Briefings in Bioinformatics, 2022, **23**, 1–13

https://doi.org/10.1093/bib/bbab493

Review

A tale of solving two computational challenges in protein science: neoantigen prediction and protein structure prediction

Ngoc Hieu Tran, Jinbo Xu and Ming Li

Corresponding author: Ming Li, David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. E-mail: mli@uwaterloo.ca

Abstract

In this article, we review two challenging computational questions in protein science: neoantigen prediction and protein structure prediction. Both topics have seen significant leaps forward by deep learning within the past five years, which immediately unlocked new developments of drugs and immunotherapies. We show that deep learning models offer unique advantages, such as representation learning and multi-layer architecture, which make them an ideal choice to leverage a huge amount of protein sequence and structure data to address those two problems. We also discuss the impact and future possibilities enabled by those two applications, especially how the data-driven approach by deep learning shall accelerate the progress towards personalized biomedicine.

Keywords: neoantigen prediction, protein structure prediction, deep learning

Introduction

Proteins are the central units of biological activities in living organisms. The functions or malfunctions of proteins are directly related to a wide range of diseases and drugs. Thanks to recent advances in sequencing technologies and community efforts to maintain public data resources [1, 2], large volumes of data have been accumulated and continue to grow at exponential rates. However, a number of computational questions remain, especially some that have been considered as the greatest challenges in science for several decades.

In 2005, in its 125th anniversary issue [3], the *Science* Magazine listed protein structure prediction as one of 125 big open questions: ‘Out of a near infinitude of possible ways to fold, a protein picks one in just tens of microseconds. The same task takes 30 years of computer time’. In 2017, on a different battlefront [4], the *Nature* journal appealed for solutions to the problem of neoantigen identification: ‘Personalized immunotherapy is all the rage but the neoantigen discovery and validation remains a daunting problem’. While it may first appear to be a new topic, the solutions to neoantigen identification are actually built on long-standing research areas that have spanned several decades, including next-generation sequencing [5], protein and peptide identification [6, 7], antigen presentation [8, 9] and immune epitope prediction [10, 11]. Once solved, both of these open questions will not only fundamentally change biomedical research, but also immediately unlock new developments of drugs and therapies. Fortunately, the progress has been significantly accelerated during the past few years and today we are closing in to the final solutions to both questions [12–18]. It is interesting to look back along these journeys, especially the common tool, deep learning, that has enabled such quick progress. The key factor responsible for this progress is the ability of deep learning to infer unknown and implicit features from an unprecedented amount of data by using models of unprecedented scales [19–21].

Major breakthroughs of deep learning in Computer Vision and Natural Language Processing have been rapidly adopted and become the cores of popular real-world applications such as autonomous vehicles, search and recommendation systems, digital personal assistants, etc. Those applications span across many industries, from automotive to banking and finance, retail and e-commerce, among others [19]. As a branch of machine learning, deep learning models essentially predict an output *y* from an input *x*, for example, detecting a pedestrian in a camera image or classifying whether an email is spam or not.

In our view, deep learning offers three unique advantages over conventional machine learning techniques. First, deep learning models are fed with raw data and

Ngoc Hieu Tran is an adjunct assistant professor at the University of Waterloo, Canada.
Jinbo Xu is a professor at the Toyota Technological Institute at Chicago, USA.
Ming Li is a professor at the University of Waterloo, Canada.

Received: September 8, 2021. **Revised:** October 11, 2021. **Accepted:** October 26, 2021

automatically learn the data representations suitable for a prediction task. This form of representation learning is fundamentally different from the traditional approach of carefully handcrafting/engineering features based on domain knowledge. Second, to learn multiple levels of representation, from simple to complex, deep learning models are composed of multiple nested layers of intermediate variables, where the outputs of one layer are fed into the inputs of the next, more abstract layer. This multi-layer architecture, which gives deep learning its name, allows highly complex functions and representations to be learned from the raw data. Lastly, deep learning performance robustly scales up with the size of data and models, and in fact, its successes have been mainly driven by the combination of large datasets, large models, and massive computational power [19–21]. Any research areas that are aligned with the above three advantages will greatly benefit from deep learning. For example, in the field of biomedical research, we are now witnessing a deep learning revolution fuelled by major advances in many sub-fields, including genomics [22], proteomics [23], protein structure determination and design [12], drug design and discovery [24], and immunotherapy [20, 21].

In this review, we focus on the applications of deep learning to address two challenging questions in protein science: neoantigen identification and protein structure prediction. These two topics share three commonalities. First, they have direct implications to the development of drugs and immunotherapies [3, 4, 25]. Second, they have been long considered as optimization problems rather than learning problems, and have not been fully addressed [6, 26–28]. Third, they have the advantage of a tremendous amount of data that has been accumulated for over decades and is ripe for deep learning applications. Protein data, from sequences to structures, are often not interpretable by the naked eye, even for human experts. Hence, representation learning directly from raw data is key to uncover nontrivial insights. Since proteins are produced at the highest level of the central dogma, one could expect a rich and highly complex amount of information encoded in proteins [29]. More importantly, in order to perform biological functions, proteins fold into four different levels of structures, including primary, secondary, tertiary, and quaternary. Thus, deep learning models with their multi-layer design are a natural choice to learn multiple levels of representation of proteins.

Indeed, the last few years have seen a number of deep learning breakthroughs that bring us close to the final solutions to the problems of neoantigen identification and protein structure prediction. For example, Figure 1 demonstrates major leaps forward by deep learning in de novo peptide sequencing and template-free protein structure prediction, two representative tasks of the two problems. The solution to the neoantigen identification problem provides a bird's-eye view of several deep learning advances in proteomics and immunopeptidomics (Figure 2), including de novo peptide sequencing [30–35], tandem mass spectrum and retention time prediction [18, 36–40], protein and peptide identification with data-independent acquisition (DIA) mass spectrometry [18, 32, 39, 40], and MHC binding and immunogenicity prediction for T-cell epitopes [9, 41–43] (MHC: major histocompatibility complex). For the problem of protein structure prediction, we shall cover protein contact and distance prediction [44], contact and distance-based tertiary structure prediction [13, 45, 46], end-to-end training [47] and protein model refinement [48]. Finally, we shall discuss the impact and future possibilities unlocked by these two technologies, with a special focus on research topics that can benefit from deep learning applications.

Neoantigen identification

Neoantigens belong to a broader class of peptides called HLA-bound peptides. HLA, or human leukocyte antigen, is a complex of genes that encode MHC proteins, which transport peptides to the cell surface and present them to T cells. Neoantigens are encoded by tumor-specific mutations; hence, they can be recognized by T cells as 'foreign' to trigger an immune response. As a result, neoantigens represent ideal targets for cancer vaccines and other types of immunotherapy, and in fact, neoantigen-based cancer vaccines are being tested in clinical trials [14, 15, 25]. Identifying neoantigens is very challenging due to their low abundance, the complex heterogeneity and the limited amount of native tumor tissues, the genetic variability and the high specificity of antigen presentation pathway and T cell immunity [16, 25]. Thus, this problem requires us to integrate different technologies from all fronts of proteomics and genomics to address, potentially giving rise to a new class of neoantigen-based cancer therapies.

