arXiv:2412.08763v2 [cs.CV] 19 Jan 2026

Beyond Knowledge Silos: Task Fingerprinting for Democratization of Medical Imaging AI

Patrick Godau[1,2,3,4*], Akriti Srivastava[2,3], Constantin Ulrich[5,6], Tim Adler[2,3], Klaus Maier-Hein[1,3,4,5,6,7], Lena Maier-Hein[1,2,3,6]

[1]National Center for Tumor Diseases (NCT), NCT Heidelberg, a partnership between DKFZ and University Hospital Heidelberg,Germany.
[2]German Cancer Research Center (DKFZ) Heidelberg, Division of Intelligent Medical Systems, Germany.
[3]Faculty of Mathematics and Computer Science, Heidelberg University, Germany.
[4]HIDSS4Health - Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany.
[5]German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany.
[6]Medical Faculty, Heidelberg University, Germany.
[7]Pattern Analysis and Learning Group, Heidelberg University Hospital.

*Corresponding author(s). E-mail(s): patrick.godau@dkfz-heidelberg.de;

Abstract

The field of medical imaging AI is currently undergoing rapid transformations, with methodical research increasingly translated into clinical practice. Despite these successes, research suffers from knowledge silos, hindering collaboration and progress: Existing knowledge is scattered across publications and many details remain unpublished, while privacy regulations restrict data sharing. In the spirit of democratizing of AI, we propose a framework for secure knowledge transfer in the field of medical image analysis. The key to our approach is dataset "fingerprints", structured representations of feature distributions, that enable quantification of task similarity. We tested our approach across 71 distinct tasks and 12 medical imaging modalities by transferring neural architectures, pretraining, augmentation policies, and multi-task learning. According to comprehensive analyses, our method outperforms traditional methods for identifying relevant knowledge and facilitates collaborative model training. Our framework fosters the democratization of AI in medical imaging and could become a valuable tool for promoting faster scientific advancement.

Keywords: Knowledge Engineering, Meta-Learning, Image Classification, Transfer Learning, Multi-task Learning

Main

Learning in isolation presents a substantial challenge in several domains of machine learning (ML). Particularly in the domain of medical imaging artificial intelligence (AI), strict regulations

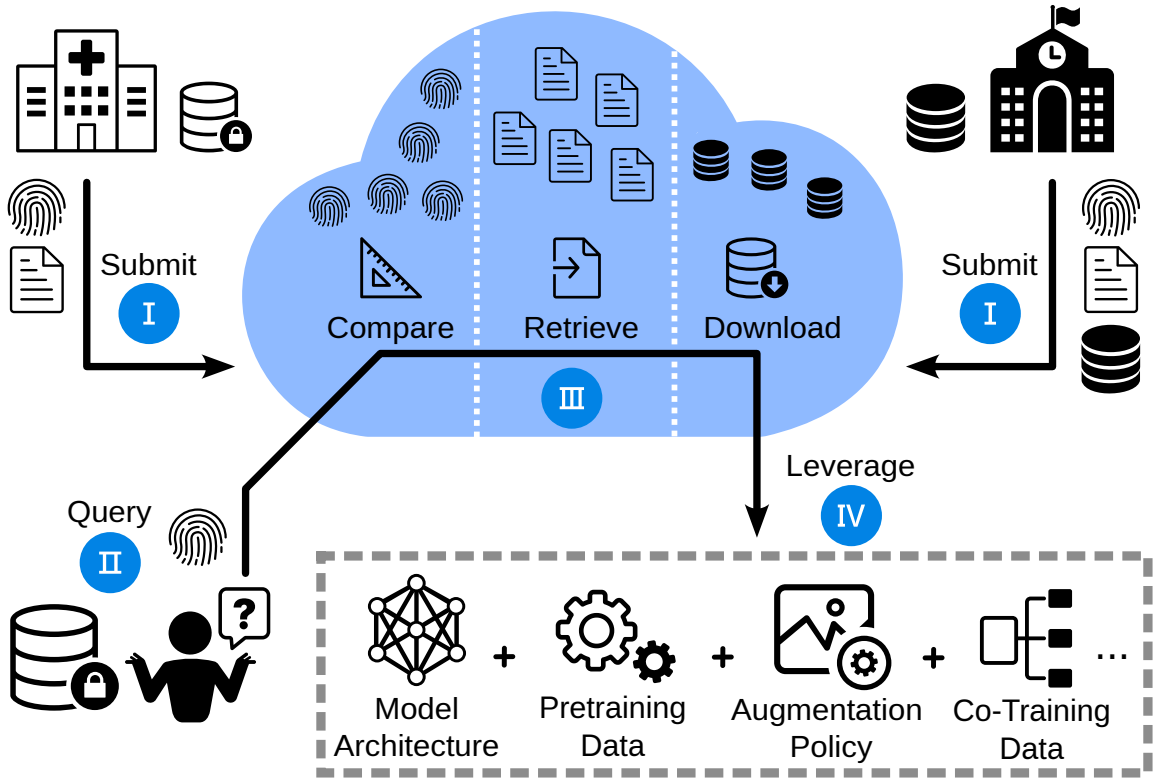

Fig. 1 Concept of task fingerprinting for knowledge transfer. (I) Participants of the knowledge cloud contribute by submitting a shareable task representation ("fingerprint"), meta information about their training strategies, and, optionally, their data. (II) A user generates the fingerprint for their task and queries the knowledge cloud. (III) Based on the most relevant tasks in the pool according to fingerprint matching, relevant training strategies and data can be retrieved. (IV) The retrieved meta information and data are used to compile a training pipeline with different components of transferred knowledge. In this study, we investigate four scenarios of knowledge transfer, namely (a) model architecture, (b) pretraining data, (c) augmentation policy, and (d) co-training data.

and high annotation costs incentivize the compilation of individual, private datasets, resulting in stakeholders creating independent **data silos**. In analogy, knowledge about successful deep learning training pipelines (including those on private data) often remains unpublished or dispersed across various publications and code repositories, making it difficult for individual researchers to effectively access and leverage past findings. The resulting partitioned landscape of **knowledge silos**, limited to individual researchers and labs, hinders overall scientific progress and puts researchers in the field at a mutual disadvantage. A recent study [1] impressively showcased this issue, reporting that researchers spend a significant amount of time selecting model architectures and augmentation strategies for each new task, underscoring the potential of improved collaboration and automated knowledge transfer in advancing the field.

The concept of **Transfer Learning** dates back to the 1970s [4] and involves any knowledge transfer from a source task to a target task [5], with its most prominent application being pretraining on large-scale datasets [6, 7]. While utilizing off-the-shelf pretrained models is a common practice to speed up model convergence [6], the vast number of available architectures, pretraining schemes, and datasets, makes selecting

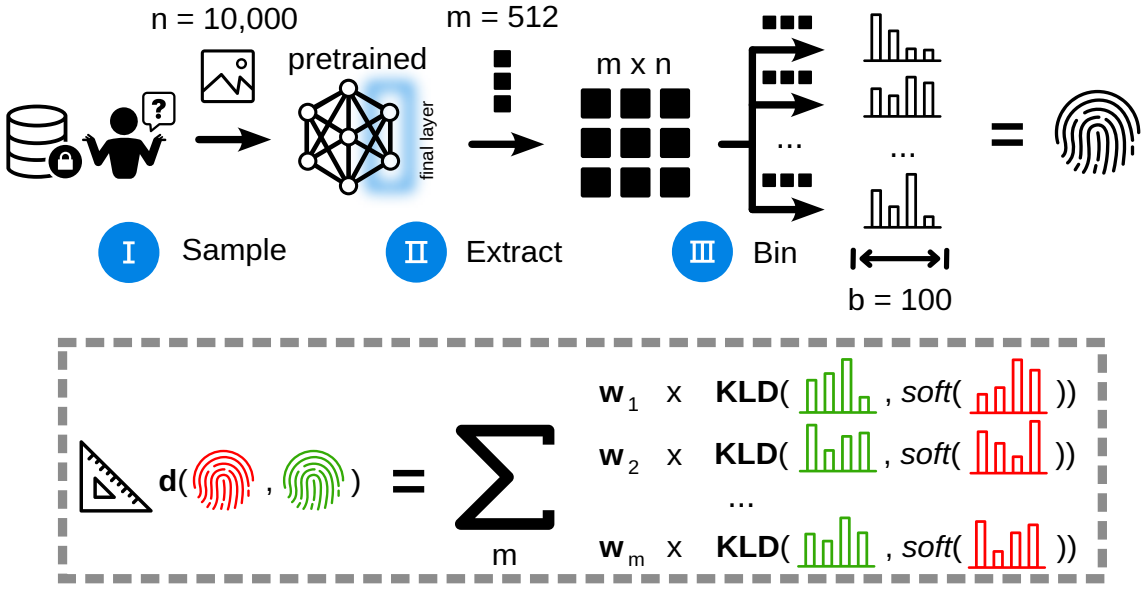

Fig. 2 Proposed fingerprinting strategy and associated task distance measure: binned Kullback-Leibler Divergence (bKLD). To compute a task fingerprint, $n = 10,000$ images are sampled (I) and passed through a pretrained backbone (we use an ImageNet [2] ResNet34 [3]) to extract $m = 512$ features per image (II). The resulting $m \times n$ features are binned into $b = 100$ bins along the features axis (III), resulting in m (normalized) histograms representing the fingerprint. To compare two fingerprints, a weighted sum of the Kullback-Leibler Divergence (KLD) across all per-feature histograms is computed. A softmax operation is applied to source task histograms to avoid empty bins.

the most suitable option a time-consuming process. This premise also holds for the nascent field of foundation models [8]. It has been shown that simply relying on benchmarks for pretrained models, such as, ImageNet [2], does not translate well to the medical domain [6]. This fueled research of **Task Transferability Estimation (TTE)**, which aims to quantify the potential of knowledge transfer between tasks [9]. While the research problem has been known for decades [10] and substantial progress has been made, challenges with respect to data privacy considerations, robustness in realistic heterogeneous data settings, and the avoidance of negative transfer remain [7, 11, 12]. Existing transferability estimation methods, which attempt to assess model suitability for a target task, often have only been applied in unrealistically homogeneous settings [13–16], lack large-scale validation (12 tasks in [17], nine tasks in [18], eight tasks in [19], seven tasks in [20], five tasks in [21]), or are incompatible with data privacy requirements [22]. Others are not suited for scalability due to computational complexity [19, 23–27] or the assumption that tasks share underlying images [28].

The core challenge of TTE lies in efficient task matching for knowledge transfer. Selecting an inappropriate source task for a new task can lead to detrimental negative transfer [11, 30]. Effective knowledge transfer requires generalizability across multiple components of the training pipeline - not just on pretraining, an aspect particularly underexplored in previous research. Additionally, creating effective task representations for knowledge transfer necessitates a balance between size (to facilitate sharing), data privacy (sufficient anonymization), and information content (retaining relevant task knowledge).

