bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ECLARE: multi-teacher contrastive learning via ensemble distillation for diagonal integration of single-cell multi-omic data

Dylan Mann-Krzisnik [1,3,*] and Yue Li [1,2,3,*]

[1]Quantitative Life Sciences, McGill University, 550 Sherbrooke West Str., H3A 1E3, QC, Canada, [2]School of Computer Science, McGill University, 3480 University Str., H3A 2A7, QC, Canada and [3]Mila – Quebec AI Institute, 6666 Saint-Urbain Str., H2S 3H1, QC, Canada
*Corresponding authors: yueli@cs.mcgill.ca, dylan.mann-krzisnik@mail.mcgill.ca
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Integrating multimodal single-cell data, such as scRNA-seq and scATAC-seq, is key for decoding gene regulatory networks but remains challenging due to issues like feature harmonization and limited quantity of paired data. To address these challenges, we introduce ECLARE, a novel framework combining multi-teacher ensemble knowledge distillation with contrastive learning for diagonal integration of single-cell multi-omic data. ECLARE trains teacher models on paired datasets to guide a student model for unpaired data, leveraging a refined contrastive objective and transport-based loss for precise cross-modality alignment. Experiments demonstrate ECLARE's competitive performance in cell pairing accuracy, multimodal integration and biological structure preservation, indicating that multi-teacher knowledge distillation provides an effective mean to improve a diagonal integration model beyond its zero-shot capabilities. Additionally, we validate ECLARE's applicability through a case study on major depressive disorder (MDD) data, illustrating its capability to reveal gene regulatory insights from unpaired nuclei. While current results highlight the potential of ensemble distillation in multi-omic analyses, future work will focus on optimizing model complexity, dataset scalability, and exploring applications in diverse multi-omic contexts. ECLARE establishes a robust foundation for biologically informed single-cell data integration, facilitating advanced downstream analyses and scaling multi-omic data for training advanced machine learning models.

Key words: single-cell multi-omics, diagonal integration, ensemble knowledge distillation, contrastive learning

Introduction

The development of multimodal single-cell technologies has enabled the acquisition of molecular data across multiple genomic modalities simultaneously (Argelaguet et al., 2021). For example, the simultaneous acquisition of single-cell RNA sequencing (scRNA-seq) and single-cell Assay for Transposase-Accessible Chromatin sequencing (scATAC-seq) provides an advantageous way to study gene regulatory networks (Badia-i Mompel et al., 2023). However, these multimodal technologies are limited in the volume of data they can generate, hampering the ability of data-hungry machine learning models to leverage multi-omic data. One potential solution is to harmonize multi-omic datasets together into a single larger dataset, which requires the harmonization of the feature sets. Although this feature set harmonization is readily feasible for scRNA-seq gene expression data, it becomes less straightforward for scATAC-seq chromatin accessibility features. The main challenge with the latter is that individual scATAC-seq datasets originating from different studies identify their own unique set of differentially accessible chromatin regions, or *peaks* (Luecken et al., 2022). The simplest approach to harmonize the set of peaks across datasets is to either obtain the intersection or the union of all peaks. However, both approaches scale poorly when increasing the number of datasets: the harmonized peak set shrinks drastically when using the intersection, leaving out important cell type-specific peaks, whereas it expands drastically when using the union and leads to even more sparse count data while further compounding the curse of dimensionality.

Nonetheless, even with a proper peak set harmonization strategy, the total volume of single-cell multi-omic data falls far below than that of single-omic datasets used (for instance) to train cell atlas foundation models (Heimberg et al., 2024). This limits the potential use of paired multi-omic data to train state-of-the-art machine learning models that are currently in development for e.g. scRNA-seq data. One way to circumvent this issue of limited volume is to combine large-scale *single*-omic datasets into a single multi-omic dataset. The main issue in combining single-omic datasets to construct one multi-omic dataset is the potential divergence in experimental conditions, sample characteristics, and importantly, the different feature sets across modalities and lack of direct pairing information between cells. Hence, at this stage, small-scale *paired* data is replaced in favor of large-scale *unpaired* data. The steps taken to integrate cells with distinct features from unpaired multi-omic data can be summarized under the umbrella of *diagonal integration* (Argelaguet et al., 2021). Although multiple methods tackle diagonal integration, only a subset of these methods give rise to models capable of aligning cells from scRNA-seq and scATAC-seq data directly into an integrated embedding space, such that pairs of cells across modalities can be explicitly derived. Being able to pair cells directly across modalities allows to imitate paired multi-omic data, such that a given pair of cells can be considered as a single cell endowed with multi-omic features. This setup is often necessary to pursue multi-omic analyses such as gene network inference (Badia-i Mompel et al., 2023).

In the field of vision-language modeling, the Contrastive Language-Image Pretraining (CLIP) framework leverages multimodal contrastive learning to learn representations that aligns images with their textual descriptions (Radford et al., 2021). The CLIP framework promotes high generalizability

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and transferability, leading to impressive zero-shot prediction capabilities on unseen data. Thus, a CLIP-like approach could be employed to learn integrated multi-omic representations on paired data, and transfer these learned representations on unpaired data. Moreover, the contrastive objective of CLIP was shown by Radford et al. (2021) to be much more efficient at zero-shot transfer than a prediction objective, requiring much less data to achieve the same performance level. As such, multimodal contrastive learning provides a means to train a model on small-scale paired data and then perform zero-shot alignment on unpaired data to achieve diagonal integration. Then, once pairs of cells have been formed via this alignment, another CLIP-like model can be trained on these newly paired cells to directly learn aligned representations on the originally unpaired data. Similarly, the model trained on paired data can be treated as a *teacher* model used to guide the training of a *student* model that is trained on the unpaired data directly. The process of distilling the knowledge contained within a teacher model into a student model is termed *knowledge distillation* (KD - Hinton et al. (2015), Yang et al. (2024)).

However, for this KD approach to be truly scalable to state-of-the-art model scales, the issue of harmonizing ATAC peak sets across datasets to achieve integration of paired multi-omic datasets into one dataset still needs to be addressed. Here, an extension of KD from single-teacher KD to multi-teacher KD (aka *ensemble* KD - Asif et al. (2020); Vasu et al. (2024)) can be used to circumvent the ATAC peak set harmonization altogether. With multi-teacher KD, a teacher model can be trained independently for each paired ATAC-RNA dataset leading to a collection of teacher models. Then, KD can be performed between the teacher models and the student model simultaneously, effectively using all paired datasets to guide the contrastive objective applied onto the unpaired dataset.

In light of all these observations, we propose **E**nsemble knowledge distillation for **C**ontrastive **L**earning of **A**TAC and **R**NA **E**mbeddings – `ECLARE` – a diagonal integration method for aligning unpaired RNA and ATAC cells using multimodal contrastive learning and multi-teacher KD. We show the benefit of adding KD to a simple base model when integrating paired and unpaired data, yielding superior pairing accuracy, integration and cell-type clustering metrics. Lastly, we perform a simple GRN analysis on unpaired MDD nuclei to demonstrate the utility of `ECLARE` for biological research.

