Published in final edited form as:
Nat Comput Sci. 2021 February ; 1(2): 143–152. doi:10.1038/s43588-021-00029-8.

Similarity-driven multi-view embeddings from high-dimensional biomedical data

Brian B. Avants, PhD, **Nicholas J. Tustison, DSc**, **James R. Stone, MD, PhD**
Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA

Abstract

Diverse, high-dimensional modalities collected in large cohorts present new opportunities for the formulation and testing of integrative scientific hypotheses. Similarity-driven multi-view linear reconstruction (SiMLR) is an algorithm that exploits inter-modality relationships to transform large scientific datasets into smaller, more well-powered and interpretable low-dimensional spaces. SiMLR contributes an objective function for identifying joint signal, regularization based on sparse matrices representing prior within-modality relationships and an implementation that permits application to joint reduction of large data matrices. We demonstrate that SiMLR outperforms closely related methods on supervised learning problems in simulation data, a multi-omics cancer survival prediction dataset and multiple modality neuroimaging datasets. Taken together, this collection of results shows that SiMLR may be applied to joint signal estimation from disparate modalities and may yield practically useful results in a variety of application domains.

Keywords

code:R; multi-modality embedding; brain; ANTs; ANTsR; genotype; depression; SiMLR; imaging genetics

1 Introduction

Healthcare – from both a prevention as well as treatment perspective – is increasingly turning to large, mixed datasets to gain a better understanding of the biological complexity that influences sensitivity or resistance to disease or injury. These studies promise new insights into disease etiology by collecting several related and complementary measurements on each subject of interest, i.e. by collecting multi-view data. In more common conditions, like Alzheimer's disease, multi-view datasets are motivated by the need to understand the

Corresponding author: Brian B. Avants, PhD, Department of Radiology and Medical Imaging, University of Virginia, 480 Ray C Hunt Drive, Box 801339, Charlottesville, VA 22903, 434-924-9585, stnava@gmail.com.

5 Author Contributions
B.B.A. and N.J.T. and J.R.S made substantial contributions to the conception and design of the work and the analysis and interpretation of data; all authors drafted the work and revised it.
B.B.A. and N.J.T. created the software used in the work.

6 Competing Interests Statement
The authors declare no competing interests.

diversity of the disease process, identify sub-groups and thereby advance personalized treatment approaches. Multi-view data can also reveal key features that drive variability within the “normal” phenotype e.g. underlying factors that contribute to difference in neurobiological age[1] or the genetic architecture of quantitative phenotypes as mediated through brain structure[2].

Multi-view (also known as multiple modality or multi-block) datasets are increasingly common in the biomedical sciences. In the idealized case, each view / modality will provide a completely unique measurement of the substrate biology. However, it is perhaps more common that each view provides a partial and not wholly independent perspective on a complex phenomenon. In this case, covariation can be exploited in order to sift through noisy measurements and better identify meaningful signal. Moreover, joint relationships across systems of the brain or across scale can form the foundation for integrative scientific hypotheses.

Pre-specified joint hypotheses allow the scientist to avoid a combinatorial explosion of tests for possible interactions. Although powerful in sufficiently large, well-understood datasets, prior multivariate hypotheses can be difficult to enumerate with sufficient detail to support implementation and testing. Fully multivariate and data-driven dimensionality reduction models provide an alternative including principal component analysis (PCA)[3,4] and independent component analysis (ICA)[5–7]. However, these popular models, applied directly, are not explicitly designed for interpretation across multiple modalities and do not provide an easy way for the scientist to regularize the solution with prior knowledge or to visualize the feature vectors which are both dense and signed (i.e. have both positive and negative weights).

Graph-regularized, imaging-focused dimensionality reduction methods emerged in recent years to address the desire for interpretable components[8–10]. Graph-net[10], similar to SCCAN[11,12], uses ℓ_1 regularization to constrain embedding vectors to be sparse and reduce over-fitting in high-dimensional problems. Relatedly, graph-regularization has been used to improve prediction in imaging genetics[10,13,14] and may be combined with canonical correlation analysis (as in SCCAN[11]). Non-negative factorization methods provide a second degree of interpretability by guaranteeing that factorizations are unsigned and, therefore, these methods allow components to be interpreted in terms of their original units (e.g. millimeters)[15,16]. Other efforts[9,17] use prior constraints to guide solutions toward familiar sparsity patterns. More generally, regularization is also critical to well-posedness[18,19].

The need for joint, interpretable modeling of several (>2) parallel but heterogenous datatypes is rapidly increasing[20–24]. Multi-block data analysis methods such as Kettering’s five offerings[25] and more recent regularized generalized canonical correlation analysis (RGCCA and its sparse variant SGCCA)[26–28] and multiway generalized canonical correlation analysis[29] extend Hotelling’s classical CCA[30,31] to multi-view (viz. multi-block) data. Joint and individual variation explained (JIVE) is another framework dedicated to data fusion[32,33] along with MultiLevel Simultaneous Component Analysis (MLSCA)[34] and Multi-Omics Factor Analysis (MOFA)[35]. A variation of JIVE, applied to convolutional network features, has also been applied to imaging genetics problems[36].

Our contribution, similarity-driven multi-view linear reconstruction (SiMLR), is a joint embedding method—targeting biomedical data—that links several of the ideas expressed in prior work. SiMLR builds on sparse canonical correlation analysis for neuroimaging (SCCAN)[12,37,38] and prior-based eigenanatomy[17,39]. SiMLR goes beyond SCCAN in that it takes two or more modalities as input, allows customized regularization models and uses a fast and memory efficient implementation appropriate for large datasets. SiMLR outputs locally optimal low-dimensional matrix embeddings for each modality that best predict its partner modalities. SiMLR achieves this by reconstructing each modality matrix from a basis set derived from the partner modalities. One important contribution of SiMLR is that the “linking” subspace is computed via a source separation algorithm, e.g. singular value decomposition (SVD) or ICA. This sub-algorithm seeks to identify latent signal sources that span modalities. The basis set can be forced to be either orthogonal (SVD) or statistically independent (ICA) where the latter option may be more appropriate for unmixing signal sources in real world data[40,41]. Simultaneously, the feature vectors may be constrained by graph-regularized sparsity and non-negativity. Furthermore, the target energy (measuring the similarity between different modalities) is also flexible and builds on classical objective functions in SVD and CCA. SiMLR is the only available framework that combines these features in an accessible and flexible joint dimensionality reduction algorithm. Although SiMLR supports path modeling, only the leave-one-modality-out approach is explored in this work.

2 Results

Figure 1 shows a general overview of how SiMLR is applied within the context of scientific data. Each evaluation below fits within this general framework. Furthermore, each study uses joint dimensionality reduction in conjunction with regression-based supervised learning in a training and testing paradigm. Table 1 summarizes the overall findings that are presented in this results section.

2.1 Simulated data

SiMLR seeks to solve a multiple modality version of the cocktail party problem[43] where the hidden source signals are distributed across each modality. Therefore, SiMLR assumes that this common latent signal exists across modalities and may be found by linear projections into a low-dimensional space. Details of the generative data simulation approach can be found in section 7.7.

The evaluation criterion then compares the ability of SiMLR to recapture the known basis with respect to: (a) regularized generalized canonical correlation analysis (RGCCA); (b) sparse generalized canonical correlation analysis (SGCCA). The primary evaluation criterion – accuracy in predicting the true latent signal – exhibits that SiMLR’s use of cross-modality information and regularization drives the solution closer to the ground truth basis in comparison to the other methods. Secondarily, we demonstrate improved robustness to data corruption.

2.1.1 Signal recovery—An overview of the results is in Figure 2. In this figure, higher scores are better and points above the diagonal dotted line show superior SiMLR performance in pair-wise fashion. Panel (a) shows that SiMLR-CCA-ICA outperforms RGCCA and SGCCA. Panel (b) shows that SiMLR-CCA-SVD achieved the best overall performance with an average R squared recovery of 0.51 while SGCCA and RGCCA score 0.45 and 0.35 respectively. Panel (c) shows that SiMLR-regression-ICA does nearly as well as the other SiMLR variants. Statistical details are available in supplementary information.

2.1.2 Sensitivity to amount of corrupted data—

RGCCA and SGCCA performance (R squared) is related to corruption with p-values 0.01179 and 0.0001525, respectively. SiMLR-CCA-ICA, SiMLR-CCA-SVD, SiMLR-Reg-ICA and SiMLR-Reg-SVD performances are impacted by corruption with p-value 0.01045, 0.04103, 0.0223 and 0.006516, respectively. As such, SGCCA (in this experiment) is most sensitive of these methods to corrupted data and SiMLR-CCA-SVD is least so. Inspection of the R squared performance plots indicates that the impact of corruption is not insubstantial with 24 of 120 RGCCA experiments leading to R squared less than 0.2. For SGCCA, SiMLR-CCA-ICA, SiMLR-CCA-SVD, SiMLR-Reg-ICA and SiMLR-Reg-SVD, these totals are 13, 6, 5, 3 and 4 respectively. From this perspective of a performance cutoff, SGCCA does slightly better than RGCCA. However, both methods still are less reliable than all variants of SiMLR. In the remaining evaluation studies, we focus on contrasting SGCCA with SiMLR because they both involve feature selection which is more appropriate for the $p >> n$ (features >> samples) cases that we investigate.

