SPEAR: a Sparse Supervised Bayesian Factor Model for Multi-omic Integration

Jeremy P. Gygi[1], Anna Konstorum[2], Shrikant Pawar[3], Steven H. Kleinstein[1,2,4,#], and Leying Guan[5,#,*]

[1]Program in Computational Biology & Bioinformatics, Yale University, New Haven, CT, USA, [2]Department of Pathology, Yale School of Medicine, New Haven, CT, USA, [3]Department of Genetics, Yale Center for Genomic Analysis (YCGA), Yale School of Medicine, New Haven, CT, USA, [4]Department of Immunobiology, Yale School of Medicine, New Haven, CT, USA, [5]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

#Co-senior authors, *To whom correspondence should be addressed.

Abstract

Motivation: Unsupervised factor modeling, which preserves the primary sources of data variation through low-dimensional factors, is commonly applied to integrate high-dimensional multi-omics data. However, the resulting factors are suboptimal for prediction tasks due to the separation between factor construction and prediction model learning. A supervised factor model that effectively utilizes the responses while accounting for structural heterogeneity across omics is needed.

Results: We present SPEAR, a supervised variational Bayesian framework that decomposes multi-omics data into latent factors with predictive power. The method adaptively determines factor rank, emphasis on factor structure, data relevance and feature sparsity. SPEAR improves reconstruction of underlying factors in synthetic examples and prediction accuracy of COVID-19 severity and breast cancer tumor subtypes.

Availability: SPEAR is a publicly available R-package hosted at https://bitbucket.org/kleinstein/SPEAR.

Contact: jeremy.gygi@yale.edu; leying.guan@yale.edu

1 Introduction

The explosion of high-throughput technologies such as RNA-seq, metabolomics, proteomics, and high-dimensional cytometry has enabled the simultaneous collection of data from multiple "omics" of the same sample[1]. While the dimensionality of these datasets is high, often, only a small number of key factors are important for a particular modeling task. For example, activation of distinct biological pathways (e.g., interferon signaling) can each drive hundreds of genes[2], proteins[3], and metabolites[4] to vary together across samples.

Unsupervised dimensionality reduction techniques for discovering factors are widely used in practice[5], including both non-probabilistic approaches such as principal component analysis (PCA), single and multi-block canonical correlation analysis (CCA)[11,49] as well as probabilistic approaches based on Bayesian factor models including Multi-omic Factor Analysis (MOFA)[6] and iClusterBayes[7]. However, in many applications, dimensionality reduction is an intermediate step toward the final goal of prediction. Separating the dimensionality reduction and model learning can be suboptimal, especially in cases where the prediction task is challenging. Supervised factor models for multi-omic integration mediate this issue, providing both response-guided dimensionality reduction and optimal prediction.

We introduce a novel Sparse suPervised bayEsian fActoR (SPEAR) model for multi-omic integration which directly identifies multi-omic factors influential to response(s) of interest and constructs a prediction model. SPEAR jointly models multi-omics data with the response(s) in a probabilistic Bayesian framework and models a variety of response types in regression and classification tasks, distinguishing itself from existing response-guided dimensionality reduction methods such as sMBPLS [9] and DIABLO[10]. Compared to these existing methods, SPEAR also efficiently and adaptively

determines the degree of emphasis on the factor structure, assay relevance and per-assay feature sparsity from the multi-omics data and the responses. Furthermore, SPEAR adaptively estimates factor rank directly from the multi-omics data.

We demonstrate SPEAR's advantages under simulated settings and validate SPEAR's improved performance on real public multi-omic datasets for breast cancer and SARS-CoV-2 (COVID-19) through balanced misclassification error and multi-class area under the receiver operating characteristic (AUROC) testing.

2 Methods

SPEAR constructs predictive factors by jointly modeling both the multi-omic assays (X) and the response of interest (Y) and approximates posteriors of the factor loadings (β) and factor scores (U) using the variational Bayes inference (Supplementary Methods B2). The method automatically balances between the construction of factors and the prediction of the response through a weight parameter (w) selected automatically via cross-validation (Online Methods - Factor Model). When w is large ($w \geq 1$), SPEAR emphasizes the construction of factors that explain the structure of the multi-omic assays. In contrast, when w is small ($w \approx 0$), SPEAR will forgo any attempt to reconstruct factors from the data and focus entirely on optimizing the prediction of the response. Thus, SPEAR constructs predictive factors influenced by the multi-omic assays only when supported by the data, otherwise returning a simple linear predictive model. By default, the largest w whose mean cross-validated error falls within one standard deviation of the overall minimum cross-validated error is selected[44], favoring higher values of w by design.

Multi-omic datasets are often high-dimensional, and different assays may capture different sources of variations. To combat this, SPEAR models assay-wise factor influence using Automatic Relevance Determination (ARD) [8] and the feature sparsity using spike-and-slab priors (Supplementary Methods A1). ARD allows for the construction of factors with asymmetrical assay contributions, and spike-and-slab priors identify key sparse predictive multi-omic features. Like alternative sparse supervised methods, feature importance is represented by a vector of coefficients for each factor, with a higher magnitude representing a higher contribution to the factor construction. In contrast to methods like MOFA and DIABLO, SPEAR returns two different types of coefficients for each multi-omic feature, denoted as regression coefficients and projection coefficients. The regression coefficients represent the sparse predictive features most important for factor construction, and the projection coefficients capture a factor's influence on all analytes. These projection coefficients are intended for downstream factor interpretation where seemingly redundant features ignored in the projection coefficients are vital to understanding the underlying biology[12]. Along with the coefficients, SPEAR will also approximate the posterior probabilities of multi-omic features being non-zero, allowing for automatic feature selection by keeping features with a posterior probability of being non-zero above 0.95, which is analogous to a false discovery rate (FDR) of 0.05. Choosing features based on this estimated posterior probability provides an advantage over other factor models, which facilitate feature selection by arbitrary cut-offs on coefficient values.

More detailed explanations of the probabilistic model, the inference procedure, parameter tuning, model rank selection, and modeling of different types of responses are provided in the online methods and supplementary methods (Supplementary Methods A) sections.

3 Results

3.1 Improved prediction of synthetic Gaussian responses

We first evaluated the ability of SPEAR to predict a Gaussian response on simulated data. In the simulation, five predetermined factor signals (U) were used to construct four multi-omic assays (X), each with 500 features (for a total of 2,000 simulated features) for 500 training and 2,000 test samples. The first two simulated factors (U_{XY}) were also used to construct a gaussian response vector (Y) for both the training and test groups, whereas the remaining three simulated factors (U_X) were purposefully only used to construct X. This procedure was repeated across a gradient of signal-to-noise ratios (low signal, moderate signal, and high signal), with 10 independent iterations for each ratio (see Supplementary Methods D for details).

