*proteomes*

*Review*

# Deep Learning in Phosphoproteomics: Methods and Application in Cancer Drug Discovery

**Neha Varshney [1,2,*] and Abhinava K. Mishra [3,*]**

[1] Division of Biological Sciences, Department of Cellular and Molecular Medicine, University of California, San Diego, CA 93093, USA
[2] Ludwig Institute for Cancer Research, La Jolla, CA 92093, USA
[3] Molecular, Cellular and Developmental Biology Department, University of California, Santa Barbara, CA 93106, USA
[*] Correspondence: nevarshney@ucsd.edu (N.V.); abhinavamishra@ucsb.edu (A.K.M.)

**Abstract:** Protein phosphorylation is a key post-translational modification (PTM) that is a central regulatory mechanism of many cellular signaling pathways. Several protein kinases and phosphatases precisely control this biochemical process. Defects in the functions of these proteins have been implicated in many diseases, including cancer. Mass spectrometry (MS)-based analysis of biological samples provides in-depth coverage of phosphoproteome. A large amount of MS data available in public repositories has unveiled big data in the field of phosphoproteomics. To address the challenges associated with handling large data and expanding confidence in phosphorylation site prediction, the development of many computational algorithms and machine learning-based approaches have gained momentum in recent years. Together, the emergence of experimental methods with high resolution and sensitivity and data mining algorithms has provided robust analytical platforms for quantitative proteomics. In this review, we compile a comprehensive collection of bioinformatic resources used for the prediction of phosphorylation sites, and their potential therapeutic applications in the context of cancer.

**Keywords:** phosphoproteomics; machine learning; deep learning; cancer; post-translational modification; personalized medicine

check for updates

**Citation:** Varshney, N.; Mishra, A.K. Deep Learning in Phosphoproteomics: Methods and Application in Cancer Drug Discovery. *Proteomes* **2023**, *11*, 16. https://doi.org/10.3390/proteomes11020016

Academic Editor: Weiguo Andy Tao

Received: 13 February 2023
Revised: 24 April 2023
Accepted: 25 April 2023
Published: 2 May 2023

## 1. Introduction

Protein phosphorylation is the most widespread post-translational modification (PTM) in eukaryotes and plays a cardinal role in regulating protein functions, such as modulating their intracellular dynamics, stability, subcellular localization, and interaction with other proteins [1,2]. Protein phosphorylation is reversibly controlled by protein kinases (PK) and protein phosphatases (PP) [3]. Protein phosphorylation regulates many cellular processes, including cellular metabolism, cell migration, cell division, proliferation and differentiation, apoptosis, etc. [4–11]. Dysregulated phosphorylation has been identified as a hallmark of many diseases, including numerous cancers, Alzheimer's disease, and diabetes [12–14]. Therefore, understanding protein phosphorylation and its effects on cell signaling is a major endeavor in the post-genomics era.

Recent advances in experimental approaches have immensely helped in the characterization of PTMs. However, the analysis and understanding of PTMs involve several challenges. Efficient and sensitive methods for the detection of PTMs are indispensable. Traditionally, techniques including Edman degradation, mutational analysis, isotopic labelling, or immunochemistry have been used for PTM such as protein phosphorylation discovery [15–17]. Recently, mass spectrometry (MS)-based approaches have shown to be useful in protein phosphorylation identification [18]. MS provides a good platform for the experimental determination of protein phosphorylation sites and high in-depth coverage, and it provides opportunities for ML-based approaches to handle large datasets in public repositories. PTM research has made remarkable progress over the years, especially after the emergence of new computational techniques. Combined with experimental methods, the application of bioinformatics tools in PTM analysis enables a more efficient exploration of the phosphorylation network, resulting in the timely analysis of datasets and providing insights for biological research and drug discovery [19].

Deep learning (DL) in phosphoproteomics refers to the application of machine learning (ML) algorithms to analyze large amounts of data generated from phosphoproteomic experiments. The aim of ML is to identify patterns, classify proteins, and make predictions about protein phosphorylation. The data analysis in phosphoproteomics involves the identification of phosphopeptides based on MS/MS spectra. This can be performed by database searches. The databases report phosphopeptide sequences along with assigned phosphorylation sites. Next, to determine the confidence of each possible phosphorylation site candidate in an identified peptide sequence, several computational algorithms or ML-based approaches can be used. A global understanding of the protein phosphorylation network using these approaches can aid in our understanding of cellular signaling pathways, disease mechanisms, disease onset prediction, drug development, and therapy response in an efficient yet comprehensive manner.

In this brief review, we survey the mainstream tools available to explore the phosphorylation network. Additionally, we present a comparative analysis of these computational tools in terms of technique used, implementation, performance, functionality and limitations from the perspective of a biologist. Finally, we discuss the applications of these phosphoproteomics-based bioinformatics tools in cancer research in identifying novel drug targets and advancing personalized medicine. Hence, this review aims to bridge the gap and emphasize the complementarity between traditional MS-based methods to study phosphoproteomics and the new cutting-edge deep-learning-based prediction methods.

## 2. Methods for Phosphorylation Site Prediction

The computational approaches provide a promising strategy for identification and understanding of phosphorylation sites. Several computational methods have been developed for phosphorylation site prediction over the years. These can be classified into two main categories: algorithm-based and more advanced ML-based methods.

### *2.1. Algorithm-Based Computational Approaches*

In the past, many studies used algorithm-based computational methods to predict phosphorylation sites in which there are no learning algorithms used to gain information directly from data. They can be further classified into simple consensus pattern-based approaches (SCPs) and sequence similarity-based clustering methods (SSs). For example, in 1988, one of the first computational approaches to predict PTM sites was developed, which used the primary sequence of the protein and SCP approach [20]. Other examples of SCPs are PROSITE [21], ELM [22], and HPRD [23], which depend upon the presence of an exact motif surrounding the phosphorylation site. However, SSs-based methods were later designed to provide a high score to a query peptide that has a high similarity score with known phosphorylation peptides, using the sequence similarity measures such as the BLOSUM62 matrix. PostMod [24] and PSEA [25] are examples of this category. These methods have been shown to be inappropriate for large-scale analyses since the performance of these methods in predicting phosphorylation sites is poorer than more advanced ML-based approaches.

### *2.2. Machine Learning (ML)-Based Computational Approaches*

Over the last decade, the integration of ML into a wide range of computational models has improved prediction accuracy and gained a better understanding of protein function and PTMs [26,27]. With the explosion of DL methods, ML-based approaches for phosphorylation site prediction have become more popular. ML is generally the ability of machines to do actions based on prior knowledge and experience [28]. ML-based methods can learn the underlying rules and signatures in the data by tuning and optimizing related parameters during the model training process, resulting in better performance as compared to SCP-based methods. A few examples of ML-based techniques for phosphorylation site prediction are neural network (NN), hidden Markov models (HMM), Bayesian decision theory (BDT), support vector machines, logistic regression (LR), random forest (RF), K-nearest neighbor (KNN), and conditional random fields (CRFs) [29–31]. A few examples of phosphorylation site prediction tools based on these techniques are NetPhos, KinasPhos, DISPHOS, and Ptpset. Most of the databases and phosphorylation-site prediction tools that use different algorithms and ML-based approaches are listed in Tables 1 and 2, respectively. The development of these models have set the benchmark for ML- and DL-based approaches for various PTM predictions.

**Table 1.** List of protein phosphorylation databases.

| Name | Technique | Organisms | Description/Functionality | Ref | Website |
|---|---|---|---|---|---|
| | | | **Kinome** | | |
| Kinomer | HMMER 2.3.2 | Eukaryotes | Annotated classifications for the protein kinase complements of 43 eukaryotic genomes | [32,33] | http://www.compbio.dundee.ac.uk/kinomer/ (accessed on 24 April 2023) |
| KinaseNET | - | Human | Comprehensive source on human kinases | - | http://www.kinasenet.ca/ (accessed on 24 April 2023) |
| | | | Phosphatome | | |
| PTP | - | - | Integrates sequence and structure with cellular and biological functions on protein tyrosine phosphatases | [34] | http://ptp.cshl.edu/ (accessed on 24 April 2023) |
| DEPOD | BLAST | | Comprehensive and informative database on human kinase-phosphatase substrate | [35] | http://depod.bioss.uni-freiburg.de/ (accessed on 24 April 2023) |
| | | | Kinome-Phosphatome | | |
| Phospho.ELM | BLAST | Multiple | Database designed to store in vivo and in vitro phosphorylation | [36] | http://phospho.elm.eu.org/ (accessed on 24 April 2023) |
| PSP (PhosphositePlus) | - | Human, mouse, and rat | Resource that comprehensively curates information about the structure and regulatory interactions of phosphorylation sites | [37] | https://www.phosphosite.org (accessed on 24 April 2023) |
| UniProt | - | Multispecies | A central hub for the collection of functional information on proteins | [38] | https://www.uniprot.org (accessed on 24 April 2023) |
| EPSD (Eukaryotic Phosphorylation site Database) | - | Multiple | A data resource for the collection, curation, integration, and annotation of p-sites in eukaryotic proteins | [39] | http://epsd.biocuckoo.cn (accessed on 24 April 2023) |
| RegPhos 2.0 (regulatory network in protein phosphorylation) | - | Human, mouse, and rat | A comprehensive tool to view intracellular signaling networks by integrating the information of metabolic pathways and protein–protein interactions | [40] | http://140.138.144.141/~RegPhos/ (accessed on 24 April 2023) |
| Phospho3D 2.0 | - | Multiple | A database for the collection of information on the residues surrounding the p-site in space (3D zones) | [41] | http://www.phospho3d.org/ (accessed on 24 April 2023) |

