

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^{2,4,†}, Selina Süzeroğlu^{2,4,†}, Moritz Jung^{1,2,3},
Mariia Seleznova^{2,3,5}, Nicole Thorisch^{1,4}, Jeanette M. Lorenz^{4,6}, Thomas Bocklitz^{7,8},
Bjoern M. Eskofier^{9,10}, Gitta Kutyniok^{3,11,12,13}, Ferenc Krausz^{1,2,4,14}, and Kosmas V. Kepesidis^{1,2,4,*}

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

²Max Planck Institute of Quantum Optics (MPQ), Garching, Germany

³Department of Mathematics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany

⁴Faculty of Physics, Ludwig-Maximilians-Universität München (LMU), Munich, Germany

⁵Signal Theory and Statistical Learning Group (STSL), Technische Universität Darmstadt (TU Darmstadt), Darmstadt, Germany

⁶Fraunhofer Institute for Cognitive Systems IKS, Munich, Germany

⁷Leibniz Institute of Photonic Technology (Leibniz-IPHT), Jena, Germany

⁸Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University (FSU), Jena, Germany

⁹Chair of AI-supported Therapy Decisions, LMU München, Munich, Germany

¹⁰Institute of AI for Health, Helmholtz Zentrum München, Neuherberg, Germany

¹¹Munich Center for Machine Learning (MCML), Munich, Germany

¹²Department of Physics and Technology, University of Tromsø, Tromsø, Norway

¹³Institute of Robotics and Mechatronics, DLR-German Aerospace Center, Germany

¹⁴Institute for Precision Prevention, The University of Hong Kong, Hong Kong SAR, China

[†]These authors contributed equally

*Correspondence: kosmas.kepesidis@cmf.hu

Additional information about (A) choice of dataset splits and the demographics of the training set and the SVCs, (B) the description of the peak ratios, (C) the noise scheduel and guidance scale tuning for CDIFF, (D) details on classification tasks, (E) the Pearson correlations between spectra and covariates, (F) Sensitivity to alternative age and BMI thresholds, (G) regression tasks for predicting age and BMI from spectra, (H) the spectral regions for dimensionality reduction visualizations, (I) classification results for rebalancing age and bmi groups, and (J) the training process of CVAE, CBEGAN, and CDIFF, is included herein to complement the main text.

A Single-Visit Cohorts

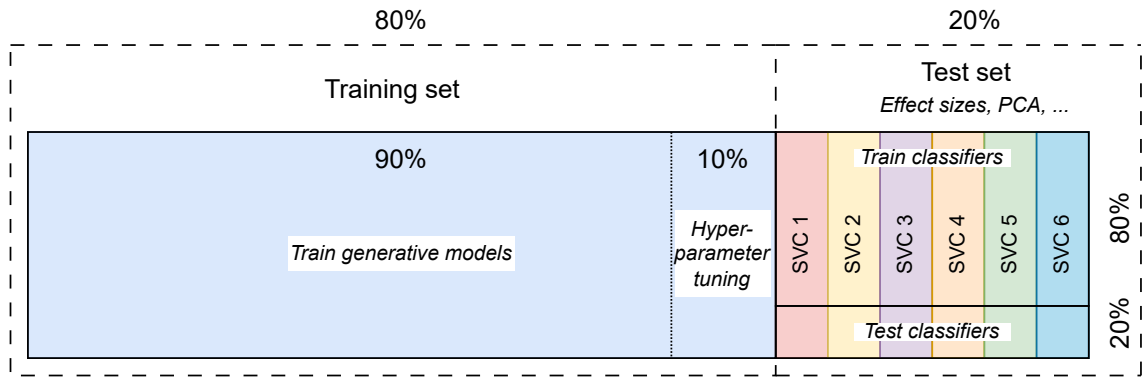


Figure S1: Overview of data splits and naming conventions. Italic text indicates use cases. Each individual can only appear in either the training, validation, or test set. An individual can appear in multiple SVCs; however, within an SVC, by construction, each measurement is from a different individual.

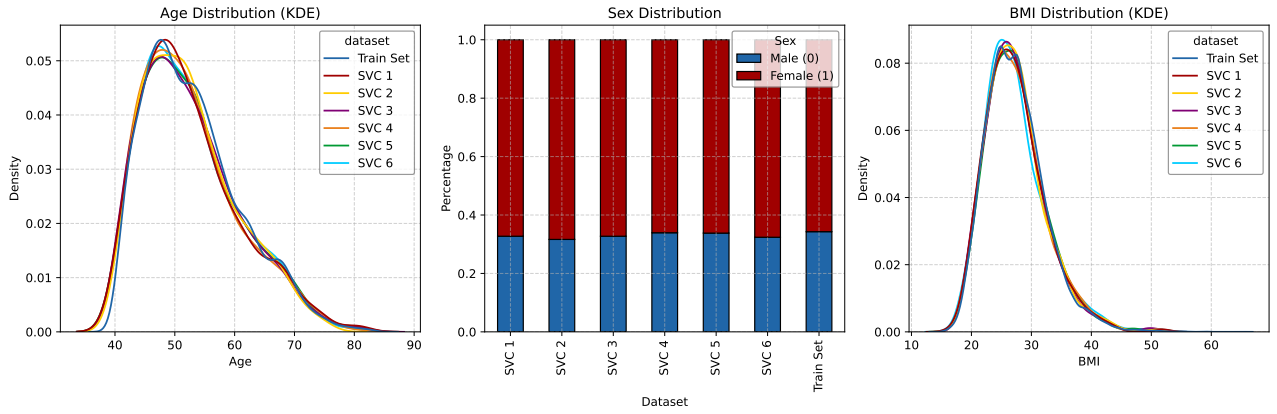


Figure S2: Distributions of age, sex, and BMI for all SVCs and the training set. SVCs were obtained by randomly assigning a `cohort_id` $\in \{1, \dots, 6\}$ to each visit without repetition per individual. Each SVC comprises all measurements sharing the same `cohort_id`. Sizes were balanced by iteratively moving a random measurement from the largest SVC to the smallest SVC whenever the individual did not already appear in the target SVC.