We trained SPEAR across a gradient of values for weight w to demonstrate the effect of w on SPEAR's performance. As a comparison, we also employed a two-step MOFA-based model that performs lasso regression[45] using factors derived from MOFA (denoted as MOFA in the following) as well as a vanilla Lasso regression model using concatenated features in lieu of factors (denoted as Lasso). Model comparison was measured by calculating the mean squared error (MSE) of the test data as well as Pearson correlation testing of the constructed factors against both Y (Fig. 2b) and U (Fig. 2c).

In scenarios with moderate signal-to-noise, SPEAR with $w = 1$ significantly outperformed MOFA (Fig. 2a). Correlation analyses of the factor scores for one iteration (denoted by the red line in Figure 2a) confirmed that SPEAR with $w = 1$ correctly identified all five simulated factors (U_{XY} and U_X), whereas MOFA only identified two of the uncorrelated factors (U_X, Factor 3 and Factor 5) likely due to the lack of supervision (Fig. 2c). SPEAR with $w = 0$ condensed all predictive signals from U_{XY} into a single factor, as evidenced by its correlation with U_{XY}.

3.2 Improved prediction of synthetic ordinal and multinomial responses

To demonstrate SPEAR's ability to model different types of responses, we also carried out additional non-gaussian simulations. Similar to the Gaussian simulation, five true factors (U) were used to construct four multi-omic assays (X) and a non-Gaussian response (Y) for both training and testing datasets. In the ordinal simulation, the simulated Factor 1 and Factor 2 scores (U_{XY}) were combined to generate nonlinear signals from which a seven-class ordinal response could be generated (Fig. 3a, Fig. 3b). This process was repeated through 10 independent iterations across a gradient of signal-to-noise ratios (low signal, moderate signal, high signal). We then trained multiple SPEAR models with each treating the response differently (Gaussian, multinomial, and ordinal).

In the moderate signal-to-noise, the moderate signal case demonstrated Gaussian SPEAR's difficulty handling the nonlinearity of the class signals in a linear representation (Fig. 3c), further seen in the Gaussian model's predictions (Fig. 3d). SPEAR using multinomial and ordinal responses achieved lower balanced misclassification errors, with the ordinal model outperforming the others.

When SPEAR constructs factors to predict a multinomial or an ordinal response, the pipeline also returns the probabilities of each class assignment per sample. These can be utilized to further investigate the performance of a SPEAR model. From the ordinal simulation, ordinal SPEAR extracted predictive signals capable of identifying all seven classes as represented by the correct median probabilities (Fig. 3e).

We also carried out a multinomial simulation (Supplementary Fig. 4), simulating data in the same manner as both the Gaussian and ordinal simulations. SPEAR modeling a multinomial response outperformed all Gaussian models as well as multinomial Lasso and MOFA (Supplementary Methods D).

3.3 SPEAR improves prediction of breast cancer tumor subtypes and COVID-19 severity

To test whether SPEAR could achieve competitive performance in real data, we applied SPEAR to two publicly available multi-omic datasets: a breast cancer dataset of tumor samples by *Singh et al.*[10] with the goal of tumor subtype prediction, and a SARS-CoV-2 patient dataset of blood samples by *Yapeng et al.*[13] with the goal of predicting disease severity. We refer to these datasets as TCGA-BC (The Cancer Genome Atlas[14] - Breast Cancer) and COVID-19, respectively. The TCGA-BC dataset is composed of 989 biopsy samples each associated with RNA-Seq, miRNA, and methylation probe data from primary solid breast cancer tumors that have been annotated according to the PAM50 subtype signature into one of four subtypes: Luminal A (LumA), Luminal B (LumB), HER2-enriched (HER2) and Basal-like (Basal)[15]. The COVID-19 dataset contains plasma protein and metabolite compositions for 254 SARS-CoV-2 positive patient samples and 124 matched healthy subject samples. Each subject is associated with a severity score based on the World Health Organization (WHO) ordinal severity score[47], binned into four ordinal classes (healthy, mild, moderate, and severe).

We applied SPEAR, Lasso, MOFA, and DIABLO to predict the response from each multi-omic dataset. Datasets were preprocessed and split into training and testing cohorts as described in the online methods section. Classification performance was evaluated via the balanced misclassification error rate and the area under the receiver operating characteristic (AUROC) curves. AUROC significance was calculated via bootstrapping procedure and compared via DeLong's test[16].

The advantage of SPEAR was clear when looking at the multi-class AUROC, measuring a classifier's ability to discriminate each class individually across a gradient of threshold values. SPEAR showed higher AUROC in discriminating between all classes and was significantly better for multiple classes: SPEAR was significantly better at predicting both the LumB subtypes from the TCGA-BC dataset (Fig. 4c), and moderate SARS-CoV-2 severity from the COVID-19 dataset (Fig. 4g) compared to all other models. Upon investigation, the improved predictive performance of SPEAR on these classes was not due to a single factor but was rather achieved through combining information from multiple multi-omic factors (Supplementary Fig. 5, Supplementary Fig. 6). The balanced misclassification errors of SPEAR (0.15, 0.21) were comparable to those of Lasso (0.15, 0.21), MOFA (0.16, 0.21), and DIABLO (0.16, 0.23) for the TCGA-BC dataset and COVID-19 dataset respectively (Fig. 4d, Fig. 4h). Overall, our results demonstrate that SPEAR outperforms current state-of-the-art methods for predicting a response using multi-omic data.

It was notable that the balanced misclassification error of SPEAR was comparable with other methods even though SPEAR does not optimize purely for predictive performance. Rather, SPEAR favors multi-omic assay influence in the construction of predictive factors by choosing the largest w whose mean cross-validated error falls within one standard deviation of the overall minimum cross-validated error. If we instead choose the w that minimizes the cross-validated error (denoted as SPEAR *min*), the balanced misclassification error is better than all other models on both the TCGA-BC and COVID-19 datasets (0.13, 0.19) (Supplementary Fig. 7a, 7d). Similarly, this approach achieves the best AUROC values for all classes (Supplementary Fig. 7c, 7f). Since the SPEAR *min* model chooses a lower w value ($w = 0.5$, $w = 0.01$) than the default SPEAR model ($w = 2.0$, $w = 2.0$) (Supplementary Fig. 8) for both the TCGA-BC and COVID-19 datasets respectively, and because the AUROC values from SPEAR *min* were not significantly better than the default SPEAR model, we opted to use the default SPEAR model for downstream interpretation. Nevertheless, SPEAR *min* can be used in cases where prediction is the key objective regardless of multi-omic assay influence in factor construction.

