Abstract

From the early stages of human development, we integrate information from multiple senses to learn and perform tasks. This type of intersensory redundancy enhances our learning capabilities. Similarly, multimodal machine learning seeks to fuse insights from diverse measurement devices or modalities to make accurate and reliable predictions. Over the past decade, many algorithms have been proposed for multimodal learning, including linear, kernel-based, and deep learning models. The recent advancements in multimodal deep learning, exemplified by models like ChatGPT, have enabled machines to "see, hear, and speak." However, multimodal biomedical data still pose significant challenges to these types of machine learning models.

In biomedicine, rapid technological progress enables researchers to collect large, high-throughput biological data across multiple modalities. Techniques such as scRNA-seq, ATAC-Seq, and CITE-seq measure high-resolution proteomic and genomic information at the single-cell level. Such datasets hold immense potential for analyzing intricate biological processes. However, they also present significant challenges to machine learning models due to the limited amount of labelled data, unpaired structure, inherent noise, and high dimensionality.

This research is dedicated to developing a comprehensive deep-learning framework for processing and analyzing multimodal biomedical data. The primary objective is to surmount core challenges associated with biomedical measurements by presenting solutions for the following multimodal learning tasks: (i) *alignment of unpaired measurements* to enable the identification of relationships and patterns across modalities and enhance predictive capabilities; (ii) *late fusion of unlabeled data* for reliable clustering to address the challenges of heterogeneity and enhance the ability to uncover biologically meaningful patterns within complex cellular systems; and (iii) *representation learning with partially overlapped observations* to extract meaningful latent biological information while attenuating modality-specific noise components.

Our framework will be implemented entirely using deep learning machinery, which consists of powerful function estimators that provide flexibility, scalability, and iterative training capabilities, and can be easily adapted to new domains. To mitigate the black-box nature of neural networks and enhance their intepretability, we will develop an unsupervised feature selection scheme to sparsify the input layer and highlight subsets of driving biological variables. Furthermore, we will accompany our algorithmic framework with theoretical guarantees that will serve as guidelines for effectively utilizing multimodal neural networks in the context of biomedical data. Our new algorithms aim to push the boundaries of biomedicine applications. These applications include cell classification, risk gene identification, and differential expression analysis. Enhancing the capabilities in these tasks holds the promise of creating more accurate models for automated diagnosis, prognosis, and drug discovery.

Part 3: Research plan

1 Scientific Background

Humans leverage complementary senses to acquire knowledge and interact with their surroundings. An illustrative example is the utilization of lip movements to aid in the discrimination of similar-sounding syllables [44]. Inspired by the advantages of integrating sensory information, researchers have developed multimodal learning techniques that leverage data acquired from diverse modalities. Each modality, denoted as $\mathcal{X}^l, l = 1, ..., L$, represents data obtained from a distinct measurement device, where $\mathcal{X}^l$ is defined as $\mathcal{X}^l = \boldsymbol{h}^l(\boldsymbol{\theta}, \boldsymbol{\psi}^l)$. Here, $\boldsymbol{h}^l$ may deform $\boldsymbol{\theta}$, the latent common variable of interest, and $\boldsymbol{\psi}^l$ encapsulates modality-specific information or measurement noise. By fusing complementary information from all measurement devices $\{\mathcal{X}^l\}_{l=1}^L$, multimodal learning can substantially enhance predictive accuracy and reliability across a wide range of applications [7, 12, 51]. For simplicity of exposition in the remainder of this section, we focus on the case of $L = 2$.

In recent years, multimodal machine learning has witnessed remarkable breakthroughs driven by deep neural network (DNN) architectures such as [16, 42]. These architectures have pushed the performance boundaries in image, text, audio analysis, and synthesis and may pave the road to artificial general intelligence (AGI) [14]. Unfortunately, existing schemes of multimodal vision–language learning are unsuitable for biomedical data. This is because many biomedical high-throughput measurements exhibit characteristics that render conventional approaches inapplicable [53]. Specifically, datasets like those seen in [37, 48] are unlabeled, unaligned, noisy, heterogeneous, imbalanced, high dimensional, or have low sample sizes. These challenges motivate the development of a comprehensive algorithmic framework capable of performing the core tasks in multimodal learning, namely, **representation learning**, **fusion**, and **alignment**. The primary goal of this proposal is to overcome these limitations by developing a coherent algorithmic framework for multimodal learning with biomedical data. In the following paragraphs, we provide a concise overview of the core tasks in multimodal learning and outline our primary goals and objectives.

