ACCSCIENCE PUBLISHING

Journal of Clinical and Translational Research

REVIEW ARTICLE

Machine learning and artificial intelligence for antigen discovery and immunogen design: A critical appraisal of computational strategies in medical microbiology

Shoaib Manzoor*

Department of Microbiology, University of Jhang, Jhang, Pakistan

Abstract

Background: Conventional vaccine development is slow, costly, and prone to attrition, with target antigen selection representing a major bottleneck. Machine learning (ML) and artificial intelligence (AI) are increasingly proposed to accelerate vaccine development. **Aim:** To critically evaluate ML- and AI-supported strategies for antigen discovery, epitope mapping, and immunogen engineering against bacterial, viral, fungal, and parasitic pathogens. **Methods:** Guided by PRISMA 2020, literature from January 2020 to April 2026 was retrieved from PubMed/MEDLINE, Web of Science, Scopus, and bioRxiv. Eligible studies described computational tools for antigen/epitope prediction or immunogen design, reported performance metrics, and, where relevant, experimental confirmation. Methods were grouped into five categories and assessed for architecture, accuracy, validation, and limitations. **Results:** Three developments shape the field: deep-learning epitope predictors (MUNIS, ROC-AUC ≈0.980; GraphBepi); AlphaFold 3-based modeling of TCR-pMHC and antibody-antigen complexes; and generative platforms (RFdiffusion, ProteinMPNN, MESODID) yielding experimentally verified antibodies and multiepitope scaffolds. Integrated pipelines (PLGDL, VaxSeer) and safety/developability filters permit end-to-end candidate ranking. Obstacles include sparse, demographically biased training data, limited conformational epitope accuracy, neglected glycosylation, and frequent laboratory failure of designed proteins. **Conclusion:** AI methods have advanced from concepts to experimentally supported tools that can accelerate antigen identification. Translation into licensed vaccines will require repeated laboratory validation, standardized reporting, and training sets reflecting population diversity. **Relevance for Patients:** By compressing antigen selection, these strategies may bring vaccines against endemic, emerging, and refractory pathogens to clinical evaluation sooner, while attention to dataset bias and *in silico* safety screening can promote equitable and safe products.

Keywords: Machine learning; Antigen discovery; Epitope prediction; Vaccine design; Deep learning; AlphaFold; Generative protein design; Computational vaccinology

***Corresponding author:**
Shoaib Manzoor
(shoaibmanzoor064@gmail.com)

Citation: Manzoor S. Machine learning and artificial intelligence for antigen discovery and immunogen design: A critical appraisal of computational strategies in medical microbiology. *J Clin Transl Res.* doi: 10.36922/JCTR026280066

Received: July 8, 2026

Revised: August 21, 2026

Accepted: September 23, 2026

Published online: October 12, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons AttributionNonCommercial 4.0 International (CC BY-NC 4.0), which permits all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Vaccination continues to be the most effective public-health measure against infectious disease. Nonetheless, the path from conceptual design to regulatory licensure typically spans approximately a decade, incurs costs approaching US$ 900 million, and sees more than 90% of candidates eliminated between preclinical evaluation and approval.[1] Identification of an appropriate antigenic target among the thousands of proteins encoded by a given pathogen is widely regarded as the principal rate-limiting step.

A decisive methodological shift occurred in the early 2000s with the introduction of reverse vaccinology, which enabled *in silico* interrogation of entire microbial genomes rather than reliance on empirical antigen screening.[2] The seminal demonstration involved *Neisseria meningitidis* serogroup B: hundreds of surface-exposed proteins were nominated computationally and 350 were advanced to murine immunogenicity testing—a throughput unattainable by classical biochemical methods.[3] That work established that computation could expand the candidate search space and uncover targets overlooked by traditional approaches.

In the subsequent two decades, machine learning and artificial intelligence have progressively occupied a central position in computational vaccinology.[4,5] Neural-network predictors of T-cell epitopes, exemplified by the NetMHC series, were among the earliest tools adopted.[6,7] Deep-learning architectures later produced substantial gains in accuracy and throughput,[8] and contemporary AI systems are now positioned to influence nearly every stage of vaccine development, from antigen selection and immune-response forecasting to manufacturing and trial design.[9]

The present review provides a systematic and critical appraisal of ML- and AI-based methods for antigen discovery and immunogen design, with particular attention to applications relevant to bacterial, viral, fungal, and parasitic pathogens. The scope extends beyond immunogenicity prediction to the downstream computational assessments that determine whether a nominated molecule is suitable for further development, including toxicity, allergenicity, inflammatory potential, solubility, stability, half-life, and overall developability. Emphasis is placed on named tools, their architectural features, published performance metrics, documented limitations, and validation requirements, thereby furnishing a practical reference for investigators intending to incorporate these methods into their workflows.

2. Review methodology

2.1. Literature search and screening

The review adhered to the PRISMA 2020 reporting guideline.[10] Four databases—PubMed/MEDLINE, Web of Science, Scopus, and bioRxiv—were queried for records published between January 2020 and April 2026. The primary Boolean string combined terms for machine learning, deep learning, or artificial intelligence with terms for epitope prediction, antigen discovery, vaccine design, or immunogen design, and further restricted results to studies reporting experimental validation or quantitative performance. Supplementary searches targeted specific tools and platforms identified during screening.

2.2. Inclusion and exclusion criteria

Studies were included if they described a computational method for antigen, epitope, or immunogen prediction/design; reported quantitative performance metrics (*e.g.*, AUC, accuracy, F1); and, where applicable, included experimental validation. Reviews, opinion pieces, and studies without performance metrics were excluded. Articles published before January 2020 were excluded to focus on recent advances.

