molecules

MDPI

Review

Intelligent Protein Design and Molecular Characterization Techniques: A Comprehensive Review

Jingjing Wang, Chang Chen, Ge Yao, Junjie Ding *, Liangliang Wang * and Hui Jiang *

State Key Laboratory of NBC Protection for Civilian, Beijing 102205, China; wjj18811039053@163.com (J.W.); chenchang15@mails.ucas.ac.cn (C.C.); bzyaoge@163.com (G.Y.)

* Correspondence: djj224@163.com (J.D.); wangliangliang0304@163.com (L.W.); ylplkmc@163.com (H.J.)

Abstract: In recent years, the widespread application of artificial intelligence algorithms in protein structure, function prediction, and de novo protein design has significantly accelerated the process of intelligent protein design and led to many noteworthy achievements. This advancement in protein intelligent design holds great potential to accelerate the development of new drugs, enhance the efficiency of biocatalysts, and even create entirely new biomaterials. Protein characterization is the key to the performance of intelligent protein design. However, there is no consensus on the most suitable characterization method for intelligent protein design tasks. This review describes the methods, characteristics, and representative applications of traditional descriptors, sequence-based and structure-based protein characterization. It discusses their advantages, disadvantages, and scope of application. It is hoped that this could help researchers to better understand the limitations and application scenarios of these methods, and provide valuable references for choosing appropriate protein characterization techniques for related research in the field, so as to better carry out protein research.

Keywords: intelligent protein design; protein characterization techniques; sequence characterization; structural characterization

Citation: Wang, J.; Chen, C.; Yao, G.; Ding, J.; Wang, L.; Jiang, H. Intelligent Protein Design and Molecular Characterization Techniques: A Comprehensive Review. *Molecules* **2023**, *28*, 7865. https://doi.org/10.3390/molecules28237865

Academic Editor: Benevides C. Pessela

Received: 21 October 2023
Revised: 13 November 2023
Accepted: 23 November 2023
Published: 30 November 2023

1. Intelligent Design for Protein Molecules

Protein molecular design refers to the comprehensive use of multidisciplinary techniques to obtain novel proteins with better target performance than natural proteins based on the structure–function relationship of proteins. This process mainly involves establishing a structural model of the target protein, studying the structure–function relationship, proposing a reasonable design and renovation plan, and further modifying the design through experimental verification, which often requires multiple iterations to achieve the desired purpose (Figure 1) [1]. The main types of protein structural designs include: (1) Minor, (2) Moderate, (3) Major modifications, which can be described, in order, as follows: (1) Artificially modifying amino acid (AA) residues of natural proteins with known structures to investigate and improve their function and properties, (2) Splicing and assembling protein structural domains from different sources to obtain protein molecules with new functions through the transfer of the corresponding functions, (3) Designing entirely new proteins with specific spatial structures and functional properties from scratch [1,2].

Early work in protein design focused on redesigning helical bundles [3], using strategies designed to generate specific hydrophobic/hydrophilic patterns to control the protein folding process without predicting specific side chain orientations [4–6]. In 1997, protein structure design methods were gradually extended to irregular geometries to increase the diversity and variability of backbone structures in protein design [7]. For example, RosettaDesign, a universal computing protocol, was used to predict the low-freedom energy sequences of nine natural protein backbones. Comparing the NMR structure of the predicted sequence with that of the natural protein, showed that RosettaDesign could reliably identify the amino acid sequence of the protein backbone [8]. In 2003, the Baker Lab

continuously iterated between sequence design and structure prediction to break the existing topologies of protein redesign, obtain novel protein folding structures, and produce the Top7 α/β topology [9]. However, the early exploratory efforts targeting computational protein design suffered from a small range of structural modifications, low success rates, and ineffective results. They relied on cyclic iterative experimental screening, which resulted in significant consumption of human, material, and time resources.

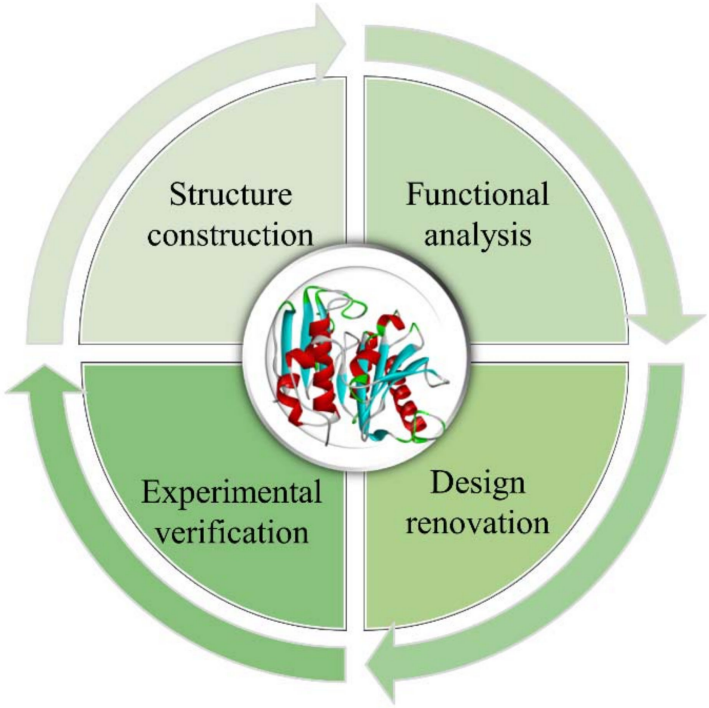

Figure 1. The flowchart for protein design.

In recent years, the updated optimization of artificial intelligence algorithms, increasing arithmetic power of computer hardware, and massive expansion of experimental protein structure data have created favorable infrastructure for intelligent protein design, resulting in many remarkable results (Figure 2). In 2019, Ingraham et al. [10] introduced a protein generation model based on a graphical representation of the three-dimensional (3D) structure of proteins, which improves both computational speed and reliability compared to traditional neural-network-based approaches due to its ability to exploit the spatial localization of dependencies in the molecular structure. In 2020, Strokach et al. [11] developed a deep graph neural network called ProteinSolver that was trained to precisely design sequences that were folded into a predetermined shape. Deep graph neural networks can rapidly design specific novel protein sequences, which are difficult to achieve using traditional computational design methods. In 2021, Anishchenko et al. [12] from the Baker Lab developed a deep neural network hallucination method, using trRosetta, which is trained in protein structure prediction and has the capability to capture protein sequences and structural information as a background network. This method generates new protein sequences with specific functions by "inducing" random sequences from the input. This facilitates an exploration of all possible protein structure spaces that is credited to the ability of deep learning to process large datasets. The following year, Wang et al. [13] from the same group developed a deep learning method based on hallucination and inpainting to enable the construction of protein binding and catalytic functional sites without pre-specifying backbone folding or secondary structure.

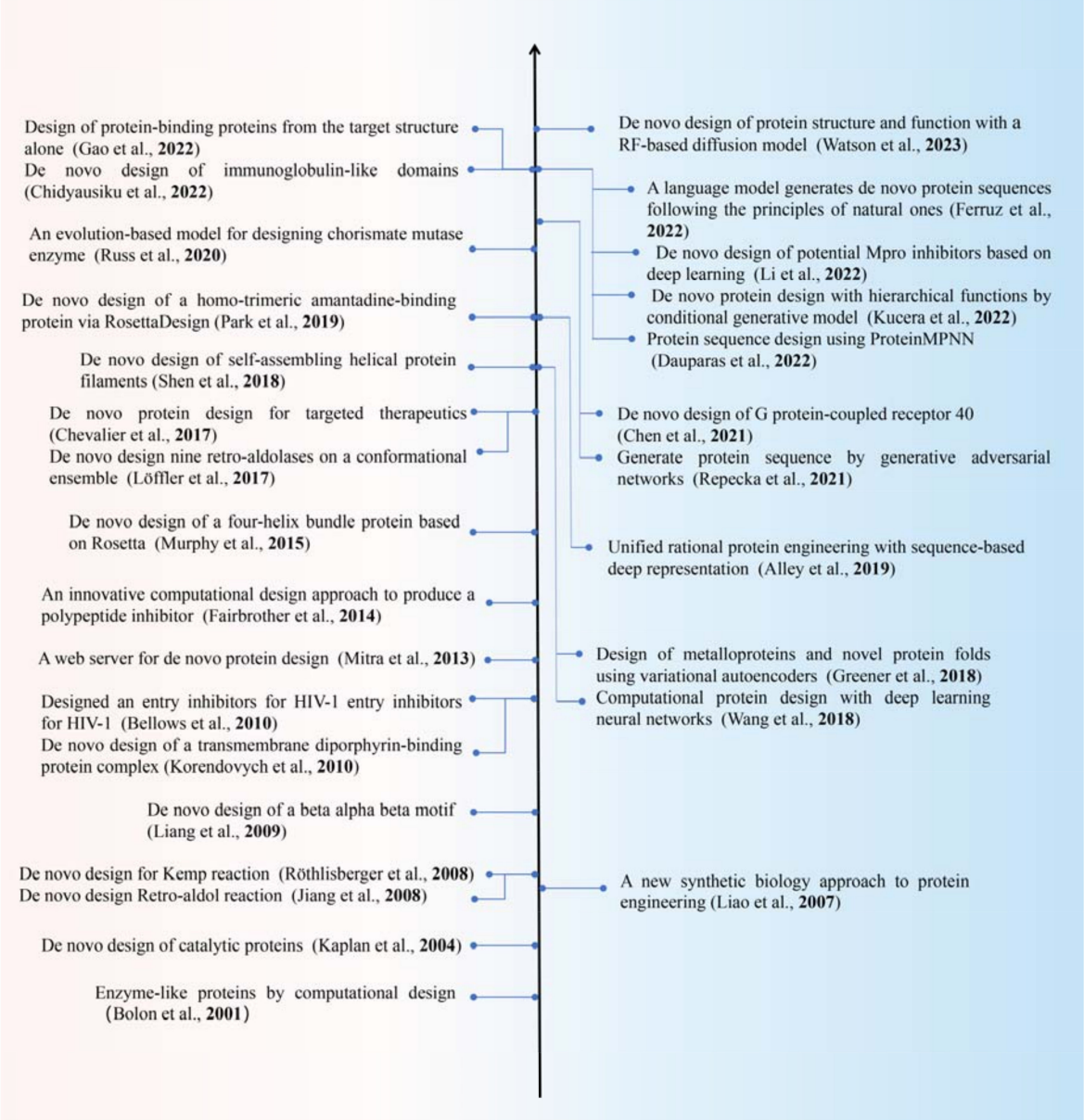

Figure 2. The summary of the classic examples of de novo protein design [14–41]. The left and right of the figure show model-based and data-driven examples, respectively.

The core of intelligent protein design involves establishing a relationship between structure and function. Therefore, the prediction of protein structure and function by artificial intelligence algorithms is also a key aspect of protein design, apart from in the above-mentioned intelligent protein design methods that directly modify protein structure to target the desired performance. AlphaFold2 [42], which has made great progress on the "protein structure prediction" problem that has plagued the academic community for five decades, has predicted structures covering 98.5% of the human proteome [43]; similarly, these data will provide a transformative impact on the intelligent design of proteins with specified functions. In February 2022, Bileschi et al. [44] used a dataset constructed from the Protein Families Database (Pfam) to train a neural network (called ProtCNN) to functionally classify protein sequences to achieve a 200-fold increase in speed, and a 9-fold reduction in error, compared to the traditional BLASTp method. This advancement in functional prediction provides a powerful tool for accelerating the intelligent design of proteins.

