernels. Each column is a different protein family dataset. We perform the test for 5 independent samples of the $N_n = 10$ mutants for four protein families and plot how often the null hypothesis was rejected at level 0.05 for increasing data n.

Evaluating Large Models Fit to Protein Families Finally, we use the KSD-B to evaluate the fit of state-of-the-art deep generative models of proteins. We considered data sets consisting of evolutionarily related protein families (Shin et al., 2021). We trained a deep autoregressive model on each data set (Wavenet; Shin et al. (2021)), and tested its goodness of fit on held-out sequences using the KSD-B. We also tested the goodness of fit of a transformer model trained on a data set of all known proteins (Tranception; Notin et al. (2022)); in Appx.C.10.1 we explain why the KSD-B is particularly suitable for evaluating such "protein universe" models. To scale the KSD-B to $n = 1000$ proteins of length roughly $L = 250$, we sample $N_n = 10$ mutants and use an embedding kernel. Despite the small N_n, we see little variation in our KSD-B estimates when we resample mutants, particularly as compared to the differences in the KSD-B between different models (Fig. 7).

The KSD-B is capable of detecting model-data mismatch for both Wavenet and Tranception, even when given fewer than 100 sequences for evaluation (Fig. 5). Moreover, the test's power does not fall on data sets with longer sequences (though this is not always true in other scenarios; see Fig. 6). In almost all cases, our vector-field KSD-B test is more powerful than a scalar-field KSD-B test (this holds even for different kernels; see Fig. 8).

9. Conclusion

In this paper we have developed the KSD-B, a novel discrepancy for distributions over biological sequences, with strong theoretical guarantees. One possible direction for future work is to further scale the KSD-B using methods for KSDs in Euclidean space (Jitkrittum et al., 2017; Huggins & Mackey, 2018; Gorham et al., 2020). Another is to apply the KSD-B to develop better samplers for biological sequences (Gorham & Mackey, 2017; Grathwohl et al., 2021). Overall, we hope the KSD-B can help ensure the accuracy, reliability, and trustworthiness of methods based on generative sequence models as they see growing use across biology, biotechnology and biomedicine.

References

10x Genomics. A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype. *10x Genomics*, 2022.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nat. Methods*, 16(12):1315–1322, December 2019.

Amin, A. N., Weinstein, E. N., and Marks, D. S. A generative nonparametric bayesian model for whole genomes. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Amin, A. N., Weinstein, E. N., and Marks, D. S. Biological sequence kernels with guaranteed flexibility. April 2023.

Barbour, A. D. Stein's method for diffusion approximations. *Probab. Theory Related Fields*, 84(3):297–322, 1990.

Barp, A., Briol, F. X., Duncan, A. B., Girolami, M., and Mackey, L. Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Baum, J., Kanagawa, H., and Gretton, A. A kernel stein test of goodness of fit for sequential models. October 2022.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. Stein points. In *International Conference on Machine Learning (ICML)*, 2018.

Chen, Y., Welling, M., and Smola, A. Super-samples from kernel herding. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2010.

Chow, S.-N., , Li, W., and Zhou, H. Entropy dissipation of Fokker-Planck equations on graphs. *Discrete Contin. Dyn. Syst. Ser. A*, 38(10):4929–4950, 2018.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pp. 2606–2615, June 2016.

Davidsen, K., Olson, B. J., DeWitt, 3rd, W. S., Feng, J., Harkins, E., Bradley, P., and Matsen, 4th, F. A. Deep generative models for T cell receptor protein sequences. *Elife*, 8, September 2019.

Douc, R., Fort, G., and Guillin, A. Subgeometric rates of convergence of f-ergodic strong markov processes. *Stochastic Process. Appl.*, 119(3):897–923, March 2009.

Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883):91–95, 2021.

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, 2017.

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928, October 2019.

Gorham, J., Raj, A., and Mackey, L. Stochastic Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Grathwohl, W., Swersky, K., Hashemi, M., Duvenaud, D., and Maddison, C. J. Oops I took a gradient: Scalable sampling for discrete distributions. 2021.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.

Hairer, M. Convergence of markov processes, 2021.

Han, J., Ding, F., Liu, X., Torresani, L., Peng, J., and Liu, Q. Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Haussler, D. Convolution kernels on discrete structures UCSC CRL. 1999.

Hodgkinson, L., Salomone, R., and Roosta, F. The reproducing stein kernel approach for post-hoc corrected sampling. January 2020.

Hopf, T. A., Ingraham, J. B., Poelwijk, F. J., Schärfe, C. P. I., Springer, M., Sander, C., and Marks, D. S. Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.*, 35(2):128–135, 2017.

Huggins, J. H. and Mackey, L. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 261–270, Red Hook, NY, USA, December 2017. Curran Associates Inc.

Jorgensen, P. and Tian, F. Discrete reproducing kernel hilbert spaces: Sampling and distribution of dirac-masses. *Journal of Machine Learning Research*, 2015.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. Mismatch string kernels for discriminative protein classification. 20(4):467–476, 2004.

Liggett, T. M. *Continuous time Markov processes: An introduction*. Graduate studies in mathematics. American Mathematical Society, Providence, RI, March 2010.

Liu, Q., Lee, J. D., and Jordan, M. A kernelized Stein discrepancy for goodness-of-fit tests. *International Conference on Machine Learning (ICML)*, 2016.

Madani, A., McCann, B., Naik, N., Keskar, N. S., Anand, N., Eguchi, R. R., Huang, P.-S., and Socher, R. ProGen: Language modeling for protein generation. March 2020.

Marcou, Q., Mora, T., and Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.*, 9 (1):561, February 2018.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. Protein 3D structure computed from evolutionary sequence variation. *PLoS One*, 6(12):e28766, December 2011.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. July 2021.

Nijkamp, E., Ruffolo, J., Weinstein, E. N., Naik, N., and Madani, A. ProGen2: Exploring the boundaries of protein language models. June 2022.

Notin, P., Dias, M., Frazer, J., Marchena-Hurtado, J., Gomez, A., Marks, D. S., and Gal, Y. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. May 2022.

Pillai, A. S., Chandler, S. A., Liu, Y., Signore, A. V., Cortez-Romero, C. R., Benesch, J. L. P., Laganowsky, A., Storz, J. F., Hochberg, G. K. A., and Thornton, J. W. Origin of complexity in haemoglobin evolution. *Nature*, 2020.

Riesselman, A. J., Ingraham, J. B., and Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*, 15(10):816–822, 2018.

Russ, W. P., Figliuzzi, M., Stocker, C., Barrat-Charlaix, P., Socolich, M., Kast, P., Hilvert, D., Monasson, R., Cocco, S., Weigt, M., and Ranganathan, R. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, July 2004.

Shi, J., Zhou, Y., Hwang, J., Titsias, M. K., and Mackey, L. Gradient estimation with discrete stein operators. *Advances in Neural Information Processing Systems (NeurIPS)*, February 2022.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. Protein design and variant prediction using autoregressive generative models. *Nat. Commun.*, 12(1): 2403, April 2021.

Sperling, A. K. and Li, R. W. Repetitive sequences. In Maloy, S. and Hughes, K. (eds.), *Brenner's Encyclopedia of Genetics (Second Edition)*, pp. 150–154. Academic Press, San Diego, January 2013.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11(50):1517–1561, 2010.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. G. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12(70):2389–2410, 2011.

Sun, H., Dai, H., Xia, W., and Ramamurthy, A. Path auxiliary proposal for MCMC in discrete space. *International Conference on Learning Representations (ICLR)*, 2022.

Thadani, N. N., Gurev, S., Notin, P., Youssef, N., Rollins, N. J., Sander, C., Gal, Y., and Marks, D. S. Learning from pre-pandemic data to forecast viral antibody escape. July 2022.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. 2020.

Weinstein, E. N. *Generative Statistical Methods for Biological Sequences*. PhD thesis, Harvard University, Ann Arbor, United States, 2022.

Weinstein, E. N. and Marks, D. A structured observation distribution for generative biological sequence prediction and forecasting. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11068–11079. PMLR, 2021.

Weinstein, E. N., Amin, A. N., Frazer, J., and Marks, D. S. Non-identifiability and the blessings of misspecification in models of molecular fitness and phylogeny. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022a.

Weinstein, E. N., Amin, A. N., Grathwohl, W., Kassler, D., Disset, J., and Marks, D. S. Optimal design of stochastic DNA synthesis protocols based on generative sequence

models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2022b.

Yang, J., Liu, Q., Rao, V., and Neville, J. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning (ICML)*, 2018a.

Yang, K. K., Wu, Z., Bedbrook, C. N., and Arnold, F. H. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, August 2018b.

Zanella, G. Informed proposals for local MCMC in discrete spaces. *J. Am. Stat. Assoc.*, 115(530):852–865, 2020.

A. Broader Impact Statement

The present work has the potential to impact a variety of procedures in biotechnology and health. Through its use in model criticism and sequence design, the KSD-B could facilitate the design of novel therapeutics. Biological sequence design may, however, also be used in applications with negative societal impact. The KSD-B may also be used to critique generative sequence models used for diagnosis and disease discovery. This could lead to more reliable models, which could lead to more accurate patient diagnoses, and a deeper understanding of the genetic underpinnings of disease. Research in this direction, however, also has the potential to exacerbate health outcome disparities that affect marginalized groups, and to do so based on the genetics of such groups.

B. Proofs

In this appendix we prove the assertions in the main text. First in Section B.1 we lay out our notation. Next, in Section B.2 we study stochastic processes on sequence space, and perform a Lyapunov function analysis of their convergence rates. In Section B.3 we show that the KSD-B is faithful, can detect convergence and non-convergence, and can be efficiently approximated. Finally, in Section B.4, we develop kernels that satisfy our theoretical requirements for detecting convergence and non-convergence.

B.1. Notation

Let our alphabet, $\mathcal{B}$, be a finite set, and let the set of all sequences be defined as $S = \cup_{L=0}^{\infty} \mathcal{B}^L$ where $\mathcal{B}^0$ is defined to only contain the empty string $\emptyset$. If p is a distribution on S let $\mathrm{supp}(p) = \{X \mid p(X) > 0\}$ and $M_{p,p} = \{(X,Y) \in M \mid X, Y \in \mathrm{supp}(p)\}$. We will say p has connected support if $\mathrm{supp}(p)$ is a connected set in the graph with vertices S and edges M. Finally, for $X \in S$, define $\mathrm{flux}_p(X) = \sum_{YMX} T_{p,X\to Y}$.

Let $C_b(S)$ be the set of bounded functions on S, let C_0 be the set of functions on S vanishing at infinity, and let $C_C(S)$ be the set of functions on S that are non-zero at only finitely many points. We also define the set of all vector fields that are non-zero on only finitely many points in M as $C_{C,vf}(M)$. We define $\|\cdot\|_\infty$ as the infinity norm on $C_b(S)$. For two distributions μ, ν on S, call $\|\nu - \mu\|_{\mathrm{TV}}$ their distance in total variation.

For two sequences of real numbers $(a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}}$, both possibly undefined for small n, we write $a_n \lesssim b_n$ to mean that there is a positive constant C such that eventually $a_n \le C b_n$. We write $a_n \sim b_n$ when $a_n \lesssim b_n$ and $a_n \gtrsim b_n$. We write $a_n = O(b_n)$ if $a_n \lesssim b_n$ and $a_n = o(b_n)$ if $\frac{|a_n|}{|b_n|} \to 0$. We define $a \wedge b$ as the minimum of a and b, and $a \vee b$ as the maximum. Define $\mathbb{1}(P)$ to be the indicator function that is 1 if P is true and 0 otherwise.

A kernel on a set H is a symmetric function $k : H \times H \to \mathbb{R}$ that is "non-negative definite", i.e. for all $X_1, \ldots, X_N \in H$, $\alpha_1, \ldots, \alpha_N \in \mathbb{R}$, $\sum_{n=1}^N \sum_{n'=1}^N \alpha_n \alpha_{n'} k(X_n, X_{n'}) \ge 0$. We also require that $k(X,X) > 0$ for all $X \in S$. For every $X \in S$ define the function $k_X = k(X, \cdot)$. Define the dot product $(\cdot|\cdot)_k$ on linear combinations of these functions with $(k_X|k_Y) = k(X,Y)$ and call the associated norm $\|\cdot\|_k$. Let $\mathcal{H}_k$ be the Hilbert space completion of the span of $\{k_X\}_{X\in H}$ under $(\cdot|\cdot)_k$ and call this the reproducing kernel Hilbert space (RKHS) of k. Elements of the RKHS can be understood as functions on H by $(f|k_X) = f(X)$.

Say k is a kernel on a space H and $A : H \to (0,\infty)$. We call $k^A(X,Y) = A(X)k(X,Y)A(Y)$ the kernel k "tilted" by A. k^A is a kernel on H and it is a well known fact that the transformation that takes $g \in \mathcal{H}_k$ to $X \mapsto g(X)A(X)$ is a unitary isomorphism to $\mathcal{H}_{k^A}$ (see for example Proposition 35 of Amin et al. (2023)).

We let χ be some non-zero, non-negative, and non-decreasing function on the non-negative real numbers $[0,\infty)$ such that $\chi(t) = t\chi(1/t)$ for all $t > 0$.

B.2. Stochastic Processes on Sequences

In this section we study stochastic processes on sequences. Our aim is to understand the convergence rate of continuous-time Markov processes. These results will be essential in proving the KSD-B detects non-convergence. They are also of much wider relevance, as Markov processes over sequences appear in many other contexts, including (a) mathematical models of evolution and (b) Markov chain Monte Carlo (MCMC) methods for sampling sequences, such as in the context of ancestral sequence reconstruction or conditional generation. We leave detailed exploration of these applications, including extensions to discrete-time Markov processes, to future work.

B.2.1. CONTINUOUS-TIME MARKOV PROCESSES

We study the continuous time Markov process on sequence space S defined by the transition rates $T_{p,X\to Y}$. Here, $T_{p,X\to Y}$ can be the transition probability of any continuous time Markov process that satisfies detailed balance, and p is a distribution on sequence space S. The operator $\mathcal{L}_p$ of the stochastic process is $\mathcal{L}_p = \mathcal{T}_p\nabla$. It acts on functions; if δ_Y is a delta function at Y, we have $\mathcal{L}_p(\delta_Y)(X) = T_{p,X\to Y}$. Notationally, to avoid switching back and forth between $\mathcal{L}_p$ and T, we will define $\mathcal{L}_{p,X,Y} = T_{p,X\to Y}$.

To simulate from a continuous time Markov process, one can first sample the sequence of distinct states the Markov process visits, and then sample how long the process stays at each state. In particular, let $K_{X\to Y} = \mathcal{L}_{p,X,Y}/\mathrm{flux}_p(X)$ if $X \neq Y$ and 0 if $X = Y$. The entries of K are positive and its rows sum to 1 so it defines a discrete-time stochastic process $(Z_0, Z_1, \dots)$, known as the "underlying stochastic process". The continuous-time process stays at each state for time $\tau_n \sim \mathrm{Exp}(\mathrm{flux}_p(Z_n))$. Thus, the continuous-time process $(X_t)_t$ with operator $\mathcal{L}_p$ is defined as $X_t = Z_n$ for $\tau_{n-1} \leq t < \tau_n$ and all n.

If we start at X_0 and run the continuous-time Markov process forward for time t, we obtain a distribution over S, denoted $P_t(X_0)$ (note $P_t(X_0)$ is a distribution; its value at $x \in S$ is denoted $P_t(X_0)(x)$). For any $f \in C_C(S)$, we define $P_tf(X)$ to be its expectation under $P_t(X)$, that is $P_tf(X) = E_{P_t(X)}f$. Note $P_tf(X)$ is continuously differentiable in t and satisfies the backwards Kolmogorov equation, i.e. $\frac{d}{dt}P_tf(X) = \mathcal{L}_pP_tf(X)$ (see section 2.5 of Liggett (2010)). If we sample the starting position X_0 from a distribution p on S, and this distribution does not change under the stochastic process – in the sense that $E_pP_tf = E_pf$ for all $f \in C_C(S)$ – then we call p stationary. We use the notation T_p and $\mathcal{L}_p$ to emphasize that the stationary distribution of the stochastic process they define is, by construction, p.

B.2.2. EXISTENCE AND OTHER USEFUL PROPERTIES

Sequence space S is infinite. Whenever the state space of a continuous time Markov process is infinite, the process may not, for any given $\mathcal{L}_p$, exist. Fundamentally, this is because $(X_t)_t$ can "explode" by transitioning infinitely many times over some finite time period. This can result in a situation where P_t is not an actual distribution (that is, $\sum_Y P_t(X)(Y) < 1$) or the forward Kolmogorov equation no longer holds (that is, $\frac{d}{dt}P_tf(X) \neq P_t\mathcal{L}_pf(X)$). To avoid these pathologies, we add an integrability condition on p, namely that $E_p\mathrm{flux}_p < \infty$. The below lemma shows that in this case the $(P_t)_t$ are valid probability distributions and the forward Kolmogorov equation holds. We also list some additional consequences that will help prove future results.

Lemma B.1. *Say p has connected support and $E_p\mathrm{flux}_p < \infty$.*

(A) There is a Markov process $(X_t)_t$ on $\mathrm{supp}(p)$ such that for all $f \in C_C(S)$, $P_tf(X)$ is continuously differentiable in t and $\frac{d}{dt}P_tf(X) = \mathcal{L}_pP_tf(X) = P_t\mathcal{L}_pf(X)$.

(B) p is stationary under P_t for all t. If q is another distribution with $E_q\mathrm{flux}_p < \infty$, then $q = p$ if and only if $E_q\mathcal{L}_pf = 0$ for all $f \in C_C(S)$.

(C) If $f \in C_C(S)$, $f(X_t) - \int_0^t \mathcal{L}_pf(X_s)ds$ is a martingale in t conditional on $X_0 = X$ for every $X \in \mathrm{supp}(p)$.

Proof. Take K, $(Z_n)_n$, $(\tau_n)_n$, $(X_t)_t$ and P_t defined as above. $(Z_n)_n$ is an irreducible Markov chain by definition as $\mathrm{supp}(p)$ is connected. To show that the P_t indeed define probability distributions, note that $\nu = \mathrm{flux}_pp$ is a finite measure on S that is stationary with respect to K since $\mathrm{flux}_p(X)p(X)K_{X\to Y} = \mathrm{flux}_p(Y)p(Y)K_{Y\to X}$. This implies that $(Z_n)_n$ will visit each $X \in \mathrm{supp}(p)$ infinitely many times almost surely. To see this, assume $(Z_n)_n$, starting at some point, visits an $X \in \mathrm{supp}(p)$ only finitely many times with positive probability. Since $(Z_n)_n$ is irreducible, every time Z_n hits X there is a fixed chance that it never hits X again, so, almost surely, Z_n hits X only finitely many times. Let $\hat{\nu} = \nu(X)/\nu(S)$ so, since $\hat{\nu}$ is stationary for K,

$$\hat{\nu}(X) = \int d\hat{\nu}(Y)(K^m)_{Y\to X} = E_{Z_0\sim\hat{\nu}}\left[\mathbb{1}(Z_m = X)\right] \to 0$$

as $m \to \infty$ by dominated convergence, a contradiction. Thus, by Corollary 2.34 (b) of Liggett (2010), P_t are distributions on S and $\sum_n \tau_n = \infty$ almost surely, that is, $(X_t)_t$ is a well defined Markov process. We also have that $P_tf(X) = E[f(X_t)|X_0 = X]$ for all X, t.

For the second claim, first note $E_q\mathrm{flux}_p < \infty$ implies $\mathrm{supp}(q) \subseteq \mathrm{supp}(p)$ since if $X \notin \mathrm{supp}(X)$, $T_{p,X\to Y}$ is defined to be ∞. By equation 2.40 of Liggett (2010), if q is a distribution on S such that $E_q\mathrm{flux}_p(X) < \infty$, q is stationary for all

P_t if and only if $E_q \mathcal{L}_p \delta_X = 0$ for all $X \in \text{supp}(p)$. In particular, p is stationary for all P_t. On the other hand, by our construction of $(X_t)_t$, since $\text{supp}(p)$ is connected, by Proposition 2.6 of Hairer (2021), each P_t has at most one stationary distribution for $t > 0$. Thus, $p = q$ if and only if $E_q \mathcal{L}_p f = 0$ for all $f \in C_C(S)$.

To show that we also have the forward Kolmogorov equation it suffices by Theorem 2.39 of Liggett (2010) to show that $P_t \text{flux}_p(X) < \infty$ for all t, X. To see this, note by the fact that p is stationary for P_t,

$$E_p \text{flux}_p \geq E_p \left(\text{flux}_p \mathbb{1}(|X| < N) \right)) = E_p P_t \left(\text{flux}_p \mathbb{1}(|X| < N) \right) \to E_p P_t \text{flux}_p$$

as $N \to \infty$ by monotone convergence so that $P_t \text{flux}_p(X) < \infty$ for all $t > 0, X \in \text{supp}(p)$.

The statement about martingales holds by the backwards and forwards Kolmogorov equations and Theorem 3.32 of Liggett (2010).

□

B.2.3. Establishing Convergence Rates

We are interested in studying the convergence of continuous time Markov processes on sequences. In other words, we would like to know whether the process defined by $\mathcal{L}_p$ approaches the stationary distribution p, and if so, how quickly. We can quantify closeness in terms of total variation distance, $\|P_t(X) - p\|_{\text{TV}}$. We are interested in how the total variation distance shrinks as a function of t. To investigate this question, we will use the following general theorem on Markov process convergence rates, which specializes Theorem 4.1 of Hairer (2021) to the infinite discrete space S. This theorem defines the convergence rate using a Lyapunov function, V.

Theorem B.2 (Theorem 4.1 of Hairer (2021))**.** *Say p has connected support and $E_p \text{flux}_p < \infty$. Say $V : S \to [1, \infty)$ is a function such that $V(X) \to \infty$ as $|X| \to \infty$. Assume $\mathcal{L}_p V \leq R - \varphi \circ V$ on $\text{supp}(p)$ for some number R and strictly concave $\varphi : [0, \infty) \to [0, \infty)$ with $\varphi(0) = 0$ and increasing to infinity. Now define $H(u) = \int_1^u \varphi(s)^{-1} ds$. Then there is a $C > 0$ such that for all $X \in \text{supp}(p)$,*

$$\|P_t(X) - p\|_{\text{TV}} \leq \frac{CV(X)}{H^{-1}(t)} + \frac{C}{\varphi \circ H^{-1}(t)}.$$

Proof. All conditions of Theorem 4.1 of Hairer (2021) are obviously satisfied except for the fact that $V(X_t) - \int_0^t ds \left(R - \varphi \circ V(X_t) \right)$ is a local super-martingale conditioned on $X_0 = X$ for some $X \in \text{supp}(p)$. This follows from Theorem 3.4 of Douc et al. (2009) if $Q_t = V(X_t) - \int_0^t \mathcal{L} V(X_s) ds$ defines a local martingale when $X_0 = X$ for all $X \in \text{supp}(p)$.

To show this, for every number N and $X \in S$, call $V^N(X) = V(X) \mathbb{1}(V(X) < N)$ so that $V_N \in C_C(S)$. Also define $T_N = \inf\{t \mid \exists Y s.t.\ Y M X_t \text{ and } V(Y) \geq N\}$. T_N is a stopping time and $T_N \to \infty$ almost surely. By Lemma B.1, $Q_t^N = V^N(X_t) - \int_0^t \mathcal{L} V^N(X_s) ds$ is a martingale conditioned on $X_0 = X$ for any $X \in \text{supp}(p)$ and, by the definition of T_N, $Q_t = Q_t^N$ for all $t \leq T_N$. Thus, $(Q_t)_t$ is a local martingale. □

B.2.4. Convergence Rates for Sequence Distributions

We now establish convergence rates for continuous time Markov processes on sequences. The fundamental challenge is to construct Lyapunov functions that are appropriate for biological sequences. In general, Lyapunov functions are constructed based on the tails of the stationary distribution p; the thinner the tails, the faster the convergence of the stochastic process. In Euclidean space, the tail of a probability distribution p refers to its value at large X. In sequence space, the tail refers to its value at *long* X (Amin et al., 2021). We will find that if p falls off quickly with sequence length, the stochastic process will be able to explore p rapidly, and so converge quickly; if p falls off slowly, convergence is slowed.