To address these issues, we propose a novel, comprehensive framework for privacy-aware, collaborative knowledge transfer in medical imaging AI (see Fig. 1). At the core of our approach lies a knowledge cloud, storing experiences of model training along with an encapsulated representation of the underlying data, which we call the **task fingerprint**. Contributors may submit training

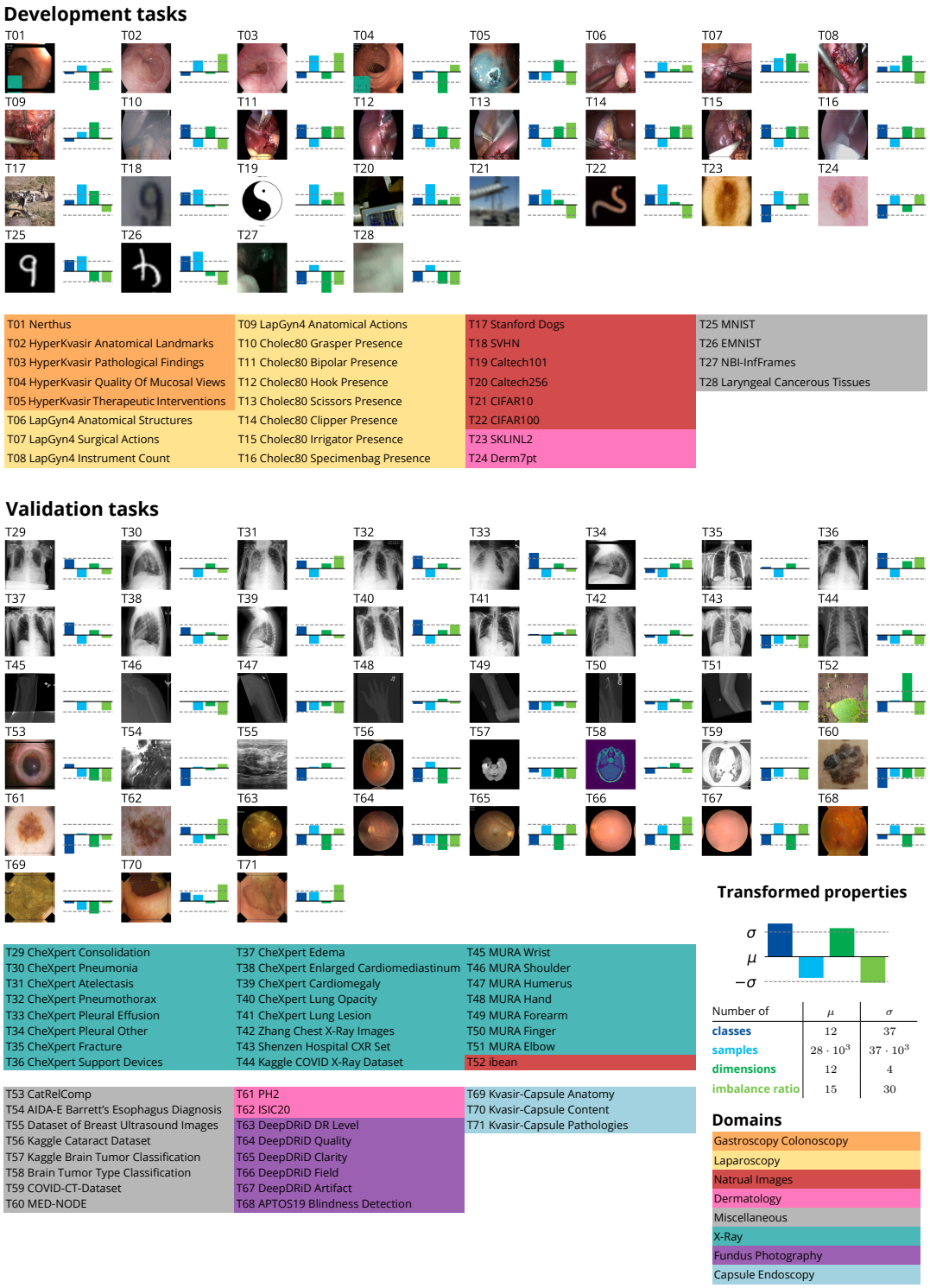

Fig. 3 Our study is based on a heterogeneous set of 71 imaging tasks. For each of the 28 tasks of the development split as well as 43 tasks of the validation split, we show a sample image next to the distribution of the following Box-Cox and Z-score transformed properties: number of classes, number of samples, intrinsic data dimension [29], and imbalance ratio (size of largest class divided by size of the smallest class). The imaging domain is encoded as the background color of the dataset name.

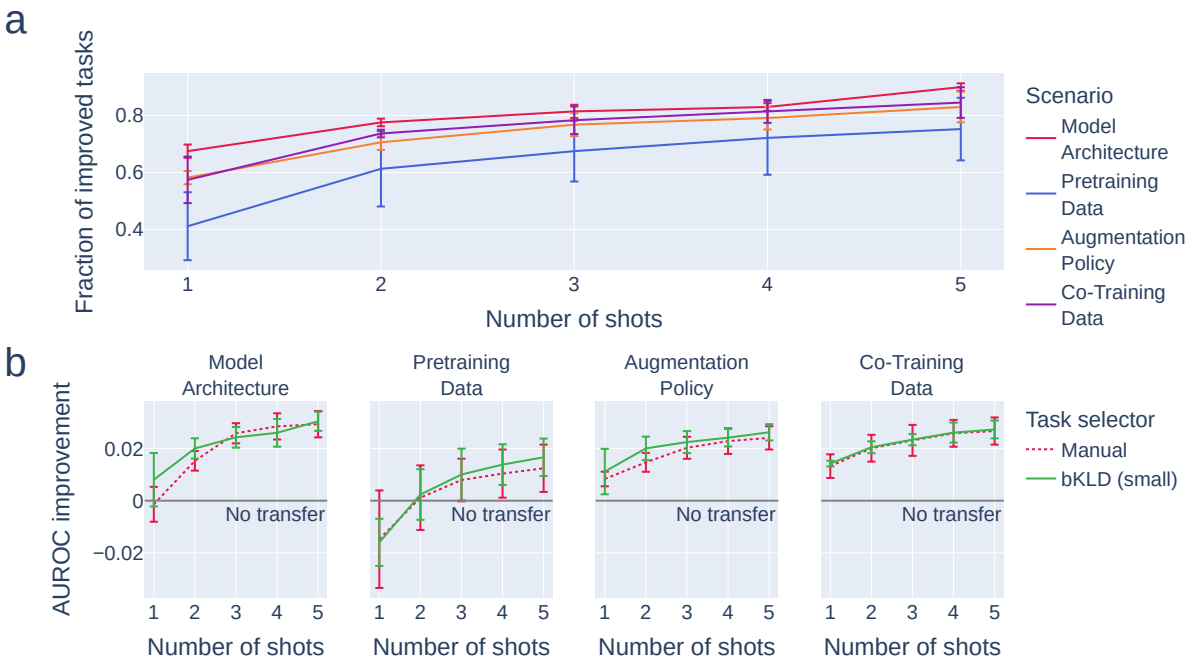

Fig. 4 Task fingerprinting benefits training pipeline configuration and beats manual knowledge transfer. (a) : Fraction of n=43 validation tasks that improve Balanced Accuracy (BA) through knowledge transfer ("gain" [28]) in four scenarios. **(b)** : Average delta in Area Under the Receiver Operator Characteristic (AUROC) across n=43 validation tasks. X-axis shows the number of shots, translating to the best of top k suggestions of our framework. Error bars correspond to standard deviation over three repetitions of all model trainings. Our proposed binned Kullback-Leibler Divergence (bKLD) fingerprint (here: the small variant) improves training for up to 90% of validation tasks.

meta information and voluntarily share the underlying data. A user may query the system for existing relevant knowledge or data by generating the fingerprint of their own task, leveraging the retrieved information during the compilation of their training pipeline. The key component of our approach is a new, efficient, and generalizable task distance measure computed from the individual task fingerprints. We refer to this distance measure as **binned Kullback-Leibler Divergence (bKLD)** (see Fig. 2), as fingerprints are generated by binning extracted image features, with similarity calculated using a weighted variant of the Kullback-Leibler Divergence. We demonstrate the effectiveness of our framework and task distance measure on the largest ever utilized set of heterogeneous tasks (n=71) in the context of medical imaging (see Fig. 3). We combine this scale in task variety with diverse knowledge transfer scenarios (n=4), namely transferring the underlying model architecture, using source task data for pretraining, leveraging automatically generated augmentation policies, and simultaneously learning a co-task. The code, data, and intermediate results of our experiments are available at https://github.com/IMSY-DKFZ/task-fingerprinting. All models and datasets are publicly available.

Results

For comprehensive validation, we generated a knowledge cloud comprising a pool of 71 tasks as shown in Fig. 3. To avoid overfitting our task fingerprints, only 28 tasks were used for development and hyperparameter tuning. The remaining 43 downstream tasks were used for prospective testing and covered the medical domains of radiology, dermatology, ophthalmology, endoscopy, and sonography.

To investigate the impact of knowledge transfer on complementary modules of the training pipeline, we queried the knowledge cloud for

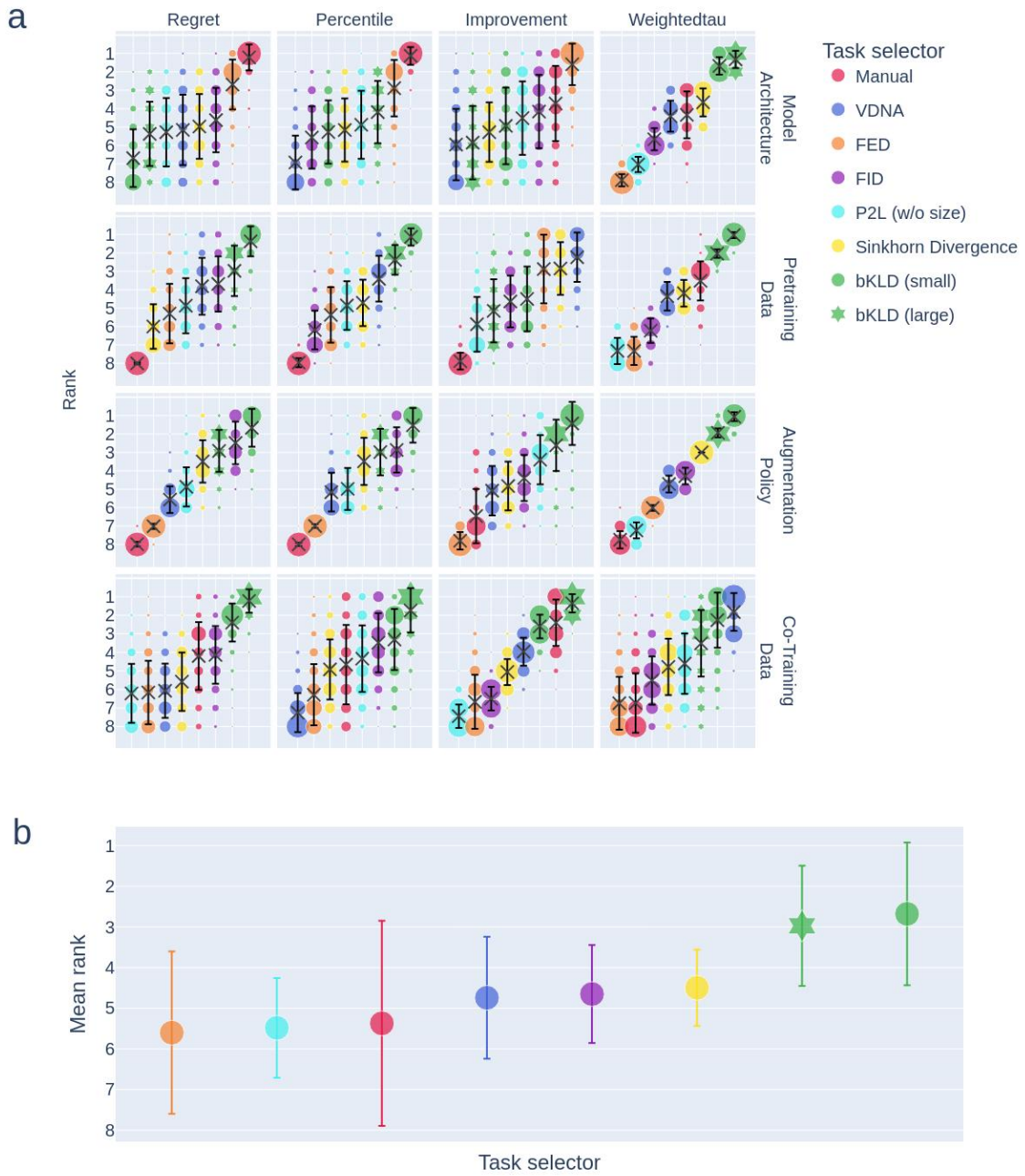

Fig. 5 binned Kullback-Leibler Divergence (bKLD) outperforms previously proposed methods for knowledge transfer. Uncertainty-aware ranking of our proposed bKLD methods for task fingerprinting versus manual task selection, VisualDNA (VDNA) [20], Fisher Embedding Distance (FED) [15, 31], Fréchet Inception Distance (FID) [32, 33], Predict To Learn (P2L) [34], and Sinkhorn Divergence [35]. **(a)**: Columns represent four meta metrics to compare distance measures, whilst rows correspond to four knowledge transfer scenarios. We average across the top three suggestions by each method (weightedtau evaluates the full candidate ranking). Blob size shows frequency of rank across 1000 bootstrap samples from 258 setups (2 base metrics, 43 validation tasks, 3 repetitions). X marks the mean rank and whiskers the standard deviation. Plot is inspired by [36] and follows the "aggregate then rank" assessment method. **(b)**: Summary of the 16 subplots below. The marker position refers to the mean rank over individual bootstrapped mean rankings (X marks) and whiskers indicate standard deviation.

relevant existing metadata, such as the best-performing model architecture and suitable augmentation policy, or for additional data samples for pretraining and co-training. As we repeated every possible knowledge transfer three times to compensate for non-determinism during model training, we trained more than 30,000 neural networks solely for the evaluation part, which resulted in roughly 10,000 GPU hours of training.

