

Semi-supervised Cooperative Learning for Multiomics Data Fusion

Daisy Yi Ding, Xiaotao Shen, Michael Snyder, and Robert Tibshirani

Stanford University, Stanford, CA 94305, USA

Abstract. Multiomics data fusion integrates diverse data modalities, ranging from transcriptomics to proteomics, to gain a comprehensive understanding of biological systems and enhance predictions on outcomes of interest related to disease phenotypes and treatment responses. Cooperative learning, a recently proposed method, unifies the commonly-used fusion approaches, including early and late fusion, and offers a systematic framework for leveraging the shared underlying relationships across omics to strengthen signals. However, the challenge of acquiring large-scale labeled data remains, and there are cases where multiomics data are available but in the absence of annotated labels. To harness the potential of unlabeled multiomics data, we introduce semi-supervised cooperative learning. By utilizing an “agreement penalty”, our method incorporates the additional unlabeled data in the learning process and achieves consistently superior predictive performance on simulated data and a real multiomics study of aging. It offers an effective solution to multiomics data fusion in settings with both labeled and unlabeled data and maximizes the utility of available data resources, with the potential of significantly improving predictive models for diagnostics and therapeutics in an increasingly multiomics world.

Keywords: Multiomics data fusion · Semi-supervised learning · Machine learning.

1 Introduction

With advancements in biotechnologies, significant progress has been made in generating and collecting a diverse range of “-omics” data on a common set of patients, including genomics, epigenomics, transcriptomics, proteomics, and metabolomics (Figure 1A). These data characterize molecular variations of human health from different perspectives and of different granularities. Fusing the multiple data modalities on a common set of observations provides the opportunity to gain a more holistic understanding of outcomes of interest such as disease phenotypes and treatment response. It offers the potential to discover hidden insights that may remain obscured in single-modality data analyses and achieve more accurate predictions of the outcomes [Kristensen et al., 2014, Ritchie et al., 2015, Robinson et al., 2017, Karczewski and Snyder, 2018, Ma et al., 2020,

The 2023 ICML Workshop on Machine Learning for Multimodal Healthcare Data.

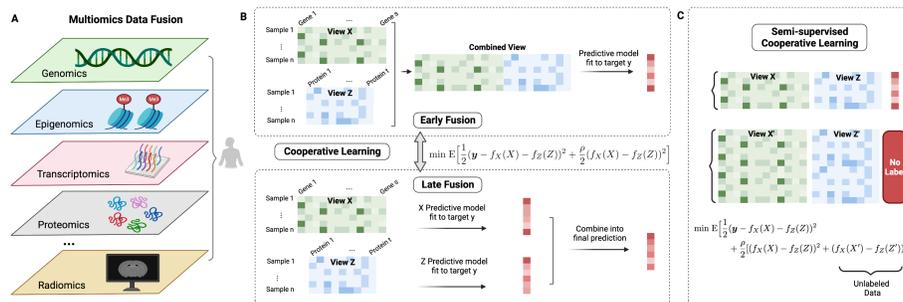


Fig. 1: Framework of semi-supervised cooperative learning for multiomics data fusion. (A) The advancements in biotechnologies have led to the generation and collection of diverse “omics” data on a common set of samples, ranging from genomics to proteomics. Fusing the data provides a unique opportunity to gain a holistic understanding of complex biological systems and enhance predictive accuracy on outcomes of interest related to disease phenotypes and treatment response (B) Commonly-used approaches to multiomics data fusion have two broad categories: early fusion involves transforming all data modalities into a unified representation before feeding it into a model of choice, while late fusion builds separate models for each data modality and combines their predictions using a second-level model. Encompassing early and late fusion, cooperative learning exploits the shared underlying relationships across omics for enhanced predictive performance. (C) The field of biomedicine faces a persistent challenge due to the scarcity of large-scale labeled data, which requires significant resources to acquire. In cases where unlabeled multiomics data are also accessible, we introduce semi-supervised cooperative learning to leverage the combined information from both labeled and unlabeled data. The agreement penalty seamlessly integrates the unlabeled samples into the learning process, effectively utilizing the shared underlying signals present in both labeled and unlabeled data and maximizing the utility of available data resources for multiomics data fusion.

