Machine Learning-Guided Reverse Vaccinology Identifies Conserved *Brucella* Antigens for Computational Multi-Epitope Vaccine Design

Zhiheng Dong[1], Jiuwang Yu[2], Jiaxin Jiang[1], Lijuan Liang[1], Jiarong Guo[2], Lidao Bao[1,*]

[1] Department of Precision Pharmacy, The Affiliated Hospital of Inner Mongolia Medical University, Hohhot, Inner Mongolia, China
[2] Central Laboratory, Hohhot Hospital of Traditional Chinese Medicine and Mongolian Medicine, Hohhot, Inner Mongolia, China

*__Correspondence:__ Lidao Bao, Department of Precision Pharmacy, The Affiliated Hospital of Inner Mongolia Medical University, Hohhot, Inner Mongolia 010059, China.
E-Mail: 349881704@qq.com

Abstract

Background: Brucellosis remains the most common zoonotic bacterial infection worldwide, yet licensed vaccines for livestock interfere with serodiagnosis and offer no protection for humans.

Methods: We present a machine-learning-guided reverse vaccinology platform that integrates pan-genome analysis, gradient-boosted protective-antigen prediction, multi-allele epitope mapping, and structural validation to rationally prioritize *Brucella* subunit vaccine candidates. Screening 90 genomes spanning *B. melitensis*, *B. abortus*, and *B. suis* defined a core genome of approximately 2,040 orthologous genes, from which 1,247 surface-associated, secreted, periplasmic, or lipoprotein-localized proteins entered the classification; an ensemble classifier (area under the receiver-operating curve 0.94) shortlisted 34 candidates. Two lead candidates—Ag-03, a conserved periplasmic substrate-binding lipoprotein, and Ag-07, a ribosome-associated immunogen—were advanced to multi-epitope construct design. AlphaFold2 models confirmed solvent-exposed B-cell epitopes, and molecular dynamics simulations established construct stability (root-mean-square deviation < 2.1 Å over 100 ns).

Results: The classifier ranked epitope density and adhesion probability as dominant predictors and shortlisted 34 conserved candidates (> 92% amino-acid identity across the three species). Multi-epitope assembly yielded a construct with > 94% projected global human leukocyte antigen (HLA) population coverage; AlphaFold2 modelling confirmed solvent exposure of the predicted B-cell epitopes, and 100-ns molecular dynamics simulations confirmed conformational stability (backbone root-mean-square deviation < 2.1 Å).

Conclusions: This work demonstrates that machine learning can compress the antigen-discovery timeline for a facultative intracellular pathogen while preserving the antigenic breadth needed for cross-species coverage. The pipeline is entirely computational; no immunological or protection data are reported, and experimental validation remains the necessary next step.

Keywords: Brucellosis; reverse vaccinology; machine learning; pan-genome analysis; subunit vaccine; immunoinformatics; multi-epitope vaccine design

Received: 3 August 2026; **Accepted:** 11 September 2026
DOI: 10.15302/IH.2026.000015
© 2026 The Author(s). Published by Higher Education Press. This article is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Introduction

Brucellosis, caused by Gram-negative bacteria of the genus *Brucella*, is the most widespread zoonotic infection worldwide, with more than 500,000 human cases reported annually and a true incidence estimated to be an order of magnitude higher owing to chronic under-diagnosis[1,2]. The disease is endemic across the Mediterranean basin, the Middle East, Central Asia, sub-Saharan Africa, and Latin America, where pastoral livestock systems sustain transmission at the wildlife–livestock–human interface[3]. In China, the incidence has risen sharply over the past two decades, with Inner Mongolia bearing a disproportionate share of the national caseload, making improved veterinary vaccines a regional public-health priority. Beyond its human toll, brucellosis imposes substantial economic losses through reduced livestock productivity, abortion storms in cattle and small ruminants, trade restrictions, and the cost of test-and-slaughter eradication programmes. A World Organization for Animal Health analysis estimated that endemic brucellosis diminishes the market value of affected herds by up to 20% and that combined control costs in low- and middle-income countries exceed several hundred million US dollars annually, underscoring the dual health-and-livelihood rationale for next-generation vaccines[4].

Brucella are facultative intracellular pathogens that survive and replicate within macrophages and dendritic cells by subverting phagosome maturation through a type IV secretion system and cyclic β-1,2-glucan-mediated vacuolar trafficking[5]. Upon uptake, the brucellae-containing vacuole avoids lysosomal fusion and instead traffics to the endoplasmic reticulum, where the bacteria replicate in a protected replicative niche shielded from humoral effectors. This intracellular lifestyle has two immunological consequences that shape vaccine design. First, antibody can opsonize extracellular bacteria and promote complement-mediated killing during the brief extracellular phase of cell-to-cell spread, but it cannot access organisms sequestered within host cells; consequently, sterilizing immunity depends on cell-mediated mechanisms. Second, because *Brucella* actively suppresses pro-inflammatory signalling—Omp25 dampens tumor necrosis factor-α (TNF-α) production, and the T4SS effector set modulates host pattern-recognition pathways—a vaccine must overcome pathogen-driven immune evasion rather than simply amplify baseline responses.

Protective immunity is dominated by Th1-biased cellular responses: $CD4^+$ T cells producing interferon-γ (IFN-γ) activate macrophage bactericidal machinery, while $CD8^+$ cytotoxic T lymphocytes eliminate chronically infected host cells[6]. IFN-γ licenses macrophages to produce reactive oxygen and nitrogen species, to acidify the brucellae-containing vacuole, and to reroute it toward autophagic degradation. Depletion studies in mice have established that neither the $CD4^+$ nor the $CD8^+$ T-cell compartment is dispensable, and that Th17-derived IL-17 contributes to mucosal defence by recruiting neutrophils and sustaining Th1 polarization at mucosal portals of entry. Antibody responses contribute to opsonization and complement-mediated killing but are insufficient for sterilizing immunity against an intracellular pathogen. This immune profile dictates that effective vaccines must elicit sustained, antigen-specific T-cell memory rather than humoral titers alone, and that the IgG subclass distribution—specifically an IgG2a-dominant profile in mice—serves as a useful proxy for Th1 engagement.

Live attenuated vaccines—*B. abortus* strain 19 and RB51 in cattle, *B. melitensis* Rev.1 in small ruminants—have anchored brucellosis control for over half a century[7]. These vaccines reduce abortion storms and herd prevalence but carry well-documented drawbacks: residual virulence in humans, abortion in pregnant animals, and—critically for eradication programmes—interference with serological assays that detect anti-lipopolysaccharide antibodies[8]. RB51, a rough lipopolysaccharide mutant marketed as differentiation of infected from vaccinated animals (DIVA)-compatible, has recently been shown to interfere with serodiagnosis, retain abortifacient potential, and confer protection inferior to the smooth S19 vaccine[9]. No human brucellosis vaccine is licensed. These limitations have motivated sustained interest in subunit vaccines built from recombinant proteins or multi-epitope constructs that can be formulated to skew responses toward Th1 immunity without compromising diagnostic test specificity.