B Peak ratio indices

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

Table S1: IR peak ratios used to evaluate the quality of higher-order moments of the synthetic spectra in Figure 2c-d. (adapted from ¹⁶).

C CDIFF noise schedule and guidance scale tuning

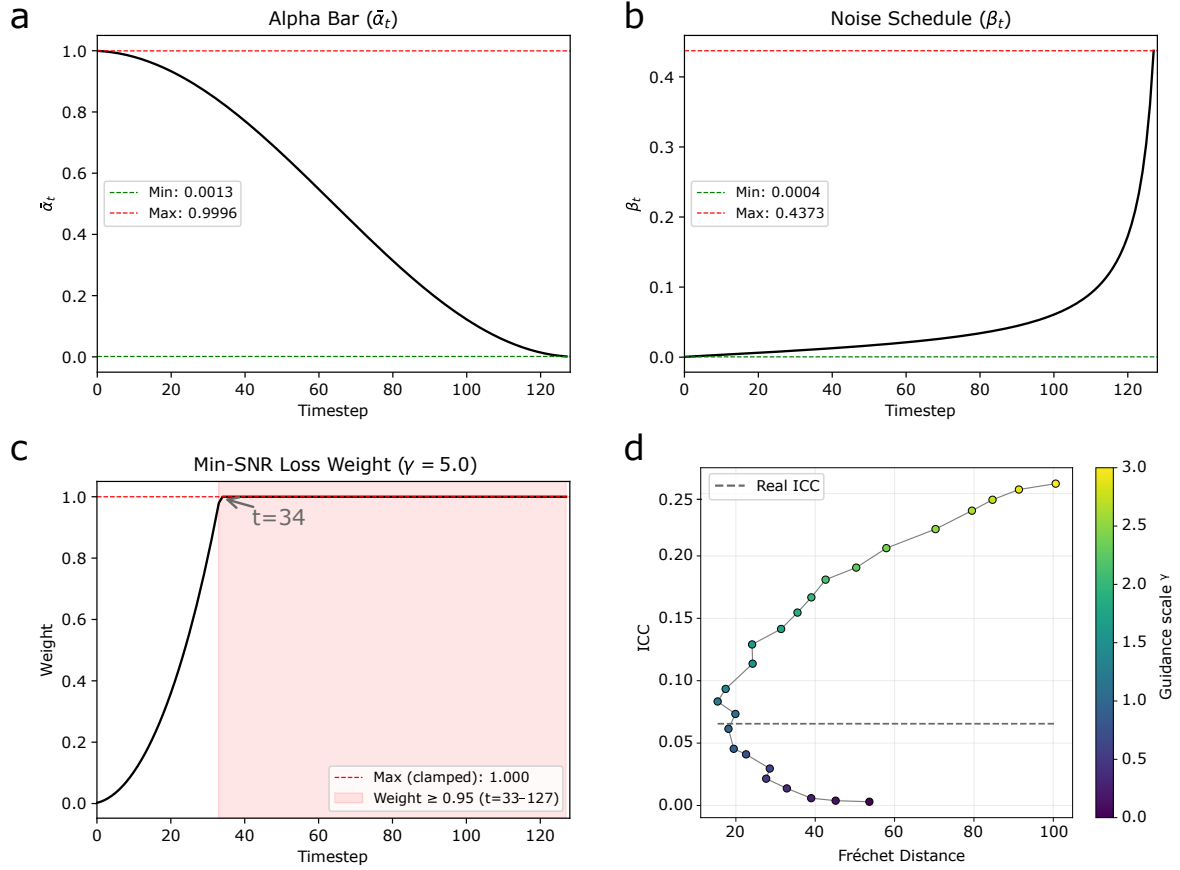


Figure S3: (a) Cumulative noise factor $\bar{\alpha}_t$ following a cosine noise schedule. (b) Resulting fraction of the input that remains β_t at timestep t . (c) Loss scaling factor w_t . (d) choice of guidance scale γ based on Fréchet distance between real and synthetic data and within-condition ICC for dichotomised covariates.

D Details on classification tasks

A separate logistic regression classifier with L2 penalty and regularization strength $C = 10$ was fit for each SVC and each covariate. The training set of a classifier is 80% of an SVC, and the test set is the remaining 20%. SVCs in turn are obtained on the held-out test set with respect to the training of the generative models (see also Figure S1). Therefore, spectra used to train the classifiers underwent the preprocessing pipeline described in the *Methods* section. Similar to the generative models, the FTIR absorbance values measured at a resolution of approximately 2cm^{-1} were used as features, while either age, sex, or BMI was the target. Resulting ROC curves in Figure 3 show the average over all SVCs.

Group	Overall	SVC 1	SVC 2	SVC 3	SVC 4	SVC 5	SVC 6
N	5062	845	843	843	843	843	845
Age high	2207	371	373	364	370	359	370
Age low	2855	474	470	479	473	484	475
BMI high	2396	395	405	390	391	413	402
BMI low	2666	450	438	453	452	430	443
Sex female	3398	571	560	572	563	561	571
Sex male	1664	274	283	271	280	282	274

Table S2: Group sizes overall and per SVC. Age and BMI were dichotomised (high/low). All experiments were conducted within cohorts. Classification tasks were performed on equally sized classes by randomly subsampling the larger class. Results were averaged across cohorts afterwards.

