BEYOND AUGMENTATION PERFORMANCE: A UTILITY, ATTRIBUTION, AND PRIVACY AUDIT FOR SYNTHETIC CLINICAL SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models for small clinical datasets are commonly evaluated by adding synthetic samples to real training data and measuring downstream performance. This mixed-training benchmark can confound generator quality with class rebalancing, does not isolate the contribution of generated samples, and provides no evidence of patient-level privacy. We address these limitations with a structured empirical audit that evaluates three distinct properties: utility through train-on-synthetic, test-on-real substitution under prespecified equivalence testing; attribution through reverse transfer and zero-signal controls; and privacy through membership-inference and copy-rate gates. On ATR-FTIR, a latent-diffusion generator with near-chance synthetic-to-real transfer matches its corrected counterpart in augmentation AUC, showing that mixed-training performance can conceal generator failure. Across ATR-FTIR ($n = 584$) and LIBS ($n = 112$), linear PCA-Gaussian matches or exceeds the individual nonlinear generators, while mixtures improve substitution stability. Post-hoc corrections create apparent utility even from label-shuffled or noise inputs, whereas SMOTE shows strong utility yet suffers from near-perfect membership discrimination under the evaluated distance-based attack. In the end we produce fully synthetic artifacts with a classifier passing equivalence criteria with a one trained on real data. However, a post-hoc full-spectrum attack detects membership signal in the assembled LIBS artifact ($\mathrm{AUC} = 0.58$, $p = 0.004$), showing that empirical privacy depends on the adversarial representation. Thus, augmentation performance alone does not establish generator utility; utility, attribution, and privacy require separate evaluation. Our audit is a falsification framework, not a formal privacy guarantee, clinical validation, or certification of release safety.

1 INTRODUCTION

Synthetic data are often evaluated by adding generated samples to real training data and comparing resulting classifier with an unaugmented baseline. This mixed-training benchmark evaluates the combined dataset but does not isolate generator's contribution: class rebalancing, duplicated observations, or post-hoc insertion of real class statistics can improve performance even when the synthetic data do not transfer to held-out real patients. We study this problem in two small clinical-spectroscopy cohorts. ATR-FTIR plasma spectroscopy has detected endometrial cancer with 87% sensitivity and 78% specificity (Paraskevaidi et al., 2020), while LIBS of plasma has distinguished SARS-CoV-2-positive from negative donors (Melikechi et al., 2024). These settings motivate synthetic cohorts that retain task-relevant structure while limiting dependence on individual training records. Our FTIR experiments expose a failure of the mixed-training benchmark. A latent-diffusion generator yields augmentation performance comparable to real-only baseline, although a classifier trained only on its samples performs near chance on held-out real patients (§4). We call such a generator *operationally failed* for the task. Correcting its sampler, training budget, and guidance substantially improves synthetic-to-real and reverse transfer, while mixed-training AUC changes little. In this setting, class rebalancing helps explain why mixed-training performance is insensitive to generator quality.

A candidate-use evaluation should therefore examine three properties separately: *substitution utility*, whether synthetic-trained classifiers transfer to held-out real patients within a prespecified margin; *attribution*, whether utility arises from the generator rather than post-hoc insertion of real class statistics; and *empirical privacy*, whether training-set membership is detectable under specified attacks and representations. We combine these tests in a falsification-oriented audit (§6). The audit

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.

is asymmetric: failure of a gate provides evidence against candidate use, whereas passing establishes compatibility only with the stated models, representations, and attacks. It does not certify anonymity, differential privacy, clinical validity, or unrestricted release safety.

Our contributions are threefold. First, we show that mixed real–synthetic augmentation performance can remain nearly unchanged when a generator moves from near-chance to informative synthetic-to-real transfer. Second, we formulate a generator-agnostic audit that separates substitution utility, attribution, and empirical membership risk through equivalence testing, reverse transfer, zero-signal and oversampling controls, and explicit gates. Third, across FTIR and LIBS, we find that a linear PCA-Gaussian control is competitive with individual nonlinear generators, post-hoc corrections can produce apparent utility from zero-signal inputs, and mixtures of PCA-Gaussian and diffusion samples can improve transfer stability. We evaluate class-conditional latent diffusion, conditional WGAN-GP, PCA-Gaussian, and a SMOTE control (§5). The supplementary material provides reproducible outputs and two synthetic artifacts with provenance manifests; neither contains real patient spectra. Privacy results remain conditional on the evaluated threat model: in particular, the LIBS artifact passes the prespecified task-representation attack but exhibits membership signal in a post-hoc full-spectrum analysis. It is therefore provided for reproducibility, not as a privacy-certified substitute for unrestricted distribution.

2 Related Work

Clinical spectroscopy and data augmentation : FTIR fingerprinting of biofluids is an established approach to diagnostic classification (Baker et al., 2014; Fadlelmoula et al., 2023; Huber et al., 2021; Butler et al., 2019; Paraskevaidi et al., 2020; Schiemer et al., 2025). However, clinical spectroscopy studies commonly involve small cohorts; a review of cancer applications identifies limited sample sizes and the need for larger clinical studies as persistent challenges (Su & Lee, 2020). Data augmentation for spectral classification includes classical oversampling methods such as SMOTE (Chawla et al., 2002; He et al., 2008; Fernández et al., 2018) and generative approaches based on adversarial networks (Chung et al., 2024; García-Esteban et al., 2023; Flanagan & Glavin, 2024). The closest clinical-spectroscopy study to ours is McHardy et al. (2023), who used Wasserstein-GAN augmentation for ATR-FTIR serum spectra and reported an increase in convolutional-classifier AUC from 0.661 to 0.757 under mixed real–synthetic training.

Diffusion models for spectra: Denoising diffusion models (Ho et al., 2020; Nichol & Dhariwal, 2021) and classifier-free guidance (Ho & Salimans, 2021) provide a flexible framework for conditional generation, while latent diffusion reduces computational cost by operating in an autoencoder representation (Rombach et al., 2022). Diffusion models have been extended from images to one-dimensional signals and time series (Kong et al., 2021; Rasul et al., 2021), and Raman spectra have also been modeled through image-based diffusion representations (Yao et al., 2024). Nevertheless, diffusion models remain comparatively underexplored in biospectroscopy (Guo et al., 2025). Existing spectral-generation studies primarily emphasize mixed-training augmentation performance. Among the clinical-spectroscopy studies reviewed, we did not identify one that jointly evaluates synthetic-only transfer under an explicit equivalence margin and compares nonlinear generators against a class-conditional linear-Gaussian baseline.

Synthetic-data utility: Synthetic data are often proposed as substitutes for restricted datasets, but their utility depends on whether models trained on synthetic observations transfer to real data (Jordon et al., 2022). Train-on-synthetic, test-on-real evaluation directly measures this transfer, whereas mixed real–synthetic training cannot isolate the information contributed by the generator. Moreover, downstream models can inherit distributional errors from the generator, and combining outputs from multiple generator fits may improve robustness (van Breugel et al., 2023). Our audit therefore treats synthetic-to-real substitution as the primary utility endpoint, compares it with the corresponding real-trained classifier under a prespecified equivalence margin, and uses reverse transfer to determine whether generated samples preserve task-relevant class structure.

Privacy and synthetic-data auditing: Synthetic data do not provide a predictable privacy advantage merely because they are generated rather than directly sampled from the original dataset (Stadler et al., 2022). Membership-inference attacks test whether an adversary can distinguish records used to train a model from records not used during training (Shokri et al., 2017; Carlini et al., 2022). For synthetic datasets, nearest-neighbor and distance-to-closest-record statistics are also commonly used to assess copying and training-record proximity (Park et al., 2018; Yale et al., 2020). Recent clinical synthetic-data benchmarks combine utility and privacy into a single aggregate score, for example by normalizing both against a random baseline and taking their harmonic mean (Sella et al., 2025).

In contrast, we treat empirical privacy as an eligibility gate: stronger utility cannot compensate for measured membership risk.

Our contribution is not a new individual utility or privacy metric. Rather, we show experimentally that mixed-training performance does not isolate generator utility and combine established tests into an audit in which substitution utility, attribution, and empirical privacy are evaluated separately. Zero-signal controls determine whether apparent utility originates from the generator or from post-hoc insertion of real class statistics, while positive and negative privacy controls calibrate the membership audit. This separation identifies which property fails and prevents strong predictive performance from masking unsupported attribution or measured membership risk. The resulting conclusions remain conditional on the evaluated classifiers, representations, and privacy attacks; the audit does not provide a formal privacy guarantee.

3 Cohorts and Evaluation Setting

ATR-FTIR: endometrial cancer: The ATR-FTIR cohort contains 242 healthy controls, 342 patients with endometrial cancer, and 68 with atypical hyperplasia (Paraskevaidi et al., 2020). We exclude atypical hyperplasia and average five replicate spectra per patient, yielding 584 patient-level spectra over the 1800–$900\,\mathrm{cm}^{-1}$ region. Following the source study, spectra undergo a Savitzky–Golay second derivative (window 5, order 2) and L_2 normalization (Savitzky & Golay, 1964), producing 235-point spectra. XGBoost is the primary classifier (Chen & Guestrin, 2016); full-spectrum logistic regression is secondary. The parsed deposit lacks the original instrument axis, but no analysis selects features by nominal wavenumber.

LIBS: COVID-19: The LIBS deposit contains 24,488-point plasma-emission spectra from 53 RT-PCR-negative and 59 positive labelled donors ($n = 112$), plus 40 unlabelled blind-test donors (Melikechi et al., 2024). Because these counts differ from those reported in the associated article, we use the deposited labels and identifiers. We discard the first of each 15-shot sequence, apply a two-sided $z > 1$ total-intensity filter, average retained shots by donor, and normalize each patient-level spectrum by total intensity. The resulting spectra match the deposited processed data to approximately 10^{-16} in a six-donor spot check. Four-feature logistic regression (LR4), using integrated Ca, K, Mg, and Na intensities, is primary; full-spectrum XGBoost is secondary. Source-study results and metric differences are summarized in Table 1.

Table 1: Reference classification results. Source values differ from ours in split, classifier, and, for LIBS, metric definition. Our entries are split-mean results over five patient-level splits and five seeds. Validation-frozen thresholds use training-side data only. Equivalence is evaluated using pooled out-of-fold AUC and is not tested against these values.

Cohort	Model	Threshold	AUC	Acc.	Sens.	Spec.
FTIR	Source PLS-DA	Kennard–Stone	0.880	0.830	0.870	0.780
FTIR	Real-trained XGB	Fixed 0.5	0.796	0.728	0.753	0.693
FTIR	Real-trained XGB	Val.-frozen	0.796	0.722	0.753	0.678
FTIR	Synthetic-trained XGB	20–40 real	0.778	0.712	0.739	0.674
LIBS	Source LR4	LOOCV / blind	–	0.883 / 0.820	0.900 / 0.794	0.867 / 0.875
LIBS	Real-trained LR4	Fixed 0.5	0.919	0.841	0.789	0.900
LIBS	Real-trained LR4	Val.-frozen	0.917	0.758	0.687	0.838

Splits, seeds, and validation: Both cohorts use five repeated stratified patient-level train–test splits, with all measurements from a patient confined to one side of each split. Each generator is trained independently for every split and five random seeds without access to test patients. Model fitting, early stopping, hyperparameter selection, and threshold selection use training-side data only, except in the explicitly identified real-data calibration analysis; all choices are frozen before test scoring. Predictions are averaged across seeds within each split and then across repeated held-out appearances, yielding one pooled prediction for 487 FTIR patients and 95 LIBS donors. Paired comparisons use identical patients, and bootstrap resampling uses unique patients as the sampling unit. The separate 409/175 FTIR Kennard–Stone split is used only for descriptive comparison (Kennard & Stone, 1969).

