Dissertation Title

This manuscript (permalink) was automatically generated from greenelab/ben_heil_dissertation@1ffd883 on October 10, 2022.

Authors

- **Benjamin J. Heil**
  0000-0002-2811-1031 · benheil · autobencoder
  Genomics and Computational Biology Graduate Group, University of Pennsylvania

✉ — Correspondence possible via GitHub Issues

Abstract

Introduction

Overall Intro

Linear models background

Disease prediction review

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 paper we review supervised 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.

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 [1] 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 [2,3,4], 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 [4]. If not, there are already diagnostics that predict various cancer aspects from biopsy gene expression [5]. It may also be possible to diagnose common diseases based on blood gene expression [6,7,8,9], or even rare ones [10].

Background

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 [11], more modern ones measure tens of thousands of genes [12].

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 [13]. 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 [14]. 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 [14]. The deep learning revolution quickly followed, with deep neural networks becoming the state of the art in any problem with enough data [15,16,17,18].

The implications of the deep learning revolution on this paper are twofold. First, almost all papers before 2014 use traditional machine learning methods, while almost all 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 [19]. 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 [20]. 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 [21] for training, but found that their model performed better when using all genes instead of just a subset [22]. Chen et al. used a different gene set. They selected the LINCS 1000 gene set [23] for an imputation method, as the LINCS landmark genes are highly correlated with the genes they were trying to impute [24].

Gene subsets can be based on prior knowledge of gene regulatory networks as well [25,26]. While very interpretable, these methods do not necessarily lead to increased performance in phenotype predictions [27].

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 [28]. 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 [29]. Instead of looking at raw differential expression, Dhruba et al. used another subsetting method called ReliefF [30] to find the top 200 genes for their source and target dataset, then kept the intersection for use in their model [31]. More recently, Li et al. used a genetic algorithm for feature selection [32].

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 [33,34]. 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 [35] and contractive [36] autoencoders in their model [37], while Danaee et al. used a stacked denoising autoencoder [38,39]. 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 [40]. In DeepType, Chen et al. use a more conventional elastic net regularization [41] to induce sparsity in the first level of their network under the assumption that most genes' expression will not affect a cancer's subtype [4].

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.

Batch Effects

When gene expression data comes from multiple studies, there are systematic differences between the samples even if they are measuring the same thing [42]. These batch effects can bias the outcome of a study, and reduce the ability of a predictive model to generalize to data outside of the dataset used in the analysis. Different studies handle this in different ways, with varying degrees of effectiveness.

Malta et al's study is a good example of addressing batch effects well [43]. They began by mean centering their data to ensure that the model didn't learn to make classifications based on the mean gene expression values. They then used Spearman correlation instead of Pearson correlation to avoid small changes in the data distributions to change their correlation measurement. Finally they evaluated their results on a different data generation method (RNA-seq) from the one they trained on (microarray).

The SAUCIE paper handles batch effects very differently [40]. They introduce a new type of regularization called maximal mean discrepancy, which penalizes the distance between the latent space representation between batches. While this regularization term is deep learning specific and depends on the model having an embedding layer, it will be interesting to see if similar ideas are used in the future.

Other studies address batch effects less comprehensively via quantile normalization and 0-1 standardization [24,44]. Using quantile normalization ensures that the different datasets have the same distribution, then 0-1 standardization makes machine learning algorithms treat all genes as equally important. Another common technique from ML is to make decisions about the model based on cross-validation. Since the feature or hyperparameter choice is validated on multiple random subsets of the data, batch effects are less likely to bias the decision [32].

Studies using contractive autoencoders [37] get some degree of batch effect protection just from their model constraints. Since contractive autoencoders are trained to ignore small perturbations in the data, they tend to be more robust to distributional changes. There are also more explicit ways of addressing batch effects. DeepType, for example, uses the method ComBat [45] to reduce batch effects as a preprocessing step for their model [4].

Unfortunately many studies don't address batch effects at all, despite operating on large multi-study datasets like the Cancer Genome Atlas (TCGA). These studies are likely to fail to generalize to real-world data, as machine learning models like to fixate on spurious correlations between data and phenotypes.

Deep Learning vs Standard ML As was discussed in the background section, 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 [46,47], others find that deep learning outperforms even domain specific models [48,49].

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 [37]. 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 [24]. On the other hand, Grewal et al. tried multiple methods and found they work roughly the same [22]. They settled this by combining a few different models into an ensemble.

Due to technical considerations [43] or other reasons, some authors only evaluate a single model [29]. 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.

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 [31,32,39], 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 [50]. 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[28,37]. 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 [24]. 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 [22]. 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 [43]. Likewise, Deng at al. initially benchmark their model on various simulated data sets, but then go on to validate their model on real data [51].

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

[52,53], 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 [54]. Similarly, Deepathology predicts tissue type, disease, and miRNA expression simultaneously [37]. 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 [20].

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 [28,29]. 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 [39]. 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 [31].

Future Directions

Upon reviewing a broad spectrum of what has been done in the field, a few opportunities for future research have become clear.

As shown in the batch effects section, authors handle batch effects in their studies with varying degrees of sophistication. The studies we have discussed use various strategies to mitigate the technical variation between studies and batches, but it may be possible to do better. Recent developments in the field of transfer learning have lead to methods that use technical variation between samples to increase the power of an analysis [55,56]. These methods exist on the bleeding edge of transfer learning, but gene expression data fits their assumptions very well. It would be interesting to see if models trained with such methods would be more successful than those using traditional batch effect correction.

While many models have been used to make predictions from gene expression, it's unclear which ones work best, and in which circumstances. One review evaluated a variety of unsupervised methods on gene regulatory network discovery, but the only supervised method that was tried was a support vector machine [57]. A large scale study comparing methods to each other would be very useful to the field. Of particular interest would be a study that determines roughly how many samples are needed before it deep learning models outperform traditional machine learning models, and how semi-supervised learning shifts that change point.

