Incorporating probabilistic domain knowledge into deep multiple instance learning

Ghadi S. Al Hajj [1] **Aliaksandr Hubin** [2 3] **Chakravarthi Kanduri** [1] **Milena Pavlovic** [1] **Knut Rand** [1]
Michael Widrich [4] **Anne Solberg** [1] **Victor Greiff** [5] **Johan Pensar** [3] **Günter Klambauer** [6] **Geir Kjetil Sandve** [1]

Abstract

Deep learning methods, including deep multiple instance learning methods, have been criticized for their limited ability to incorporate domain knowledge. A reason that knowledge incorporation is challenging in deep learning is that the models usually lack a mapping between their model components and the entities of the domain, making it a non-trivial task to incorporate probabilistic prior information. In this work, we show that such a mapping between domain entities and model components can be defined for a multiple instance learning setting and propose a framework DeeMILIP that encompasses multiple strategies to exploit this mapping for prior knowledge incorporation. We motivate and formalize these strategies from a probabilistic perspective. Experiments on an immune-based diagnostics case show that our proposed strategies allow to learn generalizable models even in settings with weak signals, limited dataset size, and limited compute.

1. Introduction

Multiple instance learning (MIL) (Dietterich et al., 1997; Maron & Lozano-Pérez, 1997; Carbonneau et al., 2018) provides a very useful framework for approaching important problems in a variety of domains such as medical image analysis (Yao et al., 2020), video analysis (Quellec et al., 2017), image segmentation (Kraus et al., 2016) and immune repertoire classification (Widrich et al., 2020). The well-defined problem structure of MIL also makes it well-suited to define modularized deep learning (DL) models (Ilse et al., 2018). However, deep learning methods, including deep multiple instance learning methods, have been criticized for their limited ability to incorporate domain knowledge (Marcus, 2018), which could improve model generalization and data efficiency.

Several aspects of modern DL can be seen as incorporating general assumptions about the world like temporal dependency (Hochreiter & Schmidhuber, 1997) or translational invariance (LeCun et al., 1989; Kayhan & Gemert, 2020) informing particular model architectures, transformational invariance inspiring data augmentation techniques (Shorten & Khoshgoftaar, 2019), and assumptions on shared representational spaces inspiring multi-task learning formulations (Caruana, 1997). Moreover, geometric deep learning is used to incorporate geometric properties into the models by leveraging properties inherent in graph-structured data in order to learn to capture and utilize the spatial relationships and connectivity within the data (Bronstein et al., 2017; Cao et al., 2020), while physics-informed models embed physical principles into DL models (Karniadakis et al., 2021; Cuomo et al., 2022). The knowledge entailed by these approaches is, however, usually highly generic, and its incorporation is conceptual. There is markedly less work on incorporating specific, quantitative knowledge from a domain into DL. For the MIL setting, the few relevant examples include the use of image annotation at low level to improve the classification performance at the image level, either by pre-annotating images (Yan et al., 2017; Li et al., 2018; Seung Yeon Shin et al., 2019) or annotating as part of an active learning approach (Melendez et al., 2016).

A primary reason why it is generally non-trivial to incorporate quantitative domain knowledge into a DL model is that the learned representational transformations of a neural network lead to parameter values and DL layer outputs that typically lack any *a priori* association with tangible entities or high-level abstractions from the application domain. Although these values collectively contribute to constructing the desired mapping function for the DL model, and

[1] Department of Informatics, University of Oslo, Oslo, Norway [2] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway [3] Department of Mathematics, University of Oslo, Oslo, Norway [4] Freenome, San Francisco, CA, USA [5] Department of Immunology, University of Oslo, Oslo, Norway [6] LIT AI Lab Institute for Machine Learning, Johannes Kepler University, Vienna, Austria. Correspondence to: Ghadi S. Al Hajj <ghadia@uio.no>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

although the neurons of inner layers are often assumed to be learning meaningful semantic representations, there is, in general, still no clear way to connect values of individual parameters or neuron outputs to entities in the problem domain.

We here show that the structured nature of MIL permits modularized DL models to have a direct mapping between model components and domain entities. We exploit this mapping to propose several strategies for incorporating probabilistic domain knowledge into a DL MIL model. We explore how domain knowledge incorporation affects model performance, including a comparison of explicit knowledge injection in the form of a prior probability to that of a multi-task learning formulation.

We apply our proposed strategies to a case of predicting disease state based on a large set of protein sequences of adaptive immune cells (immune receptors) of a person, where only a small proportion of the sequences are relevant for the disease state (Emerson et al., 2017). Emerging technology in the field allows to obtain imperfect indications on disease-relevance of individual cells (Ma et al., 2021; Setliff et al., 2019), which can serve as a source for probabilistic domain knowledge. We investigate a variety of ways to incorporate such knowledge, considering the ability to learn a generalizable model from weak signals or a few training examples using less compute.

Our contributions are:

- We demonstrate that the structure of MIL problems permits the integration of probabilistic domain knowledge directly into DL models.
- We propose and investigate strategies of knowledge incorporation through either prior injection or multi-level, multi-task learning formulations.
- On a set of large-scale experiments on semi-synthetic data, we show that for the challenging MIL problem of immune-based diagnostics, knowledge incorporation allows us to learn generalizable models at a twenty-fold lower witness rate or a thirty-fold smaller dataset size.

Structuring of this work. In Section 2.1, we provide a formal definition of MIL. Section 2.2 reviews the existing literature on deep learning-based MIL approaches. We conclude Section 2 with a discussion on methodologies that integrate supplementary knowledge through multi-task learning, detailed in Section 2.3. In Section 3, we explore prior research focused on incorporating instance-level information into MIL-based frameworks.

2. Background and notation

2.1. Multiple Instance Learning (MIL)

In MIL, input bags $\mathcal{X}^b = \{x_1^b, ..., x_{n_b}^b\}, b \in \{1, ..., \mathcal{B}\}$ consist of K dimensional instances $x_i^b \in \mathbb{R}^K, i \in \{1, ..., n_b\}$. We also observe labels $\mathcal{Y}^b$ for each bag, with unknown latent labels $y_i^b, i \in \{1, ..., n_b\}$ for each instance within each bag. According to the standard MIL assumption, a bag is considered positive if it has at least one instance with a positive label, referred to as a witness. The fraction of positive instances in a bag is called the witness rate (WR). In our case, we assume that the bag level label is binary, i.e., $\mathcal{Y}^b \in \{0, 1\}$ corresponding to healthy vs diseased individuals, respectively, and that $y_i^b \in \{0, 1\}$ is also binary corresponding to disease-irrelevant vs disease-specific sequences. This setup is a case of weakly annotated data where the classifier is typically trained only using the available coarse-grained labels, i.e., labels at the bag level.

In the embedded-space paradigm, a bag representation is built from the embeddings of its constituting instances (Amores, 2013), and then a standard ML classifier is used to classify the bag. This generalizes the collective assumption to the weighted collective assumption, where different instances contribute independently but not necessarily equally to the bag label (Foulds & Frank, 2010).

2.2. Attention-based DL for MIL

A deep learning-based MIL method, proposed by Ilse et al. (2018), employs an aggregation strategy involving a weighted average of the instance representations. The unnormalized weight of each instance is determined by an attention mechanism where a learned Key transformation for each instance is compared against a single learned Query vector. According to the MIL formulation, these instance weights can be seen as reflecting the probability that the (latent) instance label is positive (see Appendix A), allowing a direct connection to properties of instances in the domain. A priori information on the latent label of instances can thus be incorporated, e.g., through defining priors on the output values of this particular intermediate layer in the neural network (one neuron output value per instance in a bag). The internal Query vector for the attention mechanism can be seen as a template instance in a latent space, serving as a reference against which all inferred representations are compared. In the absence of any instance label information, the only way that this template vector is learned is through the gradient flowing back from the bag-level loss. It is thus a non-trivial task to learn a good attention network that correctly assigns high scores to positive instances, especially when the witness rate is low and the dataset is small. Incorporation of prior information on the outputs of the attention module could guide the model to learn a better Query vector in the backward pass (gradient flow), as well as directly produce an improved bag embedding in the forward pass (through increased weights for positive instances and an improved instance encoder).

DeepRC (Widrich et al., 2020) is also an attention-based deep MIL method, where a modern Hopfield network (MHN) is used to reconstruct patterns from a bag using attention and pooling. Widrich et al. (2020) have shown that MHNs can be used successfully for MIL problems with hundreds of thousands of instances and low witness rates. DeepRC was demonstrated in the case of immune repertoire classification, where the task is to predict a person's disease status from the repertoire of adaptive receptor sequences. We adopt DeepRC as a base model for implementing our knowledge incorporation strategies.

2.3. Multi-task learning

Multi-task learning is an approach that trains a model on multiple (typically related) tasks simultaneously to improve the overall performance of all the considered tasks or the specific performance of some target task (Crawshaw, 2020; Caruana, 1997). Particularly relevant to our work is the case where the different tasks are being performed at different layers in a model, which differs from the standard setup where all tasks are performed at the final layer, i.e., at the maximum network depth. For example, Søgaard & Goldberg (2016) train a multi-layered RNN on five different tasks where a loss term corresponding to each task was added after each layer. Hashimoto et al. (2017) and Sanh et al. (2018) follow a similar task hierarchy for related NLP problems. However, in both contexts, the model does not attribute semantic significance to the level at which the low-level tasks are performed.

