nature biomedical engineering

Article

https://doi.org/10.1038/s41551-023-01049-7

Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging

In the format provided by the authors and unedited

In the following sections, we report further experiments and analysis to understand the performance of REMEDIS. This includes ablation studies to analyze the benefits of the different components underlying REMEDIS, comparison with several supervised baselines, and other approaches to leveraging unlabelled data. Given the sensitivity of the medical domain and the importance of ensuring AI development methods do not propagate existing health equity disparities, we conduct a detailed subgroup analysis in the dermatology and mammography setting. Furthermore, we also include results on a non-classification task and visualize the learned representations and list detailed t-test statistics for all our experiments.

Supplementary figures

Supplementary tables

Ablation studies

Performance of standard supervised transfer-learning Strategy. In addition to the strong supervised baseline discussed in Fig. 3 we also evaluate our method against the standard supervised strategy which is defined as transfer learning using models pretrained on 1M natural images from ImageNet-1K dataset. Supplementary Fig. 1 shows the overview of the results demonstrating overall performance and data-efficient generalization of the proposed self-supervised learning strategy, REMEDIS in comparison to the standard supervised baseline pre-trained on ImageNet-1K and fine-tuned for the specific medical task. We observed significantly improved out-of-distribution generalization and significant reduction in the need for labelled medical data when using our proposed approach. REMEDIS exhibits significantly improved in-distribution performance with up to 11.5% relative improvement in diagnostic accuracy in comparison to the standard supervised baseline. Furthermore, REMEDIS leads to strong data-efficient generalization, matching the standard supervised baseline using 1% to 31% of retraining data from new clinical development settings across tasks. This can translate to a significant reduction in the retraining data and time required to deploy medical AI at scale and accelerate the development life cycle of these AI models.

Contribution of large-scale pre-training data. In Figure 3, Supplementary Fig. 1, Supplementary Fig. 18, Supplementary Fig. 19, Supplementary Fig. 16, and Supplementary Fig. 15, we report the comparison of REMEDIS with the widely used supervised pre-training baseline. To investigate the contribution of large-scale pre-training data, here, we also separately compare and contrast the overall performance of REMEDIS vs. these baselines.

The strong supervised baseline and REMEDIS both pre-trained on JFT-300M (BiT-L) and relies on large-scale pre-training while the standard supervised baseline has been trained on a much smaller dataset containing only 1M images. Supplementary Fig. 2 demonstrates overall in-distribution performance of REMEDIS as well as the supervised baselines trained using both ImageNet-1K and JFT-300M. Moreover, we observe that the strong supervised baseline (BiT-L) can provide significantly better in-distribution performance against the standard supervised baseline (BiT-S), showing the benefits of large-scale training as it has been reported in [4].

As discussed, the large-scale supervised pre-training (BiT-L) represents a strong supervised baseline for medical imaging [4]. Supplementary Fig. 3 shows the overview of the results demonstrating data-efficient generalization of our proposed self-supervised learning strategy, REMEDIS as well as the standard and strong supervised baseline pre-trained using both ImageNet-1K and JFT-300M. We observed significantly improved out-of-distribution generalization and significant reduction in need for annotated medical data when using our proposed approach. Moreover, comparing the data-efficient generalization of the strong and standard pretrained models, often strong supervised baseline performs significantly better than the standard supervised baseline in out-of-distribution regime. Meanwhile REMEDIS holds a steady significantly better performance against both strong and standard baseline, in some of the tasks such as Dermatology classification (T_1) and DME (T_2) the superior performance of strong supervised baseline in the data-efficient generalization regime against the standard baseline can show an unexpected pattern.

Both Supplementary Fig. 2 and Supplementary Fig. 3 indicate that large-scale supervised pre-training is a strong component and is a good starting point for developing medical imaging models [4]. To further investigate the specific significance of large-scale pretraining and how the choice of supervised pre-training impacts REMEDIS, in a new set of experiment we adapt BiT-S, M, and L as our based network and performs the self-supervised training on medical data on top of each of these models. The default REMEDIS method uses BiT-L as the base-network.

This direct comparison is only completed on the dermatology task, due to the high computational cost. The results, shown in Supplementary Fig. 4, show that the best results tend to be achieved using BiT-L, but that BiT-M can perform competitively. Given that BiT-M is openly available and is not trained on proprietary data, we believe that this is a further good indication that large-scale supervised pre-training is valuable and hope the wider medical AI community leverages this to build medical imaging models.

Contributions of BiT-L and self-supervised learning. While our general focus in this study has been to compare REMEDIS with supervised baselines (Fig. 3 and Supplementary Fig. 1), it is also of interest to understand the contributions of the representation learning strategies underlying REMEDIS and to

demonstrate the need for supervised pre-training. To this end, we ran ablation studies in which we investigated and disentangled the contribution of the large-scale pre-training on natural images and self-supervised representation learning on medical images as well as the specific architecture choices. For a fair comparison, in each case, we follow the same pre-training and fine-tuning protocol as a sour method to optimize these models.

Both SimCLR and BiT-L provide benefits over the widely used supervised JFT pre-training baseline for most tasks in both in- and out-of-distribution settings (see Supplementary Fig. 5). While for the larger architecture such as ResNet-152 (2×), BiT can provide performance gains approximately comparable to REMEDIS in some cases, this is not consistent across architectures as well as all the medical imaging tasks considered. This is also aligned with previous observations in [6]. Note that SimCLR results for mammography and DME classification are missing due to the high computational cost.

In addition, we also repeat these experiments with the MoCo variant of REMEDIS as an alternate state-of-the-art self-supervised learning method. Supplementary Fig. 6 and Supplementary Table 8 compare the MoCo variant of REMEDIS with a strong supervised baseline pre-trained on JFT data as well as pre-training using only MoCo and medical data. These results highlight the importance of self-supervision components and the superiority of REMEDIS over its component building blocks.

Comparison with self-training

REMEDIS leverages self-supervised pre-training to make use of large amounts of unlabelled medical data for learning high-quality representations. However, other approaches enable models to learn from unlabelled data. One such approach is self-training [2, 3]. In a typical self-training setting, a teacher network trained on labelled data predicts labels on the unlabelled data. Then, a second student model is fine-tuned on the original labelled data, as well as the predicted labels. We implement this by training a model on D , inferring on D_U, and then re-training the model on D and D_U. We use the predicted probabilities on D_U as soft labels and separately sweep over hyper-parameters for the teacher and student training. We otherwise do not vary the training set-up. Due to computational constraints and early evidence of superiority of REMEDIS, we only report these numbers for dermatology.

We performed this self-training cycle using models that start from BiT-L, as the most direct comparison to REMEDIS, as well as with models that start from the standard supervised baseline and the corresponding REMEDIS variant. The results, shown in Supplementary Fig. 7, indicate that self-training can produce high quality models on par with REMEDIS especially when using the larger model architecture. However, this is not consistent when using smaller architecture sizes. Self-training can also degrade performance when applied from the standard supervised baseline, perhaps because self-training relies on the quality of the teacher model. Furthermore, self-training requires the representation learning task and the downstream task to be well aligned while contrastive pre-training is agnostic to the downstream task leading to representations that can be generally applied. Nevertheless, we believe self-training is a promising approach and should be considered when appropriate for developing medical imaging AI using unlabelled data.

Performance analysis across subgroups

Given the importance of fairness in AI, when using pre-trained representations for developing medical imaging AI, we are interested in approximate parity of performance across target subgroups of interest so as to ensure the models are not amplifying existing health disparities. More specifically, for the deployment of such models in clinical settings, it is important to evaluate them comprehensively across protected subgroups. This can be of particular concern when leveraging large-scale pre-training datasets, as they may be biased towards certain subgroups without the knowledge of the model developer [6]. Thus, we also investigated the performance distribution across different subgroups of interest. We focused on subgroups in the dermatology and mammography task where we have access to metadata to disentangle them. We are particularly interested in how the introduction of large-scale pretraining, or self-supervised pre-training, affects performance across these clinically relevant subgroups.

In dermatology, we established subgroups based on age and biological sex. For sex, we considered a binary setting and compared 2,564 and 1,505 cases for the in-distribution dataset, and 3,153 and 3,486 cases of each sex for the out-of-distribution dataset. For age, we divided the data into four age groups of 18-30, 30-45,45-65, and 65+ years, which include 1,185, 1,162, 1,495, and 226 cases respectively for the in-distribution data and 186, 702, 2,560, and 3,181 cases for the out-of-distribution dataset. We compared the top-3 accuracy across these groups using the standard and strong supervised baseline and REMEDIS. We observed that while the baseline supervised pre-trained model performance drops on some subgroups, using intermediate self-supervised pre-training, the model performance is more even across the different subgroups (see Supplementary Fig. 8 (a)). This exploratory experiment suggests that the learned representations are likely general, and in most of the cases neither pick up spurious correlations during pretraining nor are they biased towards particular subgroups.

For the mammography classification task, we compared subgroups based on age and breast density (which can be correlated with age and ethnicity). The test data is divided into four age groups of 30-45, 45-65 and more than 65+ years of age which include 0, 9,901, and 2,547 cases, respectively, for the in-distribution data and 2,963, 7,109, and 109 cases for the out-of-distribution dataset. The four different density categories include 585, 3,606, 2,314, and 957 cases for the in-distribution dataset, and 109, 612, 741, and 71 cases for the out-of-distribution dataset. Density level of four is associated with a denser breast based on BI-RADS [116] assessments. The results, shown in Supplementary Fig. 8 (b), show the distribution of performance across these subgroups. With the exception of a couple of the breast density categories, REMEDIS consistently improves performance. We believe that these subgroup performance disparities are unlikely to be caused by intrinsic biases in the pretraining mechanism, but future work should investigate specific pre-training strategies such as dataset re-sampling to mitigate performance drops.

Mammography-localization results

In addition to classification tasks, we also evaluated our method and the baseline on a localization task. For this purpose, we considered the cancer localization task in mammography images (T7). In this task, the goal is to localize cancerous lesions. This task is evaluated using the mean average precision (mAP). Matches between ground truth and predicted bounding boxes are considered positive when the intersection-over-union (IOU) is higher than 10%. The pretraining setup and data were identical to T_6's. Due to the computational complexity of training these models, we report only partial results. For training the localization model, only positive cases were included. The labels in this task consisted of the coordinates of the bounding boxes derived by human radiologists a-posteriori having access to all mammograms, biopsy results, and radiology text reports. The pretrained CNN backbones were the same as inT6, but an additional feature pyramid network was added on top of the CNN [8].

The dataset contains 3,727 cases, including 5,854 ground truth bounding boxes across all mammograms. 2,909 cases were used for training, 158 for tuning, and 660 for test. When using REMEDIS, the model shows significant improvement over the baseline, moving from a mAP of 0.805 to a mAP of 0.855. We used the Adam optimizer with an exponential learning rate decay in breast cancer localization task where we performed a rigorous grid search to select the initial learning rate, decay steps, and decay factor. All of the models were trained for a maximum of 200K steps. For this task, scaled 2048×2048 pixels mammography images go through the augmentation process including random flipping, random shifting, and random color distortion. We selected the learning rate, decay steps, and decay factor after a grid search of three logarithmically spaced learning rates between $10^{-5.0}$ and $10^{-4.0}$ and three decay steps in $\{10K, 25K, 50K\}$, and three decay factor in decay steps in $\{0.1,0.25,0.5\}$.

t-SNE Visualization of representations

To gain more insight into the high dimensional embedding representations learned by models considered in this study, we use t-SNE visualization [158]. For this purpose we focus on the pathology metastases task (T_4) which has a binary label space and includes two out-of-distribution datasets. The t-SNE visualization of the best REMEDIS and best supervised model representation embeddings obtained from the test in-distribution and out-of-distribution examples of the pathology metastases task (T_4) are depicted in Supplementary Fig. 10 These models are only fine-tuned with the in-distribution train dataset and not the out-of-distribution data. The binary labels of this task are color-coded. The visualizations (Supplementary Fig. 10) qualitatively indicate that clusters associated with each class are better separated in the REMEDIS feature space compared to the supervised baseline.

Momentum contrastive learning

REMEDIS leverages the generic form of contrastive pre-training based on SimCLR [10] to learn medical domain-specific representations. Although there have been several studies investigating self-supervised representation learning for medical imaging AI applications (such as [11]), these studies consider a limited number of modalities and they do not consider how they might be combined with other representation learning strategies and often rely on task-specific design choices. In contrast, our study is comprehensive, and we demonstrate here that the core self-supervised learning method in REMEDIS can be modified with other strong alternatives.

Supplementary Fig.12 and Supplementary Fig. 13 show detailed performance of the MoCo variant of REMEDIS where the generic self-supervised learning method is replaced with improved MoCo [12]. Overall, our results indicate that REMEDIS is compatible with momentum contrastive learning as an alternative self-supervised learning technique as REMEDIS still results in data-efficient generalization and introduces significant performance improvements over the strong supervised baseline.

Supplementary Fig.11, Supplementary Fig. 14, Supplementary Table 9 and Supplementary Table 10 provide a direct comparison between SimCLR and MoCo variants of REMEDIS vs. the strong and standard supervised baseline. Overall, in this experiment, we do not observe any significant improvement for task T_1 when using the MoCo variant over the SimCLR variant. Meanwhile, in task T_2, SimCLR variant results are still significantly better than the MoCo variant. Thus, the translation of improved MoCo components to the improved discriminative properties of self-supervised pre-training in the medical domain is inconclusive.

Additional experimental results
The following figures show additional experimental results that compare the in-distribution and out-of-distribution performance of REMEDIS vs. the supervised baseline in further detail. Specifically, in this section, we investigate the performance of models for both architectures ResNet-50 (1×) and ResNet-152 (2×), and multiple additional out-of-distribution datasets for certain tasks. This section provides the followings supplementary figures:

- Supplementary Fig. 15 provides detailed in-distribution and zero-shot out-of-distribution performance for all datasets grouped by network architectures and compares REMEDIS vs. the strong supervised baseline trained on JFT-300M dataset.
- Supplementary Fig. 16 provides detailed in-distribution and zero-shot out-of-distribution performance for all datasets grouped by network architectures and compares REMEDIS vs. the standard supervised baseline trained on ImageNet-1K dataset.
- Supplementary Fig. 18 provides detailed data-efficient generalization results using common axes range for visualization and grouped by network architecture and compares REMEDIS vs. the strong supervised baseline trained on JFT-300M dataset.
- Supplementary Fig. 19 provides detailed data-efficient generalization results using a common axes range for visualization and grouped by network architecture and compares REMEDIS vs. the standard supervised baseline trained on ImageNet-1K dataset.
- Supplementary Tables 11 – 20 provide detailed performance values for zero-shot out-of-distribution gains using REMEDIS and the supervised baseline.
- Supplementary Tables 21 – 32 list detailed t-test statistics for all experiments including REMEDIS, baselines, and multiple ablation studies.
- Supplementary Fig. 17 shows a breakdown of in-distribution gains vs. zero-shot out-of-distribution gains using REMEDIS and the supervised baseline.

Specifically, Supplementary Fig. 15 and Supplementary Fig. 16 provide in-distribution performance gains vs. zero-shot out-of-distribution performance gains using REMEDIS and supervised baselines. Unlike previous visualizations in Figure 3, the results are grouped based on the base network architecture, not the best overall performing model for each task. In all plots, the 95% confidence intervals were calculated by running each experiment ten times and are shown using the error bars.

In addition, we also provide additional zero-shot out-of-distribution results for multiple tasks using additional out-of-distribution datasets in these figures. This includes: (1) an additional out-of-distribution dataset for diabetic macular edema classification (T_2) which includes 323 de-identified fundus images collected in India, (2) a non-overlapping fraction of the CAMELYON-17 dataset for pathology metastases detection (T_4), which includes 273 pathology slides that do not appear in CAMELYON-16 pathology, or the original CAMELYON-17 dataset. These datasets are considered small-scale and are not suitable for data-efficient generalization evaluations; for example, they contain only 2-3 examples at 1% data fraction which is not enough to capture all possible variability in the corresponding tasks and also not relevant in real world deployment settings. Due to computational limits, we were not able to train models using the ResNet-152 (2×) architecture for pathology tasks. These results also suggest that the ResNet-152 (2×) architecture often leads to the highest performance.

