ChAda-ViT : Channel Adaptive Attention for Joint Representation Learning of Heterogeneous Microscopy Images

Nicolas Bourriez[1*] Ihab Bendidi[1,2*] Ethan Cohen[1,3*] Gabriel Watkinson[1,3]
Maxime Sanchez[1,3] Guillaume Bollot[3] Auguste Genovesio[1]

[1]Ecole Normale Supérieure PSL, Paris, France
[2]Minos Biosciences, Paris, France
[3]Synsight, Evry, France

firstname.lastname@ens.psl.eu

*Equal Contribution

Abstract

Unlike color photography images, which are consistently encoded into RGB channels, biological images encompass various modalities, where the type of microscopy and the meaning of each channel varies with each experiment. Importantly, the number of channels can range from one to a dozen and their correlation is often comparatively much lower than RGB, as each of them brings specific information content. This aspect is largely overlooked by methods designed out of the bioimage field, and current solutions mostly focus on intra-channel spatial attention, often ignoring the relationship between channels, yet crucial in most biological applications. Importantly, the variable channel type and count prevent the projection of several experiments to a unified representation for large scale pre-training. In this study, we propose ChAda-ViT, a novel Channel Adaptive Vision Transformer architecture employing an Inter-Channel Attention mechanism on images with an arbitrary number, order and type of channels. We also introduce IDR-Cell100k, a bioimage dataset with a rich set of 79 experiments covering 7 microscope modalities, with a multitude of channel types, and counts varying from 1 to 10 per experiment. Our architecture, trained in a self-supervised manner, outperforms existing approaches in several biologically relevant downstream tasks. Additionally, it can be used to bridge the gap for the first time between assays with different microscopes, channel numbers or types by embedding various image and experimental modalities into a unified biological image representation. The latter should facilitate interdisciplinary studies and pave the way for better adoption of deep learning in biological image-based analyses.

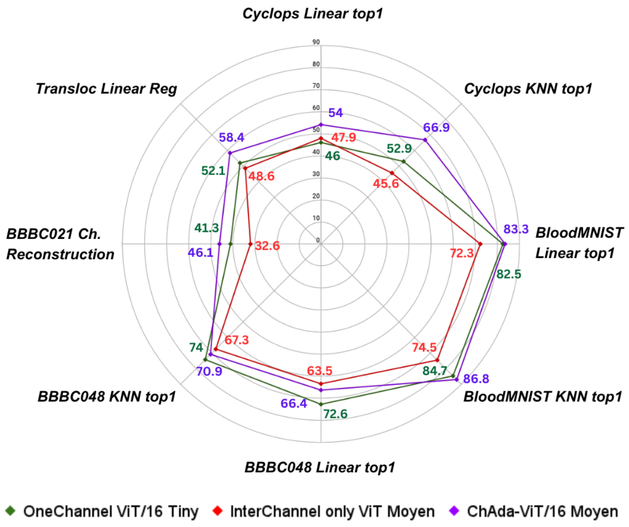

Figure 1. Performance comparison on downstream tasks, showcasing ChAda-ViT's superiority in 6 out of 8 tasks compared to existing approaches[47] using CLS token only. R^2 scores, normalized to 0-100, are presented for BBBC021 Channel Reconstruction and Nuclear Translocation prediction tasks. This success is attributed to the combined use of Intra-Channel and Inter-Channel Attention. Evaluation on all tokens is detailled in Appendix.

1. Introduction

Revolutionizing the field of image processing, Convolutional Neural Networks (CNNs) set the stage for unprecedented advancements [24, 30, 32]. Meanwhile, originally designed for Natural Language Processing (NLP), transformers emerged [43] later in computer vision as Vision Transformers (ViTs) [16]. ViTs excel in handling large datasets, detecting long-range dependencies [38] and capturing spatial correlations, often surpassing the capabilities

[‡]Code, Data & Model weights : https://github.com/nicoboou/chadavit

of their CNN counterparts [7, 15, 36, 42, 46, 50]. The adoption of ViTs has spurred important applications in many tasks[11, 14, 40, 49], as well as the creation of foundation models that have significantly improved performance across a variety of domains[18, 19], marking a new era of versatility and robustness in the application of these transformative technologies.

Bioimaging is however distinct, marked by its sparsity and lack of standardization, a contrast to the rapid advancements seen in general computer vision. Unlike conventional pictures in RGB color format, microscopy images span a plethora of specialized types of image across various channels [21, 27, 44]. Each channel, marked by a different staining or microscopy imaging technique, discloses unique biological information that, in many case, is specific to the very assay being imaged. This multimodality of images in term of channel numbers and types is pivotal as it underpins the heterogeneity inherent in bioimaging data. The capability to discern and leverage inter-channel relationships is paramount in this case as it necessarily lead to a richer capture of biological phenomena, thereby holding significant promise in advancing biological observation. [10].

However, due to the prominent use of RGB pictures in computer vision literature, transfer of current methodologies to biological images predominantly advocate for individual channel encoding [17, 29, 47]. A stance that, albeit practical for certain tasks, overlooks the potential insights harbored in the interplay between channels and the information they represent. Importantly, this approach thwarts the re-usability of pre-trained models across diverse studies[12, 13, 22, 34]. Models tailored to a specific microscopy configuration may yield compromised performance due to data scarcity, possibly driving spurious correlations over valuable biological features [39]. These approaches also overlook the opportunity to exploit the vast heterogeneous biological data available [23, 45]. A unified architecture accommodating the diverse nature of bioimaging data would not only facilitates the establishment of a common biological embedding space for various vision tasks but also heralds the potential of crafting a single pre-trained model. Such a model could serve as a linchpin for broader studies across different biological tasks, enabling comparative analyses, and studying correlations in a streamlined and unified analytical space. This consolidation could significantly accelerate and streamline analysis, fostering a quicker adoption of deep learning within the biological community for image-based studies.

