Machine learning methods for rare diseases

This manuscript (permalink) was automatically generated from jaybee84/ml-in-rd@f091575 on January 29, 2021.

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 and accelerating basic research. 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 rare disease 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

Rare disease research is increasingly dependent on high-throughput profiling of samples and would greatly benefit from applications of machine learning (ML) in their analysis. Analyzing such high dimensional data from rare diseases (fewer than 200,000 cases in the United States [1]) is challenging, as datasets typically range from 20 to 99 samples [2]. Specialized computational methods that can learn patterns from small datasets that can be generalized to newly acquired data are required [3]. Lack of statistical power and the susceptibility of ML methods to misinterpretation and unstable performance pose challenges when datasets are small. Heterogeneity in available data creates additional difficulties. 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]. Due to the limited understanding of the biology of rare diseases, the symptoms or disease labels often come with a reasonable amount label-noise leading to a silver standard dataset [5]. A systematic review of application of ML in rare disease in the last 10 years uncovered 211 human data studies in 74 different rare diseases employing ensemble methods (36.0%), support vector machines (32.2%) and artificial neural networks (31.8%) [2]. The review also showed that most studies used ML for diagnosis (40.8%) or prognosis (38.4%), but studies aiming to improve treatment were infrequent (4.7%) [2]. 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 assembled from combining data from multiple institutions from differently processed specimens collected with geographical and chronological disparity. Consequently, data analysis techniques must be robust to challenges posed by small sample sizes, as well as technical artifacts present in aggregated 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

In rare diseases, the ability to get an enormous number of measurements from a vanishingly small number of samples using high-throughput methods 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). Perhaps counterintuitively, more feature-rich data can make a prediction problem more challenging. This is because statistical interrogation of a large number of measurements requires an abundance of samples or observations, which is often not the case in rare disease. This lack of samples gives rise to the "curse of dimensionality" (i.e., few samples but many features), which can be a major impediment in analyzing feature-rich data in sample-deficient contexts such as rare disease [6] In particular, increased numbers of features results in increased sparsity (missing observations), more dissimilarity between samples, and increased redundancy between individual features or combinations of features [7]; the consequence of this additional data is a more challenging prediction problem, rather than an easier one. Furthermore, rare disease data collection and aggregation methods can add to these challenges by introducing technical variability into the data at hand. In this section, we will discuss strategies like simplifying data and addressing technical artifacts through dimension reduction which can help mitigate these challenges. Dimensionality reduction methods including unsupervised approaches like multidimensional scaling (MDS), principal components analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) [8,9,10] can help 'compress' information from a large number of features into a smaller number of features. These techniques can be applied to characterize imaging data [12], mass cytometry data [13], 'omics data, and others. These methods not only help in reducing the number of features, but can also be used to visualize structure or artifacts in the data (e.g. [14]), define subgroups within data (e.g. [15], or for feature selection/extraction during application of specific machine learning models.[16] (Figure 1)

Rare disease datasets, like many other scenarios, can contain structure unrelated to the biology of the disease; e.g. structure related to batch, sample preparation methodology, or sequencing platform [17]. The consequences of these artifacts are amplified when samples are rare and the cohort contains several 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 structures in the data, though some like t-SNE and UMAP may require tuning of hyperparameters that affect the output.[10,17] 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. Additional important considerations for using dimensionality reduction methods such as criteria for selecting a dimensionality reduction method and interpretation of results are discussed in detail by Nguyen and Holmes.[19]

Beyond dimensionality reduction, unsupervised learning approaches such as k-means clustering or hierarchical clustering have also been used to characterize the structure present in genomic or imaging data. [20,21] If non-biological heterogeneity is detectable, common approaches like reprocessing the raw data using a single analysis pipeline (if the data are obtained from different sources), application of batch correction methods [22,23], and normalization of raw values[24] may be required to obtain value from these datasets.

Dimensionality reduction is, in fact, a type of representation learning (or feature learning), a process of learning low-dimensional representations or composites of features from raw data. Each learned composite feature becomes a new variable - representing a combination of original features - thereby reducing the dimension of the dataset. Representation learning through 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 [25], and use them to interpret test input data.[18,26] Putting constraints on the features learned by the model, through regularization, can help ensure that the learned representations are generalizable and avoid overfitting of the resulting model [27]. Low-dimensional representations trained on a collection of transcriptomic data can also be used as input to supervised machine learning methods.[29,30,31]

Representation learning generally requires many samples in complex biological systems and thus may appear to aggravate the curse of dimensionality. But it can be a powerful tool to learn low-dimensional patterns from large datasets and then find those patterns in smaller, related 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.

