SIMULTANEOUS FINITE-SAMPLE CERTIFICATION OF SYNTHETIC DATA GENERATORS UNDER FIDELITY CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data generators (SDGs) are increasingly used in high-stakes settings, yet practitioners often lack a principled way to determine whether the generated data adequately represent the real data distribution for a specific dataset. While theoretical guarantees for SDGs do exist, they are often asymptotic and therefore do not directly provide finite-sample guarantees for the trained generator and dataset at hand. As a result, it can be difficult to determine which among several candidate generators satisfy a required standard of data fidelity. We address this problem by formulating the certification of SDGs as a multiple hypothesis testing problem over a collection of candidate generators. For each generator, we test whether a population-level fidelity criterion, defined through a user-specified discrepancy measure between the real and synthetic distributions, falls below a prescribed tolerance level. By constructing valid tests from finite-sample concentration bounds and applying a family-wise error rate controlling procedure, our framework guarantees, with probability at least $(1 - \alpha)$, that all selected generators satisfy the specified fidelity criterion. This provides dataset-specific, finite-sample certification of synthetic data fidelity and allows practitioners to screen multiple candidate generators and retain a statistically supported subset that meets a desired fidelity standard before considering other downstream objectives. The framework is agnostic to the generator architecture and applies to a broad class of discrepancy measures whose estimators admit suitable concentration inequalities, making it useful for applications where reliable finite-sample assessment of synthetic data is important.

1 INTRODUCTION

Synthetic data generators (SDGs) are increasingly deployed in settings where the quality of the generated data has meaningful consequences, including healthcare and finance (Altman et al. (2024); Qian et al. (2024)). The appeal of synthetic data lies in its potential to preserve privacy, augment limited datasets, and enable data sharing in settings where access to real data may be restricted (Jordon et al. (2022); Du & Li (2024); Arjovsky et al. (2017); Müller (1997); Tian & Shen (2026); El Emam et al. (2022)). However, despite rapid advances in generative modeling, there remains no principled way to certify whether the outputs of a synthetic data generator can be trusted for a particular dataset.

Modern SDGs include generative adversarial networks (GANs) (Goodfellow et al. (2020)), variational autoencoders (VAEs) (Kingma & Welling (2014)), diffusion models (Ho et al. (2020); Sohl-Dickstein et al. (2015)), and related generative architectures. Existing approaches to improving synthetic data quality primarily focus on designing better model architectures, training objectives, or optimization procedures. For example, several methods incorporate discrepancy measures such as Wasserstein distances or maximum mean discrepancy (MMD) directly into the training objective (Arjovsky et al. (2017); Zhao et al. (2017); Dziugaite et al. (2015)), while others rely on likelihood-based objectives (Song et al. (2021)) or methods based on restricted Boltzmann machines and deep Boltzmann networks (Song et al. (2021); Salakhutdinov & Hinton (2009)). Theoretical analyses of such methods are typically asymptotic, whereas empirical validation is commonly performed on benchmark datasets. While these approaches have led to substantial progress in generative modeling,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

they do not directly address a fundamental practical question: *given a specific dataset, which synthetic data generators can be statistically certified to satisfy a desired fidelity criterion?* Asymptotic convergence guarantees do not provide direct assurances in the finite-sample regimes where SDGs are trained and deployed (Hudovernik et al. (2024)), and empirical success on benchmark datasets does not imply reliable performance on a new dataset of interest. Moreover, no single SDG can be expected to perform uniformly well across all datasets and fidelity criteria. Consequently, users of synthetic data currently lack rigorous, dataset-specific guarantees regarding whether a generator faithfully represents the data at hand.

In this work, we adopt a fundamentally different perspective on synthetic data validation. Rather than attempting to design a universally optimal SDG, we formulate the problem of *certifying* SDGs as a multiple hypothesis testing problem. Given a collection of candidate generators trained on a dataset, we test whether each generator satisfies a user-specified fidelity criterion defined through a population-level discrepancy measure between the real and synthetic distributions. Under this formulation, each SDG is associated with a hypothesis corresponding to whether its discrepancy measure lies below a prescribed tolerance threshold. Our framework leverages FWER controlling procedures to select a subset of generators that satisfy the desired fidelity criterion with high probability. Specifically, if $\widehat{\mathcal{I}}$ denotes the set of generators selected by the procedure and $\mathcal{I}_a$ denotes the set of generators that truly satisfy the fidelity criterion, our framework guarantees

$$\mathbb{P}\Big(\widehat{\mathcal{I}} \subseteq \mathcal{I}_a\Big) \geq 1 - \alpha.$$

That is, with probability at least $1 - \alpha$, every selected SDG is genuinely valid according to the specified fidelity criterion. Importantly, the framework is finite-sample and distribution-free, making it particularly suitable for synthetic data settings where the underlying data distribution is unknown and only finite observations are available. A key advantage of our approach is its flexibility. The framework does not depend on a particular SDG architecture or training paradigm, and it applies broadly to discrepancy measures whose estimators satisfy concentration inequalities. In particular, while our experiments focus on fidelity measures based on maximum mean discrepancy (MMD), the framework naturally extends to other statistical measures satisfying analogous probabilistic bounds.

More broadly, our work shifts the focus of synthetic data evaluation from improving generators in expectation to providing statistically defensible, dataset-specific certification guarantees. This perspective is particularly relevant in high-stakes settings where users require rigorous assurances regarding the reliability of synthetic data before deployment.

Our Contributions:

1. We formulate synthetic data certification as a multiple hypothesis testing problem over a collection of candidate SDGs.
2. We provide finite-sample, distribution-free guarantees that all selected generators satisfy a user-specified fidelity criterion with probability at least $1 - \alpha$.
3. We show that the framework can certify multiple SDGs simultaneously, allowing practitioners to retain flexibility across downstream objectives while maintaining statistical validity.

