Multiset correlation and factor analysis enables exploration of multi-omic data

Brielin C. Brown*†[1, 2], Collin Wang†[1, 3], Silva Kasela[1, 4], François Aguet[5, 6], Daniel C. Nachun[7], Kent D. Taylor[8], Russell P. Tracy[9], Peter Durda[9], Yongmei Liu[10], W. Craig Johnson[11], David Van Den Berg[12], Namrata Gupta[6], Stacy Gabriel[6], Joshua D. Smith[13], Robert Gerzsten[14], Clary Clish[6], Quenna Wong[15], George Papanicolau[16], Thomas W. Blackwell[17], Jerome I. Rotter[8], Stephen S. Rich[18], Kristin G. Ardlie[6], David A. Knowles‡[1,2,3,4], and Tuuli Lappalainen‡[1,4,19]

[1]*New York Genome Center, New York, NY, USA*
[2]*Data Science Institute, Columbia University, New York, NY, USA*
[3]*Department of Computer Science, Columbia University, New York, NY, USA*
[4]*Department of Systems Biology, Columbia University, New York, NY, USA*
[5]*Illumina Incorporated, San Francisco, CA, USA*
[6]*The Broad Institute of MIT and Harvard, Boston, MA, USA*
[7]*Department of Pathology, Stanford University, Stanford, CA, USA*
[8]*Department of Pediatrics, The Institute for Translational Genomics and Population Sciences, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA*
[9]*Department of Pathology and Laboratory Medicine, Larner College of Medicine, University of Vermont, Burlington, VT, USA*
[10]*Department of Medicine, Duke University Medical Center, Durham, NC, USA*
[11]*Department of Biostatistics, University of Washington, Seattle, WA, USA*
[12]*Department of Clinical Preventative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA*
[13]*Northwest Genomics Center, University of Washington, Seattle, WA, USA*
[14]*Beth Israel Deaconess Medical Center, Division of Cardiovascular Medicine, Boston, Massachusetts, USA*
[15]*Department of Biostatistics, University of Washington, Seattle, WA, USA*
[16]*Division of Cardiovascular Sciences, National Heart, Lung, and Blood Institute, Bethesda, MD, USA*
[17]*Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, MI, USA*
[18]*Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA*
[19]*Science for Life Laboratory, Department of Gene Technology, KTH Royal Institute of Technology, Stockholm, Sweden.*

[*] Correspondence: bbrown@nygenome.org
[†] Co-first author
[‡] Equal contributions

Abstract

Multi-omics datasets are becoming more common, necessitating better integration methods to realize their revolutionary potential. Here, we introduce Multi-set Correlation and Factor Analysis, an unsupervised integration method that enables fast inference of shared and private factors in multi-modal data. Applied to 614 ancestry-diverse participant samples across five 'omics types, MCFA infers a shared space that captures clinically relevant molecular processes.

Main

Recent years have seen an explosion in multi-omic data, with studies simultaneously profiling RNA expression, protein levels, chromatin accessibility and more[1]. By providing complementary views into the underlying biology, these datasets promise to illuminate molecular processes and disease states that cannot be gleaned from any lone modality[2]. However, joint inference methods are lacking in either the number or type of modes that can be used, or in flexibility and efficiency[1]. Multi-omic data bring substantial challenges: distributions differ between modes, the sample size is typically small relative to features, efficient algorithms are needed, and each mode has contributions from factors that are shared between modes and unique to itself[3,4]. Canonical correlation analysis (CCA) is a statistical technique that infers shared factors between two data modes by finding correlated linear combinations of the features in each[5]. CCA has enjoyed substantial attention in genomics[6–9], however, extending CCA to additional modes is fraught; at least 10 different formulations are equivalent in the two-mode case[10], and many are challenging to fit[11]. Equivalently, CCA can be conceptualized as a probabilistic model (pCCA), revealing a connection to factor analysis[12].

We have developed Multiset Correlation and Factor Analysis (MCFA, Figure 1a), an unsupervised integration method that generalizes pCCA and factor analysis, enabling fast inference of shared and private factors in multimodal data. MCFA is based on two insights: 1) unlike traditional CCA, pCCA has only one natural extension to multi-modal data, which is both conceptually elegant and efficient to fit, and 2) after fitting pCCA, the residual in a mode represents private structure, which is well-modeled by factor analysis. Our method combines these insights to fit factors that are shared across modalities and private to each simultaneously. For efficiency and regularization, MCFA uses the top principal components (PCs) of each mode[6,7]. It allows use of random matrix techniques[13] to choose the shared dimensionality and number of PCs, eliminating tuning parameters. Finally, MCFA is a natural approach to integration: in the supplemental note, we detail a theoretical connection between our model and multiset CCA.

We have applied MCFA to 614 ancestry-diverse individuals from the Multi-Ethnic Study of Atherosclerosis (MESA)[14], which has collected comprehensive phenotypic data of its subjects. The Trans-Omics for Precision Medicine (TOPMed)[15] program instituted a multi-omics pilot study to evaluate the utility of long-term stored samples for discovery related to heart, lung, blood, and sleep disorders. MESA provided samples for five 'omics types: 1) whole genome sequencing (WGS), 2) RNA-sequencing of peripheral blood mononuclear cells (PBMCs), 3) DNA methylation array profiling from whole blood 4) protein mass spectrometry of blood plasma, and 5) metabolite mass spectrometry of blood plasma. We integrated RNA-sequencing, methylation, protein, and metabolite data from Exam 1 using MCFA, which inferred a fourteen-dimensional shared space. We found that shared structure explained a large proportion of the variance in each mode (Figure 1b, right). Protein levels had the highest sharing with 29.2% of the variance explained (VE) by the shared space, followed by RNA and metabolite levels (16.6% and 17.1%, respectively). Methylation showed the least sharing, with only 8.1% VE by the shared space. Due to the high dimensionality of the data and the limited sample size, about half of the variance in each dataset is unmodeled to reduce overfitting. Using MCFA, it is possible to further infer the variance in each modality explained by the individual factors, thus determining which modalities contribute to each (Figure 1b, left). Our top factor has contributions from all modalities, but their respective contributions to the other factors vary substantially.

We used uniform manifold approximation and projection (UMAP)[16] to construct a 2D embedding of the shared and private spaces (Figure 1c). We noticed a striking clustering of the individuals by self-reported ancestry (SRA) and sex in the shared space, even though the top PCs of individual modes do not cluster by these factors (Figure S1), and the shared space was inferred without genetic or sex chromosome features. Shared factor 1 separates Black and white individuals, with Hispanic individuals in between, while factor 3 separates Chinese individuals, and factor 2 differentiates by sex (Figure S1, S2). We validated this structure via leave-one-out cross-validation, indicating our PC selection strategy mitigated over-fitting (Figure S3).

Next, we evaluated the total phenotypic variance explained by each of our inferred spaces (Figure 1d, S2, Table S1-S3). The shared space captured 95.3% of the variation in sex, 83.3% in site, 80.0% in SRA, and 60.2% in age. The shared space also captured anthropomorphic differences such as BMI (51.0% VE) and clinical measures including those related to kidney function (creatine, 64.8% VE) and inflammation (TNF-alpha receptor-1 69.1% VE). We used CIBERSORT[17] and the Houseman method[18] to estimate the cell-type composition of our RNA (PBMC) and methylation (whole blood) samples, respectively. Both shared and privates spaces contributed to the relative proportions of PBMC-abundent cell types (e.g. T cells and NK cells) estimated from both data modalities, while the proportion of PBMC-depleted types (e.g. neutrophils) estimated from the methylation data was only captured by the methylation private space. Modality-private spaces frequently captured technical factors: 100% of the variance in sequencing center and 71.6% of the variance in 3-prime bias are captured by the RNA private space, while 76.8% of methylation array batch is captured by its private space. Many phenotypes that are themselves measurements of metabolites were captured by the metabolite private space, however the strongest association was with the month of sample collection (85.8% VE). We noticed no large associations between the protein private space and any of our metadata, despite several of our phenotypes being clinical protein markers, however, several of these factors are partially captured by the shared space.

Finally, we integrated WGS data by conducting a GWAS of the inferred factors while controlling for site, age, sex and 11 genotype PCs. Given our limited sample size, we did not expect to find genome-wide significant associations. However, we hypothesized that genetic associations with our inferred factors, which represent major axes of molecular variation, may be enriched for known GWAS hits or *trans*-eQTLs. We obtained a list of 10,174 such associations from the eQTLgen consortium[19], of which 3,854 are trans-eQTLs, and further defined a more limited set of 1,107 "influential" *trans*-eQTLs that affect at least 10 genes. We tested the GWAS of each factor for enrichment of these three categories and found 9 significant enrichments (mean $\chi^2_{cat} > 1$, FDR 5%, Figure 2a, Figure S4). Factor 7 showed the strongest enrichment for reported GWAS hits and *trans*-eQTLs. The top GWAS SNPs associated with factor 7 are from blood lipid studies and are located primarily around the FADS1 and FADS2 genes that are known to regulate lipid metabolism[20]. These include rs174541 ($p = 4.3 \times 10^{-5}$ for factor 7 association) which is also reported in GWAS of type-2 diabetes[21], rs174549 ($p = 5.6 \times 10^{-5}$) which is also reported in GWAS of white blood cell count[22], and rs1535 ($p = 8.3 \times 10^{-5}$), which is also reported in GWAS of inflammatory bowel disease[23] (Table S4). Factor 7 explains 6.7% of the modeled variation in methylation, the largest of any factor, and many of the top-associated SNPs were also cis-methylation QTLs in MESA[24] (Figure 2c). Factor 7 is anti-correlated with sample proportion of CD8 T cells and NK cells estimated from methylation data ($\rho = -0.41$ and $\rho = -0.25$), and correlated with BMI ($\rho = 0.25$) and measures of inflammation including TNF-R1 ($\rho = 0.33$) and interleukin-6 ($\rho = 0.20$) (Figure 2b). While further research is needed to establish causal relationships of these genetic effects on methylation in cis and trans as well as on diverse traits, we note that DNA methylation patterns have been previously associated with lipid metabolism and metabolic disease[25,26].

MCFA has several advantages compared to other multi-omic integration approaches. Compared to group factor analysis methods[4], MCFA separates modality-specific from dataset-shared factors. Compared to non-negative matrix factorization-based methods[3] that share a feature weight set across modalities, MCFA is able to use all data types. Due to the use of observational data and unsupervised methods, all analyses should be considered exploratory; they can find structure in the data while generating hypotheses but cannot be used to make causal claims and may reflect properties of the underlying data. For example, in MESA the sample collection site is strongly correlated with self-reported ancestry (SRA) and air-quality. We repeated our analysis of the variance explained by the learned space while additionally controlling for site (Table S3), and noticed a small decrease in the proportion of VE in SRA (from 80.0% to 71.6%) and a large decrease in the variance explained in PM25 (from 66.8% to 24.2%). In this study air quality and site are nearly co-linear and thus their independent effects cannot be distinguished. Future work with larger sample sizes may allow for network inference methods to generate directed hypotheses[27]. Genetic associations are particularly valuable in this, with the inferred axes of molecular variation providing a promising future trait for GWAS and pheWAS studies. TOPMed is among the most ambitious current efforts to collect multi-omic population-level data, thus given the results of this pilot analysis we expect future integration studies in this cohort to be fruitful.