**Table 1.** *Cont.*

| Name | Technique | Organisms | Description/Functionality | Ref | Website |
|---|---|---|---|---|---|
| dbPSP | - | Prokaryotes | Collection of p-sites in prokaryotic phosphoproteins | [42] | http://dbpsp.biocuckoo.cn (accessed on 24 April 2023) |
| LymPHOS | - | Human, mouse | A database for storage, sharing, and visualization of data related with the human T-lymphocyte phosphoproteome | [43] | http://www.lymphos.org (accessed on 24 April 2023) |
| P3DB | - | Plant species | Displays data in a relational, hierarchical manner that integrates proteins, peptides, phosphosites, and spectra for each phosphorylation event | [44] | http://www.p3db.org/ (accessed on 24 April 2023) |
| PHOSIDA (Phosphorylation site database) | - | Multiple | Structural and evolutionary investigation and prediction of phosphosites | [45] | http://phosida.de/ (accessed on 24 April 2023) |
| HPRD (human protein reference database) | BLAST | Human | A database of curated proteomic information including PTMs, kinase/phosphatase motifs, and binding motifs pertaining to human proteins | [46] | http://www.hprd.org (accessed on 24 April 2023) |
| VPTMdb | SVM, NB, RF | Virus | Predicts viral p-Ser | [47] | http://vptmdb.com:8787/VPTMdb/ (accessed on 24 April 2023) |
| pTestis | - | Mouse | Testis phosphorylation sites from various studies were analyzed, integrated with the iGPS prediction results, which present the potential kinase–substrate regulatory relationships | [48] | http://ptestis.biocuckoo.org/ (accessed on 24 April 2023) |
| PhosphoPep | BLAST | Multiple | Database of protein phosphorylation sites for systems level research in model organisms | [49] | http://www.unipep.org/phosphopep/index.php (accessed on 24 April 2023) |
| PhosphoPOINT | PPI, BLASTP | Human | Annotates interactions among kinases, with their downstream substrates and interacting phosphoproteins | [50] | http://kinase.bioinformatics.tw/ (accessed on 24 April 2023), https://bioregistry.io/registry/phosphopoint.protein (accessed on 24 April 2023) |

**Table 2.** A comprehensive list of computational tools used for phosphoproteomic data analysis, including phosphorylation site prediction, predicting kinases, and phosphoproteomic data annotation. Column headings are as follows. Name: Name of the tool; Technique: the machine learning technique used; Organisms: list of organisms the tool is applicable; Description/Functionality: important properties of the tool in terms of its functions; Reference; the paper describing that tool, website: the address of that tool's web implementation or source of access (if applicable). ANN: artificial neural network; PSSM: position-specific scoring matrices; ZSL: zero-short learning; RNN: recurrent neural network; BLR: bagged logistic regression; LR: logistic regression algorithm; DNN: deep neural network; PPI: protein–protein interaction; XGBoost: extreme gradient boosting; MLS: motif length selection (MLS); LSTM: long short-term memory.

| Name | Technique | Organisms | Description/Functionality | Ref | Website |
|---|---|---|---|---|---|
| NetPhos 3.1 | ANN | Multiple | Generates NN predictions for serine, threonine, and tyrosine phosphorylation sites in eukaryotic proteins. Utilizes sequence composition features, both generic and kinase specific predictions | [30] | https://mybiosoftware.com/tag/netphos (accessed on 24 April 2023) |
| NetPhosK | ANN | Eukaryotes | Kinase-specific phosphorylation sites prediction | [51] | https://www.hsls.pitt.edu/obrc/index.php?page=URL1117048165 (accessed on 24 April 2023) |
| Scansite 2.0 | PSSM | Human | A tool built on experimental binding and/or substrate information from oriented peptide library screening and phage display experiments, together with detailed biochemical characterization to derive a weight matrix-based scoring algorithm that predicts protein–protein interactions and sites of phosphorylation | [52] | https://scansite4.mit.edu/#home (accessed on 24 April 2023) |
| PhosphoNet | PSSM | Human | An open source of putative phosphosites predicted after improvisation of kinase substrate prediction algorithm to the primary structure of proteins | [53] | http://www.phosphonet.ca (accessed on 24 April 2023) |
| Predphospho | SVM | Human | Predicts the changes in phosphorylation sites caused by amino acid variations at intra- and interspecies levels | [54] | http://www.ngri.re.kr/proteo/PredPhospho.htm (accessed on 24 April 2023) |
| NetworkKIN | ANN, PSSM | Human | Uses probabilistic protein association network (string) to model the context of kinases and substrates, combined with consensus sequence motifs | [55] | https://networkin.info/ (accessed on 24 April 2023) |
| jEcho | Weight vector | Human | Phosphorylation sites of kinases | [56] | http://www.healthinformaticslab.org/supp/resources.php (accessed on 24 April 2023) |
| PhoScan | Scoring function | Human | Predicts kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach | [57] | http://bioinfo.au.tsinghua.edu.cn/phoscan/ (accessed on 24 April 2023) |

**Table 2.** *Cont.*

| Name | Technique | Organisms | Description/Functionality | Ref | Website |
|---|---|---|---|---|---|
| Predphos | SVM | Multiple | Structural-based prediction of phosphorylation sites, hybrid approach, which incorporates bootstrap resampling technique, SVM-based fusion classifiers and majority voting strategy | [58] | No tool link |
| NetPhosYeast | ANN | Yeast | Prediction of protein phosphorylation sites in yeast | [59] | https://services.healthtech.dtu.dk/service.php?NetPhosYeast-1.0 (accessed on 24 April 2023) |
| GPS 6.0 (group-based prediction system) | MLS, PSSM, GA | Mammalian | Protein phosphorylation sites and their cognate kinases (addresses false positive rates in prediction) | [60,61] | http://gps.biocuckoo.org/ (accessed on 24 April 2023) |
| iGPS | GPS with PPI | Human | It is a GPS algorithm with the interaction filter, or in vivo GPS mainly for the prediction of in vivo site-specific kinase-substrate relation (ssKSRs) | [62] | http://igps.biocuckoo.org/links.php (accessed on 24 April 2023) |
| PPRED | PSSM, SVM | - | Incorporates only evolutionary information of PSSM profile of the proteins in predicting phosphorylation sites | [63] | http://www.cse.univdhaka.edu/~ashis/ppred/index.php (accessed on 24 April 2023) |
| Phos3D | SVM | Human | Prediction of phosphorylation sites (p-sites) in proteins, originally designed to investigate the advantages of including spatial information in p-site prediction | [64] | https://phos3d.mpimp-golm.mpg.de/cgi-bin/index.py (accessed on 24 April 2023) |
| DAPPLE 2 | BLAST | Human | Homology-based prediction of phosphorylation sites | [65] | http://saphire.usask.ca/saphire/dapple2 (accessed on 24 April 2023) |
| EMBER | CNN + RNN | Multiple | Embedding-based multilabel prediction of phosphorylation events (EMBER), a DL method that integrates kinase phylogenetic information and motif-dissimilarity information into a multilabel classification model for the prediction of kinase motif phosphorylation events | [66] | https://github.com/gomezlab/EMBER (accessed on 24 April 2023) |
| KinomeXplorer | NetworKIN algorithm, a novel Bayesian scoring scheme | Human and major eukaryotes | Analyze phosphorylation-dependent protein interaction networks | [67] | http://kinomexplorer.info/ (accessed on 24 April 2023) |
| PhosTransfer | CNN | Info not available | Hierarchical kinase-specific phosphorylation site (KPS) prediction | [68] | https://github.com/yxu132/PhosTransfer (accessed on 24 April 2023) |

