npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

Article in Press

https://doi.org/10.1038/s41746-026-03226-9

Conditional deep generative modeling of blood-based infrared spectra enables controlled in-silico phenotyping studies

Niklas Leopold-Kerschbaumer[1,2,3,13], Timo Halenke[2,4,13], Selina Süzeroğlu[2,4,13], Moritz Jung[1,2,3], Mariia Seleznova[2,3], Nicole Thorisch[1,4], Jeanette M. Lorenz[4,5], Thomas Bocklitz[6,7], Bjoern M. Eskofier[8,9], Gitta Kutyniok[3,10,11,12], Ferenc Krausz[1,2,4] & Kosmas V. Kepesidis[1,2,4] ✉

Infrared molecular fingerprinting of blood offers a scalable, minimally invasive window into human physiology, but limited follow-up, imbalanced cohorts, and restricted access to diverse phenotypes constrain systematic studies. Here, we introduce a conditional deep generative framework for synthesizing blood-based infrared spectra that preserves individual-level structure while allowing controlled manipulation of demographic and anthropometric covariates. Using 25,308 spectra from 5,863 ostensibly healthy participants in the longitudinal Health4Hungary - Hungary4Health cohort, we train a Conditional Variational Autoencoder, a Conditional Boundary Equilibrium GAN, and a Conditional Diffusion Model to generate blood-based infrared spectra conditioned on age, sex, and body mass index. We show that the generated spectra closely match held-out real data across multiple levels and faithfully encode demographic and anthropometric information. We further demonstrate two in-silico applications: modeling of individualized healthy aging trajectories that follow cohort-level aging manifolds while retaining individual-specific characteristics, and targeted augmentation of underrepresented body mass index categories. Together, these results demonstrate the feasibility of utilizing conditional generative modeling of blood-based infrared spectra for virtual cohort construction, cohort balancing, and controlled in-silico phenotyping, paving the way toward more comprehensive and data-efficient studies in precision health.

Longitudinal molecular profiling is emerging as a powerful approach for personalized health monitoring, with the potential to support early disease detection, risk assessment, and preventive medicine. The integration of diverse "omics" technologies—including genomics, transcriptomics, proteomics, metabolomics, epigenomics, microbiomics, and exposomics—with biosensor data, medical imaging, and electronic health records has created an unprecedented wealth of health-related information[1–13]. Such integrated datasets are deepening our understanding of human biology and may transform how early warning signs of disease are identified.

Within this broader landscape, infrared (IR) spectroscopy of liquid biopsies, such as blood plasma, offers a noninvasive route to molecular health monitoring. In particular, infrared molecular fingerprinting (IMF) based on Fourier-transform infrared (FTIR) spectroscopy captures vibrational signatures of biomolecules and provides a holistic snapshot of plasma composition, including contributions from proteins, lipids, carbohydrates, metabolites, and other biochemical constituents[14–17]. More broadly, spectral and hyperspectral imaging approaches have been widely explored in biomedical analysis because they can couple molecularly informative spectral

[1]Center for Molecular Fingerprinting (CMF), Frontiers Foundation, Budapest, Hungary. [2]Max Planck Institute of Quantum Optics (MPQ), Garching, Germany. [3]Department of Mathematics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany. [4]Faculty of Physics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany. [5]Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany. [6]Leibniz Institute of Photonic Technology (Leibniz-IPHT), Jena, Germany. [7]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University (FSU), Jena, Germany. [8]Chair of AI-supported Therapy Decisions, LMU München, Munich, Germany. [9]Institute of AI for Health, Helmholtz Zentrum München, Neuherberg, Germany. [10]Munich Center for Machine Learning (MCML), Munich, Germany. [11]Department of Physics and Technology, University of Tromsø, Tromsø, Norway. [12]Institute of Robotics and Mechatronics, DLR-German Aerospace Center, Cologne, Germany. [13]These authors contributed equally: Niklas Leopold-Kerschbaumer, Timo Halenke, Selina Süzeroğlu. ✉e-mail: kosmas.kepesidis@cmf.hu

contrast with spatial or morphological information, including applications to blood-cell identification and classification[18–20]. IMF has shown high temporal stability at the individual level, an important prerequisite for longitudinal and clinical applications[21–23]. Reported applications include oncological and metabolic profiling, among others[24–37].

Despite this promise, assembling large, diverse, and well-balanced longitudinal datasets remains challenging. Repeated sampling of the same individuals over extended time periods requires long-term follow-up, substantial logistical coordination, and sustained participant engagement. Consequently, studies often face limited temporal coverage, imbalanced demographic or physiological subgroups, and modest statistical power. These constraints hinder systematic investigation of aging effects, longitudinal health trajectories, and transitions from health to disease.

Generative modeling provides a strategy to mitigate these limitations by synthesizing spectra that resemble real measurements while expanding the range of available data for analysis. Deep-learning-based generative models are particularly suitable for this task because they can learn complex, high-dimensional data distributions directly from empirical observations. Unlike simple resampling strategies or predefined parametric models, they can capture nonlinear relationships, latent structure, and higher-order dependencies among spectral features. This flexibility is especially relevant for IMF, where each spectrum represents an aggregated molecular signature of a complex biological matrix, and biologically meaningful variation may be distributed across many correlated spectral regions.

For longitudinal and precision-medicine applications, however, reproducing the overall spectral distribution is not sufficient. Synthetic spectra are most useful when their generation can be controlled with respect to biologically and clinically relevant variables. Conditional generative models address this need by incorporating covariates into the generative process, enabling spectra to be synthesized for virtual individuals or subgroups with specified characteristics. Conditioning on variables such as age, sex, and body mass index (BMI) allows models to represent structured population heterogeneity and supports controlled in-silico applications, including cohort balancing, targeted augmentation of underrepresented subgroups, and simulation of individualized longitudinal trajectories.

Previous generative approaches for blood-based spectroscopic data have mainly relied on classical statistical models, such as multivariate normal formulations[38–40], or linear combination-based schemes in which new spectra are generated by mixing existing measurements[41]. While these approaches can reproduce certain global distributional properties, their linear or parametric assumptions limit their ability to capture complex nonlinear dependencies and to generate spectra conditioned on individual-level covariates. Deep generative models have also been proposed for spectral synthesis[42–44], but in their unconditional form, they are restricted to modeling the marginal data distribution and do not provide explicit control over subject-specific attributes. Conditional deep generative models provide a more flexible alternative by learning nonlinear latent representations of spectral data while enabling controlled generation with respect to demographic and anthropometric variables. Although such models are often less interpretable than classical approaches, this limitation is less restrictive in the present setting, where FTIR plasma spectra are already highly aggregated, indirect measurements of underlying biochemistry.

In this work, we present conditional deep generative models for IMF spectra that integrate age, sex, and BMI as conditioning variables. Specifically, we evaluate a Conditional Variational Autoencoder (CVAE), a Conditional Boundary Equilibrium GAN (CBEGAN), and a Conditional Diffusion Model (CDIFF)[42,45–50]. The models are trained on spectra from healthy individuals in a longitudinal cohort and are designed to generate high-quality synthetic spectra suitable for clinical and translational research. We evaluate spectral fidelity and clinical utility and demonstrate several in-silico applications: (i) adjustable population scaling, (ii) augmentation of underrepresented subgroups, and (iii) digital-twin-style conditional projections for simulating healthy aging trajectories. Figure 1 summarizes the data type, model architectures, and main use cases. Overall, our findings highlight the potential of conditional deep generative modeling to advance in-silico clinical research and enable controlled, data-driven investigations in precision medicine.

RESULTS

Longitudinal study cohort

The Health4Hungary—Hungary4Health (H4H) program is a longitudinal clinical study launched in 2021, designed for population-scale health monitoring using infrared molecular fingerprinting. The study aims to follow ~15,000 healthy volunteers over 10 years. At the time of analysis, the dataset comprises 25,308 infrared spectral samples collected from 5863 participants, each contributing between one and six visits. The average time span for individuals with all 6 visits is 2.4 years, while the entire dataset was collected of a total time span of 3.6 years. This dataset serves as the basis for the present work.

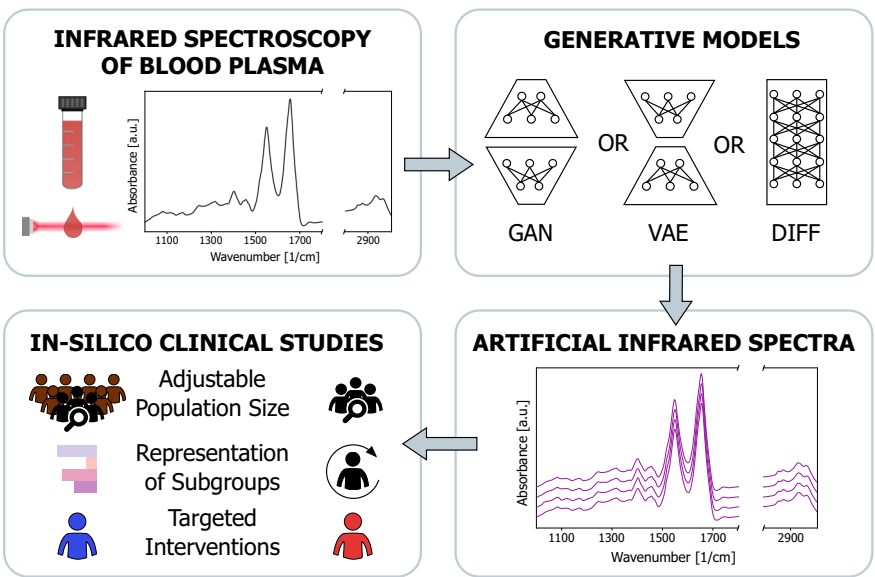

Fig. 1 | Overview of the generative framework. Schematic of the workflow for synthesizing individual-specific blood infrared spectra and in-silico clinical applications (e.g., cohort balancing, simulating healthy aging).

