Machine learning methods for rare diseases

This manuscript (permalink) was automatically generated from jaybee84/ml-in-rd@e854206 on December 7, 2020.

Authors

- **Jineta Banerjee**
  0000-0002-1775-3645 · jaybee84
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

- **Robert J Allaway**
  0000-0003-3573-3565 · allaway · allawayr
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

- **Jaclyn N Taroni**
  0000-0003-4734-4508 · jaclyn-taroni
  Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Casey Greene**
  0000-0001-8713-9213 · cgreene
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Justin Guinney**
  0000-0003-1477-1888 · jguinney
  Sage Bionetworks · Funded by Neurofibromatosis Therapeutic Acceleration Program; Children's Tumor Foundation

Synopsis

(Instructions: Describe the background, basic structure of the article, list material to be covered indicating depth of coverage, how they are logically arranged, include recent pubs in the area, 300-500 words)

Substantial technological advances have dramatically changed biomedicine by making deep characterization of patient samples routine. These technologies provide a rich portrait of genes, cellular pathways, and cell types involved in complex phenotypes. Machine learning is often a perfect fit to extract disease-relevant patterns from these high dimensional datasets. Often, these methods require many samples to identify reproducible and biologically meaningful patterns. With rare diseases, biological specimens and consequently data, are limited due to the rarity of the condition. In this perspective, we outline the challenges and emerging solutions for using machine learning in these settings. We aim to spur the development of powerful machine learning techniques for rare diseases. We also note that precision medicine presents a similar challenge, in which a common disease is partitioned into small subsets of patients with shared etiologies and treatment strategies. Advances from rare disease research are likely to be highly informative for other applications as well.

Introduction

Data collection is increasingly high-throughput, and it can now be more feasible to perform genome-wide profiling than targeted profiling in some settings. Machine learning (ML) is gaining momentum in biomedical data analysis as a means of analyzing and identifying relevant factors from such broad profiling experiments. Parsing such high dimensional data requires computational methods that can learn from the data and can also be applied to newly acquired data in the same disease space [1]. Such computational methods comprise of machine learning algorithms including (i) supervised learning (ii) unsupervised learning, or (iii) reinforcement learning [2].

Application of ML to any dataset requires careful execution, but the application to biomedical data and subsequent interpretation requires depth of knowledge not only in the biomedical domain but also a clear understanding of the methods and their underlying assumptions. Application of ML to any kind of data consists of the following major steps: (1) data evaluation and question formulation, (2) selection of normalization/dimension reduction to mitigate technical differences, (3) selection of appropriate algorithms which select features to answer the formulated question, (4) evaluation of the answers generated by the algorithm. Each of these steps require the practitioner to choose from a variety of methodologies to apply. The selection of the methodologies at each of these steps need to be based upon robust reasoning to ensure stability of the results.

A promising yet challenging application of machine learning is in the study of rare diseases - those with fewer than 200,000 cases in the United States [3]. Rare disease research has substantial constraints to consider when using ML methods, including lack of statistical power due to small dataset size, heterogeneity in available data, and sensitivity of ML methods to misinterpretation and unstable performance in view of small datasets. For example, successful training of ML models require training datasets made of “gold standard” data where the diagnosis or label of a data point has very little uncertainty (or “label-noise”) associated with it [4]. In rare disease the symptoms as well as any underlying biology often come with a reasonable amount label-noise leading to a *silver standard* dataset[5]. Moreover, in the context of rare disease, special considerations need to be made to safeguard against misinterpretation of results. Rare disease datasets are often limited in size and/or constructed at multiple institutions from different types of specimens. Such considerations include incorporation of methods that can mitigate technical disparities in the data and that are resilient to challenges posed by small datasets. Often, this will require techniques that build upon prior domain-specific knowledge and data. In this perspective, we discuss techniques for understanding the nature of rare disease data, including those that address or better tolerate the limitations of these data.

Manage complex high-dimensional rare disease data

The great irony of studying rare diseases with genomics, transcriptomics, or other similar high-throughput methods is that the ability to get an enormous number of measurements from a vanishingly small number of samples, is both the upside and the downfall of these methods. These ‘omic’ methods generate highly dimensional data - that is, data with many features (e.g. all of the mRNA transcripts in a sample). However, to be able to statistically interrogate these measurements, one needs many samples, or observations, which is often not the case in rare disease. This is known as the “curse of dimensionality,” and can be a major impediment in analyzing feature-rich data in sample-deficient contexts [6]. Additionally, identifying various correlated features and interpreting how they relate to the biological question of interest can make using highly dimensional rare disease data for discovery challenging. Furthermore, rare disease data collection and aggregation methods can further complicate these challenges by introducing technical variability into the data at hand. In this section, we will discuss strategies like simplifying data by reducing the dimensionality of high-

dimensional datasets as well as detecting and correcting technical artifacts, that can help mitigate these challenges.

Dimensionality reduction methods including unsupervised machine learning methods like multidimensional scaling (MDS), principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) can help 'compress' information from large number of features into a smaller number of features. [7,8,9,10] Use of these techniques is not restricted to 'omics' data, they can also be used in rare disease applications to characterize imaging data [11], mass cytometry data [12], and others. These methods not only help in reducing the number of features, but can also be used to visualize structure in the data (e.g. [13]), used to define subgroups within data (e.g. [14], or can be used for feature selection/extraction during application of specific machine learning models[15].

