International Journal of
Molecular Sciences

MDPI

Review

Incorporating Machine Learning into Established Bioinformatics Frameworks

Noam Auslander *,†, Ayal B. Gussow † and Eugene V. Koonin *

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA; ayal.gussow@nih.gov
* Correspondence: noam.auslander@nih.gov (N.A.); koonin@ncbi.nlm.nih.gov (E.V.K.)
† Co-first authors.

Abstract: The exponential growth of biomedical data in recent years has urged the application of numerous machine learning techniques to address emerging problems in biology and clinical research. By enabling the automatic feature extraction, selection, and generation of predictive models, these methods can be used to efficiently study complex biological systems. Machine learning techniques are frequently integrated with bioinformatic methods, as well as curated databases and biological networks, to enhance training and validation, identify the best interpretable features, and enable feature and model investigation. Here, we review recently developed methods that incorporate machine learning within the same framework with techniques from molecular evolution, protein structure analysis, systems biology, and disease genomics. We outline the challenges posed for machine learning, and, in particular, deep learning in biomedicine, and suggest unique opportunities for machine learning techniques integrated with established bioinformatics approaches to overcome some of these challenges.

Keywords: machine learning; deep learning; bioinformatics methods; phylogenetics

Citation: Auslander, N.; Gussow, A.B.; Koonin, E.V. Incorporating Machine Learning into Established Bioinformatics Frameworks. *Int. J. Mol. Sci.* **2021**, *22*, 2903. https://doi.org/10.3390/ijms22062903

Academic Editor: Jung Hun Oh

Received: 15 February 2021
Accepted: 10 March 2021
Published: 12 March 2021

1. Introduction

Over the past few decades, the advances in computational resources and computer science, combined with next-generation sequencing and other emerging omics techniques, ushered in a new era of biology, allowing for sophisticated analysis of complex biological data. Bioinformatics is evolving as an integrative field between computer science and biology, that allows the representation, storage, management, analysis and investigation of numerous data types with diverse algorithms and computational tools. The bioinformatics approaches include sequence analysis, comparative genomics, molecular evolution studies and phylogenetics, protein and RNA structure prediction, gene expression and regulation analysis, and biological network analysis, as well as the genetics of human diseases, in particular, cancer, and medical image analysis [1–3].

Machine learning (ML) is a field in computer science that studies the use of computers to simulate human learning by exploring patterns in the data and applying self-improvement to continually enhance the performance of learning tasks. ML algorithms can be roughly divided into supervised learning algorithms, which learn to map input example into their respective output, and unsupervised learning algorithms, which identify hidden patterns in unlabeled data. The advances made in machine-learning over the past decade transformed the landscape of data analysis [4–6]. In the last few years, ML and particularly deep learning (DL) have become ubiquitous in biology (Figure 1). However, clinical applications have been limited, and follow-up mechanistic investigation of ML-based predictions is often lacking, due to the difficulty in the interpretation of the results obtained with these techniques. To overcome these problems, numerous approaches have been developed to incorporate ML and DL into established bioinformatics frameworks, for training data selection and preparation, identification of informative features, or data integration. Such

integrated frameworks exploit the power of ML and DL methods, offering interpretability and reproducibility of the predictions.

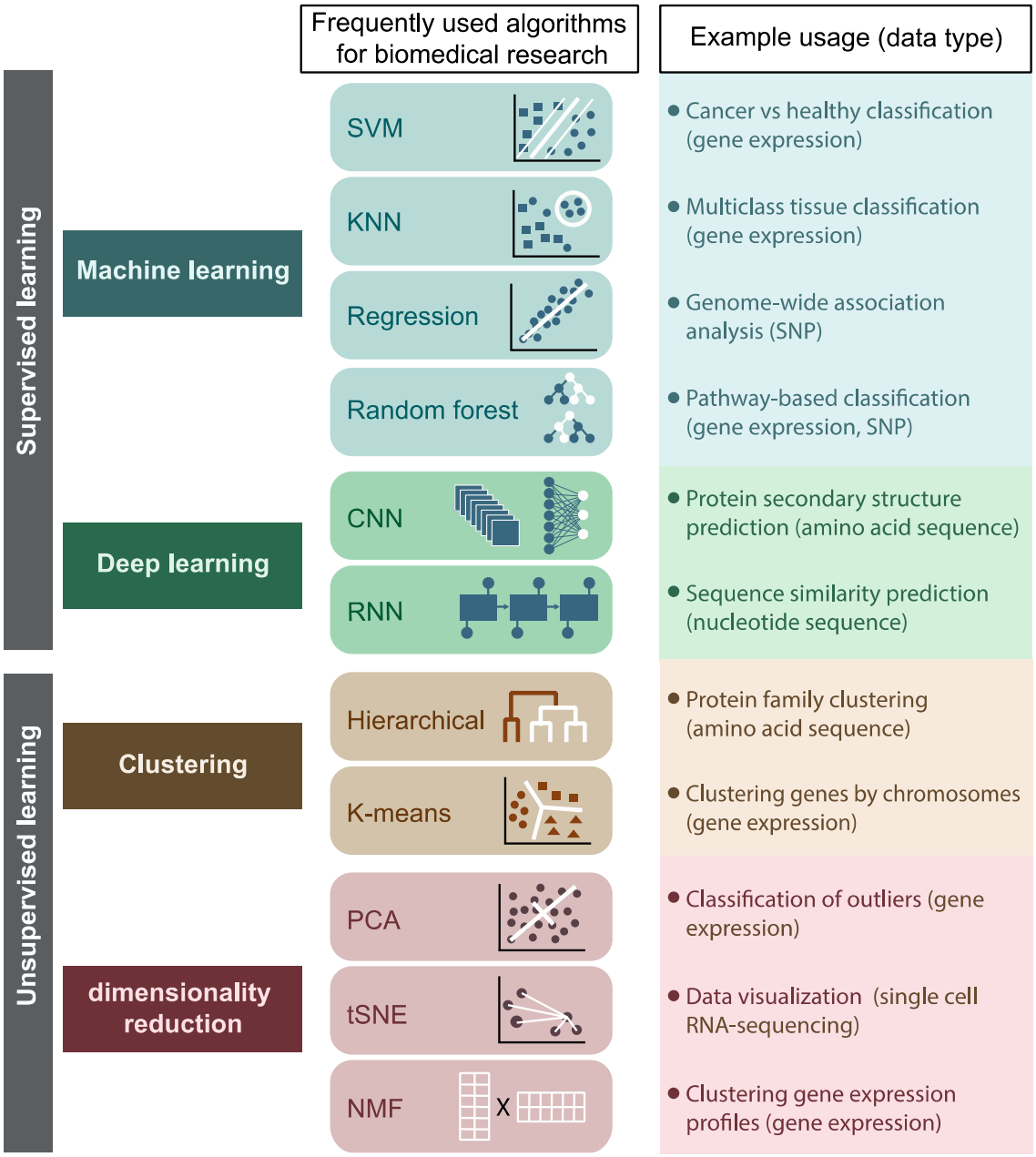

Figure 1. Machine learning algorithms frequently used in bioinformatics research. An example of the usage of each algorithm and the respective input data are indicated on the right. Abbreviations: SVM, support vector machines; KNN, K-nearest neighbors; CNN, convolutional neural networks; RNN, recurrent neural networks; PCA, principal component analysis; t-SNE, t-distributed stochastic neighbor embedding, NMF, non-negative matrix factorization.

In this brief review, we survey recent efforts to integrate ML and DL with established bioinformatic methods, across four areas in computational biology. We discuss the strengths and limitations of these integrated methods for specific applications and propose avenues to address the challenges impeding even broader application of ML techniques in biomedical research.

1.1. Integrating Machine-Learning into Molecular Evolution Research

Combining computer science approaches with principles of molecular evolution analysis has revolutionized the field of molecular evolutionary studies. Application of diverse and increasingly advanced computational methods has enabled accurate determination of evolutionary distances between species, reconstruction of evolutionary histories and ancestries, identification conserved genomic regions, functional annotation of genomes, and phylogenetics. In recent years, ML methods have been developed to address the challenges faced by molecular evolution research, in particular, by overcoming the difficulties of analyzing increasingly massive sets of sequence and other omics data. Examples of such applications include the use of autoencoders to impute incomplete data for phylogenetic tree construction [7], application of random forest for phylogenetic model selection [8], harnessing convolutional neural networks (CNNs) to infer tree topologies [9] and tumor phylogeny [10], and utilization of deep reinforcement learning for the construction of robust alignments of many sequences [11].

Evolutionary algorithms and strategies have been the most successful in solving diverse bioinformatic problems, far beyond core phylogenetic and molecular evolution tasks. Indeed, a wide range of computational techniques are founded on evolutionary strategies, including application of population-based analysis, fitness-oriented rules or variation-driven research [12,13]. For instance, genetic algorithms (GA) [14] are a type of search heuristic which is inspired by principles of biological evolution. The GA is widely used in for optimization of multiple criteria and for features selection [15–17]. Evolutionary approaches underly effectively all types of biological sequence analysis. Therefore, integrating ML with molecular evolution and phylogenetic methods is essential to uncover robust and biologically relevant patterns and discriminative features. For example, recent methods combined sequence attributes, alignment, and phylogenetic trees with ML for protein sequence analysis and clustering [18,19] for tasks as different as identification of determinants of viral pathogenicity and infectivity [20–22], prediction CRISPR-Cas9 cleavage efficiency [23] and detection of anti-CRISPR proteins [24,25].

Although numerous bioinformatics methods continue to rely on sequence alignments, the advent of ML gave rise to a variety of alignment-free methods that allow skipping the alignment step and learning directly from unaligned sequences. Alignment-free methods are especially useful, for example, for the identification of viral sequences in complex sequence datasets, where highly divergent viruses are often difficult to identify with straightforward alignment and sequence comparison. Therefore, alignment-free tools have been developed for viral sequence identification by employing ML techniques such as SVM [26], RNN [27] and CNN [28,29]. Alignment-free methods are also useful for the functional annotation of nucleic acids and proteins, where in some cases function may be inferred from particular domains or motifs that can be detected without complete nucleotide or protein alignment. In cases where sequence profiles are difficult to derive, ML and particularly DL techniques can be trained to rapidly recognize specific domains or motifs, without the need to devise explicit sequence profiles [29–31]. Several DL techniques have been employed for the annotation of functional features in nucleotide sequences, typically relying on a large, annotated sequence dataset for training, for example, using deep RNN [32,33] or CNN [34,35]. These applications include identification of promoters [36,37], enhancers [38,39], long noncoding RNAs [40–44], microRNA targets [45,46], and CRISPR arrays [47].

The key challenge in the application of ML to molecular evolution and phylogenetics, where traditional bioinformatic strategies efficiently resolve many substantial problems, is the identification of tasks that have not been yet properly addressed, but involve learnable patterns and features. This challenge stems from the difficulty of estimating the learnability of different problems, but also, from the shortage of labeled datasets of sufficient size for problems that are not easily amenable to standard bioinformatic techniques.