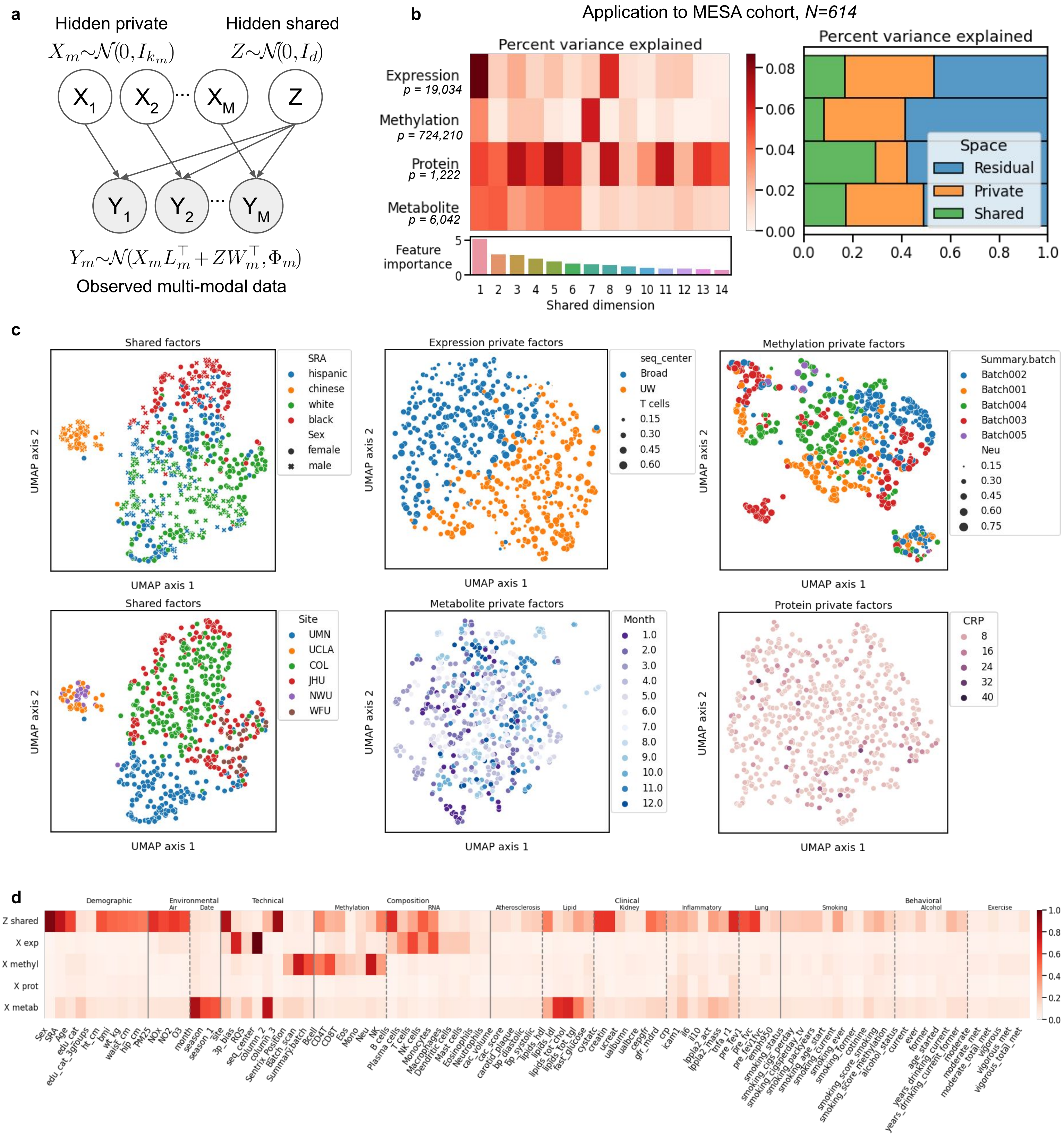
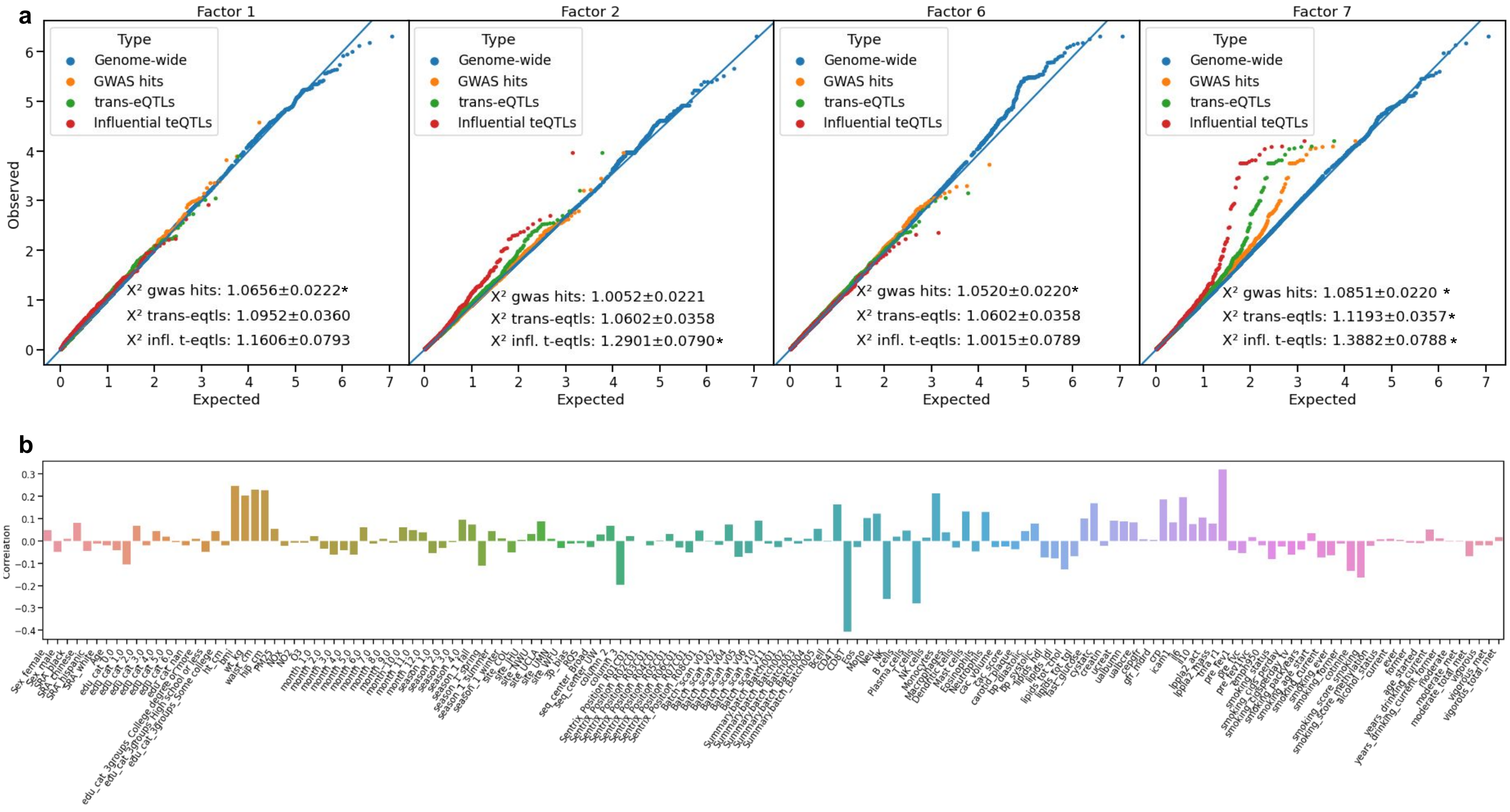


Figure 1: a) The MCFA model. b) Breakdown of the variance in four omics types captured by the inferred space. c) UMAP embedding of the shared and private spaces, annotated with the most relevant feature set. d) Variance in sample metadata explained by each learned space.



c

Trait	rsid	Chr	Position	Gene	p_val_trait	p_val_F7	CpG	p_val_mQTL	n_snps
Red blood cell fatty acid levels	rs174541	11	61565908	FADS2	3.000000e-19	0.000043	cg19610905	3.964333e-115	29
Height	rs7184046	15	75866150	PTPN9	2.000000e-10	0.000061	cg01268058	1.333136e-48	1
Systemic lupus erythematosus	rs4852324	2	74202578	DGUOK-AS1	6.000000e-14	0.000343	cg04315689	5.291427e-07	1
Mature red cell;HGB	rs1256061	14	64703593	ESR2	NaN	0.000464	cg09764150	4.629048e-13	1
IgG glycosylation	rs2186369	22	24170996	SMARCB1	3.000000e-08	0.000505	cg11916091	2.902150e-22	1
Mean platelet volume	rs7743045	6	119102271	SELENOKP3	NaN	0.001140	cg12098441	3.562536e-18	1
Polycystic ovary syndrome	rs2059807	19	7166109	INSR	1.000000e-08	0.001288	cg16940601	5.447991e-31	1
Plasma omega-6 polyunsat. fatty acid levels	rs7925523	11	61442492	DAGLA	2.000000e-59	0.001343	cg25599065	2.748742e-41	2
Height	rs7155279	14	92485881	TRIP11	1.000000e-10	0.001414	cg27159443	4.693894e-42	3
Proinsulin levels	rs11603334	11	72432985	ARAP1	3.000000e-102	0.001421	cg03713592	2.969683e-114	2

Figure 2: a) QQ-plot of a GWAS for factors 1, 2, 6, and 7. b) Correlation of factor 7 with sample metadata. c) Top unique SNP-CpG pairs for known trait-associated SNPs additionally associated with factor 7. n_snps indicates the number of SNPs suggestively-associated with factor 7 that have the same top-associated CpG.

