

<https://doi.org/10.1038/s41540-025-00644-5>

Cross-platform metabolomics imputation using importance-weighted autoencoders



Alexander Smith¹ ✉, Rui Pinto^{1,2,3}, Loukas Zagkos¹, Ioanna Tzoulaki^{1,2,4,5}, Paul Elliott^{1,2,4} & Abbas Dehghan^{1,2,4}

Metabolomics data are often generated through different platforms and quantification methods which makes their synthesis and large-scale replication challenging. This study developed an ensemble of importance-weighted autoencoders to perform cross-platform metabolomics imputation between two metabolomics platforms, Metabolon and National Phenome Centre (NPC) at Imperial College, using 979 samples from the Airwave Health Monitoring Study. The generated samples were highly correlated with real values across all metabolites ($\mu_p = 0.61$ (0.55–0.67)). The well-imputed subset contained 199 metabolites (22%), capturing $\geq 55\%$ variance ($R^2 \geq 0.55$) with minimal uncertainty (R^2 variance ≤ 0.025), including 43 metabolites unique to Metabolon. The concordance of associations in 2,971 validation samples between real and imputed metabolites with two clinical outcomes, body mass index (BMI) and C-reactive protein (CRP), were highly correlated ($\rho_{\text{BMI}} = 0.93$; $\rho_{\text{CRP}} = 0.89$) with minimal mean difference (BMI $\mu_\Delta = 0.005$ (0.04); CRP $\mu_\Delta = 0.005$ (0.04)). Similar concordance occurred with equivalent UK Biobank (BMI $\mu_\Delta = -0.007$ (0.05); CRP $\mu_\Delta = 0.01$ (0.04)) and NPC (BMI $\mu_\Delta = -0.013$ (0.04); CRP $\mu_\Delta = -0.019$ (0.04)) metabolites. This methodological innovation offers a scalable and accurate method for cross-platform imputation, enabling the aggregation of metabolomics data from different epidemiological studies for replication and meta-analyses.

Metabolomics, a critical field in systems biology, focuses on the comprehensive analysis of metabolites - important small molecules in biological systems. Metabolites can be exogenous, like nutrients and drugs, or endogenous, like cholesterol and triglycerides. They are vital building blocks for tissues and have roles as regulatory agents and messengers as well as many other functions. They help in energy production, protein synthesis, metabolism regulation, and other cellular functions, thereby maintaining an organism's internal balance and health. The study of metabolic profiles provides insights into the metabolic pathways within an organism at a given time, enhancing the understanding of the biological interplay between physiological and pathological conditions. This understanding is crucial for advancing biomarker discovery, disease diagnosis, treatment monitoring, and drug discovery^{1–3}.

In recent years a growing number of large epidemiologic studies have acquired metabolomics data⁴ but often suffer from small sample sizes and/or limited metabolite coverage. Large studies may have ten to a hundred thousand samples but small metabolite coverage, e.g., UK Biobank

lipoprotein/metabolite dataset⁵, while smaller ones contain hundreds to thousands of samples with large metabolite coverage, e.g. Multi-Ethnic Study of Atherosclerosis (MESA)⁶. The latter group of studies would greatly benefit from combining their data with additional cohorts to form a larger discovery panel or robustly externally validating their findings. However, the variability in metabolomics platforms, annotation and quantification prevents easily combining or comparing metabolomic features across studies. Several methods have been developed to combine liquid chromatography-mass spectrometry (LC-MS) metabolomics datasets that utilize the same or similar platforms^{7,8}. However, to date, none have been published which entirely recreates one dataset from another.

An innovative solution is to treat the variables of each dataset as “missing” in the others. This solution is feasible given the high correlation between metabolites. Moreover, recent advances in applying deep latent variable models, such as importance-weighted autoencoders⁹, provide a highly flexible approach to imputation in the complex, high dimensional datasets generated by metabolomics studies. The complex associations

¹Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. ²UK Dementia Research Institute at Imperial College London, Hammersmith Hospital, London, UK. ³MRC-NIHR BRC National Phenome Centre, Department of Digestion, Metabolism and Reproduction, Division of systems medicine, Imperial College London, London, UK. ⁴Medical Research Council Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. ⁵Centre for Systems Biology, Biomedical Research Foundation of the Academy of Athens, Athens, Greece.

✉ e-mail: alexander.smith@imperial.ac.uk

between metabolites caused by interconnected biological pathways can be captured by such deep learning models using their ability to model infinitely complex functions, including non-linear relationships. In addition, an importance-weighted autoencoder specifically developed for imputation has shown state-of-the-art imputation performance¹⁰, yet to date it has not been explored for cross-platform metabolomics imputation.

This study showcases the imputation of an entire dataset of the Metabolon platform from untargeted LC-MS dataset acquired by the UK National Phenome Centre (NPC), using an ensemble importance-weighted autoencoder. The imputations were then validated by comparing the observational associations of the imputed metabolites in the imputed Metabolon dataset with body mass index (BMI) and C-Reactive protein (CRP), to the associations of the real Metabolon metabolites with the same outcomes. Such an innovative approach could be very beneficial, as many epidemiological studies have metabolomics datasets using different approaches and platforms. The suggested method allows for a one-to-one comparison of the metabolites for validation purposes or aggregate analyses across studies.

Results

Metabolite clustering

Clustering the metabolites within the Metabolon-NPC dataset using the UMAP-embedded space identified five clusters (Fig. 1). Based on their majority compound classes, four clusters were labelled Fatty acids, Sphingolipids, Phospholipids, and Acyl-glycerides. The fifth cluster was labelled Others for its lack of a predominant class (see Supplementary Data 1 for full details). The clusters varied in size and composition ranging from a total of 356 metabolites for Fatty acids, including 134 Metabolon metabolites and 222 NPC metabolomic features, up to 503 metabolites for Acyl-glycerides (24 Metabolon).