2.3. Data extraction and synthesis

Data were extracted on tool architecture, input/output modalities, reported performance, validation status, and limitations. Tools were grouped into five categories: (i) T-cell epitope prediction, (ii) B-cell epitope prediction, (iii) integrated antigen prediction frameworks, (iv) safety and physicochemical profiling, and (v) generative protein design and optimization.

3. Computational strategies for antigen discovery

3.1. T-cell epitope prediction

T-cell epitope prediction has advanced considerably with deep learning. Early neural-network methods such as NetMHCpan 4.1 achieve AUC values below 0.90 and exhibit Eurocentric bias in training data.[11,12] MUNIS addresses many of these limitations through an expanded training corpus and a dual-task architecture.

Additional recent contributions include deepAntigen, a graph-convolutional framework that predicts antigen-HLA binding and antigen-TCR interactions at atomic resolution and recovers mutation-dependent alterations in T-cell responses,[13] and oncology-oriented pipelines such as CNNeoPP for neoantigen discovery.[14]

3.2. B-cell epitope prediction

B-cell epitope prediction is complicated by the predominance of conformational determinants, which are defined by three-dimensional structure rather than contiguous sequence. Accurate residue-level localization of antibody-binding sites is therefore essential for both mechanistic immunology and therapeutic antibody development.[15]

GraphBepi represents a substantial advance in this domain. The method converts AlphaFold2-predicted structures into residue graphs whose nodes carry ESM-2 embeddings; on curated benchmarks it improved AUC by more than 5.5% and AUPR by 44% relative to prior state-of-the-art approaches.[16] Experimental mutation of residues nominated by the model reduced antibody binding, supporting the biological relevance of the predictions. Complementary work demonstrated that pairing AlphaFold 3 with the epitope predictor AbEMap yielded superior results compared with alternative combinations evaluated.[15]

For parasitic pathogens, deepBCE-Parasite, a transformer-based linear B-cell epitope predictor trained on experimentally validated peptides from multiple parasite species, achieved approximately 81% accuracy and an AUC of 0.90 in cross-validation and independent testing. As proof of concept, eight epitopes of *Fasciola hepatica* leucine aminopeptidase were nominated; seven of the corresponding synthetic peptides bound positive sera in dot-blot assays.[17] A hybrid convolutional-BiLSTM network with forward filter-based feature selection performed competitively on Dengue and Zika datasets, surpassing DeepLBCEPred and LBCE-BERT.[18] Additional recent tools include RoBep, which incorporates a region-constraint mechanism to encourage spatially coherent predictions,[19] and BeeTLe, a multi-task framework combining linear epitope prediction with antibody-type classification.[20]

3.3. Integrated antigen prediction frameworks

Beyond single-epitope predictors, several frameworks now address whole-antigen selection. PLGDL (Protein Language and Geometric Deep Learning) integrates primary-sequence and three-dimensional structural signals, thereby reducing reliance on handcrafted features. The method proved robust across viral, bacterial, and eukaryotic datasets and, when applied during the Mpox outbreak, recovered known protective antigens while nominating G10R as an additional candidate.[21] VaxSeer, an AI-based evolutionary and antigenicity model, has been applied to influenza vaccine strain selection.[22]

3.4. Conformational B-cell epitope prediction: challenges and emerging solutions

More than 90% of B-cell epitopes are conformation-dependent, rendering their prediction particularly demanding.[15] Several integrative approaches have recently been proposed. DynaBCE employs a dynamic ensemble of handcrafted structural descriptors and protein-language-model embeddings, combining boosting-based modules with geometric graph neural networks.[23] AlphaEpi couples AlphaFold 3 structure prediction with graph neural networks and reports clear gains over baseline methods.[24] EpiGraph is among the first methods to receive experimental validation on antibody-antigen complexes for which no public experimental structures existed.[25] BConformerR uses a dual-branch Conformer architecture to predict both linear and conformational epitopes at residue resolution from sequence alone.[26] Future progress is expected to integrate molecular-dynamics simulations with graph neural networks so that epitope dynamics can be tracked on physiologically relevant timescales. Models that combine protein-language-model embeddings with structural features already outperform those relying on sequence embeddings alone.

3.5. Safety and physicochemical profiling of vaccine candidates

Once an epitope or antigen has been nominated, systematic *in silico* safety and physicochemical evaluation is required before experimental progression. Standard peptide-vaccine pipelines therefore incorporate multiple assessment dimensions.

Toxicity screening is typically performed first. ToxinPred, a support vector machine (SVM)–based predictor trained on toxic and non-toxic peptide collections, identifies potentially toxic stretches and can suggest mitigating mutations.[27] In a recent hepatitis C multi-epitope study, the tool was applied to every sequence variant; none was predicted toxic, supporting its utility in early candidate filtering.[28]

Allergenicity assessment constitutes the second major filter. AllerTOP v.2.0 employs amino-acid E-descriptors with auto- and cross-covariance transformation across an ensemble of classifiers and reaches 85.3% accuracy under five-fold cross-validation.[29] AlgPred 3.0 integrates machine learning with IgE epitope mapping on balanced allergen/non-allergen sets.[30] Additional resources include AllergenFP (descriptor fingerprints)[31] and AllerCatPro 2.0 (incorporation of three-dimensional structural information).[32] Comparative evaluations of these servers have been published.[33]

Immunotoxicity and inflammatory potential are

evaluated with tools such as IL6pred, which identifies peptides likely to induce interleukin-6, a cytokine linked to multiple adverse outcomes.[34],[35]

