

# Machine learning methods for rare diseases

This manuscript ([permalink](#)) was automatically generated from [jaybee84/ml-in-rd@ae446ac](#) on February 8, 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)

The advent of high-throughput profiling methods such as genomics, transcriptomics, and other technologies has accelerated basic research and made deep characterization of patient samples routine. These approaches provide a rich portrait of genes, cellular pathways, and cell types involved in complex phenotypes. Machine learning is often a perfect fit for extracting disease-relevant patterns from these high dimensional datasets. Often, machine learning 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 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, and we propose that the methods community should prioritize the development of machine learning techniques for rare disease research.

# Introduction

---

, Rare disease research projects, as with many other biomedical domains, are increasingly using high-throughput profiling methods to better understand the nature of the disease. These profiling methods, including RNA sequencing (RNA-seq, whole genome sequencing, imaging data, electronic health record data, among others, generate large and complex data. The analysis of such complex data from rare disease will likely increasingly leverage machine learning (ML)-based methodologies. Indeed, 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%) [1]. While the review points to the increasing popularity of using ML methods in rare disease, there are various hurdles that are inherent to such datasets. ML based methods benefit from using large datasets, but analyzing high dimensional data from rare diseases datasets that typically contain fewer than one hundred samples is challenging [1]. Small datasets lead to a lack of statistical power and magnify the susceptibility of ML methods to misinterpretation and unstable performance. Additionally, 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 [2]. Due to limited understanding of the biology of rare diseases, the symptoms or disease labels often come with significant label-noise (a *silver standard* dataset) [3]. Thus, specialized computational methods that can learn patterns from small datasets and can generalize to newly acquired data are required for rare disease applications [4]. In this perspective, we first highlight ML approaches that address or better tolerate the limitations of rare disease data, and then discuss the future of ML applications in rare disease.

## Manage complex high-dimensional rare disease data

In rare diseases, the high throughput ‘omic’ methods generate highly dimensional data – data with many features, such as all of the mRNA transcripts in a sample – from a vanishingly small number of samples. A lack of samples gives rise to the “curse of dimensionality” (i.e., few samples but many features), which is an impediment in analyzing feature-rich data in sample-deficient contexts such as rare disease.[5] (Figure 1A-B) 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 [6], which creates a challenging prediction problem. 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 for reducing the feature space and addressing technical artifacts through dimensionality reduction.

Dimensionality reduction 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 a large number of features into a smaller number of features in an unsupervised manner [7,8,9]. (Figure 1C) These methods not only help in reducing the number of features in various types of data [11,12], but can also be used to visualize structure or artifacts in the data (e.g. [13]), to define sample subgroups (e.g. [14], or for feature selection and extraction during application of specific machine learning models.[15] (Figure 1D)

Rare disease datasets are often combined from multiple small studies leading to the confounding of biological characteristics with technical variables such as batch, sample preparation methodology, or sequencing platform [16]. Methods like PCA, MDS, t-SNE, and UMAP can successfully identify the effect of these variables on the original data, though t-SNE and UMAP may require tuning of hyperparameters that may effect the output [9,16]. Furthermore, testing multiple dimensionality reduction methods, rather than a single method, may be necessary to obtain a more comprehensive

portrait of the data [17]. Nguyen and Holmes discuss additional important considerations for using dimensionality reduction methods such as selection criteria and interpretation of results [18]. Beyond dimensionality reduction, other unsupervised learning approaches such as k-means clustering or hierarchical clustering have been used to characterize the structure present in genomic or imaging data [19,20]. Other approaches like reprocessing the data using a single pipeline (when data are obtained from multiple sources), using batch correction methods [21,22], and normalizing raw values [23] may be necessary to obtain meaningful insights from the data.

Dimensionality reduction, or more fundamentally, representation learning, learns low-dimensional representations (composite features) from the raw data. For example, representation learning through matrix factorization can extract features from transcriptomics datasets that are made of combinations of gene expression values found in the training data [24], and use them to interpret test data [17,25]. To ensure that the learned representations are generalizable to other data, the features learned by the model can be constrained through methods like regularization [26]. Representation learning generally requires many samples when applied to complex biological systems and therefore may appear to aggravate the curse of dimensionality. However, it can be a powerful tool to learn low-dimensional patterns from large datasets and then find those patterns in smaller, related datasets. In later sections, we will discuss this method of leveraging large datasets to reduce dimensionality in smaller datasets, also known as feature-representation-transfer learning.

