UNIVERSITY OF CALIFORNIA MERCED

Optimal Transport Driven Deep Learning with Emphasis on Pathology Images

By
Mohammed Aburidi

A capstone project submitted in partial satisfaction of the requirements for the degree of Masters of Science in Applied Mathematics

Committee in charge:
Prof. Roummel Marcia
Prof. Arnold Kim

August 2023

We approve the capstone project report of Mohammed Aburidi

Date of Signature

Roummel Marcia
Professor, Department of Applied Mathematics
Thesis Advisor
University of California Merced

Arnold Kim
Professor, Department of Applied Mathematics
Thesis Committee
University of California Merced

1 Introduction

Traditionally, tissue histopathology slide examination under the microscope by a human pathologist is considered the gold standard for cancer diagnosis and determining treatment options for a patient. In recent years, digital pathology has revolutionized the clinical practice of pathology by capturing the entire tissue slide using a slide scanner to create high-resolution whole slide images (WSIs) [1]. The increasing availability and adoption of WSIs for routine diagnosis have also given rise to the new discipline of computational pathology, which aims at addressing the often time-consuming and costly diagnosis process by the development of computer vision and machine learning techniques to automatically analyze WSIs and assist pathologists in diagnostic tasks [2, 3]. Such recent advances in machine learning have rapidly improved pathology workflow by providing more objective and reproducible results, leading to better patient care [4, 5].

However, existing methods in computational pathology have suffered from major limitations [6, 7]. One of the main bottlenecks is the lack of high-quality labeled data needed for training models with high accuracy and robustness, where annotations are a labor-intensive and error-prone task to acquire and relies on medical expertise. To compensate for the scarcity of labeled datasets, the development of efficient unsupervised and self-supervised techniques is essential. Moreover, the computational complexity associated with whole slide images is considered another main challenge. Training a deep learning network on entire WSIs at full resolution is computationally intractable as the size of WSIs could reach multi-gigapixles. Another major limitation, of the current deep learning techniques, is that they are inefficient when dealing with relation-aware representations, thus, they cannot benefit from the organization and the structure of the cells and the tissues in the WSI. Such limitations have slowed the transition from research results to clinically deployed applications.

To address the computational complexity issue associated with WSI, the typical and most used approach is to sub-divided the image into small patches, where each patch is processed independently in the neural network [7], then the predicted scores for each patch within a WSI are aggregated [8, 9, 10]. However, patches provide a limited visual context, and the optimal resolution and patch size for analysis is highly problem-dependent. Further, the correlations among these patches are ignored during traditional deep learning feature learning. Consequently, machine learning methods that are based on patch-level cannot capture the overall structure and organization of the tissue in a WSI. Modern deep learning variations of graph neural networks (GNNs) have made a significant impact in many technological domains for describing relationships. Thus, it can be utilized to estimate the dependencies between patches and enhance the discriminative ability of the network features. Graphs, by definition, capture relationships between entities and can thus be used to encode relational information between variables [11]. Given the utility of graphs in modeling the histology of cancer tissue, special emphasis has been placed to exploit recent developments in deep learning for graphs in this domain. However, these applications are still in their nascent stages when compared to existing typical conventional deep learning methods. There are challenges associated with the adoption of GNNs into digital pathology. Such challenges have existed at the following levels: 1) entity graph construction, 2) the training paradigms, 3) explainability of graph models, 4) complexity of graph models, 5) the embedding of expert knowledge and many others.

Optimal Transport (OT) is a mathematical framework that defines the problem of finding the most efficient way (i.e., lowest cost) of moving an object such as probability distribution from one configuration onto another (e.g., matching two distributions or finding the similarity between two distributions). OT problems were initiated by Gaspard Monge (1746–1818), a French mathematician, in the 18th century [12]. OT has been gaining in recent years increasing attention as a promising and useful tool in the machine-learning community. This success is due to its capacity to exploit the geometric property of the samples at hand. OT methods have been successfully employed in a wide variety of machine learning applications [13, 14, 15, 16, 14, 17], computer vision [18, 19], generative adversarial networks, domain adaptation [20]. Recently, applications of OT to biology have been proposed [21, 22, 23, 24].

The main goals of this research are the following:

Goal #1: Utilize the power of OT in conjunction with deep learning to address the annotations scarcity issue of pathology WSI.

Goal #2: Develop optimization frameworks to capture relation-aware representations of different entities from whole slide pathology images.

Goal #3: Improve upon the existing automated cancer biomarkers detection AI-based methods.

To accomplish these goals, we intend to complete the following projects:

1. We aim to use contrastive learning using optimal transport to self-assign labels for non-annotated whole-slide pathology images, thus, achieving **Goal #1**.
2. We aim to build an OT-based mathematical framework to construct cell-level graphs from pathology images which would then be used to train GCN.
3. We aim to extend our graph reconstruction framework to include different types of entities such as different tissues and cells from the same image, by constructing hierarchical multi-level graphs.
4. We plan to develop a graph contrastive learning based on optimal transport to address the lack of pathology image annotations.

Projects 2 and 3 will help achieve **Goal #2.** We plan to benchmark and evaluate the performance of our developed methods using real pathology images from The Cancer Genome Atlas in breast cancer (TCGA) [25], thus, achieving **Goal #3**.

This report is organized as follows: Section 2 describes the background for optimal transport. Section 3 shows an OT-based learning framework we developed named OTCC, alongside the results. Section 4 shows our second framework named CLOT, which is built based on OTCC for cluster assignment. In sections 5 to 8, we provide brief descriptions of the future projects we aim to accomplish in the Ph.D. time. Section 9 talks about the datasets to be used to evaluate our methods and prediction tasks to be accomplished. Section 10 shows the timeline of our proposed research. In the last section, we conclude the report.

2 Background on optimal transport

Optimal Transport [26] is a mathematical framework that defines the problem of finding the most efficient way of moving an object such as probability distribution from one configuration onto another (e.g., matching two distributions or finding the similarity between two distributions). Efficient here means with a lower cost. Let $\mathcal{X} = \{x_i\}_{i=1}^N$ and $\mathcal{Y} = \{y_i\}_{i=1}^K$ be two point clouds representing the source and target samples, respectively. Let $\mathbf{p} \in \mathcal{H}_N$ and $\mathbf{q} \in \mathcal{H}_K$ to be two discretized distributions of interest, where $\mathcal{H}_N$, and $\mathcal{H}_K$ are histograms of N, K bins, with $\{\mathbf{p} \in \mathbb{R}_+^N, \sum_i p_i = 1\}$, $\{\mathbf{q} \in \mathbb{R}_+^K, \sum_i q_i = 1\}$, respectively. Thus,

$$\mathbf{q} = \sum_{i=1}^{K} q_i \delta_{y_i} \quad \textit{and} \quad \mathbf{p} = \sum_{j=1}^{N} p_j \delta_{x_j}$$

where δ is the Dirac function. Let $Q = (Q_{ij})_{i,j}$ defined as a transportation plan or the couplings matrix that describes the amount of mass p_j found at x_j to be flowed toward the mass q_i at y_i. In addition for Q to being nonnegative, it must satisfy the following two conditions

$$\sum_{j=1}^{N} Q_{ij} = q_i \quad \forall i \in \{1, .., K\} \qquad \textit{and} \qquad \sum_{i=1}^{K} Q_{ij} = p_j \quad \forall j \in \{1, .., N\}$$

Optimal Transport addresses the problem of optimally transporting $\mathbf{p}$ toward $\mathbf{q}$, given a cost C_{ij} measured as a geometric distance between x_i and y_j. The total cost of a transport plan is then:

$$\langle C, Q \rangle_F = \sum_{i=1}^{K} \sum_{j=1}^{N} C_{ij} Q_{ij} \tag{1}$$

where $\langle \cdot , \cdot \rangle$ is the Frobenius dot-product of two matrices. The optimal transport is therefore given by the following optimization problem:

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & \langle C, Q\rangle_F \\ \text{subject to} \quad & \sum_{j=1}^{N} Q_{ij} = q_i \quad \forall i \in \{1,..,K\} \\ & \sum_{i=1}^{K} Q_{ij} = p_j \quad \forall j \in \{1,..,N\} \\ & Q_{ij} \geq 0 \quad \forall (i,j) \in \{1,..,K\} \times \{1,..,N\} \end{aligned} \tag{2}$$

At each bin, the transport must distribute the exact amount of the mass p_i and must match the final amount of targeted mass q_j. We can state the optimization problem even more compactly

$$\begin{aligned} \underset{Q}{\text{minimize}} \quad & \langle C, Q\rangle_F \\ \text{subject to} \quad & Q\mathbf{1}_N = \mathbf{q} \qquad Q^T\mathbf{1}_K = \mathbf{p} \qquad Q \geq 0 \end{aligned} \tag{3}$$

where $\mathbf{1}_N$, $\mathbf{1}_K$ denote the vectors of ones in dimension N, and K, respectively. We define the set of all admissible couplings or transport plans $\mathbf{Q}(\mathbf{p}, \mathbf{q})$ between histograms is given by

$$\mathbf{Q}(\mathbf{p}, \mathbf{q}) = \{Q \in \mathbb{R}^{K\times N} \mid Q\mathbf{1}_N = \mathbf{q}, Q^T\mathbf{1}_K = \mathbf{p}\}$$

