OXFORD FEMS

FEMS Microbiology Reviews, 1–20

DOI: 10.1093/femsre/fuad003

Advance access publication date: 1 February 2023

Review Article

Elucidating the functional roles of prokaryotic proteins using big data and artificial intelligence

Zachary Ardern[1,2,†], Sagarika Chakraborty[1,†], Florian Lenk[1], Anne-Kristin Kaster[1,*]

[1]Institute for Biological Interfaces 5 (Institut für Biologische Grenzflächen IBG 5), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany

[2]Wellcome Trust Sanger Institute, Hinxton, Saffron Walden CB10 1RQ, United Kingdom

*__Corresponding author.__ Institute for Biological Interfaces 5 (Institut für Biologische Grenzflächen IBG 5), Karlsruhe Institute of Technology (KIT), Building 601, 76344 Eggenstein-Leopoldshafen, Germany. Tel: +49 721 608 23005; E-mail: kaster@kit.edu

Editor: [Ming Hammond]

[†]Contributed equally

Abstract

Annotating protein sequences according to their biological functions is one of the key steps in understanding microbial diversity, metabolic potentials, and evolutionary histories. However, even in the best-studied prokaryotic genomes, not all proteins can be characterized by classical *in vivo*, *in vitro*, and/or *in silico* methods—a challenge rapidly growing alongside the advent of next-generation sequencing technologies and their enormous extension of ‘omics’ data in public databases. These so-called hypothetical proteins (HPs) represent a huge knowledge gap and hidden potential for biotechnological applications. Opportunities for leveraging the available ‘Big Data’ have recently proliferated with the use of artificial intelligence (AI). Here, we review the aims and methods of protein annotation and explain the different principles behind machine and deep learning algorithms including recent research examples, in order to assist both biologists wishing to apply AI tools in developing comprehensive genome annotations and computer scientists who want to contribute to this leading edge of biological research.

Keywords: hypothetical proteins, annotation, omics data, machine learning, deep learning, metadata, databases

Introduction

Bacteria and archaea are the oldest, most abundant, and most diverse forms of life on Earth (Eme and Doolittle 2015, Louca et al. 2019). They dominate many functions of the biosphere and harbour a huge potential for biotechnological applications (Singh et al. 2020, Pfeifer et al. 2021). However, the task of fully characterizing microbial diversity is almost incomprehensibly vast. The ‘known unknowns’ (Logan 2009), i.e. the diversity we know is there, but which we have not characterized yet, have become increasingly apparent in the recent years, while the extent of ‘unknown unknowns’, i.e. the diversity which remains completely undiscovered, is still debated. With some approximates suggesting billions or even trillions (Locey and Lennon 2016, Larsen et al. 2017), another study estimated 0.8–1.6 million prokaryotic species (Louca et al. 2019) on Earth. Of those, only ∼2% have whole or at least partial genome sequences (Zhang et al. 2020), many with only the prokaryotic marker gene 16S rRNA known (Hugenholtz et al. 2021). Approximately, 70% belong to the so-called candidate phyla radiation (CPR), which—with very few exceptions—have no cultured representatives (Hug et al. 2016). At this point it is, therefore, it is safe to say that ∼99% of all microbial species from the environment remain uncharacterized. They are, therefore, referred to as *Microbial Dark Matter* (Bernard et al. 2018).

Strikingly, even todays’ best-studied microorganisms have not been fully functionally characterized yet. For instance, in the ‘favorite pet’ of microbiologists for over 100 years—the model organism *Escherichia coli*—the function of more than 30% of proteins has not been determined experimentally and more than 2% of protein-encoding genes have no characterization at all (Ghatak et al. 2019). These so-called hypothetical proteins (HPs), which have been referred to as ‘functional dark matter’ (Escudeiro et al. 2022), are found across all microbial species and represent an enormous gap in knowledge. In addition to their importance for the fundamental understanding of biology and evolution, these proteins might also provide novel solutions for medical treatments, bioremediation, or bioenergy production, to help solve 21st century challenges (Rehman et al. 2021, Arslan et al. 2022). However, recent analysis suggests that only 3% of the biochemical potential of bacterial genomes has been discovered (Gavriilidou et al. 2021). Genomes of taxa with no cultured representative, archaea and relatively large bacteria with a complex lifestyle, have a particularly high percentage of HPs (Makarova et al. 2011, Lobb et al. 2020) (Fig. 1). While some uncharacterized genes are conserved across species or genera, many are taxonomically restricted (Yu and Stoltzfus 2012, Tatarinova et al. 2016). Hence, since the majority of microorganisms, proteins, and products are still uncharacterized, their use for future applications is heavily constrained in comparison to what is possible (Kalkreuter et al. 2020).

This biological problem can now be addressed with the help of AI-based tools, using the large quantities of biological data avail-

Received: August 12, 2022. **Revised:** January 11, 2023. **Accepted:** January 31, 2023

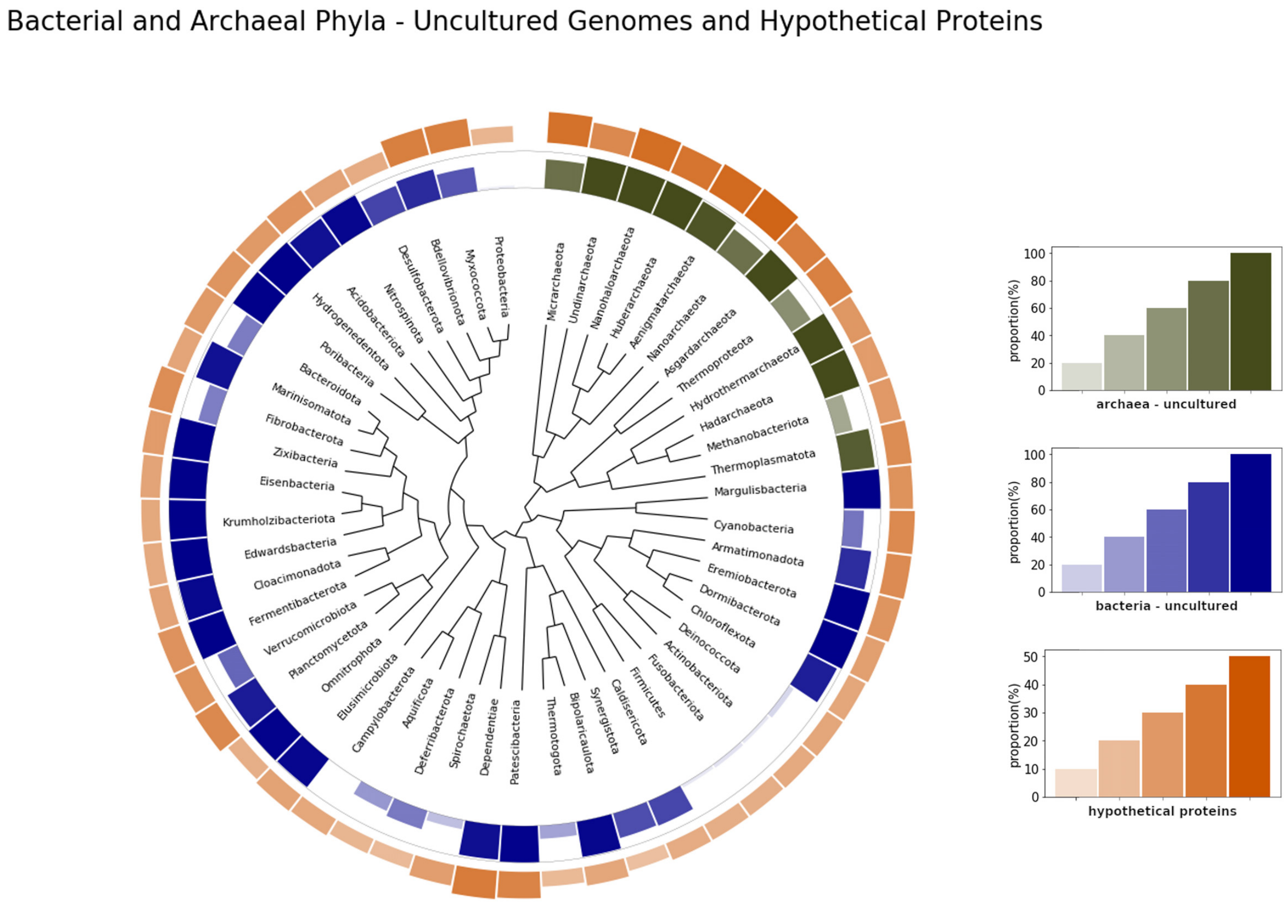

Figure 1. Continued.

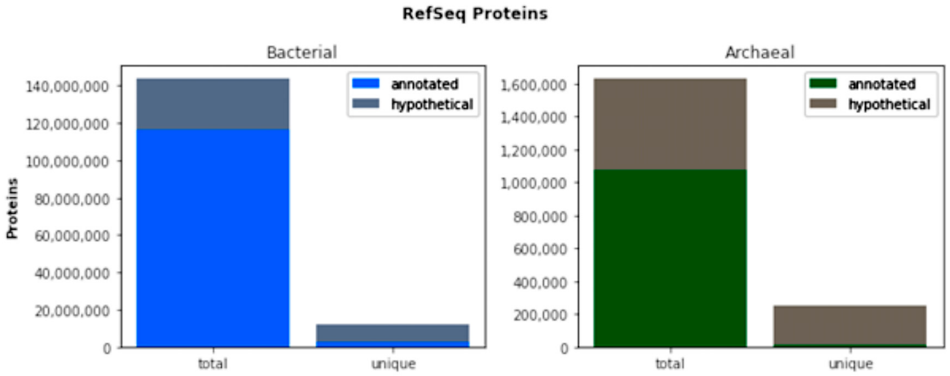

Figure 1. (A) Circular tree of archaeal and bacterial phyla, showing the percentage of uncultured genomes in the Genome Taxonomy Database (GTDB) (archaea = green, bacteria = blue) (Parks et al. 2022) and median percentage of HPs according to AnnoTree data (orange) (Mendler et al. 2019) with at least 10 (bacteria) or five genomes (archaea); phyla were renamed according to the List of Prokaryotic names with Standing in Nomenclature (LSPN) and visualized with GraPhlAn (Asnicar et al. 2015). (B) Numbers of characterized and uncharacterized (hypothetical) proteins in bacteria and archaea. Numbers after clustering to an identity of <50% are 'unique' proteins. Among unique proteins HPs include more sequence diversity than annotated proteins.

able in public databases from high-throughput experiments. A number of reviews on AI in relation to protein function prediction, genomics, or biology more generally, have recently been published (Greener et al. 2022, Whalen et al. 2022), however, all have focused on human proteins or eukaryotic model organisms (Bonetta and Valentino 2020, Mahood et al. 2020, Ofer et al. 2021). Our review particularly concentrates on the functions of proteins in bacteria and archaea, the data and algorithms available, and the difficult conceptual issues underlying the task of predicting protein function.

Functional analysis and categorization of proteins

Annotating protein-encoding genes is the process of assigning functional labels to protein products. This can be approached from different biological perspectives and the best approach to use is often simply taken for granted according to the norms of a biological subdiscipline. The concept of function is, however, multifaceted, and each aspect of function has become associated with particular experimental methods, whether *in vivo*, *in vitro*, *in silico*, or in combination. While different approaches can be complementary and integrated, they are distinct and need to be understood.

Biochemical and biophysical phenotypes

The classic microbiological account of function pertains to the phenotype of a gene. Here, researchers can either study what happens if the gene is taken away from or added to the genome, at the level of DNA or the level of protein expression. In addition, one can also study the properties of a protein in isolation. This is a subset of the causal role view, assessing the difference that the presence or absence of the protein makes to the operation of the cell, as measured for instance by population growth under different conditions, or in terms of biochemical reactions in which the protein is involved. In early molecular genetics research, gene knockouts were often achieved by transposon mutagenesis, leading to distinct phenotypes (Handfield and Levesque 1999) and multiple high throughput versions of this technique have now been developed (Cain et al. 2020). Other new technologies include CRISPR gene editing (Liu et al. 2020) and gene silencing with CRISPR interference (Zhang et al. 2021). Other traditional approaches for characterizing proteins include 2D electrophoresis for separating proteins by mass and isoelectric point, enzyme assays to determine catalytic abilities, and analysis of the 3D structure using X-rays or electron microscopy, which can help to clarify mechanisms of action (Aslam et al. 2017). These classical methods are often experimentally challenging and very time-consuming since they require bacterial cultures with sufficient biomass or means for overexpression of the protein of interest. This can be complicated or even impossible due to the aforementioned problem of not having cultures available for the majority of microbial phylotypes. New technologies using membrane diffusion, cell-sorting, or microfluidics and the use of cocultures to cultivate interacting microbes in combination are trying to circumvent some of the problems (Lewis et al. 2021). There are also ongoing developments in laboratory assays, e.g. using a high throughput system to assess the kinetic effects of more than 1000 mutations in an esterase enzyme (Markin et al. 2021).