Utility estimands and uncertainty: We report *split-mean AUC*, with between-split standard deviation, and *pooled out-of-fold AUC*, computed from one averaged prediction per held-out patient

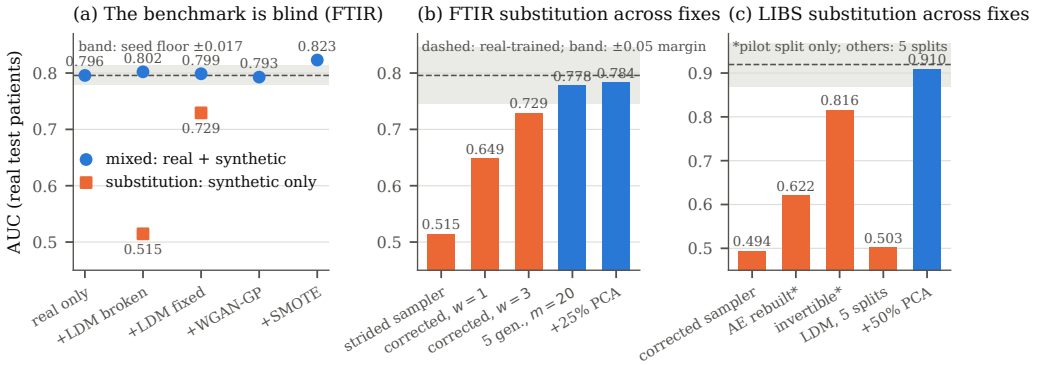


Figure 1: Mixed-training performance can conceal generator failure. **(a)** FTIR XGBoost over five patient-level splits and five seeds: mixed real–synthetic training at $1.0\times$ the class deficit (circles) and synthetic-to-real substitution (squares). Mixed-training performance is insensitive here, whereas substitution distinguishes the failed and corrected configurations. **(b–c)** Substitution AUC across successive FTIR and LIBS pipeline corrections; stars denote pilot-split results. The dashed region denotes the real-trained AUC within the prespecified 0.05 operational equivalence margin.

and used for equivalence testing. The real-trained primary classifiers achieve split-mean AUCs of 0.796 ± 0.020 for FTIR and 0.919 ± 0.025 for LIBS, with pooled AUCs of 0.817 and 0.890, respectively. Because threshold estimates are unstable in the small folds, ROC-AUC is primary; accuracy, sensitivity, and specificity are descriptive. Patient-level bootstrap intervals quantify uncertainty conditional on the observed cohorts, frozen splits, selected configurations, trained generators, and averaged predictions. They do not propagate uncertainty from repeating split construction, configuration selection, or the complete training procedure; between-split and, where applicable, between-seed variability are therefore reported separately.

4 Mixed-Training Performance Can Conceal Generator Failure

The mixed-training benchmark measures whether adding synthetic samples to real training data improves the resulting classifier; it does not isolate the generator's contribution. On FTIR, augmenting the minority class with SPECTRAL-LDM samples at $1.0\times$ the class deficit yields a split-mean XGBoost AUC of 0.802 over five patient-level splits and five seeds, compared with 0.796 for real-only training. Despite this competitive augmentation result, a classifier trained only on these synthetic spectra achieves AUC 0.515 on held-out real patients and 0.500 on the source study's Kennard–Stone split. Conversely, a real-trained classifier achieves reverse-transfer AUC 0.51 on the synthetic cohort. We therefore designate this configuration as *operationally failed* for the task; the associated implementation factors are examined in §5.1.

Evidence consistent with class rebalancing: At a fixed decision threshold, augmentation with the failed generator changes sensitivity and specificity from 0.753 and 0.693 to 0.788 and 0.674, respectively, while AUC changes from 0.796 to 0.802. These operating-point changes are consistent with class rebalancing and an altered fitted decision function. However, the experiment does not uniquely distinguish rebalancing from other effects of adding weakly informative observations. Correcting the sampler, optimization budget, and guidance increases synthetic-to-real AUC from 0.515 to 0.729 and reverse-transfer AUC from 0.51 to 0.93. In contrast, mixed-training AUC changes only from 0.802 to 0.799 at the same $1.0\times$ class deficit, a difference of 0.003. At $0.4\times$ the class deficit, the corrected generator achieves AUC 0.800; its 0.002 difference from the failed configuration at $1.0\times$ is also small relative to the observed across-seed standard deviation of 0.017. These comparisons show that mixed-training performance was insensitive to substantial changes in direct-transfer performance in this experiment, without identifying a unique causal mechanism.

Direct-transfer checks: Synthetic-to-real substitution and reverse transfer distinguish the failed and corrected configurations by approximately 0.21 and 0.42 AUC, respectively, whereas their mixed-training AUCs differ by only 0.003 at $1.0\times$ the class deficit. The two checks require training a classifier on synthetic data and evaluating it on held-out real patients (Esteban et al., 2017), or training on real data and scoring the synthetic cohort. We therefore use substitution as the primary utility endpoint in §6. Mixed-training evaluation remains appropriate when augmentation is the in-

tended use, but it evaluates the resulting classifier rather than the generator independently and does not establish that synthetic data can substitute for the real cohort.

5 GENERATORS, CONTROLS, AND CORRECTIONS

All methods use the patient-level splits, seeds, and classifiers defined in §3. Generator arms compare linear and nonlinear distributional models; controls distinguish predictive utility from generator-specific attribution and measured membership risk.

5.1 LATENT DIFFUSION AND IMPLEMENTATION ABLATIONS

Model: Spectral-LDM combines a frozen 1D convolutional autoencoder (strided convolutions, 12 latent channels, base width 64, and mean-squared reconstruction loss) with a class-conditional 1D U-Net ϵ-predictor (Ronneberger et al., 2015; Ho et al., 2020). The denoiser uses a cosine schedule with $T = 1000$ (Nichol & Dhariwal, 2021) and classifier-free guidance (Ho & Salimans, 2021):

$$\hat{\epsilon} = \epsilon_\psi(z_t, t, \varnothing) + w[\epsilon_\psi(z_t, t, y) - \epsilon_\psi(z_t, t, \varnothing)] .$$

Labels are replaced by the null class with probability 0.1 during training; $w = 1$ recovers the conditional prediction. Samples are mapped to spectra by the frozen decoder (Rombach et al., 2022). The corrected configuration uses 30,000 AdamW steps (learning rate 2×10^{-4}, EMA 0.999) and 100 stride-correct DDIM steps with $\eta = 1$ (Song et al., 2021). Separate models are trained for each cohort in domains selected by the fail-fast checks of Appendix B: the classifier domain for FTIR and globally centred linear intensity at stride 4 for LIBS.

Implementation-factor ablations: We isolate three factors underlying the operationally failed configuration in §4. First, the sampler evaluated 300 indices from the $T = 1000$ schedule but applied an ancestral update derived for $t \to t - 1$ across gaps of three or four indices. With an analytically optimal denoiser for $x_0 \sim \mathcal{N}(2.00, 0.50^2)$, this invalid update returns mean 1.19 and standard deviation 0.72, whereas stride-correct DDIM recovers mean 2.00 at both 300 and 50 steps (Appendix C). Second, 300 epochs provide only 1.5–2.1×10^3 optimizer steps on these cohorts, compared with 30,000 in the corrected configuration. Third, $w = 0.5$ targets half the conditional mean shift and, combined with the invalid update, recovers only about 30% of the original class mean. We fix $w = 3$ before confirmatory runs and report diagnostic sweep $w \in [1, 5]$ in §7.1.

5.2 BASELINES, CONTROLS, AND CORRECTIONS

Generator controls: Conditional WGAN-GP (Arjovsky et al., 2017; Gulrajani et al., 2017) operates in the classifier domain with latent dimension 100, base width 64, five critic updates per generator update, $\lambda_{\mathrm{GP}} = 10$, and 1,000 epochs; one model is trained per split and seed. It tests whether the findings are specific to latent diffusion. PCA-Gaussian fits 20 FTIR or 40 LIBS principal components to each real training fold, followed by a class-conditional diagonal Gaussian in component space; LIBS is fitted in the square-root intensity domain. This is a conceptual linear control, not a state-of-the-art generator: comparable performance would show that nonlinear capacity is unnecessary for the observed transfer. Although it samples continuous fitted distributions rather than explicitly reproducing training rows, its subspace and class parameters depend on the training data and are therefore included in the membership-risk audit. SMOTE (Chawla et al., 2002) interpolates real minority-class neighbours and serves as an oversampling and positive membership-risk control, not as a candidate generator. Its strong utility and near-perfect membership discrimination under the evaluated distance-based attack (§7.4) separate augmentation utility from candidate-use suitability.

Mixtures and attribution controls: Fixed-size mixtures contain a fraction $\phi \in \{0, 0.25, 0.5\}$ of PCA-Gaussian samples and $1 - \phi$ of Spectral-LDM samples, testing whether their distinct inductive biases provide complementary transfer. We also evaluate two post-hoc corrections (§7.3): *moment matching* aligns each synthetic class's feature-wise mean and variance with the real training fold, while *covariance matching* whitens it in a $k = 20$ PCA subspace and recolours it using the real-class covariance with shrinkage 0.1. Because these transformations insert real class-conditional statistics, we apply them identically to label-permuted Spectral-LDM samples and Gaussian noise. These zero-signal controls preserve, respectively, the generator marginal without valid class association and no learned generator structure. Improvement is attributed to the generator only when it exceeds that obtained by both controls.

6 A UTILITY, ATTRIBUTION, AND PRIVACY AUDIT

We evaluate synthetic cohorts on substitution utility, attribution, and empirical membership risk. The criteria were developed using FTIR and LIBS and frozen before the six-arm evaluation grid;

they were neither preregistered nor independently validated. Here, "confirmatory" denotes evaluation under this frozen grid, not confirmation on an external cohort. Passing therefore indicates compatibility only with the stated criteria, models, representations, and attacks; it does not establish anonymity, differential privacy, clinical validity, or release safety.

Substitution utility: The primary endpoint is train-on-synthetic, test-on-real AUC. Each synthetic training set contains m times the real class count ($m = 20$ in frozen evaluation), pooled equally across five seed-specific generators. From pooled out-of-fold predictions defined in §3, we compute

$$\Delta = \mathrm{AUC}_{\mathrm{syn}} - \mathrm{AUC}_{\mathrm{real}}.$$

A 90% confidence interval is obtained from 4,000 patient-level bootstrap resamples after repeated held-out predictions are averaged to one value per patient.

Using two one-sided tests (Schuirmann, 1987; Lakens, 2017), an arm satisfies the utility criterion when its 90% interval lies within $(-0.05, 0.05)$.

The 0.05 margin is a prespecified operational tolerance, not a clinically validated non-inferiority threshold; Appendix G evaluates margins from 0.03 to 0.10: the FTIR candidate remains equivalent at 0.03, whereas the LIBS candidate passes from 0.04. DeLong tests (DeLong et al., 1988) are diagnostic tests of $\Delta = 0$ and do not determine eligibility. Reverse-transfer AUC, obtained by applying a real-trained classifier to synthetic data, provides an additional diagnostic of preserved class structure but is not an eligibility gate.

Attribution: Label-permuted Spectral-LDM samples and Gaussian noise undergo the same correction, classifier, validation, and evaluation pipeline as each generator arm. An arm passes the point-estimate attribution criterion only if

$$\mathrm{AUC}_{\mathrm{gen}} > \max(\mathrm{AUC}_{\mathrm{perm}}, \mathrm{AUC}_{\mathrm{noise}})\,.$$

This falsification rule rejects utility reproducible without valid generator-derived class information; it is not a statistical-superiority test, and passing does not prove that all utility originates from the generator. For post-hoc corrections, we also compare

$$G_j = \mathrm{AUC}_{j,\mathrm{corrected}} - \mathrm{AUC}_{j,\mathrm{raw}}, \qquad j \in \{\mathrm{gen}, \mathrm{perm}, \mathrm{noise}\},$$

and credit generator-specific improvement only when $G_{\mathrm{gen}} > \max(G_{\mathrm{perm}}, G_{\mathrm{noise}})$. This gain analysis is diagnostic and does not replace the attribution criterion.

Empirical membership risk: Within each split, generator-training patients are members and held-out patients are non-members. The attack statistic is each patient's distance to its closest synthetic record in the prespecified classifier feature space (Park et al., 2018); attack AUC measures member–non-member discrimination. We additionally report true-positive rates at false-positive rates of 0.5–10%, using SMOTE and shuffled membership labels as positive and negative controls.

The *copy rate* is the fraction of synthetic rows closer to a training record than the fifth percentile of real-to-real nearest-neighbour distances. The descriptive *NN ratio* divides the median synthetic-to-real distance by the median real-to-real distance; unusually small or large values indicate potential copying or off-manifold generation. The operational task-space criteria are

$$\mathrm{AUC}_{\mathrm{attack}} < 0.55 \qquad \text{and} \qquad \text{copy rate} < 0.10.$$

These thresholds are prespecified but not externally validated, regulatory, or formal privacy guarantees. They assess only the specified distance-based attack and task representation. Artifact-level and ambient-space attacks are reported separately as sensitivity analyses; passing the task-space criteria does not preclude detectable membership signal in another representation.

Eligibility and selection: An arm is *task-space eligible* when (i) its 90% bootstrap interval for Δ lies within $(-0.05, 0.05)$, (ii) its attack AUC and copy rate satisfy the thresholds above, and (iii) its substitution AUC exceeds both zero-signal controls. This is a case-study designation under the evaluated threat model, not evidence of representation-robust privacy or release safety. Guidance and feature representation were fixed by nested leave-one-split-out selection before the frozen evaluation. Among eligible arms, the prespecified tie-breaker selects the lowest observed attack AUC (Table 2). Because between-arm attack-AUC differences may be small and are not tested confirmatorily, the selected arm is a representative candidate rather than a uniquely optimal configuration.