Representation learning involves learning embedding functions $\boldsymbol{f}^1(\mathcal{X}^1)$ and $\boldsymbol{f}^2(\mathcal{X}^2)$, designed to extract meaningful structures of interest, for example, the latent common ($\boldsymbol{\theta}$) or modality-specific ($\boldsymbol{\psi}^1, \boldsymbol{\psi}^2$) components. This task is unsupervised but requires access to a bijective correspondence between the realizations. In the discrete setting, the matrices $\boldsymbol{X}^1 \in \mathbb{R}^{D^1 \times N}$ and $\boldsymbol{X}^2 \in \mathbb{R}^{D^2 \times N}$ each contain N (corresponding) samples with D^1 and D^2 features from $\mathcal{X}^1$ and $\mathcal{X}^2$ respectively. Canonical Correlation Analysis (CCA) [19], along with its nonlinear extensions such as Kernel CCA [2] or Deep CCA [1], tackle the problem by embedding datasets $\boldsymbol{X}^1$ and $\boldsymbol{X}^2$ into a new coordinate system in which the observations are maximally correlated. A similar intuition is used in Contrastive Language-Image Pre-training (CLIP) [42], which extracts remarkable image–text embeddings by training a model to classify image-caption correspondences. By contrast, multimodal biomedical data are often only partially overlapped or completely lack bijective correspondence. These properties render the majority of existing representation learning schemes inapplicable. This motivates our work to develop a representation learning scheme for biomedical data.

Fusion endeavors to integrate information from all measurement devices to enable accurate and reliable predictions of a target variable y (e.g., class label or regression value). Given paired observations (with bijective correspondence), represented as $\{\boldsymbol{x}_n^1\}_{n=1}^N$ and $\{\boldsymbol{x}_n^2\}_{n=1}^N$, the goal of modality fusion, denoted as $\boldsymbol{r}(\boldsymbol{x}_n^1, \boldsymbol{x}_n^2)$, is to improve predictive capabilities or lead to a smaller empirical risk

$$R(\boldsymbol{f}, \boldsymbol{r}) = \frac{1}{N} \sum_{n=1} \mathcal{L}(\boldsymbol{f} \circ \boldsymbol{r}(\boldsymbol{x}_n^1, \boldsymbol{x}_n^2), y_n).$$

Here, $\boldsymbol{f}$ is a prediction function, and $\mathcal{L}$ denotes the desired loss, such as cross-entropy or mean squared error. In the supervised setting, late fusion techniques such as [22, 34, 38, 49] integrate information at the prediction level. Since reliable labeled data is scarce in biomedical data, there is a growing need for unsupervised fusion schemes. Existing techniques, such as [8, 58], typically require domain-specific augmentations and are less suited for biomedical data.

Alignment seeks to identify a representation that aligns samples across modalities with the same semantic meaning. Unlike the previously discussed tasks, here, no prior knowledge of sample correspondence is assumed. In other words, $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$ are not necessarily measurements of the same value of $\boldsymbol{\theta}$, even when $i = j$. The multimodal alignment objective is to learn to mapping functions $\gamma^1()$ and $\gamma^2()$ such for each $\boldsymbol{x}_i^1, i = 1, ..., N$ we can find an index j such that $\gamma^1(\boldsymbol{x}_i^1) \sim \gamma^2(\boldsymbol{x}_j^2)$. This similarity signifies that the latent representations of $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_j^2$ correspond to the same (or nearly the same) latent value $\boldsymbol{\theta}$. The quality of this alignment can also assessed by applying a distance metric to $\gamma^1(\boldsymbol{x}_i^1)$ and $\gamma^2(\boldsymbol{x}_j^2)$. Existing multimodal alignment frameworks employ techniques such as cross attention [36] or contrastive learning [11, 24] to encode the data into a shared space. Multimodal alignment of biomedical data is challenging due to the heterogeneity of the data, noise level, and data dimensionality.

The goal of this research is to tackle the main challenges in multimodal learning with biomedical data by developing a coherent deep-learning methodology accompanied by theoretical guarantees, publicly available software, and verifications on real-world applications.

Below is a short summary of our aims.

(A1) Simultaneous Alignment and Representation Learning: To address the absence of bijective correspondence in biomedical data, we will develop a method to embed and permute observations simultaneously. This approach will yield aligned multimodal data representations, enhancing our ability to work with unpaired observations.

(A2) Self-supervised Multimodal Fusion: We aim to develop a late fusion scheme to enhance the accuracy and reliability of cluster assignments derived from multi-omics data. In the absence of labels, we will exploit self-supervision to fuse information and find a more comprehensive and nuanced understanding of complex biological systems.

(A3) Representation Learning with Partially Overlapped observations. We will derive a DNN-based manifold learning framework to obtain canonical representations from partially overlapped multimodal measurements. This will enhance our ability to extract meaningful information from complex data with partial overlap.

(A4) Interpretability by Identification of Driving Biological Variable. This objective en-

hances model interpretability by identifying subsets of informative features from high-dimensional multimodal data. To achieve this goal, we will develop a multilevel, unsupervised feature selection scheme that operates at the global, local, and group levels, enabling a more flexible approach to recovering driving biological factors.

(A5) Theoretical Properties for Multimodal Learning. To offer practical guidelines for practitioners, we will establish theoretical guarantees and limitations of our proposed multimodal learning framework. Specifically, we will analyze our proposed methodology's sample complexity, convergence guarantees, and different optimization aspects.

Methodology and Datasets The objectives above are interconnected and focused on advancing our ability to integrate multimodal biomedical data. Specifically, we will focus our empirical evaluations on single-cell multi-omics data, such as RNA sequencing (scRNA-seq), assay for transposase-accessible chromatin sequencing (scATAC-seq), and cellular indexing of transcriptomes and epitopes (scCITE-seq). These technologies measure biological properties at single-cell resolution and have been valuable in advancing our capabilities in several high-impact applications, including automated diagnosis, prognosis, and personalized treatment. The proposed solutions will be based on neural network machinery, which offers several benefits, including scalability, flexibility, transferability, and adaptability to new domains.

