https://doi.org/10.1093/bib/bbad130
Problem Solving Protocol

Matching single cells across modalities with contrastive learning and optimal transport

Federico Gossi, Pushpak Pati, Panagiotis Chouvardas, Adriano Luca Martinelli, Marianna Kruithof-de Julio and Maria Anna Rapsomaniki

Corresponding authors: Federico Gossi and Maria Anna Rapsomaniki, IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland.
E-mails: federico.gossi@alumni.ethz.ch; aap@zurich.ibm.com

Abstract

Understanding the interactions between the biomolecules that govern cellular behaviors remains an emergent question in biology. Recent advances in single-cell technologies have enabled the simultaneous quantification of multiple biomolecules in the same cell, opening new avenues for understanding cellular complexity and heterogeneity. Still, the resulting multimodal single-cell datasets present unique challenges arising from the high dimensionality and multiple sources of acquisition noise. Computational methods able to match cells across different modalities offer an appealing alternative towards this goal. In this work, we propose MatchCLOT, a novel method for modality matching inspired by recent promising developments in contrastive learning and optimal transport. MatchCLOT uses contrastive learning to learn a common representation between two modalities and applies entropic optimal transport as an approximate maximum weight bipartite matching algorithm. Our model obtains state-of-the-art performance on two curated benchmarking datasets and an independent test dataset, improving the top scoring method by 26.1% while preserving the underlying biological structure of the multimodal data. Importantly, MatchCLOT offers high gains in computational time and memory that, in contrast to existing methods, allows it to scale well with the number of cells. As single-cell datasets become increasingly large, MatchCLOT offers an accurate and efficient solution to the problem of modality matching.

Keywords: single-cell data integration, modality matching, contrastive learning, optimal transport

INTRODUCTION

Single cells are complex dynamical systems where a variety of biomolecules interact in a coordinated way to produce robust and adaptive behaviors. At the same time, many of the underlying mechanisms that govern fundamental cellular functions, such as replication of DNA, transcription to RNA and translation to proteins, are intrinsically stochastic [1–3]. This stochasticity allows single cells to differentiate, forming diverse tissues, organs and, ultimately, whole organisms [4]. Understanding the interactions between these biomolecules at the intra- and inter-cellular level is a longstanding question in biology. Such a holistic view on cellular state and associated stochasticity can not only provide mechanistic insights on vital cellular functions but can also reveal how heterogeneity is linked to disease [5].

Recent advances in single-cell technologies have made it possible to quantify different combinations of (epi)genomic, transcriptomic and proteomic profiles in the same cell [6]. Integrated assays and workflows that allow simultaneous measurement of gene expression with chromatin accessibility [7], DNA methylation [8] or protein abundance [9, 10] are rapidly gaining popularity. Computational analysis of the resulting multiomic measurements has the potential to capture multiple omic views that collectively determine cellular state and thus elucidate cell complexity and heterogeneity at an unprecedented scale. However, a number of emerging challenges associated with the heterogeneity, measurement noise, batch effects and missing information in the resulting datasets limit this potential [11, 12]. As a result, to date most single-cell datasets are independently

Federico Gossi obtained his Master's Degree in Computer Science at ETH Zurich in 2022. He worked on machine learning for single-cell multi-omics with the groups at IBM Research Europe in Zurich and at the Urology Research Laboratory at the University of Bern.
Pushpak Pati is a postdoctoral researcher at IBM Research Europe in Zurich, Switzerland. He specializes in machine learning and computational biology, with a special focus on analysing the spatial multi-omics profiles of cells in tumor microenvironments using graph theory for various cancer types.
Panagiotis Chouvardas is a Computational Biology PostDoctoral Researcher at the Urology Research Laboratory, Department for BioMedical Research, University of Bern, Switzerland. He specializes in the analysis and interpretation of Next Generation Sequencing experiments with a special interest in single-cell multi-omics.
Adriano Luca Martinelli is a pre-doctoral student at IBM Research Europe in Zurich and ETH Zurich. His work focuses on analysing tumour heterogeneity in various omics modalities and deepen our understanding of how this heterogeneity influences patient outcome and therapeutic response.
Marianna Kruithof-de Julio is Professor at the Department for BioMedical Research, Urology Research Laboratory, University of Bern, Bern Switzerland. Her laboratory focuses on developing and applying tools for precision medicine. Her research group is aimed at understanding cancer from a multitude of angles: she focuses on the tumor cells, the stroma, the immune cells, and the vasculature.
Marianna Rapsomaniki is a Group Leader at IBM Research Europe in Zurich. The overarching goal of her research is modeling spatiotemporal tumor heterogeneity across different scales of biological organization, and understanding how it affects cancer initiation, progression, and response to treatment. To achieve this, her team combines artificial intelligence and machine learning approaches to develop computational methods able to extract biologically meaningful patterns from large-scale, multimodal, and noisy single-cell data, with or without spatial resolution.

Received: January 09, 2023. **Revised:** February 25, 2023. **Accepted:** March 14, 2023

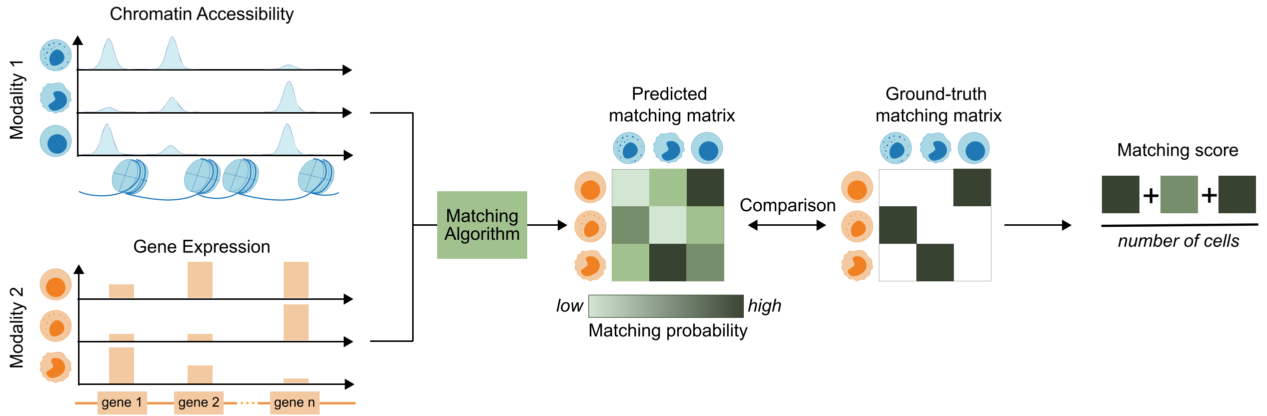

Figure 1. Modality matching task in single-cell data integration. Given n single-cell profiles of two omic modalities (here, chromatin accessibility in blue and gene expression in orange), a matching algorithm results in an $n \times n$ matrix that contains matching probabilities for all cell pairs. Comparing with the ground-truth cell correspondence allows us to compute a matching score, shown here as the average matching probability of all real cell pairs.

generated, resulting in unimodal, unpaired datasets with no cell–cell correspondence.

To address this limitation, a number of computational methods that attempt to diagonally integrate the unpaired unimodal single-cell datasets have been proposed. Several of these methods aim to align the unimodal datasets by projecting them in a joint embedding feature space using common linear dimensionality reduction approaches, such as principal component analysis (PCA) (e.g. Harmony [13]), canonical correlation analysis (CCA) (e.g. Seurat v3 [14] and bind-SC [15]) and non-negative matrix factorization (NMF) (e.g. LIGER [16] and Online iNMF [17]). An appealing alternative to overcome the need for aligning the unimodal datasets is offered by computational methods able to perform *modality matching*, i.e. pair single-cell profiles from different omic modalities (Figure 1). Notable methods in this category include UnionCom [18], Pamona [19], SCOT [20], SCOOTR [21], MMD-MA [22] and GLUE [23]. One main challenge of modality matching stems from the indistinguishability of cells of the same type, whose variations in gene expression or protein abundance are attributed to intrinsic noise and stochastic fluctuations [24]. However, if successful, such a matching could provide insights into molecular features and mechanisms that govern this stochasticity. Importantly, an effective matching prediction model could be used to computationally integrate the increasing number of generated unimodal single-cell datasets to yield valuable multimodal datasets.