In fact, dimensionality reduction is a core concept underpinning a class of ML called representation learning (or feature learning). Representation learning is the process of learning low-dimensional representations or *composites* of features from raw data, where each learned feature becomes an individual variable thus reducing the dimensions of the dataset. One such unsupervised approach, matrix factorization, can extract composite features from transcriptomics datasets made of combinations of gene expression levels, found in the training data that are descriptive in some way, and use them to interpret test input data [16,17,18]. Low-dimensional representations trained on a collection of transcriptomic data can also be used as input to supervised machine learning methods [19]. Supervised neural networks used in medical imaging studies [20] (reviewed in [21]), also use representation learning. Learned composite features in the medical imaging domain may include a series of edges representing a blood vessel formation that discriminates between disease states.

Representation learning generally tends to be data-intensive (i.e. many samples are required) and thus may seem to aggravate the curse of dimensionality. But representation learning when applied to learn low dimensional patterns from large datasets and then applying those patterns to smaller but related datasets can be a powerful tool for dimensionality reduction for smaller datasets. In the later sections of this perspective, we will discuss this method of leveraging large datasets to reduce dimensionality in smaller datasets, also known as *feature-representation-transfer* . In later sections, we will introduce methods that can leverage data that do not directly assay a rare disease of interest; representation learning underlies many of time.

Another application of dimensionality reduction methods is to assess the presence or absence of unwanted signal in the data. Rare disease datasets can sometimes contain structure unrelated to the biology of the disease, e.g. structure related to batch, sample preparation methodology, or sequencing platform [22]. The consequences of these artifacts are amplified when there are few samples and heterogeneous phenotypes. Furthermore, datasets are often combined from multiple small studies where biological characteristics are confounded by technical variables. We can leverage dimensionality reduction methods like PCA, MDS, t-SNE, and UMAP to identify the effect of these variables on the data. All of these methods can be used to identify batch effects and other structure in the data, though some (like t-SNE and UMAP) may require parameters that can affect the output [10,25]. Therefore, obtaining a clear interpretation from these methods requires understanding their underlying approach and parameters. Way, et. al. [18] further suggests that a single dimensionality reduction method alone may not be sufficient to reveal all of the technical or biological heterogeneity; thus testing multiple methods may result in a more comprehensive portrait of the data.

In addition to methodological considerations, collaboration with domain experts may result in unexpected insight into potential sources of variation. As an example, consider a study of neurofibromatosis type 1 (NF1) datasets.[26] These datasets were, unbeknownst to the computational biologists, generated from samples obtained with vastly different surgical techniques (laser ablation and excision vs standard excision), resulting in substantial biological differences that

are a consequence of process, not reality. One might expect, in this example, that this technical decision would result in profound changes in the underlying biology, such as the activation of heat shock protein related pathways, unfolded protein responses, and so on. Consequently, careful assessment of and accounting for confounding factors is critical to identifying meaningful features within a dataset.

Assessment of confounding factors and heterogeneity is perhaps most efficiently performed using unsupervised learning approaches. K-means clustering or hierarchical clustering can be used to characterize the structure present in genomic or imaging data. [27,28]

Once the nature of the non-biological heterogeneity has been established, different techniques can be used to correct the differences. Common approaches include reprocessing the raw data using a single analysis pipeline if the data are obtained from different sources, application of batch correction methods [29,30], and normalization of raw values[31]. It is also important to be realistic when working with rare disease data. For various reasons including ethical constraints, funding, and limited biospecimens, experimental design and the resulting data will often be less-than-ideal. In these cases, it may be prudent to take a step back, re-evaluate the data, and identify methods that can operate within the constraints of the data, rather than expecting the data to conform to a method of choice.

Manage model complexity while preserving the value of machine learning

Fruitful translation of patterns extracted from a dataset using machine learning into testable hypotheses requires the models to be a) stable i.e. the same predicted features should surface from the data if the model is run multiple times and, b) simple to improve interpretability and avoid misinterpretation due to technical artifacts. Fulfilling these pre-requisites become even more important in case of rare disease datasets where there is high label-uncertainty (i.e. where the label given to a data point may not be correct due to imperfect understanding of the disease). In this section we highlight a few techniques which can help improve the stability and simplicity of ML models.

Techniques like bootstrapping and ensemble learning can increase stability in machine learning predictions.

Bootstrapping is a powerful statistical technique where resampling the data with replacements can help estimate population values from datasets of limited sample size [32]. Such resampling with replacement is used in various learning methods to find the most informative models (e.g. bootstrap aggregating or *bagging* used in random forests [33,34], bootstrap in neural networks [35], or regression models [36,37]). A variation where resampling of a rare disease dataset was done *without replacement*, generated confidence intervals for the model predictions by iteratively exposing the models to incomplete datasets (mimicking the real world cases where most rare disease datasets are incomplete) [26].

Stability in predictions can also be achieved by combining various ML methods together (*ensemble learning*). Ensemble learning methods like random forests use bagging of independent decision trees that use similar parameters but different paths to form a consensus about the important predictive features [33,38,39,40,41]. But such methods have shown limited success in rare disease datasets where the label-uncertainty can be high due to imperfect understanding of the disease (i.e.silver standard datasets). This has led to the adoption of cascade learning, where multiple methods leveraging distinct underlying assumptions are used in tandem. The methods may be augmented with algorithms like AdaBoost (*boosting*) to capture stable patterns existing in the silver standard data [42,43,44]. A variation of cascade learning implemented to identify rare disease patients from

electronic health records from the general population utilized independent steps for feature extraction (using word2vec [45]), preliminary prediction (ensemble of decision trees with penalization for excessive tree-depth), and prediction refinement (using similarity of data points to resolve sample labels) [46]. Combining these three methods resulted in better performance than other methods when implemented on the silver standard dataset in isolation.

