


Machine learning in rare disease

Received: 16 March 2021

Accepted: 22 April 2023

Published online: 29 May 2023

 Check for updates

Jineta Banerjee^{1,4}, Jaclyn N. Taroni^{2,4}, Robert J. Allaway¹,
Deepashree Venkatesh Prasad², Justin Guinney¹ & Casey Greene²✉

High-throughput profiling methods (such as genomics or imaging) have accelerated basic research and made deep molecular characterization of patient samples routine. These approaches provide a rich portrait of genes, molecular pathways and cell types involved in disease phenotypes. Machine learning (ML) can be a useful tool for extracting disease-relevant patterns from high-dimensional datasets. However, depending upon the complexity of the biological question, machine learning often requires many samples to identify recurrent and biologically meaningful patterns. Rare diseases are inherently limited in clinical cases, leading to few samples to study. In this Perspective, we outline the challenges and emerging solutions for using ML for small sample sets, specifically in rare diseases. Advances in ML methods for rare diseases are likely to be informative for applications beyond rare diseases for which few samples exist with high-dimensional data. We propose that the method community prioritize the development of ML techniques for rare disease research.

Rare disease researchers increasingly depend on ML to analyze high-dimensional datasets. A systematic review of ML applications in rare diseases as defined in the European Union (fewer than 5 patients per 10,000 people) uncovered 211 human studies that used ML to study 74 rare diseases over the last 10 years¹. ML can be a powerful tool in biomedical research, but it does not come without pitfalls, some of which are magnified in a rare disease context². In this Perspective, we discuss considerations for using two types of ML, supervised and unsupervised learning, in the study of rare diseases, with a specific focus on high-dimensional molecular data.

ML algorithms are computational methods that identify patterns in data, often in the form of lower-dimensional representations that can be used to perform useful computational tasks. Supervised learning algorithms must be trained with data in which samples are ‘labeled’ with a trait of interest, such as a biological or clinical phenotype. Supervised methods can learn correlations of features (for example, expression measurements of a large number of genes) that may be associated with these labels to predict or infer these labels in unlabeled data, such as predicting which patients will or will not respond to treatment. Therefore, if a study aims to classify patients with a rare disease into disease subtypes based on high-throughput molecular profiling, a supervised

ML algorithm is appropriate to carry out this task. Conversely, unsupervised learning algorithms learn patterns or features from unlabeled data. In the absence of known disease subtypes, unsupervised ML approaches can be applied to gene expression data to identify groups of samples with similar patterns of molecular states or pathway activity³. Unsupervised approaches can also extract combinations of features (for example, genes) that may describe a certain cell type or pathway. See Box 1 for more examples of how ML can be used in rare disease research. To implement ML models in rare disease research, one also needs to consider the components and the design of ML experiments to better inform the construction of datasets appropriate for such experiments. See Boxes 2 and 3 for a deeper understanding of the different components and the design of ML experiments, respectively.

While ML can be a useful tool, there are challenges in applying ML to rare disease datasets. ML methods are generally most effective when using large datasets; analyzing high-dimensional biomedical data such as gene expression with many thousands of features from rare disease datasets that typically contain relatively few samples is challenging^{1,4}. Small-sample datasets tend to lack statistical power and magnify the susceptibility of ML to misinterpretation and unstable performance. With insufficient data, an unsupervised model will fail

¹Sage Bionetworks, Seattle, WA, USA. ²Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, PA, USA. ³Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO, USA. ⁴These authors contributed equally: Jineta Banerjee, Jaclyn N. Taroni. ✉e-mail: casey.s.greene@cuanuschutz.edu

BOX 1

Common uses for ML in rare disease

(a) Identifying patients with rare diseases

ML can be used to identify features in high-dimensional data that correlate strongly with a patient or sample phenotype and subsequently predict the presence or absence of a rare disease. For example, supervised ML models can be trained on electronic health records, genetic data or medical images to identify potential new patients with a rare disease.

(b) Drug discovery or repurposing

ML can help identify potential drug candidates for rare diseases. For example, unsupervised and supervised algorithms trained on genetic and molecular data from high-throughput screens can identify new therapeutic targets for a rare disease. Additionally, algorithms using knowledge graphs, genomic data and databases of existing approved drugs can identify potential therapeutic candidates for rare diseases.

(c) Clinical trial-design improvement

Optimized study design and identification of appropriate trial participants can greatly reduce costs while increasing the likelihood of successful outcomes for clinical trials. ML approaches can benefit clinical trial study design. For example, unsupervised ML approaches can be used to identify subgroups of patients who are more likely to respond well to a particular treatment. Supervised ML approaches can be used to predict drug response in patients with rare diseases.

(d) Molecular subtyping of disease

Rare diseases often show overlapping and heterogeneous phenotypes. ML approaches can be used to identify molecular subtypes of the disease for better understanding. For example, unsupervised ML approaches can help identify new subtypes of a rare disease using molecular and genetic data. The same approaches can help identify the important molecular features that define the subtypes.

(e) Patient prognosis prediction

Rare diseases can suffer from lack of in-depth understanding of disease mechanism. Biomarkers or clinical features that correlate strongly with adverse outcomes can be beneficial in predicting prognosis of a patient. Supervised ML algorithms can be useful in identifying factors contributing to the risk of adverse outcomes or progression to advanced disease in patients with rare diseases. Patient stratification can help identify patient subpopulations who can benefit from early and aggressive interventions.

to identify patterns that are useful for biological discovery. In the case of supervised models, the models can be adversely impacted if sample labels are uncertain or contain ‘label noise’ (ref. 5). Datasets with high label noise decrease prediction accuracy and necessitate larger sample sizes during the process by which models learn patterns that distinguish samples in different classes⁶ (model training; Box 3). Rare disease datasets often come with substantial label noise. For example, if classifications of rare disease subtypes evolve over time, researchers constructing datasets for ML research may find that cohorts collected at different time periods do not have comparable labels. Additionally, a supervised ML model is of limited utility if it can only accurately predict

sample labels in the data it was trained on, also known as overfitting. Instead, most researchers aspire to develop models that generalize (maintain performance) when applied to new data that have not yet been ‘seen’ by the model.

While we expect ML in rare disease research to continue to increase in popularity, in our opinion, the field requires methods that can learn patterns from small datasets and can generalize to newly acquired data⁷. In this Perspective, we highlight approaches that address or better tolerate the limitations of rare disease data and discuss the future of ML applications in rare disease.

Constructing ML-ready rare disease datasets

High-throughput ‘omic’ assays can generate thousands to billions of measurements from whole-transcriptome and whole-genome sequencing, respectively, resulting in high-dimensional datasets. A typical rare disease dataset consists of a small number of samples¹, leading to the ‘curse of dimensionality’ in which the feature space is much larger than the sample space, increasing the difficulty in building highly generalizable models⁸. A larger feature space can contribute to higher data missingness (sparsity), more dissimilarity between samples (variance) and increased redundancy among individual features or combinations (multicollinearity)⁹, all of which contribute to challenges in ML implementation.

An important factor in ML is model performance: the accuracy of a supervised model in identifying patterns relevant for a biological question, or the reliability of an unsupervised model in identifying hypothetical biological patterns supported by post hoc validation. When small sample sizes compromise an ML model’s performance, two approaches can be taken to manage sparsity, variance and multicollinearity: (1) increase the number of samples and (2) improve the quality of samples. In the first approach, appropriate training, evaluation and held-out validation sets could be constructed by combining multiple rare disease cohorts (Fig. 1a and Boxes 2 and 3). When combining datasets, special attention should be directed toward data harmonization, as data-collection methods can differ between cohorts. Without careful selection of aggregation methods, one may introduce technical (in contrast to biological) variability into the combined dataset and negatively impact the ML model’s ability to learn or detect meaningful signals. Steps such as reprocessing data using a single pipeline, using batch correction methods^{10,11} and normalizing raw values appropriately without affecting the underlying variance in the data¹² may be necessary to mitigate unwanted variability (Fig. 1a). Data harmonization may also entail standardization of sample labels using biomedical ontologies to normalize how samples are described across multiple datasets.