Gene expression and regulation

Another 'causal role' approach to function, is determining under what conditions a gene is expressed. This approach assumes that regulation in relation to an environmental condition or stimulus implies a functional role in that condition. Condition-dependent protein expression and interactions between proteins capture different dimensions of protein function (Morcinek-Orłowska et al. 2021). The transcriptional regulatory network to a large extent determines the gene expression responses to environmental and cellular conditions. There are multiple levels of control of expression, including post-transcriptional and post-translational factors. In general, correlations between expression of genes indicate coregulation, which may imply a functional relationship (Junier and Rivoire 2016). The genes expressed under different environmental conditions are largely determined by transcription factor binding sites. These sites can be discovered either through experimental detection of binding or computational comparisons for the specific motifs associated with particular transcription factors (Gao et al. 2018). Direct interactions between proteins similarly imply a functional relationship, with the underlying implicit assumption that interactions are specific, as nonspecific interactions tend to be toxic (Bhattacharyya et al. 2016). Protein–protein interactions can be detected with methods such as affinity purification mass spectrometry (Morris et al. 2014).

Evolution

Evolutionary considerations have a complex relationship with function. Some have proposed that to be functional simply is to be (or have recently been) a subject of natural selection (Neander 1991, Graur et al. 2013). Evolutionary biologists are interested in questions regarding historical or population dynamics processes. This gives an account of what is functional and what is not if one can detect the effects of selection; however, given our general lack of access to historical evolutionary forces, it is of little use determining what a biological entity's particular function was or is. More relevant to assigning a specific function is protein conservation across species, where it is typically assumed that the function is likewise conserved. Technically, homology—as originally used by Richard Owen (Cooper and Owen 1843), and in evolutionary theory since—refers only to a biological structure, such as a protein sequence, but it is now widely applied to functions as well (Love 2007). A related source of microbial functional insights lies in pangenomes—i.e. the patterns of gene diversity across strains within a species, genus, or other monophyletic clade (Golicz et al. 2020). This approach overlaps with the classic phenotype method—again gene presence/absence is used to infer function, but *via* presence/absence across natural strains rather than genetic manipulation in the laboratory. This kind of analysis presumes that patterns of evolution in functionally related versus functionally unrelated genes will differ. The extent to which this is the case will depend on the role of natural selection in shaping the pangenome. This topic is controversial, but there have been many cases where functional relationships between genes have been shown to leave genomic footprints (Chen et al. 2013). Other approaches integrate some of these diverse information sources to find functional clusters of genes, for instance those relating to specific metabolic processes (Psomopoulos et al. 2020).

Categorization labels

The functional concepts described above do not, however, map directly onto useful labels or gene categories for scientists to work with. Given that the concept of function is contentious and plural, many kinds of categories or labels can be assigned to proteins, each with various advantages and disadvantages, based on information from diverse sources including high throughput gene knockout or silencing, biochemical assays or biophysical studies, 'omics' expression studies, and/or evolutionary analyses. Perhaps the most popular hierarchical system for gene function annotation with a controlled vocabulary is the use of 'gene ontology' (GO) terms (Ashburner et al. 2000). The GO terms comprise three different classification systems: cellular components (e.g. an organelle), molecular functions (e.g. a particular enzyme activity), and biological processes (e.g. mismatch repair). Methods for automated function prediction using GO terms have recently been reviewed (Makrodimitris et al. 2020). Another widely used comprehensive hierarchical system is the system of enzyme commission (EC) numbers, a numerical classification scheme for enzymes with biochemical evidence (Bairoch 2000), which classify proteins into seven major types of enzymes and 6646 entries of different catalyzed reactions (as of March 2022) (Gasteiger et al. 2003). These EC numbers are very useful for well-understood enzyme families and other genes with characterized homologs. Every EC number is associated with a recommended name for the respective enzyme. If different enzymes (for instance from different organisms) catalyze the same reaction, then they receive the same EC number (Fleischmann et al. 2004). Furthermore, through convergent evolution, completely different enzymes can catalyze an identical reaction and, therefore, would be assigned an identical EC number (Omelchenko et al. 2010). Protein structures are grouped into families or folds in databases such as Pfam, SCOP, and CATH (Fox et al. 2015). Typically, members of the same protein family perform similar functions, but some ancient families or super families have significantly diverged in sequence and/or function (Jaroszewski et al. 2009).

Standard bioinformatics methods for function prediction

Many *in silico* methods have been developed for functional prediction of proteins using bioinformatics tools for classification and annotation. However, they are not high-throughput and require quite extensive computing power and time, can often only find what is already known, and do not make use of the full metadata available, such as linking gene expression data with environmental parameters. In addition, the predicted functions are usually not experimentally verified, raising the possibility of untrue annotations.

Sequence-based annotations

Comparative genomics, based on evolutionary theory, allows for the propagation of annotations across genomes (König et al. 2018). The use of 'BLAST' (Basic Local Alignment Sequence Tool) (Altschul et al. 1990) is nearly ubiquitous in biology, and for some synonymous with bioinformatics. This approach is so powerful because only a few model organisms have been probed in depth in laboratory studies, but many proteins are conserved across diverse taxonomic groups. Homology inference is applied in many annotation tools (Mahlich et al. 2018), which can also be used to identify associations between protein domains and functions (Rojano et al. 2022). However, even when a conserved domain is predicted, the function itself may not be conserved (Punta and Ofran 2008). On the other hand, proteins with different sequences/domains might be able to catalyze the same chemical reaction. In addition, characterized domains are often simply not found in HPs (Goodacre et al. 2013). Further, it is a known problem that many annotations in databases are simply wrong or suffer from 'over'-annotation (Moreno-Hagelsieb and Hudy-Yuffa 2014). Any errors in the initial assignment of function is, therefore, propagated outwards across genomes with no 'proofreading' if only the criterion of homology is used.

Structure-based annotations

Similarly, a protein's structure may be highly similar even when there is no trace of higher-than-chance similarity at the sequence level (Rost 1999). The function of a putative protein can be predicted by so called 'homology modelling' (David and Andrej 2001), in which the protein has to align with a known protein sequence whose 3D structure is known or using protein signatures, which classify proteins into families and domains based on sequence models such as hidden Markov models, with various confidence levels (Zohra Smaili et al. 2021). Further approaches include determination of protein 3D structure by structural genomics initiatives, understanding the nature and mode of prosthetic groups or metal ion binding, fold similarity with other proteins of known functions and annotating possible catalytic and regulatory sites (Myers et al. 2015). Another promising approach is structure prediction followed by biochemical function assessment by *in silico* screening for various substrates (Mills et al. 2015).

Transcription-based annotations

The transcriptional network within which a gene is located is informative of its function within the cell. These networks are currently only available in databases for a few model organisms, but in principle can now be predicted for a wider range of organisms, using high throughput data on RNA expression and transcription factor binding. The main functional units here are the operon and the regulon. The operon has been classically thought of as a cluster of colocated genes on the same strand controlled by a regulatory region (Jacob and Monod 1961), which may include positive or negative regulation. There are many additional complexities that have since been discovered, for instance recently an example of a noncontiguous operon has been proposed, where genes situated in antisense to each other are coregulated (Sáenz-Lahoya et al. 2019). Functional groupings of genes (operons and higher-level groupings) have been inferred from correlations between the expression levels of different genes in bacterial populations grown under diverse environmental conditions (Chen et al. 2018). Multiple operons can together be coregulated, and together grouped into a regulon, which can coceivably be inferred from either transcriptomic or pangenomic data. Because operons and regulons are relatively discrete functional units (Sastry et al. 2019), they may avoid some of the common pitfalls which other approaches such as GO classes face due to their complex hierarchical relations (Gaudet and Dessimoz 2017).

Pangenome-based annotations

New evolutionary or comparative genomic analyses have also been developed on genome data available across related strains and species. Differing gene content across strains mean that there is a large 'pangenome' for each species (Brockhurst et al. 2019). Studying patterns of copresence and coabsence allows for inferring functional networks of genes, which have been termed 'components' in this kind of analysis (Hall et al. 2021). Information on gene–gene relationships and combining evolutionary information with transcriptomic/translatomic data has the potential to greatly increase our knowledge of regulatory networks and gene functions.

The development of next-generation and third-generation 'long read' sequencing technologies has resulted in huge amounts of data (Fig. 1), and currently almost 200 million HPs are listed in public databases. The aforementioned *in vivo*, *in vitro*, and/or *in silico* methods of assigning functions to sequences are thus no longer able to catch up with the exponentially growing number of sequencing data and, hence, the number of HPs, leaving an enormous potential for biotechnological application uncovered (Ijaq et al. 2015). The amount of information nowadays available in public databases, however, creates opportunities for new 'Big Data'-based approaches beyond traditional analyses.

The era of Big Data

'Omics' datasets have become one of the main examples of 'Big Data' in the last few years. In general, the characteristics of Big Data can be summarized by the 6 Vs: Value, Volume, Velocity, Variety, Veracity, and Variability (Emmanuel and Stanier 2016). Aside from the possibility to obtain genomes, transcriptomes and proteomes from individual species in cultures or single cells (Dam et al. 2020, Kaster and Sobol 2020, Wiegand et al. 2021) shotgun sequencing approaches such as metagenomics and metatranscriptomics have now become the methods of choice for studying microbes from various habitats and researching phylogenetic groups that currently lack cultured representatives (Aguiar-Pulido et al. 2016). Shotgun proteomics can also be applied to whole microbial communities for metaproteomics (Karaduta et al. 2021), and the translated portion of RNA has begun to be examined *via* metatranslatomics (Fremin et al. 2020). All this data could in theory be used to improve gene annotation, but it has been argued that the application of methods designed for smaller datasets to this new wealth of data has in fact reduced the quality of annotations (Salzberg 2019).

Application of Big Data in biology therefore requires integration and analysis of complex heterogeneous data, including metadata for each dataset (Subramanian et al. 2020). This demands cost-effective, innovative forms of information processing. The quality and reliability of data are important and vary significantly across datasets, which is why maintaining curated public databases is paramount. The most widely used database for depositing sequencing and metadata is the National Center for Biotechnology Information database (NCBI), which includes GenBank (Benson et al. 2018) and RefSeq (Li et al. 2021). In addition, there are databases from the European Bioinformatics Institute (EBI), including the ENA (European Nucleotide Archive) repository (Park et al. 2017), EggNOG (Huerta-Cepas et al. 2019), and Ensembl Genomes (Howe et al. 2021). Protein sequences are stored in databases such as InterPro (Blum et al. 2021), Pfam (hosted by the Sanger Institute) (Mistry et al. 2021), or UniProtKB/Swiss-Prot (hosted by the Swiss Institute of Bioinformatics) (Bateman et al. 2021). Metabolic pathway and protein function databases include KEGG (maintained by the University of Kyoto) (Kanehisa et al. 2017) or BioCyc (maintained by SRI International) (Karp et al. 2019). The journal Nucleic Acid Research regularly publishes special issues on biological databases and currently, there are over 1640 databases listed with different purposes, some being complementary (Rigden et al. 2021). It has been estimated that the amount of omics data will double every 9–12 months (Stephens et al. 2015). Large-scale comparisons of genomes and transcriptomes have already been used for diverse analyses such as biosynthetic gene clusters (Navarro-Muñoz et al. 2020), the relative expression of different mRNAs and noncoding RNAs (Ireland et al. 2020) or 'ribosome profiling' (Steizt 1969). CHiPseq (Furey 2012) can be used to discover transcription factor binding sites, and thereby bacterial regulons (Myers et al. 2015) and a modification of IPOD-HR gives an overview of all protein binding to DNA, giving a fuller picture of gene regulation and functional components (Freddolino et al. 2021).

The next step: using artificial intelligence to characterize the proteome

With the breakthrough of technological advances over the last decades, the 'educated guesses' that had been previously used for creating specific scientific hypotheses are rapidly being replaced by the knowledge provided through untargeted high-throughput methods. Artificial intelligence (AI) provides novel opportunities to use large quantities of high-quality data on another level with a wide range of rational processes, including reasoning, learning, decisions, language processing, and perception (Oliveira 2019). We can think of AI, machine learning (ML) and deep learning (DL) as a set of concentric circles where DL is a subset for ML and ML is a subset for AI.

The massive, complex, and rapid evolution of datasets as well as computational mathematics make it now possible for AI applications to learn, make intelligent decisions, and improve pattern recognition capabilities (Perakakis et al. 2018). Manual intervention in data management and analytics are still needed, but processes that might take days or weeks (or longer) or which were not humanly possible, are now quickly achieved (Serres et al. 2001) (Fig. 2). Table 1 provides a glossary of generally used terms regarding AI, to help guide the reader through the next sections.

