SYSTEMS BIOLOGY

Domain adaptation in small-scale and heterogeneous biological datasets

Seyedmehdi Orouji[1], Martin C. Liu[2,3], Tal Korem[3,4,5]*†, Megan A. K. Peters[1,5,6]*†

Machine-learning models are key to modern biology, yet models trained on one dataset are often not generalizable to other datasets from different cohorts or laboratories due to both technical and biological differences. Domain adaptation, a type of transfer learning, alleviates this problem by aligning different datasets so that models can be applied across them. However, most state-of-the-art domain adaptation methods were designed for large-scale data such as images, whereas biological datasets are smaller and have more features, and these are also complex and heterogeneous. This Review discusses domain adaptation methods in the context of such biological data to inform biologists and guide future domain adaptation research. We describe the benefits and challenges of domain adaptation in biological research and critically explore some of its objectives, strengths, and weaknesses. We argue for the incorporation of domain adaptation techniques to the computational biologist's toolkit, with further development of customized approaches.

INTRODUCTION

In the computational biological sciences, we are interested in learning informative "truths" about biological systems through machine learning or similar quantitative modeling techniques (*1*). Contrary to "irrelevant" or "purely statistical" correlations, which find statistical idiosyncracies in data that do not reflect scientifically meaningful underlying patterns (e.g., when detecting COVID-19 from chest radiographs, a model may rely on confounding factors such as laterality markers or patient positioning, thus failing to generalize to new patients from other hospitals (*2*) and leading to misinterpretation of results within a single dataset), we expect such "truths" to generalize beyond a specific dataset or population, indicating that they offer a grounded biological meaning. However, collecting (and sometimes labeling) biological datasets is difficult, expensive, and time consuming, leading to many small but related datasets that are collected from different sources and under different environmental and experimental conditions (e.g., different laboratories, equipment, settings, humidity, etc.). For example, in the widely used Autism Brain Imaging Dataset (ABIDE), functional magnetic resonance imaging (fMRI) data were collected at multiple sites, which hindered the ability to directly aggregate data (*3*). Beyond creating challenges in data curation and metadata standards (*4*, *5*), this variability in the sources of small biological datasets creates different domains of data that have different statistical distributions.

While this variety is a strength that can facilitate discovery of generalizable truths, it also presents a major challenge to computational biology: Applying knowledge gained from one dataset (a source) to another (a target) will fail if the two datasets have highly divergent distributions—a phenomenon known as domain shift or data bias (*6*, *7*). In short, we cannot blindly apply a model (of any kind) trained on a source dataset collected under one set of conditions to new target data and expect it to perform effectively. In an age of open datasets and keen interest in adhering to FAIR principles (findability, accessibility, interoperability, and reuse of digital assets) to accelerate scientific discovery, it is increasingly urgent that we acknowledge the strengths and challenges of combining datasets.

To best extract generalizable insights while making use of all collected data from varying sources—especially in biological disciplines where data are expensive—and to apply these insights to newly collected data, we must find how to best leverage the use of all existing and continuously growing small biological datasets (*8*). Here, computational biologists can borrow insights from machine learning to leverage transfer learning, which aims to use knowledge gained from learning a task on one dataset to perform a similar task on a different but related dataset, thereby transferring knowledge across datasets (*9*–*13*). More precisely, domain adaptation (DA), a subfield of transfer learning, has been developed to address this issue of different statistical distributions by aligning the distributions of the source and target domains (Fig. 1). Of note, while there are some similarities to "batch correction" often applied in high-throughput molecular measurements (*14*, *15*), the objective is different: DA aims to learn generalizable models across domains, while batch correction is primarily aimed at removing technical variation.

DA is more than just "lining up the features" and training a model on both datasets; not only is this often impossible to do (especially if features are unlabeled), but statistical differences between the domains can often guarantee that such a brute force aggregation is doomed to failure. Instead, through DA, a model is forced to learn domain invariant features, i.e., features that are common across all domains, such that the learned model can be generalized and thus perform relatively well on a separate target domain. Another benefit of DA is that the integration of multiple datasets effectively increases the sample size, allowing for improved inference of statistical signals. This allows better use of available data and resources, reducing the need to collect and annotate expensive data (*16*–*18*). Thus, in sum, it seems clear that applying DA to biological data can potentially mitigate small sample sizes within individually collected datasets, and through transferring knowledge to other domains can ideally find generalizable truths (Table 1).

[1]Department of Cognitive Sciences, University of California Irvine, Irvine, CA, USA. [2]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA. [3]Program for Mathematical Genomics, Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA. [4]Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY, USA. [5]CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, Canada. [6]CIFAR Fellow, Program in Brain, Mind, & Consciousness, CIFAR, Toronto, Canada.
*Corresponding author. Email: tal.korem@columbia.edu (T.K.); megan.peters@uci.edu (M.A.K.P.)
†These authors contributed equally to this work.

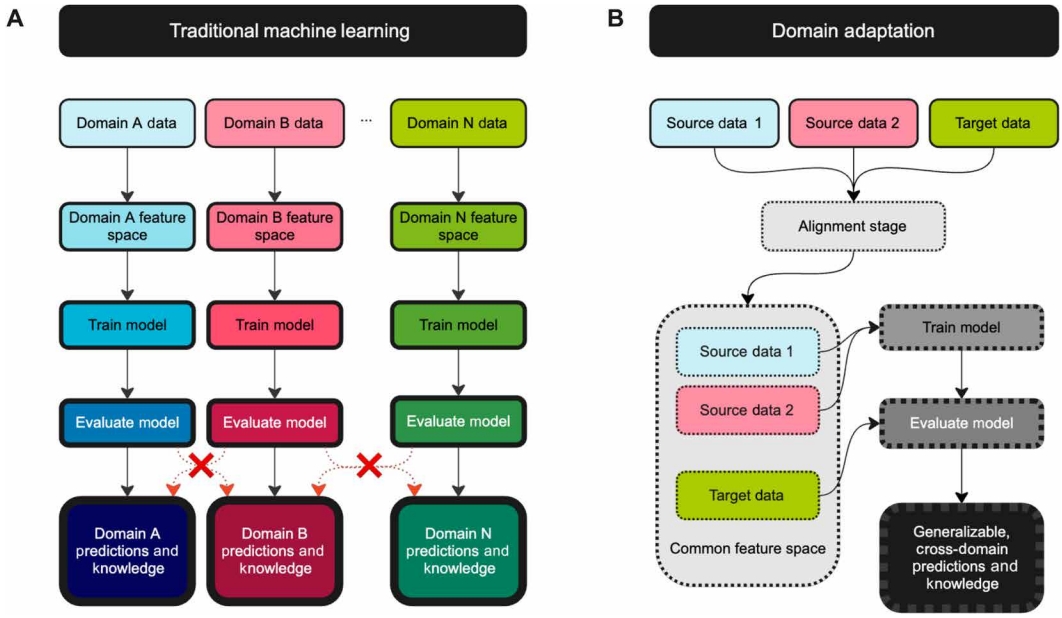

Fig. 1. Diagrammatic overview of the machine learning pipeline and modifications needed to engage in transfer learning or domain adaptation (DA). (**A**) In traditional machine learning, each domain has its own model, trained on domain-specific features. This means that the model can make predictions about data from that domain, but transferring the model to apply it to other domains is typically difficult or even impossible (indicated by red Xs). (**B**) In transfer learning or DA, data from one or more source domains are aligned (denoted by dashed outlines) with those in the target domain to find common feature spaces with similar statistical distributions such that a single model can be trained on aggregate source domain data and evaluated on target domain. This process can produce generalizable knowledge that is not domain specific. Of note, in some cases, target data will only be used after the model has been trained and not in the alignment stage (*152*).

Table 1. Specific benefits of using DA with biological datasets.

Problem	DA benefit	Example
	Mitigate poor sample-to-feature ratios	
Complex biological systems often need to be modeled with many free parameters, while training samples remain quite few.	Integrate individual datasets to increase the number of training samples, providing a larger and more diverse dataset and preventing overfitting.	Combining fMRI datasets across individuals or scanning sites (*139*).
	Transfer knowledge	
Some domains are poor in data due to either small sample size or missing labels.	Transfer knowledge from existing rich datasets to related, smaller datasets.	Transferring insights gained from MRI in adults to newborns (*155*); annotations from preclinical cell lines to more data-scarce clinical settings (*136*).
	Find generalizable patterns	
In the age of big data sharing through FAIR principles, biological datasets are often composed of many different small cohorts collected from different laboratories and under different environmental and experimental conditions (*19*, *53*). These many smaller datasets drive models to finding patterns that turn out to be statistical anomalies or idiosyncracies unique to each dataset*.	Combining as much data as possible while minimizing statistical differences between domains can minimize the risk of models finding statistical idiosyncrasies rather than patterns shared across domains.	Finding cohort-independent generalities across multiple studies of the vaginal microbiome in preterm birth (*133*, *134*) or the gut microbiome in colorectal cancer (*156*, *157*) despite the variability in microbiome profiling (*154*).

*The statistical anomalies and idiosyncracies noted here are also related to two other related concepts: batch effects (*153*), when nonbiological artifacts change the distribution of the data for a subset of an experiment (e.g., plates for DNA extraction in microbiome cohorts (*154*); days or machines for MRI data collection), and batch confounding, when batches are associated with the outcome of interest.

However, DA is not a panacea, and computational biologists should be aware of the particular challenges of using such methods to analyze biological datasets. Compared to datasets typically used to train machine learning models (*19*–*22*), many "biological-scale" datasets are smaller in sample size, have many more features than samples, and have a complicated feature space (e.g., different numbers of features in each dataset, missing values, heterogeneous features, unique feature importance distributions, etc.). Therefore, while developing effective DA techniques that can work well with these small "biological-scale" datasets to find general truths about biological systems is highly desirable, it presents a specific set of challenges to machine learning research.

In this Review, we aim to critically discuss the benefits and challenges of applying current DA methodologies and frameworks to such biological datasets. To this end, we use the token examples of fMRI and microbiome datasets, two seemingly different disciplines in biology, to show the common considerations critical to developing effective DA techniques in such data. Our goal is to lay out the key components that require consideration in selecting an effective DA technique and highlight important areas of future methodological research in DA methods that can be maximally effective in biological datasets—especially as data sharing and metadata curation continues to mature. Our hope is that this discussion and synthesis will be of value both to biologists seeking to apply DA to their own data and to machine learning researchers driving state-of-the-art advances in DA methodology.

CHALLENGES OF DOMAIN ADAPTATION IN BIO-SCALE DATA AND A PATH FORWARD

As briefly introduced above, successful application of DA to small datasets with complex features comes with substantial challenges—many of which stem from the very reasons we would want to use it in the first place. We next explore several of the most pressing limitations in greater detail, both to help researchers learn to evaluate DA approaches for appropriateness in their own research and to highlight deficiencies in current DA applications to biological questions, which may be alleviated through improved collaboration between DA researchers and computational biologists. We introduce and expand on several challenges below and summarize the challenges covered here in Table 2.

