

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

Kosmas Kepesidis

`kosmas.kepesidis@cmf.hu`

Ludwig-Maximilians-Universität München

Niklas Leopold-Kerschbaumer

Center for Molecular Fingerprinting Research Center Nonprofit Ltd.

Timo Halenke

Ludwig-Maximilians-Universität München

Selina Süzeroğlu

Ludwig-Maximilians-Universität München

Moritz Jung

Center for Molecular Fingerprinting Research Center Nonprofit Ltd.

Mariia Seleznova

Ludwig-Maximilians-Universität München

Nicole Thorisch

Center for Molecular Fingerprinting Research Center Nonprofit Ltd.

Jeanette Lorenz

Fraunhofer Institute for Cognitive Systems IKS

Thomas Bocklitz

Institute of Physical Chemistry and Abbe Center of Photonics (IPC), Friedrich-Schiller-University

Bjoern Eskofier

Institute of AI in Medicine, LMU Klinikum

Gitta Kutyniok

Ludwig-Maximilians-Universität München

Ferenc Krausz

Max Planck Institute of Quantum Optics

Article

Keywords:

Posted Date: February 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8540695/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at npj Digital Medicine on September 17th, 2026. See the published version at <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,†}, Timo Halenke^{2,3,†}, Selina Süzeroğlu^{2,3,†}, Moritz Jung^{1,2,3},
Mariia Seleznova^{3,4}, Nicole Thorisch^{1,2,3}, Jeanette M. Lorenz^{2,5}, Thomas Bocklitz^{6,7},
Bjoern M. Eskofier^{8,9}, Gitta Kutyniok^{4,10,11,12}, Ferenc Krausz^{1,2,3}, and Kosmas V. Kepesidis^{1,2,3,*}

¹Center for Molecular Fingerprinting (CMF), Budapest, Hungary

²Faculty of Physics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany

³Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), Garching, Germany

⁴Department of Mathematics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany

⁵Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany

⁶Leibniz Institute of Photonic Technology (Leibniz-IPHT), Jena, Germany

⁷Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University (FSU), Jena, Germany

⁸Chair of AI-supported Therapy Decisions, LMU München, Munich, Germany

⁹Institute of AI for Health, Helmholtz Zentrum München, Neuherberg, Germany

¹⁰Munich Center for Machine Learning (MCML), Munich, Germany

¹¹Department of Physics and Technology, University of Tromsø, Tromsø, Norway

¹²Institute of Robotics and Mechatronics, DLR-German Aerospace Center, Germany

[†]These authors contributed equally

*Correspondence: kosmas.kepesidis@cmf.hu

ABSTRACT

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 and a Conditional Boundary Equilibrium GAN 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 subject-specific characteristics, and targeted augmentation of underrepresented body mass index categories. Together, these results establish conditional generative modeling of blood-based infrared spectra as a viable approach for virtual cohort construction, cohort balancing, and controlled *in-silico* phenotyping, paving the way toward more comprehensive and data-efficient studies in precision health.

MAIN

Longitudinal molecular profiling is emerging as a powerful approach for personalized health monitoring, enabling early disease detection and preventive medicine. The integration of diverse “omics” technologies—including genomics, transcriptomics, proteomics, metabolomics, epigenomics, microbiomics, and exposomics—together with biosensor data, medical imaging, and electronic health records, has created an unprecedented wealth of health-related information [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. These integrated datasets are deepening our understanding of human biology and may transform how we assess disease risk and identify early warning signs.

Within this context, infrared (IR) spectroscopy of liquid biopsies, such as blood plasma—particularly infrared molecular fingerprinting (IMF) based on Fourier-transform infrared (FTIR) spectroscopy—offers a noninvasive method to capture molecular information linked to health and disease. By analyzing vibrational signatures of biomolecules, IMF provides a holistic snapshot of plasma composition, covering major biochemical classes such as proteins, lipids, carbohydrates, and metabolites [14, 15, 16, 17]. IMF has demonstrated high temporal stability at the individual level, a key prerequisite for longitudinal and clinical applications [18, 19]. Reported applications include oncological and metabolic profiling, among others [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32].

However, assembling large and diverse datasets remains challenging. Collecting repeated visits from the same individuals over extended time periods is particularly demanding, requiring long-term follow-up and substantial logistical coordination. As a result, studies often face limited temporal coverage, imbalanced cohorts, and modest statistical power—constraints that hinder systematic investigation of aging effects, longitudinal health trajectories, and transitions from health to disease.

Generative modeling offers a means to address these limitations by creating synthetic spectra that closely resemble real measurements. In this work, we present conditional deep generative models that integrate demographic and anthropometric variables—age, sex, and body mass index (BMI)—using a Conditional Variational Autoencoder

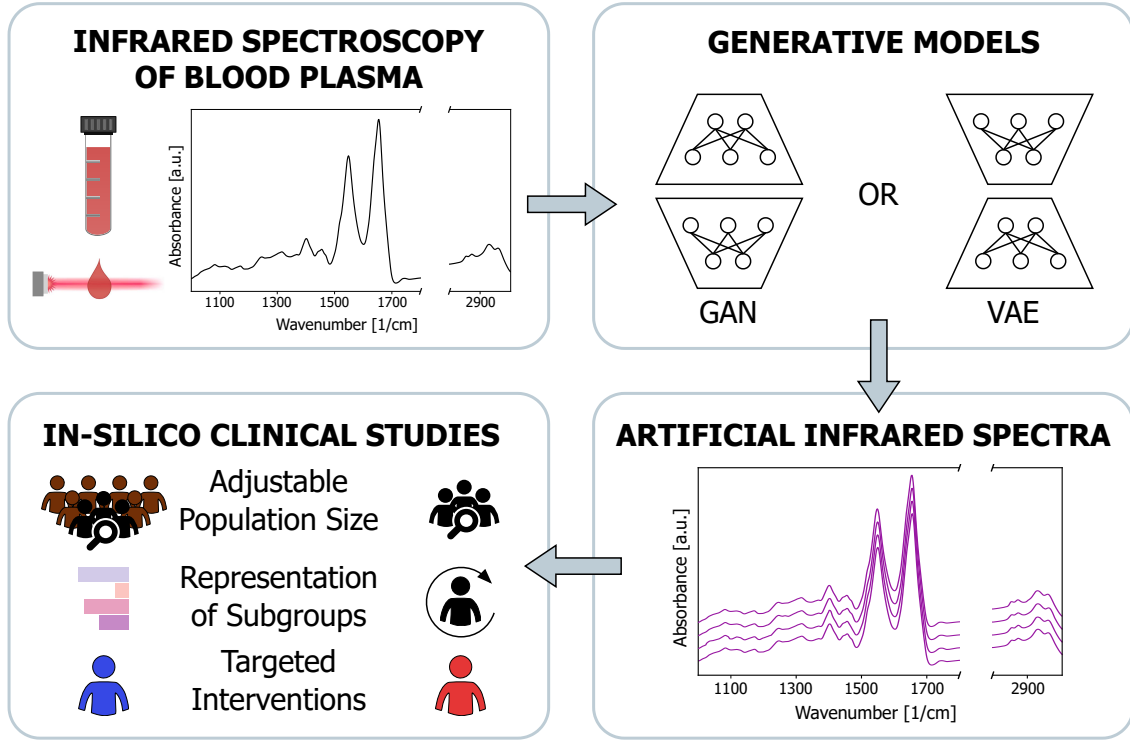


Figure 1: Overview of the generative framework for synthesizing individual-specific blood infrared spectra and *in-silico* clinical applications (e.g., cohort balancing, simulating healthy aging).