Table 1 | Overview of the H4H longitudinal dataset

Dataset	# Blood Samples	# Individuals	# Visits per Individual
H4H Total	25,308	5863	1–6
Demographics: 2009 male, 3854 female; Age: 39–84 y; BMI: 15.0–64.87			
Train Set (80%)	20,246	4698	1–6
Test Set (20%)	5062	1165	1–6
Test Set Subdivision: Single-Visit Cohorts			
SVC (6 cohorts)	5062 total	5062 total	1 per individual
(Per-cohort size range: 843–844)			

Sample counts, visit counts and demographics are reported. Further demographic details and details on SVCs in Fig. S1 and Figure S2 (Supplementary). Train and test sets are mutually exclusive at the participant level.

For model development, participants were divided into mutually exclusive training and test sets (see Table 1 and Figs. S1-S2 in Supplementary Information). The generative models were trained exclusively on the training set, while their performance and clinical relevance were evaluated on the held-out test set. To avoid potential information leakage, we ensure no individual appears in both the training and testing sets. Evaluation involved (i) comparing the statistical distributions of real and synthetic spectra and (ii) comparing clinically relevant predictive models trained on real and synthetic data, respectively.

Generative modeling architectures

To generate condition-specific spectra, we train three models: a CVAE, a CBEGAN, and a CDIFF. The CVAE encodes an input spectrum and its condition vector into a Gaussian latent distribution and decodes samples conditioned on age, sex, and BMI. The encoder and decoder use multilayer perceptrons (MLPs) and are optimized via the evidence lower bound (ELBO) method[47]. The CBEGAN generates spectra from noise and conditions, with an autoencoder discriminator using reconstruction error as the adversarial signal[45]. The CDIFF generates spectra by starting from a random noise vector and iteratively denoising it through a learned reverse process that approximates the inversion of a gradual noising (diffusion) procedure[49]. Details on the implementation and training are provided in the *Methods* section. Using the trained models, we generated three synthetic datasets, $\mathcal{D}_{\text{CVAE}}$, $\mathcal{D}_{\text{CBEGAN}}$, and $\mathcal{D}_{\text{CDIFF}}$, each with $N = 5,062$ spectra matched to the demographic distribution $\boldsymbol{y}$ of the real test set $\mathcal{D}_{\text{real}}$.

Quality of synthetic fingerprints

We compared synthetic spectra with the held-out real test set (Table 1). First, qualitative inspection showed that all three models produced smooth, continuous spectra comparable to real data (Fig. 2a). For a lower-dimensional comparison, we performed principal component analysis (PCA) on real spectra and projected the synthetic spectra onto the resulting components. The first two PCs showed overlapping clusters for real and synthetic data, indicating partial preservation of the global variance structure.

To assess potential differences between real and synthetic spectra at the spectrally resolved level, we compute the effect size (Cohen's d)[51] at each wavenumber. For a given wavenumber k, the effect size is defined as

$$d(k) = \frac{M_G(k) - M_R(k)}{s(k)}, \tag{1}$$

where $M_G(k)$ and $M_R(k)$ denote the mean absorbance values of the synthetic and real spectra, respectively, and $s(k)$ is the pooled standard deviation at that wavenumber. The pooled standard deviation is computed as

$$s(k) = \sqrt{\frac{(n_G - 1)s_G^2(k) + (n_R - 1)s_R^2(k)}{n_G + n_R - 2}}, \tag{2}$$

with $s_G(k)$ and $s_R(k)$ representing the standard deviations of the synthetic and real absorbance distributions, and n_G and n_R their corresponding sample sizes. This formulation yields a standardized, unitless measure of the magnitude of the difference between the two distributions at each spectral position.

As shown in Fig. 2b, the effect size per wavenumber remained well below 20% of the standard deviation across all wavenumbers for the CVAE and CDIFF, while the CBEGAN only rarely exceeded this threshold. This suggests that the deviations between real and synthetic spectra are small and of limited practical significance[52]. To probe higher-order structure, we extracted a set of 24 peak ratios previously reported[21] to reflect interpretable biochemical relationships. A description of these peak ratios and their biological relevance is provided in Table S1 of the Supplementary Information. Figure 2c shows that the distributions of peak ratios for real and synthetic spectra align closely. Moreover, Fig. 2d shows that effect sizes across ratios remained well below the 20% threshold[52], indicating a small effect for the CVAE and CDIFF, while the CBEGAN occasionally exceeds this threshold.

Additionally, we quantified the similarity between real and synthetic distributions using complementary distribution distance metrics (Table 2). Across all metrics, the diffusion model consistently achieved the lowest distances, indicating closer agreement with the real data distribution relative to the other models. In particular, both the global and within-condition Fréchet Distance (FD), which captures discrepancies in first- and second-order statistics (mean and covariance), were substantially reduced for the diffusion model compared to the VAE and GAN. A similar trend was observed for the Maximum Mean Discrepancy (MMD), a kernel-based metric sensitive to higher-order differences in the distributions: the diffusion model yielded markedly lower values, both globally (in PCA space) and within conditions, suggesting improved alignment beyond simple moment matching. The within-condition Intraclass Correlation Coefficient (ICC) provides a complementary perspective by quantifying the ratio of between-condition to total variance, thereby reflecting how well the conditional structure of the data is preserved. Here, all models produced ICC values in the vicinity of the real dataset (0.066), with the diffusion model being closest. However, this agreement should be interpreted with caution, as the diffusion model was explicitly tuned with respect to this metric (see supplementary Fig. S3).

Finally, following prior work on auditing generative models using nearest-neighbor distances between synthetic and training samples[53], we evaluated whether generated samples were unusually close to the training data. For each generated sample, we computed its Euclidean distance to the nearest training sample and compared the resulting distance distribution with that of held-out real test samples. Smaller nearest-neighbor distances indicate a greater potential for memorization and privacy leakage. Details on the evaluation setup are reported in section in the Supplementary Information. We found a nearest-neighbor proximity score (NNPS) of 0.45, 0.60, and 0.47 for CVAE, CBEGAN and CDIFF, respectively. Values close to 0.5 indicate that generated samples are no closer to the training data than held-out real samples, whereas larger values suggest an increased tendency toward memorization. Accordingly, CVAE and CDIFF showed no evidence of elevated privacy risk based on this analysis, while CBEGAN exhibited a modest increase in proximity to training samples. Nevertheless, given the limited magnitude of this effect and the inherently limited accuracy for predicting even the conditioned covariates, we found no indication of substantial privacy leakage in any of the evaluated models.

Taken together, these results indicate that while all models capture the coarse statistical structure of the data, the diffusion model more faithfully

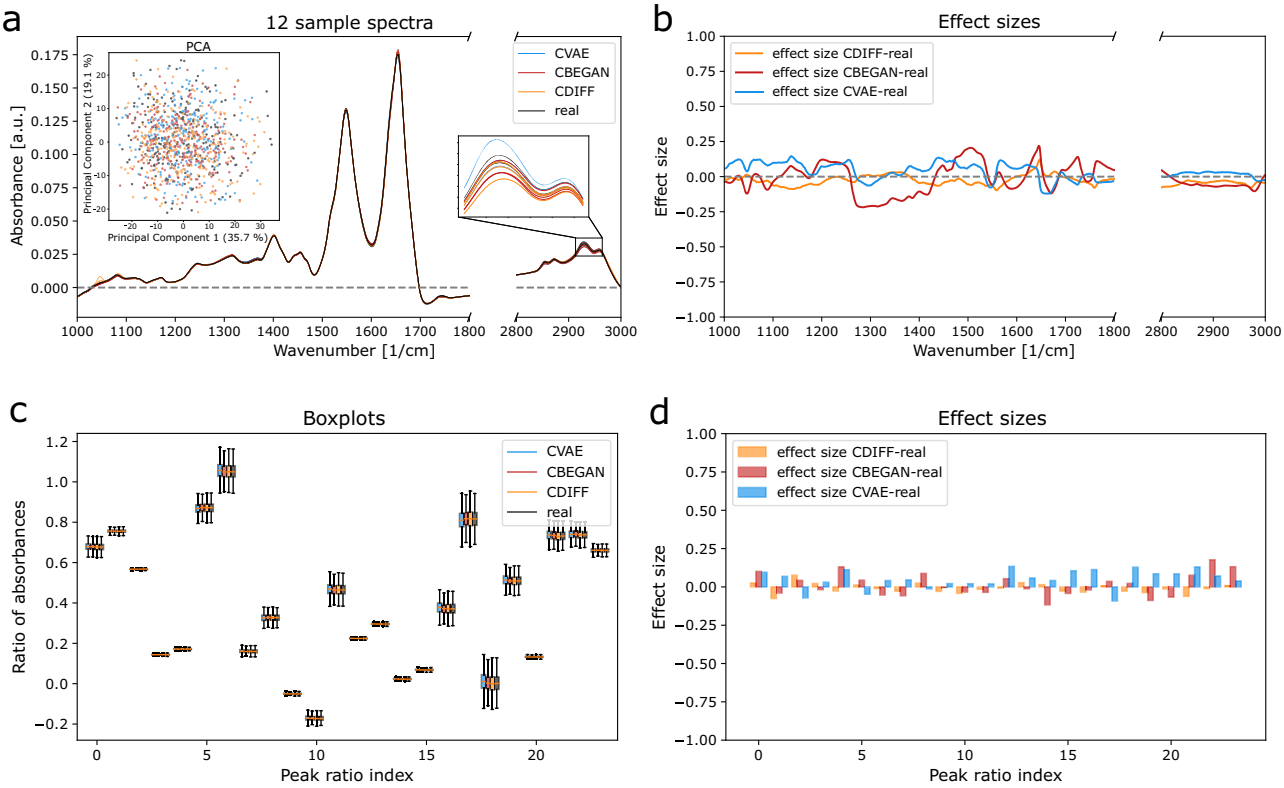

Fig. 2 | Distribution analysis. a Representative spectra (three each) from the real, CVAE-synthetic, CBEGAN-synthetic, and CDIFF-synthetic datasets. Spectra show absorbance in arbitrary units, measured via FTIR and preprocessed as described in the Methods. The inset displays a scatter plot of the first two principal components for 200 samples from each set. Inset zoom range: 2915 cm^{1}–2966 cm^{1}. **b** Cohen's d effect size (real vs. synthetic) computed at each wavenumber. **c** Box plots of peak-ratio distributions for real and synthetic data. **d** Cohen's d values for peak ratios. Ratio indices correspond to Table S1 (Supplementary).