Number of samples and features

Most DA methods have been designed in the fields of computer vision, text mining, or language processing (*23*–*26*) with reference to—and evaluation on—large-scale text and image data, where there can be tens of thousands (or even millions) of samples available for training [e.g., MNIST, CIFAR10; (*27*–*29*)]. In contrast, the number of samples in biological datasets is often small, but they simultaneously have many features, a problem known as the curse of dimensionality (*30*). For instance, in a typical fMRI or microbiome dataset, we might only have a few dozens to hundreds of samples, while the number of features could exceed thousands (*31*–*33*). This imbalance between the number of samples and features can potentially lead to overfitting problems (*34*, *35*) or cases where the model performs well on training data but fails to make accurate predictions or conclusions from any other data; this, of course, hinders the effectiveness of DA techniques on biological datasets (*30*). There do exist several datasets typically used to benchmark DA approaches that may be somewhat closer in size to biological-scale data, including Office31 (*36*), which contains image data of objects collected from three source domains with different resolutions, for a total of 4110 images from 31 object categories (132 images per category). However, while one might hope that DA methods that have shown success on Office31 (*37*–*39*) could be useful for biological data with similar sample size per category, it must be acknowledged that many biological datasets have very different properties than imaging data (*40*–*43*), and are even smaller, with only several hundred training samples in total. There is a need for DA algorithm development to specifically target success in the face of fewer training samples.

Differences in feature complexity

Simply checking that DA approaches can perform adequately on small datasets is unfortunately unlikely to be enough. Another barrier to applying DA approaches to biological data is that features in biological domains are inherently much more complex than those in image data. For example, in many machine learning datasets such as MNIST or Office-31, image data are essentially pixel luminance values in the RGB and alpha channels that can be relatively simple to aggregate with other source data, for example by resizing the image (*6*, *44*–*47*). In the case of biological datasets, however, the inherent complexity of features can substantially hinder our ability to aggregate different sources of data. For example, biological datasets often contain missing values (*48*–*51*) or have different numbers of features with unknown mapping between domains (*52*) (i.e., which features in a source are "the same" as which features in a target domain). They can also exhibit nonlinear relationships or interactions between features (*51*, *53*–*55*), and unique data preprocessing requirements for each source can substantially increase the complexity of developing DA techniques for biological datasets. In other words, in addition to feature-to-sample ratio and number of categories, we need to take into account the complexity and heterogeneity of biological domains before using DA techniques on biological datasets. This increased complexity stems from several sources, which we next discuss in more detail.

Missing values

Biological samples often contain many missing feature values. For example, microbiome data typically only consists of only a few taxa that are shared by most samples and even less so across cohorts. Many taxa are rare, a phenomenon known as zero inflation in statistics (*56*). In human neuroimaging, positron emission tomography or MRI scans combined with patients' genetic information can help with early diagnosis of Alzheimer's disease. However, the very common problem of missing values (i.e., not every subject has completed multimodality data) can impede the ability of these multimodal models to make reliable predictions (*57*–*59*). Missing data are less problematic in many traditional datasets used to train DA approaches, meaning that these approaches may not deal with missing data well; to be successful with biological data, DA algorithms need to adequately handle both small data and missing values.

Heterogeneity of features

Biological domains also often have different numbers of features, and the features also often do not lie in the same rank order across domains. For example, fMRI data from a given brain region will have different numbers of voxels from one human subject to the next, and the information represented, for example, in voxel 1 in person A is unlikely to functionally align with the information encoded by voxel 1 in person B. While functional alignment approaches have been developed (*52*, *60*), they do not explicitly perform DA operations. In microbiome research, it can be unclear whether particular taxa are the same across datasets, especially because, sometimes, the measurement techniques differ (e.g., taxa are characterized using different regions of a marker gene such that the same taxa might be represented by different features in different datasets). These examples are in stark contrast to most image-based DA approaches, which can exploit physical proximity of features (pixels) through spatial convolution or learn feature importance maps based on spatial features alone (e.g., the center of an image may often be more informative than the edges).

Table 2. Challenges of DA that are specific to biological datasets.

Challenge		Description
Poor sample-to-feature ratios		State-of-the-art DA approaches often require tens of thousands (or even millions) of samples to train [e.g. (*27*–*29*)], but biological datasets have a few dozens or hundreds of samples despite having thousands of features (*31*–*33*). DA models should be evaluated on biological-scale datasets [e.g., Office31 (*36*)].
Complex features	Missing values	Traditional DA models are not as often evaluated on data with many missing feature values, which is common in biological data [e.g., rare taxa in microbiome research (*56*–*59*)]. DA models should be evaluated in the context of missing data.
	Heterogeneous features	Biological data often have different features in different domains, only some of which are shared across domains (*61*). Traditional DA approaches often assume that features are shared and alignable (e.g., pixels with Cartesian coordinates), and in the face of poor labeling (*62*) or unavailable information about which features are shared (*12*) may simply deal with nonshared and non-alignable features by removing them from the DA model's inputs. DA for biology should focus on targeting feature alignment in the context of some shared and some unique features across domains.
	Feature importance distributions	In traditional machine learning benchmark datasets, many features can be similarly important to the performance of a model (*40*–*43*), but in biological datasets, sometimes only a few features are very important. Thus, feature importance distributions are quite different in biological data than in DA benchmark datasets. DA models should be evaluated on datasets with varying feature importance distributions, ideally matching those found in biological data.
Data collection and preprocessing contributions		Biological data must be extensively preprocessed after primary data collection. Choices about which preprocessing steps to take, and which software packages to use, can meaningfully alter statistical distributions of feature behavior. Machine learning models thus can easily fall prey to preprocessing-induced statistical idiosyncracies (*75*), even when standardization efforts are made (*76*). DA approaches should be evaluated on their robustness to preprocessing choices for biological data.
Feature interpretability		Biological dataset features can be difficult to interpret: They are not simply pixel luminances at specific image coordinates, for example. Especially in the case of latent features found through DA (*83*–*86*) or simply dimensionality reduction techniques, care should be taken to integrate DA approaches with emphasis on interpretability to maximize their utility for biology.
Theoretical limitations		DA can only be successful if the source and target domains are *adaptable*—i.e., theoretically joinable (*89*–*92*). Adaptability (*89*, *91*) is highly understudied in biology, and failures of adaptability can lead to negative transfer, or cases where DA causes more harm than benefit (*89*, *90*, *93*). Methods development and empirical study are crucial to understanding theoretical limits on adaptability in biological domains.

In addition, domains may have some overlapping features but also some nonshared (distinct) features—i.e., those that are specific to one domain but not the other (*61*). Current DA techniques may not be very effective on such datasets since domains may lack supplementary information such as labels (*62*) or information about matching features or samples between datasets (*12*). This limitation could force researchers to remove domain-specific features and hence lose the capacity of DA models to benefit from these unique features in the learning process. Ideally, DA for biology could benefit from a specific focus on both feature alignment (ideally unlabeled) and principled ways to deal with shared versus nonshared features.

Distribution of feature importance

In biological datasets, feature importance distributions can be more highly skewed than in many standard benchmarks used to test DA approaches. That is, in biology, a few features can be very important for the ultimate performance of a model; in contrast, in typical benchmark datasets, many features can have similar importance (*40*–*43*). This difference in skewness of feature importance distributions can lead to extreme challenges for many DA approaches such that DA models that succeed even on small "typical" benchmark datasets may fail in biological applications.

Contributions of data collection and preprocessing procedures

Biological datasets often require extensive preprocessing after the data collection stage, which can be inconsistent across datasets or laboratories. Preprocessing can refer to either specific steps that may be used or not to clean, align, or otherwise modify raw data, or to the specific software packages used to accomplish logically similar goals [e.g., DADA2 or deblur for 16*S* ribosomal RNA (rRNA) amplicon data (*63*, *64*), fMRIPrep (*65*) versus AFNI (*66*, *67*) or FSL (*68*–*70*) for fMRI images (*71*)]. These choices can be made because of individual laboratories' conventions or because of development of new software or algorithm versions that challenge reproducibility even within a given dataset. For example, in MRI data, it has been clearly demonstrated that software selection at multiple stages of processing, atlas selection (e.g., Desikan-Killiany-Tourville versus Destrieux versus Glasser), and idiosyncratic quality control procedural choices can strongly affect group- and individual-based inferences (*72*); in extreme cases, some later preprocessing stages can reintroduce nuisance covariates that were originally filtered out. Hence, choices regarding the stage of processing at which to perform DA could strongly affect the success of DA approaches (*73*). Such effects can also be exacerbated when full pipeline details are not included in publications to aid in reproducibility or when specific preprocessing steps are applied without appropriate attention to how they may alter statistical features specific to biological data [see (*74*) for a discussion relevant to microbiome data].

As a result, machine learning methods used in biology are typically limited to being highly context- and preprocessing-specific, requiring careful design and tailoring to test the desired hypothesis appropriately (*75*). This often occurs even despite targeted efforts in bridging this gap by the means of setting up standards in generating and preprocessing the data (*76*), since some laboratory- and individual-specific idiosyncrasies are wholly unavoidable. For example, in fMRI data correction for subject's head movement, using different scanning sequences or scanners can introduce data shifts that make applying DA techniques even more difficult (*3*, *77*–*81*). Even preprocessing methods meant specifically to correct for batch effects in microbiome research can introduce pipeline-specific factors (*82*), which may be easily overlooked. Such preprocessing idiosyncracies can thus exacerbate or interact with other batch effects, including introducing or altering interdependencies among features (*53*).

Interpretability of features and feature spaces

Interpretability is an important aspect of biological research, in contrast to at least some other machine learning applications. However, alignment steps in DA, which often require finding a latent representation of data by projecting the domains into a shared feature space (*83*–*85*), are frequently carried out by machine learning and deep learning methods. This means that DA in biological data inherits the same problem that plagues machine learning more broadly: failures in interpretability due to the black-box nature of these machine learning and deep learning methods. The shared feature space is particularly challenging to interpret (*86*) because it is defined as a latent space that bridges two or more domains rather than the latent space defined by one domain alone. Therefore, DA research can and should aim particularly at understanding how input features are related to the common feature space when using these methods (*87*, *88*).