How does one know whether a composite dataset has undergone proper harmonization and annotation? Ideally, the dominant patterns of the composite dataset reflect variables of interest, such as phenotype labels rather than technical labels. In the latter case, this suggests that the datasets used to generate the composite dataset need to be corrected to overcome differences in how the data were generated or collected. In the next section, we discuss approaches that help identify and visualize structure in datasets to determine whether composite rare disease datasets are appropriate for ML use.

Learning representations from rare disease data

Dimensionality-reduction methods help explore and visualize underlying structure in the data (for example, ref. 13), to define sample subgroups (for example, ref. 14) or for feature selection and extraction during application of specific ML models¹⁵ (Fig. 2c). Unsupervised methods, in finding low-dimensional patterns in data, can ‘compress’ information from a large number of features into a smaller number of features^{16–18} (Fig. 2). A method commonly used for dimensionality reduction is principal-component analysis (PCA). PCA identifies higher-order features, termed principal components (PCs), that are combinations of original features. The PCs are calculated in a way that maximizes the

BOX 2

Understanding components of ML experiments to inform requirements for data

ML algorithms identify patterns that explain or fit a given dataset. Every ML algorithm goes through ‘training’, during which it identifies underlying patterns in a given dataset to create a ‘trained’ algorithm (a model), and ‘testing’, during which the model applies the identified patterns to unseen data points. Typically, an ML algorithm is provided with (1) a training dataset, (2) an evaluation dataset and (3) a held-out validation dataset. These input data can be images, text, numbers or other types of data, typically encoded as a numerical representation of the input data. A training dataset is used by the model to learn underlying patterns from the features present in the data of interest. An evaluation dataset is a small and previously unused dataset that is used during the training phase to help the model iteratively update its parameters (that is, hyperparameter tuning or model tuning). In many cases, a large training set may be subdivided to form a smaller training dataset and the evaluation dataset, both of which are used to train the model. In the testing phase, a completely new or unseen test dataset or held-out validation set is used to test whether the patterns learned by the model hold true in new data (that is, they are generalizable). While the evaluation dataset helps us refine a model’s fit to patterns in the training data, the held-out validation set helps us test the generalizability of the model.

If a model is generalizable, it is able to make accurate predictions on new data. High generalizability of a model on previously unseen data suggests that the model has identified important patterns in the data that are not unique to the data used for training and tuning. Generalizability can be affected if data leakage occurs during training of the model, that is, if a model is exposed to the same or related data points in both the training set and the held-out validation set. Ensuring absence of any overlap or relatedness among data points or samples used in the training set and the evaluation set is important to avoid data leakage during model training. Specifically, in cases of rare genetic diseases in which many samples can contain familial relationships or data from the same patient could be collected by multiple specialists at different clinical facilities, special care should be taken while crafting the training and testing sets to ensure that no data leakage occurred and that the trained model has high generalizability.

amount of information (variance) that they contain and ensures that each PC is uncorrelated with the other PCs¹⁷. In practice, researchers often use the first few PCs to reduce the dimensionality without removing what may be important biological variability in the data. Nguyen and Holmes highlight the use of the ‘elbow method’ to select the number of appropriate dimensions¹⁹. Multidimensional scaling, *t*-distributed stochastic neighbor embedding and uniform manifold approximation and projection are other popular dimension-reduction methods, often used for low-dimensional visualization and interpretation of data^{18,20}. Testing multiple dimensionality-reduction methods may be necessary to obtain a more comprehensive portrait of the data²¹. Other unsupervised learning approaches such as *k* means or hierarchical clustering are used to characterize structure in genomic and imaging data^{22,23}. Dimensionality-reduction methods are a subset of a type of ML called representation learning. Representation learning methods have been

BOX 3

Understanding experimental design of training and testing ML models to inform requirements for data

The implementation of an ML experiment begins with splitting a single dataset of interest such that a large proportion of the data is used for training and the remaining data are used for testing or validation as the held-out validation dataset. The training dataset is generally subdivided into the training dataset and the evaluation dataset. Ideally, a held-out validation dataset is an entirely new study or cohort, as researchers typically aim to build models that generalize to unseen, newly generated data. In rare diseases for which multiple datasets may be combined to make a large enough training dataset, special care should be taken to standardize the features and the patterns therein. Although ML algorithms generally expect that datasets have uniform features, normalizing training and testing data together may introduce similarities between samples (causing inadvertent data leakage) that hamper the goal of training models that are highly generalizable.

The iterative training phase helps the model learn important patterns in the training dataset and then use the evaluation dataset to test for errors in prediction and update its learning parameters (hyperparameter tuning). The method by which the trained model is applied to the evaluation dataset to measure performance and update hyperparameters is called cross-validation. There are multiple approaches that can be deployed to maximally use the available data when generating training and evaluation datasets, for example, leave-*p*-out cross-validation, leave-one-out cross-validation, *k*-fold cross-validation and Monte-Carlo random subsampling cross-validation⁸³. In the case of *k*-fold cross-validation, a given dataset is shuffled randomly and split into *k* parts. One of the *k* parts is reserved as the evaluation dataset, and the rest are combined and used as the training dataset. In the next iteration, a different part is used as the evaluation dataset, while the rest are used for training. To avoid data leakage and to promote generalization of models to new studies, researchers can use study-wise cross-validation, such that all samples from a study are in the same fold and no individual study is represented in both the training and evaluation datasets. Once the model has iterated through all *k* parts of the training and evaluation datasets, it is ready to be tested on the held-out validation dataset (Fig. 1b).

The held-out validation dataset is exposed to the model only once to estimate the accuracy of the model. High accuracy of a model during cross-validation but low accuracy on the held-out validation dataset is a sign that the model has become overfit to the training set and has low generalizability. If there is evidence of overfitting, researchers should revisit the construction of the dataset to ensure that they meet the best practices outlined above.

It is important to note that accuracy alone may not be the best measure of performance in rare disease datasets. A model tested for identifying rare disease samples may still achieve high accuracy if it identifies every sample as a non-rare disease sample. Measures that are more suitable to handle class imbalance, such as the *κ* statistic or area-under-the-precision–recall curve⁸⁴, are better metrics for model performance for rare disease.