2.2 Cancer survival prediction

We compare methods in the context of a training-testing paradigm for survival prediction[47,48,49,50] based on multi-omics benchmark data in glioblastoma (GBM)[51]. The biological data includes $n = 274$ subjects with: gene expression ($p = 12{,}042$ predictors)[44], methylomics/DNA methylation ($p = 5{,}000$)[45] and transcriptomics/micro RNA expression ($p = 534$)[46]. GBM also provides survival data, i.e. the number of days since diagnosis and whether or not death has occurred at that time.

The benchmark paper above showed that multiple canonical correlation analysis (MCCA pairwise CCA across all pairs)[52] had the total best prognostic value. Thus, a comparison between SGCCA and SiMLR is pertinent. Nevertheless, these approaches should not be considered as the best strategy given that single-omic analysis did nearly as well[51].

For both SGCCA and SiMLR and over 50 runs, we split the data into 80% training and 20% testing sets. In training data, we perform supervised dimensionality reduction where the ‘omics data is jointly reduced with both death and survival time acting as a fourth matrix. We train a Cox model[47,48,49,50] with the low-dimensional bases derived from the ‘omics data in the previous step. Lastly, we predict the survival outcome in test data and evaluate accuracy with the concordance metric[49,50].

Under this design, a better method will both produce higher concordance values and produce more concordance values that meet or exceed the value of 0.6 which is considered a threshold of moderate agreement[53] and has been reported in recent studies for reasonably performing methods[49,50]. We repeat the above experiments over 50 splits of the data in order to gain an empirical estimate of the difference in performance between SiMLR and SGCCA with different input data. We also test at three different sparseness levels thereby comparing the performance of solutions that yield feature vectors with low (25%), moderate (50%) and high (75%) sparseness. The study design is further explained in 7.8.

In this evaluation, SiMLR with the reconstruction/regression energy shows an advantage over SGCCA in terms of predictive performance as measured by concordance in test data. Average concordance for SiMLR with the reconstruction error term is 0.64 (two-sided paired *t*-test comparing concordance performance, *p*-values < 0.0001 for both source separation options at the best performing sparseness levels). The covariance energy and SGCCA perform nearly identically and, on average, do not exceed 0.6 concordance (two-sided paired *t*-test, *p*-values > 0.05 for both source separation options and all sparsenesss levels). Neither method was optimized for this problem in terms of data selection, parameter or pre-processing choices. Moreover, the authors are not domain experts in this field. As such, this acts as a fairly unbiased comparison of these tools. In summary, SiMLR with reconstruction performs statistically equivalently or better on average than SGCCA in this problem with SiMLR-Reg-ICA showing the best results over all sparseness values. SiMLR-CCA performed equivalently to SGCCA. Further details may be found in supplementary information.

2.3 Brain age prediction

The pediatric template of brain perfusion (PTBP[54]) includes freely available multiple modality neuroimaging consistently collected in a cohort of subjects between ages 7 and 18 years of age. PTBP also includes a variety of demographic and cognitive measurements. A relevant reference analysis of this data is available in[55].

We provide pre-processed (machine learning ready) matrix format for three measurements taken in 97 subjects: voxelwise cortical thickness[56], fractional anisotropy (FA) derived from diffusion tensor imaging and cerebral blood flow (CBF) all at the voxel-wise level at 1mm resolution. The dimensionality of the matrices are 97 × 515,317 for thickness and CBF and 97 × 438,394 for FA. The development-related phenotype matrix consists of the subjects' sex, chronological age, total IQ score, verbal IQ score and performance IQ score. The IQ variables are highly correlated. The study design is explained in 7.9.

2.3.1 Computation time—

In this example, SGCCA and SiMLR demonstrate overall similar run-time with a few exceptions. These exceptions are caused by data-dependent longer convergence times. SGCCA runs, over each of five folds, for 235, 29, 29, 30 and 33 minutes. SiMLR with CCA and ICA runs for 105, 46, 46, 41 and 46 minutes. SiMLR with CCA and SVD runs for 78, 58, 94, 53 and 56 minutes. SiMLR with regression and ICA runs for 44, 47, 55, 39 and 60 minutes. SiMLR with regression and SVD runs for 28, 31, 41, 54 and 33 minutes. Overall differences in run-time likely depend on convergence settings as well as the variability of the energy function combined with the input data.

2.3.2 Prediction outcomes—

Figure 3 demonstrates the predictions' mean absolute error (MAE) for each algorithm that we tested. None of the methods perform well for predicting IQ-related scores. However, both SiMLR and SGCCA component regression produce reasonable predictions of brain age[57]. These values reported here are competitive with those reported in[55]. The MAE differences translate to a statistically significant improvement in performance between SiMLR (all variants) and SGCCA (best result *p*-value = 0.0002965, worst *p*-value = 0.04692). At the individual prediction level, this means that SiMLR produces a more accurate age in 61 of 97 cases for SiMLR-CCA-SVD.

2.4 Imaging-genetics data

Pediatric Imaging, Neurocognition, and Genetics (PING) data[58] offers the opportunity to jointly study two types of neuroimaging, anxiety and depression related SNPs[2] and self-reported scores of anxiety and depression. The training portion of the data is defined by subjects who have only neuroimaging and SNPs. This allows us to perform dimensionality reduction in training subjects alone (n=508) to identify a much lower dimensional space that encodes the variability induced jointly by SNPs and brain structure. The test set is distinguished by individuals who have not only imaging and genetics measurements but also self-reported measures of anxiety and depression. We perform inference in the test set (n=162) to determine which, if any, of the learned embeddings relate to these scores.

The evaluation criterion, here, is inferential i.e. we prefer the method that leads to embeddings with greater relationship to the clinical scores. This exploratory study is shared in supplementary information. Primarily, SiMLR identifies more signal related to anxiety and depression in the inferential portion of the study, when compared to SGCCA. I.e. more components relate to self-report anxiety and depression scores – with both SNPs and brain structure (thickness and white matter integrity, like PTBP) contributing – when using SiMLR compared to SGCCA. SiMLR leads to 3 components whereas SGCCA only identifies a single component related to anxiety. More importantly, however, we noted a severe difference in computation time. This study computes 40 components from high-dimensional data. SGCCA takes over 24 hours to compute these components. SiMLR (all variants) takes less than an hour. The primary difference between this study and the others included as examples is that the number of rows and the number of columns is relatively large (for training, n=508 and $p_{\text{thickness}} = 66,565$, $p_2 = 68,966$, $p_3 = 4{,}309$). As such, the advantage SGCCA gains by working in the dual space may be overwhelmed by the combined cost of relatively large covariance matrices and the need to perform deflation for each set of components. In contrast, SiMLR computes the feature matrix for each modality in one pass through the optimization.

We also provide a related supplementary result in multi-omic Alzheimer's disease neuroimaging initiative (ADNI) data. This result shows another way to relate imaging and cognition to genetic measurements: through polygenic risk measurements. Polygenic risk scores effectively reduce the dimensionality of genetic data based on an a priori weighted sum of trait-associated alleles. In supplementary information, we contrast SiMLR, RGCCA and SGCCA applied to tabular data where $n >> p$. The results of this joint reduction repeat trends shown elsewhere in this document; however, the difference between sparse and unconstrained dimensionality reduction is relatively less due to the more classical setting ($n >> p$). This demonstrates that SiMLR can be used effectively, like RGCCA, even when a dataset is already relatively well-powered.

3 Discussion

Interestingly, in both simulation and real clinical data, SiMLR extracts different signal than related methods as judged by the systematic performance trends in our stdies. This feature may relate to the method's core mathematics: high-dimensional embedding vectors are constructed purely from within modality data but the low-dimensional bases are derived from cross-modality representations determined by a user-selected source separation algorithm. If the SVD source separation method is chosen, then this representation will be orthogonal; if ICA is chosen, they will be statistically independent where independence is defined by measuring non-gaussianity[40] (one of the tenets of fastICA is "non-gaussianity is independence"). This type of approach will only be effective in datasets that exhibit some degree of cross-modality covariation that can be decoded meaningfully into multiple "true" source signals. If this is not possible, then SiMLR may obscure rather than extract hidden signal.

Performance differences could relate to two other implementation details. SiMLR uses a primal formulation that directly optimizes in the high-dimensional feature space in which the energy function is defined. In contrast, SGCCA computes solution updates in a low-dimensional space (see Algorithm 1 in[26]) and then performs soft-thresholding on the resulting vectors after transformation to the high-dimensional feature space. Secondly, SGCCA uses deflation to generate multiple components whereas SiMLR operates on full feature matrices. That is, SiMLR computes full matrix solutions all at once and uses the underlying source separation method to optimize these vectors jointly at each iteration of the algorithm. This improves computational efficiency when extracting several components (i.e. more than a few) but also marks a clear difference in the objective functions defined by SiMLR and SGCCA. These technical factors all contribute to differences in the outcomes reported here.