Table 2 | Distance metric based comparison of generative models

Metric	VAE	GAN	Diffusion
Global FD	10.8930	23.6450	3.1529
Global MMD (10 PCs)	0.0036	0.0044	0.0004
Within-Condition FD	25.5505	35.5207	13.5237
Within-Condition MMD	0.0077	0.0075	0.0014
Within-Condition ICC	0.0617	0.0799	0.0694
NNPS [%]	44.626	59.621	47.037

The Within-Condition ICC of the real dataset is 0.066. A generative model that would perfectly generalize to the test set would have an NNPS of 50%. For all other distance metrics, lower is better.

reproduces both global and condition-specific distributions, including higher-order dependencies.

Conditional information content of synthetic fingerprints

We next assess whether the synthetic spectra correctly encode conditional information. To do so, we construct case-control designs that allow comparison of the conditional distributions between real and synthetic spectra. The conditions examined in this work are based on three covariates: age, sex, and BMI. Accordingly, we define the dichotomies male/female, young/old, and high/low BMI. To minimize potential confounding, case-control groups are formed through pairwise statistical matching of controls to cases based on age and sex whenever possible[54]. Specifically, for the sex condition, we compare males and females balanced on age. For age and BMI, we dichotomize the variables at $a_0 = 52$ years and $b_0 = 27$, respectively, and then construct balanced groups by matching on age and sex. The age and BMI thresholds were chosen to ensure that the corresponding label subgroups are balanced in size. The exact number of individuals per subgroup is reported in Table S2 (Supplementary).

After defining these groups, we first evaluate conditional alignment using the effect size defined in Equation (1), which quantifies differences between the two distributions. We compute and compare the effect size for both real and synthetic data. As shown in Fig. 3a-c, the effect size from real and synthetic spectra is closely aligned across the full spectral range for all three conditions. Alternatively, the Pearson correlation between covariates and each wavenumber can serve as an additional evaluation metric without relying on dichotomization of covariates. As shown in Figure S4, these results are consistent with the effect size analysis.

To further assess conditional consistency, we train binary classification models for each dichotomy using the same case-control groupings. This evaluation is performed within a cross-validation framework in which the test set is partitioned into six single-visit cohorts (SVCs). Each SVC contains at most one visit per participant while preserving the age, sex, and BMI distributions of the full test set (see Table 1 and Supplementary Figs. S1-S2). We train logistic regression classifiers (L2 penalty, $C = 10$) to predict sex and the binarized age and BMI labels, respectively. Further details on the classification setup are described in section D in the Supplementary Information. A baseline model is trained and evaluated on real data, and its performance is compared with models trained on synthetic data (CVAE, CBEGAN, CDIFF) and evaluated on real spectra. The objective of this analysis is not to maximize classification accuracy but to assess consistency: synthetic data should support analyses that yield the same substantive conclusions as those derived from real data. The resulting cross-validation Receiver Operating Characteristic (ROC) curves are shown in Fig. 3d-f. All synthetic-trained classifiers achieve ROC performance comparable to the

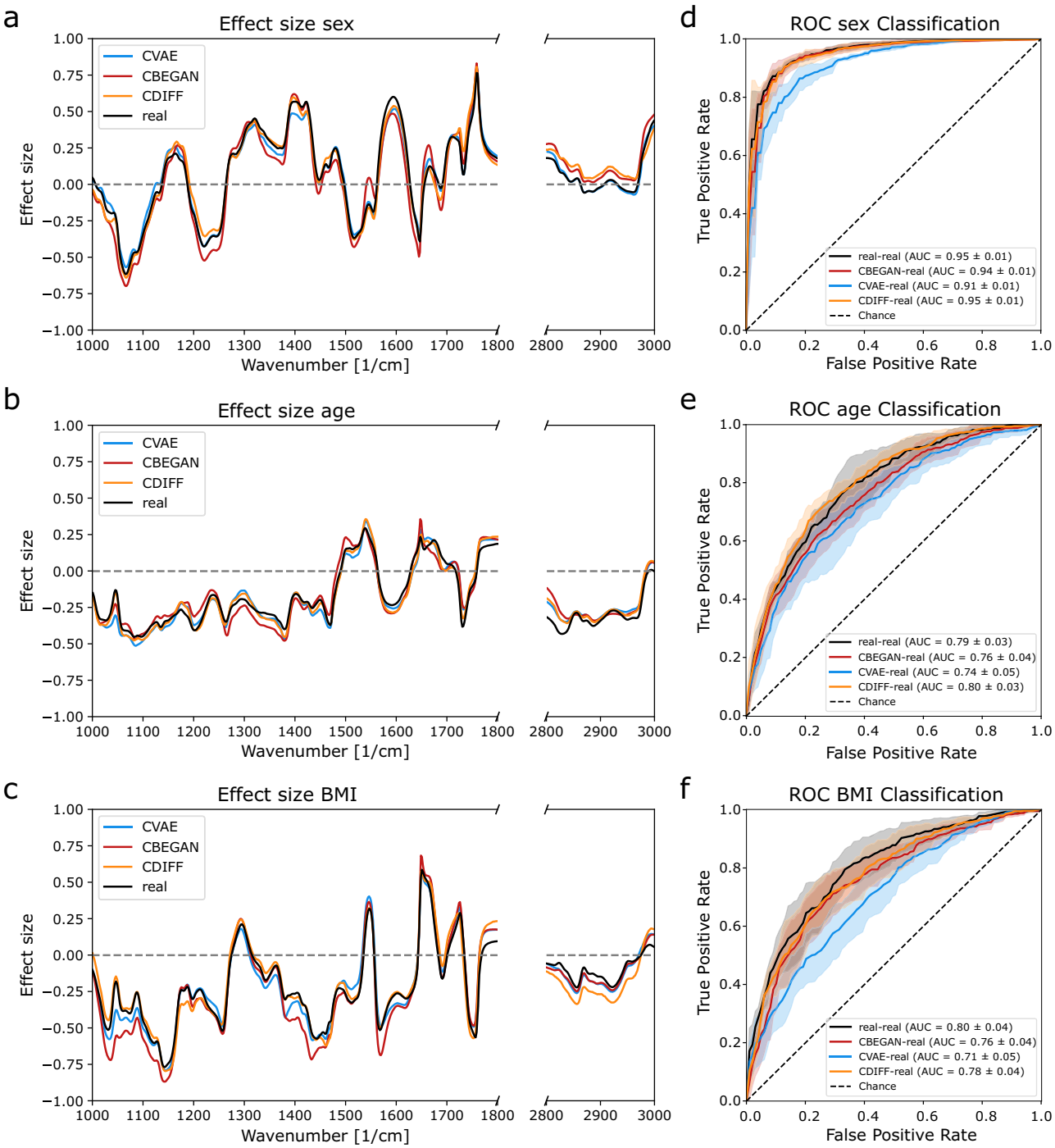

Fig. 3 | Conditional analysis. a–c Effect-size curves comparing positive and negative classes derived from binarized covariates (sex; age $\leq a_0$ vs. $> a_0$; BMI $\leq b_0$ vs. $> b_0$) for real and synthetic data. **d–f** ROC curves for classifiers trained on real spectra (black) and on synthetic spectra generated by the CVAE (blue), CBEGAN (red), or CDIFF (orange), with all models evaluated on real data. Error intervals for both ROCs and AUCs report ± 1 standard deviation. See the main text for details on label construction.

baseline, indicating preservation of the predictive signal. CDIFF exhibits the best overall alignment, followed by CBEGAN, and then CVAE. This ranking is consistent with the expected modeling characteristics of each architecture: diffusion models, widely regarded as state of the art for generative tasks, capture fine-grained distributional structure most faithfully; the CBEGAN's adversarial training enables it to reproduce sharper, condition-relevant spectral features; while the CVAE's smoother generative assumptions can temper subtle variations associated with the covariates. ROC-AUC values were compared using the existing SVC structure. For each classification task and each SVC, we computed the AUC of the classifier trained on real data and the AUC of each classifier trained on generated data, all evaluated on the same real SVC samples. Pairwise differences in AUC were then computed within each SVC. For each generated model, the six SVC-level AUC differences were tested against zero using a paired t-test. Results are reported as mean AUC difference across SVCs with 95% confidence intervals and nominal p-values. The results shown in Table 3 confirm that the AUCs obtained via CDIFF have no significant difference compared to those of the real data. CBEGAN only falls below $p < 0.05$ for the

Table 3 | AUC significance tests

Model	Covariate	Mean	SD	SE	95% CI	p
CVAE	age	−0.055	0.035	0.014	[−0.091, −0.018]	0.012
	BMI	−0.084	0.034	0.014	[−0.119, −0.049]	0.002
	sex	−0.031	0.012	0.005	[−0.043, −0.019]	0.001
CBEGAN	age	−0.030	0.032	0.013	[−0.064, 0.004]	0.076
	BMI	−0.037	0.028	0.012	[−0.067, −0.007]	0.024
	sex	−0.011	0.013	0.005	[−0.024, 0.002]	0.086
CDIFF	age	−0.006	0.018	0.007	[−0.025, 0.012]	0.429
	BMI	−0.007	0.016	0.007	[−0.024, 0.009]	0.315
	sex	0.001	0.005	0.002	[−0.005, 0.007]	0.743

Mean AUC difference (synthetic − real) across the six SVCs, with standard deviation, standard error, 95% confidence interval, and paired *t*-test *p*-value.