Figure 2 describes a personalized workflow for neoantigen identification and highlights the analysis tasks where deep learning can be applied. First, HLA-bound peptides, including neoantigens, are captured directly from native tumor tissues by immunoprecipitation and purification assays, and then sequenced by liquid chromatography tandem mass spectrometry (LC–MS/MS) [16]. Due to the low abundance of neoantigens and the limited amount of native tumor tissues, data-independent acquisition (DIA) is preferred over traditional data-dependent acquisition (DDA) as the former can produce a complete profile of all peptides in a sample, thus increasing the sensitivity of neoantigen identification. Since neoantigens carry mutations, their amino acid sequences are not presented in standard databases. They can be identified directly by de novo sequencing from MS/MS spectra. Or, in a different proteogenomic approach, tumor DNA mutations are detected by genome sequencing and translated into customized protein databases and spectral libraries, which are then used to search MS/MS spectra to find the neoantigens. Once identified, candidate neoantigens

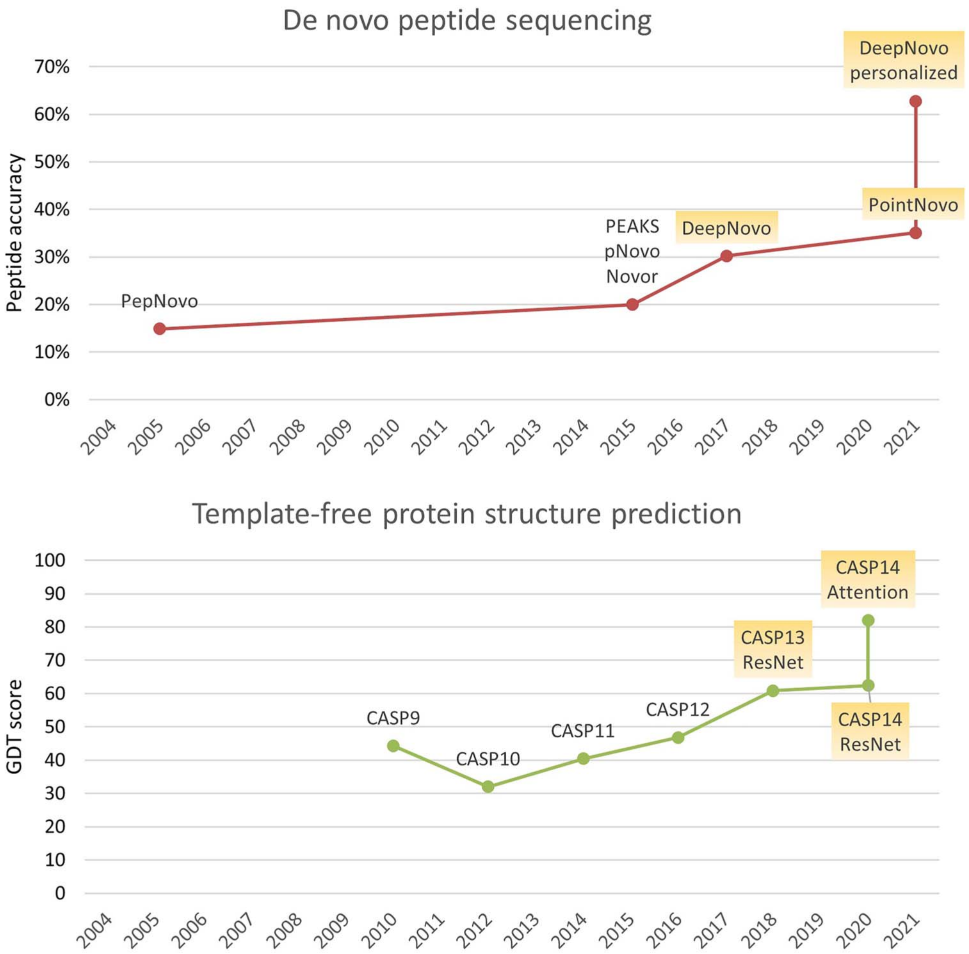

Figure 1. Accuracy improvements of de novo peptide sequencing and template-free protein structure prediction over the past 20 years. Orange boxes indicate major leaps forward by the respective deep learning methods. CASP: Critical Assessment of Protein Structure Prediction.

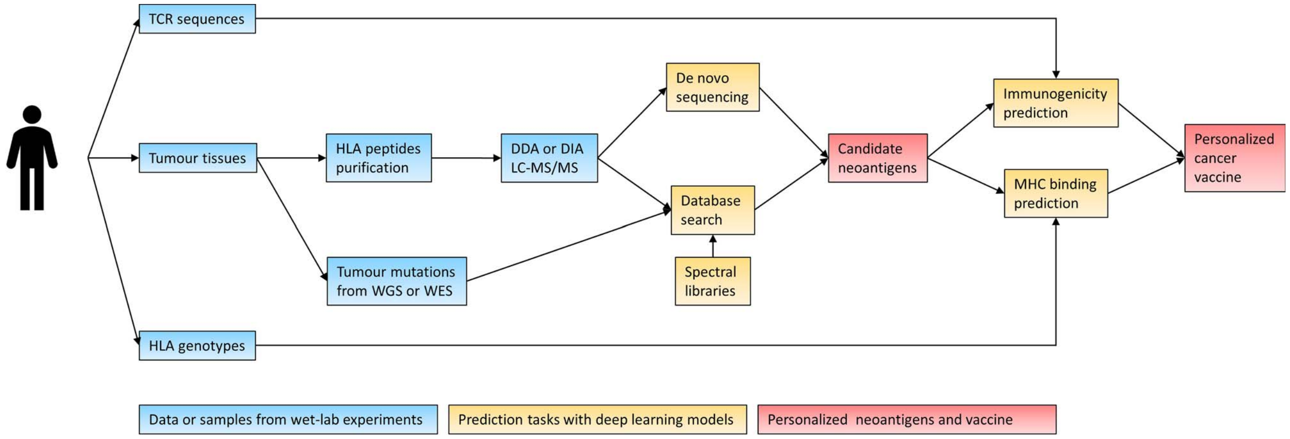
TCR sequences
Tumour tissues
HLA peptides purification
DDA or DIA LC-MS/MS
De novo sequencing
Database search
Spectral libraries
Tumour mutations from WGS or WES
HLA genotypes
Candidate neoantigens
Immunogenicity prediction
MHC binding prediction
Personalized cancer vaccine
Data or samples from wet-lab experiments
Prediction tasks with deep learning models
Personalized neoantigens and vaccine

Figure 2. Personalized workflow for neoantigen identification. TCR: T cell receptor; WGS/WES: whole genome/exome sequencing. HLA: human leukocyte antigen; MHC: major histocompatibility complex. DDA: data-dependent acquisition; DIA: data-independent acquisition. LC–MS/MS: liquid chromatography with tandem mass spectrometry.

are evaluated based on their predicted MHC binding affinity and immunogenicity and then selected for further vaccine development. In the next sections, we shall discuss deep learning applications for de novo sequencing, MS/MS spectrum and retention time prediction, database and spectra library search, and MHC binding affinity and immunogenicity prediction.

De novo sequencing of neoantigens

The task of de novo sequencing is to reconstruct the amino acid sequence of a peptide from a given MS/MS spectrum and the peptide mass, without assisting databases. It is the method of choice to identify novel or mutated peptides, such as neoantigens. An MS/MS spectrum is a collection of masses and intensities of fragment ions acquired from the peptide fragmentation inside a mass spectrometer. The mass differences between the fragment ions may correspond to the masses of amino acids and hence can help to determine their identities (Supplementary Figure S1). Thus, de novo sequencing can be formulated as an optimization problem where one needs to find the best amino acid sequence to interpret the most fragment ions in the spectrum, subject to the given peptide mass. Graph theory and dynamic programming have been used for nearly 20 years to address this problem [6, 28, 49, 50]. In the first approach [49], the spectrum is translated into a graph where nodes correspond to fragment ions and two nodes are connected if their mass difference is equal to the mass of an amino acid. Then one needs to find the highest scoring paths through the graph. In the second approach [6], a dynamic programming algorithm is used to efficiently interrogate all possible combinations of amino acids for the given peptide mass, assign reward or penalty to the observed or missing fragment ions, and find a sequence to maximize the total score recursively.