Task fingerprinting enables knowledge transfer without sharing sensible data

For a fraction of 67% of validation tasks, the resulting models exhibited an improvement by switching our baseline model architecture to the prime candidate proposed by the knowledge cloud. Replacing the augmentation strategy benefited 58% of tasks' first shot, i.e., only leveraging the augmentation strategy derived from the best matching source task in the knowledge cloud. Using shared image data of the best-aligned source task for knowledge transfer through pretraining and co-training resulted in an improvement for 41% and 57% of validation tasks. Increasing the computational budget during downstream model training and considering multiple knowledge source candidates ("multi-shot") led to up to 90% of validation tasks being improved, as shown in Fig. 4 (a). Notably, this held true when transferring model architecture (90%) and augmentation policy (83%) solely via sharing the fingerprint as a compressed identifier without any other data samples leaving the institution of the data owner. Assuming that the task data was available for pre- or co-training, we observed a similar rate of task improvement for the co-training scenario (84%), compared to our single task baseline. Considerably fewer tasks benefited from knowledge transfer via pretraining (75%), while the fluctuations in our repeated experiments due to randomness were more severe.

Task fingerprinting outperforms manual task selection

Reviewing the literature to identify previously successful training recipes has proven to be the predominant strategy chosen by participants of biomedical challenges [1]. This strategy, which we refer to as "manual" task matching, solely relies on the similarity of the semantic task description, is limited to the actual training information published on a similar task, and requires manual extraction from the publication or repository. Simulating such semantic task matching allowed us to compare our fully data-driven task fingerprinting with this prevailing strategy. The results are shown in Fig. 4 (b), which displays the average performance increase per task compared to our single task baseline. The relative performance increase by task fingerprinting was 12%, 57%, 15%, and 2% larger than manual task matching for the investigated knowledge transfer scenarios of model architecture, pretraining, augmentation policy, and co-training, respectively, when averaging across all tasks, repetitions, and the multi-shot range from 1 to 5.

Our proposed task distance measure outperforms previously suggested ones

To compare our fingerprinting technique to previously proposed task similarity measures, we instantiated our framework with previously proposed task transferability measures including VisualDNA (VDNA) [20], Predict To Learn (P2L) [34], Fisher Embedding Distance (FED) [15, 31], Sinkhorn Divergence [35], and Fréchet Inception Distance (FID) [32, 33]. Note that the underlying task representations differ for each of these, with FID and VDNA being comparable in size to our bKLD-based task fingerprints, while P2L is significantly smaller and FED as well as Sinkhorn being larger. We compare them to two variants (hyperparameter configurations) of our proposed distance measure bKLD, which obtain the top two positions in the overall ranking (see Fig. 5 b). Fig. 5 (a) shows a detailed analysis of the rankings for all mentioned methods with respect to meta metrics and knowledge transfer scenarios. Additional evaluation aggregation schemes can be found in the supplementary information (Supplementary Table 4 and Supplementary Fig. 1).

Comparing transferability measures has been shown to be a very sensitive experiment, where small variations in the setup may lead to conflicting conclusions on superiority [37]. Therefore, the **setup stability score** [37], has been proposed to

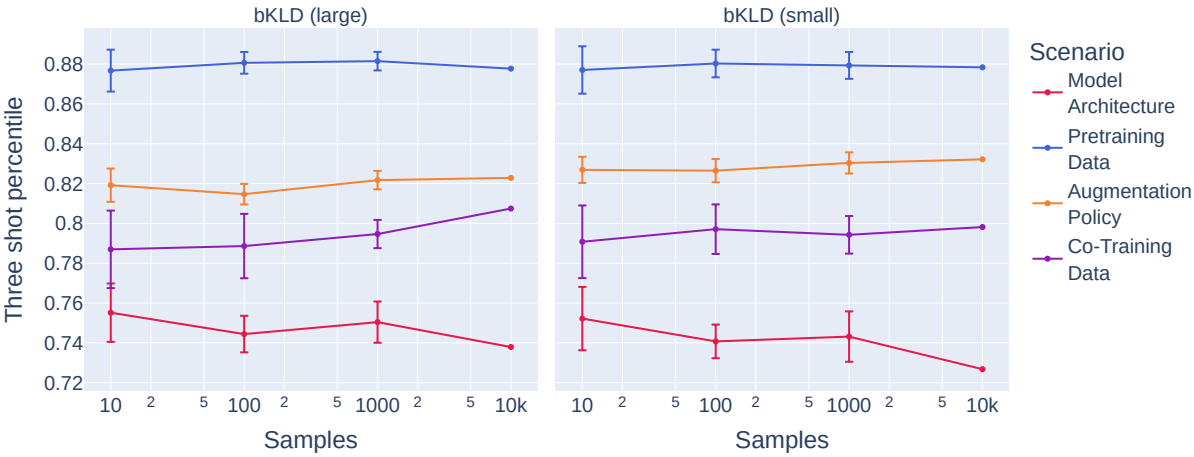

Fig. 6 binned Kullback-Leibler Divergence (bKLD) is robust with respect to dataset size. Percentile of best 3-shot source task knowledge transfer according to Balanced Accuracy (BA) averaged over n=43 validation tasks in four scenarios applying two proposed bKLD variants. X-axis shows the number of samples used from a task to generate its fingerprint in log scale. Error bars indicate standard deviation over 10 resamplings.

indicate how stable the ranking of task similarity algorithms is with respect to varying a single component of the setup, e.g. choosing a different performance metric. A value of 0 indicates no correlation between the rankings and 1 denotes perfectly identical rankings. For our experiments, the stability scores were 0.59 for the meta metrics, 0.31 for the performance metrics, 0.1 for the random seed, 0.02 for the target tasks, and only 0.01 for the knowledge transfer scenario.

Our proposed task distance measure is robust with respect to task size

As fingerprint generation is a non-deterministic process and we envision our methodology to be especially useful for small tasks, we varied the number of samples used to compute a fingerprint. While all previous experiments (including all alternative task similarity measures) evaluated 10,000 samples from a task, we tested significantly reduced sizes of up to only 10 samples. Figure 6 shows our method's stability in retrieving sufficiently beneficial task candidates in a best of three shots scenario, where random task selections would be expected to achieve a value of 0.75. Based on fingerprints generated from only 10 samples both bKLD variants show robust task matching performance. Of note, knowledge transfer for pretraining data has the best *relative* task matching of the four scenarios investigated, meaning that out of the available sources typically ones around the 88th percentile are chosen (Fig. 6). This stands in contrast to the *absolute* improvement to the validation tasks as shown in Fig. 4, where applying pretraining knowledge transfer benefited the least across the scenarios investigated.

Our proposed task distance measure generalizes to unseen task type and feature extractors

To further test the generalization capabilities of bKLD, we investigated two variations of our experiments. First, we applied task transferability estimation on the outcomes of the Medical Segmentation Decathlon [41]. In this international competition 19 participating teams submitted segmentation models for 17 segmentation tasks on 10 datasets. Data were obtained from the two most common 3D medical imaging modalities, namely 3D magnetic resonance imaging (MRI) and computed tomography (CT). Segmentation tasks were highly diverse, ranging from the segmentation of large objects (e.g., liver) to the segmentation of tubular structures (e.g., vessels). Note that

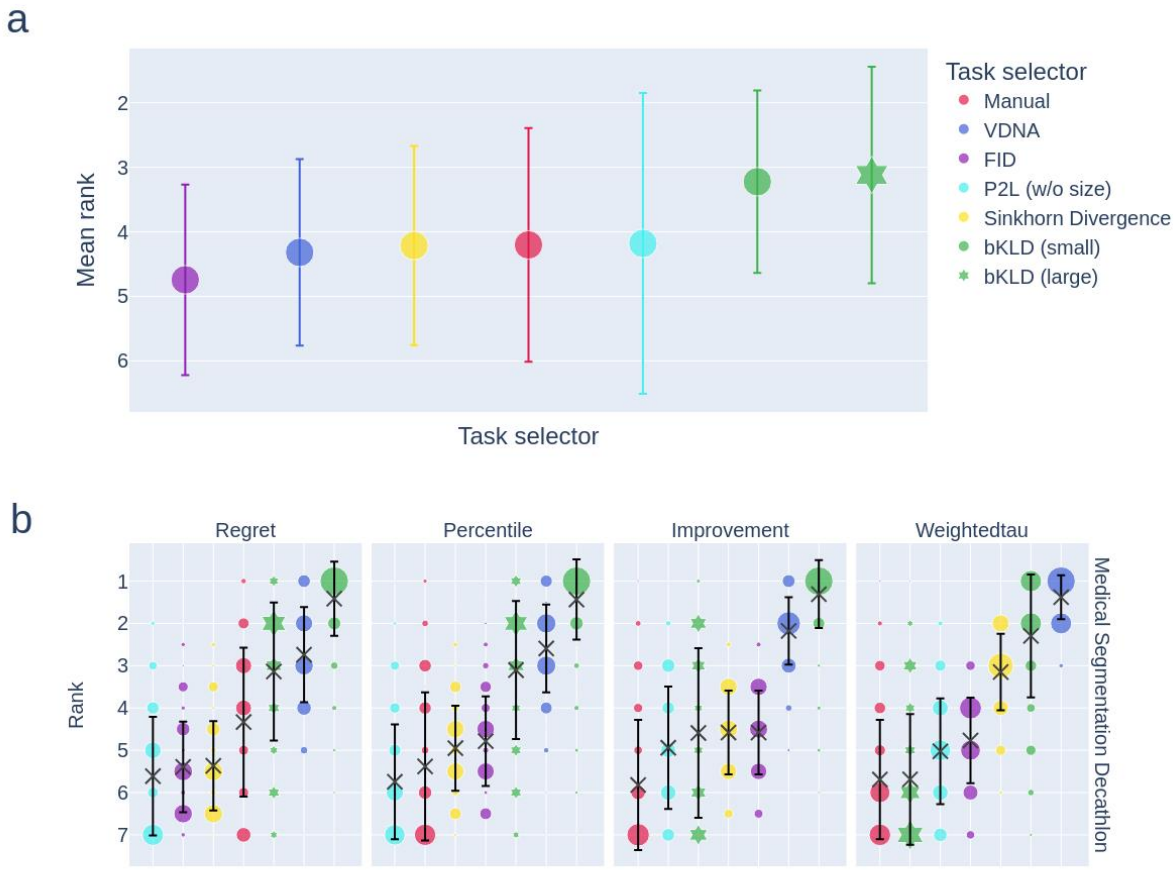

Fig. 7 binned Kullback-Leibler Divergence (bKLD) generalizes to 3D, new task types and further backbone models. Uncertainty-aware ranking for 2D classification and 3D medical image segmentation tasks demonstrating strong performance of our proposed bKLD methods for task fingerprinting compared to manual task selection, VisualDNA (VDNA) [20], Fréchet Inception Distance (FID) [32, 33], Predict To Learn (P2L) [34], and Sinkhorn Divergence [35]. **(a)**: Extension of Fig. 5 to three additional feature extraction backbones. Summary of 48 subplots (16 from a CLIP [38], DINO [39], and MAE [40] backbone respectively). The marker position refers to the mean rank over individual bootstrapped mean rankings (X marks) and whiskers indicate standard deviation. **(b)**: Extension of Fig. 5 to 3D medical imaging using data from the Medical Segmentation Decathlon [41]. Columns represent four meta metrics to compare distance measures. We average across the top three suggestions by each method (weightedtau evaluates the full candidate ranking). Blob size shows frequency of rank across 1000 bootstrap samples from 34 setups (2 base metrics, 17 validation tasks). X marks the mean rank and whiskers the standard deviation. Plot is inspired by [36] and follows the "aggregate then rank" assessment method.

the task type (segmentation versus classification), the modalities (2D versus tomographic 3D) and the target structures (e.g., vessels) all represent a severe shift compared to our previous experiments [42]. The outcomes are shown in Fig. 7 (b) and are consistent with our previous findings. Concretely, the small variant of bKLD performs best in three out of four meta metrics. Second, we replaced the feature extractor used in our previous experiments with three more recent models based on CLIP [38], DINO [39], and MAE [40]. These experiments represent shifts in the distribution of the extracted features and consequently the resulting task fingerprints. Fig. 7 (a) shows the aggregated pendant to Fig. 5 (b), i.e., the ranking of fingerprinting methods across meta metrics,

transfer scenarios, and feature extractors, and is consistent with our main experiments.