Metabolon imputation performance

IWAE imputation models, trained on each metabolite cluster, achieved low mean absolute error (MAE) on the test set samples (Fatty acids = 0.141, Sphingolipids = 0.140, Phospholipids = 0.132, Acyl-glycerides = 0.125 and Others = 0.190). The imputation model trained using the entire set of metabolites achieved a root mean squared error (RMSE) of 0.251 on the test set. The error rate (MAE or RMSE) for imputation on

the combined training-validation set achieved results similar to the test set.

After greedy selection in the test set, the imputation of 644 metabolites achieved an $R^2 \geq 0.1$ (Fig. 2A), of which 199 were equal to or surpassed an R^2 of 0.55. Out of these 199 metabolites, 43 were unique to Metabolon compared to the NPC LC-MS platform (Fig. 2B). Uncertainty of imputation showed an inverse linear relationship with overall imputation performance (Fig. 2A). The mean R^2 variance for metabolites imputed with overall $R^2 \geq 0.1$ was 0.014, which drops to 0.009 for metabolites with overall $R^2 \geq 0.55$.

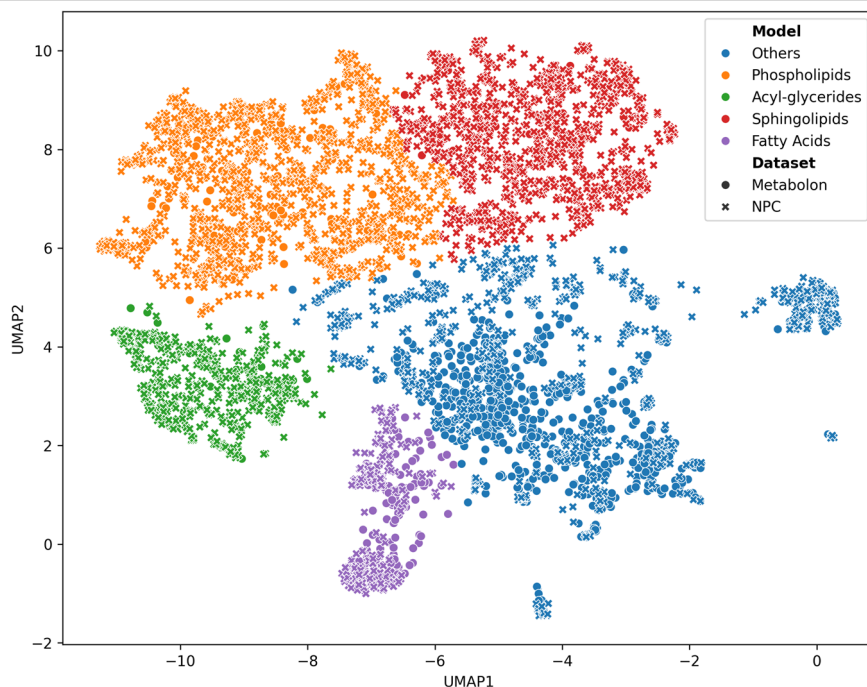
Metabolon categorises metabolites into superclasses based on their compound class. Imputation performance varied greatly across superclasses (Fig. 3). The “Lipid” superclass was the best imputed with a mean $R^2 = 0.49$, although it had the largest variance in performance (SD = 0.25). Xenobiotics (mean $R^2 = 0.13$, SD = 0.15), Amino acids, (0.14, 0.14) Peptides (0.22, 0.15), and Cofactors and Vitamins (0.24, 0.25) had a combined total of 35 metabolites which imputed over an R^2 of 0.3 but overall, they had low imputation performance. Carbohydrates, Energy and Nucleotide metabolites had no outstanding metabolites with poor mean $R^2 = 0.11$ (SD = 0.08), 0.16 (SD = 0.08) and 0.11 (SD = 0.10) respectively.

After imputation of all Metabolon metabolites, the test set samples ($n = 78$) had a mean correlation of $\rho = 0.61$ (SD = 0.09, IQR 0.55–0.67), with the real Metabolon metabolites using the entire metabolic profile (Fig. 2C). The metabolites that had an $R^2 \geq 0.55$, were taken forward for association studies (Fig. 2A). Full details of the imputation model parameters and average performance on the training sets can be found in Supplementary Data 2. Individual Metabolon metabolite imputation performances on the test set can be found in Supplementary Data 3.

Observational association concordance

The study's rigorous validation process involved comparing the imputed metabolites' associations with BMI and CRP to those of actual Metabolon metabolites using data from the combined Metabolon-NPC and Metabolon-Only datasets (Fig. 4). The beta coefficients were highly correlated between the real and imputed values with $\rho = 0.93$ and $\rho = 0.89$ for BMI and CRP respectively. Based on the standardised beta coefficient, the average association of real metabolites was 0.06 (SD = 0.12) with BMI and 0.01 (SD = 0.08) with CRP. The mean difference between the real and

Fig. 1 | Metabolite clustering within the Metabolon-NPC dataset ($N = 979$) using both NPC and Metabolon metabolites.