1.2. Integrating Machine-Learning with Protein Structure Analysis

In the study of proteins, numerous methods have been developed to process the amino acid sequence, and predict the protein structure, function and post translational modifications, such as phosphorylation and glycosylation, that are crucial to the function of many if not most proteins. ML techniques have been incorporated with traditional proteomic methods to predict and analyze post translational modifications [48,49], including CNN [50], hierarchical and K-means clustering [51,52]. The Musite suite integrated KNN with the search for local sequence similarity to known phosphorylation sites, protein disorder scoring and amino acid frequency calculation to predict general and kinase-specific phosphorylation sites [53]. EnsembleGly developed an ensemble classifier of protein glycosylation site based on curated glycosylated protein database and SVM [54]. More recently, several DL models have been incorporated with other modeling techniques and curated databases for the prediction of phosphorylation sites [50,55], and protein glycosylation [56].

Fundamental computational challenges in the field of protein analysis include prediction of protein structure from sequence, accurate estimation of structural similarity to infer homology and prediction of protein contact maps [57,58]. Solving these problems is crucial for the characterization of protein functions, localization and interactions, and can directly contribute to many research directions, from deciphering evolutionary history [59] to drug discovery [60]. Existing computational methods for protein structure prediction that rely on thermodynamics, molecular mechanics, heuristics, and similarity to previously solved structures have demonstrated varying levels of success [61–63]. ML and particularly DL techniques have recently entered this field but have already shown the potential to revolutionize protein structure prediction, inference of homology from structure comparison and estimation of contact maps.

Numerous ML methods have been developed for protein structural prediction, with particular success achieved with deep learning architectures [58,64]. The Critical Assessment of Structure Prediction (CASP), which assesses prediction methods and models [64], recently noted substantial progress in structure modeling by deep learning, in particular, template free modeling (FM), that is, modeling structure without an existing template, as opposed to homology modeling. Numerous deep learning methods now require fewer proteins in the input MSA and have demonstrated increasing success in FM modelling [65–71], primarily due to more precise prediction of contact maps and inter-residue distances [64]. Some methods are narrower in scope and focus on contact prediction [72–75]. The strongest predictor for CASP13, the most recent CASP with a published report, was AlphaFold [76,77], a deep learning predictor from DeepMind. The results from CASP14 have not been yet described in detail but are available online [78]. CASP14 was marked with the striking success of AlphaFold2, the next version of AlphaFold, which integrates established sequence search tools into a deep learning framework. AlphaFold2 employs sequence database search to construct multiple sequence alignments (MSA), and extracts MSA-based features that are given as input to a deep residual convolutional neural network [79]. This network architecture eases the training of deep networks by introducing shortcut connections with gating functions, that avail the input of lower layers to higher layer nodes in the network. In CASP14, AlphaFold2 vastly outperformed every other method, both FM and template-based modeling approaches. The results of AlphaFold2 are so impressive that there seems to be a realistic possibility that this computational approach could begin to replace the expensive and time-consuming protein crystallography and even the more efficient cryo-EM. Regardless of whether and when this promise materializes, it is becoming clear that DL has already revolutionized protein structure analysis, and rapid and broad improvements can be expected to occur in the next few years (Figure 2).

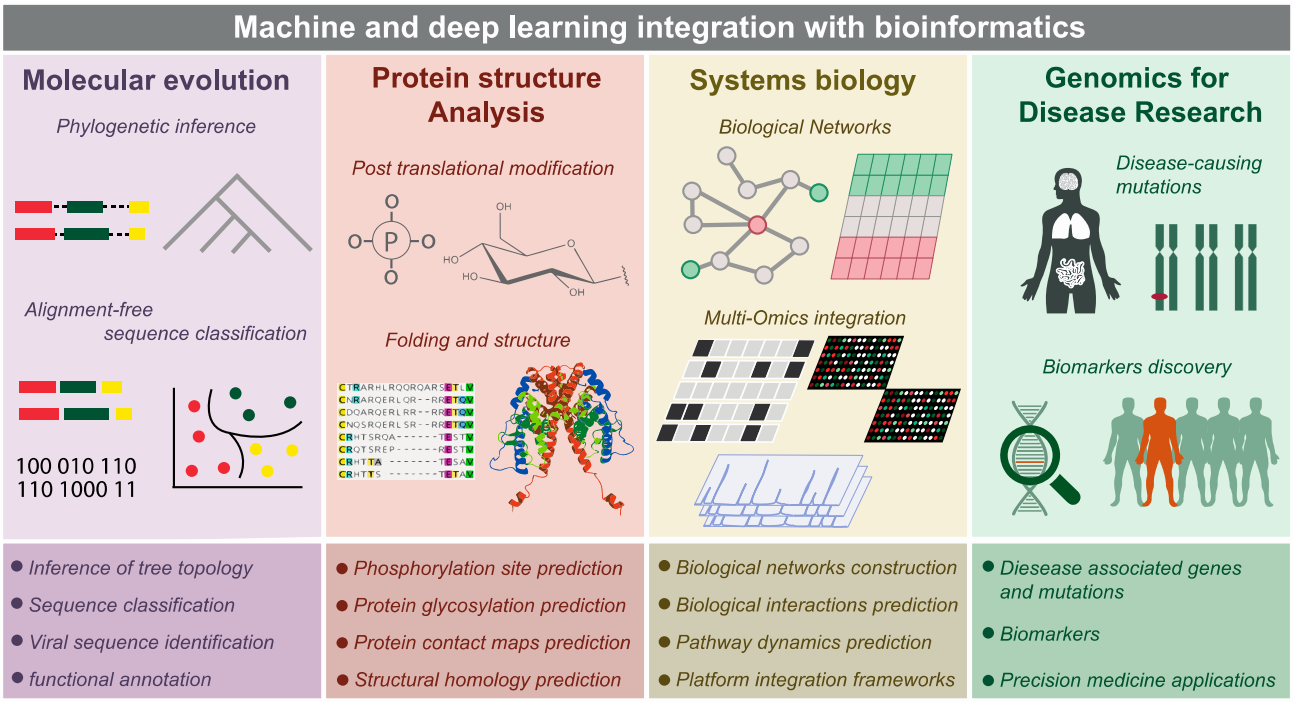

Figure 2. Applications of integrated machine learning techniques with bioinformatics in molecular evolution, protein structure analysis, systems biology, and disease genomics.

1.3. Integrating ML into Systems Biology

The rapid growth and diversification of biological data calls for an increasingly wide range of modeling and analysis techniques to be employed in systems biology. With complex omics datasets that are now incessantly accumulating, there is a growing need for techniques that can integrate different data types, incorporate datasets into established biological networks and combine different systems biology approaches to investigate multi-omics datasets. Various ML methods have been developed to utilize multi-dimensional datasets together with biological networks, study complex interactions and model biological systems. ML techniques in network biology can be classified into those that infer the network architecture and those that integrate existing network architectures with biological data measurements [80]. Consequently, some of these techniques also require sophisticated data integration methods to incorporate different data types into a model.

Different ML frameworks have been utilized for the inference of biological networks, such as the gene regulatory network (GRN) in the DREAM5 project [81] which utilizes SIRENE [82], a support vector machines-based approach for regulatory networks utilization. More recently, a transfer learning technique [83], and a single-cell RNA sequencing based ML technique [84] have also been proposed for GRN reconstruction. ML methods also have been employed for the inference of protein-protein interactions (PPI) networks, for example, by utilizing NMF [85], regression [86], PCA [87] and deep neural networks [88]. Such methods include the recently developed signed variational graph auto-encoder [89], a graph representation learning method that incorporated graph structure and sequence information to study PPI networks, PPI_SVM [90], which integrated support vectors machines with domain affinity and frequency tables, and LightGBM-PPI [91], which utilizes elastic net regression models with different protein descriptors for inference of PPI networks. In addition, several DL-based techniques have been proposed for PPI network reconstruction [88,92–95]. These methods primarily exploit recent advances in deep learning architectures to enhance the prediction of PPI networks [93]. Network inference techniques were additionally developed to advance disease research, and several ML techniques have been developed to identify drug-target interaction networks using drug similarity [96,97], by integrated K-means clustering with network analysis [98], or by integrating different

networks and data types [99,100]. Several DL based techniques have been developed to predict drug response based on cell line data [101,102], by integrating genomic profiles [103], or through multi-omics integration [104]. Some methods incorporate chemical properties of compounds with ML to predict their clinical effects [105–107] and recently, a cancer network inference technique has been proposed to identify signal linkers which coordinate oncogenic signals between mutated and differentially expressed genes [108].

ML methods have also been incorporated with established network structures to analyze diverse biological datasets. ML techniques have been incorporated with biological networks to predict anti-cancer drug efficacy [109], to model drug response by integrating prior biological knowledge with different biological data types [110], and by computing "network profiles" based on PPI networks [111]. Several strategies have been proposed to employ ML for network-based prediction of drug side effects [112–114] and drug combinations [115], for prediction of synergistic drugs [116,117] and drug repositioning [118–120]. Several studies have used machine and deep learning techniques to investigate properties of metabolic networks, such as inference of metabolic pathways [121,122], differential metabolic activity [123] and pathway reconstruction [124,125]. A variety of studies have integrated information obtained for different data types using ML methods, including the integration of network and pathway data for the discovery of drug targets [123,126,127], incorporation of a pathway-derived mechanistic model with gene expression to identify new drug targets [128], and inference of the activity of oncogenic pathways in cancer [129,130]. Recent strategies integrate multi-omics datasets with ML techniques to enhance the prediction of pathway dynamics [131] and utilize pathway based multi-omics integration for patient clustering [132].

With the recent increased availability of multiple, powerful omics techniques (that is, genomics, transcriptomics, proteomics, and metabolomics), a key emerging challenge is the integration of different omics platforms. Several methods have been developed for multi-omics integration using machine and deep learning techniques [133], including SVM [134,135], KNN [136,137], NMF [138], PCA [139] and CNN [140], for example, for cancer subtype and survival prediction [141–143] and for prediction of drug response [143,144], the paucity of studies systematically comparing different multi-omics integration methods is a serious bottleneck in the advancement of this field. Such systematic comparison was recently performed for a subset of the multi-omics techniques aimed at the prediction of tumor subtype [145]. The lack of standardized techniques and clear recommendation of methods to use for particular applications may lead to inadequate selection of analysis strategy and overfitting [146].

1.4. Integrating ML with Genomics and Biomarker Analysis for Disease Research

In recent years, molecular phenotyping using genetic and genomic information has allowed early and accurate prediction and diagnosis of many diseases, and critically improved clinical decision making [147,148]. In disease research, the key challenges are the identification of disease-associated genes and mutations for diagnosis, and prediction of the disease progression and clinical outcome as well as drug response and personalized medicine.

