Transfer Learning of Clinical Outcomes with Molecular Data, Principles and Perspectives

Axel Kowald[1], **Israel Barrantes**[1], **Steffen Möller**[1], **Daniel Palmer**[1], **Hugo Murua Escobar**[2], **Georg Fuellen**[1,3]*

1 Institute for Biostatistics and Informatics in Medicine and Ageing Research, Rostock University Medical Center, Rostock, Germany
2 Department of Medicine, Clinic III, Hematology, Oncology, Palliative Medicine, Rostock University Medical Center, Rostock, Germany
3 Centre for Transdisciplinary Neurosciences Rostock and Research Focus Oncology, Rostock and Ageing of Individuals and Society, Interdisciplinary Faculty, Rostock University, Rostock, Germany

*corresponding author, fuellen@uni-rostock.de

Keywords: transfer learning, classification, regression

Abstract. Accurate transfer learning of clinical outcomes, e.g., of the effects and side effects of drugs or other interventions, from one cellular context to another (in-vitro versus ex-vivo versus in-vivo, or across tissues), between cell-types, developmental stages, omics modalities or species, is considered tremendously useful. Ultimately, it may avoid most drug development failing in translation, despite large investments in the preclinical stages, which includes animal experiments requiring careful justification. Thus, when transferring a prediction task from a source (model) domain to a target domain, what counts is the high quality of the predictions in the target domain, requiring molecular states or processes common to both source and target that can be learned by the predictor, reflected by latent variables. These latent variables may form a compendium of knowledge that is learned in the source, to enable predictions in the target; usually, there are few, if any, labeled target training samples to learn from. Transductive learning then refers to the learning of the predictor in the source domain, transferring its outcome label calculations to the target domain, considering the same task. Inductive learning considers cases where the target predictor is performing a different yet related task as compared to the source predictor, making some labeled target data necessary. Often, there is also a need to first map the variables in the input/feature spaces (e.g. of gene names to orthologs) and/or the variables in the output/outcome spaces (e.g. by matching of labels). Transfer across omics modalities also requires that the molecular information flow connecting these modalities is sufficiently conserved. Only one of the methods for transfer learning we reviewed offers an assessment of input data, suggesting that transfer learning is unreliable in certain cases. Moreover, source domains feature their very own particularities, and transfer learning should consider these, e.g., as differences in pharmacokinetics, drug clearance or the microenvironment. In light of these general considerations, we here discuss and juxtapose various recent transfer learning approaches, specifically designed (or at least adaptable) to *predict clinical (human in-vivo) outcomes based on molecular data*, towards finding the right tool for a given task, and paving the way for a comprehensive and systematic comparison of the suitability and accuracy of transfer learning of clinical outcomes.

Introduction: Questions and issues in translation research.

“Translation” in biomedicine requires the transfer of knowledge from a source domain to a target domain. In biomedicine, molecular and clinical data become available at a large scale, with the potential to help with diagnosis, prognosis, as well as treatment monitoring, selection and development. Such data may be available as a source of knowledge from, e.g., human blood, cell cultures or animals, but they are often not available for the human target tissue, e.g. an inoperable tumor or brain tissue. More generally, translating (i.e. transferring) insights from models, across tissues or species, etc., to the human in-vivo situation has been a long-term challenge. Given cancer cell line or advanced tumor model data, can we predict human treatment success of cancer drugs? Given toxicology data from rabbits, rodents or dogs, can we predict drug mediated toxicity or tolerance in humans?

Easily accessible molecular data for the model (or proximate) situations of blood, cell cultures or animals are becoming abundant in biomedicine, often based on high-throughput technologies (omics). Organoids and multi-organ-chip models are also the source of additional model data. As a first step, we may wish to work on the successful transfer of conserved *molecular processes and associated biomarkers*. On that basis, we may be able to process and organize the model data so that there can be a successful transfer of insights about *outcomes*, referring to the human in-vivo situation. Alternatively, outcome predictors may be transferred directly from the model to humans. For the example of predicting intervention outcomes in humans, knowledge of processes and biomarkers then enables trials “enriched” with likely responders, or with co-treatment of non-responders, to counter insufficient effect (or overly-strong side-effects). Even better, we would like to estimate the chances of success of such a transfer a priori.

Accurate transfer is so valuable because insights for human in-vivo are so hard to obtain; clinical outcomes are usually expensive to establish, and the constrained accessibility of most human tissues implies that detailed data, mechanistic understanding and useful biomarkers cannot easily be obtained either. Moreover, biomedicine is a “rich” discipline, where most molecular processes, including their causal influence on phenotypes, are rarely conserved: they differ by cellular context (in-vitro versus ex-vivo versus in-vivo, or across tissues), cell-type, developmental stage, molecular entity (omics modality) and species, and even by genetics. Case in point are the on-going discussions about the similarity or dissimilarity of immune responses in mouse versus human (see, for example, Seok *et al.* (2013) versus Takao and Miyakawa (2015), and the discussion in Brubaker *et al.* (2019)). Thus, the context-dependency of cellular responses and of their high-level phenotypic implications is significant. It is one root cause of what is called the “reproducibility crisis” and it is at the heart of translational failures. Also, correlation and causality are not easy to discern. While correlative relationships are sometimes sufficient (e.g. for a biomarker to be predictive), causal relationships are telling us much more about the information flow that starts molecularly and ends up in generating the high-level outcome phenotypes that are of ultimate interest. Transfer learning based on causal relationships can thus be expected to be more successful in general. Nevertheless, we must aim for accurate transfer learning to the best of our abilities.

