NEURAL NETS ARE NOT ALL YOU NEED: EVALUATING THE EFFECTS OF DEEP LEARNING ON TRANSCRIPTOMIC ANALYSIS

Benjamin J. Heil

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the degree of Doctor of Philosophy

2023

Supervisor of Dissertation:

Casey S. Greene

Director of the Center for Health AI

Co-Supervisor of Dissertation:

John H. Holmes

Professor of Medical Informatics in Epidemiology

Graduate Group Chairperson:

Benjamin F. Voight

Associate Professor of Pharmacology

Dissertation Committee:

Marylyn D. Ritchie (Chair), Director, Institute For Biomedical Informatics

Russ B. Altman, Kenneth Fong Professor of Bioengineering, Stanford University

Konrad Kording, Professor of Neuroscience

Kai Wang, Professor of Pathology and Laboratory Medicine

NEURAL NETS ARE NOT ALL YOU NEED: EVALUATING THE EFFECTS OF DEEP LEARNING ON TRANSCRIPTOMIC ANALYSIS

To Sydney, Mom, and Dad

ACKNOWLEDGEMENT

I would not have reached this point without the support of many people. First I would like to thank my mentor Casey Greene for helping me grow from a first-year grad student with aspirations of diagnosing all human disease with a cleverly designed model to a wisened (or maybe wizened) PhD candidate who believes that data is paramount. I still remember reading papers as an undergrad trying to better understand what was going on at the intersection of computational biology and machine learning and wondering "What is the University of Pennsylvania and why does this Casey guy's papers keep showing up in my searches?" Thank you to my thesis committee: Marylyn Ritchie, Russ Altman, Konrad Kording, and Kai Wang. Your feedback has helped keep my research from going off the rails. Thank you as well to Greenelab members past and present. From grilling me to help prepare for prelims, to going on adventures with me in Colorado, to giving me tips on where to find free food, you've all helped me to better understand science and what it means to be a scientist. Thanks to Shuo Zhang and Liz Heller for collaborating with me on MousiPLIER, the project would not have been possible without you. I would also like to thank John Holmes for agreeing to be my advisor at Penn when Greenelab moved west to Colorado. In addition, I'd like to thank the GCB administration, especially Maureen Kirsch, Anne-Cara Apple, and Ben Voigt. You all do a good job of looking after students and making sure we don't fall through the cracks due to conditions beyond our control.

I've also been helped through grad school by many people outside of academia. Thanks Mom, Dad, Nana, Mary, Wes, and Sujin, your support has meant a lot even if you don't always understand what I'm talking about. Thanks as well to Rachel Ungar, and sorry that we didn't get an opportunity to collaborate (yet?) If you hadn't teamed up with me in

the audacious plan to get internships at the NIH after sophomore year, I wouldn't be where I am today. Thanks to my friends in Philly and in Texas for convincing me to get outside the lab and have fun on occasion, and for giving conflicting advice on whether or not I should drop out of grad school. Finally, thank you Sydney for helping me see that there is a world going on outside of the small bubble I interact with on a daily basis. I know that living with a PhD candidate has been frustrating at times, especially as I've gotten closer to defending and therefore progressively less interesting. I'm tempted to come up with something witty to write here, but you'd probably prefer sincerity, so: thank you.

ABSTRACT

NEURAL NETS ARE NOT ALL YOU NEED: EVALUATING THE EFFECTS OF DEEP LEARNING ON TRANSCRIPTOMIC ANALYSIS

Benjamin J. Heil

Casey S. Greene

John H. Holmes

Technologies for quantifying biology have undergone significant advances in the past few decades, leading to datasets rapidly increasing in size and complexity. At the same time, deep learning models have gone from a curiosity to a massive field of research, with their advancements spilling over into other fields. Machine learning is not new to computational biology, as machine learning models have been used frequently in the field to account for the aforementioned size and complexity of the data. This dissertation asks whether the paradigm shift in machine learning that has led to the rise of deep learning models is causing a paradigm shift in computational biology. To answer this question, we begin with chapter 1, which gives background information helpful for understanding the main thesis chapters. We then move to chapter 2, which discusses standards necessary to ensure that research done with deep learning is reproducible. We continue to chapter 3, where we find that deep learning models may not be helpful in analyzing expression data. In chapters 4 and 5 we demonstrate that classical machine learning methods already allow scientists to uncover knowledge from large datasets. Then in chapter 6 we conclude by discussing the implications of the previous chapters and their potential future directions. Ultimately we find that while deep learning models are useful to various subfields of computational biology, they have yet to lead to a paradigm shift.

Table of Contents

List of Tables

List of Illustrations

Chapter 1: Background

This chapter was prepared for this dissertation to provide background information and context for the dissertation as a whole. Parts of the “Applications of Machine Learning in Transcriptomics” were written for the class GCB752.

Contributions:

I was the sole author for this chapter. Some of the “Applications of Machine Learning in Transcriptomics” section was edited based on feedback from Kara Maxwell.

Introduction

As computational biologists, we live in exciting times. Beginning with the Human Genome Project [1], advancements in technologies for biological quantification have generated data with a scale and granularity previously unimaginable [2,3,4].

Concurrently with the skyrocketing amounts of data, the advent of deep learning has generated methods designed to make sense of large, complex datasets. These methods have led to a paradigm shift [5] in machine learning, creating new possibilities in many fields and surfacing new phenomena unexplained by classical machine learning theory [6,7,8,9].

The field of computational biology has long used machine learning methods, as they help cope with the scale of the data being generated. Accordingly, problem domains in computational biology that map well to existing research in deep learning have adopted or developed deep learning models and seen great advances [10,11]. Previous applications of classical machine learning to the field of transcriptomics have been successful. Two of the scientists who wrote the book [12] on machine learning have even

written papers [13,14,15,16] analyzing gene expression. However, the data itself is not well-suited to deep learning methods.

This dissertation explores whether the paradigm shift in machine learning will spill over to transcriptomics. That is to say, have deep learning techniques fundamentally changed transcriptomics, or are they incremental improvements over existing methods? Our thesis is that while deep learning provides valuable tools for analyzing biological datasets, it does not necessarily change the field on a fundamental level.

We begin with chapter 1, which gives background information on previous research for the other thesis chapters. We then move to chapter 2, which discusses standards necessary to ensure that research done with deep learning is reproducible. We continue to Chapter 3, where we find that deep learning models may not be helpful in analyzing expression data. In chapters 4 and 5 we demonstrate that classical machine learning methods already allow scientists to uncover knowledge from large datasets. Finally, in chapter 6 we conclude by discussing the implications of the previous chapters and their potential future directions.

Applications of machine learning in transcriptomics

The human transcriptome provides a rich source of information about both healthy and disease states. Not only is gene expression information useful for learning novel biological phenomena, it can also be used to diagnose and predict diseases. These predictions have become more powerful in recent years as the field of machine learning has developed more methods. In this section we review machine learning methods applied to predict various phenotypes from gene expression, with a focus on the challenges in the field and what is being done to overcome them. We close the review

with potential areas for future research, as well as our perspectives on the strengths and weaknesses of supervised learning for phenotype prediction in particular.

Introduction

Over the past few decades a number of tools for measuring gene expression have been developed. As proteomics is currently difficult to do at a large scale, gene expression quantification methods are our best way to measure cells' internal states. While this wealth of information is promising, gene expression data is more difficult to work with than one might think. The high dimensionality and instrument-driven variation require sophisticated techniques to separate the signal from the noise.

One such class of techniques is the set of methods from machine learning. Machine learning methods depend on the assumption that there are patterns in data that can be learned to make predictions about future data. Luckily, different people respond to the same disease in similar ways (for some diseases). Learning genes that indicate an inflammatory response, for example, can help a machine learning model learn the difference between healthy and diseased expression samples.

There are many varieties of machine learning algorithms, so the scope of this paper is limited to analysis of supervised machine learning methods for phenotype prediction. Supervised machine learning is a paradigm where the model attempts to predict labels. For example, a model that predicts whether someone has lupus based on their gene expression data [17] is a supervised learning model. In contrast, techniques for grouping data together without having phenotype labels are called unsupervised methods. While these methods are also commonly used in computational biology [18,19,20], we will not be discussing them here.

The purpose of this review is to explain and analyze the various approaches that are used to predict phenotypes. Each section of the review is centered around one of the challenges ubiquitous in using supervised machine learning techniques on gene expression data. We hope to explain what has been tried and what the consensus for handling the challenge, if one exists. The review will conclude with a section outlining promising new methods and areas where further study is needed.

If the field succeeds in addressing all the challenges, the payoffs will be substantial. Being able to predict and diagnose diseases from whole blood gene expression is particularly interesting. With sufficiently advanced analysis, invasive cancer biopsies might be able to be replaced with simple blood draws [20]. If not, there are already diagnostics that predict various cancer aspects from biopsy gene expression [21]. It may also be possible to diagnose common diseases based on blood gene expression [22,23,24,25], or even rare ones [26].

The techniques for measuring gene expression and for analyzing it have changed dramatically over the past few decades. This section aims to explain what some of those changes are and how they affect phenotype prediction.

Gene expression

Gene expression measurement methods have three main categories. This first to be created is the gene expression microarray. In a microarray, RNA is reverse transcribed to cDNA, labeled with fluorescent markers, then hybridized to probes corresponding to parts of genes. The amount of fluorescence is then quantified to give the relative amount of gene expression for each gene. While early microarrays had fewer genes and gene probes [27], more modern ones measure tens of thousands of genes [28].

While microarrays are useful, decreases in the price of genetic sequencing have made bulk RNA sequencing (RNA-seq) more common. In RNA-seq, cDNA molecules are sequenced directly after being reverse transcribed from mRNA. These cDNA fragments are then aligned against a reference exome to determine which gene, if any, each fragment maps to. The output of the bulk RNA-seq pipeline is a list of genes and their corresponding read counts. While there is not gene probe bias like in microarrays, RNA-seq has its own patterns of bias based on gene lengths and expression levels [29]. Bulk RNA-seq is also unable to resolve heterogeneous populations of cells, as it measures the average gene expression of all of cells in the sample.

Fairly recently a new method was developed called single-cell RNA sequencing. True to its name, single-cell sequencing allows gene expression to be measured at the individual cell level. This increase in precision is accompanied by an increase in data sparsity though, as genes expressed infrequently or at low levels may not be detected. The sparsity of single-cell data has led to a number of interesting methods, but as we worked with bulk RNA-sequencing single-cell papers will largely be absent from this review.

Machine Learning

Machine learning has undergone a paradigm shift in the past decade, beginning with the publication of the AlexNet paper in 2012 [30]. For decades random forests and support vector machines were the most widely used models in machine learning. This changed dramatically when the AlexNet paper showed that neural networks could vastly outperform traditional methods in some domains [30]. The deep learning revolution quickly followed, with deep neural networks becoming the state of the art in any problem with enough data [11,31,32,33].

The implications of the deep learning revolution on this paper are twofold. First, almost all papers before 2014 use traditional machine learning methods, while many papers after use deep learning methods. Second, deep neural networks' capacity to overfit the data and fail to generalize to outside data are vast. We'll show throughout the review various mistakes authors make because they don't fully understand the failure states of neural networks and how to avoid them.

Dimensionality Reduction

The most obvious challenge in working with gene expression data is its high dimensionality. That is to say that the number of features (genes) in a dataset is typically greater than the number of samples. It is common for an analysis to have tens of thousands of genes, but only hundreds (or tens) of samples. Because even simple models struggle under such circumstances, it is necessary to find a representation of the data that uses fewer dimensions.

In the traditional machine learning paradigm, this is done via manual or heuristic feature selection methods. Such methods tend to use a criterion like mutual information to select a subset of genes for the analysis [34]. In one of the earliest papers in this review, Li et al. try a eight different methods from statistics and machine learning to see if any one in particular outperformed the others [35]. Ultimately they found that no individual method rose to the top, and that the performance of different methods varies depending on the problem.

A number of other papers since then have also used manual methods. Grewal et al. chose a subset of genes from COSMIC [36] for training, but found that their model performed better when using all genes instead of just a subset [37]. Chen et al. used a

different gene set. They selected the LINCS 1000 gene set [38] for an imputation method, as the LINCS landmark genes are highly correlated with the genes they were trying to impute [39].

Gene subsets can be based on prior knowledge of gene regulatory networks as well [40,41]. While very interpretable, these methods do not necessarily lead to increased performance in phenotype predictions [42]. However, such methods can be useful in their own right. PLIER (and the associated MultiPLIER framework) use prior knowledge genes to guide the latent variables learned by a matrix factorization technique [43,44]. The resulting latent variables can then be used in differential expression analyses in lieu of raw gene counts, allowing dimensionality reduction while guiding the learned variables towards biological relevance.

Selecting gene subsets via a heuristic or a machine learning model is also popular. Sevakula et al. use decision stumps to select features then use a stacked autoencoder-type architecture to further compress the representation [45]. Xiao et al. did something similar where they reduced the data to only genes were differentially expressed between their conditions of interest, then used a stacked autoencoder architecture [46]. Instead of looking at raw differential expression, Dhruba et al. used another subsetting method called ReliefF [47] to find the top 200 genes for their source and target dataset, then kept the intersection for use in their model [48]. More recently, Li et al. used a genetic algorithm for feature selection [49].

Not all papers use a subset of the original genes in their analysis, however. It is fairly common in recent years for authors to transform the data into a new lower dimensional space based on various metrics. This used to be done via principle component analysis (PCA), a method that performs a linear transformation to maximize the variance

explained by a reduced number of dimensions [50,51]. Now scientists typically use different types of autoencoders, which learn a nonlinear mapping from the original space to a space with fewer dimensions. Deepathology uses variational [52] and contractive [53] autoencoders in their model [54], while Danaee et al. used a stacked denoising autoencoder [55,56]. Both papers compared their autoencoder dimensionality reduction to that of PCA and found that it performed better. Danaee found that kernel PCA, a nonlinear version of PCA performed equivalently though.

It is also possible to use regularization methods to perform dimensionality reduction. While they do not influence the nominal dimensionality of the data, they reduce the effective dimensionality by putting constraints on the input data or the model. For example, SAUCIE uses an autoencoder structure, but combines it with a number of exotic regularization methods to further decrease the effective dimensionality of their data [57]. In DeepType, Chen et al. use a more conventional elastic net regularization [58] to induce sparsity in the first level of their network under the assumption that most genes' expression will not affect a cancer's subtype [20].

Ultimately, there is no clear consensus in which dimensionality reduction methods perform the best. Among the methods that transform the data there is a small amount of evidence that nonlinear transformations outperform linear ones, but only a few studies have tried both. Going forward, a systematic evaluation of gene selection and dimensionality reduction methods on a variety of problems could be a huge asset to the field.

Evaluating Model Performance

Validation is another important consideration in phenotype prediction. The gold standard of validation would be a knockout and rescue assay demonstrating that the predicted mechanism or expression relationship truly exists. Since machine learning models make predictions of nonlinear relationships between thousands of genes, however, such validation isn't feasible. Instead scientists evaluate their models' efficacy by testing their performance on data they didn't train them on. Test datasets can be built in different ways, assorting roughly into three tiers based on their external validity.

The most basic method is referred to as cross-validation. In cross-validation, the training data is split into a training and validation dataset. The model is trained on the training dataset, then its performance is measured on the validation dataset. Typically this is done with a process called five-fold cross-validation, where the process is repeated five times on five different ways of splitting up the training data. This method is common [48,49,56], but isn't really a rigorous evaluation. Because the same dataset is used for both selecting a model and measuring performance, the data can 'go stale' when you test several models [59]. In the extreme case, it is possible to get 100% accuracy by testing random prediction schemes on the data.

In order to keep data fresh, some researchers use a more rigorous method called a held out test set[45,54]. In the held out test set paradigm, a portion of the dataset is set aside and effectively put in a locked box until the end of the analysis. Once the model architecture, hyperparameters, and dimensionality reduction decisions are all made via cross-validation on the training data, the lock box can be opened and the data within used for evaluation. As the lock box data is only used once, it has no risk of becoming stale due to multiple testing. The only drawback to this method is that is depends on the

assumption that the data in the real world is distributed the same as the data in your training set.

The best (and most difficult) way to evaluate a model is by using an independent dataset. Ideally, an independent dataset is created by a different group or on a different expression quantification platform. For example, once their model was trained, Chen et al. evaluated their model on a dataset from GEO, a dataset from GTEx, and a cancer cell line [39]. It is also possible to use combinations of validation methods. In their paper Grewal et al. used a held-out section of their original data, then went on to evaluate their model in an independent dataset [37]. Similarly, Malta et al. used cross-validation initially, but then evaluated their model on an external microarray dataset to ensure their data wasn't stale [60]. Likewise, Deng at al. initially benchmark their model on various simulated data sets, but then go on to validate their model on real data [61].

Ultimately researchers work with what they have, and it's not always possible to acquire an independent dataset. That being said, it is always worth keeping the different tiers of external validity in mind when evaluating papers that use machine learning.

Transfer Learning

Transfer learning is a field of machine learning that uses information from outside of the training dataset to improve model performance. Techniques from the field of transfer learning are particularly useful in the domain of gene expression, because there are large databases like GEO and TCGA that contain data that may be useful in prediction tasks. In this section we'll focus on two types of transfer learning that are particularly useful: multitask learning and semi-supervised learning.

Multitask learning involves training a model on multiple problems in order to improve the model's performance on a problem of interest. As gene expression patterns can be shared across diseases [62,63], the extra data can help increase the model's power. For example, instead of training a model to learn one drug response at a time, Yuan et al. had better results predicting all the drugs in their dataset simultaneously [64]. Similarly, Deepathology predicts tissue type, disease, and miRNA expression simultaneously [54]. It is worth noting that multitask learning works best when using a deep learning model. When using standard machine learning it is necessary to perform some difficult data transformation to do classification on multiple classes [35].

Where supervised learning uses entirely labeled data, semi-supervised learning takes advantage of unlabeled data as well. The most popular way of doing semi-supervised learning is to use an autoencoder structure to initialize your model's weights. Where most models begin training with a randomly initialized set of weights, it is possible to initially train a neural network to create a compressed representation of the input data (an encoding). The weights that it learns in the process often turn out to be a better initialization when the labeled training data is finally brought in. There are a number of ways to perform the autoencoding step. Instead of training all the layers of the network simultaneously, it is possible to train one layer to create the encoding at a time [45,46]. This is referred to as a stacked autoencoder. One can also train the whole network at the same time, as Danaee et al do with their denoising autoencoder [56]. Not all methods are autoencoder-based though. Dhruba et al. develop their own semi-supervised learning process that teaches a model to learn a latent space between classes [48].

Deep Learning vs Classical ML

Recent years have seen a dramatic shift towards deep learning methods. It is not immediately clear, however, whether this is a good decision for problems without giant datasets. While some argue that deep learning is overrated and simpler models should be used instead [65,66], others find that deep learning outperforms even domain specific models [67,68].

Because it is unclear which type of model will perform best on which dataset, it is important to try both simple and complex models. In the Deepathology paper, Azarkahlili et al. found that their deep neural networks outperformed decision tree, KNN, random forest, logistic regression, and SVM models [54]. Likewise, in gene expression imputation, Chen et al. found that their neural network classifier outperformed linear regression in 99.97 percent of genes and k-nearest neighbors in all genes [39]. On the other hand, Grewal et al. tried multiple methods and found they work roughly the same [37]. They settled this by combining a few different models into an ensemble.

Due to technical considerations [60] or other reasons, some authors only evaluate a single model [46]. While this simplifies the analysis for their papers, it makes it unclear whether they could have done better with a different model. This is particularly important for authors who are using deep learning models, because simpler models tend to be much more interpretable.

In chapters 3 and 4, we apply machine learning models to transcriptomic data. Chapter 3 has us comparing linear and deep learning models and showing that the linear models perform at least as well as the neural networks. Chapter 4 continues the idea by demonstrating that classical machine learning can be used to great effect on gene expression data.

Citation indices

Over the past century quantifying the progress of science has become popular. Even before computers made it easy to collate information about publications, work had already begun to evaluate papers based on their number of citations [69]. There is even a book about it [70].

Determining the relative “impact” of different authors and journals is a perennial question when measuring science. One of the most commonly used metrics in this space is the h-index, which balances an author’s number of publications with the number of citations each receives [71]. However, the h-index is not a perfect metric [72] and has arguably become less useful in recent years [73]. Other metrics, like the g-index[74] and the i-10 index (https://scholar.google.com/), try to improve on the h-index by placing a higher weight on more highly cited papers.

There are metrics for comparing journals as well. The Journal Impact Factor [75] is the progenitor journal metric, evaluating journals based on how many citations the average paper in that journal has received over the past few years. Other measures use a more network-based approach to quantifying journals’ importance. The most common are Eigenfactor [76] and the SCImago Journal Rank (https://www.scimagojr.com/), which use variations on the PageRank algorithm to evaluate the importance of various journals.

Academic articles are arguably the main building blocks of scientific communication, so it makes sense to try to understand which ones are the most important. Citation count seems like an obvious choice, but differences in citation practices between fields [77] make it too crude a measure of impact. Instead, many other metrics have been developed to choose which papers to read.

Many of these methods work by analyzing the graph formed by treating articles as nodes and citations as edges. PageRank[78], one of the most influential methods for ranking nodes' importance in a graph, can also be applied to ranking papers [79]. It is not the only graph-based method, though. Other centrality calculation methods, such as betweenness centrality, would make sense to use but are prohibitively computationally expensive to run. Instead, methods like the disruption index [80] and its variants [81] are more often used.

Some lines of research try to quantify other desirable characteristics of papers. For example, Foster et al. claim to measure innovation by looking at papers that create new connections between known chemical entities [82]. Likewise, Wang et al. define novel papers as those that cite papers from unusual combinations of journals [83]. The Altmetric Attention Score (https://www.altmetric.com/) goes even further, measuring the attention on a paper from outside the standard academic channels.

These metrics do not stand alone, however. Much work has gone into improving the various methods by shoring up their weaknesses or normalizing them to make them more comparable across fields. The relative citation ratio makes citation counts comparable across fields by normalizing it according to other papers in its neighborhood of the citation network [84]. Similarly, the source-normalized impact per paper normalizes article citation counts based on the total number of citations in the whole field [85]. Several methods modify PageRank, such as Topical PageRank, which incorporates topic and journal prestige information into the PageRank calculation [86], and Vaccario et al.'s page and field rescaled PageRank, which accounts for differences between papers' ages and fields [87]. There are also several variants of the disruption index [81].

Of course, these methods only work with data to train and evaluate them on. We have come a long way from Garfield's "not unreasonable" proposal to aggregate one million citations manually [69]. These days we have several datasets with hundreds of millions to billions of references (https://www.webofknowledge.com, https://www.scopus.com [88]).

Quantifying science could be better, however. In addition to the shortcomings of individual methods [89,90,91], there are issues inherent to reducing the process of science to numbers. To quote Alfred Korzybski, "the map is not the territory." Metrics of science truly measure quantitative relationships like mean citation counts, despite purporting to reflect "impact," "disruption," or "novelty." If we forget that, we can mistake useful tools for arbiters of ground truth.

In chapter 5, we dive into one such shortcoming by demonstrating differences in article PageRanks between fields. There we argue that normalizing out field-specific differences obscures useful signal and propose new directions of research for future citation metrics.

Chapter 2: Reproducibility standards for machine learning in the life sciences

This chapter was originally published in Nature Methods as “Reproducibility standards for machine learning in the life sciences” by Benjamin J. Heil, Michael M. Hoffman, Florian Markowetz, Su-In Lee, Casey S. Greene, and Stephanie C. Hicks (https://doi.org/10.1038/s41592-021-01256-7).

Contributions:

C.S.G. was responsible for conceptualization. B.J.H. was responsible for project administration. B.J.H. and S.C.H. wrote the original draft of the manuscript; and B.J.H., S.C.H., M.M.H., S.L., F.M. and C.S.G. contributed to reviewing and editing.

Abstract

Establishing reproducibility expectations focused on data, models, and code will ensure that the life sciences community can trust machine learning analyses.

Introduction

The field of machine learning has grown tremendously within the past ten years. In the life sciences, machine learning models are being rapidly adopted because they are well suited to cope with the scale and complexity of biological data. There are drawbacks to using such models though. For example, machine learning models can be harder to interpret than simpler models, and this opacity can obscure learned biases. If we are going to use such models in the life sciences, we will need to trust them. Ultimately all science requires trust [92] — no scientist can reproduce the results from every paper

they read. The question, then, is how to ensure that machine learning analyses in the life sciences can be trusted.

One attempt at creating trustworthy analyses with machine learning models revolves around reporting analysis details such as hyperparameter values, model architectures, and data splitting procedures. Unfortunately, such reporting requirements are insufficient to make analyses trustworthy. Documenting implementation details without making data, models, and code publicly available and usable by other scientists does little to help future scientists attempting the same analyses and less to uncover biases. Authors can only report on biases they already know about, and without the data, models, and code, other scientists will be unable to discover issues post-hoc.

For machine learning models in the life sciences to become trusted, scientists must prioritize computational reproducibility [93]. Specifically, using published data, models, and code, other scientists must be able to obtain the same results as the original authors. With access to published data, models, and code, a researcher can confirm that a model functions and probe how the model functions. This means that using the published model a third party can examine for themselves the accuracy of reported results and biases in the model. Analyses and models that are reproducible by third parties can be examined in depth and, ultimately, become worthy of trust. To that end, the life science community should adopt norms and standards that underlie reproducible machine learning research.

The menu

While many regard the computational reproducibility of a work as a binary property, we prefer to think of it on a sliding scale [93] reflecting the time needed to reproduce.