3. Related Work

Several MIL papers have proposed different ways to incorporate instance-level information available for a subset of the whole dataset or inferred by the model. However, to the best of our knowledge, and despite some of these papers being tangentially close to our case, no previous work has the same setup as our problem. The closest to our setup is the work of Li et al. (2018), who formulate an object detection task as a MIL problem and use the bounding boxes available for a subset of the images as instance-level labels. They then train the model using a combination of CE loss and a MIL-inspired classification loss for the images with this additional information. However, in their setup, there is no direct mapping between specific DL model components and domain semantics, and they only explore one CE loss-based approach without exploring other ways to incorporate the information. Furthermore, the instance-level information is in their case annotated directly on the main dataset, while in our case the instance-level information is treated independently from the main dataset and can be incorporated from any separate auxiliary data source.

Other works include Yan et al. (2017), which use an EM algorithm to train an object detection model with access to a small number of images annotated at the instance level, i.e., with bounding boxes. Neither of these two methods optimizes the bounding boxes directly, while Seung Yeon Shin et al. (2019) does that directly using instance-level object detection losses (Ren et al., 2016) in addition to the image-level classification loss. Melendez et al. (2016) use an active learning-enhanced MIL approach where a human expert provides instance-level labels for the instances deemed the most valuable based on a defined criterion, whereas Choi et al. (2024) and Liu et al. (2023) enforce instance-level information inferred by the model itself. However, the latter differs from our case since the instance-level information is not from an external source and thus does not fall into the prior domain knowledge category.

4. Deep Multiple Instance Learning with Instance Priors (DeeMILIP)

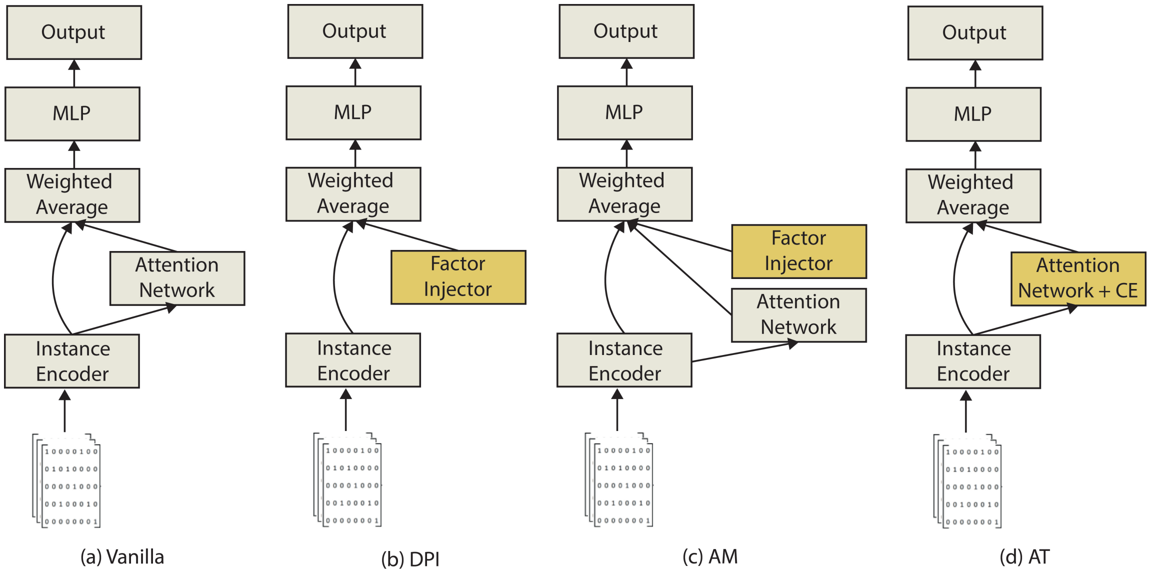
While instance-level labels are assumed to be unavailable in MIL scenarios, there are cases where (limited) information on instance labels might be obtainable. For example, in the considered case of immune-based diagnostics, some (limited and uncertain) information may be available from separate laboratory experiments regarding which instances are disease-relevant (more details in Section 5.1). The availability of such additional information raises the question of whether and how such auxiliary data can be effectively utilized to improve an MIL model. This consideration is particularly pertinent when dealing with a limited-size dataset and a low witness rate, which is the situation in, e.g., immune-based diagnostics (Greiff et al., 2020).

We here describe our proposed methodology in two steps: 1) how we define our available prior information and 2) how we formally incorporate the prior into a multiple-instance learning formulation.

4.1. Defining a prior for the instance class

In attention-based MIL, particularly DeepRC, the attention layer should output high attention values for instances indicative of the positive class (witnesses). These are the instances that discriminate between positive and negative bags and should thus get a high weight in the aggregate bag representation. The attention value cannot be directly interpreted as the probability of being a witness. Still, in a well-trained network, it should be monotonic to the probability of being positive (see Appendix A).

Here, we assume that we have an additional external source



(a) Vanilla (b) DPI (c) AM (d) AT

Figure 1. The DL components of the Vanilla DeepRC (a), as well as the three proposed strategies to incorporate instance-level probabilistic domain knowledge: Direct Prior Injection, DPI (b), Attention Modulation, AM (c), and Attention Training, AT (d).

of information regarding the class of instances. In general, we formulate this as a function π that maps an instance feature vector to a prior distribution over the probability of this instance belonging to the positive class. When the external source only indicates whether the instances have a high or low probability of being of the positive class (without information on uncertainty), this can be simplified to a specific value ρ_i^b of a Bernoulli distribution for y_i^b.

In our concrete case, we consider the simplified setting where the instances feature vectors are mapped to either of two values, referred to as ϕ^l and ϕ^h, respectively, where $0 < \phi^l < \phi^h < 1$.

The instance-level knowledge is assumed to be noisy, assigning high prior values to some non-witnesses (denoted as H$\overline{\text{W}}$) and low prior values to some witnesses (denoted as LW). The latter is more challenging since with H$\overline{\text{W}}$s, the model just needs to remove the few incorrect H$\overline{\text{W}}$s in the small pool of high-prior instances. On the other hand, with LWs, the model has to learn to identify these LWs in a much larger pool of low-prior instances (assuming a low overall witness rate, which is typical).

4.2. Incorporating instance priors into a MIL model

Conceptually speaking, the attention network learns to place higher weights on the witnesses (disease-relevant sequences) so that they contribute the most to the bag embedding. In Vanilla DeepRC, this occurs only through the gradient flowing back from the bag-level loss. In the presence of instance-level priors, several approaches could be used to achieve or facilitate this goal. The most straightforward way is to remove the attention network and directly use the prior (high or low level) as the attention score for a given instance. A more refined version of this approach is to retain the attention network and boost the attention score of positive sequences by a prior-associated factor. An alternative approach is to instead impose a cross-entropy loss on the attention network to minimize the discrepancy between the predicted and the prior probability distributions over the classes, conceptually resembling multi-level multi-task learning.

We here formally define these three ways of incorporating probabilistic domain knowledge (abbreviated as PDK from here onwards) into a MIL model, where the first two follow explicit prior injection approaches to model the output of intermediate layers, while the third follows a multi-task learning formulation.

4.2.1. General modeling

We model our observed bag labels with Bernoulli random variables Y^b

$$Y^b | \mathcal{X}^b \sim \text{Bern}(\theta(\mathcal{X}^b)). \tag{1}$$

For $\theta(\mathcal{X}^b)$, assume first trainable embeddings $h_i^b(x_i^b) = f_\upsilon(x_i^b) : \mathbb{R}^K \to \mathbb{R}^M, M < K$ modeled through an instance encoder, for example a convolutional network (LeCun et al., 1989). Then define the whole bag M-dimensional representation

$$z^b(\mathcal{X}^b) = \sum_{i=1}^{n_b} y_i^b \times \omega_i^b(x_i^b) \times h_i^b(x_i^b).$$

where ω is a strategy-dependent function that assigns an importance score to each instance in a given bag.

Here, y_i^b are random variables of the unobserved instance-level labels. We assume y_i^b to follow a Bernoulli distribution, i.e.

$$y_i^b | \rho_i^b \sim \text{Bern}(\rho_i^b)$$

with two distinct success probabilities:

$$\rho_i^b = \phi^h, i \in H^b \subseteq \{1, ..., n_b\}$$

and

$$\rho_j^b = \phi^l, j \in L^b = \{1, ..., n_b\} \setminus H^b,$$

corresponding to high and low expectations about $\Pr(y_i^b = 1)$ respectively.

Then $\theta(\mathcal{X}^b)$ is modeled as an MLP $m_\gamma(\cdot) : \mathbb{R}^M \to [0, 1]$ taking the expected representation $r(\mathcal{X}^b) = \text{E}_y[z^b(\mathcal{X}^b)]$ as input:

$$\theta(\mathcal{X}^b) = m_\gamma(r(\mathcal{X}^b)). \tag{2}$$

Motivated by the interpretability enabled by attention-based pooling (Widrich et al., 2020; Ilse et al., 2018) and by the fact that a model that effectively identifies the positive instances is more likely to achieve better bag label performance (Liu et al., 2012), we propose the following three modeling choices depicted in Figure 1: Direct Prior Injection (DPI), Attention Modulation (AM), and Attention Training (AT). As the main baselines, we consider the vanilla DeepRC model, referred to as Vanilla from here onwards, and a version that assigns equal weights to all instances, i.e., uses average pooling, referred to as AP. The latter resembles a DPI model having a PDK that assigns the same prior value to all instances. The exact equations are outlined in Appendix B for brevity.