Hao et al., 2021]. While the term “multiomics data fusion” can have various interpretations, we use it here in the context of predicting an outcome of interest by integrating different data modalities.

Commonly-used data fusion methods can be broadly categorized into early and late fusion (Figure 1B). Early fusion works by transforming the multiple data modalities into a single representation before feeding the aggregated representation into a supervised learning model of choice [Yuan et al., 2014, Gentles et al., 2015, Perkins et al., 2018, Chaudhary et al., 2018]. Late fusion refers to methods where individual models are first built from the distinct data modality, and then the predictions of the individual models are combined into the final predictor [Yang et al., 2010, Zhao et al., 2019, Chen et al., 2022, Chabon et al., 2020, Wu et al., 2020]. However, both methods do not explicitly leverage the shared underlying relationship across modalities, and a systematic framework for multiomics data fusion is lacking.

To tackle this limitation, a new method called *cooperative learning* has recently been proposed [Ding et al., 2022]. It combines the usual squared error loss of

predictions with an “agreement” penalty to encourage alignment of predictions from different data modalities (Figure 1B). By varying the weight of the agreement penalty, one can get a continuum of solutions that include early and late fusion. Cooperative learning chooses the degree of fusion in a data-adaptive manner, providing enhanced flexibility and performance. It has demonstrated effectiveness on both simulated data and real multiomics data, particularly when the different data modalities share some underlying relationships in their signals that can be exploited to boost the signals.

However, an important challenge persists in the field of biomedicine: the scarcity of large-scale labeled data. Acquiring a substantial amount of labeled data in this domain often demands considerable effort, time, and financial resources. Nonetheless, there are instances where multiomics data are available, but in the absence of corresponding labels. In such cases, it becomes imperative to leverage the available unlabeled data to enhance predictive models.

To harness the potential of unlabeled data, we propose *semi-supervised cooperative learning*. The key idea is to utilize the agreement penalty, inherent in the cooperative learning framework, as a means to leverage the matched unlabeled samples to our advantage (Figure 1C). It acts as a mechanism for incorporating the unlabeled samples into the learning process, by encouraging the predictions from different data modalities to align not only on the labeled samples but also on the unlabeled ones. Semi-supervised cooperative learning leverages the additional shared underlying signals across the unlabeled data and exploits the valuable information that would otherwise remain untapped. Through comprehensive simulated studies and a real multiomics study of aging, we showed that our method achieves consistently higher predictive accuracy on the outcomes of interest. By incorporating matched unlabeled data and thus maximizing the utility of available data, semi-supervised cooperative learning offers an effective solution to multiomics data fusion, with the potential to significantly enhance predictive models and unlock hidden insights in health and disease.

2 Approach

2.1 Cooperative learning

We begin by giving a concise overview of the recently proposed *cooperative learning* framework [Ding et al., 2022] to set the stage for the introduction of *semi-supervised cooperative learning*. Let $X \in \mathcal{R}^{n \times p_x}$, $Z \in \mathcal{R}^{n \times p_z}$ — representing two data views — and $\mathbf{y} \in \mathcal{R}^n$ be a real-valued response. Fixing the hyperparameter $\rho \geq 0$, cooperative learning aims to minimize the population quantity:

$$\min \mathbb{E} \left[\frac{1}{2} (\mathbf{y} - f_X(X) - f_Z(Z))^2 + \frac{\rho}{2} (f_X(X) - f_Z(Z))^2 \right]. \quad (1)$$

The first term is the usual prediction error loss, while the second term is an “agreement” penalty, encouraging alignment of predictions from different modalities.

