Application of machine learning on understanding biomolecule interactions in cellular machinery

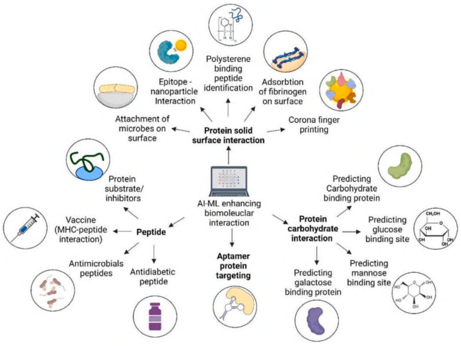
Rewati Dixit[a], Khushal Khambhati[b], Kolli Venkata Supraja[a], Vijai Singh[b], Franziska Lederer[c], Pau-Loke Show[d,e,f], Mukesh Kumar Awasthi[g], Abhinav Sharma[h], Rohan Jain[c,*]

[a] *Waste Treatment Laboratory, Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology Delhi, Haus-khas, New Delhi 110016, India*
[b] *Department of Biosciences, School of Science, Indrashil University, Rajpur, Mehsana 382715, Gujarat, India*
[c] *Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Bautzner landstrasse 400, 01328 Dresden, Germany*
[d] *Zhejiang Provincial Key Laboratory for Subtropical Water Environment and Marine Biological Resources Protection, Wenzhou University, Wenzhou 325035, China*
[e] *Department of Sustainable Engineering, Saveetha School of Engineering, SIMATS, Chennai 602105, India*
[f] *Department of Chemical and Environmental Engineering, University of Nottingham, Malaysia, 43500 Semenyih, Selangor Darul Ehsan, Malaysia*
[g] *College of Natural Resources and Environment, Northwest A&F University, Yangling 712100, China*
[h] *Institute Theory of Polymers, Leibniz Institute for Polymer Research, Hohe Strasse 6, 01069 Dresden, Germany*

HIGHLIGHTS

- ML applications in therapeutics can enhance drug delivery.
- Protein modifications through ML can improve their targeted bioactivity.
- Detection of protein-solid interactions through ML can aid in nanomedicine.
- ML can predict the interactions between carbohydrate and protein.

GRAPHICAL ABSTRACT


Polysterene binding peptide identification
Epitope - nanoparticle Interaction
Adsorbtion of fibrinogen on surface
Attachment of microbes on surface
Protein solid surface interaction
Corona finger printing
Protein substrate/ inhibitors
AI-ML enhancing biomoleuclar interaction
Predicting Carbohydrate binding protein
Vaccine (MHC-peptide interaction)
Peptide
Protein carbohydrate interaction
Predicting glucose binding site
Antimicrobials peptides
Antidiabetic peptide
Aptamer protein targeting
Predicting galactose binding protein
Predicting mannose binding site
CH₂OH
OH
HO

ABSTRACT

Machine learning (ML) applications have become ubiquitous in all fields of research including protein science and engineering. Apart from protein structure and mutation prediction, scientists are focusing on knowledge gaps with respect to the molecular mechanisms involved in protein binding and interactions with other components in the experimental setups or the human body. Researchers are working on several wet-lab techniques and generating data for a better understanding of concepts and mechanics involved. The information like biomolecular structure, binding affinities, structure fluctuations and movements are enormous which can be handled and analyzed by ML. Therefore, this review highlights the significance of ML in understanding the biomolecular interactions while assisting in various fields of research such as drug discovery, nanomedicine, nanotoxicity and material science. Hence, the way ahead would be to force hand-in hand of laboratory work and computational techniques.

[*] Corresponding author.
E-mail address: r.jain@hzdr.de (R. Jain).

1. Introduction

Studying proteins is an obligatory field of research since these macromolecules maintain and regulate all human bodily functions because of their distinct physicochemical properties. They are involved in innumerable inter- and intracellular interactions. They bind with ligands such as metals, organic molecules, inorganic molecules and other proteins for specific functions like catalysis or cell signaling (Dhakal et al., 2022; Rausell et al., 2010). These interactions happen at specific binding sites which have ignited a lot of interest amongst scientists in molecular modelling and drug design (Yang et al., 2013). When these interactions are to be predicted, a number of considerations have to be made such as free energy, enthalpy, entropy and binding kinetics (Du et al., 2016).

Protein interactions with other proteins, biomolecules, ligands or surfaces involve numerous forces acting simultaneously or one after the other. For instance, when a soluble antibody interacts with an antigen, it acts based on the lock and key model which is an amalgam of van der Waals forces, hydrogen bonding, hydrophobic interaction and steric hindrances (Leckband, 2000). Conventionally, knowledge of protein structures, interactions and functions was heavily based on structure determining techniques like X-ray crystallography, cryo-electron microscopy and biochemical assays to map functional consequences of altering protein structures (Ramanathan et al., 2021). Further, the complete genome sequencing projects have collected large data sets with respect to structure, function, genomic and biological context (Skrabanek et al., 2008). This information can be applied to investigate various protein–ligand interactions. Nevertheless, experimentation is time-consuming, expensive, and labor intensive (Skrabanek et al., 2008). Hence, molecular dynamics (MD) can be used as a computational tool to study biophysical processes (Wang et al., 2020). The major challenges with MD are where to store the huge amount of data and how to make it comprehensible. Conversely, machine learning (ML) can facilitate accurate approximation of complex relationships between variables and establish non-linear correlations (Ding et al., 2022; Huang et al., 2022; Xing et al., 2022).

Recently, artificial intelligence (AI) and ML have found groundbreaking applications in this area of research. For instance, geometric deep learning has been applied to predict interactions between the SARS-CoV-2 virus and human proteins with an accuracy of 97.76 % (Alakus and Turkoglu, 2021). Further, it has been shown that laccase and lignin compounds interact with each other by using hydrogen bond with the help of AI and ML programs (Wang et al., 2022a). Additionally, the most important step in drug discovery and design is protein folding and its interaction. Interactions based on protein binding domains can be predicted using a ML tool called hierarchical statistical mechanical modelling (Cunningham et al., 2020). Principal component analysis-ensemble extreme learning machine can give insight into the interactions by using sequence information with 87 % accuracy in significantly less time (You et al., 2013). Also, the structure-based threading regression tool is the first structure-based predictor to evaluate interaction probability (Singh et al., 2010). Computational techniques have not only been applied to study protein–protein interactions, but they can also be used to investigate hindrances in the interactions by using Naïve Bayesian, K-nearest neighbor, artificial neural network (ANN), decision tree (DT), random forest (RF) and support vector machine (SVM) (Gupta et al., 2021).

So far, the reviews focus on application of ML in protein structure prediction, protein–ligand interaction prediction, binding site prediction, or accelerating drug discovery (Dhakal et al., 2022; Y. Wang et al., 2022b; Zhao et al., 2020, 2022). Pandiyan and Wang (2022) summarized 136 papers elaborating the booming application of AI and ML in early detection and diagnosis followed by imaging and anti-cancer drug discovery. They concluded that it can support judicious utilization of resources and quality of cancer therapy (Pandiyan and Wang, 2022). However, none of the reviews incorporate how the proteins interact with other proteins and biomolecules in the body. Thus, this review focuses majorly on three aspects which are missing in the presented literature so far. The first aspect is how ML can be a vital tool in protein mutations to enhance their properties like an affinity towards a particular target or bioactivity. The next part of this review focusses on the application of ML in studying the interaction of proteins with solid surfaces and how it has been widely applied in nanomedicine. The final aspect of this review discusses the interactions between proteins and carbohydrates which is important for cell defense mechanisms.

2. ML in peptide modifications

Peptides are short chains of 2 to 50 amino acids connected via amide bonds and are naturally found in all living beings. Based on the amino acid sequence and the conformation they adopt; the peptides exhibit highly specific biological activities. Numerous functions of peptides as co-factors, activators, hormones, modulators, enzyme inhibitors and antimicrobials have been studied in the past (Hayashi et al., 2012). Apart from their diverse mode of action, due to their small size and less immunogenicity, they have been recognized as a valuable asset for human diagnosis and therapy (Apostolopoulos et al., 2021). For instance, insulin was debuted as a therapeutic peptide in 1922, and within the time span of 100 years, more than 80 functional peptides have reached the market for a wide range of diseases (Muttenthaler et al., 2021). Fig. S1 in supporting information summarizes key milestone achieved in field of therapeutic peptide from 1920 to 2020 (Muttenthaler et al., 2021).

Recently, a study has reported the use of peptides to inhibit SARS-CoV-2 (Chen et al., 2021a). Peptide driven technologies could also be advantageous in the pandemic situation considering the structural and functional versatility those peptides have to offer, along with their possible sequence combination and synthesis techniques. However, like any other drug discovery program, the development of lead peptides appears to be costly, challenging and a tedious task. In such a scenario, taking advantage of ever-growing peptide datasets, or generating a small amount of experimental data could help to run ML-based data-driven algorithms (Table 1). These algorithms assist to predict the evolution of the low bioactive peptides to peptides being highly active. It is evident that such an approach could resolve *de novo* design, directed evolution and property prediction problems (Fig. 1) (Chen et al., 2021a).