Through attempting to resolve these issues, the main contributions of this paper are threefold:

- The introduction of a heterogeneous bioimage dataset encompassing various channel types and numbers, as well as a variety of microscopy imaging techniques used to acquire these channels.
- The introduction of a backbone architecture of Vision Transformers capable of handling bioimage datasets with different numbers and types of channels through a masking strategy coupled with intra and inter-channel attention, while achieving state-of-the-art results in a number of biologically relevant tasks compared to the usual ViT based approach for biological images.
- For the first time, to the best of our knowledge, we present a unified embedding space for any microscopy image dataset, bridging the gap between different heterogeneous datasets and opening the door to cross-modal imaging studies.

2. Related Works

Microscopy Image analysis. Advancements in image processing for biological applications have significantly contributed to high-throughput assays analyses and functional genomics. Open tools such as CellProfiler [35] have been integral, facilitating cell analysis with simple and efficient approaches, as well as modular image analysis pipelines for 3D image stacks and cloud-based processing to handle the surge in *biological big data*. The introduction of CNNs has further augmented bio-imaging, providing rapid and efficient solutions to tasks such as phase unwrapping, subtle phenotype analysis and multi-parametric cell classification and analysis. [1, 4, 26, 31]. Vision Transformers have leveraged their ability to effectively learn long-range dependencies in biological data, presenting new opportunities and addressing remaining challenges in bioimage analysis [33]. Moreover, recent studies [3] highlight the significant influence of transformation design on feature learning in microscopy images, an aspect that underscores the need for biology-specific considerations in self-supervised learning (SSL). Additionally, SSL methods employing vision transformers, such as DINO [8], have outshined traditional tools and achieved superior performance in numerous biological tasks, offering better classification of chemical perturbations and clustering gene families [29], and advancing morphological profiling, with enhanced capabilities in encoding complex cellular morphology without manual supervision [17]. These developments mark a shift toward more automated and sophisticated frameworks in bioimage analysis, crucial for navigating the increasing complexity and volume of biological data.

Unified Microscopy Image Representation. In computational biology, achieving a unified representation space suitable for diverse microscopy techniques remains an ongoing challenge. While cell painting methods [9] offer a data-rich environment for in-depth analysis within their specific domain [20, 37, 51], the scarcity of data in other experimental types creates a bottleneck for wider application. Approaches like Microsnoop [47] address this by operating in a one channel encoding regime, encoding each channel independently and subsequently concatenating these into a final representation. However, this approach leads to variable-sized embeddings depending on channel configurations, which impedes the integration of data for cross-experiment studies. CytoImageNet's strategy [28] to average channel information into a single-channel dataset resolves the representation space inconsistency but at the cost of losing detailed multi-channel information. To the best of our knowledge, no method offers a joint representation inclusive of all types of microscopy and channel configurations.

3. Dataset

We introduce the IDRCell100K image dataset, a collection of biological images, purposefully curated from the extensive and varied Image Data Resource platform [45]. This section outlines the process for selecting and refining these images and provide details on these biological assays, with the explicit goal of encompassing a heterogeneous distribution of data. Our selection, based on metadata provided with these experiments, covered various microscopy techniques to encapsulate a diverse array of imaging modalities, ensuring the dataset's breadth in representing biological information. Efforts were made to minimize experimental and imaging biases, striving for a balanced representation up to a feasible extent, thereby reducing dependency on each image modality or experiment. Further details on the equitable distribution of images across different microscopy modalities within the dataset are available in the Appendix.

Data Source Heterogeneity. To create a well-rounded dataset, we focused on cell culture experiments from the Image Data Resource. We picked 79 distinct experiments conducted under different conditions and for different scientific purposes. These experiments employed 7 types of microscopy techniques and fell into 6 categories of study.

Data Selection. As the number of images differ from one experiment to the other, we carefully chose 1,300 images from each selected experiment, in order to keep the final dataset balanced. These images come from experiments using different methods and include a wide range of channels monitoring for various components of the cells. Altogether, we end up with 308,898 single channel images, which we resized to 224x224 pixels from a variety of original sizes. When combined, it resulted to 104,093 multiplexed microscopy images containing cells at various scales, with each image made from one to up to 10 different channels.

Implementation details. Due to the lack of a dedicated Application Platform Interface (API), the retrieval of these images was performed through automated authorized webscraping of the Image Data Ressource platform. This process was performed on a distributed High-Performance-Computing CPUs cluster, using HTCondor cluster manager software [41] with 10 processes per node. In this settings, creation of the IDRCell100k Dataset took two weeks.

4. Methodology

4.1. Problem Formulation

Let I denote a single image from the dataset, extracted from a union of spaces $\bigcup \mathbb{R}^{H\times W\times n_i}$, where n_i represents the number of channels in image I_i, and $1 \leq n_i \leq N_{\text{max}}$, with N_{max} being the maximum number of channels across all images.

The objective is to find a projection function Φ that maps an image I_i to a latent space $\mathbb{R}^K$, formalized as:

$$\Phi : \bigcup \mathbb{R}^{H\times W\times n_i} \to \mathbb{R}^K \tag{1}$$

$$l_i = \Phi(I_i) \tag{2}$$

where K is the dimensionality of the latent space for the image I_i.

4.2. Gold Standard Approach

Existing works in literature [28, 47] such as Microsnoop adopt a One Channel Encoding approach, which results for Image I_i with n_i channels into a generalized function mapping Φ' defined as :

$$\Phi' : \bigcup \mathbb{R}^{H\times W\times n_i} \to \bigcup \mathbb{R}^{K'\times n_i} \tag{3}$$

where K' is the dimensionality of the latent space for a single channel.