E Alternative metrics for assessing the quality of conditioning

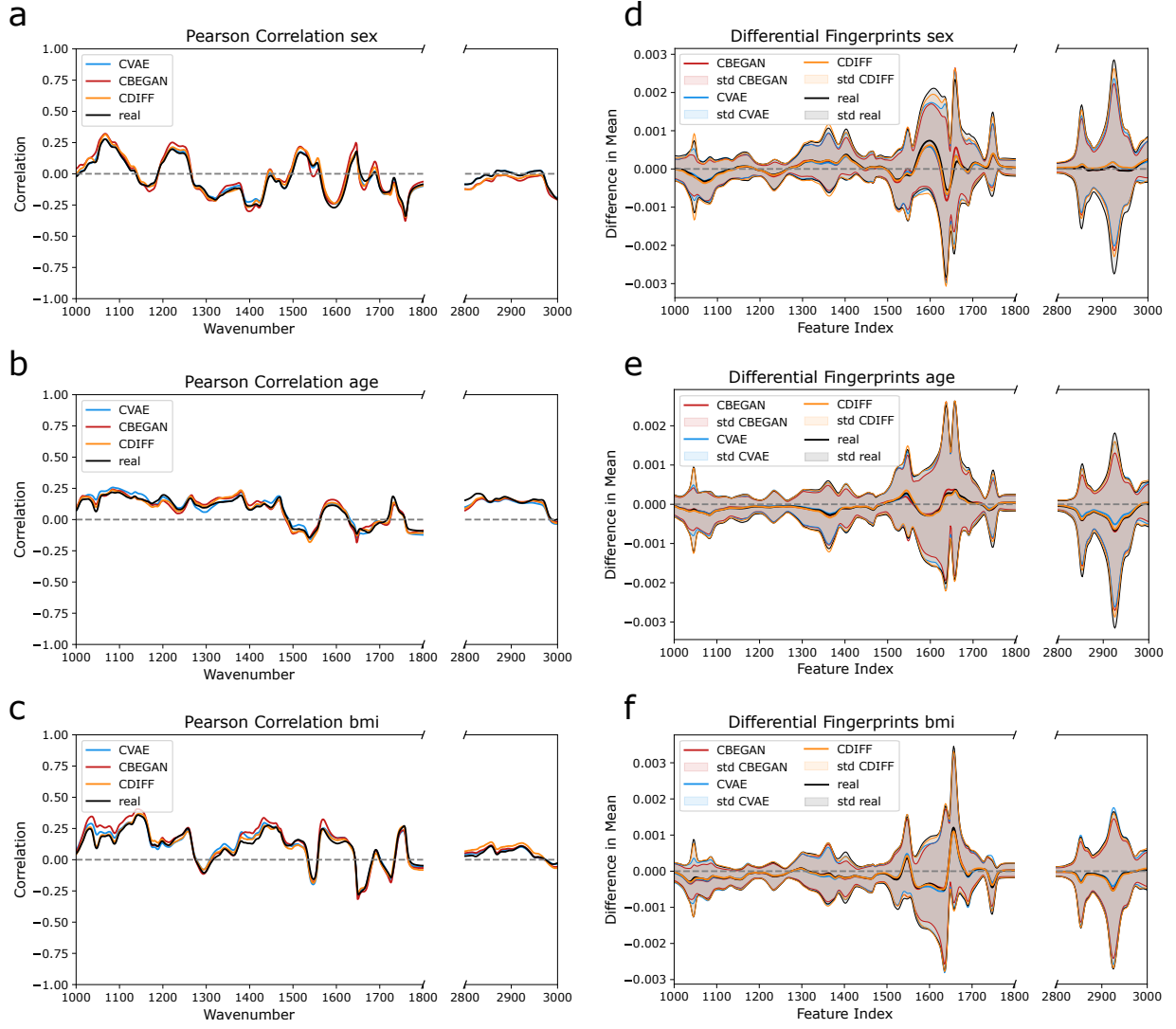


Figure S4: (a-c) Pearson correlations of real and synthetic data with the demographic covariates (non-dichotomised!). (d-f) Differential fingerprints (=difference between the mean of class 0 and the mean of class 1) between dichotomised covariates and standard deviations of real and synthetic data.