It is evident that the latest advances in artificial intelligence algorithms (especially deep learning technology) can boost the overall intelligent protein design process by assisting protein structure modification and structure and function prediction [10–13,42,43]. Structural characterization of protein molecules is a crucial part of the intelligent protein design process. The ability to represent protein structures in a comprehensive, accurate, and efficient manner in a machine-recognizable language or vector is essential for the success of downstream intelligent protein design tasks using intelligent algorithms. This review systematically described various protein characterization techniques and representative applications used in intelligent protein design and discussed their advantages, disadvantages, and application areas. We hope to provide a valuable reference for scholars to conduct relevant research in this field.

2. Examples of Applications for Intelligent Protein Design

Artificial intelligence has been used in many applications in the field of protein engineering; including protein structure, function, thermal stability [45–47], and stereoselectivity prediction [48,49]; owing to its high accuracy, fast computational speed, and independence from protein structure and function information compared with earlier protein design methods. Various deep learning algorithms and natural language processing (NLP) techniques based on deep learning were successfully used in numerous applications, apart from classical machine learning algorithms (support vector machines, decision trees, Gaussian regression, and so on) [50–54]. The following section focuses on three recent successful cases in protein structure prediction, function prediction, and de novo protein design to systematically analyze the advantages of artificial intelligence algorithms applied in protein engineering.

2.1. Protein Structure Prediction

Protein structure prediction is a critical step in the intelligent design of proteins and is a fundamental scientific problem in the field of protein computation. This problem can be traced back to the famous statement made by Christian B. Anfinsen (the Nobel laureate in chemistry in 1972), that the AA sequence of a polypeptide chain contains all the information about its 3D structure [55]. Currently, experimental techniques for obtaining 3D protein structures include X-ray crystallography [56], nuclear magnetic resonance (NMR) [57], and cryo-electron microscopy (cryo-EM) [58]. There are only about 205,000 experimentally resolved protein structures in the Protein Data Bank (PDB) as of June 2023 [59], while the UniProt database contains over 250 million sequences [60]. This means that the number of proteins with known sequences is more than 1200 times greater than the number of experimentally resolved protein structures. In contrast, the number of known protein sequences was only 160 times that of the experimentally resolved protein structures in 2011 [61]. It is evident that the number of protein structures solved is far lower than the total number of protein sequences.

To address this problem, the academic community has been organizing the critical assessment of protein structure prediction (CASP) competitions since 1994, which has greatly promoted the development of computational methods for protein structure prediction. For example, I-TASSER [62] represents a homology modeling approach that uses threading to predict structures and has won multiple championships in the CASP. In 2020, AlphaFold2 [42], developed by DeepMind, won CASP14 by a landslide using the transformer algorithm. In 2022, DeepMind released the AlphaFold protein structure library, AlphaFold DB [63], demonstrating the dominance of the AlphaFold tool for protein structure prediction. In addition, RoseTTAFold [64], developed by Baker Lab, achieved considerable prediction accuracy at CASP14, ranking only behind AlphaFold2. Novel artificial intelligence-driven protein folding prediction tools such as AlphaFold2 and RoseTTAFold provide powerful drivers for rapid and accurate protein structure prediction and subsequent protein design modifications [65–73]. Many studies were conducted using

them to further improve the accuracy and speed of protein structure predictions. Table 1 summarizes the methods, models, and functions of relevant studies.

Table 1. Several tools for protein structure prediction derived from AlphaFold2 and RoseTTAFold.

Methods	Models	Inputs	Multimeric Structure	Advantages	URLs	References
ColabFold	JAX	MSA-based	Yes	40–60 × faster prediction than AlphaFold2, and user friendly	https://github.com/sokrypton/ColabFold, accessed on 24 November 2023	[65]
OpenFold	PyTorch	MSA-based	Yes	PyTorch replication of AlphaFold, high flexibility	https://github.com/aqlaboratory/openfold, accessed on 24 November 2023	N/A
Uni-Fold	PyTorch	MSA-based	Yes	Friendly operating environment, and wide hardware adaptation	https://github.com/dptech-corp/Uni-Fold, accessed on 24 November 2023	[66]
FastFold	PyTorch	MSA-based	No	Reduced training time from 11 days to 67 h	https://github.com/hpcaitech/FastFold, accessed on 24 November 2023	[67]
HelixFold	PaddleHelix	MSA-based	No	Improved training and prediction speed, and reduced memory consumption	https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/protein_folding/helixfold, accessed on 24 November 2023	[68]
MindSpore-Fold	MindSpore	MSA-based	Yes	Based on MindSpore framework, high performance, and fast prediction speed	https://github.com/mindspore-ai/mindspore, accessed on 24 November 2023	N/A
MEGA-Fold	MindSpore	MSA-based	No	More accurate and efficient protein structure prediction than AlphaFold2	https://gitee.com/mindspore/mindscience/tree/master/MindSPONGE/applications/MEGAProtein, accessed on 24 November 2023	[69]
EMBER3D	PyTorch	pLM-based	No	Ability to visualize the effect of mutations on predicted structures and high predictive efficiency	https://github.com/kWeissenow/EMBER3D, accessed on 24 November 2023	N/A
ESM-Fold	PyTorch	pLM-based	No	Reduced dependence on MSA input, inference speed is an order of magnitude faster than AlphaFold2	N/A	[51]
HelixFold-Single	PaddleHelix	pLM-based	No	Breaking the speed bottleneck of relying on MSA retrieval models, and prediction accuracy is comparable to AlphaFold2 and nearly a thousand times faster	https://github.com/PaddlePaddle/PaddleHelix/tree/dev/apps/protein_folding/helixfold-single, accessed on 24 November 2023	[70]
OmegaFold	PyTorch	pLM-based	No	Protein homology-independent, easy to install, and overall predictive power comparable to AlphaFold2 and RoseTTAFold	https://github.com/HeliXonProtein/OmegaFold, accessed on 24 November 2023	[71]
IgFold	PyTorch	pLM-based	No	Focus on antibody structure prediction, high prediction accuracy, and prediction time less than 1 min	https://github.com/Graylab/IgFold, accessed on 24 November 2023	[72]
D-I-TASSER	PyTorch	MSA-based		Higher prediction accuracy with online server	https://zhanggroup.org/D-I-TASSER/, accessed on 24 November 2023	[73]

The advent of AlphaFold2 and RoseTTAFold has increased the accuracy of protein structure prediction to a new level. However, these methods are not effective at predicting the structure of orphan proteins because of the lack of homologous proteins. In October 2022, Chowdhury et al. [54] proposed an end-to-end recurrent geometric network computational model named RGN2 that predicts the structure of orphan proteins with better accuracy than AlphaFold2 and RoseTTAFold. It uses the protein language AminoBERT to parse the potential structural information of orphan proteins, and its computational

efficiency is 106 times faster than that of AlphaFold2. In November of the same year, Wang et al. [52] proposed a single-sequence protein structure prediction algorithm called trRosettaX-Single. The algorithm integrates sequence embeddings from the Transformer protein language model into a knowledge-distillation-enhanced multiscale network to predict two-dimensional geometric structures between residues. Then, the three-dimensional structure is reconstructed using an energy minimization approach, which improves the accuracy and efficiency of orphan protein structure prediction.

In addition to protein monomer structure prediction, multimer structure prediction has also been studied recently. In October 2021, the DeepMind team developed AlphaFold-Multimer [74], with innovative multi-chain feature extraction and symmetric replacement modules based on AlphaFold2. It achieved prediction accuracies of 67% and 69% at the contact interface of heterologous and homologous multimers, respectively. In September 2022, Tang et al. [75] proposed the first MSA pairing algorithm, ColAttn, which combines the outputs of protein language models into a joint MSA form to identify paired homologs from single chains using the attention score in the MSA Transformer, making it demonstrate the best structure prediction accuracy on heterodimers. Meanwhile, Uni-Fold v.2.0.0 [66], released by DP Technology, also added a protein multimer structure prediction function. The tool is modeled on the model AlphaFold-Multimer architecture and modifies and optimizes the model details, achieving a two-fold increase in speed and accuracy. In addition, Zhang Yang's lab proposed the DMFold-Multimer [73], which combines DeepMSA2 for searching homologous sequences from large-scale genomic and metagenomics databases with AlphaFold2-Multimer's structure model generator, leading to the champion of the protein complex structure prediction project in the CASP15. However, these studies were mainly constrained by the limited number of multimer structures used for training and the lack of accurate characterization of multimer clustering relationships. This provides limited prediction accuracy and few structure predictions of protein–ligand complexes. In conclusion, the structure prediction of protein monomers with homologs was basically solved with the advent of AlphaFold2 and Uni-Fold v.2.0.0. The accuracy of the structural predictions for orphan proteins and multimers requires improvement. Moreover, structural prediction research on protein–ligand complexes is sparse, and mainly relies on docking and dynamic simulations to predict protein–ligand binding patterns. The direct prediction of protein–ligand complex structures by artificial intelligence-based methods would receive significant attention from scholars with the increase in experimentally resolved protein–ligand complex structures, the development of protein complex characterization methods, and the further improvement of computer performance.

2.2. Protein Function Prediction

The primary sequence of a protein determines its high-level structure that determines its function according to the golden rule of sequence–structure–function correspondence. Thus, the protein sequence ultimately determines protein function. A deep understanding of the relationship between protein sequence and function enables the rapid localization of novel protein functions that facilitates de novo protein design by direct sequence modification. The advent of low-cost and efficient sequencing technologies has driven rapid growth in the number of protein sequences [76,77]. The UniProtKB database contains over 200 million sequences, with only approximately 0.25% manually annotated by March 2022 [78]. Determining the relationships between sequences and functions has become a critical issue in protein design with the growing number of protein sequences.

In 2020, Hippe et al. [79] proposed the ProLanGO2 method that follows the design principles of natural language translation and uses sequence-based recurrent neural networks for protein function prediction. Its prediction performance is comparable to that of other sequence-based methods, and even the network-based method NetGO2.0. ProLanGO2 has proven its potential for protein function prediction by converting protein function prediction into natural language translation. In 2021, Gligorijević et al. [80] proposed a graphical convolutional network model, DeepFRI, which combines deep learning with more available

sequence information to substantially improve protein function prediction. In the same year, Yong et al. [81] proposed an automated protein function prediction method based on graph neural networks, DeepGraphGO, which significantly outperformed many state-of-the-art methods by fully combining protein sequences and higher-order protein network information. In 2022, Zhang Yang's lab established a unified and efficient multi-domain protein structure and function prediction platform, I-TASSER-MTD, by integrating methods developed by his lab in recent years. This includes protein sequence structural domain delineation, deep learning spatial geometric constraint prediction, single-domain structure modeling, multi-domain structure assembly, and structure-based function annotation to achieve fully automated multi-domain protein structure and function prediction from protein sequences [51].

Various types of information were used for automated prediction of protein function in addition to sequence-based methods for intelligent prediction, such as domain-based prediction [82–84], homologous-protein-based functional transfer [85–87], and protein-network-dependent methods [88–90]. However, there is a lack of functional sites, homologous proteins, or biological network information for newly sequenced or less studied proteins. Therefore, in future protein function prediction, we can focus on three key aspects. Firstly, we need to accurately characterize the relationship between protein structure and function to enhance the overall performance of protein function prediction. Secondly, we can use correlations between different functions to aid in the precise localization of protein functions for multifunctional proteins. Lastly, we can integrate protein structural features, global and local sequence features, and genomic contextual environmental features to achieve accurate protein function prediction.

2.3. De Novo Protein Design

The emergence of de novo protein design can be traced back to the 1980s, when DeGrad et al. [91] made a preliminary attempt at protein design and successfully constructed stable four-stranded helix bundles using rule-based heuristics. In the late 1990s, Dahiyat et al. [7] pioneered the design of AA sequences using an automated optimization approach with the development of molecular mechanics energy functions, AA side-chain conformational libraries, and optimization algorithms.