For DPI

Let

$$\omega_i^b(x_i^b) = n_b^{-1}, i \in \{1, ..., n_b\}$$

effectively reducing the model to using only our PDK about y_i^b to model the attention, resulting in

$$\begin{aligned} r(\mathcal{X}^b) = \phi^h \sum_{i=1}^{n_b} n_b^{-1} \times h_i^b(x_i^b) \times \mathbb{1}\{i \in H^b\} \\ + \phi^l \sum_{i=1}^{n_b} n_b^{-1} \times h_i^b(x_i^b) \times \mathbb{1}\{i \in L^b\}. \end{aligned}$$

For AM

Let

$$\omega_i^b(x_i^b) = \frac{g_\kappa(x_i^b)}{\sum_{j=1}^{n_b} g_\kappa(x_i^b)}, i \in \{1, ..., n_b\} \tag{3}$$

effectively using the instances to model the attention weights through some neural network g with parameters κ, and then interacting with our random y_i^b in the representation of the bag b, resulting in the following expected representation:

$$\begin{aligned} r(\mathcal{X}^b) = \phi^h \sum_{i=1}^{n_b} \omega_i^b(x_i^b) \times h_i^b(x_i^b) \times \mathbb{1}\{i \in H^b\} \\ + \phi^l \sum_{i=1}^{n_b} \omega_i^b(x_i^b) \times h_i^b(x_i^b) \times \mathbb{1}\{i \in L^b\}, \end{aligned}$$

where ϕ^h and ϕ^l are implemented as hyperparameters of the model, up to a constant. For DPI, AM, Vanilla, and AP, the model is trained end-to-end by minimizing the following bag-level loss over all parameters of the network:

$$\begin{aligned} \ell_{bag} = -\sum_{b=1}^{\mathcal{B}} \Big[\mathcal{Y}^b \cdot \log p\left(\mathcal{Y}^b \mid \mathcal{X}^b\right) \\ + (1 - \mathcal{Y}^b) \cdot \log\left(1 - p(\mathcal{Y}^b \mid \mathcal{X}^b)\right) \Big], \end{aligned}$$

where $p\left(\mathcal{Y}^b \mid \mathcal{X}^b\right) = \theta(\mathcal{X}^b)$.

4.2.2. AT

We also consider a hybrid AT modeling approach where $\omega_i^b(x_i^b)$ is the same as in AM and Vanilla (3), and we further directly assume deterministic representations

$$r(\mathcal{X}^b) = \sum_{i=1}^{n_b} \omega_i^b(x_i^b) \times h_i^b(x_i^b).$$

Here, the additional knowledge of instance-level information is incorporated by putting dummy labels for all y_i^b, and the model is trained end-to-end by minimizing the following loss:

$$\ell_{total} = \ell_{bag} + \ell_{ins}$$

with

$$\begin{aligned} \ell_{ins} = -\sum_{b=1}^{\mathcal{B}} \sum_{i=1}^{n_b} \Big[\eta \cdot y_i^b \cdot \log p\left(y_i^b \mid x_i^b\right) \\ + (1 - y_i^b) \cdot \log\left(1 - p(y_i^b \mid x_i^b)\right) \Big], \end{aligned}$$

where $p\left(y_i^b \mid x_i^b\right) = \sigma(g_\kappa(x_i^b))$, σ is the sigmoid function, and η is the weight of the instances with a high prior value, set as a hyperparameter, in order to adjust for the PDK imbalance.

5. Experiments

Here, we explore how the different proposed approaches to integrating prior domain knowledge influence training and model behavior compared to a vanilla model. We consider a variety of problem settings and explore overall model performance as well as the behavior of the attention module in particular. Specifically, in Section 5.2, we investigate how Vanilla and AM perform at different witness rates, the influence of the dataset size on their performance, and their learning efficiency as a function of the number of weights updates. In Section 5.3, we compare the different strategies to explore how the attention module and PDKinteract to influence the model's performance. In Section 5.4, we study the behaviour of the different strategies at varying noise levels. Finally, in Section 5.5, we look at the behaviour of the models at the instance level.

We employed a 5-fold nested cross-validation (NCV) approach to assess each method's performance, where the hyperparameter values that give the lowest validation loss are used to evaluate the model on the test set for each outer loop. Additional details can be found in Appendix Table 6.

5.1. Dataset

We exemplify our DL MIL methodology on a case where the objective is to predict the immune status of a person (e.g., healthy vs. diseased) based on a large set of protein sequences called adaptive immune receptors (i.e., instances), comprising a person's immune repertoire (i.e., bag). We perform our experiments on a dataset that mimics the frontier of immune-based diagnostics, where recently emerged experimental technology allows us to simultaneously record a patient's immune repertoire and get (imperfect) indications on which particular immune cells are reacting to disease-relevant cells (Ma et al., 2021; Setliff et al., 2019).

As this emerging technology is not yet cost-effective for large patient cohorts, we use a hybrid dataset construction approach inspired by (Widrich et al., 2020) and MNIST-MIL (Ilse et al., 2018) to ensure control over the underlying truth, including the ability to emulate different controlled rates of signal strength (witness rate) (see Appendix D).

Specifically, we use sequences from approximately 1400 experimental repertoires of patients exposed to SARS-CoV-2 (Nolan et al., 2020) to construct 600 baseline repertoires containing 25,000 sequences on average through random sampling (all assumed to be of negative instance class). We then use the simARR package (Kanduri et al., 2022) to realistically inject a controlled proportion of immune receptor sequences experimentally annotated as binding to the EB Virus (positive instance class) (Bagaev et al., 2020) into half of the repertoires (positive repertoire class). We created six versions of the dataset, corresponding to the controlled proportion of EB Viruses (witness rate) in positive repertoires ranging from 0.02% to 2%.

To simulate in a controlled way the imperfect information on antigen binding (disease-relevant) immune cells (the noisy PDK), we assigned a high prior value to 20% of the EB Virus sequences (witnesses) (leaving the remaining 80% of witnesses with a low prior value), as well as assigning a high prior value to an equal number of randomly selected sequences (non-witnesses) to serve as $\mathrm{H\overline{W}}$s. For one experiment (Section 5.5, Figure 2), we further create 20 versions of the dataset at 0.2% witness rate, where we systematically vary what fraction of witnesses are assigned a high prior value (referred to as High-prior Witnesses Rate (HWR) and what fraction of the high-prior-valued instances are non-witnesses (referred to as High-prior Non-Witnesses Rate ($\mathrm{H\overline{W}R}$)). Similar experiments are conducted in Appendix I on a MIL version of the MNIST dataset following (Ilse et al., 2018).

5.2. Priors help learn weaker signals from less available data, with less compute

Table 1. Bag-level ROCAUC score on the test set, mean ± standard error, at different witness rates over 5-fold NCV. In the immunology setting, the witness rate reflects the proportion of a person's immune cells connected to the particular disease state being studied (here, EBV).

Witness rate	Vanilla	AM
0.02 %	0.52 ± 0.03	0.66 ± 0.04
0.05 %	0.53 ± 0.02	0.84 ± 0.05
0.1 %	0.50 ± 0.04	0.88 ± 0.05
0.2 %	0.54 ± 0.01	0.99 ± 0.00
1 %	0.78 ± 0.11	1.00 ± 0.00
2 %	0.81 ± 0.12	1.00 ± 0.00

Table 2. Bag-level ROCAUC score on the test set, mean ± standard error, at different training dataset size over 5-fold NCV at WR = 2%.

Dataset Size	Vanilla	AM
12	0.56 ± 0.03	0.98 ± 0.02
48	0.52 ± 0.03	1.00 ± 0.00
96	0.63 ± 0.08	1.00 ± 0.00
180	0.72 ± 0.11	1.00 ± 0.00
360	0.81 ± 0.12	1.00 ± 0.00

Incorporating prior information into a DL MIL model may enable it to acquire a useful bag embedding even in scenarios of sparse signals characterized by a low witness rate in a MIL formulation. As seen from Table 1, Vanilla requires a witness rate of 1% (250 positive instances expected per bag) to successfully distinguish positive from negative bags on this dataset. In contrast, incorporating prior information for the instances allows successful classification already at a witness rate of 0.05%, representing a twenty-fold reduction

Table 3. Bag-level ROCAUC score on the validation set, mean ± standard error, at different steps over 5-fold NCV at $\mathrm{WR} = 2\%$.

Step	Vanilla	AM
0	0.55 ± 0.04	0.44 ± 0.18
1000	0.59 ± 0.06	1.00 ± 0.00
5000	0.69 ± 0.08	1.00 ± 0.00
10000	0.73 ± 0.11	1.00 ± 0.00
50000	0.81 ± 0.12	1.00 ± 0.00
100000	0.81 ± 0.12	1.00 ± 0.00

compared to 1%.

The performance of DL models is also known to rely on the availability of large datasets, posing a challenge when confronted with limited dataset sizes. This is particularly pronounced in MIL, where the signal tends to be sparse, and the bag size is substantial, thereby complicating the identification of relevant instances. However, incorporating prior information allows us to learn a successful classifier based on only 12 bags in the training data set at a witness rate of 2% (a thirty-fold improvement over Vanilla, which requires 360 bags to perform well).

DL is also known to require heavy compute, which can hinder development by labs with restricted computing resources and lead to considerable energy consumption, negatively impacting the environment. The results in Table 3 show that using prior information can help the DL model learn faster, requiring fewer weight updates while also achieving a higher ROCAUC score. Similar results for AT and DPI are shown in Appendix E.