There are several limitations to this study and opportunities for future work. Primarily, we believe this approach and the current findings will be strengthened by application in related, larger datasets such as those provided by Adolescent Brain Cognitive Development (ABCD), the UK Biobank and Human Connectome Project. Furthermore, while we present methods for matrix standardization (the usual centering and scaling), this may not be a perfect solution for all cases, in particular when data deviates strongly from gaussianity. Other alternatives are available (e.g. rank transformations), but those are not explored here. While this work provides several automated or semi-automated strategies for selecting regularization parameters and the rank (k) for the feature vectors, none of these strategies are "perfect". This is unsurprising, given that technical research continues about parameter setting even in more classical methodology (PCA, CCA). While cross-validation approaches may also be used, the computational and data expense for these is relatively high and they also suffer theoretical as well as practical limitations in terms of effectiveness[60]. Despite these issues that are rather general, we believe the current implementation and interface to SiMLR, combined with guidance provided here, may yield a practically useful tool for multiple modality analysis of biomedical imaging and related data.

A second caveat to this study is that the design is explicitly multivariate and, as such, we do not interrogate the predictive value of individual embeddings. Our statistical focus is on the omnibus models. Other researchers may prefer to study individual embeddings independently. This is one known limitation within the current demonstration of SiMLR. Future work may explore this research in conjunction with extracting not just joint but also individual structure. This latter advantage is one provided by JIVE. SiMLR could also be further optimized directly for clustering problems, e.g. by implementing a multi-view clustering loss[61,62].

Two technical findings from these results are suggestive of directions for future work. First, SiMLR's performance suggests that a primal formulation for large joint matrix learning problems is feasible and can achieve competitive results in real and simulated data. Second, direct computation of feature matrices (vs. feature vectors as is done with deflation schemes) provides computational advantages in our experiments. However, further analysis of the differences between these technical approaches within a consistent framework would be needed to draw deeper conclusions. As always, we recommend interested users contact developers/authors for guidance or with issues arising in the use of this software.

7 Methods

7.1 Terminology

We outline the terminology used in the discussion that follows.

- **Multi-view:** several modalities collected in one cohort; alternatively, the same measurements taken across different studies[42]. We focus on the first case here.
- **Covariation:** we use the term in two contexts. As a general concept, we mean systematic changes in one modality are reflected in a predictable amount of change in other modalities. In the mathematical context, we use the definition of covariation for discrete random variables.
- **Latent space/embeddings:** both terms refer to an (often lower-dimensional) representation of high-dimensional data. These are also known as components in PCA. In the context of this paper, we are approximating the (hidden) latent space with the learned embeddings. Often, the true latent space cannot be known. We compute embeddings (or components), here, by multiplying feature vectors against input data matrices. Importantly, SiMLR can compute latent spaces that target either statistical independence (the ICA source separation algorithm[40]) or orthogonality (the SVD algorithm). Deflation-based schemes, on the other hand, only target orthogonality.
- **Feature vectors:** these are weights on the original features. In SiMLR, the feature vectors are the solutions that we are seeking. Projecting the feature vectors onto the original data will provide a low-dimensional representation.

These concepts are expanded upon in more detail below.

7.2 Software platform: ANTsR

The core platform, ANTsR, builds upon the powerful R language to interface and help organize raw neuroimaging, genomics and other data. ANTsR uses Rcpp[63] to wrap Insight ToolKit (ITK, now in version 5[64]) and ANTs (currently in version 2.3.3[65]) C++ tools for the R environment.

7.3 Technical background

7.3.1 Data representation—

SiMLR assumes "clean" data as input. This data has no missing values and is structured in matrix format with each modality matched along rows (the subjects/samples) while the columns represent features. Single nucleotide polymorphism (SNP) data is often formatted this way after imputing to a common reference dataset such as the HapMap. In neuroimaging, we employ region of interest measurements or spatial normalization in order to map a high-dimensional image into this common representation. For example, if a brain template has p voxels within the cortex and the population contains n subjects, then the matrix representation of the population level voxel-wise, normalized cortical thickness map will be $X_{\text{thickness}}$ with dimensions $n \times p$. SiMLR accepts > 1 matrices organized in this manner. A study of m distinct modalities would have input matrices with dimensions $n \times p_i$ (subjects × predictors), noting that p_i need not equal p_j for any $i, j \in 1, \cdots, m$.

We discuss, briefly, the primary algorithms upon which SiMLR is based. We assume data matrices, below, are standardized (columns with zero mean, unit variance) and $\| \cdot \|$ denotes the Frobenius norm.

7.3.2 Multiple regression—

Multiple regression solves a least squares problem that optimally fits several predictors (the $n \times p$ matrix X) to an outcome (y). As a quadratic minimization problem, we have:

$$\arg\min_{\beta} \|y - X\beta\|^2,$$

with optimal least squares solution:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y .$$

Above, we may also add a "ridge" penalty $\lambda\|\beta\|^2$ on the βs which is useful if $p >> n$ i.e. in the case of complex, multi-view, and multivariate datasets as we propose to model here. In this document, n refers to the number of samples or subjects and p to predictors.

7.3.3 Principal component analysis—

PCA, like multiple regression, may be formulated as the solution to an energy minimization problem. Select $k < n$, then find $U(n \times k)$, $V(p \times k)$ that minimize reconstruction error (where we add an ℓ_1 constraint as in[67–69] to illustrate sparse PCA):

$$\arg\min_{U,V} \|X - UV^T\|^2 + \sum_k \lambda_k \|V_k\|_1,$$

with additional constraints $U = XV$ and $V^T V = I$ where I is the identity matrix. The details of these constraints may vary in regularized variants of the method. Each of the columns of X is, here, expressed as a linear combination of the columns of U. For several modalities, we would compute: $\left\{X_1 = U_1 V_1^T, \cdots, X_n = U_n V_n^T\right\}$. In this case, the "predictors" are the U_i and the V_i is analogous to the β in the multiple regression case. The V_i feature vectors will be sparse if the ℓ_0 or ℓ_1 penalty is used.

7.3.4 Canonical correlation analysis—

CCA may be thought of as a generalization of multiple regression. Denoting Y as a $n \times q$ matrix, CCA seeks to find solution matrices $U(k \times p), V(k \times q)$ that maximize correlation in a low-dimensional space between X and Y:

$$\arg\max_{U,V} tr\left(Corr\left(XU^T, YV^T\right)\right),$$

where $Corr$ is Pearson correlation and tr is the trace operator. In contrast to our previous formulation for PCA, CCA evaluates the objective function (the "energy") in a reduced dimensionality space. Any of the methods above can be made sparse by enforcing the penalties on the feature weights as described for sparse PCA with the caveat that optimality constraints must be relaxed. Non-convex optimization methods such as alternating minimization and/or projected gradient descent must then be used[70–72].

7.4 Similarity-driven multi-view linear reconstruction

SiMLR is a general framework that can be specified in forms that relate to either sparse PCA (a regression-like objective) or sparse CCA (a covariance-related objective). The primary concepts are illustrated in Figure 1. We make two assumptions about datasets to which we will apply SiMLR.

- **Assumption 1:** Real latent signal(s) are independent and linearly mixed across the biological system on which we are collecting several measurements (a standard assumption for blind source separation).
- **Assumption 2:** Sparse, regularized feature vectors can relate estimated latent signals in assumption 1 to the original data matrices through linear operations.

If data matches these assumptions then methods that can combine modalities have a better chance of finding the latent signals; e.g. joint analysis from (for example) genetics, neuroimaging and cognition may provide more reliable recovery of the true latent signal influencing them all. Furthermore, it is likely that spurious signal will not be shared across all modalities – or all elements of the features within a modality – in a consistent manner. Natural filtering of noise occurs in joint analysis because (most forms of) noise does not covary across measurement instances. Adding regularization goes further in adding robustness: methods regularized with sparseness terms (ℓ_0 or ℓ_1) can down-weight (even to zero) features that do not improve the objective function. A caveat of these assumptions is that if no covariation across measurements exists – or if noise overwhelms all modalities/measurements – then these methods may not be relevant.

7.4.1 The SiMLR objective function—We first present the high-level framework and will expand upon details for similarity measurement and regularization below. The core concepts in SiMLR include the fact that it incorporates flexible approaches to measuring differences between modalities (similarity-driven), can take as input several different matrices (multi-view) and that all operations are linear algebraic in nature (linear reconstruction). First, we define X_i as a $n \times p_i$ (subjects by features) matrix for a given measurement/view/modality. The i ranges from 1 to m i.e. the number of modalities (or views). Then SiMLR optimizes an objective function that seeks to approximate each modality from its partner matrices through a sparse feature matrix (V_i) and low-dimensional representations ($U_{\neq i}$):

$$\arg\min_{V_i} \sum_{i=1}^{m} S\left(X_i, f\left(U_{\neq i}\right), V_i\right) + \text{Regularization}(V_i),$$

where:

- k denotes the rank of V_i and U_i;
- V_i is a $p_i \times k$ matrix of feature/solution vectors (analogous to βs) for the X_i modality;
- $\forall_i \; U_i = X_i V_i$;
- $U_{\neq i}$ is a $n \times (k(m-1))$ low-dimensional representation of modalities other than X_i i.e. the column-bound matrix $U_{\neq 2} = [U_1, U_3]$ if $i = 2$ and $m = 3$;
- f is a function (with output dimensionality $n \times k$) that estimates a low-rank basis set from its argument, is related to **Assumption 1**, and is described in more detail below;
- S is a function measuring the quality of the approximation of X_i from the other modalities and is related to **Assumption 2**;

The $f\left(U_{\neq i}\right) = \widetilde{U}_{\neq i}$ is a key component in the SiMLR framework and is derived by performing blind source separation over the set of $j \neq i : \{X_j V_j\}$ embeddings (the $U_{\neq i}$). We now provide details for each term and other aspects of the implementation.

