Evaluation Methods for Generative Models

Anja Kroon, supervised by Professor Mark Coates and Florence Regol
Honours EE Thesis 06 | McGill University, Department of Electrical and Computer Engineering

McGill Networks Research Laboratory

Research Objective

- Generative models (e.g. **ChatGPT, DALLE-2**) produce samples from a learned distribution
- Evaluation metrics for generative models compare p, the ground truth distribution, to q, the learned distribution
- Current metrics are approximate, insensitive to nuances in failures, and task-specific. **Better metrics** → **better models**
- Propose new evaluation metric with **statistical guarantees** in high-dim. probability space settings $\Omega \geq 10^9$ [1] which is **scalable**, **robust** to mismatches in p, q and provides **interpretable results**
- Synthetic experiments are conducted to verify claims. Current experiments in **protein sequence modeling**

Introduction

Problem Statement

- Given p evaluate **which generative model** q_1, q_2 **is closer to** p
- Closeness classically determined by total variation error $d_{TV} = \frac{1}{2}||\mathbf{p}-\mathbf{q}||$ but estimating d_{TV} scales with $|\Omega|$. (Intractable for many machine learning applications) [2]

Existing Evaluation Metrics

- **Negative Log-Likelihood:** not always accessible, doesn't guarantee good sample generation [3]
- **Task-oriented:** good sample evaluation but not general [4]
- **Coverage-based:** compares distribution coverage, p, q, does not consider sample complexity or stat. significance [5] [6]

Proposed Method

- **Partition** Ω into smaller spaces called **bins** of size k, $\mathcal{B}^k$, and perform statistical tests on the smaller probability distributions, $p^{\mathcal{B}^k}$, on these smaller spaces $\mathcal{B}^k$

