Kernel-Based Evaluation of Conditional Biological Sequence Models

Pierre Glaser [1] **Steffanie Paul** [2] **Alissa M. Hummer** [2 3] **Charlotte M. Deane** [3] **Debora S. Marks** [4] **Alan N. Amin** [5]

Abstract

We propose a set of kernel-based tools to evaluate the designs and tune the hyperparameters of conditional sequence models, with a focus on problems in computational biology. The backbone of our tools is a new measure of discrepancy between the true conditional distribution and the model's estimate, called the Augmented Conditional Maximum Mean Discrepancy (ACMMD). Provided that the model can be sampled from, the ACMMD can be estimated unbiasedly from data to quantify absolute model fit, integrated within hypothesis tests, and used to evaluate model reliability. We demonstrate the utility of our approach by analyzing a popular protein design model, ProteinMPNN. We are able to reject the hypothesis that ProteinMPNN fits its data for various protein families, and tune the model's temperature hyperparameter to achieve a better fit.

1. Introduction

Conditional sequence models constitute one of the most prominent model classes of modern machine learning. Such models have allowed progress in longstanding problems in fields ranging from natural language generation to biomedical applications such as genomics and protein design. Abstracting away the precise nature of the data, the objective common to many of these problems can be summarized as the prediction of high-dimensional discrete-valued sequences, given some possibly high-dimensional input information about the sequence. For example, in protein design, inverse folding models (Dauparas et al., 2022) seek to learn the conditional distribution of amino acid sequences (proteins) that are likely to fold to a given input protein backbone 3D geometry, or *structure*.

In such problems, it is crucial to evaluate the properties of the trained model. Model evaluation can help assess the risk of using the model's predictions in the real world, such as performing in-vitro experiments (a time-intensive process), guide hyperparameter searches, and deepen one's understanding of the model's behavior. Two properties are particularly important to measure: the first one is model *accuracy*, which describes how well the model approximates the true conditional distribution of the target variable given the input. Models with high accuracy have learned the underlying structure of the data, suggesting a high potential value in deploying them in real-world applications. However, in practice, it is likely that models will not be perfectly accurate. Inaccurate models can still be useful as long as they fall back to conservative guesses (in the extreme case, the prior distribution) when they are uncertain. From a statistical perspective, this property is known as *reliability* (Bröcker, 2008; Vaicenavicius et al., 2019; Widmann et al., 2021), and will be the second property of interest in this work.

Given a set of real samples, the standard approach to evaluate models in protein design consists in using log-likelihoods or sequence recovery (Dauparas et al., 2022; Hsu et al., 2022; Gao et al., 2022). However, log-likelihoods cannot be used to evaluate reliability, and are only *relative* measures of accuracy: these methods can only be used to compare models and would not alert the practitioner for example if all models make very poor predictions. Instead, to assess how far a model is from being optimally accurate and consistent — and thus the potential value in improving it, by for example collecting more data or increasing its complexity, one should consider *absolute* rather than *relative* metrics, that is, metrics that not only allow one to compare models to each other, but also to evaluate a single model's performance without any other point of comparison. For these metrics to have practical value, they should come with estimators computable from data samples. These estimators should be efficiently computable, recover the true metric as in the large sample size limit (i.e. be *consistent*), and preferably be centered around the true value of the metric (i.e. be *unbiased*). Finally, to factor out the statistical error coming from estimating these metrics using a finite number of samples, these metrics should be integrable into hypothesis tests built to detect *statistically significant* mismatches

[1] Gatsby Computational Neuroscience Unit, London, UK [2] Systems Biology, Harvard Medical School, Boston, USA [3] Department of Statistics, University of Oxford, Oxford, UK [4] Harvard Medical School, Broad Institute, Boston, USA [5] Courant Institute, New York University, New York, USA. Correspondence to: Pierre Glaser <pierreglaser@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

between the model and the data.

Contributions In this work, we introduce a set of absolute evaluation metrics for measuring the accuracy and the reliability of conditional sequence models. Both our metrics are grounded in a new measure of divergence between conditional probability distributions, which we call the Augmented Conditional Maximum Mean Discrepancy (ACMMD), which extends the kernel-based conditional goodness-of-fit framework of Jitkrittum et al. (2020); Glaser et al. (2023); Widmann et al. (2021) to the case of sequence-valued variables. We analyze the statistical properties of our proposed metrics, which can be estimated using samples from the data and the model. Under certain conditions, we show that the ACMMD is able to detect any mismatch between the model and the data. In addition, we integrate the ACMMD into hypothesis tests to detect such mismatches from the model and the data samples. We showcase the utility of our methods by using them in an in-depth analysis of a popular inverse folding model - ProteinMPNN (Dauparas et al., 2022). Our results demonstrate the theoretical properties of our methods, while also providing insight as to how to gauge the certainty and applicability of ProteinMPNN for designing proteins of varying topologies and evolutionary families.

2. Problem Setting

We consider the problem of predicting a discrete sequence-valued variable we are designing $Y \in \mathcal{Y}$, for example a biological sequence, conditionally on a variable $X \in \mathcal{X}$ at our disposal. The predicted sequence Y is allowed to have an arbitrary length, e.g. $\mathcal{Y} = \cup_{\ell=1}^{\infty} \mathcal{A}^{\ell}$, where $\mathcal{A}$ is a finite set. In protein design, X could be the 3D structure of a protein (e.g. $\mathcal{X} = \cup_{\ell=1}^{\infty} \mathbb{R}^{3\ell}$) and Y the sequence of amino acids making up the protein, in which case $\mathcal{A}$ is the set of amino acids. Given a large number of i.i.d measurements $\{X_i, Y_i\}_{i=1}^{N_T}$ from a distribution $\mathbb{P}(X, Y)$, for example pairs of sequences and structures from the Protein Data Bank (Ingraham et al., 2019), we train a *predictive* model $Q_{|} : x \longmapsto Q_{|x}$ that takes in a value x and outputs a distribution on Y, $Q_{|x}(Y)$ that attempts to match the true conditional $\mathbb{P}(Y|X = x)$, denoted $\mathbb{P}_{|x}(Y)$ in this work. After training, we are interested in quantifying how accurately $Q_{|}$ approximates $\mathbb{P}_{|}$ on average across all values of x after training, using a held-out set of samples $\{X_i, Y_i\}_{i=1}^{N} \sim \mathbb{P}(X, Y)$. Quantifying the accuracy of $Q_{|}$ is known as the *conditional goodness-of-fit* problem, and we address it in Section 3. Furthermore, we will also be interested in quantifying the reliability of $Q_{|}$, a task which we address in Section 4.

3. Conditional Goodness–of–Fit with ACMMD

In this section, we propose a metric that quantifies the accuracy of a predictive sequence model. We will show that this metric satisfies many desirable properties: first, it is absolute and able to detect any differences between conditional distributions. Second, it can be unbiasedly and efficiently estimated using samples from the model and the data distribution. Third, it can be used in hypothesis tests to detect statistically significant mismatches from such samples.

3.1. The Augmented Conditional MMD

We now propose a method to quantitatively evaluate the conditional goodness–of–fit of $Q_{|}$ to $\mathbb{P}_{|}$. Our approach consists in constructing a *divergence* $\mathrm{D}(\mathbb{P}_{|}, Q_{|})$, between the conditional distribution of Y given X and the model $Q_{|}$. By definition, this divergence should satisfy:

$$\begin{aligned} &(i)\ \ \mathrm{D}(\mathbb{P}_{|}, Q_{|}) \geq 0 \\ &(ii)\ \ \mathrm{D}(\mathbb{P}_{|}, Q_{|}) = 0 \iff \mathbb{P}_{|x} = Q_{|x},\ \mathbb{P}(X)\text{–a.e.} \end{aligned} \tag{1}$$

Combined, these two properties ensure that $\mathrm{D}(\mathbb{P}_{|}, Q_{|})$ is *absolute*, e.g. assigns the known value lowest value 0 to the best possible model, and is able to distinguish any mismatch between the model and the data, which is crucial to prevent blind spots in our evaluation. We borrow the idea of comparing $Q_{|}$ with $\mathbb{P}_{|}$ by comparing the joint $\mathbb{P}(X, Y)$ with a joint that keeps the same marginal $\mathbb{P}(X)$ but swaps $\mathbb{P}_{|}$ with $Q_{|}$. These two joint distributions are equal if and only if $Q_{|}$ and $\mathbb{P}_{|}$ match almost everywhere. To compare these two distributions, we will use the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) given by:

$$\mathrm{MMD}(\mathbb{Q}_1, \mathbb{Q}_2) = \sup_{\substack{f \in \mathcal{H}_{\mathcal{Z}} \\ \|f\|_{\mathcal{H}_{\mathcal{Z}}} \leq 1}} \mathbb{E}_{\mathbb{Q}_1}[f(Z)] - \mathbb{E}_{\mathbb{Q}_2}[f(Z)]. \tag{2}$$

Here, $\mathcal{Z}$ is some measurable space, $\mathbb{Q}_1$ and $\mathbb{Q}_2$ are probability measures on $\mathcal{Z}$, and $\mathcal{H}_{\mathcal{Z}}$ is a reproducing kernel Hilbert space (RKHS) of functions from $\mathcal{Z}$ to $\mathbb{R}$ with kernel $k_{\mathcal{Z}}$ (Berlinet & Thomas-Agnan, 2011). Applying this general definition to the case at hand, we obtain a measure of accuracy for $Q_{|}$, defined below.

Definition 3.1 (Augmented Conditional MMD)**.** *Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ with law $\mathbb{P}_X \otimes \mathbb{P}_{|}$. Let $Q_{|}$ be a conditional probability from $\mathcal{X}$ to $\mathcal{Y}$. We define the Augmented Conditional* MMD *(ACMMD) between $\mathbb{P}_{|}$ and $Q_{|}$ as:*

$$\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) \coloneqq \mathrm{MMD}(\mathbb{P}_X \otimes \mathbb{P}_{|}, \mathbb{P}_X \otimes Q_{|}) \tag{3}$$

where the MMD *is evaluated with a user-specified kernel $k_{\mathcal{X}\times\mathcal{Y}}$ on $\mathcal{X} \times \mathcal{Y}$. Here, $\mathbb{P}_X \otimes \mathbb{P}_{|}$ is defined by $(X, Y) \sim \mathbb{P}_X \otimes \mathbb{P}_{|} \iff X \sim \mathbb{P}_X$, $(Y|X = x) \sim \mathbb{P}_{|x}$, and similarly for $\mathbb{P}_X \otimes Q_{|}$.*

Choice of kernel for ACMMD The ACMMD requires specifying a kernel on the joint space $\mathcal{X} \times \mathcal{Y}$. In this work, we will focus on the case where $k_{\mathcal{X}\times\mathcal{Y}}$ is the *tensor product* kernel $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ of two kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ on $\mathcal{X}$ and $\mathcal{Y}$ respectively:

$$k_{\mathcal{X}\times\mathcal{Y}}((x, y), (x', y')) = k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, y') \tag{4}$$

This choice is popular in practice, and the resulting ACMMD retains its desirable properties, as we show next.

The ACMMD is a divergence between conditional probabilities The ACMMD writes as divergence (which is symmetric, e.g. a distance) between joint distributions, while we seek to use it to compare conditional distributions. The following lemma shows that the same ACMMD can be formulated in alternative manner that highlights its purpose as a conditional distribution comparator.

Lemma 3.2. *Under mild integrability conditions, we have:*

$$\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) = \left\| T_{K_{\mathcal{X}}}(\mu_{\mathbb{P}_{|}} - \mu_{Q_{|}}) \right\|_{\mathcal{H}_{\mathcal{X},\mathcal{H}_{\mathcal{Y}}}}$$

Where $\mu_{\mathbb{P}_{|}}$ and $\mu_{Q_{|}}$ are the conditional mean embeddings (Park & Muandet, 2020) of $\mathbb{P}_{|}$ and $Q_{|}$, $K_{\mathcal{X}}(x, x') := k_{\mathcal{X}}(x, x')I_{\mathcal{H}_{\mathcal{Y}}}$ (here, $I_{\mathcal{H}_{\mathcal{Y}}}$ the identity operator) is an operator-valued kernel with associated vector-valued RKHS $\mathcal{H}_{\mathcal{X},\mathcal{H}_{\mathcal{Y}}} \subset L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$, and $T_{K_{\mathcal{X}}}$ is its associated integral operator from $L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ to $\mathcal{H}_{\mathcal{X},\mathcal{H}_{\mathcal{Y}}}$. Moreover, if $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are C_0-universal [1], then it holds that:

$$\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0 \iff \mathbb{P}_{|x} = Q_{|x}, \quad \mathbb{P}_X\text{-a.e.}$$

The complete statement (with the full set of assumptions, and the definition of integral operators) and its proof can be found in Appendix A. Lemma 3.2 shows that the ACMMD can be understood as the result of a two-step procedure, given by (1) computing the conditional mean embedding $\mu_{\mathbb{P}_{|}} : x \longmapsto \mathbb{E}_{\mathbb{P}_{|x}}[k_{\mathcal{Y}}(y, \cdot)|X = x]$ of $\mathbb{P}_{|}$ (resp. of $Q_{|}$), which is a function from $\mathcal{X}$ to $\mathcal{H}_{\mathcal{Y}}$, and (2) embed the difference of these conditional mean embeddings into the *vector-valued* RKHS $\mathcal{H}_{\mathcal{X},\mathcal{H}_{\mathcal{Y}}} \subset L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{Y}})$ with kernel $K_{\mathcal{X}}$, before returning its associated RKHS norm. The second part of the lemma gives sufficient conditions for the ACMMD to discriminate between any non ($\mathbb{P}_X$–a.e) equal conditional distributions, fulfilling the requirements specified in Equation (1): these conditions are to use *universal* kernels $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$. Regarding $k_{\mathcal{Y}}$, this requirement is not very restrictive, as many universal kernels on sequences have been shown to be universal (Amin et al., 2023a). The difficulty in finding a universal $k_{\mathcal{X}}$ will depend on the space $\mathcal{X}$ (unspecified in this work) for the problem at hand.

Estimating the ACMMD from data Crucial to this work is the fact that if the model $Q_{|}$ can be sampled from for any $x \in \mathcal{X}$, ACMMD ^{2} will admit tractable unbiased estimators. To see this, we first rewrite ACMMD ^{2} in a form that will make this property apparent.

Lemma 3.3. *Let $Z := (X, Y, \tilde{Y})$ the triplet of random variables with law [2] $\mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$. Then, under the integrability assumptions of Lemma 3.2, we have that:*

$$\mathrm{ACMMD}^2(\mathbb{P}_{|}, Q_{|}) = \mathbb{E}_{Z_1,Z_2}\left[h(Z_1, Z_2)\right]$$

where Z_1, Z_2 are two independent copies of Z and h is a symmetric function given by:

$$\begin{aligned} h(Z_1, Z_2) &:= k_{\mathcal{X}}(X_1, X_2)g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) \\ g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) &:= k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) + k_{\mathcal{Y}}(Y_1, Y_2) \\ &\quad - k_{\mathcal{Y}}(\tilde{Y}_1, Y_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) \end{aligned}$$

Lemma 3.3, proved in Appendix B.2, expresses ACMMD ^{2} as a double expectation given two independent *samples* of $(X, Y, \tilde{Y}) \sim \mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$. Leveraging this fact, we can derive an unbiased and consistent estimator for ACMMD ^{2}.

Lemma 3.4. *Let $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \overset{i.i.d}{\sim} \mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$ be samples from the data and the model. Then an unbiased estimator $\widehat{\mathrm{ACMMD}}^2(\mathbb{P}_{|}, Q_{|})$ of $\mathrm{ACMMD}^2(\mathbb{P}_{|}, Q_{|})$ is given by:*

$$\frac{2}{N(N-1)} \sum_{1 \le i < j \le N} h((X_i, Y_i, \tilde{Y}_i), (X_j, Y_j, \tilde{Y}_j)) \tag{5}$$

Lemma 3.4, proved in Appendix B.2, shows that it is possible to unbiasedly estimate ACMMD^2 even when the analytical model expectations are intractable, provided that one can sample from the model. This estimator takes the form of a U-statistics (Serfling, 2009, Chapter 5) with symmetric probability kernel h, which are well-studied in the statistics literature. In particular, they provide a generic framework to obtain minimal-variance analogues of unbiased estimators (Serfling, 2009, Chapter 5, p. 176). $\widehat{\mathrm{ACMMD}}^2$ is a *consistent* estimator of ACMMD^2: under the integrability assumptions of Lemma 3.2 the strong law of large numbers applies (Serfling, 2009, Section 5.4, Theorem A), and we have: $\widehat{\mathrm{ACMMD}}^2(\mathbb{P}_{|}, Q_{|}) \xrightarrow[N\to\infty]{a.s.} \mathrm{ACMMD}^2(\mathbb{P}_{|}, Q_{|})$. We provide a more detailed characterization of the asymptotic distribution of $\widehat{\mathrm{ACMMD}}^2$ in Appendix B.