Traditional classifiers

ML includes a wide range of methods, which can be divided into supervised, unsupervised, semisupervised, semiunsupervised, and reinforcement-based learning approaches (Alloghani et al. 2020). There are several algorithms that can be used for the task of categorization—assigning functional labels to putative genes, which have already been predicted to encode a protein (Fig. 3). The technique of classification has long been used for function prediction by identifying suitable features using feature engineering and generating numeric vectors to subsequently develop suitable models.

Support vector machine (SVM) (Vapnik 1995) a supervised learning approach, which performs binary classification, with linear or nonlinear functions. It establishes a maximum-margin hyperplane within the *n*-dimensional space of the data, which separates the data into two classes. The ML algorithm achieves this by determining an appropriate kernel function (e.g. linear, polynomial, or radial basis) (Kulkarni-Kale et al. 2014).

k-nearest-neighbor (KNN) (Altman 1992) is a supervised learning algorithm that tries to classify each data point by locating the nearest 'k' neighbors with known labels and subsequently assigns a class label, i.e. determined by a majority vote among neighbours. Traditional KNN-based methods are easy to use but involve higher computation times (Borah et al. 2020).

Decision tree (DT) (Quinlan 1986) is a branch-test-based classifier, supervised algorithm, which recursively partitions the data based on its attributes, until some stopping condition is reached. This recursive partitioning gives rise to a tree-like structure. The route undertaken for classification of data can be traced from the root node to each leaf node in the tree (Schietgat et al. 2010).

Random forest (RF) (Breiman 2001) uses an ensemble of DTs to obtain a majority vote on the correct classification. Classification trees are constructed by randomly selecting from training datasets. Results from each tree can then be gathered to give a prediction for each observation.

Some studies also mention regression-based protein function annotations, which are now used as ensemble methods along with other classifiers (You et al. 2018). Although ML has gained immense popularity over the last decade, DL methods are now increasingly being explored. Unlike traditional ML algorithms that require a lot of domain expertise and human intervention, DL algorithms automatically learn from raw input data, therefore, describing highly nonlinear and complex patterns more effectively (Lv et al. 2019). DL is widely associated with artificial neural network (ANN) architectures having numerous hidden layers for fea-

Table 1. A glossary of generally used terms regarding AI.

Term	Description
Activation function	A function which is applied at each **neuron** on its summed input and which determines its output to the next **layer**. Can be linear, i.e. multiplying the neuron's input by a constant factor, or nonlinear. Nonlinear **activation functions** give **neural networks** the ability to model complex problems
Categorization	Assigning **labels** to elements of a dataset (such as functional classes to protein sequences), where **classes** can be hierarchical, i.e. related to each other, and one element can belong to multiple classes (e.g. GO terms)
Class	A discrete value returned by a **classifier**, which is mutually exclusive from all the other values obtained for all the data points of a dataset (such as EC numbers for protein sequences)
Classification	Assigning **classes** to elements of a dataset, where classes are mutually exclusive, i.e. one element can only belong to one class
Classifier	An algorithm used for the **classification** of input data into specific **classes**
Clustering	Divides the elements of a dataset into several groups based on similarities in their **features**/attributes, without defining these groups *a priori*, i.e. learning without any **labels**
Collaborative filtering	Filtering for patterns of **features** across large datasets. Similarities in feature patterns across different data points can be used to make predictions ('filtering') about missing data
Convolutional layer	**Layer** type that gives **convolutional neural networks** (CNNs) their name; it extracts Feature from **input data**, such as DNA or protein sequences in their matrix form by applying several **filters**; these decompose the input into smaller matrices **(feature Maps**), which are passed to the next layer (e.g. a **pooling layer**)
DL	Use of artificial **neural networks** comprising multiple (more than three) layers of **neurons** to identify complex patterns over large datasets
Dimension	The number of **features/**attributes of a data point/dataset (e.g. the amino acid frequencies of a protein sequence would be a 20-dimensional vector, describing/containing 20 attributes)
Ensemble classification	**Classification** based on the predictions of several independent **classifiers**. Ensemble methods usually perform better than their individual constituent classifiers. The classifiers can have different algorithms, training sets and **hyperparameters** and are all independently trained. Two methods by which a Classification may be reached by the ensemble are **unweighted** and **weighted voting.**
Feature	A single representation (typically numeric) that combines/summarizes multiple attributes/metrics of a data point, e.g. combining the frequencies of R (Arginine), H (Histidine), and K (Lysine) (three attributes) in a protein into the frequency of positively charged amino acids (one feature), or computing a **feature map** on an input matrix in a **convolutional neural network**
Feature extraction	Defining a set of **features** for a dataset (usually done by dedicated algorithms), each of which combines several attributes into one numerical representation. Effectively reduces the dimensionality of the dataset and thus speeds up computations
Feature map	The activation pattern of **neurons** for a particular **filter** across the input matrix in a **convolutional layer.**
Feature selection	Manually or algorithmically selecting a subset of **features** that best represents the structure of the dataset. Further reduces the number of **dimensions** of the dataset and thus also computational complexity
Filter	Filters define **features** in **convolutional neural networks**. Each one is a small matrix of **weights,** which is slid across the input matrix and at each position may or may not activate a **neuron**. This results in a smaller matrix of activations, called a **feature map**. Effectively, each **filter** acts as a template or a pattern for which the input is scanned
Fully connected layer	**Layer** type in a **neural network**, where each **neuron** is connected to the next one of the subsequent layer; in **CNN**-based **classifiers**, the **classification** process happens in the **fully connected layers**
Feedforward neural network	A simple network where the output of a **neuron** layer is the input of the layer below; i.e. information flows linearly from input-to-hidden-to-output layers, unlike in recurrent **neural networks** or Boltzmann machines
Hidden layer	**Layer** type in **neural networks**, i.e. 'hidden' between the **input** and the **output layers**. Here, from layer to layer, the input is transformed through mathematical functions before being passed to the output layer. In effect, the learning takes place here. The number of **hidden layers** determines the **model** complexity and in **CNNs**, how well higher-level **features** can be derived from the input features
Hyperparameter	A **parameter**, i.e. not changed/adapted during **model** training, but which describes the model or training itself. For example, this can be the number of **hidden layers** in the network, or the number of **neurons** in each layer

Table 1. Continued

Term	Description
Inference data	Data for which the true **labels** are unknown in a **supervised learning** setting, and for which the model(s) are trained in order to classify them
Input layer	**Layer** type in a **neural network** that takes in raw information or data from the user. No computation is performed at this layer
Kernel functions	When high-dimensional data are not linearly separable, **kernel functions** (such as polynomial or radial basis functions) are applied to project them into a higher-dimensional space where they become linearly separable; **kernel functions** thus facilitate **classification** tasks in **ML**
Label	A value returned by a **classifier** which is not mutually exclusive from other values obtained for all elements of a dataset (such as functional **classes** to protein sequences using GO terms)
Labelled data	Data for which each element's true **label** is known, often split 80:20 into **training data**, to train a **model** and **test data** with which final model accuracy is tested
Layer	Highest-level building block in **neural networks**; effectively a group of neurons, which are not connected to each other and which receive input (either user-provided data, or the output from the preceding **layer** of neurons), transform it using mathematical functions, and pass it to the next **layer**
ML	Field of study in which machines learn from existing data via dedicated algorithms to perform tasks or make accurate predictions without human intervention
Model	Pertaining to **ML/DL**, a **model** is the result of an algorithm being trained on data in order to make predictions about unknown data or to find specific patterns
Natural language processing (NLP)	Branch of AI, which teaches machines to understand, utilize, and generate human language with statistical, **ML** or **DL** algorithms
Neural network	Collection of artificial **neurons** that mimic a group of biological neurons; the neurons are connected to each other in patterns that generally influence the behaviour and capabilities of the algorithm
Neuron/node	The individual unit of **neural networks**, organized in **layers** of several **neurons**; each neuron receives numerical input, e.g. from neurons of the preceding layer, and produces an output via a mathematical function, i.e. applied on the sum of all its inputs, which it then passes on to all its downstream connected neurons
Neural depth	The number of neural **layers** in a neural network
Output layer	**Layer** type in **neural networks,** which brings the information learned through the **hidden layer**(s) into a form, i.e. interpretable by the user; e.g. in case of a simple **classifier**, the **output layer** contains as many **neurons** as there are **classes** (each neuron corresponding to a specific class) and the neuron with the highest output corresponds to the class predicted by the **model** for that input
Parameter	A variable, i.e. internal to the algorithm and optimized during **training**, e.g. **weights** and **biases**
Pooling layer	**Layer** type in a **neural network,** which down-samples the output **feature maps** from a **convolutional layer** (usually to half their original size). Among others, this helps in reducing the number of **parameters** and thus computational cost during training
Predicted data	It is the output generated from a **model**
Regression	Statistical method(s) for understanding and describing the relationship between two **features** or variables
Reinforcement learning	Field of **ML** in which learning is based on a trial-and-error system. Instead of training on labelled or unlabelled data, the algorithm (or 'agent') learns the cost or reward associated with the outcomes of its different choices and then tries to maximize the reward. It can be considered a separate paradigm from **supervised** and **unsupervised learning**
Supervised learning	Field of **ML,** which requires labelled data. In this type of learning, the user defines the **classes** and their corresponding **labels**, and the algorithm learns to associate data/attribute/**feature** patterns with these labels
Semisupervised learning	Field of **ML** where learning takes place by using a combination of small amount of labelled data and a large amount of unlabeled data
Test data	Subset of the **labelled data**, used to test the accuracy of the **model**, after its **parameters** have been optimized in several rounds of training and validation, and before it is fed the **inference data**

Table 1. Continued

Term	Description
Training data	Subset of the **labelled data**, which is used to train an algorithm/**classifier**, i.e. teach the algorithm to associate (patterns of) **features** with the different **classes** or class **labels**. Part of the **training data** (usually 20%) is held back as **validation data** and not used in adjusting algorithm **parameters**
Training Iteration	After each **training iteration**, algorithm **parameters** are updated. The total number of **training iterations** depends on how much of the **training data** the algorithm sees in each iteration, as well as how often it should go over the **training data** to optimize its **parameters**
Unsupervised learning	Field of **ML** that deals with unlabelled data. The algorithm learns previously undetected patterns in a dataset, without relying on user-provided associations between data points and **classes/labels**
Unweighted voting	Assigning a **class/label** to an Inference data point via an **ensemble method**, using the label, which was predicted by the majority of **classifiers**
Validation data	A subset of the labelled **training data**, not used in algorithm **parameter** tuning, but retained to check algorithm accuracy. Each **training iteration** predicts **classes** or **labels** for the **validation data**
Weights and biases	Learnable **parameters** of a **neural network**, i.e. those that are optimized during training; **weights** signify the strength of a connection between two **neurons**. **Biases** are constant values that are added at each **neuron** and influence its output; unlike weights, they are not influenced by a neuron's incoming connections
Weighted voting	Assigning a **class/label** to an **inference data** point via an **ensemble method** by weighting the prediction of each **classifier** where the **weights** are adjusted over the learning period. The algorithms emit predictions, followed by the input of correct **labels** by the user. For each error made by the best performing classifier, its **weight** decreases exponentially