BMI classification, while the AUCs of CVAE significantly differ from those of the real dataset.

To test the sensitivity to other age and BMI thresholds, we provide complementary effect-sizes and ROC-AUC scores in Fig. S5 and Tables S3–S4. Alternatively, an elastic net can be fit to predict age and BMI directly (without dichotomies). Results are shown in Fig. S6S7S8 (Supplementary).

Applications in in-silico modeling

In the preceding sections, we demonstrated that the synthetic spectra not only reproduce the characteristics of real data but also accurately encode demographic and anthropometric information. This property enables the interpretation of synthetic spectra as originating from virtual individuals with well-defined attributes such as age, sex, and BMI. Consequently, these synthetic data can serve as a basis for in-silico studies that enable controlled, targeted interventions—for example, modeling the spectral consequences of aging or disease progression in a reproducible and individualized manner. In the following, we illustrate two exemplary in-silico applications: (i) simulation of healthy aging trajectories through controlled interventions, and (ii) augmentation of limited datasets to enhance classification performance.

Simulating healthy aging. Given the low intra-individual yet high inter-individual variability observed in IMF profiles[21,23], we investigated the potential of digital-twin-style conditional spectrum generation to model individual-specific aging trajectories. The CVAE framework facilitates this by encoding an individual's baseline spectrum and then modifying the concatenated condition vector to simulate gradual changes in covariates—such as progressive aging—while preserving individual-specific features. Figure 4a shows a two-dimensional UMAP (Uniform Manifold Approximation and Projection[55]) embedding of real spectra from the study cohort, revealing a distinct non-linear manifold along which chronological age increases. This manifold was identified using spectral regions with the highest correlation with age; see Fig. S9a (Supplementary). On this manifold, we projected synthetic spectra generated for predefined combinations of age, sex, and BMI. By systematically varying the age condition, we obtained continuous trajectories of synthetic spectra that trace movement along the cohort-level aging manifold. In a complementary analysis, when the cohort was mapped onto an alternative two-dimensional UMAP embedding constructed using features selected by the index of individuality, these same trajectories were found to cluster within the same local regions as the real visits of each corresponding individual (Fig. 4b). The spectral areas used for both embeddings are shown in Fig. S9c (Supplementary). Therefore, these trajectories demonstrate that the CVAE captures both global aging trends and individual-specific spectral characteristics. This allows it to reproduce aging pathways consistent with cohort-level aging-associated spectral trends. Extending this approach to include additional clinical covariates could enable the joint modeling of healthy and pathological aging, supporting in-silico experiments that disentangle age-related and disease-related spectral effects or simulate alternative clinical scenarios.

Targeted augmentation of underrepresented populations. As a second application, we examined whether conditionally generated spectra can enhance statistical power in classification tasks, particularly when working with limited or imbalanced datasets. Specifically, we evaluated BMI classification across four categories: underweight (< 18.5), normal (18.5–25.0), overweight (25.0–30.0), and obese (> 30.0)[56]. Due to limited sample sizes, the underweight and normal categories were combined into a single class. Figure 4c presents results from a three-class classification experiment conducted under four conditions: (i) a baseline model trained on the full real dataset, (ii) a model trained on a dataset with strong class imbalance, and (iii-v) three augmented versions in which the class imbalance was resolved by adding spectra generated by the CVAE, CBEGAN, and CDIFF models, respectively. All analyses were performed using cross-validation across SVCs. A logistic regression classifier achieved an overall accuracy of 59% on the baseline real data. Artificially reducing the representation of the Normal and Obese classes by 90% led to substantial performance degradation, with predictions collapsing toward the dominant Overweight class. Augmentation with conditionally generated spectra restored balanced class representation and improved classification accuracy, approaching the baseline level with the CBEGAN and CVAE, while moderately improving with the CDIFF. Additional analyses—including comparisons with standard oversampling methods such as random oversampling and SMOTENC, as well as analogous experiments for age classification—are provided in Supplementary Fig. S10. Collectively, these results highlight the potential of synthetic data to stabilize model performance in underrepresented subgroups under the evaluated experimental conditions.

Discussion

This work demonstrates that conditional deep generative models can synthesize IMF spectra that closely resemble real measurements, preserve conditional structure for age, sex, and BMI, and enable controlled in-silico analyses. These models act as simulators of "virtual individuals," supporting cohort balancing, power scaling, and digital-twin-style trajectory studies without additional sample collection. All our generative models reproduced key spectral and feature-level properties of real data and supported classifiers that generalized to true measurements. Notably, the CVAE's explicit latent representation provides an advantage in digital-twin-style conditional generation, where smooth manipulation of latent codes enables individualized trajectories that remain anchored to individual-specific characteristics. Conditioning allows the synthetic spectra to be interpreted in relation to individual-level attributes, enabling reproducible case-control

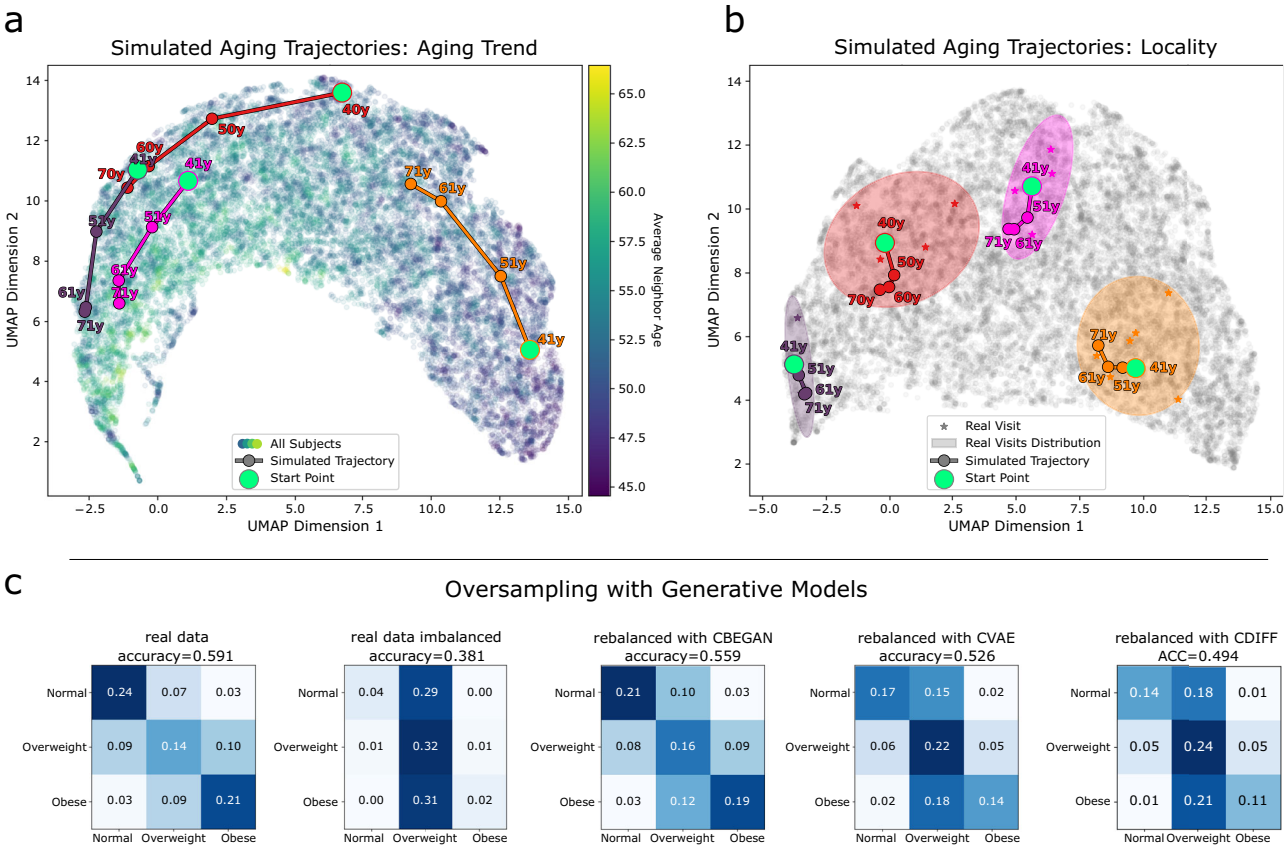

Fig. 4 | 2D UMAP embeddings for four individuals with simulated aging in four 10-year steps. Initial spectra are averaged across the first four short-interval visits. **a** Embeddings using 30 age-correlated wavenumbers show clear aging trajectories; the background is colored by the average age of the 20 nearest neighbors. **b** Embeddings using 30 wavenumbers with the highest Index of Individuality highlight individual-specific clustering. **c** Multi-class BMI classification: baseline; undersampled (Normal/Obese reduced 90%); and rebalanced with CVAE/CBEGAN/CDIFF augmentation.

