Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A dual-arm peptide vaccine against KRAS, EGFR, PD-L1 and VEGFR2 for non-small cell lung cancer in Vietnamese patients

Dung Vu Luu
Hai phong University Of Medicine and Pharmacy
Hau Nguyen Xuan
Hai phong University Of Medicine and Pharmacy
Trang Le Huyen
Hai phong University Of Medicine and Pharmacy
Giang Nguyen Huong
Hai phong University Of Medicine and Pharmacy
Trang Nguyen Ngoc
Hai phong University Of Medicine and Pharmacy
Lan Nguyen Thi Phuong
Hai phong University Of Medicine and Pharmacy
Hoang Nguyen Viet
Hai phong University Of Medicine and Pharmacy
Khanh La Van
Institute of Cosmetic Health and Regenerative Medicine
Thanh Tran Trung
Institute of Cosmetic Health and Regenerative Medicine
Giang Bui Thi Huong
Hai phong University Of Medicine and Pharmacy
Anh Do Nguyen Mai
Hai phong University Of Medicine and Pharmacy
Hieu Truong Van
Hai phong University Of Medicine and Pharmacy
Thuc Pham Van
Hai phong University Of Medicine and Pharmacy
Khai Nguyen Van
Hai phong University Of Medicine and Pharmacy
Loc Pham Thi
ptloc@hpmu.edu.vn

Hai phong University Of Medicine and Pharmacy

Article

Keywords: NSCLC, multi-epitope vaccine, KRAS/EGFR neoepitopes, XCL1–XCR1 targeting, immunoinformatics, Vietnamese HLA

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10642195/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

Abstract

Therapeutic vaccines for non-small cell lung cancer (NSCLC) are almost invariably designed against global HLA supertypes, yet the immunogenetics of Southeast Asian patients diverges substantially from the largely European allele sets underlying them. We report CRUISER, a dual-arm vaccine restricted to a Vietnamese-specific panel of 12 HLA class I alleles and 13 class II specificities. The T-cell arm is a 273-residue recombinant protein carrying eight KRAS and EGFR neoepitopes fused to the chemokine XCL1, which targets the construct to XCR1-positive conventional type 1 dendritic cells, the subset specialised in cross-presentation. The antibody arm comprises four disulfide-cyclised peptides drawn from the ligand-binding faces of PD-L1 and VEGFR2, selected on structural rather than antigenicity criteria. Seventeen of 92 candidate epitopes survived a four-layer safety and potency cascade. Combined population coverage reached 73.35% in Vietnam against 46.78% worldwide and 41.23% in Europe, quantifying the gain conferred by population restriction. Critically, triplicate 100 ns molecular dynamics simulations benchmarked against a free XCL1:XCR1 positive control showed that fusion preserves receptor engagement: the fusion complex matched the free chemokine in MM-PBSA interaction energy, maintained more interface hydrogen bonds, and distorted neither the XCL1 core nor the receptor. All findings are computational predictions requiring experimental validation.

Introduction

Lung cancer is the leading cause of cancer-related death worldwide[1]. NSCLC accounts for approximately 80–85% of cases: in 2022 adenocarcinoma, squamous cell carcinoma and large-cell carcinoma together represented 81.5% of new cases in men and 83.3% in women, against 11.5% and 9.7% for small-cell carcinoma[2]. In Viet Nam it is likewise the leading cause of cancer death, accounting for 21,340 of 111,015 cancer deaths in 2024 (19.2%) and ranking second in incidence with 23,085 new cases (13.0%); among men it is the most frequent malignancy by both measures[3]. A mortality-to-incidence ratio of 0.92 reflects late presentation and a prognosis that neither tyrosine kinase inhibitors nor checkpoint blockade has yet altered at the population level.

Two largely mutually exclusive driver genes dominate the mutational landscape of NSCLC in Vietnamese patients: activating EGFR mutations occur in approximately 35% of cases, markedly more often than in European cohorts, and KRAS mutations in approximately 22%[4]. Because they are rarely co-mutated, a construct addressing both covers most domestic patients. KRAS is particularly attractive: unlike tumour-associated antigens, neoepitopes arising from somatic mutation escape central tolerance and can elicit genuinely tumour-restricted T-cell responses[5], and the codon 12 neoepitopes are efficiently presented by HLA-A*11:01, the most frequent class I allele in this population at approximately 25%[6,7]. The convergence of a recurrent public neoantigen with a high-frequency restricting allele is a population-specific opportunity that supertype-based constructs cannot exploit.

Tumour-reactive cytotoxic T lymphocytes alone, however, are not sufficient. The PD-1/PD-L1 axis delivers inhibitory signals driving T-cell exhaustion, while VEGFR2-driven angiogenesis produces a disorganised,

poorly perfused vasculature impeding lymphocyte extravasation into solid tumours. Antibody blockade of either axis therefore complements rather than duplicates a cellular response, defining a four-target design in which each module addresses one step of a single mechanistic loop: KRAS and EGFR neoepitopes generate effector T cells, anti-VEGFR2 antibodies normalise the vasculature so that those cells can infiltrate, and anti-PD-L1 antibodies relieve checkpoint inhibition at the tumour cell surface.

Cellular and humoral immunity nevertheless demand structurally incompatible antigen formats: T-cell epitopes are linear peptides liberated by proteasomal and endolysosomal processing, whereas blocking antibodies must engage the conformationally constrained surface of a folded receptor. Fusing all four antigens into one chain would set these requirements against one another and generate junctional neo-epitopes at every linker. We therefore split the system into two independently formulated arms: Pool A, a recombinant T-cell construct, and Pool B, a set of conformationally constrained synthetic peptides.

The choice of adjuvant is a further point of departure. In place of the bacterial ribosomal proteins conventionally used, we appended the chemokine XCL1 to the amino terminus in order to deliver antigen to XCR1-positive conventional type 1 dendritic cells, the subset that dominates cross-presentation to CD8 + T cells. This is meaningful only if XCL1 retains receptor engagement once constrained within a fusion protein — an assumption that comparable designs typically test by docking alone, which yields a plausible pose but no benchmark. We therefore evaluated it by replicate molecular dynamics with MM-PBSA interaction energy calculations against a free-chemokine positive control. The anticipated clinical setting is adjuvant therapy or minimal residual disease after curative-intent resection, when tumour burden is low and immune competence best preserved.

Table 1
The four target antigens and the design rationale for each arm.

Target	UniProt	Region used	Arm	Mechanistic role
KRAS	P01116	G12C, G12D, G12V, G13D	Pool A (T cell)	Public neoantigen; generates effector T cells
EGFR	P00533	L858R, delE746_A750	Pool A (T cell)	Neoantigen; broadens patient coverage
PD-L1	Q9NZQ7	IgV domain (PD-1 contact face)	Pool B (antibody)	Checkpoint blockade; permits ADCC
VEGFR2	P35968	Ig domains 2–3 (VEGF-binding face)	Pool B (antibody)	Vascular normalisation; increases lymphocyte infiltration

Results

A Vietnamese HLA reference panel underpins epitope selection

Because the entire epitope selection strategy is restricted to the target population, we first established a Vietnamese HLA reference set. No single study has characterised this population at the resolution

required for vaccine design: the available data arise from transplantation, transfusion or pharmacogenetic objectives and each cohort covers only part of the loci — HLA-DRB1 and -DQB1 in 103 Hanoi Kinh subjects[9], five loci with haplotype estimation in a further 170[8], three-field typing of five loci in 101 southern Kinh subjects[10], and the national cord blood series reporting four loci in 3,750 units[6] and HLA-DQB1 in a further 2,076[11]. VN1K subsequently applied high-depth whole-genome sequencing to 1,011 unrelated Kinh individuals from 53 provinces, resolving 166 alleles at four-digit resolution and providing the only HLA-DPB1 frequencies reported for this population[7].

Concordance across these cohorts is high for the alleles relevant to this design (Fig. 3a). HLA-A*11:01 is the leading class I allele in every study (22.2–25.3%), followed by A*24:02 (9.6–13.8%) and A*33:03 (9.0–12.9%); B*15:02 (11.1–15.1%) and B*46:01 (10.7–11.5%) predominate at HLA-B, and C*07:02, C*01:02 and C*08:01 each fall between 12 and 17%. At class II, DRB1*12:02 is the most frequent DRB1 allele throughout (23.2–35.3%), DRB1*09:01 reaches 9.7–13.6%, DQB1*03:01 dominates its locus (30.8–40.0%) and DPB1*05:01 reaches 33%. One allele diverges systematically: HLA-A*02:01 was reported at 11.2% by low-resolution PCR-SSO typing[6] but at only 2.1–4.6% in three higher-resolution cohorts[7,8], most plausibly because such methods group it with the A*02 variants that are more common here, notably A*02:07 (9.3%) and A*02:03 (7.7%); higher-resolution data were therefore given precedence.

The final reference set comprised 12 class I alleles and 13 class II specificities (Fig. 3a, Supplementary Table S2), selected for high frequency in the Kinh population and for the availability of IEDB prediction matrices; global supertype sets were deliberately not used. No cohort reports DQA1 or DPA1 frequencies, so the values shown for the eight heterodimeric specificities are beta-chain frequencies and therefore upper bounds.

Inclusion in the reference set does not guarantee representation in the construct: the set defines the scope of binding prediction, whereas cassette composition follows from the screening cascade below. Only seven of the 25 specificities were the best-ranked binder for at least one retained epitope (Fig. 3a, Best rank strip; Supplementary Table S7). At class I, the five epitopes mapped to three alleles only: HLA-A*11:01 (three epitopes, percentile ranks 0.70–1.94), C*08:01 (1.67) and C*03:04 (1.96); the remaining nine alleles, including A*02:01, B*15:02 and B*58:01, ranked first for no epitope. At class II, two KRAS G12C epitopes mapped to DRB1*07:01 (ranks 9.27 and 8.12) and one EGFR L858R epitope to DRB1*09:01 (9.64), while two further L858R epitopes mapped to DQA1*01:01–DQB1*06:01 (strong binder, rank 1.60) and DQA1*01:02–DQB1*06:01 (weak binder, rank 4.36); these two additionally bound six and three further specificities, indicating appreciable promiscuity within the DQ group. One L858R epitope bound nothing at a rank of 10% or better. The internal C-ImmSim predictor, run independently on the full construct, corroborated two of these findings: no peptide bound HLA-A*02:01, and the sole HLA-B*15:02 binder (WVRDVVRSM) lies within the XCL1 module rather than the epitope cassette.

