PNAS RESEARCH ARTICLE | BIOCHEMISTRY

Modeling blood metabolite homeostatic levels reduces sample heterogeneity across cohorts

Danni Liu[a], G. A. Nagana Gowda[b], Zhongli Jiang[a], Kangni Alemdjrodo[a], Min Zhang[a,c,1], Dabao Zhang[a,c,1], and Daniel Raftery[b,1]

Edited by Rafael Brüschweiler, The Ohio State University, Columbus, OH; received May 6, 2023; accepted December 5, 2023 by Editorial Board Member Angela M. Gronenborn

Blood metabolite levels are affected by numerous factors, including preanalytical factors such as collection methods and geographical sites. These perturbations have caused deleterious consequences for many metabolomics studies and represent a major challenge in the metabolomics field. It is important to understand these factors and develop models to reduce their perturbations. However, to date, the lack of suitable mathematical models for blood metabolite levels under homeostasis has hindered progress. In this study, we develop quantitative models of blood metabolite levels in healthy adults based on multisite sample cohorts that mimic the current challenge. Five cohorts of samples obtained across four geographically distinct sites were investigated, focusing on approximately 50 metabolites that were quantified using ^{1}H NMR spectroscopy. More than one-third of the variation in these metabolite profiles is due to cross-cohort variation. A dramatic reduction in the variation of metabolite levels (90%), especially their site-to-site variation (95%), was achieved by modeling each metabolite using demographic and clinical factors and especially other metabolites, as observed in the top principal components. The results also reveal that several metabolites contribute disproportionately to such variation, which could be explained by their association with biological pathways including biosynthesis and degradation. The study demonstrates an intriguing network effect of metabolites that can be utilized to better define homeostatic metabolite levels, which may have implications for improved health monitoring. As an example of the potential utility of the approach, we show that modeling gender-related metabolic differences retains the interesting variance while reducing unwanted (site-related) variance.

homeostasis | metabolomics | metabolic modeling | NMR | variance reduction

Significance

Metabolite concentrations in the blood are perturbed by many factors that include clinical and demographic variables, preanalytical factors, and even genetics. Differences observed in blood metabolite levels across cohorts, accounting for more than one-third of the total variation, are especially important as they impede biomarker validation efforts. This multicohort study shows that metabolite levels can be successfully modeled to reduce cross-cohort and even within-cohort variations, using as few as two cohorts for training. The modeling also reveals a network of metabolite relationships that can be used to gain further insights into homeostasis and metabolic processes. The approach has the potential to reduce unwanted variance while retaining metabolic signals of interest.

Blood is the most widely used biospecimen in the clinic and the metabolomics field. Blood metabolite profiles obtained using both global and targeted methods have been commonly used in metabolomics to better understand biological phenotypes, decipher metabolic mechanisms, and identify biomarkers or drug targets for a variety of diseases (1–9). Advances in mass spectrometry (MS) and NMR spectroscopy have enabled the accurate measurement of metabolite levels in blood on a routine basis, which is immensely useful for metabolomics applications (10–18). However, unconnected with the analysis platform, blood metabolite levels are affected by numerous factors, including demographic, clinical, and genetic, as well as preanalytical factors such as collection methods and geographical sites. The variation caused by these factors has deleteriously influenced inferences of many metabolomics studies published so far, including those focused on discovering and validating potential metabolite biomarkers for various diseases. This major challenge in the metabolomics field needs to be addressed.

One promising approach is to understand factors that cause the variation and develop models to alleviate their contribution. However, to date, the lack of suitable mathematical models for blood metabolite levels under homeostasis has hindered progress. Traditional modeling methods use detailed kinetic models and atom balancing or flux balance based on individual enzymatic reaction rates and chemical equilibrium equations (19, 20). These approaches have worked well in discrete cellular models (21, 22) and even at the single organ level for predicting the time-dependent metabolism of metabolites such as glucose (23). In addition, cellular models and more extensive systems approaches based on genomic data have also been constructed (24–26). The challenge for developing approaches to model blood metabolite levels accurately is that they require multiorgan and multicomponent models. Nevertheless, an ability to accurately model metabolite homeostasis would benefit many metabolomics applications.

Numerous efforts have been focused on identifying factors that give rise to blood metabolite variability, including demographic, genomic, preclinical, and even microbial (27–30). We recently modeled the effects of clinical and demographic variables on metabolite levels using seemingly unrelated regression methods (31, 32) and found

Author affiliations: [a]Department of Statistics, Purdue University, West Lafayette, IN 47907; [b]Department of Anesthesiology and Pain Medicine, Northwest Metabolomics Research Center, University of Washington, Seattle, WA 98109; and [c]Department of Epidemiology and Biostatistics, University of California, Irvine, CA 92697

Author contributions: M.Z., D.Z., and D.R. designed research; D.L., G.A.N.G., Z.J., and K.A. performed research; D.L. analyzed data; and D.L., M.Z., D.Z., and D.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.B. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: minzhang@uci.edu, dabao.zhang@uci.edu, or draftery@uw.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2307430121/-/DCSupplemental.

Published February 15, 2024.

that modeling these effects revealed a more realistic performance of metabolite biomarkers after removing such bias. In the present study, we describe the development of quantitative models of blood metabolite levels in healthy adults focusing on reducing sample heterogeneity. The study utilized sample cohorts obtained from geographically dispersed locations to mimic the variation observed in many multi-site studies and help shed light on the factors that lead to disparate metabolite levels. We focused on approximately 50 aqueous metabolites whose concentrations were quantified using ^{1}H 800 MHz NMR spectroscopy. A dramatic reduction of more than 95% in the site-to-site variation of metabolite levels in the top principal component (PC) was achieved based on modeling each metabolite using demographic factors, clinical variables, and especially other metabolites. Reductions in the within-cohort variance were also observed. The results reveal that several metabolites contribute disproportionately to such variation, which could be explained by their association with biological pathways, including biosynthesis and degradation. As an example of the potential utility of the approach, we also show that modeling gender-related metabolic differences retains the interesting variance while reducing unwanted (site-related) variance. Overall, our study demonstrates the intriguing network effect of metabolites that can be utilized to better define homeostatic metabolite levels, which may have beneficial implications for metabolomics, including biomarker discovery, validation, and improved health monitoring.