One such study has predicted highly active inhibitors of α-amylase and α-glucosidase using ML algorithms (Yamashita et al., 2020). α-amylase and α-glucosidase tend to elevate blood glucose levels after consuming meals. Their inhibitors are used in managing blood glucose levels and in the treatment of postprandial hyperglycemia. Through the peptide search method, the study conducted by Yamashita et al. (2020) used physiochemical properties (such as Isoelectric point, polarity, hydropathy index, side chain contribution to protein stability, molecular weight etc.) of amino acids as input feature for regression analysis, which assist in activating peptides with a large structural contribution. Yamashita et al. (2020) constructed 1-amino acid substitution library consisting of 153 peptides using “GHWYYRCW” as a design template and its inhibitory activity against α-amylase and α-glucosidase was experimentally determined. Regression analysis was conducted using 120 physiochemical properties of amino acids as input feature and the same from each peptide was related to enzyme inhibitory activity. The data from the 1-amino acid substitution peptide library was taken as training data set, whereas the 2- and 3- amino acid substitution peptide library was used as test data for the prediction of the highly inhibitory peptide. As per the report, all the predictions of 2-amino acid substitution peptides for high inhibition activity were in line with experimental assays. However, the same was not found for peptides with 3-amino acid substitution. Out of all the 3-amino acid substitution peptides predicted to have significant inhibitory activity, 86.7 % of them correlated with experimental data (Yamashita et al., 2020). Recently, ML methods has also been applied to identify short linear peptides as novel therapeutics

Table 1
AI-ML enhancing bioactivity of peptides.

Target	Peptide attribute	Input data	Training data	Output data	Highlight	Reference
Type 2 diabetes mellitus	Anti-Diabetic peptide for control of type 2 diabetes mellitus	Peptide sequences	Structured data from public databases*, unstructured data from scientific papers and patents	Predicts novel peptides glucose uptake efficacy	• 5 peptide tested for experimental validation that were distinct from human signaling peptides • Peptides predicted were less than 16 amino acid long • Candidate peptide were able to stimulate glucose transporter type 4 translocation and glucose uptake • Reduced glycated hemoglobin, significantly lower the plasma glucose levels and improve hepatic steatosis in obese insulin-resistant mice	Casey et al. (2021)
Class II MHC complex	Peptide vaccines	82 seed sequences having some affinity for HLA-DR401 87 seed sequences having some affinity for HLA-DR402 44 seed sequences having high affinity for HLA-DR402 and some affinity for HLA-DR401	Enrichment data from a library consisting of 108 random 9-mer peptides flanked by invariant peptide flanking residues	Optimize seed sequence having affinity either to HLA-DR401, HLA-DR402 or both MHC allele	• Evaluation and optimization of peptide-MHC binding • Optimize the anchor residues from Zika, HIV and Dengue proteomes • Yeast display assay demonstrated the improvement in the peptide binding by modulating anchor residues • 44 out of the 82 seed sequences performed better then seed sequence for HLA-DR401 • 72 out of the 87 seed sequences outperformed seed sequence for HLA-DR402 • The sequence optimized for both the allele had generally performed better for HLA-DR401 whereas perform the same HLA-DR402	Dai et al. (2021)
S. epidermidis	Antimicrobial peptides	Antimicrobial peptide form APD3	Positive training dataset containing 1,274 unique sequences**, negative training dataset contains 1,440 unique sequences**	Customized active peptides	• Study incorporates a transparent machine learning algorithm and rough set theory • Improved diversity generator and evolutionary search was also incorporated • Designing peptides against specific strains along with desired properties • Out of three peptide tested, one resulted positive for clear zone of inhibition	Boone et al. (2021)
α-amylase α-glucosidase	Enzyme inhibitors for managing blood glucose levels and postprandial hyperglycemia	2- and 3- amino acid substitution peptide library using GHWYYRCW as template	Inhibitory activity from 1-aminoacid substitution peptide library using GHWYYRCW as template. Regression analysis using enzyme inhibitory activity and 120 physicochemical properties of amino acid as input feature.	Prediction of inhibitory activity	• Prediction based on 2-amino substituting peptide library were 100% accurate • Prediction based on 3-amino substituting peptide library were 86.7% accurate	Yamashita et al. (2020)

(*continued on next page*)

Table 1 (*continued*)

Target	Peptide attribute	Input data	Training data	Output data	Highlight	Reference
E. coli	Antimicrobial peptide	Peptide library distantly related to Temporin-Ali	*In vitro* evaluation data to train a generalized liner model. Regression analysis using amino acid substitutions and IC_{50} values	Evolved peptide with greater antimicrobial activity	• Improve the antimicrobial properties of Temporin-Ali against *E. coli* • Identification of 44 peptides within 3 rounds of iteration having 160-fold more antimicrobial activity compared to wild type counter part • The most potent predicted peptides displayed IC_{50} (half-maximal inhibitory concentration) in the range of 0.50 to 2 μM • Selected peptides had a change in conformation from random coil to α-helical	Yoshida et al. (2018)
Phosphopantetheinyl transferase	Peptide substrate	Known/confirmed peptide sequences	Truncated portions of the ACP from that are substrate for Sfp, AcpS and AcpH peptide substrate for PPTases truncated peptide substrate that are inactive for Sfp, AcpS and AcpH	Short peptides substrate of 8-20 amino acid long	• A pipeline to optimize the peptide substrate for enzymes via prediction and targeted experimentation • Steadily increase in number of orthogonal peptide hit with each round of iteration	Tallorin et al. (2018)
E. coli *S. aureus*	Cationic antimicrobial peptides	1000 Randomly generated peptides	Data from approximately 100 peptides with their corresponding validated bioactivity	Peptides with high bioactivity	• Efficient algorithm based on graph theory • Tested peptide had MIC in range of 2-16 μg/mL	Giguere et al. (2015)

*Bioactivity annotations, biological pathways, and structural annotations.
**the data set were taken from the study published by Xiao et al. (2013).
MHC: Major histocompatibility complex.
APD3: Antimicrobial Peptide Database.
IC_{50}: half-maximal inhibitory concentration.
Sfp: Surfactin phosphopantetheinyl transferase from *B. subtilis*.
AcpS: Holo-acyl carrier protein synthase from *S. coelicolor*.
ACP: Acyl carrier Protein.
AcpH: Enzyme known to unlabel some substrates previously labeled by PPTase from *P. fluorescens*.

to combat Type 2 diabetes mellitus (T2DM) (Casey et al., 2021). Ensemble of neural network was used to build the predictive model. For training set, structured data from public database of bioactivity annotation, pathways and structural annotation was incorporated. Whereas data from peer-reviewed scientific papers and patents were referred to build unstructured database. Apart from that, predict–test–refine loop method was applied for additional refinement and testing. Out of 10^9 peptides, 100 were classified as active at the end of refinement process. Additionally, property filtering such as cell penetrability, toxicity, peptide length, stability in blood and odd number of cysteine residues were used to narrow down the peptides to be tested during the iterations that resulted ten of peptides from 100. As per the report, the algorithm was able to determine peptides that are distinct from human signaling peptides and are less than 16 amino acids in size. Furthermore, during *in vitro* studies, the same was also able to stimulate glucose transporter type 4 (GLUT4) translocation and glucose uptake (Casey et al., 2021).

Similarly, ML has also been used to optimize the affinity of peptides against class II Major Histocompatibility Complex (MHC). Class II MHC molecules present the antigenic peptides, which are contained in the peptide vaccines that further activate *T*-cells. *T*-cells are an important component of the immune system to combat pathogens and cancer (Dai et al., 2021). In the study conducted by Dai et al. (2021), all possible changes in the anchor residues of the pathogenic peptides were used to evaluate and optimize the peptide-MHC binding. Using an *in silico* objective function, scoring of peptide was done and the best was selected. The prediction from the PUFFIN (Prediction of Uncertainty in MHC-peptide aFFInity using residual Networks) peptide-MHC binding model was used for their objective function. PUFFIN uses deep residual network based computational approach that not only results affinity prediction of given peptide-MHC pair but also aids in quantifying uncertainty in peptide-MHC affinity prediction, resulting in state-of-the-art peptide-MHC binding prediction (Zeng and Gifford, 2019). Dai et al. (2021) optimize the anchor residues drawn from Zika, HIV and Dengue proteomes. A high-throughput yeast display assay was used to demonstrate the improvement in the peptide binding by modulating anchor residues (Dai et al., 2021).