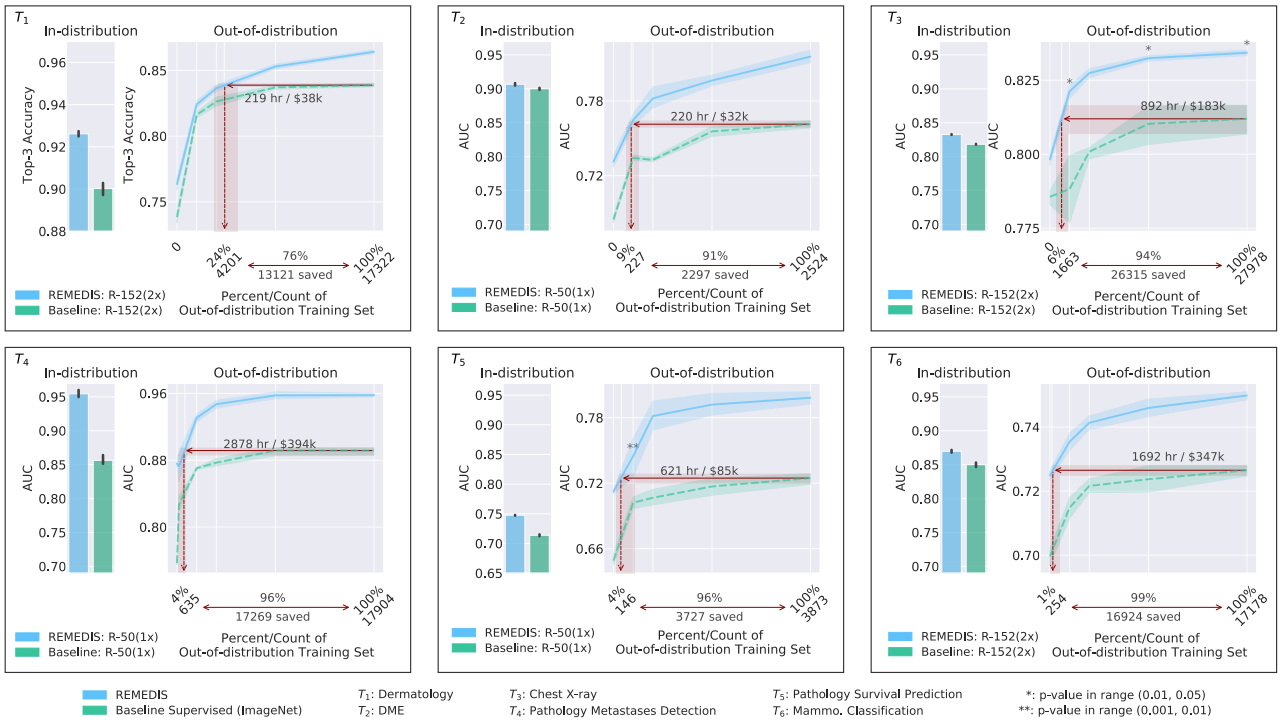
Supplementary Fig. 17 shows the relationship between in-distribution vs. zero-shot out-of-distribution performance using REMEDIS and the supervised baseline. As discussed, 95% confidence intervals of our experiments were calculated by running each experiment ten times. Each point in this plot corresponds to one of these repeated runs and the coordinates were obtained by calculating the in-distribution and zero-shot out-of-distribution for the target task. These plots confirm that dataset shift greatly impacts the performance of models when evaluated on the out-of-distribution dataset. However, our results suggest that REMEDIS improves out-of-distribution performance without decreasing in-distribution performance, and REMEDIS has higher performance for both in-distribution and out-of-distribution data.

Lastly, Supplementary Fig. 18 and Supplementary Fig. 19 show results demonstrating data-efficient generalization of our method vs. the strong supervised baseline pretrained on JFT-300M and also the standard supervised baseline pre-trained onImageNet-1K. Unlike the previous visualizations in Fig. 3 and Supplementary Fig. 1, these graphs were scaled based on a unified performance range axes and the results are grouped based on the base network architecture, not the best overall results for a given task. In particular, each graph depicts performance (measured by top-3 accuracy or area under the curve) when

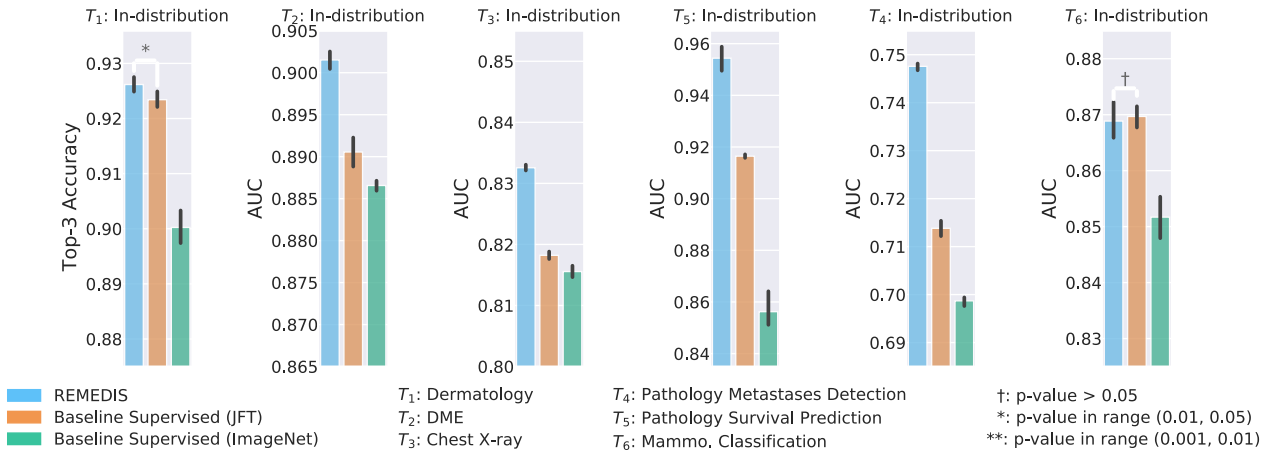
using different data fractions/data counts of out-of-distribution data to fine-tune the model for the dermatology condition classification (T_1), diabetic macular edema classification (T_2), chest X-ray condition classification (T3), pathology metastases detection (T_4), pathology colorectal survival prediction (T_5), and mammography classification (T6) as well as two architectures ResNet-50 (1×) and ResNet-152 (2×). We also calculate a 95% confidence interval by running each label fraction and experiment ten times and intervals are shown using the shaded area and error bars. A two-sided t-test was also calculated for each label fraction as well as in-distribution results and p-value for several thresholds comparing any significant improvement of our method against these baselines. We observe significantly ($p < 0.05$) improved out-of-distribution generalization and a significant reduction in the need for labelled medical data when using REMEDIS.

References

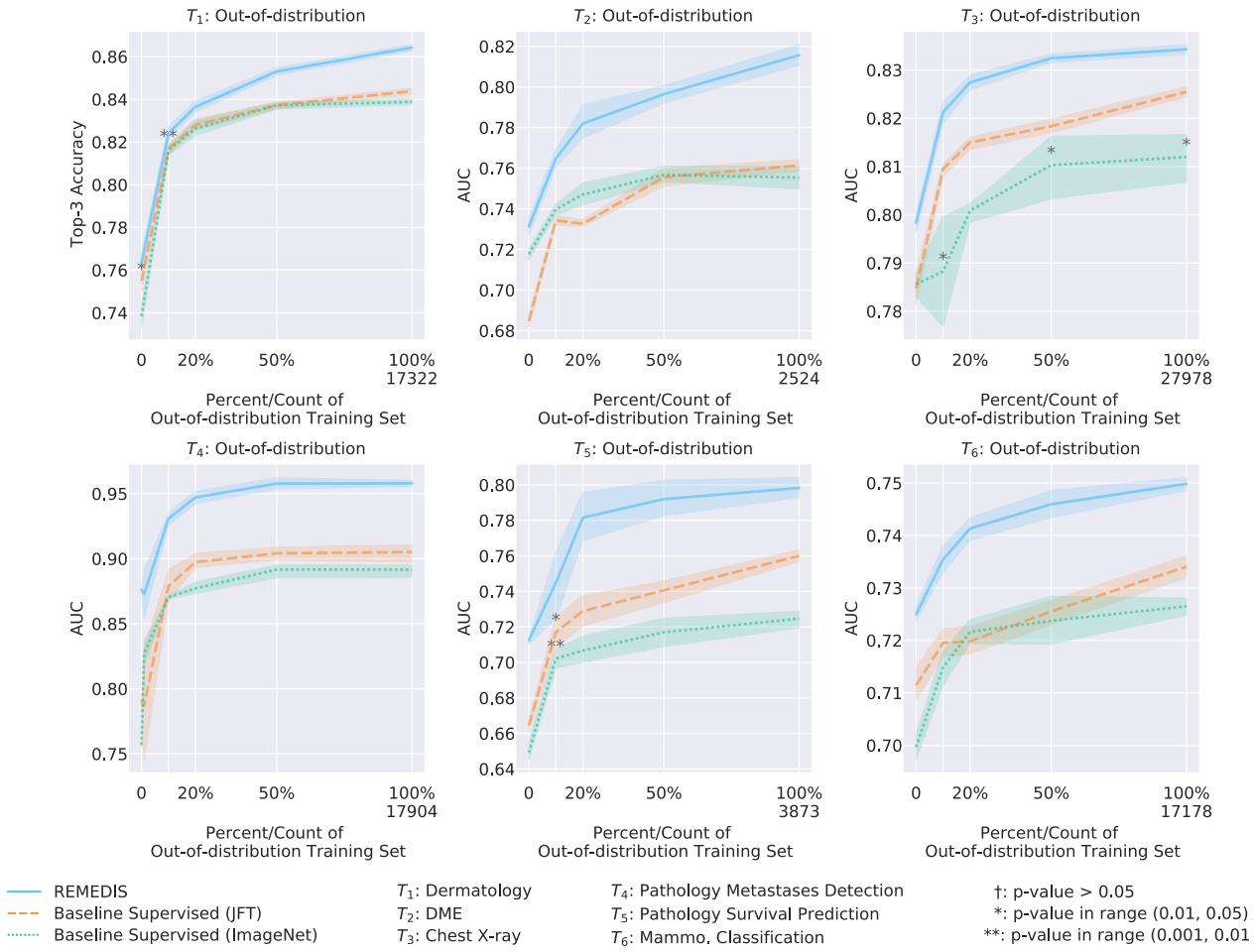
1. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S. & Houlsby, N. Big transfer (BiT): General visual representation learning. *arXiv preprint arXiv:1912.11370* **6** (2019).
2. Chen, T., Kornblith, S., Swersky, K., Norouzi, M. & Hinton, G. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029* (2020).
3. Xie, Q., Luong, M.-T., Hovy, E. & Le, Q. V. *Self-training with noisy student improves imagenet classification* in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), 10687–10698.
4. Mustafa, B., Loh, A., Freyberg, J., MacWilliams, P., Wilson, M., McKinney, S. M., Sieniek, M., Winkens, J., Liu, Y., Bui, P., *et al.* Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913* (2021).
5. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., *et al.* Big self-supervised models advance medical image classification. *ICCV 2021* (2021).
6. sano, Y. M., Rupprecht, C., Zisserman, A. & Vedaldi, A. PASS: An ImageNet replacement for self-supervised pretraining without humans. *NeurIPS Track on Datasets and Benchmarks* (2021).
7. Of Radiology, A. C., D'Orsi, C. J.,et al. *ACR BI-RADS atlas: breast imaging reporting and data system; mammography, ultrasound, magnetic resonance imaging, follow-up and outcome monitoring, data dictionary* (ACR, American College of Radiology, 2013)
8. Tan, M., Pang, R. & Le, Q. V. *EfficientDet: Scalable and Efficient Object Detection* 2020. arXiv: 1911.09070 [cs.CV].
9. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9** (2008).
10. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709* (2020).
11. Sowrirajan, H., Yang, J., Ng, A. Y. & Rajpurkar, P. *MoCo Pretraining Improves Representation and Transferability of Chest X-ray Models* in *Medical Imaging with Deep Learning* (2021), 728–744.
12. Chen, X., Fan, H., Girshick, R. & He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).



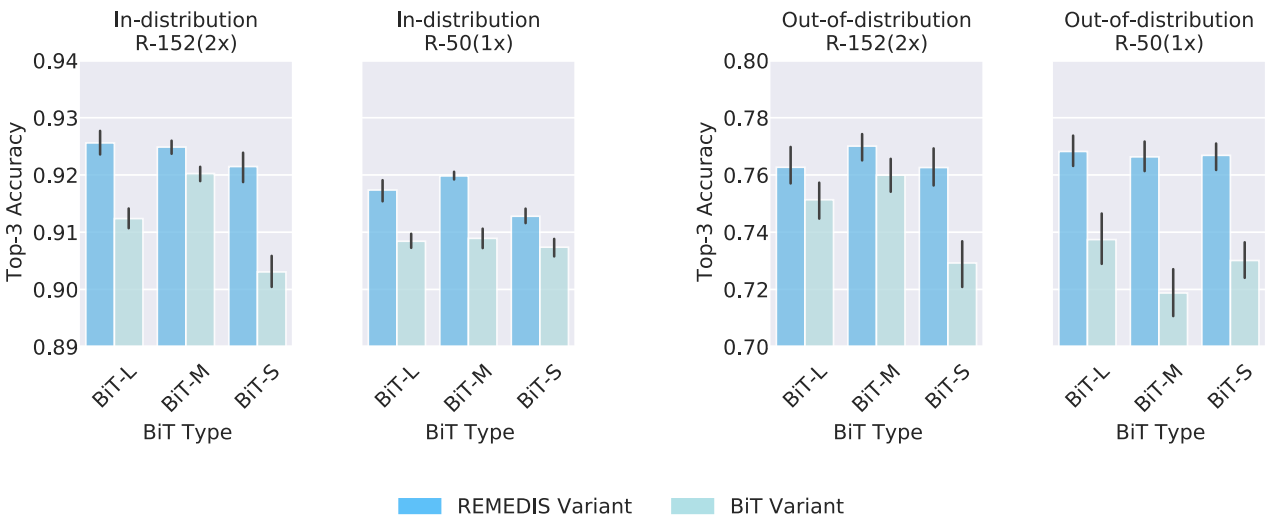
Supplementary Fig. 1 | REMEDIS vs. Standard Supervised Baseline. Overview of the results demonstrating overall performance and generalization of our proposed strategy as well as the standard supervised baseline pretrained on ImageNet-1K. We observed significantly improved out-of-distribution generalization and significant reduction in need for labelled medical data when using our proposed approach. 95% confidence intervals were calculated by running each label fraction and experiment up to ten times and intervals are shown using the shaded area and error bars. A two-sided *t*-test was also done for each experiment. If no * is shown, the *p*-value is less than 0.001, otherwise, the *p*-value is as indicated. The red lines indicate the amount of data that REMEDIS needs to match the highest standard supervised baseline performance when simulated in a new OOD clinical deployment setting and summarize the amount of annotated data and clinician hours potentially saved by using REMEDIS for each medical task.



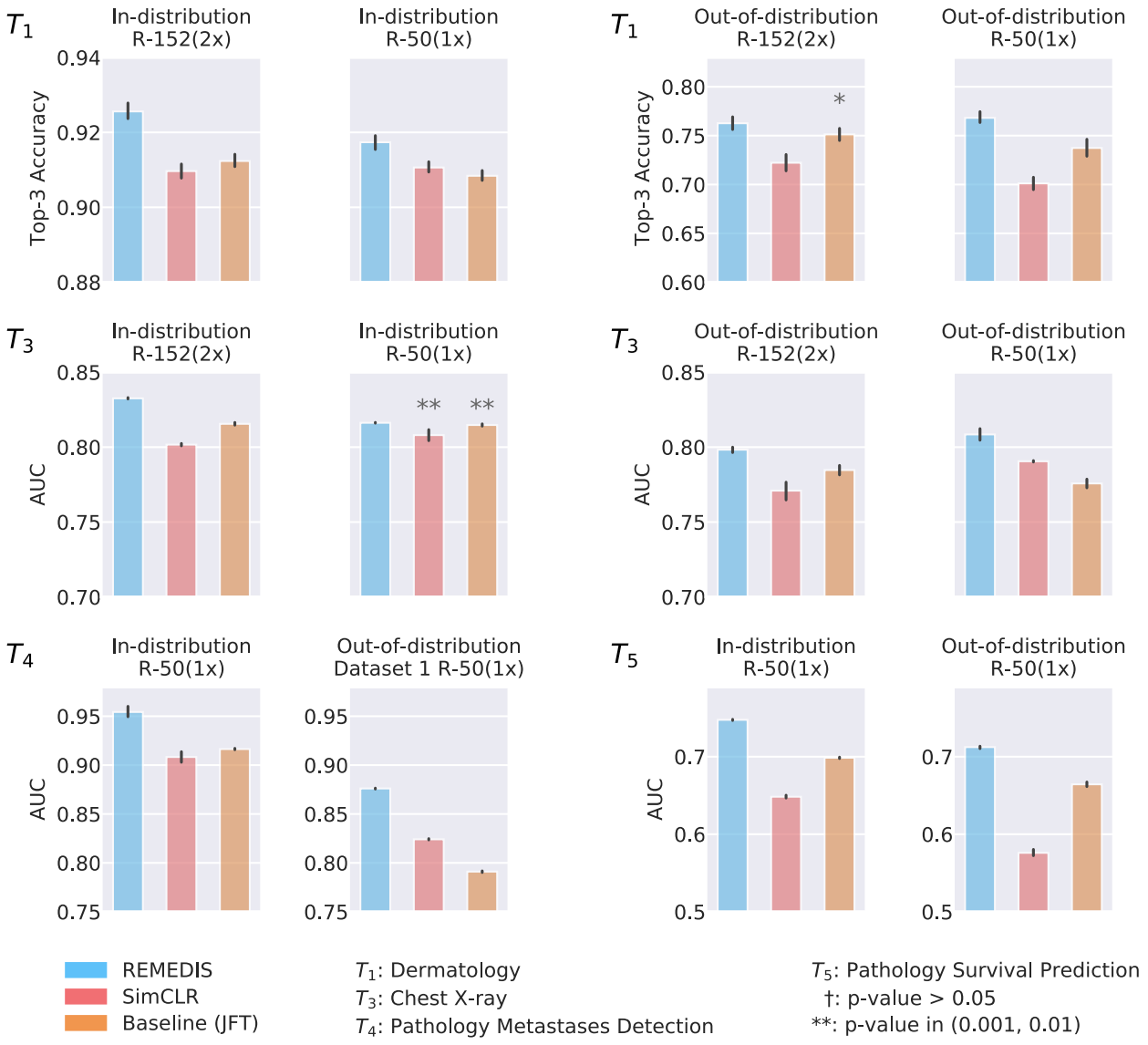
Supplementary Fig. 2 l Overall in-distribution performance. Overview of the results demonstrating overall in-distribution performance of REMEDIS as well as the standard and strong supervised baselines trained using both ImageNet-1K and JFT-300M. We observed significantly improved in-distribution performance using our proposed strategy vs. the standard and strong transfer learning strategies. Moreover, the strong supervised baseline (BiT-L) can provide significantly better in-distribution performance against the standard supervised baseline. 95% confidence intervals were calculated by running each label fraction and experiment up to ten times and intervals are shown using the error bars. A two-sided *t*t-test was also done for each label fraction as well as when computing the in-distribution results. If no * is shown, the *p*-value is less than 0.001, otherwise, the *p*-value is as indicated.