Results

Modeling Metabolite Levels Reduces Cohort-to-Cohort Variation. Variations of the metabolite profiles obtained using NMR were first revealed using principal component analysis (PCA) of the set of samples, consisting of four cohorts (S-I_A, S-I_B, S-II, and S-III) obtained from three geographically dispersed collection sites. PCA plots for these samples showed large separations (Fig. 1*A*). While cohort S-II was observed to be clearly separated from the other three cohorts in PC1 and PC2, all cohorts were clustered in PC3 and PC5 (*SI Appendix*, Fig. S3*A*). While 36.4% of the total variation of the metabolite profiles is due to between-cohort variation, Table 1 shows that the first two PCs explain 50.6% of the total variation and 86.9% of the between-cohort variation. ANOVA showed that 44 out of 47 metabolites exhibited significant differences across all four cohorts (Benjamini–Hochberg adjusted P-value < 0.05). The three metabolites that were not significantly different were betaine, tryptophan, and tyrosine, with adjusted P-values of 0.41, 0.07, and 0.06 (Dataset S1), respectively.

A linear model for each metabolite was constructed based on demographic factors and other metabolites (*Materials and Methods*). Samples from two cohorts, S-I_B and S-II, were randomly split into two parts, stratified for gender and smoking status. One part (n = 104) was used as the training set to develop the model, and the other part (n = 103) was taken as an in-cohort test set. An out-of-cohort test set consisted of the other two cohorts (S-IA and S-III, n=200) and was combined with the in-cohort test set

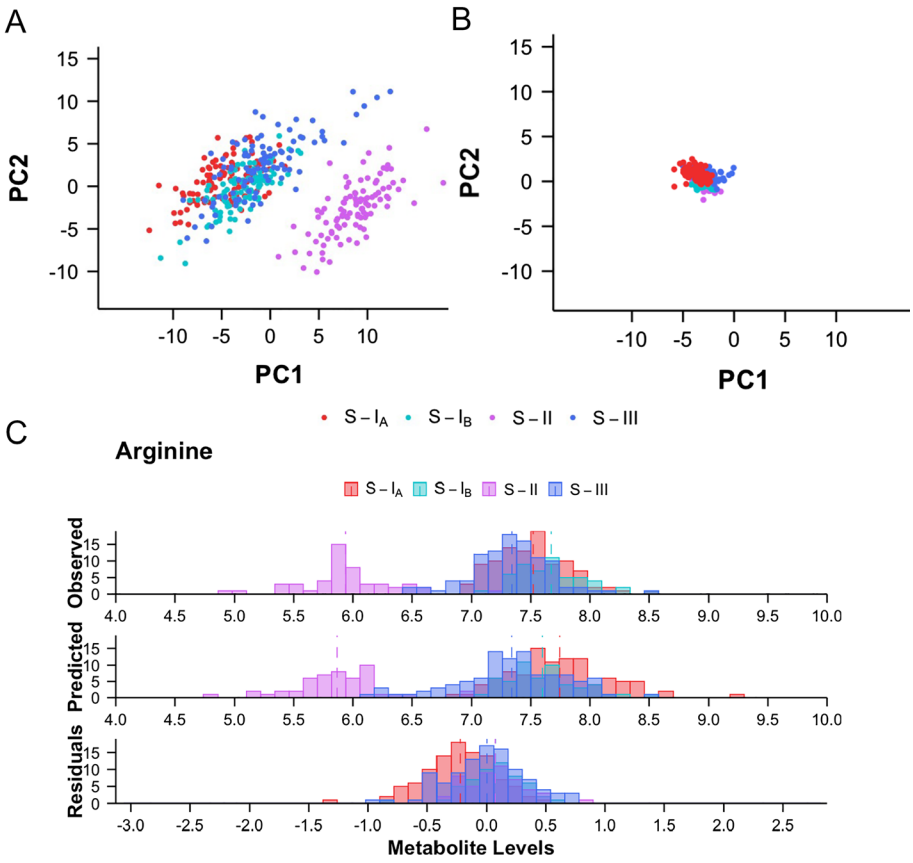

Fig. 1. Modeling metabolite homeostatic levels captures the between-cohort variation. (*A*) PC plot (PC1 vs. PC2) for metabolite data observed in all four cohorts. Each point represents an individual subject's metabolite profile, and each cohort is colored differently. Cohort S-II is clearly separated from the other three. (*B*) PC plot (PC1 vs. PC2) for residuals in the pooled test set (n = 303) following the models built with the training set (n = 104) from S-I_B and S-II. After modeling, the remaining variation is significantly smaller along both (original) PC directions, indicating that the majority of between-cohort variation can be captured by modeling metabolite homeostatic levels. (*C*) Histogram of base-2 logarithm transformed observed levels, predicted levels, and residuals of arginine, which is one of the metabolites with the largest between-cohort variation. Clear separation was shown in observed levels, which were well retained by predicted levels.

Table 1. Variances of top five PCs before and after modeling metabolite homeostatic levels using the two-cohort training set (n = 104)

	Test sets	Modeling	PC1 (38.8%)	PC2 (12.8%)	PC3 (6.2%)	PC4 (4.3%)	PC5 (3.5%)
Total variation	Pooled test set (n = 303)	Before	32.0	11.46	5.46	3.91	3.51
		After	0.82	0.66	1.66	2.51	1.80
		Reduced (%)	97.4	94.3	69.6	36.0	48.7
	In-cohort test subset (n = 103)	Before	42.94	8.32	4.88	3.31	2.30
		After	0.34	0.20	0.40	0.60	0.58
		Reduced (%)	99.2	97.6	91.7	81.8	75.0
	Out-of-cohort test subset (n = 200)	Before	16.0	10.3	4.55	3.61	4.14
		After	0.98	0.37	1.33	1.78	2.09
		Reduced (%)	93.9	96.4	70.7	50.7	49.6
	Test sets	**Modeling**	**PC1 (77.7%)**	**PC2 (9.2%)**	**PC3 (4.7%)**	**PC4 (3.0%)**	**PC5 (2.7%)**
Between-cohort variation	Pooled test set (n = 303)	Before	21.4	2.72	1.23	0.75	1.19
		After	0.18	0.35	0.74	1.13	0.99
		Reduced (%)	99.1	87.1	39.8	-50.2	16.5
	In-cohort test subset (n = 103)	Before	35.5	0.97	1.17	0.84	0.27
		After	0.002	0.006	0.003	0.005	0.010
		Reduced (%)	99.9	99.4	99.8	99.4	96.0
	Out-of-cohort test subset (n = 200)	Before	3.68	0.77	0.03	0.07	1.65
		After	0.19	0.001	0.15	0.006	1.15
		Reduced (%)	94.9	99.8	−363	92.2	30.5

The proportion of variation explained by each PC is shown in the row heads in parentheses.

to serve as a pooled test set (*SI Appendix*, Fig. S1*B*). PC scores shown in Fig. 1*A* were recalculated for the residuals of the test data using the same loadings to quantify the metabolite differences across cohorts after modeling. The results show a dramatic reduction in the distance among cohorts as well as the variation within each cohort, which is indicated by the tight clusters in the PC scores plots (Fig. 1*B* and *SI Appendix*, Fig. S3*B*) compared to the plots before modeling (Fig. 1*A* and *SI Appendix*, Fig. S3*A*). The first two PCs for the pooled test set showed >90% reduction after modeling, while PC3 and PC5 decreased by 69.6% and 48.7%, respectively (Table 1). Table 1 also shows that the in-cohort test subset performed slightly better than the out-of-cohort test subset. Although the variance of PC4 decreased by only 36.0% in the test data from all four cohorts, it decreased by 81.8% and 50.7% in the in-cohort and out-of-cohort test subsets, respectively. This implies that PC4 might explain the differences between the two groups of cohorts, i.e., S-I_B+S-II vs. S-I_A+S-III.

