Autoencoders reveal polyunsaturated fatty acids (PUFA)-Related metabolic signature linked to cancer risk

Marie Breeur,[a,b,*] Joshua Atkins,[b] Laia Peruchet-Noray,[a] Lisa Bonheme,[c] Nicolas Alcala,[c] Laure Dossus,[a] Mazda Jenab,[a] Mattias Johansson,[c] Sabina Rinaldi,[a] Ruth C. Travis,[b] Christian Bork,[d,e] Christina C. Dahm,[f] Anne Tjønneland,[g,h] Anja Olsen,[g,i] Sabine Naudin,[j] Seehyun Park,[j] Therese Truong,[j] Verena Katzke,[k] Charlotte Le Cornet,[k] Matthias B. Schulze,[l] Marcela Prada,[l] Carlotta Sacerdote,[m] Benedetta Bendinelli,[n] Claudia Agnoli,[o] Fabrizio Pasanisi,[p] José María Gálvez-Navas,[q,r,s] Marcela Guevara,[r,t,u] Alicia K. Heath,[v] James Yarmolinsky,[v] Marc J. Gunter,[a,w] Pietro Ferrari,[a] Karl Smith-Byrne,[b] and Vivian Viallon[a,**]

[a]Nutrition and Metabolism Branch, International Agency for Research on Cancer, Lyon, France
[b]Cancer Epidemiology Unit, Oxford Population Health, University of Oxford, Oxford, United Kingdom
[c]Genetic Epidemiology Branch, International Agency for Research on Cancer, Lyon, France
[d]Department of Cardiology, Aalborg University Hospital, Denmark
[e]Department of Clinical Medicine, Aalborg University, Denmark
[f]Department of Public Health, Aarhus University, Denmark
[g]Danish Cancer Institute, Denmark
[h]Department of Public Health, University of Copenhagen, Denmark
[i]Department of Public Health, University of Århus, Denmark
[j]Paris-Saclay University, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France
[k]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[l]German Institute of Human Nutrition, Department of Molecular Epidemiology, Nuthetal, Germany
[m]Unit of Cancer Epidemiology Città della Salute e della Scienza University-Hospital, Turin, Italy
[n]Clinical Epidemiology Unit, Institute for cancer research, prevention and clinical network (ISPRO) Florence, Italy
[o]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[p]Dipartimento di Medicina Clinica e Chirurgia, Federico II University, Naples, Italy
[q]Biosanitary Research Institute ibs.GRANADA, Granada, Spain
[r]Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[s]Andalusian School of Public Health, Granada, Spain
[t]Instituto de Salud Pública y Laboral de Navarra, Pamplona, Spain
[u]Navarra Institute for Health Research (IdiSNA), Pamplona, Spain
[v]Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, United Kingdom
[w]Department of Epidemiology and Biostatistics, School of Public Health, Faculty of Medicine, Imperial College London, London, UK

eBioMedicine
2026;124: 106147
Published Online xxx
https://doi.org/10.1016/j.ebiom.2026.106147

Summary

Background Metabolomics is a valuable tool for characterising biological mechanisms involved in cancer development, but produces complex datasets with intricate interdependencies. While linear dimension reduction techniques such as principal component analysis (PCA), have proven useful to summarise informative hidden patterns, biological evidence suggests metabolic relationships extend beyond linearity. Non-linear dimension reduction techniques, such as autoencoders (AEs), may identify more meaningful components.

Methods We applied AEs and PCA to metabolomic data available for 5828 matched case–control pairs from 8 cancer-specific case–control studies nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, and compared their performance. We evaluated the association between components identified by AEs and PCA with cancer risk, and explored the biological interpretation of components through their association with genetic factors and selected biomarkers.

Findings PCA and AEs showed similar reconstruction performance. PCA's first component (PCA.1) captured phosphatidylcholines (PCs) as the primary source of variability and was associated with cancer risk. Conversely, AEs decomposed PC metabolism into two components, one of which exhibited a stronger association with cancer risk than PCA.1. Unlike PCA.1, this component was strongly associated with genetic variants mapping to the *TMEM258* and *FADS* genes, key in polyunsaturated fatty acids (PUFA) biosynthesis and regulation. Consistently, the AE component demonstrated stronger associations with circulating omega-3 and omega-6 PUFA levels than PCA.1.

*Corresponding author. Nutrition and Metabolism Branch, International Agency for Research on Cancer, Lyon, France.
**Corresponding author.
E-mail addresses: marie.breeur@ndph.ox.ac.uk (M. Breeur), viallonv@iarc.who.int (V. Viallon).

Interpretation Linear methods remain adequate for general dimension reduction. However, AEs better captured specific pathways, identifying a component reflecting perturbations in PUFA metabolism associated with cancer risk.

Funding World Cancer Research Fund (IIG_FULL_2022_013).

Keywords: Metabolomics; Autoencoder; Neural networks; Dimension reduction; Cancer; Fatty acids

Research in context

Evidence before this study

Metabolomic processes are known to be complex, involving numerous non-linear interactions between metabolites. Nevertheless, commonly used frameworks for identifying metabolic patterns typically rely on linear techniques such as principal component analysis (PCA). It remains unclear whether non-linear approaches can provide more detailed or biologically meaningful insights. A previous study in 2021 investigated the use of variational autoencoders (VAEs) for this purpose, but overall, the application of non-linear methods in metabolomics remains relatively limited compared to other omics disciplines.

Added value of this study

This study evaluates the utility of autoencoders for two key tasks: dimensionality reduction (i.e., capturing essential information and enabling data reconstruction) and pathway recovery (i.e., identifying biologically meaningful components). For dimensionality reduction, autoencoders and linear methods performed comparably, suggesting that metabolic variability is well-approximated by linear models. However, autoencoders uncovered components reflecting more nuanced biological processes. Notably, we identified a distinct metabolic component associated with genetic variants involved in polyunsaturated fatty acid (PUFA) metabolism, which may help explain the observed relationship between PUFA metabolism and cancer risk.

Implications of all the available evidence

Our findings suggest that while linear methods such as PCA are adequate for general dimensionality reduction in metabolomic data, incorporating non-linear approaches like autoencoders can reveal more detailed biological insights, particularly in the context of pathway-level analysis. From a biomedical perspective, this work strengthens the evidence linking PUFA metabolism to cancer risk and further suggests that this relationship might be linked to underlying genetic factors.

Introduction

Metabolomics, the comprehensive study of metabolites in biological systems, offers a powerful approach for investigating biochemical processes and disease mechanisms.[1] In epidemiology, metabolomics has been utilised to identify biomarkers[2,3] and elucidate metabolic pathways possibly associated with disease risk, such as cancer,[4–6] diabetes,[7] and cardiovascular conditions.[8] However, metabolomics datasets are often characterised by high dimensionality and complex correlation structures, reflecting the interconnected nature of metabolites across metabolic pathways and biological processes, making it challenging to interpret results from studies that analyse metabolites individually, without accounting for their underlying relationships.

Different analytical approaches were proposed to account for the interconnections across metabolites and better describe the overall metabolic landscape possibly related to the studied outcome. Approaches such as over-representation analysis[9] typically rely on a two-step strategy, first identifying a set of metabolites associated with the outcome, then identifying significantly impacted metabolomic pathways based on that set. Other types of approaches, such as network-based approaches[10,11] integrate information on the underlying relationships while assessing the association between metabolites and the studied outcome to directly identify groups of metabolites associated with the outcome. Alternatively, dimension reduction techniques can be used to derive metabolic components, or latent variables, that summarise the set of metabolites well and capture possibly relevant biological or metabolic processes.[12,13]

Principal component analysis (PCA)[14] and its supervised extensions, partial least squares discriminant analysis or regression (PLS-DA and PLS-R),[15–17] have found widespread application in metabolomics to linearly transform metabolic features into new components.[18–21] However, evidence suggests that metabolic relationships extend beyond linearity, with metabolites exhibiting nonlinear associations with outcomes such as birth weight,[22] as well as established nonlinear pathway fluxes.[23,24] Consequently, nonlinear dimension reduction techniques, such as kernel principal component analysis[25] and variational autoencoders[26,27] might outperform their linear counterpart for the identification of metabolic

components capturing relevant biological processes.[28,29] In a recent analysis of metabolomics data in the TwinsUK study,[29] metabolic components identified by variational autoencoders performed well in terms of biological informativeness and generalisability, capturing metabolite interconnections more comprehensively and showing stronger associations with disease outcomes compared to principal components.