The ECLARE framework

We aim to achieve diagonal integration of unpaired single-cell/single-nucleus (sc/sn) RNA-seq (sc/snRNA-seq) and sc/snATAC-seq data. For this task we introduce `ECLARE`, an overview of which is provided in Fig. 1. Note that we will refer to cells and nuclei interchangeably.

Contrastive pretraining of individual teacher models

The first step of `ECLARE` shown in Fig. 1A is to independently train a set of teacher models on their respective paired RNA & ATAC multi-omic data (i.e. the *source* datasets) based on the CLIP framework by minimizing the InfoNCE loss (Radford

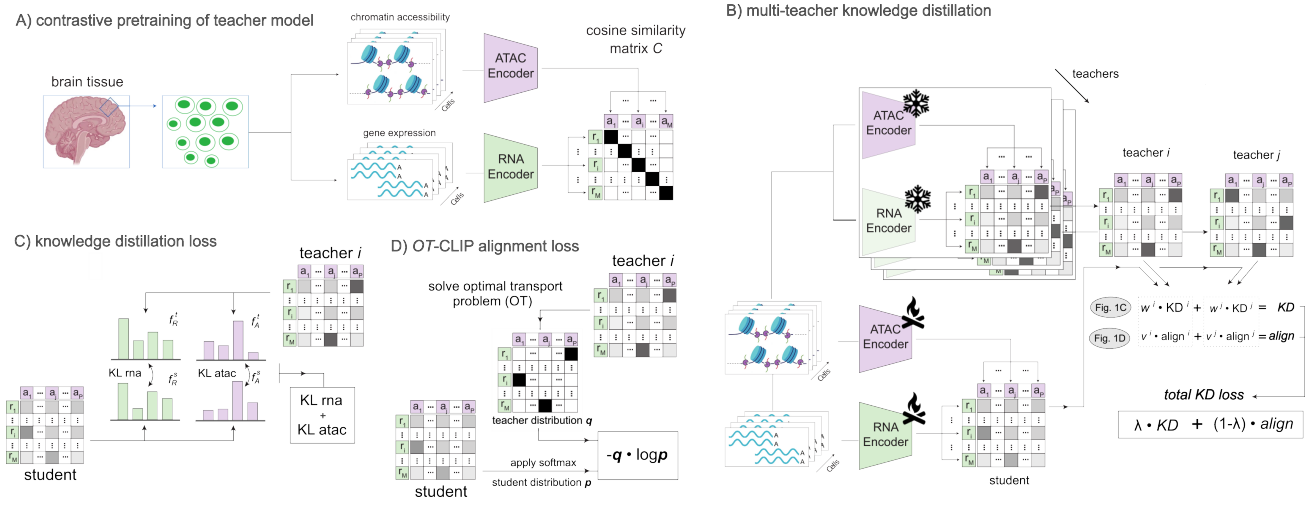


Fig. 1. Overview of `ECLARE`. **A) Contrastive pretraining of teacher model**. The first step of `ECLARE` is to train teacher models consisting of MLP encoders using paired multi-omic data (for ease of interpretation, only one teacher model is illustrated). The contrastive loss forces a diagonal structure of the cosine similarity matrix between cells & nuclei arising from ATAC-seq and RNA-seq. **B) Multi-teacher knowledge distillation**. After training the teacher models using source datasets, `ECLARE` trains a student model on a target dataset using multi-teacher knowledge distillation. For this, the target dataset is processed by the student model (trainable weights) and teacher models (frozen weights). The student model and each of the teacher models generate their respective ATAC-RNA cosine similarity matrix. The pair of cosine similarity matrices from each student-teacher pair are used to compute a weighted knowledge distillation (KD) loss and an alignment loss (*align*), where weights are teacher-dependent. These weighted KD and alignment losses are combined into a total KD loss. It is this total KD loss that is minimized to promote alignment between unpaired ATAC and RNA nuclei when embedded by the student model. **C) Knowledge distillation loss**. Computing the knowledge distillation loss involves deriving row-wise and column-wise softmax distributions from both student and teacher cosine similarity matrices, giving rise to distributions f_R^s, f_A^s, f_R^t and f_A^t. Separate Kullback-Leibler divergences (KL) are computed between the RNA distributions (f_R^s & f_R^t) and ATAC distributions (f_A^s & f_A^t). Both KL distributions are summed to give rise to the KD loss, which forces the student cosine similarity matrix to mimic the teacher one. **D) Alignment loss**. Computing the alignment loss involves using the teacher cosine similarity matrix to solve an Optimal Transport problem. This process results in a sparse similarity matrix containing unique one-to-one associations between rows (RNA nuclei) and columns (ATAC nuclei), which we call the teacher distribution (q) as its entries sum to unity. This teacher distribution is then used to compute cross-entropy with row-wise and column-wise student softmax distributions, resulting in two cross-entropy losses that are summed together (only one of these two losses is illustrated here). We name this alignment loss *OT*-CLIP.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

et al., 2021):

$$\mathcal{L} = -\frac{1}{M}\sum_{i=1}^{M}\log\frac{\exp\left(\cos(z_i, z_i^+)/\tau\right)}{\sum_{j=1}^{M}\exp\left(\cos(z_i, z_j)/\tau\right)} \tag{1}$$

where $z_i \in \{a_i, r_i\}$ represents the embedding of a given anchor cell from one modality (i.e. RNA cell r_i or ATAC cell a_i), $z_i^+ \in \{r_i, a_i\}$ represents the matching cell embedding from the other modality, z_j represents any cell from the modality opposite to the anchor (including z_i^+), cos represents the cosine similarity, τ the temperature hyperparameter (τ=1 for our experiments), and M the total number of cells for a given batch. The purpose of the InfoNCE loss is thus to optimize the encoders to maximize the diagonal entries of the cosine similarity matrix as illustrated in Fig. 1A. In this way, a given teacher model is trained to force alignment between matching RNA and ATAC cells leading to an integrated latent space, where pairs of matching cells share highly similar coordinates. Importantly, the encoders for each teacher can operate on a different set of features (i.e., genes and peaks) from each dataset and no feature harmonization is needed for this step.

Multi-teacher knowledge distillation

The second step of **ECLARE** is to distill the knowledge from this ensemble of trained teacher models into a student model using a separate *target* dataset (Fig. 1B). Importantly, this target dataset can be unpaired; this second step thus relies on the ability of the teacher models to perform multimodal alignment in a zero-shot fashion without relying on cell-pairing information. This KD step is performed by projecting the cells or nuclei from the target dataset through the encoders of the teacher and student models to construct cell-cell similarity matrices for each model. From these similarity matrices, two loss terms are computed for each student-teacher pair: a KD loss (Fig. 1C) and an alignment loss (Fig. 1D). These two losses are combined into a single total KD loss via a convex mixture (Fig. 1B), similar to the mixture of soft-labels and hard-labels losses initially proposed for KD (Hinton et al., 2015) or the mixture of KD and InfoNCE loss from CLIP (Vasu et al., 2024):

$$\mathcal{L} = \lambda \cdot \text{KD} + (1-\lambda)\cdot align \tag{2}$$

We evaluated coefficients $\lambda \in \{0.0, 0.1, 0.25, 0.5, 0.75, 0.9, 1.0\}$ and settled on λ=0.1 as the best value. Note that since the teacher model is used to project both source dataset during training and target dataset for inference, pairwise feature alignment is still required between source and target datasets. Still, the pairwise ATAC peak set alignment that we perform using intersection is customary and is much more suitable than the all-to-all alignment that we described previously.