ML can assist to screen the potential peptides for desire bioactivity by generating small amount of experimental data or using published datasets. Learning predictors could use this dataset and may lower the cost of expensive laboratory experiments. The work done by Giguere et al. (2015) focused on kernel method and ML to learn predictive model. Once a model is learned, the intense computational time is required to predict the peptides with soaring bioactivity. In order to overcome such an issue, Giguere et al (2015) have proposed an efficient algorithm that is based on graph theory. This graph theory-based algorithm proposed by them when combined with multi-target model enabled the user to predict the binding motif of the target with no prior knowledge of the ligand. As per the report, 100 and 1000 randomly generated peptides were used to find the peptides with high bioactivities. Increasing the number of the peptide from 100 to 1000 resulted predictions that are more beneficial on the bioactivity measurements. In total, the authors selected four peptides from the list of 1000 peptide candidates predicted to have high bioactivity. The selection was based on such a criteria that the peptide should at least differ by four amino acids to each other. These peptides were 15 amino acid long having 40 to 46 % similarity with their training dataset and minimal inhibitory concentration (MIC) in the range of 2 to 16 μg/mL against *Escherichia coli* and *Staphylococcus aureus* (Giguere et al., 2015). Similarly, novel antimicrobial peptides using transparent ML algorithm has been discovered that enabled better comprehension of design problems (Boone et al., 2021). The study incorporates a transparent ML algorithm and rough set theory (using datasets from Xiao et al., (2013)). This was combined with evolutionary search and an improved diversity generator, which in turn predicted peptides enabling ease of solid-state peptide synthesis while also maintaining their activity. Moreover, the proposed computer-aided molecular design approach facilitates designing of peptides against specific strains along with desired properties. Thus, the peptide with improved activities may help to address the concern of microbiome dysbiosis, immune system suppression and antimicrobial resistance simultaneously (Boone et al., 2021). The algorithm was applied to find antimicrobial peptides against *S. epidermidis* which is among the pathogens responsible for transplant infection. As per the report, three peptides were tested and only one resulted positive zone of inhibition (Boone et al., 2021). In another study, artificial evolutionary workflow was used to improve the chemical trait of antimicrobial peptides (Yoshida et al., 2018). A combination of genetic algorithm, ML and *in vitro* experimentation was applied with a closed loop approach. This artificial evolutionary workflow was challenged to improve the antimicrobial properties of Temporin-Ali against *E. coli.* Temporin-Ali is a naturally occurring 13 amino acid long antimicrobial peptide, which is known for its modest antimicrobial activity (Yoshida et al., 2018). Within three rounds of experiments, this approach enabled to identify 44 peptides having 160-fold more antimicrobial activity compared to its wild type counterpart. The most potent antimicrobial peptides were having IC_{50} (half-maximal inhibitory concentration) in the range of 0.50 to 2 μM. Furthermore, it was observed that the peptide selected had a change in conformation from random coil to α-helical during the experimentations (Yoshida et al., 2018). Thus, this approach not only gives an opportunity to study the evolution of a peptide but also allows us to quickly determine most potent functional molecules by conducting fewer sets of experiments.

Moreover, ML algorithms have also been used to discover *de novo* peptide substrates for enzymes (Tallorin et al., 2018). The same could be applied to protein labeling and protein purification. A methodology titled "Peptide Optimization with Optimal Learning" (POOL) helps to optimize the peptide substrate for enzymes via prediction and targeted experimentation. POOL has been applied to discover peptide substrates for phosphopantetheinyl transferase (PPTase) (Tallorin et al., 2018). POOL has also enabled peptide identification to meet certain criteria, such as, peptide substrate specifically for a particular class of PPTase. POOL combines a predictive model with information across enzymes, it uses Bayesian optimization to diversify selections against prediction uncertainty and comprise feedback iteratively. Thus, POOL could help to ease the complex biological problems that are faced by the conventional method used for peptide substrate discovery of post-translational modification enzymes (Tallorin et al., 2018). Table 1 summarizes the AI-ML programs used for discovery of peptide-based antimicrobials, antidiabetic, enzyme substrate, vaccines and enzyme inhibitors.

Computers have always been an asset for recognition and identification of complex patterns in text and images. The constant advancements in the era of omics technology improved biological database consisting of sequencing, biochemical and functional datasets. With the discovery of high-throughput techniques, ML has emerged as a valuable tool in *de novo* design for modification of functional peptide molecules. It is evident from the current literature that such an approach significantly eases pre-experimental screening and not only aids in improving the chemical trait but also gives an opportunity to study the evolution of improved functionality.

3. Solid surface and protein interaction

Interaction of proteins with a solid surface is a key phenomenon for implications in biomaterials, nanotechnology, and biotechnology (Gray, 2004). Protein adsorption is a first step for preparation of implant device whereas immobilization of enzymes is widely known for several bioprocesses like holocellulases immobilization on acrylic resins for bioethanol production (Gray, 2004; Vaz et al., 2016). Thus, an extensive knowledge about the underlying interactions would enable advancement of these fields. Consequently, the proteins as well as the surfaces could be tailored to produce desired affinity.

Fundamentally, the interaction involves both protein unfolding and binding (Gray, 2004). Their adsorption depends on external factors (pH, buffer composition and ionic strength), protein properties (composition, size, structure) and surface properties (polarity, charge and morphology) as shown in Fig. 2(a) (Rabe et al., 2011). Thus, performing experiments to study the mechanism requires extensive planning of resources, labor, money and advanced techniques. Computational techniques can overcome these bottlenecks. Hence, this section will summarize the applications of ML in protein and solid surface interaction.

3.1. ML in biomaterials

The immobilization of biomolecules on solid surfaces is an integral part of biological research. Important examples are enzyme immobilization in reactors and enzyme-linked immunosorbent assay (ELISA) among others. In principle, the target part of the molecule to be immobilized should be oriented appropriately and there should not be

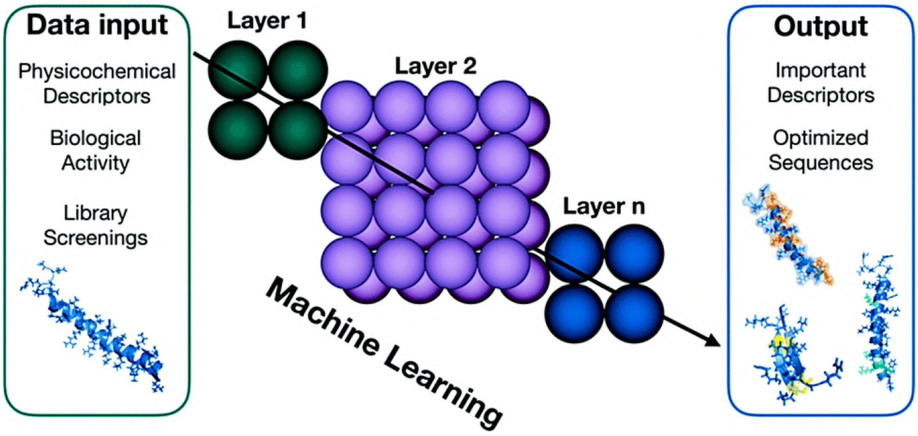


Fig. 1. Simplified schematic of machine learning used in enhancing peptide activity. The data input could include physicochemical descriptors, biological activity and library screening. The physicochemical properties such as molecular weight, net charge, hydrophobicity, hydrophobic moment, isoelectric point, aliphatic index etc. can be used as input features. The biological activity includes peptide attributes such as inhibition activity, antimicrobial activity, antidiabetic activity etc. of which optimization is desired. For training the model, this biological activity can be taken from the published literature and available database. An alternative is to perform an *in vitro* library screening and to generate a small amount of experimental dataset. Statistical methods are employed assisting computer systems to steadily learn and improve the performance to generate bioactive compounds form the input data. Figure adapted from Torres & de la Fuente-Nunez (2019) © Royal Society of Chemistry 2019.

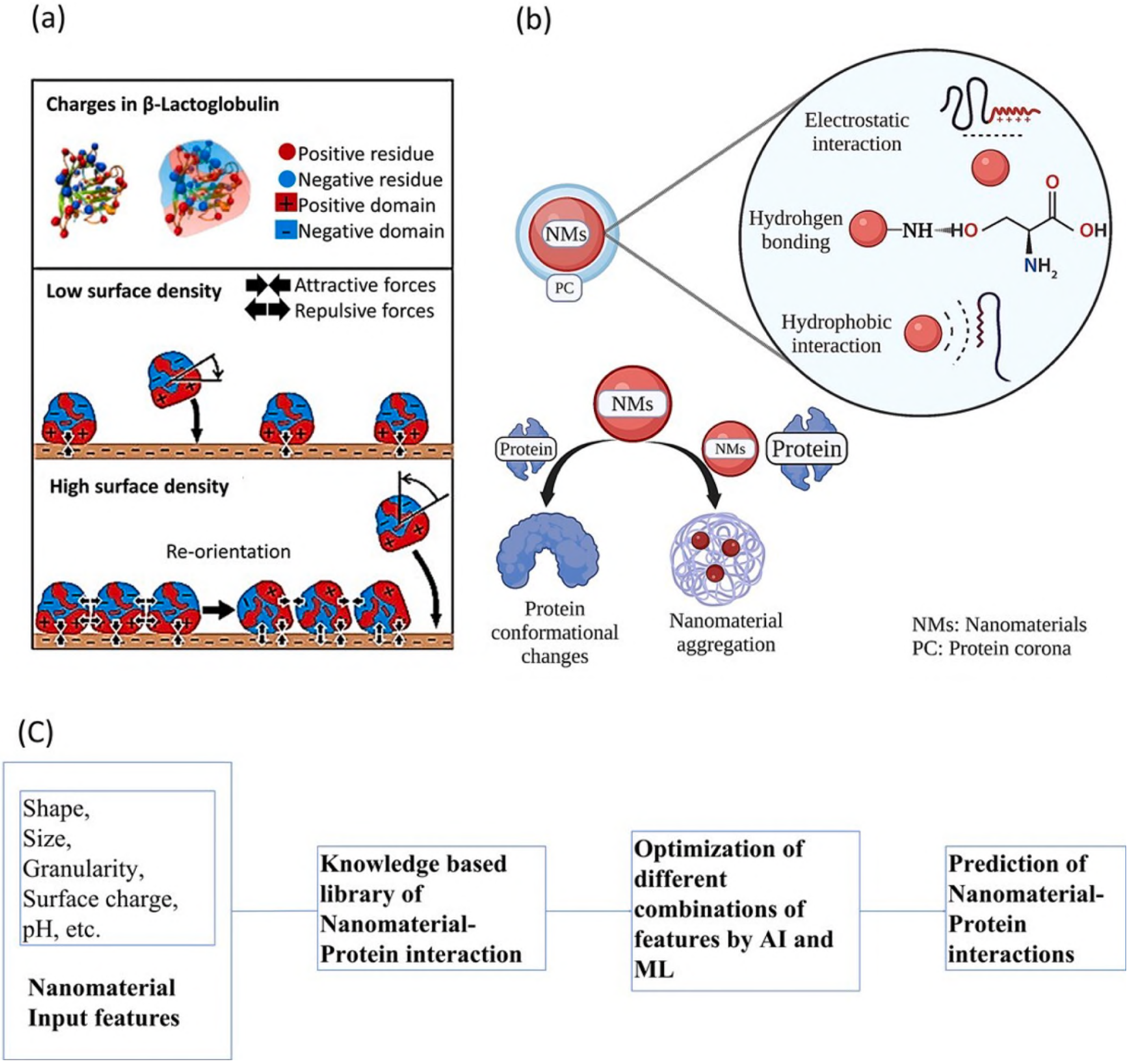


Fig. 2. (a) Reorientation of proteins while adsorbing on surfaces. Figure reproduced with permission from Rabe et al. (2011) Copyright © 2011 Elsevier. (b) Examples of interaction between nanomaterials and peptides (created with BioRender.com). (c) Flow-diagram showing the use of AI-ML in nanomaterial-protein interaction prediction.

