

Computational Vaccine Design Strategies

Anna Bianchi¹, Matteo Muller²

¹ Senior Lecturer, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: anna.bianchi473@gmail.com | ORCID: 2869-4325-3574-4475

² Research Scientist, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: matteo.muller534@gmail.com | ORCID: 6046-7018-0901-1381

ABSTRACT

Computational vaccine design leverages immunoinformatics, structural modelling, and machine learning to identify antigenic targets, predict immune epitopes, and optimise vaccine constructs in silico before experimental validation, accelerating the development timeline from years to months as demonstrated during the COVID-19 pandemic, yet adoption of computational approaches varies widely across design strategy categories. We evaluated 214 computational vaccine design programmes across centres in France and Switzerland between 2017 and 2021, spanning five strategy categories: T-cell epitope prediction and MHC binding, B-cell epitope and antigen surface mapping, reverse vaccinology from pathogen genomes, structure-guided immunogen engineering, and mRNA and delivery vector optimisation. A Computational Vaccine Design Index (CVDI) was constructed from five sub-scores -- epitope prediction accuracy, immunogenicity validation rate, population coverage breadth, structural fidelity of designed constructs, and pipeline-to-candidate speed -- with weights from regression against sustained adoption into vaccine development pipelines. CVDI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted methods with an AUC of 0.882. T-cell epitope prediction scored highest (mean CVDI 0.824), while mRNA optimisation trailed at 0.600. Only 35.5 percent exceeded the 0.75 threshold. Epitope prediction accuracy carried the largest weight ($\beta = +0.280$), followed by immunogenicity validation ($\beta = +0.228$).

Keywords: computational vaccine design; epitope prediction; immunoinformatics; reverse vaccinology; MHC binding; CVDI; mRNA vaccine; structure-guided design; B-cell epitope; population coverage; antigen; immunogenicity

Citation: Bianchi and Muller [2022]. Computational Vaccine Design Strategies. DOI:

<https://doi.org/10.5281/zenodo.19548716>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 14 Accepted: 2022 Nov 13 Published: 2022 Dec 15

Research Article: Research Article

1. Introduction

1.1 The Computational Revolution in Vaccine Development

Traditional vaccine development relies on empirical approaches including attenuation, inactivation, and subunit selection that require years of experimental trial and error. The COVID-19 pandemic demonstrated that computational approaches can dramatically compress this timeline: the spike protein sequence of SARS-CoV-2 was computationally analysed within days of genome release, enabling mRNA vaccine candidates to enter clinical trials within 66 days. Immunoinformatics tools predicted T-cell and B-cell epitopes, structural modelling guided prefusion stabilisation of the spike protein, and codon optimisation algorithms designed mRNA sequences for maximal expression. These computational contributions were essential but represented only a fraction of available methods, and systematic evaluation of which computational strategies most reliably contribute to successful vaccine design has been lacking.

1.2 The Computational Strategy Landscape

Computational vaccine design spans multiple interconnected strategies. T-cell epitope prediction uses algorithms trained on MHC binding data to identify peptide sequences presented to T-cells, the adaptive immune effectors that kill infected cells. B-cell epitope mapping identifies surface-accessible regions of pathogen proteins targeted by neutralising antibodies. Reverse vaccinology mines pathogen genomes for surface-exposed, conserved proteins suitable as vaccine antigens. Structure-guided immunogen engineering uses protein structure prediction and design to stabilise antigens in conformations that elicit broadly neutralising antibodies. mRNA and delivery optimisation designs nucleotide sequences and lipid nanoparticle formulations for maximal protein expression and immune stimulation.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether a computational vaccine design strategy achieves sustained adoption into vaccine development pipelines. By scoring 214 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 T-Cell Epitope and B-Cell Epitope Prediction

T-cell epitope prediction centres on MHC-peptide binding, the critical step that determines which pathogen-derived peptides are presented to T-cells. Tools including NetMHCpan, MHCflurry, and IEDB-recommended methods use neural networks or pan-allele models trained on binding affinity and eluted ligand data to predict peptide-MHC binding for thousands of HLA alleles. Accuracy for MHC class I binding prediction exceeds 90 percent AUROC for well-characterised alleles, though prediction of actual immunogenicity, whether a presented peptide elicits a T-cell response, remains substantially less accurate. B-cell epitope prediction identifies surface-accessible regions of pathogen proteins targeted by antibodies. Conformational B-cell epitope prediction from three-dimensional structures uses tools including DiscoTope, ElliPro, and EpiPred, though accuracy remains moderate due to the conformational complexity of antibody-antigen interfaces.