(CVAE) and a Conditional Boundary Equilibrium GAN (CBEGAN) [33, 34, 35, 36, 37]. Conditioning enables the generation of IMF spectra for virtual individuals with defined characteristics, supporting stratified data synthesis, cohort balancing, and digital-twin style analyses of person-specific effects.

Our models are trained on spectra from healthy participants in a longitudinal cohort and are designed to generate high-quality artificial 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 under-represented subgroups, and (iii) digital-twin style conditional projections for simulating healthy aging trajectories. Figure 1 summarizes the data type, model architecture, and main use cases.

Overall, our findings highlight the potential of 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 (IMF). The study aims to follow approximately 15,000 healthy volunteers over 10 years. As of now, the dataset comprises 25,308 infrared spectral samples collected from 5,863 participants, each contributing between one and six visits. This dataset serves as the basis for the present work.

For model development, participants were divided into mutually exclusive training and test sets (see Table 1). The generative models were trained exclusively on the training set, while their performance and clinical relevance were evaluated on the held-out test set. Evaluation involved (i) comparing the statistical distributions of real and generated 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 two models: a CVAE and a CBEGAN. 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 [36]. The CBEGAN generates spectra from noise and conditions, with an autoencoder discriminator

Table 1: Overview of the H4H longitudinal dataset

Dataset	# Blood Samples	# Individuals	# Visits per Person
H4H Total	25,308	5,863	1–6
Demographics¹: 2,009 male, 3,854 female; Age: 39–84 y; BMI: 15.0–64.87			
Train Set (80%)	20,246	4,698 ²	1–6
Test Set (20%)	5,062	1,165 ²	1–6
Test Set Subdivision: Single-Visit Cohorts (SVCs)			
SVC (6 cohorts)	5,062 total	5,062 total	1 per person
(Per-cohort size range: 843–844)			

¹ Further demographic details in Figure S2 (Supplementary).

² Train and test sets are mutually exclusive at the participant level.

using reconstruction error as the adversarial signal [33]. Details on the implementation and training are provided in the *Methods* section. Using the trained models, we generated two synthetic datasets, $\mathcal{D}_{\text{CVAE}}$ and $\mathcal{D}_{\text{CBEGAN}}$, each with $N=5,062$ spectra matched to the demographic distribution $\mathbf{y}$ of the real test set $\mathcal{D}_{\text{real}}$.

Quality of generated fingerprints

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

To assess potential differences between real and generated spectra at the spectrally resolved level, we compute the effect size (Cohen’s d) [38] at each wavenumber. For a given wavenumber k , the effect size is defined as

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

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

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

with $s_G(k)$ and $s_R(k)$ representing the standard deviations of the generated 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 30% of the standard deviation across all wavenumbers in both models, indicating that deviations between real and generated spectra are small and of limited practical significance.

To probe higher-order structure, we extracted a set of 24 peak ratios previously reported [18] 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 generated spectra align closely. Moreover, as shown in Figure 2d, effect size across ratios remained well below 30% of the standard deviation, supporting faithful reproduction of local and nonlinear spectral features.

Conditional information content of generated fingerprints

We next assess whether the generated spectra correctly encode conditional information. To do so, we construct case–control designs that allow comparison of the conditional distributions between real and generated 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 [39]. 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.

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

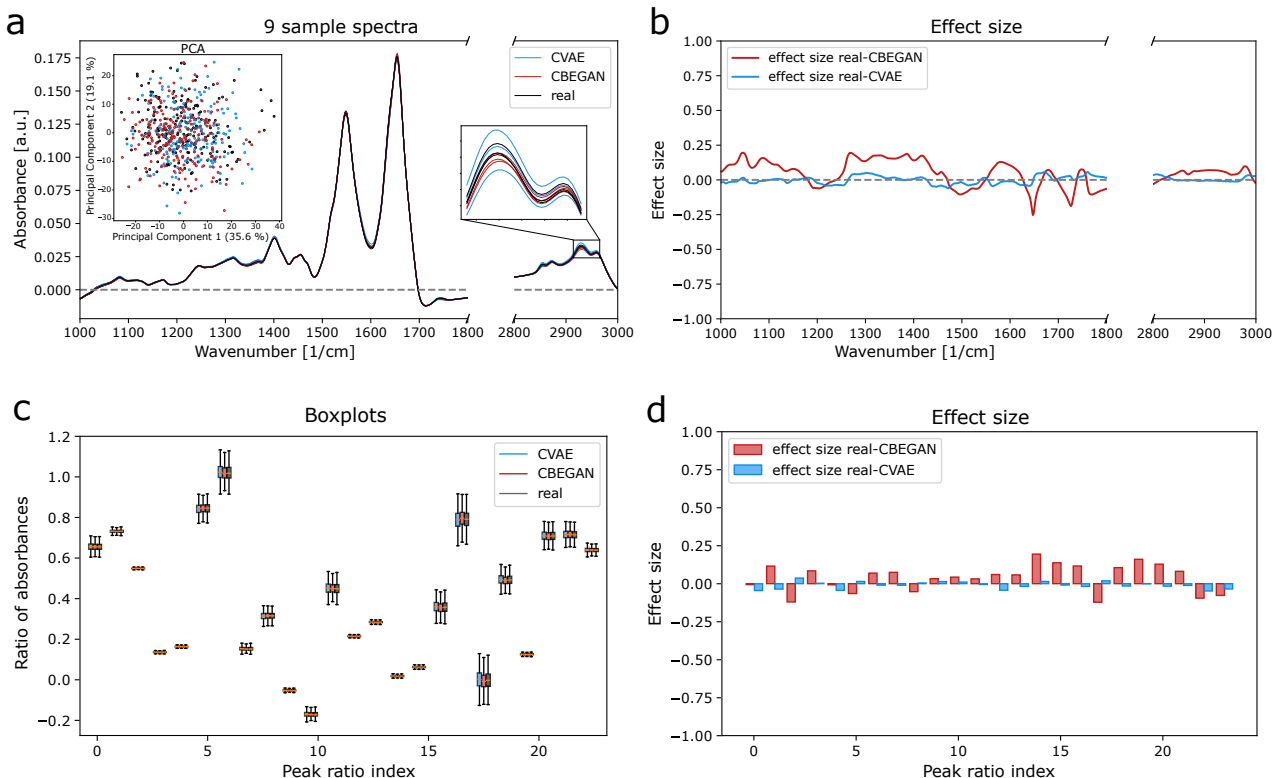


Figure 2: Distribution analysis. (a) Representative spectra (three each) from the real, CVAE-generated, and CBEGAN-generated datasets. Spectra show absorbance in arbitrary units, measured via FTIR and preprocessed as described in the Methods. The inset displays a scatter plot of the first two principal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

generated data. As shown in Figures 3a–c, the effect size from real and generated spectra is closely aligned across the full spectral range for all three conditions.

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 Figures S1–S2). We train logistic regression classifiers (L2 penalty, $C=10$) to predict sex, and the binarized age and BMI labels, respectively. A baseline model is trained and evaluated on real data, and its performance is compared with models trained on synthetic data (CVAE, CBEGAN) 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 ROC curves are shown in Figures 3d–f. Both synthetic-trained classifiers achieve ROC performance comparable to the baseline, with CBEGAN exhibiting slightly better overall alignment. This modest advantage is consistent with the adversarial model's ability to reproduce sharper, condition-relevant spectral structure, whereas the CVAE's smoother generative assumptions can temper subtle variations associated with the covariates.