Physicochemical and developability profiling addresses manufacturability. ProtParam supplies molecular weight, theoretical pI, extinction coefficient, estimated half-life, instability index, and GRAVY.[36] DeepSoluE predicts solubility in *Escherichia coli* by means of a long short-term memory (LSTM) network operating on hybrid physicochemical and distributed-representation features.[37] TemStaPro transfers learning from protein-language-model embeddings trained on more than one million sequences with annotated growth temperatures to predict thermostability.[38] PEPlife2 compiles experimental half-life data for approximately 4500 peptide and protein entries and integrates sequence-similarity search utilities.[39] Bioinf-Farma consolidates multiple solubility and thermal-stability predictors through supervised random-forest meta-learners into a single expression efficiency score, enabling joint appraisal of antigenicity and developability.[40]

4. Deep learning for protein structure prediction in vaccine design

Structure-based vaccine design depends on accurate structural models; deep-learning structure prediction has transformed this prerequisite. AlphaFold and AlphaFold 3 deliver near-atomic-resolution models of individual proteins and protein-protein complexes.[41],[42] AlphaFold 3 has been applied specifically to TCR-pMHC interactions with progressive accuracy gains, and studies have shown that AlphaFold-derived models can discriminate valid from invalid epitopes—an important step toward high-throughput *in silico* nomination of T-cell targets. AlphaFold2 has likewise underpinned design of self-assembling protein nanoparticle vaccines from predicted oligomeric building blocks.[43] Pairing of AlphaFold 3 with AbEMap further illustrates the benefit of combining structure prediction with epitope algorithms.[15]

4.1. Structural limitations of AlphaFold in immunological applications

Despite its impact, AlphaFold 3 exhibits important shortcomings in immunological contexts. Systematic evaluation of TCR and TCR-pMHC structure-prediction tools has shown that while AlphaFold2, AlphaFold3, and specialized derivatives lead on overall TCR accuracy, every examined method struggles with CDR3 loops, docking orientation, TCR-peptide interfaces, and class II MHC-peptide interfaces.[44] Long or atypical CDR3 loops remain especially difficult, with CDR3 length positively correlated with RMSD.[45]

AlphaFold 3 does not model post-translational modifications such as glycosylation, phosphorylation, or proteolytic processing. Glycosylation is a principal immune-evasion mechanism of many viruses (*e.g.*, HIV, SARS-CoV-2) and fungi; its omission therefore severely constrains predictions of epitope accessibility under biologically realistic conditions.

The model returns a single static conformation, whereas proteins under physiological conditions populate ensembles. Epitopes may interconvert among conformations relevant to immune recognition, and the flexibility of antibody and TCR CDR loops influences binding affinity and specificity—features currently outside the model's scope.[46] For viral pathogens whose glycan shields mask immunogenic surfaces, the inability to incorporate glycosylation remains one of the largest unmet needs in computational vaccinology.

5. Generative protein design for immunogen engineering

Generative AI constitutes the most recent transformative development in computational vaccinology: proteins with prescribed functions can now be designed *de novo*. RFdiffusion generates protein backbones with user-defined geometric constraints.[47] A fine-tuned derivative, RFdiffusion2, was developed specifically for antibody design. In a 2025 *Nature* study, combination of RFdiffusion2 with yeast-display screening produced *de novo* variable heavy chains (VHHs), single-chain variable fragments (scFvs), and full antibodies that engage user-specified epitopes with atomic-level precision.[48] Cryo-electron microscopy confirmed binding poses of designed VHHs targeting influenza hemagglutinin and *Clostridioides difficile* toxin B; a high-resolution structure of the influenza-directed VHH validated the atomic accuracy of the designed complementarity-determining regions. Initial designs displayed modest affinity (tens to hundreds of nanomolar Kd); subsequent OrthoRep affinity maturation yielded single-digit nanomolar binders without loss of epitope selectivity.[48]

The RFdiffusion workflow comprises three stages: backbone generation, CDR sequence design by ProteinMPNN, and structural self-consistency filtering with a fine-tuned RoseTTAFold2 model.[48] ProteinMPNN itself solves the inverse folding problem—design of sequences compatible with fixed backbones—and has been shown to expand accessible sequence space while improving solubility and stability relative to parental proteins.[49]

MESODID (Machine Learning-Driven Structural Antigen Design Pipeline) targets multiepitope vaccine

antigens. Developed around *Plasmodium falciparum* circumsporozoite protein, the pipeline couples deep-learning design with iterative refinement and produced scaffolds displaying three copies of the PfCSP minor repeat; a 3.6 Å cryo-EM structure of the lead design confirmed correct orientation of the protective monoclonal antibody L9 epitope.[50] Given the prevalence of repeated epitopes in priority targets such as HIV-1 Envelope, SARS-CoV-2 spike, and influenza hemagglutinin, the approach is expected to generalize.

5.1. Translational challenges in generative protein design

Despite computational progress, translation faces substantial obstacles. Generative models can produce sequences that appear plausible *in silico* yet fail to express, aggregate, or lose intended function at the bench. RFdiffusion can enumerate tens of thousands of backbones, yet filtering viable candidates still requires structure validation with AlphaFold2 or related models. Experimental characterization of many designs has revealed soluble aggregates or unintended oligomers rather than the desired monomers[51]; the majority of designs are classified as failures once tested in the laboratory, typically because of low expression, misfolding, aggregation, or absent function.

Designed proteins may also share sequence or structural similarity with self-antigens and thereby risk inducing cross-reactive autoimmunity, arguing for incorporation of negative-selection filters against the human proteome. Constructs intended to be minimally immunogenic can nevertheless acquire new immunodominant epitopes through linkers, non-natural residues, or aggregation states. Failure rates at the preclinical stage (expression, stability, *in vivo* immunogenicity, toxicology) substantially exceed computational success rates; rigorous experimental validation of key predictions therefore remains mandatory.

6. Machine learning–guided antigen optimization

Beyond prediction and *de novo* design, machine learning is increasingly applied to improve existing antigens with respect to immunogenicity, cross-reactivity, stability, and developability.