.
More specifically, when the cost C is a distance matrix. The optimal transport is the Wasserstein distance on $\mathcal{H}_N \times \mathcal{H}_K$ which is defined as:

$$W(\mathbf{p}, \mathbf{q}) = \min_{Q\in\mathbf{Q}(\mathbf{p},\mathbf{q})} \langle C, Q\rangle_F = \min_{Q\in\mathbf{Q}(\mathbf{p},\mathbf{q})} \sum_{i=1}^{K}\sum_{j=1}^{N} C_{ij}Q_{ij} \tag{4}$$

Although the Wasserstein distance has appealing theoretical properties and an intuitive formulation, its computation involves the resolution of a linear program and can thus be solved in polynomial time, which is nonpractical especially when histograms' dimension exceeds thousands if not millions. To address this computation issue, Cuturi [26] has proposed to smooth the classic optimal transport problem with an entropic regularization term, and show that the resulting optimum can be computed through Sinkhorn's matrix scaling algorithm at a speed that is several orders of magnitude faster than that of transport solvers. Thus, the modified optimal transport, also called dual-sinkhorn distance, is given by

$$\underset{Q\in\mathbf{Q}(\mathbf{p},\mathbf{q})}{\text{minimize}} \quad \langle C, Q\rangle_F - \frac{1}{\lambda} S(Q) \tag{5}$$

where $S(Q) = -\sum_{i=1}^{N}\sum_{j=1}^{K} Q_{ij}\log Q_{ij}$ is the entropy. This entropy regularization forms a simple structure on the optimal regularized transport. According to transport theory [27], the optimum Q^λ can be written as a rescaled version of $e^{-\lambda C}$. The existence and uniqueness of Q^λ follows from the boundedness of the set $\mathbf{Q}(\mathbf{p}, \mathbf{q})$ and the strict convexity of minus the entropy. To find Q^λ, we use the method of Lagrange multipliers, we first find the Lagrangian of equation (5) $\mathcal{L}(Q, \alpha, \beta)$ with dual variables $\alpha \in \mathbb{R}^K, \beta \in \mathbb{R}^N$ for the two equality constraints in $\mathbf{Q}(\mathbf{p}, \mathbf{q})$. The Lagrangian is given by

$$\mathcal{L}(Q, \alpha, \beta) = \sum_{i=1}^{K}\sum_{j=1}^{N} \left[\frac{1}{\lambda} Q_{ij}\log Q_{ij} + Q_{ij}C_{ij}\right] + \alpha^T(Q\mathbf{1}_N - \mathbf{q}) + \beta^T(Q\mathbf{1}_K - \mathbf{p}) \tag{6}$$

For any couple (i, j), $(\partial\mathcal{L}/Q_{ij} = 0) \Rightarrow Q_{ij} = e^{1/2-\lambda\alpha_i}e^{-\lambda Q_{ij}}e^{-1/2-\lambda\beta_j}$. Sinkhorn's theorem (1967) states that there exists a unique matrix of the form $diag(u)\ e^{-\lambda C}\ diag(v)$ that belongs to $\mathbf{Q}(\mathbf{p}, \mathbf{q})$ where $u \geq \mathbf{0}_K, v \geq \mathbf{0}_N$. Q^λ is thus necessarily that matrix, and can be computed with Sinkhorn's fixed point iteration, where u and v are updated such that the constraints are satisfied, in such a way $(u, v) \leftarrow (\mathbf{q}./(e^{-\lambda C}v), \mathbf{p}./((e^{-\lambda C})^T u))$.

Part I
Completed projects

3 Self-labeling as an optimal transport

3.1 Motivation

Deep neural networks (DNNs) have achieved considerable progress in learning strong and discriminative representations. While considerable breakthroughs have been made by DNNs in diverse applications such as computer vision and medical imaging, speech recognition, natural language processing, and time series analysis [28], learning a powerful representation often requires a large-scale dataset with manually curated ground-truth labels, which has proven to be a bottleneck for the continued development of state of the art performance and in its deployment in many application areas.

Self-supervised learning (SSL) is an increasingly popular framework that aims at obtaining features without using manual annotations [29, 30]. State-of-the-art SSL paradigms are designed in learning image representations while using a clustering algorithm in an end-to-end fashion as a means of providing pseudo labels for training a deep model in downstream applications. Typical deep learning-based clustering algorithms are based on alternation learning in which they alternate between the representation learning and clustering assignment steps [31, 32]. Such methods work iteratively to assign features in a latent space into clusters. It then uses these assignments to update the deep network. A main drawback of this approach is that clustering should pass through image features of the entire dataset (i.e., offline), which makes it less applicable to large-scale learning scenarios. Further, this approach suffers from an accumulated error during the two stages of representation learning and clustering. In order to address this issue, online clustering methods are required.

In this project, we employ optimal transport theory to assign cluster labels simultaneously (online). We develop a clustering method called Optimal Transport-based Contrastive Clustering (OTCC). Thus, providing a robust self-supervised deep training. Our framework extends the standard cross-entropy minimization to an optimal transport problem and solves it using a fast variant of the Sinkhorn-Knopp algorithm to produce the cluster assignments. Moreover, inspired by contrastive learning, we enforce consistency between the produced assignments obtained from views of the same image. The features and the labels are learned online which allows our method to scale to unlimited amounts of data. Our deep learning framework is illustrated in 1.

3.2 Proposed method

Consider a given mini-batch $\mathbf{X}$ of N images $\{x_1, ..., x_N\}$. The idea of OTCC is to compute two random augmentations $\tilde{\mathbf{X}}^a$ and $\tilde{\mathbf{X}}^b$, and compute their latent space feature vectors using an encoder network f_θ, and a projection head $g_1(\cdot)$ consists two stacked nonlinear MLP layers projects into a subspace with a dimensionality of the cluster number followed by a softmax (i.e., prototype layer). It outputs the predicted cluster assignment probabilities or the cluster-level representations $\mathbf{p}^a$, $\mathbf{p}^b \in \mathbb{R}^{N \times K}$, where K is the number of clusters.

Building upon the work of Cuturi et. al [26] on optimal transport, we encode the cluster labels as posterior distributions $q(y = k|x_i)$, and we formulate the problem of finding optimal assignments as an optimal transport optimization problem. We compare the class label predictions obtained from the projection head with assignments obtained when solving the optimization problem. If the two features p^a and p^b capture the same information, it should be possible to predict the label from the other feature vector. Thus, consistency is enforced.

To mathematically formulate the self-labeling problem as an optimal transport, we encode the labels as posterior distributions in the average cross-entropy objective [32, 33]. In this case, our loss will be

$$\mathcal{L}_{Opt}(p, q) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} q(y = k|x_i) \log p(y = k|x_i) \tag{7}$$

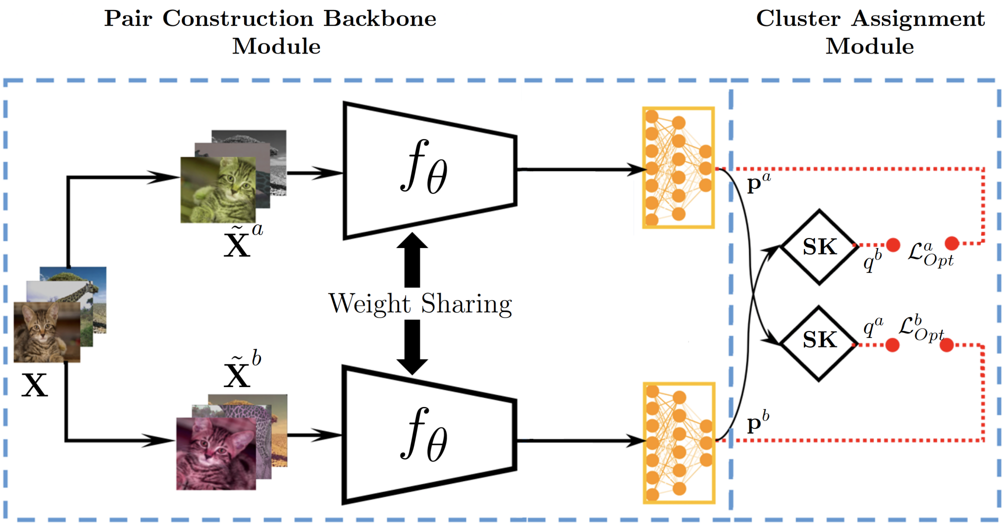


Figure 1: OTCC clustering framework. In the first module, data pairs $\tilde{\mathbf{X}}^a$, $\tilde{\mathbf{X}}^b$ are constructed using two augmentations of data $\mathbf{X}$. Then the features are extracted from the pairs using a shared encoder f_θ, followed by a projection head to obtain the class label probabilities. In the second module, the cluster assignment probabilities obtained from the head are used by the Sinkhorn-Knopp (SK) algorithm to generate ground-truth-like cluster assignments. The same head is used twice to generate the probability vector from the two views of the images. The outputs of the head are used in a way to enforce consistency in the cross-entropy loss function.