This overarching function is effectively realized by processing each channel j of an image I_i independently. A function Ψ is trained to project these individual channels into a latent space $\mathbb{R}^{K'}$:

$$\Psi : \mathbb{R}^{H\times W} \to \mathbb{R}^{K'} \tag{4}$$

$$l_{ij} = \Psi(I_{ij}) \tag{5}$$

where l_{ij} represents the latent feature of the j-th channel of the i-th image.

For images with disparate channel counts n_i, the individual latent features l_{ij} are concatenated to form a representation l_i for each image:

$$l_i = \bigoplus_{j=1}^{n_i} l_{ij} \tag{6}$$

The dimensionality of this representation l_i aligns with the shape $K' \times n_i$. Therefore it depends on the original number of channels n_i in the image ($l_i \in \bigcup \mathbb{R}^{K'\times n_i}$) which makes it impossible to train a single such a model on a large heterogeneous bioimage dataset as the one we assembled. Also it does not provide a way to encode various dataset into a single representation.

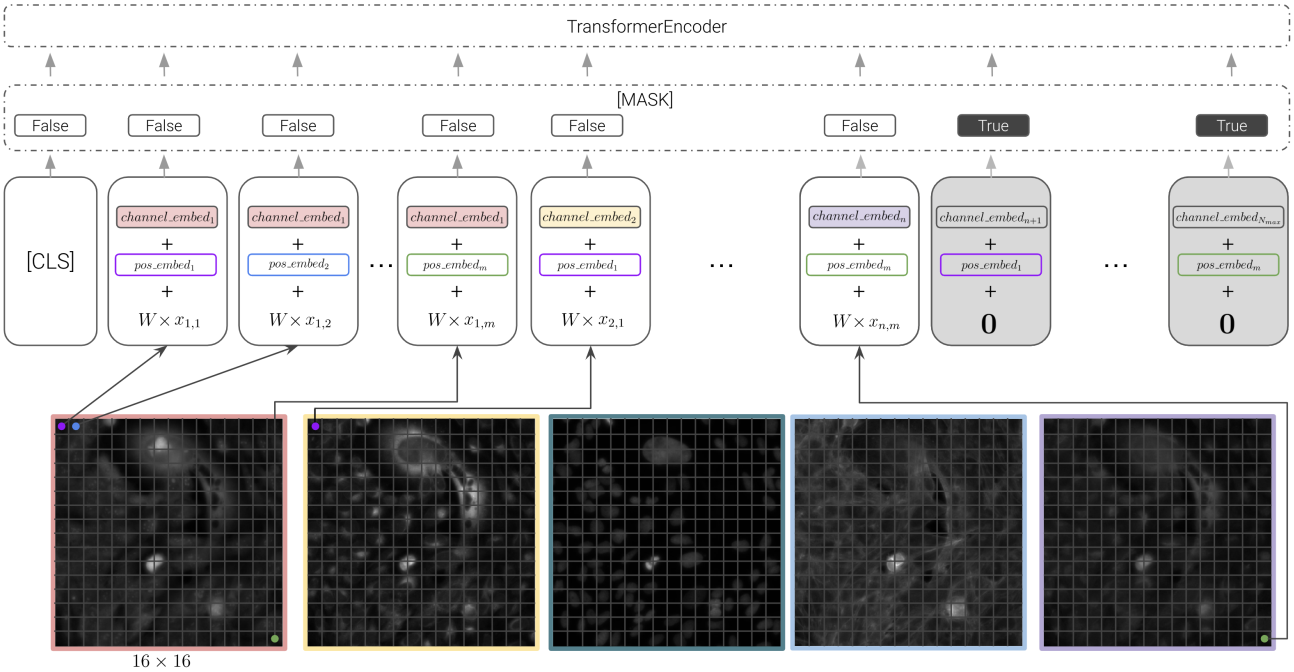

Figure 2. The ChAda-ViT model architecture, displaying the proposed channel-adaptive embedding process. This figure illustrates the token padding and masking approach for image data with an arbitrary number of channels, the split of channels into patches, and the integration of positional and channel-specific embeddings to reach a fixed size input. We use the same positional emedding for patches in the same position accross channels, and the same channel embedding for all patches of each channel.

4.3. Our Proposed Approach

We introduce ChAda-ViT, the **Ch**annel **Ada**ptive **Vi**sion **T**ransformer, a unified architecture for a model Φ as defined in Eq. 1, capable of encoding images with heterogeneous channel dimensions n_i into a single fixed size embedding space, through the introduction of Inter-Channel Attention, on top of the regular intra-channel or spatial attention. This is a consequence of leveraging the principles of token padding and masking – a technique originally established in NLP transformers [43] and partially adapted to Self Supervised learning in ViTs [2, 25] –, and introducing the concept of channel embeddings, as shown in Figure 2, to accommodate the variable number of channels present in different images.

The proposed approach patchifies each channel separately instead of considering them altogether. Each channel j of an image I with dimensions $H \times W \times n$ is split into non-overlapping patches $P_{j,x,y}$. Each patch at spatial location (x, y) and channel j is of size $p \times p$, where $p = 16$ is the preferred ViT configuration. These patches are then projected into a lower dimension with a shared 2D convolutional layer.