3.2. Testing Conditional Goodness–of-Fit with ACMMD

In the limit of infinitely many samples, a positive ACMMD means that the model and the data differ. However, in

[1] A kernel k is C_0-universal if the associated RKHS $\mathcal{H}_k$ is dense in $C_0(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$ vanishing at infinity (Sriperumbudur et al., 2010)

[2] Identifying $Q_{|}$ with its analogue Markov kernel $\tilde{Q}_{|}$ from $(\mathcal{X} \times \mathcal{Y}, \mathscr{X} \otimes \mathscr{Y})$ such that $\tilde{Q}_{|(x,y)}(dy') := Q_{|x}(dy')$.

practice, when only a finite number of samples are available, our estimate $\widehat{\mathrm{ACMMD}}^2$ is only a noisy version of the true ACMMD^2, meaning we cannot conclude whether the model fits the data by directly inspecting its value. Instead, we need a procedure that accounts for the estimation noise; we achieve this by using the ACMMD as part of a hypothesis test deciding between two different hypotheses:

$$\begin{cases} H_0 : \mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0 \\ H_1 : \mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) > 0 \end{cases}$$

In particular, we construct a test that takes as input a sample $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N$ from the data and the model and outputs a (binary) decision to reject (or not) the null hypothesis H_0 based on whether $\widehat{\mathrm{ACMMD}}^2(\mathbb{P}_{|}, Q_{|})$ exceeds a certain threshold. Because of the estimation noise arising from the use of finitely many samples, such a test cannot systematically output the right decision. Nonetheless, we build our test to ensure a *false rejection* (e.g. reject H_0 while $\mathbb{P}_{|} = Q_{|}$ a.e) rate of $\alpha \in (0, 1)$, a common practice in statistical testing (Gretton et al., 2012). To do so, we would like to set the rejection threshold $q_{1-\alpha}$ to be an estimate of the $1 - \alpha$ quantile of the distribution of $\widehat{\mathrm{ACMMD}}^2(\mathbb{P}_{|}, Q_{|})$ under H_0. However, since $q_{1-\alpha}$ is not available in closed form, we instead compute an estimate $\widehat{q}_{1-\alpha}$ using the wild bootstrap procedure (Arcones & Giné, 1992). This procedure draws B samples $\{\widetilde{\mathrm{ACMMD}}_b^2\}_{b=1}^B$ of the form:

$$\widetilde{\mathrm{ACMMD}}_b^2 := \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N}^{N} W_i^b W_j^b h(Z_i, Z_j) \quad (6)$$

where $\{W_i^b\}_{i=1 \ldots N}^{b=1 \ldots B}$ are i.i.d. Rademacher random variables independent of the data, from which we compute a quantile estimate $\widehat{q}_{1-\alpha}$ of this distribution of samples (see Appendix C for a precise definition of $\widehat{q}_{1-\alpha}$). Importantly, this procedure guarantees an exact control of the false rejection rate at level α. We prove this fact in Appendix C.3, where we cast the wild bootstrap procedure as a Monte-Carlo estimation of the distribution of $\widehat{\mathrm{ACMMD}}^2$ when $\mathbb{P}_{|} = Q_{|}$, which is valid non-asymptotically. Our test, which we call the ACMMD test, is summarized in Algorithm 1. To the best of our knowledge, this is the first conditional goodness-of-fit test that is applicable to sequence models.

Algorithm 1 ACMMD Conditional Goodness–of–fit Test

```
Input: $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$
Parameters: Level $\alpha$, kernel $k_{\mathcal{X}}$, kernel $k_{\mathcal{Y}}$
// Estimate ACMMD using Equation (5)
$\widehat{\mathrm{ACMMD}}^2 \leftarrow \frac{2}{N(N-1)} \sum_{\substack{i,j=1 \\ i<j}}^{N} h((X_i, Y_i, \tilde{Y}_i), (X_j, Y_j, \tilde{Y}_j))$
Sample $\{\widetilde{\mathrm{ACMMD}}_b^2\}_{b=1}^B$ using Equation (6)
$\widehat{q}_{1-\alpha} \leftarrow$ approx. $(1-\alpha)$-quantile of $\{\widetilde{\mathrm{ACMMD}}_b^2\}_{b=1}^B$
if $\widehat{\mathrm{ACMMD}}^2 \leq \widehat{q}_{1-\alpha}$ then
  Fail to reject $H_0$
else
  Reject $H_0$
end if
```

4. Assessing Reliability with ACMMD

In practice, our model $Q_{|}$ may not fit the data perfectly, and it is important to distinguish (at a given level of inaccuracy) models that remain consistent with their training data from ones that fail more drastically. In this section, we show how the ACMMD can be used to evaluate model reliability, a statistical property capturing model and data consistency.

Problem Setting A model $Q_{|}$ is said to be reliable (Bröcker, 2008; Vaicenavicius et al., 2019; Widmann et al., 2021) if the distribution of the target Y given that the model made a specific prediction q *is* this prediction q itself, e.g. if:

$$q = \mathbb{P}\left(Y \in \cdot \mid Q_{|X} = q\right) \qquad \mathbb{P}(Q_{|X})\text{–a.e.} \qquad (7)$$

Here, $Q_{|X} \in \mathcal{P}(\mathcal{Y})$ (the space of probability distributions on $\mathcal{Y}$) is the random variable obtained by evaluating the model $Q_{|}$ at a random value of the input variable X. Reliability differs from accuracy in that it does not require the model to learn all the information between X and Y, but only to make truthful predictions on average — thus, by assessing reliability, one may be able to detect models that hallucinate non-realistic sequences (such as repeats of the same token) in regions of the input space where they are inaccurate, instead of making a conservative guess, such as falling back to the prior disitribution. In particular, reliability can be used as an additional criterion to discriminate between models that are equally accurate. From a theoretical perspective, reliability and accuracy can be handled in a unified manner: indeed, Equation 7 shows that reliability is defined as an equality between the conditional distribution of Y given a model prediction q, $\mathbb{P}_{|q}^{Q} := \mathbb{P}(Y = \cdot | Q_{|X} = q)$ and a "model" of this conditional distribution mapping $q \in \mathcal{P}(\mathcal{Y})$ to itself, e.g. $Q_{|}^{\mathrm{Rel}} : q \longmapsto Q_{|q}^{\mathrm{Rel}} = q$. We thus propose to measure reliability using the ACMMD (a distance between conditional distributions) between $Q_{|}^{\mathrm{Rel}}$ and $\mathbb{P}_{|}^{Q}$.

Definition 4.1 (ACMMD for Reliability)**.** *The Augmented Conditional* MMD *for reliability (*ACMMD–Rel*) between* $\mathbb{P}_{|}$ *and* $Q_{|}$ *as:*

$$\begin{aligned} \mathrm{ACMMD–Rel}(\mathbb{P}_{|}, Q_{|}) &:= \mathrm{ACMMD}(\mathbb{P}_{|}^{Q}, Q_{|}^{\mathrm{Rel}}) \\ &= \mathrm{MMD}(\mathbb{P}_{|Q_{|X}} \otimes \mathbb{P}_{|}^{Q}, \mathbb{P}_{|Q_{|X}} \otimes Q_{|}) \end{aligned} \quad (8)$$

where the ACMMD *is evaluated with a user-specified kernel* $k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ *on* $\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}$*.*

As for the ACMMD, we will restrict our attention to the case where $k_{\mathcal{P}(\mathcal{Y}) \times \mathcal{Y}}$ is a tensor product kernel $k_{\mathcal{P}(\mathcal{Y})} \otimes k_{\mathcal{Y}}$

between a kernel on $\mathcal{P}(\mathcal{Y})$ and a kernel on $\mathcal{Y}$. Comparing the ACMMD–Rel with the ACMMD, we see that the former requires specifying a kernel *on the space of probability measures* on sequences $\mathcal{P}(\mathcal{Y})$ instead of a kernel on $\mathcal{X}$. Two important points must be addressed when working with such kernels. First, in order to have $\mathrm{ACMMD\text{–}Rel}(\mathbb{P}_|, Q_|) = 0$ if and only if $Q_|$ is reliable, we must find universal kernels defined on $\mathcal{P}(\mathcal{Y})$. Second, as many kernels on probabilities are intractable, we must design an approximation strategy to estimate the ACMMD–Rel from data.

ACMMD–Rel can detect any pattern of unreliability Our first goal is to ensure that ACMMD–Rel can detect any pattern of unreliability. As ACMMD–Rel is a specific instance of the ACMMD, we can apply Lemma 3.2, stating that if the kernels $k_{\mathcal{P}(\mathcal{Y})}$ and $k_{\mathcal{Y}}$ are universal, then

$$\mathrm{ACMMD\text{–}Rel}(\mathbb{P}_|, Q_|) = 0 \iff Q_| \text{ is reliable.} \qquad (9)$$

The task of finding a universal kernel $k_{\mathcal{Y}}$ on $\mathcal{Y}$ was addressed in Section 3; thus, it remains to find a universal kernel $k_{\mathcal{P}(\mathcal{Y})}$ on $\mathcal{P}(\mathcal{Y})$. However, to the best of our knowledge, none of the existing kernels defined on probability distributions (Carmeli et al., 2010; Szabó et al., 2015; 2016; Meunier et al., 2022; Glaser et al., 2023) have been shown to be universal when $\mathcal{Y}$ is the space of arbitrary-length sequences. In the next proposition, we show that many such kernels can be constructed by following a simple recipe.

Proposition 4.2. *Let $k_{\mathcal{Y}}$ be a kernel on $\mathcal{Y}$ vanishing at infinity (on $\mathcal{Y} \times \mathcal{Y}$). Suppose that $k_{\mathcal{Y}}$ has discrete masses, i.e. that $\delta_y \in \mathcal{H}_{\mathcal{Y}}$ for all sequences $y \in \mathcal{Y}$, where δ_y is the Dirac function at y, and let $\sigma > 0$. Then the kernel $k_{\mathcal{P}(\mathcal{Y})}$ on $\mathcal{P}(\mathcal{Y})$ defined as*

$$k_{\mathcal{P}(\mathcal{Y})}(q, q') := e^{-\frac{1}{2\sigma^2}\mathrm{MMD}^2(q,q')}, \qquad (10)$$

(where the MMD is computed in $\mathcal{H}_{\mathcal{Y}}$) is a C_0–universal kernel on the space of probability distributions $\mathcal{P}(\mathcal{Y})$ (under the topology of convergence in distribution or Total Variation, which are identical, see (Amin et al., 2021)).

The proof, provided in Appendix D.2, relies on an argument similar to prior work for universal kernels on probability measures (Carmeli et al., 2010), but tailored to the special case of sequences. Proposition 4.2 guarantees that any kernel on $\mathcal{Y}$ vanishing at infinity with the discrete mass property (Amin et al., 2023a) can be used to construct a universal kernel on $\mathcal{P}(\mathcal{Y})$. Kernels with discrete masses are studied in detail in (Amin et al., 2023a). In particular, the tilted Exponentiated Hamming Kernel $\frac{1}{|y||y'|}e^{-\lambda d_H(y,y')}$ (where $|y|$ is the length of the sequence y) is a kernel with discrete masses vanishing at infinity on $\mathcal{Y} \times \mathcal{Y}$, and can thus be used to construct a universal kernel on $\mathcal{P}(\mathcal{Y})$.

Estimating ACMMD–Rel from data To estimate ACMMD–Rel from the data $\{X_i, Y_i\}_{i=1}^N$ and samples from the model $\{\tilde{Y}_i \sim Q_{|X_i}\}_{i=1}^N$, one may try to use the general ACMMD estimator proposed in Lemma 3.4, which, specialized to the reliability setting, is given by:

$$\frac{2}{N(N-1)} \sum_{1 \le i < j \le N} h(Z_i, Z_j)$$
$$h(Z_i, Z_j) := k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_i}, Q_{|X_j}) g((Y_i, \tilde{Y}_i), (Y_j, \tilde{Y}_j))$$

This estimator requires evaluating $k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_i}, Q_{|X_j})$ for pairs i, j. Unfortunately, exact evaluation of these quantities for the universal kernels proposed in Proposition 4.2 is in general impossible, as $\mathrm{MMD}^2(Q_{|X_i}, Q_{|X_j})$ contains intractable expectations under $Q_{|X_i}$ and $Q_{|X_j}$. However, MMDs can be unbiasedly estimated using samples from $Q_{|X_i}$ and $Q_{|X_j}$ (Gretton et al., 2012; Schrab et al., 2022). Inspired by this fact, we propose the following estimator:

$$\begin{aligned} \widehat{\mathrm{ACMMD}}\text{–}\mathrm{Rel}^2 &:= \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \widehat{h}(Z_i, Z_j) \\ \widehat{h}(Z_i, Z_j) &:= \widehat{k}_{ij} \times g((Y_i, \tilde{Y}_i), (Y_j, \tilde{Y}_j)) \end{aligned} \qquad (11)$$

Here, $\widehat{k}_{ij}$ is an approximation of $k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_i}, Q_{|X_j})$ obtained by drawing R samples $\{\tilde{Y}_i^r\}_{r=1}^R$ and $\{\tilde{Y}_j^r\}_{r=1}^R$ from $Q_{|X_i}$ and $Q_{|X_j}$, and replacing the $\mathrm{MMD}^2(Q_{|X_i}, Q_{|X_j})$ term in $k_{\mathcal{P}}$ by an unbiased estimate $\widehat{\mathrm{MMD}}^2_{ij}$ computed from these samples. The full estimation procedure is provided in Algorithm 2. This additional approximation step has several implications: first, unlike $\widehat{\mathrm{ACMMD}}^2$, $\widehat{\mathrm{ACMMD}}\text{–}\mathrm{Rel}^2$ is not unbiased. However, the bias of this estimator can be controlled by increasing the number of samples R used to estimate the MMD. Moreover, we show in the next proposition that the estimator $\widehat{\mathrm{ACMMD}}\text{–}\mathrm{Rel}^2$ is still consistent provided that R is chosen appropriately.

Proposition 4.3. *Assume that $k_{\mathcal{Y}}$ is bounded. Then, if $R \equiv R(N)$, with $\lim_{N \to \infty} R(N) = +\infty$, $\widehat{\mathrm{ACMMD}}\text{–}\mathrm{Rel}^2$ converges in probability to $\mathrm{ACMMD\text{–}Rel}^2$ as $N \to \infty$.*

Testing for reliability with ACMMD–Rel As an ACMMD, ACMMD–Rel has the potential to be used to test whether a model is reliable given some available data: to do so, one can use Algorithm 1, replacing $\widehat{\mathrm{ACMMD}}^2$ by $\widehat{\mathrm{ACMMD}}\text{–}\mathrm{Rel}^2$, and performing quantile estimation using the $\widehat{h}(Z_i, Z_j)$ instead of the $h(Z_i, Z_j)$. A full description of the algorithm is provided in Appendix D.3.1. An important question to answer is whether the approximation of using $\widehat{h}$ instead of h affects the false-rejection rate of the test. We show in the next proposition that this is not the case.

Proposition 4.4. *Assume that $k_{\mathcal{Y}}$ is bounded, and $k_{\mathcal{P}(\mathcal{Y})}$ is a kernel of the form of Equation 10. Then a reliability test using $\widehat{h}(Z_i, Z_j)$ instead of $h(Z_i, Z_j)$ to estimate*

Algorithm 2 Estimating ACMMD–Rel

```
Input: $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_| \otimes Q_|$, model $Q_|$
Parameters: kernel $k_\mathcal{Y}$
for $i$ in 1 to $N$ do
    [$\tilde{Y}_i^r \sim Q_{|X_i}$ for $r$ in 1 to $R$ ]
end for
for $i, j$ in 1 to $N$ do
    %Use, e.g. Gretton et al. (2012, Equation 4)
    $\widehat{\mathrm{MMD}}_{ij}^2 := \texttt{estimate\_mmd}(\{\tilde{Y}_i^r\}_{r=1}^R, \{\tilde{Y}_j^r\}_{r=1}^R)$
    $\widehat{h}_{ij} := e^{-\frac{1}{2\sigma^2}\widehat{\mathrm{MMD}}_{ij}^2} \times g((Y_i, \tilde{Y}_i), (Y_j, \tilde{Y}_j))$
end for
return $\frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} \widehat{h}_{ij}$
```

ACMMD–Rel *and its* $(1-\alpha)$*–quantile under* H_0 *has a false-rejection rate of exactly* α.

5. Related Work

Goodness-of-fit methods The *goodness-of-fit* problem is a well-studied problem in the statistics and machine learning literature, for which many methods were developed (Chwialkowski et al., 2016; Gorham & Mackey, 2017; Grathwohl et al., 2020; Amin et al., 2023b; Baum et al., 2023). Impressively, these methods can operate directly from the model's analytical form, without requiring access to samples from the model – which may be hard to generate. In these works, goodness-of-fit is defined as the problem of evaluating the fit of *unconditional* models to their data, which is unlike the conditional goodness-of-fit problem we consider here. Evaluating conditional goodness-of-fit with kernels was recently studied in Jitkrittum et al. (2020). However, the proposed method requires the output space $\mathcal{Y}$ to be a subset of $\mathbb{R}^d$, and is thus unsuitable for conditional sequence models. The use of conditional goodness-of-fit metrics to evaluate reliability was also done in Glaser et al. (2023), in a method also limited to continuous output spaces. Finally, we note that ACMMD–Rel ^{2} recovers an existing calibration metric, the Squared Kernel Calibration Error (SKCE) of Widmann et al. (2021). However, the latter did not study the problems of universality, tractability and test validity in the case of sequence-valued outputs.

Deep Protein Design Models (Deep Learning–powered) conditional probability models have gained significant momentum in computational biology during the last decade. In particular, such models have revolutionized the protein design field (Johnson et al., 2023; Bennett et al., 2023). Inverse folding models are trained on protein structures and sequences in the protein data bank (PDB) (Ingraham et al., 2019; Hsu et al., 2022; Dauparas et al., 2022). They condition a sequence distribution on an input protein structure — thus learning what sequences would likely fold into that structure. The designs from these methods have been shown to be highly stable and retain function (Sumida et al., 2024). However, many of the leaps made using these models have used small, simple structural scaffolds (like loop-helix-loop motifs) (Bennett et al., 2023; Watson et al., 2023). Protein engineers interested in leveraging these tools for novel scaffolds need to know how accurate and reliable the model is on average. If the model is too imprecise, one might wish to gather more data and train more bespoke models before using the method to design experiments.

6. Experiments

We now investigate the behavior and utility of the ACMMD and ACMMD–Rel metrics and tests in practice. We start with a synthetic example showing that ACMMD is a natural measure of model distance. We then perform an extended analysis of a state-of-the-art inverse folding model, ProteinMPNN. We show that ACMMD can detect small perturbations in the model, and that it can be used to tune its temperature parameter. Finally, we analyze the absolute performance of ProteinMPNN.