Our main insight lies in recognizing that certain architectures and training procedures may be optimal for different datasets, and thus guarantees should be data-specific. We provide a framework that gives finite sample and distribution free validity which is crucial for synthetic data since we do not know the underlying distribution and train on a finite dataset. Moreover, we provide a means to certify a subset of SDG that are valid, even for empirically supported methods that may be too complicated to yield to theoretical guarantees. The subset of certified SDGs allow users to work with multiple synthetic datasets that all meet the certification threshold.

2 Related work

In this section, three main topics are covered, the first is the use of the theory of multiple hypothesis testing in the broader context of machine learning, followed by a discussion of how measures of data quality have been used as part of the training procedure in SDGs to address the problem of data

quality, then we discuss the intersection of the hypothesis testing and generative models, which is our focus, and establish how our work is different.

Multiple Hypothesis Testing in Machine Learning. The theory of multiple hypothesis testing, which has its roots in biostatistics to certify that the claims made by scientists and drug companies are true with high probability, has been used in machine learning in various different ways. It has been used in variable selection in the context of regression using FWER (Janson & Su (2016)) and also false discover rate (Bunea et al. (2006); Barber & Candès (2015)). More recently, for settings with features X associated with responses Y, and a trained predictor function, it has been used to provide calibrations for the loss function for the generalized conformal prediction task of controlling non-monotonic risks associated with predictions made by the predictor function (Angelopoulos et al. (2025)). Multiple hypothesis testing has also been explored in the context of data mining (Hanhijärvi (2011)).

Use of Distributional Discrepancy measures in SDG training. Using estimators of distributional discrepancy measures as part of the training process have been widely explored for SDGs. In particular, methods have been proposed that use the discriminator of the GAN to play the role of maximizing an estimate of the discrepancy measure such as f-divergences and maximum mean discrepancy (MMD) amongst others between the real and the synthetic data, so that the generator of the GAN produces data with the aim of minimizing such a measure (Zhao et al. (2017); Dziugaite et al. (2015)). Moreover, theoretical work has studied the gradients of these objectives, and their implications for training stability and model performance (Arjovsky et al. (2017)). Arjovsky et al. (2017) provide useful theoretical discussion of various measures for distributional discrepancy and a means of using the Wasserstein loss function in the training of generative models. Alternative approaches for improving synthetic data quality have also been proposed such as using moment matching (Hanhijärvi (2011)), and where an optimal transport map is used as the generative model (Rout et al. (2022)).

Hypothesis Testing in SDG. Finally, while there has been research into usage of hypothesis testing has been used in machine learning in general and using measures of distributional discrepancy as part of the training procedure the intersection between synthetic data and inference on population level measures, which is our focus, has not been explored much. One paper that focuses on this intersection is the Bounliphone et al. (2016) paper, which focuses on using a test for similarity based on the MMD measure that assesses whether one model is more similar than another model to the real data. Their focus is different from ours in that we are interested in selecting a subset of models that meet a criteria of data quality, rather than providing a statistical claim for which model performs the best.

3 BACKGROUND

The necessary prerequisite to understanding the framework will be introduced. We formulate the goal of FWER control and present the algorithm for sequential graphical testing, which is the multiple hypothesis testing procedure that we will use. Furthermore, we introduce the measure of fidelity that will be used.

3.1 FAMILYWISE ERROR RATE CONTROL

In multiple hypothesis testing, we consider a collection of hypotheses $\{H_1, \ldots, H_N\}$ that are tested jointly using the same testing data. A false rejection is when we reject a hypothesis when the null is actually true The goal of FWER control is to ensure that the probability of rejecting at least one hypothesis where the null is indeed true is controlled by α, the significance level.

Mathematically, let M denote the number of false rejections, i.e., $M := \left|\widehat{\mathcal{I}} \cap \mathcal{I}_0\right|$, where $\mathcal{I}_0$ is the set of indices of the true null hypotheses and $\mathcal{I}_a$ is the set of indices for the rejected hypotheses (Hochberg & Tamhane (1987)). A multiple testing procedure is said to control the FWER at level α if

$$\Pr(M > 0) \leq \alpha.$$

Algorithm 1 Bretz et al. (2009) Graph-Based Multiple Testing Procedure

1: **Initialize:** Graph G, index set $\mathcal{I} = \{1, \ldots, N\}$.
2: Let $j = \arg\min_{i \in I} \frac{p_i}{\alpha_i}$.
3: If $p_j \leq \alpha_j$, reject H_j; otherwise **Stop**.
4: Update:
- $\mathcal{I} \leftarrow \mathcal{I} \setminus \{j\}$.
- For each $\ell \in \mathcal{I}$:
$$\alpha_\ell \leftarrow \alpha_\ell + \alpha_j g_{j\ell}.$$
- For each $\ell, k \in \mathcal{I}$, $\ell \neq k$:
$$g_{\ell k} \leftarrow \frac{g_{\ell k} + g_{\ell j} g_{jk}}{1 - g_{\ell j} g_{j\ell}}.$$
- For each $\ell \in \mathcal{I}$:
$$g_{\ell j} \leftarrow 0, \quad g_{j\ell} \leftarrow 0.$$

5: If $|\mathcal{I}| > 1$, return to step 1, otherwise **Stop**.
6: **Return:** $\widehat{\mathcal{I}} = \{1, \ldots, N\} \setminus \mathcal{I}$.

A basic example is the Bonferroni procedure, which rejects hypothesis i only if its p-value is smaller than α/N. While valid, this approach is often conservative; more powerful procedures exist such as the graph-based method of Bretz et al. (2009) shown in Algorithm 1. The procedure begins with a graph $G = (g_{ij})$, where each node corresponds to a hypothesis and each edge weight $g_{ij} \in [0, 1]$ specifies the proportion of significance level that is transferred from hypothesis i to hypothesis j after H_i is rejected. Each hypothesis is also assigned a local significance level α_i, with $\sum_{i=1}^{N} \alpha_i = \alpha$, where α is the desired global FWER level.