F Sensitivity to alternative age and BMI thresholds

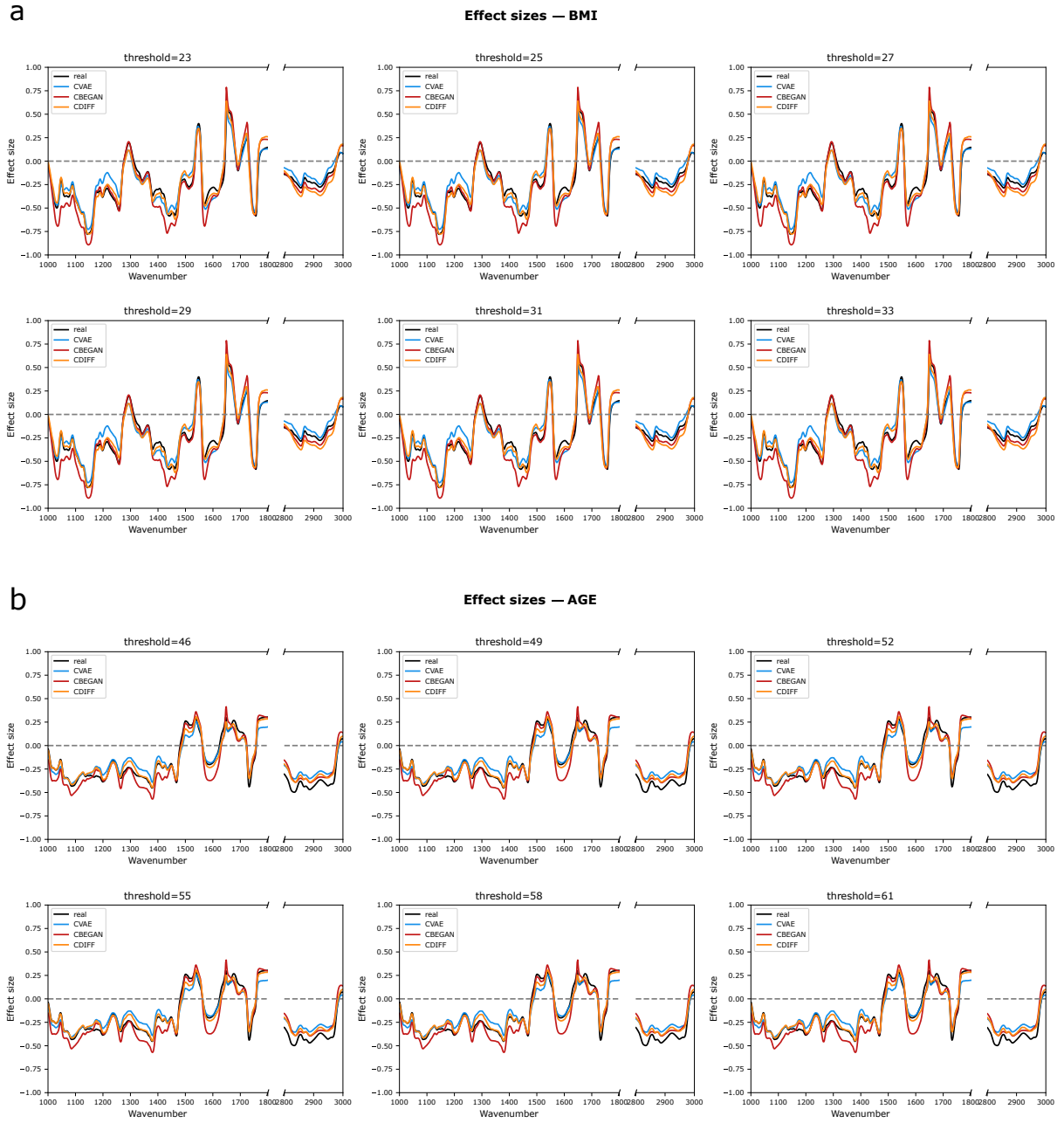


Figure S5: Sensitivity of effect sizes to alternative thresholds for age and BMI dichotomization. (a) Effect sizes when varying the BMI threshold from 23 to 33. (b) Effect sizes when varying the age threshold from 46 to 61 years old.

BMI threshold	real	CVAE	CBEGAN	CDIFF
23	0.779 ± 0.045	0.732 ± 0.054	0.759 ± 0.038	0.780 ± 0.023
25	0.793 ± 0.031	0.715 ± 0.032	0.737 ± 0.034	0.777 ± 0.016
27	0.798 ± 0.041	0.711 ± 0.042	0.753 ± 0.044	0.793 ± 0.034
29	0.789 ± 0.043	0.705 ± 0.039	0.795 ± 0.022	0.787 ± 0.042
31	0.781 ± 0.040	0.718 ± 0.056	0.789 ± 0.028	0.784 ± 0.039
33	0.769 ± 0.064	0.739 ± 0.065	0.793 ± 0.023	0.767 ± 0.051

Table S3: ROC-AUC scores for logistic regression classifiers trained on real (first column) or synthetic data to predict the BMI label when varying the BMI threshold. The larger class was subsampled to ensure a 50/50 split between classes. For every column, testing was performed on real data.

age threshold	real	CVAE	CBEGAN	CDIFF
46	0.730 ± 0.040	0.688 ± 0.026	0.715 ± 0.030	0.677 ± 0.024
49	0.749 ± 0.043	0.708 ± 0.028	0.733 ± 0.036	0.757 ± 0.026
52	0.790 ± 0.032	0.737 ± 0.052	0.761 ± 0.037	0.795 ± 0.030
55	0.780 ± 0.021	0.719 ± 0.037	0.770 ± 0.036	0.783 ± 0.023
58	0.773 ± 0.037	0.745 ± 0.038	0.775 ± 0.049	0.797 ± 0.033
61	0.760 ± 0.059	0.760 ± 0.065	0.766 ± 0.070	0.779 ± 0.043

Table S4: ROC-AUC scores for logistic regression classifiers trained on real (first column) or synthetic data to predict the age label when varying the age threshold. The larger class was subsampled to ensure a 50/50 split between classes. For every column, testing was performed on real data.

