Benchmarking Few-Shot Learning in Biomedicine: Insights from Cell Classification and Protein Function Prediction

Mika Senghaas Ludek Cizinsky Adam Barla

Abstract

Our study rigorously evaluates prominent few-shot learning algorithms, including Baseline, Matching Networks, Prototypical Networks, and MAML, in two few-shot classification tasks based on the Tabula Muris and SwissProt datasets. All algorithms demonstrate substantial efficacy, with the best performing algorithms achieving 69.1% and 91.3% accuracy on the SwissProt and Tabula Muris datasets, respectively. Additionally, we found that incorporating the Self-Optimal Transport (SOT) feature transform module enhances performance in most settings with minimal computational overhead, increasing the best performance to 70.0% (+0.9) and 93.5% (+2,2) on the SwissProt and Tabula Muris datasets, respectively. *The code and experiments are available on GitHub and W&B.*

1. Introduction

Learning from few samples remains a major challenge in machine learning, especially in biomedical tasks were data availability is often a limiting factor due to high costs of data collection and expert-dependent annotation processes. Traditional machine learning approaches often fall short in these data-limited settings as they require extensive training samples and iterations. Few-shot learning algorithms, tailored to discern distinct features from minimal data, offer a promising alternative. This study aims to assess the applicability of such methods in the biomedical domain. We examine four established few-shot learning techniques — Baseline(++) (Chen et al., 2019), Matching Networks (Vinyals et al., 2016), Prototypical Networks (Snell et al., 2017), and Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) — across two distinct biomedical tasks. The first involves predicting cell types based on gene expression, using the Tabula Muris dataset (Schaum et al., 2019). The second task focuses on predicting protein functions from their sequence embeddings, using the SwissProt dataset (Consortium, 2019).

All of the above approaches rely on meaningful embeddings of features. However, a notable challenge arises due to the potential discrepancy in data distributions between samples seen during meta-training and samples of novel classes seen during meta-testing. This discrepancy may result in embeddings that are not fully transferable, leading to suboptimal performance in downstream tasks. Self-Optimal-Transport (SOT) (Shalam & Korman, 2022) is a feature transform module that aims to mitigate this issue. We include the SOT feature transform module in all of the above mentioned few-shot learning methods to study its effectiveness in the biomedical domain.

In summary, our study presents three primary contributions. First, we train and assess leading few-shot learning algorithms on two unique biomedical tasks. We report the performance of these algorithms with and without the SOT module. Second, we study the performance in varying few-shot learning settings by differing the number of classes to distinguish and number of samples per class to learn from. Third, we ablate the hyperparameters of all models that include the SOT feature transform to understand their impact on performance. We hope that our study will serve as a benchmark for future research in the biomedical domain.

2. Data

The first dataset, **Tabula Muris** (Schaum et al., 2019) (denoted as `TM`), comprises over 100,000 mouse cells' gene expression data and annotations about the cell ontology class (cell type). The task is to predict the cell type based on the gene expression data. To address the sparsity and skewed distribution of raw gene expressions, preprocessing included gene and cell filtering, log-transformation, and mean normalisation, with zero imputation. Post-processing, the dataset features 105,960 cells across 125 cell types.

For the few-shot learning task, the focus is on generalisation across different tissues. The dataset is divided into training, validation, and testing splits, each representing distinct tissue types: 15 for training and four each for validation and testing. Despite some overlap in cell types across tissues this structure ensures diverse tissue representation and makes the task of cell type prediction in cells from novel tissues challenging.

The second dataset, **SwissProt** (Consortium, 2019) (denoted as SP), is an extensively annotated protein sequence database featuring 14,251 sequences, enriched with comprehensive information on their functions, structures, and biological roles. This project utilises pre-computed sequence embeddings obtained from ESM-2 (Lin et al., 2022), a state-of-the-art protein language model, as input data. The goal is to predict proteins' functions. In total, there are 884 unique annotated protein functions. The dataset is divided into three splits with no overlap in the targets.