Traditional algorithms for the identification of disease-associated genes and disease-causing mutations mostly rely on analysis of sequence data, which can be limited for rare diseases. In addition, some diseases are caused by epigenetic alterations, and thus are not linked to specific mutations or genetic variation. Therefore, several techniques have been developed to identify genes that are associated with complex diseases by incorporating machine and deep learning methods with different types of data, biological networks and bioinformatic techniques. For example, incorporation of network analysis of differentially expressed genes with ML allowed the prioritization of disease-genes even without disease phenotype information [149], hierarchical clustering analysis to differentially expressed genes revealed genes associated with pulmonary sarcoidosis [150], and integration of non-negative matrix factorization (NMF) with disease semantic information and miRNA

functional information uncovered new miRNA-disease association [151]. Other examples include training machine learning classifiers on gene functional similarities inferred with Gene Ontology (GO) resulting in successful identification of genes associated with the Autism Spectrum Disorder [152], and applying ML to features calculated based on protein sequences, allowing inference of the probability of a protein's involvement in disease, without considering their function or expression [153]. Furthermore, recently developed algorithms allow ML-based visualization of disease relationships, for example, of disease-phenotype similarity and disease relationships with t-SNE [154,155]. In addition, ML has been integrated with PPI networks to infer a phenotype similarity score and rank protein complexes by phenotypes that are linked to human disease [156], to identify topological features of disease-associated proteins [157], and recently, to identify host genes that are associated with infectious diseases [158]. Furthermore, ML algorithms have been employed for the detection and investigation of cancer driver genes, by incorporation of ML with statistical scoring of genomic sequencing [159], pathway-level mutations [160], mutation and gene interaction data [161], and by application of deep convolutional neural networks and random forest for analysis of mutations and gene similarity networks [162,163].

A biomarker is a biological measure that can be used as an indicator of a disease state or response to therapeutic interventions [164,165]. There are three categories of disease biomarkers. First, risk biomarkers are used to identify patients that are at risk of developing a disease. Second, diagnostic biomarkers help detect a disease state and determine the disease category. Third, prognostic biomarkers help predict disease progression, response to treatment and recurrence [166]. Various ML approaches, and in particular, feature selection methods have been applied to discover molecular biomarkers and classify clinical cases. For example, an approach for the discovery of biomarker signatures has been proposed based on a pipeline that applies feature selection through integration of different data types with biological networks [167]. Several machine learning techniques have been developed for biomarker discovery in cancer, by using protein biomarkers to classify cancer states [168], and developing biomarkers for early cancer diagnosis from microarray and gene expression data [169–172], urine metabolomics [173,174] and multidimensional omics data [175–177]. Several methods have been developed that integrate network information with omics data for biomarker discovery [167,175,178], and some methods incorporated prior knowledge into feature selection algorithms for biomarker discovery, such as diseases associated genes [179,180], evolutionary conservation [179,181], pathway information [182–184], and by applying network feature selection [185,186]. Recently, ML techniques were proposed to develop biomarkers that match patients to treatments, such as identification of markers that correlate with enhanced drug sensitivity [103,109,187], and treatment recommendations with SVM [188] and RNN [189] (Table 1).

1.5. Key Challenges and Future Directions

ML methods including, recently, DL algorithms have become a rapidly growing research area, redefining the state-of-the-art performance for a wide range of fields [4,5]. Given the rapid growth in the availability of biomedical and clinical datasets in the past decades, these techniques can be expected to similarly transform multiple avenues of biomedical research, and indications of their high efficacy are already accumulating. The success of AlphaFold2 that dramatically outperforms all other existing methods for protein structure prediction from amino acid sequences [77] is perhaps the strongest case in point. It appears more than likely that similar efforts will result in breakthroughs in a variety of biomedical fields through the integration of ML with more traditional bioinformatics approaches. However, there are several key obstacles that have to be overcome to enable the development and acceptance of ML solutions to pressing problems in biomedicine. We discuss some of the most substantial challenges and suggest means to overcome them through integration of ML frameworks with prior biological knowledge, databases, and established bioinformatics techniques (Table 2).

Table 1. Representative problems and methods addressing them by incorporating machine learning (ML) with bioinformatics tools in four areas.

Bioinformatics Area	Problem Category	Goal	ML Method	Bioinformatic Tools
Molecular evolution	Biological sequence clustering	Protein family prediction	CNN	Clusters of Orthologous Groups (COGs) and G protein-coupled receptor (GPCR) dataset [30]
		Protein function prediction	deep RNN	BLAST and HMMER search [32]
		Anti-CRISPR proteins identification	Random forest	MSA and PSI-BLAST [24]
			EXtreme Gradient Boosting	K-mer based clustering (CD-HIT), BLAST [25]
		Viral pathogenicity feature identification	SVM	MSA, phylogenetic tree construction [20,21]
	Alignment free biological sequence analysis	Identification of viral genomes	RNN	BLAST, Sequence clustering, HHPRED [27]
			CNN	BLAST [28]
protein structure analysis	Post translational modifications	Phosphorylation sites prediction	KNN	Local sequence similarity [53]
			CNN	K-mer based clustering (CD-HIT), BLAST [55]
		Glycosylation sites prediction	ensemble SVM	curated glycosylated protein database (O-GLYCBASE) [54]
	Protein structure prediction	Protein contact prediction	CNN	MSA [72]
		Prediction of distances between pairs of residues	CNN	MSA, HHPRED, PSI-BLAST [77]
systems biology	inference of biological networks	Gene regulatory network prediction	SVM	GeneNetWeaver, RegulonDB [81]
		Protein-protein interaction network prediction	SVM	Domain affinity and frequency tables [90]
			Elastic-net regression	Protein descriptors [91]
	Analysis of biological networks	Drug target prediction	K-means	Network analysis tools [98]
		Drug side effect prediction	SVM	Genome scale metabolic modeling [112]
		Drug Synergism prediction	Random Forest Ensemble	A chemical-genetic interaction matrix [117]
	Multi-omics integration	Cancer subtype prediction	Neighborhood based clustering	Similarity based integration [141]
		Drug response prediction	logistic regression	Cancer hallmarks datasets, pathway data [144]
biomarker analysis for disease research	Disease-associated genes investigation	Pulmonary sarcoidosis genes identification	Hierarchical clustering	Differential expression analysis [150]
		Identification of miRNA-disease association	NMF	Disease semantic information and miRNA functional information [151]
		Disease-phenotype visualization	t-SNE	OMIM database and human disease networks [154]
	Biomarker discovery	Cancer diagnosis	SVM	Reference gene selection [170]
		Biomarker signature identification	SVM	Network-based gene selection [167]
		Cancer outcome prediction	Random forest	Evolutionary conservation estimation [181]

Table 2. Challenges posed for ML and DL in biomedicine, existing strategies to overcome these challenges and proposed solutions by integrating ML techniques with established bioinformatics approaches.

Problem	Bottleneck	Example Solutions	Potential Integrated ML/DL and Bioinformatics Solutions
Small and dependent datasets	Data availability	Restricting the number of parameters [27,190]	Neural network architectures for small and sparse datasets
		Separating training and test sets by phylogenetic similarity [27]	Methods to evaluate data dependency by protein and sequence similarities
Biological sequence representation	Methodological	NLP with neural networks-based modeling [191–194]	Incorporating amino acid substitution and codon usage matrices to representation frameworks
			Incorporating conserved domain databases to the training framework
Incorporation of different data types	Methodological	Integration of multi-omics datasets through existing network topologies	
Reproducibility	Acceptance	Documentation and deposition of the processed data [195]	-
		Benchmarking of the processing pipeline and optimized parameters [196]	-
Interpretability	Acceptance	Incorporation of established bioinformatic methods and databases with ML and DL frameworks [128,196]	
		Generation of interpretable DL models [197–199]	

A major challenge for the application of ML and particularly DL to biological sequences is the representation of nucleotides or amino acid sequences as a sequence of numbers or vectors. Representation of biological sequences as well as feature extraction methods for genetic, molecular and clinical data are imperative for the subsequent successful application of ML and DL techniques. The leading method developed for biological sequence representation is BioVec [191], which includes GeneVec, a representation of gene sequences, and ProtVec that represents protein sequences. BioVec relies on the Word2Vec algorithm [200], a natural language processing (NLP) technique that employs a neural network-based model, and is applied to n-gram representations of the protein sequence. This approach has been applied to protein family classification and visualization of proteins [191]. More recent methods for distributed representation of biological data operate by learning gene co-expression patterns [192], representation of cancer mutations [193], and representation of residue-level sequences for kinase specific phosphorylation site prediction [194]. These efforts are almost entirely data-driven, and do not make use of the curated databases and bioinformatic tools that are widely employed for the analysis of biological sequences. For example, well established matrices that have been designed to evaluate amino acid substitutions [201] and codon usage [202] could be considered when encoding biological sequences. Furthermore, numerous manually curated conserved domains databases that document functional and structural units of proteins [203] could be integrated into the training and evaluation steps of DL frameworks for protein annotation and functional classification. Incorporation of curated databases and established bioinformatic matrices into sequence representation methods is expected to enhance the training, evaluation and interpretability of DL models.

One consequence of the lack of efficient protein sequence representation is a frequent use of the simplest, assumption-free representation, which is one-hot encoding, where each position in a sequence is represented by a 20-dimensional vector with 19 positions set to 0 and the position identifying a specific amino acid set to 1. Although the one-hot representation can sometimes outperform other scales [204], one-hot encoded protein sequences are sparse, memory-inefficient and high-dimensional [205]. In addition, one-hot encoding lacks the notion of similarity between sequences, and thus, is more appropriate

for categorical data with no relationship between the categories [205]. This could be a particularly severe problem when a one-hot representation is given to a convolutional neural network. Most convolutional layers identify spatial patterns in the data, which the one-hot encoding inherently lacks. By using a sparse, one-hot encoded protein sequences, a deep convolutional network can wrongly infer similarity patterns and spatial connections between amino acids, which could be meaningless and could lead to overfitting [206,207]. In addition, a convolution is more likely to capture local and proximal patterns and dismiss long-range patterns [208], which is problematic for any sparse representation, but especially, when long-range dependencies are known or suspected to exist in the data. Therefore, it is crucial to carefully consider the appropriate data representation and neural network architecture for every prediction problem.

Despite the advent of the big data era, for many major challenges in biomedicine, the available data are small, sparse, and highly dependent. This is a major problem for training DL models, which require massive amounts of training data and an independent test set. Biological data, and especially biological sequence databases, tend to include high proportion of duplicate or near-duplicate samples [209], which can seriously bias learning algorithms, especially when duplicates are present between the training and test datasets [210–212]. For the training and evaluation of DL algorithms on highly dependent biological data, careful data processing is needed to minimize duplicates and near-duplicates and ensure independence between the training and test sets [27,213]. With the growing availability and appeal of DL frameworks, the issues of sample size and the independence of biological data are frequently ignored, so that large-scale models are trained without data filtering and preparation, and therefore without ever being evaluated on a truly independent test. To overcome these limitations, it is necessary to develop neural network architectures that are specifically designed for small and sparse datasets [27,214,215]. In addition, there is a pressing need for the development of methods that estimate the dependencies between biological samples using existing bioinformatics techniques (such as clustering of nucleic acid and proteins by sequence similarity), with subsequent evaluation of the maximum model size and the number of parameters given the true size of independent samples.

