immunology
insights into immunity in health and disease

ORIGINAL ARTICLE

Computer Integrated Dominant Epitopes Evoke Protective Immune Response Against *Streptococcus pneumoniae*

Hitesh Harsukhbhai Chandpa | Shovan Naskar | Jairam Meena

ImmunoEngineering and Therapeutics Laboratory, Department of Pharmaceutical Engineering and Technology, Indian Institute of Technology (Banaras Hindu University), Varanasi, India

Correspondence: Jairam Meena (jairam.phe@itbhu.ac.in)

Received: 11 July 2024 | **Revised:** 22 February 2025 | **Accepted:** 25 February 2025

Funding: This work was supported by the Science and Engineering Research Board (Now Known as ANRF) (EEQ/2023/000779), and the Indian Institute of Technology (Banaras Hindu University).

Keywords: immunoinformatics | multi-epitope vaccine | pneumonia | *Streptococcus pneumoniae*

ABSTRACT

Streptococcus pneumoniae is a gram-positive bacterium responsible for various diseases like pneumonia, acute otitis media, sinusitis, meningitis and bacteraemia. These diseases cause a significant amount of morbidity and mortality. Although polysaccharide vaccines are available, the protection provided by these vaccines is serotype-dependent and not enough in sensitive populations like children and older people. Designing a subunit vaccine by using proteins that are responsible for the pathogenesis of diseases can provide better protection against bacterial infections. In this study, we present the design of a novel multi-epitope vaccine against *Streptococcus pneumoniae* using an immunoinformatic approach. More than 1170 epitopes were identified against B cells, cytotoxic T lymphocytes and helper T lymphocytes from more than 60 pneumococcal proteins. Epitopes were further screened, and potential epitopes were selected for vaccine development. Seven different vaccine combinations that harbour the 15 dominant B-cell, cytotoxic T cell and helper T cell epitopes were evaluated with linker and β-defensin adjuvant to finalise the best vaccine construct. Bioinformatics tools were used to analyse the construct's physicochemical properties, secondary and tertiary structures, allergenicity, antigenicity and immunogenicity. Docking studies with the TLR-4 receptor and molecular dynamics simulations indicated strong binding affinity and stability. In silico immune response simulations predicted robust IgG immune response generation and observed more than 200 000 $IgG_1 + IgG_2$ counts per mL. Similarly, cell-mediated immunity was also enhanced by the designed vaccine construct. The construct was codon-optimised and cloned in silico for expression in *Escherichia coli*. These findings suggest that the construct is a promising candidate for further experimental validation.

1 | Introduction

Streptococcus pneumoniae causes a range of diseases, such as otitis media, sinusitis and pneumonia, which are non-invasive as well as invasive diseases such as bacteraemia and meningitis [1]. As per the World Health Organization (WHO), pneumonia is responsible for 14% of all deaths of children below 5 years of age. In 2019, almost 740 180 children died due to pneumonia worldwide, and *S. pneumoniae* stands to be the major cause of pneumonia [2]. Mostly, this burden of pneumococcal diseases is higher in developing countries. Although the cases of pneumococcal diseases have decreased substantially after the incorporation of different types of vaccines (as per CDC shown in Figure 1) [3], because of antibiotic resistance and serotype replacement, mortality and morbidity due to pneumococcal diseases remain an issue that requires urgent attention [4].

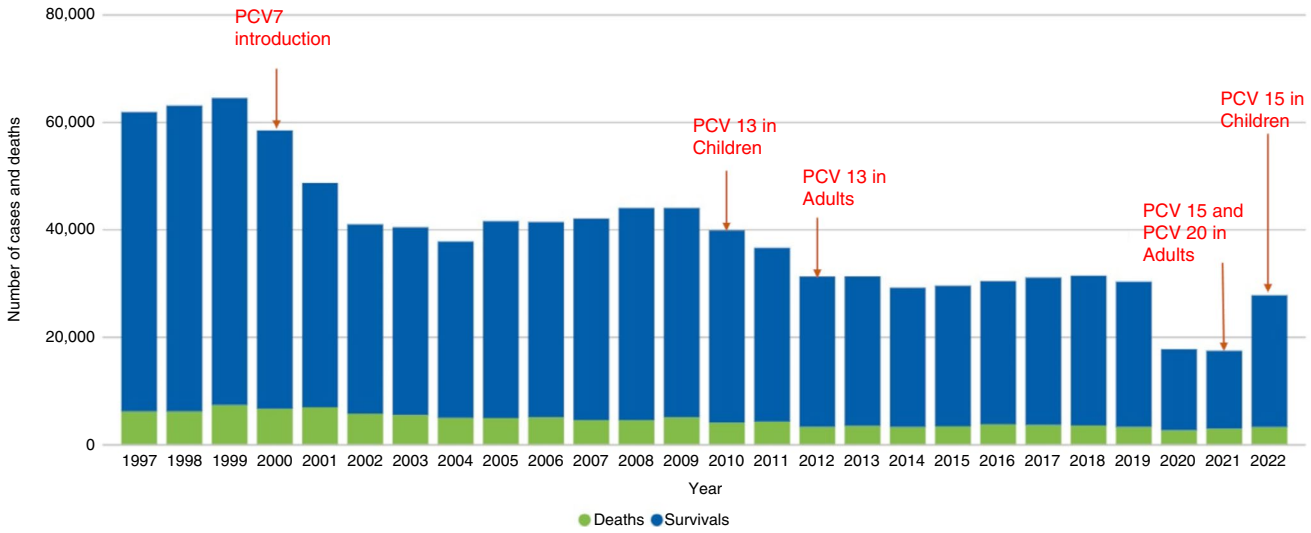


FIGURE 1 | Trend in invasive pneumococcal diseases in adults from 1997 to 2022 in the United States (https://www.cdc.gov/abcs/bact-facts/data-dashboard.html).

Based on unique glycan components and linkages in the polysaccharide capsule of bacteria, almost 98 serotypes have been identified [4]. Mainly, two different types of vaccines are being used for *S. pneumoniae*: (i) pneumococcal polysaccharide vaccine (PPV) and (ii) pneumococcal conjugated vaccine (PCV); both of these vaccines utilise the capsular polysaccharide as an antigen [5]. Although available polysaccharide vaccines have provided protection against bacterial infection, a large number of serotype availability limits the activity of the vaccines due to increasing infections by non-vaccine serotypes, non-capsular serotypes and limited immunogenicity in specific populations (children, old aged, immunocompromised, etc.) [6]. PCVs contain a conjugated protein with the polysaccharide antigen, which produces a better immunogenic response than PPVs. However, manufacturing such vaccines is difficult; serotype replacement is also an issue, and they are costly [7]. Different PPVs and PCVs have different serotype coverage, which has been given somewhere else [8–12].

Because of too many serotypes, classical methods of vaccine development have failed to provide a potential vaccine. First, second and third generations of vaccines, which have been used for a long time, are not able to provide the desired protection [13, 14]. New-generation subunit vaccines containing peptides or proteins as antigens can be a promising strategy. This strategy is being tested for many infectious diseases, as well as some chronic diseases like cancer and Alzheimer's. Although some vaccines have shown promising results, challenges remain [15]. One challenge in the case of vaccines against infectious diseases is that peptides/proteins do not produce enough immune response, and an additional adjuvant is required to produce an optimum immune response [16].

In the era of bioinformatics and immunoinformatics, vaccine development is taking a turn. These tools are used to design effective vaccines with better efficacy and improved immunogenicity while proceeding with quick results [17, 18]. A concept given by Rappuoli called 'reverse vaccinology' utilises the knowledge of bioinformatics and immunoinformatics to design the vaccine by analysing the genome sequence of the organism for the potential vaccine targets [19]. The in silico approach has already been utilised by different authors to develop effective vaccines [20–25].

One of the most recent works that adapted a similar strategy has achieved promising results, showcasing the superiority of this strategy by identifying a potential vaccine candidate against *S. pneumoniae* [26].

In the current attempt, an immunoinformatics approach has been employed to design a new pneumococcal vaccine that could be effective against diverse pneumococcal serotypes. After a literature review, different antigenic proteins of *S. pneumoniae* were identified for this study and further processed to construct the vaccine [27–32].

2 | Materials and Methods

2.1 | Retrieval of Amino Acid Sequence of Selected Proteins and 3D Structure of Receptors

Primary amino acid sequences of proteins were retrieved from NCBI and UniProt websites in FASTA format [33–35]. 3D structures of TLR-2 (2Z7X), TLR-4 (4G8A), MHC-1 (5WJL) and MHC-2 (5JLZ) were obtained from the RCBS Protein Data Bank [36].

2.2 | Identification of B-Cell Epitopes

B cells recognise antigens by utilising the receptors on their surface, and receptor–antigen interaction leads to the induction of humoral immunity [37]; thus, the inclusion of B-cell epitopes in the study becomes necessary. ABCpred server was employed with a 0.51 threshold value to identify the continuous B-cell epitopes, which provide 16 amino acid long epitopes. This server uses a recurrent neural network to identify the epitopes [38, 39].

2.3 | Identification of Helper T Lymphocyte (HTL) Epitopes and Population Coverage

HTL plays a major role in both humoral and cellular immunity by interacting with MHC-II receptors on APCs [40]; thus,

the incorporation of HTL epitopes in the vaccine construct can provide significant induction to an immune response. MHC-II Binding Prediction server was used from the IEDB website for HTL epitopes prediction. This server utilises the NetMHCIIpan-4.1 prediction method, which utilises training of a neural network based on the binding affinity of predicted epitopes as well as mass spectrometry-eluted ligands [41], where the prediction for human HTL epitopes was pursued against seven alleles of the HLA-DRB isotype [42]. This server compares the epitope score with 5 million peptides from the SwissProt database, provides 15-mer epitopes, and gives a percentile rank; the lower the percentile rank, the higher the binding affinity towards the MHC-II receptor [43, 44]. Human MHC (HLA) alleles are highly polymorphic in nature, and almost 1000 different alleles have been identified. Considering this polymorphism in alleles, the complications increase with different frequencies of expression among the different ethnicities. This makes it important to include a population coverage study of MHC epitopes. In this study, the population coverage program of the IEDB website was utilised to predict the population coverage provided by MHC-II epitopes incorporated into the vaccine construct. This web server allows for the calculation of population coverage for different ethnicities as well as for the whole world population. For the input in the web server, selected MHC-II epitopes were provided with the different HLA-DRB alleles, and the coverage was calculated for the world's population [45].

2.4 | Identification of Cytotoxic T Lymphocyte (CTL) Epitopes

CTL interacts with MHC-I receptor on APCs and based on this MHC-I molecule interaction, CTLs kill the infected cells in an antigen-specific manner [46]. Rationally, incorporating CTL epitopes in vaccine design becomes essential. For the present study, NetCTL-1.2 server was employed to identify the CTL epitopes; this server makes predictions by involving predictions of 'proteasomal breakdown of peptide', 'TAP transport efficiency' and 'binding affinity with MHC Class I' using artificial neural networks trained to predict and calculate information regarding epitope prediction [47–49]. Details of proteins and epitope screening and selection schematic are given in Figure S1.

2.5 | Cluster Evaluation of MHC Epitopes

The MHC region on the genome in humans is too polymorphic and contains thousands of alleles. Analysis of interactions between different alleles can provide information regarding the epitope's interaction and production of a potential immune response. For the cluster analysis of MHC epitopes, MHCcluster-2.0 web server was utilised. This web server uses NetMHCpan-2.8 and NetMHCIIpan-3.2 methods for the prediction of CTL and HTL epitope binding, respectively. To make the prediction regarding MHC-I clustering, all the HLA allele representatives were selected, and default parameters like total peptides to include and bootstrap calculation were kept at 50 000 and 100, respectively. For the MHC-II cluster analysis, other parameters were kept the same, but the selected allele representatives were HLA-DR [50].

2.6 | Allergenicity and Antigenicity Prediction of Epitopes

AllerTOP v2.0 server was employed to predict the allergenicity of derived epitopes. This server uses an auto cross covariance (ACC) based method that converts the protein sequence into uniform length vectors. This vector is then analysed for different properties of amino acids like hydrophobicity, molecular size, ability to form helix, capacity of amino acids to form β-strand and relative abundance to predict the allergenicity of the given polypeptide chain [51]. Only those epitopes were considered for further study that were non-allergenic in nature.