To standardize the input of the Transformer, we employ a padding strategy that compensates for images I with fewer channels than N_{max}, the maximum number of channel over the dataset. Padding tokens extend the patch sequence of each image to match the length of $N_{\text{max}} \times m$, where m is the number of patches per channel. Thus, the padded sequence, $\text{Seq}_{\text{pad}}(I)$, ensures any image is transformed into a fixed vector size to feed the the model. To maintain the integrity of the self-attention mechanism within the Transformer, we apply a binary masking strategy during attention computation. A mask is created for each image in the batch, marking the locations of padding tokens to ensure these are excluded from contributing to the self-attention mechanism. This method allows ChAda-ViT to focus solely on the meaningful data patches and preserve the inter-channel and intra-channel attention accuracy.

We also introduce the concept of channel embeddings, which focus on preserving channel information. Each patch $P_{j,x,y}$ is enriched with both *positional* and *channel-specific* embeddings to preserve its spatial context and channel identity. Positional embeddings $pos_{x,y}$ ensure spatial information is maintained across all channels, while channel embeddings $chan_j$ mark each patch with its respective channel origin. The dual embedding strategy allows the model to distinguish between patches of different channels located at the same spatial position. Both types of embeddings are learnable parameters, fine-tuned during the training process to optimize the representation of spatial and channel information within the unified embedding space.

Dataset	Downstream Task(s)	Granularity	Biological Application	Dataset size	Shape	Metric
BloodMNIST[48]	Clustering + Classification	Single-Cell	Cellular types	17092	28*28*3	Accuracy
CyclOPS[†]	Clustering + Classification	Single-Cell	Protein Localization labelling	28166	64*64*2	Accuracy
BBBC048[‡]	Clustering + Classification	Single-Cell	Cell Cycle Stages	32266	66*66*3	Accuracy
NF-kB Nuclear Transloc[31]	Regression	Single-cell	Nuclear Translocation	1000	256*256*3	R2
BBBC021[6]	Generation	Whole Slide	Imaging	13200	1024*1280*3	R2,MSE,MAE

Table 1. Overview of biological datasets used to compare the ChAda-ViT and One Channel ViT models across clustering, classification, regression, and generative tasks, highlighting the diversity in biological applications, dataset sizes, and complexity. Performance metrics include Top1 accuracy, R2, MSE and MAE, corresponding to the task-specific objectives.

4.4. Model Architecture

We use Vision Transformer (ViT) models to examine whether incorporating Inter-Channel attention – achieved by channel-specific patchification and token padding and masking – in addition to Intra-Channel attention enhances model performance on biological image tasks. We employ a ViT-Tiny architecture as the backbone. The model, employs a shared 2D convolutional layer to embed each token with an embedding dimension of 192. Due to the dataset's maximum channel count (C_{max}) being 10 and the image size being 224x224, the ChAda-ViT model processes a significantly high number of input tokens–10 times more than a standard ViT-Tiny and 50% more than a ViT-Large. To avoid confusion with traditional ViT size nomenclature, we adopt a distinct model name, dubbed ChAda-ViT Moyen (French for *Average*), reflecting its expanded width. Experiments with different ChAda-ViT architecture sizes (Grand and Petit) to confirm the scaling laws of our method are available in the Appendix. The proposed approach is compared to Microsnoop One Channel approach [47], using a standard ViT for a fair comparison, modelling the function Ψ as defined in Eq. 4, to serve as the baseline, using similar backbone, embedding size, and token per channel count to evaluate the effects of our channel-adaptive contribution. This baseline encodes each channel separately with Intra-Channel attention, and then combines the resulting *CLS* token representations into a $n_i \times 192$-dimensional image representation. Furthermore, we introduce an Inter-Channel only ViT variant as an ablation study of the inter-channel attention only, where each channel is treated as a distinct single patch of size 224x224, as opposed to the 16x16 patch size used in the one-channel ViT and ChAda ViT. Each of these full-sized channel patches is tokenized into a 192-dimensional vector, compelling the model to focus its attention solely on the features derived from the relationships between the individual channels by eschewing Intra-Channel considerations.

5. Experiments

Model Training. Given the heterogeneous nature of the data, with its assortment of unrelated experiments, diverse image types, channel configurations, cell lines, and labels, usage of standard supervised approaches presents a unique challenge due to the strong label variability and occasional label absence.. Therefore, we aim to obtain broad representations through self-supervised learning (SSL), assessing these models on specialized downstream tasks on biological images. The three models are thus trained on the IDR-Cell100k dataset we created with DINO [8] as the SSL strategy for 400 epochs. ChAda-ViT Moyen and Inter-Channel only ViT Moyen are set with a base learning rate of 0.0001, while the one-channel ViT-Tiny is trained with a base learning rate of 0.005 for the sake of stability during training. A batch size of 256 multiplexed images per GPU is maintained for both models. Training employs a cosine annealing scheduler to optimize the learning rate over time. The ChAda-ViT Moyen undergoes training on 32 A100 80GB GPUs distributed across four nodes, for a total training time of 2080 GPU hours.

Evaluation. We assess our models, trained on IDR-Cell100k, to gauge their capacity to generate versatile representations ideal for a range of biologically relevant tasks based on known benchmarking datasets unrelated to the training set, using linear probing, embedding direct evaluation (KNN), as well as image generation for evaluation with 5 different random seeds per run. Classification tasks involved differentiating cell types, protein localizations, and cell cycle stages within the BloodMNIST[48], CyclOPS, and BBBC048 datasets, respectively, each of which have varying sizes and complexities. For regression, the NF-kB Nuclear Translocation Assay dataset[31] tested the models' ability to quantify protein displacement between cancer cells compartments. The generative task with the BBBC021 dataset[6] challenged the models to reconstruct cell imaging channels from the encoded *CLS* representations. We froze the encoder and trained a simple convolutional decoder to predict the Actin channel based on other channels for this task. Table 1 presents a comprehensive view of the evaluation tasks and dataset details. The One Channel model's evaluation involved using concatenated CLS token representations from each channel for a comprehensive representation. Performance metrics, accuracy for classification and R2 for regression and generation, in addition to Mean Squared Error (MSE) and Mean Absolute Error (MAE) for generation, were selected to measure the models' efficacy

[†] The dataset can be accessed on Kaggle: CYCLoPs Dataset.