any non-essential interaction between the biomolecule and solid surface (Meng et al., 2020). A commonly used protein binding surface is polystyrene due to its biological inertia (Kumada et al., 2010). The density and biological activity of the proteins immobilized on the solid surface is less when physically adsorbed on the surface. Thus, affinity peptide tags called polystyrene binding peptides (PSBPs) were developed to mitigate these issues. These PSBPs along with the target peptide can avoid denaturation and ensure proper orientation (Meng et al., 2020). Therefore, the identification of correct PSBPs is a key first step in such applications. Wet lab experiments would be an expensive and time taking option. This paves the way for the application of ML in biomolecule immobilization. Meng et al. (2020) have developed an identifier using ML-based algorithms that can identify if the protein or the peptide is a PSBP. Firstly, the amino acid composition (AAC) and dipeptide composition (DPC) is extracted from the peptide in question. The features are then ranked according to analysis of variance scores (ANOVA) followed by 123-dimensional optimal feature set selection using incremental feature selection (IFS) considering AAC as the criterion. Then the selected feature vectors are applied to the PSBP-SVM model to identify its role as a PSBP.

3.2. *ML in nanotechnology*

Recently, nanoparticles are being investigated as a viable alternative drug and drug delivery system for multi-drug resistant strains (MDR) (Diéguez-Santana et al., 2022). However, testing each combination of drugs with the bacterial strains is a strenuous process. ML can accelerate drug discovery against such bacteria. Diéguez-Santana et al. (2022) have developed an information fusion perturbation-theory-based machine learning (IFPTML) analysis to identify antibacterial drugs (AD) and nanoparticle (NP) combinations against resistant strains. Applying ML, they could quickly study if the interaction between the epitopes and the nanoparticle-based drug could provide for the required treatment effect. However, studying the interactions at the molecular level is complicated, given, the lack of high-quality data, effective modelling methods, descriptors, dynamic alterations and biotransformation of nanomaterials (Feng et al., 2021). Nevertheless, ML is interference-resistant and requires no *a priori* functional formulae. It can embody unlimited complexities in a model and filter out the factors that actually impact the immune response to develop immunotherapies involving nanomaterials (Feng et al., 2021). Wang et al. (2017) prepared a gold nanomaterial library, characterized them and collected data on their bioactivity. Each nanomaterial was simulated against diverse biological activities like interaction or passage through the biological membranes and models were built using computational intelligence. Consequently, these models were used to predict the nanomaterials that would be suitable for the desired bioactivity. Simulation results were consistent with the experimental results. Hence, with the increased flow of data today from experimental setups, ML will definitely be indispensable for nanomaterials-based drug discovery.

Scientists have also explored the protein adsorption properties of nanomaterials using ML and guided the formation of anti-protein coating surfaces (Le et al., 2019). Preventing biofouling is a vital step towards the development of bioinert surfaces for the advancement of nanomaterial-based therapy. Although, the adsorption of proteins on the solid surface is a common mechanism, the absence of mechanistic information makes it harder to design surfaces which can prevent biological contamination. Ideally, the surfaces should not have any charge, polar groups, hydrogen bond donors or acceptor groups. The authors demonstrated that robust linear and non-linear models can be developed to accurately predict the adsorption percentage of model proteins, that is, fibrinogen or lysozyme using sparse multiple linear regression with expectation maximization (MLREM) and non-linear Bayesian regularized artificial neural networks with Bayesian prior (BRANNGP) modelling methods. They achieved a test set with r^2 value as 0.82 and a standard error of 13 %. The authors were able to distinguish between low and high adsorption polyethylene glycols. Further details can be referred to in Le et al. (2019).

3.3. *ML in nanomedicine and nanotoxicology*

The nanomaterials are generally capped by proteins before introducing them into the biological system to prevent any undue cellular damage and participate in protein exchange (Nel et al., 2009). In comparison to the bulk materials, nanomaterials have higher free energy to adsorb proteins (Mukhopadhyay et al., 2018). This free energy is fundamental to analyzing the driving force for the interaction. The free energy could be enthalpic or entropic or both. Enthalpic changes are associated with hydrogen bonds or coordination bonds, whereas, entropic changes are due to dehydration or removal of electric double layer of the surface. In general, nanomaterials can either cause conformational changes in proteins or self-aggregate as shown in Fig. 2(b). Most of the adsorption studies have been performed *in vitro*, so how adsorption would differ *in vivo* is still an enigma. Apart from adsorption, there could be other modes of interactions like self-assembly, entrapment, cross-linking and encapsulation (Mukhopadhyay et al., 2018). Therefore, AI-ML could play a pivotal role in studying the nano-bio interface as shown in Fig. 2(c).

Modelling the surfaces for biological applications require good descriptors for designing and optimizing tailored functional materials. Current methods use complex mathematical features for characterization. Mikulskis et al. (2019) used ML methods based on chemically-interpretable descriptors (dragon and signature) to model the attachment of three important hospital-acquired pathogens, *Staphylococcus aureus* (SA), *Pseudomonas aeruginosa* (PA), uropathogenic *Escherichia coli* (UPEC). Sparse feature selection methods brought down the number of dragon and signature descriptors from around 1645 and 831 to 24 and 11, respectively. They obtained predictive models with small errors for the attachment of these pathogens against a polymer library. The r^2 values for SA, PA and UPEC were 0.85, 0.88 and 0.89 respectively. Further, Singh et al. (2021) have employed ML-based graph modelling to quantify the nanomaterial and cell interaction indices. They proposed that phenotypic changes in the cells such as shape and nuclear area factors are associated with nanomaterial characteristics (Singh et al., 2021). Moreover, Findlay et al. (2018) have developed a predictive system using ML to provide protein corona fingerprinting. This model predicts the corona population by analyzing protein biophysicochemical characteristics, solution conditions and nanomaterials properties (Findlay et al., 2018). Thus, ML can be used to gain mechanistic insights into protein and solid interactions and prepare maps for the same.

4. ML in protein-carbohydrate interactions

Protein-carbohydrate interactions are crucial for biological processes that are involved in cell development, immune responses, carcinogenesis and infections. They are also important in catalytic reactions and signaling pathways including cellular adhesion, recognition and transduction (Cao et al., 2021). A thorough molecular level understanding of these interactions can lead to advancements in the field of therapeutics. Glycoproteins and glycolipids on surfaces of living cells promote cell–cell communication and during any pathogen attack, they act as the primary line of defense (Majdoul and Compton, 2022). A class of proteins called lectins can specifically identify and bind to these cell surface carbohydrates (Chettri et al., 2021). Lectins are non-immune in nature and recognized as biomarkers helpful for drug targeting. Although these interactions help in developing novel drugs that are patient specific, experimentally detecting them is challenging. This is due to the difficulty in synthesizing the specific carbohydrate molecule. Moreover, weak binding affinity of carbohydrate to the protein limits their interactions (Gattani et al.,2019). To overcome these drawbacks, computational prediction through ML tools have come into operation. This approach helps to interpret the binding sites reducing complications of the experimental procedure. Particular methods for predicting specific carbohydrate (glucose, galactose, mannose) binding proteins are also important to get an in depth understanding of protein-carbohydrate interactions taking place in cell defense mechanisms (Zhao et al., 2018). For example, mannose binding proteins (MBP) play a key role in innate immune response by binding to pathogen surfaces containing mannose and activating lectin complement pathway (Agarwal et al., 2011).

4.1. *Structure based approaches for predicting protein-carbohydrate interactions*

The first method developed using ML algorithm was based on the characteristic properties of 19 non-homologous carbohydrate binding sites (Taroni et al., 2000). In this model, six parameters of amino acids (solvation potential, hydrophobicity, relative accessible surface area, residue propensity, planarity and protrusion index) were considered. When these parameters were ranked based on their binding site scores, residue interface propensity (tendency of amino acid to be in close proximity to the sugar molecule) suited best for discriminating the carbohydrate binding sites. The presence of multiple binding sites on a single protein chain can be distinguished using a combination of protrusion index and relative accessible surface area, as their distribution varies between lectins and enzymes. A combination of only these three attributes could show an overall prediction accuracy of 65 % on a group of 40 protein-carbohydrate complexes. However, this simple model could only act as a primary filter for identifying the possible carbohydrate binding sites and further detailed analysis with more data points is suggested. Based on the three-dimensional probability density distribution of interacting atoms, an ML algorithm was developed by Tsai et al. (2012) for predicting the carbohydrate binding sites. In this algorithm, distribution patterns specific for carbohydrate binding sites from known protein structures were used to identify the tentative carbohydrate binding sites on a query protein and integrated based on the normalized prediction confidence level. This model resulted into a specificity of 97 % with a sensitivity of 47 % predicted over a dataset of 108 proteins using ANN bagging algorithm.