2 Research Objectives and Expected Significance

The overarching objective of this proposal is to formulate and implement a comprehensive deep-learning framework tailored for biomedical data, leveraging the power of deep multimodal learning. This framework will enhance data processing and analysis and enable more accurate and reliable predictions. Our research objectives have been crafted in response to the critical challenges posed by biomedical measurements, including the scarcity of labeled data, the absence of bijective correspondence, the presence of nuisance variables, and the disparity between the number of features and available samples.

Our research objectives, each of which has the potential to advance the field significantly, are presented below. The successful completion of $\sim$ 80% of these objectives would be considered a significant achievement, likely resulting in 4–6 publications.

2.1 Objective 1: Simultaneous Alignment and Representation Learning

As discussed in the scientific background, most multimodal representation learning schemes require paired datasets. Namely, there must be a bijective correspondence between samples in all modalities, e.g., sample $\boldsymbol{x}_i^1$ and $\boldsymbol{x}_i^2$ must correspond to the same observation. However, this assumption is not valid for most sequencing technologies, which cannot simultaneously profile a cell with independent modalities. The topic of multimodal representation learning for unpaired measurements is an understudied area, with only a limited number of studies such as [18] exploring this more general setting.

Under this objective, we will develop a method to simultaneously align multimodal datasets and learn representations capturing shared latent information ($\boldsymbol{\theta}$). For simplicity, we assume access to

N samples from each modality, namely $\boldsymbol{X}^1 \in \mathbb{R}^{D^1 \times N}$ and $\boldsymbol{X}^2 \in \mathbb{R}^{D^2 \times N}$. We will focus on scRNA-seq and scATAC-seq, which are typically unpaired; therefore, classic representation learning methods such as CCA or its extensions cannot be directly applied. Several recent methods offer solutions [10, 17, 60], but they are either linear or require some supervision. Instead, we propose an unsupervised approach that involves learning to embed the data into a shared space while simultaneously learning a permutation matrix $\boldsymbol{\Pi}$ that maximizes correlation in this shared space. The optimization problem can be formulated as follows:

$$\max_{\boldsymbol{\Pi} \in \mathcal{P}_N} \quad \text{corr}(\boldsymbol{f}_1(\boldsymbol{X^1\Pi; \theta^1}), \boldsymbol{f_2}(\boldsymbol{X^2}; \boldsymbol{\theta}^2)) = \frac{\boldsymbol{f}_1(\boldsymbol{X}^1\boldsymbol{\Pi}; \boldsymbol{\theta}^1)\boldsymbol{f}_2^T(\boldsymbol{X}^2; \boldsymbol{\theta}^2)}{\|\boldsymbol{f}_1(\boldsymbol{X}^1; \boldsymbol{\theta}^1)\|_2 \|\boldsymbol{f}_2^T(\boldsymbol{X}^2; \boldsymbol{\theta}^2)\|_2}, \tag{1}$$

where $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ represent neural networks with parameters $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$, respectively, and $\mathcal{P}_N$ is the set of all permutation matrices of size $N \times N$. Because of the discrete nature of $\boldsymbol{\Pi}$, traditional gradient-based optimization methods cannot be directly employed to maximize Eq. 1. To address this challenge, we propose a probabilistic relaxation for Eq. 1 (as outlined in Section 3.3) and demonstrate its applicability using synthetic data. Additionally, we will assess a more relaxed alignment objective, which involves aligning the data distributions in the latent spaces by leveraging techniques such as [5, 6, 9].

2.2 Objective 2: Late Fusion of Unlabeled Data

Obtaining reliable sample (cell) annotation of multi-omics data is an ongoing challenge. Practitioners often resort to manual cell annotation via dimensionality reduction and clustering, which induces many false annotations. These can later propagate and induce errors in downstream tasks, such as drug discovery or personalized treatment. Data fusion could mitigate these errors; however, most existing schemes rely on labeled data to perform fusion [15, 59]. Recently, there has been a growing interest in unsupervised fusion, but most solutions are dedicated to image data and require domain-specific data augmentations.

Here, we propose an innovative approach to perform representation learning and fusion without needing labeled data. Specifically, we treat the fusion problem as a self-supervised co-clustering task. We formulate an objective for learning the reduced representation via a deep CCA objective while simultaneously learning multi-modal cluster assignments using a trained prediction head. The predictions are fused using a self-supervised contrastive loss.