Towards this goal, the single-cell community launched a multimodal single-cell data integration competition at NeurIPS 2021 [24], with modality matching being one of the main tasks. The organizers generated a unique curated dataset by profiling bone marrow mononuclear cells (BMMCs) collected from 12 donors at 4 data generation sites using two multiomic single-cell technologies: cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq), which captures single-cell RNA gene expression (GEX) and surface protein levels as antibody-derived tags (ADT) [9]; and 10X Multiome assay, which integrates the assay for transposase-accessible chromatin (ATAC) for chromatin accessibility with single-nucleus RNA gene expression levels (GEX) [25]. As both the CITE-seq and 10X Multiome benchmarking datasets are paired multimodal assays, the actual ground-truth modality matchings are known and can be used to evaluate method performance. To date, this is the largest realistic benchmarking dataset available for multimodal single-cell data integration.

Among a total of 462 submissions to the competition from 23 teams, the winning model in all subtasks was proposed by Team CLUE [26]. CLUE's architecture is based on modality-specific variational autoencoders that learn low-dimensional embeddings both within and across modalities. The second best-scoring model was proposed by Team Novel [24], and was based on CLIP [27], a popular contrastive learning model that learns a common representation between text and images. Team Novel's model uses two modality-specific encoders that are trained using a contrastive learning approach to generate similar latent representations for profiles of the same cell, and orthogonal representations for non-matching profiles. A post-competition method is scMoGNN [28], a graph neural network-based method that outperformed the competition winners. scMoGNN is based on a cell-to-feature bipartite graph across modalities and uses heterogeneous graph convolutions to construct a cosine similarity matrix between all the pairs of cells. In both Novel and scMoGNN, cell matching was achieved by a maximum weight bipartite graph that generates hard matchings, i.e. each cell is paired with the top-scoring match. Additionally, since the algorithm for the maximum-weight bipartite matching is computationally intensive, it does not scale well with an increasing number of cells. Indeed, in both Novel and scMoGNN, the similarity matrix is sparsified by discarding the weights that are smaller than the row- and column-wise 0.995 and 0.95 quantiles, respectively.

In this paper, we propose MatchCLOT, a novel solution for the modality matching problem. Our work is inspired by recent promising applications of optimal transport (OT) in various single-cell data analysis tasks, such as identification of temporal cellular dynamics [29–31], integration of spatial information with gene expression [32], improvement of single-cell similarity metrics [33, 34] and alignment of single-cell multiomics datasets [20, 21]. Drawing inspiration from contrastive learning approaches, we train two modality-specific encoders to project the single-cell multimodal measurements onto a unified latent embedding space, and afterwards employ a novel OT algorithm to perform soft-matching of the cells between the modalities. MatchCLOT additionally exploits prior knowledge of the batch label, resulting in a smaller search space for cell-level modality matching, and uses a transductive setup that mitigates the effects of distribution shifts in the test data. By benchmarking MatchCLOT on two multimodal datasets from the NeurIPS competition we show that it consistently outperforms the competing best method, scMoGNN [28], achieving a 26.1% higher overall matching probability score while preserving the underlying biological structure of the integrated multimodal data. Importantly, our OT matching algorithm offers significant gains in computational time and memory compared to the existing methods and eliminates the need to discard any measurements. This advantage is an important consideration

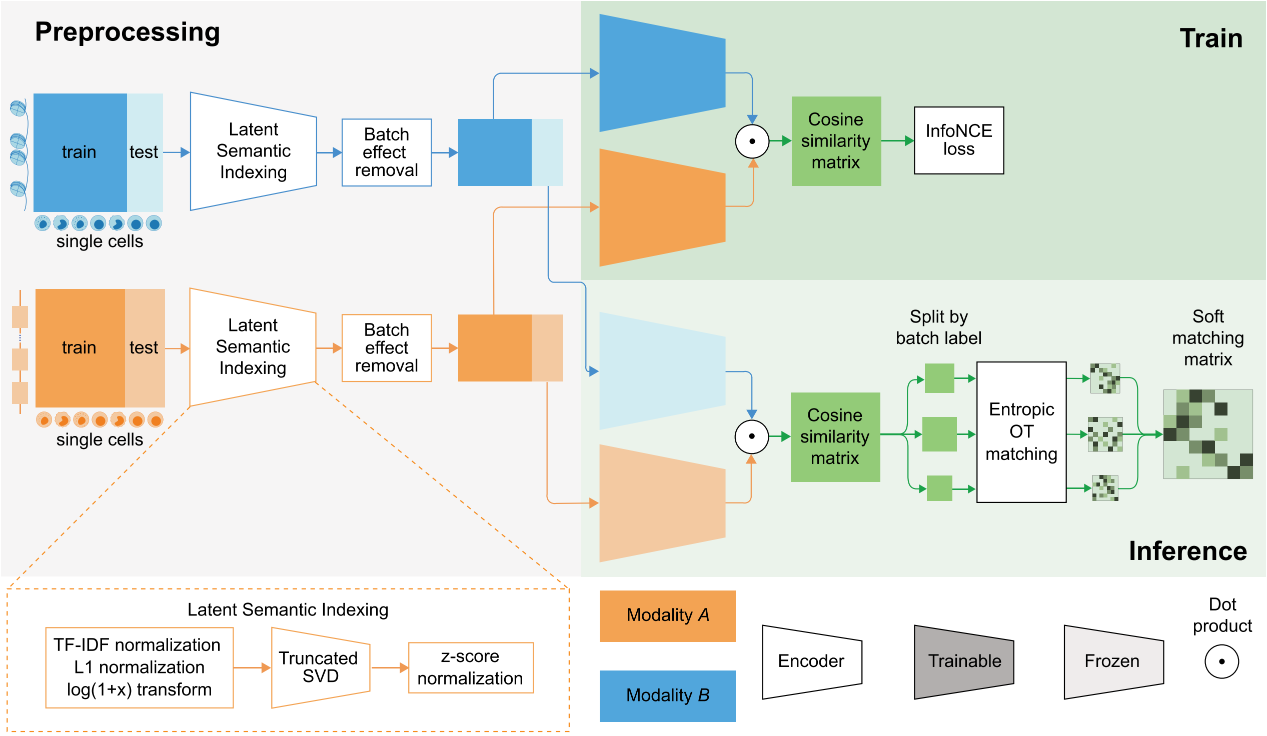

Figure 2. Overview of the MatchCLOT framework. The three primary blocks of the framework, i.e. preprocessing, train, and test, are highlighted in different colors, and the legends are presented at the bottom of the figure.

with the ever-increasing size of today's single-cell datasets. Lastly, by testing the trained model on an independent multimodal dataset, we show that MatchCLOT consistently outperforms existing methods, proving its robustness and ability to generalize to new, previously unseen single-cell multiomic data.

METHODS

MatchCLOT processes the cell measurements from A and B modalities to identify the cell correspondence. Let $\mathbf{x}_i^A$ and $\mathbf{x}_j^B$ denote the measurements of the i^{th} cell in A and j^{th} cell in B, respectively. The collective measurement for n cells is denoted by $\mathbf{A} \in \mathbb{R}^{n\times d_1}$ and $\mathbf{B} \in \mathbb{R}^{n\times d_2}$, where d_1 and d_2 are the corresponding measurement dimensions. MatchCLOT aims to predict a matching matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$, with $\mathbf{M}_{i,j}$ indicating the matching probability between the i^{th} cell in A and the j^{th} cell in B, by maximizing a matching probability score computed over $\mathbf{M}$. An overview of MatchCLOT is presented in Figure 2, which highlights its three primary blocks:

1. A *preprocessing block* that normalizes the data and projects it onto a low-dimensional latent space while correcting for the batch effect in a transductive setting.
2. A *training block* that employs a contrastive learning approach to maximize the similarity between matching cell profiles across modalities in the latent space.
3. An *inference block* that involves an entropic regularized OT and utilizes the batch labels for identifying the matching cell profiles.

Preprocessing block

The first block of MatchCLOT independently preprocesses the raw single-cell data across the modalities A and B in a transductive setting by operating on the union of the train and test sets that are available while training. For each modality, MatchCLOT normalizes and reduces the dimensions of the combined train and test set using latent semantic indexing (LSI), a common method for processing scATAC-seq data [35]. LSI consists of a term frequency-inverse document frequency (TF-IDF) normalization coupled to an L1-normalization and a logarithmic transformation, followed by a truncated SVD and a zero mean and unit variance scaling, as in Figure 2.