Applications in *in-silico* modeling

In the preceding sections, we demonstrated that the generated spectra not only reproduce the characteristics of real data but also accurately encode demographic and anthropometric information. This property enables the interpretation of generated 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 infrared metabolic fingerprint (IMF) profiles [18, 19], we investigated the potential of digital-twin style conditional spectrum

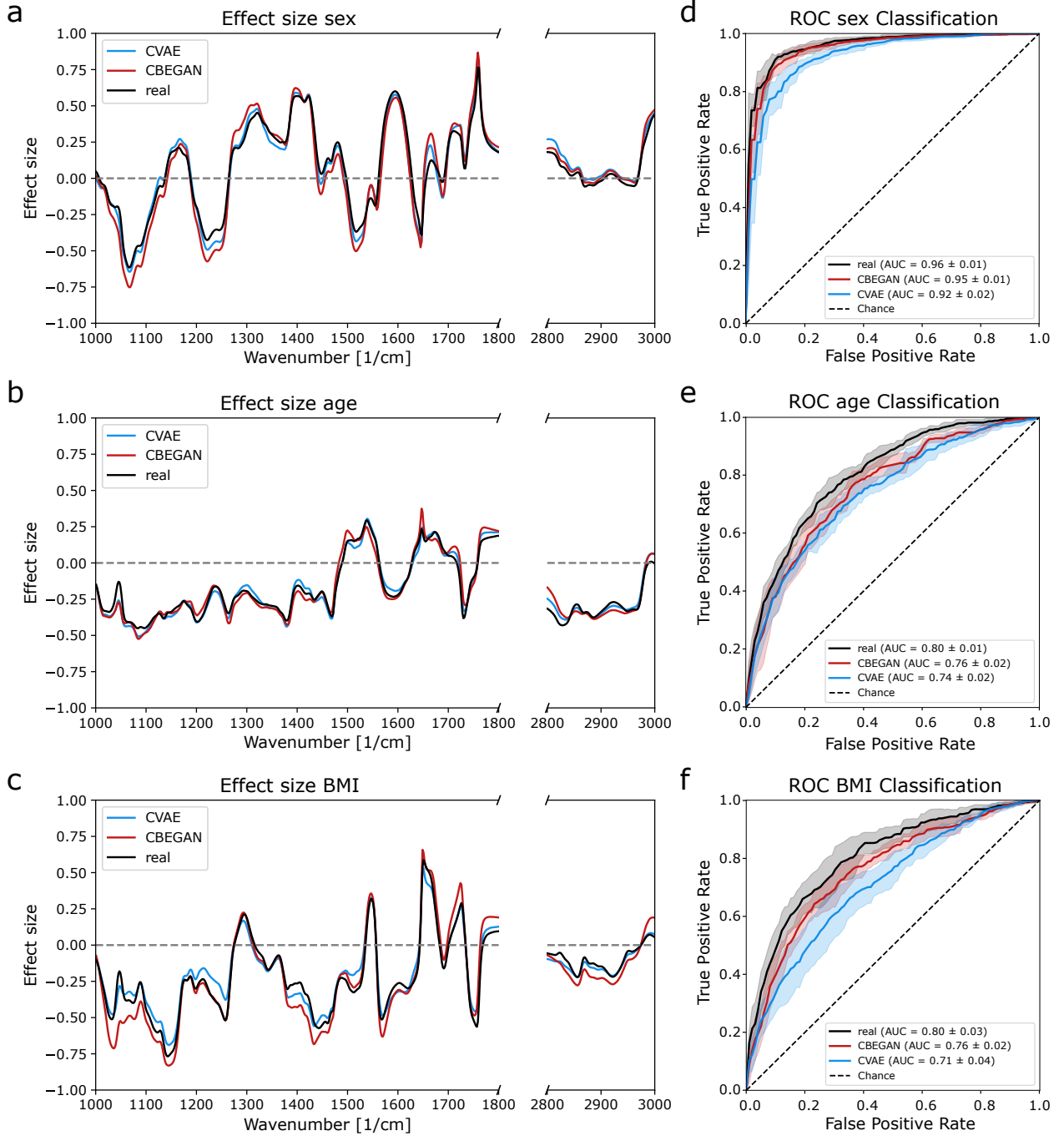


Figure 3: Conditional analysis. (a–c) Effect-size curves comparing positive and negative classes derived from binarized covariates (sex; age $\leq a_0$ vs. $> a_0$; BMI $\leq b_0$ vs. $> b_0$) for real and simulated data. (d–f) ROC curves for classifiers trained on real spectra (black) and on synthetic spectra generated by the CVAE (blue) or CBEGAN (red), with all models evaluated on real data. See the main text for details on label construction.

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 subject-specific features. Figure 4a shows a two-dimensional UMAP (Uniform Manifold Approximation and Projection [40]) 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 Figure S3a (Supplementary). Onto 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 generated 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 (Figure 4b). The spectral areas used for both embeddings are shown in Figure S3c (Supplementary). Therefore, these trajectories demonstrate