Table 4 | Qualitative comparison of the three conditional generative models

Criterion	CVAE	CBEGAN	CDIFF
Distributional fidelity (FD, MMD)	low	low	high
Conditional quality (classifier transfer)	low	moderate	high
Latent control / digital twins	high	n.a.	n.a.
Sampling efficiency	high	moderate	low
Extensibility toward causal modeling	moderate	low	moderate

Ratings denote relative qualitative rankings among the three models: high > moderate > low. "n.a." indicates that the corresponding capability is not supported by the architecture under consideration. The qualitative ratings are derived based on the metrics presented in Tables 2–3 and Figs. 2–4.

designs and individualized aging simulations. Table 4 summarizes these relative strengths of the three models across these axes. It indicates that no single architecture dominates across all axes, and that the appropriate choice depends on the intended application. The CDIFF achieved the most faithful reproduction of both the global and condition-specific data distributions (lowest FD and MMD; Table 2) together with the closest classifier transfer to real data, yielding AUC differences that were statistically indistinguishable from zero (Table 3). These advantages, however, come at the cost of comparatively expensive iterative sampling and the absence of a directly manipulable latent representation. The CVAE occupies the opposite end of this trade-off: its smoothing tendency produces the lowest distributional fidelity and the largest classifier-transfer gaps, yet it is the only one of the three architectures to provide an explicit, continuous latent space and single-pass sampling, which makes it uniquely suited to digital-twin-style interpolation and large-scale generation. The CBEGAN is intermediate, offering better conditional transfer than the CVAE but weaker distributional agreement than the CDIFF.

A recurring observation across our analyses is that the CBEGAN supports slightly better classifier transfer than the CVAE despite poorer agreement with the overall data distribution. We hypothesize that this reflects a known consequence of the two training objectives: the CVAE optimizes a Gaussian reconstruction (MSE) likelihood and averages over its posterior, which attenuates fine, locally contrasted spectral features, whereas the adversarial reconstruction objective of the CBEGAN penalizes such blurring and rewards the reproduction of sharper, condition-relevant edges that linear classifiers can readily exploit. Sharper, however, does not necessarily mean more realistic, and our data support this caveat: the CBEGAN simultaneously exhibits the poorest global distributional agreement of the three models (highest global FD and MMD) and the largest within-condition FD (Table 2), occasionally exceeding the 20% effect-size threshold at the distributional level.

Despite these promising results, several limitations remain. The models were trained on healthy individuals within a single measurement pipeline, and their generalizability to other instruments, laboratories, cohorts, or disease states has not yet been established. Cross-domain robustness is therefore an important direction for future work. In this regard, dedicated domain-adaptation strategies for blood-based IR spectroscopy have recently been developed, including data-augmentation-based harmonization

approaches[57] and Domain-Adversarial Neural Network-based cross-instrument harmonization[58]. The latter is particularly promising because it learns to map measurements acquired on different devices into a shared spectral domain. Thus, for future cross-domain applications, the intended strategy is not necessarily to make the generative models themselves directly invariant across devices or laboratories, but rather to first harmonize spectra into a common domain using dedicated domain-adaptation methods. Conditional generative modeling can then be applied within this shared domain.

Another limitation is that conditioning is currently limited to three covariates, which may lead to residual confounding. Extending the conditioning space to include routine clinical laboratory parameters—such as lipid panels, inflammatory markers, glucose or insulin levels, and blood counts—could address several of these limitations. Incorporating such biochemical information would reduce confounding by capturing key biological axes linked to IMF spectra, enabling the generation of under-represented clinical subgroups, and improving interpretability by establishing direct relationships between spectral regions and laboratory measures (for example, lipid-sensitive bands versus LDL/HDL levels). These additional covariates can be implemented through continuous encodings, multi-branch conditional networks, and missingness-aware gating mechanisms to ensure flexibility and scalability. Beyond adding covariates, emerging multimodal foundation-model frameworks enable generating biospectral data alongside other modalities—including omics profiles, vital signs, medical imaging, and electronic health records—thereby embedding IMF synthesis within a broader cross-modality generative ecosystem.

Looking ahead, a major step will be the transition from conditional to causal generative modeling[59–64]. While conditional models learn $p(\mathbf{x}|\mathbf{y})$ and support conditional simulation of spectra, they do not capture interventional distributions $p(\mathbf{x}|do(\mathbf{y}))$. Causal generative models could disentangle the effects of aging and disease, prevent implausible manipulations of correlated covariates, and simulate true counterfactual outcomes, such as lifestyle or medication changes. This transition requires specifying a structural causal model (SCM) over demographic, clinical, and latent variables; constraining decoder dependencies to causal parents (e.g., through masked conditioners or modular decoders); exploiting longitudinal and negative-control data for identifiability; and validating predictions against observed pre-post changes. Bayesian or ensemble-based decoders could further quantify uncertainty in interventional scenarios.

With these extensions, synthetic IMF data could evolve from a data-augmentation tool into a platform for controlled in-silico phenotyping. Such models would enable virtual cohort simulations, systematic stress-testing of analytical pipelines, and exploration of how modifiable risk factors influence spectral fingerprints. Expanding the conditioning space, integrating multimodal foundation-model paradigms, and introducing causal modeling directly address current limitations and represent key steps toward trustworthy, actionable synthetic cohorts in precision health.

Methods

Sample collection

This study drew on data from a large-scale longitudinal cohort consisting of participants who self-reported as healthy at the time of recruitment (study code: H4H_HU_2020_Sample Collection; ethics approval reference: 2754-11/2020/EÜIG). Informed consent was obtained from all participants prior to blood draw. The study was conducted in accordance with the principles of the Declaration of Helsinki. As part of the study protocol, IMFs were acquired at the laboratory of the Center of Molecular Fingerprinting in Szeged, Hungary.

Sample handling and FTIR measurements

As part of this study, EDTA plasma samples were collected at 20 study centers across Hungary, following harmonized protocols for blood collection and processing. The resulting plasma was divided into 8–10 aliquots of 0.5 mL each and stored at −80 °C. A second aliquotation step was performed centrally in Szeged, Hungary: one 0.5 ml aliquot per sample was thawed, gently mixed for 30 seconds, and centrifuged. From this, four 90 µl measurement aliquots were prepared and subsequently refrozen at −80 °C. For the FTIR measurements, these aliquots were thawed again, resulting in two freeze-thaw cycles before FTIR analysis.

All FTIR measurements were conducted in a fully randomized and blinded manner. The operators performing the measurements had no access to clinical information or sample identities. A commercial FTIR instrument specifically designed for liquid samples in transmission mode was used for all analyses (MIRA Analyzer, version MA7; CLADE GmbH, formerly Micro-Biolytics GmbH).

Spectra were recorded using a flow-through transmission cuvette with CaF_2 windows and an 8 μm optical path length. Data were acquired at a resolution of 4 cm^{-1} across a spectral range of 930-3050 cm^{-1}. A water reference spectrum was automatically collected after each sample to reconstruct absorbance spectra. Although the manufacturer's algorithm for this correction is not disclosed, it is known to overcompensate for highly concentrated samples like human plasma. The effective path length was also determined automatically and used for spectral adjustment. The MIRA-Analyzer does not provide access to raw measurements but only to the water-compensated absorbance spectra.

Samples were processed in batches of 28, with $DMSO_2$, water, and pooled human serum (BioWest, Nuaillé, France) included as quality controls, yielding a total of 34 samples per batch. The BioWest quality control samples enabled tracking of analytical variation and instrument stability over time.

Data preparation and preprocessing

Measured spectra were truncated to 1000 cm^1 to 3000 cm^1 to remove noisy regions caused by the low transmittance of CaF_2 below 950 cm^{-1} and the high water absorption above 3000 cm^{-1}. In addition, the uninformative region (so-called silent region) between 1800 cm^1 and 2800 cm^1 was removed. We finally applied ℓ_2 normalization to the truncated spectra to standardize the scale across samples and scaled the retained absorbance values of each wavenumber by standard scaling (z-scoring) fit on the train data set to normalize the feature distributions. This pipeline follows prior work[21,24,29,30,36,57].

Covariates were preprocessed separately. We apply min-max scaling to the covariates (age and BMI), as these variables are naturally bounded and do not exhibit extreme outliers in our dataset, while sex was encoded $\in \{0, 1\}$. All three were then mapped to d-dimensional positional encodings[49,65,66]:

$$\mathrm{PE}(pos, 2i) = \sin\left(\frac{pos}{B^{2i/d}}\right), \quad \mathrm{PE}(pos, 2i+1) = \cos\left(\frac{pos}{B^{2i/d}}\right), \tag{3}$$

with $i \in \{0, \ldots, d/2\}$ and B as the frequency base. Because age and BMI were min-max scaled to [0, 1], we used $B = 0.1$, yielding a frequency range better matched to the dynamic range of the covariates. This fixed, non-learned encoding supports unseen continuous values (e.g., new ages/BMIs) and provides a richer input than scalars. Python implementations used `PyTorch`.

Modeling background

Let $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ denote spectra drawn from unknown $p(\mathbf{x})$. Generative models learn a mapping from a simple prior (e.g., Gaussian) to the data space, approximating $p(\mathbf{x})$[67]. We consider conditional variants of the VAE[47,48], BEGAN[45,46], and diffusion model[49] that incorporate covariates $\mathbf{y}$. All models were trained on an NVIDIA A40 GPU.