LSI results in preprocessed measurements of lower dimensions that are in turn corrected for batch effects using Harmony [13], an established batch effect correction method. Since the data distribution of the test batches is independent of the training batches, a correction of the batch effects in a transductive setting is imperative, which minimizes the impact of the distribution shifts due to acquisition variations in the data. We note that during the transductive preprocessing, the two modalities are processed independently and the ground-truth labels are not used by the model. Notably, leveraging the unlabeled test data was a common practice during the NeurIPS integration competition, where several methods applied unsupervised approaches to the test data prior to the training of respective computational methods. The final preprocessed cell measurements are denoted as $\mathbf{x}^A_{i_p}$ and $\mathbf{x}^B_{j_p}$ for the cells $i \in A$ and $j \in B$, respectively.

Training block

The training block of MatchCLOT is motivated by the contrastive learning idea of CLIP [27]. The main idea is to transform the preprocessed single cell measurements $\mathbf{x}^A_{i_p}$, $\forall i \in A$ and $\mathbf{x}^B_{j_p}$, $\forall j \in B$ into a unified latent embedding space $\mathbf{E}$, where the embeddings of the same cell are pulled closer together while pushing away the embeddings of the other cells. Formally, the embeddings of the i^{th} cell are denoted as $\mathbf{e}_i^A$ and $\mathbf{e}_i^B$ across the modalities A and B. The aim is to bring together $\mathbf{e}_i^A$ and $\mathbf{e}_i^B$, while pushing away from $\mathbf{e}_j^A$ and $\mathbf{e}_j^B$ $\forall j \in A$ and B, where $j \neq i$. The training model consists of two modality-specific encoders f^A and f^B, as in Figure 2, to produce the cell-level embeddings. Both the encoders are shallow multi-layer perceptrons (MLPs) with Exponential Linear Unit (ELU) activation function and dropout layers. The resulting modality-specific embeddings are unit normalized and multiplied via a dot product to produce a cosine similarity matrix **S**. For a pair of embeddings $\mathbf{e}_i^A$ and $\mathbf{e}_j^B$, $\mathbf{S}_{ij}$ is defined as:

$$\mathbf{S}_{ij} = \frac{\mathbf{e}_i^A}{\|\mathbf{e}_i^A\|} \cdot \frac{\mathbf{e}_j^B}{\|\mathbf{e}_j^B\|}$$

where · denotes the dot product operator.

An InfoNCE [36] objective is then computed for all the embedding similarity pairs in **S**. InfoNCE is a popular contrastive objective due to its simplicity, effectiveness, and theoretical guarantees. It leverages the idea of noise contrastive estimation (NCE) [37], a probabilistic model aiming to efficiently discriminate a target datapoint from noise by utilizing the corresponding context information, *i.e.*, the distribution of the neighboring datapoints. Given an embedding $\mathbf{e}_i^A$ and its context set $C = \{\mathbf{e}_i^B\} \cup \{\mathbf{e}_j^B \mid j \neq i\}$, where $\mathbf{e}_i^B$ is a positive embedding and $\{\mathbf{e}_j^B \mid j \neq i\}$ denotes a set of $k-1$ negative embeddings, the InfoNCE loss is computed as:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}\left[\log \frac{s(\mathbf{e}_i^B, \mathbf{e}_i^A)}{\sum\limits_{\mathbf{e}_j^B \in C} s(\mathbf{e}_j^B, \mathbf{e}_i^A)}\right]$$

where s denotes the scoring function defined as $s(\mathbf{e}_i^A, \mathbf{e}_j^B) = \exp\left(\mathbf{S}_{ij}/\tau\right)$ and τ is a temperature scaling parameter optimized during training.

The InfoNCE objective matches the embedding of the i$^{\text{th}}$ cell in modality A to the true matching profile of the same cell in modality B, while pulling away from the negative profiles of other cells in modality B. During training, the contrastive objective maximizes the cosine similarity between profiles corresponding to the same cell and minimizes the similarity between non-matching profiles via backpropagation. In particular, we jointly optimize two InfoNCE objectives during training, i.e. $\mathcal{L}_{\text{InfoNCE}}^{A\to B}$ and $\mathcal{L}_{\text{InfoNCE}}^{B\to A}$, to match the cell profiles in A to B and vice versa. The InfoNCE objective is particularly suitable for the modality matching task as it encourages the model to learn a representation where the two modalities are aligned at the single-cell level.

During training, we optimize the model and the training hyperparameters, namely the LSI reduced dimensions, the dimension of the encoder hidden layers, dropout rate, learning rate, and weight decay using Wandb [38], a Bayesian hyperparameter optimization library. To eliminate the risk of information leakage from the test set, the model is trained from scratch and the hyperparameters are tuned on a validation split obtained from the labeled training set.

Inference block

During the inference on the test set, MatchCLOT involves an entropic regularized OT to expedite and improve the matching performance compared to the maximum weight bipartite hard matching approach employed by existing methodologies. We further utilize the batch labels of the cells to optimize the search-space for the cell matching. Our OT matching module is method agnostic and can be applied to any kind of similarity matrix.

OT matching

OT is a field of mathematics that studies the optimal way of transporting a source distribution to a target distribution while minimizing the costs of displacement. Given two discrete probability distributions with supports A, B of the same size ($|A| = |B| = n$), with densities $\boldsymbol{\alpha}, \boldsymbol{\beta}$, costs $c(i,j)$ and probabilistic transportation plan $\boldsymbol{\Gamma}(i,j)$ defined $\forall i \in A$, $\forall j \in B$, the linear program formulation of the OT is given as:

$$\min_{\Gamma} \sum_{(i,j)\in A\times B} c(i,j)\,\Gamma(i,j), \quad \text{subject to:}$$
$$\sum_{j\in B} \Gamma(i,j) = \alpha(i) \quad \forall i \in A,$$
$$\sum_{i\in A} \Gamma(i,j) = \beta(j) \quad \forall j \in B,$$
$$\Gamma(i,j) \geq 0 \quad \forall i \in A,\ \forall j \in B$$

We utilize the linear program formulation of the OT problem to relax the integer linear program (ILP) formulation of the max-weight bipartite matching and convert it to an OT problem. Given a bipartite graph $G = (V, E)$ with bipartition (A, B), weight function $w : E \mapsto \mathbb{R}$, a matching $M \subseteq E$, let $m(i,j) = 1$, if $(i,j) \in M$ and 0 otherwise. Then, the ILP formulation of the maximum weight bipartite perfect matching is given as:

$$\max_{m} \sum_{(i,j)\in A\times B} w(i,j)\,m(i,j) \quad \text{subject to:}$$
$$\sum_{j\in B} m(i,j) = 1 \quad \forall i \in A,$$
$$\sum_{i\in A} m(i,j) = 1 \quad \forall j \in B,$$
$$m(i,j) \in \{0,1\}, \quad \forall i \in A,\ \forall j \in B$$

By dropping the integrality constraints on the variables $m(i,j)$, the problem becomes a linear program and can be converted to an OT problem with negative weights $-w(i,j)$ as costs $c(i,j)$, a function of variables $\frac{1}{n}m(i,j)$ as transport plan $\Gamma(i,j)$, and uniform distributions over A, B with densities $\boldsymbol{\alpha}(i) = \boldsymbol{\beta}(j) = \frac{1}{n}$ $\forall i \in A$, $\forall j \in B$. With the OT formulation, the transport plan Γ can be interpreted as a soft matching, where each vertex $i \in A$ can be matched with multiple vertices $j \in B$. Adding an entropic regularization [39] can speed up the computation of OT and lead to the following final objective:

$$\min_{\Gamma} \sum_{(i,j)\in A\times B} \underbrace{c(i,j)\ \ \Gamma(i,j)}_{\text{transportation cost}} + \underbrace{\varepsilon\ \Gamma(i,j)\ \ \log\Gamma(i,j)}_{\text{entropic regularization}}$$

The term ε controls the strength of the entropic regularization, with higher values producing noisier transport plans Γ. In our case, we use $\varepsilon = 0.01$ to generate a soft matching from the cosine similarity matrix **S**. Compared to a hard matching, where every profile is matched with only one profile, a soft matching has the advantage of providing additional information by predicting multiple weighted correspondences for a given profile.

Matching with batch label information

Finally, to avoid matching cell profiles across different batches, MatchCLOT exploits the test data batch labels to reduce the search space for the matching algorithm. This is achieved by splitting the profiles by batch labels, computing the cosine similarity matrices and entropic OT matching per batch, and combining the matching matrices for the final prediction.