Couldn't load plugin.

Figure 1: Dimension reduction can help manage the curse of dimensionality in rare disease data. A) Multiple datasets (shapes) with multiple phenotypes (purple, green) are combined for an analysis. The data (e.g., transcriptomic data) are highly dimensional, having thousands of features (f1-f100000). B) Evaluating the features, it appears that a combination of features (e.g., expressed genes) partition the purple samples from the green samples. C) Applying a dimensionality reduction method (e.g., PCA) condenses these features into new features (e.g., New Feature 1, a combination of f1, f2 .... f100000, and New Feature 2, a different combination of f1, f2 .... f100000). New Feature 1 describes the difference in input dataset (shapes) while New Feature 2 describes the difference in phenotype (color). D) New features (F1-F1000) can be used to interrogate the biology of the input samples, develop classification models, or use other analytical techniques that would have been more difficult with the original dataset dimensions.

## Manage model complexity while preserving the value of machine learning

Translating machine learning findings into testable hypotheses requires the ML models to be both stable – the same predicted features should surface from the data if the model is run multiple times – and simple, as simple models guard against misinterpretation. Meeting these requirements is challenging in rare disease datasets where label-noise is abundant. 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 resampling and combining various ML methods together (ensemble learning) can help achieve stability in predictions (Figure[2]A-B). Resampling without replacement can generate 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 [28]. Alternatively, resampling with replacement (bootstrapping) helps estimate population values from

datasets of limited size, and is also commonly used to find robust models when multiple models are combined into an ensemble ([29,30,31,32,33,34]). Ensemble learning methods like random forests use *bagging* (bootstrap aggregation) of independent decision trees that use similar parameters but different paths to form a consensus about the important predictive features [30,35,36,37,38]. However, standard ensemble learning has shown limited success in rare disease datasets with substantial label-noise. [TODO: citation?] This has led to the adoption of cascade learning, a variant of ensemble learning, where multiple methods leveraging distinct underlying assumptions are used in tandem; and augmented with algorithms like AdaBoost (boosting) to capture stable patterns existing in silver standard data [39,40,41]. For example, a cascade learning approach for identifying rare disease patients from electronic health record data utilized independent steps for feature extraction (word2vec [???]), preliminary prediction with ensembled decision trees, and prediction refinement using data similarity metrics [42]. 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 also 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 [43]. (Figure[2]D)

Regularization simplifies models by making the feature space proportionate with the sample space. (Figure[2]C) Regularization can not only protect ML models from poor generalizability caused by overfitting (where the model performs well on held-out training data but poorly on new test data) [44], but also reduces model complexity and the feature space to build simpler models. Three popular regularized methods, ridge regression, LASSO regression, and elastic-net regression, differ predominantly in how they modify features of the input data. Ridge regression can minimize the magnitude of the features, but cannot entirely remove features. LASSO regression, on the other hand, works well for selecting a few important features since it can minimize the magnitude of some features more than the others [45]. A combination of LASSO and ridge, elastic-net regression [46] 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, and may be useful models for examining the utility of regularization in scenarios with limited signal. For example, ridge regression has been used to combine rare variants into a single score to increase the signal of these variants [47], while LASSO has been implemented along with group penalties to identify gene variants [48,49]. Hybrid applications of LASSO in rare variant discovery studies like capturing combinations of variants [50,51], integrating with a probabilistic logistic Bayesian approach [52], combining feature selection methods with a generalized pooling strategy [53], and incorporating prior knowledge into the regularization step to select driver genes in a pathway of interest [54] have also proven beneficial. On the other hand, in the context of rare immune cell signature discovery, elastic-net regression was found to outperform other regression approaches [46,55,56,57]. Regularization methods like LASSO or elastic-net have been methods of choice for making models simpler by reducing the feature space in data with rare observations; use of these regularization approaches should be considered while working with rare disease datasets.

By employing bootstrapping, ensembling, and regularization, researchers may be able to generate more stable and simpler models to characterize the biological phenomena underlying rare diseases.

Couldn't load plugin.

