SiMLR: Figures and Table

Brian B. Avants et al.

2020-10-17

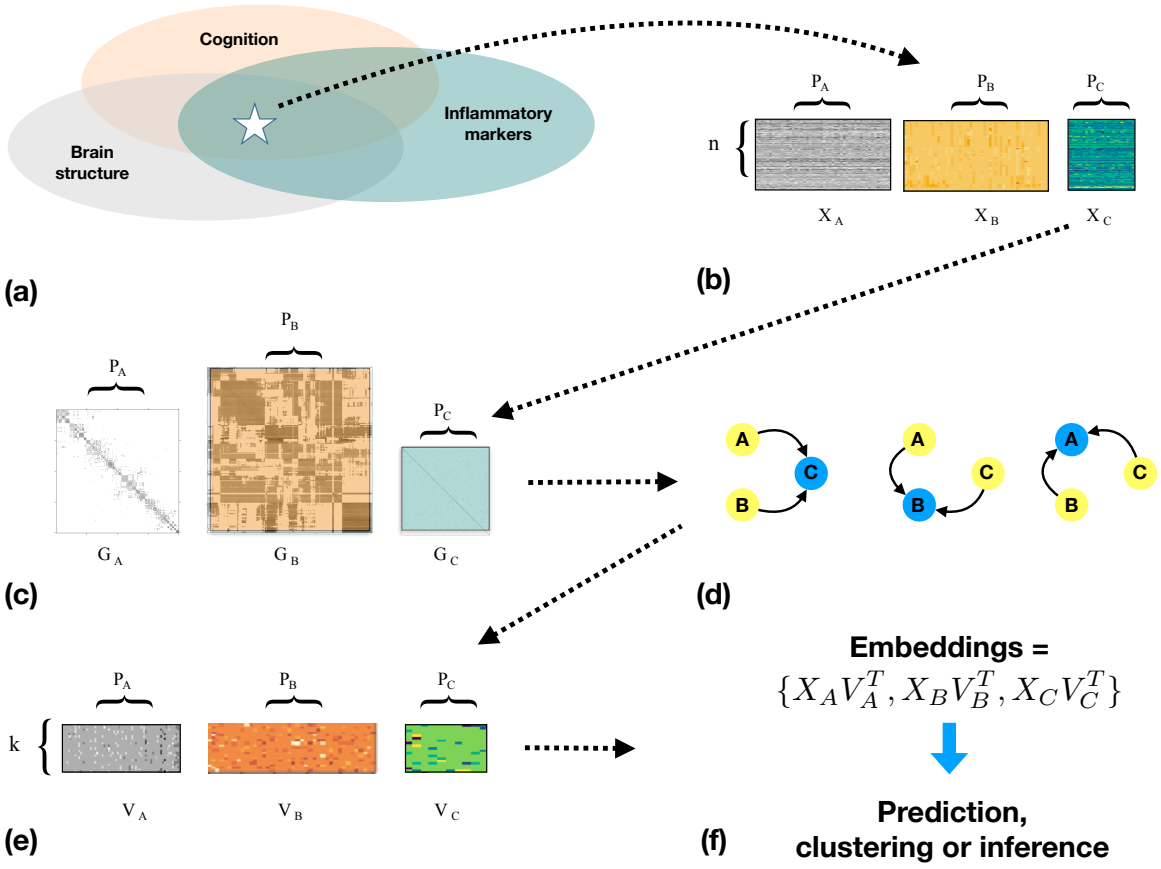


Figure 1: Example SiMLR study overview: (a) SiMLR and related methods expect a correlated multi-omic, multivariate dataset as input. (b) The data is converted to matrices; in this effort we focus on matrices with common number of subject here denoted by n and variable number of predictors ($p_{A,B,C}$). (c) Sparse regularization matrices (G) are constructed with user input of domain knowledge or via helper functions; (d) SiMLR iteratively optimizes the ability of the modalities to predict each other in leave one out fashion; (e) Sparse feature vectors emerge which are used to compute embeddings in (f) and passed to downstream analyses. Alternatively, one could permute the SiMLR solution to gain empirical statistics on its solutions.