Vaxijen v2.0 was employed to predict the antigenicity of derived peptides. The Vaxijen server provides options to select target organisms (such as viruses, bacteria, fungi, cancer and parasites); in this case, 'bacteria' was selected, and a 0.4 threshold value was kept. Vaxijen utilises an alignment-independent method for the prediction of antigenicity [52]. Here, for the selection of the best epitopes for the vaccine construct, only those epitopes were selected that have at least a twofold higher antigenic score than the threshold value.

2.7 | Multi-Epitope Vaccine Construction

After evaluation of epitopes for allergenicity and antigenicity, epitopes that are the most safe and antigenic were selected from different proteins to make a construct and to provide broader serotype coverage. For the construction of the vaccine, epitopes were linked using linkers and an adjuvant was incorporated. In the construct of the vaccine, 'KK' linkers were utilised for CTL and BCL peptides, while for HTL peptides 'GGGS' linkers were utilised. 'EAAAK' linker was added between the adjuvant and the first epitope as well as at the end of the construct [53].

2.8 | Physiochemical Properties, Allergenicity and Antigenicity Derivation

Expasy's ProtParam server was employed to compute the physiochemical parameters like molecular weight, atomic composition, estimated half-life, aliphatic index, theoretical pI, amino acid composition, extinction coefficient and grand average of hydropathicity (GRAVY) [54]. Aqueous solubility prediction for the derived construct was done using the Protein-Sol server, where, if the predicted solubility value was higher than the 0.45 threshold value, the protein construct was considered to have higher solubility in comparison to average proteins from *E. coli* [55]. For the prediction of allergenicity and antigenicity of the vaccine construct, AllerTOP [51] and Vaxijen v2.0 [52] servers were used respectively.

2.9 | Secondary Structure Prediction

The prediction of the secondary structure for linear construct was made by the PSIPRED 4.0 server. A neural network is utilised by

this server in two steps based on PSI-BLAST generated position-specific scoring matrices to create the secondary structure of the given amino acid sequence. It provides information about different types of arrangement (such as strands, helix, coils and disordered) of amino acids in the construct [56, 57].

2.10 | Tertiary Structure Prediction

I-TASSER is a widely used web server for the prediction of the tertiary structure of an amino acid sequence. Initially, this server finds templates from the protein data bank by utilising a multiple threading approach to construct a full-length atomic model using iterative template-based fragment assembly simulations, and based on this template, the server constructs the tertiary structure. I-TASSER results provide the tertiary structure as well as the root-mean-square deviation (RMSD) value, Confident-score (C-score) and TM-score. C-score is correlated with both TM-score and RMSD value, with correlation coefficients of 0.91 and 0.75, respectively, for the first model out of the given main five models [58].

2.11 | Refinement and Validation of Tertiary Structure

Refinement of the tertiary structure was done by the GalaxyRefine web server. This web server starts refinement by rebuilding side chains, which leads to repacking of sidechains followed by overall relaxation of structure using molecular dynamics (MD) simulation [59, 60].

UCLA's SAVES v6.0 server in which two programs (i) ERRAT—which provides an overall quality factor of structure— [61] and (ii) PROCHECK—which provides a Ramachandran plot showing four regions where amino acids of proteins reside—were employed to validate the tertiary structure of the designed vaccine [62]. The regions (most favoured region, additional allowed region, generously allowed region and disallowed region) in Ramachandran plots give an idea about the quality of the tertiary structure in discussion [63]. Also, ProSA (Protein Structure Analysis)-web server provided the z-score for the structures. The z-score gives an idea about overall model quality as it is a value obtained by comparing the given structure to native structures of similar size [64, 65].

2.12 | Discontinuous B-Cell Peptides Prediction

Discontinuous peptides are non-linear and made up of multiple different parts of the amino acid chain. Naturally, almost 90% of B-cell epitopes are in discontinuous form, which makes it important to consider them during vaccine design [66, 67]. Prediction of discontinuous epitopes was done using the ElliPro server from the IEDB website, which uses Thornton's method to identify linear epitopes and a residue clustering algorithm by keeping a 0.5 score as a threshold value [68].

2.13 | Protein–Protein Docking

The docking of the vaccine construct with different receptors like TLR-2, TLR-4, MHC-1 and MHC-2 was done by the ClusPro 2.0 web server [69]. The docking is performed by this server in three steps: the first is rigid body docking, the second step is forming a cluster of 1000 structures having the lowest energy based on RMSD and the third step removes steric clashes by minimising the energy [70–72].

For binding energy calculation of receptor–ligand complex, the PRODIGY server was used, which provides the contact-based binding affinity prediction as well as dissociation constant (K_a) [73, 74]. To visualise the interaction between amino acids of receptors and ligands (here constructed vaccine), the DIMPLOT program of LigPlot$^+$ software was used, which provides better insight into understanding the interactions of molecules during docking [75].

2.14 | MD Simulation

During the interaction of ligand and receptor, different types of conformations take place, which depend on the flexibility of structures as well as the conformational transitions of the interaction. Determining these properties can give useful insights into the molecular interaction of ligands and receptors [76]. In this study, to perform MD simulation, the basic mode of the iMOD server was used. iMOD is an NMA (Normal Mode Analysis) based method combined with Coarse-grained representations, providing information such as deformability, flexibility, stiffness, eigenvalue and covariance map [77–79]. MD simulation was also performed using GROMACS software, where gromacs-2024.2 version 2024.2 was used [80–86]. Input files were generated using CHARMM-GUI web server's solution builder program, where the system was built by providing the vaccine–receptor complex in protein database (.pdb) format. A rectangular box of water was generated using TIP3 model water, and the system was neutralised using Na^+ and Cl^- ions. The ions were added at a concentration of 0.15 M using the Monte Carlo method of ion placing. Once the system was generated, the temperature of 311.15 K and pressure of 1 bar were set, and the system was equilibrated using the energy minimisation step. After the minimum energy levels were obtained for the system, NVT (constant number of particles, volume and temperature) and NPT (constant number of particles, pressure and temperature) ensembles were performed following MD simulations for 100 ns (5×10^7 steps) [87–90].

2.15 | In Silico Immune Response

Estimating the immune response of any vaccine for the evaluation of efficacy is one of the most important parts of the study. Here, for simulating the immune response of the constructed vaccine, the C-ImmSim server was utilised. The C-ImmSim web server is a polyclonal agent-based model that simulates the immune system of a mammal and gives information regarding humoral and cellular responses [91–94].

2.16 | Codon Optimisation and In Silico Cloning

To convert the amino acid sequence of the vaccine construct to the codon sequence, the EMBOSS Backtranseq server was

TABLE 1 | B-cell epitopes that were incorporated in vaccine construct were identified from selected proteins and sort listed based on allergenicity and antigenicity analysis.

Proteins	Epitope	Start	End	Antigenicity	Allergenicity
Pneumococcal histidine triad protein A (PhtA)	QPSPQSTPEPSPSLQP	368	383	1.1134	Probable non-allergen
Choline-binding protein A (CbpA)	KKKAEDQKEEDRRNYP	152	167	1.6945	Probable non-allergen
Choline-binding protein F (CbpF)	GSTWYYLNASNGDMKT	290	305	1.1101	Probable non-allergen
Choline-binding protein J (CbpJ)	EKKAKDQKEEDRRNYP	311	326	1.7306	Probable non-allergen
Zinc metalloproteinase C (ZmpC)	GVNADTSLDDLYLD	1620	1632	1.4054	Probable non-allergen

utilised. This server provides the option to choose from different organisms, where *E. coli* K-12 was selected because the obtained codon sequence was supposed to be used for the expression of protein by *E. coli* as a host [95, 96]. The obtained codon sequence was then optimised using the JCat web server, which is a very fast and simple method because manually defining highly expressed genes is not required; it also provides data like %GC content and Codon Adaptation Index (CAI) value [97, 98].

Cloning of a desired protein is a method of expressing it by incorporating the genome sequence of that protein in plasmid of a host like *E. coli*. Cloning is used for the synthesis of many proteins and therapeutic molecules [99]. To mimic the cloning of designed vaccine construct SnapGene v7.0.2 software was utilised, in which a function called 'Restriction and insertion cloning' provides the platform to perform in silico cloning. pET-28a(+) plasmid was selected as a vector, and then the sequence of nucleotides was incorporated into the vector by using restriction enzyme sites (*Xho*I and *Mlu*I) [21].

2.17 | Open Reading Frame (ORF) of Nucleotide Sequence

ORF is the sequence of DNA or RNA that can undergo the translation process and indicates the coding sequence for a protein [100]. The ORF finder server from the NCBI website was employed to find the ORF in the codon sequence of the vaccine construct [101].

3 | Results

3.1 | Retrieval of Amino Acid Sequence of Selected Proteins and 3D Structure of Receptors

The study was initiated by retrieving the amino acid sequence of 60 proteins that are essential for the pathogenesis of bacteria (*S. pneumoniae*). These proteins were then screened for their homology with the mouse and human proteomes using protein–protein BLAST and allergenicity using the AllerTOP server, which led to the exclusion of 19 proteins from the study. Epitopes were identified from 41 proteins, as shown in Table S1. Based on the screening criteria, the selected dominant epitopes, which were part of the final vaccine construct, belong to the 14 proteins. Only non-allergen and non-homologous (< 30% similarity in BLAST analysis) proteins were incorporated into the study. Table S2 contains a detailed description of these 14 proteins.

3.2 | Identification of B-Cell Epitopes

After employing the ABCpred server with a 0.51 threshold value, 16 mer epitopes were obtained for each protein, which were analysed for their antigenicity and allergenicity. Based on these results, epitopes that were non-allergenic and had an antigenicity score of more than 1 were identified and selected for the construction of vaccine, which is given in Table 1.

3.3 | Identification of HTL Epitopes

Based on the NetMHCIIpan-4.1 method, the prediction of HTL epitopes was performed by the MHC-II server on the IEDB website, which provided 15 amino-acid-long epitopes. These epitopes were then analysed for their allergenicity, antigenicity and IFN-γ induction capacity. Analysis results showed epitopes having non-allergenicity, antigenicity more than 0.7 and positive IFN-γ induction (Table 2).

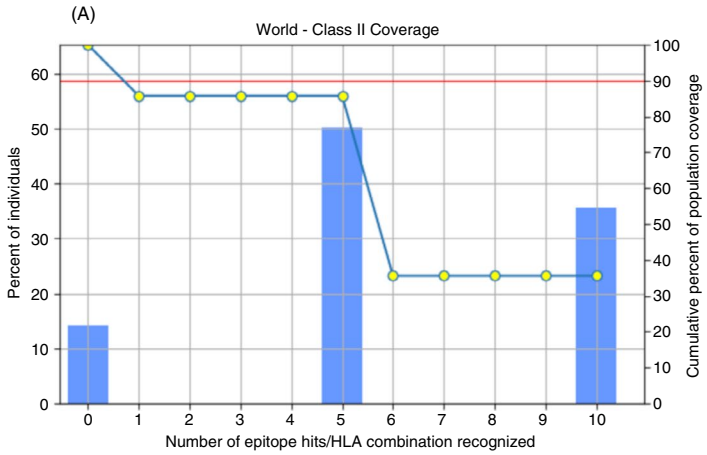
Population coverage prediction web server provided the results, where given epitopes showed 85.76% cumulative population coverage (Figure 2A) with an average hits per HLA combination of 6.07. Additionally, the minimal number of epitope hits per HLA combination that 90% of the population could recognise was 3.51 (Figure 2B).

3.4 | Identification of CTL Epitopes

CTL epitopes derived from the NetCTL 1.2 server for selected proteins by keeping a weightage of 0.15 for C-terminal cleavage and 0.05 for TAP transport efficiency, while the threshold was set at 0.75 for the identification of epitopes. Determination of possible epitopes by this server was based on a combined score of MHC-I affinity, TAP transport efficiency and C-terminal cleavage score. Epitopes with non-allergenicity and more than 1 antigenicity score were chosen for the construction of a vaccine; those selected epitopes are listed in Table 3.

TABLE 2 | HTL epitopes used for vaccine construct were identified from selected proteins and sorted listed based on allergenicity, antigenicity and IFN-γ induction capacity analysis.