Molecular data and similarity of molecular processes.

With the introduction of high-throughput technologies such as genotype-phenotype association mapping (GWAS and polygenic risk scores) and gene expression measurements by microarray or RNAseq (transcriptomics, single-cell or bulk), the molecular mechanisms of intervention effects and side effects can be investigated more and more thoroughly in-vitro (for human and animal) as well as in-vivo (mostly for animal). However, the molecules that we can measure as potential markers are just a glimpse of the intricate in-vivo situation, whereby the measurements for one molecular modality (such as mRNA) are in a complex relationship to the measurements of another (such as protein). Further, the available datasets often differ significantly in quality, lacking comprehensive sample descriptions incl. detailed source specifications as well as adequate sample processing. As described, in-vivo molecular human data are scarce due to limited tissue accessibility, though blood may be available and genetic information is readily obtainable. But at the core of the translation gap are two issues: context-dependent and incongruent data. As the authors of AITL (see below) point out, considering in-vitro to in-vivo translation: “Two major discrepancies exist between preclinical and clinical datasets: (i) in the input space, the gene expression data, due to difference in the basic biology, and (ii) in the output space, the different measures of the drug response. Therefore, training a computational model on cell lines and testing it on patients violates the i.i.d assumption that train and test data are from the same distribution” (Sharifi-Noghabi *et al.*, 2020). Other frequently encountered gaps hindering transfer are the species gap and the complex relationships between the molecular entities that may be measured (see above).

Transfer learning, terminology and examples.

By default, we follow Sharifi-Noghabi *et al.* (2020) in adopting the terminology of Pan and Yang (2010), which is a widely cited review and reference in the field of transfer learning. In their paper, transfer learning is defined in terms of source and target domains (of features with probability distributions associated with these), as well as source and target tasks (mapping features to labels using predictors) so that the predictor in the target domain uses some knowledge from the source domain. In a simple case, that “use of knowledge” means to learn a predictor for the source task in the source domain, and then to just use it, after matching the input/output variables, as the predictor for the target task in the target domain. For any real transfer to take place, the source and target domain, or the source and target tasks, must of course be distinct. For example, gene expression to phenotype relationships may be learned by a neural network in one species, and then transferred to another species. Considering the transfer gaps just described, a sufficient degree of similarity, or conservation, of the input-output relationships between the source and target is the key to accurate transfer learning from a well-sampled source domain to an under-sampled or unknown target domain. Considering the predictors, which map the input features to output labels, latent variables are calculated by these predictors in the source domain, and the conserved role of the latent variables in the target is a necessity for transfer learning to succeed. In the "simple case" just introduced, after learning in the source domain, the predictor contains knowledge about this domain. This knowledge is "latent", or "hidden" in the predictor. If the predictor is a neural net, this knowledge is expected to be represented by the weights that were learned in the source domain. These weights may then reflect how gene expression maps to phenotypes. If that map is sufficiently similar in the two domains, that is, in the two species under consideration, then the predictor can be applied successfully in the target domain.

In general, latent variables can be thought of as low-dimensional representations of the input data. As another simple example, a principal component analysis (PCA) derives these as eigenvectors of a matrix of an all-against-all covariance analysis. Often, a compendium of latent variables is learned from the source data. The transductive flavor of transfer learning then entails a target domain without any labeled samples, and it is hoped that the predictor which uses the latent variables from the source domain can repeat its success in the target domain (Pan & Yang, 2010). For this to happen, the prediction/classification task must be the same in source and target, as it was for the neural network above, learning to map gene expression to phenotype. The inductive flavor of transfer learning considers tasks that are different yet "related" (Pan & Yang, 2010). There is usually no formal definition of this relatedness; the "proof is in the pudding", that is, accurate transfer in terms of correct predictions/classifications is the indicator of sufficient relatedness of the task in the source and the task in the target domain. Moreover, for inductive transfer learning, at least a few labelled samples are needed in the target domain, so that the knowledge about the latent variables can be grounded to some true relations in the target domain. Modifying our example from above, if the phenotype is morbidity in one species and mortality in the other species, the task is different yet related and inductive transfer learning is needed. We also consider the entirely unsupervised flavor of transfer learning, where none of the samples feature any labels. In this case, the latent variables are employed to transfer, from the source to the target, information useful for clustering, for better feature representation or for dimensionality reduction (Pan & Yang, 2010); see the examples given below. In this review, we will specifically consider preclinical source and clinical target domains, towards clinical outcome prediction, as well as molecular omics data as the dominant sample features that are input for the predictors.