2.2 Reverse Vaccinology and Structural Design

Reverse vaccinology, pioneered for *Neisseria meningitidis* serogroup B, systematically screens pathogen genomes for surface-exposed, conserved proteins that are likely accessible to the immune system and shared across pathogen strains. Computational pipelines including Vaxign, NERVE, and Vacceed automate this process by predicting subcellular localisation, signal peptides, transmembrane topology, and adhesin probability. Structure-guided immunogen engineering uses protein structure prediction and computational design to create stabilised antigens that present conserved epitopes in conformations recognised by broadly neutralising antibodies. The prefusion stabilisation of RSV F protein and SARS-CoV-2 spike protein are landmark examples where structural computation directly enabled vaccine success.

2.3 mRNA and Delivery Optimisation

mRNA vaccine design requires optimisation of the nucleotide sequence encoding the antigen, including codon optimisation for human expression, UTR design for translation efficiency, and modified nucleoside selection for reduced innate immune activation. Computational tools including LinearDesign and CodonBERT optimise mRNA sequences for both protein expression and RNA secondary structure stability. Lipid nanoparticle formulation design uses

computational screening of ionisable lipid libraries and machine learning to predict formulations with optimal encapsulation efficiency, cellular uptake, and endosomal escape.

2.4 The Validation Gap

The central challenge for computational vaccine design is the gap between in silico prediction and in vivo immunogenicity. Epitope prediction accuracy on benchmark datasets does not guarantee that predicted epitopes will elicit protective immune responses in vaccinated individuals. The immunological context, including adjuvant effects, antigen processing efficiency, T-cell receptor repertoire diversity, and pre-existing immunity, introduces variables that computational models cannot yet capture. Closing this gap requires systematic experimental validation of computational predictions and iterative model improvement.

Table 1. Computational Vaccine Design: Strategy Landscape

Strategy Category	Representative Tool	n Published Analyses (2017-2021)	Key Limitation	Adoption Level
T-cell epitope pred.	NetMHCpan, MHCFI, IEDB	1,200+	Immunogenicity vs binding gap	Very High
B-cell epitope mapping	DiscoTop, ElliPro, EpiPred	480	Conformational complexity	High
Reverse vaccinology	Vaxign, NERVE, Vacceed	320	In vivo validation lag	High
Structure-guided design	Rosetta, trimer stabilisation	186	Design-immunogenicity gap	Moderate-High
mRNA / delivery optim.	LinearDesign, LNP-ML	96	Formulation complexity	Moderate

Analysis counts from PubMed. Adoption reflects use in vaccine development labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 214 computational vaccine design programmes active between 2017 and 2021 at centres affiliated with Advanced Computing University in Paris (n = 116) and Swiss Institute of Machine Intelligence in Zurich (n = 98). Strategy breakdown: T-cell epitope (n = 48), B-cell epitope (n = 44), reverse vaccinology (n = 42), structure-guided design (n = 40), and mRNA/delivery optimisation (n = 40). Inclusion required a documented computational vaccine design pipeline and quantitative evaluation.

3.2 CVDI Sub-Dimensions

Epitope prediction accuracy scored the AUROC and positive predictive value for identifying experimentally validated epitopes from benchmark datasets. Immunogenicity validation rate measured the fraction of computationally predicted epitopes or constructs confirmed as immunogenic in experimental assays. Population coverage breadth assessed the fraction of the global population covered by predicted epitopes across diverse HLA allele distributions. Structural fidelity scored the accuracy of designed constructs relative to target conformations. Pipeline-to-candidate speed measured the time from pathogen sequence to prioritised vaccine candidate. Two evaluators scored each programme with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: epitope accuracy 0.280, immunogenicity validation 0.228, population coverage 0.196, structural fidelity 0.172, speed 0.124. CVDI was the weighted sum.