Discussion

This work introduced task fingerprinting as a novel approach to AI democratization. These shareable task embeddings enable knowledge transfer estimation, allowing researchers to collaboratively collect and leverage insights during neural network experiments. In particular, our method introduces a new task similarity measure named **binned Kullback-Leibler Divergence (bKLD)**, which is both robust and efficient in computation, as well as flexible to adapt to various knowledge transfer scenarios. Importantly, it outperformed previous approaches in our experiments which were, to the best of our knowledge, conducted using the largest heterogeneous set of tasks in combination with the broadest evaluation across knowledge transfer scenarios in the domain of task transferability estimation research. The key insights drawn from this comprehensive analysis are:

1. **Task fingerprinting allows for secure knowledge transfer.** Our findings demonstrate the potential of task fingerprinting as a method for knowledge sharing, particularly in the data-sensitive domain of medical imaging. Our approach improved downstream task performance for a majority of tasks within a reasonable computational budget, while outperforming the prevailing and time-consuming practice of manual knowledge compilation.
2. **The best fingerprinting strategy varies substantially across scenarios.** The reason we did not identify a one-size-fits-all solution (see Fig. 5 a) is that the optimal source task rankings are inconsistent (see Supplementary Table 3), hence no single task distance measure might suit all transfer scenarios. Despite this, bKLD demonstrates strong generalization capabilities.
3. **Task distance measures benefit from feature binning.** Our newly introduced task distance measure bKLD outperformed existing measures . While the Predict To Learn (P2L) [34] approach computes the Kullback-Leibler Divergence (KLD) on averaged features, the binning strategy in bKLD allows a much more granular comparison of image feature distributions. The introduced weighting mechanism further provides an option for prioritization of dominant features of the target task.

While our method is straightforward to use, it comes with some important hyperparameters. The number of bins allows to control the trade-off between representation granularity on the one hand and privacy and efficiency on the other hand. The weighting mechanism allows to make use of the same representation in multiple ways, thus enables bKLD to adapt to various scenarios of knowledge transfer. Importantly, the comparability of fingerprints requires an identical feature extraction model and at least a similar preprocessing method. Otherwise the j-th feature dimension of model A is unlikely to be directly related to the j-th feature dimension of model B, and the total number of feature dimensions may differ. Although our experiments (Fig. 7) demonstrate that bKLD generalizes well beyond a single feature extractor, major considerations regarding the maintainability of knowledge in the knowledge cloud remain. In theory, it would be feasible to support multiple fingerprinting pipelines (comprising preprocessing and feature extraction) in parallel for images with widely varying natures (e.g., histopathology and 3D imaging) and occasionally update these pipelines to support multiple extraction models over time. This approach would provide sufficient flexibility to introduce more powerful feature extractors while avoiding complete erasure of existing knowledge with the introduction of a single new extraction pipeline.

Compared with previous methods, bKLD performed poor in the model architecture scenario (see Fig. 5 a), which also obtained the lowest percentile score (see Fig. 6). Furthermore, manual task selection suited this scenario better than all others, presumably indicating that granular feature distributions have a subordinate role in the fitting of model architectures. For this specific scenario, also known as **source-free model transferability estimation**, many solutions have been proposed, but the problem remains unsolved [12]. On a wider scope our generalization experiment on the Medical Segmentation Decathlon shows that the transfer of a complete model pipeline may be better predicted (Fig. 7 b). For the pre-training scenario, **bKLD(small)** proposed source

tasks ranked highest over all scenarios (see Fig. 6). Interestingly, in absolute terms, the performance improvement was low compared to the other training scenarios (see Fig. 4), i.e., for pretraining the bKLD method appeared to perform best, but the improvement is still small compared to other scenarios. We interpret this as a lack of tasks that are suited well for pretraining in our experimental data pool, as pretraining tasks ideally are large in size, and comprise sufficient variety in sample distribution to cover the downstream task [43, 44]. Within the augmentation policy scenario **bKLD**(small) outperformed all other task selectors consistently on all meta metrics (see Fig. 5), demonstrating the strength of its image representation. In the co-training scenario, where the unweighted large bKLD variant performed best, it appears intuitive that focusing primarily on one task's dominant features does not yield better transfer estimation. Though the **bKLD(small)** and **bKLD(large)** variants' meta metrics were very close to each other in most cases (see Fig. 5, 6, Supplementary Fig. 1).

While the knowledge cloud may hide the actual task fingerprints from users by returning only distances and metadata, the central service will always see the full task fingerprints. To preserve privacy, it is necessary to sufficiently obfuscate patient information within the fingerprints to prevent attacks such as membership inference or data reconstruction [45]. **bKLD** offers three mechanisms to accomplish this: aggregation, decoupling, and quantization. The binning procedure is key here. In this step feature vectors of individual images are broken up (i.e., decoupled or shuffled), and each entry is assigned to separate bins (i.e., quantized or perturbed). Moreover, the final fingerprint consists of the aggregated histograms for all images. Although we do not provide full theoretical evidence, we can point to the literature, where these techniques are common tools in differential privacy [46].

Consideration of computational costs is highly relevant in the context of both democratization of AI and environmental impact. Related to this work, two different perspectives need to be distinguished: The resources for developing and validating the present approach and the resources required for using our method. The "brute force" approach of combining every potential knowledge transfer between two tasks of our task pool was required to assess the meta metrics weightedtau, percentile and (most of the time) regret. But these are not the costs incurred during regular use: Computing task fingerprints (i.e., feature extraction and binning) took us only about 2 minutes for all tasks together on a standard workstation with 12 CPU cores and a single RTX 3090. A regular participant has to compute only a single task fingerprint once, which stored as a numpy array, makes about 0.4 MB (100 bins) to 4.1 MB (1000 bins). For every conducted experiment during model development, the outcome and hyperparameters for the corresponding fingerprint can be submitted in an automated fashion, without any noticeable overhead. The most compute intensive part remains the task distance computation in the cloud during a query, where all previously submitted fingerprints need to be considered. Fortunately, computing B (number of bins) times the Kullback-Leibler Divergence is a "cheap" operation and the computations for each existing fingerprint can be parallelized. In our implementation, computing distances for all combinations of task pairs within the task pool of size 71 took about 4 seconds for the 1000 bins bKLD variant (including data transfer to GPU). For an actual query not all internal combinations must be computed, hence our measurement roughly corresponds to a usual query with a knowledge cloud of a couple thousand distinct task submissions (except for the GPU data transfer). This computation time scales linearly with the size of the knowledge cloud. So on standard consumer hardware and without further potentially existing optimizations, our implementation should already scale to 100,000s of distinct fingerprints.

Closely related to task fingerprinting is the field of **Automated Machine Learning (AutoML)**, which aims to fully automate the configuration of ML algorithms, specifically steps such as data preprocessing, feature engineering, and hyperparameter optimization [47]. It shares the goal of AI democratization by resolving tricky steps in ML and empowers non-experts to generate AI systems. However, AutoML often raises computational expenses, especially in approaches such as Neural Architecture Search (NAS) [14, 48], and, in contrast to our approach, lacks a collaborative mechanism. Similarly, frameworks such as **nnU-net** [49] have shown promise in offering self-configuring solutions for biomedical image

segmentation tasks. Although training adjustment in nnU-net is way faster and substantially less resource-intensive than traditional AutoML, it is limited to the static rules designed by the authors (i.e., no new knowledge is gained automatically over time), focuses on a specific task type (only segmentation) and excludes crucial aspects of the training pipeline (e.g., architecture template and data augmentation are fixed). **Federated learning** [50], on the other hand, offers a path for collaborative learning while preserving data privacy, but imposes additional burdens related to synchronization, trust establishment, and preparation [51]. Task fingerprinting combines the efficiency of nnU-net with the flexibility of AutoML, while leveraging collaborative efforts such as Federated Learning. Similarly to the latter, a general limitation of task fingerprinting is that the system is vulnerable to malicious actors submitting false data in a public setting. This requires the development of preventive measures, such as detection algorithms [52] or trust scores [53]. In addition, the success of the overall concept relies on the willingness of researchers to share knowledge within the community. Perhaps most importantly and in contrast to all three mentioned approaches, however, our method has the potential to facilitate **Lifelong Learning** [54] by continuously and automatically growing knowledge, as more data and tasks are incorporated over time. This capability could lead to long-term scalability of AI which would reduce energy consumption and carbon footprint of systems [55].

It is important to consider our findings in the context of previous research on task distance measures. In line with our preliminary work [31] the FED task selector ranks high for the model architecture and pretraining scenarios and some meta metrics. However, it falls short on the other meta metrics and knowledge transfer scenarios. Prior studies have also highlighted the inconsistency of transferability estimations in general [37] and specifically in the medical domain [56]. This aligns well with our observations regarding the low stability scores of existing distance measures across target tasks and the transfer scenario and emphasizes the importance of finding a flexible approach that can adapt to multiple transfer scenarios. In fact, Standley et al. [57], also observed that task similarity measures might be suited for one kind of knowledge transfer, e.g., pretraining, but inappropriate for another, e.g., co-training. From this perspective, the consistency with bKLD is remarkable. bKLD also excels as a task selector for 3D segmentation tasks outside of our original experimental design. Our aggregation over multiple metrics and meta metrics was intended to counteract the sensitivity of evaluations to specific metrics [58] and increase the assessment robustness as well as allow for the consideration of multiple knowledge transfer scenarios. We observed that weightedtau allows for the most stable separation of task selectors, as opposed to regret, percentile, and improvement (Fig. 5 a , Supplementary Fig. 1). However, it is hard to interpret and circumvents a resource limitation-based cutoff.

The emergence of foundation models [8] has amplified the importance of AI democratization, as the necessary expertise, computational resources, and data volumes have grown substantially. These models could bridge the gap between different modalities in medical imaging through their large-scale pretraining [59]. However, two fundamental problems accompany them: first, selecting the most suitable one from the diverse landscape of available ones , and second, devising an effective adaptation strategy for specific downstream tasks. Although integrating multiple modalities into fingerprints is a topic for future research, we note that task fingerprinting is a promising approach to addressing these challenges. Notably, the computational overhead of task fingerprinting – both in terms of archiving post-experimental knowledge and querying prospective knowledge – is minimal, especially compared to the actual model training process. Therefore, task fingerprinting has the potential to accelerate the selection and adaptation of foundation models.

The actual implementation of a knowledge cloud remains a future challenge for the realization of our concept. Using existing infrastructure, such as the "Joint Imaging Platform" [60], where cross-institutional research in the field of medical imaging is already being carried out, would provide an excellent starting point. Additionally, as our experiments treated each transfer scenario in isolation, investigating the impact of transferring multiple components of a pipeline at the same time (entangled transfer scenarios) is a crucial

next step. Expanding upon our evidence by testing other task types, a larger number of target tasks, and other parts of the training pipeline is also essential.

We consider the broader implications of our research to be significant: The widespread adoption of knowledge clouds based on task fingerprinting has the potential to democratize AI research by facilitating collaboration and knowledge sharing. Shared knowledge, in turn, can lead to reduced model development times and a decrease in carbon emissions associated with extensive training processes.

In conclusion, this study provides compelling evidence that task fingerprinting offers a promising solution to overcome knowledge silos and enhance knowledge transfer in medical image analysis. Our newly proposed task distance measure bKLD demonstrates clear advantages across various knowledge transfer scenarios and exhibits large flexibility to be applied in a wide range of use cases. Future research is set to focus on expanding the framework's capabilities and exploring the broader impact on medical AI research.