6.1. Immunogenicity optimization

Naso (neoantigen immunogenicity optimization) couples simulated annealing with multi-objective training to elevate the immunogenicity of neoantigens. Unlike earlier approaches limited to HLA affinity, Naso simultaneously optimizes multiple immunological features. Experimental evaluation demonstrated that three or fewer amino-acid substitutions can convert a non-immunoreactive neoantigen into an immunoreactive one and that optimized neoantigens support cross-immunization, offering a route toward heteroclitic neoantigen vaccines.[52]

6.2. Cross-reactivity design

Machine-learning cycles have also been applied to factor H binding protein (fHbp) of *Neisseria meningitidis*. Computational models predicted antigenicity-related properties, a Gaussian-process model selected promising mutants, and experimental results were fed back into the model. The cycle generated mutants that transferred fHbp variant 1.1-specific conformational epitopes onto variant 3.28 while preserving binding to overlapping cross-reactive epitopes; the lead mutant retained the overall fHbp fold on biophysical and crystallographic characterization.[53]

6.3. Peptide stability and half-life optimization

Poor *in vivo* stability remains a decisive barrier to peptide-vaccine translation. Most peptides persist only 2–30 minutes because of proteolytic degradation, and molecules smaller than approximately 30 kDa undergo rapid renal clearance. PEPlife2 consolidates experimental half-life literature,[39] while ProtParam estimates half-life from N-end rule considerations.[36]

6.4. Developability and integrated assessment

Bioinf-Farma represents a recent advance in joint antigenicity-developability assessment. Candidates are submitted as sequences or structures; when an experimental structure is unavailable, a model is retrieved from AlphaFold DB or generated with Boltz-2. Antigenicity is quantified by combining structure-based conformational signals with sequence-based linear epitope scores. Developability is scored by random-forest meta-learners that integrate multiple solubility and thermal-stability predictors into an expression efficiency score. The modular architecture accommodates future predictors and is accessible via a web interface without programming expertise.[40]

6.5. Multi-epitope vaccine constructs: computational design versus experimental reality

Multi-epitope constructs are typically assembled with flexible or rigid linkers (*e.g.*, EAAAK, GPGPG, AAY) and often incorporate adjuvant domains such as β-defensin. A persistent gap separates computational design from experimental validation. Few studies systematically test whether common linkers preserve independent domain folding. Designed linkers themselves can become immunodominant, diverting responses from intended

epitopes. Expression yields and folding of multi-epitope fusions in hosts such as *E. coli* remain unpredictable, and no controlled benchmark comparing linker strategies has yet been published. Future work should validate linker function experimentally (circular dichroism, limited proteolysis, surface plasmon resonance), report actual expression yields and purity, and compare the immunological consequences of alternative linker designs.

7. Method selection and implementation framework

7.1. Criteria for method selection

Selection among computational methods depends on epitope type, pathogen, available data modalities, safety and developability requirements, experimental validation capacity, and computational resources. Representative tools for each scenario are summarized in Table 1.

7.2. Comparative performance of key algorithms

Several representative AI-driven epitope prediction and antigen discovery tools are summarized in Table 2.

7.3. Workflow integration

A typical computational antigen-discovery pipeline proceeds through the following stages:

(1) Genome/proteome mining to identify candidate antigens from pathogen sequences.
(2) Epitope prediction using ML methods (*e.g.*, MUNIS for T-cell epitopes, GraphBepi for B-cell epitopes).
(3) Structure prediction and epitope-accessibility modeling with AlphaFold.
(4) Safety assessment (toxicity, allergenicity, immunotoxicity) with ToxinPred, AllerTOP, and IL6pred.
(5) Immunogen design via generative methods (RFdiffusion, ProteinMPNN, MESODID).
(6) Optimization and developability profiling with Naso, Gaussian-process models, Bioinf-Farma, and PEPlife2.
(7) Experimental validation of top-ranked candidates in appropriate assays.

The PLGDL framework exemplifies this integrated philosophy by combining protein language models with geometric deep learning for protective-antigen prediction.[21]

Table 1. Factors guiding selection of computational antigen-discovery and vaccine-candidate assessment methods

Factor	Consideration	Relevant tools
Epitope type	Linear vs conformational B-cell; MHC I vs II T-cell	Linear B: BepiPred 3.0, deepBCE-Parasite; Conformational B: GraphBepi, AbEMap+AlphaFold3; T-cell: NetMHCpan 4.1, MUNIS
Pathogen type	Viral, bacterial, fungal, or parasitic	Broad: PLGDL; Parasite: deepBCE-Parasite; Influenza: VaxSeer
Data availability	Sequence only vs structure available	Sequence: VaxiJen, AllerTOP, ToxinPred; Structure: AlphaFold 3, GraphBepi
Safety assessment	Toxicity, allergenic, immunotoxicity	ToxinPred, AllerTOP v.2.0, AlgPred 3.0, IL6pred
Developability	Solubility, thermostability, half-life	DeepSoluE, TemStaPro, ProtParam, PEPlife2, Bioinf-Farma

Table 2. Comparative summary of representative AI-driven epitope prediction and antigen discovery tools

Tool	Input	Output	Performance	Limitations	Availability
MUNIS	Peptide + HLA	HLA-I epitopes	AUC ≈ 0.980	HLA-I only; allele bias	Open-source
GraphBepi	Sequence + AF2 structure	Conformational B epitopes	>5.5% AUC, 44% AUPR gain	Depends on structure quality	Open-source/web
NetMHCpan 4.1	Peptide + HLA	MHC binding	AUC < 0.90	Eurocentric bias	Free academic
deepBCE-Parasite	Peptide sequence	Parasite linear B epitopes	Acc ≈ 81%, AUC 0.90	Parasites only; linear	Web server
RFdiffusion + ProteinMPNN	Target structure/ hotspots	*De novo* binders	nM affinity post-maturation	Hallucination; expression failure	Open-source

Abbreviations: Acc: Accuracy; AUC: Area under the receiver operating characteristic curve; AUPR: Area under the precision-recall curve.