The most consequential outcome of screening is that DRB1*12:02 is absent from the final construct. No class II epitope restricted by this specificity simultaneously satisfied the binding, antigenicity, non-toxicity, non-allergenicity and IFN-γ thresholds; the same applied to DRB1*15:02 and DRB1*04:05.

Because DRB1*12:02 accounts for 23.8–35% of DRB1 alleles across Vietnamese cohorts[7], its loss is the single largest coverage penalty of the design and the direct cause of the class II coverage of only 26.76% reported below, despite a reference set built specifically for this population.

Population-restricted screening narrowed 92 candidates to 17

Prediction against the Vietnamese HLA set returned 115 class I and 167 class II neoepitopes meeting two-tool consensus; after deduplication and prioritisation of mutation-spanning peptides, 37 class I and 36 class II candidates entered screening. In parallel, ABCpred identified 34 linear B-cell epitopes on PD-L1 and 106 on VEGFR2 above threshold, of which ten and nine respectively survived intersection with the predefined ligand-binding faces (Supplementary Table S5). In total 92 candidates entered the cascade (37 class I, 36 class II, 19 B cell; by gene, 40 KRAS, 33 EGFR, 10 PD-L1, 9 VEGFR2). Notably, 1,451 class I epitopes were nested within a class II 15-mer of the same variant, a configuration allowing a single synthetic long peptide to deliver both cytotoxic and helper components[12] (Supplementary Tables S3, S4).

The cascade was deliberately conservative and only 17 candidates survived, a pass rate of 18.5% (Supplementary Table S6). Antigenicity was the most stringent filter, eliminating 48 of 92 at a VaxiJen threshold of 0.5, whereas toxicity eliminated almost none (85 of 92 passed) and interferon-gamma positivity removed 19 of 36 class II candidates, reflecting the deliberate preference for Th1 polarisation. The allergenicity layer produced marked disagreement: AllerTOP v2.1 called 51 of 92 peptides non-allergenic whereas AlgPred 2.0 accepted only 5, rejecting even peptides of human self sequence. AlgPred hybrid scores clustered between 0.18 and 0.56 with MERCI motif and BLAST components both zero, showing that the output was governed entirely by a machine-learning component operating at its decision boundary. Because allergenicity classifiers are trained largely on plant- and animal-derived allergens and are known to over-assign short peptides, we adopted AllerTOP as the decisive criterion and record the disagreement explicitly.

BLASTp screening against the human proteome detected no statistically significant similarity beyond the target proteins themselves and their paralogues: the KRAS neoepitopes matched HRAS, NRAS and RIT1 at 88–93% identity, as expected given the high conservation surrounding codon 12, but the mutated residue itself is the point of divergence from the wild-type protein and therefore preserves the tumour specificity of the neoepitope. All E values fell between 23 and 44, that is, no match reached the significance threshold (Supplementary Table S19).

CRUISER-A architecture and predicted developability

The T-cell arm is a 273-residue single-chain construct carrying five class I and three class II epitopes between the amino-terminal XCL1 adjuvant module and the carboxy-terminal foldon trimerisation domain (Table 2, Fig. 2a). Three class I positions were filled with screened epitopes and two with the mutant KRAS decamers VVVGADGVGK (G12D) and VVVGAVGVGK (G12V). Both had been eliminated at peptide-level screening, and their reinstatement is a deliberate, documented exception to the prespecified criteria rather than a screening result: they are the canonical HLA-A*11:01-presented class I

elements of mutant KRAS, have been de-risked clinically in the amphiphile vaccine ELI-002[13], and without them the cellular arm would rest on a single class I KRAS epitope. Of the three cysteines, Cys11 and Cys48 form the conserved lymphotactin disulfide bond (Supplementary Table S14), whereas Cys230 must remain free because it is the G12C mutant residue conferring tumour specificity (Fig. 2b, c).

The construct was predicted to be antigenic, non-toxic, non-allergenic by both AllerTOP v2.1 and AllergenFP v1.1, and highly soluble (SoluProt 0.913; Protein-Sol 0.815), with an instability index of 15.16 well below the threshold of 40 (Supplementary Tables S10, S11, S13). Two caveats attach to these values. The VaxiJen score of 0.5027 only marginally exceeds the 0.5 tumour-model threshold; we treat it as the absence of a contraindication rather than as independent evidence of immunogenicity. Protein-Sol flagged a possible transmembrane region (Kyte–Doolittle 1.63 against 1.6), which we interpret as a false positive arising from the hydrophobic cores of the two folded modules, since neither XCL1 nor foldon is a membrane protein. The CamSol profile showed no dominant aggregation-prone region, with local minima confined to the hydrophobic epitope cassette (Supplementary Tables S10, S11).

Table 2: Architecture of the two vaccine arms. (a) Modular architecture of the Pool A construct CRUISER-A; full epitope sequences and linker boundaries are given in Supplementary Table S8. (b) The four Pool B cyclic peptides with their targets, interaction energies, conformational stability and predicted antibody accessibility. Interaction energies were computed with gmx_MMPBSA under the Poisson–Boltzmann model; Generalized Born values give the same rank order (Supplementary Table S18). RMSD is that of the cyclic peptide alone over the second half of the trajectory.

Module	Residues	Composition	Function
Adjuvant	1–93	XCL1 (lymphotactin)	Targeting to XCR1-positive cDC1
Linker	94–98	GGSGG	Separates adjuvant from epitope cassette
Class I cassette	99–158	M1–M5 joined by AAY linkers	Activation of CD8 + T cells
Linker	159–162	GGGS	Separates the two cassettes
Universal helper	163–175	PADRE (AKFVAAWTLKAAA)	Haplotype-independent T help
Class II cassette	176–235	H1–H3 joined by GPGPG linkers	Activation of CD4 + T cells
Linker	236–240	GGSGG	Separates cassette from trimerisation domain
Trimerisation	241–267	T4 fibritin (foldon)	Multivalent presentation
Purification	268–273	Hexahistidine	Metal affinity chromatography

(a) Pool A: CRUISER-A modular architecture

Epitope	Target	Cyclic peptide sequence	ΔG_PBSA (kcal mol⁻¹)	RMSD (nm)	DiscoTope-3.0
EP092	PD-L1	CNIIQFVHGEEDLKVQHC	−19.93 ± 9.47	0.213 ± 0.033	Positive
EP093	PD-L1	CTVTVPKDLYVVEYGSNC	−19.29 ± 5.30	0.427 ± 0.051	Positive
EP076	VEGFR2	TCAASSGLMTKKNSTFC	−18.41 ± 22.52	0.335 ± 0.113	Positive
EP078	VEGFR2	CDQHGVVYITENKNKTVC	−11.74 ± 6.85	0.395 ± 0.045	Negative

Combined population coverage reaches 73.35% and is extended further by PADRE

The retained epitope set covered 63.61% of the Vietnam population at class I (mean 1.62 hits) and 26.76% at class II (0.46 hits); combining the two classes raised coverage to 73.35% with a mean of 2.08 hits, so approximately three quarters of the population are predicted to recognise at least one epitope–HLA combination (Fig. 3b–d). The gain over either class alone indicates that the two cassettes are complementary.

Submitting the identical input for three reference populations tested whether this advantage derives from population restriction rather than being a generic property of the epitope set. Combined coverage fell from 73.35% in Vietnam to 46.78% worldwide and 41.23% in Europe, almost entirely through class I (63.61% against 29.96% and 20.90%), reflecting the high East Asian frequencies of HLA-A*11:01, C*08:01 and C*03:04. The mean number of epitopes recognised per individual was 2.08 in Vietnam against 0.95 worldwide and 0.73 in Europe, and the proportion recognising no epitope was 26.65% against 53.22% and 58.77%. Class II coverage, by contrast, was almost unchanged across populations (24.01–26.76%), because the cassette is restricted by only two specificities that offset one another geographically: DRB1*09:01, common in East Asia, and DRB1*07:01, common in Europe.

The 73.35% figure is therefore a conservative estimate rather than an upper bound, because the coverage algorithm scores only epitopes with allele frequency data and cannot account for PADRE, engineered to bind most human HLA-DR molecules rather than a single specificity[14] and thus supplying essentially haplotype-independent help. Pool B extends applicability further: recognition by surface immunoglobulin is not MHC-restricted, so even individuals recognising no Pool A epitope–HLA combination may benefit from antibodies blocking PD-L1 and VEGFR2, although isotype switching and affinity maturation still require the CD4 + T-cell help supplied by PADRE and the class II cassette. Effective coverage is thus the union of three layers: HLA-restricted neoepitopes, haplotype-independent help, and MHC-unrestricted antibody recognition.

XCL1 retains XCR1 engagement within the fusion protein

The strategy of using XCL1 as an adjuvant is valid only if the chemokine can still engage its receptor once constrained within a fusion protein. Docking against XCR1 extracted from the cryo-EM structure 9AST yielded a representative complex in which the XCL1 core packs against the extracellular face of the receptor. To test whether this conformation is stable, we performed three independent molecular dynamics simulations.

The receptor was essentially unchanged across all three runs, with backbone root-mean-square deviations of 0.249, 0.245 and 0.256 nm, confirming preservation of the seven-transmembrane fold. The vaccine component fluctuated more (0.661–0.806 nm), and per-residue profiles localised the mobile peaks to the linkers and carboxy-terminal tail rather than to the XCL1 core or the interface. Two runs gave stable whole-complex trajectories (0.614 ± 0.036 and 0.494 ± 0.120 nm) while the third fluctuated markedly (1.076 ± 1.235 nm) through a rearrangement of the epitope tail. Critically, the interface hydrogen bond network persisted continuously in all three runs (means 10.44 ± 2.12, 10.09 ± 4.14 and 15.81 ± 2.57), showing that the interface withstood even that rearrangement (Fig. 4a–d).