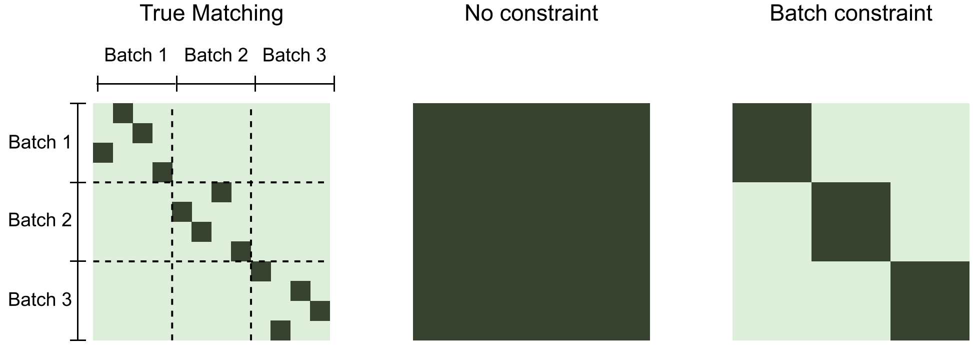

Figure 3. Search space of MATCHCLOT's matching step, without (middle) or with (right) batch constraint.

Table 1. Best hyperparameter configurations for the modality matching model found with Bayesian optimization. *The batch size was set without using Bayesian optimization.

Hyperparameter	**Search space**	**GEX ↔ ATAC**	**GEX ↔ ADT**
LSI dim mod1	$\{64, 96, 128, 192, 256, 384, 512\}$	192	192
LSI dim mod2	$\{64, 96, 128, 192, 256, 384, 512\}$	256	134
Encoder hidden dim mod1	$\{128, 256, 512, 1024, 2048, 4096\}^2$	(2048, 1024)	(256, 2048)
Encoder hidden dim mod2	$\{128, 256, 512, 1024, 2048, 4096\}^2$	(2048)	(4096, 2048)
Embedding dim	$\{128, 256, 512, 1024\}$	128	256
Dropout rates mod1	$[0.0, 0.7]^2$	(0.34, 0.47)	(0.3, 0.05)
Dropout rates mod2	$[0.0, 0.7]^2$	(0.67)	(0.4, 0.2)
Initial temperature $\log \tau$	$[1.0, 5.0]$	2.74	4.0
Learning rate	$\left[10^{-6}, 10^{-3}\right]$	$6 \cdot 10^{-4}$	$1.75 \cdot 10^{-4}$
Weight decay	$\left[10^{-6}, 10^{-3}\right]$	$1.25 \cdot 10^{-4}$	$2 \cdot 10^{-4}$
Batch size k*	–	16 384	16 384

Implementation

We implemented MATCHCLOT using PyTorch [40] and conducted the experiments on NVIDIA Tesla P100 GPU and POWER9 CPU. The comprehensive list of hyperparameter configurations is presented in Table 1. The search space of each hyperparameter was set to a reasonably large range based on the baseline model architecture and current best practices in hyperparameter tuning.

RESULTS

To evaluate the performance of our method, we tested MATCHCLOT on the CITE-seq and 10X Multiome data from the modality matching task of the NeurIPS competition that included a total of 90 000 and 70 000 cells, respectively [25]. Different combinations of the omic profiles in the dataset gave rise to a total of four subtasks, namely ATAC → GEX and GEX → ATAC for CITE-seq, GEX → ADT and ADT → GEX for Multiome. The labeled CITE-seq and Multiome data were split into two sets according to the batch labels for method validation (1 batch) and training (8–9 batches). The test data consisted of 15 066 cells for the CITE-seq subtasks GEX → ADT and ADT → GEX, and 20 009 cells for the Multiome subtasks ATAC → GEX and GEX → ATAC.

MATCHCLOT outperforms existing modality matching methods

To benchmark MATCHCLOT against other methods, we used the *matching probability score* of the NeurIPS competition that measures the modality matching performance of a method in terms of the weight/probability assigned to the correct cell pairings. Given a probability matching matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$ for n cells, the *matching probability score* is computed as:

$$\frac{1}{n} \sum_{\text{row}=1}^{n} \sum_{\text{col}=1}^{n} \mathbf{M}_{\text{row, col}} \cdot \mathbb{1}\left\{\text{row} = \text{true-match(col)}\right\}$$

To evaluate the soft matching predictions, we calculated the *Fraction of Samples Closer Than the True Match* (FOSCTTM) score [22]. Given a predicted matching matrix **M**, the FOSCTTM score evaluates how many confidence scores in **M** are higher than the score given to the true match, with lower values indicating better matching. The FOSCTTM is computed as:

$$\frac{1}{2n^2} \left(\sum_{\text{row}=1}^{n} \sum_{\text{col}=1}^{n} \mathbb{1}\left\{\mathbf{M}_{\text{row, col}} > \mathbf{M}_{\text{row, true-match(row)}}\right\} + \sum_{\text{row}=1}^{n} \sum_{\text{col}=1}^{n} \mathbb{1}\left\{\mathbf{M}_{\text{row, col}} > \mathbf{M}_{\text{true-match(col), col}}\right\} \right)$$

The matching probability scores of MATCHCLOT and the competing methods for the modality matching subtasks GEX → ATAC and GEX → ADT, as well as across all subtasks (Overall), are presented in Figure 4. MATCHCLOT achieved state-of-the-art scores for all subtasks and improved over scMoGNN by 26.1% for the overall matching score (Figure 4A). Since both CLUE and MATCHCLOT generate soft matchings, we used the FOSCTTM score to compare their performance, and found that, in all subtasks, MATCHCLOT outperformed CLUE, improving by –62% for the overall FOSCTTM score (Figure 4B). This significant drop in FOSCTTM score demonstrates the superiority of the matching matrix **M** because for each cell of modality *A* that we are trying to match, less than 1% of all other cells in modality *B* are assigned

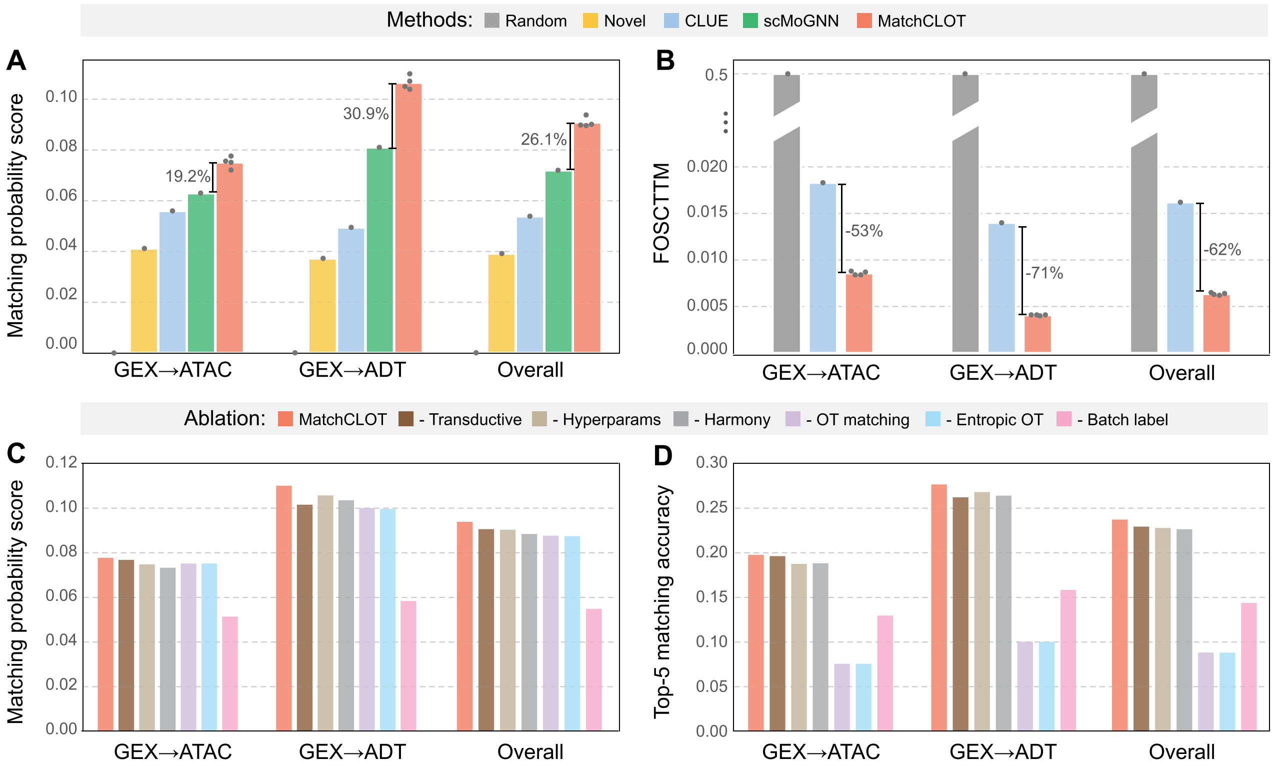

Figure 4. Performance assessment of MatchCLOT. Benchmarking MatchCLOT against existing modality matching methods using the matching probability score (**A**) and FOSCTTM score (**B**). For MatchCLOT, the height of the barplot is set to the mean of four different random seed initializations of the model, represented as dots. Ablation study of MatchCLOT using the matching probability score (**C**) and top-5 matching accuracy score (**D**).

a higher matching probability than their true match. We note that, since Novel and scMoGNN result in a hard matching, the FOSCTTM score is not defined in their case. We also note that MatchCLOT is independent of the sequence of the modality matching task, i.e. GEX → ATAC = ATAC → GEX, and ADT → GEX = GEX → ADT, similar to Team Novel and scMoGNN. To ensure the robustness of MatchCLOT, we repeated the training for three additional random initializations; the inference results (dots in Figure 4A and B) indicate that MatchCLOT's performance is fairly robust and that MatchCLOT consistently outperforms existing methods.