Techniques like regularization and *one-class-at-a-time* classifications can help simplify models by making the feature space proportionate with the sample space.

Regularization can not only protect ML models from *overfitting* (where the model performs well for the training data but poorly for new test data) [47], but also help reduce the feature space to help build simpler models using limited datasets. The three main methods of regularization include ridge regression, LASSO, and elastic-net. While ridge regression can minimize the magnitude of the features, it cannot remove unimportant features. LASSO regression, on the other hand, works well for selecting few important features since it can minimize the magnitude of some features more than the others[48]. A combination of LASSO and ridge, elastic-net regression[49] efficiently selects the most useful features, especially in presence of large number of correlated features.

While regularization has not been used extensively in rare disease yet, examples in rare variant discovery and immune cell signature discovery can provide insights into their possible application in rare disease. In rare variant discovery, ridge regression has been utilized to combine rare variants into a single score to increase the signal of rare variants [50], while LASSO was implemented along with group penalties to identify rare variants/low frequency predictors [51,52]. Hybrid applications of LASSO have also been tested in rare variant discovery, including boosting the signal of rare variants by capturing combinations of variants [53,54], integration with a probabilistic logistic Bayesian approach [55], combining feature selection methods with a generalized pooling strategy [56], and incorporating prior knowledge into the regularization step to select driver genes in a pathway of interest [57]. In immune cell signature discovery, elastic-net regression has been used to reduce the feature space and was found to outperform other regression approaches [49,58,59,60]. Regularization methods like LASSO or elastic-net have been methods of choice for making models simpler by reducing the feature space; these methods should be explored while working with rare disease datasets.

In rare diseases like neurofibromatosis, the presence of more than one phenotype (or class) further decreases the number of data-points per class and introduces additional label-uncertainty due to related phenotypes. In datasets with multiple classes, the classical ensemble or cascade classifiers approach follows a *one-classifier-at-a-time* approach where algorithms at each level predict all classes involved. But instances where the need for high prediction accuracy for one class outweighs other classes, modification of the cascade learning method into a *one-class-at-a-time* approach (where at each stage a binary classifier predicts a specific class against all others) has been found to be beneficial [61]. In this instance, the final model implemented all models together each identifying one class sequentially and then reporting the union of the predictions of all the different models as the final prediction. The cascade classifiers using the one-class-at-a-time approach were found to perform better than multi-class ensemble classifiers in most cases.

By employing bootstrapping, ensemble learning, and regularization methods, researchers may be able to better generate stable, simple models that identify reliable biological phenomena underlying rare diseases.

Utilize prior knowledge and indirectly related data

Rare diseases lack large, normalized datasets, limiting our ability to study key attributes of these diseases. A potentially powerful strategy for evaluating genotype-phenotype relationships or repurposing drugs when large datasets are scarce is to use knowledge graphs. Knowledge graphs integrate related-but-different data types, creating a rich data source. Examples of public biomedical knowledge graphs and frameworks that could be useful in rare disease include the Monarch Graph Database[62], hetionet[63], PheKnowLator[64], and the Global Network of Biomedical Relationships[65]. These graphs connect information like genetic, functional, chemical, clinical, and ontological data to enable the exploration of relationships of data with disease phenotypes through manual review[66] or computational methods[67,68].

In the academic rare disease space, there a few pioneering examples of ML-based mining of knowledge graphs to repurpose drugs[67] and classify rare diseases[68] highlight the challenges in using these in rare disease. For example, these projects rely on a gold standard datasets to validate the performance of the models which are often absent for individual rare diseases. They also evaluate rare diseases in an unbiased manner, rather than interrogating a specific disease of interest. Consequently, it is not yet clear how effective these approaches, and knowledge graphs in general, are in studying a specific disease of interest. More work needs to be done to identify methods that can provide actionable insights for a specific rare disease application.

Beyond the few academic studies, private entities (e.g. healx, Boehringer Ingelheim) are also performing an undisclosed amount of work to create proprietary rare disease knowledge graphs for ML-based drug discovery applications. The existence of private companies pursuing this idea, as well as the availability of public biomedical knowledge graphs, suggests that this may be a fruitful untapped area of rare disease research in the public arena. More work needs to be done to assess 1) which graphs and graph features capture the salient information about rare diseases, 2) the utility of ML methods to obtain actionable insights about rare diseases and 3) which problems - like drug discovery, identification of novel rare diseases, or assessment of genotype-phenotype relationships - can be interrogated using ML of knowledge graphs. However, the existence of interest in this idea in both pharma and public domain suggests that this is an untapped area of rare disease research.

Other approaches that build upon prior knowledge and large volumes of related data include transfer learning, multitask learning, and few-shot learning approaches. These approaches leverage shared features e.g. normal developmental processes that are aberrant in disease, imaging anomalies present in rare and common diseases, for advancing our understanding of rare diseases.

Transfer learning is an approach where a model trained for one task or domain (source domain) is applied to another, typically related task or domain (target domain). Transfer learning can be supervised (one or both of the source and target domains have labels), or unsupervised (both domains are unlabeled). Among the various types of transfer learning, we will mainly focus on *feature-representation-transfer*. Feature-representation-transfer approaches learn representations from the source domain and apply them to a target domain [69]. For example, low-dimensional representations can be learned from tumor transcriptomic data and transferred to describe patterns associated with genetic alterations in cell line data [18].