Although promising, these results call for replication given the limited application of autoencoders in metabolomics compared to other -omic fields.[30,31] In this work, we investigated whether autoencoders[32,33] could capture biological processes involved in carcinogenesis better than standard PCA using targeted metabolomic data available in the European Prospective Investigation into Cancer and Nutrition[34] (EPIC) study. To compare the biological relevance of the components identified by both approaches, we evaluated their associations with the risk of eight cancer types, including breast, colorectal, endometrial, gallbladder and biliary tract, kidney, localised and advanced prostate cancers, and hepatocellular carcinoma. Additionally, we explored the genetic determinants of components of interest through a Genome-Wide Association Study (GWAS) and investigated their association with dietary factors and circulating levels of selected blood biomarkers.

Methods

An overview of the methodology employed in this paper can be found in Fig. 1.

Study population

The EPIC cohort

The EPIC cohort is an ongoing multicentric prospective study involving over 500,000 men and women recruited between 1992 and 2000 from 23 centres in 10 European countries,[34] originally designed to study the relationship between diet and cancer risk. Incident cancer cases were identified through a combination of health insurance records, cancer and pathology registries, and active follow-up with participants and their next-of-kin. At recruitment, participants provided information on their diet and lifestyle through self-administered questionnaires, and blood samples were collected from approximately 386,000 individuals using a standardised protocol. Details on blood fraction storage, processing, and handling are provided in the Supplementary Materials. Fasting was not required.

Ethics statement

This study was conducted in accordance with the Declaration of Helsinki. EPIC was approved by the Ethics Committee of the International Agency for Research on Cancer (IARC) (ref IEC 14–02), Lyon, France, as well as the local ethics committees of the study centres. All participants provided written informed consent for data collection and storage as well as individual follow-up. The seven case–control studies nested within EPIC, were approved by the ethics committee at IARC (details in Supplementary Materials).

Statistics—construction of the metabolic components

Metabolomic measurements

Our analyses used metabolomics measurements from 15,948 EPIC participants across seven cancer-specific matched case–control studies nested within EPIC (Table 1), which were described in detail elsewhere.[35] Briefly, in each study, one matched control was randomly selected from the risk set of the index case,[36] and matching factors included study centre, sex (self-reported), age at blood collection, time of day of blood collection, fasting status, and for women, the use of exogenous hormones and menopausal status. Metabolite data were acquired for 117 metabolites using Biocrates AbsoluteIDQ p150 or p180 kits via liquid chromatography-tandem mass spectrometry (LC-MS/MS) for amino acids and biogenic amines, and flow injection analysis-tandem mass spectrometry (FIA-MS/MS) for other metabolites. Samples were either serum or citrate plasma, with all samples in a study using the same matrix, except for the breast cancer study (Table 1). The data was pre-processed following an established procedure[35,37] to ensure comparability across multiple studies (Supplementary Materials, section I.A.a).

Dimensionality reduction techniques

PCA is a common dimensionality reduction method that transforms the original variables into a set of uncorrelated components, ranked by explained variance. However, PCA is limited by its reliance on linear combinations of the initial variables.[12,23,25] In contrast, more flexible methods like Kernel Principal Component Analysis (KPCA) capture non-linear patterns using kernel functions,[25] while autoencoders (AE) rely on neural network-based non-linear mappings to compress high-dimensional data into lower-dimensional representations.[32,33] Additional methodological details are provided in Supplementary Materials (section I.B).

Implementation, notation and reconstruction performance

Full details about the implementation of the AEs, KPCA and PCA can be found in the Supplementary Materials, section I.B.d. Briefly, we explored the impact of neural network depth by implementing two distinct AE architectures, each comprising an input layer, with either one or two intermediate layers in the encoder, a latent layer, a decoder symmetrically mirroring the encoder and an output layer (Supplementary Figure 1). Several versions of KPCA were implemented using, in turn, cosine, polynomial, and radial basis function (RBF) kernels. Models were compared through their reconstruction performance, defined as the ability to

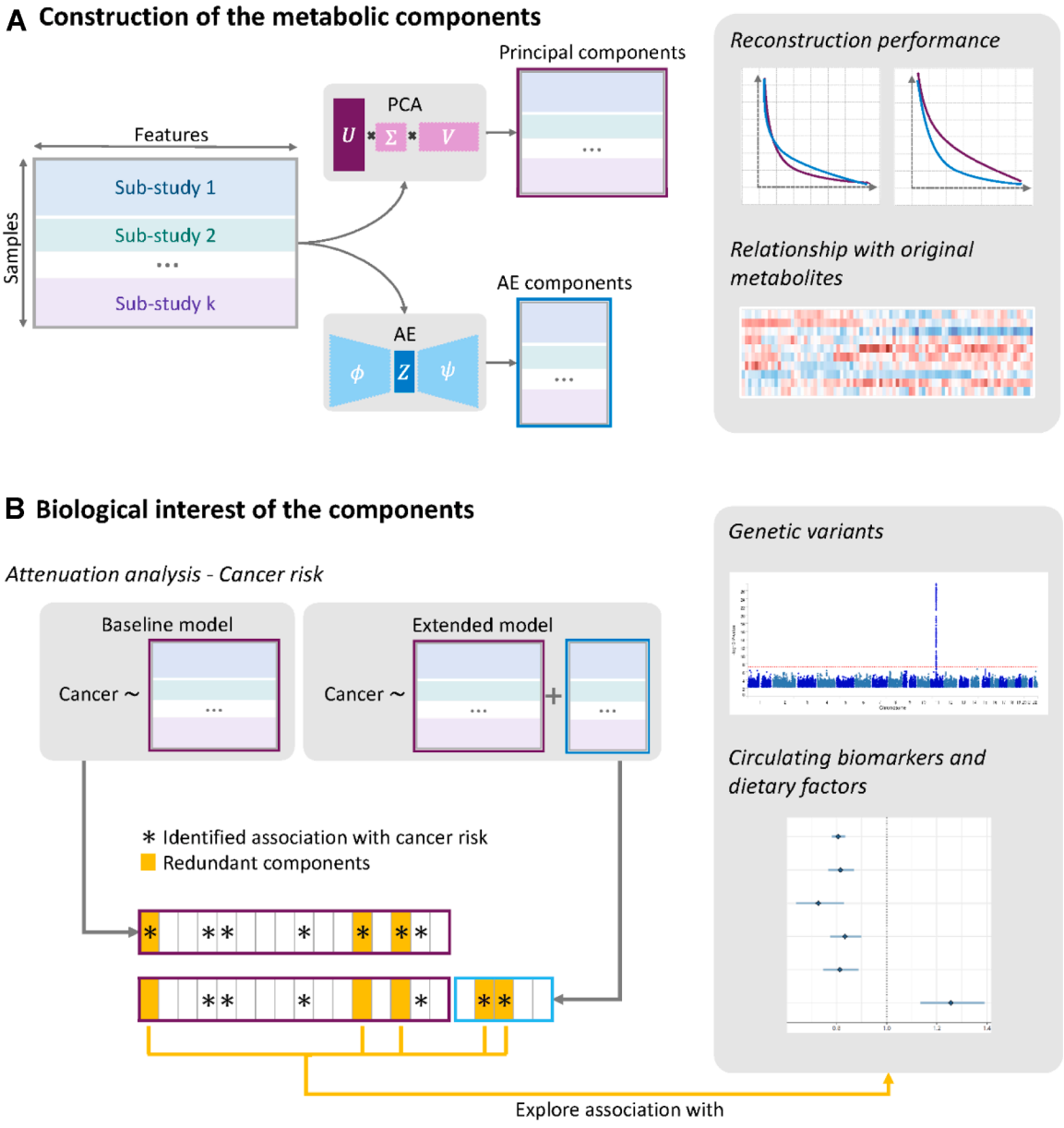

***Fig. 1:* Overview of the methodology.** Panel A: Metabolic components are derived using principal component analysis (PCA) or autoencoders (AE). Their relationships with the original features are assessed, along with information loss during dimensionality reduction. Panel B: Components are analysed for associations with eight different cancer risks in a mutually adjusted framework (left). Redundant components are further examined for associations with genetic variants, selected blood biomarkers, and dietary intakes.

reconstruct the original metabolomic data from the components (Supplementary Materials).

To differentiate the various components derived in this work, we use the notation M.k to refer to the k-th component produced by method M. For instance, the first principal component produced by PCA is denoted by PCA.1.