3.4 SPEAR identifies steroid response pathways associated with TCGA PAM50 subtypes

Several factors returned by SPEAR were associated with different biological pathways that distinguish the PAM50 subtypes. Factors 1-3 clearly distinguished one or more subtypes, which

motivated us to investigate what biological pathways each factor was associated with (Fig. 5, Supplementary Fig. 5). Factor 1 was strongly associated with the Basal subtype and moderately associated with the HER2e subtype (Fig. 5a, Fig. 5b), while the negative factor loadings were enriched for Estrogen Response (ER) (Early and Late) pathways of the Molecular Signatures Database (MSigDB) Hallmark collection[17] (Fig. 5c, Fig. 5d).

The anti-correlation between the ER pathways and the Basal subtype, the highest scoring Factor 1 subtype, reflects that it is most strongly associated with a triple-negative profile for Estrogen, Progesterone, and the HER2e receptor[18]. Whereas HER2e-classified samples are predominantly hormone receptor (HR) negative, a proportion is also HR positive[19,20]. LumA and LumB are hormone receptor positive (HR) and retain expression of ER, and the Progesterone receptor (PR) in the case of LumA, and in some proportion of LumB[20]. Indeed, genes *XBP1*, *AGR2*, and *CA12* which yielded high posterior selection probabilities as well as the strongest negative projection coefficients in Factor 1 (Fig. 5e, Supplemental Table 1) for the Estrogen Response (Late) pathway are all associated with a breast cancer Steroid Responsiveness (SR) module that indicates functional steroid response[21]. Genes *XBP1* and *MYB* were in the top 20% of the ER Early and Late pathway genes with respect to the SPEAR projection coefficient magnitude. *MYB* is a direct target of Estrogen signaling and is overexpressed in most ER+ cancers[22], and both genes were also identified as part of a luminal expression signature[14].

The positive association of Factor 1 with the MYC Targets V1 Hallmark pathway is consistent with the association of MYC signaling and Basal subtypes identified in earlier TCGA and other analyses[23]. Additionally, four of the miRNAs associated with Factor 1, *hsa-mir-18a*, *hsa-mir-10a*, *hsa-mir135b*, *hsa-mir-577*, were identified in a miRNA analysis of TCGA data to have diagnostic significance for triple-negative breast cancer[24]. *Hsa-mir-18a* has been shown in independent datasets to downregulate ERα[25] and is associated with worse overall survival[26]. Hypomethylation of *MIA*, a PAM50 gene, which has the second highest in magnitude SPEAR projection coefficient for Factor 1, has been identified as associated with the Basal subtype in TCGA and independent data[27]. We did not exclude the PAM50 genes from the methylation results since the markers are used in gene expression assays.

Factor 1 distinguished Basal-like and HER2e subtypes from LumA/LumB, while Factors 2 and 3 distinguished LumA from the other subtypes. Pathways enriched in Factor 2 are associated with genes that are more highly expressed in LumA, whereas pathways in Factor 3 are associated with genes that are downregulated in LumA (Fig. 5a, Fig. 5b). The strongest signal in Factor 2 is in the Epithelial Mesenchymal Transition (EMT) Hallmark pathway, and in Factor 3 the G2M Checkpoint Hallmark pathway (Fig. 5c). It has been previously observed that LumB tumors are enriched in proliferation and cell-cycle associated genes in comparison to LumA[20], whereas the positive association of LumA with an EMT phenotype is somewhat surprising, as LumA tumors are generally considered to be associated with a stronger epithelial phenotype in comparison to the Basal-like subtype[28]. The role of EMT has been primarily studied in non-luminal tumors[28], and therefore the presence of this pathway distinguishing the subtypes warrants further investigation with respect to the underlying biology. Interestingly, the miRNAs with the top magnitude projection coefficients in Factor 2 and which have a positive association with the factor, *hsa-mir199a/b*, have been associated with EMT[29,30]. Thus, the SPEAR factorization of the TCGA-BC data has identified both well-documented and novel associations of multi-omic features with PAM50 breast cancer subtypes.

3.5 SPEAR identifies multi-omic factors and pathways associated with COVID-19 severity

On the COVID-19 dataset, SPEAR identified several factors that were significantly associated with the WHO ordinal severity score (Fig. 6, Supplementary Fig. 6). Upon investigating the association

of each factor with the COVID-19 severity, we observed that the Factor 2 score showed a positive ordinal association (Fig. 6a). SPEAR also identified several factors that were heavily associated with identifying specific COVID-19 severities, including Factor 1 score for mild severity and Factor 8 for moderate severity (Supplementary Fig. 5). Calculation of the variance explained for these three factors revealed that Factor 2 had the largest influence from both assays (16% proteomics, 11% metabolomics) compared to Factor 1 (20% proteomics, 1% metabolomics) and Factor 8 (1% proteomics, 2% metabolomics) (Supplementary Fig. 6). We opted to further investigate Factor 2 due to its larger multi-omic influences.

Proteomic enrichment analysis into Factor 2 identified several significant pathways from the MSigDB Hallmark pathway database (Supplemental Table 2). One pathway enriched in Factor 2 is the Janus kinase Signal Transducer and Activator of Transcription (JAK-STAT) signaling pathway[31] (Fig. 6c). Pathway member Interleukin 6 (IL-6) was found to be a key contributor to the Factor 2 score, with the second highest projection coefficient (Fig. 6e, Supplemental Table 3). IL-6 is a proinflammatory cytokine proposed as an inflammatory biomarker for COVID-19 severity[32] and is an activator of the JAK-STAT signaling pathway. The JAK-STAT signaling pathway is involved in immune regulation, lymphocyte growth and differentiation, and promotes oxidative stress, serving as an attractive therapeutic target for COVID-19 treatment[33].

Automatic feature selection of the top contributing metabolites to the Factor 2 score identified members of the amino acid, cofactors and vitamins, lipid, and nucleotide super-pathways (Fig. 6f). Kynurenine, a key contributor to the Factor 2 score, is a known marker of severe/fatal COVID-19 trajectory[34,35]. The tryptophan/kynurenine (Trp/Kyn) pathway is activated by inflammatory cytokines36, which is consistent with its positive correlation with the cytokines of Factor 2 such as IL-6 (Fig. 6f).

Several plasmalogens, including phosphatidylcholines (PCs) and phosphatidylethanolamines (PEs) were also found to be inversely associated with the ordinal severity trend of Factor 2 (Fig. 6f). Plasmalogens are plasma-borne antioxidant phospholipids that provide endothelial protection during oxidative stress[37], which could account for the inverse association with the JAK-STAT signaling pathway proteins. Our multi-omic results further support the utility of plasmalogens as prognostic indicators of COVID-19 severity[38].