3. Methods

Few-shot learning algorithms can be classified into two main categories: *Transfer learning* involves a two-phase process of pre-training on a large dataset to learn general data representation, followed by fine-tuning on the target task with limited data. *Meta-learning*, conversely, leverages past experiences from a series of related tasks to efficiently tackle a new task with sparse data. These algorithms undergo meta-training, where the model encounters various tasks, mimicking the target few-shot learning scenario. Each task comprises randomly chosen support and query samples from identical class sets, training the model to adapt to support samples and classify query samples.

In the following we describe the high-level idea of all methods considered within this study. For more details on the methods please refer to the original papers.

The **Baseline** (Chen et al., 2019) model implements the fine-tuning paradigm. During meta-testing, the model is fine-tuned on support samples and then classifies query samples. Within this study, we consider two variants of the trainable classification head - one learns a traditional linear and the other a cosine similarity layer. We refer to these as Baseline (B) and Baseline++ (B++), respectively.

Prototypical Networks (Snell et al., 2017) (PN) learn an embedding space that clusters samples from the same class close together. Query samples are then classified according to the distance to the average support sample (prototype) of each class.

Matching Networks (Vinyals et al., 2016) (MN) are similar to PN. However, in MN, the distance between a query sample is computed to all support samples and then aggregated. Importantly, before the distance computation, MN contextualises both support and query samples by re-embedding them using an LSTM.

Finally, **Model Agnostic Meta Learning** (Finn et al., 2017) (MAML) is an optimisation-based meta-learning approach that aims to learn an effective weight initialisation that can be adapted to new tasks in a small number of gradient steps.

The **Self-Optimal Transport** (SOT) (Shalam & Korman, 2022) feature transform is a parameterless and fully differentiable method for transforming feature vectors. SOT embeddings are notable for their interpretability and potential to upgrade a set of features to facilitate downstream matching or grouping related tasks, as encountered frequently in few-shot learning settings.

SOT fundamentally utilises Optimal Transport (OT) on a square distance matrix (e.g. cosine similarity matrix) of input features, leading to embeddings that reflect the *direct* similarity and *third-party* agreement of samples to each other. Mathematically, SOT is a function $T : \mathbb{R}^{n\times d} \rightarrow \mathbb{R}^{n\times n}$ that maps n samples in d-dimensions to a re-embedded SOT vectors in n-dimensions. The SOT embeddings are computed from an iterative optimisation algorithm known as the Sinkhorn-Knopp algorithm (Sinkhorn & Knopp, 1967) that solves a regularised version of the OT problem.

In few-shot learning contexts, SOT helps align independently embedded support and query samples by jointly embedding them according to their similarities to each other - an example of *transductivity*. The SOT feature transform is used in state-of-the-art methods in common few-shot learning benchmarks (Shalam & Korman, 2022). Within our study we employ the SOT feature transform module on the embeddings obtained from the backbone network. Critically, we shuffle the query samples before the forward-pass to avoid learning a trivial mapping from sample position to class label.

4. Methodology

An experiment in our study is defined as a combination of a few-shot learning *method*, optionally including the *SOT* feature transform, trained and evaluated on a *dataset* within a specified few-shot learning setting, characterised by the number of classes (*n-way*) and the number of samples per class (*n-shot*).

4.1. Experiment Setup

Backbone. All experiments employ a fully-connected feed-forward neural network with batch normalisation, ReLU activation, and dropout. The network has two hidden layers, with hidden dimensions being tuned for each experiment.

Training. Training of the models is conducted for a maximum of 40 epochs, employing the Adam optimiser with varying learning rates. We implement early stopping after five epochs of no improvement in validation accuracy.

Tuning. Extensive hyperparameter tuning is performed for all models that include the SOT module, unless specified otherwise. Tuning includes the learning rate ($\lambda = 1e^{-x} \forall\ x \in [-5, -1]$) for all methods as well as the backbone's hid-