Published works fall somewhere on this scale, which is bookended by “forever”, for a completely irreproducible work, and “zero”, for a work where one can automatically repeat the entire analysis with a single keystroke. Since it makes little sense to impose a single standard dividing work into “reproducible” and “irreproducible”, we instead propose a menu of three standards with varying degrees of rigor for computational reproducibility:

The bronze standard: the authors make the data, models, and code used in the analysis publicly available. The bronze standard is the minimal standard for reproducibility. Without data, models, and code, it is not possible to reproduce a work. The silver standard: in addition to meeting the bronze standard, (1) the dependencies of the analysis can be downloaded and installed in a single command, (2) key details for reproducing the work are documented, including the order in which to run the analysis scripts, the operating system used, and system resource requirements, and (3) all random components in the analysis are set to be deterministic. The silver standard is a midway point between minimal availability and full automation. Works that meet this standard will take much less time to reproduce than ones only meeting the bronze standard. The gold standard: the work meets the silver standard, and the authors make the analysis reproducible with a single command. The gold standard for reproducibility is full automation. When a work meets this standard, it will take little to no effort for a scientist to reproduce it.

While reporting has become a recent area of focus [94,95,96], excellent reporting is akin to a nutrition information panel. It describes information about a work, but is insufficient for reproducing the work. In the best case it provides a summary of what the researchers who conducted the analysis know about biases in the data, model limitations, and other

elements. It does not, however, provide enough information for someone to fully understand how the model came to be. For these reasons, concrete standards for ensuring reproducibility should be preferred over reporting requirements.

Bronze

Data

Data are a fundamental component of analyses. Without data, models can not be trained and analyses can not be reproduced. Moreover, biases and artifacts in the data that were missed by the authors cannot be discovered if the data are never made available. For the data in an analysis to be trusted, they must be published.

To that end, all datasets used in a publication should be made publicly available when their corresponding manuscript is first posted as a preprint or published by a peer-reviewed journal. Specifically, the raw form of all data used for the publication must be published. The way the bronze standard should be met depends on the data used. Authors should deposit new data in a specialist repository designed for that kind of data [97], when possible. For example, one may deposit gene expression data in the Gene Expression Omnibus [98] or microscopy images in the BioImage Archive [99]. If no specialist repository for that data type exists, one should instead use a generalist repository like Zenodo (https://zenodo.org) for datasets of up to 50 GB or Dryad (https://datadryad.org/) for datasets larger than 50GB. When researchers use existing datasets, they must include the code required to download and preprocess the data.

Models

Sharing trained models is another critical component for reproducibility. Even if the code for an analysis were perfectly reproducible and required no extra scientist-time to run, its

corresponding model would still need to be made publicly available. Requiring people who wish to use a method on their own data to re-train a model slows the progress of science, creates an unnecessary barrier to entry, and wastes the compute and effort of future researchers. Being unable to examine a model also makes trusting it difficult. Without access to the model it is hard to say whether the model fails to generalize to other datasets, fails to make fair decisions across demographic groups, or learns to make predictions based on artifacts in the data.

Because of the importance of sharing trained models, meeting the bronze standard of reproducibility requires that authors deposit trained weights for the models used to generate their results in a public repository. However, authors do not need to publish the weights for additional models from a hyperparameter sweep if one can reproduce the results without them. When a relevant specialist model zoo such as Kipoi [100] or Sfaira [101] exists, authors should deposit the models there. Otherwise, authors can deposit the models in a generalist repository such as Zenodo. Making models available solely on a non-archived website, such as a GitHub project, does not fulfill this requirement.

Source Code

From a reproducibility standpoint, a work's source code is as critical as its methods section. Source code contains implementation details that a future author is unlikely to replicate exactly from methods descriptions and reporting tables. These small deviations can lead to different behavior between the original work and the reproduced one. That is, of course, ignoring the huge burden of having to reimplement the entire analysis from scratch. For the computational components of a study, the code is likely a better description of the work than the methods section itself. As a result, computational papers

without published code should meet similar skepticism to papers without methods sections.

To meet the bronze standard, authors must deposit code in a third-party, archivable repository like Zenodo. This includes the code used in training, tuning, and testing models, creating figures, processing data, and generating the final results. One good way of meeting the bronze standard involves creating a GitHub project and archiving it in Zenodo. Doing so gives both the persistence of Zenodo required by scholarly literature and GitHub's resources for further development and use, such as the user support forum provided by GitHub Issues.

Silver

While it is possible to reproduce an analysis with only its data, models, and code, this task is by no means easy. Fortunately there are best practices from the field of software engineering that can make reproducing analyses easier by simplifying package management, recording analysis details, and controlling randomness.

One roadblock that appears when attempting to reproduce an analysis stems from differences in behavior between versions of packages used in the analysis. Analyses that once worked with specific dependency versions can stop working altogether with later versions. Guessing which versions one must use to reproduce an analysis—or even to get it to run at all—can feel like playing a game of "package Battleship". Proper use of dependency management tools like Packrat (https://rstudio.github.io/packrat/) and Conda (https://conda.io/) can eliminate these difficulties both for the authors and others seeking to build on the work by tracking which versions of packages are used.

Authors may also wish to consider containerization for managing dependencies. Container systems like Docker [102] allow authors to specify the system state in which to run their code more precisely than just versions of key software packages. Containerization provides better guarantees of reproducing a precise software environment, but this very fact can also facilitate code that won't tolerate even modest environment changes. That brittleness can make it more difficult for future researchers to build on the original analysis. Therefore, we recommend that authors using containers also ensure that their code works on the latest version of at least one operating system distribution. Furthermore, containers do not fully insulate the running environment from the underlying hardware. Authors expecting bit-for-bit reproducibility from their containers may find that GPU-accelerated code fails to yield identical results on other machines due to the presence of different hardware or drivers.

Knowing the steps to run an analysis is a crucial part of reproducing it, yet this knowledge is often not formally recorded. It takes far less time for the original authors to document factors such as the order of analysis components or information about the computers used than for a third-party analyst attempting to reproduce the work to determine that information on their own. Accordingly, the silver standard requires that authors record the order in which one should run their analysis components, the operating system version used to produce the work, and the time taken to run the code. Authors must also list the system resources that yielded that time, such as the model and number of CPUs and GPUs and the amount of CPU RAM and GPU RAM required. Authors may record the order in which one should run components (1) in a README file within the code repository, (2) by adding numbers to the beginning of each script's name to denote their order of execution, or (3) by providing a script to run them in order.

Authors must include details on the operating system, wall clock and CPU running time, and system resources used both within the body of the manuscript and in the README.

The last challenge of this section, randomness, is common in machine learning analyses. Dataset splitting, neural network initialization, and even some GPU-parallelized math used in model training all include elements of randomness. Because models' outputs depend heavily on these factors, the pseudorandom number generators used in analyses must be seeded to ensure consistent results. How the seeds are set depends on the language, though authors need to take special care when working with deep learning libraries. Current implementations often do not prioritize determinism, especially when accelerating operations on GPUs. However, some frameworks have options to mitigate nondeterministic operation (https://pytorch.org/docs/1.8.1/notes/randomness), and future versions may have fully deterministic operation (https://github.com/NVIDIA/framework-determinism). For now, the best way to account for this type of randomness is by publishing trained models. This nondeterminism is another reason why the minimal standard requires model publication—reproducing the model using data and code alone may prove impossible.

As it is difficult to evaluate the extent to which an analysis follows best practices, we provide three requirements that must be met to achieve the silver standard in reproducibility. First, future users must be able to download and install all software dependencies for the analysis with a single command. Second, the order in which the analysis scripts should be run and how to run them should be documented. Finally, any random elements within the analysis should be made deterministic.

Gold

The gold standard for reproducibility requires the entire analysis to be reproducible with a single command. Achieving this goal requires authors to automate all the steps of their analysis, including downloading data, preprocessing data, training models, producing output tables, and generating and annotating figures. Full automation stands in addition to tracking dependencies and making their data and code available. In short, by meeting the gold standard authors make the burden of reproducing their work as small as possible.

Workflow management software such as Snakemake [103] or Nextflow [104] streamline the work of meeting the gold standard. They enable authors to create a series of rules that run all the components in an analysis. While a simple shell script can also accomplish this goal, workflow management software provides a number of advantages without extra work from the authors. For example, workflow management software can make it easy to restart analyses after errors, parallelize analyses, and track the progress of an analysis as it runs.

Caveats

Privacy

Not all data can be publicly released. Some data contain personally identifiable information or are restricted by a data use agreement. In these cases data should be stored in a controlled access repository [105], but the use of controlled access should be explicitly approved by journals to prevent it from becoming another form of “data available upon request”.

Training models on private data also poses privacy challenges. Models trained with standard workflows can be attacked to extract training data [106]. Fortunately, model training methods designed to preserve privacy exist: techniques such as differential privacy [107] can help make models resistant to attacks seeking to uncover personally identifiable information, and can be applied with open source libraries such as Opacus (https://opacus.ai/). Researchers working on data with privacy constraints should employ these techniques as a routine practice.

When data cannot be shared, models must be shared to have any hope of computational reproducibility. If neither data nor models are published, the code is nearly useless, as it does not have anything to operate on. Future authors could perhaps replicate the study by recollecting data and regenerating the models, but they will not be able to evaluate the original analysis based on the published materials. When working on data with privacy restrictions, it is important for authors to use privacy preserving techniques for model training so that model release is not impeded. Studies with only models published will not be able to be fully reproduced, but there will at least be the possibility of testing the models' behavior on other datasets.

Compute-intensive analyses

Analyses can take a long time to run. In some cases they may take so long to run that it is infeasible for them to be reproduced by a different research group. In those cases, authors should store and publish intermediate outputs. Doing so allows other users to verify the final results even if they can not reproduce the entire pipeline. Workflow management systems, as mentioned in the gold standard section, make this partial reproduction straightforward by tracking intermediate outputs and using them to reproduce the final results automatically. Setting up a lightweight analysis demonstration,

such as a web app on a small dataset or a Colab notebook (https://research.google.com/colaboratory/) running a pretrained model, can also be helpful for giving users the ability to evaluate model behavior without using large amounts of compute.

Reproducibility of packages, libraries, and software products

The standards outlined in this paper focus on the computational reproducibility of analyses using machine learning. Standards for software designed for reuse, such as software packages and utilities, would have a broader scope and encompass more topics. In addition to our standards, such software should make use of unit testing, follow code style guidelines, have clear documentation [108], and ensure compatibility across major operating systems to meet the gold standard for this type of research product.

Conclusion

If we are to make machine learning research in the life sciences trustworthy, then we must make it computationally reproducible. Authors who strive to meet the bronze, silver, and gold standards will increase the reproducibility of machine learning analyses in the life sciences. These standards can also accelerate research in the field. In the status quo, there is no explicit reward for reproducible programming practices. As a result, authors can ostensibly minimize their own programming effort by using irreproducible programming practices and leaving future authors to make up the difference. In practice, irreproducible programming practices tend to decrease short-run effort for the authors, but increase effort in the long run on both the parts of the original authors and future reproducing authors. Implementing the standards in a way that rewards reproducible science (Box 1) helps avoid these long-run costs.

Ultimately, reproducibility in computational research is comparatively easy to experimental life science research. Computers are designed to perform the same tasks repeatedly with identical results. If we can not make purely computational analysis reproducible, how can we ever manage to make truly reproducible work in wet lab research with such variable factors as reagents, cell lines, and environmental conditions? If we want life science to lead the way in trustworthy, verifiable research, then setting standards for computational reproducibility is a good place to start.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-03948 to M.M.H.); Cancer Research UK (A19274 to F.M.); and the National Institutes of Health's National Institute of General Medical Sciences (R35 GM128638 to S.L.), the National Human Genome Research Institute (R00HG009007 to S.C.H. and R01HG010067 to C.S.G), and the National Cancer Institute of the National Institutes of Health (R01CA237170 to C.S.G.)

Author contributions

Conceptualization, C.S.G. Project administration, B.J.H. Writing — original draft, B.J.H., S.C.H. Writing — review & editing, B.J.H., S.C.H., M.M.H., S.L., F.M., C.S.G.

Ethics declarations

Competing interests: M.M.H. received an Nvidia GPU Grant.

Box 1

Journals Journals can enforce reproducibility standards as a condition of publication. The bronze standard should be the minimal standard, though some journals may wish to

differentiate themselves by setting higher standards. Such journals may require the silver or gold standards for all manuscripts, or for particular classes of articles such as those focused on analysis. If journals act as the enforcing body for reproducibility standards, they can verify that the standards are met by either requiring reviewers to report which standards the work meets or by including a special reproducibility reviewer to evaluate the work.

Badging A badge system that indicates the trustworthiness of work could incentivize scientists to progress to higher standards of reproducibility. Upon completing analyses, authors could submit their work to a badging organization that would then verify which standards of reproducibility their work met and assign a badge accordingly. Such an organization would likely operate in a similar way to the Bioconductor [109] package review process. Authors could then include the badge with a publication or preprint to tout the effort the authors put in to ensure their code was reproducible. Including these badges in biosketches or CVs would make it simple to demonstrate a researcher's track record of achieving high levels of reproducibility. This would provide a powerful signal to funding agencies and their reviewers that a researcher's strengths in reproducibility would maximize the results of the investment made in a project. Universities could also promote reproducibility by explicitly requiring a track record of reproducible research in faculty hiring, annual review, and promotion.

Reproducibility Collaborators Adding "reproducibility collaborators" to manuscripts would also provide another means to make analyses more reproducible. We envision a reproducibility collaborator as someone outside the primary authors' research groups who certifies that they were able to reproduce the results of the paper from only the data, models, code, and accompanying documentation. Such collaborators would currently fall

under the “validation” role in the CRediT Taxonomy (https://casrai.org/credit/), though it should be made clear that the reproducibility coauthor should not also be collaborating on the design or implementation of the analysis.

Table 1 - Reproducibility Standards

	Bronze	Silver	Gold
Data published and downloadable	x	x	x
Models published and downloadable	x	x	x
Source code published and downloadable	x	x	x
Dependencies set up in a single command		x	x
Key analysis details recorded		x	x
Analysis components set to deterministic		x	x
Entire analysis reproducible with a single command			x

Chapter 3: The Effect of Non-Linear Signal in Classification Problems using Gene Expression

This chapter has been preprinted at bioRxiv (https://www.biorxiv.org/content/10.1101/2022.06.22.497194v2), reviewed through Review Commons, and submitted for publication at PLOS Computational Biology as “The Effects of Nonlinear Signal on Expression-Based Prediction Performance” by Benjamin J. Heil, Jake Crawford, and Casey S. Greene.

Contributions: I designed and ran the experiments, created the figures, and wrote/edited the manuscript. Jake Crawford acted as the primary code reviewer, gave feedback and guidance on experiments, and edited the manuscript. Casey S. Greene gave feedback and guidance on experiments and edited the manuscript.

Abstract

Those building predictive models from transcriptomic data are faced with two conflicting perspectives. The first, based on the inherent high dimensionality of biological systems, supposes that complex non-linear models such as neural networks will better match complex biological systems. The second, imagining that complex systems will still be well predicted by simple dividing lines prefers linear models that are easier to interpret. We compare multi-layer neural networks and logistic regression across multiple prediction tasks on GTEx and Recount3 datasets and find evidence in favor of both possibilities. We verified the presence of non-linear signal when predicting tissue and metadata sex labels from expression data by removing the predictive linear signal with Limma, and showed the removal ablated the performance of linear methods but not non-linear ones. However, we also found that the presence of non-linear signal was not

necessarily sufficient for neural networks to outperform logistic regression. Our results demonstrate that while multi-layer neural networks may be useful for making predictions from gene expression data, including a linear baseline model is critical because while biological systems are high-dimensional, effective dividing lines for predictive models may not be.

Introduction

Transcriptomic data contains a wealth of information about biology. Gene expression-based models are already being used for subtyping cancer [110], predicting transplant rejections [111], and uncovering biases in public data [112]. In fact, both the capability of machine learning models [113] and the amount of transcriptomic data available [114,115] are increasing rapidly. It makes sense, then, that neural networks are frequently being used to build predictive models from transcriptomic data [54,116,117].

However, there are two conflicting ideas in the literature regarding the utility of non-linear models. One theory draws on prior biological understanding: the paths linking gene expression to phenotypes are complex [118,119], and non-linear models like neural networks should be more capable of learning that complexity. Unlike purely linear models such as logistic regression, non-linear models can learn non-linear decision boundaries to differentiate phenotypes. Accordingly, many have used non-linear models to learn representations useful for making predictions of phenotypes from gene expression [46,120,121].

The other supposes that even high-dimensional complex systems may have linear decision boundaries. This is supported empirically: linear models seem to do as well as or better than non-linear ones in many cases [122]. While papers of this sort are harder

to come by — perhaps scientists do not tend to write papers about how their deep learning model was worse than logistic regression — other complex biological problems have also seen linear models prove equivalent to non-linear ones [123,124].

We design experiments to ablate linear signal and find merit to both hypotheses. We construct a system of binary and multiclass classification problems on the GTEx and Recount3 compendia [125,126] that shows linear and non-linear models have similar accuracy on several prediction tasks. However, when we remove any linear separability from the data, we find non-linear models are still able to make useful predictions even when the linear models previously outperformed the non-linear ones. Given the unexpected nature of these findings, we evaluate independent tasks, examine different problem formulations, and verify our models' behavior with simulated data. The models' results are consistent across each setting, and the models themselves are comparable, as they use the same training and hyperparameter optimization processes [127].

In reconciling these two ostensibly conflicting theories, we confirm the importance of implementing and optimizing a linear baseline model before deploying a complex non-linear approach. While non-linear models may outperform simpler models at the limit of infinite data, they do not necessarily do so even when trained on the largest datasets publicly available today.

Results

Linear and non-linear models have similar performance in many tasks

We compared the performance of linear and non-linear models across multiple datasets and tasks (fig. 1 A). We examined using TPM-normalized RNA-seq data to predict tissue labels from GTEx [125], tissue labels from Recount3 [126], and metadata-derived sex

labels from Flynn et al. [128]. To avoid leakage between cross-validation folds, we placed entire studies into single folds (fig. 1 B). We evaluated models on subsampled datasets to determine the extent to which performance was affected by the amount of training data.

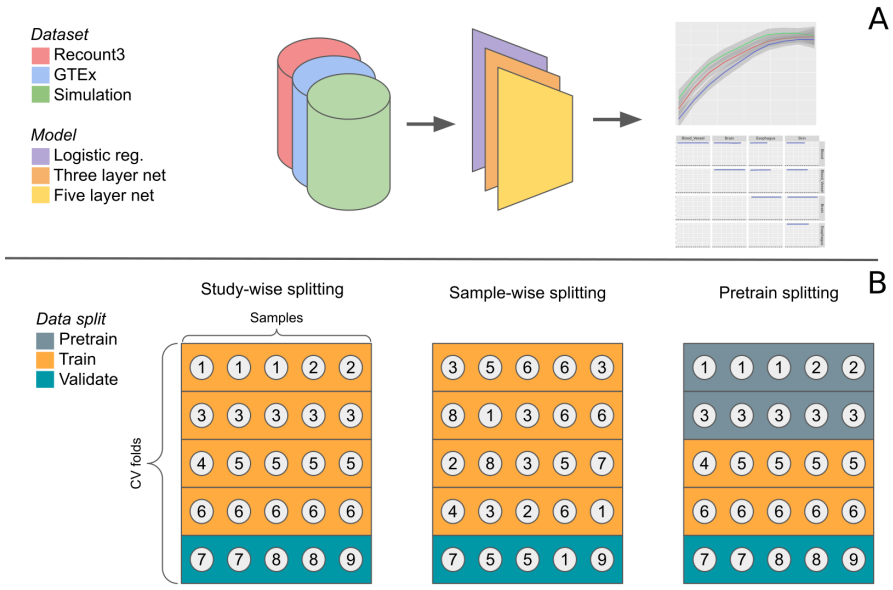

Figure 1: Schematic of the model analysis workflow. We evaluate three models on multiple classification problems in three datasets (A). We stratify the samples into cross-validation folds based on their study (in Recount3) or donor (in GTEx). We also evaluate the effects of sample-wise splitting and pretraining (B).

We used GTEx [125] to determine whether linear and non-linear models performed similarly on a well-characterized dataset with consistent experimental protocols across samples. We first trained our models to differentiate between tissue types on pairs of the five most common tissues in the dataset. Likely due to the clean nature of the data, all models were able to perform perfectly on these binary classification tasks (fig. 4 A). Because binary classification was unable to differentiate between models, we evaluated

the models on a more challenging task. We tested the models on their ability to perform multiclass classification on all 31 tissues present in the dataset. In the multitask setting, logistic regression slightly outperformed the five-layer neural network, which in turn slightly outperformed the three-layer net (fig. 2 A).

We then evaluated the same approaches in a dataset with very different characteristics: Sequence Read Archive [129] samples from Recount3 [126]. We compared the models' ability to differentiate between pairs of tissues (supp. fig. 4 B) and found their performance was roughly equivalent. We also evaluated the models' performance on a multiclass classification problem differentiating between the 21 most common tissues in the dataset. As in the GTEx setting, the logistic regression model outperformed the five-layer network, which outperformed the three-layer network (fig. 2 B).

To examine whether these results held in a problem domain other than tissue type prediction, we tested performance on metadata-derived sex labels (fig. 2 C), a task previously studied by Flynn et al. [128]. We used the same experimental setup as in our other binary prediction tasks to train the models, but rather than using tissue labels we used sex labels from Flynn et al. In this setting we found that while the models all performed similarly, the non-linear models tended to have a slight edge over the linear one.

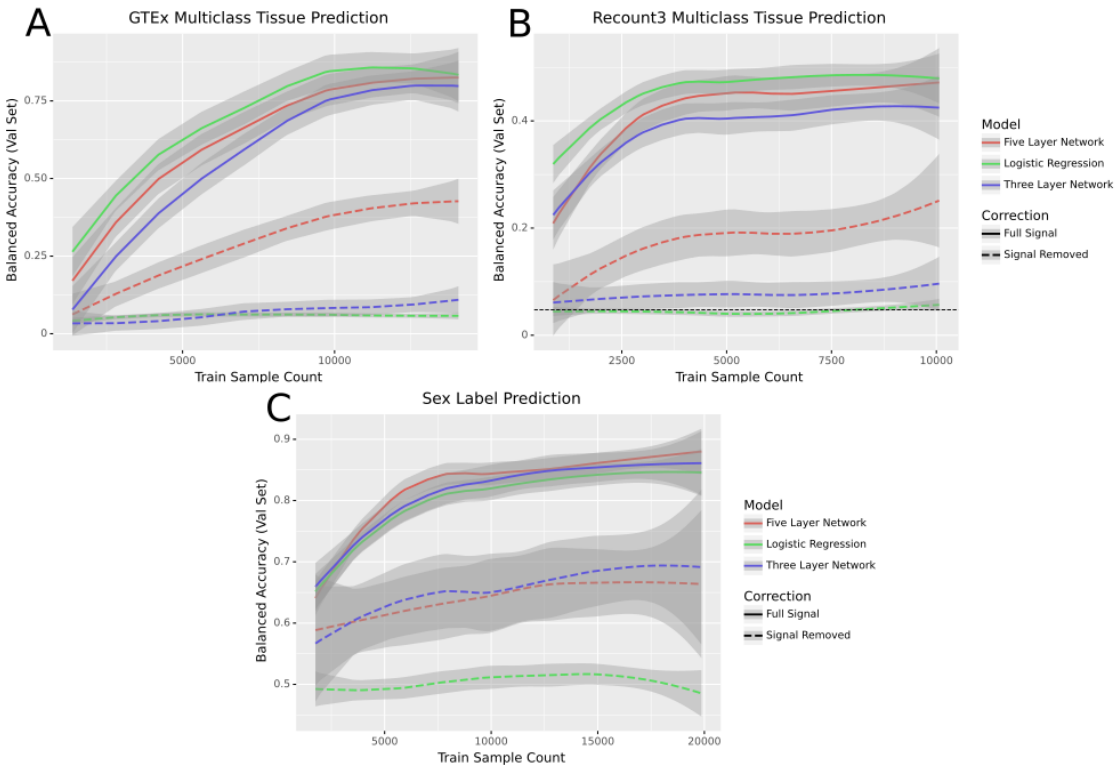

Figure 2: Performance of models across three classification tasks before and after signal removal. In each panel the loess curve and its 95% confidence interval are plotted based on points from three seeds, ten data subsets, and five folds of study-wise cross-validation (for a total of 150 points per model per panel). It is worth noting that "Sample Count" in these figures refers to the total number of RNA-seq samples, some of which share donors. As a result, the effective sample size may be lower than the sample count.

There is predictive non-linear signal in transcriptomic data

Our results to this point are consistent with a world where the predictive signal present in transcriptomic data is entirely linear. If that were the case, non-linear models like neural networks would fail to give any substantial advantage. However, based on past results

we expect there to be relevant non-linear biological signal [130]. To get a clearer idea of what that would look like, we simulated three datasets to better understand model performance for a variety of data generating processes. We created data with both linear and non-linear signal by generating two types of features: half of the features with a linear decision boundary between the simulated classes and half with a non-linear decision boundary (see Methods for more details). After training to classify the simulated dataset, all models effectively predicted the simulated classes. To determine whether or not there was non-linear signal, we then used Limma [131] to remove the linear signal associated with the endpoint being predicted. After removing the linear signal from the dataset, non-linear models correctly predicted classes, but logistic regression performed no better than random (fig. 3 B).

To confirm that non-linear signal was key to the performance of non-linear methods, we generated another simulated dataset consisting solely of features with a linear decision boundary between the classes. As before, all models were able to predict the different classes well. However, once the linear signal was removed, all models performed no better than random guessing (fig. 3 A). That the non-linear models only achieved baseline accuracy also indicated that the signal removal method was not injecting non-linear signal into data where non-linear signal did not exist.