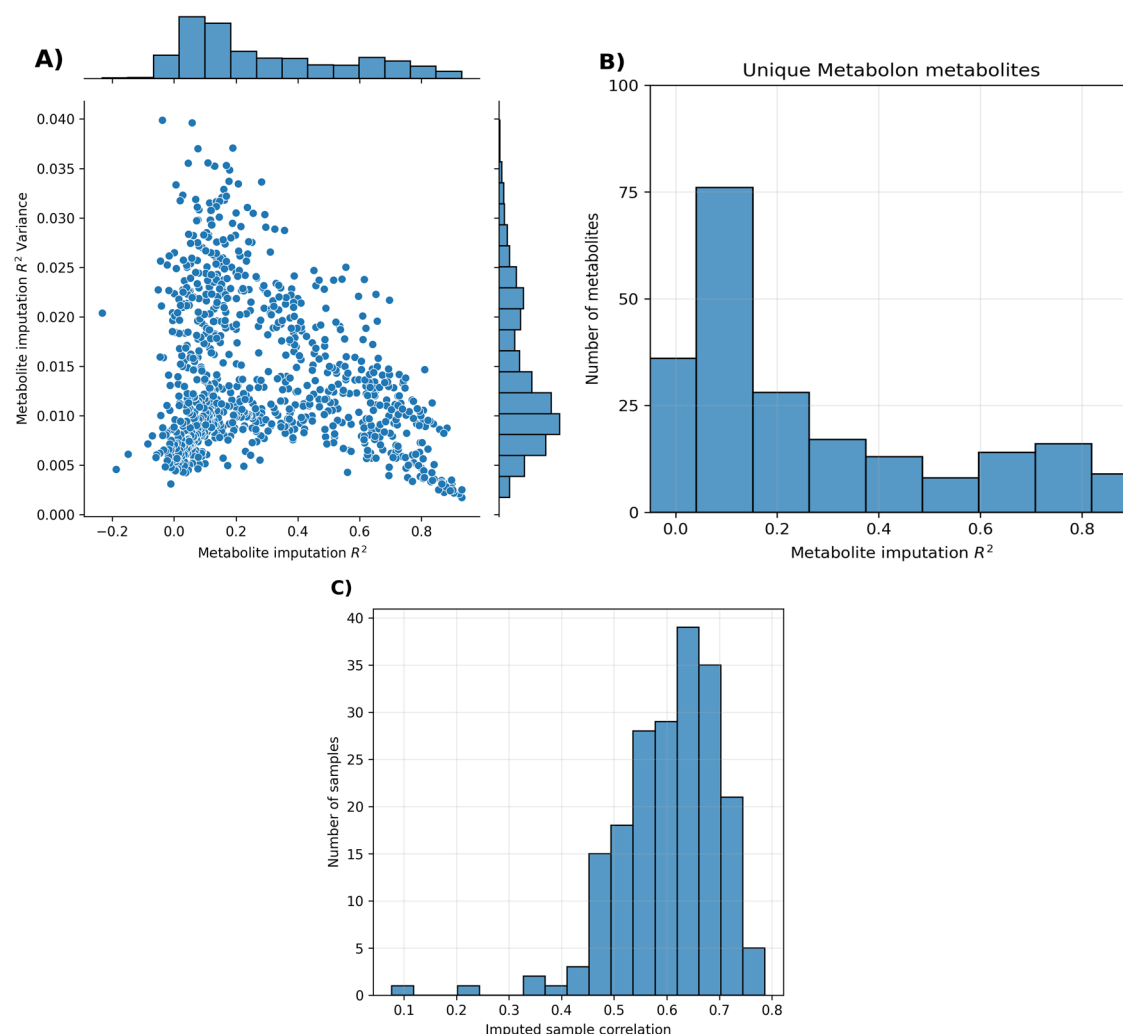


Fig. 2 | Metabolite imputation performance on the held-out test set. A metabolite R^2 vs R^2 variance (B) performance within the unique Metabolon metabolites (C) imputed held-out test set sample correlation with real values.

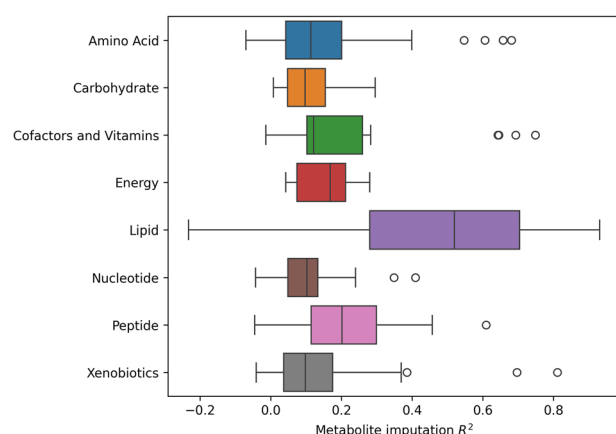


Fig. 3 | Imputation performance across metabolite compound classes based on Metabolon superclass classification.

imputed metabolite associations was 0.005 (SD = 0.04) with BMI and 0.005 (SD = 0.04) with CRP. Out of the 199 metabolites tested, 10 metabolites (5% of all metabolites) had a beta coefficient difference greater than 1.96 standard deviations from the mean for BMI whilst CRP had 6 metabolites greater than 1.96 standard deviations (3% of all metabolites). The association comparison between real NPC LC-MS metabolites with the imputed

versions had similar concordances with a mean difference of -0.013 (SD = 0.04) for BMI and -0.019 (SD = 0.04) for CRP (Fig. 5). For the four UK Biobank NMR metabolites which could be matched to the imputed Metabolon dataset (NPC-Only), the mean difference in association with the imputed metabolites was -0.007 (SD = 0.05) and 0.01 (SD = 0.04) for BMI and CRP respectively. Full results of the association effect sizes for all models can be found in Supplementary Data 4 and 5.

Discussion

In this study, deep learning techniques were applied to train a model ensemble to generate an imputed Metabolon dataset comprising 199 metabolites, representing 22% of the metabolites in the original Metabolon dataset, using LC-MS metabolomic features from three platforms. The results showed that the imputed dataset is expected to capture a minimum of 55% of the original variance ($R^2 \geq 0.55$) with low imputation uncertainty (R^2 variance ≤ 0.025). Using the trained model, an imputed Metabolon dataset of 199 metabolites was generated for 2971 samples which was combined with the Airwave Metabolon, Airwave NPC and UK Biobank datasets to run observational associations between metabolites and two important clinical outcomes: BMI and CRP. The results revealed strong alignment between associations observed using imputed Metabolon data and those found using real metabolites, underscoring the success of the model imputations.