**Table 2.** *Cont.*

| Name | Technique | Organisms | Description/Functionality | Ref | Website |
|---|---|---|---|---|---|
| MusiteDeep | CNN/CapsNet | Human | Prediction and visualization for multiple PTMs and simultaneously potential PTM cross-talks | [69] | https://www.musite.net (accessed on 24 April 2023) |
| PROSPECT | CNN | *E. coli* | Predicts histidine phosphorylation sites from sequence information | [70] | https://bio.tools/prospect-web (accessed on 24 April 2023) |
| DeepKinZero | ZSL | Human | Predicts the kinase acting on a phosphosite for kinases with no known phosphosite information | [71] | https://github.com/Tastanlab/DeepKinZero (accessed on 24 April 2023) |
| DeepPPSite | LSTM | Mammals and *Arabidopsis thaliana* | Long short-term memory (LSTM) recurrent network for predicting phosphorylation sites | [72] | https://github.com/saeed344/DeepPPSite (accessed on 24 April 2023) |
| DeepIPs | CNN + LSTM | Human | Identification of phosphorylation sites using deep learning method | [73] | https://github.com/linDing-group/DeepIPs (accessed on 24 April 2023) |
| Rice_Phospho 1.0 | SVM | Rice | Predicts protein phosphorylation sites in rice | [74] | http://bioinformatics.fafu.edu.cn/rice_phospho1.0 (accessed on 24 April 2023) |
| Yeast KID | - | Yeast | The first literature-curated database for kinases that integrates a series of HTP and LTP, genetic, physical, and biochemical experimental evidence with the goal of establishing known kinase–substrate relationships. | [75] | http://www.moseslab.csb.utoronto.ca/KID/ (accessed on 24 April 2023) |
| AutoMotif | SVM | | The service uses a supervised support vector machine approach to predict various types of phosphorylation sites in proteins | [76] | http://ams2.bioinfo.pl/ (accessed on 24 April 2023) |
| PhosIDN | Multilayer DNN | Human | An integrated DNN approach for improving protein phosphorylation site prediction by combining sequence and protein–protein interaction information | [77] | https://github.com/ustchangyuanyang/PhosIDN (accessed on 24 April 2023) |
| DeepPhos | CNN | Human | Uses densely connected CNN for kinase-specific phosphorylation site prediction | [29] | https://github.com/USTC-HIlab/DeepPhos (accessed on 24 April 2023) |
| Chlamy-EnPhosSite | CNN + LSTM | *Chlamydomonas reinhardtii* | Can predict novel sites of phosphorylation within the entire C. reinhardtii proteome | [78] | https://github.com/dukkakc/Chlamy-EnPhosSite (accessed on 24 April 2023) |

**Table 2.** *Cont.*

| Name | Technique | Organisms | Description/Functionality | Ref | Website |
|---|---|---|---|---|---|
| DeepPSP | DNN, SENet, Bi-LSTM | ? | Uses both local and global sequence information to improve phosphorylation site prediction performance | [79] | https://github.com/DeepPSP (accessed on 24 April 2023) |
| Predikin 2.0 | PSSM | Human | Utilizes the kinase sequence to build scoring matrices based on key residues in the kinase catalytic domain that are known from structural analysis to interact with the substrate phosphorylation site. | [80] | http://predikin.biosci.uq.edu.au (accessed on 24 April 2023) |
| KinasePhos2.0 | SVM | Human? | Predicts phosphorylation sites based on protein sequence profile and protein coupling pattern and the type of kinase that acts at each site | [81] | http://kinasephos2.mbc.nctu.edu.tw/document.html, https://bio.tools/kinasephos_2.0 (accessed on 24 April 2023) |
| KinasePhos 3.0 | SVM, XGBoost | Human and others | Provides comprehensive annotations of kinase-specific phosphorylation sites on multiple proteins. Shapley additive explanations (SHAP) was integrated to increase the feature interpretability | [82] | https://awi.cuhk.edu.cn/KinasePhos/index.html, https://github.com/tom-209/KinasePhos-3.0-executable-file (accessed on 24 April 2023) |
| DISPHOS (disorder-enhanced phosphorylation predictor) | BLR | Human | Position-specific amino acid frequencies and disorder information is used to improve the discrimination between phosphorylation and non-phosphorylation sites | [83] | http://www.ist.temple.edu/DISPHOS (accessed on 24 April 2023) |
| pkaPS | Scoring function | Human | Phosphorylation sites of PKA | [84] | http://mendel.imp.univie.ac.at/sat/pkaPS (accessed on 24 April 2023) |
| Quokka | Seqeunce scoring function + LR | Human | Predicts kinase-specific phosphorylation sites | [85] | http://quokka.erc.monash.edu/ (accessed on 24 April 2023) |
| PHOSIDA (phosphorylation site database) | SVM | Multiple | Structural and evolutionary investigation and prediction of phosphosites | [45] | http://phosida.de/ (accessed on 24 April 2023) |
| VPTMdb | SVM, NB, RF | Virus | Predicts viral p-Ser | [47] | http://vptmdb.com:8787/VPTMdb/ (accessed on 24 April 2023) |

## 3. Framework of ML-Based Approaches for Phosphorylation Site Prediction

Generally, ML-based computational approaches for phosphorylation site prediction are developed using the following five steps: (1) dataset preparation; (2) selection of encoding methods; (3) building prediction models; and (4) performance evaluation and development of a web-server (Figure 1).

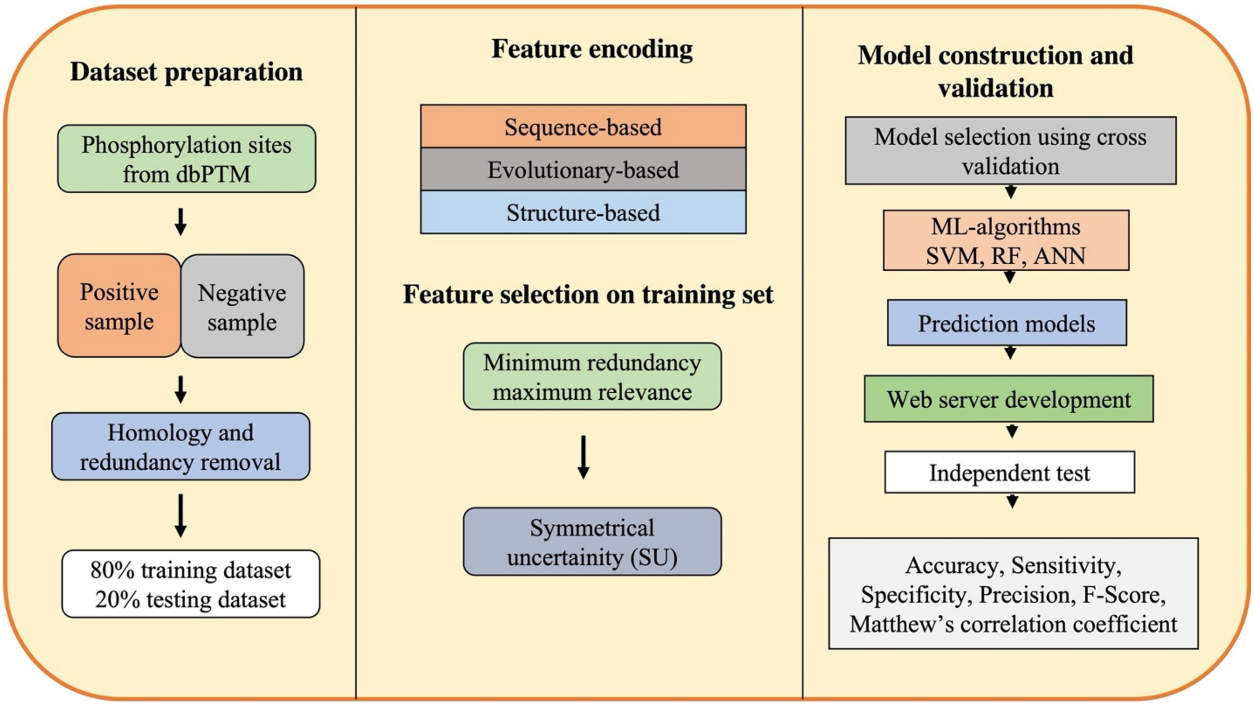
Dataset preparation
Phosphorylation sites from dbPTM
Positive sample
Negative sample
Homology and redundancy removal
80% training dataset 20% testing dataset
Feature encoding
Sequence-based
Evolutionary-based
Structure-based
Feature selection on training set
Minimum redundancy maximum relevance
Symmetrical uncertainity (SU)
Model construction and validation
Model selection using cross validation
ML-algorithms SVM, RF, ANN
Prediction models
Web server development
Independent test
Accuracy, Sensitivity, Specificity, Precision, F-Score, Matthew's correlation coefficient

**Figure 1.** Workflow of machine learning-based computational approaches for phosphorylation site prediction (SVM: support vector machine, RF: random forest, ANN: artificial neural network).

### *3.1. Dataset Preparation*

The first step for phosphorylation site prediction is dataset preparation that includes the extraction of experimentally validated phosphorylation sites from the publicly available databases, dbPTMs, and the literature [86]. A few of the protein phosphorylation databases are enlisted in Table 1. An extracted dataset must include both positive and negative datasets. The fragments or peptides that have the phosphorylated residues (S, T, Y) compiled from the aforementioned dbPTMs are considered a positive dataset. The S, T, Y amino acids in the experimental peptides with no phospho-groups on them are considered as negative datasets. Almost all studies use databases to gather positive samples, yet, selecting the negative dataset is the most challenging step. While a particular residue that can be phosphorylated can be validated experimentally, a particular residue that is not phosphorylated under any conditions is difficult to prove experimentally. Therefore, databases contain thousands of known phosphorylation sites but do not contain phosphorylation sites that are known to be unphosphorylated. A few criteria to apply while extracting a negative dataset include the selection of a site that should not have been reported as a phosphorylation site in the positive dataset, the thresholding of a solvent accessible area of the protein, etc. Following the construction of these datasets, the next step is the removal of homologous and redundant sequences. The Cluster Database at high identity with tolerance (CD-hit) is a popular program to detect and filter similar sequences [87]. The threshold of identity between sequences is considered to range from 30% to 90%. These prepared datasets are used as benchmark data, which are eventually divided into 80% training data for learning and 20% testing set for model validation. The training data are used for feature selection and ML model generation, which also includes a 5-fold internal cross-validation of the trained classifiers' performance. The testing dataset is used for further assessment and validation of ML models using various statistical measures.

