# Neural Nets Are Not All You Need: Evaluating the Effects of Deep Learning on Transcriptomic Analysis

## Acknowledgements

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

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.

# Chapter 1: Background

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 sections 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 in 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 1: 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” (https://doi.org/10.1038/s41592-021-01256-7).

## 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 [**stodden2013?**]. 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 [**stodden2013?**] 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 [93,94,95], 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 [96], when possible. For example, one may deposit gene expression data in the Gene Expression Omnibus [97] or microscopy images in the BioImage Archive [98]. 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 [99] or Sfaira [100] 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 [**docker?**] 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 [101] or Nextflow [102] 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 [103], 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 [104]. Fortunately, model training methods designed to preserve privacy exist: techniques such as differential privacy [105] 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 [106], 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 [**bioconductor?**] 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

TODO COPY OVER Table 1 WHEN CONVERTING TO WORD

# Chapter 6 - Future Directions

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 ml 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 [107], 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 RNNs 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 is 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 contain hundreds of thousands of samples. Meanwhile in machine learning, even before deep learning took off ImageNet already had more than three million images [108].

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 [109] 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.

# References


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, ...
*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. 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. 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)

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**
Blan 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. **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

94. **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

95. **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

96. **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

97. **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

98. **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

99. **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

100. **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

101. **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

102. **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

103. **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

104. **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

105. **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

106. **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

107. **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

108. **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

109. **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