To evaluate to which extent individual components of MatchCLOT contribute to its state-of-the-art performance, we ran a thorough ablation study, where we progressively removed its different components and recomputed the matching probability score for all subtasks. As observed in Figure 4C, transductive preprocessing, hyperparameter tuning, batch effect correction and OT matching had small yet noticeable contributions to the final performance. Removing the batch constraint rendered the largest loss of performance, with the matching probability score dropping from 0.094 to 0.055. We repeated the ablation study using the *top-K matching accuracy* that quantifies if the true match is in the top-K matching probability scores across the rows/columns of **M**. The top-K matching accuracy score is defined as:

$$\frac{1}{2n}\left(\sum_{\text{row}=1}^{n} \mathbb{1}\left\{\bigcup_{k=1}^{K} \text{top-}k(\text{row}) = \text{true-match(row)}\right\} + \sum_{\text{col}=1}^{n} \mathbb{1}\left\{\bigcup_{k=1}^{K} \text{top-}k(\text{col}) = \text{true-match(col)}\right\}\right)$$

As observed in Figure 4D, the top-5 matching accuracy is higher than the matching probability score for all subtasks, indicating that even in cases where the true match is missed, it is still ranked within the top 5 scoring cells. We observe the same patterns as before for transductive preprocessing, hyperparameter tuning, batch effect correction and batch label constraint. However, we now observe that OT matching and entropic OT have a much more noticeable contribution to the top-5 matching accuracy, suggesting that OT matching is an integral component of MatchCLOT that contributes to producing accurate matching probability matrices that capture cell-cell similarities across all pairs of cells in both modalities.

MatchCLOT dramatically improves computation time and memory needs

As previously discussed, a limitation of the maximum weight bipartite graph matching used by Team Novel and scMoGNN is its scalability. Indeed, to achieve feasible computation times, both teams sparsified the matching matrix **M** by discarding a very high percentage of the edges (99.5% for Team Novel and 95% for scMoGNN). To evaluate the speedup in computation time and gains in memory usage of MatchCLOT, we repeated the inference step of the GEX → ADT subtask for which $\mathbf{M} \in \mathbb{R}^{15066\times15066}$, and computed computation time and max memory usage for a max-weight bipartite matching and OT matching (Figure 5). We observe that while the max-weight bipartite matching needs to discard 99% of the edges to render a computation time and memory usage of ∼125 s and 12 GB, respectively, our OT solution uses 100% of the edges within 50 s and 16 GB. For equal memory constraint, the max-weight bipartite matching needed to discard 95% of the edges. We conclude that apart from contributing to increased performance in terms of soft-matching, the OT matching component dramatically improves the scalability of the matching step. This is important not only because discarding values in **M** might

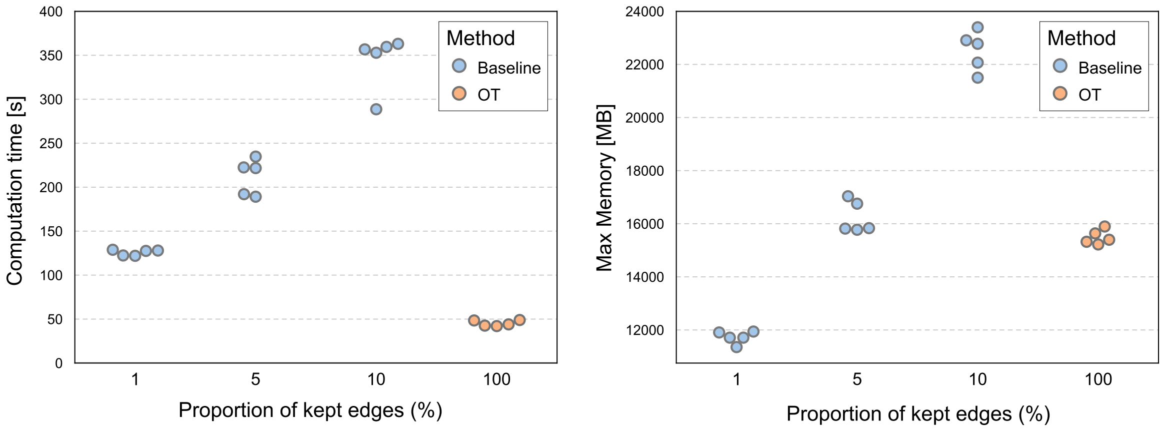

Figure 5. Comparison of computation time (left) and memory usage (right) between the baseline method using max-weight bipartite matching (baseline - blue) and OT matching (MATCHCLOT - orange) for different proportions of retained matching edges.

inhibit the performance by ignoring valuable information, but more importantly because the speedup offered by the OT matching allows MATCHCLOT to scale to the latest single-cell datasets that report up to millions of single cells at once. Moreover, our OT matching algorithm is independent of the previous components, therefore it can be applied to any representation of cell profiles for which a similarity function is defined across the two modalities.

MATCHCLOT achieves high cell type match scores

As already mentioned, matching cells of the same cell type is particularly hard due to their indistinguishability. To further understand if the mismatched cells are affected by this, we computed the cell type score, i.e. the percentage of cells matched with cells of the same cell type. We observe that MATCHCLOT achieves a very high cell type score of ≈0.88, scoring second just after CLUE (Figure 6A). Visualizing the score per cell type (Figure 6B), we observe that for 15 out of the 20 cell types the score is higher than 0.84, reaching and exceeding a score of 0.9 for 9 cell types. The cell types where the score had noticeably lower values are activated and naive CD4$^+$ cells and naive CD8$^+$ cells. To further understand between which cell types the mismatches occur, we computed a confusion matrix of all pairwise cell types (Supplementary Figure S1). We observe that several mismatchings are between closely related cell types, whose profiles are possibly very close in the epigenomic and transcriptomic space. For example, activated CD4$^+$ cells were often mismatched with naive CD4$^+$ cells, and vice-versa, and naive CD8$^+$ cells were also confused as activated or naive CD4$^+$ cells. Comparing MATCHCLOT to team Novel's model, we observe noticeable gains in the cell type scores for these frequently mismatched cell types (Figure 6B).

MATCHCLOT preserves the underlying structure of integrated single-cell datasets

We next asked how well MATCHCLOT preserves the biological patterns found in the multi-omic single-cell datasets. To evaluate the outcome of the MATCHCLOT embedding and matching, we performed dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) [41] of the 10X Multiome test data before any preprocessing (Figure 7A) and after LSI preprocessing, batch effect correction and encoder embedding (Figure 7B). As expected, although the UMAP projection of the raw data does capture some cell-type-specific clusters, after MATCHCLOT preprocessing and encoder embedding the UMAP projection is able to disentangle these coarse clusters into detailed trajectories of cell-type evolution (for example, notice the Megakaryocyte and Erythrocyte Progenitor → Proerythroblast → Erythroblast → Normoblast trajectory indicated by a black arrow in Figure 7B). Notice that activated CD4$^+$ cells, naive CD4$^+$ cells and naive CD8$^+$ cells, indicated in Figure 7 by orange, green and purple, respectively, are highly mixed in the UMAP space, in accordance with the previous observation that these cell types are often mismatched because of the similarity of their omic profiles.