4 Discussion

SPEAR extracts predictive factors from multi-omic datasets via supervised dimensionality reduction and can model multiple types of responses, including Gaussian, multinomial, and ordinal. We have compared SPEAR to state-of-art linear dimension reduction methods as well as the pure prediction model from lasso regression in both simulated and real datasets. In our evaluations, SPEAR effectively generates predictive factors that are also biologically interpretable. Unlike similar factor models such as MOFA and iClusterBayes, the supervised SPEAR framework will construct factors to emphasize a response of interest, emphasizing signals from the multi-omic assays when supported by the data. This is accomplished through the weight parameter w, which is automatically tuned across a gradient of values. SPEAR then favors higher weights, selecting the highest w within one standard deviation of the best performing w. By design, a high w ($w \geq 1$) will reconstruct predictive factor structure found in both the multi-omic assays and the response. Alternatively, a low w ($w = 0$) SPEAR model disregards the multi-omic assay factor structure entirely in pursuit of improved prediction, which may lead to less biologically interpretable factors. Ultimately, SPEAR stores the factors for all trained values of w, allowing for easy extraction and exploration of each model.

Specifying a reasonable number of factors can be critical for applying factor model-based methods. We have observed that underestimating the rank of a SPEAR model can lead to poor

performance at higher values of w. SPEAR begins by estimating the minimum number of factors required to accurately represent a multi-omic dataset. This adaptive factor number estimation approach was shown to be successful in our empirical study. This idea is not restrictive to SPEAR and can be applied to other dimension reduction methods to reduce the opportunities for human error. The procedure is outlined in Supplementary Methods A2.

While generalizable to many types of multi-omics, SPEAR does assume a linear dependence of the latent factors on the features. As such, the application of SPEAR to nonlinear analyte dependence may produce undesirable results. This could be mediated by appropriately preprocessing nonnormal multi-omic assays through transformations. A more data-adaptive solution would be to model the non-linear relationship between the features and factors via the generalized additive model[46], which we consider a future work and an extension to SPEAR.

In conclusion, the SPEAR model decomposes high-dimensional multi-omic datasets into interpretable low-dimensional factors with high predictive power without the need for parameter tuning. SPEAR returns both sparse (regression) and full (projection) coefficients as well as feature-wise posterior probabilities used to assign feature significance. SPEAR is currently hosted as a publicly available R-package at https://bitbucket.org/kleinstein/SPEAR.

Acknowledgements

We are grateful to Joann Diray-Arce for her assistance with the metabolomic enrichment analysis for the COVID-19 dataset. We also acknowledge Amrit Singh, Casey Shannon, and Scott Tebbutt for the helpful conversations and support with DIABLO.

Funding

This work was funded in part by NIH grant U19AI089992

Conflict of Interest: SHK receives consulting fees from Peraton. All other authors declare that they have no competing interests.

5.1 SPEAR Model Construction

Let X be an $N \times p$ matrix representing the concatenation of multiple multi-omics assays for N samples and p features and let Y be a length N vector representing the response of interest. We assume that X is driven by underlying low dimensional factors, coupled with unstructured Gaussian noise (E):

$$X_j = U\beta_j + E_j \text{ , for all } j = 1, \dots, p$$

The goal of latent factor model analysis is to decompose X into latent factors (U) and factor loadings (β_j for all j = 1, …, p). This has been previously accomplished by finding the Bayesian posteriors of U and β in the probability model[6] but has not yet been extended to work with a response of interest.

Let Y be a length N vector representing a univariate Gaussian response of interest (see Supplementary Methods C for ordinal, multinomial, and binomial). Like X, let Y also be modeled by a subset of the estimated factors ($U_{||}$):

$$Y = U^{||}\beta_0 + \varepsilon$$

That is, we decouple $U = U^{||} + U^{\perp}$ where $U^{||}$ represents the relevant factors for predicting Y and $U^{\perp}$ are irrelevant. In some scenarios when sample size (N) is limited or the signal sizes of $U^{||}$ on X are not sufficiently strong, it can be difficult to estimate all factors $U^{||}$ that are important for

understanding the response Y. SPEAR addresses this problem by modeling $U = ZB$ as a linear function of some covariates Z, e.g., $Z \leftarrow X$, and considering a weighted joint probability for X and Y:

$$P^w(X, Y, \beta, B|\Theta) = P^w(X|\beta, B, \Theta, Z) \times P(Y|\beta, B, \Theta, Z) \times P(\beta, B|\Theta, Z)$$

SPEAR prioritizes the construction of $U^{||}$ before $U^{\perp}$ by adjusting the importance of the fit towards X through the weight parameter w. A low w ($w \approx 0$) will yield factors that emphasize prediction of Y, while higher values for w ($w \geq 1$) will construct factors that better explain the structure of X.

5.2 Selection of SPEAR Model Rank

The model rank, or number of factors to construct, need to be determined by the user, and proper choices of them are data dependent. Alternative factor models utilize a static number such as 15 or 20 factors as a default. This is problematic in predictive settings, as underestimating the required number of predictive factors will result in worse performance. To this end, SPEAR begins by suggesting a sufficient model rank by estimating how the largest number estimable factors from the multi-omic assays whose influence on the data could not be accounted for by pure noise, based on a data randomization approach used in the regression problem[39,40]. This procedure is explained in detail in Supplementary Methods A2.

5.3 TCGA-BC Dataset Background and Preprocessing

The TCGA-BC dataset was adapted from *Singh et al.*[10], consisting of 16851 mRNAs, 349 miRNAs, and 9482 CpG methylation sites. Each sample was taken from a primary solid breast cancer tumor that has been classified according to the PAM50 subtype signature, a 50-gene signature. The PAM50 signature is one of the main intrinsic breast cancer signatures used in clinical practice to determine the course of therapy[41,42]. *Singh et al.* showed that even when the 50 genes were removed from the dataset, it was still possible to predict subtype class using the remaining multi-omic features, indicating that there are underlying biological pathways that can distinguish and drive the subtypes.

While the majority of the preprocessing was kept as described in *Singh et al.*, the least 20% variable features from each assay were removed to reduce the total number of features. Samples were then split into train ($n_{train} = 379$) and test ($n_{test} = 610$) sets as described in *Singh et al*. Finally, all three assays (mRNA, CpG, and methylation) were scaled and centered to be normally distributed ($\mu = 0, \sigma^2 = 1$).