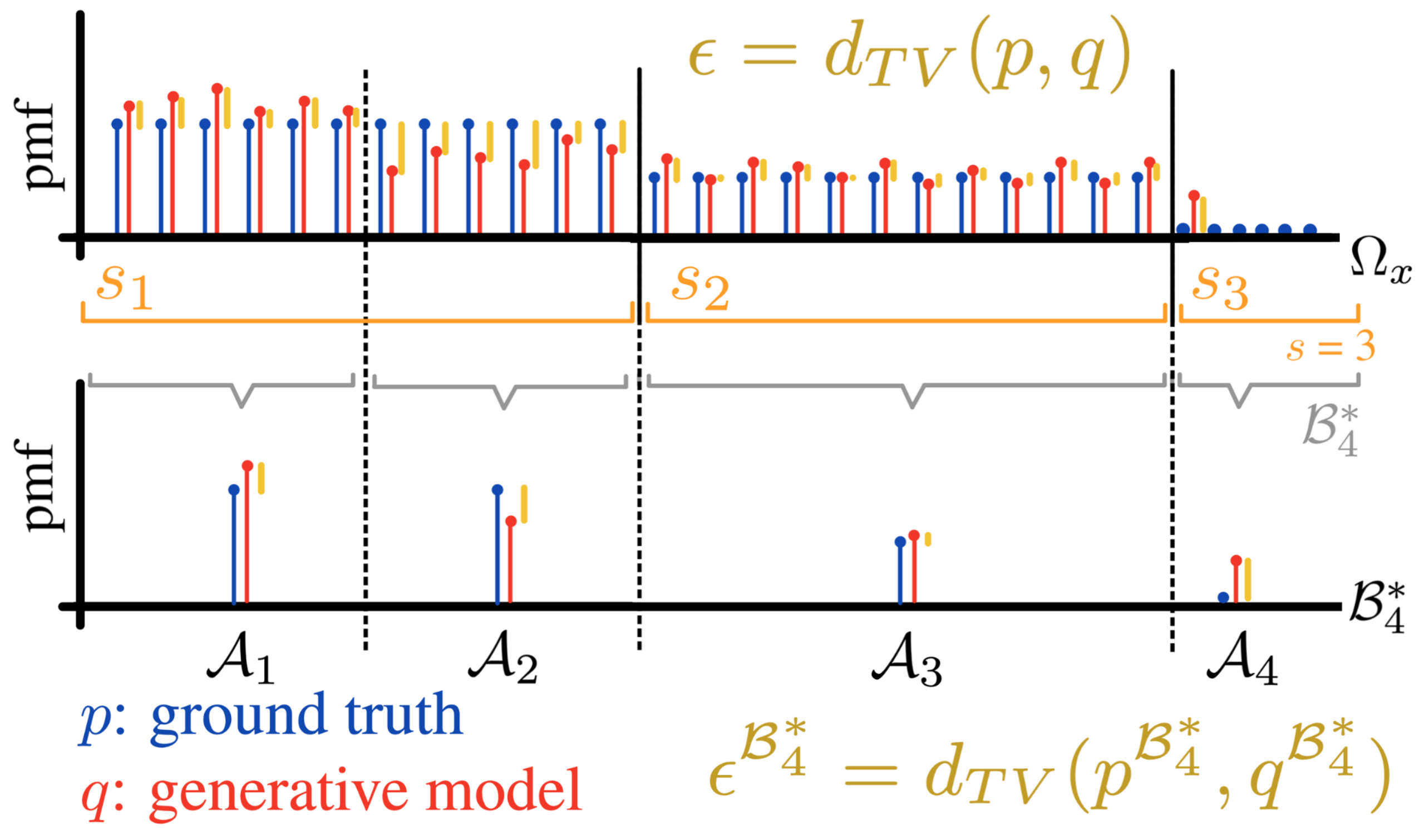

Figure 1: Overview of Proposed Procedure [1]

Methodology

Preliminaries: Ω is partitioned into bins of size k **inducing new probability distributions** on the smaller partitions

$$\rho(\Omega) = \{\{\mathcal{A}_1, \dots\} | \cup_i \mathcal{A}_i = \Omega, \mathcal{A}_j \cap \mathcal{A}_i = \emptyset \forall i \neq j\}$$

$$\rho^k(\Omega) = \{\mathcal{B} \in \rho(\Omega), |\mathcal{B}| = k\}$$

$$p_{\mathcal{A}_i}^{\mathcal{B}^k} \triangleq \sum_{x \in \mathcal{A}_i} p_x$$

By the triangle inequality, d_{TV} is constrained and increases with the granularity $|\mathcal{B}|$.

$$\mathcal{B} \in \rho(\Omega) \implies d_{TV}^{(p^{\mathcal{B}}, q^{\mathcal{B}})} \leq d_{TV}^{(p,q)}$$

$$d_{TV}(p^{\mathcal{B}^{i-1}}, q^{\mathcal{B}^{i-1}}) \leq d_{TV}(p^{\mathcal{B}^i}, q^{\mathcal{B}^i}) \cdots \leq d_{TV}(p, q)$$

Binning: Bin the space, **identifying sets where the masses associated with any two elements in the set do not differ by much**.

Setting an Error Tolerance: Error tolerance ϵ_{test} is a function of the cardinality of the probability space $|\mathcal{B}|$, number of samples m, and probability significance δ. Given a set of m samples from q ($\{\tilde{x}_i\}_{i=1}^m$ $\tilde{x}_i \sim q$), the empirical total variation estimator B^m is estimated as follows:

$$d_{TV}(p, \tilde{q}) \triangleq B^m = \frac{1}{2}\sum_{x \in \Omega} |p_x - \tilde{q}_x|, \text{where } \tilde{q}_x = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}[\tilde{x}_i = x].$$

Provided that

$$\epsilon_{test} \geq \max(\sqrt{\frac{|\mathcal{B}|}{m}}, \sqrt{\frac{2\ln(2/\delta)}{m}}),$$

we can be at least $1-\delta$ confident that the true total variation $d_{TV}(p, q)$ is within the interval $[B^m - \epsilon_{test}, B^m + \epsilon_{test}]$