Methods

Our knowledge transfer framework is based on the following concept (Fig. 1): Any contribution by a participant of the network is associated with a task fingerprint, i.e., a shareable task representation that respects individual privacy regulation (e.g., that the data itself must not be shared). Meta information about suitable training design decisions, such as model architecture (e.g., a *timm* [61] string) or augmentation policy (e.g., an *albumentations* [62] serialization), is stored alongside the fingerprint. Optionally, the data may be tagged as shareable. To learn from the current network of tasks, a user generates the fingerprint for their task and queries the knowledge cloud. Based on the most relevant tasks in the pool according to fingerprint matching, relevant training strategies and data can be retrieved. The retrieved meta information and data are used to compile a training pipeline with different components of transferred knowledge. This section details our core methodological contributions as well as the evaluation concept we developed to assess the value of our approach.

Task fingerprinting

The major challenge of task fingerprinting is the efficient extraction of relevant task information into a fingerprint without compromising private information. Formally, we define a **task** $\mathfrak{T}$ as $\mathfrak{T} = \{(x_i, y_i)\}_{i=1}^{N}$ comprising N samples, each consisting of an image x_i and a label y_i, following Achille et al. [9]. The process of **task fingerprinting** refers to computing a mapping f of a task $\mathfrak{T}$ onto a set of real valued vector from within $\mathbb{R}^n$. Generated fingerprints of tasks must be comparable to measure similarity (and thus the corresponding potential of knowledge transfer) between tasks. This is done with the help of a **distance measure** d (not necessarily symmetric or positive) that maps two fingerprints onto a real value ("the lower, the higher the similarity").

Our proposed fingerprinting method is visualized in Fig. 2. To generate the fingerprint of a task $\mathfrak{T}$ we feed a fixed amount of samples ($n = 10,000$) through an ImageNet [2] pretrained ResNet34 [3] extracting n deep feature vectors $\{\mathbf{v}_i \in \mathbb{R}^m\}_{i=1}^{n}$ from the activations ($m = 512$) prior to the classification head. Samples are drawn randomly with replacement and slight data augmentation (RandomCrop and HorizontalFlipping) is applied. Following the strategy of Ramtoula et al. [20], we distribute the activations of each feature dimension $\{1, .., m\}$ into b uniform bins - creating m activation histograms. After normalizing along the feature dimension the resulting feature distributions $\{\mathbf{p}_i \in [0,1]^b\}_{i=1}^{m}$ can be shared without revealing the features of individual samples.

To compare these fingerprints we use the Kullback-Leibler divergence [63], which has been used for this purpose before [34, 64], jointly with an optional weighting of features. The weights are hyperparameters that allow to shift focus on features that are particularly relevant for either the source or target task. Assume some weights $\mathbf{w} \in \mathbb{R}^m$ and tasks $\mathfrak{T}, \mathfrak{S}$ with $f(\mathfrak{T}) = \{\mathbf{p}_i \in [0,1]^b\}_{i=1}^{m}$ as well as $f(\mathfrak{S}) = \{\mathbf{q}_i \in [0,1]^b\}_{i=1}^{m}$. Then we define

$$
\begin{aligned}
d(f(\mathfrak{S}), f(\mathfrak{T})) &= \sum_{i=1}^{m} w^{(i)} \cdot \mathbf{KLD}(p_i, \mathrm{soft}(q_i)) \\
&= \sum_{i=1}^{m} w^{(i)} \cdot \sum_{j=1}^{b} p_i^{(j)} \cdot \log \frac{p_i^{(j)}}{\hat{q}_i^{(j)}},
\end{aligned}
$$

where soft is the **softmax** function such that $\hat{q}_i^{(j)} = \exp q_i^{(j)} / \sum_j \exp q_i^{(j)}$. The softmax is used to circumvent potentially empty bins. The key novelty of our approach is the combination of a sufficiently granular task embedding (compared to P2L [34]) with a flexible weighting to guide feature attention (compared to VDNA [20]).

Hyperparameter optimization

During the development phase, we began by exploring variations of previously proposed methods for task similarity [15, 20, 31, 33–35, 65]. We combined various embedding mappings f, including the binning scheme, with multiple distance measures d. We examined several smoothing functions (e.g., normalization, softmax, and symmetric uniform smoothing) for both the source and target embeddings, as well as weighting schemes. The list of all configurations can be found in our code repository. Based on our exploratory experiments on the development tasks (see Fig. 3), we identified two favorable settings of hyperparameters: **bKLD(small)** a "small" fingerprint with $b = 100$ bins and weighting by the softened feature average of the target task $w = \text{soft}(\sum_i^n \mathbf{v}_i/n)$ as well as **bKLD(large** with $b = 1,000$ and uniform weighting. To compare bKLD with the manual task matching approach, we extracted keywords from the semantic description of tasks, e.g., imaging modality, anatomical regions, and entities of interest, and computed the intersection over union between any pair of tasks. Ties were resolved by preferring larger source tasks.

Knowledge transfer

A **task selector** is a function s, that given a (target) task $\mathfrak{T}$, and a set of potential source tasks $\mathbf{P}$ (not including $\mathfrak{T}$), returns a subset of $\mathbf{P}$ that are called candidates. For simplicity, we also allow s to return a single candidate (instead of a singleton). Given a fingerprinting mapping f and a distance measure d, the map $s_{f,d}(\mathfrak{T}) := \text{argmin}_{\mathfrak{S} \in \mathbf{P}} \, d(f(\mathfrak{S}), f(\mathfrak{T}))$ becomes a task selector.

Assume t to be a **trainer** for a model on target task $\mathfrak{T}$ that defines all elements of the training pipeline (e.g., data augmentation, neural architecture, data sampling, ...), while potentially involving information from an additional source task $\mathfrak{S}$. The kind of knowledge transfer between source and target (e.g., pretraining, co-training) is not restricted in our framework. With $t_{\mathfrak{S}}$ we note the fully characterized training procedure of the model using $\mathfrak{S}$ as source task and t as a trainer. For a performance metric m that evaluates models on their target task based on a held-out test set, $m(t_{\mathfrak{S}}(\mathfrak{T}))$ describes the performance for task $\mathfrak{T}$ after the training of $t_{\mathfrak{S}}(\mathfrak{T})$ has taken place. As a baseline comparison $m(t(\mathfrak{T}))$ measures the performance for task $\mathfrak{T}$ after training a model with information solely given by $\mathfrak{T}$. Given m is oriented positively ("the larger the better"), then we call the knowledge transfer "negative" if $m(t_{\mathfrak{S}}(\mathfrak{T})) < m(t(\mathfrak{T}))$ [30].

Validation metrics

Following recent recommendations [66], we measured model performance with two complementary and broadly applicable **base metrics** m, namely Balanced Accuracy (BA) and (macro averaged) Area Under the Receiver Operator Characteristic (AUROC). In order to evaluate the quality of a task selector s for a given target task $\mathfrak{T}$ upon a pool of potential source tasks $\mathbf{P}$, a variety of **meta metrics** exists. Assume s chose source task $\mathfrak{S}$, and m, t are given, let $\mathbf{O} := \{m(t_{\mathfrak{S}_i}(\mathfrak{T})) | \mathfrak{S}_i \in \mathbf{P}\}$ be the set of all possible transfer outcomes, then:

Improvement measures the difference in applying the knowledge transfer via s compared to the baseline: $m(t_{\mathfrak{S}}(\mathfrak{T})) - m(t(\mathfrak{T}))$. **Percentile** measures the relative position of $m(t_{\mathfrak{S}}(\mathfrak{T}))$ within all potential transfer outcomes: $|\{o \in \mathbf{O} : o \leq m(t_{\mathfrak{S}}(\mathfrak{T}))\}|/|\mathbf{O}|$. **Regret** [67] measures how much of the remaining performance gap (from the performance determined by s to optimal performance) could be overcome by making the optimal choice within $\mathbf{P}$: $[\max \mathbf{O} - m(t_{\mathfrak{S}}(\mathfrak{T}))]/[1 - m(t_{\mathfrak{S}}(\mathfrak{T}))]$. **Gain** [28] is a weaker form of improvement and returns 0 for a negative transfer, otherwise 1.

Note that it is a realistic scenario to query for multiple shots of s and retrieve the top k candidates $\mathfrak{S}_1, ..., \mathfrak{S}_k$. For evaluation aggregation, one may either average over the individual results of selections and assess the expected quality of any suggestion (as we did in Fig. 5) or proceed only with the best suggestion to assess the expected outcome (as we did in Fig. 4 and 6). In contrast, **weightedtau** [37] considers the full

ranking of candidates in $\mathbf{P}$. It is a weighted version of the Kendall rank correlation coefficient. We use a hyperbolic drop-off in importance as it is implemented in scikit-learn [68]. This allows us to compare the ranking of candidates by s with the ranking of all potential sources via the actual outcomes $\mathbf{O}$.

Data

To guarantee broad applicability of our method, we gathered a very heterogeneous task pool from publicly available datasets. Figure 3 gives an overview of all used tasks, which amount to 71 in total. Inclusion criteria for datasets were (1) public availability, (2) use of 2D images, and (3) provided classification labels. A focus was laid on datasets from the medical domain (62 tasks), as well as a great variety in imaging domains (laparoscopy, dermatoscopy, gastroscopy, colonoscopy, ophthalmic microscopy, confocal laser endomicroscopy, MRI, X-ray, fundus photography, CT, capsule endoscopy, ultrasound). Supplementary Table 2 provides more details on the data sources [2, 69–103]. Each task was split into train and test sets ensuring proportions of 80:20 and equal class distributions between splits. All images were preprocessed to match a resolution of 256×256 pixels. Greyscale images were further transferred to RGB. For tasks that exceeded 1,000 samples, we subsampled a shrunk version of 800 train samples (following [67]) that served as a target task variant. This served to reduce compute resources, increase comparability across target tasks, and allow knowledge transfer to have a significant influence as, e.g., pretraining is the most advantageous in scenarios where the target task $\mathfrak{T}$ is small [6]. This implies that all task distances are estimated from a "complete" source task to a "shrunk" target task. We partitioned the set of all tasks into a development set $\mathbf{P}_{\text{dev}}$ and a validation set $\mathbf{P}_{\text{val}}$. We ensured that all tasks related to the tasks we used in our previous study [31] were selected as development tasks. Tasks from the validation split were masked for the process of hyperparameter selection for bKLD. Note that there is a severe (imaging) domain shift from development to validation tasks, challenging the generalizability of our task selector. Furthermore, note the high variability in task size (min:170, q1:1572, q3:40'673, max:122'138), number of classes (min:2, q1:2, q3:5, max:257), imbalance ratio (min:1, q1:1.34, q3:14.27, max:171.33), and resolutions. To the best of our knowledge, no imaging task similarity measure has been evaluated this extensively before.

Validation of knowledge transfer quality

The purpose of our experiments was to demonstrate the benefit of task fingerprinting for a variety of knowledge transfer scenarios. We therefore designed four realistic scenarios, covering both cases with only source task metadata available ("Model Architecture" and "Augmentation Policy") as well as cases with source task data available ("Pretraining Data" and "Co-Training Data"). For the former, we created separate knowledge clouds and conducted initial experiments as described below. As a baseline, we also trained models on all (shrunk) target tasks in isolation. During our search for optimal bKLD hyperparameters, we fixed $\mathbf{P}_{\text{dev}}$ both as a pool of source and target tasks. For the final evaluation, we set $\mathbf{P}_{\text{val}}$ as a pool of target tasks and allowed both development and validation tasks as a pool of available source tasks. In any case of image overlap between two tasks, we excluded those tasks as respective potential source tasks (the DeepDRiD [101], Cholec80 [72], and CheXpert [84] tasks).

To facilitate comparison, we included a selection of previously proposed task selectors based on task distances as baselines. These were chosen to balance "classic" methods, such as FID [32, 33]), with more recent methods, such as VDNA [20]). We also compared bKLD with approaches that use more dense task representations, such as Sinkhorn Divergence [35], as well as approaches that use much simpler representations, such as P2L [34]). FED [15, 31] complements these baselines because it is not based on feature extraction. The "manual" baseline was intended to capture a common and tangible approach to the problem.

