A guide to machine learning for biologists

*Joe G. Greener**
Dept of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
j.greener@ucl.ac.uk

*Shaun M. Kandathil**
Dept of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
s.kandathil@ucl.ac.uk

Lewis Moffat
Dept of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
lewis.moffat@cs.ucl.ac.uk

David T. Jones†
Dept of Computer Science, University College London, Gower Street, London WC1E 6BT, United Kingdom
d.t.jones@ucl.ac.uk

* These authors contributed equally
† Correspondence should be addressed to: d.t.jones@ucl.ac.uk

The expanding scale and inherent complexity of biological data have encouraged a growing use of machine learning in biology to build informative and predictive models of the underlying biological processes. All machine learning techniques fit models to data, however, the specific methodologies are quite varied and can at first glance seem bewildering. In this article, we aim to provide readers with a gentle introduction to a few key machine learning techniques, including those based on the most recently developed and widely used techniques involving deep neural networks. We describe how different techniques may be suited to specific types of biological data, and also discuss some best practices and points to consider when embarking on experiments involving machine learning. Some emerging directions in machine learning methodology will also be discussed.

Introduction

Humans make sense of the world around them by observing it, and learning to predict what might happen next. Consider a child learning to catch a ball: the child (usually) knows nothing about the physical laws that govern the motion of a thrown ball, yet, by a process of observation, trial and error, the child adjusts their understanding of the ball's motion, and how to move their body, until they are able to catch it reliably. In other words, the child has learned how to catch the ball by building a sufficiently accurate and useful 'model' of the process, by repeatedly testing this model against the data, and making corrections to the model to make it better.

Machine learning refers broadly to the process of fitting predictive models to data or of identifying informative groupings within data. The field of machine learning essentially attempts to approximate or imitate humans' ability to recognise patterns, albeit in an objective manner, using computation. Machine learning is particularly useful when the dataset one wishes to analyse is too large (many individual data points), or too complex (contains a large number of features) for human analysis and/or when it is desired to automate the process of data analysis to establish a reproducible and time-efficient pipeline. Data from biological experiments frequently possess these properties; biological data sets have grown enormously in both size and complexity in the last few decades, and it is becoming increasingly important not only to have some practical means of making sense of this data abundance, but also to have a sound understanding of the techniques that are used. Machine learning has been used in biology for a number of decades, but it has steadily grown in importance to the point where it is used in nearly every field of biology. However, only in the last few years has the field taken a more critical look on the available strategies and begun to assess which methods are most appropriate in different scenarios, or even if they are appropriate at all.

This Review aims to inform biologists on how they can start to understand and use machine learning techniques. We do not intend to present a thorough literature review of papers using machine learning for biological problems [1], or to describe the detailed mathematics of various machine learning methods [2,3]. Instead, we focus on linking particular techniques to different types of biological data (similar reviews are available for specific biological disciplines e.g. [4–11]). We also attempt to distil some best practices of how to practically go about the process of training and improving a model. The complexity of biological data presents pitfalls as well as opportunities for their analysis using machine learning techniques. To address these, we discuss the widespread issues that affect the validity of studies, with guidance on how to avoid them. The bulk of the article is devoted to the description of a number of machine learning techniques, and in each case, we provide examples of the appropriate application of the method and how to interpret the results. These methods discussed include traditional machine learning methods, as these are still the best choices in many cases, and deep learning with artificial neural networks, which are emerging as the most effective methods for many tasks. We finish by describing what

the future holds for incorporating machine learning in data analysis pipelines in biology.

There are two goals when using machine learning in biology. The first is to make accurate predictions where experimental data is lacking, and use these predictions to guide future research efforts. However, as scientists we seek to understand the world, and so the second goal is to use machine learning to further our understanding of biology. Throughout this guide we will discuss how these two goals often come into conflict in machine learning, and how to extract understanding from models that are often treated as "black boxes" because their inner workings are difficult to understand[12]**.**

Key concepts

We first introduce a number of key concepts in machine learning. Where possible, we illustrate these concepts with examples taken from biological literature.

General terms. A dataset comprises a number of data points or instances, each of which can be thought of as a single observation from an experiment. Each data point is described by a (usually fixed) number of features. Examples of such features might be length, time, concentration, and gene expression level. A machine learning task is an objective specification for what we want a machine learning model to accomplish. For example, for an experiment investigating the expression of genes over time, we might want to predict the rate of conversion of a specific metabolite into another species. In this case, the features “gene expression level” and “time” could be termed input features or simply inputs for the model, and “conversion rate” would be the desired output of the model, that is the quantity we are interested in predicting. A model can have any number of input and output features. Features can be either continuous (taking continuous numerical values), or categorical (taking only discrete values). Quite often, categorical features are simply binary and are either true (1) or false (0).

Supervised and unsupervised learning. Supervised machine learning refers to the fitting of a model to data (or a subset of data) that has been labelled — where there exists some ground truth property, which is usually experimentally measured or human-assigned. Examples include protein secondary structure prediction [13] and prediction of genome accessibility to genome-regulatory factors [14]. In both cases, the ground truth is derived ultimately from laboratory observations, but often this raw data is preprocessed in some way. In the case of secondary structure, for example, the ground truth data are derived from analysing protein crystal structure data in the Protein Data Bank, and in the latter case, the ground truth comes from data derived from DNA sequencing experiments. By contrast, unsupervised learning methods are able to identify patterns in unlabelled data, without the need for providing the system with the ground truth information in the form of pre-determined labels, such as finding subsets of patients with similar expression level in a gene expression study [15] or predicting mutation effects from gene sequence co-variation [16].

Sometimes the two approaches are combined in semi-supervised learning, where small amounts of labelled data are combined with large amounts of unlabelled data. This can improve performance in cases where labelled data is costly to obtain.

Classification, regression and clustering problems. When a problem involves assigning data points to a set of discrete categories (e.g. "cancerous" or "not cancerous"), the problem is called a classification problem, and any algorithm that performs such classification can be said to be a classifier. By contrast, regression models output a continuous set of values, such as predicting the free energy change of folding after mutating a residue in a protein [17]. Continuous values can be thresholded or otherwise discretized, meaning that it is often possible to reformulate regression problems as classification problems. For example, the free energy change mentioned above can be binned into ranges of values that are favourable or unfavourable for protein stability. Clustering methods are used to predict groupings of similar data points in a dataset, and are usually based on some measure of similarity between data points. They are unsupervised methods that do not require that the examples in a dataset have labels. For example, in a gene expression study, clustering could find subsets of patients with similar gene expression.

Classes and labels. The discrete set of values returned by a classifier can be made to be mutually exclusive, in which case they are called classes. Where these values need not be mutually exclusive, they are termed labels. For example, a residue in a protein structure can be in only one of multiple secondary structure classes, but could simultaneously be assigned the non-exclusive labels of being α-helical and transmembrane. Classes and labels are usually represented by an encoding, for example a one-hot encoding.

Loss or cost functions. The output(s) of a machine learning model are never ideal, and will diverge from the ground truth. The mathematical functions that measure this deviation, or in more general terms, that measure the amount of "disagreement" between the obtained and ideal outputs are referred to as loss or cost functions. In supervised learning settings, the loss function would be a measure of deviation of the output relative to the ground truth output. Examples include mean squared error loss for regression problems and binary cross entropy for classification problems.

Parameters and hyperparameters. Models are essentially mathematical functions that operate on some set of input features, and produce one or more output values or features. In order to be able to learn on training data, models contain adjustable parameters whose values can be changed over the training process to achieve the most optimal performance of the model (see below). In a simple regression model, for example, each feature has a parameter that is multiplied by the feature value and these are added together to make the prediction. Hyperparameters are adjustable values that are not considered part of the model itself in that they are not updated during training, but which still have an impact on the training of the model and its performance. A common example of a hyperparameter is the learning rate, which controls the rate or speed with which the model's

parameters are altered during training.

Training, validation and testing. Before being used to make predictions, models require training, which involves automatically adjusting the parameters of a model to improve its performance. In a supervised learning setting, this involves modifying the parameters so the model performs well on a training dataset, by minimizing the average value of the loss or cost function (described above). Usually, a separate validation dataset is used to monitor but not influence the training process, in order to detect potential overfitting (see next section). In unsupervised settings, a cost function is still minimized, though it does not operate on ground-truth outputs. Once a model is trained it can be tested on data not used for training. See Box 1 for a guide on the overall process of training and how to split the data appropriately between training and testing sets. A flowchart to help the overall process is shown in Figure 1, and some of the concepts in model training are shown in Figure 2.

Overfitting and underfitting. The purpose of fitting a model to training data is to capture the "true" relationship between the variables in the data, such that the model has predictive power on unseen (non-training) data. Models that are either overfitted or underfitted will produce poor predictions on data not in the training set (Figure 2d). An overfitted model will produce excellent results on data in the training set (usually as a result of having too many parameters), but will produce poor results on unseen data. The overfitted model in Figure 2d passes exactly through every training point, and so its prediction error on the training set will be zero. However, it is evident that this model has "memorized" the training data and is unlikely to produce good results on unseen data. By contrast, an underfitted model fails to adequately capture the relationships between the variables in the data. This could be due to an incorrect choice of model type, incomplete or incorrect assumptions about the data, too few parameters in the model, and/or an incomplete training process. The underfitted model depicted in Figure 2d is inadequate for the data it is trying to fit; in this case it is evident that the variables have a nonlinear relationship, which cannot be adequately described with a simple linear model and so a non-linear model would be more appropriate.

Inductive bias and the bias-variance tradeoff. The inductive bias of a model refers to the set of assumptions made in the learning algorithm that leads it to favour a particular solution to a learning problem over others. It can be thought of as the model's preference for a particular type of solution to a learning problem over others. This preference is often programmed into the model using its specific mathematical form, and/or by using a particular loss function. For example, the inductive bias of recurrent neural networks (discussed below) is that there are sequential dependencies in the input data such as the concentration of a metabolite over time. This dependence is explicitly accounted for in the mathematical form of a recurrent neural network. Different inductive biases in different model types make them more suitable and usually more performant for specific types of data. Another important concept is the tradeoff between bias and variance. A model with a high

bias can be said to have stronger constraints on the trained model, whereas a model with low bias makes fewer assumptions about the property being modelled, and can, in theory, model a wide variety of function types. The variance of a model describes how much the trained model changes in response to training it on different training datasets. In general, we desire models with very low bias and low variance, though these objectives are often in conflict as a model with low bias will often learn different signals on different training sets. Controlling the bias–variance tradeoff is key to avoiding overfitting or underfitting.

Traditional machine learning

We will now discuss several key machine learning methods, with an emphasis on their particular strengths and weaknesses. A comparison of different machine learning approaches is shown in Table 1. We begin with a discussion of methods not based on neural networks, sometimes called traditional machine learning. Figure 3 shows some of the methods of traditional machine learning. Various software packages can be used to train such models, including scikit-learn in Python (https://scikit-learn.org/stable) [18], caret in R (https://topepo.github.io/caret) [19] and MLJ in Julia (https://alan-turing-institute.github.io/MLJ.jl/stable) [20].