6.1. A toy synthetic setting

We first study the behavior of ACMMD and ACMMD–Rel in a synthetic setting where the data distribution and the model are simple generative models on sequences. We set the input variable X to be a single scalar p drawn from some distribution $\mathbb{P}_X$ with support in $(0.3, 0.5)$. Y is a sequence of arbitrary length with alphabet $\mathcal{A} = \{A, B, \mathrm{STOP}\}$. We set the conditional distribution of Y given p to be:

$$p(y_n | y_{0:n-1}, x = p) = \begin{cases} A & \text{with probability } p \\ B & \text{with probability } p \\ \mathrm{STOP} & \text{with probability } 1 - 2p \end{cases}$$

so long as $y_{n-1} \neq$ STOP. The model distribution $Q_|$ is the same as the data, except for the fact that the first factor $Q_{|p}(y_0)$ is perturbed by a parameter Δp:

$$Q_{|p}(y_0) = \begin{cases} A & \text{with probability } p - \Delta p \\ B & \text{with probability } p + \Delta p \\ \mathrm{STOP} & \text{with probability } 1 - 2p \end{cases}$$

We set the kernel on $\mathcal{Y}$ to be the exponentiated Hamming distance kernel $k_\mathcal{Y}(y, y') = e^{-d_H(y,y')}$, where $d_H(y, y')$ is the Hamming distance between y and y', and $k_\mathcal{X}$ to be the Gaussian kernel $k_\mathcal{X}(p, p') = e^{-\frac{1}{2}(p-p')^2}$. With such choices, it is possible to show that:

$$\mathrm{ACMMD}(\mathbb{P}_|, Q_|) = C|\Delta p|$$

for some $C > 0$ that does not depend on Δp, and is computable in closed form for discrete priors on X. The proof and expression of C are given in Appendix D.4. From this expression, we immediately see that ACMMD is 0 only if $\Delta p = 0$, a manifestation of Lemma 3.2 which guarantees that the ACMMD can detect any mismatch between the model and the data when $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are universal. Moreover, in this case, the ACMMD depends monotonically on the shift $|\Delta p|$. Since $|\Delta p|$ represents a natural measure of how different the model is from the data, this fact suggests that the ACMMD is a natural, well-behaved measure of model distance. Additionally, we plot the average rejection rate of the ACMMD test for various number of samples and shifts in Figure 1. The results for $\Delta p = 0$ confirm that our test has the correct specified type-I error rate (0.05). Moreover, we see that the power of the test increases with the number of samples, and the shift $|\Delta p|$.

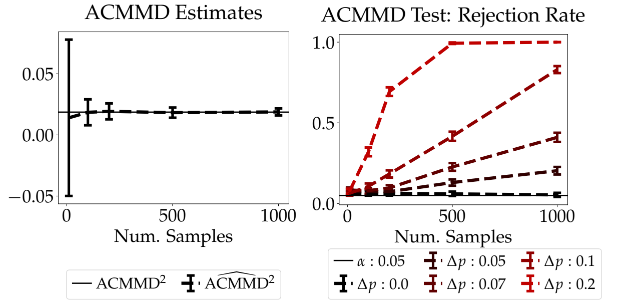

Figure 1. Left panel: ACMMD estimates for a fixed shift value $\Delta p = 0.25$ and various number of samples in the synthetic example of Section 6.1. The analytic ACMMD value is given by the horizontal line. Right panel: ACMMD test average rejection rate for various number of samples and shifts in the same setting.

6.2. ACMMD Case Study: Inverse Folding Models

To demonstrate the utility of the ACMMD measures and tests, we apply them to evaluate inverse folding models, a popular model framework used in protein design. Inverse folding models seek a distribution of amino acid sequences that are likely to fold into a given input three-dimensional structure, as discussed in Section 3. We focus our experiments on evaluating ProteinMPNN (Dauparas et al., 2022), a sampleable, commonly used model in this class. The sampling temperature T of ProteinMPNN can also be varied, letting the user control the trade-off between accuracy and diversity of the generated sequences.

Data We leveraged the CATH taxonomy to select a set of diverse (in sequence and structural topologies) protein structures to perform our ACMMD test on. CATH is a taxonomy of protein structures that categorizes proteins according to a hierarchy of structural organization (Sillitoe et al., 2021). We used the S60 redundancy filtered set which includes proteins that are at least 60% different in sequence identity from each other. Of these, we selected all single domain monomers (proteins where only one topological domain is found in the monomer), and removed any topologies that had fewer than 10 chains in its classification. This left us with 17,540 structures.

Choice of kernel Key to the performance of our metrics is the choice of the kernels $k_{\mathcal{X}}$, $k_{\mathcal{Y}}$ and $k_{\mathcal{P}(\mathcal{Y})}$. For $k_{\mathcal{Y}}$, we propose to use kernels that first embed each element – or *residue* – of a sequence y using an embedding function $\phi_{\mathcal{Y}} : \mathcal{A} \times \mathcal{Y} \longmapsto \mathbb{R}^{d_{\mathcal{Y}}}$, and evaluating a euclidean kernel on $\mathbb{R}^{d_{\mathcal{Y}}}$ on the mean of the resulting embeddings, yielding a kernel of the form:

$$k_{\mathcal{Y}}(y, y') = k_{\mathbb{R}^{d_{\mathcal{Y}}}}\left(\frac{1}{|y|}\sum_{i=1}^{|y|}\phi_{\mathcal{Y}}(y_i, y), \frac{1}{|y'|}\sum_{i=1}^{|y'|}\phi_{\mathcal{Y}}(y'_i, y')\right)$$

where we noted $y = (y_1, \ldots, y_{|y|})$. As the input space $\mathcal{X}$ is also sequence-valued, we follow the same recipe to construct our a kernel $k_{\mathcal{X}}$, using an embedding function $\phi_{\mathcal{X}} : \mathbb{R}^3 \times \mathcal{X} \longmapsto \mathbb{R}^{d_{\mathcal{X}}}$. Finally, for the kernel on $\mathcal{P}(\mathcal{Y})$, we will use a kernel of the form of Equation (10), with kernel $k_{\mathcal{Y}}$ described above to compute the inner MMD. We set our embedding functions $\phi_{\mathcal{X}}$ and $\phi_{\mathcal{Y}}$ to a pair of recent pre-trained neural networks that are commonly used for representation learning of protein sequences and structures: Gearnet (Zhang et al., 2023) for structures, and ESM-2 (Lin et al., 2023) for sequences. Such two-step kernels allow us to instill the complex structure present in the distribution of protein structures and sequences within the ACMMD maximizing the performance and meaningfulness of our evaluation pipeline. Whether Proposition 4.2 holds for these kernels is an open question, but we find that they perform well in our experiments.

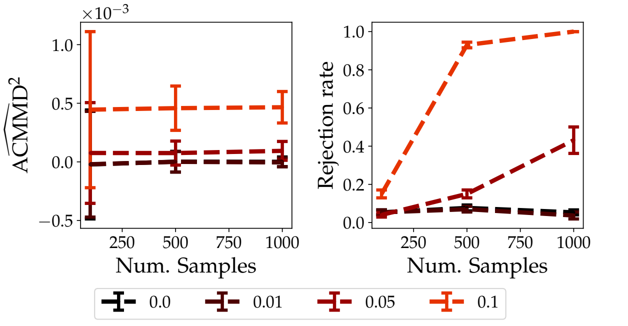

Figure 2. Values of $\widehat{\text{ACMMD}}^2$, (left) and of the average rejection rate of the ACMMD test (right) in the setting described in 6.2.1. Each line corresponds to a different value for δT.

6.2.1. The Discriminative power of ACMMD

We first propose to evaluate the behavior of the ACMMD and its associated test when comparing a known ground truth and a model distribution differing from the ground truth in a controlled manner. To this end, we set the ground truth to be a pre-trained ProteinMPNN model $Q_{|}^{T}$ with temperature T, and the model to be the same model $Q_{|}^{T+\delta T}$ with temperature $T + \delta T$. As ProteinMPNN's probability distribution is a continuous function of T, small changes in T result in small changes in the predicted distribution which will be hard to detect, translating into "low" values for $\widehat{\mathrm{ACMMD}}^{2}$ relative to larger temperature changes. Conversely, we posit that large changes in T will result in large changes in the model distribution, and will be simpler to detect by the ACMMD. To test these hypotheses, we performed an estimation of $\mathrm{ACMMD}^{\,2}$ for a ground truth temperature $T = 0.1$ (the default in the ProteinMPNN documentation) and $\delta T \in \{0, 0.01, 0.05, 0.1\}$. We used the winged helix-like DNA binding domain superfamily (CATH ID: 1.10.10.10), and performed bootstrap sampling to produce dataset sizes ranging from 100 to 1000, and 100 different random seeds in order to obtain confidence intervals of our estimates.

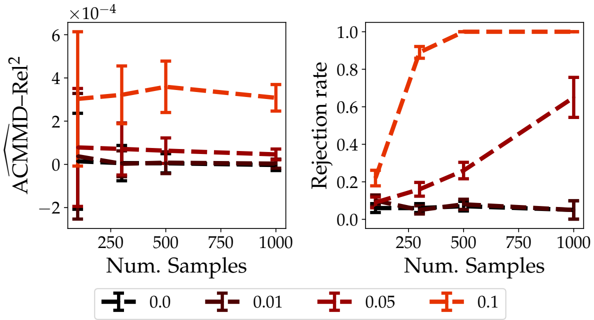

Figure 3. Values of $\widehat{\mathrm{ACMMD}}\text{–Rel}^{\,2}$, (left) and of the average rejection rate of the ACMMD–Rel test (right) in the setting described in Section 6.2.1. Each line corresponds to a different value for δT.

The results are shown in Figure 2. As expected, $\mathrm{ACMMD}^{\,2}(Q_{|}^{T}, Q_{|}^{T+\delta T})$ robustly increases with increasing values of δT. Additionally, we performed the ACMMD test of Section 3.2 with a target type-I error rate of $\alpha = 0.05$, and 100 permutations to estimate the $1 - \alpha$ quantile of the null distribution for the same values of N and δT, and computed the average rejection rate of the null hypothesis $H_0 : \mathrm{ACMMD}^{2}(Q_{|}^{T}, Q_{|}^{T+\delta T}) > 0$, which, if $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are universal, is equivalent to $H_0 : \delta T = 0$. The results, shown in Figure 2 (right), empirically confirm that the ACMMD test controls its type-I error rate and is able to detect differences in temperatures of an order relevant for ProteinMPNN. Similarly, we evaluate the behavior of the ACMMD–Rel, which is used to assess the (lack of) reliability of between model $Q_{|}^{T+\delta T}$ w.r.t the data $\mathbb{P}_{|X} \otimes Q_{|}^{T}$, the assumption being that $Q_{|}^{T+\delta T}$ is not reliable when $\delta T \neq 0$. The results are shown in Figure 3, and exhibits similar behavior.

6.2.2. Evaluation of ProteinMPNN on the CATH dataset

Now that we have confirmed the discriminative power of the ACMMD on semi-synthetic data, we use our tests to evaluate ProteinMPNN against real-world protein structures and sequences from the CATH dataset. We perform a whole-data evaluation, using samples of 5000 proteins across all families in the dataset. Then we perform a fine-grain evaluation on a subset of CATH superfamilies.

Whole-data Evaluation We first study the deviation of ProteinMPNN from the true data by computing $\widehat{\mathrm{ACMMD}}^{2}$ and estimating its mean and variance by bootstrapping over 10 random seeds. We find that ProteinMPNN with no temperature adjustment ($T = 1.0$) has an $\widehat{\mathrm{ACMMD}}^{2}$ value of 0.0916 (and a p-value < 0.01). Comparing this to the criterion values obtained on similar dataset sizes in the toy data experiments demonstrates that the model does not fit the test data. This suggests that there is still much room for improvement on solving the inverse folding problem.

On optimal temperature choices for ProteinMPNN Practitioners vary the sampling temperature as a heuristic method for sampling more certain sequences from ProteinMPNN; lower temperature settings have been found to generate sequences with fewer unrealistic artifacts (e.g. runs of alanines) which fold to more stable structures (Sumida et al., 2024). However, the relationship between sampling temperature, model reliability, and design accuracy has not been fully established. To thoroughly evaluate this, designs from different sampling temperatures conditioned on a diverse set of backbone structures would need to be experimentally characterized, which is resourse intensive in practice. We leverage the ACMMD to understand at what sampling temperature ProteinMPNN best fits the data, which gives insight as to what temperature is optimum, by computing $\widehat{\mathrm{ACMMD}}^{2}$ and $\widehat{\mathrm{ACMMD}}\text{–Rel}^{\,2}$ for varying temperature values across 10 seeds for each temperature value. The results are shown in Figure 4.

First, we observe that reducing the temperature below 1.0 improves both the model's goodness-of-fit and its reliability. This corroborates the empirical design success of lowering the sampling temperature, suggesting that greater model fit may increase the quality of samples from the model. The decrease in reliability at higher temperature shows that even though increasing the temperature increases the diversity of the model's predictions, this diversity does not necessarily capture the one of the data distribution, as for instance the prior would. The optimal temperature from the perspective of goodness-of-fit is 0.4 (which lies outside the suggested temperature range of 0.1-0.3 in the ProteinMPNN documentation (Dauparas et al., 2022)). However, we notice that model reliability continues to improve with even lower sampling temperatures while accuracy slightly increases, suggesting a trade-off between reliability and accuracy. Further experiments will determine how this trade-off manifests in the quality of designs from low-temperature settings.

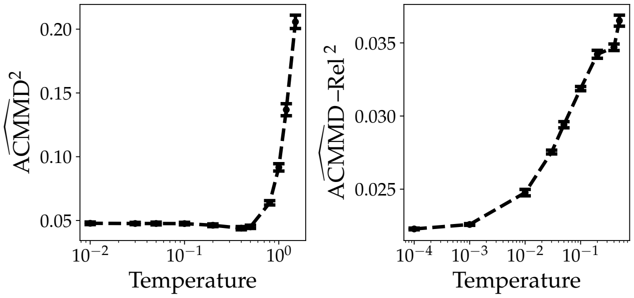

Figure 4. Evolution of $\widehat{\mathrm{ACMMD}}^2$ (left) and $\widehat{\mathrm{ACMMD}}\text{–Rel}^2$ (right) between a pre-trained ProteinMPNN model and the CATH S60 reference dataset, for varying temperature values

Structural superfamily evaluation The (H)omologous superfamily tier within the CAT(H) hierarchy groups proteins with the same similar folds and sequence identity. While ProteinMPNN has shown great performance in designing particular structural scaffolds, a practitioner aiming to leverage this model on a yet untested structural family may want some insight as to how well ProteinMPNN may fit the distribution of proteins they are interested in. Thus, we performed ACMMD evaluation separately on individual superfamilies contained in our dataset to gain insights on what types of structures ProteinMPNN does or does not fit well. We filtered the superfamilies for groupings with at least 500 proteins under a length of 100, yielding 11 families. The results are shown in Figure 5. We find that the model fit varies across families and the fit ranking is largely maintained at different temperatures. With no temperature adjustment ($T = 1.0$) the best fit superfamily (lowest $\widehat{\mathrm{ACMMD}}^2$) is the Homeodomain-like proteins (CATH ID: 1.10.10.60). These structures are largely dominated by helical bundles - a class of proteins that ProteinMPNN has demonstrated success on designing (Dauparas et al., 2022; Watson et al., 2023; Bennett et al., 2023). While the Immunoglobulin superfamily has the highest fit at lower sampling temperatures, we note that most of an immunoglobulin structure consists of the beta sandwich of the framework, while, for antibody design, engineers are often most interested in the unstructured complementarity determining regions (CDRs) of antibodies (Kunik & Ofran, 2013; Liu et al., 2020; Jin et al., 2022). As the criterion is calculated across the entire sequence, this may not reflect that ProteinMPNN has learned the distribution of CDR loops well. Further work will extend these tests to focus on subsequences of a domain to answer specific questions of model fit on regions of interest.

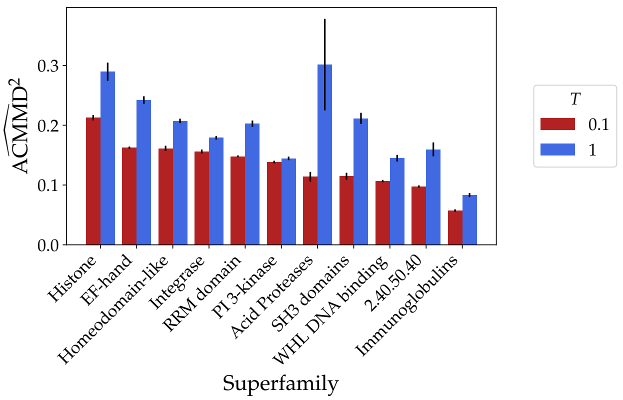

Figure 5. Value of $\widehat{\mathrm{ACMMD}}^2$ between ProteinMPNN and the CATH S60 reference dataset on a subset of 10 superfamilies for two different temperatures $T = 1.0$ and $T = 0.1$.

7. Discussion

Advancing the computational evaluation of conditional sequence models is crucial for accelerating the development of these methods for protein engineering. Given the limitations of current evaluation methods, and leveraging recent advancements in kernel methods for designing tests of goodness-of-fit and calibration, we propose a criterion and its associated test to principledly evaluate protein sequence models for how well they have learned input-conditioned sequence distributions. We discuss the statistical properties of our metrics and develop testing frameworks from them. Finally, we leverage them to investigate the performance of inverse folding models under default and temperature-adjusted settings. We develop novel insights on ideal temperature settings for ProteinMPNN and discuss the trade-off between design accuracy and model calibration that our tests demonstrate for lower temperatures. Future work can perform a more fine-grained evaluation, for example investigating which structures in particular cause the model to make unreliable predictions and what features of the model's predictions do not match the data through the use of witness functions, a by-product of MMDs (Lloyd & Ghahramani, 2015). We also note that protein engineering goals may differ from pure modeling goals, and whether performance under our metrics reflect experimental design success rates requires further investigation to determine. Yet, barring orthogonal *in silico* validation data or experimental testing, our methods offer a powerful framework to test conditional sequence models for desirable statistical properties.