Histograms of samples for individual metabolites before (observed values) and after modeling (residuals) provide a visual description of the reduction in individual metabolite variation after modeling. As an example, Fig. 1*C* shows histograms before and after modeling for arginine, which exhibited the largest difference between cohorts; as shown in the figure, the variations across the four cohorts were dramatically reduced after modeling, and a reduction in the within-cohort variation was also observed.

Fig. 2*A* depicts the total variation of each metabolite's levels obtained after modeling, quantified by the ratio of total average squared deviation, i.e., *ASD(T)*, for each metabolite defined in *SI Appendix, section S.3.2.3*. The results show reductions in total variation for 38 metabolites, 13 of which were reduced by more than 50%. Fig. 2*A* also includes between- and within-cohort variations of each metabolite, quantified by *ASD(B)* and *ASD(W)*, respectively. For variation between cohorts, the results show reductions in 31 metabolites, 22 of which showed reductions by more than 50%. For example, the considerable initial variation of myo-inositol was reduced by >99%, while the small initial variation of betaine was increased by a factor >10. However, the variance reduction within cohorts varied widely across different metabolites. Metabolites were divided into four groups according to $R^2_{(C)}$, which is the proportion of total variation explained by the cohorts (Fig. 2*B*). A group of seven metabolites exhibited the highest $R^2_{(C)}$ (> 50%) between cohorts. These metabolites include myoinositol, arginine, glycerol, aspartic acid, asparagine, sarcosine, and the sum of pyroglutamic acid and glutamine. Interestingly, these metabolites showed the highest reduction in the difference between cohorts after modeling, distinctly different from the other three groups (Fig. 2*B*).

Metabolite Predictors alone Capture Most of the Variation across Cohorts. Metabolite level variation caused by different factors was compared based on the results of regression analyses performed separately against a) cohorts alone, b) a combination of cohorts, demographic factors, and other metabolites, c) demographic factors alone, and d) other metabolites alone, as described below.

To explore the different aspects of the cohort variations, we first calculated the marginal coefficient of determination for cohort ($R^2_{(C)}$) by regressing each metabolite's levels against the cohort factor alone (Dataset S3). The $R^2_{(C)}$ values ranged from 0.027 ~ 0.804 for 44 metabolites. Seven of these metabolites had $R^2_{(C)} > 0.5$, indicating considerable sample heterogeneity across different cohorts (Fig. 2*C*). Conversely, the $R^2_{(C)}$ values were negligible for three metabolites, betaine (0.007), tryptophan (0.018), and tyrosine (0.019) indicating little sample heterogeneity across cohorts.

Metabolite levels were then regressed against the combination of the cohort factor, demographic factors, and other metabolites' levels. The coefficient of conditional determination ($R^2_{(C|M,D)}$) was calculated (Dataset S3). Unlike $R^2_{(C)}$, which quantifies the overall