Recent examples of transfer learning.

In Table 1, we collected some recent representative examples, without claim to completeness. We aimed for high-level descriptions while keeping formulas at a minimum. We describe the source domain, frequently also known as the "background model" or "compendium", and the target domain. Further, we describe the input and output of the predictor that is learned, and the kind of transfer learning methodology employed, labeling it as "transductive," "inductive" or "unsupervised", following

Pan and Yang (2010). This table lays the foundation for finding the right tool for a user's task, by conceptual similarity matching of the user's task to the entries in the table. Optimally, this matching follows a principled approach, considering the kind of transfer learning (see also the Discussion and Figure 1). It can also be seen as the *starting point* for a comprehensive and systematic comparison of transfer learning methods, considering a variety of application scenarios. In the following, we give a textual description of the examples in Table 1, describing some aspects of the methods that did not fit into the table.

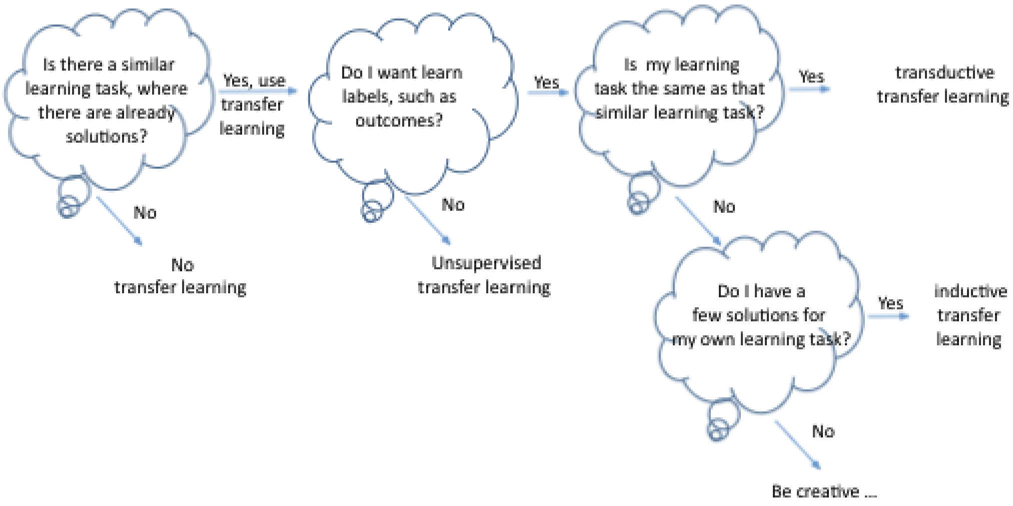
Is there a similar learning task, where there are already solutions?
Yes, use transfer learning
Do I want learn labels, such as outcomes?
Yes
Is my learning task the same as that similar learning task?
Yes
transductive transfer learning
No
No transfer learning
No
Unsupervised transfer learning
No
Do I have a few solutions for my own learning task?
Yes
inductive transfer learning
No
Be creative ...

Figure 1: How to select a transfer learning methodology for the task you have. (See also the Discussion.)

Adversarial Inductive Transfer Learning (AITL) Sharifi-Noghabi *et al.* (2020) is explicitly described using the terminology of (Pan & Yang, 2010), and it bridges in-vitro (source domain, human cancer cell line data) and in-vivo (target domain, human cancer patient data) in two ways. On one hand, it transfers gene expression knowledge based on cell-lines to patients, where the expression profiling was done to describe the response to chemotherapeutic drugs. For cell lines, the data stem from GDSC (Genomics of Drug Sensitivity in Cancer), and the labels are IC50 values. For patients, the data stem from TCGA (The Cancer Genome Atlas) and some other sources, and the labels (if available) are binary, reflecting response/non-response (yes/no). The different output labels are handled by multi-task learning. More specifically, a multi-class predictor is trained on both source and target samples, utilizing a binarized outcome in case of the source samples; this simultaneous learning on source and target data is also suggested to improve accuracy. The "biological" differences in the gene expression input data are handled by adversarial domain adaptation. In more detail, latent variables (called "extracted features" in the AITL framework) are learned in a domain-invariant fashion by employing an adversary network tasked with distinguishing the domains; its failure is rewarded. If the extracted features learned by AITL play a similar role in both source and target domains, AITL transfer learning can be successful.

Conditional out-of-distribution transfer learning by Lotfollahi *et al.* (2020) employs a transfer variational autoencoder (trVAE), which enables the transfer of conditions across domains. This makes it for instance possible to train a neural network on images of smiling and non-smiling men and non-smiling women, and then transfer the smiling condition ("style") from the male to the female domain. Similarly, the authors applied trVAE to a single-cell gene expression dataset of the gut (comprising eight different cell types) after infections with different bacteria. The method could successfully transfer the effects of the infection to cell types not included for training. As the name implies, the architecture of trVAE is based on an autoencoder scheme where the output layer is trained to reproduce the input layer while going through a bottleneck layer in between. trVAE modifies this approach by explicitly providing the first decoder layer with information about the condition of the input sample (e.g. smiling vs non-smiling). During training all samples are supplied with their correct condition, but for prediction the desired condition (e.g. smiling) is used, causing the last layer of the autoencoder to contain a representation of the input modified by the desired condition. Thus, tools like trVAE might be used to make predictions about human tissues and organs from which no biopsy can be obtained (e.g. brain) as long as data are available for another organ (e.g. blood) and from another domain (e.g. brain & blood of mice), enabling clinical outcome predictions based on preclinical data.