G Covariate regression tasks

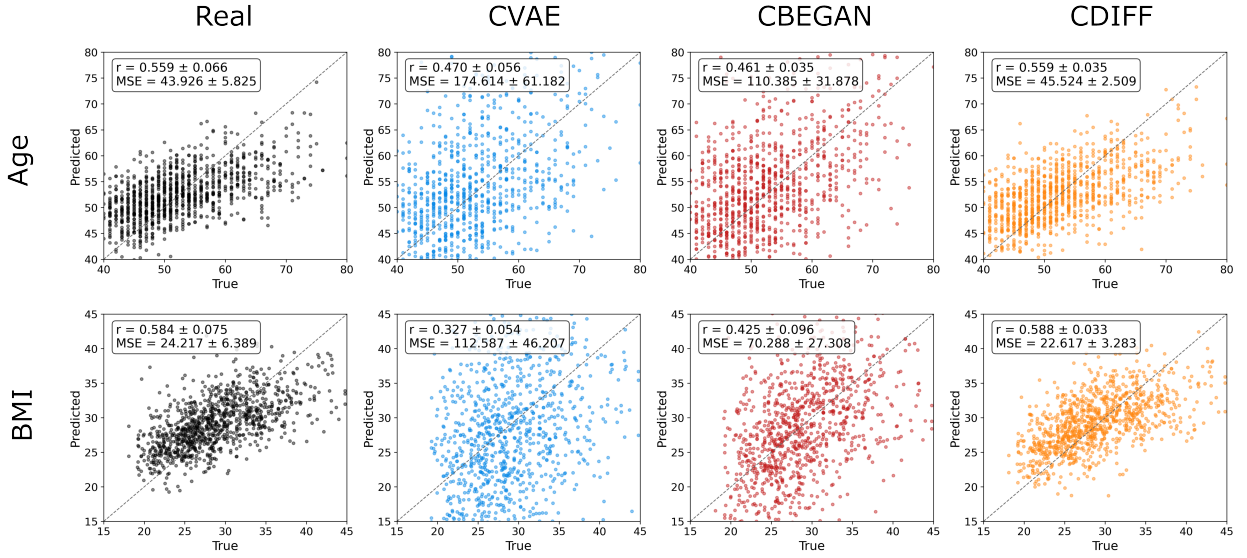


Figure S6: Elastic net regression scatter plots for predicting age and BMI. Models were trained per SVC; Scatter plots show all SVCs overlaid. Trained on real (first column) or synthetic data, tested on real data for every column. This setup is analogous to the setup used for binary classification in Figure 3(d-f).

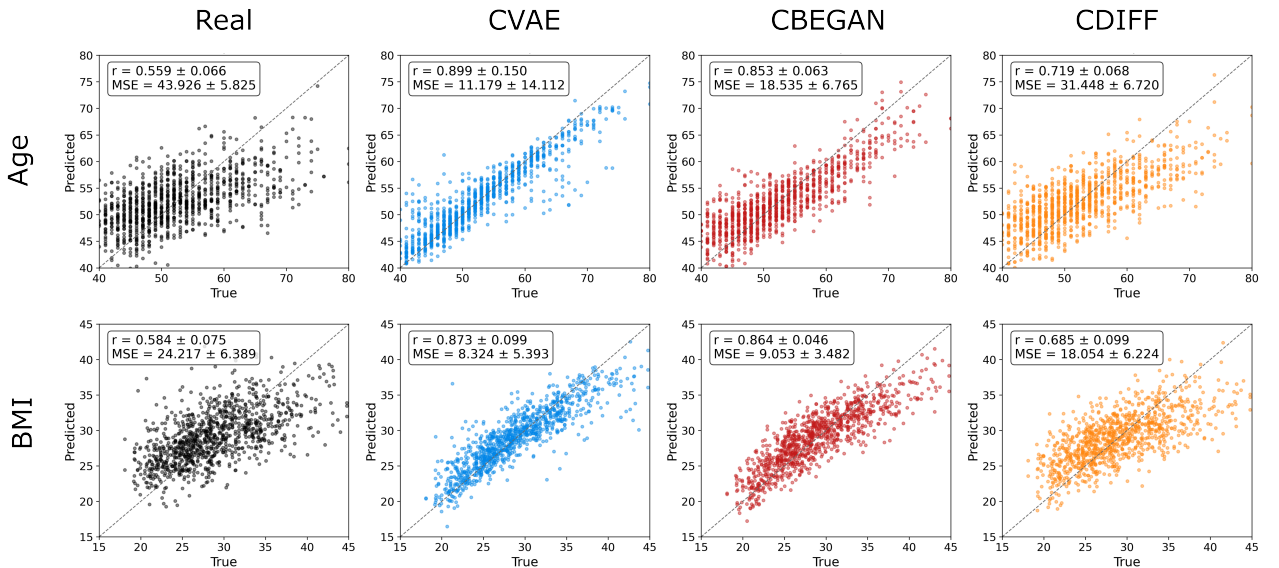


Figure S7: Elastic net regression scatter plots for predicting age and BMI. Models were trained per SVC; Scatter plots show all SVCs overlaid. In this setup, the training and test sets are from the same source, i.e., real data or from the same generative model. This setup tests how strongly each condition was encoded without comparing it to real data samples.