5.4 COVID-19 Dataset Background and Preprocessing

The SARS-CoV-2 (COVID-19) dataset contains multi-omic data from 254 samples from SARS-COV-2 positive patients and 124 samples from matched healthy subjects from *Yapeng et. al*[13]. Samples were taken from one of two timepoints, T1 and T2. At each timepoint, participants were assigned a COVID-19 severity score based on a World Health Organization (WHO) ordinal scale for clinical improvement: uninfected (0), ambulatory without (1) and with activity limitation (2), hospitalized without (3) and with oxygen therapy (4), hospitalized with non-invasive ventilation (5), intubation (6) or ventilation with additional organ support (7) and death (8). These ordinal values were condensed into four classes: healthy (0), mild (1-2), moderate (3-4), and severe (5-7).

MissForest, a non-parametric missing value imputation for mixed-type data was utilized to address missing values in both the proteomics and metabolomics[43]. The metabolite expression values required quantile normalization and log transformation. Both proteomic and metabolomic expression values were then scaled and centered to be normally distributed ($\mu = 0, \sigma^2 = 1$).

6 Figures

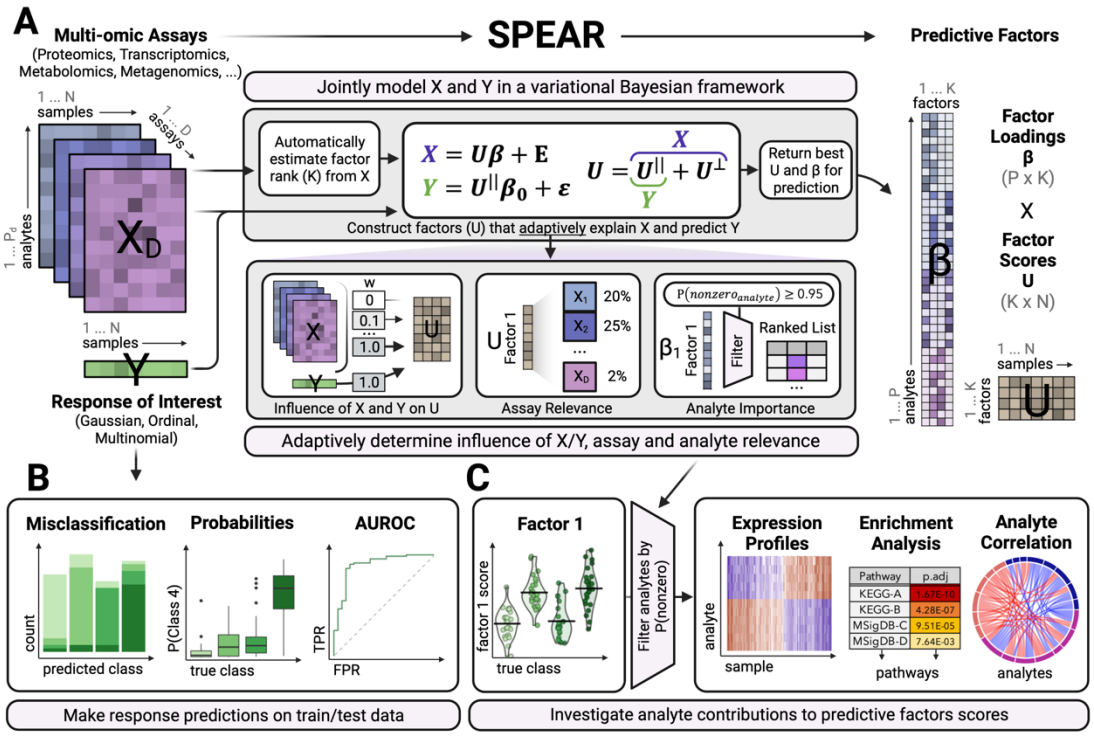

Figure 1: SPEAR workflow overview: (**A**) SPEAR takes multi-omic data (X) taken from the same N samples, as well as a response of interest (Y). SPEAR supports Gaussian, ordinal, and multinomial types of responses. From these inputs, the algorithm first automatically estimates the minimum number of factors to use in the SPEAR model. X and Y are then jointly modeled in a variational Bayesian framework to adaptively construct factor loadings (β) and scores (U) that explain variance of X and are predictive of Y. (**B**) SPEAR factors are used to predict Y and provide probabilities of class assignment for ordinal and multinomial responses. (**C**) Downstream biological interpretation of factors is facilitated via automatic feature selection, expression profile analysis, enrichment analysis, and analyte correlation.

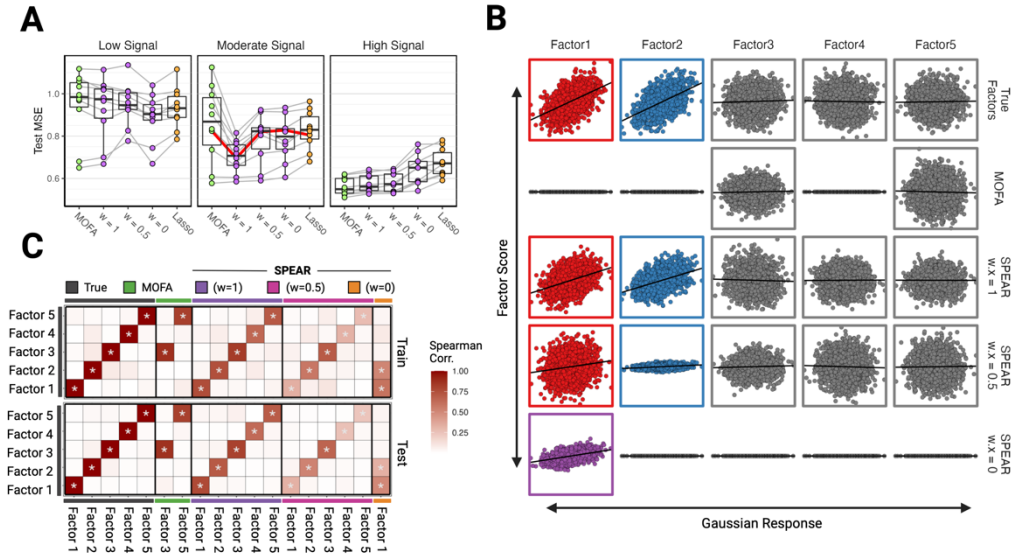

Figure 2: Gaussian simulation results. (**A**) Boxplots of mean-squared errors of the models on the testing data. MSE results for each simulated iteration are connected. Results are shown for varying signal-to-noise ratios, including low, moderate, and high signals. (**B**) Scatterplots of various factor scores (y-axis) against the true Gaussian response (x-axis) of the moderate signal test data. Color is applied to factor scores found to be correlated with true factors 1 and 2. (**C**) Correlation matrix showing the Spearman correlation between each derived factor of the true factors for both the training and testing data of the moderate signal simulated dataset. Significant correlations are denoted with '*' for $p \leq 0.001$.