The robustness of the approach was explored using a held-out test set to simulate the imputation of an external validation cohort. Various aspects of imputation performance were explored, including overall

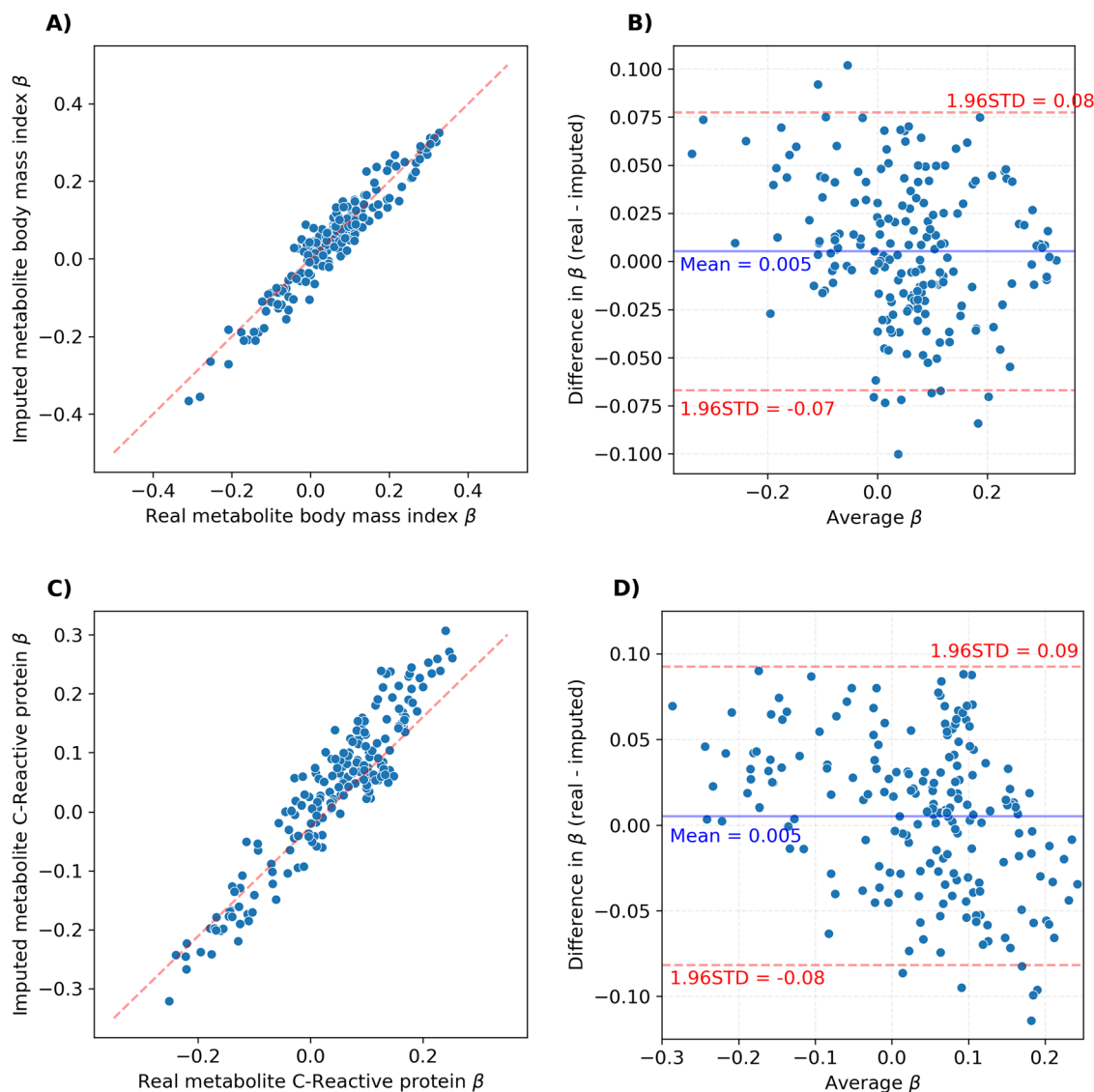


Fig. 4 | Comparison between real and imputed metabolite associations with body mass index (BMI) and C-Reactive protein (CRP). A Real vs imputed effect size for BMI (B) Bland-Altman plot of real and imputed effect size for BMI (C) Real vs. imputed effect size for CRP (D) Bland-Altman plot of real and imputed effect size for CRP.

accuracy (R^2), model uncertainty (R^2 variance), and correlation between imputed and real samples. The imputed dataset performed well across all metrics (Fig. 2), confirming that the method is reliable for cross-platform imputation in external cohorts. To validate this further, comparative observational analysis was performed using the imputed dataset generated from an external cohort (NPC-Only) with three real metabolomics datasets. This was performed on two clinically relevant outcomes, BMI and CRP, which have well-documented metabolite associations¹¹ to ensure a large set of significant metabolite associations where available for comparison. The concordance of associations for both outcomes across with all three real datasets and the imputed Metabolon dataset supported the reliability of the imputation approach (Figs. 4 and 5).

The imputed Metabolon dataset primarily consisted of lipid metabolites (Fig. 3), which was expected since the LC-MS platforms used for the imputation measure predominantly lipid compounds. This suggests that the presence of a closely related feature within the input dataset is one of the most important factors for metabolite imputation performance. This is additionally supported by the fact that small molecule classes, such as peptides and carbohydrates, showed much better imputation performance in Airwave 2 when

using an NMR dataset as the input (Fig. 6). Lipid imputation performed consistently in both settings which is likely influenced by the low intra-laboratory variability within lipidomics, leading to a higher baseline imputation performance. In addition, given the relatively small sample size ($N = 979$), constraints were applied to the size (number of layers and number of neurons) and training duration (early stopping) of the imputation models to mitigate the risk of overfitting. As a result, the imputation model trained on all metabolites and metabolomic features prioritized the identification of lipid-imputing features over other metabolite classes. However, recognizing the importance of accurately imputing non-lipid metabolites, the imputation method was expanded to an ensemble which included models trained on each of the identified metabolite clusters, as depicted in Fig. 1. The switch from a single model to an ensemble employing a greedy selection strategy significantly improved the imputation performance of non-lipid metabolites. This further suggests that incorporating additional metabolomics platforms, such as NMR or reverse phase, would further enhance imputation across a broader range of metabolites, resulting in a more diverse imputation dataset.