### *3.2. Feature Encoding and Selection*

For feature encoding, all the proteins are partitioned into polypeptides in such a way that the target residue is placed at the center of the peptide. Each polypeptide sequence (both positive and negative datasets) is encoded as a numerical feature vector according to the appropriate biological descriptors, such as amino acid composition [88], similarity score to the known motifs [89], and evolutionary and structural properties [90]. Occasionally, to enhance the prediction performance, all features are pooled, thus resulting in a combination of features to generate learning models. Feature selection methods are then used to choose the most relevant features while minimizing the redundancy in the data and further improving the model performance by reducing its computational time. The feature selection is performed at two levels: minimum redundancy maximum relevance (mRMR) approach followed by symmetrical uncertainty (SU) selection method. mRMR is a widely used feature selection method approach that ranks the features while taking into consideration their importance to the classification variable along with the redundancy among the features themselves [91]. The SU attribute evaluation method weighs the merit of an attribute by determining its uncertainty with reference to other sets of attributes [92].

### *3.3. Model Construction and Validation*

Once the features have been extracted, data are used to train a model/classifier for PTM site prediction. At this point, different classifiers are trained, and based on the performance of each classifier, a suitable classifier is selected. One of the most popular ML-based methods used for predicting sites is SVM. SVMs are a set of points in the n-dimensional space of data that define the boundaries of categories. It is a maximum margin classifier in which data are separated by a hyperplane, provided that they have the highest margin over the data. RF is one of the other well-known ML-based algorithms used for phospho-sites. RF is a supervised learning algorithm; as the name suggests, it builds forests randomly whereby forests are groups of decision trees. Once several decision trees are made, they are merged to make more stable and accurate stable predictions. The classifier is trained on a subset of assembled dataset (training dataset) after parameter optimization and, finally, the predictor is ready to be assessed for performance and compared with other methods. The prediction performance of the model is assessed by its accuracy (proportion of correct positive and negative predictions), sensitivity or true-positive rate, F-score, and Mathew's correlation coefficient (MCC). An independent test set is carried out to evaluate the performance of the classifier and further verify its practicality.

## 4. Use of Machine Learning-Based Approaches for Phosphoproteome Prediction in Cancers

Quantitative phosphoproteomics-based approaches are powerful tools to investigate the signaling pathways and cross-talk networks in cancer cells, assess disease prognosis, and develop personalized treatments [8,9,93–95]. Integrating ML and multi-omics data to classify cancer stages or accelerate the prognosis of the disease in the early stages is an active area of investigation. Many in silico approaches for predicting the phosphoproteomic profiles of cancer patients have gained attention in recent years. Sequence-based approaches to predict phosphoproteomes have limited accuracy as phosphoproteomic profiles may vary considerably across cancer patients [96]. Further, MS-based approaches are time-consuming and expensive. Therefore, new computational methods to predict phosphoproteomic profiles across cancer patients are now widely investigated. Several models have been developed and used to predict phosphoproteome in cancer cells, discover biomarkers, patient-specific drug targets, individualized prediction of drug response, and clinical outcomes and toxicity [95,97–101] (Figure 2).

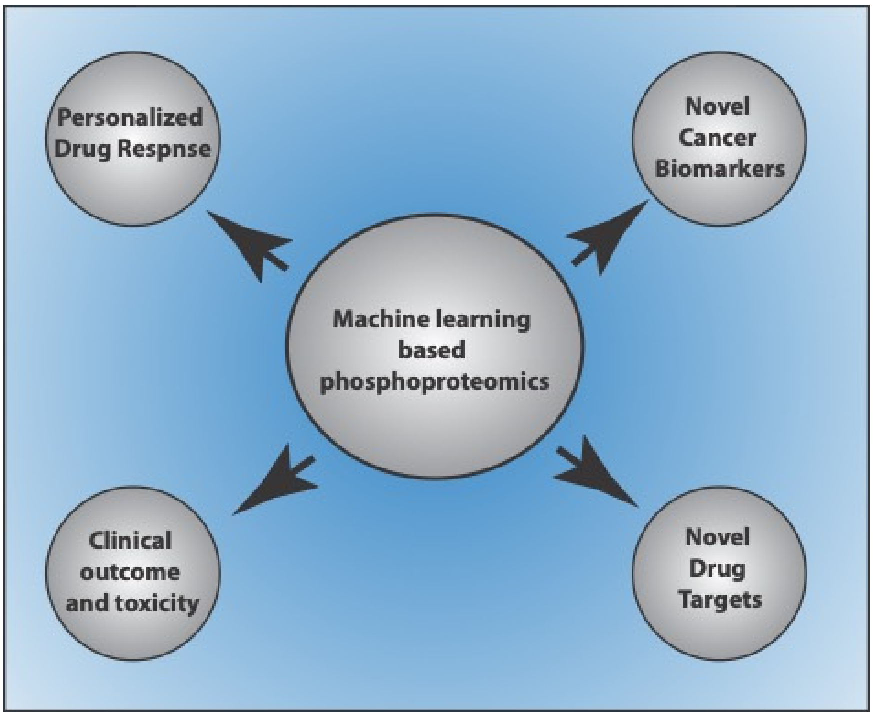
Personalized Drug Respnse
Novel Cancer Biomarkers
Machine learning based phosphoproteomics
Clinical outcome and toxicity
Novel Drug Targets

**Figure 2.** Application of machine learning-based phosphoproteome prediction studies in cancer.

### *4.1. Machine Learning-Based Approaches for Phosphoproteome-Based Biomarker Prediction*

The Cancer Genome Atlas (TCGA), the National Cancer Institute (NCI), and the Clinical Proteomic Tumor Analysis Consortium (CPTAC) are valuable resources that provide a collection of genomic, transcriptomic, proteomic, and phosphoproteomic data for a variety of cancer types. Artificial intelligence (AI) can be used to train these datasets to create algorithms that can predict patient-specific outcomes by predicting biomarkers. For example, using the Boruta algorithm to identify mutant genes involved in the vascular invasion from TCGA, the National Institute of Health, Medical Research, and AMC databases, a gene signature was identified and a recurrence prediction model for recurrence for HCC patients was established [102]. A convolutional NN algorithm was used to analyze proteomics and histology imaging datasets generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) from clear cell renal cell carcinoma patients. This study reported a robust correlation between diagnostic markers and predictions generated by the imaging-trained classification model [103]. Joint learning (JL) is a type of ML method to predict proteome from the transcriptome. This method was developed using a training dataset by NCI-CPTAC and TCGA, consisting of proteomic, phosphoproteomic, and transcriptomic data from 77 breast and 105 ovarian cancer samples. In this powerful model, a gene-specific regulatory network was trained, followed by creating a cross-tissue model by JL, the shared regulatory networks and pathways across many cancer tissues. Such a robust model can help facilitate biomarker discoveries for high- and low-risk patients in survival analyses with different clinical outcomes due to the activation of different functional pathways [104]. Further, the proteome complexity across cancer types and within the patient-specific cohort can also be effectively studied using these models, whereas the traditional approaches may have limited scope to address these issues.

Further, to predict the drug response and design rational combination therapies, a recent study used seven targeted anticancer drugs in 35 non-small cell lung cancer (NSCLC) cell lines and 16 samples of pleural effusions from NSCLC and analyzed dynamic changes in 52 phosphoproteins. They developed an orthogonal ML approach to predict drug response and rational combination therapy. Such studies can supplement the existing methods of using gene mutations to predict biomarkers by utilizing the proteomics data and predict treatment choices and therapy outcomes based on the dynamic proteome complexity [98].

### 4.2. Machine Learning-Based Approaches for Phosphoproteome-Based Patient-Specific Drug Targets and Responses

ML is becoming increasingly popular and valuable in enhancing our current understanding of established or new molecular targets in regulating stemness and cancer metastasis. These studies are key to identifying novel phosphoproteome-based drug targets for hard-to-treat cancers. In a recent in-depth global and phosphoproteomic analyses of tumor cells, using protein structure modeling and interface prediction-guided mutagenesis, the interaction between CD44 and CD81 in extracellular vesicles (EVs) secretion was identified [100]. EVs are the drivers of breast cancer stemness and metastasis in triple-negative breast cancer (TNBC). Hence, this study is seminal to identifying new molecular drug targets with the help of ML approaches. Another study analyzed the phosphoproteomes of cholangiocarcinoma cell lines and patient tumors using MS-based phosphoproteomics and computational methods to identify patient-specific drug targets. This study identified the inhibitors of histone deacetylase and PI3K pathway members as high-ranking therapies to use in primary cholangiocarcinoma by the drug ranking using machine learning (DRUML) algorithm [97]. Drug ranking using ML (DRUML) has also been successfully applied to predict the efficacy of anticancer drugs [105].

KSTAR is graph- and statistics-based algorithm that can capture patient-specific kinase activities from phosphoproteomic data. This algorithm was applied to clinical breast cancer phosphoproteomic data. The study reported that the predicted kinase activity profiles could successfully identify misclassified HER2-positive breast cancer patients. In addition, the algorithm can also identify the likelihood of clinically diagnosed HER2-negative patients to respond to HER2-targeted therapy [106]. Thus, in addition to identifying novel drug targets, ML-based studies are also actively contributing to our current understanding of patient-specific drug responses.