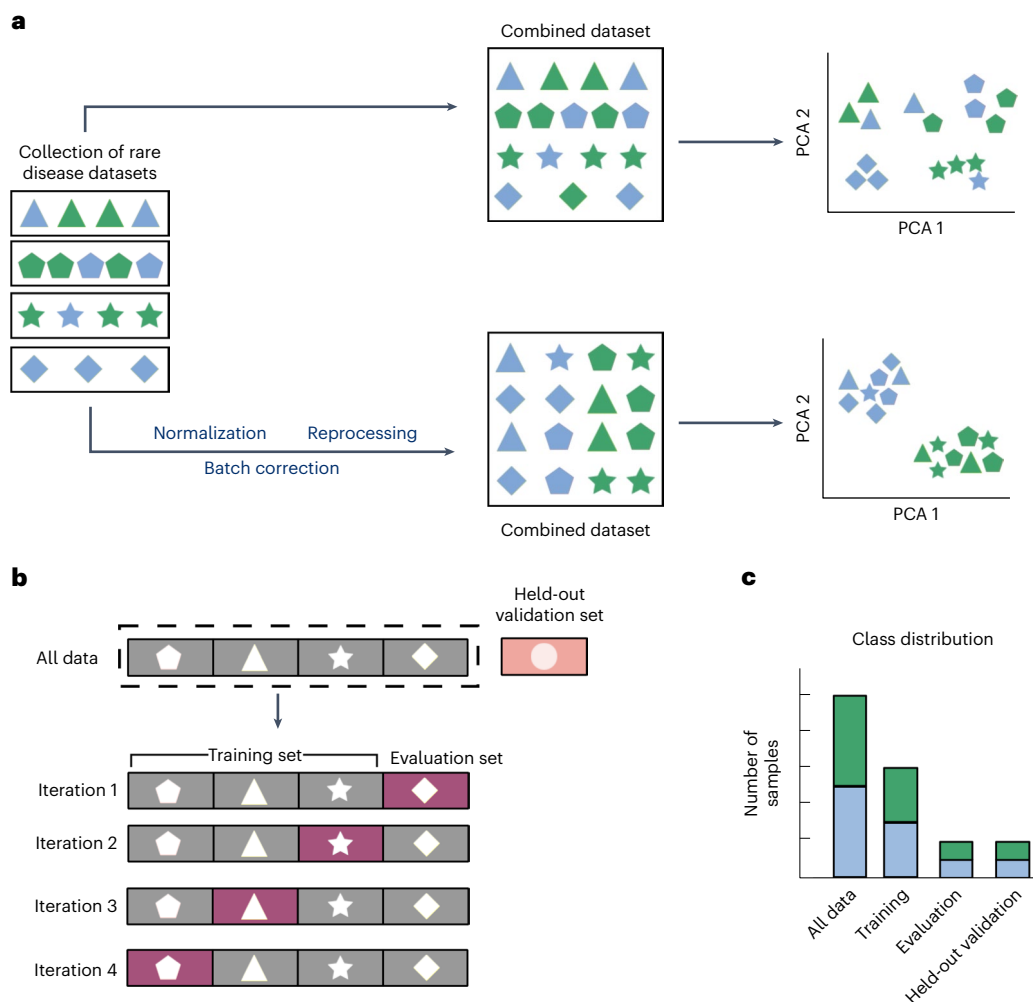


Fig. 1 | Combining datasets to increase data for training ML models.

a, Appropriate methods are required to combine smaller datasets into a larger composite dataset: on the left, multiple small, rare disease datasets that need to be combined to form a dataset of higher sample size are shown. The color of the samples suggests classes or groups present in the datasets. The shape represents the origin of dataset. In the middle, methods that may be used to combine the datasets while accounting for dataset-specific technical differences are shown. On the right, PCA of the combined datasets to verify proper integration of samples in the larger dataset is shown. **b**, Composite datasets can be used to make training, evaluation and validation datasets for ML: on the left, the division

of the composite dataset into a training dataset and a held-out validation dataset is shown (top). The held-out validation set is a separate study or cohort that has not been seen by the model. The training set is further divided into training and evaluation datasets for k -fold cross-validation (in this example, $k = 4$), where each fold contains all samples from an individual study. This approach is termed study-wise cross-validation and supports the goal of training models that generalize to unseen cohorts. **c**, Bar plot showing the class distribution of the training, evaluation and held-out validation datasets from **b**. Text in blue font above or below arrows represent overarching methodological concepts. Text in black font above or below arrows represent specific methods used for analyses.

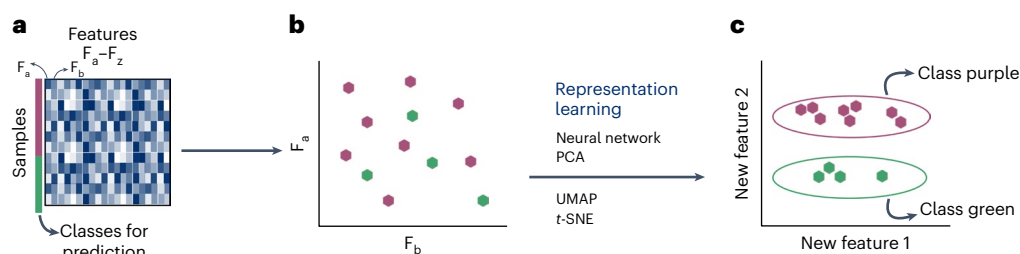


Fig. 2 | Representation learning can extract useful features from high-dimensional data.

a, The data (for example, transcriptomic data) are high dimensional, having thousands of features (displayed as F_1, F_2, \dots, F_z). Samples come from two separate classes (purple and green row annotation). **b**, In the original feature space, F_a and F_b do not separate the two classes (purple and green) well. UMAP, uniform manifold approximation and projection; t-SNE, t -distributed stochastic neighbor embedding. **c**, A representation learning approach learns new features (for example, new feature 1, a combination of F_1, F_2, \dots, F_z , and new

feature 2, a different combination of F_1, F_2, \dots, F_z). New feature 2 distinguishes class, whereas new feature 1 may capture some other variable such as batch (not represented in the figure). New features from the model 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. Text in blue font above or below arrows depict overarching methodological concepts. Text in black font above or below arrows depict specific methods used for analyses.