Single task baseline

The single task baseline comprises a "transfer-free" training recipe and subsequently simulates a model performance that is indicative of an achievable result without significant resource investment

in experimentation. Coming up with such a fair and meaningful training scheme for each task presents a major challenge given the large heterogeneity between them. Too much individual optimization per task is costly and introduces another potential factor of bias; on the other hand, applying a uniform training scheme on all tasks might lead to unrealistic performances. We aimed for a balance with the following strategy: For faster convergence and more stable training, we use ImageNet [2] pretrained models throughout [6]. Starting from the recommendations given by Wightman et al. [104], we made slight adaptations to the training pipeline to ensure model convergence, avoid overfitting, and generally improve performance on $\mathbf{P}_{\text{dev}}$. On an individual task level, we used the automated tuning of the learning rate [105] that is implemented in Pytorch Lightning [106]. Medical data often suffers from class imbalance, which can affect model performance. To address this important issue, we worked with a variety of imbalanced tasks, measuring imbalance as the ratio of the size of the largest class to the size of the smallest class, with ratios up to 171. Class-balanced sampling was used to show the models the less frequent classes with a higher probability. The pipeline was optimized collectively across tasks during development. Afterwards, it was validated on the validation tasks. It is worth noting that this approach occasionally results in non-convergence during training and frequently performs less effectively than a training procedure that incorporates more task-specific pipeline tuning. However, we argue that the primary purpose of task fingerprinting is not to immediately produce a state-of-the-art model, but rather to refine the model development process by utilizing advantageous initialization and assisted iteration. Second, in order to validate the methodology, it is sufficient to demonstrate positive trends in a suboptimal yet sensible regime. As soon as we identified a solid configuration, we kept it for all ensuing experiments. The unmodified pipeline served as a baseline to compute gain and improvement. To compute the other meta metrics, it was necessary to train all potential knowledge transfer-induced pipeline adaptations (on all tasks) - this allowed us to rank all source tasks (percentile and weightedtau) and deduce the theoretically optimal performance (regret).

Scenario 1: Model Architecture

The first knowledge transfer scenario concerns the transfer of the neural architecture (including its pretrained weights), which has previously been suggested in literature [14, 31]. Given a candidate source task $\mathfrak{S}$ and the attached meta information $m(\mathfrak{S})$ for multiple neural architectures, $t_{\mathfrak{S}}(\mathfrak{T})$ refers to using the architecture (and initialization) with the best performance $m(\mathfrak{S})$ on the source task $\mathfrak{S}$ for training target task $\mathfrak{T}$. To generate the necessary meta information for each potential source task, we fixed a set of 20 candidate architectures, based on the following criteria: (1) availability within the PyTorch Image Models [61], (2) a reported ImageNet [2] accuracy above 75% according to PyTorch Image Models [61] while using at most 30 million parameters, which is to allow for (3) the applicability of the architecture with our compiled training pipeline within the restrictions of our hardware (max. GPU VRAM of 24GB). Finally, we reduced the list to architectures (4) published from 2019 onwards to capture the current state-of-the-art. For architectures available in different sizes, we chose the single largest variant that met our hardware requirements. The list of resulting 20 neural architectures [3, 107–123] is presented in Supplementary Table 1. Note that all neural architectures have at least one source task, they perform best on (in contrast to [14]).

Scenario 2: Pretraining Data

The second knowledge transfer scenario involves the commonly applied concept of fine-tuning. Given a source task $\mathfrak{S}$, the training of $t_{\mathfrak{S}}(\mathfrak{T})$ consists of a pretraining phase, during which a model is trained on the task $\mathfrak{S}$, followed by fine-tuning (with a newly attached final classification layer) on the target task $\mathfrak{T}$. This setup has been used in previous work [28, 34, 37].

Scenario 3: Augmentation Policy

The third knowledge transfer scenario centers around automatically learned data augmentation [124, 125]. Optimal data augmentation is an important ingredient for training state-of-the-art models [104], but remains highly task dependent [126, 127]. The automation on this part of the training pipeline offers potential benefits

especially in the low data regime, when augmentation is necessary to avoid model overfitting [127]. Nevertheless the process of automatically generating augmentation policies is highly resource intense [124] even with proposed speed improvements [125]. Therefore, the transfer and reuse of learned policies have been suggested [124]. In our experimental setup, this is modeled as follows: For each potential source task $\mathfrak{S}$, we compute a task-specific augmentation policy according to Albumentations' [62] implementation of Faster-AutoAugment [125], and attach it as meta information. $t_{\mathfrak{S}}(\mathfrak{T})$ refers to the transfer of the corresponding augmentation policy of $\mathfrak{S}$ to the training of target task $\mathfrak{T}$.

Scenario 4: Co-Training Data

The fourth knowledge transfer scenario focuses on the immediate interplay of source and target tasks during neural network training. The research around multi-task learning [128] has revealed interesting insights regarding this interplay: Optimal source tasks for pretraining might differ from optimal learning partners in multi-task learning [57]. We thus explicitly include this setup as a complementary scenario, which has so far received only moderate attention in the literature on task similarity [24, 129]. Given source and target tasks $\mathfrak{S}$ and $\mathfrak{T}$, the procedure for $t_{\mathfrak{S}}(\mathfrak{T})$ involves attaching two separate classifier heads to the shared backbone of the neural network. Training samples are drawn with equal likelihood from each of the two tasks $\mathfrak{S}$ and $\mathfrak{T}$, while only the corresponding samples are used to compute the respective losses for each of the heads. Both losses are weighted equally before backpropagation.

Generalization Scenario

To investigate the generalization of our methodology to different modalities and tasks prospectively, we applied the concept developed on 2D classification tasks to 3D segmentation tasks. To this end, we leveraged one of the most widely used datasets, namely the Medical Segmentation Decathlon challenge data [42], and official results [41]. For feature extraction we used a Masked Autoencoder (MAE) [40] pretrained by Eckstein et al. [130] on two large 3D datasets [131, 132] based on the implementation by Wald et al. [133] and a ResEncL backbone [134]. Again, we extracted 10,000 feature maps per dataset from the bottleneck layer of the network, without masking the input patch. The feature maps originally had a size of $320 \times 5 \times 5 \times 5$, however for computational efficiency, we reduced all but the first dimension via the arithmetic mean. The patches were sampled randomly across patients, available imaging modalities, and spatial locations. The challenge provides the validation values for 19 participating teams, on 17 tasks for two performance measures: Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD). When analyzing the results, however, we observed that the distribution of "best models per task" was skewed (a single participant won 29 out of the 34 rankings). Hence, applying our methodology for scenario 1, which only considers the best performance on the source task, would lead to a trivial setup, where almost all source tasks would be optimal for most target tasks. To address this issue, we computed the transferability between tasks as the Pearson correlation coefficient of the participants' results on two target tasks. Note that we excluded two participants who occasionally performed with a value of zero, indicating algorithmic failure.

Acknowledgements. The authors would like to thank Maike Rees for her help on designing Fig. 1 and Fig. 3. We would also like to thank all members of the "Intelligent Systems in Surgical Endoscopy" team for fruitful discussions. We are grateful for the linguistic corrections by Minu Dietlinde Tizabi and Marcel Knopp. This project has been funded by the German Federal Ministry of Health under the reference number 2520DAT0P1 as part of the pAItient (Protected Artificial Intelligence Innovation Environment for Patient Oriented Digital Health Solutions for developing, testing and evidence based evaluation of clinical value) project. The funder played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript. The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health". This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (project NEURAL SPICING grant agreement No. 101002198).

Declarations

Conflict of interest

All authors declare no financial or non-financial competing interests.

Data availability

All datasets used for this study are publicly available. Details can be found in Supplementary Table 2. Intermediate results such as each performance measure of knowledge transfer experiments and all pairwise task distances, are provided in our code repository.

Code availability

The source code can be found at https://github.com/IMSY-DKFZ/task-fingerprinting under the MIT license.

Author contribution

P.G. and L.M.H. conceptualized the work. P.G., A.S.,C.U., T.A., K.M.H., and L.M.H. designed the study. P.G., C.U., and A.S. performed the coding and data preparation. P.G. trained all deep learning models and conducted the statistical analysis. T.A., K.M.H., and L.M.H. provided advisory support for the project. P.G. and L.M.H. prepared the initial draft of the manuscript, with all authors participating in the refinement.

References

[1] Eisenmann, M. *et al.* Why is the winner the best? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 19955–19966 (2023).

[2] Deng, J. *et al.* Imagenet: A large-scale hierarchical image database (2009).

[3] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016).

[4] Bozinovski, S. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica (Slovenia)* **44** (2020).

[5] Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* **22**, 1345–1359 (2010).

[6] Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Wallach, H. *et al.* (eds) *Transfusion: Understanding transfer learning for medical imaging.* (eds Wallach, H. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019).

[7] Zhuang, F. *et al.* A comprehensive survey on transfer learning. *Proceedings of the IEEE* **109**, 43–76 (2021).

[8] Bommasani, R. *et al.* On the opportunities and risks of foundation models. *ArXiv* **abs/2108.07258** (2021).

[9] Achille, A. & et al. The Information Complexity of Learning Tasks, their Structure and their Distance. *ArXiv* (2019).

[10] Ben-David, S., Blitzer, J., Crammer, K. & Pereira, F. Schölkopf, B., Platt, J. & Hoffman, T. (eds) *Analysis of representations for domain adaptation.* (eds Schölkopf, B., Platt, J. & Hoffman, T.) *Advances in Neural Information Processing Systems*, Vol. 19 (MIT Press, 2006).

[11] Zhang, W., Deng, L., Zhang, L. & Wu, D. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica* **10**, 305–329 (2023).

[12] Ding, Y., Jiang, B., Yu, A., Zheng, A. & Liang, J. Which model to transfer? a survey on transferability estimation. *ArXiv* **abs/2402.15231** (2024).

[13] Nguyen, C., Do, T.-T. & Carneiro, G. Similarity of classification tasks (2021).

[14] Dias, L. V. *et al.* Imagedataset2vec: An image dataset embedding for algorithm selection. *Expert Syst. Appl.* **180**, 115053 (2021).

[15] Achille, A. *et al.* Task2vec: Task embedding for meta-learning (2019).

[16] Liu, X., Bai, Y., Lu, Y., Soltoggio, A. & Kolouri, S. Wasserstein task embedding for measuring task similarities. *Neural Networks* **181**, 106796 (2025).

[17] Molina-Moreno, M., Schilling, M. P., Reischl, M. & Mikut, R. Automated style-aware selection of annotated pre-training databases in biomedical imaging. *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)* 1–5 (2023).

[18] Hatakeyama, T., Wang, X. & Yamasaki, T. Transferability prediction among classification and regression tasks using optimal transport. *Multimedia Tools and Applications* **83**, 25105–25119 (2024).

[19] Tan, Y., Li, Y. & Huang, S.-L. *Otce: A transferability metric for cross-domain cross-task representations*, 15779–15788 (2021).

[20] Ramtoula, B., Gadd, M., Newman, P. & Martini, D. Visual dna: Representing and comparing images using distributions of neuron activations. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11113–11123 (2023).

[21] Alvarez-Melis, D. & Fusi, N. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems* **33**, 21428–21439 (2020).

[22] You, K., Liu, Y., Wang, J. & Long, M. Meila, M. & Zhang, T. (eds) *Logme: Practical assessment of pre-trained models for transfer learning.* (eds Meila, M. & Zhang, T.) *Proceedings of the 38th International Conference on Machine Learning*, Vol. 139 of *Proceedings of Machine Learning Research*, 12133–12143 (PMLR, 2021).

[23] Cheplygina, V., Moeskops, P., Veta, M., Dashtbozorg, B. & Pluim, J. P. W. Cardoso, M. J. *et al.* (eds) *Exploring the similarity of medical imaging classification problems.* (eds Cardoso, M. J. *et al.*) *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 59–66 (Springer International Publishing, Cham, 2017).

[24] Fifty, C. *et al.* Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds) *Efficiently identifying task groupings for multi-task learning.* (eds Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W.) *Advances in Neural Information Processing Systems*, Vol. 34, 27503–27516 (Curran Associates, Inc., 2021).

[25] Tran, A., Nguyen, C. V. & Hassner, T. Transferability and hardness of supervised classification tasks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 1395–1405 (2019).

[26] Peng, X., Li, Y. & Saenko, K. *Domain2vec: Domain embedding for unsupervised domain adaptation*, 756–774 (Springer, 2020).

[27] Nguyen, C., Hassner, T., Seeger, M. & Archambeau, C. *Leep: A new measure to evaluate transferability of learned representations*, 7294–7305 (PMLR, 2020).

[28] Zamir, A. R. *et al.* Taskonomy: Disentangling task transfer learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3712–3722 (2018).