To be more concrete in the setting of regularized linear regression, for a fixed value of the hyperparameter $\rho \geq 0$, cooperative learning finds $\boldsymbol{\theta}_x \in \mathcal{R}^{p_x}$ and $\boldsymbol{\theta}_z \in \mathcal{R}^{p_z}$ that minimize:

$$J(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}_x - Z\boldsymbol{\theta}_z\|^2 + \frac{\rho}{2} \|(X\boldsymbol{\theta}_x - Z\boldsymbol{\theta}_z)\|^2 + \lambda_x \|\boldsymbol{\theta}_x\|_1 + \lambda_z \|\boldsymbol{\theta}_z\|_1, \quad (2)$$

where ρ is the hyperparameter that controls the relative importance of the agreement penalty term $\|(X\boldsymbol{\theta}_x - Z\boldsymbol{\theta}_z)\|^2$ in the objective, and $\lambda_x \|\boldsymbol{\theta}_x\|_1$ and $\lambda_z \|\boldsymbol{\theta}_z\|_1$ are ℓ_1 penalties*.

When $\rho = 0$, cooperative learning reduces to early fusion, where we simply use the combined set of features in a supervised learning procedure. When $\rho = 1$, we can show that it yields a simple form of late fusion. In addition, theoretical analysis has demonstrated that the agreement penalty offers an advantage in reducing the mean-squared error of the predictions under a latent factor model [Ding et al., 2022].

2.2 Semi-supervised cooperative learning

In this section, we present *semi-supervised cooperative learning*, which enables us to harness the power of both labeled and unlabeled data for multiomics data fusion. Consider feature matrices $X \in \mathcal{R}^{n \times p_x}$, $Z \in \mathcal{R}^{n \times p_z}$, with labels $\mathbf{y} \in \mathcal{R}^n$, and then additional feature matrices $X' \in \mathcal{R}^{n_{\text{unlabeled}} \times p_x}$, $Z' \in \mathcal{R}^{n_{\text{unlabeled}} \times p_z}$, without labels. The objective of semi-supervised cooperative learning is

$$\min \mathbb{E} \left[\frac{1}{2} (\mathbf{y} - f_X(X) - f_Z(Z))^2 + \frac{\rho}{2} [(f_X(X) - f_Z(Z))^2 + (f_X(X') - f_Z(Z'))^2] \right]. \quad (3)$$

The agreement penalty allows us to use the matched unlabeled samples to our advantage, by encouraging predictions from different data modalities to align on both labeled and unlabeled samples, thus leveraging the aligned signals across omics in a semi-supervised manner. This agreement penalty term is also related to “contrastive learning” [Chen et al., 2020, Khosla et al., 2020], which is an unsupervised learning technique first proposed for learning visual representations. Without the supervision of \mathbf{y} , it learns representations of images by maximizing agreement between differently augmented “views” of the same data example. While contrastive learning is unsupervised and cooperative learning is supervised, both of which have a term in the objective that encourages agreement between correlated views, semi-supervised cooperative learning combines the strengths of both paradigms to fully exploit labeled and unlabeled data simultaneously.

*We assume that the columns of X and Z have been standardized, and \mathbf{y} has mean 0 (hence we can omit the intercept). We use the commonly-used ℓ_1 penalties for illustration, while the framework generalizes to other penalty functions.

In the regularized regression setting and with a common λ^\dagger , the objective becomes

$$J(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\theta}_x - Z\boldsymbol{\theta}_z\|^2 + \frac{\rho}{2} [\|(X\boldsymbol{\theta}_x - Z\boldsymbol{\theta}_z)\|^2 + \|(X'\boldsymbol{\theta}_x - Z'\boldsymbol{\theta}_z)\|^2] + \lambda(\|\boldsymbol{\theta}_x\|_1 + \|\boldsymbol{\theta}_z\|_1), \quad (4)$$

and one can compute a regularization path of solutions indexed by λ . Problem (4) is convex, and the solution can be computed as follows. Letting