Couldn't load plugin.

Figure 1: Dimension reduction can help manage the curse of dimensionality in rare disease data

Manage model complexity while preserving the value of machine learning

Translating machine learning findings into testable hypotheses requires the applied models to be a) stable – the same predicted features should surface from the data if the model is run multiple times – and b) simple, as simple models guard against misinterpretation due to technical challenges. Meeting these requirements is challenging in 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 common ML techniques that can help improve the stability and simplicity of ML models applied to rare disease data. 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]. While resampling with replacement is commonly used to find models that are robust to overfitting ([**???**,33,34,35,36]); resampling without replacement, when applied to rare disease data generated confidence intervals for the model predictions by iteratively exposing the models to incomplete datasets (mimicking real world cases where most rare disease datasets are incomplete) [37].

Stability in predictions can also be achieved by combining various ML methods together (ensemble learning).(Figure[2]A-B) 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]. 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 [41,42,43]. Cascade learning implemented to identify rare disease patients from electronic health records from the general population utilized independent steps for feature extraction (using natural language processing based word2vec [44]), preliminary prediction (using an ensemble of decision trees with penalization for excessive tree-depth), and prediction refinement (using similarity of data points to resolve sample labels) [45]. Combining these three methods resulted in better performance than other methods when implemented on the silver standard dataset in isolation. The presence of multiple phenotypes (or classes) in rare disease datasets further decreases the available data-points per class. In such cases, a one-class-at-a-time cascade learning approach (where at each stage a binary classifier predicts a specific class against all others) has been found to produce simpler models that perform better compared to multi-class ensemble classifiers [46]. (Figure[2]D)

Simplification of models by making the feature space proportionate with the sample space can also be achieved through regularization. (Figure[2]C) Regularization can not only protect ML models from poor generalizability that results from overfitting (where the model performs well for the training data but poorly for new test data) [47], but also help penalize model complexity and reduce the feature space to build simpler models using limited datasets. Three popular regularized methods include ridge regression, LASSO, and elastic-net. Ridge regression can minimize the magnitude of the features, but 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] selects the most useful features, especially in presence of a large number of correlated features.

Rare variant discovery and immune cell signature discovery studies, like rare diseases, face challenges of the sparsity of observations (e.g. rare variants, or rare immune cells). Studies leveraging regularization in these problems can provide important insights into its 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.

Thus 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 .

Figure 2: Strategies to simplify models and stabilize predictions preserve the value of machine learning in rare disease. A-B) Strategies to build confidence in model predictions; A) schematic showing the concept of bootstrap, B) schematic showing the concept of ensemble learning to converge on reliable models; C-D) Strategies to simplify models by penalizing complexity in ML models; C) schematic showing the concept of regularization to selectively learn relevant features, D)schematic showing the concept of one-class-at-a-time learning to select few features at a time. Horizontal bars represent health of a model, models are represented as a network of nodes (features) and edges (relationships), nodes with solid edges represent real patterns, nodes with broken edges represent spurious patterns

Build upon prior knowledge and indirectly related data

Rare diseases often lack large, normalized datasets, limiting our ability to study key attributes of these diseases. Thus evaluating genotype-phenotype relationships or repurposing drugs using knowledge graphs can greatly benefit rare disease. Knowledge graphs (KGs) integrate related-but-different data types, creating a rich data source (e.g. Monarch Graph Database[61], hetionet[62], PheKnowLator[63], and the Global Network of Biomedical Relationships[64], Orphanet[65]). These graphs connect 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].(Figure[3]a) KGs may include links or nodes that are specific to the rare disease of interest (e.g. an FDA approved treatment would be a specific disease-compound link in the KG) as well as links that are more generalized (e.g. gene-gene interactions noted in the literature for a different disease).

Rare disease researchers can leverage the entities and relationships outside of the specific disease-context, including common comorbidities that are more prevalent conditions[67], for prediction. Such approaches have been used in rare disease research in areas such as drug repurposing[67] and disease classification[68]. Identifying KG encoding methods that can provide actionable insights for a specific rare disease application is an active area of 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, or an imaging anomaly present in rare and common diseases, for advancing our understanding of rare diseases.