where the values in the vector $p(y|x_i) \in \{p^a_i, p^b_i\}$, and $q(y|x_i) \in \{q^a_i, q^b_i\}$. Optimizing q is the same as reassigning the labels, which leads to a degenerate solution, i.e., (7) can be trivially minimized by assigning all data points to a single and arbitrary class label. A common way to avoid this is by adding a constraint that enforces an equally-sized partition [33]. The learning objective is thus

$$\begin{aligned} &\underset{q}{\text{minimize}} \quad \mathcal{L}_{Opt}(p, q) \\ &\text{subject to} \quad \sum_{i=1}^{N} q(y = k|x_i) = \frac{N}{K}, \quad q(y = k|x_i) \in \{0, 1\}. \end{aligned} \tag{8}$$

At this step, we only optimize the labels, keeping the predictions p fixed, given a batch of images. The constraints mean that each data point x_i is assigned to exactly one class label and the N data points is split equally among the K classes. By reforming it as an optimal transport using the notations in [26], let $P_{y,i} = p(y|x_i)$ be the $K \times N$ matrix of joint probabilities which is estimated by the model, and $Q_{y,i} = q(y|x_i)/N$ be the $K \times N$ matrix of assigned joint probabilities. Using the notation of [26], we restrict the matrix Q to the transportation polytope $\mathbf{Q} = \{Q \in \mathbb{R}^{K\times N} \mid Q\mathbf{1}_N = \frac{1}{K}\mathbf{1}_K, Q^T\mathbf{1}_K = \frac{1}{N}\mathbf{1}_N\}$,

where $\mathbf{1}_N$ denotes the vector of ones in dimension N. The constraints enforce that the matrix Q splits the data uniformly. We then can rewrite the optimization problem (8) as

$$\underset{Q\in\mathbf{Q}}{\text{minimize}} \quad \langle Q, -\log P\rangle \tag{9}$$

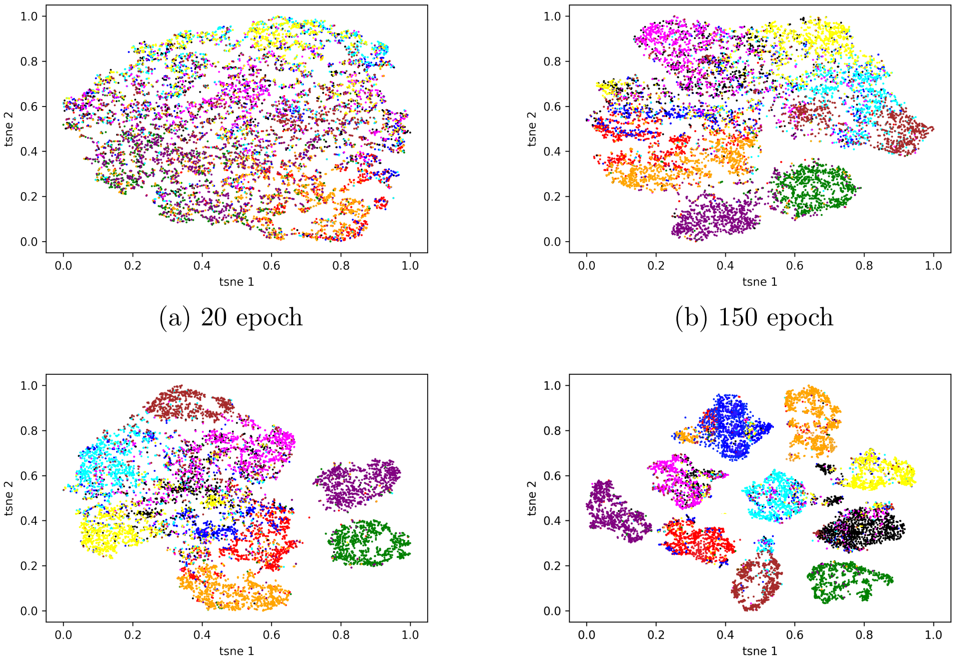
where $\langle\ \cdot\ ,\cdot\ \rangle$ is the Frobenius dot-product of two matrices. This optimization problem is linear optimization, and we would solve it using the last version of Sinkhorn-Knopp algorithm [26], which amounts by introducing a regularization term

$$\underset{Q\in\mathbf{Q}}{\text{minimize}} \quad \langle Q, -\log P\rangle - \frac{1}{\lambda}S(Q) \tag{10}$$

where $S(Q) = -\sum_{i=1}^{N}\sum_{j=1}^{K} q_{ij}\log q_{ij}$ is the entropy. This problem can be solved using the Lagrange multiplier for the entropy constraint of Sinkhorn distances [26], and its minimizer can be written as

$$Q = \text{Diag}(u)\, P^{\lambda}\, \text{Diag}(v), \tag{11}$$

where u and v are normalization vectors chosen such that the resulting matrix Q is also a probability matrix (see above for a derivation). Once Q is found, we optimize the overall objective defined next section to find the optimal P (i.e., the model parameters).



(a) 20 epoch

(b) 150 epoch

(a) 300 epoch

(b) 700 epoch

Figure 2: The evolution of features and cluster assignments across the training process on ImageNet-10. The colors indicate the cluster assignment obtained from the cluster assignment module and the features for t-SNE are computed from ResNet-34.

3.3 Results

Datasets: The proposed method was evaluated on three image datasets: CIFAR-100 [34], STL-10 [35], and ImageNet-10 [36]. Each dataset contains 10 classes except CIFAR-100, which contains 20 classes.

Implementation Details: We implement ResNet34 as an encoder backbone architecture [37] and use the Adam optimizer [38] to simultaneously optimize the two projection heads and the backbone network, with cosine learning rate scheduler [39]. The weight decay is set to 0.0001. ResNet is designed for images of size 224 × 224, so we resize all input images to this size. The projection head consists two-layer nonlinear MLP. ReLU activation was used between the two layers. Softmax activation was used int he in the cluster-level contrastive projection head to produce soft labels as in [40]. Following [41]. The batch size is set to 256 due to the memory limitation. All the models are trained from scratch for 1000 epochs. The training is carried out on UC Merced Pinnacles Cluster using one 2x NVIDIA Tesla A100 PCIe v4 40GB HBM2 Single GPU.

Data Augmentations: Following [41, 40] we use random cropping, color jittering, grayscale transformation, horizontal flipping, and Gaussian blurring for augmentation. Each transformation is applied with a certain probability.

Evaluation Metrics: We utilize three common clustering evaluation metrics including Accuracy (ACC), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI). Higher values indicate better performance.

In Table 1, we present the comparisons between our method with eight state-of-the-art clustering methods including K-means [42], CC [40], DCCM [43], PICA [44], DAC [36], DEC [45], JULE [46], VAE [47]. Clustering results of DAE, DCGAN, DeCNN, and VAE are obtained by applying k-means on the latent space features extracted from images. Results shown in Table 1 demonstrate

Table 1: The clustering performance on three image benchmarks. The best results are shown in boldface.

	CIFAR-100			STL-10			ImageNet-10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-means	0.130	0.085	0.028	0.192	0.125	0.061	0.241	0.119	0.057
JULE	0.137	0.103	0.033	0.277	0.182	0.164	0.300	0.175	0.138
VAE	0.152	0.108	0.040	0.282	0.200	0.146	0.334	0.193	0.168
DEC	0.185	0.136	0.050	0.359	0.276	0.186	0.381	0.282	0.203
DAC	0.238	0.185	0.088	0.470	0.366	0.257	0.527	0.394	0.302
DCCM	0.327	0.285	0.173	0.482	0.376	0.262	0.710	0.608	0.555
PICA	0.317	0.310	0.171	0.713	0.611	0.531	0.870	0.802	0.761
CC	0.429	0.431	0.266	0.850	0.764	0.726	0.893	0.859	0.822
OTCC(Ours)	**0.501**	**0.492**	**0.309**	**0.872**	**0.797**	**0.762**	**0.913**	**0.892**	**0.851**

the clustering ability of OTCC, which outperforms the baselines by a large margin on all of three datasets. Specifically, OTCC outperforms the closest competitor (CC) on the three datasets in terms of the three evaluation measures. The largest margin has been achieved on CIFAR-100, which is interesting as it has the largest number of classes. The results demonstrate how meaningful the cluster assignments obtained by solving the labeling problem as an optimal transport are. Figure 2 shows the evolution of features obtained from the backbone and cluster assignments across the training process on ImageNet-10. It demonstrates the ability of our method to cluster the instances.

4 Combining with contrastive learning

4.1 Motivation

To further improve the performance of OTCC, we proposed a modification to the original model by including the concepts of contrastive learning besides optimal transport. We thus propose a deep-based clustering method called Contrastive Learning driven and Optimal Transport-based (CLOT) clustering which focuses on the problem of obtaining the labels simultaneously. We contribute a new simultaneous and dual contrastive learning-based clustering framework that consists of two stages. In the first stage, instance- and cluster-level representations are learned by maximizing the similarities of the projections of positive pairs while minimizing those of negative ones, thus pushing away features from different images while pulling together those from the augmented views of the same image.