5.3. Probabilistic domain knowledge complements learned attention

Table 4. Bag-level ROCAUC score on the test set, mean ± standard error, for no PDK, for Prior Injection and for Multi-Task Learning (MTL), with or without an attention module, over 5-fold NCV, at $\mathrm{WR} = 2\%$.

	w/o Attention	w/ Attention
No PDK	AP: 0.75 ± 0.07	Vanilla: 0.81 ± 0.12
Prior injection	DPI: 0.93 ± 0.07	AM: **1.00** ± **0.00**
MTL	/	AT: **1.00** ± **0.00**

As introduced in Section 4, PDK can be incorporated explicitly through prior injection, through the definition of an auxiliary loss, or not used at all. Furthermore, the DL model can include or not include an attention learning module. This gives rise to five different formulations, as shown in Table 4.

Incorporating PDK or attention-learning in isolation (DPI and Vanilla, respectively) improved performance over the AP baseline. When combined, the performance was better than either alone, both when injecting a prior (AM) and when using a multi-task learning formulation (AT).

Additional baselines are considered in Appendix G.

5.4. The DL MIL model benefits also from noisy domain knowledge

The domain knowledge that is available at the instance level may be very noisy in the sense that only a small proportion of witnesses may be assigned a high prior (and/or many high-prior instances being non-witnesses).

To this end, we systematically explore the performance of DPI, AM, and AT when HWRs and $\mathrm{H\overline{W}R}$s vary from 5% to 100% and from 0% to 80%, respectively. Figure 2 shows that even the naive approach of incorporating PDK, i.e., DPI, already improves the performance compared to Vanilla at sufficiently high HWRs. More interestingly, AM and AT can benefit even from the least informative PDK, at $\mathrm{HWR} = 5\%$ and $\mathrm{H\overline{W}R} = 80\%$, and lead to well-performing models.

5.5. Probabilistic domain knowledge even boosts attention learning

Table 5. Average difference of instance-level PRAUC($\times$100) score against a random classifier on the test set, mean ± standard error, at $\mathrm{WR} = 2\%$, $\mathrm{HWR} = 20\%$s and $\mathrm{H\overline{W}R} = 50\%$ for different strategies over 5-fold NCV. The class in each column is considered the positive class against $\mathrm{L\overline{W}}$s, i.e., non-witnesses with low prior values, as the negative class. Higher is better in the first and third columns, and lower is better in the second.

	HW	$\mathrm{H\overline{W}}$	LW
Vanilla	1.429 ± 0.522	**0.006** ± **0.005**	1.422 ± 0.440
AM	7.715 ± 1.072	0.011 ± 0.006	1.619 ± 0.390
AT	**21.516** ± **2.603**	0.502 ± 0.180	**1.834** ± **0.455**

Since we assume that the available PDK can be noisy (low HWRs and high $\mathrm{H\overline{W}R}$s), a model's ability to recover witnesses with a low prior and ignore non-witnesses with high priors can be highly desirable for learning robust models.

Table 5 shows the ability of the AM and AT approaches to boost the high-prior witnesses through the attention module, recover witnesses with low prior values and ignore non-witnesses with high prior values. As expected, the instance-level PRAUC for high-prior witnesses against low-prior non-witnesses is substantially increased when introducing a cross-entropy loss against the PDK for the attention module output (AT). Interestingly, despite not being trained at the instance level, the attention modulation strategy (AM) also leads to a considerable increase in PRAUC for the raw attention values. It is also noteworthy that both approaches (AM

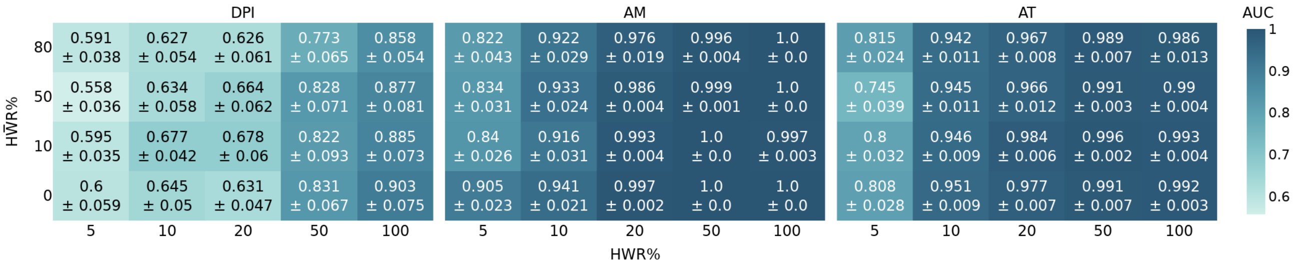


Figure 2. Test ROCAUC scores of different strategies at $\mathrm{WR} = 0.2\%$ over different HWRs and $\mathrm{H\overline{W}R}$s. In the immunology setting, the HWR reflects what fraction of the disease-relevant cells are already annotated as disease-relevant in auxiliary databases, while the $\mathrm{H\overline{W}R}$) rate reflects the fraction of false positive annotations in these databases. Text color is chosen for clarity purposes.

and AT) are able to recover low-prior witnesses better than Vanilla (more evident at $\mathrm{WR} = 0.2\%$, as shown in Table 10 in the Appendix), and that especially AM is able to almost completely ignore high-prior non-witnesses (while still assigning high attention to the high-prior witnesses). Density plots of attention scores for each of the four instance types, i.e., HW, $\mathrm{H\overline{W}}$, LW, $\mathrm{L\overline{W}}$, can be found in the Appendix for $\mathrm{WR} = 2\%$ (Figure 3) and $\mathrm{WR} = 0.2\%$ (Figure 4).

Discussion

The incorporation of quantitative domain knowledge is a promising direction for improving the generalization of DL in settings with limited data. While the Bayesian framework is well suited for knowledge integration, its application in DL, particularly through BNNs (Jospin et al., 2022), has mainly centered on quantifying uncertainty, typically through the incorporation of generic priors like zero-centered Gaussian or Laplace distributions on model parameters (Murphy, 2012; Fortuin, 2022). We have in this work focused on the integration of auxiliary quantitative information for a domain, which we refer to as probabilistic domain knowledge (PDK), into a DL model. The prior we are interested in does not operate at the parameter level, as with BNNs. Instead, it pertains to the values of intermediate features extracted by the model, which, in turn, are a function of the model's parameters.

While the learned transformations of DL make it impossible, in general, to connect entities in the domain (for which external PDK is available) to particular model components, we here rely on the particular structure of MIL, and a previously proposed modularized DL MIL model in particular (Widrich et al., 2020), to make such a connection. This allows us to inject prior knowledge into the outputs from specific components of the model. In our applied case, the attention output is connected to the probability that individual immune cells react to disease-relevant molecules, enabling us to incorporate PDK defined for individual immune cells to the attention module's output. We further compare this prior injection approach to the approach of defining multiple related subtasks in a multi-task learning formulation as a way to incorporate PDK.

As expected, incorporating PDK allows us to learn a good MIL model even in settings with weaker signals (lower witness rate), less data, and less compute. This performance is not solely due to the up-weighting of high-prior witnesses in the bag embedding - introspection of attention values showed that the module itself learns to assign high attention to high-prior (and even low-prior) witnesses. Such learning occurs at witness rates for which a model without prior does not learn any useful attention function, which for AM could be due to the prior amplifying the flow of gradients through relevant instances (HW).

While we focus on a particular setting of adaptive immunology, we believe this type of prior information may be useful for MIL in multiple settings and domains. For instance, in biomedical imaging, a microscopy image requiring final classification for the entire image typically contains many cells, some of which experts label as being disease-relevant. In quality control, products like washing machines must be categorized into different quality levels, with individual sub-components potentially labeled by experts, but the final label is determined by the quality control process or the consumer. Similarly, in drug discovery, drug combinations or mixtures are assessed for their properties, with some molecules labeled and others unlabelled. This makes our approach applicable to a wide range of applications where (partial) instance-level information is available.

Limitation and future directions Our ability to incorporate a prior relied on particular characteristics of the MIL problem (latent instance classes) and DL architecture (a separate attention module). An interesting question is which other problem formulations might provide such opportunities to connect DL model components with domain entities and whether such cases might shed light on additional aspects of the behavior of deep learning models when prior information is incorporated. For the DL MIL context, it would also be interesting to better understand the effects (in terms of gradient flow and representational spaces) on different components of the model (instance embedding, attention, bag classifier) when PDK is incorporated. Finally, the PDK we considered only indicated two prior levels (high and low) for the instance class. Considering a richer PDK could allow a more detailed exploration of the effects of prior uncertainty as well as prior level, including investigation of how such information could be incorporated as e.g. a beta-distributed hyper-prior for class probability rather than directly injecting a single class probability as done here.

Impact Statement

As the proposed strategies allow us to learn generalizable models from less data, this directly translates to minimizing both financial and human resources allocated to data collection. Additionally, since incorporating PDK, enables models to achieve high performance with fewer training steps than the vanilla model, this has the potential of reducing the amount of time and compute required to train such models. Interestingly, the proposed methods require no additional computation which contributes even more to less compute requirements. A detailed description of the compute requirements can be found in Appendix J.

Conclusion

Here, we showed that the structured nature of MIL permits a direct mapping between entities in a domain and components of a modularised DL MIL model. We proposed multiple strategies to exploit this mapping for prior knowledge incorporation. Interestingly, the injection of a prior on the output of the attention module not only complemented the learned attention but even improved the learning in the attention module itself. Compared to a vanilla model, both the injection of a prior and the definition of an additional instance-level task allowed the model to learn successful classifiers at a considerably lower witness rate, based on less available data and with less compute.

Acknowledgements

Ghadi received internationalization support from *UiO:Life Science* for his research stay at the Johannes Kepler Universität - Linz. We thank Adín Ramírez Rivera for the useful discussions throughout the project.