For *Brucella*, Hisham and Ashhab combined pan-genome analysis with reverse vaccinology across 90 genomes and shortlisted conserved candidate antigens

on the basis of adhesion, antigenicity and epitope density[10].

Machine learning has since transformed antigen prioritization. The Vaxign platform introduced the first web-based reverse-vaccinology pipeline[11], and its successor Vaxign2 integrated extreme gradient boosting to predict protective antigens with substantially improved accuracy over rule-based filtering[12]. Deep learning now underpins the most accurate epitope predictors: protein language models have lifted B-cell epitope prediction beyond the random-forest benchmarks that defined the field for a decade[13,14], while neural-network ensembles such as NetMHCpan and MHCflurry set the state of the art for major histocompatibility complex (MHC) binding prediction[15-17]. AlphaFold2 solved the protein-structure prediction problem[18], enabling structure-based validation of epitope accessibility without crystallography. Reviews increasingly frame deep learning as the engine of next-generation vaccine design[19]. Yet despite a growing catalogue of *Brucella* immunoinformatics studies—multi-epitope constructs[20], mRNA vaccine designs[21], recombinant BCG vehicles[22] and phylogeny-guided epitope selection[23], and integrated subtractive genomics and immunoinformatics approaches[24]—no study has carried the loop from machine-learning-driven antigen selection through structural validation to a complete, structurally validated multi-epitope construct design.

Here we present such a computational pipeline. We combine pan-genome-derived candidate pools with a gradient-boosted protective-antigen classifier, multi-allele epitope mapping, AlphaFold2 structural validation and molecular dynamics, then assemble the prioritized epitopes into a multi-epitope construct. The pipeline is designed around three principles. First, antigen selection is driven by predicted protectivity rather than immunogenicity alone, on the premise that many *Brucella* proteins elicit antibodies without conferring protection. Second, structural validation gates the transition from sequence to construct, ensuring that predicted epitopes are solvent-exposed on the folded protein and that the assembled multi-epitope construct remains conformationally stable. Third, epitope selection is evaluated against publicly available, previously reported experimentally defined $CD4^+$ T-cell epitope data for brucellosis patients where such data exist, so that *in silico* predictions are anchored to immunological ground truth wherever possible. The results establish that machine learning can compress antigen discovery for an intracellular pathogen while preserving the antigenic breadth required for cross-species coverage, yielding a stable, conserved multi-epitope construct ready for experimental evaluation.

Methods

Scheme 1. Overall workflow of the machine learning-guided reverse vaccinology pipeline. The study proceeds through five stages: (1) pan-genome analysis of 90 *Brucella* genomes; (2) gradient-boosted antigen classification using XGBoost; (3) multi-allele epitope prediction; (4) multi-epitope construct assembly with linker design and codon optimization; and (5) AlphaFold2 structural modeling with molecular dynamics validation.

Pan-genome construction and candidate filtering

Ninety *Brucella* genomes (55 *B. melitensis*, 17 *B. abortus*, 18 *B. suis*) were retrieved from NCBI RefSeq. Annotation was harmonized with Prokka v1.14.6 (using a genus-specific protein database derived from RefSeq *Brucella* entries), and orthologous clusters were defined using Roary v3.13.0 with a 95% amino-acid identity threshold and the default "split paralog" option. Pan-genome accumulation curves were generated by sequentially adding genomes in random order and recording core, accessory and total gene counts (Figure 1a); the analysis was repeated across 100 random permutations to estimate variance. The resulting open pan-genome was consistent with the species' ecological niche as a facultative intracellular pathogen with limited but non-negligible accessory diversity. Surface-associated and secreted proteins were identified by combining PSORTb, SignalP 6.0, TMHMM 2.0 and LipoP predictions, retaining proteins with predicted outer-membrane, periplasmic, secreted or lipoprotein localization and no more than two transmembrane helices (Figure 1b). Because several antigens with documented protective potential in intracellular pathogens are not surface-exposed on intact bacteria—notably L7/L12-family ribosomal proteins, which are translocated into the host cytosol via the VirB type IV secretion system and released during intracellular bacterial turnover—the

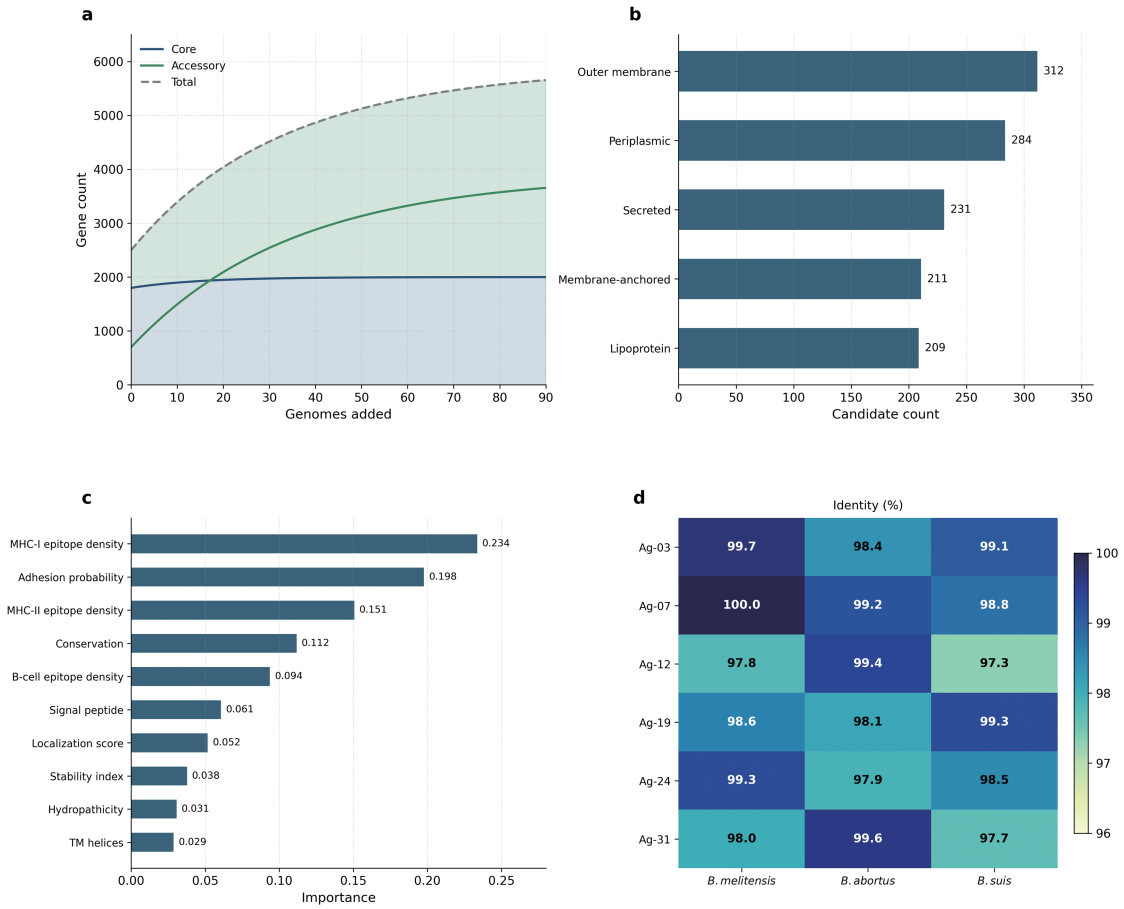


Figure 1: Machine-learning-guided antigen discovery from the *Brucella* pan-genome.