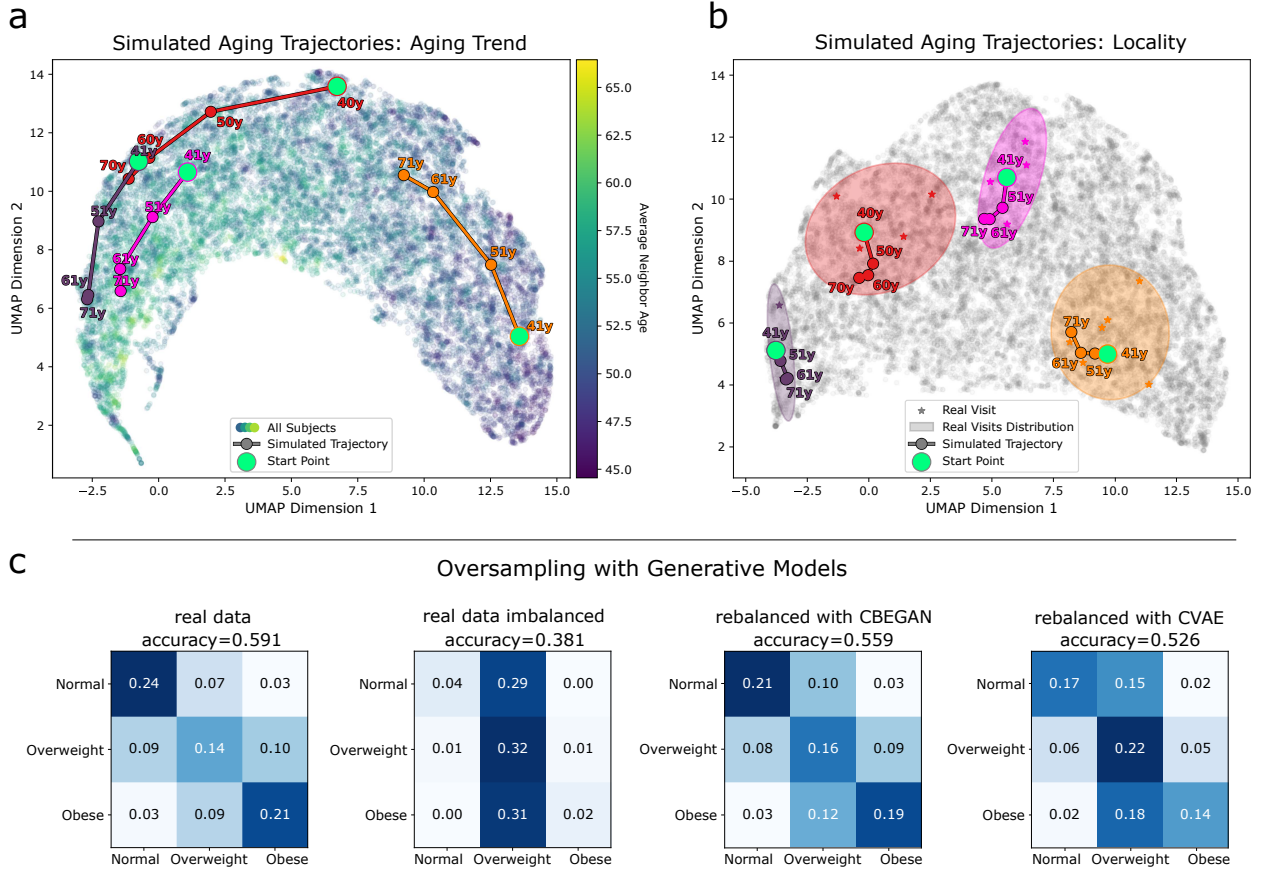


Figure 4: 2D UMAP embeddings for four patients with simulated aging in four 10-year steps. Initial spectra are averaged across the first four short-interval visits. (a) Embeddings using 30 age-correlated wavenumbers show clear aging trajectories; background colored by the average age of 20 nearest neighbors.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

that the CVAE captures both global aging trends and individual-specific spectral characteristics. This allows it to reproduce physiologically plausible aging pathways. 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 ($18.5\text{--}25.0$), overweight ($25.0\text{--}30.0$), and obese (>30.0) [41]. 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–iv) two augmented versions in which the class imbalance was resolved by adding spectra generated by the CVAE and CBEGAN 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. 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 Figure S5. Collectively, these results highlight the potential of synthetic data to stabilize model performance in underrepresented subgroups, thereby enhancing the robustness of in-silico analyses.

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. Both the CVAE and CBEGAN 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 subject-specific characteristics. Conditioning allows the generated spectra to be interpreted in relation to person-level attributes, enabling reproducible case-control designs and individualized aging simulations.

Despite these promising results, several limitations remain. The models were trained on healthy participants within a single measurement pipeline, and their generalization to other instruments, laboratories, or disease states has yet to be established. Conditioning is currently limited to three covariates, leaving potential 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, enable the generation of underrepresented clinical subgroups, and improve 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 offer the possibility of generating biospectral data jointly with other modalities—including omics profiles, vital signs, medical imaging, and electronic health records—thereby embedding IMF synthesis in a broader cross-modality generative ecosystem.

Looking ahead, a major step will be the transition from conditional to causal generative modeling [42, 43, 44, 45, 46, 47]. While conditional models learn $p(\mathbf{x} \mid \mathbf{y})$ and support conditional simulation of spectra, they do not capture interventional distributions $p(\mathbf{x} \mid do(\mathbf{y}))$. 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). As part of the study protocol, Fourier-transform infrared (FTIR) spectra 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 $930\text{--}3050\text{ cm}^{-1}$. 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.

236 The MIRA-Analyzer does not provide access to raw measurements but only to the water-compensated absorbance
 237 spectra.

238 Samples were processed in batches of 28, with DMSO₂, water, and pooled human serum (BioWest, Nuaille, France)
 239 included as quality controls, yielding a total of 34 samples per batch. The BioWest quality control samples enabled
 240 tracking of analytical variation and instrument stability over time.

241 Data preparation and preprocessing

242 Measured spectra were truncated to 1000 cm⁻¹ to 3000 cm⁻¹ to remove noisy regions caused by the low transmittance
 243 of CaF₂ below 950 cm⁻¹ and the high water absorption above 3000 cm⁻¹. In addition, the uninformative region
 244 (so-called silent region) between 1800 cm⁻¹ to 2800 cm⁻¹ was removed. We finally applied ℓ_2 normalization to the
 245 truncated spectra and standardized the retained absorbance values of each wavenumber by standard scaling (z-scoring)
 246 fit on the train data set. This pipeline follows prior work [18, 20, 26, 48, 25, 32].

247 Covariates were preprocessed separately. Age and BMI were min-max scaled to $[0, 1]$; sex was encoded $\in \{0, 1\}$.
 248 All three were then mapped to d -dimensional positional encodings [49]:

$$\text{PE}(\text{pos}, 2i) = \sin\left(\frac{\text{pos}}{0.1^{2i/d}}\right), \quad \text{PE}(\text{pos}, 2i+1) = \cos\left(\frac{\text{pos}}{0.1^{2i/d}}\right), \quad (3)$$

249 with $i \in \{0, \dots, d/2\}$. This fixed, non-learned encoding supports unseen continuous values (e.g., new ages/BMIs) and
 250 provides a richer input than scalars. Python implementations used PyTorch.

251 Modeling background

252 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
 253 (e.g., Gaussian) to the data space, approximating $p(\mathbf{x})$ [50]. We consider conditional variants of the VAE [36, 37] and
 254 BEGAN [33, 35] that incorporate covariates $\mathbf{y}$.

255 Conditional VAE

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

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

257 and generation as

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

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

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

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

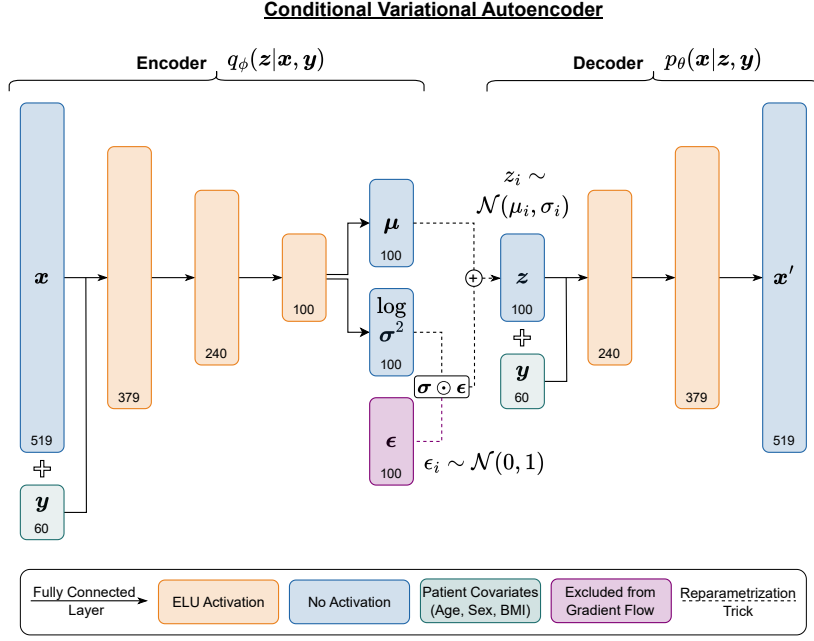


Figure 5: CVAE schematic. The encoder outputs parameters of $q_\phi(\mathbf{z} | \mathbf{x}, \mathbf{y})$; reparameterized samples are decoded to $p_\theta(\mathbf{x} | \mathbf{z}, \mathbf{y})$.

Mitigating latent space collapse. To avoid degenerate solutions with negligible KL and uninformative $\mathbf{z}$ [51], 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 S4a (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` [52].

