Multi-omics factorization illustrates the added value of deep learning approaches

Gautam Machiraju
Biomedical Informatics
Training Program,
Stanford University
gmachi@stanford.edu

David Amar
Dept. of Medicine,
Stanford University
davidama@stanford.edu

Euan Ashley
Depts. of Medicine, Genetics,
Biomedical Data Science,
Stanford University
euan@stanford.edu

1 Introduction

The biomedical community is increasingly collecting and analyzing multi-omics datasets [14] — heterogeneous panels of biological (i.e. *omics*) measurements ranging from genomics to metabolomics. Compounded with cheaper costs of sequencing and experimental assays, multi-omics studies are designed with the goal of robustly detecting potential biomarkers (e.g. genes, proteins, etc.) for states of interest (e.g. disease). The primary assumption of such studies is that the omic features are systemically interacting with one another in complex, often nonlinear [22] intra- and inter-omic networks, and that a patient observation of thousands of signals can serve as a rich snapshot of their state. Furthermore, the analysis of this broader system of networks at multiple measurement time points may coax out key drivers that are differentially expressed between states. Manifold learning of patient state spaces has large implications for anomaly detection and health forecasting.

However, there exist many challenges to integrating these datasets for downstream analyses (patient subgrouping, survival analysis, etc.), particularly due to feature-feature nonlinearities and their collective heterogeneity (varying dynamic ranges, underlying distributions, etc.), noise, high-dimensionality ($p \gg n$), and subsequent sparsity. Naturally, multi-omics data integration, usually taking the form of unsupervised, dimensionality reduction methods [12] [13], is a growing area of methods development with recent advances in group factor and neural approaches in the past year. While linear methods such as Principle Components Analysis (PCA) and Factor Analysis (FA) perform well on single omic data, these methods fail to acknowledge (1) groupings of features by omic and (2) nonlinear relationships between omic features. Group factor methods acknowledge omic heterogeneity and harness that to learn potentially richer encodings. Through our investigation, we show that state-of-the-art linear methods such as Multi-Omics Factor Analysis (MOFA) [5] present issues with dominating omics signals in projected data, motivating nonlinear approaches for group factor analysis. Accordingly, this study presents a series of experiments with Variational Autoencoders against multiple baselines performing on multi-omic datasets, analyzing the meaning of their latent features. Furthermore, we aim to provide much-needed benchmarking measures for such multi-omics integration methods [15]. Through the use of a human longitudinal omics profiling and large cancer multi-omics datasets, multiple dimensionality reduction methods (PCA, FA, Autoencoders, etc.) are studied to compute encoded loadings and dimension-reduced representations of these data for interpretation through correlation analysis and other statistical techniques. Finally, this study investigates the efficacy of nonlinear and neural group factor methods as interpretable approaches to omics integration and an improvement upon the biomarker discovery process.

2 Related Work

Current state-of-the-art methods such as Multi-Omics Factor Analysis (MOFA) [5] aim to find relevant latent variables via linear group factor analysis, but still lack the extension to nonlinear feature-feature interactions. With an increasing amount of computational resources, as well as theoretical results describing shared subspaces between PCA and linear Autoencoders (AE) [6] [17] [11], nonlinear AEs are increasingly being argued as natural extensions to linear dimensionality reduction approaches. Within the last year, their application in multi-omics has also been increasing in popularity [25] [24] [8] with the detection of nonlinear feature-feature relationships. However, unlike MOFA, this approach fails to acknowledge grouped features and instead simply concatenates these disparate data types before encoding for latent variables. After the development of Variational Autoencoders (VAE) [27] [23], recent works in variational inference in the grouped factor context have brought about the *Stacked Variational Autoencoder* (sVAE) [7] [4] for group factor analysis. Subsequently, an early application of an sVAE on heterogeneous multi-omics datasets

Preprint. Under review.

called Multi-omics Autoencoder Integration (MAUI) [20] has shown improved performance in latent representational quality in downstream analyses and interpretability. There is a growing need to interpret feature-feature loadings for these nonlinear, deep learning approaches.