Multiplicity and scope: The confirmatory equivalence family comprises the six arms in Table 2 within each cohort. Appendix G applies a one-sided Bonferroni level of $0.05/6$ to this family only. The adjustment does not validate the membership-risk thresholds, attribution rule, selection criterion, secondary classifiers, correction experiments, or other diagnostic and sensitivity analyses.

Table 2: Matched-scale comparison using pooled out-of-fold predictions. Bold rows satisfy the prespecified task-space criteria; stars mark representative candidates selected by the deterministic tie-breaker, not statistically superior configurations.

Cohort	Arm	$\mathrm{AUC}_{\mathrm{syn}}$	Δ	90% CI	TOST	Attack	Copy
FTIR	real-trained	0.817	–	–	–	–	–
FTIR	Spectral-LDM only	0.780	-0.037	$[-0.060, -0.015]$	no	0.522	0.117
FTIR	WGAN-GP	0.726	-0.091	$[-0.122, -0.060]$	no	0.508	0.000
FTIR	PCA-Gaussian	0.785	-0.032	$[-0.061, -0.005]$	no	0.531	0.015
FTIR	**25% PCA mix***	**0.811**	$\mathbf{-0.006}$	$[-0.028, +0.015]$	yes	0.523	0.094
FTIR	**50% PCA mix**	**0.802**	$\mathbf{-0.015}$	$[-0.039, +0.009]$	yes	0.527	0.064
FTIR	SMOTE (control)	0.840	$+0.023$	$[+0.010, +0.037]$	yes	1.000	0.956
LIBS	real-trained	0.890	–	–	–	–	–
LIBS	Spectral-LDM only	0.563	-0.327	$[-0.424, -0.234]$	no	0.512	0.002
LIBS	WGAN-GP	0.439	-0.452	$[-0.573, -0.330]$	no	0.512	0.000
LIBS	**PCA-Gaussian**	**0.904**	$\mathbf{+0.013}$	$[-0.015, +0.041]$	yes	0.502	0.014
LIBS	25% PCA mix	0.857	-0.033	$[-0.066, -0.003]$	no	0.507	0.005
LIBS	**50% PCA mix***	**0.884**	$\mathbf{-0.006}$	$[-0.031, +0.019]$	yes	0.476	0.008
LIBS	SMOTE (control)	0.897	$+0.007$	$[-0.020, +0.034]$	yes	0.992	0.390

7 Results

7.1 Implementation and transfer diagnostics

Correcting the FTIR sampler and increasing optimization from approximately 2×10^3 to 30,000 steps raised substitution AUC from 0.515 to 0.729 ($p = 10^{-4}$, paired over splits), reverse-transfer AUC to 0.93, and reduced the NN ratio from 9.8 to 0.89. Increasing guidance alone did not resolve the failure, and gains plateaued beyond $w = 3$ (Appendix E). Larger synthetic sets and pooling independently trained generators provided smaller additional gains (Appendix K). The corrected pipeline did not ensure stable transfer to LIBS. Diffusion predictions were inverted on three of five splits, whereas the linear PCA-Gaussian control achieved split-mean AUC 0.915, compared with 0.919 for real-trained LR4. Diagnostic analyses indicate weak conditioning rather than a label swap (Appendix D). Thus, LIBS exposed representation and conditioning failures not apparent on FTIR, despite passing the implementation checks in Appendix B.

7.2 Matched-scale generator comparison

Table 2 reports the six-arm comparison at $m = 20$ without post-hoc correction. Eligibility uses the equivalence, attribution, and task-representation membership-risk criteria of §6; comparisons among eligible arms are descriptive.

Individual generators: Neither nonlinear generator satisfies equivalence on either cohort. PCA-Gaussian performs similarly to Spectral-LDM on FTIR (AUC 0.785 versus 0.780; paired DeLong $p = 0.71$) and satisfies the unadjusted criteria on LIBS. However, LIBS PCA-Gaussian alone does not remain equivalent after adjustment of the six-arm equivalence family. These results do not support attributing transfer utility to nonlinear generative capacity alone.

Mixtures: Both FTIR mixtures satisfy the prespecified equivalence, attribution, and task-space membership-risk criteria, with $\Delta = -0.006$ and -0.015. On LIBS, the 50% mixture satisfies these criteria with $\Delta = -0.006$ and reduces split-level variability relative to either component. Both FTIR mixtures and the LIBS 50% mixture remain within the 0.05 equivalence margin after adjustment of the six-arm equivalence family (Appendix G). Diagnostic analyses indicate that this improvement is not reproduced by independently resampling diffusion alone and is consistent with complementary approximation errors (Appendix A.2).

Representative candidates: The prespecified tie-breaker selects the 25% FTIR mixture and the 50% LIBS mixture from the task-space-eligible arms. Their attack-AUC differences from alternative eligible arms are small and were not tested confirmatorily. Accordingly, these are representative candidates selected by the stated rule, not uniquely optimal mixture proportions.

7.3 Corrections fail generator-specific attribution

Covariance matching raises Gaussian noise to AUC 0.918 on LIBS and label-permuted Spectral-LDM samples to 0.797 on FTIR, although both zero-signal controls are near chance before cor-

rection. These controls generally improve at least as much as their corresponding generator arms; consequently, ten corrected LIBS arms satisfy equivalence and task-space membership-risk criteria but fail attribution. The corrections therefore insert substantial class-conditional information, but the analysis does not identify the separate contributions of the input marginal, learned generator structure, and inserted statistics. Detailed correction results are reported in Appendix A.4.

7.4 Privacy depends on the threat model

Positive control: SMOTE exceeds the real-trained FTIR AUC by 0.023 (DeLong $p = 0.006$) but permits near-perfect membership discrimination under the evaluated distance-based attack, with attack AUCs of 1.000 on FTIR and 0.992 on LIBS. It therefore demonstrates that predictive utility does not imply candidate-use suitability. Low-false-positive-rate results are reported in Appendix H.

Representation dependence: The representative LIBS candidate satisfies the prespecified four-dimensional LR4-space thresholds. In the full-spectrum representation, however, its assembled artifact has attack AUC 0.58 ($p = 0.004$), while the corresponding 50% mixture evaluated across pooled seeds reaches approximately 0.65; the signal is concentrated in training-dependent PCA-Gaussian rows (Appendix A.6). Although low-FPR performance is not distinguishable from the shuffled-label control, this does not negate discrimination over the full ROC curve. Because the full spectrum is available to an adversary, LIBS passes only the prespecified task-space gate and does not demonstrate representation-robust low membership risk.

Interpretation: Exploratory analyses show that improved fidelity can coincide with increased measured membership risk, whereas near-chance attack AUC can result from off-manifold generation with poor utility (Appendix A.7). The FTIR candidate satisfies the evaluated task-space criteria, for which the task and ambient representations coincide. Neither result establishes anonymity, differential privacy, or protection against untested attacks.

7.5 Calibration and secondary diagnostics

Equivalent ranking performance does not imply transferable operating thresholds. Synthetic-only threshold selection yields FTIR sensitivity 0.217, whereas calibration on 20 and 40 labelled real patients yields sensitivities of 0.717 and 0.765, respectively, compared with 0.753 for the real-trained classifier (Appendix J). Thus, labelled real data remain necessary for clinically meaningful threshold selection. Exploratory scarcity and optimization analyses show no consistent advantage for nonlinear generation and confirm that lower denoising loss need not improve held-out real-patient transfer. Detailed quantity, scarcity, optimization, masking, and secondary-classifier results are reported in Appendices K, L, and D.

8 Candidate Synthetic Artifacts

Two synthetic artifacts accompany the paper for reproducibility and methodological evaluation. The FTIR artifact satisfies the prespecified utility, attribution, and task-representation privacy criteria under the evaluated threat model. The LIBS artifact satisfies the prespecified LR4-space criteria but exhibits detectable membership signal under the post hoc full-spectrum attack (§7.4); it is therefore included to reproduce and study representation-dependent privacy conclusions, not as an artifact that has passed a representation-robust privacy assessment. The selected FTIR candidate contains 2,000 uncorrected spectra (825 healthy and 1,175 cancer; 235 second-derivative, L_2-normalized features) from the 25% PCA-Gaussian mixture. The LIBS reproducibility artifact contains 375 uncorrected spectra (175 negative and 200 positive; 24,488 linear, sum-normalized features) from the 50% PCA-Gaussian mixture. Both were assembled from samples pooled equally across the five splits and seeds using the prespecified selection rule (Table 2, $\star$). Each includes a manifest documenting its generation recipe, preprocessing, audit results, and row-level provenance (split, seed, and generator); neither contains real patient spectra, identifiers, or replicate acquisitions. These artifacts provide neither formal privacy guarantees nor evidence of clinical validity. In particular, the LIBS artifact should not be interpreted as suitable for unrestricted redistribution. Any use or redistribution should retain the synthetic-data designation, provenance manifest, intended-use restrictions, and disclosure that protection against attacks beyond the evaluated threat models has not been established. Neither artifact is intended for clinical decision-making, and clinically meaningful operating thresholds require calibration using labelled real data (§7.5).

9 Discussion and limitations

Central finding. The central failed-generator demonstration is established primarily in the FTIR case study: a failed and a corrected generator produced nearly identical mixed-training performance despite markedly different synthetic-to-real and reverse-transfer AUCs. Mixed real–synthetic training therefore evaluates the resulting classifier but can fail to identify generator quality or transferable class structure. The LIBS study complements this result by exposing failures in reconstruction, transformation, latent variability, and class conditioning, but does not independently replicate the FTIR failed-versus-corrected comparison.

Audit interpretation and novelty. The audit separates three failure modes: equivalence testing rejects inadequate real-patient transfer, empirical privacy gates reject SMOTE despite strong utility, and attribution controls reject corrected arms whose performance is reproduced by zero-signal inputs. Its individual components are established; the contribution is their joint use to show why mixed-training performance alone can mischaracterize generator utility. Attribution is a falsification check, not proof of causal origin: failure shows that utility need not depend on valid generator-derived class information, whereas passing does not uniquely identify the generator as its source.

Generator implications. The linear PCA-Gaussian control matches diffusion on FTIR and outperforms it on LIBS, indicating that downstream transfer should not be attributed to nonlinear capacity without a linear baseline. Mixtures reduce utility gaps and split variability relative to diffusion alone, although differences among audit-eligible arms are too small to establish a uniquely optimal recipe. These findings are best interpreted as two methodological case studies rather than evidence that a particular generator or mixture will generalize across clinical domains.

Limitations and scope. The audit is a structured falsification framework: failing a gate identifies a demonstrated problem, whereas passing establishes compatibility only with the evaluated classifiers, representations, attacks, thresholds, and datasets, leaving untested risks unresolved. The patient-bootstrap intervals are conditional on the frozen models, configurations, and repeated splits and do not capture uncertainty from pipeline selection or alternative experimental designs. Likewise, the 0.05 equivalence margin and privacy gates are operational thresholds rather than clinically or empirically validated risk tolerances; stricter equivalence margins change some conclusions. The audit was developed and evaluated on two small binary spectroscopy cohorts and was not independently validated or calibrated for release decisions beyond these case studies. The empirical privacy analysis provides neither differential privacy nor protection against untested attacks. In particular, LIBS passes the prespecified task-feature gate, whereas a post hoc ambient-space analysis detects membership signal, demonstrating dependence on the attack representation. Synthetic threshold selection also transfers poorly, so labelled real patients remain necessary to determine a clinically meaningful operating point, and unstable LIBS diffusion conditioning remains unexplained. Accordingly, the conclusions are limited to the inadequacy of mixed-training performance as a standalone generator evaluation and the value of the proposed tests as an auditable but incomplete framework; passing the audit does not establish anonymity, clinical validity, or unrestricted release safety.

10 Conclusion

In the FTIR case study, mixed real–synthetic training did not distinguish a near-chance generator from its corrected counterpart; synthetic-to-real and reverse transfer exposed the failure. Zero-signal and SMOTE controls further showed that predictive utility can arise from inserted class statistics or memorization, motivating separate evaluation of utility, attribution, and privacy. Privacy must also be assessed across plausible representations: the LIBS case shows that passing a prespecified task-space attack does not imply representation-robust privacy. The proposed audit is therefore a structured falsification framework that can expose unsupported utility, attribution, or privacy claims; passing its evaluated tests does not certify anonymity, clinical validity, or release safety.

REPRODUCIBILITY STATEMENT

Both cohorts are publicly available: the FTIR deposit (Paraskevaidi et al., 2020) and the LIBS deposit (Melikechi et al., 2024). The supplementary `code.zip` archive contains cohort preparation, fail-fast gates, generator training and sampling, evaluation and privacy analyses, frozen patient-level splits and seeds, configuration files, and table-generation scripts. The original sampler is retained as a legacy mode. The supplementary `data.zip` archive contains the per-split evaluation outputs, synthetic cohorts, and their manifests. The sampler analysis in Appendix C and the transform round-trip check in Appendix B each run on a CPU in under one minute. Neither archive contains real patient spectra.