[‡] More details on the dataset are found here : BBBC048.

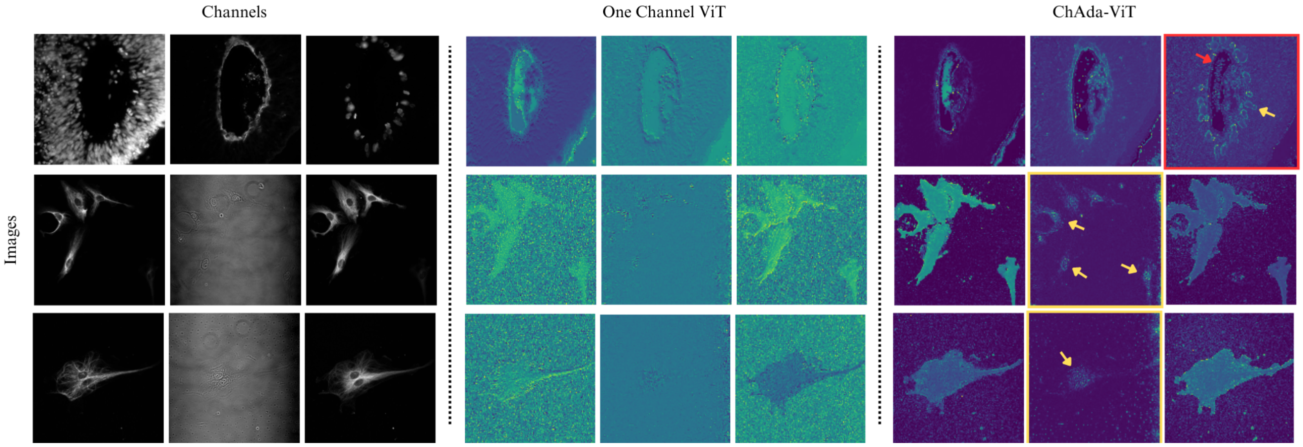

Figure 3. Comparison of the last layer self-attention maps between One Channel ViT and ChAda-ViT on IDRCell100k image channels. ChAda-ViT, utilizing Inter-Channel Attention, effectively discerns significant cross-channel correlations (red arrow), focusing on spatially relevant areas in each channel (yellow arrows). This mechanism enables ChAda-ViT to identify critical biological features that might be overlooked with a single-channel focus.

Dataset	One Channel ViT/16 Tiny	ChAda ViT/16 Moy
BloodMNIST	576	192
CyclOPS	384	192
BBBC048	576	192
NF-kB Nuclear Transloc	576	192
BBBC021	384	192

Table 2. Comparison of representation dimensions across downstream task datasets for One Channel ViT/16 Tiny and ChAda ViT/16 Moyen using CLS token only. The One Channel ViT, influenced by its dependence on the input channel counts, offers larger and more varied embedding dimensions, theorically leading to more extensive but less channel-interrelated image representations.

precisely. Detailed dataset information and biological context are provided in the appendix.

6. Results

Biologically Relevant tasks. Our experimental outcomes, as detailed in Figure 1, indicate that the proposed ChAda-ViT model, with its dual focus on Inter-Channel and Intra-Channel Attention, surpasses the standard one-channel approach in 6 out of the 8 tasks evaluated. This is notable considering the one-channel approach employs a larger representation space when using the CLS token only, as shown in Table 2. Evaluation on all output tokens in Appendix showcases a similar pattern. These results underline the effectiveness of introducing an inter-channel attention mechanism while training on microscopy images and suggest it leads to a more subtle and efficient biological image representation. Additional comparisons on Standard ViT trained on the 3 channel subset of IDRCell100K in the Appendix further validates the added value of training with Inter-channel attention.

Moreover, the experiments suggest that solely employing Inter-Channel attention, as shown by the Inter-Channel only ViT, might be insufficient for capturing the complexities of biological imaging. Instead, the amalgamation of both Inter-Channel and Intra-Channel Attention, as implemented in ChAda-ViT, yields superior representation quality compared to the application of each method in isolation. This integrative approach harnesses the strengths of both attention mechanisms to enhance the model's performance.

However, in the classification and clustering tasks within the BBBC048 dataset, the One Channel ViT approach, focusing primarily on Intra-Channel Attention, demonstrates a better performance over ChAda-ViT. This outcome could be attributed to the characteristics of the BBBC048 dataset (illustrated in Appendix). The images in this dataset reveal that certain features, crucial for classifying cell cycle stages, are already predominantly present within each single channel. With the representation power of One Channel ViT being bigger than ChAda-ViT Moyen (see Table 2), for the same input size, it performs better at this classification task, diminishing the need for Inter-Channel relationship analysis for accurate classification.