References

1. Krassowski, M., Das, V., Sahu, S. K. & Misra, B. B. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. *Frontiers in Genetics* **11,** 1598 (Dec. 2020).
2. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biology 2017 18:1* **18,** 1–15 (May 2017).
3. Welch, J. D. *et al.* Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity. *Cell* **177,** 1873–1887.e17 (June 2019).
4. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14** (June 2018).
5. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28,** 321–377 (May 1936).
6. Brown, B. C., Bray, N. L. & Pachter, L. Expression reflects population structure. *PLoS Genetics* **14,** e1007841 (Dec. 2018).
7. Soneson, C., Lilljebjörn, H., Fioretos, T. & Fontes, M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics* **11,** 1–20 (Apr. 2010).
8. Naylor, M. G., Lin, X., Weiss, S. T., Raby, B. A. & Lange, C. Using Canonical Correlation Analysis to Discover Genetic Regulatory Variants. *PLoS ONE* **5** (2010).
9. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology 2018 36:5* **36,** 411–420 (Apr. 2018).
10. Kettenring, J. R. Canonical analysis of several sets of variables. *Biometrika* **58,** 433–451 (1971).
11. Asendorf, N. A. Informative Data Fusion: Beyond Canonical Correlation Analysis (2015).
12. Bach, F. R. & Jordan, M. I. A probabilistic interpretation of canonical correlation analysis (2005).
13. Marčenko, V. A. & Pastur, L. A. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik* **1,** 457–483 (1967).
14. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *American journal of epidemiology* **156,** 871–881 (Nov. 2002).
15. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature 2021 590:7845* **590,** 290–299 (Feb. 2021).
16. Mcinnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv: `1802.03426v3` (2020).
17. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods 2015 12:5* **12,** 453–457 (Mar. 2015).
18. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13** (May 2012).
19. Võsa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics 2021 53:9* **53,** 1300–1310 (Sept. 2021).
20. Schaeffer, L. *et al.* Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Human molecular genetics* **15,** 1745–1756 (June 2006).
21. Dupuis, J. *et al.* New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature genetics* **42,** 105 (2010).
22. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167,** 1415–1429.e19 (Nov. 2016).

23. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature genetics* **47,** 979–986 (Aug. 2015).

24. Aguet, F. & Lappalainen, T. Placeholder for FA paper (2022).

25. Mittelstraß, K. & Waldenberger, M. DNA methylation in human lipid metabolism and related diseases. *Current Opinion in Lipidology* **29,** 116–124 (Apr. 2018).

26. Gomez-Alonso, M. D. C. *et al.* DNA methylation and lipid metabolism: an EWAS of 226 metabolic measures. eng. *Clinical epigenetics* **13,** 7 (Jan. 2021).

27. Brown, B. C. & Knowles, D. A. Phenome-scale causal network discovery with bidirectional mediated Mendelian randomization. *bioRxiv,* 2020.06.18.160176 (June 2020).

1 Online methods

1.1 Multiset correlation and factor analysis

Let $Y = \{Y_m\}_{m=1}^{M}$ be a set of $N \times p_m$ observed data matrices: N individuals measured in M data modalities consisting of p_m features each. We model each observed mode as having contributions from two low-dimensional hidden factors (Figure 1a, Figure S7)

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_d) \\ x_n^m &\sim \mathcal{N}(0, I_{k_m}) \\ y_n^m &\sim \mathcal{N}(W_m z_n + L_m x_n^m, \Psi_m) \end{aligned}$$

where d is the shared hidden dimensionality, k_m are the dataset-private hidden dimensionalities, W_m are $p_m \times d$ shared space loading matrices, L_m are $p_m \times k_m$ private space loading matrices and $\Psi_m = \mathrm{diag}(\psi_m^1, \ldots, \psi_m^{p_m})$ are the diagonal residual covariance matrices. Given Y, d and k_m, our goal is to infer the hidden factors Z and X_m and loading matrices W_m and L_m. This can be accomplished using a straightforward application of expectation-maximization (EM)[1]. For a derivation of the EM update equations, as well as a more detailed exposition including the relationship to pCCA, factor analysis and other multiset CCA (MCCA) methods, see the Supplemental Note. In practice, we center and scale all data variables. This is not strictly required, however it enables simple estimation of the number of PCs to include and simplifies explained variance calculations, see below.

1.2 Model initialization

An important aspect of EM optimization is choosing a good initialization. We benchmarked three approaches to initializing W: random initialization and two versions of MCCA that correspond to maximizing the sum of pairwise correlations with the average variance and average norm constraints. These MCCA fomulations can be solved via simple eigendecompositions. We found that the sum of pairwise correlations with average variance constraint produced the best initial estimates (Figure S5). This can be solved with a simple two step procedure: 1) whiten each data matrix using the singular value decomposition (SVD), 2) perform a second SVD on the concatenated whitened data matrices[2]:

```
    **Input:** $Y_1, \ldots, Y_M, d$
    **Result:** $\hat{W} = [W_1^\top : \ldots : W_M^\top]^\top$
    $U_{all} \leftarrow$ concatenate(SVD($Y_1$).$U$, ..., SVD($Y_M$).$U$) ;
    $\hat{W} \leftarrow$ SVD($U_{all}$).$V[:, 0 : d]$ ;
    $\hat{\rho} \leftarrow$ SVD($U_{all}$).$\lambda[0 : d]$ ;
    **return** $\hat{W}, \hat{\rho}$
```

We initialize L and Ψ using probabilistic PCA on the residual data matrices after fitting MCCA. Specifically:

```
    **Input:** $Y_i, W_i, N, k_i$
    **Result:** $\hat{L}_i, \hat{\Psi}_i$
    $\Sigma_i^\perp \leftarrow Y_i^\top Y_i / N - W_i W_i^\top$ ;
    $\hat{L}_i \leftarrow$ eigh($\Sigma_i^\perp$).$V[:, 0 : k_i]$ ;
    $\sigma^2 \leftarrow$ mean(eigh($\Sigma_i^\perp$).$\lambda[k_i :]$) ;
    $\hat{\Psi}_i \leftarrow \sigma^2 \mathbb{1}_{k_i}$ ;
    **return** $\hat{L}_i, \hat{\Psi}_i$
```

1.3 High dimensionality and selection of hyperparameters

There are two primary approaches to control for over-fitting in applications of CCA-type methods to high-dimensional ($N \ll p$) problems. The first is to use penalized optimization techniques, where the objective function additionally contains an l_1 constraint on the weight matrices[3]. The second is to project each dataset onto its *informative* principal components[4–6]. In this application, we choose the latter approach in order to find components with broad effects on the structure of the data, rather than specific effects on small numbers of molecular features[6]. We choose the number of principal components of each dataset using the Marchenko-Pasteur law[7], which states that for mean 0, variance 1 data, principal components with corresponding eigenvalues above $\lambda_m = 1 + \sqrt{p_m/N}$ should be considered non-noise. We are not aware of a corresponding law for the cross-covariance matrices used in CCA, however, the empirical spectral distribution of the cross-covariance of matrices of random noise can be easily estimated in practice:

```
Input: $N, k = \{k_m\}_{m=1}^{M}, n_{it}$
Result: $\bar{\rho}$
$\rho \leftarrow \mathsf{array}[n_{it}]$ ;
for $it \leftarrow 0$ to $n_{it}$ do
    for $k_m \in k$ do
        $[Y_m]_{i=1,j=1}^{N,k_m} \sim \mathcal{N}(0,1)$;
    end
    $\rho[it] \leftarrow \texttt{max(InitializeMCFA(}Y_1, \ldots, Y_M\texttt{)}.\rho\texttt{)}$;
end
return $\texttt{mean(}\rho\texttt{)}$
```

Then we keep all components where $\rho_{init} > \bar{\rho}$.

1.4 Calculating the variance explained

The linear-Gaussian nature of the model simplifies estimation of the variance explained. That is, if the features of each mode $Y_m^{(:,j)}$ are normalized to variance 1, the model $Y_m^{(:,j)} = \sum_d W_m^{(j,d)} Z^{(:,d)} + \sum_{k_m} L_m^{(j,k_m)} X_m^{(:,k_m)} + \epsilon$ implies that the variance in feature j of mode m explained by shared factor d is $W_m^{(j,d)2}$. Likewise, the variance explained by the k_m-th private factor of mode m is $L_m^{(j,k_m)2}$. The total variance in mode m explained by a given shared factor d (respectively, private factor k_m) is thus given by $\sum_j W_m^{(j,d)2}$ (respectively, $\sum_j L_m^{(j,k_m)2}$), and the total variance in the mode explained by the factors are $\sum_{j,d} W_m^{(j,d)2}$ and $\sum_{j,k_m} L_m^{(j,k_m)2}$, respectively. Note that when working in PC-space, the raw W and L features correspond to variance in PCs explained, rather than modality features. Thus, we calculate the variance explained after projecting back into the original feature space $W_m \leftarrow V_m W_m, L_m \leftarrow V_m L_m$ where V_m are the right singular vectors of mode m.

To calculate the variance in a metadata feature explained by a particular space, we regressed the trait value T on the shared or private space, $T \sim Z$ or $T \sim X_m$. For continuous-valued traits we used linear regression as implemented in SciKitLearn v1.0 `linear_model.LinearRegression`[8] and report the coefficient of determination. For discrete-valued traits, we used multinomial logistic regression as implemented in SciKitLearn v1.0 `linear_model.LogisticRegression`[8]. We fit two models: a null model including only intercept or intercept and site, and one including the factor variables. We report the variance explained as the McFadden pseudo-R^2[9], $1 - \frac{ll_{\text{alt}}}{ll_{\text{null}}}$, with ll_{null} and ll_{alt} being the model negative log-likelihood for the null and alternative model respectively.

1.5 Calculating relative feature importance

Feature importance in traditional CCA is defined by the correlation of the variables in the reduced space $\rho = \text{cor}(Y_1 f_1, Y_2 f_2)$. Unfortunately this notion breaks down in higher dimensions. As we discuss further in the supplemental note, the degree of sharing in MCCA is defined by functions of the cross-correlation matrix in the reduced space,

$$S = \text{cor}(Y_1 f_1, \ldots, Y_m f_m) \in \mathbb{R}^{m \times m}.$$

We seek to define an analogous quantity for our graphical model. In MCFA, the data in the reduced (shared) space is given by the posterior mean of Z, $\hat{Z} = \mathbb{E}[Z|W, \Psi, L, Y] = Y(WW^\top + LL^\top\Psi)^{-1}W$. We can also calculate the posterior mean of Z conditional on observing a single mode, $\hat{Z}_m = \mathbb{E}[Z|W_m, \Psi_m, L_m, Y_m] = Y_m(W_m W_m^\top + L_m L_m^\top \Psi_m)^{-1} W_m$. This latter quantity is analogous to the reduced variables $Y_m f_m$ in MCCA.

Thus we can summarize the importance of each dimension of the shared space by calculating functions of the cross-correlation of columns of $\hat{Z}_m$,

$$S_d = \mathrm{cor}(\hat{Z}_1^{(:,d)}, \ldots, \hat{Z}_m^{(:,d)}).$$

As we show in the supplemental note, the relevant function in our model is the generalized variance $|S|$. The determinant of a correlation matrix is bounded between 0 and 1, with lower values indicating *more* correlation, and higher values *less*. Thus to aid interpretability, we report $\rho_d = -\log |S_d|$ and re-order columns of Z and W with decreasing ρ_d.