Impact Statement

The tools developed in this work assess the quality of sequences predictors. As such, they have the potential to influence various procedures in protein design, and, on longer timescales, healthcare. However, the conclusions that they provide are only statistical: while they are guaranteed to hold on average, they will not hold every time. Such tools should thus be used with caution, and in conjunction with external help from domain experts to ensure that the real-world actions they will influence remain beneficial to society.

References

Amin, A., Weinstein, E. N., and Marks, D. A generative non-parametric bayesian model for whole genomes. *NeurIPS*, 34:27798–27812, 2021.

Amin, A. N., Weinstein, E. N., and Marks, D. S. Biological sequence kernels with guaranteed flexibility. *arXiv preprint arXiv:2304.03775*, 2023a.

Amin, A. N., Weinstein, E. N., and Marks, D. S. A kernelized stein discrepancy for biological sequences. In *Proceedings of the 40th ICML*, 2023b.

Arcones, M. A. and Giné, E. On the bootstrap of U and V statistics. *The Annals of Statistics*, 1992.

Baum, J., Kanagawa, H., and Gretton, A. A kernel stein test of goodness of fit for sequential models. In *ICML*, 2023.

Bennett, N. R., Coventry, B., Goreshnik, I., Huang, B., Allen, A., Vafeados, D., Peng, Y. P., Dauparas, J., Baek, M., Stewart, L., DiMaio, F., De Munck, S., Savvides, S. N., and Baker, D. Improving de novo protein binder design with deep learning. *Nat. Commun.*, 2023.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. 2011.

Bröcker, J. Some remarks on the reliability of categorical probability forecasts. *Monthly Weather Review*, 2008.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 2010.

Christmann, A. and Steinwart, I. Universal kernels on Non-Standard input spaces. In *NeurIPS*, 2010.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *ICML*, 2016.

Cuturi, M. and Blondel, M. Soft-dtw: a differentiable loss function for time-series. In *ICML*, 2017.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. A kernel for time series based on global alignments. In *ICASSP*, 2007.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 2022.

Dinculeanu, N. *Vector integration and stochastic integration in Banach spaces*. John Wiley & Sons, 2000.

Domingo-Enrich, C., Dwivedi, R., and Mackey, L. Compress then test: Powerful kernel testing in near-linear time. *AISTATS*, 2023.

Gao, Z., Tan, C., Chacón, P., and Li, S. Z. PiFold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022.

Glaser, P., Widmann, D., Lindsten, F., and Gretton, A. Fast and scalable score-based kernel calibration tests. In *UAI*, 2023.

Gorham, J. and Mackey, L. Measuring sample quality with kernels. In *ICML*, 2017.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *ICML*, 2020.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 2012.

Hoeffding, W. *On sequences of sums of independent random vectors*. Springer, 1994.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *ICML*, 2022.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. Generative models for graph-based protein design. *NeurIPS*, 2019.

Jin, W., Barzilay, R., and Jaakkola, T. Antibody-Antigen docking and design via hierarchical equivariant refinement. *ICML*, 2022.

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. Testing goodness of fit of conditional density models with kernels. In *UAI*, 2020.

Johnson, S. R., Fu, X., Viknander, S., Goldin, C., Monaco, S., Zelezniak, A., and Yang, K. K. Computational scoring and experimental evaluation of enzymes generated by neural networks. 2023.

Kriege, N. M., Johansson, F. D., and Morris, C. A survey on graph kernels. *Applied Network Science*, 2020.

Kunik, V. and Ofran, Y. The indistinguishability of epitopes from protein surface is explained by the distinct binding preferences of each of the six antigen-binding loops. *Protein Engineering, Design & Selection*, 2013.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.

Liu, G., Zeng, H., Mueller, J., Carter, B., Wang, Z., Schilz, J., Horny, G., Birnbaum, M. E., Ewert, S., and Gifford, D. K. Antibody complementarity determining region design using high-capacity machine learning. *Bioinformatics*, 2020.

Lloyd, J. R. and Ghahramani, Z. Statistical model criticism using kernel two sample tests. *Adv. Neural Inf. Process. Syst.*, 2015-Janua:829–837, 2015.

Meunier, D., Pontil, M., and Ciliberto, C. Distribution regression with sliced Wasserstein kernels. In *ICML*, 2022.

Park, J. and Muandet, K. A measure-theoretic approach to kernel conditional mean embeddings. *NeurIPS*, 2020.

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. Protein homology detection using string alignment kernels. *Bioinformatics*, 2004.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. Efficient aggregated kernel tests using incomplete u-statistics. *Advances in Neural Information Processing Systems*, 35: 18793–18807, 2022.

Serfling, R. J. *Approximation theorems of mathematical statistics*. John Wiley & Sons, 2009.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., and Orengo, C. A. CATH: increased structural coverage of functional space. *Nucleic acids research*, 2021.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. On the relation between universality, characteristic kernels and rkhs embedding of measures. In *AISTATS*, 2010.

Sumida, K. H., Núñez-Franco, R., Kalvet, I., Pellock, S. J., Wicky, B. I. M., Milles, L. F., Dauparas, J., Wang, J., Kipnis, Y., Jameson, N., Kang, A., De La Cruz, J., Sankaran, B., Bera, A. K., Jiménez-Osés, G., and Baker, D. Improving protein expression, stability, and function with ProteinMPNN. *Journal of the American Chemical Society*, 2024.

Szabó, Z., Gretton, A., Póczos, B., and Sriperumbudur, B. K. Two-stage sampled learning theory on distributions. In *AISTATS*, 2015.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. Learning theory for distribution regression. *JMLR*, 2016.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. Evaluating model calibration in classification. In *AISTATS*, 2019.

Vert, J.-P., Saigo, H., and Akutsu, T. Local alignment kernels for biological sequences. *Kernel methods in computational biology*, 2004.

Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I., Torres, S. V., Lauko, A., De Bortoli, V., Mathieu, E., Ovchinnikov, S., Barzilay, R., Jaakkola, T. S., DiMaio, F., Baek, M., and Baker, D. De novo design of protein structure and function with RFdiffusion. *Nature*, 2023.

Widmann, D., Lindsten, F., and Zachariah, D. Calibration tests beyond classification. In *ICLR*, 2021.

Zhang, Z., Xu, M., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Enhancing protein language models with structure-based encoder and pre-training. *arXiv preprint arXiv:2303.06275*, 2023.

Supplementary Material of the paper *Kernel-Based Evaluation of Conditional Biological Sequence Models*

A. Proof of Lemma 3.2

Let us first re-state the lemma in its complete form.

Lemma (Complete form of Lemma 3.2)**.** *Assume that $\mathcal{X}$ is locally-compact and second countable. Moreover, assume that $k_{\mathcal{X}\times\mathcal{Y}} = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$, and that $k_{\mathcal{X}}$, $k_{\mathcal{Y}}$ satisfy the integrability conditions $\mathbb{E}\left[k_{\mathcal{X}}(X,X)k_{\mathcal{Y}}(Y,Y)\right] < +\infty$ and $\mathbb{E}\left[k_{\mathcal{Y}}(Y,Y)\right] < +\infty$ (and similarly for $\tilde{Y}$). Then,*

$$\mathrm{ACMMD}^2(\mathbb{P}_|, Q_|) = \left\| T_{K_\mathcal{X}}(\mu_{\mathbb{P}_|} - \mu_{Q_|}) \right\|^2_{\mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}}}$$

Where $\mu_{\mathbb{P}_|}$ and $\mu_{Q_|}$ are the conditional mean embeddings (Park & Muandet, 2020) of $\mathbb{P}_|$ and $Q_|$, given by: $\mu_{\mathbb{P}_|} : x \longmapsto \mathbb{E}_{y\sim\mathbb{P}_{|x}} k_{\mathcal{Y}}(y,\cdot)$ (and similarly for $Q_|$), $K_\mathcal{X}(x,x') \coloneqq k_\mathcal{X}(x,x')I_{\mathcal{H}_\mathcal{Y}}$ is an operator-valued kernel with associated vector-valued RKHS $\mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}} \subset L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_\mathcal{Y})$, and $T_{K_\mathcal{X}}$ is its associated integral operator from $L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}})$ to $\mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}}$, defined as

$$T_{K_\mathcal{X}} f(x) = \int_\mathcal{X} K_\mathcal{X}(x,x') f(x') \mathbb{P}_X(\mathrm{d}x') \in \mathcal{H}_\mathcal{Y}$$

for all $f \in L^2_{\mathbb{P}_X}(\mathcal{X}, \mathcal{H}_\mathcal{Y})$ and $x \in \mathcal{X}$. Moreover, if $k_\mathcal{X}$ and $k_\mathcal{Y}$ are C_0-universal [3]

$$\mathrm{ACMMD}(\mathbb{P}_|, Q_|) = 0 \iff \mu_{\mathbb{P}_{|x}} = \mu_{Q_{|x}}, \quad \mathbb{P}_X\text{-}a.e.$$

Proof. Let us introduce the notations used in this proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the sample space, $X(\omega), Y(\omega), \tilde{Y}(\omega)$ being random variables on Ω corresponding to the input, target and the model. When clear, we will identify the measure $\mathbb{P}$ and the push-forwards $Y_\#\mathbb{P}$, $\tilde{Y}_\#\mathbb{P}$ and drop the dependence of $Y, \tilde{Y}$ on ω. Given $x \in \mathcal{X}$, we write K_x the linear operator from $\mathcal{H}_\mathcal{Y}$ to $\mathcal{L}(\mathcal{X}, \mathcal{H}_\mathcal{Y})$, the space of linear operators from $\mathcal{X}$ to $\mathcal{H}_\mathcal{Y}$, such that $(K_x f)(x') = K_\mathcal{X}(x,x')f \in \mathcal{H}_\mathcal{Y}$ for all $f \in \mathcal{H}_\mathcal{Y}$. When no confusion is possible, we may identify the notations $k_\mathcal{Y}(y,\cdot)$ and k_y.

The existence of the conditional mean embeddings $\mu_{\mathbb{P}_|}$ and $\mu_{Q_|}$ is guaranteed by (Park & Muandet, 2020, Definition 3.1) under the integrability assumption $\int k_\mathcal{Y}(y,y)\mathrm{d}\mathbb{P}(y) < +\infty$ and $\int k_\mathcal{Y}(\tilde{y},\tilde{y})\mathrm{d}\mathbb{P}(\tilde{y}) < +\infty$. The second integrability assumption $\int k_\mathcal{X}(x,x)k_\mathcal{Y}(y,y)\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_|)(x,y) = \int k_\mathcal{X} \otimes k_\mathcal{Y}((x,y),(x,y))\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_|)(x,y) < +\infty$ guarantees the existence of the mean embedding $\mu_{\mathbb{P}_X \otimes \mathbb{P}_|}$, defined as:

$$\int k_\mathcal{X}(x,\cdot) \otimes k_\mathcal{Y}(y,\cdot)\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_|)(x,y) \in \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$$

by (Gretton et al., 2012, Lemma 3) (and respectively for $Q_|$). Here, $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ is the tensor product Hilbert space of $\mathcal{H}_\mathcal{X}$ and $\mathcal{H}_\mathcal{Y}$, with kernel $k_\mathcal{X} \otimes k_\mathcal{Y}$. We actually prove a stronger form of the lemma, given by removing the norm from both hands of the equality and replacing it with a suitable isometric isomorphism $\phi : \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y} \longmapsto \mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}}$

$$\phi(\int k_\mathcal{X}(x,\cdot) \otimes k_\mathcal{Y}(y,\cdot)\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_|)(x,y) = \int K_x \mu_{\mathbb{P}_|} \mathrm{d}\mathbb{P}_X(x)$$

This isometric isomorphism is shown to exist in the "Currying lemma" of Carmeli et al. (2010, Example 6) regarding tensor product kernels (note that both $\mathcal{X}$ – by assumption – and $\mathcal{Y}$ are locally compact and second-countable). This lemma shows that the mapping:

$$\begin{aligned} \phi : \mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y} &\longrightarrow \mathcal{F}(\mathcal{X}, \mathcal{H}_\mathcal{Y}) \\ f \otimes g &\longmapsto \phi(f \otimes g) = (x \in \mathcal{X} \longmapsto f(x)g \in \mathcal{H}_\mathcal{Y}) \end{aligned}$$

is an isometric isomorphism between $\mathcal{H}_\mathcal{X} \otimes \mathcal{H}_\mathcal{Y}$ and $\mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}}$. This lemma gives both a representation formula for elements of $\mathcal{H}_{\mathcal{X},\mathcal{H}_\mathcal{Y}}$, and a way to formalize the currying operation, (e.g. the transformation of a function of two variables into a higher-order function of one variable and returning a function of one variable) on tensor-product spaces, since

[3] A kernel k is C_0-universal if the associated RKHS $\mathcal{H}_k$ is dense in $C_0(\mathcal{X})$, the space of continuous functions on $\mathcal{X}$ vanishing at infinity (Sriperumbudur et al., 2010)

$(f \otimes g)(x, y) = (\phi(f \otimes g)(x))(y)$. We refer to Carmeli et al. (2010, Example 6) for a proof. Proceeding with the proof of Lemma 3.2, when f and g are kernel functions $k_{\mathcal{X}}(x, \cdot)$ and $k_{\mathcal{Y}}(y, \cdot)$, the right-hand side of the equality can be related to K_x as

$$\begin{aligned} \phi(k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot))(x') &= k_{\mathcal{X}}(x, x')k_{\mathcal{Y}}(y, \cdot) \\ &= K^{\star}_{x'} K_x k_{\mathcal{Y}}(y, \cdot) \\ &= K_x k_y(x') \end{aligned}$$

where the second to last equality follows from the reproducing property of $\mathcal{H}_{\mathcal{X}, \mathcal{H}_{\mathcal{Y}}}$. Since ϕ is linear and unitary, it commutes with the mean embedding operation: (Dinculeanu, 2000, Theorem 36), yielding:

$$\begin{aligned} \phi(\int (k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot))\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y)) &= \int \phi(k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot))\mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y) \\ &= \int K_x k_y \mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y) \end{aligned}$$

To complete the proof, it remains to relate the right-hand side to the conditional mean embedding $\mu_{\mathbb{P}_{|}}$, using

$$\begin{aligned} \int K_x k_y \mathrm{d}(\mathbb{P}_X \otimes \mathbb{P}_{|})(x, y) &= \iint K_x k_y \mathrm{d}\mathbb{P}_X(x)\mathrm{d}\mathbb{P}_{|x}(y) \\ &= \int K_x \int k_y \mathrm{d}\mathbb{P}_{|x}(y)\mathrm{d}\mathbb{P}_X(x) \\ &= \int K_x \mu_{\mathbb{P}_{|}}(x)\mathrm{d}\mathbb{P}_X(x) \end{aligned}$$

as K_x is a bounded linear operator. We thus have that:

$$\phi(\int k_{\mathcal{X}}(x, \cdot) \otimes k_{\mathcal{Y}}(y, \cdot)\mathrm{d}(\mathbb{P}_{|X} \otimes \mathbb{P}_{|})(x, y) = \int K_x \mu_{\mathbb{P}_{|}} \mathrm{d}\mathbb{P}_X(x)$$

Combining this with the analogue of this result holding for $\mu_{Q_{|}}$ allows to show the stronger form of Lemma 3.2. Let us now prove the second part of the lemma. Assume $\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0$, meaning

$$T_{K_{\mathcal{X}}}(\mu_{\mathbb{P}_{|}} - \mu_{Q_{|}}) = 0$$

By Carmeli et al. (2010, Theorem 2) $K_{\mathcal{X}}$ is a C_0-universal operator-valued kernel, the operator $T_{K_{\mathcal{X}}}$ is injective. This implies that the conditional mean embeddings of $\mathbb{P}_{|x}$ and $Q_{|x}$ are equal $\mathbb{P}_X$–almost everywhere. By Park & Muandet (2020, Theorem 5.2) applied to the case where the marginals are equal, and since $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$ is C_0–universal, this implies that $\mathbb{P}_{|x} = Q_{|x}$, $\mathbb{P}_X$–almost everywhere, and in summary, $\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0$ implies $\mathbb{P}_{|x} = Q_{|x}$, $\mathbb{P}_X$–almost everywhere. To prove the reverse direction, assume that $\mathbb{P}_{|x} = Q_{|x}$, $\mathbb{P}_X$–almost everywhere Since Park & Muandet (2020, Theorem 5.2) also prove the reverse direction of the statement relied upon in the previous argument, we have that conversely $\mu_{\mathbb{P}_{|}}(x) = \mu_{Q_{|}}(x)$, $\mathbb{P}_X$–almost everywhere. By linearity of $T_{K_{\mathcal{X}}}$, we thus have that $T_{K_{\mathcal{X}}}(\mu_{\mathbb{P}_{|}} - \mu_{Q_{|}}) = 0$, and therefore $\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|}) = 0$. □

B. Asymptotic distribution of $\widehat{\mathrm{ACMMD}}^2$

As discussed in the main text, it is possible to characterize the asymptotic distribution of $N\widehat{\mathrm{ACMMD}}^2$. when $\mathbb{P}_{|} = Q_{|}$, and $\sqrt{N}(\widehat{\mathrm{ACMMD}}^2 - \mathrm{ACMMD}^2)$ when $\mathbb{P}_{|} \neq Q_{|}$. This characterization is given in the next lemma.