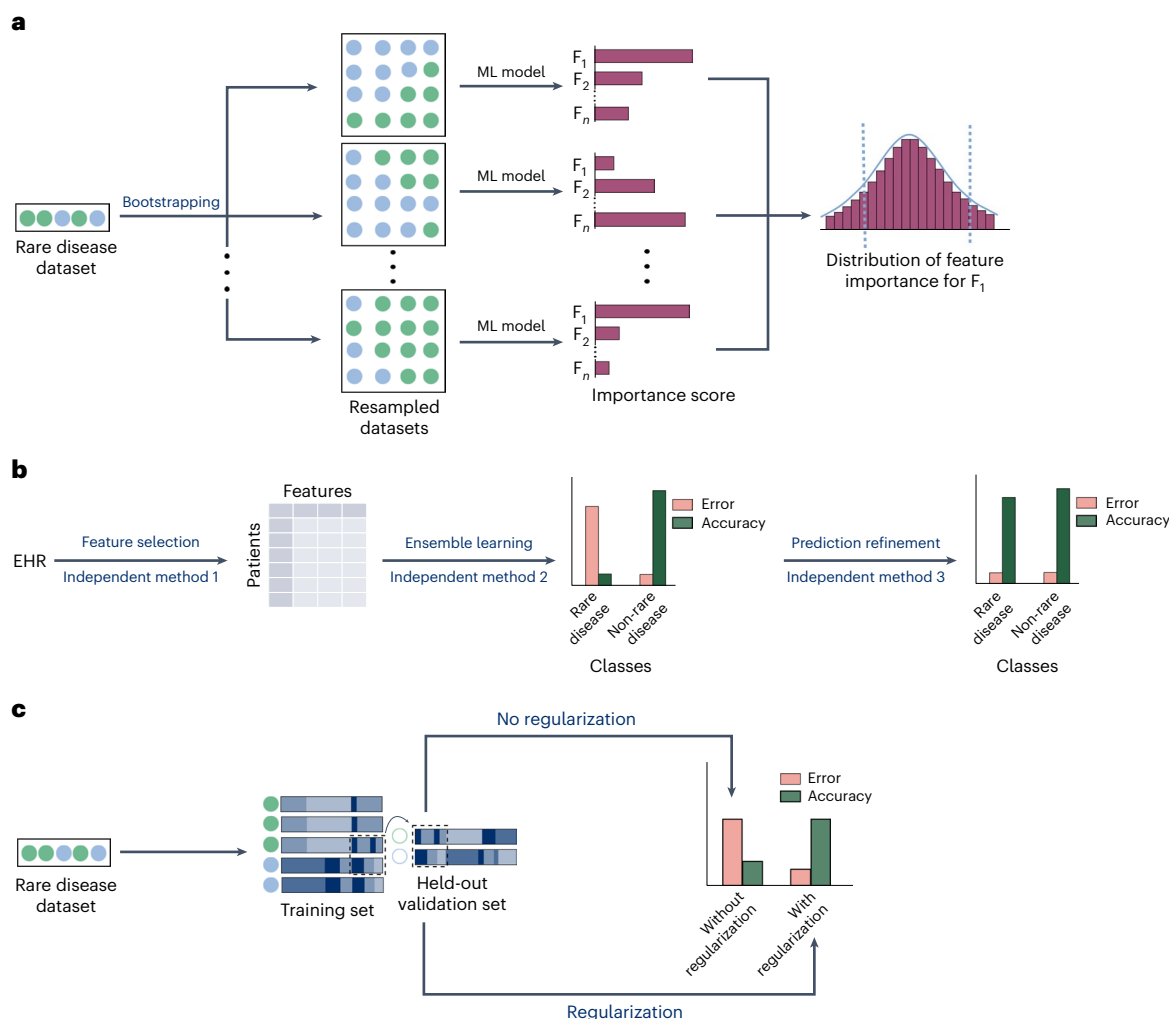


Fig. 3 | Strategies to reduce misinterpretation of ML model output in rare disease. **a**, Bootstrapping: on the left is shown a small rare disease dataset, which can be resampled with replacement using bootstrapping to form a large resampled dataset (middle). Running the same ML model on multiple resampled datasets generates a distribution of values for the importance scores for each feature used by the ML model (right). **b**, Cascade learning: a schematic showing the different steps in a cascade learning approach for identifying patients with rare diseases from electronic health record (EHR) data. The bar plot in the middle schematically represents patient-classification accuracy after ensemble learning. The accuracy is high for non-rare diseases but low for rare diseases. The bar plot on the right depicts classification accuracy after implementation of cascade learning. The accuracy is high for both non-rare and rare diseases. **c**, Regularization: a schematic showing the concept of regularization to

selectively learn relevant features. The samples (green and blue circles) in the rare disease dataset on the left can be represented as a combination of features. Each horizontal bar in the middle (training set) represents a feature-by-sample heatmap for one sample each. In the held-out validation dataset, for a sample of unknown class (open circle), some features recapitulate the pattern present in the training set, while others do not. Right, accuracy of predicting the class of the open circles with or without using regularization during implementation of the ML models on rare disease data. Without regularization, the classification accuracy is low due to the presence of only a subset of learned features (denoted by the dashed rectangle in the middle), but, with regularization, this subset of features is sufficient to gain high classification accuracy. Text in blue font above arrows represent overarching methodological concepts. Text in black font above or below arrows represent specific methods used for analyses.

used to extract features from transcriptomic datasets made of combinations of gene expression values^{21,24,25}, predict rare pathologies from images²⁶ (point 1 of Box 1) or detect cell populations associated with rare diseases²⁷.

When applied to complex systems, representation learning generally requires many samples 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 discuss this method of leveraging large datasets to reduce dimensionality in smaller datasets, also known as feature-representation-transfer learning. Once the dimensions of the training dataset have been reduced, model training can proceed using the experimental design as outlined in Box 3.

Reducing misinterpretation of model output with statistical techniques

The successful application of ML can be improved by meeting certain conditions. First, the dataset contains sufficient representation from each class such that relevant variability from that class is captured. Second, the dataset is complete; all samples have measurements for all variables in the dataset (that is, the dataset is not 'sparse'; it is not missing data for some of the samples). Third, there is no ambiguity about the labels for the samples in the dataset (that is, no 'label noise').

Rare disease datasets violate many of these assumptions. Small numbers of samples for specific classes fail to fully capture the sample variability in those classes. For example, only a few patients with a particular rare disease in a health record dataset require special consideration for evaluation (Boxes 2 and 3). The data are also often sparse, and

there may be abundant label noise due to incomplete understanding of the disease. All these factors contribute to a low signal-to-noise ratio in rare disease datasets. Applying ML to such data without addressing these shortcomings may lead to models that have poor generalizability or are hard to interpret.

Class imbalance or insufficient representation of a class in datasets can be addressed using decision tree-based ensemble learning methods (for example, random forests)²⁸ (Fig. 3a). Random forests use sampling with replacement-based techniques to form a consensus about the important predictive features identified by the decision trees (for example, point 3 of Box 1)^{29,30}. Additional approaches such as combining random forests with sampling without replacement can generate confidence intervals for the model predictions (for applications such as point 4 of Box 1) by mimicking real-world cases in which most rare disease datasets are incomplete³¹. Resampling approaches are most helpful in constructing confidence intervals for algorithms that generate the same outcome every time they are run (that is, deterministic models). For decision trees that choose features at random for selecting a path to the outcome (that is, non-deterministic ones), resampling approaches can be helpful in estimating the reproducibility of the model.

When decision tree-based ensemble methods fail for rare disease datasets, cascade learning is a viable alternative (Fig. 3b)³². In cascade learning, multiple methods leveraging distinct statistical techniques are used to identify stable patterns in the dataset^{33,34}. For example, a cascade learning approach for identifying patients with rare diseases from electronic health record data (point 1 in Box 1) incorporated independent steps for feature extraction (word2vec³⁵), preliminary prediction with ensembled decision trees and then prediction refinement using data similarity metrics³². Combining these three methods resulted in better overall prediction when implemented on a silver-standard dataset, as compared to a model that used ensemble-based prediction alone. In addition to cascade learning, approaches that better represent rare classes using class re-balancing techniques such as inverse sampling probability weighting³⁶, inverse class frequency weighting³⁷, oversampling of rare classes³⁸ or uniformly random undersampling of the majority class³⁹ may also help minimize issues associated with class imbalance.

The presence of label noise and sparsity in the data can lead to poor generalizability or overfitting, meaning that the models show high prediction accuracy on the training data but low prediction accuracy on new evaluation data. Overfit models tend to rely on patterns that are unique to the training data, such as the clinical coding practices at a hospital, and not generalize to new data such as data collected at different hospitals^{40,41}. Regularization approaches can help mitigate these scenarios by adding constraints to a model to avoid making large prediction errors. This protects ML models from poor generalizability by reducing model complexity and minimizing model feature space⁴² (Fig. 3c). Examples of ML methods with regularization include ridge regression, lasso regression and elastic net regression⁴³, among others. Lasso regularization helped select a few informative genes as features to include in models classifying patients with amyotrophic lateral sclerosis and healthy patients with high accuracy based on brain tissue gene expression, thus making the models more interpretable⁴⁴. In the context of rare immune cell signature discovery, in which a few genes or features are expected to distinguish between immune cell types, elastic net regression was able to exclude groups of uninformative genes by reducing their contribution to zero⁴⁵. In a study using a variational autoencoder (VAE; Box 4) for dimensionality reduction in gene expression data from acute myeloid leukemia (AML) samples, the Kullback–Leibler loss between the input data and its low-dimensional representation provided the regularizing penalty for the model^{46,47}. A study using a convolutional neural network to identify tubers in magnetic resonance images from patients with tuberous sclerosis (an application that can facilitate

BOX 4

Definitions

Knowledge graph

A knowledge graph (KG) is a network representation of human knowledge about a domain, abstracted into nodes and edges. Any entity of interest (for example, a gene, a disease, a protein or a cell line) can be represented as a node in a KG. All nodes can be linked through edges that represent known relationships between the nodes. Edges can be directed, indicating that the order of the nodes is important for encoding the relationship, or undirected. For example, a gene (node) can be linked to a protein (node) using a directed edge that represents the relationship that the protein is generated through transcription and translation of the gene. KGs serve to integrate data that exist in distributed sources, encode human readable knowledge in machine-readable format and evolve in a flexible manner to integrate new knowledge as it becomes available.

Machine learning

Machine learning (ML) is a scientific discipline at the intersection of computer science and statistics, which combines computational and statistical methods to identify patterns in sample data⁴⁸. In this discipline, one intends to use data as input and apply or fit predictive models to recognize patterns in the data or identify informative groups among the data using objective computational methods.

Rare disease

According to the Orphan Drug Act⁴ of the USA, diseases or conditions that impact less than 200,000 people in the USA are considered to be rare diseases. The European Union defines a disease as rare when it affects less than 1 in 2,000 people.

Regularization

Regularization is an approach to reduce overfitting of models to training data, in which a penalty or constraint is added to a model trained on a training dataset to avoid making large prediction errors on the evaluation dataset.