Semisupervised transfer learning as described by Brubaker *et al.* (2019) matches the transductive paradigm. The authors collected gene expression data for inflammatory diseases, consisting of samples labeled either "healthy" or "sick", that had been measured for mouse and human and constructed 36 matched pairs to which they applied various machine learning techniques (e.g. support-vector machines, k-nearest-neighbor classifiers, random forests & neural nets). The best result in terms of precision and recall for learning the labels, and, consequently, differentially expressed genes (DEG, contrasting "healthy" and "sick"), and pathways, were obtained by a semi-supervised neural net, which iteratively used the human data to augment the mouse data sets. (When validating the method, the ground truth comes from DEGs and pathways that were identified from human data using human labels.) Initially, the neural net classifier was exclusively trained on labeled mouse data and used to predict human labels based on human expression data. In the next step the human samples with the highest classification confidence were used to generate an augmented training set consisting of mouse and human data. After re-training, the classifier was then anew applied to the remaining human data and again the samples with the highest classification confidence were incorporated into the cross-species training set. The iteration ended when finally all human data were incorporated and classified. Note that this algorithm does not require the true human labels, the integration works by only using the predicted labels; the true human labels are used for validation. This strategy is a clever way to humanize animal data and seems to be applicable to a wide field of problems. It does, however, require a classification task and is not suitable for regression problems (like predicting age, speed or other numerical values), since only in the classification case the output of the machine learning algorithm can be used to assign a high or low confidence to the prediction (depending on how much a prediction is "between" classes).

PRECISE (Patient Response Estimation Corrected by Interpolation of Subspace Embeddings) (Mourragui *et al.*, 2019) uses preclinical models (cell lines and patient-derived xenografts) as predictors, despite their inherent differences as compared to real human tumors. To identify common molecular mechanisms (based on similarity of gene expression) in preclinical models and human tumors, PRECISE processes transcriptomic data to first find specific "factors" (based on PCA) for each set (preclinical models and human tumors) separately, and the factors from both sets are then aligned and compared to generate common factors (or principal vectors, PVs) between both sets, the most similar of which are then used to generate a consensus representation of the tumor model. This consensus representation can finally be employed to train a regression model of the preclinical gene expression data with respect to the preclinical drug response data, which is then applied on the real human tumor gene expression data to predict human tumor response.

PROGENy and DoRothEA These two approaches specifically aim to recover perturbations in mice at the pathway (PROGENy) and transcription factor (DoRothEA) level using human gene expression data. The first tool (PROGENy) was originally developed to assess the activity of human signaling pathways from human gene expression data, by finding pathway-specific transcriptomic footprints that entail targets of such pathways (Schubert *et al.*, 2018). In turn, DoRothEA was initially built to assess associations between transcription factor activities and drug responses in human transcriptomic data, and then it was reformulated as a resource of regulons, i.e. curated transcription factors and their transcriptional targets (Garcia-Alonso *et al.*, 2019); these regulons were curated and collected from different experimental and literature sources. Pathway and transcription factor footprints tend to be evolutionarily conserved between humans and mice, and since various studies have demonstrated that it is possible to estimate human gene expression from mouse gene expression data (Normand *et al.*, 2018; Brubaker *et al.*, 2019), the authors of PROGENy and DoRothEA adapted both tools to work with mouse data, finding 4,020 significant associations between pathways and transcription factors in mouse and human diseases by using human-mouse ortholog information. They demonstrated the reliability of these approaches by estimating the transcription factor and pathway activities from a large collection of mouse in-vitro experiments, such as chemical and genetic perturbations, as well as from mouse in-vivo disease-related experiments, and provided these results as an interactive web application. PROGENy and DoRothEA estimate "footprints" of a pathway or a transcription factor (TF) on gene expression, and the evolutionary conservation of footprint effects between human and mouse can be further investigated in detail (Holland *et al.*, 2020). On this basis, disease associations and perturbations can be inferred (and validated, e.g., by checking human-based predictions in mice), alongside pathway and TF activity scores for a large collection of human and mouse perturbation and disease experiments.