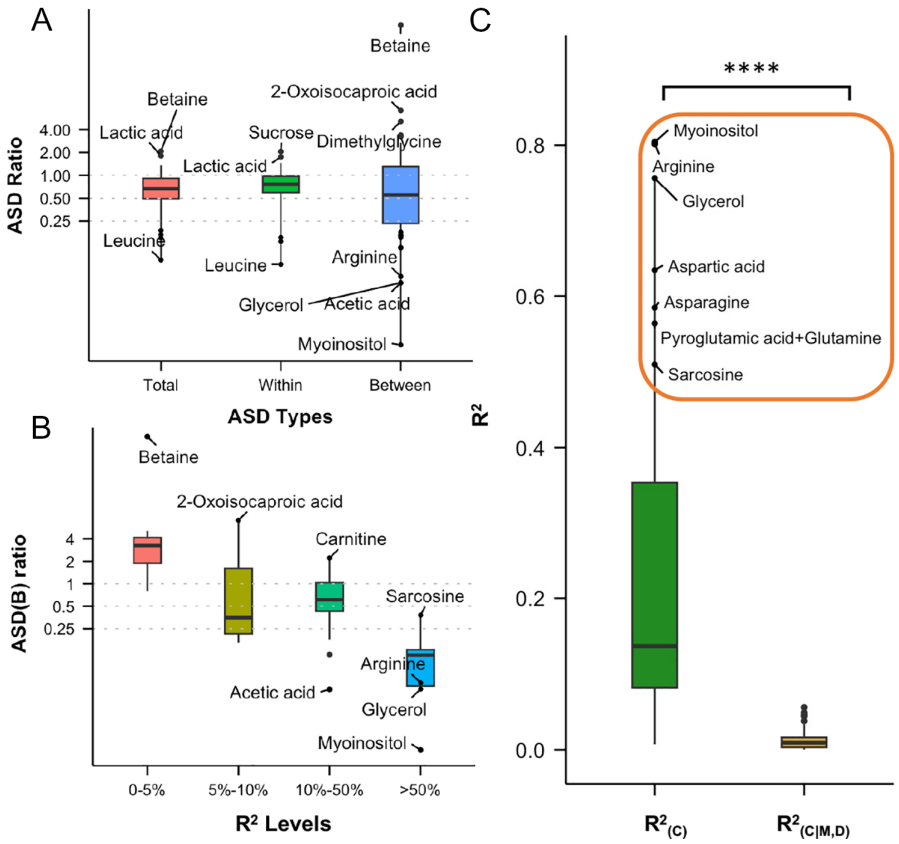

Fig. 2. Results from modeling each metabolite with the effects of demographic factors and other metabolites. (*A*) Boxplot showing the reduction of total, within-cohort, and between-cohort metabolite variations, quantified by *ASD(T), ASD(W), and ASD(B)*, respectively, by modeling. The *y* axis is the ratio of ASDs before and after modeling. Metabolites with a ratio beyond the 1.5 * interquartile range (IQR) of the 3rd quantiles are labeled as outliers. Metabolites with a ratio <0.2 are also labeled, indicating their large reduction in variance after modeling. (*B*) Boxplot showing the reduction of metabolite between-cohort variations, quantified by *ASD(B)*, by modeling for different subsets of metabolites. Under consideration are four groups of metabolites with $R^2_{(C)}$, i.e., the proportion of total variation explained by cohort when no other factors are modeled, at levels of 0 to 5%, 5 to 10%, 10 to 50% or >50%, respectively. The y-axis is the ratio of *ASD*(*B*) before and after modeling. Outlying metabolites as well as metabolites with large reduction (ratio < 0.1) are both labeled in the plot. This figure shows that metabolites with higher levels of between-cohort variation have larger reductions after modeling. (*C*) Boxplot showing the proportion of total variation explained by cohort before and after modeling effects of demographic factors and other metabolites, i.e., $R^2_{(C)}$ and $R^2_{(C|M,D)}$, respectively. Metabolites with R^2 beyond 1.5 * IQR of the 3rd quantile are labeled as outliers in the plot. Very low levels of $R^2_{(C|M,D)}$ imply that modeling the effects of demographic factors and other metabolites leaves little variation explained by cohort; therefore, the cohort variable is of little utility in fitting the predictive model (****$P < 0.0001$ from the Wilcoxon test).

sample heterogeneity across cohorts, $R^2_{(C|M,D)}$ quantifies the variation across cohorts after modeling the contributions made by other metabolites and demographic factors. Fig. 2*C* and *SI Appendix*, Fig. S4 compare $R^2_{(C)}$ and $R^2_{(C|M,D)}$. $R^2_{(C|M,D)}$ showed dramatically smaller values (≤0.056) compared to $R^2_{(C)}$. These results indicate that other metabolites and demographic variables explain the major variation across cohorts.

Separately, the marginal coefficients of determination for demographic factors ($R^2_{(D)}$) were calculated by regressing each metabolite's levels against demographic factors only. The results show much lower $R^2_{(D)}$ values (<0.1) for 41 metabolites. Creatine showed a maximum $R^2_{(D)} = 0.181$, and histidine showed a minimum $R^2_{(D)} = 0.003$. Finally, each metabolite's levels were regressed against the other metabolites' levels, and the marginal coefficient of determination for metabolites ($R^2_{(M)}$) was calculated. $R^2_{(M)}$ values were >0.5 for 44 metabolites, and 5 of these metabolites showed $R^2_{(M)} > 0.9$ (Dataset S3), indicating that strong models could be built using metabolites as predictors. Overall, the results of regression against different factors indicate that metabolites alone capture most of the variance consistently when compared to either cohorts or demographic characteristics (Dataset S3).

The Heatmap of Model Coefficients Reveals the Interactions between Metabolites. Fig. 3*A* shows a heat map of the coefficients for the metabolite models grouped by the correlation of the model coefficients. Among the 47 metabolites, 46 have a mix of predictors with both positive and negative coefficients, with the positive coefficients' percentage ranging from 43 to 86%. Aspartic acid and tryptophan have the lowest percentages of positive coefficients, while glucose has only positive coefficients. The metabolites can be classified into seven groups, i.e., amino acids, amino/imino acids, organic acids, hydroxy acids, sugars, metabolites with quaternary ammonium, and others (*SI Appendix*, Table S7). Connections of the metabolites via the built two-cohort model in Fig. 3*B* indicate the broad network of metabolites that interact across the canonical metabolite pathways defined by Kyoto Encyclopedia of Genes and Genomes (KEGG) (33).

The Addition of a Cohort Contributes to an Improved Homeostatic Model. A three-cohort study was conducted by including an additional cohort (S-III) of samples in the training set. Samples from S-III were randomly but evenly split, stratifying for gender and smoking; one half was combined with the training set in the two-cohort study to build a three-cohort model, and the other half, as well as the remaining samples in $S\text{-}I_B$ and S-II, was used as a test set. Similar to the two-cohort study (Fig. 1*B*), the PC scores were recalculated using the residuals of the test data. The PC scores from the three-cohort model showed better clustering of samples (*SI Appendix*, Fig. S5) compared to the two-cohort model (Fig. 1*B* and *SI Appendix*, Fig. S3*B*). These results show that using an additional cohort to build the model leads to reduced variation

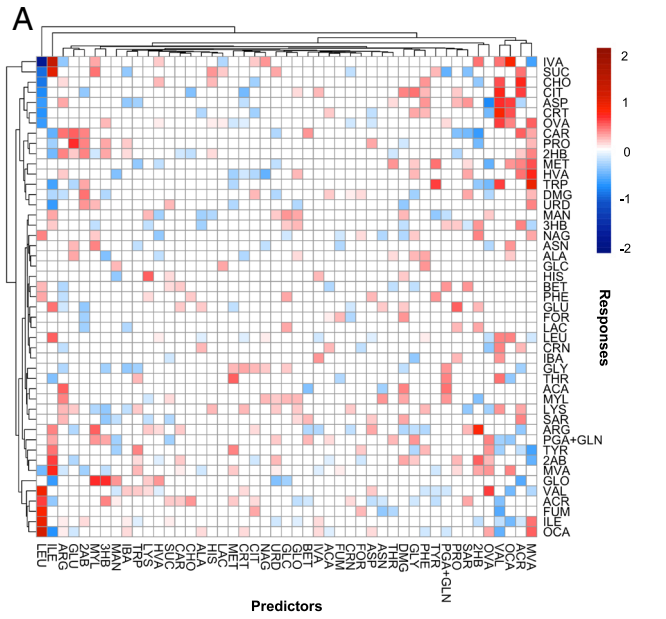

Fig. 3. Summary of the models built in the two-cohort study. (*A*) Heatmap of the models built in the two-cohort study. Shown in each row is the resultant model for a metabolite with each grid colored for the contribution of a metabolomic predictor (in column). A white square means the corresponding metabolite contribution to the model is zero. Metabolites shown in the rows and columns were grouped by (correlational) hierarchical clustering of the metabolite contributions. (*B*) Circular plot of the models built in the two-cohort study. Metabolites were grouped by chemical class. Each link shows how a (response) metabolite is affected by another metabolite in the two-cohort study and is colored the same as the chemical class of the response metabolite. Note that we have separated GLN and PGA to indicate their combined signal contributes to two different chemical classes. Abbreviations of the metabolites can be found in *SI Appendix*, Table S7.

between cohorts as well as within cohorts (*SI Appendix*, Fig. S7). Specifically, the variations of the first two PCs were reduced by 98.5% and 95.7%, respectively (*SI Appendix*, Table S3), better than the results obtained for the two-cohort model. The reduction in the fourth PC was even more significant for the three-cohort model (86.4%) compared to the two-cohort model (36.0%). However, the reductions in the third and fifth PCs were slightly smaller compared with the first two PCs, where a slight separation between cohorts was noticeable (*SI Appendix*, Fig. S5*B*).