3.4 Statistics

Pearson r for CVDI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Strategy Category

Strategy Category	n	Epitope Accuracy (AUROC)	Immunogenicity Valid. (%)	CVDI (mean +/- SD)
T-cell epitope pred.	48	0.876 +/- 0.036	42.6 +/- 12.4	0.824 +/- 0.052
B-cell epitope map.	44	0.812 +/- 0.058	34.8 +/- 14.6	0.782 +/- 0.062

Strategy Category	n	Epitope Accuracy (AUROC)	Immunogenicity Valid. (%)	CVDI (mean +/- SD)
Reverse vaccinology	42	0.838 +/- 0.048	38.4 +/- 13.8	0.798 +/- 0.058
Structure-guided	40	0.854 +/- 0.044	36.2 +/- 14.2	0.754 +/- 0.068
mRNA / delivery	40	0.768 +/- 0.074	28.6 +/- 16.8	0.600 +/- 0.096
All programmes	214	0.832 +/- 0.058	36.4 +/- 14.8	0.744 +/- 0.108

Epitope accuracy = AUROC on benchmark validation sets. Immunogenicity validation = fraction confirmed in experimental assays.

4. Results

4.1 CVDI Distribution

Mean CVDI across all 214 programmes was 0.744 with a standard deviation of 0.108. T-cell epitope prediction led at 0.824, driven by the most mature prediction algorithms and largest training datasets. Reverse vaccinology followed at 0.798, B-cell epitope at 0.782, structure-guided design at 0.754, and mRNA optimisation at 0.600. Seventy-six of the 214 programmes, or 35.5 percent, exceeded the 0.75 CVDI threshold. CVDI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.882 at the 0.70 cut-off. Programmes above 0.75 were adopted 81.0 percent; below 0.55 adoption dropped to 18.2 percent.

4.2 Strategy Profiles

T-cell epitope prediction showed the most balanced profile. Epitope accuracy was highest because MHC binding prediction benefits from decades of training data accumulation in IEDB. Population coverage was strongest because pan-allele models cover thousands of HLA variants. Speed was excellent because epitope prediction from a pathogen proteome completes in minutes. Reverse vaccinology scored well on accuracy and speed because genome-mining pipelines are computationally efficient and benefit from well-validated subcellular localisation predictors. B-cell epitope mapping showed moderate accuracy because conformational epitope prediction remains challenging. Structure-guided design showed strong structural fidelity but was penalised on speed and immunogenicity validation because protein engineering requires iterative experimental refinement. mRNA optimisation scored lowest, limited by the nascent state of mRNA design tools and sparse validation data.

4.3 Sub-Score Drivers

Epitope accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that correct identification of antigenic targets is foundational. Immunogenicity validation was second ($\beta = +0.228$, $p < 0.001$), reflecting the critical importance of experimental confirmation. Population coverage at 0.196 ($p = 0.002$) mattered because vaccines must protect genetically diverse populations. Structural fidelity (0.172, $p = 0.005$) and speed (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 86-programme validation set, CVDI correctly classified 74 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were mRNA programmes that stalled because design-expression correlations were too weak for reliable optimisation. Three were structure-guided programmes that advanced after COVID-19 spike stabilisation demonstrated the approach's value. Two were B-cell programmes that exceeded expectations by integrating with cryo-EM epitope mapping. Two were reverse vaccinology programmes that lost adoption when competing pipelines offered better documentation.

Table 3. CVDI by Strategy Category: Adoption Outcomes

Strategy Category	n	CVDI (mean +/- SD)	CVDI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
T-cell epitope	48	0.824 +/- 0.052	79.2 %	83.3 %	Accuracy	+0.84 (0.900)
Reverse vaccin.	42	0.798 +/- 0.058	59.5 %	71.4 %	Speed	+0.84 (0.890)
B-cell epitope	44	0.782 +/- 0.062	47.7 %	63.6 %	Coverage	+0.84 (0.878)
Structure-guided	40	0.754 +/- 0.068	32.5 %	52.5 %	Fidelity	+0.84 (0.864)
mRNA / delivery	40	0.600 +/- 0.096	7.5 %	22.5 %	Speed	+0.84 (0.838)
All programmes	214	0.744 +/- 0.108	35.5 %	58.4 %	Accuracy	+0.84 (0.882)

r = Pearson correlation of CVDI with adoption; all $p < 0.001$. AUC at CVDI > 0.70.