Conditional BEGAN

The CBEGAN comprises a generator and discriminator (Figure 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

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

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

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

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

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

with diversity ratio $\gamma \in [0, 1]$ [33, 35].

Training CBEGAN: hyperparameters. We trained for 500 epochs with Adam (initial learning rate 10^{-4} , cosine annealed to 10^{-6}) [53, 54]. Following [34], 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` [52]. Training curves are in Figure S4b (Supplementary).

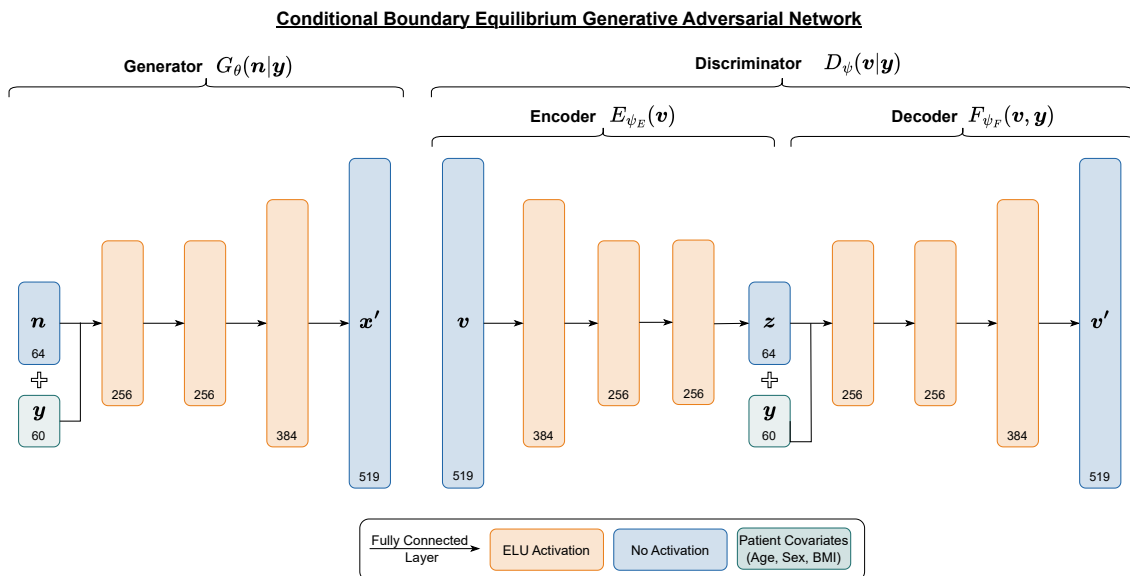


Figure 6: CBEGAN schematic. $G_{\theta}(\mathbf{n} | \mathbf{y})$ generates $\mathbf{x}'$. $D_{\psi}(\cdot | \mathbf{y})$ reconstructs both $\mathbf{x}$ and $\mathbf{x}'$; the reconstruction error drives adversarial training.

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.

CODE AVAILABILITY

Python implementations and trained weights for CBEGAN and CVAE are available on GitHub [55] (<https://github.com/Data-Science-Group-Attoworld/in-silico-infrared-modeling>).

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. *New England journal of medicine* **372**, 793–795 (2015).
- [3] Price, N. D. *et al.* A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature biotechnology* **35**, 747–756 (2017).
- [4] Piening, B. D. *et al.* Integrative personal omics profiles during periods of weight gain and loss. *Cell systems* **6**, 157–170 (2018).
- [5] Schüssler-Fiorenza Rose, S. M. *et al.* A longitudinal big data approach for precision health. *Nature medicine* **25**, 792–804 (2019).
- [6] Ahadi, S. *et al.* Personal aging markers and ageotypes revealed by deep longitudinal profiling. *Nature medicine* **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. *The Journals of Gerontology: Series A* **74**, S52–S60 (2019).
- [8] Hou, Y.-C. C. *et al.* Precision medicine integrating whole-genome sequencing, comprehensive metabolomics, and advanced imaging. *Proceedings of the National Academy of Sciences* **117**, 3053–3062 (2020).
- [9] Tebani, A. *et al.* Integration of molecular profiles in a longitudinal wellness profiling cohort. *Nature communications* **11**, 4487 (2020).
- [10] Chen, X. *et al.* Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. *Nature communications* **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. *Nature medicine* **26**, 29–38 (2020).
- [14] Baker, M. J. *et al.* Using fourier transform ir spectroscopy to analyze biological materials. *Nature protocols* **9**, 1771–1791 (2014).
- [15] Butler, H. J. *et al.* Development of high-throughput atr-ftir technology for rapid triage of brain cancer. *Nature communications* **10**, 4501 (2019).
- [16] Voronina, L. *et al.* Molecular origin of blood-based infrared spectroscopic fingerprints. *Angewandte Chemie* **60**, 17060–17069 (2021).
- [17] Paraskevaidi, M. *et al.* Clinical applications of infrared and raman spectroscopy in the fields of cancer and infectious diseases. *Applied Spectroscopy Reviews* **56**, 804–868 (2021).
- [18] Huber, M. *et al.* Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nature communications* **12**, 1511 (2021).
- [19] 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).
- [20] Huber, M. *et al.* Infrared molecular fingerprinting of blood-based liquid biopsies for the detection of cancer. *Elife* **10**, e68758 (2021).
- [21] Martin, F. L. *et al.* Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nature protocols* **5**, 1748–1760 (2010).
- [22] Ghimire, H. *et al.* Protein conformational changes in breast cancer sera using infrared spectroscopic analysis. *Cancers* **12**, 1708 (2020).
- [23] Ollesch, J. *et al.* An infrared spectroscopic blood test for non-small cell lung carcinoma and subtyping into pulmonary squamous cell carcinoma or adenocarcinoma. *Biomedical Spectroscopy and Imaging* **5**, 129–144 (2016).
- [24] Eissa, T. *et al.* Plasma infrared fingerprinting with machine learning enables single-measurement multi-phenotype health screening. *Cell Reports Medicine* **5** (2024).
- [25] Kepesidis, K. V. *et al.* Assessing lung cancer progression and survival with infrared spectroscopy of blood serum. *BMC medicine* **23**, 101 (2025).
- [26] Kepesidis, K. V. *et al.* Breast-cancer detection using blood-based infrared molecular fingerprints. *BMC cancer* **21**, 1–9 (2021).
- [27] Backhaus, J. *et al.* Diagnosis of breast cancer with infrared spectroscopy from serum samples. *Vibrational Spectroscopy* **52**, 173–177 (2010).
- [28] Elmi, F., Movaghar, A. F., Elmi, M. M., Alinezhad, H. & Nikbakhsh, N. Application of ft-ir spectroscopy on breast cancer serum analysis. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **187**, 87–91 (2017).
- [29] Ollesch, J. *et al.* It’s in your blood: spectral biomarker candidates for urinary bladder cancer from automated ftir spectroscopy. *Journal of biophotonics* **7**, 210–221 (2014).
- [30] Anderson, D., Anderson, R., Moug, S. & Baker, M. Liquid biopsy for cancer diagnosis using vibrational spectroscopy: systematic review. *BJS open* **4**, 554–562 (2020).
- [31] 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).
- [32] Wegner, C. *et al.* Toward informative representations of blood-based infrared spectra via unsupervised deep learning. *Journal of Biophotonics* e70011 (2025).
- [33] Berthelot, D., Schumm, T. & Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