Lemma B.1. *Assume that the integrability assumptions of Lemma 3.2 hold, and that $\mathbb{E}_{Z_1, Z_2} h(Z_1, Z_2)^2 < +\infty$, and that*

- *if $\mathbb{P}_{|x} = Q_{|x}$ $\mathbb{P}(X)$–a.s, then $\mathbb{E}[\widehat{\mathrm{ACMMD}}] = 0$ and*

$$N\widehat{\mathrm{ACMMD}}^2 \xrightarrow{d} \sum_{j=1}^{\infty} \lambda_j(\chi^2_{1j} - 1)$$

where $\{\chi^2_{1j}\}_{j=1}^{\infty}$ are independent random χ^2_1 variables, and λ_j are the eigenvalues of the operator defined as:

$$\phi \longmapsto \int h(z, \cdot)\phi(z)d\mathbb{P}(z)$$

- *Assume moreover that $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ are C_0-universal kernels, and that $\sigma_h^2 = 4\mathbb{V}_{Z_2}[\mathbb{E}_{Z_1} h(Z_1, Z_2)] > 0$. Then*

$$\sqrt{N}(\widehat{\mathrm{ACMMD}}^2 - \mathrm{ACMMD}\,^2]) \xrightarrow{d} \mathcal{N}(0, \sigma_h^2)$$

Proof. Since $\mathbb{E}_{Z_1,Z_2} h(Z_1, Z_2)^2 < +\infty$ we have that $\mathbb{V}_{Z_1,Z_2}\mathbb{E}_{Z_1,Z_2} h(Z_1, Z_2)^2 < +\infty$. Let us define, as in Serfling (2009, Section 5.1.5), the function $h(z) = \mathbb{E}_{Z_2} h(z, Z_2)$, and define $\zeta := \mathbb{V}_z h$. For the first point, we will show that if $\mathbb{P}_| = Q_|$, $\mathbb{P}$–a.s, then $\zeta = 0$, and the result will follow from Serfling (2009, Setion 5.5.2). Indeed, noting $k = k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$,

$$\begin{aligned} h(z) &= \mathbb{E}_{Z_2} h(z, Z_2) \\ &= \mathbb{E}_{Z_2} \Big\langle k((x,y),\cdot) - k((x,y'),\cdot), k((X_2,Y_2),\cdot) - k(X_2,\tilde{Y}_2) \Big\rangle \\ &= \Big\langle k((x,y),\cdot), -k((x,y'),\cdot), \mathbb{E}_{z_2}\Big[k((X_2,Y_2),\cdot) - k((X_2,\tilde{Y}_2),\cdot)\Big]\Big\rangle \end{aligned}$$

Where we exchanged the order of integration and inner product, which is possible since $h \longmapsto \langle k((x,y),\cdot) - k((x,\tilde{y}),\cdot), h\rangle$ is a bounded linear functional for all $(x, y, \tilde{y})$. Now,

$$\mathbb{E}_{z_2} k((X_2,Y_2),\cdot) - k((X_2,\tilde{Y}_2),\cdot) = \mathbb{E}_{\mathbb{P}_X}\Big[\mathbb{E}_{\mathbb{P}|} k((X_2,Y_2),\cdot) - \mathbb{E}_{Q_|} k((X_2,\tilde{Y}_2),\cdot)\Big] = 0$$

since $\mathbb{P}_{|x} = Q_{|x}$ $\mathbb{P}_X$–a.s. Thus, $h(z)$ is a constant function, and $\zeta = 0$. The second case follows by assumption from Serfling (2009, Section 5.1.1). □

B.1. Proof of Lemma 3.3

Let $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}} := \mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ be the tensor-product RKHS of functions from $\mathcal{X} \times \mathcal{Y}$ with kernel $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$. The result can be obtained by applying a "coupling" argument, and starting from the following object:

$$\begin{aligned} \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes \mathbb{Q}_|} &:= \int \left(k((x(\omega), y(\omega)),\cdot) - k((x(\omega), \tilde{y}(\omega)),\cdot)\right) \mathrm{d}\mathbb{P}(\omega) \\ &= \mathbb{E}_{x,y,\tilde{y}}\left[k((x,y),\cdot) - k((x,\tilde{y}),\cdot)\right] \end{aligned} \tag{12}$$

We first show that $\mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes \mathbb{Q}_|}$ is a well-defined element of $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}$. Indeed, the following operator

$$T : f \in \mathcal{H} \longmapsto \mathbb{E}_z f((x,y)) - f((x,\tilde{y}))$$

satisfies

$$\begin{aligned} |Tf| &\leq \mathbb{E}\left[|f(x,y)| + |f(x,\tilde{y})|\right] \\ &\leq \|f\|_{\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}} (\mathbb{E}\sqrt{k((x,y),(x,y))} \\ &\qquad + \mathbb{E}\sqrt{k((x,\tilde{y}),(x,\tilde{y}))}) \end{aligned}$$

and is bounded thanks to the integrability assumptions of Lemma 3.2. Applying the same argument as (Gretton et al., 2012, Lemma 3), it follows that the object in Equation (12) is well-defined and belongs to $\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}$. Furthermore, by linearity of integration, we have that:

$$\begin{aligned} &\int \left(k((x(\omega), y(\omega)),\cdot) - k((x(\omega), \tilde{y}(\omega)),\cdot)\right) \mathrm{d}\mathbb{P}(\omega) \\ &= \int k((x(\omega), y(\omega)),\cdot)\mathrm{d}\mathbb{P}(\omega) - \int k((x(\omega'), \tilde{y}(\omega')),\cdot)\mathrm{d}\mathbb{P}(\omega') \\ &= \mu_{\mathbb{P}_X \otimes \mathbb{P}_|} - \mu_{\mathbb{P}_X \otimes Q_|} \end{aligned}$$

To conclude, note that:

$$\begin{aligned} \mathrm{ACMMD}(\mathbb{P}_|, Q_|)^2 &= \left\|\mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q_|}\right\|^2_{\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}} \\ &= \left\langle \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q_|}, \mu_{\mathbb{P}_X \otimes \mathbb{P}_| - \mathbb{P}_X \otimes Q_|}\right\rangle_{\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}} \\ &= \left\langle \mathbb{E}_{x,y,\tilde{y}}\left[k((x,y),\cdot) - k((x,\tilde{y}),\cdot)\right], \mathbb{E}_{x,y,\tilde{y}}\left[k((x,y),\cdot) - k((x,\tilde{y}),\cdot)\right]\right\rangle_{\mathcal{H}_{\mathcal{X}\times\mathcal{Y}}} \\ &= \mathbb{E}_{x_1,y_1,\tilde{y}_1}\mathbb{E}_{x_2,y_2,\tilde{y}_2} h((x_1,y_1,\tilde{y}_1),(x_2,y_2,\tilde{y}_2)) \end{aligned}$$

Where the last equality was obtained by exchanging the order of integration and dot product, possible thanks to the integrability assumptions of Lemma 3.2, by using the bilinearity of the inner product and the reproducing property of the kernel k. The symmetry of h in (Z_1, Z_2) follows from the symmetry of $k_{\mathcal{X}} \otimes k_{\mathcal{Y}}$.

B.2. Proof of Lemma 3.4

Proof. The proof of the unbiasedness of $\widehat{\mathrm{ACMMD}}^2$ follows by linearity of the expectation, and that each $h((X_i, Y_i, \tilde{Y}_i), (X_j, Y_j, \tilde{Y}_j))$ is an unbiased estimator of $\mathrm{ACMMD}\,^2(\mathbb{P}_{|}, Q_{|})$. □

C. Type-I error control of the ACMMD test

The goal of this section is to show that the ACMMD test is guaranteed to control its type-I error rate at level α.

C.1. Quantile estimation and Decision Rule

We first fully specify the way we compute our quantile estimate $\widehat{q}_{1-\alpha}$. Let $b_\alpha := \lceil (1-\alpha)(B+1) \rceil$. Given B bootstrap samples $\{\widetilde{\mathrm{ACMMD}}_b^2\}_{b=1}^B$ and an $\widehat{\mathrm{ACMMD}}^2$ estimate, we order them in increasing order in a sequence of size $B+1$, with ties broken arbitrarily. Let $m = \min\{b \in [\![1, B+1]\!] \,|\, \widetilde{\mathrm{ACMMD}}_b^2 = \widetilde{\mathrm{ACMMD}}_{b_\alpha}^2\}$, and $M = \max\{b \in [\![1, B+1]\!] \,|\, \widetilde{\mathrm{ACMMD}}_b^2 = \widetilde{\mathrm{ACMMD}}_{b_\alpha}^2\}$. We set $\widehat{q}_{1-\alpha}$ to be the $(m-1)$-th element with probability $(b_\alpha - (1-\alpha)(B+1))/(M-m+1)$ (with the convention that the 0-th element is $-\infty$), and the b_α-th element otherwise The decision rule is then to reject the null hypothesis if $\widehat{\mathrm{ACMMD}}^2 > q_{1-\alpha}$.

C.2. Wild-bootstrap and permutation-based approaches are equivalent in the ACMMD test

To show that the ACMMD test is guaranteed to control its type-I error rate at level α, we show that the use of a wild bootstrap procedure in the ACMMD test can be cast as a computationally efficient way to approximate the quantiles of the random variable $\widehat{\mathrm{ACMMD}}^2$ when $\mathbb{P}_{|x} = Q_{|x}$ $\mathbb{P}_X$–a.e.

Lemma C.1. *Let $\{Z_i\}_{i=1}^N$ be i.i.d realizations of $\mathbb{P}_X \otimes \mathbb{P}_{|} \otimes Q_{|}$, and let $\{W_i^b\}_{i=1 \ldots N}^{b=1 \ldots B}$ be i.i.d. Rademacher random variables independent of the data. Given a function $\sigma : [\![1, N]\!] \longmapsto \{-1, 1\}$, define $\{Z_i^\sigma\}_{i=1}^N := \{X_i, Y_i^\sigma, \tilde{Y}_i^\sigma\}_{i=1}^N$, where $(Y_i^\sigma, \tilde{Y}_i^\sigma) = (Y_i, \tilde{Y}_i)$ if $\sigma(i) = 1$, and $(\tilde{Y}_i, Y_i)$ otherwise. Then we have:*

$$\widetilde{\mathrm{ACMMD}}_b^2 = \frac{2}{N(N-1)} \sum_{\substack{i,j=1 \\ i<j}}^{N} h(Z_i^{\sigma_b}, Z_j^{\sigma_b}) := \widehat{\mathrm{ACMMD}}_{\sigma_b}^2$$

for $\sigma_b(i) := W_i^b$.

The W_i^b should be understood as elements of a random swap σ_b, which for each i, swaps Y_i and $\tilde{Y}_i$ with probability $1/2$.

Proof. Without loss of generality, we fix $i = 1$ and $j = 2$, and fix b, dropping the b index. Note that $h(Z_1, Z_2)$ and $h(Z_1^\sigma, Z_2^\sigma)$ share the same $k_{\mathcal{X}}(X_1, X_2)$. The only differing term is

$$g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) := k_{\mathcal{Y}}(Y_1, Y_2)) + k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) - k_{\mathcal{Y}}(\tilde{Y}^1, Y^2)$$

and we only need to show that $W_1 W_2 g((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) = g((Y_1^\sigma, \tilde{Y}_1^\sigma), (Y_2^\sigma, \tilde{Y}_2^\sigma))$.

Case $W^1 = W^2 = 1$ In that case, $Z_1 = Z_1^\sigma$ and $Z_2 = Z_2^\sigma$, and $W_1 W_2 h(Z_1, Z_2) = h(Z_1, Z_2) = h(Z_1^\sigma, Z_2^\sigma)$. by definition of σ.

Case $W_1 = W_2 = -1$ In that case, we have:

$$g((Y_1^\sigma, \tilde{Y}_1^\sigma), (Y_2^\sigma, \tilde{Y}_2^\sigma)) = k(\tilde{Y}_1, \tilde{Y}_2) + k(Y_1, Y_2) k(\tilde{Y}_1, Y_2) - k(Y_1, \tilde{Y}_2)) = h(Z_1, Z_2)$$

implying again $W_1 W_2 h(Z_1, Z_2) = h(Z_1, Z_2) = h(Z_1^\sigma, Z_2^\sigma)$.

Case $W_1 = 1$ **and** $W_2 = -1$ In that case, we have:

$$\begin{aligned} h(Z_1^\sigma, Z_2^\sigma) &= k((X_1, Y_1), (X_2, \tilde{Y}_2)) + k((X_1, \tilde{Y}_1), (X_2, Y_2)) - k((X_1, Y_1), (X_2, Y_2)) - k((X_1, \tilde{Y}_1), (X_2, \tilde{Y}_2)) \\ &= -h(Z_1, Z_2) \end{aligned}$$

meaning again $W_1 W_2 h(Z_1, Z_2) = -h(Z_1, Z_2) = h(Z_1^\sigma, Z_2^\sigma)$, and the last case is proved similarly. □

C.3. Level of the ACMMD test

We now show that the ACMMD test has the desired type-I error rate.

Lemma C.2. *Assume that $\mathbb{P}_{|x} = Q_{|x}$ $\mathbb{P}_X$–a.s. Then the probability that the ACMMD test rejects the null hypothesis is exactly α.*

The proof consists in 2 steps. First, we show that the decision rule is equivalent to a simpler one. Then, we analyze the latter decision rule.

An equivalent decision rule This decision rule is equivalent to the one rejecting H_0 if the position Q (with ties broken uniformly at random) of $\widehat{\mathrm{ACMMD}}^2$ in that sequence satisfies $Q > b_\alpha$, accepting it if $Q < b_\alpha$, and rejecting it with probability $b_\alpha - (1-\alpha)(B+1)$ if $Q = b_\alpha$: Indeed, $Q > M \iff \widehat{\mathrm{ACMMD}}^2 > q_{1-\alpha}$ (we always reject), $Q < m \iff \widehat{\mathrm{ACMMD}}^2 \leq q_{1-\alpha}$ (we never reject), and for both rules, when the random position Q is in $[\![m, M]\!]$, the null is rejected with probability $(b_\alpha - \alpha(B+1))/(M - m + 1)$.

Analysis of the decision rule We derive the type-I error of our decision rule by analyzing the equivalent, latter one. Our analysis follows a similar argument, in flavor, as Domingo-Enrich et al. (2023, Appendix C). Now, recall that from Lemma C.1, the wild bootstrap quantile estimation are draws of $\widehat{\mathrm{ACMMD}}$ on swapped samples Z^σ, e.g. $\{X_i, Y_i^\sigma, \tilde{Y}_i^\sigma\}_{i=1}^N$ parameterized by $\sigma : [\![1, N]\!] \longmapsto \{-1, 1\}$ where $Y_i^\sigma = Y_i$ if $\sigma(i) := w_i$ and $Y_i^\sigma = \tilde{Y}_i$ otherwise:

$$(\widetilde{\mathrm{ACMMD}}^2_b)_{b=1}^B = (\widehat{\mathrm{ACMMD}}^2_{\sigma_b})_{b=1}^B$$

using the notation of Lemma C.1. Note that σ is a random swap operator such that $\sigma(i) = 1$ with probability 0.5, and $\sigma(i) = -1$ with probability 0.5. If $\mathbb{P}_{|x} = Q_{|x}$ a.e., then since the B swap maps $\sigma_1, \ldots, \sigma_B$ are i.i.d. let us note $\widehat{\mathrm{ACMMD}}^2_{\sigma_0} = \widehat{\mathrm{ACMMD}}^2$, e.g. $\sigma_0(i) = 1$. Then the random sequence $(\widehat{\mathrm{ACMMD}}^2_{\sigma_b})_{b=0}^B$ is exchangeable. Since Q is the position of $\widehat{\mathrm{ACMMD}}^2$ within that sorted sequence, and that all positions are equally likely under exchangeability, we have:

$$\begin{aligned} &\mathbb{P}\left[Q < m\right] = 1/(B+1) \\ &\mathbb{P}\left[Q > b_\alpha\right] = \left(B + 1 - b_\alpha\right)/(B+1) \\ &\mathbb{P}\left[Q < b_\alpha\right] = \left(b_\alpha - 1\right)/(B+1) \end{aligned}$$

Noting $\Delta((X^i, Y^i, \tilde{Y}^i)_{i=1}^N)$ the event that the null hypothesis is rejected, we have:

$$\begin{aligned} \mathbb{P}\left[\Delta((X^i, Y^i, \tilde{Y}^i)_{i=1}^N)\right] &= \mathbb{P}\left[Q > b_\alpha\right] + \mathbb{P}\left[Q = b_\alpha\right]\mathbb{P}[\text{Reject}|Q = b_\alpha] \\ &= \left(B + 1 - b_\alpha\right)/(B+1) + \left(b_\alpha - (1-\alpha)(B+1)\right)/(B+1) = \alpha, \end{aligned}$$

thus showing that the ACMMD test has the desired type-I error rate. □

D. Proofs related to ACMMD–Rel

D.1. Differences between the SKCE U-statistics and the ACMMD U-statistic

We recall the definition of the SKCE U-statistics estimator from (Widmann et al., 2021, Lemma 2):

$$\widehat{\mathrm{SKCE}} = \frac{2}{N(N-1)} \sum_{1 \leq i < j \leq N} G((Q_{|X_i}, Y_i), (Q_{|X_j}, Y_j)) \tag{13}$$

where

$$\begin{aligned} G((q,y),(q',y')) &\coloneqq k_{\mathcal{P}(\mathcal{Y})\times\mathcal{Y}}((q,y),(q',y')) - \mathbb{E}_{Y\sim q} k_{\mathcal{P}(\mathcal{Y})\times\mathcal{Y}}((q,Y),(q',y')) \\ &\quad - \mathbb{E}_{Y'\sim q'} k_{\mathcal{P}(\mathcal{Y})\times\mathcal{Y}}((q,y),(q',Y')) + \mathbb{E}_{Y\sim q}\mathbb{E}_{Y'\sim q'} k_{\mathcal{P}(\mathcal{Y})\times\mathcal{Y}}((q,Y),(q',Y')). \\ &= k_{\mathcal{P}(\mathcal{Y})}(q,q') \times (k_{\mathcal{Y}}(y,y') - \mathbb{E}_{Y\sim q} k_{\mathcal{Y}}(Y,y') - \mathbb{E}_{Y'\sim q'} k_{\mathcal{Y}}(y,Y') + \mathbb{E}_{Y\sim q}\mathbb{E}_{Y'\sim q'} k_{\mathcal{Y}}(Y,Y')) \end{aligned} \tag{14}$$

Where the second equality holds when focusing on tensor product kernels. Comparing Equation (13) and Equation (14) with the expression of the ACMMD U-statistics estimator given in Equation (5) and Lemma 3.3, we see that the SKCE population criterion equals the ACMMD. However, the SKCE U-statistics estimator is different from the ACMMD U-statistics estimator: while the ACMMD U-statistics only requires samples the conditional distributions $Q_{|X}$, the SKCE U-statistics contains expectations over the conditional distributions $Q_{|X}$, which are rarely available in practice.