<u>Similarity Options.:</u> The default similarity measurement is one of difference. This is akin to the reconstruction form for PCA, discussed above. In this case, we have:

$$S\left(X_i, \widetilde{U}_{\neq i}, V_i\right) = \left\| X_i - \widetilde{U}_{\neq i} V_i^T \right\|^2 .$$

Here, SiMLR attempts to reconstruct – in a least-error sense – each matrix X_i directly from the basis representation of the other $n - 1$ modalities.

We also implement a similarity term inspired by CCA but modified for the SiMLR objective function. In prior work, we observed that the CCA criterion – in the under-constrained form here where we expect $p >> n$ – demonstrates some sensitivity to the sign of correlations[73]. As such, we implement an absolute canonical covariance (ACC) similarity measurement expressed as:

$$\frac{tr(|\widetilde{U}^{T}_{\neq i} X_i V_i|)}{\|\widetilde{U}_{\neq i}\| \|X_i V_i\|}.$$

Both reconstruction and ACC have easily computable analytical derivatives that are amenable to projected gradient descent, as used in our prior work[11,12,17]. This similarity term is most closely related to SABSCOR and SABSCOV in multi-block data analysis[74,75]. However, it focuses only on cross-modality signal.

For the reconstruction energy, SiMLR optimizes these feature vectors to reconstruct each full matrix from a reduced representation of the other matrices (the $\widetilde{U}_{\neq i}$). For ACC, SiMLR optimizes V_i to maximize covariance of $X_i V_i$ with the low-rank basis. As such, the latter similarity term may be more appropriate for recovering signal that exists more sparsely in the input matrices. This is because the operation $X_i V_i$ is able to completely ignore large portions of the given matrix X_i due to the sparseness terms in our regularization (described below). The regression energy, on the other hand, will be more directly informed by the raw high-dimensional matrix which may have advantages in some cases. Quadratic energies also tend to have larger capture ranges.

The method's performance also depends on the selection for the basis representation. We evaluate two options in this initial work:

- $f_{svd} = svd_u([U_{\neq \mathrm{i}}])$
- $f_{ica} = ica_S([U_{\neq \mathrm{i}}])$

The notation $[U_{\neq i}]$ indicates that we bind the columns together (cbind in R). Below alg will represent svd_u or ica_S. The method ica_S indicates that we take the independent components matrix (the S matrix) from the ICA algorithm (where ICA produces $X = AS$). The method svd_u indicates that we take the U component of the SVD (where SVD produces $X = UDV^T$).

We focus our evaluation on ica_S and svd_u functions in this work as we have found that they produce useful outcomes in example experiments and they are well-proven methods applied in several domains. The assumptions underlying ICA and SVD are related. They both fit our assumption 1 above of linearly mixed independent signals. The difference is the measure of independence. SVD (or PCA) assumes independence is measured by variance which leads to orthogonal basis functions. ICA uses non-gaussianity to measure independence. Both are valid options, from the theoretical perspective, and we rely on evaluation results to make recommendations about how to choose between these in practice.

Regularization options.: Regularization occurs on the V_i i.e. our feature matrices. Denote:

- v_{ik} as the the kth feature vector in V_i;

- G_i is $p_i \times p_i$ a sparse regularization matrix with rows that sum to one;
- γ_i as a scalar weight which could be used to regularize each component differently; effectively, this controls the sparseness and varies in zero to one.

Then the regularization terms take the form:

$$\text{Regularization}(V_i) = \sum_i \sum_k \gamma_i \|G_i v_{ik}\|_{\ell_p}^{+},$$

where $\| \cdot \|_{\ell_p}^{+}$ is the positivity constrained ℓ_p norm (usually, $p = 0$ or $p = 1$). This term both enforces sparseness via ℓ_p while providing data-adaptive degrees of smoothing via the the graph regularization matrix G^i. For neuroimaging, this latter feature means that one does not need to pre-smooth images before running SiMLR. In practice, $\| \cdot \|_{\ell_p}^{+}$ induces unsigned feature vectors. I.e. all non-zero entries will be either only positive or only negative.

Regularization weights: The parameterization of the sparseness for each modality is set by γ_i in the range of zero to one, where higher values are increasingly sparse (more values of the feature vector are zero). By default, γ_i is automatically set to accept the largest 50^{th} percentile weights but the user may decide to increase or decrease this value depending on the needs of a specific study. Alternatively, one may use hyperparameter tuning methods to automatically determine γ_i. For most applications, we recommend default values.

Regularization matrices: optional G_i are currently set by the user and must be determined in a data/application/hypothesis-specific manner. In implementation, we provide helper functions that allow the user to employ k-nearest neighbors (KNN) to build the non-zero entries of the regularization matrices. We use HNSW[76] to compute sparse KNN matrix representations for the G_i. HNSW is among the most efficient methods currently available and, combined with sparse matrix representations, make graph regularization on large input matrices efficient. This aspect of regularization promotes smooth feature vectors where the nature of smoothness is typically determined by proximity either spatially or in terms of feature magnitude or feature correlation.

Although we provide default methods, choice of regularization should involve some consideration on the part of the user. Because there is no single theoretically justified answer to these questions, the best general approach would be to use hyper-parameter optimization. Alternatively, domain-specific knowledge may be used to guide parameter setting, in particular sparseness and regularization. Rules of thumb should be, for regularization, that the estimated V_i should appear to reflect biologically plausible feature sets. For sparseness, biological plausibility should also be considered although we believe our default parameters provide good general performance. As such, regularization (i.e. construction of the G_i) should perhaps be given more domain-dependent attention by users. Examples below provide clarity on how we set these terms in practice. E.g. in neuroimaging, we may use $k = 5^d$ mask-constrained neighbors for KNN where d is image dimensionality. For genomics or psychometrics data, we may set regularization simply by thresholding correlation (or linkage disequilibrium[77]) matrices.

7.5 Optimization

The overall approach to optimizing the SiMLR objective is that of projected gradient descent[78]. In this context, one derives the optimization algorithm without regularization constraints and then, at each iteration, projects to the sub-space defined by the regularization terms. The SiMLR objective function for V_i, at a given iteration, depends only on the set values for X_i and $\widetilde{U}_{\neq i}$. As such, we only need the gradient of the similarity term with respect to V_i which greatly simplifies implementation. We optimize total energy E via a projected gradient descent algorithm:

$$\begin{gathered}\text{loop until convergence:}\\ \forall_i V_i^{\text{new}} \leftarrow H(G_i \star (V_i - \partial S/\partial V_i \epsilon_i))\\ \forall_i \widetilde{U}_{j \neq i} \leftarrow f_{alg}([X_j V_j^{\text{new}}]_{\neq i})\end{gathered}$$

where:

- $[X_j V_j^{\text{new}}]_{\neq i}$ is the collection of low-dimensional projections resulting from multiplying the feature vectors onto the data matrices where $\neq i$ indicates that the i^{th} projection is held out;
- H is the thresholding operation which here is applied separately to each column of V_i (see the iterative hard/soft thresholding literature[72] and[78] which suggests that ℓ_0 penalties provide greater robustness to noise);
- ϵ_i is a gradient step parameter determined automatically by line-search over the total energy E.

Recall that f_{alg} is a dimensionality reduction step that reduces $U_{\neq i}$ to a k-column matrix. Here, we provide an example gradient calculation for our default reconstruction error:

$$\begin{aligned} S &= \left\| X_i - \widetilde{U}_{\neq i} V_i^T \right\|^2, \\ \partial S/\partial V_i &= -2(X_i^T - V_i \widetilde{U}_{\neq i}^T)\widetilde{U}_{\neq i}, \end{aligned}$$

which allows updating the full V_i at each gradient step. SiMLR only allows gradient-based updates that improve the total energy; these are arrived at by line search over the gradient step size and means that the objective function (driven primarily by the similarity term) is improved by the new candidate solution; this process is iterated until the method reaches a fixed point. A fixed point is – practically speaking – a convergent solution. I.e. if we further iterate the algorithm, the solutions do not change beyond some small numeric fluctuation.

This strategy also allows SiMLR to work directly on the feature matrices themselves even when $p >> n$. When large numbers of components are being computed, this can lead to a distinct computational advantage in comparison to deflation methods.

7.6 Parameters and initialization

We summarize default (recommended) parameters and preprocessing steps for the methodology.