References

Akbar, R., Bashour, H., Rawat, P., Robert, P. A., Smorodina, E., Cotet, T.-S., Flem-Karlsen, K., Frank, R., Mehta, B. B., Vu, M. H., Zengin, T., Gutierrez-Marcos, J., Lund-Johansen, F., Andersen, J. T., and Greiff, V. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. 14(1):2008790, 2022. ISSN 1942-0870. doi: 10.1080/19420862.2021.2008790.

Amores, J. Multiple instance classification: Review, taxonomy and comparative study. 201:81–105, 2013. ISSN 0004-3702. doi: 10.1016/j.artint.2013.06.003. URL https://www.sciencedirect.com/science/article/pii/S0004370213000581.

Bagaev, D. V., Vroomans, R. M. A., Samir, J., Stervbo, U., Rius, C., Dolton, G., Greenshields-Watson, A., Attaf, M., Egorov, E. S., Zvyagin, I. V., Babel, N., Cole, D. K., Godkin, A. J., Sewell, A. K., Kesmir, C., Chudakov, D. M., Luciani, F., and Shugay, M. VDJdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. 48:D1057–D1062, 2020. ISSN 1362-4962. doi: 10.1093/nar/gkz874.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017.

Cao, W., Yan, Z., He, Z., and He, Z. A comprehensive survey on geometric deep learning. *IEEE Access*, 8:35929–35949, 2020.

Carbonneau, M.-A., Cheplygina, V., Granger, E., and Gagnon, G. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition*, 77:329–353, 2018.

Caruana, R. Multitask learning. 28(1):41–75, 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734. URL https://doi.org/10.1023/A:1007379606734.

Choi, W.-G., Chang, J.-H., Yang, J.-M., and Moon, H.-G. Instance-level loss based multiple-instance learning framework for acoustic scene classification. 216:109757, 2024. ISSN 0003-682X. doi: 10.1016/j.apacoust.2023.109757. URL https://www.sciencedirect.com/science/article/pii/S0003682X23005558.

Crawshaw, M. Multi-task learning with deep neural networks: A survey, 2020. URL http://arxiv.org/abs/2009.09796.

Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., and Piccialli, F. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.

Dietterich, T. G., Lathrop, R. H., and Lozano-Pérez, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(96)00034-3. URL https://www.sciencedirect.com/science/article/pii/S0004370296000343.

Early, J., Cheung, G. K. C., Cutajar, K., Xie, H., Kandola, J., and Twomey, N. Inherently interpretable time series classification via multiple instance learning. *ArXiv*, abs/2311.10049, 2023. URL https://api.semanticscholar.org/CorpusID:265221436.

Emerson, R. O., DeWitt, W. S., Vignali, M., Gravley, J., Hu, J. K., Osborne, E. J., Desmarais, C., Klinger, M., Carlson, C. S., Hansen, J. A., Rieder, M., and Robins, H. S. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the t cell repertoire. 49(5):659–665, 2017. ISSN 1546-1718. doi: 10.1038/ng.3822. URL https://www.nature.com/articles/ng.3822. Number: 5 Publisher: Nature Publishing Group.

Fortuin, V. Priors in bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022.

Foulds, J. and Frank, E. A review of multi-instance learning assumptions. *The knowledge engineering review*, 25(1): 1–25, 2010.

Greiff, V., Yaari, G., and Cowell, L. G. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. 24:109–119, 2020. ISSN 2452-3100. doi: 10.1016/j.coisb.2020.10.010. URL https://www.sciencedirect.com/science/article/pii/S2452310020300524.

Hashimoto, K., Xiong, C., Tsuruoka, Y., and Socher, R. A joint many-task model: Growing a neural network for multiple NLP tasks. In Palmer, M., Hwa, R., and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1923–1933. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1206. URL https://aclanthology.org/D17-1206.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html. ISSN: 2640-3498.

Javed, S. A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., and Prakash, A. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 20689–20702. Curran Associates, Inc.

Jospin, L. V., Laga, H., Boussaid, F., Buntine, W., and Bennamoun, M. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022. doi: 10.1109/MCI.2022.3155327.

Kanduri, C., Scheffer, L., Pavlović, M., Rand, K. D., Chernigovskaya, M., Pirvandy, O., Yaari, G., Greiff, V., and Sandve, G. K. simAIRR: simulation of adaptive immune repertoires with realistic receptor sequence sharing for benchmarking of immune state prediction methods. 12:giad074, 2022. ISSN 2047-217X. doi: 10.1093/gigascience/giad074.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Kayhan, O. S. and Gemert, J. C. v. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2017.

Kraus, O. Z., Ba, J. L., and Frey, B. J. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J., and Fei-Fei, L. Thoracic disease identification and localization with limited supervision. 2018.

Liu, G., Wu, J., and Zhou, Z.-H. Key instance detection in multi-instance learning. In Hoi, S. C. H. and Buntine, W. (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 253–268, Singapore Management University, Singapore, 2012. PMLR. URL https://proceedings.mlr.press/v25/liu12b.html.

Liu, K., Zhu, W., Shen, Y., Liu, S., Razavian, N., Geras, K. J., and Fernandez-Granda, C. Multiple instance learning via iterative self-paced supervised contrastive learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3355–3365. IEEE, 2023. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.00327. URL https://ieeexplore.ieee.org/document/10205311/.

Ma, K.-Y., Schonnesen, A. A., He, C., Xia, A. Y., Sun, E., Chen, E., Sebastian, K. R., Guo, Y.-W., Balderas, R., Kulkarni-Date, M., and Jiang, N. High-throughput and high-dimensional single-cell analysis of antigen-specific CD8+ t cells. 22(12):1590–1598, 2021. ISSN 1529-2916. doi: 10.1038/s41590-021-01073-2.

Marcus, G. Deep learning: A critical appraisal, 2018.

Maron, O. and Lozano-Pérez, T. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

Melendez, J., van Ginneken, B., Maduskar, P., Philipsen, R. H. H. M., Ayles, H., and Sanchez, C. I. On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis. 35(4):1013–1024, 2016. ISSN 1558-254X. doi: 10.1109/TMI.2015.2505672.

Murphy, K. P. *Machine learning: a probabilistic perspective*. 2012.

Nolan, S., Vignali, M., Klinger, M., Dines, J. N., Kaplan, I. M., Svejnoha, E., Craft, T., Boland, K., Pesesky, M., Gittelman, R. M., Snyder, T. M., Gooley, C. J., Semprini, S., Cerchione, C., Mazza, M., Delmonte, O. M., Dobbs, K., Carreño-Tarragona, G., Barrio, S., Sambri, V., Martinelli, G., Goldman, J. D., Heath, J. R., Notarangelo, L. D., Carlson, J. M., Martinez-Lopez, J., and Robins, H. S. A large-scale database of t-cell receptor beta (TCR) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. pp. rs.3.rs–51964, 2020. doi: 10.21203/rs.3.rs-51964/v1. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418738/.

Quellec, G., Cazuguel, G., Cochener, B., and Lamard, M. Multiple-instance learning for medical image and video analysis. *IEEE reviews in biomedical engineering*, 10:213–234, 2017.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Sanh, V., Wolf, T., and Ruder, S. A hierarchical multi-task approach for learning embeddings from semantic tasks, 2018. URL http://arxiv.org/abs/1811.06031.

Setliff, I., Shiakolas, A. R., Pilewski, K. A., Murji, A. A., Mapengo, R. E., Janowska, K., Richardson, S., Oosthuysen, C., Raju, N., Ronsard, L., Kanekiyo, M., Qin, J. S., Kramer, K. J., Greenplate, A. R., McDonnell, W. J., Graham, B. S., Connors, M., Lingwood, D., Acharya, P., Morris, L., and Georgiev, I. S. High-throughput mapping of b cell receptor sequences to antigen specificity. 179(7):1636–1646.e15, 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2019.11.003.

Seung Yeon Shin, n., Soochahn Lee, n., Il Dong Yun, n., Sun Mi Kim, n., and Kyoung Mu Lee, n. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. 38 (3):762–774, 2019. ISSN 1558-254X. doi: 10.1109/TMI.2018.2872031.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Søgaard, A. and Goldberg, Y. Deep multi-task learning with low level tasks supervised at lower layers. In Erk, K. and Smith, N. A. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 231–235. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-2038. URL https://aclanthology.org/P16-2038.

Valkiers, S., de Vrij, N., Gielis, S., Verbandt, S., Ogunjimi, B., Laukens, K., and Meysman, P. Recent advances in t-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing. 5:100009, 2020. ISSN 2667-1190. doi: 10.1016/j.immuno.2022.100009. URL https://www.sciencedirect.com/science/article/pii/S2667119022000015.

Widrich, M., Schäfl, B., Pavlović, M., Ramsauer, H., Gruber, L., Holzleitner, M., Brandstetter, J., Sandve, G. K., Greiff, V., Hochreiter, S., and Klambauer, G. Modern hopfield networks and attention for immune repertoire classification. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18832–18845. Curran Associates, Inc., 2020. URL https://dl.acm.org/doi/10.5555/3495724.3497305.

Yan, Z., Liang, J., Pan, W., Li, J., and Zhang, C. Weakly- and semi-supervised object detection with expectation-maximization algorithm. 2017.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., and Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

Appendix

A. Justification for the Probabilistic Interpretation of Attention Values in MIL

According to the standard MIL assumption, instances are of two distinct latent classes (of an unobserved positive or negative label), where the distribution of feature values is conditional only on this latent instance class. Labels are only observed at the bag level, where positive bags are a mixture of positive and negative instances (of unobserved labels), while negative bags exclusively contain negative instances.