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) A schematic showing the concept of bootstrap, B) A 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) A schematic showing the concept of regularization to selectively learn relevant features, D) A 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. Evaluating genotype-phenotype relationships or repurposing drugs using knowledge graphs may greatly benefit rare disease research. Knowledge graphs (KGs) integrate related-but-different data types, creating a rich data source (e.g. Monarch Graph Database [58], hetionet [59], PheKnowLator [60], and the Global Network of Biomedical Relationships [61], Orphanet [62]). These graphs connect genetic, functional, chemical, clinical, and ontological data to enable the exploration of relationships of data with disease phenotypes through manual review [63] or computational methods [64,65]. (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 in a knowledge graph outside of the specific disease-context [64]. Such approaches have been used in rare disease research in areas such as drug repurposing [64] and disease classification [65]. 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 both rare and common diseases, to advance 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, feature-representation-transfer approaches learn representations from the source domain and apply them to a target domain [66] (Figure[3]b). 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 [17]. Alternatively, multitask and few-shot learning are forms of supervised learning that often rely on deep neural networks.

While multitask learning classifiers use shared representations to learn multiple related but individual predictions (tasks) simultaneously [67], few-shot learning 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) [68,69,70] (Figure[3]c-d). Smaller datasets tended to benefit from multitask learning (due to task relatedness, *multitask effect*) [71], and the performance gains were generally context-dependent, i.e., multitask neural networks outperformed single-task networks for predicting complex rare phenotypes from EHR data or predicting drug sensitivity in rare cancer cell lines [72,73]. In contrast, one-shot or few-shot learning used prior knowledge to generalize a distance metric learned from input data to compare with a low number of new examples for prediction [70,74]. In another study, 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 [11]. Thus, transfer, multi-task, and few-shot learning are appealing approaches for rare disease applications, but their limits and potential utility are still open research questions.

Couldn't load plugin.

Figure 3: Strategies that build upon prior knowledge help ML models learn patterns in rare disease datasets. A) Knowledge graphs integrate different data types and may allow models to learn from connections that are rare disease-specific or happen in many biomedical contexts. B) Transfer learning is when a model trained in for one task or domain is applied to another, related task. C) Multitask learning uses models that learn and leverage shared representations to predict multiple, related tasks. D) Few-shot learning generalizes a previously trained model to predict a new, related task with a limited number of samples.

## Using composite approaches can be a powerful strategy

We have described multiple approaches for maximizing the success of ML applications in rare disease, but it is rarely sufficient to use any of these techniques in isolation. Below, we highlight two recent works in the rare disease domain that draw on concepts of feature-representation-transfer, use of prior data, and regularization.

A large public dataset of acute myeloid leukemia (AML) patient samples with no drug response data and a small *in vitro* experiment with drug response data form the basis of our first example [75]. Training an ML model on the small *in vitro* dataset alone faced the *curse of dimensionality* and the dataset size prohibited representation learning. Dincer et al. trained a variational autoencoder on the large AML patient dataset (VAE; see [definitions](#)) to learn meaningful representations in an approach termed DeepProfile [??] (Figure[4]a). The representations or *encodings* learned by the VAE were then *transferred* to the small *in vitro* dataset reducing it's number of features from thousands to eight, and improving the performance of the final LASSO linear regression model. In addition to improvement in performance, the *encodings* learned by the VAE captured more biological pathways than PCA, which may be attributable to the constraints on the encodings imposed during the training process [definitions](#). Similar results were observed for prediction of histopathology in another rare cancer dataset [76].

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 and a cross-tissue analysis may require an analyst to compare representations from multiple models. 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 a Pathway-Level Information Extractor (PLIER) (a matrix factorization approach that takes prior knowledge in the form of gene sets or pathways) on a large generic collection of human transcriptomic data [77]. PLIER used constraints (regularization) that learned *latent variables* aligned with a small number of input gene sets, making it suitable for biological discovery or description of rare disease data. The authors *transferred* the representations or *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 [78]. (Figure[4]b) MultiPLIER used one model to describe multiple datasets instead of reconciling output from multiple models, thus making it possible to identify commonalities among disease manifestations or affected tissues.



Taken together, DeepProfile [76] and MultiPLIER [78] suggest a combination of the techniques discussed throughout this article can be capitalized on for rare disease research. In cases where we have a few samples from our disease of interest with the required phenotypic labels, we can leverage prior knowledge if we select the models with the right attributes. DeepProfile and MultiPLIER capitalizes on the fact that biological processes can be shared between biological contexts and that the methods underlying the approaches can effectively learn about those processes.

Couldn't load plugin.