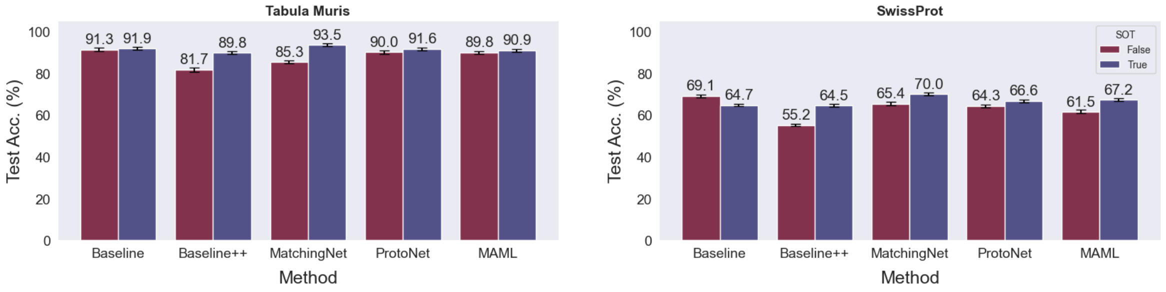

Figure 1. **Benchmark Results.** Test accuracy of all methods on TM (left) and SP (right) in the 5-way-5-shot setting. The plot shows the mean accuracy over 600 episodes and the 95% confidence interval.

den dimension size ($\kappa = \{64, 512, 1024\}$). For models including the SOT module, we adapt the hyperparameter grid of Shalam & Korman, namely the regularisation parameter ($\gamma = \{1.0, 0.1, 0.01\}$) and the choice of distance metric ($\delta = \{cosine, euclidean\}$). Whenever it was clear that further grid search would not yield better results, we stopped the tuning process. The model demonstrating the best performance on the validation split is evaluated on the test split.

Evaluation. A model performance's is reported through the mean and 95% confidence interval of the few-shot accuracy, calculated over 600 episodes with each episode utilising five query samples per class.

4.2. Experiments

Due to the impracticality of exhaustive hyperparameter grid searching across all experimental configurations, we have structured our experiments into two distinct groups. Each group fixes certain parameters, allowing us to focus on the impact of the variables of interest.

General Benchmark. This experiment group evaluates models, with and without SOT feature transform, on both datasets in a 5-way-5-shot setting, comprising 20 experiments. The aim is to analyse the influence of the method, dataset, and SOT module on few-shot learning performance.

Way-Shot Analysis. In the second group we investigate the performance in various few-shot learning settings, exploring combinations of n-way (2, 4, 6, 8, 10) and n-shot (1, 5, 10, 15, 20). Here, we fix the method to MN and the dataset to TM, resulting in 50 experiments. For this experiment, we do not fine-tune the hyperparameters but instead use the best-performing hyperparameters from the general benchmark experiment.

5. Results

Figure 1 and Table 1 shows the test accuracies for each method with and without the SOT module on both datasets.

Table 1. **Benchmark Results**. Test accuracy of all methods on TM and SP in the 5-way-5-shot setting. We depict the average accuracy and the 95% confidence interval both without (left) and with SOT (right) and the difference.

		Test Accuracy (%)		
		w/o SOT	w/ SOT	Diff
TM	B	$\mathbf{91.3 \pm 0.6}$	91.9 ± 0.7	**0.6**
	B++	81.7 ± 0.9	89.8 ± 0.7	**8.1**
	MAML	89.8 ± 0.7	90.9 ± 0.6	**1.1**
	MN	85.3 ± 0.8	$\mathbf{93.5 \pm 0.6}$	**8.2**
	PN	90.0 ± 0.7	91.6 ± 0.6	**1.6**
SP	B	$\mathbf{69.1 \pm 0.7}$	64.7 ± 0.9	-4.4
	B++	55.2 ± 0.7	64.5 ± 0.8	**9.3**
	MAML	61.5 ± 0.7	67.2 ± 0.7	**5.7**
	MN	65.4 ± 0.7	$\mathbf{70.0 \pm 0.7}$	**4.6**
	PN	64.3 ± 0.7	66.6 ± 0.8	**2.3**

Models without the SOT module reach an average accuracy of 88% on the TM dataset and 63% on the SP dataset. These results generally show that the models are capable of learning from few samples, improving significantly over the random baseline of 20% accuracy. Peak performances in this group are achieved by B on both datasets, with 91% and 69% accuracy, respectively.