MM-PBSA interaction energies reinforced this: −76.06, − 97.56 and − 96.09 kcal mol^{-1}, averaging − 89.9 ± 12.0 kcal mol^{-1}. Since the configurational entropy term was not computed, what is meaningful is the consistently negative sign and comparable magnitude across three independent runs rather than the absolute value. Decomposition showed that binding is driven principally by van der Waals interactions (− 211.51 ± 0.88 kcal mol^{-1}) with a non-polar solvation contribution (− 22.99 ± 0.07 kcal mol^{-1}), while the in vacuo electrostatic term (+ 658.26 ± 3.97) is almost entirely offset by desolvation (− 499.82 ± 3.55) — a compensation pattern characteristic of an extensive interface based on shape complementarity and consistent with chemokine binding to G protein-coupled receptors. The first two principal components accounted for 64.8%, 97.6% and 81.6% of the variance, with free energy landscapes showing well-separated wells (Fig. 4e–f).

To test whether this reflects genuine retention of receptor engagement upon fusion, a positive control comprising free, unfused XCL1 bound to XCR1 was simulated under an identical 100 ns protocol. The control also yielded a clearly negative interaction energy with the same decomposition pattern, and the receptor and the XCL1 module itself fluctuated to equivalent extents in both systems, indicating that appending the epitope cassette does not distort the chemokine core (Supplementary Table S17).

The fusion complex gave a more negative interaction energy than the free chemokine and maintained almost twice as many interface hydrogen bonds. We interpret this difference conservatively: MM-PBSA energies scale with buried interface area, and a 273-residue construct necessarily buries more surface than 93-residue XCL1, so part of the difference may reflect contacts contributed by neighbouring modules rather than any higher intrinsic affinity of the XCL1 core. The control also carries a large standard deviation that overlaps the lower bound of the fusion complex. What can be concluded firmly is precisely what the control was designed to answer: fusion does not disrupt the XCR1 interaction, since the fusion system achieves an interaction energy at least equivalent to, and an interface more stable than, that of the free chemokine. Notably, the control lost hydrogen bonds transiently between

approximately 58 and 82 ns, a phenomenon absent from the fusion complex, suggesting that the fusion context may itself stabilise the bound conformation.

Immune simulation predicts a Th1-polarised, memory-forming response

C-ImmSim simulation of the three-dose schedule reproduced the classical primary and secondary response pattern: immunoglobulin M predominated after the first dose, isotype switching to immunoglobulin G occurred at the boosters, titres peaked after the third dose well above the primary response, and antigen was cleared rapidly after each dose, with memory B-cell and memory helper T-cell compartments maintained to the end of the simulation. The decisive readouts for a cancer vaccine are cellular: the cytotoxic T-lymphocyte population was sustained throughout the activation phase before returning to rest with a long-lived memory subset, and interferon-gamma rose sharply after each dose and remained the most abundant cytokine alongside interleukin-2, confirming the intended Th1 polarisation. A low Simpson diversity index indicates a response focused on a limited set of epitopes (Fig. 5a, Supplementary Fig. S4).

The codon-adapted sequence for E. coli K12 was 819 base pairs with a codon adaptation index of 0.919 and a guanine-cytosine content of 52.26% (host genome 50.73%), both favourable for high-level heterologous expression, and back-translation reproduced the 273-residue chain exactly. In silico cloning into pET-28a(+) between NdeI and XhoI generated a 6,203 base pair plasmid under T7 control (Fig. 5b). Because the construct carries a structural disulfide bond, we propose an oxidising-cytoplasm strain such as SHuffle T7 Express rather than BL21(DE3), allowing the Cys11–Cys48 bond to form in the cytoplasm without refolding from inclusion bodies and requiring no change of vector.

Structure-first selection yields four cyclic peptides blocking PD-L1 and VEGFR2

Unlike Pool A, Pool B targets wild-type epitopes and was selected entirely on structural criteria. Placing structure ahead of antigenicity proved decisive: two of four initial candidates were eliminated for the same reason, a high VaxiJen score combined with a location outside the ligand-binding face. EP085 scored 1.10 but mapped to the IgC domain of PD-L1, away from the PD-1 contact face on IgV, and was replaced by EP093; EP082 scored 0.58 but was absent from both reference structures 3V2A and 2X1W, lying outside the VEGFR2 D2–D3 domains, and was replaced by EP076. The structural gate caught both before any modelling effort was expended. EP076 does not belong to the 17 epitopes that passed the four-layer cascade; it was introduced at the structure-guided reselection step to replace EP082. Three of four allergenicity tools classified it as allergenic whereas AllerCatPro 2.0 returned no evidence of allergenicity at 100% confidence. We retained it with an explicit risk annotation because it is the only candidate both on the VEGF-binding face and sufficiently surface-exposed, but regard this as a documented exception requiring experimental allergenicity testing before any development step.

All four peptide–receptor complexes yielded favourable interaction energies, the two PD-L1-targeting peptides achieving the most negative values (Table 2, Fig. 6), and peptide-only simulations confirmed the stabilising effect of the disulfide bond (root-mean-square deviations 0.213–0.427 nm over three 50 ns

runs). Three of four achieved positive DiscoTope-3.0 scores; EP078 scored negative, consistent with most of its residues being buried within the receptor rather than with any design error. All four were predicted non-toxic with codon adaptation indices of 0.911–1.000. The large standard deviation for EP076 reflects appreciable conformational fluctuation, so its mean should be read with caution.

Discussion

Three deliberate departures from established multi-epitope practice[15] define this design: restriction to a population-specific HLA panel with the gain quantified rather than asserted, a positive-controlled test of the assumption underlying chemokine-based dendritic cell targeting, and separation of the cellular and humoral arms so that each follows its own selection logic.

Population restriction is the first. Most published designs, including those addressing the same driver genes, rest on global supertypes or on European-weighted allele sets. Evaluated against three reference populations with identical input, the CRUISER-A epitope set achieved combined coverage of 73.35% in Vietnam but only 46.78% worldwide and 41.23% in Europe, with 2.2- to 2.8-fold more epitopes recognised per individual. The gap is concentrated at class I and traces to the East Asian frequencies of HLA-A*11:01 and two HLA-C alleles. Population optimisation is therefore not a procedural refinement but a determinant of predicted potency: a construct designed for Vietnamese patients would forfeit roughly one third of its coverage in Europe, and the converse holds equally — an argument for developing therapeutic cancer vaccines population by population rather than deploying one supertype-based design globally.

The second is direct evidence that the adjuvant module remains functional. Chemokine fusion to route antigen to cDC1 is immunologically well founded, but its value rests on a rarely interrogated assumption: that the chemokine still engages its receptor once conformationally constrained within a fusion protein. Comparable studies stop at docking, which returns a plausible pose but no reference point. Triplicate 100 ns simulations showed that XCR1 retained its seven-transmembrane fold, that the interface hydrogen bond network persisted, and that the MM-PBSA interaction energy remained consistently negative; decisively, a free XCL1:XCR1 positive control under an identical protocol established that the fusion complex matches the free chemokine while distorting neither the chemokine core nor the receptor. The assumption underlying the targeting strategy was thus tested rather than presumed. We decline to read the difference in magnitude between the two systems as higher affinity: MM-PBSA energies scale with buried interface area and the fusion construct is nearly threefold longer than free XCL1, so the defensible conclusion is that fusion does not abrogate receptor binding.

The third is the division into two arms. Because blocking antibodies must engage a folded receptor surface whereas T-cell epitopes must first be liberated by antigen processing, the arms impose conflicting structural demands and are formulated separately. For Pool B this let structural criteria take precedence over antigenicity, removing two high-scoring candidates that mapped outside the ligand-contact face. The point generalises: a high antigenicity score is no guarantee that an epitope lies on the

surface that must be blocked, and structure-first selection should be the default whenever the intended mechanism is receptor blockade.

Two validation metrics warrant context. The AlphaFold model returns a mean pLDDT of approximately 42 and a pTM of 0.31, values that appear alarming in isolation, but the low-confidence regions are the linkers and epitope cassette, designed to be flexible, whereas the XCL1 module exceeds pLDDT 80 (Supplementary Table S12, Supplementary Fig. S3). Verify3D likewise reached only 64.84% against a threshold of 80% while MolProbity, clashscore and the Ramachandran distribution were all excellent; Verify3D penalises extended unfolded segments, which is to say precisely the linkers. The model should be read as reliable in the folded core and indicative in the epitope tail.

The work is entirely computational and every conclusion is a prediction. The MM-PBSA values reported here exclude the configurational entropy term and are usable only for relative comparison, not for conversion into dissociation constants, and the detailed decomposition derives from a single representative run. In the positive control the two systems do not present equivalent interface areas, so the difference in magnitude cannot be attributed to affinity; a rigorous comparison would require normalisation to buried surface area or per-residue decomposition. The control also carries a large standard deviation overlapping the fusion complex, and one of three simulations fluctuated markedly in the epitope tail.

In design terms, class II coverage reached only 26.76% because just two DRB1 specificities survived screening, the interferon-gamma requirement alone eliminating 19 of 36 candidates. The inclusion of PADRE is the design element intended to absorb this deficit: because it engages most common HLA-DR molecules rather than a single specificity, it supplies CD4 + T-cell help largely independently of the recipient haplotype, so the functional consequence of the shortfall — insufficient help for cytotoxic T-lymphocyte priming and for antibody maturation in Pool B — is mitigated even though the deficit in tumour-specific class II neoepitopes is not itself repaired. This compensation is nonetheless partial: the promiscuity of PADRE derives from predominantly European allele sets and has not been verified against East Asian DR specificities; binding predictors are likewise trained mainly on European and North American alleles. Multi-epitope architecture also generates junctional neo-epitopes: C-ImmSim identified SAAAYVAIK, spanning the AAY linker boundary, as an HLA-A*11:01 binder that belongs to no designed epitope and carries no mutant residue, and we did not optimise linker sequences to eliminate this. C-ImmSim reproduces the general behaviour of an immune response but does not model the immunosuppressive tumour microenvironment.

In manufacturing terms, Cys230 within the KRAS G12C epitope is an obligate free cysteine; because foldon trimerises, each particle carries three free thiols, raising the risk of intermolecular disulfide formation and aggregation during purification. This requires a defined redox buffer and non-reducing electrophoresis monitoring; if intractable, secreted mammalian expression is the alternative, at the cost of codon re-optimisation and a signal peptide. Pool B targets wild-type epitopes and therefore carries a higher autoimmune risk in principle — a deliberate trade-off for blocking function that demands close

safety monitoring — and EP076 in particular lies outside the screened epitope set and returned discordant allergenicity results. Because both ligand-binding faces are largely buried, the peptides overlap only partially with the true contact surface.