ML algorithms have also been applied for more specific prediction of sugar binding proteins. Among them, a structure-based approach was developed using 18 protein-galactose complexes from 7 non-homologous protein families for determining galactose-binding sites with COTRAN program (Sujatha & Balaji, 2004). COTRAN is a C program based on 3D structure searching algorithm. This model incorporates a combination of solvent accessibility and three-dimensional structural characteristics to search for unknown galactose binding sites from known ones sharing the same fold. Furthermore, to predict the inositol and carbohydrate binding sites on protein surface, an ML tool called InCa-SiteFinder was developed and tested on 80 protein–ligand complexes (Kulharia et al., 2009). This model considered amino acid inclination behavior and van der Waals energy releasing from the interactions of protein and a methylene probe positioned at each point. These interactions form clusters and are ranked based on the spatial proximity of energetically favorable probe sites. This model was able to give 98 % specificity with a sensitivity and error rate of 73 % and 12 %, respectively. Another model for predicting glucose binding sites used RF selection coupled with SVM considering the physio-chemical properties such as hydrophobicity, charge and hydrogen bonding (Nassif et al., 2009). This classifier algorithm was able to give 93.3 % specificity with a sensitivity and error rate of 89.66 % and 8.11 %, respectively. An ML algorithm for predicting mannose binding sites was developed by Khare et al. (2012) that used ligand centroid approaches employing RF. With this algorithm, a prediction accuracy of 95 % was achieved with 10-fold cross validation. All the above mentioned are structure-based models that predict carbohydrate binding sites based on the availability of protein structures.

4.2. Sequence-based approaches for predicting protein-carbohydrate interactions

The first sequence-based approach for determining the carbohydrate binding proteins used structural features such as secondary structure conformation, solvent accessibility and packaging density as a preliminary step (Malik and Ahmad, 2007). In the next step, the corresponding binding sites were predicted based on position specific scoring matrix (PSSM) incorporating amino acid sequence profiles from their evolutionary origin collected from protein database (PDB). When evaluated for a data set of 40 protein-carbohydrate complexes, this method could predict a carbohydrate binding site with 87 % sensitivity at a specificity of 23 %. Mannose-interacting residues were predicted with PSSM as an input using the method of MOWGLI (prediction of protein-mannose interacting residues with ensemble classifiers using evolutionary information) while exploring RF and SVM as base classifiers (Pai and Mondal, 2016). This method could show a prediction accuracy of 92 % for a data set of 29 protein chains. In another study, mannose interacting and non-interacting sites were predicted using SVM based approach with a dataset of 120 protein chains (Agarwal et al., 2011). In this model, composition of peptide, segment or pattern is taken into consideration for determining the mannose binding sites and an accuracy of 86 % was achieved. The above-mentioned sequence-based methods rely on limited test datasets subjected to leave-one-out analysis and depend on sequence profiles taken only from PSSM.

A more accurate ML algorithm called SPRINT-CBH (sequence-based prediction of residual level interaction sites of carbohydrates) based on SVMs was developed to predict carbohydrate-binding sites (Taherzadeh et al., 2016). This method incorporates PSSM profiles with additional information on both the sequence and structural features beyond the evolutionary profiles. This model was able to give 99 % specificity with 18.8 % sensitivity when tested for a dataset of 102 complexes using a 10-fold cross validation (CV). However, when tested with an independent data set of 50 protein-carbohydrate complexes, this model was able to show a specificity of 98 % with 22.3 % sensitivity. Hence, it is found to yield imbalanced predictions with either a low sensitivity and high specificity or low specificity and high sensitivity. This might be due to selection of some features that affect the prediction sensitivity negatively. The secondary structure feature was eliminated by Gattani et al., (2019) as it was found to negatively affect the sensitivity. A stacking-based classifier called StackCBPred (carbohydrate binding site predictor accessible at: https://bmll.cs.uno.edu/) was built based on features extracted from PSSM. A more recent method of prediction has come forth considering effective parameters including the binding affinity and docking score. With this approach, a new ML tool cutoff scanning matrix (CSM)-carbohydrate was developed considering the biophysical data and structural features of 370 protein-carbohydrate complexes (Nguyen et al., 2022). Considering the drawbacks of other previous models, the CSM-carbohydrate model (accessible at: https://biosig.unimelb.edu.au/csm_carbohydrate/) is made user-friendly by allowing the data available in a web interface as well as an application programming interface (Nguyen et al., 2022). All the above-mentioned details have been summarized in Table 2.

5. ML for aptamer design and protein targeting

Aptamers are short sequences of 25–80 bases of oligonucleotides (DNA or RNA) that bind specifically to target molecules with high affinity. They are considered as a replacement for monoclonal antibodies in therapeutics and used as biorecognition elements in sensors and nanoscale devices (Ni et al., 2020). Designing an aptamer through experimental procedure with the method of systematic evolution of ligands by exponential enrichment (SELEX) is complex and time consuming with low reproducibility (Zhou and Rossi, 2017). Moreover, the specific aptamer characterization from an enriched pool of oligonucleotides is laborious. The half-life of an aptamer can be improved *in vivo* by mutations, substitutions or chemical modifications of its natural bases (Yang et al., 2022). Hence, focus has been shifted towards advanced and rapid screening of aptamers with AI using ML algorithms. Based on the information of target protein, AI can aid in designing the aptamer under *de novo* conditions following a stepwise procedure (Navien et al., 2021; Kohlberger and Gadermaier, 2021). In the first step, potential binding sites present on the targeted protein are considered and analyzed. This is followed by designing the specific sequence of ssDNA or RNA oligonucleotides that could show highest binding affinities with the targeted protein. The final step involves the automatic speculation of aptamer sequence from the predicted structure (Fig. 3). With this approach, time and cost required for aptamer design can be reduced while screening them from large sets of oligonucleotide libraries. Furthermore, appropriate mutations or modification of bases for specific targeting can also be done through ML tools (Bashir et al., 2021).

Some of the conventional methods in bioinformatics for designing an aptamer used structural information as a tool to predict the binding affinity. For determining the 2D and 3D structures of ssDNA and RNA molecules, online platforms like RNAComposer and RNAfold have been proposed (Chen et al., 2021b). These servers also provide structural information of short oligonucleotides and hence can be used for identifying the structures of specific aptamers. The novel computational methods for selection of an aptamer to predict the binding affinity rely on virtual screening and molecular docking scores (Buglak et al., 2020; Chen et al., 2021b). In the *in silico* methods for aptamer design, several steps are followed. Initially the 2D and 3D structures are predicted for studying the folding of desired aptamers and calculating the minimal amount of free energy generated. In the next step, docking scores are determined by calculating the binding energy through studying the interaction between target molecule and aptamer. Later, through molecular dynamics, the binding affinity is evaluated for the formed target molecule aptamer complex. The final step involves the aptamer/ligand complex analysis followed by data interpretation that can allow us to perform mutations or any required chemical modifications to the predicted aptamers.

Various ML algorithms for determining aptamer-protein binding affinities through structural and sequence-based clustering has already been described by Chen et al. (2021b). The latest approach for designing an aptamer employed extreme gradient boosting and RF classifiers for

Table 2
Sequence and structural based prediction models for determining carbohydrate protein interactions.