For the KD loss illustrated in Fig. 1C, softmax is computed for all similarity matrices to obtain probabilities across rows and columns, such that the Kullback-Leibler (KL) divergence can be computed between the rows and columns of the student similarity matrix with those of each teacher similarity matrices. The KL divergences for RNA and ATAC are then summed to obtain a single KD loss per teacher:

$$\text{KD}_R^t = D_{\text{KL}}(f_R^t \parallel f_R^s) = \sum_{i=1}^{M} f_R^t(r_i)\log\frac{f_R^t(r_i)}{f_R^s(r_i)} \tag{3}$$

$$\text{KD}_A^t = D_{\text{KL}}(f_A^t \parallel f_A^s) = \sum_{j=1}^{P} f_A^t(a_j)\log\frac{f_A^t(a_j)}{f_A^s(a_j)} \tag{4}$$

$$\text{KD}^t = \text{KD}_R^t + \text{KD}_A^t \tag{5}$$

where f_R^t and f_R^s are the distributions obtained by computing row-wise softmax on the teacher and student cosine similarity matrices, f_A^t and f_A^s are the distributions obtained by computing column-wise softmax on the teacher and student cosine similarity matrices, and M and P are the number of RNA and ATAC nuclei (generally, $M = P$).

For the alignment loss illustrated in Fig. 1D, the teacher similarity matrix is refined by solving a classical exact Optimal Transport (OT) problem using the ot.solve() function from the Python OT library (POT - Flamary et al. (2021)). This is done by minimizing the Wasserstein distance between the scRNA-seq cells and scATAC-seq cells based on the cosine similarity computed by the teacher CLIP model:

$$W(u, v) = \min_{q\in\Pi(u,v)}\sum_{i=1}^{M}\sum_{j=1}^{P} c(r_i, a_j)\, q_{ij}, \tag{6}$$

$$\text{where } \Pi(u, v) = \left\{ q \in \mathbb{R}_+^{M\times P} \;\middle|\; q\,\mathbf{1}_M = u,\; q^\top \mathbf{1}_P = v \right\} \tag{7}$$

where c corresponds to the cosine distance defined as $c = 0.5[1 - \cos(x, y)]$. The constraints in Fig. (7) indicate that the elements of q (the *transport plan*) are positive-definite real numbers, and that summing across rows and columns matches the marginals u and v, which we both set as uniform distibutions for simplicity. Also, r_i and a_j correspond to embedded RNA and ATAC cells.

This process sparsifies the similarity matrix into a binary matrix (i.e. the transport plan q) to define unique pairs of cells across modalities. This teacher transport plan is then used as ground truth to compute the InfoNCE loss from the student similarity matrix, normalized via the softmax operation. By treating the teacher transport plan q as a bivariate teacher distribution and the student RNA and ATAC softmax distributions f_R^s and f_A^s, the alignment loss is computed with cross-entropy:

$$OT\text{-CLIP}_R^t = -\frac{1}{M\cdot P}\sum_{i=1}^{M}\sum_{j=1}^{P} q_{ij}(r_i, a_j)\log f_R^s(r_i) \tag{8}$$

$$OT\text{-CLIP}_A^t = -\frac{1}{M\cdot P}\sum_{i=1}^{M}\sum_{j=1}^{P} q_{ij}(r_i, a_j)\log f_A^s(a_j) \tag{9}$$

$$OT\text{-CLIP}^t = OT\text{-CLIP}_R^t + OT\text{-CLIP}_A^t \tag{10}$$

In other words, we compute the InfoNCE loss similarly to CLIP, but compute cross-entropy for entries of the student similarity matrix that correspond to the teacher transport plan rather than simply using the diagonal entries. Since we solve for an OT problem to obtain the teacher transport plan and then compute a loss similar to that of CLIP, we call this loss variant the OT-CLIP loss.

Computing teacher-specific weights

Both the teacher-specific KD and alignment losses from Eqs. (5) & (10) are combined across teachers to create ensemble multi-teacher KD and alignment losses via weighted averaging, as illustrated in Fig. 1B.

For KD, we use the teacher similarity matrices $C^t\ \forall\ t \in \{1, \ldots, T\}$ to quantify how well a given teacher is capable of aligning nuclei across modalities, where teachers with higher cosine similarity values in C^t see their KD loss upweighted, and vice-versa. Specifically, consider C^t and the transport plan q derived from C^t (i.e. Eq. (6)). The values in C^t that intersect

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the predicted matches in q – i.e. wherever $q > 0$ – can be extracted as follows:

$$\mathbb{I}(q > 0)_{ij} = \begin{cases} 1 & \text{if } q_{ij} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

$$\mathbf{c}^t = \text{vec}(C^t \odot \mathbb{I}(q > 0)) \tag{12}$$

where $\mathbf{c}^t$ represents a vector of cosine similarity values for the pairs predicted by teacher t. On the one hand, we can obtain weights based on the average value across the M pairs of nuclei as follows:

$$\bar{c}^t = \frac{1}{M} \sum_{i=1}^{M} \mathbf{c}_i^t \tag{13}$$

$$w^t = \text{softmax}(\bar{c}^t) = \frac{\exp(\bar{c}^t)}{\sum_{t'=1}^{T} \exp(\bar{c}^{t'})} \tag{14}$$

which provides a single weight for each teacher that is shared between all pairs of nuclei. On the other hand, the averaging across pairs from Eq. (13) can be withheld, giving rise to a weight for each teacher at each pair, indexed by e.g. RNA nuclei $i \in \{1, \ldots, M\}$:

$$w_i^t = \text{softmax}(\mathbf{c}_i^t) = \frac{\exp(\mathbf{c}_i^t)}{\sum_{t'=1}^{T} \exp(\mathbf{c}_i^t)} \tag{15}$$

such that $\mathbf{w}^t = (w_1^t, w_2^t, \ldots, w_M^t)$. In the same fashion, one can derive weights by indexing the ATAC nuclei $j \in \{1, \ldots, P\}$, leading to teacher weights $\mathbf{v}^t = (v_1^t, v_2^t, \ldots, v_P^t)$.