In summary, an immunoinformatics workflow (Fig. 1) restricted to the Vietnamese HLA panel yielded CRUISER, a dual-arm vaccine against KRAS, EGFR, PD-L1 and VEGFR2 in which 17 of 92 candidate epitopes survived four-layer screening, combined population coverage reached 73.35%, and a positive-controlled molecular dynamics test established that XCL1 fusion preserves XCR1 engagement — evidence that goes materially beyond the docking-only standard of the field. Validation should proceed through peptide–HLA binding and stability assays, interferon-gamma ELISpot and intracellular cytokine staining on HLA-matched Vietnamese peripheral blood mononuclear cells, immunogenicity testing in HLA-A*11:01 transgenic mice, and, for Pool B, ELISA with ligand-competition blocking. Until those data exist, every finding here remains a prediction.

Materials and methods

All predictions were performed between July and August 2026. Server versions, access addresses, job identifiers and access dates for every tool are compiled in Supplementary Table S1, and the complete parameter set for each step in the Supplementary Methods. The study is entirely computational and involved no human participants, human material or animals; no ethical approval was therefore required. Throughout, values expressed as mean ± are standard deviations unless stated otherwise, and molecular dynamics statistics are computed across frames of the equilibrated trajectory segment rather than across independent biological replicates.

Target sequences, HLA set and peptide library

Canonical human protein sequences were retrieved from UniProt[16] for KRAS (P01116), EGFR (P00533), PD-L1 (Q9NZQ7) and VEGFR2 (P35968). Four KRAS hotspot substitutions (G12C, G12D, G12V, G13D) and two EGFR alterations (L858R, delE746_A750) were introduced in silico into the canonical sequences. For the antibody arm the epitope search was restricted a priori to the IgV domain of PD-L1 and to Ig domains 2–3 of VEGFR2, that is, the regions forming the ligand-contact faces, because only epitopes located on those surfaces can support functional blockade. A population-specific allele set comprising 12 class I alleles and 13 class II specificities was compiled from published HLA typing studies of the Vietnamese population and from the Allele Frequency Net Database[17] (Supplementary Table S2); alleles were selected on the joint criteria of high frequency in the Kinh population and availability of a validated prediction matrix in IEDB. This set was applied uniformly to every subsequent prediction, and global supertype sets were deliberately not used. Overlapping peptides were generated with a one-residue sliding window (8–11-mers for class I and 15-mers for class II), retaining 224 mutation-bearing peptides for KRAS and 1,225 for EGFR after removal of sequences identical to the wild-type proteome.

Epitope prediction

MHC class I binding was predicted independently by two algorithms to reduce single-tool bias: MHCflurry 2.0[18] run locally with antigen processing and presentation scoring, and NetMHCpan-4.1[19]. Class II binding was assessed with NetMHCIIpan-4.3[19] and MHCnuggets[20]. A peptide was carried forward only if both tools of the same class predicted binding at the prespecified percentile-rank thresholds: for class I, strong binders were defined as %Rank < 0.5 and weak binders as %Rank ≤ 2.0; for class II, strong binders as %Rank < 2 and weak binders as %Rank ≤ 10. Peptide–MHC complex stability, a better correlate of immunogenicity than affinity alone, was estimated with NetMHCstabpan[21]. Linear B-cell epitopes were predicted across the full PD-L1 and VEGFR2 sequences with ABCpred[22] (16-mer window, threshold 0.51) and BepiPred-3.0[23] at default settings, then intersected with the predefined ligand-binding faces. Conformational epitopes were predicted with ElliPro[24] (minimum score 0.5; maximum distance 6 Å) and DiscoTope-3.0[25] on the reference complexes PD-1:PD-L1 (4ZQK) and VEGF:VEGFR2 (3V2A, 2X1W).

Safety and efficacy screening

Candidate epitopes passed through a four-layer screening cascade applied in fixed order: antigenicity (VaxiJen v2.0[26], tumour model, threshold 0.5), toxicity (ToxinPred 1.0[27] and 3.0[28] and ToxDL[29], with a consensus criterion of at least two of three tools), allergenicity (AllerTOP v2.1[30] and AlgPred 2.0[31]) and, for class II epitopes only, cytokine-inducing capacity (IFNepitope[32], IL4pred[33]) so as to favour Th1 polarisation. Thresholds were fixed before screening began and were not revised in the light of the results; the two documented departures from these criteria are stated explicitly in the Results. Where the two allergenicity classifiers disagreed, AllerTOP was treated as decisive and the disagreement recorded rather than resolved silently. Each retained epitope was compared against the human proteome by local BLASTp[34] (BLOSUM62, parameters adjusted for short sequences) to assess the risk of autoimmune cross-reactivity, and hits were considered significant only below the standard E-value threshold. Population coverage was computed with IEDB Population Coverage v3.0.2[35] run locally, separately for MHC class I, MHC class II and the combined set. To test whether the construct was genuinely optimised for the target population rather than broadly immunogenic, the identical epitope–allele input was submitted for three reference populations — Vietnam, World and Europe — and the outputs compared directly.

Construct assembly, modelling and validation

CRUISER-A was assembled as a single chain in the following amino-to-carboxy order: the XCL1 adjuvant module[36]; five MHC class I epitopes joined by AAY linkers, which match proteasomal cleavage specificity and so favour correct epitope liberation; the universal helper epitope PADRE[14]; three MHC class II epitopes joined by GPGPG linkers, which disrupt junctional neo-epitope formation and promote independent processing of adjacent epitopes; the T4 fibritin trimerisation domain (foldon)[37], included to confer multivalent antigen display; and a carboxy-terminal hexahistidine tag for affinity purification. Cassettes were separated by flexible GGSGG or GGGS linkers. The native XCL1 secretory signal peptide

was deliberately removed because the construct is intended for cytoplasmic expression in bacteria. The resulting chain is 273 residues. Construct-level antigenicity, toxicity and allergenicity were evaluated with VaxiJen[26], ToxinPred 2.0[38], AllerTOP v2.1[30] and AllergenFP v1.1[39]; physicochemical properties with ExPASy ProtParam[40]; solubility with SoluProt v1.0[41] and Protein-Sol[42]; and aggregation propensity with CamSol[43] in intrinsic mode. Secondary structure was predicted with SOPMA[44,45]. The three-dimensional model was built with AlphaFold2[46] through ColabFold 1.6.2 implementing AlphaFold2-Multimer, model type alphafold2_multimer_v3[47], with multiple sequence alignments generated through the MMseqs2 server (msa_mode mmseqs2_uniref_env), a single model, three recycles and one seed, no template information and no Amber relaxation (num_relax = 0); models were ranked by the multimer confidence metric and confidence assessed by pLDDT, pTM and interface predicted TM-score (ipTM). Peptide–MHC complexes, each comprising the peptide, the HLA-A*11:01 α-chain and β2-microglobulin, were modelled under the same settings. The highest-ranking model was refined in two successive steps with ModRefiner[48] and GalaxyRefine[49], and the refined model was validated with SAVES v6.1 (ERRAT[50], Verify3D[51], PROCHECK[52]), ProSA-web[53] and MolProbity[54]. Candidate disulfide bonds were identified with Disulfide by Design 2.13[55] using the geometric criteria $\chi3 = -87°$ or $+97°$ ($\pm 30°$) and Cα–Cβ–Sγ = 114.6° ($\pm 10°$), with an energy cut-off of 2.2 kcal mol^{-1}.

Docking, molecular dynamics and interaction energy

The refined CRUISER-A model was docked with the human XCR1 receptor extracted from the cryo-EM structure of the XCL1:XCR1 complex (PDB 9AST, chain R, residues 21–303)[56] using GRAMM[57], with binding-pocket residues identified directly from that structure applied as restraints (Supplementary Tables S15, S16) Protein–protein docking of the CRUISER-A construct (XCL1 arm) onto its receptor XCR1 was performed with GRAMM v2.0 (Global Range Molecular Matching) in global docking mode. A rigid-body fast Fourier transform search was run with a grid step of 3.5 Å, a grid size of 64, and a rotational sampling interval of 10°, generating 30,000 candidate matches. Poses were clustered, and the ten top-ranked cluster representatives were retained for inspection; the highest-ranked cluster was selected as the representative receptor–ligand complex and carried forward to molecular dynamics refinement. The representative pose of the largest cluster was carried forward. Simulations were run in GROMACS 2023.5[58] with the AMBER99SB-ILDN force field in explicit solvent with neutralising counter-ionsBecause XCR1 is a seven-transmembrane receptor, the complex was embedded in a POPC:cholesterol (4:1) bilayer using PACKMOL-Memgen, solvated in a periodic box with a minimum solute–box distance of 15 Å (17.5 Å for the water layer) and neutralised with KCl at a physiological concentration of 0.15 M. After energy minimisation the system was equilibrated for 500 ps in the NVT ensemble (V-rescale thermostat, 310 K) and a further 500 ps in the NPT ensemble (C-rescale barostat, semi-isotropic coupling, 1.0 bar). Three independent production runs of 100 ns were then performed from different initial velocity seeds, giving 300 ns of sampling, with a 2 fs integration time step. Temperature was held at 310 K (V-rescale, $\tau = 0.1$ ps) with the protein–membrane and solvent groups coupled separately, and pressure at 1.0 bar (C-rescale, semi-isotropic, $\tau = 5.0$ ps). Bonds to hydrogen were constrained with LINCS, and long-range electrostatics were treated by particle-mesh Ewald with real-space Coulomb and van der Waals cut-offs

of 1.0 nm. Three independent production runs of 100 ns were performed from different initial velocity seeds, giving 300 ns of sampling. Trajectories were analysed with gmx rms, rmsf, gyrate, sasa and hbond, and all reported statistics were computed over the final 50 ns of each trajectory, by which point the backbone deviation had plateaued. Interaction energies were estimated with gmx_MMPBSA[59] under the Poisson–Boltzmann model on frames extracted from the equilibrated segment. Binding free energies were estimated by the MM-PBSA method using gmx_MMPBSA 1.6.5. Calculations used the ff14SB force field and sampled 154 evenly spaced frames from the final production trajectory (every 13th frame between frames 1 and 2000). The Poisson-Boltzmann equation was solved with an interior dielectric constant of 1.0, an exterior dielectric constant of 80.0, and an ionic strength of 0.15 M; the non-polar solvation free energy was estimated from the solvent-accessible surface area (default SASA-based model). It must be emphasised that the configurational entropy term (− TΔS) was not computed; the reported values are therefore effective interaction energies suitable for relative comparison between systems and between replicates, and are not absolute binding free energies comparable with experimental dissociation constants. Collective motions were examined by principal component analysis with free energy landscapes derived from the first two components (gmx covar, anaeig, sham). To provide a benchmark for the claim that fusion does not disrupt the receptor interaction, a positive control was run in parallel under an identical protocol: free, unfused XCL1 was docked and simulated with XCR1 for 100 ns and its interaction energy computed with the same gmx_MMPBSA settings. The two systems were compared on equivalent metrics — interaction energy, interface hydrogen bond count, and per-component root-mean-square deviation — chosen so that any difference could be attributed to the fusion context rather than to differences in protocol.