Theoretical limitations of DA

It is also important that we discuss a critical theoretical limitation of DA, especially as it might affect biological data. The primary driver of DA's potential success is the adaptability between the source and target domains (*89*, *90*)—essentially, the theoretically maximal ability of an ideal model to jointly model them (*91*, *92*). Failure of adaptability is thus a potentially fatal concern. While considering that additional source domains provide the benefits of a larger and more diverse sample set (or additional labels), these domains might have inherently different distributions of features or different joint distribution with the labels, which could mean that applying DA might ultimately bring more cost than benefit (*90*). In these worst-case scenarios, applying DA can result in what is known as negative transfer, which is when the application of knowledge from a source domain negatively affects the performance of a model in a target domain (*89*, *93*). For instance, Wang and colleagues (*93*) applied a domain-adversarial neural network (*94*) to transfer knowledge from product images as source domain to real-world images as target domain but found that the models' accuracy on the target domain decreased by 10% because of divergence in lighting, angles, and photo backgrounds between domains. Crucially, the potential for negative transfer can be amplified when working with biological data due to its already-heterogeneous nature and the smaller sample size of each dataset, and due to unknown adaptability between biological domains. Therefore, it is imperative that the adaptability of the particular biological datasets in question be explicitly quantified or estimated before applying DA methods. Unfortunately, while there exist a few methods to quantify adaptability between domains (*89*, *91*), analysis in the context of different biological subfields is exceedingly rare. The development of adaptability analysis methods thus may be a fruitful and critical area of future research into DA application to biological datasets.

CONSIDERATIONS FOR SELECTING AND APPLYING DA APPROACHES

Despite the challenges noted above, even in their current state, DA approaches can still provide benefit in biological data at this critical expansion of data sharing and open science practices in biology.

However, there are a great many methods to choose from. How should a scientist select the best DA approaches for their own datasets or scientific questions? In this section, we outline specific considerations for biologists in selecting and applying DA approaches in their own research.

We begin this section by presenting a formal definition of domain and DA. We then present a taxonomy that can be useful in gaining a better understanding of what to search for in the literature. In this Review, we focus on the primary subcategory of DA that addresses data bias or covariate shift; this DA subcategory tries to align shifts in the feature spaces between domains (or the change in the marginal distribution of data samples across domains). Other specialized subcategories of domain shift include label shift (*95*), which indicates that different domains contain different number of labels for each class, and concept shift (*96*), in which the data distribution remains the same but the conditional distribution changes [i.e., $P_s(y \mid X) \neq P_t(y \mid X)$]. Interested readers should refer to these surveys (*97*, *98*) for a comprehensive overview of the different types of shifts in the DA field.

What is a domain?

A domain can be defined as $D = \{\chi, P(X)\}$, where χ is a feature space, $X = \{x_1, x_2, \ldots, x_n\}$ is an instance set with x_i denoting a given feature, n denotes the number of features or dimensions in the data (e.g., in fMRI data voxel activities or taxa in microbiome data), and $P(X)$ denotes the marginal probability distribution of all samples in that dataset. This formal definition is typically used in discussions of DA across a wide variety of disciplines (*99*, *100*).

The terminology of DA

For a specific domain, we define the task (e.g., predicting what image a subject is looking at from neuroimaging data or predicting a disease state from microbiome composition) as $T = \{y, f(\cdot)\}$, where y denotes the labels to be predicted and $f(\cdot)$ denotes a decision function [i.e., the posterior probability distribution of $P(y \mid X)$ of the joint distribution $P(X, y)$] that needs to be learned to map input features to the corresponding labels. Given these definitions, DA is faced with the following problem, in which the distributions or relative alignment of features across domains are different but the task remains approximately the same. Thus, a DA problem with covariate shift can be formally defined as follows

$$P(X_{s_1}) \neq P(X_{s_2}) \neq \ldots \neq P(X_{s_k}) \neq P(X_t)$$

$$T_{s_1} \approx T_{s_2} \approx \ldots \approx T_{s_k} \approx T_t$$

where s denotes the source domain, t denotes the target domain, k is the number of source domains, $P(X)$ is the marginal distribution of a specific instance set in a given domain, and T is the task performed in each domain. Here, the goal of DA is to improve the performance of target decision function $f(\cdot)_t$ in target domain D_t by leveraging the information from source domain D_s and decision function $f(\cdot)_s$ (which is learned on the source domain after the source and target domains are aligned). In other words, DA intends to adapt the model(s) trained from a source (or sources) to a different, but related, target dataset. It does this by aligning the distributions of features and samples belonging to different domains so that the models emphasize learning domain invariant features that are not dependent on a specific dataset (Fig. 2). The methods by which DA accomplishes this alignment differ depending on algorithm specifics; interested readers can refer to the Supplementary Materials for details, in which we catalog a number of different algorithms and their various applications.

A taxonomy of DA

In general, when undertaking a DA analysis, we should consider three main factors:

1) The data used to train a model may be collected from multiple sources or just from a single source.

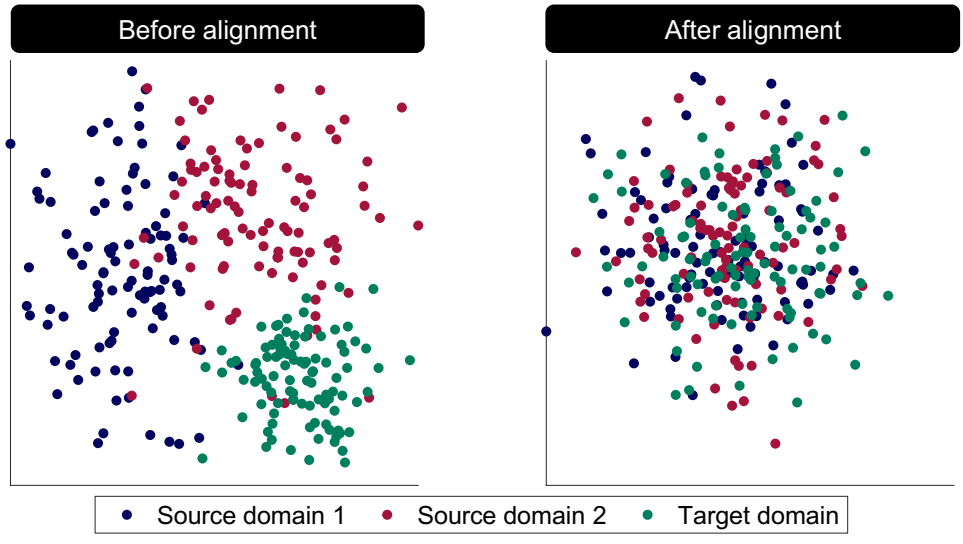

Fig. 2. A cartoon representation of source and target domains before and after alignment. In this cartoon, features vary in their values along two dimensions, and each domain's features take on a different mean and covariance. Unless the domains are aligned, these differences could both obscure other meaningful variation in the data that are shared across domains and prevent models trained on one domain from generalizing to another.

2) Depending on the availability of labels in the target domain, we might choose supervised, semi-supervised, or unsupervised models.

3) The feature spaces in the source(s) and target domains can be homogenous, meaning that they have the same dimensionality and "meaning," e.g., feature A in source 1 represents the same "type" of information as feature A in source 2, or heterogenous, meaning that the feature spaces may differ in terms of dimensionality and/or meaning.

In the following, we discuss these three factors in more detail. Table 3 also shows a summary of these categories accompanied by mathematical annotations.

Single versus multisource

In selecting a DA method, one question you will want to ask is how many domains are present. As mentioned above, DA techniques can be divided into two categories of "single source" and "multisource" (*101*). In single-source DA, the source domain is usually labeled, while the target domain belongs to another domain that has a different distribution (*13*, *85*). Single-source DA is simpler than multisource DA since there are only two distributions of data—source and target. Therefore, single-source DA is a good technique when there is enough data available in both the source and target domains to effectively train a model that can perform well on the target domain (*38*, *102*–*104*).

However, in modern real-world data sharing initiatives, most biological data come from many sources (*105*, *106*), and using these data to their full extent can facilitate novel insights. Therefore it is advantageous to develop models that leverage all available sources. This problem can be addressed through multisource DA, which aims to combine multiple sources of labeled data to make predictions about a similar task on a target dataset (*101*, *105*, *107*, *108*). A naive way to solve this problem is to combine multiple sources into one big source domain and then approach the problem as a single-source DA (*101*, *109*). However, these methods can show very limited improvement in performance—and sometimes even worse performance—in comparison to using only one source (*110*), specifically stemming from challenges of aligning the sources to begin with. Another way to tackle this problem could be to train a model on each source independently, apply each trained model to the target domain, and then vote for the "correct" label in the target domain based on the prediction across sources (*111*). One could also attempt to first find domain-invariant features among all source and target domains (*112*) or use a two-stage alignment technique that first tries to find domain-invariant feature spaces for each source-target pairing and then align model outputs across these spaces (*110*). In all cases, though, multisource DA is more challenging than single-source DA—a problem

Table 3. Difference among traditional machine learning, transfer learning, and various kinds of DA. ML, machine learning; DA, domain adaptation. χ represents feature space, and $P(X)$ is the marginal distribution of instance set X, T denotes the performed task, and $f(\cdot)$ is the decision function to map each sample to the corresponding label. s denotes the source domain, t denotes the target domain, and k is the number of source domains.

	Categories | definitions	Domains, $D = \{\chi, P(X)\}$ and tasks, $T = \{Y, f(\cdot)\}$	Verbal description
Traditional ML versus transfer learning	Traditional ML	$D_s = D_t$ and $T_s = T_t$	When the source (i.e., training set) and target (i.e., test set) have the same distribution and the task is exactly the same.
	Transfer learning (TL)	$D_s \neq D_t$ or $T_s \neq T_t$ or both	When the source and target domains have different distributions or the performed task on source and target are different, or both.
Single- versus multisource DA	Single-source DA	$P(X_s) \neq P(X_t)$ and $T_s \approx T_t$	When there is only one source domain and the marginal distribution of the feature space between source and target domain is different. The task in the target domain is similar to that in the source domain.
	Multisource DA	$P(X_{s_1}) \neq P(X_{s_2}) \neq \ldots \neq P(X_{s_k}) \neq P(X_t)$, and $T_{s_1} \approx T_{s_2} \approx \ldots \approx T_{s_k} \approx T_t$	When there are multiple sources available that can have different distributions, and when these distributions differ from that of the target domain. The task is similar across all domains.
Supervised, semi-, or unsupervised	Supervised	$P(X_s) \neq P(X_t)$, with all target labels	When source and target domains are both labeled.
	Semi-supervised	$P(X_s) \neq P(X_t)$, with some target labels	When source is labeled but target is partially labeled.
	Unsupervised	$P(X_s) \neq P(X_t)$, with no target labels	When source is labeled but target is not labeled.
Homogeneous versus heterogeneous	Homogeneous DA	$P(X_s) \neq P(X_t)$ and $\chi_s = \chi_t$ and $T_s \approx T_t$	When the feature spaces have the same dimensionality and the same meaning.
	Heterogeneous DA	$P(X_s) \neq P(X_t)$ and $\chi_s \neq \chi_t$ and $T_s \approx T_t$	When the feature spaces have different dimensionality or different meanings.

made worse by the particular characteristics of biological data, as discussed above.