Our focus is on clustering multimodal data points, denoted as $\boldsymbol{X}^\ell, \ell = 1, ..., L$, where $\boldsymbol{X}^\ell = \{\boldsymbol{x}_i^\ell\}_{i=1}^N$, into matching clusters, denoted as $\boldsymbol{Y} = \{\boldsymbol{y}_i\}_{i=1}^N$. Here, $\mathbf{x}_i^\ell \in R^{D^\ell}$ represents D^ℓ-dimensional vector-valued observations of general type, i.e., tabular data that do not adhere to any specific feature structure. We aim to establish an end-to-end deep learning model that seamlessly combines embedding and clustering. We will learn encoders $\boldsymbol{h}^\ell(\boldsymbol{x}_i^\ell) = \boldsymbol{\psi}_i^\ell$ and clustering heads $\boldsymbol{f}^\ell(\boldsymbol{\psi}_i^\ell) = \hat{y}_i$, where $\hat{y}_i \in 1, 2, ..., K$, represent an accurate clustering assignment. Our conceptual innovation is to learn the parameters of $\boldsymbol{h}^\ell$ and $\boldsymbol{f}^\ell$ by employing a representation learning objective on $\boldsymbol{\psi}_i^\ell$, while leveraging self-supervised techniques for late fusion. This will allow us to reliably predict cluster assignments based on embedded information from all modalities.

2.3 Objective 3: Multimodal Representation Learning with Partial Overlap

In multi-omics data, each modality may have a good resolution of a different subset of the biological system. In such cases, integrating all modalities can yield a more comprehensive understanding of biological properties. To accomplish this, we aim to develop a method for integrating partially overlapping modalities while learning a representation that aligns with the geometry of the latent factors of interest. In this context, we make certain assumptions: (i) The latent domain of interest is a d-dimensional path-connected manifold $\mathcal{M}$. (ii) The data is obtained with K different measurement devices that capture specific regions of $\mathcal{M}$, denoted by $\mathcal{M}^1, \ldots, \mathcal{M}^K \subset \mathcal{M}$, and the union of these regions is path connected. (iii) Each measurement device is characterized by a smooth and injective function that maps the respective region $\mathcal{M}^i$ to its observation space. These functions are denoted as $\boldsymbol{f}^1, \ldots, \boldsymbol{f}^K$, and the observation spaces are $\mathcal{X}^1 \subset \mathbb{R}^{D_1}, \ldots, \mathcal{X}^K \subset \mathbb{R}^{D_K}$, with $D_1, \ldots, D_k \geqslant d$.

We present an illustration of the problem in Fig. 1. The brown area represents the latent manifold, which is observed through multiple measurement devices or "modalities." These devices capture the system's states using a perturbed sampling mechanism, where multiple observations are captured for each state, referred to as a "burst" (depicted as points within black circles). These bursts represent sets of samples within the neighborhood of the captured state in the latent space. This strategy was used in prior work on manifold learning [41, 46]. Our primary objective is to integrate information from all modalities, represented as the projected oval shapes, and discover a representation that faithfully represents the underlying latent manifold.

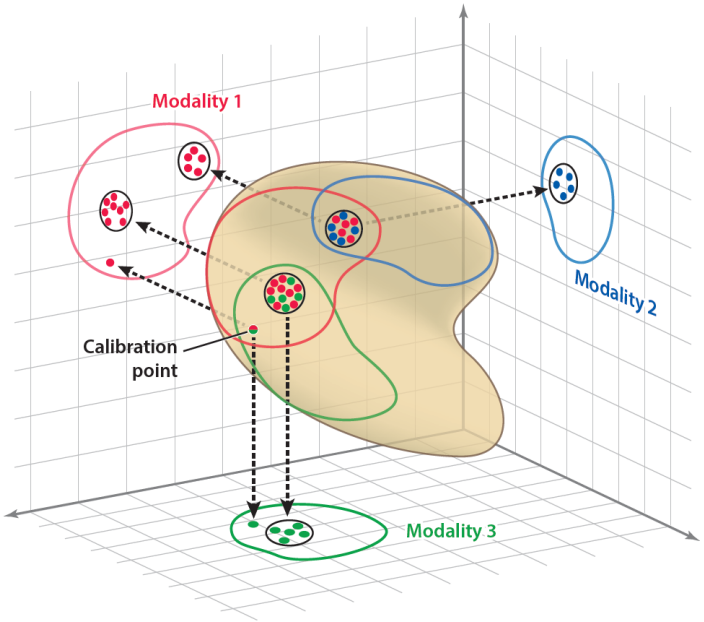

Figure 1: The latent representation of the data (center, three-dimensional) is observed by three different modalities/measurement devices (on the coordinate planes, two-dimensional). As depicted in the figure, each modality is capable of capturing only a specific subset of the latent domain and introduces its own unique deformation to the data. Local neighborhoods of points in the latent space are transformed into elliptical shapes when observed in the modalities. Within the intersection regions, some points are observed by more than one modality.

2.4 Objective 4: Global, Local, and Group Unsupervised Feature Selection

In high-throughput biological observations, many observed variables are nuisance variables and do not carry information about the phenomenon of interest. In such cases, it is vital to remove nuisance variables to prevent overfitting of commonly used multimodal learning schemes [1, 19]. Furthermore, identifying subsets of "driving variables" is important for enhancing model interpretability and analyzing biological effects. To address these challenges, several authors have proposed using unsupervised feature selection to attenuate the influence of nuisance features and improve model interpretability.

Under this objective, we aim to develop a deep learning framework for unsupervised feature selection (FS) in the context of multimodal observations. Our primary objective is to provide a feature selection mechanism that operates at three distinct levels of granularity:

1. **Global FS**: This represents the classic setting in which the selected features are shared across all samples, providing a global sparsification of the feature space.