In a pioneering study published in 2017 [31], Tran et al. introduced deep learning to de novo sequencing. Their model, DeepNovo, is fundamentally different from the conventional optimization approach. DeepNovo sequences a peptide by iteratively predicting one amino acid after another, similarly to composing a sentence by predicting one word after another in Natural Language Processing [19]. Since its iterative framework is simple and errors can be accumulated, DeepNovo relies on two neural networks to make highly accurate predictions at each iteration. First, a convolutional neural network [19] (CNN) coupled with amino acid embedding is used to model the masses and intensities of fragment ions. This model takes into account both fragment ion and amino acid types, whereas previous scoring methods only considered the former. Learning the amino acid representation is very important because the intensities of fragment ions are determined by the bonding between adjacent amino acids. The second model of DeepNovo uses a recurrent neural network (RNN) and amino acid embedding to learn sequential patterns of peptides. The key idea here is to treat protein sequences as a language, where 20 amino acid letters represent its alphabet. The sequential patterns contribute a new dimension of information on top of the spectrum information and can help to overcome the problem of noisy or missing fragment ions in the spectrum. Overall, the deep learning approach by DeepNovo significantly increased the accuracy of de novo sequencing, resulting in 38.1–64.0% more accurate peptides than previous methods [31].

This learning approach has shifted the focus from algorithms to data and models, which can be selected and trained to become application-specific, species-specific, or individual specific (i.e. personalized). This is especially advantageous for de novo sequencing of new neoantigens and antibodies because they are just different from those in existing databases by only a few amino acids, and hence, a model can learn from the existing ones to predict the new ones. Indeed, Tran et al. [17] proposed to train a personalized deep learning model on normal HLA-bound peptides of an individual patient and use it to predict neoantigens of that patient. They were able to expand the immunopeptidomes of five melanoma patients by 5–15% and discover novel neoantigens with T-cell responses. Similar results were demonstrated for antibodies [30, 31] where a deep learning model was trained specifically on known antibodies and then used for de novo assembly of new antibodies, achieving 97.5–100% coverage and 97.2–99.5% accuracy.

A number of deep learning models have been proposed to further improve de novo sequencing. SMSNet [35] used mass tags and an assisting database to refine de novo sequencing results. pNovo3 [33] predicted MS/MS spectra of de novo peptides and compared them to the experimental ones to rank de novo peptides. PointNovo [34] applied a compact representation of MS/MS data and an order-invariant neural network to keep the computational complexity unchanged, regardless of the resolution of the mass spectrometers.

All of those deep learning tools are open source and implemented in Python and Tensorflow or Pytorch (Supplementary Table S1). They also include pre-trained models that were used in the respective publications. However, those models are often not updated after publication and may not be good enough for new datasets. The best way to use those tools is to retrain their models to adapt to new sources of data, rather than expecting a general model that works for different datasets. Indeed, training/retraining is an important feature of deep learning, given massive amounts of data coming from several types of instruments, diverse species and different experiment designs. One can even train a personalized deep learning model to predict neoantigens for each individual patient, as we have discussed above.

Accurate prediction of MS/MS spectra and retention times to improve neoantigen identification

Neoantigens can also be identified using a proteogenomic approach [16, 51]: genome mutations are used to build a customized database consisting of both normal and mutated protein sequences; MS/MS spectra are then searched against that database to identify neoantigens. Accurate prediction of theoretical MS/MS spectra and retention times for candidate database peptides is critical because a search engine compares those theoretical ones to an experimental MS/MS spectrum and its retention time to identify the best candidate peptide. The intensities and retention times of fragment ions depend on many factors, including fragment ion types, precursor charges, instruments, fragmentation methods and collision energies, amino acids and their peptide bonds. Traditional methods such as MassAnalyzer [52] and MS-Simulator [53] were developed based on the kinetic model and mobile proton hypothesis to simulate the peptide fragmentation process. However, their accuracies are limited at ~80% and they are difficult to apply to different fragmentation methods like higher-energy collisional dissociation (HCD), electron-transfer dissociation (ETD) or electron-transfer and higher-energy collision dissociation (EThcD). A machine learning-approach was proposed in PeptideART [54] using a two-layer feed-forward neural network and handcrafted features, including amino acid compositions, ion masses, N- and C-termini, and physicochemical properties such as hydrophobicity or helicity. The accuracy was improved to near 89% for major b- and y-ions, but overall, it is still lower than that of within-experiment replicates.

Inspired by the successes of deep learning in Natural Language Processing, several studies have applied deep learning models to predict MS/MS spectra and retention times from peptide sequences, including pDeep [36], DeepRT [37], Prosit [18, 38], DeepDIA [39], Dia-NN[40], etc. Among them, pDeep was the first deep learning tool and outperformed traditional tools across different instruments and fragmentation methods. One of the key ideas is learning the representation vectors of amino acids, which can automatically capture physicochemical properties of amino acids rather than manually selecting features such as hydrophobicity or helicity as in traditional methods. Another key idea is using bidirectional long short-term memory networks to automatically capture the dependencies of sequential patterns in the whole peptide, rather than manually selecting which amino acid positions to consider. These two models together improved the accuracy of predicted fragment ions to ~95%. Furthermore, the models were able to reveal the fragmentation behaviors between amino acids and to distinguish isobaric amino acids, e.g. isoleucine and leucine, which have the same mass and cannot be differentiated by traditional search engines.

In addition to model architectures, significant improvements of deep learning also come from the massive amount of training data. For instance, Prosit was trained on synthetic peptide libraries of 550 000 tryptic peptides [38] and 300 000 non-tryptic peptides [18] from the ProteomeTools project, making it one of the most comprehensive deep learning models for spectrum and retention time prediction. Predicted MS/MS spectra and retention times were reported to be nearly identical to the experimental ones, with Pearson correlation coefficients >0.99. Those accurate predictions of Prosit substantially increased the sensitivity of database search engines, resulting in 5–35% more peptide identifications. Moreover, Wilhelm et al. [18] showed that such accurate predictions improved the identification of HLA-bound peptides by up to 7-fold. Dozens of additional immunogenic neoantigens were also discovered from melanoma patients, much higher than previously reported by standard database search [16, 18].

A list of deep learning tools for MS/MS spectrum and retention time prediction and their availability are provided in Supplementary Table S2. Notably, in addition to the open source code, an online tool is also available for Prosit as part of the ProteomeTools project. The predicted spectra and retention times can also be used to rescore the database search results with MaxQuant [55] and Percolator [56]. The online tool comes in handy for those who need quick predictions for small datasets, without having to go through the troubles of model training and implementation.

DIA mass spectrometry to boost sensitivity of neoantigen detection

DIA strategies allow unbiased fragmentation of all precursor ions within a wide window of *m/z* and retention time, thus producing a complete profile of all peptides in a sample [57–59]. This is crucial to address the problem of low abundance of neoantigens in tumor tissues [60]. However, DIA spectra are highly multiplexed as they contain fragment ions of multiple peptides. The prevalent approach for DIA analysis is to search DIA spectra against spectral libraries of known peptides to identify the best matching peptides [61–63]. Thus, its performance depends significantly on the availability of spectral libraries, which in turn may depend on a variety of experimental conditions and are costly to obtain. In silico spectral libraries, which include MS/MS spectra and retention times predicted by deep learning, can be built directly from protein databases at much less cost than experimental spectral libraries. They are also easier to recalibrate and transfer across different types of peptides and instruments. Gessulat *et al.* [38] showed that Prosit in silico spectral libraries could replace experimental spectral libraries and give the same DIA identification results on different species and instruments. More importantly, Pak et al. [60] showed that DIA immunopeptidomics workflows using Prosit in silico spectral libraries of HLA-bound peptides could achieve higher sensitivity and better neoantigen discovery than the DDA approach. In silico spectral libraries of HLA-bound peptides are especially valuable because existing HLA spectral libraries are limited while HLA alleles are among the most genetically diverse regions of human genome.