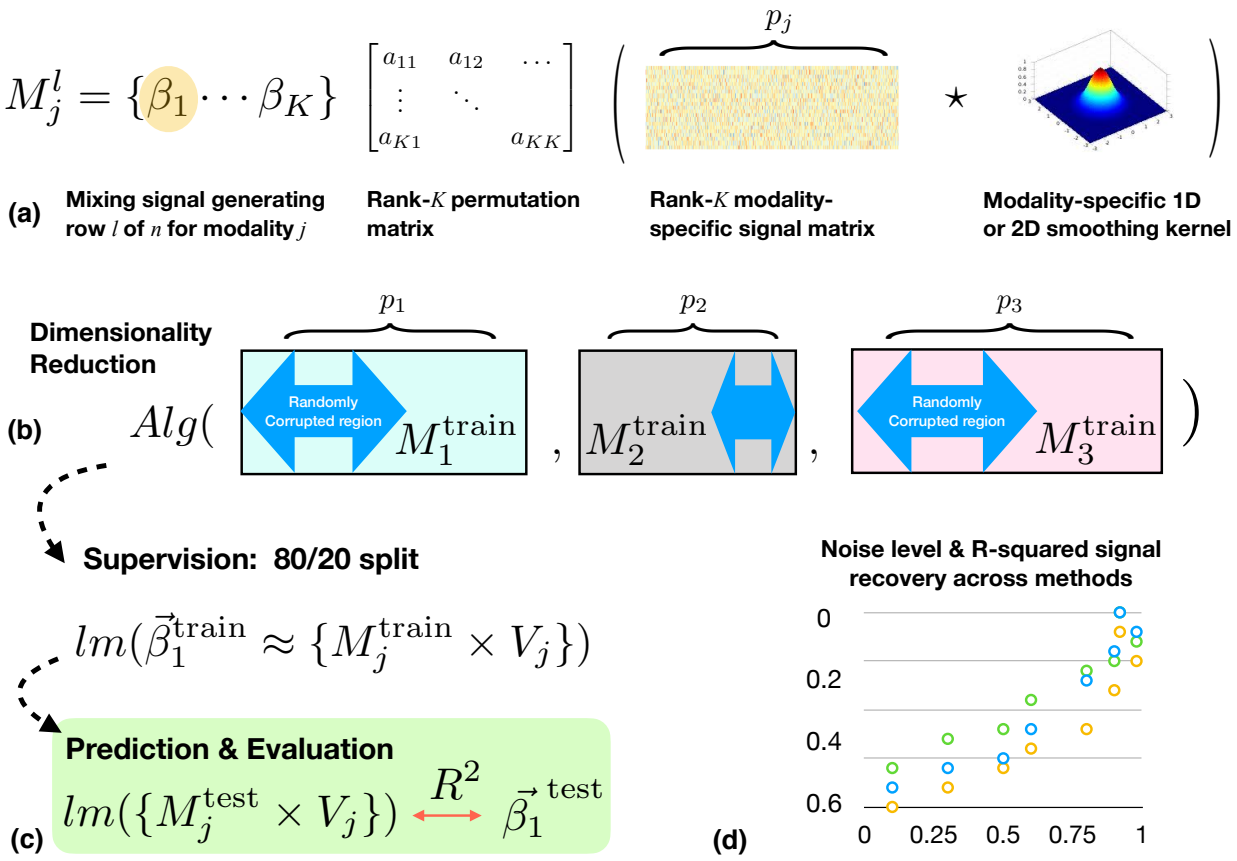


Figure 2: Conceptual overview of the SiMLR simulation study which defines an inverse problem for signal recovery with known ground truth. (a) shows the generation of a given row, M_j^l, in a single simulated matrix, M_j. The k-rank generating basis set is smoothed to induce signal-specific covariation as is present in many types of real biological data. The highlighted beta is the common signal that we seek to recover. (b) The study also randomly corrupts each of the three generated matrices by eliminating any true signal in some fraction of the matrix. The fraction of corruption is drawn from a uniform distribution between 0.1 and 0.9 (i.e. 10 and 90 percent corruption). (c) We define a random 80/20 split for each simulation and learn from the 80 percent of training data. RGCCA, SGCCA and SiMLR are each run in the dimensionality reduction step. A linear regression method takes the low-dimensional embeddings, the ground truth signal in the training data and then predicts the test data signal. (d) This process enables us to evaluate the signal recovery performance and how it is impacted by the corruption process.