Transfer learning can be supervised or unsupervised, but the related approaches multitask and few-shot learning are forms of supervised learning that generally rely on deep neural networks. Multitask learning is an approach where classifiers use shared representations to learn multiple *related but individual predictions* (tasks) simultaneously [70]. Few-shot learning on the other hand generalizes a model trained on related tasks to a new task with limited labeled data (e.g., the detection of a patient with a rare disease from a low number of examples of that rare disease). While the variety of

approaches and architectures that underlie multitask and few-shot learning are beyond this scope of this work (see [71,72] and [73] for an overview), we will delve into a few selected studies to illustrate potential uses and limitations of these approaches in rare disease.

Multitask neural networks (which predict multiple tasks simultaneously) improve model performance over single task models by learning a shared representation, effectively being exposed to more training data than single task models [70,74]. Examination of the effects of dataset size and task relatedness on multitask learning performance improvements ("multitask effect") in drug discovery showed that smaller datasets tended to benefit most from multitask learning and the addition of more training data did not guarantee improved performance for multitask models [74]. Another study demonstrated that performance gains were context-dependent, i.e. multitask neural networks outperformed single-task networks for predicting complex rare phenotypes from EHR data, but not common phenotypes [75].

In contrast, one-shot or few-shot learning uses prior knowledge to generalize a distance metric learned from input data to compare with a low number of new examples for prediction [73,76], e.g. a method developed for predicting small molecule activity learned a meaningful distance metric over the properties of various compounds [76]. But the authors' results also suggest that also suggest limited performance boost or generalizability of the models relative to baseline models in absence of structural similarity among compounds. Moreover, few-shot learning approaches outperform multitask learning in contexts where predicting common conditions simultaneously results in a loss of performance [77].

Thus transfer, multi-task, and few-shot learning are appealing for the study of rare diseases, conditions, or phenotypes, but their limits and potential utility are still open research questions. Nevertheless, selecting an appropriate model for a given task and evaluations that are well-aligned with a research question are crucial for applying these approaches in rare diseases.

Multiple approaches are required

In the rare disease domain, Dincer et al. leveraged publicly available acute myeloid leukemia (AML) gene expression data to improve the prediction of *in vitro* drug responses [78]. The authors trained a variational autoencoder (an unsupervised neural network that learns a series of representations from data), or VAE, on AML data that had been collected over time without the desired phenotypic information (drug response). The authors used the learned attributes to encode a low-dimensional representation of held-out AML data with phenotype labels of interest, and used this representation as input to a classifier that predicted *in vitro* drug response.

Though there were over 6500 AML samples from many different studies used as part of the training set in Dincer et al. [78], we expect that in other rare diseases considerably fewer samples will be available or may be from different tissues in systemic diseases. The study by Dincer and colleagues highlights another challenge: samples collected as part of multiple studies may not be associated with the deep phenotypic information that would maximize their scientific value.

Feature-representation-transfer is embodied in Dincer et al., where features are learned from unlabeled AML data and then used to encode a low-dimensional representation of AML data with in vitro drug response labels [78]. The authors then used this low-dimensional representation as input to predict drug response labels–a supervised example.

In an unsupervised case, Taroni et al. trained Pathway-Level Information ExtractoR (PLIER) [79] on a large generic collection of human transcriptomic data (recount2 [80]) and used the latent variables learned by the model to describe transcriptomic data from the unseen rare diseases antineutrophil cytoplasmic antibody (ANCA)-associated vasculitis (AAV) and medulloblastoma in an approach termed

MultiPLIER [22]. (Here “unseen” refers to the fact that these diseases were not in the training set.) PLIER is a matrix factorization approach that takes prior knowledge in the form of gene sets or pathways and gene expression data as input; some latent variables learned by the model will align with input gene sets [79]. Training on larger collections of randomly selected samples produced models that captured a larger proportion of input gene sets and better distinguished closely related signals, which suggests that larger training sets produced models that are more suitable for biological discovery [22].

Though models trained on generic compendia had appealing properties, we need to also examine the relevance of learned features to the disease under study. In Taroni et al., we found that the expression of latent variables that could be matched between the MultiPLIER model and a dataset-specific model were well-correlated, particularly when latent variables were associated with input gene sets [22]. Despite the absence of AAV from the training set, MultiPLIER was able to learn a latent variable where the genes with the highest contributions encode antigens that the ANCA form against in AAV and with higher expression in more severe disease [81]. The utility of this approach stems from the fact that biological processes are often shared between conditions–the same ANCA antigen genes are components of neutrophilic granule development that is likely captured or assayed in the collection of transcriptomic data used for training. MultiPLIER has additional attributes that make it practical for studying rare diseases: latent variables that are not associated with input gene sets may capture technical noise separately from biological signal and we can use one model to describe multiple datasets instead of reconciling output from multiple models (see 03.heterogeneity.md).

Taken together, DeepProfile [78] and MultiPLIER [22] suggest transfer learning can be beneficial for studying rare diseases. In the natural images field, researchers have demonstrated that the transferability of features depends on relatedness of tasks [82]. The limits of transfer learning for and the concept of relatedness in high-dimensional biomedical data assaying rare diseases are open research questions. In the authors’ opinion, selecting an appropriate model for a given task and evaluations that are well-aligned with a research goal are crucial for applying these approaches in rare diseases.

Conclusions

> We will conclude by discussing the potential of the above-mentioned approaches in rare diseases and other biomedical areas where data is scarce.