Transfer learning, where a model trained for one task or domain (source domain) is applied to another related task or domain (target domain), can be supervised or unsupervised. Among various types of transfer learning we will mainly focus on feature-representation-transfer. (Figure[3]b) 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].

Other approaches related to transfer learning, multitask and few-shot learning, are forms of supervised learning that often rely on deep neural networks. In multitask learning classifiers use shared representations to learn multiple related but individual predictions (tasks) simultaneously [70]. (Figure[**???**]c-d) 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 various approaches and architectures underlie multitask and few-shot learning (see [71,72,73] for an overview), we will delve into a few selected studies to illustrate potential uses and limitations of these approaches in rare disease.

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]. The top-performer in a recent DREAM challenge for predicting drug sensitivity in cancer cell lines, including cell lines from rare cancers, was a multitask learning approach [TODO: CTD-squared Chemogenomic DREAM Challenge citation]. From these studies and others, it is clear that multitask learning is a promising approach for rare disease research albeit with some important, context-specific limitations. 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], 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 structural similarity among compounds was a requirement for this desired performance boost. In a study of rare pathologies in fundus photographs, a few-shot learning approach had a performance advantage over multitask learning, since predicting common conditions simultaneously resulted in a loss of performance for the multitask learner [12]. 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.

Figure 3: Strategies that build upon prior knowledge help ML models learn patterns in rare disease datasets.

Using composite approaches can be a powerful strategy

We have described multiple approaches for maximizing the success of machine learning applications in rare disease research throughout. In practice, it is rarely sufficient to use one of these techniques in isolation. Below, we highlight two recent works in the rare disease domain that draw on multiple concepts covered in the earlier sections. Feature-representation-transfer, which incorporates dimension reduction through representation learning, prior data, and regularization underlie both approaches.

Thousands of acute myeloid leukemia (AML) patient gene expression samples have been collected over time and are publicly available. Not all of these samples include the most relevant clinical or phenotypic data such as drug response. However, these publicly available data include an *in vitro* experiment that examined the response to 160 drugs for 30 AML patient samples, measured using genome-wide arrays which have tens of thousands of features [77]. Training on this drug response dataset alone poses challenges raised throughout – many features combined with small sample size can result in ML models that are of limited utility and sample size can also be prohibitive when performing representation learning. Dincer et al. trained a variational autoencoder on over 6500 AML samples (VAE; see definitions) [TODO: link between sections?] to reduce the dimensionality of the test set in an approach termed DeepProfile [78]. (Figure[4]a) The 30 AML test samples with drug response information were encoded using the VAE's low-dimensional representation or *transferred*, reducing the number of features from thousands to eight. LASSO linear regression models that used encodings as features had better performance than models that used individual gene expression values as features on average. In addition, the low-dimensional representation learned by the VAE captured more biological pathways than PCA, which may be attributable to the constraints on encodings imposed during the training process definitions [TODO: link between sections or cite what is in definitions?]. Similar results were observed for prediction of histopathology in another rare cancer (ovarian) [78].

While DeepProfile was centered on training on an individual disease and tissue combination, some rare diseases affect multiple tissues that a researcher may be interested in studying together for the

purpose of biological discovery. Studying multiple tissues poses significant challenges – features that can be appreciably measured may be tissue-specific even when looking at learned features or representations, a cross-tissue analysis may require an analyst to compare representations from multiple models and models trained on a low number of samples may learn representations that "lump together" multiple biological signals, reducing the interpretability of the results. To address these challenges, Taroni et al. trained Pathway-Level Information ExtractoR (PLIER) [79] on a large generic collection of human transcriptomic data (recount2 [80]). The authors 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 [81]. (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 [79]. PLIER includes constraints (regularization) such that some latent variables learned by the model will align with input gene sets and ideally latent variables will only be associated with a low number of related gene sets [79], which make it suitable for biological discovery or description of rare disease data. MultiPLIER allows us to use one model to describe multiple datasets instead of reconciling output from multiple models, which is highly beneficial when identifying commonalities among disease manifestations or affected tissues is a research goal. (Figure[4]b) (This benefit extends to studying multiple cohorts with a different model.) The inclusion of a large number of samples from diverse biological conditions in the training set results in models with desirable features (e.g., similar pathways are disentangled or separated out in the learned representations).