Transfer learning

Transfer learning is an approach in which a model trained for one task or domain (source domain) is applied to another, typically related task or domain (target domain), for example, a model pretrained with natural images from the ImageNet dataset can potentially be used to classify medical images⁴⁹. Transfer learning can be supervised (one or both of the source and target domains have labels) or unsupervised (both domains are unlabeled).

Variational autoencoder

Variational autoencoders (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 because the distribution of the encodings are regularized such that they are close to a normal distribution, which may contribute to learning more biologically relevant signals²¹.

point 1 in Box 1) minimized overfitting using the dropout regularization method, which removed randomly chosen network nodes in each iteration of the convolutional neural network model, generating

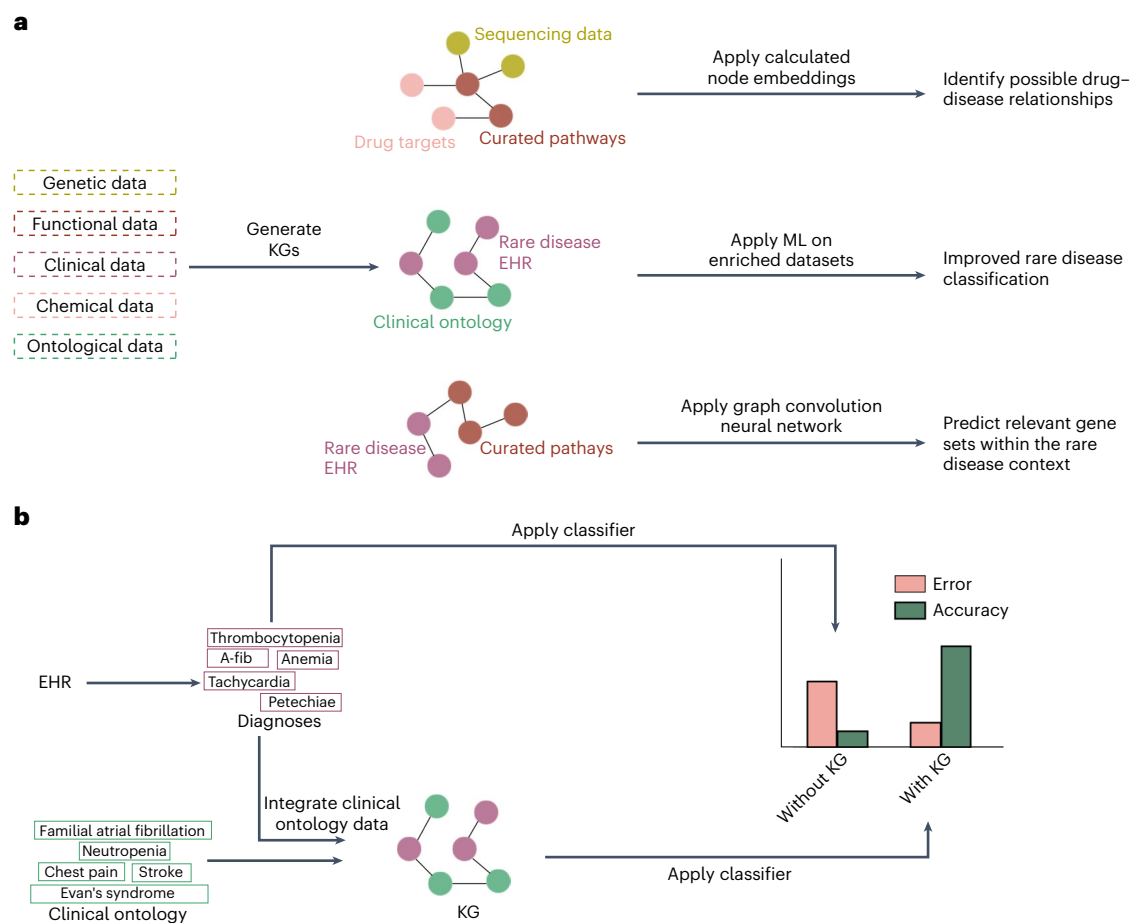


Fig. 4 | Application of KGs can improve ML in rare disease. **a**, KGs integrate different data types (for example, genetic, functional, clinical, chemical and ontological data) and may allow models to learn from connections that are specific to rare diseases or happen in many biomedical contexts. There are a variety of possible applications of this approach, including identifying new disease–drug relationships⁵⁷, augmenting data to improve accuracy of models trained on the data⁵⁸ or mining prior knowledge to discover important gene sets and pathways in rare diseases⁵⁹. **b**, KGs can also be used to augment data. Li et al.⁵⁶

applied a classifier to an electronic health record corpus to identify patients with rare diseases. They trained a classifier on the electronic health record data alone (for example, thrombocytopenia, anemia, atrial fibrillation (A-fib)) and trained another classifier on data augmented with medically related concepts from a KG (for example, neutropenia, stroke). The classifier trained on KG-augmented data has lower error and higher accuracy (right). Text in black font above or below arrows represent specific methods used for analyses.

simpler models in each iteration⁴⁸. Thus, in our opinion, depending on the learning method used, regularization approaches should be considered when working with rare disease datasets.

Building upon prior knowledge and indirectly related data

One strategy to overcome the paucity of data in rare disease is to combine a variety of data types and explore rare disease data alongside other existing knowledge. By using several data modalities, such as curated pathways, genetic data or drug–target relationships, it may be possible to gain a better understanding of rare diseases. Knowledge graphs (KGs), which integrate related but different data types, provide a rich multimodal data source, for example, Monarch Graph Database⁴⁹, Hetionet⁵⁰, PheKnowLator⁵¹, the global network of biomedical relationships⁵² and Orphanet⁵³. These graphs connect genetic, functional, chemical, clinical and ontological data so that relationships of data with disease phenotypes can be explored through manual review⁵⁴ or computational methods^{55,56} (Fig. 4a). KGs may include links (that is, edges) or nodes that are specific to a rare disease of interest (for example, a Food and Drug Administration-approved treatment would be a specific disease–compound edge in the KG) and more generalized information (for example, gene–gene interactions noted in the literature for a different disease) (Fig. 4a).

Rare disease researchers can repurpose general biological or chemical KGs that are not disease specific to answer rare disease-based research questions⁵⁷ (for example, point 2 in Box 1). One tactic to sift through the large amounts of data encoded in KGs is to calculate the distances between nodes of interest (for example, diseases and drugs for point 2 in Box 1 (ref. 57)), often done by determining the ‘embeddings’ (lower-dimension vector representations of the position and connections of a particular point in the graph for nodes in the KG) and calculating the similarity between these embeddings. Effective methods to calculate node embeddings that can generate actionable insights for rare diseases are an active area of research⁵⁷.

Another application of KGs is to augment a dataset⁵⁸. Li et al.⁵⁶ used a KG to identify linked terms in a medical corpus from a large number of patients, some with rare disease diagnoses. They augmented their text dataset by mapping related clinical terms together, for example, mapping ‘cancer’ and ‘malignancy’ in different patients to the same clinical concept. With this enhanced dataset, they trained and tested a variety of text-classification algorithms to identify patients with rare diseases within their corpus (Fig. 4b and point 1 in Box 1).