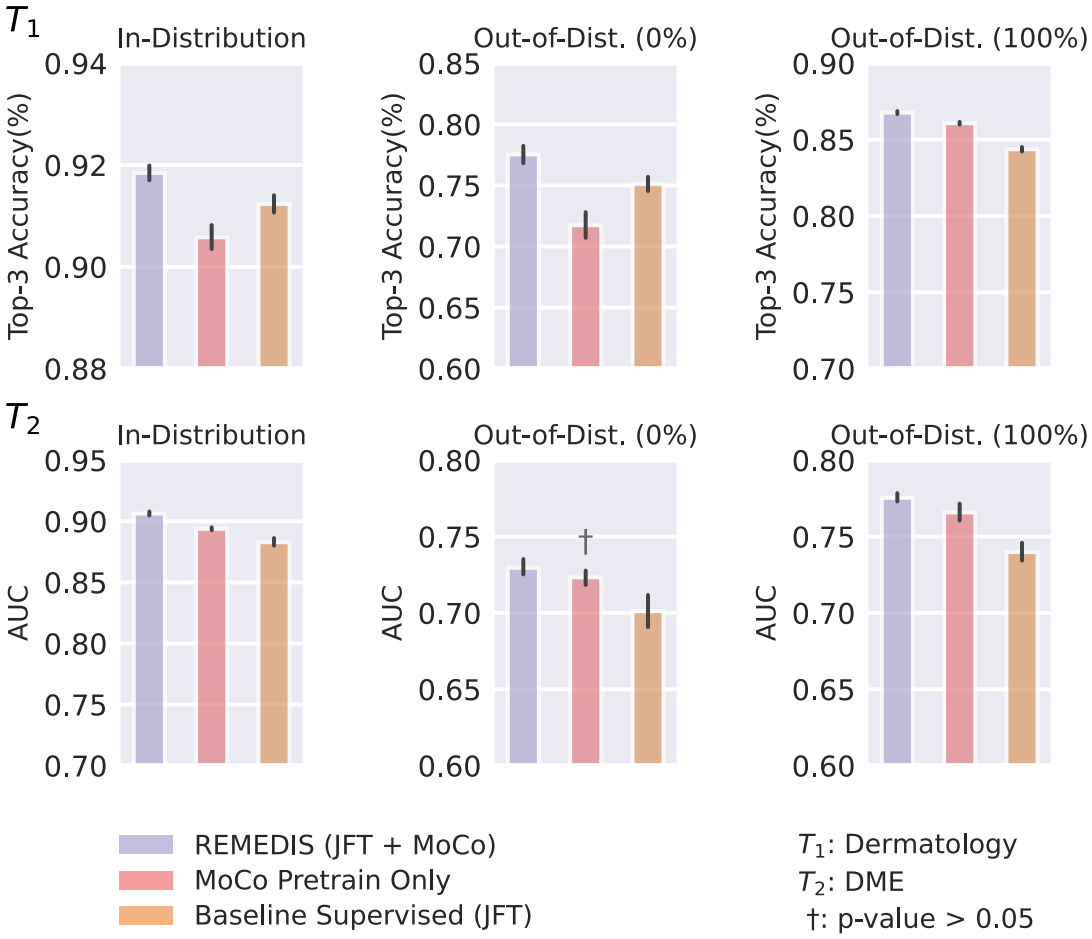
Supplementary Fig. 3 | REMEDIS overall data-efficiency performance. Overview of the results demonstrating data-efficient generalization of our proposed self-supervised learning strategy, REMEDIS as well as the standard and strong supervised baseline pretrained using both ImageNet-1K and JFT-300M. We observed significantly improved out-of-distribution generalization and significant reduction in need for annotated medical data when using our proposed approach. Moreover, comparing the data- efficient generalization of the strong and standard pretrained models, often string supervised baseline performs significantly better than the standard supervised baseline in out-of-distribution regime. 95% confidence intervals were calculated by running each label fraction and experiment up to ten times and intervals are shown using the shaded area. A two-sided t-test was also done for each label fraction. If no * is shown, the p-value is less than 0.001, otherwise, the p-value is as indicated.



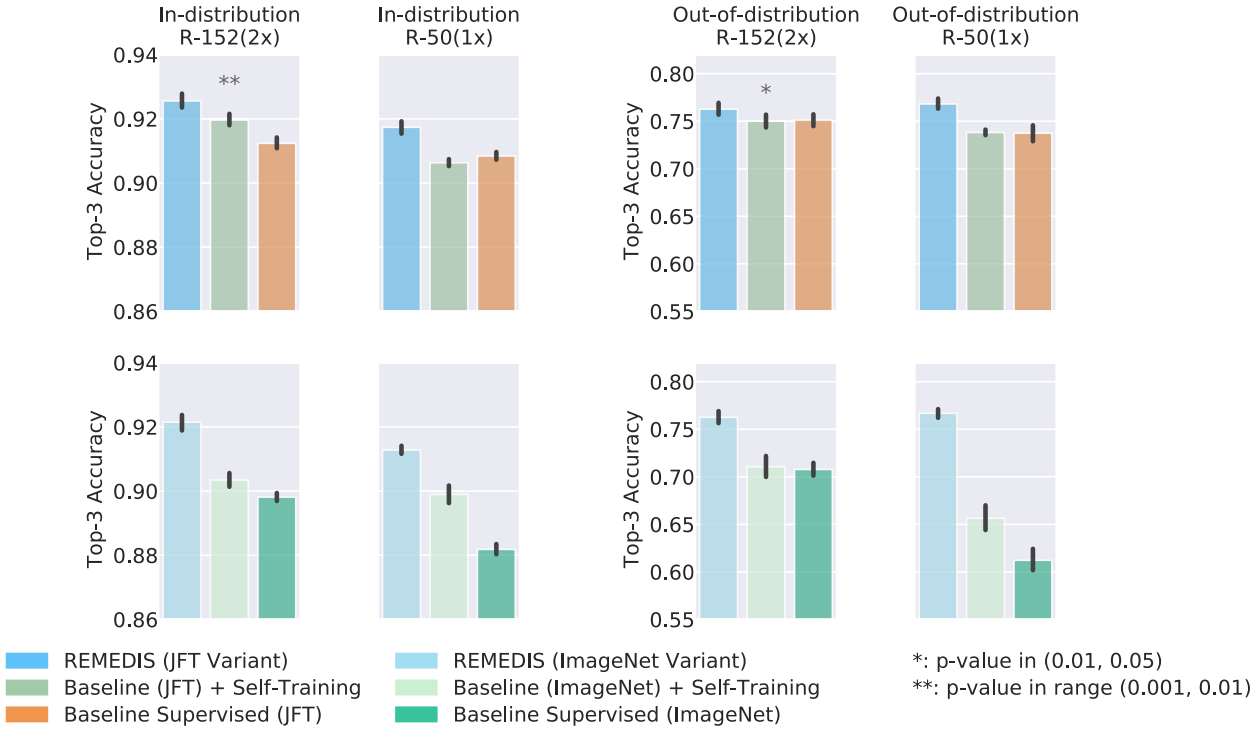
Supplementary Fig. 4 | Ablation of different BiT models used in both REMEDIS and the BiT Baseline, for the dermatology Task, T_1. The BiT model type shown is used in both REMEDIS and BiT for this comparison. BiT-L is the largest BiT model, trained on JFT, and is the default BiT model used in REMEDIS. BiT-M is trained on ImageNet 21k, while BiT-S is trained on ImageNet.



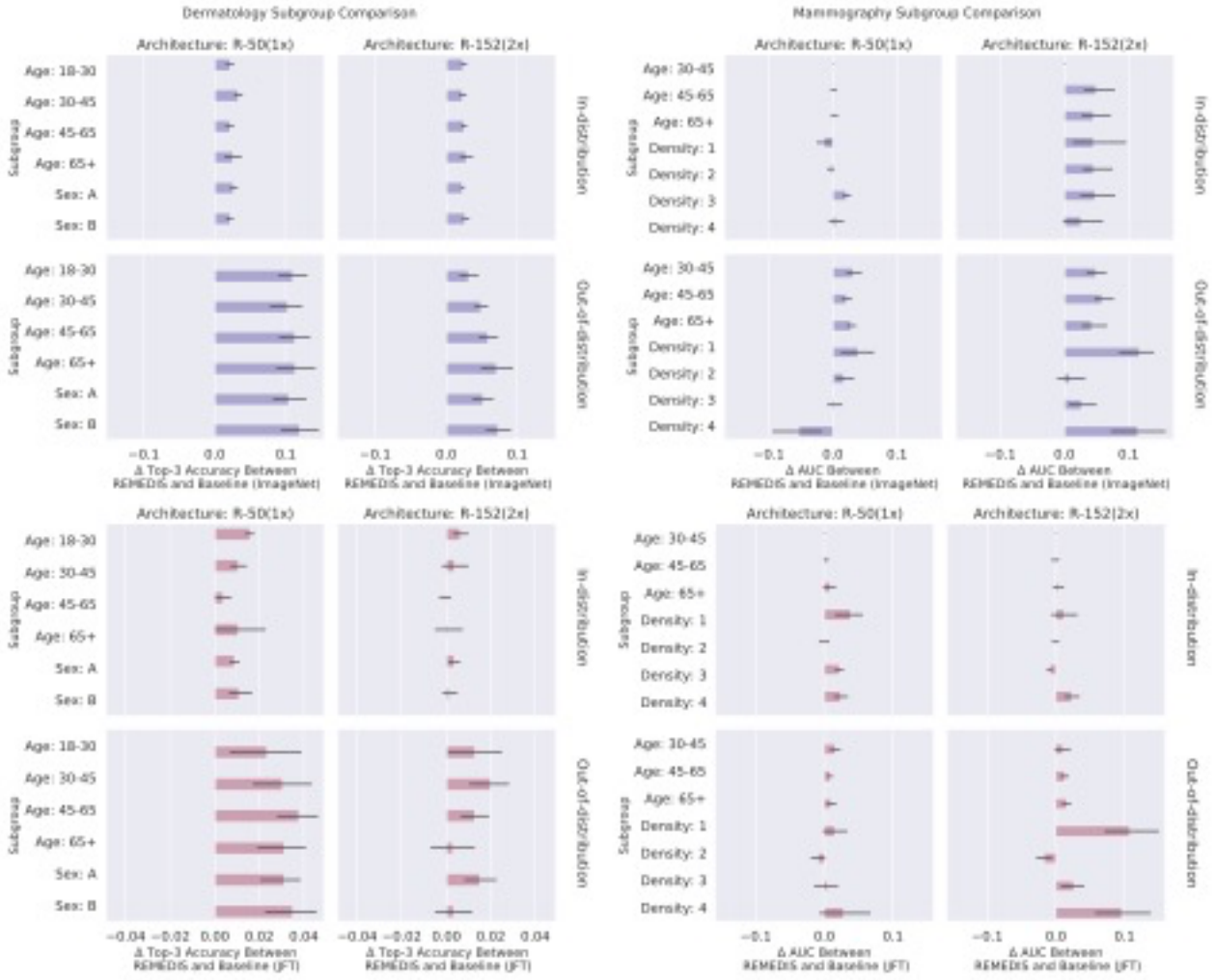
Supplementary Fig. 5 | Contribution of BiT-L and SimCLR. Both large-scale supervised pretraining and self-supervised pretraining separately provide benefits. A two-sided t-test is performed between each baseline model and REMEDIS, and a symbol above the bar being compared to REMEDIS shows the relevant p-value range. If no symbol is shown, the p-value is less than 0.001. REMEDIS outperforms both of its building components, BiT-L and SimCLR.



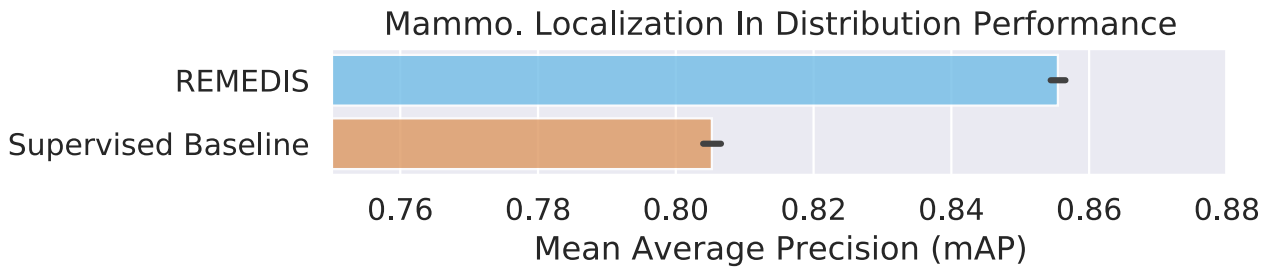
Supplementary Fig. 6 | Contribution of BiT-L and MoCo in REMEDIS MoCo variant. When using ResNet-152 (2×), both large-scale supervised pretraining and self-supervised pretraining separately provide benefits. A two-sided *t*-test is performed between each baseline model and REMEDIS MoCo variant, and a symbol above the bar being compared to REMEDIS shows the relevant *p*-value range. If no symbol is shown, the *p*-value is less than 0.001. REMEDIS MoCo variant outperforms both of its component building blocks, BiT-L and MoCo.



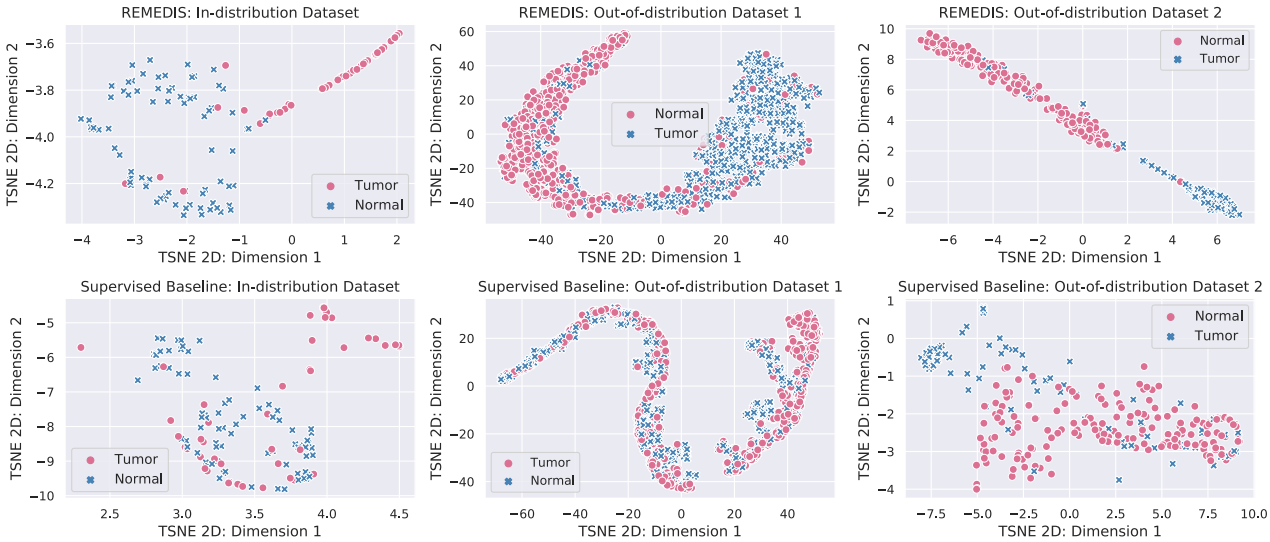
Supplementary Fig. 7 | Self-training vs. REMEDIS. Comparison of our proposed self-supervised method, REMEDIS as well as the strong and standard supervised baseline with self-training approach [12] for leveraging unlabelled medical data on the dermatology task. We observe that while self-training can produce high performance models on par with REMEDIS when using large architecture, it does not consistently provide benefits across all the settings considered. This is possibly due to the requirement for a good teacher model when using self-training. Furthermore, self-training requires the representation learning task and the downstream task to be well aligned while contrastive pretraining is agnostic to the downstream task leading to representations that can be generally applied.



