SELF-SUPERVISED EXTREME COMPRESSION OF GIGAPIXEL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Whole slide images (WSI) are microscopy images of stained tissue slides routinely prepared for diagnosis and treatment selection in clinical practice. WSI are very large (gigapixel size) and complex (made of up to millions of cells). The current state of the art (SoTA) approach to classify WSI subdivides them into tiles, encodes them by pre-trained networks, and applies Multiple Instance Learning (MIL) to train for specific downstream tasks. However, annotated datasets are often small, typically a few hundred to a few thousand WSI, which may cause overfitting and underperforming models. On the other hand, the number of unannotated WSI is ever increasing, with datasets of tens of thousands (soon to be millions) of images available. Nevertheless, using unannotated WSI is limited due to the challenges of extending self-supervised learning from natural images to WSI. We propose a strategy of slide-level self-supervised learning (SSL) to leverage the large number of images without annotations to infer powerful slide representations. The resulting embeddings allow compression of the whole public WSI dataset available at the Cancer-Genome Atlas (TCGA), one of the most widely used data resources in cancer research, from 16 TB to 23 MB, thus dramatically simplifying future studies in the field of computational pathology in terms of data storage and processing. We show that a linear classifier trained on top of these embeddings maintains or improves previous SoTA performances on various benchmark WSI classification tasks. Finally, we observe that training a classifier on these representations with tiny datasets (e.g. 50 slides) improved performances over SoTA by an average of +6.3 AUC points over all downstream tasks. We further analyze the conditions necessary for such a training framework to succeed, bringing insights into WSI processing.

1 INTRODUCTION

Whole Slide Images (WSI) are microscopy images of diseased tissue. Taken in routine in cancer treatment centers, they are used by clinicians for diagnosis, patient stratification and treatment selection. They contain complex and abundant information on thousands of individual cells, their environments and the overall tissue architecture.

The most important task in computational pathology is to make predictions directly from the WSI. For example, we might want to predict cancer subtype, survival of the patient or response to treatment. The major challenges in building predictive models operating on WSI are:

- prohibitive memory requirements (typically 15GB uncompressed per WSI)
- We do not know a priori which parts of the WSI are relevant for the output variable.
- WSI datasets tend to be small. The large amount of noise thus often leads to overfitting and poor performance.
- Dealing with WSI is technically demanding. This represents a considerable barrier for multi-modal analyses of genomic and pathology data.

Today, the leading methods for WSI classification rely on Multiple Instance Learning (MIL): WSI are tessellated into small images, called tiles, which are encoded by an instance embedder. Tile embedders are usually pre-trained, either on natural images or - more recently and with great effect

- by self-supervised learning (SSL). WSI are then seen as bags of tiles, and the slide representation is thus obtained by combining the tile representations, which are then used as input for the slide classification network. The agglomeration strategy comes in different flavors and usually relies on tile selection or weighted averaging of tile embeddings (Courtiol et al. (2019)Ilse et al. (2018); Lu et al. (2021); Rymarczyk et al. (2020); Li & Eliceiri (2020)). The slide classification network is usually trained from scratch on the specific classification task.

While these methods successfully predict a large variety of output variables, such as grade, cancer subtype, gene signatures, mutations or response to treatment (Campanella et al. (2019); Coudray et al. (2018); Kather et al. (2020); Lazard et al. (2021); Naylor et al. (2022); Echle et al. (2021); Qu et al. (2021)), the performances remain highly dependent on the size of the training dataset (Campanella et al. (2019)). Indeed, MIL performance reaches saturation when using thousands of slides with associated ground truth for training.This might be realistic for the most frequent cancer types and routinely acquired output variables, but we would rather have tens to hundreds of WSI with known ground truth in most real-world projects. On the other hand, with the digitalization of many pathology facilities, we have access to hundreds of thousands of WSI, scanned in clinical routine. Following the SSL paradigm that has been successfully applied at the tile level (Dehaene et al. (2020); Lazard et al. (2021); Ciga et al. (2021); Saillard et al. (2021)), there is a challenging opportunity to make use of these unannotated data at the slide level to derive powerful slide representations. These would be particularly useful for small cohorts and non-standard output variables, such as prognosis for rare cancer types or prediction of treatment response in clinical trials.

Another limitation of current methods is inherent to the MIL approach: slides are modeled as bags of tiles, their spatial relationship is not taken into consideration. However, depending on the classification task, the spatial interconnections between different tiles might be informative.

Here, we propose Giga-SSL, a strategy to perform SSL for Gigapixel images. Designed for pathology data, our method is capable of leveraging large datasets, such as TCGA to learn powerful representations at the WSI level without using any ground truth data. Our contributions are:

- Giga-SSL, an efficient self-supervised contrastive learning framework for Gigapixels images.
- We show that a linear classifier on top of these embeddings maintains or improves SOTA performances on most of the downstream classification tasks we tackled. The performance gains are particularly important for small datasets.
- We provide the WSI embeddings of the whole TCGA dataset, compressing it by a factor of almost 1 million from 16Tb to 23Mb, and thus making complex and large image datasets amenable for future research.

We expect that this method will have an important impact on the field of computational pathology: with the representations provided by Giga-SSL, we can generate predictive models for small datasets and we can make image data accessible to a larger community of researchers in cancer bioinformatics, in order to investigate the complex relationships between genetic, transcriptomic and phenotypic data.