- Matrix pre-processing is performed automatically. Unless the user overrides default behavior, we transform each matrix such that: $\forall X_i : X_i \leftarrow \frac{sc(X_i)}{np_i}$ where sc denotes scaling and centering applied to the matrix columns. Normalizing by np controls the relative scale of the eigenvalues of each matrix.
- Number of components (k) – The practice for setting these values is very similar to practice in PCA or SVD; it may be determined via statistical power considerations, cross-validation or set to be $k = n - 1$, one less than the number of subjects. This is a problem that is currently under active research[60].
- Similarity measurement – evaluation and comparison of similarity choices is ongoing. Trade-offs are comparable to choosing correlation versus Euclidean distance for vectors and better performance may be gained in a data-dependent manner. ACC is faster to compute on a per-iteration basis but may require more iterations to converge. This latter comment is an empirical observation based on our studies which, again, may be data dependent.
- Source separation algorithm – Trade-offs are comparable to choosing between SVD and ICA in general. Effectively, ICA should force the multiple component solutions toward statistical independence in a non-gaussian sense. SVD would be more appropriate for separating purely gaussian sources that are mixed linearly.

The nature of the feature space is impacted by the constraints on the U_i which are determined by the user-selected source separation algorithm. SVD produces an orthogonal latent space whereas ICA does not. ICA seeks a latent space that demonstrates statistical independence, that is, that are maximally non-gaussian[40,41]. It is an empirical question about which is "best" for a given dataset; neither is right or wrong in an absolute sense. SVD and ICA are both used in many practical applications in machine learning and statistics. Our experiments confirm that both options can produce results that outperform RGCCA. Overall, the reconstruction error with ICA source separation appears to give good general performance in our experiments.

SiMLR may be initialized with several different approaches:

- random matrices for all or for each individual modality;
- a joint ICA or SVD across concatenated modalities (recommended and default behavior);
- Any other initial low-rank basis set e.g. derived from RGCCA, etc which may be passed to the algorithm by the user;

Due to the fact that SiMLR cannot guarantee convergence to a global optimum (sparse selection is a NP-hard problem), several different starting points should be evaluated when using SiMLR in new problems. This is in concordance with the theory of multi-start global optimization which we can only approximate in practice[79]. Other joint reduction methods such as SGCCA suffer the same limitation. Our recommended default behavior avoids forcing users to explore multiple starting points but does not eliminate this fundamental issue that is general to the field of feature selection in high-dimensional spaces.

7.7 Generation of simulated data

We construct simulated data that matches this setting by constructing 3 matrices from different (modality-specific) multivariate distributions. Each matrix contains a common low-dimensional basis (the true latent signal) which can be recovered by joint dimensionality reduction. Matrices are generated by the following steps.

- Generate a rank-K basis set (S_j^K of size $K \times p_j$ where $j \in 1, 2, 3$) of gaussian distributed signal that is smoothed by a different amount over each simulation run; K and p_j vary over simulations.
- Generate ground truth latent signal matrix $B = [\vec{\beta}_1, \cdots, \vec{\beta}_K]$ with n rows that will weight each basis matrix S_j^K. B is consistent across all modalities but n varies across simulations.
- Generate each $n \times p_j$ data matrix by computing $M_j = B\,S_j^K$.
- Replace a percentage of the columns of each matrix M_j with random noise.
- Split the data into 80% train and 20% test and run the candidate algorithms on each M_j^{train} matrix. Lastly, use linear regression to relate the learned embeddings to the true source signal ($\vec{\beta}_1^{\text{train}}$) and predict $\vec{\beta}_1^{\text{test}}$ in the test data from the learned embeddings.

The above steps produce data where each of the three 3 matrices is generated from very different distributions but that contain a common latent signal. The key is that the latent signal is at least partially consistent ($\vec{\beta}_1$) and can, in some cases, be recovered by joint analysis. Recoverability varies across simulations due to both corruption and the intrinsic variability of the underlying generating distributions. Supplementary Figure 1 illustrates the overall design of the simulation study.

7.7.1 Signal recovery—For each experiment, we run 120 simulations and evaluate the quality of the recovered signal by training a linear regression algorithm to relate the learned basis to the true basis. We then predict the latent signal in held-out test data (80 percent of subjects are used for training and 20 percent for testing). In this scenario, better performing methods will lead to more accurate predictions of the latent signal in the tesing subjects. We can evaluate, by two-tailed paired t-test on the recovery (measured by R squared of the fit), whether SiMLR performs better than, equal to or worse than other methods.

7.7.2 Sensitivity to amount of corrupted data—As above, for each of the 120 simulations, a varying degree of corruption to each matrix is performed. That is, a random percentage of the matrix that contains true signal is replaced with noise signal with no relationship to the latent ground truth. The amount of corruption varies between 10 and 90 percent of the column entries. This enables us to test the degree to which recovery performance can be predicted from the amount of corruption where corruption is represented as a 3-vector for each experiment where each entry in the vector codifies the amount of corruption.

7.8 Cancer survival prediction

The hypothesis is that gene expression, methylomics and transcriptomics, which track the biological/genetic dynamics of tumor activity, will improve prediction of patient-specific outcomes. However, these data are fairly high-dimensional relative to the number of subjects. As such, targeted dimensionality reduction is needed to overcome the $p >> n$ problem (where p, here, refers to predictors) in order to allow low-dimensional versions of these predictors (i.e. embeddings) to be used in a classical regression context[47,48,49,50].

We selected the GBM (glioblastoma, brain) set from the multi-omic benchmark collection. GBM allows a train-test split with sufficient variability in survival in both train and test groups. These data were compiled by the multi-omic cancer benchmark organizers from The Cancer Genome Atlas (TCGA).

The statistical model used for training and prediction is a Cox proportional hazards regression model implemented in the coxph function in the survival package. We evaluate concordance in test data via the survcomp package. Concordance is similar to a rank correlation method and is used to assess agreement of the predicted outcomes with true outcomes. Its value under the null hypothesis of no predictive value is 0.5. Values greater than roughly 0.6 show some evidence of predictive power[47,48,49,50].

In the evaluation, graph-based regularization parameters are selected to include roughly 2.5% of the predictors in each predictor ‘omics matrix (see the call to the regularizeSimlr function). As such, regularization is present but neither overwhelming nor optimized for this data. I.e. this value was chosen based on the desire for a small amount of denoising in the solution space. Neither method was optimized for this problem in terms of data selection, parameter or pre-processing choices. As such, this acts as a fairly unbiased comparison of these tools.

7.9 Brain Age prediction

Supplementary Figure 2 summarizes the design of the study. First, a 5-fold cross-validation grouping of subjects is defined. For each fold, SiMLR and SGCCA are run with parameters that are set to select interpretable “network”-like components. In this example, we choose these parameters specifically at higher sparseness levels to facilitate interpretability. We record computation time as well as the embedding vectors for each modality. We then train, within each fold, a linear regression model to predict age and IQ-related variables from the neuroimaging embeddings. This is a form of principal component regression. These predictions are stored for each fold to facilitate a final comparison of performance across all folds. We use this technique, rather than repeated resampling as in prior studies, in part because the run-time for this problem can be relatively long, up to 235 minutes for SGCCA.

Data Availability—All visualized plots in the main manuscript are generated from the code capsule which contains both the specific data sources and software calls necessary to reproduce the figures [80].

Simulation data: The simulation data is built on the fly in ‘R’. The scripts that generate the data are publicly available [80].

Multi-omic cancer benchmark: We downloaded evaluation data from the multi-omic cancer benchmark[51] website http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. As with other results in the main body of the paper, data is available in our code capsule [80] along with the relevant statistical details and calls needed to reproduce results reported here. The data is free to use with no restrictions.

Brain age: The brain age data used in preparation of this article were obtained from the Pediatric Template of Brain Perfusion (PTBP) [81]. These data were originally downloaded from https://figshare.com/articles/dataset/The_Pediatric_Template_of_Brain_Perfusion_PTBP_/923555. The relevant subset is available in our code capsule [80]. The data is free to use with no restrictions.

PING for imaging genetics study: Supplementary data used in preparation of this article were obtained from the Pediatric Imaging, Neurocognition and Genetics Study (PING) database (http://ping.chd.ucsd.edu). PING requires a user to register and request data. The review of the request may also require institutional support and justification of data use. We originally gained access to these data in 2013 as part of the now defunct “PING-in-a-box” service.

ADNI for imaging genetics study: Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer’s disease (AD). For up-to-date information, see www.adni-info.org. As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

ADNI requires a user to register and request data. The review of the request may also require institutional support and justification of data use. We originally gained access to these data in 2008. The version used in our supplementary study was downloaded in august 2020 from LONI.

Code Availability—ANTsR is open source and freely available at https://github.com/ANTsX/ANTsR. The github version of the code is typically in development. The specific release version of the code and scripts used in the analyses and generation of figures in the main body of this paper are available in the code capsule [80].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work is supported by the Office of Naval Research N00014-18-1-2440 and K01-ES025432-01. Correspondence and request for materials should be directed to Dr. Avants.