3 Datasets & Features

We utilize three longitudinal datasets: the Integrated Personal Omics Profiling (iPOP) [9] and the NIH's The Cancer Genome Atlas (TCGA) [2]. iPOP contains approximately 600 samples over 100 individuals, each with over 15K features comprised of transcriptomics, proteomics, metabolomics, and clinical data. The subset of TCGA used for this study contains approximately 600 samples with 1300 features comprised of gene expression, point mutations, and copy number variation (CNV). While iPOP is longitudinal, autocorrelation was observed to be fairly low, allowing for the assumption to treat these data as independent and identically distributed (i.i.d.) within respective omics samples for the purposes of capturing distinct states. Preprocessing this numeric data included normalization among each omics and overall, followed by filtering out basally-expressed markers using the coefficient of variation. We express our design matrix $\mathbf{X}$ by *omics groups* $\mathbf{X}_{\{j\}} \in \mathbb{R}^{n \times p_j}\ \forall j = 1 \dots m$ as seen in Figure 1:

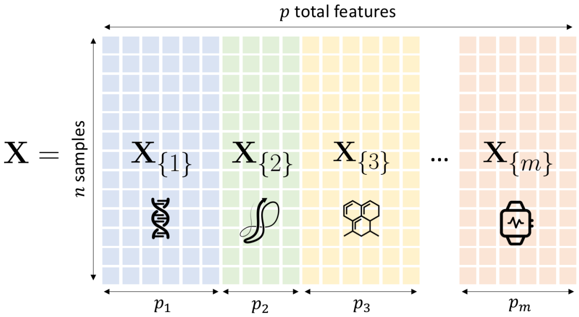


Figure 1: An example design matrix of a multi-omics dataset, in which the features are denoted by different colored data groupings, representing their omic types. This example contains transcriptomic, proteomic, metabolomic, and clinical data.

4 Methods

To evaluate performance against the grouped neural approach of the sVAE (as implemented in MAUI), this study compared the sVAE with several baselines, including linear ungrouped methods (PCA, FA, linear-transfer AE), a linear grouped method (group factor analysis via MOFA), and a nonlinear neural method (ReLU-transfer AE). Dimensionality reduction methods can be expressed in terms of their compression, as we show in this section. Given a design matrix $\mathbf{X}$, linear ungrouped methods like PCA and FA perform the following reconstruction:

$$\begin{aligned} \hat{\mathbf{X}} &= f(\mathbf{X}) \approx \mathbf{X} \\ \text{s.t.} \quad f &\triangleq [\mathbf{W} Z(\mathbf{X})]^\intercal \end{aligned}$$

where f denotes the reconstruction function, $\mathbf{W} \in \mathbb{R}^{p \times L}$ are the loadings or weights of the reconstruction, and $Z(\cdot) : \mathbb{R}^{n \times p} \to \mathbb{R}^{L \times n}$ is a matrix function that compresses our design matrix. For example, in PCA, $Z(\cdot)$ is the eigendecomposition of the design matrix, where $Z(\mathbf{X})$ make up the principle components computed from that decomposition. MOFA's use of group factor analysis simply extends this architecture to the omics groups of the design matrix in Figure 1:

$$\begin{aligned} \hat{\mathbf{X}}_{\{j\}} &= f(\mathbf{X}_{\{j\}}) \approx \mathbf{X}_{\{j\}} \\ \text{s.t.} \quad f &\triangleq [\mathbf{W}_{\{j\}} Z(\mathbf{X}_{\{j\}})]^\intercal \end{aligned}$$

where $\mathbf{W}_{\{j\}} \in \mathbb{R}^{p_j \times L}$ are the j^{th} set of loadings and $Z(\cdot) : \mathbb{R}^{n \times p_j} \to \mathbb{R}^{L \times n}$ is the j^{th} matrix function that compresses omics group $\mathbf{X}_{\{j\}}$. In the case of AEs, the reconstruction model looks similar to the linear architectures above. For a single training example $\mathbf{x}^{(i)} \in \mathbb{R}^p\ \forall i = 1 \dots n$:

$$\begin{aligned} \hat{\mathbf{x}}^{(i)} &= f(\mathbf{x}^{(i)}) \approx \mathbf{x}^{(i)} \\ \mathbf{h}^{(i)} = z(\mathbf{x}^{(i)}) &= \psi(\mathbf{W}_{\text{enc}} \cdot \mathbf{x}^{(i)} + \mathbf{b}_{\text{enc}}) \\ \text{s.t.} \quad f[z(\mathbf{x}^{(i)})] &\triangleq \phi(\mathbf{W}_{\text{dec}} \cdot \mathbf{h}^{(i)} + \mathbf{b}_{\text{dec}}) \end{aligned}$$