Figure 4: Combining multiple strategies strengthens the performance of ML models in rare disease. A) The authors of DeepProfile trained a variational autoencoder (VAE) to learn a representation from acute myeloid leukemia data without phenotype labels, transferred those representations to a small dataset with phenotype labels, and found that it improved prediction performance [76]. B) The authors of MultiPLIER trained a Pathway-Level Information Extractor (PLIER) model on a large, heterogeneous collection of expression data and transferred the representations to multiple datasets from unseen rare diseases [77].

## Outlook

Throughout this perspective, we highlighted various challenges in applying ML methods to rare disease data as well as examples of approaches that address these challenges. Small sample size, 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. Leveraging prior knowledge and transfer learning methods to appropriately interpret data is also required. Furthermore, we posit that researchers applying machine learning methods on rare disease data should use techniques that increase confidence (i.e., bootstrapping) and penalize complexity of the resultant models (i.e., regularization) to enhance the generalizability of their work.

All of the approaches highlighted in this perspective come with weaknesses that may undermine investigators' confidence in using these techniques for rare disease research. We believe that the challenges in applying ML to rare disease are opportunities for data generation and method development going forward. In particular, we identified two areas where that the field should explore to increase the utility of machine learning in rare disease.

*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, low-quality data may hurt the end goal even if it adds to the size of the dataset. In our experience, collaboration with domain experts has proved to be critical in gaining insight into potential sources of variation in the datasets. An anecdotal example from the authors' personal experience: conversations with a rare disease clinician revealed that samples in a particular tumor dataset were collected using vastly different surgical techniques (laser ablation and excision vs standard excision). This information that was not readily available to non-experts, but was obvious to the clinician. Such instances underline the fact that continuous collaboration with domain experts is needed to generate robust datasets in the future.

In addition to sample scarcity, there is a dearth of comprehensive phenotypic-genotypic databases in rare disease. Rare disease studies that collect genomic and phenotypic data are becoming more common. [79,80,81] However, an important next step is to develop comprehensive genomics-based genotype-phenotype databases that prioritize clinical and genomics data standards in order to 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. [82]

#### *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.[83] 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.[83] Another approach is to provide robust *error analysis* for newly developed models to help users understand the strengths and weaknesses of a model.[84,85,86] Adoption of these approaches into biomedical ML is quickly becoming necessary as machine learning approaches become mainstream in research and clinical settings.

Finally, methods that can reliably integrate disparate datasets will likely always remain a need in rare disease research. 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.[87,88,89] We speculate that this an important and burgeoning area of research, and we are optimistic about the future of applying machine learning approaches to rare diseases.

# 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, and k-means clustering [[7](#),[8](#),[9](#)].

## 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 [[17](#)].

# References

---

**1. 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](https://doi.org/10.1186/s13023-020-01424-6) · PMID: [32517778](https://pubmed.ncbi.nlm.nih.gov/32517778/) · PMCID: [PMC7285453](https://pubmed.ncbi.nlm.nih.gov/PMC7285453/)

**2. 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](https://doi.org/10.1093/jamia/ocw028) · PMID: [27174893](https://pubmed.ncbi.nlm.nih.gov/27174893/) · PMCID: [PMC5070523](https://pubmed.ncbi.nlm.nih.gov/PMC5070523/)

**3. 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/f5zdgg>  
DOI: [10.1109/tnnls.2013.2292894](https://doi.org/10.1109/tnnls.2013.2292894) · PMID: [24808033](https://pubmed.ncbi.nlm.nih.gov/24808033/)

**4. 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](https://doi.org/10.1016/j.ebiom.2019.08.027) · PMID: [31466916](https://pubmed.ncbi.nlm.nih.gov/31466916/) · PMCID: [PMC6796516](https://pubmed.ncbi.nlm.nih.gov/PMC6796516/)

**5. 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](https://doi.org/10.1038/nrc2294) · PMID: [18097463](https://pubmed.ncbi.nlm.nih.gov/18097463/) · PMCID: [PMC2238676](https://pubmed.ncbi.nlm.nih.gov/PMC2238676/)

**6. 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](https://doi.org/10.1038/s41592-018-0019-x) · PMID: [29855577](https://pubmed.ncbi.nlm.nih.gov/29855577/)