Deep learning models have also been applied to learn coeluting patterns of precursor and fragment ions in DIA spectra, which are then used for direct DIA database search [40] and DIA de novo sequencing [32]. The capability to recognize shapes and objects of deep learning gives it advantages over other methods, e.g. using Pearson correlation, for discriminating target and decoy precursors, or for detecting and removing interfering fragment ions [40, 59]. The combination of DIA and de novo sequencing is an ideal solution to simultaneously address both the low abundance and the mutations of neoantigens. Tran et al. reported nearly 2x more HLA-bound peptides by applying de novo sequencing on top of standard DIA database search [32]. However, DIA de novo sequencing is very challenging due to the highly multiplexed nature of DIA spectra, and currently deep learning remains the only solution to this problem.

MHC binding affinity and immunogenicity prediction

Once identified, candidate neoantigens are evaluated based on their predicted MHC binding affinity and immunogenicity, as these two properties determine the likelihoods that a neoantigen is presented on the cell surface and is recognized by T cells, respectively. Neural networks have been used for MHC binding prediction for nearly 20 years [8]. There is a great body of literature on this topic and readers can refer to comprehensive reviews [64, 65] and common tools such as NetMHC [8, 9], MHCflurry [42], MixMHCpred [66] for more details. NetMHC was one of the earliest and currently is still one of the most popular tools for MHC binding prediction. In this method, neural networks were applied to combine multiple encoding schemes of peptide sequences, including sparse encoding, Blosum encoding, and hidden Markov model encoding. The authors found that different encoding schemes provided complementary representations of amino acids, while multi-layer neural networks could capture high-order sequence correlations between amino acids. NetMHC is available as an online tool that takes input protein sequences and HLA alleles of interest, and then predicts a list of peptides, their binding affinities and their ranks compared to random peptides.

Initially, MHC binding prediction tools were used to predict candidate neoantigens from somatic mutations obtained from RNA-seq of the tumor [14, 15, 25], without mass spectrometry involved. Recent improvements of immunoassays and mass spectrometry for immunopeptidomics have enabled direct identification of HLA-bound peptides from the cell surface, both for cell lines and for native tumor tissues [16]. This has prompted a large number of studies that used mass spectrometry to improve MHC binding prediction [42, 43, 66–69]. Most of them focused on two approaches: (i) using endogenous (naturally presented) peptides identified by mass spectrometry to account for the whole antigen presentation pathway instead of just MHC binding and (ii) using pan-allele models to address the unbalanced and the limited availability of data for many HLA alleles. Notably, neural networks still remain as the core model in most of those studies.

Protein structure prediction

Computational protein structure prediction is a very challenging problem and many methods have been developed in the past decades. They can be broadly divided into two categories: template-based modeling (TBM) [70–75] and template-free modeling (FM) [46, 76–78]. As their names suggest, TBM predicts the structure of a protein by copying and refining the experimental structures of one or multiple similar proteins (called templates), while FM predicts the protein structure without explicitly copying from a whole template. Even just a few years ago, TBM was the most common method for protein structure prediction as evidenced by several popular tools [79, 80], whereas the accuracy of FM on most proteins was far from satisfactory [81]. Machine learning has been applied to protein structure prediction for a long time [82–85], but effective deep learning methods for FM have been developed only recently [13, 44, 46].

Inter-residue contact prediction for tertiary structure prediction

The revolution of protein structure prediction started with inter-residue contact prediction. Two residues are assumed to form a contact if they are spatially close to each other. In principle if a reasonable number of native contacts can be identified for a protein, its tertiary structure can be rebuilt with accurate topology [86]. Contact prediction has been challenging for a long time and its precision was very low before, e.g. 27% for top L/5 precision in CASP11[87] (the 11th Critical Assessment of Structure Prediction) by a meta-predictor [88] that integrates several global statistical methods through supervised machine learning where L is protein sequence length and 'top L/5 precision' denotes the percentage of correctly predicted contacts among the top L/5 predicted ones. Global statistical methods [89] shed some light on contact prediction for proteins with thousands of sequence homologs. However, they do not fare well on a large number of protein families without any experimental structures since these families do not have so many sequence homologs. Even for proteins with many sequence homologs, global statistical methods usually can only predict a small percentage of correct contacts, which alone may not be sufficient to yield high-resolution tertiary structure prediction [85, 90]. Supervised machine learning in principle shall outperform global statistical methods since the former integrates both sequence and structure information while the latter purely makes use of sequence information, but back then the advantage of supervised learning over global statistical methods is small if there is any especially when the protein under prediction has many sequence homologs [85, 88, 91].

Deep belief network (DBN) was attempted for contact prediction in 2012 [92, 93], but it performed similarly to traditional machine learning methods and thus drew little attention from the community. Only until 2016 in the RaptorX-Contact program [94], Xu's group demonstrated a major step forward in contact prediction and contact-based tertiary structure prediction by using a fully deep convolutional residual neural network (ResNet) [44], as shown in Figure 3a. By learning from thousands of experimentally solved protein structures, ResNet greatly reduced the number of sequence homologs needed for