To describe the tails of probability distributions over sequences, we introduce the quantities,

$$\begin{aligned}\mathrm{del}_p(X) &= \sum_{|Y|=L-1, XMY} T_{p,X\to Y}\\ \mathrm{ins}_p(X) &= \sum_{|Y|=L+1, XMY} T_{p,X\to Y}\\ \mathrm{gap}_p(L) &= \inf_{X\in S \mid |X|=L} \mathrm{del}_p(X) - \mathrm{ins}_p(X).\end{aligned}$$

(If $X \notin \mathrm{supp}(p)$, take $\mathrm{del}_p(L) = \infty$, and $\mathrm{ins}_p(X) = 0$.) Now, $\mathrm{del}_p(X)$ describes the propensity to gain a deletion, $\mathrm{ins}_p(L)$ the propensity to gain an insertion, and $\mathrm{gap}_p(L)$ describes the difference between the two. The intuition is that $\mathrm{gap}_p(L)$ characterizes how much probability mass will move towards shorter sequences under the stochastic process. If p falls off quickly with L, then at long sequence lengths L, $\mathrm{gap}_p(L)$ will be big (for p to to be the stationary distribution, the stochastic process must be very likely to head back towards shorter sequences). Conversely, if p falls off slowly with L, then at long sequence lengths L, $\mathrm{gap}_p(L)$ will be small. We can thus use $\mathrm{gap}_p(L)$ to describe the tail of p.

We now translate our description of the tail of p into a Lyapunov function V_p, and from there into a convergence rate.

Assumption B.3. We assume p has connected support, $E_p \mathrm{flux}_p < \infty$, and there is some concave function $V_p : [0,\infty) \to [0,\infty)$ such that $\lim_{L\to\infty} V_p(L) = \infty$ and

$$\mathrm{gap}_p(L) \gtrsim \frac{V_p(L)^{\frac{1+\epsilon_V}{2+\epsilon_V}}}{V_p(L) - V_p(L-1)} \tag{6}$$

for some $\epsilon_V > 0$.

If a function V_p exists that satisfies this assumption, we can guarantee convergence. If V_p is small, we can guarantee fast convergence.

Theorem B.4. *Recall $P_t(X)$ is the distribution of a stochastic process with operator $\mathcal{L}_p = \mathcal{T}_p \nabla$, after being initialized at X and evolving for time t. Say the stationary distribution p obeys Assumption B.3. Then, the stochastic process converges to the stationary distribution in total variation. It does so with rate,*

$$\|P_t(X) - p\|_{\mathrm{TV}} \lesssim t^{-(1+\epsilon)} + V_p(|X|) t^{-(2+\epsilon)}.$$

Proof. Define $\Delta V_{p,L} = V_p(L) - V_p(L-1)$ and define $V_p(X)$ as $V_p(|X|)$. If $X \in \mathrm{supp}(p)$ with $|X| = L$,

$$\begin{aligned}\mathcal{L}_p V_p(X) =& \sum_{YMX, |Y|=|X|+1} T_{p,X\to Y} \Delta V_{p,L+1} - \sum_{YMX, |Y|=|X|-1} T_{p,X\to Y} \Delta V_{p,L}\\ =& \mathrm{ins}_p(X) \Delta V_{p,L+1} - \mathrm{del}_p(X) \Delta V_{p,L}\\ \leq& \mathrm{ins}_p(X) \left(\Delta V_{p,L+1} - \Delta V_{p,L}\right) - \mathrm{gap}_p(L) \Delta V_{p,L}\end{aligned}$$

Since V_p in concave, the first term is negative. As well, by Assumption B.3, $\mathrm{gap}_p(L)\Delta V_{p,L} \gtrsim \varphi(V_p(L-1))$ where $\varphi(x) = x^{(1+\epsilon)/(2+\epsilon)}$. Thus there are constants C_1, C_2 such that for all $X \in \mathrm{supp}(p)$,

$$\mathcal{L}_p V_p(X) \leq C_1 - C_2 \varphi \circ V_p(X).$$

By Theorem B.2, with $H = \int_1^u ds \varphi^{-1}(s) = C_3(u^{\frac{1}{2+\epsilon}} - 1)$, we have

$$\|P_t(X) - p\|_{\mathrm{TV}} \lesssim V_p(X) t^{-(2+\epsilon)} + t^{-(1+\epsilon)}.$$

□

Theorem B.4 tells us that the rate of convergence of the stochastic process depends on $V_p(|X|)$, the value of the Lyapunov function at the initialization point X. Larger values of $V_p(|X|)$ translate into a looser bound on the total variation, and thus slower convergence rates.

To understand the connection between the tail of p (as quantified by gap_p) and the convergence rate of the stochastic process (as quantified by V_p) in greater depth, we investigate Equation 6 further. We are interested in the smallest value of V_p that satisfies Equation 6 for a given value of gap_p; this tells us how fast a convergence rate we can guarantee. Note first that since V_p is concave and goes to ∞, the right hand side of Equation 6 is eventually less than

$$\frac{V_p(L)^{\frac{1+\epsilon_V}{2+\epsilon_V}}}{V_p'(L)} = \frac{V_p(L)}{V_p'(L)} V_p(L)^{-\frac{1}{2+\epsilon_V}} = \left((\log V_p)'(L) V_p^{\frac{1}{2+\epsilon_V}}(L) \right)^{-1}. \tag{7}$$

where V_p' is the derivative of V_p. This quantity is larger when V_p grows more slowly with L. Let's now consider three example choices of gap_p and V_p. First, consider V_p of the form $V_p(L) = L^\alpha$. Now, Equation 7 is proportional to $L^{1-\frac{\alpha}{2+\epsilon_V}}$. So, we can satisfy Equation 6 if $\text{gap}_p(L) \gtrsim L^\beta$ for some $\beta > 1/2$, if we choose $V_p = L^\alpha$ with $0 < \alpha \leq 1$. Alternatively, we can consider V_p of the form $V_p(L) = (\log(L))^\alpha$, which is slower growing. In this case Equation 7 is proportional to $L \log(L)^{1-\frac{\alpha}{2+\epsilon_V}}$. We can thus satisfy Equation 6 for $\text{gap}_p(L) \gtrsim L$ if we choose $V_p(L) = (\log(L))^\alpha$ with $\alpha > 2$. Finally, define $\log^{\circ N}$ as $\log$ composed with itself N times, and consider the very slow growing $V_p(L) = (\log^{\circ N}(L))^\alpha$, which corresponds to $\text{gap}_p(L) \gtrsim L \log(L) \log\log(L) \ldots \log^{\circ(N-1)}(L) (\log^{\circ N}(L))^{1-\frac{\alpha}{2+\epsilon_V}}$. Now, if $\text{gap}_p(L) \gtrsim L^\beta$ for some $\beta > 1$, then we can pick a V_p that grows as slowly as desired, ensuring very fast convergence. Also note that this is satisfied by $\text{supp}(p)$ being finite. Thus, in general, the faster gap_p increases, the slower we can make V_p increase, and the faster convergence rate we can guarantee.

In summary, in this section, we have studied the implications of a basic biological fact: sequences come in different lengths. We have found that length variation has a major impact on the convergence rate of stochastic processes over sequences. If the tail of the probability distribution falls off quickly with sequence length, convergence is rapid; if it falls off more gradually, convergence slows; if it falls off very gradually, convergence is no longer guaranteed at all. We will later find that in order for the KSD-B to detect non-convergence to a distribution p, the stochastic process used in the KSD-B (as defined by the Stein operator) must in fact converge to p.

B.2.5. Tails of Common Sequence Distributions

We now consider some examples of distributions p on sequence space, and study their tail behavior. This gives us the rate of convergence to p of the stochastic process defined by the Zanella Stein operator. It will also tell us whether or not the KSD-B can detect non-convergence to p.

We start with some relatively simple examples. We then study profile hidden Markov models (pHMMs), a type of model which is ubiquitous in biological sequence analysis. We show that we can guarantee convergence to any pHMM, and quantify the rate. pHMMs are often successful models of real biological sequence distributions, especially distributions over evolutionarily related proteins or protein domains. Our results are thus informative not just about pHMMs specifically, but also about biological sequence distributions found in nature.

Examples with Convergence We start with some simple examples of distributions p for which we can prove convergence and quantify the rate. Consider $p(X) \propto |\mathcal{B}|^{-L} e^{-\mu L}$, which falls off exponentially with sequence length. We have,

$$\text{gap}_p(L) = L\chi(|\mathcal{B}|e^\mu) - (L+1)|\mathcal{B}|\chi(|\mathcal{B}|^{-1}e^{-\mu}) = L\chi(|\mathcal{B}|e^\mu)\left(1 - \frac{L+1}{L}e^{-\mu}\right) \sim L.$$

Thus, by the discussion above, we can satisfy Equation 7 with $V_p(L) = (\log(L))^{2+\epsilon}$ for any $\epsilon > 0$. This translates into a convergence rate of $t^{-(1+\epsilon)} + L^\epsilon t^{-(2+\epsilon)}$ by Theorem B.4.

Alternately, consider $p(X) \propto |\mathcal{B}|^{-L} L!^{-1}$ where $L!$ is L factorial; this distribution falls off even faster with sequence length. Choose χ such that $\chi(t) = t^\alpha$ when $t \leq 1$ and $\chi(t) = t^{1-\alpha}$ when $t \geq 1$ for some $0 < \alpha < 1$. Then, whenever $|X| = |Y| - 1$, we have $p(X)/p(Y) = L|\mathcal{B}|$. As a result $\text{gap}_p(L) \sim L^{2-\alpha}$. We can now satisfy Equation 7 with $V_p(L) = \log^{\circ N}(L)$ for any N, ensuring very fast convergence of the stochastic process.

Example without Convergence We next consider an example of a distribution p for which convergence is not guaranteed. The distribution p is defined by an autoregressive model, with lag 2. The idea is that, for letters $A, B \in \mathcal{B}$, the motif ABA is high probability while AAA is low probability. Thus a sequence such as $X = AAAA$ may increase in probability by

gaining an insertion of B and so $\mathrm{del}_p(X) < \mathrm{ins}_p(X)$. This results in a situation where Assumption B.3 cannot be satisfied, and so we cannot guarantee convergence by Theorem B.4.

Proposition B.5. *Let $A \neq B \in \mathcal{B}$. Let p be a distribution on S such that $X \sim p$ all start with two A's, i.e. $X_{(0:2)} = AA$, and the rest of the sequence is sampled autoregressively with lag 2 as $X_{(l)} \sim p(b|X_{(l-2:l)})$. We set $p(A|AA) = 0.1$, $p(B|AA) = 0.8$, $p(\$|AA) = 0.1$, $p(A|AB) = 0.8$, $p(B|AB) = 0.1$, $p(\$|AB) = 0.1$, $p(A|BA) = 0.8$, $p(B|BA) = 0.1$, and $p(\$|BA) = 0.1$, and $p(\cdot|BB)$ to be anything, where $\$$ represents the end of the sequence. Then $\mathrm{gap}_p(L) < 0$ for large enough L and so p does not satisfy Assumption B.3.*

Proof. Call $X = L \times A$. $p(X) = 0.1^{L-1}$, so $\mathrm{del}_p(X) = L\chi(0.1^{-1})$. However, for $L_1, L_2 \geq 2$, $p(L_1 \times A + B + L_2 \times A) = 0.1^{L_1+L_2-2-1}0.8^3$, so, if $L_1 + L_2 = L$, $p(L_1 \times A + B + L_2 \times A)/p(L \times A) = 0.8^3 0.1^{-2} > 0.1^{-1}$. Thus,

$$\begin{aligned}
\mathrm{del}_p(L \times A) - \mathrm{ins}_p(L \times A) \leq & L\chi(0.1^{-1}) - \left(\sum_{L_1, L_2} \chi\left(\frac{p(L_1 \times A + B + L_2 \times A)}{p(L \times A)} \right) + (L+1)\chi\left(\frac{p((L+1) \times A)}{p(L \times A)} \right) \right) \\
\leq & L\chi(0.1^{-1}) - \left((L-3)\chi(0.1^{-1}) + (L+1)\chi(0.1) \right) \\
= & 3\chi(0.1^{-1}) - (L+1)\chi(0.1)
\end{aligned}$$

where we have used our assumption that χ is non-decreasing. Thus, $\mathrm{gap}_p(L) \leq 0$ for large enough L. □

Profile Hidden Markov models We now study the tails of pHMMs, a widely used probabilistic model of biological sequences. In this section, for a sequence $X \in S$ we define $X_{(l)}$ as its l-th letter, starting counting at $l = 0$, and $X_{(l:l')}$ as the sequence of $l' - l$ letters $X_{(l)}, X_{(l+1)}, \ldots, X_{(l'-1)}$. For an $X \in S$ and a number L define $L \times X$ as X concatenated to itself L times. Let $X + Y$ for $X, Y \in S$ be their concatenation.

To define a pHMM, we start with a Markov model with "match" states $\mathcal{J}_s = \{s_1, s_2, \ldots, s_{\tilde{L}}\}$, "insertion" states $\mathcal{J}_i = \{i_0, i_1, \ldots, i_{\tilde{L}}\}$, a start state s_0, and a termination state Δ. s_l and i_l may only transfer to $s_{l'}$ for $l' > l$ or $i_{l'}$ for $l' \geq l$. Then each of these hidden states, except s_0 and Δ, emits a $b \in \mathcal{B}$ with probability $p(b|Z)$ for a state Z. Thus a probability of a sequence X with $|X| = L$ can be written as

$$p(X) = \sum_{Z \in \mathcal{I}_L} p(Z)p(X|Z) = \sum_{Z \in \mathcal{I}_L} p(Z_L|Z_{L-1}) \prod_{l=0}^{L-1} p(Z_l|Z_{l-1})p(X_{(l)}|Z_l)$$

where we define $\mathcal{I}_L = \{(Z_{-1}, Z_0, Z_1, \ldots, Z_L) \mid Z_i \in \mathcal{J}_s \cup \mathcal{J}_i \text{ for } 1 \leq i \leq L, Z_L = \Delta, Z_{-1} = s_0\}$. We add a few mild conditions to our pHMM. The first is that infinite length insertions are not allowed, i.e. $\sup_l p(i_l|i_l) \leq e^{-\mu}$ for some $\mu > 0$. We also require that emission probabilities are non-zero, i.e. $p(b|Z) > 0$ for all states Z and $b \in \mathcal{B}$. Call $\eta = \min_{b,Z} p(b|Z)$. Finally, we require that if $p(Z|i_l) > 0$ for some state Z and $p(i_l|s_{l'}) > 0$, then $p(Z|s_{l'}) > 0$, that is, if a state can be reached by $s_{l'}$ by first adding an insertion, then it can be reached by $s_{l'}$ directly as well. This last condition guarantees that removing an insertion from any sequence of states Z does not make the sequence probability 0.

Before our proof let us build some intuition. For long sequences X we will see that the latent alignment $p(Z|X)$ has almost all its mass on Z for which almost all states have $Z_l = i_{l^*}$ where i_{l^*} is the insertion state that maximizes $p(i_l|i_l)$. In this case, $p(X) \approx p(X||X| \times i_{l^*})p(|X| \times i_{l^*}) = \left(\prod_{l=0}^{|X|-1} p(X_{(l)}|i_{l^*}) \right) e^{-\mu|X|}$. Let us thus consider the toy situation where $p(X) = e^{-\mu|X|} \prod_{l=0}^{|X|-1} q(X_{(l)})$ for some distribution q over $\mathcal{B}$ (note this is also technically a pHMM). For every sequence X in this case,

$$\mathrm{ins}_p(X) = (L+1) \sum_{b \in \mathcal{B}} \chi(e^{-\mu} q(b)) = (L+1)e^{-\mu} \sum_b q(b)\chi(e^{\mu} q(b)^{-1}) = (L+1)e^{-\mu} E_{b \sim q} \chi(e^{\mu} q(b)^{-1}),$$

$$\mathrm{del}_p(X) = \sum_{l=0}^{|X|-1} \chi(e^{\mu} q(X_{(l)})^{-1}).$$

Thus, if b^* maximizes $q(b)$,

$$\mathrm{gap}_p(L) = L\chi(e^{\mu} q(b^*)^{-1}) - (L+1)e^{-\mu} E_{b \sim q} \chi(e^{\mu} q(b)^{-1}) = L \left(\chi(e^{\mu} q(b^*)^{-1}) - \frac{L+1}{L} e^{-\mu} E_{b \sim q} \chi(e^{\mu} q(b)^{-1}) \right).$$

If $q(b) = |\mathcal{B}|^{-1}$ for all b then $E_{b\sim q}\chi(e^{\mu}q(b)^{-1}) = \chi(e^{\mu}q(b^*)^{-1})$ and we recover the situation of our example with $p(X) \propto |\mathcal{B}|^{-L}e^{-\mu L}$. However if $q(b)$ is not uniform and χ is strictly increasing, we can have $E_{b\sim q}\chi(e^{\mu}q(b)^{-1}) > \chi(e^{\mu}q(b^*)^{-1})$. In this case, if μ is sufficiently small, then $\mathrm{gap}_p(L) < 0$ for large enough L. Thus for general pHMMs p and χ, whether or not p has uniformly decreasing tails can depend on μ and the emission probabilities at the most likely insertion. Note however, by selecting $\chi(t) = t \wedge 1$, $E_{b\sim q}\chi(e^{\mu}q(b)^{-1}) = \chi(e^{\mu}q(b^*)^{-1}) = 1$ so $\mathrm{gap}_p(L) \sim L$, and thus p satisfies Assumption B.3 regardless of q and μ. We will therefore take $\chi(t) = t \wedge 1$, and show that in this case pHMMs always satisfy Assumption B.3.

We now characterize the tails of pHMMs by lower bounding $\mathrm{gap}_p(L)$. Note, our analysis also allows us to prove that pHMMs are subexponential, i.e. if p is a pHMM then $E_p e^{t|X|} < \infty$ for any t small enough (Amin et al., 2021); this will be useful later for proving the pHMM is p, k-integrable.

Proposition B.6. *If p is a pHMM and $\chi(t) = t \wedge 1$ then $\mathrm{gap}_p(L) \gtrsim L$. Also, $\mathrm{ins}_p(X) \lesssim \mathrm{del}_p(X) \sim \mathrm{flux}_p(X) \sim |X|$ and $E_p e^{t|X|} < \infty$ for any $t < \mu$.*

Proof. Let $|X| = L$. Applying our choice of χ, we have

$$\mathrm{ins}_p(X) = \sum_{|Y|=L+1, XMY} T_{p,X\to Y} \leq \frac{1}{p(X)} \sum_{l=0}^{L} \sum_{b\in\mathcal{B}} p(X_{b,+l})$$

where $X_{b,+l}$ is the sequence X with an inserted letter b at position l. Now we use the sum over $\mathcal{B}$ to marginalize out the emission at position l:

$$\begin{aligned}
\sum_{b\in\mathcal{B}} p(X_{b,+l}) &= \sum_{b\in\mathcal{B}} \sum_{Z\in\mathcal{I}_{L+1}} p(Z)p(X_{b,+l}|Z) \\
&= \sum_{b\in\mathcal{B}} \sum_{Z\in\mathcal{I}_{L+1}} p(Z) \left(\prod_{l'=0}^{l-1} p(X_{b,+l,(l')}|Z_{l'}) \prod_{l'=l}^{L} p(X_{b,+l,(l'+1)}|Z_{l'+1}) \right) p(X_{b,+l,(l)}|Z_l) \\
&= \sum_{Z\in\mathcal{I}_{L+1}} p(Z) \prod_{l'=0}^{l-1} p(X_{(l')}|Z_{l'}) \prod_{l'=l}^{L} p(X_{(l')}|Z_{l'+1}) \left(\sum_{b\in\mathcal{B}} p(b|Z_l) \right) \\
&= \sum_{Z\in\mathcal{I}_{L+1}} p(Z)p(X|\tilde{Z})
\end{aligned}$$

where, for $Z \in \mathcal{I}_{L+1}$, $\tilde{Z} \in \mathcal{I}_L$ is defined to be Z but with Z_l removed. The idea of the proof is to show that the leading terms of the last sum are ones in which Z_l is in the middle of a multiple insertion. For these Z, $Z \mapsto \tilde{Z}$ is an injection and we can replace $p(Z)$ with its upper bound $e^{-\mu}p(\tilde{Z})$. Then summing over $\tilde{Z}$ will give us $e^{-\mu}p(X)$ and finally summing over l and dividing by $p(X)$ will give our bound $|X|e^{-\mu}$ on $\mathrm{ins}_p(X)$.

We take L to be sufficiently large such that $L > 3\tilde{L}$, where recall $\tilde{L}$ is the number of match states and the number of insertion states in the pHMM. For every $\hat{L}, l$, define $\mathcal{I}_{\hat{L},s}(l) = \{Z \in \mathcal{I}_{\hat{L}} \mid Z_l \in \mathcal{J}_s\}$ and $\mathcal{I}_{\hat{L},i}(l) = \{Z \in \mathcal{I}_{\hat{L}} \mid Z_l \in \mathcal{J}_i\}$. First we consider Z with a match state at position l, i.e. $Z \in \mathcal{I}_{L+1,s}(l)$. For each $Z \in \mathcal{I}_{L+1,s}(l)$ pick a position l_Z such that $Z_{l_Z} = Z_{l_Z+1} = Z_{l_Z-1} \in \mathcal{J}_i$, i.e. l_Z is in the middle of a multiple insertion. Define $\hat{Z}$ to be Z with l_Z removed; note that $\hat{Z} \in \mathcal{I}_{L,s}(l-1) \cup \mathcal{I}_{L,s}(l)$. First note, by our choice of l_Z, $p(Z)/p(\hat{Z}) < 1$. Next, note $\hat{Z}$ differs from $\tilde{Z}$ in at most $2\tilde{L}$ positions, since there are $2\tilde{L}$ states (excluding the start and termination states). Using the fact that the emission probability is lower bounded by η, we have $p(X|\tilde{Z}) \leq p(X|\hat{Z})\eta^{-2\tilde{L}}$. Finally, note that at most $\tilde{L}+1$ different Z map to the same $\hat{Z}$,

i.e. $\hat{Z}$ has at most $\tilde{L}+1$ multiple insertions (since there are $\tilde{L}+1$ insertion states). Now write

$$\begin{aligned}
\sum_{Z\in\mathcal{I}_{L+1,s}(l)} p(Z)p(X|\tilde{Z}) =& \sum_{Z\in\mathcal{I}_{L+1,s}(l)} \frac{p(Z)}{p(\hat{Z})}\frac{p(X|\tilde{Z})}{p(X|\hat{Z})}p(\hat{Z})p(X|\hat{Z}) \\
\leq& \eta^{-2\tilde{L}} \sum_{Z\in\mathcal{I}_{L+1,s}(l)} p(\hat{Z})p(X|\hat{Z}) \\
\leq& \eta^{-2\tilde{L}}(\tilde{L}+1) \sum_{Z'\in\mathcal{I}_{L,s}(l-1)\cup\mathcal{I}_{L,s}(l)} p(Z')p(X|Z') \\
=& \eta^{-2\tilde{L}}(\tilde{L}+1)p(X, Z'\in\mathcal{I}_{L,s}(l-1)\cup\mathcal{I}_{L,s}(l)) \\
\leq& \eta^{-2\tilde{L}}(\tilde{L}+1)\left(p(X, Z_l\in\mathcal{J}_s) + p(X, Z_{l-1}\in\mathcal{J}_s)\right).
\end{aligned}$$

For the first term in the sum write

$$\begin{aligned}
\frac{1}{p(X)}\sum_{l=0}^{L}\eta^{-2\tilde{L}}(\tilde{L}+1)p(X, Z_l\in\mathcal{J}_s) =& \eta^{-2\tilde{L}}(\tilde{L}+1)\sum_{l=0}^{L} p(Z_l\in\mathcal{J}_s|X) \\
=& \eta^{-2\tilde{L}}(\tilde{L}+1)\sum_{l=0}^{L} E\left[\mathbb{1}(Z_l\in\mathcal{J}_s)|X\right] \\
=& \eta^{-2\tilde{L}}(\tilde{L}+1)E\left[\sum_{l=0}^{L}\mathbb{1}(Z_l\in\mathcal{J}_s)\middle|X\right] \\
\leq& \eta^{-2\tilde{L}}(\tilde{L}+1)\tilde{L} = O(1)
\end{aligned}$$

Where the last inequality follows from the fact that if $p(Z)>0$, then at most $\tilde{L}$ states are letters. The second term is similar.