XGSEA aims to directly predict gene sets of interest in a target species under a given condition based solely on gene expression of another (source) species under the same or a comparable condition (Cai *et al.*, 2021). Shortly, gene set enrichment analysis (GSEA) is performed on the gene expression data of the source species, say mice, comparing a condition of interest to a control and thus determining significantly enriched gene sets for that phenotype. The gene sets of the target species (e.g. human gene ontology terms), are then subjected to domain adaptation based on sequence homology between their constituent genes, minimizing divergence between source and target domains while maintaining the pairwise distance between the gene sets across the domains. After domain adaptation, the tool then offers multiple options for gene set enrichment prediction, using either 1) logistic regression to predict the p-values of each gene set in the target domain, based on those in the source domain, 2) regression on the enrichment values for each gene set, then calculating the p-values directly from those or 3) regression on the enrichment values for each gene set in each direction (over and under enrichment) before calculating p-values as before. These methods were evaluated against three naïve methods (all three based on mapping target genes to source genes based on sequence homology) in four different datasets, three mouse to human and one zebrafish to human. Generally, XGSEA outperformed the naïve methods for these datasets (evaluated by comparing area under ROC curves at a range of enrichment p-value thresholds), in particular when performing regression on enrichment values for each direction of enrichment separately. To test the method further, XGSEA was used to analyze a CD8+ T cell ATACseq dataset, predicting the enriched pathways in human solid tumors from murine tumor data. The method identified gene expression and immune system terms, in addition to a large number of Notch signaling terms, as likely being enriched in the human tissue. A naïve approach performed on the same data returned a larger number of more diverse terms, so that in this case XGSEA gave more focused results.

Found In Translation (FIT) Normand *et al.* (2018) follows the unsupervised paradigm, aiming to transfer the property of being a high-effect gene from mouse to human, where a high-effect gene is characterized by a high fold change for RNA-seq datasets or a high z-test value for microarray datasets. The authors assembled mouse and human gene expression datasets from GEO that each compared a disease condition vs healthy, created 170 cross species pairings (CSP) spanning 28 human diseases (and the corresponding mouse models), and constructed a model for each CSP that aims to predict human expression values H based on mouse expression values M according to a linear relationship (for each gene ‘g’, the model is, $H_g= \alpha_g+ \beta_g*M_g$). Parameters α_g and β_g are then determined using a least-squares optimization algorithm including penalty terms for regularization. The resulting model parameters are used to predict human gene expression from mouse gene expression, in a disease-specific fashion. The accuracy of the transfer is estimated from human disease-specific datasets (disjunct from the ones on which the CSP are based), by checking whether the predicted high-effect human genes match already known ones. In fact, the ‘Found In Translation’ approach increases the mouse-human overlap of differentially expressed genes by 20-50% compared to direct cross-species extrapolation (Normand *et al.*, 2018). Furthermore, it is possible to predict whether new mouse data can be extrapolated to human by FIT using a support vector machine (SVM) classifier. The SVM basically tests if the new mouse data bear enough resemblance to the mouse data of the 170 CSPs or not.

MultiPlier (Taroni *et al.*, 2019) is an unsupervised learning approach, aiming to transfer feature representations (latent variables, that is, “patterns” based on correlation of gene expression calculated by PLIER (Mao *et al.*, 2019)) from the source to the target domain. The source domain is derived from recount2 (Collado-Torres *et al.*, 2017), a collection of disease-related gene expression datasets generated by next-generation sequencing, where all raw data were processed in a unified way, reflecting a wide variety of biological processes and pathways. The target domain can entail any gene expression dataset that is expected to feature at least some of these processes; rare disease datasets are the use case, since they feature few samples (almost) by definition. The latent variables (“patterns” of correlated genes) are calculated for the source by matrix factorization, so that each latent variable partly associates with some known pathways and cell-type-specific gene sets. Once the latent variables are learned, a new gene expression dataset can be projected into the space defined by these, and the authors show that this projection is effective in revealing biological processes related to disease severity.

Translatable components regression (TransComp-R) by Brubaker *et al.* (2020) presents an application of transfer learning to predict resistance to inflammatory bowel disease treatment with infliximab. The authors aim not only to transfer knowledge from one species to another, but also from the space of transcriptomics to proteomics. Labelled human transcriptomics data are used to infer which mouse proteomics data are predictive for responder vs non-responder phenotype in humans. First, human gene expression data are selected for genes associated with the responder phenotype. Next, mouse proteomics data are chosen for genes homologous to the human ones selected in the previous step and a PCA analysis is performed on these. Finally, the human transcriptomics data are projected into the PCA space and a regression against the human responder phenotype is performed, allowing to identify new mouse proteins that might be predictive for the human phenotype. Using this approach the authors predict a collagen-binding integrin to be involved in resistance to treatment, supported in vitro (in human) using anti-integrin antibodies.

Table 1. Transfer learning, tools and techniques.