Relationship with the original metabolites

Because the principal components are defined as linear combinations of the initial variables, their relationship with the original metabolites is fully determined by the weights assigned to each metabolite. In contrast, there is typically no simple mapping from the components identified by non-linear methods back to the original features. Here, we adopted the following strategy to assess the relationship between the original metabolites and the components identified by AEs and KPCA. We evaluated the nonlinearity of the components by computing the coefficient of determination (R^2) from a linear regression of each component on the original metabolites. Components with large R^2 are well

approximated by a linear combination of the original metabolites, which can be used to characterise their relationships with the original metabolites.

For AE components with a low R^2, we used integrated gradients (IG) to characterise their relationship with the original metabolites.[38] This technique determines the contribution of each metabolite to a given component by measuring how much the component changes as the metabolite level is linearly interpolated from 0 to its actual value (Supplementary Materials, section I.B.e). The IG contribution was calculated separately for samples with high and low values of the components of interest, specifically those in the 1st and 9th deciles.

Statistics—biological relevance of the components

To better understand the metabolic and biological significance of the components, we evaluated their associations with a range of factors.

Association with cancer risk

We examined the association between the components identified by the various approaches and cancer risk through multivariate conditional logistic regression models. A data-shared lasso penalty[39,40] was applied to identify components associated with overall and/or type-specific cancer risk. The data shared lasso decomposes each component's odds-ratio into an overall odds-ratio and type-specific deviations around this overall odds-ratio and, under technical assumptions, allows the identification of non-zero overall association with cancer risk as well as non-zero cancer-specific deviations from this overall association (see Supplementary Materials, section I.C.a). These models were additionally adjusted for BMI using the residual method via component specific linear models.

We first considered approach-specific models assessing the associations between the metabolic components and cancer risk for each method, referring to each model as the method-baseline model. For instance, the model examining the associations between cancer risk and the principal components is termed the PCA-baseline model. Additionally, we considered an extended model incorporating all the metabolic components to evaluate potential redundancies across methods and determine whether some components were stronger cancer risk predictors than those identified by other approaches (Fig. 1, lower panel).

To assess the robustness of both the baseline and extended models, we applied all three approaches repeatedly to 50 bootstrap samples generated from the original dataset.[41]

Associations with genetic variants

To explore the potential biological meaning of the metabolic components, we investigated whether they were associated with specific genetic variants.

Cancer site	Number of samples	Matrix	Laboratory	Kit Used
Breast	3172	Citrate plasma[a]	IARC	p180
Colorectal (Study 1)	946	Citrate plasma	IARC	p180
Colorectal (Study 2)	2295	Serum	HZM[b]	p150
Endometrial	1706	Citrate plasma	ICL[c]	p180
Liver	662	Serum	IARC	p180
Kidney	1213	Citrate plasma	IARC	p180
Prostate	6020	Citrate plasma	IARC	p180

[a]Except Swedish participants (n = 101; EDTA plasma). [b]Helmhotz Zentrum München. [c]Imperial College London.

Table 1: **Description of the original seven cancer-specific matched case–control studies nested within EPIC.**

Genotype data was assessed in the EPIC cohort across 14 studies in the form of Single Nucleotide Polymorphisms (SNPs). SNPs represent a common form of genetic variation wherein a single nucleotide in the DNA sequence of a gene differs among individuals or populations. They are the most common type of genetic variation in humans and can be found throughout the genome. A total of 5400 individuals had both SNP and metabolomic data available. DNA samples were genotyped via various platforms a (Supplementary Table 1), and genome-wide data were normalised and imputed to the 1000 Genomes Project Phase 3 v5 in Genome Reference Consortium Human Build 37 using an automated pipeline detailed elsewhere[42] and in the Supplementary Materials.

We examined the connection between the metabolic components and SNPs, using a two-step approach: *(i)* we conducted GWAS within each individual sub-study; and *(ii)* we then meta-analysed the results using an inverse-variance weighted framework. Additional examination of the results was carried out using the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA GWAS) platform.[43] In addition, colocalisation analysis was performed using coloc + SuSiE[44], a Bayesian method that estimates the probability of shared causal variants between traits and allows for the comparison of significant loci across metabolic components. We report posterior probabilities of colocalisation (PP4) across credible sets derived by SuSiE.

Additional details about the models, thresholds and softwares used for implementation can be found in the Supplementary Materials.

Associations with fatty acids and dietary intake

To further evaluate the biological relevance of the metabolic components, we examined their associations with circulating fatty acids and dietary intake.

Levels of 31 fatty acids were measured in eight EPIC studies using gas chromatography, following standardised protocols.[45–53] Due to differences in data preprocessing, a meta-analysis approach was used to assess the association between circulating fatty acid levels and

metabolic components, with linear models adjusted for BMI, centre of recruitment, age, alcohol intake, smoking status, waist circumference, height, and physical activity.[46,47,49] Further details on data acquisition and preprocessing, and analytical procedures are provided in the Supplementary Materials, section I.A.c. Sample sizes are listed in Supplementary Table 2.

Associations between metabolic components and self-reported intake of five pre-defines dietary factors (vegetables; fruits, nuts, and seeds; dairy; meat; and fish) were examined using linear regression models. These dietary variables were derived directly from the EPIC food-frequency questionnaires completed at recruitment. Models were adjusted for BMI, age at blood collection, country, sex, alcohol intake, fasting status, and total energy intake. Analyses were limited to control participants to reduce selection bias.

Role of the funding source

The funders had no role in the study design, data collection, data analysis, interpretation or writing of this report.

Results

Study population

After the preprocessing and exclusions described in the Supplementary Materials, the study population consisted of 11,398 EPIC participants, forming 5699 matched case–control pairs. Cases were diagnosed at an average age of 67.6 years, 11.6 years after blood collection. Supplementary Table 3 presents the main characteristics of the cases and controls for each study. The primary analysis concentrated on 117 metabolites retained after preprocessing, as detailed in Supplementary Table 4.

Construction of the latent components

Supplementary Figure 2 illustrates the reconstruction error with regard to the dimension d of the latent space. Overall, PCA and AEs performed similarly well in terms of reconstruction (Supplementary Figure 2A), while KPCA reconstruction performance depended on the choice of the kernel (Supplementary Figure 2B). The RBF kernel performed poorly, suggesting that this kernel, designed to enforce non-linearity, may not be optimal. Cosine and polynomial kernels yielded better results, but their components were predominantly linear (Supplementary Figure 3) and nearly identical to the PCA components (Supplementary Figure 4). Given these findings, we excluded KPCA from further analysis and focused on PCA and AEs.

Notably, the value of the latent dimension d affected the linearity of the AE components (Supplementary Figure 3), with median R^2 decreasing from 0.96 at $d = 20$ to 0.81 at $d = 5$, indicating increased non-linearity. To assess the possible added value of non-linear components compared to those derived from PCA, we decided to consider components produced by a double-layer AE with a latent dimensionality of 5, referred to as DAE5. For PCA, we opted for $d = 20$, accounting for 85% of the data variance.

Association with cancer risk

The results from our analyses based on the data-shared lasso are presented in Fig. 2. In the PCA baseline model, we identified four overall associations with cancer risk, comprising two inverse (PCA.1 and PCA.17) and two positive associations (PCA.9 and PCA.20). Type-specific associations were also identified with breast cancer (PCA.4, PCA.6), colorectal cancer (PCA.17), endometrial cancer (PCA.18), kidney cancer (PCA.10, PCA.12), HCC (PCA.8, PCA.19), and advanced and localised prostate cancer (PCA.3 and PCA.10, respectively). Furthermore, the AE-baseline model showed an inverse association between DAE5.4 and overall cancer risk, except for localised prostate cancer where the association with DAE5.4 was positive.

In the extended model comprising the first 20 principal components and the five DAE5 components, the associations between the components and overall cancer risk identified in the two baseline models all persisted, save for the negative association between overall cancer risk and PCA.1. Supplementary Table 5 illustrates that the association between PCA.1 and overall cancer risk weakened after the inclusion of component DAE5.4, specifically. This was further supported by our bootstrap analysis. In the extended model, the association between PCA.1 and overall cancer risk was identified in only 20% of bootstrap samples, compared to 56% in the PCA-baseline model. Meanwhile, the association between DAE5.4 and overall cancer risk was identified in 62% of bootstrap samples in the extended model, whereas in the AE-baseline model, it was identified in 88% of bootstrap samples. This suggests that DAE5.4 might better capture information on certain metabolomic mechanisms linked to overall cancer risk than the first principal component PCA.1.