1.6 SNP set enrichment analysis

For SNP set enrichment analysis, we broadly follow the approach of CAMERA[10]. In brief, enrichment statistics can be inflated due to correlations in the sample - in this case, linkage disequilibrium between two GWAS SNPs. This results in an under-estimate of the standard error of the enrichment test statistic and an increase in false positives. We calculate the variance inflation factor[10] by using plink v1.9[11,12] to estimate linkage disequilibrium between annotation SNPs in $337,781$ unrelated individuals from the UK Biobank[13]. The variance inflation factor is $\nu = 1 + (p_A - 1)\bar{\rho}_A$, with $\bar{\rho}_A$ the average person correlation between features in set A. We test the known GWAS mean χ^2 statistic $h_0 : \bar{\chi}^2_A = 1$ against the alternative $h_1 : \bar{\chi}^2_A > 1$. The standard error of the test statistic is $\sigma_t = \sigma\sqrt{\frac{\nu}{p_A} - \frac{1}{p_m}}$ with σ the pooled empirical standard deviation of the test statistics.

1.7 Preprocessing of the MESA multi-omics pilot dataset

The Multi-Ethnic Study of Atherosclerosis (MESA) is a prospective cohort study with the goal to identify progression of subclinical atherosclerosis[14]. MESA recruited 6,814 participants, ages 45-84 years and free of clinical cardiovascular disease, during 2000-2002. The participants are 53% female, 38% non-Hispanic white, 28% Black, 22% Hispanic and 12% Asian-American. The Multi-Omics pilot dataset includes 30x whole genome sequencing (WGS) through the Trans-Omics for Precision Medicine (TOPMed) Project[15]. Blood samples for multi-omic analysis of participants were collected at two time points (exam 1 and exam 5) and assayed for transcriptomics (RNA-seq in PBMCs, monocytes and T cells), Illumina EPIC methylomics data (whole blood), targeted and untargeted metabolomics data (plasma), and proteomics data (plasma). The MESA Multi-Omics pilot biospecimen collection, molecular phenotype data production and quality control (QC) are described in detail in Aguet et al[16].

We analyzed individuals from Exam 1 where all five data types were collected and pass QC. All data modalities were inverse rank normalized prior to sample filtering based on the availability of other data types. There were 614 individuals with observations of WGS, RNA-seq, methylation, metabolomics and proteomics that all pass QC. We further removed all features (CpGs, genes, proteins) located on sex-chromosomes, 0-variance features, CpGs with missing data, and CpGs where the probe was within 5 bases of a SNP, leaving us with $6,042$ metabolites, $1,222$ proteins, $19,034$ genes, and $724,210$ CpGs. We analyzed 28 PCs of RNA expression, 39 PCs of methylation, 27 PCs of protein expression and 63 PCs of metabolite, as determined using the aforementioned method. For sample metadata, we leveraged the rich phenotype data available in MESA that were harmonized by the TOPMed Data Coordinating Center[17]. For details on the estimation of sample cell-type proportions from methylation and RNA-seq data, see Kasela et al[18].

1.8 Cross-validation

We used leave-one-out cross-validation (CV) to evaluate our model. The primary reason we chose leave-one-out CV over k-fold CV is that our hyperparameter selection method depends on the sample size. With $n-1$ individuals, the same parameters used for the full inference procedure are likely to be valid. For small k, fitting with $\frac{k-1}{k}n$ individuals while using the same number of PCs may result in over-fitting in the training set, and using a smaller number of PCs may not capture the same variation as the full model.

To perform cross-validation we hold out a set of individuals, fit the MCFA model, then project the held out individuals into the learned space. If W_{tr}, L_{tr} and Φ_{tr} are the model parameters learned from the training set, the projections of the test data into the learned spaces are given by

$$\hat{Z}_{te} = Y_{te}(W_{tr}W_{tr}^{\top} + L_{tr}L_{tr}^{\top}\Psi_{tr})^{-1}W_{tr}$$
$$\hat{X}_{te} = Y_{te}(W_{tr}W_{tr}^{\top} + L_{tr}L_{tr}^{\top}\Psi_{tr})^{-1}L_{tr}$$

The full data reconstruction is

$$\hat{Y}_{te} = \hat{Z}_{te}W_{tr}^{\top} + \hat{X}_{te}L_{tr}^{\top}$$

We evaluate model fit by calculating the normalized root mean squared error (NRMSE). In order to provide a fair evaluation across modes with a highly variable number of features, we calculate NRMSE on a per mode basis

$$NRMSE = \sqrt{\frac{1}{p_m}\sum_{i=1}^{p_m}\frac{\left(Y_m^{(:,i)} - \hat{Y}_m^{(:,i)}\right)^2}{\operatorname{var} Y_m^{(:,i)}}}$$

and potential over-fitting can be assessed by comparing the median training set NRMSE against the median test set NRMSE over many cross-validation iterations.

References

1. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39,** 1–22 (1977).
2. Parra, L. C. *Multiset Canonical Correlation Analysis simply explained* tech. rep. (). arXiv: 1802.03759v1.
3. Witten, D. M. & Tibshirani, R. J. *Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data ** tech. rep. 1 (2009).
4. Soneson, C., Lilljebjörn, H., Fioretos, T. & Fontes, M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics* **11,** 1–20 (Apr. 2010).
5. Asendorf, N. A. Informative Data Fusion: Beyond Canonical Correlation Analysis (2015).
6. Brown, B. C., Bray, N. L. & Pachter, L. Expression reflects population structure. *PLoS Genetics* **14,** e1007841 (Dec. 2018).
7. Marčenko, V. A. & Pastur, L. A. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of the USSR-Sbornik* **1,** 457–483 (1967).
8. Pedregosa, F. *et al. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot* tech. rep. (2011), 2825–2830.
9. McFadden, D. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics,* 105–142 (1973).
10. Wu, D. & Smyth, G. K. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research* **40,** 1–12 (2012).
11. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4,** 1–16. arXiv: 1410.4803 (2015).
12. Purcell, S. & Chang, C. *PLINK [1.9]*
13. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562,** 203–209 (Oct. 2018).

14. Bild, D. E. *et al.* Multi-Ethnic Study of Atherosclerosis: objectives and design. *American journal of epidemiology* **156,** 871–881 (Nov. 2002).

15. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature 2021 590:7845* **590,** 290–299 (Feb. 2021).

16. Aguet, F. & Lappalainen, T. Placeholder for FA paper (2022).

17. Stilp, A. M. *et al.* A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. eng. *American journal of epidemiology* **190,** 1977–1992 (Oct. 2021).

18. Kasela, S. *et al.* Interaction molecular QTL mapping discovers cellular and environmental modifiers of genetic regulatory effects. *Forthcoming* (2022).

2 Acknowledgements

BCB would like to thank Lior Pachter and Nicholas Bray for numerous insightful conversations about CCA over the years. Funding for DAK and BCB is provided by NIA U01AG068880. Funding for BCB is provided by NHGRI K99HG012373 and the Columbia Data Science Institute. Funding for TL and SK is provided by NHLBI R01HL142028. Funding for TL is provided by NIH R01AG057422 and NIMH R01MH106842. TL is a paid adviser or consultant of Variant Bio, GSK, Pfizer and Goldfinch Bio. Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)” (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1), and TOPMed MESA Multi-Omics (HHSN2682015000031/HSN26800004). The MESA projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for the Multi-Ethnic Study of Atherosclerosis (MESA) projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1TR001881, DK063491, and R01HL105756. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutes can be found at http://www.mesa-nhlbi.org. This publication was developed under a STAR research assistance agreement, No. RD831697 (MESA Air), awarded by the U.S Environmental protection Agency. It has not been formally reviewed by the EPA. The views expressed in this document are solely those of the authors and the EPA does not endorse any products or commercial services mentioned in this publication.

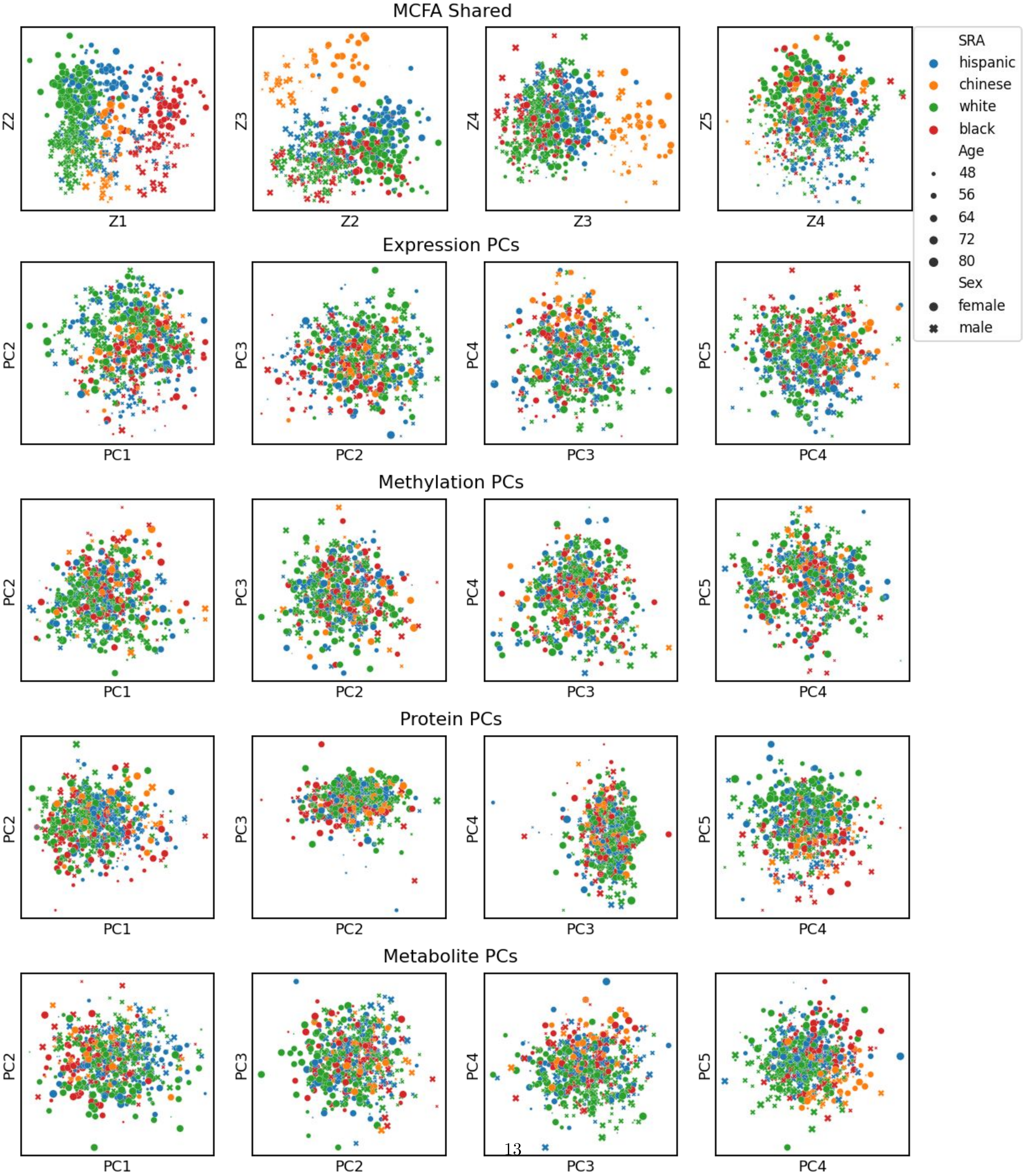


Figure S1: The top shared components learned from the MCFA model clearly reflect self-reported ancestry (SRA), age and sex, while none of the top PCs of any of the datasets show this structure.