Because a teacher weight is defined for each RNA and ATAC nucleus, the KD losses defined in Eqs. (3)–(5) can be reweighted as follows:

$$\tilde{\text{KD}}_R^t = \sum_{i=1}^{M} w_i^t \cdot f_R^t(r_i) \log \frac{f_R^t(r_i)}{f_R^s(r_i)} \tag{16}$$

$$\tilde{\text{KD}}_A^t = \sum_{j=1}^{P} v_j^t \cdot f_A^t(a_j) \log \frac{f_A^t(a_j)}{f_A^s(a_j)} \tag{17}$$

$$\tilde{\text{KD}}^t = \tilde{\text{KD}}_R{}^t + \tilde{\text{KD}}_A{}^t \tag{18}$$

We adopt this nucleus-level teacher weighting in ECLARE and obtain the ensemble multi-teacher KD loss as follows:

$$\text{KD} = \sum_{t=1}^{T} \tilde{\text{KD}}^t \tag{19}$$

Lastly, the OT-CLIP alignment losses also require their teacher-specific weights. For computing these weights, we extract for each teacher the linear Wasserstein distance associated with solving the OT problem (Eq. 6) that is provided by the solver in the POT library, giving rise to a Wasserstein distance l^t for teacher t. Teacher weights are then obtained by computing $w_t = \text{softmax}(1 - l^t)$ for teachers $t \in \{1, \ldots, T\}$. Note that teacher reweighting of the alignment losses is performed for the whole batch of nuclei, such that the unweighted teacher alignment losses from Eq. (10) are weighted and summed:

$$OT\text{-CLIP} = \sum_{t=1}^{T} w^t \cdot OT\text{-CLIP}^t \tag{20}$$

Finally, the ensemble KD and OT-CLIP losses from equations Eqs. (19) & (20) are combined into the total KD loss of Eq. 2.

Experiments

To demonstrate the effectiveness of ECLARE for multi-omic integration of paired and unpaired data, we performed experiments to assess i) whether training a pair of RNA and ATAC encoders using contrastive learning on paired data similar to CLIP outperforms other methods in terms of zero-shot alignment of unseen data, ii) whether integrating KD improves upon the basic zero-shot approach, iii) whether KD further benefits from a multi-teacher approach compared to a single-teacher one, and iv) whether these observations are applicable to unpaired multi-omic data.

Single-nucleus multi-omic data

For these experiments, we make use of five multi-omic snATAC-seq and snRNA-seq datasets from brain tissue, summarized in Table 1. Of these six datasets, four are publicly-available *paired* datasets acquired using snMultiome or similar technologies and thus provide direct matching information between nuclei across both modalities. The other two datasets correspond to *unpaired* datasets from case and control subjects of a prior study on MDD for which no direct pairing information is available, thus requiring diagonal integration and to perform the case study described in a later section.

Baseline methods

CLIP The base model that serves as the building blocks of our ECLARE framework is a lightweight variant of scCLIP (Xiong et al., 2023). scCLIP trains ATAC and RNA encoders in a contrastive manner by maximizing similarity between a nucleus's RNA and ATAC embeddings, but it relies on artificially matched atlas-level data based on shared cell-type annotations. By contrast, we use genuinely paired multi-omic data (though in smaller quantity) and opt for simpler MLP encoders to maintain true cell-matching fidelity. We simply refer to this MLP-base model as CLIP.

Ablated ECLARE We also implemented an ablated version of ECLARE where the multi-teacher setup is replaced by a single-teacher one. We refer to this single-teacher variant as KD-CLIP.

MOJITOO (Cheng et al., 2022) trains and filters canonical correlation analysis (CCA) components based on paired data having undergone PCA (RNA) or LSI (ATAC) dimensionality reduction. **MultiVI** (Ashuach et al., 2023) trains a variational autoencoder model for each modality while also promoting alignment between variational disitributions for joint modeling of counts data. **GLUE** (Cao and Gao, 2022) also performs multimodal autoencoding while guiding the data reconstruction to mimic the feature-by-feature associations (e.g. genes-by-peaks) encoded within a knowledge graph. **scDART** (Zhang et al., 2022) performs diagonal integration with an MLP that directly maps an ATAC input layer onto an RNA input layer and forces manifold alignment between the embedded representations arising from the feature sets of both modalities. **scJoint** (Lin et al., 2022) is a two-step method where an integrated embeddings space is first used to transfer cell-type labels from RNA cells onto ATAC cells, followed by a refinement step to consolidate the ATAC embeddings afforded by the transferred labels.

Evaluation metrics

We assess the performance of ECLARE and other integration methods using metrics that capture nuclei pairing accuracy, degree of multimodal integration and conservation of biological

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Name of Dataset	Num. Nuclei	Tissue Source	Num. Cell Types	Reference
DLPFC_Anderson	56,737	DLPFC	8	Anderson et al. (2023)
Midbrain_Adams	50,596	midbrain	29	Adams et al. (2024)
DLPFC_Ma	56,938	DLPFC	28	Ma et al. (2022)
PFC_Zhu	31,213	DLPFC & APFC	15	Zhu et al. (2023)
MDD – RNA (n=37, m=44)	156,911	DLPFC	7	Maitra et al. (2023); Nagy et al. (2020)
MDD – ATAC (n=34, m=40)	201,456	DLPFC	7	Chawla et al. (2023)

Table 1. Summary of datasets. The horizontal dividing line serves to delineate paired (above) from unpaired (below) datasets. For MDD datasets, we report the number of case (n) and control subjects (m). DLPFC: dorsolateral prefrontal cortex, APFC: anterior prefrontal cortex

information. We provide an overview of these metrics here, although the interested reader can find more details in Luecken et al. (2022); Cao and Gao (2022). All metrics are computed based on embedded features, and most metrics are computed using the scib-metrics library (Luecken et al., 2022).

FOSCTTM The fraction of cells closest to the true match (FOSCTTM) quantifies for each cell from one modality (e.g. RNA cell r_i) the fraction of non-matching cells from the other modality ($a_j \ \forall \ j \neq i$) that are closest to the matching cell (a_i). This process is repeated for all cells from both modalities followed by global averaging of individual FOSCTTM scores. Because FOSCTTM requires ground-truth information on the true matches (or pairings) across modalities, it can only be evaluated on paired datasets. Note that we report 1-FOSCTTM, such that higher 1-FOSCTTM represents better pairing accuracy.

iLISI The integration local inverse Simpson's Index (iLISI) metric quantifies the amount of mixing occurring within local cell neighborhoods with regards to a fixed label associated with each cell. For example, the most common label for iLISI is the batch label such that high iLISI values correspond to strong mixing between cells from different batches, a good indication of batch effect removal. In the current study, we use the modality of origin as labels such that high iLISI corresponds to strong mixing between cells from different modalities, a good indication of multimodal integration. We use the KNN variant of iLISI using 30 neighbors, and opt for scaled iLISI to bound its value between 0 and 1.

ARI, NMI & **ASW** ARI and NMI are reference-based metrics that compare learned cluster labels to reference cell-type labels, differing in how similarity is computed: ARI assesses pairwise label consistency after adjusting for chance, while NMI measures mutual information between label distributions, normalized by entropy. ASW, unlike ARI and NMI, evaluates cluster compactness using only the data-driven labels, without reference labels. All three metrics rely on Leiden clustering with 30 neighbors.