To avoid that the bag aggregation will drown the signals of (systematic feature value differences for) positive instances in sampling noise of feature values contributed by negative instances in bags, the aggregation can instead be a weighted average of instances. The weights can be learned as a function from instance feature values to weight. In deep learning terminology, this can be described as a learned attention for the set of instance feature values.

Since only positive instances contribute a discriminative signal for the bag embedding, the exclusive aim for such a function is to assign high weights to positive instances. A higher estimated probability for an instance to be positive should thus correspond to a higher weight. If we define $d(x)$ as the (true) probability of an instance being positive given its feature values, i.e., $d(x) = P(y = 1|x)$, then our aim is to learn some function g, such that $\sigma(g_\kappa(x))$ is monotonic to $d(x)$, where σ is the sigmoid function.

B. Vanilla DeepRC and Average Pooling

As with AT, we directly assume deterministic representations in Vanilla and AP. The pooling equations of the baseline methods, then, become as follows:

Vanilla

$$r(\mathcal{X}^b) = \sum_{i=1}^{n_b} \omega_i^b(x_i^b) \times h_i^b(x_i^b)$$

i.e., the same as AT, except that the objective does not include a term for the instance-level loss.

Average Pooling (AP)

$$r(\mathcal{X}^b) = \sum_{i=1}^{n_b} n_b^{-1} \times h_i^b(x_i^b)$$

C. Training details

In each of the main experiments, 5-fold nested cross-validation was used where, in each outer loop, one training-validation split was made for model selection, and then the best model was evaluated with the best HP set. Each model was trained for 30,000 updates, unless otherwise specified, using the Adam optimizer (Kingma & Ba, 2017) with a learning rate of $1\text{e}{-4}$.

Table 6. Tested hyperparameter values for each of the different strategies.

Strategy	Hyperparameter	Value
Vanilla	λ_{ℓ_2}	$\{1\text{e}{-4}, 1\text{e}{-3}, 0\}$
DPI	ϕ^h	$\{20, 100, 500\}$
AM	ϕ^h	$\{20, 100, 500\}$
AT	η	$\{100, 500, 1000\}$
AP	None	/

For DPI and AM, ϕ^l is set to 1. This is equivalent to modelling ϕ^l and ϕ^h as:

$$\phi^l = \frac{1}{value + 1} \text{ and } \phi^h = \frac{value}{value + 1}$$

to match their probabilistic interpretation, proposed in Section 4.2.1.

D. Additional Dataset Information

As the technology to simultaneously record a patient's immune repertoire and get (imperfect) indications on which immune receptors are reacting to disease-relevant cells (Ma et al., 2021; Setliff et al., 2019) is not yet cost-effective for large patient cohorts, available experimental data is currently of two separate types (Greiff et al., 2020): A) repertoires: large sets of (largely) non-annotated immune receptor sequences from individuals having a particular immune state (disease) (Emerson et al., 2017), B) immune receptor sequences: sets of immune receptor sequences labeled with their particular (disease-relevant) target molecule (antigen) (e.g., (Bagaev et al., 2020)).

Since the set of antigen-binding immune receptors (type B) is still vastly undersampled due to technological limitations (Akbar et al., 2022; Valkiers et al., 2020), directly screening for these in patient repertoires (type A) currently gives incomplete annotations of disease-relevant immune receptors. For this reason, we used a hybrid dataset construction approach to have control of the underlying ground truth in our experiments. This dataset construction approach also allowed us to vary the signal strength (witness rate) in a controlled way in the experiments.

E. Additional Results

Table 7. Bag-level ROCAUC score on the test set, mean ± standard error, at different witness rates over 5-fold NCV.

Witness rate	AT	DPI
0.02 %	0.74 ± 0.05	0.55 ± 0.02
0.05 %	0.79 ± 0.04	0.54 ± 0.03
0.1 %	0.95 ± 0.01	0.54 ± 0.04
0.2 %	0.97 ± 0.01	0.67 ± 0.06
1 %	0.99 ± 0.01	0.89 ± 0.05
2 %	1.00 ± 0.00	0.93 ± 0.07

Table 8. Bag-level ROCAUC score on the test set, mean ± standard error, at different training data set sizes over 5 runs at WR=2%.

Dataset Size	AT	DPI
12	0.98 ± 0.01	0.87 ± 0.04
48	0.95 ± 0.03	0.91 ± 0.06
96	0.99 ± 0.01	0.88 ± 0.07
180	0.99 ± 0.01	0.94 ± 0.06
360	1.00 ± 0.07	0.93 ± 0.00

Table 9. Bag-level ROCAUC score on the validation set, mean ± standard error, at different steps over 5-fold NCV at WR $= 2\%$.

Step	AT	DPI
0	0.55 ± 0.04	0.50 ± 0.00
1000	0.52 ± 0.10	0.84 ± 0.04
5000	0.99 ± 0.01	0.84 ± 0.07
10000	1.00 ± 0.00	0.85 ± 0.06
50000	1.00 ± 0.00	0.85 ± 0.09
100000	1.00 ± 0.00	0.85 ± 0.07

Table 10. Average difference of instance-level PRAUC($\times 100$) score against a random classifier on the test set, mean ± standard error, at $\text{WR} = 0.2\%$ for different strategies over 5-fold NCV.

	HW	H$\overline{\text{W}}$	LW
Vanilla	0.000 ± 0.002	0.008 ± 0.006	-0.002 ± 0.008
AM	8.984 ± 1.932	0.011 ± 0.005	0.305 ± 0.220
AT	36.35 ± 1.623	0.074 ± 0.029	0.733 ± 0.447

F. Density Plots

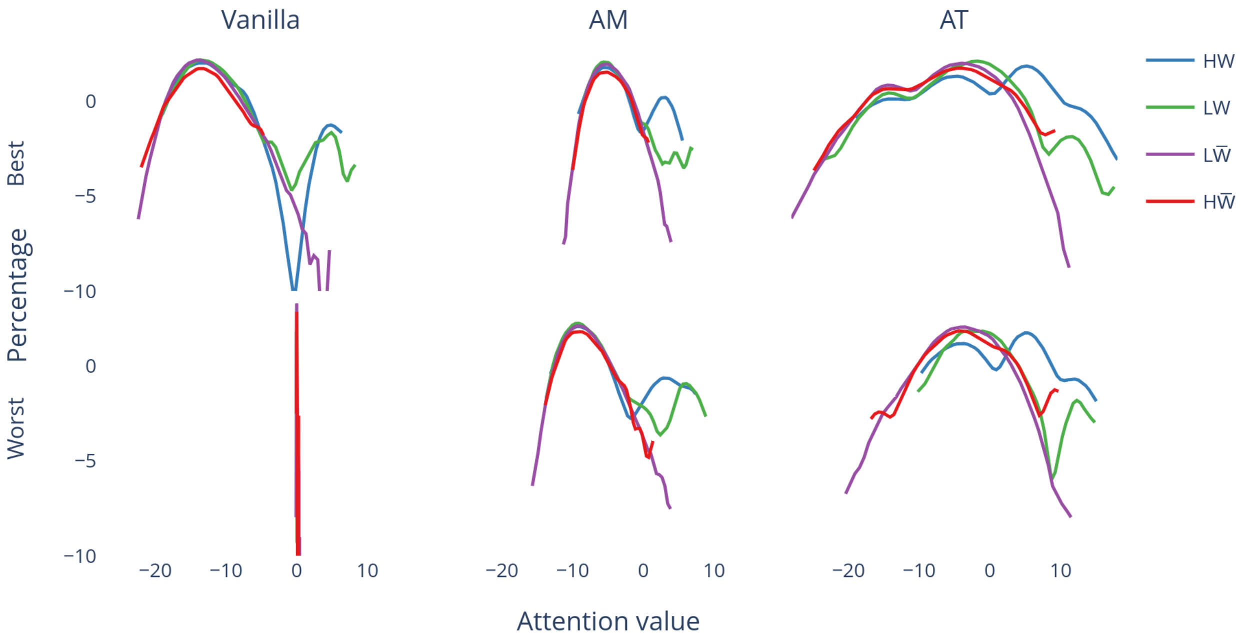


Figure 3. Density plots of the attention scores on the test set for the best and the worst performing models, in terms of cross-entropy loss on the test set, the Vanilla, AM, and AT strategies at $\text{WR} = 2\%$. The log scale is used for the y-axis for clarity, i.e., log(percentage).

G. Additional Baselines

To compare the results of ROCAUC scores in a broader context, we added four baselines. Namely, a logistic regression model with L1 regularization (L1 LogReg), a logistic regression model with L2 regularization (L2 LogReg), a support vector classifier with L2 regularization (L2 SVC), and a variation of DeepRC that uses Conjunctive pooling (Conj) Early et al. (2023).

Table 11. Bag-level ROCAUC score on the test set, mean ± standard error, at different witness rates over 5-fold NCV.