The Study of an Additional Cohort Further Verifies the Utility of Homeostatic Models. An additional cohort, S-IV with 253 samples, was used as a new out-of-cohort test set to further evaluate the two- and three-cohort predictive models. *SI Appendix*, Fig. S9 *A* and *C* show the cohort separations when S-IV is projected onto the space of the previous four cohorts. Samples in S-IV lie between S-II and the other three cohorts, with most samples located closer to S-II. PC1 is still the major PC that distinguishes the five cohorts (*SI Appendix*, Fig. S9*A*). But this cohort distance was mostly removed in the residuals after model prediction using the established two- and three-cohort models, and all five cohorts aggregated closely together (*SI Appendix*, Figs. S9*B* and S10*A*). Compared with the PC variances reported in Table 1 and *SI Appendix*, Table S3 for the previous four cohorts, the reduction percentages in PC variances after adding S-IV were similar, with a reduction of over 90% in the first two PCs for both two- and three-cohort models (*SI Appendix*, Tables S4 and S5). The reduction in the fourth PC was improved from 36.0 to 49.6% in the two-cohort model and was similar in the three-cohort model with 86.4 to 85.9%, respectively (Table 1 and *SI Appendix*, Tables S3–S5).

A Study of Gender Differences in Metabolite Profiles Shows the Potential of Homeostatic Models. As an example to demonstrate the possible application of homeostatic models, we investigated gender metabolic differences in the first four cohorts ($S\text{-}I_A$, $S\text{-}I_B$, S-II, and S-III). As shown in *SI Appendix*, Table S6, 73% of the variation in the top PC could be explained by cohort, and for the other top four PCs, each has over one-quarter of its variation explained by cohort. However, gender explained little variation in each of the top five PCs, with the most at 3% in PC1. Although age accounts for 10% of the total variation in PC1, adding it together with gender and smoking status only improves the explained variation by 2% besides cohort, indicating slight complications between cohort and other demographic factors. These results suggested that the top five PCs in this study capture the cohort variation well and may also capture some other demographic factors, which motivated us to identify gender-differentiated metabolites by controlling the effects of the top five PCs. In comparison, we also identified such metabolites without controlling any confounding factors or by controlling cohort only. The p-values, as well as adjusted *P*-values (for multiplicity), from the Wilcoxon test are presented in Dataset S6. Controlling for the top 5 PCs resulted in a somewhat different list of significant gender-differentiating metabolites than the other two tests (*SI Appendix*, Fig. S11). In particular, lysine showed the largest increase in adjusted *P*-value, from 0.460 (no factors controlled) and 0.506 (controlling cohort only) in observed data to 3.99E-08 when controlling the top five PCs, followed by acetylcarnitine and tyrosine which increased by a factor of 5.

Alternatively, we identified gender-differentiated metabolites through two-cohort models rebuilt by excluding a set of eight gender metabolites. These eight metabolites were the most significant gender-specific and also stable across cohorts based on the Wilcoxon test on gender and ANOVA results on cohort, including creatinine, dimethylglycine, betaine, creatine, 2-oxoisocaproic acid (2-OA), leucine, glycine, and tryptophan (Datasets S1 and S6). All of these metabolites, except 2-OA which was not measured, were found to show significant gender differences in a recent large study by Krumsiek et al. (34). We obtained 25 metabolites with significant gender differences after modeling, with 13 metabolites showing increased significance. Twenty metabolites that were

gender-significant in the observed data still showed significance after modeling, with 5 additional metabolites (methionine, phenylalanine, 3-methyl-2-oxovaleric acid, myoinositol, and combined pyroglutamic acid and glutamine) becoming significant. Out of the 25 significant metabolites, 22 were measured in the gender study by Krumsiek et al. (34), and all but two metabolites (dimethylglycine and arginine) were also significant.

Discussion

One of the persistent challenges in metabolomics is the fact that metabolite levels are sensitive to many exogenous factors that are unrelated to the biological stresses of interest, such as disease. In human studies, metabolite levels often vary across collection sites, making validation of biomarker candidates very challenging. Here, using a set of five sample cohorts containing over 650 samples from healthy adults acquired from four geographically dispersed collection sites, we have explored approaches to model the metabolite levels and find that by incorporating other metabolites as predictors, much of the observed site-to-site variation can be removed (Fig. 1 *B* and *C*). The sample set contained one cohort, S-I_A, which was not frozen immediately but nevertheless could still be modeled well along with the other cohorts. Models built using other metabolites were much more explanatory than models using only demographic variables such as age or gender.

The differences in metabolite profiles among cohorts do not appear to cause mostly random effects on the levels of specific metabolites but instead, affect the network of metabolites in predictable ways. Thus, the metabolite relationships can be used to predict profiles in cohorts not included in model training, as shown in Fig. 1*B*, where the model was built using 104 samples from only two of the cohorts. The predictive power of using other metabolites in the models is further evidenced by the fact that differences (both between and within) are reduced when a metabolite is modeled by other metabolites (Fig. 2*A*).

The derived models included 11 significant metabolite variables on average (mean = 10.6 ± 3.6, range 3 to 18), indicating that these models draw upon a range of metabolites. In addition, it was found that metabolites from a variety of different metabolite pathways contribute to the models for each metabolite (Fig. 3*B*), which indicates the broad nature of the network. While it would be interesting to explore models with much more restricted numbers of variables, we have left this work for future studies.

Seven metabolites were particularly affected by site-to-site differences (Fig. 2), but interestingly, they were all modeled quite well. A number of these metabolites are associated with common metabolic pathways, as defined by the KEGG pathway database (33). For example, arginine and sarcosine are both part of arginine and proline metabolism; aspartic acid, glutamine, and arginine are part of arginine biosynthesis; aspartic acid, asparagine, glutamine, and arginine are a part of the biosynthesis of amino acids pathway; glycerol and myoinositol are a part of galactose metabolism, and acetic acid and aspartate are a part of the carbon metabolism pathway.

Not surprisingly, the models were more accurate when more cohorts were used to build the models. Improvements in the clustering of the testing set in PCA plots (*SI Appendix*, Fig. S5) and reduced PCA variance were observed (*SI Appendix*, Table S3). While we did observe reduced variance overall, it was not possible to create a robust model using a single cohort(*SI Appendix*, section S5 and *SI Appendix*, Fig. S8), indicating that much of the intercohort variance is not explained by intracohort metabolite level differences. Nevertheless, it is encouraging that even using two cohorts, we can predict the metabolite levels in a new cohort quite well.