Immune simulation and codon optimisation

The theoretical immune response was simulated using C-ImmSim v10.1[60] with a three-dose schedule at time steps 1, 84 and 168 (each step corresponding to eight hours, that is, four-week intervals). The host haplotype (HLA-A*02:01, A*11:01, B*15:02, B*58:01, DRB1*07:01, DRB1*15:02) was selected from the allele list supported by C-ImmSim, prioritising alleles that were also present in the Vietnamese reference set; the platform permits only four MHC class I and two MHC class II molecules to be specified and does not support the HLA-C locus, DRB1*09:01 or DQ specificities, so the simulated haplotype does not fully coincide with the epitope-restricting specificities reported in the Results. Remaining parameters were: random seed 12345, volume 10, 1,050 total steps, 1,000 antigen molecules per dose, adjuvant 100, and no lipopolysaccharide. The sequence was codon-adapted for Escherichia coli K12 using JCat[61] and cloned in silico into pET-28a(+) between NdeI and XhoI under the control of the T7 promoter. Because the construct contains a structural disulfide bond within the XCL1 module, the proposed expression strategy uses a strain with an oxidising cytoplasm such as SHuffle T7 Express, in which the trxB/gor genetic background combined with cytoplasmic disulfide isomerase DsbC permits disulfide bond formation directly in the cytoplasm without a refolding step; this option preserves the vector backbone and changes only the host strain.

[TO BE UPDATED] The C-ImmSim simulation is being re-run with an adjusted host haplotype prioritising the specificities that genuinely restrict the construct epitopes (HLA-A*11:01 and DRB1*07:01, with DRB1*09:01 added if supported by the platform) and excluding HLA-A*02:01, which bound no peptide. Once the new results are available, this sentence, the immune simulation subsection of the Results and Fig. 5a must be updated consistently; if the response curves are essentially unchanged between the two haplotypes, this should be stated in the Discussion as a robustness check on the conclusions.

Pool B design

Pool B comprises four peptides derived from the ligand-binding contact faces of PD-L1 and VEGFR2. Unlike Pool A these derive from wild-type sequence and were therefore selected on structural rather than antigenicity criteria, since a peptide can only support functional blockade if it reproduces part of the true contact surface. Contact residues were defined directly from the reference crystal complexes using a 5 Å heavy-atom cut-off, and only segments that overlapped this contact face, were solvent-exposed and achieved a B-cell epitope score were retained. Each peptide was cyclised through a disulfide bond between two terminal cysteines in order to restrain it towards the receptor-bound conformation. Structures were built with PEP-FOLD4[62], docked with the corresponding receptor using GRAMM, and simulated in GROMACS over three independent 50 ns runs under the protocol described above. Interaction energies were computed with gmx_MMPBSA under both the Generalized Born and Poisson–Boltzmann models, with the same caveat regarding the omitted entropy term; both models were reported so that the rank order of the four peptides could be shown to be independent of the solvation treatment. Developability was assessed with ProtParam[40], PepCalc and AGGRESCAN3D[63].

Declarations

Funding

No funding was received.

Competing interests

The authors declare no competing interests.

Data availability

The complete predicted epitope datasets, screening tables, construct sequences, structural models, simulation trajectories, and docking outputs are publicly available in the Zenodo repository at https://doi.org/10.5281/zenodo.21853299. The source protein sequences were retrieved from UniProt under accessions P01116, P00533, Q9NZQ7, and P35968. The XCR1 receptor structure is available from the Protein Data Bank under accession 9AST.

Author Contribution

L.P.T. conceived and designed the study, supervised the work and administered the project. D.L.V. performed epitope prediction and screening, analysed the data and wrote the original draft. H.N.X. and T.L.H. compiled the Vietnamese population HLA reference set and computed population coverage. G.N.H. and T.N.N. assembled the CRUISER-A construct and assessed construct-level physicochemical, antigenicity, toxicity and allergenicity properties. L.N.T.P. and H.N.V. performed three-dimensional modelling, refinement, structure validation and disulfide design. K.L.V. and T.T.T. performed molecular docking, molecular dynamics simulations and MM-PBSA interaction energy calculations. G.B.T.H. performed immune simulation and codon optimisation. A.D.N.M. and H.T.V. designed and evaluated the Pool B cyclic peptides. K.N.V. and T.P.V. curated the data and compiled the supplementary material. L.P.T. and D.L.V. reviewed and edited the manuscript. All authors read and approved the final manuscript.

References

1. Sung, H. et al. Global cancer statistics 2024: GLOBOCAN estimates of incidence and mortality worldwide for 34 cancers in 186 countries. *CA Cancer J. Clin.* **76**, e70090 (2026).
2. Luo, G. et al. Estimated worldwide variation and trends in incidence of lung cancer by histological subtype in 2022 and over time: a population-based study. *Lancet Respir Med.* **13**, 348–363 (2025).
3. Ervik, M. *et al. Global Cancer Observatory: Cancer Today (2024 Estimates, Version 1.0) — Viet Nam Fact Sheet* (International Agency for Research on Cancer, 2026); https://gco.iarc.who.int/today (accessed 8 August 2026).
4. Dang, A. T. H. et al. Actionable Mutation Profiles of Non-Small Cell Lung Cancer patients from Vietnamese population. *Sci. Rep.* **10**, 2707 (2020).
5. Thạch, N. H. & Lộc, P. T. Vaccine tân kháng nguyên (neoantigen) trị liệu ung thư – triển vọng tương lai và thách thức. *YHCĐ* 67, (2026).
6. Que, T. N. et al. Allele and Haplotype Frequencies of HLA-A, -B, -C, and -DRB1 Genes in 3,750 Cord Blood Units From a Kinh Vietnamese Population. *Front. Immunol.* **13**, 875283 (2022).
7. Tran, T. T. H. et al. VN1K is a pangenome-informed multi-omics and phenomics resource for the Vietnamese population. *Nat. Commun.* **17**, 6679 (2026).
8. Hoa, B. K. et al. HLA-A, -B, -C, -DRB1 and -DQB1 alleles and haplotypes in the Kinh population in Vietnam. *Tissue Antigens* 71, 127–134 (2008).
9. Vu-Trieu, A. et al. HLA-DR and -DQB1 DNA polymorphisms in a Vietnamese Kinh population from Hanoi. *Eur. J. Immunogenet.* **24**, 345–356 (1997).
10. Do, M. D. et al. High-Resolution HLA Typing of HLA-A, -B, -C, -DRB1, and -DQB1 in Kinh Vietnamese by Using Next-Generation Sequencing. *Front Genet* 11, 383 (2020).
11. Que, T. N. et al. Frequency and distribution of HLA-DQB1 alleles from 2076 cord blood samples of the Vietnamese cohort. *Int. J. Immunogenet.* **49**, 340–344 (2022).

12. Melief, C. J. M. & van der Burg, S. H. Immunotherapy of established (pre)malignant disease by synthetic long peptide vaccines. *Nat. Rev. Cancer*. **8**, 351–360 (2008).
13. Pant, S. et al. Lymph-node-targeted, mKRAS-specific amphiphile vaccine in pancreatic and colorectal cancer: the phase 1 AMPLIFY-201 trial. *Nat. Med.* **30**, 531–542 (2024).
14. Alexander, J. et al. Development of high potency universal DR-restricted helper epitopes by modification of high affinity DR-blocking peptides. *Immunity* **1**, 751–761 (1994).
15. Mahafujul Alam, S. S. et al. Immunoinformatics based designing of a multi-epitope cancer vaccine targeting programmed cell death ligand 1. *Sci. Rep.* **15**, 12420 (2025).
16. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).
17. Gonzalez-Galarza, F. F. et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783–D788 (2020).
18. O'Donnell, T. J., Rubinsteyn, A. & Laserson, U. MHCflurry 2.0: Improved Pan-Allele Prediction of MHC Class I-Presented Peptides by Incorporating Antigen Processing. *Cell. Syst.* **11**, 42–48e7 (2020).
19. 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.* **48**, W449–W454 (2020).
20. Shao, X. M. et al. High-Throughput Prediction of MHC Class I and II Neoantigens with MHCnuggets. *Cancer Immunol. Res.* **8**, 396–408 (2020).
21. Rasmussen, M. et al. Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J. Immunol.* **197**, 1517–1524 (2016).
22. Saha, S. & Raghava, G. P. S. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* **65**, 40–48 (2006).
23. Clifford, J. N. et al. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci.* **31**, e4497 (2022).
24. Ponomarenko, J. et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinform.* **9**, 514 (2008).
25. Høie, M. H. et al. DiscoTope-3.0: improved B-cell epitope prediction using inverse folding latent representations. *Front. Immunol.* **15**, 1322712 (2024).
26. Doytchinova, I. A. & Flower, D. R. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinform.* **8**, 4 (2007).
27. Gupta, S. et al. In silico approach for predicting toxicity of peptides and proteins. *PLoS One*. **8**, e73957 (2013).
28. Rathore, A. S., Choudhury, S., Arora, A., Tijare, P. & Raghava, G. P. ToxinPred 3.0: An improved method for predicting the toxicity of peptides. *Comput. Biol. Med.* **179**, 108926 (2024).