Witness rate	L1 LogReg	L2 LogReg	L2 SVC	Conj
0.01 %	0.57 ± 0.00	0.55 ± 0.02	0.54 ± 0.01	0.51 ± 0.01
0.02 %	0.50 ± 0.02	0.53 ± 0.03	0.51 ± 0.02	0.54 ± 0.02
0.1 %	0.49 ± 0.02	0.51 ± 0.02	0.46 ± 0.02	0.49 ± 0.03
0.2 %	0.63 ± 0.02	0.60 ± 0.03	0.59 ± 0.02	0.47 ± 0.01
1 %	0.91 ± 0.01	0.91 ± 0.01	0.82 ± 0.01	0.63 ± 0.06
2 %	0.99 ± 0.00	0.99 ± 0.00	0.95 ± 0.00	0.85 ± 0.09

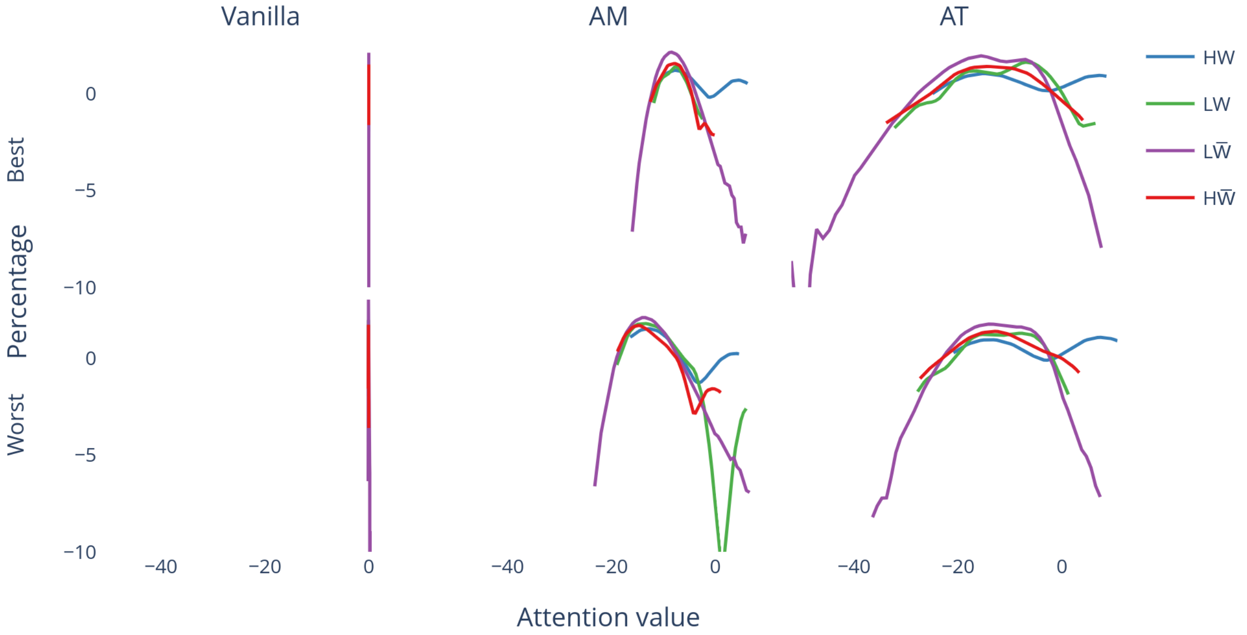


Figure 4. Density plots of the attention scores on the test set for the best and the worst performing models, in terms of cross-entropy loss on the test set, the Vanilla, AM, and AT strategies at $\mathrm{WR} = 0.2\%$. The log scale is used for the y-axis for clarity, i.e., log(percentage).

To compare how the proposed strategies compare to a model that predicts solely on the prior knowledge, we train a logistic regression model that predicts the disease state based on the counts of instances with a high prior, regardless of whether they are witnesses. We do that across a combination of HWRs and $\mathrm{H\overline{W}R}$s in addition to having no prior at all. In the case of no prior, AT and AM default back to Vanilla since AT then has no instance-level loss, and AM would scale all attention weights in the same way, which is equivalent to no scaling since the softmax operation cancels the common factor. DPI, on the other hand, defaults back to AP since all the instances would be given the same pre-softmax attention weight.

Table 12. Bag-level ROCAUC score on the test set, mean ± standard error, at combinations of HWR and $\mathrm{H\overline{W}R}$ (represented as $\mathrm{HWR}/\mathrm{H\overline{W}R}$) over 5-fold NCV at $\mathrm{WR} = 0.2\%$.

Model	Vanilla	AT	AM	DPI	LogReg
No Prior	0.54 ± 0.01	0.54 ± 0.01	0.54 ± 0.01	0.50 ± 0.01	0.50 ± 0.00
5/80	0.54 ± 0.01	0.82 ± 0.02	0.82 ± 0.02	0.59 ± 0.04	0.69 ± 0.01
20/80	0.54 ± 0.01	0.97 ± 0.01	0.99 ± 0.00	0.66 ± 0.06	0.88 ± 0.02
50/50	0.54 ± 0.01	0.99 ± 0.00	1.00 ± 0.00	0.83 ± 0.07	1.00 ± 0.00
100/0	0.54 ± 0.01	0.99 ± 0.00	1.00 ± 0.00	0.90 ± 0.08	1.00 ± 0.00

H. Further Hyperparameter Optimization for Vanilla DeepRC

For compute resources reasons, we have restricted the number of hyperparameter values to three and the number of updates in all the experiments to 30,000. Nonetheless, to ensure that these are not the restricting factors that prevent the vanilla model from performing well, we conduct a more extensive hyperparameter search for the vanilla model with 100,000 updates. The tested HP values are adopted from Table A15 in the original DeepRC paper (Widrich et al., 2020), reproduced here for ease of reference in Table 13. We exclude the number of attention heads from the set of optimized hyperparameters.

For computational reasons, instead of using a grid search, we use the middle value of each tuple as the default value and then try the other values of each hyperparameter while keeping the others at the default value, giving eight different hyperparameter combinations, in addition to the default values of all hyperparameters, resulting in a total of nine combinations per fold. Following (Widrich et al., 2020), we run three folds of a 5-fold NCV setup. The validation ROCAUC of these three folds are shown in Figures 5, 6, 7 with smoothing using an exponential moving average for clarity purposes with smoothing constant $\alpha = 0.35$. The test ROCAUC score has a mean of 0.4696 and a standard error of 0.0158.

Table 13. Hyperparameters of the model with the values to test as proposed in (Widrich et al., 2020).

learning rate	10^{-4}
number of attention heads	$\{1; 16; 64\}$
β of attention softmax	$\{0.1; 1.0; 10.0\}$
l2 weight penalty	$\{1.0; 0.1; 0.01\}$
number of kernels	$\{8; 32; 128\}$
number of CNN layers	$\{1\}$
number of layers in key-NN	$\{2\}$
number of units in key-NN	$\{32\}$
kernel size	$\{5; 7; 9\}$
subsampled sequences	10,000
batch size	4

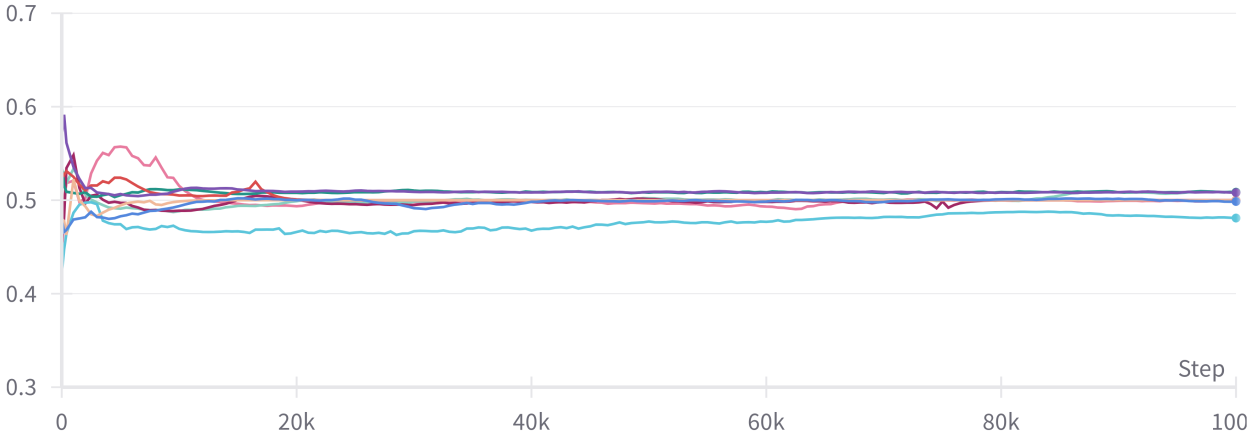


Figure 5. Validation ROCAUC score for the first fold as a function of the number of updates.

I. MNIST-MIL

For these experiments, bags were constructed following Ilse et al. (2018), where each bag consists of several MNIST images, and images holding the number 9 are assigned as the positive instance class. To make the problem more challenging, we use a bag size of mean 500 and variance 5 compared to the smaller bags used in Ilse et al. (2018). Additionally, to allow for more precise investigation, we manually specify the witness rate instead of using the default value of 0.1 (since the classes in MNIST are in equal proportions) and use low values to make the data set challenging. All the experiments are run using $\mathrm{HWR} = 20\%$ and $\mathrm{H\overline{W}R} = 50\%$.

For these experiments, we use the attention pooling-based ADMIL model (Ilse et al., 2018) as the base model. Additionally, in order to showcase the versatility of our proposed approaches and their applicability to any embedding-space or instance-space MIL method, we run the same experiments on two variations of the ADMIL model, one with conjunctive pooling (Early et al., 2023) and one with additive pooling (Javed et al.). Note that in the Additive pooling method, we already apply the prior information at the attention network level. On the other hand, since the Conjunctive pooling method inherently admits a classifier, we inject the prior information into the classifier instead.