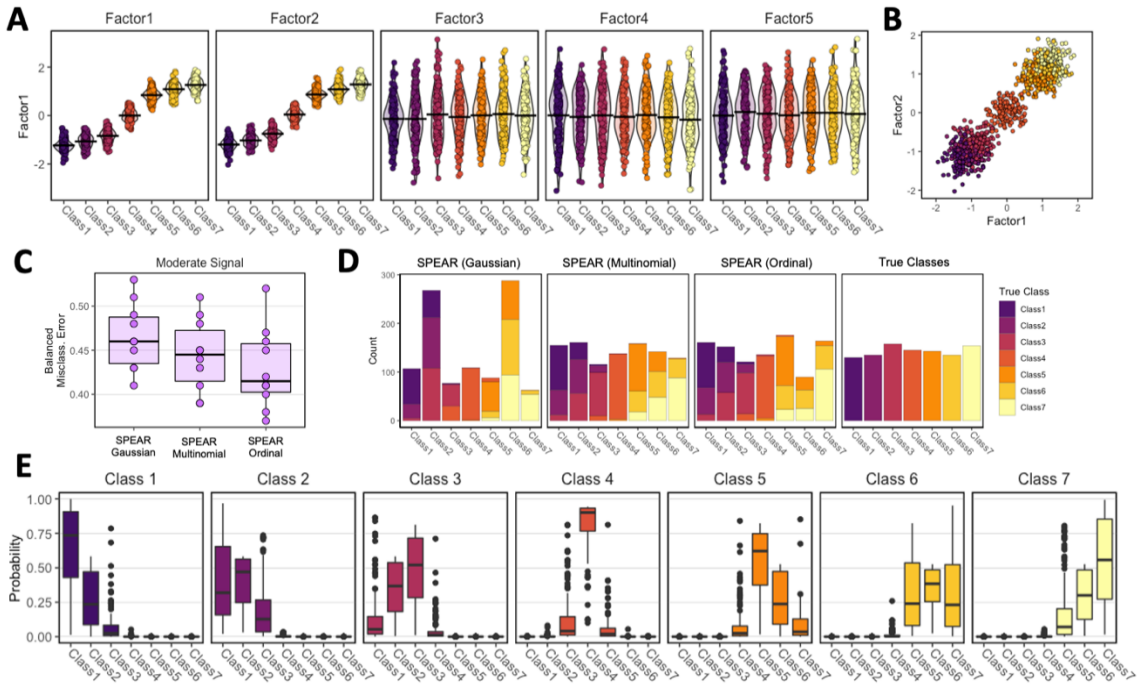

Figure 3: Ordinal simulation results. (**A**) Grouped violin plots of the simulated samples by true factor scores (y-axis) and their ordinal response class (x-axis). (**B**) Scatter plot of simulated samples by true factor 1 score (x-axis) and true factor 2 score (y-axis). (**C**) Grouped boxplots of balanced misclassification error results from SPEAR models on simulated test data. SPEAR models are differentiated by how the response was treated. (**D**) Histogram of simulated test data ordinal class predictions. Plots are sorted by predicted class (x-axis) and colored by true class. (**E**) Boxplots of ordinal class assignment probabilities derived from the SPEAR ordinal model. Samples are separated by their true ordinal class.

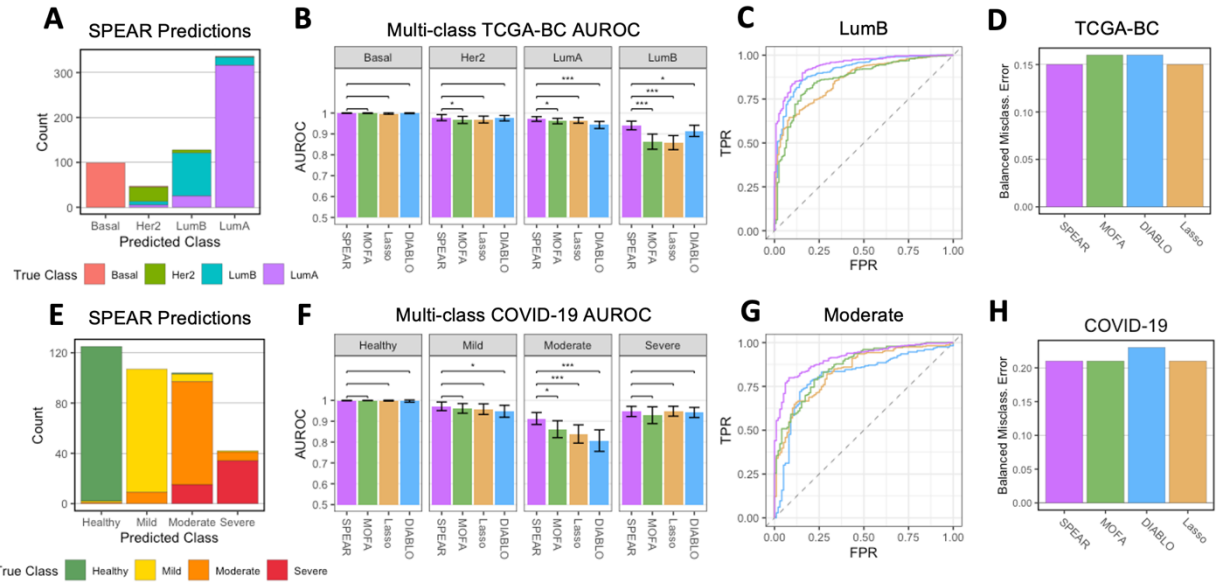

Figure 4: TCGA-BC Tumor Subtype and COVID-19 Severity Prediction Results. (**A, E**) Test sample class predictions of the SPEAR model, colored by true class. (**B, F**) Multi-class AUROC statistics for each model. Error bars show the 95% confidence interval found via 2,000 stratified bootstrapping replicates. Significance testing is denoted as * ($p \leq .05$), ** ($p \leq .005$), and *** ($p \leq .0005$). (**C**) AUROC plot for all models predicting LumB subtype. (**G**) AUROC plot for all models predicting the moderate severity class. (**D, H**) Balanced misclassification errors of SPEAR, MOFA, DIABLO, and Lasso on test samples from the (**D**) TCGA (Breast Cancer) dataset and (**H**) COVID-19 dataset.