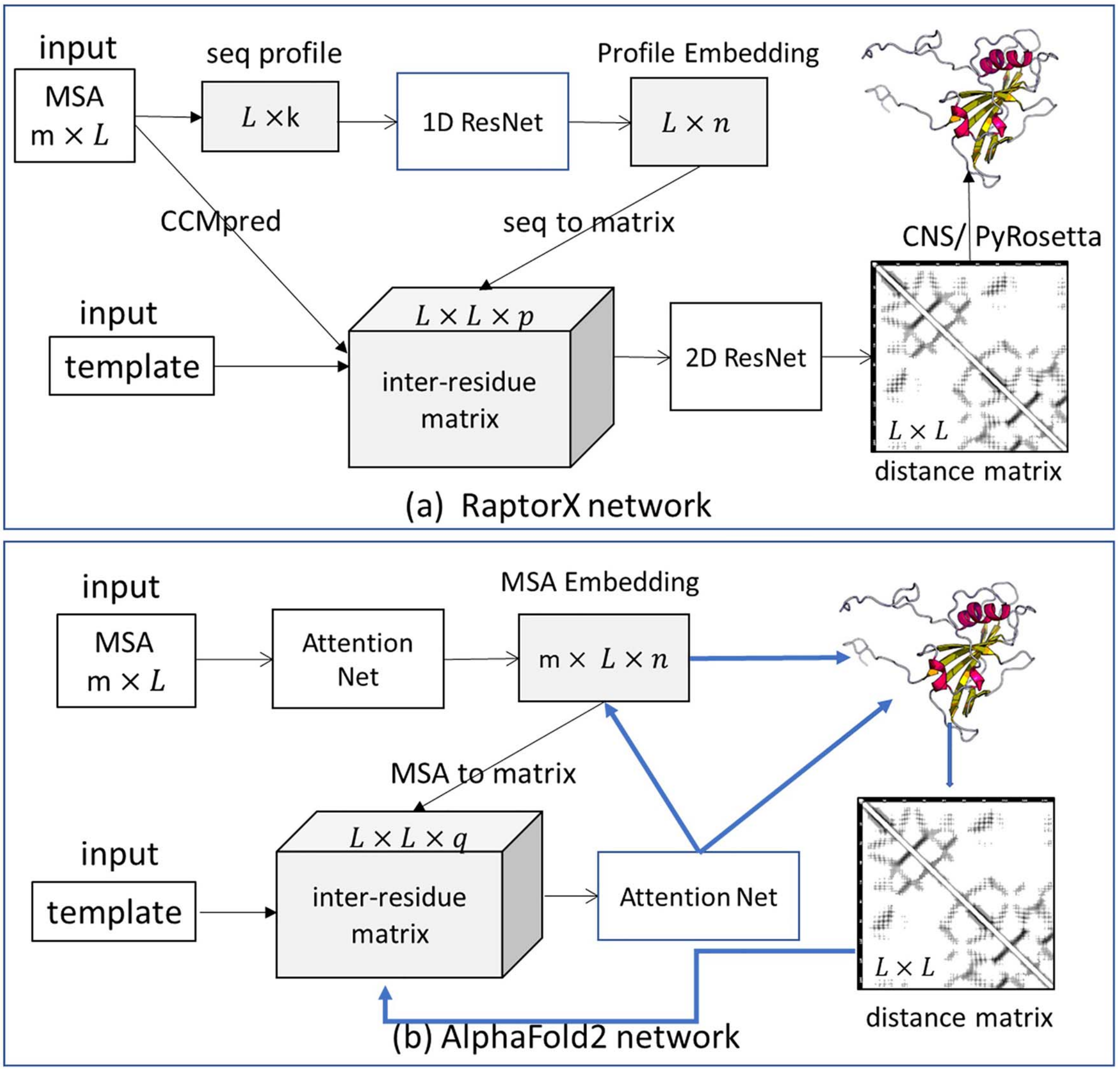

Figure 3. Deep network architectures of (a) RaptorX and (b) AlphaFold2 for protein structure prediction. The three blue arrows in (b) show important differences of AlphaFold2 from RaptorX and other methods.

satisfactory contact prediction, doubling or even tripling the precision over traditional methods on the CASP13 hard test proteins [45]. Recent studies have shown that ResNet was able to predict accurate contacts and correct folds for most proteins with more than 30 non-redundant sequence homologs [95, 96]. One of the major differences between DBN and RaptorX's ResNet is that the former predicts inter-residue contacts one by one while the latter predicts the whole contact matrix simultaneously. That is, ResNet predicts contacts by making use of protein global information while DBN does not. The residual learning module in ResNet is critical for the construction of a very deep neural network to capture protein global information.

Xu's group has shown that the contacts predicted by ResNet can be used to build tertiary structures of correct folds for many (large) proteins without good templates in PDB [44, 94]. In addition, this group has also shown that a ResNet model trained without any membrane proteins worked well on membrane protein structure prediction [97] and that a ResNet model trained on individual protein chains could be applied to protein complex contact prediction [98, 99]. These results suggest that ResNet predicts protein contact and tertiary structure not simply based upon sequence similarity and that protein complex structure prediction may also be addressed by a similar deep learning method.

Inter-residue distance prediction for tertiary structure prediction

Protein distance matrix encodes finer-grained information than contact matrix and thus may lead to more accurate tertiary structure prediction. Further, inter-residue distance is metric while contact is not, so a deep learning method trained by distance matrices can learn more physical constraints than a method trained by contact matrices. However, for a long time inter-residue distance prediction has not received as much attention as contact prediction, possibly because back then even contact prediction was already very challenging. In 2012, Xu's group applied shallow neural networks (≤5 layers) to predict inter-residue discrete distance distribution and then convert it to distance potential for protein decoy ranking [100], protein alignment [101] and conformation sampling [102]. After successfully applying deep ResNet to contact prediction, Xu extended his RaptorX-Contact program to discrete distance prediction and showed that ResNet-predicted distance might greatly improve both template-based [103] and template-free protein structure modeling [45, 46]. In CASP13, DeepMind's AlphaFold1 predicted on average the best protein 3D models for the hard targets using ResNet-predicted distance potential [13]. With accurate distance prediction, protein 3D models can be built within minutes or hours on a single Linux workstation or even a laptop computer using distance geometry or gradient descent optimization instead of time-consuming conformation sampling [46, 104]. The latest study shows that a better-engineered RaptorX may predict correct folds for 80% of the CASP13 hard test proteins [95]. Some studies have capitalized on ResNet's inherent ability to predict arbitrary inter-residue relationships, such as inter-residue orientation [104] and hydrogen bonds [105], which may help improve protein tertiary structure prediction. Other studies have also tried real-valued distance prediction [106–109], but it did not show better protein 3D models than discrete distance prediction.

Attention-based deep learning for protein structure prediction

Attention-based deep neural networks (e.g. transformer) were first applied to self-supervised learning of protein sequences [110, 111]. The sequence embedding or attention matrices produced by the self-supervised learning models can be used to deduce inter-residue contacts with accuracy close to that of the global statistical methods. In CASP14, DeepMind's AlphaFold2 achieved a very impressive result by using attention-based neural networks and supervised end-to-end training [112], predicting correct folds for more than 90% of the CASP14 hard test proteins. As shown in Figure 3b, AlphaFold2 used attention-based neural networks to model both multiple sequence alignments (MSAs) and inter-residue interaction matrix (e.g. contact/distance matrix). A few CASP14 groups have also added attention layers into ResNet [113] for distance prediction. In principle, attention-based neural networks are better than ResNet in capturing extra long-range inter-residue interactions and thus yield more accurate tertiary structure prediction. AlphaFold2 differed from RaptorX in that the inter-residue interaction matrix was also used to regenerate MSA embedding iteratively. To explicitly model the triangle inequality of distance (i.e. distance is metric), AlphaFold2 used a novel triangle attention mechanism to model the correlation of three residues.

A major innovation in AlphaFold2 is that instead of using predicted inter-residue distance matrix to build protein 3D structure, it employs an attention-based neural network (and ResNet) to directly predict (backbone and side chain) atom 3D coordinates from the protein sequence embedding and inferred inter-residue interaction matrix, which makes it possible to conduct an end-to-end training. By doing so, physical constraints of a protein structure can be better learned by deep learning models and the prediction error of a protein 3D model can be directly propagated back to the network input and thus, greatly improve prediction accuracy. End-to-end training was also explored by RGN [47], NEMO [114] and Jones' group [115], but AlphaFold2 first demonstrated that this strategy indeed worked well on protein structure prediction. AlphaFold2 has also integrated protein model refinement (called recycling in AlphaFold2) into its end-to-end pipeline. Such a recycling strategy takes the currently predicted structure model (and other information such as MSAs and templates) as input and produces an improved structure model. Finally, AlphaFold2 employs a self-distillation strategy to retrain its deep model by using a large number of well-predicted protein structures. Nevertheless, it needs a large amount of computing resources to implement and train a complete AlphaFold2 pipeline, which are not available for most protein structure prediction research groups. Very recently, Baker's group implemented a similar network as AlphaFold2, with accuracy better than current ResNet-based methods but still worse than AlphaFold2 [116].

Template-based modeling and integrating templates into deep learning

Template-based modeling (TBM) consists of two major steps: sequence-template alignment and 3D structure modeling from alignment. Lately, several deep learning models (e.g. ResNet) have been developed to substantially improve sequence-template alignment for remotely similar templates [74, 103, 117]. ResNet-predicted contact and distance have also been used to improve sequence-template alignment [74, 103, 117–119]. With very similar templates, traditional methods such as HHblits [73] and CNFpred [120] may already perform well on sequence-template alignment and thus deep learning is not essential for this step. A few years ago TBM was usually the first choice for protein structure prediction, but now FM may outperform TBM for many proteins unless very similar templates are available [74]. As such, improving TBM of a protein may not lead to substantial improvement in its final structure prediction since FM may produce a better prediction. However, when similar templates are available, integrating sequence-template alignment and template (backbone angle and distance matrix) information into deep networks may help improve 3D structure prediction, especially when the protein under prediction is large and does not have many sequence homologs. Such an idea was first implemented in the RaptorX-DeepModeller server in CASP13 [45, 74] and then adopted by a few other groups in CASP14 such as AlphaFold2 and Baker's group [121]. It was shown that with template information AlphaFold2 may improve by a good margin the structure prediction of some proteins with fewer than 100 sequence homologs [112].