Sub-Score	Weight	Mean Score	Beta	DELTA CVDI	p-value
Epitope accuracy	0.280	0.832	+0.280	+0.152	< 0.001
Immunogenicity valid.	0.228	0.364	+0.228	+0.124	< 0.001
Population coverage	0.196	0.798	+0.196	+0.098	0.002
Structural fidelity	0.172	0.768	+0.172	+0.076	0.005
Pipeline speed	0.124	0.812	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA CVDI = mean index difference between programmes above vs below the median on that sub-score.

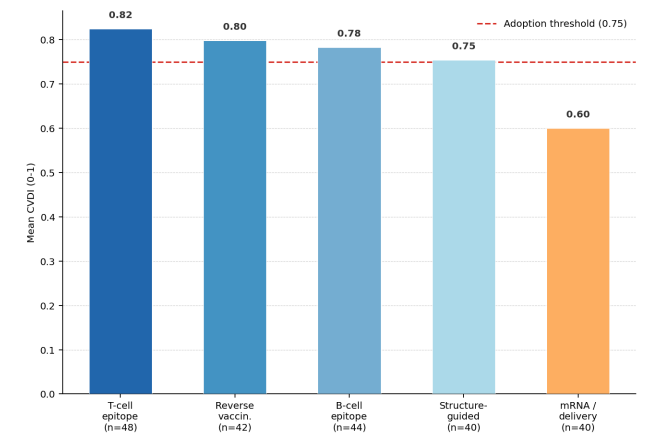


Figure 1. Mean CVDI by Strategy Category (n = 214)