7.4. Toward standardized reporting: the MINIMAL proposal

Modeled on the MIQE guidelines for quantitative PCR, we propose MINIMAL (Minimum Information for AI in Immunology Research) recommendations for AI-immunology publications. Dataset descriptions should report source, size, composition, and racial/ethnic representation of training, validation, and test sets. Model training should document hyperparameter optimization, final configuration, k-fold cross-validation ($k \geq 5$ recommended), and performance on independent, non-overlapping test sets. Performance metrics should include at minimum AUC-ROC, area under the precision-recall curve, F1 score, Matthew's correlation coefficient, and accuracy, together with justification of the chosen classification threshold. Experimental validation must be clearly distinguished from *in silico* prediction, with detailed methods and explanation of any discrepancies. Code and model weights should be released publicly where feasible, and accessible web servers or software packages provided when applicable.

7.5. Validation and benchmarking

Experimental validation remains indispensable. Core approaches include enzyme-linked immunosorbent and surface-plasmon-resonance binding assays, HLA-peptide stability assays for T-cell epitopes, enzyme-linked immunospot and intracellular cytokine staining for T-cell immunogenicity, cryo-electron microscopy for designed protein structures, and X-ray crystallography for high-resolution confirmation. Standardized benchmarking datasets and metrics (AUC-ROC, precision-recall, accuracy) are prerequisites for fair inter-method comparison.

Figure 1 depicts integrated computational and experimental workflow for antigen discovery and immunogen design.

8. Critical limitations and challenges

8.1. Current method limitations

Training datasets remain skewed toward well-studied pathogens and HLA alleles. Large-scale immunopeptidomics resources such as the 651,237 HLA-I ligands used for MUNIS mark progress, yet comparable datasets are absent for many organisms. Conformational B-cell epitopes continue to be difficult to predict: methods trained on generic protein-protein interfaces transfer poorly to antibody-antigen recognition. Most algorithms ignore glycosylation, phosphorylation, and proteolytic processing, modifications that can occlude

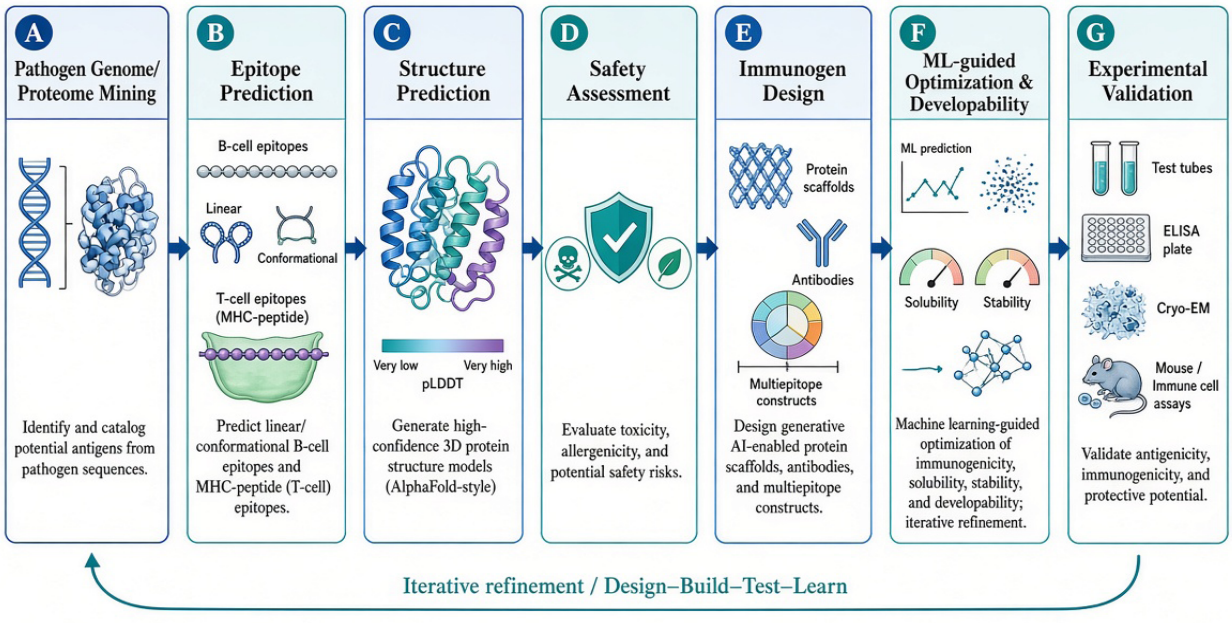


Figure 1. Integrated computational and experimental workflow for antigen discovery and immunogen design. The workflow proceeds from (A) pathogen genome/proteome mining through (B) epitope prediction, (C) structure prediction, (D) safety assessment, (E) immunogen design, (F) ML-guided optimization and developability, and (G) experimental validation. Feedback loops enable iterative refinement of candidates.

epitopes or reshape antigen structure. Models trained on one pathogen or allele frequently fail to generalize. Deep-learning structure prediction and generative design impose substantial computational costs. Although standalone tools exist for toxicity, allergenicity, and solubility, platforms that jointly optimize immunogenicity, safety, and manufacturability are only beginning to appear; Bioinf-Farma constitutes a notable step in that direction.[40]

8.2. Data bias and demographic equity

Model performance is bounded by the quality and representativeness of training data. Current MHC-peptide binding predictors inherit substantial demographic bias: training sets over-represent individuals of European ancestry relative to Asian and Pacific Islander populations.[54,55] HLA-A*02:01, the most common allele in European-ancestry cohorts, appears 5- to 15-fold more frequently than many alleles prevalent in non-European populations, producing unequal predictive accuracy across global populations.