Taken together, DeepProfile [**???**] and MultiPLIER [81] suggest a combination of the techniques discussed throughout this article can be capitalized on for rare disease research. In cases where we have few samples from our disease of interest with the required phenotypic labels, we can leverage existing collections of data and knowledge if we select the models with the right attributes. The utility of DeepProfile and MultiPLIER stem from the fact that biological processes can be shared between biological contexts and that the methods underlying the approaches can effectively learn about those processes. 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 .

Figure 4: Combining multiple strategies strengthens the performance of ML models in rare disease

Outlook

Throughout this perspective, we have highlighted various challenges in applying ML methods to rare disease data as well as examples of approaches that address these challenges. Scarcity of samples, while significant, is not the only roadblock towards application of ML in rare disease data. The high dimensionality of modern data requires creative approaches, such as learning new representations of the data, to manage the curse of dimensionality. It further requires leveraging prior knowledge and transfer learning methods to appropriately interpret data. Additionally, anyone applying machine

learning methods on rare disease data should use techniques that increase confidence (such as bootstrapping) and penalize complexity of the resultant models (such as regularization) to enhance the generalizability of their work.

All of the approaches highlighted in this perspective come with certain challenges or inadequacies that breed mistrust in using these powerful techniques in rare disease. We believe that the same challenges that are currently considered major pitfalls in applying ML to rare disease can be great opportunities for data generation as well as method development in moving the field forward. During our journey through the various challenges, we identified two major areas where mindful strategies can immeasurably enhance the power of machine learning in rare disease and move the field forward.

Emphasis on not just "more n" but "more meaningful n"

Mindful addition of data is key for powering the next generation of analysis in rare disease data. While there are many techniques to collate rare data from different sources, incorrect data generation may hurt the end goal even if it adds to the size of the dataset. In our experience collaboration with domain experts have proved to be critical in gaining insight into potential sources of variation in the datasets. As an example, an neurofibromatosis type 1 (NF1) dataset was found to contain samples collected using vastly different surgical techniques (laser ablation and excision vs standard excision). [37] While the integrative analysis in the study using transfer learning techniques was able to minimize technique related signals [81], a more traditional analysis may have resulted in surfacing of substantial biological differences that are a consequence of process (e.g. activation of heat shock protein related pathways), not disease related biology. Such instances underline the fact that continuous collaboration with domain experts is needed to generate robust datasets in the future. A few such collaborations are beginning to show promise in generating valuable datasets for future use. [83]

In addition to sample scarcity, there is a dearth of comprehensive phenotypic-genotypic databases in rare disease. With the ubiquity of sequencing platforms, genomic data has been, relatively speaking, easy to gather for rare disease patients.[84,85,86] An important next step is to develop comprehensive comprehensive genomics-driven genotype-phenotype databases that can fuel interpretation of features extracted using ML methods. Finally, mindful sharing of data with proper metadata and attribution to enable prompt data reuse is of utmost important in building datasets that can be of great value in rare disease. [87]

Development of methods that reliably support mechanistic interrogation of specific rare diseases

The majority of ML methods for rare disease that we have investigated are applied to classification tasks. Conversely, we've found few examples of methodologies that interrogate biological mechanisms of rare diseases. This is likely a consequence of a dearth of methods that can tolerate the constraints imposed by rare disease research such as phenotypic heterogeneity and limited data. An intentional push towards developing methods or analytical workflows that address this will be critical to apply machine learning approaches to rare disease data.

Method development with rare disease applications in mind requires the developers to bear the responsibility of ensuring that the resulting model is *trustworthy*. The field of natural language processing has a few examples of how this can be achieved.[88] One way to increase trust in a developed model is by helping users understand the behavior of the developed model through providing explanations regarding why a certain model made certain predictions.[88] Another approach is to provide robust *error analysis* for newly developed models to help users understand the strengths and weaknesses of a model.[89,90,91] Adoption of these kind of approaches into

biological data and analysis is still rare but is quickly becoming necessary as machine learning approaches become mainstream in biomedicine.