Cellular immunotherapies are a form of personalized medicine that has revolutionized cancer treatment. However, only a subset of patients responds to immunotherapy; hence, there is vast room for improvement. In a recent study, ML-based algorithms were applied to MS-based serum proteomics signatures to predict the response and toxicity of immunotherapy. Datasets from advanced non-small cell lung cancer and malignant melanoma patients were used in this study. Interestingly, the algorithm was able to effectively categorize patients into groups with good and poor treatment outcomes independent of the biomarker signatures [99].

To understand the disease progression and therapy outcome and to identify new drug targets, a holistic understanding of the complex phosphoproteome in cancer is required. This will involve a combination of mass spectrometry-based phosphoproteomics, together with databases and bioinformatics tools to capture the actual, real-time activity of kinases. Such tools could be valuable to establish a phosphoproteomics-based personalized medicine platform for hard-to-treat cancers.

## 5. Conclusions and Future Perspective

The function of a protein is strongly affected by the post-translational chemical modifications that play important functions in a myriad of cellular processes. Therefore, PTM identification is critical for the understanding of molecular functions and diseases. The considerable amount of PTM data generated from the in-depth MS-based experimental approaches could be used to support the development of downstream computational identification methods. DL is a highly effective computational approach to understand large and complex datasets to predict PTMs. In recent years, several DL methods have been developed to predict PTM sites with high efficiency. While these tools have shed light on the quicker, efficient, and less labor-intensive ways on the discovery of phosphorylation site prediction, there are some common weaknesses in assessing these methods, and various factors should be considered in deciding which tool to choose. The most critical factor relevant for the evaluation of prediction tools is the motif size and proper biological context. Another important factor relevant for consideration in PTM predictor construction is the quality of underlying data, including the amount and redundancy of example substrate protein sequences and the level of authenticity. There are several DL algorithms employed; however, each model has its own advantages and disadvantages. In many models, PTM sites are predicted based on sequence information, physical properties, chemical properties, and protein structure properties, but there is still room for approaches that are based on reduced amino acid compositions [107–109].

Thousands of phosphorylation sites have been identified for different proteins by MS; however, the kinase responsible for the phosphorylation of that amino acid in a few of the reported datasets is missing. Therefore, there is a need to develop databases which could bridge the gap between the number of experimentally identified phosphorylation sites and the number of phosphorylation sites for which the modifying kinase is known. While PTM identification can be implemented with DL-based methods in a non-invasive, efficient, and low-cost way, there is still a caveat if these algorithms can be directly used for diseases diagnoses. The over-arching problem is the false-positive rate, which is not ideal for its application in healthcare studies where every misdiagnosis can pose a danger to a patient's health. An ideal model is characterized by high sensitivity and a very low false-positive prediction rate. Therefore, further research is required to evaluate more state-of-the-art frameworks so that these techniques could be applied in clinical practice more effectively.

A phosphorylation event is dynamic and cell type-specific and cannot be traced in a heterogenous cell population, highlighting the importance of analyzing phosphorylation events at the single-cell level for complex samples, such as tissues and organs. With the advent of single-cell proteomics, the adaptation of phosphoproteomics profiling to single-cells has revolutionized the field in uncovering the heterogeneity in signaling networks, complementing single-cell genomics and transcriptomics [110–112]. Therefore, we believe that an integration of computational and biochemical approaches will form the basis for the future development of methods that can reconstruct trans-regulatory networks for heterogeneous cells in single-cell multi-omics data [113]. Another forth-coming area of research in this field is the characterization of cross-talk between different types of PTMs.

Mass spectrometry is one of the key platforms for proteomic analyses that involves either a 'bottom-up' or a 'top-down' proteomics approach. The traditional 'bottom-up' approach employs the digestion of intact proteins into peptides, followed by introduction into the mass spectrometer for fragmentation/detection. Majority of the ML-based methods run smoothly on the bottom-up proteomics data. In the 'top-down' approach, the proteins are ionized directly and the intact fragmented proteins rather than digestive peptides are used in the analysis [114]. Many phosphoproteins have been studied using the top-down approach [115–117]. However, one of the major challenges in top-down proteomics data analysis is the complexity of the high-resolution top-down mass spectra that involves centroiding, deconvolution, proteoform identification, and quantification [118]. A number of algorithm- and ML-based approaches are now actively being developed to enhance the predictions in the top-down proteomics. These methods will be extremely valuable resources that will aid into our understanding of proteoform complexity and improve the performance of disease diagnosis and drug target discovery.

Recently, ensembled learning-based feature selection methods were employed to explore the nature of the phosphorylation of SARS-CoV-2 to contribute to SARS-CoV-2 drug discovery [119]. Finally, in the era of personalized medicine, ML-based approaches in phosphoproteome studies will play an instrumental role both in understanding the disease mechanisms and in identifying new therapy targets. ML-based approaches will be valuable in discovering novel biomarkers, advance our current understanding of patient-specific drug targets and drug responses, and facilitate cancer stage classification.

**Author Contributions:** Conceptualization, N.V. and A.K.M.; writing—original draft preparation, N.V. and A.K.M.; writing—review and editing, N.V. and A.K.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We would like to acknowledge the work of those researchers that could not be cited in the present article due to space limitations. The study used PubMed and Google Scholar search engines for the literature search.

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

## References

1. Ubersax, J.A.; Ferrell, J.E. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 530–541. [CrossRef] [PubMed]
2. Deribe, Y.L.; Pawson, T.; Dikic, I. Post-translational modifications in signal integration. *Nat. Struct. Mol. Biol.* **2010**, *17*, 666–672. [CrossRef] [PubMed]
3. Hunter, T. Protein kinases and phosphatases: The Yin and Yang of protein phosphorylation and signaling. *Cell* **1995**, *80*, 225–236. [CrossRef]
4. Varshney, N.; Schaekel, A.; Singha, R.; Chakraborty, T.; van Wijlick, L.; Ernst, J.F.; Sanyal, K. A surprising role for the Sch9 protein kinase in chromosome segregation in Candida albicans. *Genetics* **2015**, *199*, 671–674. [CrossRef] [PubMed]
5. Varshney, N.; Sanyal, K. Aurora kinase Ipl1 facilitates bilobed distribution of clustered kinetochores to ensure error-free chromosome segregation in Candida albicans. *Mol. Microbiol.* **2019**, *112*, 569–587. [CrossRef]
6. Varshney, N.; Som, S.; Chatterjee, S.; Sridhar, S.; Bhattacharyya, D.; Paul, R.; Sanyal, K. Spatio-temporal regulation of nuclear division by Aurora B kinase Ipl1 in Cryptococcus neoformans. *PLoS Genet.* **2019**, *15*, e1007959. [CrossRef]
7. Humphrey, S.J.; James, D.E.; Mann, M. Protein phosphorylation: A major switch mechanism for metabolic regulation. *Trends Endocrinol. Metab.* **2015**, *26*, 676–687. [CrossRef]
8. Mishra, A.K.; Sharma, V.; Mutsuddi, M.; Mukherjee, A. Signaling cross-talk during development: Context-specific networking of Notch, NF-κB and JNK signaling pathways in Drosophila. *Cell. Signal.* **2021**, *82*, 109937. [CrossRef]
9. Mishra, A.K.; Sachan, N.; Mutsuddi, M.; Mukherjee, A. Kinase active Misshapen regulates Notch signaling in Drosophila melanogaster. *Exp. Cell Res.* **2015**, *339*, 51–60. [CrossRef]
10. Cohen, P. The origins of protein phosphorylation. *Nat. Cell Biol.* **2002**, *4*, E127–E130. [CrossRef]
11. Cohen, P. The role of protein phosphorylation in the hormonal control of enzyme activity. *Eur. J. Biochem.* **1985**, *151*, 439–448. [CrossRef]
12. Meyerovitch, J.; Backer, J.M.; Kahn, C.R. Hepatic phosphotyrosine phosphatase activity and its alterations in diabetic rats. *J. Clin. Investig.* **1989**, *84*, 976–983. [CrossRef]
13. Hijazi, M.; Smith, R.; Rajeeve, V.; Bessant, C.; Cutillas, P.R. Reconstructing kinase network topologies from phosphoproteomics data reveals cancer-associated rewiring. *Nat. Biotechnol.* **2020**, *38*, 493–502. [CrossRef] [PubMed]
14. Blume-Jensen, P.; Hunter, T. Oncogenic kinase signalling. *Nature* **2001**, *411*, 355–365. [CrossRef] [PubMed]
15. Kettenbach, A.N.; Rush, J.; Gerber, S.A. Absolute quantification of protein and post-translational modification abundance with stable isotope-labeled synthetic peptides. *Nat. Protoc.* **2011**, *6*, 175–186. [CrossRef] [PubMed]
16. Ross, F.E.; Zamborelli, T.; Herman, A.C.; Yeh, C.-H.; Tedeschi, N.I.; Luedke, E.S. Detection of acetylated lysine residues using sequencing by edman degradation and mass spectrometry. In *Techniques in Protein Chemistry*; Elsevier: Amsterdam, The Netherlands, 1996; Volume 7, pp. 201–208. ISBN 9780124735569.
17. Fuchs, S.M.; Strahl, B.D. Antibody recognition of histone post-translational modifications: Emerging issues and future prospects. *Epigenomics* **2011**, *3*, 247–249. [CrossRef]
18. Witze, E.S.; Old, W.M.; Resing, K.A.; Ahn, N.G. Mapping protein post-translational modifications with mass spectrometry. *Nat. Methods* **2007**, *4*, 798–806. [CrossRef]
19. Paul, P.; Muthu, M.; Chilukuri, Y.; Haga, S.W.; Chun, S.; Oh, J.-W. In silico tools and phosphoproteomic software exclusives. *Processes* **2019**, *7*, 869. [CrossRef]
20. Nakai, K.; Kanehisa, M. Prediction of in-vivo modification sites of proteins from their primary structures. *J. Biochem.* **1988**, *104*, 693–699. [CrossRef]
21. Sigrist, C.J.A.; de Castro, E.; Cerutti, L.; Cuche, B.A.; Hulo, N.; Bridge, A.; Bougueleret, L.; Xenarios, I. New and continuing developments at PROSITE. *Nucleic Acids Res.* **2013**, *41*, D344–D347. [CrossRef]