Prediction	Data sets for optimization	Parameters considered	Test datasets	Accuracy of the model	References
Structural characteristics					
Patch prediction algorithm for carbohydrate binding sites	19 non-homologous carbohydrate binding proteins	Relative accessible surface area, hydrophobicity, planarity, protrusion, residue propensity, solvation potential	40 protein-carbohydrate complexes	65 %	Taroni et al. (2000)
ANN_BAGGING algorithm	36 non-covalent interacting atoms	Three-dimensional probability density maps of non-covalent interacting atoms	108 proteins	Sensitivity of 49 % and specificity of 97 %	Tsai et al. (2012)
COTRAN (C computer program)	18 protein galactose complexes from 7 non-homologous families	Secondary structure type and solvent accessibility	—	—	Sujatha and Balaji (2004)
InCa-SiteFinder for carbohydrate binding sites	30,000 protein–ligand complexes	Van der Waals energy of interaction between protein and probe, amino acid propensities	40 carbohydrate binding sites	Sensitivity of 73 % and specificity of 98 % with error rate of 12 %	Kulharia et al. (2009)
Glucose binding site classifier program	29 protein glucose binding sites	Hydrophobicity, charge, hydrogen bonding and nature of amino acid side chains	14 protein glucose binding sites	Sensitivity of 89 % and specificity of 93 % with error rate of 8 %	Nassif et al. (2009)
Mannose binding sites	55 mannose binding sites derived from 11 proteins	Hydrophobicity, charge, hydrogen bonds, nature of amino acid side chains, accessible surface area	—	95 %	Khare et al. (2012)
Sequential characteristics					
Galactose binding proteins	20 residue types	Amino acid composition, solvent accessibility, packing density, accessible surface area, secondary structure	18 galactose specific proteins	63 % sensitivity and 79 % specificity	Malik and Ahmad (2007)
Ensemble classifier (MOWGLI) for mannose binding residues	128 residue types	Ensemble of base classifiers, evolutionary information	29 mannose binding proteins	92 %	Pai and Mondal (2016)
Web server (PreMieR) for mannose binding sites	120 residue types	Binary profile and local composition of patterns, evolutionary information	1029 mannose interacting residues	86 %	Agarwal et al. (2011)
SPRINT-CBH (sequence-based prediction of residual level interaction sites of carbohydrates)	113 protein carbohydrate complexes	Sequence information, evolutionary information, solvent accessible surface area, secondary structure, helix probability, steric parameter, polarizability, hydrophobicity, isoelectric point, volume, sheet probability	50 protein carbohydrate complexes	Sensitivity of 22 % and specificity of 98 %	Taherzadeh et al. (2016)
StackCBPred for prediction of carbohydrate binding proteins	100 carbohydrate binding proteins	Accessible surface area, secondary structure, polarizability, hydrophobicity, helix probability, volume, isoelectric point, sheet probability, molecular recognition features	49 protein carbohydrate complexes	Accuracy of 80 % and sensitivity of 70 %	Gattani et al. (2019)
Cutoff scanning matrix (CSM)-Carbohydrate algorithm	370 carbohydrate binding proteins	protein-carbohydrate interatomic interactions, graph-based signatures, molecular surface area of the interaction	43 protein carbohydrate complexes	—	Nguyen et al. (2022)

— missing data.

selecting the dominant features of the amino acids (Manju et al., 2022). 50 principal components were selected in this approach and a 98 % accuracy in detecting aptamer protein interactions was achieved. Another approach using RF was investigated by Emami and Ferdousi, (2021) that incorporated k-mer and complement k-mer frequency to predict aptamer protein interactions. A deep neural network tool called AptaNet was developed that used both interacting and non-interacting pairs of aptamer-protein. Although this model could provide an accuracy of 99 % on training dataset and 91 % on test dataset, a web server is required to further carry forward the research on aptamer-protein interactions. Recently, due to the outbreak of novel coronavirus, an ML algorithm was developed for screening the high affinity aptamers towards SARS-CoV-2 Receptor binding domain (RBD) (Song et al., 2020). During infection, an interaction of glycoprotein (S protein) of SARS-CoV-2 RBD with angiotensin-converting enzyme II (ACE2) expressed on the host cells was found. Through this study, two aptamers were recognized to have identical binding sites with the virus and new probes can be generated for recognizing the virus ultimately assisting the treatment of the infection (Song et al., 2020). Hence, ML algorithms provide a better opportunity to build unique and efficient aptamers that can bind to specific protein targets for usage in diagnostics, therapeutics and biosensing of disease-causing pathogens.

6. Research needs and future directions

The application of AI and ML has evolved to mimic human behavior and process huge data in short span of time due to its high computing capability and efficient algorithm. In this review, it has been shown that ML has been widely applied in drug discovery, nanomedicine, biomedical sciences and immunotherapy. Regression analysis and neural network based-ML algorithms appear to be an attractive option for modifying and predicting sequences for enhanced peptide functionality. Further, Quantitative structure–property relationship (QSPR) methods with ML and quantitative nanostructure activity relationship (QNAR) models are currently prominent in material science and nano-bio interactions. It enables delving into the details of the various molecular interactions. RF and SVM classifiers can be used for ML methods while predicting protein-carbohydrate interactions. Furthermore, ML algorithms rely on the datasets derived from the available database and they are as good as the training and test datasets provided. Thus, datasets from varying experimental conditions and from multiple sources could greatly increase the reliability of the predictions. Therefore, there is a need to harmonize and connect the data sets. Apart from this, another important challenge is the validation of models to be accepted by regulatory authorities. Using mechanistic information from peptide structures may help to guide the model better. Moreover, for it to be applied

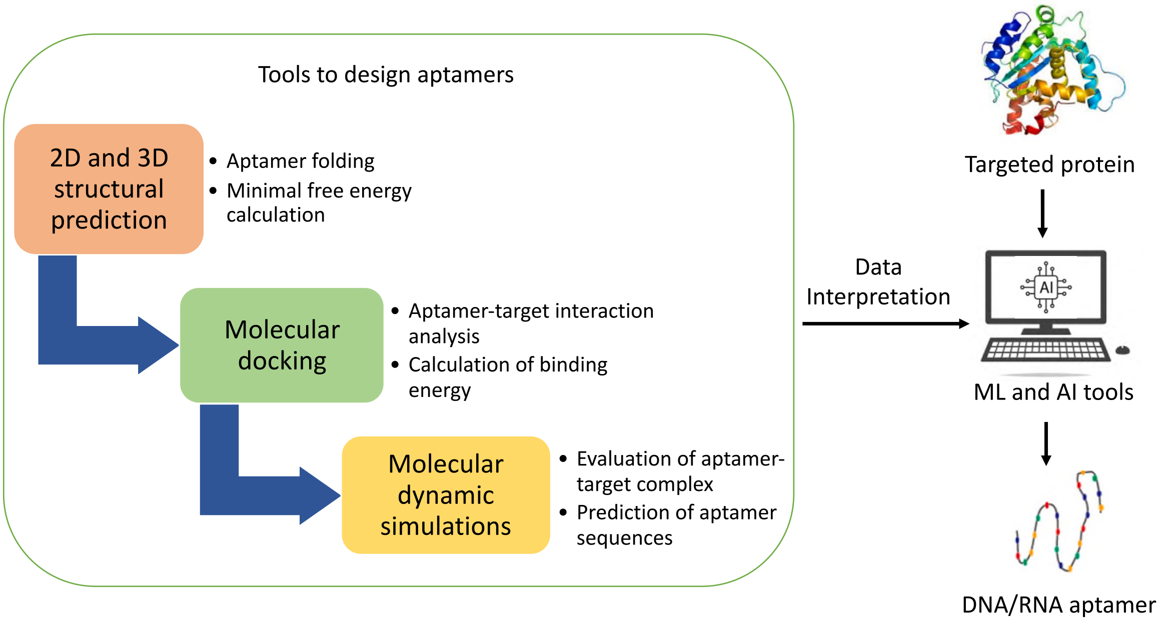

Tools to design aptamers
2D and 3D structural prediction
• Aptamer folding
• Minimal free energy calculation
Molecular docking
• Aptamer-target interaction analysis
• Calculation of binding energy
Molecular dynamic simulations
• Evaluation of aptamer-target complex
• Prediction of aptamer sequences
Data Interpretation
Targeted protein
AI
ML and AI tools
DNA/RNA aptamer

Fig. 3. *In-silico* methods for aptamer design using AI and ML.

in drug designing, it has to overcome five major challenges, namely, appropriate datasets, generating new hypothesis, optimize in multi-objective way, reduce the timelines involved and research culture. However, with the potential of generating high throughput structural and functional data and considering the advancements made in ML algorithms, it can be deduced that ML would play a key role in improving peptide functionality, especially in peptide-based therapeutic programs. Hence, modern ML and AI methods are useful tools in designing targeted and precision medicines.

7. Conclusion

This review highlights the various applications of ML which have propelled the progress of biology and biotechnology. Not only has it facilitated faster molecular docking, it has been employed to repurpose drugs against SARS-CoV-2 virus amidst the deadly pandemic. Researchers have also used ML to understand the molecular interactions for purposes such as studying protein-target binding, designing materials with desired characteristics for adsorption or to prepare biocompatible nanomaterials as therapeutics. Hence, applications are myriad but scientific development would be better if ML go hand-in-hand with wet-laboratory experiments.

Declaration of Competing Interest

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

Data availability

No data was used for the research described in the article.

Acknowledgements

Rewati Dixit and Kolli Venkata Supraja thank Ministry of Education, Government of India for the Prime Minister's research fellowship to support their doctoral research (PMRF ID: 1400791, 1401210, respectively). Khushal Khambhati acknowledges the Indian Council of Medical Research of the Government of India for financial assistance in the form of a Senior Research Fellowship (File No. AMR/Fellowship/36/2022-ECD-II).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.biortech.2022.128522.

References

Agarwal, S., Mishra, N.K., Singh, H., Raghava, G.P., 2011. Identification of mannose interacting residues using local composition. PLoS One 6 (9), e24039. https://doi.org/10.1371/journal.pone.0024039.

Alakus, T.B., Turkoglu, I., 2021. A novel protein mapping method for predicting the protein interactions in COVID-19 disease by deep learning. Interdiscip. Sci. Comput. Life Sci. 13, 44–60. https://doi.org/10.1007/s12539-020-00405-4.

Apostolopoulos, V., Bojarska, J., Chai, T.T., Elnagdy, S., Kaczmarek, K., Matsoukas, J., New, R., Parang, K., Lopez, O.P., Parhiz, H., Perera, C.O., 2021. A global review on short peptides: frontiers and perspectives. Molecules 26 (2), 430. https://doi.org/10.3390/molecules26020430.

Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., Ferguson, B.S., 2021. Machine learning guided aptamer refinement and discovery. Nat. Commun. 12 (1), 1–11. https://doi.org/10.1038/s41467-021-22555-9.