Rare disease researchers have also integrated multiple KGs and applied neural network-based algorithms optimized for graph data, such as a graph convolutional neural network. Rao and colleagues⁵⁹ describe the construction of a KG using phenotype information

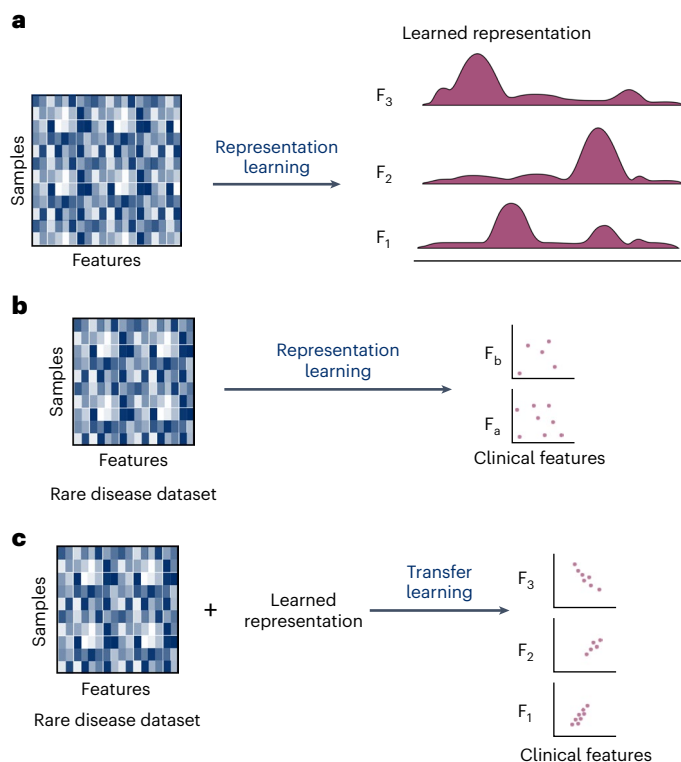


Fig. 5 | Feature-representation-transfer approaches learn representations from a source domain and apply them to a target domain. a, Combination of features representing samples of a large dataset (transcriptomic data from tumors) are learned by an ML model through representation learning. **b**, When applied to a small cell line dataset, the representations extracted by an ML model tend to be incomplete and correlate poorly with clinical or drug-sensitivity features. **c**, When a representation learning model trained on the large dataset (**a**) is applied to the small cell line dataset to extract consistent combinations of features based on the combinations found in the larger training dataset, the extracted representations correlate strongly with the clinical or drug-sensitivity features. Text in blue font above arrows represent overarching methodological concepts.

from the Human Phenotype Ontology and rare disease information from Orphanet and curated gene interaction and pathway data from Lit-BM-13 and WikiPathways^{60–62}. They trained a spectral graph convolution neural network on this KG to identify and rank potentially causal genes for the rare diseases from Orphanet and used this model to accurately predict causal genes for a ground truth dataset of rare diseases with known causal genes. While several groups have used KGs to study rare diseases, we expect that better multimodal datasets and ML methods to analyze KGs will make them a more popular and important tool in rare disease.

Another approach that builds on prior knowledge and large volumes of related data is transfer learning, a modeling technique that ‘borrows strength’ across datasets with both similar and distinct properties, such as an imaging anomaly present in both rare and common diseases, to advance our understanding of rare diseases. Transfer learning, in which 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 domain⁶³ (Fig. 5a–c). That is, representation learning, as discussed earlier, does not need to be applied only to describe the dataset on which the algorithm was trained; it can also be used to elucidate signals in sufficiently similar data (Fig. 5c) and may offer an improvement in descriptive capability over models trained only on small rare disease datasets (Fig. 5c).

For instance, low-dimensional representations can be learned from tumor transcriptomic data and transferred to describe patterns in genetic alterations in cell lines²¹ (Fig. 5c). In the next section, we summarize specific instances of applying transfer learning, along with other techniques, to the study of rare diseases.

Successful applications often combine multiple ML techniques

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 examples in the rare disease domain that use concepts of feature-representation transfer, use of prior data and regularization.

Our first example includes a large dataset of AML patient samples with no drug-response data and a small in vitro experiment with drug-response data⁶⁴. 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 VAE (Box 4) on a reasonably large dataset of AML patient samples from 96 independent studies to learn meaningful representations in an approach termed DeepProfile⁴⁶ (Fig. 6a). The representations or encodings learned by the VAE were then transferred to the small in vitro dataset, reducing its number of features from thousands to eight and improving the performance of the final lasso linear regression model (point 2 of Box 1). In addition to improving performance, the encodings learned by the VAE captured more biological pathways than PCA, possibly due to the constraints on the encodings imposed during training (Box 4). Similar results were observed for prediction of histopathology in another rare cancer dataset⁴⁶.

While DeepProfile was centered on training on an individual disease-and-tissue combination, some rare diseases affect multiple tissues that a researcher may wish to study (for example, point 4 in Box 1). Studying multiple tissues poses substantial challenges, and a cross-tissue analysis may require comparing 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)⁶⁵ on a large generic collection of human transcriptomic data⁶⁶. PLIER used constraints (regularization) that learned latent variables aligned with a small number of input gene sets, making it suitable for rare disease data. The authors transferred the representations or latent variables learned by the model to describe transcriptomic data from the unseen rare diseases anti-neutrophil cytoplasmic antibody-associated vasculitis and medulloblastoma in an approach termed MultiPLIER⁶⁶ (Fig. 6b). MultiPLIER used one model to describe multiple datasets instead of reconciling output from multiple models, making it possible to identify commonalities among disease manifestations or affected tissues.

DeepProfile⁴⁶ and MultiPLIER⁶⁶ exemplify modeling approaches incorporating prior knowledge, thereby constraining the model space according to plausible or expected biology, or sharing information across datasets. These two methods capitalize on similar biological processes observed across different biological contexts and the fact that the methods underlying the approaches can effectively learn about those processes.

Outlook

This Perspective highlights challenges in applying ML to rare disease data and approaches that address these challenges. Small sample size, while important, is not the only roadblock. The high dimensionality of modern data requires creative approaches, such as learning new representations of the data, to manage the curse of dimensionality. In our opinion, leveraging prior knowledge and transfer learning methods

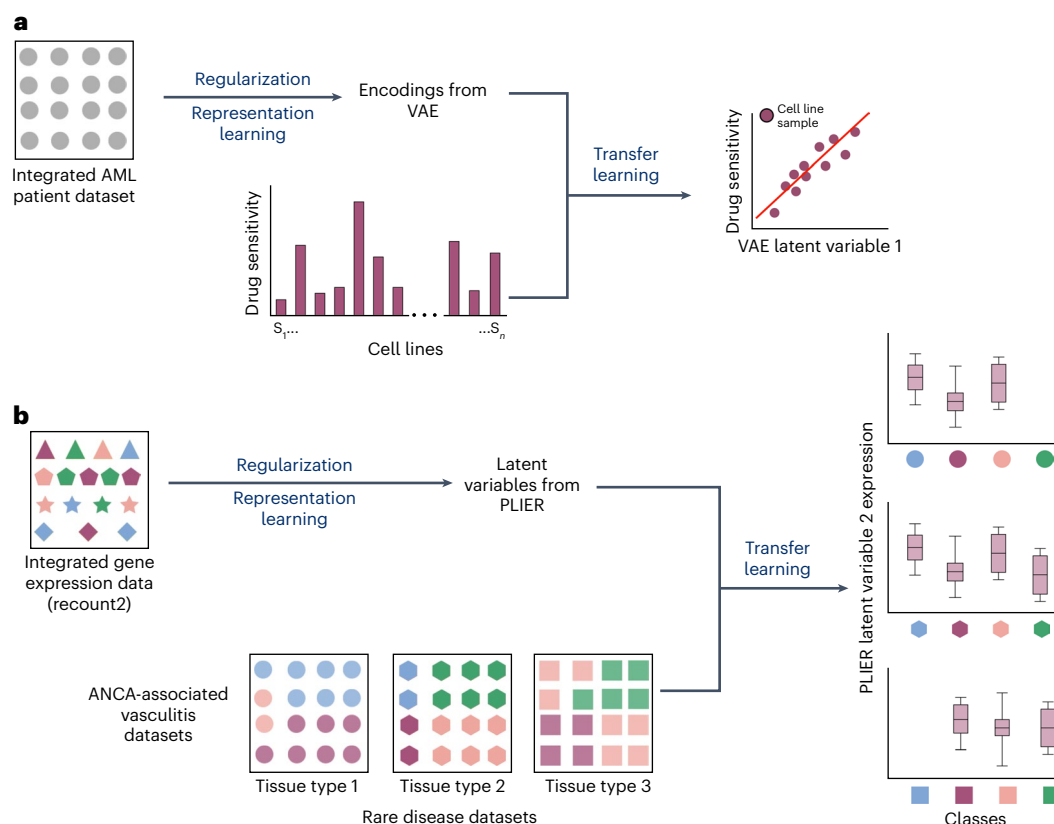


Fig. 6 | Combining multiple strategies strengthens the performance of ML models in rare disease. **a**, The authors of DeepProfile trained a VAE to learn a representation from AML data without phenotype labels, transferred those representations to a small dataset with phenotype labels and found that it improved prediction performance in a drug-sensitivity prediction task⁴⁶. **b**, The authors of MultiPLIER trained a PLIER model on a large, heterogeneous collection

of expression data (recount2 (ref. 82)) and transferred the representations (termed latent variables) to multiple datasets from rare diseases that were not in the training set⁶⁵. Expression of PLIER latent variables can be used to check for concordance between datasets, among other applications. ANCA, anti-neutrophil cytoplasmic antibody. Text in blue font above or below arrows represent overarching methodological concepts.