D.2. Proof of Proposition 4.2

Proof. We will show that the image of $q \mapsto \mu_q$ is compact, and the result will follow from (Christmann & Steinwart, 2010). Let $\mathcal{M}(\mathcal{Y})$ the Banach space of measures of sequences endowed with the total variation norm:

$$\|q\|_{\mathrm{TV}} \coloneqq q_+(\mathcal{Y}) + q_-(\mathcal{Y})$$

We recall that by the Riesz-Markov theorem, $(\mathcal{M}(\mathcal{Y}), \|\cdot\|_{\mathrm{TV}})$ can be identified with the topological dual of $C_0(\mathcal{Y})$, $(C_0(\mathcal{S})^\star, \|\cdot\|_{\mathrm{op}})$ through an isometric isomorphism $q \in \mathcal{M}(\mathcal{Y}) \longmapsto \tilde{q} \in C_0(\mathcal{Y})^\star$, and for which the following holds:

$$\tilde{q}(f) = \int_{\mathcal{Y}} f \mathrm{d}q, \quad \forall f \in C_0(\mathcal{Y}).$$

Let $B \coloneqq \{q \in \mathcal{M}(\mathcal{Y}) \mid \|q\|_{\mathrm{TV}} \leq 1\}$. As a unit ball, by the Banach-Alaoglu theorem, B is compact under the weak–$\star$ topology and contains all distributions on sequences. We will show that $q \mapsto \mu_q$ is continuous on B and the result will follow. Given that this mapping is linear, it is sufficient to show continuity at 0. Moreover, since $\{B(0_{\mathcal{H}}, r)\}_{r>0}$ is a neighborhood basis of $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, it suffices to show that there is a neighborhood $\mathcal{V}$ of the null measure in the weak-$\star$ topology such that $\| \int k_{\mathcal{Y}}(y,\cdot) dq(y)\|_{\mathcal{H}}^2 = \int k_{\mathcal{Y}}(y,y')\mathrm{d}(q\otimes q)(y,y') < 1$ for all q in $\mathcal{V}$. Since the family

$$\{q \in \mathcal{M}(\mathcal{Y}), \int f_i(x)\mathrm{d}q(x) < \epsilon,\ i \in 1,\ldots k, f_i \in C_0(\mathcal{Y})\}$$

form a neighborhood basis of the weak-$\star$ topology, we can consider candidates of this form for $\mathcal{V}$. In particular, let us set $\{f_i\} = \{x \longmapsto \sqrt{k_{\mathcal{Y}}(x,x)} \in \mathcal{C}_0(\mathcal{Y})\}$, since $k_{\mathcal{Y}} \in C_0(\mathcal{Y}\times\mathcal{Y})$, and $\epsilon = 0.5$. On this neighborhood, we have:

$$\int k(y,y')\mathrm{d}q(y)\mathrm{d}q(y') \leq \int \sqrt{k(y,y) \times k(y',y')}\mathrm{d}q(y)\mathrm{d}q(y') \leq 0.5^2 < 1, \quad \forall q \in \mathcal{V}$$

showing the continuity of the map in question. As a consequence, the image of $B_{\mathcal{M}}(S)(0,1)$ by the map $q \longmapsto \mu_q$ is compact, implying from (Christmann & Steinwart, 2010) that the kernel

$$\tilde{k}(f,g) \coloneqq \exp(-\frac{1}{2\sigma^2}\|f-g\|_{\mathcal{H}}^2)$$

is universal on that set. Thus, we have shown that $\tilde{k}$ is universal on $\mathcal{H}$ under the strong topology (e.g. the norm topology in $\mathcal{H}$). This is equivalent to the TV topology of $\mathcal{P}(S)$ since k has discrete masses by proposition 9 of (Amin et al., 2023a), and thus $k_{\mathcal{P}(\mathcal{Y})}$ is universal on $\mathcal{P}(\mathcal{Y})$. □

D.3. Proofs regarding the impact of approximate kernels

To prove the convergence of the ACMMD–Rel estimator and the validity of its test, we rely on an augmented U-statistics formulation. Let:

$$U \coloneqq (Q_{|X}, \tilde{Y}^1, \ldots, \tilde{Y}^R, \tilde{Y}, Y) \sim \mathbb{P}_{Q_{|X}} \otimes \mathbb{Q}_{|}^{\otimes r} \otimes \mathbb{Q}_{|} \otimes \mathbb{P}_{|}^{Q} \coloneqq \mathbb{U}$$

U is the random variable which, for each model $Q_{|X}$, concatenates the synthetic samples $(\tilde{Y}^1, \ldots, \tilde{Y}^R)$ used to perform the kernel approximation $\widehat{k}_{\mathcal{P}(\mathcal{Y})}(q,q')$, the synthetic sample $\tilde{Y}$ used to evaluate h, and $\tilde{Y}$, a sample from $\mathbb{P}_{|}^{Q}$ the conditional

Algorithm 3 ACMMD–Rel Conditional Goodness of fit Test

```
Input: $\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N \overset{\text{i.i.d.}}{\sim} \mathbb{P}_X \otimes \mathbb{P}_| \otimes Q_|$
Parameters: Level $\alpha$, kernel $k_\mathcal{X}$, kernel $k_\mathcal{Y}$
// Estimate ACMMD–Rel using Algorithm 2 and collect the $\widehat{h}(Z_i, Z_j)$ of Equation (11)
$\widehat{\text{ACMMD}}\text{–Rel}^2, \{h(Z_i, Z_j)\}_{1\le i<j\le N} \leftarrow$ estimate_acmmd_rel($\{X_i, Y_i, \tilde{Y}_i\}_{i=1}^N, Q_{|X_i}$)
[$W_i^b \sim$ Rademacher for $i \in 1, \dots, N$ for $b \in 1, \dots, B$]
[$\widetilde{\text{ACMMD}}_b^2 \leftarrow \frac{2}{N(N-1)} \sum_{1\le i<j\le N} W_i^b W_j^b \widehat{h}(Z_i, Z_j)$, for $b$ in $1, \dots, B$]
// See Appendix C.1 for how to compute $\widehat{q}_{1-\alpha}$
$\widehat{q}_{1-\alpha} \leftarrow$ approx. $(1-\alpha)$-quantile of $\{\widetilde{\text{ACMMD}}_b^2\}_{b=1}^B$
if $\widehat{\text{ACMMD}}^2 \le \widehat{q}_{1-\alpha}$ then
  Fail to reject $H_0$
else
  Reject $H_0$
end if
```

distribution of Y given $Q_{|X}$. Then, given N realizations $\{U_i\}_{i=1}^N$, of U, the estimator $\widehat{\text{ACMMD}}\text{–Rel}^2$ can be written as a U-statistics on the $\{U_i\}_{i=1}^N$

$$\widehat{\text{ACMMD}}\text{–Rel}^2 = \frac{2}{N(N-1)} \sum_{1\le i<j\le N} h_a(U_i, U_j)$$

where

$$h_a(U_i, U_j) \coloneqq \widehat{k}(\{\tilde{Y}_i^r\}_{r=1}^R, \{\tilde{Y}_j^r\}_{r=1}^R) \times (k_\mathcal{Y}(Y_i, Y_j) + k_\mathcal{Y}(\tilde{Y}_i, \tilde{Y}_j) - k_\mathcal{Y}(Y_i, \tilde{Y}_j) - k_\mathcal{Y}(\tilde{Y}_i, Y_j))$$

We also will note

$$\text{ACMMD}_\text{a}^2 \coloneqq \mathbb{E}_{U_1, U_2 \sim \mathbb{U} \otimes \mathbb{U}}\ h_a(U_1, U_2)$$

D.3.1. Proof of Proposition 4.4

With this formalism, we now prove that the ACMMD–Rel test has the specified type-I error rate of $\alpha \in (0, 1)$, e.g. rejects H_0 when $\mathbb{P}_{|x} = Q_{|x}$ with probability α. Indeed, straightforward adaptations of the arguments in Appendix C.2 show that that doing a wild bootstrap using the $\widehat{h}(Z_i, Z_j)$ is equivalent to estimating

$$\frac{2}{N(N-1)} \sum_{1\le i<j\le N} h_a(U_i^\sigma, U_j^\sigma) \coloneqq \widehat{\text{ACMMD}}\text{–Rel}_{\sigma_b}^2$$

where $U_i^\sigma \coloneqq (Q_{|X_i}, \tilde{Y}_i^1, \dots, \tilde{Y}_i^R, \tilde{Y}_i^\sigma, Y_i^\sigma)$, where $(Y_i^\sigma, \tilde{Y}_i^\sigma) = (Y_i, \tilde{Y}_i)$ if $\sigma(i) = 1$ and $(\tilde{Y}_i, Y_i)$ otherwise. The same argument to show that ACMMD test has the desired type-I error rate follows in this case too: Under $\mathbb{P}_{|x} = Q_{|x}$, the sequence $\{\widehat{\text{ACMMD}}\text{–Rel}_{\sigma_b}^2\}_{b=0}^B$ is exchangeable (noting $\widehat{\text{ACMMD}}\text{–Rel}_{\sigma_0}^2 = \widehat{\text{ACMMD}}\text{–Rel}^2$, e.g. $\sigma_0(i) = 1$ for all $1 \le i \le N$), and we can repeat the derivations of the proof of Lemma C.2 to show that the ACMMD–Rel test has the desired type-I error rate.

D.3.2. Proof of Proposition 4.3

We prove a slightly more general version of the proposition, for kernels of the form $\phi(d(q, q')^2)$, where ϕ is a Lipschitz function and d is a distance on $\mathcal{P}(\mathcal{Y})$. Setting $\phi = e^{-\frac{\cdot}{\sigma^2}}$, we recover the kernels of Proposition 4.3, which include the exponentiated MMD kernel.

Proposition D.1. *Assume that $k_\mathcal{Y}$ and $k_{\mathcal{P}(\mathcal{Y})}$ is a kernel of the form $k_{\mathcal{P}(\mathcal{Y})}(q, q') = \phi(d(q, q'))$, for a Lipschitz function ϕ and a function $d(q, q')$ admitting an unbiased estimator of the form $\widehat{d}(\{y_1^r\}_{r=1}^R, \{y_2^r\}_{r=1}^R)$ where $\{y_1^r\}_{r=1}^r$ and $\{y_2^r\}_{r=1}^r$ are R i.i.d samples of q and q' respectively, with variance $O(\frac{1}{R})$ (the bound in uniform in q and q'). Then, assuming $R \equiv R(N)$, with $\lim_{N\to\infty} R(N) = +\infty$, $\widehat{\text{ACMMD}}\text{–Rel}^2$ converges in probability to ACMMD–Rel^2 as $N \to \infty$.*

Proof. As discussed above, the estimator $\widehat{\mathrm{ACMMD}}\,\text{–Rel}^{\,2}$ can be written as a U-statistics on the $\{U_i\}_{i=1}^N$, where $U_i = (Q_{|X_i}, \tilde{Y}_i^1, \ldots, \tilde{Y}_i^R, \tilde{Y}_i, Y_i)$, and using the kernel h_a defined as (accounting approximating $k_{\mathcal{Y}}$ through d directly)

$$h_a(U_i, U_j) := \phi(\widehat{d}(\{\tilde{Y}_i^r\}_{r=1}^R, \{\tilde{Y}_j^r\}_{r=1}^R)) \times (k_{\mathcal{Y}}(Y_i, Y_j) + k_{\mathcal{Y}}(\tilde{Y}_i, \tilde{Y}_j) - k_{\mathcal{Y}}(Y_i, \tilde{Y}_j) - k_{\mathcal{Y}}(\tilde{Y}_i, Y_j))$$

e.g.

$$\widehat{\mathrm{ACMMD}}\,\text{–Rel}^{\,2} = \frac{2}{N(N-1)} \sum_{i<j} h_a(U_i, U_j)$$

To study the convergence in probability of $\widehat{\mathrm{ACMMD}}_a^2$ to ACMMD_a^2, we use finite-sample bounds on U-statistcs Hoeffding (1994):

$$P\left(|\widehat{\mathrm{ACMMD}}_a^2 - \mathrm{ACMMD}\,_a^2| > \|h_a\|_\infty \sqrt{\frac{\log(2/\delta)}{2\lfloor N/2 \rfloor}} \right) \leq \delta$$

for all $\delta > 0$, where, by assumption, $k_{\mathcal{Y}}$ bounded, and $\widehat{k}_{\mathcal{P}(\mathcal{Y})}$ is of the form $\phi(\widehat{d}(\{y_1^r\}_{r=1}^R, \{y_2^r\}_{r=1}^R))$ for some bounded function ϕ, implying that h_a is bounded. To show the dependence in R, we bound the difference ACMMD_a and ACMMD.

$$\begin{aligned}
|\mathrm{ACMMD}\,_a^2 - \mathrm{ACMMD}\,^2| = \mathbb{E}_{\mathbb{U},\mathbb{U}} \Big[&(\widehat{k}_{\mathcal{P}(\mathcal{Y})}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_1}, Q_{X_2})) \times (k_{\mathcal{Y}}(Y_1, Y_2) + k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) \\
&- k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) - k_{\mathcal{Y}}(\tilde{Y}_1, Y_2)) \Big] \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \, |\mathbb{E}_{\mathbb{P}_{Q_|} \otimes \mathbb{Q}^{\otimes r} \times \mathbb{P}_{Q_|} \otimes \mathbb{Q}^{\otimes r}} \widehat{k}_{\mathcal{P}(\mathcal{Y})}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - k_{\mathcal{P}(\mathcal{Y})}(Q_{|X_1}, Q_{X_2})| \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \, \|\phi\|_{\mathrm{Lip}} \, \mathbb{E}_{\mathbb{P}_{Q_|} \otimes \mathbb{Q}^{\otimes r} \times \mathbb{P}_{Q_|} \otimes \mathbb{Q}^{\otimes r}} |\widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - d(Q_{|X_1}, Q_{X_2})| \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \, \|\phi\|_{\mathrm{Lip}} \times \\
&\quad \mathbb{E}_{\mathbb{P}_{Q_|} \times \mathbb{P}_{Q_|}} \left[\mathbb{E}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} |\widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) - d(Q_{|X_1}, Q_{X_2})| \;\middle|\; Q_{|X_1}, Q_{X_2} \right] \\
&\leq 4 \|k_{\mathcal{Y}}\|_\infty \, \|\phi\|_{\mathrm{Lip}} \, \mathbb{E}_{\mathbb{P}_{Q_|} \times \mathbb{P}_{Q_|}} \left[\sqrt{\mathbb{V}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} \widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R)} \middle| \, Q_{|X_1}, Q_{X_2} \right]
\end{aligned}$$

Where the last inequality follows from Jensen's inequality and the unbiasedness of $\widehat{d}$. The result follows by applying the assumption on the variance of $\widehat{d}$ (a bound which we assume is uniform in $Q_{|X)1}, Q_{|X_2}$). □

The term $\mathbb{V}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} \left[\widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) | Q_{|X_1}, Q_{|X_2} \right]$ can be more precisely characterized depending on $\widehat{d}$. For instance, we have, when $\widehat{d}$ is a U-statistics (for instance, using the MMD estimator of Gretton et al. (2012, Lemma 6, Equation 4)) , that (Serfling, 2009, section 5.2.1) $\mathbb{V}_{\mathbb{Q}^{\otimes r} \times \mathbb{Q}^{\otimes r}} \widehat{d}(\{Y_1^r\}_{r=1}^R, \{Y_2^r\}_{r=1}^R) < \zeta(Q_1, Q_{|X_2})/R$, where $\zeta(Q_{|X_1}, Q_{|X_2}) := \mathbb{V}_{(Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2) \sim Q_{|X_1} \otimes Q_{|X_2}} (\tilde{h}((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2))$ and $\tilde{h}((Y_1, \tilde{Y}_1), (Y_2, \tilde{Y}_2)) = k_{\mathcal{Y}}(Y_1, Y_2) + k_{\mathcal{Y}}(\tilde{Y}_1, \tilde{Y}_2) - k_{\mathcal{Y}}(Y_1, \tilde{Y}_2) - k_{\mathcal{Y}}(\tilde{Y}_1, Y_2)$, which is uniformly bounded by $4 \|k_{\mathcal{Y}}\|_\infty$ for bounded kernels. Putting the two parts together, we thus have that:

$$P\left(\left\{ \widehat{\mathrm{ACMMD}}_a^2 - \mathrm{ACMMD}^2 \right\} > 4 \|k_{\mathcal{Y}}\|_\infty \sqrt{\frac{\log(2/\delta)}{2\lfloor N/2 \rfloor}} + \frac{16 \|k_{\mathcal{Y}}\|_\infty^2 \|\phi\|_{\mathrm{Lip}}}{\sqrt{R}} \right) \leq \delta$$

for all $\delta > 0$, showing the convergence in probability of $\widehat{\mathrm{ACMMD}}_a$ to ACMMD.

D.4. Additional Details for ACMMD and ACMMD–Rel in the synthetic example

D.4.1. Derivations of ACMMD in the synthetic example

We first prove that ACMMD is proportional to Δp.