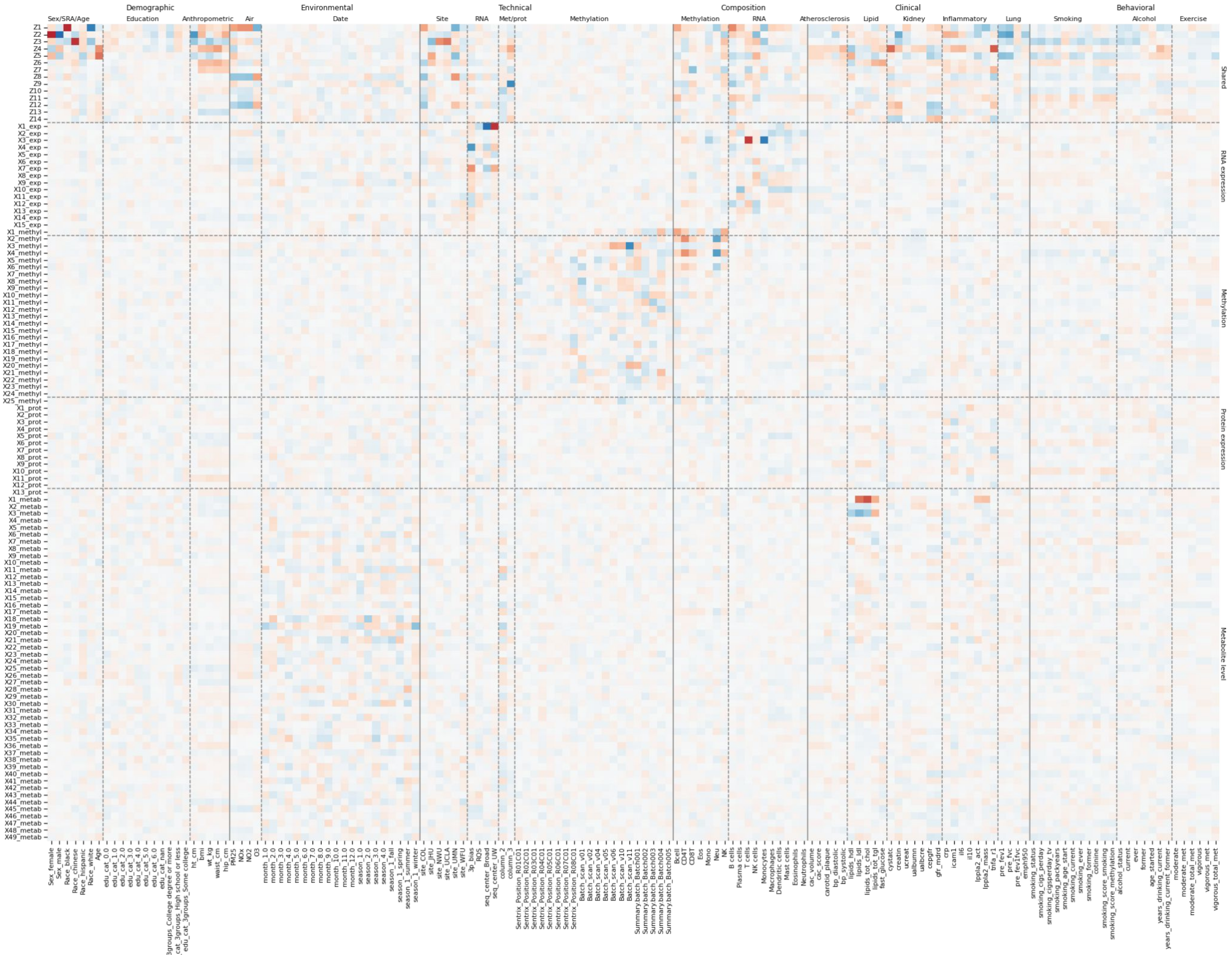


Figure S2: Correlation of each dimension of each learned space (rows) with each metadata factor (columns). Red values indicate positive correlation and blue values negative correlation.

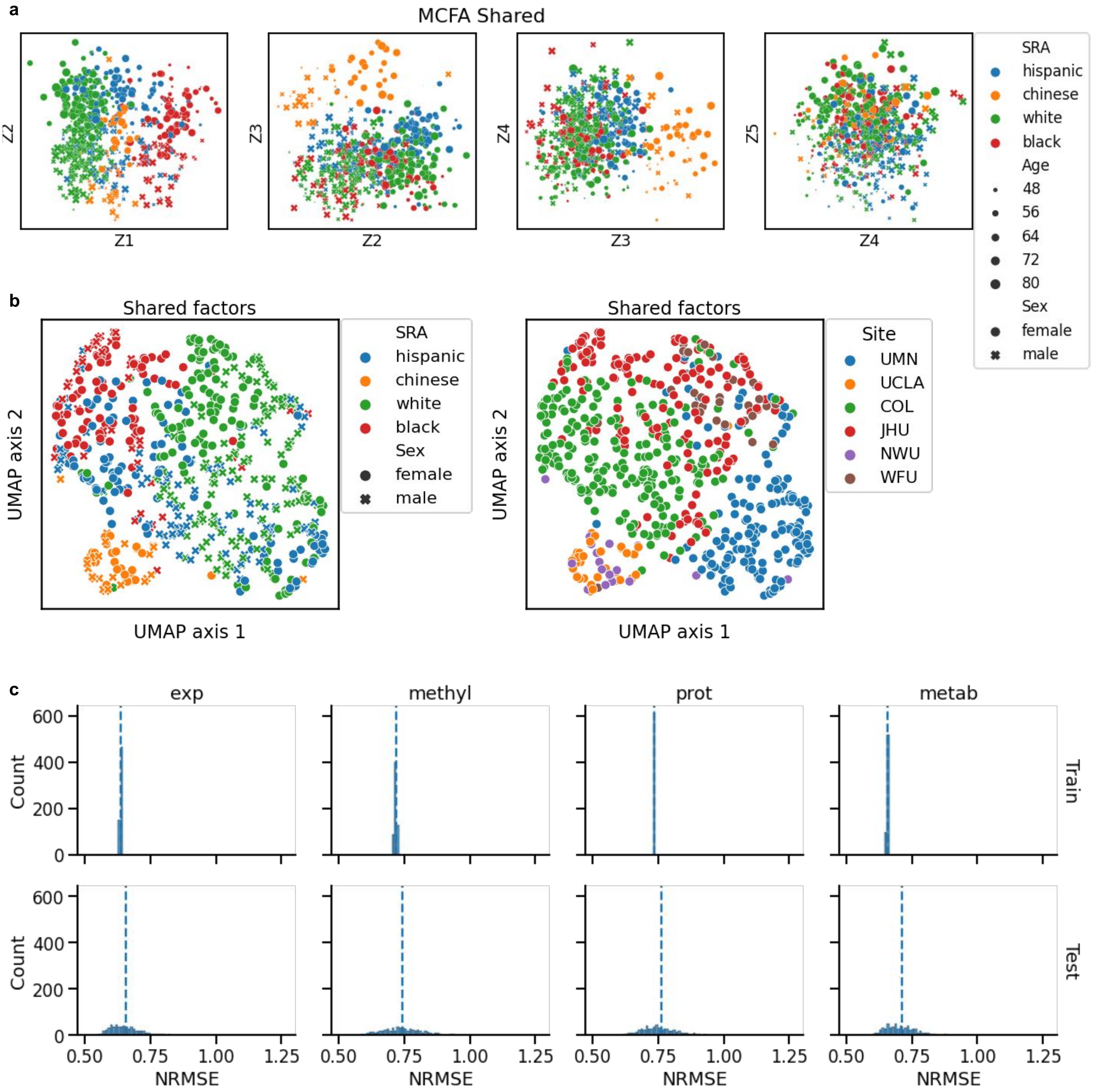


Figure S3: a) The top shared cross-validated MCFA components, annotated by self-reported ancestry (SRA), age and sex. Each point is creating by holding that individual out, fitting MCFA on the remaining individuals, then projecting the held-out individual into the shared space (see Online Methods). b) UMAP embeddings of the cross-validated MCFA components. c) Normalized root mean square error of the 613 training individuals for each dataset (top), versus the held-out individual (bottom), split by data type. Blue dashed line indicates the median.