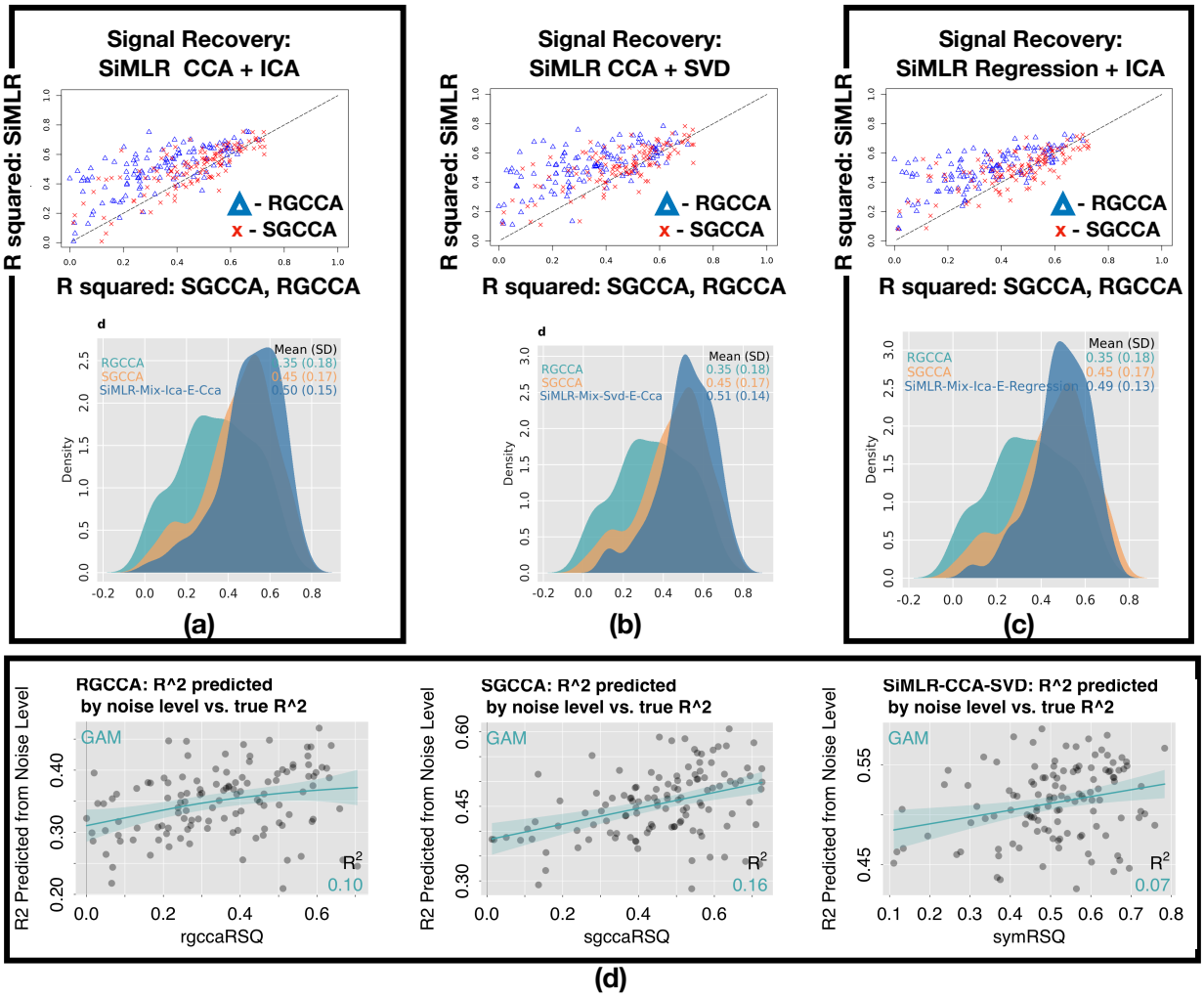


Figure 3: SiMLR simulation study results: sensitivity to noise and ability to recover signal. In each panel, (a-c), the SiMLR signal recovery performance in terms of R^2 is plotted against RGCCA and SGCCA performance. Thus, higher scores are better and points above the diagonal dotted line show superior SiMLR performance in pair-wise fashion. (a) Demonstrates performance of signal recovery of SiMLR with the CCA energy and ICA mixing method. (b) Demonstrates performance of signal recovery of SiMLR with the CCA energy and SVD mixing method. (c) Demonstrates performance of signal recovery of SiMLR with the regression energy and ICA mixing method. The lower plots in (d) show how performance is impacted by the amount of matrix corruption. Here, lower scores are better. In this simulation study, SiMLR with CCA and SVD mixing does best in terms of both raw scores and pairwise test statistics.