3.2 Maximum Mean Discrepancy as a Fidelity Measure

Let $(\mathcal{X}, d)$ be a metric space, and let $X \sim p$ and $\tilde{X} \sim q$ be random variables taking values in $\mathcal{X}$, where p and q are Borel probability measures on $\mathcal{X}$. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$. The MMD is defined as

$$\mathrm{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left(\mathbb{E}_X[f(X)] - \mathbb{E}_{\tilde{X}}[f(\tilde{X})] \right).$$

In the context of synthetic data, X denotes the real data and $\tilde{X}$ denotes the synthetic data. We let the function class be the reproducing kernel Hilbert space (RKHS) of real valued functions, and let $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denote the inner product on $\mathcal{H}$. By the reproducing property, there exists a reproducing kernel $k \in \mathcal{H}$ such that every $f \in \mathcal{H}$ can be expressed as

$$f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = \sum_i \alpha_i \, k(x, x_i).$$

We say the function class induced by a kernel k is the class of functions in the closure of the span of the set $\{k(\cdot, x)\}$. The function class induced by the kernel k is a RKHS. For the extent to which we will use the MMD, if we use the Gaussian kernel, we know $\mathrm{MMD}(\mathcal{F}, p, q) = 0 \iff p = q$(*Grettonet al.* (2012)).

As presented in Gretton et al. (2012), let $Z := (z_1, \ldots, z_c)$ be i.i.d. random variables of size c, where $z := (x, y) \sim p \times q$. An unbiased estimate of MMD^2 is

$$\mathrm{MMD}_u^2[\mathcal{F}, X, \tilde{X}] = \frac{1}{c(c-1)} \sum_{i \neq j}^{c} h(z_i, z_j),$$

which is a one-sample U-statistic with $h(z_i, z_j) := k(x_i, x_j) + k(\tilde{x}_i, \tilde{x}_j) - k(x_i, \tilde{x}_j) - k(x_j, \tilde{x}_i)$..

4 PROBLEM SETUP

This section contains the mathematical notation and formulation of the problem. First, we will introduce the problem setting generally, then adapt it to the fidelity by using a measure of discrepancy. We also explain which parameters within our setup will be chosen by the user of the synthetic data.

4.1 GENERAL SETTING

Suppose we observe i.i.d. data $\{X_j\}_{j=1}^{n_{\text{tot}}} \subset \mathbb{R}^k$ drawn from a distribution p, which is not known. We partition the data into a training set $\mathcal{D}_{\text{train}} = \{X_j\}_{j=1}^{n}$ and a holdout set $\mathcal{D}_{\text{H}} = \{X_j\}_{j=n+1}^{n+c}$, where $n_{\text{tot}} = n + c$. Let $\mathcal{I} = \{1, \ldots, N\}$ denote the index set, and let $\mathcal{G} = \{G_i \mid i \in \mathcal{I}\}$ be the collection of SDGs that are being considered. We use $\hat{G}_i$ to denote the generator obtained by training G_i on $\mathcal{D}_{\text{train}}$, and let q_i denote the distribution of samples from the trained generator $\hat{G}_i$.

Let θ_i denote the population-level measure associated with trained generator $\hat{G}_i$ that we aim to make a claim about. To carry out the procedure, we need an estimator, so we let $\hat{\theta}_i$ denote an estimator of θ_i computed using the holdout data and samples generated from $\hat{G}_i$. Let Θ_a denote the ideal set for the measure to be in, defined by the user or a regulatory body. Define

$$\mathcal{I}_a = \{i \in \mathcal{I} \mid \theta_i \in \Theta_a\},$$

which is the set of SDGs that actually satisfy the desired criterion for the dataset. Furthermore, let $\hat{\mathcal{I}}$ denote the subset of indices corresponding to the generators selected by the procedure. We desire

$$\mathbb{P}\big(\hat{\mathcal{I}} \subseteq \mathcal{I}_a\big) \geq 1 - \alpha.$$

This means that, with high probability, all the SDGs that have been selected by the procedure are truly ideal. The probability here is with respect to the held out data. An important aspect to highlight is that while $\mathcal{I}_a$ is integral to the formulation of our claim, it is not necessary to know what it contains, and such information is never used in the procedure to select $\hat{\mathcal{I}}$.

4.2 FIDELITY

We note that SDGs perfectly matching the data distribution of the original data is not possible to achieve in general with an arbitrary dataset. Thus, the objective here is not to guarantee synthetic data that emulates perfectly the real data distribution, but rather that the distribution of the synthetic data generated by the selected data generators reside within an acceptable threshold, ϵ that the user of the data or a regulatory body will specify. Let $D(p, q)$ denote a theoretical measure of distributional discrepancy. In the methodology, we will specify a particular choice (e.g., MMD^2), but for the problem formulation, we work with a general discrepancy measure.

In our setting, each generator induces its own discrepancy by letting the real data follow p and the synthetic data follow q_i. Thus, for each generator G_j, the corresponding discrepancy is $D(p, q_j)$. We choose the θ_i, introduced earlier, as a measure of distributional discrepancy.

The goal is to select a subset $\widehat{\mathcal{I}} \subseteq \mathcal{I}$ of the indexes of the trained generators $(\hat{G}_1, \ldots, \hat{G}_N)$ such that

$$D(p, q_i) < \epsilon \quad \forall i \in \widehat{\mathcal{I}}.$$

with high probability. More formally, define the collection of ideal indices for the context of fidelity as

$$\mathcal{I}_a = \{i \in \mathcal{I} \mid D(p, q_i) < \epsilon\}.$$

In particular, for our experiments, this measure is $\text{MMD}^2(\mathcal{F}, p, q_i)$, where p denotes the distribution of the real data, q_i denotes the distribution induced by the trained generator $\hat{G}_i$, and $\mathcal{F}$ is the RKHS associated with a Gaussian kernel. The specific measure we have chosen for the experiment is not necessary to use, but what is necessary is that the estimator admits a suitable concentration inequality as described in section 5.