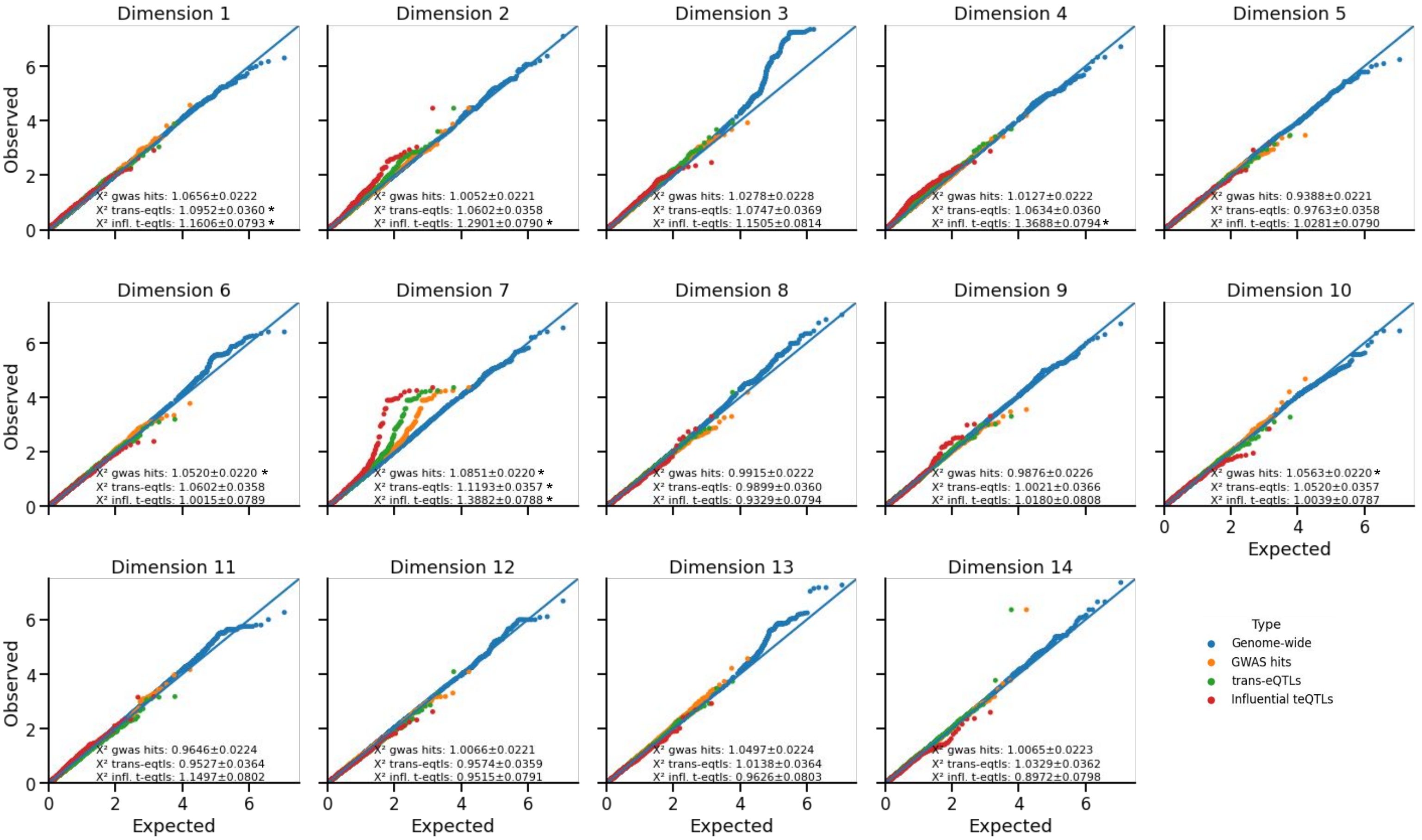


Figure S4: Q-Q plot of GWAS results for each shared factor.

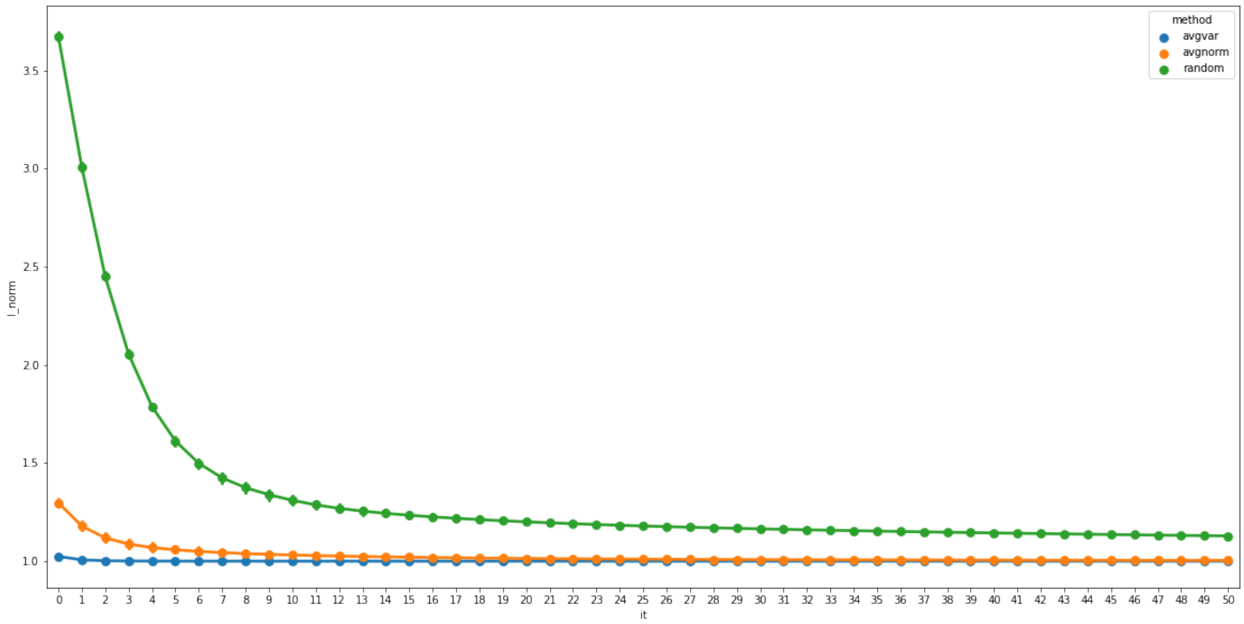


Figure S5: Average normalized model likelihood (y-axis, negative log-likelihood divided by minimum negative log-likelihood) as a function of EM step iteration. We simulated three datasets with $p = 30, 40, 50$ observed features generated by $k_m = 8, 11, 15$ private and $d = 10$ shared factors and $N = 1000$ individuals in 100 simulations. We compared random, SUMCORR-AVGVAR, and SUMCORR-AVGNORM model initializers. SUMCORR-AVGVAR produced good initial estimates resulting in fast convergence, while other approaches took longer to converge.

Supplemental Note

PCA, pPCA and FA

Principal components analysis[1] is a classic technique for dimensionality reduction. Assume we have N samples measured at p features each. Let y_n be a p-vector denoting the observations for sample n and let $Y = [y_1 : \dots : y_N]^\top$ be the corresponding $N \times p$ data matrix. For ease of notation, we assume throughout that each feature has mean 0 but note that this is not a requirement.

There are many ways to derive PCA, but perhaps the most common is to consider the problem of finding a unit projection vector v_i that maximizes the variance in the reduced space $T_i = Yv_i$. The first k principal axes $v_1, \dots, v_k$ are a sequence of orthonormal vectors which successively maximize the variance in the reduced space $T_i = Yv_i$. Let $\hat{\Sigma} = Y^\top Y/N$ be the empirical covariance matrix of Y. PCA solves the following problem:

$$\begin{aligned} &\texttt{max}_v v^\top \hat{\Sigma} v \\ &\text{s. t. } v^\top v = 1 \end{aligned}$$

The top-k principal axes are thus given by the eigenvectors of $\hat{\Sigma}$ that have the k highest eigenvalues. The data in PC space is thus given by the linear projection of the data into this space. Specifically, let $V_k = [v_1, \dots, v_k]$ be the projection matrix, and let $Y = U\Lambda V^\top$ be the singular value decomposition of Y. Notice that the eigenvectors of the covariance matrix $\hat{\Sigma}$ and the right singular values of the data matrix Y are the same. The points in PC-space are thus given by $T_k = YV_k = U_k\Lambda_k$ where U_k are the the left singular vectors with the k-highest singular values.

Tipping and Bishop[2] introduced a graphical model called probabilistic Principal Components Analysis that provides a generative framework for understanding PCA. The model is as follows:

$$\begin{aligned} x_n &\sim \mathcal{N}(0, I_k) \\ y_n &\sim \mathcal{N}(Lx_n, \sigma^2 I) \end{aligned}$$

where L is a $p \times k$ weight matrix and $\sigma^2 \geq 0$ is the residual noise. They show that the maximum likelihood estimate of the parameters W and σ^2 are given by

$$\begin{aligned} L_{ML} &= V_k(\Lambda_k^2 - \sigma^2 I_k)^{1/2} R \\ \sigma_{ML}^2 &= \frac{1}{p-d} \sum_{j=k+1}^{d} \Lambda_j^2 \end{aligned}$$

where R is an arbitrary $k \times k$ orthogonal rotation matrix. Thus, as $\sigma^2 \to 0$, W represents an orthogonal projection into standard PC space. This defines an equivalence relationship between pPCA and PCA.

Factor analysis is a very similar model, with the only difference being the form of the noise term. Rather than force an isotropic noise model $\sigma^2 I$, factor analysis allows for an arbitrary diagonal positive semi-definite matrix $\Psi = \text{diag}(\psi_1, \dots, \psi_p) \succeq 0$. This allows each observed feature to have it's own error variance.

CCA and pCCA

Now assume each sample is measured on two different sets of conceptually distinct features y_n^1 and y_n^2 with corresponding $N \times p_1$ and $N \times p_2$ data matrices Y_1 and Y_2. As before let $\hat{\Sigma}_{11} = Y_1^\top Y_1/N$ and $\hat{\Sigma}_{22} = Y_2^\top Y_2/N$ be the empirical covariance matrices for modalities 1 and 2, and let $\hat{\Sigma}_{12} = Y_1^\top Y_2/N$ be empirical cross-covariance matrix between the features in each mode. The first set of canonical vectors f_1, f_2 are those that maximize the correlation

$$\text{cor}(Y_1 f_1, Y_2 f_2) = \frac{f_1^\top \hat{\Sigma}_{12} f_2}{\sqrt{f_1^\top \hat{\Sigma}_{11} f_1}\sqrt{f_2^\top \hat{\Sigma}_{22} f_2}}$$

Note that, similarly to PCA, this definition reveals that CCA is a constrained optimization problem:

$$\begin{aligned}&\texttt{max}_{f_1,f_2} f_1^\top \hat{\Sigma}_{12} f_2\\ \text{s.t. } & f_1^\top \hat{\Sigma}_{11} f_1 = f_2^\top \hat{\Sigma}_{22} f_2 = 1\end{aligned}$$

but note that rather than the unit norm constraint $f^\top f = 1$ used in PCA, we have a unit variance constraint $f^\top \hat{\Sigma} f = 1$. This unit variance constraint allows there to be correlation within the features of a dataset that is not explained by correlation between the features across datasets.

Successive components can be found by projecting out the first canonical component and again maximizing the correlation of the residuals[3]. Equivalently, all components can be found by solving an eigenvalue problem. To see this consider the change of variables $g_1 = V_1 \Lambda_1^{-1} f_1$ and $g_2 = V_2 \Lambda_2^{-1} f_2$. The correlation is now given by:

$$\operatorname{cor}(Y_1 f_1, Y_2 f_2) = \frac{g_1^\top U_1^\top U_2 g_2}{\sqrt{g_1^\top g_1}\sqrt{g_2^\top g_2}}$$

which indicates that g_1, g_2 are the top pair of left-right singular vectors of the matrix $U_1^\top U_2$. Further components are further singular vectors of $U_1^\top U_2$, and it's singular values are the correlations. This also reveals that CCA is equivalent to using PCA to whiten the variables of each data matrix, concatenating them, and then performing PCA again on the whitened, concatenated data matrix.