Next we consider $Z\in\mathcal{I}_{L+1,i}(l)$. Note that at most $\tilde{L}+1$ elements Z of $\mathcal{I}_{L+1,i}(l)$ map to the same $\tilde{Z}$. (The bound $\tilde{L}+1$ comes from considering the $\tilde{L}+1$ possible values of the deleted state. For example, consider a Z with entries $Z_{l'}=i_{\tilde{L}}$ for all $l'\in\{1,\ldots,\tilde{L}\}$. If we delete the zeroth entry of Z to obtain $\tilde{Z}$, i.e. we set $l=0$, then we always obtain the same value of Z, regardless of Z_0. Since Z_0 can take any value $i_{l'}$ for $l'\in\{0,1,\ldots,\tilde{L}\}$, we have $\tilde{L}+1$ possibilities.) Note also that by the fact that $p(Z)=p(Z_0|Z_{-1})\times\ldots p(Z_{L+1}|Z_L)$ and our assumption that removing an insertion does not make the sequence probability 0, there is a $\gamma>0$ such that $p(Z)/p(\tilde{Z})\leq\gamma$ for all $Z\in\mathcal{I}_{L+1,i}(l)$. We will split $\mathcal{I}_{L+1,i}(l)$ into two parts: define $A_1=\{Z\in\mathcal{I}_{L+1,i}(l)\mid Z_{l-1}\neq Z_{l+1}\}$ and $A_2=\{Z\in\mathcal{I}_{L+1,i}(l)\mid Z_{l-1}=Z_{l+1}\}$. That is, if $Z\in A_2$ then $Z_{l-1}=Z_{l+1}$, so position l is in a multiple insertion and $Z_l=Z_{l-1}$. Thus, if $Z\in A_2$, then $p(Z)/p(\tilde{Z})\leq e^{-\mu}$, and, $Z\mapsto\tilde{Z}$ is injective on A_2. On the other hand, if $Z\in A_1$ then $\tilde{Z}_{l-1}\neq\tilde{Z}_l$. Thus,

$$\begin{aligned}
\sum_{Z\in A_1} p(Z)p(X|\tilde{Z}) \leq& \gamma\sum_{Z\in A_1} p(\tilde{Z})p(X|\tilde{Z}) \leq (\tilde{L}+1)\gamma p(X, Z_{l-1}\neq Z_l) \\
\sum_{Z\in A_2} p(Z)p(X|\tilde{Z}) \leq& e^{-\mu}\sum_{Z\in A_2} p(\tilde{Z})p(X|\tilde{Z}) \leq e^{-\mu}p(X)
\end{aligned}$$

and, since there are $2\tilde{L}+3$ total states in the Markov chain,

$$\begin{aligned}
\frac{1}{p(X)}\sum_{l=0}^{L}\sum_{Z\in A_1} p(Z)p(X|\tilde{Z}) \leq& (\tilde{L}+1)\gamma E\left[\sum_{l=0}^{L}\mathbb{1}(Z_{l-1}\neq Z_l)\middle|X\right] \leq (\tilde{L}+1)\gamma(2\tilde{L}+3) = O(1) \\
\frac{1}{p(X)}\sum_{l=0}^{L}\sum_{Z\in A_2} p(Z)p(X|\tilde{Z}) \leq& e^{-\mu}L.
\end{aligned}$$

Combining the above results we finally have

$$\mathrm{ins}_p(X) \leq Le^{-\mu} + O(1)$$

Considering deletions now, we have

$$\mathrm{del}_p(X) = \sum_{|Y|=L-1, XMY} T_{p,X\to Y} = \sum_{l=0}^{L-1} \frac{p(X_{-l})}{p(X)} \wedge 1$$

with X_{-l} defined to be X with position l deleted. In this case,

$$p(X_{-l}) = \sum_{Z\in\mathcal{I}_{L-1}} p(Z) \prod_{l'=0}^{l-1} p(X_{(l')}|Z_{l'}) \prod_{l'=l+1}^{L-1} p(X_{(l')}|Z_{l'-1}).$$

For $Z \in \mathcal{I}_{L-1,i}(l-1)$, let $\tilde{Z}$ be Z but with an extra i_k in position l if $Z_{l-1} = i_k$. For $Z \in \mathcal{I}_{L-1,i}(l-1)$, $p(Z)/p(\tilde{Z}) \geq e^{\mu} \geq e^{\mu} p(X_l|\tilde{Z}_l)$ and $Z \mapsto \tilde{Z}$ is a bijection to elements Z of $\mathcal{I}_L$ such that $Z_{l-1} = Z_l \in \mathcal{J}_i$. Thus we have,

$$\frac{1}{p(X)} p(X_{-l}) \geq e^{\mu} \frac{1}{p(X)} \sum_{Z\in\mathcal{I}_{L-1,i}(l-1)} p(\tilde{Z}) p(X|\tilde{Z}) = e^{\mu} p(Z_{l-1} = Z_l \in \mathcal{J}_i|X).$$

Now let $R = \sum_{l=0}^{L} \left(e^{\mu} \mathbb{1}(Z_{l-1} = Z_l \in \mathcal{J}_i)\right) \wedge 1 = \sum_{l=0}^{L} \mathbb{1}(Z_{l-1} = Z_l \in \mathcal{J}_i)$, which is lower bounded by $L - 3\tilde{L}$. Thus

$$\mathrm{del}_p(X) = \sum_{|Y|=L-1, XMY} T_{p,X\to Y} \geq E_p\left[R|X\right] \geq \left(L - 3\tilde{L}\right) = L - O(1).$$

On the other hand we clearly have $\mathrm{del}_p(X) \leq L$. Note we similarly have $\mathrm{flux}_p(X)$ must be less than the number of neighbours of X, $L + (|B|-1)L + |B|(L+1)$. Thus we have $\mathrm{ins}_p(X) \lesssim \mathrm{del}_p(X) \sim \mathrm{flux}_p(X) \sim |X|$ and $\mathrm{gap}_p(L) \geq (L - O(1)) - (Le^{-\mu} + O(1)) \gtrsim L$.

Finally, recall out bound for any sequence X,

$$\mathrm{ins}_p(X) \leq \frac{1}{p(X)} \sum_{l=0}^{L} \sum_{b\in\mathcal{B}} p(X_{b,+l}) \leq Le^{-\mu} + O(1) \leq (L+1)(e^{-\mu} + o(1)).$$

Thus,

$$\sum_{|X|=L} p(X) \geq \sum_{|X|=L} p(X) \frac{1}{(L+1)(e^{-\mu}+o(1))} \left(\frac{1}{p(X)} \sum_{l=0}^{L} \sum_{b\in\mathcal{B}} p(X_{b,+l})\right) = (e^{-\mu}+o(1))^{-1} \sum_{|X|=L+1} p(X)$$

so $p(|X| = L) \lesssim e^{-tL}$ if $e^{-t} > e^{-\mu}$. In particular, if $t < \mu$, then $E_p e^{t|X|} = \sum_L p(|X| = L) e^{tL} < \infty$. □

Now, for any pHMM, we can guarantee convergence of the stochastic process and characterize its convergence rate.

Corollary B.7. *If p is a pHMM and $\chi(t) = t \wedge 1$ then $E_p \mathrm{flux}_p(X) < \infty$ and Assumption B.3 is satisfied with $V_p(L) = (\log L)^{2+\epsilon}$ for any $\epsilon > 0$.*

This convergence speed is faster than that we obtained in the scenario where p decayed exponentially with L, but slower than when p decayed as $L!$ or when p had zero probability past a certain length. It suggests that for many real biological sequence distributions – those for which the pHMM is a good model – we can expect reasonably fast convergence of the stochastic process.

B.3. Proofs of the KSD-B's Properties

In this section we prove the results described in the main text for KSD-Bs. Section B.3.1 shows how to tractably compute the KSD-B. Section B.3.2 shows that scalar field KSD-Bs can be written as a special case of vector field KSD-Bs. Section B.3.3 introduces the property of discrete masses and describes its generalization to vector field kernels. Section B.3.4 establishes conditions under which the KSD-B is faithful and detects tight non-convergence. Section B.3.5 establishes conditions under which the KSD-B can detect non-convergence more generally, and also proves that the KSD-B detects convergence. Section B.3.6 details examples of scenarios where the KSD-B can fail, if the kernel is not chosen appropriately or p is pathological. Section B.3.7 shows how to accurately approximate the KSD-B.

B.3.1. CALCULATING THE KSD-B

In this section we develop a tractable expression for computing the KSD-B. First note that by our definition that $T_{p,X\to Y} = \infty$ if $p(X) = 0$, any p, k integrable distribution q must have $\text{supp}(q) \subseteq \text{supp}(p)$.

Proposition B.8. *(Proof of Eq. 1 and Proposition 4.1) Say k is a vector field kernel and q is a p, k-integrable distribution on S. Then for all $f \in \mathcal{H}_k$,*

$$E_q \mathcal{T}_p f = \frac{1}{2} \sum_{(X,Y)\in M_{p,p}} p(Y) T_{p,Y\to X} \left(\frac{q(X)}{p(X)} - \frac{q(Y)}{p(Y)} \right) f(X,Y). \tag{8}$$

Note when $q(X) > 0$, for $(X, Y) \in M_{p,p}$,

$$p(Y) \left(\frac{q(X)}{p(X)} - \frac{q(Y)}{p(Y)} \right) = q(X) \left(\frac{p(Y)}{p(X)} - \frac{q(Y)}{q(X)} \right)$$

and we recover Eq. 1. As well,

$$\text{KSD-B}_{p,k}(q)^2 = E_{X,X'\sim q} \sum_{YMX,Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k((X,Y),(X',Y')).$$

If p is p, k-integrable, then for all $f \in \mathcal{H}_k$, $E_p \mathcal{T}_p f = 0$.

Proof. Say q is p, k-integrable. Define $\phi_q(f) = E_q \mathcal{T}_p f$. For $f \in \mathcal{H}_k$,

$$\begin{aligned} \phi_q(f) =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} (f|k_{(X,Y)})_k \\ \leq& \|f\|_k E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \sqrt{k((X,Y),(X,Y))}. \end{aligned} \tag{9}$$

by Cauchy Schwarz. Thus ϕ_q is a bounded linear operator on $\mathcal{H}_k$ and is thus a member of $\mathcal{H}_k$, by the Riesz representation theorem. As well, $\text{KSD-B}_{p,k}(q) = \|\phi_k\|_k$. We now have,

$$\begin{aligned} (\phi_q|\phi_q)_k =& \phi_q(\phi_q) \\ =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \phi_q(k_{(X,Y)}) \\ =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} E_{X'\sim q} \sum_{Y'MX'} T_{p,X'\to Y'} k_{(X,Y)}(X',Y') \\ =& E_{X,X'\sim q} \sum_{YMX} \sum_{Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k((X,Y),(X',Y')). \end{aligned}$$

Note that since all quantities in the expectation and sum are positive, Eqn. 9 shows the absolute integrability of the expectation and sum. Thus we can rearrange terms to get

$$\begin{aligned} \phi_q(f) =& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} f(X,Y) \\ =& \sum_{(X,Y)\in M_{p,p}} q(X) T_{p,X\to Y} f(X,Y) \\ =& \frac{1}{2} \sum_{(X,Y)\in M_{p,p}} (q(X) T_{p,X\to Y} f(X,Y) + q(Y) T_{p,Y\to X} f(Y,X)) \\ =& \frac{1}{2} \sum_{(X,Y)\in M_{p,p}} p(Y) T_{p,Y\to X} \left(\frac{q(X)}{p(X)} - \frac{q(Y)}{p(Y)} \right) f(X,Y) \end{aligned}$$

where the last line follows from detailed balance, $T_{p,X\to Y} p(X) = T_{p,Y\to X} p(Y)$. If we set $p = q$ we have $\frac{q(X)}{p(X)} = \frac{q(Y)}{p(Y)}$ for all $(X, Y) \in M_{p,p}$, so $E_p \mathcal{T}_p f = 0$. □

B.3.2. SCALAR FIELD KSD-BS AS AN INSTANCE OF VECTOR FIELD KSD-BS

Here we demonstrate that every scalar field KSD-B can be written as a special case of a vector field KSD-B. Recall that scalar field Stein discrepancies take the form $\sup_{g\in\nabla\mathcal{F}}|E_q\mathcal{T}g|$, whereas vector field Stein discrepancies take the more general form $\sup_{g\in\mathcal{G}}|E_q\mathcal{T}g|$. We will show that if the family of functions $\mathcal{F}$ is an RKHS $\mathcal{H}_k$ with kernel k, then the set of functions $\nabla\mathcal{F}$ – that is, the set of gradients of functions in $\mathcal{H}_k$ – is itself an RKHS with kernel k^∇. This implies that the scalar field Stein discrepancy with kernel k is equivalent to a vector field Stein discrepancy with kernel k^∇. To show this, we start by defining k^∇.

Proposition B.9. *For any given scalar field kernel k on S,*

$$k^\nabla((X,Y),(X',Y')) = (k_Y - k_X|k_{Y'} - k_{X'})_k = k(Y,Y') - k(X,Y') - k(Y,X') + k(X,X')$$

for $(X,Y),(X',Y')\in M$ defines a vector field kernel. For every $f\in\mathcal{H}_{k^\nabla}$ there is a $g\in\mathcal{H}_k$ with $f=\nabla g$ and $\|f\|_{k^\nabla}=\|g\|_k$.

Proof. k^∇ is non-negative definite as if $(X_1,Y_1).\ldots,(X_N,Y_N)\in M$ and $\alpha_1,\ldots,\alpha_N\in\mathbb{R}$ then, calling $f=\sum_n\alpha_n k_{X_n}$ and $g=\sum_n\alpha_n k_{Y_n}$,

$$\sum_{n,m}\alpha_n\alpha_m k^\nabla((X_n,Y_n),(X_m,Y_m)) = (g|g)_k - (f|g)_k - (g|f)_k + (f|f)_k = \|f-g\|_k \geq 0.$$

One can also verify that $k^\nabla_{(X,Y)} = -k^\nabla_{(Y,X)}$ for all $(X,Y)\in M$, so for every $f\in\mathcal{H}_{k^\nabla}$,

$$f(X,Y) = (f|k^\nabla_{(X,Y)})_{k^\nabla} = -(f|k^\nabla_{(Y,X)})_{k^\nabla} = -f(Y,X).$$

Thus k^∇ is a vector field kernel.

Let $f=\sum_{i=1}^n\alpha_i k^\nabla_{(X_i,Y_i)}$ and $g=\sum_{i=1}^n\alpha_i(k_{Y_i}-k_{X_i})$.

$$f(X,Y) = \sum_{i=1}^n\alpha_i(k_{Y_i}-k_{X_i}|k_Y-k_X)_k = g(Y)-g(X) = \nabla g(X).$$

As well,

$$\|f\|^2_{k^\nabla} = \sum_{i=1}^n\sum_{j=1}^n\alpha_i\alpha_j(k_{Y_i}-k_{X_i}|k_{Y_j}-k_{X_j})_k = (g|g)_k = \|g\|^2_k.$$

Now say $f\in\mathcal{H}_{k^\nabla}$ and $(f_n)_n$ is a sequence of finite linear combinations of $(k^\nabla_{(X,Y)})_{(X,Y)\in M}$ such that $f_n\to f$. Say $g_n\in\mathcal{H}_k$ such that $\nabla g_n = f_n$. Since $\|f_n-f_m\|_{k^\nabla}=\|g_n-g_m\|_k$, $(g_n)_n$ is a Cauchy sequence and thus converges to a $g\in\mathcal{H}$. $\|g\|_k=\lim_n\|g_n\|_k=\lim_n\|f_n\|_{k^\nabla}=\|f\|_{k^\nabla}$ and finally,

$$f(X,Y) = (f|k^\nabla_{(X,Y)})_{k^\nabla} = \lim_n(f_n|k^\nabla_{(X,Y)})_{k^\nabla} = \lim_n(g_n|k_Y-k_X)_k = (g|k_Y-k_X)_k = \nabla g(X,Y).$$

□

Now we show that k^∇ defines a vector field KSD-B that is identical to a scalar field KSD-B with k.

Proposition B.10. *Say k is a scalar field kernel on S. If q is a p,k^∇-integrable distribution on S, then,*

$$\sup_{\|f\|_{k^\nabla}\leq 1} E_q\mathcal{T}_p f = \sup_{\|f\|_k\leq 1} E_q\mathcal{T}_p\nabla f.$$

Proof. Define, similar to Proposition B.8, $\tilde{\phi}_q:\mathcal{H}_k\to\mathbb{R}\mid f\mapsto E_q\mathcal{T}_p\nabla f$. For $f\in\mathcal{H}_k$,

$$\begin{aligned}\tilde{\phi}_q(f) =& E_{X\sim q}\sum_{YMX}T_{p,X\to Y}(f|k_Y-k_X)_k\\ \leq& \|f\|_k E_{X\sim q}\sum_{YMX}T_{p,X\to Y}\|k_Y-k_X\|_k\\ \leq& \|f\|_k E_{X\sim q}\sum_{YMX}T_{p,X\to Y}\sqrt{k^\nabla((X,Y),(X,Y))}.\end{aligned}$$

Thus $\tilde{\phi}_q$ is a bounded linear operator on $\mathcal{H}_k$ and, by the Riesz representation theorem, is a member of $\mathcal{H}_k$. As well, $\left(\sup_{\|f\|_k \leq 1} E_q \mathcal{T}_p \nabla f\right)^2 = \|\phi_k\|_k^2$. Now,

$$\begin{aligned}
(\tilde{\phi}_q|\tilde{\phi}_q)_k =& \tilde{\phi}_q(\tilde{\phi}_q) \\
=& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \left(\tilde{\phi}_q(k_Y - k_X)\right) \\
=& E_{X\sim q} \sum_{YMX} T_{p,X\to Y} \\
& \times \left(E_{X'\sim q} \sum_{Y'MX'} T_{p,X'\to Y'} \left((k_Y(Y') - k_X(Y')) - (k_Y(X') - k_X(X'))\right)\right) \\
=& E_{X,X'\sim q} \sum_{YMX,Y'MX'} T_{p,X\to Y} T_{p,X'\to Y'} k^{\nabla}((X,Y),(X',Y')) \\
=& \text{KSD-B}_{p,k}(q)^2.
\end{aligned}$$

□

B.3.3. Kernels with Discrete Masses

We want the KSD-B to be able to be faithful, i.e. $\text{KSD-B}_{p,k}(q) = 0$ if and only if $p = q$. For this to hold, the set of test functions the KSD-B uses for comparing q_n and p – that is, $\mathcal{H}_k$ – must be sufficiently large. In Euclidean space, one can make sure $\mathcal{H}_k$ is sufficiently large by using a kernel that is universal (Gorham & Mackey, 2017). In our infinite discrete case, we will use kernels with discrete masses, i.e. kernels whose RKHS contains delta functions (Jorgensen & Tian, 2015). In this section we introduce the property of discrete masses, generalize it to vector fields, and discuss its implications. Later we will show how the discrete mass property, together with additional conditions on the kernel's tail, ensures the KSD-B can detect non-convergence (Section B.3.4-B.3.5); we then construct practical kernels with discrete masses (Section B.4).

A standard, scalar field kernel has discrete masses if for every point $X \in S$, the RKHS contains a delta function at X. For vector field kernels, we generalize this idea from single points to pairs of neighboring points.

Definition B.11 (Kernel with discrete masses)**.** (A) If k is a scalar field kernel on S, then we say k has discrete masses if $\delta_X \in \mathcal{H}_k$ for all $X \in S$, where δ_X is the function that is 1 at X and 0 elsewhere.

(B) If k is a vector field kernel, then we say k has discrete masses if $\delta_{(X,Y)} \in \mathcal{H}_k$ for all $(X,Y) \in M$, where $\delta_{(X,Y)}$ is the vector field on M that is 1 at (X,Y), -1 at (Y,X), and 0 elsewhere.

To see that having discrete masses implies that $\mathcal{H}_k$ is large in some absolute sense, note that a scalar field kernel has discrete masses if and only if $C_C(S) \subset \mathcal{H}_k$, and a vector field kernel k has discrete masses if and only if $C_{C,vf}(M) \subset \mathcal{H}_k$. Moreover, $\mathcal{H}_k$ is dense in any space for which $C_{C,vf}(M)$ or $C_C(S)$ are dense. Thus, a kernel with discrete masses is C_0 and L^p-universal, meaning that any function in C_0 or L_p-space can be approximated arbitrarily well by a function in $\mathcal{H}_k$ (Sriperumbudur et al., 2011; Amin et al., 2023). Note also that kernels on Euclidean space cannot have discrete masses (Amin et al., 2023).

Kernels with discrete masses are guaranteed to detect non-convergence when used in a maximum mean discrepancy (MMD) (Sriperumbudur et al., 2010; Amin et al., 2023). Although the KSD-B is closely related to the MMD, the same results do not transfer directly. In the MMD, the set of test functions is the RKHS $\mathcal{H}_k$ itself, i.e. we have $\sup_{\|f\|_k \leq 1 : f \in \mathcal{H}_k} |E_q f - E_p f|$. In the KSD-B, however, the set of test functions is the Stein operator applied to the RKHS, $\sup_{\tilde{f} \in \mathcal{T}_p(\{\|f\|_k \leq 1 : f \in \mathcal{H}_k\})} |E_q \tilde{f} - E_p \tilde{f}|$.

We introduced vector field KSD-Bs as a generalization of scalar field KSD-Bs with a larger set of test functions. One way in which this notion of a larger set of test functions manifests itself is in terms of discrete masses. Consider a scalar field kernel k; even if this kernel has discrete masses, its corresponding vector field kernel k^{∇} (Proposition B.10) cannot have discrete masses. In other words, even if k can describe a very large set of functions on S, k^{∇} cannot describe a very large set of vector field functions on $M \subset S \times S$. This is one important way in which scalar field KSD-Bs are limited.

Proposition B.12. *Say k is a kernel on S. Then k^{∇} does not have discrete masses.*

Proof. Let X_1, X_2, X_3 be three distinct sequences in S such that $X_1MX_2MX_3MX_1$. For any $(X, Y) \in M$, calling $f = k^{\nabla}_{(X,Y)}$, we have

$$\begin{aligned} f(X_1, X_2) + f(X_2, X_3) + f(X_3, X_1) = \\ &= k^{\nabla}((X, Y), (X_1, X_2)) + k^{\nabla}((X, Y), (X_2, X_3)) + k^{\nabla}((X, Y), (X_3, X_1)) \\ &= (k_Y - k_X | (k_{X_2} - k_{X_1}) + (k_{X_3} - k_{X_2}) + (k_{X_1} - k_{X_3}))_k \\ &= 0. \end{aligned}$$

Thus, for all $f \in \mathcal{H}_{k^{\nabla}}$, $f(X_1, X_2) + f(X_2, X_3) + f(X_3, X_1) = 0$. However,

$$\delta_{(X_1,X_2)}(X_1, X_2) + \delta_{(X_1,X_2)}(X_2, X_3) + \delta_{(X_1,X_2)}(X_3, X_1) = 1.$$

□

B.3.4. Faithfulness and Tight Non-convergence

In this section we show that the KSD-B is faithful, meaning that $\text{KSD-B}_{p,k}(q) = 0$ if and only if $p = q$. We also show that the KSD-B can detect *tight* non-convergence, meaning $\text{KSD-B}_{p,k}(q_n) \not\to 0$ if $q_n \not\to p$ and $(q_n)_n$ is uniformly tight. Intuitively, uniform tightness says that the distributions q_n do not become too diffuse or spread out as $n \to \infty$ – a scenario that may occur, for instance, if q_n is the empirical distribution of samples drawn by a biased sampler that "spins out of control" (Gorham & Mackey, 2017). In the next section, we will relax the tightness assumption, guarding against such a possibility.

The basic idea behind our proof is that, if we use a kernel with discrete masses, $\text{KSD-B}_{p,k}(q) = 0$ implies $E_q f = 0$ for all f in $\mathcal{T}_p(C_{C,vf}(M))$ (or, for a scalar field kernel, for all f in $\mathcal{T}_p\nabla(C_C(S))$). This in turn implies $q = p$, as we show in the following lemma. A brief technical point: when k is a scalar field kernel, we will rely on the fact that the only distribution stationary for the stochastic process induced by $\mathcal{L}_p$ is p (Lemma B.1 (B)). Thus we must add additional integrability assumptions to ensure that this process exists, namely that $E_p\text{flux}_p$ and $E_q\text{flux}_p$ are finite. In the vector field case, such extra assumptions are unnecessary as we can appeal directly to Equation 1.

Lemma B.13. *Say p has connected support and q is a distribution on S. If $E_q\mathcal{T}_p f \neq \infty$ for all $f \in C_{C,vf}(M)$, or if $E_q\mathcal{T}_p\nabla f \neq \infty$ for all $f \in C_C(S)$, then $\text{supp}(q) \subseteq \text{supp}(p)$. If $E_q\mathcal{T}_p f = 0$ for all $f \in C_{C,vf}(M)$, then $q = p$. Or, if $E_q\text{flux}_p < \infty$, $E_p\text{flux}_p < \infty$ and $E_q\mathcal{T}_p\nabla f = 0$ for all $f \in C_C(S)$, then $q = p$.*

Proof. Assume $E_q\mathcal{T}_p f$ is well defined and finite for all $f \in C_{C,vf}(M)$. If $\text{supp}(q) \not\subseteq \text{supp}(p)$ then there is a $X \in \text{supp}(q) \setminus \text{supp}(p)$ such that there is a YMX where either $q(Y) = 0$ or $Y \in \text{supp}(p)$. In either case, we have $q(Y)T_{p,Y\to X} = 0$ (recall we define $0 \times \infty = 0$). Thus, from the definition of $\mathcal{T}_p$,

$$E_q\mathcal{T}_p\delta_{(X,Y)} = q(X)T_{p,X\to Y} - q(Y)T_{p,Y\to X} = \infty,$$

since $T_{p,X\to Y}$ is defined to be ∞ when $X \notin \text{supp}(p)$. This contradicts the assumption that $E_q\mathcal{T}_p f$ is finite, so $\text{supp}(q) \subseteq \text{supp}(p)$.