Table 1: Summary of RGCCA-SGCCA-SiMLR comparison results. 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 mixing method; SVD = SVD mixing method. Best results are highlighted in cadet blue; worst in antiquewhite. The second ranking approach is in pink. SiMLR with the absolute canonical covariance similarity measurement and SVD (SimCCASVD) as a mixing method performs best overall. For the multi-omics example, SiMLR with the regression energy and ICA mixing method (SimRegICA) outperforms SGCCA most consistently across sparseness levels, provides closely competitive performance overall, and is highlighted in pink. In brainAge, all SiMLR variants perform substantially better than SGCCA. 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. The last row summarizes our overall ranking and explains the relationship of cell color to the ranking system. The rank is calculated by counting instances of a given rank across columns.

study	**RGCCA**	**SGCCA**	**SimCCAICA**	**SimCCASVD**	**SimRegICA**	**SimRegSVD**	**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	0.62 +/ 0.01	0.64 +/ 0.03	0.65 +/ 0.03	0.65 +/ 0.04	0.61 +/ 0.03	Concordance
brainAge	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.	5 comp.	N/A	Inferential
PING-Dep	N/A	1 comp.	N/A	1 comp.	5 comp.	N/A	Inferential
Overall:	6th (worst)	5th	3rd	1st	2nd	4th	rank count