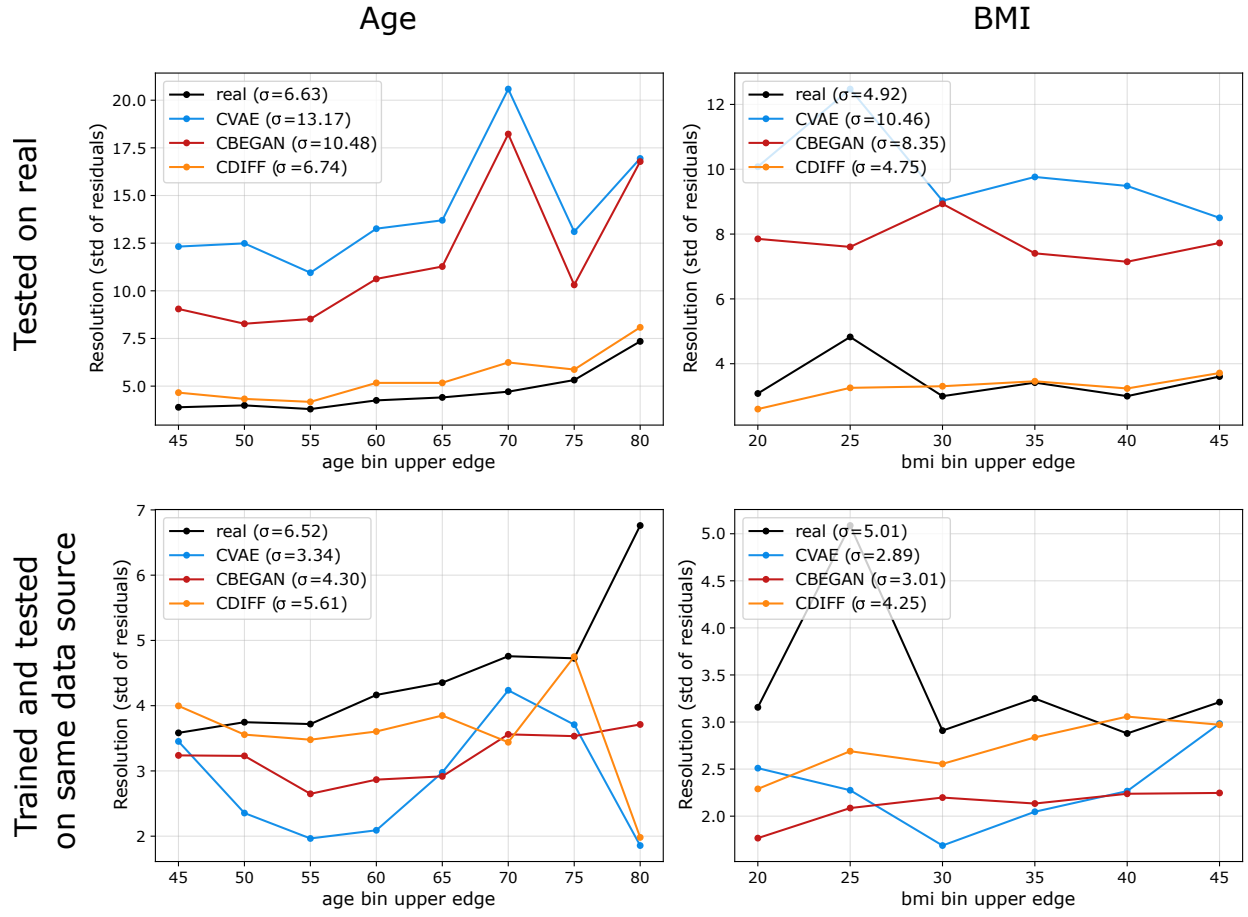


Figure S8: Resolutions at a binning range of 5 for both age and BMI. The top row corresponds to the regression tasks in Figure S6, and the bottom row corresponds to the regression tasks in Figure S7. "Data source" refers to either real data or synthetic data of a generative model.