2 BACKGROUND

2.1 MULTIPLE INSTANCE LEARNING FOR GIGAPIXEL IMAGES

Because gigapixel images do not fit in GPU memory with modern deep learning architectures, most of the community has used the multiple instance learning (MIL) paradigm to handle them. In the MIL paradigm, objects (called bags) contain other objects (called instances). For gigapixel images, the object is a gigapixel image its instances are subimages (also called tiles or patches) extracted throughout the gigapixel image. While traditional MIL assumes independent and identically distributed (i.i.d.) instances within each bag. this assumption is relaxed for gigapixel images because instances are extracted from the same image, and are therefore not independent. Given a gigapixel image X made of n_x instances $(x_1, \ldots, x_{n_x})$, MIL is implemented as a combination of three modules i.e. (i) an instance embedder $e_{\theta_1}(\cdot)$, (ii) a pooling operator $p_{\theta_2}(\cdot)$ and (iii) a classifier $c_{\theta_3}(\cdot)$ such that a decision $\hat{y}$ is obtained with

$$\hat{y} = c_{\theta_3}\Big(p_{\theta_2}\big(\{e_{\theta_1}(x_1), \ldots, \{e_{\theta_1}(x_n)\}\big)\Big).$$

Most of the MIL architectures differ from the design of the pooling operator $p_{\theta_2}(\dot{})$. There are two families of operators: (i) those that consider instances as i.i.d. and (ii) those that exploit the relationship between instances of a bag. Architectures that consider instances as i.i.d. are either parameterless (e.g. using the operators average, maximum, a concatenation of both (Lerousseau et al. (2020)), or a noisy-OR function), or trainable such as attention based neural networks (Ilse et al. (2018)). While these architectures obtain good performance, instances of gigapixel images are dependent and contain information that can be leveraged to produce accurate predictions. Modern MIL architecture for gigapixel images have been designed to exploit the spatial relationship of instances. For instance, transformer-based MIL approaches (Shao et al. (2021)) extend the attention mechanism of (Ilse et al. (2018)) by incorporating the positions of instances for decision prediction. Of particular interest in this work, the SparseConvMIL (Lerousseau et al. (2021)) architecture bridges the gap between MIL and convolutional neural networks by building a sparse map from both the embeddings and the spatial locations of the samples instances. This map is further processed by a sparse-input convolutional neural network that outputs a latent vector to be further classified by a generic classifier $c_{\theta_3}(\dot{})$.

2.2 Self-supervised learning in classical image processing

The SSL framework aims at learning useful representations of objects by solving pretext tasks that do not need external annotation. SSL has achieved massive improvement on standard classification benchmarks, most notable with contrastive or similarity learning (Chen et al. (2020a); He et al. (2020)). The foundation of these algorithms is to learn representation of images that are invariant to a set of augmentations. For instance, in SimCLR (Chen et al. (2020a)), images are transformed by two transformations drawn from a distribution designed to change the aspect of the resulting views but not their semantic content. These views are then projected on a lower dimensional space by a neural network. By optimizing the contrastive loss, two views of the same image should become close while distance between views of different images should be increased. The resulting representation is therefore supposed to be robust w.r.t. to the transformations, while being informative about the semantic content. These algorithms have led to a tremendous amount of studies and improvement in image processing tasks. Specifically, models pre-trained with SSL present a strong label-efficiency (Chen et al. (2020b)) when fine-tuned on downstream supervised tasks.

2.3 Self-supervised learning for Gigapixel Images

Self-supervised learning for Gigapixel Images has not yet been largely explored: there are only two articles addressing this idea, both published very recently. The work of (Fashi et al. (2022))proposes self-supervised learning in the context of attention MIL. The slide labels are defined as the organ from which the slide originates. Consequently, this approach instead qualifies as traditional transfer learning than SSL. The proposed method has two drawbacks: the pretext task seems to be too easy to lead to powerful representations. Furthermore, this strategy can only be employed in pan-cancer settings, where data from many cancer types is available. (Chen et al. (2022)) used visual transformers (Dosovitskiy et al. (2020)) (ViT), adapting the DINO (Caron et al. (2021)) self-supervised framework to pathology data. Their strategy exploits the hierarchical nature of WSI by stacking three ViTs. Each of these models are then trained sequentially using the Dino algorithm. When the first network takes as input 16x16 pixels images to learn representations of 256x256 pixel-wide regions, the second takes as input the output of the first network and learns representations of 4096x4096 pixel-wide regions. To classify WSI, they then need to train from scratch a last transformer layer that aggregates the information from the 4096x4096 patches. Results are encouraging, improving SoTA performances on many classification or survival prediction tasks. However, as the last ViT layer cannot be trained with DINO, a fine-tuning step is required. Consequently, this method cannot provide universal slide representations, and our main objective could therefore not be reached with this method.

3 METHODS

3.1 ARCHITECTURE

As illustrated in Figure 1, Giga-SSL comprises the following five components:

- A gigapixel image-level augmentation operator transforms a gigapixel image into two randomly augmented views X_1 and X_2. In this work, we consider the following transformations at the slide level: tile sampling, slide rotations, slide flips, and slide scaling.
- A tile-level augmentation operator transforms each sampled tile. These transformations can be applied either independently to all the tiles (not shared) or the same transformation can be applied to all tiles sampled from the same view (shared). We used popular image augmentation functions, including color jitter, rotations, flips, scaling, blur.
- A tile embedder $e_{\theta_1}(\cdot)$ that concurrently and independently embeds the augmented tiles sampled from X_1 and X_2 into a latent space of dimension F.
- A pooling operator $p_{\theta_2}(\cdot)$ that processes the resulting embeddings of X_1 (resp. X_2) into a single representation vector W_1 (resp. W_2).
- A loss function appropriate for self-supervision, such as a contrastive loss or a similarity loss. This loss aims to pull together the representations of the two transformed views of the same gigapixel image and potentially pull apart view representations originating from different gigapixel images.