to appropriately interpret data is also required. Furthermore, we posit that researchers applying ML methods on rare disease data should use techniques that increase confidence (that is, bootstrapping) and penalize complexity of the resultant models (that is, regularization) to enhance the generalizability of their work. It should be noted that the line between classical statistical methods and ML is fuzzy. Multiple statistical techniques that were considered to be out of the scope of this article (for example, hierarchical models, Bayesian frameworks, association tests)^{67–70} may have substantial potential to enhance the accuracy and generalizability of models and should be considered in the rare disease study-design process.

The approaches highlighted in this Perspective come with challenges 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 to improve data generation and method development going forward. The following two areas are particularly important for the field to explore.

Datasets should be designed with ML-based research in mind

While many techniques exist to collate rare data from different sources, low-quality data may hurt the end goal even if it increases 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: conversations with a clinician revealed that samples in a particular tumor dataset were collected using vastly different surgical techniques (laser ablation and excision

versus standard excision). This information, not readily available to non-experts, was obvious to the clinician. Such instances suggest that collaboration with domain experts and sharing of well-annotated data are needed to generate robust datasets.

In addition to sample scarcity, comprehensive phenotypic-genotypic databases are also lacking. While rare disease studies that collect genomic and phenotypic data are becoming more common^{71–73}, we believe that developing comprehensive genomics-based genotype-phenotype databases that prioritize clinical and genomic data standards is key to fueling interpretation of ML methods. This method can be bolstered by funding or fostering collaboration between biobanking projects and patient registry initiatives. In our opinion, mindful sharing of data with proper metadata and attribution enabling prompt data reuse is important in building valuable datasets for rare disease research⁷⁴. Finally, federated learning methods, such as those used in mobile health⁷⁵ and electronic healthcare record studies⁷⁶, may allow researchers to develop ML models on data from larger numbers of people with rare diseases while protecting patient privacy.

Methods to probe mechanisms of rare diseases are needed

Most ML methods for rare diseases are used for classification tasks. Not many methods investigate biological mechanisms; this is likely due to a lack of methods that can handle the limitations of rare disease data described throughout this Perspective. We believe that developing methods to address this will be critical for applying ML to rare disease data.

For example, development of methods with a focus on explainability of the model can identify features that may be related to the underlying disease mechanism⁷⁷. Alternatively, representation learning or regularization methods may help identify multiple correlated features, which can be interrogated to identify biologically meaningful pathways. Additionally, robust error analysis for newly developed models to help users understand how a feature influences the performance of a model can provide insight into its potential contribution to the underlying disease mechanism⁷⁸. Interrogating disease mechanisms by adopting modifications of these approaches is necessary as ML applications become mainstream in research and clinical settings.

Finally, in our view, methods that can reliably integrate disparate datasets will likely always remain a need in rare disease research. Methods that rely on finding structural correspondence between datasets ('anchors') may be able to transform the status quo of using ML in rare disease^{79–81}. We speculate that this is an important and burgeoning area of research, and we are optimistic about the future of applying ML approaches to rare diseases.

References

- Schaefer, J., Lehne, M., Schepers, J., Prasser, F. & Thun, S. The use of machine learning in rare diseases: a scoping review. *Orphanet J. Rare Dis.* **15**, 145 (2020).
- Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A. & Sangiorgi, L. Opportunities and challenges for machine learning in rare diseases. *Front. Med.* **8**, 747612 (2021).
- Li, A. et al. Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. *Cancer Res.* **69**, 2091–2099 (2009).
- Senate and House of Representatives of the United States of America in Congress. *Orphan Drug Act* (1983).
- Agarwal, V. et al. Learning statistical models of phenotypes using noisy labeled training data. *J. Am. Med. Inform. Assoc.* **23**, 1166–1173 (2016).
- Frénay, B. & Verleysen, M. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Netw. Learn. Syst.* **25**, 845–869 (2014).
- Toh, T. S., Dondelinger, F. & Wang, D. Looking beyond the hype: applied AI and machine learning in translational medicine. *EBioMedicine* **47**, 607–615 (2019).
- Clarke, R. et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer* **8**, 37–49 (2008).
- Altman, N. & Krzywinski, M. The curse(s) of dimensionality. *Nat. Methods* **15**, 399–400 (2018).
- Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
- Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
- Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
- Kobak, D. & Berens, P. The art of using t-SNE for single-cell transcriptomics. *Nat. Commun.* **10**, 5416 (2019).
- Dorrity, M. W., Saunders, L. M., Queitsch, C., Fields, S. & Trapnell, C. Dimensionality reduction by UMAP to visualize physical and genetic interactions. *Nat. Commun.* **11**, 1537 (2020).
- Chellappa, R. & Turaga, P. Feature selection. In *Computer Vision: a Reference Guide* 1–5 (Springer International, 2020).
- Chen, C.-H., Härdle, W. & Unwin, A. *Handbook of Data Visualization* (Springer, 2008).
- Jolliffe, I. T. & Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* **374**, 20150202 (2016).
- McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1802.03426> (2018).
- Nguyen, L. H. & Holmes, S. Ten quick tips for effective dimensionality reduction. *PLoS Comput. Biol.* **15**, e1006907 (2019).
- Wattenberg, M., Viégas, F. & Johnson, I. How to use t-SNE effectively. *Distill* **1**, <https://doi.org/10.23915/distill.00002> (2016).
- Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S. & Greene, C. S. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol.* **21**, 109 (2020).
- de Souto, M. C. P., Costa, I. G., de Araujo, D. S. A., Ludermit, T. B. & Schliep, A. Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics* **9**, 497 (2008).
- Kothari, S. et al. Removing batch effects from histopathological images for enhanced cancer diagnosis. *IEEE J. Biomed. Health Inform.* **18**, 765–772 (2014).
- Dwivedi, S. K., Tjärnberg, A., Tegnér, J. & Gustafsson, M. Deriving disease modules from the compressed transcriptional space embedded in a deep autoencoder. *Nat. Commun.* **11**, 856 (2020).
- Fertig, E. J., Ding, J., Favorov, A. V., Parmigiani, G. & Ochs, M. F. CoGAPS: an R/C++ package to identify patterns and biological process activity in transcriptomic data. *Bioinformatics* **26**, 2792–2793 (2010).
- Quelleg, G., Lamard, M., Conze, P.-H., Massin, P. & Cochener, B. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Med. Image Anal.* **61**, 101660 (2020).
- Arvaniti, E. & Claassen, M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat. Commun.* **8**, 14825 (2017).
- Chaabane, I., Guermazi, R. & Hammami, M. Enhancing techniques for learning decision trees from imbalanced data. *Adv. Data Anal. Classif.* **14**, 677–745 (2020).
- Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
- Köpcke, F. et al. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Med. Inform. Decis. Mak.* **13**, 134 (2013).
- Banerjee, J. et al. Integrative analysis identifies candidate tumor microenvironment and intracellular signaling pathways that define tumor heterogeneity in NF1. *Genes* **11**, 226 (2020).
- Colbaugh, R., Glass, K., Rudolf, C., & Tremblay, M. Learning to identify rare disease patients from electronic health records. *AMIA Annu. Symp. Proc.* **2018**, 340–347 (2018).
- Heiselet, B., Serre, T., Pontil, M. & Poggio, T. Component-based face detection. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition I (CPRV, 2001)*.
- Kasinski, A. & Schmidt, A. The architecture of the face and eyes detection system based on cascade classifiers. In *Computer Recognition Systems 2* (ed. Kurzynski, M. et al.) 124–131 (Springer, 2007).
- Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1301.3781> (2013).
- Han, S., Williamson, B. D. & Fong, Y. Improving random forest predictions in small datasets from two-phase sampling designs. *BMC Med. Inform. Decis. Mak.* **21**, 322 (2021).
- Ambert, K. H. & Cohen, A. M. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *J. Am. Med. Inform. Assoc.* **16**, 590–595 (2009).
- Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