Interpretation of DAE5.4 and comparison with PCA.1

As illustrated in Supplementary Figure 4, PCA.1 was recovered by multiple methods. This is in line with the fact that it captures the main source of variability in the metabolic data. Interestingly, DAE5 was the only method that did not directly encode PCA.1 but rather deconstructed it into two components, DAE5.4 and DAE5.5.

Relationship with the original metabolites

DAE5.4 was relatively well approximated by a linear combination of the original metabolites (with an R^2 of 0.891) and showed strong positive correlations with several phosphatidylcholines (Fig. 3, upper panel). The top 10 features with the highest IG contributions for

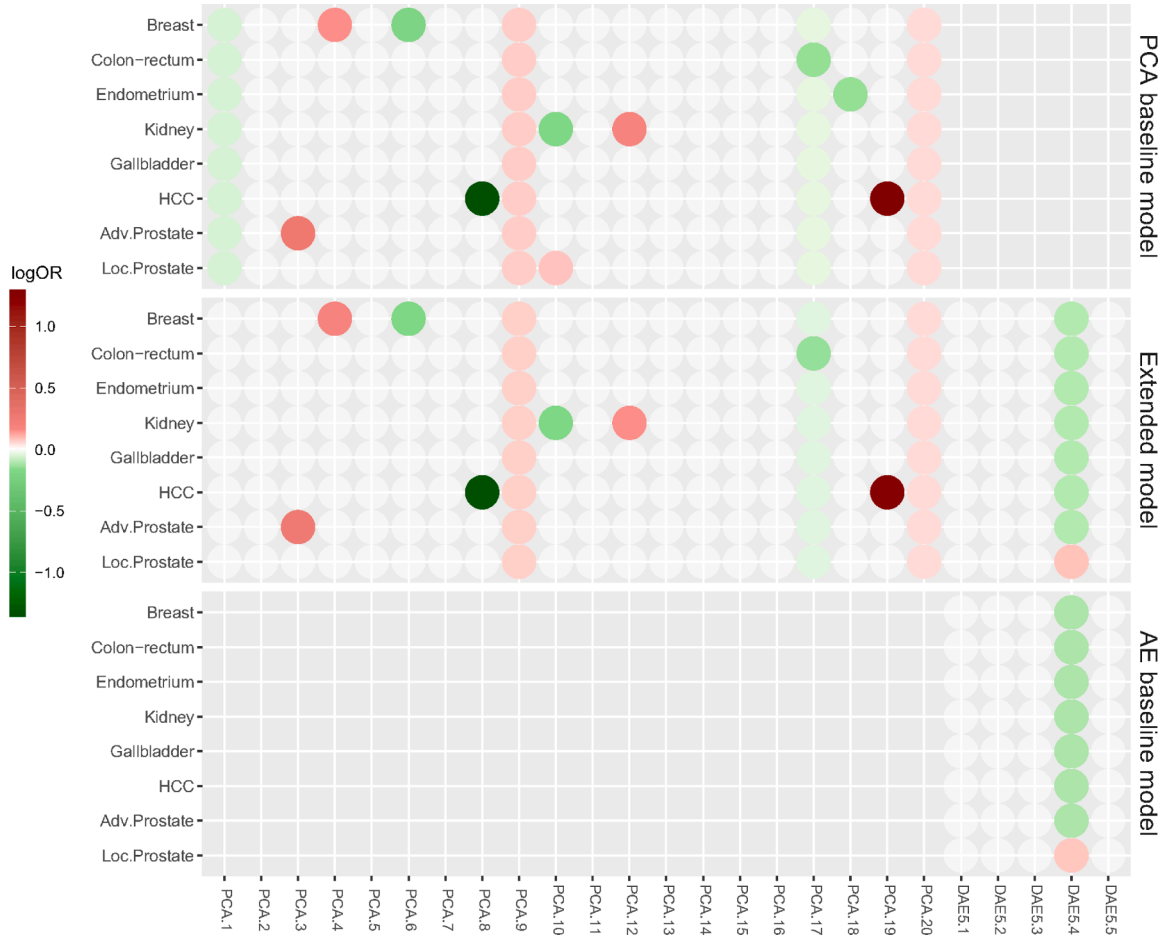

Fig. 2: **Mutually adjusted associations between latent components and the risk of the eight cancer types, using a data shared lasso penalty.** The x-axis represents 25 metabolic components (20 principal components and 5 components derived from an autoencoder [AE]), while the y-axis represents the eight cancer types of interest. PCA and AE baseline models examine associations between cancer risk and components derived from PCA and AE, respectively, whereas the extended model considers all 25 components. White cells indicate no detected associations, while green and red cells represent inverse and positive associations, respectively. Colour intensity corresponds to the absolute value of the log-odds ratio, re-estimated using multivariate, unpenalised conditional regression models (Supplementary Materials). HCC stands for hepatocellular carcinoma, while Adv. Prostate and Loc. Prostate stand for advanced prostate cancer and localised prostate cancer respectively.

both high and low levels of DAE5.4 are shown in Fig. 3, lower panel. The primary contributor, for both high and low DAE5.4 levels, was PC ae C38:6, whose correlation with DAE5.4 was 0.771. Notably, all the top 10 contributors to high levels of DAE5.4 were phosphatidylcholines. In the case of low DAE5.4 levels, phosphatidylcholines remained prominent, accounting for seven of the top 10 features, but acylcarnitines also appeared as contributors. By contrast PCA.1 broadly captured PCs and Sphingomyelins (SMs). As shown in Supplementary Figure 5, the correlation between PCA.1 and PCs and SMs was consistently above 0.4, and generally high, compared to the more selective profile of DAE5.4

Association with genetic variants

Fig. 4 displays the results of the meta-analysed GWAS results for PCA.1 and DAE5.4. No significant heterogeneity across studies was detected for either component.

Both PCA.1 and DAE5.4 exhibited genome-wide significant loci, with a notably shared peak located on 11q12.2–12.3. This region is the most prominent signal for DAE5.4, reaching a minimum p-value of 2.10^{-28}, whereas the corresponding peak in PCA.1 was less significant (minimum p-value 2.10^{-11}). The difference in effect sizes and the presence of an additional locus (15q21.3) associated with PCA.1 suggest that while PCA.1 and DAE5.4 likely capture overlapping

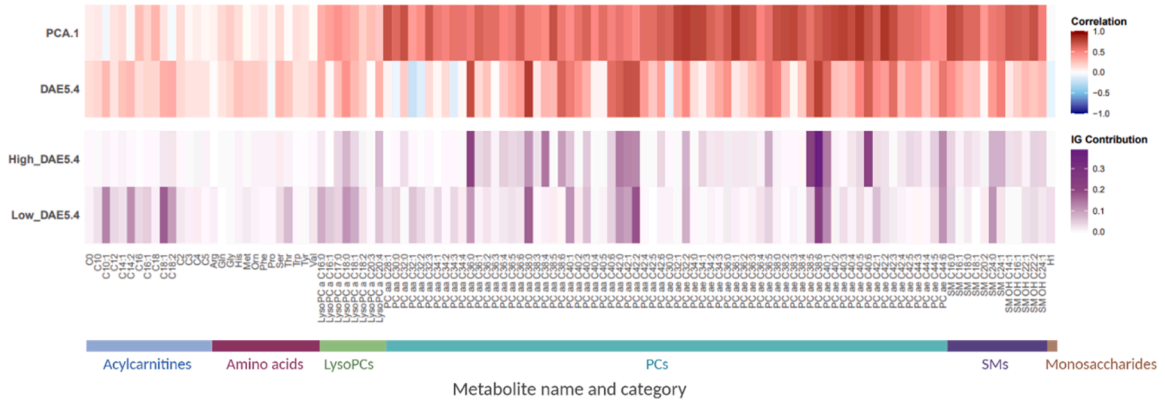

***Fig. 3:* Relationship between PCA.1, DAE5.4, and the original features.** Upper panel: Pearson correlations between the original metabolites and the first principal component (PCA.1) as well as DAE5.4. Lower panel: Contribution of the original metabolites to low and high values of DAE5.4, estimated using Integrated Gradients (IG; see Supplementary Materials). PC and SM stand for phosphatidylcholine and sphingomyelin, respectively.

processes, DAE5.4 might capture a mechanism more strongly associated with baseline genetics.