**7. 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](https://doi.org/10.1007/978-3-540-33037-0)

**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](https://doi.org/10.1098/rsta.2015.0202) · PMID: [26953178](https://pubmed.ncbi.nlm.nih.gov/26953178/) · PMCID: [PMC4792409](https://pubmed.ncbi.nlm.nih.gov/PMC4792409/)

**9. 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>

**10. Visualizing Data using t-SNE**

Laurens van der Maaten, Geoffrey Hinton

*Journal of Machine Learning Research* (2008) <http://jmlr.org/papers/v9/vandermaaten08a.html>

**11. Automatic detection of rare pathologies in fundus photographs using few-shot learning**

Gwenolé Quéllec, 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](https://doi.org/10.1016/j.media.2020.101660) · PMID: [32028213](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1038/ncomms14825) · PMID: [28382969](https://pubmed.ncbi.nlm.nih.gov/28382969/) · PMCID: [PMC5384229](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1038/s41467-019-13056-x) · PMID: [31780648](https://pubmed.ncbi.nlm.nih.gov/31780648/) · PMCID: [PMC6882829](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1038/s41467-020-15351-4) · PMID: [32210240](https://pubmed.ncbi.nlm.nih.gov/32210240/) · PMCID: [PMC7093466](https://pubmed.ncbi.nlm.nih.gov/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](https://doi.org/10.1007/978-3-030-03243-2_299-1)

**16. 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](https://doi.org/10.23915/distill.00002)

**17. 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](https://doi.org/10.1186/s13059-020-02021-3) · PMID: [32393369](https://pubmed.ncbi.nlm.nih.gov/32393369/) · PMCID: [PMC7212571](https://pubmed.ncbi.nlm.nih.gov/PMC7212571/)

**18. 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](https://doi.org/10.1371/journal.pcbi.1006907) · PMID: [31220072](https://pubmed.ncbi.nlm.nih.gov/31220072/) · PMCID: [PMC6586259](https://pubmed.ncbi.nlm.nih.gov/PMC6586259/)

**19. 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/dqgqbn6>

DOI: [10.1186/1471-2105-9-497](https://doi.org/10.1186/1471-2105-9-497) · PMID: [19038021](https://pubmed.ncbi.nlm.nih.gov/19038021/) · PMCID: [PMC2632677](https://pubmed.ncbi.nlm.nih.gov/PMC2632677/)

**20. 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](https://doi.org/10.1109/jbhi.2013.2276766) · PMID: [24808220](https://pubmed.ncbi.nlm.nih.gov/24808220/) · PMCID: [PMC5003052](https://pubmed.ncbi.nlm.nih.gov/PMC5003052/)

**21. 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](https://doi.org/10.1093/biostatistics/kxj037) · PMID: [16632515](https://pubmed.ncbi.nlm.nih.gov/16632515/)

**22. 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](https://doi.org/10.1093/nar/gku864) · PMID: [25294822](https://pubmed.ncbi.nlm.nih.gov/25294822/) · PMCID: [PMC4245966](https://pubmed.ncbi.nlm.nih.gov/PMC4245966/)

**23. 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](https://doi.org/10.1186/gb-2010-11-3-r25) · PMID: [20196867](https://pubmed.ncbi.nlm.nih.gov/20196867/) · PMCID: [PMC2864565](https://pubmed.ncbi.nlm.nih.gov/PMC2864565/)

**24. 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](https://doi.org/10.1038/s41467-020-14666-6) · PMID: [32051402](https://pubmed.ncbi.nlm.nih.gov/32051402/) · PMCID: [PMC7016183](https://pubmed.ncbi.nlm.nih.gov/PMC7016183/)

**25. 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](https://doi.org/10.1093/bioinformatics/btq503) · PMID: [20810601](https://pubmed.ncbi.nlm.nih.gov/20810601/) · PMCID: [PMC3025742](https://pubmed.ncbi.nlm.nih.gov/PMC3025742/)

**26. 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](https://doi.org/10.1371/journal.pgen.1004754) · PMID: [25393026](https://pubmed.ncbi.nlm.nih.gov/25393026/) · PMCID: [PMC4230844](https://pubmed.ncbi.nlm.nih.gov/PMC4230844/)

**27. 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](https://doi.org/10.1002/sim.6782) · PMID: [26514699](https://pubmed.ncbi.nlm.nih.gov/26514699/) · PMCID: [PMC4982098](https://pubmed.ncbi.nlm.nih.gov/PMC4982098/)