22. Puntervoll, P.; Linding, R.; Gemünd, C.; Chabanis-Davidson, S.; Mattingsdal, M.; Cameron, S.; Martin, D.M.A.; Ausiello, G.; Brannetti, B.; Costantini, A.; et al. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **2003**, *31*, 3625–3630. [CrossRef] [PubMed]
23. Peri, S.; Navarro, J.D.; Amanchy, R.; Kristiansen, T.Z.; Jonnalagadda, C.K.; Surendranath, V.; Niranjan, V.; Muthusamy, B.; Gandhi, T.K.B.; Gronborg, M.; et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **2003**, *13*, 2363–2371. [CrossRef] [PubMed]
24. Jung, I.; Matsuyama, A.; Yoshida, M.; Kim, D. PostMod: Sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC Bioinform.* **2010**, *11* (Suppl. S1), S10. [CrossRef] [PubMed]
25. Suo, S.-B.; Qiu, J.-D.; Shi, S.-P.; Chen, X.; Liang, R.-P. PSEA: Kinase-specific prediction and analysis of human phosphorylation substrates. *Sci. Rep.* **2014**, *4*, 4524. [CrossRef]
26. Avery, C.; Patterson, J.; Grear, T.; Frater, T.; Jacobs, D.J. Protein Function Analysis through Machine Learning. *Biomolecules* **2022**, *12*, 1246. [CrossRef]
27. Auslander, N.; Gussow, A.B.; Koonin, E.V. Incorporating Machine Learning into Established Bioinformatics Frameworks. *Int. J. Mol. Sci.* **2021**, *22*, 2903. [CrossRef]
28. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef]
29. 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]
30. Blom, N.; Gammeltoft, S.; Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **1999**, *294*, 1351–1362. [CrossRef]
31. Plewczyński, D.; Tkacz, A.; Godzik, A.; Rychlewski, L. A support vector machine approach to the identification of phosphorylation sites. *Cell. Mol. Biol. Lett.* **2005**, *10*, 73–89.
32. Miranda-Saavedra, D.; Barton, G.J. Classification and functional annotation of eukaryotic protein kinases. *Proteins* **2007**, *68*, 893–914. [CrossRef] [PubMed]
33. Martin, D.M.A.; Miranda-Saavedra, D.; Barton, G.J. Kinomer v. 1.0: A database of systematically classified eukaryotic protein kinases. *Nucleic Acids Res.* **2009**, *37*, D244–D250. [CrossRef] [PubMed]
34. Andersen, J.N.; Del Vecchio, R.L.; Kannan, N.; Gergel, J.; Neuwald, A.F.; Tonks, N.K. Computational analysis of protein tyrosine phosphatases: Practical guide to bioinformatics and data resources. *Methods* **2005**, *35*, 90–114. [CrossRef] [PubMed]
35. Damle, N.P.; Köhn, M. The human DEPhOsphorylation Database DEPOD: 2019 update. *Database* **2019**, *2019*, baz133. [CrossRef]
36. Dinkel, H.; Chica, C.; Via, A.; Gould, C.M.; Jensen, L.J.; Gibson, T.J.; Diella, F. Phospho.ELM: A database of phosphorylation sites–update 2011. *Nucleic Acids Res.* **2011**, *39*, D261–D267. [CrossRef]
37. Hornbeck, P.V.; Zhang, B.; Murray, B.; Kornhauser, J.M.; Latham, V.; Skrzypek, E. PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Res.* **2015**, *43*, D512–D520. [CrossRef]
38. Bairoch, A.; Apweiler, R.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. The universal protein resource (uniprot). *Nucleic Acids Res.* **2005**, *33*, D154–D159. [CrossRef]
39. Lin, S.; Wang, C.; Zhou, J.; Shi, Y.; Ruan, C.; Tu, Y.; Yao, L.; Peng, D.; Xue, Y. EPSD: A well-annotated data resource of protein phosphorylation sites in eukaryotes. *Brief. Bioinform.* **2021**, *22*, 298–307. [CrossRef]
40. Huang, K.-Y.; Wu, H.-Y.; Chen, Y.-J.; Lu, C.-T.; Su, M.-G.; Hsieh, Y.-C.; Tsai, C.-M.; Lin, K.-I.; Huang, H.-D.; Lee, T.-Y.; et al. RegPhos 2.0: An updated resource to explore protein kinase-substrate phosphorylation networks in mammals. *Database* **2014**, *2014*, bau034. [CrossRef]
41. Zanzoni, A.; Carbajo, D.; Diella, F.; Gherardini, P.F.; Tramontano, A.; Helmer-Citterich, M.; Via, A. Phospho3D 2.0: An enhanced database of three-dimensional structures of phosphorylation sites. *Nucleic Acids Res.* **2011**, *39*, D268–D271. [CrossRef]
42. Shi, Y.; Zhang, Y.; Lin, S.; Wang, C.; Zhou, J.; Peng, D.; Xue, Y. dbPSP 2.0, an updated database of protein phosphorylation sites in prokaryotes. *Sci. Data* **2020**, *7*, 164. [CrossRef] [PubMed]
43. Nguyen, T.D.; Vidal-Cortes, O.; Gallardo, O.; Abian, J.; Carrascal, M. LymPHOS 2.0: An update of a phosphosite database of primary human T cells. *Database* **2015**, *2015*, bav115. [CrossRef]
44. Yao, Q.; Bollinger, C.; Gao, J.; Xu, D.; Thelen, J.J. P(3)DB: An Integrated Database for Plant Protein Phosphorylation. *Front. Plant Sci.* **2012**, *3*, 206. [CrossRef] [PubMed]
45. Gnad, F.; Ren, S.; Cox, J.; Olsen, J.V.; Macek, B.; Oroshi, M.; Mann, M. PHOSIDA (phosphorylation site database): Management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.* **2007**, *8*, R250. [CrossRef]
46. Keshava Prasad, T.S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; et al. Human Protein Reference Database–2009 update. *Nucleic Acids Res.* **2009**, *37*, D767–D772. [CrossRef] [PubMed]
47. Xiang, Y.; Zou, Q.; Zhao, L. VPTMdb: A viral posttranslational modification database. *Brief. Bioinform.* **2021**, *22*, bbaa251. [CrossRef] [PubMed]
48. Qi, L.; Liu, Z.; Wang, J.; Cui, Y.; Guo, Y.; Zhou, T.; Zhou, Z.; Guo, X.; Xue, Y.; Sha, J. Systematic analysis of the phosphoproteome and kinase-substrate networks in the mouse testis. *Mol. Cell. Proteom.* **2014**, *13*, 3626–3638. [CrossRef]
49. Bodenmiller, B.; Campbell, D.; Gerrits, B.; Lam, H.; Jovanovic, M.; Picotti, P.; Schlapbach, R.; Aebersold, R. PhosphoPep—A database of protein phosphorylation sites in model organisms. *Nat. Biotechnol.* **2008**, *26*, 1339–1340. [CrossRef]