ETHICS STATEMENT

Nothing here supports clinical use; an equivalent classifier is not a validated diagnostic. Likewise, passing the empirical audit should not be interpreted as regulatory, institutional, or ethical authorization for unrestricted data release. This study is a secondary analysis of two previously published, de-identified datasets. Recruitment, consent, and ethical approval were obtained by the original investigators and are documented in the source publications; no new human-subject data were collected. The synthetic spectra must remain identified as synthetic and must not enter clinical records or reference spectral libraries. Any redistribution should include the accompanying manifest. Because generative models can reproduce limitations of their training cohorts, we make no claim of representativeness beyond the source populations.

THE USE OF LARGE LANGUAGE MODELS

Generative AI tools assisted with feedback on evaluation methodology, implementation and review of analysis code, literature search, interpretation discussions, and manuscript drafting and editing. The authors independently verified cited sources, code, analyses, and text and take responsibility for the submission. No dataset, synthetic spectrum, experimental result, numerical value, table, or figure was generated or altered by a language model.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 214–223, 2017.

Matthew J. Baker, Júlio Trevisan, Paul Bassan, Rohit Bhargava, Holly J. Butler, Konrad M. Dorling, Peter R. Fielden, Simon W. Fogarty, Nigel J. Fullwood, Kelly A. Heys, Caryn Hughes, Peter Lasch, Pierre L. Martin-Hirsch, Blessing Obinaju, Ganesh D. Sockalingum, Josep Sulé-Suso, Rebecca J. Strong, Michael J. Walsh, Bayden R. Wood, Peter Gardner, and Francis L. Martin. Using Fourier transform IR spectroscopy to analyze biological materials. *Nature Protocols*, 9(8): 1771–1791, 2014. doi: 10.1038/nprot.2014.110.

Holly J. Butler, Paul M. Brennan, James M. Cameron, Duncan Finlayson, Mark G. Hegarty, Michael D. Jenkinson, David S. Palmer, Benjamin R. Smith, and Matthew J. Baker. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nature Communications*, 10:4501, 2019. doi: 10.1038/s41467-019-12527-5.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2022. doi: 10.1109/SP46214.2022.9833649.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Jihoon Chung, Junru Zhang, Amirul Islam Saimon, Yang Liu, Blake N Johnson, and Zhenyu Kong. Imbalanced spectral data analysis using data augmentation based on the generative adversarial network. *Scientific Reports*, 14(1):13230, 2024. doi: 10.1038/s41598-024-63285-4.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595.

Khaled El Emam, Lucy Mosquera, and Jason Bass. Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *Journal of Medical Internet Research*, 22(11):e23139, 2020. doi: 10.2196/23139.

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv preprint arXiv:1706.02633*, 2017.

Ahmed Fadlelmoula, Susana O. Catarino, Graça Minas, and Vitor Carvalho. A review of machine learning methods recently applied to FTIR spectroscopy data for the analysis of human blood cells. *Micromachines*, 14(6):1145, 2023. doi: 10.3390/mi14061145.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

Aaron R Flanagan and Frank G Glavin. A comparative analysis of data synthesis techniques to improve classification accuracy of Raman spectroscopy data. *Journal of Chemical Information and Modeling*, 64(7):2311–2322, 2024. doi: 10.1021/acs.jcim.3c00761.

Ketan Gajjar, Júlio Trevisan, Gemma Owens, Patrick J. Keating, Nicholas J. Wood, Helen F. Stringfellow, Pierre L. Martin-Hirsch, and Francis L. Martin. Fourier-transform infrared spectroscopy coupled with a classification machine for the analysis of blood plasma or serum: a novel diagnostic approach for ovarian cancer. *Analyst*, 138(14):3917–3926, 2013. doi: 10.1039/C3AN36654E.

Juan José García-Esteban, Juan Carlos Cuevas, and Jorge Bravo-Abad. Generative adversarial networks for data-scarce spectral applications. *arXiv preprint arXiv:2307.07454*, 2023.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS 2017)*, 2017.

Kehan Guo, Yili Shen, Gisela Abigail Gonzalez-Montiel, Yue Huang, Yujun Zhou, Mihir Surve, Zhichun Guo, Prayel Das, Nitesh V Chawla, Olaf Wiest, et al. Artificial intelligence in spectroscopy: Advancing chemistry from prediction to generation and beyond. *arXiv preprint arXiv:2502.09897*, 2025.

Shanshan Guo, Gongxiang Wei, Wenqiang Chen, Chengbin Lei, Cong Xu, Yu Guan, Te Ji, Fuli Wang, and Huiqiang Liu. Fast and deep diagnosis using blood-based ATR-FTIR spectroscopy for digestive tract cancers. *Biomolecules*, 12(12):1815, 2022. doi: 10.3390/biom12121815.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the 2008 International Joint Conference on Neural Networks (IJCNN)*, pp. 1322–1328, 2008.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. doi: 10.48550/arXiv.2207.12598. arXiv:2207.12598.

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

Marinus Huber, Kosmas V. Kepesidis, Liudmila Voronina, Frank Fleischmann, Ernst Fill, Jacqueline Hermann, Ina Koch, Katrin Milger-Kneidinger, Thomas Kolben, Gerald B. Schulz, Friedrich Jokisch, Juergen Behr, Nadia Harbeck, Maximilian Reiser, Christian Stief, Ferenc Krausz, and Mihaela Zigman. Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *eLife*, 10:e68758, 2021. doi: 10.7554/eLife.68758.

James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how? Technical report, The Royal Society and The Alan Turing Institute, 2022.

R. W. Kennard and L. A. Stone. Computer aided design of experiments. *Technometrics*, 11(1): 137–148, 1969. doi: 10.1080/00401706.1969.10490666.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2021.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. doi: 10.1177/1948550617697177.

Adriana Martinez-Cuazitl, Gustavo J. Vazquez-Zapien, Miguel Sanchez-Brito, Jorge H. Limon-Pacheco, Melissa Guerrero-Ruiz, Francisco Garibay-Gonzalez, Raul J. Delgado-Macuil, Maria G. Garcia de Jesus, Miguel A. Corona-Perezgrovas, Armando Pereyra-Talamantes, and Monica M. Mata-Miranda. ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients. *Scientific Reports*, 11:19980, 2021. doi: 10.1038/s41598-021-99529-w.

Rose G. McHardy, Georgios Antoniou, Justin J. A. Conn, Matthew J. Baker, and David S. Palmer. Augmentation of FTIR spectral datasets using Wasserstein generative adversarial networks for cancer liquid biopsies. *Analyst*, 148(16):3860–3869, 2023. doi: 10.1039/d3an00669g.

Noureddine Melikechi, Helmar G. Adler, Ali Safi, Joshua E. Landis, Farhad Pourkamali-Anaraki, Kemal Efe Eseller, Kim Berlo, Danielle Bonito, Gregory R. Chiklis, and Weiming Xia. Blood metal analysis of plasmas from donors with and without sars-cov-2 using laser-induced breakdown spectroscopy and logistic regression. *Biomedical Optics Express*, 15(1):446–459, 2024. doi: 10.1364/BOE.513558.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171, 2021.

Marcelo Saito Nogueira, Leonardo Barbosa Leal, Wena Dantas Marcarini, Raquel Lemos Pimentel, Matheus Muller, Paula Frizera Vassallo, Luciene Cristina Gastalho Campos, Leonardo dos Santos, Wilson Barros Luiz, José Geraldo Mill, Valerio Garrone Barauna, and Luis Felipe das Chagas e Silva de Carvalho. Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning. *Scientific Reports*, 11:15409, 2021. doi: 10.1038/s41598-021-93511-2.

Lyra O'Leary, Dougal Ferguson, Claire A. Hart, Mick Brown, Pedro Oliveira, Noel Clarke, Ashwin Sachdeva, Peter Gardner, and Hujun Yin. Spatial-spectral deep learning for prostate cancer tissue classification in infrared spectroscopy. *Analytical Chemistry*, 98(4):2743–2755, 2026. doi: 10.1021/acs.analchem.5c04765.

Maria Paraskevaidi, Camilo L. M. Morais, Katherine M. Ashton, Helen F. Stringfellow, Rhona J. McVey, Neil A. J. Ryan, Helena O'Flynn, Vanitha N. Sivalingam, Sarah J. Kitson, Michelle L. MacKintosh, Abigail E. Derbyshire, Cecilia Pow, Olivia Raglan, Kássio M. G. Lima, Maria Kyrgiou, Pierre L. Martin-Hirsch, Francis L. Martin, and Emma J. Crosbie. Detecting endometrial cancer by blood spectroscopy: A diagnostic cross-sectional study. *Cancers*, 12(5):1256, May 2020. ISSN 2072-6694. doi: 10.3390/cancers12051256. URL http://dx.doi.org/10.3390/cancers12051256.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018. doi: 10.14778/3231751.3231757.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868, 2021. URL https://proceedings.mlr.press/v139/rasul21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28.

Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964. doi: 10.1021/ac60214a047.

Roberta Schiemer, Jessica Grant, Mohamad N. Shafiee, Sendy Phang, David Furniss, Radu Boitor, Angela B. Seddon, Ioan Notingher, William Atiomo, Nia W. Jones, and Ketankumar B. Gajjar. Infrared and raman spectroscopy of blood plasma for rapid endometrial cancer detection. *British Journal of Cancer*, 133(2):194–207, 2025. doi: 10.1038/s41416-025-03050-0.

Donald J. Schuirmann. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. doi: 10.1007/BF01068419.

Nadir Sella, Florent Guinot, Nikita Lagrange, Laurent-Philippe Albou, Jonathan Desponds, and Hervé Isambert. Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. *npj Digital Medicine*, 8(1):49, 2025. doi: 10.1038/s41746-025-01431-6.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, 2017.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468. USENIX Association, 2022.

Kar-Yan Su and Wai-Leng Lee. Fourier transform infrared spectroscopy as a cancer screening and diagnostic tool: A review and prospects. *Cancers*, 12(1):115, 2020. doi: 10.3390/cancers12010115.

Jiayi Tang, Alex Henderson, and Peter Gardner. Exploring AdaBoost and random forests machine learning approaches for infrared pathology on unbalanced data sets. *Analyst*, 146(19):5880–5891, 2021. doi: 10.1039/D0AN02155E.

Rock Christian Tomas, Anthony Jay Sayat, Andrea Nicole Atienza, Jannah Lianne Danganan, Ma. Rollene Ramos, Allan Fellizar, Kin Israel Notarte, Lara Mae Angeles, Ruth Bangaoil, Abegail Santillan, and Pia Marie Albano. Detection of breast cancer by ATR-FTIR spectroscopy using artificial neural networks. *PLOS ONE*, 17(1):e0262489, 2022. doi: 10.1371/journal.pone.0262489.

Boris van Breugel, Zhaozhi Qian, and Mihaela van der Schaar. Synthetic data, real errors: how (not) to publish and use synthetic data. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020. doi: 10.1016/j.neucom.2019.12.136.

Haiming Yao, Wei Luo, Ang Gao, Tao Zhou, and Xue Wang. DiffRaman: A Conditional Latent Denoising Diffusion Probabilistic Model for Bacterial Raman Spectroscopy Identification Under Limited Data Conditions. *arXiv preprint arXiv:2412.08131*, 2024.

Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020. doi: 10.1109/JBHI.2020.2980262.

A Additional Results and Diagnostic Analyses

Unless otherwise stated, the analyses in this section are diagnostic or exploratory and are not members of the six-arm confirmatory equivalence family. They do not alter the eligibility criteria in §6.

A.1 Synthetic-set size and ensembling

For FTIR diffusion at $w = 5$, increasing the synthetic-set multiplier from $m = 1$ to $m = 16$ improves substitution AUC by 0.029 ($p = 0.008$, paired over splits). Pooling samples from five independently trained generators instead of one improves AUC by 0.025–0.043 across the evaluated multipliers. Under the configuration frozen before the six-arm evaluation ($w = 3$, $m = 20$), pooled diffusion achieves split-mean AUC 0.778, compared with 0.765 for PCA-Gaussian. These analyses motivated the use of pooled generators but are diagnostic because the quantity grid was not part of the confirmatory equivalence family.