Semi-supervised learning is a technique that began being applied to gene expression data only recently. While the technique has been useful when applied to large amounts of unlabeled data, the effects of which unlabeled dataset(s) are used hasn't been measured. Due to the large differences between RNA-seq and microarray data, it may make sense to do pretraining with just GEO or Recount3 [58] depending on whether the labeled data is primarily from microarrays or RNA-seq. A study looking at whether more data is always better, and whether using data from a different platform helps or hurts would be a useful reference for those using semi-supervised learning to train their models.

Looking closer at how to do multitask learning could also help the field. While several studies in this review have analyzed multitask learning, there is not a study that we know of that determines exactly how similar the classes should be for gene expression data. Testing various methods from Sebastian

Ruder’s multitask learning review paper could help find a heuristic for how similar phenotypes should be in multitask learning [59].

For the most part the studies in this review either learn how to diagnose a specific phenotype with a small dataset, or learn more classes by studying TCGA data. We believe that there is an opportunity for datasets to be created from Recount3 and Refine.bio [58,60] data that would be able to predict phenotypes other than just cancer on a large dataset. The consistent preprocessing for these resources makes their gene expression data much easier to use with machine learning methods.

Conclusion and Perspectives

Making predictions from gene expression information holds great promise, and is already being used in some cases. Because the problem space lies between the fields of machine learning and computational biology, however, it inherits pitfalls from both fields. Frequently, biologists who want to attempt to make models will fail to understand how to do model validation and hyperparameter tuning in a way that doesn’t invalidate their results. Likewise, machine learning researchers often will leak information between the training and the testing set by blindly randomizing all their samples, or will fail to account for the batch effects inherent to muli-study datasets.

In addition to the challenges from working across disciplines, the approaches used in making predictions are largely fragmented. Researchers make decisions about their model architecture, dimensionality reduction, and batch effect correction largely based on their intuition. There have been few papers evaluating methods across several problems, and even less consensus about which methods work the best. Moving forward, the field will need to consolidate and determine a set of best practices to reduce the model search space for new papers. Likewise, researchers will need to begin working with clinicians and wet-lab scientists to validate whether their models work in vivo as well as in-silico. Ultimately, phenotype predictions from gene expression appear to have have a bright future. In order to get there, however, there are many challenges that need to be addressed.

Reproducible research background

Citation indices background

Talk briefly about conclusion chapter (probably write this part once it’s done)

References

1. **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

2. **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

3. **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

4. **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

5. **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

6. **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

7. **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

8. **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

9. **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

10. **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 &amp; Rheumatism (2007) https://doi.org/chxcfh
DOI: 10.1002/art.22981 · PMID: 17968951

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

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

18. **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

19. **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

20. **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

21. **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

22. **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

23. **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

24. **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

25. **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

26. **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

27. **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

28. **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

29. **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

30. Igor Kononenko, Edvard Šimec, Marko Robnik-Šikonja
Applied Intelligence (1997) https://doi.org/fdm4r3
DOI: 10.1023/a:1008280620621

31. **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

32. **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

33. **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

34. https://www.researchgate.net/publication/281857285_Using_deep_learning_to_enhance_cancer_diagnosis_and_classification

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

36. **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

37. **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

38. **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) http://jmlr.org/papers/v11/vincent10a.html

39. **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

40. **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

41. https://www.jstor.org/stable/3647580

42. **Detecting and correcting systematic variation in large-scale RNA sequencing data**
Sheng Li, Paweł P Łabaj, Paul Zumbo, Peter Sykacek, Wei Shi, Leming Shi, John Phan, Po-Yen Wu, May Wang, Charles Wang, … Christopher E Mason
Nature Biotechnology (2014-08-24) https://doi.org/f6j2gj
DOI: 10.1038/nbt.3000 · PMID: 25150837 · PMCID: PMC4160374

43. **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

44. **&lt;p&gt;The blood transcriptional signature for active and latent tuberculosis&lt;/p&gt;**
Min Deng, Xiao-Dong Lv, Zhi-Xian Fang, Xin-Sheng Xie, Wen-Yu Chen
Infection and Drug Resistance (2019-01) https://doi.org/gqzq6v
DOI: 10.2147/idr.s184640 · PMID: 30787624 · PMCID: PMC6363485

45. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
WEvan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2006-04-21) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

46. **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

47. **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

48. **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

49. **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

50. **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 &amp; Biobehavioral Reviews (2020-12) https://doi.org/ghkskv
DOI: 10.1016/j.neubiorev.2020.09.036 · PMID: 33035522

51. **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

52. **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

53. **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

54. **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

55. **Invariant Risk Minimization**
Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, David Lopez-Paz
arXiv (2020-03-31) https://arxiv.org/abs/1907.02893

56. **Invariant Risk Minimization Games**
Kartik Ahuja, Karthikeyan Shanmugam, Kush R Varshney, Amit Dhurandhar
arXiv (2020-03-20) https://arxiv.org/abs/2002.04692

57. **Supervised, semi-supervised and unsupervised inference of gene regulatory networks**
SR Maetschke, PB Madhamshettiwar, MJ Davis, MA Ragan
Briefings in Bioinformatics (2013-05-21) https://doi.org/gccv5w
DOI: 10.1093/bib/bbt034 · PMID: 23698722 · PMCID: PMC3956069

58. **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

59. **An Overview of Multi-Task Learning in Deep Neural Networks**
Sebastian Ruder
arXiv (2017-06-19) https://arxiv.org/abs/1706.05098

60. **refine.bio**
Refine.bio
https://www.refine.bio