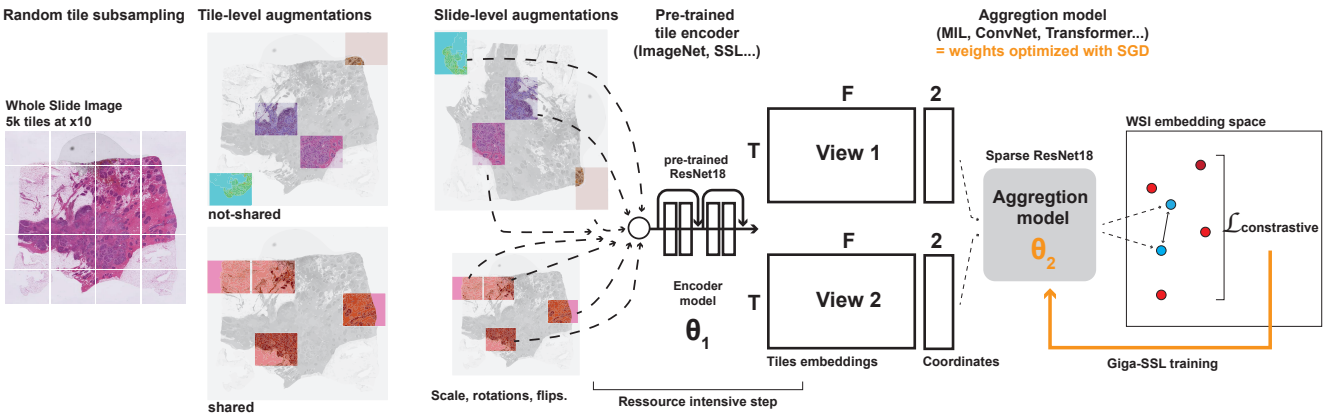

Figure 1: Overview of the method. Presentation of the architecture and different steps of Giga-SSL.

The major computational bottleneck of this strategy is the online computation of the tiles embeddings. A batch size of B_s WSI will be composed of 2 views for each WSI, each WSI itself being composed of T sampled images of size 256x256. A batch therefore contains $B_s \times 2 \times T$ images. Given the limited memory of GPU cards, this puts constraints on B_s and N_t and effectively limits the batch size we can use. However, it has been shown (Chen & He (2020); Chen et al. (2020a; 2021)) that a high batch size is necessary for contrastive learning. Furthermore, a high number of sampled tiles bring better results in WSI classification. Finally, computing tile representations online is time-consuming and therefore leads to very long training times.

We overcome this problem by pre-computing augmented representations of the tiles. If the tile-level augmentations are not shared among views, we perform a random transformation on each tile of the WSI and store their stacked embeddings. If the tile-level augmentations are shared among the same view, for each WSI we sample $N_{aug} = 50$ different augmentations and apply them each to a batch of $N_t = 256$ tiles randomly sampled in the WSI. Embeddings resulting from different transformations are stored in different files. At training time, we just randomly sample an augmentation and then sample T tiles in the corresponding file (see appendix 5). Performing offline finite random augmentations does not degrade the quality of the Giga-SSL representations compared to online and infinite augmentations. It allows at the same time to train a full Giga-SSL model in 10 hours of GPU-time

(which would take a week with online tiles augmentations, or roughly 20 times longer on the same GPU device) as well as experimenting with bigger batch and tile sizes.

At training time, we sample a minibatch of N gigapixel images, and sample two views X_i^1, X_i^2 for each of them such that $X_i = (E_i, L_i)$ with the tile embeddings E_i being a $T \times F$ matrix and the locations L_i a $T \times 2$ matrix. Slide level transformations are applied to L_i and finally the views are forwarded into the pooling operator producing two representation vectors $W_i^1 = p_{\theta_2}(X_i^1)$ and $W_i^2 = p_{\theta_2}(X_i^2)$.

Finally, at inference time, because the final aggregation layer needs at test time the same number of tiles as at training time, we devise an ensembling strategy (appendix figure 4). For a slide X we bootstrap $R = 50$ views without tile augmentation (i.e. differing only in the sampled tiles), compute their embedding W_r and output the average $W_X = \sum_{r=1}^{R} W_r / R$.

3.2 Training dataset

Self-supervised pre-training of Giga-SSL is done using The Cancer Genome Atlas (TCGA) Weinstein et al. (2013), a public dataset that comprises 11754 whole slide images, i.e, digitized images of glass slides containing cancer tissue from virtually all types of solid cancers. These slides are crucial for patient care since they are the basis of diagnosis and treatment selection. On average, images have a width of 93000 pixels and a height 67500 pixels, for an average of 6.5e9 pixels per image. Fully compressed, TCGA weighs more than 16 Terabytes, i.e. 3 orders of magnitude more than ImageNet (Deng et al. (2009)).

4 Experimental validation

4.1 Downstream classification tasks

Datasets Similarly to the works on natural images, we measured the quality of the learned representations by performing linear probing (Chen et al. (2020a); He et al. (2020); Caron et al. (2021),) either with all the labels available for a given task or by artificially reducing the number of labels to simulate a semi-supervised setting. To do so, one representation was extracted for each WSI after SSL pretraining. These representations were then used as input data to train a logistic regression for each task. This protocol was applied for 6 diagnostic tasks highly pertinent for clinical practice: 3 tasks as performed in (Chen et al. (2022)) aiming at automating the diagnostic routine of pathologists (in lung -NSCLC-, breast -BRCA-, and kidney -CRC-), and 3 additional tasks aiming at going beyond human knowledge by inferring molecular properties from tissue slides, which can have a significant positive impact on patient care with reduced diagnostic time, cost and accessibility. Table 1 displays the main statistics for these experiments. These classification tasks are benchmark standards in the field of computational pathology (Kather et al. (2020); Shao et al. (2021); Chen et al. (2022); Schirris et al. (2021)). For each experiment, following the protocol of (Chen et al. (2022)), results were computed on 10 bootstrapped splits of the data.

	BRCA Subtyping	Kidney Subtyping	NSCLC Subtyping	BRCA Molecular Profiling	mHRD-BRCA	tHRD-BRCA
Number of training labels	1041	924	1033	595	912	634
Number of classes	2	3	2	2	2	2
label repartition	831 / 210	510 / 294 / 120	528 / 505	129 / 466	447 / 465	318 / 316

Table 1: Presentation of the 6 downstream WSI classification settings.