Models incorporating the SOT module exhibit enhanced accuracy, achieving 90% and 67% average accuracy on the TM and SP dataset. All models benefit from the inclusion of the SOT feature transform module, except for B on SP.

The addition of the SOT module increases the peak accuracy by 0.9 percentage points on SP and 2.2 on TM. Both achieved by MN with 70% and 93% accuracy, respectively.

Hyperparameter Ablation. Figure 2 shows the pair-wise interactions between all tuned hyperparameters for both datasets separately. We add a column for the methods to study how they react to certain hyperparameters configurations.

The plots indicate a clear trend across all methods and

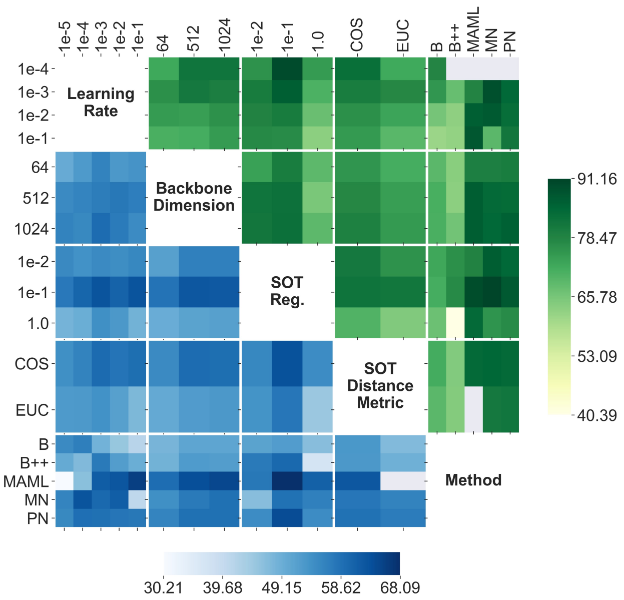

Figure 2. **Hyperparameter Ablation.** Average test accuracies on the SP dataset for all pairs of hyperparameter settings for SP in blue (bottom left) and TM in green (top right). The grey areas were not explored during tuning.

datasets regarding the SOT hyperparameters: We find that using the cosine distance metric and regularisation with $\gamma = 0.1$ consistently yields the best performance. Furthermore, we find that the choice of the learning rate has a significant impact on the performance of all methods. B, B++ and MN generally prefer lower learning rates, while MAML in contrast prefers higher learning rates. PN's performance is robust to the choice of learning rate showing no significant variation across tuning runs. Finally, the plot reveals that MN, PN, and MAML are generally more robust to the choice of hyperparameters, yielding higher mean accuracies across tuning runs, while B++ was found especially sensitive.

Way-Shot Analysis. Figure 3 illustrates PN's way-shot analysis on the TM dataset, comparing scenarios with and without the SOT module. The left subplot depicts test accuracy versus the number of classes (ways), while the right subplot relates accuracy to the number of samples per class (shots). In both SOT and non-SOT contexts, a consistent trend emerges: accuracy diminishes linearly with more classes and grows with additional samples per class up to some limit. Notably, having access to more than ten samples per class yields no substantial accuracy gains. As expected, the model's performance with the SOT module is consistently higher for higher numbers of classes and samples.

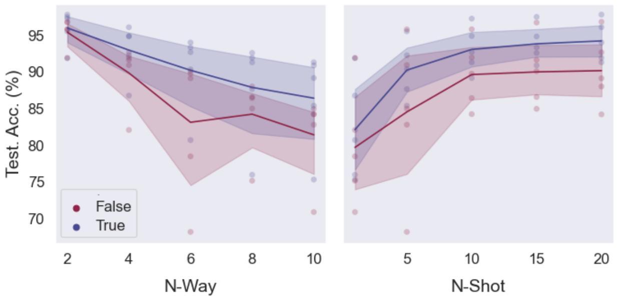

Figure 3. **Way-Shot Analysis.** Test accuracy of PN on the TM dataset with and without the SOT module in various few-shot learning settings for fixed n-way (left) and n-shot (right). Individual points represent a single experiment. We show the regression line with a 95% confidence interval.