In the second stage, CLOT extends the standard cross-entropy minimization to an optimal transport problem and solves it using a fast variant of the Sinkhorn-Knopp algorithm to produce the cluster assignments. The cluster-level representation notion is used for the first time in [40], where it's learned beside the instance-level representation which is obtained by optimizing the typical contrastive objective. Our work improves upon both objectives by considering a third objective that compares the class assignments obtained from solving the self-labeling in an online fashion as an optimal transport and enforces consistency between the produced assignments obtained from views of the same image. Our framework thus allows contrasting different image views not only in terms of features but also in terms of cluster assignments. CLOT is similar to OTCC, however, we add a second project head to output the feature vector in a latent space besides two contrastive loss functions. The modified framework is illustrated in figure 3.

4.2 Proposed method

Contrastive learning maximizes the similarities of positive pairs (i.e., the transformed views of the same image) while minimizing those of negative ones by pushing away features from different images while pulling together those from the augmented views of the same image.

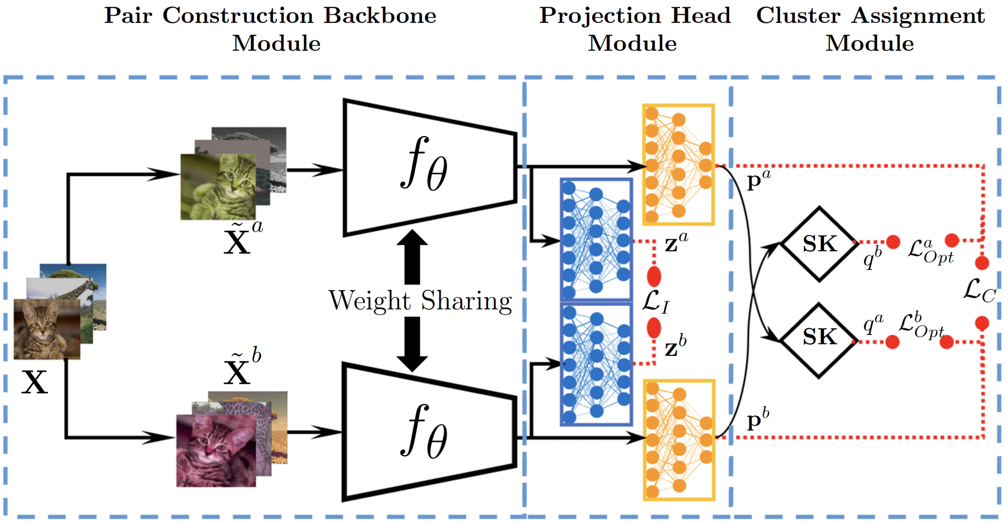


Figure 3: CLOT clustering framework. In the first module, data pairs $\tilde{\mathbf{X}}^a$, $\tilde{\mathbf{X}}^b$ are constructed using two augmentations of data $\mathbf{X}$. Then the features are extracted from the pairs using a shared encoder f_θ. In the second module, two different MLPs are used as projection heads to project the features into latent space (the blue head) and into subspace with a dimensionality of the cluster number followed by a softmax (the yellow head). Finally, the cluster assignment probabilities obtained from the yellow head are used by the Sinkhorn-Knopp (SK) algorithm to generate ground-truth-like cluster assignment, consistency is enforced as in OTCC. Losses are used in three levels, at the instance-level, cluster-level (soft labels), and at the output of the optimal transport problem.

To get the most out of contrastive learning power, we compute the latent space feature vectors $\mathbf{z}^a, \mathbf{z}^b \in \mathbb{R}^{N\times D}$ using an encoder network f_θ, and a projection head $g_2(\cdot)$ (two stacked nonlinear MLP layers). For a specific sample x_i^a, there are $2N-1$ pairs in total, among which we choose its corresponding augmented sample x_i^b to construct the positive pair $\{x_i^a, x_i^b\}$, and leave the rest $2N-2$ to be negative. The features $\mathbf{z}^a, \mathbf{z}^b$ in this case are the instance representations. We utilize an extra two loss functions. The first loss for a given sample x_i^a is of the form

$$\mathcal{L}^a_{I,i} = -\log \frac{\exp\left(\dfrac{s(z_i^a, z_i^b)}{\tau_I}\right)}{\sum_{j=1}^{N}\left\{\exp\left(\dfrac{s(z_i^a, z_j^a)}{\tau_I}\right) + \exp\left(\dfrac{s(z_i^a, z_i^b)}{\tau_I}\right)\right\}}, \tag{12}$$

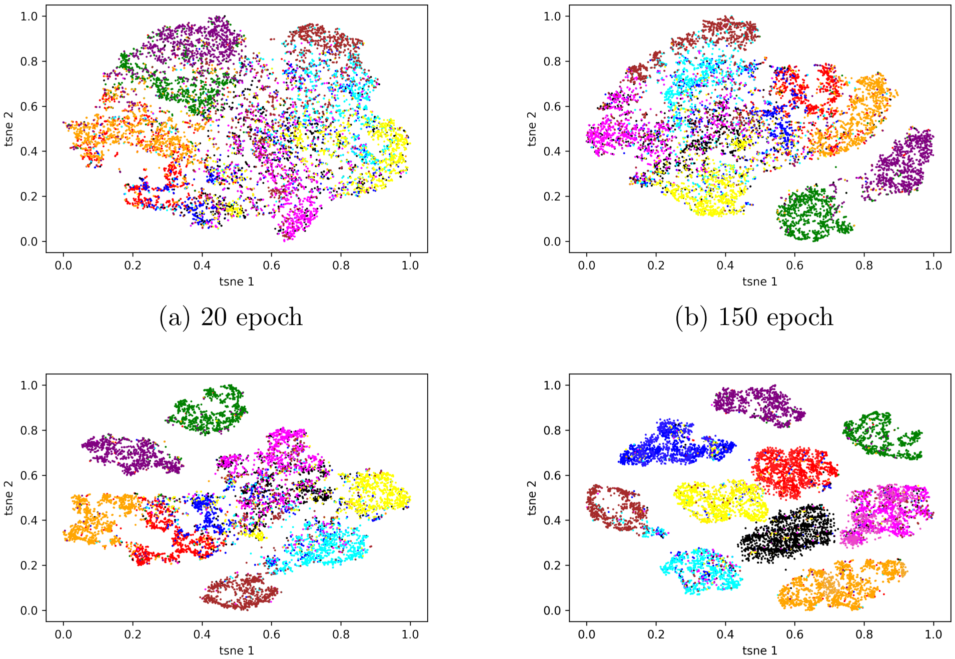
where $s(\cdot,\cdot)$ is the pair-wise cosine distance, and z_i^a and z_i^b are two corresponding rows from the feature matrices $\mathbf{z}^a$ and $\mathbf{z}^b$, respectively. Here, τ_I is the instance-level temperature parameter [48] that is used to control the "softness" of this loss function.

Similarly, the cluster-level representation loss is utilized to distinguish cluster-level representations of positive pairs from the rest as follows

$$\mathcal{L}^a_{C,i} = -\log \frac{\exp\left(\frac{s(p_i^a, p_i^b)}{\tau_c}\right)}{\sum_{j=1}^{K}\left\{\exp\left(\dfrac{s(p_i^a, p_i^a)}{\tau_c}\right) + \exp\left(\dfrac{s(p_i^a, p_i^b)}{\tau_c}\right)\right\}}, \tag{13}$$

where p_i^a and p_i^b are two corresponding columns from the probability matrices $\mathbf{p}^a$ and $\mathbf{p}^b$, respectively, that comes from the second projection head. Here τ_c is the cluster-level temperature parameter. To include every possible positive pair across the dataset, the instance-level contrastive loss, and the cluster-level contrastive loss are as follows:

$$\mathcal{L}_I = \tfrac{1}{2N}\sum_{i=1}^{N}(\mathcal{L}^a_{I,i} + \mathcal{L}^b_{I,i}) \text{ and } \mathcal{L}_C = \tfrac{1}{2K}\sum_{i=1}^{K}(\mathcal{L}^a_{C,i} + \mathcal{L}^b_{C,i}) - S(\mathbf{p}),$$



(a) 20 epoch

(b) 150 epoch

(a) 300 epoch

(b) 700 epoch

Figure 4: The evolution of features and cluster assignments across the training process on ImageNet-10. The colors indicate the cluster assignment obtained from the cluster assignment module and the features for t-SNE are computed from ResNet-34.

where $S(\mathbf{p}) = -\sum_{i=1}^{K}[p_i^a \log p_i^a + p_i^b \log p_i^b]$ is the entropy of cluster assignment probabilities added to prevent assigning all instances within the mini-batch to the same cluster [49]. The functions $\mathcal{L}_{I,i}^b$ and $\mathcal{L}_{C,i}^b$ are defined similarly as in (12) and (13), respectively.