Finally, methods that can reliably integrate disparate datasets will always remain a need of rare diseases. Moreover, combining data that originated from diverse modalities to create a complete picture of the disease related biology is increasingly becoming common. To facilitate such analyses in rare disease, methods that rely on finding structural correspondences between datasets (“anchors”) may be able to transform the status-quo of using machine learning methods in rare disease.[92; 93/~mahadeva/papers/IJCAI2011-DA.pdf; https://www.cell.com/cell/fulltext/S0092-8674(19)30559-8] Overall, we speculate that this an important burgeoning area of research, and we are optimistic about the future of applying machine learning approaches to rare disease.

Definitions

Unsupervised learning:

Machine learning algorithms which can learn features from unlabeled training data (e.g. datasets where the samples do not have disease or phenotype labels) to predict the class or phenotype of new or unseen test data are part of unsupervised learning. Examples of unsupervised learning include principal component analyses, multidimensional scaling, UMAP, t-SNE, k-means clustering etc [TODO - add REFs].

Supervised learning:

Machine learning algorithms that require training data with specific phenotype labels are part of supervised learning. Such algorithms learn correlations of features with the phenotype labels and use the learned correlations to predict the phenotype labels of unseen or new test data.

VAE:

Variational Autoencoders or VAEs are unsupervised neural networks that use hidden layers to learn or encode representations from available data while mapping the input data to the output data. VAEs are distinct from other autoencoders since the distribution of the encodings are regularized such that they are close to a normal distribution, which may contribute to learning more biologically relevant signals [18].

References

1.
Potomac Publishing
(2018-10-08) https://www.fda.gov/media/99546/download

2. **The use of machine learning in rare diseases: a scoping review**
Julia Schaefer, Moritz Lehne, Josef Schepers, Fabian Prasser, Sylvia Thun
Orphanet Journal of Rare Diseases (2020-06-09) https://doi.org/ghb3wx
DOI: 10.1186/s13023-020-01424-6 · PMID: 32517778 · PMCID: PMC7285453

3. **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

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. **The curse(s) of dimensionality**
Naomi Altman, Martin Krzywinski
Nature Methods (2018-05-31) https://doi.org/ghrqhp
DOI: 10.1038/s41592-018-0019-x · PMID: 29855577

8. **Handbook of Data Visualization**
Chun-houh Chen, Wolfgang Härdle, Antony Unwin
Springer Science and Business Media LLC (2008) https://doi.org/ckmkfp
DOI: 10.1007/978-3-540-33037-0

9. **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

10. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**
Leland McInnes, John Healy, James Melville
arXiv:1802.03426 [cs, stat] (2020-09-17) http://arxiv.org/abs/1802.03426

11. (2020-06-01) https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

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. **Ten quick tips for effective dimensionality reduction**
Lan Huong Nguyen, Susan Holmes
PLOS Computational Biology (2019-06-20) https://doi.org/gf3583
DOI: 10.1371/journal.pcbi.1006907 · PMID: 31220072 · PMCID: PMC6586259

20. **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

21. **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

22. **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

23. **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

24. **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

25. **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

26. **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

27. **Regularized Machine Learning in the Genetic Prediction of Complex Traits**
Sebastian Okser, Tapio Pahikkala, Antti Airola, Tapio Salakoski, Samuli Ripatti, Tero Aittokallio
PLoS Genetics (2014-11-13) https://doi.org/ghrqhq
DOI: 10.1371/journal.pgen.1004754 · PMID: 25393026 · PMCID: PMC4230844

28. **Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events**
Menelaos Pavlou, Gareth Ambler, Shaun Seaman, Maria De Iorio, Rumana Z Omar
Statistics in Medicine (2015-10-29) https://doi.org/ggn9zg
DOI: 10.1002/sim.6782 · PMID: 26514699 · PMCID: PMC4982098

29. **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

30. **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

31. **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

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. **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

36. **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

37. **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

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. **Component-based face detection**
B. Heiselet, T. Serre, M. Pontil, T. Poggio
Institute of Electrical and Electronics Engineers (IEEE) (2005-08-25) https://doi.org/c89p2b
DOI: 10.1109/cvpr.2001.990537

42. **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

43. **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

44. https://arxiv.org/abs/1301.3781v343

45. **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

46. **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 (2018-11) https://doi.org/gfh874
DOI: 10.1093/jamia/ocy109 · PMID: 30380082 · PMCID: PMC6213089

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. **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