Table 3: Quantity and ensembling, FTIR XGB, split-mean ± split s.d. of substitution AUC. Left: w=5, one generator ("single") or five seed-trained generators pooled ("ensemble"), validation fold carved from the synthetic training set. Right: five pooled generators at $m \in \{16, 20\}$ over the guidance sweep (the release recipe is w=3, m=20). Real-trained reference under the same validation protocol: 0.781 ± 0.029.

m	single	ensemble	w	m=16	m=20
1	0.712±0.024	0.737±0.048	1	0.733±0.038	0.729±0.033
2	0.715±0.028	0.744±0.036	1.5	0.745±0.029	0.751±0.028
4	0.718±0.032	0.762±0.023	2	0.764±0.032	0.770±0.025
8	0.729±0.028	0.763±0.027	3	0.768±0.035	0.778±0.034
16	0.730±0.033	0.767±0.035	4	0.777±0.019	0.777±0.022
			5	0.774±0.034	0.774±0.037

A.2 Mixture complementarity

Mixtures reduce split-to-split variation relative to diffusion alone. On FTIR, the split-level standard deviation decreases from 0.034 for Spectral-LDM to 0.014 and 0.011 for the 25% and 50% PCA-Gaussian mixtures, respectively. On LIBS, the 50% mixture scores 0.89–0.92 across all five splits and reduces the standard deviation from 0.288 for Spectral-LDM and 0.018 for PCA-Gaussian to 0.010.

Increasing diffusion diversity without introducing a second generator family does not reproduce the FTIR mixture result: independently resampling diffusion for each classifier seed changes Δ from -0.037 to -0.042, whereas adding 25% PCA-Gaussian samples yields $\Delta = -0.006$. A diagnostic real-versus-synthetic discriminator distinguishes Spectral-LDM and PCA-Gaussian samples from real spectra with AUCs of 0.998 and 0.882, respectively, compared with 0.933 and 0.904 for the 25% and 50% mixtures. Together with the nearest-neighbour diagnostics, these results are consistent with complementary approximation errors: Spectral-LDM produces patient-scale distances but incompletely represents within-class fine structure, whereas PCA-Gaussian captures covariance within a training-dependent linear subspace. This interpretation does not establish that either component is individually sufficient or that a particular mixture is uniquely optimal.

A.3 Visualization of the six-arm comparison

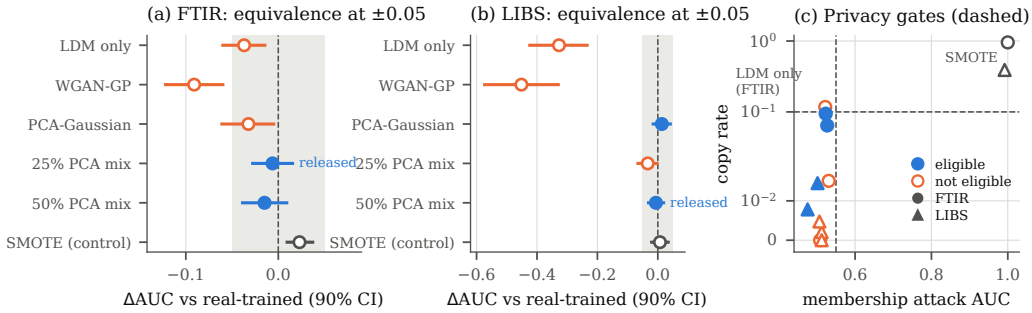


Figure 2: Six-arm equivalence and task-space membership-risk results. Filled markers denote arms satisfying the prespecified task-space criteria, and stars denote representative candidates selected by the deterministic tie-breaker. Neither notation implies statistically supported superiority over another eligible arm.

A.4 Correction and attribution diagnostics

Before correction, label-permuted Spectral-LDM samples and Gaussian noise produce substitution AUCs of 0.49–0.53. Moment and covariance matching raise these zero-signal controls by 0.24–0.43 AUC and generally improve them at least as much as the corresponding generator arms. The sole exception is the LIBS XGBoost diffusion comparison, whose uncorrected AUC is 0.419.

Corrections can also improve genuine generator arms. On FTIR, covariance matching raises one Spectral-LDM arm to AUC 0.815 while increasing its copy rate from 0.117 to 0.431. On LIBS, covariance matching changes the WGAN-GP utility difference from $\Delta = -0.452$ to $\Delta = 0.000$, but corrected zero-signal inputs reproduce comparable utility. In total, ten corrected LIBS arms satisfy equivalence and the task-space membership-risk criteria but fail the point-estimate attribution rule.

These controls falsify generator-specific attribution for the evaluated correction pipeline; they do not prove that the generators contribute no useful structure. In particular, they do not decompose utility among the input marginal, learned generator structure, and class statistics inserted by the correction. All correction and secondary-classifier comparisons are diagnostic.

A.5 Low false-positive-rate behavior

At 1% false-positive rate, the distance-based attack identifies 99.95% of FTIR members and 85.5% of LIBS members for SMOTE, confirming its role as a positive control. In the prespecified task representations, the representative FTIR and LIBS candidates yield lifts of 0.93 and 1.95 over chance, respectively, compared with 1.28 and 1.85 under shuffled membership labels. These results indicate limited low-FPR separation under the task-space attack, but do not override signal detected in another representation.

Table 4: Membership inference at low false-positive rate. Distance-to-closest-record attack in the audit space of Table 2, pooled over five patient-level splits × five seeds (FTIR: 10,200 member and 4,400 non-member observations; LIBS: 1,950 and 850), true-positive rate at the largest achievable FPR at or below the target, 90% patient-level cluster-bootstrap CI (2,000 resamples). Lift is TPR/FPR; the last column is the lift of the same attack after membership labels are shuffled (negative control). Per split, 1% FPR corresponds to 8.8 (FTIR) and 1.7 (LIBS) non-members, which is why the figures are pooled. Attack AUCs here pool five seeds per split, so they differ from the single-seed per-split means of Table 2 (released LIBS recipe: 0.505 against 0.476); the released files themselves give 0.520 (FTIR) and 0.516 (LIBS, LR4 space; Appendix I). Released recipes marked †.

Cohort	Arm	Attack AUC	TPR@1%FPR	90% CI	Lift	TPR@5%FPR	Shuffled lift@1%
FTIR	Spectral-LDM only	0.522	0.0093	[0.0039, 0.0186]	0.93	0.0461	1.21
FTIR	WGAN-GP	0.504	0.0103	[0.0081, 0.0131]	1.03	0.0582	0.99
FTIR	PCA-Gaussian	0.521	0.0093	[0.0060, 0.0141]	0.93	0.0490	1.17
FTIR	25% PCA mix†	0.519	0.0093	[0.0042, 0.0174]	0.93	0.0507	1.28
FTIR	50% PCA mix	0.520	0.0085	[0.0049, 0.0210]	0.85	0.0563	1.29
FTIR	SMOTE (control)	1.000	0.9995	[0.9990, 0.9999]	99.95	0.9997	1.03
LIBS	Spectral-LDM only	0.515	0.0077	[0.0025, 0.0279]	0.77	0.0462	1.03
LIBS	WGAN-GP	0.504	0.0128	[0.0077, 0.0287]	1.28	0.0513	1.08
LIBS	PCA-Gaussian	0.506	0.0092	[0.0051, 0.0152]	0.92	0.0497	0.87
LIBS	25% PCA mix	0.504	0.0231	[0.0106, 0.0305]	2.31	0.0559	1.03
LIBS	50% PCA mix†	0.505	0.0195	[0.0125, 0.0277]	1.95	0.0564	1.85
LIBS	SMOTE (control)	0.987	0.8549	[0.8231, 0.9140]	85.49	0.9333	0.97

A.6 Representation sensitivity

The prespecified LIBS audit operates in the four-dimensional LR4 feature space. In that representation, the assembled candidate artifact gives five-split mean attack AUC 0.516, below the permutation-null 95th percentile of 0.54 ($p = 0.27$). Across the full 24,488-channel spectrum, however, the same artifact gives split AUCs of 0.50, 0.58, 0.54, 0.62, and 0.69, for a mean of 0.584 against a null 95th percentile of 0.54 ($p = 0.004$).

When pooled over splits, the full-spectrum true-positive rate is 1.0 times chance at 1% FPR and 2.1 times chance at 5% FPR, compared with 2.3 and 1.0 for shuffled membership labels. Thus, low-FPR performance is not distinguishable from the negative control, whereas discrimination over the full ROC curve is detectable. These findings are complementary rather than contradictory because they evaluate different operating regions.

The full-spectrum signal is concentrated in PCA-Gaussian rows. Across five pooled seeds, the full-spectrum attack AUC is 0.68 for PCA-Gaussian, 0.51 for diffusion, and approximately 0.65 for the 50% mixture. The PCA-Gaussian samples lie in the training-fold subspace in their fitting domain, with relative residual at most 2×10^{-7}, compared with 0.020–0.040 for held-out donors. An ambient-space distance attack can therefore detect training-dependent subspace geometry that is absent from the four LR4 features.

Because the candidate artifact exposes the complete spectrum, the ambient space is a plausible adversarial representation and receives equal interpretive weight. The LIBS artifact satisfies the prespecified LR4-space criteria but does not establish representation-robust low membership risk. These analyses are post-hoc privacy sensitivities and do not constitute a formal privacy assessment.

For FTIR, the classifier and ambient representations coincide. The assembled FTIR candidate gives attack AUC 0.520, while the 20-component PCA-Gaussian arm also gives 0.52. These results remain conditional on the evaluated distance-based attack and do not address stronger adversary families.

A.7 Fidelity and geometric diagnostics

Retraining the FTIR autoencoder with reflection padding, a longer schedule, and an edge-weighted loss improves all evaluated fidelity measures: the round-trip Kolmogorov–Smirnov distance decreases from 0.23 to 0.12, the edge artifact score decreases from 0.39 to 0.12, and the variance ratio increases from 0.80 to 0.89. Substitution changes little ($\Delta = -0.029$ to $\Delta = -0.033$), but attack AUC increases from 0.551 to 0.653. Copy rate decreases from 0.117 to 0.024, and the NN ratio increases from 0.94 to 1.18. This exploratory experiment associates improved fidelity with increased

measured membership risk for this autoencoder; it does not establish a universal fidelity–privacy trade-off.

Adding 50% PCA-Gaussian samples to the FTIR diffusion set reduces copy rate from 0.117 to 0.064 and the utility gap from 0.037 to 0.015, while task-space attack AUC remains between 0.52 and 0.53. Conversely, WGAN-GP produces near-chance attack AUC but NN ratios of 5.6 on FTIR and 20 on LIBS, together with poor substitution utility. Near-chance membership discrimination can therefore result from off-manifold generation and is not, by itself, evidence of a favorable privacy–utility profile. These fidelity and geometric analyses are exploratory.

A.8 Scarcity analysis

Table 16 compares real-only classifiers trained on a fraction f of the available patients with classifiers whose class deficit is filled using Spectral-LDM, PCA-Gaussian, or SMOTE. At $f = 0.10$, FTIR AUC improves by 0.153 with SMOTE ($p = 0.001$), 0.075 with PCA-Gaussian ($p = 0.010$), and 0.040 with Spectral-LDM ($p = 0.24$). On LIBS, the diffusion fill is unstable at the same fraction ($\Delta\mathrm{AUC} = -0.080$, $p = 0.24$). Linear and diffusion fillers do not differ significantly at any evaluated fraction ($p > 0.05$). These comparisons are exploratory and do not alter candidate eligibility.

Table 5: Exploratory scarcity analysis. Entries are AUC gains over real-only training on the same fraction f of patients after filling the class deficit. Values are means over five splits and five seeds.

	FTIR			LIBS		
f	Spectral-LDM	PCA-Gaussian	SMOTE	Spectral-LDM	PCA-Gaussian	SMOTE
0.10	$+0.040$	$+0.075$	$+0.153$	-0.080	$+0.052$	$+0.055$
0.25	$+0.042$	$+0.060$	$+0.111$	$+0.015$	$+0.014$	$+0.017$
0.50	$+0.025$	$+0.002$	$+0.051$	$+0.013$	$+0.014$	$+0.011$
0.75	$+0.015$	$+0.018$	$+0.035$	$+0.004$	$+0.002$	$+0.003$

A.9 Optimization and negative results

Extended optimization. Continuing FTIR denoiser training from 30,000 to a median of 68,000 steps, with early stopping on a training-time proxy, changes the utility difference from -0.019 to -0.050 (DeLong $p = 0.006$), despite lower denoising loss and a higher internal proxy. Better optimization objectives therefore do not necessarily imply improved transfer to held-out real patients.

Feature masking. Dropping the 10, 20, 40, or 60 features most predictive of real-versus-synthetic status, trimming spectral edges, or applying both changes FTIR substitution by at most 0.006 relative to the unmasked result (0.772–0.779 versus 0.778). This analysis does not support a small set of discriminator-selected features as the primary cause of the remaining transfer gap.

Secondary classifiers. On raw FTIR diffusion samples, full-spectrum logistic regression loses 0.222 AUC relative to its real-trained counterpart (0.626 versus 0.849), logistic regression on 20 principal components loses 0.087 (0.716 versus 0.803), and XGBoost loses 0.017 (0.778 versus 0.796). These secondary-classifier results are diagnostic and show that transfer depends on the downstream inductive bias.