Inter-Channel Attention. The comparative analysis of the last-layer self-attention maps between One Channel ViT and ChAda-ViT on the IDRCell100k image channels, as depicted in Figure 3, reveals significant insights into the models' focus and interpretability. ChAda-ViT's use of Inter-Channel Attention is a pivotal aspect that distinguishes its performance from the One Channel ViT. Specifically, ChAda-ViT demonstrates a heightened ability to establish meaningful correlations across different channels, as indicated by the red arrow. This capability allows it to associate biological information from various channels, effec-

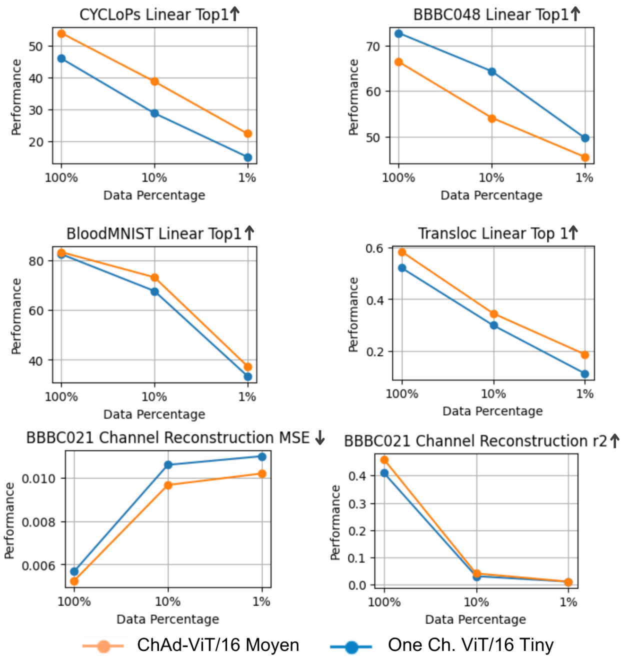

Figure 4. Evaluation of ChAda-ViT's performance in linear probing low-data regimes, utilizing 100%, 10%, and 1% of the training data for linear probing. The results consistently demonstrate ChAda-ViT's relative performance across tasks, maintaining the performance trends in the same tasks, regardless of the data volume.

tively enhancing focus on spatial locations in a channel that might otherwise appear information-scarce. Similarly, yellow arrows highlight the ChAda-ViT's strategic focus on spatially relevant areas, such as intra-nuclei features in the yellow-framed images, even in channels with limited information—contrasting with the One Channel ViT's approach. This targeted approach allows ChAda-ViT to unveil key biological features, which could potentially be missed in a single-channel analysis. Such capability is especially valuable in complex biological imaging where multiple channels convey different but interconnected information.

While qualitative, the implications of these findings are substantial in the context of biological image analysis. Evidences in ChAda-ViT's proficiency to recognize and emphasize spatially relevant areas across channels underscores its utility in deciphering complex biological structures and functions. This multi-channel attention helps the model to construct a more holistic and nuanced understanding of the cellular components and their interactions. This intricate understanding could lead to more accurate and comprehensive interpretations of biological data, particularly in scenarios where multiple channels contribute to the overall picture. The attention maps are thus consistent with the effectiveness of ChAda-ViT's design but also offer a window into the model's operational dynamics, highlighting its potential to significantly enhance the analysis of multi-channel biological imaging.

Low Data Regime. We then delved into how the performance of ChAda-ViT models compares to one-channel ViT models under constrained data conditions. Specifically, we conducted linear probing using 100%, 10%, and 1% of the available training data for each downstream task. This aspect of the study is crucial for evaluating model behavior in real world laboratory context, where data related to a specific experiment is often limited. Such situations commonly involve either fine-tuning a pre-trained model with the available data or employing the model directly as is.

As illustrated in Figure 4, our findings reveal that the ChAda-ViT model maintains a consistent performance trend across various data regimes. This persistence in performance, regardless of the amount of data used, is particularly significant. It indicates that once pre-trained, our ChAda-ViT model is robust and capable of achieving high-quality results, even in low-data regime scenarios commonly encountered in biological research. This consistency underscores the practical applicability and reliability of our models in diverse real-world biological research settings.

Single Joint Embedding Space. In addition, a pivotal contribution of the ChAda-ViT model lies in its ability to unify different datasets, each with distinct channel types and counts, into a consistent embedding space. This feature could end up being particularly advantageous for cross-experimental studies in biology, where diverse experiments often produce sparse data with varying imaging techniques and channels. Such a unified approach would highly benefit fields such as drug discovery, where integrating varied experimental results can provide deeper insights.

For instance, the BBBC021 dataset, composed of 3 channels per image, showcases whole-slide images of cells responding to specific chemical compounds, captured using the Cell Painting technique. Similarly, the Bray dataset[5], composed of 5 channels per image, as well as different experimental conditions, displays cells affected by various compounds, but imaged with a higher channel counts encoding distinct biological information than the first dataset. Notably, these datasets share at least 22 common compound treatments that could benefit from cross-dataset comparisons. By examining these shared compound treatments, it might be possible shed light on the interconnection of different mechanisms of action and understand the broader impact of these chemical compounds. However, existing approaches are unable to properly compare these different experimental settings, due different learnt embedding spaces caused by differences in channel count. This common representation space, uniquely facilitated by ChAda-ViT architecture, allows us to leverage data from other compounds or experiments, providing a bridge between numerous but disconnected biological datasets currently available. This