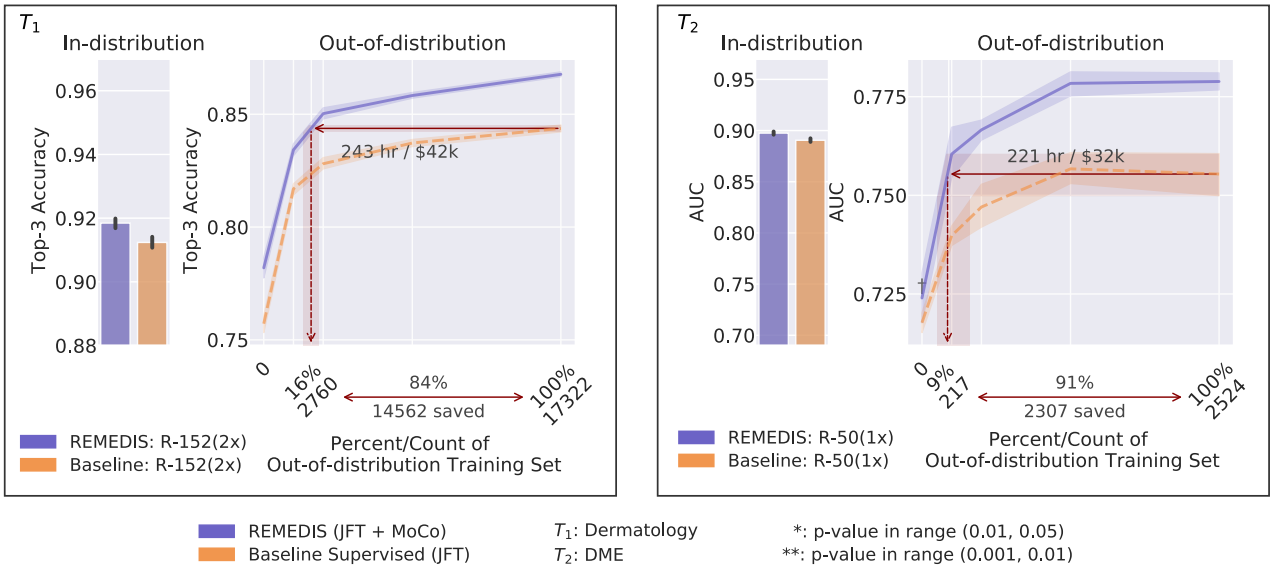
Supplementary Fig. 8 | Performance analysis across subgroups. Comparison of REMEDIS and the supervised baseline on subgroups of interest in the dermatology and mammography tasks. 95% confidence intervals were calculated by running each label fraction and experiment ten times and is shown with the error bars. In particular, we note that our method leads to improvement consistently across all subgroups of interest across both the tasks.



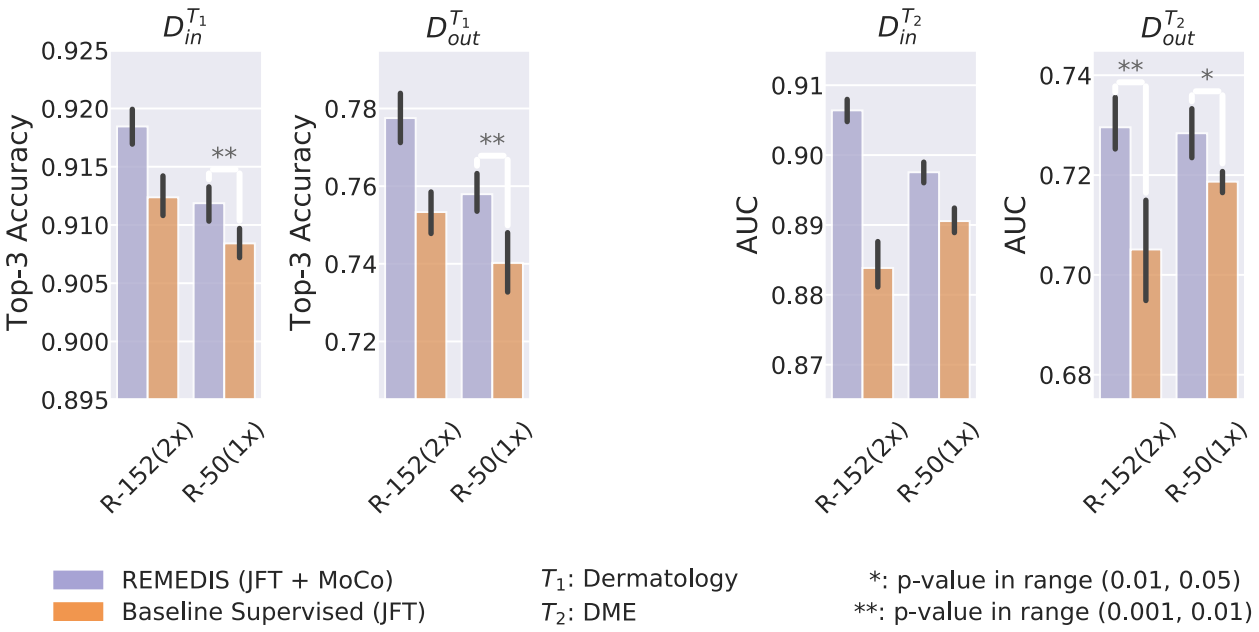
Supplementary Fig. 9 | Mammography localization performance, measured using the mean average precision of the localized cancer. 95% confidence intervals were calculated by running each label fraction and experiment ten times and is shown with the shaded area and error bars. A two-sided *t*-test was also done, which showed a *p*-value less than 0.001. Specifically, the t-Statistic was 57.14, the *p*-value is 3.54e-52, and the degree of freedom is 57.



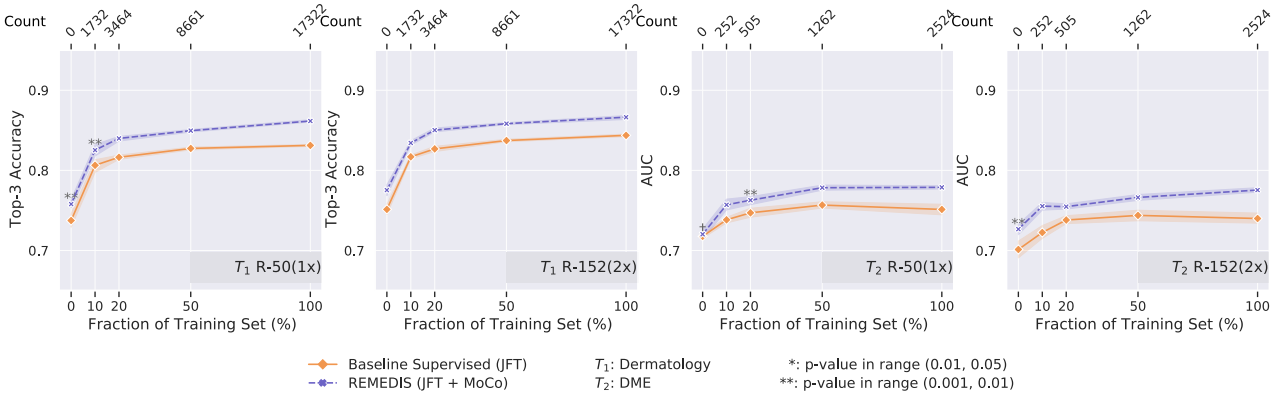
Supplementary Fig. 10 l t-SNE visualization of representations. The embedding representations obtained using REMEDIS and supervised baseline models are visualized for both the in-distribution and out-of-distribution datasets of the pathology metastases task. Clusters associated with various classes are better separated in the REMEDIS feature space as compared to the baseline.



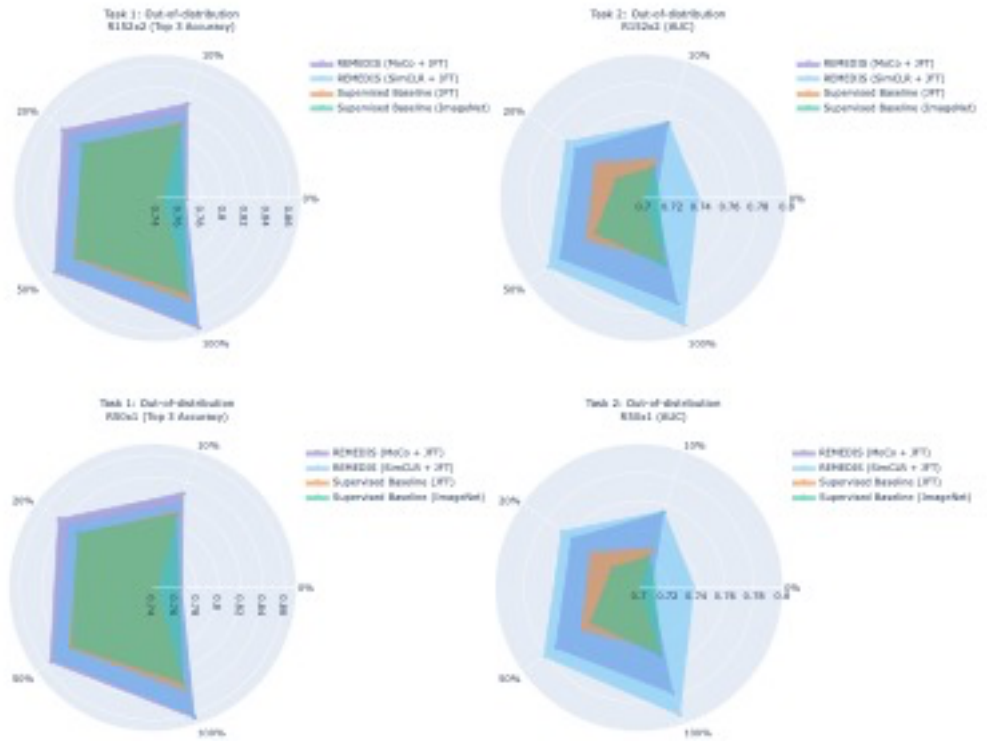
Supplementary Fig. 11 | Data-efficient generalization results of REMEDIS using MoCo for self-supervised learning. Overview of the results demonstrating overall performance and data-efficient generalization of our proposed self-supervised learning method, MoCo variant of REMEDIS as well as the strong supervised baseline pretrained on JFT-300M for the dermatology condition classification (T_1) and the diabetic macular edema classification (T_2). We observed that REMEDIS is compatible with momentum contrastive learning as an alternative self-supervised learning technique. If no * is shown, the *p*-value is less than 0.001, otherwise, the *p*-value is as indicated. The red lines indicate the amount of data that MoCo variant REMEDIS needs to match the best supervised AI baseline performance when simulated in a new clinical deployment setting and summarizes the savings in data annotation and clinician hours.



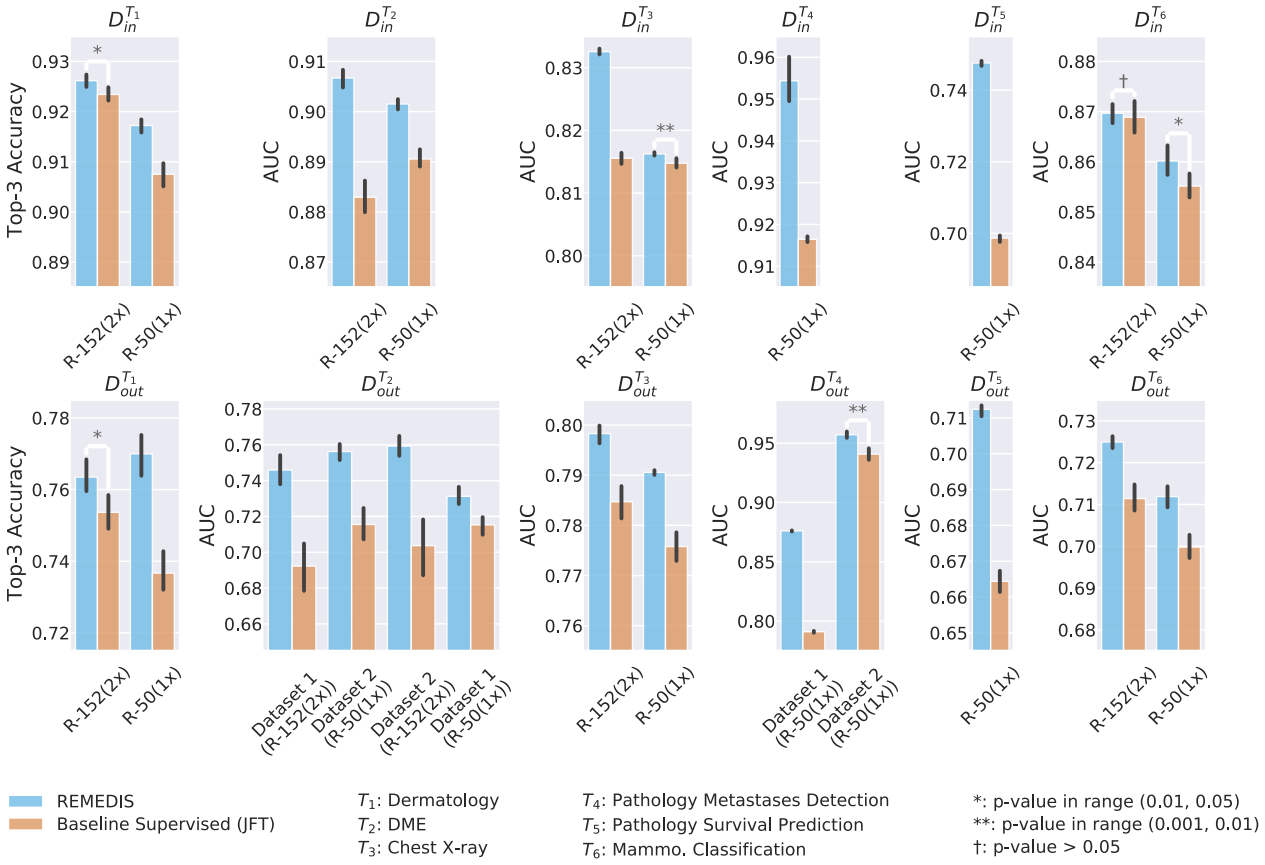
Supplementary Fig. 12 | Detailed in-distribution and zero-shot out-of-distribution performance for all architectures. We show the superior in-distribution performance of MoCo variant REMEDIS across T_1 and T_2 vs. strong supervised baseline trained on JFT-300M. The 95% confidence intervals were calculated by running each label fraction and experiment ten times and are shown with the shaded area and error bars. A two-sided t-test was also conducted for each pair of results. If no * is shown the p-value is less than 0.001, otherwise, the p-value is as indicated. Unlike previous visualizations, here we group the results based on the base network architecture not the architecture with the overall best performance.



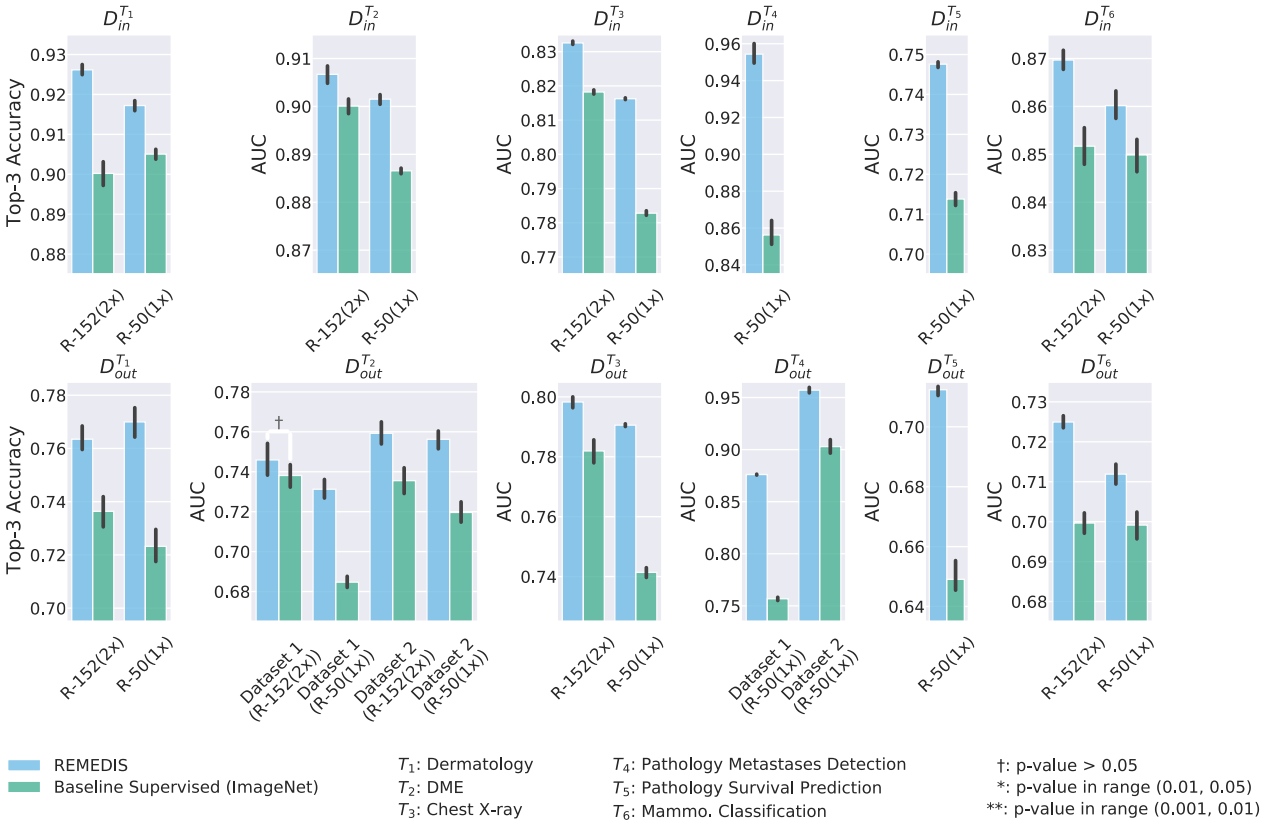
Supplementary Fig. 13 | Detailed generalization results for the MoCo variant REMEDIS vs. strong supervised baseline. Overview of the results demonstrating data-efficient generalization of MoCo variant REMEDIS vs. the strong supervised baseline pretrained on JFT-300M for all architectures. This includes the dermatology condition classification (T_1), diabetic macular edema classification (T_2) as well as two architectures ResNet-50 (1×) and ResNet-152 (2×). In particular, we observe significantly improved out-of-distribution generalization and a significant reduction in the need for labelled medical data when using our REMEDIS. 95% confidence intervals were calculated by running each label fraction and experiment ten times and intervals are shown using the shaded area and error bars. A two-sided t-test was also done for each label fraction as well as in-distribution results. If no * is shown, the p-value is less than 0.001, otherwise, the p-value is as indicated. Unlike previous visualizations, here we scale all of the graphs using a unified range and group the results based on the base network architecture.