The proof for the scalar field kernel proceeds analogously. Assume $E_q\mathcal{T}_p\nabla f \neq \infty$ for all $f \in C_C(S)$, and say $\text{supp}(q) \not\subseteq \text{supp}(p)$. Again pick $X \in \text{supp}(q) \setminus \text{supp}(p)$ such that there is a YMX such that $q(Y) = 0$ or $Y \in \text{supp}(p)$. We have,

$$E_q\mathcal{T}_p\nabla\delta_Y = q(Y)\mathcal{T}_p\nabla\delta_Y(Y) + \sum_{ZMY} q(Z)\mathcal{T}_p\nabla\delta_Y(Z) = -q(Y)\text{flux}_p(Y) + \sum_{ZMY} q(Z)T_{p,Z\to Y},$$

and in either case the first term is finite and the second is ∞, a contradiction.

Now say $E_q\mathcal{T}_p f = 0$ for all $f \in C_{C,vf}(M)$. If $X \in \text{supp}(q), Y \in \text{supp}(p)$ and YMX, we have from Equation 1,

$$0 = E_q\mathcal{T}_p\delta_{(X,Y)} = q(X)T_{p,Y\to X}\left(\frac{p(Y)}{p(X)} - \frac{q(Y)}{q(X)}\right).$$

Thus, $q(Y)/q(X) = p(Y)/p(X)$. Thus $\text{supp}(q) = \text{supp}(p)$ and $q(Y)/q(X) = p(Y)/p(X)$ for all $(X, Y) \in M_{p,p}$. Since the support of p in connected this implies that $q = p$.

Now if $E_p\text{flux}_p < \infty$ and $E_q\mathcal{T}_p\nabla f = 0$ for all $f \in C_C(S)$ then $q = p$ by Lemma B.1 (B).

□

We now show the KSD-B is faithful and detects tight non-convergence, proving Proposition 5.1 in the main text.

Proposition B.14. *Say $\mathrm{supp}(p)$ is connected. Assume either (a) k is a vector field kernel with discrete masses or (b) k is a scalar field kernel on S with discrete masses and $E_p \mathrm{flux}_p < \infty$.*

(A) ***The KSD-B is faithful.*** *If k is a scalar field kernel on S, assume further that $E_q \mathrm{flux}_p < \infty$. Now, $\text{KSD-B}_{p,k}(q) = 0$ only if $p = q$.*

(B) ***The KSD-B detects tight non-convergence.*** *Say $(q_n)_n$ is a tight sequence of distributions on S satisfying $\text{KSD-B}_{p,k}(q_n) \to 0$ as $n \to \infty$. If k is a scalar field kernel, assume further that $\sup_n E_{q_n} \mathrm{flux}_p < \infty$. Then, $q_n \to p$ in distribution.*

Proof. Assume k is a vector field kernel with discrete masses. Say $\text{KSD-B}_{p,k}(q_n) \to 0$ as $n \to \infty$ but $(q_n)_n$ does not converge in distribution to p. Since $(q_n)_n$ is tight, we can, by Prokhorov's theorem, pass to a sub-sequence $(q_{n_k})_k$ that converges in distribution to a distribution q on S. Now, for all $f \in C_{C,vf}(M)$, $\mathcal{T}_p f$ is non-zero on only finitely many points, so $E_q \mathcal{T}_p f = \lim_k E_{q_{n_k}} \mathcal{T}_p f$. Recall that having discrete masses implies we also have $f \in \mathcal{H}_k$. Therefore, $E_{q_{n_k}} \mathcal{T}_p f \leq \|f\|_k \text{KSD-B}_{p,k}(q_{n_k})$, and so $\lim_k E_{q_{n_k}} \mathcal{T}_p f = 0$. We thus have $E_q \mathcal{T}_p f = 0$, which by Lemma B.13 implies $q = p$, a contradiction.

If k is a kernel on S with discrete masses by the same logic as above we have that for all $f \in C_C(M)$, $E_q \mathcal{T}_p f = 0$. By Fatou's lemma we also have

$$E_q \mathrm{flux}_p \leq \liminf_k E_{q_{n_k}} \mathrm{flux}_p < \infty.$$

By Lemma B.13 we again have $q = p$, a contradiction. □

If the support of p is finite – for instance, if p describes only sequences of fixed length – then any sequence of distributions $(q_n)_n$ that sends $\text{KSD-B}_{p,k}(q_n) \to 0$ must be uniformly tight. The reason is that if $\text{KSD-B}_{p,k}(q_n) \to 0$, Lemma B.13 implies we must have $\mathrm{supp}(q_n) \subset \mathrm{supp}(p)$ eventually. Therefore, if $\mathrm{supp}(p)$ is finite, the KSD-B can detect non-convergence in general.

B.3.5. Detecting Convergence and Non-convergence in General

We now establish conditions under which the KSD-B can detect convergence and non-convergence for any sequence of distributions $(q_n)_n$, no matter whether it is tight or not. In this more general setting, we require additional assumptions on p; we must also choose our kernel k more carefully.

Conditions on p For the KSD-B to detect non-convergence to p for any sequence $(q_n)_n$, the stochastic process the KSD-B uses – namely, the continuous time Markov process on sequences defined by the Zanella Stein operator – must in fact converge to p quickly. Intuitively, the KSD-B is evaluating how the expectation of functions in the test set $\mathcal{H}_k$ changes under an infinitesimal step of the stochastic process. If the stochastic process is guaranteed to converge to p quickly, we know the KSD-B can only become very small when q_n is very near to the stationary distribution p. We will therefore require that p satisfies Assumption B.3. Recall that in Corollary B.7, we saw that Assumption B.3 is satisfied for the pHMM, a biological sequence model that is widely successful in practice.

Concretely, to see that an assumption like Assumption B.3 is in fact necessary for detecting non-convergence, consider the following example. We construct a p that does not satisfy Assumption B.3, along with a sequence of distributions $(q_n)_n$ that does not converge to p, such that the KSD-B nonetheless goes to zero.

Proposition B.15. *Say $\alpha_1, \alpha_2, \ldots$ is a decreasing positive sequence such that $\chi(\alpha_L) = L^{-1}|\mathcal{B}|^{-2L}$. For a distribution $\tilde{p}$ on $\mathbb{N}$, for a sequence $X \in S$ with $L = |X|$, let $p(X) \propto |\mathcal{B}|^{-L/2}\alpha_{L+1}\tilde{p}(L/2)$ if L is even and $p(X) \propto |\mathcal{B}|^{-(L-1)/2}\tilde{p}((L-1)/2)$ if L is odd. Say k is a bounded vector field kernel. Then there is a sequence $(q_n)_n$ such that $\text{KSD-B}_{p,k}(q_n) \to 0$ and q_n does not converge to p in distribution.*

The proof is in Section B.3.6. The intuition is that the tail of p is not "uniformly decreasing": $p(X)$ alternates back and forth between large and small values as $|X|$ increases. Since only sequences that differ in length by one letter (rather than two) are related by M, the KSD-B is fooled into detecting convergence. Note that Assumption B.3 demands that the tail of p falls off not only sufficiently slowly, but also uniformly, as $\mathrm{gap}_p(L)$ depends on the difference between the stochastic process's

propensity to delete and to insert a letter. Biologically, this result suggests we should be cautious when applying the KSD-B to distributions over sequences with variable-length tandem repeats: if having a complete repeat motif is much more likely than a partial repeat, it may produce a non-uniform tail for p, breaking Assumption B.3 (Sperling & Li, 2013).

Conditions on k Next, we describe what kinds of kernels we must use to guarantee detection of non-convergence. Recall that for Euclidean KSDs to detect non-convergence, one needs kernels that are heavy tailed, such that their RKHS includes "coercive" functions that have thick tails (Gorham & Mackey, 2017). Intuitively, the situation in sequence space is analogous: we will need RKHSs that include coercive functions.

To motivate our conditions, we consider examples of scalar and vector field kernels that fail to detect non-convergence. First, we show that if we use a scalar field kernel that is bounded, the KSD-B cannot detect convergence.

Proposition B.16. *Say k is a kernel on S that is bounded. Then there exists a distribution p on S that satisfies Assumption B.3, and a sequence of distributions q_n that does not converge to p, such that* $\text{KSD-B}_{p,k}(q_n) \to 0$.

The proof is in Section B.3.6. The distribution p in our construction does not have particularly heavy or non-uniform tails; in fact, p decays exponentially in sequence length, and satisfies $\text{gap}_p(L) \sim L$. The issue therefore is the kernel, not p.

Next we give an example of a vector field kernel that fails to detect non-convergence. The construction is similar in idea to the example in Theorem 6 of Gorham & Mackey (2017).

Proposition B.17. *Let $p(X) \propto e^{-\mu|X|}|\mathcal{B}|^{-|X|}$ for some $\mu > 0$ and k be a vector field kernel such that, for $(X, Y), (X', Y') \in M$ with $|X| = |X'|$,*

$$|k((X, Y), (X', Y'))| \leq C(d_H(X, X') + 1)^{-4-\epsilon}$$

for some $C, \epsilon > 0$ where d_H is the Hamming distance. Then there is a sequence of distributions $(q_n)_n$ in S such that $\text{KSD-B}_{p,k}(q_n) \to 0$ *but q_n doesn't converge to p.*

The proof is in Section B.3.6. Here again, the distribution p decays exponentially in sequence length; the problem is the kernel.

Motivated by these examples, we will require that the RKHS of our kernel includes functions with thick tails. More precisely, we require that there is a $\tilde{f} \in \mathcal{H}_k$ such that $\mathcal{T}_p\tilde{f}$ increases sufficiently quickly with respect to the tail of p.

Assumption B.18. Say p is a distribution on S that satisfies Assumption B.3 with V_p. We assume either:

(A) k is a vector field kernel such that there is a $\tilde{f} \in \mathcal{H}_k$ with $\lim_{|X|\to\infty} \mathcal{T}_p\tilde{f}(X) = \infty$ and

$$\sum_L \frac{\inf_{|X|=L} \mathcal{T}_p\tilde{f}(X)}{\left(\sup_{|X|=L} \text{ins}_p(X)\right) V_p(L+1)} = \infty. \tag{10}$$

(B) k is a kernel on S such that $\text{supp}(p)$ is finite or there is a $\tilde{f} \in \mathcal{H}_k$ with $\lim_{|X|\to\infty} \mathcal{T}_p\nabla\tilde{f}(X) = \infty$ and

$$\sum_L C_L \wedge C_{L+1} = \infty \text{ where } C_L = \frac{\inf_{|X|=L} \mathcal{T}_p\nabla\tilde{f}(X)}{\left(\sup_{|X|=L} \text{flux}_p(X)\right) V_p(L+1)}. \tag{11}$$

To understand this condition, first consider part (A). The denominator in the sum is the maximum propensity for insertions, $\sup_{|X|=L} \text{ins}_p(X)$, multiplied by our Lyapunov function, $V_p(L+1)$. In Section B.4.3, we will construct $\tilde{f}$ such that $\inf_{|X|=L} \mathcal{T}_p\tilde{f}(X) \gtrsim \text{gap}_p(|X|)\tilde{f}(X)$. If $\text{gap}_p(|X|) \gtrsim \text{ins}_p(X)$, then the assumption is satisfied if $\sum_L \frac{\inf_{|X|=L} \tilde{f}(X)}{V_p(L+1)} = \infty$. Recall that V_p must diverge to infinity with increasing L to meet Assumption B.3. Thus, part (A) in essence requires that $\tilde{f}$ has thick tails. Note also that if p has thinner tails, V_p can be smaller, and this assumption is easier to satisfy.

Part (B) is similar to (A), except (1) it uses of the operator $\mathcal{T}_p\nabla$ instead of $\mathcal{T}_p$, (2) the ins_p terms have been replaced by a flux_p term, which can be much larger, and (3) the sum over L takes the minimum of sequential terms. The last difference (3) implies that the sequence $C_1, C_2, \ldots$ cannot alternate between large and small values.

Assumption B.18 is analogous to the coercitivity assumption used in Euclidean KSDs, which similarly requires that $\mathcal{H}_k$ includes functions $\tilde{f}$ such that $\mathcal{T}_p\tilde{f}$ increases sufficiently quickly; see Theorem 8 of Gorham & Mackey (2017) and Theorem 3.2 of Huggins & Mackey (2018).

Non-convergence With Assumption B.3 on p and Assumption B.18 on k, we can prove the the KSD-B detects non-convergence. Our proof strategy is inspired by that of Theorem 8 of Gorham & Mackey (2017). We start by using the fact that the stochastic process converges to p (Theorem B.4) to prove the following lemma, which is similar to Theorem 5 of (Gorham et al., 2019).

Lemma B.19. *Say p is a distribution on S obeying Assumption B.3. If $g \in C_b(S)$ and $g(X) = 0$ for $X \notin \text{supp}(p)$, then there is a $f_g : S \to \mathbb{R}$ such that $f_g(X) = 0$ for $X \notin \text{supp}(p)$, $\mathcal{T}_p \nabla f_g = g - E_p g$, and $f_g(X) \leq C V_p(X) \|g\|_\infty$ for a universal constant C.*

Proof. Recall that by Theorem B.4, we have

$$\|P_t(X) - p\|_{\text{TV}} \lesssim V_p(X) t^{-(2+\epsilon)} + t^{-(1+\epsilon)}.$$

Now since $g \in C_b(S)$,

$$|P_t g(X) - E_p g| \leq \|g\|_\infty \|P_t(X) - p\|_{\text{TV}},$$

so $\int_0^\infty dt |P_t g(X) - E_p g| \leq C' \|g\|_\infty V_p(X)$ for some large enough $C' > 0$. Thus we can define

$$f_g(X) = \int_0^\infty dt \, (E_p g - P_t g(X))$$

with $|f_g|(X) \leq C' \|g\|_\infty V_p(X)$. Because we have absolute integrability, and by Lemma B.1 (A), we can also write

$$\mathcal{L}_p f_g(X) = \int_0^\infty dt \, (-\mathcal{L}_p P_t g(X)) = \int_0^\infty dt \left(-\frac{d}{dt} P_t g(X) \right) = g(X) - E_p g.$$

□

We now show that the KSD-B can detect non-convergence for any sequence of distributions q_n, giving Theorem 5.2.

Theorem B.20. *Say p is a distribution on S obeying Assumption B.3 and k is a scalar or vector field kernel with discrete masses obeying Assumption B.18. Say $(q_n)_n$ is a sequence of distributions on S. If $\text{KSD-B}_{p,k}(q_n) \to 0$ then q_n converges to p in distribution.*

Proof. First note that by Lemma B.13, $\text{supp}(q_n) \subseteq \text{supp}(p)$ for all n eventually. Let $g \in C_b(S)$ with $g(X) = 0$ for $X \notin \text{supp}(p)$ and $\|g\|_\infty \leq 1$, so by Lemma B.19, there is an $f_g : S \to \mathbb{R}$ such that $f_g \leq \tilde{C} V_p$ for some $\tilde{C} > 0$ and $\mathcal{T}_p \nabla f_g = g - E_p g$. We will show that $E_{q_n} g - E_p g = E_{q_n} \mathcal{T}_p \nabla f_g \to 0$, which will be enough to prove the theorem, since it implies q_n converges to p in total variation. We will do so by picking a sequence of $h_m \in \mathcal{H}_k$ such that $\sup_n E_{q_n} |\mathcal{T}_p h_m - \mathcal{T}_p \nabla f_g| \to 0$ as $m \to \infty$. This will show that

$$|E_{q_n} \mathcal{T}_p \nabla f_g| \leq |E_{q_n} \mathcal{T}_p h_m| + E_{q_n} |\mathcal{T}_p h_m - \mathcal{T}_p \nabla f_g| \leq \|h_m\|_k \text{KSD-B}_{p,k}(q_n) + E_{q_n} |\mathcal{T}_p h_m - \mathcal{T}_p \nabla f_g|,$$

which goes to zero as $n \to \infty$ and as $m \to \infty$ slowly enough.

First assume k is a scalar field kernel with discrete masses. For a sequence $v = (v_1, v_2, \dots)$ of numbers $0 \leq v_n \leq 1$ such that v_n is eventually equal to 0, define the vector field on M given by $h_v(X, Y) = v_{|X| \vee |Y|} \nabla f_g(X, Y)$. Since v is eventually zero, $h_v(X, Y) > 0$ for only a finite set of $X, Y \in S$; since k has discrete masses, $h_v \in \mathcal{H}_k$. As well,

$$\begin{aligned}
\mathcal{T}_p h_v(X) &= \sum_{YMX} T_{p, X \to Y} v_{|X| \vee |Y|} \nabla f_g(X, Y) \\
&= v_{|X|} \sum_{YMX \mid |Y| \leq |X|} T_{p, X \to Y} \nabla f_g(X, Y) + v_{|X|+1} \sum_{YMX \mid |Y| = |X|+1} T_{p, X \to Y} \nabla f_g(X, Y) \\
&= v_{|X|} \mathcal{T}_p \nabla f_g(X) + (v_{|X|+1} - v_{|X|}) \sum_{YMX, |Y| = |X|+1} T_{p, X \to Y} \nabla f_g(X, Y).
\end{aligned}$$

The first term is a better and better approximation of $\mathcal{T}_p \nabla f_g$ as $v \to 1$. We will bound the second term using Assumption B.18 (A) and the fact

$$\left| \sum_{YMX, |Y| = |X|+1} T_{p, X \to Y} \nabla f_g(X, Y) \right| \leq 2 \tilde{C} V_p(|X| + 1) \text{ins}_p(X).$$

Assume k satisfies Assumption B.18 (A). Let $\tilde{f} \in \mathcal{H}_k$ satisfy Equation 10 and have $\mathcal{T}_p\tilde{f}(X) \to \infty$ as $|X| \to \infty$. There is thus a $\zeta \in \mathbb{R}$ such that $\mathcal{T}_p\tilde{f}(X) + \zeta > 0$ for all $X \in S$. Now call $\Delta v_L = |v_{L+1} - v_L|$ and $R_L := \frac{V_p(L+1)\sup_{|X|=L}\text{ins}_p(X)}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta}$, so,

$$\begin{aligned}
E_{q_n}|\mathcal{T}_p h_v - \mathcal{T}_p\nabla f_g| \leq& E_{q_n}[(1-v_{|X|})|\mathcal{T}_p\nabla f_g|] + E_{q_n}\left[\Delta v_{|X|} 2\tilde{C}V_p(|X|+1)\text{ins}_p(X)\right] \\
\leq& 2\|g\|_\infty E_{q_n}[1-v_{|X|}] + 2\tilde{C}E_{q_n}\left[\left(\mathcal{T}_p\tilde{f}+\zeta\right)\Delta v_{|X|}\frac{V_p(|X|+1)\text{ins}_p(X)}{\mathcal{T}_p\tilde{f}+\zeta}\right] \\
\leq& 2E_{q_n}[1-v_{|X|}] + 2\tilde{C}E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\sup_L(\Delta v_L R_L) \\
\leq& 2E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta} + 2\tilde{C}E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\sup_L(\Delta v_L R_L) \\
=& E_{q_n}\left[\mathcal{T}_p\tilde{f}+\zeta\right]\left(2\sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta} + 2\tilde{C}\sup_L(\Delta v_L R_L)\right) \\
\leq& \left(\|\tilde{f}\|_k\text{KSD-B}_{p,k}(q_n)+\zeta\right)\left(2\sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta} + 2\tilde{C}\sup_L(\Delta v_L R_L)\right) \\
\lesssim& \sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta} + \sup_L(\Delta v_L R_L).
\end{aligned}$$

By assumption $\mathcal{T}_p\tilde{f}+\zeta \to \infty$ and $\sum_L R_L^{-1} = \infty$. For $\epsilon, L' > 0$ define $v_L^{\epsilon,L'} = 1$ for $L \leq L'$ and $\Delta v_L^{\epsilon,L'} = \epsilon R_L^{-1} \wedge (v_L^{\epsilon,L'})$ for $L' \geq L$. By assumption $\sum_L R_L^{-1} = \infty$ so $v^{\epsilon,L'}$ is eventually 0. We thus have $\sup_L\left(\Delta v_L^{\epsilon,L'} R_L\right) = \epsilon$ and $\sup_L \frac{1-v_L^{\epsilon,L'}}{\inf_{|X|=L}\mathcal{T}_p\tilde{f}(X)+\zeta} \leq \frac{1}{\inf_{|X|\geq L}\mathcal{T}_p\tilde{f}+\zeta}$. By our assumption that $\mathcal{T}_p\tilde{f} \to \infty$, both of these quantities go to 0 as $L' \to \infty$ and $\epsilon \to 0$.

Now assume k is a vector field kernel with discrete masses obeying Assumption B.18 (B). The case that $\text{supp}(p)$ is finite was shown in Proposition 5.1 so assume $\text{supp}(p)$ is infinite. The proof is very similar. This time, for a sequence $v = (v_1, v_2, \dots)$ of decreasing numbers $0 \leq v_n \leq 1$ such that v_n is eventually equal to 0, define the function on S, $h_v(X) = v_{|X|}f_g(X)$. Since v is eventually 0, and since k has discrete masses, $h_v \in \mathcal{H}_k$. Then, by similar reasoning to the previous case,

$$\begin{aligned}
\mathcal{T}_p\nabla h_v(X) =& v_{|X|}\mathcal{T}_p\nabla f_g(X) + (v_{|X|+1} - v_{|X|})\sum_{YMX,|Y|=|X|+1} T_{p,X\to Y}\nabla f_g(X,Y) \\
&+ (v_{|X|-1} - v_{|X|})\sum_{YMX,|Y|=|X|-1} T_{p,X\to Y}\nabla f_g(X,Y).
\end{aligned}$$

Note that since V_p is increasing, the sum of the later two terms is upper bounded by

$$2\tilde{C}\tilde{\Delta}v_L V_p(|X|+1)\text{flux}_p(X),$$

defining $\tilde{\Delta}v_L = |v_{L+1} - v_L| \vee |v_L - v_{L-1}|$. Now call $\tilde{R}_L := \frac{V_p(L+1)\sup_{|X|=L}\text{flux}_p(X)}{\inf_{|X|=L}\mathcal{T}_p\nabla\tilde{f}(X)+\zeta}$, so,

$$\begin{aligned}
E_{q_n}|\mathcal{T}_p\nabla h_v - \mathcal{T}_p\nabla f| \leq& E_{q_n}[(1-v_{|X|})|\mathcal{T}_p\nabla f_g|] + E_{q_n}\left[\tilde{\Delta}v_{|X|}2\tilde{C}V_p(|X|+1)\text{flux}_p(X)\right] \\
\leq& 2E_{q_n}\left[\mathcal{T}_p\nabla\tilde{f}+\zeta\right]\left(\sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\nabla\tilde{f}(X)+\zeta} + 2\tilde{C}\sup_L\left(\tilde{\Delta}v_L\tilde{R}_L\right)\right) \\
\leq& \left(\|\tilde{f}\|_k\text{KSD-B}_{p,k}(q_n)+\zeta\right)\left(2\sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\nabla\tilde{f}(X)+\zeta} + 2\tilde{C}\sup_L\left(\tilde{\Delta}v_L\tilde{R}_L\right)\right) \\
\lesssim& \sup_L\frac{1-v_L}{\inf_{|X|=L}\mathcal{T}_p\nabla\tilde{f}(X)+\zeta} + \sup_L\left(\tilde{\Delta}v_L\tilde{R}_L\right).
\end{aligned}$$

By assumption $\mathcal{T}_p\tilde{f}+\zeta \to \infty$ and $\sum_L \tilde{R}_L^{-1} \wedge \tilde{R}_{L+1}^{-1} = \infty$. For $\epsilon, L' > 0$ define $v_L^{\epsilon,L'} = 1$ for $L \leq L'$ and $v_L^{\epsilon,L'} = v_{L-1}^{\epsilon,L'} - \epsilon\tilde{R}_{L-1}^{-1} \wedge \tilde{R}_L^{-1} \wedge (v_{L-1})$ for $l \geq L$. Thus $\tilde{\Delta}v_L \leq \epsilon\tilde{R}_L^{-1}$. By assumption $\sum_L \tilde{R}_L^{-1} \wedge \tilde{R}_{L+1}^{-1} = \infty$ so $v^{\epsilon,L'}$ is

eventually 0. We thus have $\sup_L \left(\tilde{\Delta} v_L^{\epsilon, L'} \tilde{R}_L\right) = \epsilon$ and $\sup_L \frac{1 - v_{|X|}^{\epsilon, L'}}{\inf_{|X|=L} \mathcal{T}_p \nabla \tilde{f}(X) + \zeta} \leq \frac{1}{\inf_{|X| \geq L} \mathcal{T}_p \tilde{f} + \zeta}$. By our assumption that $\mathcal{T}_p \nabla \tilde{f} \to \infty$, both of these quantities go to 0 as $L' \to \infty$ and $\epsilon \to 0$. □

Convergence Finally, we prove that the KSD-B can detect convergence, giving Proposition 5.3.