$$\tilde{X} = \begin{pmatrix} X & Z \\ -\sqrt{\rho}X & \sqrt{\rho}Z \\ -\sqrt{\rho}X' & \sqrt{\rho}Z' \end{pmatrix}, \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \tilde{\boldsymbol{\beta}} = \begin{pmatrix} \boldsymbol{\theta}_x \\ \boldsymbol{\theta}_z \end{pmatrix}, \quad (5)$$

then the equivalent problem to (4) is

$$\frac{1}{2} \|\tilde{\mathbf{y}} - \tilde{X}\tilde{\boldsymbol{\beta}}\|^2 + \lambda(\|\boldsymbol{\theta}_x\|_1 + \|\boldsymbol{\theta}_z\|_1). \quad (6)$$

This is a form of the lasso, and can be computed, for example by the `glmnet` package [Friedman et al., 2010].

Let $\text{Lasso}(X, \mathbf{y}, \lambda)$ denote the generic problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{y} - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1. \quad (7)$$

We outline the algorithm for semi-supervised cooperative learning in Algorithm 1.

Algorithm 1 *Semi-supervised cooperative learning.*

Input: $X \in \mathcal{R}^{n \times p_x}$ and $Z \in \mathcal{R}^{n \times p_z}$, the response $\mathbf{y} \in \mathcal{R}^n$, and the unlabeled data $X' \in \mathcal{R}^{n_{\text{unlabeled}} \times p_x}$ and $Z' \in \mathcal{R}^{n_{\text{unlabeled}} \times p_z}$, and a grid of hyperparameter values $(\rho_{\min}, \dots, \rho_{\max})$.

for $\rho \leftarrow \rho_{\min}, \dots, \rho_{\max}$ **do**
 Set

$$\tilde{X} = \begin{pmatrix} X & Z \\ -\sqrt{\rho}X & \sqrt{\rho}Z \\ -\sqrt{\rho}X' & \sqrt{\rho}Z' \end{pmatrix}, \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

 Solve $\text{Lasso}(\tilde{X}, \tilde{\mathbf{y}}, \lambda)$ over a decreasing grid of λ values.

end

Select the optimal value of ρ^* based on the CV error and get the final fit.

[†]It was shown in Ding et al. [2022] that there is generally no advantage to allowing different λ values for different modalities.

3 Experiments

3.1 Simulated studies

We first compare semi-supervised cooperative learning with vanilla cooperative learning, early and late fusion methods in simulation studies. We generated Gaussian data with $n = 200$ and $p = 500$ in each of two views X and Z , and created correlation between them using latent factors. The response \mathbf{y} was generated as a linear combination of the latent factors, corrupted by Gaussian noise. We then generated an additional set of unlabeled data X' and Z' with $n_{\text{unlabeled}} = 200$ and $p = 500$.

The simulation is set up as follows. Given values for parameters $n, n_{\text{unlabeled}}, p_x, p_z, p_u, s_u, t_x, t_z, \beta_u, \sigma$, we generate data according to the following procedure:

1. $x_j \in \mathcal{R}^n$ and $x'_j \in \mathcal{R}^n$ distributed i.i.d. $\text{MVN}(0, I_n)$ for $j = 1, 2, \dots, p_x$.
2. $z_j \in \mathcal{R}^n$ and $z'_j \in \mathcal{R}^n$ distributed i.i.d. $\text{MVN}(0, I_n)$ for $j = 1, 2, \dots, p_z$.
3. For $i = 1, 2, \dots, p_u$ (p_u corresponds to the number of latent factors, $p_u < p_x$ and $p_u < p_z$):
 - (a) $u_i \in \mathcal{R}^n$ and $u'_i \in \mathcal{R}^n$ distributed i.i.d. $\text{MVN}(0, s_u^2 I_n)$;
 - (b) $x_i = x_i + t_x * u_i$, $x'_i = x'_i + t_x * u'_i$;
 - (c) $z_i = z_i + t_z * u_i$, $z'_i = z'_i + t_z * u'_i$.
4. $X = [x_1, x_2, \dots, x_{p_x}]$, $Z = [z_1, z_2, \dots, z_{p_z}]$.
5. $X' = [x'_1, x'_2, \dots, x'_{p_x}]$, $Z' = [z'_1, z'_2, \dots, z'_{p_z}]$.
6. $U = [u_1, u_2, \dots, u_{p_u}]$, $\mathbf{y} = U\beta_u + \epsilon$ where $\epsilon \in \mathcal{R}^n$ distributed i.i.d. $\text{MVN}(0, \sigma^2 I_n)$.