We also trained the models on a dataset where all features were Gaussian noise as a negative control. As expected, the models all performed at baseline accuracy both before and after the signal removal process (fig. 3 C). This experiment supported our decision to perform signal removal on the training and validation sets separately. One potential failure state when using the signal removal method would be if it induced new

signal as it removed the old. Such a state can be seen when removing the linear signal in the full dataset (supp. fig. 5).

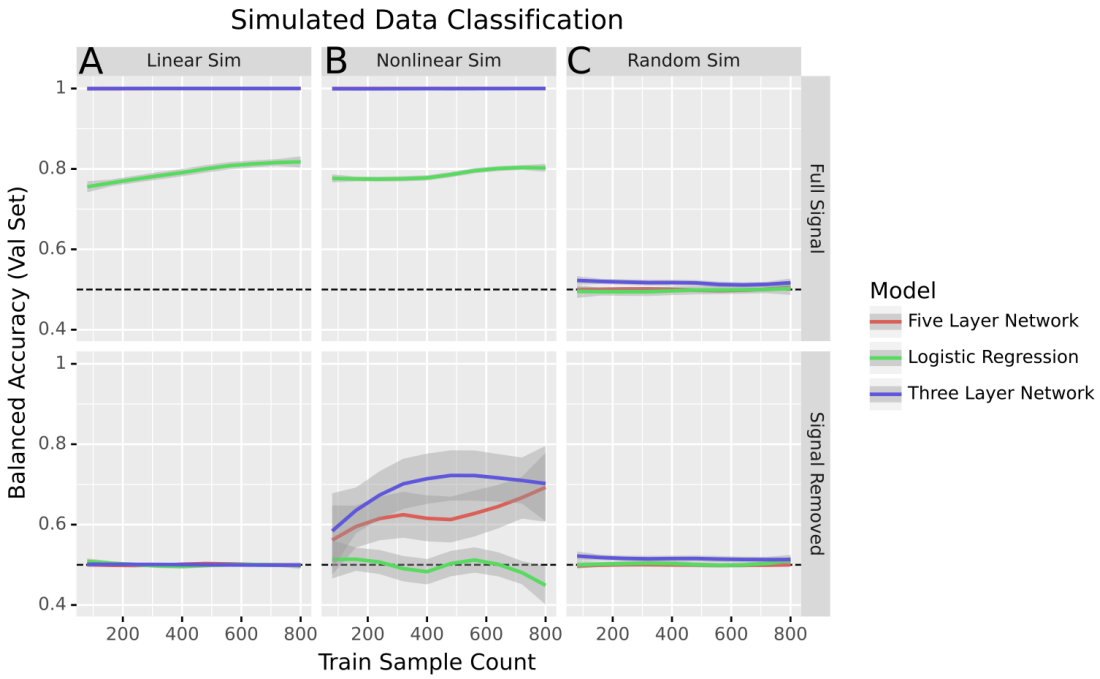

Figure 3: Performance of models in binary classification of simulated data before and after signal removal. Dotted lines indicate expected performance for a naive baseline classifier that predicts the most frequent class.

We next removed linear signal from GTEx and Recount3. We found that the neural nets performed better than the baseline while logistic regression did not (fig. 2, fig. 4). For multiclass problems logistic regression performed poorly while the non-linear models had performance that increased with an increase in data while remaining worse than before the linear signal was removed (fig. 2 A, B) Likewise, the sex label prediction task showed a marked difference between the neural networks and logistic regression: only the neural networks could learn from the data (fig. 2 C). In each of the settings, the models

performed less well when run on data with signal removed, indicating an increase in the problem's difficulty. Logistic regression, in particular, performed no better than random.

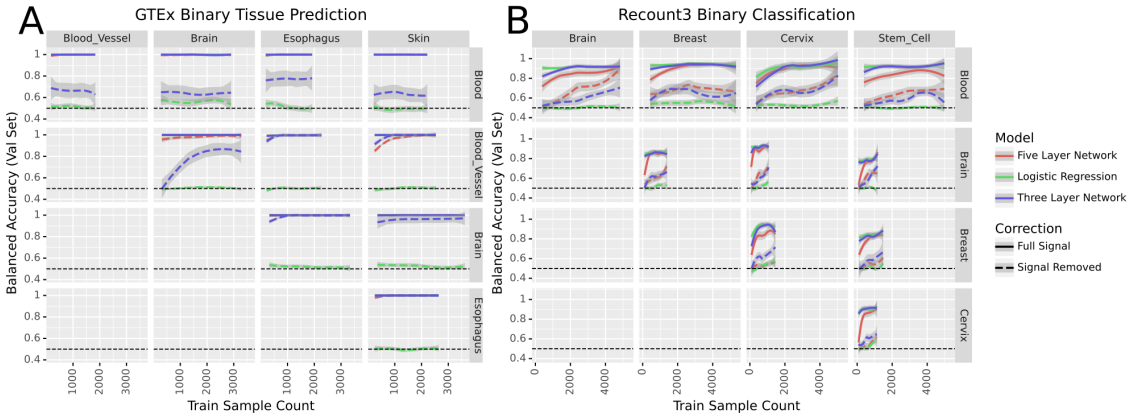

Figure 4: Models' performance across binary classification tasks before and after signal removal in the Recount and GTEx datasets.

To verify that our results were not an artifact of our decision to assign studies to cross-validation folds rather than samples, we compared the study-wise splitting that we used with an alternate method called sample-wise splitting. Sample-wise splitting (see Methods) is common in machine learning, but can leak information between the training and validation sets when samples are not independently and identically distributed among studies - a common feature of data in biology [132]. We found that sample-wise splitting induced substantial performance inflation (supp. fig. 7). The relative performance of each model stayed the same regardless of the data splitting technique, so the results observed were not dependent on the choice of splitting technique.

Another growing strategy in machine learning, especially on biological data where samples are limited, is training models on a general-purpose dataset and fine-tuning them on a dataset of interest. We examined the performance of models with and without

pretraining (supp. fig. 8). We split the Recount3 data into three sets: pretraining, training, and validation (fig. 1 B), then trained two identically initialized copies of each model. One was trained solely on the training data, while the other was trained on the pretraining data and fine-tuned on the training data. The pretrained models showed high performance even when trained with small amounts of data from the training set. However, the non-linear models did not have a greater performance gain from pretraining than logistic regression, and the balanced accuracy was similar across models.

Methods

Datasets

GTEx

We downloaded the 17,382 TPM-normalized samples of bulk RNA-seq expression data available from version 8 of GTEx. We zero-one standardized the data and retained the 5000 most variable genes. The tissue labels we used for the GTEx dataset were derived from the ‘SMTS’ column of the sample metadata file.

Recount3

We downloaded RNA-seq data from the Recount3 compendium [133] during the week of March 14, 2022. Before filtering, the dataset contained 317,258 samples, each containing 63,856 genes. To filter out single-cell data, we removed all samples with greater than 75 percent sparsity. We also removed all samples marked ‘scrna-seq’ by Recount3’s pattern matching method (stored in the metadata as ‘recount_pred.pattern.predict.type’). We then converted the data to transcripts per

kilobase million using gene lengths from BioMart [134] and performed standardization to scale each gene's range from zero to one. We kept the 5,000 most variable genes within the dataset.

We labeled samples with their corresponding tissues using the 'recount_pred.curated.tissue' field in the Recount3 metadata. These labels were based on manual curation by the Recount3 authors. A total of 20,324 samples in the dataset had corresponding tissue labels. Samples were also labeled with their corresponding sex using labels from Flynn et al. [112]. These labels were derived using pattern matching on metadata from the European Nucleotide Archive [135]. A total of 23,525 samples in our dataset had sex labels.

Data simulation

We generated three simulated datasets. The first dataset contained 1,000 samples of 5,000 features corresponding to two classes. Of those features, 2,500 contained linear signal. That is to say that the feature values corresponding to one class were drawn from a standard normal distribution, while the feature values corresponding to the other were drawn from a Gaussian with a mean of 6 and unit variance.

We generated the non-linear features similarly. The values for the non-linear features were drawn from a standard normal distribution for one class, while the second class had values drawn from either a mean six or negative six Gaussian with equal probability. These features are referred to as "non-linear" because two dividing lines are necessary to perfectly classify such data, while a linear classifier can only draw one such line per feature.

The second dataset was similar to the first dataset, but it consisted solely of 2,500 linear features. The final dataset contained only values drawn from a standard normal distribution regardless of class label.

Model architectures

We used three representative models to demonstrate the performance profiles of different model classes. The first was a linear model, ridge logistic regression, selected as a simple linear baseline to compare the non-linear models against. The next model was a three-layer fully-connected neural network with ReLU non-linearities [136] and hidden layers of size 2500 and 1250. This network served as a model of intermediate complexity: it was capable of learning non-linear decision boundaries, but not the more complex representations a deeper model might learn. Finally, we built a five-layer neural network to serve as a (somewhat) deep neural net. This model also used ReLU non-linearities, and had hidden layers of sizes 2500, 2500, 2500, and 1250. The five-layer network, while not particularly deep compared to, e.g., state of the art computer vision models, was still in the domain where more complex representations could be learned, and vanishing gradients had to be accounted for.

Model training

We trained our models via a maximum of 50 epochs of mini-batch stochastic gradient descent in PyTorch [137]. Our models minimized the cross-entropy loss using an Adam [138] optimizer. They also used inverse frequency weighting to avoid giving more weight to more common classes. To regularize the models, we used early stopping and gradient clipping during the training process. The only training differences between the models

were that the two neural nets used dropout [139] with a probability of 0.5, and the deeper network used batch normalization [140] to mitigate the vanishing gradient problem.

We ensured the results were deterministic by setting the Python, NumPy, and PyTorch random seeds for each run, as well as setting the PyTorch backends to deterministic and disabling the benchmark mode. The learning rate and weight decay hyperparameters for each model were selected via nested cross-validation over the training folds at runtime, and we tracked and recorded our model training progress using Neptune [141].

We also used Limma [131] to remove linear signal associated with tissues in the data. We ran the 'removeBatchEffect' function on the training and validation sets separately, using the tissue labels as batch labels. This function fits a linear model that learns to predict the training data from the batch labels, and uses that model to regress out the linear signal within the training data that is predictive of the batch labels.

Model Evaluation

In our analyses we used five-fold cross-validation with study-wise data splitting. In a study-wise split, the studies are randomly assigned to cross-validation folds such that all samples in a given study end up in a single fold (fig. 1 B).

Hardware

Our analyses were performed on an Ubuntu 18.04 machine and the Colorado Summit compute cluster. The desktop CPU used was an AMD Ryzen 7 3800xt processor with 16 cores and access to 64 GB of RAM, and the desktop GPU used was an Nvidia RTX 3090. The Summit cluster used Intel Xeon E5-2680 CPUs and NVidia Tesla K80 GPUs. From initiating data download to finishing all analyses and generating all figures, the full Snakemake [103] pipeline took around one month to run.

Recount3 tissue prediction

In the Recount3 setting, the multi-tissue classification analyses were trained on the 21 tissues (see Supp. Methods) that had at least ten studies in the dataset. Each model was trained to determine which of the 21 tissues a given expression sample corresponded to.

To address class imbalance, our models' performance was then measured based on the balanced accuracy across all classes. Unlike raw accuracy, balanced accuracy (the mean across all classes of the per-class recall) isn't predominantly determined by performance on the largest class in an imbalanced class setting. For example, in a binary classification setting with 9 instances of class A and 1 instance of class B, successfully predicting 8 of the 9 instances of class A and none of class B yields an accuracy of 0.8 and a balanced accuracy of 0.44.

The binary classification setting was similar to the multiclass one. The five tissues with the most studies (brain, blood, breast, stem cell, and cervix) were compared against each other pairwise. The expression used in this setting was the set of samples labeled as one of the two tissues being compared.

The data for both settings were split in a stratified manner based on their study.

GTEx classification

The multi-tissue classification analysis for GTEx used all 31 tissues. The multiclass and binary settings were formulated and evaluated in the same way as in the Recount3 data. However, rather than being split study-wise, the cross-validation splits were stratified according to the samples' donors.

Simulated data classification/sex prediction

The sex prediction and simulated data classification tasks were solely binary. Both settings used balanced accuracy, as in the Recount3 and GTEx problems.

Pretraining

When testing the effects of pretraining on the different model types, we split the data into three sets. Approximately forty percent of the data went into the pretraining set, forty percent went into the training set, and twenty percent went into the validation set. The data was split such that each study's samples were in only one of the three sets to simulate the real-world scenario where a model is trained on publicly available data and then fine-tuned on a dataset of interest.

To ensure the results were comparable, we made two copies of each model with the same weight initialization. The first copy was trained solely on the training data, while the second was trained on the pretraining data, then the training data. Both models were then evaluated on the validation set. This process was repeated four more times with different studies assigned to the pretraining, training, and validation sets.

Discussion and Conclusion

We performed a series of analyses to determine the relative performance of linear and non-linear models across multiple tasks. Consistent with previous papers [122,123], linear and non-linear models performed roughly equivalently in a number of tasks. That is to say that there are some tasks where linear models perform better, some tasks where non-linear models have better performance, and some tasks where both model types are equivalent.

However, when we removed all linear signal in the data, we found that residual non-linear signal remained. This was true in simulated data as well as GTEx and Recount3 data across several tasks. These results also held in altered problem settings, such as using a pretraining dataset before the training dataset and using sample-wise data splitting instead of study-wise splitting. This consistent presence of non-linear signal demonstrated that the similarity in performance across model types was not due to our problem domains having solely linear signals.

One limitation of our study is that the results likely do not hold in an infinite data setting. Deep learning models have been shown to solve complex problems in biology and tend to significantly outperform linear models when given enough data. However, we do not yet live in a world in which millions of well-annotated examples are available in many areas of biology. Our results are generated on some of the largest labeled expression datasets in existence (Recount3 and GTEx), but our tens of thousands of samples are far from the millions or billions used in deep learning research.

We are also unable to make claims about all problem domains or model classes. There are many potential transcriptomic prediction tasks and many datasets to perform them on. While we show that non-linear signal is not always helpful in tissue or sex prediction, and others have shown the same for various disease prediction tasks, there may be problems where non-linear signal is more important. It is also possible that other classes of models, be they simpler non-linear models or different neural network topologies, are more capable of taking advantage of the non-linear signal present in the data.

Ultimately, our results show that task-relevant non-linear signal in the data, which we confirm is present, does not necessarily lead non-linear models to outperform linear ones. Additionally, our results suggest that scientists making predictions from expression

data should always include simple linear models as a baseline to determine whether more complex models are warranted.

Code and Data Availability

The code, data, and model weights to reproduce this work can be found at https://github.com/greenelab/linear_signal. Our work meets the bronze standard of reproducibility [142] and fulfills aspects of the silver and gold standards including deterministic operation and an automated analysis pipeline.

Acknowledgements

We would like to thank Alexandra Lee and Jake Crawford for reviewing code that went into this project. We would also like to thank the past and present members of GreeneLab who gave feedback on this project during lab meetings. This work utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236), the University of Colorado Boulder, and Colorado State University.

Funding

This work was supported by grants from the National Institutes of Health's National Human Genome Research Institute (NHGRI) under award R01 HG010067 and the Gordon and Betty Moore Foundation (GBMF 4552) to CSG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary Materials

Results

Recount binary classification - Signal removal

While it's possible to remove signal in the full dataset or the train and validation sets independently, we decided to do the latter. We made this decision because we observed potential data leakage when removing signal from the entire dataset in one go (supp. fig. 5).

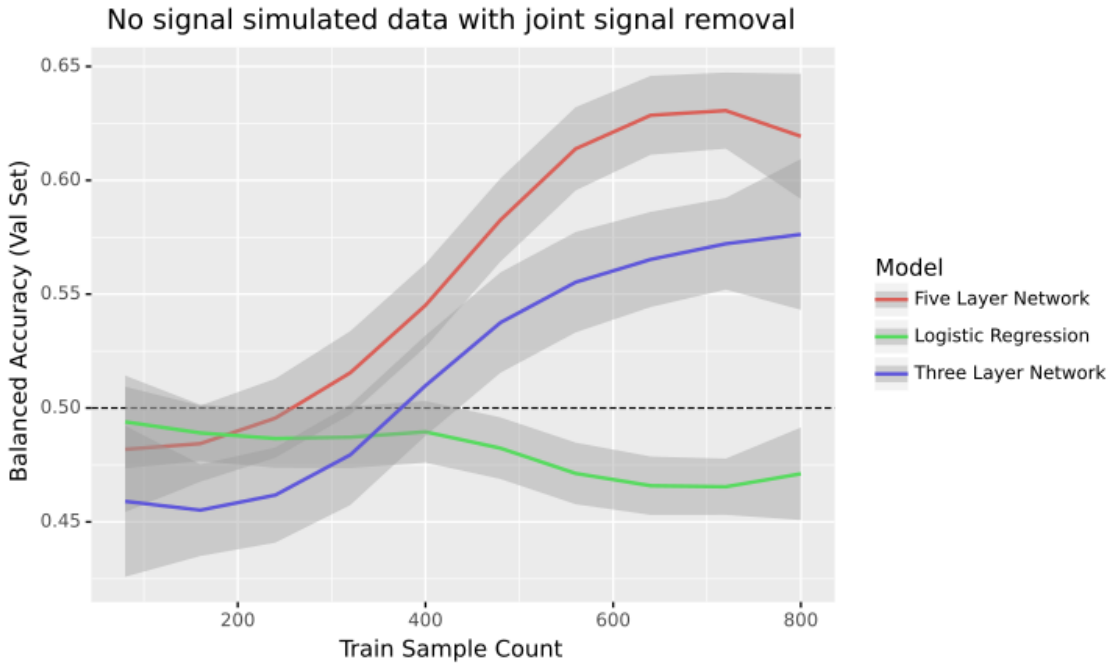

Figure 5:

Full dataset signal removal in a dataset without signal

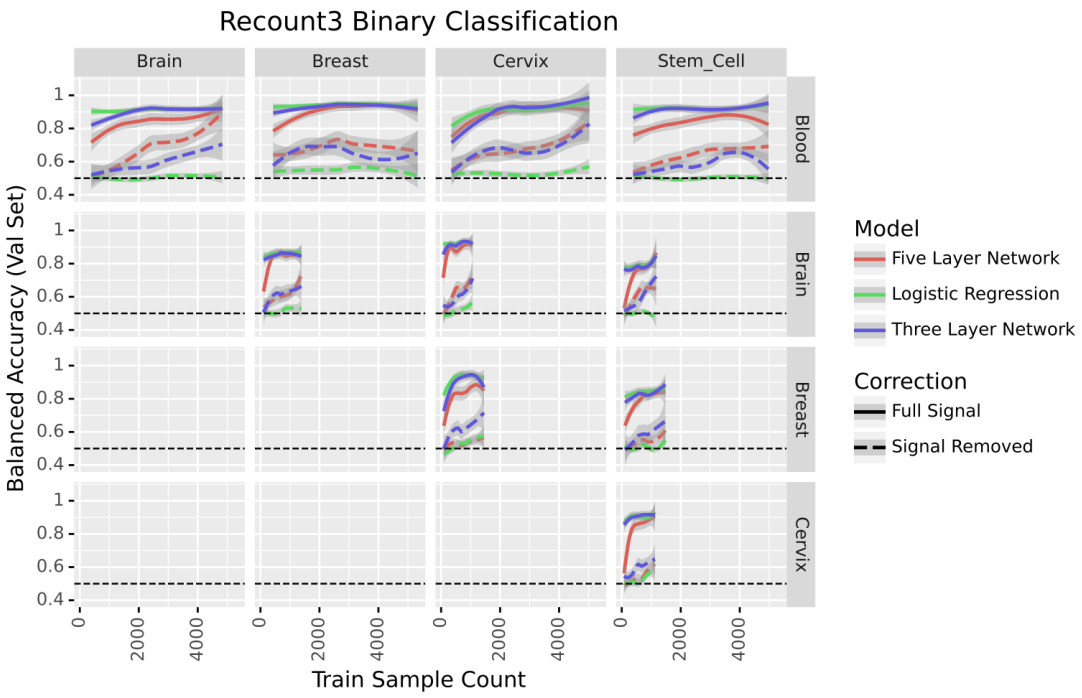

Figure 6:

Comparison of models' binary classification performance before and after removing linear signal

Samplewise splitting

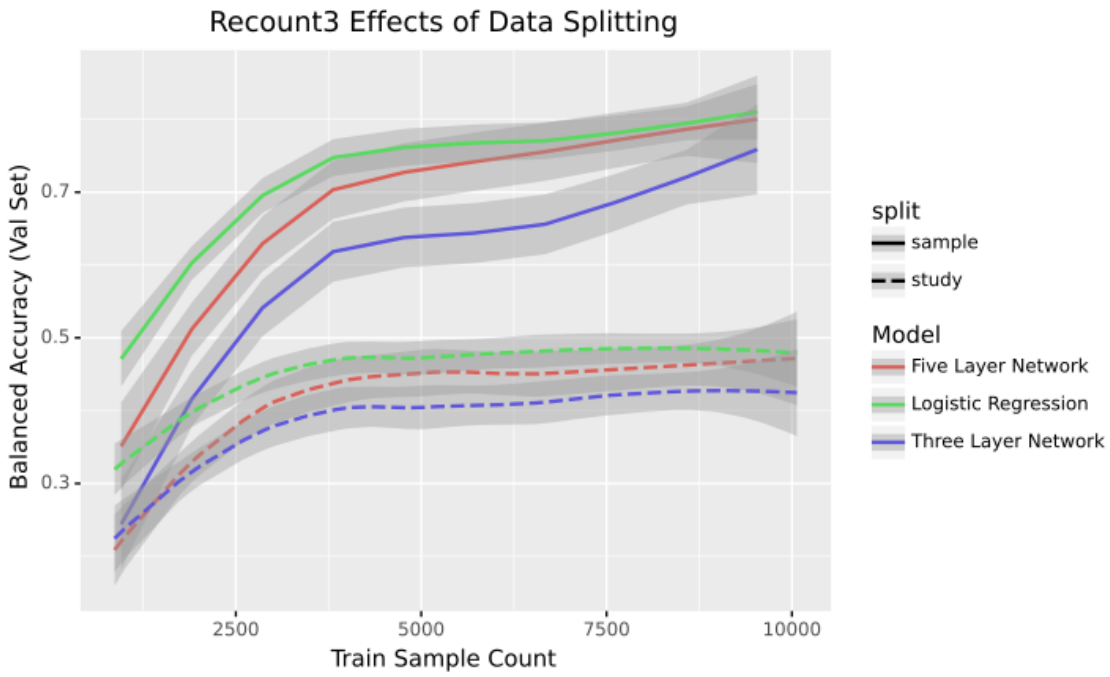

Figure 7:

Performance of Recount3 multiclass prediction with samplewise train/val splitting

Recount3 Pretraining

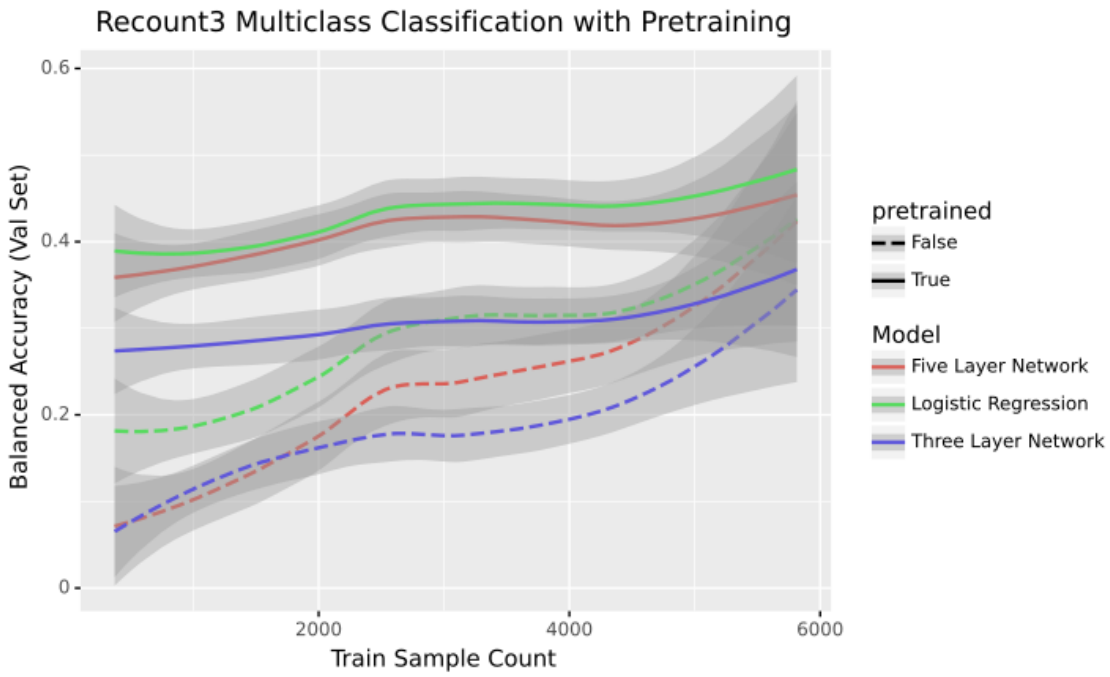

Figure 8:

Performance of Recount3 multiclass prediction with pretraining

Methods

Recount3 tissues used

The tissues used from Recount3 were blood, breast, stem cell, cervix, brain, kidney, umbilical cord, lung, epithelium, prostate, liver, heart, skin, colon, bone marrow, muscle, tonsil, blood vessel, spinal cord, testis, and placenta.