Proposition B.21. *Say k is a vector field kernel and $p, q_1, q_2, \ldots$ are p, k-integrable distributions on S. Call $A(X) = \sum_{YMX} T_{p,Y \to X} \sqrt{k((X,Y),(X,Y))}$.*

$$\sum_X |p(X) - q_n(X)| A(X) \to 0 \implies \text{KSD-B}_{p,k}(q_n) \to 0.$$

Proof. Say $f \in \mathcal{H}_k$. Now,

$$|E_p \mathcal{T}_p f - E_{q_n} \mathcal{T}_p f| \leq \|f\|_k \sum_X |p(X) - q_n(X)| \sum_{YMX} T_{p,Y \to X} \sqrt{k((X,Y),(X,Y))}$$

which proves the result. □

One implication of this result is that when the kernel is large, in the sense that $A(X)$ is big, convergence will be harder to detect. To ensure that the KSD-B can reliably detect convergence, we will want to choose kernels that are not very large.

B.3.6. Proofs of Examples

Here, we give the proofs of the examples discussed in the previous section.

Proposition B.22. *(Proposition B.15) Say $\alpha_1, \alpha_2, \ldots$ is a decreasing positive sequence such that $\chi(\alpha_L) = L^{-1} |\mathcal{B}|^{-2L}$. For a distribution $\tilde{p}$ on $\mathbb{N}$, for a sequence $X \in S$ with $L = |X|$, let $p(X) \propto |\mathcal{B}|^{-L/2} \alpha_{L+1} \tilde{p}(L/2)$ if L is even and $p(X) \propto |\mathcal{B}|^{-(L-1)/2} \tilde{p}((L-1)/2)$ if L is odd. Say k is a bounded vector field kernel. Then there is a sequence $(q_n)_n$ such that $\text{KSD-B}_{p,k}(q_n) \to 0$ and q_n does not converge to p in distribution.*

Proof. We will consider a sequence of distributions indexed by even sequence lengths L, that is, q_L for $L \in \{2, 4, 6, \ldots\}$. Define $\tilde{q}_L = p \mathbb{1}_{|X| > L}$ and $q_L = \tilde{q}_L / \sum_X \tilde{q}_L(X)$. Call $q_L(L') = q_L(X)$ for any $|X| = L'$. Call $N_L = \{(X,Y) \in M \mid |X| = L+1, |Y| = L\}$. The terms of the sum in Equation 8 are non-zero only for $(X,Y) \in N_L$. Thus,

$$\begin{aligned}
\text{KSD-B}_{p,k}(q_L)^2 &= \left(\sup_{\|f\|_k \leq 1} \sum_{(X,Y) \in N_L} q_L(X) T_{p,X \to Y} f(X,Y) \right)^2 \\
&= q_L(L+1)^2 \left(\sup_{\|f\|_k \leq 1} \left(f \middle| \sum_{(X,Y) \in N_L} T_{p,X \to Y} k_{(X,Y)} \right)_k \right)^2 \\
&= q_L(L+1)^2 \left\| \sum_{(X,Y) \in N_L} T_{p,X \to Y} k_{(X,Y)} \right\|_k^2 \\
&= q_L(L+1)^2 \sum_{(X,Y) \in N_L} \sum_{(X',Y') \in N_L} T_{p,X \to Y} T_{p,X' \to Y'} k((X,Y),(X',Y')).
\end{aligned}$$

If $(X,Y) \in N_L$, then $T_{p,X \to Y} \leq (L+1) \chi(\alpha_{L+1}) = |\mathcal{B}|^{-2(L+1)}$; this comes from the fact that the maximum number of mutations which can take X to Y is $L+1$ (corresponding to the case where $X = L \times A$ and $Y = (L+1) \times A$, where $A \in \mathcal{B}$, i.e. X and Y are homopolymers). Thus, if k is bounded by a number $C > 0$,

$$\text{KSD-B}_{p,k}(q_L)^2 \leq q_L (L+1)^2 |\mathcal{B}|^{2(L+1)} |\mathcal{B}|^{-4(L+1)} C \leq |\mathcal{B}|^{-2(L+1)} C \to 0.$$

□

Proposition B.23. *(Proposition B.16) Say k is a kernel on S that is bounded, such that $k(X,Y) \leq N < \infty$ for all $X, Y \in S$. Then there exists a distribution p on S that satisfies Assumption B.3, and a sequence of distributions q_n that does not converge to p, such that* $\text{KSD-B}_{p,k}(q_n) \to 0$.

Proof. Note that since k is bounded by N, we have for all $X \in S$ and $f \in \{f \in \mathcal{H}_k : \|f\|_k \leq 1\}$, $f(X) = (f|k_X)_k \leq \|f\|_k \sqrt{k(X,X)} \leq N$. Thus $\|f\|_\infty \leq N$ for all $f \in \{f \in \mathcal{H}_k : \|f\|_k \leq 1\}$. So, to prove the result, it is sufficient to find p and q_n such that $\sup_{\|f\|_\infty \leq 1} E_{q_n} \mathcal{T}_p \nabla f \to 0$.

Let p be the distribution supported on $\{\emptyset, A, AA, AAA, \dots\}$ for $A \in \mathcal{B}$ with $p(L) = p(L \times A) = 2^{-(L+1)}$ for any number L. Note that this distribution satisfies $\text{gap}_p(L) \sim L$ (as discussed in Section B.2.5). As a consequence, Assumption B.3 is met.

Now, define $r = \chi\left(\frac{p(L)}{p(L-1)}\right) = \chi(1/2)$ for any L and $\tilde{r} = \chi\left(\frac{p(L-1)}{p(L)}\right) = \chi(2) = 2\chi(1/2) = 2r > r$ for any L. Let $\tilde{r}_0 = 0$. Say q is a distribution supported on finitely many $\{\emptyset, A, AA, AAA, \dots\}$, and f is a function on S with $\|f\|_\infty \leq 1$,

$$\begin{aligned}
E_q \mathcal{T}_p \nabla f &= \sum_{L=0}^{\infty} q(L) \left((L+1)r(f(L+1) - f(L)) + L\tilde{r}(f(L-1) - f(L))\right) \\
&= \sum_{L=0}^{\infty} f(L) \Big(q(L+1)(L+1)\tilde{r} + q(L-1)Lr \\
&\qquad - q(L)\left(L\tilde{r} + (L+1)r\right) \Big) \\
&= \sum_{L=0}^{\infty} f(L) \Big(q(L+1)(L+1)\tilde{r} - q(L)L\tilde{r} \\
&\qquad + q(L-1)Lr - q(L)(L+1)r \Big).
\end{aligned}$$

Let $\tilde{q}_{m,n}(L) = L^{-1}$ for $m \leq L < n$ and $\tilde{q}_{m,n}(L) = 0$ for $L \geq n$ and $L < m$. Now let $q_{m,n} = \tilde{q}_{m,n}/Z_{m,n}$ where $Z_{m,n} = \sum_{L=m}^{n-1} L^{-1}$ which goes to ∞ as $n \to \infty$. Thus,

$$\begin{aligned}
E_{q_{m,n}} \mathcal{T}_p \nabla f =& f(m-1)Z_{m,n}^{-1}\tilde{r} - f(n-1)Z_{m,n}^{-1}\tilde{r} \\
&+ \sum_{L=m+1}^{n-1} f(L)\left(q_{m,n}(L-1)Lr - q_{m,n}(L)(L+1)r\right) \\
&- f(m)q_{m,n}(m)(m+1)r + f(n)q_{m,n}(n-1)nr \\
=& Z_{m,n}^{-1}\left(\tilde{r}f(m-1) - \tilde{r}f(n) - rf(m)\frac{m+1}{m} + rf(n)\frac{n}{n-1}\right) \\
&+ \sum_{L=m+1}^{n-1} q_{m,n}(L-1)f(L)r\left(L - \frac{L-1}{L}(L+1)\right) \qquad (12) \\
\leq& 6\tilde{r}{Z_{m,n}}^{-1} + r\sum_{L=m+1}^{n-1} q_{m,n}(L-1)L\left|1 - \frac{L^2-1}{L^2}\right| \\
=& 6\tilde{r}{Z_{m,n}}^{-1} + r\sum_{L=m+1}^{n-1} q_{m,n}(L-1)L^{-1} \\
\leq& 6\tilde{r}{Z_{m,n}}^{-1} + r(m+1)^{-1}
\end{aligned}$$

This expression goes to 0 as $n, m \to \infty$.

□

Proposition B.24. *(Proposition B.17) Let $p(X) \propto e^{-\mu|X|}|\mathcal{B}|^{-|X|}$ for some $\mu > 0$ and k be a vector field kernel such that,*

for $(X,Y),(X',Y') \in M$ *with* $|X| = |X'|$,

$$|k((X,Y),(X',Y'))| \leq C(d_H(X,X')+1)^{-4-\epsilon}$$

for some $C, \epsilon > 0$ *where* d_H *is the Hamming distance. Then there is a sequence of distributions* $(q_n)_n$ *in* S *such that* $\text{KSD-B}_{p,k}(q_n) \to 0$ *but* q_n *doesn't converge to* p.

Proof. First note that for $(X,Y) \in M$, calling $\chi\left(e^{\mu}|\mathcal{B}|\right) = c$, $T_{p,X\to Y} \leq c(|X|+1)$. For distinct points $X_1,\ldots,X_N \in \mathcal{B}^L$ let $q = \frac{1}{N}\sum_{n=1}^N \delta_{X_n}$. Call $R = \min_{n\neq m} d_H(X_n,X_m) > 0$. Then by equation 4,

$$\begin{aligned}
\text{KSD-B}_{p,k}(q)^2 \leq & \frac{c^2(L+1)^2}{N^2}\sum_{n=1}^N\sum_{m=1}^N \sum_{Y MX_n}\sum_{Y' MX_m} |k((X_n,Y),(X_m,Y'))| \\
= & \frac{c^2(L+1)^2}{N^2}\Bigg(\sum_{n=1}^N \sum_{Y MX_n}\sum_{Y' MX_n} |k((X_n,Y),(X_n,Y'))| \\
& \qquad + \sum_{n\neq m}\sum_{Y MX_n}\sum_{Y' MX_m} |k((X_n,Y),(X_m,Y'))|\Bigg) \\
\lesssim & \frac{(L+1)^2}{N^2}\left(NL^2 + N^2L^2R^{-(4+\epsilon)}\right) \\
= & O\left(L^4\left(N^{-1} + R^{-(4+\epsilon)}\right)\right).
\end{aligned}$$

We now set $R_L = L(1-|\mathcal{B}|^{-1})/20$ and pick, for each L, $X_1,\ldots,X_{N_L} \in \mathcal{B}^L$ to be the largest set of sequences such that $\min_{n\neq m} d_H(X_n,X_m) > R_L$. We will show $L^4 N_L^{-1} \to 0$, so that we will have $L^4\left(N_L^{-1} + R_L^{-(4+\epsilon)}\right) \to 0$ and the proof will be complete.

For $X \in \mathcal{B}^L$, $r > 0$, define the Hamming ball $B(X,r) = \{Y \in \mathcal{B}^L \mid d_H(X,Y) \leq r\}$. Thus $\mathcal{B}^L = \cup_n B(X_n,R_L)$, otherwise we could add another sequence to $(X_n)_n$. Thus $|\mathcal{B}|^L \leq \sum_n |B(X_n,R_L)| = N_L|B(X_1,R_L)|$. Now, consider the Hamming distance from X_1 to a sequence drawn uniformly at random from $\mathcal{B}^L$. This distance is a random variable Z distributed as a Binomial with parameters L and $1-|\mathcal{B}|^{-1}$. Then $N_L^{-1} \leq |B(X_1,R_L)|/|\mathcal{B}|^L = P(Z \leq R_L)$. On the other hand, calling $\gamma = 1-|\mathcal{B}|^{-1}$ and $t = -\log\left(\frac{R_L}{L\gamma}\right) = \log 20$, and using the moment generating function of the Binomial distribution,

$$\begin{aligned}
P(Z \leq R_L) = & P(e^{-tZ} \geq e^{-tR_L}) \\
\leq & e^{tR_L}Ee^{-tZ} \\
= & e^{tR_L}\left(\gamma e^{-t} + (1-\gamma)\right)^L \\
= & e^{tR_L}\left(1+\gamma(e^{-t}-1)\right)^L \\
\leq & \exp\left(tR_L + L\gamma(e^{-t}-1)\right) \\
= & \exp\left(R_L(1+t) - L\gamma\right) \\
= & \exp\left(-L\gamma\left(1 - \frac{1}{20}\left(1+\log 20\right)\right)\right) \\
\leq & \exp\left(-\frac{1}{2}L\gamma\right).
\end{aligned}$$

Thus, $N_L \geq e^{\frac{1}{2}L(1-|\mathcal{B}|^{-1})}$ so that $L^4N_L^{-1} \to 0$ as $L \to \infty$. □

B.3.7. Efficient Approximate Kernelized Stein Discrepancies

In this section prove Proposition 6.1, which establishes an efficient approximation for the KSD-B.

Proposition B.25. *(Proposition 6.1) Let* p *be a distribution on* S*, and* $(q_n)_n$ *a sequence of distributions on* S *with* $\sup_n E_{q_n}\text{flux}_p < \infty$*. Say* k *is a bounded vector field kernel. Let* $(N_{n,X})_{X\in S,n}$ *be a family of numbers. For each* X, n*, let*

$(Y^n_{X,m})_{m=1}^{N_{n,X}}$ be a set of iid samples, each drawn by taking a single step of a Markov chain with the transition matrix $K_{X\to Y}$ initialized at X. Define the approximate KSD-B*,*

$$\widehat{\text{KSD-B}}^n_{p,k}(q_n)^2 = E_{X,X'\sim q}\text{flux}_p(X)\text{flux}_p(X')\frac{1}{N_{n,X}N_{n,X'}}\sum_{m,m'}k((X,Y^n_{X,m}),(X',Y^n_{X',m'})).$$

If $N_{n,X}/(\log(n)+|X|)\to\infty$ then almost surely

$$\left|\text{KSD-B}_{p,k}(q_n) - \widehat{\text{KSD-B}}^n_{p,k}(q_n)\right| \to 0.$$

Proof. Call $p(Y|X) = K_{X\to Y}$. Sample $(Y^n_{X,m})_{m=1}^{N_{n,X}}$ iid from $p(Y|X)$ for all X, n. Call $\hat{p}_n(Y|X) = \frac{1}{N_{n,X}}\sum_{m=1}^{N_{n,X}}\delta_{Y^n_{X,m}}$. We have

$$E_{X\sim q_n}\text{flux}_p(X)E_{\hat{p}_n(Y|X)}\sqrt{k((X,Y),(X,Y))} < \infty$$

by assumption since k is bounded. Thus the functional $\phi_n : \mathcal{H}_k \to \mathbb{R} \mid f \mapsto E_{X\sim q_n}\text{flux}_p(X)E_{\hat{p}_n(Y|X)}f(X,Y)$ is bounded, and is thus in $\mathcal{H}_k$ by the Reisz representation theorem. Applying the definition of $K_{X\to Y}$, we can derive, as in the proof of Proposition B.8,

$$\widehat{\text{KSD-B}}^n_{p,k}(q_n) = \|\phi_n\|_k = \sup_{f\in\mathcal{F}}E_{X\sim q_n}\text{flux}_p(X)E_{\hat{p}_n(Y|X)}f(X,Y).$$

We will show that $\sup_X\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}}\to 0$ as $n\to\infty$ almost surely below; for now, assume this is the case. Say k is bounded by a number c^2, so $\|f\|_k\le 1$ implies $\|f\|_\infty \le \sup_{(X,Y)\in M}|(f|k_{(X,Y)})_k| \le c$. Thus, using the definition of the total variation metric as an integral probability metric over bounded functions,

$$\begin{aligned}
&\left|\text{KSD-B}_{p,k}(q_n) - \widehat{\text{KSD-B}}^n_{p,k}(q_n)\right| \\
&\quad\le \sup_{\|f\|_k\le 1}E_{X\sim q_n}\text{flux}_p(X)\left|E_{p(Y|X)}f(X,Y) - E_{\hat{p}_n(Y|X)}f(X,Y)\right| \\
&\quad\le cE_{X\sim q_n}\text{flux}_p(X)\,\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}} \\
&\quad\le c\left(E_{X\sim q_n}\text{flux}_p(X)\right)\sup_X\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}} \\
&\quad\to 0.
\end{aligned}$$

Now we will show that $\sup_X\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}}\to 0$ almost surely. Let $X\in\text{supp}(p)$ and call $M|_X = \{Y\in S \mid YMX\}$. Call $\mathcal{F}_X = \{h : M|_X\to\{-1,1\}\}$ so

$$\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}} = \frac{1}{2}\sum_{Y\in M|_X}|p(Y|X)-\hat{p}_n(Y|X)| = \frac{1}{2}\max_{h\in\mathcal{F}_X}E_{p(Y|X)}h(Y) - E_{\hat{p}_n(Y|X)}h(Y).$$

Note for each $h\in\mathcal{F}_X$, $E_{\hat{p}_n(Y|X)}[h(Y)-E_{p(Y'|X)}h(Y')]$ is an average of $N_{n,X}$ mean-zero iid random variables that take values $[-2,2]$ and are therefore sub-Gaussian; the average is thus also a sub-Gaussian random variable, with variance-proxy $C'/\sqrt{N_{n,X}}$ for some C' (Vershynin, 2020, Prop. 2.6.1). Then by a union bound, since $|\mathcal{F}_X|\le 2^{C''|X|}$ for some $C''>0$,

$$P(\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}} > \epsilon_n) \le e^{C_1|X|}\exp\left(-C_2N_{n,X}\epsilon_n^2\right)$$

for some constants $C_1, C_2 > 0$. Pick a sequence of positive numbers $\epsilon_1,\epsilon_2,\dots$ and choose $N_{n,X}$ such that $N_{n,X}/\left(|X|+(\log n)\right)\to\infty$. If ϵ_n decreases slowly enough, then eventually $C_2N_{n,X}\epsilon_n^2 \ge (C_1+\log|\mathcal{B}|+1)|X|+2\log n$, so,

$$\begin{aligned}
\sum_{X,n}P(\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}} > \epsilon_n) &\le \sum_n\sum_X e^{C_1|X|}\exp\left(-C_2N_{n,X}\epsilon_n^2\right) \\
&\lesssim \sum_n\sum_L\sum_{|X|=L}|\mathcal{B}|^{-L}e^{-L}\exp\left(-2\log n\right) \\
&\lesssim \sum_n n^{-2} < \infty.
\end{aligned}$$

By the Borel-Cantelli lemma, the probability that $\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}} > \epsilon_n$ for infinitely many X, n is 0. Thus, with probability 1, as $n\to\infty$,

$$\sup_X\|p(Y|X)-\hat{p}_n(Y|X)\|_{\text{TV}}\to 0$$

□

B.4. Designing Kernels for the KSD-B

In this section we design practical kernels for the KSD-B. Our theoretical results suggest that we want kernels that (1) have discrete masses, (2) have thick tails (Assumption B.18) and (3) are not very large (Proposition B.21). We also want our kernels to capture a sensible biological notion of sequence similarity, so that sequences which are closer together, as judged by the kernel, are more likely to be functionally similar and evolutionarily related.

Section B.4.1 reviews existing results on scalar field sequence kernels that (a) have discrete masses and (b) capture biological notions of sequence similarity. Section B.4.2 develops vector field kernels with the same virtues. Section B.4.3 adds thick tails; it relies in part on a combinatoric analysis of alignment kernels, which is deferred to Section B.4.5. Section B.4.4 confirms that the proposed kernels are not too large.

B.4.1. Scalar Field Biological Sequence Kernels with Discrete Masses

In this section we review some scalar field kernels for biological sequences with discrete masses, proposed in Amin et al. (2023).

Position-wise Comparison Kernels We start by introducing kernels that compare sequences position-by-position. Such kernels have strong biological justification for many problems; for instance, amino acids at the same position in related proteins are likely to have similar biological functions. A standard position-wise measure of sequence similarity is the Hamming distance $d_H(X, Y)$ which counts the number of positions l at which $X_{(l)}$ does not match $Y_{(l)}$; if Y is longer than X, positions past the length of X are counted as mismatches (i.e. we treat each sequence as ending with an infinite tail of stop symbols \$). We consider two kernels that rely on the Hamming distance to measure sequence similarity: the exponential Hamming kernel (Exp-H),

$$k_{\mathrm{Exp-H}}(X, Y) = \exp(-\lambda d_H(X, Y))$$

with $\lambda > 0$, and the inverse multiquadric Hamming kernel (IMQ-H),

$$k_{\mathrm{IMQ-H}}(X, Y) = (C + d_H(X, Y))^{-\beta}$$

with $C, \beta > 0$. Amin et al. (2023) showed that both have discrete masses.

Theorem B.26. *(Theorem 21 and Example 6 of Amin et al. (2023))* $k_{\mathrm{Exp-H}}$ *and* $k_{\mathrm{IMQ-H}}$ *have discrete masses.*

Alignment Kernels We next consider alignment kernels, which compare sequences based on pairwise alignments (Haussler, 1999). Position-wise comparison kernels will judge two sequences to be very different even if they differ by a single insertion; in many biological settings, however, a small insertion is unlikely to change a sequence's biological function dramatically. Alignment kernels, by contrast, consider sequences that differ by a small number of inserted or deleted letters to be similar.

To define alignment kernels, we first introduce two simpler kernels that will later be combined. The first is for comparing letters; the second is for penalizing insertions. To compare letters, let $k_s(X, Y) = \sigma^{-1}|\mathcal{B}|\delta_X(Y) \times \mathbb{1}(|X| = 1)\mathbb{1}(|Y| = 1)$; note that k_s is only non-zero if X and Y both consist of the same single letter. To penalize insertions, let $k_I(X, Y) = \exp\left(-\mu(|X| + |Y|) - \Delta\mu\left(\mathbb{1}(|X| \geq 1) + \mathbb{1}(|Y| \geq 1)\right)\right)$ for $0 < \mu < \infty$ and $0 \leq \Delta\mu \leq \infty$. Sequences compared by k_I are interpreted as insertions, and penalized with an insertion start penalty $\Delta\mu$ and insertion length penalty μ. We also consider a variant without the start penalty, $\tilde{k}_I(X, Y) = \exp\left(-\mu(|X| + |Y|)\right)$.

The alignment kernel sums over all possible pairwise alignments of two sequences, and for each pairwise alignment scores matched positions with k_s and insertions with k_I. It can be written as,

$$\tilde{k}_{\mathrm{ali}}(X, Y) = \sum_{l, X^{(1)}+\cdots+X^{(2l+1)}=X, Y^{(1)}+\cdots+Y^{(2l+1)}=Y} k_I(X^{(1)}, Y^{(1)}) \prod_{i=1}^{l} k_s(X^{(2i)}, Y^{(2i)}) k_I(X^{(2i+1)}, Y^{(2i+1)}) \quad (13)$$

where the sum is over all numbers l and partitions of X and Y into $2l + 1$ substrings. The even substrings $X^{(2)}, X^{(4)}, \ldots, Y^{(2)}, Y^{(4)}, \ldots$ correspond to the matched positions, while the odd substrings $X^{(1)}, X^{(3)}, \ldots, Y^{(1)}, Y^{(3)}, \ldots$ are the intervening insertions. (See e.g. Amin et al. (2023) for a detailed explanation.) We

also define the local alignment kernel, which does not penalize the creation of an insertion at the beginning or end of the alignment,

$$\tilde{k}_{\mathrm{la}}(X,Y) = \sum \tilde{k}_I(X^{(1)},Y^{(1)}) \left(\prod_{i=1}^{l-1} k_s(X^{(2i)},Y^{(2i)}) k_I(X^{(2i+1)},Y^{(2i+1)}) \right) k_s(X^{(2l)},Y^{(2l)}) \tilde{k}_I(X^{(2l+1)},Y^{(2l+1)}).$$

Amin et al. (2023) showed that k_{ali} and k_{la} have discrete masses, provided the hyperparameters $\Delta\mu$ and $\zeta = 2\mu - \log\sigma + \log|\mathcal{B}|$ are set appropriately.

Theorem B.27. *(Theorems 23 and 25 of Amin et al. (2023)) $\tilde{k}_{\mathrm{ali}}$ and $\tilde{k}_{\mathrm{la}}$ have discrete masses if and only if $\Delta\mu = \infty$; or $\Delta\mu > 0$ and $\zeta \geq \log|\mathcal{B}|$; or $\Delta\mu = 0$ and $\zeta > \log|\mathcal{B}|$.*

It is common in practice to work with tilted versions of the alignment kernel, for instance to normalize the kernel. To enable our theoretical analysis of the alignment kernel's tails, we will use the tilting $\tilde{A}(X) = \exp(\mu|X|)$, which gives the kernels $k_{\mathrm{ali}}(X,Y) = \tilde{A}(X)\tilde{k}_{\mathrm{ali}}(X,Y)\tilde{A}(Y)$ and $k_{\mathrm{la}} = \tilde{A}(X)\tilde{k}_{\mathrm{la}}(X,Y)\tilde{A}(Y)$. Note that this particular tilting is equivalent to setting $\mu = 0$. Thus k_{ali} and k_{la} in effect have only two parameters, $\Delta\mu$ and ζ, since the choice of ζ determines σ.