There is sparsity in the solution since a subset of columns of X and Z are independent of the latent factors used to generate \mathbf{y} . We use 10-fold CV to select the optimal values of hyperparameters. We compare the following methods: (1) separate X and separate Z on the labeled data: the standard lasso is applied on the separate data modalities of X and Z with 10-fold CV; (2) early fusion on the labeled data: the standard lasso is applied on the concatenated data modalities of X and Z with 10-fold CV (note that this is equivalent to cooperative learning with $\rho = 0$); (3) late fusion on the labeled data: separate lasso models are first fitted on X and Z independently with 10-fold CV, and the two resulting predictors are then combined through linear least squares for the final prediction; (4) cooperative learning on the labeled data; (5) semi-supervised cooperative learning on both the labeled and unlabeled data[‡].

Overall, the simulation results can be summarized as follows:

- Semi-supervised cooperative learning performs the best in terms of test MSE across the range of SNR and correlation settings. It is most helpful when the data views are correlated and both contain signals, as shown in Figure 2A.
- When there is no correlation between data views but each data view carries signals, semi-supervised cooperative learning still offers performance advantages as it utilizes the signals in both labeled and unlabeled data, as shown in Figure 2C.

[‡]Traditional supervised learning models are not directly applicable to scenarios involving both labeled and unlabeled data.