Benchmarking study

For our benchmarking study, we compare single-cell multimodal integration methods on their ability to perform zero-shot alignment on unseen data. The motivation here is to see whether the base CLIP model behind `ECLARE` outperforms other published multi-omic integration methods. In this way, any improvements beyond this base model would showcase the effectiveness of multi-teacher KD even for models with strong transferability.

We first compare CLIP with the baseline methods MOJITOO, MultiVI, GLUE, scDART and scJoint using the metrics described in Section 3.3 on the 4 paired datasets described in Table 1. Each dataset is used as either a source dataset for training or a target dataset for evaluation. Thus, the task consists of performing all 12 unique combinations of source-to-target evaluations. Moreover, each experiment from each of the 12 source-to-target combinations is repeated 3 times by resampling the source and target datasets, similar to cross-validation folds. Hence, a total of 36 experiments are performed for each method. For each experiment, the evaluation metrics are applied onto RNA and ATAC nuclei that have been embedded by the method in question.

We then make similar comparisons between CLIP, KD-CLIP and `ECLARE` on the same 4 paired datasets. Note that for a given target dataset, `ECLARE` consumes all the other 3 source datasets simultaneously due to its multi-teacher mechanism, thus reducing its number of experiments from 36 to 12. These experiments serve to compare the performance of single-teacher and multi-teacher KD compared to performing zero-shot alignment with CLIP.

Finally, we repeat experiments with CLIP, KD-CLIP and `ECLARE` by treating the unpaired MDD data as the target dataset and using the 4 paired datasets as source datasets. Here, the number of source-to-target combinations reduces from 12 to 4 for CLIP and KD-CLIP and from 4 to 1 for `ECLARE`. Due to the 3 validation folds, a total of 12 experiments are performed for CLIP and KD-CLIP and 3 experiments for `ECLARE`. Note that we do not report 1-FOSCTTM for these experiments due to the lack of ground-truth nuclei pairing information.

Gene regulatory network analysis on MDD data

Once `ECLARE` is trained on MDD, it can be used to project ATAC and RNA nuclei into a common embedding space and derive nuclei pairings by solving the OT problem formalized in Eq. (6) and identifying unique pairs contained within the transport plan. Cosine similarities are then computed for all gene-peak combinations across the paired nuclei as a measure of the degree of association between genes and peaks, creating a gene regulatory network (GRN). For each pair of peak p and gene g, we averaged the cosine similarity scores for the MDD subjects $\bar{s}_{g,p}^{(case)}$ and for the controls subjects $\bar{s}_{g,p}^{(ctrl)}$ and computed log fold-change with 1 padding:

$$\log \mathrm{FC}_{g,p} = \log \frac{\bar{s}_{g,p}^{(case)} + 1}{\bar{s}_{g,p}^{(ctrl)} + 1} \tag{21}$$

The absolute values of $\log \mathrm{FC}$ are summed across peaks giving rise to a summed $\log \mathrm{FC}$ value per gene denoted as ϕ_g, i.e., $\phi_g = \sum_{p \in \mathcal{P}} |\log \mathrm{FC}_{gp}|$ for ATAC peak-set $\mathcal{P}$. The value of ϕ_g measures gene-level MDD-association. We focused our analysis specifically on microglia from female subjects as strong evidence has been found in Maitra et al. (2023) for this cell-type and sex intersection. We thus isolated for microglia 1049 nuclei sampled from female cases and 957 nuclei sampled from female controls. We verify whether ϕ_g for the most important genes

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reported previously are ranked highly with respect to the overall distribution of ϕ_g values across all genes contained in the MDD data. Specifically, we investigated genes *SLIT3*, *ROBO2*, *ADAMTSL1*, *THSD4*, *SOCS3*, *SPP1*, *GAS6* and *MERTK*.

Results

As an overview, we first show how a CLIP base model compares to other baseline methods (Fig. 2). We then show the benefit of integrating KD via single-teacher (KD-CLIP) and multi-teacher (`ECLARE`) mechanisms when applied onto paired data (Fig. 3). We also provide results for unpaired MDD data to explore how KD and its extension to multiple teachers is beneficial to diagonal integration of unpaired snATAC- and snRNA-seq (Fig. 4) and its applicability to GRN analysis (Fig. 6).

Benchmarking study of zero-shot alignment task

In Fig. 2, we compare CLIP with MOJITOO, multiVI, GLUE, scDART and scJoint on nucleus pairing, integration and clustering metrics using paired data.

Given the paired nature of the datasets for these experiments, we consider 1-FOSCTTM as the key ground truth metrics, which quantifies the quality of nuclei pairing across ATAC and RNA modalities with respect to ground truth pairings. We see from Fig. 2A that CLIP provides high-quality pairings based on a median value of 0.88, outperforming all other baseline methods.

We also resort to iLISI to assess the quality of the broader integration and alignment of nuclei from both modalities within the shared latent space (Fig. 2B). One striking feature of the boxplots comparing iLISI values across methods is the abundance of (near-)zero values for most baseline methods. Indeed, this suggests that most methods do not (in these instances) inherently force alignment between modalities when investigating the held-out target nuclei. Otherwise, only scDART shows a median iLISI value exceeding that of CLIP.

We also compare methods in light of clustering methods to investigate how well the cell type-based biological structure is conserved when applying methods on target data in a zero-shot manner. Clearly, CLIP outperforms all other methods in terms of ARI (Fig. 2C) and NMI (Fig. 2D) reference-based clustering metrics. However, CLIP underperforms GLUE, scDART and scJoint on the reference-free ASW metric (Fig. 2E).

Based on these results, we conclude that CLIP is a suitable base model for KD in terms of pairing quality, integration quality and biological conservation.

Benchmarking of KD using paired data

We investigate whether incorporating single-teacher (KD-CLIP) and multi-teacher (`ECLARE`) KD improves the performance of the base CLIP model. Results are shown in Fig. 3.

The baseline CLIP achieves median 1-FOSCTTM score of 0.88 on target data and is further increased by single-teacher and multi-teacher KD (Fig. 3A). Both of the latter achieve median scores around 0.91, although `ECLARE` demonstrates lower variability than KD-CLIP. Note that the maximal 1-FOSCTTM score is 1, around which the CLIP-based models are approaching a saturation point.

On the other hand, the median iLISI score of 0.31 for CLIP shows much room for improvement. Indeed, KD-CLIP and `ECLARE` show a marked increase in iLISI, respectively attaining values of 0.66 and 0.64 (Fig. 3B).

In terms of ARI and NMI clustering metrics, there is a cumulative increase when incorporating single-teacher (KD-CLIP) and then multi-teacher (`ECLARE`) KD from median values of 0.55 to 0.58 and 0.69 for ARI and from 0.63 to 0.65 and 0.70 for NMI (Figs. 3C & 3D). It is noteworthy that the distributions of ARI and NMI values are dominated by inter-target differences where notable differences can be observed between targets, i.e. the trends across methods with respect to color. Lastly, we also observed improvements in ASW for both KD-CLIP and `ECLARE` compared to CLIP from 0.49 to 0.53 and 0.54 (Fig. 3E), only slightly lower than the ASW for GLUE, scDART and scJoint from Fig. 2E.