Name/acronym/reference	Source domain	Target domain	Input	Output	Transfer method	Remarks/ weblink
Adversarial Inductive Transfer Learning (AITL) (Sharifi-Noghabi *et al.*, 2020)	application-area-specific in-vitro (cell line) gene expression & quantitative outcome (IC50) data	application-area-specific in-vivo (patient) gene expression & qualitative outcome (yes/no) data	in-vitro gene expression data (GDSC)	in-vivo outcomes (TCGA)	*inductive*: adversarial domain adaptation & multi-task learning (predicting outcomes for both source & target) using deep neural networks	code available at https://github.com/hosseinshn/AITL.
transfer variational autoencoder, trVAE (Lotfollahi *et al.*, 2020)	gene expression data, or image data (or similar) under a specific (first) condition	gene expression data, or image data (or similar) under a different (second) condition	data under the first condition and a label specifying the second condition	data transformed to the second condition	*transductive:* based on an autoencoder neural network.	Available from https://github.com/theislab/trvae_reproducibility
Semisupervised transfer learning (Brubaker *et al.*, 2019)	application-area-specific mouse phenotype-outcome -labeled gene expression data, iteratively augmented by unlabeled human gene expression data	human gene expression data	human gene expression data	human phenotype data (and subsequently DEGs and enriched pathways inferred from these)	*transductive*: supervised modeling (mouse) amended iteratively by semi-supervised retraining (adding unlabeled human data)	Matlab code available from www.mathworks.com/matlabcentral/fileexchange/69718-semisupervised-learning-functions
Patient Response Estimation Corrected by Interpolation of Subspace Embeddings (PRECISE) (Mourragui *et al.*, 2019)	gene expression data from preclinical models (cell lines, patient- derived xenografts) and drug response	human gene expression data	human gene expression data	human drug response	*transductive*: similarity-based identification of shared mechanisms between large datasets from preclinical models and a small number of human samples, focussed on cancer	Available as python package; example protocols provided as Jupyter notebooks
Pathway RespOnsive GENes (PROGENy) (Schubert *et al.*, 2018) and Discriminant Regulon	two curated resources, of footprint pathway perturbations (PROGENy), and another of footprint regulons (transcription factor - target interactions in	the mouse equivalent of the source (human) resources	mouse gene expression data	human pathway activity (PROGENy), or transcription factor activity and enrichment (DoRothEA)	*transductive*: supervised prediction of mouse pathways (PROGENy) and regulons (DoRothEA)	both available as R (Bioconductor) and python packages

Expression Analysis (DoRothEA) (Garcia-Alonso *et al.*, 2019)	DoRothEA), all from human data; human-mouse orthologs					
XGSEA (Cai *et al.*, 2021)	gene sets, from a source species, e.g., gene ontology annotations from mouse	gene sets, from a target species, e.g. gene ontology annotations from human	gene expression for specific condition(s) from source species and gene sets for enrichment analysis from both source and target species	gene sets significantly associated with the condition(s) of interest in target species	*transductive*: domain adaptation followed by regression (logistic on p-values, linear on enrichment scores or linear on positive and negative enrichment scores separately)	
Found In Translation (FIT) (Normand *et al.*, 2018)	precompiled datasets of mouse gene expression, disease vs healthy	precompiled datasets of human gene expression, disease vs healthy	mouse gene expression experiment	human genes for matching condition, with high effect	*unsupervised (dimensionality reduction):* gene-level lasso regression	available at http://www.mouse2man.org including pre-test for transferability
MultiPlier	preprocessed disease-related datasets of human gene expression, highlighting latent variables (characteristic patterns of correlated genes)	human (rare disease) gene expression data	human (rare disease) gene expression data	characteristic expression patterns of correlated rare disease genes	*unsupervised (feature representation):* constrained matrix factorization highlighting latent variables, then projection of input into latent space	
Translatable components regression (TransComp-R) (Brubaker *et al.*, 2020)	human gene expression (pretreatment), drug response data	mouse protein abundance data	human gene expression (pretreatment) and drug response data (the latter are given, not to be predicted)	mouse proteins (and corresponding pathway enrichments) with association to human drug response	*unsupervised (feature representation):* PCA-based regression	Matlab code available from https://de.mathworks.com/matlabcentral/fileexchange/77987-transcompr

Discussion, Conclusion and Perspectives

The ultimate goal in biomedical research is often to understand and tackle a disease or dysfunction in humans. What are the molecular foundations of a certain disease? Which are the diagnostic, prognostic or predictive biomarkers? Is a certain intervention effective in humans? What are the pharmacokinetics and pharmacodynamics of a drug? Unfortunately, it is often not possible to perform the necessary experiments and measurements in humans for ethical, financial or technical reasons. Also, measuring outcomes may simply take too long, given the long lifespan of humans. For these reasons, researchers often use alternative model systems in the hope that the insights from those systems can be directly applied to humans, often leading to failures. Accurate transfer learning is expected to improve this situation.