Another important challenge in biomedical applications of ML is the difficulty in incorporating different data types. With the growing availability of multi-omics datasets that combine genomics, transcriptomics, metabolomics and proteomics data, there is a pressing need for systematic evaluation of the strategies for multi-omics integration techniques, and for the assessment and development of learning algorithms that can be applied to integrated datasets. In particular, methods are required for data reduction, visualization, and feature selection that allow a combined view and evaluation of integrated multi-omics datasets. Integration of multi-omics datasets through incorporation of curated network topology can enhance the development of multi-omics ML pipelines, and provide means for feature connection, selection and reduction based on established biological networks.

Reproducibility is another major issue that has been extensively discussed in the context of biomedical applications of ML and other computational techniques [216]. Code sharing and open-source licensing and sufficient documentation and additional recommended practices are crucial factors to allow reproducibility of computational biology methods [195]. In bioinformatics research, poor reproducibility can also be attributed to data processing, where different pipelines can differ even in estimations for the same dataset [217–219]. Documentation and deposition of the processed data are imperative, and when possible, benchmarking of the processing pipeline and optimized parameters can substantially increase the reproducibility of ML approaches.

Last but not least, the lack of interpretability is a principal issue impeding the widespread usage and adaptation of ML, and especially DL techniques in bioinformatics research. Investigation of the biological mechanisms underlying the success of predictive models and features is highly desirable for the acceptance and use of these techniques, and particularly for clinical applications. Despite several important efforts to improve interpretability of DL models in biomedicine [197–199], model interpretability research in

genomic and medicine is highly underdeveloped. Common techniques to address the interpretation of concepts learned by a deep neural network include activation maximization, which identifies input patterns that maximize a desired model response [35,220]; sensitivity analysis or network function decomposition, aimed to explain the network's decisions and input representation [220–222]; and layer-wise backpropagation, which propagates the prediction to highlight the supporting input features [223]. Use of bioinformatic techniques, for example, for input representation, will enhance the interpretation of these analyses by revealing biological implications of the input patterns. Therefore, incorporation of established bioinformatics methods and curated databases into ML frameworks is a powerful way to increase the interpretability of these approaches, enhance their utility and use in biomedicine, and allow for follow-up investigation and derivation of hypotheses.

2. Conclusions

Machine learning and deep learning in particular are powerful computational tools that have already revolutionized many domains of research. With the recent expansive growth of genomic, molecular, and clinical data, ML offers unique solutions for the interrogation, analysis, and processing of these data, and for extracting substantial new knowledge on the underlying processes. The ML techniques are especially appealing in computational biology because of their ability to rapidly derive predictive models in the absence of strong assumptions about the underlying mechanisms, which is typical of some of the most pressing challenges in biomedicine. However, this unique ability also imposes serious obstacles for the development and widespread acceptance of the ML and particularly DL methods, impeding the reproducibility and interpretability of predictive models. Researchers in biomedical fields often lack the background and skills to perform or evaluate ML and especially DL analysis, which may lead to erroneous practices and conclusions [224]. The development of ML frameworks for biomedicine requires expertise in biology or clinical research, to comprehend and evaluate the strengths and limitations of intricate biological and clinical data, to be combined with a strong background in data mining and computational techniques.

Incorporation of ML techniques into established bioinformatics and computational biology frameworks has already notably facilitated the development of predictive models and powerful tools in molecular evolution, proteomics, systems biology, and disease genomics. The reliance on bioinformatics frameworks for data processing, training and evaluation of predictive models has been instrumental for the use and acceptance of these techniques in biomedicine, and such integrated approaches present promising solutions for many of the major obstacles for machine learning in biology and medicine.

Author Contributions: N.A., A.B.G. and E.V.K. wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: The authors' research is supported by the intramural research program at the National Institutes of Health (National Library of Medicine). This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable for a review article.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pevsner, J. Funtional Genomics. In *Bioinformatics and Functional Genomics*; John Wiley & Sons: Hoboken, NJ, USA, 2015; ISBN 9781118581780.
2. Ayyildiz, D.; Piazza, S. Introduction to Bioinformatics. In *Methods in Molecular Biology*; Oxford University Press: Oxford, UK, 2019.
3. Wodarz, D.; Komarova, N. *Computational Biology of Cancer*; World Scientific: Singapore, 2005.
4. Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

5. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nat. Cell Biol.* **2018**, *559*, 547–555. [CrossRef] [PubMed]
6. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's NMT. *arXiv* **2016**, arXiv:1609.08144v2.
7. Bhattacharjee, A.; Bayzid, M.S. Machine Learning Based Imputation Techniques for Estimating Phylogenetic Trees from Incomplete Distance Matrices. *BMC Genom.* **2020**, *21*, 497. [CrossRef]
8. Abadi, S.; Avram, O.; Rosset, S.; Pupko, T.; Mayrose, I. ModelTeller: Model Selection for Optimal Phylogenetic Reconstruction Using Machine Learning. *Mol. Biol. Evol.* **2020**, *37*, 3338–3352. [CrossRef]
9. Suvorov, A.; Hochuli, J.; Schrider, D.R. Accurate Inference of Tree Topologies from Multiple Sequence Alignments Using Deep Learning. *Syst. Biol.* **2019**, *69*, 221–233. [CrossRef]
10. Azer, E.S.; Ebrahimabadi, M.H.; Malikić, S.; Khardon, R.; Sahinalp, S.C. Tumor Phylogeny Topology Inference via Deep Learning. *iScience* **2020**, *23*, 101655. [CrossRef]
11. Jafari, R.; Javidi, M.M.; Kuchaki Rafsanjani, M. Using Deep Reinforcement Learning Approach for Solving the Multiple Sequence Alignment Problem. *SN Appl. Sci.* **2019**, *1*, 592. [CrossRef]
12. Yu, X. *Introduction to Evolutionary Algorithms*; Springer: Berlin/Heidelberg, Germany, 2010.
13. Fortin, F.A.; De Rainville, F.M.; Gardner, M.A.; Parizeau, M.; Gagńe, C. DEAP: Evolutionary Algorithms Made Easy. *J. Mach. Learn. Res.* **2012**, *13*, 2171–2175.
14. Whitley, D. A genetic algorithm tutorial. *Stat. Comput.* **1994**, *4*, 65–85. [CrossRef]
15. Pal, S.K.; Bandyopadhyay, S.; Ray, S.S. Evolutionary Computation in Bioinformatics: A Review. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2006**, *36*, 601–615. [CrossRef]
16. Sivanandam, S.N.; Deepa, S.N. *Introduction to Genetic Algorithms*; Springer: Berlin/Heidelberg, Germany, 2008; ISBN 9783540731894.
17. Audet, C.; Hare, W. Genetic Algorithms. In *Springer Series in Operations Research and Financial Engineering*; Springer: Berlin/Heidelberg, Germany, 2017.
18. Strodthoff, N.; Wagner, P.; Wenzel, M.; Samek, W. UDSMProt: Universal deep sequence models for protein classification. *Bioinformatics* **2020**, *36*, 2401–2409. [CrossRef] [PubMed]
19. Liu, B. BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Briefings Bioinform.* **2017**, *20*, 1280–1294. [CrossRef]
20. Gussow, A.B.; Auslander, N.; Faure, G.; Wolf, Y.I.; Zhang, F.; Koonin, E.V. Genomic Determinants of Pathogenicity in SARS-CoV-2 and Other Human Coronaviruses. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 15193–15199. [CrossRef]
21. Auslander, N.; Wolf, Y.I.; Shabalina, S.A.; Koonin, E.V. A unique insert in the genomes of high-risk human papillomaviruses with a predicted dual role in conferring oncogenic risk. *F1000Research* **2019**, *8*, 1000. [CrossRef]
22. Gussow, A.B.; Auslander, N.; Wolf, Y.I.; Koonin, E.V. Prediction of the incubation period for COVID-19 and future virus disease outbreaks. *BMC Biol.* **2020**, *18*, 186. [CrossRef] [PubMed]
23. Abadi, S.; Yan, W.X.; Amar, D.; Mayrose, I. A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS Comput. Biol.* **2017**, *13*, e1005807. [CrossRef]
24. Gussow, A.B.; Park, A.E.; Borges, A.L.; Shmakov, S.A.; Makarova, K.S.; Wolf, Y.I.; Bondy-Denomy, J.; Koonin, E.V. Machine-Learning Approach Expands the Repertoire of Anti-CRISPR Protein Families. *Nat. Commun.* **2020**, *11*, 3784. [CrossRef] [PubMed]
25. Eitzinger, S.; Asif, A.; Watters, K.E.; Iavarone, A.T.; Knott, G.J.; Doudna, J.A.; Minhas, F.; Ul, A.A. Machine Learning Predicts New Anti-CRISPR Proteins. *Nucleic Acids Res.* **2020**, *48*, 4698–4708. [CrossRef]
26. Solis-Reyes, S.; Avino, M.; Poon, A.; Kari, L. An Open-Source k-Mer Based Machine Learning Tool for Fast and Accurate Subtyping of HIV-1 Genomes. *PLoS ONE* **2018**, *13*, e0206409. [CrossRef]
27. Auslander, N.; Gussow, A.B.; Benler, S.; Wolf, Y.I.; Koonin, E.V. Seeker: Alignment-Free Identification of Bacteriophage Genomes by Deep Learning. *Nucleic Acids Res.* **2020**, *48*, e121. [CrossRef] [PubMed]
28. Fang, Z.; Tan, J.; Wu, S.; Li, M.; Xu, C.; Xie, Z.; Zhu, H. PPR-Meta: A tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience* **2019**, *8*. [CrossRef] [PubMed]
29. Ren, J.; Song, K.; Deng, C.; Ahlgren, N.A.; Fuhrman, J.A.; Li, Y.; Xie, X.; Poplin, R.; Sun, F. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* **2020**, *8*, 64–77. [CrossRef]
30. Seo, S.; Oh, M.; Park, Y.; Kim, S. DeepFam: Deep Learning Based Alignment-Free Method for Protein Family Modeling and Prediction. *Bioinformatics* **2018**, *34*, i254–i262. [CrossRef] [PubMed]
31. Kumar, M.; Thakur, V.; Raghava, G.P.S. COPid: Composition Based Protein Identification. *In Silico Biol.* **2008**, *8*, 121–128.
32. Liu, X.L. Deep Recurrent Neural Network for Protein Function Prediction from Sequence. *arXiv* **2017**, arXiv:1701.08318.
33. Hamid, M.-N.; Friedberg, I. Identifying antimicrobial peptides using word embedding with deep recurrent neural networks. *Bioinformmatics* **2019**, *35*, 2009–2016. [CrossRef]
34. Zacharaki, E.I. Prediction of protein function using a deep convolutional neural network ensemble. *PeerJ Comput. Sci.* **2017**, *3*. [CrossRef]
35. Nguyen, A.; Dosovitskiy, A.; Yosinski, J.; Brox, T.; Clune, J. Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks. *arXiv* **2016**, arXiv:1605.09304.

