Machine learning derived embeddings of bulk multi-omics data enable clinically significant representations in a pan-cancer cohort

Sanjay Nagaraj
Insitro Inc.
279 E Grand Ave., South San Francisco, CA 94080
snagaraj@insitro.com

Zachary McCaw
Insitro Inc.
279 E Grand Ave., South San Francisco, CA 94080
zmccaw@insitro.com

Theofanis Karaletsos
Insitro Inc.
279 E Grand Ave., South San Francisco, CA 94080
theofanis@insitro.com

Daphne Koller
Insitro Inc.
279 E Grand Ave., South San Francisco, CA 94080
daphne@insitro.com

Anna Shcherbina
Insitro Inc.
279 E Grand Ave., South San Francisco, CA 94080
annashch@insitro.com

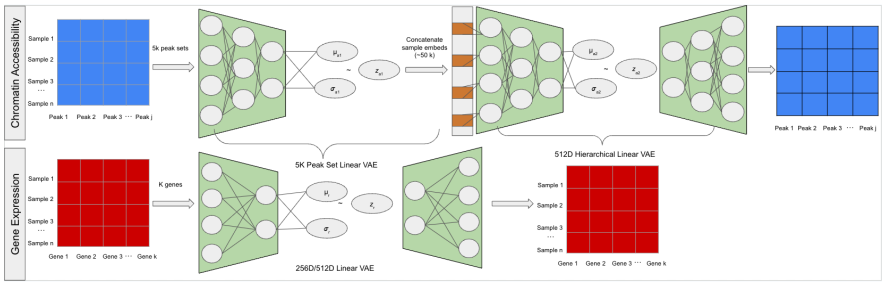
Abstract

Bulk multiomics data provides a comprehensive view of tissue biology, but datasets rarely contain matched transcriptomics and chromatin accessibility data for a given sample. Furthermore, it is difficult to identify relevant genetic signatures from the high-dimensional, sparse representations provided by omics modalities. Machine learning (ML) models have the ability to extract dense, information-rich, denoised representations from omics data, which facilitate finding novel genetic signatures. To this end, we develop and compare generative ML models through an evaluation framework that examines the biological and clinical relevance of the underlying latent embeddings produced. We focus our analysis on pan-cancer multiomics data from a set of 21 diverse cancer metacohorts across three datasets. We additionally investigate if our framework can generate robust representations from oncology imaging modalities (i.e. histopathology slides). Our best performing models learn clinical and biological signals and show improved performance over traditional baselines in our evaluations, including overall survival prediction.

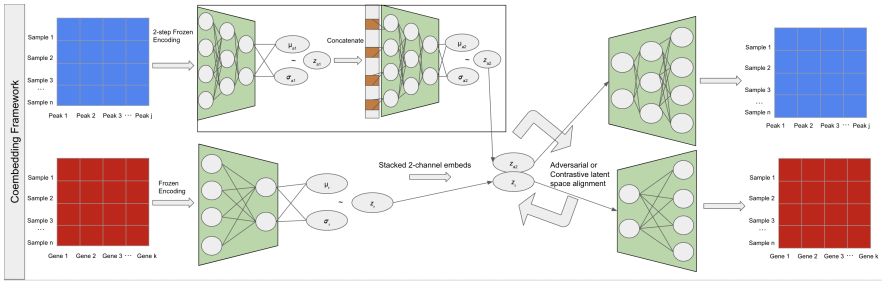
1 Introduction

Bulk RNA and ATAC sequencing have been routinely utilized for the past decade to understand tissue-level changes in gene expression and chromatin accessibility. These sequencing technologies have been critical in understanding the causal relationships between genetic variation and phenotype, especially in oncology [1]. Furthermore, integrated analysis of bulk ATAC and RNA-seq profiles can reveal deeper biological insights around biological pathways and tumor state, with methods being developed to quantify peak (i.e. enhancer)-gene interactions [2][3]. However, a number of challenges exist in directly working with bulk sequencing data (and omics data more generally), ranging from sparseness and high-dimensionality to the technical noise inherent in the data collection process[4]. Therefore, we propose a novel generative ML framework to create 1) modality-specific

NeurIPS 2023 Generative AI and Biology Workshop.