Likewise to probabilistic PCA, probabilistic CCA is a graphical model that provides a generative framework for thinking about CCA[4]. The model is as follows:

$$\begin{aligned}z_n &\sim \mathcal{N}(0, I_d)\\ y_n^1 &\sim \mathcal{N}(W_1 x_n, \Psi_1)\\ y_n^2 &\sim \mathcal{N}(W_2 x_n, \Psi_2)\end{aligned}$$

similarly to FA and pPCA, we sample a d-dimensional random normal hidden vector, pass it through a weight matrix, and add random noise. We have two weight matrices W_1 and W_2 of shape $p_1 \times d$ and $p_2 \times d$, and two noise matrices Ψ_1 and Ψ_2, however in this case these noise matrices are arbitrary positive semi-definite matrices ($\Psi_\bullet \succeq 0$). Bach and Jordan show that the maximum likelihood estimate of the parameters of pCCA can be determined from the CCA solution:

$$\begin{aligned}W_{1,ML} &= \hat{\Sigma}_{11} F_{1d} M_1\\ W_{2,ML} &= \hat{\Sigma}_{22} F_{2d} M_2\\ \Psi_{1,ML} &= \hat{\Sigma}_{11} - W_{1,ML} W_{1,ML}^\top\\ \Psi_{2,ML} &= \hat{\Sigma}_{22} - W_{2,ML} W_{2,ML}^\top\end{aligned}$$

where $F_{\bullet d} = [f_{\bullet 1}; \ldots; f_{\bullet d}]$ are the first d canonical directions and M_1, M_2 are arbitrary matrices with spectral norm less than 1 such that $M_1 M_2 = \rho_d$.

Multi-set canonical correlation analysis

Now rather than having two sets of conceptually distinct features for each sample, assume we have M different conceptually distinct sets of features $\{y_n^m\}$ with corresponding $N \times p_m$ data matrices $\{Y_m\}$. In MCCA, we are still interested in finding projection vectors $\{f_m\}$ which map our high dimensional data into a one-dimensional space, however there are many formulations that are equivalent to classical CCA with two datasets. Let $\hat{\Sigma}_{kl} = Y_k^\top Y_l / N$ be the empirical cross-covariance matrix between the features in dataset k and dataset l. The covariance of the data in the reduced space is given by

$$S = \begin{bmatrix} f_1^\top \hat{\Sigma}_{11} f_1 & \ldots & f_1^\top \hat{\Sigma}_{1M} f_M \\ \vdots & \ddots & \vdots \\ f_M^\top \hat{\Sigma}_{M1} f_1 & \ldots & f_M^\top \hat{\Sigma}_{MM} f_M \end{bmatrix}$$

The various formulations of MCCA correspond to optimizing different objective functions $J(S)$ subject to the a constraint function $h(f, \hat{\Sigma})$[5,6]. In brief, possible objective functions include:

- SUMCOR: Maximize the sum of pairwise correlations: $J = \sum_{i,j} f_i^\top \hat{\Sigma}_{i,j} f_j = 1^\top S 1$
- SUMSQCOR: Maximize the sum of squares of pairwise correlations: $J = \sum_{i,j} (f_i^\top \hat{\Sigma}_{i,j} f_j)^2 = ||S||_F^2$
- MAVAR: Maximize the largest eigenvalue of S: $J = \lambda_1(S)$
- MINVAR: Minimize the smallest eigenvalue of S: $J = \lambda_d(S)$
- GENVAR: Minimize the determinant of S, also known as the generalized variance: $J = |S| = \prod_i \lambda_i(S)$

while possible constraints include:

- VAR: The canonical directions each have unit variance: $h : \forall_i f_i^\top \hat{\Sigma}_{ii} f_i = 1$
- AVGVAR: The canonical directions have unit variance on average: $h : \sum_i f_i^\top \hat{\Sigma}_{ii} f_i = d$
- NORM: The canonical directions each have unit norm: $h : \forall_i f_i^\top f_i = 1$
- AVGNORM: The canonical directions have unit norm on average: $h : \sum_i f_i^\top f_i = d$

It is straightforward to see that any of the 5 listed objective functions could be combined with either of the first two constraints to create 10 optimization problems that are equivalent to CCA in the two-dataset case. The final two constraints correspond to relaxations of the unit variance constraint which can reveal a simpler optimization problem in some cases, and which does not suffer from being trivially satisfiable when $p > N$.

Some of these can be fit by solving eigenvalue problems, while others require more complicated iterative methods. Of particular note are the SUMCOR and GENVAR objectives. The GENVAR objective was the first considered MCCA approach[7], where a simple solution for the $M = 3$ and $p_1 = p_2 = p_3 = 2$ case is given. GENVAR is a particularly natural way of thinking about MCCA - it is a single value that represents the multidimesnioal scatter of points in space[8]. Smaller values of the generalized variance indicate less scatter, and thus higher "correlation" of the points in the reduced space. Despite this, is has received relatively little attention as a method for MCCA, perhaps because it is challenging to fit[6]. On the other hand, most attention has been focused on the SUMCOR objective[9], which can be solved easily with the AVGVAR and AVGNORM constraints. SUMCOR with AVGVAR constraint can be solved via a simple two-stage procedure: first whiten each data matrix, concatenate the whitened features, and then perform PCA on the whitened, concatenated features. SUMCOR with AVGNORM constraint is even simpler to solve: simply perform PCA on the concatenated feature set. In the two dataset case, this latter method is sometimes called "diagonal CCA" and forms the basis of the original integration approach used in Seurat[10] as well as many sparse CCA approaches[11]. This is also closely related to group factor analysis approaches for multi-modal data, see for example[12] and references therein. The equivalence to PCA on the concatenated feature set makes it straightforward to see that the NORM-based constraints involve an implicit assumption that shared factors are responsible for both covariation across features in different modes and between features within a mode.

Probabilistic graphical model for multi-set CCA

Here, we describe a probabilistic graphical model for multi-set CCA (pMCCA). Note that while this model is an option in our software package, we derive and discuss it primarily to draw connection to traditional multiset CCA. The full model which includes simultaneous factor analysis of the private spaces is described in the next section. Unlike traditional CCA, pCCA has one single obvious generalization to multiple datasets:

$$z_n \sim \mathcal{N}(0, I_d) \tag{1}$$

$$y_n^m \sim \mathcal{N}(W_m z_n, \Psi_m) \tag{2}$$

where again we have weight matrices W_m of shape $p_m \times d$ and arbitrary positive semi-definite noise matrices $\Psi_m \succeq 0$. This model is illustrated as a plate diagram in Figure S6.

We will now see that pMCCA is related to MCCA in the following ways:

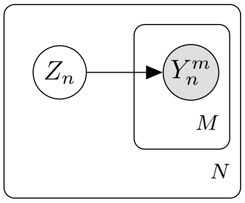


Figure S6: Probabilistic multi-set canonical correlation analysis (pMCCA). For each individual n (the outer plate), we sample z_n from the d-dimensional unit Gaussian, $z_n \sim \mathcal{N}(0, I_d)$. For each dataset m (the inner plate) we sample the features of that dataset from a p_m-dimensional unit Gaussian, $y_n^m \sim \mathcal{N}(W_m z_n, \Psi_m)$. Here $W_m \in \mathbb{R}^{p_m \times d}$ is the transformation weight matrix and $\Psi_m \in \mathbb{R}^{p_m \times p_m} \succeq 0$ is the residual covariance matrix.

Observation 1. *The maximum likelihood solution to the pMCCA model corresponds to the GENVAR objective with VAR constraint in the $M = 3$ dataset case.*

Observation 2. *The maximum likelihood solution to the pMCCA model does not correspond to any of the listed MCCA formulations in the $M \geq 4$ case.*

Let $W = [W_1^\top : \dots : W_M^\top]^\top$ be the stacked weight matrices and $\Psi = \text{diag}(\Psi_1, \dots, \Psi_M)$ be the block diagonal covariance matrix. The model covariance is given by:

$$\Sigma = \begin{bmatrix} W_1 W_1^\top + \Psi_1 & W_1 W_2^\top & \dots & W_1 W_M^\top \\ W_2 W_1^\top & W_2 W_2^\top + \Psi_2 & \dots & W_2 W_M^\top \\ \vdots & \vdots & \ddots & \vdots \\ W_M W_1^\top & W_M W_2^\top & \dots & W_M W_M^\top + \Psi_M \end{bmatrix} = WW^\top + \Psi \tag{3}$$

Let $Y = [Y_1 : \dots : Y_M]$ so that the empirical covariance matrix can be written $\hat{\Sigma} = Y^\top Y / N$. The model negative log-likelihood is given by:

$$l(\Sigma | \hat{\Sigma}) = \frac{Np}{2} \log 2\pi + \frac{N}{2} \log |\Sigma| + \frac{N}{2} \text{Tr}(\Sigma^{-1} \hat{\Sigma}) \tag{4}$$

where $p = \sum_m p_m$ is the total number of features from all datasets. It is straightforward to see that many arguments from Bach and Jordan[4] carry over to the multi-set case, thus we refer readers there for proofs. In particular we have

Lemma 1. *At a stationary point of the likelihood $\Sigma_{mm} = W_m W_m^\top + \Psi_m = \hat{\Sigma}_{mm}$.*