29. Pan, X. et al. ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. *Bioinformatics* **36**, 5159–5168 (2021).
30. Dimitrov, I., Bangov, I., Flower, D. R. & Doytchinova, I. AllerTOP v.2–a server for in silico prediction of allergens. *J. Mol. Model.* **20**, 2278 (2014).
31. Sharma, N. et al. AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes. *Brief. Bioinform*. **22**, bbaa294 (2021).
32. Dhanda, S. K., Vir, P. & Raghava, G. P. Designing of interferon-gamma inducing MHC class-II binders. *Biol. Direct*. **8**, 30 (2013).
33. Dhanda, S. K., Gupta, S., Vir, P. & Raghava, G. P. S. Prediction of IL4 inducing peptides. *Clin Dev Immunol* 263952 (2013). (2013).
34. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
35. Bui, H. H. et al. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinform.* **7**, 153 (2006).
36. Dorner, B. G. et al. Selective expression of the chemokine receptor XCR1 on cross-presenting dendritic cells determines cooperation with CD8 + T cells. *Immunity* **31**, 823–833 (2009).
37. Güthe, S. et al. Very fast folding and association of a trimerization domain from bacteriophage T4 fibritin. *J. Mol. Biol.* **337**, 905–915 (2004).
38. Sharma, N., Naorem, L. D., Jain, S. & Raghava, G. P. ToxinPred2: an improved method for predicting toxicity of proteins. *Brief. Bioinform*. **23**, bbac174 (2022).
39. Dimitrov, I., Naneva, L., Doytchinova, I. & Bangov, I. AllergenFP: allergenicity prediction by descriptor fingerprints. *Bioinformatics* **30**, 846–851 (2014).
40. Gasteiger, E. et al. Protein Identification and Analysis Tools on the ExPASy Server. in The Proteomics Protocols Handbook (ed Walker, J. M.) 571–607 (Humana, (2005).
41. Hon, J. et al. SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **37**, 23–28 (2021).
42. Hebditch, M., Carballo-Amador, M. A., Charonis, S., Curtis, R. & Warwicker, J. Protein-Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics* **33**, 3098–3100 (2017).
43. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
44. Geourjon, C. & Deléage, G. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* **11**, 681–684 (1995).
45. Combet, C., Blanchet, C., Geourjon, C. & Deléage, G. NPS@: network protein sequence analysis. *Trends Biochem. Sci.* **25**, 147–150 (2000).
46. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

47. Mirdita, M. et al. ColabFold: making protein folding accessible to all. *Nat. Methods*. **19**, 679–682 (2022).
48. Xu, D. & Zhang, Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. *Biophys. J.* **101**, 2525–2534 (2011).
49. Heo, L., Park, H., Seok, C. & GalaxyRefine Protein structure refinement driven by side-chain repacking. *Nucleic Acids Res.* **41**, W384–388 (2013).
50. Colovos, C. & Yeates, T. O. Verification of protein structures: patterns of nonbonded atomic interactions. *Protein Sci.* **2**, 1511–1519 (1993).
51. Lüthy, R., Bowie, J. U. & Eisenberg, D. Assessment of protein models with three-dimensional profiles. *Nature* **356**, 83–85 (1992).
52. Laskowski, R. A. et al. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
53. Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res.* **35**, W407–410 (2007).
54. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
55. Craig, D. B. & Dombkowski, A. A. Disulfide by Design 2.0: a web-based tool for disulfide engineering in proteins. *BMC Bioinform.* **14**, 346 (2013).
56. Zhang, X. et al. Molecular basis for chemokine recognition and activation of XCR1. *Proc. Natl. Acad. Sci. U S A*. **121**, e2405732121 (2024).
57. Singh, A., Copeland, M. M., Kundrotas, P. J. & Vakser, I. A. GRAMM Web Server for Protein Docking. *Methods Mol. Biol.* **2714**, 101–112 (2024).
58. Abraham, M. J. et al. High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX 1–2*. **GROMACS**, 19–25 (2015).
59. Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A. & Moreno, E. gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *J. Chem. Theory Comput.* **17**, 6281–6291 (2021).
60. Rapin, N., Lund, O., Bernaschi, M. & Castiglione, F. Computational immunology meets bioinformatics: the use of prediction tools for molecular binding in the simulation of the immune system. *PLoS One*. **5**, e9862 (2010).
61. Grote, A. et al. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.* **33**, W526–531 (2005).
62. Rey, J., Murail, S., de Vries, S., Derreumaux, P. & Tuffery, P. PEP-FOLD4: a pH-dependent force field for peptide structure prediction in aqueous solution. *Nucleic Acids Res.* **51**, W432–W437 (2023).
63. Kuriata, A. et al. Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility. *Nucleic Acids Res.* **47**, W300–W307 (2019).

Figures

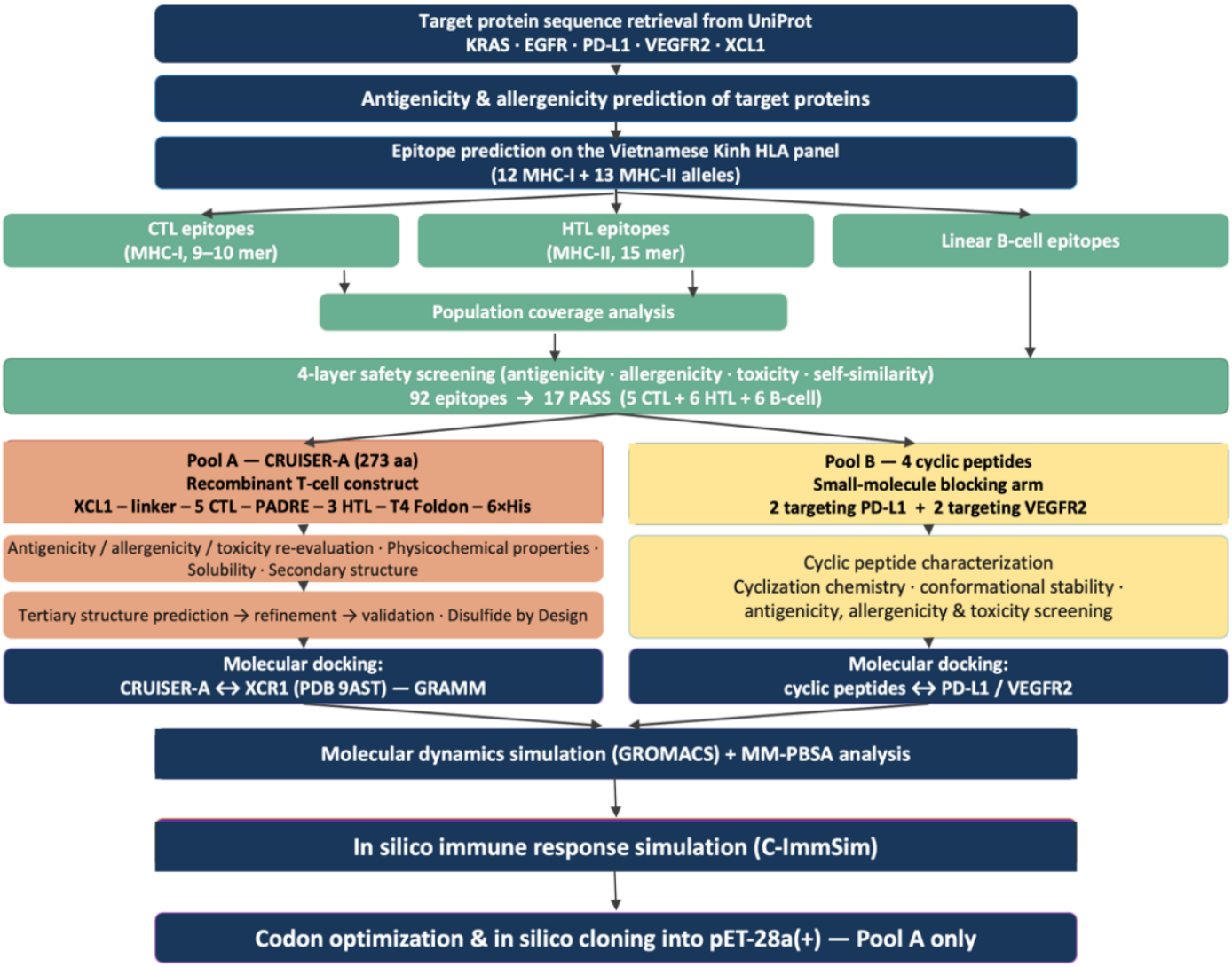


Figure 1

Immunoinformatics workflow used to design the CRUISER dual-arm vaccine system. Target protein sequences were retrieved and screened for antigenicity and allergenicity, and epitopes were predicted exclusively against the Vietnamese Kinh HLA panel of 12 MHC class I alleles and 13 MHC class II specificities. Cytotoxic T-lymphocyte, helper T-lymphocyte and linear B-cell epitopes were carried through a four-layer safety cascade, from which 17 of 92 candidates were retained. The two arms then diverge: Pool A is assembled into the 273-residue recombinant construct CRUISER-A and taken through physicochemical characterisation, tertiary structure prediction, refinement, validation and disulfide design before docking to XCR1; Pool B comprises four disulfide-cyclised peptides characterised separately and docked to PD-L1 and VEGFR2. Both arms converge on molecular dynamics simulation with MM-PBSA interaction energy analysis, followed by in silico immune simulation and, for Pool A only, codon optimisation and cloning into pET-28a(+).