Closer examination of the 11q12.2–12.3 locus showed that genome-wide significant SNPs for DAE5.4 and PCA.1 mapped to the *FADS* genes cluster, as well as *TMEM258*, *MYRF* and *FEN1* nearby genes. Initial LD-based grouping with FUMA identified nine significant SNPs associated with DAE5.4 (see Table 2A), and three significant SNPs associated with PCA.1 (Table 2B). Fine-mapping with SuSiE identified two 95% credible sets for DAE5.4 (led by rs102274 and rs7394579, mapped to TMEM258 and FADS2 respectively), and for PCA.1 (rs174592 and rs7394579, both mapped to *FADS2*), with strong evidence of colocalisation between the two traits (PP4 of 0.91 and 0.99 for matched credible sets). These results indicate that DAE5.4 and PCA.1 share two fine-mapped causal signals at 11q12.2–12.3, centred on *TMEM258* and *FADS2*.

Both *TMEM258* and *FADS2* have established links to the biosynthesis and regulation of polyunsaturated fatty acids (PUFAs). Consistent with this, DAE5.4 showed strong positive associations with circulating omega-3 PUFAs, including docosahexaenoic acid (DHA) and eicosapentaenoic acid (EPA), as well as inverse associations with certain omega-6 PUFAs, such as dihomo-γ-linolenic acid. Additionally, DAE5.4 was associated with dietary intake of fish and shellfish, further implicating this component in PUFA-related metabolic pathways. In line with the GWAS findings, PCA.1 also showed similar but weaker associations with fish intake and circulating omega-3 and omega-6 PUFAs (see Supplementary Materials, section II.C for a more detailed description of those results, Supplementary Figure 6, Supplementary Tables 6 and 7).

Discussion

We used EPIC metabolomic data to derive and compare metabolic latent components using AEs, standard PCA and KPCA. In our analysis, a consistent observation across various methodologies was the absence of a clear plateau in the reconstruction curves with regard to latent dimensionality d, making the selection of the most appropriate latent dimensionality non-trivial (Supplementary Figure 2). We observed that AEs tended to reconstruct linear latent variables when the latent dimensionality was sufficiently large, and that KPCA consistently recovered the majority of the principal components (Supplementary Materials section II.A, Supplementary Figure 4). Consequently, in the context of typical dimensionality reduction scenarios, our results suggested that linear dimension reduction methods were appropriate if the primary objective was data reconstruction.

The architecture choice for the autoencoder (double-layered, d = 5) was deliberate to explore non-linear representations departing from PCA. Increasing d leads to representations that closely approximate the linear subspace spanned by the first principal components, effectively reproducing PCA.[54] A lower latent dimensionality (here, 5) is thus essential to enforce non-linearity. The architecture choice intentionally emphasises non-linear structure rather than maximising reconstruction accuracy, which would otherwise favour a linear (PCA-like) solution. This setting, while suboptimal from a dimensionality reduction point of view, allowed DAE5 to capture relevant components. For instance, DAE5 was the only method that decomposed the PC metabolism into two components, DAE5.4 and DAE5.5, whereas all other methods recovered a single dominant component involving all of the PCs (PCA.1, as shown in Supplementary Figure 4), while the relationship between PCs and DAE5.4 was notably sparser. This shows that DAE5 was able to disentangle distinct biological signals into two more specific mechanisms, including DAE5.4 which captures the PUFA metabolism. Additionally, DAE5.4 reflected

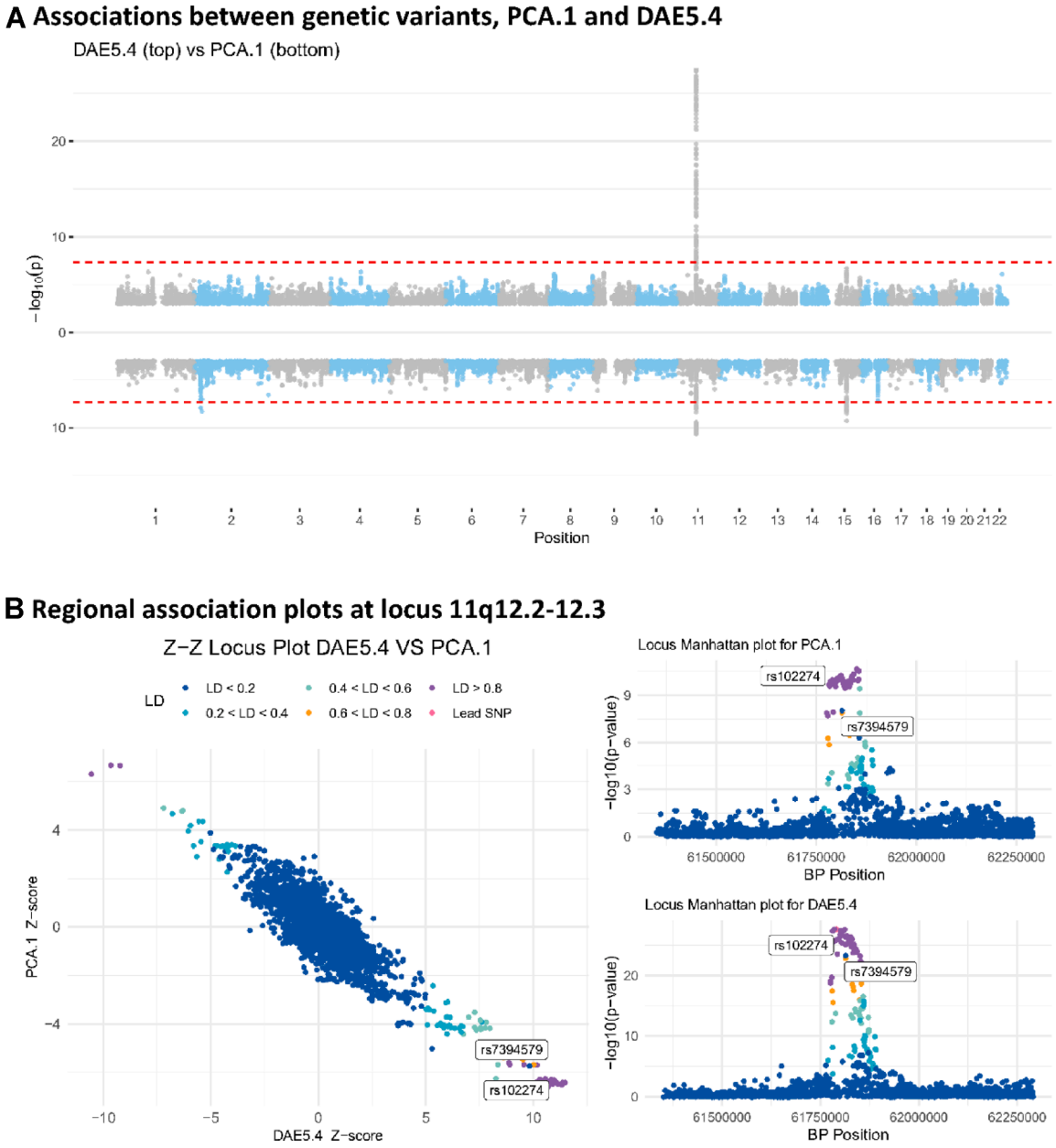

***Fig. 4:* Associations between genetic variants, PCA.1 and DAE5.4.** Panel A: The x-axis displays chromosomal positions of SNPs associated with the latent component of interest. The y-axis displays -log10(P-values) obtained from a meta-analysis of genome-wide summary statistics from linear regression models using genetic variants as exposures and standardised component levels as outcomes run within each contributing study (Methods, Supplementary Materials). The dashed red line indicates the significance threshold of 5.10^{-8}. Panel B: Regional association plots at 11q12.2–12.3, showing the main locus of association. Plots for PCA.1 and DAE5.4 are stacked on the right-hand side, and a z–z plot of DAE5.4 versus PCA.1 is displayed on the left, where the z-score is defined as effect estimate divided by standard error. The two highlighted SNPs correspond to the lead SNPs tagging the SuSiE credible sets identified for DAE5.4. SNPs are coloured according to linkage disequilibrium (LD) with the lead SNP (rs102274), defined as the variant with the lowest P-value in the region. BP stands for base pair.