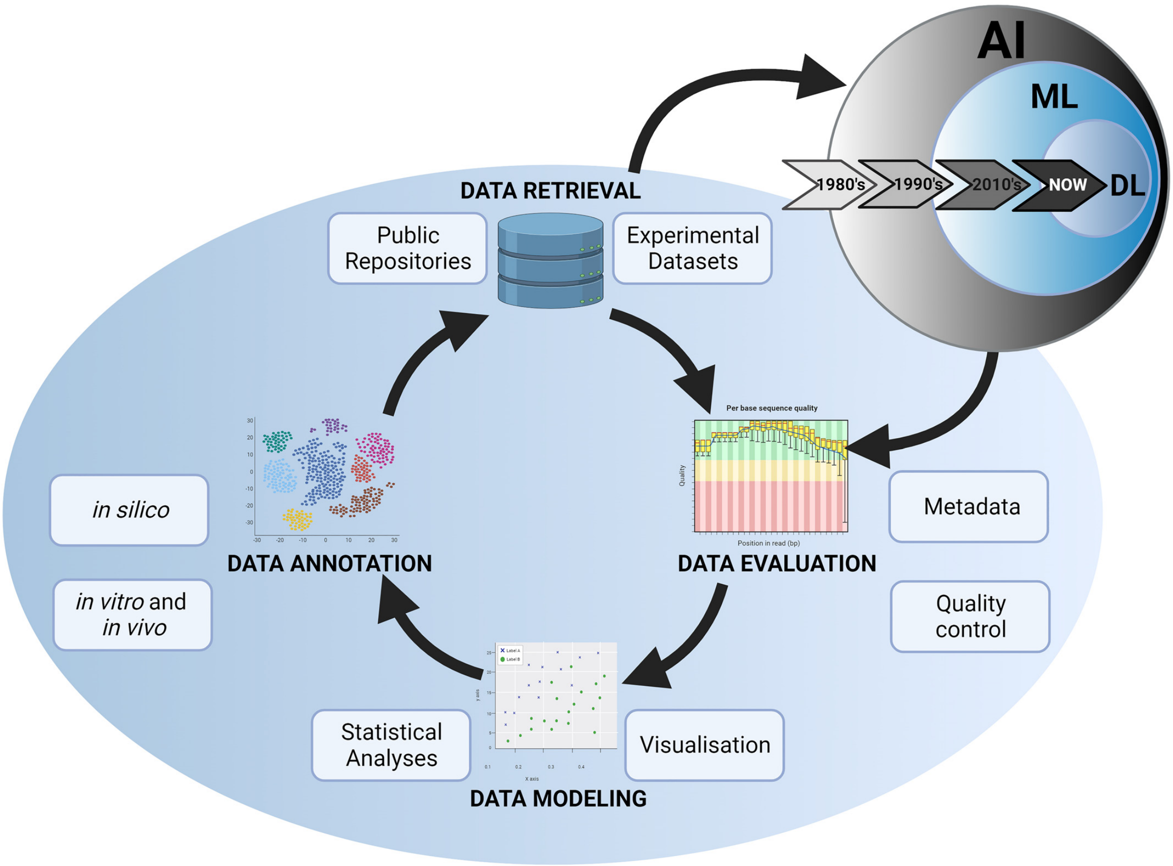

Figure 2. Annotation of data using classical methods and AI. AI surpasses the traditional methods from data curation to data annotation. ML-based techniques have now paved the way for DL neural networks. AI can be used directly on data acquired from experimental or publicly available repositories and extract meaningful features, ensuring a reduction in total number of dimensions, which translates to reduction in the number of input variables for training datasets, hence, reducing computational load.

ture extraction. Unlike ML, which is limited to predicting discrete outputs obtained by counting the data or continuous outputs obtained by measuring input data, DL-based methods are able to learn data representations i.e. feature learning (Fig. 3). These models often have the capability to automatically obtain useful information from input datasets and bypass traditional feature engineering and selection processes (Bonetta and Valentino 2020).

ANNs (Mcculloch and Pitts 1943) can be simply described as being similar to biological neurons where the learning process is due to synaptic connections. The data passes through input, hidden, and output layers. The input layer feeds in the data, from which meaningful information is extracted by hidden layers, which lead to the prediction of data for the classification problem in hand. All the layers comprising this network architecture are together known as a Deep neural network (DNN). ANNs can process nonlinear data and handle noisy data but are prone to overfitting (Kulkarni-Kale et al. 2014). ANNs include RNNs, CNNs, and GCNs (see below).

Recurrent neural networks (RNN) (Sperduti and Starita 1997) can be used for supervised learning with artificial neurons having one or more feedback loops. These loops are recurrent cycles over data. Input-target pairs are provided by the user. RNNs are expected to optimize the networks by minimizing the difference between the target-output pairs (Salehinejad et al. 2018).

Convolutional neural networks (CNN) (Lecun et al. 1998) are comprised of convolutional, pooling, and fully connected layers. They can identify relevant features without human supervision and resemble a feedforward NN. The pooling layer is involved in dimension reduction and the results are forwarded into the fully connected layers. Its massive parallelism yields a great amount of computational efficiency (Alzubaidi et al. 2021).

Graph convolutional networks (GCN) (Miller et al. 1989) have numerous spectral or spatial convolutional layers. Input data featurization and elaborate architectures render them suitable for complicated problems. The graphs can extract meaningful information from their own structures (Zhou et al. 2022).

Transformers are novel neural network architectures based on self-attention mechanism, which are most widely used in the field of natural language processing (NLP). When performing text translation tasks, these DL models are not required to process each sentence from the beginning but refer to the context of each word in a sentence, which improves parallelization of the classification task. These, however, suffer from the problem of sequential computation (Vaswani et al. 2017).

Ensemble classifiers (Dietterich 2000) ensembles make individual decisions by different classifiers, which are combined either by weighted or unweighted voting for classification of new instances. Multiple learning algorithms make the classification model more robust. Methods like this increase efficiency but can be biased as performance heavily depends on weights in a weighted voting (Gokalp and Tasci 2019). These can involve a combination of any of the classifiers described above.

Clustering-based methods are capable of exploiting the direct, as well as the indirect, interaction proteins of the unannotated

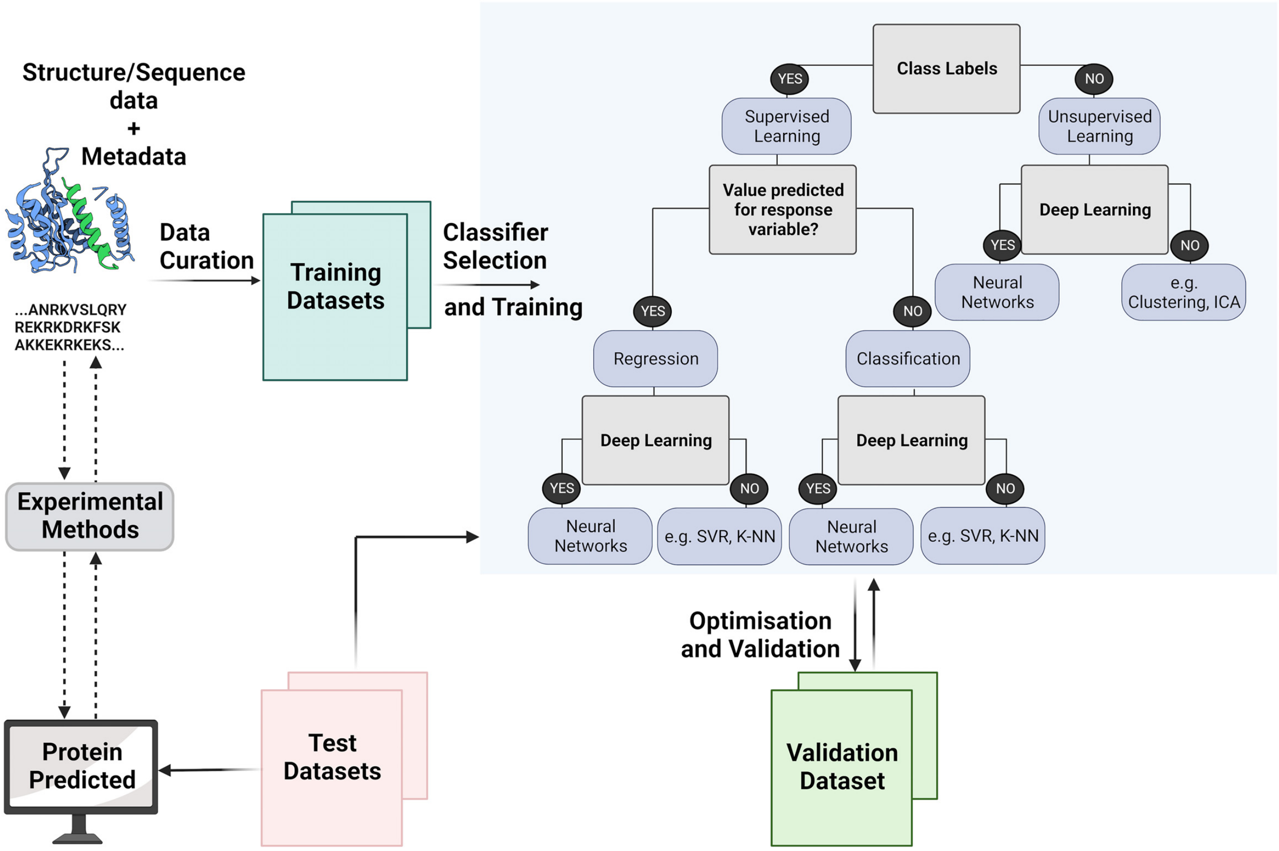

Figure 3. A workflow for functional annotation of proteins using AI. Specific features from sequence and/or structure data are extracted and used as training data. Apart from classification-based tasks, ML algorithms can also perform tasks pertaining to regression that predict the relationship between two known variables. Other methods involve algorithms that group the input data into specific clusters, the output data. A validation dataset is then used to test the efficiency of the model. Test datasets (in this case HPs) are then fed to the optimized model. An experimental feedback loop could also be used for the validation of correct assignment for function. Partially adapted from (Mahood et al. 2020).

protein to predict functions, and of more effectively interpreting the protein interaction relationships in the prediction process (Hou 2017). One of the oldest and most used techniques is the k-means clustering. This clustering algorithm follows an unsupervised approach and does not require the user to define the output clusters.

Reinforcement learning (RF) or reward-based approaches have often been utilized in DL techniques. The famous AlphaFold2 software (Jumper et al. 2021), developed by DeepMind, has used this approach to identify protein structures, which has aided in deciphering how a protein folds in its natural environment (also see below). RF also finds its use in data augmentation, i.e. generating artificial data points or sequences, which helps in balancing the datasets by compensating the lack of adequate protein sequences as described in a recent study (Eftekhar 2020) for the prediction of subcellular localization of proteins.

By using traditional ML or DL classifiers, it is possible to predict the function of HPs without using homology information (Han et al. 2006). There are several other data clustering algorithms such as DBSCAN (density-based spatial clustering of applications with noise) (Ester et al. 1996), which along with ICA (independent component analysis) (Bell and Sejnowski 1995)—a blind-source separation algorithm—help in computing context-specific activities (e.g. metadata) of gene modules, determining their relative strengths, followed by an unsupervised approach and do not require a predetermined number of clusters to perform classification. As an example, DBSCAN and ICA for clustering gene expression data into so called i-modulons (Sastry et al. 2019). Summary aspects and usage of each classifier is given in Table 2, where also the pros and cons for each type of method are listed.

Current computational challenges and resources

One of the biggest challenge in AI is the cost of algorithm training, which scales up with the number of parameters and data points while the computational requirements grow exponentially with number of data points. In order to address the problem emanating with Big Data processing, scientists require dedicated servers with high computing power (Central and Graphics Processing Units) (Thompson et al. 2020).

For research groups with no or very limited funding in terms of computational infrastructure, there are some free services available such as Amazon Web Services (https://aws.amazon.com/machine-learning/ai-services/), which provide processing power for ML-based models but are rather ineffective for training DL models when using the free tier which currently allows 750 hours total computing time for 1 month with two cores, but excludes GPUs required for faster computing. Google cloud (https://cloud.google.com/products/ai) and Azure (https://azure.microsoft.com/en-us/solutions/ai/) are other options but have similar limited capabilities for free accounts. Researchers might also be interested in GPU-based Jupyter notebook servers to train their ML/DL models. These include some freely available platforms like Google Co-

Table 2. Aspects and usage of classifiers. All classifiers/methods used for classification or clustering of protein sequence or structure data. Each classifier is listed with pros, cons, and problem-specific usages, to guide researchers in choosing a method, which is best suited to their datasets and biological question.