5 METHODOLOGY

This section provides an overview of the method and its theoretical foundations.

5.1 PROCEDURE

The following provides an overview of the steps that are taken in the procedure.

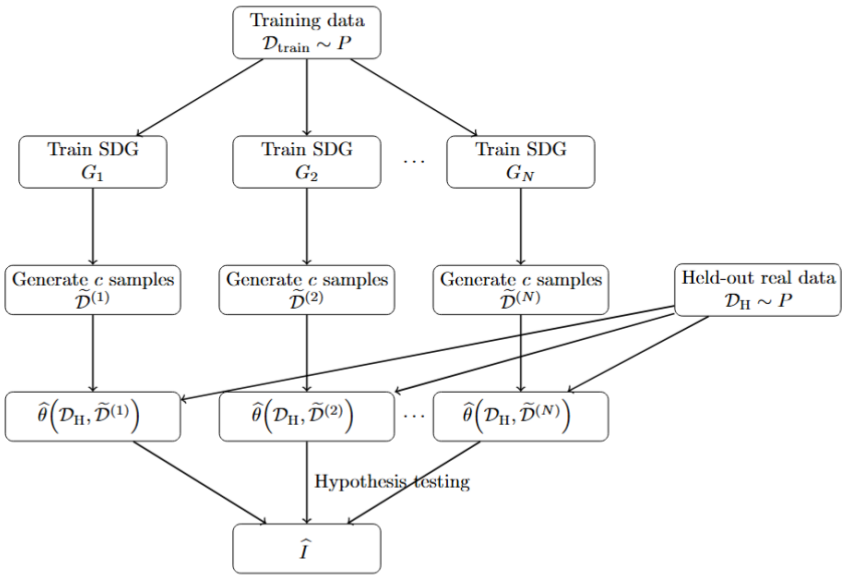

Figure 1: This figure provides a diagram of the procedure. Candidate models are first constructed and trained, after which samples are generated from each trained candidate model. The hypothesis testing procedure is then applied to obtain the certified subset $\hat{\mathcal{I}}$.

1. **Candidate Generator Construction:** In this step, the training dataset $\mathcal{D}_{\text{train}}$, is used to construct a collection of candidate generator procedures $\{G_i\}_{i=1}^N$. The particular procedure used to obtain the resulting trained generators $\{\hat{G}_i\}_{i=1}^N$ is left at the liberty of the practitioner and may involve arbitrary training, filtering, ranking, or model selection strategies based solely on $\mathcal{D}_{\text{train}}$. The requirement is that the hold-out dataset, $\mathcal{D}_{\text{H}}$, must remain unused during this stage, as the statistical guarantees are derived entirely from this hold-out dataset.
2. **Hold-Out Evaluation:** Generate a synthetic dataset of size c from each of the trained generator models $\{\hat{G}_i\}_{i=1}^N$ to obtain estimates, $\hat{\theta}_i$, of the measure of interest, θ_i, between the generated data $\widetilde{\mathcal{D}}_i$ and the real data from the hold-out set $\mathcal{D}_{\text{H}}$.
3. **Multiple Hypothesis Testing:** Using the estimators $\{\hat{\theta}_i\}_{i=1}^N$ obtained for the trained generators $\{\hat{G}_{\lambda_i}\}_{i=1}^N$, compute valid p-values for the corresponding hypotheses and apply a FWER controlling multiple testing procedure.

The following is the multiple hypothesis test that we construct: For each candidate generator G_i, $i = 1, \ldots, N$, we test

$$H_i : \text{MMD}^2(\mathcal{F}, p, q_i) \geq \epsilon \qquad \text{vs.} \qquad H_i' : \text{MMD}^2(\mathcal{F}, p, q_i) < \epsilon.$$

Thus, rejecting H_i corresponds to certifying that generator G_i has population discrepancy smaller than the desired threshold ϵ. The p-values associated with these are elaborated in the next section and Appendix B.

5.2 THEORY FOR CERTIFICATION GUARANTEE

Theorem 1 (Distribution-free certification guarantee for SDG subset). Suppose that a collection of synthetic data generators (SDGs) $\{G_i\}_{i\in\mathcal{I}}$ is trained using the training dataset $\mathcal{D}_{\text{train}}$, where each generator G_i induces a synthetic data distribution q_i. Let p denote the distribution of the real data.

For each generator $i \in \mathcal{I}$, let $\theta(i) := D(p, q_i)$ denote a population measure of discrepancy between the real and synthetic distributions, and suppose that its estimator $\widehat{\theta}_c(i)$, computed using a hold-out dataset of size c, satisfies the concentration inequality

$$\mathbb{P}\Big(\theta(i) - \widehat{\theta}_c(i) > t\Big) \le B(t, c),$$

for all $t > 0$, where for all $k \in \mathbb{R}$, $\sup_{t\in(k,\infty)} B(t, c) < \infty \quad$ and $\quad \lim_{c\to\infty} B(t, c) = 0 \quad \forall t$.

Define the null and alternative hypotheses as $H_{0,i} : \theta(i) \in \Theta_0$ and $H_{1,i} : \theta(i) \in \Theta_a$, where $\Theta_a = (-\infty, \epsilon)$ and $\Theta_0 = [\epsilon, \infty)$.