Supervised versus semi-supervised versus unsupervised

It is also important to assess what kinds of labels are available for your data, across all the domains you need to align; this will dictate whether you should select a supervised, semi-supervised, or unsupervised DA method. These labels have been applied in varying ways (*13*, *101*, *113*–*115*). Here, we have chosen a categorization based strictly on the usage of target labels: In unsupervised DA, no label is available in the target domain (*38*, *85*, *116*, *117*); in semi-supervised DA (*118*–*120*), some labels are available to use; and in supervised DA, labels in the target domain are available for most samples (*97*). Although most DA techniques in existing literature focus on unsupervised DA (since it is often used for the purpose of annotating unlabeled data in the target domain), in the case of biological data, any of the supervised, semi-supervised, or unsupervised scenarios is possible. This is because the primary goal of DA in biological settings is to uncover insights about biological systems that generalize across domains. Thus, even when labeled data are available in the target domain, one can still benefit from using DA techniques on different datasets to find generalizable patterns across domains.

Homogeneous versus heterogeneous

Last, it is important to understand how the features are related across your different domains. DA can be divided into two categories based on the relationships between these features: homogeneous or heterogeneous (*97*, *99*, *101*). In homogeneous DA, the source and target domains have the same feature space, $\chi_s = \chi_t$, but the data distributions of instances of these feature spaces are different, $P(X_s) \neq P(X_t)$. That is, feature 1 in domain 1 represents the same meaning as feature 1 in domain 2—for example, they both represent a specific voxel at a specific coordinate in the brain or represent the same microbe. (Note that $\chi_s = \chi_t$ means that the feature space in both domains is homogenous, but if $X_s = X_t$, then, it means that X_s and X_t are identical datasets such that there is no difference between the source and target datasets at all.) In heterogeneous DA, conversely, the feature space is related but different between the domains. Many DA techniques that have been developed so far tend to focus on homogeneous DA (*84*, *121*–*129*). For instance, the source data could be the fMRI data obtained from a subject with one scanner and the target domain is the fMRI data obtained from the same subject with the same protocol but a different scanner. Alternatively, different domains could contain gut metagenomic sequencing data from different studies aligned against the same reference database. Addressing the domain shift in a homogeneous DA problem is relatively simpler since it is possible to perform the feature alignment directly on the original instances of the domains without the need to project them into a common feature space.

Unfortunately, however, most biological datasets are heterogeneous in nature (*51*, *53*) since these data are collected in different laboratories, under different environmental and experimental conditions, and sometimes even for answering different but related questions. In other words, neither the feature spaces nor the marginal distributions are the same [i.e., $\chi_s \neq \chi_t$,] $P(X_s) \neq P(X_t)$]. As a result, biological datasets very often have different feature dimensionalities, and, sometimes, these features even have different labels or come from different modalities of data collection (e.g., fMRI versus another neuroimaging modality like electroencephalography). For instance, the fMRI data from the brains of two individuals have different numbers of voxels (features), which also are not meaningfully aligned across individuals with respect to their functional properties (e.g., voxel 1 in person A is unlikely to encode the same information as voxel 1 in person B)—even when the scanner, protocol, and performed task are exactly the same.

Case studies and practical examples

Given the nature of most biological datasets, which often contain limited samples and originate from many different sources, the most common DA setting in this field is multisource heterogenous DA settings. For instance, aggregating fMRI data from multiple subjects or even multiple sites (*130*–*132*) can be considered a multisource heterogeneous DA. It is multisource because the data are coming from multiple subjects or multiple sites with different MRI scanners, and it is heterogeneous because the number of voxels (i.e., features) from each subject and the information they represent is different. (Note that the number of voxels can be equated through spatial normalization to a standardized template, but this does not address that each voxel will still represent different information across individuals.). In the microbiome field, integration of data from multiple microbiome datasets in order to predict a phenotype on a held-out study (*133*–*135*) is once again multisource and heterogeneous, as data are often amplicons of different regions of the 16*S* rRNA gene. To illustrate the utility of existing DA approaches and explore their categorization with the taxonomy discussed above, here, we select several methods to discuss in slightly more detail (summarized in Table 4).

One DA method, the PRECISE method (*136*), has been used to predict patients' drug response based on available preclinical datasets such as cell lines and patient-driven xenografts (PDXs). To achieve this, the authors first extracted factors from cell lines, PDXs, and human tumors using principal component analysis (PCA). Then, they aligned these subspaces from human tumor data with preclinical data using geometric transformations and extracted common features associated with biological processes followed by training a regression model using consensus genes and validated with known biomarker-drug associations to accurately predict drug response in patients. In this study, DA was homogenous, as the features (genes) in the source and target domains were the same; multisource, as various source domains were used (i.e., cell lines), and supervised, as the labels of all samples were used.

Another method, Adversarial Inductive Transfer Learning (AITL) (*137*), similarly aims to use largely available source domains such as cell lines and clinical trials to predict drug responses on small and hard-to-obtain gene expression data from patients. To this end, researchers first used a feature extractor network to map the source and target into a common feature space. This mapping aimed to alleviate the domain shift by using a global discriminator to learn domain-invariant features. Then, these domain-invariant features were used to build a regression model for the source task (i.e., predicting median inhibitory concentration) and a classification network to make predictions on the target task (i.e., predicting whether there is reduction in the size of the tumor). This study aimed to address both prior and covariate shifts in the source and target domains. The data used in this study came from multiple heterogeneous sources including thousands of cell lines from different cancer types. Last, the target samples were labeled. This study can thus be characterized as a multisource and supervised heterogeneous (i.e., drug response is categorized differently between preclinical and clinical settings) DA scenario.

Table 4. Case studies and their categorization according to our DA taxonomy.

Method	Goal	Single or multi source?	Supervised, semi-, or unsupervised?	Homogeneous or heterogeneous?
PRECISE (*136*)	Predict patients' drug Response based on preclinical datasets	Multi	Supervised	Homogenous
Adversarial inductive transfer learning (AITL) (*137*)	Predict drug responses on small and hard-to-obtain gene expression data	Multi	Supervised	Homogenous
WENDA (*138*)	Predict a human's age using DNA methylation data, which are known to differ across tissues	Multi	Unsupervised	Homogenous
Li and colleagues' (*3*)	Improve classification accuracy of autism diagnosis by detecting biomarkers in resting-state fMRI Autism Brain Imaging Data Exchange (ABIDE) datasets (*139*) from multiple sites	Multi	Supervised	Homogenous
Deep cross-subject adaptation decoding (DCAD) (*142*)	Learn common spatiotemporal patterns within a source fMRI domain (person) to generate labels for another person	Single	Unsupervised	Heterogeneous

Other methods such as WENDA (*138*) (Weighted Elastic Net for unsupervised DA) aim to predict a human's age using DNA methylation data, which are known to be different across different tissues. WENDA aims to use the available DNA methylation data from some tissues (source domains) to predict the age of the human subject using DNA methylation from a different tissue (target domain) by giving more importance to features that are more robust and behave in a similar fashion across source and target domains. In this study, data from 19 different tissues with chronological age ranging from 0 to 103 years old were used as the source domain. The target domain came from 13 different tissues, with chronological age ranging from 0 to 70 years old. In the application of WENDA, the source domain remained unchanged, while each tissue type was viewed as a distinct target domain. This thus represents a multisource, unsupervised, homogenous DA scenario.

In another study, Li and colleagues (*3*) propose a multisource DA approach by using resting-state fMRI ABIDE datasets (*139*) from multiple academic sites (UMI, NYU, USM, and UCLA). Their goal was to improve the classification accuracy of autism diagnosis by detecting biomarkers. In this study, the feature space, denoted as χ, was extracted features from fMRI sites such that $\chi_i = \chi_j$, with i and j representing different institutions (the data can be spatially normalized across participants by warping to MNI space). From this perspective, this problem is a homogeneous DA scenario. Subsequently, the authors used a Mixture of Experts (*140*, *141*), combining multiple neural networks—each of which is specialized in solving a specific task—to improve the overall performance of the model, and adversarial domain alignment methods to minimize the discrepancies between the domains, and successfully demonstrated the advantage of using federated DA techniques in using multisite fMRI dataset to classify autism. In addition, they were able to reveal possible biomarkers in the brain for autism classification. Therefore, in this framing, this can be considered as a multisource and supervised homogeneous DA problem.

Last, Gao and colleagues (*142*) proposed the deep cross-subject adaptation decoding (DCAD) method: a single-source, unsupervised, and heterogeneous DA technique. DCAD uses a three-dimensional (3D) feature extraction framework using 3D convolution and pooling operations based on volume fMRI data to learn common spatiotemporal patterns within a source domain to generate labels (*142*). Subsequently, an unsupervised DA method minimizes the discrepancy between source and target distributions. This process considers different subjects as different sources and aids in the precise decoding of cognitive states (in working memory tasks) across subjects. To validate the approach, they applied task-fMRI data from the Human Connectome Project (*143*) dataset. The experimental outcomes revealed exceptional decoding performance, achieving state-of-the-art accuracy rates of 81.9 and 84.9% under two conditions (four brain states and nine brain states, respectively) during working memory tasks. In addition, this study demonstrated that unsupervised DA effectively mitigates data distribution shifts, offering an excellent solution to enhance cross-subject decoding performance without relying on annotations.

FUTURE DIRECTIONS

What is missing from DA approaches in biological applications?

Despite these exciting successes, continued development of DA approaches tailored to the challenges of biological data is critically needed. This is especially important in light of the increasing availability of curated open datasets, complemented by increasing metadata standardization (*4*, *5*). We thus hope that the machine learning community will continue to develop techniques that can address relevant limitations of biological datasets, including:

1) Models must be able to capture the nonlinear and complex patterns in biological systems, ideally with minimal or no assumptions.

Therefore, many linear-based or parametric DA techniques (usually focused on some sort of predetermined transformation from source to target domain) might not be adequate. See the Supplementary Materials for detailed descriptions of some existing DA techniques, many of which rely on such predetermined parametric assumptions. We recommend a concerted research program to catalog the successes and failures of existing DA approaches with respect to different types of biological datasets, with attention to the impact of predetermined parametric assumptions.

2) Ideally, we want to use DA to find the underlying mechanisms of biological phenomena rather than simply aggregating data for automatic annotation. Unfortunately, many existing techniques are primarily developed for addressing automatic annotation of unlabeled data. Therefore, to fully unleash the power of DA in biological systems, we must focus on methods that seek to find domain-invariant features that are common across datasets. This usually happens by mapping all domains into a common feature space. We recommend that DA research for biological application should prioritize discovery of latent or shared spaces between domains and ideally those which are "interpretable" or "explainable." ["Interpretability" or "explainability" in machine learning may be defined as "how well a human could understand the decisions in the given context" (*144*).]