4.3 Objective Function

In our method, the optimization is done in an end-to-end process. The parameters θ of the backbone and the two heads are simultaneously optimized. Thus, the overall objective function consists of (1) the instance-level contrastive loss, (2) the cluster-level contrastive loss, and (3) the two cross-entropy loss functions that enforce the consistency:

$$\mathcal{L}(z,p) = \mathcal{L}_I + \mathcal{L}_C + \mathcal{L}_{Opt}^a + \mathcal{L}_{Opt}^b \tag{14}$$

Our objective enables robust training at both the latent feature and the code assignment levels. In general, we solve two optimization problems: the first is to find the labels and the second is to find the predictions of the model (i.e., the model parameters). We do so by alternating between two steps:

1. Given the current model's parameters θ, we first compute the log probabilities P, then, we find Q using (11).
2. Given the current label assignments Q, we optimize the model parameters θ by minimizing (14). This step is the same as training the model but with a multi-loss function.

4.4 Results

We used the same implementation details as in the previous section. However, here we added a second projection head to output the feature vectors. Following [41] we set the dimension of the latent vector to 128 and the temperatures parameters to 0.5. The same evaluation metrics were used (ACC, NMI, and ARI). Comparisons were done using the baselines.

Table 2: The clustering performance on three image benchmarks. The best results are shown in boldface.

	CIFAR-100			STL-10			ImageNet-10		
	ACC	NMI	ARI	ACC	NMI	ARI	ACC	NMI	ARI
K-means	0.130	0.085	0.028	0.192	0.125	0.061	0.241	0.119	0.057
JULE	0.137	0.103	0.033	0.277	0.182	0.164	0.300	0.175	0.138
VAE	0.152	0.108	0.040	0.282	0.200	0.146	0.334	0.193	0.168
DEC	0.185	0.136	0.050	0.359	0.276	0.186	0.381	0.282	0.203
DAC	0.238	0.185	0.088	0.470	0.366	0.257	0.527	0.394	0.302
DCCM	0.327	0.285	0.173	0.482	0.376	0.262	0.710	0.608	0.555
PICA	0.317	0.310	0.171	0.713	0.611	0.531	0.870	0.802	0.761
CC	0.429	0.431	0.266	0.850	0.764	0.726	0.893	0.859	0.822
OTCC(Ours)	**0.501**	**0.492**	**0.309**	**0.872**	**0.797**	**0.762**	**0.913**	**0.892**	**0.851**
CLOT(Ours)	**0.559**	**0.570**	**0.352**	**0.898**	**0.863**	**0.822**	**0.927**	**0.901**	**0.875**

Again, results shown in Table 2 demonstrate the clustering ability of CLOT, which outperforms the baselines by a large margin on all three datasets. It also outperforms OTCC by a good margin in terms of the three evaluation measures. The results demonstrate that this robustness is a result of combining both contrastive learning and contrasting cluster assignments obtained by solving the labeling problem as an optimal transport. Figure 4 shows the evolution of instance features and cluster assignments across the training process on ImageNet-10. It's clear that CLOT has a higher ability to separate the instances and put them into clusters than OTCC.

Part II
Proposed (future) projects

5 Project I: An optimal transport contrastive learning for clustering tissues from histopathology images

5.1 Motivation

Manual annotation of gigapixel WSIs is a labor-intensive and error-prone task and requires domain knowledge from experts. While seasoned pathologists are busy diagnosing tens of hundreds of slides each day, it is infeasible to ask pathologists to label routine cases for supervised CNN training. To solve the annotation shortage problem, many efforts have been made by researchers to develop annotation-efficient CNN training methods for WSI analysis. Current popular solutions can be weakly supervised, semi-supervised, or self-supervised methods. However, the above methods still need a certain number of manual annotations. Therefore, fully annotation-free WSI analysis methods are yet to be explored. Recent successes of contrastive learning methods [37, 41, 50] in natural image classification have largely advanced the progress of unsupervised visual representation learning and sheds the light on annotation-free WSI analysis. The contrastive learning approaches leverage input data themselves as the supervision via multiple image augmentations to train an encoder for a discriminative visual embedding generation. Many contrastive learning-based approaches have shown effectiveness in WSI analysis. However, these works still rely on annotations in fine-tuning process for downstream tasks and thus fall in the category of semi-supervised methods. In contrast to existing methods, we propose to learn from annotation-free histopathology images. We aim to adopt CLOT, the OT-based framework we developed to cluster patches of the same WSIs based on different tissue types.

5.2 The proposed method

In order to build the association between the patches of a WSI with the corresponding tissue types without manual labels. We need to build a mapping from WSI patches to a discriminative embedding space where different tissues can be distinguished based on semantic distance. To this end, we will use CLOT, a framework based on optimal transport to cluster the patches. To better extract pathology-specific contextual features from WSIs, we plan to make a change on the encoder of CLOT output features at a multi-scale, thus, including low-level and high-level features in the final image embedding. We argue that the local tissue texture (i.e., low-level features) could tell critical information about WSI diagnosis such as cell malignancy. In the end, a global average pooling operation can be employed to average the features obtained at a multi-scale.

6 Project II: Optimal transport approach for capturing cellular topology in whole slide pathology images

6.1 Motivation

Due to the large size of the multi-resolution and multi-gigapixel WSIs, existing methods in computational pathology suffer from the associated computational complexity [51, 52]. This makes training a deep learning network on entire WSIs at full resolution computationally intractable. To tackle the computational complexity issue, the typical approach is to sub-divided the image into small patches, where each patch is processed independently in the neural network [52, 53, 54, 55]. Then, the predicted scores for each patch within a WSI are aggregated by combining their results with one of the aggregating strategies. However, patch-level-based analysis has serious drawbacks. First, patches provide a limited visual context, and determining the optimal resolution and patch size depends of the problem. For example, patches drawn at a high magnification level lead to less contextual and spatial information whereas patches at lower magnification levels may not capture cell-level features [51]. As a result, patch-level machine learning methods cannot capture the overall structure of the tissue in a

WSI. Secondly, in most prediction problems in computational pathology, the available labels are only at the WSI level. Secondly, in most prediction problems in computational pathology, only WSI-level labels are often available. It is non-trivial to assign the association of different patches with a target class, which makes the patch-level approach less applicable.

The cell-graph technique was introduced to learn the structure-function relationship by modeling the geometric structure of the tissue using graph theory [56, 57]. It is based on the assumption that structural and spatial patterns of cell organizations in a tissue are not random but associated with the underlying functional state. As a result, cell-graph constructions have been successfully used to characterize the spatial proximity of histopathologic primitives in tasks [58, 59]. However, those graph-based methods with deep learning classifiers were all trained on a per-patch basis which has limited visual context. Extra patch-based voting methods are necessary to assess the functional state of a given WSI.

In this project, we propose a cell-graph-based model to handle these limitations of existing methods. Instead of extracting small patches from the WSI and doing analysis on a limited visual field for prediction, we introduce a new method based on optimal transport which constructs a graph from the nuclei level to the entire WSI level. A graph convolutional neural network is then used for WSI-level prediction. This method accounts for both cell-level information and contextual information by modeling cellular architecture and interactions in the form of a graph.

6.2 The proposed method

The proposed graph generation from images method first builds a graph representation $G_i = G(x_i)$ of a WSI and then uses a graph convolutional neural network (GCN) to generate slide level predictions. The framework consists three steps, first, we plan to use HoVer-Net to segment nuclei simultaneously [60] and extract nuclear features. Second, we use optimal transport principals to self-match the nuclei and its corresponding features, and to find the connections between the nuclei. Solving the problem as an optimal transport will capture cellular topology of the WSI. Lastly, the graph built on the entire WSI is taken as an input to a GCN to do predictions.

HoVer-Net is a convolutional neural network for simultaneous nuclear segmentation. For a given WSI, it results in a set of N nuclei in conjunction with the type and morphological features of each nucleus. We consider each nucleus as a node, thus we have a vertex set V. We formulate the graph-matching task as an optimal transport problem that leads to correspondences between the graph nodes. Specifically, we want to determine the optimal transportation plan $\mathbf{A}$ represented by a matrix of size $N \times N$ with $N = |V|$ that matches the set of the same nodes V (i.e., self-matching). In other words, each coefficient A_{ij} of $\mathbf{A}$ provides the transfer route for conveying a certain quantity of mass from a node $v_i \in V$ to another one $v_j \in V$. A trivial solution to this problem is the node itself. Therefore, we add a constraint on the diagonal of the transport matrix $\mathbf{A}$ to stay zero. Mathematically, we can model the self-matching problem by solving the following optimization problem:

$$\begin{aligned}
\underset{\mathbf{A}}{\text{minimize}} \quad & \langle \mathbf{C}, \mathbf{A} \rangle_F \\
\text{subject to} \quad & \sum_{j=1}^{N} A_{ij} = q(v_i) \quad \forall i \in \{1, .., N\} \\
& \sum_{i=1}^{N} A_{ij} = p(v_j) \quad \forall j \in \{1, .., N\} \\
& A_{ii} = 0 \quad \forall i \in \{1, .., N\} \\
& A_{ij} \geq 0 \quad \forall (i,j) \in \{1, .., N\} \times \{1, .., N\}
\end{aligned} \tag{15}$$

where $q(v_i)$ and $p(v_j)$ are the masses located at the nodes. Each value C_{ij} represents the coefficient of a cost matrix $\mathbf{C}$, which accounts for the transportation cost to move a portion of mass from v_i to v_j. The values for $q(v_i)$ and $p(v_j)$ are user-specified, however, we set it to 100, which means that there are 100 mass units that existed at each node that needs to be transposed to the rest of the nodes. In terms of biology, this means that the most related nuclei (i.e., the ones that share similar morphological features) share the larger portion of this number. Thus, $\mathbf{A}$ is weighted not a binary matrix. We use the MILP solver as provided by Vielma [61] to get a solution for the resulting optimization problem. The cost matrix $\mathbf{C}$ can be generated by measuring the dissimilarities between morphological features of the nodes (e.g., negative cross-correlation, cosine, or Euclidean distance).