Infinite kmer Spectrum Kernels Next we consider kmer spectrum kernels, which compare sequences based on how many times substrings (kmers) occur in each sequence (Leslie et al., 2004). Like alignment kernels, kmer spectrum kernels judge two sequences to be similar even if they differ by insertions or deletions rather than just substitutions. In fact, Amin et al. (2023) showed that for a particular tilting and parameter choice, the local alignment kernel is equivalent to a kmer spectrum kernel. The kmer counts of a sequence are its features under the local alignment kernel.

Proposition B.28. *(Proposition 27 in Amin et al. (2023)) Say $\Delta\mu = \infty$ and $\zeta = 0$. For $X, Z \in S$ call $\phi_Z(X)$ the number of times Z appears in X. Then, $k_{\mathrm{la}}(X,Y) = \sum_{Z\in S} \phi_Z(X)\phi_Z(Y)$. This kernel has discrete masses.*

We refer to this special case of the local alignment kernel, with $\Delta\mu = \infty$ and $\zeta = 0$, as an *infinite* kmer spectrum kernel k_{ISK}, since it sums over an infinite number of kmers $Z \in S$ (typical kmer spectrum kernels just consider all kmers shorter than a given length, see Leslie et al. (2004)).

Embedding Kernels Finally we consider embedding kernels. These kernels are built using a learned embedding of sequences into Euclidean space $F : S \to \mathbb{R}^D$. We compare embedded sequences using a translation invariant kernel $k_E(z,z') = \Psi(z-z')$ with Ψ a positive continuous function on $\mathbb{R}^D$ that has a strictly positive Fourier transform. The embedding kernel is defined as $k_{F,\mathrm{Emb}}(X,Y) = k_E(F(X),F(Y))$. In this paper we always use Unirep64 as F, for which $D = 64$ (Alley et al., 2019).

Now we look at when embedding kernels have discrete masses. Amin et al. (2023) proved that k has discrete masses if the image of F doesn't have accumulation points.

Proposition B.29. *(Proposition 31 in Amin et al. (2023)) k_E has discrete masses if and only if $F(S)$ has no accumulation points, that is, there is no $X \in S$ such that $F(X)$ is in the closure of $F(S \setminus \{X\})$.*

Amin et al. (2023) suggests that F from regularized representation learning methods may struggle to avoid accumulation points in their image as S is infinite and F outputs representations with small norm. However they suggest this can be solved by rescaling embeddings so that longer sequences have embeddings with larger norm. They demonstrate this theoretically for a random embedding:

Proposition B.30. *Consider an embedding $\tilde{F}$ where each $\tilde{F}(X)$ for $X \in S$ is drawn from the uniform distribution on the sphere, $\{x \in \mathbb{R}^D \mid \|x\| \leq 1\}$. Then, a kernel using the rescaled embedding $F(X) = |\mathcal{B}|^{(1+\epsilon)|X|/D}\tilde{F}(X)$, for $\epsilon > 0$, has discrete masses almost surely.*

To make it likely that our kernels have discrete masses, we use a rescaled embedding below $F_{\mathrm{rescaled}}(X) = |\mathcal{B}|^{1.1\times|X|/64}F(X)$.

By picking different k_E we can build different Embedding kernels. For example we define 1) the IMQ embedding kernel $k_{F,IMQ,\sigma}(X,Y) = (1 + \|F_{\mathrm{rescaled}}(X) - F_{\mathrm{rescaled}}(Y)\|^2/\sigma^2)^{-0.5}$ and 2) the EXP embedding kernel $k_{F,EXP}(X,Y) = \exp(-\|F_{\mathrm{rescaled}}(X) - F_{\mathrm{rescaled}}(Y)\|^2/(2\sigma^2))$, where the scaling σ is a bandwidth parameter.

Kernel Transformations Tilting a kernel with discrete masses preserves discrete masses. So do several other common kernel transformations.

Proposition B.31. *(Section 6.2 in Amin et al. (2023)) If k is a kernel with discrete masses and $A : S \to (0, \infty)$, then the tilted kernel $k^A(X, Y) = A(X)k(X, Y)A(Y)$ has discrete masses. If k, k' are kernels with discrete masses, the tensorized kernel $k \otimes k'((X, Y), (X', Y')) = k(X, X')k'(X', Y')$ has discrete masses. If k is a kernel with discrete masses and k' is another kernel then $k + k'$ has discrete masses. If k is a kernel with discrete masses and $S' \subseteq S$ then k restricted to S' has discrete masses.*

We can use these transformations to craft new scalar field kernels with discrete masses out of existing ones.

B.4.2. Vector Field Kernels with Discrete Masses

In this section we construct vector field kernels that have discrete masses and that capture biological notions of sequence similarity. The basic idea is to develop transformations from scalar field to vector field kernels that preserve the discrete mass property. Recall that we cannot simply take the gradient of the scalar field kernel (Proposition B.9). Instead, our approach is to tensorize scalar field kernels so that they can be applied to pairs of sequences, and then to enforce anticommutativity.

We first explain how anticommutativity is enforced. The idea is to first define a canonical ordering of sequences; once we have chosen a value of the kernel for the canonical ordering, its value for all other orderings follows by anticommutativity. The canonical ordering itself is defined in terms of a sign.

Definition B.32. A sign on M is a $\sigma : M \to \{-1, 1\}$ such that $\sigma(X, Y) = -\sigma(Y, X)$ for all $(X, Y) \in M$. Define $M^\sigma = \{(X, Y) \in M \mid \sigma(X, Y) = 1\}$. For a $(X, Y) \in M$, define $(X, Y)^\sigma = (X, Y)$ if $\sigma(X, Y) = 1$ and (Y, X) otherwise. We say σ is "proper" if $\sigma(X, Y) = 1$ if $|Y| = |X| - 1$ for $(X, Y) \in M$.

Once we have chosen a value of the kernel for the canonical ordering (i.e. M^σ) we can extend it to all orderings (i.e. M) by symmetry. If the kernel has discrete masses on M^σ then its extension to M will be a vector field kernel with discrete masses.

Proposition B.33. *Let σ be a sign on M. There is a bijective correspondence between kernels on M^σ and vector field kernels, such that a kernel k on M^σ corresponds to the vector field kernel*

$$((X, Y), (X', Y')) \mapsto \sigma(X, Y)\sigma(X', Y')k((X, Y)^\sigma, (X', Y')^\sigma) \tag{14}$$

and a vector field kernel corresponds to its restriction to M^σ. Kernels k on M^σ with discrete masses, i.e. kernels such that $\delta_{(X,Y)} \in \mathcal{H}_k$ for all $(X, Y) \in M^\sigma$, correspond to vector field kernels with discrete masses.

Proof. We first show that any given kernel on M^σ corresponds to a well-defined, non-negative definite vector field kernel. Provided this holds, it is clear that the correspondence indeed describes a bijection between kernels on M^σ and vector field kernels, since every vector field kernel must fit the form of Eqn. 14 due to the anticommutivity property (Eqn. 3). Let k be a kernel on M^σ, $(Z_n)_{n=1}^N \subset M$ be distinct, and $(\alpha_n)_{n=1}^N \subset \mathbb{R}$. For $Z \in M$, call $\alpha_Z = \alpha_n$ if $Z = Z_n$ and 0 if $Z \neq Z_n$ for any n. We have non-negative definiteness of the proposed kernel on M from:

$$\begin{aligned}\sum_n \sum_m \sigma(Z_n)\sigma(Z_m)\alpha_n\alpha_m k(Z_n^\sigma, Z_m^\sigma) &= \sum_{Z \in M} \sum_{Z' \in M} \sigma(Z)\sigma(Z')\alpha_Z\alpha_{Z'}k(Z^\sigma, Z'^\sigma) \\ &= \sum_{Z \in M^\sigma} \sum_{Z' \in M^\sigma} (\alpha_Z - \alpha_{Z^{-\sigma}})(\alpha_{Z'} - \alpha_{Z'^{-\sigma}})\, k(Z, Z') \geq 0.\end{aligned}$$

Next we need to check that the kernel on M is indeed a vector field kernel, which satisfies anticommutivity. Call $\tilde{k}$ the extension of the kernel k to M. Then if $\tilde{f} \in \mathcal{H}_{\tilde{k}}$,

$$\tilde{f}(X, Y) = \left\langle \tilde{f} \middle| \tilde{k}((X, Y), \cdot) \right\rangle_{\tilde{k}} = -\left\langle \tilde{f} \middle| \tilde{k}((Y, X), \cdot) \right\rangle_{\tilde{k}} = -\tilde{f}(Y, X).$$

The first equality follows from the fact that for any $f \in \mathcal{H}_k$ there is a $\tilde{f} \in \mathcal{H}_{\tilde{k}}$ such that $\tilde{f}(X, Y) = \sigma(X, Y)f((X, Y)^\sigma)$. To see this, note that $k_{(X,Y)} \mapsto \tilde{k}_{(X,Y)}$ defines a unitary linear transformation on finite linear combinations of $\{k_{(X,Y)}\}_{(X,Y) \in M^\sigma}$. This transformation takes f that are finite linear combinations of $\{k_{(X,Y)}\}_{(X,Y) \in M^\sigma}$ to $\tilde{f}$ as defined above, and can be extended to all of $\mathcal{H}_k$ to obey the same property.

The discrete mass property of the vector field kernel on M follows by setting $f = \delta_{(X,Y)}$, in which case the corresponding $\tilde{f}$ is a delta function on M (Def. B.11) □

Now, to construct a kernel on M^σ with discrete masses, we can tensorize two scalar field kernels with discrete masses. In the following proposition, we also give some examples of kernels on M^σ that are valid kernels (they are non-negative definite) but are not guaranteed to have discrete masses.

Proposition B.34. *Let k, k' be scalar field kernels on S. The following are all valid kernels on M^σ.*

$$((X,Y),(X',Y')) \mapsto k(X,X')k'(Y,Y')$$
$$((X,Y),(X',Y')) \mapsto (k(X,X') + k'(Y,Y'))^2$$
$$((X,Y),(X',Y')) \mapsto k(X+Y,X'+Y')$$
$$((X,Y),(X',Y')) \mapsto k(X,X')$$
$$((X,Y),(X',Y')) \mapsto k(Y,Y')\mathbb{1}(|X| \neq |Y|, |X'| \neq |Y'|).$$

If k, k' have discrete masses, then the first two kernels have discrete masses on M^σ. By Proposition B.33, their extension to M is a vector field kernel with discrete masses.

Proof. The first four of these kernels are non-negative definite because they are restrictions of non-negative definite kernels on $S \times S$. The last kernel can be constructed by first defining the kernel $((X,Y),(X',Y')) \mapsto k(Y,Y')$ on $S \times S$, restricting to $\{(X,Y) \in M^\sigma \mid |X| \neq |Y|\}$ and then extending to the rest of M^σ by setting $k_{(X,Y)} = 0$ if $|X| = |Y|$. Each of these operations preserved non-negative definiteness.

If k, k' have discrete masses, the first kernel described above has discrete masses by Proposition B.31 as it is the restriction of $k \otimes k'$ on $S \times S$. The second kernel also has discrete masses by Proposition B.31 as $(k(X,X') + k'(Y,Y'))^2 = \left(k(X,X')^2 + k'(Y,Y')^2\right) + 2k \otimes k'((X,Y),(X',Y'))$ so that is the sum of two kernels, one with discrete masses. □

We can now employ scalar field kernels with discrete masses to construct vector field kernels with discrete masses. If the scalar field kernel we use captures sensible biological notions of sequence similarity – such as Hamming distance or alignment distance – then the resulting vector field kernel will also.

B.4.3. Thick Tailed Vector Field Kernels with Discrete Masses

So far, we have shown how to construct vector field kernels with discrete masses that satisfy biological notions of sequence similarity. The next step is to add thick tails. Recall that the thickness of a kernel's tails is measured with respect to the base distribution p (Assumption B.18). Our analysis in this section will focus on the setting where p is a pHMM, and take $\chi(t) = t \wedge 1$, so that Proposition B.6 and Corollary B.7 hold.

Establishing Thick Tails in Practice For the KSD-B to be guaranteed to detect non-convergence, the kernel must satisfy Assumption B.18. To prove this holds for our actual proposed kernels, we will use the following technical lemma. The lemma asks for an $h \in \mathcal{H}_k$ that can be written as a small perturbation $g(X,Y)$ of a vector field $f(X,Y)$ that is zero if X and Y are the same length, and which depends only on $|X|$ when X and Y have different lengths. So long as f does not increase or decrease too quickly with $|X|$, we will be guaranteed that $\mathcal{T}_p h$ is large, and Assumption B.18 satisfied.

Lemma B.35. *Say p is a pHMM and k is either (a) a vector field kernel with a $h \in \mathcal{H}_k$ such that $h = f + g$ for vector fields f, g which satisfy the following conditions, or (b) a scalar field kernel with a $h \in \mathcal{H}_k$ such that $\nabla h = f + g$ for f, g that satisfy the following conditions.*

1. *(f only detects differences of sequence length) $f(X,Y) = 0$ if $|X| = |Y|$ and $f(X,Y)$ depends only on $|X|$ if $|X| > |Y|$. Call $f(L) = f(X,Y)$ for $|X| = L$ and $|Y| = L - 1$.*
2. *(g is small) $g(X,Y) = o(f(|X|))$ as $|X| \to \infty$. Note $g = 0$ satisfies this condition.*
3. *(f does not increase too fast) As $L \to \infty$, $f(L)$ is eventually positive. Moreover, for small enough $c > 0$, $(f(L+1) - f(L)) \leq cf(L)$ eventually. Note this latter condition is satisfied if f is non-increasing in L.*

4. *(f does not decrease too fast) $f(L) \gtrsim L^{-(1-\delta)}$ for some $\delta > 0$. Note this condition is satisfied if f is non-decreasing in L.*

Then, k satisfies Assumption B.18 for p.

Proof. Note first that if c is small enough, since $\mathrm{gap}_p(L) \gtrsim \sup_{|X|=L} \mathrm{ins}_p(X)$ by Proposition B.6, we eventually have $\mathrm{gap}_p(L) \geq c \sup_{|X|=L} \mathrm{ins}_p(X)$. Thus, eventually,

$$\left(\sup_{|X|=L} \mathrm{ins}_p(X)\right)(f(L+1) - f(L)) \leq \mathrm{gap}_p(L) f(L).$$

Say $X \in S$ and $|X| = L$. Since $\mathrm{flux}_p(X) \sim \mathrm{gap}_p(|X|)$ by Proposition B.6, for large enough L,

$$\begin{aligned}\mathcal{T}_p(f+g)(X) =& -\mathrm{ins}_p(X) f(L+1) + \mathrm{del}_p(X) f(L) + \mathrm{flux}_p(X) o(f(L)) \\ \geq& \mathrm{gap}_p(L) f(L) + \mathrm{ins}_p(X)\left(f(L) - f(L+1)\right) + \mathrm{flux}_p(X) o(f(L)) \\ \geq& \mathrm{gap}_p(L) f(L)(1 + o(1)).\end{aligned}$$

Since $\mathrm{ins}_p(X) \lesssim \mathrm{gap}_p(|X|) \sim \mathrm{flux}_p(X)$ and we can set $V_p(X) = (\log|X|)^{2+\epsilon}$ for some $\epsilon > 0$ by Proposition B.6 and Corollary B.7,

$$\sum_L \frac{\inf_{|X|=L} \mathcal{T}_p(f+g)(X)}{\left(\sup_{|X|=L} \mathrm{flux}_p(X)\right) V_p(L)} \gtrsim \sum_L \frac{f(L)}{(\log L)^{2+\epsilon}} \gtrsim \sum_L (\log L)^{-(2+\epsilon)} L^{-(1-\delta)} = \infty$$

and the same is true replacing flux_p with ins_p. □

Alignment Kernel Tails In the case of the alignment kernel, finding an h that satisfies Lemma B.35 is non-trivial. In Section B.4.5 we deploys tools from combinatorics to unpack the asymptotic behavior of the alignment kernel; this allows us to construct such an h. Here, we summarize the key conclusions of Section B.4.5. Define $\xi = 1 - e^{-\Delta\mu} < 1$ and the function,

$$r_1(x, \xi) = \frac{1}{2}\left(1 + x + \sqrt{(1+x)^2 - 4\xi x}\right).$$

Note that to ensure discrete masses, we need $\zeta \geq \log|\mathcal{B}|$ (Theorem B.27); in this case, $r_1(e^{\zeta/2}, \xi) \geq r_1(e^{\zeta/2}|\mathcal{B}|^{-1/2}, \xi) > 1$. We first study the alignment and local alignment kernels.

Proposition B.36. *Say $\Delta\mu < \infty$. Then,*

$$L^{1/2} r_1(e^{\zeta/2}, \xi)^{|X|} \leq \sqrt{k_{\mathrm{ali}}(X, X)} \leq r_1(e^{\zeta/2}, \xi)^{|X|}$$

and for any $\pi < 1$, there is a $h \in \mathcal{H}_k$ such that $h(X)$ depends only on $|X|$ and

$$h(X) = r_1(\pi e^{\zeta/2}|\mathcal{B}|^{-1/2}, \xi)^{|X|} + O(|X|).$$

As well, for any $X \in S$, $k_{\mathrm{ali},X} \lesssim h$. The same proposition is true replacing k_{ali} with k_{la}.

We next study the infinite kmer spectrum kernel. (Recall that this is equivalent to a local alignment kernel with $\Delta\mu = \infty$ and $\zeta = 0$.)

Proposition B.37. *Let $A(X) = |X|^{-3/2}$. k^A_{ISK} is a bounded C_0 kernel, i.e. for all $f \in \mathcal{H}_k$, $f \in C_0(S)$. k^A_{ISK} is also non-vanishing, i.e. $\sqrt{k^A_{\mathrm{ISK}}(X, X)} \not\to 0$ as $|X| \to \infty$. Moreover, if we set $h = \sum_{Y \in \mathcal{B}} k^A_{\mathrm{ISK},Y}$, then $h(X)$ depends only on $|X|$ and $h(X) = |X|^{-1/2} + |\mathcal{B}||X|^{-3/2}$.*

Scalar Field Kernels with Thick Tails We now describe our proposed scalar field kernels. These kernels have discrete masses, capture biological notions of sequence similarity, and possess thick tails.

Proposition B.38. *The following kernels have discrete masses and satisfy Assumption B.18 for any pHMM p with $\chi(t) = t \wedge 1$.*

1. *Unbounded IMQ-H (**IMQ-H (U)**): $k(X,Y) = A(X)k_{\mathrm{IMQ-H}}(X,Y)A(Y)$ for $A(X) = (|X| + C)^{\beta+1}$.*
2. *Unbounded alignment kernel (**Ali (U)**): $k(X,Y) = A(X)k_{\mathrm{ali}}(X,Y)A(Y)$ for $0 < \Delta\mu < \infty, \zeta \geq \log|\mathcal{B}|$ and $A(X) = r_1(e^{\zeta/2}|\mathcal{B}|^{-1/2}, \xi)^{(1-\epsilon)|X|}$ for some $\epsilon > 0$. One can replace k_{ali} with k_{la}.*
3. *Unbounded infinite kmer spectrum kernel (**ISK (U)**): $k(X,Y) = k_{\mathrm{ISK}}(X,Y)$.*

Proof. First note all three of these kernels have discrete masses by Theorems B.26, B.27 and B.28, since tilting preserves discrete masses (Proposition B.31). Now we show they satisfy the conditions of Lemma B.35.

1. **IMQ-H (U)** Let $h = -k_\emptyset$, so $h(X) = -(|X| + C)$. $\nabla h(X,Y) = |X| - |Y|$ depends only on $|X|$, $|Y|$ and is 0 if $|X| = |Y|$. Setting $f = \nabla h$ and $g = 0$ we have $f(L) = 1$ for all L and we satisfy the conditions of Lemma B.35.

2. **Ali (U)** Let h be as defined in Proposition B.36, picking $\pi < 1$ such that $\frac{r_1(\pi e^{\zeta/2}|\mathcal{B}|^{-1/2},\xi)}{r_1(e^{\zeta/2}|\mathcal{B}|^{-1/2},\xi)^{1-\epsilon}} = 1 + \delta$ for a small δ. h is a function only of the length of the sequence and $h(X) = (1+\delta)^{|X|} + o(1)$. Call $f = -\nabla\left(X \mapsto (1+\delta)^{|X|}\right)$ and $g = f - \nabla(-h) = o(1)$. Now,

$$f(X,Y) = 0 \text{ if } |X| = |Y|$$
$$f(X,Y) = \delta(1+\delta)^{|X|-1} \text{ if } |Y| = |X| - 1$$
$$f(Z,X) - f(X,Y) = \delta^2(1+\delta)^{|X|-1} = \delta(1+\delta)^{-1}f(X,Y) \text{ if } |Y| < |X| < |Z|.$$

Clearly f and g satisfy conditions 1, 2, and 4 of Lemma B.35 and, picking small enough δ, condition 3 is also satisfied.

3. **ISK (U)** Let h be as defined in Proposition B.37 so that h is a function only of sequence length and $A(X)h(X) = |X| + 4$. By similar reasoning to the IMQ-H (U), setting $f = \nabla h$ and $g = 0$ we satisfy the conditions of Lemma B.35. □

Vector Field Kernels with Thick Tails We now describe our proposed vector field kernels. These kernels have discrete masses, capture biological notions of sequence similarity, and possess thick tails. To construct the kernels, we add together a thick tailed kernel that does not have discrete masses and a thin tailed kernel that does. In this section, we assume σ is a *proper* ordering (Definition B.32). We may for example let σ be the lexicographic ordering for some ordering of the letters in $\mathcal{B}$.

Proposition B.39. *Let $k = k_{\mathrm{HT}} + k_\delta$ for vector field kernels $k_{\mathrm{HT}}, k_\delta$. Any of the following choices of k_{HT} and k_δ result in a vector field kernel k that is bounded and satisfies Assumption B.18 for a pHMM p and $\chi(t) = t \wedge 1$. k_{HT} can be,*

1. *IMQ-H (**IMQ-H**): $k_{\mathrm{HT}}((X,Y),(X',Y')) = k_{\mathrm{IMQ-H}}(Y,Y')\mathbb{1}(|X| \neq |Y|, |X'| \neq |Y'|)$ for $(X,Y) \in M^\sigma$, with $\beta < 1$ in $k_{\mathrm{IMQ-H}}$.*
2. *Infinite kmer spectrum kernel (**ISK**): $k_{\mathrm{HT}}((X,Y),(X',Y')) = A(Y)k_{\mathrm{ISK}}(Y,Y')A(Y')\mathbb{1}(|X| \neq |Y|, |X'| \neq |Y'|)$ for $(X,Y) \in M^\sigma$ with $A(X) = |X|^{-3/2}$.*

k_δ can be,

1. *Alignment kernel (**Ali**): $k_\delta((X,Y),(X',Y')) = (k^A_{\mathrm{ali}}(X,X') + k^A_{\mathrm{ali}}(Y,Y'))^2$ for $(X,Y) \in M^\sigma$, with $0 < \Delta\mu < \infty$ and $\zeta \geq \log|\mathcal{B}|$, and $A(X) = \left(r_1(e^{\zeta/2},\xi)\right)^{-|X|}$. One can also use k_{la} instead of k_{ali}.*
2. *Exponential Hamming kernel (**Exp-H**): $k_\delta((X,Y),(X',Y')) = \left(k_{\mathrm{Exp-H}}(X,X') + k_{\mathrm{Exp-H}}(Y,Y')\right)^2$ for $(X,Y) \in M^\sigma$.*

Proof. Note the vector field alignment kernel **Ali** has discrete masses by Propositions B.33 and B.34, since the tilted scalar field alignment kernel has discrete masses by Theorem B.27 and tilting preserves discrete masses by Proposition B.31. Similarly, by Theorem B.26 the vector field exponential Hamming kernel **Exp-H** has discrete masses. Finally, since k_δ has discrete masses, by Proposition B.31, $k = k_{\mathrm{HT}} + k_\delta$ has discrete masses when restricted to M^σ. Finally by Proposition B.33 k has discrete masses as a vector field kernel on M.

Also note that all of these kernels are bounded: the alignment vector field kernel is bounded by Proposition B.36 and the ISK kernel is bounded by Proposition B.37.