Mitigation strategies include:

(1) Fairness-aware modeling frameworks that explicitly balance performance across alleles and population strata.
(2) Transfer learning: pre-training on data-rich alleles followed by fine-tuning on limited experimental data for under-represented alleles.
(3) Active learning that prioritizes experiments maximally reducing model uncertainty for under-represented populations.
(4) Synthetic data generation with generative models to fill experimental gaps for diverse HLA alleles.

8.3. Disparities in pathogen representation

AI-driven antigen discovery is distributed unevenly across pathogen types. Viral pathogens—HIV, influenza, and SARS-CoV-2 in particular—dominate the literature owing to compact genomes, rich datasets, and pandemic-driven funding, whereas fungal and parasitic pathogens lag substantially. Eukaryotic pathogens present unique challenges: larger genomes with extensive redundancy, stage-specific antigen expression (*e.g.*, *Plasmodium* life-cycle stages), and glycosylation patterns distinct from those of mammals. Multi-omics integration (genomics, transcriptomics, proteomics, glycomics) together with single-cell and spatial transcriptomics will be required to address these complexities.

9. Future directions

Priority research avenues include application of protein language models (ESM-2, ESM-3) to epitope prediction and antigen design[55]; multimodal frameworks that unify sequence, structural, and evolutionary information; active-learning cycles that alternate computational prediction with targeted experiments; explainable AI that reveals the biological basis of predictions; methods that generalize across pathogen types; tighter integration with mRNA, nanoparticle, and other delivery platforms; and end-to-end models that simultaneously predict immunogenicity, safety, stability, solubility, and half-life.

Antigen discovery is shifting from pathogen-sequence analysis alone toward a more complete picture of the host microenvironment. Single-cell multi-omics integrating transcriptomic, epigenomic, proteomic, and metabolomic readouts is transforming clinical immunology and now touches every stage of vaccine development—from infection-informed discovery to high-resolution efficacy assessment. Spatial transcriptomics coupled with single-cell sequencing reveals the spatial organization of immune cells within tissues. Future models should fuse genomic, transcriptomic, proteomic, metabolomic, and spatial data to predict in-situ antigen presentation, T-cell infiltration, and B-cell responses. Large single-cell and spatial multi-omics cohorts are expected to expand rapidly, creating training opportunities for more capable predictive systems.

10. Conclusions

AI-driven computational methods have matured from theoretical concepts into experimentally validated instruments that accelerate antigen discovery and immunogen design. Deep-learning epitope predictors such as MUNIS (AUC ≈ 0.980) and GraphBepi have substantially outperformed earlier tools by exploiting large-scale immunopeptidomics data and fusing sequence with structural information. Structure prediction, most prominently AlphaFold 3, now models TCR-pMHC and antibody-antigen interfaces at near-atomic resolution, although CDR3-loop modeling, glycosylation, and conformational dynamics remain unresolved. Generative platforms—RFdiffusion and ProteinMPNN foremost among them—have made *de novo*, epitope-specific antibodies and multiepitope antigens a practical reality, yet hallucination, aggregation, expression failure, and autoimmunity risk still limit translation. ML-guided optimization methods refine antigen immunogenicity and cross-reactivity iteratively, while computational safety and physicochemical tools have become integral to ensuring that designed candidates are not merely immunogenic but also safe, stable, and manufacturable.

Considerable obstacles persist: data scarcity and demographic bias, conformational epitope prediction, post-translational modifications, limited cross-pathogen

generalization, uneven pathogen representation, and the non-negotiable requirement for experimental validation. The field is converging on integrated frameworks that couple multiple computational methods with high-throughput validation in iterative design-build-test-learn cycles.

For investigators in medical microbiology, the practical recommendations are: (1) selecting methods according to epitope type, pathogen, available data, and required safety and developability assessments; (2) combining complementary methods within a coherent workflow; (3) validating predictions experimentally; and (4) adopting open-science practices that share data, code, and protocols. The computational tools needed to accelerate antigen discovery and immunogen design are now available; the remaining challenge is rigorous and equitable implementation.

Acknowledgments

None.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Conflict of interest

The author declares no competing interests.

Author contributions

This is a single-authored article.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

Not applicable.

References

1. Pronker ES, Weenen TC, Commandeur H, Claassen EHJHM, Osterhaus ADME. Risk in vaccine research and development quantified. *PLoS ONE*. 2013;8(3):e57755.

   doi: 10.1371/journal.pone.0057755

2. Rappuoli R. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*. 2001;19(17-19):2688-2691.

   doi: 10.1016/S0264-410X(00)00554-5

3. Pizza M, Scarlato V, Masignani V, *et al*. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*. 2000;287(5459):1816-1820.

   doi: 10.1126/science.287.5459.1816

4. Bravi B. Development and use of machine learning algorithms in vaccine target selection. *npj Vaccines*. 2024;9:15.

   doi: 10.1038/s41541-023-00795-8

5. Sanchez-Trincado JL, Gomez-Perosanz M, Reche PA. Fundamentals and methods for T- and B-cell epitope prediction. *J Immunol Res*. 2017;2017:2680160.

   doi: 10.1155/2017/2680160

6. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 2020;48(W1):W449-W454.

   doi: 10.1093/nar/gkaa379

7. Jurtz V, Paul S, Andreatta M, Marcatili P, Peters B, Nielsen M. NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol*. 2017;199(9):3360-3368.

   doi: 10.4049/jimmunol.1700893

8. Villanueva-Flores F, Sanchez-Villamil JI, Garcia-Atutxa I. AI-driven epitope prediction: a systematic review, comparative analysis, and practical guide for vaccine development. *npj Vaccines*. 2025;10:207.