Default settings For all the following experiments, except when explicitly said otherwise, we use fixed setting parameters. We selected ResNet18 (He et al. (2015)) up to its third block as the architecture for the tile encoder. To measure the impact of image-level pretraining, we first trained the tile embedder p_{θ_1} using the Momentum Contrast (He et al. (2020)) framework. We set the number T of sampled tiles per slide to 5, and the number R of bootstrapped WSI at inference to 20. In addition, tile augmentations are shared among views. Both Giga-SSL and Average embeddings are evaluated linearly: a logistic regression is used on top of the frozen embeddings extracted respectively through Giga-SSL or averaging the tiles MoCo embeddings over the WSI. Average embeddings

undergo a Standard scale normalization and Giga-SSL embeddings an L2 unit normalization. All Giga-SSL models are trained for 1000 epochs. More details about training parameters are given in the appendix.

4.2 RESULTS

Classification results on benchmarked tasks Table 2 synthesizes the results on all tasks for 5 models i.e. average, an attention-based MIL Ilse et al. (2018) on top of a ResNet18 pretrained with MoCo , DeepSMILE Schirris et al. (2021) and HIPTChen et al. (2022). Results from HIPT and DeepSMILE are taken from their respective articles, and constitute the SOTA on the task on which they are cited. Both Giga-SSL and Average are linear: a logistic regression is used on top of the frozen embeddings extracted respectively through Giga-SSL or averaging the MoCo embeddings of the tiles over the WSI. Chen et al. (2022) made available the train/test split used to benchmark the 3 subtyping tasks when using 100% of the available data. We used these splits when available and created new ones when none were provided. We provide these splits in order to allow future fair benchmarking.

	Method name	Giga-SSL	Average	AttnDeepMIL	HIPT	DeepSMILE
	From	proposed	proposed	Ilse et al. (2018)	Chen et al. (2022)	Schirris et al. (2021)
	linear probing?	✓	✓	X	X	X
Task	% of train set					
NSCLC Subtyping	100	**0.952**	0.913	0.948	**0.952**	-
	25	**0.939**	0.885	0.922	0.92	-
BRCA Subtyping	100	**0.905**	0.859	0.874	0.874	-
	25	**0.890**	0.825	0.860	0.821	-
RCC Subtyping	100	0.982	0.973	**0.986**	0.98	-
	25	**0.975**	0.959	0.97	0.974	-
BRCA Molecular Profiling	100	**0.938**	0.920	0.924	-	-
	25	**0.853**	0.799	0.810	-	-
BRCA mHRD	100	**0.756**	0.706	0.736	-	0.727
	25	**0.743**	0.643	0.660	-	-
BRCA tHRD	100	**0.855**	0.799	0.836	-	0.838
	25	**0.781**	0.698	0.721	-	-

Table 2: Benchmark study reporting the 10-fold cross-validated AUC performances of a logistic regression fed by Giga-SSL features and other WSI classification models. For each task, we evaluate the methods with two data budgets: the training set contains either 100% or 25% of the available data.

Using 100% of the available training labels, the proposed approach achieved SOTA performance on 2 of the 3 tasks benchmarked in (Chen et al. (2022)), i.e. for NSCLC, BRCA subtyping, increasing the AUC by 3 points for BRCA subtyping, and obtained a slightly lower performance for RCC subtyping with AUC of 0.982 vs 0.986 for the attention-based MIL architecture. The proposed approach achieves superior performances for all the other remaining tasks (mHRD, tHRD and BRCA molecular profiling). Notwithstanding these promising results on datasets with more than 500 labels, the power of the proposed approach seems to be in the low data regime, as highlighted by the results obtained by using only 25% of available labels. Indeed, the proposed approach obtained the best results on all tasks in this semi-supervised regime. While this finding may be expected when comparing Giga-SSL to methods without pretraining , Giga-SSL obtained superior results compared to the other SSL-based approach HIPT e.g. with a gain of 6.9 AUC points for BRCA subtyping.

Besides the overall gain of performance of Giga-SSL over attention-based MIL and HIPT, the proposed approach works in a linear regime, while HIPT and attention-based are respectively fine-tuned and learned from scratch. Consequently, their training pipeline is extremely intensive compared to the proposed approach. For instance, training for BRCA subtyping with 100% of the training data on 10 bootstrapped splits took 1.25 CPU-seconds for the proposed approach versus 150 GPU-minutes for attention-based MIL, i.e. 7200 times less time for the proposed approach – while obtaining superior performances.

Training a linear model on the learned representations reduces the risk of overfitting, which is particularly critical for small dataset regimes. We think that this explains the favourable comparison of the Giga-SSL approach w.r.t. Sota approaches. We also observe that the results obtained from

the Average model encodings (also in combination with logistic regression) do not show the same favourable tendency, which is a strong indication of the power of the Giga-SSL encodings. Our proposed strategy, therefore, avoids both overfitting and underfitting.

Tiny datasets In practice, pathological datasets can be tiny for treatment response prediction datasets. For instance, phase II clinical trials typically involve 50 patients. A significant proportion of drugs are discarded during phase II trials since the proportion of responding patients is not judged enough for a subsequent phase III trial. Training a model to better select responding patients is therefore challenging due to the low number of labels.

We challenged the Giga-SSL method in training settings with extremely low data budgets (from 250 slides down to 50) by randomly subsampling the training set of each classification dataset. We compare Giga-SSL to the DeepAttnMIL model, which performances are on par with SOTA algorithms (Table 2).

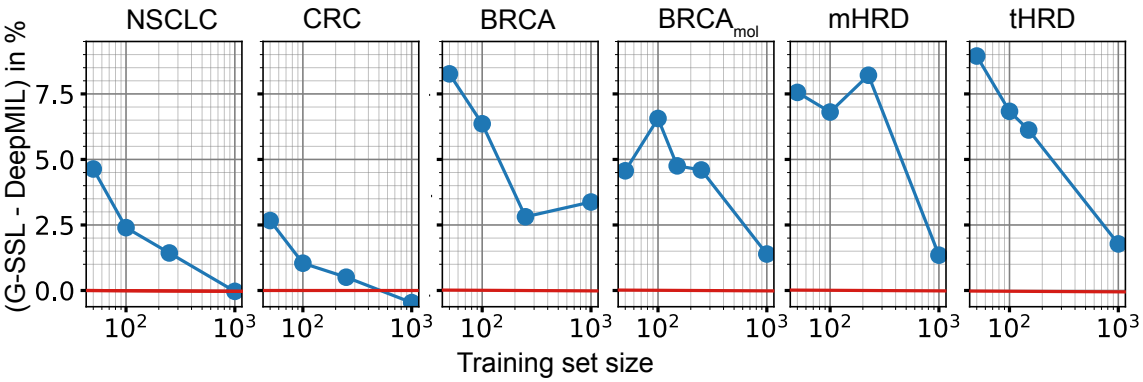

Figure 2: Difference between the average AUC performances of Giga-SSL and AttnDeepMIL (in %) as a function of the training set size. The red line represents equal performance. Above the red line, the advantage is given to Giga-SSL