We prove that $k = k_{\mathrm{HT}} + k_\delta$ satisfies the conditions of Lemma B.35 when k_δ is chosen to be the alignment kernel **Ali**, using Proposition B.36. The logic is similar when we choose k_δ to be the exponential Hamming kernel **Exp-H**.

1. Let k_{HT} be the **IMQ-H** kernel. Let $h = k_{(A,\emptyset)}$ for some $A \in \mathcal{B}$. Now let $f = k_{\mathrm{HT},(A,\emptyset)}$ so that $f(X,Y) = (|Y| + C')^{-\beta} = (|X| - 1 + C')^{-\beta}$ if $|Y| < |X|$ and $f(X,Y) = 0$ if $|X| = |Y|$, which satisfies conditions 1, 3, and 4 of Lemma B.35. Define $g = h - f = k_{\delta,(A,\emptyset)}$. Let $\tilde{h}$ be as defined as the h in Proposition B.36 for some $0 < \pi < 1$. We have, applying the bound on k_{ali} from Proposition B.36, that

$$k^A_{\mathrm{ali}}(X,X') \lesssim \tilde{h}(X)A(X) \sim \left(\frac{r_1(\pi e^{\zeta/2}|\mathcal{B}|^{-1/2},\xi)}{r_1(e^{\zeta/2},\xi)}\right)^{|X|} + O(|X|r_1(e^{\zeta/2},\xi)^{-|X|}) \sim \exp(-c|X|)$$

for some $c > 0$ when $|X'| = 0$ or $|X'| = 1$. Thus, $g(X,Y) = k_{\delta,(A,\emptyset)}(X,Y) = O(e^{-2c|X|})$. Thus, condition 2 of Lemma B.35 is also satisfied.

2. Let k_{HT} be the **ISK** kernel. Let $h = \sum_{b\in\mathcal{B}} k_{(b+b,b)}$. Now let $h_{\mathrm{HT}} = \sum_{b\in\mathcal{B}} k_{\mathrm{HT},(b+b,b)}$ so that, by Proposition B.37, $h_{\mathrm{HT}}(X,Y) = (|X|-1)^{-1/2} + Co(|X|^{-1/2})$ if $|Y| < |X|$ and $h_{\mathrm{HT}}(X,Y) = 0$ if $|X| = |Y|$. Let $f(X,Y) = (|X|-1)^{-1/2}$ if $|Y| < |X|$ and $f(X,Y) = 0$ if $|X| = |Y|$. Finally define $g(X,Y) = (h-f)(X,Y) = \sum_{b\in\mathcal{B}} k_{\delta,(b+b,b)}(X,Y) + o(|X|^{-1/2})$. As in the previous example, $g(X,Y) = O(e^{-c|X|}) + o(|X|^{-1/2})$. Thus, k satisfied the conditions of Lemma B.35. □

B.4.4. Kernel Integrability

For the KSD-B to reliably detect convergence to p, Proposition 5.3 says we need k to not be too large. At a minimum, the KSD-B should be zero when evaluated exactly at p itself. Proposition 5.1 says that for $\text{KSD-B}_{p,k}(p) = 0$ we need p to be p,k-integrable. In this section we show that for all of our proposed vector field kernels, if $\chi(t) = 1 \wedge t$ then any subexponential p is p,k-integrable (recall p is subexponential if $E_{X\sim p}e^{t|X|} < \infty$ for small enough t (Amin et al., 2021)). All pHMMs are subexponential (Proposition B.6). Another important class of subexponential distributions are autoregressive models (Proposition 5 of Amin et al. (2021)).

Only some of our proposed scalar field kernels, however, have the same guarantee. This reflects a fundamental disadvantage of scalar field kernels: they must be unbounded, and hence very large, to detect non-convergence.

Our results rely on the following lemma, which gives conditions on the kernel that ensure p,k-integrability when p is sub-exponential.

Lemma B.40. *Say $\chi(t) = t \wedge 1$ and p is subexponential. If $\sqrt{k((X,Y),(X,Y))} \le e^{t'|X|}$ for small enough t' then p is p,k integrable.*

Proof. Note that since $\chi(t) = t \wedge 1$, $T_{p,X\to Y} \le |X|$ for all XMY. Now we have

$$E_{X\sim p}\sum_{YMX} T_{p,X\to Y}\sqrt{k((X,Y),(X,Y))} \le E_{X\sim p}|X|^2 e^{t'|X|} < \infty$$

if t' is small enough. □

We can now guarantee integrability for all of our proposed kernels, besides **Ali (U)**.

Corollary B.41. *Say $\chi(t) = t\wedge 1$ and p is subexponential. If k is any of the vector field kernels considered in Proposition B.39 then p is p,k integrable. If k is the **IMQ-H (U)** or **ISK (U)** kernels considered in Proposition B.38 then p is p,k^∇ integrable.*

Proof. The first statement follows from Lemma B.40 and the fact that the kernels in Proposition B.39 are bounded. The second statement follows from the fact that for the **IMQ-H (U)** kernel, $\sqrt{k(X,X)} = (1+|X|)^{1+\beta}$ and for the **ISK (U)** kernel $\sqrt{k(X,X)} \leq (1+|X|)^{3/2}$ by Proposition B.37. □

Next we consider the unbounded scalar field alignment kernel, **Ali (U)**. If p is a pHMM and k is **Ali (U)**, it is possible that under certain conditions p is not p,k-integrable. By Proposition B.36,

$$L^{-1/2}\left(\frac{r_1(e^{\zeta/2},\xi)}{r_1(e^{\zeta/2}|\mathcal{B}|^{-1/2},\xi)^{1-\epsilon}}\right)^L \leq \sup_{|X|=L}\sqrt{k(X,X)}.$$

One can check that the ratio on the left hand side is minimized in the limit $\epsilon = 0, \xi = 0, \zeta = \log|\mathcal{B}|$ in which case

$$\frac{r_1(e^{\zeta/2},\xi)}{r_1(e^{\zeta/2}|\mathcal{B}|^{-1/2},\xi)^{1-\epsilon}} \geq \frac{r_1(|\mathcal{B}|^{1/2},0)}{r_1(1,0)} = \frac{|\mathcal{B}|^{1/2}+1}{2}.$$

Now, $\frac{1}{2}\left(|\mathcal{B}|^{1/2}+1\right)$ is 3/2 in the case when $\mathcal{B}$ is the set of nucleotides (where $|\mathcal{B}|=4$) and approximately 2.74 in the case when $\mathcal{B}$ is the set of amino acids (where $|\mathcal{B}|=20$). So on real biological sequence data, $\sup_{|X|=L}\sqrt{k(X,X)}$ grows exponentially, and thus the unbounded scalar field alignment kernel may be too large to ensure p,k integrability.

B.4.5. Proofs for the Alignment Kernel

In this section we bound $\sqrt{k(X,X)}$ and find a thick tailed $h \in \mathcal{H}_k$ for the alignment kernel k.

Let us review some results for the case when $|\mathcal{B}|=1$ that will be useful. If $\mathcal{B}=\{A\}$, call $k(L,L') = k(L\times A, L'\times A)$. Section 9 of Amin et al. (2023) showed that there is an orthogonal basis $(u_L)_L$ such that $\|u_L\|_k = e^{-\zeta L/2}$ where $\zeta = 2\mu + \log k(A,A)$. In this case, $(u_{L'},k_L)_k \geq 0$ for all L, L' and $(u_{L'},k_L)_k = 0$ if $L' > L$. Then, defining the infinite upper triangular matrix Q such that $Q_{L',L} = (u_{L'},k_L)_k$, we get

$$k(L,L') = (k_L|k_{L'})_k = \left(\sum_{L''} Q_{L'',L}e^{\zeta L''}u_{L''}\middle|\sum_{L''}Q_{L'',L'}e^{\zeta L''}u_{L''}\right)_k = \sum_{L''=0}^{\infty} Q_{L'',L}Q_{L'',L'}e^{L''\zeta}. \tag{15}$$

The same equation holds for k_{la} for another matrix Q_{la}. The exact values of the entries of the matrix Q and the matrix Q_{la} will be important to achieve bounds on the tails of the alignment kernel. Amin et al. (2023) showed in Appendix I and J that if we define $\xi = 1-e^{-\Delta\mu}$, $f_\xi(y) = \frac{1-\xi y}{1-y}$, and the formal power series

$$F_\xi(x,y) = \frac{f_\xi(y)}{1-xyf_\xi(y)} = \frac{1-\xi y}{1-(1+x)y+\xi xy^2}$$

$$F_{\xi,\mathrm{la}}(x,y) = xy\frac{\left(\frac{1}{1-y}\right)^2}{1-xyf_\xi(y)} + \frac{1}{1-y} = \frac{xy}{(1-y)\left(1-(1+x)y+\xi xy^2\right)} + \frac{1}{1-y}$$

then $Q_{L',L} = [x^{L'}y^L]F_\xi(x,y)$ and $Q_{\mathrm{la},L',L} = [x^{L'}y^L]F_{\xi,\mathrm{la}}(x,y)$ where $[x^{L'}y^L]$ denotes the coefficient in front of the term $x^{L'}y^L$ of the formal power series.

We now show that we can use these formal power series to describe the size of $k(X,X)$.

Proposition B.42. *Calling $C_L = [y^L]F_\xi(e^{\zeta/2},y)$, we have $L^{-1/2}C_L \leq \sup_{|X|=L}\sqrt{k(X,X)} \leq C_L$. The same inequality is true for k_{la} and $F_{\xi,\mathrm{la}}$.*

Proof. First, by equation 13, if $A \in \mathcal{B}$, we clearly have $k(X,X) \leq k(|X|\times A, |X|\times A)$, since the alignment kernel takes its largest value if every letter in the sequence matches every other. $k_s(A,A) = \sigma^{-1}|\mathcal{B}|$, so k restricted to $\{\emptyset, A, AA, AAA, \dots\}$ is identical to the string kernel in the case $|\mathcal{B}|=1$ and with $\zeta = 2\mu - \log\sigma + \log|\mathcal{B}|$. Thus, by

equation 15,

$$\begin{aligned}
k(L,L) &= \sum_{L'=0}^{L} e^{L'\zeta} Q_{L',L}^2 \\
&\leq \left(\sum_{L'=0}^{\infty} e^{L'\zeta/2} Q_{L',L} \right)^2 \\
&= \left(\sum_{L'=0}^{\infty} \left(e^{\zeta/2}\right)^{L'} [x^{L'} y^L] F_\xi(x,y) \right)^2 \\
&= \left([y^L] F_\xi(e^{\zeta/2}, y) \right)^2 .
\end{aligned}$$

The result is identical with $F_{\xi,\mathrm{la}}$. On the other hand, using Jensen's inequality,

$$\begin{aligned}
k(L,L) =& L \left(\frac{1}{L} \sum_{L'=0}^{L} \left(e^{L'\zeta} Q_{L',L}\right)^2 \right) \\
\geq& L \left(\frac{1}{L} \sum_{L'=0}^{L} e^{L'\zeta/2} Q_{L',L} \right)^2 \\
=& \frac{1}{L} \left([y^L] F_\xi(e^{\zeta/2}, y) \right)^2 .
\end{aligned}$$

□

Now we build a thick tailed $h \in \mathcal{H}_k$.

Proposition B.43. *Say $0 < \pi < 1$. For the alignment kernel, there is an $h \in \mathcal{H}_k$ such that $(h|k_X)_k = [y^{|X|}] F_\xi\left(\pi e^{\zeta/2} |\mathcal{B}|^{-1/2}, y\right)$. For the local alignment kernel, there is a $h \in \mathcal{H}_{k_{\mathrm{la}}}$ such that $(h|k_{\mathrm{la},X})_k = C + [y^{|X|}] F_{\xi,\mathrm{la}}\left(\pi e^{\zeta/2} |\mathcal{B}|^{-1/2}, y\right)$ for some constant C. In both cases, for any $X \in S$, we have $k_X \lesssim h$.*

Proof. Define $k_L = |\mathcal{B}|^{-L} \sum_{|X|=L} k_X$. If $Y, Y' \in S$ and $|Y| = |Y'| = L'$, one can check that $(k_L|k_Y)_k = (k_L|k_{Y'})_k$, thus $(k_L|k_Y)_k = (k_L|k_{L'})_k$. One can show that k restricted to $\{k_0, k_1, \dots\}$ is identical to the string kernel in the case $|\mathcal{B}| = 1$ with $\zeta = -\log|\mathcal{B}|$. We will create a h for this kernel with $(h|k_L)_k = [y^L] F_\xi\left(e^{\zeta/2} |\mathcal{B}|^{-1/2} \pi, y\right)$ and the Proposition will follow from the fact that $(h|k_Y)_k = (h|k_{|Y|})_k$.

We define $h = \sum_L \alpha^L k_L$ for some $\alpha > 0$. Note that since $k(X,Y) \geq 0$ for all X, Y, $k_X \lesssim h$ for any X. Now write

$$(h|u_{L'})_k = \sum_L \alpha^L [x^{L'} y^L] F_\xi(x,y) = [x^{L'}] F_\xi(x, \alpha).$$

Thus, since $F_\xi(x,y) = f_\xi(y)(1 - xy f_\xi(y))^{-1} = f_\xi(y) \sum_{L=0}^{\infty} x^L (y f_\xi(y))^L$,

$$\|h\|_k^2 = \sum_L e^{\zeta L} |\mathcal{B}|^{-L} \left([x^L] F_\xi(x,\alpha) \right)^2 = f_\xi(\alpha)^2 \sum_L e^{\zeta L} |\mathcal{B}|^{-L} \left(\alpha f_\xi(\alpha)\right)^{2L}$$

which is finite as long as $\pi = \alpha f_\xi(\alpha) e^{\zeta/2} |\mathcal{B}|^{-1/2} < 1$. We can pick α to let π be any positive value < 1. In this case

$$\begin{aligned}
(h|k_{L'})_k =& \sum_L e^{\zeta L} |\mathcal{B}|^{-L} (h|u_L)_k Q_{L,L'} \\
=& \sum_L e^{\zeta L} |\mathcal{B}|^{-L} \left(f_\xi(\alpha) (\alpha f_\xi(\alpha))^L \right) [x^L y^{L'}] F_\xi(x,y) \\
=& f_\xi(\alpha) \sum_L \left(\pi e^{\zeta/2} |\mathcal{B}|^{-1/2} \right)^L [x^L y^{L'}] F_\xi(x,y) \\
=& f_\xi(\alpha) [y^{L'}] F_\xi\left(\pi e^{\zeta/2} |\mathcal{B}|^{-1/2}, y \right).
\end{aligned}$$

We now turn to the very similar case of k_{la}. The norm of $h = \sum_L \alpha^L k_{\mathrm{la},L}$ is

$$\sum_L e^{\zeta L}|\mathcal{B}|^{-L}\left([x^L]F_{\xi,\mathrm{la}}(x,\alpha)\right)^2 = \left(\frac{1}{1-\alpha}\right)^2 + \left(\frac{\alpha}{1-\alpha}\right)^2 \sum_{L=1}^{\infty} e^{\zeta L}|\mathcal{B}|^{-L}\left(\alpha f_\xi(\alpha)\right)^{2(L-1)}$$

which is finite again as long as $\pi = \alpha f_\xi(\alpha) e^{\zeta/2}|\mathcal{B}|^{-1/2} < 1$.

$$\begin{aligned}(h|k_{\mathrm{la},L'})_k =& \sum_L e^{\zeta L} Q_{\mathrm{la},L,L'}\left([x^L]F_{\xi,\mathrm{la}}(x,\alpha)\right)\\ =& \frac{1}{1-\alpha}Q_{\mathrm{la},0,L'} + \frac{\alpha}{(1-\alpha)^2\alpha f_\xi(\alpha)}\sum_{L=1}^{\infty} e^{\zeta L/2}|\mathcal{B}|^{-L/2}Q_{\mathrm{la},L,L'}\pi^L\\ =& \frac{1}{1-\alpha} - \frac{\alpha}{(1-\alpha)\alpha f_\xi(\alpha)}Q_{\mathrm{la},0,L} + \frac{\alpha}{(1-\alpha)\alpha f_\xi(\alpha)}[y^{L'}]F_{\xi,\mathrm{la}}(e^{\zeta/2}|\mathcal{B}|^{-1/2}\pi, y).\end{aligned}$$

Finally note $Q_{\mathrm{la},0,L} = 1$. □

Thus, to analyze the tails of the alignment kernel, we will need to analyze $[y^L]F_\xi(x,y)$ and $[y^L]F_{\xi,\mathrm{la}}(x,y)$. The coefficients will turn out to depend on the polynomial $1-(1+x)y+\xi y^2$. We rewrite the polynomial $1-(1+x)y+\xi y^2 = (1-r_1 y)(1-r_2 y)$ for roots $r_1(\xi,x) \geq r_2(\xi,x)$, which are

$$\frac{1}{2}\left(1+x \pm \sqrt{(1+x)^2 - 4\xi x}\right).$$

These values are decreasing with ξ, positive, and distinct when $\xi < 1$ since $(1+x)^2 - 4\xi x > (1+x)^2 - 4x = (x-1)^2 \geq 0$. When $\xi < 1$, r_1 is also always > 1 since it is $> \frac{1}{2}(1+x+|x-1|) = x \vee 1$. When $\xi = 0$, $r_1 = x+1, r_2 = 0$. We now see that if $\Delta\mu < \infty$ then the coefficients grow exponentially. However, if $\Delta\mu = \infty$ the coefficients may grow or shrink exponentially or, in the case of $F_{\xi,\mathrm{la}}$ grow exponentially or polynomially.

Proposition B.44. *If $\xi < 1$ and $x > 0$, both $[y^L]F_\xi(x,y)$ and $[y^L]F_{\xi,\mathrm{la}}(x,y)$ are equal to $Cr_1(x,\xi)^L + O(L)$ for some (different) $C > 0$. If $\xi = 1$ then for the alignment kernel, $[y^L]F_1(x,y) = x^L$; for the local alignment kernel, if $x > 1$, then $[y^L]F_{1,\mathrm{la}}(x,y) = x^L + O(L)$; if $x < 1$, then $[y^L]F_{1,\mathrm{la}}(x,y) = CL + C' + o(1)$ for some $C > 0, C'$; if $x = 1$, then $[y^L]F_{1,\mathrm{la}}(x,y) = L(L+1)/2 + 1$.*

Proof. First let us consider the case of $\xi = 0$.

$$F_0(x,y) = \frac{1}{1-(1+x)y} = \sum_{L=0}^{\infty}(1+x)^L y^L$$

$$F_{0,\mathrm{la}}(x,y) = \frac{xy}{(1-y)(1-(1+x)y)} + \frac{1}{1-y}.$$

By partial fraction decomposition, for some A, B with $A, B \neq 0$ and , constant c_1, c_2,

$$\begin{aligned}F_{0,\mathrm{la}}(x,y) =& \frac{Axy}{1-(1+x)y} + \frac{Bxy}{1-y} + \frac{1}{1-y}\\ =& c_1 + c_2 y + \sum_{L=2}^{\infty}\left(Ax(1+x)^{L-1} + Bx + 1\right)y^L.\end{aligned}$$

The leading term in the brackets is $Ax(1+x)^{L-1}$ and, since the coefficients of $F_{0,\mathrm{la}}$ are positive, $A > 0$.

Now we consider the case when $0 < \xi < 1$.

$$F_\xi(x,y) = \frac{(1-\xi y)}{(1-r_1 y)(1-r_2 y)}$$

so by partial fraction decomposition, for $A, B \neq 0$,

$$F_\xi(x,y) = (1-\xi y)\left(\frac{A}{1-r_1 y} + \frac{B}{1-r_2 y}\right) = c_1 \sum_{L=0}^{\infty} \left(Ar_1^L - A\xi r_1^{L-1} + Br_2^L - B\xi r_2^{L-1}\right) y^L.$$

Since $r_1 > 1 > \xi$, the leading term in the brackets is $A(1-\xi/r_1)r_1^L$. Similarly, $[y^L]F_{\xi,\mathrm{la}} = Cr_1^L + O(L)$ for some $C > 0$.

Now we look at when $\xi = 1$. Here $f_\xi(y) = 1$. Thus,

$$F_1(x,y) = \frac{1}{1-xy} = \sum_{L=0}^{\infty} x^L y^L$$

$$F_{1,\mathrm{la}}(x,y) = \frac{xy}{(1-y)^2(1-xy)} + \frac{1}{1-y}.$$

If $x \neq 1$, again, by partial fraction decomposition, for some A, B, C with $A, B \neq 0$ and $C \neq 1$, and constants c_1, c_2,

$$\begin{aligned} F_{1,\mathrm{la}}(x,y) =& \frac{Axy}{1-xy} + \frac{Bxy(y-C)}{(1-y)^2} + \frac{1}{1-y} \\ =& c_0 + c_1 y + \sum_{L=2}^{\infty} \left(Ax^L + Bx\binom{L+1-2}{1} - BCx\binom{L+1-1}{1} + 1\right) y^L \end{aligned}$$

So that the leading term is Ax^L if $x > 1$ or $Bx\binom{L-1}{1} - BCx\binom{L}{1}$ if $x < 1$. Since $C \neq 1$, the latter term is $= CL + C'$ for some $C, C' > 0$. If $x = 1$,

$$F_{1,\mathrm{la}}(x,y) = \frac{1}{1-y} + \frac{y}{(1-y)^3} = 1 + \sum_{L=1}^{\infty} \left(\binom{L+2-1}{2} + 1\right) y^L$$

so that $[y^L]F_{1,\mathrm{la}}(x,y) = L(L+1)/2 + 1$. □

Now, combining the results of these last three propositions, we have proven Proposition B.36. To begin proving Proposition B.37, we first tighten our estimate of $\sqrt{k(X,X)}$ in the case $\Delta\mu = \infty$.

Proposition B.45. *Say $\Delta\mu = \infty$. $\sup_{|X|=L} \sqrt{k(X,X)} = e^{L\zeta/2}$ and $\sup_{|X|=L} \sqrt{k_{\mathrm{la}}(X,X)}$ is $\sim e^{L\zeta/2}$ if $\zeta > 0$, is $\sim L^{3/2}$ if $\zeta = 0$, and is $\sim L$ if $\zeta < 0$.*

Proof. When $\Delta\mu = \infty$, Q is the identity matrix, so that

$$k(L \times A, L \times A) = e^{L\zeta}.$$

On the other hand, $Q_{\mathrm{la},0,L} = 1$ for all L and, for $L \geq L' > 0$,

$$[x^{L'} y^L] F_{1,\mathrm{la}}(x,y) = [y^L] \frac{y}{(1-y)^2} y^{L'-1} = [y^{L-L'}](1-y)^{-2} = L - L' + 1.$$

Thus,

$$\begin{aligned} k(L \times A, L \times A) =& 1 + \sum_{L'=1}^{L} e^{L'\zeta}(L - L' + 1)^2 \\ =& 1 + e^{(L+1)\zeta} \sum_{L'=1}^{L} e^{-(L-L'+1)\zeta}(L - L' + 1)^2 \\ =& 1 + e^{(L+1)\zeta} \sum_{L'=1}^{L} e^{-L'\zeta} L'^2. \end{aligned}$$

If $e^\zeta > 1$, the sum is increasing and bounded, so, $k(L \times A, L \times A) = 1 + Ce^{L\zeta}(1 + o(1))$ for some $C > 0$. If $e^\zeta = 1$, we have $k(L \times A, L \times A) = 1 + CL^3(1 + o(1))$ for some $C > 0$. Finally, if $e^\zeta < 1$, since

$$k(L \times A, L \times A) = 1 + L^2 \sum_{L'=1}^{L} e^{L'\zeta} \left(1 - \frac{L' - 1}{L}\right)^2,$$

the sum is increasing and bounded with L so that $k(L \times A, L \times A) = 1 + CL^2(1 + o(1))$ for some $C > 0$. □

Next we must look at when a tilted alignment kernel is C_0.

Proposition B.46. *Say $\tilde{A} : \mathbb{N} \to (0, \infty)$ and $A(X) = \tilde{A}(|X|)$. If k^A is a bounded kernel, then it is C_0 if and only if $\tilde{A}(L)[y^L]F_\xi(\pi e^{\zeta/2}|\mathcal{B}|^{-1/2}, y) \to 0$ for some $\pi < 1$. If the latter condition holds for some value of π then it holds for any value of $0 < \pi < 1$. The same is true with k_{la} and $F_{\xi,\mathrm{la}}$.*

Proof. Let h be defined as in Proposition B.43 for some $1 > \pi > 0$. $h \in \mathcal{H}_k$ so $Ah \in \mathcal{H}_{k^A}$. $k^A_X \lesssim Ah$ for all X by Proposition B.43 so k^A is C_0 if and only if $Ah \in C_0(S)$. Finally, $Ah \in C_0(S)$ if and only if $\tilde{A}(L)[y^L]F_\xi(\pi e^{\zeta/2}|\mathcal{B}|^{-1/2}, y) \to 0$. Similar logic proves the same for k_{la} and $F_{\xi,\mathrm{la}}$. □

Finally we prove Proposition B.37. Recall that k_{ISK} is the special case of k_{la} with $\zeta = 0$ and $\Delta\mu = \infty$.