Thus, at a stationary point

$$\text{Tr}\,(\Sigma^{-1} \hat{\Sigma}) = \text{Tr} \left(\begin{bmatrix} \hat{\Sigma}_{11} & \dots & W_1 W_M^\top \\ \vdots & \ddots & \vdots \\ W_M W_1^\top & \dots & \hat{\Sigma}_{MM} \end{bmatrix}^{-1} \hat{\Sigma} \right) = p \tag{5}$$

so that the models minimum negative log-likelihood is proportional to the log generalized variance of the model:

$$l(\Sigma | \hat{\Sigma}) \propto \log |\Sigma| \tag{6}$$

Moreover, Lemma 1 allows us to further factorize Σ

$$\Sigma = \begin{bmatrix} \hat{\Sigma}_{11}^{1/2} & 0 & \dots & 0 \\ 0 & \hat{\Sigma}_{22}^{1/2} & \dots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\Sigma}_{MM}^{1/2} \end{bmatrix} \begin{bmatrix} I & \tilde{W}_1 \tilde{W}_2^\top & \dots & \tilde{W}_1 \tilde{W}_M^\top \\ \tilde{W}_2 \tilde{W}_1^\top & I & \dots & \tilde{W}_2 \tilde{W}_M^\top \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{W}_M \tilde{W}_1^\top & \tilde{W}_M \tilde{W}_2^\top & \dots & I \end{bmatrix} \begin{bmatrix} \hat{\Sigma}_{11}^{1/2} & 0 & \dots & 0 \\ 0 & \hat{\Sigma}_{22}^{1/2} & \dots 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\Sigma}_{MM}^{1/2} \end{bmatrix} \tag{7}$$

and so

$$l(\Sigma|\hat{\Sigma}) \propto \begin{vmatrix} I & \tilde{W}_1\tilde{W}_2^\top & \dots & \tilde{W}_1\tilde{W}_M^\top \\ \tilde{W}_2\tilde{W}_1^\top & I & \dots & \tilde{W}_2\tilde{W}_M^\top \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{W}_M\tilde{W}_1^\top & \tilde{W}_M\tilde{W}_2^\top & \dots & I \end{vmatrix} \tag{8}$$

where $\tilde{W}_m = \hat{\Sigma}_{mm}^{-1/2}W_m$. Notice that the off-diagonal blocks are the cross-covariance matrices in the model of individually-whitened datasets. If there exists a set of projection vectors f_m such that $f_m^\top\hat{\Sigma}_{mm}f_m = 1$ and the projection of the data into the space spanned by f_m has covariance equal to the above, then minimizing the above determinant solves the GENVAR MCCA objective with VAR constraint. Let $\tilde{Y}_m = Y_m\hat{\Sigma}_{mm}^{-1/2}$ be the whitened datasets and let $g_m = \hat{\Sigma}_{mm}^{1/2}f_m$ be the change of variables that gives $g_m^\top g_m = 1$. The projection in MCCA space is given by $\tilde{Y}_{||}^m = \tilde{Y}^m g_m g_m^\top$. We seek g_m such that

$$\begin{aligned} \tilde{Y}_{||}^{k\top}\tilde{Y}_{||}^l &= g_k g_k^\top \tilde{Y}^{k\top}\tilde{Y}^l g_l g_l^\top & (9)\\ &= g_k g_k^\top \hat{\Sigma}_{kk}^{-1/2}\hat{\Sigma}_{kl}\hat{\Sigma}_{ll}^{-1/2} g_l g_l^\top & (10)\\ &= g_k f_k^\top \hat{\Sigma}_{kl} f_l g_l^\top & (11)\\ &= \tilde{W}_k\tilde{W}_l^\top & (12) \end{aligned}$$

Thus we must satisfy

$$g_k^\top \tilde{W}_k\tilde{W}_l^\top g_l = f_k^\top\hat{\Sigma}_{kl}f_l \tag{13}$$

Notice that for $d = 1$, the left and right side are scalars. This means we can express our necessary criterion as

$$q_k q_l = c_{kl} \tag{14}$$

For $M = 3$, this results in a system of 3 equations in 3 unknowns, which has solutions of the form $q_1 = \sqrt{\frac{c_{12}c_{13}}{c_{23}}}$. Note that these can be found by setting $g_k = \tilde{W}_k/||\tilde{W}_k||_2$ which yields $q_k = ||\tilde{W}_k||_2$. For $M > 3$, there are more equations than unknowns and they cannot be mutually satisfied in general. Note also that a similar argument can be used in the $d > 1$ case to show that this system is not satisfiable even for $M = 3$. Thus, unlike CCA and pCCA, fitting $d > 1$ components is not equivalent to iteratively fitting single components and projecting them out.

Multiset Correlation and Factor Analysis

The residual covariance matrices Ψ_d deserve additional attention. Put simply, these matrices represent the residual structure in each modality after accounting for shared structure across modes. Instead of allowing this matrix to be arbitrary, we can instead think of this matrix as having some additional structure. For example, Ψ_d might be the sum of a low rank and an isotropic covariance matrix. This suggests that we add an additional latent variable to each dataset, corresponding to a factor model for the "private" structure (e.g. not shared with other datasets). Specifically, we modify pMCCA such that for each dataset, we additionally sample a latent variable from a k_m-dimensional unit Gaussian. The observed data are then sampled from a multi-variate Gaussian where the mean is a linear combination of both private variables, but the residual covariance matrix is now diagonal.

$$\begin{aligned} z_n &\sim \mathcal{N}(0, I_d) & (15)\\ x_n^m &\sim \mathcal{N}(0, I_{k_m}) & (16)\\ y_n^m &\sim \mathcal{N}(W_m z_n + L_m x_n^m, \Psi_m) & (17) \end{aligned}$$

where L_m are the $k_m \times p_m$ private space loading matrices and $\Psi_m = \mathrm{diag}(\psi_m^1, \dots, \psi_m^{p_m})$ are the diagonal residual covariance matrices. Note that in general we allow the entries on the diagonal of Ψ_m to take different

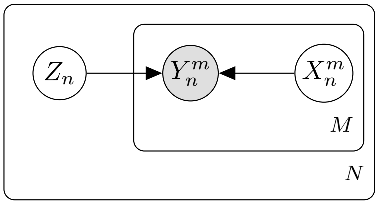


Figure S7: Multset correlation and factor analysis as a plate diagram. For each individual n (the outer plate), we sample z_n from the d-dimensional unit Gaussian, $z_n \sim \mathcal{N}(0, I_d)$. For each mode m (the inner plate) and individual n we sample x_n^m from a k_m-dimensional unit Gaussian. The observed features of that mode are then sampled from a p_m-dimensional unit Gaussian, $y_n^m \sim \mathcal{N}(W_m z_n + L_m x_n^m, \Psi_m)$. Here $W_m \in \mathbb{R}^{p_m \times d}$ is the transformation weight matrix, L_m is the private-space loadings matrix, and $\Psi_m = \mathrm{diag}(\Psi_m^1, \ldots, \Psi_m^{p_m})$ is the diagonal residual covariance matrix.

values, similarly to factor analysis. One could additionally constrain the entries of the diagonal to be the same, $\Psi = \sigma^2 I_p$, similar to pPCA.

This model can be fit via a straightforward application of the expectation-maximization (EM) algorithm[13]. We first derive conditional expectation of the log-likelihood, $\mathcal{L}$ for the model under the generative process specified in Figure S7. For convenience, let

$$y_n = [y_n^{1\top} : \ldots : y_n^{m\top}]^\top \in \mathbb{R}^p \tag{18}$$

$$x_n = [x_n^{1\top} : \ldots : x_n^{m\top}]^\top \in \mathbb{R}^k \tag{19}$$

$$W = [W_1^\top : \ldots : W_M^\top]^\top \in \mathbb{R}^{p \times d} \tag{20}$$

$$L = \mathrm{diag}(L_1, \ldots, L_m) \in \mathbb{R}^{p \times k} \tag{21}$$

$$\Psi = \mathrm{diag}(\Psi_1, \ldots, \Psi_M) \in \mathbb{R}^{p \times p} \tag{22}$$

where $k = \sum_m k_m$. At a given time step t during the computation of the EM algorithm, let the conditional expectation for given latent variables z_i and x_i be $\mathbb{E}[\cdot | W_t, \Psi_t, L_t, y_i] = \langle \cdot \rangle$.

The conditional expectation of the log-likelihood (E-step) is:

$$\langle \mathcal{L} \rangle = -\sum_{i=1}^{N} \tilde{C} + \frac{1}{2} \ln |\Psi| + \frac{1}{2} \mathrm{Tr}\left(\Psi^{-1} y_i y_i^\top\right) + \frac{1}{2} \mathrm{Tr}\left(L^\top \Psi^{-1} L \left\langle x_i x_i^\top \right\rangle\right) \tag{23}$$

$$+ \frac{1}{2} \mathrm{Tr}\left(W^\top \Psi^{-1} W \left\langle z_i z_i^\top \right\rangle\right) + \mathrm{Tr}\left(L^\top \Psi^{-1} W \left\langle z_i x_i^\top \right\rangle\right) - y_i^\top \Psi^{-1} W \langle z_i \rangle \tag{24}$$

$$- y_i^\top \Psi^{-1} L x_i + \frac{1}{2} \mathrm{Tr} \left\langle x_i x_i^\top \right\rangle + \frac{1}{2} \mathrm{Tr} \left\langle z_i z_i^\top \right\rangle \tag{25}$$

At a given timestep t, we compute the update of parameters $t+1$ by differentiating $\mathcal{L}$ with respect to W_t, L_t, and Ψ_t, and setting the derivative of the corresponding expected log-likelihood to 0. The following update steps are derived using standard matrix differentiation results[14].

$$W_{t+1} = \left(\sum_{i=1}^{N} y_i \left\langle z_i^\top \right\rangle - L_t \left\langle x_i z_i^\top \right\rangle \right) \left(\sum_{i=1}^{N} \left\langle z_i z_i^\top \right\rangle \right)^{-1} \tag{26}$$

$$L_{t+1} = \left(\sum_{i=1}^{N} y_i \left\langle x_i^\top \right\rangle - W_t \left\langle z_i x_i^\top \right\rangle \right) \left(\sum_{i=1}^{N} \left\langle x_i x_i^\top \right\rangle \right)^{-1} \Psi_{t+1} \tag{27}$$

$$= \frac{1}{N} \sum_{i=1}^{N} y_i y_i^\top + L \left\langle x_i x_i^\top \right\rangle L^\top + W \left\langle z_i z_i^\top \right\rangle W^\top + 2L \left\langle x_i z_i^\top \right\rangle W^\top - 2 y_i \left\langle z_i^\top \right\rangle W^\top - 2 y_i \left\langle x_i^\top \right\rangle L^\top \tag{28}$$

References

1. Pearson, K. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2,** 559–572 (1901).
2. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **61,** 611–622 (1999).
3. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28,** 321–377 (May 1936).
4. Bach, F. R. & Jordan, M. I. A probabilistic interpretation of canonical correlation analysis (2005).
5. Kettenring, J. R. Canonical analysis of several sets of variables. *Biometrika* **58,** 433–451 (1971).
6. Asendorf, N. A. Informative Data Fusion: Beyond Canonical Correlation Analysis (2015).
7. Steel, R. G. D. Minimum Generalized Variance for a set of Linear Functions. *https://doi.org/10.1214/aoms/1177729594* **22,** 456–460 (Sept. 1951).
8. Kocherlakota, S. & Kocherlakota, K. Generalized Variance. *Encyclopedia of Statistical Sciences* (Oct. 2004).
9. Parra, L. C. *Multiset Canonical Correlation Analysis simply explained* tech. rep. (). arXiv: `1802.03759v1`.
10. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology 2018 36:5* **36,** 411–420 (Apr. 2018).
11. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10,** 515–534 (July 2009).
12. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology* **14** (June 2018).
13. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* **39,** 1–22 (1977).
14. Petersen, K. B., Pedersen, M. S., *et al.* The matrix cookbook. *Technical University of Denmark* **7,** 510 (2008).