Protein model refinement

Protein model refinement is the last step in protein structure prediction and popular methods, such as Feig's, are built upon molecular dynamics simulation [122]. Very recently, Baker's group applied ResNet to predict the local and global quality of a protein model and inter-residue distance error, and then used it to guide protein model refinement [48]. However, this method needs extensive conformation sampling, possibly because it does not fare well in predicting inter-residue distance information. Xu's group developed GNNRefine [123] that applied deep graph neural networks to predict inter-residue distance distribution from an initial model and then rebuilt the refined 3D models using the predicted distance information. GNNRefine achieved comparable accuracy as Baker's method, and both of them slightly underperformed Feig's method on the CASP14 test set. However, GNNRefine is two to three orders of magnitude faster than the other two methods, because it is able to predict inter-residue distance more accurately and thus does not need extensive protein conformation sampling to produce better refined models. Both Baker's method and GNNRefine may refine a protein model iteratively. That is, given an initial protein model GNNRefine outputs one or several refined models, which then can be fed into GNNRefine again for further refinement. The recycling module in AlphaFold2 can also be interpreted as a model refinement module. This recycling module is better than GNNRefine (and Baker's method) is that the former is an end-to-end system while GNNRefine is not. Another difference is that the recycling module directly outputs atom 3D coordinates while GNNRefine outputs inter-residue distance and then use PyRosetta [124] to generate atom 3D coordinates (which is also why GNNRefine is not an end-to-end system).

In summary, protein structure prediction has been revolutionized by deep learning including ResNet and attention-based networks. For a very good percentage of proteins AlphaFold2 may predict their structures with accuracy comparable to that of experimental techniques. In particular, AlphaFold2 may yield a confident prediction for 58% of human protein residues [125]. However, current successful structure prediction methods including AlphaFold2 still have some limitations. In particular, they fail on many orphan proteins, which do not have any sequence homologs [126]. They do not fare well on some very large, multi-domain proteins [125] and intrinsically disordered proteins either.

As mentioned before, a deep learning model trained on individual protein chains may be used to predict inter-protein contacts [98, 99]. However, complex contact and structure prediction is not fully resolved by deep learning although recently a few AlphaFold2-based methods are developed to predict complex structures [127, 128]. One major bottleneck with complex structure prediction is how to generate a reasonable number of complex sequence homologs (called interlogs) for inter-protein co-evolution information extraction. Xu's group has described two methods for generating interlogs, but they fail on some complexes [98]. It is worth mentioning that a number of good interlogs are not always needed for complex structure prediction because deep learning sometimes works well even without co-evolution information [95].

Accurate protein structure prediction has many applications, such as assisting experimental structure determination [129] and facilitating understanding life processes at atom level instead of residue level. Accurate protein structure is also very useful for drug discovery in several aspects [130] such as protein target selection for a specific disease, pocket selection, virtual screening [131] and determining the binding affinity of drug molecules [132].

Future perspective of deep learning for personalized biomedicine

In this article, we have reviewed major deep learning applications to address two open, challenging questions in protein science: neoantigen identification and protein structure prediction. Both topics have seen significant leaps forward just within the past five years, which immediately unlocked new developments of drugs and immunotherapies. In addition, the rapid adoption of deep learning has shifted the focus from algorithm-centric to data-centric. Many real-world applications, e.g. Google searches or Tesla autonomous systems, are now mainly relying on billion-parameter models trained on huge amounts of data. The same thing is also happening in biomedicine. More importantly, we believe that this approach shall progress towards the personalized trend of biomedicine, where personal models are trained on each individual patient's data to identify optimal treatments for the patient [21]. We have discussed such an example in a previous section where the genome and immunopeptidome of an individual patient can be used to train a personal model to predict neoantigens of that patient.

We envision that the trio MHC-neoantigen-TCR (TCR: T cell receptor) will play the central role of future personalized cancer immunotherapies, and deep learning methods for neoantigen and protein structure prediction will be a major driving force. A neoantigen needs to be presented on the cell surface by MHC proteins and recognized by T cells, thus it needs to bind to both MHC and TCR [133]. MHC and TCR are well known for their great diversity and specificity [134–136]. Neoantigens arise from somatic mutations that are also specific to each individual patient. Thus, the trio MHC-neoantigen-TCR are patient-specific and identifying them requires a personalized approach, such as deep learning on individual patient's data.

More importantly, most current prediction tools mainly focus on the amino acid sequences of neoantigens and do not take into account the structure information of MHC and TCR. Thus, with recent breakthroughs of deep learning methods for neoantigen and protein structure prediction, a fascinating future research direction is to develop personalized deep learning models to predict the structures of TCR, MHC, and neoantigens of an individual patient and precisely pinpoint which combinations of MHC-neoantigen-TCR are optimal to design a vaccine for that patient. This shall be one of the most quintessential achievements of personalized medicine and shall immensely benefit from the data-driven approach of deep learning.

Key Points

- A comprehensive review of deep learning advances in proteomics and immunopeptidomics to solve the problem of neoantigen prediction.
- A complete review of deep learning breakthroughs in protein structure prediction.
- A future perspective of personalized deep learning models for MHC-neoantigen-TCR binding prediction to design cancer vaccines for each individual patient.

Funding

The NSERC (grant OGP0046506 to M.L.); the Canada Research Chair programme (to M.L.); the NSFC (grant 61832019 to M.L.); the Leading innovative and Entrepreneur Team Introduction Program of Zhejiang (to M.L.); the NIH (grant R01GM089753 to J.X.).

References

1. A celebration of structural biology. *Nat Methods* 2021;**18**:427.
2. Deutsch EW, Csordas A, Sun Z, *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* 2017;**45**:D1100–6.
3. Kennedy D. 125. *Science* 2005;**309**:19.
4. The problem with neoantigen prediction. *Nat Biotechnol* 2017; **35**:97.
5. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;**11**:31–46.
6. Ma B, Zhang K, Hendrie C, *et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* 2003;**17**:2337–42.
7. Zhang J, Xin L, Shan B, *et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* 2012;**11**:M111.010587.
8. Nielsen M, Lundegaard C, Worning P, *et al.* Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 2003;**12**:1007–17.
9. Andreatta M, Nielsen M. Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 2016;**32**:511–7.
10. Sette A. The immune epitope database and analysis resource: from vision to blueprint. *Genome Inform* 2004;**15**:299.
11. Vita R, Mahajan S, Overton JA, *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**: D339–43.
12. Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* 2020;**588**: 203–4.
13. Senior AW, Evans R, Jumper J, *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10.
14. Ott PA, Hu Z, Keskin DB, *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 2017;**547**:217–21.
15. Sahin U, Derhovanessian E, Miller M, *et al.* Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 2017;**547**:222–6.
16. Bassani-Sternberg M, Bräunlein E, Klar R, *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* 2016;**7**:13404.
17. Tran NH, Qiao R, Xin L, *et al.* Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence* 2020;**2**:764–71.
18. Wilhelm M, Zolg DP, Graber M, *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat Commun* 2021;**12**:3346.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**: 436–44.
20. Wainberg M, Merico D, Delong A, *et al.* Deep learning in biomedicine. *Nat Biotechnol* 2018;**36**:829–38.
21. Esteva A, Robicquet A, Ramsundar B, *et al.* A guide to deep learning in healthcare. *Nat Med* 2019;**25**:24–9.
22. Eraslan G, Avsec Ž, Gagneur J, *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403.
23. Wen B, Zeng WF, Liao Y, *et al.* Deep learning in proteomics. *Proteomics* 2020;**20**:e1900335.
24. Stokes JM, Yang K, Swanson K, *et al.* A deep learning approach to antibiotic discovery. *Cell* 2020;**180**:688–702.e13.
25. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat Rev Immunol* 2018;**18**: 168–82.
26. Kolinski A, Skolnick J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* 1994;**18**: 338–52.
27. Xu J, Li M, Kim D, *et al.* RAPTOR: optimal protein threading by linear programming. *J Bioinform Comput Biol* 2003;**1**:95–117.
28. Dancík V, Addona TA, Clauser KR, *et al.* novo peptide sequencing via tandem mass spectrometry. *J Comput Biol* 1999;**6**: 327–42.
29. Li M, Vitányi P. *An Introduction to Kolmogorov Complexity and Its Applications*. Cham: Springer, 2019.
30. Tran NH, Rahman MZ, He L, *et al.* Complete de novo assembly of monoclonal antibody sequences. *Sci Rep* 2016;**6**:31730.
31. Tran NH, Zhang X, Xin L, *et al.* novo peptide sequencing by deep learning. *Proc Natl Acad Sci U S A* 2017;**114**:8247–52.
32. Tran NH, Qiao R, Xin L, *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat Methods* 2019;**16**:63–6.
33. Yang H, Chi H, Zeng W-F, *et al.* pNovo 3: precise de novo peptide sequencing using a learning-to-rank framework. *Bioinformatics* 2019;**35**:i183–90.