where the encoder $z(\cdot) : \mathbb{R}^p \to \mathbb{R}^L$ is a vector function that compresses our training example with the use of nonlinear activation function, ψ, and $\mathbf{W}_{\text{enc}} \in \mathbb{R}^{p \times L}$ as the weights of encoding $\mathbf{x}^{(i)}$ into a reduced L-dimensional space. The decoder f is the reconstructing vector function that utilizes nonlinearity ϕ and $\mathbf{W}_{\text{dec}} \in \mathbb{R}^{L \times p}$ as the weights of decoding $\mathbf{h}^{(i)}$ back into a p-dimensional space. Through enforcement of $L \ll p$, AEs have an undercomplete hidden layer $\mathbf{h}^{(i)} = z(\mathbf{x}^{(i)})$ and thereby learn vector functions outside of the identity map. VAEs and sVAEs use a similar architecture and a generative framework to learn means (μ) and standard deviations (Σ) of the data, with the latter learning the aforementioned distributional statistics for each omic group. Figure 2 depicts the sVAE architecture:

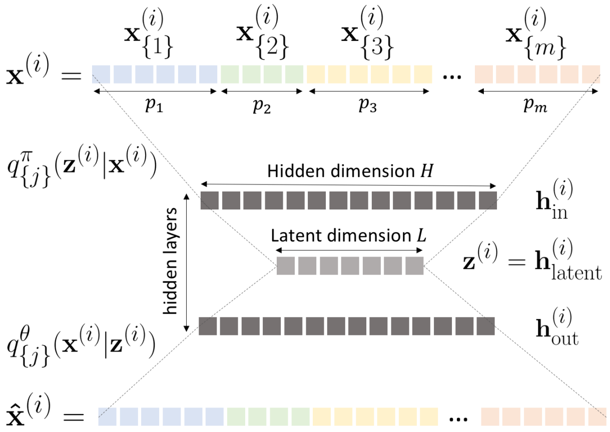

$\mathbf{x}^{(i)}_{\{1\}}$
$\mathbf{x}^{(i)}_{\{2\}}$
$\mathbf{x}^{(i)}_{\{3\}}$
$\mathbf{x}^{(i)}_{\{m\}}$
$\mathbf{x}^{(i)} =$
...
p_1
p_2
p_3
p_m
$q^{\pi}_{\{j\}}(\mathbf{z}^{(i)}|\mathbf{x}^{(i)})$
Hidden dimension H
$\mathbf{h}^{(i)}_{\text{in}}$
hidden layers
Latent dimension L
$\mathbf{z}^{(i)} = \mathbf{h}^{(i)}_{\text{latent}}$
$\mathbf{h}^{(i)}_{\text{out}}$
$q^{\theta}_{\{j\}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i)})$
$\hat{\mathbf{x}}^{(i)} =$
...

Figure 2: sVAE architecture.

where the approximative distribution $q^{\pi}_{\{j\}}(\cdot)$ generates $\mu^{\mathbf{z}|\mathbf{x}}_{\{j\}}$ and $\Sigma^{\mathbf{z}|\mathbf{x}}_{\{j\}}$, while $q^{\theta}_{\{j\}}(\cdot)$ generates $\mu^{\mathbf{x}|\mathbf{z}}_{\{j\}}$ using variational inference measures such as Kullback-Leibler Divergence ($\mathbb{D}_{\text{KL}}$) as a term in the objective function [4] [5] [20]:

$$\mathbb{E}_{q^{\pi}(\mathbf{z}|\mathbf{x})}[\ln q^{\theta}(\mathbf{x}|\mathbf{z})] - \mathbb{D}_{\text{KL}}[q^{\pi}(\mathbf{z}|\mathbf{x})||q^{\theta}(\mathbf{z})]$$