Boone, K., Wisdom, C., Camarda, K., Spencer, P., Tamerler, C., 2021. Combining genetic algorithm with machine learning strategies for designing potent antimicrobial peptides. BMC Bioinform. 22 (1), 1–17. https://doi.org/10.1186/s12859-021-04156-x.

Buglak, A.A., Samokhvalov, A.V., Zherdev, A.V., Dzantiev, B.B., 2020. Methods and applications of in silico aptamer design and modeling. Int. J. Mol. Sci. 21 (22), 8420. https://doi.org/10.3390/ijms21228420.

Cao, Y., Park, S.J., Im, W., 2021. A systematic analysis of protein–carbohydrate interactions in the Protein Data Bank. Glycobiol. 31 (2), 126–136. https://doi.org/10.1093/glycob/cwaa062.

Casey, R., Adelfio, A., Connolly, M., Wall, A., Holyer, I., Khaldi, N., 2021. Discovery through machine learning and preclinical validation of novel anti-diabetic peptides. Biomedicines. 9 (3), 276. https://doi.org/10.3390/biomedicines9030276.

Chen, Z., Hu, L., Zhang, B.T., Lu, A., Wang, Y., Yu, Y., Zhang, G., 2021b. Artificial intelligence in aptamer–target binding prediction. Int. J. Mol. Sci. 22 (7), 3605.

Chen, X., Li, C., Bernards, M.T., Shi, Y., Shao, Q., He, Y., 2021a. Sequence-based peptide identification, generation, and property prediction with deep learning: a review. Mol. Syst. Des. Eng. 6 (6), 406–428.

Chettri, D., Boro, M., Sarkar, L., Verma, A.K., 2021. Lectins: Biological significance to biotechnological application. Carbohydr. Res. 506, 108367. https://doi.org/10.1016/j.carres.2021.108367.

Cunningham, J.M., Koytiger, G., Sorger, P.K., AlQuraishi, M., 2020. Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. Nat. Methods 17, 175–183. https://doi.org/10.1038/s41592-019-0687-1.

Dai, Z., Huisman, B.D., Zeng, H., Carter, B., Jain, S., Birnbaum, M.E., Gifford, D.K., 2021. Machine learning optimization of peptides for presentation by class II MHCs.

Bioinformatics 37 (19), 3160–3167. https://doi.org/10.1093/bioinformatics/btab131.

Der Torossian Torres, M., De La Fuente-Nunez, C., 2019. Reprogramming biological peptides to combat infectious diseases. Chem. Commun. 55, 15020–15032. https://doi.org/10.1039/c9cc07898c.

Dhakal, A., McKay, C., Tanner, J.J., Cheng, J., 2022. Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions. Brief. Bioinform. 23, 1–23. https://doi.org/10.1093/bib/bbab476.

Diéguez-Santana, K., Rasulev, B., González-Díaz, H., 2022. Towards rational nanomaterial design by predicting drug-nanoparticle system interaction vs. bacterial metabolic networks. Environ. Sci. Nano. 9, 1391–1413. https://doi.org/10.1039/d1en00967b.

Ding, S., Huang, W., Xu, W., Wu, Y., Zhao, Y., Fang, P., Hu, B., Lou, L., 2022. Improving kitchen waste composting maturity by optimizing the processing parameters based on machine learning model. Bioresour. Technol. 360, 127606 https://doi.org/10.1016/j.biortech.2022.127606.

Du, X., Li, Y., Xia, Y.L., Ai, S.M., Liang, J., Sang, P., Ji, X.L., Liu, S.Q., 2016. Insights into protein–ligand interactions: mechanisms, models, and methods. Int. J. Mol. Sci. 17, 1–34. https://doi.org/10.3390/ijms17020144.

Emami, N., Ferdousi, R., 2021. AptaNet as a deep learning approach for aptamer–protein interaction prediction. Sci. Rep. 11 (1), 1–19. https://doi.org/10.1038/s41598-021-85629-0.

Feng, R., Yu, F., Xu, J., Hu, X., 2021. Knowledge gaps in immune response and immunotherapy involving nanomaterials: Databases and artificial intelligence for material design. Biomaterials. 266, 120469. https://doi.org/10.1016/j.biomaterials.2020.120469.

Findlay, M.R., Freitas, D.N., Mobed-Miremadi, M., Wheeler, K.E., 2018. Machine learning provides predictive analysis into silver nanoparticle protein corona formation from physicochemical properties. Environ. Sci. Nano 5, 64–71. https://doi.org/10.1039/c7en00466d.

Gattani, S., Mishra, A., Hoque, M.T., 2019. StackCBPred: a stacking based prediction of protein-carbohydrate binding sites from sequence. Carbohydr. Res. 486, 107857. https://doi.org/10.1016/j.carres.2019.107857.

Giguere, S., Laviolette, F., Marchand, M., Tremblay, D., Moineau, S., Liang, X., Biron, É., Corbeil, J., 2015. Machine learning assisted design of highly active peptides for drug discovery. PLoS Comput. Biol. 11 (4), e1004074 https://doi.org/10.1371/journal.pcbi.1004074.

Gray, J.J., 2004. The interaction of proteins with solid surfaces. Curr. Opin. Struct. Biol. 14, 110–115. https://doi.org/10.1016/j.sbi.2003.12.001.

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R.K., Kumar, P., 2021. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol. Divers. 25 (3), 1315–1360. https://doi.org/10.1007/s11030-021-10217-3.

Hayashi, M.A., Ducancel, F., Konno, K., 2012. Natural peptides with potential applications in drug development, diagnosis, and/or biotechnology. Int. J. Pept. 2012, 757838 https://doi.org/10.1155/2012/757838.

Huang, N., Gao, K., Yang, W., Pang, H., Yang, G., Wu, J., Zhang, S., Chen, C., Long, L., 2022. Assessing sediment organic pollution via machine learning models and resource performance. Bioresour. Technol. 361, 127710 https://doi.org/10.1016/j.biortech.2022.127710.

Khare, H., Ratnaparkhi, V., Chavan, S., Jayraman, V., 2012. Prediction of protein-mannose binding sites using random forest. Bioinformation 8 (24), 1202. https://doi.org/10.6026/97320630081202.

Kohlberger, M., Gadermaier, G., 2021. SELEX: Critical factors and optimization strategies for successful aptamer selection. Biotechnol. Appl. Biochem. https://doi.org/10.1002/bab.2244.

Kulharia, M., Bridgett, S.J., Goody, R.S., Jackson, R.M., 2009. InCa-SiteFinder: a method for structure-based prediction of inositol and carbohydrate binding sites on proteins. J. Mol. Graph. Model. 28 (3), 297–303. https://doi.org/10.1016/j.jmgm.2009.08.009.

Kumada, Y., Kuroki, D., Yasui, H., Ohse, T., Kishimoto, M., 2010. Characterization of polystyrene-binding peptides (PS-tags) for site-specific immobilization of proteins. J. Biosci. Bioeng. 109, 583–587. https://doi.org/10.1016/j.jbiosc.2009.11.005.

Le, T.C., Penna, M., Winkler, D.A., Yarovsky, I., 2019. Quantitative design rules for protein-resistant surface coatings using machine learning. Sci. Rep. 9 (1), 1–12. https://doi.org/10.1038/s41598-018-36597-5.

Leckband, D., 2000. Measuring the forces that control protein interactions. Annu. Rev. Biophys. Biomol. 29, 1–26. https://doi.org/10.1146/annurev.biophys.29.1.1.

Majdoul, S., Compton, A.A., 2022. Lessons in self-defence: inhibition of virus entry by intrinsic immunity. Nat. Rev. Immunol. 22 (6), 339–352. https://doi.org/10.1038/s41577-021-00626-8.

Malik, A., Ahmad, S., 2007. Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Struct. Biol. 7 (1), 1–14. https://doi.org/10.1186/1472-6807-7-1.

Manju, N., Samiha, C.M., Kumar, S.P., Gururaj, H.L., Flammini, F., 2022. Prediction of aptamer protein interaction using random forest algorithm. IEEE Access 10, 49677–49687. https://doi.org/10.1109/ACCESS.2022.3172278.

Meng, C., Hu, Y., Zhang, Y., Guo, F., 2020. PSBP-SVM: a machine learning-based computational identifier for predicting polystyrene binding peptides. Front. Bioeng. Biotechnol. 8, 1–9. https://doi.org/10.3389/fbioe.2020.00245.

Mikulskis, P., Alexander, M.R., Winkler, D.A., 2019. Toward Interpretable Machine Learning Models for Materials Discovery. Adv. Intell. Syst. 1, 1900045. https://doi.org/10.1002/aisy.201900045.

Mukhopadhyay, A., Basu, S., Singha, S., Patra, H.K., 2018. Inner-view of nanomaterial incited protein conformational changes: Insights into designable interaction. Research (Wash D C). 2018, 9712832.

Muttenthaler, M., King, G.F., Adams, D.J., Alewood, P.F., 2021. Trends in peptide drug discovery. Nat. Rev. Drug Discov. 20, 309–325. https://doi.org/10.1038/s41573-020-00135-8.

Nassif, H., Al-Ali, H., Khuri, S., Keirouz, W., 2009. Prediction of protein-glucose binding sites using support vector machines. Proteins Struct. Funct. Genet. 77 (1), 121–132. https://doi.org/10.1002/prot.22424.

Navien, T.N., Thevendran, R., Hamdani, H.Y., Tang, T.H., Citartan, M., 2021. In silico molecular docking in DNA aptamer development. Biochimie 180, 54–67. https://doi.org/10.1016/j.biochi.2020.10.005.