36. Le, N.Q.K.; Yapp, E.K.Y.; Nagasundaram, N.; Yeh, H.-Y. Classifying Promoters by Interpreting the Hidden Information of DNA Sequences via Deep Learning and Combination of Continuous FastText N-Grams. *Front. Bioeng. Biotechnol.* **2019**, *7*, 305. [CrossRef] [PubMed]
37. Umarov, R.K.; Solovyev, V.V. Recognition of Prokaryotic and Eukaryotic Promoters Using Convolutional Deep Learning Neural Networks. *PLoS ONE* **2017**, *12*, e0171410. [CrossRef]
38. Le, N.Q.K.; Ho, Q.-T.; Nguyen, T.-T.-D.; Ou, Y.-Y. A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information. *Briefings Bioinform.* **2021**. [CrossRef] [PubMed]
39. Min, X.; Zeng, W.; Chen, S.; Chen, N.; Chen, T.; Jiang, R. Predicting Enhancers with Deep Convolutional Neural Networks. *BMC Bioinform.* **2017**, *18*, 478. [CrossRef]
40. Xu, Y.; Zhao, X.; Liu, S.; Zhang, W. Predicting Long Non-Coding RNAs through Feature Ensemble Learning. *BMC Genom.* **2020**, *21*, 865. [CrossRef] [PubMed]
41. Sun, L.; Liu, H.; Zhang, L.; Meng, J. LncRScan-SVM: A Tool for Predicting Long Non-Coding RNAs Using Support Vector Machine. *PLoS ONE* **2015**, *10*, e0139654. [CrossRef] [PubMed]
42. Schneider, H.W.; Raiol, T.; Brigido, M.M.; Walter, M.E.M.T.; Stadler, P.F. A Support Vector Machine Based Method to Distinguish Long Non-Coding RNAs from Protein Coding Transcripts. *BMC Genom.* **2017**, *18*, 804. [CrossRef]
43. Hu, L.; Xu, Z.; Hu, B.; Lu, Z.J. COME: A Robust Coding Potential Calculation Tool for LncRNA Identification and Characterization Based on Multiple Features. *Nucleic Acids Res.* **2017**, *45*, e2. [CrossRef]
44. Zhao, J.; Song, X.; Wang, K. lncScore: Alignment-free identification of long noncoding RNA from assembled novel transcripts. *Sci. Rep.* **2016**, *6*, 34838. [CrossRef]
45. Wen, M.; Cong, P.; Zhang, Z.; Lu, H.; Li, T. DeepMirTar: A Deep-Learning Approach for Predicting Human MiRNA Targets. *Bioinformatics* **2018**, *34*, 3781–3787. [CrossRef]
46. Zheng, X.; Chen, L.; Li, X.; Zhang, Y.; Xu, S.; Huang, X. Prediction of MiRNA Targets by Learning from Interaction Sequences. *PLoS ONE* **2020**, *15*, e0232578. [CrossRef] [PubMed]
47. Mitrofanov, A.; Alkhnbashi, O.S.; Shmakov, S.A.; Makarova, K.S.; Koonin, E.V.; Backofen, R. CRISPRidentify: Identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res.* **2021**, *49*, e20. [CrossRef]
48. Blom, N.; Sicheritz-Pontén, T.; Gupta, R.; Gammeltoft, S.; Brunak, S. Prediction of Post-Translational Glycosylation and Phosphorylation of Proteins from the Amino Acid Sequence. *Proteomics* **2004**, *4*, 1633–1649. [CrossRef]
49. Huang, G.; Li, J. Feature Extractions for Computationally Predicting Protein Post- Translational Modifications. *Curr. Bioinform.* **2018**, *13*, 387–395. [CrossRef]
50. Wang, D.; Liu, D.; Yuchi, J.; He, F.; Jiang, Y.; Cai, S.; Li, J.; Xu, D. MusiteDeep: A deep-learning based webserver for protein post-translational modification site prediction and visualization. *Nucleic Acids Res.* **2020**, *48*, W140–W146. [CrossRef] [PubMed]
51. Duan, G.; Walther, D. The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLoS Comput. Biol.* **2015**, *11*, e1004049. [CrossRef] [PubMed]
52. Jia, C.; Zuo, Y.; Zou, Q. O-GlcNAcPRED-II: An integrated classification algorithm for identifying O-GlcNAcylation sites based on fuzzy undersampling and a K-means PCA oversampling technique. *Bioinformatics* **2018**, *34*, 2029–2036. [CrossRef] [PubMed]
53. Gao, J.; Thelen, J.J.; Dunker, A.K.; Xu, D. Musite, a Tool for Global Prediction of General and Kinase-specific Phosphorylation Sites. *Mol. Cell. Proteom.* **2010**, *9*, 2586–2600. [CrossRef] [PubMed]
54. Caragea, C.; Sinapov, J.; Silvescu, A.; Dobbs, D.; Honavar, V. Glycosylation site prediction using ensembles of Support Vector Machine classifiers. *BMC Bioinform.* **2007**, *8*, 438. [CrossRef]
55. Luo, F.; Wang, M.; Liu, Y.; Zhao, X.-M.; Li, A. DeepPhos: Prediction of protein phosphorylation sites with deep learning. *Bioinformatics* **2019**, *35*, 2766–2773. [CrossRef]
56. Kotidis, P.; Kontoravdi, C. Harnessing the potential of artificial neural networks for predicting protein glycosylation. *Metab. Eng. Commun.* **2020**, *10*, e00131. [CrossRef]
57. Hameduh, T.; Haddad, Y.; Adam, V.; Heger, Z. Homology Modeling in the Time of Collective and Artificial Intelligence. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 3494–3506. [CrossRef] [PubMed]
58. Torrisi, M.; Pollastri, G.; Le, Q. Deep Learning Methods in Protein Structure Prediction. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1301–1310. [CrossRef] [PubMed]
59. Shakhnovich, B.E. Protein Structure and Evolutionary History Determine Sequence Space Topology. *Genome Res.* **2005**, *15*, 385–392. [CrossRef]
60. Muhammed, M.T.; Aki-Yalcin, E. Homology Modeling in Drug Discovery: Overview, Current Applications, and Future Perspectives. *Chem. Biol. Drug Des.* **2019**, *93*, 12–20. [CrossRef]
61. Lazaridis, T.; Karplus, M. Effective Energy Functions for Protein Structure Prediction. *Curr. Opin. Struct. Biol.* **2000**, *10*, 139–145. [CrossRef]
62. Snow, C.D.; Sorin, E.J.; Rhee, Y.M.; Pande, V.S. How Well Can Simulation Predict Protein Folding Kinetics and Thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 43–69. [CrossRef]
63. Spassov, V.Z.; Flook, P.K.; Yan, L. LOOPER: A molecular mechanics-based algorithm for protein loop prediction. *Protein Eng. Des. Sel.* **2008**, *21*, 91–100. [CrossRef] [PubMed]
64. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XIII. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1011–1020. [CrossRef]

65. Xu, J.; Wang, S. Analysis of Distance-based Protein Structure Prediction by Deep Learning in CASP13. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1069–1081. [CrossRef]
66. Zheng, W.; Li, Y.; Zhang, C.; Pearce, R.; Mortuza, S.M.; Zhang, Y. Deep-learning Contact-map Guided Protein Structure Prediction in CASP13. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1149–1164. [CrossRef]
67. Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: High-Accuracy Protein Contact Prediction by Coupling Precision Matrix with Deep Residual Neural Networks. *Bioinformatics* **2019**, *35*, 4647–4655. [CrossRef]
68. Hou, J.; Wu, T.; Guo, Z.; Quadir, F.; Cheng, J. The MULTICOM Protein Structure Prediction Server Empowered by Deep Learning and Contact Distance Prediction. In *Protein Structure Prediction*; Humana Press: New York, NY, USA, 2020; pp. 13–26.
69. Jones, D.T.; Kandathil, S.M. High Precision in Protein Contact Prediction Using Fully Convolutional Neural Networks and Minimal Sequence Features. *Bioinformatics* **2018**, *34*, 3308–3315. [CrossRef] [PubMed]
70. Adhikari, B.; Hou, J.; Cheng, J. DNCON2: Improved Protein Contact Prediction Using Two-Level Deep Convolutional Neural Networks. *Bioinformatics* **2018**, *34*, 1466–1472. [CrossRef]
71. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Protein Structure Prediction Using Multiple Deep Neural Networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1141–1148. [CrossRef]
72. Fukuda, H.; Tomii, K. DeepECA: An End-to-End Learning Framework for Protein Contact Prediction from a Multiple Sequence Alignment. *BMC Bioinform.* **2020**, *21*, 10. [CrossRef] [PubMed]
73. Kandathil, S.M.; Greener, J.G.; Jones, D.T. Prediction of Interresidue Contacts with DeepMetaPSICOV in CASP13. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1092–1099. [CrossRef] [PubMed]
74. Stahl, K.; Schneider, M.; Brock, O. EPSILON-CP: Using Deep Learning to Combine Information from Multiple Sources for Protein Contact Prediction. *BMC Bioinform.* **2017**, *18*, 303. [CrossRef]
75. Gao, M.; Zhou, H.; Skolnick, J. DESTINI: A Deep-Learning Approach to Contact-Driven Protein Structure Prediction. *Sci. Rep.* **2019**, *9*, 3514. [CrossRef] [PubMed]
76. AlQuraishi, M. AlphaFold at CASP13. *Bioinformatics* **2019**, *35*, 4862–4865. [CrossRef] [PubMed]
77. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, *577*, 706–710. [CrossRef] [PubMed]
78. Liu, J.; Wu, T.; Guo, Z.; Hou, J.; Cheng, J. Improving Protein Tertiary Structure Prediction by Deep Learning and Distance Prediction in CASP14. *bioRxiv* **2021**, *1*, 1–10. [CrossRef]
79. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
80. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-Generation Machine Learning for Biological Networks. *Cell* **2018**, *173*, 1581–1592. [CrossRef]
81. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Kellis, M.; Collins, J.J.; Stolovitzky, G. Wisdom of crowds for robust gene network inference. *Nat. Methods* **2012**, *9*, 796–804. [CrossRef]
82. Mordelet, F.; Vert, J.P. SIRENE: Supervised Inference of Regulatory Networks. *Bioinformatics* **2008**, *24*, i76–i82. [CrossRef]
83. Mignone, P.; Pio, G.; D'Elia, D.; Ceci, M. Exploiting transfer learning for the reconstruction of the human gene regulatory network. *Bioinformatics* **2019**, *36*, 1553–1561. [CrossRef] [PubMed]
84. Jackson, C.A.; Castro, D.M.; Saldi, G.-A.; Bonneau, R.; Gresham, D. Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *eLife* **2020**, *9*. [CrossRef] [PubMed]
85. Greene, D.; Cagney, G.; Krogan, N.; Cunningham, P. Ensemble non-negative matrix factorization methods for clustering protein–protein interactions. *Bioinformatics* **2008**, *24*, 1722–1728. [CrossRef]
86. Huang, D.-S.; Zhang, L.; Han, K.; Deng, S.; Yang, K.; Zhang, H. Prediction of Protein-Protein Interactions Based on Protein-Protein Correlation Using Least Squares Regression. *Curr. Protein Pept. Sci.* **2014**, *15*, 553–560. [CrossRef] [PubMed]
87. You, Z.-H.; Lei, Y.-K.; Zhu, L.; Xia, J.; Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinform.* **2013**, *14*, S10. [CrossRef]
88. Zhang, L.; Yu, G.; Xia, D.; Wang, J. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* **2019**, *324*, 10–19. [CrossRef]
89. Yang, F.; Fan, K.; Song, D.; Lin, H. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinform.* **2020**, *21*, 323. [CrossRef]
90. Chatterjee, P.; Basu, S.; Kundu, M.; Nasipuri, M.; Plewczynski, D. PPI_SVM: Prediction of protein-protein interactions using machine learning, domain-domain affinities and frequency tables. *Cell. Mol. Biol. Lett.* **2011**, *16*, 264–278. [CrossRef]
91. Chen, C.; Zhang, Q.; Ma, Q.; Yu, B. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemom. Intell. Lab. Syst.* **2019**, *191*, 54–64. [CrossRef]
92. Wang, Y.-B.; You, Z.-H.; Li, X.; Jiang, T.-H.; Chen, X.; Zhou, X.; Wang, L. Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. *Mol. BioSyst.* **2017**, *13*, 1336–1344. [CrossRef]
93. Du, X.; Sun, S.; Hu, C.; Yao, Y.; Yan, Y.; Zhang, Y. DeepPPI: Boosting Prediction of Protein–Protein Interactions with Deep Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 1499–1510. [CrossRef]
94. Lei, H.; Wen, Y.; You, Z.; ElAzab, A.; Tan, E.-L.; Zhao, Y.; Lei, B. Protein–Protein Interactions Prediction via Multimodal Deep Polynomial Network and Regularized Extreme Learning Machine. *IEEE J. Biomed. Heal. Inform.* **2018**, *23*, 1290–1303. [CrossRef]