   doi: 10.1038/s41541-025-01258-y

9. Gasperini G, Baylor N, Black S, *et al*. Vaccinology in the artificial intelligence era. *Sci Transl Med*. 2025;17(794):eadu3791.

   doi: 10.1126/scitranslmed.adu3791

10. Page MJ, McKenzie JE, Bossuyt PM, *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.

    doi: 10.1136/bmj.n71

11. Wohlwend J, Nathan A, Shalon N, *et al*. Deep learning enhances the prediction of HLA class I-presented CD8+ T cell epitopes in foreign pathogens. *Nat Mach Intell*. 2025;7(2):232-243.

    doi: 10.1038/s42256-024-00971-y

12. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: from synthetic to naturally processed epitopes. *PLoS Comput Biol*. 2018;14(11):e1006457.

    doi: 10.1371/journal.pcbi.1006457

13. Que J, Xue G, Wang T, *et al*. Identifying T cell antigen at the atomic level with graph convolutional network. *Nat*

Commun. 2025;16:5171.

doi: 10.1038/s41467-025-60461-6

14. Cai Y, Chen R, Song M, *et al*. CNNeoPP: a large language model-enhanced deep learning pipeline for personalized neoantigen prediction and liquid biopsy applications. *Front Immunol*. 2026;17:1722117.

doi: 10.3389/fimmu.2026.1722117

15. Yu H, Joseph-McCarthy D, Vajda S. Improving B-cell epitope prediction. *Drug Discov Today*. 2025;30(11):104489.

doi: 10.1016/j.drudis.2025.104489

16. Zeng Y, Wei Z, Yuan Q, *et al*. Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics*. 2023;39(4):btad187.

doi: 10.1093/bioinformatics/btad187.

17. Hu RS, Gu K, Ehsan M, *et al*. Transformer-based deep learning enables improved B-cell epitope prediction in parasitic pathogens: a proof-of-concept study on *Fasciola hepatica*. *PLoS Negl Trop Dis*. 2025;19(4):e0012985.

doi: 10.1371/journal.pntd.0012985

18. Rajmane PR, Khiani SR. Hybrid deep learning framework for enhanced prediction of linear B-cell epitopes using forward-based feature filtering. *Discov Artif Intell*. 2025;5:311.

doi: 10.1007/s44163-025-00588-z

19. Xu Y, Wei G, Zhou J, *et al*. RoBep: a region-oriented deep learning model for B-cell epitope prediction. *Bioinformatics*. 2026;42(1):btag006.

doi: 10.1093/bioinformatics/btag006

20. Yuan X. BeeTLe: An Imbalance-Aware Deep Sequence Model for Linear B-Cell Epitope Prediction and Classification with Logit-Adjusted Losses. *arXiv*. Preprint posted online 2023.

doi: 10.48550/arXiv.2309.02071

21. Zai Y, Zhao Y, Wang X, *et al*. Integrating protein language and geometric deep learning models for enhanced vaccine antigen prediction. *Nat Commun*. 2026;17(1):1033.

doi: 10.1038/s41467-025-67778-2

22. Shi W, Wohlwend J, Wu M, Barzilay R. Influenza vaccine strain selection with an AI-based evolutionary and antigenicity model. *Nat Med*. 2025;31(11):3862-3870.

doi: 10.1038/s41591-025-03917-y

23. Shen Y, Jiang Z, Liu R. Dynamic integration of feature- and template-based methods improves the prediction of conformational B cell epitopes. *Structure*. 2025;33(4):798-807.e4.

doi: 10.1016/j.str.2025.01.018

24. Jiang F, Guo Y, Ma H, *et al*. AlphaEpi: enhancing B cell epitope prediction with AlphaFold 3. In: Proceedings of the 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. Association for Computing Machinery; 2024:1-8.

doi: 10.1145/3698587.3701389

25. Choi S, Kim D. B cell epitope prediction by capturing spatial clustering property of the epitopes using graph attention network. *Sci Rep*. 2024;14:27496.

doi: 10.1038/s41598-024-78506-z

26. You Z, Ma J, Li H, Hu YF, Huang JD. BConformeR: A Conformer Based on Mutual Sampling for Unified Prediction of Continuous and Discontinuous Antibody Binding Sites. *arXiv*. Preprint posted online 2025.

doi: 10.48550/arXiv.2508.12029

27. Gupta S, Kapoor P, Chaudhary K, *et al*. In silico approach for predicting toxicity of peptides and proteins. *PLoS One*. 2013;8(9):e73957.

doi: 10.1371/journal.pone.0073957

28. Muhammad AM, Salum GM, Abd El Meguid M, Fotouh BE, Dawood RM. Bioinformatics analysis of multi-epitope peptide vaccines against hepatitis C virus: a molecular docking study. *J Genet Eng Biotechnol*. 2023;21(1):123.

doi: 10.1186/s43141-023-00583-w

29. Dimitrov I, Bangov I, Flower DR, Doytchinova I. AllerTOP v.2—a server for in silico prediction of allergens. *J Mol Model*. 2014;20(6):2278.

doi: 10.1007/s00894-014-2278-5

30. AlgPred 3.0. A web server for predicting allergenic peptides and mapping IgE epitopes. Indian Institute of Technology Delhi. Accessed September 15, 2026. https://webs.iiitd.edu.in/raghava/algpred3/?utm_source=chatgpt.com

31. Dimitrov I, Naneva L, Doytchinova I, Bangov I. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics*. 2014;30(6):846-851.

doi: 10.1093/bioinformatics/btt619

32. Nguyen MN, Krutz NL, Limviphuvadh V, Lopata AL, Gerberick GF, Maurer-Stroh S. AllerCatPro 2.0: a web server for predicting protein allergenicity potential. *Nucleic Acids Res*. 2022;50(W1):W36-W43.