When developing machine learning methods for use with biological data, traditional machine learning should generally be seen as the first area to explore in finding the most appropriate method for a given task. Deep learning can be a powerful tool, and is undeniably trendy right now. However it is still limited in the application areas in which it excels: when large amounts of data are available (e.g. millions of data points); when each data point has many features; and, when the features are highly structured (the features have clear relationships with one another, such as adjacent pixels in images) [21]. Data like DNA, RNA and protein sequences [22,23] and microscopy images [24,25] are examples of biological data where these requirements can be met and deep learning has been successfully applied. However, the requirement for large amounts of data can make deep learning a poor choice even when the other two requirements are met.

Traditional methods, in comparison to deep learning, are much faster to develop and test on a given problem. Developing the architecture of a deep neural network and then training it can be a time- and computation-expensive task to undertake [26] compared to traditional models like support vector machines (SVMs) and random forests [27]. Although some approaches exist, with deep neural networks it is still non-trivial to estimate feature importance, [28] (that is,. how important each feature is for contributing to the prediction), or the confidence of predictions of the model [1,28,29], both of which are often essential in biological settings. Even if deep learning appears technically feasible for a particular biological prediction task, it is often still prudent to train a traditional method to compare against a neural network-based model, if possible [30].

Traditional methods typically expect that each example in the dataset has the same number of features, so this is not always possible. An obvious biological example of this is when protein/RNA/DNA sequences are being used and each example has a different length. In order to use traditional methods with this data, the data can be altered so it is all the same size using simple techniques like padding and windowing. Padding means taking each example and adding additional values containing zero until it is the same size as the largest example in the data set. By contrast, windowing shortens individual examples to a given size, e.g. only using the first 100 residues of each protein in a dataset of protein sequences with lengths ranging from 100 upwards.

Use of classification and regression models. For regression problems such as those shown in Figure 3a, ridge regression (linear regression with a regularization term) is often a good starting point for developing a model, as it can provide a fast and well understood benchmark for a given task. Other variants of linear regression like LASSO regression [31] and elastic net regression [32] are also worth considering when there is a desire for a model to rely on a minimal number of features within the available data. Unfortunately, the relationships between features in the data are often non-linear and so using a model like a SVM is often a more appropriate choice for these cases [33]. SVMs are a powerful type of regression and classification model that uses kernel functions to transform a non-separable problem into a separable problem that is easier to solve. SVMs can be used to perform both linear and non-linear regression depending on the kernel function employed [34–37]. A good approach to developing a model is to train a linear SVM and a SVM with a radial basis function (RBF) kernel (a general-purpose non-linear type of SVM) in order to quantify what, if any, gain can be had from a non-linear model. Non-linear approaches can provide more powerful models but at the cost of easy interpretation of which features are influencing the model, a tradeoff mentioned in the introduction.

Many of the models that are commonly used in regression are also used for classification. Training a linear SVM and an SVM with a RBF kernel is also a good default starting point for a classification task. An additional method that can be tried is *k*-nearest neighbours classification [38]. Being one of the simplest classification methods, *k*-nearest neighbours provides a useful baseline performance marker against which other more complex models, like SVMs, can be compared. Another class of robust non-linear methods are ensemble-based models like random forests [39] and XGBoost [40,41]. Both methods are powerful non-linear models that have the added benefits of providing feature importance estimates and often require minimal hyperparameter tuning. Due to the assignment of feature importance values and the decision tree structure, these models are a good choice if understanding which features contributed the most to a prediction is essential for biological understanding.

For both classification and regression, the many available models tend to have a bewildering variety of flavours and variants. Trying to predict how well-suited a particular method will be to a particular problem *a priori* can be deceptive and instead taking an empirical, trial-and-error

approach to finding the best model is generally the most prudent approach. With modern machine learning suites such as scikit-learn [18], changing between these model variants often requires changing just one line of code, so a good overall strategy for selecting the best method is to train and optimize a variety of the aforementioned methods and choose the one with the best performance on the validation set, before finally comparing their performance on a separate test set.

Use of clustering models. The use of clustering algorithms (Figure 3e) is pervasive within biology [42,43]. *k*-means is a strong general-purpose approach to clustering that, like many other clustering algorithms, requires the number of clusters to be set as a hyperparameter [44]. DBSCAN is an alternative method that does not require the number of clusters to be predefined, but has the trade-off of other hyperparameters to set, [45]. Dimensionality reduction can also be performed prior to clustering to improve performance for datasets with a large number of features.

Dimensionality reduction. Dimensionality reduction techniques are used to transform data with a large number of attributes (or dimensions) into a lower-dimensional form while preserving the different relationships between the data points as much as possible. For example, data points that are similar (e.g. two homologous protein sequences) should also be similar in their lower-dimensional form and vice versa, dissimilar data points (e.g. unrelated protein sequences) should remain dissimilar [46,47]. Two or three dimensions are often chosen to allow visualisation of the data on a set of axes, though larger numbers of dimensions have uses in machine learning too. These techniques comprise both linear and nonlinear transformations of the data. Examples common in biology include principal component analysis (PCA) as shown in Figure 3d, Uniform Manifold Approximation and Projection (UMAP)) and t-distributed stochastic neighbour embedding (t-SNE) [48]. The technique to use depends on the situation: PCA retains global relationships between data points and is interpretable because each component is a linear combination of input features, meaning it is easy to understand which features give rise to variety in the data. t-SNE more strongly preserves local relationships between data points and is a flexible method that can reveal structure in complex datasets. Applications include single-cell transcriptomics for t-SNE [49] and molecular dynamics trajectory analysis for PCA.

Artificial neural networks

Artificial neural network models get their name from the fact that the form of the mathematical model that is being fit is inspired by the connectivity and behaviour of neurons in the brain and was originally designed to learn about brain function [50]. However, the neural networks in common use in data science are obsolete as brain models, and are now just machine learning models that can offer state-of-the-art performance in certain applications. Interest in neural network models has grown in recent decades owing to rapid advances in the architectures and

training of deep neural networks [26]. In this section, we describe basic neural networks, as well as varieties that are widely used in biological studies. Some of these are shown in Figure 4.

Basic principles of neural networks. A key property of neural networks is that they are universal function approximators, which means that, with very few assumptions, a correctly configured neural network can approximate any mathematical function to an arbitrary level of accuracy. In other words, if any process (biological or otherwise) can be thought of as some function of a set of variables, then that process can be modelled to any arbitrary degree of accuracy, governed by just the size or complexity of the model. The above definition of universal approximation is not mathematically rigorous, but does highlight one reason why interest in neural networks has persisted for decades. However, this guarantee does not provide a way of finding the optimal parameters of a neural network model that will produce the best approximation for a given dataset. There is also no guarantee that the model will provide accurate predictions for new data [51].

Artificial neurons are the building blocks of all neural network models. An artificial neuron is simply a mathematical function that maps (converts) inputs to outputs in a specific way. A single artificial neuron takes in any number of input values, applies a specific mathematical function to them, and returns an output value. The function used is usually represented as

$$y = \sigma\left(\sum_{i=1}^{n}(w_i x_i) + b\right),$$

where x_i represents a single input variable or feature (there are n such inputs), w_i represents a learnable weight for that input, b represents a learnable bias term, and σ represents a non-linear activation function that takes a single input, and returns a single output. To create a network, artificial neurons are arranged in layers, with the output of one layer being the input to the next. The nodes of the network can be thought of as holding the y values from the above equation, which become the x values for the next layer. We describe various approaches to arranging artificial neurons in the subsections below, which are called neural network architectures. It is also common to combine the different architecture types, for example in a convolutional neural network (CNN) used for classification, fully connected layers are usually used to produce the final classification output.

Multilayer perceptrons. The most basic layout of a neural network model is that of layers of artificial neurons arranged in a fully connected fashion, as shown in Figure 4a. In this layout, a fixed number of “input neurons" represent the input feature values calculated from the data that are fed to the network, and each connection between a pair of neurons represents one trainable weight parameter. These weights are the main adjustable parameters in a neural network, and optimizing these weights is what is meant by neural network training. At the other end of the network, a number of output neurons represent the final output values from the network. Such a network, when correctly configured, can be used to make complex, hierarchical decisions about the input, since each neuron in a given layer receives inputs from all neurons in the previous layer.

Layers of neurons in this simple arrangement are often called multilayer perceptrons and were the first networks useful for bioinformatics applications [52,53]. They are still widely used in a number of biological modelling applications today due to their ease and speed of training [13,54]. In many other applications, however, these simple architectures have been surpassed by newer model architectures discussed below, though some of these newer architectures still often make use of fully connected layers as subcomponents.

Convolutional neural networks. CNNs are ideally suited for image-like data, where the data possesses some type of local structure, and where the recognition of such structure is a key objective of the analysis. Using the example of images, this local structure could relate to specific types of objects in a field of view (e.g. cells in a microscopy image), represented by specific local patterns of colours and/or edges in spatially close pixels in an input image.

CNNs are composed of one or more convolutional layers (see Figure 4b), in which the output is the result of applying a small, one-layer fully connected neural network, called a filter or kernel, to local groups of features in the input. In the case of image-like inputs, this local area would be a small patch of pixels in the image. The outputs of a convolutional layer are also image-like arrays, carrying the result of "sliding" the filter over the entire input and computing an output at each position. Crucially, the same filter is used across all pixels, allowing the filters to learn local structure in the input data. It is common in deeper CNNs to use skip connections that allow the input signal to bypass one or more layers in addition to passing through the processing units in the layer. This type of network is called a residual network (ResNet) and allows training to converge more quickly on performant solutions.

CNNs can be configured to operate effectively on data of different spatial structure. For example, a 1D CNN would have filters that slide in just one direction (say left to right); this type of CNN would be suitable for data that have only one spatial dimension (such as text or biological sequences). 2D CNNs operate on data with 2 spatial dimensions, such as digital images. 3D CNNs operate on volumetric data, such as magnetic resonance imaging (MRI) scans.

CNNs have seen significant success in biology for a variety of data types. Recent advances in protein structure prediction have used information on the co-evolution of residue pairs in related protein sequences to extract information on residue pair contacts and distances, allowing predictions of 3D protein structures to be built at unprecedented accuracy [23,55]. In this case, the network learns to pick out direct coupling interactions, and accurate predictions can be made even for sequences with few or no related sequences [56]. CNNs have also been applied successfully to identifying variants in genetic sequence data [57], 3D genome folding [58], DNA–protein interactions [22,59], cryogenic electron microscopy (cryo-EM) image analysis [60,61] and image classification in medically-important contexts (such as detection of malignancy), where they often now rival expert human performance [24,62].