Chapter 4: MousiPLIER: the Largest and Most Murine PLIER Model Ever Trained

This chapter is from a manuscript in progress that is a collaboration between the Heller and Greene labs.

Contributions

I am a co-first author on the manuscript. I trained the MousiPLIER model, wrote and edited the manuscript, wrote code to make the results easier to use in Python, and ran the analyses leading to figures 9, 10, 12, and 15. Shuo Zhang is the other co-first author on the manuscript. He edited the manuscript, provided biological expertise useful in selecting latent variables and gene sets, and ran the analyses leading to figures 11, 13, and 14. Wayne Mao provided guidance for training PLIER in a large dataset. Casey S. Greene gave feedback and guidance on the experiments run, and is a co-corresponding author on the manuscript. Elizabeth A. Heller edited the manuscript, gave feedback and guidance on the experiments run, and helped interpret enriched genes in MousiPLIER latent variables.

Abstract

Differential expression analysis is widely used to learn from gene expression data. However, it suffers from the curse of dimensionality — RNA-sequencing experiments tend to have tens of thousands of genes with only tens or hundreds of samples. Many unsupervised learning models are designed to reduce dimensionality, and the PLIER model in particular fits expression data well. In this paper we describe the training of the

Mouse MultiPLIER (MousiPLIER) model, the first PLIER model trained on a mouse compendium and the PLIER model with the most training samples. We then go on to show that the model's latent variables contain biologically relevant information by finding enrichment for a striatally-associated latent variable in a mouse brain aging study and using the latent variable to uncover studies in the training data corresponding to mouse brain processes. This new model can assist mouse researchers in understanding the biological processes involved in their study and finding other studies in which these processes are relevant.

Introduction

A common way to gain knowledge from gene expression is by examining which genes are differentially expressed in an experiment [143,144,145,146]. For example, one can find which genes' expression values change in mouse tissue samples before and after some biological perturbation. This approach can be challenging, however, as studies analyzing gene expression tend to have few samples compared to the number of genes. The resulting lack of statistical power can be addressed by increasing the number of samples in an experiment, which is expensive. Alternatively, one can reduce the dimensionality or use information from outside the study.

Unsupervised learning does both. It is a category of methods from the field of machine learning that learn the structure of data without need for any biological labels denoting which experimental conditions are present. Such methods are well-suited for gene expression data, and are often used for tasks such as reducing the dimensionality of expression datasets [147,148,149], clustering samples [20,150], or learning shared expression patterns across experiments [18,120]. That being said, while unsupervised models are capable of using large amounts of unlabeled expression data, many of them

don’t explicitly encode prior biological knowledge to encourage the model to learn biological patterns over technical ones.

The Pathway-level information extractor (PLIER) models do [43]. They are built explicitly to work on expression data, and use matrix factorization to incorporate prior knowledge in the form of sets of genes specified by the user corresponding to biological pathways or cell type markers. PLIER models are also capable of learning diverse biological pathways from entire compendia of expression data and transferring that knowledge to smaller studies as seen in MultiPLIER [44]. However, PLIER models are largely trained on a single dataset rather than a compendium [151,152,153], and past MultiPLIER runs have only trained models with up to tens of thousands of samples [44,154].

In this paper we train a PLIER model on a compendium of mouse gene expression to convert the data into a series of values called “latent variables” that correspond to potentially biologically relevant combinations of genes. In doing so we train the largest (in terms of training data) PLIER model to date and the first one trained on a mouse compendium. We have named this model MousiPLIER, short for Mouse MultiPLIER. We demonstrate that not only is training such a model possible, it also surfaces interesting biology in a study of mouse brain aging. When looking at a novel view of the resulting data (k-means clusters in the latent variable space), we find that the microglia-associated latent variables from our study of interest also correspond to aging-related changes in the training data. Finally, we build a web server to allow others to visualize the results and find patterns in the data based on their own latent variables of interest. Going forward, this model and its associated web server will be a useful tool for better understanding mouse gene expression.

Results

MousiPLIER learns latent variables with ideal pathway level and gene-level sparsity

To determine the utility of a mouse multiPLIER model, we first trained the model. We then examined the latent variables that our model learned and found which were significantly enriched in a mouse brain aging study. Next, we looked deeper into the significant wild-type microglia-associated latent variables and determined that they were learning striatal signals potentially relevant to aging. Ultimately, we found that our trained PLIER model is able to uncover relevant signal in individual studies, and will be useful going forward in uncovering the biological processes present in mouse transcriptomic experiments. Taken together this study provides proof of concept that the mouse PLIER model can surface meaningful biological processes in mouse transcriptomic studies.

We trained MousiPLIER by using an on-disk PCA implementation to initialize PLIER, modifying the pipeline to work with mouse data, and using a high-mem compute node to manage the size of the matrix decomposition (see Methods for more details). The resulting model had 196 latent variables, where the per-latent variable distribution had an average of around 65% sparsity, which is to say that the latent variables tended to use only around 35% of the genes in the training data (Fig. 9). While many of the latent variables corresponded to no pathways, indicating signals in the training data not passed in as prior knowledge, those that remained corresponded to few pathways (Fig. 10). This pathway-level and gene-level sparsity is ideal, as it allows us to interrogate individual latent variables that correspond to a small number of biological functions.

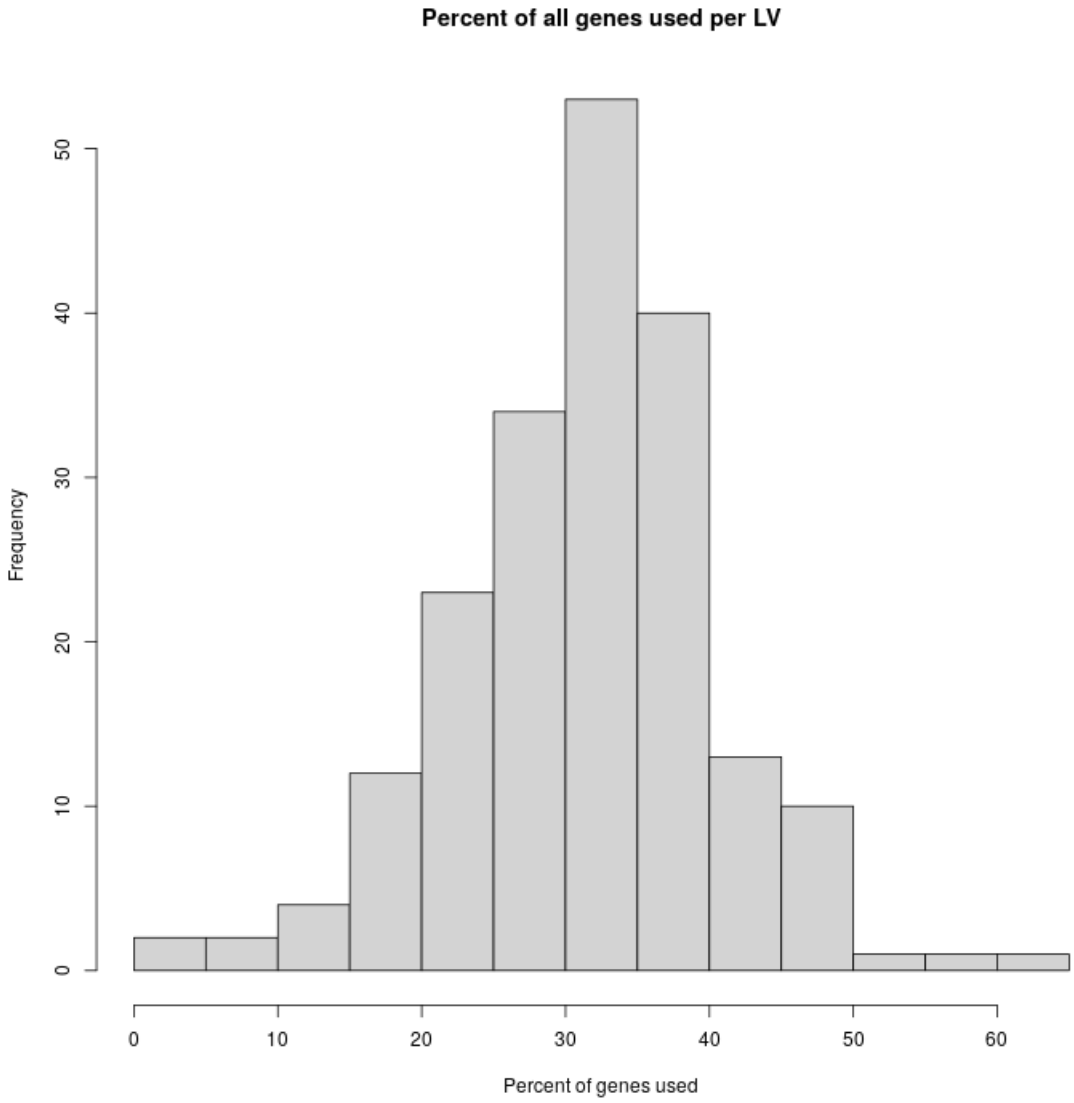

Figure 9: The distribution of the percentage of genes from the training set which had nonzero loadings for the latent variable.

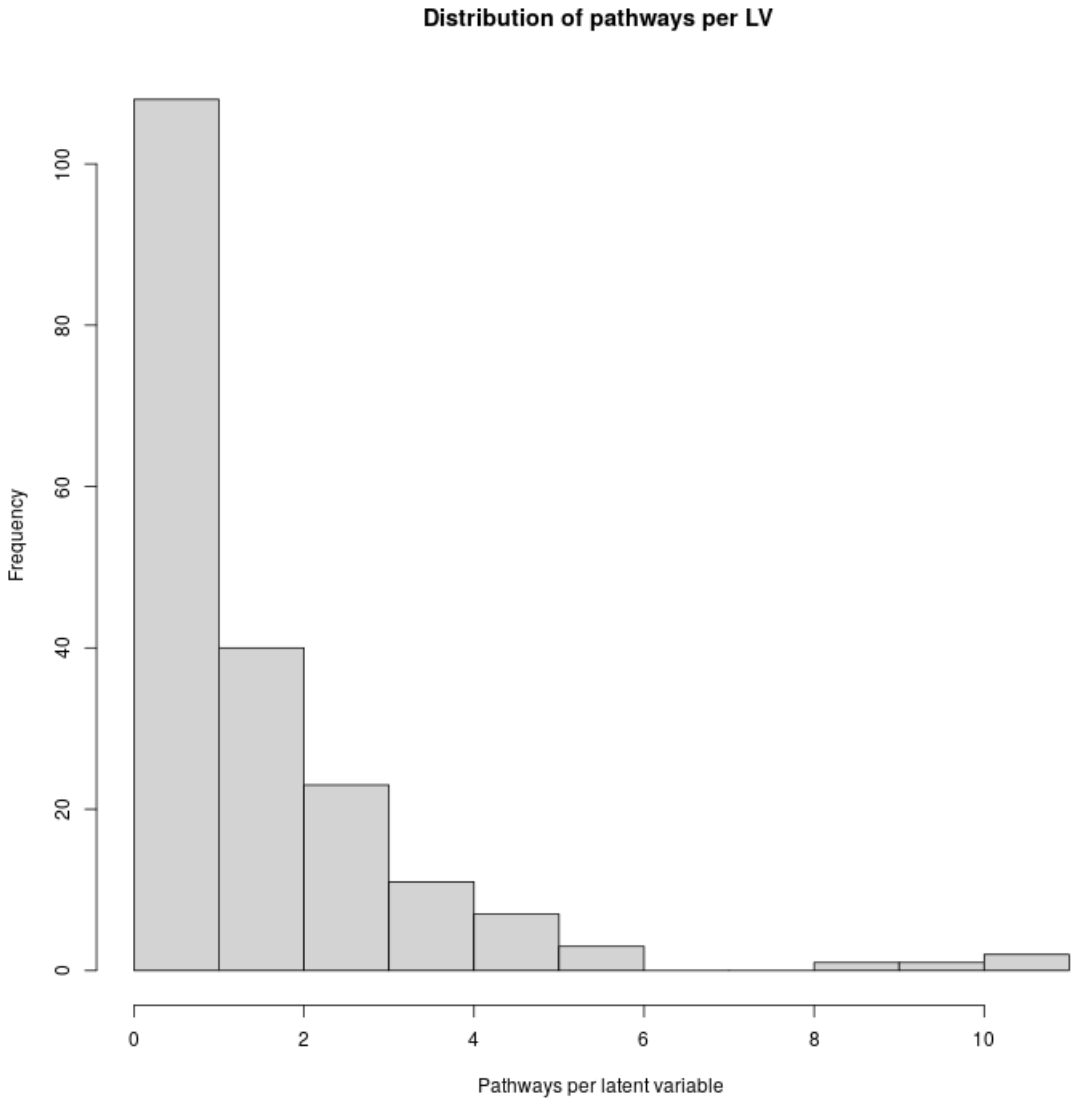

Figure 10: The distribution of the number of prior knowledge gene sets used per latent variable.

Some latent variables in MousiPLIER are enriched in wild-type microglia

With our model trained, we began interrogating our latent variables. Because we were interested in brain-relevant latent variables that our model had learned from the compendium, we analyzed a study on mouse brain aging from Pan et al. [155]. In this

study, microglia and astrocytes from five ages of mice were sequenced to see how their gene expression changed over time. To determine which (if any) latent variables were changed across developmental aging in the study, we used a linear model to find the latent variables that changed significantly as the cells aged. We found that each condition in the study had a set of significant latent variables, but that they were largely disjoint (Fig. 11). To narrow down the scope of the analysis, we decided to validate the biological relevance of the latent variables associated with wild-type microglial cells.

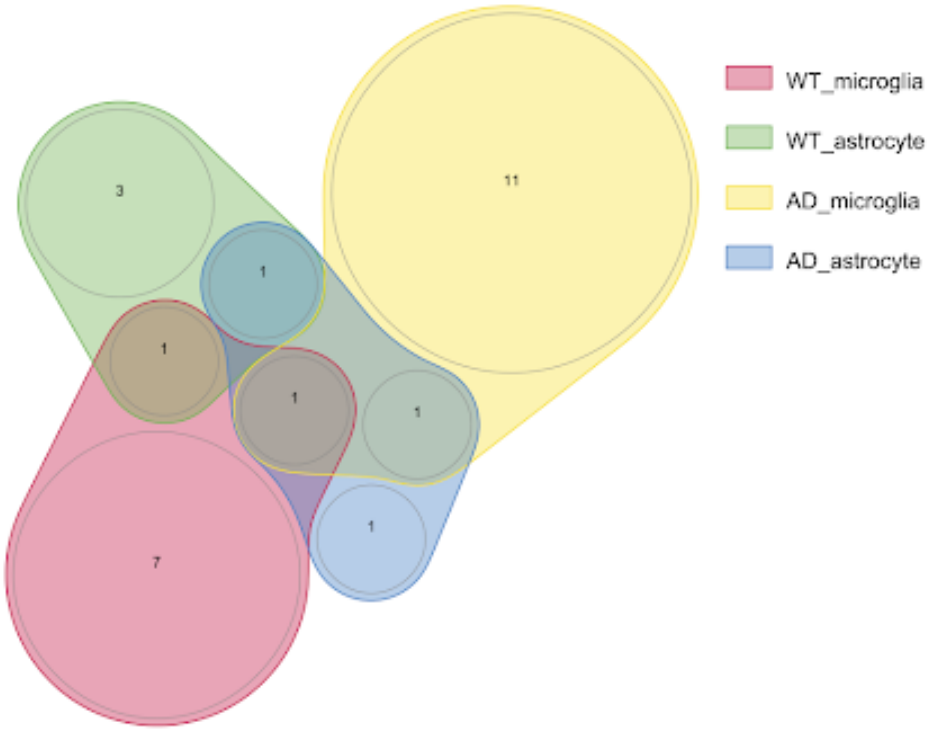

Figure 11: The overlap in significantly enriched latent variables across types and experimental conditions.

Latent variable 41 demonstrates the biological relevance of mousiplier latent variables

Once we had microglia-associated latent variables of interest, we set out to find which experiments in the training data responded strongly to them. To do so, we developed a novel method of ranking experiments based on their latent variable weights. More precisely, we performed *k*-means clustering with a *k* of two on each experiment in each latent variable space, and ranked experiments by their silhouette scores. This procedure allowed us to uncover experiments where there were two groups of samples with distinct sets of values for our latent variables of interest.

We focused this approach on latent variable 41, which contains genes functionally associated with striatal cell type specificity. Upon examining the top ranked studies, we found that several with high silhouette scores for latent variable 41 corresponded to processes occurring in the brain (Fig. 12). We dug deeper into which samples in particular were present in each clustered experiment and found our latent variable was in fact reflecting a biological process occurring in the striatum and cortex but not the cerebellum or non-brain tissues (Fig. 13). For example, in study SRP070440 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE76872) the striatal samples clearly stand apart from the other neuronal tissues. Similarly, we found that the two distinct groups in SRP047452 [156] were made up of brain samples from embryonic and adult mice, supporting the association between latent variable 41 and aging we found in the study we used to derive the latent variables (Fig. 14).

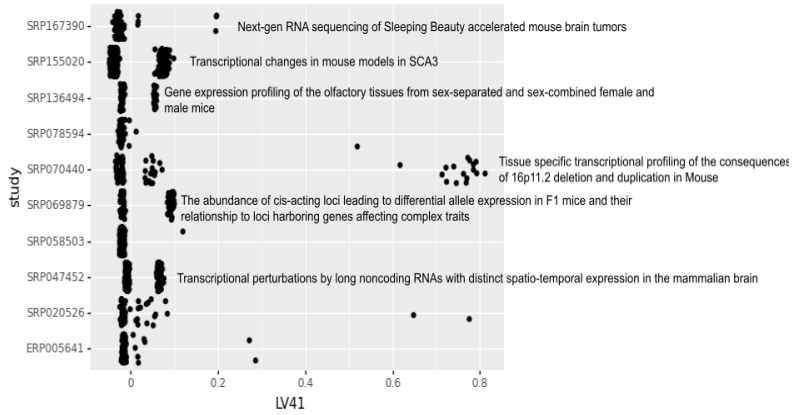

Figure 12: Latent variable 41 expression values are associated with brain-related studies

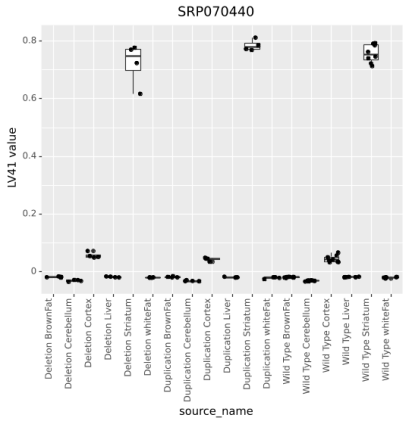

Figure 13: Striatal values for latent variable 41 compared to other tissues and brain regions.

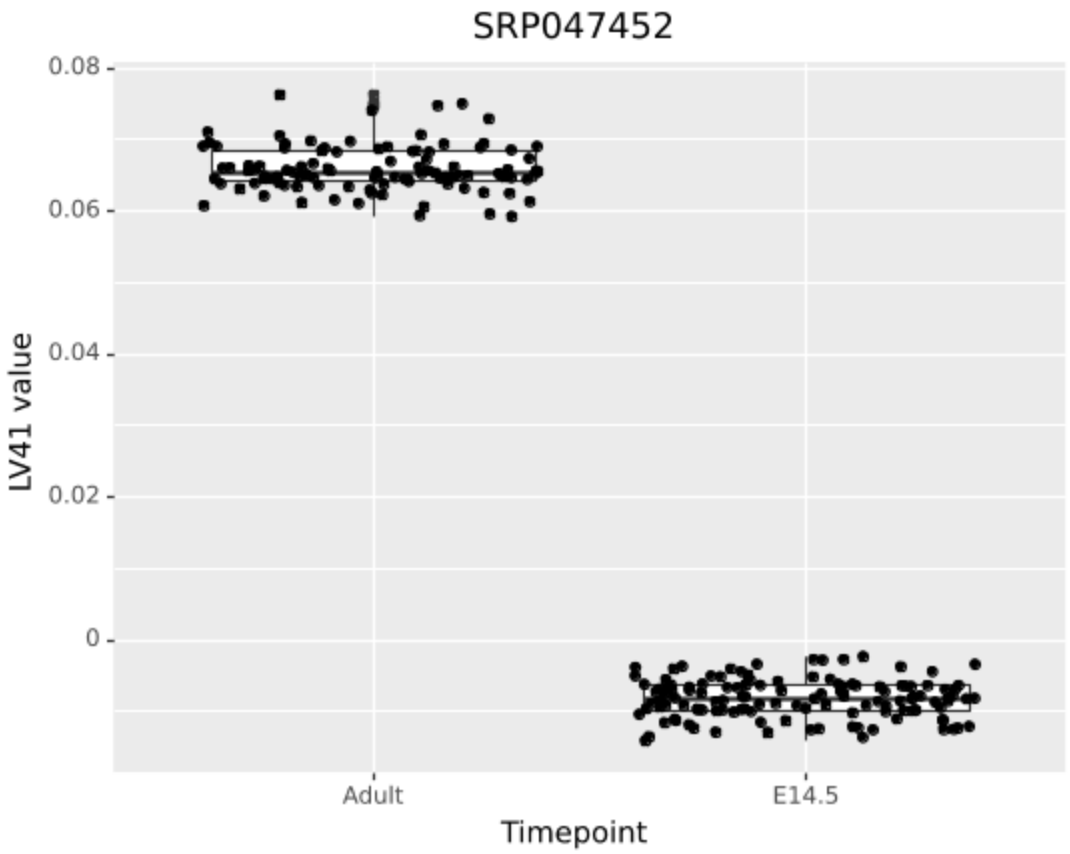

Figure 14: The effects of aging on latent variable 41.

To allow others to look at the learned latent variables on their own, we have set up a web server at https://www.mousiplier.greenelab.com. This server allows users to list the genes present in, visualize which experiments had high cluster scores for, and see which biological pathways participate in each latent variable (Fig. 15).

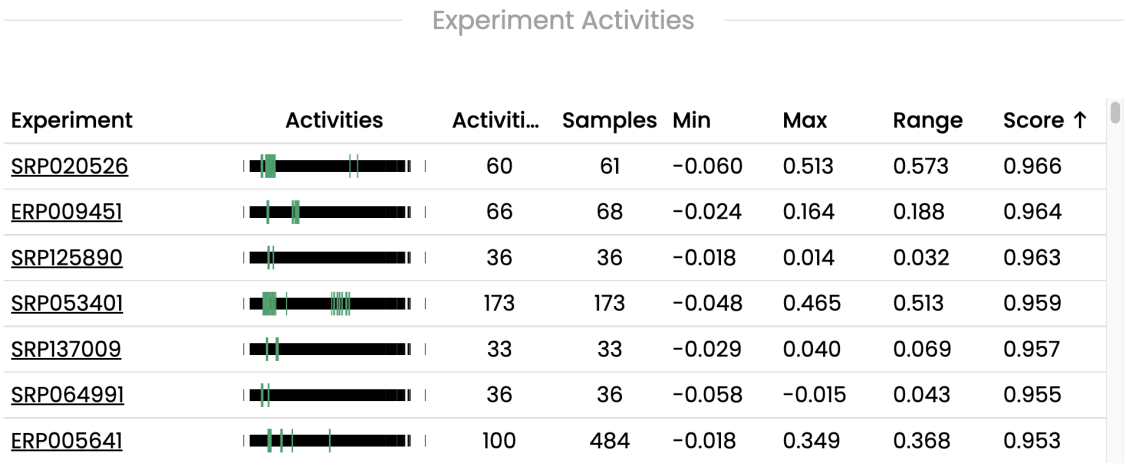

Figure 15: An image of the webserver displaying the per-latent variable experiment ranking feature.

Methods

Data

We began by downloading all the mouse gene expression data in Recount3, along with its corresponding metadata [126]. We then removed the single-cell RNAseq data from the dataset to ensure our data sources were fairly consistent across samples and studies. Next, we filtered the expression data, keeping only genes that overlapped between Recount3 and our prior-knowledge genesets. Finally, we RPKM transformed the expression using gene lengths from the Ensembl BioMart database [157] and Z-scored the expression data to ensure a consistent range for the downstream PLIER model.

For our prior knowledge gene sets we used cell type marker genes from CellMarker [158], pathway gene sets from Reactome [159], and manually curated brain marker genes. We selected cell type marker genes corresponding to all available mouse cell

types within the CellMarker database. For mouse biological pathways, we downloaded pathway information from the Reactome database. More specifically, we processed the files “Ensembl2Reactome_All_Levels.txt”, “ReactomePathways.txt”, and “ReactomePathwaysRelation.txt”, selecting only pathways using mouse genes, filtering out all pathways with fewer than 5 genes present, and keeping only pathways that were leaf nodes on the pathway hierarchy. Because we were interested in mouse brains in particular, we rounded out our set of prior information by manually selecting marker genes for the striatum, midbrain, and cerebellum. In total, we used 1,003 prior knowledge pathways when training our model.

PLIER

We began the PLIER pipeline by precomputing the initialization for PLIER with incremental PCA in scikit-learn [160]. We then used the expression compendium, prior knowledge genesets, and PCA initializations to train a PLIER model. The resulting task took two days to run and yielded 196 latent variables.

Latent variable significance

To determine which latent variables were associated with experimental conditions, we used a linear model. To correct the p-values for multiple testing, we used the Benjamani-Hochberg procedure [161].