Classifier/Method	Task	Pros	Cons	Usage
Support Vector Machine (SVM)	Classification/Regression	Efficient when classes are distinctly separable, relatively memory efficient, suited for binary classification and for data with very high dimensions	Not suitable for large datasets or noisy data, data preprocessing required, overfitting risk, computation can be costly, interpretability of output is low	Protein-protein interaction prediction using sequence data (Ma et al. 2020) Identification of intrinsically disordered proteins using sequence data (Kandemir Cavas and Yildirim 2016)
k-nearest-neighbor (k-NN)	Classification/Regression	Easy to implement for multi-class problems, selected Hyperparamerter remains the same, non-parametric algorithm	Slow algorithm, works well with small number of input variables, not ideal for imbalanced data, sensitive to outliers	Protein structure classification into distinct classes by using structure data (Mirceva et al. 2020) Discrimination of Membrane Transporters using sequence data (Zuo et al. 2015)
Decision Tree (DT)	Classification/Regression	Data pre-processing is easy and does not require scaling or normalization, good for visual representation of output, can utilize numerical and categorical features	Higher training times, overfitting, inadequate prediction for large complex datasets, not ideal for imbalanced datasets	Classification of proteins from PDB into families based on surface roughness using structure data (Satpute and Yadav 2019) Protein-protein interaction prediction using different physicochemical properties by utilising protein sequence data (Zhou et al. 2017)
Random Forest (RF)	Classification/Regression	Scaling and transformation of input variables is not required, ideal for working with large number of features, lower overfitting risk, works well with non-linear data	Not easily interpretable, computationally intensive, training is slow, limited performance with regeression while dealing with linear varibales	Prediction of small encoded proteins and their functions in a species-specific context by protein sequence data (Miravet-Verde et al. 2019). Prediction of protean segments from sequence of an Intrinsically disordered protein (Basu et al. 2017)
Recurrent Neural Network (RNN)	Classification/Regression	Dynamic neural network, computationally powerful, useful for non-linear systems	Rough to train long sequences due to gradient vanishing problem and exploding gradient problem, slow and complex training	Antibiotic resistance class prediction using protein sequences (Hamid and Friedberg, 2020) Prediction of essential proteins by integrating information from protein-interaction networks, gene expression profiles and subcellular localization (Zeng et al. 2021)
Convolutional Neural Network (CNN)	Classification/Regression	Unsupervised learning, high accuracy, weight sharing, reduce dimensionality in neural network	Long training period, requires large datasets, fails to encode the position and orientation of objects	Prediction of metal binding sites using structure data (Mohamadi et al. 2022) Identification of efflux proteins in transporters using sequence data (Taju et al. 2018)
Graphical Neural Network (GCN)	Classification/Regression /Clustering	Uses same parameters in the training iteration, inexpensive data storage, adaptively learn the importance of neighbours in a graph-based system	The ‘Black Box‘ problem, which makes algorithm’s processes untraceable, computation cost	Graph based prediction of PPIs using raw sequence data (Yang et al. 2020) Protein-protein interactions prediction using structure data (Pancino et al. 2020)

Table 2. Continued

Classifier/Method	Task	Pros	Cons	Usage
Ensemble Classifier	Classification/Regression /Clustering	Higher predictive accuracy than individual models, can handle both linear and non-linear data, Bias/overfitting can be reduced, less noisy and more stable	Model that is closest to the 'true' data generating process beats other methods, lack of interpretability, computationally expensive, memory intensive	Sequence-based classification of antioxidant proteins by RF and SVM (Meng et al. 2019) Protein function prediction using Sequence, taxonomic, structural domains and amino acid index based classification by CNN and RF (Hakala et al. 2022) Protein secondary structure prediction from sequence data by DT and SVM (Afify et al. 2021) Complex data clustering using General adversial networks for assigning proteins to different sub-families by RNN and Clustering (Bitard-Feildel, 2021)
Transformer Model	Classification/Regression /Clustering	Self-attention, universal and flexible architecture, do not provide specific structure to input data (for eg. The need for specifying nodes and edges as in GCN)	Limited to higher level representation to data, limited memory span, overfitting	Utilization of sequence statistics, chemical and biological features to generate biologically active sequences (Rives et al. 2020) Structure, remote homology prediction and protein engineering using sequence data (Rao et al. 2019) Protein Language model for prediction of secondary structure and subcellular localization (Heinzinger et al. 2019)
Clustering Model	k-means Mean shift DBSCAN Gaussian mixture Hierarchical clustering	Fast, few computations No pre-set clusters, intuitive No pre-set clusters, identifies outliers More flexible than k-means Not sensitive to the choice of distance metric, visualize hierarchy	Inconsistent, outliers Selection of a clusters Inconsistent for high dimensions Uses all available data points Lower efficiency, time complexity	Construction of protein networks using sequence data (Keel et al. 2018) Unsupervised clustering based prediction of protein structure and function using relative solvent accessibility of amino acid residues (Teletin et al. 2018) Identification of antibiotic and virulence resistance genes in pathogenic bacteria using genomic sequence data (Li et al. 2018a)

lab (Bisong 2019) and Kaggle (https://www.kaggle.com), again with limitations for storage and processing power. For instance, Collab has a space limitation of 15 GB and allows users to run their notebooks for 12 hours per day after which the user is required to pay for additional hours. Amazon Sagemaker (https://aws.amazon.com/sagemaker/) is another option for large-scale training of ML models. It is time efficient, having built-in algorithms and optimized frameworks making it easier to use, but currently requires an hourly fee of $1.125 to train large datasets. Other considerable platforms are DataCrunch (https://datacrunch.io) and Paperspace (https://www.paperspace.com), which have lower GPU costs of $1.1 and $0.18 per hour, respectively as compared to Sagemaker (https://vitalflux.com/deep-learning-top-5-online-jupyter-notebooks-servers/).

Examples of AI used in protein function prediction of prokaryotes

Table 3 gives an overview of all tools/software that are used for prokaryotic protein function prediction, with the type of input data required by the model and the generated output. Tables 2 and 3 can, therefore, aid researchers in choosing the model best suited for their data and biological question. Based on the most cited tools, we here, discuss a few examples, using methods including KNN, ensemble approaches, DL, and NLP.

NetGO (You et al. 2019) is a GO-term prediction tool, which builds on a previous ensemble learning framework, GOLabeler (You et al. 2018), by adding a module using KNN, a supervised classification algorithm. The protein features used for learning are

Table 3: Machine and Deep Learning-based tools/software forprotein function prediction. ARGs, antibiotic resistance genes; BGCs, bacterial gene clusters; CS, contribution scores; DBP, DNA binding protein; EC, enzyme commission IDs; emb, embeddings; EP/NEP, essential/non-essential proteins; FD, fingerprint descriptors; func, functional classes; GA, gene annotations; GE, gene expression data; GC, gene cluster; Ge, genome; GO, gene ontology annotations; HbL, haemoglobin; Kace/Non-Kace, lysine acetylation site/non-lysine acetylation site; LBS, ligand binding sites; Nt seq, nucleotide sequences; ORF, open reading frame; PC, protein clusters; Pfams, protein families; PI, protein interactions; PS, probability scores; PSC/PRS, per sequence/per residue scores, RP, ribosomal profiling data; SA, species abundance; Sbfam, subfamily; Seq, protein sequence data; SB; source biome; Sdr, short DNA reads; Strc, protein structure data; SL, subcellular localizations; TA: taxonomic abundance score.

Tool/Software	Classifier /method	Type of Input(s)	Type of Output(s)	References
SVMProt	Support Vector Machine	Seq, Strc	Pfams	Li et al. (2016)
BacHbpred		Seq	HbL	Krishnan et al. (2016)
iProEP		Seq	Pfams	Lai et al. (2019)
GrAPFI	k-Nearest Neighbour	Seq	EC	Sarker et al. (2020)
PseAAC		Seq	SL	Jiang et al. (2019)
CrowdGO	Decision Trees	GO	GO	Reijnders and Waterhouse (2021)
PPI & Gabor filter		Seq	PI	Zhan et al. (2020)
P2Rank	Random Forest	Seq	LBS	Krivák and Hoksza (2018)
FEAST		TA	SB	Shenhav et al. (2019)
DEEPred		Seq	GO	Rifaioglu et al. (2019)
UDSMProt	Recurrent Neural Network	Seq	GO, EC, Pfams	Strodthoff et al. (2020)
ProDec-BLSTM		Seq	PS	Li et. al (2017)
PARROT		Seq	PSC/PRS	Griffith and Holehouse (2021)
DeepBGC		Ge	BGCs	Hannigan et al. (2019)
LookingGlass		Sdr	EC, Emb	Hoarfrost et al. (2020)
UniRep		Seq	Emb	Alley et al. (2019)
ProLanGO		Seq	GO	Cao et al. (2017)
MultiPredGO		Seq, Strc	GO	Giri et al. (2021)
DeepGOPlus	Convolutional Neural Network	Seq	GO	Kulmanov and Hoehndorf (2020)
DeepAdd		Seq, PI	GO	
Balrog		Seq	Genes	Du et al. (2020)
ProtCNN		Seq	Pfams	Sommer and Salzberg (2021)
DeepHiFam		Nt seq	CS	Bileschi et al. (2019)
SmORFinder		PI, GE	EP/NEP	Sandaruwan and Wannige (2021)
DeepEP		Str	FD	
MaSIF		SA	SB	Durrant et al. (2020)
ONN4MST		Seq, ARGs	GA	Zeng et al. (2019)
ON4ARG		Seq	GO	Gainza et al. (2020)
DeeProtGO		SA	SB	Zha et al. (2020)
EXPERT		Seq	EC	Zha et al. (2021)
ProteInfer		Seq	EC, GO	Merino et al. (2022)
SeqVec		Seq	Pfams, Sbfam	Chong et al. (2021) Sandesron et al. (2021) Heinzinger et al. (2019)
AlphaFold		Seq, Strc	Strc	Jumper et al. (2021)
DeepFri	Graphical Neural Network	Strc	GO	Gligorijević et al. (2021)
PersGNN		Strc	GO	Swenson et al. (2020)
PANDA2		Seq, GO	GO	Zhao et al. (2022)
DEEPre		Seq	EC	Li et al. (2018)
ECPred	Ensemble classifiers	EC	EC	Dalkiran et al. (2018)
BioSeq-Analysis		Seq	DBPs	Liu (2019)
DeepRibo		Seq, RP	ORFs	Clauwaerts et al. (2018)
LargeGOPred		Seq	GO	Wang et al. (2020)
DeepGraphGO		Seq, PI	GO	You et al. (2021)
GOLabeler		Seq, PI, GO	GO	You et al. (2018)
NetGO		Seq	GO	You et al. (2019)
ProPythia		Seq	EC	Sequeira et al. (2022)
DeepMicrobes		Nt seq	Emb	Liang et al. (2019)
STALLION		Seq	Kace/ Non-Kace	Basith et al. (2022)
Deep_CNN_LSTM_GO		Seq	GO	Abdou et al. (2021)
PFmulDL		Seq	GO	Xia et al. (2022)
TALE	Transformer	Seq	GO	Cao and Shen (2021)
ProteinBERT		Seq, GO	GO	Brandes (2022)
iModulonDB		GE	GC	Rychel et al. (2021)
AGNOSTOS DB	Clustering	Nt seq	GC	Vanni et al. (2021)
PPI-GA		PI	PC	Shirmohammady et al. (2019)

the networks in which a protein is involved in, as found in the 'STRING' database of protein–protein interactions. If the protein is not present in this database, then data for the closest homolog is used, if one is available. GO-terms from the closest neighbours in the (potentially multiple) relevant networks are applied to the new protein by aggregating weights from different networks followed by the kNN approach of plurality voting. Significant limitations of this method are genes, which lack homology to entries in the STRING database, and examples where homologous proteins with low similarity have different functions.

Another ensemble approach, SVM-Prot 2016 (Li et al. 2016), an update to an earlier tool (Cai et al. 2003), aims to predict what it calls 'functional families' from protein sequences without relying on detectable homology. Physicochemical features of proteins such as solubility are derived from sequences and used to train a SVM algorithm to classify sequences into functional classes. The functional groupings of proteins for training purposes (labels) are derived from GO and other databases. A negative training set is derived by choosing some Pfam families with no members included in the positive training set.

There is also a growing set of tools for protein structure prediction which use DL approaches. Two of the most prominent are AlphaFold2 from Google's DeepMind (Jumper et al. 2021) and ESMfold from Meta AI Research (Lin et al. 2022). Alphafold2 uses information obtained from a multiple sequence alignment on evolutionary couplings between amino acid residues to infer pairwise distances between the residues, and from there infers a 3D protein structure. While the earlier version of AlphaFold used statistics calculated from multiple sequence alignment input, AlphaFold2 embeds the full sequence information, and uses a transformer neural network, which is able to take into account relationships between residues and apply an 'attention' mechanism to take into account those previously learned from training sets to be most important. The neural network then feeds the information into a structural model neural network, which uses another attention mechanism transformer to take into account the most important pairwise relationships and output a structural model, i.e. the 3D co-ordinates of each of the residues' atoms (Jones and Thornton 2022). ESMfold, on the other hand, uses a very large natural language model designed for proteins to capture key aspects of an input sequence, which relate to structure as an attention map, predicts pairwise distances from this, and uses both to predict structures (Lin et al. 2022). A protein language model is useful here because structural features are emergent properties of sequences, which can be learned with large input datasets. The overall architecture of the approach is based on Alphafold2, but it replaces the transformer neural network that Alphafold2 uses to process a multiple sequence alignment with a protein language model, which takes a single input sequence. High confidence structures predicted with Alphafold2 are used as part of the training data. ESMfold is less accurate than AlphaFold2, but more than an order of magnitude faster, without the need for building a multiple sequence alignment. Protein structures are not identical to protein functions, but there is often a close relationship between them, which is why structures can then be used by ML models to predict protein function. E.g. DeepFRI is a GCN for predicting functions, which uses protein structures as input along with sequence features derived from a protein language model (Gligorijević et al. 2021).

Major developments have occurred regarding inferring protein structures from sequence data for function prediction. Nevertheless, there are still obstacles, e.g. when applying algorithms for human HPs to microbial HPs (Suravajhala and Sundararajan 2012) since these are based on finding specific domains whilst the majority of the latter have uncharacterized domains. Additionally, almost 20% of the total proteins from 944 bacterial species have no identifiable domains at all (Wang et al. 2019). Feed-forward DNN-based modelling approaches for large-scale automated protein function prediction are also probably not a good choice for functional terms with low or moderate number of annotated proteins and it is currently not feasible to carry out a fold-based cross-validation analysis, especially when the number of model training operations are high, since it usually requires extremely high computational power (Sureyya Rifaioglu et al. 2019). Furthermore, many state-of-the-art ML and DL models, have not yet been extensively explored for protein function prediction.