- [34] Zhu, D. *et al.* Synthetic spectra generated by boundary equilibrium generative adversarial networks and their applications with consensus algorithms. *Optics Express* **28**, 17196–17208 (2020).
- [35] 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).
- [36] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
- [37] 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).
- [38] Sawilowsky, S. S. New effect size rules of thumb. *Journal of modern applied statistical methods* **8**, 26 (2009).
- [39] Rosenbaum, P. R., Rosenbaum, P. B. & Briskman. *Design of observational studies*, vol. 10 (Springer, 2010).
- [40] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
- [41] 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).
- [42] Sanchez, P. *et al.* Causal machine learning for healthcare and precision medicine. *Royal Society Open Science* **9**, 220638 (2022).
- [43] 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).
- [44] Schölkopf, B. *et al.* Toward causal representation learning. *Proceedings of the IEEE* **109**, 612–634 (2021).
- [45] Pawlowski, N., Coelho de Castro, D. & Glocker, B. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems* **33**, 857–869 (2020).
- [46] Sanchez, P. & Tsaftaris, S. A. Diffusion causal models for counterfactual estimation. *arXiv preprint arXiv:2202.10166* (2022).
- [47] Yang, M. *et al.* Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697* (2020).
- [48] Nemeth, F. B. *et al.* Bridging spectral gaps: Cross-device model generalization in blood-based infrared spectroscopy. *Analytical Chemistry* (2025).
- [49] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).
- [50] Foster, D. *Generative deep learning* (" O'Reilly Media, Inc.", 2022).
- [51] Fu, H. *et al.* Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145* (2019).
- [52] 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).
- [53] Kingma, D. P. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [54] Loshchilov, I. & Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
- [55] 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).
- [56] Blat, A. *et al.* Fourier transform infrared spectroscopic signature of blood plasma in the progression of breast cancer with simultaneous metastasis to lungs. *Journal of Biophotonics* **12**, e201900067 (2019).
- [57] Ghimire, H., Jayaweera, P. & Perera, A. U. Longitudinal analysis of molecular alteration in serum samples of dextran sodium sulfate-induced colitis mice by using infrared spectroscopy. *Infrared Physics & Technology* **97**, 33–37 (2019).
- [58] Ghimire, H., Venkataramani, M., Bian, Z., Liu, Y. & Perera, A. U. Atr-ftir spectral discrimination between normal and tumorous mouse models of lymphoma and melanoma from serum samples. *Scientific reports* **7**, 16993 (2017).

- [59] Wang, X., Shen, X., Sheng, D., Chen, X. & Liu, X. Ftir spectroscopic comparison of serum from lung cancer patients and healthy persons. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **122**, 193–197 (2014).
- [60] Lasch, P., Beekes, M., Fabian, H. & Naumann, D. Antemortem identification of transmissible spongiform encephalopathy (tse) from serum by mid-infrared spectroscopy. *Handbook of vibrational spectroscopy* (2006).
- [61] Sahu, R. K. *et al.* Continuous monitoring of wbc (biochemistry) in an adult leukemia patient using advanced ftir-spectroscopy. *Leukemia research* **30**, 687–693 (2006).
- [62] Staniszewska-Slezak, E., Mateuszuk, L., Chlopicki, S., Baranska, M. & Malek, K. Alterations in plasma biochemical composition in no deficiency induced by l-name in mice analysed by fourier transform infrared spectroscopy. *Journal of Biophotonics* **9**, 1098–1108 (2016).
- [63] Sheng, D. *et al.* Comparison of serum from gastric cancer patients and from healthy persons using ftir spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* **116**, 365–369 (2013).
- [64] Staniszewska-Slezak, E. *et al.* A possible fourier transform infrared-based plasma fingerprint of angiotensin-converting enzyme inhibitor-induced reversal of endothelial dysfunction in diabetic mice. *Journal of Biophotonics* **11**, e201700044 (2018).
- [65] Paraskevaïdi, M., Martin-Hirsch, P. L. & Martin, F. L. Atr-ftir spectroscopy tools for medical diagnosis and disease investigation. In *Nanotechnology Characterization Tools for Biosensing and Medical Diagnosis*, 163–211 (Springer, 2018).
- [66] Paraskevaïdi, M. *et al.* Differential diagnosis of alzheimer’s disease using spectrochemical analysis of blood. *Proceedings of the National Academy of Sciences* **114**, E7929–E7938 (2017).
- [67] Movasaghi, Z., Rehman, S. & ur Rehman, D. I. Fourier transform infrared (ftir) spectroscopy of biological tissues. *Applied Spectroscopy Reviews* **43**, 134–179 (2008).
- [68] Gazi, E. *et al.* Applications of fourier transform infrared microspectroscopy in studies of benign prostate and prostate cancer. a pilot study. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* **201**, 99–108 (2003).
- [69] Smith, B. R. *et al.* Combining random forest and 2d correlation analysis to identify serum spectral signatures for neuro-oncology. *Analyst* **141**, 3668–3678 (2016).
- [70] Carmona, P., Molina, M., López-Tobar, E. & Toledano, A. Vibrational spectroscopic analysis of peripheral blood plasma of patients with alzheimer’s disease. *Analytical and bioanalytical chemistry* **407**, 7747–7756 (2015).
- [71] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

ACKNOWLEDGMENTS

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 Debrecei, 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 Selezanova:** 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.

448 ETHICS DECLARATIONS

449 Competing interests

450 All authors declare no financial or non-financial competing interests.

451 SUPPLEMENTARY INFORMATION

452 Supplementary Information covering (A) the demographics of the training set and the SVCs, (B) spectral regions for
453 dimensionality reduction visualizations, (C) the description of the peak ratios, (D) the training process of CVAE and
454 CBEGAN, and (E) classification results for rebalancing age and BMI groups is available online at *URL*.

Supplementary Information

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

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

¹ Center for Molecular Fingerprinting (CMF), Budapest, Hungary

² Faculty of Physics, Ludwig-Maximilians-Universität München (LMU), Garching, Germany

³ Laboratory for Attosecond Physics, Max Planck Institute of Quantum Optics (MPQ), Garching, Germany

⁴ Department of Mathematics, 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, Germany

[†] These authors contributed equally

* Correspondence: kosmas.kepesidis@cmf.hu

Additional information about (A) the demographics of the training set and the SVCs, (B) spectral regions for dimensionality reduction visualizations, (C) the description of the peak ratios, (D) the training process of CVAE and CBEGAN, and (E) classification results for rebalancing age and bmi groups, is included herein to complement the main text.