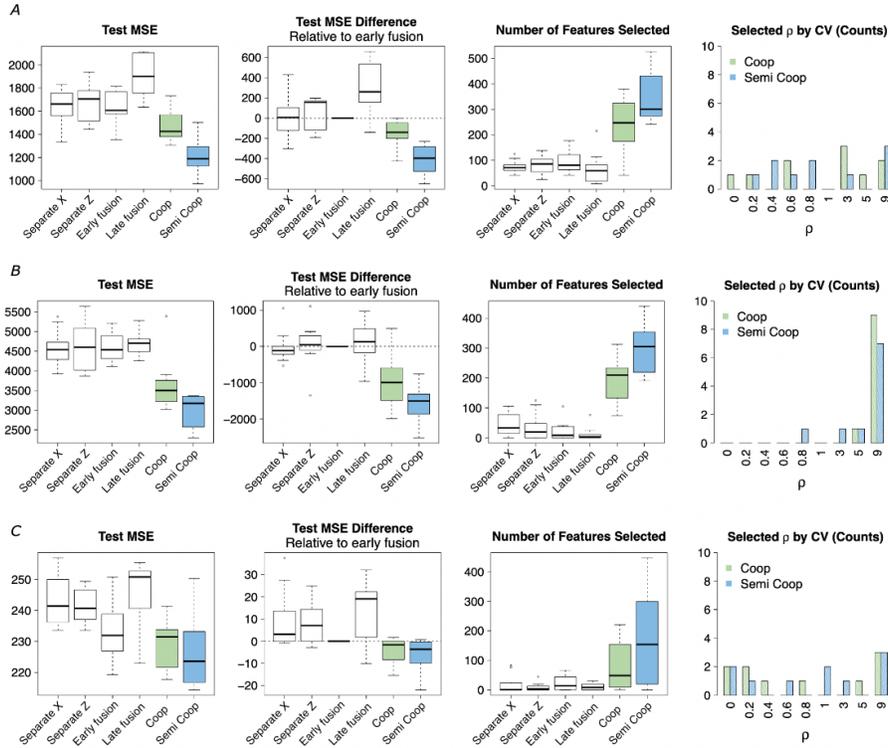


Fig. 2: Simulation studies on semi-supervised cooperative learning. (A) Simulation results when X and Z have a medium level of correlation ($t = 2, s_u = 1$); both X and Z contain signal ($b_x = b_z = 2$), $n = 200, n_{\text{unlabel}} = 200, p = 1000, \text{SNR} = 1.8$. The first panel shows MSE on a test set; the second panel shows the MSE difference on the test set relative to early fusion; the third panel shows the number of features selected; the fourth panel shows the ρ values selected by CV in cooperative learning and semi-supervised cooperative learning. Here “Coop” refers to cooperative learning and “Semi Coop” refers to semi-supervised cooperative learning. (B) Simulation results when X and Z have a high level of correlation ($t = 6, s_u = 1$); only X contains signal ($b_x = 2, b_z = 0$), $n = 200, n_{\text{unlabel}} = 200, p = 1000, \text{SNR} = 0.5$. (C) Simulation results when X and Z have no correlation ($t = 0, s_u = 1$); both X and Z contain signal ($b_x = b_z = 2$), $n = 200, n_{\text{unlabel}} = 200, p = 1000, \text{SNR} = 1.0$.

- When the correlation between data views is higher, higher values of ρ are more likely to be selected, as shown in Figure 2B compared to Figure 2A. In addition, cooperative learning-based methods tend to select more features.

3.2 Real data example

We applied semi-supervised cooperative learning to a real multiomics dataset of aging, collected from a cohort of 100 healthy individuals and individuals with prediabetes, as described in Zhou et al. [2019]. Proteomics and transcriptomics

Table 1: *Multiomics studies on aging.* The first two columns in the table show the mean and standard deviation (SD) of mean absolute error (MAE) on the test set across different splits of the training and test sets; the third and fourth columns show the MAE difference relative to early fusion. The methods include (1) separate proteomics: the standard lasso is applied on the proteomics data only; (2) separate transcriptomics: the standard lasso is applied on the transcriptomics data only; (3) early fusion: the standard lasso is applied on the concatenated data of proteomics and transcriptomics data; (4) late fusion: separate lasso models are first fit on proteomics and transcriptomics independently and the predictors are then combined through linear least squares; (5) cooperative learning; (6) semi-supervised cooperative learning.

Methods	Test MAE		Relative to Late Fusion	
	Mean	Std	Mean	Std
Separate Proteomics	8.49	0.40	-0.04	0.12
Separate Transcriptomics	8.44	0.35	-0.08	0.20
Early fusion	8.52	0.32	0	0
Late fusion	8.53	0.29	0.01	0.13
Cooperative learning	8.16	0.40	-0.37	0.16
Semi-supervised cooperative learning	7.85	0.47	-0.67	0.26

were measured on the cohort: the proteomics data contained measurements for 302 proteins and the transcriptomics data contained measurements for 8,556 genes. The goal of the analysis is to predict age using proteomics and transcriptomics data and uncover molecular signatures associated with the aging process.

We split the data set of 100 individuals into training and test sets of 75 and 25 individuals, respectively. We artificially masked the labels for half of the training samples to create a mix of labeled and unlabeled data. Both the proteomics and transcriptomics measurements were screened by their variance across the subjects. We averaged the expression levels for each individual across time points in the longitudinal study and predicted the corresponding age. We conducted the same set of experiments across 10 different random splits of the training and test sets.

The results are shown in Table 1. The model fit on the transcriptomics data achieves lower test MAE than the one fit on the proteomics data. Early and late fusion hurt performance as compared to the model fit on only proteomics or transcriptomics. Cooperative learning outperforms both early and late fusion by encouraging the predictions to align with each other. Semi-supervised cooperative learning gives further performance gains by utilizing both the labeled and unlabeled data. Moreover, it selects important features not identified by the other methods for predicting age, including PDK1, MYSM1, ATP5A1, APOA4, MST, A2M, which have been previously demonstrated to be associated with the aging process [An et al., 2020, Tian et al., 2020, Choi et al., 2019, Goldberg et al.,

2018, Blacker et al., 1998, Garasto et al., 2003, Lee et al., 2013, Shang et al., 2022].