2. **Local FS**: This level of granularity is designed to handle the inherent heterogeneity often observed in biomedical data. By enabling sample-specific feature selection, the FS model can learn the unique characteristics of different subsets in the population.

3. **Group FS**: In this approach, we aim to identify correlated feature groups and perform feature selection at the group level. This approach is particularly useful for identifying clusters of related variables and selecting the clusters of the most informative features.

By providing these three levels of granularity for the feature selection mechanism, we aim to enhance the flexibility and adaptability of the framework, making it well-suited for various scenarios and datasets in the realm of high-throughput biological observations. To address these challenges, we intend to extend our recently proposed stochastic gates (STGs) [55]. The STGs are continuously relaxed Bernoulli variables that have been demonstrated effective for nonlinear supervised [55, 56] and unsupervised feature selection [30]. Under this proposal, we aim to generalize the STGs to the multimodal setting and enable operation in the three levels of granularity described above.

2.5 Objective 5: Theoretical Foundation of Deep Multimodal Learning

In recent years, researchers have significantly advanced our understanding of deep learning, yielding several theoretical explanations for its success. These explanations encompass vital concepts such as the double descent phenomenon [3], neural collapse [40], and various optimization aspects associated with stochastic gradient descent (SGD) [52, 62]. However, most of these works primarily concentrate on supervised learning settings, with only a limited number of studies delving into the theoretical aspects of multimodal deep learning.

In the context of multimodal high-throughput biomedical observations, a common challenge arises from the fact that the number of variables often exceeds the number of actual measurements. In such a scenario, most conventional multimodal learning schemes face difficulties and may overfit. In this context, our goal is to gain a deeper understanding of the capabilities and limitations of deep multimodal learning when applied to high-dimensional biomedical data. We focus on sparse extensions

of the powerful DCCA model [1]. Specifically, we will address the following two fundamental questions regarding sample complexity and batch size:

(Q1) What is the sample complexity of sparse-DCCA?

The sample complexity is the number of training-samples needed by a deep-learning model to successfully learn a task. We start by presenting the sparse CCA objective under a linear data model assumption. Using modalities $\boldsymbol{X}^1 \in \mathbb{R}^{D^1 \times N}$ and $\boldsymbol{X}^2 \in \mathbb{R}^{D^2 \times N}$, which are centered and have N samples with D^1 and D^2 features, respectively, the goal of CCA is to find canonical vectors $\boldsymbol{a} \in \mathbb{R}^{D^1}$, and $\boldsymbol{b} \in \mathbb{R}^{D^2}$, such that , $\boldsymbol{u} = \boldsymbol{a^T X^1}$, and $\boldsymbol{v} = \boldsymbol{b^T X^2}$, will maximize the sample correlations between the *canonical variates*, i.e.

$$\max_{\boldsymbol{a}, \boldsymbol{b} \neq 0} \quad \text{corr}(\boldsymbol{a^T X}, \boldsymbol{b^T X^2}) = \frac{\boldsymbol{a^T X^1 (X^2)^T b}}{\|\boldsymbol{a^T X}\|_2 \|\boldsymbol{b^T Y}\|_2}. \tag{2}$$

To study the sample complexity of the solution we follow [50] using the data generated from the following distribution

$$\begin{pmatrix} \boldsymbol{X}^1 \\ \boldsymbol{X}^2 \end{pmatrix} \sim N(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_2 \end{pmatrix}), \text{ where } \boldsymbol{\Sigma}_{12} = \rho_0 \boldsymbol{\Sigma}_1 (\boldsymbol{\phi \eta}^T) \boldsymbol{\Sigma}_2.$$

Based on this data model, the canonical vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ maximizing the correlation objective in Eq. 2 are $\boldsymbol{\phi} \in \mathbb{R}^D$ and $\boldsymbol{\eta} \in \mathbb{R}^D$, respectively (see Proposition 1 in [50]).

In many biological datasets, only a small subset of variables capture the common latent variables. Therefore, we consider vectors $\boldsymbol{\phi}, \boldsymbol{\eta}$ that are sparse with only k nonzero elements. The indices of the active elements are chosen randomly with values equal to $1/\sqrt{n}$, and ρ_0 controls the total correlation between modalities. In this setting, we will study the consistency of the sparse ℓ_0-CCA estimator [29]. Namely, for a sparse estimate of the canonical vector $\hat{\boldsymbol{\phi}}$ (and similarly for $\hat{\boldsymbol{\eta}}$) we will study how N affects the probability $\mathbb{P}\left[\mathbb{E}\left[\|\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}\|_2^2\right] > \delta\right]$ for some $\delta > 0$ (and similarly for $\boldsymbol{\eta}$).

To answer this question, we will use similar techniques as in [47]. If successful, we will attempt to extend the sample complexity analysis to a more general setting of a nonlinear data model with a DCCA objective.