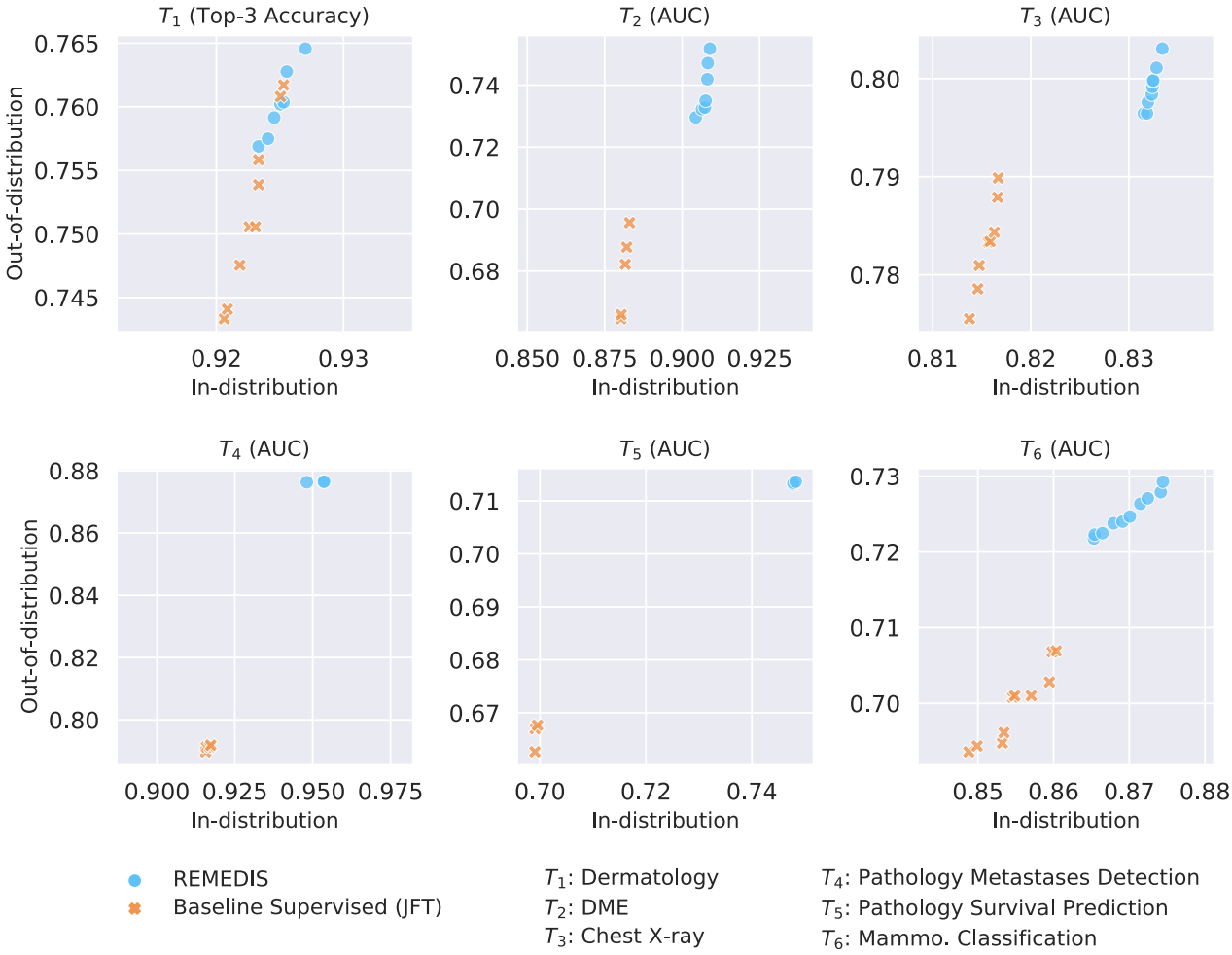
Supplementary Fig. 14 l Comparison of Performance of REMEDIS SimCLR and MoCo variant. Comparison of MoCo variant REMEDIS, SimCLR variant REMEDIS as well as the strong supervised baseline for the dermatology condition classification (T_1) and diabetic macular edema classification (T_2) for two architectures ResNet-50 (1×) and ResNet-152 (2×). 95% confidence intervals were calculated by running each label fraction and experiment ten times and is shown with the error bars. In particular, we note that our method leads to improvement consistently across all subgroups of interest across both the tasks.



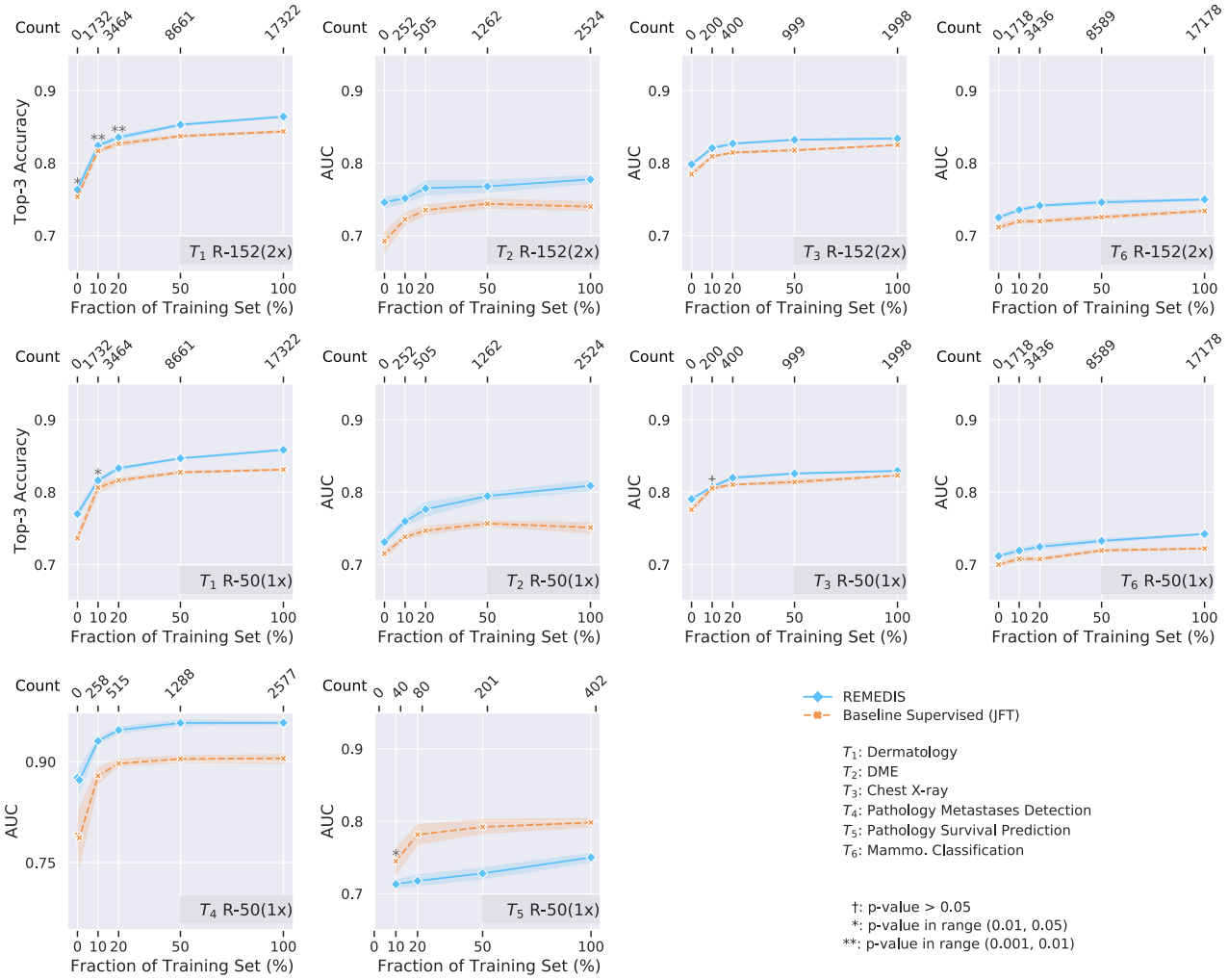
Supplementary Fig. 15 | Detailed in-distribution and zero-shot out-of-distribution performance for all architectures and all datasets considered in this study. We show the superior in-distribution performance of REMEDIS across all tasks and all datasets vs. strong supervised baseline trained on JFT-300M. The 95% confidence intervals were calculated by running each label fraction and experiment ten times and are shown with the shaded area and error bars. A two-sided *t*-test was also conducted for each pair of results. If no * is shown the *p*-value is less than 0.001, otherwise, the *p*-value is as indicated. Unlike previous visualizations, here we group the results based on the base network architecture not the architecture with the overall best performance.



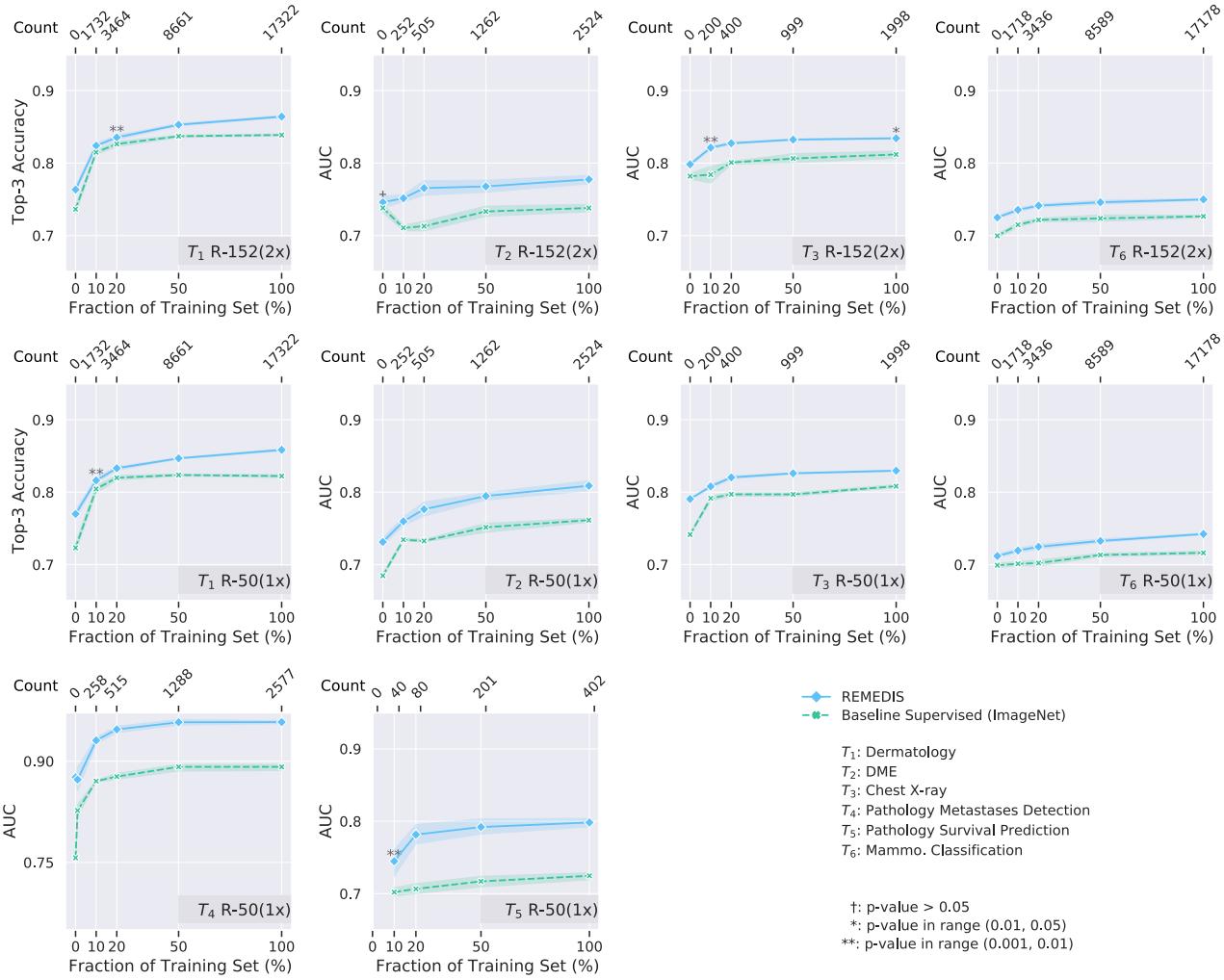
Supplementary Fig. 16 | Detailed in-distribution and zero-shot out-of-distribution performance for all architectures and all datasets considered in this study. We show the superior in-distribution performance of REMEDIS across all tasks and all datasets vs. the standard supervised baseline pretrained on ImageNet-1K. The 95% confidence intervals were calculated by running each label fraction and experiment ten times and are shown with the shaded area and error bars. A two-sided *t*-test was also conducted for each pair of results. If no * is shown the *p*-value is less than 0.001, otherwise, the *p*-value is as indicated. Unlike previous visualizations, here we group the results based on the base network architecture not the architecture with the overall best performance.



Supplementary Fig. 17 | In-distribution Performance vs. Zero-shot Out-of-distribution Performance. We show that REMEDIS produces a consistently superior model performance in comparison to the strong supervised baseline trained using JFT-300M both in- and out-of-distribution. 95% confidence intervals of our experiments were calculated by running each experiment ten times. Each point in this plot corresponds to one of these repeated runs and its coordinates are obtained by calculating the in-distribution and zero-shot out-of-distribution for the target task. These plots suggest REMEDIS improves out-of-distribution performance without decreasing in-distribution performance and REMEDIS models have higher in-distribution and out-of-distribution performance.



Supplementary Fig. 18 | Detailed generalization results for REMEDIS vs. strong supervised baseline. Overview of the results demonstrating data-efficient generalization of our method vs. the strong supervised baseline pretrained on JFT-300M for all tasks and architectures. This includes the dermatology condition classification (T_1), diabetic macular edema classification (T_2), chest X-ray condition classification (T_3), pathology metastases detection (T_4), pathology colorectal survival prediction (T_5), and mammography classification (T_6) as well as two architectures ResNet-50 (1×) and ResNet-152 (2×). In particular, we observe significantly improved out-of-distribution generalization and a significant reduction in the need for labelled medical data when using our proposed approach. 95% confidence intervals were calculated by running each label fraction and experiment ten times and intervals are shown using the shaded area and error bars. A two-sided t-test was also done for each label fraction as well as in-distribution results. If no * is shown, the p-value is less than 0.001, otherwise, the p-value is as indicated. Unlike previous visualizations, here we scale all of the graphs using a unified range and group the results based on the base network architecture.



Supplementary Fig. 19 | Detailed generalization results for REMEDIS vs. standard supervised baseline. Overview of the results demonstrating data-efficient generalization of our method vs. the standard supervised baseline pretrained on ImageNet-1K for all tasks and architectures. This includes the dermatology condition classification (T_1), diabetic macular edema classification (T_2), chest X-ray condition classification (T_3), pathology metastases detection (T_4), pathology colorectal survival prediction (T_5), and mammography classification (T_6) as well as two architectures ResNet-50 (1×) and ResNet-152 (2×). In particular, we observe significantly improved out-of-distribution generalization and a significant reduction in the need for labelled medical data when using our proposed approach. 95% confidence intervals were calculated by running each label fraction and experiment ten times and intervals are shown using the shaded area and error bars. A two-sided *t*-test was also done for each label fraction as well as in-distribution results. If no * is shown, the *p*-value is less than 0.001, otherwise, the *p*-value is as indicated. Unlike previous visualizations, here we scale all of the graphs using a unified range and group the results based on the base network architecture.

Supplementary Table 1 | Pretraining hyper-parameter details. We pre-trained the models using the following hyperparameter ranges for self-supervised learning with learning rate (lr) in {0.1 ,0.3}, temperature (τ) in {0.1, 0.2}, and batch size (B) in {1024, 2048, 4096}. We used random cropping (C), random color distortion (D), rotation (R), random Gaussian blur (G), histogram equalization (H), and elastic deformation (E) as the data augmentation strategies. We use LARS optimizer [115] and our experiments suggest that in all tasks, pretraining for 1000 epochs using a lr = 0.3 and τ = 0.1 tends to lead to optimal performance.