Clustering

We selected the latent variables significantly associated with aging in mouse microglia as a biological starting point. We then used these latent variables to query the training data and see which studies seemed associated with the same biological signals. To do

so, we used *k*-means clustering with a *k* of 2, to look for experiments where there was some experimental condition that affected the latent variable. We then ranked the top ten studies based on their silhouette scores, and looked to see which conditions were associated with relevant experimental variables.

Hardware

The PLIER model training was performed on the Penn Medicine high performance computing cluster. The full pipeline takes around two weeks to run, with the main bottlenecks being the Recount3 data download, which takes one week to run, and training the PLIER model, which takes two days on a compute node with 250GB of RAM.

Web Server

The web server for visualizing the results was built on top of the ADAGE web app framework [162]. The main changes we made were to substitute the latent variables and gene sets from our trained PLIER model, to use clusters' silhouette scores for ranking experiments, and to forgo uploading the input expression data as the mouse compendium we used was much larger than the input expression for ADAGE.

Discussion/Conclusion

In this paper we demonstrate that it is possible to train large PLIER models on mouse data. We then show that the learned latent variables map to various biological processes and cell types. We also describe a novel approach for surfacing latent-variable relevant experiments from an expression compendium. Namely, we cluster them based on latent variable values, allowing us to query a large compendium for experiments pertaining to

mouse striatal aging. Finally, we create a web server to make the model's results more easily accessible to other scientists.

Our study is not without its limitations though. We show how a study from outside the training data can be transformed into the latent space to see which of the learned latent variables have significant differences in the study. However, not all studies will have significant changes in latent variables between their experimental conditions. This may be due to lack of similar samples in training compendium, too few samples in the study of interest, or other factors. In these cases, there isn't a good way to select which latent variables should be used for downstream analyses.

Additionally, PLIER is a linear model. If there are non-linear relationships between the genes used to train the model and the learned biological pathways, at best PLIER can approximate them. While we do not expect this to have a large impact [163], incorporating prior knowledge into non-linear models such as neural networks is an exciting field of research and a potential improvement for the MultiPLIER framework we use.

Going forward, our model and web server will allow scientists to explore the latent space of their own experiments and learn about relevant biological pathways and cell types.

Code and Data Availability

The code, and model weights to reproduce this work can be found at https://github.com/greenelab/mousiplier. The data used in our analyses is publicly available and can be downloaded with the code above or is already stored in the repository. Our work meets the bronze standard of reproducibility [142].

Acknowledgements

We would like to thank Jake Crawford for reviewinging code that went into this project. We would also like to thank Faisal Alquaddoomi and Vincent Rubinetti for their assistance in developing the web server accompanying this project. This work utilized resources from the University of Pennsylvania PMACS/DART computer cluster funded by NIH grant 1S10OD012312.

Funding

This work was supported by grants from the National Institutes of Health's National Human Genome Research Institute (NHGRI) under award R01 HG010067 and the Gordon and Betty Moore Foundation (GBMF 4552) to CSG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 5 - The Field-Dependent Nature of PageRank Values in Citation Networks

Contributions:

I ran the experiments, generated the figures, and wrote the manuscript for this chapter. Casey S. Greene gave advice and feedback, helped design experiments, and edited the manuscript.

Abstract

There are more academic papers than any human can read in a lifetime, so several article-level and journal-level metrics have been devised to rank papers. One challenge when creating such metrics is the differences in citation practices between fields. To account for these differences, scientists have devised normalization schemes to make metrics more comparable across fields. In this paper, we argue that these normalization schemes obscure useful signals about fields' preferences for articles. We use PageRank as an example metric and begin by demonstrating that there are, in fact, differences in journals' PageRanks between fields. We then show that even papers shared between fields have different PageRanks depending on which field's citation network the metric is calculated in. Finally, we find that some of these differences are caused by field-specific preferences by using a degree-preserving graph shuffling algorithm to generate a null distribution of similar networks. Our results demonstrate that while differences exist between fields' metric distributions, applying metrics in a field-aware manner rather than using normalized global metrics avoids losing important information about article preferences.

Introduction

There are more academic papers than any human can read in a lifetime. Attention has been given to ranking papers, journals, or researchers by their “importance,” assessed via various metrics. Citation count assumes the number of citations determines a paper’s importance. The h-index and Journal Impact Factor focus on secondary factors like author or journal track records. Graph-based methods like PageRank or disruption index use the context of the citing papers to evaluate an article’s relevance [71,75,78,80]. Each of these methods has its strengths, and permutations exist that attempt to shore up specific weaknesses [81,84,85,86].

One objection to such practices is that “importance” is subjective. The San Francisco Declaration on Research Assessment (DORA) argues against using Journal Impact Factor, or any journal-based metric, to assess individual manuscripts or scientists [164]. DORA further argues in favor of evaluating the scientific content of articles and notes that any metrics used should be article-level (https://sfdora.org/read/). However, even article-level metrics often ignore that the importance of a specific scientific output will fundamentally differ across fields. Even Nobel prize-winning work may be unimportant to a cancer biologist if the prize-winning article is about astrophysics.

Because there are differences between fields’ citation practices [165], scientists have developed strategies including normalizing the number of citations based on nearby papers in a citation network, rescaling fields’ citation data to give more consistent PageRank results, and so on [84,166,167,168]. Such approaches normalize away field-specific effects, which might help to compare one researcher with another in a very different field. However, they do not address the difference in the relevance of a topic between fields. This phenomenon of field-specific importance has been observed at the

level of journal metrics. Mason and Singh recently noted that depending on the field, the journal *Christian Higher Education* is either ranked as a Q1 (top quartile) journal or a Q4 (bottom quartile) journal [169].

It is possible that, while global journal-level metrics fail to capture field-specific importance, article-level metrics are sufficiently granular that the importance of a manuscript remains constant across fields. We investigate the extent to which article-level metrics generalize between fields. We examine this using MeSH terms to define fields and use field-specific citation graphs to assess their importance within the field. While it is trivially apparent that journals or articles that do not have cross-field citations will have variable importance, we ignore these cases and include only those with citations in both fields, where we expect possible consistency. We first replicate previous findings that journal-level metrics can differ substantially among fields. We also find field-specific variability in importance at the article level. We make our results explorable through a web app that shows metrics for overlapping papers between pairs of fields.

Our results show that even article-level metrics can differ substantially among fields. We recommend that metrics used for assessing research outputs include field-specific, in addition to global, ones. While qualitative assessment of the content of manuscripts remains time-consuming, our results suggest that within-field and across-field assessment remains key to assessing the importance of research outputs.

Results

Journal rankings differ between fields

In an attempt to quantify the relative importance of journals, scientists have created rankings using metrics the Journal Impact Factor, which is essentially based on citations per article, and those that rely on more complex representations like Eigenfactor [76]. It has previously been reported that journal rankings differ substantially between fields using metrics based on citation numbers [169]. We calculated a PageRank-based score for the journal as the median PageRank of manuscripts published in that journal for that field (Fig. 16 A). We first sought to understand the extent to which journal ranking differences replicated using PageRank.

To begin, we compared the differences in ranking between the top fifty journals in nanotechnology and their corresponding ranks in microscopy. While the ranks were somewhat correlated there was a great deal of variance, especially for journals outside the top 20 in nanotechnology (Fig. 16 B). We then made use of the scale of the data by examining the top ranked journal in each of our 45 fields to determine whether the top ranking journal would be consistent across fields (Fig. 16 C). We found that the most common top-ranked journal was *Science*. This was unsurprising, given that it tends to rank highly among global journal level metrics such as eigenfactor. However, the ranking was very field-dependent, with only 20% of fields having *Science* as their top ranked journal.

One could argue that while general journals may have differing influence by field, specialty journals correspond to a single field so field-aware metrics are irrelevant. That turns out to be untrue. Of the 5,178 journals with at least 50 articles present in our dataset, the median number of fields publishing in a given journal is 15 (Fig. 16 D). This result confirms that while useful [170], MeSH headings reflect a different type of aggregation than journals do [171].

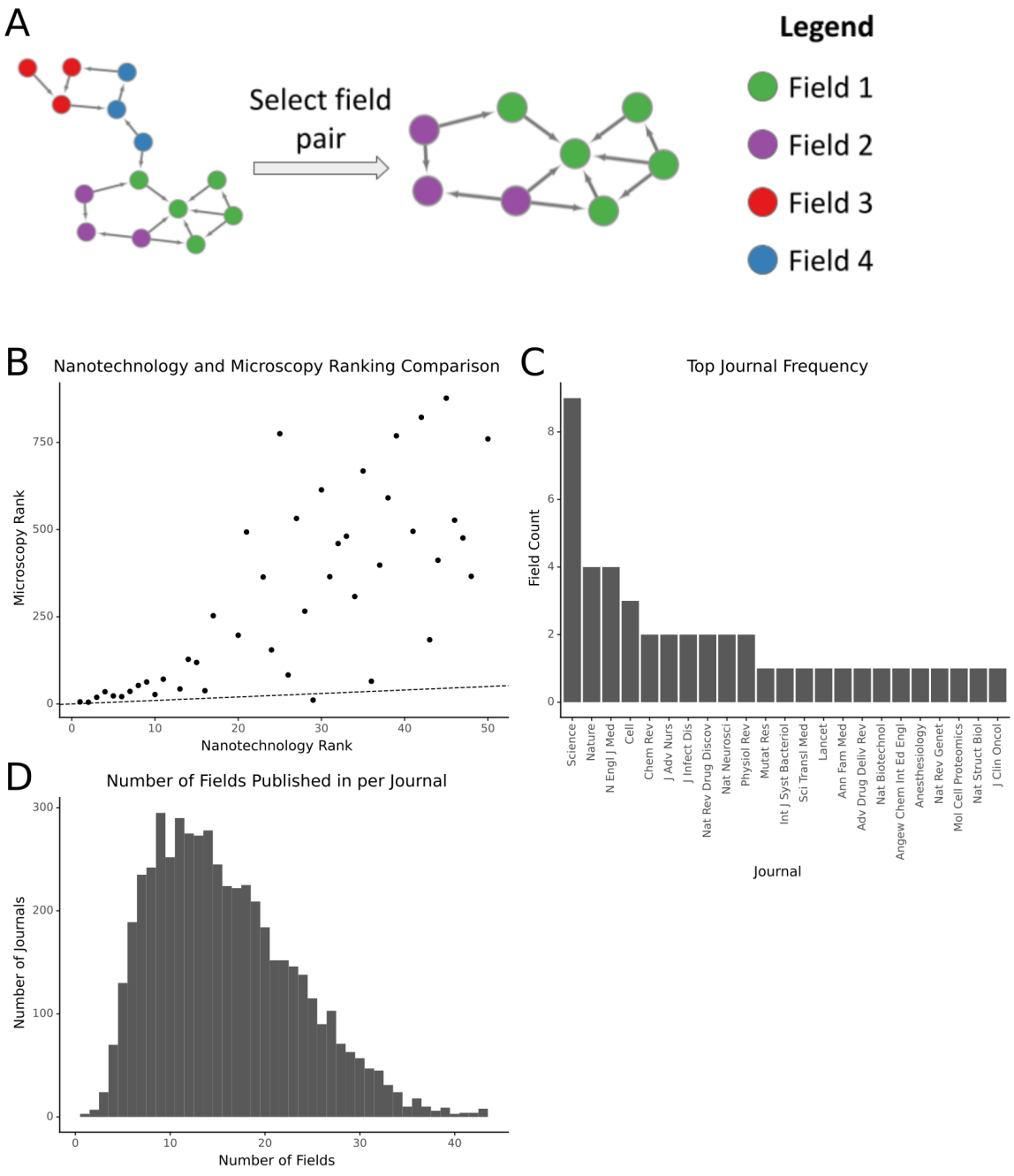

Figure 16: Journals' PageRank derived rankings differ between fields. A) A schematic showing how paired networks are derived from the full citation network. B) A comparison of the ranks of the top 50 journals by PageRank in nanotechnology and their rank in microscopy. Top-50 nanotechnology journals with no papers in microscopy have been

omitted. C) The frequency with which journals in the dataset are the top journal for a field. D) The distribution of fields published per journal. The X-axis corresponds to the number of fields for which a journal has at least one paper within the field. All plots restrict the set of journals to those with at least 50 papers in the dataset.

Manuscript PageRanks differ between fields

We split the citation network into its component fields and calculated the PageRank for each article (Fig. 17 A). We then examined the distribution of PageRanks across fields and found that they differed greatly (Fig. 17 B). These differences were driven by the size and citation practices of the fields themselves, as the papers shared by pairs of fields had distributions matching the field context they were in (Fig. 17 B, C). Given the differences in distributions in articles shared by these fields (Fig. 17 D), we found it difficult to determine whether correspondence between fields was random or due to different degrees of interest in certain articles (Fig. 17 E).

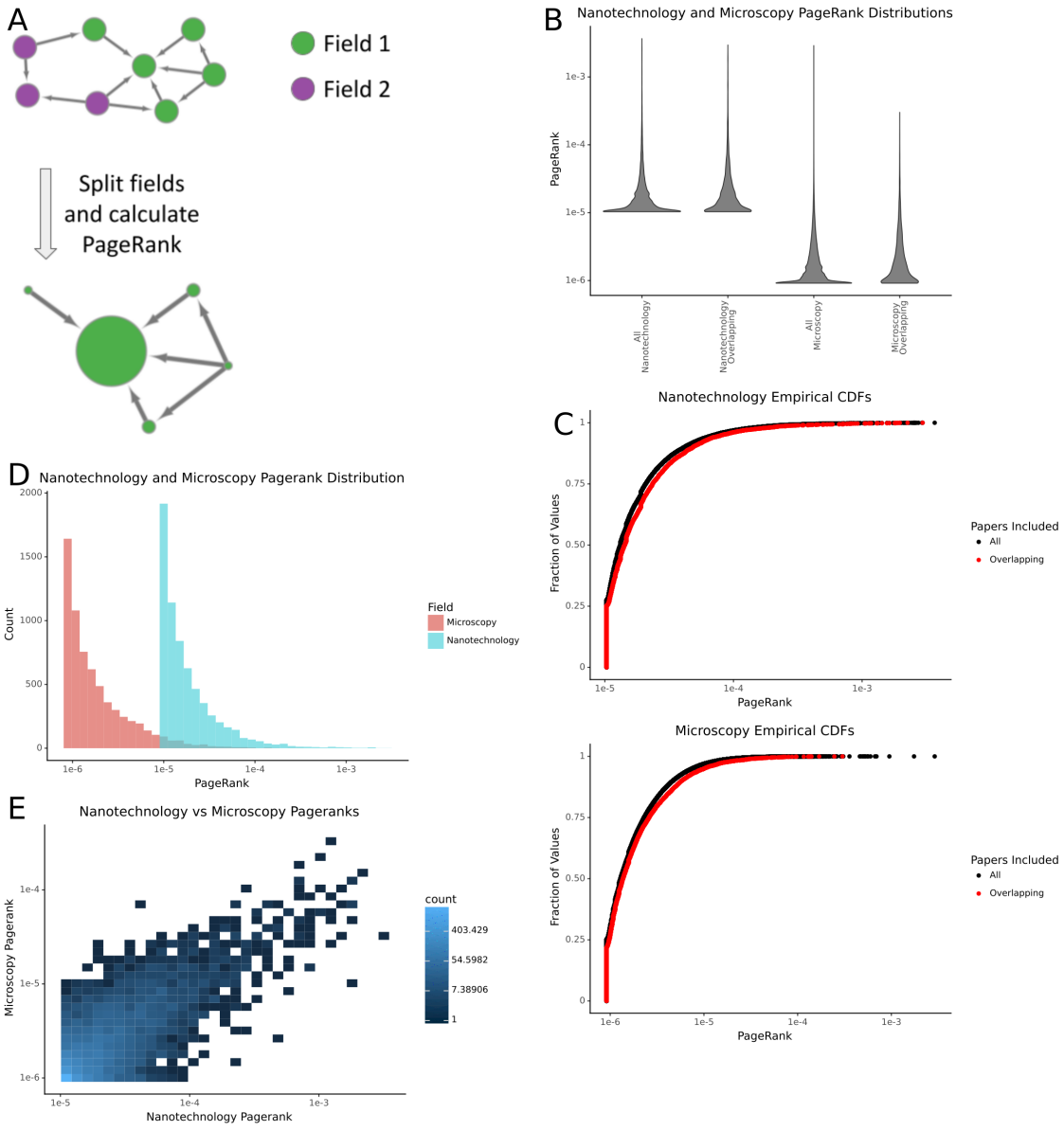

Figure 17: Differences in the distribution of PageRanks between fields. A) A schematic showing how field pairs are split and their PageRanks are calculated. B) The distribution of article PageRanks for nanotechnology and microscopy. The distributions marked with 'All' contain all the papers for the given field in the dataset, while those marked 'overlapping' contain only articles present in both fields. C) The empirical cumulative density functions of nanotechnology and microscopy. D) The differences in distribution of

the PageRanks of articles shared by nanotechnology and microscopy. E) A density plot showing the joint distribution of PageRanks for papers overlapping in nanotechnology and microscopy.

Fields' differences are not solely driven by differences in citation practices

We devised a strategy to generate an empirical null for a field pair under the assumption that the field pair represented a single, homogeneous field (Fig. 18 A). For each field-pair intersection, we performed a degree-distribution preserving permutation. We created 100 permuted networks for each field pair. We then split the networks into their constituent fields and calculated a percentile using the number of permuted networks with a lower PageRank for a manuscript than the true PageRank. A manuscript with a PageRank higher than all networks has a percentile of 100, and one lower than all permuted networks has a percentile of zero. We used the difference in the percentile in each field as the field-specific affinity for a given paper. This percentile score allowed us to control for the differing degree distributions between fields by comparing papers based on their expected PageRank in a random network with the same node degrees.

We selected field pairs with varying degrees of correlation between their PageRanks (Fig. 18 B). By examining the fields' PageRank percentiles, we found that many articles had large differences in their perception between fields (Fig. 18 C). In nanotechnology and microscopy, papers with high nanotechnology percentiles and low microscopy percentiles tended towards applications of nanotechnology, while their counterparts with high microscopy percentiles and low nanotechnology percentiles were often papers about technological developments in microscopy (Fig. 18 A, Table 1). Immunochemistry-favored papers are largely applications of immunochemical methods, while anatomy-favored articles tend to focus experiments on a single anatomical region

(Fig. 18 B, Table 2). Proteomics and metabolomics tend to use similar methods, so the fields on either end are largely (though not entirely) field-specific applications of those methods (Fig. 18 C, Table 3). Computational biology is similarly applications-focused, though human genetics tends towards policy papers due to its MeSH heading (H01.158.273.343.385) excluding fields like genomics, population genetics, and microbial genetics (Fig. 18 D, Table 4). In addition to papers with large differences between fields, each field also has papers with high PageRanks and similar percentiles in both fields. Overall it is clear that while some papers may be influential in multiple fields, others have more field-specific import.

It is not possible to describe all the field-pairs and relevant differences between fields within the space of a journal article. Instead, we have developed a web server that displays the percentiles for all pairs of fields in our dataset with at least 1000 shared articles (Fig. 18 D), which can be accessed at https://www.indices.greenelab.com. We hope that the availability of the web server and the reproducibility of our code will assist other scientists in uncovering new insights from this dataset.

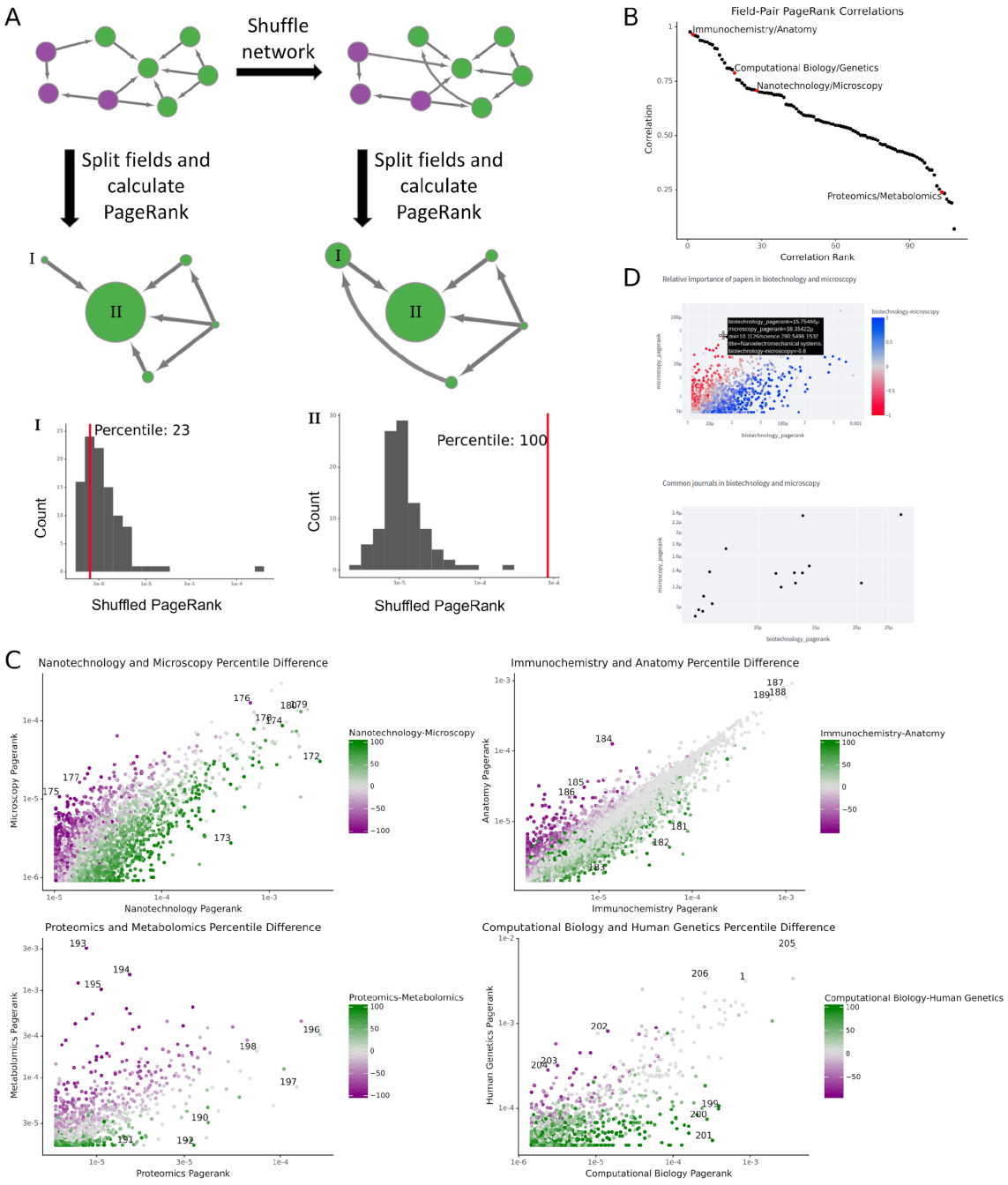

Figure 18: Field-specific preferences in papers. A) A schematic showing how networks are shuffled and how articles' percentile scores are calculated. The histograms at the bottom of the figure correspond to the distribution of PageRanks for the shuffled networks, while the red lines correspond to an article's PageRank in the true citation

network. B) The Pearson correlation of PageRanks between fields. The red points are the field pairs expanded in panel C. C) The percentile scores and PageRanks for overlapping articles in various fields. Points are colored based on the difference in percentile scores in the fields e.g. “Nanotechnology-Microscopy” corresponds to the difference between the nanotechnology and microscopy percentile scores. The numbers next to points are the reference number for the article in the bibliography. D) A screenshot of the web server showing the percentile score difference and journal median PageRank plot functionality.