The automatic design method based on energy functions is not limited by the type of main chain structure compared with the purely heuristic design method. Furthermore, the specific spatial accumulation between residues and the quantitative calculation of hydrogen bond interactions improves the success rate of the design. In the 21st century, Baker first designed protein folding that does not exist in nature, leading to the de novo design of protein backbones. In 2008, Baker proposed an inside-out protein design strategy to artificially create several non-natural enzymes (such as Diels–Alder synthase [92], Kemp eliminase [14], and Aldolase [15]) through theoretical computational design.

In recent years, algorithms emerging from the de novo design of proteins were gradually applied to the structural-functional remodeling of natural proteins. In 2015, the David Lab group re-engineered formaldehyde polymerase (FLS) to catalyze the polymerization of formaldehyde using a specific natural benzaldehyde lyase (BAL) unearthed from a database and employing Foldit and RosettaDesign tools [93]. Further modification of the FLS design increased its activity 4.7-fold in 2021. This makes it a key enzyme in the in vitro pathway converting inorganic carbon to organic carbon in the synthesis of starch from CO_2 [94]. Most of the above studies used energy functions as indicators for protein design evaluation or tools, such as Foldit and RosettaDesign, for de novo protein design, collectively referred to as model-based de novo protein design. Classic de novo protein design examples are presented in Figure 2.

Data-driven approaches to de novo protein design (including structural data and massive protein sequences) have also emerged in recent years [32–41,95] along with the wave of big data and artificial intelligence development, the development of high-throughput data collection methods, and the accumulation of available data. Liu and co-authors made a sig-

nificant contribution to the development of data-driven protein design methods [96]. The authors constructed the SCUBA model for the de novo design of protein backbone structures, using neural network energy functions and the statistical energy model ABACUS. This method is critical for designing AA sequences for a given backbone structure, and it is the only fully experimentally validated method for the de novo design of proteins besides RosettaDesign. However, this approach to sequence design by optimizing the energy function has limited success rates and computational efficiency. The Baker Lab proposed a multi-stranded and symmetry-aware model architecture, ProteinMPNN, which generates sequences that fold more reliably and accurately into the natural protein backbone than the original natural sequences [40]. This tool significantly improves computational efficiency compared to physically based methods, such as Rosetta. It is widely used in protein design, owing to its high design success rate, low time consumption, and applicability to almost all protein sequence designs [97,98]. In addition, Baker Lab also constructed a versatile protein design framework based on an RF-based diffusion model, RFdiffusion, which enables de novo binder design and the design of higher-order symmetric architectures [41].

De novo protein design using computational design has entered an unprecedented era, wherein the structural and functional design of increasingly complex proteins would be possible with the continuous iterative optimization of energy functions, main chain design, and side chain optimization. A recent review by Ovchinnikov and Huang described how structural information can replace traditional backbone design, side-chain optimization, and energy functions [99]. Huang used structural features of AA neighbors to construct "higher-order soft potential energy functions" [100]. Comparing traditional methods against deep learning methods is an important issue in the anticipation of new methods.

3. Macromolecular Characterization Techniques and Their Application in Intelligent Protein Design

Molecular characterization refers to measuring molecular properties in a certain aspect that is either the basic physical and chemical properties of molecules, or numerical indicators or vectors derived from the molecular structure using various algorithms to describe the structural information of different layers of molecules [101]. They can be divided into small and macromolecular characterizations depending on the size of the molecular system. The threshold and difficulty of characterizing biomacromolecules is significantly higher compared with the characterization techniques of small molecules, owing to their higher molecular mass and higher structural complexity. Protein intelligence design involves extracting and encoding the structural features of biological macromolecules, such as DNA, proteins, and RNA, as quantitative vectors. These vectors are then used for machine learning-based modeling tasks, including predicting protein binding regions [102–104], functions [105–107], physical and chemical properties [108–111], and more. This review characterized techniques for protein macromolecules as divided into four categories, according to the degree of description for the structure information: (1) Characterization based on traditional molecular descriptors, (2) Sequence-based characterization, (3) Structure-based characterization, (4) Hybrid sequence–structure-based characterization. The subsequent sections focus on these four aspects of macromolecular characterization techniques and the corresponding application cases, as well as systematically analyzing the characteristics, advantages, and limitations of each characterization method.

3.1. Characterization Based on Traditional Molecular Descriptors

In the early years, computer development was relatively delayed and hardware standards were low. Traditional classical descriptors were widely used for crude characterization of protein macromolecules, owing to their simplicity, ease of understanding, and low arithmetic requirements. These traditional descriptors typically quantitatively describe the intrinsic properties of a macromolecule based on its molecular composition and physicochemical properties, including the frequency of AA occurrences in the protein composition [112], the isoelectric point used to determine the charge of the protein

in different pH solutions [113,114], the hydrophilicity and hydrophobicity (which plays a major role in maintaining protein conformation) [115–117], the absolute charge of the protein [118], the sequence entropy to reflect the conservation and variability of the protein AA sequence [119], the sequence length and molecular weight to reflect the protein length and size [120], the solvent accessible surface area (SASA) [121,122] to indicate the degree of AA exposure of a protein, and the dipole moment [123,124] (used to determine the spatial conformation of a molecule), and so on. Characterization methods can be divided into two categories: sequence-based and structure-based. The characteristics, categories, and applications of each traditional descriptor representation are discussed in detail in Table 2.

Table 2. Summary of characteristics, properties, and applications of traditional descriptor characterization methods.

	Encoding	Description	Characteristic	Main Category	Application
Based on the sequence	k-mer	K-mer is a subsequence of length k that is used to minimize the effects of arbitrary starting points, where k is an integer, ranging from 1 to hundreds.	Reflects the frequency of k-conjoined AAs in the protein sequence.	Based on AA information	[125–127]
	PSSM	Logarithm of the probability of all possible molecular types occurring at each position in a given biological sequence.	Powerful, but neglects the interactions between different residues.	Based on evolutionary information	[128,129]
	BLOSUM	Reflects the exchange probability of AA pairs.	Research results vary with the type of matrix.	Based on evolutionary information	[130]
	Autocorrelation	The interdependence of AAs in a given sequence.	Reduces the feature space and standardize the sequence length.	Based on physicochemical properties	[131]
	CTD	The composition, transition, and distribution (CTD) of AAs in a given sequence.	Reflects the distribution of AAs with diverse structures and physicochemical characters in a given sequence.	Based on physicochemical properties	[132–134]
	CTriad	The conjoint triad (CTriad) is generally regarded to consist of a combination of three adjacent AAs.	AAs were divided into 7 groups based on the side chain volume and dipole.	Based on physicochemical properties	[135,136]
	Z-scales	The Z-scales obtained from the field of quantitative sequence-activity modeling (QSAM).	The most widely used descriptor set in proteochemometric modeling,	Based on physicochemical properties	[137]
	VHSE	Vectors of hydrophobic, steric, and electronic properties (VHSE) are derived from principal components analysis (PCA) of independent families of 18 hydrophobic properties, 17 steric properties, and 15 electronic properties, respectively.	VHSE is of relatively definite physicochemical meaning, easy interpretation, and contains more information compared with z scales.	Based on physicochemical properties	[138,139]
	ProtFP	Protein Fingerprint (ProtFP) is based on a selection of different AA properties obtained from the AAindex database.	The descriptor was obtained using recursive elimination of the most co-varying properties after starting with the full set of indices.	Based on physicochemical properties	[139,140]
	FASGAI	The factor analysis scales of generalized AA information (FASGAI) are derived from 335 physicochemical properties of the 20 natural AAs.	Applying a factor analysis rather than a PCA.	Based on physicochemical properties	[139,141]

Table 2. *Cont.*

	Encoding	Description	Characteristic	Main Category	Application
Based on the structure	T-scale	Derived from PCA on the 67 kinds of structural and topological variables of 135 AAs.	The 3D properties of each structure are not explicitly considered.	Topology-based representation method	[142]
	ST-scale	Structural topology scale (ST-scale) was recruited as a novel structural topological descriptor derived from PCA on 827 structural variables of 167 AAs.	The molecular structure was optimized, and 3D information of AAs was used.	Topology-based representation method	[143,144]
	MSWHIM	The MSWHIM descriptor set is derived from 36 electrostatic potential properties obtained from the 3D molecule structure.	The number of indicators is simple, easy to calculate, and invariant to the coordinate system.	Geometric-based representation method	[145]

Early traditional descriptors in intelligent protein design were mostly used in studies of protein–macromolecule interactions, protein–small molecule interactions, and protein functional site predictions [125,126,146–149]. Liu et al. [128] proposed a model called aPRBind to predict the binding residues of RNA in proteins by convolutional neural networks, that integrates the sequence features based on the spatial neighbor-based position-specific score matrix (SNB-PSSM) and structural features (including residue-kinetic properties and residue-nucleotide propensities), based on the I-TASSER model, to achieve superior predictive performance compared to other advanced methods. However, the best sensitivity, specificity, accuracy, and Mathew's correlation coefficient were 0.65, 0.82, 0.74, and 0.48, respectively, indicating that there is still room for improvement in protein–RNA binding site prediction. Traditional descriptors do not provide a comprehensive characterization of the global information of an RNA/protein. Therefore, the accuracy of more complex prediction tasks (such as functional sites) requires improvement. Consequently, these methods are inappropriate for more complex protein design.

3.2. Sequence-Based Characterization

Protein sequence determines the three-dimensional structure. Therefore, the protein sequence contains advanced structural information. Most protein-related studies have employed sequence information to characterize the proteins when protein structures were difficult to resolve and computing power was insufficient. One-hot and K-mer characterization methods were used extensively, owing to their simplicity and ease of understanding, low computational effort, and high efficiency [150–154]. In addition, protein sequence characterization methods such as word2vec [155,156], seq2vec [157], BioVec [158], doc2vec [159], and N-gram [160,161] were proposed and applied based on the intrinsic similarities between protein sequences and natural languages.

A large variety of NLP models have emerged with a profound impact on the study of intelligent protein design following advances in sequencing technology, the development of deep learning algorithms, and significant improvements in computing power. In 2017, Google released transformers based on the attention model that started a new era of NLP [162]. This greatly improved the performance of various tasks, including clinical diagnosis, image recognition, and protein–ligand affinity prediction [163–168]. Countless adaptations of pre-trained language models have emerged, including the Bidirectional Encoder Representation from Transformers (BERT) based on the transformer encoder structure [169], the Generative Pre-trained Transformer (GPT) and the successors GPT-2 and GPT-3 [170–172], the Evolutionary Scale Modeling (ESM) family for predicting protein structure and function (ESM-1b, ESM-MSA-1b, & ESM-1v) [173–175], the ProtTrans with the largest training dataset [176], and the ProGen language model [177] that can control protein generation.

Advances in the transformer era inspired several studies to apply the concept of language models to protein design. In February 2019, Yu et al. [161] applied n-gram modeling to generate a probabilistic protein language model. In October 2019, Alley et al. [34] applied a multiplicative long short-term memory network (mLSTM) to learn a language model that predicted protein sequence stability with higher accuracy. In July 2022, Höcker et al. [37] proposed a language model trained on protein space ProtGPT2 to generate new protein sequences according to natural principles. In December 2022, Rives et al. [178,179] found that the ESM2 language model can generate new proteins beyond natural proteins and generate complex and modular protein structures by learning and programming deep grammar.