Proteins	Epitope	Start	End	Antigenicity	IFN-γ induction	Allergenicity
Autolysin A (LytA)	EADYHWRKDPELGFF	38	52	0.7777	0.15599044	Probable non-allergen
Pneumococcal histidine triad protein E (PhtE)	YELFKPEEGVAKKEG	574	588	0.8538	0.36836895	Probable non-allergen
Choline-binding protein G (CbpG)	KEFQNTASNELTTYD	164	178	0.8697	0.64277099	Probable non-allergen
Pneumococcal Serine-Rich Repeat Protein (PsrP)	TTSQSLSQSKSLSVS	379	393	0.8549	0.14449709	Probable non-allergen
Iron uptake transporter permease protein C (PiuC)	TAQILGLDVEKEQKE	211	225	0.9212	0.15027014	Probable non-allergen



(B)

Population/ Area	Class II		
	Coverage[a]	average_hit[b]	pc90[c]
World	85.76%	6.07	3.51
Average	85.76	6.07	3.51

a - projected population coverage

b - average number of epitope hits / HLA combinations recognized by the population

c - minimum number of epitope hits / HLA combinations recognized by 90% of the population

FIGURE 2 | Population coverage prediction. (A) Chart showing cumulative population coverage gives idea about the population that expresses the HLA alleles with which the HTL epitopes from vaccine construct can interact and (B) shows coverage, average hit, and pc90 for the selected alleles interacting with the HTL epitopes in construct.

TABLE 3 | CTL epitopes identified from selected proteins and sorted listed based on allergenicity and antigenicity analysis. These epitopes were finally used for the construction of a vaccine along with HTL and BCL.

Proteins	Epitope	Start	End	Antigenicity	Allergenicity
Choline-binding protein G (CbpG)	GTGATITGY	206	214	1.5918	Probable non-allergen
Pneumococcal choline-binding protein A (PcpA)	TTSLNMLML	244	252	1.2886	Probable non-allergen
Pneumococcal lipoprotein (pneumococcal iron acquisition, PiaA)	TSKDPRANY	226	234	1.4052	Probable non-allergen
Pneumococcal serine-rich repeat protein (PsrP)	RTDRIGINY	358	366	1.5121	Probable non-allergen
Iron uptake transporter permease protein B (PiuB)	VTEVIAYRF	150	158	1.2189	Probable non-allergen

3.5 | Cluster Evaluation of MHC Epitopes

Cluster analysis of MHC alleles done by MHCcluster-2.0 web server provided a heat map and allele interaction tree. In the heat map of allelic interaction, the red part shows the stronger interaction, while the yellow part indicates the weaker interaction. The allelic interaction tree shows the possible interaction of the epitopes and alleles, where the nearer the epitope to an allele, the better the interaction. The output given by the web server for the given peptides is shown in Figure 3.

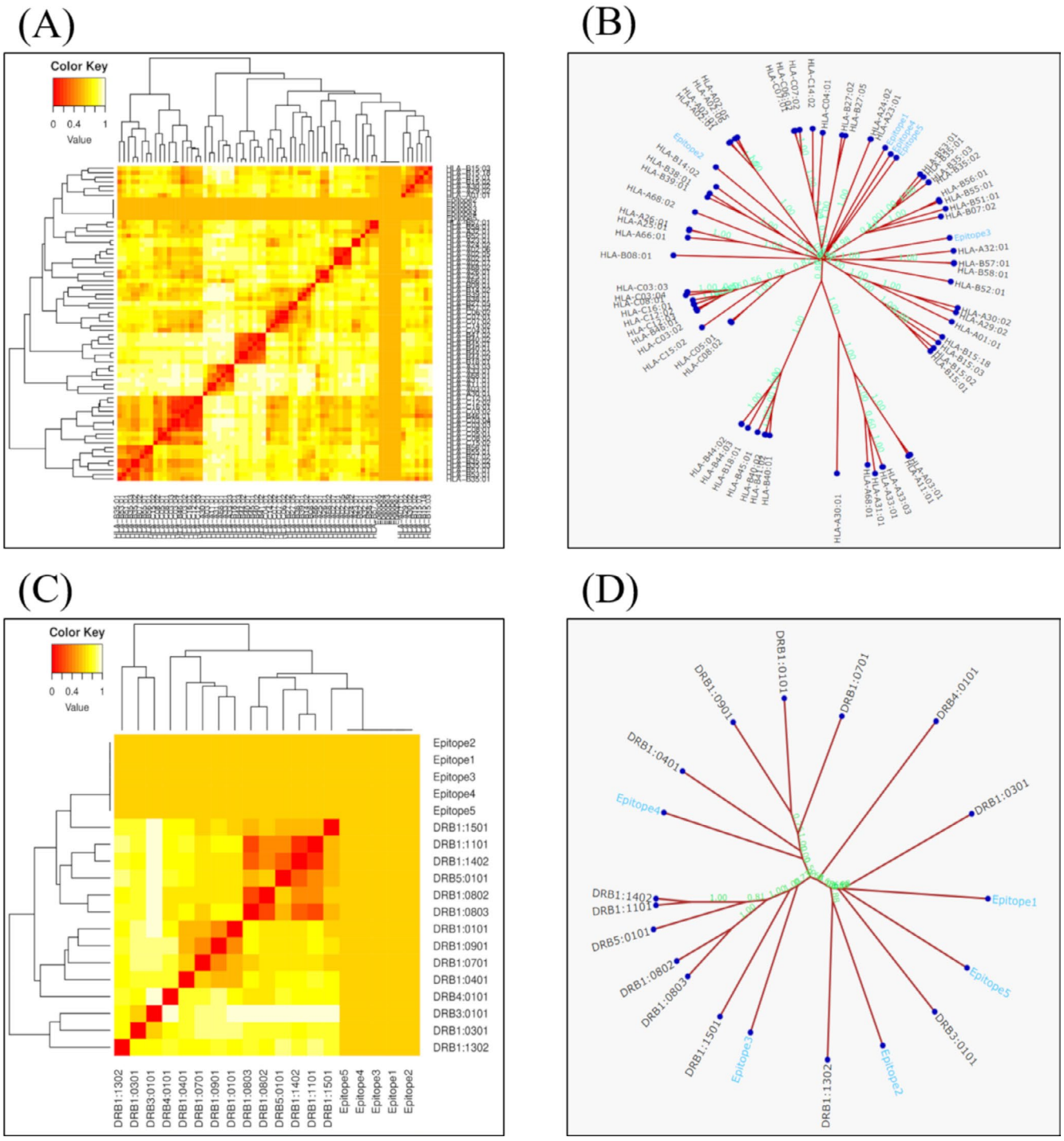


FIGURE 3 | MHC cluster evaluation, where (A) heat map of MHC-1 peptide interaction, (B) interaction tree of MHC-1 alleles, (C) heat map of MHC-II peptide interaction and (D) interaction tree of MHC-II alleles with selected epitopes. In heat map, the red area shows stronger interaction, while the white area shows weaker interaction. In case of allelic interaction tree, nearer the epitope to an allele, better the interaction.

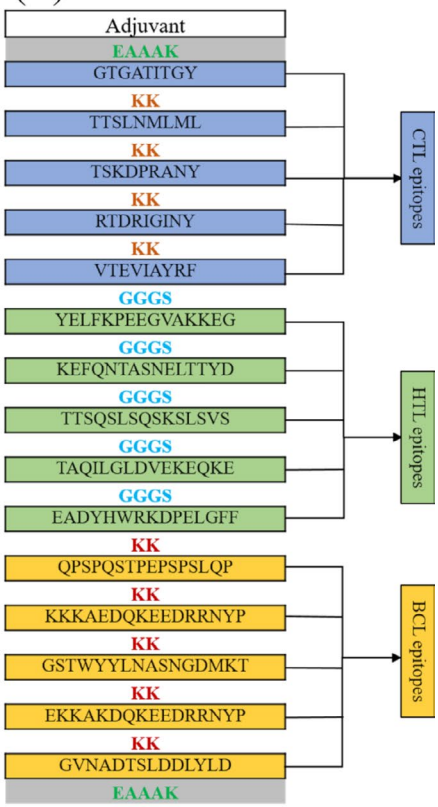
3.6 | Multi-Epitope Vaccine Construction

After the evaluation of all the epitopes, a total of 15 epitopes (5 epitopes of each type) were chosen to make multi-epitope vaccine constructs. These chosen epitopes were linked together with linkers, and an adjuvant was added to improve the immunogenicity. To link the adjuvant, an EAAAK linker was used, which was also added at the end of the construct to improve stability. A total of seven different vaccine constructs were designed and evaluated using the same set of epitopes and linkers, as well as the same adjuvant. Details of the different constructs are given in Table S3. The best construct out of seven was further studied. Figure 4B showcases the selected vaccine construct, which has 291 amino acids, while Figure 4A includes the amino acid sequence of the construct.

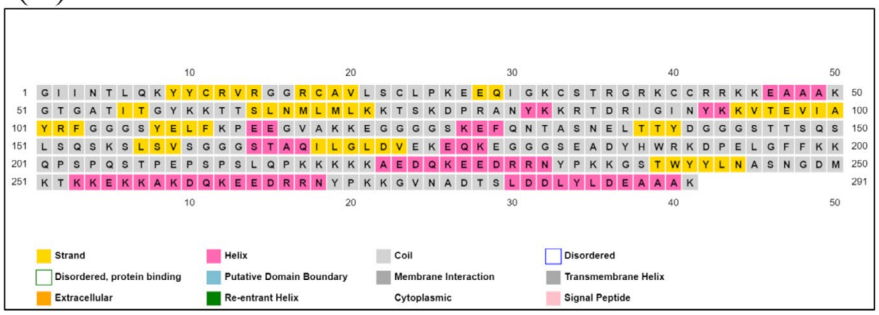
(A)

V1	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKKEAAAK GTGATITGYKKTTSLNMLMLKKTSKDPRANYKKRTDRIGINYKKVTEVI AYRFGGGSYELFKPEEGVAKKEGGGGSKEFQNTASNELTTYDGGGSTTS QSLSQSKSLSVSGGGSTAQILGLDVEKEQKEGGGSEADYHWRKDPELGF FKKQPSPQSTPEPSPSLQPKKKKKAEDQKEEDRRNYPKKGSTWYYLNAS NGDMKTKKEKKAKDQKEEDRRNYPKKGVNADTSLDDLYLDEAAAK

(B)



(C)



(D)

	No. of Amino acids	Molecular weight (gm/mole)	Theoretical pI	Aliphatic index	GRAVY	Predicted half-life (mammalian reticulocytes, in vitro)	Predicted Solubility	Antigenicity	Allergenicity
V1	291	32248.31	9.59	49	-1.168	30 hours	0.818	1.2307	Probable Non-allergen

FIGURE 4 | Vaccine construct and its evaluation. (A) Amino acid sequence of vaccine construct. (B) Vaccine constructs showing adjuvant in white box, CTL epitopes in blue box connected with 'KK linker', HTL epitopes in green box connected with "GGGS linker" and BCL epitopes in yellow box connected with 'KK linker'. (C) Secondary structure of vaccine where yellow box represents strands, pink represents helix and grey represents coil in structure. (D) Estimated physiochemical properties of construct.

3.7 | Physiochemical Properties, Allergenicity and Antigenicity Derivation of Vaccine Construct

Constructed vaccine was analysed for its physicochemical properties like molecular weight (32 248.31 g/mol), theoretical pI (9.59), estimated half-life, aliphatic index [49], GRAVY (−1.168) using Expasy's ProtParam server and solubility was estimated using Protein-Sol server. Construct was also evaluated for its antigenicity and allergenicity. Summary of all these parameters is shown in Figure 4D.

3.8 | Secondary Structure Prediction

Secondary structure of construct was predicted using PSIPRED 4.0, which is shown in Figure 4C. For the given vaccine construct, the percentage of strands, helix and coils predicted by the server was 18.56%, 19.93% and 61.51%, respectively.

3.9 | Tertiary Structure Prediction

The tertiary structure of the constructed sequence was predicted by the I-TASSER web server, which provided five different tertiary models. The model with the highest C-score was selected for further study. Also, the I-TASSER web server provided the TM-score for each model. For the representation of the selected tertiary structure, its Ramachandran plot, C-score, TM-score and values from the Ramachandran plot are shown in Figure 5.

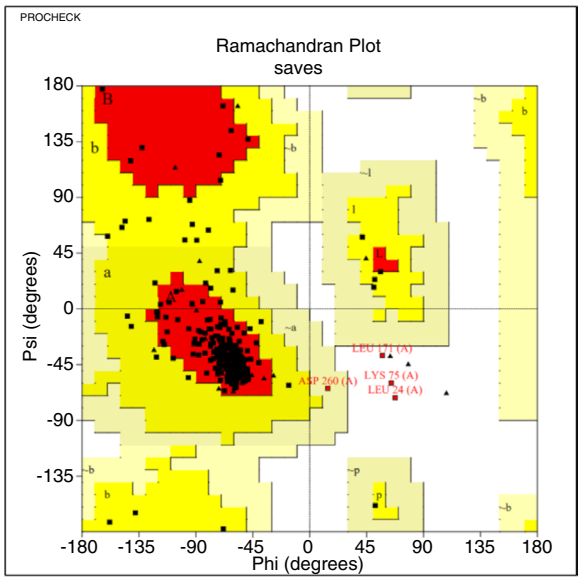
3.10 | Refinement and Validation of Tertiary Structure

Using GalaxyRefine web server, the tertiary structure was refined. This refinement of structure is required because, even after utilisation of the best tertiary structure building web server, due to the novelty in construct, the lack of a template to build on poses some difficulties, which may lead to an error in structure. RMSD value is a measure of the difference between two structures, which was also provided by the GalaxyRefine server.