Nanotechnology Percentile	Microscopy Percentile	Title	Reference
100	4	A robust DNA mechanical device controlled by hybridization topology	[172]
100	5	Bioadhesive poly(methyl methacrylate) microdevices for controlled drug delivery	[173]
99	2	DNA-templated self-assembly of protein arrays and highly conductive nanowires	[174]
0	100	Photostable luminescent nanoparticles as biological label for cell recognition of system lupus erythematosus patients	[175]

Nanotechnology Percentile	Microscopy Percentile	Title	Reference
5	90	WSXM: a software for scanning probe microscopy and a tool for nanotechnology	[176]
0	77	Measuring Distances in Supported Bilayers by Fluorescence Interference-Contrast Microscopy: Polymer Supports and SNARE Proteins	[177]
100	99	Toward fluorescence nanoscopy	[178]
100	86	In vivo imaging of quantum dots encapsulated in phospholipid micelles	[179]
100	99	Water-Soluble Quantum Dots for Multiphoton Fluorescence Imaging in Vivo	[180]

Table 2: Nanotechnology/microscopy papers of interest

Immunochemistry Percentile	Anatomy Percentile	Title	Reference
100	45	Immunoelectron microscopic exploration of the Golgi complex	[181]
100	14	Immunocytochemical and electrophoretic analyses of changes in myosin gene expression in cat posterior temporalis muscle during postnatal development	[182]
98	5	Electron microscopic demonstration of calcitonin in human medullary carcinoma of thyroid by the immuno gold staining method	[183]
12	100	Grafting genetically modified cells into the rat brain: characteristics of E. coli β-galactosidase as a reporter gene	[184]
12	100	Vitamin-D-dependent calcium-binding-protein and parvalbumin occur in bones and teeth	[185]

Immunochemistry Percentile	Anatomy Percentile	Title	Reference
3	100	Mapping of brain areas containing RNA homologous to cDNAs encoding the alpha and beta subunits of the rat GABAA gamma-aminobutyrate receptor	[186]
100	100	Studies of the HER-2/neu Proto-Oncogene in Human Breast and Ovarian Cancer	[187]
100	100	Expression of c-fos Protein in Brain: Metabolic Mapping at the Cellular Level	[188]
100	100	Proliferating cell nuclear antigen (PCNA) immunolocalization in paraffin sections: An index of cell proliferation with evidence of deregulated expression in some neoplasms	[189]

Table 3: Immunochemistry/anatomy papers of interest

Proteomics Percentile	Metabolomics Percentile	Title	Reference
67	2	Proteomics Standards Initiative: Fifteen Years of Progress and Future Work	[190]
99	0	Limited Environmental Serine and Glycine Confer Brain Metastasis Sensitivity to PHGDH Inhibition	[191]
100	0	A high-throughput processing service for retention time alignment of complex proteomics and metabolomics LC-MS data	[192]
0	100	MeltDB: a software platform for the analysis and integration of metabolomics experiment data	[193]
0	98	In silico fragmentation for computer assisted identification of metabolite mass spectra	[194]

Proteomics Percentile	Metabolomics Percentile	Title	Reference
0	100	The Metabonomic Signature of Celiac Disease	[195]
91	70	Visualization of omics data for systems biology	[196]
0	16	FunRich: An open access standalone functional enrichment and interaction network analysis tool	[197]
0	5	Proteomic and Metabolomic Characterization of COVID-19 Patient Sera	[198]

Table 4: Proteomics/metabolomics papers of interest

Computational Biology Percentile	Human Genetics Percentile	Title	Reference
99	0	Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans	[199]

Computational Biology Percentile	Human Genetics Percentile	Title	Reference
100	1	A database for post-genome analysis	[200]
100	1	Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases	[201]
12	100	Genetic Discrimination: Perspectives of Consumers	[202]
0	81	Committee Opinion No. 690: Carrier Screening in the Age of Genomic Medicine	[203]
23	100	Public health genomics: The end of the beginning	[204]
100	99	Initial sequencing and analysis of the human genome	[1]
100	100	An STS-Based Map of the Human Genome	[205]

Computational Biology Percentile	Human Genetics Percentile	Title	Reference
100	100	A New Five-Year Plan for the U.S. Human Genome Project	[206]

Table 5: Computational biology/human genetics papers of interest

Methods

COCI

We used the March 2022 version of the COCI citation index [88] as the source of our citation data. This dataset contains around 1.3 billion citations from ~73 million bibliographic resources.

Selecting fields

To differentiate between scientific fields, we needed a way to map papers to fields. Fortunately, all the papers in Pubmed Central (https://www.ncbi.nlm.nih.gov/pmc/) have corresponding Medical Subject Headings (MeSH) terms. While MeSH terms are varied and numerous, the subheadings of the Natural Science Disciplines (H01) category fit our needs. However, MeSH terms are hierarchical, and vary greatly in their size and specificity. To extract a balanced set of terms we recursively traversed the tree and selected headings that have least 10000 DOIs and don't have multiple children that also meet the cutoff. Our resulting set of headings contained 45 terms, from "Acoustics" to "Water Microbiology".

Constructing citation networks

The COCI dataset consists of pairs of Digital Object Identifiers (DOIs). To change these pairs into a form we could run calculations on, we needed to convert them into networks. To do so, we created 45 empty networks, one for each MeSH term we selected previously. We then iterated over each pair of DOIs in COCI, and added them to a network if the DOIs corresponded to two journal articles written in english, both of which were tagged with the corresponding MeSH heading.

Because we were interested in the differences between fields, we also needed to build networks from pairs of MeSH headings. These networks were built via the same process, except that instead of keeping articles corresponding to a single DOI we added a citation to the network if both articles were in the pair of fields, even if the citation occurred across fields. Running this network-building process yielded 990 two-heading networks.

Sampling a graph from the degree distribution while preserving the distribution of degrees in the network turned out to be challenging. Because citation graphs are directed, it's not possible to simply swap pairs of edges and end up with a graph that is uniformly sampled from the space. Instead, a more sophisticated three-edge swap method must be used [207]. Because this algorithm had not been implemented yet in NetworkX [208], we wrote the code to perform shuffles and submitted our change to the library. With the shuffling code implemented, we created 100 shuffled versions of each of our combined networks to act as a background distribution to compare metrics against.

Once we had a collection of shuffled networks, we needed to split them into their constituent fields. To do so, we reduced the network to solely the nodes that were present in the single heading citation network, and kept only citations between these nodes.

Metrics

We used the NetworkX implementation of PageRank with default parameters to evaluate paper importance within fields. To determine the degree to which the papers' PageRank values were higher or lower than expected, we compared the PageRank values calculated for the true citation networks to the values in the shuffled networks for each paper. We then recorded the fraction of shuffled networks where the paper had a lower PageRank than in the true network to derive a single number that described these values. For example, if a paper had a higher PageRank in the true network than in all the shuffled networks it received a score of 1. Likewise, if it had a lower PageRank in the true network than in all the shuffled networks it received a score of 0. Papers in between the two extremes had fractional values, like .5 (a paper that fell in the middle of the pack) and so on.

A convenient feature of the percentile scores is that they're directly comparable between fields. If a paper is present in two fields, the difference in scores between the two fields can be used to estimate its relative importance. For example, if a paper has a score of 1 in field A (indicating a higher PageRank in the field than expected given its number of citations and the network structure) and a score of 0 in field B (indicating a lower than expected PageRank), then the large difference in scores indicates the paper is more highly valued in field A than field B. If the paper has similar scores in both fields, it indicates that the paper is similarly valued in the two fields.

Hardware/runtime

The analysis pipeline was run on the RMACC Summit cluster. The full pipeline, from downloading the data to analyzing it to visualizing it took about a week to run. However,

that number is heavily dependent on details such as the number of CPU nodes available and the network speed.

Our web server is built by visualizing our data in Plotly (https://plotly.com/python/plotly-express/) on the Streamlit platform (https://streamlit.io/). The field pairs made available by the frontend are those with at least 1000 shared papers after filtering out papers with more than a 5% missingness level of their PageRanks after shuffling. The journals available for visualization are those with at least 25 papers for the given field pair.

Discussion/Conclusion

We analyze hundreds of field-pair citation networks to examine the extent to which article-level importance metrics vary between fields. As previously reported, we find systematic differences in PageRanks between fields [86,209] that would warrant some form of normalization when making cross-field comparisons with global statistics. However, we also find that field-specific differences are not driven solely by differences in citation practices. Instead, the importance of individual papers appears to differ meaningfully between fields. Global rankings or efforts to normalize out field-specific effects obscure meaningful differences in manuscript importance between communities.

As with any study, this research has certain limitations. One example is our selection of MeSH terms to represent fields. We used MeSH because it is a widely-annotated set of subjects in biomedicine and thresholded MeSH term sizes to balance having enough observations to calculate appropriate statistics with having sufficient granularity to capture fields. This selection process resulted in fields at the granularity of “biophysics” and “ecology.” We also have to select a number of swaps to generate a background

distribution of PageRanks for each field pair. We selected three times as many swaps as edges, where each swap modifies three edges, but certain network structures may require a different number.

We also note that there are inherent issues with the premise of ranking manuscripts' importance. We sought to understand the extent to which such rankings were stable between fields after correcting for field-specific citation practices. We found limited stability between fields, mostly between closely-related fields, suggesting that the concept of a universal ranking of importance is difficult to justify. In the way that reducing a distribution to a Journal Impact Factor distorts assessment, attempting to use a single universal score to represent importance across fields poses similar challenges at the level of individual manuscripts. Furthermore, this work's natural progression would extend to estimating the importance of individual manuscripts to individual researchers. Thus, a holistic measure of importance would need to include a distribution of scores not only across fields but across researchers. It may ultimately be impossible to calculate a meaningful importance score. The lack of ground truth for importance is an inherent feature, not a bug, of science's step-wide progression.

Shifting from the perspective of evaluation to discovery can reveal more appropriate uses for these types of statistics. Field-pair calculations for such metrics may help with self-directed learning of new fields. An expert in one field, e.g., computational biology, who aims to learn more about genetics may find manuscripts with high importance in genetics and low importance in computational biology to be important reads. These represent manuscripts not currently widely cited in one's field but highly influential in a target field. Our application can reveal these manuscripts for MeSH field pairs, and our source code allows others to perform our analysis with different granularity.

Code and Data Availability

The code to reproduce this work can be found at https://github.com/greenelab/indices. The data used for this project is publicly available and can be downloaded with the code provided above. Our work meets the bronze standard of reproducibility [142] and fulfills aspects of the silver and gold standards including deterministic operation.

Acknowledgements

We would like to thank Jake Crawford for reviewing code that went into this project and Faisal Alquaddoomi for figuring out the web server hosting. We would also like to thank the past and present members of GreeneLab who gave feedback on this project during lab meetings. This work utilized resources from the University of Colorado Boulder Research Computing Group, which is supported by the National Science Foundation (awards ACI-1532235 and ACI-1532236).

Funding

This work was supported by grants from the National Institutes of Health's National Human Genome Research Institute (NHGRI) under award R01 HG010067 and the Gordon and Betty Moore Foundation (GBMF 4552) to CSG. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Chapter 6 - Future Directions

Contributions: I wrote and edited this chapter for the purpose of being included in this dissertation.

Introduction

In this dissertation, we have examined whether deep learning has led to a paradigm shift in computational biology. We established standards for reproducible research when using deep learning models in chapter 2, showed that deep learning is not always preferable to other techniques in chapter 3, then demonstrated the effectiveness of classical machine learning methods in chapters 4 and 5. Ultimately we concluded that while deep learning has been a useful tool in some areas, it has yet to lead to a paradigm shift in computational biology. However, deep learning models' impact may grow as the fields develop, so we would like to discuss future areas where we expect interesting developments.

Deep learning representations of biology

Different areas of computational biology research have seen different effects from deep learning. Deep learning has already had a significant impact on biomedical imaging [210], and seems poised to do so in protein structure [11]. These advances were likely successful because of their similarity to well-researched fields in that they can be framed as similar problems. Biomedical images are not the same as those from a standard camera, but the inductive bias of translational equivariance and various image augmentation methods are still applicable. Similarly, while protein sequences may not seem to share much with written language, models like recurrent neural networks and transformers that look at their input as a sequence of tokens do not care whether those tokens are words or amino acids.

Not all subfields of computational biology have convenient ways to represent their data, though. Gene expression, in particular, is difficult because of its high dimensionality.

Expression data does not have spatial locality to take advantage of, so convolutional networks cannot be used to ignore it. It is not a series of tokens either; the genes in an expression dataset are listed lexicographically, so their order does not have meaning. Self-attention seems well-suited for gene expression since learning which subsets of genes interact with others would be useful. The high dimensionality makes vanilla self-attention infeasible though, due to the quadratic scaling. This issue cannot even be sidestepped with standard dimensionality reduction methods without losing predictive performance.

Do any deep learning representations work for gene expression, then? Fully-connected networks work, though they do not tend to be the best way to accomplish most tasks. An interesting potential research direction would be to apply sparse self-attention methods to gene expression data and reduce the number of comparisons made by only attending within prior knowledge gene sets. Alternatively, because expression is often thought of in terms of coregulation networks or sets of genes with shared functions, a graph representation may be more suitable. It is also possible that someone will develop a representation specifically for gene expression that will work better than anything we know about today.

To what extent is biology limited by challenges in looking at the data

An essential first step when working with data is to look at it. In images or generated text, a human can judge how good generated data is. In the classification world, a human labeler can look at an image and say, “that is a dog,” or a sentence and say, “that is grammatically correct English.” While these labels are somewhat fuzzy, a group of

humans can at least look at the label and say, “that is reasonable” or “that is mislabeled.” A human looking at a gene expression microarray or a table of RNA-seq counts cannot do the same.

Our brains are built to recognize objects, not parse gene expression perturbations corresponding to septic shock. This issue is not insurmountable; scientists can do research in quantum physics, after all. It simply serves as a hindrance to our ability to sanity-check data. Because we cannot see whether the relevant signals are distorted by batch effect normalization or a preprocessing step, we must be more careful and try more options. Perhaps in the future, as we understand more about the relevant biology, scientists will be able to create views of the data that are more human-intuitive and easier to use.

The scale of biological data

Biological data (or at least transcriptomic data) is not actually that big. The largest uniformly processed compendia of bulk human expression data contains hundreds of thousands of samples. Meanwhile in machine learning, even before deep learning took off ImageNet already had more than three million images [211].

Worse, many biological domains have strict upper bounds on the amount of data available. Even if one somehow recruited the entire world for a study, they would only be able to collect around eight billion human genomes. Given the complexity of biology, it seems unlikely that “only” eight billion genomes would be sufficient to effectively sample the space of plausible relevant mutations in the human genome. Based on recent research into neural network scaling laws [212] and machine learning intuition, it seems likely that Rich Sutton’s “Bitter Lesson”

(http://www.incompleteideas.net/IncIdeas/BitterLesson.html) would break down in a domain where there is a hard cap on the available data. This data cap probably is not true of all domains in computational biology, though. Gene expression changes with variables like cell type, time, and biological state, so the space of transcriptomic data that could be measured is much larger.

While we have shown that deep learning has not led to a paradigm shift in computational biology so far, will that always be true? As with many scientific questions, the answer is probably “it depends.” While there may be caps on individual aspects of biological data, there are always more angles of attack.

The promise of multiomics has always been that multiple views of the same system may reveal something that no single view picks up. The challenge is that the data types are different, their relationships are not well-characterized, and the methods for working in such a system have not been fully developed yet. Transformer architectures, and more specifically their self-attention mechanism, seem like a good fit for learning relationships between different ’omes. Such models are data-hungry, though, and self-attention gets expensive in problems with high dimensionality. Perhaps one day we will have the data and compute to train multiomic biological transformers. Or maybe by then the state of the art in machine learning will have moved along, rendering them irrelevant.

Conclusion

Whether deep learning takes over or simply becomes another tool in our toolbelt, the future of computational biology looks bright. These are exciting times indeed.

Bibliography

1. **Initial sequencing and analysis of the human genome** Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, … Michael J Morgan *Nature* (2001-02-15) https://doi.org/bfpgjh DOI: 10.1038/35057062 · PMID: 11237011

2. **Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets** Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, … Steven A McCarroll *Cell* (2015-05) https://doi.org/f7dkxv DOI: 10.1016/j.cell.2015.05.002 · PMID: 26000488 · PMCID: PMC4481139

3. **An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites** Peter J Skene, Steven Henikoff *eLife* (2017-01-16) https://doi.org/gfkh8w DOI: 10.7554/elife.21856 · PMID: 28079019 · PMCID: PMC5310842

4. **The second decade of 3C technologies: detailed insights into nuclear organization** Annette Denker, Wouter de Laat *Genes & Development* (2016-06-15) https://doi.org/gdcfmg DOI: 10.1101/gad.281964.116 · PMID: 27340173 · PMCID: PMC4926860

5. **The structure of scientific revolutions** Thomas S Kuhn, Ian Hacking *The University of Chicago Press* (2012) ISBN: 9780226458113

6. **Mastering the game of Go with deep neural networks and tree search** David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George van den

Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, … Demis Hassabis *Nature* (2016-01-27) https://doi.org/f77tw6 DOI: 10.1038/nature16961 · PMID: 26819042

7. **High-Resolution Image Synthesis with Latent Diffusion Models** Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer *arXiv* (2022-04-14) https://arxiv.org/abs/2112.10752

8. **Deep Double Descent: Where Bigger Models and More Data Hurt** Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, Ilya Sutskever *arXiv* (2019-12-06) https://arxiv.org/abs/1912.02292

9. **Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets** Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, Vedant Misra *arXiv* (2022-01-07) https://arxiv.org/abs/2201.02177

10. **U-Net: Convolutional Networks for Biomedical Image Segmentation** Olaf Ronneberger, Philipp Fischer, Thomas Brox *Lecture Notes in Computer Science* (2015) https://doi.org/gcgk7j DOI: 10.1007/978-3-319-24574-4_28

11. **Highly accurate protein structure prediction with AlphaFold** John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, … Demis Hassabis *Nature* (2021-07-15) https://doi.org/gk7nfp DOI: 10.1038/s41586-021-03819-2 · PMID: 34265844 · PMCID: PMC8371605

12. **The elements of statistical learning: data mining, inference, and prediction** Trevor Hastie, Robert Tibshirani, JH Friedman *Springer* (2009) ISBN: 9780387848570

13. **Diagnosis of multiple cancer types by shrunken centroids of gene expression** Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, Gilbert Chu *Proceedings of the National Academy of Sciences* (2002-05-14) https://doi.org/d2h5n3 DOI: 10.1073/pnas.082099299 · PMID: 12011421 · PMCID: PMC124443

14. **Averaged gene expressions for regression** MY Park, T Hastie, R Tibshirani *Biostatistics* (2006-05-11) https://doi.org/czxtxj DOI: 10.1093/biostatistics/kxl002 · PMID: 16698769

15. **'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns** Trevor Hastie, Robert Tibshirani, Michael B Eisen, Ash Alizadeh, Ronald Levy, Louis Staudt, Wing C Chan, David Botstein, Patrick Brown *Genome Biology* (2000) https://doi.org/fsmp4p DOI: 10.1186/gb-2000-1-2-research0003 · PMID: 11178228 · PMCID: PMC15015

16. **Outlier sums for differential gene expression analysis** R Tibshirani, T Hastie *Biostatistics* (2006-05-15) https://doi.org/cn72qh DOI: 10.1093/biostatistics/kxl005 · PMID: 16702229

17. **Machine learning approaches to predict lupus disease activity from gene expression data** Brian Kegerreis, Michelle D Catalina, Prathyusha Bachali, Nicholas S Geraci, Adam C Labonte, Chen Zeng, Nathaniel Stearrett, Keith A Crandall, Peter E Lipsky, Amrie C Grammer *Scientific Reports* (2019-07-03) https://doi.org/gh33ng DOI: 10.1038/s41598-019-45989-0 · PMID: 31270349 · PMCID: PMC6610624

18. **Weighted elastic net for unsupervised domain adaptation with application to age prediction from DNA methylation data** Lisa Handl, Adrin Jalali, Michael Scherer,

Ralf Eggeling, Nico Pfeifer *Bioinformatics* (2019-07) https://doi.org/gf5d8b DOI: 10.1093/bioinformatics/btz338 · PMID: 31510704 · PMCID: PMC6612879

19. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction** Leland McInnes, John Healy, James Melville *arXiv* (2018) https://doi.org/gqzqzn DOI: 10.48550/arxiv.1802.03426

20. **Deep-learning approach to identifying cancer subtypes using high-dimensional genomic data** Runpu Chen, Le Yang, Steve Goodison, Yijun Sun *Bioinformatics* (2019-10-11) https://doi.org/gpfzxm DOI: 10.1093/bioinformatics/btz769 · PMID: 31603461 · PMCID: PMC8215925

21. **Applications of liquid biopsies for cancer** Austin K Mattox, Chetan Bettegowda, Shibin Zhou, Nickolas Papadopoulos, Kenneth W Kinzler, Bert Vogelstein *Science Translational Medicine* (2019-08-28) https://doi.org/gjsfw9 DOI: 10.1126/scitranslmed.aay1984 · PMID: 31462507

22. **Histopathologic variables predict Oncotype DX™ Recurrence Score** Melina B Flanagan, David J Dabbs, Adam M Brufsky, Sushil Beriwal, Rohit Bhargava *Modern Pathology* (2008-03-21) https://doi.org/d27rv3 DOI: 10.1038/modpathol.2008.54 · PMID: 18360352

23. **Analysis of blood-based gene expression in idiopathic Parkinson disease** Ron Shamir, Christine Klein, David Amar, Eva-Juliane Vollstedt, Michael Bonin, Marija Usenovic, Yvette C Wong, Ales Maver, Sven Poths, Hershel Safer, … Dimitri Krainc *Neurology* (2017-09-15) https://doi.org/gcnb67 DOI: 10.1212/wnl.0000000000004516 · PMID: 28916538 · PMCID: PMC5644465

24. **Blood Transcriptional Biomarkers for Active Tuberculosis among Patients in the United States: a Case-Control Study with Systematic Cross-Classifier Evaluation** Nicholas D Walter, Mikaela A Miller, Joshua Vasquez, Marc Weiner, Adam Chapman, Melissa Engle, Michael Higgins, Amy M Quinones, Vanessa Rosselli, Elizabeth Canono, … Mark W Geraci *Journal of Clinical Microbiology* (2016-02) https://doi.org/gqzq2r DOI: 10.1128/jcm.01990-15 · PMID: 26582831 · PMCID: PMC4733166

25. **A Transcriptomic Biomarker to Quantify Systemic Inflammation in Sepsis — A Prospective Multicenter Phase II Diagnostic Study** Michael Bauer, Evangelos J Giamarellos-Bourboulis, Andreas Kortgen, Eva Möller, Karen Felsmann, Jean Marc Cavaillon, Orlando Guntinas-Lichius, Olivier Rutschmann, Andriy Ruryk, Matthias Kohl, … Konrad Reinhart *EBioMedicine* (2016-04) https://doi.org/gqzq2q DOI: 10.1016/j.ebiom.2016.03.006 · PMID: 27211554 · PMCID: PMC4856796

26. **Gene expression profiling of peripheral blood from patients with untreated new-onset systemic juvenile idiopathic arthritis reveals molecular heterogeneity that may predict macrophage activation syndrome** Ndate Fall, Michael Barnes, Sherry Thornton, Lorie Luyrink, Judyann Olson, Norman T Ilowite, Beth S Gottlieb, Thomas Griffin, David D Sherry, Susan Thompson, … Alexei A Grom *Arthritis & Rheumatism* (2007) https://doi.org/chxcfh DOI: 10.1002/art.22981 · PMID: 17968951

27. **Light-Directed, Spatially Addressable Parallel Chemical Synthesis** Stephen PA Fodor, JLeighton Read, Michael C Pirrung, Lubert Stryer, Amy Tsai Lu, Dennis Solas *Science* (1991-02-15) https://doi.org/dw6f5b DOI: 10.1126/science.1990438 · PMID: 1990438

28. **Expression monitoring by hybridization to high-density oligonucleotide arrays** David J Lockhart, Helin Dong, Michael C Byrne, Maximillian T Follettie, Michael V Gallo, Mark S Chee, Michael Mittmann, Chunwei Wang, Michiko Kobayashi, Heidi Norton, Eugene L Brown *Nature Biotechnology* (1996-12) https://doi.org/bpmwzt DOI: 10.1038/nbt1296-1675 · PMID: 9634850

29. **Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells** Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, Xuejun Liu *PLoS ONE* (2014-01-16) https://doi.org/f5tvg3 DOI: 10.1371/journal.pone.0078644 · PMID: 24454679 · PMCID: PMC3894192

30. **ImageNet classification with deep convolutional neural networks** Alex Krizhevsky, Ilya Sutskever, Geoffrey E Hinton *Communications of the ACM* (2017-05-24) https://doi.org/gbhhxs DOI: 10.1145/3065386

31. **Understanding Back-Translation at Scale** Sergey Edunov, Myle Ott, Michael Auli, David Grangier *arXiv* (2018) https://doi.org/gqzq2v DOI: 10.48550/arxiv.1808.09381

32. **Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer** Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu *arXiv* (2019-10-23) https://arxiv.org/abs/1910.10683v3

33. **U-Net: Convolutional Networks for Biomedical Image Segmentation** Olaf Ronneberger, Philipp Fischer, Thomas Brox *arXiv* (2015-05-19) https://arxiv.org/abs/1505.04597

34. **A review of feature selection methods based on mutual information** Jorge R Vergara, Pablo A Estévez *Neural Computing and Applications* (2013-03-13) https://doi.org/gj7fzd DOI: 10.1007/s00521-013-1368-0

35. **A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression** T Li, C Zhang, M Ogihara *Bioinformatics* (2004-04-15) https://doi.org/b3kzzp DOI: 10.1093/bioinformatics/bth267 · PMID: 15087314

36. **COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer** Simon A Forbes, Gurpreet Tang, Nidhi Bindal, Sally Bamford, Elisabeth Dawson, Charlotte Cole, Chai Yin Kok, Mingming Jia, Rebecca Ewing, Andrew Menzies, … PAndrew Futreal *Nucleic Acids Research* (2009-11-10) https://doi.org/fhkk8s DOI: 10.1093/nar/gkp995 · PMID: 19906727 · PMCID: PMC2808858

37. **Application of a Neural Network Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers** Jasleen K Grewal, Basile Tessier-Cloutier, Martin Jones, Sitanshu Gakkhar, Yussanne Ma, Richard Moore, Andrew J Mungall, Yongjun Zhao, Michael D Taylor, Karen Gelmon, … Steven JM Jones *JAMA Network Open* (2019-04-26) https://doi.org/gf84h2 DOI: 10.1001/jamanetworkopen.2019.2597 · PMID: 31026023 · PMCID: PMC6487574

38. **A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles** Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, …

Todd R Golub *Cell* (2017-11) https://doi.org/cgwt DOI: 10.1016/j.cell.2017.10.049 · PMID: 29195078 · PMCID: PMC5990023

39. **Gene expression inference with deep learning** Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, Xiaohui Xie *Bioinformatics* (2016-02-11) https://doi.org/f8vmtt DOI: 10.1093/bioinformatics/btw074 · PMID: 26873929 · PMCID: PMC4908320

40. **Robust clinical outcome prediction based on Bayesian analysis of transcriptional profiles and prior causal networks** Kourosh Zarringhalam, Ahmed Enayetallah, Padmalatha Reddy, Daniel Ziemek *Bioinformatics* (2014-06-11) https://doi.org/f58bp2 DOI: 10.1093/bioinformatics/btu272 · PMID: 24932007 · PMCID: PMC4058945