Supplementary data used in preparation of this article were obtained from the Pediatric Imaging, Neurocognition and Genetics Study (PING) database (http://ping.chd.ucsd.edu). As such, the investigators within PING contributed to the design and implementation of PING and/or provided data but did not participate in analysis or writing of this report. A complete listing of PING investigators can be found at https://ping-dataportal.ucsd.edu/sharing/Authors10222012.pdf

Supplementary data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.;Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.;Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

1. Cole JH, Marioni RE, Harris SE & Deary IJ Brain age and other bodily 'ages': implications for neuropsychiatry. (2019) doi:10.1038/s41380-018-0098-1.
2. Wray NR et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. Nature Genetics (2018) doi:10.1038/s41588-018-0090-3.
3. Habeck C, Stern Y & Alzheimer's Disease Neuroimaging Initiative. Multivariate data analysis for neuroimaging data: overview and application to Alzheimer's disease. Cell Biochem Biophys 58, 53–67 (2010). [PubMed: 20658269]
4. Shamy JL et al. Volumetric correlates of spatiotemporal working and recognition memory impairment in aged rhesus monkeys. Cereb Cortex 21, 1559–1573 (2011). [PubMed: 21127015]
5. McKeown MJ et al. Analysis of fMRI data by blind separation into independent spatial components. Hum Brain Mapp 6, 160–188 (1998). [PubMed: 9673671]
6. Calhoun VD, Adali T, Pearlson GD & Pekar JJ A method for making group inferences from functional {MRI} data using independent component analysis. Hum Brain Mapp 14, 140–151 (2001). [PubMed: 11559959]
7. Calhoun VD, Liu J & Adali T A review of group {ICA} for f{MRI} data and {ICA} for joint inference of imaging, genetic, and {ERP} data. Neuroimage 45, S163–72 (2009). [PubMed: 19059344]
8. Avants B Relating high-dimensional structural networks to resting functional connectivity with sparse canonical correlation analysis for neuroimaging. vol. 136 (2018).
9. Pierrefeu A. de et al. Structured Sparse Principal Components Analysis With the TV-Elastic Net Penalty. IEEE transactions on medical imaging 37, 396–407 (2018). [PubMed: 28880163]
10. Du L et al. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. Bioinformatics (Oxford, England) 32, 1544–1551 (2016).
11. Avants B et al. Sparse unbiased analysis of anatomical variance in longitudinal imaging. vol. 6361 LNCS (2010).
12. Avants B et al. Sparse canonical correlation analysis relates network-level atrophy to multivariate cognitive measures in a neurodegenerative population. NeuroImage 84, (2014).

13. Du L et al. GN-SCCA: Graphnet based sparse canonical correlation analysis for brain imaging genetics. in Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) (2015). doi:10.1007/978-3-319-23344-4_27.

14. Guigui N et al. Network regularization in imaging genetics improves prediction performances and model interpretability on Alzheimer's disease. in Proceedings - international symposium on biomedical imaging (2019). doi:10.1109/ISBI.2019.8759593.

15. Lee DD & Seung HS Learning the parts of objects by non-negative matrix factorization. Nature (1999) doi:10.1038/44565.

16. Chalise P & Fridley BL Integrative clustering of multi-level 'omic data based on non-negative matrix factorization algorithm. PLoS ONE (2017) doi:10.1371/journal.pone.0176278.

17. Dhillon P et al. Subject-specific functional parcellation via Prior Based Eigenanatomy. NeuroImage 99, (2014).

18. Tikhonov AN On the stability of inverse problems. Doklady Akademii Nauk Sssr (1943).

19. Bell JB, Tikhonov AN & Arsenin VY Solutions of Ill-Posed Problems. Mathematics of Computation (1978) doi:10.2307/2006360.

20. Smilde AK, Westerhuis JA & De Jong S A framework for sequential multiblock component methods. Journal of Chemometrics (2003) doi:10.1002/cem.811.

21. Tenenhaus A & Tenenhaus M Regularized Generalized Canonical Correlation Analysis. Psychometrika (2011) doi:10.1007/s11336-011-9206-8.

22. Tenenhaus M, Tenenhaus A & Groenen PJ Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods. Psychometrika (2017) doi:10.1007/s11336-017-9573-x.

23. Zhan Z, Ma Z & Peng W Biomedical Data Analysis Based on Multi-view Intact Space Learning with Geodesic Similarity Preserving. Neural Processing Letters 1 (2018) doi:10.1007/s11063-018-9874-9.

24. Baltrusaitis T, Ahuja C & Morency LP Multimodal Machine Learning: A Survey and Taxonomy. (2018) doi:10.1109/TPAMI.2018.2798607.

25. Kettenring JR Canonical analysis of several sets of variables. Biometrika (1971) doi:10.1093/biomet/58.3.433.

26. Tenenhaus A et al. Variable selection for generalized canonical correlation analysis. Biostatistics (2014) doi:10.1093/biostatistics/kxu001.

27. Rohart F, Gautier B, Singh A & Lê Cao KA mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Computational Biology (2017) doi:10.1371/journal.pcbi.1005752.

28. Garali I et al. A strategy for multimodal data integration: Application to biomarkers identification in spinocerebellar ataxia. Briefings in Bioinformatics (2017) doi:10.1093/bib/bbx060.

29. Gloaguen Arnaud, Philippe Cathy, Frouin Vincent, Gennari Giulia, Dehaene-Lambertz Ghislaine, Laurent Le Brusquet AT Multiway Generalized Canonical Correlation Analysis. Biostatistics In Press, (2020).

30. Hotelling H Canonical Correlation Analysis (CCA). J. Educ. Psychol (1935).

31. Hotelling H Relations between two sets of variants. Biometrika 321–377 (1936).

32. Lock EF, Hoadley KA, Marron JS & Nobel AB Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. Annals of Applied Statistics (2013) doi:10.1214/12-AOAS597.

33. Yu Q, Risk BB, Zhang K & Marron JS JIVE integration of imaging and behavioural data. NeuroImage (2017) doi:10.1016/j.neuroimage.2017.02.072.

34. Ceulemans E, Wilderjans TF, Kiers HA & Timmerman ME MultiLevel simultaneous component analysis: A computational shortcut and software package. Behavior Research Methods (2016) doi:10.3758/s13428-015-0626-8.

35. Argelaguet R et al. Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Molecular Systems Biology (2018) doi:10.15252/msb.20178124.

36. Carmichael I et al. Joint and individual analysis of breast cancer histologic images and genomic covariates. arXiv preprint arXiv:1912.00434 (2019).

37. McMillan C et al. White matter imaging helps dissociate tau from TDP-43 in frontotemporal lobar degeneration. Journal of Neurology, Neurosurgery and Psychiatry 84, (2013).

38. McMillan C et al. Genetic and neuroanatomic associations in sporadic frontotemporal lobar degeneration. Neurobiology of Aging 35, (2014).

39. Cook P et al. Relating brain anatomy and cognitive ability using a multivariate multimodal framework. NeuroImage 99, (2014).

40. Hyvärinen A & Oja E Independent component analysis: A tutorial. Notes for International Joint Conference on Neural Networks (IJCNN'99), Washington DC (1999).

41. Hyvärinen A & Oja E Independent component analysis: Algorithms and applications. Neural Networks (2000) doi:10.1016/S0893-6080(00)00026-5.

42. De Vito R, Bellio R, Trippa L & Parmigiani G Multi-study factor analysis. Biometrics (2019) doi:10.1111/biom.12974.

43. Haykin S & Chen Z The cocktail party problem. (2005) doi:10.1162/0899766054322964.

44. Goodwin S, McPherson JD & McCombie WR Coming of age: Ten years of next-generation sequencing technologies. (2016) doi:10.1038/nrg.2016.49.

45. Yong WS, Hsu FM & Chen PY Profiling genome-wide DNA methylation. (2016) doi:10.1186/s13072-016-0075-3.

46. Ozsolak F & Milos PM RNA sequencing: Advances, challenges and opportunities. (2011) doi:10.1038/nrg2934.

47. Andersen PK & Gill RD Cox's Regression Model for Counting Processes: A Large Sample Study. The Annals of Statistics (1982) doi:10.1214/aos/1176345976.

48. Fox J & Weisberg S Cox Proportional-Hazards Regression for Survival Data in R. Most (2011).

49. Huang L et al. Development and validation of a prognostic model to predict the prognosis of patients who underwent chemotherapy and resection of pancreatic adenocarcinoma: A large international population-based cohort study. BMC Medicine (2019) doi:10.1186/s12916-019-1304-y.

50. Neums L, Meier R, Koestler DC & Thompson JA Improving survival prediction using a novel feature selection and feature reduction framework based on the integration of clinical and molecular data. in Pacific symposium on biocomputing (2020). doi:10.1142/9789811215636_0037.

51. Rappoport N & Shamir R Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research (2018) doi:10.1093/nar/gky889.

52. Witten DM, Tibshirani R & Hastie T A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics (2009) doi:10.1093/biostatistics/kxp008.

53. Barnhart HX, Haber M & Song J Overall concordance correlation coefficient for evaluating agreement among multiple observers. (2002) doi:10.1111/j.0006-341X.2002.01020.x.

54. Avants BB et al. The pediatric template of brain perfusion. Scientific data (2015) doi:10.1038/sdata.2015.3.

55. Kandel B, Wang D, Detre J, Gee J & Avants B Decomposing cerebral blood flow MRI into functional and structural components: A non-local approach based on prediction. NeuroImage 105, (2015).