Validation of refined structure was done by UCLA's SAVES v6.0 server, which provided data on the overall quality factor of the structure as well as the Ramachandran plot. After the refinement of the structure, residues in the most favoured region increased to 97.2% from 87%. The z-score (−3.2) and its plot for the given construct were obtained from the ProSA-web server. The refined tertiary structure, Ramachandran plot, overall model quality plot (z-plot) and local model quality plot for that structure are given in Figure 6.

(A)

(B)



(C)

C-score	TM-score	Ramachandran Plot			
		Most favored regions	Additional allowed regions	Generously allowed regions	Disallowed regions
-3.18	0.36±0.12	87.0%	11.3%	0.4%	1.2%

FIGURE 5 | Tertiary structure of vaccine construct (A) obtained from the I-TASSER web server, (B) its Ramachandran plot (C) showing the C-score based on which best structure was selected from I-TASSER and parameters obtained from Ramachandran plot.

3.11 | Discontinuous B-Cell Epitopes Prediction

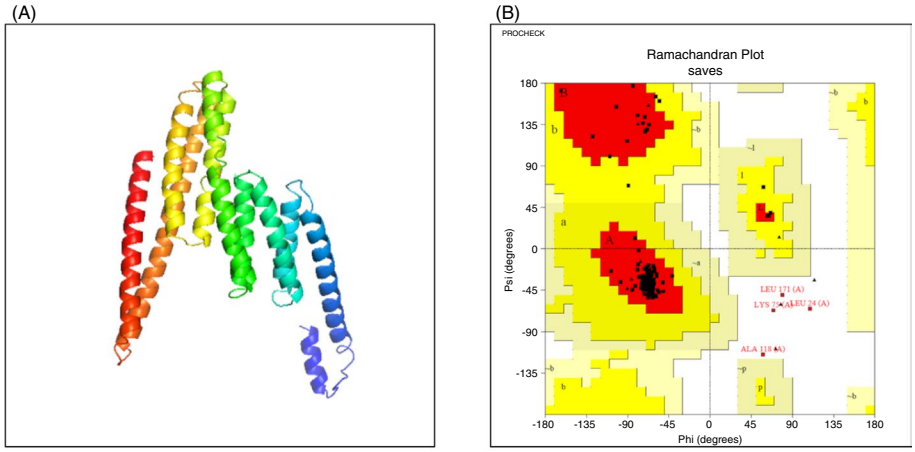
As the importance of discontinuous B-cell epitopes is higher due to their versatility, their prediction in constructed structure is important and that was done using the Ellipro server. This server provided different B-cell discontinuous epitope sequences as well as their locations in the 3D structure, where six different discontinuous epitopes were predicted by the server from the vaccine construct, which is shown in Figure 7.

3.12 | Protein–Protein Docking

As this construct is made up of different epitopes that have the ability to bind to different receptors and exert the effect in terms of immune stimulation against *S. pneumoniae*, the binding affinity toward some important receptors like TLR-2, TLR-4, MHC-1 and MHC-2 was checked by using ClusPro 2.0 web server, which provided the receptor–ligand complexes Figure 8A–D. These complexes were utilised to calculate the binding energy and dissociation constant using the PRODIGY web server, which is given in Figure 8E. As TLR-4 is a pattern recognition receptor (PRR) that is engaged by bacterial endotoxins for educating the immune system, and as the current vaccine is a bacterial vaccine where *S. pneumoniae* secretes LPS endotoxin, which is a TLR-4 ligand, we performed the docking and further evaluation by MD simulation with the TLR-4 receptor. Upon analysing a complex of TLR-4 and the vaccine construct for the interactions of amino acids, it was observed that the interaction of the protein and ligand (Figure 9) includes almost 32 hydrogen bonds and 4 salt bridges, details of which are given in Table 4.

3.13 | MD Simulation

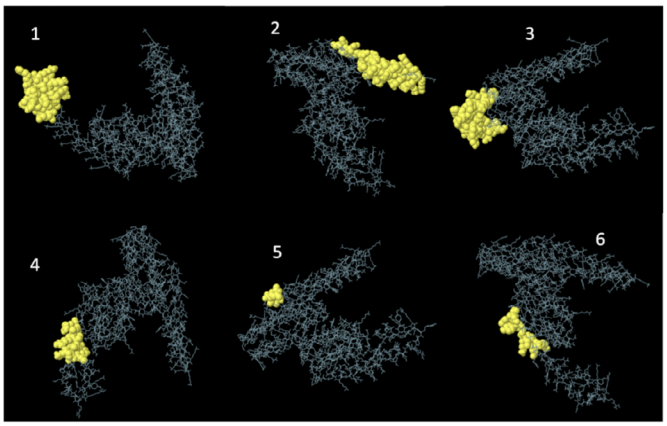
MD simulation, which gives an idea about the structure's flexibility and deformability, was performed using the iMOD server's Basic program. The obtained results provided different plots for different parameters as well as the eigenvalue which is the indicator of energetic contribution for the motion of a particular complex during simulation. The results of the MD simulation performed for the complex of TLR-4 and the vaccine construct show an eigenvalue of 6.917089×10^{-6} and other plots can be seen in Figure 10. The RMSD (root mean square deviation) value for the complex was never higher than 4 nm, which is an acceptable outcome considering the size of the complex is very large. The RMSD of the protein was within the range of 0.2–0.6 nm, which suggests the stability of the protein was not disturbed by the interaction with the vaccine construct. The root mean square fluctuation (RMSF) of different chains of the protein as well as the vaccine construct chain was within the range of 1.5–3 nm, which is parallel



(E)

	z-score	RMSD	Ramachandran plot			
			Most favored regions	Additional allowed regions	Generously allowed regions	Disallowed regions
VI	-3.2	0.279	97.2%	1.2%	0.0%	1.6%

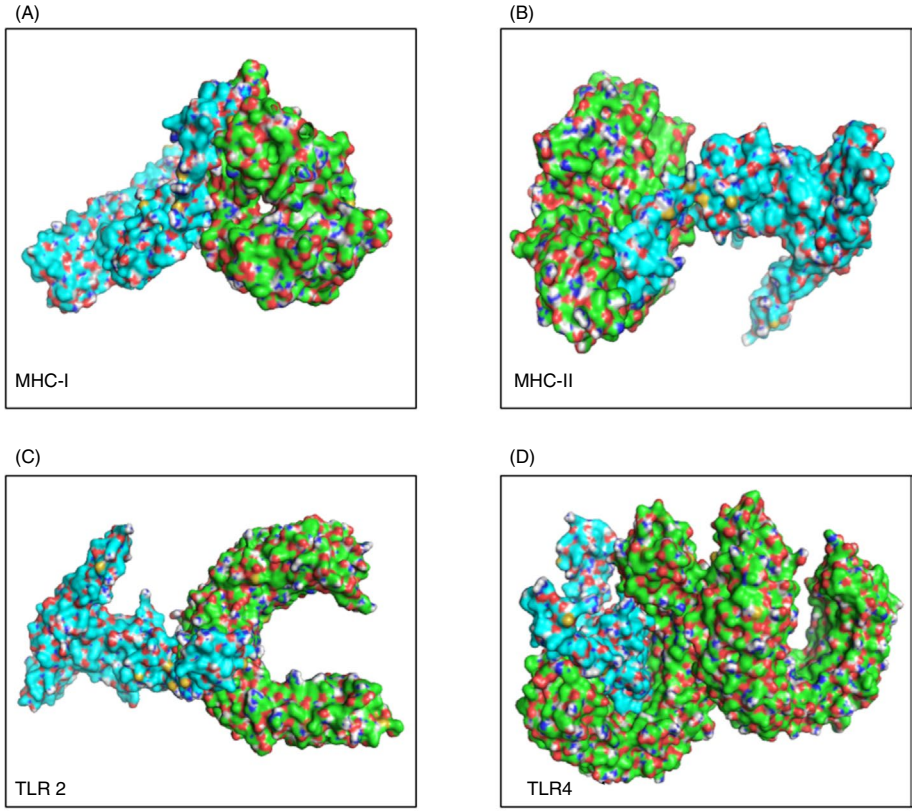
FIGURE 6 | Refinement and validation of tertiary structure of vaccine construct. (A) Refined tertiary structure, (B) Ramachandran plot showing 97.2% residues of proteins in most favoured region, (C) overall model quality plot (*z*-plot) obtained from the ProSA-web server having *z*-score of −3.2 suggests that the protein have *z*-score similar to native proteins, (D) local model quality plot indicates the stability of structure and (E) values for tertiary structure refinement and validation obtained from ProSA-web as well as Ramachandran plot.



(B)

Sr. no	Peptide sequence	Number of amino acid	Score
1	A:G1, A:I2, A:I3, A:N4, A:T5, A:L6, A:Q7, A:K8, A:Y9, A:Y10, A:C11, A:R12, A:V13, A:R14, A:G15, A:G16, A:R17, A:C18, A:A19, A:V20, A:L21, A:S22, A:C23, A:L24, A:P25, A:K26, A:E27, A:E28	28	0.907
2	A:S238, A:T239, A:Y241, A:Y242, A:L243, A:N244, A:A245, A:S246, A:N247, A:G248, A:D249, A:M250, A:K251, A:T252, A:K253, A:E255, A:K257, A:A258, A:K259, A:D260, A:Q261, A:K262, A:E263, A:E264, A:D265, A:R266, A:R267, A:N268, A:Y269, A:P270, A:K271, A:K272, A:G273, A:V274, A:N275, A:D277, A:T278, A:D281, A:D282, A:L285	40	0.76
3	A:G142, A:G144, A:S145, A:T146, A:T147, A:S148, A:Q149, A:S150, A:L151, A:S152, A:Q153, A:S154, A:K155, A:S156, A:L157, A:S158, A:V159, A:S160, A:F198, A:K200, A:Q201, A:P202, A:S203, A:P204, A:Q205, A:S206, A:T207, A:P208, A:E209, A:P210, A:S211, A:P212, A:S213, A:L214, A:Q215, A:P216, A:K217, A:K218	38	0.708
4	A:Q29, A:I30, A:G31, A:K32, A:C33, A:S34, A:T35, A:R36, A:G37, A:R38, A:K39, A:C40, A:R42, A:R43	14	0.684
5	A:A288, A:A289, A:A290, A:K291	4	0.601
6	A:K45, A:E46, A:A47, A:A49, A:K50, A:G51, A:G53, A:A54, A:T55, A:I56, A:T57, A:G58, A:Y59	13	0.57

FIGURE 7 | Discontinuous B-cell epitopes obtained from ElliPro web server. (A) 3D representation and (B) details of the location with number of residues.



(E)

Predicted binding affinity (ΔG, kcal/mole)				Predicted dissociation constant (M)			
MHC-I	MHC-II	TLR-2	TLR-4	MHC-I	MHC-II	TLR-2	TLR-4
-13.1	-20.5	-15.4	-8.4	2.30e-10	9.20e-16	5.40e-12	7.20e-07

FIGURE 8 | Docking of V1 with different receptors where green are receptors and blue is V1 construct. (A) Docking complex with MHC-I receptor, (B) docking complex with MHC-II receptor, (C) docking complex with TLR 2 receptor, (D) docking complex with TLR 4 receptor and (E) predicted binding affinity and dissociation constant for different complexes obtained using PRODIGY web server.

to the results of the deformability plot in Figure 10C. More results of MD simulation by GROMACS have been added to Figure S2A–E.