The generative deep learning models implemented in this study have shown state of the art imputation for many data types such as imaging and

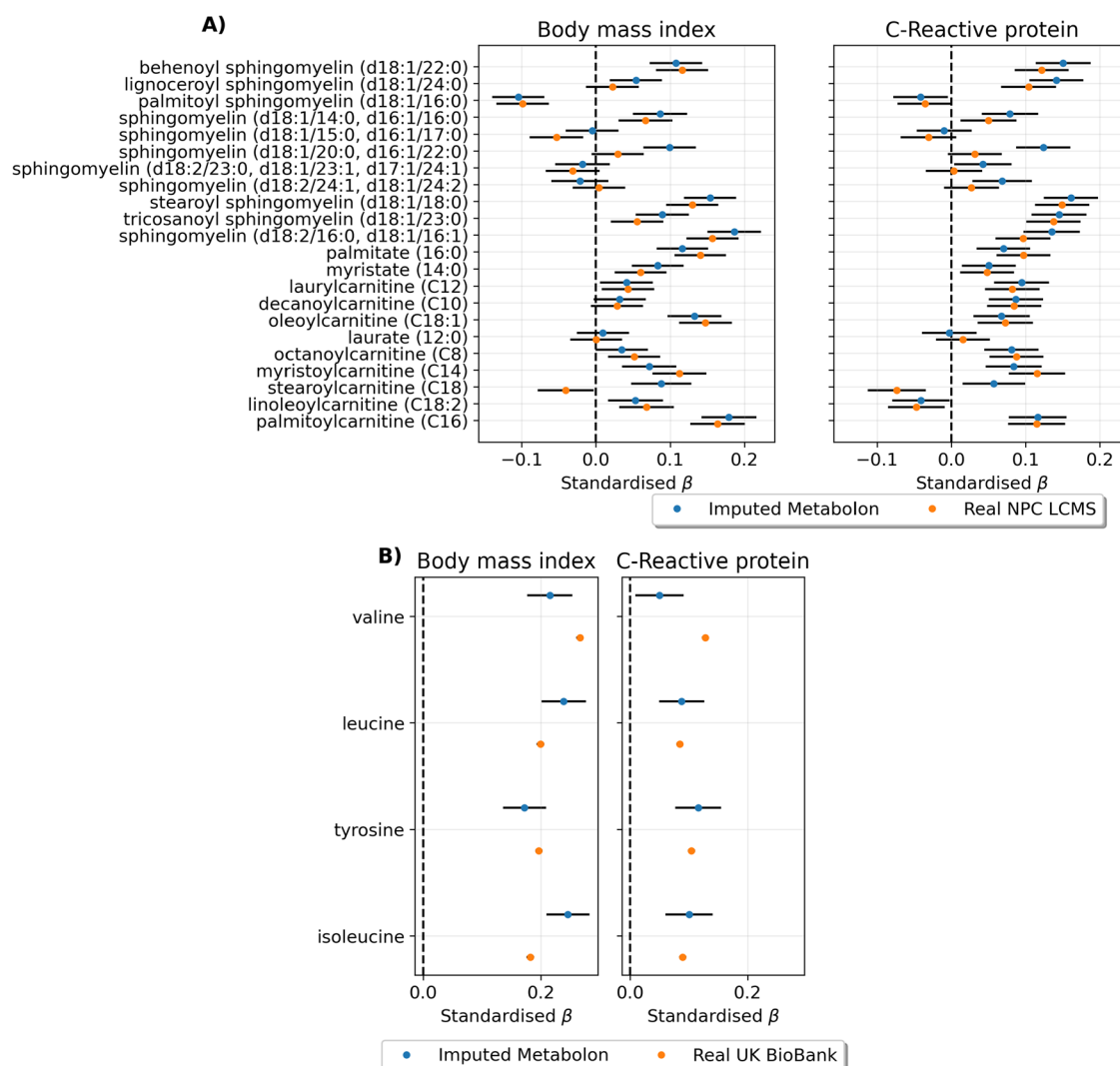


Fig. 5 | Comparison between real and imputed metabolite associations in the NPC-Only and UK BioBank NMR datasets. A Name-matched real NPC LC-MS (B) Real UK BioBank NMR.

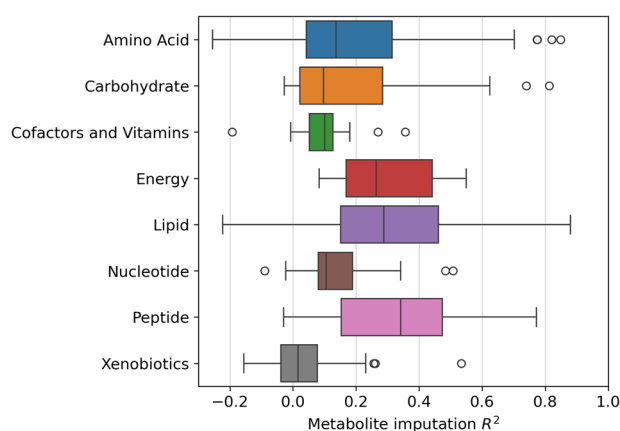


Fig. 6 | Imputation performance across metabolite compound classes based on Metabolon superclass classification using only features from NMR.

RNA-seq data¹². This success stems from the maximisation of a tighter lower bound of the observed data compared to other imputation methods, combined with their ability to make use of powerful multiple imputation methods with incomplete datasets. The study results demonstrate their

extension to cross-platform metabolomics imputation by effectively identifying key features in an LC-MS dataset capable of inferring Metabolon metabolites. The findings also highlights the potential application of IWAE for annotating metabolic features. By applying feature importance methods, such as SHapley Additive exPlanation (SHAP) values¹³, it would be possible to identify the specific LC-MS features used to infer each Metabolon metabolite. With this information, given that there seems to be a strong relationship between input features and imputed metabolites, the calculated feature importance could be used infer the annotation of unknown LC-MS features. However, further investigation into the feasibility and accuracy of this method for metabolomic feature annotation is warranted.