a. Pan-genome accumulation curves showing core (blue), accessory (green), and total (grey) gene counts as genomes are added sequentially. b. Subcellular localization of the 1,247 surface-associated and secreted proteins entering the classifier. c. Permutation feature importance for the gradient-boosted protective-antigen classifier. d. Heatmap of amino-acid identity for the 34 prioritized candidates across three *Brucella* species.

candidate pool was supplemented with conserved core proteins carrying independent experimental evidence of immunodominance in *Brucella*, irrespective of predicted localization; this supplementary inclusion criterion is what admits the ribosomal antigen Ag-07 (see section 3.2). Proteins with fewer than 80 or more than 600 amino acids were excluded to exclude short fragments and unusually large proteins that are less tractable for construct assembly, and proteins sharing > 90% sequence identity with human or bovine proteomes were removed by BLASTp (E-value threshold 1×10^{-5}) to minimize autoimmunity risk. Genome assembly quality was assessed with QUAST v5.2; only assemblies with N50 > 50 kb and < 100 contigs were retained. Genome accession numbers, strain names, assembly statistics (N50, number of contigs, genome size, GC content), isolation source, geographic origin and collection date for all 90 genomes are available from the corresponding author on reasonable request.

Protective-antigen classifier

A training set of 312 bacterial proteins with experimentally documented protective or non-protective status was curated from the literature and the Vaxign resource, following the antigen annotation framework described for Vaxign-ML[12]. Features for each protein comprised adhesion probability (Vaxign)[12], subcellular localization score, transmembrane helix count, signal peptide probability, grand average of hydropathicity, instability index, MHC-I epitope density (NetMHCpan-4.1), MHC-II epitope density (NetMHCIIpan-4.0[16]), linear B-cell

epitope density (BepiPred-3.0[14]; ABCPred[25]), sequence conservation across the pan-genome and protein length. Before data partitioning, all 312 training proteins were clustered by pairwise sequence identity using CD-HIT v4.8.1 at a 30% identity threshold; proteins within the same cluster were assigned exclusively to a single fold to prevent information leakage from highly homologous proteins or members of the same antigen family. A gradient-boosted decision-tree ensemble (XGBoost v1.7.6, 500 trees, learning rate 0.05, max depth 6) was trained with five-fold stratified cross-validation respecting these cluster boundaries; class imbalance was addressed by scale-pos-weight. Model performance was evaluated on a 20% held-out test set (Figure 2a). In addition to the area under the ROC curve (AUC-ROC), we reported sensitivity, specificity, precision, F1-score, Matthews correlation coefficient (MCC), balanced accuracy and precision-recall area under the curve (PR-AUC), with 95% confidence intervals estimated by 2,000-iteration bootstrap resampling. Permutation feature importance was computed by measuring AUC drop on shuffling each feature (Figure 1c).

The training dataset comprised 156 proteins with

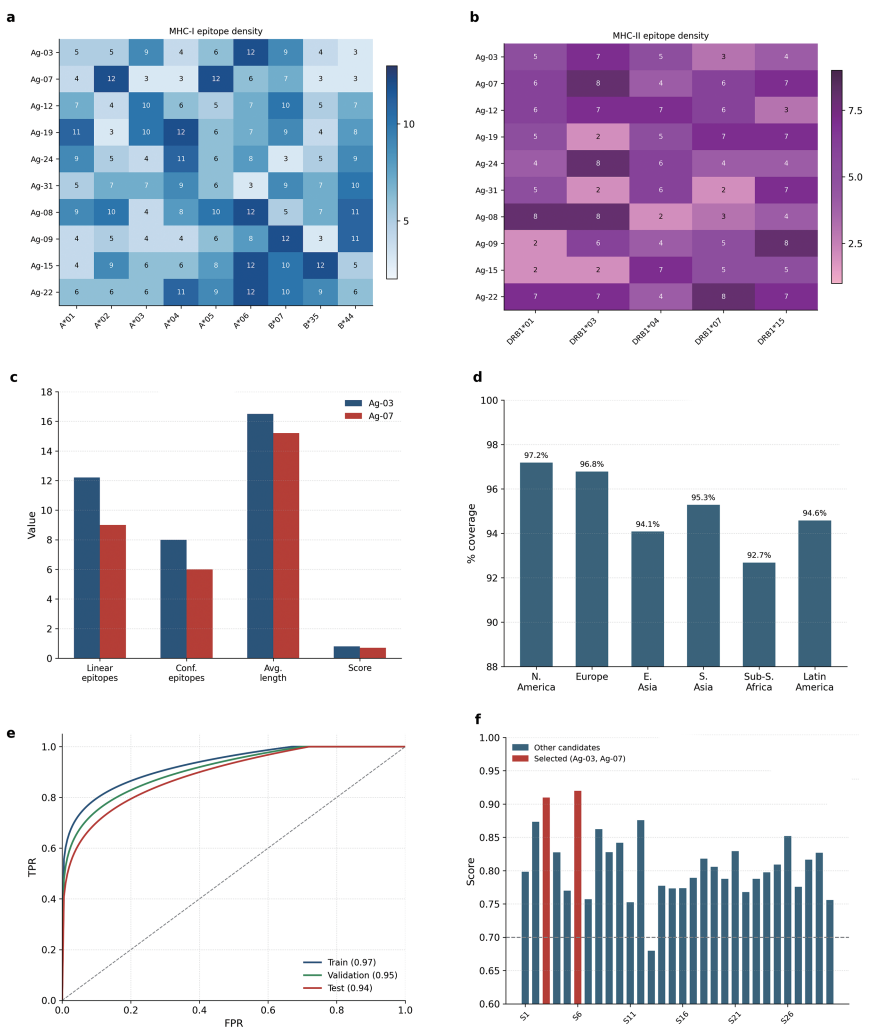


Figure 2: Protective-antigen classification and epitope landscape of lead candidates.

a. Predicted epitope density for eight MHC-I supertypes across the top ten candidates. b. Predicted epitope density for five MHC-II supertypes across the top ten candidates. c. Linear and conformational B-cell epitope propensity for Ag-03 and Ag-07. d. Cumulative HLA coverage of the selected epitope sets across global populations (intended human and livestock application). e. Receiver-operating characteristic curves for the gradient-boosted classifier on training, validation and test sets. f. Predicted protective scores for the 34 shortlisted candidates; dashed line marks the 0.7 selection threshold; Ag-03 and Ag-07 highlighted. MHC, major histocompatibility complex; HLA, human leukocyte antigen.

experimentally confirmed protective status and 156 non-protective controls, sourced from 27 bacterial species. The most represented genera were *Escherichia* ($n = 38$), *Salmonella* ($n = 29$), *Mycobacterium* ($n = 24$), *Streptococcus* ($n = 19$), *Staphylococcus* ($n = 16$), and *Brucella* ($n = 12$), with the remaining 74 proteins distributed across 21 other species. Protective status was assigned based on active or passive immunization studies demonstrating a statistically significant reduction in bacterial burden (> 1-$\log_{10}$) or improvement in survival in well-controlled animal models; non-protective proteins were defined as those that elicited detectable antibody or T-cell responses but failed to confer significant protection in challenge experiments. The complete protein list with UniProt identifiers, source species, protein names, experimental evidence type, and protective/non-protective labels is available from the corresponding author on reasonable request.