34. Qiao R, Tran NH, Xin L, *et al.* Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence* 2021;**3**:420–5.
35. Karunratanakul K, Tang H-Y, Speicher DW, *et al.* Uncovering thousands of new peptides with sequence-mask-search hybrid de novo peptide sequencing framework. *Mol Cell Proteomics* 2019;**18**:2478–91.
36. Zhou X-X, Zeng WF, Chi H, *et al.* pDeep: Predicting MS/MS spectra of peptides with deep learning. *Anal Chem* 2017;**89**: 12690–7.
37. Ma C, Ren Y, Yang J, *et al.* Improved peptide retention time prediction in liquid chromatography through deep learning. *Anal Chem* 2018;**90**:10881–8.
38. Gessulat S, Schmidt T, Zolg DP, *et al.* Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat Methods* 2019;**16**:509–18.
39. Yang Y, Liu X, Shen C, *et al.* In silico spectral libraries by deep learning facilitate data-independent acquisition proteomics. *Nat Commun* 2020;**11**:146.
40. Demichev V, Messner CB, Vernardis SI, *et al.* DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat Methods* 2020;**17**:41–4.
41. Larsen MV, Lundegaard C, Lamberth K, *et al.* Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* 2007;**8**:424.
42. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC Class I-presented peptides by incorporating antigen processing. *Cell Syst* 2020;**11**: 42–48.e7.
43. Bulik-Sullivan B, Busby J, Palmer CD, *et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat Biotechnol* 2018;**37**: 55–63. https://doi.org/10.1038/nbt.4313.
44. Wang S, Sun S, Li Z, *et al.* Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**:e1005324.
45. Xu J, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins* 2019;**87**:1069–81. https://doi.org/10.1101/624460.
46. Xu J. Distance-based protein folding powered by deep learning. *Proc Natl Acad Sci U S A* 2019;**116**:16856–65.
47. AlQuraishi M. End-to-end differentiable learning of protein structure. *Cell Syst* 2019;**8**:292–301.e3.
48. Hiranuma N, Park H, Baek M, *et al.* Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun* 2021;**12**:1340.
49. Frank A, Pevzner P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* 2005;**77**:964–73.
50. Chi H, Sun RX, Yang B, *et al.* pNovo: de novo peptide sequencing and identification using HCD spectra. *J Proteome Res* 2010;**9**: 2713–24.
51. Laumont CM, Vincent K, Hesnard L, *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci Transl Med* 2018;**10**:eaau5516.
52. Zhang Z. Prediction of low-energy collision-induced dissociation spectra of peptides with three or more charges. *Anal Chem* 2005;**77**:6364–73.
53. Wang Y, Yang F, Wu P, *et al.* OpenMS-Simulator: an open-source software for theoretical tandem mass spectrum prediction. *BMC Bioinformatics* 2015;**16**:110.
54. Arnold RJ, Jayasankar N, Aggarwal D, *et al.* A machine learning approach to predicting peptide fragmentation spectra. *Pac Symp Biocomput* 2006;219–30.
55. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008;**26**: 1367–72.
56. Käll L, Canterbury JD, Weston J, *et al.* Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* 2007;**4**:923–5.
57. Doerr A. DIA mass spectrometry. *Nat Methods* 2014;**12**:35–5.
58. Caron E, Espona L, Kowalewski DJ, *et al.* An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife* 2015;**4**:e07661.
59. Tsou C-C, *et al.* DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics. *Nat Methods* 2015;**12**:258–647 p following 264.
60. Pak H, Michaux J, Huber F, *et al.* Sensitive immunopeptidomics by leveraging available large-scale multi-hla spectral libraries, data-independent acquisition, and MS/MS prediction. *Mol Cell Proteomics* 2021;**20**:100080.
61. Bruderer R, Bernhardt OM, Gandhi T, *et al.* Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. *Mol Cell Proteomics* 2015;**14**: 1400–10.
62. Röst HL, Rosenberger G, Navarro P, *et al.* OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat Biotechnol* 2014;**32**:219–23.
63. MacLean B, Tomazela DM, Shulman N, *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 2010;**26**:966–8.
64. Mei S, Li F, Leier A, *et al.* A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2020;**21**:1119–35.
65. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. *PLoS Comput Biol* 2018;**14**:e1006457.
66. Bassani-Sternberg M, Chong C, Guillaume P, *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput Biol* 2017;**13**:e1005725.
67. Reynisson B, Alvarez B, Paul S, *et al.* NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res* 2020;**48**: W449–54.
68. Sarkizova S, Klaeger S, le PM, *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat Biotechnol* 2020;**38**:199–209.
69. Marty R, Kaabinejadian S, Rossell D, *et al.* MHC-I genotype restricts the oncogenic mutational landscape. *Cell* 2017;**171**: 1272–1283.e15.
70. Fiser A. Template-based protein structure modeling. *Methods Mol Biol* 2010;**673**:73–94.
71. Martí-Renom MA, Stuart AC, Fiser A, *et al.* Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 2000;**29**:291–325.
72. Cozzetto D, Kryshtafovych A, Fidelis K, *et al.* Evaluation of template-based models in CASP8 with standard measures. *Proteins* 2009;**77**(Suppl 9):18–28.
73. Remmert M, Biegert A, Hauser A, *et al.* HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2011;**9**:173–5.
74. Wu F, Xu J. Deep template-based protein structure prediction. *PLoS Comput Biol* 2021;**17**:e1008954.