doi: 10.1093/nar/gkac446

33. Limviphuvadh V, Ruethers T, Nguyen MN, *et al*. Fish isoallergens and variants: database compilation, in silico allergenicity prediction challenges, and epitope-based threshold optimization. *Front Bioinform*. 2025;5:1669237.

doi: 10.3389/fbinf.2025.1669237

34. Dhall A, Patiyal S, Sharma N, Usmani SS, Raghava GPS. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform*. 2021;22(2):936-945.

doi: 10.1093/bib/bbaa259

35. Dhall A, Patiyal S, Sharma N, Usmani SS, Raghava GPS. A web-based method for the identification of IL6-based immunotoxicity in vaccine candidates. *Methods Mol Biol*. 2023;2673:317-327.

doi: 10.1007/978-1-0716-3239-0_22

36. Gasteiger E, Hoogland C, Gattiker A, *et al*. Protein identification and analysis tools on the ExPASy server. In: Walker JM, ed. *The Proteomics Protocols Handbook*. Humana Press; 2005:571-607.

doi: 10.1385/1-59259-890-0:571

37. Wang C, Zou Q. Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with DeepSoluE. *BMC Biol*. 2023;21:12.

doi: 10.1186/s12915-023-01510-8

38. Pudžiuvelytė I, Olechnovič K, Godliauskaite E, Sermokas K, Urbaitis T, Gasiunas G, Kazlauskas D. TemStaPro: protein thermostability prediction using sequence representations from protein language models. *Bioinformatics*. 2024;40(4):btae157.

doi: 10.1093/bioinformatics/btae157

39. Alam U, Chaudhary K, Kumar N, Tomer R, Patiyal S, Raghava GPS. PEPlife2: an updated repository of the half-life of peptides and proteins. *Immuno*. 2026;6(2):26.

doi: 10.3390/immuno6020026

40. Bondi H, Crespi M, Orlando M, *et al*. Bioinf-Farma: supervised integration of epitope prediction and recombinant protein developability for automated vaccine candidate prioritization. *bioRxiv*. Preprint posted online June 18, 2026.

doi: 10.64898/2026.06.15.732271

41. Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596(7873):583-589.

doi: 10.1038/s41586-021-03819-2

42. Abramson J, Adler J, Dunger J, *et al*. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*. 2024;630(8016):493-500.

doi: 10.1038/s41586-024-07487-w

43. Haas CM, Jasti N, Dosey A, *et al*. From sequence to scaffold: computational design of protein nanoparticle vaccines from AlphaFold2-predicted building blocks. *Proc Natl Acad Sci U S A*. 2025;122(45):e2409566122.

doi: 10.1073/pnas.2409566122

44. Shi Y, Parks JM, Smith JC. Comparative analysis of TCR and TCR-pMHC complex structure prediction tools. *J Chem Inf Model*. 2025;65(13):7156-7173.

doi: 10.1021/acs.jcim.5c00298

45. Le HN, de Freitas MV, Antunes DA. Strengths and limitations of web servers for the modeling of TCRpMHC complexes. *Comput Struct Biotechnol J*. 2024;23:2938-2948.

doi: 10.1016/j.csbj.2024.06.028

46. Spoendlin FC, Fernández-Quintero ML, Raghavan SSR, *et al*. Predicting the conformational flexibility of antibody and T cell receptor complementarity-determining regions. *Nat Mach Intell*. 2025;7(10):1755-1767.

doi: 10.1038/s42256-025-01131-6

47. Watson JL, Juergens D, Bennett NR, *et al*. De novo design of protein structure and function with RFdiffusion. *Nature*. 2023;620(7976):1089-1100.

doi: 10.1038/s41586-023-06415-8

48. Bennett NR, Watson JL, Ragotte RJ, *et al*. Atomically accurate de novo design of antibodies with RFdiffusion. *Nature*. 2026;649(8095):183-193.

doi: 10.1038/s41586-025-09721-5

49. Nelson JAD, Garfinkle SE, Lin ZJ, *et al*. Machine learning enables de novo multiepitope design of Plasmodium falciparum circumsporozoite protein to target trimeric L9 antibody. *Proc Natl Acad Sci U S A*. 2025;122(49):e2512358122.

doi: 10.1073/pnas.2512358122

50. Anishchenko I, Pellock SJ, Chidyausiku TM, *et al*. De novo protein design by deep network hallucination. *Nature*. 2021;600(7889):547-552.

doi: 10.1038/s41586-021-04184-w

51. Zhang J, Zhou W, Yu J, Yan C, Liu J. A machine learning-based method to optimize the immunogenicity of human leukocyte antigen class I-restricted neoantigens. *Brief Bioinform*. 2025;26(6):bbaf582.

doi: 10.1093/bib/bbaf582

52. Chesterman C, Desautels T, Sierra L-J, *et al*. Design of cross-reactive antigens with machine learning and high-throughput experimental evaluation. *Front Bioinform*. 2025;5:1580967.

doi: 10.3389/fbinf.2025.1580967

53. Atkins TK, Solanki A, Vasmatzis G, Cornette J, Riedel M. Evaluating NetMHCpan performance on non-European HLA alleles not present in training data. *Front Immunol*. 2024;14:1288105.

doi: 10.3389/fimmu.2023.1288105

54. Atkins TK, Solanki A, Vasmatzis G, Cornette J, Riedel M. Geographically Biased Composition of NetMHCpan Training Datasets and Evaluation of MHC-Peptide Binding Prediction Accuracy on Novel Alleles. *bioRxiv*. Preprint posted online September 6, 2023.

doi: 10.1101/2023.09.03.556092

55. Lin Z, Akin H, Rao R, *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*. 2023;379(6637):1123-1130.

doi: 10.1126/science.ade2574