Figure 2 shows that the performance gap between the proposed approach and the standard WSI classification method strengthens as the number of samples shrinks. The average improvement over all tasks brought by Giga-SSL features is of 5.1 AUC points when using 100 WSI and up to 6.3 AUC points when using only 50 WSI.

5 Ablation Study and Sensitivity Analyses

To better understand the success conditions of Giga-SSL, we led extensive experiments. In this section, we aim to understand the impact of some of Giga-SSL design choices over the predictive power of the learned representations. For all subsequent experiments, the models were trained with the same conditions (including hyperparameters, epochs, and training dataset) as in the previous experiments, except when explicitly mentioned.

Sharing tile augmentations within views improves performance Table 3 reports the performance of Giga-SSL when removing one component at a time, i.e. (i) with a tile embedder pre-trained on ImageNet rather than pre-trained with MoCo on histopathological data (Giga-SSL$_{im}$), (ii) without slide-level augmentation during the WSI-level SSL pretraining including rotations, flips and scaling; (iii) without shared augmentations across all tiles of a view, i.e. each tile is transformed by a randomly and independently sampled augmentation.

Using a tile-level SSL algorithm to pretrain the tile encoder e_{θ_1} brings improvement to the WSI-level representations: the Giga-SSL trained with MoCo features outperforms its ImageNet (Giga-SSL$_{im}$) counter part on all tasks.

The slide-level augmentation, on the contrary, does not seem to be extremely important for the SSL task, as removing it has a small to no impact on performances.

	100% data			50 WSI		
	NSCLC	CRC	BRCA	NSCLC	CRC	BRCA
Giga-SSL	0.95	0.98	0.906	0.892	0.960	0.793
Giga-SSL - no slide aug.	0.935	0.973	0.894	0.86	0.951	0.80
Giga-SSL - aug. tiles not shared	0.933	0.971	0.875	0.847	0.939	0.774
Giga-SSL$_{im}$	0.922	0.978	0.888	0.813	0.952	0.751
Giga-SSL$_{im}$ - aug. tiles not shared	0.897	0.975	0.853	0.777	0.935	0.707

Table 3: 10-fold cross-validated AUC performances of ablated Giga-SSL models. Giga-SSL$_{im}$ stands for a Giga-SSL model trained with tiles embeddings transfered from an ImageNet pretraining.

However, applying independent transformations to each tile (*not shared*) degrades substantially the performances with an average decrease of 1.9 AUC points using 100% of the data down to 2.8 AUC points when using only 50 WSI. When ablating the shared transformations from a Giga-SSL model trained with ImageNet tiles features, the drop of performances compared to a Giga-SSL$_{im}$ is even more important: 2.1 AUC points with 100% of the data, 3.2 AUC points with 50 WSI.

Using shared augmentation thus allows the learning of useful features in abundant and scarce data regimes. We hypothesize key features linked to the slide preparation and shared by all the tiles on the slide are still available for shortcut learning if the tile-level augmentations are not shared. It seems that these shortcut features may be more present in ImageNet than in MoCo. Highlighting such features and finding even more stringent ways to hide them when learning Giga-SSL would improve even more its performances.

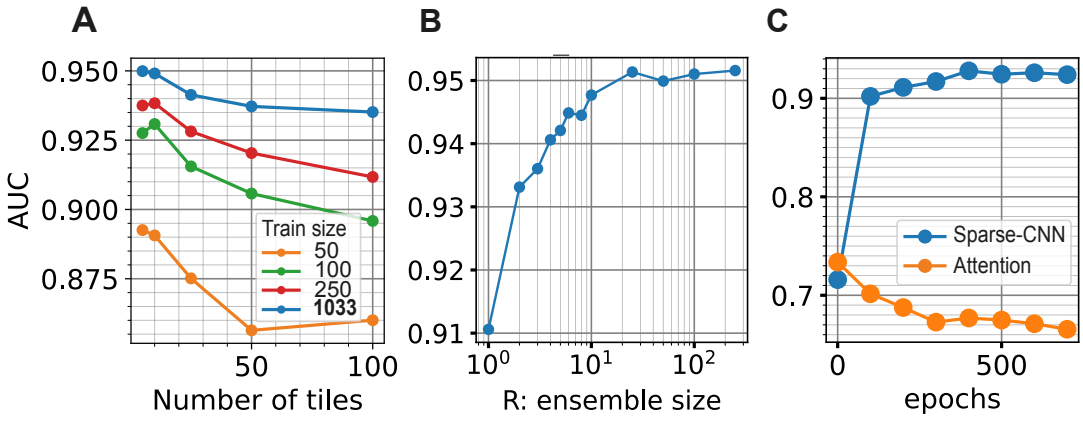

Figure 3: Experiments on key parameters of Giga-SSL. Each point is a 10-fold cross-validated AUC performance of a logistic regression fed with Giga-SSL features. The classification task is NSCLC subtyping for the three experiments. **A**. Effect of the number of sampled tiles T per WSI during training. **B.** Effect of the number R of bootstrapped non-augmented views of WSI to feed Giga-SSL at inference time (see appendix figure 4). **C.** Evolution of the performances of a Giga-SSL with a sparse CNN (blue line, normal situation) or an attention-MIL network (orange line) as an aggregator.