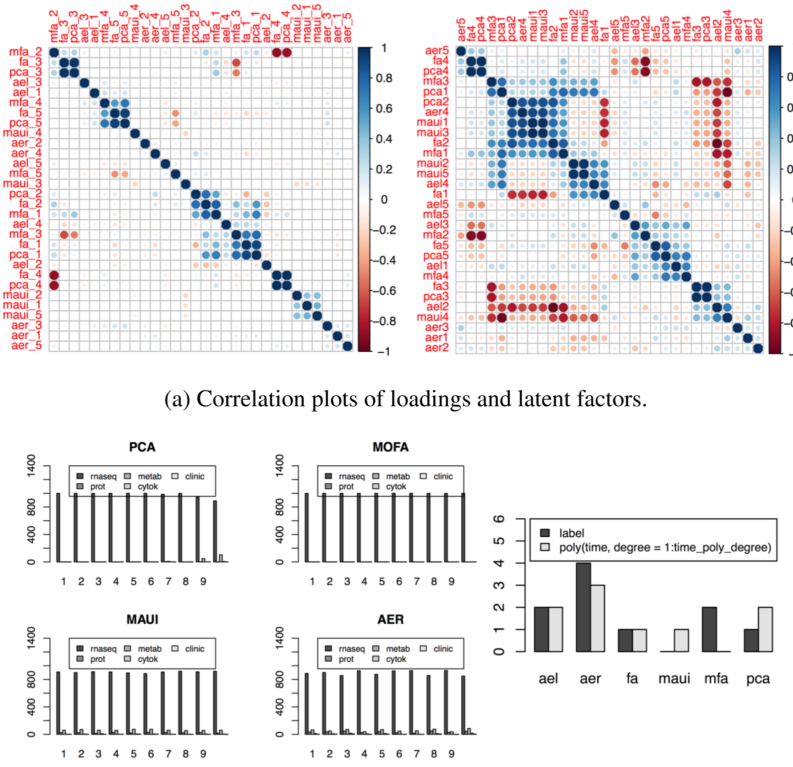
5 Experiments, Results, & Discussion

All neural models, including the MAUI implementation, were developed in Keras and trained using an Adam optimizer with learning rate of $\alpha = 1 \times 10^{-4}$, which tends to be fairly robust and does not require significant tuning. We trained using a mini-batch size of 50, which was a size large enough to accurately generate gradients, but small enough to fit in CPU memory. Due to training demands for both the AEs, sVAE, and GFA, I trained this pipeline of models on Stanford's Sherlock CPU compute clusters. For training variational inference models (sVAE and MOFA), the Kullback-Leibler Divergence was used as a term in the objective function. For both AE models (linear and ReLU transfer function variants), an MSE loss was used. MOFA was run over 1000 iterations with $\Delta\,\text{ELBO} := 0.1$, while the sVAE and AE models were all run for 300 epochs.The batch-size for MOFA, sVAE, and VAE were all set to 50. The latent variable dimensionality L was treated as a hyperparameter used to study its effect on compression methods. After training and model selection through hyperparameter tuning, we perform (1) correlation analysis between loadings for statistical benchmarking between the different learned representations, (2) analysis of omics *mixing* in these learned representations, and (3) downstream clustering and classification tasks on the transformed data to determine if we could predict health status of patients.

5.1 Correlation analysis

iPOP For this dataset, a latent dimension of $L = 20$ was chosen due to the high dimensionality of the feature space. After training our models, correlation analysis was performed on both the loadings and latent factors, resulting in similarity matrices based on the Pearson correlations and plotting them pairwise (as seen in Figure 3a). Through these visualizations, we observe that neural methods have loadings that are uncorrelated with the linear methods, while most methods' top factors are seem to be correlated. This seems to imply the learning of very different representations of the feature space.

Additionally, *omics mixing*, or the levels of contribution of each omic type toward a latent factor, was analyzed. This was done by mapping the latent factors back to features, giving us a sense of how many measurements of each omic type are primarily correlated with a particular latent factor. Specifically, given our loadings $\mathbf{W}$: for each factor, we sorted the loading by their absolute score, took the top $K = 1000$ features and their omics types, and counted how many were selected for each omic type. Through this analysis, we can see that all methods fail to mix omics. They seem to be utilizing similar omics proportions in their encodings with RNA-seq as the dominant omic group. Furthermore, we performed a correlation analysis based on factor versus label or time. The loadings of the ReLU-activated AE are most correlated with time and label compared to other methods. Associations were estimated with Linear Mixed Models with fixed effects and up to a 5-degree polynomial for time, where labels were binarized (infected vs. healthy) and time was measured in days passed from last infection for each subject. These results can be seen in Figure 3b.