75. Källberg M, Wang H, Wang S, *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat Protoc* 2012;**7**:1511–22.
76. Ben-David M, Noivirt-Brik O, Paz A, *et al.* Assessment of CASP8 structure predictions for template free targets. *Proteins* 2009;**77**(Suppl 9):50–65.
77. Rohl CA, Strauss CEM, Misura KMS, *et al.* Protein structure prediction using rosetta. *Methods Enzymol* 2004;**383**:66–93.
78. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* 2012;**80**:1715–35.
79. Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics* **Chapter 5**, 2006, Unit–5.6.
80. Waterhouse A, Bertoni M, Bienert S, *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* 2018;**46**:W296–303.
81. Abriata LA, Tamò GE, Monastyrskyy B, *et al.* Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins* 2018;**86**(Suppl 1):97–112.
82. Zhao F, Li S, Sterner BW, *et al.* Discriminative learning for protein conformation sampling. *Proteins* 2008;**73**:228–40.
83. Wang Z, Zhao F, Peng J, *et al.* Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* 2011;**11**: 3786–92.
84. Peng J, Xu J. Boosting protein threading accuracy. *Res Comput Mol Biol* 2009;**5541**:31–45.
85. Ma J, Wang S, Wang Z, *et al.* Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 2015;**31**:3506–13.
86. Kim DE, Dimaio F, Yu-Ruei Wang R, *et al.* One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins* 2014;**82**(Suppl 2):208–18.
87. Monastyrskyy B, 'Andrea D D, Fidelis K, *et al.* New encouraging developments in contact prediction: Assessment of the CASP11 results. *Proteins* 2016;**84**(Suppl 1):131–44.
88. Jones DT, Singh T, Kosciolek T, *et al.* MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 2015;**31**: 999–1006.
89. Weigt M, White RA, Szurmant H, *et al.* Identification of direct residue contacts in protein–protein interaction by message passing. *Proc Natl Acad Sci U S A* 2009;**106**:67–72.
90. Jones DT, Buchan DWA, Cozzetto D, *et al.* PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;**28**:184–90.
91. Wang Z, Xu J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* 2013;**29**:i266–73.
92. Eickholt J, Cheng J. Predicting protein residue-residue contacts using deep networks and boosting. *Bioinformatics* 2012;**28**: 3066–72.
93. Eickholt J, Cheng J. A study and benchmark of DNcon: a method for protein residue-residue contact prediction using deep networks. *BMC Bioinformatics 14 Suppl* 2013;**14**:S12.
94. Wang S, Sun S, Xu J. Analysis of deep learning methods for blind protein contact prediction in CASP12. *Proteins* 2018;**86**(Suppl 1): 67–77.
95. Xu J, McPartlon M, Li J. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nat Mach Intell* 2021;**3**:601–9.
96. Ju F, Zhu J, Shao B, *et al.* CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nat Commun* 2021;**12**:2535.
97. Wang S, Li Z, Yu Y, *et al.* Folding membrane proteins by deep transfer learning. *Cell Syst* 2017;**5**:202–211.e3.
98. Zeng H, Wang S, Zhou T, *et al.* ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res* 2018;**46**:W432–7.
99. Zhou T-M, Wang S, Xu J. Deep learning reveals many more inter-protein residue-residue contacts than direct coupling analysis-bioRxiv. 2018;240754. https://doi.org/10.1101/240754.
100. Zhao F, Xu J. A position-specific distance-dependent statistical potential for protein structure and functional study. *Structure* 2012;**20**:1118–26.
101. Ma J, Wang S, Wang Z, *et al.* MRFalign: protein homology detection through alignment of Markov random fields. *PLoS Comput Biol* 2014;**10**:e1003500.
102. Wang Z. Knowledge-based machine learning methods for macromolecular 3D structure prediction arXiv:1609.05061 [q-bio.BM]. 2016.
103. Zhu J, Wang S, Bu D, *et al.* Protein threading using residue co-variation and deep learning. *Bioinformatics* 2018;**34**: i263–73.
104. Yang J, Anishchenko I, Park H, *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc Natl Acad Sci U S A* 2020;**117**:1496–503.
105. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat Commun* 2019;**10**: 3977.
106. Li J, Xu J. Study of real-valued distance prediction for protein structure prediction with deep learning. *Bioinformatics* 2021;**37**: 3197–203.
107. Adhikari B. A fully open-source framework for deep learning protein real-valued distances. *Sci Rep* 2020;**10**:13374.
108. Ding W, Gong H. Predicting the real-valued inter-residue distances for proteins. *Adv Sci* 2020;**7**:2001314.
109. Wu T, Guo Z, Hou J, *et al.* DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. *BMC Bioinformatics* 2021;**22**:30.
110. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**: e2016239118.
111. Rao R, Meier J, Sercu T, *et al.* Transformer protein language models are unsupervised structure learners*bioRxiv* 2020.12.15.422761. 2020. https://doi.otg/10.1101/2020.12.15.422761.
112. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9. 10.1038/s41586-021-03819-2.
113. Shen T, Wu J, Lan H, *et al.* When homologous sequences meet structural decoys: Accurate contact prediction by tFold in CASP14-(tFold for CASP14 contact prediction). *Proteins* 2021;**1**:10. https://doi.otg/10.1002/prot.26232.
114. Ingraham J, Riesselman A, Sander C, *et al.* Learning protein structure with a differentiable simulator. *International Conference on Learning Representations*, 2018.
115. Kandathil SM, Greener JG, Lau AM, *et al.* Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments*bioRxiv* 2020.11.27.401232. 2020. https://doi.otg/10.1101/2020.11.27.401232.

116. Baek M, *et al.* Accurate prediction of protein structures and interactions using a 3-track network*bioRxiv* 2021.06.14.448402. 2021. https://doi.otg/10.1101/2021.06.14.448402.
117. Kong L, *et al.* ProALIGN: Directly learning alignments for protein structure prediction via exploiting context-specific alignment motifs*bioRxiv* 2020.12.28.424539. 2020. https://doi.otg/10.1101/2020.12.28.424539.
118. Du Z, Pan S, Wu Q, *et al.* CATHER: a novel threading algorithm with predicted contacts. *Bioinformatics* 2020;**36**:2119–25.
119. Zheng W, Zhang C, Wuyun Q, *et al.* LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res* 2019;**47**:W429–36.
120. Ma J, Peng J, Wang S, *et al.* A conditional neural fields model for protein threading. *Bioinformatics* 2012;**28**:i59–66.
121. Anishchenko I, Baek M, Park H, *et al.* Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins* 2021;**1**:12.
122. Feig M, Heo L. Protein structure refinement via molecular dynamics simulations. *Biophys J* 2018;**114**:575a.
123. Jing X, Xu J. Fast and effective protein model refinement by deep graph neural networks*bioRxiv* 2020.12.10.419994. 2020. https://doi.otg/10.1101/2020.12.10.419994.
124. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* 2010;**26**:689–91.
125. Tunyasuvunakool K, Adler J, Wu Z, *et al.* Highly accurate protein structure prediction for the human proteome. *Nature* 2021;**596**:590–6.
126. Chowdhury R, *et al.* Single-sequence protein structure prediction using language models from deep learning *bioRxiv* 2021.08.02.454840. 2021. https://doi.otg/10.1101/2021.08.02.454840.
127. Evans R, *et al.* Protein complex prediction with AlphaFold-Multimer*bioRxiv* 2021.10.04.463034. 2021. https://doi.otg/10.1101/2021.10.04.463034.
128. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2 and extended multiple-sequence alignments*bioRxiv* 2021.09.15.460468. 2021. https://doi.otg/10.1101/2021.09.15.460468.
129. Kryshtafovych A, Moult J, Albrecht R, *et al.* Computational models in the service of X-ray and cryo-electron microscopy structure determination. *Proteins* 2021;**1**:14. https://doi.otg/10.1002/prot.26223.
130. Mullard A. What does AlphaFold mean for drug discovery? *Nat Rev Drug Discov* 2021;**20**:725–7. https://doi.otg/10.1038/d41573-021-00161-0.
131. Rester U. From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Curr Opin Drug Discov Devel* 2008;**11**:559–68.
132. Shim J, Hong Z-Y, Sohn I, *et al.* Prediction of drug-target binding affinity using similarity-based convolutional neural network. *Sci Rep* 2021;**11**:4416.
133. Hennecke J, Wiley DC. T cell receptor-MHC interactions up close. *Cell* 2001;**104**:1–4.
134. Robins HS, Campregher PV, Srivastava SK, *et al.* Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 2009;**114**:4099–107.
135. Robins HS, *et al.* Overlap and effective size of the human CD8+ T cell receptor repertoire. *Sci Transl Med* 2010;**2**:47ra64.
136. Emerson RO, DeWitt WS, Vignali M, *et al.* Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet* 2017;**49**:659–65.