Recurrent neural networks. Recurrent neural networks (RNNs) are most suited to data that is in the form of ordered sequences, such that there exists (at least notionally) some dependence or correlation between one point in the sequence and the next. Probably their main application outside biology is in natural language processing, where text is treated as a sequence of words or characters. As shown in Figure 4c, recurrent networks can be thought of as a block of neural network layers that take as input the data corresponding to each entry (or time step) in a sequence, and produce an output for each entry that is dependent on entries that have previously been processed. They can also be used to generate a representation of the whole sequence that is passed to later layers of the network to generate the output. This is useful as sequences of any length can be converted to a fixed size representation and input to a multilayer perceptron. Obvious examples for the use of RNNs in biology include analysis of gene or protein sequences, with tasks including identifying promoter regions from gene sequences, predicting protein secondary structure, or modelling gene expression levels over time; in the latter case, the value at a given time point would count as one entry in a sequence. The more advanced long short-term memory (LSTM) or gated recurrent unit (GRU) variants of RNNs, which have many uses in biology including protein structure prediction [63,64], peptide design [65] and predicting clinical diagnosis from health records [66]. These more advanced methods are often used in combination with CNNs, which can increase accuracy [67]. RNNs can be very robust in analyzing sequence-based data. For example, RNNs trained on millions of protein sequences have shown an ability to capture evolutionary and structural information, and can be applied to a variety of supervised tasks including tasks related to the design of novel protein sequences [68].

Role of attention mechanisms and the use of transformers. A problem identified with RNNs is the difficulty they have in examining specific parts of an input sequence, which is important in order to generate a highly accurate output. The addition of an attention mechanism to RNNs, which allows the model to access all parts of the input sequence when calculating each output, was introduced to alleviate this problem. Recently it has been shown that the RNN is not even required at all, and that attention alone can be used by itself; the resulting models, called transformers, have obtained state-of-the-art results on many natural language processing benchmarks [69]. Transformer models have recently shown better accuracy than RNNs for tasks on biological sequences, but it remains to be seen whether these methods, which are often trained on billions of sequences using thousands of GPUs, will be able to outperform existing alignment-based methods of sequence analysis in bioinformatics [70]. The outstanding success of AlphaFold2 at the Critical Assessment of protein Structure Prediction 14 (CASP14) experiment, a blind assessment of computational approaches to predict protein structure from sequence, suggests that models using attention also hold promise for tasks in structural biology [71].

Graph convolutional networks. Graph convolutional networks (GCNs) are particularly suitable for data that, whilst not having any obvious visible structure like an image, are nonetheless composed

of entities connected by arbitrary specified relationships, or interactions [72]. Examples of such data relevant to biology are molecules (composed of atoms and bonds) [73–76] and protein–protein interaction networks (composed of proteins and interactions) [77]. A graph, in computational terms, is just a representation of data of this type, with each graph having a set of vertices or nodes, and a set of edges that represent various types of relationships or connections between the nodes. Using the examples given above, representations of atoms or proteins might be classed as node features, and bonds or interactions might be classed as edge features. GCNs use the structure of the resulting graph to determine the flow of information in the neural network model. As shown in Figure 4d, adjacent nodes are considered when updating the features of each node throughout the network, with the node features in the last layer being used as the output (e.g. interacting residues on a protein) or combined to form an output for the whole graph (e.g. fold type of the protein). Graphs representing different associations can combine different sources of information when making predictions, such as combining drug–gene and food–gene relationship graphs to predict foods for cancer prevention [78]. Software for training GCNs includes PyTorch Geometric (https://pytorch-geometric.readthedocs.io/en/latest) [79] and Graph Nets (https://github.com/deepmind/graph_nets) [72].

Autoencoders. As the name suggests, autoencoders are neural network architectures designed to self (auto) encode a collection of data points by representing them as points in a new space of pre-determined dimensionality, usually far fewer than the number of input dimensions. One neural network (the encoder) is trained to convert the input into a compact internal representation, called a latent vector or latent representation, representing a single point in the new space. The second part of an autoencoder, called the decoder, takes the latent vector as input and is trained to produce as output the original data with its original dimensionality (Figure 4e). Another way of looking at it is that the encoder tries to compress the input, and the decoder tries to decompress it back again. The encoder, latent representation, and decoder are trained at the same time. Although this sounds like a pointless exercise, where the output just mimics the input, the idea is to learn a new representation of the input data that compactly encodes desirable features, such as similarity between the data points, while still retaining the ability to reconstruct the original data using the learned latent representation. Applications include predicting how closely related two data points are and enforcing some structure on the latent space that is useful for further prediction tasks. Another benefit of the encoder–decoder architecture is that, once trained, the decoder can be used in isolation to generate predictions of new, synthetic data samples which can be tested in the lab and contribute to synthetic biology efforts [80]. Autoencoders have been applied to a range of biological problems including predicting DNA methylation state [81], the engineering of gene and protein sequences [82,83] and single-cell RNA-sequencing analysis [84].

Training and improving neural networks. The general procedure for training machine learning models has been outlined in Box 1. However, as neural networks are structurally much more complex than the traditional machine learning algorithms, there are some concerns that are specific

to neural networks. Having picked a neural network as an appropriate model for the intended application (Figure 1) it is often a good idea to train it on just a single training example, e.g. a single image or gene sequence. This trained model is not useful for making predictions, but the training is good at revealing programming errors. The training loss function should very quickly go to zero as the network simply memorizes the input; if it doesn't, there is likely an error in the code, or the algorithm is not complex enough to model the input data. Once the network has passed this basic debugging test, training on the whole training set can proceed where the training loss function is minimized. This may require tuning of hyperparameters such as the learning rate (Figure 2e). By monitoring loss on the training and validation sets, overfitting of the network can be detected where the training loss continues to drop lower and the loss on the validation set starts to increase. Training is usually stopped at that point, a process known as early stopping (Figure 2f). Overfitting of a neural network (or any machine learning model), visualised in Figure 2d, means that the model is starting to simply memorize features of the training set and thus starting to lose its ability to generalize to new data. Early stopping is a good way of preventing this, but other techniques can be used during training, such as regularisation of the model or dropout techniques, where nodes in the network are randomly ignored to force the network to learn a more robust prediction strategy involving multiple nodes.

Popular software packages used to train neural networks include PyTorch (https://pytorch.org) [85] and Tensorflow (https://www.tensorflow.org) [85]. Training neural networks is computationally demanding, usually requiring a GPU (Graphics Processing Unit) or TPU (Tensor Processing Unit) with sufficient memory, as these devices can provide a 10-100x speedup over using the standard, central processing unit (CPU). This speedup is required when training the larger models that have shown success in recent years, and when training on large datasets. However, running an already trained model is usually considerably faster and is often feasible on just an ordinary CPU. Cloud computing solutions from common providers exist for those without access to a GPU for training, and it is worth noting that for small tasks, the Colaboratory (Colab) (https://research.google.com/colaboratory) allows Python code to be tested on either GPUs or TPUs free of charge. Using Colab is an excellent way of getting started with Python-based deep learning.

Challenges for biological applications

Perhaps the single biggest challenge of modelling biological data is its sheer variety [1]. Biologists work with data such as gene and protein sequences, gene expression levels over time, evolutionary trees, microscopy images, 3D structures, and interaction networks, to name but a few. We have summarised some best practices and important considerations for specific biological data types in Table 2. Owing to the diversity of data types encountered, biological data often require somewhat bespoke solutions for handling them effectively, and this makes it difficult to recommend off-the-shelf tools or even general guidelines for the use of machine learning in these problem domains,

as the choice of model, training and test data will depend heavily on the exact questions one wants to answer. Nevertheless, there are some common issues that need to be considered for the successful use of machine learning in biology, but also more generally.

Data availability. Biology is somewhat unique in that there exist some problem domains that have very large quantities of data publicly available, whereas other problem domains have very little. An example is the relative abundance of biological sequence data in public databases such as NCBI GenBank and UniProt, whereas reliable data on protein interactions is much harder to come by. The quantity of data available for a given problem has a profound impact on the choice of techniques that can effectively be used. As a very rough guideline, when only small amounts of data (hundreds or a few thousand examples) are available, one is essentially forced to use more traditional machine learning methods, as these are more likely to produce robust predictions. When larger quantities are available, one can start to consider more highly parameterized models such as deep neural networks. In supervised machine learning, the relative proportions of each ground truth label in the dataset should also be considered, with more data required for machine learning to work if some labels are rare [86].

Data leakage. Although the scale and complexity of biological data may make it seem ideal for machine learning, there are some important considerations that need to be borne in mind [21,87,88]. One key concern is how to validate the performance of a model. The common setup of training, validation and test sets can lead to problems such as researchers repeatedly testing on the same test set with a variety of models to obtain the best accuracy, and hence risking overestimating performance on it without generalizing to other test sets or new data. However, biological data presents a further non-trivial question: in a large dataset with related entries (e.g. as a result of familial relationships, or evolutionary relationships), how does one ensure that two closely related entries don't end up split between the training and test sets? If this occurs then the ability of the model to remember specific cases is tested, rather than its ability to predict the property in question. This is one type of issue often called data leakage and leads to results that appear better than they are, which is perhaps one reason researchers are reluctant to be rigorous about the issue. Other types of data leakage are possible, for example, using any data or features during training that would not be available during testing. Here, we focus on the problem of having related samples in the training and testing sets.

What we mean by "related" here depends on the nature of the study. It might be a case of sampling data from the same patient, or the same organism. However, the classic situation where data leakage occurs in biology is seen in studies on protein sequences and structures. Typically, but usually not correctly, researchers try to ensure that no protein in the training set has sequence identity above a certain threshold to any protein in the test set, usually at a threshold of 30% or 25%. This is enough to exclude many homologous pairs of proteins, but it has been known for decades that some homologous proteins can have virtually no sequence similarity [89,90], which

would mean that simply filtering by sequence identity would be insufficient to prevent data leakage. This is particularly important for models that operate on sequence alignments or sequence profiles as input, since, although two individual protein sequences may not share any obvious similarity, their profiles could be virtually identical. This means that for a machine learning model, these two profiles would essentially be the same data point —. both will be describing the same protein family. For protein sequences, one solution to avoid this problem is to search the test data with a sensitive hidden Markov model (HMM) profile comparison tool such as HH-suite, which can find sequences distantly related to the training data [91]. In the common case where protein structure is being used as input or output, structural classifications such as CATH [92] or ECOD [93] can be used to exclude similar folds or homologous proteins. Similar issues affect studies predicting protein–ligand binding affinity [94].

To be clear, data leakage is not an intrinsic issue with any particular type of data, but rather it is a problem with how the data are used when training and evaluating machine learning models. One would certainly expect a trained model to produce very good results on data that are similar to the training set. The issue of data leakage becomes a problem when a model that appears performant on some benchmark set performs poorly on new data that are actually different to the training set; in other words, the model does not generalise, likely because it has not modelled the true relationship between the variables, but rather remembered hidden associations present in the data.

Thanks to frequent complaints from reviewers, some journals are now starting to require rigorous benchmarking to be carried out before a paper can be considered for publication. Without proper testing, the performance of a model will very likely not be representative of real-world performance on unseen data, which undermines user confidence in the model. Worse, authors of future studies may be misled into thinking that inadequate testing is defensible simply because it has already appeared in (possibly several) peer-reviewed papers, even though it is not. As mentioned in Box 2, authors, peer reviewers and journal editors all have a responsibility for making sure that data leakage has been avoided. Knowingly leaving these kinds of errors in place is really little better than fabricating data at the end of the day.