41. **Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes** Kourosh Zarringhalam, David Degras, Christoph Brockel, Daniel Ziemek *Scientific Reports* (2018-01-19) https://doi.org/gcwzdn DOI: 10.1038/s41598-018-19635-0 · PMID: 29352257 · PMCID: PMC5775343

42. **Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions** Yupeng Cun, Holger Fröhlich *BMC Bioinformatics* (2012-05-01) https://doi.org/f4cb5r DOI: 10.1186/1471-2105-13-69 · PMID: 22548963 · PMCID: PMC3436770

43. **Pathway-level information extractor (PLIER) for gene expression data** Weiguang Mao, Elena Zaslavsky, Boris M Hartmann, Stuart C Sealfon, Maria Chikina *Nature Methods* (2019-06-27) https://doi.org/gf75g6 DOI: 10.1038/s41592-019-0456-1 · PMID: 31249421 · PMCID: PMC7262669

44. **MultiPLIER: A Transfer Learning Framework for Transcriptomics Reveals Systemic Features of Rare Disease** Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, Casey S Greene *Cell Systems* (2019-05) https://doi.org/gf75g5 DOI: 10.1016/j.cels.2019.04.003 · PMID: 31121115 · PMCID: PMC6538307

45. **Transfer Learning for Molecular Cancer Classification Using Deep Neural Networks** Rahul K Sevakula, Vikas Singh, Nishchal K Verma, Chandan Kumar, Yan Cui *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2019-11-01) https://doi.org/gqzq3p DOI: 10.1109/tcbb.2018.2822803 · PMID: 29993662

46. **A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data** Yawen Xiao, Jun Wu, Zongli Lin, Xiaodong Zhao *Computer Methods and Programs in Biomedicine* (2018-11) https://doi.org/gfnm5c DOI: 10.1016/j.cmpb.2018.10.004 · PMID: 30415723

47. **Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF** Igor Kononenko, Edvard Šimec, Marko Robnik-Šikonja *Applied Intelligence* (1997) https://doi.org/fdm4r3 DOI: 10.1023/a:1008280620621

48. **Application of transfer learning for cancer drug sensitivity prediction** Saugato Rahman Dhruba, Raziur Rahman, Kevin Matlock, Souparno Ghosh, Ranadip Pal *BMC Bioinformatics* (2018-12) https://doi.org/gh4mnw DOI: 10.1186/s12859-018-2465-y · PMID: 30591023 · PMCID: PMC6309077

49. **A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data** Yuanyuan Li, Kai Kang, Juno M Krahn, Nicole Croutwater, Kevin Lee, David M Umbach, Leping Li *BMC Genomics* (2017-07-03)

https://doi.org/gfv37q DOI: 10.1186/s12864-017-3906-0 · PMID: 28673244 · PMCID: PMC5496318

50. **Gene expression microarray classification using PCA–BEL** Ehsan Lotfi, Azita Keshavarz *Computers in Biology and Medicine* (2014-11) https://doi.org/gqzq5p DOI: 10.1016/j.compbiomed.2014.09.008 · PMID: 25282708

51. **Using deep learning to enhance cancer diagnosis and classification** Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber *Proceedings of the international conference on machine learning* (2013)

52. **Auto-Encoding Variational Bayes** Diederik P Kingma, Max Welling *arXiv* (2014-05-02) https://arxiv.org/abs/1312.6114

53. **Higher Order Contractive Auto-Encoder** Salah Rifai, Grégoire Mesnil, Pascal Vincent, Xavier Muller, Yoshua Bengio, Yann Dauphin, Xavier Glorot *Machine Learning and Knowledge Discovery in Databases* (2011) https://doi.org/bfpkgr DOI: 10.1007/978-3-642-23783-6_41

54. **DeePathology: Deep Multi-Task Learning for Inferring Molecular Pathology from Cancer Transcriptome** Behrooz Azarkhalili, Ali Saberi, Hamidreza Chitsaz, Ali Sharifi-Zarchi *Scientific Reports* (2019-11-11) https://doi.org/gpg7vc DOI: 10.1038/s41598-019-52937-5 · PMID: 31712594 · PMCID: PMC6848155

55. **Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion** Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol *Journal of Machine Learning Research* (2010)

56. **A Deep Learning Approach For Cancer Detection And Relevant Gene Identification** Padideh Danaee, Reza Ghaeini, David A Hendrix *Biocomputing 2017* (2016-11-22) https://doi.org/gqzq5q DOI: 10.1142/9789813207813_0022 · PMID: 27896977 · PMCID: PMC5177447

57. **Exploring single-cell data with deep multitasking neural networks** Matthew Amodio, David van Dijk, Krishnan Srinivasan, William S Chen, Hussein Mohsen, Kevin R Moon, Allison Campbell, Yujiao Zhao, Xiaomei Wang, Manjunatha Venkataswamy, … Smita Krishnaswamy *Nature Methods* (2019-10-07) https://doi.org/gf9rsg DOI: 10.1038/s41592-019-0576-7 · PMID: 31591579

58. **Regularization and variable selection via the elastic net** Hui Zou, Trevor Hastie *Journal of the royal statistical society: series B (statistical methodology)* (2005)

59. **I tried a bunch of things: The dangers of unexpected overfitting in classification of brain data** Mahan Hosseini, Michael Powell, John Collins, Chloe Callahan-Flintoft, William Jones, Howard Bowman, Brad Wyble *Neuroscience & Biobehavioral Reviews* (2020-12) https://doi.org/ghkskv DOI: 10.1016/j.neubiorev.2020.09.036 · PMID: 33035522

60. **Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation** Tathiane M Malta, Artem Sokolov, Andrew J Gentles, Tomasz Burzykowski, Laila Poisson, John N Weinstein, Bożena Kamińska, Joerg Huelsken, Larsson Omberg, Olivier Gevaert, … Armaz Mariamidze *Cell* (2018-04) https://doi.org/gc93hh DOI: 10.1016/j.cell.2018.03.034 · PMID: 29625051 · PMCID: PMC5902191

61. **Massive single-cell RNA-seq analysis and imputation via deep learning** Yue Deng, Feng Bao, Qionghai Dai, Lani F Wu, Steven J Altschuler *Cold Spring Harbor Laboratory* (2018-05-06) https://doi.org/gfgrpm DOI: 10.1101/315556

62. **A global immune gene expression signature for human cancers** Yuexin Liu *Oncotarget* (2019-03-08) https://doi.org/gqzq8j DOI: 10.18632/oncotarget.26773 · PMID: 30956779 · PMCID: PMC6443003

63. **A Four-Biomarker Blood Signature Discriminates Systemic Inflammation Due to Viral Infection Versus Other Etiologies** DL Sampson, BA Fox, TD Yager, S Bhide, S Cermelli, LC McHugh, TA Seldon, RA Brandon, E Sullivan, JJ Zimmerman, … RB Brandon *Scientific Reports* (2017-06-06) https://doi.org/gc4zdw DOI: 10.1038/s41598-017-02325-8 · PMID: 28588308 · PMCID: PMC5460227

64. **Multitask learning improves prediction of cancer drug sensitivity** Han Yuan, Ivan Paskov, Hristo Paskov, Alvaro J González, Christina S Leslie *Scientific Reports* (2016-08-23) https://doi.org/f8zbhk DOI: 10.1038/srep31619 · PMID: 27550087 · PMCID: PMC4994023

65. **Don't Rule Out Simple Models Prematurely: A Large Scale Benchmark Comparing Linear and Non-linear Classifiers in OpenML** Benjamin Strang, Peter van der Putten, Jan N van Rijn, Frank Hutter *Advances in Intelligent Data Analysis XVII* (2018) https://doi.org/gqzq6q DOI: 10.1007/978-3-030-01768-2_25

66. **Does deep learning always outperform simple linear regression in optical imaging?** Shuming Jiao, Yang Gao, Jun Feng, Ting Lei, Xiaocong Yuan *arXiv* (2020-02-19) https://arxiv.org/abs/1911.00353 DOI: 10.1364/oe.382319

67. **Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set** Eelke B Lenselink, Niels ten Dijke, Brandon Bongers, George Papadatos, Herman WT van Vlijmen, Wojtek Kowalczyk, Adriaan P IJzerman, Gerard JP van Westen *Journal of Cheminformatics* (2017-08-14) https://doi.org/gbwq98 DOI: 10.1186/s13321-017-0232-0 · PMID: 29086168 · PMCID: PMC5555960

68. **Benchmarking deep learning models on large healthcare datasets** Sanjay Purushotham, Chuizheng Meng, Zhengping Che, Yan Liu *Journal of Biomedical Informatics* (2018-07) https://doi.org/gd97qc DOI: 10.1016/j.jbi.2018.04.007 · PMID: 29879470

69. **Citation Indexes for Science** Eugene Garfield *Science* (1955-07-15) https://doi.org/fnkc4f DOI: 10.1126/science.122.3159.108 · PMID: 14385826

70. **The science of science** Dashun Wang, Albert-László Barabási *Cambridge University Press* (2021) ISBN: 9781108492669

71. **An index to quantify an individual's scientific research output** JE Hirsch *Proceedings of the National Academy of Sciences* (2005-11-07) https://doi.org/cbq6dz DOI: 10.1073/pnas.0507655102 · PMID: 16275915 · PMCID: PMC1283832

72. **The h-index Debate: An Introduction for Librarians** Cameron Barnes *The Journal of Academic Librarianship* (2017-11) https://doi.org/gcjpk2 DOI: 10.1016/j.acalib.2017.08.013

73. **The h-index is no longer an effective correlate of scientific reputation** Vladlen Koltun, David Hafner *PLOS ONE* (2021-06-28) https://doi.org/gkzfnr DOI: 10.1371/journal.pone.0253397 · PMID: 34181681 · PMCID: PMC8238192

74. **Theory and practise of the g-index** Leo Egghe *Scientometrics* (2006-10) https://doi.org/dgj3tc DOI: 10.1007/s11192-006-0144-7

75. **New Tools for Improving and Evaluating The Effectiveness of Research** Irving H Sher, Eugene Garfield *Research Program Effectiveness* (1965-06-27)

76. **Eigenfactor: Measuring the value and prestige of scholarly journals** Carl Bergstrom *College & Research Libraries News* (2007-05-01) https://doi.org/gf24tg DOI: 10.5860/crln.68.5.7804

77. **A systematic empirical comparison of different approaches for normalizing citation impact indicators** Ludo Waltman, Nees Jan van Eck *Journal of Informetrics* (2013-10) https://doi.org/f5jdr5 DOI: 10.1016/j.joi.2013.08.002

78. **The PageRank Citation Ranking: Bringing Order to the Web.** Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd *Stanford InfoLab* (1999)

79. **Maps of random walks on complex networks reveal community structure** Martin Rosvall, Carl T Bergstrom *Proceedings of the National Academy of Sciences* (2008-01-29) https://doi.org/fw5xcm DOI: 10.1073/pnas.0706851105 · PMID: 18216267 · PMCID: PMC2234100

80. **Large teams develop and small teams disrupt science and technology** Lingfei Wu, Dashun Wang, James A Evans *Nature* (2019-02) https://doi.org/gfvnb9 DOI: 10.1038/s41586-019-0941-9 · PMID: 30760923

81. **Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers** Lutz Bornmann, Sitaram Devarakonda, Alexander Tekles, George Chacko *Quantitative Science Studies* (2020-08) https://doi.org/gq2ts5 DOI: 10.1162/qss_a_00068

82. **Tradition and Innovation in Scientists' Research Strategies** Jacob G Foster, Andrey Rzhetsky, James A Evans *American Sociological Review* (2015-09-01) https://doi.org/f7tzm5 DOI: 10.1177/0003122415601618

83. **Bias against novelty in science: A cautionary tale for users of bibliometric indicators** Jian Wang, Reinhilde Veugelers, Paula Stephan *Research Policy* (2017-10) https://doi.org/gb22vw DOI: 10.1016/j.respol.2017.06.006

84. **Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level** BIan Hutchins, Xin Yuan, James M Anderson, George M Santangelo *PLOS Biology* (2016-09-06) https://doi.org/f88zk2 DOI: 10.1371/journal.pbio.1002541 · PMID: 27599104 · PMCID: PMC5012559

85. **Measuring contextual citation impact of scientific journals** Henk F Moed *Journal of Informetrics* (2010-07) https://doi.org/dpbgj9 DOI: 10.1016/j.joi.2010.01.002

86. **Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers** Yongjun Zhang, Jialin Ma, Zijian Wang, Bolun Chen, Yongtao Yu *Scientometrics* (2017-12-23) https://doi.org/gc4b2s DOI: 10.1007/s11192-017-2626-1

87. **Quantifying and suppressing ranking bias in a large citation network** Giacomo Vaccario, Matus Medo, Nicolas Wider, Manuel Sebastian Mariani *arXiv* (2017-08-30) https://arxiv.org/abs/1703.08071 DOI: 10.1016/j.joi.2017.05.014

88. **Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations** Ivan Heibi, Silvio Peroni, David Shotton *Scientometrics* (2019-09-14) https://doi.org/ggzz8b DOI: 10.1007/s11192-019-03217-6

89. **Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks** S Maslov, S Redner *Journal of Neuroscience* (2008-10-29) https://doi.org/fsfh8w DOI: 10.1523/jneurosci.0002-08.2008 · PMID: 18971452 · PMCID: PMC6671494

90. **Standardizing the Evaluation of Scientific and Academic Performance in Neurosurgery—Critical Review of the “h” Index and its Variants** Salah G Aoun, Bernard R Bendok, Rudy J Rahme, Ralph G Dacey Jr., HHunt Batjer *World Neurosurgery* (2013-11) https://doi.org/fxtz98 DOI: 10.1016/j.wneu.2012.01.052 · PMID: 22381859

91. **Understanding The Limitations Of The Journal Impact Factor** Andrew P Kurmis *The Journal of Bone and Joint Surgery-American Volume* (2003-12) https://doi.org/gh6fph DOI: 10.2106/00004623-200312000-00028 · PMID: 14668520

92. **Why trust science?** Naomi Oreskes *Princeton University Press* (2021) ISBN: 9780691212265

93. **Setting the default to reproducible** Victoria Stodden, Jonathan Borwein, David H Bailey *ICERM Workshop on Reproducibility in Computational and Experimental Mathematics* (2013)

94. **Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist** Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, … Atul J Butte *Nature Medicine* (2020-09) https://doi.org/ghfzhk DOI: 10.1038/s41591-020-1041-y · PMID: 32908275 · PMCID: PMC7538196

95. **MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care** Tina Hernandez-Boussard, Selen Bozkurt, John PA Ioannidis, Nigam H Shah *Journal of the American Medical Informatics Association* (2020-06-28) https://doi.org/gmns84 DOI: 10.1093/jamia/ocaa088 · PMID: 32594179 · PMCID: PMC7727333

96. **Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers** John Mongan, Linda Moy, Charles E Kahn Jr *Radiology: Artificial Intelligence* (2020-03-01) https://doi.org/gg9f65 DOI: 10.1148/ryai.2020200029 · PMID: 33937821 · PMCID: PMC8017414

97. **Sharing biological data: why, when, and how** Samantha L Wilson, Gregory P Way, Wout Bittremieux, Jean‐Paul Armache, Melissa A Haendel, Michael M Hoffman *FEBS Letters* (2021-04) https://doi.org/gmmq7d DOI: 10.1002/1873-3468.14067 · PMID: 33843054

98. **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository** R Edgar *Nucleic Acids Research* (2002-01-01) https://doi.org/fttpkn DOI: 10.1093/nar/30.1.207 · PMID: 11752295 · PMCID: PMC99122

99. **A call for public archives for biological image data** Jan Ellenberg, Jason R Swedlow, Mary Barlow, Charles E Cook, Ugis Sarkans, Ardan Patwardhan, Alvis Brazma, Ewan Birney *Nature Methods* (2018-10-30) https://doi.org/gfgphs DOI: 10.1038/s41592-018-0195-8 · PMID: 30377375 · PMCID: PMC6884425

100. **The Kipoi repository accelerates community exchange and reuse of predictive models for genomics** Žiga Avsec, Roman Kreuzhuber, Johnny Israeli, Nancy Xu, Jun Cheng, Avanti Shrikumar, Abhimanyu Banerjee, Daniel S Kim, Thorsten Beier, Lara Urban, … Julien Gagneur *Nature Biotechnology* (2019-05-28) https://doi.org/gf3fmq DOI: 10.1038/s41587-019-0140-0 · PMID: 31138913 · PMCID: PMC6777348

101. **Sfaira accelerates data and model reuse in single cell genomics** David S Fischer, Leander Dony, Martin König, Abdul Moeed, Luke Zappia, Lukas Heumos, Sophie Tritschler, Olle Holmberg, Hananeh Aliee, Fabian J Theis *Genome Biology* (2021-08-25) https://doi.org/gq8drj DOI: 10.1186/s13059-021-02452-6 · PMID: 34433466 · PMCID: PMC8386039

102. **Docker: lightweight Linux containers for consistent development and deployment** Dirk Merkel *Linux Journal* (2014)

103. **Snakemake--a scalable bioinformatics workflow engine** J Koster, S Rahmann *Bioinformatics* (2012-08-20) https://doi.org/gd2xzq DOI: 10.1093/bioinformatics/bts480 · PMID: 22908215

104. **Nextflow enables reproducible computational workflows** Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, Cedric Notredame *Nature Biotechnology* (2017-04) https://doi.org/gfj52z DOI: 10.1038/nbt.3820 · PMID: 28398311

105. **Responsible, practical genomic data sharing that accelerates research** James Brian Byrd, Anna C Greene, Deepashree Venkatesh Prasad, Xiaoqian Jiang, Casey S Greene *Nature Reviews Genetics* (2020-07-21) https://doi.org/gg7c57 DOI: 10.1038/s41576-020-0257-5 · PMID: 32694666 · PMCID: PMC7974070

106. **Extracting Training Data from Large Language Models** Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, … Colin Raffel *arXiv* (2021-06-16) https://arxiv.org/abs/2012.07805

107. **Deep Learning with Differential Privacy** Martin Abadi, Andy Chu, Ian Goodfellow, HBrendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (2016-10-24) https://doi.org/gcrnp3 DOI: 10.1145/2976749.2978318

108. **Top considerations for creating bioinformatics software documentation** Mehran Karimzadeh, Michael M Hoffman *Briefings in Bioinformatics* (2017-01-14) https://doi.org/bzmp DOI: 10.1093/bib/bbw134 · PMID: 28088754 · PMCID: PMC6054259

109. **Bioconductor: open software development for computational biology and bioinformatics** Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff

Gentry, … Jianhua Zhang *Genome Biology* (2004) https://doi.org/c2xm5v DOI: 10.1186/gb-2004-5-10-r80 · PMID: 15461798 · PMCID: PMC545600

110. **Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes** Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, … Philip S Bernard *Journal of Clinical Oncology* (2009-03-10) https://doi.org/c2688w DOI: 10.1200/jco.2008.18.1370 · PMID: 19204204 · PMCID: PMC2667820

111. **Gene Expression Profiling for the Identification and Classification of Antibody-Mediated Heart Rejection** Alexandre Loupy, Jean Paul Duong Van Huyen, Luis Hidalgo, Jeff Reeve, Maud Racapé, Olivier Aubert, Jeffery M Venner, Konrad Falmuski, Marie Cécile Bories, Thibaut Beuscart, … Philip F Halloran *Circulation* (2017-03-07) https://doi.org/f9vfvw DOI: 10.1161/circulationaha.116.022907 · PMID: 28148598

112. **Large-scale labeling and assessment of sex bias in publicly available expression data** Emily Flynn, Annie Chang, Russ B Altman *BMC bioinformatics* (2021-03-30) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8011224/ DOI: 10.1186/s12859-021-04070-2 · PMID: 33784977 · PMCID: PMC8011224

113. **Compute Trends Across Three Eras of Machine Learning** Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Pablo Villalobos *arXiv* (2022-03-11) https://arxiv.org/abs/2202.05924

114. **Massive mining of publicly available RNA-seq data from human and mouse** Alexander Lachmann, Denis Torre, Alexandra B Keenan, Kathleen M Jagodnik, Hoyjin J Lee, Lily Wang, Moshe C Silverstein, Avi Ma'ayan *Nature Communications* (2018-04-10)

https://doi.org/gc92dr DOI: 10.1038/s41467-018-03751-6 · PMID: 29636450 · PMCID: PMC5893633

115. **A curated database reveals trends in single-cell transcriptomics** Valentine Svensson, Eduardo da Veiga Beltrame, Lior Pachter *Database* (2020) https://doi.org/gjnr3h DOI: 10.1093/database/baaa073 · PMID: 33247933 · PMCID: PMC7698659

116. **Bias-invariant RNA-sequencing metadata annotation** Hannes Wartmann, Sven Heins, Karin Kloiber, Stefan Bonn *GigaScience* (2021-09) https://doi.org/gph9xp DOI: 10.1093/gigascience/giab064 · PMID: 34553213 · PMCID: PMC8559615

117. **Improved prediction of smoking status via isoform-aware RNA-seq deep learning models** Zifeng Wang, Aria Masoomi, Zhonghui Xu, Adel Boueiz, Sool Lee, Tingting Zhao, Russell Bowler, Michael Cho, Edwin K Silverman, Craig Hersh, … Peter J Castaldi *PLOS Computational Biology* (2021-10-11) https://doi.org/gph9xq DOI: 10.1371/journal.pcbi.1009433 · PMID: 34634029 · PMCID: PMC8530282

118. **The evolution of gene expression and the transcriptome–phenotype relationship** Peter W Harrison, Alison E Wright, Judith E Mank *Seminars in Cell & Developmental Biology* (2012-04) https://doi.org/fxqd2g DOI: 10.1016/j.semcdb.2011.12.004 · PMID: 22210502 · PMCID: PMC3378502

119. **Nonlinear Dynamics in Gene Regulation Promote Robustness and Evolvability of Gene Expression Levels** Arno Steinacher, Declan G Bates, Ozgur E Akman, Orkun S Soyer *PLOS ONE* (2016-04-15) https://doi.org/f8xrrq DOI: 10.1371/journal.pone.0153295 · PMID: 27082741 · PMCID: PMC4833316

120. **ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions** Jie Tan, John H Hammond, Deborah A Hogan, Casey S Greene *mSystems* (2016-02-23) https://doi.org/gcgmbq DOI: 10.1128/msystems.00025-15 · PMID: 27822512 · PMCID: PMC5069748

121. **A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data** Tianyu Kang, Wei Ding, Luoyan Zhang, Daniel Ziemek, Kourosh Zarringhalam *BMC Bioinformatics* (2017-12) https://doi.org/gf8cm6 DOI: 10.1186/s12859-017-1984-2 · PMID: 29258445 · PMCID: PMC5735940

122. **Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data** Aaron M Smith, Jonathan R Walsh, John Long, Craig B Davis, Peter Henstock, Martin R Hodge, Mateusz Maciejewski, Xinmeng Jasmine Mu, Stephen Ra, Shanrong Zhao, … Charles K Fisher *BMC Bioinformatics* (2020-03-20) https://doi.org/ggpc9d DOI: 10.1186/s12859-020-3427-8 · PMID: 32197580 · PMCID: PMC7085143

123. **A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models** Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, Ben Van Calster *Journal of Clinical Epidemiology* (2019-06) https://doi.org/gfzstd DOI: 10.1016/j.jclinepi.2019.02.004 · PMID: 30763612

124. **Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets** Marc-Andre Schulz, BTThomas Yeo,

Joshua T Vogelstein, Janaina Mourao-Miranada, Jakob N Kather, Konrad Kording, Blake Richards, Danilo Bzdok *Nature Communications* (2020-08-25) https://doi.org/gg9njw DOI: 10.1038/s41467-020-18037-z · PMID: 32843633 · PMCID: PMC7447816

125. **The Genotype-Tissue Expression (GTEx) project** John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, … Helen F Moore *Nature Genetics* (2013-05-29) https://doi.org/gd5z68 DOI: 10.1038/ng.2653 · PMID: 23715323 · PMCID: PMC4010069

126. **recount3: summaries and queries for large-scale RNA-seq expression and splicing** Christopher Wilks, Shijie C Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P Ling, Eddie Luidy Imada, David Zhang, Lance Joseph, Jeffrey T Leek, … Ben Langmead *Genome Biology* (2021-11-29) https://doi.org/gnm7zc DOI: 10.1186/s13059-021-02533-6 · PMID: 34844637 · PMCID: PMC8628444

127. **Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics** Qiwen Hu, Casey S Greene *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2019) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417816/ PMID: 30963075 · PMCID: PMC6417816

128. **Large-scale labeling and assessment of sex bias in publicly available expression data** Emily Flynn, Annie Chang, Russ B Altman *BMC Bioinformatics* (2021-03-30) https://doi.org/gpjt3n DOI: 10.1186/s12859-021-04070-2 · PMID: 33784977 · PMCID: PMC8011224

129. **The Sequence Read Archive** R Leinonen, H Sugawara, M Shumway *Nucleic Acids Research* (2010-11-09) https://doi.org/c652z5 DOI: 10.1093/nar/gkq1019 · PMID: 21062823 · PMCID: PMC3013647

130. **An efficient not-only-linear correlation coefficient based on machine learning** Milton Pividori, Marylyn D Ritchie, Diego H Milone, Casey S Greene *Cold Spring Harbor Laboratory* (2022-06-17) https://doi.org/gqcvbw DOI: 10.1101/2022.06.15.496326

131. **limma powers differential expression analyses for RNA-sequencing and microarray studies** Matthew E Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, Gordon K Smyth *Nucleic Acids Research* (2015-01-20) https://doi.org/f7c4n5 DOI: 10.1093/nar/gkv007 · PMID: 25605792 · PMCID: PMC4402510