**28. 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](https://doi.org/10.3390/genes11020226) · PMID: [32098059](https://pubmed.ncbi.nlm.nih.gov/32098059/) · PMCID: [PMC7073563](https://pubmed.ncbi.nlm.nih.gov/PMC7073563/)

**29. 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](https://doi.org/10.1080/01621459.1997.10474007)

**30. Random Forests**

Leo Breiman

*Machine Learning* (2001-10-01) <https://doi.org/10.1023/A:1010933404324>

DOI: [10.1023/a:1010933404324](https://doi.org/10.1023/a:1010933404324)

**31. 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](https://doi.org/10.1198/0003130043277)

**32. Bootstrap for neural model selection**

Riadh Kallel, Marie Cottrell, Vincent Vigneron

*Neurocomputing* (2002-10) <https://doi.org/c8xpgz>

DOI: [10.1016/s0925-2312\(01\)00650-6](https://doi.org/10.1016/s0925-2312(01)00650-6)

**33. 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](https://doi.org/10.1016/j.neucom.2004.11.017)

**34. 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](https://doi.org/10.1002/sim.4780111607) · PMID: [1293671](https://pubmed.ncbi.nlm.nih.gov/1293671/)

**35. 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](https://doi.org/10.1186/1472-6947-13-134) · PMID: [24321610](https://pubmed.ncbi.nlm.nih.gov/24321610/) · PMCID: [PMC4029400](https://pubmed.ncbi.nlm.nih.gov/PMC4029400/)

**36. Analyzing bagging**

Peter Bühlmann, Bin Yu

*The Annals of Statistics* (2002-08) <https://doi.org/btmtjp>

DOI: [10.1214/aos/1031689014](https://doi.org/10.1214/aos/1031689014)

**37. 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](https://doi.org/10.1177/2045894019890549) · PMID: [31798836](https://pubmed.ncbi.nlm.nih.gov/31798836/) · PMCID: [PMC6868581](https://pubmed.ncbi.nlm.nih.gov/PMC6868581/)

**38. 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](https://doi.org/10.1016/s0031-3203(02)00169-3)

**39. 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](https://doi.org/10.1109/cvpr.2001.990537)

**40. 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](https://doi.org/10.1007/978-3-540-75175-5_16)

**41. 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](https://doi.org/10.1109/icpr.2004.1334680)

**42. 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](https://pubmed.ncbi.nlm.nih.gov/30815073/) · PMCID: [PMC6371307](https://pubmed.ncbi.nlm.nih.gov/PMC6371307/)

**43. 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](https://doi.org/10.1093/jamia/ocy109) · PMID: [30380082](https://pubmed.ncbi.nlm.nih.gov/30380082/) · PMCID: [PMC6213089](https://pubmed.ncbi.nlm.nih.gov/PMC6213089/)

**44. 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](https://doi.org/10.1073/pnas.1900654116) · PMID: [31619572](https://pubmed.ncbi.nlm.nih.gov/31619572/) · PMCID: [PMC6825274](https://pubmed.ncbi.nlm.nih.gov/PMC6825274/)

**45. Regularization**

Jake Lever, Martin Krzywinski, Naomi Altman

*Nature Methods* (2016-09-29) <https://doi.org/gf3zrr>

DOI: [10.1038/nmeth.4014](https://doi.org/10.1038/nmeth.4014)

**46. 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](https://doi.org/10.1111/j.1467-9868.2005.00503.x)

**47. 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](https://doi.org/10.1371/journal.pone.0044173) · PMID: [22952918](https://pubmed.ncbi.nlm.nih.gov/22952918/) · PMCID: [PMC3429469](https://pubmed.ncbi.nlm.nih.gov/PMC3429469/)

**48. 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](https://doi.org/10.1038/nrg2867) · PMID: [20940738](https://pubmed.ncbi.nlm.nih.gov/20940738/) · PMCID: [PMC3743540](https://pubmed.ncbi.nlm.nih.gov/PMC3743540/)

**49. 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](https://doi.org/10.1093/bioinformatics/btq448) · PMID: [20693321](https://pubmed.ncbi.nlm.nih.gov/20693321/) · PMCID: [PMC3025646](https://pubmed.ncbi.nlm.nih.gov/PMC3025646/)