The fewer tiles, the better Figure 3.B presents the performances of 5 G-SSL models trained with different numbers of sampled tiles per view. The fewer tiles we sample, the better the resulting WSI representations. This behaviour strengthens when the downstream problem has a smaller training set and is comparable among all the downstream classification tasks. Interestingly, we can observe the opposite effect when using a DeepAttnMIL to classify a WSI: the fewer tiles used at training time, the worse the performances. A very small number T of sampled tiles per view when training Giga-SSL can be seen as an aggressive augmentation. It has been reported (Chen et al. (2020a)) that SSL benefits from stronger augmentations more than classification tasks, and (Tian et al. (2020)) showed that there is an optimal strength of augmentation for each downstream task. This optimum trading-off between keeping enough information to solve the downstream task and minimizing irrelevant features.

As sampling 5 tiles per WSI is enough to learn useful information to solve all the proposed downstream tasks, we can deduce that the signal relative to these problems is distributed among most of the tiles of the WSI. It would be very interesting to test the performances of Giga-SSL on a classification task for which we know that the signal is highly concentrated on a few instances.

Ensembling representations brings improvement We show in appendix (6) that a Giga-SSL model with a sparse-CNN aggregation module must use the same number of tiles-per-WSI at inference and training. We therefore decided to bootstrap R views of a WSI at inference time before averaging the Giga-SSL embeddings of these R views. Figure 3.B investigates the effect of R on the downstream performances of the Giga-SSL representations. It shows that without this ensembling strategy, Giga-SSL lose up to 4 AUC points on NSCLC subtyping. The gain in performance saturates around $R = 50$.

Attention-deep-MIL unlearns when trained with SSL Instead of using a sparse-CNN as a tiles features aggregator, one could choose any other MIL model. We trained a Giga-SSL model with a AttnDeepMIL aggregation module and evaluated its downstream linear performances on the NSCLC dataset. Figure 3.C shows that the performances of such a model decrease while the SSL training is in progress. While the AttnDeepMIL shows SOTA classification performances 2 when trained from scratch, this architecture seems not suitable for Giga-SSL pretraining. We suspect that the DeepAttnMIL has too easily access to shortcuts features to learn the WSI identity. Understanding what causes its collapse may highlight key pitfall for Giga-SSL training and therefore allow to improve it.

6 Conclusion

Limitations Our work shares part of the limitations of the HIPT model from (Chen et al. (2022)). Namely, the Giga-SSL models have been trained on the whole TCGA dataset, and downstream classification tasks are also taken from the TCGA. In the future, we will validate these results on in-house datasets. Also, an obvious drawback of working with frozen embeddings of WSI removes any possibility of building explainable models.

To conclude We have explored self-supervised learning for whole slide images with a versatile design based on specific data augmentation tailored for the multiple instance learning framework. Our proposed approach achieves or beats state-of-the-art performance over a wide range of clinically impactful tasks in both high and low data regimes. Ablation study and sensitivity analyses highlighted the key components of our approach – including tile encoder pretraining and how to apply augmentations to tiles – to better understand the pitfalls of self-supervised whole slide image representation learning.
The public release of the learned representations for all diagnostic slides of The Cancer Genome Atlas in a manageable size has the potential to unlock new knowledge about cancer and to develop new tools for diagnosis assistance and treatment response prediction towards improved patient survival.

References

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, August 2019. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-019-0508-1. URL http://www.nature.com/articles/s41591-019-0508-1.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers, May 2021. URL http://arxiv.org/abs/2104.14294. arXiv:2104.14294 [cs].

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning, June 2022. URL http://arxiv.org/abs/2206.02647. arXiv:2206.02647 [cs].

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv:2002.05709 [cs, stat]*, February 2020a. URL http://arxiv.org/abs/2002.05709. arXiv: 2002.05709.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big Self-Supervised Models are Strong Semi-Supervised Learners. *arXiv:2006.10029 [cs, stat]*, October 2020b. URL http://arxiv.org/abs/2006.10029. arXiv: 2006.10029.

Ting Chen, Calvin Luo, and Lala Li. Intriguing Properties of Contrastive Losses. *arXiv:2011.02803 [cs, stat]*, October 2021. URL http://arxiv.org/abs/2011.02803. arXiv: 2011.02803.

Xinlei Chen and Kaiming He. Exploring Simple Siamese Representation Learning. *arXiv:2011.10566 [cs]*, November 2020. URL http://arxiv.org/abs/2011.10566. arXiv: 2011.10566.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive Learning. *arXiv:2003.04297 [cs]*, March 2020c. URL http://arxiv.org/abs/2003.04297. arXiv: 2003.04297.

Ozan Ciga, Tony Xu, and Anne L. Martel. Self supervised contrastive learning for digital histopathology. *arXiv:2011.13971 [cs, eess]*, September 2021. URL http://arxiv.org/abs/2011.13971. arXiv: 2011.13971.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, October 2018. ISSN 1546-170X. doi: 10.1038/s41591-018-0177-5. URL https://www.nature.com/articles/s41591-018-0177-5.

Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, Olivier Elemento, Andrew G. Nicholson, Jean-Yves Blay, Françoise Galateau-Sallé, Gilles Wainrib, and Thomas Clozel. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, October 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0583-3. URL http://www.nature.com/articles/s41591-019-0583-3. Number: 10 Publisher: Nature Publishing Group.

Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology. *arXiv:2012.03583 [cs, eess]*, December 2020. URL http://arxiv.org/abs/2012.03583. arXiv: 2012.03583.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

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:2010.11929 [cs]*, October 2020. URL http://arxiv.org/abs/2010.11929. arXiv: 2010.11929.

Amelie Echle, Narmin Ghaffari Laleh, Peter L. Schrammen, Nicholas P. West, Christian Trautwein, Titus J. Brinker, Stephen B. Gruber, Roman D. Buelow, Peter Boor, Heike I. Grabsch, Philip Quirke, and Jakob N. Kather. Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: A systematic literature review. *ImmunoInformatics*, 3-4: 100008, December 2021. ISSN 2667-1190. doi: 10.1016/j.immuno.2021.100008. URL https://www.sciencedirect.com/science/article/pii/S2667119021000082.