132. **Navigating the pitfalls of applying machine learning in genomics** Sean Whalen, Jacob Schreiber, William S Noble, Katherine S Pollard *Nature Reviews Genetics* (2021-11-26) https://doi.org/gnm4r9 DOI: 10.1038/s41576-021-00434-9 · PMID: 34837041

133. **Phosphorylation of ETS transcription factor ER81 in a complex with its coactivators CREB-binding protein and p300** S Papoutsopoulou, R Janknecht *Molecular and cellular biology* (2000-10) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC86284/ DOI: 10.1128/mcb.20.19.7300-7310.2000 · PMID: 10982847 · PMCID: PMC86284

134. **BioMart--biological queries made easy** Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, Arek Kasprzyk *BMC*

genomics (2009-01-14) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2649164/ DOI: 10.1186/1471-2164-10-22 · PMID: 19144180 · PMCID: PMC2649164

135. **The European Nucleotide Archive** Rasko Leinonen, Ruth Akhtar, Ewan Birney, Lawrence Bower, Ana Cerdeno-Tárraga, Ying Cheng, Iain Cleland, Nadeem Faruque, Neil Goodgame, Richard Gibson, … Guy Cochrane *Nucleic acids research* (2011-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3013801/ DOI: 10.1093/nar/gkq967 · PMID: 20972220 · PMCID: PMC3013801

136. **Rectified linear units improve restricted boltzmann machines** Vinod Nair, Geoffrey E Hinton *Proceedings of the 27th International Conference on International Conference on Machine Learning* (2010-06-21) https://dl.acm.org/doi/10.5555/3104322.3104425 ISBN: 9781605589077

137. **PyTorch: An Imperative Style, High-Performance Deep Learning Library** Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, … Soumith Chintala *arXiv* (2019-12-05) https://arxiv.org/abs/1912.01703

138. **Adam: A Method for Stochastic Optimization** Diederik P Kingma, Jimmy Ba *arXiv* (2017-01-31) https://arxiv.org/abs/1412.6980

139. **Dropout: A Simple Way to Prevent Neural Networks from Overfitting** Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov *Journal of Machine Learning Research* (2014) http://jmlr.org/papers/v15/srivastava14a.html

140. **Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift** Sergey Ioffe, Christian Szegedy *Proceedings of the 32nd International Conference on Machine Learning* (2015-06-01) https://proceedings.mlr.press/v37/ioffe15.html

141. **Neptune: Experiment management and collaboration tool** neptune.ai (2020) https://neptune.ai

142. **Reproducibility standards for machine learning in the life sciences** Benjamin J Heil, Michael M Hoffman, Florian Markowetz, Su-In Lee, Casey S Greene, Stephanie C Hicks *Nature Methods* (2021-08-30) https://doi.org/gmnnqh DOI: 10.1038/s41592-021-01256-7 · PMID: 34462593 · PMCID: PMC9131851

143. **High throughput analysis of differential gene expression** John P Carulli, Michael Artinger, Pamela M Swain, Colleen D Root, Linda Chee, Craig Tulig, Jennifer Guerin, Mark Osborne, Gary Stein, Jane Lian, Peter T Lomedico *Journal of Cellular Biochemistry* (1998) https://doi.org/dcz39r DOI: 10.1002/(sici)1097-4644(1998)72:30/31+<286::aid-jcb35>3.0.co;2-d

144. **Differential expression analysis for sequence count data** Simon Anders, Wolfgang Huber *Genome Biology* (2010-10) https://doi.org/btmbk5 DOI: 10.1186/gb-2010-11-10-r106 · PMID: 20979621 · PMCID: PMC3218662

145. **RNA-Seq differential expression analysis: An extended review and a software tool** Juliana Costa-Silva, Douglas Domingues, Fabricio Martins Lopes *PLOS ONE* (2017-12-21) https://doi.org/gf4p49 DOI: 10.1371/journal.pone.0190152 · PMID: 29267363 · PMCID: PMC5739479

146. **Statistical approaches for differential expression analysis in metatranscriptomics** Yancong Zhang, Kelsey N Thompson, Curtis Huttenhower, Eric A Franzosa *Bioinformatics* (2021-07-01) https://doi.org/grcgxx DOI: 10.1093/bioinformatics/btab327 · PMID: 34252963 · PMCID: PMC8275336

147. **Analysis of a complex of statistical variables into principal components.** H Hotelling *Journal of Educational Psychology* (1933-09) https://doi.org/fb5435 DOI: 10.1037/h0071325

148. **Visualizing data using t-SNE.** Geoffrey Hinton, Laurens Van der Maaten *Journal of machine learning research* (2008)

149. **UMAP: Uniform Manifold Approximation and Projection** Leland McInnes, John Healy, Nathaniel Saul, Lukas Großberger *Journal of Open Source Software* (2018-09-02) https://doi.org/gf6k3s DOI: 10.21105/joss.00861

150. **Clustering Algorithms: Their Application to Gene Expression Data** Jelili Oyelade, Itunuoluwa Isewon, Funke Oladipupo, Olufemi Aromolaran, Efosa Uwoghiren, Faridah Ameh, Moses Achas, Ezekiel Adebiyi *Bioinformatics and Biology Insights* (2016-01) https://doi.org/gm72kz DOI: 10.4137/bbi.s38316 · PMID: 27932867 · PMCID: PMC5135122

151. **Single-cell transcriptional profiles in human skeletal muscle** Aliza B Rubenstein, Gregory R Smith, Ulrika Raue, Gwénaëlle Begue, Kiril Minchev, Frederique Ruf-Zamojski, Venugopalan D Nair, Xingyu Wang, Lan Zhou, Elena Zaslavsky, … Stuart C Sealfon *Scientific Reports* (2020-01-14) https://doi.org/ggr7ms DOI: 10.1038/s41598-019-57110-6 · PMID: 31937892 · PMCID: PMC6959232

152. **Pyramidal neuron subtype diversity governs microglia states in the neocortex** Jeffrey A Stogsdill, Kwanho Kim, Loïc Binan, Samouil L Farhi, Joshua Z Levin, Paola Arlotta *Nature* (2022-08-10) https://doi.org/gqpg36 DOI: 10.1038/s41586-022-05056-7 · PMID: 35948630

153. **Single nucleus transcriptome and chromatin accessibility of postmortem human pituitaries reveal diverse stem cell regulatory mechanisms** Zidong Zhang, Michel Zamojski, Gregory R Smith, Thea L Willis, Val Yianni, Natalia Mendelev, Hanna Pincas, Nitish Seenarine, Mary Anne S Amper, Mital Vasoya, … Frederique Ruf-Zamojski *Cell Reports* (2022-03) https://doi.org/grcg62 DOI: 10.1016/j.celrep.2022.110467 · PMID: 35263594 · PMCID: PMC8957708

154. **Integrative Analysis Identifies Candidate Tumor Microenvironment and Intracellular Signaling Pathways that Define Tumor Heterogeneity in NF1** Jineta Banerjee, Robert J Allaway, Jaclyn N Taroni, Aaron Baker, Xiaochun Zhang, Chang In Moon, Christine A Pratilas, Jaishri O Blakeley, Justin Guinney, Angela Hirbe, … Sara JC Gosline *Genes* (2020-02-21) https://doi.org/gg4rbj DOI: 10.3390/genes11020226 · PMID: 32098059 · PMCID: PMC7073563

155. **Transcriptomic profiling of microglia and astrocytes throughout aging** Jie Pan, Nana Ma, Bo Yu, Wei Zhang, Jun Wan *Journal of Neuroinflammation* (2020-04-01) https://doi.org/gjnx7s DOI: 10.1186/s12974-020-01774-9 · PMID: 32238175 · PMCID: PMC7115095

156. **Spatiotemporal expression and transcriptional perturbations by long noncoding RNAs in the mouse brain** Loyal A Goff, Abigail F Groff, Martin Sauvageau, Zachary Trayes-Gibson, Diana B Sanchez-Gomez, Michael Morse, Ryan D Martin, Lara

E Elcavage, Stephen C Liapis, Meryem Gonzalez-Celeiro, … John L Rinn *Proceedings of the National Academy of Sciences* (2015-06) https://doi.org/f7fzxj DOI: 10.1073/pnas.1411263112 · PMID: 26034286 · PMCID: PMC4460505

157. **Ensembl 2021** Kevin L Howe, Premanand Achuthan, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, MRidwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, … Paul Flicek *Nucleic Acids Research* (2020-11-02) https://doi.org/gk68s9 DOI: 10.1093/nar/gkaa942 · PMID: 33137190 · PMCID: PMC7778975

158. **CellMarker: a manually curated resource of cell markers in human and mouse** Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, … Yun Xiao *Nucleic Acids Research* (2018-10-05) https://doi.org/ggnktb DOI: 10.1093/nar/gky900 · PMID: 30289549 · PMCID: PMC6323899

159. **The reactome pathway knowledgebase 2022** Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, … Peter D'Eustachio *Nucleic Acids Research* (2021-11-12) https://doi.org/gpm2r5 DOI: 10.1093/nar/gkab1028 · PMID: 34788843 · PMCID: PMC8689983

160. **Scikit-learn: Machine learning in Python** Fabian Pendregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion *The Journal of Machine Learning Research* (2011)

161. **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing** Yoav Benjamini, Yosef Hochberg *Journal of the Royal Statistical*

Society: Series B (Methodological) (1995-01) https://doi.org/gfpkdx DOI: 10.1111/j.2517-6161.1995.tb02031.x

162. **ADAGE signature analysis: differential expression analysis with data-defined gene sets** Jie Tan, Matthew Huyck, Dongbo Hu, René A Zelaya, Deborah A Hogan, Casey S Greene *BMC Bioinformatics* (2017-11-22) https://doi.org/gg7m57 DOI: 10.1186/s12859-017-1905-4 · PMID: 29166858 · PMCID: PMC5700673

163. **The Effects of Nonlinear Signal on Expression-Based Prediction Performance** Benjamin J Heil, Jake Crawford, Casey S Greene *Cold Spring Harbor Laboratory* (2022-06-26) https://doi.org/gqhgwk DOI: 10.1101/2022.06.22.497194

164. **Impact Factor Distortions** Bruce Alberts *Science* (2013-05-17) https://doi.org/mjm DOI: 10.1126/science.1240319 · PMID: 23687012

165. **Citation patterns in economics and beyond** Matthias Aistleitner, Jakob Kapeller, Stefan Steinerberger *Science in Context* (2019-12) https://doi.org/gq62s8 DOI: 10.1017/s0269889720000022 · PMID: 32202238

166. **Disruptive papers published in Scientometrics: meaningful results by using an improved variant of the disruption index originally proposed by Wu, Wang, and Evans (2019)** Lutz Bornmann, Sitaram Devarakonda, Alexander Tekles, George Chacko *Scientometrics* (2020-03-14) https://doi.org/ggzzxd DOI: 10.1007/s11192-020-03406-8

167. **Quantifying and suppressing ranking bias in a large citation network** Giacomo Vaccario, Matúš Medo, Nicolas Wider, Manuel Sebastian Mariani *Journal of Informetrics* (2017-08) https://doi.org/gbzjdh DOI: 10.1016/j.joi.2017.05.014

168. **Quantitative evaluation of alternative field normalization procedures** Yunrong Li, Filippo Radicchi, Claudio Castellano, Javier Ruiz-Castillo *Journal of Informetrics* (2013-07) https://doi.org/f48tvt DOI: 10.1016/j.joi.2013.06.001

169. **When a journal is both at the 'top' and the 'bottom': the illogicality of conflating citation-based metrics with quality** Shannon Mason, Lenandlar Singh *Scientometrics* (2022-05-25) https://doi.org/gq2468 DOI: 10.1007/s11192-022-04402-w

170. **Field delineation using medical subject headings (MeSH)-An alternative way to aggregate data in the web of science** Håkan Carlsson, Ed CM Noyons *12th International Conference on Scientometrics and Informetrics* (2009)

171. **Cited references and Medical Subject Headings (MeSH) as two different knowledge representations: clustering and mappings at the paper level** Loet Leydesdorff, Jordan A Comins, Aaron A Sorensen, Lutz Bornmann, Iina Hellsten *Scientometrics* (2016-10-08) https://doi.org/gc8zk4 DOI: 10.1007/s11192-016-2119-7 · PMID: 27942085 · PMCID: PMC5124055

172. **A robust DNA mechanical device controlled by hybridization topology** Hao Yan, Xiaoping Zhang, Zhiyong Shen, Nadrian C Seeman *Nature* (2002-01) https://doi.org/czh8hg DOI: 10.1038/415062a · PMID: 11780115

173. **Bioadhesive poly(methyl methacrylate) microdevices for controlled drug delivery** Sarah L Tao, Michael W Lubeley, Tejal A Desai *Journal of Controlled Release* (2003-03) https://doi.org/c7fpg4 DOI: 10.1016/s0168-3659(03)00005-1 · PMID: 12628329

174. **DNA-Templated Self-Assembly of Protein Arrays and Highly Conductive Nanowires** Hao Yan, Sung Ha Park, Gleb Finkelstein, John H Reif, Thomas H LaBean *Science* (2003-09-26) https://doi.org/bfgvgf DOI: 10.1126/science.1089389 · PMID: 14512621

175. **Photostable Luminescent Nanoparticles as Biological Label for Cell Recognition of System Lupus Erythematosus Patients** Xiaoxiao He, Kemin Wang, Weihong Tan, Jun Li, Xiaohai Yang, Shasheng Huang, Dan Xiao *Journal of Nanoscience and Nanotechnology* (2002-07-01) https://doi.org/dcj5cg DOI: 10.1166/jnn.2002.105 · PMID: 12908257

176. **WSXM: A software for scanning probe microscopy and a tool for nanotechnology** I Horcas, R Fernández, JM Gómez-Rodríguez, J Colchero, J Gómez-Herrero, AM Baro *Review of Scientific Instruments* (2007-01) DOI: 10.1063/1.2432410

177. **Measuring Distances in Supported Bilayers by Fluorescence Interference-Contrast Microscopy: Polymer Supports and SNARE Proteins** Volker Kiessling, Lukas K Tamm *Biophysical Journal* (2003-01) https://doi.org/dqsg2c DOI: 10.1016/s0006-3495(03)74861-9 · PMID: 12524294 · PMCID: PMC1302622

178. **Toward fluorescence nanoscopy** Stefan W Hell *Nature Biotechnology* (2003-10-31) https://doi.org/dnzt3b DOI: 10.1038/nbt895 · PMID: 14595362

179. **In Vivo Imaging of Quantum Dots Encapsulated in Phospholipid Micelles** Benoit Dubertret, Paris Skourides, David J Norris, Vincent Noireaux, Ali H Brivanlou, Albert Libchaber *Science* (2002-11-29) https://doi.org/dd6sqp DOI: 10.1126/science.1077194 · PMID: 12459582

180. **Water-Soluble Quantum Dots for Multiphoton Fluorescence Imaging in Vivo** Daniel R Larson, Warren R Zipfel, Rebecca M Williams, Stephen W Clark, Marcel P Bruchez, Frank W Wise, Watt W Webb *Science* (2003-05-30) https://doi.org/cn9j76 DOI: 10.1126/science.1083780 · PMID: 12775841

181. **Immunoelectron microscopic exploration of the Golgi complex.** JW Slot, HJ Geuze *Journal of Histochemistry & Cytochemistry* (1983-08) https://doi.org/dxxzxg DOI: 10.1177/31.8.6863900 · PMID: 6863900

182. **Immunocytochemical and electrophoretic analyses of changes in myosin gene expression in cat posterior temporalis muscle during postnatal development** JFY Hoh, S Hughes, C Chow, PT Hale, RB Fitzsimons *Journal of Muscle Research and Cell Motility* (1988-02) https://doi.org/d72278 DOI: 10.1007/bf01682147 · PMID: 3392187

183. **Electron microscopic demonstration of calcitonin in human medullary carcinoma of thyroid by the immuno gold staining method** J Dämmrich, W Ormanns, R Schäffer *Histochemistry* (1984) https://doi.org/ct253c DOI: 10.1007/bf00514331 · PMID: 6511490

184. **Grafting genetically modified cells into the rat brain: characteristics of E. coli β-galactosidase as a reporter gene** S Shimohama, MB Rosenberg, AM Fagan, JA Wolff, MP Short, XO Breakefield, T Friedmann, FH Gage *Molecular Brain Research* (1989-06) https://doi.org/dptnbm DOI: 10.1016/0169-328x(89)90061-2 · PMID: 2501628

185. **Vitamin-D-dependent calcium-binding-protein and parvalbumin occur in bones and teeth** MR Celio, AW Norman, CW Heizmann *Calcified Tissue International* (1984-12) https://doi.org/fdnfdg DOI: 10.1007/bf02405306 · PMID: 6423230

186. **Mapping of brain areas containing RNA homologous to cDNAs encoding the alpha and beta subunits of the rat GABAA gamma-aminobutyrate receptor.** JM Séquier, JG Richards, P Malherbe, GW Price, S Mathews, H Möhler *Proceedings of the National Academy of Sciences* (1988-10) https://doi.org/fv2p49 DOI: 10.1073/pnas.85.20.7815 · PMID: 2845424 · PMCID: PMC282284

187. **Studies of the HER-2/neu Proto-Oncogene in Human Breast and Ovarian Cancer** Dennis J Slamon, William Godolphin, Lovell A Jones, John A Holt, Steven G Wong, Duane E Keith, Wendy J Levin, Susan G Stuart, Judy Udove, Axel Ullrich, Michael F Press *Science* (1989-05-12) https://doi.org/cngtqx DOI: 10.1126/science.2470152 · PMID: 2470152

188. **Expression of c-fos Protein in Brain: Metabolic Mapping at the Cellular Level** SM Sagar, FR Sharp, T Curran *Science* (1988-06-03) https://doi.org/b39h2t DOI: 10.1126/science.3131879 · PMID: 3131879

189. **Proliferating cell nuclear antigen (PCNA) immunolocalization in paraffin sections: An index of cell proliferation with evidence of deregulated expression in some, neoplasms** PA Hall, DA Levison, AL Woods, CC-W Yu, DB Kellock, JA Watkins, DM Barnes, CE Gillett, R Camplejohn, R Dover, … DP Lane *The Journal of Pathology* (1990-12) https://doi.org/cntmbr DOI: 10.1002/path.1711620403 · PMID: 1981239

190. **Proteomics Standards Initiative: Fifteen Years of Progress and Future Work** Eric W Deutsch, Sandra Orchard, Pierre-Alain Binz, Wout Bittremieux, Martin Eisenacher, Henning Hermjakob, Shin Kawano, Henry Lam, Gerhard Mayer, Gerben Menschaert, … Andrew R Jones *Journal of Proteome Research* (2017-09-15)

https://doi.org/gbw99d DOI: 10.1021/acs.jproteome.7b00370 · PMID: 28849660 · PMCID: PMC5715286

191. **Limited Environmental Serine and Glycine Confer Brain Metastasis Sensitivity to PHGDH Inhibition** Bryan Ngo, Eugenie Kim, Victoria Osorio-Vasquez, Sophia Doll, Sophia Bustraan, Roger J Liang, Alba Luengo, Shawn M Davidson, Ahmed Ali, Gino B Ferraro, … Michael E Pacold *Cancer Discovery* (2020-09-01) https://doi.org/ghf85j DOI: 10.1158/2159-8290.cd-19-1228 · PMID: 32571778 · PMCID: PMC7483776

192. **A high-throughput processing service for retention time alignment of complex proteomics and metabolomics LC-MS data** Isthiaq Ahmad, Frank Suits, Berend Hoekman, Morris A Swertz, Heorhiy Byelas, Martijn Dijkstra, Rob Hooft, Dmitry Katsubo, Bas van Breukelen, Rainer Bischoff, Peter Horvatovich *Bioinformatics* (2011-02-23) https://doi.org/cxsszv DOI: 10.1093/bioinformatics/btr094 · PMID: 21349866

193. **MeltDB: a software platform for the analysis and integration of metabolomics experiment data** Heiko Neuweger, Stefan P Albaum, Michael Dondrup, Marcus Persicke, Tony Watt, Karsten Niehaus, Jens Stoye, Alexander Goesmann *Bioinformatics* (2008-09-02) https://doi.org/fds6vt DOI: 10.1093/bioinformatics/btn452 · PMID: 18765459

194. **In silico fragmentation for computer assisted identification of metabolite mass spectra** Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, Steffen Neumann *BMC Bioinformatics* (2010-03-22) https://doi.org/d7gpf5 DOI: 10.1186/1471-2105-11-148 · PMID: 20307295 · PMCID: PMC2853470

195. **The Metabonomic Signature of Celiac Disease** Ivano Bertini, Antonio Calabrò, Valeria De Carli, Claudio Luchinat, Stefano Nepi, Berardino Porfirio, Daniela Renzi, Edoardo Saccenti, Leonardo Tenori *Journal of Proteome Research* (2008-12-11) https://doi.org/c6sdnp DOI: 10.1021/pr800548z · PMID: 19072164

196. **Visualization of omics data for systems biology** Nils Gehlenborg, Seán I O'Donoghue, Nitin S Baliga, Alexander Goesmann, Matthew A Hibbs, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, Anne-Claude Gavin *Nature Methods* (2010-03) https://doi.org/cp9zgj DOI: 10.1038/nmeth.1436 · PMID: 20195258

197. **FunRich: An open access standalone functional enrichment and interaction network analysis tool** Mohashin Pathan, Shivakumar Keerthikumar, Ching-Seng Ang, Lahiru Gangoda, Camelia YJ Quek, Nicholas A Williamson, Dmitri Mouradov, Oliver M Sieber, Richard J Simpson, Agus Salim, … Suresh Mathivanan *PROTEOMICS* (2015-06-17) https://doi.org/f278rp DOI: 10.1002/pmic.201400515 · PMID: 25921073

198. **Proteomic and Metabolomic Characterization of COVID-19 Patient Sera** Bo Shen, Xiao Yi, Yaoting Sun, Xiaojie Bi, Juping Du, Chao Zhang, Sheng Quan, Fangfei Zhang, Rui Sun, Liujia Qian, … Tiannan Guo *Cell* (2020-07) https://doi.org/gg2cck DOI: 10.1016/j.cell.2020.05.032 · PMID: 32492406 · PMCID: PMC7254001

199. **Development of Human Protein Reference Database as an Initial Platform for Approaching Systems Biology in Humans** Suraj Peri, JDaniel Navarro, Ramars Amanchy, Troels Z Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, TKB Gandhi, Mads Gronborg, … Akhilesh

Pandey *Genome Research* (2003-10) https://doi.org/bc8cnv DOI: 10.1101/gr.1680803 · PMID: 14525934 · PMCID: PMC403728

200. **A database for post-genome analysis** Minoru Kanehisa *Trends in Genetics* (1997-09) https://doi.org/cfgb98 DOI: 10.1016/s0168-9525(97)01223-7 · PMID: 9287494

201. **Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases** Matthias Mann, Akhilesh Pandey *Trends in Biochemical Sciences* (2001-01) https://doi.org/ch565r DOI: 10.1016/s0968-0004(00)01726-6 · PMID: 11165518

202. **Genetic Discrimination: Perspectives of Consumers** E. Virginia Lapham, Chahira Kozma, Joan O Weiss *Science* (1996-10-25) https://doi.org/df7k88 DOI: 10.1126/science.274.5287.621 · PMID: 8849455

203. **Committee Opinion No. 690: Carrier Screening in the Age of Genomic Medicine** Obstetrics & Gynecology *Ovid Technologies (Wolters Kluwer Health)* (2017-03) https://doi.org/f92g56 DOI: 10.1097/aog.0000000000001951 · PMID: 28225425

204. **Public health genomics: The end of the beginning** Muin J Khoury *Genetics in Medicine* (2011-03) https://doi.org/bjsxzk DOI: 10.1097/gim.0b013e31821024ca · PMID: 21311338

205. **An STS-Based Map of the Human Genome** Thomas J Hudson, Lincoln D Stein, Sebastian S Gerety, Junli Ma, Andrew B Castle, James Silva, Donna K Slonim, Rafael Baptista, Leonid Kruglyak, Shu-Hua Xu, … Eric S Lander *Science* (1995-12-22) https://doi.org/ftf DOI: 10.1126/science.270.5244.1945 · PMID: 8533086

206. **A New Five-Year Plan for the U.S. Human Genome Project** Francis Collins, David Galas *Science* (1993-10) https://doi.org/fwkrnb DOI: 10.1126/science.8211127 · PMID: 8211127

207. **A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs** Péter L Erdős, István Miklós, Zoltán Toroczkai *arXiv* (2010-01-21) https://arxiv.org/abs/0905.4913

208. **Exploring Network Structure, Dynamics, and Function using NetworkX** Aric A Hagberg, Daniel A Schult, Pieter J Swart *Proceedings of the 7th Python in Science conference* (2008)

209. **Topic-based Pagerank: toward a topic-level scientific evaluation** Erjia Yan *Scientometrics* (2014-05-06) https://doi.org/f6br99 DOI: 10.1007/s11192-014-1308-5

210. **A bird's-eye view of deep learning in bioimage analysis** Erik Meijering *Computational and Structural Biotechnology Journal* (2020) https://doi.org/gk5mtd DOI: 10.1016/j.csbj.2020.08.003 · PMID: 32994890 · PMCID: PMC7494605

211. **ImageNet: A large-scale hierarchical image database** Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009-06) https://doi.org/cvc7xp DOI: 10.1109/cvpr.2009.5206848

212. **Training Compute-Optimal Large Language Models** Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, … Laurent Sifre *arXiv* (2022-03-30) https://arxiv.org/abs/2203.15556