3) This domain-invariant mapping should be done using methods that work with limited data in individual cohorts. Although deep learning models are great tools to uncover highly nonlinear and complex relations in data with no specific assumptions, they often require many samples. Recently, simpler neural network architectures such as TRACE (*145*) and Fader networks (*78*) have shown promise with small fMRI datasets. However, many of the powerful neural network architectures such as generative adversarial networks might not be suitable for biological datasets as they usually require vast amounts of data (*146*, *147*). We recommend that DA research focuses on performance under data-scarce regimes, including explicit truncation of training datasets to evaluate DA methods under highly undesirable sample-to-feature ratios.

4) While some methods do exist to quantify adaptability between domains (*89*, *91*), limited attention has been paid to how such methods may fare in biological contexts. We recommend that DA research develops adaptability assessment methodologies with specific focus on biological datasets.

In sum, it is incumbent upon us in the biological disciplines to challenge machine learning research to design more flexible and broadly applicable DA methods that can perform under the constraints of real-world biological datasets. An important step toward this goal will be to test and evaluate existing approaches on our own data and on data available through broad and consistently annotated shared data repositories, to comprehensively explore and categorize their current shortcomings. Thus, we hope that, with the help of the topics discussed in this Review, researchers in biological disciplines will feel empowered to try out existing DA approaches and to help catalog their successes and shortcomings, which can then support the efforts of DA researchers to maximize the utility of such methods for biology.

If you would like to use DA techniques to augment your own data processing pipeline, we urge you to begin by gaining a comprehensive perspective on your data using the definitions and taxonomy described above. For example, How many sources do you have available? What is the sample size in each source? Do these sources contain equal amounts of features? If not, what are the nature of features in each source? Are these features in each source known and have a label? What task are you trying to achieve? Depending on the answers, you can choose the appropriate DA approaches and set about examining their successes or failures. We hope that the tools and information provided in this Review will encourage you to do so and to report your findings so that iterative improvements in DA approaches can be made to best serve our fields.

Promises for the future

In this piece, we have focused on human neuroimaging (specifically fMRI) and microbiome sciences as token examples to speculate the potential promises of DA in computational biology as a whole. We hope that these selected case studies have helped to show off the potential of DA in numerous and varied biological disciplines, from electrophysiology, multi-omics, DNA sequencing, and single-cell RNA sequencing to protein localization—all of which face similar challenges in data collection and labeling to the case study fields discussed here. Differences in equipment, experimental setup, or even individuals can lead to a shift in the distribution of data, even when the task is identical. In all cases, however, our goal as researchers and clinicians is to go beyond domain-specific or dataset-specific models to find domain-general and informative "truths" about biological systems.

Thus, DA could be extremely useful to aggregate diverse biological datasets available across the Open Science Framework, OpenNeuro, Neurosynth, Dryad, CEDAR, and more in search of meaningful and even clinically relevant outcomes (*148*–*151*). However, much work is needed to address the existing challenges. It is the intention of this paper to help and facilitate these processes by bringing more awareness of DA and the need to develop new techniques that are compatible with the limitations of biological datasets in order to make it accessible to biologists. If we are successful in identifying the challenges of performing DA on biological data, we are optimistic that DA and transfer learning methodologies can greatly benefit biologists.

Supplementary Materials

This PDF file includes:

Supplementary Text
Table S1
References

REFERENCES AND NOTES

1. L. N. Ross, D. S. Bassett, Causation in neuroscience: Keeping mechanism meaningful. *Nat. Rev. Neurosci.* **25**, 81–90 (2024).
2. A. J. DeGrave, J. Janizek, S.-I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
3. X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, J. S. Duncan, Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **65**, 101765 (2020).
4. M. Zizienová, New OSF metadata to support data sharing policy compliance. (2023).
5. M. A. Musen, C. A. Bean, K.-H. Cheung, M. Dumontier, K. A. Durante, O. Gevaert, A. Gonzalez-Beltran, P. Khatri, S. H. Kleinstein, M. J. O'Connor, Y. Pouliot, P. Rocca-Serra, S.-A. Sansone, J. A. Wiser, CEDAR team, The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc.* **22**, 1148–1152 (2015).
6. S. Zhao, X. Zhao, G. Ding, K. Keutzer, EmotionGAN: Unsupervised domain adaptation for learning discrete probability distributions of image emotions, in *Proceedings of the 26th ACM International Conference on Multimedia* (Association for Computing Machinery, 2018) *MM '18*, pp. 1319–1327.
7. A. Torralba, A. A. Efros, Unbiased look at dataset bias, in *CVPR 2011* (2011), pp. 1521–1528.
8. H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, D. K. Bhattacharyya, Big data analytics in bioinformatics: A machine learning perspective. arXiv:1506.05101 [cs.CE] (2015).

9. L. Duan, D. Xu, I. Tsang, Learning with augmented features for heterogeneous domain adaptation. arXiv:1206.4660 [cs.LG] (2012).
10. M. Harel, S. Mannor, Learning from multiple outlooks. arXiv:1005.0027 [cs.LG] (2010).
11. P. Prettenhofer, B. Stein, Cross-language text classification using structural correspondence learning, in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2010), pp. 1118–1127.
12. J. Zhou, S. Pan, I. Tsang, Y. Yan, Hybrid heterogeneous transfer learning through deep learning. *AAAI* **28**, 10.1609/aaai.v28i1.8961 (2014).
13. S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
14. W. Ling, J. Lu, N. Zhao, A. Lulla, A. M. Plantinga, W. Fu, A. Zhang, H. Liu, H. Song, Z. Li, J. Chen, T. W. Randolph, W. L. A. Koay, J. R. White, L. J. Launer, A. A. Fodor, K. A. Meyer, M. C. Wu, Batch effects removal for microbiome data via conditional quantile regression. *Nat. Commun.* **13**, 5418 (2022).
15. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
16. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
17. Y. Zhang, Y. Wei, Q. Wu, P. Zhao, S. Niu, J. Huang, M. Tan, Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Trans. Image Process.* **29**, 7834–7844 (2020).
18. C. Chen, Q. Dou, H. Chen, J. Qin, P.-A. Heng, Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. *AAAI* **33**, 865–872 (2019).
19. B. Berger, N. M. Daniels, Y. W. Yu, Computational biology in the 21st century: Scaling with compressive algorithms. *Commun. ACM* **59**, 72–80 (2016).
20. K. Lan, D.-T. Wang, S. Fong, L.-S. Liu, K. K. L. Wong, N. Dey, A survey of data mining and deep learning in bioinformatics. *J. Med. Syst.* **42**, 139 (2018).
21. K. A. Shastry, H. A. Sanjay, Machine learning for bioinformatics, in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, K. G. Srinivasa, G. M. Siddesh, S. R. Manisekhar, Eds. (Springer Singapore, 2020), pp. 25–39.
22. M. W. Libbrecht, W. S. Noble, Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
23. P. Liu, X. Qiu, X. Huang, Adversarial multi-task learning for text classification. arXiv:1704.05742 [cs.CL] (2017).
24. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
25. H. Venkateswara, J. Eusebio, S. Chakraborty, S. Panchanathan, Deep hashing network for unsupervised domain adaptation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5018–5027.
26. X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, Moment matching for multi-source domain adaptation, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 1406–1415.
27. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
28. A. Torralba, R. Fergus, W. T. Freeman, 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**, 1958–1970 (2008).
29. A. Krizhevsky, G. Hinton, Others, Learning multiple layers of features from tiny images. (2009).
30. N. Altman, M. Krzywinski, The curse(s) of dimensionality. *Nat. Methods* **15**, 399–400 (2018).
31. P. D. Schloss, Identifying and overcoming threats to reproducibility, replicability, robustness, and generalizability in microbiome research. *MBio* **9**, e00525-18 (2018).
32. B. O. Turner, E. J. Paul, M. B. Miller, A. K. Barbey, Small sample sizes reduce the replicability of task-based fMRI studies. *Commun Biol* **1**, 62 (2018).
33. S. Zhou, C. R. Cox, H. Lu, Improving whole-brain neural decoding of fMRI with domain adaptation, in *Machine Learning in Medical Imaging* (Springer International Publishing, 2019), pp. 265–273.
34. M. Wasikowski, X.-W. Chen, Combating the small sample class imbalance problem using feature selection. *IEEE Trans. Knowl. Data Eng.* **22**, 1388–1400 (2010).
35. H. He, E. A. Garcia, Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009).
36. K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains. in Computer Vision–ECCV 2010 (Springer, 2010), pp. 213–226.
37. M. Long, H. Zhu, J. Wang, M. I. Jordan, Deep transfer learning with joint adaptation networks, arXiv:1605.06636 [cs.LG] (2016).
38. M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach, D. Blei, Eds. (PMLR, 2015) vol. 37 of *Proceedings of Machine Learning Research*, pp. 97–105.
39. E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, T. Darrell, Deep domain confusion: Maximizing for domain invariance, arXiv:1412.3474 [cs.CV] (2014).
40. I. Shavitt, E. Segal, Regularization learning networks: Deep learning for tabular datasets. arXiv:1805.06440 [stat.ML] (2018).
41. S. Ö. Arik, T. Pfister, TabNet: Attentive interpretable tabular learning. *AAAI* **35**, 6679–6687 (2021).
42. D. McElfresh, S. Khandagale, J. Valverde, C. V. Prasad, G. Ramakrishnan, M. Goldblum, C. White, When do neural nets outperform boosted trees on tabular data? *Adv. Neural Inf. Process. Syst.* **36**, 76336–76369 (2023).
43. L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data? in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, Eds. (Curran Associates Inc., 2022) vol. 35, pp. 507–520.
44. Y. Yang, S. Soatto, FDA: Fourier domain adaptation for semantic segmentation, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2020), pp. 4085–4095.
45. W. Hong, Z. Wang, M. Yang, J. Yuan, Conditional generative adversarial network for structured domain adaptation, in 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 1335–1344.
46. S. Motiian, Q. Jones, S. Iranmanesh, G. Doretto, Few-shot adversarial domain adaptation. *Adv. Neural Inf. Process. Syst.* **30**, (2017).
47. K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, M. Chandraker, Unsupervised domain adaptation for face recognition in unlabeled videos, arXiv:1708.02191 [cs.CV] (2017).
48. B. Ghosh-Dastidar, J. L. Schafer, Multiple edit/multiple imputation for multivariate continuous data. *J. Am. Stat. Assoc.* **98**, 807–817 (2003).
49. N. Eisemann, A. Waldmann, A. Katalinic, Imputation of missing values of tumour stage in population-based cancer registration. *BMC Med. Res. Methodol.* **11**, 129 (2011).
50. D. van Dijk, J. Nainys, R. Sharma, P. Kaithail, A. J. Carr, K. R. Moon, L. Mazutis, G. Wolf, S. Krishnaswamy, D. Pe'er, MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. bioRxiv 111591 [Preprint] (2017). https://doi.org/10.1101/111591.
51. M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, M. M. Hoffman, Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Inf. Fusion* **50**, 71–91 (2019).
52. J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, P. J. Ramadge, A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
53. C. Xu, S. A. Jackson, Machine learning and complex biological data. *Genome Biol.* **20**, 76 (2019).
54. V. Z. Marmarelis, Identification of nonlinear biological systems using Laguerre expansions of kernels. *Ann. Biomed. Eng.* **21**, 573–589 (1993).
55. D. Singh, H. Climente-Gonzalez, M. Petrovich, E. Kawakami, M. Yamada, FsNet: Feature selection network on high-dimensional biological data, in *2023 International Joint Conference on Neural Networks (IJCNN)* (IEEE, 2023), pp. 1–9.
56. A. Y. Pan, Statistical analysis of microbiome data: The challenge of sparsity. *Curr. Opin. Endocr. Metab. Res.* **19**, 35–40 (2021).
57. T. Zhou, M. Liu, K.-H. Thung, D. Shen, Latent representation learning for Alzheimer's disease diagnosis with incomplete multi-modality neuroimaging and genetic data. *IEEE Trans. Med. Imaging* **38**, 2411–2422 (2019).
58. P. Samartsidis, S. Montagna, A. R. Laird, P. T. Fox, T. D. Johnson, T. E. Nichols, Estimating the prevalence of missing experiments in a neuroimaging meta-analysis. *Res. Synth. Methods* **11**, 866–883 (2020).
59. T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, D. Shen, Multi-modal latent space inducing ensemble SVM classifier for early dementia diagnosis with neuroimaging data. *Med. Image Anal.* **60**, 101630 (2020).
60. E. L. Busch, L. Slipski, M. Feilong, J. S. Guntupalli, M. V. di Oleggio Castello, J. F. Huckins, S. A. Nastase, M. I. Gobbini, T. D. Wager, J. V. Haxby, Hybrid Hyperalignment: A single high-dimensional model of shared information embedded in cortical patterns of response and functional connectivity. *Neuroimage* **233**, 117975 (2021).
61. P. Wei, Y. Ke, C. K. Goh, A general domain specific feature transfer framework for hybrid domain adaptation. *IEEE Trans. Knowl. Data Eng.* **31**, 1440–1451 (2019).
62. C. Wang, S. Mahadevan, Heterogeneous domain adaptation using manifold alignment, in *Twenty-Second International Joint Conference on Artificial Intelligence* (2011), pp. 1541–1546. https://aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/viewPaper/3207.
63. B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, S. P. Holmes, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
64. A. Amir, D. McDonald, J. A. Navas-Molina, E. Kopylova, J. T. Morton, Z. Zech Xu, E. P. Kightley, L. R. Thompson, E. R. Hyde, A. Gonzalez, R. Knight, Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* **2**, e00191-16 (2017).
65. O. Esteban, C. J. Markiewicz, R. W. Blair, C. A. Moodie, A. I. Isik, A. Erramuzpe, J. D. Kent, M. Goncalves, E. DuPre, M. Snyder, H. Oya, S. S. Ghosh, J. Wright, J. Durnez, R. A. Poldrack,