[29] Pope, P., Zhu, C., Abdelkader, A., Goldblum, M. & Goldstein, T. ICLR (ed.) *The intrinsic dimension of images and its impact on learning.* (ed.ICLR) *International Conference on Learning Representations* (OpenReview.net, 2021).

[30] Wang, Z., Dai, Z., Póczos, B. & Carbonell, J. G. Characterizing and avoiding negative transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11285–11294 (2018).

[31] Godau, P. & Maier-Hein, L. de Bruijne, M. *et al.* (eds) *Task fingerprinting for meta learning inbiomedical image analysis.* (eds de Bruijne, M. *et al.*) *Medical Image Computing and Computer Assisted Intervention*

– *MICCAI 2021*, 436–446 (Springer International Publishing, 2021).

[32] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Guyon, I. *et al.* (eds) *Gans trained by a two time-scale update rule converge to a local nash equilibrium.* (eds Guyon, I. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 30 (Curran Associates, Inc., 2017).

[33] Ding, Y., Wang, L. & Gong, B. Analyzing deep neural network's transferability via fréchet distance. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* 3931–3940 (2021).

[34] Bhattacharjee, B. *et al.* P2l: Predicting transfer learning for images and semantic relations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 3284–3293 (2020).

[35] Feydy, J. *et al. Interpolating between optimal transport and mmd using sinkhorn divergences*, 2681–2690 (PMLR, 2019).

[36] Wiesenfarth, M. *et al.* Methods and open-source toolkit for analyzing and visualizing challenge results. *Scientific Reports* **11**, 2369 (2021). URL https://doi.org/10.1038/s41598-021-82017-6.

[37] Agostinelli, A., Pándy, M., Uijlings, J., Mensink, T. & Ferrari, V. Avidan, S., Brostow, G., Cissé, M., Farinella, G. M. & Hassner, T. (eds) *How stable are transferability metrics evaluations?* (eds Avidan, S., Brostow, G., Cissé, M., Farinella, G. M. & Hassner, T.) *Computer Vision – ECCV 2022*, 303–321 (Springer Nature Switzerland, Cham, 2022).

[38] Radford, A. *et al. Learning transferable visual models from natural language supervision*, 8748–8763 (PmLR, 2021).

[39] Caron, M. *et al. Emerging properties in self-supervised vision transformers*, 9650–9660 (2021).

[40] He, K. *et al. Masked autoencoders are scalable vision learners*, 16000–16009 (2022).

[41] Antonelli, M. *et al.* The medical segmentation decathlon. *Nature communications* **13**, 4128 (2022).

[42] Simpson, A. L. *et al.* A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019).

[43] Kolesnikov, A. *et al.* Big transfer (bit): General visual representation learning (2019).

[44] Mensink, T., Uijlings, J. R. R., Kuznetsova, A., Gygli, M. & Ferrari, V. Factors of influence for transfer learning across diverse appearance domains and task types. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 9298–9314 (2021).

[45] El Mestari, S. Z., Lenzini, G. & Demirci, H. Preserving data privacy in machine learning systems. *Computers & Security* **137**, 103605 (2024).

[46] Kaissis, G. A., Makowski, M. R., Rückert, D. & Braren, R. F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2**, 305–311 (2020).

[47] He, X., Zhao, K. & Chu, X. Automl: A survey of the state-of-the-art. *Knowledge-based systems* **212**, 106622 (2021).

[48] Le, C. P., Soltani, M., Dong, J. & Tarokh, V. Fisher task distance and its application in neural architecture search. *IEEE Access* **10**, 47235–47249 (2022).

[49] Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* (2020).

[50] Li, Q., Wen, Z. & He, B. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *ArXiv* **abs/1907.09693** (2021).

[51] Rieke, N. *et al.* The future of digital health with federated learning. *npj Digital Medicine* **3** (2020).

[52] Zhang, Z., Cao, X., Jia, J. & Gong, N. Z. *Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients*, 2545–2555 (2022).

[53] Cao, X., Fang, M., Liu, J. & Gong, N. *Fltrust: Byzantine-robust federated learning via trust bootstrapping* (2021).

[54] Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. & Wermter, S. Continual lifelong learning with neural networks: A review. *Neural Networks* **113**, 54–71 (2019).

[55] Soltoggio, A. *et al.* A collective ai via lifelong learning and sharing at the edge. *Nat. Mac. Intell.* **6**, 251–264 (2024).

[56] Chaves, L., Bissoto, A., Valle, E. & Avila, S. Koch, L. *et al.* (eds) *The performance of transferability metrics does not translate to medical tasks.* (eds Koch, L. *et al.*) *Domain Adaptation and Representation Transfer*, 105–114 (Springer Nature Switzerland, Cham, 2024).

[57] Standley, T. *et al.* III, H. D. & Singh, A. (eds) *Which tasks should be learned together in multi-task learning?* (eds III, H. D. & Singh, A.) *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119 of *Proceedings of Machine Learning Research*, 9120–9132 (PMLR, 2020).

[58] Reinke, A. *et al.* Understanding metric-related pitfalls in image analysis validation. *Nature methods* **21**, 182–194 (2024).

[59] Azad, B. *et al.* Foundational models in medical imaging: A comprehensive survey and future vision. *ArXiv* **abs/2310.18689** (2023).

[60] Scherer, J. *et al.* Die joint imaging platform (jip) des deutschen konsortiums für translationale krebsforschung (dktk) (2020).

[61] Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models (2019).

[62] Buslaev, A. *et al.* Albumentations: Fast and flexible image augmentations. *Information* **11** (2020). URL https://www.mdpi.com/2078-2489/11/2/125.

[63] Kullback, S. & Leibler, R. A. On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86 (1951).

[64] Sugiyama, M. *et al.* Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* **60**, 699–746 (2008).

[65] Sun, B. & Saenko, K. *Deep coral: Correlation alignment for deep domain adaptation*, 443–450 (Springer, 2016).

[66] Maier-Hein, L. *et al.* Metrics reloaded: recommendations for image analysis validation. *Nature methods* (2022).

[67] Renggli, C. *et al.* IEEE (ed.) *Which model to transfer? finding the needle in the growing haystack.* (ed.IEEE) *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9195–9204 (2022).

[68] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[69] Pogorelov, K. *et al.* Nerthus: A bowel preparation quality video dataset. *Proceedings of the 8th ACM on Multimedia Systems Conference* 170–174 (2017).

[70] Borgli, H. *et al.* Hyper-kvasir: A comprehensive multi-class image and video dataset for gastrointestinal endoscopy (2019). URL https://osf.io/mh9sj/.

[71] Leibetseder, A. *et al.* Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology (2018).

[72] Twinanda, A. P. *et al.* Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging* **36**, 86–97 (2017).

[73] Aditya Khosla, B. Y., Nityananda Jayadevaprakash & Fei-Fei, L. Novel dataset for fine-grained image categorization (2011).

[74] Netzer, Y. *et al.* Reading digits in natural images with unsupervised feature learning (2011). URL http://ufldl.stanford.edu/housenumbers.

[75] Fei-Fei, L., Fergus, R. & Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *2004 Conference on Computer Vision and Pattern Recognition Workshop* 178–178 (2004).

[76] Griffin, G., Holub, A. & Perona, P. Caltech-256 object category dataset (2007). URL https://doi.org/10.22002/D1.20087.

[77] Krizhevsky, A. Learning multiple layers of features from tiny images (2009). URL https://www.cs.toronto.edu/~kriz/cifar.html.

[78] Faria, S. M. M. *et al.* Light field image dataset of skin lesions. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 3905–3908 (2019).

[79] Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics* **23**, 538–546 (2019).

[80] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).

[81] Cohen, G., Afshar, S., Tapson, J. C. & van Schaik, A. Emnist: Extending mnist to handwritten letters. *2017 International Joint Conference on Neural Networks (IJCNN)* 2921–2926 (2017).

[82] Moccia, S., Vanone, G. O., Momi, E. D. & Mattos, L. S. Nbi-inframes (2018). URL https://doi.org/10.5281/zenodo.1162784.

[83] Moccia, S., Momi, E. D. & Mattos, L. S. Laryngeal dataset (2017). URL https://doi.org/10.5281/zenodo.1003200.

[84] Irvin, J. A. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison (2019).

[85] Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131.e9 (2018).

[86] Jaeger, S. *et al.* Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery* **4**, 475 (2014).

[87] Hemateja, A. V. N. M. Covid19 x-ray classification dataset on kaggle. https://www.kaggle.com/ahemateja19bec1025/covid-xray-dataset (2021). Accessed: 2022-01-13.

[88] Rajpurkar, P. *et al.* MURA dataset: Towards radiologist-level abnormality detection in musculoskeletal radiographs (2018).

[89] Lab, M. A. Bean disease dataset (2020). URL https://github.com/AI-Lab-Makerere/ibean/.

[90] Ghamsarian, N., Amirpourazarian, H., Timmerer, C., Taschwer, M. & Schöffmann, K. Relevance-based compression of cataract surgery videos using convolutional neural networks (2020).

[91] Analysis of images to detect abnormalities in endoscopy (aida-e) challenge. https://aidasub-clebarrett.grand-challenge.org/home/.

[92] Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data in Brief* **28**, 104863 (2020).

[93] Cataract dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset.

[94] Bohaju, J. Brain tumor (2020). URL https://doi.org/10.34740/KAGGLE/DSV/1370629.

[95] Cheng, J. *et al.* Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one* **10**, e0140381 (2015).

[96] Yang, X., Zhao, J., Zhang, Y., He, X. & Xie, P. Covid-ct-dataset: A ct scan dataset about covid-19. *ArXiv* (2020).

[97] Giotis, I. *et al.* Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications* **42** (2015).

[98] Mendonça, T. *et al. PH2: A Public Database for the Analysis of Dermoscopic Images*, 419–439 (2015).

[99] Collaboration, I. S. I. Siim-isic 2020 challenge dataset (2020). URL https://doi.org/10.34970/2020-ds01.

[100] Rotemberg, V. *et al.* A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data* **8** (2021).

[101] Liu, R. *et al.* Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns* 100512 (2022).

[102] Aptos 2019 blindness detection challenge. https://www.kaggle.com/c/aptos2019-blindness-detection/data.

[103] Smedsrud, P. H. *et al.* Kvasir-Capsule, a video capsule endoscopy dataset. *Scientific Data* **8**, 142 (2021).

[104] Wightman, R., Touvron, H. & J'egou, H. Resnet strikes back: An improved training procedure in timm. *ArXiv* **abs/2110.00476** (2021).

[105] Smith, L. N. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 464–472 (2017).

[106] Falcon, W. & The PyTorch Lightning team. PyTorch Lightning (2019). URL https://github.com/Lightning-AI/lightning.

[107] Xie, Q., Hovy, E. H., Luong, M.-T. & Le, Q. V. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10684–10695 (2020).

[108] Zhang, H. *et al.* Resnest: Split-attention networks. *ArXiv* **abs/2004.08955** (2020).

[109] Wang, Q. *et al.* Eca-net: Efficient channel attention for deep convolutional neural networks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 11531–11539 (2020).

[110] Yalniz, I. Z., Jégou, H., Chen, K., Paluri, M. & Mahajan, D. K. Billion-scale semi-supervised learning for image classification. *ArXiv* **abs/1905.00546** (2019).

[111] Xie, C. *et al.* Adversarial examples improve image recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 816–825 (2020).

[112] Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv* **abs/1905.11946** (2019).

[113] Bochkovskiy, A., Wang, C.-Y. & Liao, H.-Y. M. Yolov4: Optimal speed and accuracy of object detection. *ArXiv* **abs/2004.10934** (2020).

[114] Wang, C.-Y. *et al.* Cspnet: A new backbone that can enhance learning capability of cnn. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1571–1580 (2020).

[115] Lee, Y., won Hwang, J., Lee, S., Bae, Y. & Park, J. An energy and gpu-computation efficient backbone network for real-time object detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 752–760 (2019).

[116] Tan, M. & Le, Q. V. Mixconv: Mixed depthwise convolutional kernels. *ArXiv* **abs/1907.09595** (2019).

[117] Radosavovic, I., Kosaraju, R. P., Girshick, R. B., He, K. & Dollár, P. Designing network design spaces. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10425–10433 (2020).

[118] Gao, S. *et al.* Res2net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 652–662 (2021).