**50. 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](https://doi.org/10.1016/j.ajhg.2008.06.024) · PMID: [18691683](https://pubmed.ncbi.nlm.nih.gov/18691683/) · PMCID: [PMC2842185](https://pubmed.ncbi.nlm.nih.gov/PMC2842185/)

**51. 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](https://doi.org/10.1186/1753-6561-5-s9-s113) · PMID: [22373209](https://pubmed.ncbi.nlm.nih.gov/22373209/) · PMCID: [PMC3287837](https://pubmed.ncbi.nlm.nih.gov/PMC3287837/)

**52. 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](https://doi.org/10.4137/cin.s17290) · PMID: [25733797](https://pubmed.ncbi.nlm.nih.gov/25733797/) · PMCID: [PMC4332044](https://pubmed.ncbi.nlm.nih.gov/PMC4332044/)

**53. 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](https://doi.org/10.1371/journal.pone.0041694) · PMID: [22916111](https://pubmed.ncbi.nlm.nih.gov/22916111/) · PMCID: [PMC3420665](https://pubmed.ncbi.nlm.nih.gov/PMC3420665/)

**54. 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](https://doi.org/10.1080/10618600.2012.681250)

**55. 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.combiomed.2015.10.008](https://doi.org/10.1016/j.combiomed.2015.10.008) · PMID: [26520484](https://pubmed.ncbi.nlm.nih.gov/26520484/)

**56. 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](https://doi.org/10.1186/1471-2105-14-198) · PMID: [23777239](https://pubmed.ncbi.nlm.nih.gov/23777239/) · PMCID: [PMC3718705](https://pubmed.ncbi.nlm.nih.gov/PMC3718705/)

**57. 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. Kline

*BMC Bioinformatics* (2019-08-22) <https://doi.org/gg5hmj>

DOI: [10.1186/s12859-019-2994-z](https://doi.org/10.1186/s12859-019-2994-z) · PMID: [31438843](https://pubmed.ncbi.nlm.nih.gov/31438843/) · PMCID: [PMC6704630](https://pubmed.ncbi.nlm.nih.gov/PMC6704630/)

**58. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species**

Christopher J. Mungall, Julie A. McMurtry, 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](https://doi.org/10.1093/nar/gkw1128) · PMID: [27899636](https://pubmed.ncbi.nlm.nih.gov/27899636/) · PMCID: [PMC5210586](https://pubmed.ncbi.nlm.nih.gov/PMC5210586/)

**59. 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](https://doi.org/10.7554/elife.26726) · PMID: [28936969](https://pubmed.ncbi.nlm.nih.gov/28936969/) · PMCID: [PMC5640425](https://pubmed.ncbi.nlm.nih.gov/PMC5640425/)

**60. 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](https://doi.org/10.1101/2020.04.30.071407)

**61. 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](https://doi.org/10.1093/bioinformatics/bty114) · PMID: [29490008](https://pubmed.ncbi.nlm.nih.gov/29490008/) · PMCID: [PMC6061699](https://pubmed.ncbi.nlm.nih.gov/PMC6061699/)

**62. Orphanet** <https://www.orpha.net/consor/cgi-bin/index.php>

**63. 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](https://doi.org/10.1093/database/baaa015) · PMID: [32283553](https://pubmed.ncbi.nlm.nih.gov/32283553/) · PMCID: [PMC7153956](https://pubmed.ncbi.nlm.nih.gov/PMC7153956/)

**64. 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](https://doi.org/10.1101/727925)

**65. 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](https://doi.org/10.1186/s12911-019-0938-1) · PMID: [31801534](https://pubmed.ncbi.nlm.nih.gov/31801534/) · PMCID: [PMC6894101](https://pubmed.ncbi.nlm.nih.gov/PMC6894101/)

**66. 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](https://doi.org/10.1109/tkde.2009.191)

**67. Multitask Learning**

Rich Caruana  
*Machine Learning* (1997-07-01) <https://doi.org/10.1023/A:1007379606734>  
DOI: [10.1023/a:1007379606734](https://doi.org/10.1023/a:1007379606734)