NLP-based learning models for protein sequence representation have achieved remarkable results in protein design [34,37,161,178,179]. However, there are deep grammatical structural differences between modeling languages and protein representations. It is estimated that a native American English speaker uses approximately 46,200 words on average and multi-word expressions. However, only 20 different AAs are processed in proteins by representation models in a manner similar to a linguistic lemma. Moreover, these language models have relatively high spatial and temporal complexities. For example, the ESM-2 model with 15 billion parameters, requires significant computational time and powerful computing equipment for training.

It is anticipated that NLP models will be further improved by simplification and reducing their dependence on computing devices. New protein characterization methods will be developed to better represent the relationship between protein and natural language. Alternatively, we may see the continued growth of protein sequences and the implementation of quantum computers that will allow protein design models to achieve human-like thinking and precisely achieve the second law proposed by Manfred Reetz: "You get what you designed" [180].

3.3. Structure-Based Characterization

Proteins are composed of one or more peptide chains, and the connections and folding patterns of each peptide chain constitute their special three-dimensional spatial structure [181]. The unique spatial structure determines the specific biological functions. In theory, obtaining the structural information of proteins could lead to a better understanding of the relationship between the structure and function of proteins, which could lead to a better intelligent protein design. Therefore, protein intelligent design and functional studies require structural characterization. Structural characterizations can be divided into graph structure-based and geometric structure-based characterizations according to the manner in which they are performed. Graph structure-based characterization methods can be divided into topology and distance-graph-based protein characterization methods.

3.3.1. Graph Structure-Based Characterization

Topology Structure-Based Protein Characterization

Topology-based protein characterizations describe AAs based on the atomic linkage indices generated from molecular graphs. These mainly include traditional T-scale and ST-scale topology descriptors, and newer meta-graph and circuit topology descriptors.

In 2007, T-scale was proposed by Tian et al. [107] based on a computer program. generating 67 generic topological descriptors based on 135 AAs. However, these descriptors do not explicitly consider the 3D features of each structure, and they are based only on the strength of the AA linkage table. In 2009, ST-scale proposed by Yang et al. [182] used the 3D information of 167 AAs and PCA based on 827 structural dimensions. The chemical structure of a set of peptides and their analogs can be characterized by describing the position of each AA using eight ST-scale values based on ST-scale.

A meta-graph is a newly proposed graph structure that differs from the traditional network themes or sub-graphs. It captures specific topological arrangements involving interactions and associations between proteins and keywords. Each protein can be described by a series of meta-graphs illustrating its interactions with other proteins and

their associations with keywords. Proteins with similar functions often exhibit similar meta-graph representations [183].

Circuit topology is a newly proposed descriptor that theoretically assesses the relationship between contact pairs on the protein backbone and provides information about the protein structure (such as the order of residues and residue contacts). The use of circuit topology to predict the folding rate of proteins has improved pathogenicity prediction of missense mutations [184].

Distance Map-Based Protein Characterization

Protein distance graphs can be obtained by calculating the distance between Cα atoms or neighboring residues. A protein of length n can be represented as a $n \times n$ matrix, and descriptor values can then be obtained using matrix decomposition or image processing techniques. The contact graph is a binary graph obtained by setting a distance threshold on the distance graph. Distance and contact graphs have the advantage of rotational or translational invariance of the protein structure and low dimensionality, which makes them computationally efficient. Currently, contact, and distance graphs were extensively used in protein structure prediction methods, such as AlphaFold [185], trRosetta [186], C-I-TASSER [187], C-QUARK [188], DeepFold [189], and so on.

3.3.2. Geometry-Based Characterization

Geometry-based protein characterization is related to indicators representing the structural features of a protein, such as the locations of atoms in space, and the shape and size of the protein. These include point clouds [50], three-dimensional tessellation [190], three-dimensional convolutional neural network (3D-CNN) [191], and GVP-GNN [192].

A point cloud is a set of points representing object-space partitioning and external attributes in the same spatial reference system. It is a group of isolated nodes with a given position in 3D space, called a 3D point cloud [50]. Point clouds are significantly faster than other procedures in terms of data processing. They can be directly processed by rotation and other variable operations, thereby avoiding extension of the data. Currently, point clouds are used in areas involving protein–ligand binding affinity prediction and protein–ligand binding site prediction [50,193].

Three-dimensional tessellations allow graphical representation of proteins by dividing the three-dimensional space into cells with specific properties. Each node represents a cell and any contact between two cells is represented by each edge. A Voronoi diagram is a typical type of tessellation that describes the structure and interactions of proteins and is mostly applied in structural bioinformatics [190]. For example, it is used to estimate the deviation between the predicted and native protein structures [194], and to analyze the structure of protein–protein interactions [195]. An effective programming representation of Voronoi graphs requires quite a complex data structure. The high cost of developing and maintaining these data structures is a notable barrier to fully utilizing this powerful mathematical concept in practice.

The 3DCNN divides 3D space into multiple grids, allowing direct manipulation of atomic positions in space by vowelizing the structure and facilitating the capture of the local microenvironment of the protein structure. Thus, the 3DCNN automatically extracts protein structural features and has powerful structural characterization capabilities that are compatible with the detection of structural patterns, binding pockets, and other important structural features of specific shapes. Li et al. [191] used deep 3D convolutional neural networks to predict the changes in the thermodynamic stability of proteins upon point mutations. Zhao et al. [196] predicted the binding sites of metal ions on RNA by 3DCNN.

The GVP-GNN introduces Geometric Vector Perceptrons (GVPs) and extends the standard dense layer to enable manipulation of a collection of Euclidean vectors [192]. By introducing GVPs, GVP-GNN can incorporate protein 3D structure vectors into GNNs that satisfy rotational translation covariance and conveniently capture spatial neighborhood information to enhance the ability of the GNN to represent proteins. GVP-GNN can

also accomplish covariant and invariant representation of biomolecular geometry with lightweight parameters. It is well suited for biomolecules and biomolecular complexes and is expected to be further developed in the field of intelligent protein design.

3.4. Hybrid Sequence–Structure-Based Characterization

Protein design often relies on three-dimensional structural data to fully capture the functional information of proteins. It is typically richer than the information provided by sequence data. However, current models predominantly use sequence features, owing to the lack of proper 3D structure characterization methods. Most are computationally expensive and cannot avoid information loss when dimensional reduction is performed. Furthermore, deep learning models may not fully explore the hidden information in high-dimensional data [197]. Consequently, multi-scale representation methods that incorporate sequence and structural information have emerged for protein design. Currently, there is a lack of direct representation methods for multimodal data; therefore, researchers mainly separately use the sequence and structural representation methods described above, and then merge the extracted feature vectors using downstream models. Sequence information provides complementary information that is not fully covered by three-dimensional structure data. This can improve the accuracy of predicting protein–small molecule interactions and protein functional sites [125,128].

4. Conclusions and Outlook

At present, intelligent protein design is in a boom period, and several intelligent protein design models were developed, including SCUBA, ABACUS, ProteinMPNN, and RFdiffusion. This significantly improved the success rate and computational efficiency of protein design. However, the accurate and rapid protein design concept of 'You get what you designed' is yet to be realized in practice.

Effective protein characterization is essential for intelligent protein design. Four protein characterization methods (namely, traditional descriptor-based, sequence-based, structure-based, and hybrid sequence-structure-based methods), were introduced. Traditional protein representation methods were applied in the early days due to their simplicity and ease of understanding. However, they could not comprehensively represent proteins. The similarity between natural language and protein sequences resulted in sequence-based protein characterization methods based on NLP becoming the main method for protein sequence characterization. Structure-based protein characterization methods, such as point clouds based on spatial coordinates and GVP-GNNs based on geometric vectors, have also received widespread attention with the rapid development of protein structure prediction methods and artificial intelligence algorithms. However, their applications are limited because of their high computational requirements. Researchers have attempted to integrate sequence and structural information to represent proteins to comprehensively consider computational power and protein characterization; however, determining the best combination of multiple features remains still an open question.

Although the representation of proteins for intelligent model construction is largely resolved, there is no consensus on which representation is most appropriate for characterizing proteins. We believe that a large amount of protein structure resolution and the development of intelligent algorithms will inspire new efforts to improve protein characterization. This would promise to accurately extract useful information from the vast amount of data, and associate sequence structure information with functional phenotypes to enable efficient and accurate protein design with new functions.

Author Contributions: J.W.: Conception and writing the manuscript. C.C.: Image modification. G.Y. and J.D.: Research Grants. L.W.: Conception and revision of the manuscript. H.J.: Research Grants and academic supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially support by State Key Laboratory of NBC Protection for Civilian, Beijing, and the National Key R&D Program of China (Grant No. 2018YFA0900400).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Woolfson, D.N. A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. *J. Mol. Biol.* **2021**, *433*, 167160. [CrossRef]
2. Meinen, B.A.; Bahl, C.D. Breakthroughs in Computational Design Methods Open up New Frontiers for De Novo Protein Engineering. *Protein Eng. Des. Sel.* **2021**, *34*, gzab007. [CrossRef] [PubMed]
3. Hill, R.B.; Raleigh, D.P.; Lombardi, A.; DeGrado, W.F. De Novo Design of Helical Bundles as Models for Understanding Protein Folding and Function. *Acc. Chem. Res.* **2000**, *33*, 745–754. [CrossRef] [PubMed]
4. Simons, K.T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of Protein Tertiary Structures from Fragments with Similar Local Sequences Using Simulated Annealing and Bayesian Scoring Functions. *J. Mol. Biol.* **1997**, *268*, 209–225. [CrossRef] [PubMed]
5. Gibney, B.R.; Rabanal, F.; Skalicky, J.J.; Wand, A.J.; Dutton, P.L. Design of a Unique Protein Scaffold for Maquettes. *J. Am. Chem. Soc.* **1997**, *119*, 2323–2324. [CrossRef]
6. Gibney, B.R.; Rabanal, F.; Skalicky, J.J.; Wand, A.J.; Dutton, P.L. Iterative Protein Redesign. *J. Am. Chem. Soc.* **1999**, *121*, 4952–4960. [CrossRef]
7. Dahiyat, B.I.; Mayo, S.L. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **1997**, *278*, 82–87. [CrossRef]
8. Dantas, G.; Kuhlman, B.; Callender, D.; Wong, M.; Baker, D. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. *J. Mol. Biol.* **2003**, *332*, 449–460. [CrossRef]
9. Kuhlman, B.; Dantas, G.; Ireton, G.C.; Varani, G.; Stoddard, B.L.; Baker, D. Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. *Science* **2003**, *302*, 1364–1368. [CrossRef]
10. Ingraham, J.; Garg, V.K.; Barzilay, R.; Jaakkola, T. Generative Models for Graph-Based Protein Design. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
11. Strokach, A.; Becerra, D.; Corbi-Verge, C.; Perez-Riba, A.; Kim, P.M. Fast and Flexible Protein Design Using Deep Graph Neural Networks. *Cell Syst.* **2020**, *11*, 402–411.e4. [CrossRef]
12. Anishchenko, I.; Pellock, S.J.; Chidyausiku, T.M.; Ramelot, T.A.; Ovchinnikov, S.; Hao, J.; Bafna, K.; Norn, C.; Kang, A.; Bera, A.K.; et al. De Novo Protein Design by Deep Network Hallucination. *Nature* **2021**, *600*, 547–552. [CrossRef] [PubMed]
13. Wang, J.; Lisanza, S.; Juergens, D.; Tischer, D.; Watson, J.L.; Castro, K.M.; Ragotte, R.; Saragovi, A.; Milles, L.F.; Baek, M.; et al. Scaffolding Protein Functional Sites Using Deep Learning. *Science* **2022**, *377*, 387–394. [CrossRef] [PubMed]
14. Röthlisberger, D.; Khersonsky, O.; Wollacott, A.M.; Jiang, L.; DeChancie, J.; Betker, J.; Gallaher, J.L.; Althoff, E.A.; Zanghellini, A.; Dym, O.; et al. Kemp Elimination Catalysts by Computational Enzyme Design. *Nature* **2008**, *453*, 190–195. [CrossRef] [PubMed]
15. Jiang, L.; Althoff, E.A.; Clemente, F.R.; Doyle, L.; Röthlisberger, D.; Zanghellini, A.; Gallaher, J.L.; Betker, J.L.; Tanaka, F.; Barbas, C.F.; et al. De Novo Computational Design of Retro-Aldol Enzymes. *Science* **2008**, *319*, 1387–1391. [CrossRef]
16. Bolon, D.N.; Mayo, S.L. Enzyme-like Proteins by Computational Design. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 14274–14279. [CrossRef]
17. Kaplan, J.; DeGrado, W.F. De Novo Design of Catalytic Proteins. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 11566–11570. [CrossRef]
18. Liang, H.; Chen, H.; Fan, K.; Wei, P.; Guo, X.; Jin, C.; Zeng, C.; Tang, C.; Lai, L. De Novo Design of a Beta Alpha Beta Motif. *Angew. Chem. Int. Ed.* **2009**, *48*, 3301–3303. [CrossRef]
19. Bellows, M.L.; Taylor, M.S.; Cole, P.A.; Shen, L.; Siliciano, R.F.; Fung, H.K.; Floudas, C.A. Discovery of Entry Inhibitors for HIV-1 via a New De Novo Protein Design Framework. *Biophys. J.* **2010**, *99*, 3445–3453. [CrossRef]
20. Korendovych, I.V.; Senes, A.; Kim, Y.H.; Lear, J.D.; Fry, H.C.; Therien, M.J.; Blasie, J.K.; Walker, F.A.; DeGrado, W.F. De Novo Design and Molecular Assembly of a Transmembrane Diporphyrin-Binding Protein Complex. *J. Am. Chem. Soc.* **2010**, *132*, 15516–15518. [CrossRef]
21. Mitra, P.; Shultis, D.; Zhang, Y. EvoDesign: De Novo Protein Design Based on Structural and Evolutionary Profiles. *Nucleic Acids Res.* **2013**, *41*, W273–W280. [CrossRef]
22. Fairbrother, W.J.; Ashkenazi, A. Designer Proteins to Trigger Cell Death. *Cell* **2014**, *157*, 1506–1508. [CrossRef] [PubMed]
23. Murphy, G.S.; Sathyamoorthy, B.; Der, B.S.; Machius, M.C.; Pulavarti, S.V.; Szyperski, T.; Kuhlman, B. Computational De Novo Design of a Four-Helix Bundle Protein—DND_4HB. *Protein Sci.* **2015**, *24*, 434–445. [CrossRef] [PubMed]
24. Chevalier, A.; Silva, D.-A.; Rocklin, G.J.; Hicks, D.R.; Vergara, R.; Murapa, P.; Bernard, S.M.; Zhang, L.; Lam, K.-H.; Yao, G.; et al. Massively Parallel De Novo Protein Design for Targeted Therapeutics. *Nature* **2017**, *550*, 74–79. [CrossRef]