95. Hashemifar, S.; Neyshabur, B.; Khan, A.A.; Xu, J. Predicting Protein-Protein Interactions through Sequence-Based Deep Learning. *Bioinformatics* **2018**, *34*, i802–i810. [CrossRef]
96. Lu, Y.; Guo, Y.; Korhonen, A. Link prediction in drug-target interactions network using similarity indices. *BMC Bioinform.* **2017**, *18*, 39. [CrossRef]
97. Nascimento, A.C.A.; Prudêncio, R.B.C.; Costa, I.G. A Drug-Target Network-Based Supervised Machine Learning Repurposing Method Allowing the Use of Multiple Heterogeneous Information Sources. In *Methods in Molecular Biology*; Springer: Berlin/Heidelberg, Germany, 2019.
98. Aghakhani, S.; Qabaja, A.; Alhajj, R. Integration of k-means clustering algorithm with network analysis for drug-target interactions network prediction. *Int. J. Data Min. Bioinform.* **2018**, *20*, 185. [CrossRef]
99. Madhukar, N.S.; Khade, P.K.; Huang, L.; Gayvert, K.; Galletti, G.; Stogniew, M.; Allen, J.E.; Giannakakou, P.; Elemento, O. A Bayesian machine learning approach for drug target identification using diverse data types. *Nat. Commun.* **2019**, *10*, 5221. [CrossRef]
100. Zeng, X.; Zhu, S.; Hou, Y.; Zhang, P.; Li, L.; Li, J.; Huang, L.F.; Lewis, S.J.; Nussinov, R.; Cheng, F. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* **2020**, *36*, 2805–2812. [CrossRef] [PubMed]
101. Liu, P.; Li, H.; Li, S.; Leung, K.-S. Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network. *BMC Bioinform.* **2019**, *20*, 408–414. [CrossRef] [PubMed]
102. Chang, Y.; Park, H.; Yang, H.-J.; Lee, S.; Lee, K.-Y.; Kim, T.S.; Jung, J.; Shin, J.-M. Cancer Drug Response Profile scan (CDRscan): A Deep Learning Model That Predicts Drug Effectiveness from Cancer Genomic Signature. *Sci. Rep.* **2018**, *8*, 8857. [CrossRef]
103. Chiu, Y.-C.; Chen, H.-I.H.; Zhang, T.; Zhang, S.; Gorthi, A.; Wang, L.-J.; Huang, Y.; Chen, Y. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Med Genom.* **2019**, *12*, 143–155. [CrossRef]
104. Sharifi-Noghabi, H.; Zolotareva, O.; Collins, C.C.; Ester, M. MOLI: Multi-Omics Late Integration with Deep Neural Networks for Drug Response Prediction. *Bioinformatics* **2019**, *35*, i501–i509. [CrossRef]
105. Duran-Frigola, M.; Pauls, E.; Guitart-Pla, O.; Bertoni, M.; Alcalde, V.; Amat, D.; Juan-Blanco, T.; Aloy, P. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat. Biotechnol.* **2020**, *38*, 1087–1096. [CrossRef] [PubMed]
106. Kaushik, A.C.; Mehmood, A.; Dai, X.; Wei, D.-Q. A comparative chemogenic analysis for predicting Drug-Target Pair via Machine Learning Approaches. *Sci. Rep.* **2020**, *10*, 6870. [CrossRef]
107. Li, Z.; Han, P.; You, Z.-H.; Li, X.; Zhang, Y.; Yu, H.; Nie, R.; Chen, X. In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci. Rep.* **2017**, *7*, 11174. [CrossRef]
108. Bari, M.G.; Ung, C.Y.; Zhang, C.; Zhu, S.; Li, H. Machine Learning-Assisted Network Inference Approach to Identify a New Class of Genes that Coordinate the Functionality of Cancer Networks. *Sci. Rep.* **2017**, *7*, 6993. [CrossRef] [PubMed]
109. Kong, J.; Lee, H.; Kim, D.; Han, S.K.; Ha, D.; Shin, K.; Kim, S. Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. *Nat. Commun.* **2020**, *11*, 5485. [CrossRef]
110. Ammad-Ud-Din, M.; Khan, S.A.; Wennerberg, K.; Aittokallio, T. Systematic Identification of Feature Combinations for Predicting Drug Response with Bayesian Multi-View Multi-Task Linear Regression. *Bioinformatics* **2017**, *33*, i359–i368. [CrossRef]
111. Stanfield, Z.; Coşkun, M.; Koyutürk, M. Drug Response Prediction as a Link Prediction Problem. *Sci. Rep.* **2017**, *7*, 40321. [CrossRef]
112. Shaked, I.; Oberhardt, M.A.; Atias, N.; Sharan, R.; Ruppin, E. Metabolic Network Prediction of Drug Side Effects. *Cell Syst.* **2016**, *2*, 209–213. [CrossRef]
113. Zhao, X.; Chen, L.; Lu, J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* **2018**, *306*, 136–144. [CrossRef]
114. Zhang, W.; Liu, F.; Luo, L.; Zhang, J. Predicting drug side effects by multi-label learning and ensemble learning. *BMC Bioinform.* **2015**, *16*, 365. [CrossRef] [PubMed]
115. Cheng, F.; Kovács, I.A.; Barabási, A.-L. Network-based prediction of drug combinations. *Nat. Commun.* **2019**, *10*, 1197. [CrossRef] [PubMed]
116. Singh, H.; Rana, P.S.; Singh, U. Prediction of drug synergy in cancer using ensemble-based machine learning techniques. *Mod. Phys. Lett. B* **2018**, *32*. [CrossRef]
117. Wildenhain, J.; Spitzer, M.; Dolma, S.; Jarvik, N.; White, R.; Roy, M.; Griffiths, E.; Bellows, D.S.; Wright, G.D.; Tyers, M. Prediction of Synergism from Chemical-Genetic Interactions by Machine Learning. *Cell Syst.* **2015**, *1*, 383–395. [CrossRef]
118. Zeng, X.; Zhu, S.; Liu, X.; Zhou, Y.; Nussinov, R.; Cheng, F. deepDR: A network-based deep learning approach to in silico drug repositioning. *Bioinformatics* **2019**, *35*, 5191–5198. [CrossRef]
119. Gottlieb, A.; Stein, G.Y.; Ruppin, E.; Sharan, R. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011**, *7*, 496. [CrossRef]
120. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; E Baranzini, S. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **2017**, *6*, e26726. [CrossRef] [PubMed]
121. Dale, J.M.; Popescu, L.; Karp, P.D. Machine learning methods for metabolic pathway prediction. *BMC Bioinform.* **2010**, *11*, 15. [CrossRef]