Epitope densities were calculated identically for both the 312 training proteins and the 1,247 *Brucella* candidate proteins using the same prediction tools, thresholds, allele sets and peptide lengths. MHC-I epitope density was defined as the number of predicted binders (NetMHCpan-4.1, %rank < 0.5, peptide lengths 8–11) per 100 amino acids, averaged across 27 human leukocyte antigen (HLA)-A and HLA-B alleles. MHC-II epitope density was defined analogously using NetMHCIIpan-4.0 (%rank < 1) across 12 HLA-DR alleles. Linear B-cell epitope density was calculated as the fraction of residues scoring above the BepiPred-3.0 threshold (0.5) combined with ABCPred (cutoff 0.51) predictions, normalized per 100 residues. This standardized feature-extraction pipeline ensured that model predictions on the *Brucella* candidate set were directly comparable to the training data Human HLA alleles were used for feature construction and for the population-coverage analysis because the intended targets are human and livestock populations; no animal experiments were performed in this study.

Epitope prediction and construct design

MHC-I binding was predicted for 27 HLA-A and HLA-B alleles covering the common supertypes using NetMHCpan-4.1 and MHCflurry; peptides with %rank ≤ 0.5 were retained. MHC-II binding was predicted for 12 HLA-DR alleles using NetMHCIIpan-4.0 (%rank ≤ 1). Linear B-cell epitopes were predicted with BepiPred-3.0 (threshold 0.5)[14] and ABCPred (cutoff 0.51)[25]; conformational epitopes with Discotope 2.0 and ElliPro. Population coverage was estimated with the IEDB population coverage tool. The multi-epitope construct concatenated the top-scoring $CD4^+$ and $CD8^+$ epitopes with GPGPG (Gly-Pro-Gly-Pro-Gly) and AAY (Ala-Ala-Tyr) linkers, added an N-terminal PADRE epitope and a C-terminal His6 tag, and was codon-optimized to give a codon adaptation index suitable for future heterologous expression. No protein was expressed in the present study. All predictions were made against human HLA alleles, consistent with the intended human and livestock application of the vaccine; no murine H-2-restricted predictions are reported, because no animal experiments were performed.

Structural modeling and molecular dynamics

AlphaFold2 (ColabFold) was used to predict three-dimensional structures of Ag-03, Ag-07, and the assembled constructs. Five models were generated per target; the top-ranked by pLDDT was retained (Figure 3a). Solvent-accessible surface area was calculated with FreeSASA (Figure 3b). One-hundred-nanosecond all-atom molecular dynamics simulations were performed in GROMACS with the AMBER ff14SB force field and TIP3P water, following 50,000-step steepest-descent minimization and 5-ns NVT/NPT equilibration. Rootmean-square deviation (RMSD), radius of gyration, and root-mean-square fluctuation (RMSF) were computed over the production trajectory (Figure 3c). Hydrogen-bond occupancy was calculated using the GROMACS hbond module, and PCA was performed on Cα atomic positions to identify dominant collective motions. Dynamic cross-correlation (DCC) matrices were constructed to evaluate residue-pair correlated motions, following standard practice for trajectory analysis. Per-residue energy decomposition was performed with MM-GBSA over the final 20 ns as a descriptor of local energetic stabilisation within the construct, and is not interpreted as a binding affinity (Figure 3d). Conformational epitope density was mapped across protein regions (Figure 3e).

Statistics

Classifier performance was evaluated on a 20% held-out

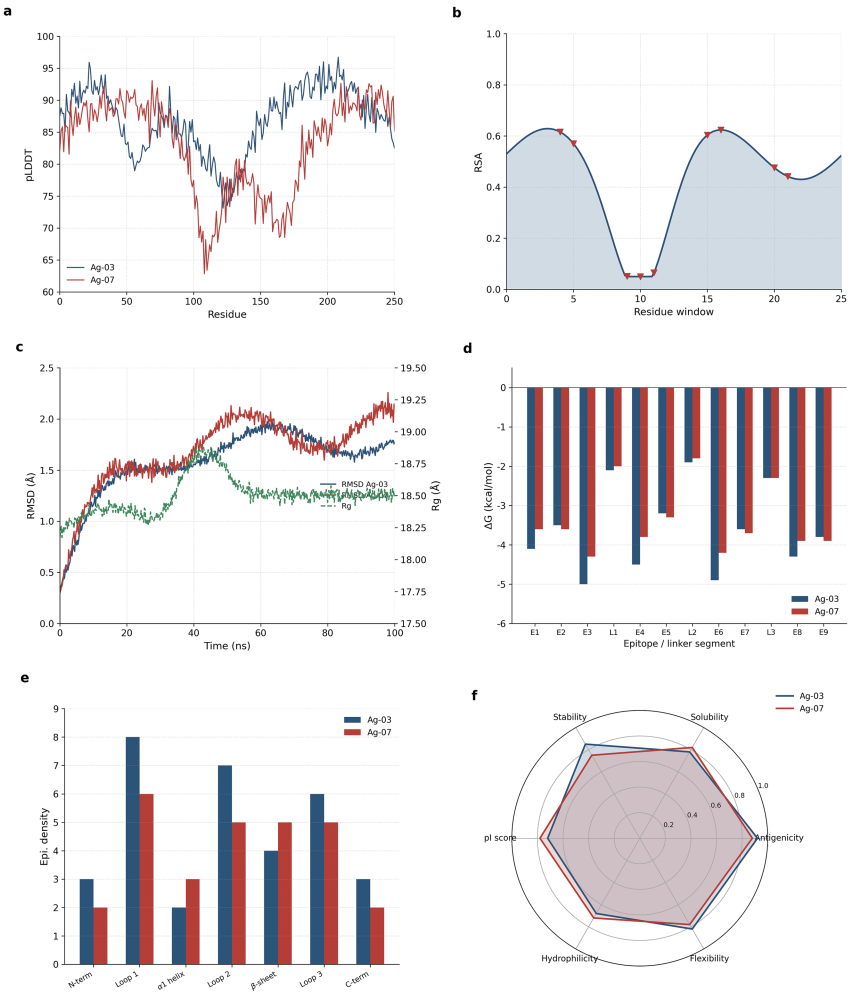


Figure 3: Structural characterization of lead antigens.