Applications to unpaired MDD multi-omic data

Due to the lack of ground-truth knowledge of nuclei pairing for unpaired MDD data, we could not use 1-FOSCTTM to assess pairing accuracy and instead consider iLISI as the prime indicator of the quality of integration. KD provides a net increase in iLISI that is more pronounced for KD-CLIP than it is for `ECLARE` (Fig. 4A), from 0.31 for CLIP to 0.61 and 0.55, respectively. However, `ECLARE` outperforms KD-CLIP in terms of ARI, NMI and ASW (Figs. 4B-4D). Therefore, `ECLARE` provides better integration than CLIP while also maintaining compact and biologically-meaningful clusters.

To support the numerical metrics, we also provide UMAP embeddings of held-out ATAC and RNA nuclei integrated within `ECLARE`'s embedding space in Fig. 5 where we observe that nuclei align closely across modality and preserve important cell-type information.

We then examined the distribution of gene-level MDD-association scores (i.e., ϕ_g) computed from nuclei from microglia amongst female subjects (Fig. 6). Some of the known MDD-associated genes (Maitra et al., 2023) such as *THSD4*, *ROBO2* and *GAS6* exhibit very high gene scores and rank among the top percentiles – 96%, 98% & 99%, respectively – indicating that the nuclei pairings obtained with `ECLARE` lead to strong MDD-related biological signals.

Discussion

We demonstrated the effectiveness of `ECLARE` by first establishing that a simple MLP trained with contrastive learning (CLIP) serves as a good base model when evaluated on a zero-shot alignment task (Fig. 2). We then evaluated the merits of single-teacher (KD-CLIP) and multi-teacher (`ECLARE`) KD as improvements beyond zero-shot alignment on both paired (Fig. 3) and unpaired (Fig. 4) datasets. We then showcased a simple GRN-inference analysis on unpaired MDD data to demonstrate a real-world application using our method (Fig. 6). Hence, we extend the general research direction of contrastive learning and KD for multi-omic integration (Xiong et al., 2023; Gossi et al., 2023; Yang et al., 2022) by showing via `ECLARE` that incorporating multi-teacher KD leads to a more capable and reliable solution for achieving diagonal integration.

We can explore several future directions. First, we can develop more effective algorithms for weighing the different teachers as a function of cell-type or class imbalances, and implement more sophisticated architectures such as Transformer (Xiong et al., 2023) to increase model capability and transferability.

Second, `ECLARE` can be adapted to other multi-omics datasets that further incorporate other modalities such as proteomics and methylomics. It is also worth investigating whether `ECLARE`

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

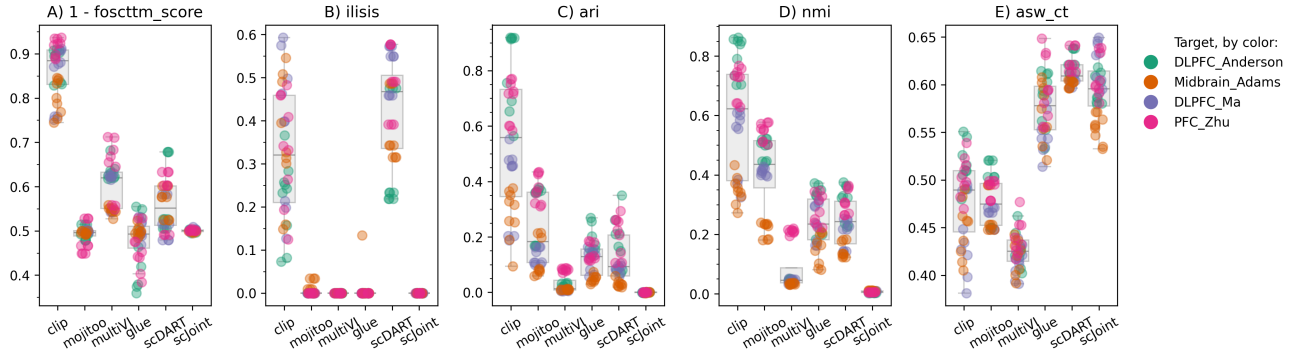


Fig. 2. Benchmarking results on zero-shot alignment task with paired data to compare different baseline methods. Each panel corresponds to an evaluation metric (Section 3.3). Given a total of 4 paired datasets used for these experiments, there are 12 unique combinations of source-to-target zero-shot alignment experiments. Data points are color-coded by the target dataset at hand for a given experiment. Each of the 12 experiments are repeated 3 times to give rise to different train/test splits of the source and target data, giving rise to a total of 36 experiments per method.

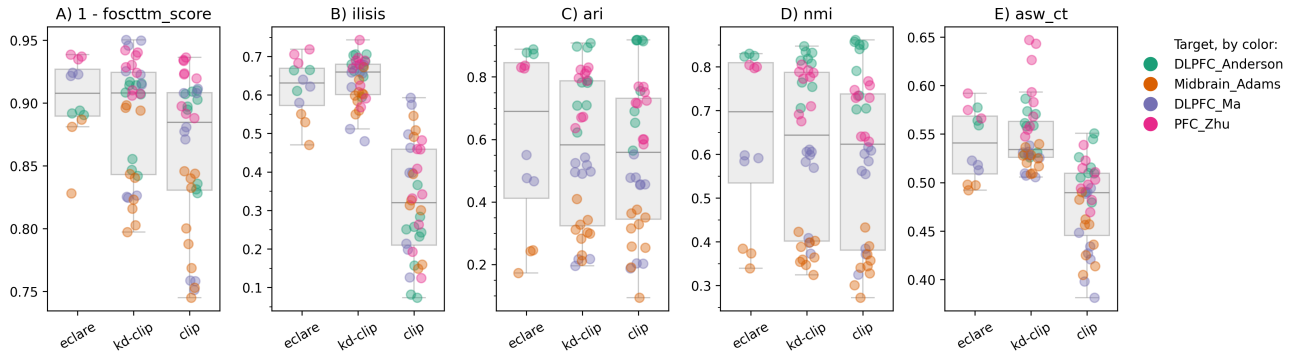


Fig. 3. Comparison of ECLARE, KD-CLIP & CLIP methods on paired data. Each panel corresponds to an evaluation metric (Section 3.3). Since ECLARE's multi-teacher approach consumes all source datasets for training, the number of source-to-target combinations are limited to the number of target datasets, i.e. n=4. Considering the 3 train/test splits, the total number of experiments is 36 for CLIP & KD-CLIP and 12 for ECLARE.