(a) Single-modality embedding models



(b) Coembedding models

Figure 1: Overview of our generative modeling framework. Decoder of the first ATAC VAE is omitted for simplicity.

embeddings for bulk ATAC and RNA-seq, 2) co-embeddings that utilize knowledge transfer across both modalities, and 3) an evaluation framework for benchmarking models with computational validation as well as clinical and biological benchmarks. Similar generative modeling ideas have emerged in the single-cell space that focus on heterogeneous single-cell dataset integration [5][6][7] or co-embedding of scRNA-seq and sc/snATAC-seq data for imputation, automatic batch correction, and cell-type analysis [8][9][10]. Unfortunately, these methods are not easily extensible to the bulk-seq regime. Therefore, our modeling framework is the first of its kind for producing generalized clinical representations directly from unpaired bulk RNA and ATAC-seq data.

2 Methods

Data and Preprocessing Bulk ATAC count data were obtained from TCGA[11] (800 ATAC samples across 33 cancer types), while bulk RNA data was obtained from TCGA (11,000 samples across 33 cancer subtypes), ENCODE[12] (40 healthy samples), and a large, US-based real-world oncology source (2000 samples from 182 unique pan-cancer tumor sites). The corresponding peak-by-sample and gene-by-sample matrices were pre-processed to remove zero counts data and peaks/genes that showed less than 1% variance in counts across all samples. The remaining data was quantile-normalized and batch corrected via edgeR's `removeBatchEffect`[13] algorithm to remove confounding variables (i.e. demographics, library prep, sequencing differences) while preserving cohort-level biological signals.

Training Setup and Baselines Samples were divided into a 80-20 train-evaluation split, with a fixed hold-out set. The performance of models on the hold-out set was benchmarked against a baseline of 512 principal components (PCs) fit on the training set.

Single Modality Architecture and Training Procedure For the single-modality initial ATAC variational autoencoder (VAE) (Figure 1a), we used a three layer encoder-decoder architecture with three fully connected modules, where we define such a module to contain a linear layer, a leaky ReLU nonlinearity, and 1D BatchNorm, to generate a $\sim$ 50k dimensional embedding. A forward pass of the encoder saw overlapping blocks of 5000 contiguous peaks from training samples and an epoch consisted of training until all 563,000 peaks were seen. This approach enables the model to learn

biological spatial constraints driving chromatin accessibility mediated patterns of gene regulation (i.e. cooperative and additive peak effects)[14]. Between passes, we employed a 1000 peak sliding window across the 5k peak set as a form of data augmentation and to prevent model forgetting of prior training. Training was performed until convergence (i.e. 2 continuous epochs of no validation loss improvement). The resulting model generated a 512D embedding for each peak set, which were concatenated and provided as input to a hierarchical VAE, consisting of a convolutional layer with kernel size 3 followed by two fully connected modules, to generate a final 512D embedding. This hierarchical approach was further necessitated by the high dimensionality of the sparse input ATAC profiles, which led to unstable training when directly encoding into our final embedding space. Both ATAC VAEs were optimized with a reconstruction loss that summed the Huber loss between samples and reconstructions with the KL divergence from the latent distribution.

For the single-modality RNA VAE (Figure 1a), a shallower architecture was used to avoid overfitting on the comparatively smaller feature space (two fully connected layers with a single ReLU nonlinearity between) with the matrix of 19k relevant genes by $\sim$ 13k patients in the training set fed directly. The capacity of the architecture was sufficient to avoid a need for gene set batching. The RNA VAE was optimized with the same losses as the ATAC VAEs.