Parsa Ashrafi Fashi, Sobhan Hemati, Morteza Babaie, Ricardo Gonzalez, and H. R. Tizhoosh. A self-supervised contrastive learning approach for whole slide image representation in digital pathology. *Journal of Pathology Informatics*, pp. 100133, August 2022. ISSN 2153-3539. doi: 10.1016/j.jpi.2022.100133. URL https://www.sciencedirect.com/science/article/pii/S2153353922007271.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv: 1512.03385.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning, March 2020. URL http://arxiv.org/abs/1911.05722. arXiv:1911.05722 [cs].

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. *arXiv:1802.04712 [cs, stat]*, June 2018. URL http://arxiv.org/abs/1802.04712. arXiv: 1802.04712.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, February 2015. URL http://arxiv.org/abs/1502.03167. arXiv: 1502.03167.

Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M. Niehues, Kai A. J. Sommer, Peter Bankhead, Loes F. S. Kooreman, Jefree J. Schulte, Nicole A. Cipriani, Roman D. Buelow, Peter Boor, Nadina Ortiz-Brüchle, Andrew M. Hanby, Valerie Speirs, Sara Kochanny, Akash Patnaik, Andrew Srisuwananukorn, Hermann Brenner, Michael Hoffmeister, Piet A. van den Brandt, Dirk Jäger, Christian Trautwein, Alexander T. Pearson, and Tom Luedde. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer*, 1(8):789–799, August 2020. ISSN 2662-1347. doi: 10.1038/s43018-020-0087-6. URL http://www.nature.com/articles/s43018-020-0087-6. Number: 8 Publisher: Nature Publishing Group.

Tristan Lazard, Guillaume Bataillon, Peter Naylor, Tatiana Popova, François-Clément Bidard, Dominique Stoppa-Lyonnet, Marc-Henri Stern, Etienne Decencière, Thomas Walter, and Anne Vincent Salomon. Deep Learning identifies new morphological patterns of Homologous Recombination Deficiency in luminal breast cancers from whole slide images. Technical report, September 2021. URL https://www.biorxiv.org/content/10.1101/2021.09.10.459734v1. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

Marvin Lerousseau, Eric Deutsh, and Nikos Paragios. Multimodal brain tumor classification, October 2020. URL http://arxiv.org/abs/2009.01592. arXiv:2009.01592 [cs, eess].

Marvin Lerousseau, Maria Vakalopoulou, Eric Deutsch, and Nikos Paragios. SparseConvMIL: Sparse Convolutional Context-Aware Multiple Instance Learning for Whole Slide Image Classification, August 2021. URL http://arxiv.org/abs/2105.02726. arXiv:2105.02726 [cs].

Bin Li and Kevin W. Eliceiri. Dual-stream Maximum Self-attention Multi-instance Learning. *arXiv:2006.05538 [cs]*, June 2020. URL http://arxiv.org/abs/2006.05538. arXiv: 2006.05538.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, pp. 1–16, March 2021. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL http://www.nature.com/articles/s41551-020-00682-w. Publisher: Nature Publishing Group.

Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1107–1110, Boston, MA, USA, June 2009. IEEE. ISBN 978-1-4244-3931-7.

doi: 10.1109/ISBI.2009.5193250. URL http://ieeexplore.ieee.org/document/5193250/.

Peter Naylor, Tristan Lazard, Guillaume Bataillon, Marick Lae, Anne Vincent-Salomon, Anne-Sophie Hamy, Fabien Reyal, and Thomas Walter. Neural network for the prediction of treatment response in Triple Negative Breast Cancer *, January 2022. URL https://www.biorxiv.org/content/10.1101/2022.01.31.478433v1. Pages: 2022.01.31.478433 Section: New Results.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, pp. 6.

Hui Qu, Mu Zhou, Zhennan Yan, He Wang, Vinod K. Rustgi, Shaoting Zhang, Olivier Gevaert, and Dimitris N. Metaxas. Genetic mutation and biological pathway prediction based on whole slide images in breast carcinoma using deep learning. *npj Precision Oncology*, 5(1):87, December 2021. ISSN 2397-768X. doi: 10.1038/s41698-021-00225-9. URL https://www.nature.com/articles/s41698-021-00225-9.

Dawid Rymarczyk, Jacek Tabor, and Bartosz Zieliński. Kernel Self-Attention in Deep Multiple Instance Learning. *arXiv:2005.12991 [cs, stat]*, May 2020. URL http://arxiv.org/abs/2005.12991. arXiv: 2005.12991.

Charlie Saillard, Olivier Dehaene, Tanguy Marchand, Olivier Moindrot, Aurélie Kamoun, Benoit Schmauch, and Simon Jegou. Self supervised learning improves dMMR/MSI detection from histology slides across multiple cancers. *arXiv:2109.05819 [cs, eess]*, September 2021. URL http://arxiv.org/abs/2109.05819. arXiv: 2109.05819.

Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. DeepSMILE: Self-supervised heterogeneity-aware multiple instance learning for DNA damage response defect classification directly from H&E whole-slide images, July 2021. URL http://arxiv.org/abs/2107.09405. arXiv:2107.09405 [cs, eess].

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification, October 2021. URL http://arxiv.org/abs/2106.00908. arXiv:2106.00908 [cs].

Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What Makes for Good Views for Contrastive Learning? *arXiv:2005.10243 [cs]*, December 2020. URL http://arxiv.org/abs/2005.10243. arXiv: 2005.10243.

John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna M. Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nature genetics*, 45(10):1113–1120, October 2013. ISSN 1061-4036. doi: 10.1038/ng.2764. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3919969/.

A Appendix

A.1 Learning and architectural parameters.

Tile-embedder pre-training We used the official MoCo repository with default parameters. Augmentations used are the augmentations detailed in MoCoV2 (Chen et al. (2020c)). We used a ResNet18 as encoder. At inference, tiles features are averages of the activation map before, after the third residual block, they result in embeddings of dimensions 256. During training as well as during inference of tiles embeddings, the tiles are normalized with the Macenko (Macenko et al. (2009)) normalization. The model was trained on a subset of 6 million tiles extracted from a random set of 3000 slides from the TCGA at a magnification of 10x for 200 epochs, on 4 Nvidia P100 during 2 weeks.