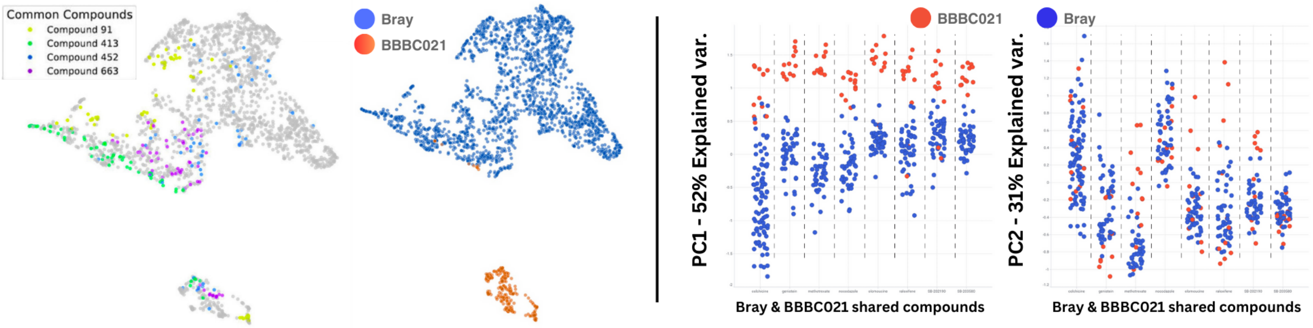

Figure 5. UMAP & PCA projections of the BBBC021 (3 channels) and Bray (5 channels) datasets in a unified representation space, derived from ChAda-ViT Moyen. The left UMAP projection highlights the possibility of projecting structurally different experiments into the same space. The right PCA projection and components are labeled by dataset, with a linear separation of the structural differences of different experiments (PC1) and different compounds (PC2) using only the two datasets' common compounds.

approach could potentially enhance our understanding of a wider range of compounds, serving as a preliminary step that capitalizes on the variety of open experiments in this domain, and laying the groundwork for more extensive future research.

Figure 5 presents a UMAP projection of these two datasets, highlighting their common compounds, within the same representation (left), achieved through ChAda-ViT Moyen, while similarly highlighting that the model, trained on only 100k images before projecting the representatons of the two datasets' common compound using a PCA, can linearly separate the two datasets through only the PCA first Principal Component (PC1), while differentiating the different compounds from each other in the PCA second component (PC2) (right). While aligning various labels (such as compound and cell types) in the representation is necessary for some cross-dataset tasks, this shared representation space is a significant initial step. It serves as a foundation for further exploration and potential breakthroughs in cross-experimental biological research.

7. Conclusion

In this study, we introduced ChAda-ViT, a Channel-Adaptive Vision Transformer designed specifically for multichannel image data. ChAda-ViT integrates token padding, masking as well as channel and positional embeddings within a patch-based framework, making it highly effective for handling diverse types of imaging data. On top of the added inter-channel attention, the standout feature of ChAda-ViT is the ability to bring disparate microscopy images into a single, joint embedding space, facilitating comprehensive comparisons across varied datasets with distinct types, channel counts, and imaging techniques.

Our model demonstrates superior performance over existing methods in the literature across most tasks, including normal and low data regime linear probing. It also provides more insightful attention maps at the channel level, paving the way for new explorations into cellular components and their interactions. The potential of ChAda-ViT extends to aligning various biological datasets within its embedding space, offering novel avenues for cross-experimental studies and new biological insights. This capability offers a significant potential to the field of biological image analysis, e.g. leading to augment datasets with additional channels based on learnings from other datasets, thereby reducing experimental costs.

Another key contribution of our work is the introduction of the IDRCell100k dataset, a first-of-its-kind collection featuring heterogeneous biological image sources and a range of multi-channel configurations. This dataset not only demonstrates the adaptability and robustness of ChAda-ViT but also serves as a critical asset for ongoing and future research in representation learning for biological images.

Nevertheless, our research is not without limitations. Although ChAda-ViT excels in many tasks, it falls short in tasks heavily reliant on intra-channel information. Addressing this shortfall could position our approach as the go-to method for encoding biological images, encapsulating both inter- and intra-channel dynamics. Furthermore, exploring ChAda-ViT as basis for a foundation model, potentially leveraging massive datasets, could unveil new capabilities and applications. Bridging this gap and expanding its dataset foundation remain key objectives for future enhancements, with the aim of solidifying ChAda-ViT's role as a cornerstone technology in biological imaging analysis.

8. Acknowledgments

This work was supported by ANR–10–LABX–54 MEMOLIFE and ANR–10 IDEX 0001-02 PSL* Université Paris and was granted access to the HPC resources of IDRIS under the allocations 2020-AD011011495 and 2023-AD011014487 made by GENCI.

References

[1] Cédric Allier, Lionel Hervé, Chiara Paviolo, Ondrej Mandula, Olivier Cioni, William Pierre, Francesca Andriani, Kiran Padmanabhan, and Sophie Morales. Cnn-based cell analysis: From image to quantitative representation. *Frontiers in Bioengineering and Biotechnology*, 9:673840, 2021. 2

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *ICCV*, 2023. 4

[3] Ihab Bendidi, Adrien Bardes, Ethan Cohen, Alexis Lamiable, Guillaume Bollot, and Auguste Genovesio. No free lunch in self supervised representation learning, 2023. 2

[4] Anis Bourou and Kévin Daupin et al. Unpaired image-to-image translation with limited data to reveal subtle phenotypes. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 2

[5] Mark-Anthony Bray and Sigrun M et al. Gustafsdottir. A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *GigaScience*, 6(12):giw014, 2017. 7, 4

[6] Peter D. Caie, Rebecca E. Walls, and Alexandra Ingleston-Orme et al. High-Content Phenotypic Profiling of Drug Response Signatures across Distinct Cancer Cells. *Molecular Cancer Therapeutics*, 9(6):1913–1926, 2010. 5

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 5

[9] Srinivas Niranj Chandrasekaran and Jeanelle Ackerman et al. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023. 2

[10] Srinivas Niranj Chandrasekaran, Beth A. Cimini, and Amy Goodale et al. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *bioRxiv*, 2022. 2

[11] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan Loddon Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *ArXiv*, abs/2102.04306, 2021. 2

[12] Ethan Cohen and Virginie Uhlmann. Aura-net: Robust segmentation of phase-contrast microscopy images with few annotations. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 640–644, 2021. 2

[13] Ethan Cohen, Maxime Corbé, Cláudio Areias Franco, Francisca F. Vasconcelos, Franck Perez, Elaine Del Nery, Guillaume Bollot, and Auguste Genovesio. Cell painting transfer increases screening hit rate. *bioRxiv*, 2022. 2