(Q2) Should small batches be used for multimodal learning? The choice of batch size in neural network training, specifically its effects on the training dynamics, is a crucial aspect. Our research will explore how small batch training, which relies on Stochastic Gradient Descent (SGD), influences multimodal deep learning. This fundamental question is rooted in the understanding that small batches impact the training dynamics and shape the stochastic gradient noise. Multiple studies have analyzed theoretical and empirical properties involved in small-batch training for supervised learning [31, 32]. Here, we intend to investigate how small-batch training can affect multimodal deep learning.

Addressing **(Q1)** and **(Q2)** will provide valuable guidelines for practitioners, offering insights into effectively employing DCCA models for multimodal learning in the challenging landscape of high-dimensional biomedical data.

2.6 Impact and Significance:

This research is driven by the emergence of many high-throughput technologies enabling the collection of multimodal information about complex biological systems. Examples of such multimodal measurements include SHARE-seq [37], DBiT-seq [35], and CITE-seq [48], which have provided biological insights and advancements in applications such as transcription factor characterization [21], cell type identification in the human hippocampus [54], and immune cell profiling [23]. These types of modalities, commonly formed as tables, still pose a significant challenge to standard multimodal techniques. This proposal is geared towards offering a complete deep-learning framework for multimodal biomedical data. We expect our contributions to impact the following aspects:

Algorithmic framework: The methodology developed under this research will serve as a reliable neural-network framework for analyzing multimodal biomedical data. This framework will offer several advantages over existing linear models or kernel methods. Neural networks are known for their flexibility, scalability to large datasets, iterative training capabilities, adaptability to new domains, and extensibility for incorporating additional modalities. One significant implication of this work is the potential to establish a foundation multimodal model for biomedical data. Foundation models have recently revolutionized various fields, including natural language processing (NLP) and computer vision. Applying similar principles to biomedical data can lead to groundbreaking advancements in the understanding and application of complex biological systems.

Theory: One hurdle in advancing deep learning stems from a lack of a complete theoretical understanding of frequently used modules. A crucial component of this research is the accompanying theoretical analysis. By delving into the theoretical underpinnings of multimodal deep learning, we aim to contribute to a better understanding of the critical modules commonly used in this field. This understanding can help break current barriers and provide valuable insights into the interplay between sample size, feature count, and model performance. The resulting theoretical guarantees will serve as guidelines for effectively utilizing multimodal neural networks in the context of biomedical data. Furthermore, such theoretical insights can enhance trust in neural network-based predictions, a critical quality in biomedicine.

Application: The impact of this proposal extends to the practical application of multimodal learning in the analysis of high-throughput biological data. Even partial success has the potential to revolutionize the way researchers approach the analysis of such data. The ability to reliably integrate diverse data types, including genomics, proteomics, and imaging, will enable a more comprehensive understanding of complex biological systems. In genomics, the framework can contribute to predicting risk genes, identifying regulatory elements, and uncovering gene-to-gene interactions, paving the way for significant advancements in genetics research. Applications in proteomics could include automated diagnosis [43], prognosis, and personalized treatment [4], which have substantial implications for improving human healthcare and personalized medicine [39].

Impact: advancing the state of the art in multimodal biomedical data analysis and providing powerful tools and insights that will benefit a wide range of scientific and medical applications.

3 Detailed Description of the Proposed Research

3.1 Working Hypothesis

Multimodal biomedical data integrates information from diverse sources, providing a complementary view of biological processes. Such measurements typically consist of nonlinear interconnections between the observed variables; therefore, linear models can fail to capture these complex interactions. Deep learning is a powerful machinery that is an exceptional non-linear function estimator. Our main working hypothesis is that **a multimodal deep-learning framework will enhance the analysis and interpretation of complex biomedical data by integrating information from multiple sources**, improving disease diagnosis, treatment planning, and patient outcomes. This hypothesis induces the goal of our research, which is to develop a complete DNN methodology for the representation, fusion, and alignment of multimodal biomedical observations. Our methods will be accompanied by a theoretical analysis and application to real-world use cases.

In the following subsections, we provide a mathematical description of our methodological strategy for solving each posed objective. Some of these subsections include empirical results supporting the presented solutions. We note that most of the results are based on synthetic or simplified settings; therefore, there is still much work to be done in the development, evaluation, and analysis of all methods.

3.2 Research Design and Methodologies

We now provide more technical details about our strategy for achieving our goals. Throughout the following section, we focus for simplicity on the coupled setting of two modalities ($L = 2$). We are given realizations (observations) from two modalities $\{\boldsymbol{x}_n^1\}_{n=1}^N$ and $\{\boldsymbol{x}_n^2\}_{n=1}^N$ either paired (with bijective correspondence) or unpaired.

3.3 Preliminary Results

4 Infrastructure and Human Resources

The research will be carried out at Bar Ilan University. Dr. Ofir Lindenbaum is a senior lecturer in the Faculty of Engineering. He has had very productive collaborations with biologists, physicians, applied mathematicians, data scientists, and engineers. Driven by real-world problems, his research primarily focuses on developing supervised and unsupervised machine learning methods for identifying meaningful parameters from raw empirical measurements. In the past decade, he extensively studied the problems of multimodal learning, sparse recovery, and feature selection. He is an expert in multimodal learning and has published several articles on the problem [25, 27, 28, 29, 33, 45, 57]. He has also worked extensively on the feature selection problem [30, 31, 32, 55]. Furthermore, he has an ongoing collaboration in several biomedical studies [13, 20, 26, 61].