Co-embedding Architecture and Training Procedure We trained two different types of co-embedding VAEs (Figure 1b) to perform adversarial latent space alignment and contrastive latent space alignment respectively. Alignment was performed under the assumption that shared information between the two modalities (ATAC and RNA) can improve representations. The adversarial alignment was performed by training a 3-layer fully-connected discriminator to distinguish between modality embeddings by binarizing modality type into a learnt label. Training used the single-modality ATAC and RNA reconstruction losses, a discriminator binary cross-entropy loss, generator loss (opposite of discriminator loss), and a cross-modality cycle consistency loss. This loss measures per-modality alignment between the latent space of a sample embedding versus an embedding of a reconstruction of the same sample. Formally, given E_a as the frozen ATAC encoding, E_r as the frozen RNA encoding, D_a as the ATAC decoder, and D_r as the RNA decoder, the loss can be written as the quantity $\mathbb{E}_{(a\sim P_{ATAC})}$ MSE$(E_a(a), E_r(D_r(E_a(a))))$ + $\mathbb{E}_{(r\sim P_{RNA})}$ MSE$(E_r(r), E_a(D_a(E_r(r))))$ where MSE represents the mean-squared error loss [15].

The contrastive latent space alignment was performed by adding an additional in-modality contrastive term to the modality-specific reconstruction losses alongside the cross-modality cyclical consistency loss. Given d_i as the cancer cohort label of sample i, the contrastive loss is calculated as $\mathbb{E}_{(a\sim P_{ATAC}, p\sim P_{(ATAC|d=d_a)}, n\sim P_{(ATAC|d\neq d_a)})} \max(||a-p||_2 - ||a-n||_2 + 1, 0)$, if $s \leq 0.5$ otherwise $\mathbb{E}_{(a\sim P_{RNA}, p\sim P_{(RNA|d=d_a)}, n\sim P_{(RNA|d\neq d_a)})} \max(||a-p||_2 - ||a-n||_2 + 1, 0)$ where $s \sim U(0,1)$. This enforces high embedding similarity between samples with the same cohort label and low embedding similarity with embeddings from a randomly sampled disease cohort for the ATAC and RNA modalities.

Each of these losses were optimized with learned hyper-parameters that were found using a grid search with Wandb, a Bayesian hyperparameter optimization service [16].

Based on these architectures, co-embedding VAEs were trained to create joint representations of bulk omics data by leveraging a heuristic matching algorithm to create train and hold-out sets that best mimic the population distributions of the individual modalities. The hold-out cohort contained 160 samples with ATAC and RNA data from the same donor, but not necessarily from the same sample. The training set was initially created by matching ATAC and RNA samples on cancer subtype, such as BRCA (breast) and COAD (colorectal), age, sex, and self-reported ethnicity. Additional augmentation was performed by matching each ATAC sample with the 4 most highly correlated RNA samples (in the gene space). The performance of these models was benchmarked against the highest-performing single-modality ATAC and RNA models.

Evaluation Framework The latent embeddings from the VAEs were evaluated via four benchmarks. To avoid variability in evaluations caused by sampling, we sampled distribution means as our latent embeddings at test-time.

Two technical benchmarks were used to assess how well the embeddings captured information content within the original input data. The first, cancer cohort classification accuracy, refers to the accuracy of the cancer predictions from an XgBoost classifier trained and evaluated in a 5-fold manner on the