3.14 | In Silico Immune Response

In silico immune response predicted by the C-ImmSim server showed an increased IgM. Eventually, the IgM + IgG response was further increased after the booster dose on the 18th day of the first immunisation (Figure 11A). Similarly, cytokine secretion such as IFN-γ, TGF-β, IL-10 and IL-2 was also shown to be increased with the first injection as well as the booster dose (Figure 11B). Additionally, B cells and their isotypes, specifically memory B cells, were shown to increase drastically after the booster dose Figure 11C,D. The number of plasma B-cell isotypes secreting IgG1 and IgM was also elevated (Figure 11E) after the first and booster immunisations. The number of active T-helper cells and Th-memory cells was also predicted to be significantly higher, and Th1 cells were also increased in number in comparison to Th2 and Th17 Figure 11F–H. Role of macrophages (MAs) and dendritic cells (DCs) is significant during the activation of the immune system against bacterial infection. As shown in Figure 11I, after the first immunisation, the number of active MAs increased dramatically and maintained its level for up to 50 days, while in the case of DCs, the level was not that high, but it was maintained throughout the simulation, which lasted about 120 days (Figure 11J).

3.15 | Codon Optimisation and In Silico Cloning

As the constructed sequence is a novel protein and for the synthesis of this protein, it needs to be expressed in *E. coli*. For the expression, the codon sequence needs to be inserted into the plasmid of *E. coli* and for that, conversion of the amino acid sequence to the codon sequence was needed, which was done using the EMBOSS Backtranseq server. The codon sequence is shown in Table 5 with its GC content and CAI value.

Optimisation of this codon sequence was done by Jcat web server to improve the expression of protein from *E. coli*. After

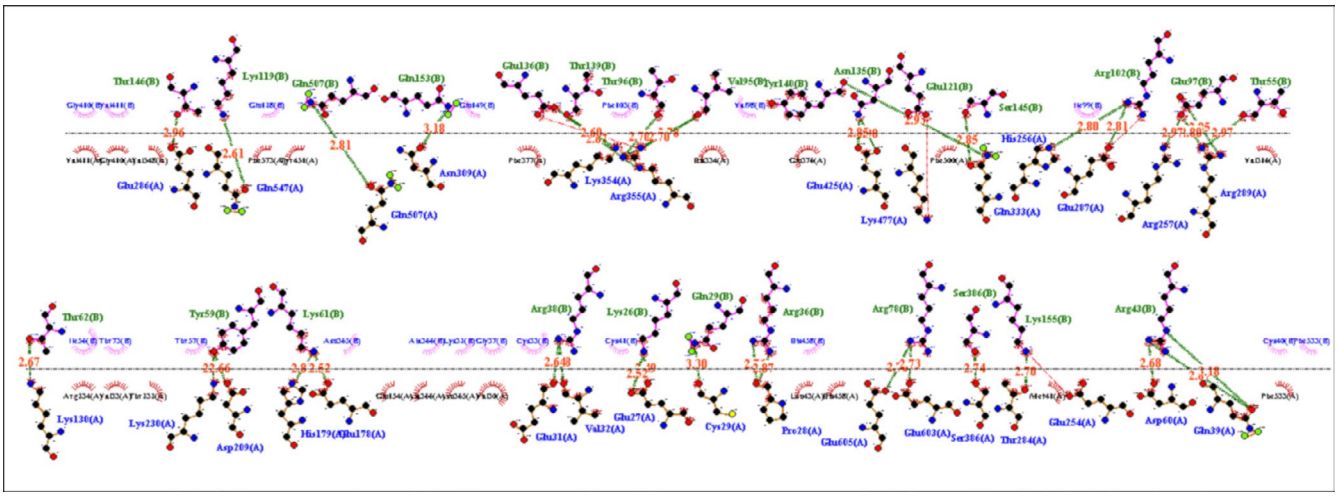


FIGURE 9 | Amino acid interactions in a complex of TLR-4 and V1 construct. Green bonds are hydrogen bonds, and red bonds are salt bridges. Chain (A) is receptor (TLR-4) and Chain (B) is vaccine construct.

TABLE 4 | List of hydrogen bonds and salt bridges present in amino acid interactions in a complex of TLR-4 and vaccine construct.

TLR-4	Vaccine construct		TLR-4	Vaccine construct	
Residue (location)	**Residue (location)**	**Bond length (Å)**	**Residue (location)**	**Residue (location)**	**Bond length (Å)**
Hydrogen bonds					
Glu (286)	Thr (146)	2.91	Lys (230)	Tyr (59)	2.66
Gln (547)	Lys (119)	2.61	Asp (209)	Tyr (59)	2.66
Gln (507)	Gln (507)	2.81	His (179)	Lys (61)	2.81
Asn (309)	Gln (153)	3.18	Glu (178)	Lys (61)	2.52
Arg (355)	Glu (136)	2.87	Glu (31)	Arg (38)	2.64
Arg (355)	Thr (96)	2.70	Val (32)	Arg (38)	2.68
Arg (355)	Val (95)	2.70	Glu (27)	Lys (26)	2.52
Glu (425)	Asn (135)	2.85	Cys (29)	Gln (29)	3.30
Gln (333)	Tyr (140)	2.93	Pro (28)	Arg (36)	2.87
Gln (333)	Ser (145)	2.85	Glu (605)	Arg (78)	2.76
His (256)	Arg (102)	2.80	Glu (603)	Arg (78)	2.73
Glu (287)	Arg (102)	2.81	Ser (386)	Ser (386)	2.74
Arg (257)	Glu (97)	2.97	Thr (284)	Lys (155)	2.70
Arg (289)	Glu (97)	2.80	Asp (60)	Arg (43)	2.68
Arg (289)	Thr (55)	2.97	Gln (39)	Arg (43)	3.18
Lys (354)	Thr (139)	2.60	Lys (130)	Thr (62)	2.67
Salt bridges					
Arg (355)	Glu (136)		Glu (287)	Arg (102)	
Lys (477)	Glu (121)		Glu (254)	Lys (155)	

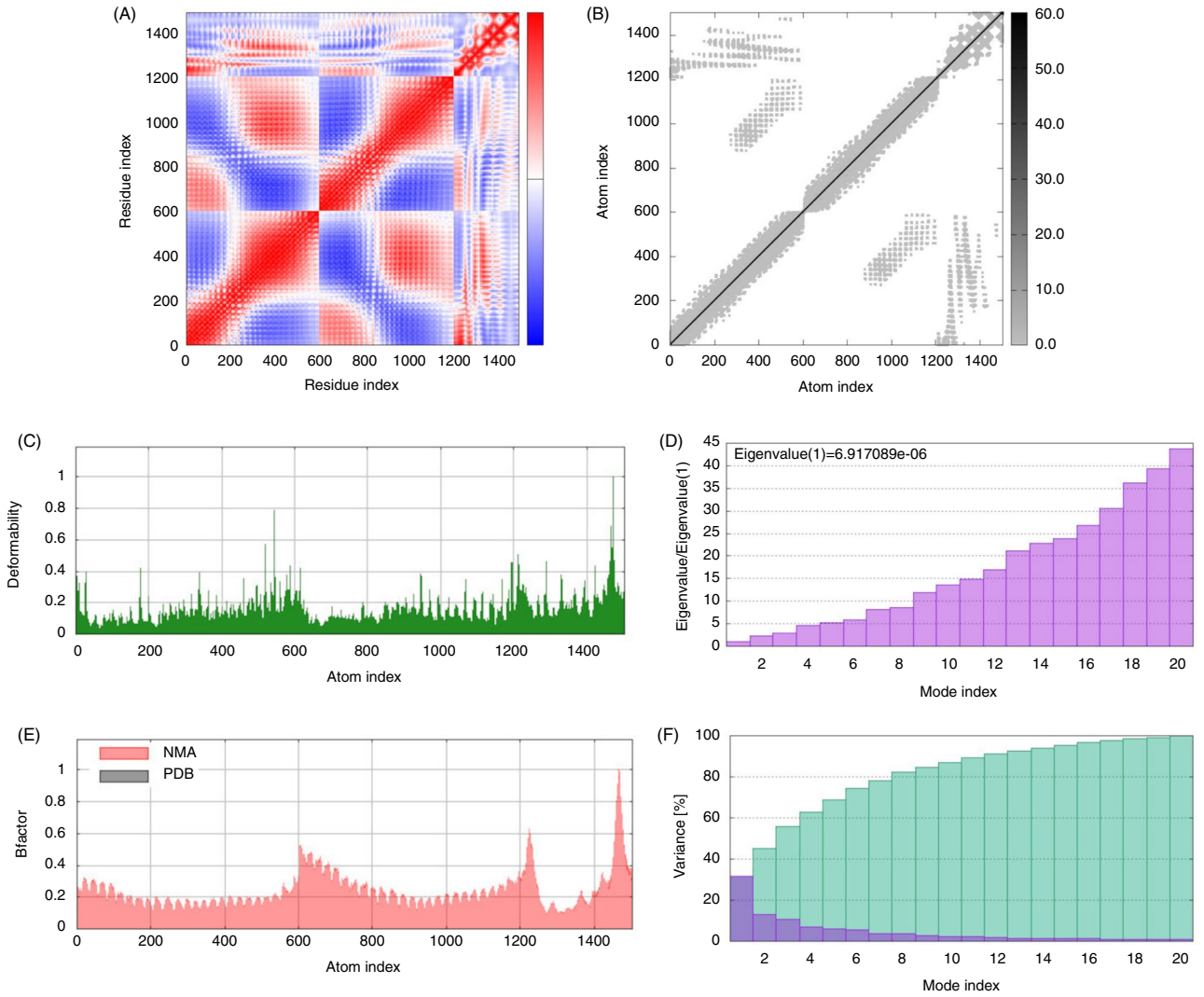


FIGURE 10 | Results of MD simulation. (A) Residue index plot, (B) atom index plot, (C) deformability, (D) eigenvalue plot, (E) Bfactor plot and (F) variance plot.

optimisation, the CAI value increased to 0.98 and %GC content decreased to 47.19% (Table 6).

This optimised codon sequence was used for the in silico cloning by SnapGene software, where *Xho*I (158) and *Mlu*I (1035) restriction enzyme sites were selected for the insertion of the codon sequence into the pET-28a(+) plasmid (which is one of the most commonly used vectors and has the size of 5265 bp). As these sites were not present in the nucleotide sequence of the construct, at the start and end of the nucleotide sequence, restriction sites were added for the insertion of the codon sequence into the plasmid. The codon sequence inserted was about 879 bp to make a recombinant plasmid (Figure 12).

3.16 | ORF of Nucleotide Sequence

ORF-finder from the NCBI website gave a total of 16 different ORFs from the codon sequence of the constructed vaccine. The first ORF was from the start to the end of the sequence, which shows that the expression of the protein sequence from the obtained codon sequence is possible when expressed. Figure 13 shows the ORFs from the codon sequence.

4 | Discussion

The immunoinformatics approach to designing novel vaccines against *S. pneumoniae* has been utilised by many researchers [17, 20–25]. Unlike available polysaccharide and conjugate vaccines, the development of a vaccine that can provide protection against diverse serotypes could be an interesting and viable approach. Such approaches will not only help in cost reduction but also in serotype-independent protection [7]. In this study, after a literature review, 14 proteins (Table S2) were identified at the surface of *S. pneumoniae*. All these proteins have < 30% homology with human and mouse proteins as per the BLAST analysis, which means that these proteins are not similar to any human or mouse proteins. This analysis is necessary as homologous proteins used as vaccines can initiate an autoimmune response that can result in any acute to severe adverse effects in a healthy body. For the construction of a vaccine, the sequences of amino acids