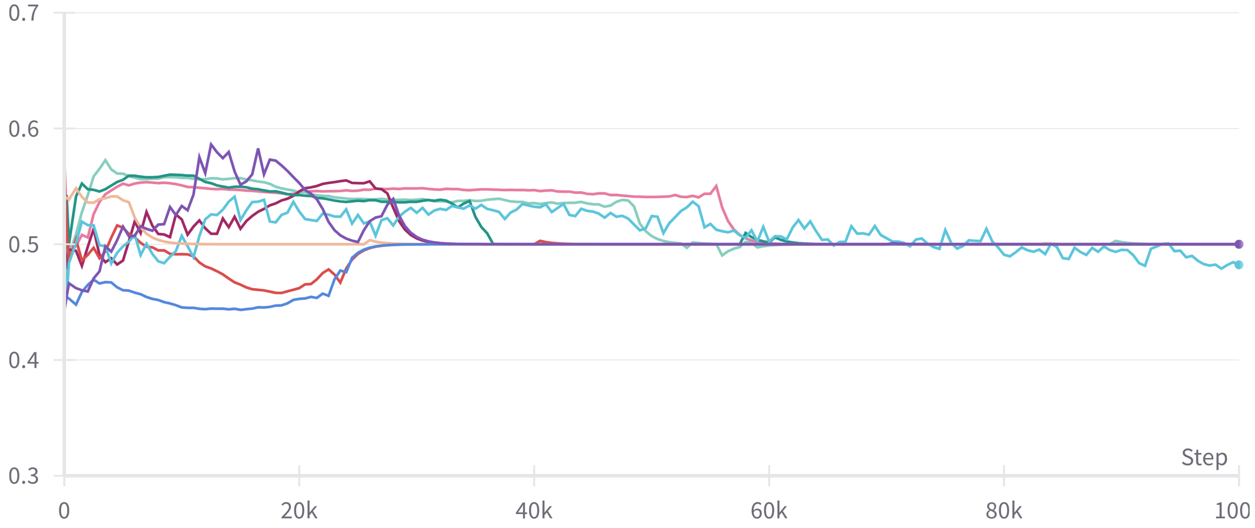


Figure 6. Validation ROCAUC score for the second fold as a function of the number of updates.

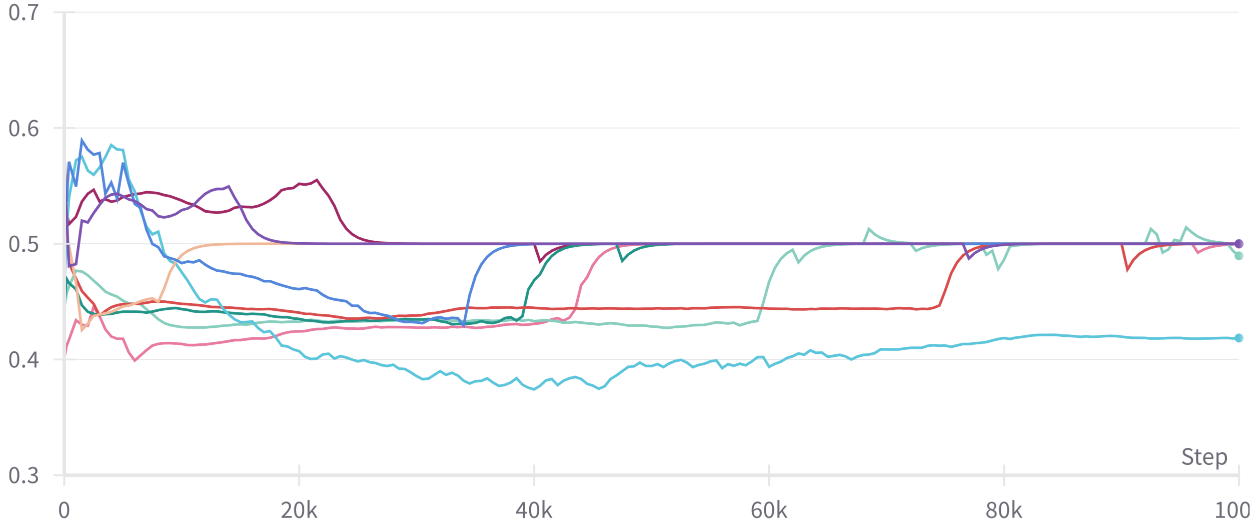


Figure 7. Validation ROCAUC score for the third fold as a function of the number of updates.

I.1. Attention pooling

Table 14. Bag-level ROCAUC score on the test set, mean ± standard error, at different witness rates over 5 NCV runs for the Attention pooling-based model on the MNIST-MIL dataset.

Witness rate	Vanilla	AM	AT	DPI
0.5 %	0.57 ± 0.05	0.55 ± 0.08	0.92 ± 0.02	0.00 ± 0.00
1.0 %	0.64 ± 0.06	0.62 ± 0.09	0.95 ± 0.02	0.00 ± 0.00
1.5 %	0.67 ± 0.04	0.62 ± 0.11	0.97 ± 0.01	0.00 ± 0.00
2.0 %	0.69 ± 0.04	0.88 ± 0.09	0.99 ± 0.01	0.00 ± 0.00
3.0 %	0.87 ± 0.07	1.00 ± 0.00	1.00 ± 0.00	0.00 ± 0.00

Table 15. Bag-level ROCAUC score on the test set, mean ± standard error, at different training set sizes over 5 NCV runs, at a witness rate of 3% for the Attention pooling-based model on the MNIST-MIL dataset.

Dataset size	Vanilla	AM	AT	DPI
18	0.65 ± 0.09	0.68 ± 0.09	0.98 ± 0.02	0.64 ± 0.06
27	0.79 ± 0.08	0.73 ± 0.17	0.99 ± 0.00	0.78 ± 0.05
54	0.68 ± 0.16	0.89 ± 0.11	1.00 ± 0.00	0.85 ± 0.05
81	0.76 ± 0.09	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.06
108	0.87 ± 0.07	1.00 ± 0.00	1.00 ± 0.00	0.92 ± 0.06

I.2. Additive pooling

Table 16. Bag-level ROCAUC score on the test set, mean ± standard error, at different witness rates over 5 NCV runs for the Additive pooling-based model on the MNIST-MIL dataset.

Witness rate	Vanilla	AM	AT
0.5 %	0.52 ± 0.07	0.73 ± 0.05	0.58 ± 0.06
1.0 %	0.61 ± 0.06	0.76 ± 0.04	0.69 ± 0.10
1.5 %	0.52 ± 0.07	0.78 ± 0.05	0.81 ± 0.02
2.0 %	0.65 ± 0.06	0.80 ± 0.05	0.93 ± 0.00
3.0 %	0.74 ± 0.08	0.87 ± 0.02	0.99 ± 0.01

Table 17. Bag-level ROCAUC score on the test set, mean ± standard error, at different training set sizes over 5 NCV runs, at a witness rate of 3% for the Additive pooling-based model on the MNIST-MIL dataset.

Dataset size	Vanilla	AM	AT
18	0.70 ± 0.09	0.65 ± 0.14	0.97 ± 0.01
27	0.71 ± 0.08	0.77 ± 0.11	0.98 ± 0.02
54	0.71 ± 0.08	0.68 ± 0.20	1.00 ± 0.00
81	0.74 ± 0.10	0.75 ± 0.16	1.00 ± 0.00
108	0.74 ± 0.08	0.87 ± 0.02	0.99 ± 0.01

I.3. Conjunctive pooling

Table 18. Bag-level ROCAUC score on the test set, mean ± standard error, at different witness rates over 5 NCV runs for the Conjunctive pooling-based model on the MNIST-MIL dataset.

Witness rate	Vanilla	AM	AT	DPI
0.5 %	0.50 ± 0.07	0.71 ± 0.04	0.76 ± 0.08	0.70 ± 0.04
1.0 %	0.50 ± 0.08	0.74 ± 0.03	0.90 ± 0.03	0.73 ± 0.04
1.5 %	0.65 ± 0.04	0.75 ± 0.02	0.93 ± 0.02	0.74 ± 0.03
2.0 %	0.54 ± 0.08	0.79 ± 0.04	0.97 ± 0.01	0.77 ± 0.04
3.0 %	0.57 ± 0.11	0.80 ± 0.02	0.99 ± 0.01	0.83 ± 0.03

Table 19. Bag-level ROCAUC score on the test set, mean ± standard error, at different training set sizes over 5 NCV runs, at a witness rate of 3% for the Conjunctive pooling-based model on the MNIST-MIL dataset.

Dataset size	Vanilla	AM	AT	DPI
18	0.58 ± 0.09	0.55 ± 0.13	0.95 ± 0.03	0.68 ± 0.05
27	0.71 ± 0.15	0.65 ± 0.12	1.00 ± 0.00	0.77 ± 0.05
54	0.88 ± 0.10	0.63 ± 0.17	1.00 ± 0.00	0.87 ± 0.05
81	0.69 ± 0.14	0.85 ± 0.08	1.00 ± 0.00	0.90 ± 0.05
108	0.57 ± 0.11	0.80 ± 0.02	0.99 ± 0.01	0.83 ± 0.00

J. Compute Requirements

All the proposed strategies in DeeMILIP do not add trainable parameters to the model. The only additional computation in AT is an optimization step for the instance-level loss since the attention network is used as the pseudo-classifier for probing the instance classes. However, no additional matrix operations are involved in the model itself. The same can be said about AM in terms of having no additional parameters, though there is an extra vector-vector dot product step when combining the model's attention vector with the labels vector from the PDK. On the other hand, in DPI, the entire attention network is actually removed, so there are actually fewer parameters in this case.