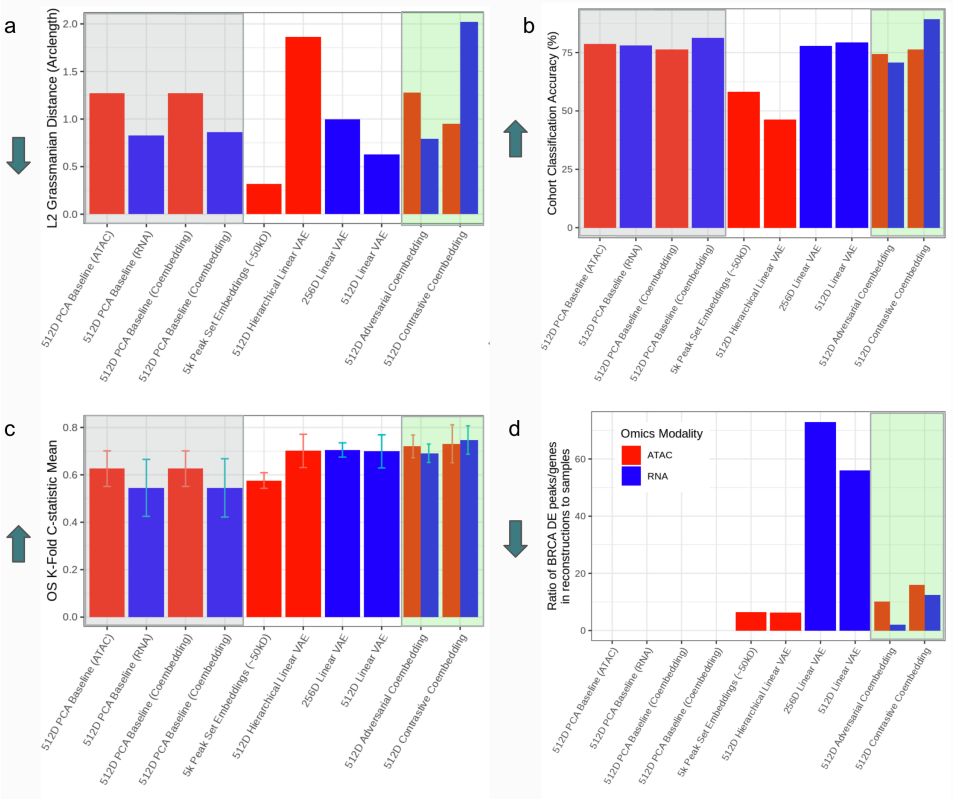


Figure 2: Comparisons of single-modality versus co-embedding models suggest co-embedding ATAC and RNA yields improved overall performance. Grey highlighting indicates PCA baselines. Green highlighting denotes the highest performing models across the four metrics. Arrows represent the direction of increased performance for a metric. Error bars indicate standard error (SE).

embeddings[17]. The second measures alignment of the ground truth cohort label subspace with the embedding subspace using the L2 Norm of the Grassmanian manifold distance [18]. We compute this quantity by taking the PCA of each subspace and computing the L2 norm of the arclength from the top n (n = 100) principal angles between subspaces. Two additional benchmarks were employed to assess the utility of the embeddings in preserving meaningful biology. The ability of the embeddings to predict patient-level overall survival (OS) was assessed by fitting Cox proportional hazards using the top K (hyper-parameter) embedding PCs and evaluated by the mean and standard deviation, across 5-folds, of the C-statistic[19]. Finally, differential expression/accessibility between observed and reconstructed peaks/genes within the test patients was calculated as peaks/genes with abs$(\log_2(\text{fold change}))$ > 1 and FDR < 0.05. The numbers of differentially expressed genes and differentially accessible peaks are reported for the largest cohort – BRCA (breast cancer).

3 Results

Embeddings from our best single modality models show strong performance on the alignment and OS tasks (Figure 2). The hierarchical ATAC VAE embeddings improve the PCA baseline on the OS task and perform well on the DE task, but struggle on the alignment and cancer classification tasks. Both RNA embedding models perform similarly well across the metrics, with the 512D RNA Linear VAE showing slightly stronger technical and biological performance. Additionally, the 512D RNA VAE improves the L2 Norm by 0.2 on the alignment task and the c-statistic by 0.15 on OS over the PCA baseline, while slightly exceeding the classification performance of the PCA baseline. Differential peak and gene ratios are reported relative to pseudoreplicate technical noise baselines of 5813 differential peaks and 2 differential genes, from a starting set of 19,421 expressed genes and 562,709 accessible peaks.