Interpretability of models. It is usually the case that biologists want to know why a particular model is making a particular prediction, i.e. what features of the input data is the model responding to and how, and why it works in some cases but not others. In other words, biologists are often interested in discovering mechanisms and the factors responsible for modelling output, rather than just accurate modelling, as mentioned previously. The ability to interpret a model depends on the machine learning method used and the input data. Interpretation is usually easier for non-neural network methods, since these have feature sets more amenable to direct meaningful interpretation and generally have fewer learnable parameters. In the case of, say, a simple linear regression model, the parameter assigned to each input feature gives a direct indication of how that feature affects the prediction.

The low cost of training non-neural network methods means that it is advisable to carry out ablation studies, where the effect on performance of removing defined features of the input is measured. Ablation studies can reveal which features are most useful for the modelling task at hand, and are one way to possibly discover more robust, efficient and interpretable models.

Interpreting a neural network (particularly a deep neural network) is generally much harder due to the frequently large number of input features and parameters in the model. It is still possible to identify, for example, regions in an input image most responsible for a particular classification by building a saliency map [28]. Although saliency maps show which regions of an image are important, it can be more difficult to pinpoint which properties of the data in these locations were responsible for the prediction, particularly when the inputs are not easily human-interpretable like images and text. Nevertheless, saliency maps and similar representations can be useful as a 'sanity check' to ensure that the model is indeed looking at the relevant parts of an image. This can help avoid situations where models make unintended connections, such as classifying medical images based on hospital or department labels in the corner of the image rather than the medical content of the image itself [95]. Generating adversarial examples, synthetic inputs that cause a neural network to produce confident incorrect predictions, can also be a good way of providing information on which features are being most used for prediction [96]. For example, CNNs often use textures (such as stripes in animal fur) to classify objects in images where humans would primarily make use of the shapes [51].

Privacy-preserving machine learning. Some biological data, most notably human genomics data and commercially sensitive pharmaceutical data, has data privacy implications. There have been a number of efforts to allow sharing of data and distributed training of machine learning models in the context of data privacy. For example, modern cryptographic techniques allow training of a drug-target interaction model where the data and results are provably secure [97]. Simulated, synthetic participants that closely resemble real participants in a clinical trial can lead to results that are accurate for real participants without revealing identifying data [98]. Algorithms have been developed for efficient federated model training with data stored in different places [99].

The need for interdisciplinary collaborations. Unless publicly available data is being used, it is rare that one research group will have the expertise and resources to both collect data for a machine learning study and also apply the most appropriate machine learning method effectively. It is common for experimental biologists to collaborate with computer scientists, with such collaborations often obtaining excellent results. It is, however, important in such collaborations that each side has some working knowledge of the other. In particular, computer scientists should make an effort to understand the data, such as the expected degree of noise and reproducibility, and biologists should understand the limitations of the machine learning algorithms being used.

Building such understanding takes time and effort, but is important to prevent the often unintentional dissemination of poor models and misleading results.

Future directions

The increasing use of machine learning in biological studies looks set to continue for the foreseeable future. This increased uptake has been enabled by important advances in methodology, software and hardware, all of which keep on developing. A number of large technology companies are using their technical expertise and considerable resources to assist academic researchers, or even carry out their own research in biology with innovative machine learning strategies. To date, however, most success has come from applying algorithms developed in other fields directly to biological data. For example, CNNs and RNNs were developed for applications such as image analysis (for face recognition or in self-driving cars) and natural language processing, respectively. One of the most exciting prospects for machine learning in biology is algorithms tailored specifically to biological data and biological questions [100,101]. If the known structure of a biological system can be exploited and neural networks used to learn the unknown parts, then increasingly heavily parameterised models can be replaced with simpler ones that are more amenable to interpretation and more robust on new data [102]. Applications include biological reaction systems and pharmacokinetics, where systems of known differential equations can be exploited. This will also assist in the move from predictive machine learning to generative models that can create new entities, such as designing proteins with novel structures and functions [103,104].

As the variety of useful architectures and input data types increases, the paradigm of differentiable programming is emerging from the field of deep learning [105]. Differentiable programming is the use of automatic differentiation, the central concept in training neural networks, to calculate gradients and improve parameters in any desired algorithm. This shows promise for physical models of biological systems in protein structure prediction [63,106], and for learning force field parameters for molecular dynamics simulations [107,108]. The development of differentiable software packages such as Jax [109] and packages tailored to specific areas of biology such as Selene [110], Janggu [111] and Jax MD [112] will assist the development of such methods.

The progress in biological data analysis with machine learning has also been enabled owing to deposition of trained models in publically-available repositories. In this way, researchers working on similar problems can use these models without the need for training, and a variety of different models can be used with only minimal effort required for switching between them [113]. The field has also seen an expansion of automated machine learning pipelines, which train and tune a variety of models without user input and return the best performing model. These may assist non-experts in training models [114]. However, these resources cannot replace a thorough understanding of the method being used, which is important for choosing the appropriate inductive bias and interpreting

the predictions of the model. It remains to be seen whether in the future automated machine learning will be reliable and flexible enough to allow experimentalists to routinely use complex machine learning algorithms independently, or whether machine learning expertise will remain a necessity.

As has been discussed, rigorous validation of models and comparison of different models is challenging but remains necessary to identify the best performing models and inform future research directions. For the field to progress it will be necessary to develop benchmark datasets and validation tasks [115], such as ProteinNet [116], ATOM3D [117] and TAPE [118], and for these to become widely used. Of course, over-optimizing to a particular benchmark can occur and it is important that researchers resist the temptation to do this in order to make their results seem better. Blind assessments such as CASP [119] and Critical Assessment of Functional Annotation (CAFA) [120] will continue to play an important role in assessing which models perform best.

Overall, the variety of biological data makes it hard to provide general guidelines for machine learning in biology. Hence, we have aimed here to give biologists an overview of the different methods available and to provide them with some ideas about how to carry out effective machine learning with their data. Of course, machine learning is not suited to every problem, and it is just as important to know when to avoid it: when there is not sufficient data, when understanding rather than prediction is required, or when it is unclear how to assess performance in a fair manner. The boundaries of when machine learning is useful in biology are still being explored and will continue to change in accordance with the nature and volume of available experimental data. Undeniably, though, machine learning has had a huge impact on biology and will continue to do so.

Acknowledgements

The authors thank members of the UCL Bioinformatics Group for valuable discussions and comments. This work was supported by the European Research Council Advanced Grant "ProCovar" (project ID 695558).

Author contributions

All authors researched data for the article, contributed substantially to discussion of the content, wrote the article and reviewed the manuscript before submission.

Competing interests

The authors declare no competing interests.

Table 1 - Comparing different machine learning methods.

Method	Type of data	Example applications	Advantages	Disadvantages
Ridge (& LASSO/Elastic) Regression	Labelled Fixed number of features	Protein variant effect prediction [121] Chemical/biochemical reaction kinetics [122]	Easy to interpret Easy to train Good benchmark	Can't learn complex feature relationships Overfits with a large number of features
Support-vector machine (SVM)	Labelled Fixed number of features	Protein function prediction [123] Transmembrane protein topology prediction [124]	Can perform both linear and non-linear classification and regression	Scaling to large datasets is often difficult
Random forest	Labelled Fixed number of features	Prediction of disease-associated genome mutations[125] Scoring of protein-ligand interactions [39]	Learns how important each feature is to the prediction Individual decision trees are human-readable allowing interpretation of how a decision is made Less sensitive to feature scaling and normalization so easier to train and tune	Less appropriate for regression Many decision trees are hard to interpret
Gradient boosting (e.g. XGBoost)	Labelled Fixed number of features	Gene expression profiling [126]	Learns how important each feature is to the prediction Decision trees are human-readable allowing interpretation of how a decision is made	Can struggle to learn underlying signal if noise is present Less appropriate for regression

			Less sensitive to feature scaling and normalization so easier to train and tune	
Clustering	Unlabelled Fixed number of features	Differential gene expression analysis [15] Model selection in protein structure prediction [127]	For low-dimensional data, good clustering is easily identifiable Cluster validation metrics are available to assess performance	Scaling to large datasets is difficult for some methods Noisy datasets sometimes yield contradictory results
Dimensionality reduction	Unlabelled Large and fixed number of features	Single-cell transcriptomics [49] Analysis of molecular dynamics trajectories [128]	Provides visual representation of data Goodness-of-fit evaluations usually available to assess performance	Hard to preserve both global and local differences in data Scaling to large numbers of samples is difficult for some methods
Multilayer perceptrons	Labelled Fixed number of features	Protein secondary structure prediction [13] Drug toxicity prediction [54]	Can fit datasets with fewer layers than architectures such as CNNs, making them easier and faster to train	Easy to overfit Large number of parameters Hard to interpret
Convolutional neural networks (CNNs)	Spatial data arranged in a grid e.g. 2D image (pixels) or 3D volumes (voxels) Allows variable input size	Protein residue–residue contact and distance prediction [23] Medical image recognition [24,57]	Variable input size Learns patterns irrespective of location in input	Receptive field, the amount of the input that is considered when predicting the output for each pixel, can be limited Hard to train deeper architectures that use many layers to increase the receptive field and make more complex predictions
Recurrent neural networks	Sequential data, e.g. biological sequences	Protein engineering [68]	Variable input size	Long training times

(RNNs)	or time-series data Allows variable input size	Predicting clinical events [66]	Sequences are found in many areas of biology	High computing memory requirements
Graph convolutional networks (GCNs)	Data characterised by connections between entities (spatial, interaction or association) Allows variable input size	Predicting drug properties [77] Interpreting molecular structures [73,74] Knowledge extraction [129]	Variable graph sizes supported, which is important because most graphs in biology have variable size Learns patterns by following graph connectivity so predictor uses most relevant associations	High computing memory requirements for large, densely connected graphs Hard to train deeper architectures
Autoencoders	Labelled or unlabelled data Fixed or variable input size depending on architecture	Protein and gene engineering [82] Prediction of DNA methylation [81] Neural population dynamics [130]	Latent space provides low-dimensional representation that can be used to visualise input data Can generate new samples, which is useful in areas such as protein design	Latent space specific to data in training set and may not be appropriate to other datasets Testing newly generated samples often requires wet lab experiments

Each method has types of data and applications to which it is best suited, along with advantages and disadvantages when compared to other methods.

Table 2 - Recommendations for the use of machine learning strategies for different biological data types.