Interestingly, the UMAP projection of the data using the predicted MATCHCLOT matching (Figure 7C) is highly similar to that of the real matching (Figure 7B), with all major clusters and trajectories maintained. This observation indicates that single-cell multi-omic integration using MATCHCLOT fully preserves the true underlying biological structure and topology of the data, and suggests that even for the cases where MATCHCLOT misses the real match, the predicted match is close enough to the omic profile of the real match that there are no noticeable effects that could disturb observed biological patterns. To further validate this result and ensure that the UMAP projection was driven by the integrated multi-omic profile of each cell and not dominated by a single modality, we performed a UMAP projection of randomly matched single cells (Supplementary Figure S2), where we clearly observe that the previous structure is distorted and the cell-type-specific clusters and trajectories are now split and dispersed in different locations of the UMAP space.

MATCHCLOT efficiently generalizes to unseen datasets

To assess the generalization ability of MATCHCLOT, we tested the model trained on the competition 10X Multiome dataset on an independent, non-competition 10X Multiome dataset in a zero-shot fashion. While the competition dataset contains paired ATAC and GEX measurements from bone marrow mononuclear (BMMC) cells, the new dataset contains paired ATAC and GEX measurements from peripheral blood mononuclear cells (PBMCs) [42]. This makes it of particular interest to test the generalization ability of the model: although PBMC and BMMC samples share many cell types, we expect notable differences in both the presence and frequency of cell types between the two. We first preprocessed the 10X Multiome PBMC dataset as described in the Supplementary Methods and then fed it to MATCHCLOT that matched the cells from the two modalities. We compared our results with several single-cell data integration methods based on a variety of approaches, such as PCA (Harmony [13]), CCA

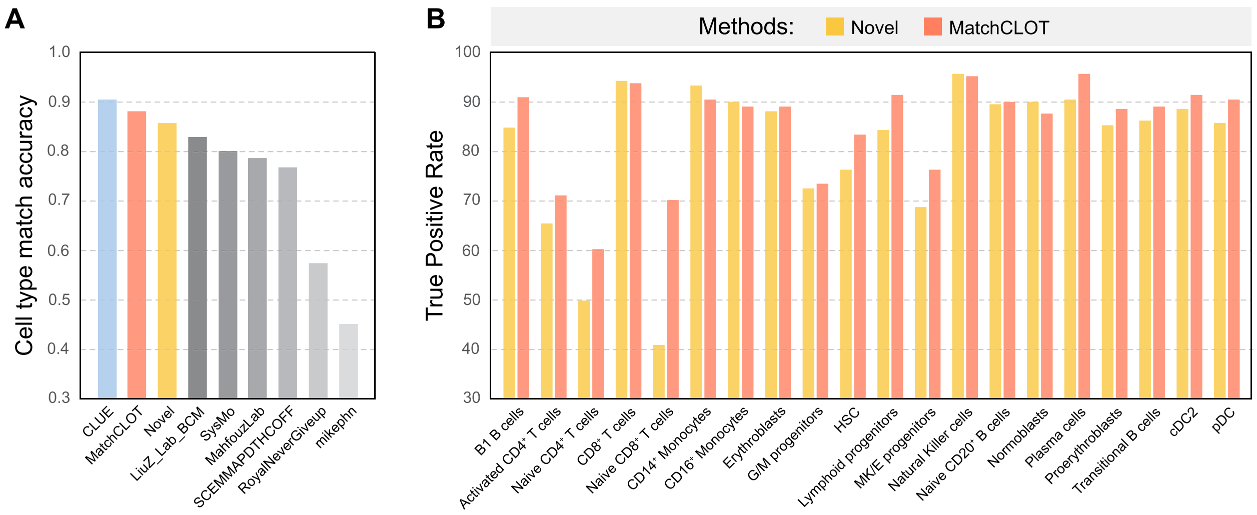

Figure 6. Matching score at the cell type level. (**A**) Cell type match score for the ATAC → GEX subtasks by top scoring teams. (**B**) Comparison of Team Novel and MatchCLOT scores for each cell type independently. Cell type abbreviations: HSC: Hematopoietic stem cell, MK/E: Megakaryocyte and Erythrocyte, cDC2: classical dendritic cells type 2, pDC: plasmacytoid dendritic cells, G/M: Granulocyte–Macrophage.

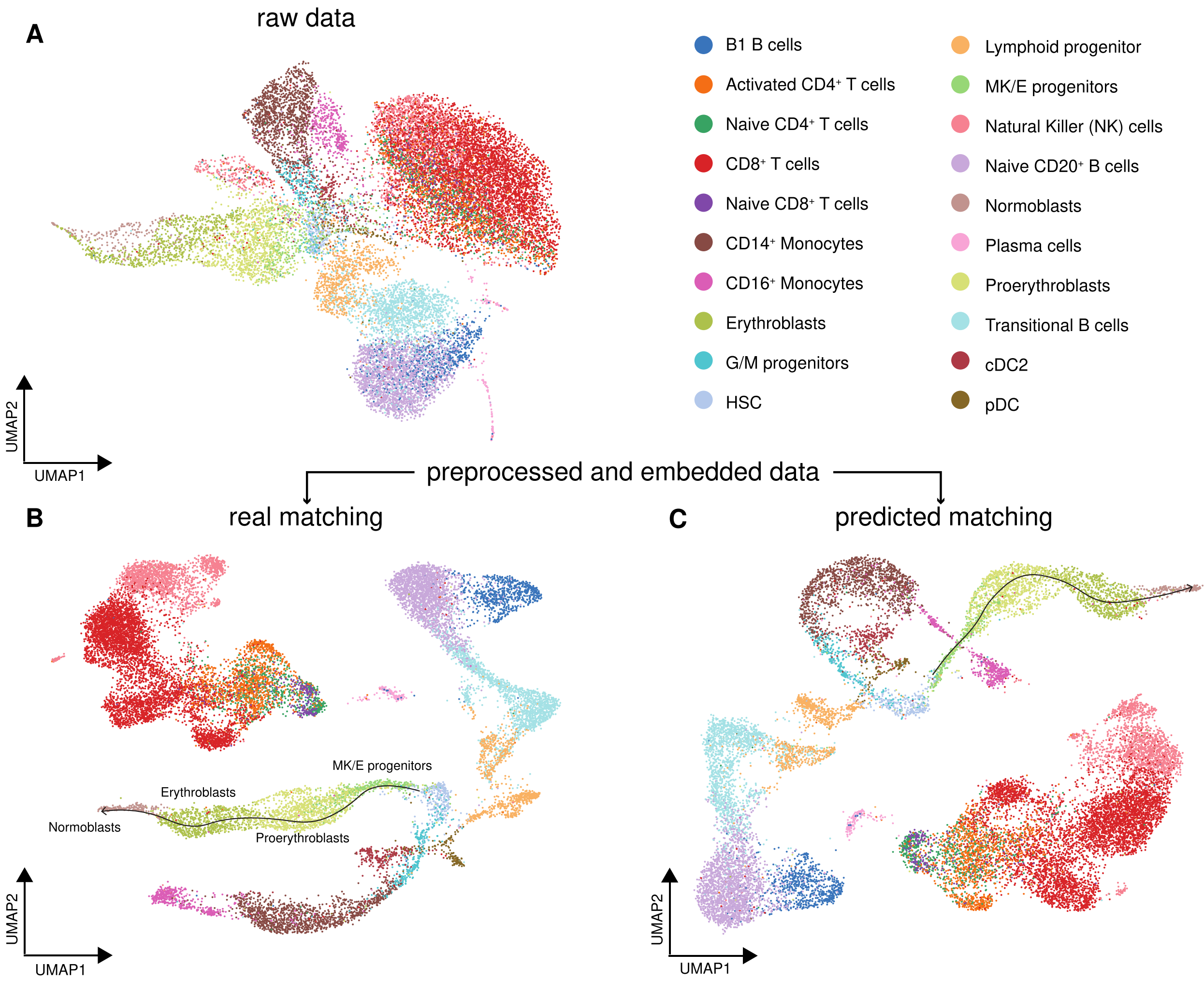

Figure 7. Dimensionality reduction of the 10X Multiome (RNA & ATAC) test data. UMAP embeddings of the raw (**A**) and preprocessed and embedded data using the real (**B**) and predicted (**C**) matching. Cell-type abbreviations as in Figure 6.