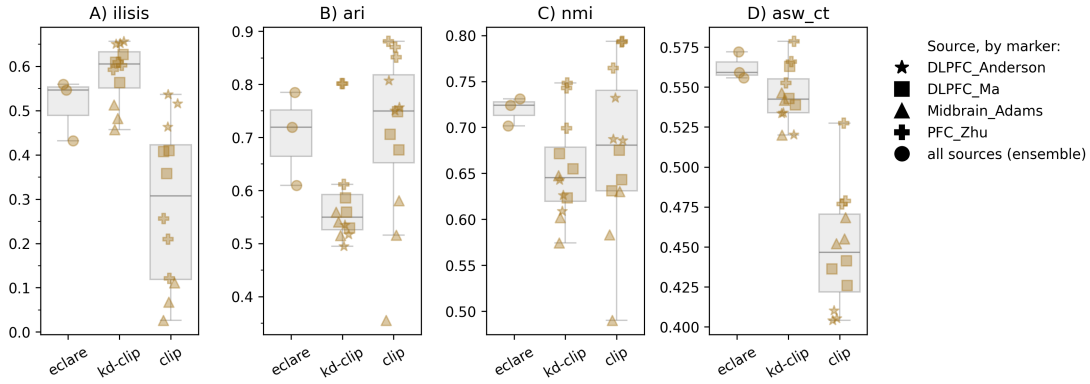


Fig. 4. Comparison of ECLARE, KD-CLIP & CLIP methods on unpaired MDD data. Each panel corresponds to an evaluation metric (Section 3.3 - excluding 1-FOSCTTM). Here, data points are coded by source dataset via marker shape since MDD is the sole target data. For ECLARE, we refer to the source dataset as "all sources (ensemble)" since all source datasets are consumed. Experiments are repeated for 3 train/test splits, giving rise to 12 experiments for KD-CLIP & CLIP and 3 experiments for ECLARE.

would further benefit from a mosaic integration approach, where datasets are only partially unpaired and thus contain overlapping cells and/or features (Argelaguet et al., 2021).

Third, other training paradigms can be incorporated to further boost the performance of ECLARE. For example, the contrastive teacher pretraining step could be preceded by a fully unsupervised reconstruction task to better learn the within-modality data distribution (Rosen et al., 2024).

Lastly, GRN analysis can be expanded to identify specific peak-gene linkages. Similar to GLUE, prior knowledge about feature associations such as in-cis regulatory interactions could be incorporated within the training paradigm to stabilize training and providing more interpertable results. Standard

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

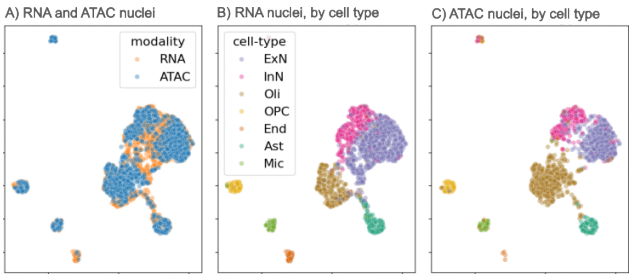


Fig. 5. UMAP visualizations of held-out MDD nuclei embedded using ECLARE. A) RNA and ATAC nuclei colored by modality. B) RNA nuclei colored by cell-type labels. C) ATAC nuclei colored by cell-type labels.

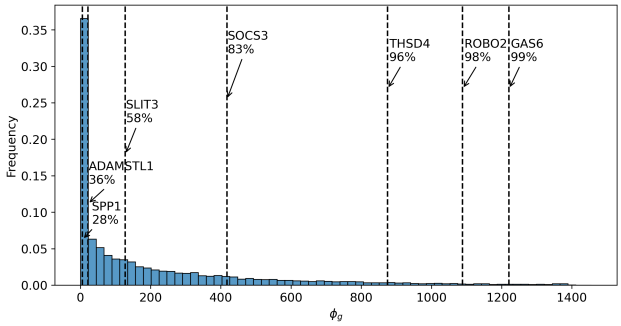


Fig. 6. Gene regulatory network analysis of MDD data. This histogram shows the distribution of MDD-association scores ϕ_g across all genes for microglia from female subjects. Vertical lines demarcate ϕ_g values for genes reported to be affected by female MDD subjects in microglia, with annotations for gene name and the percentile of that gene's ϕ_g value.

scATAC-seq processing librairies such as ArchR (Granja et al., 2021) and Signac (Stuart et al., 2021) contain commands for inferring peak-gene links, facilitating the use of such prior information for developing better diagonal integration methods.

In conclusion, we demonstrate that KD is a useful paradigm for improving single-nuclei integration quality beyond zero-shot transfer on paired and unpaired multi-omic data. Moreover, extending KD to a multi-teacher setup provides a more biologically reliable integration strategy by using a principled way of leveraging multiple paired datasets. This solution also mitigates the ATAC peak alignment issue, therefore enabling to scale KD with increasing data volumes needed to train state-of-the-art multiomic machine learning models. aling with increasing data volume

Code & data availability

Code is made available at github.com/li-lab-mcgill/ECLARE. Data are publicly accessible via the references listed in Table 1.

References

L. Adams, M. K. Song, S. Yuen, Y. Tanaka, and Y.-S. Kim. A single-nuclei paired multiomic analysis of the human midbrain reveals age- and Parkinson's disease–associated glial changes. *Nature aging*, 4(3):364–378, Mar. 2024. ISSN 2662-8465. doi: 10.1038/s43587-024-00583-6. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11361719/.

A. G. Anderson, B. B. Rogers, J. M. Loupe, I. Rodriguez-Nunez, S. C. Roberts, L. M. White, J. N. Brazell, W. E. Bunney, B. G. Bunney, S. J. Watson, J. N. Cochran, R. M. Myers, and L. F. Rizzardi. Single nucleus multiomics identifies ZEB1 and MAFB as candidate regulators of Alzheimer's disease-specific cis-regulatory elements. *Cell Genomics*, 3(3), Mar. 2023. ISSN 2666-979X. doi: 10.1016/j.xgen.2023.100263. URL https://www.cell.com/cell-genomics/abstract/S2666-979X(23)00019-8. Publisher: Elsevier.

R. Argelaguet, A. S. E. Cuomo, O. Stegle, and J. C. Marioni. Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10):1202–1215, Oct. 2021. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-00895-7. URL https://www.nature.com/articles/s41587-021-00895-7.

T. Ashuach, M. I. Gabitto, R. V. Koodli, G.-A. Saldi, M. I. Jordan, and N. Yosef. MultiVI: deep generative model for the integration of multimodal data. *Nature Methods*, 20 (8):1222–1231, Aug. 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01909-9. URL https://www.nature.com/articles/s41592-023-01909-9. Publisher: Nature Publishing Group.

U. Asif, J. Tang, and S. Harrer. Ensemble Knowledge Distillation for Learning Improved and Efficient Networks, Apr. 2020. URL http://arxiv.org/abs/1909.08097. arXiv:1909.08097.

P. Badia-i Mompel, L. Wessels, S. Müller-Dott, R. Trimbour, R. O. Ramirez Flores, R. Argelaguet, and J. Saez-Rodriguez. Gene regulatory network inference in the era of single-cell multi-omics. *Nature Reviews Genetics*, 24(11):739–754, Nov. 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00618-5. URL https://www.nature.com/articles/s41576-023-00618-5. Publisher: Nature Publishing Group.

Z.-J. Cao and G. Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, Oct. 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01284-4. URL https://www.nature.com/articles/s41587-022-01284-4. Publisher: Nature Publishing Group.