25. Löffler, P.; Schmitz, S.; Hupfeld, E.; Sterner, R.; Merkl, R. Rosetta:MSF: A Modular Framework for Multi-State Computational Protein Design. *PLoS Comput. Biol.* **2017**, *13*, e1005600. [CrossRef]
26. Shen, H.; Fallas, J.A.; Lynch, E.; Sheffler, W.; Parry, B.; Jannetty, N.; Decarreau, J.; Wagenbach, M.; Vicente, J.J.; Chen, J.; et al. De Novo Design of Self-Assembling Helical Protein Filaments. *Science* **2018**, *362*, 705–709. [CrossRef] [PubMed]
27. Park, J.; Selvaraj, B.; McShan, A.C.; Boyken, S.E.; Wei, K.Y.; Oberdorfer, G.; DeGrado, W.; Sgourakis, N.G.; Cuneo, M.J.; Myles, D.A.; et al. De Novo Design of a Homo-Trimeric Amantadine-Binding Protein. *eLife* **2019**, *8*, e47839. [CrossRef] [PubMed]
28. Russ, W.P.; Figliuzzi, M.; Stocker, C.; Barrat-Charlaix, P.; Socolich, M.; Kast, P.; Hilvert, D.; Monasson, R.; Cocco, S.; Weigt, M.; et al. An Evolution-Based Model for Designing Chorismate Mutase Enzymes. *Science* **2020**, *369*, 440–445. [CrossRef]
29. Chidyausiku, T.M.; Mendes, S.R.; Klima, J.C.; Nadal, M.; Eckhard, U.; Roel-Touris, J.; Houliston, S.; Guevara, T.; Haddox, H.K.; Moyer, A.; et al. De Novo Design of Immunoglobulin-like Domains. *Nat. Commun.* **2022**, *13*, 5661. [CrossRef]
30. Cao, L.; Coventry, B.; Goreshnik, I.; Huang, B.; Sheffler, W.; Park, J.S.; Jude, K.M.; Marković, I.; Kadam, R.U.; Verschueren, K.H.G.; et al. Design of Protein-Binding Proteins from the Target Structure Alone. *Nature* **2022**, *605*, 551–560. [CrossRef]
31. Liao, J.; Warmuth, M.K.; Govindarajan, S.; Ness, J.E.; Wang, R.P.; Gustafsson, C.; Minshull, J. Engineering Proteinase K Using Machine Learning and Synthetic Genes. *BMC Biotechnol.* **2007**, *7*, 16. [CrossRef]
32. Greener, J.G.; Moffat, L.; Jones, D.T. Design of Metalloproteins and Novel Protein Folds Using Variational Autoencoders. *Sci. Rep.* **2018**, *8*, 16189. [CrossRef]
33. Wang, J.; Cao, H.; Zhang, J.Z.H.; Qi, Y. Computational Protein Design with Deep Learning Neural Networks. *Sci. Rep.* **2018**, *8*, 6349. [CrossRef] [PubMed]
34. Alley, E.C.; Khimulya, G.; Biswas, S.; AlQuraishi, M.; Church, G.M. Unified Rational Protein Engineering with Sequence-Based Deep Representation Learning. *Nat. Methods* **2019**, *16*, 1315–1322. [CrossRef]
35. Chen, X.; Chen, Z.; Xu, D.; Lyu, Y.; Li, Y.; Li, S.; Wang, J.; Wang, Z. De Novo Design of G Protein-Coupled Receptor 40 Peptide Agonists for Type 2 Diabetes Mellitus Based on Artificial Intelligence and Site-Directed Mutagenesis. *Front. Bioeng. Biotechnol.* **2021**, *9*, 694100. [CrossRef] [PubMed]
36. Repecka, D.; Jauniskis, V.; Karpus, L.; Rembeza, E.; Rokaitis, I.; Zrimec, J.; Poviloniene, S.; Laurynenas, A.; Viknander, S.; Abuajwa, W.; et al. Expanding Functional Protein Sequence Spaces Using Generative Adversarial Networks. *Nat. Mach. Intell.* **2021**, *3*, 324–333. [CrossRef]
37. Ferruz, N.; Schmidt, S.; Höcker, B. ProtGPT2 Is a Deep Unsupervised Language Model for Protein Design. *Nat. Commun.* **2022**, *13*, 4348. [CrossRef]
38. Li, S.; Wang, L.; Meng, J.; Zhao, Q.; Zhang, L.; Liu, H. De Novo Design of Potential Inhibitors against SARS-CoV-2 Mpro. *Comput. Biol. Med.* **2022**, *147*, 105728. [CrossRef]
39. Kucera, T.; Togninalli, M.; Meng-Papaxanthos, L. Conditional Generative Modeling for De Novo Protein Design with Hierarchical Functions. *Bioinformatics* **2022**, *38*, 3454–3461. [CrossRef]
40. Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R.J.; Milles, L.F.; Wicky, B.I.M.; Courbet, A.; de Haas, R.J.; Bethel, N.; et al. Robust Deep Learning-Based Protein Sequence Design Using ProteinMPNN. *Science* **2022**, *378*, 49–56. [CrossRef]
41. Watson, J.L.; Juergens, D.; Bennett, N.R.; Trippe, B.L.; Yim, J.; Eisenach, H.E.; Ahern, W.; Borst, A.J.; Ragotte, R.J.; Milles, L.F.; et al. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100. [CrossRef]
42. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
43. Tunyasuvunakool, K.; Adler, J.; Wu, Z.; Green, T.; Zielinski, M.; Žídek, A.; Bridgland, A.; Cowie, A.; Meyer, C.; Laydon, A.; et al. Highly Accurate Protein Structure Prediction for the Human Proteome. *Nature* **2021**, *596*, 590–596. [CrossRef] [PubMed]
44. Bileschi, M.L.; Belanger, D.; Bryant, D.H.; Sanderson, T.; Carter, B.; Sculley, D.; Bateman, A.; DePristo, M.A.; Colwell, L.J. Using Deep Learning to Annotate the Protein Universe. *Nat. Biotechnol.* **2022**, *40*, 932–937. [CrossRef] [PubMed]
45. Charoenkwan, P.; Chotpatiwetchkul, W.; Lee, V.S.; Nantasenamat, C.; Shoombuatong, W. A Novel Sequence-Based Predictor for Identifying and Characterizing Thermophilic Proteins Using Estimated Propensity Scores of Dipeptides. *Sci. Rep.* **2021**, *11*, 23782. [CrossRef] [PubMed]
46. Jia, L.; Sun, T.; Wang, Y.; Shen, Y. A Machine Learning Study on the Thermostability Prediction of (R)-ω-Selective Amine Transaminase from Aspergillus Terreus. *BioMed Res. Int.* **2021**, *2021*, 2593748. [CrossRef] [PubMed]
47. Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0: Predicting Stability Changes upon Mutation from the Protein Sequence or Structure. *Nucleic Acids Res.* **2005**, *33*, W306–W310. [CrossRef] [PubMed]
48. Trimble, J.S.; Crawshaw, R.; Hardy, F.J.; Levy, C.W.; Brown, M.J.B.; Fuerst, D.E.; Heyes, D.J.; Obexer, R.; Green, A.P. A Designed Photoenzyme for Enantioselective [2+2] Cycloadditions. *Nature* **2022**, *611*, 709–714. [CrossRef] [PubMed]
49. Sun, N.; Huang, J.; Qian, J.; Zhou, T.-P.; Guo, J.; Tang, L.; Zhang, W.; Deng, Y.; Zhao, W.; Wu, G.; et al. Enantioselective [2+2]-Cycloadditions with Triplet Photoenzymes. *Nature* **2022**, *611*, 715–720. [CrossRef]
50. Tubiana, J.; Schneidman-Duhovny, D.; Wolfson, H.J. ScanNet: An Interpretable Geometric Deep Learning Model for Structure-Based Protein Binding Site Prediction. *Nat. Methods* **2022**, *19*, 730–739. [CrossRef]
51. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction. *bioRxiv* **2022**. [CrossRef]
52. Wang, W.; Peng, Z.; Yang, J. Single-Sequence Protein Structure Prediction Using Supervised Transformer Protein Language Models. *Nat. Comput. Sci.* **2022**, *2*, 804–814. [CrossRef]