7 Project III: Hierarchical graph representation from digital pathology using optimal transport

7.1 Motivation

In project II, the proposed method represents pathological images by cell graphs. Though a cell graph efficiently encodes the cell microenvironment, cellular morphology, and topology, it cannot extensively capture the tissue microenvironment, i.e., the distribution of tissue regions such as necrosis, stroma, epithelium, etc., thus, it discards tissue distribution information that is vital for appropriate representation of histopathological structures. Similarly, a tissue graph comprising the set of tissue regions cannot depict the cell microenvironment. Cellular or tissue interactions alone are insufficient to fully represent pathological structures. Therefore, an entity-graph representation using a single type of entity set is insufficient to comprehensively describe the tissue composition. Thus, to learn the intrinsic characteristics of cancerous tissue it is necessary to aggregate multilevel structural information.

To address this issue, we use optimal transport theory to construct a multi-level hierarchical entity graph representation consisting of multiple types of entity sets, i.e., cells and tissue regions to encode both cell and tissue microenvironment. The proposed construction method encodes individual entity attributes and intra- and inter-entity relationships to hierarchically describe a histology image.

7.2 The proposed method

Here we detail our proposed methodology for hierarchical graph representation. Given a pre-processed whole slide histology image, we first identify pathologically relevant entities (cells and tissues) and construct a graph representation of the region of interest by incorporating the morphological and topological distribution of the entities. Then, we employ a GCN to map the hierarchical graph to a corresponding category, e.g., cancer subtype.

More specifically, we consider nuclei and tissue regions as entities. Therefore, the Hierarchical graph consists of three components: 1) a low-level cell graph, capturing cell morphology and interactions, 2) a high-level tissue graph, capturing morphology and spatial distribution of tissue regions, and 3) cells-to-tissue hierarchies, encoding the relative spatial distribution of cells with respect to the tissue distribution. The cell-graph and the tissue-graph topology configuration can be constructed using our proposed approach in project III, that's based on optimal transport. However, for the cells-to-tissue hierarchies, we here describe a modified version of our proposed approach in Project III.

To find cell-to-tissue interactions, we match the two graphs by optimal transportation. Let G_c and G_t be the cell and tissue graph computed by solving the optimization problem described in the previous section, respectively. The two graphs are not the same size, cell graph has a much larger number of nodes than the tissue graph. Therefore, we formulate the graph matching task as an OT problem that leads to "many-to-one" correspondences between the graph nodes. OT will find the group of cell nodes that best matches a corresponding tissue. Specifically, we want to determine the optimal transportation plan $\mathbf{A}$ (the adjacency matrix) represented by a matrix of size $K \times N$ with $K = |V_c|$ and $N = |V_t|$, that matches the set of nodes V_c to V_t. In other words, each coefficient A_{ij} of $\mathbf{A}$ provides the transfer route for conveying a certain quantity of mass from a node $v_i^c \in V_c$ to another one $v_j^t \in V_t$. Mathematically, we can model this matching problem by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{A}}{\text{minimize}} \quad & \langle C, \mathbf{A} \rangle_F \\ \text{subject to} \quad & \sum_{j=1}^{N} A_{ij} = q(v_i^c) \quad \forall i \in \{1, .., K\} \\ & \sum_{i=1}^{K} A_{ij} = p(v_j^t) \quad \forall j \in \{1, .., N\} \\ & A_{ij} \geq 0 \quad \forall (i,j) \in \{1, .., K\} \times \{1, .., N\} \end{aligned} \tag{16}$$

where $q(v_i^c)$ and $p(v_j^c)$ are the masses located at the nodes of the cell and tissue graph, respectively. The values for $q(v_i^t)$ and $p(v_j^t)$ are user-specified, however, we set it to 100, which means that there are 100 mass units existed at each cell node that needs to be transposed to the corresponding tissue nodes. In terms of biology, this means that the most related nuclei (i.e. the ones that share similar morphological features with the corresponding tissue) share a larger portion of this number with the corresponding tissue. Thus, **A** is weighted not a binary matrix. Each value C_{ij} represents the coefficient of a cost matrix **C**, which accounts for the transportation cost to move a portion of mass from v_i to v_j.

We use the MILP solver as provided by Vielma [61] to get a solution for the resulting optimization problem. The cost matrix **C** can be generated by measuring the dissimilarities between morphological features of the nodes (e.g., negative cross-correlation, cosine, or Euclidean distance).

7.3 Framework description

The framework consists of three steps, first, nuclei and tissue segmentation step. Second, hierarchical graph topology configuration by optimal transportation. Finally, graph learning using GCN.
For cell graphs, nodes denote cells and encode cell morphology, and edges denote cellular interactions and encode cell topology. It is constructed in three steps: i) nuclei detection, ii) nuclei feature extraction, and iii) topology configuration. We plan to use HoVer-Net to segment nuclei simultaneously [60] and extract nuclear features. Then we mathematically formulate the problem as described above as an optimal transportation to configure the cellular topology of the graph.

On the other hand, a tissue graph depicts a high-level tissue microenvironment, where the nodes and edges denote tissue regions and their interactions, respectively. A tissue graph is constructed by first identifying tissue regions (e.g., epithelium, stroma, lumen, necrosis), followed by encoding the tissue regions, and finally the topology building. to detect and semantically segment tissue regions in histology images, and to extract feature representations of tissue regions, we plan to use the approach used by Mercan et al. [62]. Tissue graph topology is then configured by optimal transport as described in the previous section.

The last stage is graph learning. We first plan to test the power of GCN in handling hierarchical graphs. In case GCN shows limited performance, we might come up with our new graph neural network architecture or we may use an existing architecture from the literature.

8 Project IV: Optimal transport contrastive graph convolutional networks for pathology images

8.1 Motivation

Very few Contrastive learning methods have been proposed for applications on graph data [63, 64, 65]. This is due to the challenges brought by the complex, non-Euclidean structure of graph data, which limits the direct analogizing of traditional augmentation operations on other types of images, video or text data. To the best of our knowledge, no one utilizes contrastive learning on graphs from digital pathology images. Contrastive learning is a self-supervised approach to address unlabeled data, in our case, unlabeled graphs. In this project, we examine the capability of graph contrastive learning and optimal transport for predictions on whole slide histopathology images. Typically, a graph contrastive learning framework includes three main components: a GDA module that generates different views of the given graph data, a GNN-based encoder to compute the representations, and a contrastive learning objective to train the model. Besides the contrastive learning objective, we aim to come up with new optimal transport-based objectives for robust training procedures.

8.2 The proposed method

More specifically, GDA modules will perform graph augmentations to generate different views of the same graph, which encompasses techniques of increasing/generating training data without directly collecting or labeling more data. In graph machine learning, in contrast to regular and Euclidean data such as grids (e.g., images) and sequences (e.g., sentences), the graph structure is encoded by node connectivity, which is non-Euclidean and irregular. Thus, most structured augmentation operations are used frequently in computer vision or and NLP cannot be easily analogized to graph data. However, some rule-based augmentation approaches can be used to generate many views of the same graph, such as edge dropping [66], data interpolation, counterfactual augmentations [67], attribute augmentation [68], subgraph Substituting, feature masking [69], subgraph cropping [70]. Another class of graph augmentation is based on learning approaches, where no learnable parameters are involved during data augmentation (see [70] for more details).

In the second module, the GCN-based encoder is aimed to be used as a backbone to compute the latent space graph representations. Because each graph is augmented twice, GCN outputs representations for each augmented view. Then the cost matrix is obtained using a similarity measure between the two representations. The last module is to apply OT in order to find the best match (lowest cost) between the two representations. The output cost is used in a loss function to update the network parameters. Thus, unsupervised learning is achieved with the help of OT.

9 Datasets and prediction tasks

In this section, we explain the prediction problems aimed to be achieved to validate our proposed methods on real histopathology images. We explain five cancer-related prediction takes. Each proposed project is assigned one or more tasks. We also explain the corresponding datasets we plan to use to accomplish these tasks.