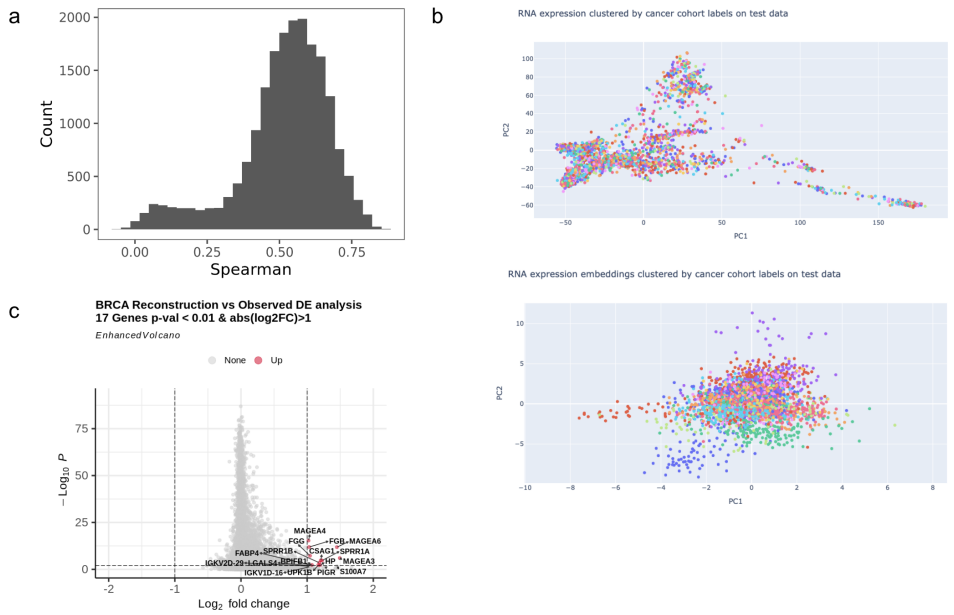


Figure 3: Evaluation of our H&E imaging to RNA embedding pipeline. a) High spearman correlation between predicted and actual gene expression counts across 19,421 genes. b) PCA of ground truth expression and VAE-derived embeddings shows denoising and clustering by cancer cohort type. c) Volcano plot of BRCA reconstructions versus observed expression counts in gene space shows 17 differentially expressed genes.

Co-embedding the ATAC and RNA modalities via contrastive and adversarial approaches yields improved or comparable performance to single-modality models on the two technical benchmarks (Figure 2). ATAC cohort classification accuracy shows a significant improvement due to co-embedding, and the Grassmannian is improved by co-embedding except against the best single-modality RNA VAE. However, the best cohort classification accuracy from co-embedding is similar to that of the PCA baseline.

Furthermore, we see comparable or improved performance with co-embedding over single-modality embedding for the C-statistic mean and the differential expression metric. Specifically, both adversarial and contrastive co-embedding show competitive performance to OS analysis with age, sex, and tumor type (C-statistic mean: 0.715, std: 0.014). Co-embedding also shows improvements over the baseline for the alignment and clinical tasks. We also note that Gene Ontology (GO) term enrichment analysis with GOrilla (RNA)[20] and GREAT (ATAC)[21] for co-embedding showed no enrichment for the differential gene/peak sets, suggesting that the peaks/genes are largely driven by stochastic noise and hence the reconstructions are not systematically biased.

We also investigated applying our modeling framework on imputed RNA profiles derived from H&E whole slide images (WSIs). First, we ran HIPT [22] on WSIs to generate tile-level embeddings, from which a supervised learning model was trained to predict tile-level bulk RNA-seq expression. RNA-seq expression was then averaged across the tiles to produce an imputed RNA profile. This profile is validated by ensuring that the imputed gene expression values have moderate to high correlation with the actual bulk RNA-seq profiles (Figure 3a). Using a 80/20 train-test split, we re-trained and evaluated our RNA VAE model. The retrained model achieves a 80.38% cohort classification accuracy and an average spearman correlation coefficient of 0.788 between the reconstructed and ground truth samples. Furthermore, the model's embeddings show strong biologically, relevant denoising (Figure 3b), but the OS C-statistic has a mean of only 0.52 and standard deviation of 0.02. This indicates that the embedding model is able to learn biologically important structure in the latent space, but is not expressive enough yet to accurately predict mortality. This should be expected given that cell morphology from imaging is insufficient to fully predict RNA signal. Lastly, the differential expression between sample-level gene expression predictions and ground truth expression levels was calculated, with only 17 genes of 19421 being differentially expressed (Figure 3c). This indicates that