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[2] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[3] M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[4] D. Bertsimas, A. Orfanoudaki, and R. B. Weiner. Personalized treatment for coronary artery disease patients: a machine learning approach. *Health care management science*, 23:482–506, 2020.

[5] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[6] Y. Cao, L. Fu, J. Wu, Q. Peng, Q. Nie, J. Zhang, and X. Xie. Integrated analysis of multimodal single-cell data with structural similarity. *Nucleic acids research*, 50(21):e121–e121, 2022.

[7] Y. Chang and Y. Bisk. Webqa: A multimodal multihop neurips challenge. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 232–245. PMLR, 2022.

[8] J. Chen, H. Mao, W. L. Woo, and X. Peng. Deep multiview clustering by contrasting cluster assignments. *arXiv preprint arXiv:2304.10769*, 2023.

[9] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

[10] J. Dou, S. Liang, V. Mohanty, Q. Miao, Y. Huang, Q. Liang, X. Cheng, S. Kim, J. Choi, Y. Li, et al. Bi-order multimodal integration of single-cell data. *Genome biology*, 23(1):1–25, 2022.

[11] J. Duan, L. Chen, S. Tran, J. Yang, Y. Xu, B. Zeng, and T. Chilimbi. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15651–15660, 2022.

[12] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik. Multimodal learning with graphs. *Nature Machine Intelligence*, 5(4):340–350, 2023.

[13] S. F. Farhadian, O. Lindenbaum, J. Zhao, et al. Hiv viral transcription and immune perturbations in the cns of people with hiv despite art. *JCI insight*, 7(13), 2022.

[14] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022.

[15] K. Gadzicki, R. Khamsehashari, and C. Zetzsche. Early vs late fusion in multimodal convolutional neural networks. In *2020 IEEE 23rd international conference on information fusion (FUSION)*, pages 1–6. IEEE, 2020.

[16] A. Guzhov, F. Raue, J. Hees, and A. Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[17] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[18] Y. Hoshen and L. Wolf. Unsupervised correlation analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3319–3328, 2018.

[19] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[20] L. Irshaid, J. Bleiberg, E. Weinberger, J. Garritano, R. M. Shallis, J. Patsenker, O. Lindenbaum, Y. Kluger, S. G. Katz, and M. L. Xu. Histopathologic and machine deep learning criteria to predict lymphoma transformation in bone marrow biopsies. *Archives of Pathology & Laboratory Medicine*, 146(2):182–193, 2022.

[21] J. Joung, S. Ma, T. Tay, K. R. Geiger-Schuller, P. C. Kirchgatterer, V. K. Verdine, B. Guo, M. A. Arias-Garcia, W. E. Allen, A. Singh, et al. A transcription factor atlas of directed differentiation. *Cell*, 186(1):209–229, 2023.

[22] S. Kumar, S. K. Gupta, V. Kumar, M. Kumar, M. K. Chaube, and N. S. Naik. Ensemble multimodal deep learning for early diagnosis and accurate classification of covid-19. *Computers and Electrical Engineering*, 103:108396, 2022.

[23] N. Leblay, R. Maity, E. Barakat, S. McCulloch, P. Duggan, V. Jimenez-Zepeda, N. J. Bahlis, and P. Neri. Cite-seq profiling of t cells in multiple myeloma patients undergoing bcma targeting car-t or bites immunotherapy. *Blood*, 136:11–12, 2020.

[24] Z. Lin, Z. Zhang, M. Wang, Y. Shi, X. Wu, and Y. Zheng. Multi-modal contrastive representation learning for entity alignment. *arXiv preprint arXiv:2209.00891*, 2022.

[25] O. Lindenbaum, Y. Bregman, N. Rabin, and A. Averbuch. Multiview kernels for low-dimensional modeling of seismic events. *IEEE Transactions on Geoscience and Remote Sensing*, 56(6):3300–3310, 2018.

[26] O. Lindenbaum, N. Nouri, Y. Kluger, and S. H. Kleinstein. Alignment free identification of clones in b cell receptor repertoires. *Nucleic acids research*, 49(4):e21–e21, 2021.

[27] O. Lindenbaum, N. Rabin, Y. Bregman, and A. Averbuch. Multi-channel fusion for seismic event detection and classification. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)*, pages 1–5. IEEE, 2016.

[28] O. Lindenbaum, N. Rabin, Y. Bregman, and A. Averbuch. Seismic event discrimination using deep cca. *IEEE Geoscience and Remote Sensing Letters*, 17(11):1856–1860, 2019.

[29] O. Lindenbaum, M. Salhov, A. Averbuch, and Y. Kluger. L0-sparse canonical correlation analysis. In *International Conference on Learning Representations*, 2021.

[30] O. Lindenbaum, U. Shaham, J. Svirsky, E. Peterfreund, and Y. Kluger. Differentiable unsupervised feature selection based on a gated laplacian. *Conference on Neural Information Processing Systems (NerIPS)*, 2021.

[31] O. Lindenbaum and S. Steinerberger. Randomly aggregated least squares for support recovery. *Signal Processing*, 180:107858, 2021.