K. J. Gorgolewski, fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
66. R. W. Cox, AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
67. R. W. Cox, J. S. Hyde, Software tools for analysis and visualization of fMRI data. *NMR Biomed.* **10**, 171–178 (1997).
68. M. W. Woolrich, S. Jbabdi, B. Patenaude, M. Chappell, S. Makni, T. Behrens, C. Beckmann, M. Jenkinson, S. M. Smith, Bayesian analysis of neuroimaging data in FSL. *Neuroimage* **45**, S173–S186 (2009).
69. S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. J. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, R. K. Niazy, J. Saunders, J. Vickers, Y. Zhang, N. De Stefano, J. M. Brady, P. M. Matthews, Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, S208–S219 (2004).
70. M. Jenkinson, C. F. Beckmann, T. E. J. Behrens, M. W. Woolrich, S. M. Smith, FSL. *Neuroimage* **62**, 782–790 (2012).
71. A. Andronache, C. Rosazza, D. Sattin, M. Leonardi, L. D'Incerti, L. Minati, Coma Research Centre (CRC)–Besta Institute, Impact of functional MRI data preprocessing pipeline on default-mode network detectability in patients with disorders of consciousness. *Front. Neuroinform.* **7**, 16 (2013).
72. N. Bhagwat, A. Barry, E. W. Dickie, S. T. Brown, G. A. Devenyi, K. Hatano, E. DuPre, A. Dagher, M. Chakravarty, C. M. T. Greenwood, B. Misic, D. N. Kennedy, J.-B. Poline, Understanding the impact of preprocessing pipelines on neuroimaging cortical surface analyses. *Gigascience* **10**, giaa155 (2021).
73. M. A. Lindquist, S. Geuter, T. D. Wager, B. S. Caffo, Modular preprocessing pipelines can reintroduce artifacts into fMRI data. *Hum. Brain Mapp.* **40**, 2358–2376 (2019).
74. E. Ibrahimi, M. B. Lopes, X. Dhamo, A. Simeon, R. Shigdel, K. Hron, B. Stres, D. D'Elia, M. Berland, L. J. Marcos-Zambrano, Overview of data preprocessing for machine learning applications in human microbiome research. *Front. Microbiol.* **14**, 1250909 (2023).
75. G. Cammarota, G. Ianiro, A. Ahern, C. Carbone, A. Temko, M. J. Claesson, A. Gasbarrini, G. Tortora, Gut microbiome, big data and machine learning to promote precision medicine for cancer. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 635–648 (2020).
76. Y.-M. Kim, J.-B. Poline, G. Dumas, Experimenting with reproducibility: A case study of robustness in bioinformatics. *Gigascience* **7**, (2018).
77. X. Liu, J. Wu, W. Li, Q. Liu, L. Tian, H. Huang, Domain adaptation via low rank and class discriminative representation for autism spectrum disorder identification: A multi-site fMRI study. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 806–817 (2023).
78. M. Pominova, E. Kondrateva, M. Sharaev, A. Bernstein, E. Burnaev, Fader networks for domain adaptation on fMRI: ABIDE-II study, in *Thirteenth International Conference on Machine Vision* (SPIE, 2021) vol. 11605, pp. 570–577.
79. X. Liu, H. Huang, Alterations of functional connectivities associated with autism spectrum disorder symptom severity: A multi-site study using multivariate pattern analysis. *Sci. Rep.* **10**, 4330 (2020).
80. A. van Opbroek, M. A. Ikram, M. W. Vernooij, M. de Bruijne, Transfer learning improves supervised image segmentation across imaging protocols. *IEEE Trans. Med. Imaging* **34**, 1018–1030 (2015).
81. W. M. Kouw, M. Loog, L. W. Bartels, A. M. Mendrik, MR acquisition-invariant representation learning, arXiv:1709.07944 [cs.CV] (2017).
82. B. Wang, F. Sun, Y. Luan, Comparison of the effectiveness of different normalization methods for metagenomic cross-study phenotype prediction under heterogeneity. *Sci. Rep.* **14**, 7024 (2024).
83. X. Yang, C. Deng, T. Liu, D. Tao, Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 1992–2003 (2022).
84. F. Liu, G. Zhang, J. Lu, Heterogeneous domain adaptation: An unsupervised approach. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 5588–5602 (2020).
85. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, in *Domain Adaptation in Computer Vision Applications* (Springer International Publishing, 2017) *Advances in computer vision and pattern recognition*, pp. 189–209.
86. R. Vinuesa, B. Sirmacek, Interpretable deep-learning models to help achieve the sustainable development goals. *Nat. Mach. Intell.* **3**, 926–926 (2021).
87. D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, S. Lacoste-Julien, A closer look at memorization in deep networks, in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 06--11 Aug. 2017) vol. 70 of *Proceedings of Machine Learning Research*, pp. 233–242.
88. P. W. Koh, P. Liang, Understanding Black-box Predictions via Influence Functions, in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 06--11 Aug 2017) vol. 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894.
89. A. Mehra, B. Kailkhura, P.-Y. Chen, J. Hamm, Understanding the limits of unsupervised domain adaptation via data poisoning. *Adv. Neural Inf. Process. Syst.* **1327**, 17347–17359 (2021).
90. S. Ben-David, T. Lu, T. Luu, D. Pál, Impossibility theorems for domain adaptation. *AISTATS* **9**, 129–136 (2010).
91. I. Redko, A. Habrard, M. Sebban, On the analysis of adaptability in multi-source domain adaptation. *Mach. Learn.* **108**, 1635–1652 (2019).
92. H. Liu, M. Long, J. Wang, M. Jordan, Transferable adversarial training: A general approach to adapting deep classifiers, in *Proceedings of the 36th International Conference on Machine Learning*, K. Chaudhuri, R. Salakhutdinov, Eds. (PMLR, 09–15 Jun 2019) vol. 97 of *Proceedings of Machine Learning Research*, pp. 4013–4022.
93. Z. Wang, Z. Dai, B. Póczos, J. Carbonell, Characterizing and avoiding negative transfer, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019), pp. 11285–11294.
94. H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, Domain-adversarial neural networks. arXiv:1412.4446 [stat.ML] (2014).
95. Y. S. Chan, H. T. Ng, Word sense disambiguation with distribution estimation. https://ijcai.org/Proceedings/05/Papers/1543.pdf.
96. W. M. Kouw, M. Loog, An introduction to domain adaptation and transfer learning. arXiv:1812.11806 [cs.LG] (2018).
97. M. Wang, W. Deng, Deep visual domain adaptation: A survey. *Neurocomputing* **312**, 135–153 (2018).
98. G. Wilson, D. J. Cook, A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* **11**, 1–46 (2020).
99. G. Csurka, Domain adaptation for visual applications: A comprehensive survey. arXiv:1702.05374 [cs.CV] (2017).
100. X. Liu, C. Yoo, F. Xing, H. Oh, G. El Fakhri, J.-W. Kang, J. Woo, Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing* **11**, (2022).
101. S. Zhao, B. Li, C. Reed, P. Xu, K. Keutzer, Multi-source domain adaptation in the deep learning era: A systematic survey. arXiv:2002.12169 [cs.LG] (2020).
102. B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, Unsupervised visual domain adaptation using subspace alignment, in *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 2960–2967.
103. B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (ieeexplore.ieee.org, 2012), pp. 2066–2073.
104. B. Sun, K. Saenko, Deep CORAL: Correlation alignment for deep domain adaptation, in *Computer Vision–ECCV 2016 Workshops* (Springer International Publishing, 2016), pp. 443–450.
105. S. Sun, H. Shi, Y. Wu, A survey of multi-source domain adaptation. *Inf. Fusion* **24**, 84–92 (2015).
106. H. S. Bhatt, A. Rajkumar, S. Roy, Multi-source iterative adaptation for cross-domain classification, in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*, pp. 3691–3697. https://ijcai.org/Proceedings/16/Papers/519.pdf.
107. T. Matsuura, T. Harada, Domain generalization using a mixture of multiple latent domains. *AAAI* **34**, 11749–11756 (2020).
108. E. F. Montesuma, F. M. N. Mboula, Wasserstein barycenter for multi-source domain adaptation, in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2021), pp. 16785–16793.
109. Y. Mansour, M. Mohri, A. Rostamizadeh, Domain adaptation with multiple sources. *Adv. Neural Inf. Process. Syst.* **21**, 1041–1048 (2008).
110. Y. Zhu, F. Zhuang, D. Wang, Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. *AAAI* **33**, 5989–5996 (2019).
111. R. Xu, Z. Chen, W. Zuo, J. Yan, L. Lin, Deep cocktail network: Multi-source unsupervised domain adaptation with category shift, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3964–3973.
112. H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, G. J. Gordon, Adversarial multiple source domain adaptation. *Adv. Neural Inf. Process. Syst.* **31**, 8559–8570 (2018).
113. H. Guan, M. Liu, Domain adaptation for medical image analysis: A survey. *I.E.E.E. Trans. Biomed. Eng.* **69**, 1173–1185 (2022).
114. H. Daumé III, Frustratingly easy domain adaptation. arXiv:0907.1815 [cs.LG] (2009).
115. K. Saito, Y. Ushiku, T. Harada, Asymmetric tri-training for unsupervised domain adaptation, in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup, Y. W. Teh, Eds. (PMLR, 06--11 Aug 2017) vol. 70 of *Proceedings of Machine Learning Research*, pp. 2988–2997.
116. A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training. arXiv:1612.07828 [cs.CV] (2016).
117. J. Zhuo, S. Wang, W. Zhang, Q. Huang, Deep unsupervised convolutional domain adaptation, in *Proceedings of the 25th ACM International Conference on Multimedia* (Association for Computing Machinery, 2017) *MM '17*, pp. 261–269.