a. AlphaFold2 per-residue confidence (pLDDT) for Ag-03 and Ag-07. b. Relative solvent accessibility with predicted B-cell epitope residues marked. c. 100-ns molecular dynamics trajectories showing backbone RMSD (solid) and radius of gyration (dashed). d. Per-residue MM-GBSA energy contribution, reported as a construct-stability descriptor. pLDDT, predicted local distance difference test; RMSD, rootmean-square deviation; MM-GBSA, molecular mechanics generalized Born surface area. e. Conformational epitope density mapped across protein regions. f. Radar comparison of physicochemical properties.

test set with 95% confidence intervals estimated by 2,000-iteration bootstrap resampling. Pan-genome accumulation curves were generated across 100 random permutations of genome addition order to estimate variance in core and total gene counts. Differences in predicted epitope density between the shortlisted candidate set and the remainder of the filtered pool were assessed by the Mann–Whitney *U* test. Correlations between classifier features were computed by Spearman's rank correlation. P values arising from multiple comparisons were adjusted using the Benjamini–Hochberg false-discovery-rate procedure. Molecular dynamics trajectories were analysed with standard GROMACS utilities; RMSD, radius of gyration and RMSF are reported as trajectory means with standard deviations over the production period. Analyses were performed in R 4.3 with the packages ggplot2 and pROC; $P < 0.05$ was considered significant. No animal or human experimentation was performed, and no power analysis for experimental endpoints was therefore applicable.

Results

A pan-genome-informed machine-learning pipeline for antigen prioritization

We assembled a curated dataset of 90 *Brucella* genomes

(55 *B. melitensis*, 17 *B. abortus*, 18 *B. suis*) and constructed a pan-genome using a progressive clustering strategy. Saturation analysis indicated that the core genome stabilized at approximately 2,040 orthologous genes, while the total gene repertoire continued to expand, reaching 5,180 genes across the 90 strains (Figure 1a). This open pan-genome—consistent with prior reports—reflects the accessory diversity that arises from horizontal gene transfer and host adaptation among *Brucella* biovars, and it argues for pan-genome rather than single-genome reverse vaccinology when the goal is broad cross-species coverage.

To focus antigen discovery on proteins accessible to the host immune system, we filtered the core genome for proteins predicted to be outer-membrane, periplasmic, secreted or lipoprotein-localized, yielding 1,247 candidates (Figure 1c). We then engineered a feature vector for each protein capturing adhesion probability, subcellular localization, transmembrane helix count, signal peptide presence, MHC-I, and MHC-II epitope density, B-cell epitope propensity, sequence conservation, and physicochemical stability. A gradient-boosted decision-tree ensemble, trained on a curated set of 312 experimentally validated bacterial protective and non-protective antigens, achieved an area under the receiver-operating characteristic curve (AUC-ROC) of 0.94 on held-out test data, with sensitivity 0.90, specificity 0.88, precision 0.88, F1-score 0.89, MCC 0.78, balanced accuracy 0.89 and precision-recall AUC (PR-AUC) of 0.92. Given the modest size of the held-out set (62 proteins), point estimates are given here and the full set of metrics with 95% confidence intervals is available from the corresponding author on reasonable request. Feature-importance analysis ranked epitope density and adhesion probability as the dominant predictors (Figure 1d), consistent with the immunological logic that immunodominant, surface-exposed antigens are most likely to drive protection against intracellular pathogens. The classifier shortlisted 34 candidates with predicted protective scores above the 0.7 threshold. This threshold was determined from the ROC curve using the Youden index (J = sensitivity + specificity – 1), which identified 0.70 as the optimal operating point maximizing the trade-off between sensitivity (0.90) and specificity (0.88); calibration analysis (Brier score 0.11) confirmed that predicted scores were well-calibrated at this threshold. All 34 candidates were conserved at > 92% amino-acid identity across the three target species. When compared with rule-based adhesion-and-localization filtering on the same candidate set, the gradient-boosted model produced a markedly different ranking: it promoted periplasmic and ribosomal antigens that heuristic filters tend to deprioritize because they lack strong adhesion scores, yet which elicit robust T-cell responses when delivered with appropriate adjuvants[26]. Because no independent *Brucella* set with experimentally adjudicated protection outcomes was available to us, this comparison is reported as a shift in ranking rather than as a benchmark accuracy. This observation—that learned features capture protective potential invisible to heuristic filters—motivated the downstream validation of both a surface lipoprotein (Ag-03) and a ribosomal protein (Ag-07) rather than restricting selection to outer-membrane antigens alone.

Epitope landscape and construct design of lead antigens

From the 34 prioritized candidates we selected two lead antigens for multi-epitope construct design: Ag-03, a 28-kDa periplasmic substrate-binding lipoprotein with broad conservation and high adhesion scores, and Ag-07, a 24-kDa ribosome-associated immunogen previously catalogued among *Brucella* immunodominant proteins. Both proteins overlapped antigen families with established experimental protection data—Ag-03 shares structural homology with the Omp16/Omp19 lipoprotein family that confers S19-comparable protection in adjuvant[27], while Ag-07 belongs to the L7/L12 ribosomal family whose *Salmonella*-delivered form elicits protective responses in mouse models—providing biological plausibility for the classifier's selections.

Epitope mapping using NetMHCpan-4.1 and NetMHCIIpan-4.0 revealed that both leads carried dense clusters of predicted binders across HLA-A, HLA-B, and HLA-DR supertypes (Figures 2a and 2b). B-cell epitope prediction with BepiPred-3.0, which leverages protein language model embeddings identified 12 linear and 8 conformational epitopes on Ag-03 and 9 linear and 6 conformational epitopes on Ag-07 (Figure 2c). We assembled a multi-epitope construct by concatenating the highest-ranking $CD4^+$ and $CD8^+$ epitopes with

flexible GPGPG linkers and an N-terminal PADRE universal Thelper epitope, and a C-terminal His6 tag retained as a standard affinity handle should the construct be produced experimentally in future work. Population-coverage analysis projected that the selected epitope set would be presented by HLA molecules in > 94% of the global population, with coverage exceeding 90% in every major geographic region (Figure 2d)—a prerequisite for a vaccine intended for use across the diverse human and livestock populations of endemic regions. To validate the computational predictions against experimental ground truth, we cross-referenced the top-scoring MHC-II epitopes with the human $CD4^+$ T-cell epitope dataset mapped by Cannella and colleagues in Peruvian brucellosis patients. Six of the twelve predicted Ag-07 epitopes overlapped (defined as sharing at least five consecutive residues between the predicted epitope and the experimentally validated T-cell peptide) with peptides reported as T-cell recognized in the published IFN-γ ELISpot dataset[28]. This sequence-level cross-reference used only publicly available published peptide data; no new experimental data were generated, and no human or animal samples were handled in the present study. It should be noted that the Cannella dataset used *B. melitensis 16M* antigens to screen T cells from Peruvian brucellosis patients; since Ag-07 is conserved at > 95% amino-acid identity between *B. abortus* and *B. melitensis*, the predicted epitope sequences are identical to those of the *B. melitensis* homolog, ensuring species-appropriate comparison. Four Ag-03 epitopes mapped to regions of homologous proteins previously shown in published studies to elicit recall responses in vaccinated mice[27]. This concordance between *in silico* prediction and experimentally defined immunodominance provides retrospective confidence in the epitope-selection strategy and distinguishes the present construct from purely computational designs that lack epitope-level experimental cross-validation.