more complex dependencies, evidenced by its IG contribution scores, which varied across high and low DAE5.4 values, reflecting the nonlinear relationship between DAE5.4 and its main contributors. The link between fatty acids and DAE5.4 also supports specific PCs as main contributors. The link between PCs and fatty acids is well-established, as PCs consist of two fatty acid chains attached to a glycerol backbone.[55] For

A. DAE5.4

rsID	Chromosome	Position (hg19)	P-value	nSNPs	Mapped gene
rs102274	11	61557826	2.36E-28	62	TMEM258
rs174604	11	61626270	3.01E-17	28	FADS2
rs174593	11	61618831	1.39E-16	18	FADS2
rs61896141	11	61556039	1.76E-14	7	TMEM258/MYRF
rs174621	11	61630104	7.75E-11	19	FADS2
rs174620	11	61629747	3.58E-10	19	FADS2
rs2727270	11	61603237	4.21E-09	11	FADS2
rs509360	11	61548559	6.96E-09	1	MYRF
rs174460	11	61657110	1.39E-08	21	FADS3

B. PCA.1

rsID	Chromosome	Position (hg19)	P-value	nSNPs	Mapped gene
rs174592	11	61618608	2.21E-11	60	FADS2
rs34212714	11	61625723	1.44E-10	21	FADS2
rs7394579	11	61581450	9.52E-09	55	FADS2

Independent SNPs are defined as genome-wide significant variants (p-value < 5e-8) that are not in high linkage disequilibrium (LD) with one another ($r^2 < 0.6$). Genomic positions are based on the hg19 reference genome. nSNPs refers to the number of genome-wide significant SNPs in high LD ($r^2 \geq 0.6$) with the SNP of interest. Gene annotations are assigned using ANNOVAR. Multiple mappings indicate intergenic SNPs annotated to the nearest gene.

Table 2: **Independent significant SNPs associated with DAE5.4 and with PCA.1 at the 11q12.2–12.3 locus.**

instance, the primary DAE5.4 contributor, PC ae C38:6 (as per the Biocrates naming convention) can refer to molecular structures like PC(O-16:0/22:6), PC(22:6/15:0), or PC(P-18:0/20:5) that include fatty acid chains formed from DHA (22:6) and EPA (20:5), both associated with DAE5.4. More generally, DAE5.4 is more strongly associated with PUFAs specifically that its PCA.1 counterpart, with more than half of the significant associations observed being between DAE5.4 and a PUFA (Supplementary Table 6). This supports DAE5.4 capturing a more specific mechanism than bulk fatty acid metabolism. The genetics also supports this, with DAE5.4 showing notably stronger associations with the *FADS/TMEM258* genes than PCA.1. *FADS1* and *FADS2* genes encode enzymes crucial for converting precursor fatty acids into long-chain PUFAs,[56] while *TMEM258* is a transmembrane protein involved in the regulation of endoplasmic reticulum (ER) stress and protein glycosylation, implicated in PUFA metabolism through its role in coordinating lipid biosynthesis and inflammatory signalling pathways. Those three genes have previously been shown to be associated with PC levels in previous GWAS of various ancestries,[57–60] with the stronger association with DAE5.4 suggesting that DAE5.4 may better reflect genetically influenced variation in PCs and PUFAs metabolism.

The association between DAE5.4 and cancer risk was also consistent with previous literature, as phosphatidylcholines have been linked with cancer risk by numerous studies.[5,6,61–64] Omega-3 PUFAs, notably found in fish, have been shown to have anti-inflammatory effects, while high intake of omega-6 PUFAs may promote inflammation and tumour growth.[65,66] These findings align with the direction of associations observed in our study, which examined the relationships between blood levels of FAs, dietary factors, and cancer risk through DAE5.4 and, to lesser extent, PCA.1. Supporting the specificity of DAE5.4, a prior pan-cancer study[35] using the same dataset identified a possible association between overall cancer risk and a specific cluster of phosphatidylcholines, notably including the main contributor to DAE5.4, PC ae C38:6. A similar type-specific deviation was already observed in localised prostate cancer, mirroring the DAE5.4 pattern. Differences have been observed previously across metabolomic profiles associated with advanced or localised prostate cancer,[6,67] possibly attributable to reverse causation.[68] The association between cancer and DAE5.4 specifically, rather than PCA.1, suggests that genetic factors implicated in PUFA metabolism may contribute more strongly to cancer risk. Consistent with this, a study in a Chinese population reported associations between serum fatty acid levels and seven SNPs (six of which were present in our analysis and significantly associated with DAE5.4), suggesting that these effects may be mediated by both gene transcription and DNA methylation.[69] These findings support the hypothesis that genetic variation may modulate the relationship between dietary PUFA intake and cancer susceptibility. Further support comes from a Mendelian randomisation study that investigated the causal relevance of PUFAs for risk of cancer,[70] and revealed significant positive associations between genetically proxied PUFA desaturase activity and colorectal, skin, and lung cancers. Positive, though non-significant, associations were also observed for liver, kidney, breast,

and early-onset prostate cancers, as well as overall cancer risk. Negative associations were reported for endometrial cancer and advanced prostate cancer, still not reaching statistical significance. Recent studies have also linked increased expression of *FADS2* and *TMEM258* with the regulation of CD4+ T-cell expression in colorectal cancer.[71] Complementary evidence from CRISPR-based loss-of-function screening suggests that *TMEM258* may be essential for the survival of certain cancer cell lines, although the majority of lines showed only borderline dependency.[72] *FADS1-FADS2* polymorphisms have been linked to the regulation of fatty acid metabolism,[56] with experimental studies showing that perturbation of *FADS*-mediated desaturation can induce ER stress and UPR activation,[73] while induced desaturation via *FADS2* may help cells mitigate ER stress.[74] *TMEM258*, which lies in the same regulatory region, participates in the oligosaccharyltransferase complex and can influence ER homoeostasis,[75] which supports the link between the DAE5.4 genetic signal and cancer-related cellular stress response.

While decomposing the PC metabolism via DAE5 provided deeper insights into PUFAs and cancer risk, it is worth noting that PCA identified several cancer-associated components missed by the DAE5 model. At the model level, PCA with a similar level of compression as DAE5 (i.e restricted to the first five principal components) identified two components associated with advanced prostate cancer and breast cancer, that were not captured by DAE5. Conversely, DAE5 captured the association with localised prostate cancer that PCA missed. This highlights that neither method is universally superior, with each approach offering distinct advantages depending on the biological context. While DAE5 effectively captured nuanced signals within PC metabolism, its performance may not extend equally to other pathways or metabolite classes where linear approaches such as PCA perform better.

Our analysis has several strengths and limitations. It leverages a large population of over 11,000 individuals to derive metabolomic components, providing a robust basis for analysis. Additionally, by integrating multiple data sources, including genetic variants, blood fatty acid levels, and dietary questionnaires, the study offers a comprehensive interpretation of the components, yielding valuable insights into cancer biology. However, a limitation stems from the non-uniqueness of the AE components,[76] as highlighted for instance in the linear case by Baldi et al.[54] Additionally, the representations learnt by AEs are inherently architecture dependent, meaning that different network designs may yield distinct latent structures even when trained on the same data, which complicates generalisation. This could be addressed by investigating supervised approaches, akin to the methodology proposed by Tan et al.,[77] offering greater control over the learning process and potentially yielding more robust representations, and by exploring more complex architectures such as variational autoencoders (VAE)[26] or β-VAE.[78] Finally, because the study is observational, reverse causation cannot be ruled out, and observed metabolic changes may reflect early disease processes.

In conclusion, our analysis highlights the strength of using non-linear dimension reduction techniques such as AEs in cancer metabolomics. Despite some limitations related to instability and non-uniqueness, AEs identified a metabolic component that reflected metabolic perturbations linked to both PUFA metabolism and cancer risk more finely than principal components.

Contributors

Conceptualisation: M.B., V.V.; formal analysis: M.B.; funding acquisition: V.V.; investigation: M.B., V.V.; methodology: M.B., V.V., K.S.-B.; resources: L.D., M.Je., M.Jo., S.R., R.C.T., M.J.G., P.F., C.B, C.C.D, A.T, A.O. S.N, S.P, T.T, V.K, C.L.C, M.B.S, M.P., C.S., B.B., C.A., F.P., J.M.G.-N., M.G., A.K.H, J.Y; supervision: V.V.; validation: M.B., V.V.; visualisation: M.B., V.V.; writing original draft: M.B., V.V.; writing, review and editing: all; accessed and verified the underlying data: M.B., J.A., V.V. All authors read and approved the final version of the manuscript.