The pioneering use of importance-weighted autoencoders to build an ensemble method for cross-platform imputation represents a meaningful step forward in untargeted LC-MS metabolomics for epidemiologic studies. The method shows robustness and precision in generating imputed datasets for external cohorts which is crucial for downstream analyses, including associations with clinical outcomes. At this stage, the approach is best suited for applications such as cross-cohort harmonisation and replication of well-established metabolite–phenotype associations (e.g., BMI, CRP), rather than for substituting direct metabolite quantification. However, this study encountered limitations inherent to the LC-MS dataset used to impute Metabolon metabolites, resulting in suboptimal performance in the imputation of non-lipid metabolites. Additionally, the limited number of training

samples necessitated restrictions on the deep learning model's capacity and training time to prevent overfitting, which likely contributed to lower imputation performance. The imputations also require a large number of features which can likely be drastically reduced given the high co-linearity of the features, which should be explored in future studies. Despite these limitations, this study highlights the potential of importance-weighted autoencoders and ensemble methods to improve cross-platform imputation and their potential application to assist with metabolomics feature annotation.

In conclusion, this study demonstrates the efficacy of using IWAEs to impute metabolomics data across different platforms, specifically between Metabolon and LC-MS datasets. The high concordance observed between real and imputed metabolites, particularly in associations with BMI and CRP, supports the robustness of this method. This approach offers a scalable solution to harmonize metabolomics data from diverse analytical platforms, facilitating data integration for replication studies and meta-analyses in large-scale epidemiological research.

Methods

Data sources

The Airwave Health Monitoring Study is a longitudinal cohort study on the UK police forces, with recruitment and baseline measurements done from 2005 to 2014, and includes medical, biomedical, occupational and lifestyle information. All participants were voluntary and provided with written informed consent. The Airwave Health Monitoring Study is approved by the National Health Service Multi-Site Research Ethics Committee (MREC/13/NW/0588). All methods followed the relevant guidelines and regulations.

Airwave samples were divided into two separate sample sets. The first set consisted of 3000 samples that utilised lithium heparin plasma samples and underwent ultra performance LC-MS profiling analysis for lipids and small metabolites at the NPC, hereafter known as the "NPC-Only dataset". The second set consisted of 2250 ethylenediaminetetraacetic (EDTA) plasma samples and were analysed by Metabolon, Inc. (Morrisville, NC, USA) with 1000 of these samples also analysed by the NPC. Hereafter, the subset of 1000 individuals that included measurements by both platforms are known as the "Metabolon-NPC dataset", whilst the remaining 1250 samples, are known as the "Metabolon-Only dataset". NPC metabolomic features were matched using the "M2S" computational methodology⁸, yielding the same 3585 metabolomic features across the two NPC-based datasets, 914 of which were annotated (306 unique annotations). In both Metabolon datasets, 1148 metabolites were measured, comprising 828 annotated and 320 unknowns. Metabolon metabolites which were invariant across all samples or had more than 25% missing values been removed, leaving 915 metabolites. Metabolon-NPC dataset was additionally quantified using nuclear magnetic resonance (NMR) by NPC within the entire 1000 samples. Data description, sample preparation, metabolite annotation and data processing have been described elsewhere for Airwave NPC¹⁴ and for Airwave Metabolon¹⁵. Outliers for all datasets were removed using local outlier factor¹⁶ with default settings across all measured metabolites resulting in the removal of 67 samples across all datasets.

Study overview

The overall analysis plan is shown in Fig. 7. Using the Metabolon-NPC dataset, five clusters of metabolites were identified which grouped based on biological pathway / chemical class. Next, the Metabolon-NPC dataset samples were split into three subsets; training ($n = 706$), validation ($n = 78$) and held-out test set ($n = 195$). The training and validation sets were used to train an importance-weighted autoencoder for each of the five metabolite clusters and an additional autoencoder with all metabolites and metabolomic features. These attempted to recreate the entire set of Metabolon measurements within the held-out test set samples based on greedy selection across all models, creating an imputed Metabolon dataset. To assess the quality of the imputations, R^2 and R^2 variance across imputation iterations were explored within the test-set samples.

The Metabolon metabolites imputed with a quality above a defined threshold in the held-out test set were selected, and their values were imputed in the NPC-Only dataset. Finally, the quality of the generated NPC-Only imputed Metabolon dataset was evaluated by comparing metabolite observational associations with body mass index (BMI) and C-reactive protein (CRP) to similar associations with the real metabolite measurements by Metabolon, NPC and UK Biobank.

Metabolite clustering

To identify clusters of metabolites within the Metabolon-NPC dataset, Uniform manifold approximation and projection (UMAP)¹⁷ was used for dimensionality reduction followed by a centroid-based clustering, a method that clusters items by similarity using central points for organization. UMAP was employed to map each of the NPC metabolomic features and Metabolon metabolites into the same 2-dimensional space, based on their values across the Metabolon-NPC dataset. UMAP was performed using cosine distance, two components, 35 neighbours and a minimum distance of 0.3. K-means clustering was used as the centroid-based clustering to identify the clusters of metabolites within the UMAP-embedded 2-dimensional space using the elbow method with silhouette score to determine the optimal number of clusters. Clusters were labelled using their most predominant chemical compound class or biological pathway.

Metabolite imputation

The Metabolon-NPC dataset ($N = 979$) was separated into training, validation and held-out test set (70/10/20 split) prior to model training. To normalize the data, min-max scaling technique was applied, adjusting the value within the range of -1 to 1 using population estimates of mean and standard deviation calculated from only the training set. The core of the study involved training an importance-weighted autoencoder (IWAE) to impute the Metabolon metabolites by randomly masking single values and imputing a percentage of Metabolon metabolites during training.