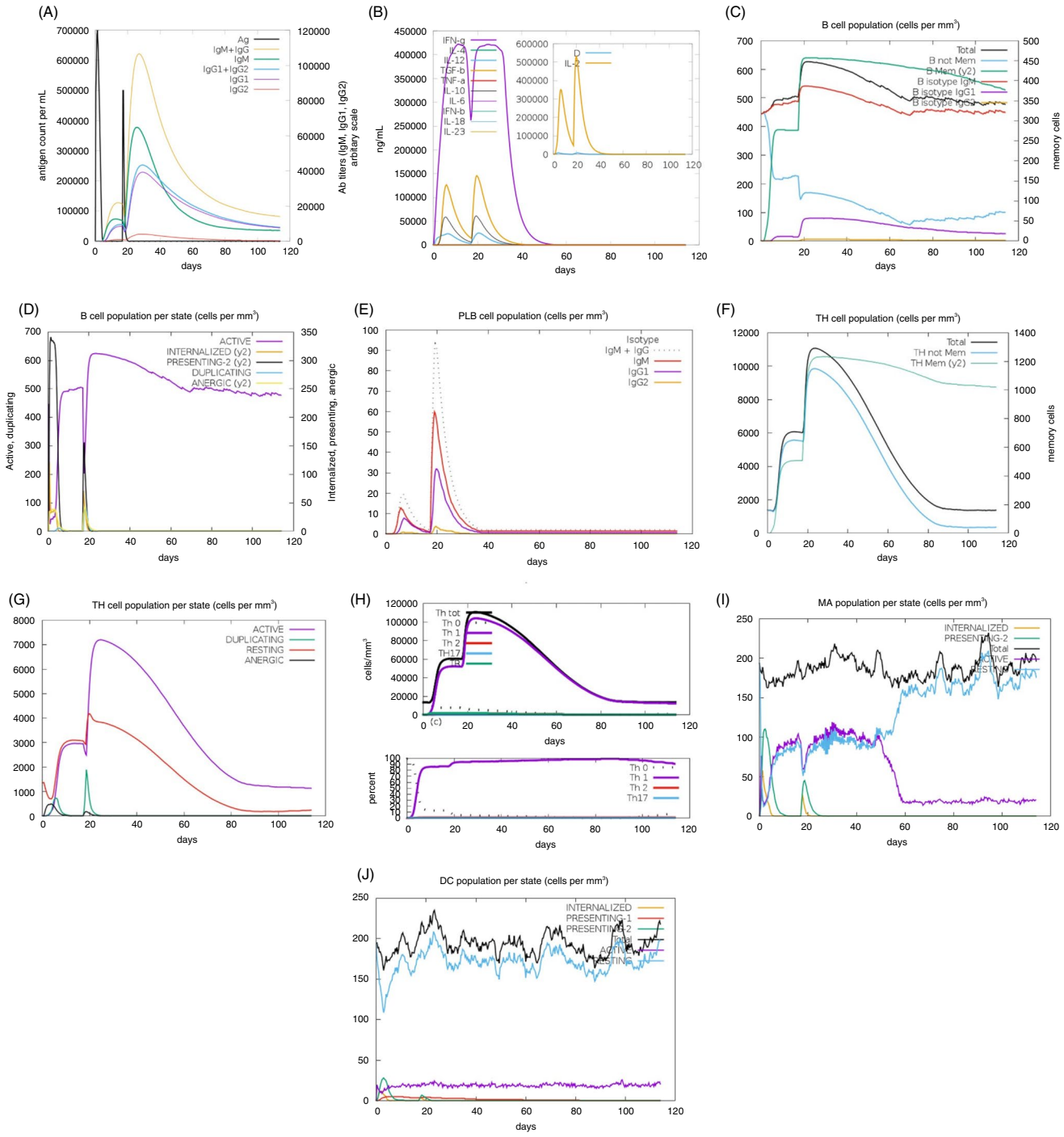


FIGURE 11 | Immune simulation showcasing the stimulation of immune system after first immunisation on Day 0 and after booster dosing on Day 18. (A) Amount of antibodies, (B) cytokines secretion, (C) B-cell population, (D) B-cell population per state, (E) plasma B-cell population, (F) Th-cell population, (G) Th-cell population per state, (H) Th cell type, (I) macrophage population and (J) dendritic cell population.

for all non-homologous proteins were retrieved from the NCBI and UniProt websites. Also, all these proteins are non-allergenic, which was predicted by using the AllerTOP v2.0 web server.

The selected BCL and CTL epitopes have an antigenicity score of more than 1 and are non-allergenic. For the HTL epitope selection, the antigenicity score was kept at 0.7, and these were further evaluated for the IFN-γ inducing ability. Based on this analysis, a total of 15 epitopes were chosen to include in the vaccine construct, which are enlisted in Tables 1–3. By incorporating these epitopes in the construct, the probability of induction of immunity increases as these epitopes were predicted to bind and activate the B and T cells with a higher threshold. In addition, these epitopes were chosen from the proteins that are expressed in diverse serotypes, and many of them fall into the conserved category, which suggests that the designed vaccine construct could be effective against diverse pneumococcal strains. Further, population coverage prediction for MHC-II epitopes showed 85.76% coverage (Figure 2), which indicates that MHC-II peptides incorporated into the vaccine would be able to interact with different types of HLA-DRB alleles expressed

TABLE 5 | Codon sequence of vaccine obtained from the Backtranseq web server with its CAI value and %GC content.

Codon sequence	CAI value	GC content
GGCATTATTAACACCCTGCAGAAATATTATTGCCGCGTGCGCGGCGGCCGCTGCGCG GTGCTGAGCTGCCTGCCGAAAGAAGAACAGATTGGCAAATGCAGCACCCGCGGCCG CAAATGCTGCCGCCGCAAAAAAGAAGCGGCGGCGAAAGGCACCGGCGCGACCATT ACCGGCTATAAAAAAACCACCAGCCTGAACATGCTGATGCTGAAAAAAACCAGCA AAGATCCGCGCGCGAACTATAAAAAACGCACCGATCGCATTGGCATTAACTATAAA AAAGTGACCGAAGTGATTGCGTATCGCTTTGGCGGCGGCAGCTATGAACTGTTTAA ACCGGAAGAAGGCGTGGCGAAAAAAGAAGGCGGCGGCGGCAGCAAAGAATTTCA GAACACCGCGAGCAACGAACTGACCACCTATGATGGCGGCGGCAGCACCACCAGCC AGAGCCTGAGCCAGAGCAAAAGCCTGAGCGTGAGCGGCGGCGGCAGCACCGCGCA GATTCTGGGCCTGGATGTGGAAAAAGAACAGAAAGAAGGCGGCGGCAGCGAAGCG GATTATCATTGGCGCAAAGATCCGGAACTGGGCTTTTTTAAAAAACAGCCGAGCCC GCAGAGCACCCCGGAACCGAGCCCGAGCCTGCAGCCGAAAAAAAAAAAAAAAGC GGAAGATCAGAAAGAAGAAGATCGCCGCAACTATCCGAAAAAAGGCAGCACCTGG TATTATCTGAACGCGAGCAACGGCGATATGAAAACCAAAAAAGAAAAAAAAGCG AAAGATCAGAAAGAAGAAGATCGCCGCAACTATCCGAAAAAAGGCGTGAACGCGG ATACCAGCCTGGATGATCTGTATCTGGATGAAGCGGCGGCGAAA	0.6505	53.15

TABLE 6 | Optimised codon sequence with its improved CAI value and %GC content.

Codon sequence	CAI value	GC content
GGCATTATTAACACCCTGCAGAAATATTATTGCCGCGTGCGCGGCGGCCGCTGCG CGGTGCTGAGCTGCCTGCCGAAAGAAGAACAGATTGGCAAATGCAGCACCCGCG GCCGCAAATGCTGCCGCCGCAAAAAAGAAGCGGCGGCGAAAGATGCGGATAGC GAAGGCTGGAAATTTAAAAAAGTGAGCAGCATTGTGTTTACCCTGTTTAAAAAA AGCAGCGAAGAACTGGCGGGCGGCTATAAAAAAGCGGTGAACACCAGCATTGAT GGCTATAAAAAAACCGCGGCGGTGATTCTGGCGGCGTATGGCGGCGGCAGCATTA GCAGCTATGGCAGCAACAACAGCAGCACCGTGGGCTGGAAAGGCGGCGGCAGCA GCGCGACCGTGAACGTGTATGGCAACAAAGATGGCAAACCGGATGGGGGGGGCA GCCGCTTCAAAATAAAAACCGATAACAAAGTGGGCATTGCGAAAGTGGGCGGCG GCAGCCTGAAAGTGATTAGCGTGACCAGCGTGAACCCGGGCGAAGGCAAAGGCGG CGGCAGCCCGAGCACCGTGGGCACCACCGAAGCGGCGAAACTGGGCATGGTGAAA AAAACCGATCATCAGGATAGCGGCAACACCGAAGCGAAAGGCGCGGAAGCGAAAA AAGGCCTGACCCCGCCGAGCACCGATCATCAGGATAGCGGCAACACCGAAAAAAA AAACAGCACCAGCAACAGCACCCTGGAAGAAGTGCCGACCGTGGATCCGAAAAAAG AAGGCAGCACCCAGGCGGCGACCAGCAGCAACATGGCGAAAACCGAAAAAAAAGA TATGGATATTCAGCTGTATAACTATGAAAACGAAACCGATAGCAGCGAAGCGGCGGCGAAA	0.98	47.19%

by 85.76% of the world population. For the construction of the vaccine using different epitopes, CTL and BCL epitopes were linked using 'KK' linkers, while for HTL epitopes, 'GGGS' linkers were utilised, and β-defensin (an adjuvant) was incorporated into the vaccine to increase the immunogenicity. To link the adjuvant with the first epitope and for better stability of the construct, the 'EAAAK' linker was added at the corresponding position of the construct. The vaccine construct (Figure 4B) has 291 amino acids, a stable structure, predicted solubility in water, a non-allergenic nature and a 1.2 antigenicity score. All these physiochemical parameters were determined using Expasy's ProtParam, Protein-Sol, AllerTOP and Vaxijen servers Figure 4D.

The predicted secondary structure of the construct includes mainly three types of amino acid arrangement: strands (18.56%), helix (19.93%) and coils (61.51%), where the number of coils in the structure was highest. Additionally, the tertiary structure was evaluated by Ramachandran plot; it presented 87% of amino acids in the most favoured region, which was increased to 97.2% after refinement of the structure. Also, the *z*-score for the refined structure was −3.2, which suggests the good stability of the structure (Figure 6C).

The presence of discontinuous B-cell peptides in the vaccine construct suggests that the construct has the ability to engage the B cells and thus can improve the overall immunogenicity. The actual ability of a vaccine to stimulate the immune system can be measured by its affinity to bind with some important receptors like TLR-2, TLR-4, MHC-1 and MHC-2. Docking studies with these receptors show the good binding affinity of the construct (Figure 8E) as well as interactions of amino acids in complex with the TLR-4 receptor showing 32 hydrogen bonds and 4 salt bridges formation

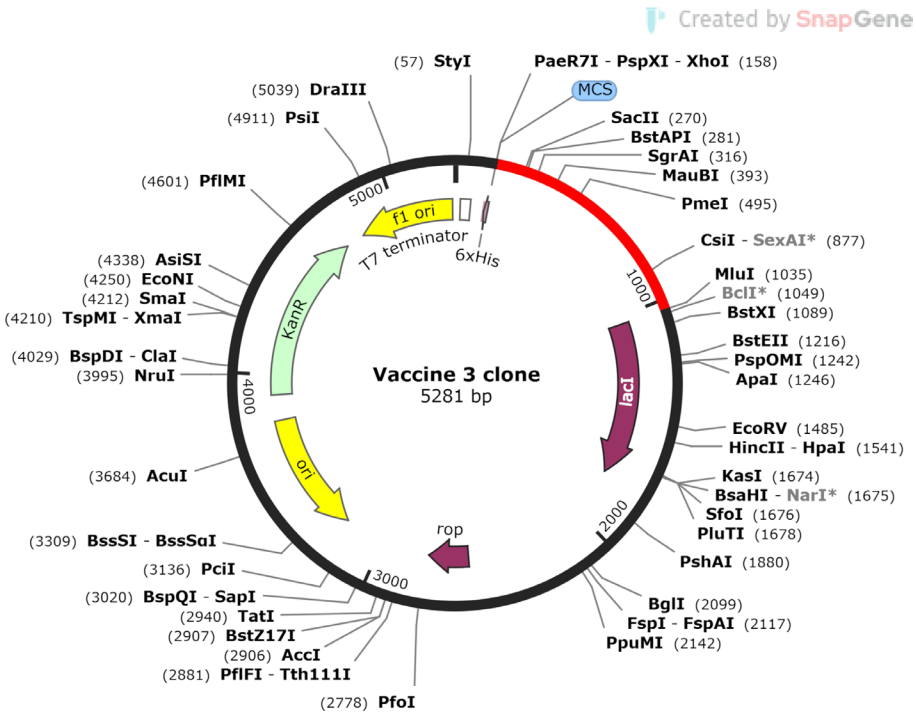


FIGURE 12 | In silico cloning with pET-28a(+) vector. The red part is inserted vaccine sequence at the restriction sites *Xho*I and *Mlu*I.

FIGURE 13 | ORFs of codon sequence where given sequence of 879 nucleotides is acting as one ORF from start to end, which is highlighted with blue colour.

(Table 4), which provides the idea about good interaction of the receptor–ligand complex. Further analysis of this complex by MD simulation revealed the eigenvalue of 6.91×10^{-6}, which describes the flexibility of the complex, where the bigger the eigenvalue, the larger the fluctuation amplitude [102]. Also, in the residue index plot, which shows the contact of residues during simulation, the

higher amount of red colour in Figure 10A suggests better residue interaction. The atom index plot shows the connection of atoms and springs, where an increase in darker areas suggests a higher stiffness of the structure (Figure 10B).