**68. 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>

69. **A Survey on Multi-Task Learning**  
Yu Zhang, Qiang Yang  
*arXiv:1707.08114 [cs]* (2018-07-26) <http://arxiv.org/abs/1707.08114>
70. **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>
71. **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>
72. **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>
73. **A Community Challenge for Pancancer Drug Mechanism of Action Inference from Perturbational Profile Data**  
Eugene F. Douglass, Robert J Allaway, Bence Szalai, Wenyu Wang, Tingzhong Tian, Adrià Fernández-Torras, Ron Realubit, Charles Karan, Shuyu Zheng, Alberto Pessia, ... DREAM CTD-squared Pancancer Drug Activity Challenge Consortium  
*Cold Spring Harbor Laboratory* (2020-12-22) <https://doi.org/ghxxk4>  
DOI: [10.1101/2020.12.21.423514](https://doi.org/10.1101/2020.12.21.423514)
74. **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](https://doi.org/10.1021/acscentsci.6b00367) · PMID: [28470045](https://pubmed.ncbi.nlm.nih.gov/28470045/) · PMCID: [PMC5408335](https://pubmed.ncbi.nlm.nih.gov/PMC5408335/)
75. **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](https://doi.org/10.1038/s41467-017-02465-5) · PMID: [29298978](https://pubmed.ncbi.nlm.nih.gov/29298978/) · PMCID: [PMC5752671](https://pubmed.ncbi.nlm.nih.gov/PMC5752671/)
76. **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](https://doi.org/10.1101/278739)
77. **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](https://doi.org/10.1038/s41592-019-0456-1) · PMID: [31249421](https://pubmed.ncbi.nlm.nih.gov/31249421/) · PMCID: [PMC7262669](https://pubmed.ncbi.nlm.nih.gov/PMC7262669/)
78. **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](https://doi.org/10.1016/j.cels.2019.04.003) · PMID: [31121115](https://pubmed.ncbi.nlm.nih.gov/31121115/) · PMCID: [PMC6538307](https://pubmed.ncbi.nlm.nih.gov/PMC6538307/)

79. **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/ghvhdsd>  
DOI: [10.1038/nrg3555](https://doi.org/10.1038/nrg3555) · PMID: [23999272](https://pubmed.ncbi.nlm.nih.gov/23999272/)
80. **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](https://doi.org/10.1038/nrg.2017.116) · PMID: [29398702](https://pubmed.ncbi.nlm.nih.gov/29398702/)
81. **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](https://doi.org/10.1056/nejmra1711801) · PMID: [30281996](https://pubmed.ncbi.nlm.nih.gov/30281996/)
82. **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](https://doi.org/10.1038/s41576-020-0257-5)
83. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**  
Marco Ribeiro, Sameer Singh, Carlos Guestrin  
*Association for Computational Linguistics (ACL)* (2016) <https://doi.org/gg8gggh>  
DOI: [10.18653/v1/n16-3020](https://doi.org/10.18653/v1/n16-3020)
84. **Errudite: Scalable, Reproducible, and Testable Error Analysis**  
Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, Daniel Weld  
*Association for Computational Linguistics (ACL)* (2019) <https://doi.org/ggb9kk>  
DOI: [10.18653/v1/p19-1073](https://doi.org/10.18653/v1/p19-1073)
85. **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](https://doi.org/10.1162/COLI_a_00072)  
DOI: [10.1162/coli\\_a\\_00072](https://doi.org/10.1162/coli_a_00072)
86. **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](https://doi.org/10.1093/bioinformatics/bth060) · PMID: [14871877](https://pubmed.ncbi.nlm.nih.gov/14871877/)
87. **Domain Adaptation with Structural Correspondence Learning**  
John Blitzer, Ryan McDonald, Fernando Pereira  
*Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (2006-07) <https://www.aclweb.org/anthology/W06-1615>
88. **Heterogeneous domain adaptation using manifold alignment**  
Chang Wang, Sridhar Mahadevan  
*Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two* (2011-07-16) <https://dl.acm.org/doi/10.5555/2283516.2283652>  
ISBN: [9781577355144](https://doi.org/9781577355144)
89. **Comprehensive Integration of Single-Cell Data**  
Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M.

Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, Rahul Satija

*Cell* (2019-06) <https://doi.org/gf3sxv>

DOI: [10.1016/j.cell.2019.05.031](https://doi.org/10.1016/j.cell.2019.05.031) · PMID: [31178118](https://pubmed.ncbi.nlm.nih.gov/31178118/) · PMCID: [PMC6687398](https://pubmed.ncbi.nlm.nih.gov/PMC6687398/)

90. **Some methods for classification and analysis of multivariate observations**

J. MacQueen

*The Regents of the University of California* (1967)

<https://projecteuclid.org/euclid.bsm/1200512992>