53. Zhou, X.; Zheng, W.; Li, Y.; Pearce, R.; Zhang, C.; Bell, E.W.; Zhang, G.; Zhang, Y. I-TASSER-MTD: A Deep-Learning-Based Platform for Multi-Domain Protein Structure and Function Prediction. *Nat. Protoc.* **2022**, *17*, 2326–2353. [CrossRef] [PubMed]
54. Chowdhury, R.; Bouatta, N.; Biswas, S.; Floristean, C.; Kharkar, A.; Roy, K.; Rochereau, C.; Ahdritz, G.; Zhang, J.; Church, G.M.; et al. Single-Sequence Protein Structure Prediction Using a Language Model and Deep Learning. *Nat. Biotechnol.* **2022**, *40*, 1617–1623. [CrossRef] [PubMed]
55. Anfinsen, C.B. Principles That Govern the Folding of Protein Chains. *Science* **1973**, *181*, 223–230. [CrossRef] [PubMed]
56. Acharya, K.R.; Lloyd, M.D. The Advantages and Limitations of Protein Crystal Structures. *Trends Pharmacol. Sci.* **2005**, *26*, 10–14. [CrossRef] [PubMed]
57. Cavalli, A.; Salvatella, X.; Dobson, C.M.; Vendruscolo, M. Protein Structure Determination from NMR Chemical Shifts. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 9615–9620. [CrossRef]
58. Yip, K.M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-Resolution Protein Structure Determination by Cryo-EM. *Nature* **2020**, *587*, 157–161. [CrossRef]
59. Burley, S.K.; Bhikadiya, C.; Bi, C.; Bittrich, S.; Chao, H.; Chen, L.; Craig, P.A.; Crichlow, G.V.; Dalenberg, K.; Duarte, J.M.; et al. RCSB Protein Data Bank (RCSB.Org): Delivery of Experimentally-Determined PDB Structures alongside One Million Computed Structure Models of Proteins from Artificial Intelligence/Machine Learning. *Nucleic Acids Res.* **2023**, *51*, D488–D508. [CrossRef]
60. UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523–D531. [CrossRef]
61. Magrane, M. UniProt Consortium UniProt Knowledgebase: A Hub of Integrated Protein Data. *Database* **2011**, *2011*, bar009. [CrossRef]
62. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12*, 7–8. [CrossRef] [PubMed]
63. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. [CrossRef] [PubMed]
64. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373*, 871–876. [CrossRef] [PubMed]
65. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat. Methods* **2022**, *19*, 679–682. [CrossRef] [PubMed]
66. Li, Z.; Liu, X.; Chen, W.; Shen, F.; Bi, H.; Ke, G.; Zhang, L. Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. *bioRxiv* **2022**. [CrossRef]
67. Cheng, S.; Wu, R.; Yu, Z.; Li, B.; Zhang, X.; Peng, J.; You, Y. FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours. *arXiv* **2022**, arXiv:2203.00854.
68. Wang, G.; Fang, X.; Wu, Z.; Liu, Y.; Xue, Y.; Xiang, Y.; Yu, D.; Wang, F.; Ma, Y. HelixFold: An Efficient Implementation of AlphaFold2 Using PaddlePaddle. *arXiv* **2022**, arXiv:2207.05477.
69. Liu, S.; Zhang, J.; Chu, H.; Wang, M.; Xue, B.; Ni, N.; Yu, J.; Xie, Y.; Chen, Z.; Chen, M.; et al. PSP: Million-Level Protein Sequence Dataset for Protein Structure Prediction. *arXiv* **2022**, arXiv:2206.12240.
70. Fang, X.; Wang, F.; Liu, L.; He, J.; Lin, D.; Xiang, Y.; Zhang, X.; Wu, H.; Li, H.; Song, L. HelixFold-Single: MSA-Free Protein Structure Prediction by Using Protein Language Model as an Alternative. *arXiv* **2022**, arXiv:2207.13921.
71. Wu, R.; Ding, F.; Wang, R.; Shen, R.; Zhang, X.; Luo, S.; Su, C.; Wu, Z.; Xie, Q.; Berger, B.; et al. High-Resolution De Novo Structure Prediction from Primary Sequence. *bioRxiv* **2022**. [CrossRef]
72. Ruffolo, J.A.; Chu, L.-S.; Mahajan, S.P.; Jeffrey, J. Gray Fast, Accurate Antibody Structure Prediction from Deep Learning on Massive Set of Natural Antibodies. *bioRxiv* **2022**. [CrossRef]
73. Zheng, W.; Wuyun, Q.; Freddolino, P.L.; Zhang, Y. Integrating Deep Learning, Threading Alignments, and a multi-MSA Strategy for High-quality Protein Monomer and Complex Structure Prediction in CASP15. *Proteins* **2023**, *12*, 1684–1703. [CrossRef] [PubMed]
74. Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; et al. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv* **2021**. [CrossRef]
75. Chen, B.; Xie, Z.; Qiu, J.; Ye, Z.; Xu, J.; Tang, J. Improved the Protein Complex Prediction with Protein Language Models. *bioRxiv* **2022**. [CrossRef]
76. Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef] [PubMed]
77. Steinegger, M.; Söding, J. Clustering Huge Protein Sequence Sets in Linear Time. *Nat. Commun.* **2018**, *9*, 2542. [CrossRef]
78. UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef]
79. Hippe, K.; Gbenro, S.; Cao, R. ProLanGO2: Protein Function Prediction with Ensemble of Encoder-Decoder Networks. In Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, ACM, Virtual Event, 21–24 September 2020; pp. 1–6. [CrossRef]
80. Gligorijević, V.; Renfrew, P.D.; Kosciolek, T.; Leman, J.K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B.C.; Fisk, I.M.; Vlamakis, H.; et al. Structure-Based Protein Function Prediction Using Graph Convolutional Networks. *Nat. Commun.* **2021**, *12*, 3168. [CrossRef] [PubMed]

81. You, R.; Yao, S.; Mamitsuka, H.; Zhu, S. DeepGraphGO: Graph Neural Network for Large-Scale, Multispecies Protein Function Prediction. *Bioinformatics* **2021**, *37*, i262–i271. [CrossRef]
82. Schug, J.; Diskin, S.; Mazzarelli, J.; Brunk, B.P.; Stoeckert, C.J. Predicting Gene Ontology Functions from ProDom and CDD Protein Domains. *Genome Res.* **2002**, *12*, 648–655. [CrossRef]
83. Das, S.; Lee, D.; Sillitoe, I.; Dawson, N.L.; Lees, J.G.; Orengo, C.A. Functional Classification of CATH Superfamilies: A Domain-Based Approach for Protein Function Annotation. *Bioinformatics* **2015**, *31*, 3460–3467. [CrossRef] [PubMed]
84. Koo, D.C.E.; Bonneau, R. Towards Region-Specific Propagation of Protein Functions. *Bioinformatics* **2019**, *35*, 1737–1744. [CrossRef] [PubMed]
85. Wass, M.N.; Barton, G.; Sternberg, M.J.E. CombFunc: Predicting Protein Function Using Heterogeneous Data Sources. *Nucleic Acids Res.* **2012**, *40*, W466–W470. [CrossRef] [PubMed]
86. Guan, Y.; Myers, C.L.; Hess, D.C.; Barutcuoglu, Z.; Caudy, A.A.; Troyanskaya, O.G. Predicting Gene Function in a Hierarchical Context with an Ensemble of Classifiers. *Genome Biol.* **2008**, *9*, S3. [CrossRef] [PubMed]
87. Törönen, P.; Medlar, A.; Holm, L. PANNZER2: A Rapid Functional Annotation Web Server. *Nucleic Acids Res.* **2018**, *46*, W84–W88. [CrossRef]
88. Mostafavi, S.; Ray, D.; Warde-Farley, D.; Grouios, C.; Morris, Q. GeneMANIA: A Real-Time Multiple Association Network Integration Algorithm for Predicting Gene Function. *Genome Biol.* **2008**, *9*, S4. [CrossRef]
89. 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]
90. Gligorijević, V.; Barot, M.; Bonneau, R. deepNF: Deep Network Fusion for Protein Function Prediction. *Bioinformatics* **2018**, *34*, 3873–3881. [CrossRef]
91. Regan, L.; DeGrado, W.F. Characterization of a Helical Protein Designed from First Principles. *Science* **1988**, *241*, 976–978. [CrossRef]
92. Siegel, J.B.; Zanghellini, A.; Lovick, H.M.; Kiss, G.; Lambert, A.R.; St. Clair, J.L.; Gallaher, J.L.; Hilvert, D.; Gelb, M.H.; Stoddard, B.L.; et al. Computational Design of an Enzyme Catalyst for a Stereoselective Bimolecular Diels-Alder Reaction. *Science* **2010**, *329*, 309–313. [CrossRef] [PubMed]
93. Siegel, J.B.; Smith, A.L.; Poust, S.; Wargacki, A.J.; Bar-Even, A.; Louw, C.; Shen, B.W.; Eiben, C.B.; Tran, H.M.; Noor, E.; et al. Computational Protein Design Enables a Novel One-Carbon Assimilation Pathway. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 3704–3709. [CrossRef] [PubMed]
94. Cai, T.; Sun, H.; Qiao, J.; Zhu, L.; Zhang, F.; Zhang, J.; Tang, Z.; Wei, X.; Yang, J.; Yuan, Q.; et al. Cell-Free Chemoenzymatic Starch Synthesis from Carbon Dioxide. *Science* **2021**, *373*, 1523–1527. [CrossRef] [PubMed]
95. Ferguson, A.L.; Ranganathan, R. 100th Anniversary of Macromolecular Science Viewpoint: Data-Driven Protein Design. *ACS Macro Lett.* **2021**, *10*, 327–340. [CrossRef] [PubMed]
96. Huang, B.; Xu, Y.; Hu, X.; Liu, Y.; Liao, S.; Zhang, J.; Huang, C.; Hong, J.; Chen, Q.; Liu, H. A Backbone-Centred Energy Function of Neural Networks for Protein Design. *Nature* **2022**, *602*, 523–528. [CrossRef]
97. An, L.; Hicks, D.R.; Zorine, D.; Dauparas, J.; Wicky, B.I.M.; Milles, L.F.; Courbet, A.; Bera, A.K.; Nguyen, H.; Kang, A.; et al. Hallucination of Closed Repeat Proteins Containing Central Pockets. *Nat. Struct. Mol. Biol.* **2023**, *30*, 1755–1760. [CrossRef] [PubMed]
98. Doyle, L.A.; Takushi, B.; Kibler, R.D.; Milles, L.F.; Orozco, C.T.; Jones, J.D.; Jackson, S.E.; Stoddard, B.L.; Bradley, P. De Novo Design of Knotted Tandem Repeat Proteins. *Nat. Commun.* **2023**, *14*, 6746. [CrossRef] [PubMed]
99. Ovchinnikov, S.; Huang, P.-S. Structure-Based Protein Design with Deep Learning. *Curr. Opin. Chem. Biol.* **2021**, *65*, 136–144. [CrossRef]
100. Anand, N.; Eguchi, R.; Mathews, I.I.; Perez, C.P.; Derry, A.; Altman, R.B.; Huang, P.-S. Protein Sequence Design with a Learned Potential. *Nat. Commun.* **2022**, *13*, 746. [CrossRef]
101. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminform.* **2020**, *12*, 56. [CrossRef]
102. Zhang, F.; Zhao, B.; Shi, W.; Li, M.; Kurgan, L. DeepDISOBind: Accurate Prediction of RNA-, DNA- and Protein-Binding Intrinsically Disordered Residues with Deep Multi-Task Learning. *Brief. Bioinform.* **2022**, *23*, bbab521. [CrossRef]
103. Lee, I.; Nam, H. Sequence-Based Prediction of Protein Binding Regions and Drug-Target Interactions. *J. Cheminform.* **2022**, *14*, 5. [CrossRef]
104. Basu, S.; Kihara, D.; Kurgan, L. Computational Prediction of Disordered Binding Regions. *Comput. Struct. Biotechnol. J.* **2023**, *21*, 1487–1497. [CrossRef] [PubMed]
105. Kulmanov, M.; Zhapa-Camacho, F.; Hoehndorf, R. DeepGOWeb: Fast and Accurate Protein Function Prediction on the (Semantic) Web. *Nucleic Acids Res.* **2021**, *49*, W140–W146. [CrossRef]
106. Kulmanov, M.; Hoehndorf, R. DeepGOPlus: Improved Protein Function Prediction from Sequence. *Bioinformatics* **2020**, *36*, 422–429. [CrossRef] [PubMed]
107. Yunes, J.M.; Babbitt, P.C. Effusion: Prediction of Protein Function from Sequence Similarity Networks. *Bioinformatics* **2019**, *35*, 442–451. [CrossRef]
108. Magliery, T.J. Protein Stability: Computation, Sequence Statistics, and New Experimental Methods. *Curr. Opin. Struct. Biol.* **2015**, *33*, 161–168. [CrossRef]