The correlation coefficients for metabolite models, as seen in the heat map (Fig. 3*A*), indicate the magnitude of the association between metabolites. Thirty-nine metabolites have more positive coefficients than negative ones, showing a positive correlation between these metabolites. Given that many of the detected metabolites are amino acids, it is not surprising to see these positive correlations as the amino acid levels fall and rise together with protein synthesis and degradation. Similarly, negative correlations may be understood based on the conversion of one metabolite to the other. For example, the high negative coefficient (−1.7) between leucine and isovaleric acid is due to the conversion of leucine to isovaleric acid in the gut (35). It is well known that a high protein diet leads to high isovaleric acid concentration (36) and isovaleric acid is considered a marker of protein fermentation (37). Generally, branched-chain amino acids (BCCA), leucine, isoleucine, and valine exhibit stronger correlation coefficients. Metabolic pathway analyses involving these amino acids indicate that the majority of the BCCA and other metabolites with high correlations between them are a part of common pathways, including the biosynthesis of amino acids and BCCA, degradation of BCCA, cyanoamino acid metabolism, and 2-oxocarboxylic acid metabolism. Many of the correlation coefficients among other metabolites can similarly be explained based on their association with common metabolite pathways.

Relationships among the metabolites revealed by the modeling can be seen in Fig. 3*B*, in which metabolites contributing to the individual metabolite models are connected to the metabolite levels (the response variable) they help predict. Metabolite models draw from contributions across a number of different classes of metabolites (amino acids, organic acids, etc.), with some particular metabolite types (e.g., BCAAs) participating in many models. Conversely, glucose shows less participation in the models, perhaps because it serves primarily as an input to downstream metabolism (although it is also the product of gluconeogenesis). Similarly, lactate, an endpoint in glycolysis, also participates in only a few of the models.

An important aspect of the modeling is the fact that we can retain interesting variations while reducing uninteresting or deleterious variations. As a test case, we investigated two approaches for retaining metabolic signals in the context of gender-related metabolite differences. Both at the metabolite profile level as well as for single metabolites, we were able to reduce the site contribution while maintaining or even improving the gender metabolite differences. Almost all the significant metabolites observed here were in concordance with those seen in a recent large study on gender metabolic differences (34). We anticipate there will likely be a number of possible approaches for retaining interesting variations that could be developed along these lines as they are explored further.

Finally, there are some differences in performance in prediction between the cohorts. For example, the model built and tested using the samples from the S-I_A and S-III cohorts does not predict as well as the model built using S-I_B and S-II (Table 1). This is most likely because some metabolites in the S-I_A and S-III cohorts have levels beyond the normal range, which may be caused by poor sample treatment (especially S-I_A) or potentially other factors. It may be possible in the future to incorporate more sophisticated models that explicitly take into account sample treatment and resulting effects on metabolite levels that result from enzymatic conversion before the metabolites are extracted (38). Some samples can be identified as problematic; for example, plasma samples with very high levels of hypoxanthine may indicate hemolysis (*SI Appendix*, Fig. S2).

Several limitations in the current analysis are acknowledged, including the use of moderate-size cohorts and a relatively small number of metabolites detected by NMR. Perhaps more importantly, we

have not considered samples from subjects with a particular disease, which is challenging in terms of sample acquisition across multiple cohorts, and thus, we haven't considered the effects of these models on disease signals. Nevertheless, our results showing the site-to-site variation can be reduced while maintaining the signal from gender-related metabolic differences is encouraging.

In conclusion, we have shown that differences observed in blood metabolite levels across cohorts can be successfully modeled to reduce cross-cohort variation. Our results show that even the use of two cohorts in the training set can dramatically reduce this variation and that the inclusion of additional cohorts is useful to significantly reduce the variation. Intracohort variation is also reduced in this manner. A number of future analyses and improvements can be contemplated based on the initial results of this study. Further work in this direction, including the use of MS to expand the number of measurable metabolites, is planned, which should improve the modeling results. In addition, more sophisticated approaches that take into account other biological knowledge, such as metabolite pathway information or even causal network approaches, might be of utility to further refine the approach and to better reveal the relationships among these metabolites and their effects on one another. Future work may also be helpful to reveal and expand some of the relationships of metabolites under homeostasis beyond the current understanding. Finally, the ability to model blood metabolite levels with accuracy should provide useful information to better understand the metabolic status of subjects for a variety of applications.

Materials and Methods

Chemicals. Methanol, sodium phosphate, monobasic (NaH_2PO_4), sodium phosphate, dibasic (Na_2HPO_4), and 3-(trimethylsilyl) propionic acid-2,2,3,3-d_4 sodium salt (TSP) were obtained from Sigma-Aldrich (St. Louis, MO). Deuterium oxide (D_2O) was obtained from Cambridge Isotope Laboratories, Inc. (Andover, MA). Deionized water was purified using an in-house Synergy Ultrapure Water System from Millipore (Billerica, MA).

Biospecimens. Human plasma samples (4 cohorts, total n = 407) from healthy controls were obtained from three geographically dispersed sites. Two cohorts of samples (S-I_A, n = 100; S-I_B, n = 107) were collected at different times (nearly a year apart) and storage conditions from Solomon Park Research Lab (Kirkland, WA), another cohort of samples (S-II, n = 100) was obtained from BioIVT (Westbury, NY) and a fourth (S-III, n = 100) from Innovative Research (Novi, MI). Separately, an additional set of samples (S-IV, n = 253) was procured from a geographically distinct site (BioIVT, Pennsylvania) to evaluate the new methods. Biospecimens used in this study were procured from commercial sources and the research activity was determined not to involve human subjects by the local IRB (STUDY00010094) at the University of Washington. Samples, along with demographic parameters, were obtained based on a custom metabolomics-centric protocol. Briefly, blood samples were collected using sodium heparin tubes and centrifuged for 20 min at 15,000 × g and 4 °C. Supernatant plasma was pipetted into 2 mL cryovials and stored at –80 °C until they were shipped to the Northwest Metabolomics Research Center under dry ice. The samples were kept frozen at –80 °C until used for analysis. *SI Appendix*, Table S1 shows the summary of demographic parameters for all sample cohorts.