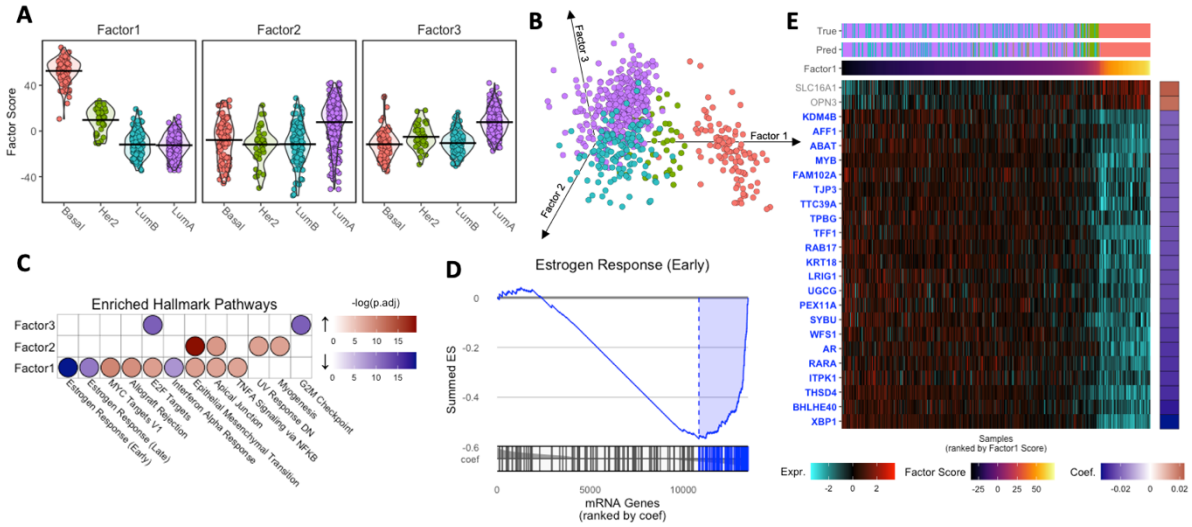

Figure 5: Downstream TCGA-BC Analysis. (**A**) Grouped violin plot of factors 1-3 scores (y-axis) and tumor subtype (x-axis), with group means marked with a line. (**B**) 3D scatter plot, embedding samples by factors 1, 2, and 3 scores. Samples are colored by tumor subtype. (**C**) Dotplot of GSEA results on mRNA features for factors 1-3. Points are shaded by -log(p.adjusted) with color representing enrichment direction. (**D**) GSEA plot for Estrogen Response (Early) Hallmark pathway for SPEAR Factor 1. mRNA genes are ranked by their assigned projection coefficient from SPEAR factor 1. (**E**) Heatmap showing normalized expressions for the top 24 mRNA genes involved in the Estrogen Response (Early) Hallmark pathway. mRNA genes were selected with a factor loading (projection coefficient) magnitude $\geq .02$. Samples were ranked by Factor 1 score (x-axis) and genes were ranked by projection coefficient (y-axis). Also shown are corresponding true tumor subtypes (True) and SPEAR-predicted tumor subtypes (Pred).

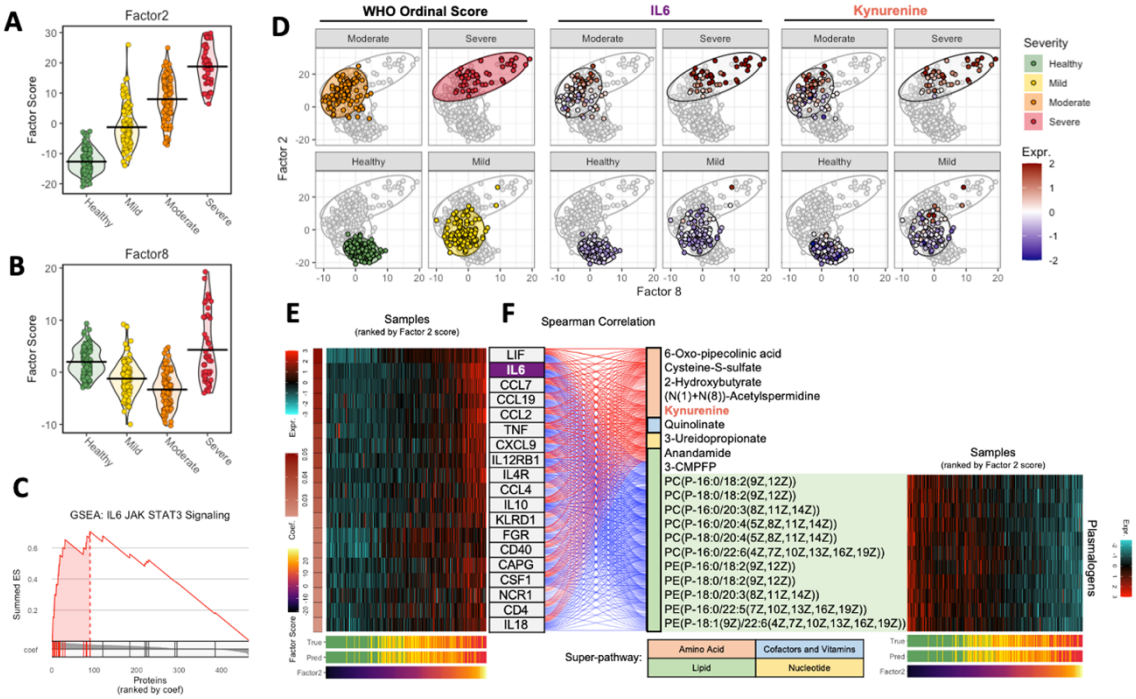

Figure 6: **Downstream COVID-19 Dataset Analysis**. (**A**) Grouped violin plot of factor 2 scores (y-axis) and simplified WHO score (x-axis), with group means marked with a line. (**B**) Grouped violin plot of factor 2 scores (y-axis) and simplified WHO score (x-axis), with group means marked with a line. (**C**) GSEA plot for IL6 JAK STAT3 Signaling Hallmark pathway for SPEAR Factor 2. Proteins are ranked by their assigned projection coefficient from SPEAR factor 2. (**D**) Embedding of samples by factor 8 (x-axis) and factor 2 (y-axis) scores. Samples are colored by WHO Ordinal Score, normalized IL6 expression, and normalized kynurenine expression. (**E**) Heatmap showing normalized expressions for proteins involved in the IL6 JAK STAT3 Signaling Hallmark pathway. Samples were ranked by Factor 2 score (x-axis) and proteins were ranked by projection coefficient (y-axis). Also shown are corresponding true patient severity scores (True) and SPEAR (sd) predicted severity scores (Pred). (**F**) Alluvial plot showing correlation between the IL6 JAK STAT3 proteins and the top metabolite features contributing to SPEAR Factor 2. Metabolites are grouped by positive/negative correlation and super-pathway. Also shown are normalized plasmalogen expressions of samples ranked by Factor 2 score (x-axis).