Outlook

Accurate prediction of microbial protein functions has the potential to revolutionize multiple fields of biological research. The accelerating expansion in biological Big Data presents many opportunities but also challenges for researchers, such as computational power and storage limitations and properly integrating diverse data types. DL algorithms have now paved the way for a better and more competent prediction of HPs. Furthermore, new genetic elements and sources of genetic information will open the door to an unexplored world of coding and noncoding complexity in prokaryote genomes (Grainger 2016, Kirchberger et al. 2020). This new world, if it was included in annotation, would further expand the class of 'hypothetical proteins'. Gene annotation tools, however, have typically excluded short ORFs and overlapping ORFs, largely as a matter of practicality. In recent years, many of both kinds have been discovered to be true protein-coding genes (Storz et al. 2014, Ardern et al. 2020). The number of 'dual-coding' RNAs, where there are separate functions for a transcript at the level of RNA and protein structure, remains unknown. Further, the phenomenon of 'proteoforms' (Smith et al. 2013), i.e. alternative start sites for genes, has shown the terrain of prokaryotic protein coding to be yet more complex. Nevertheless, direct laboratory analysis will still remain the gold standard of functional annotation for the foreseeable future. An experimental feedback loop is consequently of great importance to further test and train the models for accurate predictions and progress will come from creative and efficient integration of existing data with careful experimental validation. Metaservers or servers, which can be used to input query data simultaneously to extract specific features from the data keep improving with time, hence facilitating developers and users alike, by making the training and testing datasets as well as benchmarking results available in public domain. While there is a need to refine the methods to achieve higher accuracy, there is also a growing need to bring various aspects of protein function prediction under the realm of AI.

One of the key takeaway points from our review of the available resources and methods is that at least something informative can be said about essentially all protein sequences. The label 'uncharacterized' or 'hypothetical' is completely uninformative, but for most proteins it should be possible to derive at least some characteristics from their general sequence properties (e.g. do they have a transmembrane domain?), whether there is evidence of expression, their possible cellular location, details of homologs, and whether they are predicted or demonstrated to be expressed as part of an operon and/or regulon. New approaches to storing gene annotation data should take these diverse lines of evidence into account.

Authors’ contributions

S.C., Z.A., and A.K.K. wrote the manuscript, F.L. helped with the revision, and A.K.K. had the idea for this review and acquired the funding.

Competing interests. The authors declare no competing interests.

Funding

This work was supported by the Future Fields DigeteLiSe grant of the Karlsruhe Institute of Technology (5207.0011.5096).

References

Afify HM, Abdelhalim MB, Mabrouk MS *et al.* Protein secondary structure prediction (PSSP) using different machine algorithms. *Egypt J Med Hum Genet* 2021;**22**:54.

Aguiar-Pulido V, Huang W, Suarez-Ulloa V *et al.* Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. *Evol Bioinform Online* 2016;**12**:5–16.

Alley E, Khimulya G, Biswas S *et al.* Unified rational protein engineering with sequence-based deep representation learning. *Nat Methods* 2019;**16**. doi: 10.1038/s41592-019-0598-1.

Alloghani M, Al-Jumeily D, Mustafina J *et al.* A systematic review on supervised and unsupervised machine learning algorithms for data science. In: *Supervised and Unsupervised Learning for Data Science*. Cham: Springer, 2020, 3–21.

Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. *Am Stat* 1992;**46**. doi: 10.1080/00031305.1992.10475879.

Altschul SF, Gish W, Miller W *et al.* Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.

Alzubaidi L, Zhang J, Humaidi AJ *et al.* Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 2021;**8**. doi: 10.1186/s40537-021-00444-8.

Ardern Z, Neuhaus K, Scherer S. Are antisense proteins in prokaryotes functional?. *Front Mol Biosci* 2020;**7**. doi: 10.3389/fmolb.2020.00187.

Arslan M, Müller JA, Gamal El-Din M. Aerobic naphthenic acid-degrading bacteria in petroleum-coke improve oil sands process water remediation in biofilters: dNA-stable isotope probing reveals methylotrophy in Schmutzdecke. *Sci Total Environ* 2022;**815**:151961.

Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.

Aslam B, Basit M, Nisar MA *et al.* Proteomics: technologies and their applications. *J Chromatogr Sci* 2017;**55**:182–96.

Asnicar F,Weingart G,Tickle T *et al.*.Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ* 2015;**3**:e1029. 10.7717/peerj.1029.

Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;**28**:304–5.

Basith S, Lee G, Manavalan B. STALLION: a stacking-based ensemble learning framework for prokaryotic lysine acetylation site prediction. *Briefings Bioinf* 2022;**23**:bbab376.

Basu S, Söderquist F, Wallner B. Proteus: a random forest classifier to predict disorder-to-order transitioning binding regions in intrinsically disordered proteins. *J Comput Aided Mol Des* 2017;**31**:1–14.

Bateman A, Martin M-J, Orchard S *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**. doi: 10.1093/nar/gkaa1100.

Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* 1995;**7**:1129–59.

Benson D,Cavanaugh M,Clark K *et al.* GenBank. *Nucleic Acids Research*, 2018;**46**:D41–D47. 10.1093/nar/gkx1094.

Bernard G, Pathmanathan J, Lannes R *et al.* Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery.*Genome Biol Evol* 2018;**10**. doi: 10.1093/gbe/evy031.

Bhattacharyya S, Bershtein S, Yan J *et al.* Transient protein-protein interactions perturb *E. coli* metabolome and cause gene dosage toxicity.*Elife* 2016;**5**. doi: 10.7554/eLife.20309.

Bileschi ML, Belanger D, Bryant D *et al.* Using deep learning to annotate the protein universe. *Nat Biotechnol* 2019;**40**:626507.

Bisong E. Google colaboratory. In: Bisong E (ed.), *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*. Berkeley: Apress, 2019, 59–64.

Bitard-Feildel T. Navigating the amino acid sequence space between functional proteins using a deep learning framework. *PeerJ Comput Sci* 2021;**7**:e684.

Blum M, Chang H-Y, Chuguransky S *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* 2021;**49**. doi: 10.1093/nar/gkaa977.

Bonetta R, Valentino G. Machine learning techniques for protein function prediction. *Proteins Struct Funct Bioinf* 2020;**88**:397–413.

Borah P, Teja A, Jha SA *et al.* TUKNN: a parallel KNN algorithm to handle large data. In: Patgiri R, Bandyopadhyay S, Borah MD *et al.* (eds). *Big Data, Machine Learning, and Applications*. Cham: Springer International Publishing, 2020, 1–13.

Brandes N, Ofer D, Peleg Y *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**: 2102–10.

Breiman L. Random forests. *Mach Learn* 2001;**45**. doi: 10.1023/A:1010933404324.

Brockhurst MA, Harrison E, Hall JPJ *et al.* The ecology and evolution of pangenomes. *Curr Biol* 2019;**29**. doi: 10.1016/j.cub.2019.08.012.

Cai C, Han LY, Ji Z-L *et al.* SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res* 2003;**31**:3692–7.

Cain AK, Barquist L, Goodman AL *et al.* A decade of advances in transposon-insertion sequencing. *Nat Rev Genet* 2020;**21**:526–40.

Cao R, Freitas C, Chan L *et al.* ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 2017;**22**:1732.

Cao Y, Shen Y. TALE: transformer-based protein function annotation with joint sequence–label embedding. *Bioinformatics* 2021;**37**:2825–33.

Chen X, Ma A, McDermaid A *et al.* RECTA: regulon identification based on comparative genomics and transcriptomics analysis. *Genes* 2018;**9**. doi: 10.3390/genes9060278.

Chen Y, Yang L, Ding Y *et al.* Tracing evolutionary footprints to identify novel gene functional linkages. *PLoS ONE* 2013;**8**. doi: 10.1371/journal.pone.0066817.

Chong H, Yu Q, Zha Y *et al.* Enabling technology for microbial source tracking based on transfer learning: from ontology-aware general knowledge to context-aware expert systems. *Biorxiv* 2021. doi: 10.1101/2021.01.29.428751.

Clauwaert J, Menschaert G, Waegeman W. DeepRibo: a neural network for precise gene annotation of prokaryotes by combining ribosome profiling signal and binding site patterns. *Nucleic Acids Res* 2019;**47**: e36.

Cooper W, Owen R. *Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals: Delivered at the Royal College of Surgeons in 1843*. London : Longman, Brown, Green, and Longmans, 1843.

Dalkiran A, Rifaioglu AS, Martin MJ *et al.* ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinf* 2018;**19**:334.

Dam HT, Vollmers J, Sobol MS *et al.* Targeted cell sorting combined with single cell genomics captures low abundant microbial dark matter with higher sensitivity than metagenomics. *Front Microbiol* 2020;**11**. doi: 10.3389/fmicb.2020.01377.

David B, Andrej S. Protein structure prediction and structural genomics. *Science* 2001;**294**:93–6.

Dietterich TG. Ensemble methods in machine learning. In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer, 2000, 1–15.

Du Z, He Y, Li J *et al.* DeepAdd: protein function prediction from k-mer embedding and additional features.*Comput Biol Chem* 2020;**89**:107379.

Durrant MG, Bhatt AS. Automated prediction and annotation of small proteins in microbial genomes. *Cell Host Microbe* 2020;**29**. doi: 10.1016/j.chom.2020.11.002.

Eftekhar MG. Prediction of protein subcellular localization using deep learning and data augmentation. *Biorxiv* 2020. doi: 10.1101/2020.05.19.068122.

Elhaj-Abdou MEM, El-Dib H, El-Helw A *et al.* Deep_CNN_LSTM_GO: protein function prediction from amino-acid sequences. *Comput Biol Chem* 2021;**95**:107584.

Eme L, Doolittle WF. Archaea. *Curr Biol* 2015;**25**:R851–5.

Emmanuel IA, Stanier C. Defining big data. In: *Proceedings of the International Conference on Big Data and Advanced Wireless Technologies BDAW '16*. New York: ACM Digital Library, 2016, 1–6.

Escudeiro P, Henry CS, Dias RPM. Functional characterization of prokaryotic dark matter: the road so far and what lies ahead. *Curr Res Microb Sci* 2022;**3**:100159.

Ester M, Kriegel H-P, Sander J *et al.* A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Palo Alto: AAAI Press, 1996, 226–31.

Fleischmann A, Darsow M, Degtyarenko K *et al.* IntEnz, the integrated relational enzyme database. *Nucleic Acids Res* 2004;**32**. doi: 10.1093/nar/gkh119.

Fox N, Brenner S, Chandonia J-M. The value of protein structure classification information—surveying the scientific literature. *Proteins Struct Funct Bioinf* 2015;**83**:2025–38. doi: 10.1002/prot.24915.

Freddolino PL, Amemiya HM, Goss TJ *et al.* Dynamic landscape of protein occupancy across the *Escherichia coli* chromosome. *PLoS Biol* 2021;**19**. doi: 10.1371/journal.pbio.3001306.

Fremin BJ, Sberro H, Bhatt AS. MetaRibo-Seq measures translation in microbiomes. *Nat Commun* 2020;**11**. doi: 10.1038/s41467-020-17081-z.

Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* 2012;**13**:840–52.

Gainza P, Sverrisson F, Monti F *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* 2020;**17**:184–92.

Gao Y, Yurkovich JT, Seo SW *et al.* Systematic discovery of uncharacterized transcription factors in *Escherichia coli* K-12 MG1655. *Nucleic Acids Res* 2018;**46**: 10682–96.

Gasteiger E, Gattiker A, Hoogland C *et al.* ExPASy: the proteomics server for in-depth protein knowledge and analysis.*Nucleic Acids Res* 2003;**31**: 3784–8.

Gaudet P, Dessimoz C. Gene ontology: pitfalls, biases, and remedies. In: Dessimoz C, Škunca N (eds). *The Gene Ontology Handbook*. New York: Springer, 2017, 189–205.

Gavriilidou A, Mackenzie TA, Sánchez P *et al.* Bioactivity screening and gene-trait matching across marine sponge-associated bacteria. *Mar Drugs* 2021;**19**. doi: 10.3390/md19020075.

Ghatak S, King ZA, Sastry A *et al.* The y-ome defines the 35% of *Escherichia coli* genes that lack experimental evidence of function. *Nucleic Acids Res* 2019;**47**:2446–54.

Giri SJ, Dutta P, Halani P *et al.* MultiPredGO: deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information. *IEEE J Biomed Health Inform* 2021;**25**:1832–8.

Gligorijević V, Renfrew PD, Kosciolek T *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**. doi: 10.1038/s41467-021-23303-9.