Plasma Protein Precipitation. Plasma protein precipitation was performed based on the protocol developed in our laboratory (39, 40). Briefly, frozen plasma samples were thawed at 4 °C, mixed, and then, 200 μL was pipetted into Eppendorf tubes. Methanol was then added in a 1:2 ratio (v sample/v methanol) to precipitate protein, vortexed, and incubated at −20 °C for 20 min. The mixtures were centrifuged at 13,400 × g for 30 min to pellet proteins. Supernatants were collected in fresh vials and dried using a Vacufuge centrifuge concentrator (Eppendorf, Enfield, CT). The dried samples were dissolved in 200 μL phosphate buffer in D_2O (100 mM, pH 7.4) containing 25.0 μM TSP and transferred to 3 mm NMR tubes. The buffer solution was prepared by dissolving 1,124 mg anhydrous NaH_2PO_4 and 250 mg anhydrous Na_2HPO_4 in 100 g D_2O and used without further pH correction.

Analysis of Plasma Samples Using NMR Spectroscopy. NMR experiments were performed at 298 K on a Bruker Avance III 800 MHz spectrometer equipped with a ^{1}H{13C,15N} cryogenically cooled probe and Z-gradients suitable for inverse detection. The Carr-Purcell-Meiboom-Gill pulse sequence with water suppression using presaturation was used to obtain one-dimensional (1D) ^{1}H NMR spectra. A spectral width of 9,615 Hz, 6-s recycle delay, 128 transients, and 32,768 time domain points were used for ^{1}H 1D NMR experiments. The free induction decay signals were Fourier transformed after zero filling by a factor of two and multiplied using an exponential window function with a line broadening of 0.5 Hz. After baseline correction, chemical shifts were referenced to the internal TSP signal. Bruker Topspin version 3.1 software package was used for NMR data acquisition and processing.

Metabolite Identification and Quantitation. Metabolite peaks were identified based on the established literature, specifically, the publications from our laboratory on the blood metabolome (40, 41), the human metabolome database (42), and the biological magnetic resonance data bank (43). Areas of the characteristic metabolite peaks and the reference compound (TSP) peak were integrated using the Bruker AMIX software package and used (along with the number of ^{1}H represented by each peak) to compute concentrations.

Metabolite Data Preprocessing. Of the 51 metabolites detected and quantified by NMR analysis, pyroglutamic acid and glutamine were combined since glutamine undergoes cyclization to form a variable amount of pyroglutamic acid during sample preparation (40) and three metabolites, 1,2-propanediol, ornithine and hypoxanthine, were excluded from the modeling (*SI Appendix, section S1*). Subsequent statistical analyses, therefore, focused on 47 metabolites, along with three demographic factors: age, gender, and smoking status (see details provided in *SI Appendix*, Table S1). Metabolite peaks that were below the detection limit were truncated to zero. Metabolite data were scaled using base-2 logarithm transformation after shifting up one. Any metabolite with a level more than six times its median absolute deviation was considered an outlier and treated as missing in subsequent analysis. All missing values were imputed using the MissForest algorithm (44). After scaling and imputation, metabolite levels and demographic variables were further standardized based on the means and SD estimated from the cohort, S-I_B, which was used as a reference. Additional information on the data preprocessing is provided in *SI Appendix, section S1*.

Metabolite Data Analysis and Modeling. The workflow for metabolite data analysis and modeling is summarized in *SI Appendix*, Fig. S1. Briefly, PCA and ANOVA were performed using data from the first four cohorts (S-I_A, S-I_B, S-II, and S-III, n = 407) to investigate the variations of metabolite levels across cohorts. Linear regression models were fit for each metabolite, accounting for impacts by different demographic variables and other metabolites as follows:

$$\text{Level of } k^{th} \text{ metabolite} \sim \text{Levels of other metabolites} + \text{Demographic factors.} \quad [1]$$

The coefficient of determination (R^2) was calculated to quantify the proportion of variation explained by each set of variables in Eq. **1** (see *SI Appendix, section S2* for details). Samples from different cohorts were randomly split into training and test sets to further evaluate the predictability of metabolite profiles with these variables. Initially, we constructed a two-cohort model using samples from S-I_B and S-II. Samples of 207 in S-I_B and S-II were randomly split into training (n = 104) and in-cohort test sets (n = 103) using stratified splitting, and a predictive linear model for each metabolite was constructed using the training set followed by backward elimination to select important metabolomic predictors (with a significance level of 0.05). These predictive models were further applied to the in-cohort test set (n = 103) from S-I_B and S-II, as well as an out-of-cohort test set (n = 200) from S-I_A and S-III, to evaluate the reduction in sample heterogeneity by modeling homeostatic metabolite levels. Detailed information on model selection and evaluation is provided in *SI Appendix, section S3*. For each metabolite *k* in the test set, three types of ASDs were calculated, $ASD(T_k)$, $ASD(W_k)$, and $ASD(B_k)$, to quantify the metabolite's total variation, within-cohort variation, and between-cohort variation, respectively. The three types of ASDs and their reductions were recalculated using residuals from the

test set to illustrate the utility of modeling metabolite homeostatic levels. Similarly, we carried out a three-cohort study with a 50% training set (n = 154) from S-I_B, S-II, and S-III. Detailed information can be found in *SI Appendix, section S4*. We used S-IV (n = 253) as an independent test cohort to further evaluate our proposed predictive models in a similar fashion. Detailed evaluation methods and results with S-IV can be found in *SI Appendix, section S6*.

Analysis of Gender Metabolic Differences. The previously constructed top five PCs were regressed against demographic factors including cohort and gender for decomposition of their variations. Two approaches were explored to study gender-differentiated metabolites. The first approach used all 407 samples from the first four cohorts (S-I_A, S-I_B, S-II, and S-III) and conducted Wilcoxon tests on each metabolite by controlling for no confounding factors, cohort only, or the top five PCs instead. The second approach reconstructed the two-cohort models by excluding the eight most significant gender-specific metabolites, which were also stable across the training cohorts. The residuals were calculated from the resultant models and used for Wilcoxon tests of gender difference. false discovery rate (FDR)-adjusted *P*-values were used for determining significance with a 0.05 cutoff. More detailed information can be found in *SI Appendix, section S7*.

Data, Materials, and Software Availability. All metabolomic data and codes for all analyses presented in the manuscript and supporting information have been deposited at Purdue University (45). All other data are included in the article and/or supporting information.

ACKNOWLEDGMENTS. We gratefully acknowledge the financial support from NIH grants R01GM131491, R01GM131491-02S2, and P30DK035816.