50. Yang, C.-Y.; Chang, C.-H.; Yu, Y.-L.; Lin, T.-C.E.; Lee, S.-A.; Yen, C.-C.; Yang, J.-M.; Lai, J.-M.; Hong, Y.-R.; Tseng, T.-L.; et al. PhosphoPOINT: A comprehensive human kinase interactome and phospho-protein database. *Bioinformatics* **2008**, *24*, i14–i20. [CrossRef]
51. 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]
52. Obenauer, J.C.; Cantley, L.C.; Yaffe, M.B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.* **2003**, *31*, 3635–3641. [CrossRef] [PubMed]
53. Safaei, J.; Maňuch, J.; Gupta, A.; Stacho, L.; Pelech, S. Prediction of 492 human protein kinase substrate specificities. *Proteome Sci.* **2011**, *9* (Suppl. S1), S6. [CrossRef] [PubMed]
54. Kim, J.H.; Lee, J.; Oh, B.; Kimm, K.; Koh, I. Prediction of phosphorylation sites using SVMs. *Bioinformatics* **2004**, *20*, 3179–3184. [CrossRef]
55. Linding, R.; Jensen, L.J.; Pasculescu, A.; Olhovsky, M.; Colwill, K.; Bork, P.; Yaffe, M.B.; Pawson, T. NetworKIN: A resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.* **2008**, *36*, D695–D699. [CrossRef] [PubMed]
56. Zhao, M.; Zhang, Z.; Mai, G.; Luo, Y.; Zhou, F. jEcho: An Evolved weight vector to CHaracterize the protein's posttranslational modification mOtifs. *Interdiscip. Sci.* **2015**, *7*, 194–199. [CrossRef]
57. Li, T.; Li, F.; Zhang, X. Prediction of kinase-specific phosphorylation sites with sequence features by a log-odds ratio approach. *Proteins* **2008**, *70*, 404–414. [CrossRef]
58. Gao, Y.; Hao, W.; Gu, J.; Liu, D.; Fan, C.; Chen, Z.; Deng, L. PredPhos: An ensemble framework for structure-based prediction of phosphorylation sites. *J. Biol. Res. (Thessalon)* **2016**, *23*, 12. [CrossRef]
59. Ingrell, C.R.; Miller, M.L.; Jensen, O.N.; Blom, N. NetPhosYeast: Prediction of protein phosphorylation sites in yeast. *Bioinformatics* **2007**, *23*, 895–897. [CrossRef]
60. Xue, Y.; Ren, J.; Gao, X.; Jin, C.; Wen, L.; Yao, X. GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. *Mol. Cell. Proteom.* **2008**, *7*, 1598–1608. [CrossRef]
61. Wang, C.; Xu, H.; Lin, S.; Deng, W.; Zhou, J.; Zhang, Y.; Shi, Y.; Peng, D.; Xue, Y. GPS 5.0: An Update on the Prediction of Kinase-specific Phosphorylation Sites in Proteins. *Genom. Proteom. Bioinform.* **2020**, *18*, 72–80. [CrossRef]
62. Song, C.; Ye, M.; Liu, Z.; Cheng, H.; Jiang, X.; Han, G.; Songyang, Z.; Tan, Y.; Wang, H.; Ren, J.; et al. Systematic analysis of protein phosphorylation networks from phosphoproteomic data. *Mol. Cell. Proteom.* **2012**, *11*, 1070–1083. [CrossRef] [PubMed]
63. Biswas, A.K.; Noman, N.; Sikder, A.R. Machine learning approach to predict protein phosphorylation sites by incorporating evolutionary information. *BMC Bioinform.* **2010**, *11*, 273. [CrossRef] [PubMed]
64. Durek, P.; Schudoma, C.; Weckwerth, W.; Selbig, J.; Walther, D. Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinform.* **2009**, *10*, 117. [CrossRef] [PubMed]
65. Trost, B.; Maleki, F.; Kusalik, A.; Napper, S. DAPPLE 2: A Tool for the Homology-Based Prediction of Post-Translational Modification Sites. *J. Proteome Res.* **2016**, *15*, 2760–2767. [CrossRef] [PubMed]
66. Kirchoff, K.E.; Gomez, S.M. EMBER: Multi-label prediction of kinase-substrate phosphorylation events through deep learning. *Bioinformatics* **2022**, *38*, 2119–2126. [CrossRef]
67. Horn, H.; Schoof, E.M.; Kim, J.; Robin, X.; Miller, M.L.; Diella, F.; Palma, A.; Cesareni, G.; Jensen, L.J.; Linding, R. KinomeXplorer: An integrated platform for kinome biology studies. *Nat. Methods* **2014**, *11*, 603–604. [CrossRef]
68. Xu, Y.; Wilson, C.; Leier, A.; Marquez-Lago, T.T.; Whisstock, J.; Song, J. PhosTransfer: A Deep Transfer Learning Framework for Kinase-Specific Phosphorylation Site Prediction in Hierarchy. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, 11–14 May 2020, Proceedings, Part II*; Lauw, H.W., Wong, R.C.-W., Ntoulas, A., Lim, E.-P., Ng, S.-K., Pan, S.J., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2020; Volume 12085, pp. 384–395. ISBN 978-3-030-47435-5.
69. Wang, D.; Zeng, S.; Xu, C.; Qiu, W.; Liang, Y.; Joshi, T.; Xu, D. MusiteDeep: A deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* **2017**, *33*, 3909–3916. [CrossRef]
70. Chen, Z.; Zhao, P.; Li, F.; Leier, A.; Marquez-Lago, T.T.; Webb, G.I.; Baggag, A.; Bensmail, H.; Song, J. PROSPECT: A web server for predicting protein histidine phosphorylation sites. *J. Bioinform. Comput. Biol.* **2020**, *18*, 2050018. [CrossRef]
71. Deznabi, I.; Arabaci, B.; Koyutürk, M.; Tastan, O. DeepKinZero: Zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases. *Bioinformatics* **2020**, *36*, 3652–3661. [CrossRef]
72. Ahmed, S.; Kabir, M.; Arif, M.; Khan, Z.U.; Yu, D.-J. DeepPPSite: A deep learning-based model for analysis and prediction of phosphorylation sites using efficient sequence information. *Anal. Biochem.* **2021**, *612*, 113955. [CrossRef]
73. Lv, H.; Dao, F.-Y.; Zulfiqar, H.; Lin, H. DeepIPs: Comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Brief. Bioinform.* **2021**, *22*, bbab244. [CrossRef]
74. Lin, S.; Song, Q.; Tao, H.; Wang, W.; Wan, W.; Huang, J.; Xu, C.; Chebii, V.; Kitony, J.; Que, S.; et al. Rice_Phospho 1.0: A new rice-specific SVM predictor for protein phosphorylation sites. *Sci. Rep.* **2015**, *5*, 11940. [CrossRef] [PubMed]
75. Sharifpoor, S.; Nguyen Ba, A.N.; Youn, J.-Y.; van Dyk, D.; Friesen, H.; Douglas, A.C.; Kurat, C.F.; Chong, Y.T.; Founk, K.; Moses, A.M.; et al. A quantitative literature-curated gold standard for kinase-substrate pairs. *Genome Biol.* **2011**, *12*, R39. [CrossRef] [PubMed]

76. Plewczynski, D.; Tkacz, A.; Wyrwicz, L.S.; Rychlewski, L.; Ginalski, K. AutoMotif Server for prediction of phosphorylation sites in proteins using support vector machine: 2007 update. *J. Mol. Model.* **2008**, *14*, 69–76. [CrossRef] [PubMed]
77. Yang, H.; Wang, M.; Liu, X.; Zhao, X.-M.; Li, A. PhosIDN: An integrated deep neural network for improving protein phosphorylation site prediction by combining sequence and protein-protein interaction information. *Bioinformatics* **2021**, *37*, 4668–4676. [CrossRef] [PubMed]
78. Thapa, N.; Chaudhari, M.; Iannetta, A.A.; White, C.; Roy, K.; Newman, R.H.; Hicks, L.M.; Kc, D.B. A deep learning based approach for prediction of Chlamydomonas reinhardtii phosphorylation sites. *Sci. Rep.* **2021**, *11*, 12550. [CrossRef]
79. Guo, L.; Wang, Y.; Xu, X.; Cheng, K.-K.; Long, Y.; Xu, J.; Li, S.; Dong, J. DeepPSP: A Global-Local Information-Based Deep Neural Network for the Prediction of Protein Phosphorylation Sites. *J. Proteome Res.* **2021**, *20*, 346–356. [CrossRef]
80. Saunders, N.F.W.; Brinkworth, R.I.; Huber, T.; Kemp, B.E.; Kobe, B. Predikin and PredikinDB: A computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinform.* **2008**, *9*, 245. [CrossRef]
81. Wong, Y.-H.; Lee, T.-Y.; Liang, H.-K.; Huang, C.-M.; Wang, T.-Y.; Yang, Y.-H.; Chu, C.-H.; Huang, H.-D.; Ko, M.-T.; Hwang, J.-K. KinasePhos 2.0: A web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.* **2007**, *35*, W588–W594. [CrossRef]
82. Ma, R.; Li, S.; Li, W.; Yao, L.; Huang, H.-D.; Lee, T.-Y. KinasePhos 3.0: Redesign and expansion of the prediction on kinase-specific phosphorylation sites. *Genom. Proteom. Bioinform.* 2022; *in press*. [CrossRef]
83. Iakoucheva, L.M.; Radivojac, P.; Brown, C.J.; O'Connor, T.R.; Sikes, J.G.; Obradovic, Z.; Dunker, A.K. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **2004**, *32*, 1037–1049. [CrossRef]
84. Neuberger, G.; Schneider, G.; Eisenhaber, F. pkaPS: Prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. *Biol. Direct* **2007**, *2*, 1. [CrossRef] [PubMed]
85. Li, F.; Li, C.; Marquez-Lago, T.T.; Leier, A.; Akutsu, T.; Purcell, A.W.; Ian Smith, A.; Lithgow, T.; Daly, R.J.; Song, J.; et al. Quokka: A comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics* **2018**, *34*, 4223–4231. [CrossRef] [PubMed]
86. Lee, T.-Y.; Huang, H.-D.; Hung, J.-H.; Huang, H.-Y.; Yang, Y.-S.; Wang, T.-H. dbPTM: An information repository of protein post-translational modification. *Nucleic Acids Res.* **2006**, *34*, D622–D627. [CrossRef]
87. Li, W.; Godzik, A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **2006**, *22*, 1658–1659. [CrossRef]
88. Kawashima, S.; Ogata, H.; Kanehisa, M. Aaindex: Amino acid index database. *Nucleic Acids Res.* **1999**, *27*, 368–369. [CrossRef]
89. Li, T.; Du, P.; Xu, N. Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. *PLoS ONE* **2010**, *5*, e15411. [CrossRef] [PubMed]
90. Lins, L.; Thomas, A.; Brasseur, R. Analysis of accessible surface of residues in proteins. *Protein Sci.* **2003**, *12*, 1406–1417. [CrossRef] [PubMed]
91. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef] [PubMed]
92. Jamal, S.; Ali, W.; Nagpal, P.; Grover, A.; Grover, S. Predicting phosphorylation sites using machine learning by integrating the sequence, structure, and functional information of proteins. *J. Transl. Med.* **2021**, *19*, 218. [CrossRef]
93. Scheidt, T.; Alka, O.; Gonczarowska-Jorge, H.; Gruber, W.; Rathje, F.; Dell'Aica, M.; Rurik, M.; Kohlbacher, O.; Zahedi, R.P.; Aberger, F.; et al. Phosphoproteomics of short-term hedgehog signaling in human medulloblastoma cells. *Cell Commun. Signal.* **2020**, *18*, 99. [CrossRef]
94. Rubbi, L.; Titz, B.; Brown, L.; Galvan, E.; Komisopoulou, E.; Chen, S.S.; Low, T.; Tahmasian, M.; Skaggs, B.; Müschen, M.; et al. Global phosphoproteomics reveals crosstalk between Bcr-Abl and negative feedback mechanisms controlling Src signaling. *Sci. Signal.* **2011**, *4*, ra18. [CrossRef] [PubMed]
95. Li, H.; Guan, Y. Machine learning empowers phosphoproteome prediction in cancers. *Bioinformatics* **2020**, *36*, 859–864. [CrossRef] [PubMed]
96. Zhang, H.; Liu, T.; Zhang, Z.; Payne, S.H.; Zhang, B.; McDermott, J.E.; Zhou, J.-Y.; Petyuk, V.A.; Chen, L.; Ray, D.; et al. CPTAC Investigators Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **2016**, *166*, 755–765. [CrossRef] [PubMed]
97. Khorsandi, S.E.; Dokal, A.D.; Rajeeve, V.; Britton, D.J.; Illingworth, M.S.; Heaton, N.; Cutillas, P.R. Computational Analysis of Cholangiocarcinoma Phosphoproteomes Identifies Patient-Specific Drug Targets. *Cancer Res.* **2021**, *81*, 5765–5776. [CrossRef] [PubMed]
98. Coker, E.A.; Stewart, A.; Ozer, B.; Minchom, A.; Pickard, L.; Ruddle, R.; Carreira, S.; Popat, S.; O'Brien, M.; Raynaud, F.; et al. Individualized Prediction of Drug Response and Rational Combination Therapy in NSCLC Using Artificial Intelligence-Enabled Studies of Acute Phosphoproteomic Changes. *Mol. Cancer Ther.* **2022**, *21*, 1020–1029. [CrossRef]
99. Park, Y.; Kim, M.J.; Choi, Y.; Kim, N.H.; Kim, L.; Hong, S.P.D.; Cho, H.-G.; Yu, E.; Chae, Y.K. Role of mass spectrometry-based serum proteomics signatures in predicting clinical outcomes and toxicity in patients with cancer treated with immunotherapy. *J. Immunother. Cancer* **2022**, *10*, e003566. [CrossRef]