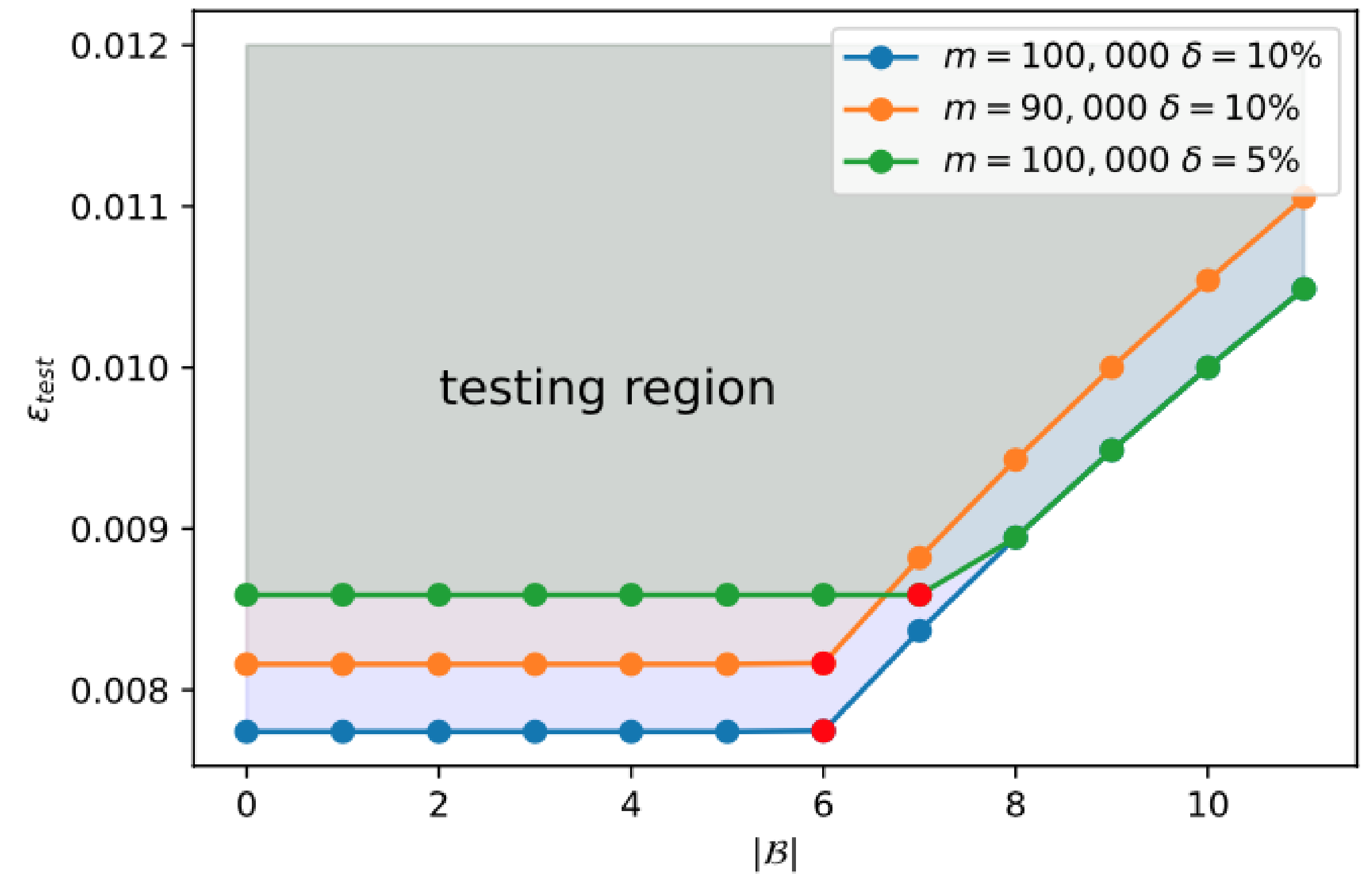

Figure 2: Error threshold that we can obtain over varying $k = |\mathcal{B}|$, δ and m.

Evaluation Procedure

- If we have e.g. $d_{TV}(p^{\mathcal{B}^i}, q_1^{\mathcal{B}^i}) \leq \epsilon_{thresh}$, but model q_2 is not: $d_{TV}(p^{\mathcal{B}^i}, q_2^{\mathcal{B}^i}) \geq \epsilon_{thresh}$, we can say q_1 is better than q_2 in $\mathcal{B}^i$.