Guidance and over-separation. Increasing w beyond 3 does not improve FTIR substitution, although reverse-transfer AUC continues to rise and synthetic class separation exceeds that of the real data. Conversely, $w = 1$, which gives the most real-like class-separation ratio, has the weakest substitution AUC. Over-separation alone therefore does not explain the residual transfer gap; the complete guidance grid appears in Appendix E.

LIBS inversion. The three inverted LIBS splits are not consistent with a simple label swap. Real-trained LR4 scores the diffusion pools near chance, synthetic class differences in the four elemental features are only 0.2–0.6 of the corresponding standardized real-data differences, and seed pools on the same split vary from 0.13 to 0.86 in substitution AUC. Moreover, the 25% mixture, which

remains 75% diffusion, is not correspondingly inverted. These diagnostics instead indicate weak and unstable class conditioning. This interpretation remains exploratory because the underlying cause is not identified.

B FAIL-FAST PIPELINE GATES

The LIBS analysis of §7.1 yielded three gates, ordered by cost, that are now applied before generator training. Each retrospectively identified an implementation failure. These gate-development analyses are diagnostic and are not part of the six-arm confirmatory equivalence family.

Gate 1: transform round-trip with an identity autoencoder (seconds, no model). Replace the autoencoder by the identity and pass a real spectrum through the forward transform to the model domain and its inverse to the classifier domain. Relative L_2 error must be below 10^{-2}. Before correction, the log-domain transform failed this test (0.94 over the full spectrum) because per-spectrum L_2 normalization was not inverted and $\exp$ is not homogeneous; all six domain×stride selection cells then tied at the score of a uniform spectrum (0.9594). After correction, the error is 2.6×10^{-4} in the linear, square-root, and log domains.

Gate 2: autoencoder round-trip in both domains. We measure relative L_2 reconstruction error in the model domain and, separately, after inverse transformation to the classifier domain, with a model-domain threshold of 0.30. The first LIBS autoencoder scored 0.95 (FTIR: 0.076); the rebuilt model scored 0.15 in its own domain but 0.96 in the classifier domain, motivating Gate 1. The final model scored 0.026–0.040 and 0.003–0.005, respectively, across the five splits.

Gate 3: latent spread. We measure the between-spectrum standard deviation of the encoded training set relative to unit variance, with threshold 0.1. In the uncentred, per-spectrum-normalized log domain, between-donor variation was 0.6% of the signal (mean pairwise cosine 0.994), diffusion loss was 0.0006 against unit-variance noise, and the conditioning signal decreased from 0.007 to 0.003; every stride-5 autoencoder had zero spread. With global centring and a stored scale, spread is 0.11–0.44 across splits and diffusion loss at 30,000 steps is 0.018–0.028. Domain and stride are selected using training-internal validation only. The selected LIBS configuration is the linear domain at stride 4; its first diffusion model achieved LR4 substitution AUC 0.816 on the pilot split, with NN ratio 1.00 and copy rate 0.01. This pilot result is diagnostic rather than confirmatory.

C SAMPLER ANALYSIS WITH AN ANALYTICALLY OPTIMAL DENOISER

For a latent coordinate $x_0 \sim \mathcal{N}(\mu, \sigma^2)$ under the forward process $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, the optimal noise predictor is $\mathbb{E}[\epsilon \mid x_t] = \sqrt{1-\bar{\alpha}_t}\,(x_t - \sqrt{\bar{\alpha}_t}\mu)/(\bar{\alpha}_t\sigma^2 + 1 - \bar{\alpha}_t)$. Because this expression is affine in x_t, sample moments after any sequence of affine updates can be propagated exactly. Table 6 reports exact values and a Monte Carlo run (2×10^5 samples, $x_T \sim \mathcal{N}(0,1)$) for the implemented cosine schedule ($s = 0.008$, $T = 1000$, and β_t clamped to $[10^{-8}, 0.999]$), with $\mu = 2.00$ and $\sigma = 0.50$.

The invalid loop visited `linspace(999,0,300)` and applied at each visited t the one-step update $x_{t-1} = \alpha_t^{-1/2}(x_t - \beta_t(1-\bar{\alpha}_t)^{-1/2}\hat{\epsilon}) + \sqrt{\beta_t}z$, retaining full-schedule coefficients despite gaps of three or four indices. The stride-correct DDIM $\eta = 1$ update instead uses $\bar{\alpha}$ at consecutive visited indices. Under a linear schedule (β from 10^{-4} to 0.02), the corresponding results remain 1.19/0.72 for the invalid strided update and mean 2.00 for DDIM at 300 and 50 steps. Applying the ancestral update at all 1,000 indices recovers the target mean, showing that the failure arises from striding rather than the one-step update itself. Monte Carlo estimates vary in the third decimal; the supplementary script reports 1.187 and 0.713.

With classifier-free guidance $\hat{\epsilon} = \hat{\epsilon}_\varnothing + w(\hat{\epsilon}_y - \hat{\epsilon}_\varnothing)$ and an unconditional model $\mathcal{N}(0, \sigma^2)$, the guided optimum corresponds to $\mathcal{N}(w\mu, \sigma^2)$. The invalid strided sampler recovers 59.4% of this target at every w; at $w = 0.5$, the recovered mean is 0.59, or 30% of μ. DDIM recovers the target mean but not the variance exactly at coarse strides (0.49 at 300 steps and 0.47 at 50 steps, versus 0.50), because it conditions on the point estimate $\hat{x}_0$. We therefore claim recovery of the class mean, not the complete distribution. This sampler analysis is a diagnostic implementation ablation, not part of the confirmatory family.

Table 6: Samples from the analytically optimal denoiser, $\mu = 2.00$, $\sigma = 0.50$, cosine schedule $T = 1000$. Exact moment propagation and Monte Carlo (2×10^5 samples).

Sampler	Steps	Mean (exact)	S.d. (exact)	Mean (MC)	% of μ
Ancestral, full schedule (reference)	1000	2.000	0.500	2.000	100.0
Ancestral, strided (invalid)	300	1.189	0.716	1.188	59.4
Ancestral, strided	500	1.574	0.617	1.577	78.7
Ancestral, strided	50	0.276	0.937	0.278	13.8
DDIM $\eta = 1$, stride-correct	300	2.000	0.494	2.000	100.0
DDIM $\eta = 1$, stride-correct	50	2.000	0.467	2.000	100.0

D Further diagnostic results

All analyses in this section are diagnostic or exploratory and are outside the six-arm confirmatory equivalence family.

Masking is ineffective. Dropping the top 10, 20, 40, or 60 features of the real-versus-synthetic discriminator, trimming spectral edges, or both, leaves FTIR substitution within 0.006 of the unmasked value (0.772–0.779 versus 0.778; five splits $\times$ five seeds, $w = 3$, $m = 20$).

Linear classifiers transfer worse than trees on raw FTIR diffusion samples. Full-spectrum logistic regression trained on synthetic data loses 0.222 AUC relative to its real-trained counterpart (0.626 versus 0.849), logistic regression on 20 principal components loses 0.087 (0.716 versus 0.803), and XGBoost loses 0.017 (0.778 versus 0.796). These secondary-classifier results are diagnostic. Appendix E indicates that over-sharpening does not explain the residual gap.

The LIBS inversion is not explained by a label swap. On every split, a real-trained LR4 scores the diffusion pools near chance (reverse AUC 0.44–0.59, averaged over five seed pools). Their class differences in the four elemental features are 0.2–0.6 of the real fold's standardized differences, and seed pools on the same inverted split range from 0.13 to 0.86 in substitution AUC. Training and sampling use the same integer labels and split files, without a label encoder, sorting step, or split-specific path. A label swap would also invert the 25% mixture, which contains 75% diffusion rows but scores 0.83–0.91 on the three inverted splits. These observations do not identify the cause, but they support weak and unstable conditioning rather than a deterministic label reversal.

E Diagnostic guidance analysis

The guidance, geometry, and optimization analyses below are diagnostic and are not included in the six-arm confirmatory equivalence family.

Table 7: Diagnostic FTIR guidance sweep with corrected sampler and training budget, one generator ($m = 1$), and XGBoost; mean over five splits and five seeds. Copy rate is the fraction of synthetic rows below the fifth percentile of real-to-real nearest-neighbour distance. The invalid strided sampler at $w = 0.5$ gave substitution AUC 0.515, reverse AUC 0.51, and NN ratio 9.8.

w	Substitution AUC	Reverse AUC	NN ratio	Copy rate	Δ vs real
1	0.649	0.716	0.830	0.199	-0.147
2	0.708	0.860	0.856	0.162	-0.088
3	0.729	0.931	0.893	0.129	-0.066
4	0.729	0.964	0.947	0.093	-0.067
5	0.727	0.979	1.014	0.069	-0.069

At $m = 20$ with five pooled generators, the $w \times m$ grid gives split-mean substitution AUCs of 0.729, 0.751, 0.770, 0.778, 0.777, and 0.774 at $w = 1, 1.5, 2, 3, 4, 5$. Nested leave-one-split-out selection chose $w = 3$ on three splits and $w = 4$ on two. The value $w = 3$, prespecified before the confirmatory runs, and the nested diagnostic choice give out-of-fold differences of -0.037 and -0.043, respectively; neither satisfies equivalence. Per-split membership-attack AUC for diffusion-only samples is approximately 0.55 across w, while NN ratio increases from 0.86 to 1.10 and copy

rate decreases from 0.19 to 0.06 between $w = 1$ and $w = 5$. Over-sharpening does not explain the residual transfer gap: the most real-like class geometry occurs at $w = 1$ (synthetic-to-real class-separation ratio 0.88), which has the lowest substitution AUC, whereas ratios of 2.5–4.0 at $w = 3$–5 coincide with the highest AUCs.

F CONFIRMATORY AND DIAGNOSTIC RESULT GRID

Table 8 reports every evaluated arm by correction and classifier on both cohorts as split means with split standard deviations. Only the uncorrected six-arm comparison using each cohort's primary classifier constitutes the confirmatory equivalence family in Table 2. Secondary classifiers and all correction experiments are diagnostic or exploratory and are not covered by the confirmatory multiplicity adjustment.

Table 8: Release-run grid: substitution AUC, split-mean ± split s.d. over five patient-level splits (each split value is the mean over five seeds), every arm × correction × classifier. Real-trained: FTIR XGB 0.796 ± 0.020, FTIR LR 0.849 ± 0.026; LIBS LR4 0.919 ± 0.025, LIBS XGB 0.881 ± 0.037.

	FTIR / XGB			FTIR / LR		
Arm	raw	moments	cov	raw	moments	cov
Spectral-LDM only	0.778±0.034	0.775±0.018	0.815±0.018	0.626±0.047	0.792±0.023	0.750±0.034
25% PCA mix	0.784±0.014	0.772±0.009	0.799±0.032	0.628±0.033	0.774±0.018	0.713±0.025
50% PCA mix	0.776±0.011	0.774±0.029	0.801±0.026	0.617±0.041	0.740±0.042	0.713±0.026
PCA-Gaussian	0.763±0.023	0.782±0.029	0.793±0.029	0.782±0.022	0.707±0.029	0.724±0.036
WGAN-GP	0.722±0.030	0.734±0.039	0.785±0.026	0.702±0.039	0.707±0.040	0.726±0.035
SMOTE (control)	0.849±0.020	0.843±0.019	0.837±0.014	0.805±0.031	0.811±0.039	0.724±0.036
Control: label-shuffled	0.504±0.024	0.790±0.025	0.797±0.028	0.508±0.037	0.791±0.019	0.749±0.033
Control: noise	0.507±0.021	0.710±0.040	0.783±0.025	0.519±0.051	0.697±0.035	0.727±0.035
	LIBS / LR4			LIBS / XGB		
Arm	raw	moments	cov	raw	moments	cov
Spectral-LDM only	0.503±0.288	0.901±0.035	0.917±0.033	0.419±0.065	0.829±0.029	0.847±0.049
25% PCA mix	0.877±0.035	0.904±0.032	0.929±0.013	0.766±0.060	0.841±0.046	0.891±0.014
50% PCA mix	0.910±0.010	0.907±0.024	0.928±0.013	0.802±0.055	0.848±0.051	0.884±0.007
PCA-Gaussian	0.922±0.018	0.923±0.019	0.928±0.021	0.831±0.043	0.841±0.031	0.871±0.029
WGAN-GP	0.496±0.023	0.891±0.026	0.917±0.027	0.525±0.035	0.470±0.079	0.863±0.033
SMOTE (control)	0.916±0.010	0.914±0.019	0.920±0.019	0.876±0.044	0.878±0.023	0.838±0.033
Control: label-shuffled	0.489±0.119	0.904±0.033	0.920±0.030	0.504±0.010	0.843±0.032	0.850±0.020
Control: noise	0.534±0.050	0.891±0.023	0.918±0.028	0.487±0.026	0.875±0.029	0.861±0.044

G MARGIN SENSITIVITY, MULTIPLICITY, AND PER-SPLIT RESULTS

The main analysis uses an absolute AUC equivalence margin of 0.05, prespecified before the confirmatory runs, as an operational tolerance for this methodological evaluation. It is not a clinically validated non-inferiority margin. Table 9 re-evaluates the pooled out-of-fold confidence intervals in Table 2 at margins from 0.03 to 0.10. The selected FTIR candidate artifact satisfies equivalence from 0.03 upward; the selected LIBS candidate artifact satisfies equivalence from 0.04 upward and fails at 0.03 because its lower confidence bound is -0.031.