[14] Jan Oscar Cross-Zamirski, Elizabeth Mouchet, Guy B. Williams, Carola-Bibiane Schönlieb, Riku Turkki, and Yinhai Wang. Label-free prediction of cell painting from brightfield images. *Scientific Reports*, 12, 2021. 2

[15] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1610, 2020. 2

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 1

[17] Michael Doron et al. Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023. 2

[18] Maxime Oquab et al. Dinov2: Learning robust visual features without supervision, 2023. 2

[19] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. 2

[20] Watkinson Gabriel, Cohen Ethan, Bourriez Nicolas, Bendidi Ihab, Bollot Guillaume, and Genovesio Auguste. Weakly supervised cross-model learning in high-content screening, 2023. 2

[21] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil S. Bhate, Matthew B. Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174:968 – 981.e15, 2017. 2

[22] John H Harkness, Jacob J. Theis, Will M. O'Keefe, Grant W. Wade, and Kristy J. Lawton. Leveraging ai transfer learning for rapid and accurate identification and quantification of cellular biomarkers in microscopy images. *The FASEB Journal*, 36, 2022. 2

[23] Matthew Hartley, Gerard J. Kleywegt, Ardan Patwardhan, Ugis Sarkans, Jason R. Swedlow, and Alvis Brazma. The bioimage archive - building a home for life-sciences microscopy data. *bioRxiv*, 2021. 2

[24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 1

[25] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 4

[26] L. Hervé, D. C. A. Kraemer, O. Cioni, O.and Mandula, M. Menneteau, S. Morales, and C. Allier. Alternation of inverse problem approach and deep learning for lens-free microscopy image reconstruction. *Scientific Reports*, 10(1): 20207, 2020. 2

[27] John W. Hickey, Elizabeth K. Neumann, and Andrea J. Radtke et al. Spatial mapping of protein composition and tissue organization: a primer for multiplexed antibody-based imaging. *Nature Methods*, 19:284 – 295, 2021. 2

[28] Stanley Hua et al. Cytoimagenet: A large-scale pretraining dataset for bioimage transfer learning. In *NeuIPS LMRL Workshop*, 2021. 3

[29] Vladislav Kim, Nikolaos Adaloglou, Marc Osterland, Flavio M Morelli, and Paula A Marin Zapata. Self-supervision advances morphological profiling by unlocking powerful image representations. *bioRxiv*, 2023. 2

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. 1

[31] Alexis Lamiable, Tiphaine Champetier, Francesco Leonardi, Ethan Cohen, Peter Sommer, David Hardy, Nicolas Argy, Achille Massougbodji, Elaine Del Nery, Gilles Cottrell, et al. Revealing invisible cell phenotypes with conditional generative modeling. *Nature Communications*, 14(1):6386, 2023. 2, 5

[32] Yann LeCun and Bernhard E. Boser et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989. 1

[33] Xinyang Li, Yuanlong Zhang, Jiamin Wu, and Qionghai Dai. Challenges and opportunities in bioimage analysis. *Nature Methods*, 20(4):367–377, 2023. https://www.nature.com/articles/s41592-023-01900-4. 2

[34] Umar Masud, Ethan O. Cohen, Ihab Bendidi, Guillaume Bollot, and Auguste Genovesio. Comparison of semi-supervised learning methods for high content screening quality control. In *ECCV Workshops*, 2022. 2

[35] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A Cimini, Kyle W Karhohs, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS Biol*, 16(7):e2005970, 2018. 2

[36] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12308, 2021. 2

[37] Nikita Moshkov, Tim Becker, Kevin Yang, Peter Horvath, Vlado Dancik, Bridget K Wagner, Paul A Clemons, Shantanu Singh, Anne E Carpenter, and Juan C Caicedo. Predicting compound activity from phenotypic profiles and chemical structures. *Nature Communications*, 14(1):1967, 2023. 2

[38] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *ArXiv*, abs/2106.02034, 2021. 1

[39] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634, 2021. 2

[40] Zongyao Sha and Jianfeng Li. Mitformer: A multiinstance vision transformer for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 2

[41] Douglas Thain, Todd Tannenbaum, and Miron Livny. Distributed computing in practice: the condor experience. *Concurrency - Practice and Experience*, 17(2-4):323–356, 2005. 3

[42] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *PMLR*, 2021. 2

[43] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 1, 4

[44] Lucas von Chamier, Romain F. Laine, and Johanna Jukkala et al. Democratising deep learning for microscopy with zerocostdl4mic. *Nature Communications*, 12, 2021. 2

[45] Eleanor Williams and Josh et al. Moore. Image data resource: a bioimage data integration and publication platform. *Nature Methods*, 14(8):775–781, 2017. 2, 3

[46] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *ArXiv*, abs/2105.15203, 2021. 2

[47] Dejin Xun, Rui Wang, and Yi Wang. Microsnoop: a generalist tool for the unbiased representation of heterogeneous microscopy images. *bioRxiv*, 2023. 1, 2, 3, 5

[48] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2 - a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. 5, 2

[49] Shuang Yu, Kai Ma, Qi Bi, Cheng Bian, Munan Ning, Nanjun He, Yuexiang Li, Hanruo Liu, and Yefeng Zheng. Mil-vt: Multiple instance learning enhanced vision transformer for fundus image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. 2

[50] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis E. H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 538–547, 2021. 2

[51] Shuangjia Zheng, Jiahua Rao, Jixian Zhang, Ethan Cohen, Chengtao Li, and Yuedong Yang. Cross-modal graph contrastive learning with cellular images. *bioRxiv*, pages 2022–06, 2022. 2