Outlook

References

1. **Looking beyond the hype: Applied AI and machine learning in translational medicine**
   Tzen S. Toh, Frank Dondelinger, Dennis Wang
   EBioMedicine (2019-09) https://doi.org/gg9dcx
   DOI: 10.1016/j.ebiom.2019.08.027 · PMID: 31466916 · PMCID: PMC6796516

2. (2014-08-15) http://cds.cern.ch/record/998831/files/9780387310732_TOC.pdf

3.
   Potomac Publishing
   (2018-10-08) https://www.fda.gov/media/99546/download

4. **Learning statistical models of phenotypes using noisy labeled training data**
   Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, Nigam H Shah
   Journal of the American Medical Informatics Association (2016-11) https://doi.org/f9bxf9
   DOI: 10.1093/jamia/ocw028 · PMID: 27174893 · PMCID: PMC5070523

5. **Classification in the Presence of Label Noise: A Survey**
   Benoit Frenay, Michel Verleysen
   IEEE Transactions on Neural Networks and Learning Systems (2014-05) https://doi.org/f5zdgq
   DOI: 10.1109/tnnls.2013.2292894 · PMID: 24808033

6. **The properties of high-dimensional data spaces: implications for exploring gene and protein expression data**
   Robert Clarke, Habtom W. Ressom, Antai Wang, Jianhua Xuan, Minetta C. Liu, Edmund A. Gehan, Yue Wang
   Nature Reviews Cancer (2008-01) https://doi.org/ffksnf
   DOI: 10.1038/nrc2294 · PMID: 18097463 · PMCID: PMC2238676

7. **Multidimensional Scaling**
   Michael A. A. Cox, Trevor F. Cox
   Springer Science and Business Media LLC (2008) https://doi.org/dg9m4f
   DOI: 10.1007/978-3-540-33037-0_14

8. **Principal component analysis: a review and recent developments**
   Ian T. Jolliffe, Jorge Cadima
   Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences (2016-04-13) https://doi.org/gcsfk7
   DOI: 10.1098/rsta.2015.0202 · PMID: 26953178 · PMCID: PMC4792409

9. (2020-06-01) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

10. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
    Leland McInnes, John Healy, James Melville
    arXiv (2020-09-21) https://arxiv.org/abs/1802.03426

11. **Automatic detection of rare pathologies in fundus photographs using few-shot learning**
    Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, Béatrice Cochener

Medical Image Analysis (2020-04) https://doi.org/ggsrc7
DOI: 10.1016/j.media.2020.101660 · PMID: 32028213

12. **Sensitive detection of rare disease-associated cell subsets via representation learning**
Eirini Arvaniti, Manfred Claassen
Nature Communications (2017-04-06) https://doi.org/gf9t7w
DOI: 10.1038/ncomms14825 · PMID: 28382969 · PMCID: PMC5384229

13. **The art of using t-SNE for single-cell transcriptomics**
Dmitry Kobak, Philipp Berens
Nature Communications (2019-11-28) https://doi.org/ggdrfz
DOI: 10.1038/s41467-019-13056-x · PMID: 31780648 · PMCID: PMC6882829

14. **Dimensionality reduction by UMAP to visualize physical and genetic interactions**
Michael W. Dorrity, Lauren M. Saunders, Christine Queitsch, Stanley Fields, Cole Trapnell
Nature Communications (2020-03-24) https://doi.org/ggqcqp
DOI: 10.1038/s41467-020-15351-4 · PMID: 32210240 · PMCID: PMC7093466

15. **Feature Selection**
Rama Chellappa, Pavan Turaga
Springer Science and Business Media LLC (2020) https://doi.org/ghgqb9
DOI: 10.1007/978-3-030-03243-2_299-1

16. **Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder**
Sanjiv K. Dwivedi, Andreas Tjärnberg, Jesper Tegnér, Mika Gustafsson
Nature Communications (2020-02-12) https://doi.org/gg7krm
DOI: 10.1038/s41467-020-14666-6 · PMID: 32051402 · PMCID: PMC7016183

17. **CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data**
Elana J. Fertig, Jie Ding, Alexander V. Favorov, Giovanni Parmigiani, Michael F. Ochs
Bioinformatics (2010-11-01) https://doi.org/cwqsv4
DOI: 10.1093/bioinformatics/btq503 · PMID: 20810601 · PMCID: PMC3025742

18. **Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations**
Gregory P. Way, Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, Casey S. Greene
Genome Biology (2020-05-11) https://doi.org/gg2mjh
DOI: 10.1186/s13059-020-02021-3 · PMID: 32393369 · PMCID: PMC7212571

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

20. **Convolutional Neural Networks for Diabetic Retinopathy**
Harry Pratt, Frans Coenen, Deborah M. Broadbent, Simon P. Harding, Yalin Zheng
Procedia Computer Science (2016) https://doi.org/gcgk75
DOI: 10.1016/j.procs.2016.07.014

21. **Opportunities and obstacles for deep learning in biology and medicine**
Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, … Casey S. Greene
Journal of The Royal Society Interface (2018-04-04) https://doi.org/gddkhn
DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

22. **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

23. **A scoping review and proposed workflow for multi-omic rare disease research**
Katie Kerr, Helen McAneney, Laura J. Smyth, Caitlin Bailie, Shane McKee, Amy Jayne McKnight
Orphanet Journal of Rare Diseases (2020-04-28) https://doi.org/ggzkhf
DOI: 10.1186/s13023-020-01376-x · PMID: 32345347 · PMCID: PMC7189570