109. Scarabelli, G.; Oloo, E.O.; Maier, J.K.X.; Rodriguez-Granillo, A. Accurate Prediction of Protein Thermodynamic Stability Changes upon Residue Mutation Using Free Energy Perturbation. *J. Mol. Biol.* **2022**, *434*, 167375. [CrossRef] [PubMed]
110. Wu, X.; Yu, L. EPSOL: Sequence-Based Protein Solubility Prediction Using Multidimensional Embedding. *Bioinformatics* **2021**, *37*, 4314–4320. [CrossRef]
111. Wang, C.; Zou, Q. Prediction of Protein Solubility Based on Sequence Physicochemical Patterns and Distributed Representation Information with DeepSoluE. *BMC Biol.* **2023**, *21*, 12. [CrossRef] [PubMed]
112. Zhou, C.; Yu, H.; Ding, Y.; Guo, F.; Gong, X.-J. Multi-Scale Encoding of Amino Acid Sequences for Predicting Protein Interactions Using Gradient Boosting Decision Tree. *PLoS ONE* **2017**, *12*, e0181426. [CrossRef]
113. Kirkwood, J.; Hargreaves, D.; O'Keefe, S.; Wilson, J. Using Isoelectric Point to Determine the pH for Initial Protein Crystallization Trials. *Bioinformatics* **2015**, *31*, 1444–1451. [CrossRef]
114. Perez-Riverol, Y.; Audain, E.; Millan, A.; Ramos, Y.; Sanchez, A.; Vizcaíno, J.A.; Wang, R.; Müller, M.; Machado, Y.J.; Betancourt, L.H.; et al. Isoelectric Point Optimization Using Peptide Descriptors and Support Vector Machines. *J. Proteom.* **2012**, *75*, 2269–2274. [CrossRef]
115. Aftabuddin, M.; Kundu, S. Hydrophobic, Hydrophilic, and Charged Amino Acid Networks within Protein. *Biophys. J.* **2007**, *93*, 225–231. [CrossRef]
116. Sengupta, D.; Kundu, S. Role of Long- and Short-Range Hydrophobic, Hydrophilic and Charged Residues Contact Network in Protein's Structural Organization. *BMC Bioinform.* **2012**, *13*, 142. [CrossRef] [PubMed]
117. Durell, S.R.; Ben-Naim, A. Hydrophobic-Hydrophilic Forces in Protein Folding. *Biopolymers* **2017**, *107*, e23020. [CrossRef] [PubMed]
118. Oehme, D.P.; Brownlee, R.T.C.; Wilson, D.J.D. Effect of Atomic Charge, Solvation, Entropy, and Ligand Protonation State on MM-PB(GB)SA Binding Energies of HIV Protease. *J. Comput. Chem.* **2012**, *33*, 2566–2580. [CrossRef] [PubMed]
119. Hebditch, M.; Carballo-Amador, M.A.; Charonis, S.; Curtis, R.; Warwicker, J. Protein–Sol: A Web Tool for Predicting Protein Solubility from Sequence. *Bioinformatics* **2017**, *33*, 3098–3100. [CrossRef] [PubMed]
120. Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G.-Y.; Bensmail, H.; Mall, R. DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction. *Bioinformatics* **2018**, *34*, 2605–2613. [CrossRef]
121. Munteanu, C.R.; Pimenta, A.C.; Fernandez-Lozano, C.; Melo, A.; Cordeiro, M.N.D.S.; Moreira, I.S. Solvent Accessible Surface Area-Based Hot-Spot Detection Methods for Protein-Protein and Protein-Nucleic Acid Interfaces. *J. Chem. Inf. Model.* **2015**, *55*, 1077–1086. [CrossRef]
122. Faraggi, E.; Zhou, Y.; Kloczkowski, A. Accurate Single-Sequence Prediction of Solvent Accessible Surface Area Using Local and Global Features. *Proteins* **2014**, *82*, 3170–3176. [CrossRef]
123. Houghtaling, J.; Ying, C.; Eggenberger, O.M.; Fennouri, A.; Nandivada, S.; Acharjee, M.; Li, J.; Hall, A.R.; Mayer, M. Estimation of Shape, Volume, and Dipole Moment of Individual Proteins Freely Transiting a Synthetic Nanopore. *ACS Nano* **2019**, *13*, 5231–5242. [CrossRef]
124. Pathak, P.; Shvartsburg, A.A. Assessing the Dipole Moments and Directional Cross Sections of Proteins and Complexes by Differential Ion Mobility Spectrometry. *Anal. Chem.* **2022**, *94*, 7041–7049. [CrossRef]
125. Wang, J.; Zhao, Y.; Gong, W.; Liu, Y.; Wang, M.; Huang, X.; Tan, J. EDLMFC: An Ensemble Deep Learning Framework with Multi-Scale Features Combination for ncRNA–Protein Interaction Prediction. *BMC Bioinform.* **2021**, *22*, 133. [CrossRef] [PubMed]
126. Suresh, V.; Liu, L.; Adjeroh, D.; Zhou, X. RPI-Pred: Predicting ncRNA-Protein Interaction Using Sequence and Structural Information. *Nucleic Acids Res.* **2015**, *43*, 1370–1379. [CrossRef] [PubMed]
127. Su, X.-R.; Hu, L.; You, Z.-H.; Hu, P.-W.; Zhao, B.-W. Multi-View Heterogeneous Molecular Network Representation Learning for Protein-Protein Interaction Prediction. *BMC Bioinform.* **2022**, *23*, 234. [CrossRef]
128. Liu, Y.; Gong, W.; Zhao, Y.; Deng, X.; Zhang, S.; Li, C. aPRBind: Protein–RNA Interface Prediction by Combining Sequence and I-TASSER Model-Based Structural Features Learned with Convolutional Neural Networks. *Bioinformatics* **2021**, *37*, 937–942. [CrossRef]
129. Hong, X.; Lv, J.; Li, Z.; Xiong, Y.; Zhang, J.; Chen, H.-F. Sequence-Based Machine Learning Method for Predicting the Effects of Phosphorylation on Protein-Protein Interactions. *Int. J. Biol. Macromol.* **2023**, *243*, 125233. [CrossRef] [PubMed]
130. Jandrlić, D.R. SVM and SVR-Based MHC-Binding Prediction Using a Mathematical Presentation of Peptide Sequences. *Comput. Biol. Chem.* **2016**, *65*, 117–127. [CrossRef]
131. Chen, C.; Zhang, Q.; Yu, B.; Yu, Z.; Lawrence, P.J.; Ma, Q.; Zhang, Y. Improving Protein-Protein Interactions Prediction Accuracy Using XGBoost Feature Selection and Stacked Ensemble Classifier. *Comput. Biol. Med.* **2020**, *123*, 103899. [CrossRef]
132. Gu, X.; Chen, Z.; Wang, D. Prediction of G Protein-Coupled Receptors With CTDC Extraction and MRMD2.0 Dimension-Reduction Methods. *Front. Bioeng. Biotechnol.* **2020**, *8*, 635. [CrossRef]
133. Meher, P.K.; Sahu, T.K.; Mohanty, J.; Gahoi, S.; Purru, S.; Grover, M.; Rao, A.R. nifPred: Proteome-Wide Identification and Categorization of Nitrogen-Fixation Proteins of Diaztrophs Based on Composition-Transition-Distribution Features Using Support Vector Machine. *Front. Microbiol.* **2018**, *9*, 1100. [CrossRef]
134. Yang, S.; Wang, Y.; Lin, Y.; Shao, D.; He, K.; Huang, L. LncMirNet: Predicting LncRNA-miRNA Interaction Based on Deep Learning of Ribonucleic Acid Sequences. *Molecules* **2020**, *25*, 4372. [CrossRef]
135. Ma, X.; Guo, J.; Sun, X. Sequence-Based Prediction of RNA-Binding Proteins Using Random Forest with Minimum Redundancy Maximum Relevance Feature Selection. *BioMed Res. Int.* **2015**, *2015*, 425810. [CrossRef]