Input data	Example prediction tasks	Recommended models	Challenges
Gene sequence	Genome accessibility [14] 3D genome organisation [58] Enhancer–promoter interactions [40]	1D convolutional neural networks (CNNs) Recurrent neural networks (RNNs) Transformers	Repetitive regions in genome Sparse regions of interest Very long sequences
Protein sequence	Protein structure [23,55] Protein function [131] Protein–protein interaction [132]	2D CNNs and residual networks (ResNets) using covariation data Multilayer perceptrons with windowing Transformers	Metagenome data stored in many places and therefore hard to access Data leakage (from homology) can make validation difficult
Protein 3D structure	Protein model refinement [133] Protein model quality assessment [134] Change in stability on mutation [135]	Graph convolutional networks (GCNs) using molecular graph 3D CNNs using coordinates Traditional methods using structural features Clustering	Lack of data, particularly on protein complexes Lack of data on disordered proteins
Gene expression	Intergenic interactions or co-expression [136] Organisation of transcription machinery [137]	Clustering CNNs Autoencoders	Unclear link between co-expression and function High dimensionality High noise
Mass spectrometry	Detecting peaks in spectra [138]	CNNs using spectral data Traditional methods	Lack of standardised benchmarks [140]

	Metabolite annotation [139]	using derived features	Normalisation* required between different datasets
Images	Medical image recognition [24,62] Cryo-EM image reconstruction [60,141] RNA-Seq profiles [142]	2D CNNs and ResNets Autoencoders Traditional methods using image features	Systematic differences in data collection affect prediction Hard to obtain large datasets of consistent data
Molecular structure	Antibiotic activity [73] Drug toxicity [54] Protein-ligand docking [39] Novel drug generation [143]	GCNs using molecular graph Traditional methods or multilayer perceptrons using molecular properties RNNs using text-based representations of molecular structure such as SMILES Autoencoders	Only a tiny fraction of possible small molecules have experimental data available
Protein–protein interaction network	Polypharmacology side effects [77] Protein function [144]	GCNs Graph embedding	Interaction networks can be incomplete Cellular location affects whether proteins interact High number of possible combinations

Each type of biological data has prediction tasks where it has been used effectively, machine learning models that are appropriate and specific challenges when using machine learning. Some challenges, such as high dimensionality, affect most biological data types.

* Normalisation means rescaling or otherwise transforming variables from different data sets with the intention that their contributions should carry roughly equal weight and their ranges are so that they become comparable on a joint common scale of values. The most common way of achieving this is by subtracting the means of each variable and dividing by their standard deviations, which can also be called “standardisation”. This is required because different instruments, experimental protocols etc. can produce systematic differences in measurements of the same quantities, rendering it difficult or impossible to compare trends between experiments.

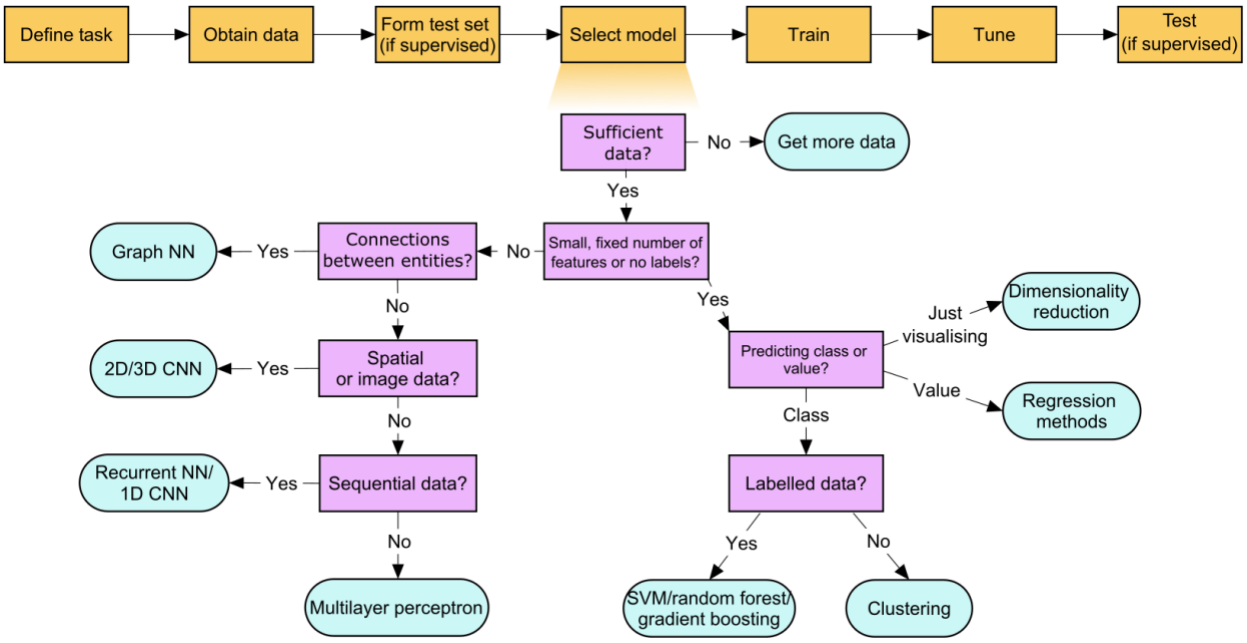

Figure 1 - Choosing and training a machine learning method. The overall procedure for training a machine learning method is shown along the top. A decision tree to assist researchers in selecting a model is given below. This flowchart is intended to be used as a visual guide linking the concepts outlined in the paper. However, a simple overview such as this cannot cover every case. For example, the number of data points required for machine learning to become applicable depends on the number of features available for each data point, with more features requiring more data points, and also depends on the model being used. There are also deep learning models that work on unlabelled data.

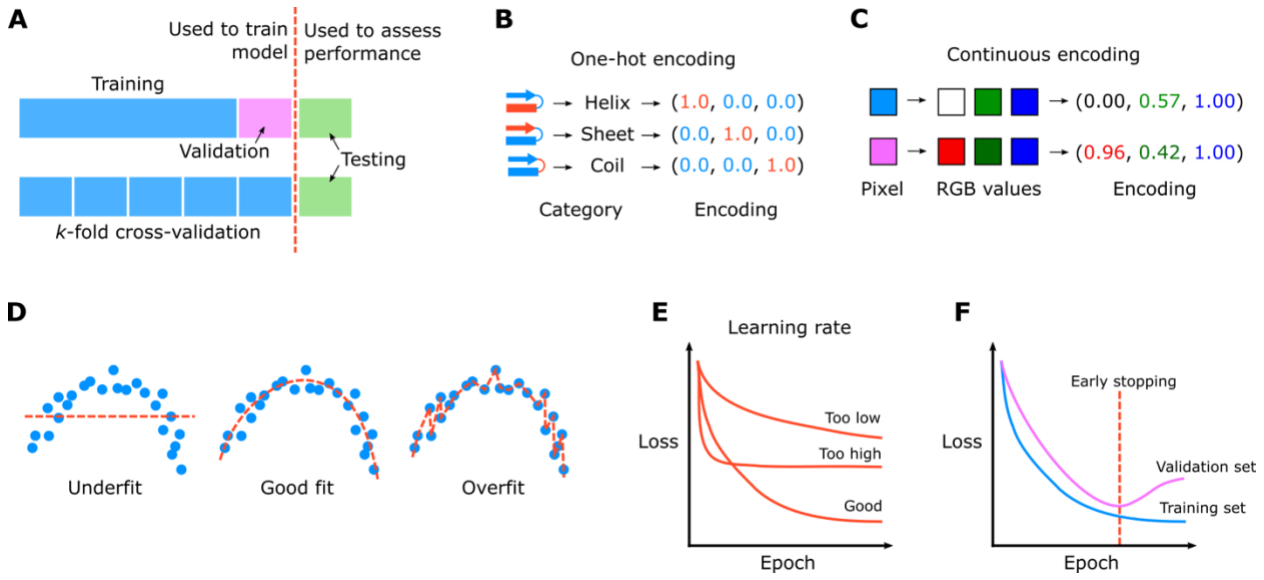

Figure 2 - Training machine learning methods. (a) Available data is often split into training, validation and test sets. The training set is directly used to train the model, the validation set is used to monitor training, and the test set is used to assess the performance of the model. *k*-fold cross-validation with a test set can also be used. (b) One-hot encoding is a common approach for representing categorical inputs where a single choice is permitted from a number of possibilities, in this case three possible protein secondary structure classes. The result of the encoding is a vector with three numbers, all equal to 0 except the occupied class which is set to 1. This vector is used by the machine learning model. (c) Continuous encoding represents numerical inputs, in this case the red, green and blue values of a pixel in an image. Again the result is a vector with three numbers corresponding to the amount of red, green and blue in the pixel. (d) Failing to learn the underlying relationship between the variables is called underfitting, whereas learning the noise in the training data is called overfitting. Underfitting can be caused by using a model without sufficient complexity to describe the signal. Overfitting can be caused by using a model with too many parameters or by continuing training after it has learned an approximation of the true relationship between the variables. (e) The learning rate of the model determines how quickly learned parameters are adjusted when training a neural network or some traditional methods such as gradient boosting. A low learning rate can lead to slow training, which is time consuming and requires considerable computing power. By contrast, a high learning rate can lead to quick convergence on a non-optimal solution and poor performance of the model. (f) Early stopping is the process of terminating training at the point where the loss function on the validation set starts to increase, even if the loss function on the training set is still decreasing. Use of early stopping can prevent overfitting.