H Features with high correlation and Index of Individuality

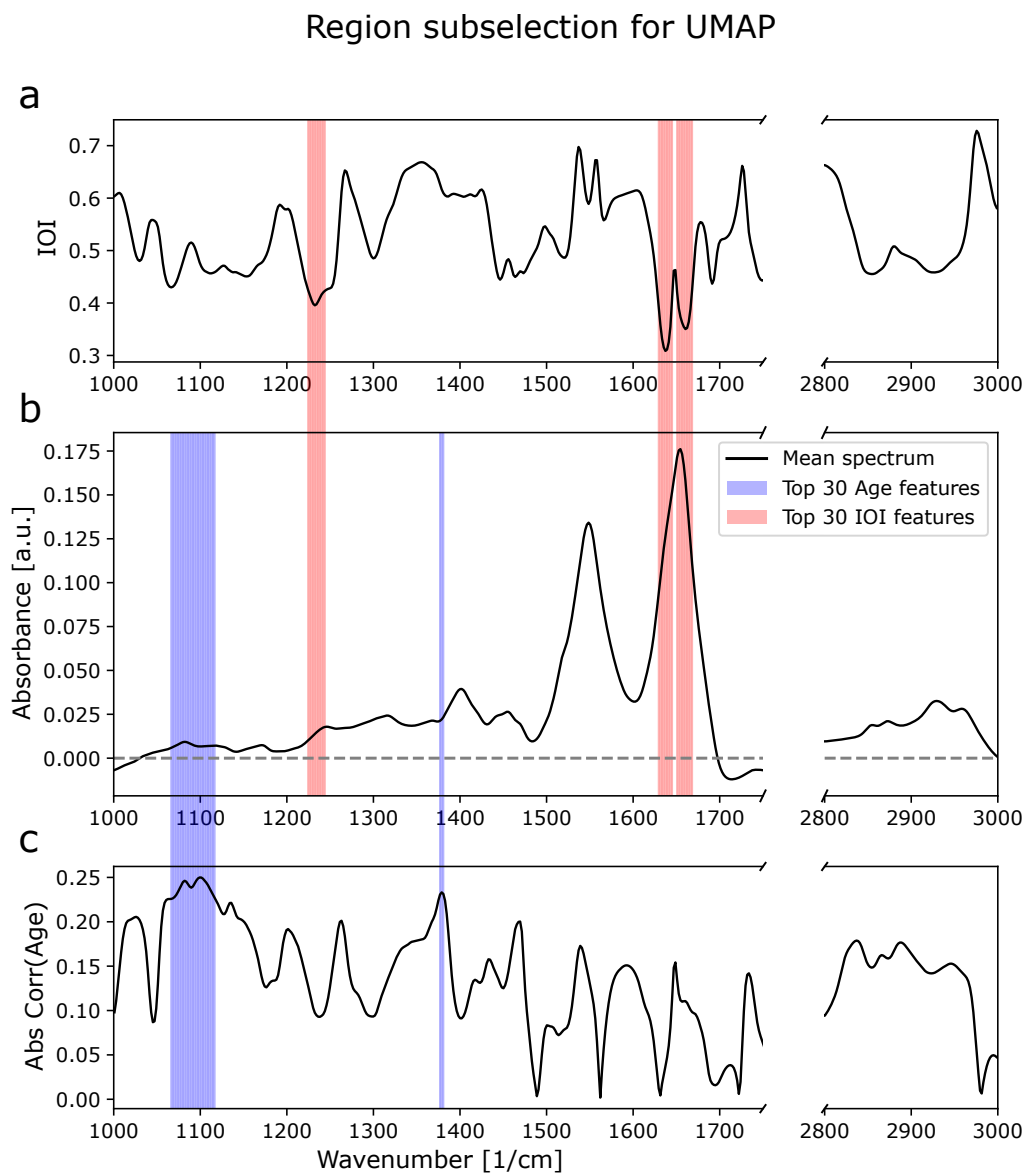


Figure S9: Regions selected for dimensionality reduction in Figure 4. (a) Index of Individuality, (b) mean FTIR spectrum, (c) absolute Pearson correlation with age per wavenumber.

I All classification tasks for underrepresented groups

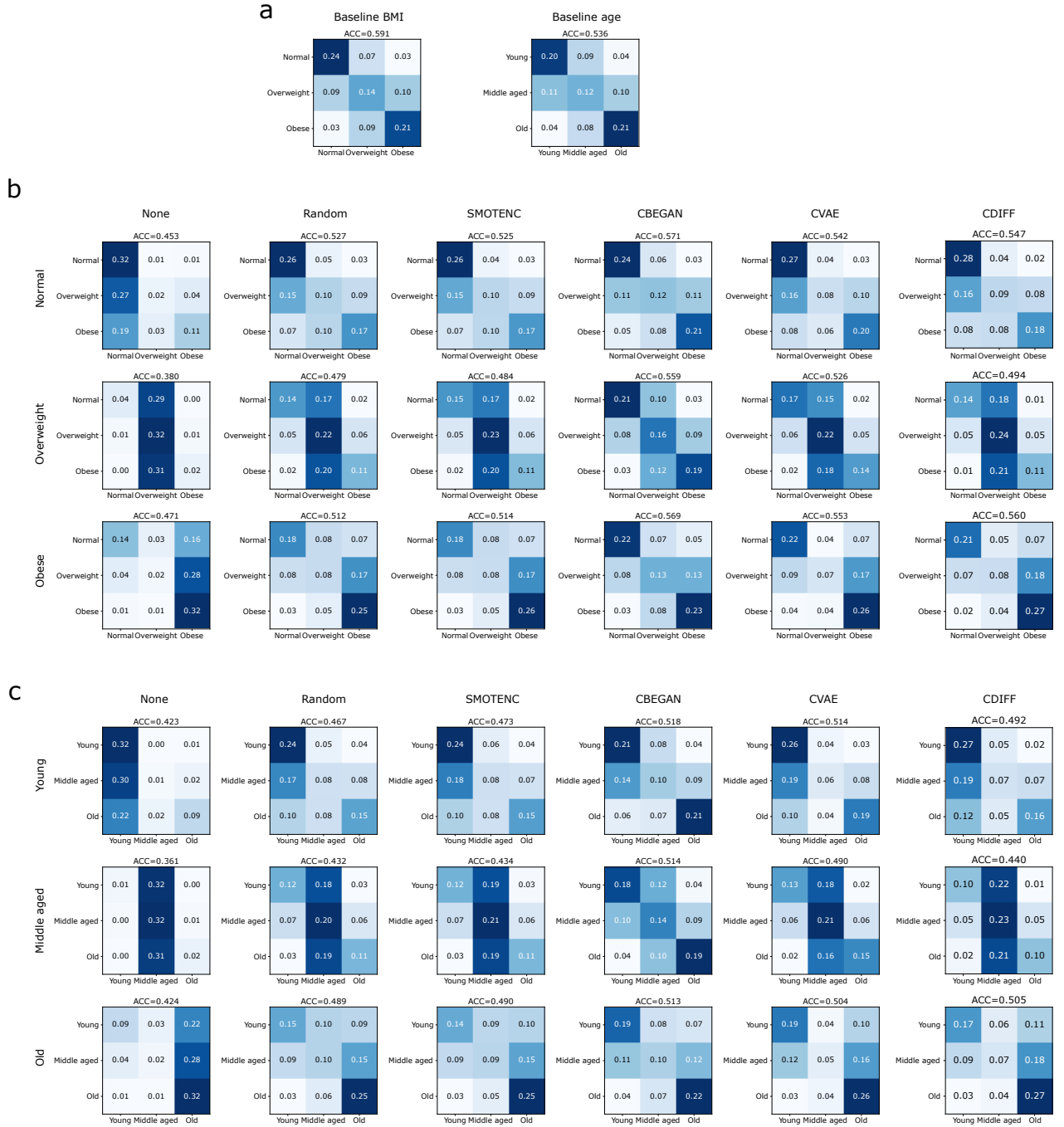


Figure S10: Evaluation of oversampling methods for underrepresented groups. Random and SMOTENC were implemented via scikit-learn¹⁷. In each matrix, the x-axis shows the predicted label and the y-axis shows the true label. (a) Baseline without undersampling. (b) Age classification (thresholds 49, 55). Similar to the ROC curves in Figure 3, the age thresholds were chosen to enable balanced classes of the largest possible size. (c) BMI classification (thresholds 25, 30). Thresholds according to the common separation between normal, overweight, and obese¹⁸