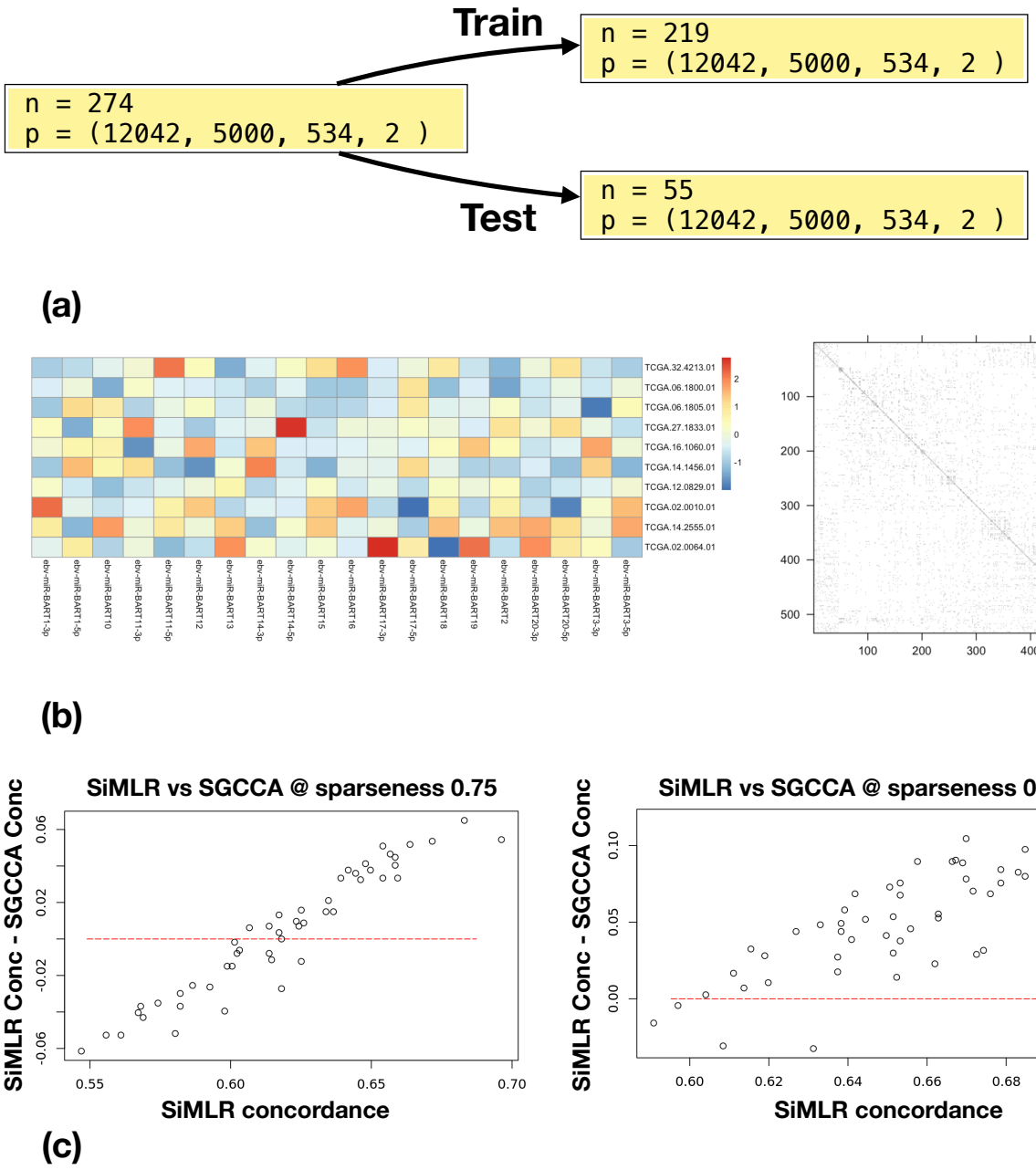
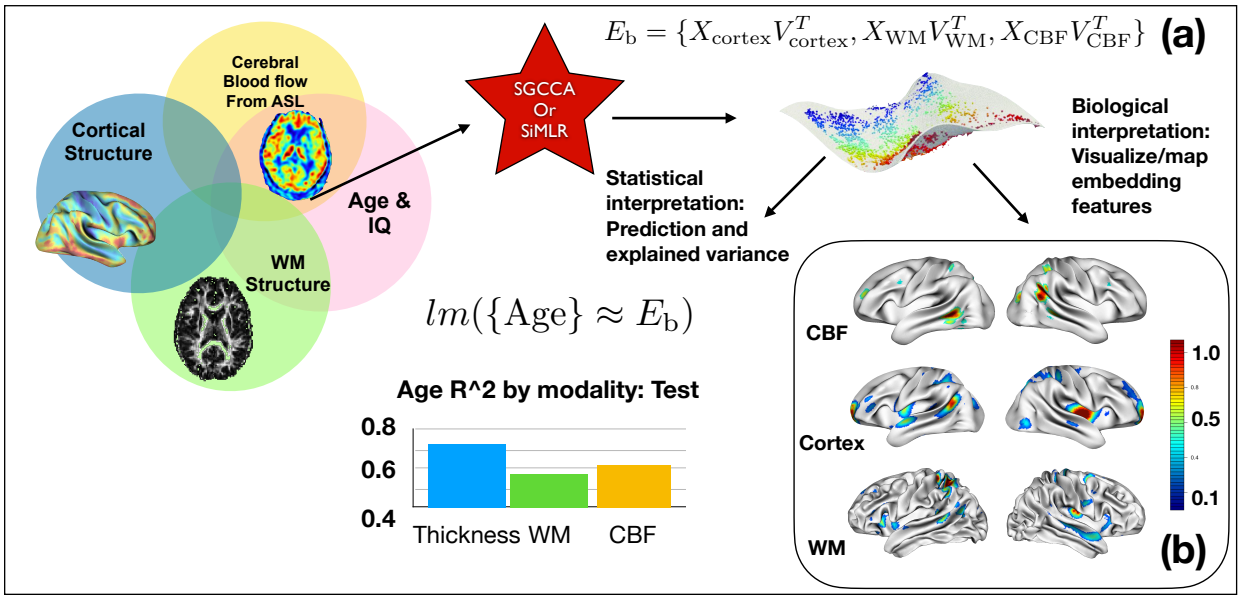