1. N. Psychogios *et al.*, The human serum metabolome. *PLoS One* **6**, e16957 (2011).
2. W. B. Dunn *et al.*, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6**, 1060–1083 (2011).
3. J. C. Lindon, J. K. Nicholson, The emergent role of metabolic phenotyping in dynamic patient stratification. *Expert Opin. Drug Metab. Toxicol.* **10**, 915–919 (2014).
4. E. P. Rhee, R. E. Gerszten, Metabolomics and cardiovascular biomarker discovery. *Clin. Chem.* **58**, 139–147 (2012).
5. T. J. Wang *et al.*, Metabolite profiles and the risk of developing diabetes. *Nat. Med.* **17**, 448–453 (2011).
6. G. A. Nagana Gowda *et al.*, Metabolomics-based methods for early disease diagnostics. *Expert Rev. Mol. Diagn.* **8**, 617–633 (2008).
7. G. J. Patti, O. Yanes, G. Siuzdak, Metabolomics: The apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **13**, 263–269 (2012).
8. S. L. Robinette, E. Holmes, J. K. Nicholson, M. E. Dumas, Genetic determinants of metabolism in health and disease: From biochemical genetics to genome-wide associations. *Genome Med.* **4**, 30 (2012).
9. D. S. Wishart, Emerging applications of metabolomics in drug discovery and precision medicine. *Nat. Rev. Drug Discov.* **15**, 473–484 (2016).
10. N. J. Serkova, T. J. Standiford, K. A. Stringer, The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. *Am. J. Respir. Crit. Care Med.* **184**, 647–655 (2011).
11. G. A. Nagana Gowda, Y. N. Gowda, D. Raftery, Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. *Anal. Chem.* **87**, 706–715 (2015).
12. P. Soininen, A. J. Kangas, P. Würtz, T. Suna, M. Ala-Korpela, Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.* **8**, 192–206 (2015).
13. A.-H. Emwas *et al.*, NMR spectroscopy for metabolomics research. *Metabolites* **9**, 123 (2019).
14. W. B. Dunn, D. I. Broadhurst, H. J. Atherton, R. Goodacre, J. L. Griffin, Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.* **40**, 387–426 (2011).
15. T. Kind, O. Fiehn, Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinf.* **8**, 105 (2007).
16. D. Raftery, Ed., *Mass Spectrometry in Metabolomics: Methods and Protocols* (Humana Press, 2014).
17. S. Alseekh *et al.*, Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nat. Methods* **18**, 747–756 (2021).
18. M. Jacob, A. L. Lopata, M. Dasouki, A. M. Abdel Rahman, Metabolomics toward personalized medicine. *Mass Spectrom. Rev.* **38**, 221–238 (2019).
19. R. Heinrich, S. Schuster, *The Regulation of Cellular Systems* (Chapman & Hall, 1996).
20. G. Stephanopoulos, J. J. Vallino, Network rigidity and metabolic engineering in metabolite overproduction. *Science* **252**, 1675–1681 (1991).
21. R. L. Schiek, E. E. May, BioXyce: An engineering platform for the study of cellular systems. *IET Syst. Biol.* **3**, 77–89 (2009).
22. A. Mardinoglu, J. Nielsen, New paradigms for metabolic modeling of human cells. *Curr. Opin. Biotechnol.* **34**, 91–97 (2015).
23. M. König, H.-G. Holzhütter, Kinetic modeling of human hepatic glucose metabolism in type 2 diabetes mellitus predicts higher risk of hypoglycemic events in rigorous insulin therapy. *J. Biol. Chem.* **287**, 36978–36989 (2012).
24. N. C. Duarte *et al.*, Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1777–1782 (2007).
25. I. Thiele *et al.*, A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.* **31**, 419–425 (2013).
26. P. Ghaffari, A. Mardinoglu, J. Nielsen, Cancer metabolism: A modeling perspective. *Front. Physiol.* **6**, 382 (2015).
27. S. L. Navarro *et al.*, Demographic, health and lifestyle factors associated with the metabolome in older women. *Metabolites* **13**, 514 (2023).
28. V. L. Stevens, E. Hoover, Y. Wang, K. A. Zanetti, Pre-analytical factors that affect metabolite stability in human urine, plasma, and serum: A review. *Metabolites* **9**, 156 (2019).
29. J. Tokarz, J. Adamski, "Confounders in metabolomics" in *Metabolomics for Biomedical Research*, J. Adamski (Elsevier, 2020), pp. 17–32.
30. N. Bar *et al.*, A reference map of potential determinants for the human serum metabolome. *Nature* **588**, 135–140 (2020).
31. C. Chen *et al.*, Exploring metabolic profile differences between colorectal polyp patients and controls using seemingly unrelated regression. *J. Proteome Res.* **14**, 2492–2499 (2015).
32. C. Chen *et al.*, Altered metabolite levels and correlations in patients with colorectal cancer and polyps detected using seemingly unrelated regression analysis. *Metabolomics* **13**, 125 (2017).
33. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
34. J. Krumsiek *et al.*, Gender-specific pathway differences in the human serum metabolome. *Metabolomics* **11**, 1815–1833 (2015).
35. A. Thierry, M.-B. Maillard, M. Yvon, Conversion of l-leucine to isovaleric acid by propionibacterium freudenreichii TL 34 and ITGP23. *Appl. Environ. Microbiol.* **68**, 608–615 (2002).
36. M. Aguirre *et al.*, Diet drives quick changes in the metabolic activity and composition of human gut microbiota in a validated in vitro gut model. *Res. Microbiol.* **167**, 114–125 (2016).
37. D. Rios-Covian *et al.*, An overview on fecal branched short-chain fatty acids along human life and as related with body mass index: Associated dietary and anthropometric factors. *Front. Microbiol.* **11**, 973 (2020).
38. C. Brunius *et al.*, Prediction and modeling of pre-analytical sampling errors as a strategy to improve plasma NMR metabolomics data. *Bioinformatics* **33**, 3567–3574 (2017).
39. G. A. Nagana Gowda, Y. N. Gowda, D. Raftery, Expanding the limits of human blood metabolite quantitation using NMR spectroscopy. *Anal. Chem.*, **87**, 706–715 (2015).
40. G. A. Nagana Gowda, Y. N. Gowda, D. Raftery, Massive glutamine cyclization to pyroglutamic acid in human serum discovered using NMR spectroscopy. *Anal. Chem.* **87**, 3800–3805 (2015).
41. G. A. Nagana Gowda, D. Raftery, Whole blood metabolomics by 1 H NMR spectroscopy provides a new opportunity to evaluate coenzymes and antioxidants. *Anal. Chem.* **89**, 4620–4627 (2017).
42. D. S. Wishart *et al.*, HMDB 5.0: The human metabolome database for 2022. *Nucleic Acids Res.* **50**, D622–D631 (2022).
43. E. L. Ulrich *et al.*, BioMagResBank. *Nucleic Acids Res.* **36**, D402–D408 (2007).
44. D. J. Stekhoven, P. Buhlmann, MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).
45. D. Liu *et al.*, Across-site predictability of metabolite profiles. Metabolic Data Analysis. https://www.zstats.org/metabolomics/projects1/across_site_model/across_site_model.html. Deposited 23 October 2023.