Proposition B.47. *(Proof of Proposition B.37) Let $A(X) = |X|^{-3/2}$. k^A_{ISK} is a bounded C_0 kernel, i.e. for all $f \in \mathcal{H}_k$, $f \in C_0(S)$. k^A_{ISK} is also non-vanishing, i.e. $\sqrt{k^A_{\mathrm{ISK}}(X, X)} \not\to 0$ as $|X| \to \infty$. Moreover, if we set $h = \sum_{X \in \mathcal{B}} k^A_{\mathrm{ISK},X}$, then $h(X)$ depends only on $|X|$ and $h(X) = |X|^{-1/2} + 4|X|^{-3/2}$.*

Proof. Note by proposition B.45, $\sup_{|X|=L} \sqrt{k_{\mathrm{ISK}}(X, X)} \sim L^{3/2}$. Thus, k^A_{ISK} is bounded and non-vanishing. On the other hand, if $\pi < 1$, $[y^L]F_\xi(\pi e^{\zeta/2}|\mathcal{B}|^{-1/2}, y) \sim L$ Proposition B.44, so, by Proposition B.46, k^A_{ISK} is C_0.

Finally, letting $h = \sum_{X \in \mathcal{B}} k_{\mathrm{ISK},X}$ and noting that if $X \in \mathcal{B}$, $k_{\mathrm{ISK},X}(Y) = \#(X \text{ in } Y) + 1$ (the plus one for $\phi_\emptyset$), we have that $h(Y) = |Y| + |\mathcal{B}|$. After tilting by $A(X)$, we obtain the proposition statement. □

C. Experimental Details

In this section we describe the details of our experiments. Note we used $\chi(t) = t \wedge 1$ in all cases unless otherwise specified.

C.1. Goodness of Fit Test Bootstrap

To get p-values for our goodness of fit tests, we used a bootstrap procedure as in Liu et al. (2016). Given $X_1, \ldots, X_n \in S$, for each i we sampled $Y_{X_i,1}, \ldots, Y_{X_i,N_n}$ drawn by taking a single step of a Markov chain with transition matrix $K_{X \to Y}$ defined in Section B.3.7. We defined our test U-statistic as

$$U = \frac{1}{n^2} \sum_{i \neq j} \mathrm{flux}_p(X_i)\mathrm{flux}_p(X_j) \frac{1}{N_n^2} \sum_{m,m'} k((X_i, Y_{X_i,m}), (X_j, Y_{X_j,m})).$$

We bootstrapped by sampling, for each $b = 1, \ldots, B$, $(w_{(b),i})_{i=1}^n \sim \mathrm{Multinomial}((1/n, \ldots, 1/n), n)$ and defining

$$U_{(b)} = \frac{1}{n^2} \sum_{i \neq j} (w_{(b),i} - 1)(w_{(b),j} - 1)\mathrm{flux}_p(X_i)\mathrm{flux}_p(X_j) \frac{1}{N_n^2} \sum_{m,m'} k((X_i, Y_{X_i,m}), (X_j, Y_{X_j,m})).$$

Then we defined a p-value $p = \#\{b \mid U_{(b)} \geq U\}/B$. Throughout we use $B = 1000$.

C.2. Kernel Parameters

In every case we used $\lambda = 1/5$ for the Exp-H kernel, $\beta = 1/2$ for the IMQ-H kernel, $\zeta = \log|\mathcal{B}|$ and $\Delta\mu = 0.2$ for the alignment kernel, and $\epsilon = 0.2$ for the tilting parameter of the alignment kernel. We set $C = 1$ for the IMQ-H kernel when in a vector field kernel and $C = 3$ when in a scalar field kernel. For embedding kernels, we set the bandwidth parameter σ to be the median distance between rescaled embeddings.

Given a kernel k we define its *normalized tilting* as $\tilde{k}(X, Y) = k(X, Y)/\sqrt{k(X, X)k(Y, Y)}$. This is how we define the IMQ (N), Ali (N) and ISK (N) kernels.

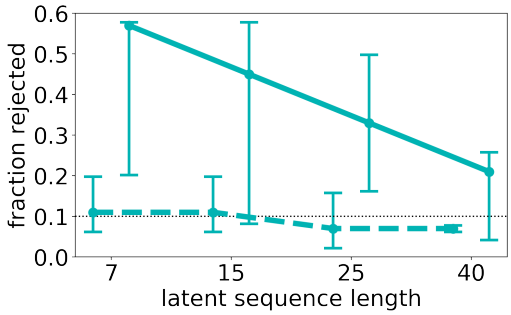


Figure 6. **Dependence of the Power of the Goodness of Fit Test on Sequence Length**Performance of the KSD-B, using a vector field kernel (**vf KSD-B**), with increasing sequence length. Dotted lines show the results for data sampled from the unperturbed model p, to evaluate calibration.

C.3. pHMM Model

Throughout our simulations we used a random profile hidden Markov models (pHMMs). We now define the prior from which we draw random pHMMs.

We first defined a site-wise independent model for sequences of length L. First we sampled the logits

$$h_{l,b}^{-2} \sim \beta_h^{-1}\text{Gamma}(\alpha_h).$$

Then we defined our site-wise independent model as

$$\tilde{p}(X) \propto \exp\left(\sum_{l,b} h_{l,b}\mathbb{1}(X_{(l)} = b)\right).$$

$\tilde{p}$ is supported on the set of sequences of length L.

Next we added insertions and deletions using a MuE model we sampled with with `indel_prior_bias = 5.0` (Weinstein & Marks, 2021). In particular, we set the latent sequence of the MuE model to be a sample from $\tilde{p}$. The final distribution we call p. p is a pHMM and its likelihoods can be calculated in closed form (Weinstein & Marks, 2021). To sample from and calculate likelihoods for this pHMM we used code from `https://github.com/pyro-ppl/pyro/tree/dev/pyro/contrib/mue` under the MIT licence.

C.4. Powerful Goodness of Fit Test with the KSD-B (Fig. 2(a))

For this experiment only we used $\chi(t) = \sqrt{t}$. We sampled a pHMM p with $L = 20, |\mathcal{B}| = 4, \alpha_h = 1, \beta_h = 0.5$. We then set $h_4 = 500\mathbb{1}(X_{(4)} = C)$ so that the 5-th letter of the latent sequence very likely a C (recall indexing of the sequence begins at 0). To make position 5 of the latent sequence coincide with position 5 of the sampled sequence, we subtracted 5 from the first 5 positions of the insertion and deletion logit matrices of the MuE indel model. We finally made a C in position 6 unlikely by subtracting 500 from $h_{5,C}$.

We then perturbed the pHMM by making other letters more probable as the 5-th letter: we defined, for perturbation weight γ, $\tilde{h}_{4,b} = \log(\gamma/3)$ if $b \neq C$ and $\tilde{h}_{4,C} = \log(1-\gamma)$ so that the probability of the 5-th letter being a C is $1-\gamma$. Finally we sampled and tested from a pHMM with $\tilde{h}$ for a variety of perturbation weights γ. We sampled and tested 25 times from the same models and reported the fraction of null hypotheses rejected.

C.5. Testing pHMM Models (Figures 2(b) and 6)

We now define the Potts model for sequences of length L that we used in the experiment. First we sampled the mean field parameters $h_{l,b}^{-2} \sim \beta_h^{-1}\text{Gamma}(\alpha_h)$ and the couplings parameters $e_{l,l',b,b'} \sim \sigma_{l,l'}N(0, 1.2^2)$ where $\sigma_{l,l'} \sim \text{Bern}(0.9)$.

Then, for a cross-term perturbation weight γ, we defined our Potts model as

$$\tilde{p}(X) \propto \exp\left(\sum_{l,b} h_{l,b}\mathbb{1}(X_{(l)} = b) + \gamma \sum_{l,l',b,b'} e_{l,l',b,b'}\mathbb{1}(X_{(l)} = b, X_{(l')} = b')\right).$$

$\tilde{p}$ is supported on the set of sequences of length L. We then sampled from a Potts model using code from `https://github.com/debbiemarkslab/plmc` under the MIT licence (Hopf et al., 2017).

We used $\alpha_h = 1, \beta_h = 3, |\mathcal{B}| = 4$ to define a Potts model. We added insertions and deletions using a MuE model we sampled with with `indel_prior_bias = 5.0` to get a distribution p that depends on γ. Note when $\gamma = 0$, p is a pHMM and we can calculate its likelihoods. Finally, we sample and test for a variety of perturbation weights γ, and we repeat the experiment 25 times. We sampled and tested 50 times from the same models and reported the fraction of null hypotheses rejected.

To produce Fig. 6 we first sampled 3 pHMMs for each latent sequence length with the parameters above. For $\gamma = 0, 0.3$ we performed the same testing procedure this time using $N_n = 10$ and repeating the sampling and testing 25 times.

C.6. Testing Autoregressive Models (Fig. 2(c))

We defined a lag 2 linear autoregressive model as

$$X_1 = C$$
$$X_L \sim \text{Categorical}\left(\mathcal{B} \cup \{\$\}, (q(X_{(L-2:L)}, b''))_{b'' \in \mathcal{B}\cup\{\$\}}\right)$$

where, for tensors A, B and perturbation weight $\gamma > 0$

$$(q(X_{(L-2:L)}, b''))_{b''\in\mathcal{B}\cup\{\$\}} = \text{softmax}\Bigg(\sum_{l=1}^{2}\sum_{b\in\mathcal{B}} A_{l,b,b''}\mathbb{1}(X_{(L-l)} = b) + \gamma\sum_{l=1}^{2}\sum_{l'=1}^{2}\sum_{b,b'\in\mathcal{B}} B_{l,l',b,b',b''}\mathbb{1}(X_{(L-l)} = b \text{ and } X_{(L-l')} = b')\Bigg)_{b''\in\mathcal{B}\cup\{\$\}}$$

where \$ represents a stop. We set $X_1 = C$ to avoid empty sequences. We set $A_{l,b,\$} = 5$ and $B_{l,l',b,b',\$} = 1/2$. We sampled

$$(A_{l,b,b'})_{b'\in\mathcal{B}} \sim \frac{5}{2}\text{Multinomial}((1/|\mathcal{B}|)_{b\in\mathcal{B}}),$$

$$(B_{l,l',b,b',b''})_{b''\in\mathcal{B}} \sim \frac{5}{4}\text{Multinomial}((1/|\mathcal{B}|)_{b\in\mathcal{B}}).$$

We set p to be this autoregressive model and perturb p by increasing γ. We sampled and tested 100 times from the same models and reported the fraction of null hypotheses rejected.

C.7. Testing without Normalized Likelihoods (Fig. 2(d))

First we sampled a pHMM model with $L = 10, |\mathcal{B}| = 4, \alpha_h = 1, \beta_h = 0.5$ which we call π. We then sampled $X_0 \sim \pi$.

We now define a distribution for sequences "descended from X_0" given a parameter t that controls the substitution rate, $\kappa(\cdot|X, t)$. We let $\kappa(\cdot|X, t)$ be a pHMM with the same indel process as π but with

$$h_{l,b}|X, t = \log\left(1 + t^{-1}\mathbb{1}(X_{(l)} = b)\right)$$

so that sequences are less likely to have substitutions when t is smaller, i.e. "the sequence is more closely related to X".

Next we sampled $Y_1, \ldots, Y_5 \sim \kappa(\cdot|X_0, t = 1)$. Finally, we defined the posterior reconstruction of X_0 as $p(X|Y_1, \ldots, Y_5, t) \propto \pi(X)\prod_{i=1}^{5}\kappa(Y_i|X, t)$. To sample from this model, we performed path-auxiliary MCMC sampling (Sun et al., 2022). Finally we sampled from $p(X|Y_1, \ldots, Y_5, t)$ for a variety of perturbation weights $\gamma = t - 1$ and test whether sequences came from the correct posterior $p(X|Y_1, \ldots, Y_5, t = 1)$ for which we can calculate unnormalized likelihoods. We sampled and tested 25 times from the same models and reported the fraction of null hypotheses rejected.

C.8. Comparing Approximating the KSD-B and Shrinking the Graph (Fig. 3)

Baum et al. (2022) suggested constructing a computationally tractable KSD-B by replacing the graph M with a smaller altered graph. To do so, we identify each letter in $\mathcal{B}$ with a number modulo $|\mathcal{B}|$, $0, 1, \ldots, 19 \in Z_{|\mathcal{B}|}$. Conceptually, we assume letters that are closer to 0 in $Z_{|\mathcal{B}|}$, that is $0, 1, -1 = 19, 2, -2 = 18, \ldots$ are more hydrophobic and those close to 10 are hydrophilic. Next we define the altered graph $M_{(\tau)}$ so that $XM_{(\tau)}Y$ if X and Y differ by a single substitution of a letter b to b' with $|b - b'| \leq \tau$. For instance, if $\tau = 2$, $b = 1$ is connected only to $b' \in \{-1, 0, 1, 2, 3\}$ in $M_{(\tau)}$. Then the KSD according to this method is

$$E_{X,X' \sim q} \sum_{Y M_{(\tau)} X} \sum_{Y' M_{(\tau)} X'} T_{p,X \to Y} T_{p,X' \to Y'} k((X, Y), (X', Y')).$$

We sampled a pHMM p with $L = 15, |\mathcal{B}| = 20, \alpha_h = 0.1, \beta_h = 0.5$. We then perturbed p to generate sequences with more hydrophobic residues by defining a perturbed mean field parameter $\tilde{h}$ with $\tilde{h}_{l,b} = h_{l,b} + 0.16|b - 10|$, so that $\tilde{h}_{l,b}$ were made larger when b is far from 10, i.e. close to 0. We then sampled and tested 100 samples from the perturbed and unperturbed pHMM. We then calculated the KSD as in Baum et al. (2022) with $\tau = 1, 2, 3$. We sampled and tested 100 times from the same model and reported the fraction of null hypotheses rejected.

C.9. Designing Synthesis Procedures (Fig. 4)

First we downloaded 115 thousand CDR3 protein sequences varying in length from 10 to 27 from patient 1 from 10x Genomics (2022). We train a MuE model with latent sequence length 17 with $|\mathcal{B}| = 20$ using the code from `https://github.com/pyro-ppl/pyro/tree/dev/pyro/contrib/mue`. We then train 16 stochastic synthesis models described in Weinstein et al. (2022b) varying $N_{\text{templates}} = 1, 10, 100, 1000$ and the synthesis strategy from finite nucleotide mixtures with alphabet size 8, enzymatic mutagenesis with mutazymeII, finite codon mixtures with alphabet size 24 and arbitrary codon mixtures. We trained these models using the code at `https://github.com/debbiemarkslab/variational-synthesis` under the MIT liscence. After training each model, we sampled 100 sequences from each and performed a goodness of fit test on each set of 100 sequences. We sampled and tested 25 times using $N_n = 20$ from the same models and reported the fraction of null hypotheses rejected.

C.10. Evaluating Large Models Fit to Protein Families (Fig. 5)

We first gathered data sets of evolutionarily related sequences of four protein regions studied in Shin et al. (2021): YAP1_HUMAN (36 AA), IF1_ECOLI (72 AA), CALM1_HUMAN (149 AA), and TRPC_THEMA (252 AA). We held out 20% of the sequence from each set and trained a deep generative autoregressive model (Wavenet; Shin et al. (2021)) on the held-in sequences. To do so we used code from `https://github.com/debbiemarkslab/SeqDesign/`. We then performed the KSD-B goodness of fit test, comparing the trained model to the held-out sequences. We also tested the goodness of fit of a large transformer model trained on a data set of all known proteins (Tranception; Notin et al. (2022)) using code from `https://github.com/OATML-Markslab/Tranception`. For n varying from 10 to 1000, we sample 5 sets of $N_n = 10$ mutants for each observed sequence and calculate the KSD-B for Wavenet and Tranception. We then performed a goodness of fit test for each set of sequences and models.

We chose k to be a scalar or vector field embedding kernel in our KSD-B. In particular, we considered the scalar field kernel $k_{F,\text{IMQ},\sigma}$. We also build a vector field kernel $k_{F,\text{vf}} = k_\delta + k_{\text{HT}}$ with $k_{\text{HT}}((X, Y), (X', Y')) = k_{F,\text{IMQ},sigma}(Y, Y')\mathbb{1}(|X| \geq |Y|)\mathbb{1}(|X'| \geq |Y'|)$ and $k_\delta((X, Y), (X', Y')) = (k_{F,\text{EXP},\sigma}(X, X') + k_{F,\text{EXP},\sigma}(Y, Y'))^2$ for $(X, Y), (X', Y') \in M^\sigma$. Note that our theoretical results on discrete masses in vector field kernels show that these kernels, like the scalar field kernels, are likely to have discrete masses; moreover, though we have not proven that the proposed kernel satisfies the assumptions of Thm. 5.2 (detecting non-convergence), its form is chosen based on our theoretical analysis in Section 7 and B.4.3.

We also compare the power of goodness of fit tests of two scalar field kernels $k_{F,\text{IMQ},\sigma}$ and $k_{F,\text{vf}}$ in Fig. 5. In Fig. 7 we evaluate how accurate our KSD-B approximations are. Finally we also compare the power of goodness of fit tests of two scalar field kernels $k_{F,\text{IMQ},\sigma}$ and $k_{F,\text{EXP},\sigma}$ in Fig. 8.

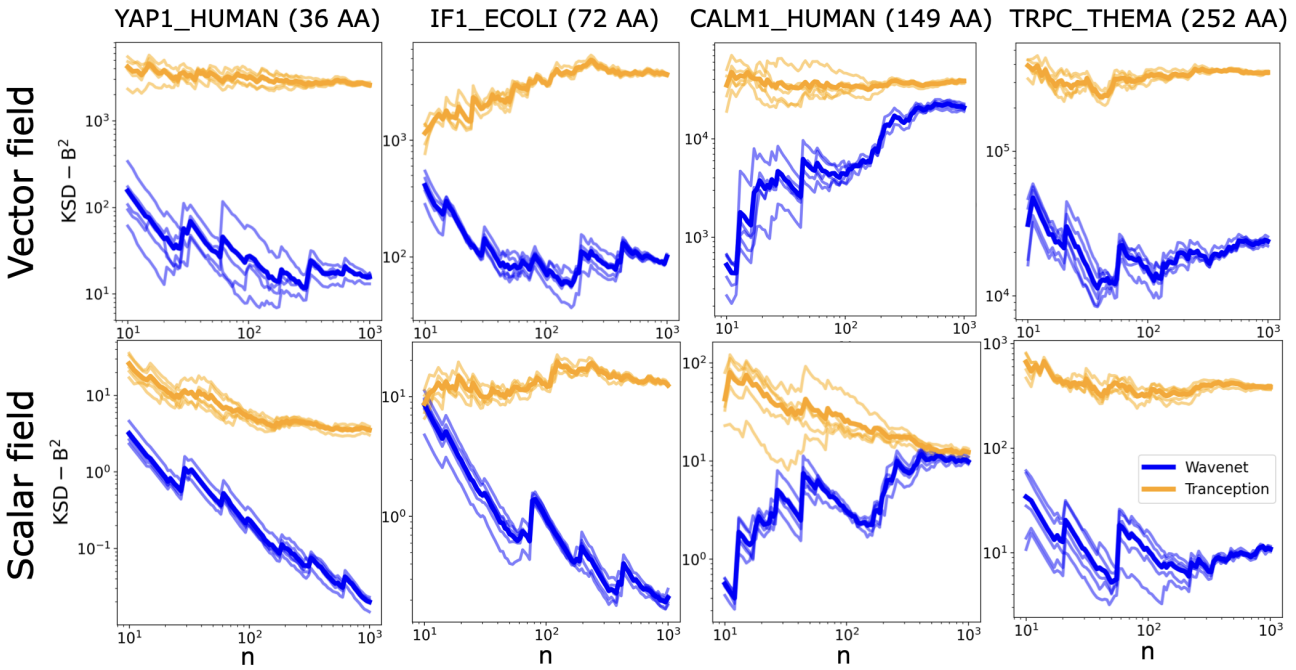


Figure 7. **Reliability of KSD-B Approximations** We examine the variance of KSD-B estimated across 5 independent samples of the $N_n = 10$ mutants for four protein families. We plot our estimates for different n for Wavenet (Blue) and Tranception (Orange) and for the vector field KSD-B (top row) and the scalar field KSD-B (bottom row).

C.10.1. Testing Protein Universe Models with the KSD-B

There has been some debate over the value of family-specific protein models versus models of the entire "protein universe", such as Tranception (Notin et al., 2022), Progen2 (Nijkamp et al., 2022), ESM1v (Meier et al., 2021), and UniRep (Alley et al., 2019). Of course, a model trained to fit the set of all proteins will achieve much smaller likelihoods on a particular protein family than a model trained to only fit that family. Despite their low likelihoods, "protein universe" models are able to predict the effects of mutations on a protein just as well as generative models trained on a single protein family (Notin et al., 2022). It is hypothesised that this is due to the fact that protein-universe models fit the distribution of sequences well "locally".

The KSD-B can be used to help compare family-specific and protein universe models, and understand their relative performance. In particular, the KSD-B is an excellent tool for evaluating protein universe models' performance on specific protein families, because it uses only local differences in likelihood (the likelihood's "slope") to evaluate goodness of fit, rather than the likelihood itself (Eqn. 1). Formally, one can hypothesise that for a particular protein family, say YAP1, the distribution learned by a protein universe model may be written as $p_{\text{ProteinUniverse}} = \alpha\mu + (1-\alpha)\nu$, where μ is a distribution just over the YAP1 protein family, ν is a distribution over everything else (which has little mass on YAP1), and α is a very small number that represents the fraction of the protein universe that is in the YAP1 family. We can think of μ as describing the "local fitness landscape" of YAP1; it is the part of the model responsible for accurate mutation effect predictions for YAP1. Now, if q is the true distribution of YAP1 sequences found in nature, we have $\text{KSD-B}_{p_{\text{ProteinUniverse}},k}(q) \approx \text{KSD-B}_{\mu,k}(q)$, where equality holds when the support of ν is not connected to μ at all, i.e. the protein family is completely isolated in sequence space. Thus, the KSD-B can be used to check the local fit of the model, μ, to the protein family distribution q.

In Fig.5 we see no evidence that the protein universe model (Tranception) learns a better model of local, family sequence distributions than a family-specific model (Wavenet).

D. Supplementary Code

The supplementary code (`https://github.com/AlanNawzadAmin/KSD-B/`) provides a Jupyter notebook (`KSD-B theory example.ipynb`) recreating Fig. 1(a) and 1(b) using the IMQ-H (U), IMQ-H (N), and IMQ-H+Exp-H kernels.

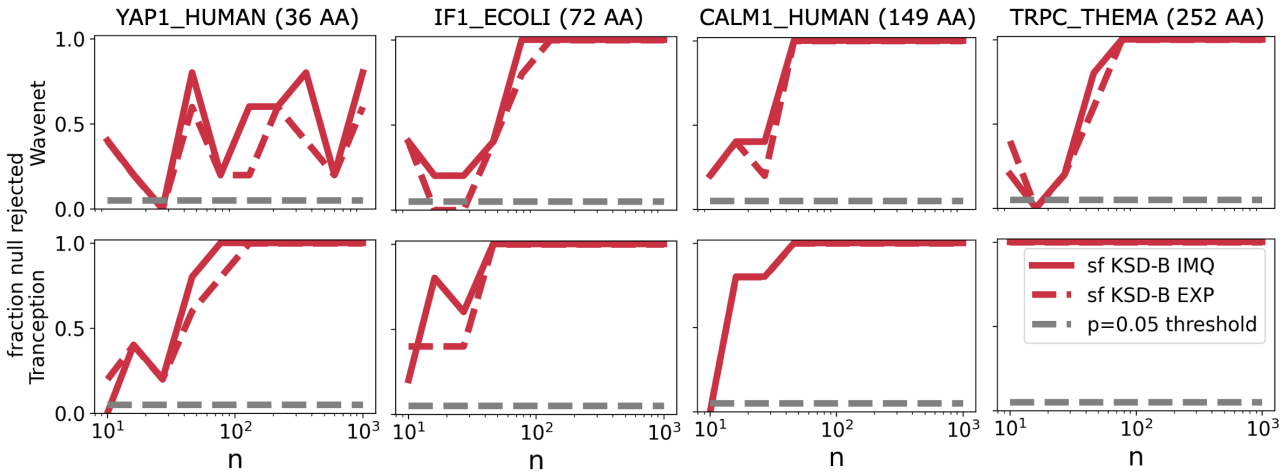


Figure 8. **Scalar Field KSD-B Power using Embedding Kernels** We perform a goodness of fit test with two scalar field embedding kernels across 5 independent samples of the $N_n = 10$ mutants for four protein families. We use the IMQ (solid) and EXP (dashed) embedding kernels and see that our goodness of fit tests have similar power for the two kernels.