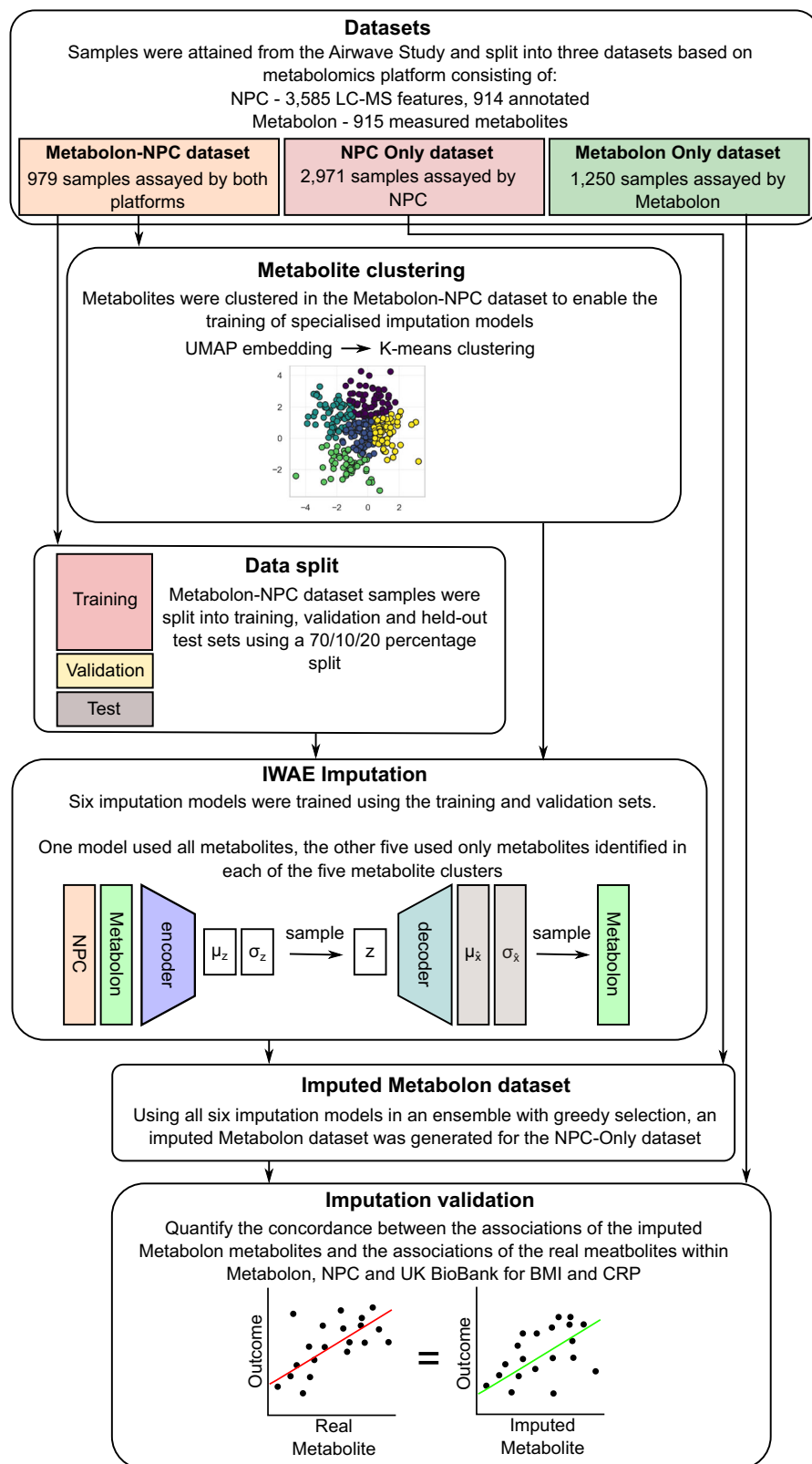
Hyperparameters for each model were selected using coarse-to-fine grid searching including the type of IWAE^{10,18}, the percentage of random masking, presence of mix-up regularisation¹⁹, dropout rate²⁰, presence of batch normalisation²¹ and error function. Model performance was measured using the imputation error calculated after masking the entire set of Metabolon metabolites on the validation set with early stopping. After hyperparameter tuning, the final imputation model was trained using the combined training and validation set. This process was completed six times, once for each of the five metabolite clusters using only the metabolites and metabolomic features present in the cluster and once using all metabolites and metabolomic features.

For the evaluation phase, the model's performance was tested by masking and imputing all Metabolon metabolites within the held-out test set. A selection process (greedy selection) determined the most appropriate imputation model for each metabolite based on the imputation performance within the validation set during model tuning. To assess the reliability and accuracy of the imputation, the test set underwent a total of fifty-one imputation cycles. The first cycle provided a baseline R^2 performance for each metabolite, while the remaining fifty cycles were averaged to explore prediction uncertainty via R^2 variance of the predicted vs observed values. Spearman's rank correlation was used to compare the imputed values in the test set against the corresponding actual values. Finally, leveraging the finalised imputation model, an imputed Metabolon dataset was generated using the NPC-Only dataset ($N = 2971$), employing the same greedy selection rules as applied to the test set.

Observational analysis

To further evaluate the accuracy of the imputations, metabolite associations with BMI and CRP were examined using linear regression adjusted for age and sex. Before fitting the models, both exposures and outcomes were standardised (mean = 0, standard deviation (SD) = 1) before fitting the models. Only Metabolon metabolites which were imputed with an overall $R^2 \geq 0.55$ in the test set were tested

Fig. 7 | Schematic diagram of the study.



for association using the imputed Metabolon (NPC-Only) dataset. This was selected as a reasonable trade-off point between metabolite imputation performance and uncertainty based on the test-set imputation performance (Fig. 2A). The same set of metabolites was tested for association using the real Metabolon metabolites within the combined

Metabolon-NPC and Metabolon-Only datasets ($N = 2229$). Additionally, metabolites tested for association in the imputed Metabolon (NPC-Only) dataset which had a matching LC-MS annotation in the combined Metabolon-NPC and NPC-Only dataset ($N = 3950$), or UK Biobank NMR ($N = 110,068$) were tested for associations (Fig. 5).

Software

NPC metabolomic features were matched across Airwave subsets using the “M2S” MATLAB (R2019b) package available at <https://github.com/rjdossan/M2S>⁸. Imputation was performed using TensorFlow v2.4.1²² and Python v3.7.9. Observational analyses and data processing were performed using sci-kit learn v0.23.2²³.

Data Availability

The Airwave and NPC datasets used and/or analysed during the study are available from the corresponding author on reasonable request. UK Bio-Bank datasets are available upon the submission and acceptance of a suitable research application to UK BioBank.