In Figure 1, we thus provide a simple decision tree from the users' perspective, regarding the question whether transfer learning may help for a task at hand. As described, application areas of transfer learning are motivated by identifying two distinct domains or tasks: the one at hand where few or no solutions of the task are given, and a different yet related one where many more solutions are available. In the biomedical application areas we consider here, the distinct domains or tasks reflect different cellular contexts (in-vitro versus ex-vivo versus in-vivo, or across tissues), cell-types, developmental stages, omics modalities or species. Very often, the source is "preclinical" but the target is "clinical". Thus, for a user faced with a task that she thinks could be solved by transfer learning, the first and foremost goal is to answer the question: "Is there a problem domain, and a learning task in that domain, different from what I am looking at, where there are already (many) solutions that may be of relevance to the task I have?". If yes, for our user, the second question then is: "Is my task a case for inductive, transductive or unsupervised transfer?". In principle, any need to learn labels (such as diagnoses, or outcomes, be they a disease prognosis or the prediction of the success or failure of an intervention) requires a supervised approach, which may be inductive or transductive. Further, an inductive method will be needed if the learning tasks are different, comparing source to target; in this case, some labeled samples are required for the target domain. For example, in the case of AITL, the source task is to predict the IC50 (a quantity), and the target task is to predict patient response (yes or no). If the tasks are the same, a transductive method would be sufficient, and "domain adaptation", e.g. by relabelling, may be the way to go; furthermore, no labeled samples are required for the target domain. For example, in case of the semi-supervised transfer learning of Brubaker *et al.* (2019), the task is phenotype/outcome prediction for human (the target domain), transferred from the same task for mouse (the source domain), and domain adaptation entails the mapping of homologous genes. Finally, if no labels need to be learned, our user may explore whether unsupervised transfer learning, e.g. of association or enrichment data, is possible and useful.

While the principled approach just outlined is the best way to start, the devil is in the details, and these details are usually very much dependent on the exact application area. Thus, apart from identifying the kind of transfer learning as just described, ideas may be borrowed from one kind of transfer learning to tackle another, as long as the application areas are the same or similar. Case in point is the (trivial) observation that homology matching is useful if the input variables are genes or proteins and the transfer is done from one species to another. In the next section, we therefore motivate transfer learning for the various application areas in some detail, and discuss some specifics to be considered. Notably, none of the methods we described is presented for more than one application area, and we are not aware of a method that underwent the “meta-transfer” from one application area to another in any follow-up work. Nevertheless, we could assign all methods to one of the three fundamental kinds of transfer learning, that is, inductive, transductive or unsupervised. However, it is hard to identify further methodological similarities among, e.g., all the transductive methods that we presented.

Transfer from **species to species** is important since animals are usually short-lived and allow experiments under controlled conditions that are not possible in humans. The more closely related the species is to humans, the more likely the transfer is expected to succeed, given an appropriate transfer learning approach. “Found in Translation" by Normand *et al.* (2018) is specifically designed to transfer results from mouse to human, for 28 disease models. Similarly, the semi-supervised method of Brubaker *et al.* (2019) also transfers between mice and humans, where the authors focus on inflammatory diseases (other diseases were not investigated). In both cases, gene/protein homology information is needed. A second relevant area of transfer is from one omics to another, specifically from **transcriptomics to proteomics**. With modern NGS techniques, transcriptomics data can easily and reproducibly be measured, but one would like to infer information about the proteome since proteins are the biomechanical machines that eventually perform most tasks in the cell. TransComp-R (Brubaker *et al.*, 2020) is such a tool that in addition also transfers information between species (mouse & human). *In-vitro* studies are also used to approximate the human *in-vivo* situation. Thus **in-vitro to in-vivo** transfer is another important area for transfer learning. AITL by Sharifi-Noghabi *et al.* (2020) is one such example, which transfers knowledge from the in-vitro to the in-vivo situation (and from quantitative output to a binary output). Not surprisingly, the tasks are different because the outcomes are not the same in cells versus humans, and inductive learning must be done. Finally, it is very helpful to transfer knowledge from **tissue to tissue**. In humans it is often not possible to obtain a sample from the affected tissue (e.g. brain or pancreas), but a blood sample can be collected easily. Moreover, the flow of blood connects most tissues of the body in one way or another, so we can expect to find the traces of many organ-specific processes in blood. Also, in cancer for example, disease-specific nucleic acids can be traced and monitored in blood samples. Using the blood transcriptome as a proxy for disease processes in other organs opens the way for a personalised medicine approach that is complementary to genetics-based approaches. For transfer learning from tissue to tissue, trVAE may be employed, as described above.

For molecular data, some structuring of the latent variables may enhance the success rates of transfer learning. For one, deep neural-net based learning is essentially a black-box approach frequently employed in transfer learning. However, structuring neural networks based on hierarchical knowledge (such as the gene ontology, GO) gained momentum and acceptance (Kuenzi *et al.*, 2020; Holzscheck *et al.*, 2021), and ontologies may be a key to structure latent spaces yielding not just better accuracy in predictor performance, but also better interpretability of the prediction/classification process. Here, knowledge about master regulators and (signaling) pathways may be encoded by gene/protein interaction and regulation subnetworks which may enable an even better structuring than the GO hierarchy, and in fact, the development of the GO is heading in a similar direction, towards investigating GO-Causal-Activity-Models (Gene Ontology, 2021).

In this review, we provided a sample of available tools and algorithms for transfer learning in biomedicine, focusing on outcome prediction, yet there are a large number of approaches and software packages. It is quite challenging to adequately compare all these tools in a coherent and fair way, but we hope to have provided a starting point. Moreover, we hope that Table 1, Figure 1 and the text may be of help for anyone facing a learning task that may profit from transfer learning.