The immune stimulation of the vaccine shows the induction of humoral and cellular immunity. Increased amounts of antigen-specific IgM, IgG and IgM + IgG levels after the first immunisation as well as after the booster dose suggest the plasma and memory B-cell generation. Further, a higher level of IgG1 is an indication of Th2-driven immune responses; in addition, Th1-type cells can also be observed in Figure 11H. Other types of T-cell populations in the active state and the T-memory cells increased and maintained the levels for almost up to 80 days in simulation, which indicates the ability of the vaccine construct to create the memory as well as the cytotoxic response against the antigen Figure 11F,G. Plasma B-cell secreting IgM and IgG1 antibody-secreting cells can be observed in Figure 11E, and other isotypes like memory cells were also seen to increase (Figure 11C), which suggests the induction of a long-lasting memory response by the vaccine construct. Cytokine secretion and activation of DCs as well as MAs show the induction of cellular immunity.

The optimisation of codon sequence for the production of vaccine construct using suitable vector (here *E. coli*) was performed. In silico cloning using the restriction enzyme technique showed that the vaccine construct can be converted and cloned by expressing in *E. coli*. For that, the cloned pET-28a (+) vector in which the codon sequence of the vaccine was inserted is given in Figure 12. Figure 13 represents the possible ORFs from the codon sequence, and the ORF1 is for the whole length of the sequence, which means the gene expression of this codon sequence is possible from start to end.

5 | Conclusion

The meticulous design and computational analysis of the vaccine construct, incorporating 15 carefully selected non-allergic epitopes from diverse serotypes, showcases a promising potential candidate vaccine which can be effective against diverse serotypes. The structural refinement, molecular docking with TLR-4, and MD simulation gave information regarding the construct's stability and interaction dynamics. The results of in silico cloning with the pET-28a (+) plasmid suggest a feasible production method through *E. coli* expression. Furthermore, the immune simulation predicts the induction of a robust immune response, encompassing humoral, cellular and innate immunity. While these findings are encouraging, rigorous experimental analysis is imperative to ascertain the vaccine's efficacy and safety in real-world applications.

Author Contributions

H.H.C. performed the data curation, research experimentation, manuscript writing and data analysis. S.N. performed the data curation. J.M. designed the study, analysed the data and edited the manuscript.

Acknowledgements

We thank the Indian Institute of Technology (Banaras Hindu University), Varanasi, and paramshivay super computer facility, IIT-BHU, Varanasi, India for providing the necessary facilities for the current work. We also thank the Anusandhan National Research Foundation for providing the research grant for this work.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All the data used to draw the results and conclusions are available in the manuscript or available at the different web servers cited in the manuscript.

References

1. M. Masomian, Z. Ahmad, L. Ti Gew, and C. L. Poh, "Development of Next Generation *Streptococcus Pneumoniae* Vaccines Conferring Broad Protection," *Vaccine* 8, no. 1 (2020): 132.

2. WHO, *Pneumonia in Children* (World Health Organization, 2022), https://www.who.int/news-room/fact-sheets/detail/pneumonia.

3. Centers for Disease Control and Prevention, "Pneumococcal Diseases: Surveillance and Reporting,", https://www.cdc.gov/abcs/bact-facts/data-dashboard.html.

4. D. Briles, J. Paton, R. Mukerji, E. Swiatlo, and M. Crain, "Pneumococcal Vaccines," *Microbiology Spectrum* 7, no. 6 (2019): 2.

5. A. De Roux and H. Lode, "Pneumococcal Vaccination," *European Respiratory Journal* 26 (2005): 982–983.

6. E. Swiatlo and D. Ware, "Novel Vaccine Strategies With Protein Antigens of *Streptococcus Pneumoniae*," *FEMS Immunology and Medical Microbiology* 38, no. 1 (2003): 1–7.

7. S. Tereziu and D. A. Minter, "Pneumococcal Vaccine," 2018.

8. S. Nagaraj, B. S. Kalal, A. Manoharan, and A. Shet, "*Streptococcus Pneumoniae* Serotype Prevalence and Antibiotic Resistance Among Young Children With Invasive Pneumococcal Disease: Experience From a Tertiary Care Center in South India," *Germs* 7, no. 2 (2017): 78–85.

9. D. Hurley, C. Griffin, M. Young, Jr., et al., "Safety, Tolerability, and Immunogenicity of a 20-Valent Pneumococcal Conjugate Vaccine (PCV20) in Adults 60 to 64 Years of Age," *Clinical Infectious Diseases* 73, no. 7 (2021): e1489–e1497.

10. H. L. Stacey, J. Rosen, J. T. Peterson, et al., "Safety and Immunogenicity of 15-Valent Pneumococcal Conjugate Vaccine (PCV-15) Compared to PCV-13 in Healthy Older Adults," *Human Vaccines & Immunotherapeutics* 15, no. 3 (2019): 530–539.

11. M. Mirsaeidi and D. E. Schraufnagel, "Pneumococcal Vaccines: Understanding Centers for Disease Control and Prevention Recommendations," *Annals of the American Thoracic Society* 11, no. 6 (2014): 980–985.

12. S. J. Vandecasteele, S. Ombelet, S. Blumental, and W. E. Peetermans, "The ABC of Pneumococcal Infections and Vaccination in Patients With Chronic Kidney Disease," *Clinical Kidney Journal* 8, no. 3 (2015): 318–324.

13. J. Grabenstein and K. Klugman, "A Century of Pneumococcal Vaccination Research in Humans," *Clinical Microbiology and Infection* 18 (2012): 15–24.

14. D. M. Musher, R. Anderson, and C. Feldman, "The Remarkable History of Pneumococcal Vaccination: An Ongoing Challenge," *Pneumonia* 14, no. 1 (2022): 1–15.

15. R. J. Malonis, J. R. Lai, and O. Vergnolle, "Peptide-Based Vaccines: Current Progress and Future Challenges," *Chemical Reviews* 120, no. 6 (2019): 3210–3229.

16. K. Moffitt and R. Malley, "Rationale and Prospects for Novel Pneumococcal Vaccines," *Human Vaccines & Immunotherapeutics* 12, no. 2 (2016): 383–392.

17. S. S. Rawat, A. K. Keshri, R. Kaur, and A. Prasad, "Immunoinformatics Approaches for Vaccine Design: A Fast and Secure Strategy for Successful Vaccine Development," *Vaccines (Basel)* 11, no. 2 (2023): 221.

18. A. N. Oli, W. O. Obialor, M. O. Ifeanyichukwu, et al., "Immunoinformatics and Vaccine Development: An Overview," *ImmunoTargets and Therapy* 9 (2020): 13–30.

19. R. Rappuoli, "Reverse Vaccinology," *Current Opinion in Microbiology* 3, no. 5 (2000): 445–450.

20. E. Afshari, R. A. Cohan, F. Sotoodehnejadnematalahi, and S. F. Mousavi, "In-Silico Design and Evaluation of an Epitope-Based Serotype-Independent Promising Vaccine Candidate for Highly Cross-Reactive Regions of Pneumococcal Surface Protein A," *Journal of Translational Medicine* 21, no. 1 (2023): 13.

21. Z. Bahadori, M. Shafaghi, H. Madanchi, M. M. Ranjbar, A. A. Shabani, and S. F. Mousavi, "In Silico Designing of a Novel Epitope-Based Candidate Vaccine Against *Streptococcus Pneumoniae* With Introduction of a New Domain of PepO as Adjuvant," *Journal of Translational Medicine* 20, no. 1 (2022): 389.

22. L. Mazumder, M. Shahab, S. Islam, et al., "An Immunoinformatics Approach to Epitope-Based Vaccine Design Against PspA in *Streptococcus Pneumoniae*," *Journal of Genetic Engineering and Biotechnology* 21, no. 1 (2023): 57.

23. M. Munia, S. Mahmud, M. Mohasin, and K. K. Kibria, "In Silico Design of an Epitope-Based Vaccine Against Choline Binding Protein A of *Streptococcus Pneumoniae*," *Informatics in Medicine Unlocked* 23 (2021): 100546.

24. M. Shafaghi, Z. Bahadori, H. Madanchi, M. M. Ranjbar, A. A. Shabani, and S. F. Mousavi, "Immunoinformatics-Aided Design of a New Multi-Epitope Vaccine Adjuvanted With Domain 4 of Pneumolysin Against *Streptococcus Pneumoniae* Strains," *BMC Bioinformatics* 24, no. 1 (2023): 1–27.

25. S. Tarahomjoo and S. Ghaderi, "In Silico Design of a Novel Serotype Independent Vaccine Against *Streptococcus Pneumoniae* Based on B-Cell Epitope Regions of Fibronectin Binding Protein, Choline Binding Protein D, and D-Alanyl-D-Alanine Carboxypeptidase," *Letters in Drug Design & Discovery* 16, no. 4 (2019): 372–381.

26. M. Nahian, M. Shahab, M. R. Khan, et al., "Development of a Broad-Spectrum Epitope-Based Vaccine Against *Streptococcus Pneumoniae*," *PLoS One* 20, no. 1 (2025): e0317216.

27. J. Aceil and F. Y. Avci, "Pneumococcal Surface Proteins as Virulence Factors, Immunogens, and Conserved Vaccine Targets," *Frontiers in Cellular and Infection Microbiology* 12 (2022): 832254.

28. C. Morsczeck, T. Prokhorova, J. Sigh, et al., "*Streptococcus Pneumoniae*: Proteomics of Surface Proteins for Vaccine Development," *Clinical Microbiology and Infection* 14, no. 1 (2008): 74–81.

29. I. Pérez-Dorado, S. Galan-Bartual, and J. Hermoso, "Pneumococcal Surface Proteins: When the Whole Is Greater Than the Sum of Its Parts," *Molecular Oral Microbiology* 27, no. 4 (2012): 221–245.

30. S. Talukdar, S. Zutshi, K. Prashanth, K. K. Saikia, and P. Kumar, "Identification of Potential Vaccine Candidates Against *Streptococcus Pneumoniae* by Reverse Vaccinology Approach," *Applied Biochemistry and Biotechnology* 172 (2014): 3026–3041.

31. K. K. Gosink, E. R. Mann, C. Guglielmo, E. I. Tuomanen, and H. R. Masure, "Role of Novel Choline Binding Proteins in Virulence of *Streptococcus Pneumoniae*," *Infection and Immunity* 68, no. 10 (2000): 5690–5695.

32. J. Brown, S. Hammerschmidt, and C. Orihuela, *Streptococcus Pneumoniae: Molecular Mechanisms of Host-Pathogen Interactions* (Academic Press, 2015).

33. Uniprot, https://www.uniprot.org/.

34. National Center for Biotechnology Information, National Library of Medicine, https://www.ncbi.nlm.nih.gov/.

35. E. W. Sayers, E. E. Bolton, J. R. Brister, et al., "Database Resources of the National Center for Biotechnology Information," *Nucleic Acids Research* 50, no. D1 (2022): D20.

36. RCBS Protein Data Bank, https://www.rcsb.org/.

37. J. A. Owen, J. Punt, S. A. Stranford, and P. P. Jones, *Kuby Immunology* (WH Freeman, 2013).

38. S. Saha and G. P. Raghava, "Prediction Methods for B-Cell Epitopes," in *Imunoinformatics: Predicting Immunogenicity In Silico* (Humana Press, 2007), 387–394.

39. S. Saha and G. P. S. Raghava, "Prediction of Continuous B-Cell Epitopes in an Antigen Using Recurrent Neural Network," *Proteins: Structure, Function, and Bioinformatics* 65, no. 1 (2006): 40–48.

40. J. Alexander, J. Fikes, S. Hoffman, et al., "The Optimization of Helper T Lymphocyte (HTL) Function in Vaccine Development," *Immunologic Research* 18 (1998): 79–92.

41. B. Reynisson, B. Alvarez, S. Paul, B. Peters, and M. Nielsen, "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 Research* 48, no. W1 (2020): W449–W454.

42. P. Wang, J. Sidney, Y. Kim, et al., "Peptide Binding Predictions for HLA DR, DP and DQ Molecules," *BMC Bioinformatics* 11, no. 1 (2010): 1–12, https://doi.org/10.1186/1471-2105-11-568.

43. K. K. Jensen, M. Andreatta, P. Marcatili, et al., "Improved Methods for Predicting Peptide Binding Affinity to MHC Class II Molecules," *Immunology* 154, no. 3 (2018): 394–406.