[119] Han, D., Yun, S., Heo, B. & Yoo, Y. J. Rethinking channel dimensions for efficient model design. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 732–741 (2021).

[120] Yang, B., Bender, G., Le, Q. V. & Ngiam, J. Wallach, H. *et al.* (eds) *Condconv: Conditionally parameterized convolutions for efficient inference.* (eds Wallach, H. *et al.*) *Advances in Neural Information Processing Systems*, Vol. 32 (Curran Associates, Inc., 2019).

[121] Li, X., Wang, W., Hu, X. & Yang, J. Selective kernel networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 510–519 (2019).

[122] Sun, K. *et al.* High-resolution representations for labeling pixels and regions. *ArXiv* **abs/1904.04514** (2019).

[123] Howard, A. G. *et al.* Searching for mobilenetv3. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 1314–1324 (2019).

[124] Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V. & Le, Q. V. Autoaugment: Learning augmentation policies from data. *ArXiv* **abs/1805.09501** (2018).

[125] Hataya, R., Zdenek, J., Yoshizoe, K. & Nakayama, H. Faster autoaugment: Learning augmentation strategies using backpropagation. *ArXiv* **abs/1911.06987** (2020).

[126] Yang, S. *et al.* Image data augmentation for deep learning: A survey. *ArXiv* **abs/2204.08610** (2022).

[127] Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6**, 1–48 (2019).

[128] Crawshaw, M. Multi-task learning with deep neural networks: A survey. *ArXiv* **abs/2009.09796** (2020).

[129] Zhou, F. *et al.* Task similarity estimation through adversarial multitask neural network. *IEEE Transactions on Neural Networks and Learning Systems* **32**, 466–480 (2021).

[130] Eckstein, K. *et al.* *The missing piece: A case for pre-training in 3d medical object detection*, 615–626 (Springer, 2025).

[131] Hamamci, I. E. *et al.* Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834* (2024).

[132] Casey, B. J. *et al.* The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience* **32**, 43–54 (2018).

[133] Wald, T. *et al.* An openmind for 3d medical vision self-supervised learning. *arXiv preprint arXiv:2412.17041* (2024).

[134] Isensee, F. *et al.* *nnu-net revisited: A call for rigorous validation in 3d medical image segmentation*, 488–498 (Springer, 2024).

Supplementary information

We provide details and references for the used neural architectures (see Supplementary Table 1) and datasets (see Supplementary Table 2). Supplementary Table 3 displays the weightedtau of measured transferability outcomes between knowledge transfer scenarios. Supplementary Table 4 presents the win rates across compared task selectors and investigated knowledge transfer scenarios. In addition we provide the complementary approach of "rank then mean" to Fig. 5 of the main paper in Supplementary Fig. 1.

Table 1 Overview on Neural Architectures used for model architecture transfer experiments. ImageNet Accuracy has been provided by the timm library [61]. Train parameters only comprise the shared backbone.

Architecture	**timm reference**	**ImageNet Acc. (%)**	**train params (mio.)**
EfficientNet B2 noisy student [107]	tf_efficientnet_b2_ns	82.38	7.7
ResNet50 SWSL [110]	swsl_resnet50	81.166	23.5
ResNeSt50 [108]	resnest50d	80.974	25.4
ECA ResNet50 [109]	ecaresnet50d	80.592	23.5
ResNeXt50 SSL [110]	ssl_resnext50_32x4d	80.318	23.0
EfficientNet B2 AdvProp [111]	tf_efficientnet_b2_ap	80.3	7.7
EfficientNet B2 [112]	tf_efficientnet_b2	80.086	7.7
CSP DarkNet53 [113]	cspdarknet53	80.058	26.6
CSPResNeXt50 [114]	cspresnext50	80.04	18.5
CSPResNet50 [114]	cspresnet50	79.574	20.6
VoVNet [115]	ese_vovnet39b	79.32	23.5
MixNet-L [116]	mixnet_l	78.976	5.8
RegNetY [117]	regnety_032	78.886	17.9
RegNetX [117]	regnetx_032	78.172	14.3
Res2Net50 [118]	res2net50_26w_4s	77.964	23.7
RexNet100 [119]	rexnet_100	77.858	3.5
EfficientNet B0 CondCov [120]	tf_efficientnet_cc_b0_4e	77.306	12.0
SK ResNet34 [121]	skresnet34	76.912	21.8
HRNetV2-W18 [122]	hrnet_w18	76.758	19.3
MobileNetV3-Large [123]	mobilenetv3_large_100	75.766	4.2
ResNet34 [3]	resnet34	75.11	21.3

Table 2 Overview on datasets and tasks used within out task pool. More properties can be found in Fig. 3.

Dataset	Tasks	Reference(s)
Nerthus	T01	[69]
HyperKvasir	T02-T05	[70]
LapGyn4	T06-T09	[71, 72]
Cholec80	T10-T16	[72]
Stanford Dogs	T17	[2, 73]
SVHN	T18	[74]
Caltech101	T19	[75]
Caltech256	T20	[76]
CIFAR	T21, T22	[77]
SKLIN2	T23	[78]
Derm7pt	T24	[79]
MNIST	T25	[80]
EMNIST	T26	[81]
NBI-InfFrames	T27	[82]
Laryngeal Cancerous Tissues	T28	[83]
CheXpert	T29-T41	[84]
Zhang Chest X-Ray Images	T42	[85]
Shenzen Hospital CXR Set	T43	[86]
Kaggle COVID X-Ray Dataset	T44	[87]
MURA	T45-T51	[88]
ibean	T52	[89]
CatRelComp	T53	[90]
AIDA-E Barrett's Esophagus Diagnosis	T54	[91]
Dataset of Breast Ultrasound Images	T55	[92]
Kaggle Cataract Dataset	T56	[93]
Kaggle Brain Tumor Classification	T57	[94]
Brain Tumor Type Classification	T58	[95]
COVID-CT-Dataset	T59	[96]
MED-NODE	T60	[97]
PH2	T61	[98]
ISIC20	T62	[99, 100]
DeepDRiD	T63-T67	[101]
APTOS19	T68	[102]
Kvasir-Capsule	T69-T71	[103]

Table 3 Mean and standard deviation of weightedtau on pairwise transfer scenario outcomes over 43 tasks, 3 repetitions and 3 performance metrics for the four knowledge transfer scenarios we investigate: Model Architecture (M. A.), Pretraining Data (P. D.), Augmentation Policy (A. P.) and Co-Training Data (C. D.).

	Model Architecture	**Pretraining Data**	**Augmentation Policy**	**Co-Training Data**
M. A.	1.000 ± 0.000	0.084 ± 0.147	0.078 ± 0.141	0.052 ± 0.137
P. D.	0.084 ± 0.147	1.000 ± 0.000	0.037 ± 0.164	0.038 ± 0.136
A. P.	0.078 ± 0.141	0.037 ± 0.164	1.000 ± 0.000	-0.002 ± 0.140
C. D.	0.052 ± 0.137	0.038 ± 0.136	-0.002 ± 0.140	1.000 ± 0.000

Table 4 Win rates [37] for task selectors described in Fig. 5 in percent. Shows the fraction of 1032 individual cases (43 tasks, 3 repetitions, 4 meta metrics, 2 base metrics) that a specific task selector performs best. Columns may sum above 100 because of ties, to reduce such occurrences we averaged the meta metrics of top 3 suggestions. Next to the win rates for four knwoledge transfer scenarios we also provide the average across them.

	M. A.	**P. D.**	**A. P.**	**C. D.**	**mean**
FID	15.02	13.86	12.60	10.66	13.03
P2L	15.21	14.92	13.66	10.17	13.49
FED	15.89	15.21	11.14	12.79	13.76
VDNA	13.66	18.41	16.09	12.21	15.09
Manual	20.64	19.86	11.34	16.96	17.20
bKLD(small,target)	14.53	**25.58**	23.84	17.25	20.30
bKLD(large,unweighted)	14.53	25.39	22.19	20.64	20.69
bKLD(large,source)	**34.98**	19.86	**27.71**	**32.66**	**28.80**

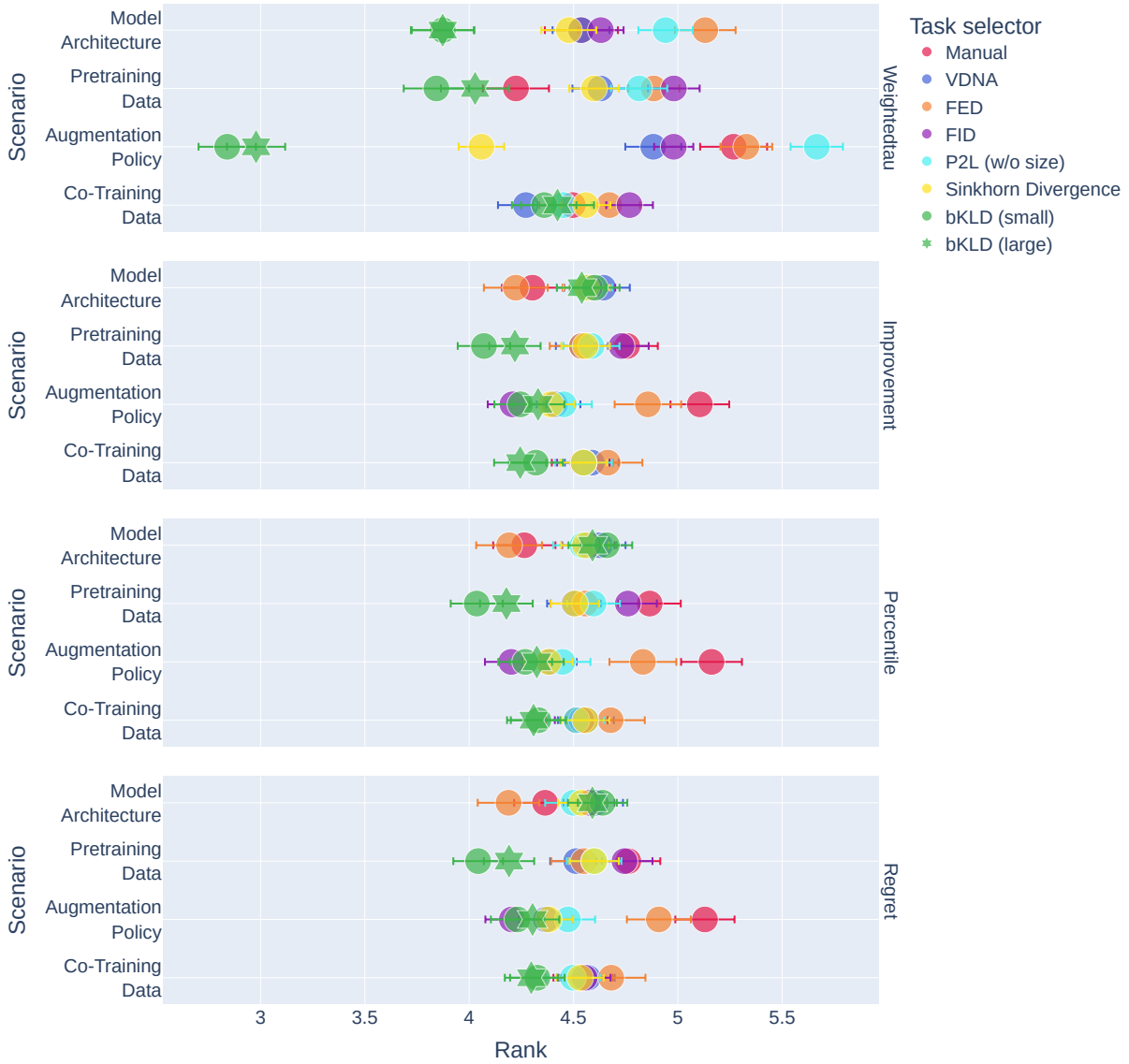

Fig. 1 binned Kullback-Leibler Divergence (bKLD) outperforms previously proposed methods for knowledge transfer. In contrast to the evaluation shown in Fig. 5, this figure shows the evaluation according to "rank then mean" [36], using 258 setups (2 base metrics, 43 validation tasks, 3 repetitions). The marker position refers to the mean over 1000 bootstraps, with the whiskers indicating standard deviation. For each setup, the improvement, percentile, and regret of the top three suggestions are averaged, while weightedtau is evaluated on the full ranking of suggested knowledge transfer sources.