24. **Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts**
Laure Frésard, Craig Smail, Nicole M. Ferraro, Nicole A. Teran, Xin Li, Kevin S. Smith, Devon Bonner, Kristin D. Kernohan, Shruti Marwaha, Zachary Zappala, … Care4Rare Canada Consortium
Nature Medicine (2019-06-03) https://doi.org/gf3jnd
DOI: 10.1038/s41591-019-0457-8 · PMID: 31160820 · PMCID: PMC6634302

25. **How to Use t-SNE Effectively**
Martin Wattenberg, Fernanda Viégas, Ian Johnson
Distill (2016-10-13) https://doi.org/gffk7g
DOI: 10.23915/distill.00002

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

27. **Clustering cancer gene expression data: a comparative study**
Marcilio CP de Souto, Ivan G Costa, Daniel SA de Araujo, Teresa B Ludermir, Alexander Schliep
BMC Bioinformatics (2008-11-27) https://doi.org/dqqbn6
DOI: 10.1186/1471-2105-9-497 · PMID: 19038021 · PMCID: PMC2632677

28. **Removing Batch Effects From Histopathological Images for Enhanced Cancer Diagnosis**
Sonal Kothari, John H. Phan, Todd H. Stokes, Adeboye O. Osunkoya, Andrew N. Young, May D. Wang
IEEE Journal of Biomedical and Health Informatics (2014-05) https://doi.org/gdm9jd
DOI: 10.1109/jbhi.2013.2276766 · PMID: 24808220 · PMCID: PMC5003052

29. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
W. Evan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2007-01) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

30. **svaseq: removing batch effects and other unwanted noise from sequencing data**
Jeffrey T. Leek
Nucleic Acids Research (2014-12-01) https://doi.org/f8k8kf
DOI: 10.1093/nar/gku864 · PMID: 25294822 · PMCID: PMC4245966

31. **A scaling normalization method for differential expression analysis of RNA-seq data**
Mark D Robinson, Alicia Oshlack
Genome Biology (2010) https://doi.org/cq6f8b
DOI: 10.1186/gb-2010-11-3-r25 · PMID: 20196867 · PMCID: PMC2864565

32. **Improvements on Cross-Validation: The 632+ Bootstrap Method**
Bradley Efron, Robert Tibshirani
Journal of the American Statistical Association (1997-06) https://doi.org/gfts5c
DOI: 10.1080/01621459.1997.10474007

33. **:(unav)**
Leo Breiman
Machine Learning (2001) https://doi.org/d8zjwq
DOI: 10.1023/a:1010933404324

34. **Bootstrap Methods for Developing Predictive Models**
Peter C Austin, Jack V Tu
The American Statistician (2004-05) https://doi.org/bzjjxt
DOI: 10.1198/0003130043277

35. **Bootstrap for neural model selection**
Riadh Kallel, Marie Cottrell, Vincent Vigneron
Neurocomputing (2002-10) https://doi.org/c8xpqz
DOI: 10.1016/s0925-2312(01)00650-6

36. **Fast bootstrap methodology for regression model selection**
A. Lendasse, G. Simon, V. Wertz, M. Verleysen
Neurocomputing (2005-03) https://doi.org/dx5c3p
DOI: 10.1016/j.neucom.2004.11.017

37. **A bootstrap resampling procedure for model building: Application to the cox regression model**
Willi Sauerbrei, Martin Schumacher
Statistics in Medicine (1992) https://doi.org/cnpg3d
DOI: 10.1002/sim.4780111607 · PMID: 1293671

38. **Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data**
Felix Köpcke, Dorota Lubgan, Rainer Fietkau, Axel Scholler, Carla Nau, Michael Stürzl, Roland Croner, Hans-Ulrich Prokosch, Dennis Toddenroth
BMC Medical Informatics and Decision Making (2013-12-09) https://doi.org/f5jqvh
DOI: 10.1186/1472-6947-13-134 · PMID: 24321610 · PMCID: PMC4029400

39. **Analyzing bagging**
Peter Bühlmann, Bin Yu
The Annals of Statistics (2002-08) https://doi.org/btmtjp
DOI: 10.1214/aos/1031689014

40. **Utilising artificial intelligence to determine patients at risk of a rare disease: idiopathic pulmonary arterial hypertension**
David G. Kiely, Orla Doyle, Edmund Drage, Harvey Jenner, Valentina Salvatelli, Flora A. Daniels, John Rigg, Claude Schmitt, Yevgeniy Samyshkin, Allan Lawrie, Rito Bergemann
Pulmonary Circulation (2019-11-20) https://doi.org/gg4jc7
DOI: 10.1177/2045894019890549 · PMID: 31798836 · PMCID: PMC6868581

41. **Double-bagging: combining classifiers by bootstrap aggregation**
Torsten Hothorn, Berthold Lausen
Pattern Recognition (2003-06) https://doi.org/btzfh6
DOI: 10.1016/s0031-3203(02)00169-3

42. **Component-based face detection**
B. Heiselet, T. Serre, M. Pontil, T. Poggio
Institute of Electrical and Electronics Engineers (IEEE) (2005-08-24) https://doi.org/c89p2b
DOI: 10.1109/cvpr.2001.990537

43. **The Architecture of the Face and Eyes Detection System Based on Cascade Classifiers**
Andrzej Kasinski, Adam Schmidt
Advances in Soft Computing (2007) https://doi.org/cbzq9n
DOI: 10.1007/978-3-540-75175-5_16

44. **Real time facial expression recognition with AdaBoost**
Yubo Wang, Haizhou Ai, Bo Wu, Chang Huang
Institute of Electrical and Electronics Engineers (IEEE) (2004) https://doi.org/crv3sq
DOI: 10.1109/icpr.2004.1334680