56. Tustison N et al. Logical circularity in voxel-based analysis: Normalization strategy may induce statistical bias. Human Brain Mapping 35, (2014).

57. Franke K & Gaser C Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained? (2019) doi:10.3389/fneur.2019.00789.

58. Jernigan TL et al. The Pediatric Imaging, Neurocognition, and Genetics (PING) Data Repository. NeuroImage (2016) doi:10.1016/j.neuroimage.2015.04.057.

59. Manola J et al. Prognostic model for survival in patients with metastatic renal cell carcinoma: Results from the international kidney cancer working group. Clinical Cancer Research (2011) doi:10.1158/1078-0432.CCR-11-0553.

60. Bro R, Kjeldahl K, Smilde AK & Kiers HA Cross-validation of component models: A critical look at current methods. Analytical and Bioanalytical Chemistry (2008) doi:10.1007/s00216-007-1790-1.
61. Bickel S & Scheffer T Multi-view clustering. in Proceedings - fourth ieee international conference on data mining, icdm 2004 (2004). doi:10.1109/ICDM.2004.10095.
62. Wang Y, Wu L, Lin X & Gao J Multiview Spectral Clustering via Structured Low-Rank Matrix Factorization. IEEE Transactions on Neural Networks and Learning Systems (2018) doi:10.1109/TNNLS.2017.2777489.

Methods-only References

63. Eddelbuettel D & Balamuta JJ Extending R with C++: A Brief Introduction to Rcpp. American Statistician (2018) doi:10.1080/00031305.2017.1375990.
64. Avants B, Johnson H & Tustison N Neuroinformatics and the the insight toolkit. Frontiers in Neuroinformatics 9, (2015).
65. Avants B et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. NeuroImage 54, (2011).
66. Muschelli J et al. Neuroconductor: An R platform for medical imaging analysis. Biostatistics (2019) doi:10.1093/biostatistics/kxx068.
67. Zou H, Hastie T & Tibshirani R Sparse principal component analysis. Journal of Computational and Graphical Statistics (2006) doi:10.1198/106186006X113430.
68. Shen H & Huang JZ Sparse principal component analysis via regularized low rank matrix approximation. Journal of Multivariate Analysis (2008) doi:10.1016/j.jmva.2007.06.007.
69. Jolliffe IT, Trendafilov NT & Uddin M A Modified Principal Component Technique Based on the LASSO. Journal of Computational and Graphical Statistics (2003) doi:10.1198/1061860032148.
70. Lin CJ Projected gradient methods for nonnegative matrix factorization. Neural Computation (2007) doi:10.1162/neco.2007.19.10.2756.
71. Jain P, Netrapalli P & Sanghavi S Low-rank matrix completion using alternating minimization. in Proceedings of the annual acm symposium on theory of computing (2013). doi:10.1145/2488608.2488693.
72. Blumensath T & Davies ME Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis (2009) doi:10.1016/j.acha.2009.04.002.
73. Pustina D, Avants B, Faseyitan OK, Medaglia JD & Coslett HB Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. Neuropsychologia 115, 154–166 (2018). [PubMed: 28882479]
74. Hanafi M PLS Path modelling: Computation of latent variables with the estimation mode B. Computational Statistics (2007) doi:10.1007/s00180-007-0042-3.
75. Tenenhaus A, Philippe C & Frouin V Computational Statistics and Data Analysis Kernel Generalized Canonical Correlation Analysis. Computational Statistics and Data Analysis (2015) doi:10.1016/j.csda.2015.04.004.
76. Malkov YA & Yashunin DA Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) doi:10.1109/TPAMI.2018.2889473.
77. Hill WG & Robertson A Linkage disequilibrium in finite populations. Theoretical and Applied Genetics (1968) doi:10.1007/BF01245622.
78. Bahmani S & Raj B A unifying analysis of projected gradient descent for lp- constrained least squares. Applied and Computational Harmonic Analysis (2013) doi:10.1016/j.acha.2012.07.004.
79. Martí R, Resende MG & Ribeiro CC Multi-start methods for combinatorial optimization. European Journal of Operational Research (2013) doi:10.1016/j.ejor.2012.10.012.
80. Avants BB, Tustison NJ & Stone JR SiMLR in ANTsR: Interpretable, similarity-driven multi-view embeddings from high-dimensional biomedical data. Code Ocean(2021) [Source Code]. doi: 10.24433/CO.3087836.v2.

81. Avants BB, Tustison NJ & Wang DJJ The Pediatric Template of Brain Perfusion (PTBP). figshare (2013). Dataset. 10.6084/m9.figshare.923555.v20

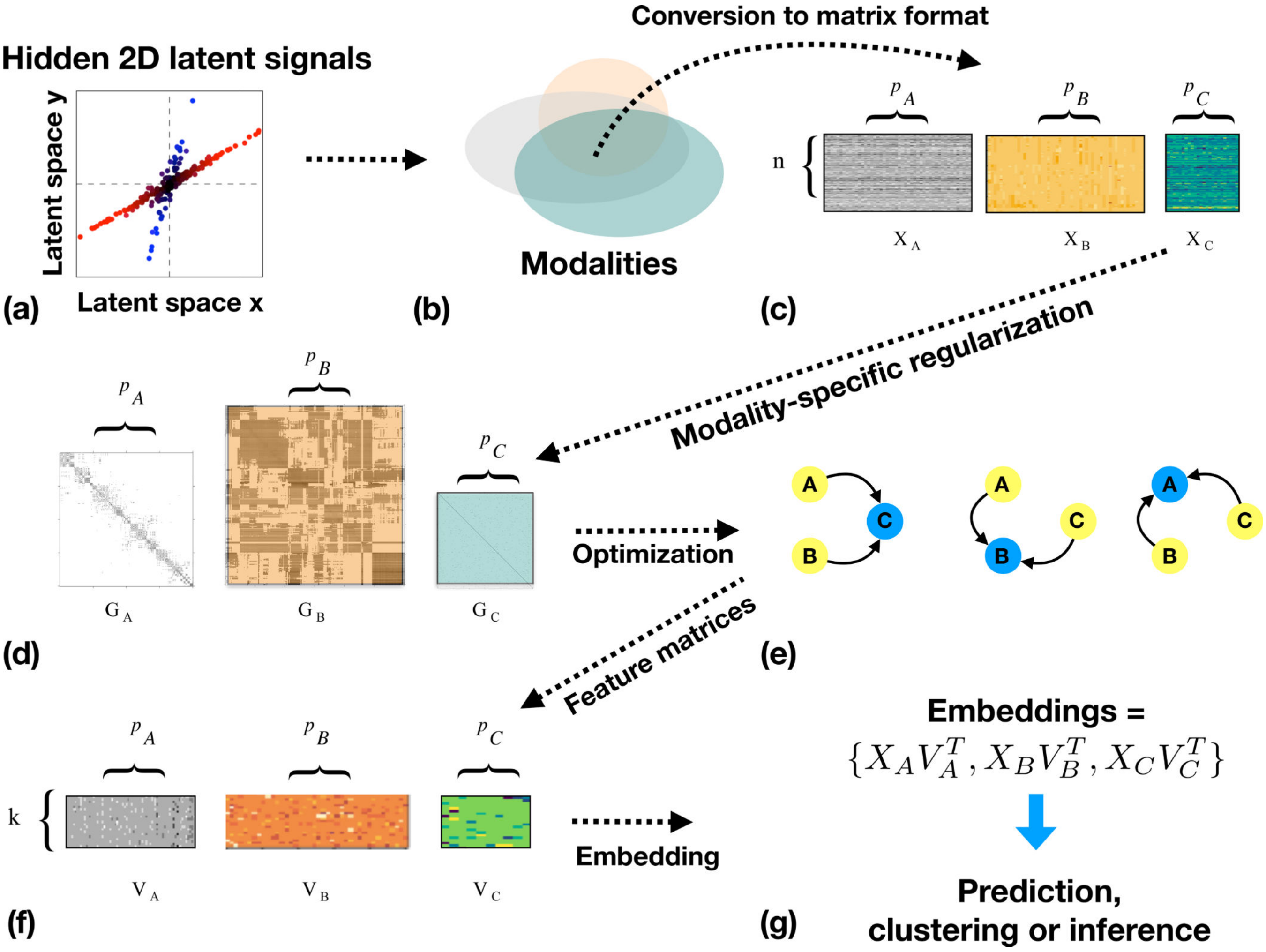

Figure 1:

An overview of SiMLR's workflow: (a) Two statistically independent signals are shown here to represent the hidden latent signal potentially two components of a disease process; (b) The latent signal is manifested across three different modalities (each represented by an oval) where the joint component of the signal is represented in the overlap. (c) This three-view data is converted to three matrices X_A,X_B,X_C; in this effort we focus on matrices with common number of subject here denoted by n and variable number of predictors (p_A, p_B, p_C). (d) Sparse regularization matrices (G_A,G_B,G_C) are constructed with user input of domain knowledge or via helper functions; (e) SiMLR iteratively optimizes the ability of the modalities to predict each other in leave one out fashion; (f) Sparse feature vectors emerge which can be interpreted as weighted averages over selected columns of the input matrices that maintain the original units of the data. These are used to compute embeddings in (g) and passed to downstream analyses. Alternatively, one could permute the SiMLR solution to gain empirical statistics on its solutions.