122. Baranwal, M.; Magner, A.; Elvati, P.; Saldinger, J.; Violi, A.; Hero, A.O. A deep learning architecture for metabolic pathway prediction. *Bioinformatics* **2019**, *36*, 2547–2553. [CrossRef] [PubMed]
123. Çubuk, C.; Hidalgo, M.R.; Amadoz, A.; Rian, K.; Salavert, F.; Pujana, M.A.; Mateo, F.; Herranz, C.; Carbonell-Caballero, J.; Dopazo, J. Differential metabolic activity and discovery of therapeutic targets using summarized metabolic pathway models. *NPJ Syst. Biol. Appl.* **2019**, *5*, 7. [CrossRef] [PubMed]
124. Auslander, N.; Wagner, A.; Oberhardt, M.; Ruppin, E. Data-Driven Metabolic Pathway Compositions Enhance Cancer Survival Prediction. *PLoS Comput. Biol.* **2016**, *12*. [CrossRef]
125. Kim, J.Y.; Lee, H.; Woo, J.; Yue, W.; Kim, K.; Choi, S.; Jang, J.-J.; Kim, Y.; Park, I.A.; Han, D.; et al. Reconstruction of pathway modification induced by nicotinamide using multi-omic network analyses in triple negative breast cancer. *Sci. Rep.* **2017**, *7*, 3466. [CrossRef]
126. Fu, G.; Ding, Y.; Seal, A.; Chen, B.; Sun, Y.; Bolton, E. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinform.* **2016**, *17*, 1–10. [CrossRef] [PubMed]
127. Yang, M.; Simm, J.; Lam, C.C.; Zakeri, P.; Van Westen, G.J.P.; Moreau, Y.; Saez-Rodriguez, J. Linking drug target and pathway activation for effective therapy using multi-task learning. *Sci. Rep.* **2018**, *8*, 8322. [CrossRef]
128. Esteban-Medina, M.; Peña-Chilet, M.; Loucera, C.; Dopazo, J. Exploring the druggable space around the Fanconi anemia pathway using machine learning and mechanistic models. *BMC Bioinform.* **2019**, *20*, 1–15. [CrossRef] [PubMed]
129. Way, G.P.; Sanchez-Vega, F.; Konnor Cancer Genome Atlas Research Network; Armenia, J.; Chatila, W.K.; Luna, A.; Sander, C.; Cherniack, A.D.; Mina, M.; Ciriello, G.; et al. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep.* **2018**, *23*, 172–180.e3. [CrossRef]
130. Huang, E.; Ishida, S.; Pittman, J.; Dressman, H.; Bild, A.; Kloos, M.; D'Amico, M.; Pestell, R.G.; West, M.; Nevins, J.R. Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nat. Genet.* **2003**, *34*, 226–230. [CrossRef]
131. Costello, Z.; Martin, H.G. A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. *NPJ Syst. Biol. Appl.* **2018**, *4*, 1–14. [CrossRef]
132. Tepeli, Y.I.; Ünal, A.B.; Akdemir, F.M.; Tastan, O. PAMOGK: A pathway graph kernel-based multiomics approach for patient clustering. *Bioinformatics* **2021**, *36*, 5237–5246. [CrossRef]
133. Cho, H.; Berger, B.; Peng, J. Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell Syst.* **2016**, *3*, 540–548.e5. [CrossRef]
134. Auslander, N.; Yizhak, K.; Weinstock, A.; Budhu, A.; Tang, W.; Wang, X.W.; Ambs, S.; Ruppin, E. A joint analysis of transcriptomic and metabolomic data uncovers enhanced enzyme-metabolite coupling in breast cancer. *Sci. Rep.* **2016**, *6*, 29662. [CrossRef] [PubMed]
135. Katzir, R.; Polat, I.H.; Harel, M.; Katz, S.; Foguet, C.; Selivanov, V.A.; Sabatier, P.; Cascante, M.; Geiger, T.; Ruppin, E. The landscape of tiered regulation of breast cancer cell metabolism. *Sci. Rep.* **2019**, *9*, 1–12. [CrossRef]
136. Lan, L.; Djuric, N.; Guo, Y.; Vucetic, S. MS-k NN: Protein function prediction by integrating multiple data sources. *BMC Bioinform.* **2013**, *14*, S8. [CrossRef] [PubMed]
137. Yao, Z.; Ruzzo, W.L. A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinform.* **2006**, *7*, S11. [CrossRef] [PubMed]
138. Yang, Z.; Michailidis, G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **2015**, *32*, 1–8. [CrossRef]
139. Kim, S.; Kang, D.; Huo, Z.; Park, Y.; Tseng, G.C. Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* **2017**, *34*, 1321–1328. [CrossRef] [PubMed]
140. Wang, T.; Shao, W.; Huang, Z.; Tang, H.; Zhang, J.; Ding, Z.; Huang, K. MORONET: Multi-Omics Integration via Graph Convolutional NETworks for Biomedical Data Classification. *bioRxiv* **2020**, *1*, 1–10. [CrossRef]
141. Rappoport, N.; Shamir, R. NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics* **2019**, *35*, 3348–3356. [CrossRef]
142. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef]
143. Chaudhary, K.; Poirion, O.B.; Lu, L.; Garmire, L.X. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin. Cancer Res.* **2017**, *24*, 1248–1259. [CrossRef] [PubMed]
144. Xu, Y.; Dong, Q.; Li, F.; Xu, Y.; Hu, C.; Wang, J.; Shang, D.; Zheng, X.; Yang, H.; Zhang, C.; et al. Identifying subpathway signatures for individualized anticancer drug response by integrating multi-omics data. *J. Transl. Med.* **2019**, *17*, 1–16. [CrossRef] [PubMed]
145. Sathyanarayanan, A.; Gupta, R.; Thompson, E.W.; Nyholt, D.R.; Bauer, D.C.; Nagaraj, S.H. A comparative study of multi-omics integration tools for cancer driver gene identification and tumour subtyping. *Briefings Bioinform.* **2020**, *21*, 1920–1936. [CrossRef]
146. McCabe, S.D.; Lin, D.Y.; Love, M.I. Consistency and Overfitting of Multi-Omics Methods on Experimental Data. *Brief. Bioinform.* **2019**, *21*, 1277–1284. [CrossRef] [PubMed]
147. Haendel, M.A.; Chute, C.G.; Robinson, P.N. Classification, Ontology, and Precision Medicine. *New Engl. J. Med.* **2018**, *379*, 1452–1462. [CrossRef] [PubMed]
148. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From Big Data to Precision Medicine. *Front. Med.* **2019**, *6*, 34. [CrossRef]

149. Nitsch, D.; Gonçalves, J.P.; Ojeda, F.; De Moor, B.; Moreau, Y. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinform.* **2010**, *11*, 1–16. [CrossRef]
150. Li, H.; Zhao, X.; Wang, J.; Zong, M.; Yang, H. Bioinformatics analysis of gene expression profile data to screen key genes involved in pulmonary sarcoidosis. *Gene* **2017**, *596*, 98–104. [CrossRef] [PubMed]
151. Xiao, Q.; Luo, J.; Liang, C.; Cai, J.; Ding, P. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* **2018**, *34*, 239–248. [CrossRef] [PubMed]
152. Asif, M.; Martiniano, H.F.M.C.M.; Vicente, A.M.; Couto, F.M. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. *PLoS ONE* **2018**, *13*, e0208626. [CrossRef] [PubMed]
153. Lopez-Bigas, N.; Ouzounis, C.A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.* **2004**, *32*, 3108–3114. [CrossRef] [PubMed]
154. Xu, W.; Jiang, X.; Hu, X.; Li, G. Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization. *BMC Med. Genom.* **2014**, *7*, S1. [CrossRef]
155. Shen, X.; Zhu, X.; Jiang, X.; He, T.; Hu, X. Visualization of Disease Relationships by Multiple Maps T-SNE Regularization Based on Nesterov Accelerated Gradient. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, Kansas City, MO, USA, 13–16 November 2017.
156. Lage, K.; Karlberg, L.O.E.; Størling, M.Z.; Ólason, P.Í.; Pedersen, A.G.; Rigina, O.; Hinsby, A.M.; Tümer, Z.; Pociot, F.; Tommerup, N.; et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* **2007**, *25*, 309–316. [CrossRef]
157. Xu, J.; Li, Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics* **2006**, *22*, 2800–2805. [CrossRef]
158. Barman, R.K.; Mukhopadhyay, A.; Maulik, U.; Das, S. Identification of infectious disease-associated host genes using machine learning techniques. *BMC Bioinform.* **2019**, *20*, 736. [CrossRef]
159. Han, Y.; Yang, J.; Qian, X.; Cheng, W.-C.; Liu, S.-H.; Hua, X.; Zhou, L.; Yang, Y.; Wu, Q.; Liu, P.; et al. DriverML: A machine learning algorithm for identifying driver genes in cancer sequencing studies. *Nucleic Acids Res.* **2019**, *47*, e45. [CrossRef]
160. Auslander, N.; Wolf, Y.I.; Koonin, E.V. Interplay between DNA damage repair and apoptosis shapes cancer evolution through aneuploidy and microsatellite instability. *Nat. Commun.* **2020**, *11*, 1234. [CrossRef]
161. Collier, O.; Stoven, V.; Vert, J.-P. LOTUS: A single- and multitask machine learning algorithm for the prediction of cancer driver genes. *PLoS Comput. Biol.* **2019**, *15*, e1007381. [CrossRef]
162. Luo, P.; Ding, Y.; Lei, X.; Wu, F.-X. deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. *Front. Genet.* **2019**, *10*, 13. [CrossRef]
163. Agajanian, S.; Oluyemi, O.; Verkhivker, G.M. Integration of Random Forest Classifiers and Deep Convolutional Neural Networks for Classification and Biomolecular Modeling of Cancer Driver Mutations. *Front. Mol. Biosci.* **2019**, *6*, 44. [CrossRef]
164. Califf, R.M. Biomarker definitions and their applications. *Exp. Biol. Med.* **2018**, *243*, 213–221. [CrossRef]
165. Ray, P.; Manach, Y.L.; Riou, B.; Houle, T.T. Statistical Evaluation of a Biomarker. *Anesthesiology* **2010**, *112*, 1023–1040. [CrossRef]
166. McDermott, J.E.; Wang, J.; Mitchell, H.; Webb-Robertson, B.J.; Hafen, R.; Ramey, J.; Rodland, K.D. Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data. *Expert Opin. Med. Diagn.* **2013**, *7*, 37–51. [CrossRef]
167. Cun, Y.; Fröhlich, H. netClass: An R-package for network based, integrative biomarker signature discovery. *Bioinformatics* **2014**, *30*, 1325–1326. [CrossRef]
168. Yasui, Y.; Pepe, M.; Thompson, M.L.; Adam, B.; Wright, G.L.; Qu, Y.; Potter, J.D.; Winget, M.; Thornquist, M.; Feng, Z. A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **2003**, *4*, 449–463. [CrossRef] [PubMed]
169. Statnikov, A.; Tsamardinos, I.; Dosbayev, Y.; Aliferis, C.F. GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *Int. J. Med. Inform.* **2005**, *74*, 491–503. [CrossRef] [PubMed]
170. Abeel, T.; Helleputte, T.; Van de Peer, Y.; Dupont, P.; Saeys, Y. Robust Biomarker Identification for Cancer Diagnosis with Ensemble Feature Selection Methods. *Bioinformatics* **2010**, *26*, 392–398. [CrossRef] [PubMed]
171. Kossenkov, A.V.; Qureshi, R.; Dawany, N.B.; Wickramasinghe, J.; Liu, Q.; Majumdar, R.S.; Chang, C.; Widura, S.; Kumar, T.; Horng, W.-H.; et al. A Gene Expression Classifier from Whole Blood Distinguishes Benign from Malignant Lung Nodules Detected by Low-Dose CT. *Cancer Res.* **2019**, *79*, 263–273. [CrossRef]
172. Gal, O.; Auslander, N.; Fan, Y.; Meerzaman, D. Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression. *Cancer Inform.* **2019**, *18*. [CrossRef] [PubMed]
173. Ganti, S.; Weiss, R.H. Urine Metabolomics for Kidney Cancer Detection and Biomarker Discovery. In *Urologic Oncology: Seminars and Original Investigations*; Elsevier: Amsterdam, The Netherlands, 2011.
174. Shen, C.; Sun, Z.; Chen, D.; Su, X.; Jiang, J.; Li, G.; Lin, B.; Biaoyang, L. Developing Urinary Metabolomic Signatures as Early Bladder Cancer Diagnostic Markers. *OMICS A J. Integr. Biol.* **2015**, *19*, 1–11. [CrossRef] [PubMed]
175. Leclercq, M.; Vittrant, B.; Martin-Magniette, M.L.; Boyer, M.P.S.; Perin, O.; Bergeron, A.; Fradet, Y.; Droit, A. Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. *Front. Genet.* **2019**, *10*, 452. [CrossRef]