Let $\widehat{\mathcal{I}}$ denote the set of rejected hypotheses obtained from an FWER-controlling multiple testing procedure at level α, and define

$$\mathcal{I}_a = \{i \in \mathcal{I} : \theta(i) \in \Theta_a\}.$$

Then

$$\mathbb{P}\Big(\widehat{\mathcal{I}} \subseteq \mathcal{I}_a\Big) \ge 1 - \alpha.$$

The result follows by constructing valid p-values from the concentration bounds associated with the discrepancy estimators and applying an FWER-controlling multiple testing procedure. Since we are using multiple hypothesis testing, these p-values using the concentration inequality are formulated using finite-sample concentration inequalities. The complete details are deferred to Appendix A.

The requirement for the existence of the concentration bound is widely applicable. In particular, for distribution free guarantees, there are a few available readily. If the estimator is bounded, Hoeffding inequality is available Hoeffding (1963). The McDiarmid inequality also provides a way to give distribution free bounds of the sort that are required for the procedure (McDiarmid et al. (1989)). However, if available, it is best to use concentration bounds that are sharpest as this will enable rejections to be made when the measure truly is within the ideal range.

p-value for the MMD^2 when using Gaussian kernel In order to apply FWER-controlling procedures such as the graphical procedure by Bretz et al. (2009), valid p-values must first be constructed for each hypothesis. Since our experiments use the MMD^2 measure, we construct a valid p-value for testing $H_0 : \mathrm{MMD}^2(\mathcal{F}, p, q) \ge \epsilon$ against $H_1 : \mathrm{MMD}^2(\mathcal{F}, p, q) < \epsilon$.

Let $m_2 = \lfloor c/2 \rfloor$. A valid p-value for this test using the observed estimator $\widehat{\mathrm{MMD}}^2_{u,\text{obs}}$ is

$$p_B\Big(\widehat{\mathrm{MMD}}^2_{u,\text{obs}}\Big) = \exp\left(-\frac{m_2\left(\epsilon - \widehat{\mathrm{MMD}}^2_{u,\text{obs}}\right)^2}{8}\right).$$

6 EXPERIMENTAL RESULTS

The objective of these experiments is not to demonstrate the superiority of any particular generative model architecture, but rather to verify that the SDG selected by the proposed procedures can achieve guarantees on the population level MMD^2 and also how the choice of N can change the outcome of how many SDGs can be certified.

6.1 EXPERIMENTAL PROCEDURES

All our experiments used GAN-based synthetic data generators and the training sample size was fixed at 20,000 and the hold-out sample size at 4,500. The models differed through their generator and discriminator architectures. Details of the architectures for GAN can be found in the Appendix B. SDG candidate selection was performed by first training a collection of 6 models and then

choosing the top N models according to the empirical MMD^2 estimate between generated samples and a random subset of the training data. For the Gaussian kernel used in the MMD^2 computations and hypothesis testing bounds, the bandwidth parameter σ was selected using the median heuristic computed from a subset of the training data. These are the N candidates that we will perform the multiple hypothesis test on. We report results for $N = 3$ and $N = 4$.

To this end, we use simulation studies where the underlying data-generating mechanism is known, and we are able to generate around 100,000 extra samples for each of the datasets for each of the trained SDG. Details of the architectures for the simulated datasets can be found in the Appendix B.

6.2 Results

A table of the reference values for each model–dataset pair, together with details on how the estimates were computed, is reported in Appendix B. In both Figures 2 and 3, each dataset is divided into two columns.

The first column summarizes the reference MMD^2 estimates: the symbol "O" denotes values smaller than $\epsilon = 0.15$, "X" denotes values larger than 0.15, and "--" denotes cases where the estimate was not statistically distinguishable from 0.15.

The second column compares the claims made by the procedure, represented by $\hat{\mathcal{I}}$, regarding the MMD^2 between the trained SDG and the real data relative to the reference estimates reported in Appendix B. In this column, "O" denotes models selected by the procedure whose reference MMD^2 values were also below 0.15. The symbol "--" denotes cases where the models were not selected and the reference estimate was either inconclusive or larger than 0.15, while "——" denotes models that were not selected despite having reference MMD^2 values below 0.15. Red "O" markers were reserved for cases where a selected model had a reference MMD^2 value larger than 0.15 or inconclusive; this did not occur in the experiments. Blank entries in the second column correspond to models that were not selected as candidates.

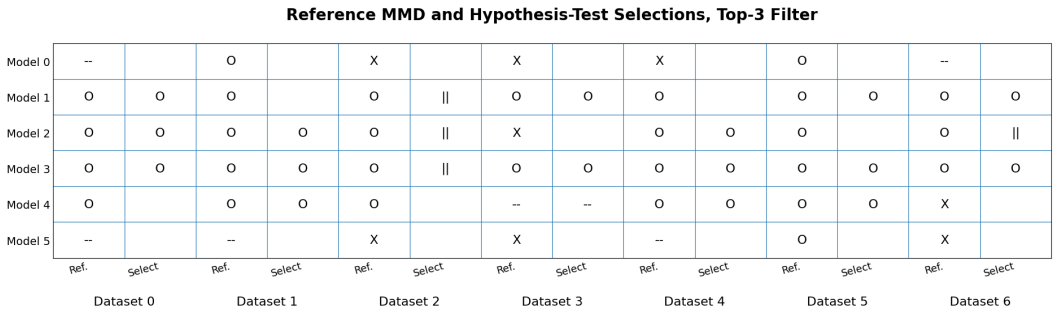

Figure 2: Result for $N = 3$ selection procedure across datasets. Each dataset is divided into two columns: the first indicates the candidate SDGs selected in the top 3, while the second shows the claims made by the procedure, represented by $\hat{\mathcal{I}}$, regarding whether the trained SDG satisfies $\mathrm{MMD}^2 < 0.15$.).

The guarantees claimed for the selected models possessing MMD^2 values within the desired region are supported by these results, as no red "O" markers appear. Furthermore, we can see that for some datasets, such as dataset 2, there were models that satisfied the criterion based on the reference MMD^2 estimates, but the procedure refrained from making claims about them by not selecting them. Additionally, we highlight that while it is possible to increase the number of selections, as occurred for datasets 1 and 4, it is also possible that some models that would have been selected when using only 3 candidates are no longer selected, such as model 2 for dataset 0. This occurs because, as N increases, the multiple testing procedure generally becomes more conservative.