Conditional VAE

The CVAE comprises an encoder and decoder (Fig. 5). The inference factorizes as

$$q_\phi(\mathbf{x}, \mathbf{y}, \mathbf{z}) = q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})\, q(\mathbf{x}, \mathbf{y}), \tag{4}$$

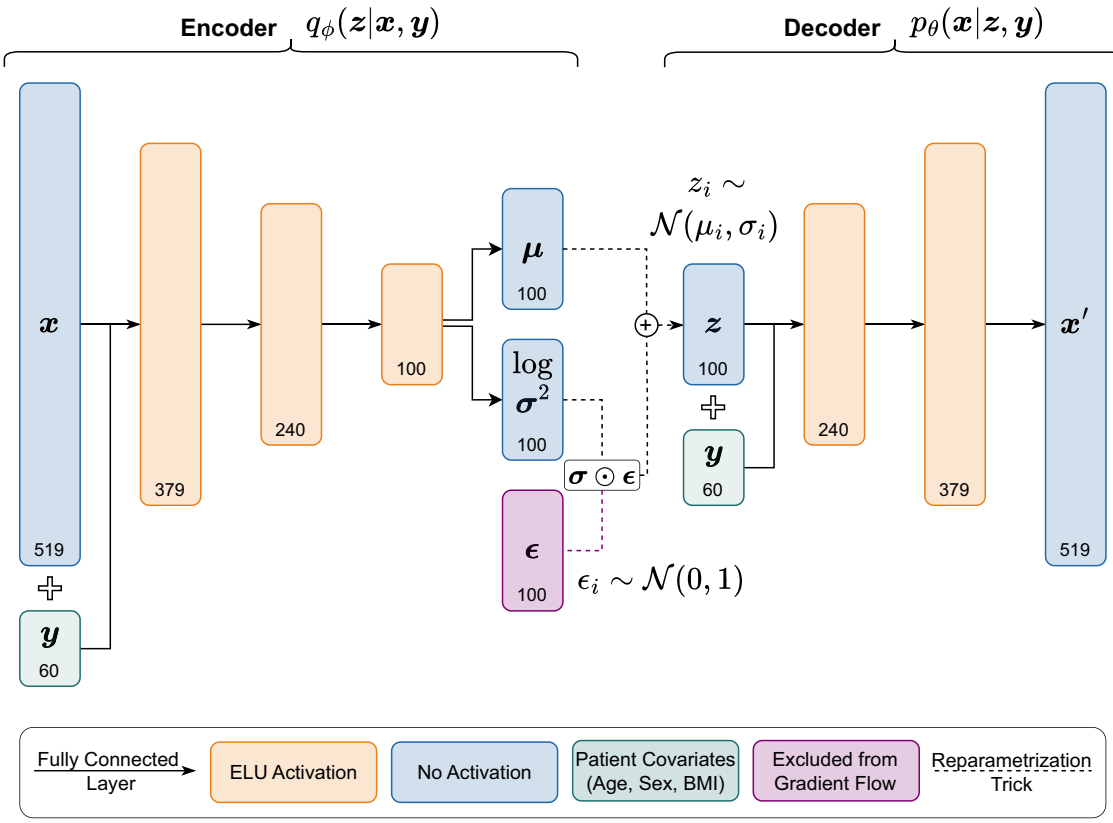

Fig. 5 | Schematic of the CVAE architecture. The encoder outputs parameters of $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})$; reparameterized samples are decoded to $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})$.

and generation as

$$p_\theta(\mathbf{x}, \mathbf{y}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})\, p(\mathbf{z}) p(\mathbf{y}), \tag{5}$$

with Gaussian $p(\mathbf{z})$. The ELBO is

$$\mathcal{L}_{\text{CVAE}} = \alpha\ \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\big[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{y})\big] - \beta\ \text{KL}\left(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y})\ \|\ p(\mathbf{z})\right), \tag{6}$$

reducing to MSE reconstruction and closed-form KL under diagonal Gaussians.

Mitigating latent space collapse. To avoid degenerate solutions with negligible KL and uninformative $\mathbf{z}$[68], we use cyclic annealing of $\beta \in [0, 1]$ with a logistic schedule. Early epochs emphasize reconstruction (small β); later epochs regularize the latent space (larger β), alternating to balance both objectives. See training curves in Figure S11a (Supplementary).

Training CVAE: hyperparameters. We trained for 500 epochs with Adam (10^{-3} learning rate), $\alpha = 10$, cyclic β with 50-epoch period, and positional encoding dimension $d = 20$. Encoder/decoder: three linear layers; the encoder branches to mean and log-variance heads. ELU activations except for mean/log-variance and decoder output. Batch size: 524. Hyperparameters were tuned with `optuna`[69].

Conditional BEGAN

The CBEGAN comprises a generator and discriminator (Fig. 6). Given $(\mathbf{x}, \mathbf{y}) \sim q(\mathbf{x}, \mathbf{y})$ and $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the generator $G_\theta(\mathbf{n}|\mathbf{y})$ produces $\mathbf{x}'$. The discriminator $D_\psi(\cdot\,|\mathbf{y})$ is an autoencoder with reconstruction loss

$$\mathcal{L}_{\text{CAE}}(\mathbf{v}|\mathbf{y}) = \|\ \mathbf{v} - D_\psi(\mathbf{v}|\mathbf{y})\|_1, \quad \mathbf{v} \in \{\mathbf{x}, \mathbf{x}'\}. \tag{7}$$

CBEGAN enforces the matching of real and generated reconstruction-loss distributions by updating

$$\mathcal{L}_D = \mathcal{L}_{\text{CAE}}(\mathbf{x}|\mathbf{y}) - k_t\, \mathcal{L}_{\text{CAE}}(G_\theta(\mathbf{n}|\mathbf{y})), \tag{8}$$

$$\mathcal{L}_G = \mathcal{L}_{\text{CAE}}(G_\theta(\mathbf{n}|\mathbf{y})), \tag{9}$$

$$k_{t+1} = k_t + \lambda_k\big(\gamma\, \mathcal{L}_{\text{CAE}}(\mathbf{x}|\mathbf{y}) - \mathcal{L}_{\text{CAE}}(G_\theta(\mathbf{n}|\mathbf{y}))\big), \tag{10}$$

with diversity ratio $\gamma \in [0, 1]$[45,46].

Training CBEGAN: hyperparameters. We trained for 500 epochs with Adam (initial learning rate 10^{-4}, cosine annealed to 10^{-6})[70,71]. Following[42], the discriminator was pretrained for 10 epochs. We set $\lambda_k = 0.001$ and $\gamma = 0.9$. Batch size: 32. Hyperparameters were tuned with `optuna`[69]. Training curves are in Figure S11b (Supplementary).

Conditional diffusion model

CDIFF is based on a denoising diffusion probabilistic model with classifier-free guidance[50]. It consists of a noise-conditioned network $\epsilon_\theta(\mathbf{x}_t, \mathbf{y}, t)$ that predicts the Gaussian noise added to data at diffusion step t (Fig. 7). The forward (diffusion) process progressively corrupts data $\mathbf{x}_0 \sim q(\mathbf{x}|\mathbf{y})$ into noise via

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1 - \bar{\alpha}_t)\mathbf{I}\right), \tag{11}$$

where $\{\bar{\alpha}_t\}_{t=1}^T$ is a predefined variance schedule derived from a cosine schedule. Sampling from this process can be written as

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{12}$$

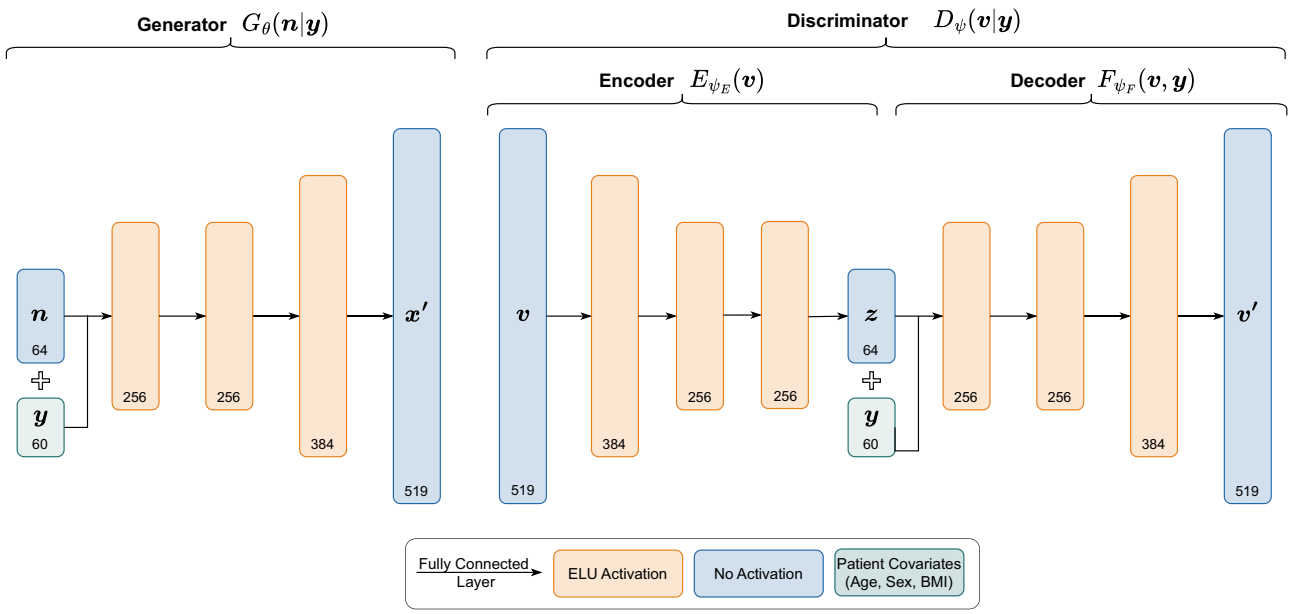

Fig. 6 | Schematic of the CBEGAN architecture. $G_\theta(\mathbf{n}|\mathbf{y})$ generates $\mathbf{x}'$. $D_\psi(\cdot\,|\mathbf{y})$ reconstructs both $\mathbf{x}$ and $\mathbf{x}'$; the reconstruction error drives adversarial training.