Lemma D.2. *In the setting described in Section 6.1, we have*

$$\mathrm{ACMMD}^2(\mathbb{P}_|, Q_|) = C \times \Delta p^2$$

for

$$C := \iint k_{\mathcal{X}}(p,p')2(1-e^{-\lambda})\frac{(1-2p)(1-2p')}{1-4pp'(1+e^{-\lambda})/2}\left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}}+\frac{2pe^{-\lambda}}{1-2pe^{-\lambda}}+1\right)\mathrm{d}\mathbb{P}_X(p)\mathrm{d}\mathbb{P}_X(p')$$

Proof. Recall that we have

$$\mathrm{ACMMD}^2 = \mathrm{MMD}^2(\mathbb{P}_X \otimes \mathbb{P}_|, \mathbb{P}_X \otimes Q_|)^2 \quad = \int k_{\mathcal{X}}(p,p')\left(T_{11}+T_{22}-2T_{12}\right)\mathrm{d}\mathbb{P}_X(p)\mathrm{d}\mathbb{P}_X(p')$$

where

$$T_{12} = \int k_{\mathcal{Y}}(y,y')p(y|p)q(y'|p')\mathrm{d}(y)\mathrm{d}(y')$$

and T_{22} and T_{12} are defined similarly. For a sequence y, we define the the function len given by $\mathrm{len}(y) := \min\{i \in \mathbb{N} | y_i = \text{STOP}\}$, which intuitively returns the length of the sequence.

Computing T_{ij} As we will see, a lot of the computations are agnostic to whether we are computing T_{11}, T_{22} or T_{12}. Note that the exponentiated hamming distance kernel on $\mathcal{Y}$ writes as a product

$$k_{\mathcal{Y}}(y,y') = e^{-\lambda d_H(y,y')} = e^{-\lambda_y \sum_{i=0}^{\infty}\delta(y_i \neq y'_i)} = \prod_{i=0}^{\infty} e^{-\lambda\delta(y_i\neq y'_i)} = \prod_{i=0}^{\max(\mathrm{len}(y),\mathrm{len}(y'))} e^{-\lambda\delta(y_i\neq y'_i)}$$

let us define the following events

$$\begin{aligned} F(m) &:= \Big\{\min(\text{len}(y), \text{len}(y')) = m\Big\} \\ G(m,\delta m) &:= \Big\{\max(\text{len}(y), \text{len}(y')) = m + \delta m\Big\} \end{aligned}$$

which we further break down as

$$\begin{aligned} F_1(m) &= \{\mathrm{len}(y) = m\} \cap \{\text{len}(y') > m\} \\ F_2(m) &= \{\text{len}(y) > m\} \cap \{\mathrm{len}(y') = m\} \\ F_3(m) &= \{\mathrm{len}(y) = m\} \cap \{\mathrm{len}(y') = m\} \\ \implies F(m) &= F_1(m) \cup F_2(m) \cup F_3(m) \end{aligned}$$

For which the following probabilities hold:

$$\begin{aligned} P(F_1(m)) &= P(\text{len}(y) = m) \times P(\text{len}(y') > m) = ((2p)^m \times (1-2p)) \times (2p')^{m+1} \\ P(F_2(m)) &= P(\text{len}(y') = m) \times P(\text{len}(y) > m) = ((2p')^m \times (1-2p')) \times (2p)^{m+1} \\ P(F_3(m)) &= P(\text{len}(y') = m) \times P(\text{len}(y) = m) = ((2p')^m \times (1-2p')) \times ((2p)^m \times (1-2p)) \\ P(G(m,\delta m)|F_1(m)) &= P(\text{len}(y') = m+\delta m | \text{len}(y) = m, \text{len}(y') > m) = (2p')^{\delta m - 1} \times (1-2p')\delta_{(\delta m \geq 1)} \\ P(G(m,\delta m)|F_2(m)) &= P(\text{len}(y) = m+\delta m | \text{len}(y') = m, \text{len}(y) > m) = (2p)^{\delta m - 1} \times (1-2p)\delta_{(\delta m \geq 1)} \\ P(G(m,\delta m)|F_3(m)) &= \delta(\delta m = 0) \end{aligned}$$

Let us note

$$E(m,\delta m, i) := F_i(m) \cap G(m, \delta m)$$

We have that $E(m, \delta_m, i) \cap E(m', \delta m', j) = \emptyset$ if $(m, \delta m, i) \neq (m', \delta m', j)$.

$$\Omega = \bigcup_{m=0}^{+\infty}\bigcup_{i=1}^{3}\bigcup_{\delta m=0}^{+\infty} E(m, \delta m, i)$$

Using the law of total probability, we have that Thus, using the law of total probability:

$$
\begin{aligned}
T_{ij}(p,p') &= \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\sum_{\delta m=0}^{+\infty} \mathbb{P}(E(m,\delta m,i))\mathbb{E}(e^{-\lambda d_H(y,y')}|E(m,\delta m,i)) \\
&= \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\sum_{\delta m=0}^{+\infty} \mathbb{P}(F_i(m)\cap G(m,\delta m))\mathbb{E}(e^{-\lambda d_H(y,y')}|E(m,\delta m,i)) \\
&= \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\mathbb{P}(F_i(m))\sum_{\delta m=0}^{+\infty} P(G(m,\delta m)|F_i(m))\mathbb{E}(e^{-\lambda d_H(y,y')}|E(m,\delta m,i)) \\
&= \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\mathbb{P}(F_i(m))\mathbb{E}(e^{-\lambda d_H(y_{:m},y'_{:m})}|F_i(m)) \\
&\times \sum_{\delta m=0}^{+\infty} P(G(m,\delta m)|F_i(m))\mathbb{E}(e^{-\lambda d_H(y_{m+1:m+\max(\delta m,1)},y'_{m+1:m+\max(\delta m,1)})}|E(m,\delta m,i),p,p') \\
&= \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\mathbb{P}(F_i(m))\mathbb{E}(e^{-\lambda d_H(y_{:m},y'_{:m})}|F_i(m)) \times \sum_{\delta m=0}^{+\infty} P(G(m,\delta m)|F_i(m))e^{-\lambda(\max(0,\delta m-1)+\delta(m>0))} \\
&= \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\mathbb{P}(F_i(m))\mathbb{E}(e^{-\lambda d_H(y_{:m},y'_{:m})}|F_i(m)) \left(\prod_{i=1}^{\max(m-1,1)} \mathbb{E}(e^{-\lambda\delta(y_i\neq y'_i)}|F_i(m))\right)^{\delta(m\geq 2)} \\
&\times \sum_{\delta m=0}^{+\infty} P(G(m,\delta m)|F_i(m))e^{-\lambda(\max(0,\delta m-1)+\delta(m>0))}
\end{aligned}
$$

where we break down the factorized hamming distance over the sequence into the sum of the hamming distances over each coordinate, and made use of the fact that

$$
d_H(y_{m:m+\delta m}, y'_{m:m+\delta m}) = \max(0,\delta m-1)+\delta(m>0)
$$

conditioned on $F_i(m)$ and $G(m,\delta m)$. The disjunction of cases is necessary in order to not count the term 0^{th} term twice in the event when $m=0$. This representation is convenient since whenever $m\geq 2$, for any $1\leq i\leq m-1$,

$$
P(\delta(y_i,y'_i)=1|F_i(m)) = \frac{(pp')+(pp')}{(p+p)\times(p'+p')} = \frac{1}{2} = P(\delta(y_i,y'_i)=0|F_i(m))
$$

meaning we have

$$
\begin{aligned}
T_{ij}(p,p') = \sum_{m=0}^{+\infty}\sum_{i=1}^{3}\mathbb{E}(e^{-\lambda\delta(y_0\neq y'_0)}|F_i(m))\mathbb{P}(F_i(m))\left(\frac{1+e^{-\lambda}}{2}\right)^{\max(m-1,0)} \\
\times \sum_{\delta m=0}^{+\infty} P(G(m,\delta m)|F_i(m))e^{-\lambda(\max(0,\delta m-1)+\delta(m>0))}
\end{aligned}
$$

Inserting the relevant event probabilities into the expression for T_{ij}, we have

$$
\begin{aligned}
T_{ij}(p,p') &= \sum_{m=0}^{+\infty} \left(\frac{1+e^{-\lambda}}{2}\right)^{\max(m-1,0)} \\
&\times \Bigg(\mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_1(m))(2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p')e^{-\lambda\delta(m>0)} \sum_{\delta m=1}^{+\infty} e^{-\lambda(\delta m-1)}(2p')^{\delta m-1} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_2(m))(2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p)e^{-\lambda\delta(m>0)} \sum_{\delta m=1}^{+\infty} e^{-\lambda(\delta m-1)}(2p)^{\delta m-1} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_3(m))(2p')^m \times (1-2p')(2p)^m(1-2p) \Bigg) \\
&= \sum_{m=0}^{+\infty} \left(\frac{1+e^{-\lambda}}{2}\right)^{\max(m-1,0)} \\
&\times \Bigg(\mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_1(m))(2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p')e^{-\lambda\delta(m>0)} \sum_{\delta m=0}^{+\infty} e^{-\lambda\delta m}(2p')^{\delta m} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_2(m))(2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p)e^{-\lambda\delta(m>0)} \sum_{\delta m=0}^{+\infty} e^{-\lambda\delta m}(2p)^{\delta m} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_3(m))(2p')^m \times (1-2p')(2p)^m(1-2p) \Bigg) \\
&= \sum_{m=0}^{+\infty} \left(\frac{1+e^{-\lambda}}{2}\right)^{\max(m-1,0)} \\
&\times \Bigg(\mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_1(m))(2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p') \times \frac{e^{-\lambda\delta(m>0)}}{1-2p'e^{-\lambda}} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_2(m))(2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p) \times \frac{e^{-\lambda\delta(m>0)}}{1-2pe^{-\lambda}} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F_3(m))(2p')^m \times (1-2p')(2p)^m(1-2p) \Bigg)
\end{aligned}
$$

Now, some simplification arise when $m \geq 1$. Indeed, in that case, $\mathbb{E}(e^{-\lambda\delta(y_0,y_0')}|F_i(m))$ is independent of i. Noting $T_{ij}^1(p,p')$ the sum of the terms for $m \geq 1$, we thus have

$$
\begin{aligned}
T_{ij}^1(p,p') = \sum_{m=1}^{+\infty} & \mathbb{E}(e^{-\lambda\delta(y_0\neq y_0')}|F(m)) \left(\frac{1+e^{-\lambda}}{2}\right)^{m-1} \\
\times & \Bigg((2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p') \times \frac{e^{-\lambda}}{1-2p'e^{-\lambda}} \\
+ & \ (2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p) \times \frac{e^{-\lambda}}{1-2pe^{-\lambda}} \\
+ & \ (2p')^m \times (1-2p')(2p)^m(1-2p) \Bigg)
\end{aligned}
$$

Noting A_{ij} the term $\mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F(m))$, which is constant for all $m \geq 1$

$$\begin{aligned}
T_{ij}^1(p,p') &= A_{ij} \sum_{m=1}^{+\infty} \left(\frac{1+e^{-\lambda}}{2}\right)^{m-1} \\
&\times \Bigg((2p)^m \times (1-2p)(2p')^{m+1} \times (1-2p') \times \frac{e^{-\lambda}}{1-2p'e^{-\lambda}} \\
&+ (2p')^m \times (1-2p')(2p)^{m+1} \times (1-2p) \times \frac{e^{-\lambda}}{1-2pe^{-\lambda}} \\
&+ (2p')^m \times (1-2p')(2p)^m(1-2p)\Bigg) \\
&= A_{ij}(1-2p)(1-2p')4pp'(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1) \sum_{m=0}^{+\infty} (4pp'(1+e^{-\lambda})/2)^m \\
&= A_{ij} \times \frac{(1-2p)(1-2p')4pp'}{1-4pp'(1+e^{-\lambda})/2} \left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1\right) \\
&= C \times A_{ij}
\end{aligned}$$

where

$$C(p,p') = \frac{(1-2p)(1-2p')4pp'}{1-4pp'(1+e^{-\lambda})/2} \left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1\right)$$

is a constant that does not depend on i, j. We compute the $m = 0$ sum, noted $T_{ij}^0(p,p')$. We have

$$\begin{aligned}
T_{ij}^0(p,p') = \Bigg(&\mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_1(0)) \times (1-2p)(2p') \times (1-2p') \times \frac{1}{1-2p'e^{-\lambda}} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_2(0)) \times (1-2p')(2p) \times (1-2p) \times \frac{1}{1-2pe^{-\lambda}} \\
&+ \mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_3(0)) \times (1-2p')(1-2p)\Bigg)
\end{aligned}$$

And we need to compute the terms $\mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_i(0))$ indivdually.

i=1, i=2 For $i = 1$, we must have $y_0 \neq y_0'$, since $y_0 = \text{STOP}$, and $\text{len}(y') > 0$. Thus, $\mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_1(0)) = e^{-\lambda}$. Similarly, $\mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_2(0)) = e^{-\lambda}$.

i=3 In that case, we must have $y_0 = y_0' = \text{STOP}$, since $\text{len}(y) = \text{len}(y') = 0$. Thus, $\mathbb{E}(e^{-\lambda\delta(y_0 \neq y_0')}|F_3(0)) = 1$.

Putting this together, we have

$$T_{ij}^0(p,p') = (1-2p)(1-2p')\left(\frac{2p'e^{-\lambda}}{1-2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1-2pe^{-\lambda}} + 1\right)$$

With that notation, we have:

$$\begin{aligned}
\text{ACMMD}^2(\mathbb{P}_|, \mathbb{Q}_|) &= \int k_{\mathcal{X}}(p,p')C(p,p')(A_{11} + A_{22} - 2A_{12})\mathrm{d}\mathbb{P}_X(p)\mathrm{d}\mathbb{P}_X(p') \\
&+ \int k_{\mathcal{X}}(p,p')(T_{11}^0(p,p') + T_{22}^0(p,p') - 2T_{12}^0(p,p'))\mathrm{d}\mathbb{P}_X(p)\mathrm{d}\mathbb{P}_X(p') \\
&= \int k_{\mathcal{X}}(p,p')C(p,p')(A_{11} + A_{22} - 2A_{12})\mathrm{d}\mathbb{P}_X(p)\mathrm{d}\mathbb{P}_X(p')
\end{aligned}$$

since T_{ij}^0 does not depend on i, j. We can narrow the variation down even further: by noting $p_{ij}^A = P(\delta(y_i \neq y_i') = 0|F(m))$ (resp $p_{ij}^B = P(\delta(y_i \neq y_i') = 0|F(0))$), since $\mathbb{E}(e^{-\lambda\epsilon}) = p(\epsilon = 0)(1-e^{-\lambda}) + e^{-\lambda}$ if ϵ is a Bernoulli random variable,

$$\text{ACMMD}^2(\mathbb{P}_|, \mathbb{Q}_|) = \int k_{\mathcal{X}}(p,p')C(p,p')(1-e^{-\lambda})(p_{11}^A + p_{22}^A - 2p_{12}^A)\mathrm{d}\mathbb{P}_X(p)\mathrm{d}\mathbb{P}_X(p')$$

We now compute the probabilities p_{ij}^A for $i, j \in \{1, 2\}$. In every case, such p_{ij}^A can be written as:

$$p_{ij}^A = \frac{P(y_0 = y_0' = A) + P(y_0 = y_0' = B)}{P(\{y_0 \in \{A, B\}\} \cap \{y_0' \in \{A, B\}\})} \quad \frac{P(y_0 = y_0' = A) + P(y_0 = y_0' = B)}{4pp'}$$

and we have

$$\begin{aligned}
p_{11}^A &= \frac{pp' + pp'}{4pp'} = \frac{1}{2} \\
p_{22}^A &= \frac{(p + \Delta p)(p' + \Delta p) + (p - \Delta p)(p' - \Delta p)}{4pp'} = \frac{2pp' + 2\Delta p^2}{4pp'} \\
p_{12}^A &= \frac{(p)(p' + \Delta p) + (p)(p' - \Delta p)}{4pp'} = \frac{1}{2} \\
\implies p_{11}^A + p_{22}^A - 2p_{12}^A &= \frac{2pp' + 2\Delta p^2}{4pp'} - \frac{1}{2} = \frac{2\Delta p^2}{4pp'}
\end{aligned}$$

Putting it together We thus have

$$\text{ACMMD}(\mathbb{P}_{|}, \mathbb{Q}_{|}) = \int C(p, p') k(p, p')(1 - e^{-\lambda}) \frac{2\Delta p^2}{4pp'} \mathrm{d}\mathbb{P}_X(p) \mathrm{d}\mathbb{P}_X(p')$$

Recalling that

$$C(p, p') = \frac{(1 - 2p)(1 - 2p')4pp'}{1 - 4pp'(1 + e^{-\lambda})/2} \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right)$$

yields the desired result. □

D.4.2. Closed-form ACMMD–Rel evaluation

Assuming the same model, it is also possible to evalute ACMMD–Rel$(\mathbb{P}_{|}, Q_{|})$ in closed form. Indeed, ACMMD–Rel becomes a special case of the ACMMD formula given above, with the conditionned variable X set to be the models $Q_{|X}$. It is thus possible to show:

Lemma D.3. *We have*

$$\text{ACMMD–Rel}^2(\mathbb{P}_{|}, Q_{|}) = C \times \Delta p^2$$

for

$$C = \iint k_{\mathcal{P}(\mathcal{Y})}(q_{|p}, q_{|p'}) 2(1 - e^{-\lambda}) \frac{(1 - 2p)(1 - 2p')}{1 - 4pp'(1 + e^{-\lambda})/2} \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right) \mathrm{d}\mathbb{P}_X(p) \mathrm{d}\mathbb{P}_X(p')$$

The above lemma leaves the choice of the kernel $k_{\mathcal{P}(\mathcal{Y})}$ open: the tractability of this expression will follow only if such kernel can be tractably computed. In the next lemma, we derive a closed form solution for $k_{\mathcal{P}(\mathcal{Y})}(q, q')$ when $k_{\mathcal{P}(\mathcal{Y})}(q, q') = e^{-\frac{\text{MMD}^2(q,q')}{2\sigma^2}}$, where the MMD is computed with an Exponentiated Hamming kernel on $\mathcal{Y}$.