9.1 Detection of Ki67 hot-spots in whole tumor slide images: (projects I, IV)

The high resolution of WSIs opens a wide range of possibilities for addressing challenging image analysis problems, including the identification of tissue-based biomarkers. In this task, we aim to detect proliferating activity patterns in tumor WSIs based on Ki67 immunohistochemistry. Hot spots (HSs) are tumor regions that usually exhibit high proliferating activity. Pathologists need tools that can quantitatively characterize these HS patterns. To respond to this clinical need, we plan to use our proposed clustering methods with the aim of identifying Ki67 HSs in whole tumor slide images.

We aim to conduct our experiments using the AIDPATH breast cancer database [71], which is composed of breast tissue cohorts from four institutions and pathology labs around Europe. The dataset includes: 501 WSI breast cancer specimens stained with Ki67/MIB-1 antibody and counterstained with hematoxylin (blue). It also contains 509 WSI breast cancer specimens stained with HE (violet).

9.2 Prediction of the status of growth factor receptor HER2 and PR (projects II, III)

Here, we consider two prediction problems: prediction of the status of human epidermal growth factor receptor 2 (HER2) and progesterone receptor (PR) expression from WSIs of HE-stained tissue slides of breast cancer. HER2 is a growth-promoting biomarker/protein that helps breast cells grow, divide, and repair themselves and breast cancer cells that over-express HER2 are called HER2-positive. HER2-positive breast cancers grow and spread faster than HER2-negative cancers but are much more likely to respond to treatment with specific drugs. Similarly, PR is a prognostic biomarker for determining survival, drug response, and progression [72, 73]. Our method will be used to predict, thus, determine the HER2 and PR status from histopathology images.

We plan to evaluate the performance of our proposed network on the same HE stained cohort from The Cancer Genome Atlas in breast cancer (TCGA-BRCA) [25]. This dataset has 710 WSIs. Among them, in HER2 status differentiation, there are 608 HER2 negative and 101 HER2 positive images while in PR status differentiation, 452 PR positive and 256 PR negative images are included.

9.3 Survival and death time prediction from whole slide pathological images (projects II, III, IV)

In the medical analysis domain, survival and death analysis aims to predict the time of death, cardiac arrest, or occurrence of a specific disease. Accurate survival and death prediction can help doctors make correct diagnoses with fewer mistakes, thereby improving the treatment quality and quality of life among patients, it also helps them to evaluate the progression of the disease and how critical is the situation of the patient. Most of the existing methods either focused on ranking the death occurrences of patients or predicting survival times, which does not provide sufficient information in practical diagnosis since the prediction of the death order among patients and the survival time for each patient are both essential. In this task, we aim to take advantage of both risk or death prediction and survival time prediction, thereby enabling the prediction of survival times with higher accuracy in the correct sequence.

We plan to conduct our experiments on two datasets, bladder and brain cancer datasets obtained from TCGA [25]: bladder urothelial carcinoma (TCGA-BLCA), and glioblastoma multiforme (TCGA-GBM).

10 Timeline of proposed research

Here we provide our proposed research plan from Summer 2023 to Spring 2025 with specific research activities and targeted conferences/journals that we expect to present the results. Project I is aimed to be done in summer 2023. From fall 2023 to summer 2024, we plan to investigate the ability of OT in capturing the different topological properties from whole slide pathology images, which is corresponding to projects II and III. Project IV is planned to be investigated in the fall of 2024. Time in the last semester (spring 2025) will be spent writing the Ph.D. dissertation and conducting the defense. The following table summarizes two years plan, including (but not limited to) the proposed work:

Timeline	
Semester	**Proposed work**
Summer 2023	1) Evaluate CLOT framework on pathology images, and perform clustering of tissues for detection and quantitative assessment of Ki67 Hot-Spots. 2) Write a journal paper to be submitted to the Scientific Reports
Fall 2023	1) Design an OT-based mathematical optimization framework for constructing cell-level graphs from pathology images. 2) Complete a Python implementation of the framework and obtain results. 3) Write and submit a conference paper to CVPR 2024.
Spring 2024	1) Design an OT-based mathematical optimization framework for constructing hierarchical multi-level graphs from pathology images. 2) Complete a Python implementation of the framework.
Summer 2024	1) Implement a version of a graph convolutional network that handles hierarchical graphs and complete the training and testing. 2) Write and submit a journal paper to the IEEE Journal of Biomedical and Health Informatics.
Fall 2024	1) Design a graph contrastive learning framework based on OT for unlabeled graphs. 2) Implement, train, and test, the proposed framework. 3) Write and submit a paper to the International Journal of Medical Informatics.
Spring 2025	1) Start writing the PhD dissertation. 2) Defend the Ph.D. dissertation at end of the semester.

Table 3: Timeline of the proposed research.

11 Conclusion

In this report, we have shown our achieved work and explored the utility of optimal transport in developing deep-learning methods for digital whole-slide histopathology images. OT has been gaining in recent years increasing attention as a promising and useful tool in the machine-learning community. This success is due to its capacity to exploit the geometric property of the samples at hand.

We presented the results of our first method on natural scene images. Our method "CLOT" benefits from the power of contrastive learning and OT to address the cluster assignment problem and to self-generate the labels. In contrast to other methods, ours optimizes three objectives during feature learning and during clustering, thus providing a robust training setting. Our method which is based on OT outperformed the state-of-the-art methods in terms of three evaluation metrics. This promising performance motivates us to apply our method to real pathology images, for the sake of addressing the time-consuming and costly diagnosis process and assisting pathologists in diagnostic tasks.

Furthermore, we showed in detail our proposed research for the rest of the Ph.D. We illustrated four projects on employing OT in conjunction with graph-deep neural networks. More specifically, in our first project, we aim to use the CLOT framework on pathology images and perform clustering of tissues for detection and quantitative assessment of Ki67 Hot-Spots. In the second project, we plan to design an OT-based mathematical optimization framework for constructing cell-level graphs from pathology images. In the third project, we aim to improve our third method and design an OT-based mathematical optimization framework for constructing hierarchical multi-level graphs from pathology images, that incorporates tissue-level graphs besides cell-level graphs. In the last project, we aim to explore and develop a graph contrastive learning framework based on OT for unlabeled graphs.

Further, we explained the prediction problems aimed to be achieved to validate our proposed methods on real histopathology images. We explained five cancer-related prediction takes. Each proposed project is aimed to be assigned to solve at least one task.

References

[1] Farzad Ghaznavi, Andrew Evans, Anant Madabhushi, and Michael Feldman, "Digital imaging in pathology: whole-slide imaging and beyond," *Annu Rev Pathol*, vol. 8, pp. 331–359, Nov. 2012.

[2] Thomas J Fuchs, Peter J Wild, Holger Moch, and Joachim M Buhmann, "Computational pathology analysis of tissue microarrays predicts survival of renal clear cell carcinoma patients," *Med Image Comput Comput Assist Interv*, vol. 11, no. Pt 2, pp. 1–8, 2008.

[3] Anant Madabhushi and George Lee, "Image analysis and machine learning in digital pathology: Challenges and opportunities," *Med Image Anal*, vol. 33, pp. 170–175, July 2016.

[4] Alexander J J Smits, J Alain Kummer, Peter C de Bruin, Mijke Bol, Jan G van den Tweel, Kees A Seldenrijk, Stefan M Willems, G Johan A Offerhaus, Roel A de Weger, Paul J van Diest, and Aryan Vink, "The estimation of tumor cell percentage for molecular testing by pathologists is not accurate," *Mod Pathol*, vol. 27, no. 2, pp. 168–174, July 2013.

[5] Hollis Viray, Kevin Li, Thomas A Long, Patricia Vasalos, Julia A Bridge, Lawrence J Jennings, Kevin C Halling, Meera Hameed, and David L Rimm, "A prospective, multi-institutional diagnostic trial to determine pathologist accuracy in estimation of percentage of malignant cells," *Arch Pathol Lab Med*, vol. 137, no. 11, pp. 1545–1549, Nov. 2013.

[6] Le Hou, Dimitris Samaras, Tahsin M. Kurç, Yi Gao, James E. Davis, and Joel H. Saltz, "Efficient multiple instance convolutional neural networks for gigapixel resolution image classification," *CoRR*, vol. abs/1504.07947, 2015.

[7] Hamid Reza Tizhoosh and Liron Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities," *J Pathol Inform*, vol. 9, pp. 38, Nov. 2018.

[8] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 9041, 02 2014.

[9] Andrew Janowczyk and Anant Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J Pathol Inform*, vol. 7, pp. 29, July 2016.

[10] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens, "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Trans Med Imaging*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

[11] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[12] Cédric Villani, "Optimal transport: Old and new," 2008.

[13] David Alvarez-Melis, Tommi S. Jaakkola, and Stefanie Jegelka, "Structured optimal transport," 2017.

[14] Guillermo D. Cañas and Lorenzo Rosasco, "Learning probability measures with respect to optimal transport metrics," *CoRR*, vol. abs/1209.1077, 2012.

[15] Gabriel Peyré and Marco Cuturi, "Computational optimal transport," 2020.