Figure 4: Multi-omic study: supervised survival prediction comparison to SGCCA. SiMLR outperforms SGCCA in predicting survival time in test data when dimensionality reduction is performed jointly on all four modalities in training data. The performance metric is concordance, i.e. the degree to which the rank-ordering of the predicted surivival outcomes match the ground truth. In (a), we show the train-test split numbers for each of the simulation runs as well as the n and p for each measurement type: gene expression, methylomics, transcriptomics and survival data. This real data example was drawn from the multi-omic benchmark collection. (b) Shows an example subset of the real data and its associated regularization matrix, the latter of which was generated by one of our helper functions simply based on the data correlation matrix. (c) Shows plots of SiMLR concordance minus SGCCA concordance against SiMLR concordance. As such, points above the red dotted line show where SiMLR does better. At left in (c) shows a case where SiMLR and SGCCA demonstrate close performance and (right) where SiMLR outperforms SGCCA the most. Both sparseness levels and objective function impact performance. SiMLR regression energy with ICA mixing outperformed SGCCA across all sparseness levels.



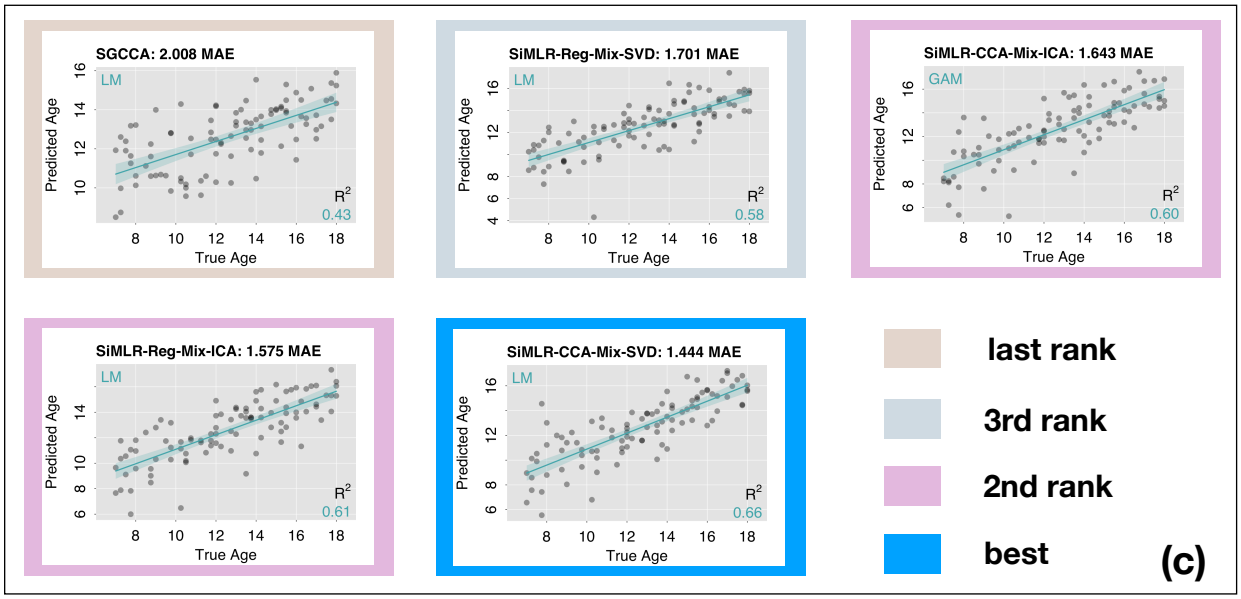


Figure 5: PTBP fully supervised brain age prediction: comparison to SGCCA. Brain age is the subject's age predicted from neuroimaging data. Four matrices are input where cortical thickness, white matter integrity and cerebral blood flow derive from different types of neuroimages; the fourth modality describes brain maturation in terms of age and IQ measurements. Panel (a) shows the overall study design where embeddings are computed as in prior examples and then passed downstream to facilitate statistical and biological interpretation. The first phase of statistical interpretation (the bar plots) compares the ability of each modality to predict age independently and suggests thickness is most predictive (when acting alone); WM and CBF have close performance to each other (data drawn from best performing method). In (b), we show the feature vectors from the best performing method noting that the weights are relative to each feature vector where its values are scaled to zero to one. In (c) we show the ability to predict real age from the brain where SiMLR-CCA-SVD does best; none of the methods perform well for IQ prediction.