J Training progress

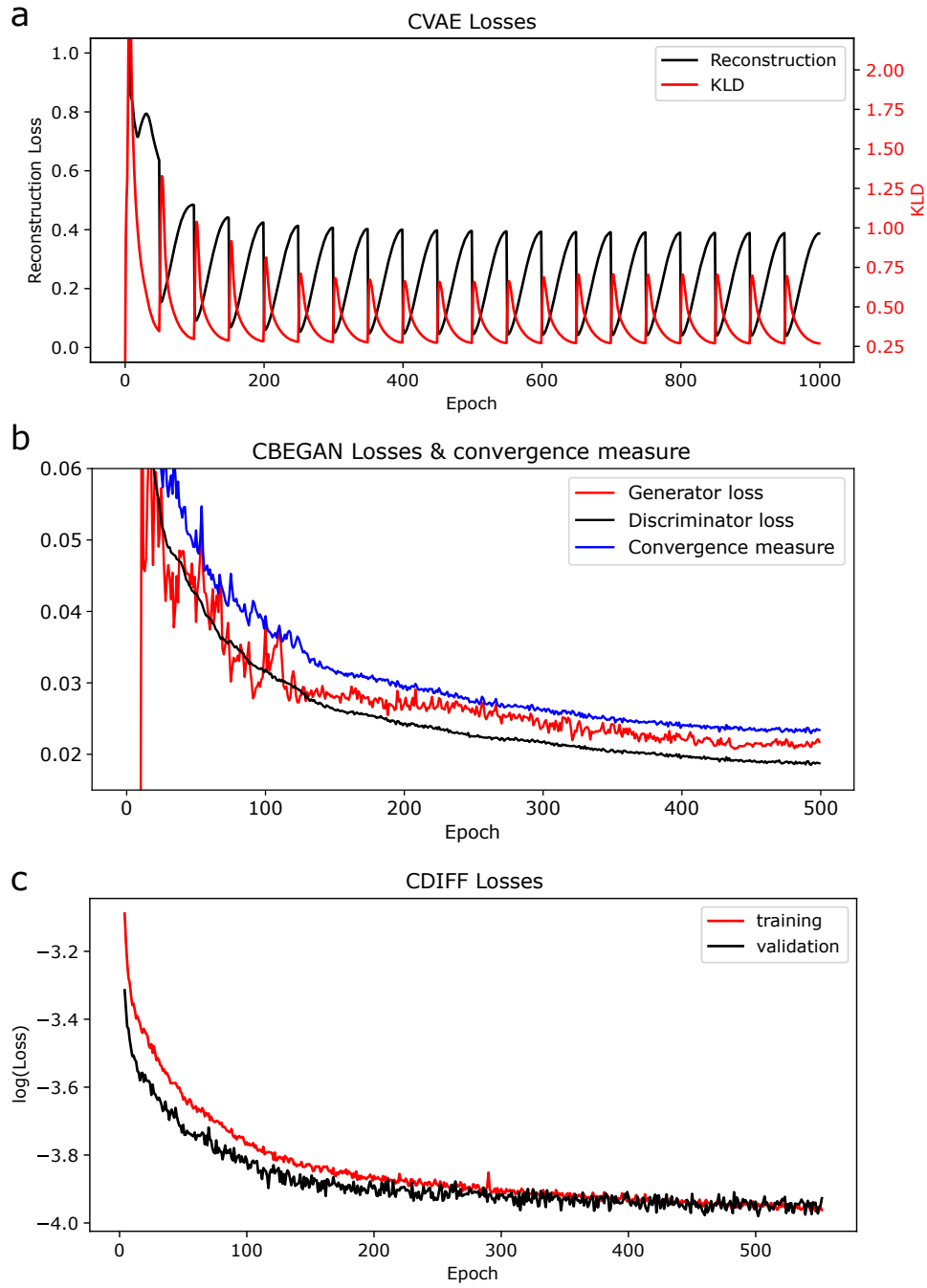


Figure S11: Training progress. (a) CVAE reconstruction (MSE) and KL loss with cyclic β . (b) CBEGAN generator/discriminator losses and convergence measure. (c) CDIFF training and validation loss.

References

- [1] 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).
- [2] 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).
- [3] 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).
- [4] 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).
- [5] 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).
- [6] 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).
- [7] 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).
- [8] 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).
- [9] 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).
- [10] 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).
- [11] 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).
- [12] Movasaghi, Z., Rehman, S. & ur Rehman, D. I. Fourier transform infrared (ftir) spectroscopy of biological tissues. *Applied Spectroscopy Reviews* **43**, 134–179 (2008).
- [13] 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).
- [14] 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).
- [15] 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).
- [16] Huber, M. *et al.* Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. *Nature communications* **12**, 1511 (2021).
- [17] Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).
- [18] 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).