[16] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein gan," 2017.

[17] Rémi Flamary, Marco Cuturi, Nicolas Courty, and Alain Rakotomamonjy, "Wasserstein discriminant analysis," *Machine Learning*, vol. 107, no. 12, pp. 1923–1945, may 2018.

[18] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol, "Regularized discrete optimal transport," *CoRR*, vol. abs/1307.5551, 2013.

[19] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu, "Optimal mass transport for shape matching and comparison," *IEEE Trans Pattern Anal Mach Intell*, vol. 37, no. 11, pp. 2246–2259, Nov. 2015.

[20] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, "Optimal transport for domain adaptation," *CoRR*, vol. abs/1507.00504, 2015.

[21] Riccardo Bellazzi, Andrea Codegoni, Stefano Gualandi, Giovanna Nicora, and Eleonora Vercesi, "The gene mover's distance: Single-cell similarity via optimal transport," 2021.

[22] Kai Cao, Yiguang Hong, and Lin Wan, "Manifold alignment for heterogeneous single-cell multi-omics data integration using pamona," *Bioinformatics*, vol. 38, no. 1, pp. 211–219, Dec. 2021.

[23] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh, "Gromov-wasserstein optimal transport to align single-cell multi-omics data," *bioRxiv*, 2020.

[24] Geert-Jan Huizing, Laura Cantini, and Gabriel Peyré, "Unsupervised ground metric learning using wasserstein singular vectors," 2022.

[25] Cancer Genome Atlas Network et al, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012.

[26] Marco Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, Eds. 2013, vol. 26, Curran Associates, Inc.

[27] Sven Erlander and Neil F. Stewart, "The gravity model in transportation analysis - theory and extensions," 1990.

[28] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[29] Longlong Jing and Yingli Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2021.

[30] Pedro O O. Pinheiro, Amjad Almahairi, Ryan Benmalek, Florian Golemo, and Aaron C Courville, "Unsupervised learning of dense visual representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 4489–4500, Curran Associates, Inc.

[31] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, "Deep clustering for unsupervised learning of visual features," *CoRR*, vol. abs/1807.05520, 2018.

[32] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi, "Self-labelling via simultaneous clustering and representation learning," *CoRR*, vol. abs/1911.05371, 2019.

[33] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *CoRR*, vol. abs/2006.09882, 2020.

[34] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," Tech. Rep. 0, University of Toronto, Toronto, Ontario, 2009.

[35] Adam Coates, Andrew Ng, and Honglak Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Geoffrey Gordon, David Dunson, and Miroslav Dudík, Eds., Fort Lauderdale, FL, USA, 11–13 Apr 2011, vol. 15 of *Proceedings of Machine Learning Research*, pp. 215–223, PMLR.

[36] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Deep adaptive image clustering," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5880–5888.

[37] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick, "Momentum contrast for unsupervised visual representation learning," *CoRR*, vol. abs/1911.05722, 2019.

[38] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[39] Ilya Loshchilov and Frank Hutter, "SGDR: stochastic gradient descent with restarts," *CoRR*, vol. abs/1608.03983, 2016.

[40] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng, "Contrastive clustering," *CoRR*, vol. abs/2009.09687, 2020.

[41] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton, "A simple framework for contrastive learning of visual representations," *CoRR*, vol. abs/2002.05709, 2020.

[42] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symposium on Math., Stat., and Prob*, 1967.

[43] Vivek Sharma, Makarand Tapaswi, M. Saquib Sarfraz, and Rainer Stiefelhagen, "Clustering based contrastive learning for improving face representations," *CoRR*, vol. abs/2004.02195, 2020.

[44] Jiabo Huang, Shaogang Gong, and Xiatian Zhu, "Deep semantic clustering by partition confidence maximisation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8846–8855.

[45] Junyuan Xie, Ross B. Girshick, and Ali Farhadi, "Unsupervised deep embedding for clustering analysis," *CoRR*, vol. abs/1511.06335, 2015.

[46] Jianwei Yang, Devi Parikh, and Dhruv Batra, "Joint unsupervised learning of deep representations and image clusters," *CoRR*, vol. abs/1604.03628, 2016.

[47] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," 2013.

[48] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin, "Unsupervised feature learning via non-parametric instance-level discrimination," *CoRR*, vol. abs/1805.01978, 2018.

[49] Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama, "Learning discrete representations via information maximizing self-augmented training," 2017.

[50] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *CoRR*, vol. abs/2006.07733, 2020.

[51] Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi, "A dataset and a technique for generalized nuclear segmentation for computational pathology," *IEEE Transactions on Medical Imaging*, vol. 36, no. 7, pp. 1550–1560, 2017.

[52] Hamid Reza Tizhoosh and Liron Pantanowitz, "Artificial intelligence and digital pathology: Challenges and opportunities," *J Pathol Inform*, vol. 9, pp. 38, Nov. 2018.

[53] Andrew Janowczyk and Anant Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," *J Pathol Inform*, vol. 7, pp. 29, July 2016.

[54] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens, "From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge," *IEEE Trans Med Imaging*, vol. 38, no. 2, pp. 550–560, Feb. 2019.

[55] Angel Cruz-Roa, Ajay Basavanhally, Fabio González, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi, "Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks," *Progress in Biomedical Optics and Imaging - Proceedings of SPIE*, vol. 9041, 02 2014.

[56] Cigdem Gunduz, Bülent Yener, and S Humayun Gultekin, "The cell graphs of cancer," *Bioinformatics*, vol. 20 Suppl 1, pp. i145–51, Aug. 2004.

[57] Cigdem Demir, S Humayun Gultekin, and Bülent Yener, "Augmented cell-graphs for automated cancer diagnosis," *Bioinformatics*, vol. 21 Suppl 2, pp. ii7–12, Sept. 2005.

[58] Cheng Lu, Xiangxue Wang, Prateek Prasanna, Germán Corredor, Geoffrey Sedor, Kaustav Bera, Vamsidhar Velcheti, and Anant Madabhushi, *Feature Driven Local Cell Graph (FeDeG): Predicting Overall Survival in Early Stage Lung Cancer*, pp. 407–416, 09 2018.

[59] Jon Whitney, German Corredor, Andrew Janowczyk, Shridar Ganesan, Scott Doyle, John Tomaszewski, Michael Feldman, Hannah Gilmore, and Anant Madabhushi, "Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer," *BMC Cancer*, vol. 18, no. 1, pp. 610, May 2018.

[60] Simon Graham, Quoc Dang Vu, Shan e Ahmed Raza, Jin Tae Kwak, and Nasir M. Rajpoot, "XY network for nuclear segmentation in multi-tissue histology images," *CoRR*, vol. abs/1812.06499, 2018.

[61] Juan Pablo Vielma, "Mixed integer linear programming formulation techniques," *SIAM Rev.*, vol. 57, pp. 3–57, 2015.

[62] Ezgi Mercan, Sachin Mehta, Jamen Bartlett, Donald Weaver, Joann Elmore, and Linda Shapiro, "Automated diagnosis of breast cancer and pre-invasive lesions on digital whole slide images," 01 2018, pp. 60–68.

[63] Yanqiao Zhu, Yichen Xu, Feng Yu, Qiang Liu, Shu Wu, and Liang Wang, "Deep graph contrastive representation learning," *CoRR*, vol. abs/2006.04131, 2020.

[64] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen, "Graph contrastive learning with augmentations," *CoRR*, vol. abs/2010.13902, 2020.

[65] Yaochen Xie, Zhao Xu, Zhengyang Wang, and Shuiwang Ji, "Self-supervised learning of graph neural networks: A unified review," *CoRR*, vol. abs/2102.10757, 2021.

[66] Yu Rong, Wenbing Huang, Tingyang Xu, and Junzhou Huang, "The truly deep graph convolutional networks for node classification," *CoRR*, vol. abs/1907.10903, 2019.

[67] Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang, "Counterfactual graph learning for link prediction," *CoRR*, vol. abs/2106.02172, 2021.

[68] Songtao Liu, Rex Ying, Hanze Dong, Lanqing Li, Tingyang Xu, Yu Rong, Peilin Zhao, Junzhou Huang, and Dinghao Wu, "Local augmentation for graph neural networks," 2022.

[69] Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L. Dyer, Rémi Munos, Petar Veličković, and Michal Valko, "Large-scale representation learning on graphs via bootstrapping," 2023.

[70] Tong Zhao, Wei Jin, Yozen Liu, Yingheng Wang, Gang Liu, Stephan Günnemann, Neil Shah, and Meng Jiang, "Graph data augmentation for graph machine learning: A survey," 2023.

[71] Database AIDPATH, "Academia and industry collaboration for digital pathology," .

[72] Lisa K Dunnwald, Mary Anne Rossing, and Christopher I Li, "Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients," *Breast Cancer Research*, vol. 9, no. 1, pp. R6, Jan. 2007.

[73] Xiaojiang Cui, Rachel Schiff, Grazia Arpino, C Kent Osborne, and Adrian V Lee, "Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy," *Journal of clinical oncology*, vol. 23, no. 30, pp. 7721–7735, 2005.