62. **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

63. **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

64. **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

65. **Orphanet** https://www.orpha.net/consor/cgi-bin/index.php

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:1706.05098 [cs, stat] (2017-06-15) http://arxiv.org/abs/1706.05098

72. **A Survey on Multi-Task Learning**
Yu Zhang, Qiang Yang
arXiv:1707.08114 [cs] (2018-07-26) http://arxiv.org/abs/1707.08114

73. **Generalizing from a Few Examples: A Survey on Few-Shot Learning**
Yaqing Wang, Quanming Yao, James Kwok, Lionel M. Ni
arXiv:1904.05046 [cs] (2020-03-29) http://arxiv.org/abs/1904.05046

74. **Modeling Industrial ADMET Data with Multitask Networks**
Steven Kearnes, Brian Goldman, Vijay Pande
arXiv:1606.08793 [stat] (2017-01-12) http://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:1808.03331 [cs, stat] (2019-01-05) http://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. **A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia**
Su-In Lee, Safiye Celik, Benjamin A. Logsdon, Scott M. Lundberg, Timothy J. Martins, Vivian G. Oehler, Elihu H. Estey, Chris P. Miller, Sylvia Chien, Jin Dai, … Pamela S. Becker
Nature Communications (2018-01-03) https://doi.org/gcpx72
DOI: 10.1038/s41467-017-02465-5 · PMID: 29298978 · PMCID: PMC5752671

78. **DeepProfile: Deep learning of cancer molecular profiles for precision medicine**
Ayse Berceste Dincer, Safiye Celik, Naozumi Hiranuma, Su-In Lee
bioRxiv (2018-05-26) https://www.biorxiv.org/content/10.1101/278739v2
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. **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

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

83. **A clinically and genomically annotated nerve sheath tumor biospecimen repository**
Kai Pollard, Jineta Banerjee, Xengie Doan, Jiawan Wang, Xindi Guo, Robert Allaway, Shannon Langmead, Bronwyn Slobogean, Christian F. Meyer, David M. Loeb, … Christine A. Pratilas
Scientific Data (2020-06-19) https://doi.org/ghv6ch
DOI: 10.1038/s41597-020-0508-5 · PMID: 32561749 · PMCID: PMC7305302

84. **Rare-disease genetics in the era of next-generation sequencing: discovery to translation**
Kym M. Boycott, Megan R. Vanstone, Dennis E. Bulman, Alex E. MacKenzie
Nature Reviews Genetics (2013-09-03) https://doi.org/ghvhsd
DOI: 10.1038/nrg3555 · PMID: 23999272

85. **Paediatric genomics: diagnosing rare disease in children**
Caroline F. Wright, David R. FitzPatrick, Helen V. Firth
Nature Reviews Genetics (2018-02-05) https://doi.org/gcxbr8
DOI: 10.1038/nrg.2017.116 · PMID: 29398702

86. **Next-Generation Sequencing to Diagnose Suspected Genetic Disorders**
David R. Adams, Christine M. Eng
New England Journal of Medicine (2018-10-04) https://doi.org/gf49m7
DOI: 10.1056/nejmra1711801 · PMID: 30281996

87. **Responsible, practical genomic data sharing that accelerates research**
James Brian Byrd, Anna C. Greene, Deepashree Venkatesh Prasad, Xiaoqian Jiang, Casey S. Greene
Nature Reviews Genetics (2020-10) https://www.nature.com/articles/s41576-020-0257-5
DOI: 10.1038/s41576-020-0257-5

88. (2016-06-16) https://www.aclweb.org/anthology/N16-3020.pdf

89. (2019-07-15) https://www.aclweb.org/anthology/P19-1073.pdf

90. **Towards Automatic Error Analysis of Machine Translation Output**
Maja Popović, Hermann Ney
Computational Linguistics (2011-07-14) https://doi.org/10.1162/COLI_a_00072
DOI: 10.1162/coli_a_00072

91. **Recognizing names in biomedical texts: a machine learning approach**
G. Zhou, J. Zhang, J. Su, D. Shen, C. Tan
Bioinformatics (2004-02-10) https://doi.org/bxts7r
DOI: 10.1093/bioinformatics/bth060 · PMID: 14871877

92. (2010-06-15) https://www.aclweb.org/anthology/W06-1615.pdf

93. **College of Information & Computer Sciences**
College of Information & Computer Sciences
https://www.cics.umass.edu/