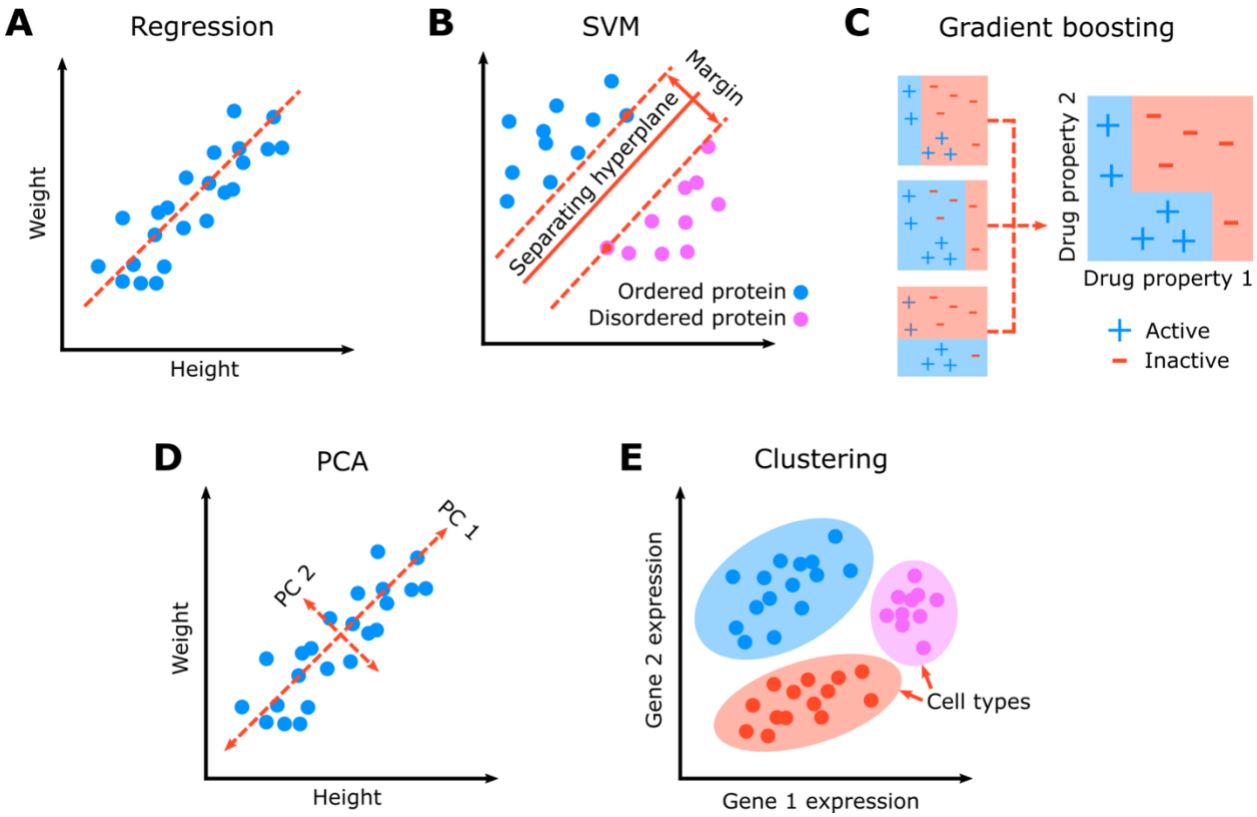

Figure 3 - Traditional machine learning methods. (a) Regression finds the relationship between a dependent variable (the observed property) and one or more independent variables (features), for example predicting the weight of a person from their height. (b) A support-vector machine (SVM) transforms the original input data such that in their transformed versions (called the latent representation), data belonging to separate categories are divided by a clear gap that is made as wide as possible. In this case we show a prediction of whether a protein is ordered or disordered, with the axes representing dimensions of the transformed data (c) Gradient boosting uses an ensemble of weak prediction models, typically decision trees, to make predictions. For example, active drugs can be predicted from molecular descriptors such as molecular weight and the presence of particular chemical groups. Individual predictors are combined in a stage-wise manner to make the final prediction. (d) Principal component analysis (PCA) finds a series of feature combinations that best describe the data whilst being orthogonal to each other. It is commonly used for dimensionality reduction. In the case of the height and weight of a person, the first principal component (PC1), corresponding to a linear combination of height and weight, describes the strong positive correlation, whereas PC2 might describe another variables that do not correlate strongly with those, such as percentage body fat or muscle mass. (e) Clustering uses one of various algorithms to group sets of similar objects, for example grouping cell types based on gene expression profiles.

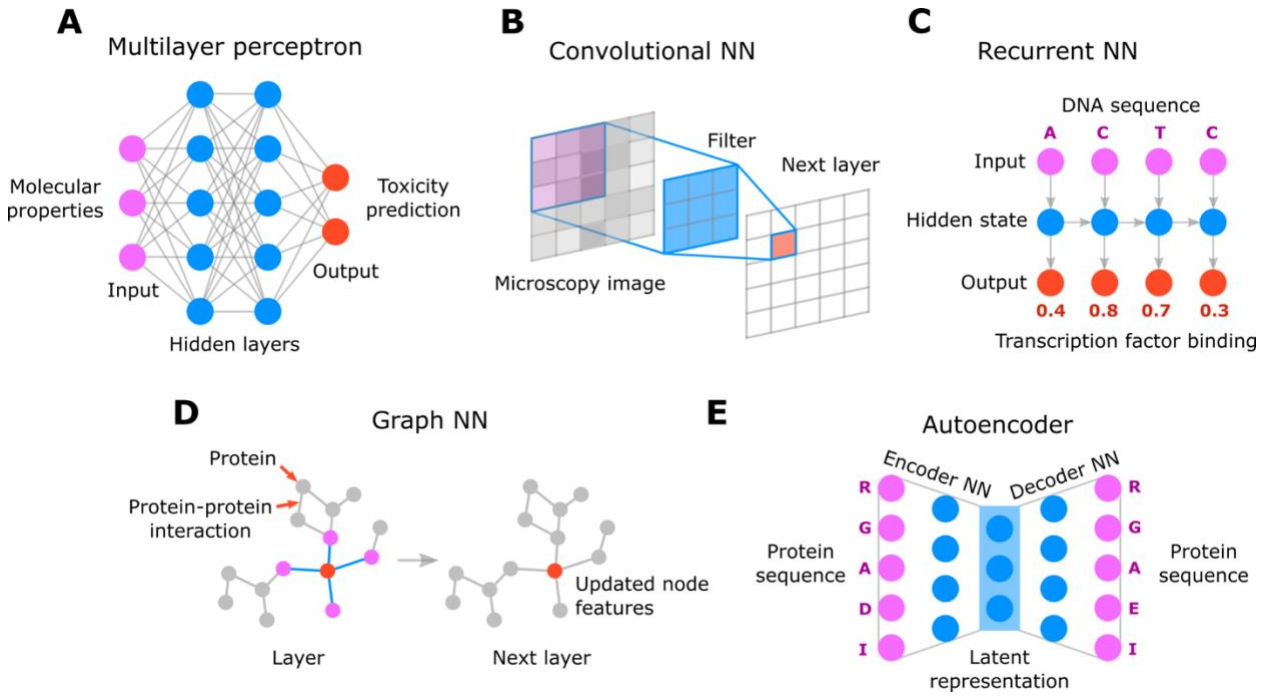

Figure 4 - Neural network methods. (a) A multilayer perceptron consists of nodes (shown as circles) that represent numbers: either an input value, an output value, or an internal (hidden) value. Nodes are arranged in layers with connections, indicating learned parameters, between every node of a layer and every node of the next layer. For example, molecular properties can be used to predict drug toxicity since the prediction can be made from some complicated combination of independent input features. (b) A convolutional neural network (CNN) uses filters that move across the input layer and are used to calculate the values in the next layer. The filters operating across the whole layer mean that parameters are shared, allowing similar entities to be detected regardless of location. A 2D CNN is shown operating on a microscopy image but 1D and 3D CNNs also find applications in biology. The dimensionality, in this case, refers to how many spatial dimensions there are in the data, and the connectivity within the CNNs can be configured accordingly. For example, biological sequences can be considered 1D and Magnetic Resonance Imaging data can be considered 3D. (c) A recurrent neural network (RNN) processes each part of a sequential input using the same learned parameters , giving an output and an updated hidden state for every input. The hidden state is used to carry information about the preceding parts of the sequence. In this case the probability of transcription factor binding is predicted for each base in a DNA sequence. The RNN is expanded to show how each output is generated using the same layers; this should not be confused with using different layers for each output. (d) A graph convolutional network (GCN) uses information from connected nodes in a graph, such as a protein–protein interaction network, to update node properties the network by combining predictions from all neighbouring nodes. The updated node properties form the next layer in the network and predict the desired property in the output layer. (e) An autoencoder consists of an encoder neural network, which converts an input into a lower-dimensional latent representation, and a decoder neural network, which converts this latent representation back to the original input form. For example,

protein sequences can be encoded and the latent representation used to generate novel protein sequences. In the example 4 of the 5 residues are the same as the input after encoding and decoding by the autoencoder, indicating an accuracy of 80% on this sequence.

Box 1 - Doing machine learning.

In this box we outline the steps that should be taken when training a machine learning model. There is surprisingly little guidance available on the model selection and training process [145,146], with descriptions of the stepping stones and failed models rarely making it into published research papers. The first step, before touching any machine learning code, should be to fully understand the data (inputs) and prediction task (outputs) at hand. This means a biological understanding of the question, such as knowing the origin of the data and the sources of noise, and having an idea of how the output could theoretically be predicted from the input using biological principles. For example, it can be reasoned that different amino acids might have preferences for particular secondary structures in proteins, so it makes sense to predict secondary structure from amino acid frequencies at each position in a protein sequence. It is also important to know how the inputs and outputs are stored computationally. Are they normalised to prevent one feature having an unduly large influence on prediction? Are they encoded as binary variables or continuously? Are there duplicate entries? Are there missing data elements?

Next, the data should be split to allow training, validation and testing. There are a number of ways to do this, two of which are shown in Figure 2a. The training set is used to directly update the parameters of the model being trained. The validation set, usually around 10% of the available data, is used to monitor training, select hyperparameters and prevent the model overfitting to the training data. Often *k*-fold cross-validation is used: the training set is split into *k* evenly sized partitions (e.g. 5 or 10) to form *k* different training and validation sets, and the performance compared across each partition to select the best hyperparameters. The test set, sometimes called the hold-out set, typically also around 10% of the available data, is used to assess the performance of the model on data not used for training or validation (i.e. estimate its expected real-world performance). The test set should be used only once at the very end of the study or as little as possible [27,38] to avoid tuning the model to fit the test set. See the "Data leakage" subsection of the article for issues to consider when making a fair test set.

The next step is model selection, which depends on the nature of the data and prediction task and is summarised in Figure 1. The training set is used to train the model following best practices of the software framework being used. Most methods have a handful of hyperparameters that need to be tuned in order to achieve the best performance. This can be done using random search or grid search, and be combined with *k*-fold cross-validation as outlined above [27]. Model ensembling should be considered, where the outputs of a number of similar models are simply averaged to give a relatively reliable way to boost overall accuracy of the modelling task. Finally, the accuracy of the model on the test set (see above) should be assessed.

Box 2 - Evaluating articles that use machine learning.

Here are some questions to consider when reading or reviewing papers that use machine learning on biological data. It is useful to bear these considerations in mind, even if the answers are not apparent, and these questions can be used as the basis for a discussion with collaborators with the required expertise. A surprising number of papers do not pass these criteria [147].

Is the dataset adequately described? Complete steps to assemble the dataset should be provided, ideally with the dataset or summary data (e.g. biological database IDs) available at a persistent URL. In our experience, a thorough description of the machine learning methodology but with only a cursory reference to the data is a red flag. If a standard dataset or dataset from another study is being used, then this should be adequately justified in the paper.

Is the test set valid? Based on discussion in " Challenges for biological applications " section, check that the test set is sufficient to benchmark the property under investigation. There should be no data leakage between the training and test sets, the test set should be of a large enough size to give reliable results, and the test set should mirror the range of examples a standard user of the tool would be likely to use it on. The composition and size of the training and test sets should be discussed in detail. Authors have a responsibility to ensure that all steps have been taken to avoid data leakage, and these steps should be described in the article, along with the rationale behind them. Journal editors and peer reviewers should also ensure that these tasks have been performed to a good standard, and certainly should never just assume that they have been.

Is the model choice justified? Reasons should be given for the choice of machine learning method. Neural networks should be used because they are appropriate to the data and question in hand, and not just because everyone else is using them. Discussion of models that were tried and didn't work should be encouraged as it may help others; too often a complex model is presented without any discussion of the inevitable trial and error that will have been required to end up with that model.

Have they compared to other methods? A novel method should be compared to existing methods that show good performance and are used in the community. Ideally methods using a variety of model types should be compared, which can aid in interpreting results. It is surprising how many complex models can be matched in performance by simple regression methods.

Are the results too good to be true? Claims of 99%+ accuracy are not uncommon in machine learning papers for biology. Usually, this is a sign of a problem with the cross-validation rather than an amazing breakthrough. Both authors and reviewers should take note of this point.