Tasks	Augmentations	Max Iteration (*M*)	Batch Size (*B*)	Architectures
T_1	C, D, R, G	202K	1024	ResNet-50 (1×) and ResNet-152 (2×)
T_2	C, D, R, G	2,229K	1024	ResNet-50 (1×) and ResNet-152 (2×)
T_3	C, D, R, G, H	210K	1024	ResNet-50 (1×) and ResNet-152 (2×)
T_4 , T_5	C, D, R, G	1,220K	4096	ResNet-50 (1×) Only
T_6	C, D, R, G, H, E	302K	1024	ResNet-50 (1×) and ResNet-152 (2×)

Supplementary Table 2 | Fine-tuning hyper-parameter search details. In the fine-tuning step, we investigated learning rate (lr),weight decay (w), the choice of optimizer and the decay step. In each case, we performed a grid search of logarithmically spaced samples for learning rate and weight decay and performed model selection based on the performance on the validation set in both ID and OOD settings

Tasks	Optimizer	Learning rate (*lr*)	Weight decay (*w*)	Max steps (*M*)	Decay step
T_1	Adam Linear	7 samples $\in [10^{-6.0}, 10^{-4.0}]$	$\{0, 10^{-6.0}, 10^{-5.0}, 10^{-4.0}\}$	150K	N/A
T_2	SGD Exponential	8 samples $\in [10^{-4.0}, 10^{-0.5}]$	$\{0, 10^{-5.0}, 10^{-4.0}, 10^{-3.0}\}$	1K	{*50*, *100*}
T_3	SGD Exponential	5 samples $\in [10^{-5.0}, 10^{-2.0}]$	$\{0, 10^{-6.0}, 10^{-5.0}\}$	250K	{*10K*, *25K*}
T_4 , T_5	Adam Linear	4 samples $\in [10^{-7.0}, 10^{-4.0}]$	N/A	25K	N/A
T_6	SGD Exponential	5 samples $\in [10^{-4.0}, 10^{-2.0}]$	$\{0, 10^{-6.0}, 10^{-5.0}, 10^{-4.0}\}$	100K	{*10K*, *25K*}

Supplementary Table 3 | Fine-tuning hyper-parameter details. In our hyper-parameter search, we investigated the choice of optimizer, learning rate, weight decay, decay step, and the network architecture. The table summarizes the selected hyper-parameters for fine-tuning REMEDIS and both strong and standard supervised baseline models pretrained on JFT-300M and ImageNet-1K dataset for both the ID and OOD settings.

	Parameters	T_1	T_2	T_3	T_4	T_5	T_6
Data	Input size	448×448	587×587	224×224	224×224	224×224	2048×2048
	Batch size	16	8	64	8	8	1
	Shuffle buffer	256	64	256	N/A	N/A	256
Optimization	Optimizer	Adam	SGD	Adam	Adam	Adam	SGD
	Schedule	Linear	Exponential	Exponential	Linear	Linear	Exponential
	Max training	150K	100K	250K	25K	25K	100K
Hyper-parameters REMEDIS (D_{in})	Architecture	R-152 (2×)	R-152 (2×)	R-152 (2×)	R-50 (1×)	R-50 (1×)	R-152 (2×)
	Learning rate	0.0003	0.3	0.001	0.0001	0.0001	0.0003
	Decay factor	N/A	0.99	0.9	N/A	N/A	0.1
	Decay step	N/A	50	10K	N/A	N/A	10K
	Weight decay	$10^{-5.0}$	0.0001	$10^{-5.0}$	N/A	N/A	0.001
Hyper-parameters REMEDIS (D_{out})	Architecture	R-152 (2×)	R-152 (2×)	R-152 (2×)	R-50 (1×)	R-50 (1×)	R-152 (2×)
	Learning rate	0.0001	0.3	0.001	0.0001	0.0001	0.0001
	Decay factor	N/A	0.99	0.9	N/A	N/A	0.1
	Decay step	N/A	100	10K	N/A	N/A	10K
	Weight decay	$10^{-5.0}$	0	$10^{-5.0}$	N/A	N/A	0.0001
Hyper-parameters Strong Baseline(D_{in})	Architecture	R-152 (2×)	R-152 (2×)	R-152 (2×)	R-50 (1×)	R-50 (1×)	R-152 (2×)
	Learning rate	0.0001	0.01	0.0001	$10^{-4.0}$	$10^{-5.0}$	0.01
	Decay factor	N/A	0.99	0.9	N/A	N/A	0.1
	Decay step	N/A	50	10K	N/A	N/A	10K
	Weight decay	0.0001	0.0001	$10^{-5.0}$	N/A	N/A	0
Hyper-parameters Strong Baseline(D_{out})	Architecture	R-152 (2×)	R-152 (2×)	R-152 (2×)	R-50 (1×)	R-50 (1×)	R-152 (2×)
	Learning rate	0.001	0.1	$10^{-5.0}$	$10^{-5.0}$	$10^{-7.0}$	0.0001
	Decay factor	N/A	0.99	0.9	N/A	N/A	0.1
	Decay step	N/A	100	N/A	N/A	N/A	10K
	Weight decay	$10^{-6.0}$	0	$10^{-5.0}$	N/A	N/A	0.0001
Hyper-parameters Standard Baseline(D_{in})	Architecture	R-152 (2×)	R-152 (2×)	R-152 (2×)	R-50 (1×)	R-50 (1×)	R-152 (2×)
	Learning rate	0.0001	0.003	0.001	$10^{-5.0}$	$10^{-6.0}$	0.01
	Decay factor	N/A	0.99	0.9	N/A	N/A	0.1
	Decay step	N/A	50	10K	N/A	N/A	10K
	Weight decay	0	$10^{-5.0}$	0	N/A	N/A	0
Hyper-parameters Standard Baseline(D_{out})	Architecture	R-152 (2×)	R-152 (2×)	R-152 (2×)	R-50 (1×)	R-50 (1×)	R-152 (2×)
	Learning rate	0.001	0.1	0.001	$10^{-5.0}$	$10^{-6.0}$	0.0001
	Decay factor	N/A	0.99	0.9	N/A	N/A	0.1
	Decay step	N/A	100	N/A	N/A	N/A	10K
	Weight decay	0.001	0	$10^{-5.0}$	N/A	N/A	0

Supplementary Table 4 | Clinically applicable performance. Summary of clinically applicable

performance range across the medical imaging tasks considered in this study in the OOD clinical setting wherever available.

Tasks	Task name	Clinician performance
T_1	Dermatology	0.650 (95% CI 0.545–0.755)
T_3	Chest-X-ray classification	0.869 (95% CI 0.843–0.894)
T_5	Survival prediction (pathology)	0.684 (95% CI 0.639-0.716)

Supplementary Table 5 | Dataset fingerprints. The above table illustrates the size and characteristics of the labelled, unlabelled and OOD dataset across the different medical imaging tasks we considered in this study.

Task	D_u	D_{in}			D_{out}			Secondary D_{out}
		Training	Validation	Test	Training	Validation	Test	
T_1	207,032	15,340	1,190	4,146	17,322	4,339	6,639	–
T_2	2,287,716	3,874	973	1,192	2,524	643	612	323
T_3	215,695	201,055	9,027	13,332	27,978	17,723	1,998	–
T_4	10,705	216	54	129	2,577	1,295	1,289	273
T_5	10,705	2,236	1,128	1,132	402	101	168	–
T_6	77,340	26,739	49,831	12,448	17,178	11,551	12,314	–

Supplementary Table 6 | Analysis of distribution shifts. The three most frequent sources of data distribution shifts in medical ML datasets are population shifts, technology shifts, and behavioral shifts, each arising from various influencing factors such as changes in acquisition devices, disease prevalence, etc. [27]. The check-mark (✓) and dash sign (–) indicate the existence or absence of the shift factor, and U indicates whether the evidence showing the factor is unknown or undefined in the metadata information associated with the datasets considered for the given medical imaging task.

Tasks		T_1	T_2	T_3	T_4	T_5	T_6
Technology shift	Acquisition device shift	✓	✓	U	✓	✓	✓
	IT practice, software, terminology shift	✓	✓	✓	✓	–	✓
Population Shift	Demographic shift	✓	✓	✓	–	–	✓
	Clinical setting shift	✓	–	–	–	–	✓
	Disease prevalence shift	✓	✓	–	✓	U	✓
	Seasonal shift	✓	U	U	–	–	–
Behaviour shift	Clinical behaviour and incentives shift	U	✓	–	–	✓	✓
	Patient behaviour change	U	U	–	U	✓	U
	Clinical practice change	U	U	–	U	✓	U
	Clinical nomenclature shift	U	U	–	U	✓	U

Supplementary Table 7 | Clinical cost analysis of data acquisition and annotation. The table below provides a summary of clinical costs associated with the collection of the OOD dataset for all the medical imaging tasks considered in this study. In all cases, we focus on the train splits of the dataset. The acquisition time for each task approximates the time that it took to collect each dataset starting from the first patient recruitment. The cost and hour savings are calculated based on the percentage of the data that REMEDIS requires to match the performance of the strong supervised baseline pretrained on JFT-300M dataset as depicted in Fig. 3. The overall annotation cost of a dataset is equal to (average annotation cost per image) × (number of training images) and the overall clinician hours for each dataset is equal to (average annotation time per image) × (number of training images).

Tasks	T_1	T_2	T_3	T_4	T_5	T_6
Number of training images	17,322	2,524	27,978	17,904	3,873	17,178
Average annotation time per image (second)	60	345	122	600	600	360
Average hourly wage of clinician ($)	$172	$147	$205	$138	$138	$205
Average cost of annotation per image ($)	$2.86	$14	$6.95	$23	$23	$20.5
Clinical hours cost of dataset (hour)	289	242	948	2,984	645	1,718
Annotation cost of dataset ($)	$49K	$35K	$194K	$411K	$89K	$352K
Acquisition Time (Years)	9	7	23	23	5	17
Acquisition Period	2007-2016	2003-2020	1992-2015	1984-2007	2008-2013	2001-2018
Percentage of data saved using REMEDIS	67%	93%	83%	94%	86%	91%
Amount of data saved using REMEDIS	11,578	2,342	23,278	16,872	3,325	15,689
Clinical annotation hours saved (hour)	193	224	789	2,812	554	1,569
Approximate annotation cost saved ($)	$33K	$33K	$162K	$385K	$76K	$322K

Supplementary Table 8 | Contribution of BiT-L and MoCo in REMEDIS MoCo variant. The table contains the exact metrics for Supplementary Figure 6. When using ResNet-152 (2×) and for T_1, T_2 , both large-scale supervised pretraining and self-supervised pretraining separately provide benefits. A two-sided t-test is performed between each baseline model and REMEDIS MoCo variant and the p-value more than 0.05 has been indicated with † and If no symbol is shown, the p-value is less than 0.001.

Task (Metric)	Method	In-distribution	Out-of-dist. (0%)	Out-of-dist. (100%)
Task 1 (Top-3 Acc.)	REMEDIS (JFT + MoCo)	0.918 (0.917,0.920)	0.775 (0.767,0.784)	0.868 (0.866,0.869)
	MoCo Pretrain Only	0.906 (0.903,0.909)	0.717 (0.704,0.731)	0.861 (0.860,0.862)
	Baseline Supervised (JFT)	0.912 (0.910,0.914)	0.755 (0.750,0.760)	0.844 (0.842,0.845)
Task 2 (AUC)	REMEDIS (JFT + MoCo)	0.906 (0.904,0.908)	0.730 (0.723,0.736)	0.775 (0.772,0.779)
	MoCo Pretrain Only	0.894 (0.892,0.895)	0.723 (0.716,0.730) †	0.766 (0.759,0.772)
	Baseline Supervised (JFT)	0.883 (0.879,0.887)	0.701 (0.686,0.717)	0.740 (0.732,0.748)

Supplementary Table 9 | Comparison of in-distribution improvement between MoCo and SimCLR variant REMEDIS vs. the strong supervised baseline. The absolute and relative improvement in the main task metrics between two variants of REMEDIS over the strong supervised baseline (JFT) for the in-distribution dataset is calculated. This data is represented in Figure 3 and Supplementary Fig. 11 for T_1 and T_2.

Task (Metric)	Method	Absolute Improvement	Relative Improvement (%)
Task 1 (Top-3 Accuracy)	REMEDIS (MoCo)	0.006 (0.003, 0.009)	0.7 (0.3, 1.0)
	REMEDIS (SimCLR)	0.003 (0.000, 0.005)	0.3 (0.0, 0.6)
Task 2 (AUC)	REMEDIS (MoCo)	0.007 (0.004, 0.01)	0.8 (0.4, 1.1)
	REMEDIS (SimCLR)	0.024 (0.018, 0.028)	2.7 (2.1, 3.2)

Supplementary Table 10 | Comparison of out-of-distribution improvement between MoCo and SimCLR variant REMEDIS vs. the strong supervised baseline. The absolute and relative improvement in the metrics between two variants of REMEDIS over the strong supervised baseline (JFT), using 0% and 100% of the out-of-distribution dataset is calculated. This data is represented in Figure 3 and Supplementary Fig.13 for T_1 and T_2.

Task (Metric)	Percentage	Method	Absolute Improvement	Relative Improvement (%)
Task 1 (Top-3 Accuracy)	0	REMEDIS (MoCo)	0.025 (0.017, 0.033)	3.3 (2.2, 4.4)
	0	REMEDIS (SimCLR)	0.025 (0.016, 0.035)	3.4 (2.2, 4.7)
	100	REMEDIS (MoCo)	0.024 (0.021, 0.027)	2.8 (2.5, 3.2)
	100	REMEDIS (SimCLR)	0.025 (0.022, 0.028)	3.0 (2.7, 3.4)
Task 2 (AUC)	0	REMEDIS (MoCo)	0.006 (-0.003, 0.015)	0.9 (-0.4, 2.1)
	0	REMEDIS (SimCLR)	0.047 (0.039, 0.054)	6.8 (5.7, 7.9)
	100	REMEDIS (MoCo)	0.023 (0.016, 0.031)	3.1 (2.1, 4.2)
	100	REMEDIS (SimCLR)	0.054 (0.046, 0.062)	7.1 (6.1, 8.2)

Supplementary Table 11 | Detailed in-distribution results. The table contains the numeric results displayed in Fig. 3 and Supplementary Fig. 1, specifically the average in-distribution performance values, with 95% confidence intervals in parentheses.

Task and Metric	Method	Metric
Task 1 (Top-3 Accuracy)	Baseline Supervised (ImageNet)	0.900 (0.897,0.903)
	Baseline Supervised (JFT)	0.923 (0.922,0.925)
	REMEDIS	0.926 (0.925,0.928)
Task 2 (AUC)	Baseline Supervised (ImageNet)	0.887 (0.886,0.887)
	Baseline Supervised (JFT)	0.883 (0.880,0.886)
	REMEDIS	0.902 (0.900,0.902)
Task 3 (AUC)	Baseline Supervised (ImageNet)	0.818 (0.818,0.819)
	Baseline Supervised (JFT)	0.816 (0.815,0.816)
	REMEDIS	0.833 (0.832,0.833)
Task 4 (AUC)	Baseline Supervised (ImageNet)	0.856 (0.851,0.864)
	Baseline Supervised (JFT)	0.916 (0.916,0.917)
	REMEDIS	0.954 (0.950,0.960)
Task 5 (AUC)	Baseline Supervised (ImageNet)	0.714 (0.712,0.715)
	Baseline Supervised (JFT)	0.699 (0.698,0.699)
	REMEDIS	0.748 (0.747,0.748)
Task 6 (AUC)	Baseline Supervised (ImageNet)	0.852 (0.848,0.856)
	Baseline Supervised (JFT)	0.855 (0.853,0.858)
	REMEDIS	0.870 (0.868,0.872)

Supplementary Table 12 | In-distribution improvement between REMEDIS and the strong supervised baseline. The absolute and relative improvement in the main task metrics between REMEDIS and the strong supervised baseline (JFT) for the in-distribution dataset. This data is represented in Fig. 3.

Task (Metric)	Absolute Improvement	Relative Improvement (%)
Task 1 (Top-3 Accuracy)	0.003 (0.000, 0.005)	0.3 (0.0, 0.6)
Task 2 (AUC)	0.024 (0.018, 0.028)	2.7 (2.1, 3.2)
Task 3 (AUC)	0.017 (0.016, 0.018)	2.1 (1.9, 2.3)
Task 4 (AUC)	0.038 (0.032, 0.044)	4.1 (3.5, 4.8)
Task 5 (AUC)	0.049 (0.047, 0.051)	7.0 (6.8, 7.2)
Task 6 (AUC)	0.001 (-0.005, 0.006)	0.1 (-0.5, 0.7)

Supplementary Table 13 | In-distribution improvement between REMEDIS and the standard supervised baseline. The absolute and relative improvement in the main task metrics between REMEDIS and the standard supervised baseline (ImageNet) for the in-distribution dataset. This data is represented in Supplementary Fig. 1.

Task (Metric)	Absolute Improvement	Relative Improvement (%)
Task 1 (Top-3 Accuracy)	0.026 (0.022, 0.03)	2.9 (2.4, 3.4)
Task 2 (AUC)	0.015 (0.013, 0.017)	1.7 (1.5, 1.9)
Task 3 (AUC)	0.014 (0.013, 0.016)	1.8 (1.6, 1.9)
Task 4 (AUC)	0.098 (0.085, 0.109)	11.5 (9.9, 12.8)
Task 5 (AUC)	0.034 (0.031, 0.036)	4.7 (4.4, 5.1)
Task 6 (AUC)	0.018 (0.012, 0.024)	2.1 (1.4, 2.8)

Supplementary Table 14 | Out-of-distribution data efficiency metrics vs. Clinician. This table contains the percentage and absolute numbers of the training set necessary for REMEDIS to match the strong supervised baseline (JFT) performance, as shown in Fig. 3, as well as the estimated clinician hours saved.