The Bonferroni analysis applies only to the six-arm confirmatory equivalence family within each cohort. It does not adjust or validate the operational privacy gates, attribution checks, secondary-classifier analyses, correction experiments, optimization studies, or privacy sensitivity analyses. Under this adjustment, both selected candidate artifacts remain within the prespecified ± 0.05 operational margin, whereas LIBS PCA-Gaussian alone does not. Table 10 reports the selected artifacts by split; the LIBS mixture achieves AUC 0.89–0.92 on all five splits, including the three on which diffusion alone is inverted.

Table 9: Margin sensitivity and multiplicity, raw arms at m=20 (Table 2). Equivalence at each margin δ requires the 90% bootstrap interval inside $(-\delta, \delta)$; ± 0.05 is the pre-registered margin. The Bonferroni column is the percentile bootstrap interval at a one-sided level of $0.05/6$ (six arms per cohort), from a recomputation of the out-of-fold predictions in which the arms with PCA-Gaussian rows are redrawn (their Δ differs from the third column by at most 0.002; LIBS WGAN-GP, resampled on a CPU, by 0.008). Released recipes marked †.

Cohort	Arm	Δ	90% CI	± 0.03	± 0.04	± 0.05	± 0.10	Bonferroni CI	± 0.05
FTIR	Spectral-LDM only	-0.037	$[-0.060, -0.015]$	no	no	no	yes	$[-0.071, -0.005]$	no
FTIR	WGAN-GP	-0.091	$[-0.122, -0.060]$	no	no	no	no	$[-0.139, -0.045]$	no
FTIR	PCA-Gaussian	-0.032	$[-0.061, -0.005]$	no	no	no	yes	$[-0.072, +0.008]$	no
FTIR	25% PCA mix†	-0.006	$[-0.028, +0.015]$	yes	yes	yes	yes	$[-0.036, +0.027]$	yes
FTIR	50% PCA mix	-0.015	$[-0.039, +0.009]$	no	yes	yes	yes	$[-0.047, +0.023]$	yes
FTIR	SMOTE (control)	$+0.023$	$[+0.010, +0.037]$	no	yes	yes	yes	$[+0.004, +0.043]$	yes
LIBS	Spectral-LDM only	-0.327	$[-0.424, -0.234]$	no	no	no	no	$[-0.470, -0.194]$	no
LIBS	WGAN-GP	-0.452	$[-0.573, -0.330]$	no	no	no	no	$[-0.628, -0.285]$	no
LIBS	PCA-Gaussian	$+0.013$	$[-0.015, +0.041]$	no	no	yes	yes	$[-0.028, +0.058]$	no
LIBS	25% PCA mix	-0.033	$[-0.066, -0.003]$	no	no	no	yes	$[-0.082, +0.014]$	no
LIBS	50% PCA mix†	-0.006	$[-0.031, +0.019]$	no	yes	yes	yes	$[-0.042, +0.035]$	yes
LIBS	SMOTE (control)	$+0.007$	$[-0.020, +0.034]$	no	yes	yes	yes	$[-0.032, +0.049]$	yes

Table 10: Released arms split by split: substitution AUC on the real test patients of each split, mean over five seeds (FTIR: XGB; LIBS: LR4), with the components of each released mixture and the real-trained classifier.

Cohort	Training set	Split 1	Split 2	Split 3	Split 4	Split 5	Mean	s.d.
FTIR	Real-trained (XGB)	0.780	0.776	0.827	0.793	0.802	0.796	0.020
FTIR	25% PCA mix†	0.782	0.762	0.795	0.793	0.789	0.784	0.014
FTIR	Spectral-LDM only	0.774	0.721	0.788	0.807	0.802	0.778	0.034
FTIR	PCA-Gaussian	0.803	0.748	0.757	0.749	0.759	0.763	0.023
LIBS	Real-trained (LR4)	0.878	0.934	0.917	0.924	0.944	0.919	0.025
LIBS	50% PCA mix†	0.893	0.915	0.920	0.912	0.909	0.910	0.010
LIBS	Spectral-LDM only	0.788	0.267	0.240	0.837	0.382	0.503	0.288
LIBS	PCA-Gaussian	0.928	0.899	0.907	0.931	0.942	0.922	0.018

H DIAGNOSTIC MEMBERSHIP INFERENCE AT LOW FALSE-POSITIVE RATES

Table 11 reports the attack in §6 at 1% and 5% false-positive rates for every arm, with SMOTE as a positive control and shuffled membership labels as a negative control. These low-FPR analyses are privacy sensitivities rather than part of the confirmatory equivalence family. For the selected FTIR candidate artifact (one draw per split; 2,040 member and 880 non-member observations), attack AUC is 0.520 and the true-positive rate at 1% FPR is 0.0127 [0.0039, 0.0191], or 1.27 times chance, compared with 1.47 for the shuffled-label control. These results indicate low measured membership risk at the reported operating point under this attack; they do not establish anonymity or protection against other attacks.

Table 11: Membership inference at low false-positive rate. Distance-to-closest-record attack in the audit space of Table 2, pooled over five patient-level splits × five seeds (FTIR: 10,200 member and 4,400 non-member observations; LIBS: 1,950 and 850), true-positive rate at the largest achievable FPR at or below the target, 90% patient-level cluster-bootstrap CI (2,000 resamples). Lift is TPR/FPR; the last column is the lift of the same attack after membership labels are shuffled (negative control). Per split, 1% FPR corresponds to 8.8 (FTIR) and 1.7 (LIBS) non-members, which is why the figures are pooled. Attack AUCs here pool five seeds per split, so they differ from the single-seed per-split means of Table 2 (released LIBS recipe: 0.505 against 0.476); the released files themselves give 0.520 (FTIR) and 0.516 (LIBS, LR4 space; Appendix I). Released recipes marked †.

Cohort	Arm	Attack AUC	TPR@1%FPR	90% CI	Lift	TPR@5%FPR	Shuffled lift@1%
FTIR	Spectral-LDM only	0.522	0.0093	[0.0039, 0.0186]	0.93	0.0461	1.21
FTIR	WGAN-GP	0.504	0.0103	[0.0081, 0.0131]	1.03	0.0582	0.99
FTIR	PCA-Gaussian	0.521	0.0093	[0.0060, 0.0141]	0.93	0.0490	1.17
FTIR	25% PCA mix†	0.519	0.0093	[0.0042, 0.0174]	0.93	0.0507	1.28
FTIR	50% PCA mix	0.520	0.0085	[0.0049, 0.0210]	0.85	0.0563	1.29
FTIR	SMOTE (control)	1.000	0.9995	[0.9990, 0.9999]	99.95	0.9997	1.03
LIBS	Spectral-LDM only	0.515	0.0077	[0.0025, 0.0279]	0.77	0.0462	1.03
LIBS	WGAN-GP	0.504	0.0128	[0.0077, 0.0287]	1.28	0.0513	1.08
LIBS	PCA-Gaussian	0.506	0.0092	[0.0051, 0.0152]	0.92	0.0497	0.87
LIBS	25% PCA mix	0.504	0.0231	[0.0106, 0.0305]	2.31	0.0559	1.03
LIBS	50% PCA mix†	0.505	0.0195	[0.0125, 0.0277]	1.95	0.0564	1.85
LIBS	SMOTE (control)	0.987	0.8549	[0.8231, 0.9140]	85.49	0.9333	0.97

I EXTERNAL QUALITY AND PRIVACY BENCHMARK CORRESPONDENCE

Clinical synthetic-data benchmarks commonly evaluate quality and empirical privacy and then aggregate the two axes (Sella et al., 2025; Yoon et al., 2020; El Emam et al., 2020). The audit in §6 uses related distance statistics but a different decision rule. Table 12 summarizes the correspondence, and Table 13 reports exploratory external scores for the selected LIBS candidate artifact. These comparisons and the representation-sensitivity analyses below are diagnostic; they are not part of the six-arm confirmatory equivalence family.

Table 12: Clinical synthetic-data benchmark quantities and their counterparts in this audit. The statistics overlap, but their aggregation and interpretation differ.

Quantity	**Benchmark**	**This audit**
Re-identification	Identifiability score (Yoon et al., 2020): fraction of real records whose nearest synthetic record is closer than their nearest real neighbour, using an entropy-weighted distance	Copy rate and NN ratio, with the operational copy-rate gate < 0.10
Membership disclosure	Membership-inference score (El Emam et al., 2020): distance from a real record to its closest synthetic record, thresholded on a hold-out partition	The same distance family, reported as member–non-member attack AUC and TPR at 0.5–10% FPR, with shuffled-label and SMOTE controls; the operational attack-AUC gate is < 0.55
Utility	Predictive performance of a model trained on synthetic data and evaluated on a real hold-out set	Substitution AUC under two one-sided equivalence tests at the 0.05 operational margin prespecified before the confirmatory runs, supplemented by reverse transfer and zero-signal attribution controls
Fidelity	Univariate tests, correlation and mutual-information matrix distances, and Wasserstein distance, normalized against a random-in-range generator	Round-trip Kolmogorov–Smirnov distance, variance ratio, edge artifact, and discriminator AUC (Appendix B)
Aggregation	Quality–privacy score formed as a harmonic mean of normalized quality and privacy scores	Separate operational gates; among audit-eligible arms, the prespecified selection rule identifies the candidate with the lowest measured attack AUC

Three differences are deliberate. First, the unit is the patient: members and non-members are patients rather than spectra, and all acquisitions from one patient remain on the same side of a split. Second, the membership statistic is computed per split and pooled across five splits rather than estimated from one hold-out partition. Sella et al. (2025) note that their membership score could not be estimated for complete datasets without a hold-out and recommend cross-validation; the split structure provides that separation. Third, empirical privacy is treated as an operational gate rather than a term that can be offset by utility. These gates are not validated risk tolerances.

This distinction changes the interpretation of two controls. SMOTE provides the highest synthetic-training utility on FTIR but permits near-perfect membership discrimination under the evaluated distance-based attack (§7.4). Retraining the autoencoder improves every reported fidelity measure while increasing attack AUC from 0.551 to 0.653. An aggregate quality–privacy score may rank such arms favorably, whereas a gate exposes the measured membership risk. This does not establish general privacy for arms that pass; it only prevents the evaluated utility measure from compensating for a demonstrated failure under the specified attack.

Table 13: External scores on the two released cohorts, five patient-level splits, mean ± split s.d. **Top:** the scores of Sella et al. (2025) computed on the released files (400 FTIR and 75 LIBS synthetic spectra per split, the rows drawn from that split's generators) against the real patients of the split; the reference generator is Random-in-range, and its identifiability is 0 on every split of both cohorts, so the normalised privacy uses the ratio form. Distances are z-scored Euclidean over all channels (the entropy weights of Yoon et al. (2020) are uniform on these cohorts); the LIBS quality matrices use 256 uniformly spaced channels of 24,488. **Bottom:** the membership attack and the geometric measures in the full-width space of the top block ("full") and in the audit space of Table 2 ("audit": the 235-point classifier domain on FTIR, which is the full width; the four element features of LR4 on LIBS). The controls contain no patient: each is the pair of per-class mean spectra of the named patients.