Acknowledgements:

We acknowledge the financial support by the Federal Ministry of Education and Research (BMBF) of Germany for the SASKit study (FKZ 01ZX1903A) and the AntifibrotiX (FKZ 03VP06230) study, as well as by the European Commission (Grant agreement 633589). The funders had no role in the design of the study.

References

Brubaker, D. K., Kumar, M. P., Chiswick, E. L., Gregg, C., Starchenko, A., Vega, P. N., Southard-Smith, A. N., Simmons, A. J., Scoville, E. A., Coburn, L. A., Wilson, K. T., Lau, K. S., Lauffenburger, D. A., 2020. An interspecies translation model implicates integrin signaling in infliximab-resistant inflammatory bowel disease. Sci Signal 13

Brubaker, D. K., Proctor, E. A., Haigis, K. M., Lauffenburger, D. A., 2019. Computational translation of genomic responses from experimental model systems to humans. PLoS Comput Biol 15, e1006286.

Cai, M., Hao Nguyen, C., Mamitsuka, H., Li, L., 2021. XGSEA: CROSS-species gene set enrichment analysis via domain adaptation. Brief Bioinform 22, bbaa406.

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., Leek, J. T., 2017. Reproducible RNA-seq analysis using recount2. Nat Biotechnol 35, 319-321.

Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D., Saez-Rodriguez, J., 2019. Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res 29, 1363-1375.

Gene Ontology, C., 2021. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 49, D325-D334.

Holland, C. H., Szalai, B., Saez-Rodriguez, J., 2020. Transfer of regulatory knowledge from human to mouse for functional genomics analysis. Biochim Biophys Acta Gene Regul Mech 1863, 194431.

Holzscheck, N., Falckenhayn, C., Sohle, J., Kristof, B., Siegner, R., Werner, A., Schossow, J., Jurgens, C., Volzke, H., Wenck, H., Winnefeld, M., Gronniger, E., Kaderali, L., 2021. Modeling transcriptomic age using knowledge-primed artificial neural networks. NPJ Aging Mech Dis 7, 15.

Kuenzi, B. M., Park, J., Fong, S. H., Sanchez, K. S., Lee, J., Kreisberg, J. F., Ma, J., Ideker, T., 2020. Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. Cancer Cell 38, 672-684 e6.

Lotfollahi, M., Naghipourfar, M., Theis, F. J., Wolf, F. A., 2020. Conditional out-of-distribution generation for unpaired data using transfer VAE. Bioinformatics 36, i610-i617.

Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C., Chikina, M., 2019. Pathway-level information extractor (PLIER) for gene expression data. Nat Methods 16, 607-610.

Mourragui, S., Loog, M., van de Wiel, M. A., Reinders, M. J. T., Wessels, L. F. A., 2019. PRECISE: a domain adaptation approach to transfer predictors of drug response from pre-clinical models to tumors. Bioinformatics 35, i510-i519.

Normand, R., Du, W., Briller, M., Gaujoux, R., Starosvetsky, E., Ziv-Kenet, A., Shalev-Malul, G., Tibshirani, R. J., Shen-Orr, S. S., 2018. Found In Translation: a machine learning model for mouse-to-human inference. Nat Methods 15, 1067-1073.

Pan, S. J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345-1359.

Schubert, M., Klinger, B., Klunemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M. J., Bluthgen, N., Saez-Rodriguez, J., 2018. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun 9, 20.

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., Richards, D. R., McDonald-Smith, G. P., Gao, H., Hennessy, L., Finnerty, C. C., Lopez, C. M., Honari, S., Moore, E. E., Minei, J. P., Cuschieri, J., Bankey, P. E., Johnson, J. L., Sperry, J., Nathens, A. B., Billiar, T. R., West, M. A., Jeschke, M. G., Klein, M. B., Gamelli, R. L., Gibran, N. S., Brownstein, B. H., Miller-Graziano, C., Calvano, S. E., Mason, P. H., Cobb, J. P., Rahme, L. G., Lowry, S. F., Maier, R. V., Moldawer, L. L., Herndon, D. N., Davis, R. W., Xiao, W., Tompkins, R. G., Inflammation, Host Response to Injury, L. S. C. R. P., 2013. Genomic responses in mouse models poorly mimic human inflammatory diseases. Proc Natl Acad Sci U S A 110, 3507-12.

Sharifi-Noghabi, H., Peng, S., Zolotareva, O., Collins, C. C., Ester, M., 2020. AITL: Adversarial Inductive Transfer Learning with input and output space adaptation for pharmacogenomics. Bioinformatics 36, i380-i388.

Takao, K., Miyakawa, T., 2015. Genomic responses in mouse models greatly mimic human inflammatory diseases. Proc Natl Acad Sci U S A 112, 1167-72.

Taroni, J. N., Grayson, P. C., Hu, Q., Eddy, S., Kretzler, M., Merkel, P. A., Greene, C. S., 2019. MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease. Cell Syst 8, 380-394 e4.