Results

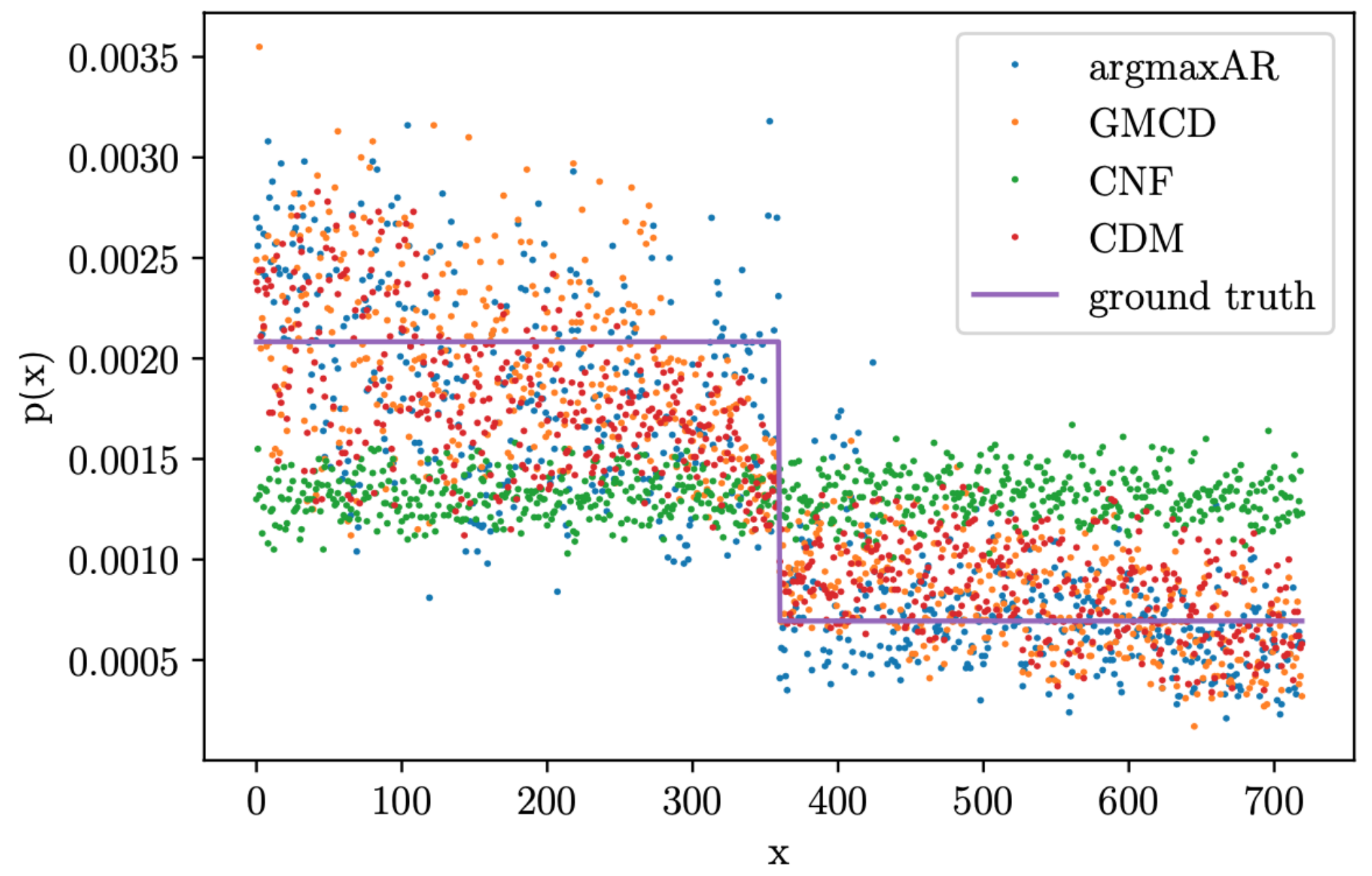

Figure 3: Empirical pmf q ($m = 100,000$) of generative models on Ω^+ with sorted ground truth. Only 700 samples are non-zero. 10k samples are used for model training. 10k samples are used for model testing.