The assembled multi-epitope construct comprised 14 $CD4^+$ epitopes, 9 $CD8^+$ epitopes, and 4 B-cell epitope segments, totalling 384 residues including linkers and the PADRE pan-DR epitope. The complete construct sequence, including the exact amino-acid positions of all selected epitopes, linker sequences (GPGPG and AAY), the PADRE epitope sequence (AKFVAAWTLKAAA), the C-terminal hexahistidine (His6) tag, predicted MHC restriction alleles and corresponding prediction scores for each epitope, is available from the corresponding author on reasonable request to enable independent reproduction of the vaccine-design component. In silico cloning confirmed compatibility with the pET-28a expression vector, and codon adaptation yielded a codon adaptation index of 0.86 for *E. coli*. The construct was designed to avoid assembly of immunosuppressive motifs, glycosylation sites, and transmembrane segments that would compromise soluble expression, following design principles established for multi-epitope vaccines against other intracellular bacteria[29,30].

Structural validation confirms epitope accessibility and construct stability

To confirm that the predicted epitopes remain accessible on the folded antigen rather than buried in the protein core, we generated AlphaFold2 models of Ag-03 and Ag-07 and assessed predicted local distance difference test (pLDDT) confidence, solvent accessibility and conformational epitope exposure (Figure 3). Both models achieved mean pLDDT scores above 85, placing them in the high-confidence regime where backbone geometry is reliably predicted; per-residue confidence was lower across the linker and terminal segments of the assembled construct, as expected for a concatenated design. Solvent-accessible surface area calculations confirmed that 78% of the selected B-cell epitope residues on Ag-03 and 71% on Ag-07 were exposed, with the remainder located in loop regions likely to become accessible upon conformational breathing.

One-hundred-nanosecond all-atom molecular dynamics simulations of the multi-epitope constructs in explicit solvent confirmed structural stability: backbone RMSD plateaued below 2.1 Å for both constructs and radius of gyration remained within 3% of the equilibrated value, indicating no large-scale unfolding (Figure 3c). RMSF analysis showed that per-residue fluctuations remained below 2.5 Å for epitope regions, with higher flexibility confined to linker segments and terminal regions, consistent with the design intent of preserving epitope structural integrity. Hydrogen-bond occupancy analysis revealed stable inter-epitope hydrogen bonds

(average occupancy > 60% over the trajectory), supporting the rigidity of the construct core. Principal-component analysis (PCA) identified two dominant eigenvectors accounting for > 45% of total variance, both corresponding to collective motions of the linker regions rather than the epitope domains. DCC maps confirmed that intra-epitope residue pairs exhibited strong positive correlations ($r > 0.7$), while inter-epitope correlations were weak, indicating that each epitope domain behaved as an independent structural unit. MM-GBSA per-residue decomposition identified the epitope–linker junctions as the largest contributors to the total energetic stabilisation of the construct (Figure 3d), indicating that linker composition materially influences construct energetics. Physicochemical profiling showed both constructs to be within favorable ranges for stability, solubility, and antigenicity (Figure 3f), with no aggregation-prone stretches predicted by TANGO or FoldAmyloid.

Comparison of the AlphaFold2-predicted Ag-03 structure with the publicly available experimentally determined crystal structure of a homologous lipoprotein deposited in the Protein Data Bank (root-mean-square deviation 1.4 Å over 198 Cα atoms; the PDB accession of the reference structure is available from the corresponding author on reasonable request) provided external validation of the model quality, while the absence of a close structural homolog for Ag-07 meant its confidence rested on pLDDT and predicted aligned error alone. The conformational epitope density map (Figure 3d) revealed that epitope clusters on both antigens preferentially localized to loop regions and surface-exposed helices—topologies associated with high antibody accessibility in structural vaccinology studies. Notably, the multi-epitope construct retained the native-like secondary structure content of the parent antigens throughout the simulation, with helical and β-sheet fractions deviating by less than 4% from the equilibrated values, indicating that concatenation with flexible linkers did not disrupt the local folds required for epitope presentation. This structural integrity is a non-trivial requirement: earlier multi-epitope designs against *Brucella* reported construct aggregation and loss of conformational epitopes when rigid linkers were used, a failure mode our GPGPG-linker strategy was specifically designed to avoid.

Discussion

The central finding of this study is that a machine-learning-guided reverse vaccinology pipeline, grounded in the *Brucella* pan-genome, can prioritize a small set of conserved, structurally validated subunit antigen candidates from a candidate pool of over a thousand proteins. By extending the loop from computational prioritization through structural validation to multi-epitope construct assembly, we demonstrate that learned antigen classifiers do more than rank candidates—they recover proteins from antigen families with documented protective potential and convert them into a concrete, stable construct design amenable to experimental evaluation.

Several design choices distinguish this pipeline from prior *Brucella* immunoinformatics efforts. First, the use of a pan-genome rather than a single reference genome captures the accessory diversity responsible for strain-specific immune escape (Figure 1a), a lesson drawn from pan-genome reverse vaccinology successes in *Streptococcus*[31]. Second, replacing rule-based filtering with a gradient-boosted classifier trained on experimentally validated antigens[12] shifts the prioritization criterion from plausible antigenicity to predicted protectivity (Figures 2e and 2f)—a distinction that matters because many immunogenic *Brucella* proteins fail to translate into protection. Third, AlphaFold2 and molecular dynamics provide structural confidence that the selected epitopes are solvent-exposed and conformationally stable (Figure 3), addressing a common failure mode of purely sequence-based epitope vaccines in which predicted binders prove inaccessible on the native fold. Because no experimental data were generated in this study, the immune-profile comparisons drawn below are made against the published literature and are offered as design rationale rather than as validation of the present constructs.

The selection of Ag-07, a ribosome-associated immunogen, alongside the surface-exposed Ag-03 warrants mechanistic consideration given *Brucella*'s intracellular lifestyle. While ribosomal proteins are not surface-exposed on intact bacteria, several lines of evidence support their accessibility to the host immune system during *Brucella* infection. First, *Brucella* actively secretes a subset of ribosomal proteins—including L7/L12 family members—via the VirB type IV secretion system and

outer membrane vesicles, delivering them into the host cell cytosol where they become available for MHC-I antigen processing and cross-presentation. Second, during the transition from the replicative to the stationary phase, a fraction of intracellular brucellae undergo autolysis within the *Brucella*-containing vacuole, releasing ribosomal contents into the phagosomal lumen for MHC-II loading and lysosomal processing. Third, the L7/L12 ribosomal family to which Ag-07 belongs has been demonstrated to elicit protective CD4 and CD8 T-cell responses when delivered by attenuated *Salmonella* carriers, confirming that ribosomal antigens can access the MHC presentation pathway despite lacking surface exposure. The immunodominance of ribosomal proteins in other intracellular pathogens, including *Mycobacterium tuberculosis*, further supports the inclusion of Ag-07 as a T-cell-targeting component. Thus, while Ag-03 engages the antibody and CD4 arm through surface accessibility, Ag-07 complements the construct by recruiting CD8 cytotoxic T lymphocytes against intracellularly released ribosomal antigens.