most important genetic drivers and relationships in our pan-cancer cohort can be preserved within our modeling framework. According to GOrilla, 6 of these genes are not involved in known cancer pathways while the other 11 are. Out of the 11 DE genes, the average spearman correlation between their predicted and ground truth expression profiles is 0.55. Improving the predictive capacity of our RNA imputation model can improve embedding quality and decrease the number of DE genes. Future work combining the imaging and imputed multi-omic profiles with stronger priors can also help in deriving higher fidelity clinical signals.

Together, these analyses indicate that ML-derived embeddings can model complex biology, predict clinical outcomes, and serve as a useful source to impute additional phenotypes and possibly identify novel genetic associations.

4 Conclusions and Future Work

We show that generative ML models can create rich, clinically-relevant representations of bulk multi-omics data. Notably, we create co-embedding methods to produce rich representations of biology and discuss a robust evaluation framework using orthogonal metrics to training. This result supports the potential for generative ML models to create disease-relevant phenotypes that can uncover genetic insights and enable novel therapeutics. We see possible avenues for future work leveraging peak-gene association scores directly, experimenting with attention-based encoding mechanisms, and augmenting generative model training with healthy samples.

References

[1] A. Kashyap, M. A. Rapsomaniki, V. Barros, A. Fomitcheva-Khartchenko, A. L. Martinelli, A. F. Rodriguez, M. Gabrani, M. Rosen-Zvi, and G. Kaigala, "Quantification of tumor heterogeneity: from data acquisition to metric generation," *Trends in Biotechnology*, vol. 40, no. 6, pp. 647–676, Jun. 2022. [Online]. Available: https://doi.org/10.1016/j.tibtech.2021.11.006

[2] L. T. Kagohara, F. Zamuner, E. F. Davis-Marcisak, G. Sharma, M. Considine, J. Allen, S. Yegnasubramanian, D. A. Gaykalova, and E. J. Fertig, "Integrated single-cell and bulk gene expression and ATAC-seq reveals heterogeneity and early changes in pathways associated with resistance to cetuximab in HNSCC-sensitive cell lines," *British Journal of Cancer*, vol. 123, no. 1, pp. 101–113, May 2020. [Online]. Available: https://doi.org/10.1038/s41416-020-0851-5

[3] C. Uhler and G. V. Shivashankar, "Machine learning approaches to single-cell data integration and translation," *Proceedings of the IEEE*, vol. 110, no. 5, pp. 557–576, May 2022. [Online]. Available: https://doi.org/10.1109/jproc.2022.3166132

[4] J. Gustafsson, F. Held, J. L. Robinson, E. Björnson, R. Jörnsten, and J. Nielsen, "Sources of variation in cell-type RNA-seq profiles," *PLOS ONE*, vol. 15, no. 9, p. e0239495, Sep. 2020. [Online]. Available: https://doi.org/10.1371/journal.pone.0239495

[5] M. Lotfollahi, A. Litinetskaya, and F. J. Theis, "Multigrate: single-cell multi-omic data integration," *bioRxiv*, Mar. 2022. [Online]. Available: https://doi.org/10.1101/2022.03.16.484643

[6] W. Kopp, A. Akalin, and U. Ohler, "Simultaneous dimensionality reduction and integration for single-cell ATAC-seq data using deep learning," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 162–168, Feb. 2022. [Online]. Available: https://doi.org/10.1038/s42256-022-00443-1

[7] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models," *Molecular Systems Biology*, vol. 17, no. 1, Jan. 2021. [Online]. Available: https://doi.org/10.15252/msb.20209620

[8] Y. Xu, E. Begoli, and R. P. McCord, "sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network," *npj Systems Biology and Applications*, vol. 8, no. 1, Sep. 2022. [Online]. Available: https://doi.org/10.1038/s41540-022-00245-6