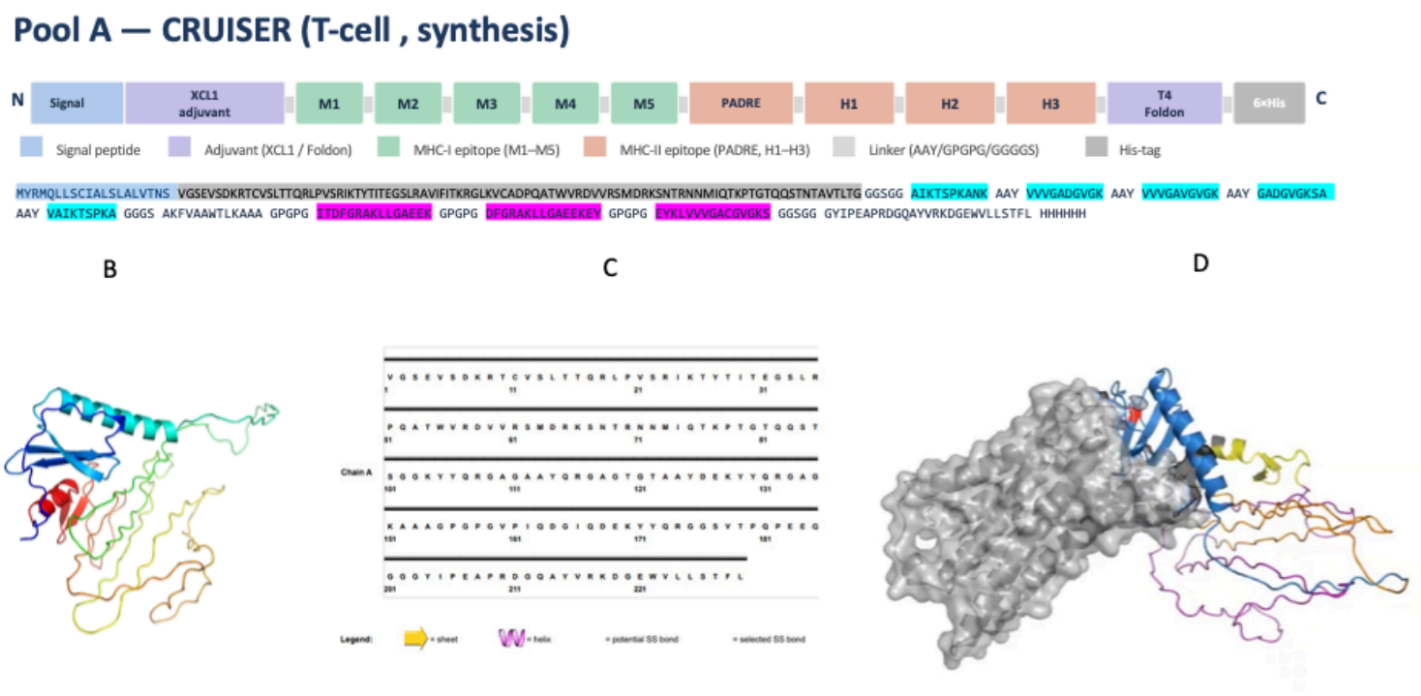


Figure 2

Architecture and structure of CRUISER-A. (a) Sequential module diagram and complete amino acid sequence, comprising the signal peptide, the XCL1 adjuvant module, five MHC class I epitopes (M1–M5) joined by AAY linkers, the universal helper epitope PADRE, three MHC class II epitopes (H1–H3) joined by GPGPG linkers, the T4 fibritin trimerisation domain and the hexahistidine tag. (b) Refined three-dimensional model, coloured along the sequence from the amino terminus (blue) to the carboxy terminus (red). (c) Secondary structure and disulfide bond diagram generated by PDBsum. (d) Docked complex between CRUISER-A and the XCR1 receptor (grey surface), showing the XCL1 core packed against the extracellular face of the receptor while the epitope cassette projects into solvent.

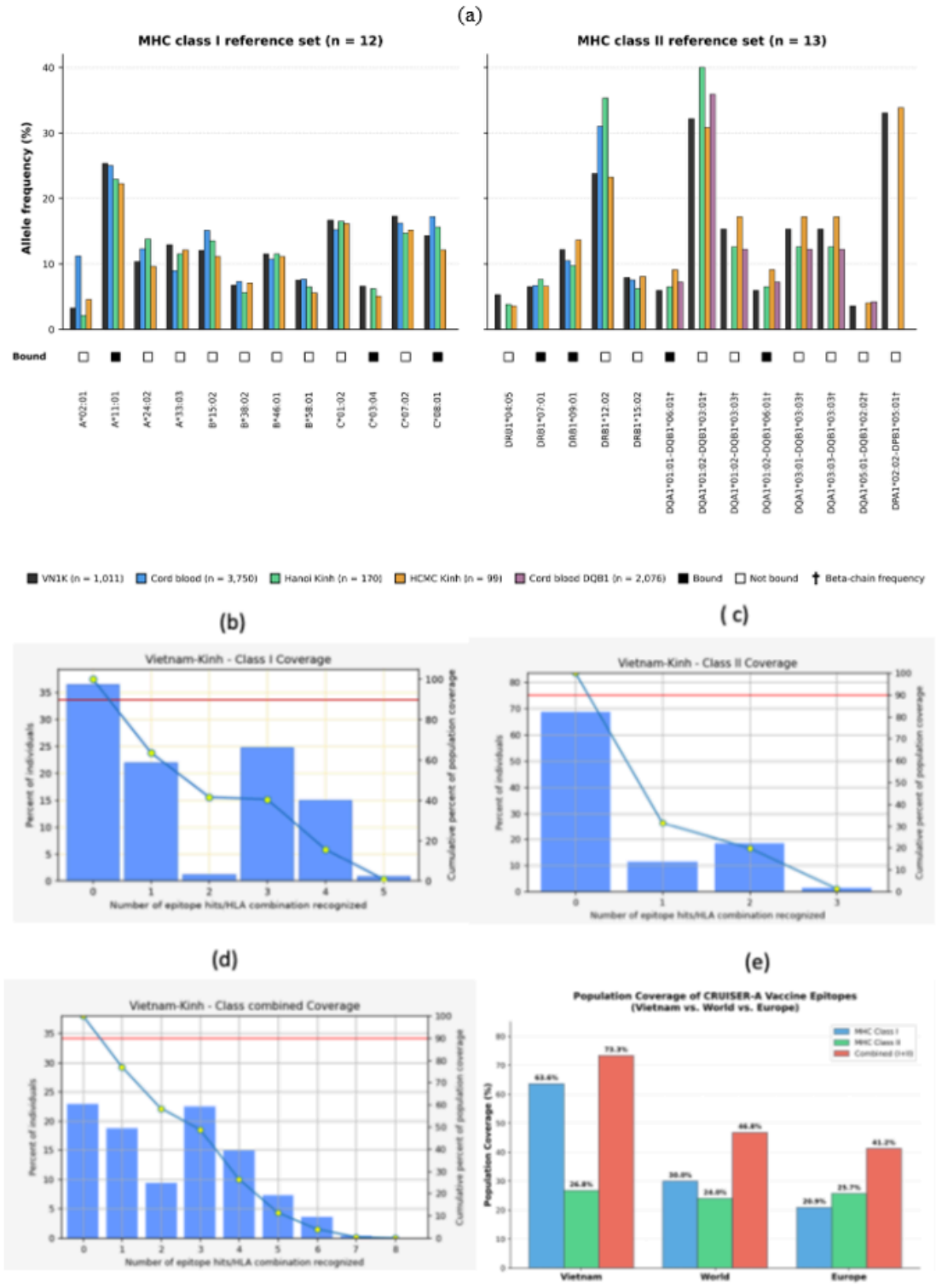


Figure 3

Vietnamese HLA reference allele set and population coverage of the CRUISER-A T-cell epitope set. (a) All 25 specificities of the reference set — 12 MHC class I alleles (left) and 13 MHC class II specificities (right) — plotted by allele frequency in five independent Vietnamese cohorts: VN1K whole-genome sequencing (n = 1,011), the cord blood series for HLA-A, -B, -C and -DRB1 (n = 3,750) and for HLA-DQB1 (n = 2,076), the Hanoi Kinh series (n = 170) and 99 Kinh individuals from Ho Chi Minh City, the last compiled in ref. 7.

*Missing bars indicate loci not reported by that cohort. The dagger marks heterodimeric specificities for which the displayed value is the frequency of the beta chain (DQB1 or DPB1), since no cohort reports DQA1 or DPA1 frequencies; specificities sharing a beta chain therefore carry identical displayed values and true heterodimer frequencies are equal to or lower than these. The Bound annotation strip below the abscissa marks specificities recorded in Supplementary Table S7 as the best-ranked binder for at least one CRUISER-A epitope (filled square: yes; open square: no); seven are marked, namely A*11:01, C*08:01, C*03:04, DRB1*07:01, DRB1*09:01, DQA1*01:01–DQB1*06:01 and DQA1*01:02–DQB1*06:01. This reflects only the top-ranked specificity for each epitope, not the full set bound: the two DQ-restricted epitopes additionally bound six and three further specificities (Supplementary Tables S4, S7). (b) Distribution of MHC class I coverage in the Vietnam population. (c) Distribution of MHC class II coverage in the Vietnam population. (d) Distribution of combined class I and class II coverage in the Vietnam population. In panels (c) and (d), bars indicate the proportion of individuals recognising the stated number of epitope–HLA combinations, the line indicates cumulative coverage on the right-hand axis, and the horizontal line marks the 90% level. (e) Comparison of coverage across three reference populations (Vietnam, World, Europe) for the same epitope–allele input, shown separately for MHC class I, MHC class II and the combined set. Complete data are given in Supplementary Table S9.*