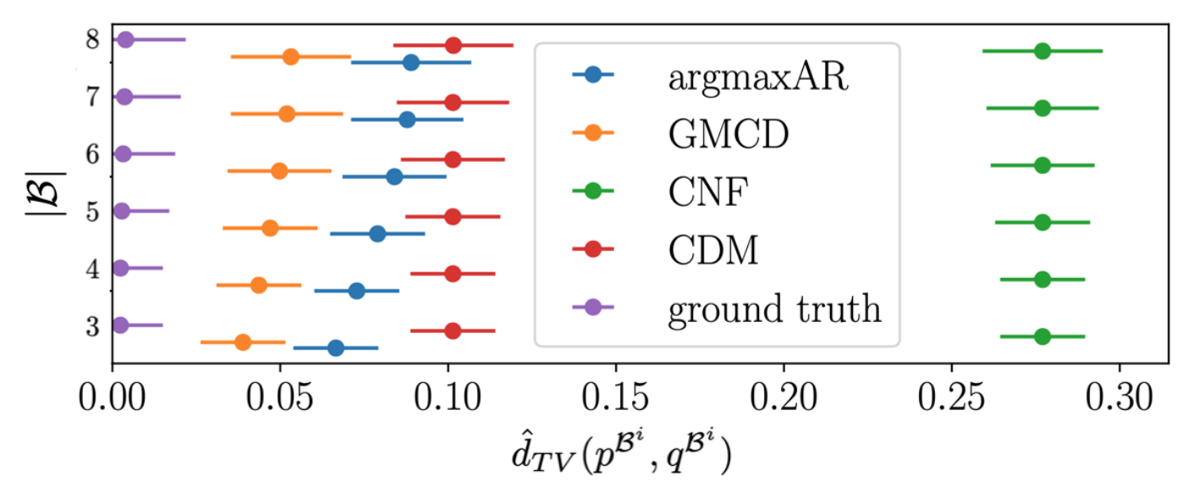

Figure 4: $d_{TV}(p^{\mathcal{B}^i}, q^{\mathcal{B}^i})$ metric reported for the generative models. Left leaning means better performance.

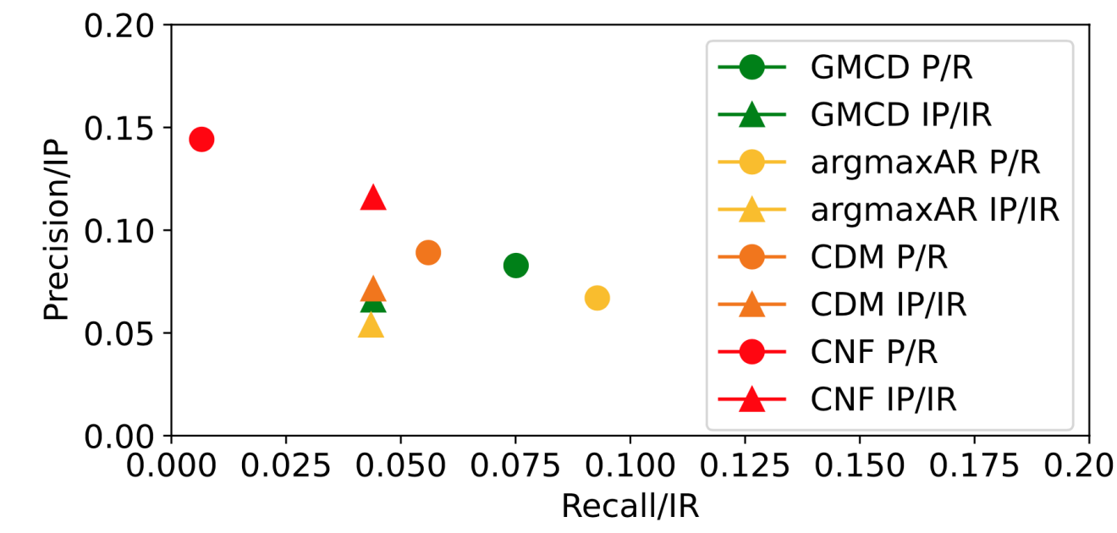

Figure 5: Coverage-based metrics, precision, recall, IP_α, IR_β for generative models [5] [6]

Discussion

- Figure 4 provides interpretable understanding → models closer to the ground truth are better models
- Figure 5 provides the less interpretable existing coverage metrics → no clear better model

Conclusion

- The proposal provides **interpretable results** with **statistical guarantees**, is **scalable to high dimensions** and offers a comparative **performance evaluation**.
- **Current and future work** involves applying the proposed metric to the real-task of **protein sequence modelling** on the order of $|\Omega| = 21^{100}$ (21 amino acids and sequence length 100).

References

[1] F. Regol, A. Kroon, and M. Coates, "Evaluation of categorical generative models – bridging the gap between real and synthetic data," in *in Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[2] I. Diakonikolas, T. Gouleakis, D. M. Kane, J. Peebles, and E. Price, "Optimal testing of discrete distributions with high probability," in *Proc. ACM SIGACT Symposium on Theory of Computing*, 2021, p. 542–555.

[3] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2016.

[4] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Info. Process. Syst. (NeurIPS)*, 2016.

[5] M. F. Naeem, S. J. Oh, Y. Uh, Y. Choi, and J. Yoo, "Reliable fidelity and diversity metrics for generative models," in *Proc. Int. Conf. Machine Learning ICML*, 2020.

[6] A. Alaa, B. Van Breugel, E. S. Saveliev, and M. van der Schaar, "How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models," in *Proc. Int. Conf. Machine Learning ICML*, 2022.