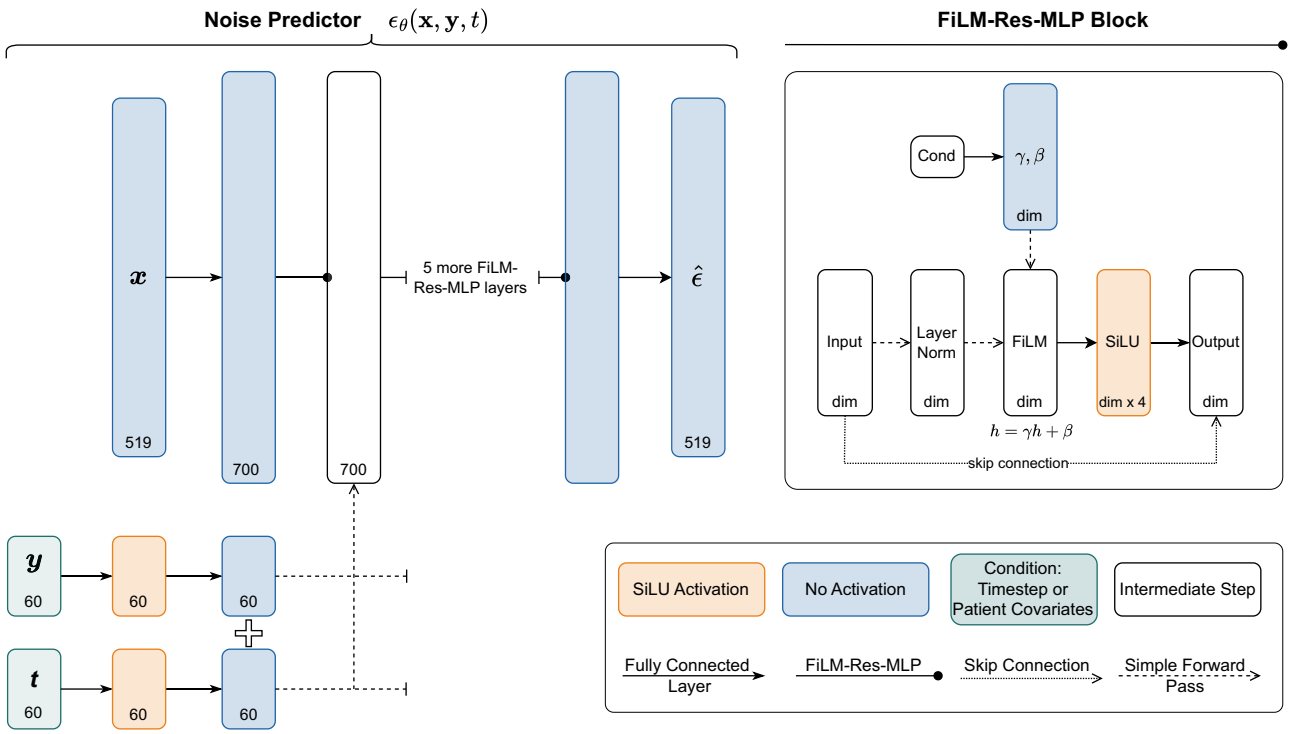

Fig. 7 | Schematic of the CDIFF architecture. The noise predictor $\epsilon_\theta(x, y, t)$ predicts the noise that was added to x at timestep t given conditioning vector y. The noise predictor consists of two fully multilayer perceptron (MLP) layers and six residual MLP layers with Feature-wise Linear Modulation (FiLM).

The model is trained to predict $\boldsymbol{\epsilon}$ from $(\mathbf{x}_t, \mathbf{y}, t)$ using a reweighted mean squared error objective

$$\mathcal{L}_{\text{CDiff}} = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[w_t \parallel \boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, \mathbf{y}, t)\parallel_2^2\right], \tag{13}$$

where w_t is a signal-to-noise ratio (SNR)-based weighting term. Specifically, using $\text{SNR}_t = \bar{\alpha}_t/(1-\bar{\alpha}_t)$, we apply the min-SNR weighting[72]

$$w_t = \frac{\min(\text{SNR}_t, 5)}{\text{SNR}_t}, \tag{14}$$

which stabilizes training by down-weighting high-SNR (low-noise) timesteps. After training the sampling was performed using DDPM[49].

Classifier-free guidance. To enable conditional generation without an explicit classifier, we employ classifier-free guidance by randomly dropping conditioning labels during training with probability p_{uncond}. At inference, conditional and unconditional predictions are combined as

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\boldsymbol{x_t}, \varnothing, t) + \gamma\big(\boldsymbol{\epsilon}_\theta(\boldsymbol{x_t}, \boldsymbol{y}, t) - \boldsymbol{\epsilon}_\theta(\boldsymbol{x_t}, \varnothing, t)\big), \tag{15}$$

where γ is the guidance scale controlling the strength of conditioning and $\varnothing$ indicates unconditional prediction. The concrete choice for γ was based on Fig. S3d.

Architecture. The noise predictor ϵ_θ is implemented as a multi-layer perceptron with residual connections and feature-wise linear modulation (FiLM). Inputs consist of: (i) the noisy spectrum $\mathbf{x}_t$, (ii) a sinusoidal embedding of timestep t, and (iii) a positional encoding of conditioning variables $\mathbf{y}$. These embeddings are combined and injected into the network via FiLM conditioning[73].

Training CDIFF: hyperparameters. We trained for 553 epochs (early-stopping with a patience of 50 epochs) using AdamW with learning rate 8×10^{-4}, weight decay 10^{-3}, and cosine learning rate scheduling with linear warmup (500 steps). The diffusion process used $T = 128$ timesteps with a cosine variance schedule, details on which are shown in Fig. S3. The network employed six FiLM-residual layers with hidden dimension 700, timestep embedding dimension 42, and positional encoding dimension 24. Classifier-free guidance used $p_{\text{uncond}} = 0.2$. Exponential moving average (EMA) of parameters was applied with decay 0.9. Batch size was 128. Hyperparameters were tuned with `optuna`[69]. Training curves are shown in Figure S11c (Supplementary).

Neighbor analysis for assessing risk of privacy leakage

For each sample in each test set $\hat{X}_i$ with $i \in \{\text{real, CVAE, CBEGAN, CDIFF}\}$, we calculated the Euclidean distance to the nearest neighbor in the training set X_{train}. For a sample $\mathbf{x} \in \hat{X}_i$, the nearest-neighbor distance is defined as

$$d_{\text{NN}}(\mathbf{x}, X_{\text{train}}) = \min_{\mathbf{z} \in X_{\text{train}}} \|\mathbf{x} - \mathbf{z}\|_2. \tag{16}$$

$\{d_{\text{NN}}(\mathbf{x}, X_{\text{train}})\}_{\mathbf{x} \in \hat{X}_i}$ for each dataset. The same procedure was applied to samples from the held-out real test set, producing the reference distribution D_{real}.

To assess the extent to which synthetic samples lie unusually close to training examples, we computed the NNPS. Following previous work on privacy assessment in generative models, NNPS is defined as the fraction of synthetic samples whose nearest-neighbor distance to the training set is smaller than the corresponding distance observed for real test samples[53]:

$$\text{NNPS}_i = \frac{1}{N} \sum_{n=1}^{N} \left(d_{\text{NN}}^{(\text{real},n)} > d_{\text{NN}}^{(i,n)} \right), \tag{17}$$

where N denotes the number of samples. An NNPS value of 0.5 indicates that synthetic samples are, on average, no closer to the training data than real held-out samples. Values substantially larger than 0.5 suggest that generated samples tend to be closer to training observations, which may indicate memorization and an elevated risk of privacy leakage. Conversely, values below 0.5 indicate that generated samples are generally farther from the training set than real test samples.

Uncited figures/tables

Label missing in caption for figures/Tables: apx:classification_tasks

Data availability

All rights to the blood measurements used in this study are reserved by the Center for Molecular Fingerprinting Research Nonprofit LCC (company registration number: 01-09-344208). Data may be requested by contacting the corresponding author. Python implementations and trained weights for CBEGAN and CVAE are available on GitHub (https://github.com/Data-Science-Group-Attoworld/in-silico-infrared-modeling).

Code availability

Python implementations and trained weights for CVAE, CBEGAN and CDIFF are available on GitHub[74] (https://github.com/Data-Science-Group-Attoworld/in-silico-infrared-modeling).

Received: 7 January 2026; Accepted: 29 August 2026;
Published online: 17 September 2026

References

1. Chen, R. et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).
2. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
3. Price, N. D. et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nat. Biotechnol.* **35**, 747–756 (2017).
4. Piening, B. D. et al. Integrative personal omics profiles during periods of weight gain and loss. *Cell Syst.* **6**, 157–170 (2018).
5. Schüssler-Fiorenza Rose, S. M. et al. A longitudinal big data approach for precision health. *Nat. Med.* **25**, 792–804 (2019).
6. Ahadi, S. et al. Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nat. Med.* **26**, 83–90 (2020).
7. Earls, J. C. et al. Multi-omic biological age estimation and its correlation with wellness and disease phenotypes: a longitudinal study of 3,558 individuals. *J. Gerontol. Ser. A* **74**, S52–S60 (2019).
8. Hou, Y.-C. C. et al. Precision medicine integrating whole-genome sequencing, comprehensive metabolomics, and advanced imaging. *Proc. Natl. Acad. Sci. USA* **117**, 3053–3062 (2020).
9. Tebani, A. et al. Integration of molecular profiles in a longitudinal wellness profiling cohort. *Nat. Commun.* **11**, 4487 (2020).
10. Chen, X. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nat. Commun.* **11**, 3475 (2020).
11. Topol, E. J. Individualized medicine from prewomb to tomb. *Cell* **157**, 241–253 (2014).
12. Mason, C. E., Porter, S. G. & Smith, T. M. Characterizing multi-omic data in systems biology. *Systems analysis of human multigene disorders* 15–38 (2013).
13. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).
14. Baker, M. J. et al. Using Fourier transform IR spectroscopy to analyze biological materials. *Nat. Protoc.* **9**, 1771–1791 (2014).
15. Butler, H. J. et al. Development of high-throughput ATR-FTIR technology for rapid triage of brain cancer. *Nat. Commun.* **10**, 4501 (2019).
16. Voronina, L. et al. Molecular origin of blood-based infrared spectroscopic fingerprints. *Angew. Chem.* **60**, 17060–17069 (2021).
17. Paraskevaidi, M. et al. Clinical applications of infrared and Raman spectroscopy in the fields of cancer and infectious diseases. *Appl. Spectrosc. Rev.* **56**, 804–868 (2021).
18. Li, Q. et al. Review of spectral imaging technology in biomedical engineering: achievements and challenges. *J. Biomed. Opt.* **18**, 100901–100901 (2013).
19. Li, Q. et al. Leukocyte cells identification and quantitative morphometry based on molecular hyperspectral imaging technology. *Computerized Med. Imaging Graph.* **38**, 171–178 (2014).
20. Wang, Q. et al. A 3d attention networks for classification of white blood cells from microscopy hyperspectral images. *Opt. Laser Technol.* **139**, 106931 (2021).
21. Huber, M. et al. Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nat. Commun.* **12**, 1511 (2021).
22. Zarandy, Z. I. et al. Individuality and information content of infrared molecular profiles: insights from a large longitudinal health-profiling study. *bioRxiv* 2026–04 (2026).