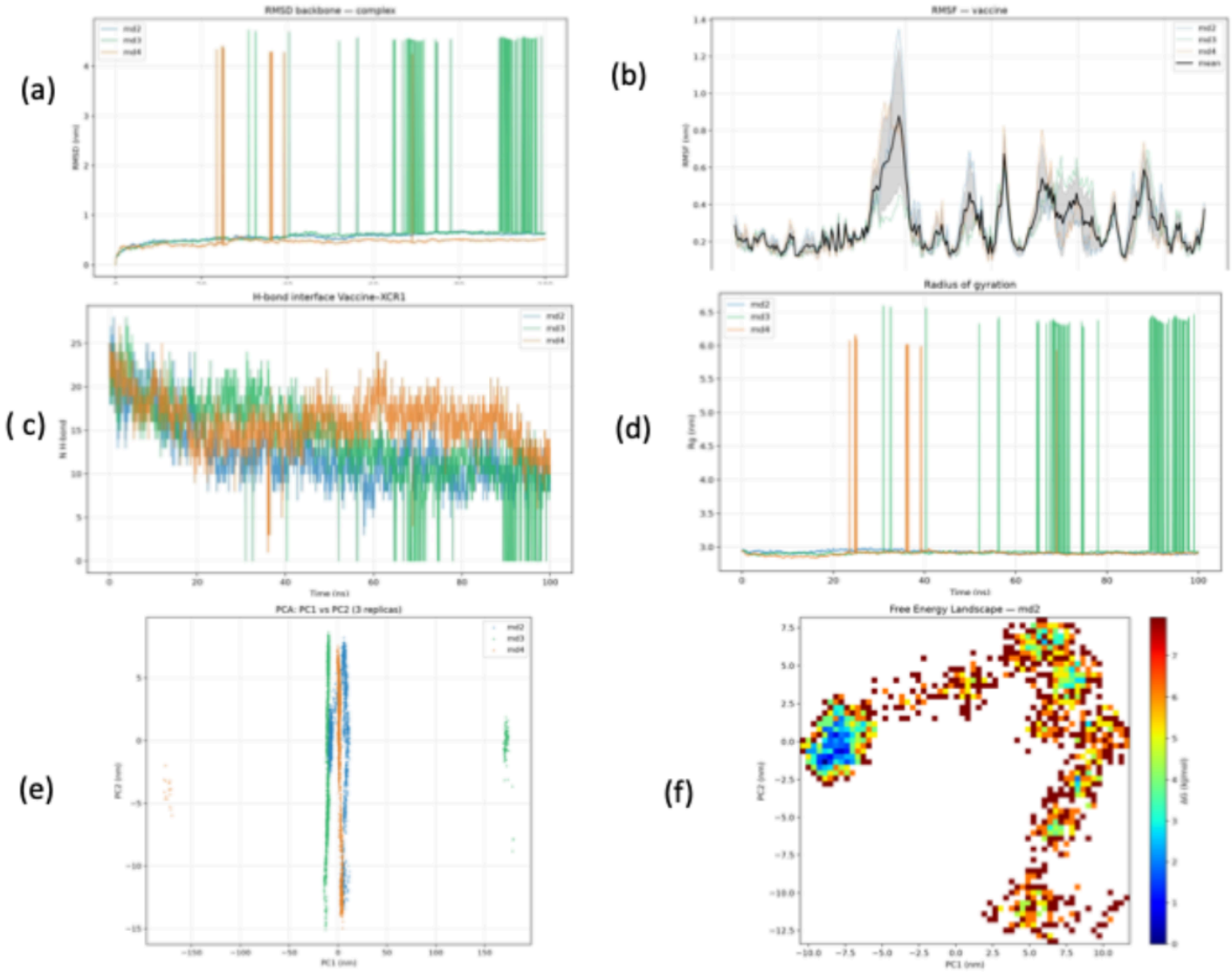


Figure 4

Dynamics and interaction energetics of the CRUISER-A:XCR1 complex across three independent simulations. (a) Root-mean-square deviation of the whole complex. (b) Per-residue fluctuation of the vaccine component. (c) Number of hydrogen bonds at the interface. (d) Radius of gyration. (e) Projection of the trajectories onto the first two principal components. (f) Free energy landscapes. Solvent-accessible surface areas and the remaining energy landscapes are given in Supplementary Fig. S2.

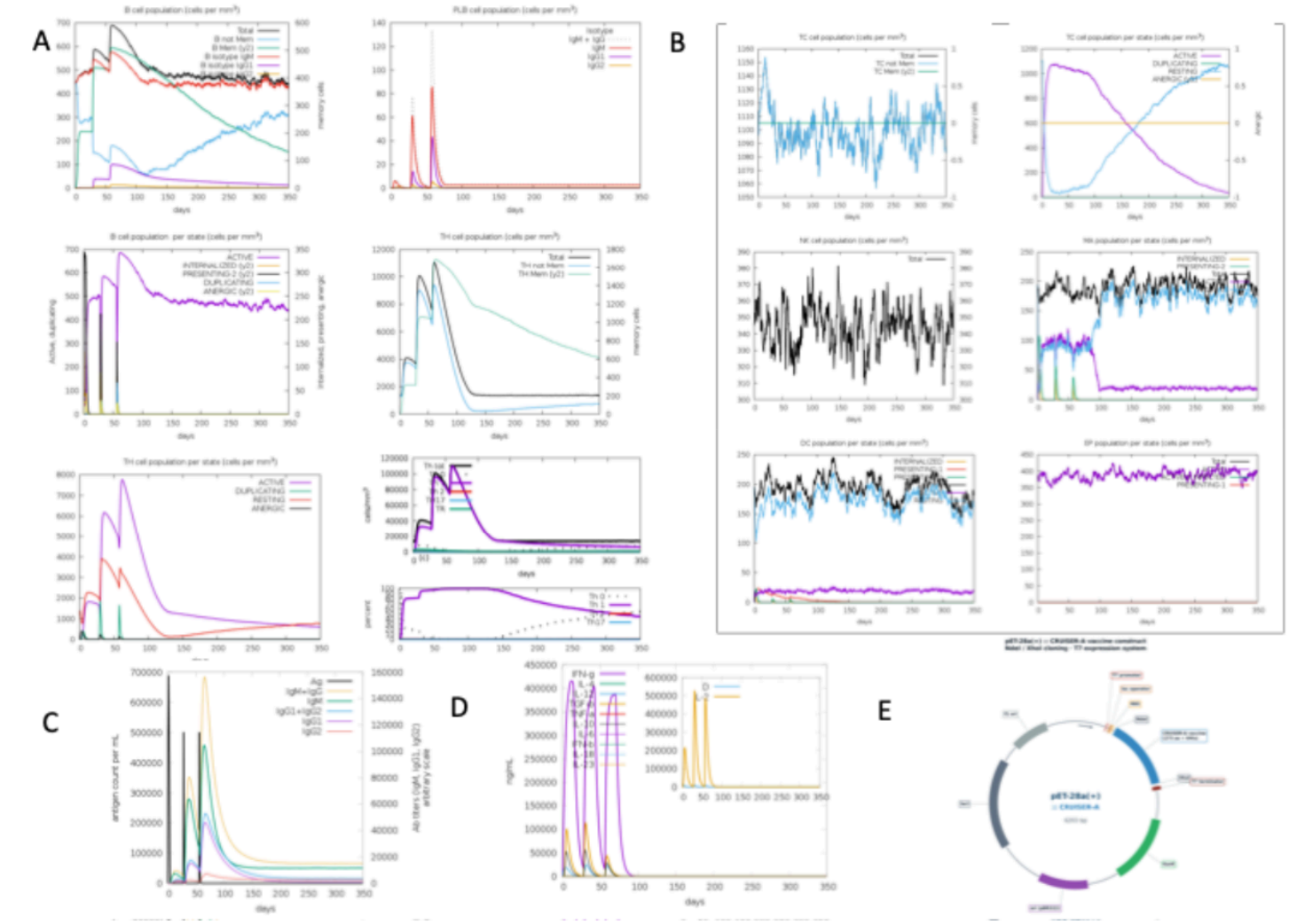


Figure 5

Immune simulation and in silico expression. (a–d) Antigen and immunoglobulin kinetics, B-cell populations, helper and cytotoxic T-cell populations, and cytokine profiles across three doses at time steps 1, 84 and 168. (e) Map of the pET-28a(+)::CRUISER-A plasmid showing the NdeI and XhoI cloning sites, the T7 promoter and the kanamycin resistance gene.

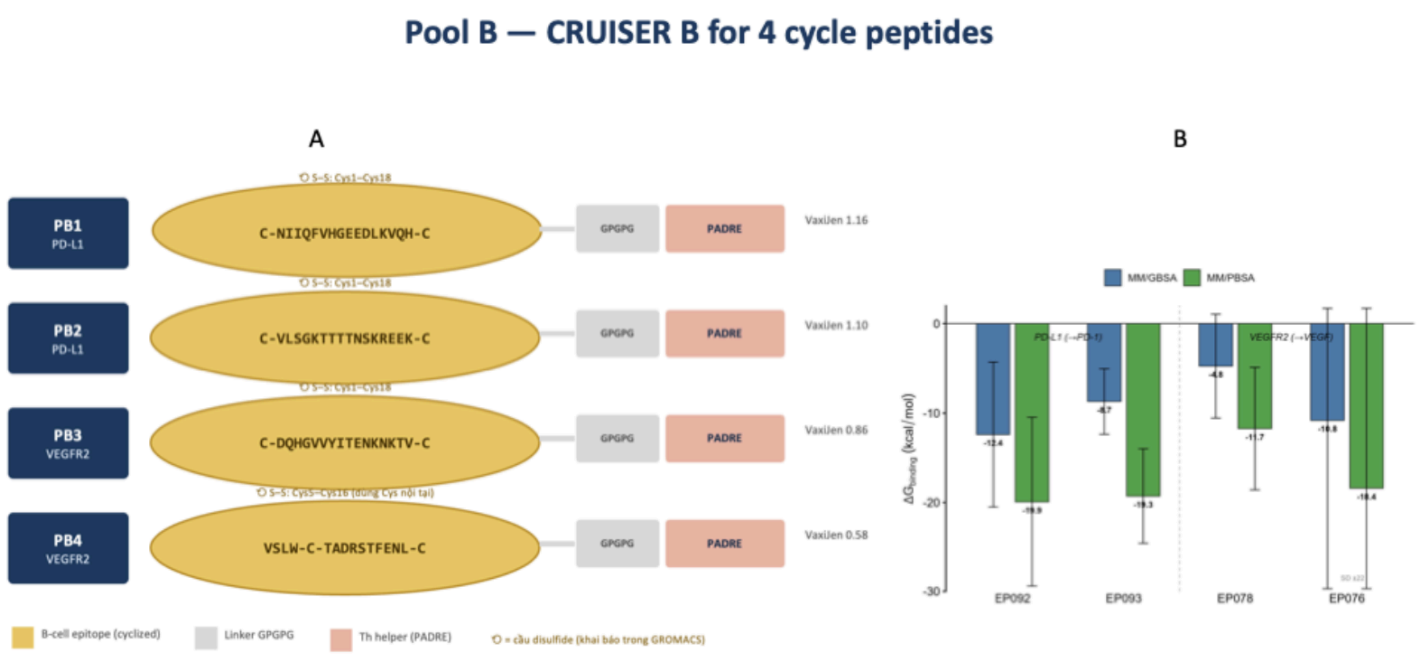


Figure 6

The Pool B antibody arm. (a) Structures of the four disulfide-cyclised peptides targeting PD-L1 and VEGFR2. (b) MM-PBSA interaction energies of the four peptide–receptor complexes under both solvation models; error bars denote standard deviations across trajectory frames. Values exclude the configurational entropy term and are intended for relative comparison between complexes.