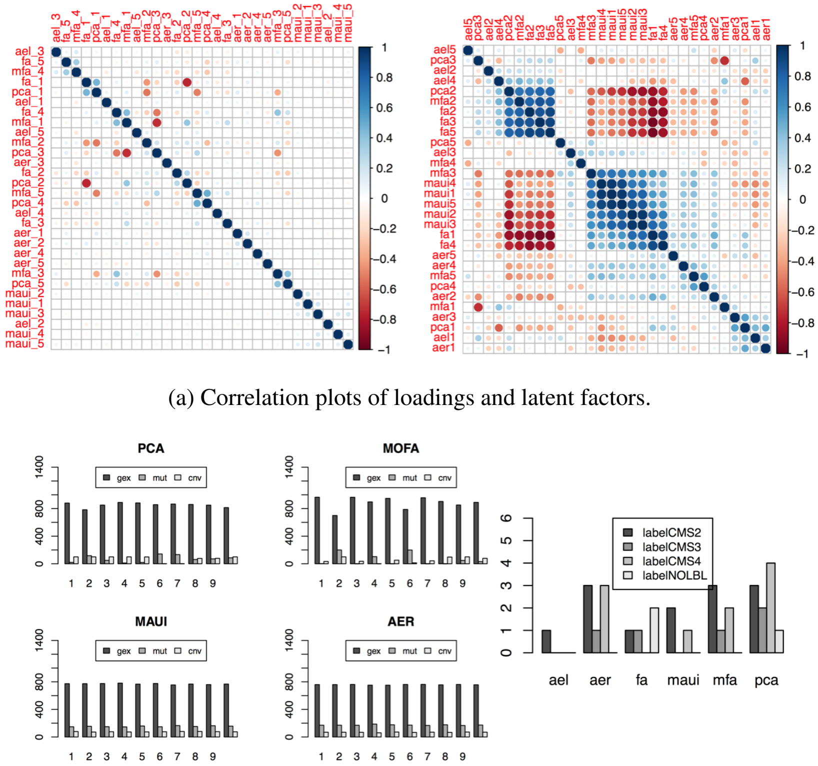
(a) Correlation plots of loadings and latent factors.

(b) Omics mixing per latent factor and correlation of latent factors with label and time.

TCGA For this dataset, a latent dimension of $L = 30$ was chosen due to the realtively lower dimensionality of the feature space. Correlation analysis shows that neural methods again have loadings that are uncorrelated with the linear methods and that again most methods' top latent factors are correlated. These results can be seen in Figure 4a. Omics mixing shows in this case that neural methods mix their omics better than the linear methods, including a current state-of-the-art method like MOFA. Even while RNA-seq is still dominant, all methods here seem to be using different omics in their encodings. Interestingly enough, we observe that correlation analysis based on factor versus label is highest in PCA, with MOFA and ReLU-AE also realtively high. These results can be seen in Figure 4b.

5.2 Downstream prediction tasks

To assess predictive power of the latent factors (transformed data), we performed classification (Support Vector Machine with regularization) and clustering (K-means with AMI-optimized clusters). This resulted in poor multi-class AUROC for all models (approximately 0.50-0.54) when computed for both iPOP and TCGA datasets. This result also held with a latent dimension $L = 70$ for TCGA. There may be multiple reasons for this low predictive performance, including the assumed qualitative separation of multi-class classification, whereas binary classification may perform better. Furthermore, both high feature dimension (p) and batch effects in iPOP's longitudinal study design may lead to low signal to noise ratio. Finally, by forcing all methods to have same latent dimension L for the sake of comparison may hinder or benefit each mode's performance differently. In temrs of clustering, both datasets observed a higher number of predicted clusters than that of label types, potentially hinting at transition states. Other analyses of our models included scatterplots of our projected data (by label and patient) and scree analysis to examine variance explained by all methods. Please refer to the associated code repository's Jupyter notebooks for additional experiments and figures.



(a) Correlation plots of loadings and latent factors.

(b) Omics mixing per latent factor and correlation of latent factors with label.

6 Conclusions & Future Work

Unsupervised methods in multi-omics molecular profiling is of utmost importance as more longitudinal and time-series datasets spring up from labs, large research consortiums, and even the healthcare industry [1]. The utility provided by these data and novel methods can help us understand and characterize disease states, with translational applications of providing potential biomarkers for early detection. Despite poor performance on the prediction tasks on latent factors, in this study we demonstrated the utility of neural group factor representations for future benchmarking through correlation analysis of loadings.