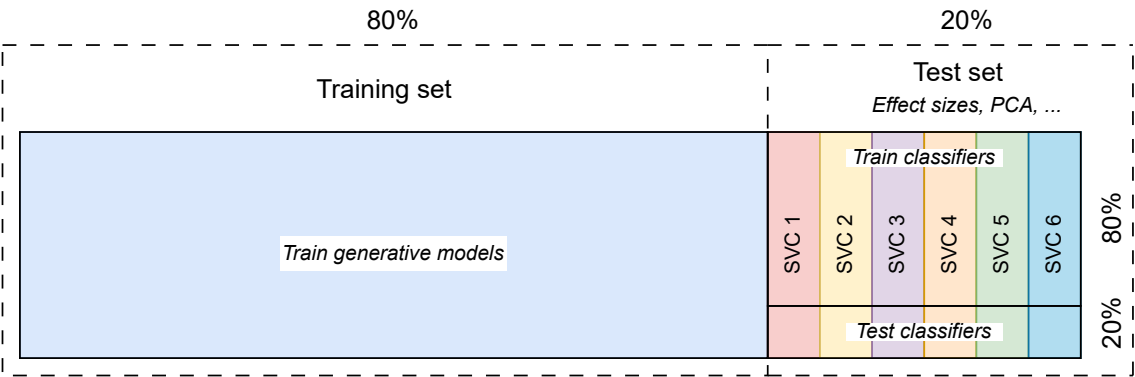


Figure S1: Overview of data splits and naming conventions. *Italic text indicates use cases.*

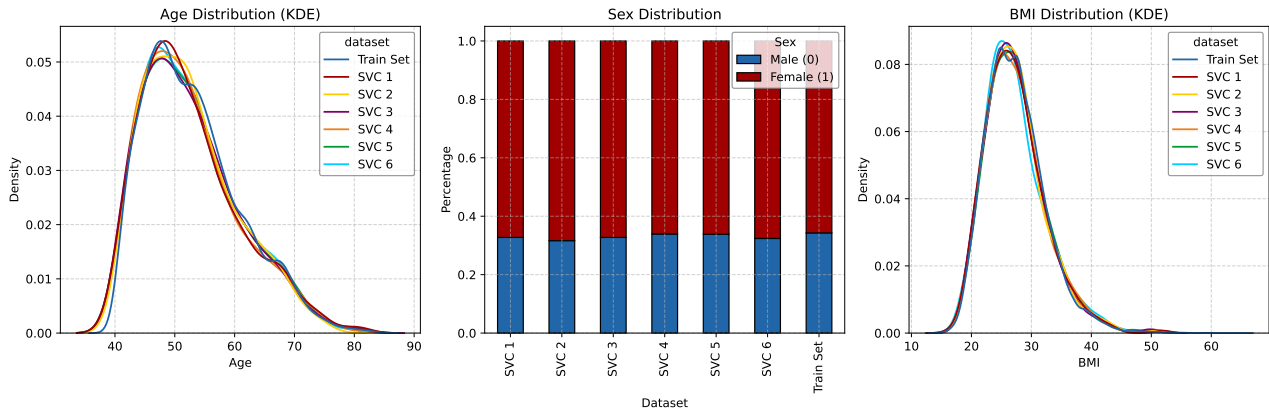


Figure S2: Distributions of age, sex, and BMI for all SVCs and the training set.

480 SVCs were obtained by randomly assigning a `cohort_id` $\in \{1, \dots, 6\}$ to each visit without repetition per individual.
481 Each SVC comprises all measurements sharing the same `cohort_id`. Sizes were balanced by iteratively moving a
482 random measurement from the largest SVC to the smallest SVC whenever the subject did not already appear in the
483 target SVC.

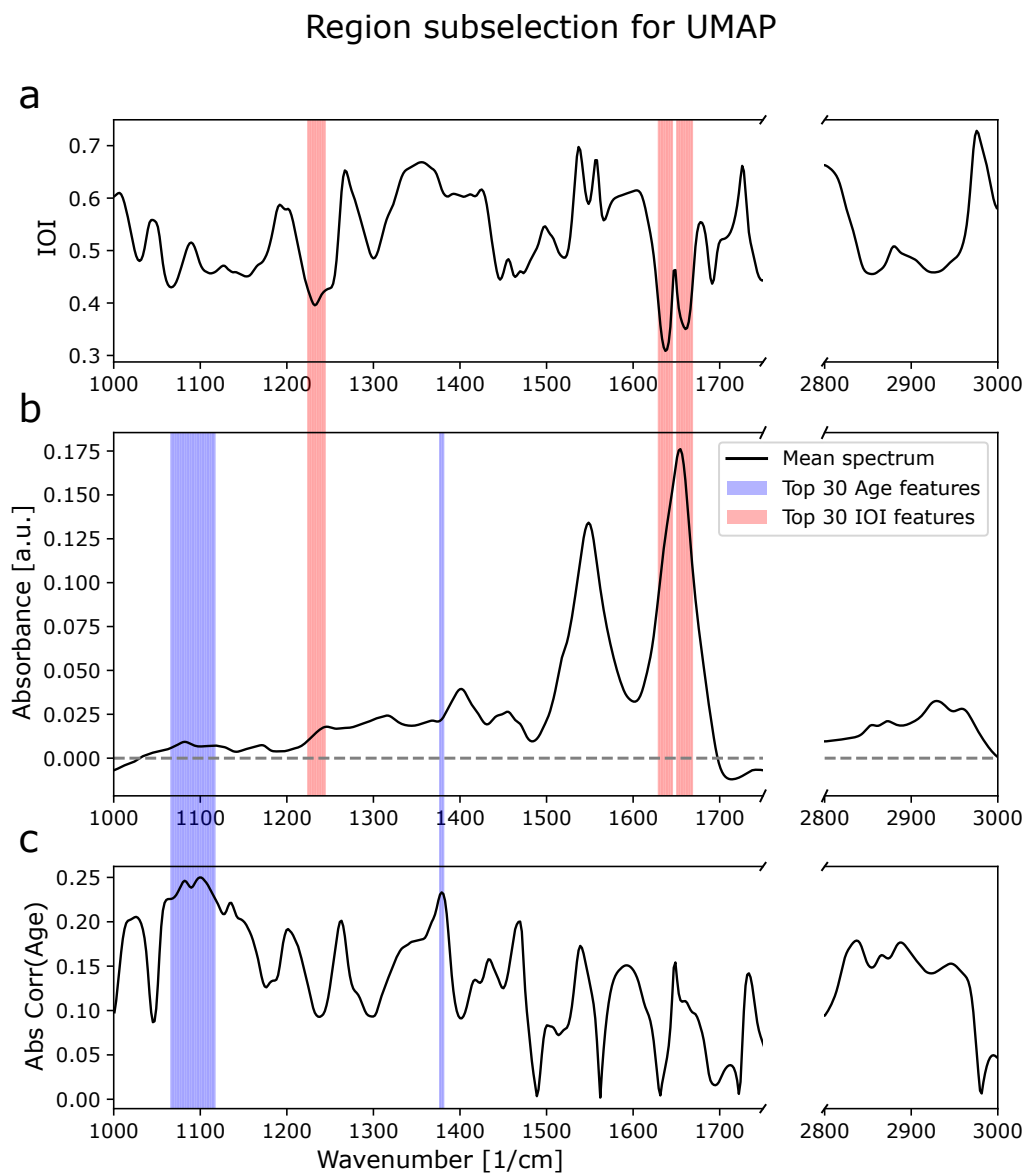


Figure S3: Regions selected for dimensionality reduction in Figure 4. (a) Index of Individuality, (b) Mean spectrum, (c) Absolute correlation with age per wavenumber.