45. **Efficient Estimation of Word Representations in Vector Space**
Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
arXiv (2013-01-16) https://arxiv.org/abs/1301.3781v3

46. **Learning to Identify Rare Disease Patients from Electronic Health Records.**
Rich Colbaugh, Kristin Glass, Christopher Rudolf, Mike Tremblay Volv Global Lausanne Switzerland
AMIA ... Annual Symposium proceedings. AMIA Symposium (2018-12-05)
https://www.ncbi.nlm.nih.gov/pubmed/30815073
PMID: 30815073 · PMCID: PMC6371307

47. **Definitions, methods, and applications in interpretable machine learning**
W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, Bin Yu
Proceedings of the National Academy of Sciences (2019-10-29) https://doi.org/ggbhmq
DOI: 10.1073/pnas.1900654116 · PMID: 31619572 · PMCID: PMC6825274

48. **Regularization**
Jake Lever, Martin Krzywinski, Naomi Altman
Nature Methods (2016-09-29) https://doi.org/gf3zrr
DOI: 10.1038/nmeth.4014

49. **Regularization and variable selection via the elastic net**
Hui Zou, Trevor Hastie
Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2005-04)
https://doi.org/b8cwwr
DOI: 10.1111/j.1467-9868.2005.00503.x

50. **Adaptive Ridge Regression for Rare Variant Detection**
Haimao Zhan, Shizhong Xu
PLoS ONE (2012-08-28) https://doi.org/f36tm5
DOI: 10.1371/journal.pone.0044173 · PMID: 22952918 · PMCID: PMC3429469

51. **Statistical analysis strategies for association studies involving rare variants**
Vikas Bansal, Ondrej Libiger, Ali Torkamani, Nicholas J. Schork
Nature Reviews Genetics (2010-10-13) https://doi.org/dn4jtz
DOI: 10.1038/nrg2867 · PMID: 20940738 · PMCID: PMC3743540

52. **Association screening of common and rare genetic variants by penalized regression**
H. Zhou, M. E. Sehl, J. S. Sinsheimer, K. Lange
Bioinformatics (2010-08-06) https://doi.org/c7ndkx
DOI: 10.1093/bioinformatics/btq448 · PMID: 20693321 · PMCID: PMC3025646

53. **Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data**
Bingshan Li, Suzanne M. Leal
The American Journal of Human Genetics (2008-09) https://doi.org/d4jpcb
DOI: 10.1016/j.ajhg.2008.06.024 · PMID: 18691683 · PMCID: PMC2842185

54. **Comparison of statistical approaches to rare variant analysis for quantitative traits**
Han Chen, Audrey E Hendricks, Yansong Cheng, Adrienne L Cupples, Josée Dupuis, Ching-Ti Liu
BMC Proceedings (2011-11-29) https://doi.org/b9mf4x
DOI: 10.1186/1753-6561-5-s9-s113 · PMID: 22373209 · PMCID: PMC3287837

55. **An Improved Version of Logistic Bayesian LASSO for Detecting Rare Haplotype-Environment Interactions with Application to Lung Cancer**
Yuan Zhang, Swati Biswas
Cancer Informatics (2015-02-09) https://doi.org/ggxxfp
DOI: 10.4137/cin.s17290 · PMID: 25733797 · PMCID: PMC4332044

56. **Multiple Regression Methods Show Great Potential for Rare Variant Association Tests**
ChangJiang Xu, Martin Ladouceur, Zari Dastani, J. Brent Richards, Antonio Ciampi, Celia M. T. Greenwood
PLoS ONE (2012-08-08) https://doi.org/f35726
DOI: 10.1371/journal.pone.0041694 · PMID: 22916111 · PMCID: PMC3420665

57. **A Sparse-Group Lasso**
Noah Simon, Jerome Friedman, Trevor Hastie, Robert Tibshirani
Journal of Computational and Graphical Statistics (2013-04) https://doi.org/gcvjw8
DOI: 10.1080/10618600.2012.681250

58. **Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification**
Zakariya Yahya Algamal, Muhammad Hisyam Lee
Computers in Biology and Medicine (2015-12) https://doi.org/f73xvj
DOI: 10.1016/j.compbiomed.2015.10.008 · PMID: 26520484

59. **Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification**
Yong Liang, Cheng Liu, Xin-Ze Luan, Kwong-Sak Leung, Tak-Ming Chan, Zong-Ben Xu, Hai Zhang
BMC Bioinformatics (2013-06-19) https://doi.org/gb8v2x
DOI: 10.1186/1471-2105-14-198 · PMID: 23777239 · PMCID: PMC3718705

60. **An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets**
Arezo Torang, Paraag Gupta, David J. Klinke
BMC Bioinformatics (2019-08-22) https://doi.org/gg5hmj
DOI: 10.1186/s12859-019-2994-z · PMID: 31438843 · PMCID: PMC6704630

61. **Machine learning for psychiatric patient triaging: an investigation of cascading classifiers.**
Vivek Kumar Singh, Utkarsh Shrivastava, Lina Bouayad, Balaji Padmanabhan, Anna Ialynytchev, Susan K Schultz
Journal of the American Medical Informatics Association : JAMIA (2018-11-01)
https://www.ncbi.nlm.nih.gov/pubmed/30380082
DOI: 10.1093/jamia/ocy109 · PMID: 30380082 · PMCID: PMC6213089