In future work, we plan to implement further filtering of our data to get more signal, while making sure not to prune large swathes of our system snapshot. Additionally, we hope to implement other methods such as Multi-dimensional scaling (MDS), Kernel PCA, UMAP, MCMC-EM, as well as other multi-omics state-of-the-art methods like iCluster [21]. For further benchmarking metrics, we plan to also compare reconstruction errors between methods. Due to potential low signal to noise ratio (e.g. due to dirty data, "weaker" signals from acute states like infection, etc.) we plan on deploying this pipeline on other datasets including single cell omics [18]. Furthermore, we plan to explore further hyperparameter tuning via grid search, whereby model selection will be performed by measuring loss on the validation set after each epoch of training and selecting the model with the lowest loss. Biomarker enrichment (similar to Gene Set Enrichment Analysis) will follow soon afterward to determine whether sets of features are significantly different between healthy and diseased groups to gather additional insights when mapping our loadings back to features. For the benefit of benchmarking, we also plan to further explore theoretical guarantees of these new neural methods. Finally, we hope to extend these methods to the time-domain with recurrent architectures, online learning, semi-supervised models. While TCGA data are sampled at one time point (assumed i.i.d.) and iPOP data were assumed i.i.d. due to potential batch effects resulting in low autocorrelation (despite longitudinal sampling), recent access to dense time-series profiling from NIH's Molecular Transducers of Physical Activity Consortium (MoTrPAC) [3] has created opportunities for such methods. Through semi-supervision, sVAEs would be exposed to health status labels of in a transfer learning framework [26], while with recurrent architectures and online learning, the implementation would include recurrence via an LSTM-VAE [10]. As benchmarks, we could compare these approaches to supervised linear methods [19] and more classical time-series models like dynamic time warping (DTW) clustering [16], respectively.

7 Contributions

I would like to thank Drs. David Amar and Euan Ashley for help with problem ideation and mentorship, as well as access to the datasets. Dr. Amar was instrumental in correlation analysis and other statistical framings for benchmarking. The Github repository for the codebase and extra figures and experiments contained on Jupyter notebooks can be found at: https://github.com/gmachiraju/cs229_dl_multiomics

References

[1] The baseline study. *https://clinicaltrials.gov/ct2/show/NCT03154346?term=baseline+studyrank=1*.

[2] The cancer genome atlas homepage. *http://cancergenome.nih.gov/abouttcga*.

[3] Molecular transducers of physical activity consortium (motrpac). *https://www.motrpac.org/*.

[4] Ainsworth. oi-vae: Output interpretable vaes for nonlinear group factor analysis. *arXiv*, 2018.

[5] Argelaguet et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 2018.

[6] Baldi et al. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 1989.

[7] Bouchacourt et al. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

[8] Chaudhary et al. Deep learning based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 2018.

[9] Chen et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, 2012.

[10] Chung et al. Unsupervised classification of multi-omics data during cardiac remodeling using deep learning. *Methods*, 2019.

[11] Hinton et al. Reducing the dimensionality of data with neural networks. *Science*, 2006.

[12] Huang et al. More is better: Recent progress in multi-omics data integration methods. *Frontiers in Genetics*, 2017.

[13] Kim et al. Data integration and predictive modeling methods for multi-omics datasets. *Molecular Omics*, 2018.

[14] Noor et al. Biological insights through omics data integration. *Current opinion in systems biology*, 2019.

[15] Peters et al. Putting benchmarks in their rightful place: The heart of computational biology. *PLOS Computational Biology*, 2018.

[16] Petitjean et al. Summarizing a set of time series by averaging: From steiner sequence to compact multiple alignment. *Theoretical Computer Science*, 2012.

[17] Plaut et al. From principal subspaces to principal components with linear autoencoders. *arXiv*, 2018.

[18] Price et al. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology*, 2017.

[19] Rohart et al. mixomics: An r package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 2017.

[20] Ronen et al. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *BioarXiv*, 2018.

[21] Shen et al. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 2009.

[22] Steinacher et al. Nonlinear dynamics in gene regulation promote robustness and evolvability of gene expression levels. *PLOS One*, 2016.

[23] Tschannen et al. Recent advances in autoencoder-based representation learning. *NeurIPS*, 2018.

[24] Xie et al. A deep auto-encoder model for gene expression prediction. *BMC Genomics*, 2013.

[25] Zhang et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Frontiers in Genetics*, 2018.

[26] Zhuang et al. Supervised representation learning: Transfer learning with deep autoencoders. *IJCAI*, 2015.

[27] Kingma. Auto-encoding variational bayes. *arXiv*, 2013.