Index	Peak Ratio	Description
0	I_{1635}/I_{1654}	Ratio of β -sheet to α -helix secondary structures; [56, 57, 58, 59, 60]
1	I_{1546}/I_{1655}	Amide I to amide II ratio; secondary structure/fibrils [61, 62, 59, 63]
2	$I_{1655}/(I_{1655} + I_{1548})$	α -helix to total proteins [62]
3	$I_{1684}/(I_{1655} + I_{1548})$	Antiparallel β -sheet to total proteins [62]
4	$I_{1515}/(I_{1655} + I_{1548})$	Tyrosine-rich proteins to total proteins [62]
5	I_{2959}/I_{2931}	$\nu_{as}(\text{CH}_3)/\nu_{as}(\text{CH}_2)$; lipid chain length [63]
6	$(I_{2855} + I_{2927})/(I_{2962} + I_{2871})$	Fatty-acid elongation [56, 64]
7	$(I_{2851} + I_{2927})/(I_{1655} + I_{1548})$	Lipid-to-protein ratio [62]
8	$I_{1239}/(I_{2851} + I_{2927})$	Phospholipids to total lipids [62]
9	I_{1741}/I_{1640}	Lipid-to-protein ratio [65]
10	I_{1740}/I_{1400}	Lipid-to-protein ratio [61, 65]
11	I_{2852}/I_{1400}	Lipid-to-protein ratio [61]
12	I_{1450}/I_{1539}	Lipid-to-protein ratio [66]
13	I_{1240}/I_{1517}	Tyrosine phosphorylation degree [62]
14	I_{1045}/I_{1545}	Phosphate-to-carbohydrate [66]
15	I_{1080}/I_{1550}	Phosphate-to-amide II [61, 59, 63]
16	I_{1060}/I_{1230}	$\nu_s(\text{PO}_2^-)/\nu_{as}(\text{PO}_2^-)$ [66]
17	I_{1170}/I_{1080}	Relative nucleic-acid content [59]
18	I_{1030}/I_{1080}	Glycogen/phosphate; metabolic turnover [67, 68, 69]
19	I_{1080}/I_{1243}	$\nu_s(\text{PO}_2^-)/\nu_{as}(\text{PO}_2^-)$
20	$I_{1587}/(I_{1655} + I_{1548})$	Free amino acids to proteins [59]
21	I_{1156}/I_{1171}	Carbohydrate moieties in plasma globulins [70]
22	I_{1243}/I_{1314}	Protein/nucleic acid changes [63]
23	I_{1453}/I_{1400}	$\delta_{as}(\text{CH}_3)/\delta_s(\text{CH}_3)$ [63]

Table S1: Peak ratios used to evaluate simulation quality (adapted from [18]).

D Training progress

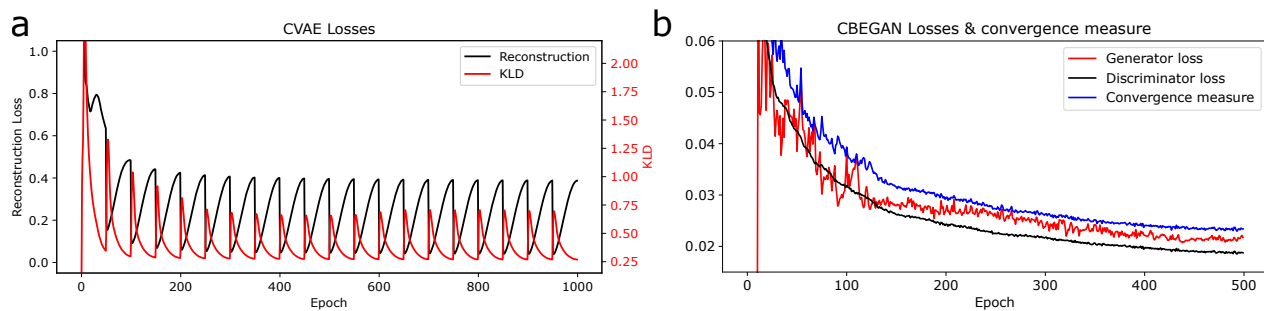


Figure S4: Training progress. (a) CVAE reconstruction (MSE) and KL loss with cyclic β . (b) CBEGAN generator/discriminator losses and convergence measure.

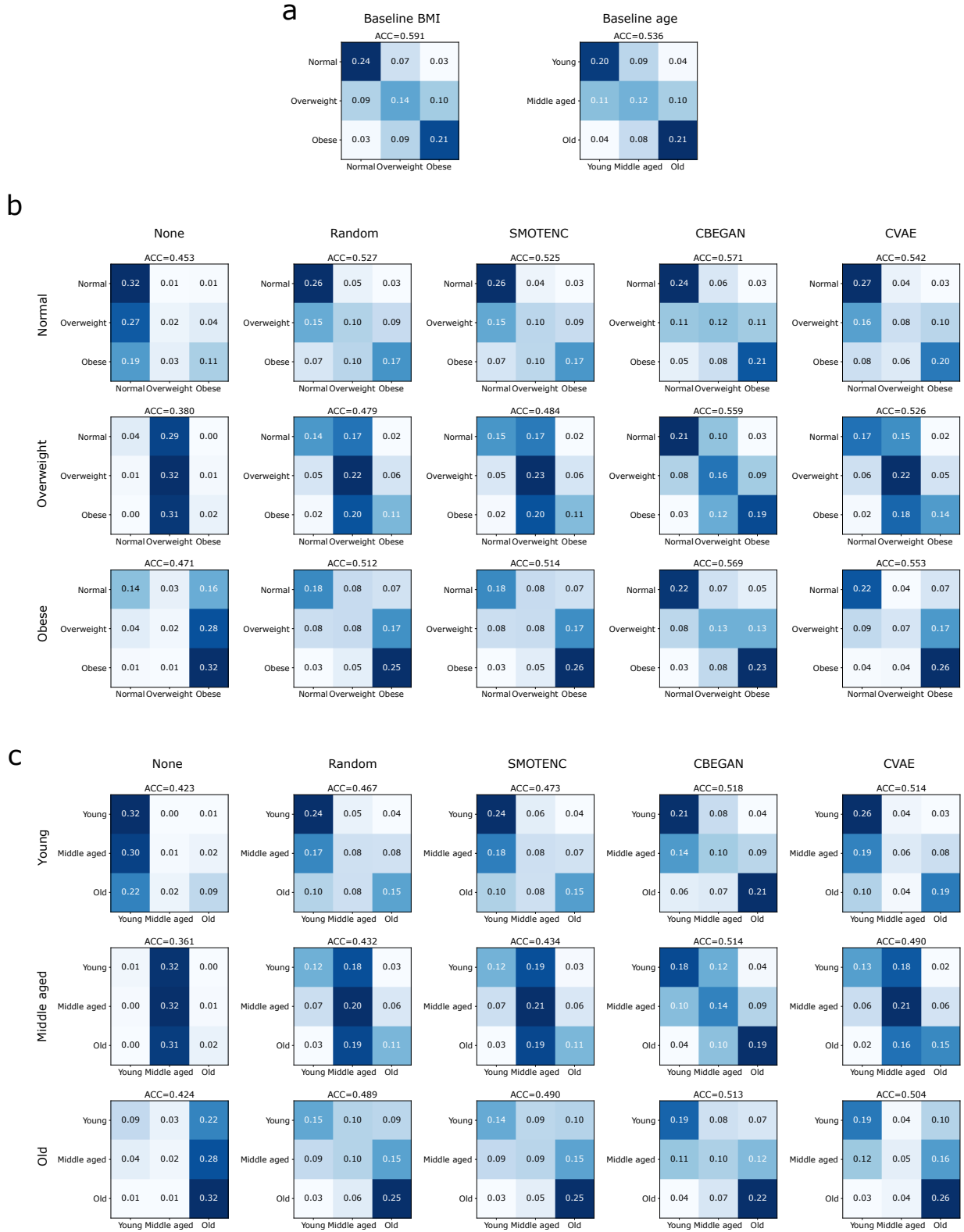


Figure S5: Evaluation of oversampling methods for underrepresented groups. Random and SMOTENC implemented via scikit-learn [71]. In each matrix, x-axis: predicted label; y-axis: true label. (a) Baseline without undersampling. (b) Age classification (thresholds 49, 55). (c) BMI classification (thresholds 25, 30).