Is the method available? At the very least, someone who wants to use a trained model from a paper should be able to run a prediction using a web service or binary file. Ideally, at least source code and the trained model should be available at a persistent URL and under a common license [148,149]. Also making the training code available is the ideal scenario, as this further increases the reproducibility of the paper and allows other researchers to build on the method without essentially having to start from scratch. Journals should bear some responsibility here to ensure that this becomes the norm.

Online links

Scikit-learn: https://scikit-learn.org/stable

Caret: https://topepo.github.io/caret

MLJ: https://alan-turing-institute.github.io/MLJ.jl/stable

PyTorch Geometric: https://pytorch-geometric.readthedocs.io/en/latest

Graph Nets: https://github.com/deepmind/graph_nets

PyTorch: https://pytorch.org

Tensorflow: https://www.tensorflow.org

Colaboratory: https://research.google.com/colaboratory

Glossary

Deep learning. Machine learning methods based on neural networks. The adjective "deep" refers to the use of many hidden layers in the network, two hidden layers as a minimum but usually many more than that. Deep learning is a subset of machine learning, and hence of artificial intelligence more broadly.

Artificial neural network. A collection of connected nodes loosely representing neuron connectivity in a biological brain. Each node is part of a layer and represents a number calculated from the previous layer. The connections, or edges, allow a signal to flow from the input layer to the output layer via hidden layers.

Ground truth. The true value that the output of a machine learning model is compared to in order to train the model and test performance. This data usually comes from experimental data (e.g. accessibility of a region of DNA to transcription factors) or expert human annotation (e.g. healthy or pathological medical image).

Encoding. Any scheme for numerically representing (often categorical) data in a form suitable for use in a machine learning model. An encoding can be a fixed numerical representation, e.g. one-hot or continuous encoding, or can be defined using parameters that are trained along with the rest of a model.

One-hot encoding. An encoding scheme that represents a fixed set of *n* categorical inputs using *n* unique *n*-dimensional vectors, each with one element set to 1 and the rest to 0. For example, the set of three letters (A,B,C) could be represented by the three vectors [1,0,0], [0,1,0] and [0,0,1], respectively.

Mean squared error. A loss function which calculates the average squared difference between the predicted values and the ground truth. This function heavily penalises outliers because it increases rapidly as the difference between a predicted value and the ground truth grows.

Binary cross entropy. The most common loss function for training a binary classifier — that is for tasks aimed to answer a question with only two choices (such as cancer versus non-cancer); sometimes called log loss.

Linear regression. A model that assumes that the output can be calculated from a linear combination of inputs, i.e. each input feature is multiplied by a single parameter and these values are added. It is easy to interpret how these models make their predictions.

Kernel functions. Refers to transformations applied to each data point to map the original points into a space in which they become separable with respect to their class.

Non-linear regression. A model where the output is calculated from a non-linear combination of inputs, i.e. the input features can be combined during prediction using operations such as multiplication. These models can describe more complex phenomena than linear regression.

k-nearest neighbours. A classification approach where a data point is classified based on the known (ground-truth) classes of the *k* most similar points in the training set, using a majority voting rule. *k* is a parameter that can be tuned. Can also be used for regression by averaging the property value over the k nearest neighbours.

Regularisation. Restricting the values of parameters to prevent the model from overfitting to the training data. For example, penalising high parameter values in regression models reduces the flexibility of the model and can stop it fitting to noise in the training data.

Cloud computing. On-demand computing services, including processing power and data storage, typically available via the internet. A pay-as-you-go model is usually used. Use of cloud

computing minimises up-front IT infrastructure costs.

Hidden Markov model. A statistical model that can be used to describe the evolution of observable events that depend on factors that are not directly observable. They have various uses in biology including representing protein sequence families.

Saliency map. In the context of machine learning, a saliency map is an image generated to show which pixels in an input image contribute to the prediction made by a model. They are useful in interpreting models.

Automatic differentiation. A set of techniques to automatically calculate the gradient of a function in a computer program. Used to train neural networks where it is called backpropagation.

Gradients. A gradient is the rate of change of one property as another property changes. In neural networks the set of gradients of the loss function with respect to the neural network parameters, computed via a process known as backpropagation, is used to adjust the parameters and thus train the model.

eTOC

Machine learning is becoming a widely used tool for the analysis of biological data. However, for experimentalists, proper use of machine learning methods can be challenging. This article overviews machine learning techniques and provides guidance on their applications in biology.

References

1. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15.

   A thorough review of applications of deep learning to biology and medicine including many references to the literature.

2. Mitchell TM. Machine Learning. McGraw Hill; 1997.

3. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

4. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16: 321–332.

5. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51: 12–18.

6. Myszczynska MA, Ojamies PN, Lacoste AMB, Neil D, Saffari A, Mead R, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Rev Neurol. 2020;16: 440–456.

7. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. Nature

Methods. 2019;16: 687–694.

8. Tarca AL, Carey VJ, Chen X-W, Romero R, Drăghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007;3: e116.

   An introduction to machine learning concepts and applications in biology with a focus on traditional machine learning methods.

9. Silva JCF, Teixeira RM, Silva FF, Brommonschenkel SH, Fontes EPB. Machine learning approaches and their current application in plant molecular biology: A systematic review. Plant Sci. 2019;284: 37–47.

10. Kandoi G, Acencio ML, Lemke N. Prediction of Druggable Proteins Using Machine Learning and Systems Biology: A Mini-Review. Front Physiol. 2015;6: 366.

11. Marblestone AH, Wayne G, Kording KP. Toward an Integration of Deep Learning and Neuroscience. Front Comput Neurosci. 2016;10: 94.

12. Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence. 2020;2: 573–584.

13. Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res. 2019;47: W402–W407.

14. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26: 990–999.

15. Altman N, Krzywinski M. Clustering. Nat Methods. 2017;14: 545–546.

16. Hopf TA, Ingraham JB, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35: 128–135.

17. Zhang Z, Wang L, Gao Y, Zhang J, Zhenirovskyy M, Alexov E. Predicting folding free energy changes upon single point mutations. Bioinformatics. 2012;28: 664–671.

18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

19. Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;28: 1–26.

20. Blaom AD, Kiraly F, Lienart T, Simillides Y, Arenas D, Vollmer SJ. MLJ: A Julia package for composable machine learning. Journal of Open Source Software. 2020;5: 2704.

21. Jones DT. Setting the standards for machine learning in biology. Nat Rev Mol Cell Biol. 2019;20: 659–660.

22. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33: 831–838.

23. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577: 706–710.

    Technology company DeepMind entered the CASP13 assessment in protein structure prediction and their method using deep learning was the most accurate of the methods entered.

24. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542: 115–118.

25. Tegunov D, Cramer P. Real-time cryo-electron microscopy data preprocessing with Warp. Nat Methods.

2019;16: 1146–1152.

26. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521: 436–444.

    A review of deep learning by some of the major figures in the deep learning revolution.

27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media; 2009.

28. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity Checks for Saliency Maps. NeurIPS. 2018. Available: https://arxiv.org/abs/1810.03292

29. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. PMLR. 2016;48: 1050–1059.

30. Smith AM, Walsh JR, Long J, Davis CB, Henstock P, Hodge MR, et al. Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. BMC Bioinformatics. 2020;21: 119.

31. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B. 1996;58: 267–288.

32. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B. 2005;67: 301–320.

33. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24: 1565–1567.

34. Ben-Hur A, Weston J. A User's Guide to Support Vector Machines. Methods in Molecular Biology. 2010; 223–239.

35. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support Vector Machines and Kernels for Computational Biology. PLoS Computational Biology. 2008;4: e1000173.

    An introduction to support vector machines with a focus on biological data and prediction tasks.

36. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014;46: 310–315.

37. Driscoll MK, Welf ES, Jamieson AR, Dean KM, Isogai T, Fiolka R, et al. Robust and automated detection of subcellular morphological motifs in 3D microscopy images. Nat Methods. 2019;16: 1037–1044.

38. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. Nat Methods. 2018;15: 5–6.

39. Wang C, Zhang Y. Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. J Comput Chem. 2017;38: 169–177.

40. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genomics. 2018;19: 84.

41. Olson RS, Cava WL, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. Pac Symp Biocomput. 2018;23: 192–203.

42. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res. 2019;47: 1044.

43. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028.

44. Jain AK. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2010;31: 651–666.

45. Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996;96: 226–231.

46. Nguyen LH, Holmes S. Ten quick tips for effective dimensionality reduction. PLoS Comput Biol. 2019;15: e1006907.

47. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019;37: 1482–1492.

48. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9: 2579–2605.

49. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. Nat Commun. 2019;10: 5416.

    Discussion and tips for using t-SNE as a dimensionality reduction technique on single-cell transcriptomics data.

50. Crick F. The recent excitement about neural networks. Nature. 1989;337: 129–132.

51. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020;2: 665–673.

    A discussion of a common problem in deep learning called shortcut learning, where the model uses decision rules that do not transfer to real-world data.

52. Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol. 1988;202: 865–884.

53. deFigueiredo RJ, Shankle WR, Maccato A, Dick MB, Mundkur P, Mena I, et al. Neural-network-based classification of cognitively normal, demented, Alzheimer disease and vascular dementia from single photon emission with computed tomography image data from brain. Proc Natl Acad Sci U S A. 1995;92: 5530–5534.

54. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Frontiers in Environmental Science. 2016;3.

55. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci U S A. 2020;117: 1496–1503.

56. Xu J, Mcpartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. bioRxiv. 2020. Available: https://www.biorxiv.org/content/10.1101/2020.10.12.336859v1

57. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36: 983–987.

58. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. Nat Methods. 2020;17: 1111–1117.

59. Zeng H, Edwards MD, Liu G, Gifford DK. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics. 2016;32: i121–i127.

60. Yao R, Qian J, Huang Q. Deep-learning with synthetic data enables automated picking of cryo-EM particle images of biological macromolecules. Bioinformatics. 2020;36: 1252–1259.

61. Si D, Moritz SA, Pfab J, Hou J, Cao R, Wang L, et al. Deep Learning to Predict Protein Backbone Structure from High-Resolution Cryo-EM Density Maps. Sci Rep. 2020;10: 4282.

62. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2: 158–164.

63. AlQuraishi M. End-to-End Differentiable Learning of Protein Structure. Cell Syst. 2019;8: 292–301.e3.

64. Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics. 2017;33: 2842–2849.

65. Müller AT, Hiss JA, Schneider G. Recurrent Neural Network Model for Constructive Peptide Design. J Chem Inf Model. 2018;58: 472–479.

66. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. JMLR Workshop Conf Proc. 2016;56: 301–318.

67. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44: e107.

68. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16: 1315–1322.

69. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. arXiv. 2017;1706.03762.

70. Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, et al. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. bioRxiv. 2020. Available: https://www.biorxiv.org/content/10.1101/2020.07.12.199554v1

71. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Tunyasuvunakool K, et al. High Accuracy Protein Structure Prediction Using Deep Learning. Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book). 2020.

72. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez A, Zambaldi V, Malinowski M, et al. Relational inductive biases, deep learning, and graph networks. arXiv. 2018;1806.01261. Available: https://arxiv.org/abs/1806.01261

73. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 2020;181: 475–483.