Task	Proportion required to meet baseline performance	Count of samples saved	Estimated clinician hours saved
Task 1	33.2% (25.7%, 39.3%)	11,578 (10510, 12862)	193 (175, 214)
Task 2	7.2% (4.0% , 10.3%)	2,342 (2263, 2423)	224 (217, 232)
Task 3	16.8% (12.0%, 22.8%)	23,278 (21588, 24620)	789 (732, 834)
Task 4	5.7% (2.3% 7.3%)	16,872 (16596, 17482)	2,812 (2766, 2914)
Task 5	14.1% (9.0% , 18.8%)	3,325 (3145, 3522)	554 (524, 587)
Task 6	8.7% (4.6%, 15.1%)	15,689 (14575, 16390)	1,569 (1457, 1639)

Supplementary Table 15 | Out-of-distribution data efficiency metrics vs. Clinician. This table contains the percentage and absolute numbers of the training set necessary for REMEDIS to match the standard supervised (ImageNet) performance, as shown in Supplementary Fig. 1, as well as the estimated clinician hours saved.

Task	Proportion required to meet baseline performance	Count of samples saved	Estimated clinician hours saved
Task 1	30.7% (25.14%, 36.9%)	11,990 (10920, 12967)	200 (182, 216)
Task 2	8.9% (6.6%, 12.9%)	2,297 (2196, 3356)	220 (210, 226)
Task 3	5.9% (2.8%, 8.9%)	26,315 (25467, 25467)	219 (212, 226)
Task 4	3.5% (0.6%, 4.5%)	17,269 (17095, 17788)	2,878 (2849, 2965)
Task 5	3.7% (1.2%, 10.7%)	3,727 (3460, 3826)	621 (577, 638)
Task 6	1.4% (0.0%, 4.9%)	16,924 (16323, 17178)	1,692 (1632, 1718)

Supplementary Table 16 | Out-of-distribution Best-vs-Best Metrics - Part 1 The table contains the exact metrics for Fig. 3 and Supplementary Fig. 1, specifically the out-of-distribution data efficiency metrics values for REMEDIS vs. the strong and standard supervised baselines for tasks T_1, T_2, T_3.

Task (Metric)	Method	Percentage	Metric
Task 1 (Top-3 Accuracy)	Baseline Supervised (ImageNet)	0	0.738 (0.734,0.743)
		10	0.816 (0.814,0.819)
		20	0.826 (0.823,0.830)
		50	0.837 (0.836,0.838)
		100	0.839 (0.838,0.840)
	Baseline Supervised (JFT)	0	0.755 (0.750,0.760)
		10	0.817 (0.814,0.819)
		20	0.828 (0.826,0.831)
		50	0.837 (0.836,0.839)
		100	0.844 (0.842,0.845)
	REMEDIS	0	0.763 (0.760,0.769)
		10	0.824 (0.822,0.827)
		20	0.836 (0.834,0.839)
		50	0.853 (0.851,0.855)
		100	0.864 (0.863,0.866)
Task 2 (AUC)	Baseline Supervised (ImageNet)	0	0.685 (0.682,0.688)
		10	0.734 (0.732,0.737)
		20	0.733 (0.731,0.735)
		50	0.756 (0.751,0.760)
		100	0.761 (0.759,0.764)
	Baseline Supervised (JFT)	0	0.718 (0.715,0.720)
		10	0.740 (0.737,0.742)
		20	0.747 (0.742,0.753)
		50	0.757 (0.753,0.761)
		100	0.755 (0.750,0.761)
	REMEDIS	0	0.731 (0.727,0.736)
		10	0.765 (0.760,0.770)
		20	0.782 (0.774,0.791)
		50	0.796 (0.792,0.801)
		100	0.816 (0.811,0.821)
Task 3 (AUC)	Baseline Supervised (ImageNet)	0	0.786 (0.783,0.788)
		10	0.788 (0.777,0.800)
		20	0.801 (0.798,0.802)
		50	0.810 (0.803,0.816)
		100	0.812 (0.807,0.817)
	Baseline Supervised (JFT)	0	0.785 (0.781,0.788)
		10	0.809 (0.808,0.810)
		20	0.815 (0.813,0.816)
		50	0.818 (0.817,0.819)
		100	0.825 (0.824,0.826)
	REMEDIS	0	0.798 (0.796,0.800)
		10	0.822 (0.819,0.824)
		20	0.828 (0.826,0.830)
		50	0.833 (0.832,0.833)
		100	0.835 (0.834,0.836)

Supplementary Table 17 | Out-of-distribution Best-vs-Best Metrics - Part 2 The table contains the exact metrics for Fig. 3 and Supplementary Fig. 1, specifically the out-of-distribution data efficiency metrics values for REMEDIS vs. the strong and standard supervised baselines for tasks T_4, T_5, T_6.

Task (Metric)	Method	Percentage	Metric
Task 4 (AUC)	Baseline Supervised (ImageNet)	0	0.757 (0.755,0.758)
		10	0.870 (0.869,0.872)
		20	0.877 (0.873,0.882)
		50	0.892 (0.885,0.895)
		100	0.892 (0.886,0.895)
	Baseline Supervised (JFT)	0	0.791 (0.790,0.792)
		10	0.879 (0.868,0.891)
		20	0.897 (0.893,0.904)
		50	0.904 (0.899,0.909)
		100	0.905 (0.897,0.911)
	REMEDIS	0	0.876 (0.876,0.876)
		10	0.931 (0.926,0.935)
		20	0.947 (0.942,0.952)
		50	0.958 (0.954,0.962)
		100	0.958 (0.956,0.960)
Task 5 (AUC)	Baseline Supervised (ImageNet)	0	0.649 (0.645,0.655)
		10	0.702 (0.697,0.708)
		20	0.707 (0.700,0.715)
		50	0.717 (0.710,0.725)
		100	0.725 (0.719,0.729)
	Baseline Supervised (JFT)	0	0.664 (0.661,0.667)
		10	0.717 (0.709,0.726)
		20	0.729 (0.720,0.737)
		50	0.741 (0.733,0.746)
		100	0.760 (0.757,0.763)
	REMEDIS	0	0.712 (0.710,0.714)
		10	0.745 (0.726,0.761)
		20	0.782 (0.768,0.795)
		50	0.792 (0.783,0.803)
		100	0.798 (0.792,0.804)
Task 6 (AUC)	Baseline Supervised (ImageNet)	0	0.700 (0.697,0.702)
		10	0.715 (0.711,0.718)
		20	0.722 (0.719,0.724)
		50	0.724 (0.720,0.728)
		100	0.727 (0.725,0.728)
	Baseline Supervised (JFT)	0	0.711 (0.709,0.715)
		10	0.720 (0.717,0.722)
		20	0.720 (0.717,0.723)
		50	0.726 (0.723,0.728)
		100	0.734 (0.732,0.736)
	REMEDIS	0	0.725 (0.724,0.726)
		10	0.735 (0.733,0.738)
		20	0.741 (0.739,0.743)
		50	0.746 (0.743,0.749)
		100	0.750 (0.749,0.751)

Supplementary Table 18 | Out-of-distribution Best-vs-Best Metrics for the MoCo variant of REMEDIS. The table contains the exact metrics values for Supplementary Fig. 11, specifically the out-of-distribution data efficiency metrics values for the MoCo variant of REMEDIS vs. the strong supervised baselines for tasks T_1, T_2.

Task (Metric)	Method	Percentage	Metric
Task 1 (Top-3 Accuracy)	Baseline Supervised (JFT)	0	0.755 (0.750,0.760)
		10	0.817 (0.814,0.819)
		20	0.828 (0.826,0.831)
		50	0.837 (0.836,0.839)
		100	0.844 (0.842,0.845)
	REMEDIS (MoCo + JFT)	0	0.782 (0.777,0.787)
		10	0.834 (0.832,0.837)
		20	0.850 (0.848,0.853)
		50	0.858 (0.857,0.860)
		100	0.868 (0.867,0.869)
Task 2 (AUC)	Baseline Supervised (JFT)	0	0.718 (0.715,0.720)
		10	0.740 (0.737,0.742)
		20	0.747 (0.742,0.753)
		50	0.757 (0.753,0.761)
		100	0.755 (0.750,0.761)
	REMEDIS (MoCo + JFT)	0	0.724 (0.718,0.730)
		10	0.760 (0.755,0.768)
		20	0.767 (0.764,0.769)
		50	0.778 (0.775,0.781)
		100	0.779 (0.777,0.781)

Supplementary Table 19 | Out-of-distribution relative improvement between REMEDIS and the strong supervised baseline. The absolute and relative improvement in the metrics between REMEDIS and the strong supervised baseline (JFT), using 0% and 100% of the Out-of-distribution dataset. This data is represented in Figure 3.

Task (Metric)	Percentage	Absolute Improvement	Relative Improvement (%)
Task 1 (Top-3 Accuracy)	0	0.009 (0.000, 0.018)	1.2 (0.0, 2.5)
	100	0.020 (0.017, 0.023)	2.4 (2.1, 2.8)
Task 2 (AUC)	0	0.014 (0.007, 0.021)	1.9 (0.9, 3.0)
	100	0.060 (0.050, 0.071)	8.0 (6.6, 9.5)
Task 3 (AUC)	0	0.014 (0.008, 0.019)	1.7 (1.1, 2.4)
	100	0.009 (0.007, 0.011)	1.1 (0.9, 1.3)
Task 4 (AUC)	0	0.085 (0.084, 0.086)	10.7 (10.6, 10.9)
	100	0.053 (0.045, 0.063)	5.8 (5.0, 7.0)
Task 5 (AUC)	0	0.048 (0.043, 0.052)	7.2 (6.5, 7.9)
	100	0.038 (0.029, 0.047)	5.0 (3.8, 6.2)
Task 6 (AUC)	0	0.014 (0.009, 0.018)	1.9 (1.2, 2.5)
	100	0.016 (0.012, 0.019)	2.2 (1.7, 2.6)

Supplementary Table 20 | Out-of-distribution relative improvement between REMEDIS and the standard supervised baseline. The absolute and relative improvement in the metrics between REMEDIS and the standard supervised baseline (ImageNet), using 0% and 100% of the Out-of-distribution dataset. This data is represented in Supplementary Fig. 1.

Task (Metric)	Percentage	Absolute Improvement	Relative Improvement (%)
Task 1 (Top-3 Accuracy)	0	0.025 (0.016, 0.035)	3.4 (2.2, 4.7)
	100	0.025 (0.022, 0.028)	3.0 (2.7, 3.4)
Task 2 (AUC)	0	0.047 (0.039, 0.054)	6.8 (5.7, 7.9)
	100	0.054 (0.046, 0.062)	7.1 (6.1, 8.2)
Task 3 (AUC)	0	0.013 (0.009, 0.017)	1.6 (1.1, 2.2)
	100	0.022 (0.017, 0.028)	2.7 (2.0, 3.5)
Task 4 (AUC)	0	0.119 (0.117, 0.121)	15.8 (15.5, 16.1)
	100	0.066 (0.061, 0.074)	7.4 (6.8, 8.4)
Task 5 (AUC)	0	0.063 (0.055, 0.068)	9.8 (8.4, 10.5)
	100	0.074 (0.064, 0.084)	10.2 (8.8, 11.7)
Task 6 (AUC)	0	0.025 (0.021, 0.029)	3.6 (3.0, 4.2)
	100	0.023 (0.02, 0.026)	3.2 (2.8, 3.6)

Supplementary Table 21 | Ablation Study Statistics The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 5, specifically the metrics produced for REMEDIS vs. each of the different techniques. This t-test was done without the assumption that the variances are equal.

Task	Distribution	Architecture	Comparison Method	T-Statistic	*p*-value	Degrees of Freedom
Task 1	In-distribution	R-152(2x)	BiT-L	1.83	8.848414e-02	14.06
Task 1	In-distribution	R-152(2x)	SimCLR	10.29	6.815352e-09	17.69
Task 1	In-distribution	R-50(1x)	BiT-L	7.22	2.571443e-06	15.40
Task 1	In-distribution	R-50(1x)	SimCLR	5.36	7.154525e-05	15.46
Task 1	Out-of-distribution	R-152(2x)	BiT-L	1.54	1.411247e-01	17.58
Task 1	Out-of-distribution	R-152(2x)	SimCLR	6.95	2.933571e-06	16.34
Task 1	Out-of-distribution	R-50(1x)	BiT-L	5.59	5.035536e-05	15.06
Task 1	Out-of-distribution	R-50(1x)	SimCLR	14.68	1.041017e-10	16.02
Task 3	In-distribution	R-152(2x)	BiT-L	31.20	2.842220e-14	13.91
Task 3	In-distribution	R-152(2x)	SimCLR	62.10	1.797997e-19	14.96
Task 3	In-distribution	R-50(1x)	BiT-L	3.52	4.608371e-03	11.28
Task 3	In-distribution	R-50(1x)	SimCLR	3.95	3.307544e-03	9.09
Task 3	Out-of-distribution	R-152(2x)	BiT-L	6.75	7.829244e-06	14.47
Task 3	Out-of-distribution	R-152(2x)	SimCLR	8.04	7.225264e-06	10.74
Task 3	Out-of-distribution	R-50(1x)	BiT-L	9.19	5.085610e-06	9.46
Task 3	Out-of-distribution	R-50(1x)	SimCLR	-8.48	1.152971e-05	9.27
Task 4	In-distribution	R-50(1x)	BiT-L	3.61	2.847966e-02	3.49
Task 4	In-distribution	R-50(1x)	SimCLR	0.66	5.547361e-01	3.15
Task 4	Out-of-distribution	R-50(1x))	BiT-L	27.16	9.455735e-05	3.06
Task 4	Out-of-distribution	R-50(1x))	SimCLR	13.51	8.170055e-04	3.04
Task 5	In-distribution	R-50(1x)	BiT-L	10.81	1.523292e-03	3.07
Task 5	In-distribution	R-50(1x)	SimCLR	19.42	2.716788e-04	3.04
Task 5	Out-of-distribution	R-50(1x)	BiT-L	3.52	3.805765e-02	3.05
Task 5	Out-of-distribution	R-50(1x)	SimCLR	12.04	1.120045e-03	3.06

Supplementary Table 22 | Task Statistics REMEDIS vs. the strong supervised baseline (JFT). The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 15, specifically the metrics produced for REMEDIS vs. the strong supervised baseline (JFT). This t-test was done without the assumption that the variances are equal.

Task	Distribution	Architecture	T-Statistic	*p*-value	Degrees of Freedom
Task 1	In-distribution	R-152(2x)	2.74	1.333964e-02	17.97
Task 1	In-distribution	R-50(1x)	6.53	1.478805e-05	13.71
Task 1	Out-of-distribution	R-152(2x)	2.76	1.331786e-02	16.98
Task 1	Out-of-distribution	R-50(1x)	7.97	2.626604e-07	17.98
Task 2	In-distribution	R-152(2x)	11.91	8.600512e-08	11.42
Task 2	In-distribution	R-50(1x)	9.91	8.610633e-08	14.31
Task 2	Out-of-distribution	Dataset 1 (R-152(2x))	6.27	4.792115e-05	11.61
Task 2	Out-of-distribution	Dataset 1 (R-50(1x))	4.24	4.932287e-04	17.93
Task 2	Out-of-distribution	Dataset 2 (R-152(2x))	5.87	2.889648e-04	8.56
Task 2	Out-of-distribution	Dataset 2 (R-50(1x))	7.49	3.946623e-06	13.32
Task 3	In-distribution	R-152(2x)	31.20	2.842220e-14	13.91
Task 3	In-distribution	R-50(1x)	3.52	4.608371e-03	11.28
Task 3	Out-of-distribution	R-152(2x)	6.75	7.829244e-06	14.47
Task 3	Out-of-distribution	R-50(1x)	9.19	5.085610e-06	9.46
Task 4	In-distribution	R-50(1x)	12.74	8.643737e-04	3.11
Task 4	Out-of-distribution	Dataset 1 (R-50(1x))	154.09	1.267567e-10	5.14
Task 4	Out-of-distribution	Dataset 2 (R-50(1x))	5.11	4.485801e-03	4.69
Task 5	In-distribution	R-50(1x)	71.75	1.512674e-09	5.62
Task 5	Out-of-distribution	R-50(1x)	23.65	4.808754e-06	4.67
Task 6	In-distribution	R-152(2x)	0.40	6.933459e-01	15.07
Task 6	In-distribution	R-50(1x)	2.50	2.251765e-02	17.35
Task 6	Out-of-distribution	R-152(2x)	7.24	6.630100e-06	12.99
Task 6	Out-of-distribution	R-50(1x)	5.87	1.624941e-05	17.61

Supplementary Table 23 | Task Statistics MoCo variant of REMEDIS vs. the strong supervised baseline (JFT). The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 12, specifically the metrics produced for REMEDIS vs. the strong supervised baseline (JFT). This t-test was done without the assumption that the variances are equal.

Task	Distribution	Architecture	T-Statistic	*p*-value	Degrees of Freedom
Task 1	In-distribution	R-152(2x)	4.98	1.001209e-04	17.76
Task 1	In-distribution	R-50(1x)	3.34	3.755625e-03	17.59
Task 1	Out-of-distribution	R-152(2x)	5.25	8.540898e-05	15.59
Task 1	Out-of-distribution	R-50(1x)	3.58	3.075927e-03	13.75
Task 2	In-distribution	R-152(2x)	11.86	7.279249e-07	9.16
Task 2	In-distribution	R-50(1x)	5.50	3.976754e-05	16.87
Task 2	Out-of-distribution	R-152(2x)	3.83	9.103610e-03	5.84
Task 2	Out-of-distribution	R-50(1x)	3.18	2.204511e-02	5.42

Supplementary Table 24 | Task Statistics REMEDIS vs. the standard supervised baseline (ImageNet). The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 16, specifically the metrics produced for REMEDIS vs. the standard supervised baseline (ImageNet). This t-test was done without the assumption that the variances are equal.