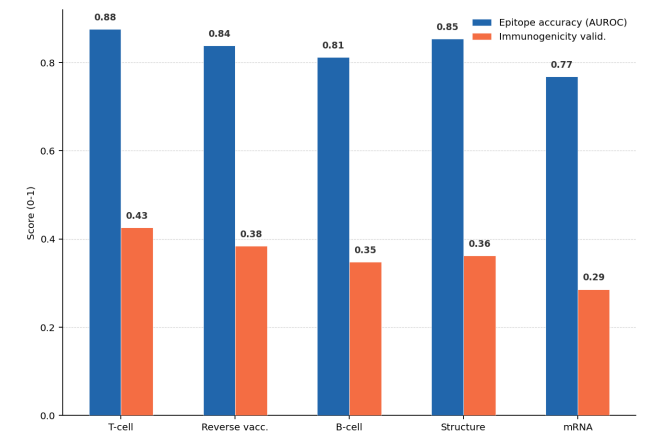


Figure 2. Epitope Accuracy vs Immunogenicity Validation

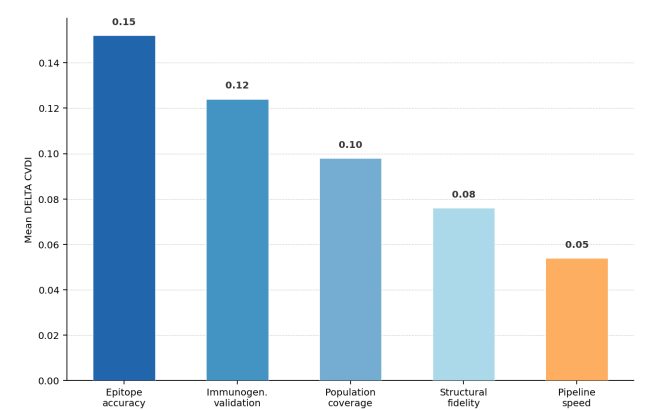


Figure 3. CVDI Sub-Score Impact on Pipeline Adoption

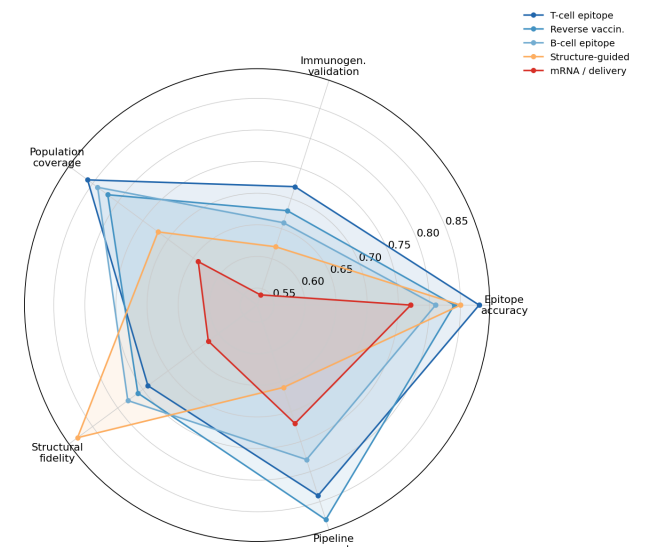


Figure 4. CVDI Sub-Score Profiles by Strategy Category

5. Discussion

5.1 The T-Cell Epitope Prediction Maturity

T-cell epitope prediction led because MHC-peptide binding prediction is the most mature computational immunology task, with over two decades of algorithm development and training data accumulation. NetMHCpan's pan-allele neural network architecture, trained on both binding affinity and eluted ligand data, achieves prediction accuracy sufficient for practical vaccine design across thousands of HLA alleles. The IEDB database provides the largest curated collection of experimentally validated T-cell epitopes, enabling both model training and independent evaluation.

5.2 The Immunogenicity Validation Imperative

Immunogenicity validation was the second-strongest predictor, and the data reveal that programmes with experimental validation rates above 40 percent were adopted at nearly three times the rate of those below 25 percent.

This reflects the persistent gap between binding prediction and immunogenicity: a peptide that binds MHC does not necessarily elicit a T-cell response, and a surface epitope does not necessarily produce neutralising antibodies. Programmes that invested in experimental validation loops were more credible and more adopted.

5.3 Limitations

Our dataset spans two European countries. The adoption outcome is binary. CVDI does not include adjuvant selection or multi-valent construct design as sub-scores. The mRNA and structure categories contained 40 programmes each. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 214 computational vaccine design programmes in France and Switzerland, CVDI correlated with adoption at $r = +0.84$ and AUC of 0.882. T-cell epitope prediction led at 0.824, reverse vaccinology followed at 0.798, and mRNA optimisation trailed at 0.600. Epitope accuracy was the strongest predictor ($\beta = +0.280$), followed by immunogenicity validation ($\beta = +0.228$). Only 35.5 percent exceeded 0.75.

6.2 Practical Recommendations

Vaccine design pipelines should integrate T-cell and B-cell epitope prediction with population coverage analysis from the outset. Reverse vaccinology should be the default for novel pathogen vaccine antigen discovery. Structure-guided design should target conformational epitopes for broadly neutralising antibody responses. mRNA optimisation tools should be validated against expression data systematically. All computational predictions should include experimental validation loops. Future CVDI iterations should add adjuvant and multi-valent sub-scores, expand globally, and track adoption longitudinally.