[9] Z.-J. Cao and G. Gao, "Multi-omics single-cell data integration and regulatory inference with graph-linked embedding," *Nature Biotechnology*, vol. 40, no. 10, pp. 1458–1466, May 2022. [Online]. Available: https://doi.org/10.1038/s41587-022-01284-4

[10] S. Kim and J. Wysocka, "Deciphering the multi-scale, quantitative cis-regulatory code," *Molecular Cell*, vol. 83, no. 3, pp. 373–392, Feb. 2023. [Online]. Available: https://doi.org/10.1016/j.molcel.2022.12.032

[11] M. R. Corces, J. M. Granja, S. Shams, B. H. Louie, J. A. Seoane, W. Zhou, T. C. Silva, C. Groeneveld, C. K. Wong, S. W. Cho, A. T. Satpathy, M. R. Mumbach, K. A. Hoadley, A. G. Robertson, N. C. Sheffield, I. Felau, M. A. A. Castro, B. P. Berman, L. M. Staudt, J. C. Zenklusen, P. W. Laird, C. Curtis, W. J. Greenleaf, H. Y. Chang, R. Akbani, C. C. Benz, E. A. Boyle, B. M. Broom, A. D. Cherniack, B. Craft, J. A. Demchok, A. S. Doane, O. Elemento, M. L. Ferguson, M. J. Goldman, D. N. Hayes, J. He, T. Hinoue, M. Imielinski, S. J. M. Jones, A. Kemal, T. A. Knijnenburg, A. Korkut, D.-C. Lin, Y. Liu, M. K. A. Mensah, G. B. Mills, V. P. Reuter, A. Schultz, H. Shen, J. P. Smith, R. Tarnuzzer, S. Trefflich, Z. Wang, J. N. Weinstein, L. C. Westlake, J. Xu, L. Yang, C. Yau, Y. Zhao, and J. Z. and, "The chromatin accessibility landscape of primary human cancers," *Science*, vol. 362, no. 6413, Oct. 2018. [Online]. Available: https://doi.org/10.1126/science.aav1898

[12] e. a. Darryl Leja, Ewan Birney, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012. [Online]. Available: https://doi.org/10.1038/nature11247

[13] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edger: a bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, Nov. 2009. [Online]. Available: https://doi.org/10.1093/bioinformatics/btp616

[14] C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, and J. M. Engreitz, "Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations," *Nature Genetics*, vol. 51, no. 12, pp. 1664–1669, Nov. 2019. [Online]. Available: https://doi.org/10.1038/s41588-019-0538-0

[15] X. Wang, Z. Hu, T. Yu, Y. Wang, R. Wang, Y. Wei, J. Shu, J. Ma, and Y. Li, "Con-aae: contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration," *Bioinformatics*, vol. 39, no. 4, Mar. 2023. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btad162

[16] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: https://www.wandb.com/

[17] C. Wade and K. Glynn, *Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python*. Packt Publishing Ltd, Oct. 2020.

[18] P. Griffiths and J. Harris, *Principles of Algebraic Geometry*. John Wiley & Sons, Aug. 2014.

[19] H. Uno, T. Cai, M. J. Pencina, R. B. D'Agostino, and L. J. Wei, "On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data," *Statistics in Medicine*, vol. 30, no. 10, pp. 1105–1117, Jan. 2011. [Online]. Available: https://doi.org/10.1002/sim.4154

[20] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini, "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," *BMC Bioinformatics*, vol. 10, no. 1, Feb. 2009. [Online]. Available: https://doi.org/10.1186/1471-2105-10-48

[21] C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, and G. Bejerano, "GREAT improves functional interpretation of cis-regulatory regions," *Nature Biotechnology*, vol. 28, no. 5, pp. 495–501, May 2010. [Online]. Available: https://doi.org/10.1038/nbt.1630

[22] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, and F. Mahmood, "Scaling vision transformers to gigapixel images via hierarchical self-supervised learning," 2022. [Online]. Available: https://arxiv.org/abs/2206.02647