Task	Distribution	Architecture	T-Statistic	*p*-value	Degrees of Freedom
Task 1	In-distribution	R-152(2x)	14.60	4.509936e-09	12.16
Task 1	In-distribution	R-50(1x)	12.70	2.120881e-10	17.93
Task 1	Out-of-distribution	R-152(2x)	6.89	2.967478e-06	16.64
Task 1	Out-of-distribution	R-50(1x)	10.51	4.458786e-09	17.86
Task 2	In-distribution	R-152(2x)	5.10	8.239890e-05	17.48
Task 2	In-distribution	R-50(1x)	23.57	6.420586e-13	14.42
Task 2	Out-of-distribution	Dataset 1 (R-152(2x))	1.47	1.609929e-01	16.13
Task 2	Out-of-distribution	Dataset 1 (R-50(1x))	15.63	1.932572e-10	14.42
Task 2	Out-of-distribution	Dataset 2 (R-152(2x))	5.18	6.678261e-05	17.63
Task 2	Out-of-distribution	Dataset 2 (R-50(1x))	9.97	1.046290e-08	17.79
Task 3	In-distribution	R-152(2x)	33.52	7.022141e-17	16.87
Task 3	In-distribution	R-50(1x)	87.58	4.931182e-18	11.85
Task 3	Out-of-distribution	R-152(2x)	7.09	7.861702e-06	13.08
Task 3	Out-of-distribution	R-50(1x)	53.01	4.713106e-14	10.46
Task 4	In-distribution	R-50(1x)	19.97	2.152548e-06	5.56
Task 4	Out-of-distribution	Dataset 1 (R-50(1x))	115.92	1.737752e-07	3.55
Task 4	Out-of-distribution	Dataset 2 (R-50(1x))	12.32	3.173486e-04	3.84
Task 5	In-distribution	R-50(1x)	34.87	2.018287e-06	4.28
Task 5	Out-of-distribution	R-50(1x)	19.40	9.161187e-05	3.59
Task 6	In-distribution	R-152(2x)	7.77	2.269294e-06	13.64
Task 6	In-distribution	R-50(1x)	4.29	4.640856e-04	17.51
Task 6	Out-of-distribution	R-152(2x)	15.93	1.118564e-10	14.70
Task 6	Out-of-distribution	R-50(1x)	5.61	3.515448e-05	16.49

Supplementary Table 25 | Self-Training Statistics The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 7, specifically the metrics produced for REMEDIS vs. strong supervised baseline (JFT) and standard supervised baseline (ImageNet) further improved using the self-training strategy. This t-test was done without the assumption that the variances are equal.

Distribution	Architecture	Comparison Method	T-Statistic	*p*-value	Degrees of Freedom
In-distribution	R-152(2x)	Baseline (ImageNet) + Self-training	20.11	3.165953e-12	14.93
In-distribution	R-50(1x)	Baseline (ImageNet) + Self-training	26.10	1.893045e-15	17.48
In-distribution	R-152(2x)	Baseline (JFT) + Self-training	3.83	1.242247e-03	17.74
In-distribution	R-50(1x)	Baseline (JFT) + self-training	9.22	2.136911e-07	14.30
Out-of-distribution	R-152(2x)	Baseline (ImageNet) + Self-training	11.24	1.445689e-09	17.98
Out-of-distribution	R-50(1x)	Baseline (ImageNet) + Self-training	23.11	7.005811e-12	12.90
Out-of-distribution	R-152(2x)	Baseline (JFT) + Self-training	2.54	2.067761e-02	17.92
Out-of-distribution	R-50(1x)	Baseline (JFT) + Self-training	9.00	3.058183e-07	14.19

Supplementary Table 26 | Data Efficiency Statistics. The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 18, specifically the metrics produced for REMEDIS vs. the Supervised Baseline (JFT). This t-test was done without the assumption that the variances are equal.

Task	Architecture	Data Percentage	T-Statistic	*p*-value	Degrees of Freedom
Task 1	R-152(2x)	0	2.76	1.331786e-02	16.98
Task 1	R-152(2x)	20	3.63	1.933656e-03	17.92
Task 1	R-152(2x)	50	12.77	2.331362e-10	17.68
Task 1	R-152(2x)	100	17.72	7.793245e-13	17.99
Task 1	R-50(1x)	0	7.97	2.626604e-07	17.98
Task 1	R-50(1x)	10	2.18	4.994976e-02	11.99
Task 1	R-50(1x)	20	7.86	3.178482e-07	17.97
Task 1	R-50(1x)	50	18.86	8.870438e-11	12.91
Task 1	R-50(1x)	100	26.29	7.120434e-14	14.86
Task 2	R-152(2x)	0	6.27	4.792115e-05	11.61
Task 2	R-152(2x)	20	4.69	3.102928e-04	14.61
Task 2	R-152(2x)	50	4.40	3.894299e-04	17.09
Task 2	R-152(2x)	100	8.21	1.751132e-07	17.92
Task 2	R-50(1x)	0	4.24	4.932287e-04	17.93
Task 2	R-50(1x)	10	5.77	3.875014e-05	14.82
Task 2	R-50(1x)	20	4.85	3.352278e-04	12.78
Task 2	R-50(1x)	50	11.09	5.125399e-09	16.33
Task 2	R-50(1x)	100	10.71	3.319600e-09	17.86
Task 3	R-152(2x)	0	6.75	7.829244e-06	14.47
Task 3	R-152(2x)	20	12.56	5.229848e-09	14.00
Task 3	R-152(2x)	50	17.66	1.508450e-10	13.16
Task 3	R-152(2x)	100	12.34	4.003740e-07	9.41
Task 3	R-50(1x)	0	9.19	5.085610e-06	9.46
Task 3	R-50(1x)	20	8.08	2.927787e-05	8.47
Task 3	R-50(1x)	50	6.95	7.986308e-05	8.69
Task 3	R-50(1x)	100	9.26	7.933072e-06	8.80
Task 4	R-50(1x)	0	154.09	1.267567e-10	5.14
Task 4	R-50(1x)	20	11.76	3.351396e-06	7.72
Task 4	R-50(1x)	50	14.39	5.197559e-08	10.00
Task 4	R-50(1x)	100	13.23	9.434235e-06	6.16
Task 5	R-50(1x)	10	2.50	4.056904e-02	7.08
Task 5	R-50(1x)	20	5.90	1.593510e-04	9.88
Task 5	R-50(1x)	50	7.92	1.644582e-05	9.59
Task 5	R-50(1x)	100	10.58	1.012535e-07	12.89
Task 6	R-152(2x)	0	7.24	6.630100e-06	12.99
Task 6	R-152(2x)	20	11.68	1.294562e-09	17.24
Task 6	R-152(2x)	50	10.88	2.767563e-09	17.77
Task 6	R-152(2x)	100	11.77	1.061911e-08	14.15
Task 6	R-50(1x)	0	5.87	1.624941e-05	17.61
Task 6	R-50(1x)	20	7.68	6.774580e-07	16.82
Task 6	R-50(1x)	50	7.66	6.151014e-07	17.17
Task 6	R-50(1x)	100	23.96	4.955856e-15	17.87

Supplementary Table 27 | Data Efficiency Statistics. The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 19, specifically the metrics produced for REMEDIS vs. the Supervised Baseline (ImageNet). This t-test was done without the assumption that the variances are equal.

Task	Architecture	Data Percentage	T-Statistic	*p*-value	Degrees of Freedom
Task 1	R-152(2x)	0	6.89	2.967478e-06	16.64
Task 1	R-152(2x)	20	3.65	1.907543e-03	17.61
Task 1	R-152(2x)	50	13.12	1.832681e-10	17.43
Task 1	R-152(2x)	100	23.52	7.950471e-15	17.75
Task 1	R-50(1x)	0	10.51	4.458786e-09	17.86
Task 1	R-50(1x)	20	5.93	1.300896e-05	17.96
Task 1	R-50(1x)	50	27.75	2.010226e-14	15.18
Task 1	R-50(1x)	100	26.63	5.431880e-15	16.50
Task 2	R-152(2x)	0	1.47	1.609929e-01	16.13
Task 2	R-152(2x)	20	8.16	8.407498e-07	14.52
Task 2	R-152(2x)	50	5.97	1.253138e-05	17.85
Task 2	R-152(2x)	100	9.66	1.970495e-08	17.49
Task 2	R-50(1x)	0	15.63	1.932572e-10	14.42
Task 2	R-50(1x)	20	8.17	2.718045e-05	8.45
Task 2	R-50(1x)	50	9.98	3.059094e-08	15.85
Task 2	R-50(1x)	100	11.16	1.388359e-07	11.69
Task 3	R-152(2x)	0	7.09	7.861702e-06	13.08
Task 3	R-152(2x)	20	18.57	7.939195e-09	9.57
Task 3	R-152(2x)	50	7.31	6.422479e-04	5.19
Task 3	R-152(2x)	100	7.73	1.315057e-02	2.16
Task 3	R-50(1x)	0	53.01	4.713106e-14	10.46
Task 3	R-50(1x)	20	13.78	3.765202e-10	15.61
Task 3	R-50(1x)	50	21.43	2.470986e-10	11.02
Task 3	R-50(1x)	100	18.36	1.472381e-08	9.20
Task 4	R-50(1x)	0	115.92	1.737752e-07	3.55
Task 4	R-50(1x)	20	18.89	3.125792e-06	5.52
Task 4	R-50(1x)	50	16.96	4.186965e-05	4.30
Task 4	R-50(1x)	100	20.71	2.223103e-05	4.19
Task 5	R-50(1x)	0	4.05	6.974446e-03	5.91
Task 5	R-50(1x)	20	8.73	1.006661e-05	9.12
Task 5	R-50(1x)	50	10.91	1.699557e-07	11.74
Task 5	R-50(1x)	100	18.17	4.321376e-12	15.97
Task 6	R-152(2x)	0	15.93	1.118564e-10	14.70
Task 6	R-152(2x)	20	11.85	6.268180e-10	17.98
Task 6	R-152(2x)	50	8.24	6.353943e-07	14.86
Task 6	R-152(2x)	100	20.90	2.851323e-13	16.44
Task 6	R-50(1x)	0	5.61	3.515448e-05	16.49
Task 6	R-50(1x)	20	8.36	1.530384e-07	17.62
Task 6	R-50(1x)	50	10.50	4.249521e-09	17.97
Task 6	R-50(1x)	100	19.89	4.476769e-11	12.92

Supplementary Table 28 | Data Efficiency Statistics. The table contains the corresponding two-sided t-test statistics for Supplementary Fig. 12, specifically the metrics produced for the MoCo variant of REMEDIS vs. the Supervised Baseline (JFT). This t-test was done without the assumption that the variances are equal.

Task	Architecture	Data Percentage	T-Statistic	p-value	Degrees of Freedom
Task 1	R-152(2x)	0	4.76	1.618803e-04	17.70
Task 1	R-152(2x)	10	8.90	6.387647e-08	17.57
Task 1	R-152(2x)	20	10.83	6.111205e-09	16.60
Task 1	R-152(2x)	50	20.14	2.002568e-13	17.25
Task 1	R-152(2x)	100	14.09	6.071464e-10	14.72
Task 1	R-50(1x)	0	3.82	1.824086e-03	14.24
Task 1	R-50(1x)	10	3.44	2.974323e-03	17.58
Task 1	R-50(1x)	20	10.85	2.714267e-09	17.87
Task 1	R-50(1x)	50	18.12	7.510653e-13	17.65
Task 1	R-50(1x)	100	31.26	1.572047e-16	17.10
Task 2	R-152(2x)	0	3.56	6.465063e-03	8.66
Task 2	R-152(2x)	10	6.61	1.614792e-05	13.10
Task 2	R-152(2x)	20	5.34	1.100929e-04	13.80
Task 2	R-152(2x)	50	5.79	4.804796e-05	13.92
Task 2	R-152(2x)	100	9.72	3.987116e-07	12.30
Task 2	R-50(1x)	0	0.66	5.213446e-01	10.39
Task 2	R-50(1x)	10	4.58	4.530154e-04	13.73
Task 2	R-50(1x)	20	3.85	1.296325e-03	17.00
Task 2	R-50(1x)	50	8.20	4.806925e-07	15.60
Task 2	R-50(1x)	100	7.24	1.873378e-05	10.77

Supplementary Table 29 | In-distribution Best-vs-Best Statistics. The table contains the corresponding two-sided t-test statistics for Fig. 3, specifically the metrics produced for REMEDIS vs. the Baseline Supervised (JFT) for in-distribution. This t-test was done without the assumption that the variances are equal.

Task	T-Statistic	*p*-value	Degrees of Freedom
Task 1	2.74	1.333964e-02	17.97
Task 2	11.91	8.600512e-08	11.42
Task 3	31.20	2.842220e-14	13.91
Task 4	12.74	8.643737e-04	3.11
Task 5	71.75	1.512674e-09	5.62
Task 6	8.71	8.879738e-08	17.53

Supplementary Table 30 | In-distribution Best-vs-Best Statistics. The table contains the corresponding two-sided t-test statistics for Fig. 3, specifically the metrics produced for REMEDIS vs. the Baseline Supervised (ImageNet) for in-distribution. This t-test was done without the assumption that the variances are equal.

Task	T-Statistic	*p*-value	Degrees of Freedom
Task 1	9.56	5.179113e-07	12.18
Task 2	5.10	8.239890e-05	17.48
Task 3	33.52	7.022141e-17	16.87
Task 4	19.97	2.152548e-06	5.56
Task 5	34.87	2.018287e-06	4.28
Task 6	9.31	1.612389e-07	14.57

Supplementary Table 31 | Out-of-distribution Best-vs-Best Statistics. The table contains the corresponding two-sided t-test statistics for Fig. 3, specifically the metrics produced for REMEDIS vs. the Baseline Supervised (JFT) for out-of-distribution. This t-test was done without the assumption that the variances are equal.

Task	Percentage	T-Statistic	*p*-value	Degrees of Freedom
Task 1	0	2.44	2.661145e-02	15.99
Task 1	10	3.80	1.334593e-03	17.76
Task 1	20	4.29	5.605730e-04	15.99
Task 1	50	12.77	2.331362e-10	17.68
Task 1	100	17.72	7.793245e-13	17.99
Task 2	0	4.60	4.244625e-04	13.83
Task 2	10	8.47	1.155962e-05	9.27
Task 2	20	6.28	6.872213e-05	10.67
Task 2	50	12.92	7.661154e-10	15.88
Task 2	100	14.78	1.658754e-09	13.00
Task 3	0	6.75	7.829244e-06	14.47
Task 3	10	8.06	6.161851e-06	10.98
Task 3	20	12.56	5.229848e-09	14.00
Task 3	50	17.66	1.508450e-10	13.16
Task 3	100	12.34	4.003740e-07	9.41
Task 4	0	154.09	1.267567e-10	5.14
Task 4	10	7.44	7.674238e-05	7.93
Task 4	20	11.76	3.351396e-06	7.72
Task 4	50	14.39	5.197559e-08	10.00
Task 4	100	13.23	9.434235e-06	6.16
Task 5	0	23.65	4.808754e-06	4.67
Task 5	10	2.50	4.056904e-02	7.08
Task 5	20	5.90	1.593510e-04	9.88
Task 5	50	7.92	1.644582e-05	9.59
Task 5	100	10.58	1.012535e-07	12.89
Task 6	0	7.24	6.630100e-06	12.99
Task 6	10	7.66	4.733065e-07	17.89
Task 6	20	11.68	1.294562e-09	17.24
Task 6	50	10.88	2.767563e-09	17.77
Task 6	100	11.77	1.061911e-08	14.15

Supplementary Table 32 | Out-of-distribution Best-vs-Best Statistics. The table contains the corresponding two-sided t-test statistics for Fig. 3, specifically the metrics produced for REMEDIS vs. the Baseline Supervised (ImageNet) for out-of-distribution. This t-test was done without the assumption that the variances are equal.

Task	Percentage	T-Statistic	*p*-value	Degrees of Freedom
Task 1	0	7.02	2.936831e-06	15.98
Task 1	10	4.20	6.026055e-04	16.92
Task 1	20	4.30	5.412103e-04	16.15
Task 1	50	13.12	1.832681e-10	17.43
Task 1	100	23.52	7.950471e-15	17.75
Task 2	0	15.63	1.932572e-10	14.42
Task 2	10	10.63	3.205746e-06	8.56
Task 2	20	10.32	3.102720e-05	6.43
Task 2	50	12.60	6.866748e-09	13.61
Task 2	100	16.99	1.627972e-08	9.66
Task 3	0	7.67	4.755322e-06	12.38
Task 3	10	4.91	1.384528e-02	3.20
Task 3	20	18.57	7.939195e-09	9.57
Task 3	50	5.49	1.104863e-02	3.09
Task 3	100	7.73	1.315057e-02	2.16
Task 4	0	115.92	1.737752e-07	3.55
Task 4	10	21.63	9.742293e-06	4.52
Task 4	20	18.89	3.125792e-06	5.52
Task 4	50	16.96	4.186965e-05	4.30
Task 4	100	20.71	2.223103e-05	4.19
Task 5	0	19.40	9.161187e-05	3.59
Task 5	10	4.05	6.974446e-03	5.91
Task 5	20	8.73	1.006661e-05	9.12
Task 5	50	10.91	1.699557e-07	11.74
Task 5	100	18.17	4.321376e-12	15.97
Task 6	0	15.93	1.118564e-10	14.70
Task 6	10	9.09	5.692101e-08	17.16
Task 6	20	11.85	6.268180e-10	17.98
Task 6	50	8.24	6.353943e-07	14.86
Task 6	100	20.90	2.851323e-13	16.44