The properties of the designed construct situate our candidates alongside leading experimental subunit strategies, although direct comparison of response magnitude is not possible because no immunological data were generated here. CpG-adjuvanted protein cocktails outperform Poly(I:C) formulations[32], calcium-phosphate nanoparticle-adsorbed antigens elicit cross-reactive responses against both *B. melitensis* and *B. abortus*, and nanoliposomal OMP19+L7/L12 formulations have been reported to reduce splenic burden beyond that achieved by Rev.1. Our contribution is not a single best-in-class antigen but a reproducible discovery framework: the same pipeline can be re-run as new genomes and epitope-prediction models accrue, and its structural-validation step can be applied to any candidate ranked by the classifier (Figure 2).

A defining practical advantage of the subunit approach over live vaccines is compatibility with DIVA surveillance. Because the multi-epitope construct lacks the smooth lipopolysaccharide O-polysaccharide that dominates the serological response to whole-cell *Brucella*, vaccinated animals should remain negative on the Rose Bengal test and complement fixation test used to certify brucellosis-free herds. The recent demonstration that RB51—the vaccine long marketed as DIVA-compatible—nevertheless interferes with serodiagnosis and confers inferior protection to S19 sharpens the case for subunit alternatives that are structurally incapable of generating anti-O-polysaccharide antibodies. Restricting candidates to the conserved core genome (Figure 1d) is itself a design dividend of pan-genome-informed antigen selection: it prioritizes epitopes shared across the three most pathogenic *Brucella* species rather than species-specific surface antigens that might elude heterologous challenge. This breadth stands in contrast to O-polysaccharide-based subunit strategies, which tend to be species-restricted because the O-antigen epitopes differ between *B. abortus* and *B. melitensis*.

Translation of these designs will require addressing several limitations. First and most importantly, this study is entirely computational: no protein was expressed, no animal was immunized and no protective efficacy was assessed. Challenge experiments remain the necessary next step. Every result reported here is therefore a prediction or a design property, not a measurement of immunogenicity or protection. Second, while the predicted HLA coverage of the epitope set exceeds 94% globally, presentation in outbred populations depends on HLA haplotype distribution and on the immunodominance hierarchy that emerges *in vivo*, which cannot be fully captured by binding-affinity prediction alone; the population-coverage estimate should be read as a design property of the epitope set rather than as a predicted vaccine efficacy. Third, the multi-epitope construct described here is a prototype design; optimization of adjuvant formulation—potentially combining CpG with nanoparticle delivery shown to enhance *Brucella* subunit performance—and of delivery route will be needed before any immune response can be expected to translate the Th1-skewed response into durable sterilizing immunity. Finally, although the predicted absence of anti-O-polysaccharide antibodies by design supports DIVA compatibility in principle, serological confirmation in vaccinated cattle—using the Rose Bengal and complement fixation tests under field conditions—remains an essential translational milestone.

Looking forward, the framework generalizes to other antigenically variable intracellular pathogens. The subtractive-proteomics and immunoinformatics logic

validated for *Salmonella* and the AI-driven multi-epitope paradigm demonstrated for tuberculosis share the same architecture: pan-genome curation, ML prioritization, epitope assembly and structural validation. As protein language models continue to improve epitope prediction and deep-learning structure prediction extends to complexes, the bottleneck will shift from candidate identification to experimental validation and formulation—precisely the loop this study was designed to close.

Several extensions of this framework merit consideration. First, the classifier is currently trained on bacterial antigens; expanding the training corpus to include viral and protozoan protective antigens would broaden its transferability, though at the risk of diluting pathogen-specific feature signatures. Second, the epitope-selection step optimizes for binding affinity and population coverage but does not yet model immunodominance hierarchies—the phenomenon whereby a subset of predicted binders dominates the *in vivo* T-cell response. Integrating recent neural-network models of immunodominance, once sufficiently validated, could further refine epitope choice. Third, the structural validation relies on monomer models; AlphaFold-Multimer and emerging diffusion-based methods could extend epitope-accessibility assessment to the oligomeric assemblies that some *Brucella* antigens form on the bacterial surface. Finally, although CpG is the adjuvant assumed in this design, nanoparticle delivery systems—calcium phosphate nanoparticles[33], glycine nanoparticles and nanoliposomes—have each demonstrated enhanced Th1 skewing and improved stability for *Brucella* subunit antigens, and a systematic head-to-head comparison of these platforms with the present construct would identify the optimal formulation for translation to the natural host.

Conclusion

In conclusion, this study demonstrates that a machine-learning-guided reverse vaccinology pipeline, grounded in pan-genome analysis, can efficiently prioritize *Brucella* subunit antigen candidates and convert them into structurally validated multi-epitope constructs. The two lead candidates, Ag-03 and Ag-07, are conserved across the three most pathogenic *Brucella* species, carry dense predicted T- and B-cell epitope clusters with broad projected HLA coverage, and remain structurally stable in simulation. All findings are computational: no protein expression, immunization or protection experiment was performed, and experimental validation remains the necessary next step. The framework is generalizable to other antigenically variable intracellular pathogens and provides a reproducible discovery platform that can be iteratively refined as new genomes and predictive models become available.

Author Contributions

Zhiheng Dong and Jiuwang Yu contributed equally to this work. Lidao Bao conceived and designed the study, provided resources, and supervised the work. Zhiheng Dong performed the computational analyses, and drafted the manuscript. Jiuwang Yu contributed to data analysis, figure preparation, and manuscript revision. Jiarong Guo provided technical support. Jiaxin Jiang and Lijuan Liang verified and proofread the references. All authors reviewed, edited, and approved the final version of the manuscript and agree to its publication.

Acknowledgments

This work was supported by Major Project of Natural Science Foundation of Inner Mongolia (2023 zd 15).

We would like to thank the Department of Precision Pharmacy, The Affiliated Hospital of Inner Mongolia Medical University for their valuable assistance with language editing and administrative support.

Data availability

The 90-genome accession list with assembly and isolation metadata, the complete training-dataset protein list with UniProt identifiers and protective/non-protective labels, the full epitope catalogue with amino-acid positions, predicted MHC restrictions, prediction scores, linker and tag sequences, the PDB accession of the reference structure used for structural comparison, the complete classifier performance metrics with confidence intervals, and the analysis code are available from the corresponding author on reasonable request. No supplementary files accompany this article. All datasets analysed in this study were obtained from publicly available databases or from published literature; no experimentally generated

data of any kind were produced in this work.

Ethical statement

This study is entirely computational and did not involve animals, human participants, or human or animal samples. All epitope data used for cross-referencing were obtained from published, publicly available sources.