References

1. Bhattacharya, S. et al. ImmPort: disseminating data to the public for the future of immunology. Immunol. Res. 58, 234–239 (2014).
2. Bolen, C. R., Ding, S., Robek, M. D. & Kleinstein, S. H. Dynamic expression profiling of type I and type III interferon-stimulated hepatocytes reveals a stable hierarchy of gene expression. Hepatology 59, 1262–1272 (2014).
3. Lazear, H. M., Schoggins, J. W. & Diamond, M. S. Shared and Distinct Functions of Type I and Type III Interferons. Immunity 50, 907–923 (2019).
4. Banoth, B. & Cassel, S. L. Mitochondria in innate immune signaling. Transl Res 202, 52–68 (2018).
5. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat. Commun. 12, 124 (2021).
6. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 21, 111 (2020).
7. Mo, Q. et al. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostat. Oxf. Engl. 19, 71–86 (2018).
8. Wipf, D. & Nagarajan, S. A New View of Automatic Relevance Determination. in Advances in Neural Information Processing Systems vol. 20 (Curran Associates, Inc., 2007).
9. Li, W., Zhang, S., Liu, C.-C. & Zhou, X. J. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinforma. Oxf. Engl. 28, 2458–2466 (2012).
10. Singh, A. et al. DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinforma. Oxf. Engl. 35, 3055–3062 (2019).
11. Tenenhaus, A. et al. Variable selection for generalized canonical correlation analysis. Biostat. Oxf. Engl. 15, 569–583 (2014).
12. Furman, D. & Davis, M. M. New approaches to understanding the immune response to vaccination and infection. Vaccine 33, 5271–5281 (2015).
13. Su, Y. et al. Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. Cell 183, 1479-1495.e20 (2020).
14. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).
15. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 27, 1160–1167 (2009).
16. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics 44, 837–845 (1988).
17. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 1, 417–425 (2015).
18. Prat, A. et al. Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. The Oncologist 18, 123–133 (2013).
19. Bastien, R. R. L. et al. PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. BMC Med. Genomics 5, 44 (2012).
20. Prat, A. et al. Clinical implications of the intrinsic molecular subtypes of breast cancer. Breast Edinb. Scotl. 24 Suppl 2, S26-35 (2015).
21. Fredlund, E. et al. The gene expression landscape of breast cancer is shaped by tumor protein p53 status and epithelial-mesenchymal transition. Breast Cancer Res. BCR 14, R113 (2012).
22. Gonda, T. J., Leo, P. & Ramsay, R. G. Estrogen and MYB in breast cancer: potential for new therapies. Expert Opin. Biol. Ther. 8, 713–717 (2008).
23. Xu, J., Chen, Y. & Olopade, O. I. MYC and Breast Cancer. Genes Cancer 1, 629–640 (2010).
24. Fan, C. & Liu, N. Identification of dysregulated microRNAs associated with diagnosis and prognosis in triple-negative breast cancer: An in silico study. Oncol. Rep. 41, 3313–3324 (2019).
25. Klinge, C. M. miRNAs and estrogen action. Trends Endocrinol. Metab. TEM 23, 223–233 (2012).
26. Luengo-Gil, G. et al. Clinical and biological impact of miR-18a expression in breast cancer after neoadjuvant chemotherapy. Cell. Oncol. Dordr. 42, 627–644 (2019).
27. Bardowell, S. A. et al. Differential methylation relative to breast cancer subtype and matched normal tissue reveals distinct patterns. Breast Cancer Res. Treat. 142, 365–380 (2013).
28. Felipe Lima, J., Nofech-Mozes, S., Bayani, J. & Bartlett, J. M. S. EMT in Breast Carcinoma-A Review. J. Clin. Med. 5, E65 (2016).
29. Drago-García, D., Espinal-Enríquez, J. & Hernández-Lemus, E. Network analysis of EMT and MET micro-RNA regulation in breast cancer. Sci. Rep. 7, 13534 (2017).
30. Wang, Q. et al. Overview of microRNA-199a Regulation in Cancer. Cancer Manag. Res. 11, 10327–10335 (2019).
31. Seif, F. et al. The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. Cell Commun. Signal. 15, 23 (2017).
32. Azevedo, R. B. et al. Covid-19 and the cardiovascular system: a comprehensive review. J. Hum. Hypertens. 35, 4–11 (2021).
33. Luo, W. et al. Targeting JAK-STAT Signaling to Control Cytokine Release Syndrome in COVID-19. Trends Pharmacol. Sci. 41, 531–543 (2020).
34. Danlos, F.-X. et al. Metabolomic analyses of COVID-19 patients unravel stage-dependent and prognostic biomarkers. Cell Death Dis. 12, 1–11 (2021).
35. Mangge, H. et al. Increased Kynurenine Indicates a Fatal Course of COVID-19. Antioxidants 10, 1960 (2021).
36. Almulla, A. F. et al. The tryptophan catabolite or kynurenine pathway in COVID-19 and critical COVID-19: a systematic review and meta-analysis. BMC Infect. Dis. 22, 615 (2022).
37. Messias, M. C. F., Mecatti, G. C., Priolli, D. G. & de Oliveira Carvalho, P. Plasmalogen lipids: functional mechanism and their involvement in gastrointestinal cancer. Lipids Health Dis. 17, 41 (2018).
38. Pike, D. P. et al. Plasmalogen Loss in Sepsis and SARS-CoV-2 Infection. Front. Cell Dev. Biol. 10, (2022).
39. Guan, L. & Tibshirani, R. Post model-fitting exploration via a “Next-Door” analysis. Can. J. Stat. 48, 447–470 (2020).
40. Tian, X. Prediction error after model search. Ann. Stat. 48, 763–784 (2020).
41. Wallden, B. et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med. Genomics 8, 54 (2015).
42. Goldhirsch, A. et al. Personalizing the treatment of women with early breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2013. Ann. Oncol. Off. J. Eur. Soc. Med. Oncol. 24, 2206–2223 (2013).
43. Stekhoven, D. J. & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118 (2012).
44. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: Springer (2009).
45. Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288 (1996).
46. Hastie, T. J. Generalized additive models. In Statistical models in S (pp. 249-307). Routledge (2017).
47. Rubio-Rivas, M. et al. WHO Ordinal Scale and Inflammation Risk Categories in COVID-19. Comparative Study of the Severity Scales. Journal of general internal medicine, 37(8), 1980–1987 (2022).
48. Friedman J, Hastie T, Tibshirani R. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” Journal of Statistical Software, 33(1), 1–22 (2010).

49 Tenenhaus, A. & Tenenhaus, M. Regularized Generalized Canonical Correlation Analysis. Psychometrika 76, 257–284 (2011).Bardet, G. (1920) Sur un syndrome d’obesite infantile avec polydactylie et retinite pigmentaire (contribution a l’etude des formes cliniques de l’obesite hypophysaire). PhD Thesis, name of institution, Paris, France.