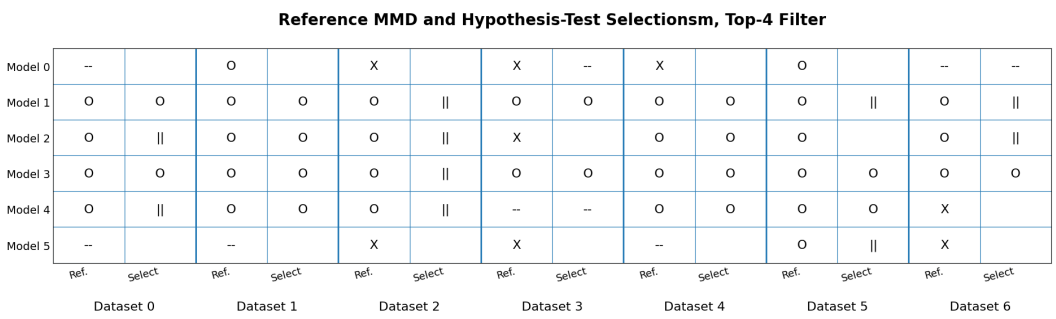

Figure 3: Result for $N = 4$ selection results across datasets. This figure depicts the same results as Figure 2, except that the top 4 models were selected as candidate SDGs instead.

7 Conclusions, Limitations and Future work

The paper introduced a framework for certifying the quality of a subset of SDGs through distribution-free and nonasymptotic multiple hypothesis testing procedures. Rather than relying solely on empirical results regarding data quality or asymptotic theoretical guarantees, we were able to construct a means by which probabilistic guarantees could be provided regarding whether the population-level discrepancy measure between the real data and the data generated by the SDG lies below a desired threshold. The methodology itself is highly flexible since it is not tied to a particular generative model architecture or discrepancy measure, provided that the estimator admits a suitable concentration inequality.

A key limitation is that the holdout data must remain unused after certification. Retraining a certified SDG on this data may invalidate the guarantee, since its fidelity could change. This can be costly when data are scarce, as part of the sample must be reserved exclusively for certification.

A non-selected SDG should not be interpreted as failing the fidelity criterion; it only means there is insufficient evidence to certify it. This conservatism is inherent to finite-sample, distribution-free guarantees and can reduce power when the holdout sample is small or the estimated discrepancy is close to the threshold. The framework also incurs additional computational and memory costs because multiple candidate generators must be trained and stored.

There are several approaches to mitigate some of these issues. For instance, reducing the total number of candidate models that will be tested alleviates the multiple testing burden. Therefore, it is of interest to select a relatively small collection of candidate SDGs for testing. We also wish to note that our experiments were intended primarily to demonstrate the concept, and there are several approaches one can take to improve the proposed candidates. We did not focus extensively on the engineering aspects of candidate model selection, even though such considerations may substantially improve the number of SDGs that can be safely deemed acceptable by the procedure. For instance, even though our choice was to select the top-N models based on a random subset of the training data, there is no issue with starting from a much larger class of models, using additional training and validation splits, or incorporating learned weighting schemes within the graphical testing procedure. The indispensable requirement is that the holdout data must remain unused for any decision made regarding the candidate models proposed for the hypothesis test. On the theoretical side, the development of improved FWER controlling procedures, discrepancy estimators, and tighter concentration inequalities would help alleviate the conservatism of the procedure.

Overall, the framework demonstrates that it is possible to make rigorous finite-sample claims regarding the fidelity of a subset of SDGs while remaining largely agnostic to the underlying model architecture in a way that is dependent on the data. Although such guarantees necessarily introduce conservatism and additional computational requirements, they provide a principled alternative to relying solely on empirical observations or asymptotic arguments when evaluating synthetic data quality.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the conceptual framing and research methodology, support the interpretation and presentation of experimental results. We have not used generative AI tools to generate synthetic datasets, implement the proposed methods, assist in mathematical proofs, clean or reformat datasets, translate research material, or conduct qualitative or thematic data analysis; the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to draft and edit portions of the manuscript for clarity and concision, suggest manuscript structure, titles and keywords, summarize and organize relevant literature, and assist with brainstorming and identifying potential gaps in presentation. All AI-assisted mathematical statements and methodological suggestions were independently checked by the authors against the underlying derivations and established literature, and all descriptions of experimental results were verified against the actual outputs of our experiments. The authors reviewed and revised all AI-assisted text and take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops statistical methodology for certifying the fidelity of synthetic data generators and does not involve human subjects or sensitive personal data. The proposed certificate concerns only a prespecified fidelity criterion and should not be interpreted as guaranteeing privacy, fairness, downstream utility, or absence of memorization. Care is therefore required when using certified generators in high-stakes applications.

REPRODUCIBILITY STATEMENT

We provide code for reproducing all experiments, including synthetic data generation, candidate SDG training, discrepancy estimation, p-value computation, and multiple-testing certification. The manuscript and appendix describe the experimental setup, model architectures, data splits, and computational resources, while the accompanying code contains the complete training and evaluation pipeline.

REFERENCES

Erik Altman, Jovan Blanuša, Luc Von Niederhäusern, Béni Egressy, Andreea Anghel, and Kubilay Atasu. Realistic synthetic financial transactions for anti-money laundering models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, June 2025. doi: 10.1214/24-AOAS1998.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.

Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015. doi: 10.1214/15-AOS1337.

Wacha Bounliphone, Eugene Belilovsky, Matthew B. Blaschko, Ioannis Antonoglou, and Arthur Gretton. A test of relative similarity for model selection in generative models. In *International Conference on Learning Representations (ICLR)*, 2016.

Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. A graphical approach to sequentially rejective multiple test procedures. *Statistics in Medicine*, 28(4):586–604, 2009. doi: 10.1002/sim.3495.

Florentina Bunea, Marten H. Wegkamp, and Anna Auguste. Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, 136 (12):4349–4364, 2006. doi: 10.1016/j.jspi.2005.03.011.

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2 edition, 2002. ISBN 0-534-24312-6.

Yuntao Du and Ninghui Li. Systematic assessment of tabular data synthesis. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2024.

Gintare Karolina Dziugaite, Daniel M. Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 258–267, 2015.

Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. Utility metrics for evaluating synthetic health data generation methods: Validation study. *JMIR Medical Informatics*, 10(4): e35734, 2022. doi: 10.2196/35734.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. doi: 10.1145/3422622.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Sami Hanhijärvi. Multiple hypothesis testing in pattern discovery. In *International Conference on Discovery Science*, pp. 122–134. Springer, 2011.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Yosef Hochberg and Ajit C. Tamhane. *Multiple Comparison Procedures*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, New York, 1987. ISBN 0-471-82222-2.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Valter Hudovernik, Martin Jurkovič, and Erik Štrumbelj. Benchmarking the fidelity and utility of synthetic relational data. *arXiv preprint arXiv:2410.03411*, 2024.

Lucas Janson and Weijie Su. Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1):960 – 975, 2016. doi: 10.1214/16-EJS1129. URL `https://doi.org/10.1214/16-EJS1129`.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1): 148–188, 1989.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Zhaozhi Qian, Rob Davis, and Mihaela van der Schaar. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations (ICLR)*, 2022.

Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *Journal of Machine Learning Research: Workshop and Conference Proceedings*, pp. 448–455, 2009.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 2256–2265, 2015.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428, 2021.

Xinyu Tian and Xiaotong Shen. Conditional data synthesis augmentation. *Journal of the American Statistical Association*, pp. 1–13, 2026.

Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. In *International Conference on Learning Representations (ICLR)*, 2017.

A PROOF OF MAIN RESULTS

Valid p-values from Concentration Bounds To apply FWER controlling multiple testing procedures, valid p-values must first be constructed for each hypothesis. In our framework, these p-values are obtained directly from concentration inequalities satisfied by the discrepancy estimators.

that an estimator $\widehat{\theta}$ of a population quantity θ satisfies the concentration inequality

$$\mathbb{P}_\theta\Big(\theta - \widehat{\theta} > t\Big) \le B(t, c).$$

Consider testing

$$H_0 : \theta \in \Theta_0, \qquad H_1 : \theta \in \Theta_a.$$

We wish to construct a valid p-value $p_B(\widehat{\theta})$ satisfying

$$\mathbb{P}_\theta\Big(p_B(\widehat{\theta}) < \alpha\Big) \le \alpha, \qquad \forall \theta \in \Theta_0.$$

Define $p_B(\widehat{\theta}_{\text{obs}}) := \sup_{\theta \in [\epsilon, \infty)} B(\theta - \widehat{\theta}_{\text{obs}}, c)$.

Then, for all $\theta \in [\epsilon, \infty)$,

$$\begin{aligned}\mathbb{P}_\theta\Big(p_B(\widehat{\theta}_{\text{obs}}) < \alpha\Big) &\le \mathbb{P}_\theta\left(\sup_{\theta' \in [\epsilon, \infty)} p_\theta\Big(\widehat{\theta} < \widehat{\theta}_{\text{obs}}\Big) < \alpha\right) \\ &\le \alpha.\end{aligned}$$

The first inequality follows since the p-value based on the unknown true distribution is $p(\widehat{\theta}_{\text{obs}}) = \mathbb{P}_\theta\Big(\widehat{\theta} < \widehat{\theta}_{\text{obs}}\Big)$,$itfollowsthat$

$$p(\widehat{\theta}_{\text{obs}}) = \mathbb{P}_\theta\Big(\theta - \widehat{\theta} > \theta - \widehat{\theta}_{\text{obs}}\Big) \le B\Big(\theta - \widehat{\theta}_{\text{obs}}, c\Big).$$

$Thus, \sup_{\theta' \in [\epsilon, \infty)} p_\theta\Big(\widehat{\theta} < \widehat{\theta}_{\text{obs}}\Big) \le \sup_{\theta' \in [\epsilon, \infty)} B(\theta' - \widehat{\theta}_{\text{obs}}, c) = p_B(\widehat{\theta}_{\text{obs}})$.

The second inequality follows from the probability integral transform and the standard validity of p-values (see Casella & Berger (2002) for details). The above describes the construction of a valid p -value for each of the SDG in the collection. In particular, we associate the certification criterion for generator G_i with the hypothesis pair

$$H_{0,i} : \theta(i) \in \Theta_0, \qquad H_{1,i} : \theta(i) \in \Theta_a,$$

where $\theta(i) = D(p, q_i)$ denotes the population discrepancy between the real data distribution p and the synthetic distribution q_i induced by generator G_i. Consequently, rejecting $H_{0,i}$ corresponds to certifying, with controlled error probability, that the discrepancy induced by G_i lies below the prescribed tolerance threshold ϵ.

Here, we demonstrate that the FWER procedure is equivalent to our claim that

$$\mathbb{P}(M > 0) \leq \alpha$$

is equivalent to

$$\mathbb{P}(\hat{\mathcal{I}} \subseteq \mathcal{I}_a) \geq 1 - \alpha.$$

Recall the definition

$$M := \left|\hat{\mathcal{I}} \cap \mathcal{I}_0\right|.$$

Thus,

$$\begin{aligned}
&\{\omega \in \Omega : M(\omega) > 0\}^c = \{\omega \in \Omega : \hat{\mathcal{I}}(\omega) \cap \mathcal{I}_0 \neq \emptyset\}^c \\
&= \{\omega \in \Omega : \hat{\mathcal{I}}(\omega) \cap \mathcal{I}_0 = \emptyset\} \\
&= \{\omega \in \Omega : \hat{\mathcal{I}}(\omega) \subseteq \mathcal{I}_0^c\} \\
&= \{\omega \in \Omega : \hat{\mathcal{I}}(\omega) \subseteq \mathcal{I}_a\}
\end{aligned}$$

where $\mathcal{I}_a := \mathcal{I}_0^c$ denotes the indices corresponding to the alternative hypotheses.