176. Wang, J.; Zuo, Y.; Man, Y.G.; Avital, I.; Stojadinovic, A.; Liu, M.; Yang, X.; Varghese, R.S.; Tadesse, M.G.; Ressom, H.W. Pathway and Network Approaches for Identification of Cancer Signature Markers from Omics Data. *J. Cancer* **2015**, *6*, 54. [CrossRef] [PubMed]
177. Long, N.P.; Jung, K.H.; Anh, N.H.; Yan, H.H.; Nghi, T.D.; Park, S.; Yoon, S.J.; Min, J.E.; Kim, H.M.; Lim, J.H.; et al. An Integrative Data Mining and Omics-Based Translational Model for the Identification and Validation of Oncogenic Biomarkers of Pancreatic Cancer. *Cancers* **2019**, *11*, 155. [CrossRef]
178. Rohart, F.; Gautier, B.; Singh, A.; Cao, K.-A.L. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **2017**, *13*, e1005752. [CrossRef]
179. Guan, X.; Runger, G.; Liu, L. Dynamic incorporation of prior knowledge from multiple domains in biomarker discovery. *BMC Bioinform.* **2020**, *21*, 1–10. [CrossRef]
180. Foroughi Pour, A.; Dalton, L.A. Integrating Prior Information with Bayesian Feature Selection. In Proceedings of the 8th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB), Boston, MA, USA, 20–23 August 2017; p. 610. [CrossRef]
181. Liu, L.; Chang, Y.; Yang, T.; Noren, D.P.; Long, B.; Kornblau, S.; Qutub, A.; Ye, J. Evolution-informed modeling improves outcome prediction for cancers. *Evol. Appl.* **2016**, *10*, 68–76. [CrossRef]
182. Johannes, M.; Fröhlich, H.; Sültmann, H.; Beissbarth, T.; Beißbarth, T. pathClass: An R-package for integration of pathway knowledge into support vector machines for biomarker discovery. *Bioinformatics* **2011**, *27*, 1442–1443. [CrossRef]
183. Haider, S.; Yao, C.Q.; Sabine, V.S.; Grzadkowski, M.; Stimper, V.; Starmans, M.H.W.; Wang, J.; Nguyen, F.; Moon, N.C.; Lin, X.; et al. Pathway-based subnetworks enable cross-disease biomarker discovery. *Nat. Commun.* **2018**, *9*, 4746. [CrossRef]
184. Fujita, N.; Mizuarai, S.; Murakami, K.; Nakai, K. Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. *Sci. Rep.* **2018**, *8*, 9743. [CrossRef] [PubMed]
185. Abbas, M.; Matta, J.; Le, T.; Bensmail, H.; Obafemi-Ajayi, T.; Honavar, V.; El-Manzalawy, Y. Biomarker discovery in inflammatory bowel diseases using network-based feature selection. *PLoS ONE* **2019**, *14*, e0225382. [CrossRef] [PubMed]
186. Zhang, J.; Xiang, Y.; Ding, L.; Keen-Circle, K.; Borlawsky, T.B.; Ozer, H.G.; Jin, R.; Payne, P.; Huang, K. Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia. *BMC Bioinform.* **2010**, *11*, S5. [CrossRef] [PubMed]
187. Lee, S.-I.; Celik, S.; Logsdon, B.A.; Lundberg, S.M.; Martins, T.J.; Oehler, V.G.; Estey, E.H.; Miller, C.P.; Chien, S.; Dai, J.; et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.* **2018**, *9*, 1–13. [CrossRef]
188. Cheerla, N.; Gevaert, O. MicroRNA based Pan-Cancer Diagnosis and Treatment Recommendation. *BMC Bioinform.* **2017**, *18*, 1–11. [CrossRef]
189. Wang, L.; He, X.; Zhang, W.; Zha, H. Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, London, UK, 19–23 August 2018.
190. Samala, R.K.; Chan, H.-P.; Hadjiiski, L.M.; A Helvie, M.; Richter, C.; Cha, K. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Phys. Med. Biol.* **2018**, *63*, 095005. [CrossRef]
191. Asgari, E.; Mofrad, M.R.K. Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. *PLoS ONE* **2015**, *10*, e0141287. [CrossRef]
192. Du, J.; Jia, P.; Dai, Y.; Tao, C.; Zhao, Z.; Zhi, D. Gene2vec: Distributed representation of genes based on co-expression. *BMC Genom.* **2019**, *20*, 7–15. [CrossRef]
193. Kim, S.; Lee, H.; Kim, K.; Kang, J. Mut2Vec: Distributed representation of cancerous mutations. *BMC Med. Genom.* **2018**, *11*, 57–69. [CrossRef]
194. Xu, Y.; Song, J.; Wilson, C.; Whisstock, J.C. PhosContext2vec: A distributed representation of residue-level sequence contexts and its application to general and kinase-specific phosphorylation site prediction. *Sci. Rep.* **2018**, *8*, 1–14. [CrossRef] [PubMed]
195. Sandve, G.K.; Nekrutenko, A.; Taylor, J.; Hovig, E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput. Biol.* **2013**, *9*, e1003285. [CrossRef]
196. Patel-Murray, N.L.; Adam, M.; Huynh, N.; Wassie, B.T.; Milani, P.; Fraenkel, E. A Multi-Omics Interpretable Machine Learning Model Reveals Modes of Action of Small Molecules. *Sci. Rep.* **2020**, *10*, 1–14. [CrossRef]
197. Jha, A.; Aicher, J.K.; Gazzara, M.R.; Singh, D.; Barash, Y. Enhanced Integrated Gradients: Improving interpretability of deep learning models using splicing codes as a case study. *Genome Biol.* **2020**, *21*, 1–22. [CrossRef]
198. Hao, J.; Kosaraju, S.C.; Tsaku, N.Z.; Song, D.H.; Kang, M. PAGE-Net: Interpretable and Integrative Deep Learning for Survival Analysis Using Histopathological Images and Genomic Data. In Proceedings of the Pacific Symposium on Biocomputing, Fairmont Orchid, HI, USA, 3–7 January 2020.
199. Dey, S.; Luo, H.; Fokoue, A.; Hu, J.; Zhang, P. Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinform.* **2018**, *19*, 476. [CrossRef] [PubMed]
200. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations Ofwords and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems, Proceedings of the Twenty-Seventh Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 December 2013*; NeurIPS: San Diego, CA, USA, 2013.

201. Henikoff, S.; Henikoff, J.G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919. [CrossRef]
202. Nakamura, Y. Codon Usage Tabulated from International DNA Sequence Databases: Status for the Year. *Nucleic Acids Res.* **2000**, *28*, 292. [CrossRef] [PubMed]
203. Marchler-Bauer, A.; Derbyshire, M.K.; Gonzales, N.R.; Lu, S.; Chitsaz, F.; Geer, L.Y.; Geer, R.C.; He, J.; Gwadz, M.; Hurwitz, D.I.; et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **2015**, *43*, D222–D226. [CrossRef] [PubMed]
204. Raimondi, D.; Orlando, G.; Vranken, W.F.; Moreau, Y. Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. *Sci. Rep.* **2019**, *9*, 16932. [CrossRef]
205. Yang, K.K.; Wu, Z.; Bedbrook, C.N.; Arnold, F.H. Learned protein embeddings for machine learning. *Bioinformatics* **2018**, *34*, 2642–2648. [CrossRef]
206. Rodríguez, P.; Bautista, M.A.; Gonzàlez, J.; Escalera, S. Beyond one-hot encoding: Lower dimensional target embedding. *Image Vis. Comput.* **2018**, *75*, 21–31. [CrossRef]
207. Zhang, W.; Du, T.; Wang, J. Deep Learning over Multi-field Categorical Data. In *Advances in Information Retrieval, Proceedings of the 38th European Conference on IR Research, ECIR 2016, Padua, Italy, 20–23 March 2016*; Springer: Berlin/Heidelberg, Germany, 2016.
208. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
209. Chen, Q.; Britto, R.; Erill, I.; Jeffery, C.J.; Liberzon, A.; Magrane, M.; Onami, J.I.; Robinson-Rechavi, M.; Sponarova, J.; Zobel, J.; et al. Quality Matters: Biocuration Experts on the Impact of Duplication and Other Data Quality Issues in Biological Databases. *Genom. Proteom. Bioinform.* **2020**, *18*, 91. [CrossRef] [PubMed]
210. Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; Oliva, A. Learning Deep Features for Scene Recognition Using Places Database. In *Advances in Neural Information Processing Systems, Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014 (NIPS), Montreal, QC, Canada, 8–13 December 2014*; MIT Press: Cambridge, MA, USA, 2014.
211. Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; Xie, X. Gene expression inference with deep learning. *Bioinformatics* **2016**, *32*, 1832–1839. [CrossRef]
212. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting Unreasonable Effectiveness of Data in Deep Learning Era. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
213. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
214. Shaikhina, T.; Khovanova, N.A. Handling limited datasets with neural networks in medical applications: A small-data approach. *Artif. Intell. Med.* **2017**, *75*, 51–63. [CrossRef] [PubMed]
215. Auslander, N.; Wolf, Y.I.; Koonin, E.V. In Silico Learning of Tumor Evolution through Mutational Time Series. *Proc. Natl. Acad. Sci. USA* **2019**, *116*. [CrossRef] [PubMed]
216. Stodden, V.; McNutt, M.; Bailey, D.H.; Deelman, E.; Gil, Y.; Hanson, B.; Heroux, M.A.; Ioannidis, J.P.A.; Taufer, M. Enhancing Reproducibility for Computational Methods. *Science* **2016**, *354*, 1240–1241. [CrossRef]
217. Arora, S.; Pattwell, S.S.; Holland, E.C.; Bolouri, H. Variability in estimated gene expression among commonly used RNA-seq pipelines. *Sci. Rep.* **2020**, *10*, 2734. [CrossRef]
218. Hong, C.S.; Singh, L.N.; Mullikin, J.C.; Biesecker, L.G. Assessing the reproducibility of exome copy number variations predictions. *Genome Med.* **2016**, *8*, 82. [CrossRef]
219. Sandmann, S.; De Graaf, A.O.; Karimi, M.; Van Der Reijden, B.A.; Hellström-Lindberg, E.; Jansen, J.H.; Dugas, M. Evaluating Variant Calling Tools for Non-Matched Next-Generation Sequencing Data. *Sci. Rep.* **2017**, *7*, srep43169. [CrossRef]
220. Montavon, G.; Samek, W.; Müller, K.R. Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal Process. A Rev. J.* **2018**, *73*, 1–15. [CrossRef]
221. Bazen, S.; Joutard, X. The Taylor Decomposition: A Unified Generalization of the Oaxaca Method to Nonlinear Models. In Proceedings of the French Econometrics Conference, Toulouse, France, 14–15 November 2013.
222. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2014**, arXiv:1312.6034v2.
223. Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K.R. Layer-Wise Relevance Propagation: An Overview. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2019.
224. Chicco, D. Ten Quick Tips for Machine Learning in Computational Biology. *BioData Min.* **2017**, *10*, 35. [CrossRef] [PubMed]