23. Kepesidis, K. V. et al. Integration of infrared molecular fingerprinting data in a longitudinal health profiling cohort. In *International Conference on Artificial Intelligence in Medicine*, 180–190 (Springer, 2025).
24. Huber, M. et al. Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *Elife* **10**, e68758 (2021).
25. Martin, F. L. et al. Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nat. Protoc.* **5**, 1748–1760 (2010).
26. Ghimire, H. et al. Protein conformational changes in breast cancer sera using infrared spectroscopic analysis. *Cancers* **12**, 1708 (2020).
27. Ollesch, J. et al. An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma. *Biomed. Spectrosc. Imaging* **5**, 129–144 (2016).
28. Eissa, T. et al. Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. *Cell Reports Medicine* **5** (2024).
29. Kepesidis, K. V. et al. Assessing lung cancer progression and survival with infrared spectroscopy of blood serum. *BMC Med.* **23**, 101 (2025).
30. Kepesidis, K. V. et al. Breast-cancer detection using blood-based infrared molecular fingerprints. *BMC Cancer* **21**, 1–9 (2021).
31. Backhaus, J. et al. Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vibrational Spectrosc.* **52**, 173–177 (2010).
32. Elmi, F., Movaghar, A. F., Elmi, M. M., Alinezhad, H. & Nikbakhsh, N. Application of ft-ir spectroscopy on breast cancer serum analysis. *Spectrochim. Acta Part A Mol. Biomolecular Spectrosc.* **187**, 87–91 (2017).
33. Ollesch, J. et al. It's in your blood: spectral biomarker candidates for urinary bladder cancer from automated FTIR spectroscopy. *J. Biophotonics* **7**, 210–221 (2014).
34. Anderson, D., Anderson, R., Moug, S. & Baker, M. Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review. *BJS Open* **4**, 554–562 (2020).
35. Gigou, L., Kepesidis, K. V. & Krausz, F. Towards precision medicine with infrared molecular profiles: Identifying and explaining subgroups. In *International Conference on Artificial Intelligence in Medicine*, 176–180 (Springer, 2025).
36. Wegner, C. et al. Toward informative representations of blood-based infrared spectra via unsupervised deep learning. *J. Biophotonics* e70011 (2025).
37. Zigman, M. et al. 90p infrared molecular fingerprinting: A new in vitro diagnostic platform technology for cancer detection in blood-based liquid biopsies. *Ann. Oncol.* **33**, S580 (2022).
38. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. & Popp, J. Sample size planning for classification models. *Analytica Chim. acta* **760**, 25–33 (2013).
39. Leopold-Kerschbaumer, N., Feiler, N. & Kepesidis, K. Synthetic blood-based infrared molecular fingerprints: artificial cohorts for methodological research. *Analytical Methods* (2026).
40. Leopold-Kerschbaumer, N., Nemeth, F. B., Feiler, N. & Kepesidis, K. V. Multivariate Gaussian modeling of blood-based FTIR spectroscopy: framework and applications. In *Data Science for Photonics and Biophotonics II*, vol. 14099, 14 (SPIE, 2026).
41. Eissa, T., Kepesidis, K. V., Zigman, M. & Huber, M. Limits and prospects of molecular fingerprinting for phenotyping biological systems revealed through in silico modeling. *Anal. Chem.* **95**, 6523–6532 (2023).
42. Zhu, D. et al. Synthetic spectra generated by boundary equilibrium generative adversarial networks and their applications with consensus algorithms. *Opt. Express* **28**, 17196–17208 (2020).
43. Luo, X.-Y., Fan, X.-R., Zhang, X.-M., Chen, T.-Y. & Huang, C.-J. AE-began-based synthetic data augmentation for sample-limited high-dimensional problems with application to nir spectral data. In *Journal of Physics: Conference Series*, vol. 2594, 012029 (IOP Publishing, 2023).
44. Gonzalez, C. et al. Anomaly detection in Fourier transform infrared spectroscopy of geological specimens using variational autoencoders. *Ore Geol. Rev.* **158**, 105478 (2023).
45. Berthelot, D., Schumm, T. & Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).
46. Marzouk, A., Barros, P., Eppe, M. & Wermter, S. The conditional boundary equilibrium generative adversarial network and its application to facial attributes. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–7 (IEEE, 2019).
47. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
48. Yan, X., Yang, J., Sohn, K. & Lee, H. Attribute2image: Conditional image generation from visual attributes. In *European conference on computer vision*, 776–791 (Springer, 2016).
49. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
50. Ho, J. & Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).
51. Sawilowsky, S. S. New effect size rules of thumb. *J. Mod. Appl. Stat. methods* **8**, 26 (2009).
52. Cohen, J. *Statistical power analysis for the behavioral sciences* (Routledge, 2013).
53. Alaa, A., Van Breugel, B., Saveliev, E. S. & Van Der Schaar, M. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, 290–306 (PMLR, 2022).
54. Rosenbaum, P. R., Rosenbaum, P. B. & Briskman. *Design of observational studies*, vol. 10 (Springer, 2010).
55. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
56. Organization, W. H. et al. Surveillance of chronic disease risk factors: country level data and comparable estimates. In *Surveillance of chronic disease risk factors: country level data and comparable estimates* (2005).
57. Nemeth, F. B. et al. Bridging spectral gaps: Cross-device model generalization in blood-based infrared spectroscopy. *Analytical Chem.* (2025).
58. Németh, F. B. & Kepesidis, K. Domain-adversarial neural networks for cross-instrument harmonization in blood-based infrared spectroscopy. In *Data Science for Photonics and Biophotonics II*, vol. 14099, 41 (SPIE, 2026).
59. Sanchez, P. et al. Causal machine learning for healthcare and precision medicine. *R. Soc. Open Sci.* **9**, 220638 (2022).
60. Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J. & Silva, R. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475* (2022).
61. Schölkopf, B. et al. Toward causal representation learning. *Proc. IEEE* **109**, 612–634 (2021).
62. Pawlowski, N., Coelho de Castro, D. & Glocker, B. Deep structural causal models for tractable counterfactual inference. *Adv. Neural Inf. Process. Syst.* **33**, 857–869 (2020).
63. Sanchez, P. & Tsaftaris, S. A. Diffusion causal models for counterfactual estimation. *arXiv preprint arXiv:2202.10166* (2022).
64. Yang, M. et al. CausalVAE: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697* (2020).
65. Vaswani, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems* **30** (2017).
66. Mildenhall, B. et al. NeRF: representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**, 99-106 (2021).
67. Foster, D.*Generative deep learning* (" O'Reilly Media, Inc.", 2022).

68. Fu, H. et al. Cyclical annealing schedule: A simple approach to mitigating KL vanishing. *arXiv preprint arXiv:1903.10145* (2019).
69. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631 (2019).
70. Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
71. Loshchilov, I. & Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
72. Hang, T. et al. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7441–7451 (2023).
73. Perez, E., Strub, F., De Vries, H., Dumoulin, V. & Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).
74. Moritz Jung, Timo Halenke, Selina Süzeroglu, Niklas Leopold-Kerschbaumer. in-silico-infrared-modeling. https://github.com/Data-Science-Group-Attoworld/in-silico-infrared-modeling (2025).

Acknowledgements

We thank all members of the Center for Molecular Fingerprinting (CMF) for fostering a research environment that made this work possible. We are grateful to Diána Debreceni, Csilla Lakatos, Dóra Nagy, Kamilla Mileant, and Ágnes Majárné Kovács for assistance with FTIR measurements in the H4H study. KVK acknowledges insightful discussions with Lea Gigou and Andreas Döpp. This work is part of project no. 2020-2.1.1-ED-2022-00213. Project no. 2020-2.1.1-ED-2022-00213 has been implemented with support from the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the 2020-2.1.1-ED funding scheme.

Author contributions

Niklas Leopold-Kerschbaumer: Investigation; Formal analysis; Visualization; Methodology; Writing—original draft; Writing—review & editing. Timo Halenke: Investigation; Formal analysis; Visualization; Methodology; Writing—original draft; Writing—review & editing. Selina Süzeroğlu: Investigation; Formal analysis; Visualization; Methodology; Writing—original draft; Writing—review & editing. Moritz Jung: Investigation; Visualization; Validation; Writing—review & editing. Mariia Seleznova: Investigation; Writing—review & editing. Nicole Thorisch: Investigation; Writing—review & editing. Jeanette M. Lorenz: Writing—review & editing. Thomas Bocklitz: Writing—review & editing. Bjoern M. Eskofier: Writing—review & editing. Gitta Kutyniok: Writing—review & editing. Ferenc Krausz: Supervision; Project administration; Writing—review & editing. Kosmas V. Kepesidis: Conceptualization; Supervision; Project administration; Methodology; Writing—original draft; Writing—review & editing. All authors have read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-03226-9.

Correspondence and requests for materials should be addressed to Kosmas V. Kepesidis.