Note: The dependence on ω appears only through $\hat{\mathcal{I}}(\omega)$, since the rejected set is random, while $\mathcal{I}_0$ (and hence $\mathcal{I}_a$) is fixed. This inclusion emphasizes that the probability is taken over the randomness of $\hat{\mathcal{I}}$, which is a result of the hold-out data.

Therefore,

$$\mathbb{P}(M > 0) \leq \alpha \quad \Longleftrightarrow \quad 1 - \alpha \leq \mathbb{P}\big(\hat{\mathcal{I}} \subseteq \mathcal{I}_a\big).$$

Lastly, to show that the p-value for MMD^2 used in the paper is valid, we use the Hoeffding-based concentration bound from Gretton et al. (2012):

$$B(t, c) = \exp\left(-\frac{t^2 m_2}{8}\right), \qquad m_2 = \lfloor m/2 \rfloor.$$

Since $B(t, c)$ is monotone decreasing in t, we obtain

$$\sup_{\mathrm{MMD}^2 \in [\epsilon, \infty)} B\Big(\mathrm{MMD}^2 - \widehat{\mathrm{MMD}}^2_{u,\mathrm{obs}},\, c\Big) = \exp\left(-\frac{m_2\left(\epsilon - \widehat{\mathrm{MMD}}^2_{u,\mathrm{obs}}\right)^2}{8}\right).$$

B Experimental Details and Supplemental Figures

We perform all our experiment on a laptop with a 13th Gen Intel(R) Core(TM) i7-13700H, 2400 Mhz, 14 Core(s), 20 Logical Processor(s) and 16.0 GB of RAM. The time taken to run all the experiments together was around 40 minutes.

Table 1: Generator and discriminator architectures used in the experiments.

Model	Generator Architecture	Discriminator Architecture
1	Dense(256, LeakyReLU) → Dense(512, LeakyReLU)	Dense(32, LeakyReLU) → Dense(64, LeakyReLU) → Dense(32, LeakyReLU)
2	Dense(512, ELU) → Dense(1024, GELU)	Dense(32, ELU) → Dense(64, LeakyReLU) → Dense(32, ReLU)
3	Dense(128, ReLU) → Dense(128, ReLU)	Dense(32, ReLU) → Dense(16, ReLU) → Dense(16, ReLU)
4	Dense(64, ReLU) → Dense(64, ReLU) → Dense(32, ReLU)	Dense(32, ReLU) → Dense(32, ReLU)
5	Dense(32, GELU) → Dense(16, ELU) → Dense(32, GELU) → Dense(32, ReLU)	Dense(32, ELU) → Dense(32, ELU) → Dense(32, ELU) → Dense(64, ELU)
6	Dense(512, LeakyReLU) → Dense(1024, ReLU)	Dense(16, LeakyReLU) → Dense(32, LeakyReLU)

Table 2: Architectures used for constructing the real data generating distributions.

Model	Architecture
1	Dense(512, GELU) $\rightarrow$ Dense(256, LeakyReLU) $\rightarrow$ Dense(512, GELU)
2	Dense(64, LeakyReLU) $\rightarrow$ Dense(32, GELU) $\rightarrow$ Dense(32, ELU)
3	Dense(512, ReLU) $\rightarrow$ Dense(32, GELU)
4	Dense(512, GELU) $\rightarrow$ Dense(512, ReLU) $\rightarrow$ Dense(64, ReLU)
5	Dense(4, ReLU) $\rightarrow$ Dense(64, ReLU) $\rightarrow$ Dense(64, ReLU)
6	Dense(1024, ELU) $\rightarrow$ Dense(512, GELU)
7	Dense(1024, GELU) $\rightarrow$ Dense(1024, LeakyReLU) $\rightarrow$ Dense(1024, GELU) $\rightarrow$ Dense(1024, ReLU)

Note that the generators have a layer of batch normalization before each activation function and at the end for all of the generators, and use a final tanh activation function at the output layer. The discriminator have a dropout layer with probability 0.1 after each activation function.

The real data distributions were generated using randomly initialized fully connected neural networks. Each model maps Gaussian noise vectors to the data space through a sequence of dense layers and activation functions, followed by a final tanh output layer so that the generated data lies in $(-1, 1)^d$. The network weights were independently initialized from Gaussian distributions with variance scaled inversely proportional to the input dimension of each layer, while biases were initialized to zero.

Reference MMD for (Dataset, Generator) pairs

	Dataset 0	Dataset 1	Dataset 2	Dataset 3	Dataset 4	Dataset 5	Dataset 6
Model 0	--	O	X	X	X	O	--
Model 1	O	O	O	O	O	O	O
Model 2	O	O	O	X	O	O	O
Model 3	O	O	O	O	O	O	O
Model 4	O	O	O	--	O	O	X
Model 5	--	--	X	X	--	O	X

Figure 4: For computational feasibility, the computations for the use the linear-time MMD^2 estimator from Gretton et al. (2012). Using the probability bound provided in Gretton et al. (2012), we construct a 95% confidence interval around the estimated MMD^2 value. We then compare this interval to the certification threshold of 0.15. If the interval contains 0.15, the corresponding entry is marked with "--". Otherwise, the entry is marked with "O" if the estimated MMD^2 is below 0.15, and with "X" if it exceeds 0.15.