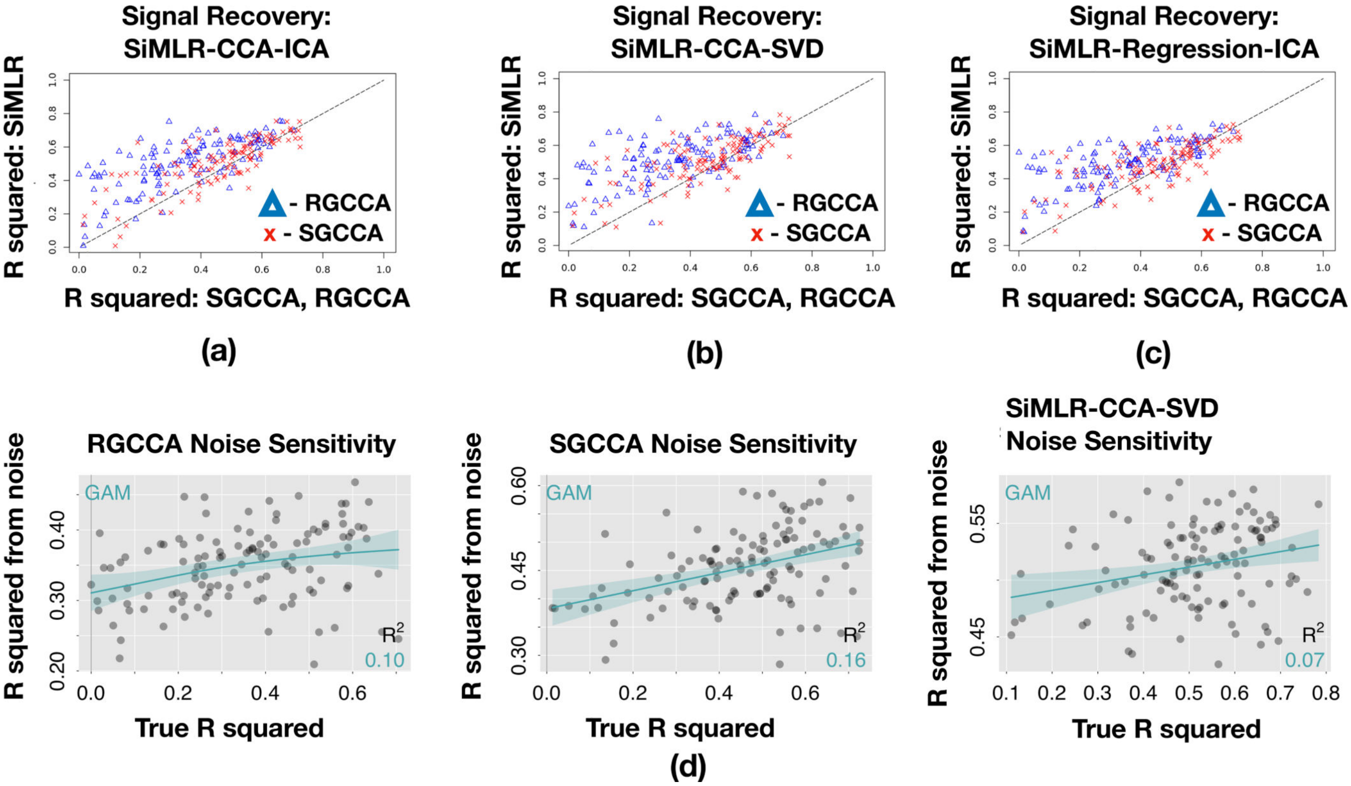

Figure 2:

SiMLR simulation study results: sensitivity to noise and ability to recover signal. In each panel, (a-c), the SiMLR signal recovery performance (120 simulations) in terms of R squared is plotted against RGCCA and SGCCA performance. (a) Demonstrates performance of signal recovery of SiMLR with the CCA energy and ICA source separation method. (b) Demonstrates performance of signal recovery of SiMLR with the CCA energy and SVD source separation method. (c) Demonstrates performance of signal recovery of SiMLR with the regression energy and ICA source separation method. Plots in (d) show how well signal recovery (R squared) can be predicted from the amount of matrix corruption. In this case, ideally, matrix corruption would minimally impact performance; therefore, lower scores are better. The best fit line (computed by generalized additive model (GAM)) is shaded with 95% confidence intervals.

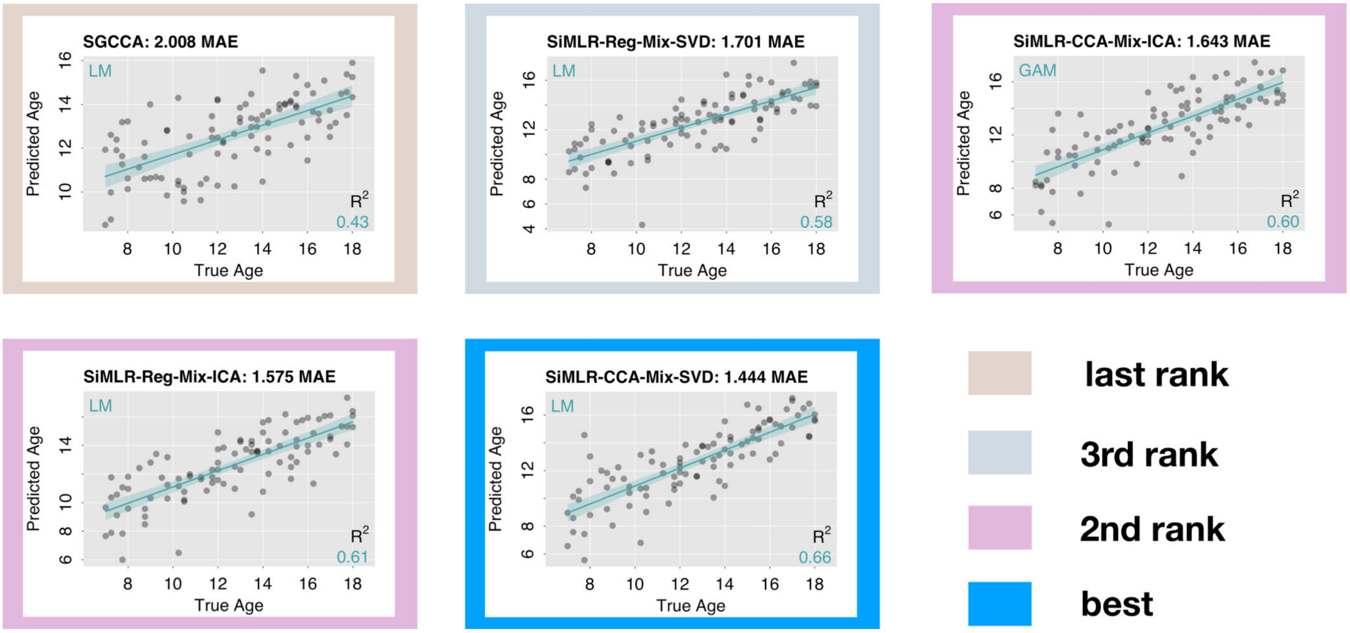

Figure 3:

PTBP fully supervised brain age prediction: comparison to SGCCA. In each panel, we show the ability to predict chronological age from the brain. Confidence intervals are shown as gray shaded regions around a best-fit linear regression line. *R* squared for the predicted model fit is also shown. Performance ranking is provided on the figure’s right and is based on the mean absolute error between the predicted and real age.

Table 1:

Summary of experimental results. The $x+/-y$ indicates mean and standard deviation values in the table below. RGCCA = regularized generalized canonical correlation analysis; SGCCA = sparse generalized canonical correlation analysis; Sim = similarity-driven multivariate linear reconstruction (SiMLR); Reg = regression; CCA = absolute canonical covariance; ICA = ICA blind source separation (BSS) method; SVD = SVD (BSS) method. Best results are bold. The PING examples are exploratory analyses described in the supplementary information as we cannot directly share the data. The “n comp” description in the PING table refers to the number of significant components related to either anxiety or depression.

study	RGCCA	SGCCA	SiMLR-CCA-ICA	SiMLR-CCA-SVD	SiMLR-Reg-ICA	SiMLR-Reg-SVD	metric
Signal-Sens.	0.35+/−0.18	0.45+/−0.17	0.5+/−0.15	**0.51+/−0.14**	0.49+/−0.13	0.49+/−0.14	R-squared
Noise-Sens.	0.09	0.16	0.09	**0.06**	0.07	0.1	R-squared
Multi-omic	N/A	0.56+/−0.12	0.56+/−0.13	0.56+/−0.14	**0.64+/−0.08**	0.64+/−0.11	Concordance
Brain Age	N/A	2+/−1.5	1.6+/−1.2	**1.4+/−1.2**	1.6+/−1.3	1.7+/−1.2	MAE
PING-Anx	N/A	1 comp.	N/A	3 comp.	**3 comp.**	N/A	Inferential
PING-Dep	N/A	0 comp.	N/A	1 comp, (trend)	**1 comp, (trend)**	N/A	Inferential