118. W. Li, L. Duan, X. Dong, I. W. Tsang, Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 1134–1148 (2014).
119. E. Tzeng, J. Hoffman, T. Darrell, K. Saenko, Simultaneous deep transfer across domains and tasks. arXiv:1510.02192 [cs.CV] (2015).
120. K. Saito, D. Kim, S. Sclaroff, T. Darrell, K. Saenko, Semi-supervised domain adaptation via minimax entropy, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (IEEE, 2019), pp. 8050–8058.
121. H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* **90**, 227–244 (2000).
122. J. Yang, R. Yan, A. G. Hauptmann, Cross-domain video concept detection using adaptive svms, in *Proceedings of the 15th ACM International Conference on Multimedia* (Association for Computing Machinery, 2007) *MM '07*, pp. 188–197.
123. L. Duan, I. W. Tsang, D. Xu, T.-S. Chua, Domain adaptation from multiple sources via auxiliary classifiers, in *Proceedings of the 26th Annual International Conference on Machine Learning* (Association for Computing Machinery, 2009) *ICML '09*, pp. 289–296.
124. L. Duan, D. Xu, S.-F. Chang, Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach, in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 1338–1345.
125. J. T. Zhou, I. W.Tsang, S. J. Pan, M. Tan, Heterogeneous domain adaptation for multiple classes, in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, S. Kaski, J. Corander, Eds. (PMLR, 2014) vol. 33 of *Proceedings of Machine Learning Research*, pp. 1095–1103.
126. B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation. *AAAI* **30**, (2016).
127. J. Guo, D. Shah, R. Barzilay, Multi-source domain adaptation with mixture of experts. arXiv:1809.02256 [cs.CL] (218).
128. M. Mancini, L. Porzi, S. R. Bulo, B. Caputo, E. Ricci, Boosting domain adaptation by discovering latent domains, in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 3771–3780.
129. S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, K. Keutzer, Multi-source distilling domain adaptation. *AAAI* **34**, 12975–12983 (2020).
130. M. Wang, D. Zhang, J. Huang, P.-T. Yap, D. Shen, M. Liu, Identifying autism spectrum disorder with multi-site fMRI via low-rank domain adaptation. *IEEE Trans. Med. Imaging* **39**, 644–655 (2020).
131. A. Mensch, J. Mairal, D. Bzdok, B. Thirion, G. Varoquaux, Learning neural representations of human cognition across many fMRI studies. arXiv:1710.11438 [stat.ML] (2017).
132. H. Zhang, P.-H. Chen, P. Ramadge, Transfer learning on fMRI datasets, in *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey, F. Perez-Cruz, Eds. (PMLR, 2018) vol. 84 of *Proceedings of Machine Learning Research*, pp. 595–603.
133. C. Huang, C. Gin, J. Fettweis, B. Foxman, B. Gelaye, D. A. MacIntyre, A. Subramaniam, W. Fraser, N. Tabatabaei, B. Callahan, Meta-analysis reveals the vaginal microbiome is a better predictor of earlier than later preterm birth. *BMC Biol.* **21**, 199 (2023).
134. J. L. Golob, T. T. Oskotsky, A. S. Tang, A. Roldan, V. Chung, C. W. Y. Ha, R. J. Wong, K. J. Flynn, A. Parraga-Leo, C. Wibrand, S. S. Minot, B. Oskotsky, G. Andreoletti, I. Kosti, J. Bletz, A. Nelson, J. Gao, Z. Wei, G. Chen, Z.-Z. Tang, P. Novielli, D. Romano, E. Pantaleo, N. Amoroso, A. Monaco, M. Vacca, M. De Angelis, R. Bellotti, S. Tangaro, A. Kuntzleman, I. Bigcraft, S. Techtmann, D. Bae, E. Kim, J. Jeon, S. Joe, Preterm Birth DREAM Community, K. R. Theis, S. Ng, Y. S. Lee, P. Diaz-Gimeno, P. R. Bennett, D. A. MacIntyre, G. Stolovitzky, S. V. Lynch, J. Albrecht, N. Gomez-Lopez, R. Romero, D. K. Stevenson, N. Aghaeepour, A. L. Tarca, J. C. Costello, M. Sirota, Microbiome preterm birth DREAM challenge: Crowdsourcing machine learning approaches to advance preterm birth research. *Cell Rep. Med.* **5**, 101350 (2024).
135. G. I. Austin, A. B. Kav, H. Park, J. Biermann, A.-C. Uhlemann, T. Korem, Processing-bias correction with DEBIAS-M improves cross-study generalization of microbiome-based prediction models. bioRxiv 579716 [Preprint] (2024). https://doi.org/10.1101/2024.02.09.579716.
136. S. Mourragui, M. Loog, M. A. van de Wiel, M. J. T. Reinders, L. F. A. Wessels, PRECISE: A domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. *Bioinformatics* **35**, i510–i519 (2019).
137. H. Sharifi-Noghabi, S. Peng, O. Zolotareva, C. C. Collins, M. Ester, AITL: Adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics. *Bioinformatics* **36**, i380–i388 (2020).
138. L. Handl, A. Jalali, M. Scherer, R. Eggeling, N. Pfeifer, Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data. *Bioinformatics* **35**, i154–i163 (2019).
139. A. Di Martino, C.-G. Yan, Q. Li, E. Denio, F. X. Castellanos, K. Alaerts, J. S. Anderson, M. Assaf, S. Y. Bookheimer, M. Dapretto, B. Deen, S. Delmonte, I. Dinstein, B. Ertl-Wagner, D. A. Fair, L. Gallagher, D. P. Kennedy, C. L. Keown, C. Keysers, J. E. Lainhart, C. Lord, B. Luna, V. Menon, N. J. Minshew, C. S. Monk, S. Mueller, R.-A. Müller, M. B. Nebel, J. T. Nigg, K. O'Hearn, K. A. Pelphrey, S. J. Peltier, J. D. Rudie, S. Sunaert, M. Thioux, J. M. Tyszka, L. Q. Uddin, J. S. Verhoeven, N. Wenderoth, J. L. Wiggins, S. H. Mostofsky, M. P. Milham, The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
140. S. Masoudnia, R. Ebrahimpour, Mixture of experts: A literature survey. *Artif. Intell. Rev.* **42**, 275–293 (2014).
141. N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, J. Dean, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv:1701.06538 [cs.LG] (2017).
142. Y. Gao, Y. Zhang, Z. Cao, X. Guo, J. Zhang, Decoding brain states from fMRI signals by using unsupervised domain adaptation. *IEEE J. Biomed. Health Inform.* **24**, 1677–1685 (2020).
143. D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. E. J. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss, S. Della Penna, D. Feinberg, M. F. Glasser, N. Harel, A. C. Heath, L. Larson-Prior, D. Marcus, G. Michalareas, S. Moeller, R. Oostenveld, S. E. Petersen, F. Prior, B. L. Schlaggar, S. M. Smith, A. Z. Snyder, J. Xu, E. Yacoub, WU-Minn HCP Consortium, The human connectome project: A data acquisition perspective. *Neuroimage* **62**, 2222–2231 (2012).
144. T. Miller, Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell* **267**, 1–38 (2019).
145. S. Orouji, V. Taschereau-Dumouchel, A. Cortese, Task-relevant autoencoding enhances machine learning for human neuroscience. arXiv:2208.08478 [q-bio.NC] (2022).
146. L. Hou, Regularizing label-augmented generative adversarial networks under limited data. *IEEE Access* **11**, 28966–28976 (2023).
147. R. Webster, J. Rabin, L. Simon, F. Jurie, Detecting overfitting of deep generative networks via latent recovery. arXiv:1901.03396 [cs.LG] (2019).
148. A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, J. G. Sanders, J. Shorenstein, H. Holste, S. Petrus, A. Robbins-Pianka, C. J. Brislawn, M. Wang, J. R. Rideout, E. Bolyen, M. Dillon, J. G. Caporaso, P. C. Dorrestein, R. Knight, Qiita: Rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
149. E. Pasolli, L. Schiffer, P. Manghi, A. Renson, V. Obenchain, D. T. Truong, F. Beghini, F. Malik, M. Ramos, J. B. Dowd, C. Huttenhower, M. Morgan, N. Segata, L. Waldron, Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14**, 1023–1024 (2017).
150. A. Amir, E. Ozel, Y. Haberman, N. Shental, Achieving pan-microbiome biological insights via the dbBact knowledge base. *Nucleic Acids Res.* **51**, 6593–6608 (2023).
151. R. J. Abdill, S. P. Graham, V. Rubinetti, F. W. Albert, C. S. Greene, S. Davis, R. Blekhman, Integration of 168,000 samples reveals global patterns of the human gut microbiome. bioRxiv 560955 [Preprint] (2023). https://doi.org/10.1101/2023.10.11.560955.
152. K. Muandet, D. Balduzzi, B. Schölkopf, Domain generalization via invariant feature representation. *Proc. Mach. Learn.* **28**, 10–18 (2013).
153. J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, R. A. Irizarry, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).
154. P. I. Costea, G. Zeller, S. Sunagawa, E. Pelletier, A. Alberti, F. Levenez, M. Tramontano, M. Driessen, R. Hercog, F.-E. Jung, J. R. Kultima, M. R. Hayward, L. P. Coelho, E. Allen-Vercoe, L. Bertrand, M. Blaut, J. R. M. Brown, T. Carton, S. Cools-Portier, M. Daigneault, M. Derrien, A. Druesne, W. M. de Vos, B. B. Finlay, H. J. Flint, F. Guarner, M. Hattori, H. Heilig, R. A. Luna, J. van Hylckama Vlieg, J. Junick, I. Klymiuk, P. Langella, E. Le Chatelier, V. Mai, C. Manichanh, J. C. Martin, C. Mery, H. Morita, P. W. O'Toole, C. Orvain, K. R. Patil, J. Penders, S. Persson, N. Pons, M. Popova, A. Salonen, D. Saulnier, K. P. Scott, B. Singh, K. Slezak, P. Veiga, J. Versalovic, L. Zhao, E. G. Zoetendal, S. D. Ehrlich, J. Dore, P. Bork, Towards standards for human fecal sample processing in metagenomic studies. *Nat. Biotechnol.* **35**, 1069–1076 (2017).
155. N. Herndon, D. Caragea, Naive bayes domain adaptation for biological sequences, in *Proceedings of the 4th International Conference on Bioinformatics Models, Methods and Algorithms, BIOINFORMATICS* (2013), pp. 62–70.
156. J. Wirbel, P. T. Pyl, E. Kartal, K. Zych, A. Kashani, A. Milanese, J. S. Fleck, A. Y. Voigt, A. Palleja, R. Ponnudurai, S. Sunagawa, L. P. Coelho, P. Schrotz-King, E. Vogtmann, N. Habermann, E. Niméus, A. M. Thomas, P. Manghi, S. Gandini, D. Serrano, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, L. Waldron, A. Naccarati, N. Segata, R. Sinha, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork, G. Zeller, Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. *Nat. Med.* **25**, 679–689 (2019).
157. A. M. Thomas, P. Manghi, F. Asnicar, E. Pasolli, F. Armanini, M. Zolfo, F. Beghini, S. Manara, N. Karcher, C. Pozzi, S. Gandini, D. Serrano, S. Tarallo, A. Francavilla, G. Gallo, M. Trompetto, G. Ferrero, S. Mizutani, H. Shiroma, S. Shiba, T. Shibata, S. Yachida, T. Yamada, J. Wirbel, P. Schrotz-King, C. M. Ulrich, H. Brenner, M. Arumugam, P. Bork, G. Zeller, F. Cordero, E. Dias-Neto, J. C. Setubal, A. Tett, B. Pardini, M. Rescigno, L. Waldron, A. Naccarati, N. Segata, Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat. Med.* **25**, 667–678 (2019).