[32] O. Lindenbaum and S. Steinerberger. Refined least squares for support recovery. *Signal Processing*, 195:108493, 2022.

[33] O. Lindenbaum, A. Yeredor, and M. Salhov. Learning coupled embedding using multiview diffusion maps. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 127–134. Springer, 2015.

[34] K. Liu, Y. Li, N. Xu, and P. Natarajan. Learn to combine modalities in multimodal deep learning. *arXiv preprint arXiv:1805.11730*, 2018.

[35] Y. Liu, M. Yang, Y. Deng, G. Su, A. Enninful, C. C. Guo, T. Tebaldi, D. Zhang, D. Kim, Z. Bai, et al. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. *Cell*, 183(6):1665–1681, 2020.

[36] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[37] S. Ma, B. Zhang, L. M. LaFave, A. S. Earl, Z. Chiang, Y. Hu, J. Ding, A. Brack, V. K. Kartha, T. Tay, et al. Chromatin potential identified by shared single-cell profiling of rna and chromatin. *Cell*, 183(4):1103–1116, 2020.

[38] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021.

[39] M. E. Ozer, P. O. Sarica, and K. Y. Arga. New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines. *Omics: a journal of integrative biology*, 24(5):241–246, 2020.

[40] V. Papyan, X. Han, and D. L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, 2020.

[41] E. Peterfreund, O. Lindenbaum, F. Dietrich, T. Bertalan, M. Gavish, I. G. Kevrekidis, and R. R. Coifman. Local conformal autoencoder for standardized data coordinates. *Proceedings of the National Academy of Sciences*, 117(49):30918–30927, 2020.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[43] U. Raghavendra, U. R. Acharya, and H. Adeli. Artificial intelligence techniques for automated diagnosis of neurological disorders. *European neurology*, 82(1-3):41–64, 2020.

[44] T. Raij, K. Uutela, and R. Hari. Audiovisual integration of letters in the human brain. *Neuron*, 28(2):617–625, 2000.

[45] M. Salhov, O. Lindenbaum, Y. Aizenbud, A. Silberschatz, Y. Shkolnisky, and A. Averbuch. Multi-view kernel consensus for data analysis. *Applied and Computational Harmonic Analysis*, 49(1):208–228, 2020.

[46] A. Singer and R. R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

[47] J. Soham, H. Li, Y. Yamada, and O. Lindenbaum. Support recovery with stochastic gates: Theory and application for linear models. *arXiv preprint arXiv:2110.15960*, 2021.

[48] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14(9):865–868, 2017.

[49] D. Sun, M. Wang, and A. Li. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. *IEEE/ACM transactions on computational biology and bioinformatics*, 16(3):841–850, 2018.

[50] X. Suo, V. Minden, B. Nelson, R. Tibshirani, and M. Saunders. Sparse canonical correlation analysis. *arXiv preprint arXiv:1705.10865*, 2017.

[51] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[52] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019.

[53] S. Wang, M. E. Celebi, Y.-D. Zhang, X. Yu, S. Lu, X. Yao, Q. Zhou, M.-G. Miguel, Y. Tian, J. M. Gorriz, et al. Advances in data preprocessing for biomedical data fusion: An overview of the methods, challenges, and prospects. *Information Fusion*, 76:376–421, 2021.

[54] Y. Xiao, G. Su, Y. Liu, C. A. Sissoko, Y.-y. Huang, A. N. Santiago, A. J. Dwork, G. B. Rosoklija, U. D. Mark, V. Arango, et al. Spatially resolved transcriptomes in human hippocampus. *Biological Psychiatry*, 91(9):S18, 2022.

[55] Y. Yamada, O. Lindenbaum, S. Negahban, and Y. Kluger. Feature selection using stochastic gates. In *International Conference on Machine Learning*, pages 10648–10659. PMLR, 2020.

[56] J. Yang, O. Lindenbaum, and Y. Kluger. Locally sparse neural networks for tabular biomedical data. In *International Conference on Machine Learning*, pages 25123–25153. PMLR, 2022.

[57] J. Yang, O. Lindenbaum, Y. Kluger, and A. Jaffe. Multi-modal differentiable unsupervised feature selection. *arXiv preprint arXiv:2303.09381*, 2023.

[58] X. Yang, C. Deng, Z. Dang, and D. Tao. Deep multiview collaborative clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[59] C. Zhang, Z. Yang, X. He, and L. Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[60] R. Zhang, L. Meng-Papaxanthos, J.-p. Vert, and W. S. Noble. Multimodal single-cell translation and alignment with semi-supervised learning. *Journal of Computational Biology*, 29(11):1198–1212, 2022.

[61] J. Zhao, A. Jaffe, H. Li, O. Lindenbaum, E. Sefik, R. Jackson, X. Cheng, R. Flavell, and Y. Kluger. Detection of differentially abundant cell subpopulations discriminates biological states in scrna-seq data. *bioRxiv*, page 711929, 2020.

[62] Y. Zhou, Y. Liang, and H. Zhang. Understanding generalization error of sgd in nonconvex optimization. *Machine Learning*, pages 1–31, 2022.