136. Firoz, A.; Malik, A.; Ali, H.M.; Akhter, Y.; Manavalan, B.; Kim, C.-B. PRR-HyPred: A Two-Layer Hybrid Framework to Predict Pattern Recognition Receptors and Their Families by Employing Sequence Encoded Optimal Features. *Int. J. Biol. Macromol.* **2023**, *234*, 123622. [CrossRef]
137. Collantes, E.R.; Dunn, W.J. Amino Acid Side Chain Descriptors for Quantitative Structure-Activity Relationship Studies of Peptide Analogs. *J. Med. Chem.* **1995**, *38*, 2705–2713. [CrossRef] [PubMed]
138. Mei, H.; Liao, Z.H.; Zhou, Y.; Li, S.Z. A New Set of Amino Acid Descriptors and Its Application in Peptide QSARs. *Biopolymers* **2005**, *80*, 775–786. [CrossRef]
139. Van Westen, G.J.; Swier, R.F.; Cortes-Ciriano, I.; Wegner, J.K.; Overington, J.P.; Ijzerman, A.P.; van Vlijmen, H.W.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. *J. Cheminformatics* **2013**, *5*, 42. [CrossRef] [PubMed]
140. Zhou, P.; Tian, F.; Wu, Y.; Li, Z.; Shang, Z. Quantitative Sequence-Activity Model (QSAM): Applying QSAR Strategy to Model and Predict Bioactivity and Function of Peptides, Proteins and Nucleic Acids. *CAD* **2008**, *4*, 311–321. [CrossRef]
141. Liang, G.; Li, Z. Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *QSAR Comb. Sci.* **2007**, *26*, 754–763. [CrossRef]
142. Tian, F.; Zhou, P.; Li, Z. T-Scale as a Novel Vector of Topological Descriptors for Amino Acids and Its Application in QSARs of Peptides. *J. Mol. Struct.* **2007**, *830*, 106–115. [CrossRef]
143. Yang, L.; Shu, M.; Ma, K.; Mei, H.; Jiang, Y.; Li, Z. ST-Scale as a Novel Amino Acid Descriptor and Its Application in QSAM of Peptides and Analogues. *Amino Acids* **2010**, *38*, 805–816. [CrossRef] [PubMed]
144. Yue, Z.-X.; Yan, T.-C.; Xu, H.-Q.; Liu, Y.-H.; Hong, Y.-F.; Chen, G.-X.; Xie, T.; Tao, L. A Systematic Review on the State-of-the-Art Strategies for Protein Representation. *Comput. Biol. Med.* **2023**, *152*, 106440. [CrossRef] [PubMed]
145. Zaliani, A.; Gancia, E. MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 525–533. [CrossRef]
146. Muppirala, U.K.; Honavar, V.G.; Dobbs, D. Predicting RNA-Protein Interactions Using Only Sequence Information. *BMC Bioinform.* **2011**, *12*, 489. [CrossRef] [PubMed]
147. Peng, Z.; Kurgan, L. High-Throughput Prediction of RNA, DNA and Protein Binding Regions Mediated by Intrinsic Disorder. *Nucleic Acids Res.* **2015**, *43*, e121. [CrossRef] [PubMed]
148. Soleymani, F.; Paquet, E.; Viktor, H.; Michalowski, W.; Spinello, D. Protein-Protein Interaction Prediction with Deep Learning: A Comprehensive Review. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 5316–5341. [CrossRef]
149. Zhao, L.; Zhu, Y.; Wang, J.; Wen, N.; Wang, C.; Cheng, L. A Brief Review of Protein-Ligand Interaction Prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 2831–2838. [CrossRef]
150. Singh, V.K.; Maurya, N.S.; Mani, A.; Yadav, R.S. Machine Learning Method Using Position-Specific Mutation Based Classification Outperforms One Hot Coding for Disease Severity Prediction in Haemophilia "A". *Genomics* **2020**, *112*, 5122–5128. [CrossRef]
151. Shen, H.; Zhang, Y.; Zheng, C.; Wang, B.; Chen, P. A Cascade Graph Convolutional Network for Predicting Protein-Ligand Binding Affinity. *Int. J. Mol. Sci.* **2021**, *22*, 4023. [CrossRef]
152. Béroud, C.; Joly, D.; Gallou, C.; Staroz, F.; Orfanelli, M.T.; Junien, C. Software and Database for the Analysis of Mutations in the VHL Gene. *Nucleic Acids Res.* **1998**, *26*, 256–258. [CrossRef]
153. Mei, S.; Fei, W. Amino Acid Classification Based Spectrum Kernel Fusion for Protein Subnuclear Localization. *BMC Bioinform.* **2010**, *11* (Suppl. S1), S17. [CrossRef]
154. Li, L.; Luo, Q.; Xiao, W.; Li, J.; Zhou, S.; Li, Y.; Zheng, X.; Yang, H. A Machine-Learning Approach for Predicting Palmitoylation Sites from Integrated Sequence-Based Features. *J. Bioinform. Comput. Biol.* **2017**, *15*, 1650025. [CrossRef] [PubMed]
155. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *arXiv* **2013**, arXiv:1310.4546.
156. Wu, C.; Gao, R.; Zhang, Y.; De Marinis, Y. PTPD: Predicting Therapeutic Peptides by Deep Learning and Word2vec. *BMC Bioinform.* **2019**, *20*, 456. [CrossRef] [PubMed]
157. Miao, Y.; Liu, F.; Hou, T.; Liu, Y. Virtifier: A Deep Learning-Based Identifier for Viral Sequences from Metagenomes. *Bioinformatics* **2022**, *38*, 1216–1222. [CrossRef]
158. Abrahamsson, E.; Plotkin, S.S. BioVEC: A Program for Biomolecule Visualization with Ellipsoidal Coarse-Graining. *J. Mol. Graph. Model.* **2009**, *28*, 140–145. [CrossRef] [PubMed]
159. Yang, X.; Yang, S.; Li, Q.; Wuchty, S.; Zhang, Z. Prediction of Human-Virus Protein-Protein Interactions through a Sequence Embedding-Based Machine Learning Method. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 153–161. [CrossRef] [PubMed]
160. Li, F.; Yuan, L.; Lu, H.; Li, G.; Chen, Y.; Engqvist, M.K.M.; Kerkhoven, E.J.; Nielsen, J. Deep Learning-Based Kcat Prediction Enables Improved Enzyme-Constrained Model Reconstruction. *Nat. Catal.* **2022**, *5*, 662–672. [CrossRef]
161. Yu, L.; Tanwar, D.K.; Penha, E.D.S.; Wolf, Y.I.; Koonin, E.V.; Basu, M.K. Grammar of Protein Domain Architectures. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 3636–3645. [CrossRef]
162. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
163. Yang, X.; Bian, J.; Hogan, W.R.; Wu, Y. Clinical Concept Extraction Using Transformers. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1935–1942. [CrossRef]

164. Chen, Z.-M.; Cui, Q.; Zhao, B.; Song, R.; Zhang, X.; Yoshie, O. SST: Spatial and Semantic Transformers for Multi-Label Image Recognition. *IEEE Trans Image Process* **2022**, *31*, 2570–2583. [CrossRef] [PubMed]
165. Monteiro, N.R.C.; Oliveira, J.L.; Arrais, J.P. DTITR: End-to-End Drug-Target Binding Affinity Prediction with Transformers. *Comput. Biol. Med.* **2022**, *147*, 105772. [CrossRef] [PubMed]
166. Mazuz, E.; Shtar, G.; Shapira, B.; Rokach, L. Molecule Generation Using Transformers and Policy Gradient Reinforcement Learning. *Sci. Rep.* **2023**, *13*, 8799. [CrossRef]
167. Wang, H.; Guo, F.; Du, M.; Wang, G.; Cao, C. A Novel Method for Drug-Target Interaction Prediction Based on Graph Transformers Model. *BMC Bioinform.* **2022**, *23*, 459. [CrossRef]
168. Rodriguez, M.A.; AlMarzouqi, H.; Liatsis, P. Multi-Label Retinal Disease Classification Using Transformers. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 2739–2750. [CrossRef] [PubMed]
169. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
170. Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. OpenAI Blog. 2018. Available online: https://openai.com/research/language-unsupervised (accessed on 20 October 2023).
171. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. OpenAI Blog. 2019. Available online: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (accessed on 20 October 2023).
172. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
173. Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. *bioRxiv* **2021**. [CrossRef]
174. Rao, R.M.; Liu, J.; Verkuil, R.; Meier, J.; Canny, J.; Abbeel, P.; Sercu, T.; Rives, A. MSA Transformer. *bioRxiv* **2021**. [CrossRef]
175. Rives, A.; Meier, J.; Sercu, T.; Goyal, S.; Lin, Z.; Liu, J.; Guo, D.; Ott, M.; Zitnick, C.L.; Ma, J.; et al. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2016239118. [CrossRef]
176. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7112–7127. [CrossRef]
177. Madani, A.; Krause, B.; Greene, E.R.; Subramanian, S.; Mohr, B.P.; Holton, J.M.; Olmos, J.L.; Xiong, C.; Sun, Z.Z.; Socher, R.; et al. Large Language Models Generate Functional Protein Sequences across Diverse Families. *Nat. Biotechnol.* **2023**, *41*, 1099–1106. [CrossRef]
178. Verkuil, R.; Kabeli, O.; Du, Y.; Wicky, B.I.M.; Milles, L.F.; Dauparas, J.; Baker, D.; Sercu, T.; Ovchinnikov, S.; Rives, A. Language Models Generalize beyond Natural Proteins. *bioRxiv* **2022**. [CrossRef]
179. Hie, B.; Candido, S.; Lin, Z.; Kabeli, O.; Rao, R.; Smetanin, N.; Sercu, T.; Alexander Rives, A. A High-Level Programming Language for Generative Protein Design. *bioRxiv* **2022**. [CrossRef]
180. Qu, G.; Li, A.; Acevedo-Rocha, C.G.; Sun, Z.; Reetz, M.T. The Crucial Role of Methodology Development in Directed Evolution of Selective Enzymes. *Angew. Chem. Int. Ed.* **2020**, *59*, 13204–13231. [CrossRef] [PubMed]
181. Cho, S.Y.; Yun, Y.S.; Jang, D.; Jeon, J.W.; Kim, B.H.; Lee, S.; Jin, H.-J. Ultra Strong Pyroprotein Fibres with Long-Range Ordering. *Nat. Commun.* **2017**, *8*, 74. [CrossRef] [PubMed]
182. Yuan, P.; Bartlam, M.; Lou, Z.; Chen, S.; Zhou, J.; He, X.; Lv, Z.; Ge, R.; Li, X.; Deng, T.; et al. Crystal Structure of an Avian Influenza Polymerase PAN Reveals an Endonuclease Active Site. *Nature* **2009**, *458*, 909–913. [CrossRef]
183. Kircali Ata, S.; Fang, Y.; Wu, M.; Li, X.-L.; Xiao, X. Disease Gene Classification with Metagraph Representations. *Methods* **2017**, *131*, 83–92. [CrossRef]
184. Woodard, J.; Iqbal, S.; Mashaghi, A. Circuit Topology Predicts Pathogenicity of Missense Mutations. *Proteins* **2022**, *90*, 1634–1644. [CrossRef] [PubMed]
185. 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]
186. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved Protein Structure Prediction Using Predicted Interresidue Orientations. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1496–1503. [CrossRef] [PubMed]
187. Zheng, W.; Zhang, C.; Li, Y.; Pearce, R.; Bell, E.W.; Zhang, Y. Folding Non-Homologous Proteins by Coupling Deep-Learning Contact Maps with I-TASSER Assembly Simulations. *Cell Rep.* **2021**, *1*, 100014. [CrossRef] [PubMed]
188. Mortuza, S.M.; Zheng, W.; Zhang, C.; Li, Y.; Pearce, R.; Zhang, Y. Improving Fragment-Based Ab Initio Protein Structure Assembly Using Low-Accuracy Contact-Map Predictions. *Nat. Commun.* **2021**, *12*, 5011. [CrossRef]
189. Pearce, R.; Li, Y.; Omenn, G.S.; Zhang, Y. Fast and Accurate Ab Initio Protein Structure Prediction Using Deep Learning Potentials. *PLoS Comput. Biol.* **2022**, *18*, e1010539. [CrossRef]
190. Olechnovič, K.; Venclovas, Č. Voronota: A Fast and Reliable Tool for Computing the Vertices of the Voronoi Diagram of Atomic Balls. *J. Comput. Chem.* **2014**, *35*, 672–681. [CrossRef]
191. Li, B.; Yang, Y.T.; Capra, J.A.; Gerstein, M.B. Predicting Changes in Protein Thermodynamic Stability upon Point Mutation with Deep 3D Convolutional Neural Networks. *PLoS Comput. Biol.* **2020**, *16*, e1008291. [CrossRef]

192. Jing, B.; Eismann, S.; Suriana, P.; Townshend, R.J.L.; Dror, R. Learning from Protein Structure with Geometric Vector Perceptrons. *arXiv* **2021**, arXiv:2009.01411.
193. Wang, Y.; Wu, S.; Duan, Y.; Huang, Y. A Point Cloud-Based Deep Learning Strategy for Protein-Ligand Binding Affinity Prediction. *Brief. Bioinform.* **2022**, *23*, bbab474. [CrossRef]
194. Igashov, I.; Olechnovič, K.; Kadukova, M.; Venclovas, Č.; Grudinin, S. VoroCNN: Deep Convolutional Neural Network Built on 3D Voronoi Tessellation of Protein Structures. *Bioinformatics* **2021**, *37*, 2332–2339. [CrossRef]
195. Dapkūnas, J.; Timinskas, A.; Olechnovič, K.; Margelevičius, M.; Dičiūnas, R.; Venclovas, Č. The PPI3D Web Server for Searching, Analyzing and Modeling Protein–Protein Interactions in the Context of 3D Structures. *Bioinformatics* **2017**, *33*, 935–937. [CrossRef] [PubMed]
196. Zhao, Y.; Wang, J.; Chang, F.; Gong, W.; Liu, Y.; Li, C. Identification of Metal Ion-Binding Sites in RNA Structures Using Deep Learning Method. *Brief. Bioinform.* **2023**, *24*, bbad049. [CrossRef] [PubMed]
197. Defresne, M.; Barbe, S.; Schiex, T. Protein Design with Deep Learning. *Int. J. Mech. Sci.* **2021**, *22*, 11741. [CrossRef] [PubMed]