4 Conclusion

We introduce semi-supervised cooperative learning for multiomics data fusion in the presence of both labeled and unlabeled data. By exploiting the shared underlying relationships across omics through an agreement penalty in both labeled and unlabeled data, our proposed approach demonstrates improved predictive accuracy on simulated studies and a real multiomics study of aging. The agreement penalty allows us to effectively incorporate the unlabeled samples in the learning process and leverage them to our advantage. To our knowledge, our work represents a pioneering effort in multi-omics data fusion that unlocks the untapped potential of unlabeled data, enabling us to harness the valuable information that would otherwise remain unused for the discovery of novel insights and enhanced predictive modeling of diagnostics and therapeutics.

Bibliography

- Vessela N Kristensen, Ole Christian Lingjærde, Hege G Russnes, Hans Kristian M Volla, Arnaldo Frigessi, and Anne-Lise Børresen-Dale. Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer*, 14(5):299–313, 2014.
- Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.
- Dan R Robinson, Yi-Mi Wu, Robert J Lonigro, Pankaj Vats, Erin Cobain, Jessica Everett, Xuhong Cao, Erica Rabban, Chandan Kumar-Sinha, Victoria Raymond, et al. Integrative clinical genomics of metastatic cancer. *Nature*, 548(7667):297–303, 2017.
- Konrad J Karczewski and Michael P Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19(5):299, 2018.
- Anjun Ma, Adam McDermaid, Jennifer Xu, Yuzhou Chang, and Qin Ma. Integrative methods and practical challenges for single-cell multi-omics. *Trends in biotechnology*, 38(9):1007–1022, 2020.
- Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.
- Yuan Yuan, Eliezer M Van Allen, Larsson Omberg, Nikhil Wagle, Ali Amin-Mansour, Artem Sokolov, Lauren A Byers, Yanxun Xu, Kenneth R Hess, Lixia Diao, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature Biotechnology*, 32(7):644–652, 2014.
- Andrew J Gentles, Scott V Bratman, Luke J Lee, Jeremy P Harris, Weiguo Feng, Ramesh V Nair, David B Shultz, Viswam S Nair, Chuong D Hoang, Robert B West, et al. Integrating tumor and stromal gene expression signatures with clinical indices for survival stratification of early-stage non-small cell lung cancer. *JNCI: Journal of the National Cancer Institute*, 107(10), 2015.
- Bradley A Perkins, C Thomas Caskey, Pamela Brar, Eric Dec, David S Karow, Andrew M Kahn, Ying-Chen Claire Hou, Naisha Shah, Debbie Boeldt, Erin Coughlin, et al. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proceedings of the National Academy of Sciences*, 115(14):3686–3691, 2018.
- Kumardeep Chaudhary, Olivier B Poirion, Liangqun Lu, and Lana X Garmire. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259, 2018.