Data sharing statement

The EPIC dataset is not publicly available; however, access requests can be submitted to the Steering Committee. For information on how to request access to EPIC data and/or biospecimens, please refer to https://epic.iarc.fr/access/.

The code used for metabolomic data preprocessing is available at: https://code.iarc.fr/viallonv/pipeline_biocrates. The implementation of conditional logistic regression with the data-shared lasso penalty can be found at: https://github.com/NadimBLT/SL1CLR. Code for the implementation and evaluation of dimensionality reduction methods (autoencoders, PCA, and KPCA), as well as for generating the figures and results presented in this paper, is available at: https://github.com/BreeurM/MetAE.

Declaration of interests

C.B is chairperson of the Danish Knowledge Network on familial hypercholesterolaemia, member of the steering committee of the Danish familial hypercholesterolaemia registry, chairperson of the Danish lipid guidelines, and declares having received funding from the Karen Elise Jensens Foundation The remaining authors declare no competing interests.

Acknowledgements

We acknowledge the use of data and biological samples from the EPIC-Ragusa, EPIC-Utrecht, EPIC-Bilthoven, EPIC-Barcelona, EPIC-Asturias, EPIC-Gipuzkoa, EPIC-Murcia, and EPIC-Cambridge cohorts.

This work was financially supported by the World Cancer Research Fund (UK) through the World Cancer Research Fund International grant program (grant number: IIG_FULL_2022_013).

The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC) and by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are supported by: Danish Cancer Society (Denmark); Ligue Nationale Contre le Cancer, Institut Gustave-Roussy, Mutuelle Générale de l'Education Nationale (MGEN), Institut National de la Santé et de la Recherche Médicale (INSERM), French National Research Agency (ANR, reference ANR-10-COHO-0006), French Ministry for Higher Education (subsidy 2102918823, 2103236497, and

2103586016) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Italian Ministry of Health, Italian Ministry of University and Research (MUR), Compagnia di San Paolo (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), the Netherlands Organisation for Health Research and Development (ZonMW), World Cancer Research Fund (WCRF), (The Netherlands); UiT The Arctic University of Norway; Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (C864/A14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (MR/N003284/1, MC-UU_12015/1 and MC_UU_00006/1 to EPIC-Norfolk (DOI 10.22025/2019.10.105.00004); MR/Y013662/1 to EPIC-Oxford) (United Kingdom). Previous support has come from "Europe against Cancer" Programme of the European Commission (DG SANCO). Funding for cancer specific studies is detailed in the Supplementary Materials.

During the preparation of this work, the authors used DeepSeek in order to improve the writing quality. The authors have reviewed and confirmed the validity of the text and take full responsibility for the content of the publication.

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2026.106147.

References

1 Wishart DS. Metabolomics for investigating physiological and pathophysiological processes. *Physiol Rev*. 2019;99(4):1819–1875.
2 Loftfield E, Stepien M, Viallon V, et al. Novel biomarkers of habitual alcohol intake and associations with risk of pancreatic and liver cancers and liver disease mortality. *J Natl Cancer Inst*. 2021;113(11):1542–1550.
3 Rothwell JA, Keski-Rahkonen P, Robinot N, et al. A metabolomic study of biomarkers of habitual coffee intake in four European countries. *Mol Nutr Food Res*. 2019;63(22):1900659.
4 Cross AJ, Moore SC, Boca S, et al. A prospective study of serum metabolites and colorectal cancer risk. *Cancer*. 2014;120(19):3049–3057.
5 His M, Viallon V, Dossus L, et al. Prospective analysis of circulating metabolites and breast cancer in EPIC. *BMC Med*. 2019;17(1):1–13.
6 Schmidt JA, Fensom GK, Rinaldi S, et al. Pre-diagnostic metabolite concentrations and prostate cancer risk in 1077 cases and 1077 matched controls in the European Prospective Investigation into Cancer and Nutrition. *BMC Med*. 2017;15(1):1–14.
7 Sun Y, Gao HY, Fan ZY, He Y, Yan YX. Metabolomics signatures in type 2 diabetes: a systematic review and integrative analysis. *J Clin Endocrinol Metab*. 2020;105(4):1000–1008.
8 McGarrah RW, Crown SB, Zhang GF, Shah SH, Newgard CB. Cardiovascular metabolomics. *Circ Res*. 2018;122(9):1238–1258.
9 Wieder C, Frainay C, Poupin N, et al. Pathway analysis in metabolomics: recommendations for the use of over-representation analysis. *PLoS Comput Biol*. 2021;17(9):e1009105.
10 Amara A, Frainay C, Jourdan F, et al. Networks and graphs discovery in metabolomics data analysis and interpretation. *Front Mol Biosci*. 2022;9:841373.
11 Do KT, Rasp DJP, Kastenmüller G, Suhre K, Krumsiek J. MoDentify: phenotype-driven module identification in metabolomics networks at different resolutions. *Bioinformatics*. 2019;35(3):532–534.
12 Van Der Maaten L, Postma E, Van den Herik J. Others. Dimensionality reduction: a comparative. *J Mach Learn Res*. 2009;10(66–71).
13 Hastie T, Tibshirani R, Friedman JH, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. vol. 2. Springer; 2009.
14 Hotelling H. Relations between two sets of variates. In: *Breakthroughs in statistics: methodology and distribution*. Springer; 1992:162–190.
15 Wold H. *Estimation of principal components and related models by iterative least squares. Multivariate Analysis*. New York: Academic Press; 1966:391–420.
16 Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: a Monte Carlo study. *J Chemom*. 1987;1(3):185–196.
17 Barker M, Rayens W. Partial least squares for discrimination. *J Chemom J Chemom Soc*. 2003;17(3):166–173.
18 Nyamundanda G, Brennan L, Gormley IC. Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics*. 2010;11:1–11.
19 Yamamoto H, Yamaji H, Abe Y, et al. Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. *Chemom Intell Lab Syst*. 2009;98(2):136–142.
20 Szymańska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*. 2012;8:3–16.
21 Bujak R, Daghir-Wojtkowiak E, Kaliszan R, Markuszewski MJ. PLS-based and regularization-based methods for the selection of relevant variables in non-targeted metabolomics data. *Front Mol Biosci*. 2016;3:35.
22 Colicino E, Ferrari F, Cowell W, et al. Non-linear and non-additive associations between the pregnancy metabolome and birthweight. *Environ Int*. 2021;156:106750.
23 Lo-Thong-Viramoutou O, Charton P, Cadet XF, et al. Non-linearity of metabolic pathways critically influences the choice of machine learning model. *Front Artif Intell*. 2022;5:744755.
24 Schwahn K, Beleggia R, Omranian N, Nikoloski Z. Stoichiometric correlation analysis: principles of metabolic functionality from metabolomics data. *Front Plant Sci*. 2017;8:312638.
25 Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput*. 1998;10(5):1299–1319.
26 Kingma DP, Welling M. Auto-encoding variational bayes. *arXiv preprint*. 2022. arXiv:1312.6114.
27 Kingma DP, Welling M. An introduction to variational autoencoders. *Found Trends® Mach Learn*. 2019;12(4):307–392.
28 Shiokawa Y, Date Y, Kikuchi J. Application of kernel principal component analysis and computational machine learning to exploration of metabolites strongly associated with diet. *Sci Rep*. 2018;8(1):3426.
29 Gomari DP, Schweickart A, Cerchietti L, et al. Variational autoencoders learn transferrable representations of metabolomics data. *Commun Biol*. 2022;5(1):645.
30 Pomyen Y, Wanichthanarak K, Poungsombat P, Fahrmann J, Grapov D, Khoomrung S. Deep metabolome: applications of deep learning in metabolomics. *Comput Struct Biotechnol J*. 2020;18:2818–2825.
31 Sen P, Lamichhane S, Mathema VB, et al. Deep learning meets metabolomics: a methodological perspective. *Brief Bioinform*. 2021;22(2):1531–1542.
32 Kramer MA. Nonlinear principal component analysis using autoassociative neural networks. *AIChE J*. 1991;37(2):233–243.
33 Kramer MA. Autoassociative neural networks. *Comput Chem Eng*. 1992;16(4):313–328.
34 Riboli E, Hunt K, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;5(6b):1113–1124.
35 Breeur M, Ferrari P, Dossus L, et al. Pan-cancer analysis of pre-diagnostic blood metabolite concentrations in the European Prospective Investigation into Cancer and Nutrition. *BMC Med*. 2022;20(1):351.
36 Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect*. 1994;102(suppl 8):47–51.
37 Viallon V, His M, Rinaldi S, et al. A new pipeline for the normalization and pooling of metabolomics data. *Metabolites*. 2021;11(9):631.
38 Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *International conference on machine learning*. PMLR; 2017:3319–3328.