Lemma D.4. *We have*

$$\text{MMD}^2(q_{|p}, q_{|p'}) = T(p, p) + T(p', p') - 2T(p, p')$$

Where

$$\begin{aligned}
T(p, p') &= C(p, p')A(p, p') + T^0(p, p') \\
C(p, p') &= \frac{(1 - 2p)(1 - 2p')4pp'}{1 - 4pp'(1 + e^{-\lambda})/2} \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right) \\
A(p, p') &= \frac{2pp' + 2\Delta p^2}{4pp'} \times (1 - e^{-\lambda}) + e^{-\lambda} \\
T^0(p, p') &= (1 - 2p)(1 - 2p') \left(\frac{2p'e^{-\lambda}}{1 - 2p'e^{-\lambda}} + \frac{2pe^{-\lambda}}{1 - 2pe^{-\lambda}} + 1 \right)
\end{aligned}$$

Combining the two lemmas allows us to obtain a computable expression for ACMMD–Rel$(\mathbb{P}_{|}, Q_{|})$.

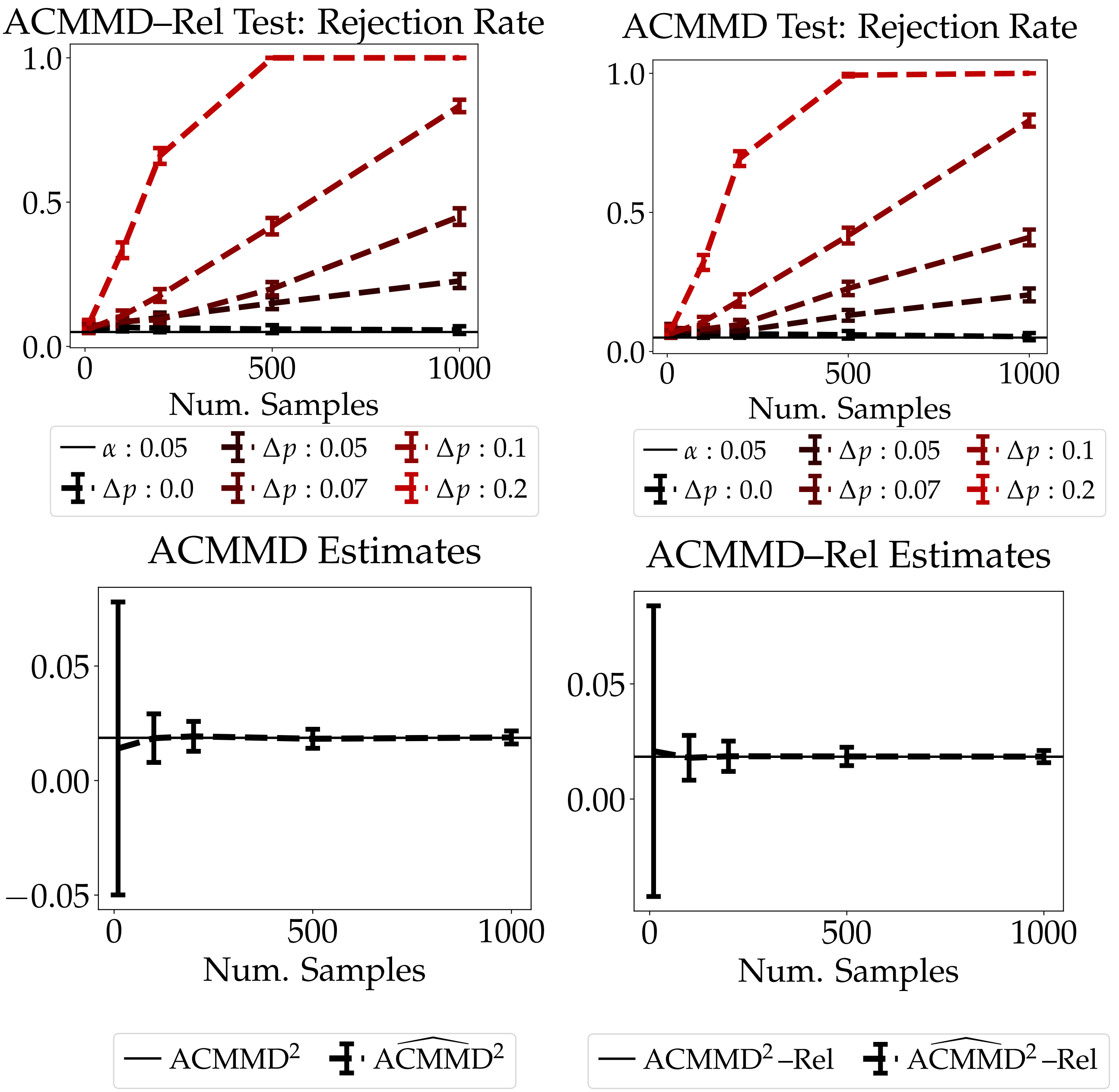

Figure 6. Top left: Rejection Rate of the ACMMD test as a function of the dataset size, for different values of Δp. Top right: Rejection Rate of the ACMMD test as a function of Δp, for different dataset sizes. Bottom left: Estimated ACMMD as a function of the dataset size. Bottom right: Estimated ACMMD–Rel as a function of Δp. To compute these estimates, we use dataset sizes of $\{10, 100, 200, 500, 1000\}$, used $m = 5$ atoms for the prior on p between $p_1 = 0.3$, $p_2 = 0.45$, used $\lambda = 1$, $\Delta p = 0.25$, and average over 300 runs. In addition, we plot the true value $\mathrm{ACMMD}(\mathbb{P}_{|}, Q_{|})$ using the closed-form expression derived above.

E. Additional Experiments

E.1. Additional Experiments for the semi-synthetic ProteinMPNN data

In addition to the figures of Section 6.2.1, which use $T = 0.1$ to plot the estimates and rejection rates of ACMMD and ACMMD–Rel on the ProteinMPNN synthetic data, we provide here the same plots for $T = 1.0$ the value used to train ProteinMPNN. We notice that detecting a given change in temperature is sligtly simpler for $T = 1.0$ than for $T = 0.1$.

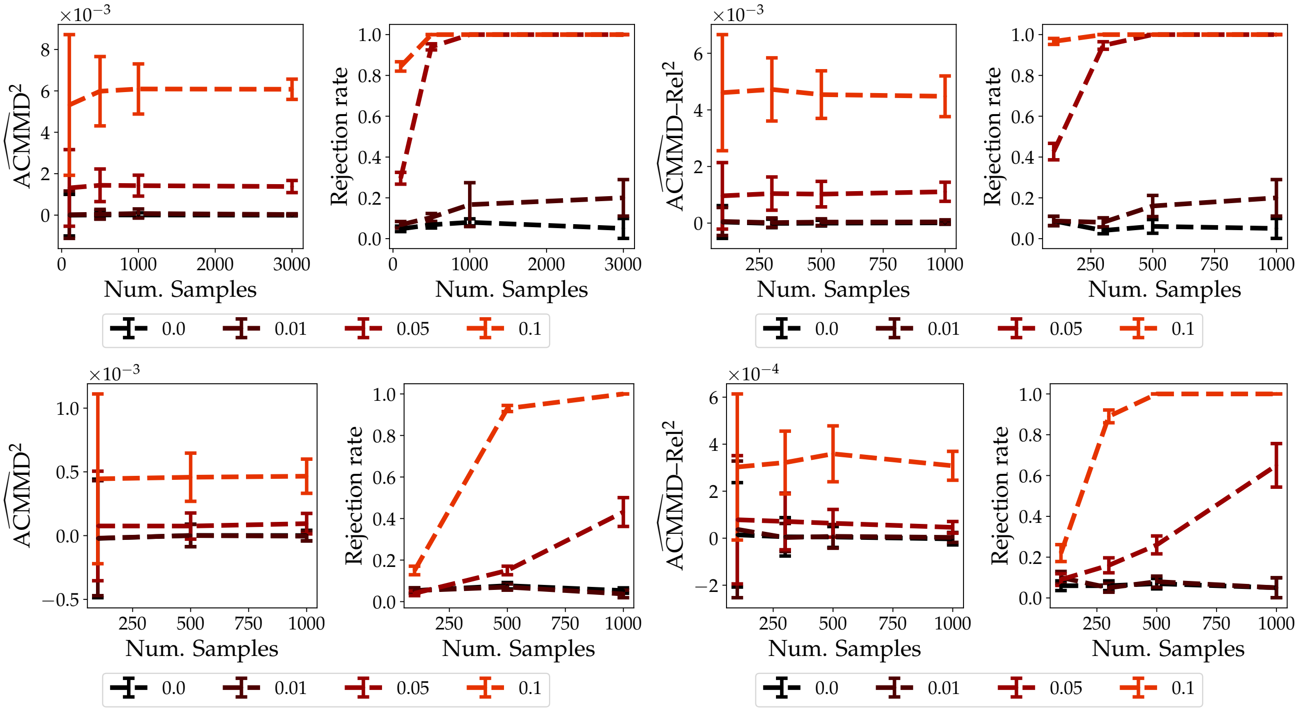

Figure 7. ACMMD and ACMMD–Rel estimates and rejection rate in the semi-synthetic setting of Section 6.2.1. The different lines indicate a different temperature shift between the two MPNN models. Top panel shows uses a base temperature of $T = 1.0$, while the bottom panel uses $T = 0.1$.

E.2. Additional Experiments for the structural superfamily evalution

We include in Figure 8 the values of $\widehat{\mathrm{ACMMD}}\,–\mathrm{Rel}^2$ for different superfamilies (which was not included in Section 6.2.2), and compare it with the values of $\widehat{\mathrm{ACMMD}}^2$. In line with the hyperparameter tuning results of Section 6.2.2, we notice that high temperature are highly detrimental from a reliability perspective. Intuitively, increasing the temperature of ProteinMPNN makes the model "underconfident". Since a reliable model is neither over– nor underconfident, this decrease of confidence is penalized by ACMMD–Rel. This also shows that increasing the temperature of a model does not make the model fallbak to its prior distribution (otherwise the model would be more reliable). Instead, it just increases the uncertainty of the model in a detrimental fashion.

F. Known Kernels for protein sequences and structures

In the context of inverse folding, computing the ACMMD requires a kernel on sequences $k_\mathcal{Y}$ and a kernel on protein structures. This section contains a brief overview of non neural-network based, known kernels for protein sequences and structures. The main desiderata to achieve when choosing kernels for computing goodness-of-fit criterion is to find kernels that are able to detect (up to statistical noise) any deviation from a perfect fit between the model and the data. Such kernels are referred to as *universal*.

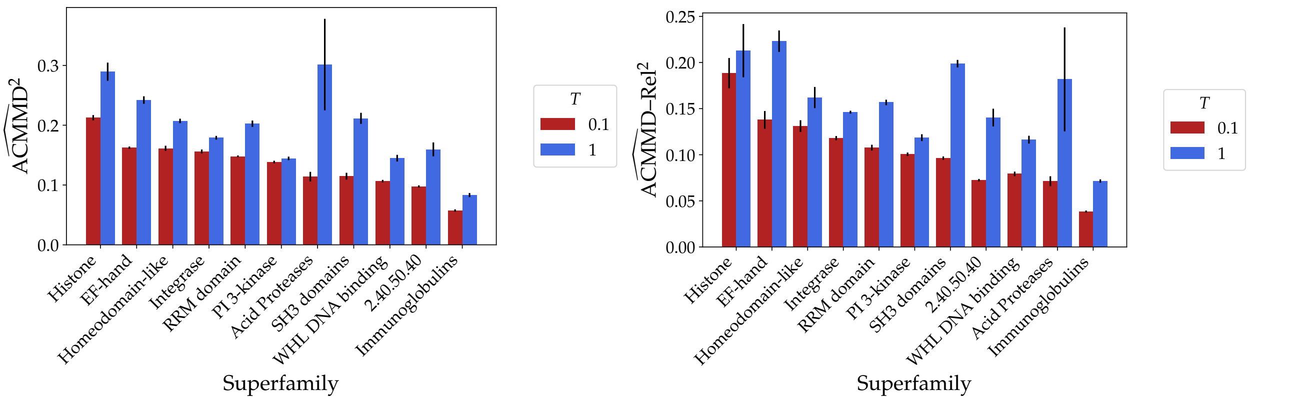

Figure 8. Value of $\widehat{\mathrm{ACMMD}}^2$ (left) and $\widehat{\mathrm{ACMMD}}\,{-}\mathrm{Rel}^{\,2}$ (right) between ProteinMPNN and the CATH S60 reference dataset on a subset of 10 superfamilies for two different temperatures $T = 1.0$ and $T = 0.1$.

The protein formalism The most general formalism for the space protein structures is the set of equivalence classes of graphs where the equivalence relationship is defined to the existence of a graph isomorphism. The need for equivalence classes stems from the fact that different labelling policies exist for a given protein, meaning that a single protein can be associated to multiple graphs. However, this labelling will in practice not be completely arbitrary: first, the set of candidate labelling can be restricted to the ones consistent with covalent bounds. But in the inverse folding problem, the setting is even simpler: the protein structure is restricted to its backbone, which is sequential by nature. This limits the set of covalent-bound consistent labelling policies to two (the forward and the backward one), and my vague understanding is that there is a terminal atom in protein, which suggests the existence of a canonical direction: thus, only one labelling policy remain, and protein structures can thus be associated to the set of atom locations $\bigcup_{i=1}^{+\infty} \mathbb{R}^i$. This set differs from the set of protein sequences $\bigcup_{i=1}^{+\infty} \mathcal{A}$ in that the “alphabet” is the real line instead of a finite set of symbols. The restriction from the space of graphs to the space of variable-length sequences since there it is known that no graph kernels commonly in use are even characteristic (Kriege et al., 2020). The space $\bigcup_{i=1}^{+\infty} \mathcal{X}$ (for arbitrary $\mathcal{X}$ have been investigated by the time series community), which have developed a set of kernels to carry out data analysis on it. We provide some background on such kernels below.

Background: alignment kernels for real-valued sequences of arbitrary length Alignment kernels (Cuturi et al., 2007; Cuturi & Blondel, 2017; Saigo et al., 2004; Vert et al., 2004) refer to a diverse set of variety of kernels constitute a family of kernels on $\bigcup_{i=1}^{+\infty} \mathcal{X}^i$ that are computed based on aggregating the similarities between all possible “alignment candidates” between two inputs x_1 and x_2. There are two main subfamilies of alignment kernels, which both use slightly different alignment definitions: local alignment kernels, and global alignment kernels.

Local alignment kernels Local alignment kernels (Saigo et al., 2004; Vert et al., 2004) are kernels of the form

$$k_{\mathrm{LA}}(x, y) = \sum_{\pi \in \Pi(x,y)} \exp(\beta s(x, y, \pi)) \tag{15}$$

Where

$$s(x, y, \pi) = \sum_{i=1}^{|\pi|} s(x_1^{(\pi_1(i))}, x_2^{(\pi_2(i))}) + \sum_{i=1}^{|\pi|-1} g(\pi_1(i+1) - \pi_1(i)) + g(\pi_2(i+1) - \pi_2(i))$$

and $\Pi(x, y)$ is the set of all possible *alignments* of x and y, e.g. the set of all 2-tuple of p-long sequences

$$\pi \coloneqq ((\pi_1(1), \ldots, \pi_1(p)), (\pi_2(1), \ldots, \pi_2(p))$$

where

$$1 \leq \pi_1(1) < \pi_1(2) \cdots < \pi_2(p) \leq n$$
$$1 \leq \pi_2(1) < \pi_1(2) \cdots < \pi_2(p) \leq m$$

Importantly, local alignment kernels involve a gap function, and thus give a specific status to insertions and deletions, unlike global alignment kernels, as we will see below. The local alignment kernel can be seen as computing the (soft) minimum of a discrepancy within the set of all possible alignments. The use of a soft minimum (and not a hard one) is crucial to ensure positive definiteness. Local alignment kernels seem to have been designed initially for finite alphabets target. When $g = 0$, the necessary and sufficient condition on s to ensure that k_{LA} is a positive definite is for s to be a conditionally positive definite kernel [4]. This is in particular verified if $(s(x^i, y^i))_{1 \leq i,j \leq |\mathcal{A}|}$ is positive definite. I need further reading to investigate whether the case of infinite $\mathcal{X}$ was studied.

Global alignment kernels Global alignment kernels (Cuturi et al., 2007; Cuturi & Blondel, 2017) also perform a softmin over alignment, but do not incorporate gaps in their score, and use a slightly different notion of alignment, namely:

$$\pi \coloneqq ((\pi_1(1), \pi_2(1), \ldots, (\pi_1(p), \pi_2(p))$$

where now, the constraints on π_1 and π_2 are

$$\begin{aligned} 1 = \pi_1(1) < \pi_1(2) \cdots < \pi_2(p) = n \\ 1 = \pi_2(1) < \pi_1(2) \cdots < \pi_2(p) = m \\ \pi_1(i+1) \leq \pi_1(i) + 1 \quad \text{unitary increments} \\ (\pi_1(i+1) - \pi_1(i)) + (\pi_2(i+1) - \pi_2(i)) \geq 1 \quad \text{no repetitions} \end{aligned}$$

Unlike the previous alignment definition, this one explicitly maps each item in each sequence with another item in the other sequence, and does not try to account for potential gaps. Let us call $\mathcal{A}$ the set of all alignment. The final definition for a global alignment kernel is then:

$$k_{\text{GA}}(x, y) = \sum_{\pi \in \mathcal{A}(x,y)} \exp(\sum_{i=1}^{\pi} s(x_1^{(\pi_1(i))}, x_2^{(\pi_2(i))}) \tag{16}$$

As stated in Cuturi et al. (2007, Theorem 1), k_{GA} will be positive definite if $k(x, y) \coloneqq \exp(s(x, y))$ is a positive definite kernel such that $\frac{k}{(1-k)}$ is positive definite.

[4] A kernel is c.p.d if $\sum_{i,j=1}^{n} c_i c_j s(x^{(i)}, x^{(j)}) \geq 0 \forall c_1, \ldots, c_n, c_1 + \cdots + c_n = 0$.