44. M. Andreatta, E. Karosiene, M. Rasmussen, A. Stryhn, S. Buus, and M. Nielsen, "Accurate Pan-Specific Prediction of Peptide-MHC Class II Binding Affinity With Improved Binding Core Identification," *Immunogenetics* 67 (2015): 641–650.

45. H. H. Bui, J. Sidney, K. Dinh, S. Southwood, M. J. Newman, and A. Sette, "Predicting Population Coverage of T-Cell Epitope-Based Diagnostics and Vaccines," *BMC Bioinformatics* 7 (2006): 153.

46. J. K. Actor, *Chapter 4—T-Cell Immunity* (Elsevier, 2012).

47. M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, and M. Nielsen, "Large-Scale Validation of Methods for Cytotoxic T-Lymphocyte Epitope Prediction," *BMC Bioinformatics* 8 (2007): 1–12.

48. M. Nielsen, C. Lundegaard, P. Worning, et al., "Reliable Prediction of T-Cell Epitopes Using Neural Networks With Novel Sequence Representations," *Protein Science* 12, no. 5 (2003): 1007–1017.

49. B. Peters, S. Bulik, R. Tampe, P. M. Van Endert, and H.-G. Holzhutter, "Identifying MHC Class I Epitopes by Predicting the TAP Transport Efficiency of Epitope Precursors," *Journal of Immunology* 171, no. 4 (2003): 1741–1749.

50. M. Thomsen, C. Lundegaard, S. Buus, O. Lund, and M. Nielsen, "MHCcluster, a Method for Functional Clustering of MHC Molecules," *Immunogenetics* 65 (2013): 655–665.

51. I. Dimitrov, I. Bangov, D. R. Flower, and I. Doytchinova, "AllerTOP v.2—A Server for In Silico Prediction of Allergens," *Journal of Molecular Modeling* 20 (2014): 1–6.

52. I. A. Doytchinova and D. R. Flower, "VaxiJen: A Server for Prediction of Protective Antigens, Tumour Antigens and Subunit Vaccines," *BMC Bioinformatics* 8, no. 1 (2007): 1–7.

53. K. Rawal, R. Sinha, B. A. Abbasi, et al., "Identification of Vaccine Targets in Pathogens and Design of a Vaccine Using Computational Approaches," *Scientific Reports* 11, no. 1 (2021): 17626.

54. E. Gasteiger, C. Hoogland, A. Gattiker, et al., *Protein Identification and Analysis Tools on the ExPASy Server* (Springer, 2005).

55. M. Hebditch, M. A. Carballo-Amador, S. Charonis, R. Curtis, and J. Warwicker, "Protein–Sol: A Web Tool for Predicting Protein Solubility From Sequence," *Bioinformatics* 33, no. 19 (2017): 3098–3100.

56. D. T. Jones, "Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices," *Journal of Molecular Biology* 292, no. 2 (1999): 195–202.

57. D. W. Buchan and D. T. Jones, "The PSIPRED Protein Analysis Workbench: 20 Years on," *Nucleic Acids Research* 47, no. W1 (2019): W402–W407.

58. J. Yang and Y. Zhang, "I-TASSER Server: New Development for Protein Structure and Function Predictions," *Nucleic Acids Research* 43, no. W1 (2015): W174–W181.

59. G. R. Lee, L. Heo, and C. Seok, "Effective Protein Model Structure Refinement by Loop Modeling and Overall Relaxation," *Proteins: Structure, Function, and Bioinformatics* 84 (2016): 293–301.

60. L. Heo, H. Park, and C. Seok, "GalaxyRefine: Protein Structure Refinement Driven by Side-Chain Repacking," *Nucleic Acids Research* 41, no. W1 (2013): W384–W388.

61. C. Colovos and T. O. Yeates, "Verification of Protein Structures: Patterns of Nonbonded Atomic Interactions," *Protein Science* 2, no. 9 (1993): 1511–1519.

62. R. A. Laskowski, M. W. MacArthur, D. S. Moss, and J. M. Thornton, "PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures," *Journal of Applied Crystallography* 26, no. 2 (1993): 283–291.

63. O. V. Sobolev, P. V. Afonine, N. W. Moriarty, et al., "A Global Ramachandran Score Identifies Protein Structures With Unlikely Stereochemistry," *Structure* 28, no. 11 (2020): 1249–1258.e2.

64. M. Wiederstein and M. J. Sippl, "ProSA-Web: Interactive Web Service for the Recognition of Errors in Three-Dimensional Structures of Proteins," *Nucleic Acids Research* 35, no. Suppl_2 (2007): W407–W410.

65. M. J. Sippl, "Recognition of Errors in Three-Dimensional Structures of Proteins," *Proteins: Structure, Function, and Bioinformatics* 17, no. 4 (1993): 355–362.

66. S. Ferdous, S. Kelm, T. S. Baker, J. Shi, and A. C. Martin, "B-Cell Epitopes: Discontinuity and Conformational Analysis," *Molecular Immunology* 114 (2019): 643–650.

67. G. N. Sivalingam and A. J. Shepherd, "An Analysis of B-Cell Epitope Discontinuity," *Molecular Immunology* 51, no. 3–4 (2012): 304–309.

68. J. Ponomarenko, H.-H. Bui, W. Li, et al., "ElliPro: A New Structure-Based Tool for the Prediction of Antibody Epitopes," *BMC Bioinformatics* 9 (2008): 514, https://doi.org/10.1186/1471-2105-9-514.

69. D. Kozakov, D. R. Hall, B. Xia, et al., "The ClusPro Web Server for Protein–Protein Docking," *Nature Protocols* 12, no. 2 (2017): 255–278.

70. D. Kozakov, D. Beglov, T. Bohnuud, et al., "How Good Is Automated Protein Docking?," *Proteins: Structure, Function, and Bioinformatics* 81, no. 12 (2013): 2159–2166, https://doi.org/10.1002/prot.24403.

71. S. Vajda, C. Yueh, D. Beglov, et al., "New Additions to the C Lus P Ro Server Motivated by CAPRI," *Proteins: Structure, Function, and Bioinformatics* 85, no. 3 (2017): 435–444, https://doi.org/10.1002/prot.25219.

72. I. T. Desta, K. A. Porter, B. Xia, D. Kozakov, and S. Vajda, "Performance and Its Limits in Rigid Body Protein-Protein Docking," *Structure* 28, no. 9 (2020): 1071–1081.e3.

73. A. Vangone and A. M. Bonvin, "Contacts-Based Prediction of Binding Affinity in Protein–Protein Complexes," *eLife* 4 (2015): e07454.

74. L. C. Xue, J. P. Rodrigues, P. L. Kastritis, A. M. Bonvin, and A. Vangone, "PRODIGY: A Web Server for Predicting the Binding Affinity of Protein–Protein Complexes," *Bioinformatics* 32, no. 23 (2016): 3676–3678.

75. A. C. Wallace, R. A. Laskowski, and J. M. Thornton, "LIGPLOT: A Program to Generate Schematic Diagrams of Protein-Ligand Interactions," *Protein Engineering, Design and Selection* 8, no. 2 (1995): 127–134.

76. R. Shukla and T. Tripathi, "Molecular Dynamics Simulation of Protein and Protein–Ligand Complexes," in *Computer-Aided Drug Design*, ed. D. B. Singh (Springer, 2020), 133–161.

77. J. R. López-Blanco, J. I. Aliaga, E. S. Quintana-Ortí, and P. Chacón, "iMODS: Internal Coordinates Normal Mode Analysis Server," *Nucleic Acids Research* 42, no. W1 (2014): W271–W276.

78. J. R. Lopéz-Blanco, J. I. Garzón, and P. Chacón, "iMod: Multipurpose Normal Mode Analysis in Internal Coordinates," *Bioinformatics* 27, no. 20 (2011): 2843–2850.

79. J. A. Kovacs, P. Chacón, and R. Abagyan, "Predictions of Protein Flexibility: First-Order Measures," *Proteins: Structure, Function, and Bioinformatics* 56, no. 4 (2004): 661–668.

80. M. Abraham, A. Alekseenko, V. Basov, et al., "GROMACS 2024.3 Manual," 2024.

81. M. Abraham, A. Alekseenko, V. Basov, et al., "GROMACS 2024.2 Source Code," Zenodo, 2024.

82. H. Bekker, H. Berendsen, E. Dijkstra, et al., "GROMACS—A Parallel Computer for Molecular-Dynamics Simulations," in *4th International Conference on Computational Physics (PC 92)* (World Scientific Publishing, 1993).

83. H. J. Berendsen, D. van der Spoel, and R. van Drunen, "GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation," *Computer Physics Communications* 91, no. 1–3 (1995): 43–56.

84. A. Bondi, "Van der Waals Volumes and Radii," *Journal of Physical Chemistry* 68, no. 3 (1964): 441–451.

85. S. Pronk, S. Páll, R. Schulz, et al., "GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit," *Bioinformatics* 29, no. 7 (2013): 845–854.

86. D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen, "GROMACS: Fast, Flexible, and Free," *Journal of Computational Chemistry* 26, no. 16 (2005): 1701–1718.

87. B. R. Brooks, C. L. Brooks, III, A. D. Mackerell, Jr., et al., "CHARMM: The Biomolecular Simulation Program," *Journal of Computational Chemistry* 30, no. 10 (2009): 1545–1614.

88. S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM," *Journal of Computational Chemistry* 29, no. 11 (2008): 1859–1865.

89. J. Lee, X. Cheng, J. M. Swails, et al., "CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field," *Journal of Chemical Theory and Computation* 12, no. 1 (2016): 405–413.

90. J. Lee, M. Hitzenberger, M. Rieger, N. R. Kern, M. Zacharias, and W. Im, "CHARMM-GUI Supports the Amber Force Fields," *Journal of Chemical Physics* 153, no. 3 (2020): 035103, https://doi.org/10.1063/5.0012280.

91. F. Castiglione, "C-ImmSim: A Model of Immunological Models."

92. P. Stolfi, F. Castiglione, E. Mastrostefano, et al., "In-Silico Evaluation of Adenoviral COVID-19 Vaccination Protocols: Assessment of Immunological Memory up to 6 Months After the Third Dose," *Frontiers in Immunology* 13 (2022): 998262.

93. N. Rapin, O. Lund, M. Bernaschi, and F. Castiglione, "Computational Immunology Meets Bioinformatics: The Use of Prediction Tools for Molecular Binding in the Simulation of the Immune System," *PLoS One* 5, no. 4 (2010): e9862.

94. C. Ragone, C. Manolio, B. Cavalluzzo, et al., "Identification and Validation of Viral Antigens Sharing Sequence and Structural Homology With Tumor-Associated Antigens (TAAs)," *Journal for Immunotherapy of Cancer* 9, no. 5 (2021): e002694, https://doi.org/10.1136/jitc-2021-002694.

95. P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," *Trends in Genetics* 16, no. 6 (2000): 276–277.

96. F. Madeira, N. Madhusoodanan, J. Lee, A. R. Tivey, and R. Lopez, "Using EMBL-EBI Services via Web Interface and Programmatically via Web Services," *Current Protocols in Bioinformatics* 66, no. 1 (2019): e74.

97. A. Grote, K. Hiller, M. Scheer, et al., "JCat: A Novel Tool to Adapt Codon Usage of a Target Gene to Its Potential Expression Host," *Nucleic Acids Research* 33, no. Suppl_2 (2005): W526–W531.

98. P. M. Sharp and W.-H. Li, "The Codon Adaptation Index—A Measure of Directional Synonymous Codon Usage Bias, and Its Potential Applications," *Nucleic Acids Research* 15, no. 3 (1987): 1281–1295.

99. G. L. Rosano and E. A. Ceccarelli, "Recombinant Protein Expression in *Escherichia coli*: Advances and Challenges," *Frontiers in Microbiology* 5 (2014): 172.

100. P. Sieber, M. Platzer, and S. Schuster, "The Definition of Open Reading Frame Revisited," *Trends in Genetics* 34, no. 3 (2018): 167–170.

101. S. H. Rangwala, A. Kuznetsov, V. Ananiev, et al., "Accessing NCBI Data Using the NCBI Sequence Viewer and Genome Data Viewer (GDV)," *Genome Research* 31, no. 1 (2021): 159–169.

102. N. Andrusier, E. Mashiach, R. Nussinov, and H. J. Wolfson, "Principles of Flexible Protein–Protein Docking," *Proteins: Structure, Function, and Bioinformatics* 73, no. 2 (2008): 271–289.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.