- Pengyi Yang, Yee Hwa Yang, Bing B Zhou, and Albert Y Zomaya. A review of ensemble methods in bioinformatics. *Current Bioinformatics*, 5(4):296–308, 2010.
- Juan Zhao, QiPing Feng, Patrick Wu, Roxana A Lupu, Russell A Wilke, Quinn S Wells, Joshua C Denny, and Wei-Qi Wei. Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Scientific Reports*, 9(1):1–10, 2019.
- Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F. K. Williamson, Scott J. Rodig, Neal I. Lindeman, and Faisal Mahmood. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022. <https://doi.org/10.1109/TMI.2020.3021387>.
- Jacob J Chabon, Emily G Hamilton, David M Kurtz, Mohammad S Esfahani, Everett J Moding, Henning Stehr, Joseph Schroers-Martin, Barzin Y Nabet, Binbin Chen, Aadel A Chaudhuri, et al. Integrating genomic features for non-invasive early lung cancer detection. *Nature*, 580(7802):245–251, 2020.
- Lang Wu, Yaohua Yang, Xingyi Guo, Xiao-Ou Shu, Qiuyin Cai, Xiang Shu, Bingshan Li, Ran Tao, Chong Wu, Jason B Nikas, et al. An integrative multi-omics analysis to identify candidate dna methylation biomarkers related to prostate cancer risk. *Nature Communications*, 11(1):1–11, 2020.
- Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan, and Robert Tibshirani. Cooperative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119, 2022.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschiot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020.
- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.
- Wenyu Zhou, M Reza Sailani, Kévin Contrepois, Yanjiao Zhou, Sara Ahadi, Shana R Leopold, Martin J Zhang, Varsha Rao, Monika Avina, Tejaswini Mishra, et al. Longitudinal multi-omics of host–microbe dynamics in prediabetes. *Nature*, 569(7758):663–671, 2019.
- Sugyun An, Si-Young Cho, Junsoo Kang, Soobeom Lee, Hyung-Su Kim, Dae-Jin Min, EuiDong Son, and Kwang-Hyun Cho. Inhibition of 3-phosphoinositide-dependent protein kinase 1 (pdk1) can revert cellular senescence in human

- dermal fibroblasts. *Proceedings of the National Academy of Sciences*, 117(49):31535–31546, 2020.
- Mingfu Tian, Yuqing Huang, Yunting Song, Wen Li, Peiyi Zhao, Weiyong Liu, Kailang Wu, and Jianguo Wu. Mysm1 suppresses cellular senescence and the aging process to prolong lifespan. *Advanced Science*, 7(22):2001950, 2020.
- So Yoen Choi, Rodrigo Lopez-Gonzalez, Gopinath Krishnan, Hannah L Phillips, Alissa Nana Li, William W Seeley, Wei-Dong Yao, Sandra Almeida, and Fen-Biao Gao. C9orf72-als/ftd-associated poly (gr) binds atp5a1 and compromises mitochondrial function in vivo. *Nature neuroscience*, 22(6):851–862, 2019.
- Joshua Goldberg, Antonio Currais, Marguerite Prior, Wolfgang Fischer, Chandramouli Chiruta, Eric Ratliff, Daniel Daugherty, Richard Dargusch, Kim Finley, Pau B Esparza-Moltó, et al. The mitochondrial atp synthase is a shared drug target for aging and dementia. *Aging cell*, 17(2):e12715, 2018.
- Deborah Blacker, Marsha A Wilcox, Nan M Laird, Linda Rodes, Steven M Horvath, Rodney CP Go, Rodney Perry, Bracie Watson, Susan S Bassett, Melvin G McInnis, et al. Alpha-2 macroglobulin is genetically associated with alzheimer disease. *Nature genetics*, 19(4):357–360, 1998.
- S Garasto, G Rose, F Derango, M Berardelli, A Corsonello, E Feraco, V Mari, R Maletta, A Bruni, C Franceschi, et al. The study of apoa1, apoc3 and apoa4 variability in healthy ageing people reveals another paradox in the oldest old subjects. *Annals of human genetics*, 67(1):54–62, 2003.
- Jae Keun Lee, Jin Hee Shin, Sang Gil Hwang, Byoung Joo Gwag, Ann C McKee, Junghee Lee, Neil W Kowall, Hoon Ryu, Dae-Sik Lim, and Eui-Ju Choi. Mst1 functions as a key modulator of neurodegeneration in a mouse model of als. *Proceedings of the National Academy of Sciences*, 110(29):12066–12071, 2013.
- Huayu Shang, Trisha A VanDusseldorp, Ranggui Ma, Yan Zhao, Jason Cholewa, Nelo Eidy Zanchi, and Zhi Xia. Role of mst1 in the regulation of autophagy and mitophagy: implications for aging-related diseases. *Journal of physiology and biochemistry*, pages 1–11, 2022.