(Seurat v3 [14] and bind-SC [15]), NMF (LIGER [16] and Online iNMF [17]), graph neural networks (GLUE [23]), OT (Pamona [19]) and optimization (UnionCom [18], MMD-MA [22]). As observed in Figure 8A, MatchCLOT consistently outperforms all methods, with a mean FOSCTTM score of only 0.015, an improvement of 44.4% over the top performing method GLUE [23] (mean FOSCTTM equal to 0.0342). This is a very encouraging result considering that, unlike GLUE, MatchCLOT has not been trained on the PBMC

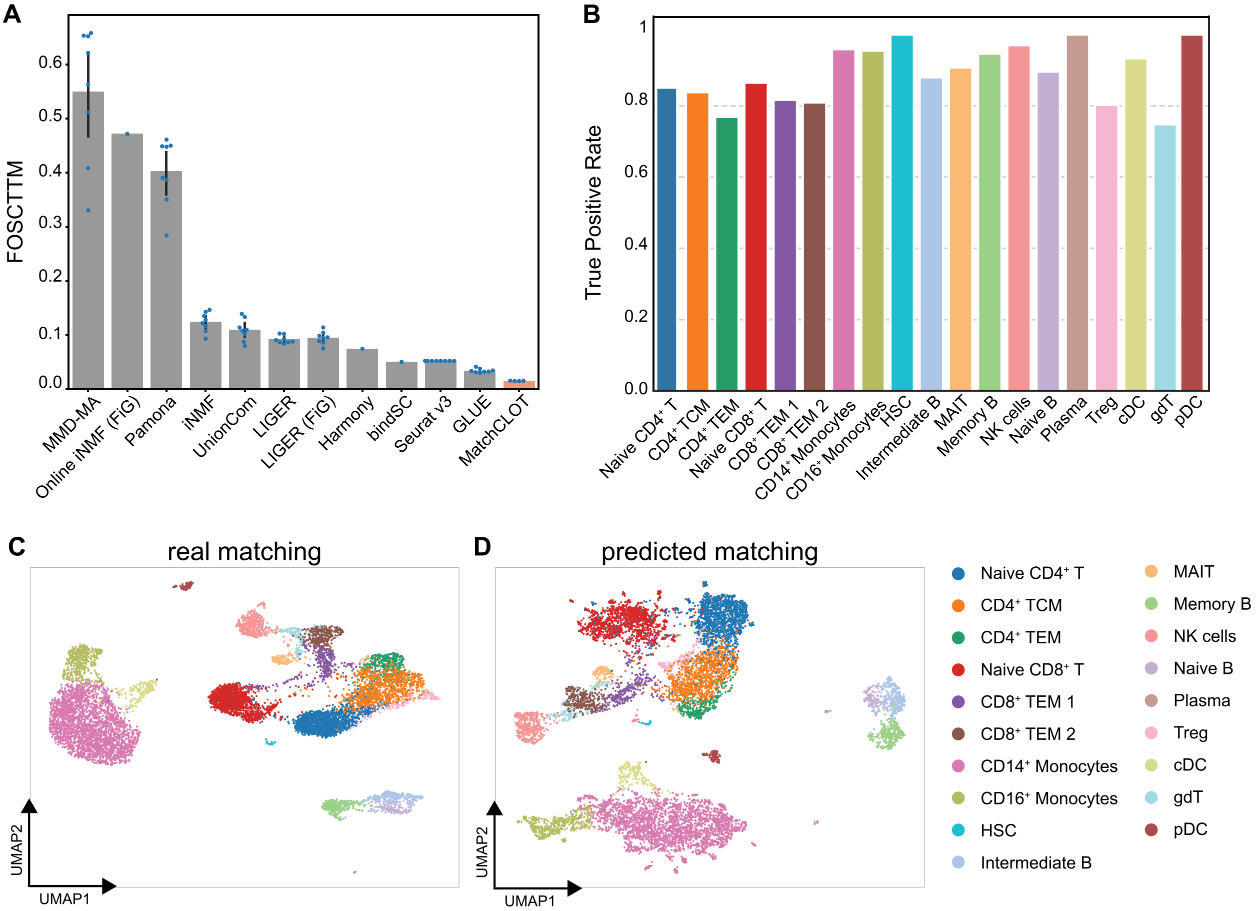

Figure 8. Testing MatchCLOT on an unseen 10X Multiome test dataset of PBMC cells from a healthy donor. (**A**) Benchmarking MatchCLOT against several existing algorithms using the FOSCTTM score. Bar heights correspond to the mean across independent runs of the models with random initializations shown as dots, and errorbars indicate 0.95 confidence intervals. For fair comparison, we used the same preprocessed data and scores of the previous methods from [23] (**B**) Matching score per cell type for all PBMC cell types. (**C** and **D**) UMAP embeddings of the preprocessed and embedded data using the real (C) and predicted (D) matching. Cell-type abbreviations: TCM/TEM: central memory/effector memory T cells, MAIT: mucosal-associated invariant T cells, Treg: Regulatory T cells, gdT: Gamma Delta T cells; all other cell type abbreviations as in Figure 6.

dataset. Assessing the match score per cell type (Figure 8B and Supplementary Figure S4), we observe that, for 17 out of the 19 PBMC cell types, MatchCLOT achieves a matching score higher that 0.8. As expected, many of the cell types with relatively lower matching scores correspond to novel cell types that did not exist in the BMMC dataset used to train MatchCLOT (e.g. gdT cells, CD4$^+$ TCM and TEM).

To evaluate how well MatchCLOT preserves the underlying structure of the new PBMC dataset, we again performed UMAPs of the preprocessed and embedded multimodal data using the real and predicted matching. We observe that the UMAP projection of the single-cell data as matched by MatchCLOT (Figure 8D) is highly similar to that of the real matching (Figure 8C). In both cases, clusters reflecting different PBMC cell types are visible and highly preserved between the two UMAPs, indicating that MatchCLOT does not distort the underlying topology. Interestingly, the MatchCLOT UMAP projection appears more noisy, with several small clusters appearing in the periphery of larger clusters, such as CD14$^+$ Monocytes (pink), Naive CD8$^+$ T cells (red) and Naive CD4$^+$ T cells (blue). A potential explanation of this effect is that the soft matching that comes as a result of the entropic regularization allows some weakly matched cells to simultaneously match to the same cell, creating small and tight sub-clusters. To reduce the noise level in the UMAP, one could set the entropic regularization to zero at the cost of a slightly worse matching score (Supplementary Figure S5). We also observe that several of the novel cell types with low matching scores appear to be mixed the UMAP embedding (e.g. gdT cells (light cyan) with CD8$^+$ TEM 2 cells (brown) and CD4$^+$ TEM (green) with CD4$^+$ TCM (orange)). As this observation is true also in the UMAP of the real matching, it suggests that the mismatchings can also be attributed to a high similarity between these cell types in the multi-omic feature space.

CONCLUSION

In this work, we proposed a novel computational framework that addresses the problem of matching cells across multiomic single-cell data. Our method MatchCLOT employs a contrastive learning setup that maximizes the similarity between matching cell profiles across modalities in the latent space, and a novel entropic regularized OT matching algorithm that replaces the common step of maximum weight bipartite graph matching to identify the matching cell profiles. MatchCLOT achieves state-of-the-art performance on two multiomic benchmarking datasets, achieving high overall and cell-type-level scores while preserving the underlying topology of the integrated data. Importantly, MatchCLOT is significantly more efficient in terms of computational time and memory usage, a critical advantage at a time where single-cell datasets routinely profile millions of cells at once.

Moreover, MatchCLOT achieves state-of-the-art zero-shot performance when evaluated on an unseen multiomic dataset containing cells of different origins than the original training dataset, clearly outperforming even previous methods that were trained on the unseen dataset. To our knowledge, this makes MatchCLOT the first example of CLIP-based zero-shot generalization in the domain of single-cell multiomic integration. Importantly, the ability of MatchCLOT to generalize well to unseen data of different biological origins suggests that the model learns biologically relevant relations governing the behavior of the measured biomolecules that are common across different biological contexts. This suggests that creating a collection of MatchCLOT models pretrained on curated benchmarking multiomic datasets can serve as a basis to enable modality matching across different organs, species or conditions. This ability to generalize is strengthened by the fact that, in contrast to several existing methods, at inference time MatchCLOT does not require any prior knowledge or cell-type annotation, making it highly applicable to any single-cell multiomic data without the need for manual labeling.

Despite its many advantages, further improvements could address some of MatchCLOT's current limitations. A key challenge is to define reasonable data augmentations of single-cell omics, crucial for contrastive learning [43]. Future works in this direction can address this challenge by exploring the recent advances in contrastive learning techniques. Another limitation stems from the fact that MatchCLOT's architecture is designed to enable the integration of only two single-cell omic modalities. However, emerging data acquisition methods that profile three or more omics at once [44] present an opportunity to train models able to match more than two modalities. This can be achieved by future adaptations of MatchCLOT that employ more sophisticated architectures and contrastive learning losses, similar to recent applications integrating language, audio and visual in multimodal sentiment analysis [45]. At the same time, more powerful encoding models based on transformers [46] could be employed together with larger datasets to potentially improve the results. Finally, approaches that are not only able to match cells but also identify interpretable omics features that drive the generated matchings are a very promising direction for future research [21]. Still, our vision is that MatchCLOT will be widely adopted to match the ever-growing single-cell dataset and open new avenues for integrated single-cell analysis.