62. **The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species**
Christopher J. Mungall, Julie A. McMurry, Sebastian Köhler, James P. Balhoff, Charles Borromeo, Matthew Brush, Seth Carbon, Tom Conlin, Nathan Dunn, Mark Engelstad, … Melissa A. Haendel
Nucleic Acids Research (2017-01-04) https://doi.org/f9v7bz
DOI: 10.1093/nar/gkw1128 · PMID: 27899636 · PMCID: PMC5210586

63. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

64. **A Framework for Automated Construction of Heterogeneous Large-Scale Biomedical Knowledge Graphs**
Tiffany J. Callahan, Ignacio J. Tripodi, Lawrence E. Hunter, William A. Baumgartner
Cold Spring Harbor Laboratory (2020-05-02) https://doi.org/gg338z
DOI: 10.1101/2020.04.30.071407

65. **A global network of biomedical relationships derived from text**
Bethany Percha, Russ B Altman
Bioinformatics (2018-08-01) https://doi.org/gc3ndk
DOI: 10.1093/bioinformatics/bty114 · PMID: 29490008 · PMCID: PMC6061699

66. **Structured reviews for data and knowledge-driven research**
Núria Queralt-Rosinach, Gregory S Stupp, Tong Shu Li, Michael Mayers, Maureen E Hoatlin, Matthew Might, Benjamin M Good, Andrew I Su
Database (2020) https://doi.org/ggsdkj
DOI: 10.1093/database/baaa015 · PMID: 32283553 · PMCID: PMC7153956

67. **A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases**
Daniel N. Sosa, Alexander Derry, Margaret Guo, Eric Wei, Connor Brinton, Russ B. Altman
Cold Spring Harbor Laboratory (2019-08-08) https://doi.org/gg5j64
DOI: 10.1101/727925

68. **Improving rare disease classification using imperfect knowledge graph**
Xuedong Li, Yue Wang, Dongwu Wang, Walter Yuan, Dezhong Peng, Qiaozhu Mei
BMC Medical Informatics and Decision Making (2019-12-05) https://doi.org/gg5j65
DOI: 10.1186/s12911-019-0938-1 · PMID: 31801534 · PMCID: PMC6894101

69. **A Survey on Transfer Learning**
Sinno Jialin Pan, Qiang Yang
IEEE Transactions on Knowledge and Data Engineering (2010-10) https://doi.org/bc4vws
DOI: 10.1109/tkde.2009.191

70.**:(unav)**
Rich Caruana
Machine Learning (1997) https://doi.org/d3gsgj
DOI: 10.1023/a:1007379606734

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

72. **A Survey on Multi-Task Learning**
Yu Zhang, Qiang Yang
arXiv (2017-07-25) https://arxiv.org/abs/1707.08114v2

73. **Generalizing from a Few Examples: A Survey on Few-Shot Learning**
Yaqing Wang, Quanming Yao, James Kwok, Lionel M. Ni
arXiv (2019-04-10) https://arxiv.org/abs/1904.05046v3

74. **Modeling Industrial ADMET Data with Multitask Networks**
Steven Kearnes, Brian Goldman, Vijay Pande
arXiv (2017-01-16) https://arxiv.org/abs/1606.08793

75. **The Effectiveness of Multitask Learning for Phenotyping with Electronic Health Records Data**
Daisy Yi Ding, Chloé Simpson, Stephen Pfohl, Dave C. Kale, Kenneth Jung, Nigam H. Shah
arXiv (2019-01-08) https://arxiv.org/abs/1808.03331

76. **Low Data Drug Discovery with One-Shot Learning**
Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, Vijay Pande
ACS Central Science (2017-04-03) https://doi.org/f95dnd
DOI: 10.1021/acscentsci.6b00367 · PMID: 28470045 · PMCID: PMC5408335

77. **Automatic detection of rare pathologies in fundus photographs using few-shot learning**
Gwenolé Quellec, Mathieu Lamard, Pierre-Henri Conze, Pascale Massin, Béatrice Cochener
arXiv (2019-07-22) https://arxiv.org/abs/1907.09449v3
DOI: 10.1016/j.media.2020.101660

78. **DeepProfile: Deep learning of cancer molecular profiles for precision medicine**
Ayse Berceste Dincer, Safiye Celik, Naozumi Hiranuma, Su-In Lee
Cold Spring Harbor Laboratory (2018-05-26) https://doi.org/gdj2j4
DOI: 10.1101/278739

79. **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

80. **Reproducible RNA-seq analysis using recount2**
Leonardo Collado-Torres, Abhinav Nellore, Kai Kammers, Shannon E Ellis, Margaret A Taub, Kasper

D Hansen, Andrew E Jaffe, Ben Langmead, Jeffrey T Leek
Nature Biotechnology (2017-04-11) https://doi.org/gf75hp
DOI: 10.1038/nbt.3838 · PMID: 28398307 · PMCID: PMC6742427

81. **Transcription of proteinase 3 and related myelopoiesis genes in peripheral blood mononuclear cells of patients with active Wegener's granulomatosis**
Chris Cheadle, Alan E. Berger, Felipe Andrade, Regina James, Kristen Johnson, Tonya Watkins, Jin Kyun Park, Yu-Chi Chen, Eva Ehrlich, Marissa Mullins, ... Stuart M. Levine
Arthritis & Rheumatism (2010-02-12) https://doi.org/chfbtv
DOI: 10.1002/art.27398 · PMID: 20155833 · PMCID: PMC2887718

82. **How transferable are features in deep neural networks?**
Jason Yosinski, Jeff Clune, Yoshua Bengio, Hod Lipson
arXiv (2014-12-09) https://arxiv.org/abs/1411.1792