39 Gross S, Tibshirani R. Data shared lasso: a novel tool to discover uplift. *Comput Stat Data Anal*. 2016;101:226–235.
40 Ollier E, Viallon V. Regression modelling on stratified data with the lasso. *Biometrika*. 2017;104(1):83–96.
41 Bach FR. Bolasso: model consistent lasso estimation through the bootstrap. In: *Proceedings of the 25th international conference on Machine learning*. 2008:33–40.
42 Peruchet-Noray L, Dimou N, Cordova R, et al. Nature or nurture: genetic and environmental predictors of adiposity gain in adults. *EBioMedicine*. 2025;111.
43 Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8(1):1826.
44 Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet*. 2021;17(9): e1009440.
45 Chajès V, Thiébaut A, Rotival M, et al. Association between serum trans-monounsaturated fatty acids and breast cancer risk in the E3N-EPIC Study. *Am J Epidemiol*. 2008;167(11):1312–1320.
46 Chajès V, Assi N, Biessy C, et al. A prospective evaluation of plasma phospholipid fatty acids and breast cancer risk in the EPIC study. *Ann Oncol*. 2017;28(11):2836–2842.
47 Aglago EK, Murphy N, Huybrechts I, et al. Dietary intake and plasma phospholipid concentrations of saturated, monounsaturated and trans fatty acids and colorectal cancer risk in the European Prospective Investigation into Cancer and Nutrition cohort. *Int J Cancer*. 2021;149(4):865–882.
48 Crowe FL, Allen NE, Appleby PN, et al. Fatty acid composition of plasma phospholipids and risk of prostate cancer in a case-control analysis nested within the European Prospective Investigation into Cancer and Nutrition. *Am J Clin Nutr*. 2008;88(5):1353–1363.
49 Dahm CC, Gorst-Rasmussen A, Crowe FL, et al. Fatty acid patterns and risk of prostate cancer in a case-control study nested within the European Prospective Investigation into Cancer and Nutrition. *Am J Clin Nutr*. 2012;96(6):1354–1361.
50 Khaw KT, Friesen MD, Riboli E, Luben R, Wareham N. Plasma phospholipid fatty acid concentration and incident coronary heart disease in men and women: the EPIC-Norfolk prospective study. *PLoS Med*. 2012;9(7):e1001255.
51 Shi F, Chowdhury R, Sofianopoulou E, et al. Association of circulating fatty acids with cardiovascular disease risk: analysis of individual-level data in three large prospective cohorts and updated meta-analysis. *Eur J Prev Cardiol*. 2025;32:zwae315.
52 Saadatian-Elahi M, Norat T, Bueno-de-Mesquita H, et al. Plasma concentrations of fatty acids in nine European countries: cross-sectional study within the European Prospective Investigation into Cancer and Nutrition (EPIC). 2002;156:215–218.
53 Forouhi NG, Koulman A, Sharp SJ, et al. Differences in the prospective association between individual plasma phospholipid saturated fatty acids and incident type 2 diabetes: the EPIC-InterAct case-cohort study. *Lancet Diabetes Endocrinol*. 2014;2(10):810–818.
54 Baldi P, Hornik K. Neural networks and principal component analysis: learning from examples without local minima. *Neural Netw*. 1989;2(1):53–58.
55 Vance DE, Vance JE. *Biochemistry of Lipids, Lipoproteins and Membranes*. Amsterdam, the Netherlands: Elsevier; 1996.
56 Santana JDM, Pereira M, Carvalho GQ, et al. FADS1 and FADS2 gene polymorphisms modulate the relationship of omega-3 and omega-6 fatty acid plasma concentrations in gestational weight gain: a NISAMI cohort study. *Nutrients*. 2022;14(5):1056.
57 Long T, Hicks M, Yu HC, et al. Whole-genome sequencing identifies common-to-rare variants associated with human blood metabolites. *Nat Genet*. 2017;49(4):568–578.
58 Lotta LA, Pietzner M, Stewart ID, et al. A cross-platform approach identifies genetic regulators of human metabolism and health. *Nat Genet*. 2021;53(1):54–64.
59 Ottensmann L, Tabassum R, Ruotsalainen SE, et al. Genome-wide association analysis of plasma lipidome identifies 495 genetic associations. *Nat Commun*. 2023;14(1):6934.
60 Yang C, Veenstra J, Bartz TM, et al. Genome-wide association studies and fine-mapping identify genomic loci for n-3 and n-6 polyunsaturated fatty acids in Hispanic American and African American cohorts. *Commun Biol*. 2023;6(1):852.
61 Guida F, Tan VY, Corbin LJ, et al. The blood metabolome of incident kidney cancer: a case–control study nested within the MetKid consortium. *PLoS Med*. 2021;18(9):e1003786.
62 Stepien M, Keski-Rahkonen P, Kiss A, et al. Metabolic perturbations prior to hepatocellular carcinoma diagnosis: findings from a prospective observational cohort study. *Int J Cancer*. 2021;148(3):609–625.
63 Shu X, Xiang YB, Rothman N, et al. Prospective study of blood metabolites associated with colorectal cancer risk. *Int J Cancer*. 2018;143(3):527–534.
64 Shu X, Zheng W, Yu D, et al. Prospective metabolomics study identifies potential novel blood metabolites associated with pancreatic cancer risk. *Int J Cancer*. 2018;143(9):2161–2167.
65 Abel S, Riedel S, Gelderblom WCA. Dietary PUFA and cancer. *Proc Nutr Soc*. 2014;73(3):361–367.
66 D'Angelo S, Motti ML, Meccariello R. ω-3 and ω-6 polyunsaturated fatty acids, obesity and cancer. *Nutrients*. 2020;12(9). Available from: https://www.mdpi.com/2072-6643/12/9/2751.
67 Schmidt JA, Fensom GK, Rinaldi S, et al. Patterns in metabolite profile are associated with risk of more aggressive prostate cancer: a prospective study of 3,057 matched case–control sets from EPIC. *Int J Cancer*. 2020;146(3):720–730.
68 Grenville ZS, Noor U, Rinaldi S, et al. Perturbations in the blood metabolome up to a decade before prostate cancer diagnosis in 4387 matched case–control sets from the European Prospective Investigation into Cancer and Nutrition. *Int J Cancer*. 2025;156(5):943–952.
69 He Z, Zhang R, Jiang F, et al. FADS1-FADS2 genetic polymorphisms are associated with fatty acid metabolism through changes in DNA methylation and gene expression. *Clin Epigenetics*. 2018;10:1–13.
70 Haycock P, Borges M, Burrows K, et al. The association between genetically elevated polyunsaturated fatty acids and risk of cancer. *EBioMedicine*. 2023;91:104510.
71 Deslandes B, Wu X, Lee MA, et al. Transcriptome-wide Mendelian randomization exploring dynamic CD4+ T cell gene expression in colorectal cancer development. *J Leukoc Biol*. 2025;117(10):qiaf131. https://doi.org/10.1093/jleuko/qiaf131.
72 DepMap, Broad. DepMap public 25Q2. Available from: depmap.org; 2025.
73 Heravi G, Liu Z, Herroon M, et al. Targeting polyunsaturated fatty acids desaturase FADS1 inhibits renal cancer growth via ATF3-mediated ER stress response. *Biomed Pharmacother*. 2025;186: 118006.
74 Ikeda T, Katoh Y, Hino H, et al. FADS2 confers SCD1 inhibition resistance to cancer cells by modulating the ER stress response. *Sci Rep*. 2024;14(1):13116.
75 Graham DB, Lefkovith A, Deelen P, et al. TMEM258 is a component of the oligosaccharyltransferase complex controlling ER stress and intestinal inflammation. *Cell Rep*. 2016;17(11):2955–2965.
76 Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(8):1798–1828.
77 Tan K, Huang W, Hu J, Dong S. A multi-omics supervised autoencoder for pan-cancer clinical outcome endpoints prediction. *BMC Med Inform Decis Mak*. 2020;20:1–9.
78 Higgins I, Matthey L, Pal A, et al. beta-vae: learning basic visual concepts with a constrained variational framework. *ICLR Poster*. 2017;3.