Gokalp O, Tasci E. Weighted voting based ensemble classification with hyper-parameter optimization. In: *Proceedings of the 2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*. Piscataway: IEEE, 2019, 1–4.

Golicz AA, Bayer PE, Bhalla PL *et al.* Pangenomics comes of age: from bacteria to plant and animal applications. *Trends Genet* 2020;**36**:132–45.

Goodacre NF, Gerloff DL, Uetz P. Protein domains of unknown function are essential in bacteria. *Mbio* 2013;**5**. doi: 10.1128/mBio.00744-13.

Grainger DC. Structure and function of bacterial H-NS protein. *Biochem Soc Trans* 2016;**44**. doi: 10.1042/BST20160190.

Graur D, Zheng Y, Price N *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 2013;**5**. doi: 10.1093/gbe/evt028.

Greener JG, Kandathil SM, Moffat L *et al.* A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 2022;**23**:40–55.

Griffith D, Holehouse AS. PARROT: a flexible recurrent neural network framework for analysis of large protein datasets. *Elife* 2021;**10**. doi: 10.7554/eLife.70576.

Hakala K, Kaewphan S, Björne J *et al.* Neural network and random forest models in protein function prediction. *IEEE/ACM Trans Comput Biol Bioinf* 2022;**19**:1772–81.

Hall RJ, Whelan FJ, Cummins EA *et al.* Gene-gene relationships in an *Escherichia coli* accessory genome are linked to function and mobility. *Microb Genom* 2021;**7**:000650. doi: 10.1099/MGEN.0.000650.

Hamid M-N, Friedberg I. Transfer learning improves antibiotic resistance class prediction. *Biorxiv* 2020. doi: 10.1101/2020.04.17.047316.

Han L, Cui J, Lin H *et al.* Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* 2006;**6**. doi: 10.1002/pmic.200500938.

Handfield M, Levesque RC. Strategies for isolation of in vivo expressed genes from bacteria. *FEMS Microbiol Rev* 1999;**23**: 69–91.

Hannigan GD, Prihoda D, Palicka A *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res* 2019;**47**:e110.

Heinzinger M, Elnaggar A, Wang Y *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinf* 2019;**20**:723.

Hoarfrost A, Aptekmann A, Farfañuk G *et al.* Shedding light on microbial dark matter with a universal language of life. *Bioinformatics* 2020;**12**. doi: 10.1101/2020.12.23.424215.

Hou J. Chapter 3 - clustering-based protein function prediction. In: Hou J (ed.), *New Approaches of Protein Function Prediction from Protein Interaction Networks*. Cambridge: Academic Press, 2017, 37–58.

Howe KL, Achuthan P, Allen J *et al.* Ensembl 2021. *Nucleic Acids Res* 2021;**49**:D884–91.

Huerta-Cepas J, Szklarczyk D, Heller D *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019;**47**:D309–14.

Hug LA, Baker BJ, Anantharaman K *et al.* A new view of the tree of life. *Nat Microbiol* 2016;**1**:16048.

Hugenholtz P, Chuvochina M, Oren A *et al.* Prokaryotic taxonomy and nomenclature in the age of big sequence data. *ISME J* 2021;**15**. doi: 10.1038/s41396-021-00941-x.

Ijaq J, Chandrasekharan M, Poddar R *et al.* Annotation and curation of uncharacterized proteins- challenges. *Front Genet* 2015;**6**:119.

Ireland WT, Beeler SM, Flores-Bautista E *et al.* Deciphering the regulatory genome of *Escherichia coli*, one hundred promoters at a time. *Elife* 2020;**9**:1–76.

Jacob F, Monod J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 1961;**3**. doi: 10.1016/S0022-2836(61)80072-7.

Jaroszewski L, Li Z, Krishna SS *et al.* Exploration of uncharted regions of the protein universe. *PLoS Biol* 2009;**7**. doi: 10.1371/journal.pbio.1000205.

Jiang Z, Wang D, Wu P *et al.* Predicting subcellular localization of multisite proteins using differently weighted multi-label k-nearest neighbors sets. *Technol Health Care* 2019;**27**:185–93.

Jones DT, Thornton JM. The impact of AlphaFold2 one year on. *Nat Methods* 2022;**19**:15–20.

Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**. doi : 10.1038/s41586-021-03819-2.

Junier I, Rivoire O. Conserved units of co-expression in bacterial genomes: an evolutionary insight into transcriptional regulation. *PLoS ONE* 2016;**11**. doi : 10.1371/journal.pone.0155740.

Kalkreuter E, Pan G, Cepeda AJ *et al.* Targeting bacterial genomes for natural product discovery. *Trends Pharmacol Sci* 2020;**41**. doi : 10.1016/j.tips.2019.11.002.

Kandemir Çavaş Ç, Yildirim S. Classifying ordered-disordered proteins using linear and kernel support vector machines. 2016;**41**:431–6.

Kanehisa M, Furumichi M, Tanabe M *et al.* KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2017;**45**:D353–61.

Karaduta O, Dvanajscak Z, Zybailov B. Metaproteomics—an advantageous option in studies of host–microbiota interaction. *Microorganisms* 2021;**9**. doi : 10.3390/microorganisms9050980.

Karp PD, Billington R, Caspi R *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Briefings Bioinf* 2019;**20**:1085–93.

Kaster A-K, Sobol M. Microbial single-cell omics: the crux of the matter. *Appl Microbiol Biotechnol* 2020;**104**. doi : 10.1007/s00253-020-10844-0.

Keel BN, Deng B, Moriyama EN. MOCASSIN-prot: a multi-objective clustering approach for protein similarity networks. *Bioinformatics* 2018;**34**:1270–7.

Kirchberger PC, Schmidt ML, Ochman H. The ingenuity of bacterial genomes. *Annu Rev Microbiol* 2020;**74**. doi : 10.1146/annurev-micro-020518-115822.

König S, Romoth L, Stanke M. Comparative genome annotation. In: Setubal JC, Stoye J, Stadler PF (eds). *Comparative Genomics: Methods and Protocols*. New York: Springer, 2018, 189–212.

Krishnan M, Puri M, Dikshit K *et al.* BacHbpred: support vector machine methods for the prediction of bacterial hemoglobin-like proteins. *Adv Bioinformatics* 2016;**2016**:1–11.

Kriváк R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J Cheminform* 2018;**10**:39.

Kulkarni-Kale U, Sawant S, Kadam K *et al.* Prediction of protein function based on machine learning methods: an overview. In: *Genomics III Methods, Techniques and Applications*. 1st edn. Hong Kong: iConcept Press Ltd, 2014.

Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 2020;**36**:422–9.

Lai H-Y, Zhang Z-Y, Su Z-D *et al.* iProEP: a computational predictor for predicting promoter. *Mol Ther Nucleic Acids* 2019;**17**. doi : 10.1016/j.omtn.2019.05.028.

Larsen BB, Miller EC, Rhodes MK *et al.* Inordinate fondness multiplied and redistributed: the number of species on Earth and the new pie of life. *Q Rev Biol* 2017;**92**:229–65.

Lecun Y, Bottou L, Bengio Y *et al.* Gradient-based learning applied to document recognition. *Proc IEEE* 1998;**86**:2278–324.

Lewis WH, Tahon G, Geesink P *et al.* Innovations to culturing the uncultured microbial majority. *Nat Rev Microbiol* 2021;**19**. doi : 10.1038/s41579-020-00458-8.

Li J, Tai C, Deng Z *et al.* VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Brief Bioinform* 2018a;**19**:566–74.

Li S, Chen J, Liu B. Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinf* 2017;**18**: 443.

Li W, O'Neill KR, Haft DH *et al.* RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 2021;**49**. doi : 10.1093/nar/gkaa1105.

Li Y, Wang S, Umarov R *et al.* DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* 2018b;**34**:760–9.

Li YH, Xu JY, Tao L *et al.* SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. *PLoS ONE* 2016;**11**:e0155290.

Liang Q, Bible PW, Liu Y *et al.* DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genom Bioinform* 2019;**2**:694851.

Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic level protein structure with a language model. *Biorxiv* 2022. doi: 10.1101/2022.07.20.500902.

Liu B. BioSeq-analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief Bioinf* 2019;**20**:1280–94.

Liu Z, Dong H, Cui Y *et al.* Application of different types of CRISPR/Cas-based systems in bacteria. *Microb Cell Fact* 2020;**19**:172.

Lobb B, Tremblay BJ-M, Moreno-Hagelsieb G *et al.* An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom* 2020;**6**. doi : 10.1099/mgen.0.000341.

Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci* 2016;**113**. doi : 10.1073/pnas.1521291113.

Logan DC. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *J Exp Bot* 2009;**60**. doi : 10.1093/jxb/erp043.

Louca S, Mazel F, Doebeli M *et al.* A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol* 2019;**17**. doi : 10.1371/journal.pbio.3000106.

Love AC. Functional homology and homology of function: biological concepts and philosophical consequences. *Biol Philos* 2007;**22**. doi : 10.1007/s10539-007-9093-7.

Lv Z, Ao C, Zou Q. Protein function prediction: from traditional classifier to deep learning. *Proteomics* 2019;**19**:1900119.

Ma W, Cao Y, Bao W *et al.* ACT-SVM: prediction of protein-protein interactions based on support vector basis model. Huang C (ed.), *Sci Prog* 2020;**2020**:8866557.

Mahlich Y, Steinegger M, Rost B *et al.* HFSP: high speed homology-driven function annotation of proteins. *Bioinformatics* 2018;**34**. doi : 10.1093/bioinformatics/bty262.

Mahood EH, Kruse LH, Moghe GD. Machine learning: a powerful tool for gene function prediction in plants. *Appl Plant Sci* 2020;**8**:e11376.

Makarova KS, Wolf YI, Snir S *et al.* Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol* 2011;**193**. doi : 10.1128/JB.05535-11.

Makrodimitris S, van Ham RCHJ, Reinders MJT. Automatic gene function prediction in the 2020's. *Genes* 2020;**11**. doi : 10.3390/genes11111264.

Markin CJ, Mokhtari DA, Sunden F *et al.* Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. *Science* 2021;**373**. doi : 10.1126/science.abf8761.

McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys* 1943;**5**:115–33.

Mendler K,Chen H,Parks D *et al.*.AnnoTree: visualization and exploration of a functionally annotated microbial tree of life. *Nucleic Acids Research* 2019;**47**:4442–4448. https://doi.org/10.1093/nar/gkz246.

Meng C, Jin S, Wang L *et al.* AOPs-SVM: a sequence-based classifier of antioxidant proteins using a support vector machine. *Front Bioeng Biotechnol* 2019;**7**. doi : 10.3389/fbioe.2019.00224.

Merino GA, Saidi R, Milone DH *et al.* Hierarchical deep learning for predicting GO annotations by integrating protein knowledge. *Bioinformatics* 2022;**38**:btac536.

Miller GF, Todd PM, Hegde SU. Designing neural networks using genetic algorithms. In: *Proceedings of the Third International Conference on Genetic Algorithms*. San Francisco: Morgan Kaufmann Publishers Inc., 1989, 379–84.

Mills CL, Beuning PJ, Ondrechen MJ. Biochemical functional predictions for protein structures of unknown or uncertain function. *Comput Struct Biotechnol J* 2015;**13**:182–91.

Miravet-Verde S, Ferrar T, Espadas-García G *et al.* Unraveling the hidden universe of small proteins in bacterial genomes. *Mol Syst Biol* 2019;**15**:e8290.

Mirceva G, Naumoski A, Kulakov A. Classifying protein structures by using protein ray based descriptor, KNN and FuzzyKNN classification methods. In: *Proceedings of the 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*. Piscataway: IEEE, 2020.

Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**. doi : 10.1093/nar/gkaa913.

Mohamadi A, Cheng T, Jin L *et al.* An ensemble 3D deep-learning model to predict protein metal-binding site. *Cell Rep Phys Sci* 2022;**3**:101046.

Morcinek-Orłowska J, Walter BM, Forquet R *et al.* Protein interaction network analysis reveals growth conditions-specific crosstalk between chromosomal DNA replication and other cellular processes in *E. coli*. *Biorxiv* 2021. doi: 10.1101/2021.12.08.471875.

Moreno-Hagelsieb G, Hudy-Yuffa B. Estimating overannotation across prokaryotic genomes using BLAST+, UBLAST, LAST and BLAT. *BMC Res Notes* 2014;**7**. doi : 10.1186/1756-0500-7-651.

Morris JH, Knudsen GM, Verschueren E *et al.* Affinity purification-mass spectrometry and network analysis to understand protein-protein interactions. *Nat Protoc* 2014;**9**:2539–54.

Myers KS, Park DM, Beauchene NA *et al.* Defining bacterial regulons using ChIP-seq. *Methods* 2015;**86**. doi : 10.1016/j.ymeth.2015.05.022.

Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol* 2020;**16**. doi : 10.1038/s41589-019-0400-9.

Neander K. The teleological notion of 'function'. *Austr J Philos* 1991;**69**. doi : 10.1080/00048409112344881.

Ofer D, Brandes N, Linial M. The language of proteins: NLP, machine learning & protein sequences.*Comput Struct Biotechnol J* 2021;**19**:1750–8.

Oliveira AL. Biotechnology, big data and artificial intelligence. *Biotechnol J* 2019;**14**. doi : 10.1002/biot.201800613.

Omelchenko Mv, Galperin MY, Wolf YI *et al.* Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biol Direct* 2010;**5**. doi : 10.1186/1745-6150-5-31.

Pancino N, Rossi A, Ciano G *et al.* Graph neural networks for the prediction of protein-protein interfaces. In: *Proceedings of the 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. At: Bruges, Belgium2020.

Leinonen R, Akhtar R, Birney E *et al.* The European Nucleotide Archive. *Nucleic Acids Res* 2011;**39**:D28–31.

Parks DH,Chuvochina M,Rinke C *et al.*. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* 2022;**50**: D785–D794. https://doi.org/10.1093/nar/gkab776.

Perakakis N, Yazdani A, Karniadakis GE *et al.* Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. *Metabolism* 2018;**87**:A1–9.

Pfeifer K, Ergal İ, Koller M *et al.* Archaea biotechnology. *Biotechnol Adv* 2021;**47**:107668.

Psomopoulos FE, van Helden J, Médigue C *et al.* Ancestral state reconstruction of metabolic pathways across pangenome ensembles. *Microb Genom* 2020;**6**:1–14.

Punta M, Ofran Y. The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS Comput Biol* 2008;**4**. doi : 10.1371/journal.pcbi.1000160.

Quinlan JR Induction of decision trees. *Mach Learn* 1986;**1**. doi : 10.1007/BF00116251.

Rao R, Bhattacharya N, Thomas N *et al.* Evaluating protein transfer learning with TAPE. *Adv Neural Inf Process Syst* 2019;**32**:676825.

Rehman K, Arslan M, Müller JA *et al.* Bioaugmentation-enhanced remediation of crude oil polluted water in pilot-scale floating treatment wetlands. *Water* 2021;**13**. doi : 10.3390/w13202882.

Reijnders MJMF, Waterhouse RM. CrowdGO: machine learning and semantic similarity guided consensus gene ontology annotation. *PLoS Comput Biol* 2021;**18**:731596.

Rigden DJ, Xos´ X, Fernández XM *et al.* The 2021 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res* 2021;**49**:1–9.

Rives A, Meier J, Sercu T *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci USA* 2020;**118**:622803.

Rojano E, Jabato FM, Perkins JR *et al.* Assigning protein function from domain-function associations using DomFun. *BMC Bioinf* 2022;**23**:43.

Rost B. Twilight zone of protein sequence alignments. *Protein Eng Design Select* 1999;**12**. doi : 10.1093/protein/12.2.85.

Rychel K, Decker K, Sastry Av *et al.* iModulonDB: a knowledgebase of microbial transcriptional regulation derived from machine learning. *Nucleic Acids Res* 2021;**49**:D112–20.

Sáenz-Lahoya S, Bitarte N, García B *et al.* Noncontiguous operon is a genetic organization for coordinating bacterial gene expression. *Proc Natl Acad Sci* 2019;**116**. doi : 10.1073/pnas.1812746116.

Salehinejad H, Baarbe J, Sankar S *et al.* Recent advances in recurrent neural networks. *ArXiv* 2018. doi: 10.48550/ARXIV.1801.01078.

Salzberg SL. Next-generation genome annotation: we still struggle to get it right. *Genome Biol* 2019;**20**. doi : 10.1186/s13059-019-1715-2.

Sandaruwan PD, Wannige CT. An improved deep learning model for hierarchical classification of protein families. *PLoS ONE* 2021;**16**:e0258625–.

Sanderson T, Bileschi ML, Belanger D *et al.* ProteInfer: deep networks for protein functional inference. *Biorxiv* 2021. doi: 10.1101/2021.09.20.461077.

Sarker B, Ritchie DW, Aridhi S. GrAPFI: predicting enzymatic function of proteins from domain similarity graphs. *BMC Bioinf* 2020;**21**:168.

Sastry Av, Y Gao, Szubin R *et al.* The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat Commun* 2019;**10**. doi: 10.1038/s41467-019-13483-w.

Satpute BS, Yadav R. Decision tree classifier for classification of proteins using the Protein Data Bank. In: Krishna AN, Srikantaiah KC, Naveena C (eds). *Integrated Intelligent Computing, Communication and Security*. Singapore: Springer, 2019, 71–8.

Schietgat L, Vens C, Struyf J *et al.* Predicting gene function using hierarchical multi-label decision tree ensembles. *BMC Bioinf* 2010;**11**:2.

Sequeira AM, Lousa D, Rocha M. ProPythia: a Python package for protein classification based on machine and deep learning. *Neurocomputing* 2022;**484**:172–82.

Serres MH, Gopal S, Nahum LA *et al.* A functional update of the *Escherichia coli* K-12 genome. *Genome Biol* 2001;**2**. doi : 10.1186/gb-2001-2-9-research0035.

Shenhav L, Thompson M, Joseph TA *et al.* FEAST: fast expectation-maximization for microbial source tracking. *Nat Methods* 2019;**16**:627–32.

Shirmohammady N, Izadkhah H, Isazadeh A. PPI-GA: a novel clustering algorithm to identify protein complexes within protein-protein interaction networks using genetic algorithm. *Complex* 2021;**2021**:1–14.

Singh J, Vyas A, Wang S *et al. Microbial Biotechnology: Basic Research and Applications*. Berlin: Springer, 2020.

Smith LM, Kelleher NL, Consortium for Top Down Proteomics, Proteoform: a single term describing protein complexity. *Nat Methods* 2013;**10**:186–7.

Sommer MJ, Salzberg SL. Balrog: a universal protein model for prokaryotic gene prediction. *PLoS Comput Biol* 2021;**17**:e1008727.

Sperduti A, Starita A. Supervised neural networks for the classification of structures. *IEEE Trans Neural Netw* 1997;**8**:714–35.

Ingolia N. Ghaemmaghami S Newman J *et al.* Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science* 2009;**324**:218–223.

Stephens ZD, Lee SY, Faghri F *et al.* Big data: astronomical or genomical?. *PLoS Biol* 2015;**13**. doi : 10.1371/journal.pbio.1002195.

Storz G, Wolf YI, Ramamurthi KS. Small proteins can no longer be ignored. *Annu Rev Biochem* 2014;**83**. doi : 10.1146/annurev-biochem-070611-102400.

Strodthoff N, Wagner P, Wenzel M *et al.* UDSMProt: universal deep sequence models for protein classification. *Bioinformatics* 2020;**36**:2401–9.

Subramanian I, Verma S, Kumar S *et al.* Multi-omics data integration, interpretation, and its application. *Bioinf Biol Insights* 2020;**14**:1177932219899051.

Suravajhala P, Sundararajan V. A classification scoring schema to validate protein interactors. *Bioinformation* 2012;**8**:34–9.

Sureyya Rifaioglu A, Doğan T, Jesus Martin M *et al.* DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. *Sci Rep* 2019;**9**. doi : 10.1038/s41598-019-43708-3.

Swenson N, Krishnapriyan AS, Buluc A *et al.* PersGNN: applying topological data analysis and geometric deep learning to structure-based protein function prediction. *Arxiv e-prints* 2020. doi: 10.48550/arXiv.2010.16027.

Taju SW, Nguyen T-T-D, Le N-Q-K *et al.* DeepEfflux: a 2D convolutional neural network model for identifying families of efflux proteins in transporters. *Bioinformatics* 2018;**34**:3111–7.

Tatarinova Tv, Chekalin E, Nikolsky Y *et al.* Nucleotide diversity analysis highlights functionally important genomic regions. *Sci Rep* 2016;**6**. doi : 10.1038/srep35730.

Teletin M, Czibula G, Albert S *et al.* Using unsupervised learning methods for enhancing protein structure insight. *Proc Comput Sci* 2018;**126**:19–28.

Thompson NC, Greenewald K, Lee K *et al.* The computational limits of deep learning. arXiv preprint. 2020. doi: 10.48550/arXiv.2007.05558.

Vanni C, Schechter MS, Delmont TO *et al.* AGNOSTOS-DB: a resource to unlock the uncharted regions of the coding sequence space. *Biorxiv* 2021. doi: 10.1101/2021.06.07.447314.

Vapnik VN. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

Vaswani A, Shazeer N, Parmar N *et al.* Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Red Hook: Curran Associates Inc., 2017, 6000–10.

Wang L, Law J, Murali TM *et al.* Data integration through heterogeneous ensembles for protein function prediction. *Biorxiv* 2020. doi: 10.1101/2020.05.29.123497.

Wang L, Yang J, Xu Y *et al.* Domain-based comparative analysis of bacterial proteomes: uniqueness, interactions, and the dark matter. *Curr Genomics* 2019;**20**. doi : 10.2174/1389202920666190320134438.

Whalen S, Schreiber J, Noble WS *et al.* Navigating the pitfalls of applying machine learning in genomics. *Nat Rev Genet* 2022;**23**:169–81.

Wiegand S, Dam HT, Riba J *et al.* Printing microbial dark matter: using single cell dispensing and genomics to investigate the patescibacteria/candidate phyla radiation. *Front Microbiol* 2021;**12**. doi : 10.3389/fmicb.2021.635506.

Xia W, Zheng L, Fang J *et al.* PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput Biol Med* 2022;**145**:105465.

Yang F, Fan K, Song D *et al.* Graph-based prediction of protein-protein interactions with attributed signed graph embedding. *BMC Bioinf* 2020;**21**:323.

You R, Yao S, Mamitsuka H *et al.* DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics* 2021;**37**:i262–71.

You R, Yao S, Xiong Y *et al.* NetGO: improving large-scale protein function prediction with massive network information. *Nucleic Acids Res* 2019;**47**:W379–87.

You R, Zhang Z, Xiong Y *et al.* GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 2018;**34**:2465–73.

Yu G, Stoltzfus A. Population diversity of orfan genes in *Escherichia coli*. *Genome Biol Evol* 2012;**4**. doi: 10.1093/gbe/evs081.

Zeng M, Li M, Fei Z *et al.* A deep learning framework for identifying essential proteins by integrating multiple types of biological information. *IEEE/ACM Trans Comput Biol Bioinform* 2021;**18**: 296–305.

Zeng M, Li M, Wu F-X *et al.* DeepEP: a deep learning framework for identifying essential proteins. *BMC Bioinf* 2019;**20**:506.

Zha Y, CChen , Jiao Q *et al.* Ontology-Aware Deep Learning Enables Novel Antibiotic Resistance Gene Discovery Towards Comprehensive Profiling of ARGs. *Biorxiv* 2021. doi: 10.1101/2021.07.30.454403.

Zha Y, Chong H,Qiu H *et al.* Ontology-aware deep learning enables ultrafast and interpretable source tracking among sub-million microbial community samples from hundreds of niches. *Genome Med* 2022. doi: 10.1186/s13073-022-01047-5.

Zhan X-K, You Z-H, Li L *et al.* Using random forest model combined with Gabor feature to predict protein-protein interaction from protein sequence. *Evol Bioinformatics* 2020;**16**:117693432093449.

Zhang R,Xu W,Shao S *et al.* .Gene Silencing Through CRISPR Interference in Bacteria: Current Advances and Future Prospects , *Frontiers in Microbiology* 2021;**12**. 10.3389/fmicb.2021.635227 .

Zhang Z, Wang J, Wang J *et al.* Estimate of the sequenced proportion of the global prokaryotic genome. *Microbiome* 2020;**8**. doi : 10.1186/s40168-020-00903-z.

Zhao C, Liu T, Wang Z. PANDA2: protein function prediction using graph neural networks. *NAR Genom Bioinform* 2022;**4**:lqac004.

Zhou C, Yu H, Ding Y *et al.* Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree. *PLoS ONE* 2017;**12**:e0181426-.

Zhou Y, Zheng H, Huang X. Graph neural networks: taxonomy, advances, and trends. *ACM Trans Intell Syst Technol* 2022;**13**:1–54.

Zohra Smaili F, Tian S, Roy A *et al.* QAUST: protein function prediction using structure similarity, protein interaction, and functional motifs. *Genomics Proteomics Bioinformatics* 2021;**19**. doi: 10.1016/j.gpb.2021.02.001.

Zuo Y-C, Su W-X, Zhang S-H *et al.* Discrimination of membrane transporter protein types using K-nearest neighbor method derived from the similarity distance of total diversity measure. *Mol Biosyst* 2015;**11**: 950–7.