158. H. Daumé III, A. Kumar, A. Saha, Frustratingly easy semi-supervised domain adaptation, in *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing* (2010), pp. 53–59.
159. M. Schneider, L. Wang, C. Marr, Evaluation of domain adaptation approaches for robust classification of heterogeneous biological data sets, in *Artificial Neural Networks and Machine Learning–ICANN 2019: Deep Learning* (Springer International Publishing, 2019), pp. 673–686.
160. J. Blitzer, R. McDonald, F. Pereira, Domain adaptation with structural correspondence learning, in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (2006), pp. 120–128.
161. M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, M. Salzmann, Domain adaptation on the statistical manifold, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 2481–2488.
162. B. Gong, K. Grauman, F. Sha, Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation, in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta, D. McAllester, Eds. (PMLR, 2013) vol. 28 of *Proceedings of Machine Learning Research*, pp. 222–230.
163. H. Hotelling, Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936).
164. D. R. Hardoon, S. Szedmak, J. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* **16**, 2639–2664 (2004).
165. Y.-R. Yeh, C.-H. Huang, Y.-C. F. Wang, Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Trans. Image Process.* **23**, 2009–2018 (2014).
166. P. L. Lai, C. Fyfe, Kernel and nonlinear canonical correlation analysis. *Int. J. Neural Syst.* **10**, 365–377 (2000).
167. F. R. Bach, Kernel independent component analysis. *J. Mach. Learn. Res.* **3**, 1–48 (2002).
168. J. V. Haxby, J. S. Guntupalli, S. A. Nastase, M. Feilong, Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
169. J. S. Guntupalli, M. Feilong, J. V. Haxby, A computational model of shared fine-scale structure in the human connectome. *PLOS Comput. Biol.* **14**, e1006120 (2018).
170. C. Wang, S. Mahadevan, Manifold alignment without correspondence, in *Twenty-First International Joint Conference on Artificial Intelligence* (2009; https://aaai.org/ocs/index.php/IJCAI/IJCAI-09/paper/viewPaper/446).
171. P.-H. C. Chen, J. Chen, Y. Yeshurun, U. Hasson, J. Haxby, P. J. Ramadge, A reduced-fimension fMRI shared response ,odel, in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett, Eds. (Curran Associates Inc., 2015) vol. 28, pp. 460–468.
172. M. R. McLaren, A. D. Willis, B. J. Callahan, Consistent and correctable bias in metagenomic sequencing experiments. *eLife* **8**, e46923 (2019).
173. H. Wang, W. Yang, Z. Lin, Y. Yu, TMDA: Task-specific multi-source domain adaptation via clustering embedded adversarial training, in *2019 IEEE International Conference on Data Mining (ICDM)* (2019), pp. 1372–1377.
174. E. Tzeng, J. Hoffman, K. Saenko, T. Darrell, Adversarial discriminative domain adaptation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (openaccess.thecvf.com, 2017), pp. 7167–7176.
175. M.-Y. Liu, O. Tuzel, Coupled Generative Adversarial Networks, in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates Inc., 2016; https://proceedings.neurips.cc/paper/2016/file/502e4a16930e414107ee22b6198c578f-Paper.pdf), vol. 29.
176. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2014; https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf), vol. 27.
177. Y. Bengio, Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2**, 1–127 (2009).
178. P. Vincent, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **11**, 3371–3408 (2010).
179. K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, in *Advances in Neural Information Processing Systems*. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates Inc., 2016; https://proceedings.neurips.cc/paper/2016/file/45fbc6d3e05ebd93369ce542e8f2322d-Paper.pdf) vol. 29.
180. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.
181. J. Hoffman, M. Mohri, N. Zhang, Algorithms and theory for multiple-source adaptation. arXiv:1805.08727 [cs.LG] (2018).
182. P.-H. Chen, X. Zhu, H. Zhang, J. S. Turek, J. Chen, T. L. Willke, U. Hasson, P. J. Ramadge, A convolutional autoencoder for multi-subject fMRI data aggregation. arXiv:1608.04846 [stat.ML] (2016).
183. X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, H. Wang, An efficient LSTM network for emotion recognition from multichannel EEG signals. *IEEE Trans. Affect. Comput.* **13**, 1528–1540 (2020).
184. I. Sosin, D. Kudenko, A. Shpilman, Continuous gesture recognition from sEMG sensor data with recurrent neural networks and adversarial domain adaptation, in *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)* (IEEE, 2018). pp. 1436–1441.
185. J. Liu, X. Gong, Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. *BMC Bioinformatics* **20**, 609 (2019).
186. M. Ahmed, J. Islam, M. R. Samee, R. E. Mercer, Identifying protein-protein interaction using tree LSTM and structured attention, in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)* (2019). pp. 224–231.
187. Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation. in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach, D. Blei, Eds. (PMLR, 07–09 Jul 2015) vol. 37 of *Proceedings of Machine Learning Research*, pp. 1180–1189.
188. M. Baktashmotlagh, M. Salzmann, U. Dogan, M. Kloft, F. Orabona, T. Tommasi, Distribution-matching embedding for visual domain adaptation. *J. Mach. Learn. Res.* **17**, 1–30 (2016).
189. J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, CyCADA: Cycle-consistent adversarial domain adaptation, in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy, A. Krause, Eds. (PMLR, 10–15 Jul 2018) vol. 80 of *Proceedings of Machine Learning Research*, pp. 1989–1998.
190. S. Rakshit, B. Banerjee, G. Roig, S. Chaudhuri, Unsupervised multi-source domain adaptation driven by deep adversarial ensemble learning, in *Pattern Recognition* (Springer International Publishing, 2019), pp. 485–498.
191. A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel method for the two-sample-problem. *Adv. Neural Inf. Process. Syst.* **19**, (2007).
192. G. Kang, L. Jiang, Y. Yang, A. G. Hauptmann, Contrastive adaptation network for unsupervised domain adaptation, in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2019), pp. 4893–4902.
193. H. Guo, R. Pasunuru, M. Bansal, Multi-source domain adaptation for text classification via DistanceNet-bandits. *AAAI* **34**, 7830–7838 (2020).
194. B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, N. Courty, DeepJDOT: Deep joint distribution optimal transport for unsupervised domain adaptation. arXiv:1803.10081 [cs.CV] (2018).
195. M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, W. Li, Deep reconstruction-classification networks for unsupervised domain adaptation, in *Computer Vision–ECCV 2016* (Springer International Publishing, 2016), pp. 597–613.
196. M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, Domain generalization for object recognition with Multi-task autoencoders. arXiv:1508.07680 [cs.CV] (2015).
197. Y.-H. Tsai, W.-C. Hung, S. Schulter, K. Sohn, M.-H. Yang, M. Chandraker, Learning to adapt structured output space for semantic segmentation. arXiv:1802.10349 [cs.CV] (2018).
198. J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *2017 IEEE International Conference on Computer Vision (ICCV)* (IEEE, 2017), pp. 2223–2232.
199. S. Zhao, C. Lin, P. Xu, S. Zhao, Y. Guo, R. Krishna, G. Ding, K. Keutzer, CycleEmotionGAN: Emotional semantic consistency preserved CycleGAN for adapting image emotions. *AAAI* **33**, 2620–2627 (2019).

Acknowledgments: We thank H. H. Lee for initial discussions on this topic. **Funding:** This work was supported by two Canadian Institute for Advanced Research Azrieli Global Scholars Fellowships [in Brain, Mind, & Consciousness (to M.A.K.P.) and in Humans and the Microbiome (to TK)], and a Canadian Institute for Advanced Research Catalyst Grant (to M.A.K.P. and T.K.). The funders had no involvement in the design or content of this work. **Author contributions:** S.O.: Methodology, investigation, data curation, writing—original draft, and writing—review and editing. M.C.L.: Methodology, investigation, data curation, writing—original draft, and writing—review and editing. T.K.: Conceptualization, methodology, investigation, resources, data curation, writing—original draft; writing—review and editing, supervision, project administration, and funding acquisition. M.A.K.P.: Conceptualization, methodology, investigation, resources, data duration, writing—original draft, writing—review and editing, supervision, project administration, and funding acquisition. **Competing interests:** The authors declare that they have no competing interests.

Submitted 9 May 2024
Accepted 15 November 2024
Published 20 December 2024
10.1126/sciadv.adp6040