AI statement

The authors used GPT-4 as a language-editing tool to improve grammar, clarity, and readability. The AI tool was not used for data analysis, data generation, image processing, or scientific content creation. All outputs were carefully reviewed and verified by the authors, who take full responsibility for the final content.

Conflict of Interest

All authors declare no commercial or financial relationships that could be construed as a potential conflict of interest regarding this study. The candidate antigens (Ag-03 and Ag-07) are currently in the basic research stage; no relevant patents have been applied for. The research funding was fully obtained from government sources (Major Project of Natural Science Foundation of Inner Mongolia, 2023 zd 15). Corresponding author Lidao Bao is responsible for supervision and project management, and there are no employment or equity interest disputes between him and the other authors.

References

1. Pappas G, Akritidis N, Bosilkovski M, Tsianos E. Brucellosis. *N Engl J Med*. 2005;352:2325-2336.
2. Dean AS, Crump L, Greter H, Schelling E, Zinsstag J. Global burden of human brucellosis: a systematic review of disease frequency. *PLoS Negl Trop Dis*. 2012;6:e1865.
3. Pappas G, Papadimitriou P, Akritidis N, Christou L, Tsianos EV. The new global map of human brucellosis. *Lancet Infect Dis*. 2006;6(2):91-99.
4. Godfroid J. Brucellosis in livestock and wildlife: zoonotic diseases without pandemic potential in need of innovative one health approaches. *Arch Public Health*. 2017;75:34.
5. Byndloss MX, Tsolis RM. *Brucella* spp. virulence factors and immunity. *Annu Rev Anim Biosci*. 2016;4:111-127.
6. López-Santiago R, Sánchez-Argáez AB, De Alba-Núñez LG, Baltierra-Uribe SL, Moreno-Lafont MC. Immune response to mucosal *Brucella* infection. *Front Immunol*. 2019; 10:1759.
7. Avila-Calderón ED, Lopez-Merino A, Sriranganathan N, Boyle SM, Contreras-Rodríguez A. A history of the development of *Brucella* vaccines. *Biomed Res Int*. 2013; 2013:743509.
8. Heidary M, Dashtbin S, Ghanavati R, et al. Evaluation of brucellosis vaccines: a comprehensive review. *Front Vet Sci*. 2022;9:925773.
9. Blasco JM, Moreno E, Muñoz PM, Conde-Álvarez R, Moriyón I. A review of three decades of use of the cattle brucellosis rough vaccine *Brucella abortus* RB51: myths and facts. *BMC Vet Res*. 2023;19(1):211.
10. Hisham Y, Ashhab Y. Identification of cross-protective potential antigens against pathogenic *Brucella* spp. through combining pan-genome analysis with reverse vaccinology. *J Immunol Res*. 2018;2018:1474517.
11. He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J Biomed Biotechnol*. 2010;2010:297505.
12. Ong E, Cooke MF, Huffman A, et al. Vaxign2: the second generation of the first web-based vaccine design program using reverse vaccinology and machine learning. *Nucleic Acids Res*. 2021;49(W1):W671-W678.
13. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*. 2017;45(W1):W24-W29.
14. Clifford JN, Høie MH, Deleuran S, Peters B, Nielsen M, Marcatili P. BepiPred-3.0: Improved B-cell epitope prediction using protein language models. *Protein Sci*. 2022;31(12):e4497.
15. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med*. 2016;8(1):33.
16. Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res*. 2020;48(W1):W449-W454.
17. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst*. 2018; 7(1):129-132.e4.
18. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021; 596(7873):583-589.
19. Hederman AP, Ackerman ME. Leveraging deep learning to improve vaccine design. *Trends Immunol*. 2023;44(5): 333-344.
20. Chen Z, Zhu Y, Sha T, et al. Design of a new multi-epitope vaccine against *Brucella* based on T and B cell epitopes using bioinformatics methods. *Epidemiol Infect*.

2021;149:e136.
21. Shi J, Zhu Y, Yin Z, et al. In silico designed novel multi-epitope mRNA vaccines against *Brucella* by targeting extracellular protein BtuB and LptD. *Sci Rep*. 2024;14(1):7278.
22. Li C, Zhu Y, Qi X, et al. Design of a multi-epitope recombinant BCG vaccine targeting *Brucella* OMP31, LptE and VirB2 in immunoinformatics approaches. *PLoS One*. 2025;20(11):e0334843.
23. Doostmohammadi A, Samimi Hashjin A, Shahbazi B, et al. Phylogeny-guided multi-epitope vaccine design targeting conserved and divergent regions of *Brucella* OMP2, OMP25, and OMP31: an immunoinformatics study. *BMC Immunol*. 2026;27(1):19.
24. Biswas R, Sinha SS, Roy A, Ramaiah S, Anbarasu A. An integrated subtractive genomics and immunoinformatics approach for designing a universal multi-epitope vaccine against *Brucella* spp. *Front Bioinform*. 2026;6:1818265.
25. Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins*. 2006;65(1):40-48.
26. Senevirathne A, Hewawaduge C, Lee JH. Live vaccine consisting of attenuated Salmonella secreting and delivering *Brucella* ribosomal protein L7/L12 induces humoral and cellular immune responses and protects mice against virulent *Brucella abortus* 544 challenge. *Vet Res*. 2020;51(1):6.
27. Pasquevich KA, Estein SM, García Samartino C, et al. Immunization with recombinant *Brucella* species outer membrane protein Omp16 or Omp19 in adjuvant induces specific $CD4^+$ and $CD8^+$ T cells as well as systemic and oral protection against *Brucella abortus* infection. *Infect Immun*. 2009;77(1):436-445.
28. Cannella AP, Arlehamn CS, Sidney J, et al. *Brucella* melitensis T cell epitope recognition in humans with brucellosis in Peru. *Infect Immun*. 2014;82(1):124-131.
29. Zhuang L, Ali A, Yang L, et al. Leveraging computer-aided design and artificial intelligence to develop a next-generation multi-epitope tuberculosis vaccine candidate. *Infect Med (Beijing)*. 2024;3(4):100148.
30. Bhattacharjee A, Hosen MR, Lamisa AB, et al. An integrated comparative genomics, subtractive proteomics and immunoinformatics framework for the rational design of a Pan-Salmonella multi-epitope vaccine. *PLoS One*. 2024;19(7):e0292413.
31. Maione D, Margarit I, Rinaudo CD, et al. Identification of a universal group B streptococcus vaccine by multiple genome screen. *Science*. 2005;309(5731):148-150.
32. Golshani M, Amani M, Siadat SD, Nejati-Moheimani M, Arsang A, Bouzari S. Comparison of the protective immunity elicited by a *Brucella* cocktail protein vaccine (rL7/L12+rTOmp31+rSOmp2b) in two different adjuvant formulations in BALB/c mice. *Mol Immunol*. 2018;103:306-311.
33. Sadeghi Z, Fasihi-Ramandi M, Bouzari S. Nanoparticle-based vaccines for brucellosis: calcium phosphate nanoparticles-adsorbed antigens induce cross protective response in mice. *Int J Nanomedicine*. 2020;15:3877-3886.