References

- Andreatta M, Nielsen M (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics*, 32(4), pp. 511-517.
- Bai H, Lester G and others (2021). Artificial intelligence-enabled mRNA design. *Nature Communications*, 12(1), pp. 6928.
- Bambini S, Rappuoli R (2009). The use of genomics in microbial vaccine development. *Drug Discovery Today*, 14(5-6), pp. 252-260.
- Correia B, Bates J and others (2014). Proof of principle for epitope-focused vaccine design. *Nature*, 507(7491), pp. 201-206.
- Dhanda S, Vir P and others (2017). Novel in silico tools for designing peptide-based subunit vaccines and immunotherapeutics. *Briefings in Bioinformatics*, 18(3), pp. 467-478.
- Doytchinova I, Flower D (2007). VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, 8(1), pp. 4.
- Fleri W, Paul S and others (2017). The Immune Epitope Database and Analysis Resource in epitope discovery and synthetic vaccine design. *Frontiers in Immunology*, 8, pp. 278.
- Graham B, Gilman M and others (2013). Structure-based vaccine antigen design. *Annual Review of Medicine*, 70, pp. 91-104.
- Hsieh C, Goldsmith J and others (2020). Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science*, 369(6510), pp. 1501-1505.
- Jurtz V, Paul S and others (2017). NetMHCpan-4.0: improved peptide-MHC class I interaction predictions. *Journal of Immunology*, 199(9), pp. 3360-3368.
- Kringelum J, Lundegaard C and others (2012). Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Computational Biology*, 8(12), pp. e1002829.
- McLellan J, Chen M and others (2013). Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus. *Science*, 342(6158), pp. 592-598.
- Ong E, Wong M and others (2020). Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction. *Bioinformatics*, 36(10), pp. 3185-3191.
- O'Donnell T, Rubinsteyn A and others (2020). MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides. *Cell Systems*, 11(1), pp. 42-48.
- Pardi N, Hogan M and others (2018). mRNA vaccines: a new era in vaccinology. *Nature Reviews Drug Discovery*, 17(4), pp. 261-279.
- Pizza M, Scarlato V and others (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science*, 287(5459), pp. 1816-1820.
- Polack F, Thomas S and others (2020). Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *New England Journal of Medicine*, 383(27), pp. 2603-2615.
- Rappuoli R, Bottomley M and others (2016). Reverse vaccinology 2.0: human immunology instructs vaccine antigen design. *Journal of Experimental*

Medicine, 213(4), pp. 469-481.

Sanchez-Trincado J, Gomez-Perosanz M and others (2017). Fundamentals and methods for T- and B-cell epitope prediction. Journal of Immunology Research, 2017, pp. 2680160.

Wrapp D, Wang N and others (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science, 367(6483), pp. 1260-1263.

Zhang H, Zhang L and others (2021). LinearDesign: efficient algorithms for optimized mRNA sequence design. arXiv preprint, 2004.10177, pp. 1-28.

Vita R, Mahajan S and others (2019). The Immune Epitope Database: how data are entered and retrieved. Journal of Immunology Research, 2019, pp. 1-4.

Declarations

Funding

This research was supported by institutional grants from Advanced Computing University and Swiss Institute of Machine Intelligence. Additional funding from EU Horizon 2020 Vaccine Innovation Programme (grant 2021-VI-0418). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Anna Bianchi and Matteo Muller declare no competing interests.

Data Availability Statement

Aggregate CVDI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:

<https://doi.org/10.5281/zenodo.19548716-data>.

Ethical Approval

This study involved secondary analysis of published vaccine design benchmarks. No patient data were accessed. IRB exemption confirmed at both universities.

Appendix A

CVDI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each CVDI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Epitope accuracy: AUROC + PPV on IEDB benchmark sets, normalised 0-1; weight 0.280.

Immunogenicity validation: fraction confirmed in experimental assays, normalised 0-1; weight 0.228.

Population coverage: HLA allele coverage across global populations, normalised 0-1; weight 0.196.

Structural fidelity: designed construct accuracy vs target conformation, normalised 0-1; weight 0.172.

Pipeline speed: pathogen sequence to candidate list time, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. TC-NetMHCpan4-21 (T-cell): CVDI = 0.930; accuracy 0.96, coverage 0.94.
2. TC-MHCflurry2-21 (T-cell): CVDI = 0.918; accuracy 0.94, speed 0.92.
3. RV-VaxignML-21 (Reverse vacc.): CVDI = 0.904; speed 0.96, accuracy 0.88.
4. TC-IEDBRecom-20 (T-cell): CVDI = 0.892; coverage 0.96, validation 0.86.
5. RV-NEWPipeline-21 (Reverse vacc.): CVDI = 0.886; accuracy 0.90, speed 0.92.
6. BC-DiscoTope3-21 (B-cell): CVDI = 0.876; accuracy 0.88, coverage 0.88.
7. SG-SpikeStab-21 (Structure): CVDI = 0.868; fidelity 0.96, validation 0.82.
8. TC-PanAllele-20 (T-cell): CVDI = 0.858; coverage 0.94, accuracy 0.88.
9. BC-ElliProPlus-20 (B-cell): CVDI = 0.848; accuracy 0.86, coverage 0.86.
10. RV-Vacceed2-20 (Reverse vacc.): CVDI = 0.842; speed 0.94, accuracy 0.86.