A. Chawla, D. Cakmakci, W. Zhang, M. Maitra, R. Rahimian, H. Mitsuhashi, M. Davoli, J. Yang, G. G. Chen, R. Denniston, D. Mash, N. Mechawar, M. Suderman, Y. Li, C. Nagy, and G. Turecki. Differential Chromatin Architecture and Risk Variants in Deep Layer Excitatory Neurons and Grey Matter Microglia Contribute to Major Depressive Disorder, Oct. 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.10.02.560567.

M. Cheng, Z. Li, and I. G. Costa. MOJITOO: a fast and universal method for integration of multimodal single-cell data. *Bioinformatics*, 38(Supplement_1):i282–i289, June 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac220. URL https://doi.org/10.1093/bioinformatics/btac220.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. H. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. ISSN 1533-7928. URL http://jmlr.org/papers/v22/20-451.html.

F. Gossi, P. Pati, P. Chouvardas, A. L. Martinelli, M. Kruithof-de Julio, and M. A. Rapsomaniki. Matching single cells across modalities with contrastive learning and optimal transport. *Briefings in Bioinformatics*, 24(3):bbad130, May 2023. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbad130. URL https://academic.oup.com/bib/article/doi/10.1093/bib/bbad130/7147026.

bioRxiv preprint doi: https://doi.org/10.1101/2025.01.24.634799; this version posted January 27, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

J. M. Granja, M. R. Corces, S. E. Pierce, S. T. Bagdatli, H. Choudhry, H. Y. Chang, and W. J. Greenleaf. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3): 403–411, Mar. 2021. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-021-00790-6. URL https://www.nature.com/articles/s41588-021-00790-6.

G. Heimberg, T. Kuo, D. J. DePianto, O. Salem, T. Heigl, N. Diamant, G. Scalia, T. Biancalani, S. J. Turley, J. R. Rock, H. Corrada Bravo, J. Kaminker, J. A. Vander Heiden, and A. Regev. A cell atlas foundation model for scalable search of similar human cells. *Nature*, Nov. 2024. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-024-08411-y. URL https://www.nature.com/articles/s41586-024-08411-y.

G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network, Mar. 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531 [stat].

Y. Lin, T.-Y. Wu, S. Wan, J. Y. H. Yang, W. H. Wong, and Y. X. R. Wang. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nature Biotechnology*, 40(5):703–710, May 2022. ISSN 1087-0156, 1546-1696. doi: 10.1038/s41587-021-01161-6. URL https://www.nature.com/articles/s41587-021-01161-6.

M. D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and F. J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, Jan. 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8. URL https://www.nature.com/articles/s41592-021-01336-8. Publisher: Nature Publishing Group.

S. Ma, M. Skarica, Q. Li, C. Xu, R. D. Risgaard, A. T. N. Tebbenkamp, X. Mato-Blanco, R. Kovner, Krsnik, X. De Martin, V. Luria, X. Martí-Pérez, D. Liang, A. Karger, D. K. Schmidt, Z. Gomez-Sanchez, C. Qi, K. T. Gobeske, S. Pochareddy, A. Debnath, C. J. Hottman, J. Spurrier, L. Teo, A. G. Boghdadi, J. Homman-Ludiye, J. J. Ely, E. W. Daadi, D. Mi, M. Daadi, O. Marín, P. R. Hof, M.-R. Rasin, J. Bourne, C. C. Sherwood, G. Santpere, M. J. Girgenti, S. M. Strittmatter, A. M. M. Sousa, and N. Sestan. Molecular and cellular evolution of the primate dorsolateral prefrontal cortex. *Science*, 377(6614):eabo7257, Sept. 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abo7257. URL https://www.science.org/doi/10.1126/science.abo7257.

M. Maitra, H. Mitsuhashi, R. Rahimian, A. Chawla, J. Yang, L. M. Fiori, M. A. Davoli, K. Perlman, Z. Aouabed, D. C. Mash, M. Suderman, N. Mechawar, G. Turecki, and C. Nagy. Cell type specific transcriptomic differences in depression show similar patterns between males and females but implicate distinct cell types and genes. *Nature Communications*, 14(1):2912, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38530-5.

C. Nagy, M. Maitra, A. Tanti, M. Suderman, J.-F. Théroux, M. A. Davoli, K. Perlman, V. Yerko, Y. C. Wang, S. J. Tripathy, P. Pavlidis, N. Mechawar, J. Ragoussis, and G. Turecki. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nature Neuroscience*, 23(6):771–781, June 2020. ISSN 1546-1726. doi: 10.1038/s41593-020-0621-y. URL https://www.nature.com/articles/s41593-020-0621-y. Publisher: Nature Publishing Group.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Y. Rosen, M. Brbić, Y. Roohani, K. Swanson, Z. Li, and J. Leskovec. Toward universal cell embeddings: integrating single-cell RNA-seq datasets across species with SATURN. *Nature Methods*, 21(8):1492–1500, Aug. 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02191-z. URL https://www.nature.com/articles/s41592-024-02191-z. Publisher: Nature Publishing Group.

T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, and R. Satija. Single-cell chromatin state analysis with Signac. *Nature Methods*, 18(11):1333–1341, Nov. 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01282-5. URL https://www.nature.com/articles/s41592-021-01282-5. Publisher: Nature Publishing Group.

P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel. MobileCLIP: Fast Image-Text Models through Multi-Modal Reinforced Training, Apr. 2024. URL http://arxiv.org/abs/2311.17049. arXiv:2311.17049.

L. Xiong, T. Chen, and M. Kellis. scCLIP: Multi-modal Single-cell Contrastive Learning Integration Pre-training. Oct. 2023. URL https://openreview.net/forum?id=KMtM5ZHxct.

C. Yang, Z. An, L. Huang, J. Bi, X. Yu, H. Yang, B. Diao, and Y. Xu. CLIP-KD: An Empirical Study of CLIP Model Distillation, May 2024. URL http://arxiv.org/abs/2307.12732. arXiv:2307.12732.

M. Yang, Y. Yang, C. Xie, M. Ni, J. Liu, H. Yang, F. Mu, and J. Wang. Contrastive learning enables rapid mapping to multimodal single-cell atlas of multimillion scale. *Nature Machine Intelligence*, 4(8):696–709, Aug. 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00518-z. URL https://www.nature.com/articles/s42256-022-00518-z. Publisher: Nature Publishing Group.

Z. Zhang, C. Yang, and X. Zhang. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. *Genome Biology*, 23(1):139, June 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02706-x. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02706-x.

K. Zhu, J. Bendl, S. Rahman, J. M. Vicari, C. Coleman, T. Clarence, O. Latouche, N. M. Tsankova, A. Li, K. J. Brennand, D. Lee, G.-C. Yuan, J. F. Fullard, and P. Roussos. Multi-omic profiling of the developing human cerebral cortex at the single-cell level. *Science Advances*, 9(41):eadg3754, Oct. 2023. doi: 10.1126/sciadv.adg3754. URL https://www.science.org/doi/10.1126/sciadv.adg3754. Publisher: American Association for the Advancement of Science.