100. Ramos, E.K.; Tsai, C.-F.; Jia, Y.; Cao, Y.; Manu, M.; Taftaf, R.; Hoffmann, A.D.; El-Shennawy, L.; Gritsenko, M.A.; Adorno-Cruz, V.; et al. Machine learning-assisted elucidation of CD81-CD44 interactions in promoting cancer stemness and extracellular vesicle integrity. *eLife* **2022**, *11*, e82669. [CrossRef]
101. Rodrigues-Ferreira, S.; Nahmias, C. Predictive biomarkers for personalized medicine in breast cancer. *Cancer Lett.* **2022**, *545*, 215828. [CrossRef]
102. Shen, J.; Qi, L.; Zou, Z.; Du, J.; Kong, W.; Zhao, L.; Wei, J.; Lin, L.; Ren, M.; Liu, B. Identification of a novel gene signature for the prediction of recurrence in HCC patients by machine learning of genome-wide databases. *Sci. Rep.* **2020**, *10*, 4435. [CrossRef]
103. Azuaje, F.; Kim, S.-Y.; Perez Hernandez, D.; Dittmar, G. Connecting Histopathology Imaging and Proteomics in Kidney Cancer through Machine Learning. *J. Clin. Med.* **2019**, *8*, 1535. [CrossRef]
104. Li, H.; Siddiqui, O.; Zhang, H.; Guan, Y. Joint learning improves protein abundance prediction in cancers. *BMC Biol.* **2019**, *17*, 107. [CrossRef] [PubMed]
105. Gerdes, H.; Casado, P.; Dokal, A.; Hijazi, M.; Akhtar, N.; Osuntola, R.; Rajeeve, V.; Fitzgibbon, J.; Travers, J.; Britton, D.; et al. Drug ranking using machine learning systematically predicts the efficacy of anti-cancer drugs. *Nat. Commun.* **2021**, *12*, 1850. [CrossRef] [PubMed]
106. Crowl, S.; Jordan, B.T.; Ahmed, H.; Ma, C.X.; Naegle, K.M. KSTAR: An algorithm to predict patient-specific kinase activities from phosphoproteomic data. *Nat. Commun.* **2022**, *13*, 4283. [CrossRef] [PubMed]
107. Pan, Y.; Wang, S.; Zhang, Q.; Lu, Q.; Su, D.; Zuo, Y.; Yang, L. Analysis and prediction of animal toxins by various Chou's pseudo components and reduced amino acid compositions. *J. Theor. Biol.* **2019**, *462*, 221–229. [CrossRef]
108. Zuo, Y.; Li, Y.; Chen, Y.; Li, G.; Yan, Z.; Yang, L. PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* **2017**, *33*, 122–124. [CrossRef]
109. Meng, L.; Chan, W.-S.; Huang, L.; Liu, L.; Chen, X.; Zhang, W.; Wang, F.; Cheng, K.; Sun, H.; Wong, K.-C. Mini-review: Recent advances in post-translational modification site prediction based on deep learning. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 3522–3532. [CrossRef]
110. Schoof, E.M.; Furtwängler, B.; Üresin, N.; Rapin, N.; Savickas, S.; Gentil, C.; Lechman, E.; Keller, U.A.d.; Dick, J.E.; Porse, B.T. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. *Nat. Commun.* **2021**, *12*, 3341. [CrossRef]
111. Lun, X.-K.; Bodenmiller, B. Profiling Cell Signaling Networks at Single-cell Resolution. *Mol. Cell. Proteom.* **2020**, *19*, 744–756. [CrossRef]
112. Wei, W.; Shin, Y.S.; Xue, M.; Matsutani, T.; Masui, K.; Yang, H.; Ikegami, S.; Gu, Y.; Herrmann, K.; Johnson, D.; et al. Single-Cell Phosphoproteomics Resolves Adaptive Signaling Dynamics and Informs Targeted Combination Therapy in Glioblastoma. *Cancer Cell* **2016**, *29*, 563–573. [CrossRef]
113. Pérez-Mejías, G.; Velázquez-Cruz, A.; Guerra-Castellano, A.; Baños-Jaime, B.; Díaz-Quintana, A.; González-Arzola, K.; Ángel De la Rosa, M.; Díaz-Moreno, I. Exploring protein phosphorylation by combining computational approaches and biochemical methods. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1852–1863. [CrossRef]
114. Smith, L.M.; Kelleher, N.L. Consortium for Top Down Proteomics Proteoform: A single term describing protein complexity. *Nat. Methods* **2013**, *10*, 186–187. [CrossRef] [PubMed]
115. Chen, J.; Shiyanov, P.; Green, K.B. Top-down mass spectrometry of intact phosphorylated β-casein: Correlation between the precursor charge state and internal fragments. *J. Mass Spectrom.* **2019**, *54*, 527–539. [CrossRef] [PubMed]
116. Gafken, P.R.; Lampe, P.D. Methodologies for characterizing phosphoproteins by mass spectrometry. *Cell Commun. Adhes.* **2006**, *13*, 249–262. [CrossRef] [PubMed]
117. Zabrouskov, V.; Ge, Y.; Schwartz, J.; Walker, J.W. Unraveling molecular complexity of phosphorylated human cardiac troponin I by top down electron capture dissociation/electron transfer dissociation mass spectrometry. *Mol. Cell. Proteom.* **2008**, *7*, 1838–1849. [CrossRef] [PubMed]
118. McIlwain, S.J.; Wu, Z.; Wetzel, M.; Belongia, D.; Jin, Y.; Wenger, K.; Ong, I.M.; Ge, Y. Enhancing Top-Down Proteomics Data Analysis by Combining Deconvolution Results through a Machine Learning Strategy. *J. Am. Soc. Mass Spectrom.* **2020**, *31*, 1104–1113. [CrossRef] [PubMed]
119. Liu, S.; Cui, C.; Chen, H.; Liu, T. Ensemble learning-based feature selection for phosphorylation site detection. *Front. Genet.* **2022**, *13*, 984068. [CrossRef] [PubMed]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.