Code availability

Code used to generate the models and explore performance used in the study are available from the corresponding author on reasonable request.

Received: 7 January 2025; Accepted: 29 December 2025;

Published online: 10 January 2026

References

- Johnson, C. H., Ivanisevic, J. & Siuzdak, G. Metabolomics: beyond biomarkers and towards mechanisms. *Nat. Rev. Mol. Cell Biol.* **17**, 451–459 (2016).
- Zhang, A., Sun, H., Yan, G., Wang, P. & Wang, X. Metabolomics for Biomarker Discovery: Moving to the Clinic. *Biomed. Res Int* **2015**, 354671 (2015).
- Aderemi, A. V., Ayeleso, A. O., Oyedapo, O. O. & Mukwevho, E. Metabolomics: A Scoping Review of Its Role as a Tool for Disease Biomarker Discovery in Selected Non-Communicable Diseases. *Metabolites* **11**, 418 (2021).
- Tzoulaki, I., Ebbels, T. M. D., Valdes, A., Elliott, P. & Ioannidis, J. P. A. Design and analysis of metabolomics studies in epidemiologic research: A primer on-omic technologies. *Am. J. Epidemiol* **180**, 129–139 (2014).
- Julkunen, H. et al. Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nat. Commun.* **14**, 604 (2023).
- Bild, D. E. et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am. J. Epidemiol.* **156**, 871–881 (2002).
- Habra, H. et al. metabCombiner: Paired Untargeted LC-HRMS Metabolomics Feature Matching and Concatenation of Disparately Acquired Data Sets. *Anal. Chem.* **93**, 5028–5036 (2021).
- Climaco Pinto, R. et al. Finding Correspondence between Metabolomic Features in Untargeted Liquid Chromatography-Mass Spectrometry Metabolomics Datasets. *Anal. Chem.* **94**, 5493–5503 (2022).
- Burda, Y., Grosse, R. & Salakhutdinov, R. Importance Weighted Autoencoders. *CoRR* <https://www.semanticscholar.org/paper/Importance-Weighted-Autoencoders-Burda-Grosse/3e47c4c2dd98c49b7771c7228812d5fd9eee56a3> (2015).
- Mattei, P.-A. & Frellsen, J. missWAE: Deep Generative Modelling and Imputation of Incomplete Data. *ArXiv* <https://www.semanticscholar.org/paper/missWAE%3A-Deep-Generative-Modelling-and-Imputation-Mattei-Frellsen/eb799e5161d895ed773717f9f6c197b2547c13dc> (2018).
- Pietzner, M. et al. Comprehensive metabolic profiling of chronic low-grade inflammation among generally healthy individuals. *BMC Med* **15**, 210 (2017).
- Xu, Y. et al. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res* **48**, e85 (2020).
- Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Neural Information Processing Systems* (2017).
- Dehghan, A. et al. Metabolome-wide association study on ABCA7 indicates a role of ceramide metabolism in Alzheimer’s disease. *Proc. Natl. Acad. Sci.* **119**, e2206083119 (2022).

- Kojouri, M. et al. Metabolome-wide association study on physical activity. *Sci. Rep.* **13**, 2374 (2023).
- Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. LOF: identifying density-based local outliers. in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* 93–104 (Association for Computing Machinery, <https://doi.org/10.1145/342009.335388>).
- McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* <https://www.semanticscholar.org/paper/UMAP%3A-Uniform-Manifold-Approximation-and-Projection-McInnes-Healy/3a288c63576fc385910cb5bc44eaa75b442e62e> (2018).
- Ipsen, N. B., Mattei, P.-A. & Frellsen, J. not-MIWAE: Deep Generative Modelling with Missing not at Random Data. *ArXiv* <https://www.semanticscholar.org/paper/not-MIWAE%3A-Deep-Generative-Modelling-with-Missing-Ipsen-Mattei/a6bf259a29f9be2593720d1b62697edc825292c14> (2020).
- Zhang, H., Cissé, M., Dauphin, Y. & Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *ArXiv* <https://www.semanticscholar.org/paper/mixup%3A-Beyond-Empirical-Risk-Minimization-Zhang-Cisse%3A%2F4feef0fd284feb1233399b400eb897f59ec92755> (2017).
- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* <https://www.semanticscholar.org/paper/Dropout%3A-a-simple-way-to-prevent-neural-networks-Srivastava-Hinton/34f25a8704614163c4095b3ee2fc969b60de4698> (2014).
- Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ArXiv* <https://www.semanticscholar.org/paper/Batch-Normalization%3A-Accelerating-Deep-Network-by-Ioffe-Szegedy/995c5f5e62614fcb4d2796ad2faab969da51713e> (2015).
- Abadi, M. et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv* <https://www.semanticscholar.org/paper/TensorFlow%3A-Large-Scale-Machine-Learning-on-Systems-Abadi-Agarwal/9c9d7247f8c51ec5a02b0d911d1d7b9e8160495d> (2016).
- Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* <https://www.semanticscholar.org/paper/Scikit-learn%3A-Machine-Learning-in-Python-Pedregosa-Varoquaux/168f28ac3c8c7ea63bf7ed25f2288e8b67e2fe74> (2011).

Acknowledgements

This work is supported by the UK Dementia Research Institute at Imperial College, which receives its funding from UK Dementia Research Institute Ltd., funded by the UK Medical Research Council (MRC), Alzheimer’s Society, and Alzheimer’s Research UK. A.D. is funded by a Wellcome Trust seed award (206046/Z/17/Z). This work was supported by the Medical Research Council and National Institute for Health Research [grant number MC_PC_12025] and an ITMAT Push for Impact grant from the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre (BRC). Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre.

Author contributions

A.S., A.D., and R.P. were responsible for conceptualising the project. R.P. processed the data and A.S. conducted all the analysis. A.S. and A.D. wrote the main manuscript text. All authors reviewed 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/s41540-025-00644-5>.

Correspondence and requests for materials should be addressed to Alexander Smith.

Reprints and permissions information is available at
<http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026