Linear classification The Average and Giga-SSL embeddings are classified both with a logistic regression, implemented with Scikit-learn, as well as the pre-processing applied to WSI representations -StandardScaler and Normalizer- (Pedregosa et al.).

DeepAttnMIL architecture We use the architecture defined in **?**, adding a classification head after the pooling operation. This classification head contains 3 layers, each layer being a fully-connected layer with 128 neurons, a batch-normalisation Ioffe & Szegedy (2015) layer and a dropout layer.

Logistic Regression is set with $C = 10$, and class weight=balanced.

A.2 Ensembling at inference

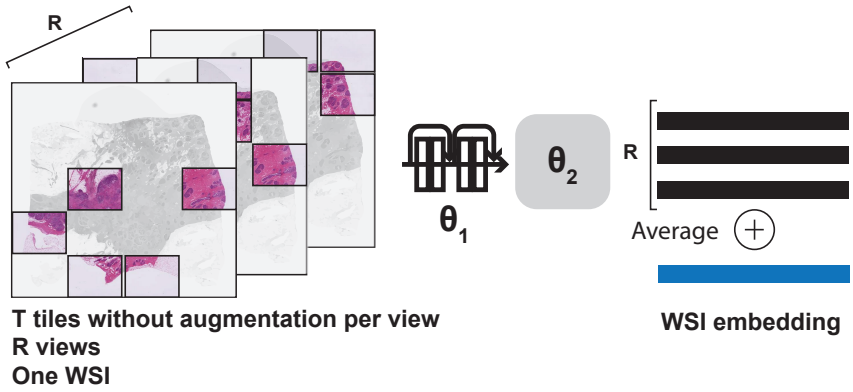

Figure 4: Illustration of the ensembling strategy at inference, to tackle the fact that the sparse-CNN cannot take more tokens at training and inference.

A.3 Pre-computation of a dataset with shared augmentation

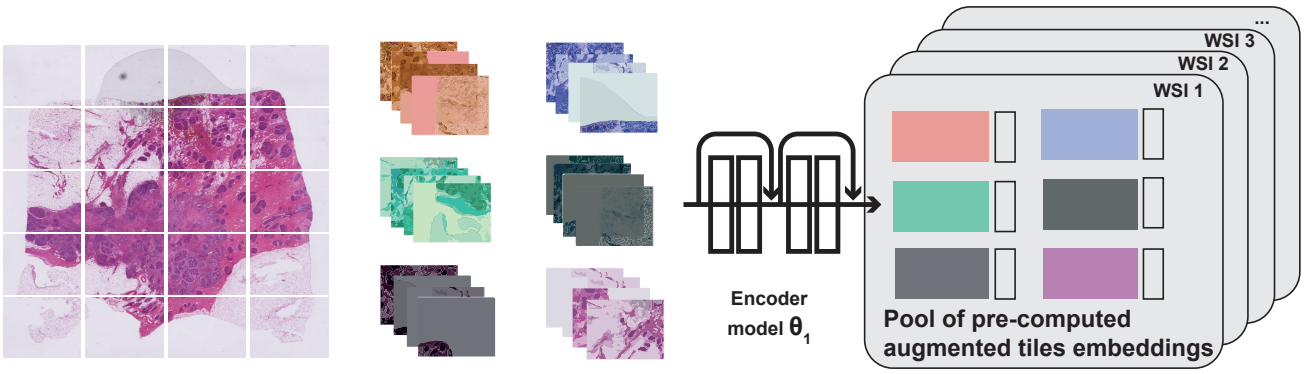

Figure 5: Visual explanation of the constitution of an embedded dataset with shared augmentation for the Giga-SSL training. For each WSI, 50 random augmentations are sampled. We augment with each of these augmentation 256 tiles, encode them, and store them in an augmentation subfolder. This discrete approximation of online shared augmentation is enough to train Giga-SSL and reduce by a factor of 20 the training computation time.

A.4 EFFECTS OF THE NTILES PARAMETER

We study here the impact of the number of sampled tiles at inference. The same Giga-SSL model has been tested, trained with the default settings and therefore with ntiles = 5, and stopped at epoch = 1000. At inference time, we investigate the impact of increasing this parameter to take into account a bigger context of the WSI.

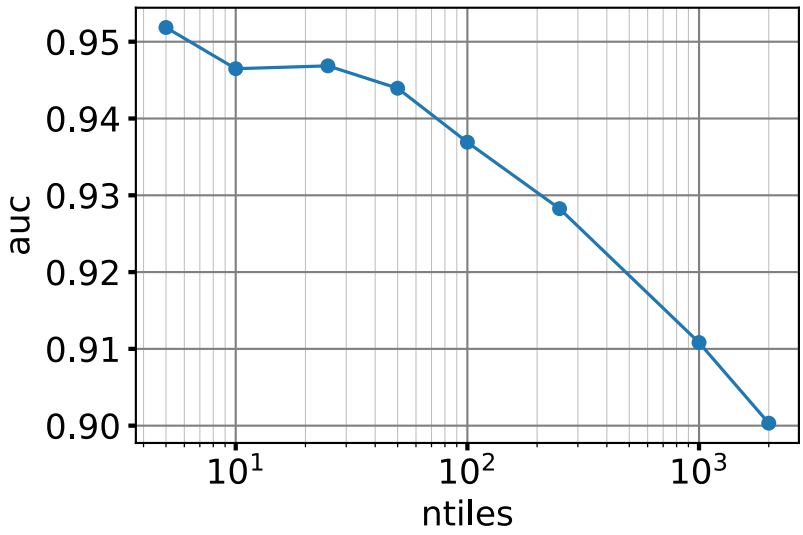

Figure 6: Linear performances of embeddings from the same Giga-SSL model, used at inference with different T sampled tiles per WSI. Results are 10-fold cross-validated.