Nel, A.E., Mädler, L., Velegol, D., Xia, T., Hoek, E.M.V., Somasundaran, P., Klaessig, F., Castranova, V., Thompson, M., 2009. Understanding biophysicochemical interactions at the nano-bio interface. Nat. Mater. 8, 543–557. https://doi.org/10.1038/nmat2442.

Nguyen, T.B., Pires, D.E., Ascher, D.B., 2022. CSM-carbohydrate: protein-carbohydrate binding affinity prediction and docking scoring function. Brief. Bioinform. 23 (1), 512. https://doi.org/10.1093/bib/bbab512.

Ni, S., Zhuo, Z., Pan, Y., Yu, Y., Li, F., Liu, J., Zhang, G., 2020. Recent progress in aptamer discoveries and modifications for therapeutic applications. ACS Appl. Mater. Interfaces 13 (8), 9500–9519. https://doi.org/10.1021/acsami.0c05750.

Pai, P.P., Mondal, S., 2016. MOWGLI: prediction of protein–Mannose interacting residues with ensemble classifiers using evolutionary information. J. Biomol. Struct. Dyn. 34 (10), 2069–2083. https://doi.org/10.1080/07391102.2015.1106978.

Pandiyan, S., Wang, L., 2022. A comprehensive review on recent approaches for cancer drug discovery associated with artificial intelligence. Comput. Biol. Med. 150, 106140 https://doi.org/10.1016/j.compbiomed.2022.106140.

Rabe, M., Verdes, D., Seeger, S., 2011. Understanding protein adsorption phenomena at solid surfaces. Adv. Colloid Interface Sci. 162, 87–106. https://doi.org/10.1016/j.cis.2010.12.007.

Ramanathan, A., Ma, H., Parvatikar, A., Chennubhotla, S.C., 2021. Artificial intelligence techniques for integrative structural biology of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 66, 216–224. https://doi.org/10.1016/j.sbi.2020.12.001.

Rausell, A., Juan, D., Pazos, F., Valencia, A., 2010. Protein interactions and ligand binding: From protein subfamilies to functional specificity. PNAS 107, 1995–2000. https://doi.org/10.1073/pnas.0908044107.

Singh, R., Park, D., Xu, J., Hosur, R., Berger, B., 2010. Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. Nucleic acids Res. W508–W515. https://doi.org/10.1093/nar/gkq481.

Singh, A.V., Maharjan, R.S., Kanase, A., Siewert, K., Rosenkranz, D., Singh, R., Laux, P., Luch, A., 2021. Machine-Learning-Based Approach to Decode the Influence of Nanomaterial Properties on Their Interaction with Cells. ACS Appl. Mater. Interfaces. 13, 1943–1955. https://doi.org/10.1021/acsami.0c18470.

Skrabanek, L., Saini, H.K., Bader, G.D., Enright, A.J., 2008. Computational prediction of protein-protein interactions. Mol. Biotechnol. 38, 1–17. https://doi.org/10.1007/s12033-007-0069-2.

Song, Y., Song, J., Wei, X., Huang, M., Sun, M., Zhu, L., Yang, C., 2020. Discovery of aptamers targeting the receptor-binding domain of the SARS-CoV-2 spike glycoprotein. J. Anal. Chem. 92 (14), 9895–9900. https://doi.org/10.1021/acs.analchem.0c01394.

Sujatha, M.S., Balaji, P.V., 2004. Identification of common structural features of binding sites in galactose-specific proteins. Proteins Struct. Funct. Genet. 55 (1), 44–65. https://doi.org/10.1002/prot.10612.

Taherzadeh, G., Zhou, Y., Liew, A.W.C., Yang, Y., 2016. Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. J. Chem. Inf. Model. 56 (10), 2115–2122. https://doi.org/10.1021/acs.jcim.6b00320.

Tallorin, L., Wang, J., Kim, W.E., Sahu, S., Kosa, N.M., Yang, P., Thompson, M., Gilson, M.K., Frazier, P.I., Burkart, M.D., Gianneschi, N.C., 2018. Discovering de novo peptide substrates for enzymes using machine learning. Nat. Commun. 9 (1), 1–10. https://doi.org/10.1038/s41467-018-07717-6.

Taroni, C., Jones, S., Thornton, J.M., 2000. Analysis and prediction of carbohydrate binding sites. Protein Eng. 13 (2), 89–98. https://doi.org/10.1093/protein/13.2.89.

Torres, M.D.T., de la Fuente-Nunez, C., 2019. Reprogramming biological peptides to combat infectious diseases. Chem. Comm. 55 (100), 15020–15032. https://doi.org/10.1039/C9CC07898C.

Tsai, K.C., Jian, J.W., Yang, E.W., Hsu, P.C., Peng, H.P., Chen, C.T., Yang, A.S., 2012. Prediction of carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. PLoS One 7 (7), e40846. https://doi.org/10.1371/journal.pone.0040846.

Vaz, R.P., de Souza Moreira, L.R., Ferreira Filho, E.X., 2016. An overview of holocellulose-degrading enzyme immobilization for use in bioethanol production. J. Mol. Catal. B Enzym. 133, 127–135. https://doi.org/10.1016/j.molcatb.2016.08.006.

Wang, Y., Lamim Ribeiro, J.M., Tiwary, P., 2020. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. Curr. Opin. Struct. Biol. 61, 139–145. https://doi.org/10.1016/j.sbi.2019.12.016.

Wang, W., Sedykh, A., Sun, H., Zhao, L., Russo, D.P., Zhou, H., Yan, B., Zhu, H., 2017. Predicting Nano-Bio Interactions by Integrating Nanoparticle Libraries and Quantitative Nanostructure Activity Relationship Modeling. ACS Nano 11, 12641–12649. https://doi.org/10.1021/acsnano.7b07093.

Wang, Y., Wu, S., Duan, Y., Huang, Y., 2022b. A point cloud-based deep learning strategy for protein-ligand binding affinity prediction. Brief. Bioinform. 23, 1–11. https://doi.org/10.1093/bib/bbab474.

Wang, L., Xue, C., Owens, G., Chen, Z., 2022a. Artificial intelligence modeling and molecular docking to analyze the laccase delignification process of rice straw by Comamonas testosteroni FJ17. Bioresour. Technol. 345, 126565 https://doi.org/10.1016/j.biortech.2021.126565.

Xiao, X., Wang, P., Lin, W.Z., Jia, J.H., Chou, K.C., 2013. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. 436 (2), 168–177. https://doi.org/10.1016/j.ab.2013.01.019.

Xing, J., Kurose, R., Luo, K., Fan, J., 2022. Chemistry-Informed Neural Networks modelling of lignocellulosic biomass pyrolysis. Bioresour. Technol. 355, 127275 https://doi.org/10.1016/j.biortech.2022.127275.

Yamashita, H., Fujitani, M., Shimizu, K., Kanie, K., Kato, R., Honda, H., 2020. Machine learning-based amino acid substitution of short peptides: acquisition of peptides with enhanced inhibitory activities against α-amylase and α-glucosidase. ACS Biomater Sci. Eng. 6 (11), 6117–6125. https://doi.org/10.1021/acsbiomaterials.0c01010.

Yang, C., Jiang, Y., Hao, S.H., Yan, X.Y., Naranmandura, H., 2022. Aptamers: An emerging navigation tool of therapeutic agents for targeted cancer therapy. J. Mater. Chem. B. https://doi.org/10.1039/D1TB02098F.

Yang, J., Roy, A., Zhang, Y., 2013. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29, 2588–2595. https://doi.org/10.1093/bioinformatics/btt447.

Yoshida, M., Hinkley, T., Tsuda, S., Abul-Haija, Y.M., McBurney, R.T., Kulikov, V., Mathieson, J.S., Reyes, S.G., Castro, M.D., Cronin, L., 2018. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. Chem 4 (3), 533–543. https://doi.org/10.1016/j.chempr.2018.01.005.

You, Z.H., Lei, Y.K., Zhu, L., Xia, J., Wang, B., 2013. May. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. In: BMC Bioinform., 14 BioMed Central, pp. 1–11. https://doi.org/10.1186/1471-2105-14-S8-S10.

Zeng, H., Gifford, D.K., 2019. Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide selection for therapeutic design. Cell Syst. 9 (2), 159–166. https://doi.org/10.1016/j.cels.2019.05.004.

Zhao, J., Cao, Y., Zhang, L., 2020. Exploring the computational methods for protein-ligand binding site prediction. Comput. Struct. Biotechnol. J. 18, 417–426. https://doi.org/10.1016/j.csbj.2020.02.008.

Zhao, H., Taherzadeh, G., Zhou, Y., Yang, Y., 2018. Computational prediction of carbohydrate-binding proteins and binding sites. Curr. Protoc. Protein Sci. 94 (1), 75. https://doi.org/10.1002/cpps.75.

Zhao, L., Zhu, Y., Wang, J., Wen, N., Wang, C., Cheng, L., 2022. A brief review of protein–ligand interaction prediction. Comput. Struct. Biotechnol. J. 20, 2831–2838. https://doi.org/10.1016/j.csbj.2022.06.004.

Zhou, J., Rossi, J., 2017. Aptamers as targeted therapeutics: current potential and challenges. Nat. Rev. Drug Discov. 16 (3), 181–202. https://doi.org/10.1038/nrd.2016.19.