Key Points

- MatchCLOT is a computational framework that is able to match single-cells measured using different omic modalities.
- MatchCLOT outperforms existing modality matching methods in terms of matching score and computational efficiency.
- MatchCLOT is able to preserve the underlying topology and structure of the multiomic data.

ACKNOWLEDGEMENTS

We would like to thank Anne Marie Cromack for critical feedback on the manuscript, and Dr Susanne Unger and all members of the AI for Single-Cell Research team for constructive comments, discussions and suggestions.

FUNDING

This work was supported by the Swiss National Science Foundation (SNSF) Sinergia Grant (CRSII5_202297).

DATA AVAILABILITY

The BMMC CITE-seq and 10X Multiome datasets used in this study are publicly available from the Gene Expression Omnibus (GEO) repository under the accession number GSE194122. The PBMC 10X Multiome dataset is available through 10X Genomics (see reference [42]). MatchCLOT is implemented as a pip-installable Python package and is publicly available under an open-source license at https://github.com/AI4SCR/MatchCLOT. Detailed instructions on how to reproduce all training, inference and ablation study steps, as well as pretrained models can also be found in the above link.

AUTHOR CONTRIBUTIONS

Conceptualization: F.G. and M.A.R; Methodology: F.G.; Software: F.G., P.P., P.C. and A.L.M., Formal analysis and Visualization: F.G., P.P., P.C. and M.A.R., Writing – Original draft: F.G., P.P. and M.A.R., Writing – Review and Editing: all authors; Supervision and Funding acquisition: M.K.d.J. and M.A.R.

References

1. Lygeros J, Koutroumpas K, Dimopoulos S, *et al*. Stochastic hybrid modeling of dna replication across a complete genome. *Proc Natl Acad Sci USA* 2008;**105**(34):12295–300.
2. Elowitz MB, Levine AJ, Siggia ED, *et al*. Stochastic gene expression in a single cell. *Science* 2002;**297**(5584):1183–6.
3. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. *Nature* 2010;**467**(7312):167–73.
4. Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. *Nat Rev Genet* 2019;**20**(9): 536–48.
5. Kashyap A, Rapsomaniki MA, Barros V, *et al*. Quantification of tumor heterogeneity: from data acquisition to metric generation. *Trends Biotechnol* 2022;**40**(6):647–76.
6. Stuart T, Satija R. Integrative single-cell analysis. *Nat Rev Genet* 2019;**20**(5):257–72.
7. Cao J, Cusanovich DA, Ramani V, *et al*. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* 2018;**361**(6409):1380–5.
8. Angermueller C, Clark SJ, Lee HJ, *et al*. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods* 2016;**13**(3):229–32.
9. Stoeckius M, Hafemeister C, Stephenson W, *et al*. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**(9):865–8.
10. Peterson VM, Zhang KX, Kumar N, *et al*. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* 2017;**35**(10):936–9.
11. Argelaguet R, Cuomo ASE, Stegle O, *et al*. Computational principles and challenges in single-cell data integration. *Nat Biotechnol* 2021;**39**(10):1202–15.
12. Efremova M, Teichmann SA. Computational methods for single-cell omics across modalities. *Nat Methods* 2020;**17**(1):14–7.
13. Korsunsky I, Millard N, Fan J, *et al*. Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**(12):1289–96.

14. Stuart T, Butler A, Hoffman P, *et al*. Comprehensive integration of single-cell data. *Cell* 2019; **177**(7):1888–1902.e21.
15. Dou J, Liang S, Mohanty V, *et al*. Bi-order multimodal integration of single-cell data. *Genome Biol* 2022;**23**(1):112.
16. Welch JD, Kozareva V, Ferreira A, *et al*. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 2019;**177**(7):1873–87.
17. Gao C, Liu J, Kriebel AR, *et al*. Iterative single-cell multi-omic integration using online learning. *Nat Biotechnol* 2021;**39**(8):1000–7.
18. Cao K, Bai X, Hong Y, Wan L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* 2020;**36**(Supplement 1):i48–56.
19. Cao K, Hong Y, Wan L. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics* 2021;**38**(1):211–9.
20. Demetci P, Santorella R, Sandstede B, *et al*. SCOT: single-cell multi-omics alignment with optimal transport. *J Comput Biol* 2022;**29**(1):3–18.
21. Demetci P, Tran QH, Redko I, *et al*. Jointly aligning cells and genomic features of single-cell multi-omics data with co-optimal transport. *bioRxiv* 2022. https://doi.org/10.1101/2022.11.09.515883.
22. Singh R, Demetci P, Bonora G, *et al*. Unsupervised manifold alignment for single-cell multi-omics data. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2020.
23. Cao Z-J, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat Biotechnol* 2022;**40**:1458–66.
24. Lance C, Luecken MD, Burkhardt DB, *et al*. Multimodal single cell data integration challenge: results and lessons learned. In: *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, PMLR 2022;**176**:162–76.
25. Luecken MD, Burkhardt DB, Cannoodt R, *et al*. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* 2021.
26. Xinming T, Cao Z-J, Xia C-R, *et al*. Cross-linked unified embedding for cross-modality representation learning. In: *Advances in Neural Information Processing Systems* 2022;**35**:15942–55.
27. Radford A, Kim JW, Hallacy C, *et al*. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning* 2021;**139**: 8748–63.
28. Wen H, Ding J, Jin W, *et al*. Graph neural networks for multimodal single-cell data integration. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 2022; 4153–63.
29. Schiebinger G, Shu J, Tabaka M, *et al*. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell* 2019;**176**(4):928–43.
30. Tong A, Huang J, Wolf G, *et al*. Trajectorynet: a dynamic optimal transport network for modeling cellular dynamics. In: *Proceedings of the 37th International Conference on Machine Learning*, PMLR 2000;**119**:9526–36.
31. Bunne C, Papaxanthos L, Krause A, *et al*. Proximal optimal transport modeling of population dynamics. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, PMLR 2022;**151**:6511–28.
32. Moriel N, Senel E, Friedman N, *et al*. Novosparc: flexible spatial reconstruction of single-cell gene expression with optimal transport. *Nat Protoc* 2021;**16**(9):4177–200.
33. Bellazzi R, Codegoni A, Gualandi S, *et al*. The gene mover's distance: single-cell similarity via optimal transport. arXiv:2102.01218 [q-bio.GN], 2021.
34. Huizing G-J, Peyré G, Cantini L. Optimal transport improves cell–cell similarity inference in single-cell omics data. *Bioinformatics* 2022;**38**(8):2169–77.
35. Cusanovich DA, Daza R, Adey A, *et al*. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 2015;**348**(6237):910–4.
36. Oord AVD, Li Y, Vinyals O. Representation learning with contrastive predictive coding. arXiv:1807.03748 [cs.LG], 2018.
37. Gutmann M, Hyvärinen A. Noise-contrastive estimation: a new estimation principle for unnormalized statistical models. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, PMLR 2010;**9**: 297–304.
38. Biewald L. Experiment tracking with weights and biases, 2020. Software available from. wandb.com.
39. Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. In: *Advances in Neural Information Processing Systems* 2013;**26**:2292–300.
40. Paszke A, Gross S, Massa F, *et al*. Pytorch: an imperative style, high-performance deep learning library. *In: Advances in Neural Information Processing Systems* 2019;**32**:8026–42.
41. McInnes L, Healy J, Melville J. Umap: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML], 2018.
42. 10X Genomics. *PBMC from a healthy donor, single cell multiome atac gene expression demonstration data by Cell Ranger ARC 1.0.0.* https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k, 2020.
43. Chen T, Kornblith S, Norouzi M, *et al*. A simple framework for contrastive learning of visual representations. In: *Proceedings of the 37th International Conference on Machine Learning*, PMLR 2020;**119**:1597–1607.
44. Ma A, McDermaid A, Jennifer X, *et al*. Integrative methods and practical challenges for single-cell multi-omics. *Trends Biotechnol* 2020;**38**(9):1007–22.
45. Mai S, Zeng Y, Zheng S, *et al*. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Trans Affect Comput* 2022.
46. Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. *In: Advances in Neural Information Processing Systems* 2017;**30**: 5998–6008.