	FTIR		LIBS	
Score	Released	Random-in-range	Released	Random-in-range
Identifiability score IS (reference: all real records)	0.283 ± 0.022	0 on every split	0.534 ± 0.068	0 on every split
IS (reference: training members, as in Yoon et al.)	0.352 ± 0.020	0 on every split	0.656 ± 0.088	0 on every split
Membership-inference score MIS (F1, median rule, full width)	0.590 ± 0.013	0.589 ± 0.016	0.627 ± 0.052	0.591 ± 0.047
MIS in the 10-component PCA space of the training fold	0.593 ± 0.018	0.575 ± 0.016	0.573 ± 0.023	0.579 ± 0.045
Normalised quality Q_{MI_d} / Q_{C_d} / Q_W	0.158 / 0.489 / 0.869	0 by construction	0.503 / 0.162 / 0.000	0 by construction
Normalised privacy NP_{IS} / NP_{MIS} (ratio form)	0.717 / 0.981	1 by construction	0.466 / 0.900	1 by construction
$\mathrm{metaQPS}_{\mathrm{am}}$ / $\mathrm{metaQPS}_{\mathrm{hm}}$	0.629 / 0.324	0 by construction	0.333 / 0.000	0 by construction

Cohort	Synthetic set	Attack AUC (full)	Attack AUC (audit)	IS (full)	Copy rate (full)
FTIR	Released cohort (25% PCA-Gaussian mixture)	0.520 ± 0.022	= full	0.283 ± 0.022	0.094 ± 0.041
FTIR	PCA-Gaussian, 20 components	0.523 ± 0.014	= full	0.350 ± 0.028	0.016 ± 0.008
FTIR	PCA-Gaussian, 10 components	0.519 ± 0.011	= full	0.527 ± 0.034	0.084 ± 0.018
FTIR	Control: class means of the training patients	0.504 ± 0.025	= full	0.087 ± 0.007	0.917 ± 0.186
FTIR	Control: class means of the held-out patients	0.495 ± 0.028	= full	0.085 ± 0.010	0.917 ± 0.186
LIBS	Released cohort (50% PCA-Gaussian mixture)	0.584 ± 0.073	0.516 ± 0.031	0.534 ± 0.068	0.136 ± 0.035
LIBS	PCA-Gaussian, 40 components	0.583 ± 0.072	0.503 ± 0.051	0.675 ± 0.046	0.280 ± 0.065
LIBS	PCA-Gaussian, 10 components	0.562 ± 0.068	0.476 ± 0.028	0.880 ± 0.015	0.603 ± 0.073
LIBS	Control: class means of the training donors	0.518 ± 0.083	0.489 ± 0.082	0.211 ± 0.024	1.000 ± 0.000
LIBS	Control: class means of the held-out donors	0.471 ± 0.075	0.480 ± 0.056	0.241 ± 0.013	1.000 ± 0.000

Exploratory scores for the selected candidate artifacts. Table 13 applies the published benchmark definitions using a z-scored Euclidean distance over all channels. Entropy weights are uniform in these cohorts because equal-frequency binning gives every tie-free continuous channel the same entropy. The random-in-range reference has identifiability zero on every split: in 235 or 24,488 correlated dimensions, a uniform per-channel draw is never closer to a real spectrum than that spectrum's nearest real neighbour. Equations (7) and (8) of Sella et al. (2025), which divide by the reference score, are therefore undefined here and become negative whenever the reference is the lowest-risk arm. The accompanying prose instead describes normalization as a ratio to the random reference; Table 13 uses that interpretation because it is consistent with the reported magnitudes.

The identifiability score also reflects geometric centrality. On LIBS, two controls containing no individual donor—the two training-member class means and the corresponding held-out class means—score $\mathrm{IS} = 0.21$ and 0.24 with copy rate 1.00. A 10-component linear-Gaussian generator, which does not explicitly reproduce a training row, scores 0.88 versus 0.68 for the 40-component model. On FTIR, the class-mean controls score 0.09 with copy rate 0.92, and the 10-component model scores 0.53 versus 0.35 for the 20-component model. The selected artifacts' scores of 0.28 (FTIR) and 0.53 (LIBS) must therefore be interpreted as distance-based geometry rather than direct evidence of anonymity. The external membership scores are 0.59 and 0.63, compared with 0.59 for the reference and a chance level of 0.58.

The exploratory quality axis disagrees with substitution on LIBS. The sliced Wasserstein distance between the selected LIBS candidate artifact and real spectra is 1.1–6.4 across splits, versus 0.5–0.7 for the random-in-range reference. Consequently, normalized Wasserstein quality Q_W and the harmonic-mean summary $\mathrm{metaQPS}_{\mathrm{hm}}$ are zero on every split, although the same rows train a classifier satisfying the prespecified substitution-equivalence criterion in the task representation. The measures answer different questions: substitution evaluates the four element-line features used by LR4, whereas full-width distance evaluates all 24,488 channels, over which the diffusion rows remain distant from real spectra. On FTIR, $Q_W = 0.87$, while $Q_{\mathrm{MI}_d} = 0$ on two splits and $\mathrm{metaQPS}_{\mathrm{hm}} = 0.32 \pm 0.33$. Thus, aggregate quality scores do not substitute for task utility, attribution, or empirical privacy tests.

Representation dependence of the LIBS membership result. The prespecified LIBS audit operates in the four-feature LR4 task space. For the selected candidate artifact, the per-split candidate attack AUC is 0.48; on the assembled artifact, the same attack yields 0.52 ± 0.03 (mean 0.516, permutation-null 95th percentile 0.54, $p = 0.27$). These findings indicate low measured membership risk under the specified task-space attack only.

The full-spectrum result changes that interpretation. Across all 24,488 channels, the assembled LIBS artifact yields attack AUC 0.58 ± 0.07 (split values 0.50, 0.58, 0.54, 0.62, and 0.69), exceeding the permutation-null 95th percentile of 0.54 ($p = 0.004$). Pooled across splits (390 members and 170 non-members), the full-width TPR is 1.0 times chance at 1% FPR [0.0, 6.1] and 2.1 times chance at 5% FPR [0.6, 3.9], compared with 2.3 and 1.0 for the shuffled control. Thus, low-FPR rates are not distinguishable from the control, but the aggregate AUC detects membership signal.

Diagnostic decomposition attributes this signal primarily to PCA-Gaussian rows: full-width attack AUC is 0.68 for the evaluated PCA-Gaussian arm, 0.51 for diffusion alone, and 0.65 for the 50% mixture, whose 1%-FPR TPR is 6.6 times chance [3.3, 13.5]. PCA-Gaussian samples lie in the span fitted from the training fold (relative residual $\leq 2 \times 10^{-7}$, versus 0.020–0.040 for held-out donors), which is detectable in ambient space but not in the four-feature task space; zero-donor class-mean controls score 0.52 and 0.47. The LIBS candidate therefore satisfies the prespecified task-space gate but does not establish representation-robust privacy. The task-space and ambient-space findings have equal interpretive status for claim limitation: neither alone certifies anonymity, and the ambient result requires privacy conclusions to remain representation-conditional.

For FTIR, the classifier and ambient domains coincide. The selected FTIR artifact yields attack AUC 0.52 (five-split mean 0.520, null 95th percentile 0.520, $p = 0.047$), while the 20-component PCA-Gaussian arm yields 0.52 and the class-mean controls 0.50 and 0.49. Per-split null 95th percentiles are 0.54–0.55 for FTIR and 0.59–0.61 for LIBS, so the 0.55 operational gate is interpretable for the pooled five-split mean rather than for a single split. The copy rate of held-out real patients against each training fold is 0.060 for FTIR and 0.035 for LIBS, providing context for the operational 0.10 copy-rate gate. All analyses in this section are sensitivity analyses and provide no formal privacy guarantee.

J EXPLORATORY THRESHOLD ANALYSIS

Table 14 reports operating points for thresholds selected using real validation data, synthetic validation data, or 20–40 labeled real patients withheld from the test set. This calibration analysis is exploratory and is outside the confirmatory equivalence family.

Table 14: Operating points on real test patients, FTIR XGB, five splits × five seeds. Real-trained classifier with thresholds selected on its own real validation fold; synthetic-trained classifier (diffusion only, w=3, m=20) with thresholds selected on its synthetic validation fold ("own") or on n real labelled patients withheld from the test set.

Training set	Threshold rule	AUC	Acc	Sens	Spec
Real	fixed 0.5	0.796	0.728	0.753	0.693
Real	sensitivity ≥ 0.80 on own validation	0.796	0.722	0.753	0.678
Real	Youden J on own validation	0.796	0.705	0.655	0.774
Synthetic	fixed 0.5	0.778	0.695	0.607	0.819
Synthetic	sensitivity ≥ 0.80 on synthetic validation	0.778	0.526	0.217	0.962
Synthetic	Youden J on synthetic validation	0.778	0.682	0.570	0.840
Synthetic	sensitivity ≥ 0.80 on n=10 real patients	0.778	0.675	0.621	0.751
Synthetic	sensitivity ≥ 0.80 on n=20 real patients	0.778	0.707	0.717	0.694
Synthetic	sensitivity ≥ 0.80 on n=40 real patients	0.778	0.711	0.765	0.634
Synthetic	sensitivity ≥ 0.80 on n=80 real patients	0.778	0.721	0.744	0.689

K DIAGNOSTIC QUANTITY AND ENSEMBLING GRID

Table 15 reports the quantity and ensembling analysis of §7.1 and the guidance sweep at the confirmatory sample scale. These optimization analyses are diagnostic and are outside the six-arm confirmatory equivalence family.

Table 15: Quantity and ensembling, FTIR XGB, split-mean ± split s.d. of substitution AUC. Left: w=5, one generator ("single") or five seed-trained generators pooled ("ensemble"), validation fold carved from the synthetic training set. Right: five pooled generators at $m \in \{16, 20\}$ over the guidance sweep (the release recipe is w=3, m=20). Real-trained reference under the same validation protocol: 0.781 ± 0.029.

m	single	ensemble	w	m=16	m=20
1	0.712±0.024	0.737±0.048	1	0.733±0.038	0.729±0.033
2	0.715±0.028	0.744±0.036	1.5	0.745±0.029	0.751±0.028
4	0.718±0.032	0.762±0.023	2	0.764±0.032	0.770±0.025
8	0.729±0.028	0.763±0.027	3	0.768±0.035	0.778±0.034
16	0.730±0.033	0.767±0.035	4	0.777±0.019	0.777±0.022
			5	0.774±0.034	0.774±0.037

L EXPLORATORY SCARCITY ANALYSIS

Table 16 reports the scarcity analysis of §7.5 for both cohorts and all three fillers. This analysis is exploratory and is not part of the confirmatory equivalence family.

Table 16: Exploratory scarcity analysis. AUC gain over real-only training on the same fraction f of real patients, with the class deficit filled by two generators or the SMOTE control; mean over five splits and five seeds (FTIR XGBoost, LIBS LR4).

	FTIR			LIBS		
f	Spectral-LDM	PCA-Gaussian	SMOTE	Spectral-LDM	PCA-Gaussian	SMOTE
0.10	+0.040	+0.075	+0.153	−0.080	+0.052	+0.055
0.25	+0.042	+0.060	+0.111	+0.015	+0.014	+0.017
0.50	+0.025	+0.002	+0.051	+0.013	+0.014	+0.011
0.75	+0.015	+0.018	+0.035	+0.004	+0.002	+0.003

M SCREENED EXTERNAL COHORTS

Table 17 records external FTIR cohorts considered for a second oncological validation study and the reason each was excluded. "Deposited" means that the spectral matrix itself is downloadable, not merely that the publication is open access. The criteria are: (C1) benchtop ATR-FTIR; (C2) blood-derived biofluid; (C3) patient-level labels with a binarizable oncological endpoint; (C4) deposited spectra; and (C5) a cohort sufficiently large to add power. No candidate satisfied all five criteria. Consequently, the audit has not been independently validated on a second oncological FTIR cohort, and the LIBS analysis should be interpreted as a second methodological case study in a different modality and endpoint rather than direct replication.

Table 17: External FTIR cohorts screened as candidate second validation sets.

Cohort / source	**Modality, specimen**	**Scale**	**Fails**	**Reason for exclusion**
Brain cancer serum (Butler et al., 2019)	ATR-FTIR, dried serum	724 patients	C4	Benchtop, serum, patient-level binary; spectra available on request, not deposited.
Pancreatic and colorectal serum (McHardy et al., 2023)	ATR-FTIR, dried serum	625 / 659 patients	C4	Closest prior work and cohort match; spectra not deposited.
Ovarian and endometrial plasma and serum (Gajjar et al., 2013)	ATR-FTIR, plasma and serum	$\sim$90 patients	C5	Correct modality and biofluid; smaller than the cohort in use.
Digestive-tract cancer serum (Guo et al., 2022)	Synchrotron FTIR, wet serum	$\sim$210 patients	C1, C4	Synchrotron, measured in solution; available on request only; 252 specimens from 210 patients, patient-level separation not stated.
Breast tissue microarray (Tang et al., 2021)	FTIR imaging, tissue	207 cores (15 non-malignant)	C2, C3	Deposited, but pixel-level tissue imaging; a minority class of fifteen at core level.
Prostate tissue microarray (O'Leary et al., 2026)	FTIR imaging, tissue	1,045 cores	C2, C3	Deposited and large at core level; tissue imaging rather than a patient-level biofluid.
Breast cancer repository with code (Tomas et al., 2022)	ATR-FTIR, tissue	166 samples	C4	Provides a pseudo-spectrum generator based on class eigenvectors; a measured patient-level absorbance matrix is not documented.
COVID-19 swab fluid (Nogueira et al., 2021)	ATR-FTIR, swab medium	65 / 178 patients	C2–C4	Not blood-derived or oncological; no deposited matrix; replicates treated as independent.
COVID-19 saliva (Martinez-Cuazitl et al., 2021)	ATR-FTIR, saliva	1,209 / 255 patients	C2–C4	Not blood-derived or oncological; not deposited; cases and controls collected in different periods.