    A deep learning model predicts antibiotic activity with one candidate showing broad-spectrum antibiotic activities in mice.

74. Gainza P, Sverrisson F, Monti F, Rodolà E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods. 2020;17: 184–192.

75. Strokach A, Becerra D, Corbi-Verge C, Perez-Riba A, Kim PM. Fast and Flexible Protein Design Using Deep Graph Neural Networks. Cell Syst. 2020;11: 402–411.e4.

76. Gligorijevic V, Douglas Renfrew P, Kosciolek T, Leman JK, Cho K, Vatanen T, et al. Structure-Based Function Prediction using Graph Convolutional Networks. bioRxiv. 2019. Available: https://www.biorxiv.org/content/10.1101/786236v1

77. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34: i457–i466.

78. Veselkov K, Gonzalez G, Aljifri S, Galea D, Mirnezami R, Youssef J, et al. HyperFoods: Machine intelligent mapping of cancer-beating molecules in foods. Sci Rep. 2019;9: 9237.

79. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. ICLR. 2019. Available: https://arxiv.org/abs/1903.02428

80. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. 2019;37: 1038–1040.

81. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo Y-Y, et al. Predicting DNA Methylation State of CpG Dinucleotide Using Genome Topological Features and Deep Networks. Scientific Reports. 2016;6: 19598.

82. Linder J, Bogard N, Rosenberg AB, Seelig G. A Generative Neural Network for Maximizing Fitness and Diversity of Synthetic DNA and Protein Sequences. Cell Syst. 2020;11: 49–62.e16.

83. Greener JG, Moffat L, Jones DT. Design of metalloproteins and novel protein folds using variational autoencoders. Sci Rep. 2018;8: 16189.

84. Wang J, Ma A, Chang Y, Gong J, Jiang Y, Qi R, et al. scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. Nat Commun. 2021;12: 1882.

85. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. Tensorflow: A system for large-scale machine learning. 12th USENIX Symposium on Operating Systems Design and Implementation. 2016; 265–283.

86. Wei Q, Dunbrack RL Jr. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One. 2013;8: e67863.

87. Walsh I, Pollastri G, Tosatto SCE. Correct machine learning on protein sequences: a peer-reviewing perspective. Brief Bioinform. 2016;17: 831–840.

    A discussion of how peer reviewers can assess machine learning methods in biology, and by extension how scientists can design and carry out such studies properly.

88. Schreiber J, Singh R, Bilmes J, Noble WS. A pitfall for machine learning methods aiming to predict across cell types. Genome Biology. 2020;21.

89. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. The EMBO Journal. 1986;5: 823–826.

90. Söding J, Remmert M. Protein sequence comparison and fold recognition: progress and good-practice benchmarking. Curr Opin Struct Biol. 2011;21: 404–411.

91. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20: 473.

92. Sillitoe I, Dawson N, Lewis TE, Das S, Lees JG, Ashford P, et al. CATH: expanding the horizons of structure-based functional annotations for genome sequences. Nucleic Acids Res. 2019;47: D280–D284.

93. Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, et al. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014;10: e1003926.

94. Li Y, Yang J. Structural and Sequence Similarity Makes a Significant Impact on Machine-Learning-Based Scoring Functions for Protein-Ligand Interactions. J Chem Inf Model. 2017;57: 1007–1012.

95. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med. 2018;15: e1002683.

96. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. ICLR. 2014. Available: https://arxiv.org/abs/1312.6199

97. Hie B, Cho H, Berger B. Realizing private and practical pharmacological collaboration. Science. 2018;362: 347–350.

98. Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing. Circ Cardiovasc Qual Outcomes. 2019;12: e005122.

99. Konečný J, Brendan McMahan H, Ramage D, Richtárik P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. arXiv. 2016;1610.02527. Available: https://arxiv.org/abs/1610.02527

100. Pérez A, Martínez-Rosell G, De Fabritiis G. Simulations meet machine learning in structural biology. Curr Opin Struct Biol. 2018;49: 139–144.

101. Noé F, Olsson S, Köhler J, Wu H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Science. 2019;365: 6457.

102. Shrikumar A, Greenside P, Kundaje A. Reverse-complement parameter sharing improves deep learning models for genomics. bioRxiv. 2017. Available: https://www.biorxiv.org/content/10.1101/103663v1

103. Lopez R, Gayoso A, Yosef N. Enhancing scientific discoveries in molecular biology with deep generative models. Mol Syst Biol. 2020;16: e9198.

104. Anishchenko I, Chidyausiku TM, Ovchinnikov S, Pellock SJ, Baker D. De novo protein design by deep network hallucination. bioRxiv. 2020. Available: https://www.biorxiv.org/content/10.1101/2020.07.22.211482v1

105. Innes M, Edelman A, Fischer K, Rackauckas C, Saba E, Shah VB, et al. A differentiable programming system to bridge machine learning and scientific computing. arXiv. 2019;1907.07587. Available: https://arxiv.org/abs/1907.07587

106. Ingraham J, Riesselman AJ, Sander C, Marks DS. Learning Protein Structure with a Differentiable Simulator. ICLR. 2019. Available: https://openreview.net/forum?id=Byg3y3C9Km

107. Jumper JM, Faruk NF, Freed KF, Sosnick TR. Trajectory-based training enables protein simulations with accurate folding and Boltzmann ensembles in cpu-hours. PLoS Comput Biol. 2018;14: e1006578.

108. Wang Y, Fass J, Chodera JD. End-to-End Differentiable Molecular Mechanics Force Field Construction. arXiv. 2020;2010.01196. Available: http://arxiv.org/abs/2010.01196

109. Bradbury J, Frostig R, Hawkins P, Johnson MJ, Leary C, Maclaurin D, et al. JAX: composable transformations of Python+NumPy programs. 2018. Available: http://github.com/google/jax

110. Chen KM, Cofer EM, Zhou J, Troyanskaya OG. Selene: a PyTorch-based deep learning library for sequence data. Nat Methods. 2019;16: 315–318.

A software library based on PyTorch providing functionality for biological sequences.

111. Kopp W, Monti R, Tamburrini A, Ohler U, Akalin A. Deep learning for genomics using Janggu. Nat Commun. 2020;11: 3488.

112. Schoenholz SS, Cubuk ED. JAX, M.D.: End-to-End Differentiable, Hardware Accelerated, Molecular Dynamics in Pure Python. arXiv. 2019;1912.04232. Available: https://arxiv.org/abs/1912.04232

113. Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nature Biotechnol. 2019;37: 592–600.

114. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2020.

115. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 2020;16: e9380.

116. AlQuraishi M. ProteinNet: a standardized data set for machine learning of protein structure. BMC Bioinformatics. 2019;20: 311.

117. Townshend RJL, Vögele M, Suriana P, Derry A, Powers A, Laloudakis Y, et al. ATOM3D: Tasks On Molecules in Three Dimensions. arXiv. 2020;2012.04035.

118. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen P, Canny J, et al. Evaluating Protein Transfer Learning with TAPE. Adv Neural Inf Process Syst. 2019;33.

119. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics. 2019;87: 1011–1020.

120. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol. 2019;20: 244.

121. Munro D, Singh M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. Bioinformatics. 2020;36: 5322–5329.

122. Haario H, Taavitsainen V-M. Combining soft and hard modelling in chemical kinetic models. Chemometrics Intellig Lab Syst. 1998;44: 77–98.

123. Cozzetto D, Minneci F, Currant H, Jones DT. FFPred 3: feature-based function prediction for all Gene Ontology domains. Sci Rep. 2016;6: 31865.

124. Nugent T, Jones DT. Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics. 2009;10: 159.

125. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005;33: W480–2.

126. Li W, Yin Y, Quan X, Zhang H. Gene Expression Value Prediction Based on XGBoost Algorithm. Front Genet. 2019;10: 1077.

127. Zhang Y, Skolnick J. SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem. 2004;25: 865–871.

128. Teodoro ML, Phillips GN Jr, Kavraki LE. Understanding protein flexibility through dimensionality reduction. J Comput Biol. 2003;10: 617–634.

129. Schlichtkrull M, Kipf TN, Bloem P, van den Berg R, Titov I, Welling M. Modeling Relational Data with Graph Convolutional Networks. The Semantic Web. Springer International Publishing; 2018. pp. 593–607.

130. Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. Nat Methods. 2018;15: 805–815.

131. Antczak M, Michaelis M, Wass MN. Environmental conditions shape the nature of a minimal bacterial genome. Nat Commun. 2019;10: 3100.

132. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics. 2017;18: 277.

133. Hiranuma N, Park H, Baek M, Anishchenko I, Dauparas J, Baker D. Improved protein structure refinement guided by deep learning based accuracy estimation. Nat Commun. 2021;12: 1340.

134. Pagès G, Charmettant B, Grudinin S. Protein model quality assessment using 3D oriented convolutional neural networks. Bioinformatics. 2019;35: 3313–3319.

135. Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Research. 2014;42: W314–W319.

136. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. Proc Natl Acad Sci U S A. 2019;116: 27151–27158.

137. Chen L, Cai C, Chen V, Lu X. Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model. BMC Bioinformatics. 2016;17.

138. Kantz ED, Tiwari S, Watrous JD, Cheng S, Jain M. Deep Neural Networks for Classification of LC-MS Spectral Peaks. Anal Chem. 2019;91: 12407–12413.

139. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods. 2019;16: 299–302.

140. Liebal UW, Phan ANT, Sudhakar M, Raman K, Blank LM. Machine Learning Applications for Mass Spectrometry-Based Metabolomics. Metabolites. 2020;10: 243.

141. Zhong ED, Bepler T, Berger B, Davis JH. CryoDRGN: reconstruction of heterogeneous cryo-EM structures using neural networks. Nat Methods. 2021;18: 176–185.

142. Schmauch B, Romagnoni A, Pronier E, Saillard C, Maillé P, Calderaro J, et al. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. Nat Commun. 2020;11: 3877.

143. Das P, Sercu T, Wadhawan K, Padhi I, Gehrmann S, Cipcigan F, et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nat Biomed Eng. 2021.

144. Gligorijevic V, Barot M, Bonneau R. deepNF: deep network fusion for protein function prediction. Bioinformatics. 2018;34: 3873–3881.

145. Karpathy A. A Recipe for Training Neural Networks. 2019. Available: https://karpathy.github.io/2019/04/25/recipe

146. Bengio Y. Practical Recommendations for Gradient-Based Training of Deep Architectures. Lecture Notes in Computer Science. 2012; 437–478.

147. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nature Machine Intelligence. 2021;3: 199–217.

This study assessed 62 machine learning studies that analyse medical images for COVID-19 and none are found to be of clinical use, indicating the difficulties of training a useful model.

148. List M, Ebert P, Albrecht F. Ten Simple Rules for Developing Usable Software in Computational Biology. PLoS Comput Biol. 2017;13: e1005265.

149. Sonnenburg SÃ, Braun ML, Ong CS, Bengio S. The need for open source software in machine learning. Journal of Machine Learning Research. 2007;8: 2443–2466.