39. More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1608.06048> (2016).
40. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT, 2016).
41. Futoma, J., Simons, M., Doshi-Velez, F. & Kamaleswaran, R. Generalization in clinical prediction models: the blessing and curse of measurement indicator variables. *Crit. Care Explor.* **3**, e0453 (2021).
42. Okser, S. et al. Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet.* **10**, e1004754 (2014).
43. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B Stat. Methodol.* **67**, 301–320 (2005).
44. Founta, K. et al. Gene targeting in amyotrophic lateral sclerosis using causality-based feature selection and machine learning. *Mol. Med.* **29**, 12 (2023).
45. Torang, A., Gupta, P. & Klinke, D. J. 2nd An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets. *BMC Bioinformatics* **20**, 433 (2019).
46. Dincer, A. B., Celik, S., Hiranuma, N. & Lee, S.-I. DeepProfile: deep learning of cancer molecular profiles for precision medicine. Preprint at *bioRxiv* <https://doi.org/10.1101/278739> (2018).
47. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.1312.6114> (2013).
48. Sánchez Fernández, I. et al. Deep learning in rare disease. Detection of tubers in tuberous sclerosis complex. *PLoS ONE* **15**, e0232376 (2020).
49. Mungall, C. J. et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.* **45**, D712–D722 (2017).
50. Himmelstein, D. S. et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **6**, e26726 (2017).
51. Callahan, T. J., Tripodi, I. J., Hunter, L. E. & Baumgartner, W. A. A framework for automated construction of heterogeneous large-scale biomedical knowledge graphs. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.04.30.071407> (2020).
52. Percha, B. & Altman, R. B. A global network of biomedical relationships derived from text. *Bioinformatics* **34**, 2614–2624 (2018).
53. Orphanet <https://www.orpha.net/consor/cgi-bin/index.php> (2023).
54. Queralt-Rosinach, N. et al. Structured reviews for data and knowledge-driven research. *Database* **2020**, baaa015 (2020).
55. Moon, C. et al. Learning drug–disease–target embedding (DDTE) from knowledge graphs to inform drug repurposing hypotheses. *J. Biomed. Inform.* **119**, 103838 (2021).
56. Li, X. et al. Improving rare disease classification using imperfect knowledge graph. *BMC Med. Inform. Decis. Mak.* **19**, 238 (2019).
57. Sosa, D. N. et al. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. In *Biocomputing 2020* 463–474 (World Scientific, 2019).
58. Shen, F. et al. Rare disease knowledge enrichment through a data-driven approach. *BMC Med. Inform. Decis. Mak.* **19**, 32 (2019).
59. Rao, A. et al. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Med. Genomics* **11**, 57 (2018).
60. Köhler, S. et al. The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
61. Rolland, T. et al. A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
62. Martens, M. et al. WikiPathways: connecting communities. *Nucleic Acids Res.* **49**, D613–D621 (2021).
63. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
64. Lee, S.-I. et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.* **9**, 42 (2018).
65. Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **16**, 607–610 (2019).
66. Taroni, J. N. et al. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst.* **8**, 380–394 (2019).
67. Greene, D., NIHR BioResource, Richardson, S. & Turro, E. Phenotype similarity regression for identifying the genetic determinants of rare diseases. *Am. J. Hum. Genet.* **98**, 490–499 (2016).
68. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
69. Ionita-Laza, I., Capanu, M., De Rubeis, S., McCallum, K. & Buxbaum, J. D. Identification of rare causal variants in sequence-based studies: methods and applications to *VPS13B*, a gene involved in Cohen syndrome and autism. *PLoS Genet.* **10**, e1004729 (2014).
70. Greene, D., NIHR BioResource, Richardson, S. & Turro, E. A fast association test for identifying pathogenic variants involved in rare diseases. *Am. J. Hum. Genet.* **101**, 104–114 (2017).
71. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat. Rev. Genet.* **14**, 681–691 (2013).
72. Wright, C. F., FitzPatrick, D. R. & Firth, H. V. Paediatric genomics: diagnosing rare disease in children. *Nat. Rev. Genet.* **19**, 253–268 (2018).
73. Adams, D. R. & Eng, C. M. Next-generation sequencing to diagnose suspected genetic disorders. *N. Engl. J. Med.* **379**, 1353–1362 (2018).
74. Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X. & Greene, C. S. Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.* **21**, 615–629 (2020).
75. Rieke, N. et al. The future of digital health with federated learning. *NPJ Digit. Med.* **3**, 119 (2020).
76. Yan, Y. et al. A continuously benchmarked and crowdsourced challenge for rapid development and evaluation of models to predict COVID-19 diagnosis and hospitalization. *JAMA Netw. Open* **4**, e2124946 (2021).
77. Lundberg, S. M. et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* **2**, 749–760 (2018).
78. Zhou, G., Zhang, J., Su, J., Shen, D. & Tan, C. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics* **20**, 1178–1190 (2004).
79. Blitzer, J., McDonald, R. & Pereira, F. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (eds. Jurafsky, D. & Gaussier, E.) 120–128 (Association for Computational Linguistics, 2006).
80. Wang, C. & Mahadevan, S. Heterogeneous domain adaptation using manifold alignment. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence* **2** (ed. Walsh, T.) 1541–1546 (AAAI, 2011).
81. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
82. Collado-Torres, L. et al. Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).
83. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer, 2013).

84. Davis, J. & Goadrich, M. The relationship between precision–recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (eds. Cohen, W. W. & Moore, A.) 233–240 (Association for Computing Machinery, 2006).
85. Hastie, T., Friedman, J. & Tibshirani, R. *The Elements of Statistical Learning* (Springer, 2001).
86. Shin, H.-C. et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).

Acknowledgements

This work was supported in part by Alex’s Lemonade Stand Foundation, the National Human Genome Research Institute (R01 HG010067), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01 HD109765), the Neurofibromatosis Therapeutic Acceleration Program, the Children’s Tumor Foundation and the Gilbert Family Foundation.

Author contributions

Authorship was determined using ICMJE recommendations. Conceptualization: J.B., J.N.T., R.J.A., C.G., J.G.; data curation: not applicable; formal analysis: not applicable; funding acquisition: R.J.A.; investigation: J.B., J.N.T., R.J.A.; methodology: J.B., J.N.T., R.J.A.; project administration: J.B.; resources: J.B., J.N.T., R.J.A.; software: not applicable; supervision: J.B., C.G.; validation: not applicable; visualization: D.V.P.; writing (original draft): J.B., J.N.T., R.J.A.; writing (review and editing): J.B., J.N.T., R.J.A., C.G., J.G.

Competing interests

J.G. is currently employed at Tempus Labs, a precision medicine company. J.N.T. and D.V.P. are employed with Alex’s Lemonade Stand Foundation, a research funder. The remaining authors declare no competing interests.

Additional information

Correspondence should be addressed to Casey Greene.

Peer review information *Nature Methods* thanks Pejman Mohammadi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lin Tang, in collaboration with the *Nature Methods* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2023