6. Discussion

Our analysis underscores the competence of all methods in extracting knowledge from few-shot biomedical datasets. Notably, the TM dataset seems to be more amenable to few-shot learning, with all methods achieving an accuracy of at least 80%. In contrast, the SP dataset poses a more challenging task, with the highest performance achieved being 70% accuracy. This discrepancy may stem from variations in the quality of the features or the potential disparities in the overlap of classes in the training and test sets observed in the TM dataset.

Across almost all methods and few-shot settings examined, the SOT module consistently enhances performance. This is in line with the claims and empirical findings of the original paper, making SOT a model-agnostic plug-in module with potential to improve performance in a wide range of few-shot learning tasks with little added computational overhead.

Our study shows conclusive evidence that the SOT module performs best in few-shot classification tasks for a regularisation parameters $\gamma = 0.1$ and using the cosine distance metric. However, careful tuning of other hyper-parameters, such as the learning rate or the size of the hidden dimension of the backbone network, are crucial to achieve peak performance and vary depending on the method and dataset.

7. Conclusion

In summary, we demonstrate the effectiveness of both transfer and meta-learning algorithms for learning effective feature representations for classification tasks with limited data in the biomedical domain. Notably, employing the transductive SOT feature transform emerged as an effective approach, consistently improving performance across models, datasets and few-shot settings. Extensive hyperparameter tuning was found crucial to achieve optimal performance.

8. Appendix

8.1. Dataset Statistics

Table 2. **Dataset Statistics.** The Table shows the number of samples and targets in each split of the two datasets.

	Split	#Samples (%)	#Targets	Overlap
TM	Train.	65,846 (62%)	59	N/A
	Val.	15,031 (14%)	47	10
	Test.	25,083 (24%)	37	4
SP	Train.	12,141 (85%)	636	N/A
	Val.	1,407 (10%)	159	0
	Test.	703 (5%)	89	0

8.2. Way-Shot Analysis Heatmap

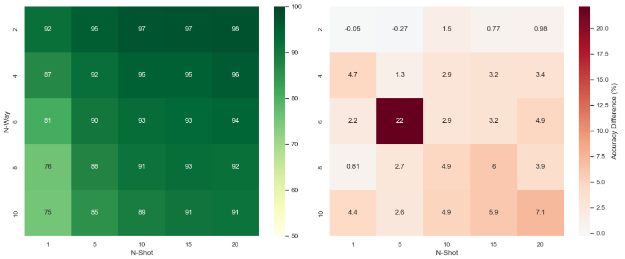

Figure 4. **Way-Shot Analysis Heatmap.** Test accuracy of PN with SOT on the TM for varying numbers of classes and samples per class (left). The right plot shows the difference to the same method without SOT.

References

Chen, W., Liu, Y., Kira, Z., Wang, Y. F., and Huang, J. A closer look at few-shot classification. *CoRR*, abs/1904.04232, 2019. URL http://arxiv.org/abs/1904.04232.

Consortium, T. U. Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Research*, 47(D1):D506–D515, 2019. doi: 10.1093/nar/gky1049. URL https://doi.org/10.1093/nar/gky1049.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *CoRR*, abs/1703.03400, 2017. URL http://arxiv.org/abs/1703.03400.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Schaum, N., Karkanias, J., Neff, N. F., May, A. P., Quake, S. R., Wyss-Coray, T., Darmanis, S., Batson, J., Botvinnik, A., Chen, M., et al. Tabula muris: single-cell transcriptomics of 20 mouse organs. *Genome Biology*, 20(1): 1–16, 2019. doi: 10.1186/s13059-019-1824-x.

Shalam, D. and Korman, S. The self-optimal-transport feature transform. *arXiv preprint arXiv:2204.03065*, 2022.

Sinkhorn, R. and Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Snell, J., Swersky, K., and Zemel, R. S. Prototypical networks for few-shot learning. *CoRR*, abs/1703.05175, 2017. URL http://arxiv.org/abs/1703.05175.

Vinyals, O., Blundell, C., Lillicrap, T. P., Kavukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. *CoRR*, abs/1606.04080, 2016. URL http://arxiv.org/abs/1606.04080.